Statistical Issues in Modelling the Ancestry from Y-Chromosome and Surname Data

by

Maarya Sharif M.Sc., B.Sc. (Hons)

A thesis submitted in fulfillment for the
degree of Doctor of Philosophy

in the

College of Science and Engineering
School of Mathematics and Statistics
University of Glasgow

May 2012

Abstract

A considerable industry has grown-up around genealogical inference from genetic testing, supplementing more traditional genealogical techniques but with very limited quantification of uncertainty. In many societies Y-chromosomes are co-inherited with surnames and as such passed down from father to son. This thesis seeks to explore what the correlation can say about ancestry. In particular it is concerned with estimation of the time to the most recent common paternal ancestor (TMRCA) for pairs of males who are not known to be directly related but share the same surname, based on the repeat number at short tandem repeat (STR) markers on their Y-chromosomes.

We develop a model of TMRCA estimation based on the difference in repeat numbers in pairs of male haplotypes using a Bayesian framework and Markov-Chain Monte-Carlo techniques, such as adaptive Metropolis-Hastings algorithm. The model incorporates the process of STR discovery and the calibration of mutation rates, which can differ across STRs. In simulation studies, we find that the estimates of TMRCA are rather robust to the ascertainment process and the way in which it is modelled. However, they are affected by the site-specific mutation rates at the typed STRs. Indeed sequencing the fastest mutating STRs yields a lower error in the estimated TMRCA than random STRs. In the British context, we extend our model to include additional information such as the haplogroup status (as determined from single nucleotide polymorphisms, SNPs) of the pair of males, as well as the frequency and origin of the surname. In general, the effect of this is to reduce estimates of the TMRCA for pairs of males with an older TMRCA, typically outwith the period of surname establishment (about 500-700 years ago). In the genealogical context, incorporating surname frequency (within the prior distribution) results in lower estimates of TMRCA for pairs of males who appear to have diverged from a common male ancestor since the period of surname establishment. In addition, we include uncertainty in the years per generation conversion factor in our model.

Keywords: Y-chromosome, surname, most recent common ancestor, haplotype, haplogroup, British, genealogy, short tandem repeat, generation.

Acknowledgements

Praise and thanks be to Allah, Lord of the Worlds, with a praise that is adequate to His favours and equal to His increase. Any mistakes that follow are mine and the praise for the merits of this thesis belongs to Allah.

There are many people whom I wish to thank for their support of my 'uni work'. To my supervisor, Dr Vincent Macaulay, who sat through my ramblings (often times accompanied by general hand-waving), patiently taught me the art of debugging code and was always kind in his supervision: you have enriched my research experience and have set the foundations for what I hope will be a fulfilling career in academia. For the DTA Ph.D. studentship, I am indebted to the Engineering and Physical Sciences Research Council. I am grateful for the support from the School of Mathematics and Statistics and *all* its staff. It has been a pleasure to study in such a friendly and understanding environment. A particular thank you also to Dr Agostino Nobile and Beverley Dixon for their assistance over the years. To the postgrads: I take it back, childbirth is far more painful! And to Joanna: the notepad proved invaluable.

To all my family, may Allah reward you for all your prayers and encouragement. For the free child care, my sincere gratitude to my Mum, Auntie Sharifa, Taee-Amma and the numerous others. To my beloved children, Noor and Zayn, I missed you too because I didn't see you for a long time. Finally to Fahad, thank you for helping with the laundry, wiping away my tears and believing in me. I couldn't have done it without you.

Contents

Abstract iii

Acknowledgements iv

List of Tables ix

List of Figures xi

Statistical Distributions xvii

Abbreviations xix

1 Introduction 1
- 1.1 Context . . . 1
  - 1.1.1 Genetics . . . 1
  - 1.1.2 Genetic typing of STRs . . . 5
  - 1.1.3 Genealogy . . . 11
  - 1.1.4 Surnames . . . 12
  - 1.1.5 Surname-based Genetic Genealogy . . . 15
- 1.2 Y-Chromosome Databases . . . 20
- 1.3 Bayesian Inference . . . 23

2 Background 25
- 2.1 Estimate of Time to the Most Recent Common Ancestor . . . 25
- 2.2 Y-chromosome and Surnames . . . 37
- 2.3 Estimation of STR Mutation Rates . . . 47
  - 2.3.1 Initial Mutation Rate Estimates . . . 49
  - 2.3.2 Intermediary Mutation Rate Estimates . . . 50
  - 2.3.3 Final Mutation Rate Estimates . . . 51

3 Maximum Likelihood Estimation of TMRCA 61
- 3.1 Preliminary Analysis . . . 61
  - 3.1.1 Materials and Methods . . . 61

3.1.2 Preliminary Results . . . 65
3.1.3 Preliminary Conclusions . . . 70
3.2 Development . . . 70
3.3 Materials and Methods . . . 71
3.3.1 Stepwise Mutational Model . . . 71
3.3.2 Infinite Sites Model . . . 71
3.3.3 Data Simulation . . . 72
3.4 Results . . . 75
3.4.1 Mutation Rate Distribution 1 . . . 75
3.4.1.1 Stepwise Mutation Model Analysis 1 . . . 75
3.4.1.2 Infinite Sites Model Analysis 1 . . . 79
3.4.2 Mutation Rate Distribution 2 . . . 80
3.4.2.1 Stepwise Mutation Model Analysis 2 . . . 80
3.4.2.2 Infinite Sites Model Analysis 2 . . . 84
3.4.3 Mutation Rate Distribution 3 . . . 85
3.4.3.1 Stepwise Mutation Model Analysis 3 . . . 87
3.4.3.2 Infinite Sites Model Analysis 3 . . . 89
3.5 Discussion . . . 91
3.6 Conclusions . . . 97

4 Modelling Mutational Mechanisms 99
4.1 Preliminary Analysis . . . 99
4.1.1 Discussions . . . 103
4.1.2 Conclusions . . . 105
4.2 Introduction . . . 105
4.3 Materials and Methods . . . 106
4.3.1 Data Simulation . . . 106
4.3.2 Bayesian Modelling . . . 109
4.3.3 Markov-Chain Monte-Carlo Sampling . . . 110
4.3.3.1 Metropolis-Hastings Algorithm . . . 111
4.3.3.2 Adaptive MCMC . . . 113
4.3.3.3 Implementation: Modelling Mutational Mechanisms 114
4.3.4 Data Simulation and Analysis Program . . . 116
4.3.5 Software . . . 119
4.3.6 Analysis . . . 119
4.4 Results . . . 122
4.4.1 Real Data . . . 122
4.4.2 Simulated Data Analysis . . . 128
4.4.2.1 Varying the Information . . . 130
4.4.2.2 Misspecification . . . 130
4.5 Discussion . . . 134
4.6 Conclusions . . . 139

5 Modelling TMRCA 141

5.1 Introduction . . . 141
5.2 Materials and Methods . . . 142
5.2.1 Simulation Study . . . 142
5.2.2 Bayesian Modelling . . . 144
5.2.3 Markov-Chain Monte-Carlo Sampling . . . 146
5.2.3.1 Implementation: TMRCA Model . . . 147
5.2.4 Data Simulation and Analysis Program . . . 149
5.2.5 Software . . . 152
5.2.6 Analysis . . . 152
5.3 Results . . . 153
5.3.1 Varying the Proportion of Typed STRs . . . 153
5.3.2 Use of Fast Mutating STRs . . . 161
5.3.3 Misspecification of the Proportion of Non-Ascertained Loci . 162
5.4 Discussion . . . 169
5.5 Conclusions . . . 171

6 TMRCA Estimation: Real Data Applications 173
6.1 Haplogroup- and Surname-Based Priors . . . 173
6.1.1 Materials and Methods . . . 173
6.1.2 Same haplogroup . . . 178
6.1.2.1 Surname Origins and Frequency . . . 178
6.1.2.2 Surname Frequency . . . 183
6.1.2.3 Surname Origins . . . 184
6.1.2.4 No Frequency or Surname Origin . . . 187
6.1.3 Random/No Haplogroup . . . 189
6.1.3.1 Surname Origin and Frequency . . . 189
6.1.3.2 Surname Frequency . . . 191
6.1.3.3 Surname Origins . . . 193
6.1.3.4 No Frequency or Surname Origin . . . 195
6.1.4 Different Haplogroup . . . 197
6.1.4.1 Surname Origin and Frequency . . . 197
6.1.4.2 Surname Origin . . . 197
6.1.5 Discussion . . . 198
6.2 Additional Priors . . . 207
6.2.1 Absence of Surname and Haplogroup Information . . . 207
6.2.2 Generation Time in Years . . . 207
6.3 Assessment of Priors on TMRCA . . . 209
6.3.1 Materials and Methods . . . 209
6.3.2 Same haplogroup . . . 212
6.3.3 Random Haplogroup . . . 213
6.3.4 Different Haplogroup . . . 214
6.3.5 Discussion . . . 215
6.4 Conclusions . . . 217

7 Discussion **219**

8 Conclusions **231**

A The Delta Method **235**

Bibliography **237**

List of Tables

1.1 STR data . . . 18
1.2 Y-chromosome testing companies . . . 19
1.3 Y-DNA databases . . . 22

2.1 Sources for initial mutation rate review . . . 50
2.2 Initial mutation rate estimates . . . 50
2.3 Sources for intermediary mutation rate review . . . 51
2.4 Additional sources for final mutation rate review . . . 51
2.5 Final mutation rate estimates . . . 53
2.6 STR markers one proportion test results . . . 54
2.7 Summary of STR markers properties . . . 56
2.8 Multi-step mutations . . . 58

3.1 Mutation rate distribution 1: ISM FSE < SMM FSE using true and empirical site-specific mutation rates . . . 93
3.2 Mutation rate distribution 2: ISM FSE < SMM FSE using true and empirical site-specific mutation rates . . . 94
3.3 Mutation rate distribution 3: ISM FSE < SMM FSE using true and empirical site-specific mutation rates . . . 94

4.1 Empirical and Bayesian mutation rate estimates . . . 123
4.2 Real data Bayesian parameter estimates . . . 124
4.3 Regression of posterior mean against percentage of non-ascertained loci . . . 128

6.1 Model for validating ANCOVA permutation test . . . 177
6.2 Simulated data error rates . . . 177
6.3 Same haplogroup: surname origin and frequency fitted model parameters . . . 180
6.4 Same haplogroup: surname origin and frequency permutation test p-values . . . 181
6.5 Same haplogroup: significant Tukey multiple comparisons between surname origin . . . 185
6.6 Same haplogroup: exponential surname origin fitted model parameters . . . 186
6.7 Same haplogroup: gamma surname origin fitted model parameters . 186

6.8 Thinned same haplogroup: exponential surname origin fitted model parameters . . . 187
6.9 Random haplogroup: permutation test p-values . . . 190
6.10 Random haplogroup: fitted surname origin and frequency model parameters . . . 190
6.11 Random haplogroup: significant Tukey multiple comparisons between surname origin . . . 194
6.12 Random haplogroup: exponential surname origin fitted model parameters . . . 194
6.13 Random haplogroup: gamma surname origin fitted model parameters 195
6.14 Thinned random haplogroup: exponential surname origin fitted model parameters . . . 195
6.15 Different haplogroup: significant Tukey multiple comparisons between surname origin . . . 198
6.16 Same haplogroup: permutation test p-values . . . 199
6.17 Thinned same haplogroup: gamma surname origin fitted model parameters . . . 199
6.18 Thinned random haplogroup: gamma surname origin fitted model parameters . . . 201
6.19 Published estimates of the ages of nodes AT and BT . . . 205
6.20 Different haplogroup: prior parameters . . . 205
6.21 TMRCA Priors for pairs of males . . . 210
6.22 Rate parameters for priors 3 and 6 . . . 211
6.23 Parameter values for priors 4, 7 and 8 . . . 211

List of Figures

1.1 Karyotype of human male . . . 2
1.2 SNP and STR highlighted in Y-chromosome . . . 3
1.3 Non-recombining portion of the Y-chromosome . . . 3
1.4 Recombination of the sex chromosomes . . . 4
1.5 Y-chromosome SNP tree . . . 5
1.6 Y-chromosome R1 haplogroup tree . . . 6
1.7 STR primer . . . 7
1.8 Gel electropherogram . . . 7
1.9 CE multiplex PCR results . . . 8
1.10 Promega PowerPlex ®Y allelic ladders . . . 8
1.11 Null allele . . . 9
1.12 STR copy types . . . 9
1.13 STRs complexity types . . . 10
1.14 Family tree . . . 11
1.15 Surname-haplotype transmission . . . 15
1.16 Male lineage . . . 17
1.17 Time to MRCA . . . 17

2.1 The time to the most recent common ancestor . . . 25
2.2 Parallel mutation . . . 26
2.3 Back mutation on male 2 lineage . . . 27
2.4 Directionality of SMM . . . 29
2.5 Random walk of an STR . . . 29
2.6 Schematic of methods for estimating mutation rates . . . 48
2.7 Histogram of mutation rates . . . 52
2.8 Estimated mutation rate vs. STR marker . . . 54
2.9 Estimated mutation rate vs. proportion of increase mutations . . . 55
2.10 Estimated mutation rate by length of STR repeat unit . . . 57
2.11 Estimated mutation rate by STR complexity . . . 58
2.12 Estimated mutation rate vs. size and direction of multi-step mutation 59
2.13 Proportion of increase mutations vs. size and direction of multi-step mutation . . . 60

3.1 TMRCA estimates: SMM vs. IAM . . . 66
3.2 SMM and IAM: TMRCA estimates vs. log surname frequency . . . 66

3.3 SMM haplogroup status: TMRCA estimates vs. log surname frequency . . . 67
3.4 IAM haplogroup status: TMRCA estimates vs. log surname frequency 67
3.5 TMRCA estimates: site-specific rates vs. average rate . . . 68
3.6 Site-specific and average mutation rates: TMRCA estimates vs. log surname frequency . . . 69
3.7 Site-specific mutation rates haplogroup status: TMRCA estimates vs. log surname frequency . . . 69
3.8 Mutation rates densities . . . 73
3.9 SMM mutation rate distribution 1: FSE and FV of $\hat{t}$ vs. number of loci . . . 76
3.10 SMM mutation rate distribution 1: FBSQ of $\hat{t}$ vs. number of loci . . 78
3.11 SMM mutation rate distribution 1: FB of $\hat{t}$ vs. number of loci . . . 78
3.12 ISM mutation rate distribution 1: FSE and FV of $\hat{t}$ vs. number of loci . . . 80
3.13 ISM mutation rate distribution 1: FBSQ of $\hat{t}$ vs. number of loci . . 81
3.14 ISM mutation rate distribution 1: FB of $\hat{t}$ vs. number of loci . . . 81
3.15 SMM mutation rate distribution 2: FSE and FV of $\hat{t}$ vs. number of loci . . . 82
3.16 SMM mutation rate distribution 2: FBSQ of $\hat{t}$ vs. number of loci . . 83
3.17 SMM mutation rate distribution 2: FB of $\hat{t}$ vs. number of loci . . . 83
3.18 ISM mutation rate distribution 2: FSE and FV of $\hat{t}$ vs. number of loci . . . 85
3.19 ISM mutation rate distribution 2: FBSQ of $\hat{t}$ vs. number of loci . . 86
3.20 ISM mutation rate distribution 2: FB of $\hat{t}$ vs. number of loci . . . 86
3.21 SMM mutation rate distribution 3: FSE and FV of $\hat{t}$ vs. number of loci . . . 87
3.22 SMM mutation rate distribution 3: FBSQ of $\hat{t}$ vs. number of loci . . 88
3.23 SMM mutation rate distribution 3: FB of $\hat{t}$ vs. number of loci . . . 88
3.24 ISM mutation rate distribution 3: FSE and FV of $\hat{t}$ vs. number of loci . . . 89
3.25 ISM mutation rate distribution 3: FBSQ of $\hat{t}$ vs. number of loci . . 90
3.26 ISM mutation rate distribution 3: FB of $\hat{t}$ vs. number of loci . . . 91
3.27 SMM mutation rate distribution 1: FB of $\hat{t}$ vs. number of loci . . . 92
3.28 SMM mutation rate distribution 2: FB of $\hat{t}$ vs. number of loci . . . 92
3.29 SMM mutation rate distribution 3: FB of $\hat{t}$ vs. number of loci . . . 92
3.30 Comparison of SMM and ISM estimates of true mutation rates . . . 95
3.31 Comparison of SMM and ISM estimates of empirical mutation rates 95
3.32 SMM: FSE and FV of $\hat{t}$ v. number of meioses . . . 96
3.33 SMM: FB of $\hat{t}$ v. number of meioses . . . 97

4.1 Schematic of USMM . . . 100
4.2 Simulation: USMM and SMM data . . . 104
4.3 Categories of markers in the mutational mechanisms model . . . 108

4.4 Comparison of Bayesian and empirical estimates of (empirically) variable mutation rates . . . 122
4.5 Comparison of Bayesian and empirical estimates of (empirically) non-variable mutation rates . . . 124
4.6 MCMC graphical output using real data . . . 125
4.7 Posterior mean α and β vs. percentage of non-ascertained loci . . . 126
4.8 Posterior mean of gamma mean and variance vs. percentage of non-ascertained loci . . . 126
4.9 Posterior mean L and N_e vs. percentage of non-ascertained loci . . . 126
4.10 Average posterior mean mutation rate vs. percentage of non-ascertained loci . . . 127
4.11 Posterior mean α and β vs. percentage of non-ascertained loci . . . 129
4.12 Posterior mean of gamma mean and variance vs. percentage of non-ascertained loci . . . 129
4.13 Posterior mean L and N_e vs. percentage of non-ascertained loci . . . 129
4.14 Posterior mean α and β vs. percentage of calibrated loci . . . 131
4.15 Posterior mean of gamma mean and variance vs. percentage of calibrated loci . . . 131
4.16 Posterior mean L and N_e vs. percentage of calibrated loci . . . 131
4.17 Posterior mean α and β vs. percentage of non-ascertained loci . . . 133
4.18 Posterior mean of gamma mean and variance vs. percentage of non-ascertained loci . . . 133
4.19 Posterior mean L and N_e vs. percentage of non-ascertained loci . . . 133
4.20 Simulated data: average posterior mean mutation rate vs. percentage of non-ascertained loci . . . 134
4.21 Intermediary mutation rate review: histogram of the number of meioses per locus . . . 135
4.22 Simulated data 2: average posterior mean mutation rate vs. percentage of non-ascertained loci . . . 136
4.23 Simulated data 3: average posterior mean mutation rate vs. percentage of non-ascertained loci . . . 138

5.1 The time to the MRCA . . . 141
5.2 Categories of markers in the TMRCA model . . . 142
5.3 Fractional squared error and variance of $\hat{t}$ vs. percentage of typed loci 154
5.4 Fractional bias of $\hat{t}$ vs. percentage of typed loci . . . 154
5.5 Mean squared error and variance of $\hat{\alpha}$ vs. percentage of typed loci . 156
5.6 Bias of $\hat{\alpha}$ vs. percentage of typed loci . . . 156
5.7 Mean squared error and variance of $\hat{\beta}$ vs. percentage of typed loci . 156
5.8 Bias of $\hat{\beta}$ vs. percentage of typed loci . . . 157
5.9 Mean squared error and variance of gamma mean vs. percentage of typed loci . . . 157
5.10 Bias of gamma mean vs. percentage of typed loci . . . 157
5.11 Mean squared error and variance of gamma variance vs. percentage of typed loci . . . 159

5.12 Bias of gamma variance vs. percentage of typed loci 159
5.13 squared error of $\hat{L}$ vs. percentage of typed loci 159
5.14 Bias of $\hat{L}$ vs. percentage of typed loci 160
5.15 Mean squared error and variance of $\hat{N}_e$ vs. percentage of typed loci 160
5.16 Bias of $\hat{N}_e$ vs. percentage of typed loci 160
5.17 Fractional squared error and variance of $\hat{t}$ vs. percentage of typed loci 161
5.18 Fractional bias of $\hat{t}$ vs. percentage of typed loci 161
5.19 Fractional squared error and variance of $\hat{t}$ vs. percentage of non-ascertained loci 163
5.20 Fractional bias of $\hat{t}$ vs. percentage of non-ascertained loci 163
5.21 Mean squared error and variance of $\hat{\alpha}$ vs. percentage of non-ascertained loci 164
5.22 Bias of $\hat{\alpha}$ vs. percentage of non-ascertained loci 164
5.23 Mean squared error and variance of $\hat{\beta}$ vs. percentage of non-ascertained loci 164
5.24 Bias of $\hat{\beta}$ vs. percentage of non-ascertained loci 165
5.25 Mean Squared Error and Variance of Gamma mean vs. percentage of non-ascertained loci 165
5.26 Bias of Gamma mean vs. percentage of non-ascertained loci 165
5.27 Mean squared error and variance of gamma variance vs. percentage of non-ascertained loci 167
5.28 Bias of gamma variance vs. percentage of non-ascertained loci . . . 167
5.29 Mean squared error and variance of $\hat{L}$ vs. percentage of non-ascertained loci 167
5.30 Bias of $\hat{L}$ vs. percentage of non-ascertained loci 168
5.31 Mean squared error and variance of $\hat{N}_e$ vs. percentage of non-ascertained loci 168
5.32 Bias of $\hat{N}_e$ vs. percentage of non-ascertained loci 168
5.33 Histogram of percentage of non-ascertained loci 169
5.34 Fast and random markers: Fractional squared error and variance of $\hat{t}$ vs. percentage of typed loci 170
5.35 Fast and random markers: Fractional bias of $\hat{t}$ vs. percentage of typed loci 170

6.1 Same haplogroup: Box-Cox TMRCA vs. log surname frequency . . 179
6.2 Same haplogroup: mean Box-Cox TMRCA vs. log surname frequency 179
6.3 Same haplogroup: a. residuals vs. fitted values b. normal Q-Q plot of residuals 181
6.4 Same haplogroup: surname origin and frequency fitted model 182
6.5 Same haplogroup: standard deviation for patronymic TMRCA surname origin and frequency priors 182
6.6 Same haplogroup: surname frequency fitted model 183
6.7 Same haplogroup: standard deviation for surname frequency prior . 184
6.8 Same haplogroup: boxplot of Box-Cox TMRCA by surname origin . 185

6.9 Same haplogroup: a. boxplot of TMRCA b. boxplot of log TMRCA c. histogram of log TMRCA . . . 187
6.10 Ten thinned same haplogroup samples . . . 188
6.11 Same haplogroup: comparison of exponential, gamma and half-normal distributions . . . 188
6.12 Random haplogroup: mean Box-Cox estimated TMRCA vs. log surname frequency . . . 189
6.13 Random haplogroup: fitted surname origin and frequency main effects model . . . 191
6.14 Random haplogroup: a. residuals vs. fitted values b. normal Q-Q plot of residuals . . . 192
6.15 Random haplogroup: surname frequency fitted model . . . 192
6.16 Random haplogroup: standard deviation for surname frequency prior 193
6.17 Random haplogroup: boxplot of Box-Cox TMRCA vs. surname origin . . . 193
6.18 Random haplogroup: histogram of TMRCA . . . 196
6.19 Random haplogroup: fitted densities of random samples with full data density . . . 196
6.20 Different haplogroup: boxplot of TMRCA by surname origin . . . 198
6.21 Distribution of random haplogroup for same and different haplogroup status . . . 200
6.22 Distribution of random haplogroup for same and different haplogroup status within surname origin . . . 201
6.23 Y-chromosome SNP tree . . . 202
6.24 Different haplogroup: boxplot of TMRCA by haplogroup node . . . 203
6.25 Years per generation distribution . . . 209
6.26 Same haplogroup TMRCA (new prior) vs. TMRCA (standard prior) 212
6.27 Random haplogroup TMRCA (new prior) vs. TMRCA (standard prior) . . . 213
6.28 Random haplogroup TMRCA (prior 4) vs. TMRCA (standard prior) 214
6.29 Different Haplogroup: TMRCA (new prior) vs. TMRCA (standard prior) . . . 215
6.30 Different haplogroup: TMRCA ancestral node vs. TMRCA (prior 9) 215
6.31 Different haplogroup TMRCA (new prior) vs. TMRCA (standard prior) . . . 216
6.32 Different haplogroup (complete data): TMRCA (new prior) vs. TMRCA (standard prior) . . . 216

7.1 Real and simulated data comparison: meioses vs. empirical mutation rate . . . 220
7.2 Fractional squared error and variance of $\hat{t}$ vs. percentage of typed loci: a. real meioses b. multi-stage calibration meioses . . . 222
7.3 Bias of $\hat{t}$ vs. percentage of typed loci: a. real meioses, b. multi-stage calibration meioses . . . 222
7.4 Comparison of mutation rate distributions . . . 223

7.5 Comparison of prior distribution on N_e 226

Statistical Distributions

Binomial

$$X \sim Bi(n, \theta): \qquad p_X(x) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Exponential

$$X \sim Ex(\beta): \qquad p_X(x) = \frac{e^{-x/\beta}}{\beta}$$

$$X \sim Ex(\theta): \qquad p_X(x) = \theta e^{-\theta x}$$

Gamma

$$X \sim Ga(\alpha, \beta): \qquad p_X(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)}$$

$$X \sim Ga(k, \theta): \qquad p_X(x) = \frac{x^{k-1} e^{-\theta x} \theta^k}{\Gamma(k)}$$

Half-Normal

$$X \sim HN(\theta): \qquad p_X(x) = \frac{2\theta}{\pi} \exp\left(-\frac{x^2\theta^2}{\pi}\right)$$

Normal

$$X \sim N(\mu, \sigma^2): \qquad p_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

Poisson

$$X \sim Po(\lambda): \qquad p_X(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Abbreviations

AMOVA	Analysis of molecular variance
ANOVA	Analysis of variance
BCE	Before the common era
BP	Before present
CE	Common era
FB	Fractional bias
FSE	Fractional squared error
FV	Fractional variance
IAM	Infinite alleles model
IMH	Irish modal haplotype
ISM	Infinite sites model
MSE	Mean squared error
MJ	Median joining
NPT	Non-paternity transmissions
NSG	Norse surname group
MRCA	Most recent common ancestor
RCC	Revised correlation coefficient
SMM	Stepwise Mutation Model
STR	Short tandem repeat
SNP	Single nucleotide polymorphism
TMRCA	Time to a most recent common ancestor
USMM	Unequal stepwise mutation model
YHRD	Y-chromosome haplotype reference database
YPG	Years per generation

Dedicated to Amma Meena

Chapter 1

Introduction

"Where do we come from?"

It is a question that has been asked throughout the ages whether on philosophical, religious, scientific or even artistic grounds. Modern man seems no less interested in this question, but may often seek a genealogical response in order to learn about their ancestry. In many cultures there is a correlation between Y-chromosomes with surnames allowing the inference of the time to a most recent common ancestor (TMRCA) for pairs of males. This thesis aims to develop a model for the estimation of TMRCA and explore the factors which affect this parameter.

1.1 Context

1.1.1 Genetics

The genetic make-up of humans consists of 23 pairs of chromosomes (fig. 1.1); 22 pairs of homologous chromosomes called autosomes and a pair of sex chromosomes which define the sex of an individual. Hence females possess two X chromosomes (XX) whilst males have one X chromosome and one Y chromosome (XY) thus making the Y-chromosome exclusively male. Additionally there are organelles called mitochondria, which help provide energy for the cell they are within. Their DNA is referred to as mitochondrial DNA (mtDNA). In contrast to DNA from the Y-chromosome (Y-DNA), which is passed down the paternal line, mtDNA is maternally inherited. Y-DNA will be the focus of this thesis.

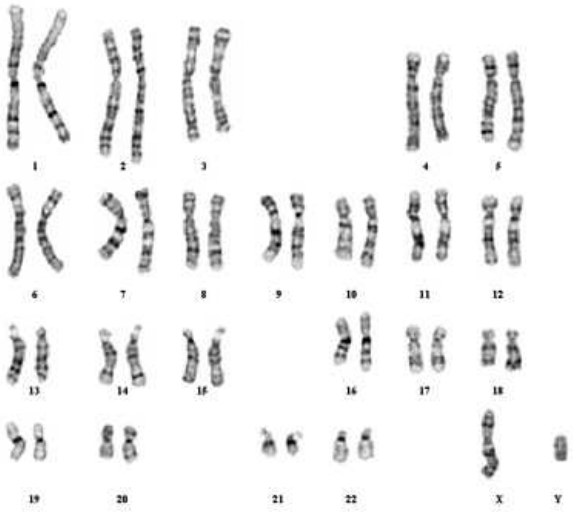

FIGURE 1.1: Karyotype of human male (Martin, 2003)

The double helix structure in a single chromosome may be represented by two rows (or 'strands') of the letters A, C, G and T, respectively the nucleotides adenine, cytosine, guanine and thymine, with the base pairs A with T always appearing together on the two strands and also C with G, i.e. the first row is complementary to the second row of letters. Note usually only one strand of each chromosome is typed or sequenced and the other is known due to the complementary base pairing. In figure 1.2 we use this representation to depict a short length of Y-DNA from two different males and highlight two types of markers, namely a microsatellite and a single nucleotide polymorphism (SNP, pronounced as "snip"), which are both variable at the population level. The SNP is the single position highlighted in blue in figure 1.2, i.e. a difference 10 bases from the left where male 1 possesses A-T, whilst male 2 has G-C. SNPs are usually bi-variate e.g. there are two forms ('alleles') A and G on the typed strand, which due to the double helix structure are equivalent to T and C on the complementary strand. The longer length microsatellite or short tandem repeat (STR) is shown in red, with male 2 possessing seven repeats of TAG while male 1 only has five repeats.

In general STRs are DNA sequences that involve up to 50 repeats of sequences of length 2-6 base pairs, e.g. GAT_n would mean n repeats of the bases GAT. These were discovered in humans in 1989, although identified as a subset of variable number of tandem repeats (VNTRs) found in 1985 (Butler, 2005; Jobling et al., 2004). SNPs were also discovered during the 1980s as biallelic restriction fragment length polymorphisms (RFLPs) (Collins et al., 1999). A SNP is simply a single base polymorphism due to a base substitution or the insertion or deletion of a single base (Jobling et al., 2004). Crucially SNPs have a much lower average mutation rate

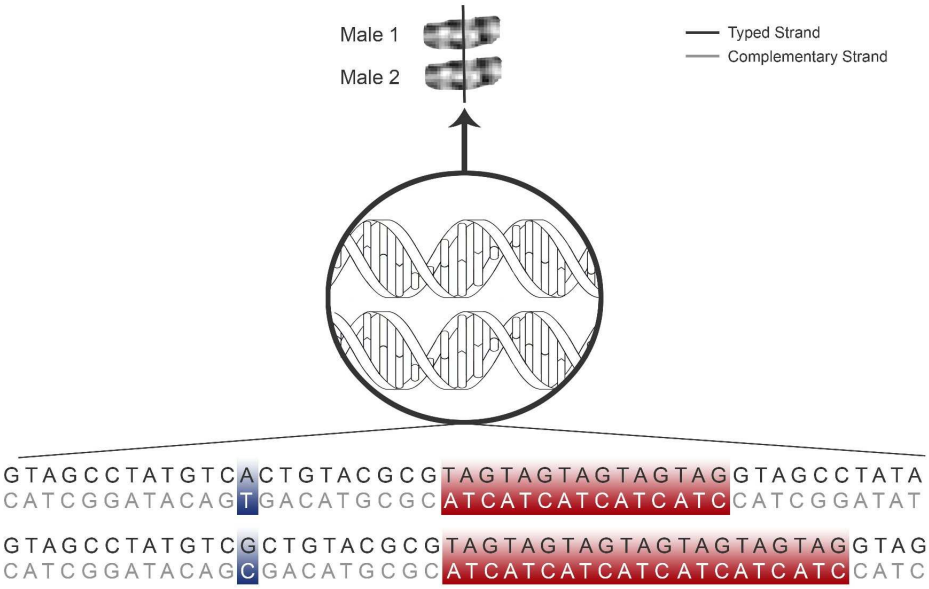

FIGURE 1.2: DNA sequence with SNP and STR highlighted in Y-chromosome. Adapted from Sharif (2007)

than STRs, 2.5×10^{-8} versus 2×10^{-3} mutations per marker per generation (Jobling et al., 2004), an aspect that will be discussed and examined in depth in this thesis.

The Y-chromosome is one of the smallest chromosomes composed of an average 60 million base pairs (Jobling et al., 2004) which importantly contains a non-recombining portion (NRY, fig. 1.3). To appreciate the significance of this it must be realised that, of the two copies of each chromosome in humans, one has been transmitted from the mother via her egg and the other from the father via his sperm. In forming the egg and sperm cells, the parents' normal cells undergo meioses during which recombination may occur. It is essentially the process by which segments of DNA are exchanged between homologous pairs of chromosome and also the sex chromosomes as shown in figure 1.4. This occurs in every generation and more often at certain places along chromosomes called recombination hotspots (Jobling et al., 2004). Importantly however in the case of

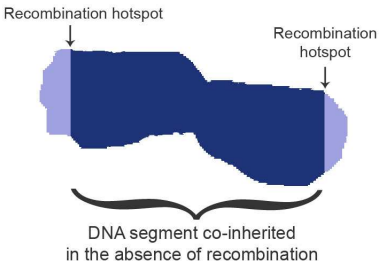

FIGURE 1.3: Non-recombining portion of the Y-chromosome. Adapted from Sharif (2007)

males, recombination of the XY-chromosomes is limited to the pseudoautosomal region on the tips (Cooke et al., 1985; Gusmão and Carracedo, 2003), thus the Y-DNA in the NRY passed down from a father to his sons will largely remain identical but for the variability of markers such as STRs or SNPs.

The lack of recombination means that Y-DNA is usually treated as being neutral in population studies but Jobling and Tyler-Smith (2000) argue it may be positively selected due to its effect on fecundity and conversely negatively selected through infertility.

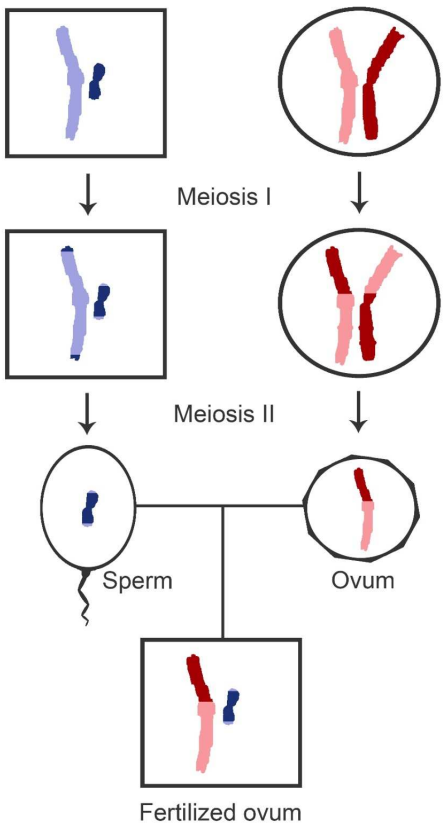

FIGURE 1.4: Recombination of the sex chromosomes. Adapted from Sharif (2007)

SNP sites where data may be collected from each male are referred to as SNP loci and sampled men can be separated into two groups according to the typed allele present at a particular SNP locus. Typing many SNPs can segregate men with similar profiles into clusters referred to as haplogroups. For example based on Y-DNA SNPs there are over 150 haplogroups: the major haplogroups/clades are shown in figure 1.5 (Karafet et al., 2008). This is a phylogenetic tree i.e. a branching diagram of the inferred relatedness of the haplogroups based on commonly shared SNPs. It is analogous to a family tree, thus the tips represent the descendants, whilst internal connections ('nodes') are treated as the most recent common ancestor. In Europe the R1 haplogroup is common and its tree is shown in figure 1.6. In particular the haplogroup R1b1 shows a gradient ('cline') across Europe with a relative frequency near one ('fixation') in western Ireland (Hill et al., 2000). SNP profiling informs us of the deep evolutionary ancestry of males 'stretching back to tens of thousands of years' (Pomery, 2007). On the other hand, STR profiles called haplotypes can describe

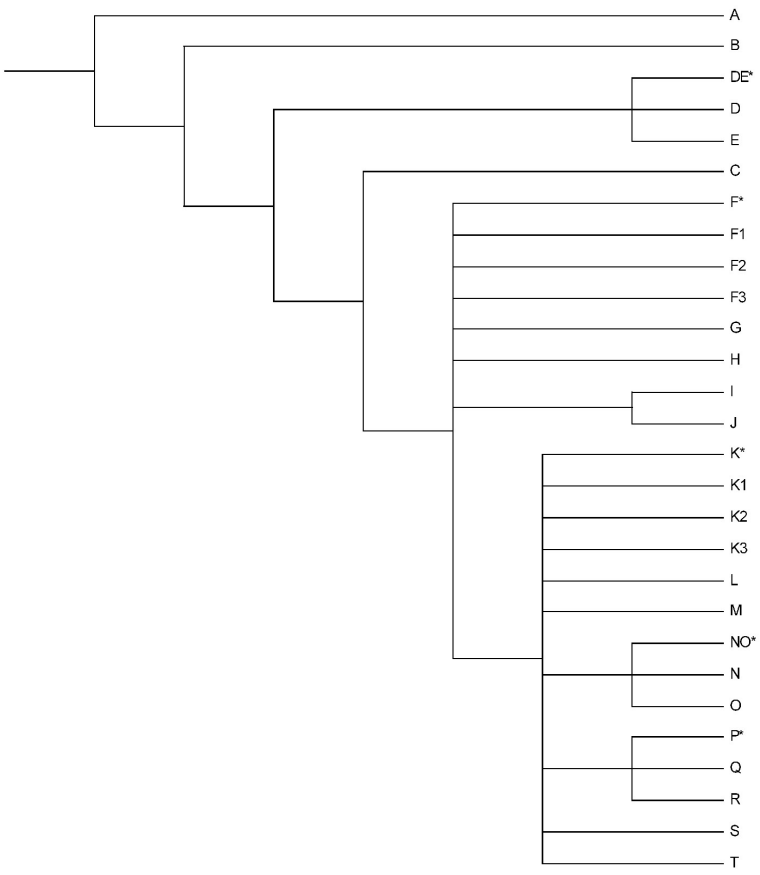

FIGURE 1.5: Y-chromosome SNP tree (Karafet et al., 2008)

more recent ancestry. In addition it is possible to infer haplogroup membership based on the haplotype for some populations (Moore et al., 2006). Both SNPs and STRs are of potential use to those wishing to research their genealogy although the latter are particularly informative for those researching their *historical* ancestry to the 16th century (Heyer et al., 1997; Pomery, 2007).

1.1.2 Genetic typing of STRs

DNA typing involves the following steps:

1. Collection
2. Extraction
3. Quantification
4. Amplification
5. Separation
6. Detection

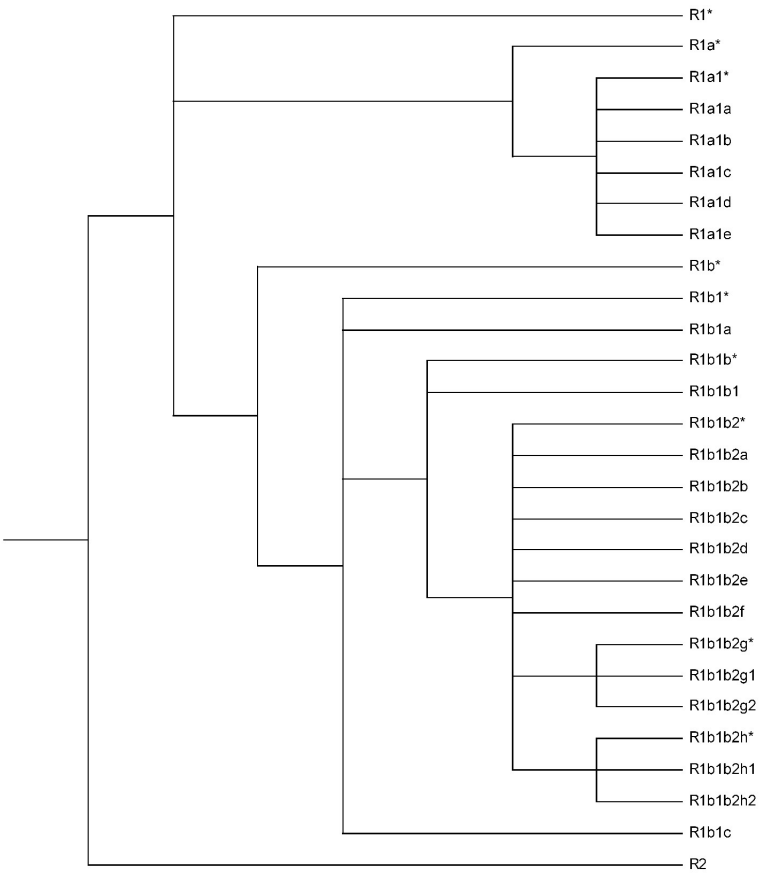

FIGURE 1.6: Y-chromosome R1 haplogroup tree (Karafet et al., 2008)

Genetic testing involves the typing of DNA from cells, usually obtained by sweeping a swab on the inside of the cheek (buccal swab) but also from samples of blood, semen and saliva in a forensic context (Butler, 2005; Vermeulen et al., 2009). The DNA may then be extracted from the samples by several methods: organic extraction, Chelex extraction, FTA (Fitzco/Flinder technology agreement) or solid-phase extraction. The latter seems to be favoured due to the ease of automation involved. The resultant DNA is then suitable for Polymerase Chain Reaction (PCR) amplification, a method used to increase the amount of DNA from a specific region. Prior to this, the DNA may undergo quantification in order to determine the amount of useful DNA, particularly when dealing with forensic samples, though for the purpose of lineage testing this may not be necessary. However due to the need for the amount of DNA to be within an optimum range some commercial PCR methods have been developed that allow integral quantification such as TaqMan (Butler, 2009).

PCR is an enzymatic process that requires many components and reagents which cause sets of specially designed primers to bind onto the complementary non-variable regions of DNA that flank an STR which are then replicated over and

over (fig. 1.7; Butler (2009)). Usually labelled 3' (pronounced "3 prime") and 5' for each end surrounding the marker, primers are typically of length 18-24 bases. Theoretically PCR can produce over 1 billion copies of the target DNA by a number of cycles of rapid cooling and heating. Multiplex PCR involves simultaneous amplification of several loci.

FIGURE 1.7: STR primer

The resulting PCR products may then be separated using either gel or capillary electrophoresis. Electrophoresis essentially allows separation of different sized PCR products by applying an electric field, which will pull shorter fragments further along than longer fragments. Gel electrophoresis involves placing PCR products along with a dye into loading wells contained on a gel slab which is covered in an electrophoresis buffer. Once the electric field is switched off, the samples remain fixed and the dye allows the end result to be detected and photographed although traditionally radioactive labels would have been used to produce images such as figure 1.8.

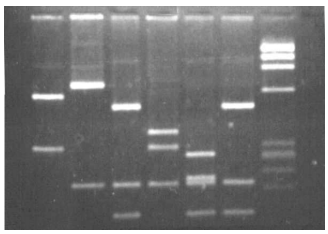

FIGURE 1.8: Gel electropherogram (Tucker, 2010)

Today, capillary electrophoresis (CE) is widely used due to its automated collection of samples, ease of use and reduction in work due to cleaning and preparing gel slabs (Butler, 2009). Also unlike using gels, CE measures the time taken for a fluorescent dye-labelled sample to pass though the capillary to the laser detection point. Often the dye is attached to the primers used to amplify the region and different coloured dyes may be used to detect carefully chosen multiplex PCR products (Butler et al., 2004). The

data collected by laser is then processed by computer software which separates the data from different dyes (in the case of multiplex PCR) and plots the spectrally resolved relative fluorescence intensity (RFU) against PCR product size which has been converted from the measured time (fig. 1.9).

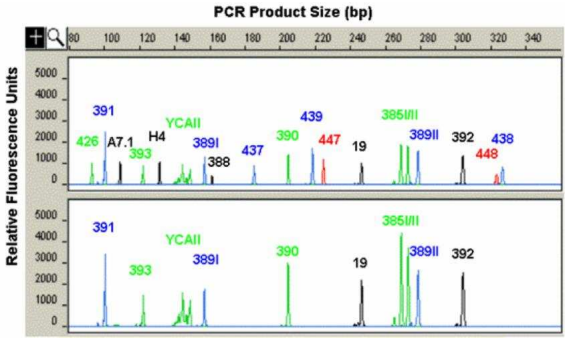

FIGURE 1.9: CE multiplex PCR results (Contexo, 2009)

Importantly, both gel and CE requires the use of allelic ladders: the PCR results are compared to those obtained from fragments of known size to allow the number of repeats in samples to be accurately reported. Figure 1.10 shows the Y allelic ladders for the Promega PowerPlex ®. Sometimes results may be reported that are outwith the ladder or off-ladder, i.e. below-ladder, above-ladder or between-ladder, representing alleles that are not exact multiples of the length of the repeat unit (Butler, 2009). Given that the range of allelic ladders may differ across commercial companies and laboratories, results reported as off-ladder by one company/lab may be designated a specified repeat number by another. Also 'stutters' may also be present in an electropherogram, which are essentially artefacts of PCR slippage and typically appear to be one repeat length shorter than the true allele (Walsh et al., 1996) but with only 10-20% the intensity/height of the true allele. Laboratories usually apply universal stutter filters to allow results to be clearly interpreted without masking the potential of mixed samples being detected (Butler, 2009).

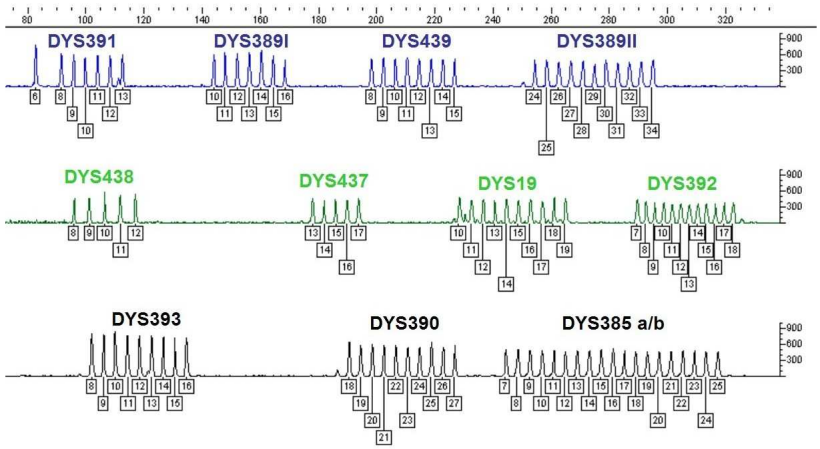

FIGURE 1.10: Promega PowerPlex ®Y allelic ladder (Butler and McCord, 2006)

In addition, the phenomenon of 'null alleles' may occur whereby samples fail to amplify due to a mutation in the primer binding site, usually at the 3' end which may manifest itself as allele dropout, i.e. no result is reported (fig. 1.11). This may be detected by use of alternative primers (Butler, 2009).

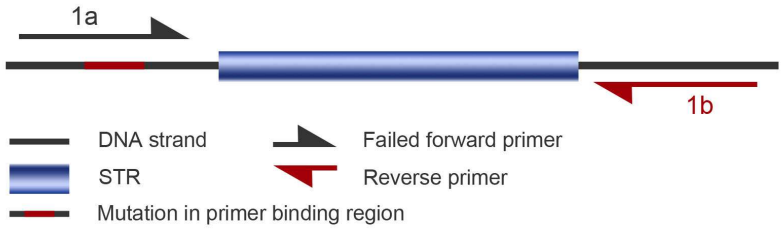

FIGURE 1.11: Null allele

Thus, in general, well-designed primers are key to the successful amplification of STRs. In particular STRs may be classified into different categories based on the results obtained by PCR and subsequent detection. Broadly speaking, STRs may be defined as either single or multi-copy markers. A single result is reported for a single-copy marker based on specific primers (fig. 1.12a). On the other hand, when more than one repeat length is reported, this will be referred to as a multi-copy marker(fig. 1.12b). In addition the lengths may not all involve the same repeating unit. Consequently there may be the problem of locus assignment particularly when the range of the repeat lengths reported overlap for different repeat units.

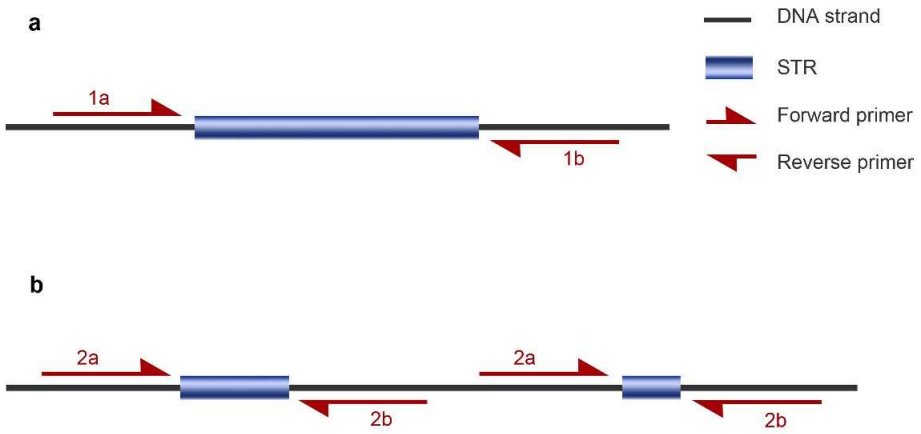

FIGURE 1.12: STR copy types: a. single copy marker b. Multi-copy marker

In addition, STRs may be classified as simple, complex or multiple complex markers based on the nature of the repeating unit. A simple marker is one which consists of a single repeating unit with no interruptions and thus is easily interpreted. For

example DYS388 consists simply of the repeating unit (ATT)n as shown in figure 1.13a (Gusmão et al., 2006). A complex (also referred to as 'compound' by Gunn (2009)) marker also consists of a single repeating unit but it may contain another repeat pattern or an interruption. DYS19 is an example of such a marker which has the unit TAGA repeated three times with the interruption TAGG thereafter the variable repeats of TAGA (fig. 1.13b).

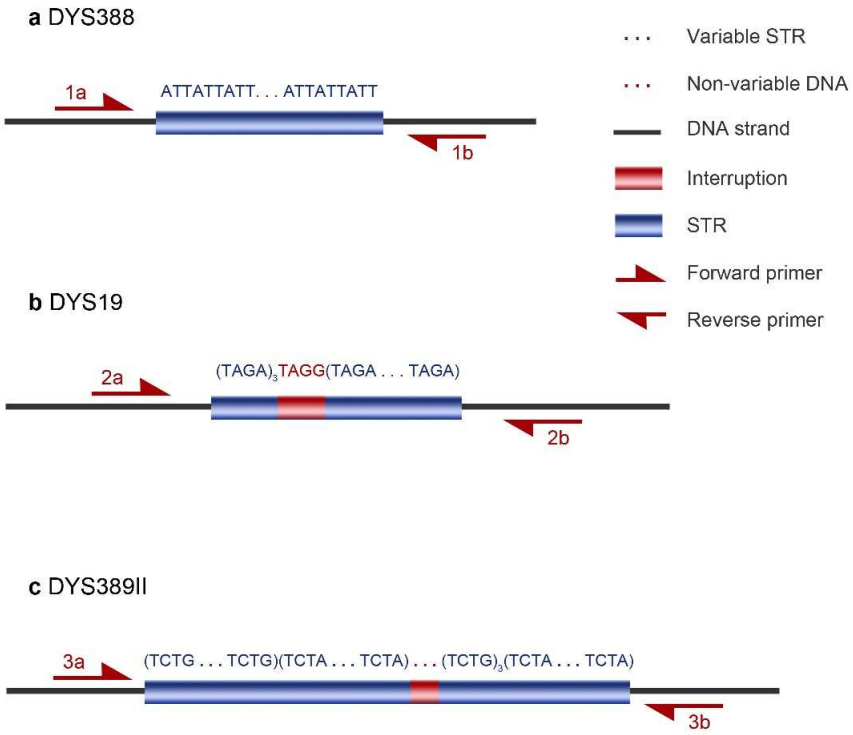

FIGURE 1.13: Varying complexity of STRs

Multiple complex markers (referred to as 'complex' by Gunn (2009)), on the other hand, consist of two of more repeating units which may or may not be interrupted or contain another repeat pattern. In figure 1.13c DYS389II illustrates this type of marker. Due to the complexity of this class of markers it may be difficult to assign which reported lengths are associated with the various repeat units in the marker without doing sequence analysis, particularly when the lengths reported are the same for each repeat unit.

Many favour the use of commercial kits such as AmpFlSTR©Yfiler^{TM}(Applied Biosystems) which amplify 17 Y-STRs (DYS456, DYS389I, DYS390, DYS389II, DYS458, DYS19, DYS385 a/b, DYS393, DYS391, DYS439, DYS635, DYS392, Y GATA H4, DYS437, DYS438, DYS448) and PowerPlex Y system (Promega Corporation) which amplifies the European minimal haplotype (DYS19, DYS385a/b,

DYS389I/II, DYS390, DYS391, DYS392, DYS393: Kayser et al. (1997)) as well as DYS437, DYS438 and DYS439.

However, there has been a trend to type simple single-copy Y-STRs recently (Lim et al., 2007; Vermeulen et al., 2009), for several reasons: the structure of the markers are simple, only one result is reported per marker and the European minimal haplotype is not as variable in other populations. Hence results are relatively straightforward and easy to interpret.

1.1.3 Genealogy

Genealogy may be described as "the study of the history and lineage of families" (Robinson and Davidson, 2003). Historically it has remained the domain of the ruling classes or nobility usually for the purpose of the distribution of inheritance or power. However by the late 20th century genealogy became of wider public interest. The role of the World Wide Web in the last 20 years has been profound.

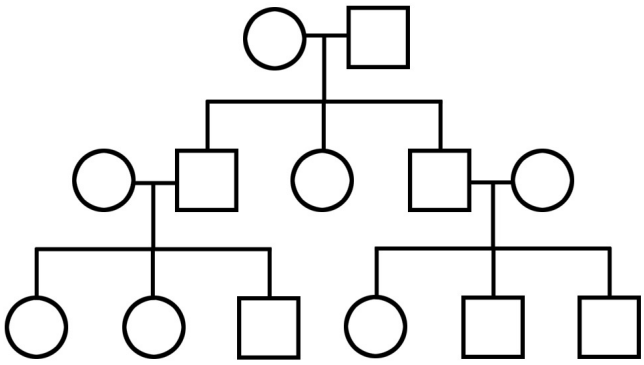

FIGURE 1.14: Graphical representation of a family tree (circle: female, square: male)

Traditional, genealogy involved reconstructing family trees of a living descendent(s) back in time, such as that shown in figure 1.14, based on studying written records such as a family Bible, parish records, wills, census returns, poll books and electoral registers, often involving manual cross referencing of names alongside dates and places of birth, marriage and death found in civil registers (Willis, 1970). Online genealogy operates in much the same way, with several benefits for both genealogists and holders of public archives. The Internet is accessible by over 1.7 billion people today, having grown over a hundred fold since 1995 (Internet

World Stats, 2010), making it an ideal conduit for the release of public archives without the disadvantage of damage to primary sources. Crucially it provides relatively easy and affordable access to a far wider audience without the need for travel, which was almost a necessary feature of genealogy traditionally (Christian, 2009).

Genealogy online began as early as 1983 with the newsgroup nets.roots and the mailing list ROOT-L in 1987 (Christian, 2009). Since the public launch of the World Wide Web in 1991 (GENUKI, 2010), there has been a gradual increase in the availability of information useful for genealogists largely due to commercial efforts but also with pressure from governments for public access. For example the 1911 UK census saw an early release online (Powell, 2009). Despite having to pay for access, these websites appear popular e.g. Ancestry.com, the market leader in online family history allowing searchable access to e.g. census returns and birth, marriage and death records, has almost a million subscribers worldwide, with nine regional websites (Ancestry.com, 2010) and web traffic of over five million unique visitors per month (Compete Inc., 2010). Furthermore Time magazine argues that 'root seeking' is as popular as searching 'sex, finance and sports' on the Internet (Hornblower et al., 1999).

Yet the popularity of genealogy is not limited to the Internet. In the UK television programmes such as the BBC's "Who do you think you are?", where celebrities are on the quest to trace part of their ancestry, have proved popular with audiences since 2004 (Rodgers, 2009) giving rise to various international versions including one in the US (NBC, 2010).

1.1.4 Surnames

Patrilineal surnames are surnames passed down from a father to his children with his sons then repeating this process and so on. In this respect surnames are cultural markers akin to genetic markers on the Y-chromosome, indeed Manni et al. (2005) argue that patrilineal surnames are like 'neutral alleles of a gene' on the Y-chromosome. However not all surnames are passed down in this manner. For example, in Iceland, surnames are based on the father's forename (Jobling, 2001). The establishment of surnames has varied across the world e.g. the Emperor Fu Xi standardised surnames in China approximately 5000 years ago while in Turkey surnames were legally imposed on its citizens as part of Ataturk's reforms in 1934

(Aslan, 2009). On the other hand in Japan only the ruling classes had surnames which were adopted ∼800 years ago with the masses obtaining surnames ∼140 years ago (Jobling, 2001).

Surnames in Britain were introduced after 1066 following the Norman Conquest, though the surnames were not necessarily hereditary, rather used as a means of identification and referred to as by-names. In Normandy, surnames had only been hereditary for a few generations prior to the Conquest and even then were limited to the nobles. Thus surnames first spread through the wealthy in Britain and by the mid 13th century most large and medium landowning families possessed hereditary surnames largely derived from place names (McKinley, 1990). Often, however, younger branches of such families may have established new surnames of their own. In the other classes, although surnames were recorded between 1150-1250, it was not till the mid 14th century that over 50% of the population had surnames. By the 13/14th century surnames, which began as by-names started to become hereditary. Thus, hereditary surnames were established in Britain between 500-700 years ago. The evolution of surnames prior to the 16th century was largely traced through taxation lists, title deeds and manorial records which were much more thorough in England than in Wales or Scotland. By 1538 parishes were required to keep records of baptisms, marriages and burials by order of Thomas Cromwell (Willis, 1970). It is important to note that since education was limited in the 17th and 18th centuries few knew how to read or write and as a consequence surnames were often written phonetically: for example 'Willis', 'Willes', 'Wyllys', 'Willys', 'Wilis', 'Willowes', 'Willic', 'Wilice' refer to the same name (Willis, 1970). Indeed even by the early 19th century spellings of surnames were not standardised (McKinley, 1990). Subsequently for those tracing a particular surname, variant spellings need to be considered, but with caution: it is possible to confuse different surnames as being derivatives when they are not (McKinley, 1990).

Britain has over 1.6 million surnames in current use (King and Jobling, 2009*b*) though earlier research puts this figure at just over 800,000 (McElduff et al., 2008). Nonetheless this is largely attributed to recent migrations with only 420,000 in use at the time of the 1881 census (King and Jobling, 2009*b*). The origin of a surname is said to be the most important factor influencing the geographic distribution of a surname and by extension the frequency (Plant, 2009). Broadly speaking surnames in Britain may be classified into six categories according to McKinley (1990):

- Locative surnames: These are based on place names e.g. London, Doncaster, Kendal. Whilst some locative surnames may be from unique place names, others may be based on the name of more than one locality, for example Norton, Kirkby or Ashby. However a great many locative surnames are based on the names of single places and are rare even today (McKinley, 1990).
- Topographical surnames: Surnames derived from natural or man-made landscape features. For example, the surnames Hill, Brooks and Marsh fall into the former class whilst Fields, Styles and Bridge belong to the latter.
- Personal surnames: These are based on first names, such as Paul, Peter and James, including their pet forms and diminutives. They also include surnames based on male forenames i.e. patronymics. For example, those surnames ending in 'son' or beginning with 'Fitz' or 'Mac/Mc'. Welsh surnames beginning with 'Pr' such as Probert, Pritchard and Price also fall into this class. Metronymics are also included here e.g. Maud, Eve and Margetson. Personal surnames tend to be rather common and are unlikely to have originated from single families (McKinley, 1990).
- Occupational surnames: Usually based on crafts or trades e.g. Smith, Taylor or Cooper, which are very common and early on had a rather dispersed distribution across the country since every town/village would usually require at least one individual to fulfil such crafts or trades. Rank/ status or office-based surnames are also included in this category e.g. Freeman, Squire, Bailliff and Hayward. Names of those in high Church or state positions are also incorporated here, surnames such as Bishop, Abbot or King. However with the relative prevalence of such surnames, it is often difficult to decide whether they were based on nicknames rather than official jobs.
- Nickname-based surnames: These are derived on either physical or moral characteristics as well as expressions for example Long, Malvoisin (French: 'Bad neighbour') and Goodenough, respectively. This class also includes those surnames based on terms for birds, mammals and fish e.g. Hawke, Fox and Herring.
- Surnames of relationship: This minor category of surnames is based on familial relationships. Examples include, Cousins, Brothers or Fadder.

Importantly many surnames origins may be unknown or indeed may not fall into any of the above categories.

Surname studies are not just of interest to genealogists but also to linguists, historians and more recently geneticists often in a genealogical setting though not exclusively (Plant, 2009). Surnames are often population or geographically specific. As such in healthcare, marketing and epidemiological studies, surnames are often used as an indicator of ethnic origin (King and Jobling, 2009*b*). In conjunction with Y-DNA, surnames aide in the study of historical migrations (Bedoya et al., 2006; Hill et al., 2000; King et al., 2007), admixture (Bowden et al., 2008; Manni et al., 2005) and non paternity rates (King and Jobling, 2009*a*; McEvoy and Bradley, 2006; Sykes and Irven, 2000). In addition there is the potential forensic application of the association between surname and Y-STRs (Jobling, 2001).

1.1.5 Surname-based Genetic Genealogy

Since patrilineal surnames in males are inherited in the same manner as Y-chromosomes, from father to son, we expect there may be an association between surname and DNA type (Jobling and Tyler-Smith, 1995).

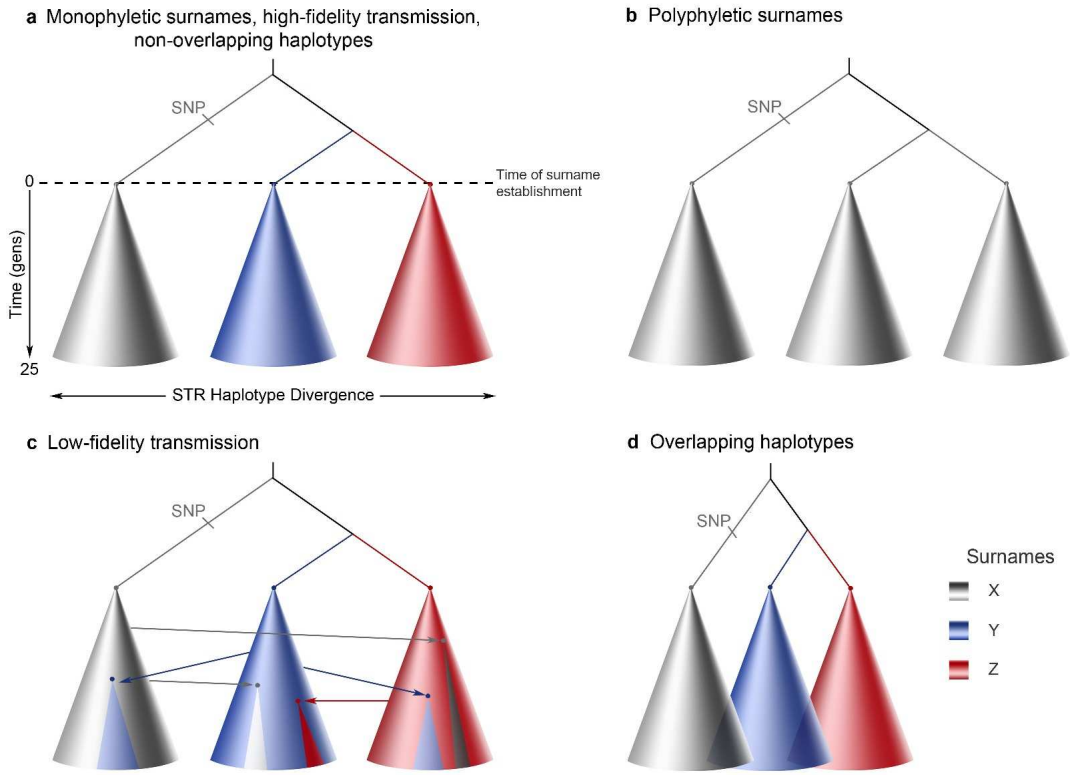

FIGURE 1.15: Surname-haplotype transmission. Adapted from Jobling (2001)

To illustrate this we begin by assuming unique ancestral haplotypes have established the surnames X, Y and Z. There will be a close relationship between the

haplotypes profile of the descendants with each surname despite divergence from the ancestral haplotypes profile due to mutations (fig. 1.15a). The presence of particular SNPs may further clarify the origins of the descendants. Indeed this may be the case for common surnames, e.g. those surnames based on occupations may have multiple origins each with distinct descendant haplotype profiles which may or may not overlap. In figure 1.15b we see that those from the first founder will have the SNP of interest, whilst those descended from the other founders will not. In this case the haplotype profiles across founders do not overlap. However in practice they may overlap, even for different surnames despite initially unique ancestral haplotypes due to the accumulation of mutations (fig. 1.15d). False paternity, where a child is given a surname other than that of his biological father either through illegitimacy, adoption or in some cases where the mother's surname is passed down (perhaps due to the passing down of land or wealth (McKinley, 1990)), may results in the introduction of different haplotypes or haplogroups to those of the founder of a surname (fig. 1.15c). The reality is often a combination of the aforementioned scenarios, though, for rarer surnames, it may be easier to discern the relationship between males since such surnames may only have a single founder.

Thus the Y-DNA surname correlation may be affected by:

- Non-paternity transmissions (NPT) due to illegitimacy, adoption, deliberate surname change or adoption of mother's surname.
- More than one surname founder.
- Mutations (on STR markers in particular).
- Genetic drift which can result in increasing the frequency of some haplotypes or equally causing others to become extinct.

In addition, Jobling and Tyler-Smith (2000) generalise that Y-chromosomes within a population tend to be closely related as a consequence of sons usually living closer to their parents than their daughters. Hence there may be both geographical and social clustering of haplotypes.

This conjunction of shared surname with Y-DNA provides a powerful tool to genetic genealogists and may often be used to complement conventional genealogical techniques whereby genealogies may be extended further or even excluded by comparison of genes from male descendants. Statistical models have been developed which aim to quantify the time to a most recent common ancestor (MRCA) for

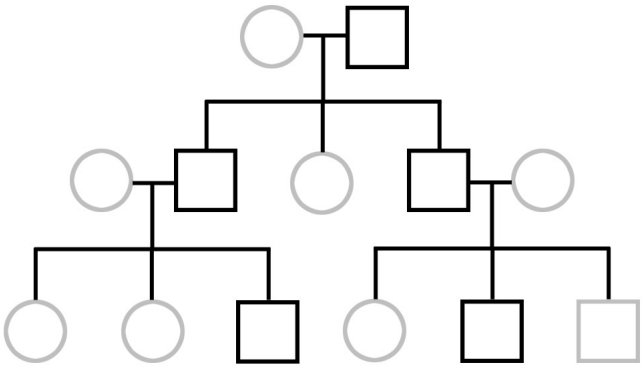

FIGURE 1.16: Graphical representation of the tracing of two male lineages to their MRCA (circle: female, square: male)

pairs of males using Y-STRs particularly in the absence of genealogical information. In a trivial example, in the family tree shown in figure 1.16 the extent of the genetic relationship between the two male grandsons highlighted in blue could be examined. This would reduce to the graphical representation in figure 1.17 and the time to the MRCA is 2 generations.

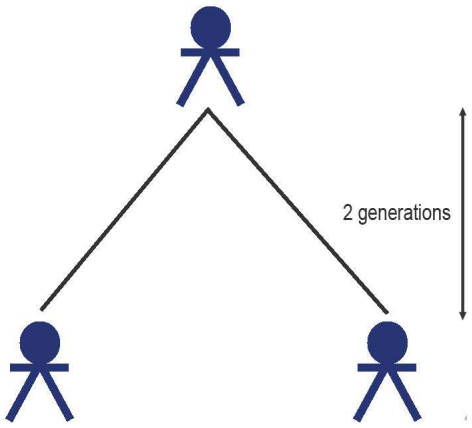

FIGURE 1.17: Graphical representation of the time to MRCA

In order to quantify the time to the MRCA for pair of males, Y-STRs are typed such that the data will be the absolute difference in the number of repeats of each STR as shown in table 1.1. Here we see that both males have the same number of repeats of DYS392 (14), whilst at DYS19 the males differ by 1 repeat. The data in this instance is 0-1-0 for the three markers and in general the more closely matched the markers are the higher chance of sharing a recent male common ancestor. Thus in surname based genetic genealogy the time to the MRCA (TMRCA, t) is usually the quantity of interest. Estimates of TMRCA can be affected by a number of factors. For example, the number of markers typed, the rate at which mutations occur and the number of surname founders.

TABLE 1.1: STR data

	DYS392	DYS19	DYS388
Repeats in male 1	14	16	13
Repeats in male 2	14	15	13
Difference in repeats	0	1	0

As a consequence of this novel use and the general popularity of recreational genealogy there has been a proliferation of genetic testing companies many specialising in typing Y-DNA and examining its association with surnames; over 15 were researched for the purpose of this thesis (see Cyndi's List (2012); Herbert (2009) for current comparisons/companies). A summary of the major companies at the start of 2010 is given in table 1.2. Although the range of the number of markers typed varies considerably e.g. African Ancestry only offers 8 Y-STRs (African Ancestry, 2010) whilst Genebase offers as many as 91 markers (Genebase, 2010), the average number of Y-STRs tested is 32 (31.77) across the companies detailed. Some companies only disclose which markers they type to customers, though most include the European minimal haplotypes as well as those markers typed on commercial Y-STR kits by either Applied Biosystems or Promega.

Of the five genetic testing companies listed that run surname projects, three provide some measure of TMRCA. DNA Heritage and GeneBase both appear to provide estimates based on the research by Walsh (2001) which will be discussed further in chapter 2. Family Tree DNA (FTDNA) is affiliated with both Bruce Walsh and Michael Hammer (Family Tree DNA, 2010). Their TMRCA calculator, FTDNATiP, operates using specific mutation rates for 37 markers whilst an average mutation rate is used for the remaining 22 markers. The specific rates have been based on over 130,000 meioses. FTDNA claim to have just as powerful results using only the 37 markers with specific mutation rate to using 56 markers with an average mutation rate of 0.004 mutations per marker per generation or indeed using 110 markers with average mutation rate of 0.002. In addition FTDNATiP allows the user to specify genealogical information such as time in generations when the earliest possible common ancestor may have lived.

In addition Ancestry.com provide estimates of TMRCA for their users based on both the number of matches and an average mutation rate of 0.0028 across all markers (Ancestry.com, 2010).

TABLE 1.2: Y-chromosome testing companies

Genetic Testing Company	No of Markers	Cost	Surnames Projects
Family Tree DNA	12 (inc. Family Finder)	$299	Yes
	37	$169	
	67	$268	
DNA Heritage	23	$137.77	Yes
	43	$199.00	
	'A la Carte' min. 23 Markers	$5.99 per marker	
Oxford Ancestors	15	£180	Yes
African Ancestry	8 and YAP	$ 299	No
Gene Tree DNA Testing Center	33	$149	No
	46	$179	
Ethnoancestry	27 plus SNPs	$399	No
National Genographic Project	12	$99.95 +pp	No
AncestrybyDNA	14	$99	No
iGENEA	12	129 Euro	Yes
	37	169 Euro	
	67	259 Euro	
Genebase	20	$119	Yes
	44	$ 199	
	67	$ 269	
	91	$ 339	
Paternity Experts	17	$79	No
Roots for Real	11	£195	No
Cambridge DNA Services	11	£150	No
Ancestry.com	33	$99	No
	46	$149	

There is the potential for collaboration between recreational genealogy, with its abundance of data, and academia. For example, Sims et al. (2009) made use of a genetic genealogy database to refine the haplogroup G phylogeny. However sampling bias may be inherent in the non-academic setting and, although commercial Y-STR tests have a higher resolution, this is coupled with the increased chance of typing errors and discovering mutations between close relatives (King and Jobling, 2009*b*). Downsides to the use of genetic genealogy particularly in a recreational setting include the unwitting identification of anonymous DNA donors, of non-paternity events (Williams, 2005), of infertility (King et al., 2005) and of erroneous membership to historical Y-lineages (King and Jobling, 2009*b*).

A well-publicised case of a fifteen-year-old boy discovering the identity of his biological sperm donor father highlights the first drawback (Motluk, 2005). The use of the donor father's date and place of birth and college degree together with finding same surname matches to his Y-DNA haplotype from a genetic genealogy database led him to find his father. This use would be considered a violation of

privacy especially in the US, where sperm donors are afforded anonymity. Indeed this cross referencing of DNA haplotypes with public databases has also been used to identify the surnames of DNA donors for genetic studies (Gitschier, 2009).

Deletions of part of the AZF gene on the Y-chromosome are associated with male infertility. Since the AZFa region contains DYS434, DYS435, DYS388, DYS436 and the popularly typed STRs DYS389, DYS437, DYS438 and DYS439 infertility may be inadvertently revealed to males returning a null allele at the given markers. Tracing lineages to well-known historical individuals or populations is often a key selling point by many Y-DNA testing companies which exploit academic inferences about the DNA of, for example, Thomas Jefferson, Genghis Khan, the Irish Ui Neill and the Jewish Cohanim (Foster et al., 1998; Moore et al., 2006; Thomas et al., 1998; Zerjal et al., 2003). In order to ascertain the ancestral Y-DNA, STRs are typed from putative living descendants typically, though typing STRs from archaeological skeletal human remains has also proved fruitful in some contexts (Gerstenberger et al., 1999; Marjanovic et al., 2009). This is however generally disapproved of in the wider genealogical community.

1.2 Y-Chromosome Databases

Eight different Y-DNA databases were researched of which four were genealogical databases, three were forensic and one academic (table 1.3). There may be some overlap in the databases: YMatch (YMatch, 2010) was developed under the auspices of DNA-Fingerprint which is now a subsidiary of FamilyTree DNA and the PowerPlex Y-Haplotype database maintained by Promega has been added to the US National Y-STR haplotype database though the original database is still available to search online. The forensic databases all return population group affiliation. The only academic database (Y Chromosome Haplotype Reference Database, 2008) has a multi-stage process for data inclusion, including that the results are published academically. The database by Sorensen Molecular Genealogy Foundation (SMGF, 2010) has an inherent verification process for data to be included. This requires DNA to be typed either by themselves via Sorensen Genomics or through GeneTree, a genetic testing company. In addition they request participants to include a detailed family tree which they verify. SMGF also provide TMRCA estimates which are based on the number of matches and does not take into account the size of any mismatches. It also appears to incorporate

site-specific mutation rates, though details of this are unclear. Access to their database is restricted to registered users unlike all the other databases. All the remaining genealogical databases allow unverified public submissions subject to human error. In general all the genealogical databases have more markers available to search compared to the non-recreational databases which allow at most 17 Y-STRs.

In addition ancestry.com has a publicly searchable surname database with very limited data. Members are afforded further information of matches and are allowed to input DNA results obtained from elsewhere. A members-only database is also available on Oxford Ancestors. Given the nature of the restricted access of these databases it is difficult to assess their content clearly. Other genetic genealogy websites may offer database access though this is not always clearly advertised.

TABLE 1.3: Y-DNA databases

Name	Affiliations/Sponsors	No. of Y-STRs	Type[a]	Search Fields[b]	Database Submissions	Populations[c]
YSearch	FamilyTreeDNA	100	G	S, Ht, Hg	Unverified Public	W
YBase	DNAHeritage.com	49	G	S, Ht,	Unverified Public	W
YMatch	DNA-FingerPrint.com	99	G	Ht, C, Hg	Not available	N
Y Chromosome Haplotype Reference Database		17 & 57 Y-SNPs	A	Ht, Hg, P, Ct	Academic publications	E
YFiler	Applied BioSystems	17	F	Ht	Forensic	W
PowerPlex Y Haplotype Database	Promega	12	F	Ht, P	Forensic	N
SMGF	GeneTree.com	42	G	S, Ht	Verified via GeneTree	W
US Y-STR Database		11-17	F	Ht, P	Verified Forensic Laboratories & Institutions	US

[a]G: genealogical, A: academic, F: forensic

[b]S: surname, Ht: haplotype, Hg: haplogroup, C: country, P: population, Ct: contributor

[c]W: worldwide, E: primarily European, US: primarily US, N: not specified

1.3 Bayesian Inference

At this stage it is important to outline the statistical framework in which we aim to infer the parameters in the model we develop since it is employed throughout this thesis.

Bayesian statistics requires forming a full probability model, which expresses the joint probability distribution for the observed data (including effects arising from how the data were collected) and unobserved parameters. From this we may derive the conditional probability of the unobserved parameters given the observed data, i.e. the posterior distribution (Gelman et al., 2004). We will use the following notation to outline standard Bayesian theory:

- y: the vector of observed data
- θ: the vector of unobserved or unknown parameters
- $p(\theta)$: the prior or marginal distribution of θ
- $p(y)$: the marginal distribution of y
- $p(\theta, y)$: the joint distribution of both θ and y
- $p(y|\theta)$: the sampling distribution, i.e. the conditional probability of y given θ
- $p(\theta|y)$: the posterior distribution, i.e. the conditional probability of θ given y

The joint probability of the data and parameters may be written as:

$$p(\theta, y) = p(y|\theta)p(\theta). \tag{1.1}$$

Bayes' theorem allows the determination of the posterior distribution and may be written as follows:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(y|\theta)p(\theta)}{p(y)} \propto p(y|\theta)p(\theta). \tag{1.2}$$

We aim to learn about the parameters of interest through $p(\theta|y)$. The observed data may be modelled in terms of certain parameters, θ, which in turn are conditional on further parameters, ϕ, so called hyperparameters (Gelman et al., 2004). These have a separate prior distribution, $p(\phi)$, distinct from the prior on θ, which

is conditioned on ϕ. The joint prior on θ and ϕ is:

$$p(\phi, \theta) = p(\theta|\phi)p(\phi). \tag{1.3}$$

The joint posterior distribution is written as:

$$\begin{aligned} p(\phi, \theta|y) &\propto p(y|\theta, \phi)p(\phi, \theta) \\ &= p(y|\theta)p(\phi, \theta). \end{aligned} \tag{1.4}$$

This may be naturally extended to include several levels in the hierarchy necessary in hierarchical modelling.

Point estimates for the parameters in θ may be obtained by calculating the mean or mode of the posterior distribution. In addition we may compute the credible region or posterior standard deviation to quantify the uncertainty in the point estimates.

Chapter 2

Background

2.1 Estimate of Time to the Most Recent Common Ancestor

Walsh (2001) has been the main source cited amongst the genealogical circles to estimate the time to the most recent common ancestor (TMRCA) for pairs of males using microsatellite data (DNA Heritage, 2010; Family Tree DNA, 2010; Genebase, 2010). The author presents two different methodologies, which incorporate the coalescent model, to estimate the time to the MRCA for haploid markers such as those found on Y-DNA and mtDNA. Both methods assume that n loci are typed on the pair of individuals who have diverged from a common ancestor t generations ago (fig. 2.1).

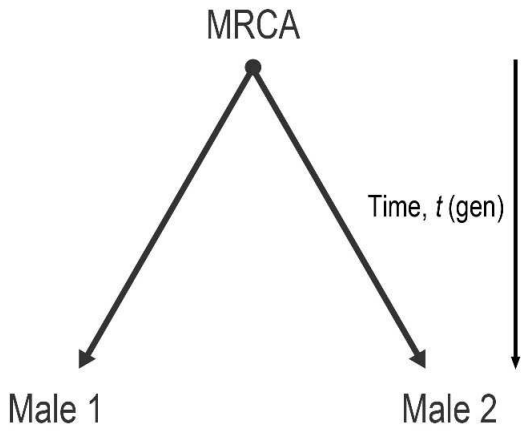

FIGURE 2.1: The time to the most recent common ancestor

The first model outlined is the infinite alleles model (IAM) which uses the following notation:

- n total number of microsatellite loci,
- k total number of matches (such that $k \leq n$),
- μ probability of a mutation (per generation per locus),
- t time to the MRCA.

The model considers only the number of matches and not the size of the difference (in the number of repeats) between the pairs of markers that do not match and importantly assumes that markers match only when there have been no mutations. Consequently recurrent mutation such as parallel and back mutations are not considered. Parallel mutations are mutations with the same size and direction occurring on both lineages from the MRCA as shown in figure 2.2 where an increase mutation occurs at a single marker on both lineages, which as a result have the same repeat length once typed.

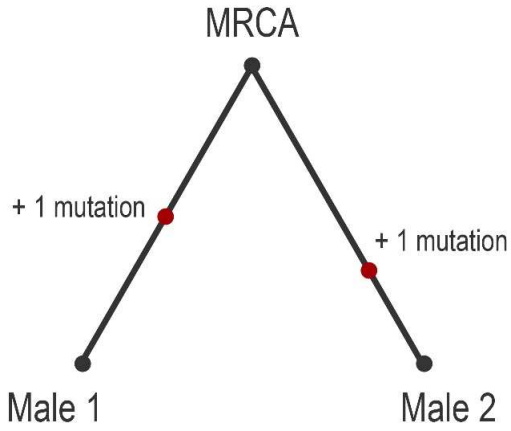

FIGURE 2.2: Parallel mutation

On the other hand, back mutations involve two mutations of the same size but opposite direction occurring on a single lineage. For example, in figure 2.3, we see that an increase mutation in lineage 2 is wiped out by a later decrease mutation. In this case the typed marker on both lineages does not differ despite the mutations.

The number of mutations between the two males at one locus in the IAM, Y, is distributed as a Poisson distribution with rate $2t\mu$:

$$P(Y = y) = e^{-2t\mu}\frac{(2t\mu)^y}{y!}.$$

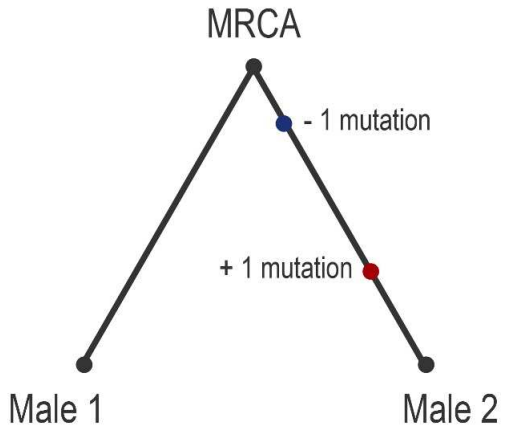

FIGURE 2.3: Back mutation on male 2 lineage

A marker matches when $Y = 0$, so that

$$P(\text{match}) = P(Y = 0) = e^{-2t\mu}\frac{(2t\mu)^0}{0!} = e^{-2t\mu}. \tag{2.1}$$

Walsh (2001) models the number of matches k with a binomial distribution, i.e. $K \sim Bi(n, e^{-2t\mu})$. So

$$P(K = k|n, t) = \binom{n}{k}(e^{-2t\mu})^k(1 - e^{-2t\mu})^{n-k} \tag{2.2}$$

This is the likelihood of t. The maximum likelihood estimator (MLE) of t, $\hat{t}$, is

$$\hat{t} = -2\mu \ln \frac{k}{n} \tag{2.3}$$

The MLE of t does not appear to be very informative as it is upwardly biased, has high variance and produces asymmetric confidence intervals around $\hat{t}$. In particular, $k = n$, the $\hat{t} = 0$. Hence Walsh employs a Bayesian approach to form the full posterior distribution for t under the IAM.

Assuming the coalescent model, we have the following prior for t,

$$p(t|\lambda) = \lambda e^{-\lambda t}, \tag{2.4}$$

where $\lambda = N_e^{-1}$ and N_e is the effective population size (Hein et al., 2005).

So the posterior in this case is:

$$\begin{aligned} p(t|k) &\propto P(k|n,t)p(t|\lambda) \\ &\propto e^{-2t\mu k}(1-e^{-2t\mu k})^{n-k}\lambda e^{-\lambda t} \end{aligned} \tag{2.5}$$

from (2.2) and (2.4).

The normalised posterior density can be obtained by integration:

$$p(t|k,\lambda) = \frac{\prod_{i=0}^{n-k}[\lambda + 2\mu(n-i)]}{2^{n-k}(n-k)!\mu^{n-k}} \frac{(1-e^{-2t\mu k})^{n-k}}{e^{(-2t\mu k+\lambda)t}}. \tag{2.6}$$

Walsh graphs the posteriors and tabulates summaries of the posteriors for $n = 5, 10, 20, 50, 100$ and various numbers $(n-k)$ of mismatches, using $\mu = 0.002$ (per locus per generation) and a flat prior ($\lambda = 0$).

Next Walsh extends this model to allow for variable mutation rates across the n loci where the data at the k^{th} locus ($k = 1 \ldots n$) are represented as:

$$x_k = \begin{cases} 1 & \text{match at } k^{th} \text{ locus,} \\ 0 & \text{mismatch at } k^{th} \text{ locus.} \end{cases}$$

The likelihood of t is:

$$L(x_1, \ldots, x_n|t) = \prod_{k=1}^{n} q_k(t)^{x_k} \left[1 - q_k(t)\right]^{1-x_k}, \tag{2.7}$$

where

$$q_k(t) = (1-\mu_k)^{2t} \simeq e^{-2t\mu_k} \tag{2.8}$$

is the probability of a match at locus k. Substituting (2.8) into (2.7) gives:

$$L(x_1, \ldots, x_n|t) = \exp\left[-2t\sum_{k=1}^{n}\mu_k x_k\right] \prod_{k=1}^{n}\left(1-e^{-2t\mu_k}\right)^{1-x_k} \tag{2.9}$$

So that the posterior density using the prior (2.4) is

$$\begin{aligned} p(t|k) &\propto P(k|n,t)p(t|\lambda) \\ &\propto \exp\left[-t\left(\lambda + 2\sum_{k=1}^{n}\mu_k x_k\right)\right] \prod_{k=1}^{n}\left(1-e^{-2t\mu_k}\right)^{1-x_k}. \end{aligned} \tag{2.10}$$

Now, since the IAM does not take into account mutations that are essentially wiped out, the method will underestimate t for older times so Walsh determines to what extent the IAM underestimates by using an alternative method based on the stepwise mutation model (SMM) of Ohta and Kimura (1973). Rather than modelling the data as having descended from the MRCA, this model assumes a directionality in the formation of the mutations from one male to the other as shown in figure 2.4, where this process occurs from male 1 to male 2. As such the model treats mutations at STRs as a random walk with an equal probability of increasing or decreasing the number of repeats. The notation used in this case is:

- μ probability of a mutation (per generation per locus)
- n_+ total number of increase mutations
- n_- total number of decrease mutations
- t time to the MRCA

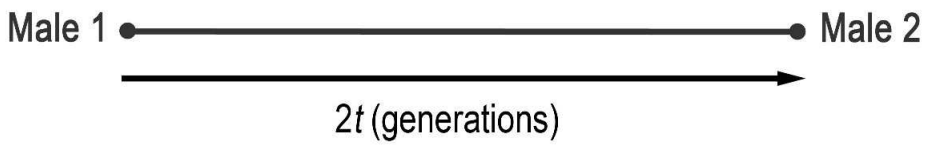

FIGURE 2.4: Directionality of SMM

For example in the graphical representation in figure 2.5 we have observed data for Male 1 of 15 repeats at a particular locus. Moving from male 1 to male 2, there are in total 6 decrease mutations and 4 increase mutation resulting in 13 repeats of the STR in Male 2.

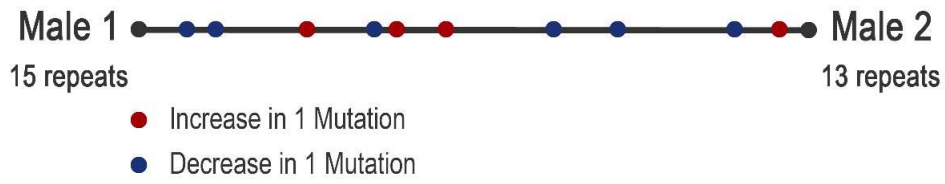

FIGURE 2.5: Random walk of an STR from male 1 to male 2

In the SMM the probability of an increase mutation and the probability of a decrease mutation are equal. So the transition probabilities are:

$$\begin{aligned} \Pr\left(X(t+1) = i-1 | X(t) = i\right) = \Pr\left(X(t+1) = i+1 | X(t) = i\right) &= \frac{\mu}{2}, \\ \Pr\left(X(t+1) = i | X(t) = i\right) &= 1-\mu, \\ \Pr\left(|X(t+1) - X(t)| \geq 2 | X(t) = i\right) &= 0, \end{aligned} \tag{2.11}$$

where $X(t)$ is the repeat length at time point t.

Next the probability of a match given that the MRCA lived t generations ago, $q(t)$, is derived. This would only be possible if an even number of mutations occurred i.e. $|n_+ - n_-| = 0$ only when $n_+ + n_- = 2m$ where m is an integer ≥ 0. Supposing that X is the total number of n_+ mutations then $X \sim Bi(2m, \frac{1}{2})$. So the probability of $n_+ = m$ is

$$\begin{aligned} P(X = m) &= \binom{2m}{m} \frac{1}{2}^m \left(1 - \frac{1}{2}\right)^{2m-m} \\ &= \frac{1}{2^{2m}} \frac{(2m)!}{m!^2}. \end{aligned} \tag{2.12}$$

In addition, since the probability of $2m$ mutations occurring in $2t$ generations is distributed as a poisson with parameter $2t\mu$ we have:

$$\begin{aligned} P(|n_+ - n_-| = 0|t) &= \sum_{m=0}^{\infty} \Pr(|n_+ - n_-| = 0|2m) \Pr(2m|t) \\ &= \sum_{m=0}^{\infty} \left(\frac{1}{2^{2m}} \frac{(2m)!}{m!^2}\right) \left(\frac{(2t\mu)^{2m}}{(2m)!}\right) \exp(-2t\mu) \\ &= \exp(-2t\mu) \left(\sum_{m=0}^{\infty} \frac{(\mu t)^{2m}}{m!^2}\right). \end{aligned} \tag{2.13}$$

Given that

$$\sum_{k=0}^{\infty} \frac{x^{2k}}{k!^2} = I_0(2x), \tag{2.14}$$

where I_0 denotes the zero-order modified type I Bessel function (Olver, F. W. J. and National Institute of Standards and Technology (U.S.), 2010). So, under the SMM, the probability of a match between two individuals after $\tau = 2\mu t$ generations at one marker is

$$q(\tau) = \exp(-\tau) I_0(\tau). \tag{2.15}$$

Walsh (2001) compares this match/mismatch SMM to the IAM already discussed by computing the ratios of their posterior means and SDs for $\hat{t}$. In general these statistics are larger under the SMM. If λ is increased from 0 to $\frac{1}{500}$, the difference decreases so the ratios tend to 1. This effect is explained as follows: increasing λ essentially means decreasing the effective population size N_e. This reduces the probability of multiple mutations since the time is reduced.

The ratios also tend to 1 on increasing the number of markers typed from $n = 5$ to 100. Walsh argues that happens because under the SMM when there are no mismatches there is still the possibility of 2 or more mutations having occurred in one or more markers. The probability of 2 mutations in the same marker is only $\frac{1}{n}$ so as n increases the probability of multiple markers being masked decreases since the occurrence of 2 mutations in the same marker is less likely.

This match/mismatch version of the SMM is sufficient under short times to the MRCA ($t \ll \frac{1}{2\mu}$) but for longer timescales the information of the size of the difference at mismatched markers would provide additional information about the time to the MRCA. Walsh extends the SMM model outlined by again considering the probability of an even number of differences which will only occur when the total number of mutations is even, i.e. $|n_+ - n_-| = 2k$ only when $n_+ + n_- = 2m$, where k and m are non-negative integers. For this to be satisfied $n_+ = m - k$ or $n_+ = m + k$. So, suppose again that X is the total number of n_+ mutations then $X \sim Bi(2m, \frac{1}{2})$ and

$$P(X = m - k) = \binom{2m}{m-k}\left(\frac{1}{2}\right)^{m-k}\left(1 - \frac{1}{2}\right)^{2m-(m-k)} = \binom{2m}{m-k}\left(\frac{1}{2}\right)^{2m} \tag{2.16}$$

and

$$P(X = m + k) = \binom{2m}{m+k}\left(\frac{1}{2}\right)^{m+k}\left(1 - \frac{1}{2}\right)^{2m-(m+k)} = \binom{2m}{m+k}\left(\frac{1}{2}\right)^{2m}. \tag{2.17}$$

Thus we have

$$\begin{aligned}\Pr(|n_+ - n_-| = 2k|2m) &= Pr(X = m - k \text{ or } X = m + k|m, k) \\ &= 2\frac{(2m)!}{(m+k)!(m-k)!}\left(\frac{1}{2}\right)^{2m}.\end{aligned} \tag{2.18}$$

Now in addition the total number of mutations $Y = n_+ + n_-$ where $Y \sim Po(2t\mu)$ so that

$$\Pr(Y = 2m|t) = e^{-2t\mu}\frac{(2t\mu)^{2m}}{(2m)!}. \tag{2.19}$$

The probability of a $2k$ difference given t is formed by combining (2.18) and (2.19), i.e.

$$\begin{aligned}\Pr(|n_+ - n_-| = 2k|t) &= \sum_{m=k}^{\infty} \Pr(|n_+ - n_-| = 2k|2m) \Pr(2m|t) \\ &= \sum_{m=k}^{\infty} 2\binom{2m}{m-k}\left(\frac{1}{2}\right)^{2m} e^{-2t\mu}\frac{(2t\mu)^{2m}}{(2m)!} \\ &= 2e^{-2t\mu}\sum_{m-k=0}^{\infty}\frac{(t\mu)^{2m}}{(m-k)!(m+k)!}. \end{aligned} \tag{2.20}$$

Letting $m' = m - k$, we can rewrite (2.20) as

$$\Pr(|n_+ - n_-| = 2k|t) = 2e^{-2t\mu}\sum_{m'=0}^{\infty}\frac{(t\mu)^{2m'+2k}}{\Gamma(m'+1)\Gamma(m'+2k+1)}. \tag{2.21}$$

The summation in 2.21 has the form of as a νth-order type I modified Bessel function,

$$\mathrm{I}_\nu(z) = \sum_{s=0}^{\infty}\frac{(z/2)^{\nu+2s}}{\Gamma(s+1)\Gamma(s+\nu+1)} \tag{2.22}$$

(Lebedev and Silverman, 1972; Olver, F. W. J. and National Institute of Standards and Technology (U.S.), 2010).

Thus we can rewrite (2.21) as:

$$\Pr(|n_+ - n_-| = 2k|t) = 2e^{-2t\mu}I_{2k}(2t\mu). \tag{2.23}$$

A similar argument can be made when $|n_+ - n_-|$ is odd. Consequently the probability that at a particular locus the observed absolute difference in the number of repeats, j, of the microsatellites between two individuals is given by

$$q^{(j)}(2t\mu) = 2e^{-2t\mu}I_j(2t\mu) \qquad j \geq 1. \tag{2.24}$$

Given n loci, we redefine x_j to be the number of times we observe j differences across the n loci. Then the likelihood of t is

$$L(x_0, \ldots, x_J|t) = \frac{n!}{x_0! \ldots, x_J!}\prod_{j=0}^{J}\left[q^{(j)}(2t\mu)\right]^{x_j}, \tag{2.25}$$

where J is the maximum number of differences observed.

Applying an exponential prior to t, as before, gives the posterior distribution:

$$\begin{aligned}\Pr(t|x_0,\ldots,x_J) &\propto \prod_{j=0}^{J}\left[q^{(j)}(2t\mu)\right]^{x_j} e^{-\lambda t} \\ &\propto \prod_{j=0}^{J} e^{-2t\mu x_j}\left[\mathrm{I}_j(2t\mu)\right]^{x_j} e^{-\lambda t} \\ &= e^{-(2\mu n+\lambda)t}\prod_{j=0}^{J}\left[\mathrm{I}_j(2t\mu)\right]^{x_j}. \end{aligned} \tag{2.26}$$

This model, along with the other two (IAM and match/mismatch SMM), are then used to examine $\hat{t}$, i.e. the posterior mode and median, of the modal haplotype of the Lemba and Cohen from Thomas et al. (2000) employing $\mu = 0.00245$ per locus per generation and $\lambda = 1/5000$ per generation. As well as other markers, six Y-STRs were typed: DYS19, DYS388, DYS390, DYS391, DYS392, and DYS393. Four markers matched whilst two mismatched; one by 1 repeat another by 2 repeats. The posterior mean $\hat{t}$ using (2.26) is 352.5 generations, a slightly lower value than the match/mismatch SMM value of 394.4 generations. Nonetheless both are much higher than the IAM $\hat{t}$ of 124.8 generations which also has a much narrower credible region (which overlaps with the credible regions for the other two models).

Walsh (2001) cautions in using the normalised posterior (2.26) since evidence exists to suggest μ increases as the STR repeat unit length increases. In addition there may be also a bias to increasing the number of repeats and of mutations of more than one repeat unit occurring. Finally there may be an interplay of many molecular processes that will affect the overall mutation rate and the correction that is required for multiple mutations. Potentially the SMM may introduce a bias since it may wrongly overcorrect for multiple mutations. Walsh (2001) suggests using the Fu and Chakraborty (1998) approach of minimising the chi-square for estimation of the general SMM, which allows for multi-step mutation, as well as allowing unequal probabilities for an increase and decrease mutation.

The results presented by Walsh (2001) suggest that the forensic use of STRs on Y-DNA or mtDNA may be limited to exclusion purposes. Though for the latter this may be extremely limited given the low number of STRs (pers. comm. Martin B. Richards). Otherwise, a match at 10 markers between a suspect's DNA and sample DNA gives a 90% chance that the MRCA lived no more than 58 generations ago, using (2.3), assuming a mutation rate per locus of 0.002 per generation. In

order for this to be reduced to a time to MRCA of 1 generation, more the 580 markers would have to match. A 50% probability of a MRCA no more than 1 generation would need a complete match of around 340 markers.

The use of $\lambda = 0$ in the prior has little effect in the IAM unless $N_e < 200$ or $k/n \ll 1$. On the other hand, it has a bigger impact in the SMM, especially when n is small and $k/n \ll 1$. Increasing λ reduces the chance of t being high since N_e is reduced.

There is also possible bias of t due to the ascertainment of markers, i.e. only using variable markers and ignoring the contribution of invariable or less variable markers. Walsh (2001) argues this is more important and needs correcting when estimating coalescent times for populations. However, in the case of estimating parameters for pairs of individuals, Walsh argues that the methodology presented innately corrects for this since it uses estimated mutation rates that have been ascertained on polymorphic markers. For random STRs the author suggests replacing the mutation rate with $c\mu$ where $c > 1$ is an ascertainment correction based on the process of how the markers were ascertained. Yet it may be possible to explicitly model the ascertainment process rather than just scaling μ (Nicholson et al., 2002).

The method may also be extended to SNP data by using the IAM for SNPs whilst using the SMM for STRs. Another extension is comparing an individual's DNA to an ancestral haplotype (whether actual or inferred) by replacing $2t$ by t so that in effect here only a single lineage descending from an ancestor, for example only considering the Male 1's lineage from the MRCA in figure 2.1.

Walsh outlines the various methods previously used for estimating the time to MRCA of more than two individuals. In particular, we note the use of the Poisson distribution for the number of segregating sites as opposed to the binomial which is used in the methods presented. Walsh argues this is justified given that there is roughly 1 STR for every 10Kb of DNA and that when n is large and μ small both models should give the very similar $\hat{t}$ (Poisson approximation to binomial). Walsh motivates a model incorporating the mutational process by including a prior on μ. So in the Bayesian framework we have:

$$P(t|\lambda, \text{marker information}) \propto \int L(t|\text{marker information})p(\lambda)p(\mu)d\mu.$$

However, Walsh stresses the need for a sensible prior on μ since a poorly chosen prior may introduce even more bias than it corrects. This is only an issue when the data itself is uninformative about μ.

Nordtvedt (2008) describes a method of estimating the TMRCA, t, for pairs of males where they mismatch by only 1 step at n out of N markers and therefore match at exactly $N - n$ markers. The nature of the data means that males typed will be quite similar to each other. In this paper Nordtvedt focuses on a model which incorporates the frequency of the ancestral haplotypes arguing that pairs of males with a similar haplotype to the most frequent ancestral haplotype, i.e. the modal, may have a higher t than $\hat{t}$ estimated simply on the basis of the similarity of the male haplotypes to each other since time is required for the ancestral haplotype to increase in frequency. The converse would also apply i.e. rare haplotypes compared to the modal are probably more recent and pairs of males close to each other and a rarer ancestral haplotype are likely to be more closely related. As such t would be expected to be lower. The narrow restriction of only a one-step mismatch at markers means that Nordtvedt's model cannot be used for random males, where the size of mismatches may be greater than one.

Howard (2009*a*) presents an alternative method of estimating the time to a MRCA based on reducing the haplotypes of pairs of males to a single value. For each pair of males the Pearson correlation coefficient between the repeat numbers at all markers, r, is computed which is then rescaled to form the revised correlation coefficient (RCC):

$$\text{RCC} = \left(\frac{1}{r} - 1\right) \times 10,000.$$

Values of RCC close to zero would imply a close relationship between the pairs. Arbitrary time slices or clusters can be constructed for the pairs of males within RCCs of 0-5 (close relatives) and 0-10 (pedigree). Indeed this process can be carried out to values up to the 1000s. Calibration of RCC with times to a MRCA was also carried out using data from 363 pairs of males who fell into four known pedigrees. The ratio of the two values was found to be skewed to higher values and so the author used the Hodges Lehmann estimator to give the conversion of 1 RCC to 43.3 years. In addition Howard outlines three methods of estimating the time to a common ancestor (TCA) for a cluster of RCCs whose genealogy is unknown. All the methods are based on various summary statistics of the pairwise RCC: average RCC, the RCC standard deviation and point P, i.e. the point at which the

histogram of the RCCs initially encounters base noise at high RCC. The following estimators were derived by examining 273 pairs known to have descended from 7 known common ancestors:

$$\widehat{TCA} = 52.7 \times \text{RCC},$$
$$\widehat{TCA} = 102.0\text{SD} \times (\text{RCC}),$$
$$\widehat{TCA} = 26.4 \times \text{Point P}.$$

Each of these methods inherently utilizes the RCC-to-years conversion of 43.3 years. Next Howard compares the RCC to the mutation rates estimated by Chandler (2006). Across 37 markers, the mutation rate is 0.00492 per locus per generation. Hence based on this method 1 RCC $\approx$ 46 years. Based on simulation data the number of mutations is linear to the RCC in the range of 0-40/50 but outside this range it is not. This work is extended further to consider surname clusters by examining the histogram of RCC values where peaks possibly indicate classification of clades or subclades within haplogroups (Howard, 2009*b*). For this to be successful everyone included in the analysis must be a true descendant, i.e. there must be no NPTs. Nonetheless, reliance on a single summary statistic, RCC, is likely to lose information when estimating the TMRCA.

Several authors also propose methods of estimating the TMRCA for groups of males usually belonging to the same haplogroup and based on STR data and these are briefly mentioned. This includes the work of Wilson and Balding (1998) which is an extension of the method proposed by Walsh (2001) but involves in addition reconstructing a genealogy based on the data. They use the SMM and coalescent theory in a Bayesian framework and implement the analysis using Markov-chain Monte-Carlo techniques. Importantly the authors allow diffuse priors for the effective population size and mutation rate. They also introduce nuisance parameters for the internal nodes and implement a branch-swapping algorithm to allow faster movement through tree space. Klyosov (2009) on the other hand use first-order kinetics on the basis of an inferred ancestral haplotype. Adamov and Karzhavin (2010) present a method of estimating the time to a MRCA for groups which takes into account the population size whilst allowing for population growth over time.

2.2 Y-chromosome and Surnames

The use of surnames information to estimate some genetic variable is not a recent phenomenon: in 1875 the son of Charles Darwin, George, estimated inbreeding rate by the method of isonomy i.e. computing the frequency of individuals marrying with the same surname. The earliest association between the Y-chromosome and surnames began with 'satellited' Y chromosomes (Yqs) in 1973 (Jobling, 2001; Schmid et al., 1984). In this case 17 out of 50 men sharing the same surname possessed the same Yqs, an otherwise rare marker which arises when a detached part of chromosome 15 or 22 is attached to the Y-Chromosome, i.e. a translocation occurs. These males were found to have descended from the same French barrel maker who emigrated to Canada in 1665. A later case involved four Colombian families who shared the same Yqs (Giraldo et al., 1981). In this instance three of the families had the same surname though they were not known to be related despite living in the same city.

The discovery of binary polymorphisms and in-depth examination of STRs on the NRY (Kayser et al., 1997) led to their initial use in determining male genetic prehistory (Jobling, 2001). For example Thomas et al. (1998), typed six binary polymorphisms and six Y-STRs (DYS19, DYS388, DYS390, DYS391, DYS392, DYS393) to examine the ancestry of 106 Jewish men said to have patrilineal descend from Moses' brother Aaron and as such members of the priesthood (Cohanim). They were compared with 200 Levite males, since Moses was said to be a member of the Levi tribe. Cohanim and Levite males should have different haplotypes to other Jewish males and have a TMRCA no more recent than the Temple period (3,000-2,000 years before present). They defined a Cohen modal haplotype found to be much more frequent in the Cohen chromosomes ($> 60\%$) compared to the Levite chromosomes. Using the stepwise mutation model basis the TMRCA of the Cohanim using the Y-STRs data excluding DYS388 (due to possible violations of the SMM) was computed at 2,650 years before present (BP) assuming a mutation rate of 0.0021 per marker per generation and a generation time of 25 years per generation (ypg). Increasing this to 30 ypg the TMRCA rose to 3,180 years BP. Importantly however the surnames of the Cohanim were different, though many were named Cohen, Kahn or Kane (Jobling, 2001).

The subsequent paper by Foster et al. (1998), though not incorporating surnames, aimed to examine genetically the historical debate that the third US president

Thomas Jefferson fathered his slave Sally Hemings' children. Descendants of Thomas Woodson, Hemings' first son, believe him to have been Jefferson's son, though no historical evidence exists for this. However there is oral tradition suggesting that Hemings' last son, Eston Hemings Jefferson, was fathered by a son of Jefferson's sister, either Samuel or Peter Carr. Thus Foster et al. typed 7 SNPs, one mini-satellite and 11 Y-STRs (DYS19, DYS388, DYS389I, DYS389II, DYS389III, DYS389IV,DYS390, DYS391, DYS392, DYS393 and DXYS156Y) on the descendants of interest. Since Jefferson did not have any surviving sons, DNA was typed from his patriline namely descendants of his uncle Field Jefferson. The results were surprising: Eston's offspring appeared to match the Jefferson haplotype whilst the Woodson's did not. The Woodson's did however match each other with one exception: a case of presumed illegitimacy due to mismatches at seven of the 11 Y-STRs. The Carr descendants also closely matched each other though not any of the other samples. This evidence led Foster et al. to conclude that Jefferson did indeed father Sally Hemings' son Eston.

A key paper to examine the association between Y-STRs and surnames in the genealogical setting was by Sykes and Irven (2000) involving only 4 Y-STRs: DYS19, DYS390, DYS391 and DYS393. 48 males with the surname Sykes were sampled along with two sets of controls: 21 'Neighbors' recruited by the Sykes males and 139 'English' natives. Linguistically the Sykes surname is suggested to have the topographical origin of 'spring', 'stream' or 'boundary ditch', thus pointing to a multiple founders surname establishment (Sykes and Irven, 2000). However a single haplotype (15-23-11-14) and two one-step neighbour haplotypes were found in over 52% of the Sykes males and only 5% of the controls ('Neighbors' and 'English' natives). Assuming the stepwise mutational model of Kimura and Ohta (1978), there is little doubt that the two one-step neighbour haplotypes evolved from the modal Sykes haplotype. Additionally, given that the remaining haplotypes are not sufficiently frequent when compared to the controls, Sykes and Irven suggest that these haplotypes have been introduced as non-paternity events rather than from other 'Sykes' founders. Using all haplotypes different to the modal Sykes haplotype, the average non-paternity rate is calculated as 1.3% per generation assuming 23 generations have passed since the Sykes MRCA. Importantly the authors attribute their findings to the relatively low frequency of Sykes in Britain, consistent with there having been only one founder, and suggest that the Y-chromosome and surname association could possibly extend to other rare surnames. This in turn could lead to use in both a genealogical and forensic framework.

In the Irish context, Hill et al. (2000) typed 7 SNPs and 6 STRs (DYS19, DYS389I, DYS390, DYS391, DYS392, DYS393) in 221 Y-chromosomes, comparing those with historical Irish and non-Irish surname roots. The SNPs defined nine distinct haplogroups with one found to be most frequent in Ireland at 78.1%. In addition, the Irish surnames falling into this haplogroup, when segregated according to their prehistoric geographical origin, gradually increased in frequency moving westwards. Furthermore, the STR data from the Irish DNA suggest common ancestry for those in the most frequent haplogroup though this was not the case in the other haplogroups.

Outwith the British Isles, Manni et al. (2005) apply the clustering method of self-organizing maps (SOMs) to just under 10,000 distinct Dutch surnames (of frequency greater than 40 in the population) from 226 sample locations in order to find their geographic origin in the Netherlands. The SOMs for each surname were grouped according to those with a similar geographic pattern. By inferring the geographical origin of a given surname and sampling close to that locality it is much more likely that one will sample living descendants of surname founder(s) providing a sampling tool for population genetic studies. However given that as few as 20% of the living descendants may remain in the area of origin, migration is thought to have played a key role in altering the ancient genetic makeup of a region. With this being the case even in the Netherlands which has a comparatively recent surname establishment (from 1796 in the south to only 1811 in the north) the effect of migration may be more extreme in other European countries including Britain.

In King et al. (2006), a broad study of British surnames was carried out. It involved initially recruiting a sample of 150 males of varying surnames. The surnames were chosen to reflect the distribution of surnames of the English, Scottish and Welsh amongst the population of Great Britain. Subsequently, a second sample of men was recruited with matched surnames to the first sample. This resulted in 150 pairs of males with the same surname but chosen such that those with recent name changes or origins out with Britain were excluded. Also patrilineal relatives were excluded on the basis of a questionnaire. Each male was typed for 11 binary markers and 17 STRs: DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS434, DYS435, DYS436, DYS437, DYS438, DYS439, DYS460, DYS461, DYS462. On the basis of the SNP profiles the male-pairs were partitioned

into those sharing their haplogroup or not. The former had their TMRCA estimated for each pair of males based on the method of Walsh (2001) using a mean mutation rate of 0.002 per locus per generation and a mean generation time of 35 years. Although information on the frequency of each surname in 1996 was available, the authors chose to examine a relationship between TMRCA estimates and the frequency rank *within* the sample of the surnames. There appeared to be a negative relationship between the variables with up to 24% of the sample possibly sharing common ancestry through surname on the basis of the 95% credible regions lower bound overlapping the surname establishment period. The authors discuss the potential for a surname based Y-chromosome database which may aide in unsolved crimes concluding that it would be most successful for less common surnames.

Based on the unusual findings of Hill et al. (2000) which suggest Irish history is comparatively undisturbed by migration, Moore et al. (2006) sampled 796 Y-chromosomes from Ireland, typing eleven binary polymorphisms and the same 17 Y-STRs as King et al. (2006) except that Moore et al. defined marker DYS389B as the difference between the number of repeats between DYS389II and DYS389I. STR analysis revealed an 'Irish modal haplotype' (IMH) in males belonging to haplogroup R1b3, itself over 85% of the males studied. The IMH with its one step neighbour (IMH + 1) accounted for 8.2% of the samples. Furthermore, the frequency of IMH appears to increase unevenly across Ireland peaking to 16.9% (21.5% for IMH + 1) in northwestern Ireland within the historical region of the Ui Neill, a powerful royal lineage in ancient Gaelic Ireland. Comparison with British data, Capelli et al. (2003), shows that the truncated IMH (6 Y-STRs) is almost absent, except that in western and central parts of Scotland it reaches 7.3% (16.7% for the truncated IMH +1). Furthermore, using 7 Y-STRs, the truncated IMH is found at very low frequency (0.13%) across worldwide samples judging from the YHRD website (YHRD, 2010). An additional 59 males with surnames of Ui Neill origin were typed for the 17 Y-STRs. Haplogroup membership to R1b3 or not was inferred from the haplotype profiles. Those in R1b3 showed reduced mutational divergence patterns from the IMH compared to the R1b3 general Irish data. Furthermore the Ui Neill appeared to have had an impact on the frequency distribution of the mutational steps away from the IMH for the northwestern R1b3 Irish data, whose distribution is disrupted compared to the smooth distribution observed for the general Irish R1b3 data. Indeed the Ui Neill samples are significantly different to both a general R1b3 northwestern

Irish sample ($P < 0.001$) and a subset of this excluding males with a surname of Ui Neill origin ($P = 0.006$). The paper also applied a reduced median algorithm before constructing median joining networks of the Ui Neill sample to reveal a large cluster around the IMH. This cluster's TMRCA of 1,730 years BP places it at the early medieval period. It was computed from the ρ statistic using the mutation rate from Zhivotovsky et al. (2004), i.e. 0.00069 per locus per generation, where a generation is 25 years long. The later dating (1,090 years BP) of the corresponding cluster for the northwestern samples fits with the conclusion that the IMH rose in frequency due to social selection associated with the Ui Neill dynasty rather than merely reflecting its general predominance in northwest Ireland.

McEvoy and Bradley (2006) also focus on Irish males extending the ideas developed by Sykes and Irven (2000). Surname establishment in Ireland began in the early 10th century and typically involved patronymic surnames. The English conquest of Ireland from the late 12th century resulted in the change from Gaelic Irish to the English language. 1,125 Irish males were typed for 6 binary polymorphisms and the same 17 STRs as typed by Moore et al. (2006), though DYS385A and DYS385B were typed but excluded from any analysis. In addition, the 765 Irish chromosomes from Moore et al. (2006) acted as controls. R1b3 was the most predominant haplogroup, 90%, consistent with previous work Hill et al. (2000); Moore et al. (2006). The remaining largely belonged to IxI1b2. Thus only STRs were used in the analysis including computing the match probability of two males sharing the same surname. This was carried out only for those surnames with ≥ 50 samples. The average match probability across surnames was 8.15% compared to only 0.2% for the controls. The match probability varied greatly, reaching as much as 12.5% for 'Ryan' to 0.9% for 'Kelly'. An analysis of molecular variance (AMOVA, Excoffier et al. (1992)) shows that 19.6% of variability was due to between-surname differences with the remaining variation occurring between bearers of the same surname for 43 surnames. Indeed the relationship between surnames and paternal ancestry holds even when considering geographical substructuring. For example, AMOVA analysis applied to surnames from the North East showed higher between-surname variability at 30.6% compared to 20.4% for surnames in the midlands of Ireland.

For the 11 surnames with samples ≥ 50, MJ networks were constructed. The presence of a single descent cluster indicated a single founder for some surnames.

Others, in line with historical evidence of multiple founders, e.g. due to Anglicisation of Irish names such as McEvoy, showed two or more descent clusters. However, on other occasions the STR data was contrary to documentary evidence. For example, 'McCarthy' and 'McGuinness' have two descent clusters contrary to historical sources suggesting a single founder. For the most common surnames, 'Murphy' and 'Kelly', historically with multiple founder origins, the data showed a more diffuse network with more diversity as anticipated. Importantly, there was no correlation between the match probabilities, used as a measure of the number of ancestors, and the current numbers of bearers. Even though the genetic and historical evidence suggests a monogenic origin for 'O'Sullivan', there are >40,000 bearers. Contrast this with similar rare surnames also with a single founder such as 'O'Gara' (<1000 bearers). This suggests that historical social power may have resulted in the proliferation of particular surnames over others. McEvoy and Bradley (2006) then estimate the TMRCA of the 15 major descent clusters based on the MJ networks. They employ the same method as Moore et al. (2006) and use the mutation rate of Zhivotovsky et al. (2004) with a generation time of 25 years. Allowing for uncertainty in the point estimates of TMRCA, most clusters included the Irish surname establishment period (900-1200CE) within their estimates bar a cluster for 'McCarthy' which has a much earlier TMRCA. Even most convincing monogenic surnames ('Ryan' and 'O'Sullivan') have only 50-55% bearers derived from the single founder with McEvoy and Bradley citing the usual reasons for generation of minor lineages in the MJ networks namely NPTs. In addition they cite 'horizontal absorption' as another cause of minor linages. Historically, distinct Gaelic surnames were anglicised into several forms. For example, 'McGuinness' is derived from the Gaelic 'Mac Aonghusa' and historically so have 'McCreesh', 'Neeson' and 'McGarton'. The data from all of the anglicised surnames derived from 'Mac Aonghusa' fall nicely into the original MJ network for 'McGuinness'. The authors end by computing a NPT rate of 1.6% per generation for 'O'Sullivan' on the basis of a 35 year generation time thus setting the surname establishment 30 generations ago. Importantly, the years per generation conversion used here is much greater than that used to estimate the TMRCA of the cluster (25 years).

McEvoy et al. (2006) carry out admixture analysis on Irish males to examine the extent of Viking influence, which is documented from $\sim$ 795 BCE, though there are few linguistic remnants of this. Nonetheless, 47 males with 26 different surnames were recruited from areas historically associated with Viking settlements in Ireland making up the Norse surname group (NSG). Six binary polymorphisms

were examined and 13 Y-STRs typed: DYS19, DYS385 A/B, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS460, DYS461 and DYS462. The NSG was treated as an admixture of Irish and Scandinavian populations, samples of DNA from which were obtained from other published work, and authors considered comparisons with British data. The admixture coefficient was computed using six different methods using either the haplogroup or haplotype data. All indicated a large Irish contribution (>80%) except one estimate based on the binary markers which estimated 49%. McEvoy et al. conclude that current Irish DNA shows little Viking contribution through the effects of either genetic drift, multiple surname founders, non paternity and indeed there may have been English contribution.

In mainland Europe the work of Immel et al. (2006) used AMOVA (Excoffier et al., 1992) to determine the extent of correlation between haplotypes and surname in East Germany. 419 males were typed at 8 Y-STRS (DYS19, DYS385, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393) excluding DYS385 in the analysis due to ambiguous allele assignment. The surnames of the males were divided into three groups according to their origins, i.e. German (G), Slavic (S) or a mixture of German and Slavic (M). Based on AMOVA analysis, the G and M groups were virtually identical and were thus amalgamated into a single group (G+M) which was significantly different to the S group. This result is attributed to recent (19^{th} century) integration of Slavs into Germany rather than the earlier migration from 950-1100CE.

The 2007 work of King et al. discovers the presence in Britain of a haplogroup rarely found outwith Africa. Their initial study typed 11 binary markers in 421 males of native British origin discovering a potentially 'African' male. Typing an additional two SNPs determined a A1 haplogroup for the male concerned who knew of no genetic connection with Africa. Furthermore the male possessed a rare locative surname (anonymyized as 'R') with only 121 bearers. This led the authors to recruit 18 unrelated males bearing the same surname or a close variant. On these males they typed 12 binary markers and 17 Y-STRs (DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS434. DYS435, DYS436, DYS437, DYS438, DYS439, DYS460, DYS461, DYS462). Seven of the males carried haplogroup A1 and possessed close haplotypes. The estimated TMRCA of the males was 440 ± 330 years based on the ρ statistic calculated from the MJ network. This assumed a mutation rate of 0.002 per locus per generation and

a generation time of 35 years. Genealogical research resolved the men into two lineages with MRCAs born in 1788 and 1789. Typing a further 60 Y-STRs revealed only one additional mismatch to the three already found. On the basis of 73 markers the TMRCA was estimated much more recently at 140 ± 80 years. Adding the average age of the subjects to this estimate gives 1734 CE as the estimated date of coalescence for the seven males which overlaps with the initial estimate of TMRCA but not the second.

A study of the Viking presence in England was carried out by Bowden et al. (2008). They investigated by means of surname-based sampling and STR data the extent of the genetic contribution of the Viking presence over a thousand years ago in North West England. Linguistic evidence shows that there is a high presence of place names of Scandinavian origin across the parts of England known to have been under Viking control. In particular there is both archaeological and linguistic evidence of the Vikings in Wirral and West Lancashire. It was anticipated that simply typing a modern sample of residents would show a weak signal of Scandinavian DNA. 'Modern' DNA ($n = 149$) was collected on the basis of a two-generation residence typical of genetic studies whilst 'medieval' DNA ($n = 79$) was collected on the basis of both this criterion and the possession of a surname known to have existed prior to 1572, although possibly modernized since then. 13 binary markers and 6 Y-STRs (DYS19, DYS388, DYS390, DYS391, DYS392, DYS393) were typed in these participants. Additional control data were also used from previously published work (Capelli et al., 2003). Population differentiation tests were significant for the 'medieval' and 'modern' data for both Wirral and West Lancashire (p-value=0.032 and 0.006). Furthermore, the 'medieval' DNA showed more haplogroup diversity. Comparison to various British and Scandinavian controls, using the SNPs only, indicated a higher proportion of Viking DNA in the 'medieval' compared to the 'modern' samples. In addition admixture analysis, where Norwegian and British/Irish DNA were considered the parental populations, revealed a higher contribution from the Scandinavian DNA for the 'medieval' samples compared to the 'modern'. The difference was more substantial for Wirral (0.47 vs. 0.38) than for West Lancashire (0.51 vs. 0.48). Indeed further subsampling of less frequent surnames (<20,000 bearers) within the 'medieval' slightly increased the Scandinavian proportion further (Wirral: 0.51; West Lancashire: 0.53). Given that genetic drift will have played a key role in altering the ancient genetic structure, ancient rather than modern parental populations should

also be used in estimating admixture. This could be achieved by surname-based sampling rather than random sampling of the modern-day populations.

The study of surnames and Y-chromosomes relationship is not limited to Europe. For example Bedoya et al. (2006) typed 6 Y-STRs (DYS19, DYS388, DYS390, DYS391, DYS392, DYS393) in males from Oriente, Colombia. The males possessed one of five currently common surnames each said to have originated from a single Spanish founder. There was a clear relationship between haplotype profiles and surnames: for two surnames 94% and 85% of the males possessed the 'surname modal haplotype'. Extending this to include one-step differences to the modal haplotypes increased the percentage to 100 in both surnames. For the remaining three surnames between 50-82% of males shared the same modal haplotypes. This increased to between 73-91% when allowing for one-step differences. For these surnames there were several cases of non-paternity since some males had a three or more step difference to their surname modal haplotype. Hence an NPT rate of 0.74×10^{-2} per generation was computed. This was on the basis of an inferred mutation rate of 1.86×10^{-3} per locus per generation (using the one-step differences across all surnames) and assuming a 25 year generation time from the time of surname establishment in the mid 17th century. Contrast this with the more recent work of Oliveira et al. (2008), who assessed the correlation of Brazilian DNA with surnames with purported poor success. The study involved 55 matched surname pairs of males having 11 binary markers, the ALU YAP insertion and 3 Y-STRs (DYS19, DYS391 and DYS393) typed. No correlation between haplogroup and surname, was found though three pairs of males shared the same haplotype. However, the authors do not discuss the frequency of surnames as a confounding factor, the historical establishment of surnames in Brazil or take into account the low haplotype resolution.

The more recent work of King and Jobling (2009*a*) focussed largely on less frequent surnames. 40 surnames were examined each with ten or more samples and the degree of coancestry within each surname was the primary quantity of interest. As in King et al. (2006), an exclusion criterion on males with recent name changes, non-UK origins and patrilineal relatives in the study was applied. A further 110 males with names different from each other and from the surnames under study acted as controls. 17 binary markers were typed as well as 17 STRs (DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS434, DYS435, DYS436, DYS437, DYS438, DYS439, DYS460, DYS461, DYS462). The

origin of each surname was assigned to one of six categories: ambiguous/unknown, locative, nickname, occupational, patronymic/matronymic and topographic. The frequency of each surname based on the 1996 UK electoral register was also provided and this was the sum of all bearers of the surname and its variant spellings. Based on the haplogroup distribution the authors found that 70% of surnames were significantly different to the controls whilst common surnames such as Smith were very similar to the controls. In addition, the distribution of haplotypes was significantly different to the controls in 85% of the surnames, with haplotypes being shared across variant spellings of the surname. MJ networks were formed within each surname resulting in one or more descent cluster based on the two sets of markers. The proportion of males within each surname falling into a cluster was significantly correlated to the surname frequency rank. The TMRCAs of descent clusters were calculated using the ρ statistic based on a 35-year generation time and a mean mutation rate of 0.0015 per locus per generation adding on the mean age of males in the sample, 60 years. A high degree of common ancestry was found with 65% of TMRCA estimates for clusters falling in the time of surname establishment, increasing to 95% when considering the lower bound of TMRCA (based on the standard deviation of the ρ statistic). Three surnames' TMRCAs predate surname establishment completely, although this was for more common surnames. Conversely, for some clusters within surnames, the TMRCA upper limits were well below the time period where documentary evidence indicates those names were in use. The authors argue that this would indicate strong genetic drift, resulting in the MRCA of a descent cluster being considerably after the founding of the surname. For surnames showing a single dominant descent cluster the nonpaternity rate was estimated based on the assumption of a single founder and by constructing extreme genealogies but with the caveat that multiple founders would in fact result in a lower nonpaternity rate. A comparison of surnames likely to have single founders versus those with multiple founders shows that the two are not significantly different in the proportion of haplotypes lying in the largest cluster. The authors also conclude that the number of founders is a poor predictor of the number of descent clusters via a forward simulation study. A comparison with similar data from Ireland (McEvoy and Bradley, 2006) reveals stark differences to the British surnames. In particular several common surnames show clear clusters and there appears to be no significant correlation between the proportion of haplotypes in clusters and the frequency of the surname in the Irish data. However the mean TMRCA of the clusters within the Irish surnames, ∼990

years, is consistent with the time period of surname establishment in Ireland (∼900 years). Similarly, the mean TMRCA of the British clusters of ∼650 years lies in the British surname establishment period. The high degree of coancestry in Irish surnames compared to British surnames may be attributed to recent population drift due to the "Great Hunger" of the 19th century in Ireland and/or the proliferation of patrilineal dynasties in medieval Ireland. However it is important to note that the Irish results are based on both a lower mutation rate (0.00069) and years per generation conversion. The former would inflate estimates of TMRCA whilst the latter results in lower estimates of TMRCA in years thus the direct comparison with the British data may not be valid. The authors next justify further their suggestion of a Y-chromosome surname database made in King et al. (2006) based on a strong correlation in the study between the match probability and surname frequency rank. For the less common surnames an average match probability of 14.5% is achieved which increases to 23.5% and 28.5% when allowing for one and two-step STR mutational step differences, respectively. The use of public databases is also suggested as a potential resource though bias from self-reporting and self-selection of males who may be closely related may affect results. The authors end with the possible use of Y haplotyping combined with surnames to identify males who share common ancestry at a time depth that is intermediate between that of pedigrees and of the general population, to aid in the detection of disease causing genes.

2.3 Estimation of STR Mutation Rates

It is clear from the review in section 2.1 that dating of the time to a most recent common ancestor based on DNA will depend on the mutation rates of the loci. Additionally these rates will affect forensic match probabilities (Holtkemper et al., 2001) as well as other factors. As such, a thorough review of the literature on calibrating Y-STR mutation rates was made. Although no conclusive evidence exists, it is thought that STR mutations are due to DNA replication slippage given the absence of recombination in the Y-chromosome (Butler, 2009; Jobling et al., 2004).

Mutation rates at Y-STRs may be estimated by several methods: examining father-son pairs with known paternity, comparing males from deep rooting pedigrees of known ancestry and assessing samples of sperms from males. Figure 2.6 depicts the typing of an STR using the three methods mentioned.

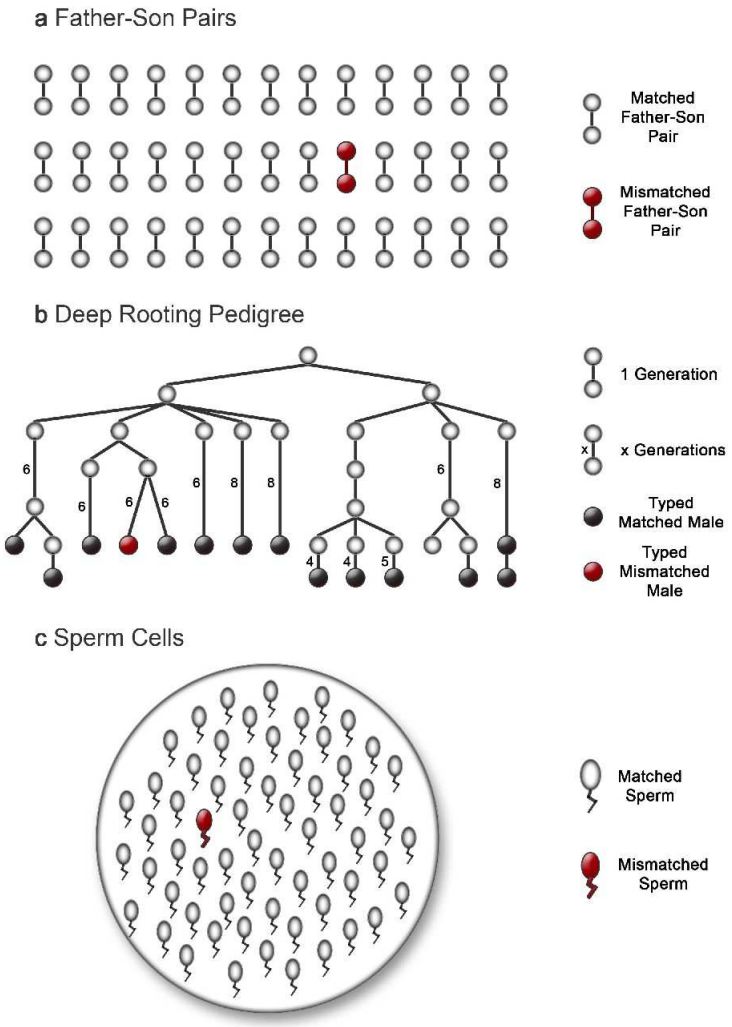

FIGURE 2.6: Schematic of methods for estimating mutation rates

In the case of father-son pairs, the estimated mutation rate is simply the number of mismatches divided by the total number of pairs examined, typically referred to as the number of meioses. Based on the example in figure 2.6a the mutation rate would be $1/39 = 0.0256$ mutations per marker per generation. Since mutation rates are rather low, in order to get precise mutation rate estimates, a large number of meioses are required and therefore a large number of father-son pairs. Typing data from pedigrees does exactly this. Although comparatively fewer males may be typed, the number of meioses is equivalent to the number of

links in the pedigree, greatly increasing the accuracy of mutation rate estimates. It operates by inferring an ancestral haplotype based on the male descendants' haplotypes and then computing the mutation rate by dividing the number of mutations inferred to have occurred in the pedigree by the overall number of meioses, for example giving a rate of $1/104 = 0.0096$ in figure 2.6b. However such mutation rates will be affected by any NPT (Kayser and Sajantila, 2001) and even in well documented pedigrees, cases of non paternity do occur (Heyer et al., 1997) which may artificially raise mutation rates. Furthermore within-pedigree differences may be considered either germline or somatic mutations. Somatic mutations are not passed down from parent to offspring in which case they should be excluded when computing the mutation rate. Typing samples of sperm cells is a strategy which also greatly increases the number of meioses. Figure 2.6c illustrates this and here a single sperm shows a difference across the 63 sperms typed from a single male giving a rate of $1/63 = 0.0159$. This method is rarely employed (Holtkemper et al., 2001), since single sperm typing is challenging.

In contrast to this there have been some estimates of evolutionary mutation rates based on simulation studies, such as that of Zhivotovsky et al. (2004) at 0.00069 mutations per locus per generation, though many in the genealogical circles argue that these rates are greatly underestimated (Athey, 2006; Chandler, 2006).

Three mutation rate reviews were carried out at various stages of the research. Since the results presented in later chapters were based on a particular set of mutation rate estimates, the three reviews are clearly detailed below.

2.3.1 Initial Mutation Rate Estimates

My first set of estimates for nine loci were taken directly from the online Y Chromosome Haplotype Reference Database (YHRD; Y Chromosome Haplotype Reference Database (2008)), which was based on research from 12 published articles and two unpublished sources (table 2.1). Heyer et al. (1997) based their estimates on pedigree data from the Saguenay Region in Canada, whilst the remaining used data from father-son pairs. No information was given on the methods of estimation for the two unpublished papers. The resulting estimates for the nine STRs are given in table 2.2.

TABLE 2.1: Sources for initial mutation rate review

Author(s)	Year of Publication
Heyer et al.	1997
Bianchi et al.	1998
Kayser et al.	2000
Dupuy et al.	2004
Kurihara et al.	2004
Ballard et al.	2005
Budowle et al.	2005
Gusmão et al.	2005
Hohoff et al.	2007
Domingues et al.	2007
Lee et al.	2007
Decker et al.	2008
Sergey Kravchenko	N/A
Gerhard Baessler	N/A

We compared the TMRCA results for the data from King et al. (2006) based on these site-specific rates and on their average mutation rate of 0.002390 mutations per locus per generation.

TABLE 2.2: Initial mutation rate estimates (per locus per generation)

STR Marker	Mutation Rate (per locus per generation)
DYS19	0.002449
DYS389I	0.002370
DYS389II	0.003430
DYS390	0.002367
DYS391	0.002834
DYS392	0.000455
DYS393	0.000792
DYS438	0.000468
DYS439	0.006348

2.3.2 Intermediary Mutation Rate Estimates

The second set of estimates for 86 markers was based on research from 29 published papers (table 2.3) and the two unpublished results from the YHRD mentioned above. Of the published papers, only one employed sperm typing (Holtkemper et al., 2001), four used data from deep rooting pedigrees (Bonné-Tamir et al., 2003; Heyer et al., 1997; Toscanini et al., 2008; Vermeulen et al., 2009) whilst the remaining 24 papers estimated mutation rates based on father-son pairs. Three additional papers were considered but excluded from the review: data from de Souza Goes et al. (2005) overlapped with Gusmão et al. (2005) (pers. comm. Gusmao); Kayser et al. (1997) overlapped with Kayser et al. (2000); data from Liu et al. (2007) was excluded as kinship could not be confirmed due to low LR values (pers. comm. Gusmao).

TABLE 2.3: Sources for intermediary mutation rate review

Author(s)	Year of Publication
Heyer et al.	1997
Bianchi et al.	1998
Lessig and Edelmann	1998
Pestoni et al.	1999
Schneider et al.	1998
Kayser et al.	2000
Dupuy et al.	2001
Holtkemper et al.	2001
Tsai et al.	2002
Bonné-Tamir et al.	2003
Dupuy et al.	2004
Kurihara et al.	2004
Ballard et al.	2005
Berger et al.	2005
Budowle et al.	2005
Gusmão et al.	2005
Turrina et al.	2006
Hohoff et al.	2007
Domingues et al.	2007
Lee et al.	2007
Pontes et al.	2007
Shi et al.	2007
Decker et al.	2008
Sanchez-Diz et al.	2008
Toscanini et al.	2008
Ge et al.	2009
Goedbloed et al.	2009
Kim et al.	2009
Vermeulen et al.	2009
Sergey Kravchenko	N/A
Gerhard Baessler	N/A

2.3.3 Final Mutation Rate Estimates

The final mutation rates review produced estimates for 94 Y-STRs based on the intermediary rates and an additional ten published papers (table 2.4), a total of 41 sources. Inadvertently, the data from de Souza Goes et al. (2005) was included.

TABLE 2.4: Additional sources for final mutation rate review

Author(s)	Year of Publication
Foster et al.	1998
de Souza Goes et al.	2005
Mulero et al.	2006
King et al.	2007
Padilla-Gutierrez et al.	2008
Pollin et al.	2008
Farfán and Prieto	2009
King and Jobling	2009*a*
Onofri et al.	2009
Vieira-Silva et al.	2009

In figure 2.7 the distribution of the mutation rates across all the markers is shown. There is a right-skewed distribution with a peak at zero. A total of 620 mutations out of 264,672 meioses were found in this review resulting in an average mutation

rate of 0.002343 (per marker per generation) with a 95% CI of 0.002162 to 0.002534, based on the Poisson approximation to the binomial.

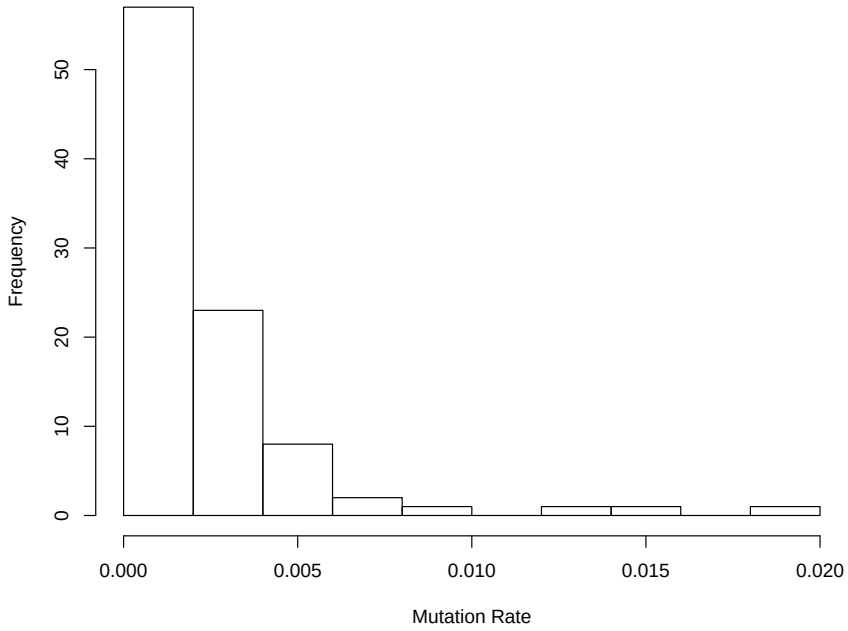

FIGURE 2.7: Histogram of mutation rates

In addition, the directionality of 516 mutations was known: 296 were increase mutations, i.e. 57.36% of the mutations increased the number of repeats of the STR. Applying a one proportion test of the null hypothesis that the proportion of increase (ν_+) and decrease (ν_-) mutations are equal i.e. $\nu_+ = \nu_- = 0.5$ returns a p-value lower than 0.05 at 0.001. Thus we can reject the null hypothesis and the 95% CI of the proportion of increase mutations lies between 0.530 and 0.617.

The empirical mutation rates at each STR is now detailed (table 2.5). For some markers, e.g. DYS425, as few as 30 meioses were examined, whereas for more commonly used markers, such DYS19 and DYS390, this figure was around 25,000. The most mutable marker was DYS449 with a mutation rate estimate of 0.018970 (95% Poisson CI: 0.007627 to 0.039086). A graphical summary of the markers with mutations in order of decreasing estimated mutation rate is given in figure 2.8. 49 markers had an estimated mutation rate of zero, based on no mutations observed in a low number of meioses ranging from 1058 to only 30 in some instances.

Of the 45 markers which were found to be variable, only 26 had information on the directionality of some of their mutations. A summary of the proportion of their

TABLE 2.5: Final mutation rate estimates (per locus per generation)

STR Marker	Mutations	Meioses	Mutation Rate (per site per generation)	Lower 95% CI	Upper 95% CI
DYS449	7	369	0.018970	0.007627	0.039086
DYS576	9	573	0.015707	0.007182	0.029816
DYS570	7	573	0.012216	0.004912	0.025170
YCAIII	1	100	0.010000	0.000253	0.055716
DYS481	3	433	0.006928	0.001429	0.020248
DYS458	48	7181	0.006684	0.004928	0.008862
Y GATA A4	56	11078	0.005055	0.003819	0.006564
DYS508	2	433	0.004619	0.000559	0.016685
DYS565	2	433	0.004619	0.000559	0.016685
DYS573	2	433	0.004619	0.000559	0.016685
DYS640	2	433	0.004619	0.000559	0.016685
GATA A10	5	1145	0.004367	0.001418	0.010190
DYS447	3	688	0.004360	0.000899	0.012743
DYS456	31	7162	0.004328	0.002941	0.006144
DYS460	6	1582	0.003793	0.001392	0.008255
DYS389II	53	14699	0.003606	0.002700	0.004716
Y GATA C4	28	7932	0.003530	0.002346	0.005102
DYS533	2	573	0.003490	0.000422	0.012609
DYS464	5	1476	0.003388	0.001100	0.007905
DYS446	2	658	0.003040	0.000368	0.010980
DYS389I	40	14728	0.002716	0.001940	0.003698
Y GATA H4	22	8167	0.002694	0.001688	0.004078
DYS391	40	16869	0.002371	0.001694	0.003229
DYS461	3	1266	0.002370	0.000488	0.006925
DYS19	61	25811	0.002363	0.001808	0.003036
DYS485	1	433	0.002309	0.000058	0.012866
DYS487	1	433	0.002309	0.000058	0.012866
DYS497	1	433	0.002309	0.000058	0.012866
DYS511	1	433	0.002309	0.000058	0.012866
DYS537	1	433	0.002309	0.000058	0.012866
DYS554	1	433	0.002309	0.000058	0.012866
DYS572	1	433	0.002309	0.000058	0.012866
DYS575	1	433	0.002309	0.000058	0.012866
DYS636	1	433	0.002309	0.000058	0.012866
DYS638	1	433	0.002309	0.000058	0.012866
DYS385a,b	60	28159	0.002131	0.001626	0.002743
DYS390	52	24998	0.002080	0.001554	0.002728
DYS549	1	573	0.001745	0.000044	0.009724
DYS448	11	7324	0.001502	0.000750	0.002687
YCAII	3	2356	0.001273	0.000263	0.003721
DYS437	12	10864	0.001105	0.000570	0.001929
DYS393	15	14662	0.001023	0.000573	0.001687
DYS388	2	2865	0.000698	0.000085	0.002522
DYS392	9	16000	0.000563	0.000257	0.001068
DYS438	5	10963	0.000456	0.000148	0.001064
DXYS156	0	1058	0.000000	0.000000	0.003487
DYS643	0	573	0.000000	0.000000	0.006438
DYS522	0	543	0.000000	0.000000	0.006794
DYS531	0	513	0.000000	0.000000	0.007191
DYS435	0	435	0.000000	0.000000	0.008480
DYS472	0	433	0.000000	0.000000	0.008519
DYS476	0	433	0.000000	0.000000	0.008519
DYS480	0	433	0.000000	0.000000	0.008519
DYS488	0	433	0.000000	0.000000	0.008519
DYS490	0	433	0.000000	0.000000	0.008519
DYS491	0	433	0.000000	0.000000	0.008519
DYS492	0	433	0.000000	0.000000	0.008519
DYS494	0	433	0.000000	0.000000	0.008519
DYS495	0	433	0.000000	0.000000	0.008519
DYS505	0	433	0.000000	0.000000	0.008519
DYS525	0	433	0.000000	0.000000	0.008519
DYS530	0	433	0.000000	0.000000	0.008519
DYS540	0	433	0.000000	0.000000	0.008519
DYS556	0	433	0.000000	0.000000	0.008519
DYS567	0	433	0.000000	0.000000	0.008519
DYS568	0	433	0.000000	0.000000	0.008519
DYS569	0	433	0.000000	0.000000	0.008519
DYS578	0	433	0.000000	0.000000	0.008519
DYS579	0	433	0.000000	0.000000	0.008519
DYS580	0	433	0.000000	0.000000	0.008519
DYS583	0	433	0.000000	0.000000	0.008519
DYS589	0	433	0.000000	0.000000	0.008519
DYS590	0	433	0.000000	0.000000	0.008519
DYS594	0	433	0.000000	0.000000	0.008519
DYS617	0	433	0.000000	0.000000	0.008519
DYS618	0	433	0.000000	0.000000	0.008519
DYS641	0	433	0.000000	0.000000	0.008519
DYS434	0	274	0.000000	0.000000	0.013463
DYS436	0	274	0.000000	0.000000	0.013463
DYS462	0	274	0.000000	0.000000	0.013463
DYS426	0	169	0.000000	0.000000	0.021828
YCAI	0	150	0.000000	0.000000	0.024593
GGAATIB07	0	119	0.000000	0.000000	0.030999
DYS557	0	110	0.000000	0.000000	0.033535
DYS443	0	80	0.000000	0.000000	0.046111
DYS444	0	80	0.000000	0.000000	0.046111
DYS520	0	80	0.000000	0.000000	0.046111
DYS622	0	80	0.000000	0.000000	0.046111
DYS630	0	80	0.000000	0.000000	0.046111
DYS709	0	80	0.000000	0.000000	0.046111
DYF386SI	0	30	0.000000	0.000000	0.122963
DYF390SI	0	30	0.000000	0.000000	0.122963
DYF406SI	0	30	0.000000	0.000000	0.122963
DYS425	0	30	0.000000	0.000000	0.122963

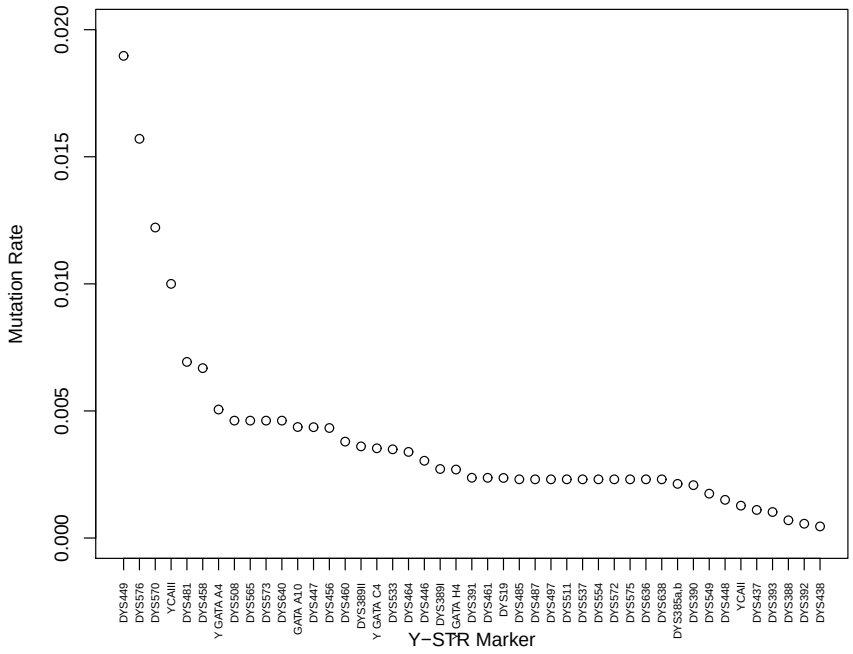

FIGURE 2.8: Estimated mutation rate (per locus per generation) vs. STR marker

increase mutations and associated one proportion p-values and 95% CI is given in table 2.6.

TABLE 2.6: STR markers one proportion test results

STR Marker	Increase Mutations	Total Mutations	Proportion of Increase Mutations	P-value	Lower 95% CI	Upper 95% CI
DYS19	38	53	0.7167	0.0025	0.5744	0.8280
DYS385a,b	39	50	0.7800	0.0001	0.6367	0.8801
DYS388	0	1	0.0000	1.0000	0.0000	0.9454
DYS389I	14	33	0.4242	0.4862	0.2595	0.6061
DYS389II	22	45	0.4889	1.0000	0.3394	0.6402
DYS390	33	46	0.7174	0.0051	0.5632	0.8354
DYS391	24	39	0.6154	0.2002	0.4465	0.7619
DYS392	4	6	0.6667	0.6831	0.2411	0.9400
DYS393	8	15	0.5333	1.0000	0.2742	0.7772
DYS437	9	12	0.7500	0.1489	0.4284	0.9331
DYS438	1	2	0.5000	1.0000	0.0945	0.9055
Y GATA A4	24	53	0.4528	0.5827	0.3180	0.5945
DYS446	1	2	0.5000	1.0000	0.0945	0.9055
DYS447	0	3	0.0000	0.2482	0.0000	0.6900
DYS448	5	11	0.4545	1.0000	0.1814	0.7544
DYS449	5	7	0.7143	0.4497	0.3026	0.9489
DYS456	15	28	0.5357	0.8501	0.3421	0.7199
DYS458	25	45	0.5556	0.5510	0.4012	0.7005
DYS460	1	5	0.2000	0.3711	0.0105	0.7012
DYS461	2	3	0.6667	1.0000	0.1253	0.9823
DYS464	3	4	0.7500	0.6171	0.2194	0.9868
DYS576	1	1	1.0000	1.0000	0.0546	1.0000
Y GATA C4	9	23	0.3913	0.4042	0.2047	0.6122
GATA A10	2	5	0.4000	1.0000	0.0726	0.8296
Y GATA H4	10	21	0.4762	1.0000	0.2639	0.6966
YCAII	1	3	0.3333	1.0000	0.0177	0.8747

For these 26 markers, we plot their estimated mutation rates versus the proportion of increase mutation (figure 2.9) to find that there may be a positive linear relationship between the variables. For example, DYS576 has both a high proportion of increase mutations and a high estimated mutation rate. Conversely, DYS447 has a low proportion of increase mutations and a low mutation rate. Applying a

linear regression results in the following fitted line:

$$\widehat{\mu} = 0.0005532 + 0.0062972 \times \nu_{+}, \tag{2.27}$$

where $\hat{\mu}$ is the estimated, per locus per generation, mutation rate and ν_{+} is the proportion of increase mutations. However, the explanatory variable is not significant ($\mathrm{p} - \mathrm{value} = 0.0925$).

In addition the assumptions of constant variance and linearity are doubtful and the normal Q-Q plot shows curvature indicating the violation of normality (data not shown).

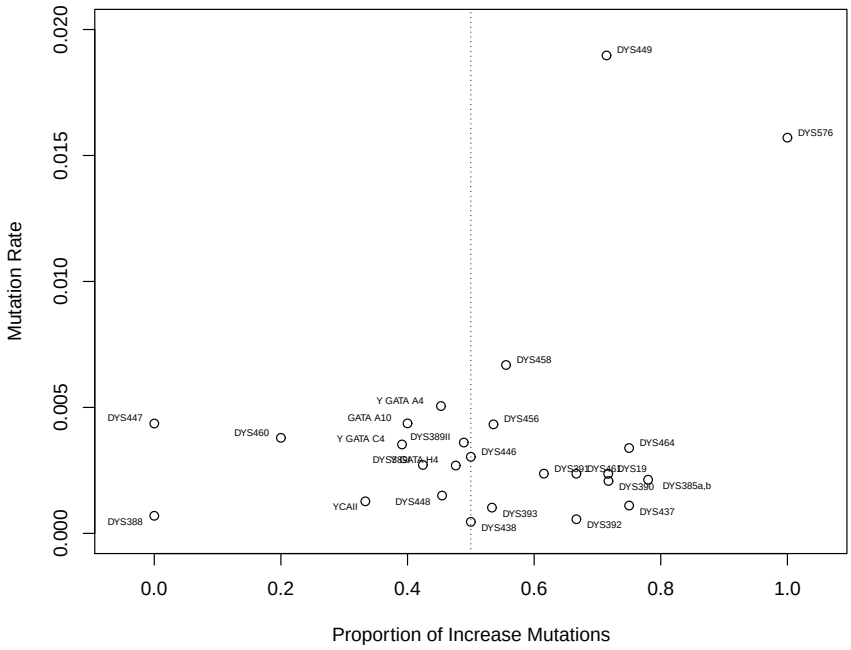

FIGURE 2.9: Estimated mutation rate (per locus per generation) vs. proportion of increase mutations

Next we looked for any relationship between the mutation rate across individual markers and the length and complexity of the repeat sequence and also whether the marker is a single or multi-copy (Burgarella and Navascues, 2011; Kayser et al., 1997, 2004; Malaspina et al., 1997; Torres-Rodriguez et al., 2006; US National Institute of Standards and Technology, 2009). This information is detailed in table 2.7.

Just over 3% of the markers involved dinucleotide repeats, 21% were trinucleotide, just under 65% were tetranucleotide markers, under 10% were pentanucleotide markers and a little over 1% were hexanucleotide markers. In figure 2.10 there is

TABLE 2.7: Summary of STR markers properties

STR Marker	Repeat Unit Length	Copy Type	Complexity
DXYS156	5		
DYF386SI	3	single	simple
DYF390SI	4	single	simple
DYF406SI	4	single	simple
DYS19	4	single	complex
DYS385a,b	4	multi	complex
DYS388	4	single	simple
DYS389I	4	single	complex
DYS389II	4	single	complex
DYS390	4	single	complex
DYS391	4	single	complex
DYS392	3	single	complex
DYS393	4	single	simple
DYS425	3	single	simple
DYS426	3	single	simple
DYS434	4	single	simple
DYS435	4	single	simple
DYS436	3	single	simple
DYS437	4	single	complex
DYS438	5	single	simple
Y GATA A4	4	single	complex
DYS443	4	multi	complex
DYS444	4	single	complex
DYS446	5		simple
DYS447	5		complex
DYS448	6		complex
DYS449	4		complex
DYS456	4		simple
DYS458	4		simple
DYS460	4	single	simple
DYS461	4	single	simple
DYS462	4		simple
DYS464	4		
DYS472	3	single	simple
DYS476	3	single	simple
DYS480	3	single	simple
DYS481	3	single	simple
DYS485	3	single	simple
DYS487	3	single	simple
DYS488	3	single	simple
DYS490	3	single	simple
DYS491	3	single	simple
DYS492	3	single	simple
DYS494	3	single	simple
DYS495	3	single	simple
DYS497	3	single	simple
DYS505	4	single	simple
DYS508	4	single	simple
DYS511	4	single	simple
DYS520	4	single	complex
DYS522	4	single	simple
DYS525	4	single	simple
DYS530	4	single	simple
DYS531	4	single	simple
DYS533	4	single	simple
DYS537	4	single	simple
DYS540	4	single	simple
DYS549	4	single	simple
DYS554	4	single	simple
DYS556	4	single	simple
DYS557	4	single	complex
DYS565	4	single	simple
DYS567	4	single	simple
DYS568	4	single	simple
DYS569	4	single	simple
DYS570	4	single	simple
DYS572	4	single	simple
DYS573	4	single	simple
DYS575	4	single	simple
DYS576	4	single	simple
DYS578	4	single	simple
DYS579	4	single	simple
DYS580	4	single	simple
DYS583	4	single	simple
DYS589	5	single	simple
DYS590	5	single	simple
DYS594	5	single	simple
DYS617	3	single	simple
DYS618	3	single	simple
DYS622	4	single	complex
DYS630	4	single	complex
Y GATA C4	4	single	complex
DYS636	4	single	simple
DYS638	4	single	simple
DYS640	4	single	simple
DYS641	4	single	simple
DYS643	5	single	simple
DYS709	4		complex
GATA A10	4	single	complex
GGAATIB07	5		
Y GATA H4	4	single	complex
YCAI	2		
YCAII	2		
YCAIII	2		

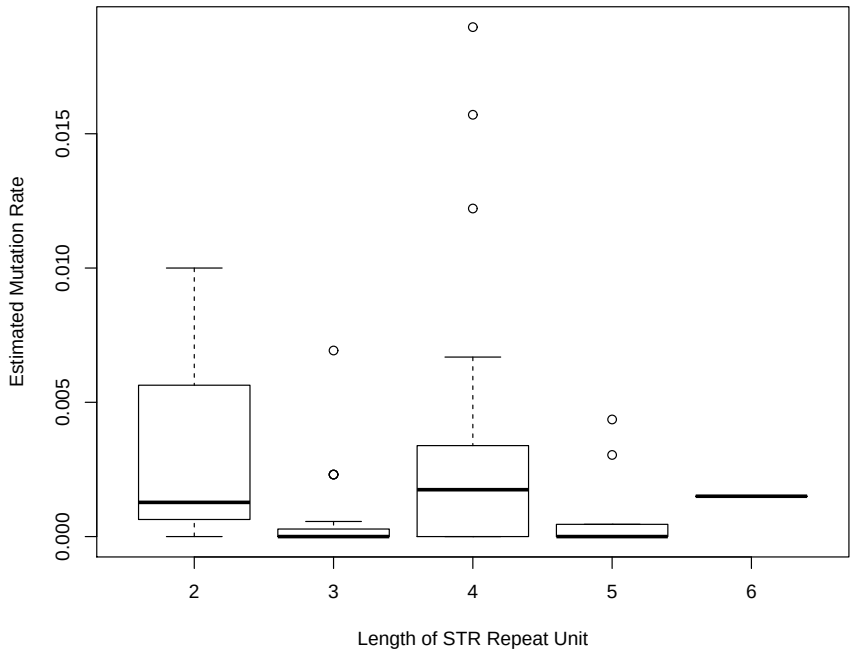

FIGURE 2.10: Estimated mutation rate by length of STR repeat unit

no clear relationship between the mutation rate and the length of the repetitive unit within the STR. Applying a one-way ANOVA shows that the repeat unit length is not a statistically significant predictor of the estimated mutation rate ($p - value = 0.2307$), though with very questionable statistical assumptions (data not shown).

Of the 94 STRs examined, 80 were classified either as single or multi-copy markers. Of these only DYS385a,b and DYS443 were a multi-copy markers, thus 97.87% of the classified markers were single copy markers. Given the low number of multi-copy markers additional analysis was not carried out.

88 of the STRs were classified according to their complexity: 66 were simple and 22 were complex markers. For the simple markers 152 mutations occurred in 69992 meioses, whilst for the complex markers these figures were 459 and 189421 respectively. Histograms for mutation rate estimates for the two types of markers (fig. 2.11) show right-skewed distributions for both. For the simple markers we have an average mutation rate of 0.002171 (95% Poisson CI: 0.001840, 0.002546) and a slightly higher rate for the complex markers of 0.002423 (95% Poisson CI: 0.002207, 0.002655). In order to test for differences in the average mutation rate between the two types of complexities, we apply an Exact Rate Ratio test using the function `rateratio.test()` in R (Lehmann, 1986). Here we test the null that the two average mutation rates are equal, i.e. that $\mu_{\text{simple}}/\mu_{\text{complex}} = 1$. In this

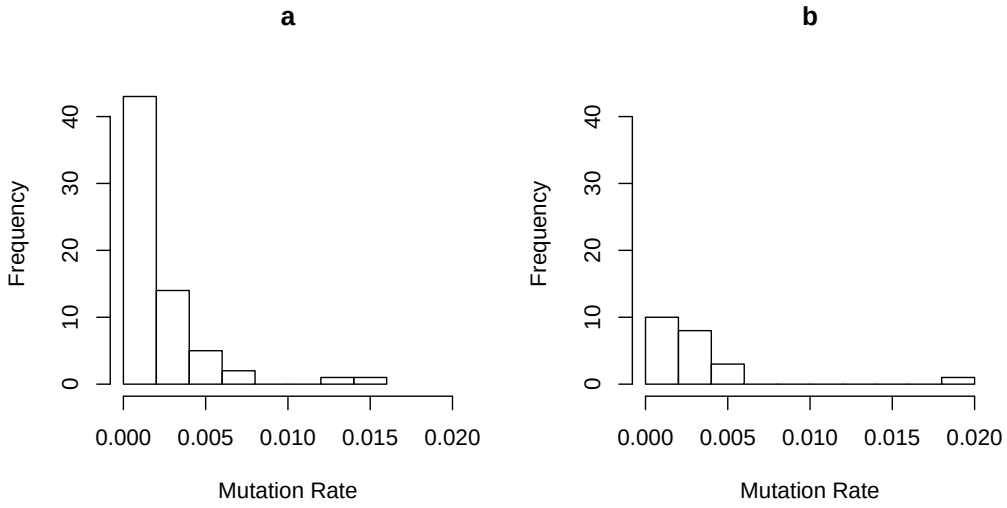

FIGURE 2.11: Estimated mutation rate (per locus per generation) by STR complexity: a. simple markers b. complex markers

case we produce a rate ratio of 0.8962 with a 95% CI of 0.7410 and 1.0789. We cannot reject the null hypothesis that the average rates are equal.

TABLE 2.8: Multi-step mutations

STR Marker(s)	Size	Direction
DYS19	2	Decrease
DYS385	2	Increase
DYS385	2	Decrease
DYS390	2	Increase
DYS438	4	Decrease
DYS456	3	Decrease
DYS458	2	Decrease
YCAIII	2	Decrease

As have most studies on STR mutation rates, we found that single-step mutations were predominant: only 8 mutations of the 620 mutations were multi-step mutations i.e. 1.29% (95%CI $0.60 - 2.63\%$). Further details of the direction and size of these mutations are given in table 2.8. Five of the multi-step mutations involve reductions in the number of repeats of the STRs, suggesting a slightly higher proportion of decrease multi step mutations. Subjectively, there is a positive linear relationship between the average mutation rate at a given multi-step STR and the nature of the multi-step marker (fig. 2.12). However, fitting a straight line to the data show a negative linear relationship between the variables such that:

$$\hat{\mu} = 0.0035087 - 0.0003006 s_{\text{multi}}, \tag{2.28}$$

where $\hat{\mu}$ is the estimated, per locus per generation, mutation rate and s_{multi} is the size and direction of multi-step mutation, and albeit insignificant ($\text{P} - \text{value} > 0.05$) and with questionable model assumptions and indeed excluding all one-step mutations. Next we excluded marker YCAIII (which has no data on the number

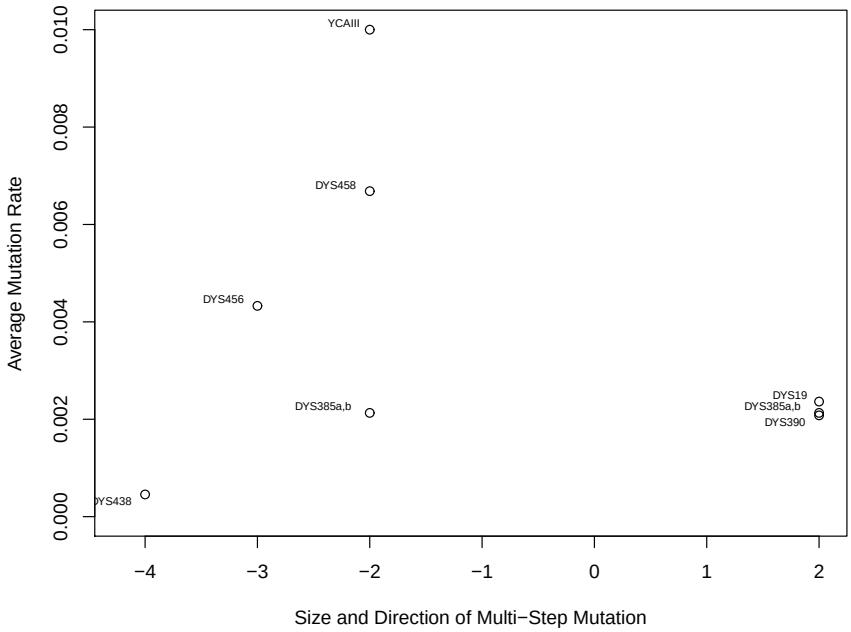

FIGURE 2.12: Estimated mutation rate (per locus per generation) vs. size and direction of multi-step mutation

of increase and decrease mutations) and examined the relationship between the proportion of increase mutations versus the size and direction of the multi-step markers. In figure 2.13 we see a positive linear relationship between the variables. Fitting a straight line in this context gives the fitted line:

$$\hat{\nu_+} = 0.67968 - 0.03442 s_{\text{multi}}, \tag{2.29}$$

where ν_+ is the proportion of increase mutations. Although the regression is not quite significant ($\text{P} - \text{value} = 0.052$). Here too we exclude all the one-step increase and decrease mutations.

In summary, our final mutation rate review shows a right-skewed distribution for the estimated mutation rate across 94 Y-STRs. The average estimated mutation rate is 0.002534 (per locus per generation) and the percentage of increase mutations was significantly higher at 57%. In addition, the majority of the mutations were single rather than multi-step ($> 98\%$). The length of the repeat unit, complexity

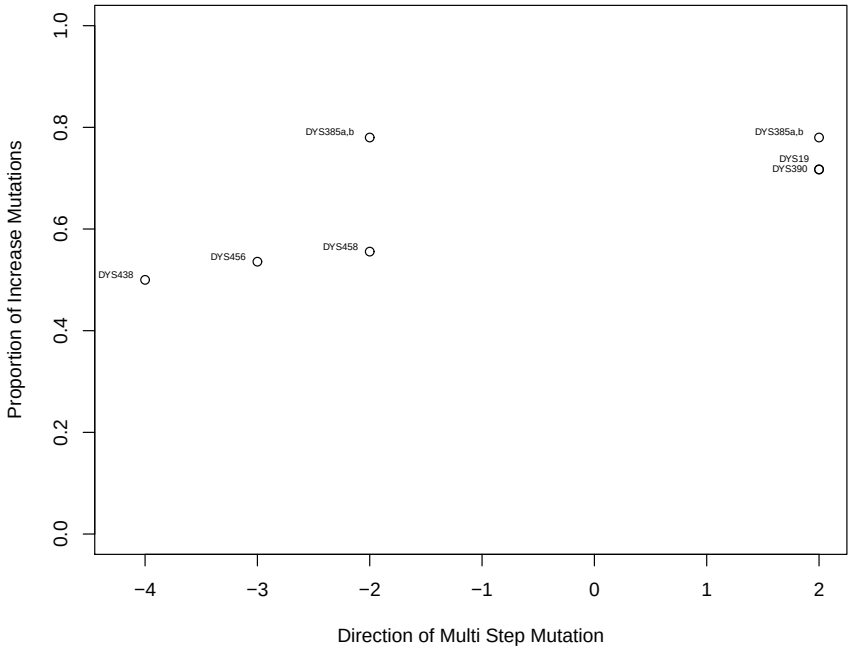

FIGURE 2.13: Proportion of increase mutations vs. size and direction of multi-step mutation

of marker and proportion of increase mutations were each not useful predictors of the estimated mutation rate. This was also the case for the size and direction of multi-step mutations.

Chapter 3

Maximum Likelihood Estimation of TMRCA

This chapter begins with a preliminary examination of the data from King et al. (2006), outlined in section 2.2, using the models developed by Walsh (2001). Thereafter a detailed study of simulated STR profiles of pairs of males will used to compare the competing models used to estimate the TMRCA.

3.1 Preliminary Analysis

In order to provide a foundation for subsequent analysis, this preliminary analysis is concerned with evaluating the SMM and IAM estimates of TMRCA on real data. Thus the methodology used is based on the work of Walsh (2001) and the data is from King et al. (2006) which involved typing Y-STRs on pairs of males who shared the same surname.

3.1.1 Materials and Methods

In this section we implemented two different models based on the stepwise mutation model (SMM) of Walsh (2001) and his infinite alleles model (IAM) both detailed in 2.1.

To recap, we suppose we have data for a pair of males across n loci, i.e. the number of microsatellite repeats at each locus. Based on this data we aim to estimate the

time t to a most recent common ancestor (MRCA). In the IAM, we simplify the data to either a match or mismatch such that for $i = 1, \ldots, n$:

$$x_i = \begin{cases} 1 & \text{match at } i^{th} \text{ locus;} \\ 0 & \text{mismatch at } i^{th} \text{ locus.} \end{cases}$$

Since we are ignoring the possibility of parallel or back mutations, the probability of a match is the probability that no mutations have occurred from the ancestor to male 1, i.e. $(1-\mu_i)^t$, where μ_i is the mutation rate at the i^{th} locus. Thus the probability of a match after t generations at the i^{th} locus across both lineages, $q_i(t)$, is:

$$q_i(t) = (1-\mu_i)^{2t}. \tag{3.1}$$

Unlike Walsh (2001), we do not approximate this probability with $e^{-2t\mu_i}$ (which is obtained when modelling the number of mutations as a Poisson distribution with parameter $2t\mu_i$). So the match/mismatch at the i^{th} locus is distributed as a $Bi(1, q_i(t))$ so

$$P(X_i = x_i|\mu_i, t) = q_i(t)^{x_i}\big(1-q_i(t)\big)^{1-x_i}. \tag{3.2}$$

Thus the likelihood of t for the data across n loci can be written as:

$$L(x_1, \ldots, x_n|t) = \prod_{i=1}^{n}(1-\mu_i)^{2tx_i}\big(1-(1-\mu_i)^{2t}\big)^{1-x_i} \tag{3.3}$$

and the posterior distribution with an exponential prior on t (of rate λ) as:

$$\begin{aligned} P(t|x_1, \ldots, x_n) &\propto L(x_1, \ldots, x_n|t)p(t) \\ &\propto e^{-\lambda t}\prod_{i=1}^{n}(1-\mu_i)^{2tx_i}\big(1-(1-\mu_i)^{2t}\big)^{1-x_i}. \end{aligned} \tag{3.4}$$

We also allow for different mutation rates at each locus in our SMM, which is otherwise similar to Walsh's. The microsatellite repeat counts in this case are modelled as a random walk from male 1 to male 2 (fig. 2.5) and is summarised again. The observed data are the differences of the number of repeats of the microsatellites for each male, which are also the differences in the numbers of increase mutations, n_+, and decrease mutations, n_-, i.e. $|n_+ - n_-|$. This may be

either even or odd, i.e.

$$|n_+ - n_-| = 2b, \text{or}$$
$$|n_+ - n_-| = 2b + 1, \text{ where } b \text{ is a positive integer.}$$

Because both increase and decrease mutations can occur, the observed difference will not be equal in general to the actual number of total mutations that occur. We now detail the likelihood for the even case. Since the observed difference is even it follows that the sum of n_+ and n_- is even:

$$n_+ + n_- = 2a, \text{ where } a \text{ is a positive integer.}$$

Supposing that increase and decrease mutations are equally probable, n_+ is distributed as $Bi(2a, \frac{1}{2})$ as is n_- mutations. Thus we have:

$$P(|n_+ - n_-| = 2b | n_+ + n_- = 2a) = 2\frac{(2a)!}{(a+b)!(a-b)!}\left(\frac{1}{2}\right)^{2a}. \tag{3.5}$$

We also suppose that the total number of mutations $n_+ + n_-$ is distributed as $Po(2t\mu)$ and so:

$$P(n_+ + n_- = 2a|t) = e^{-2t\mu}\frac{(2t\mu)^{2a}}{(2a)!}. \tag{3.6}$$

Hence

$$\begin{aligned} P(|n_+ - n_-| = 2b|t) &= \sum_{a=b}^{\infty} P(|n_+ - n_-| = 2b | n_+ + n_- = 2a) P(n_+ + n_- = 2a|t) \\ &= \sum_{a=b}^{\infty} 2\binom{2a}{a-b}\left(\frac{1}{2}\right)^{2a} e^{-2t\mu}\frac{(2t\mu)^{2a}}{(2a)!} \\ &= 2e^{-2t\mu}\sum_{c=0}^{\infty}\frac{(2t\mu/2)^{2c+2b}}{\Gamma(c+1)\Gamma(c+2b+1)} \end{aligned} \tag{3.7}$$

where $c = a - b$.

The summation in the last line is a modified type I Bessel function (Lebedev and Silverman, 1972; Olver, F. W. J. and National Institute of Standards and Technology (U.S.), 2010), so for the even case we have the likelihood at each microsatellite:

$$P(|n_+ - n_-| = 2b|t) = 2e^{-2t\mu}I_{2b}(2t\mu). \tag{3.8}$$

Similarly, for the odd case, the likelihood at each microsatellite is:

$$P(|n_+ - n_-| = 2b + 1|t) = 2e^{-2t\mu} I_{2b+1}(2t\mu). \tag{3.9}$$

Since the microsatellites are independent we can take the product of the likelihoods across microsatellites to get the overall likelihood and from it compute the maximum likelihood estimate of t. It is this that is examined in the results for some simulated data and also in some of the analysis of real data.

By applying an exponential prior of rate λ to TMRCA, we form the posterior distribution across n microsatellites. Here we use the notation $x_{j,i}$ to refer to the number of repeats of the i^{th} microsatellite ($i = 1, \ldots, n$) for the j^{th} male ($j = 1, 2$) and we also allow for site-specific mutation rates across microsatellites, i.e. μ_i. Thus we can write the posterior as follows:

$$\begin{aligned} P(t|x_{1,1}, \ldots, x_{1,n}, x_{2,1}, \ldots, x_{2,n}) &\propto e^{-\lambda t} \prod_{i=1}^{n} e^{-2t\mu_i} I_{|x_{1,i}-x_{2,i}|}(2t\mu_i) \\ &\propto \exp\left(-t(\lambda + 2\sum_{i=1}^{n} \mu_i)\right) \prod_{i=1}^{n} I_{|x_{1,i}-x_{2,i}|}(2t\mu_i). \end{aligned} \tag{3.10}$$

Our analysis begins by comparing estimates of TMRCA using the stepwise mutation model to those using the infinite alleles model for the data from King et al. (2006). Recall that King et al. typed 17 STRs (DYS19, DYS388, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS434, DYS435, DYS436, DYS437, DYS438, DYS439, DYS460, DYS461 and DYS462) in 150 pairs of males with each pair sharing the same surname. The surnames covered a wide range of frequencies, varying from common surnames to rarer ones, allowing the opportunity to model TMRCA against some measure of the frequency of the surname. Despite the availability of the raw frequency of surnames across Great Britain in 1996, King et al. examined their estimates of TMRCA against the *rank* of the surnames in the sample. Thus the most frequent surname was assigned 1 whilst the least frequent was ranked 150. Any relationship based on sample rank is not ideal; any new surname would not fit into this measure naturally and indeed the relationship itself would be altered. So we wished to model the relation of t against the UK frequency of the surnames or alternatively the rank of surnames in the UK population, a measure which was not included in the supplementary information of King

et al. (2006). Given the non-linear relationship between the frequency and population rank (figure 2A in King et al., 2006), we wished to consider both measures in our analysis. Professor Kevin Schurer, one of the contributing authors (King et al., 2006), kindly provided us with UK surname rank information and granted us permission to use it in this research (pers. comm. Schurer). This information involved ranking the surnames by frequency based on information in the 1996 UK electoral register covering those aged 18 or over, making up just under 44 million people. We also consider dividing the data according to those pairs of males who share the same haplogroup status and those who do not share haplogroup.

The posterior distribution in equations 3.10 and 3.4 for the SMM and IAM, respectively, with a common mutation rate of 0.002 mutation per locus per generation and $\lambda = 0.0002$ were maximised using the `optimize()` function in R for the King et al. (2006) data.

Following the initial mutation rate review (section 2.3.1) we compared the results using the SMM using a single average mutation rate, 0.002390, to using the site-specific mutation rates in the SMM (with $\lambda = 0.0002$). This analysis examined the results for nine of the available 17 Y-STRs, i.e. DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS438 and DYS439 in accordance with the initial mutation review which was limited to these Y-STRs.

3.1.2 Preliminary Results

We begin by comparing $\hat{t}$ using the SMM and IAM directly (fig. 3.1). For times up to about 75 generations the points straddle the line of equality roughly equally above and below suggesting there is not much difference in the SMM and IAM estimates of TMRCA. However, thereafter, the SMM produces substantially higher values than the IAM.

In figure 3.2, we plot the SMM and IAM estimates versus the reverse surname frequency and find there is a difference in the estimates. In general it appears that IAM produces lower estimates of TMRCA than SMM. Although this difference appears almost negligible under 100 generations, higher estimates of TMRCA produce a greater difference between the estimates from the two models. The reason for this more pronounced difference the further back in time is because more mutations will have occurred, and as the IAM does not differentiate between a

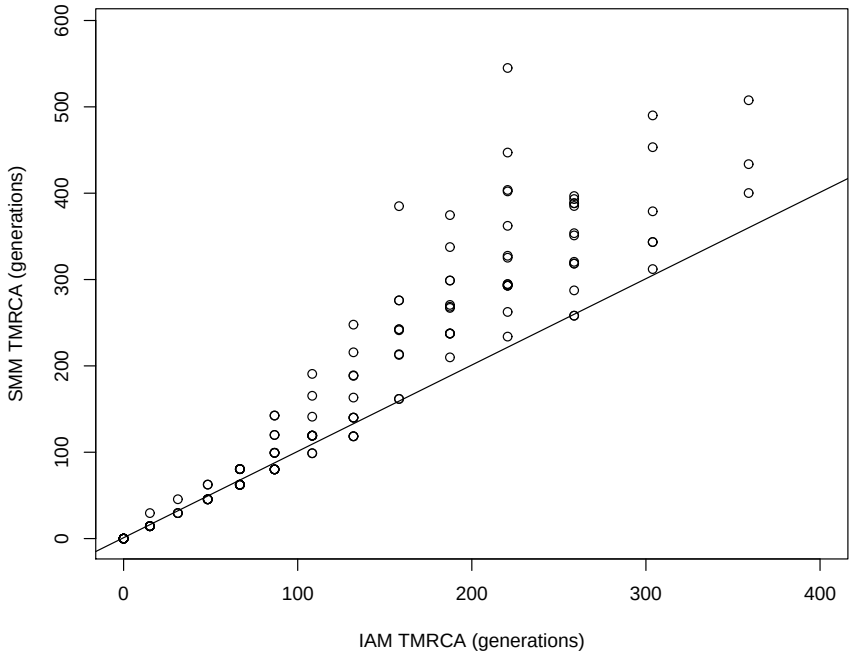

FIGURE 3.1: TMRCA estimates: SMM vs. IAM with the line of equality superimposed

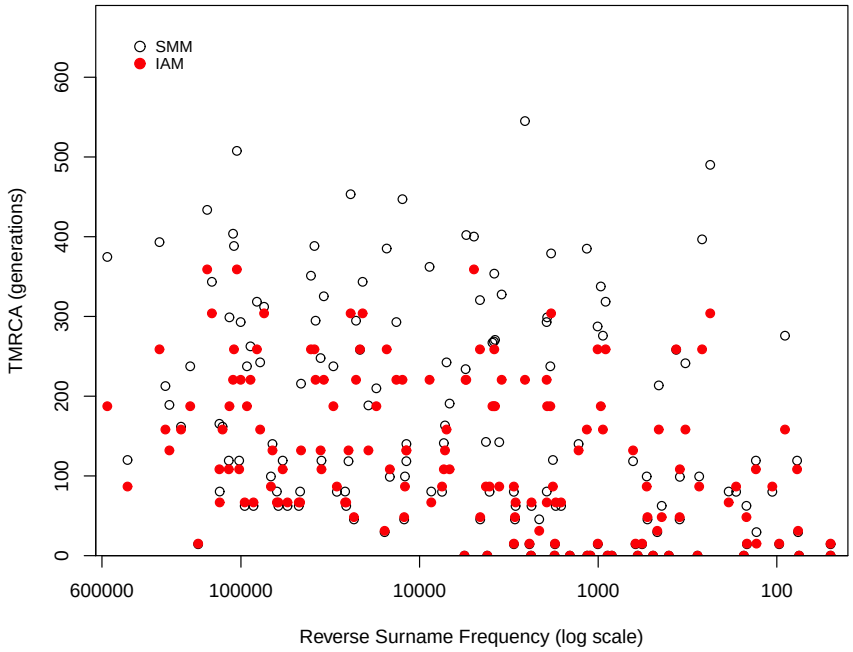

FIGURE 3.2: SMM and IAM: TMRCA estimates vs. log surname frequency

1-step mismatch or $>$ 1-step one, it produces lower estimates than the SMM. Nonetheless, we see that there may be some negative relationship between estimated TMRCA and the reverse log surname frequency. Also the SMM models mutations being hidden by increase and decrease mutations which IAM does not model.

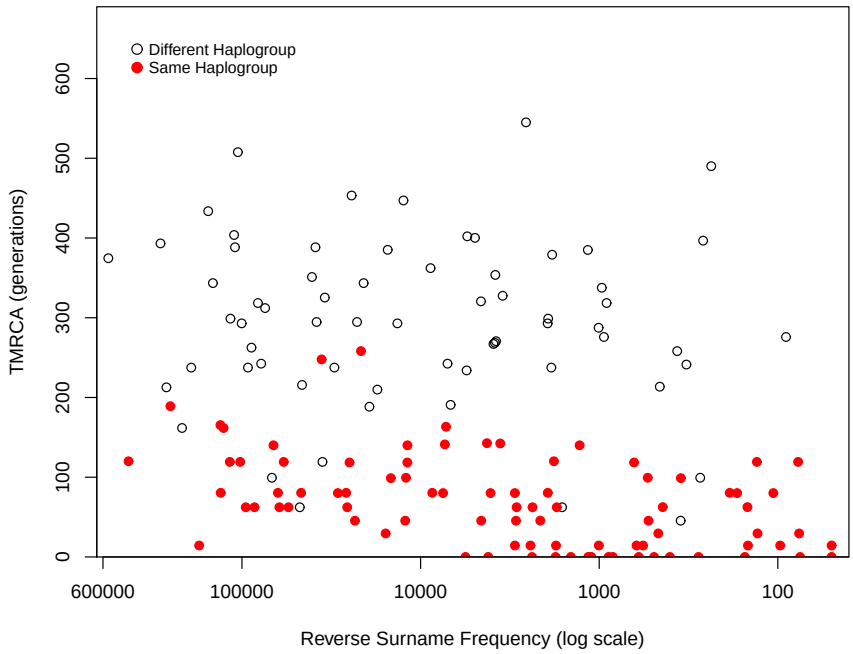

FIGURE 3.3: SMM haplogroup status: TMRCA estimates vs. log surname frequency

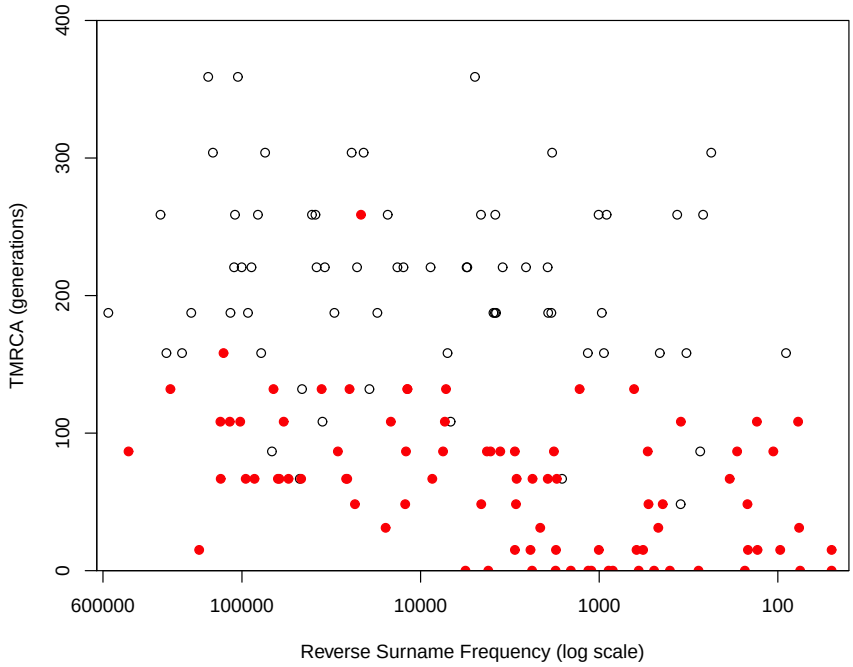

FIGURE 3.4: IAM haplogroup status: TMRCA estimates vs. log surname frequency

We next consider the haplogroup data for each model separately. In figure 3.3, the estimated TMRCA using the SMM versus the log surname frequency, the same haplogroup points lie lower than the different haplogroup ones (● versus ○ respectively). The same haplogroup results show a much more obvious negative relationship between the variables. This does not appear to be the case for the

different haplogroup data; the estimates appear to have similar spread across the range of log surname frequency and show little sign of decrease. A similar picture emerges for the IAM (fig. 3.4).

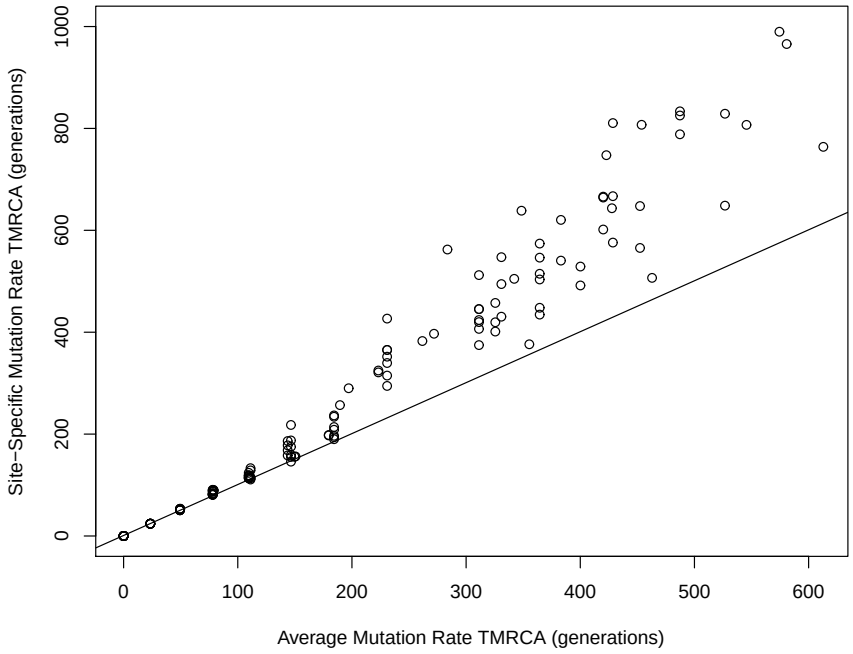

FIGURE 3.5: TMRCA estimates: site-specific rates vs. average rate with the line of equality superimposed

Now we show the results when examining data from only nine STRs from King et al. (2006) using site-specific mutation rates compared to their average mutation rate in the SMM. Comparing the estimates of TMRCA directly in figure 3.5, we see that there appears to be little difference between the estimates for estimated TMRCA below 100 generations with the points lying close to the line of equality. However for greater values of estimated TMRCA the points all lie above the line of equality showing that the site-specific rates produce higher estimates than the average rate and the spread of the difference appears to increase as time increases.

When plotting the results against the log surname frequency in figure 3.6, it is also evident that the average rate produces lower values of $\hat{t}$ than the site-specific rate, (● versus ○ respectively). Although there appears to be little difference in estimates under 200 years, there is a considerable difference in some higher estimates. In addition there still appears to be a negative linear relationship between estimated TMRCA and the reverse log surname frequency.

Decomposing the results according to the haplogroup status helps to accentuate the linear relationship between $\hat{t}$ and the log surname frequency. Figure 3.7 shows

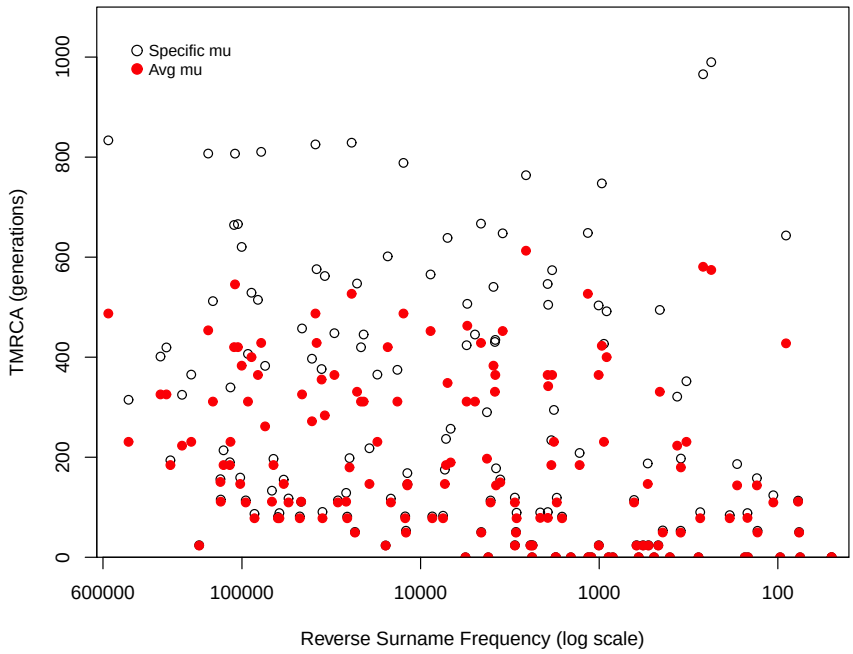

FIGURE 3.6: Site-specific and average mutation rates: TMRCA estimates vs. log surname frequency

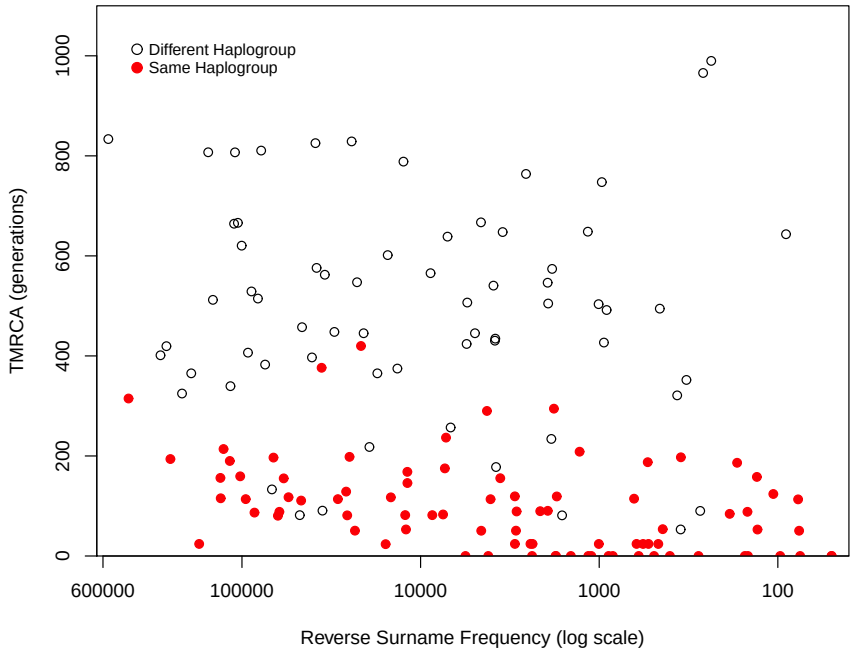

FIGURE 3.7: Site-specific mutation rates haplogroup status: TMRCA estimates vs. log surname frequency

the site-specific mutation rate results. The same haplogroup estimates (●) lie much lower than the different haplogroup estimates (○) and form a negative linear relationship between the variables which is not so clear for the different haplogroup estimates. A similar picture emerges when employing the average mutation rate in the SMM (data not shown).

3.1.3 Preliminary Conclusions

In summary, we find that the IAM produces lower estimates of TMRCA than the SMM. In addition, when subdividing the data according to haplogroup status, both models' estimates show a negative linear relationship with the reverse log surname frequency for the same haplogroup data, which does not appear as striking for the different haplogroup data.

It also appears that using an average mutation rate produces a lower estimate of TMRCA than using site-specific mutation rates particularly for times greater than 200 generations due to the effect of "hidden mutations".

3.2 Development

From section 3.1, it is evident that the model used to analyse Y-STR data from pairs of males who shared the same surname affected the estimates of TMRCA, t. This was also the case when site-specific and average mutation rates were employed. However in both comparisons we did not know which model or mutation rates was better at estimating t. This required data where this parameter was known. In the absence of real data of this kind, we examine simulated data where t could be specified. This has the advantage that we can produce statistics such as the mean squared error and the bias to compare the various estimators over replicate data sets.

The aim of this analysis is to compare the use of site-specific mutation rates with an average mutation rate both using the stepwise mutation model (SMM) and initially the infinite alleles model (IAM). The validity of comparing the IAM to the SMM was in doubt when using simulated data as the nature of the simulation would presumably favour the model from which the data were simulated, in this

case the SMM. Consequently we developed the infinite sites model (ISM) described below to compare to the SMM under varying conditions such as the number of loci (n) typed and the TMRCA, t, for a pair of Y-chromosomes.

3.3 Materials and Methods

The following notation will be used in this section:

- μ is the per locus per generation mutation rate. In general it will be indexed by i referring to the i^{th} locus;
- n_+ is the total number of increase mutations. This may be indexed as $n_{+,i}$ where i refers to the locus;
- n_- is the total number of decrease mutations. This may be indexed as $n_{-,i}$ where i refers to the locus;
- x_1, x_2 are the number of repeats of an STR for male 1 and 2, respectively. These may be indexed $x_{j,i}$ where $j = 1, 2$ refers to which male and i refers to the locus.
- m is the number of meioses used to calibrate the mutation rate. This may also be indexed by i to indicate the locus.

3.3.1 Stepwise Mutational Model

This model has already been outlined in section 3.1. The likelihood can be written as:

$$L(x_{1,1}, \ldots, x_{1,n}, x_{2,1}, \ldots, x_{2,n}|t) \propto \prod_{i=1}^{n} e^{-2t\mu_i} I_{|x_{1,i}-x_{2,i}|}(2t\mu_i). \tag{3.11}$$

The MLE of t cannot be analytically derived given the use of the Bessel function. As such we use the function `optim()` in R to find $\hat{t}$ that maximises L, or rather its logarithm. This function operates by searching for the maximum (or minimum) of a function given lower and upper values of the parameter to be estimated.

3.3.2 Infinite Sites Model

We now outline the model which we call the infinite sites model (ISM) in which we assume that no back mutations occur (fig. 2.3). In this case $|x_1 - x_2|$ is the

number of mutations at the locus of interest and is distributed as $Po(2\mu t)$. Hence the full likelihood is:

$$L(x_{1,1},\ldots,x_{1,n},x_{2,1},\ldots,x_{2,n}|t) = \prod_{i=1}^{n} e^{-2t\mu}\frac{(2t\mu)^{|x_{1,i}-x_{2,i}|}}{|x_{1,i}-x_{2,i}|!}, \tag{3.12}$$

which is maximised when:

$$\hat{t} = \frac{\sum_{i=1}^{n}|x_{1,i}-x_{2,i}|}{2\sum_{i=1}^{n}\mu_i}. \tag{3.13}$$

Given the form of the MLE of t for the ISM, the average mutation rate will produce the same result as the site-specific mutation rates.

3.3.3 Data Simulation

In the previous chapter we outlined various ways in which estimates of mutation rates can be obtained by examining DNA. Typically this is achieved by examining DNA from many father-son pairs or alternatively from DNA from several males whose ancestry is known, i.e. deep-rooting pedigrees. In both cases the number of meioses, i.e. the total number of times the Y-chromosome has been transmitted from any father to his son, m, is counted along with the number of times the number of repeats at each microsatellite differs, r. Thus in general at the i^{th} microsatellite the mutation rate can be estimated as $\hat{\mu}_i = r_i/m_i$.

Now we could simply use these point estimates. However we can suppose that these rates have an underlying distribution, which we would like to model. In this case a gamma distribution, with shape α and scale β, is probably flexible enough. The use of this distribution to model mutation rates has been well established (Nei et al., 1976). In particular, it has been used to model the variation of autosomal STRs (Xu et al., 2005). As such we also use the gamma distribution to model the distribution of Y-STR mutations rates. The shape and scale parameters can be estimated e.g. by the method of moments, equating the mean and variance of the observed estimated mutation rates to the mean and variance of the gamma distribution, and solving for $\hat{\alpha}$ and $\hat{\beta}$.

In the initial mutation rate review using the estimates at nine microsatellites in section 2.3.1, we observed a mean mutation rate of 0.002390 per generation with variance 0.000003368, for which $\hat{\alpha} = 1.697$ and $\hat{\beta} = 0.001409$ per generation.

Alternatively maximum likelihood estimates of the parameters in the gamma distribution may be produced e.g. by using the `optimize()` function in R. These are $\hat{\alpha} = 1.710$ and $\hat{\beta} = 0.001398$ per generation. The gamma distribution of the average of the two estimates is shown in figure 3.8 (black line). Using this gamma distribution with $\hat{\alpha} = 1.703$ and $\hat{\beta} = 0.001404$ per generation we can sample random mutation rates. Next, we apply a combined ascertainment and calibration step consisting of m simulated meioses. Only those microsatellites that are found to be variable, i.e. for which $\hat{\mu} \neq 0$ will be retained thus providing an empirical estimate for the mutation rate which has been both ascertained and calibrated. We end by simulating microsatellite data for a pair of Y-chromosomes given a specified number of loci n and t.

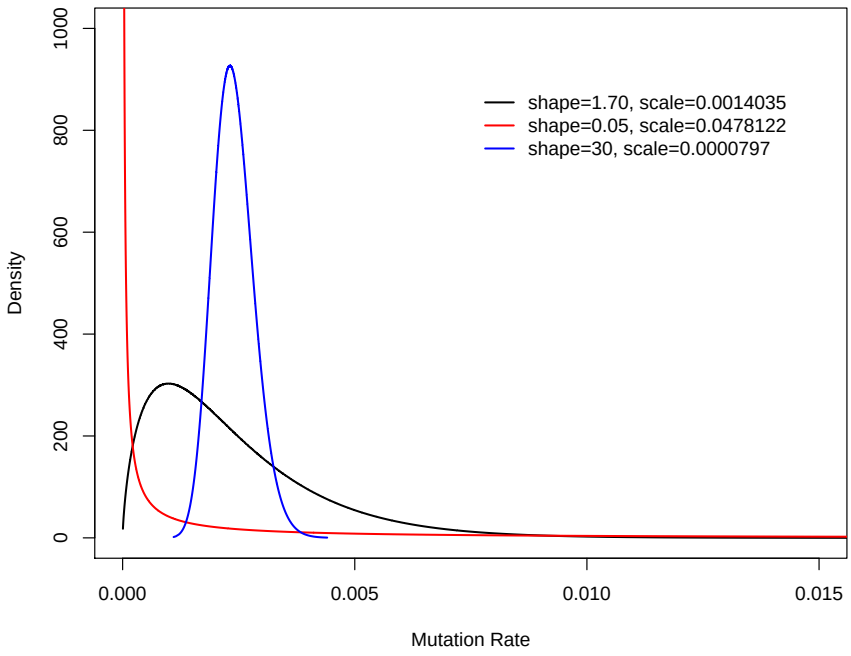

FIGURE 3.8: Three densities used to simulate mutation rates

The steps in the data simulation process can be summarised as follows assuming we have n loci.

1. Generate mutation rates: $\mu_i \sim Ga(\hat{\alpha}, \hat{\beta})$.
2. Simulate mutations in m_i meioses: $r_i \sim Bi(m_i, \mu_i)$.

3. For $r_i \neq 0$, compute the empirical mutation rate: $\hat{\mu}_i = r_i/m_i$.
4. Generate the total number of mutations between two males: $n_{+,i} + n_{-,i} \sim Po(2t\mu_i)$.
5. Generate the total number of increase mutations: $n_{+,i} \sim Bi(n_{+,i}+n_{-,i}, 0.5)$.
6. Compute the total number of decrease mutations: $n_{-,i}$.
7. Standardise male 1 as having zero STR repeats: $x_{1,i} = 0$.
8. Compute the number of STR repeats for male 2 relative to male 1: $x_{2,i} = n_{+,i} - n_{-,i}$.
9. Compute the observed data, i.e. the absolute difference in the number of STR repeats between males 1 and 2: $|x_{1,i} - x_{2,i}| = |n_{-,i} - n_{+,i}|$.

In our analysis we will simulate data for $n = 50$ calibrated mutation rates using m_i=10,000. The observed data are then generated for t=10, 25, 50, 100, 200, 400, 800, 1000 generations using the calibrated mutation rates 10,000 times. We subsample from the $n = 50$ simulated loci in order to obtain data for n =5, 10, 20, 30, 40. Thereafter we compute the MLE of t under the SMM for each run using:

- the true mutation rates,
- the average of the true rates,
- the empirical rates, and
- the average of the empirical rates.

Each estimator's mean squared error (MSE) is also computed. Recall that the MSE of any estimator can be broken down as follows:

$$\text{MSE} = \text{variance} + \text{bias}^2.$$

To allow a meaningful comparison across different t we compute the fractional squared error (FSE), i.e. MSE/t^2, comparing it for the four different mutation rates used. Similarly the fractional variance (variance/t^2) and the fractional bias squared (bias^2/t^2) will be produced. Where possible when plotting the results, the scales will be kept the same unless they prevent a clear interpretation of the results. Furthermore, these statistics will be produced for estimates of t using the ISM, based on the true mutation rates and the empirical rates.

We also examine how the underlying distribution from which the mutation rates are drawn affects the estimation of TMRCA using the SMM and ISM by considering two rather extreme cases of the gamma distribution such that the mean

mutation rate is fixed at 0.00239 per generation. The first case will involve mutation rates generated from a very right-skewed distribution with $\hat{\alpha} = 0.05$ and $\hat{\beta} = 0.04781$. The second case will have $\hat{\alpha} = 30$ and $\hat{\beta} = 0.00008$ where only a narrow range of mutation rates can be drawn. These two distributions are shown in figure 3.8 (red and blue lines, respectively). The three distributions have the following variances:

- 0.0000034 ($\hat{\alpha} = 1.710$ and $\hat{\beta} = 0.001398$, black line)
- 0.0001143 ($\hat{\alpha} = 0.05$ and $\hat{\beta} = 0.04781$, red line)
- 0.0000002 ($\hat{\alpha} = 30$ and $\hat{\beta} = 0.00008$, blue line)

The analysis will be followed by a discussion of possible bias due to the calibration process and any other factors that affect the estimation of TMRCA.

3.4 Results

3.4.1 Mutation Rate Distribution 1

In this first set of results, we will examine the simulated data in which simulated mutation rates are drawn from a $Ga(1.703, 0.001404)$ which are calibrated with 10,000 meioses. The data considers all combinations of n=5, 10, 20, 30, 40, 50 with t=10, 25, 50, 100, 200, 400, 800, 1000 generations. For each combination of n and t, 10,000 independent simulated datasets will be analysed.

3.4.1.1 Stepwise Mutation Model Analysis 1

We begin by analysing the data using the SMM. In figure 3.9 we plot the FSE of the estimates of TMRCA, $\hat{t}$, against the number of loci, n, at each value of t.

Note that in figure 3.9 the scale is different at each t. It is clear that increasing n decreases the FSE of $\hat{t}$ for each given t. This is true for all four types of mutation rates used. Also as t increases the FSE decreases substantially. Moreover the true average and site-specific rates (◇ and ○) produce consistently lower FSE values than their empirical counterparts (◇ and ○).

However, surprisingly, for values of t less than 400 (figs. 3.9a-e), the average rates (◇ and ◇) perform better than the site-specific rates (○ and ○) particularly when

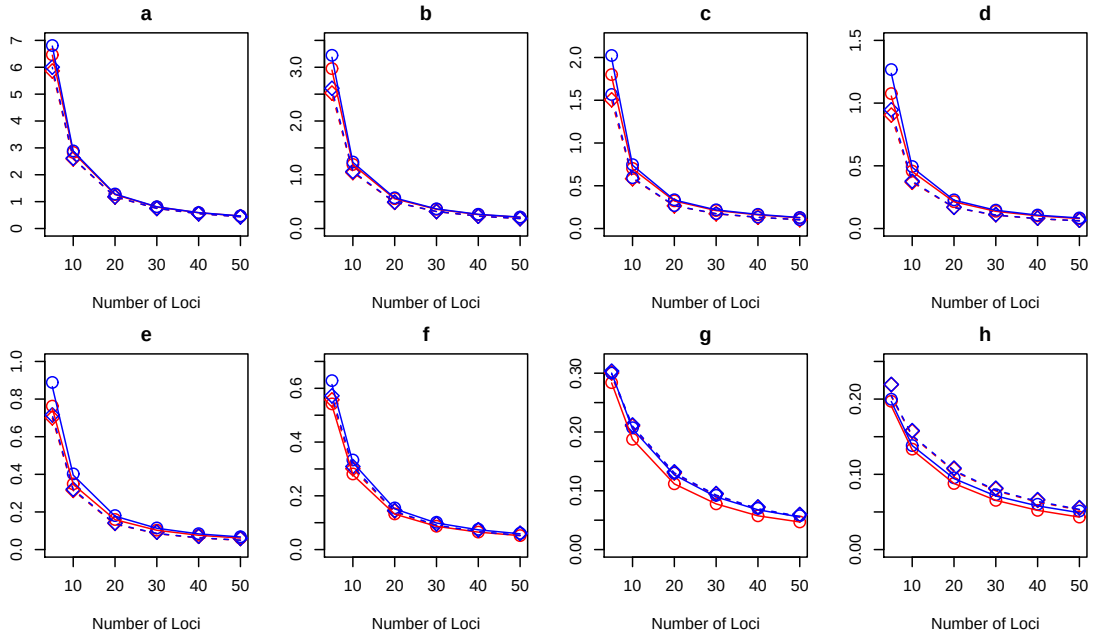

FIGURE 3.9: Fractional squared error and variance of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$ (○/solid red line - FSE/FV of true rates, ◇/dashed red line - FSE/FV of average of true rates, ○/solid blue line - FSE/FV of empirical rates, ◇/dashed blue line - FSE/FV of average of empirical rates)

$n \leq 20$. Nonetheless, the FSEs for all four rates converge for these five cases of t as the number of loci increase. Additionally, the true and empirical averages (◇ and ◇) converge quickly, whilst the true and empirical site-specific rates (○ and ○) converge more gradually with increasing n.

In figure 3.9f, for $t = 400$, the FSE for the true site-specific rates (○) is lowest although this is clearest when $n \leq 20$. Conversely, the FSE for the empirical site-specific rates (○) is highest throughout. The FSEs for the average of the true and empirical rates (◇ and ◇) are very similar even for low values of n. Despite the initial differences at low value of n, all four rates FSEs converge quickly.

In contrast, in the case $t = 800$ (fig. 3.9g) the FSE for the true site-specific (○) appears distinctly lower than the three other mutation rates (○, ◇ and ◇), which all appear to have very similar FSE values. The difference decreases as n increases but still remains evident at high n. This may imply a slower convergence between the rates in comparison to lower values of t. For low n, at $t = 1000$, the true and empirical site-specific rates have similar FSEs (○ and ○) although the former are slightly lower (fig. 3.9h). However, as n increases, their FSEs gradually diverge, with the true site-specific rates remaining lower. In contrast, although the average of the true and empirical rates perform poorly for small n, they converge with the empirical site-specific rates FSE with increasing n.

Now we examine the fractional variance (FV) in the same cases. The FV of the MLE estimates for the four types of mutation rates (fig. 3.9) appears to follow the same pattern as the FSE. Again note that the FV scales are different across t. The FV decreases as the number of loci increases with the four rates converging by $n = 50$ at each t except $t = 800$ and 1000 (fig. 3.9gh) where the convergence is slower.

The fractional bias squared (FBSQ) $\hat{t}$ against the number of loci (n) at each value of t is shown in figure 3.10. In general, the FBSQ is largest when sampling few loci and it decreases with increasing n throughout for the true and empirical site-specific rates (○ and ○). However, for the remaining two rates (◇ and ◇) this is true for $t = 10$, 25 and 1000 (figs. 3.10abh). At other times, although the FBSQ drops initially, it gradually increases as n increases further for the true and empirical averages (◇ and ◇). For these cases, although they have lower FBSQ for low n, they have much higher FBSQ values than their site-specific counterparts at high n. This is also the case for the empirical site-specific rates at $t = 800$ (fig. 3.10g). Here the FBSQ is negligible at $n = 5$ yet it progressively increases as n increases. A similar pattern exists for $t = 1000$ although there is decrease from $n = 5 - 10$ in FBSQ (fig. 3.10h). Also for the true and empirical averages FBSQ decreases as n increases.

In general, the true site-specific rates produce lower FBSQ when n is high. To understand further the unusual behaviour for the average rates, we examine the fractional bias (FB) of the corresponding results (fig. 3.11), where:

$$\text{FB} = \frac{\overline{\hat{t}} - t}{t}.$$

At $t = 10$ generations we have a straightforward result: as n increases the positive FB for $\hat{t}$ reduces by the four types of mutation rates. The true and empirical averages (◇ and ◇) produce consistently lower positive FB than their site-specific counterparts; they produce only a small positive FB when n is high.

At $t = 25$ (fig. 3.11b), although there is a decaying trend in FB in general, the averages (◇ and ◇) produce an increasing negative FB after $n = 30$. The true and empirical site-specific rates (○ and ○) continue with the trend of decreasing positive FB as n increases. For these two rate, this behaviour continues throughout $t = 10 - 400$ (figs. 3.11a-f). However, as t increases, the difference in their FB also increases with the true site-specific remaining less biased throughout.

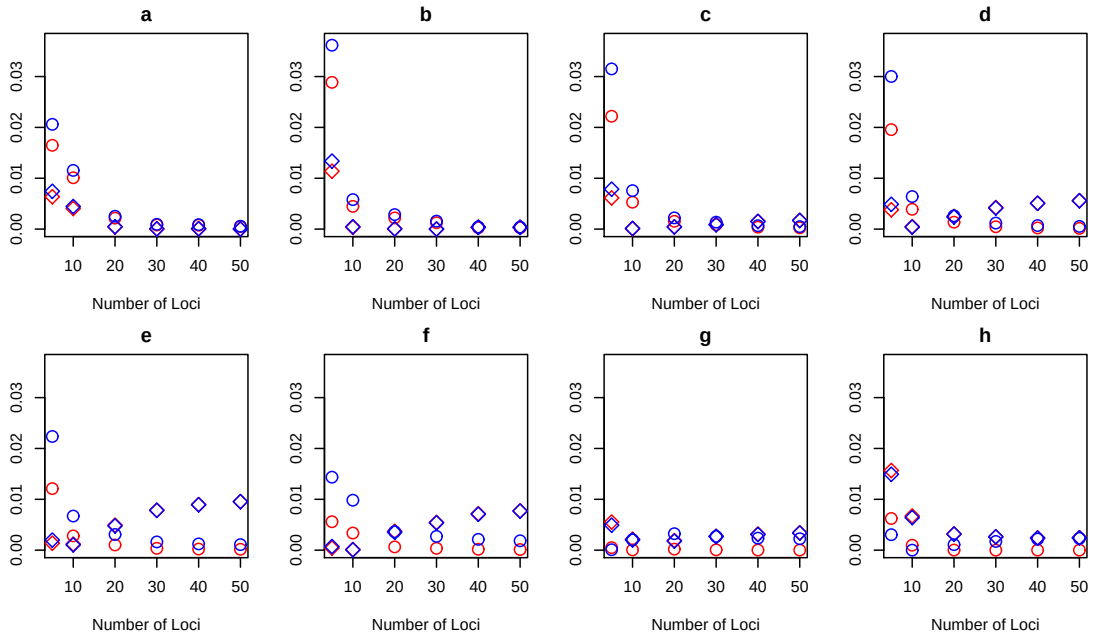

FIGURE 3.10: Fractional bias squared of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

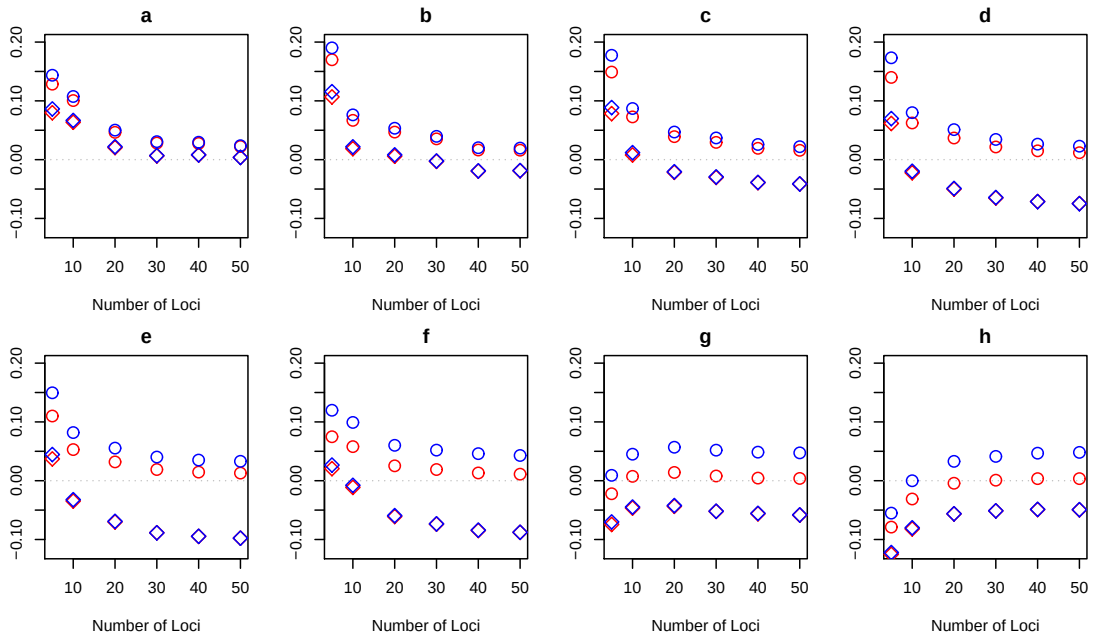

FIGURE 3.11: Fractional bias of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

For both the averages (◇ and ◇), increasing n continues to monotonically decrease FB when $t = 25 - 400$ (figs. 3.11b-f). For small n, FB is positive but becomes increasingly negative as n increases. There is little difference between the FB of these two average rates for every t (figs. 3.11a-h). At $t = 800$ and 1000, they both give a negative FB across n, which reduces as n increases.

In contrast, for the true site-specific rates (○) at $t = 800$, $\hat{t}$ begins with a negative FB, but this quickly (by $n = 10$) moves very close to zero (fig. 3.11g). This is also the case when $t = 1000$ (fig. 3.11h) but the estimates become virtually unbiased when $n \geq 20$. At t=800 (fig. 3.11g), the empirical site-specific rates (○) produce unbiased estimates for small n but a positive FB as n increases. This too is the case for t=1000; however the estimates have an initial negative FB, which then become unbiased around $n = 10$ and later remain positive (fig. 3.11h). Furthermore, there appears to be a gradual increase in the absolute difference between the fractional bias of these empirical site-specific rates and the true site-specific rates (○, ○).

3.4.1.2 Infinite Sites Model Analysis 1

Now we look at the ISM results for the data described in section 3.4.1. Firstly, we examine the FSE of the $\hat{t}$ against the n for the true and empirical site-specific rates where the scales are different across t (○ and ○, fig. 3.12). Note that these results are identical to those obtained by using the corresponding average rates due to the ISM estimator only depending on the average rate. We note that there is little difference in the FSE between the true or empirical rates across t thus the points are superimposed in most cases.

In figures 3.12a-f, there is a clear trend of decreasing FSE as n increases although the amount of decrease reduces relative to the initial FSE as t increases. At $t = 800$ and 1000, there is little change in FSE as n increases (fig. 3.12gh). However, as t increases, FSE decreases until $t = 200$, after which it begins to increase.

Turning to FV (fig. 3.12) again the lines are superimposed for both the true and empirical rates. At low times ($t = 10$, 25, 50), there is little difference between FV and FSE. However, the FV component of the FSE reduces as t increases further, so much so that at t=1000, less than 7% of the FSE is contributed by the FV.

The form of the FBSQ is shown in figure 3.13. It shows an increase as t increases. In addition, we find the FBSQ of both the empirical and true rates are virtually

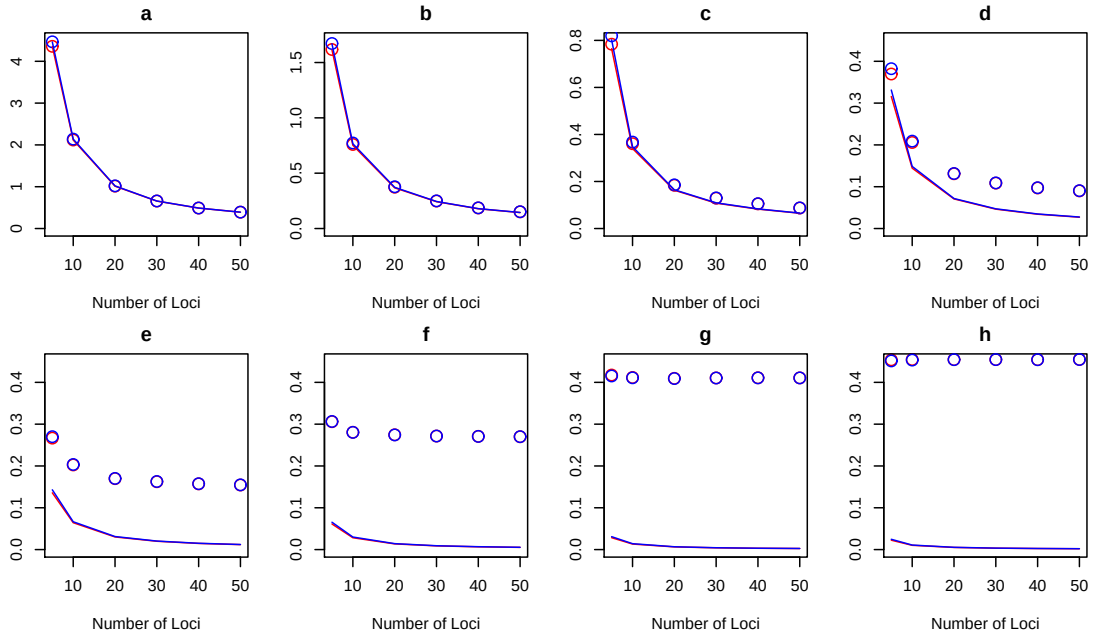

FIGURE 3.12: Fractional squared error and variance of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$ (○/solid red line - FSE/FV of true rates, ○/solid blue line - FSE/FV of empirical rates)

identical and that there is little change in FBSQ as n increases. In figure 3.14 the FB against n is plotted for each time. For all times there is a negative FB indicating that the ISM underestimates $\hat{t}$, and the magnitude of FB increases, as t increases. Furthermore, at each time, FB is fairly constant as n increases.

3.4.2 Mutation Rate Distribution 2

This second set of results was obtained using data simulated using mutation rates drawn from a $Ga(0.05, 0.04781)$, which are estimated with 10,000 meioses. This has significant mass at a wider range of mutation rates than those sampled in section 3.4.1. The data were, otherwise, simulated as before.

3.4.2.1 Stepwise Mutation Model Analysis 2

As before we commence by analysing the data using the SMM. Figure 3.15 shows FSE against n at each value of t. Note that the scales of FSE across t are not identical. At each t, there is little difference in the FSE produced using the true and empirical site-specific rates (○ and ○) with the latter overprinting the former in the figure and likewise for the true and empirical averages (◇ and ◇). Furthermore, the FSE decreases both as n increases at each time and also as t

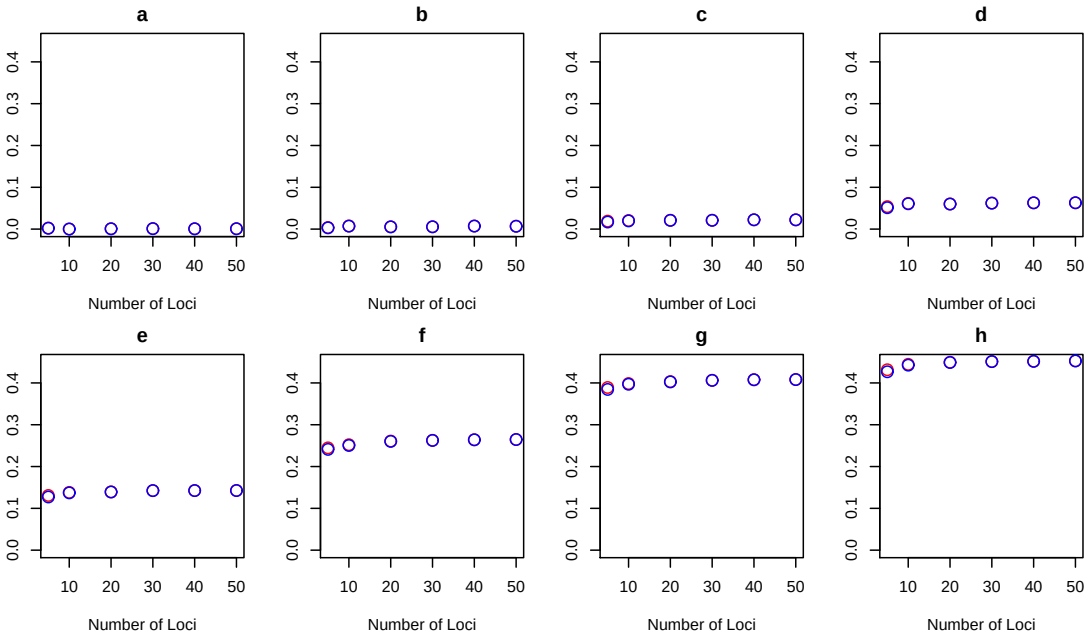

FIGURE 3.13: Fractional bias squared of $\hat{t}$ Estimates vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ○ - empirical rates)

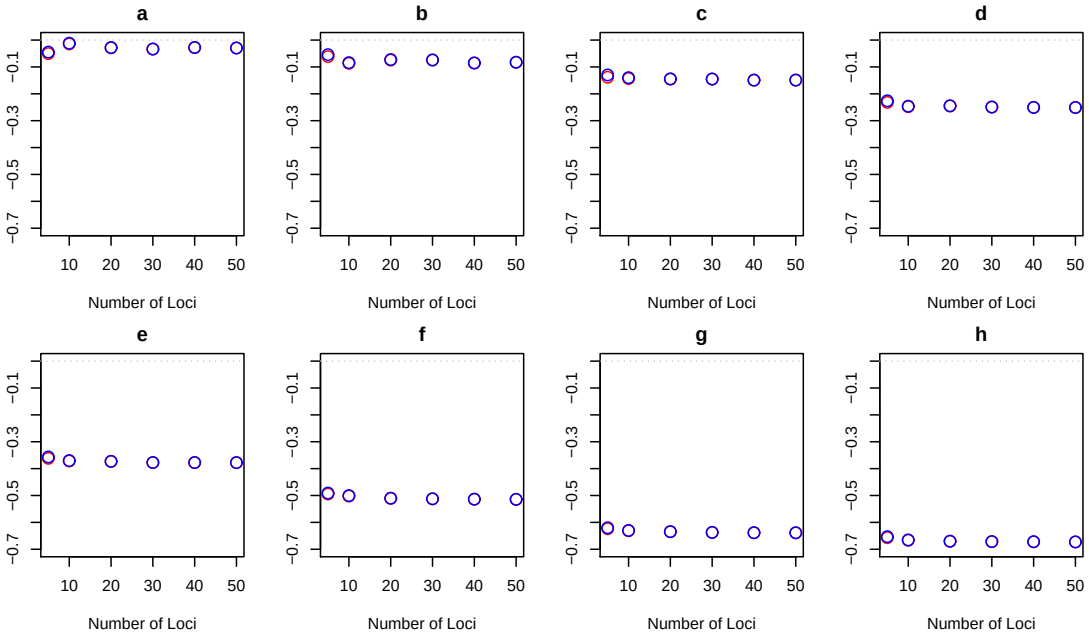

FIGURE 3.14: Fractional bias of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ○ - empirical rates)

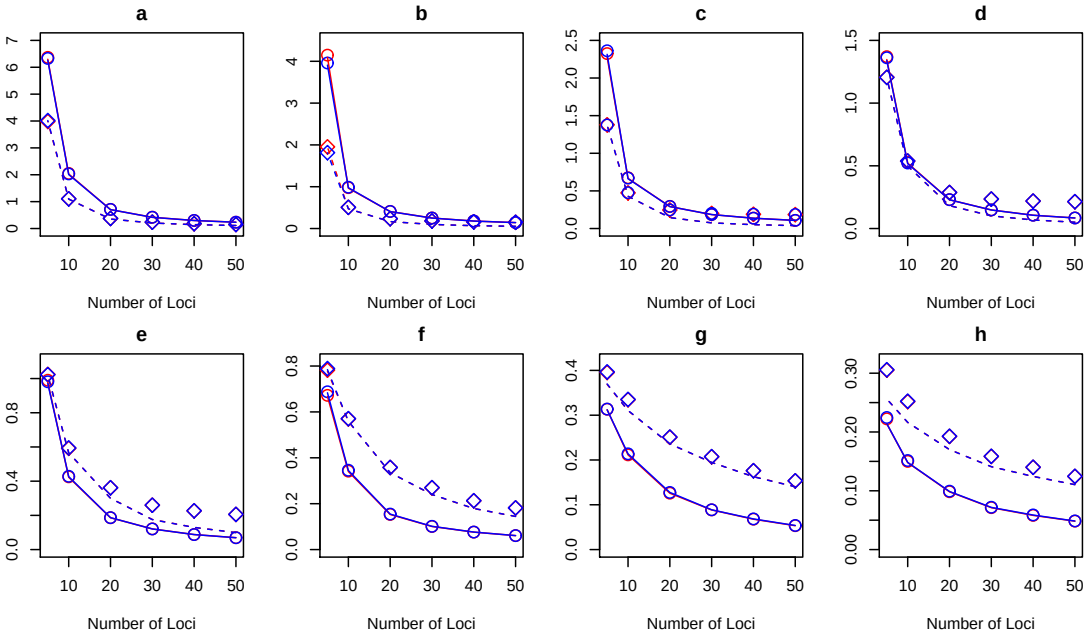

FIGURE 3.15: Fractional squared error and variance of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$ (○/solid red line - FSE/FV of true rates, ◇/dashed red line - FSE/FV of average of true rates, ○/solid blue line - FSE/FV of empirical rates, ◇/dashed blue line - FSE/FV of average of empirical rates)

increases. At $t = 10$ and 25, the average rates tend to lower FSE that their site-specific counterparts, although the absolute difference between them decreases as n increases (fig. 3.15ab). Indeed by $n = 50$, their FSEs appear to have converged. In figure 3.15c, the averages' FSEs are still lower than those for the site-specific rates, although there appears to be quicker convergence (by $n = 30$) than at lower times. Thereafter as n increases, there appears to be a slight divergence with the site-specific rates giving a slightly lower FSE than the averages.

At the next time, $t = 100$, the averages' FSE is lower only at n=5 (fig. 3.15d). By $n = 10$ the FSE is indistinguishable for all four rates. For higher n, the site-specific rates have lower FSE than the averages. Furthermore, the difference appears to increase as n increases. In figure 3.15e, the absolute difference between the site-specific rates and average rates appears constant across n with the exception of n=5. The site-specific rates have lower FSE than the averages. This is also the case for $t = 400$, 800 and 1000. Again there is almost constant absolute difference between the averages and site-specific rates. Turning to the FV component, the general trend is as before: FV has an inverse relationship to n. In addition, the FV decreases as t increases. There is little difference in FV between the true and empirical site-specific (solid red and blue lines) and similarly between the averages (dashed red and blue lines). There is less FV in the estimates based on the average rates for $t = 10 - 100$ (figs. 3.15a-d) than for the site-specific rates. However, for

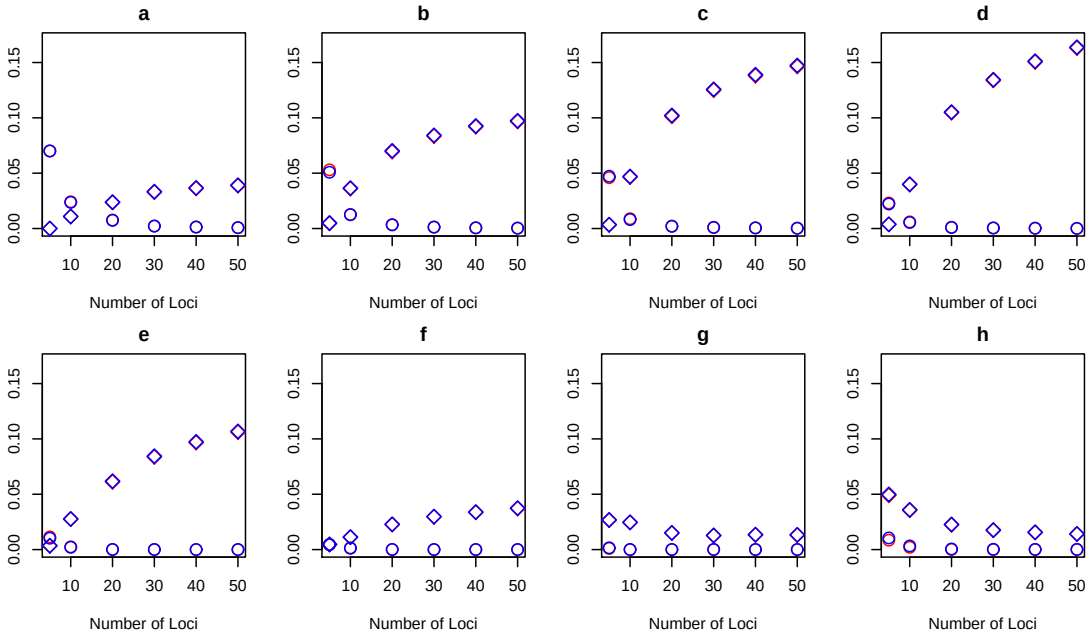

FIGURE 3.16: Fractional bias squared of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

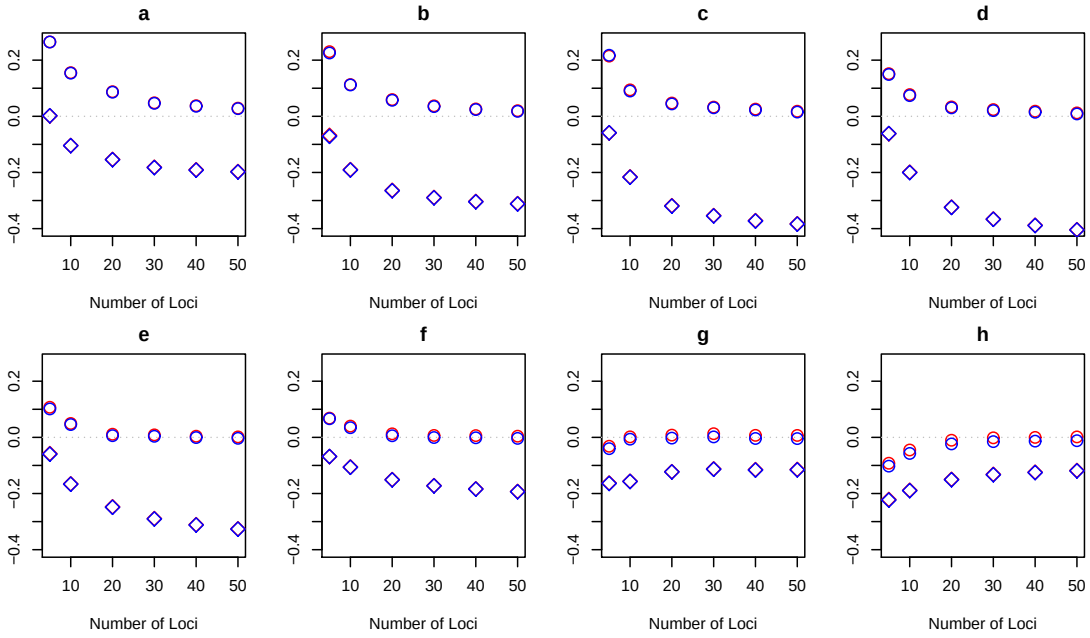

FIGURE 3.17: Fractional bias of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

the remaining times, the reverse is true (figs. 3.15e-h). Furthermore there is much less contribution of FV to the FSE for the averages (◇ and ◇) in particular for $t = 50 - 400$ when n is high (figs. 3.15c-f), and a little less for high t (figs. 3.15gh). For the true rates (○ and ○), there is little difference FV and FSE.

In figure 3.16, we focus on FBSQ, for which, as before, there is almost no difference between the true and empirical site-specific rates or between the average rates and

thus the empirical superimpose their true counterparts. We do find however that whilst FBSQ for the site-specific rates decreases with increasing n, the reverse holds true for the FBSQ for the average rates when $t = 10 - 400$ (figs. 3.16a-f). In addition, the rate at which the FBSQ increases with n also increases as t increases from 10 to 100, thereafter decreasing in these cases (figs. 3.16a-d). On the other hand, for the site-specific rates whether true or empirical, the rate at which the FBSQ decreases also increases with increasing time with the exception of t=1000, where the FBSQ appears slightly higher. For t=800 and 1000, the FBSQ produced using the averages decreases with increasing n. Furthermore, FBSQ is less in general for the site-specific rates than for the average rates.

To complete this section we comment on the fractional bias of $\hat{t}$ versus the number of loci (figure 3.17). As there is little difference in FB between the true and empirical rates, we will again refer simply to the site-specific rates and the average rates. For the former (○ and ○), there is a positive FB which decreases, approaching zero as n increases at each $t = 10 - 400$ (figs. 3.17a-f). The initial FB ($n = 5$) also decreases, as time increases, as does the rate of convergence to zero. However, although the average rates (◇ and ◇) follow a similar monotonically decreasing trend, the FB becomes increasingly negative both as n increases and also as time increases from 10-400. For $t = 800$ and 1000, the site-specific rates tend to a negative FB for small n which increases quickly to zero, i.e. producing little bias (figs. 3.17gh). There is slightly less FB for the true site-specific rates than the empirical rates (○ and ○). For the average rates there is also a reduction in FB as n increases. However, there is still a substantial negative bias compared to the FB arising from the site-specific rates.

3.4.2.2 Infinite Sites Model Analysis 2

We now examine the simulated data described in section 3.4.2 using the ISM. Firstly, we view the FSE of $\hat{t}$ against n at each t using the true and empirical site-specific mutation rates. As before the scales of FSE are different for each time. In general, there is no difference in FSE between the true and empirical rates (○ and ○). At $t = 10 - 100$ the FSE decreases with increasing n (figs. 3.18a-d). Conversely, the FSE increases with increasing n for $t = 200 - 1000$ (figs. 3.18e-h). Furthermore, in those cases the overall FSE progressively increases with increasing t.

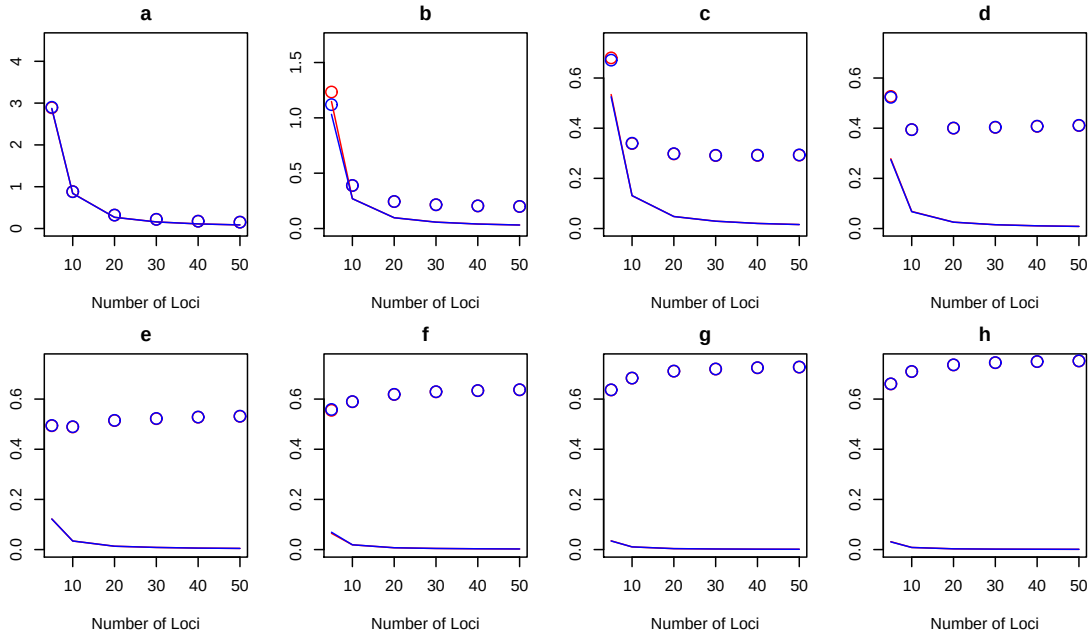

FIGURE 3.18: Fractional squared error and variance of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$ (○/solid red line - FSE/FV of true rates, ○/solid blue line - FSE/FV of empirical rates)

The FV component is overlaid on figure 3.18 also. There is virtually no difference in FV using either the true or empirical rates, the latter (blue line) superimposes the former. As n increases, FV decreases at each t. However the contribution that FV makes to FSE decreases with increasing time. At t=10, almost all the FSE is from the FV, but by t=1000 there is very little FV in the FSE. Figure 3.19 shows FBSQ versus the n at each t. Again the true and empirical rates have indistinguishable FBSQ (○ and ○), which increases slowly as n increases. Also it increases gradually from approximately 0.05 to over 0.7 as t increases.

The corresponding plot of fractional bias (FB) is shown in figure 3.20. Here we find that both types of mutation rates produce negative FB, i.e. they underestimate of t. The amount by which the rates underestimate increases as n increases at each time and also as t increases.

3.4.3 Mutation Rate Distribution 3

In this final set of results, we analyse data simulated using mutation rates drawn from a $Ga(30, 0.00008)$. As before the mutation rates have been drawn using 10,000 meioses. This gamma distribution allows a narrower range of mutation rates to be drawn compared to the distribution mentioned in section 3.4.1.

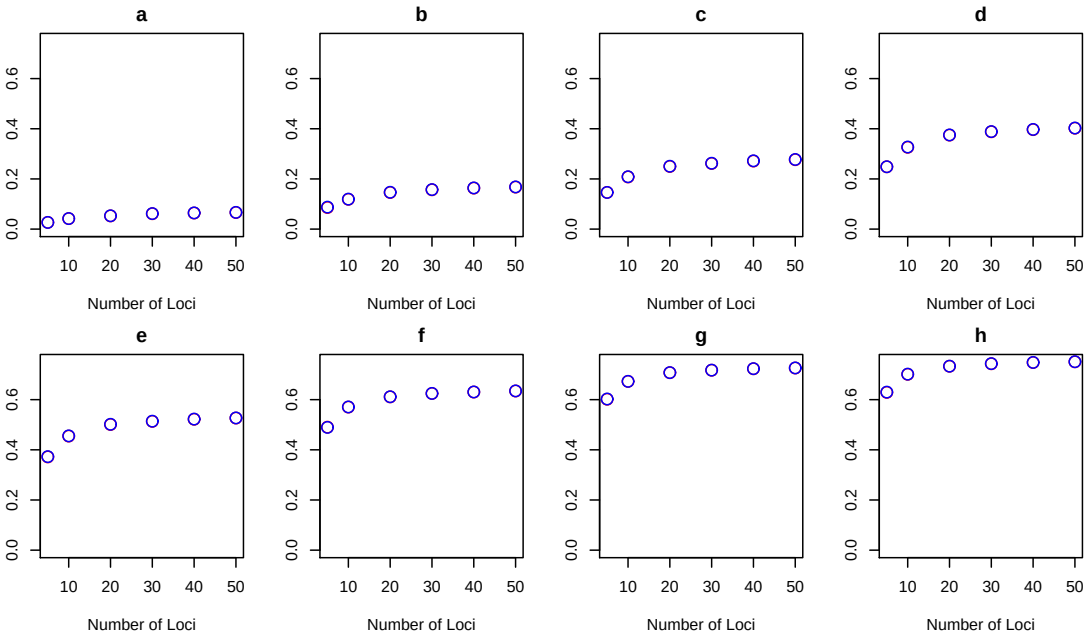

FIGURE 3.19: Fractional bias squared of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ○ - empirical rates)

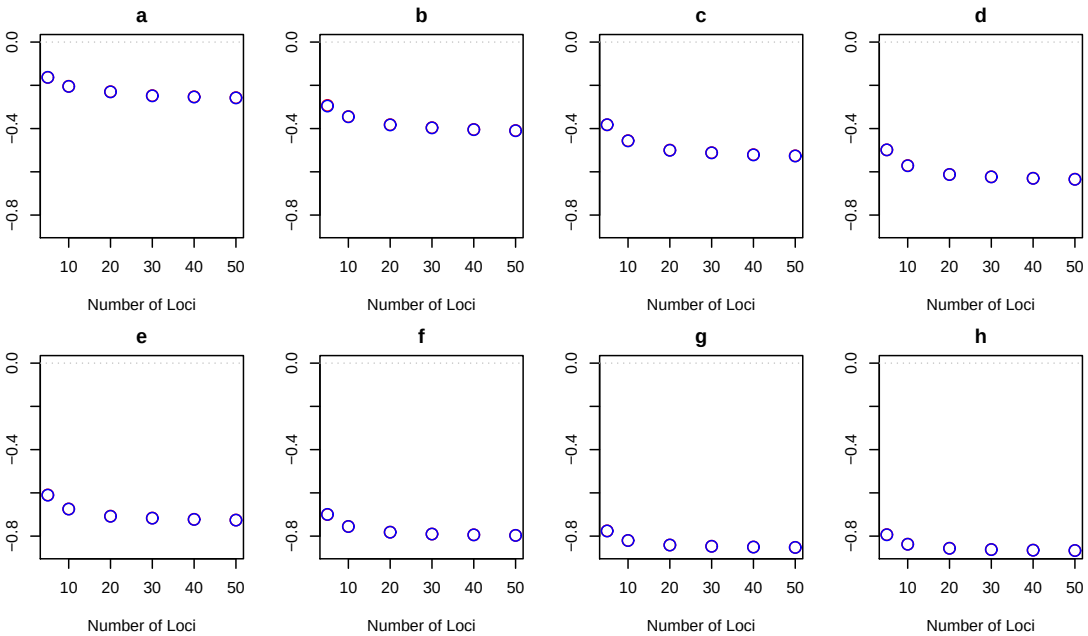

FIGURE 3.20: Fractional bias of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ○ - empirical rates)

3.4.3.1 Stepwise Mutation Model Analysis 3

In this section, we use the SMM to analyse the above data. Figure 3.21 is a plot of FSE of $\hat{t}$ against n for each t, using the four types of mutation rates described in section 3.4.1. Note that the scale of FSE reduces as times increases.

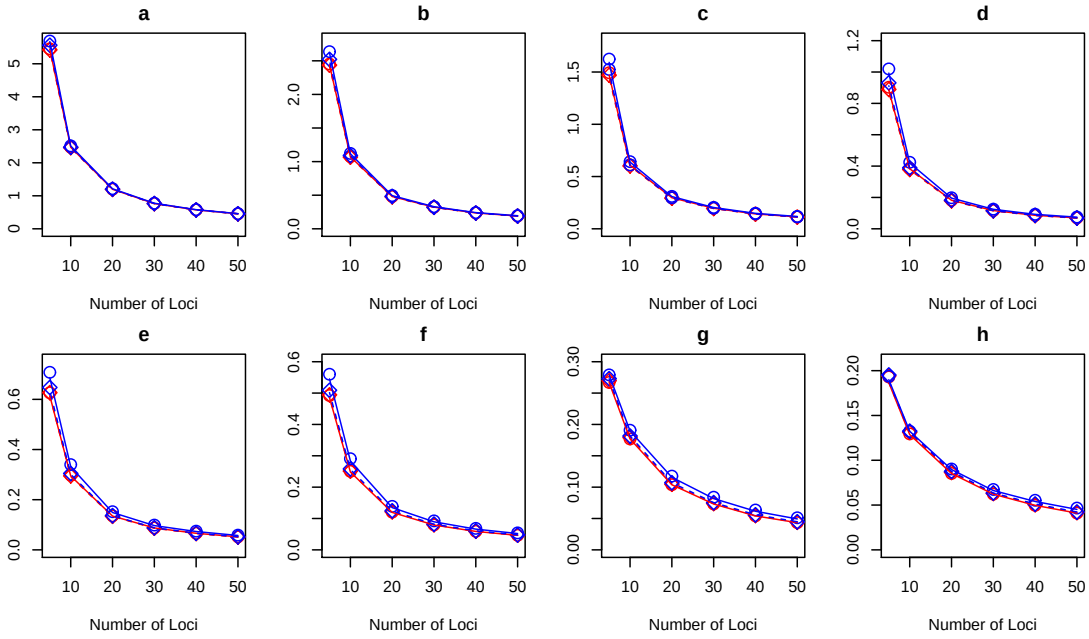

FIGURE 3.21: Fractional squared error and variance of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$ (○/solid red line - FSE/FV of true rates, ◇/dashed red line - FSE/FV of average of true rates, ○/solid blue line - FSE/FV of empirical rates, ◇/dashed blue line - FSE/FV of average of empirical rates)

Here we find a decreasing trend of FSE as n increases at each time. Also, as t increases, FSE decreases. Furthermore, there are only minor differences between the four rates, with the empirical site-specific rates' FSE (○) just slightly higher than the other three rates' FSEs (○, ◇ and ◇). In addition, at certain times, e.g. t=100, 200 and 400, these empirical site-specific rates lie distinctly higher at low values of n (figs. 3.21). Examining the FV component, it is clear that most of the FSE is contributed by the FV across all times, which similarly has a decreasing trend with increasing n and increasing t. The FBSQ is shown in figure 3.22. Here we find that the FBSQ component is very small and, although there is a pattern of increasing FBSQ as t increases from 10 to 50, it is still negligible compared to the FSE (fig. 3.21). From $t = 100 - 1000$, FBSQ decreases. We find that the true site-specific, true average and empirical averages rates (○, ◇ and ◇) have very similar FBSQ throughout. From $t = 10 - 400$, the absolute difference of the FBSQ between them and the empirical site-specific rates increases (○, figs. 3.22a-f). However the latter's FBSQ appears to increase with n at t=800 (fig. 3.22g).

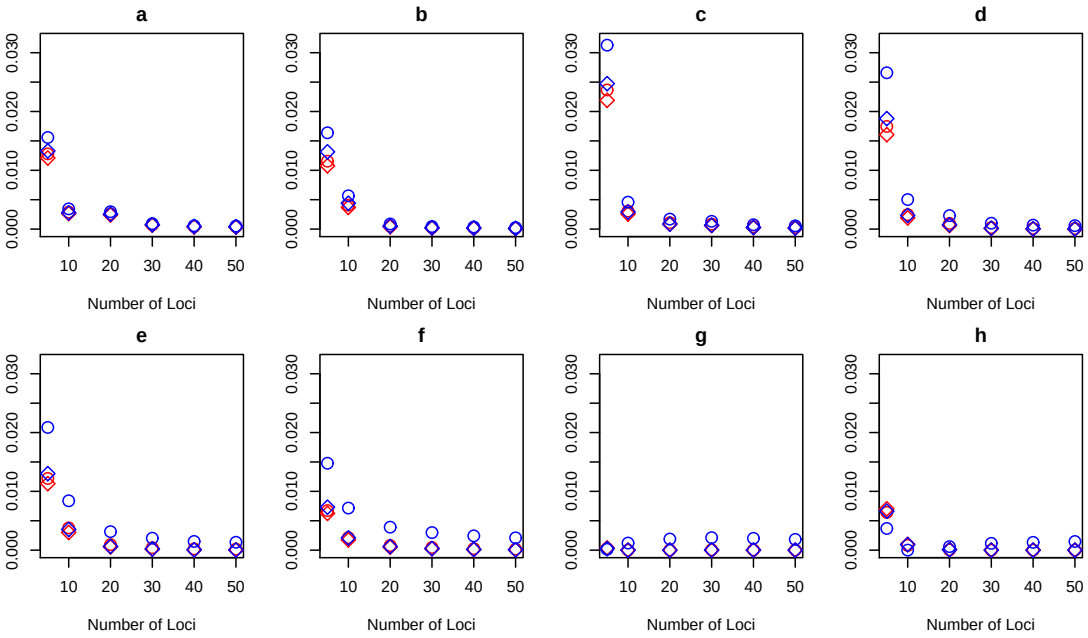

FIGURE 3.22: Fractional bias squared of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

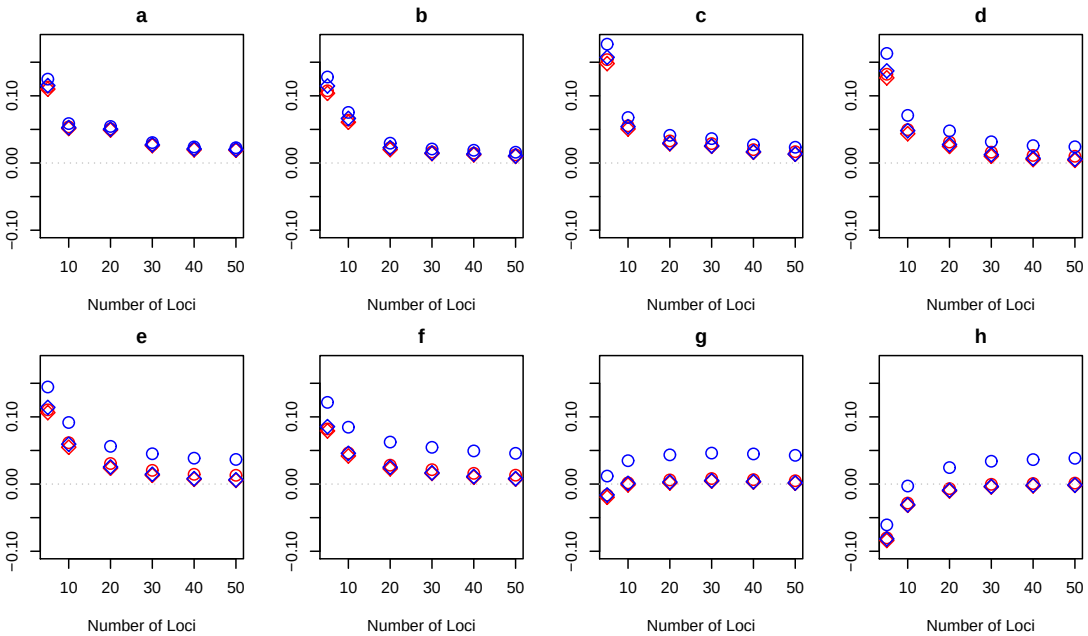

FIGURE 3.23: Fractional bias of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

At this time, the three remaining rates FBSQ appear constant as n increases. At t=1000, these rates show a decreasing trend with increasing n and although the empirical site-specific rates' FBSQ initially decreases, it shows a very slight increase as n increases further.

The final figure in this subsection is of FB of $\hat{t}$ against the n for each t (fig. 3.23). For $t = 10-400$, all four rates' FB is positive (figs. 3.23a-f). It decreases in value as n increases, approaching zero. However, as t increases, there is a greater difference between the empirical site-specific rates (○) FB and those for the other rates (○, ◇ and ◇). At $t = 800$, these other rates have an initial negative bias which quickly decreases in magnitude to near zero (fig. 3.23g). Although the empirical rates' FB increases, it is positive throughout (○). At $t = 1000$, the rates all exhibit a FB which increases with n (fig. 3.23h). But whilst the empirical rates begins with a negative FB which reaches zero and then remains positive, the FB for the three other rates increases but remains negative or near zero.

3.4.3.2 Infinite Sites Model Analysis 3

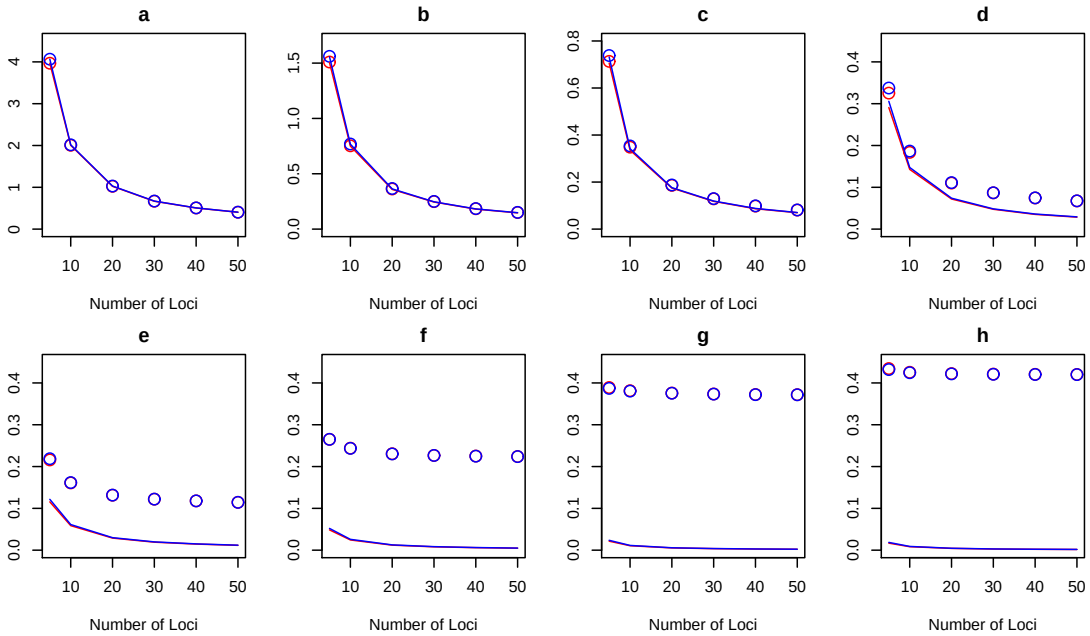

FIGURE 3.24: Fractional squared error and variance of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$ (○/solid red line - FSE/FV of true rates, ○/solid blue line - FSE/FV of empirical rates)

This section reanalyses the data outlined in section 3.4.3 using the ISM. Figure 3.24 shows the ISM estimates using the empirical and true rates against n. Firstly there is virtually no difference between the FSE for either type of rates (○ and ○). Secondly there is a strong decreasing trend in FSE as n increases for $t < 400$

(figs. 3.24f-h). The rate of decrease as n increases is substantial for low t but reduces as time increases. Also, with increasing time, although there is a decrease in FSE from $t = 10 - 200$ (figs. 3.24a-e), thereafter the overall FSE increases. For $t = 10 - 50$, FV is the major component to FSE, but as time increases further, the contribution reduces to a very small fraction.

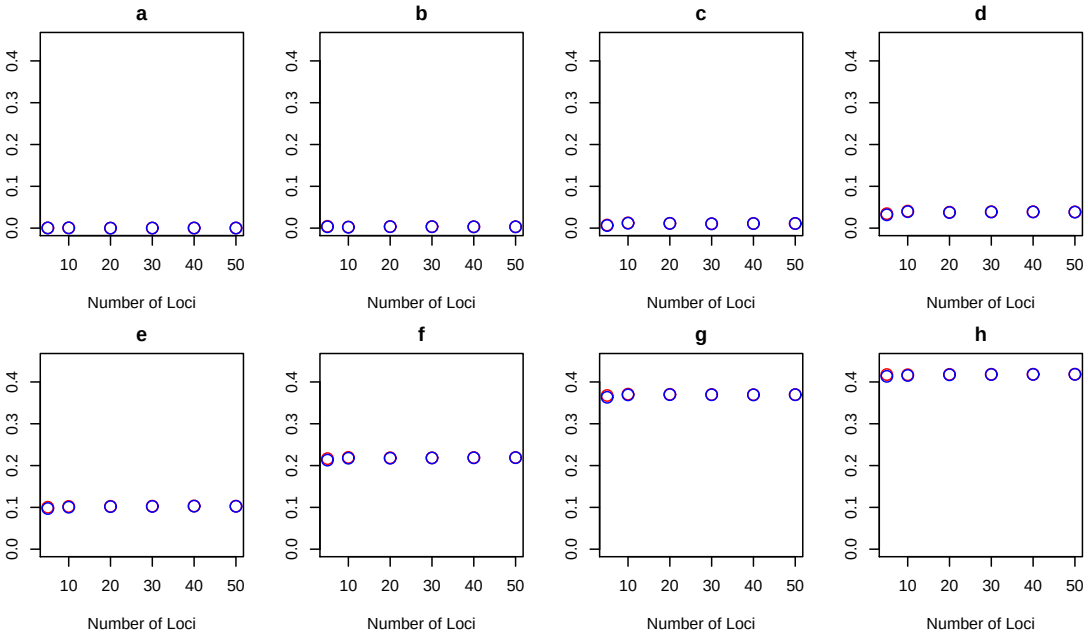

FIGURE 3.25: Fractional bias squared of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ○ - empirical rates)

The inverse relationship between FV and time is also evident between FBSQ and t in figure 3.25. Here, as t increases so does FBSQ, although there is little difference in FBSQ as n increases at each time. Conversely, we see that the FB of the ISM estimates decrease as t increases (fig. 3.26). Again, the FB is roughly constant as n increases but is consistently negative.

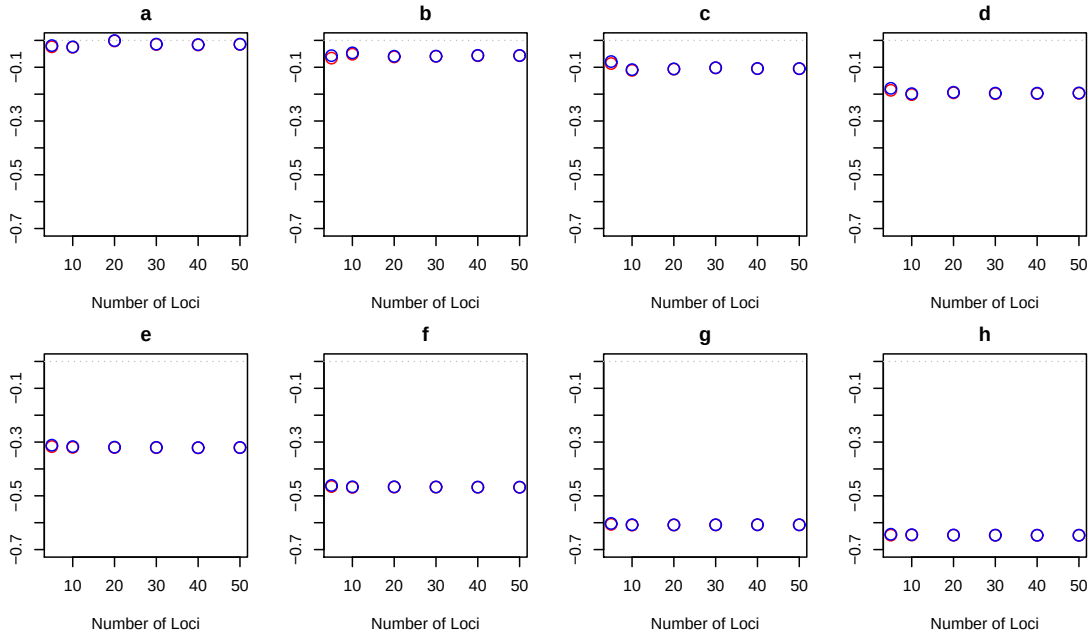

FIGURE 3.26: Fractional bias of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
(○ - true rates, ○ - empirical rates)

3.5 Discussion

In this discussion we will compare the results from the site-specific rates and the average mutation rates using the SMM. Next the results for the different mutation rate distributions will be compared. We will then discuss differences between the SMM and the ISM. We will finish by examining three selected times to compare the results using different mutation rate distributions and also the type of mutation rates used, i.e. $t = 10$, 200 and 1000 generations.

For mutation rate distribution 1, the right-skewed distribution, at very short times the estimates of t using average mutation rates (◇ and ◇) have lower FSE and FB compared to their site-specific counterparts (○ and ○, fig 3.27a). This may seem counterintuitive: site-specific rates should give more informative results. However using site-specific rates, some of which are high, the SMM overcorrects for recurrent mutation that essentially do not occur for small t, where the data is essentially compatible with the ISM.

For higher times, for example $t = 200$ generations, although the average rate produces a monotonically decreasing FB, that bias actually becomes negative. Thus, at high n, the magnitude of the FB can be high. This is consistent with the fact that for high t, the data are affected by repeat mutation. Since the correction

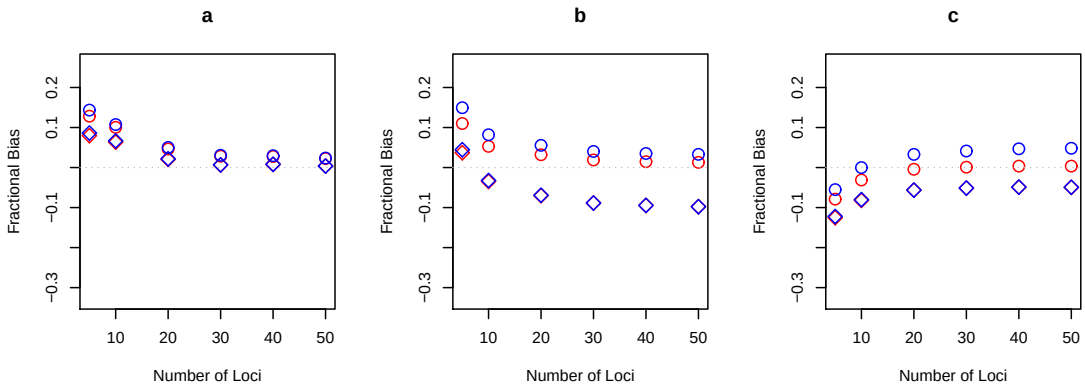

FIGURE 3.27: Mutation rate distribution 1: fractional bias of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 200$, c. $t = 1000$
(○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

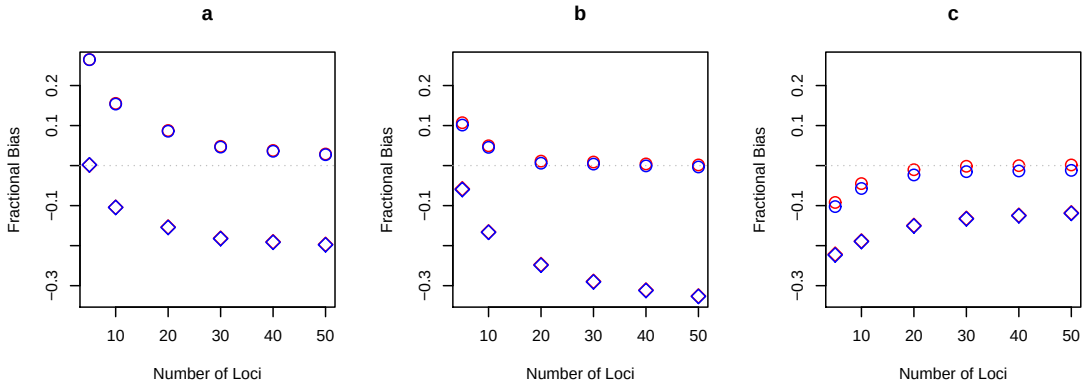

FIGURE 3.28: Mutation rate distribution 2: fractional bias of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 200$, c. $t = 1000$
(○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

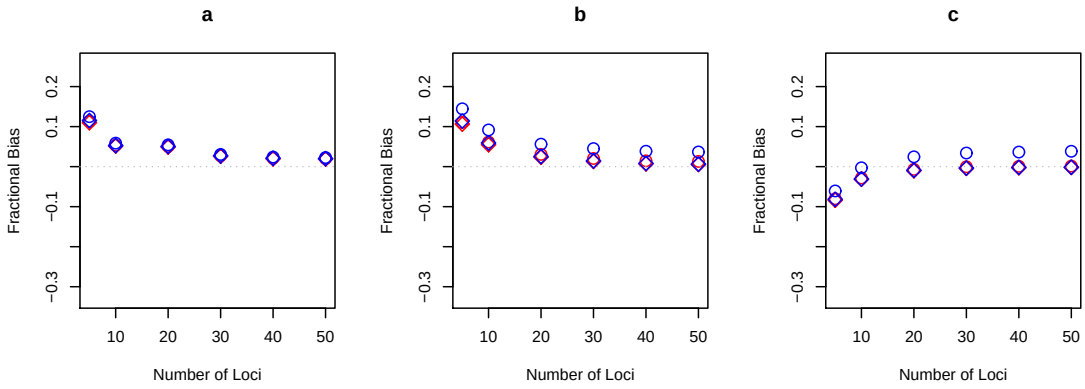

FIGURE 3.29: Mutation rate distribution 3: fractional bias of $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 200$, c. $t = 1000$
(○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

for this depends on each marker's rate, not the average, the average rate is not adequate, to allow for recurrent mutation.

When we consider mutation rate distribution 2 (fig. 3.28), a very right-skewed distribution, the magnitude of the difference in FB based on the site-specific and average rates is much greater than for mutation rate distribution 1 (fig. 3.27). This is consistent with there being more high rates generated in this case, for which repeat mutation is more of an issue, and which cannot be corrected solely based on the average rate.

Furthermore, considering mutation rate distribution 3, the narrow almost symmetric distribution, we find in figure 3.29 there appears to be little or no difference between FB when using the true site-specific rates or either of the average rates (the difference in these and the empirical site-specific rates will be discussed later). Given the low variance in marker rates, most rates are close to the mean rate. So it is unsurprising that use of the average rate in estimation is as effective as having all the individual rates.

In summary, as the variance of the mutation rates increases, the average becomes less and less able to correct for the impact of repeat mutation at markers with a high rate.

We focus next on the results from the SMM and the ISM, examining the FSE of estimates using the true mutation rates and then the empirical mutation rates. For mutation rate distribution 1, table 3.1 shows when the FSE for the ISM is lower than that for the SMM for all combinations of n and t.

TABLE 3.1: Mutation rate distribution 1: ISM FSE < SMM FSE using true and empirical site-specific mutation rates (T=true, F=false)

Parameters		t(generations)							
		10	25	50	100	200	400	800	1000
n	5	T	T	T	T	T	T	F	F
	10	T	T	T	T	T	T	F	F
	20	T	T	T	T	T/F[1]	F	F	F
	30	T	T	T	T	F	F	F	F
	40	T	T	T	T	F	F	F	F
	50	T	T	T	F	F	F	F	F

The ISM outperforms the SMM particularly at lower times (< 100 generations) for mutation rate distribution 1, when there is very little repeat mutation to correct

[1]True/empirical site-specific mutation rate

for. As the variance of the mutation rate distribution increases, the SMM tends to outperform the ISM as demonstrated by mutation rate distribution 2 (table 3.2). The converse holds when the variance of the mutation rate distribution is narrowed (table 3.3).

TABLE 3.2: Mutation rate distribution 2: ISM FSE < SMM FSE using True and Empirical Site-Specific Mutation Rates (T=true, F=false)

Parameters		t(generations)							
		10	25	50	100	200	400	800	1000
n	5	T	T	T	T	T	T	F	F
	10	T	T	T	T	F	F	F	F
	20	T	T	F	F	F	F	F	F
	30	T	T	F	F	F	F	F	F
	40	T	F	F	F	F	F	F	F
	50	T	F	F	F	F	F	F	F

TABLE 3.3: Mutation rate distribution 3: ISM FSE < SMM FSE using True and Empirical Site-Specific Mutation Rates (T=true, F=false)

Parameters		t(generations)							
		10	25	50	100	200	400	800	1000
n	5	T	T	T	T	T	T	F	F
	10	T	T	T	T	T	T	F	F
	20	T	T	T	T	T/F[2]	F	F	F
	30	T	T	T	T	F	F	F	F
	40	T	T	T	T	F	F	F	F
	50	T	T	T	F	F	F	F	F

However, for every case of n and t and for all mutational distributions, the FV is lower for the SMM estimates than for the ISM estimates (data not shown). Also, the perceived success of the ISM falls into doubt when FB is considered.

It is clear (fig. 3.30) that the ISM systematically underestimates t. However it does better at short times, when the differences in repeat number are almost the total number of mutation, i.e. no wiping out of mutations occurs, consistent with the ISM likelihood. However, once time increases, when the chance of mutations being wiped out is increased, the model continues to underestimate t and the FB does not reduce with additional information from more markers.

Next we consider the results of the SMM and ISM when using empirical mutation rates for the three mutation rate distributions. As before, tables 3.1, 3.2 and 3.3 demonstrate the dependence of the performance of the ISM FSE according to the mutation rate distribution. However the fractional variance is lower for the

[2]True/empirical site-specific mutation rate

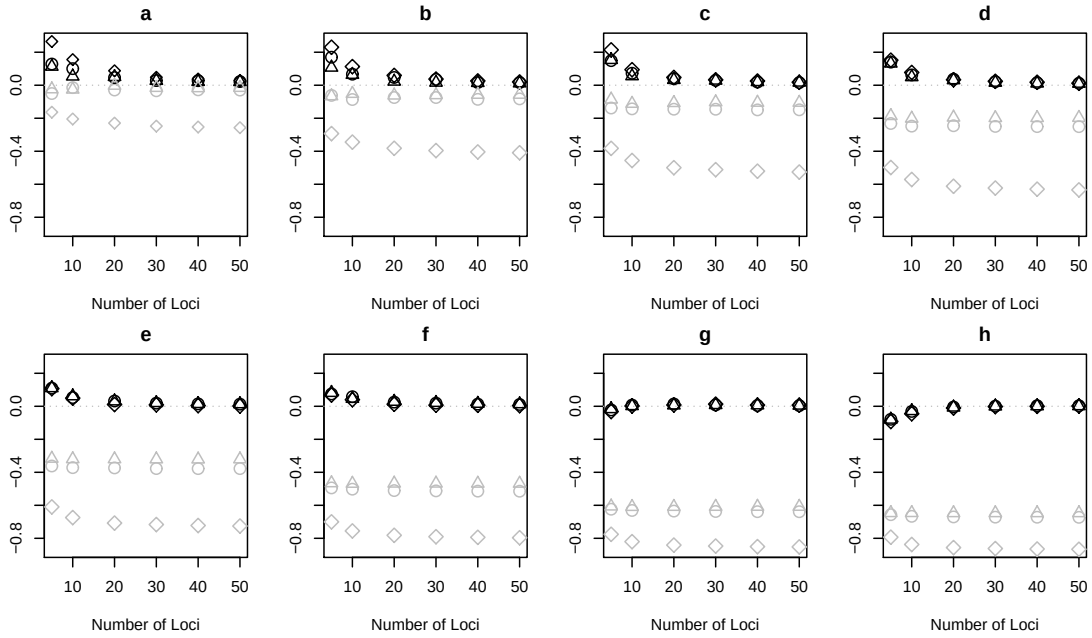

FIGURE 3.30: Fractional bias of true mutation rates $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
○/○- SMM/ISM estimate mutation rate distribution 1, ◇/◇ - SMM/ISM estimate mutation rate distribution 2, △/△- SMM/ISM estimate mutation rate distribution 3,

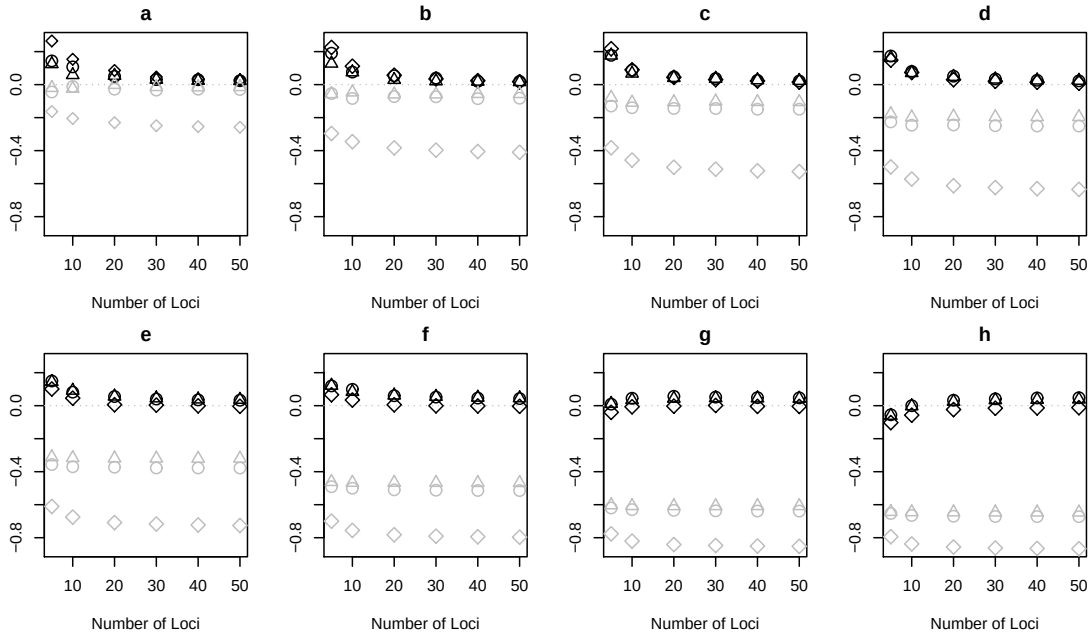

FIGURE 3.31: Fractional bias of empirical mutation rates $\hat{t}$ vs. number of loci: a. $t = 10$, b. $t = 25$, c. $t = 50$, d. $t = 100$, e. $t = 200$, f. $t = 400$, g. $t = 800$, h. $t = 1000$
○/○- SMM/ISM estimate mutation rate distribution 1, ◇/◇ - SMM/ISM estimate mutation rate distribution 2, △/△- SMM/ISM estimate mutation rate distribution 3,

SMM than the ISM (data not shown). Examining the FB based on the empirical mutation rates (fig. 3.31) similar conclusions can be drawn for the ISM estimates. However, for the SMM, the FB for mutational distribution 2 (◇) gradually diverges from the FB of the other two distributions (○ and △) as t increases. This is also due to mutation rate distribution 2 having the highest variance.

We end by exploring the effect of the number of meioses for all three mutation rate distributions using all four types of mutations rates when $n = 50$ and $t = 1000$, this being the most pronounced case of the difference mentioned above using the SMM. The number of meioses varies from 100-1,000,000 and for each number of meioses 10,000 runs are simulated for each mutation rate distribution.

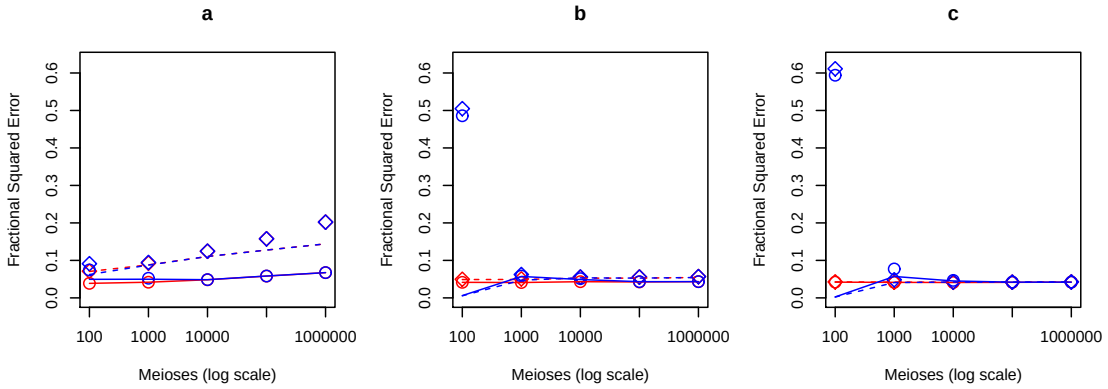

FIGURE 3.32: Fractional squared error and variance of $\hat{t}$ vs. number of meioses (log scale) a. mutation rate distribution 2, b. mutation rate distribution 1, c. mutation rate distribution 3
(○/solid red line - FSE/FV of true rates, ◇/dashed red line - FSE/FV of average of true rates, ○/solid blue line - FSE/FV of empirical rates, ◇/dashed blue line - FSE/FV of average of empirical rates)

Figure 3.32 shows the FSE and FV for the three distributions. Note that ordering of the graphs is according to decreasing variance of the mutation rate distribution. For mutational distribution 2, there is some difference in FSE at low numbers of meioses between the four types of mutation rates. However, it reduces as the number of meioses increases for the site-specific rates (○ and ○). The average rates' FSE (◇ and ◇) increases with increasing number of meioses with the former being superimposed by the latter (fig. 3.32a). The FV is fairly close to the FSE but contributes less for the average rates when there are more than 10,000 meioses. For mutation rate distributions 1, in figure 3.32b, the empirical site-specific and average rates (○ and ◇) have very high FSE with very low FV contribution at low meioses with the former reducing a lot thereafter and the FV contributing more. The true rates produce fairly similar FSE and fractional variance across all meioses (○ and ◇). A similar picture emerges for mutation rate distribution

3, although the FSE is slightly greater for the empirical rates (○ and ◇) at low meioses (fig. 3.32c).

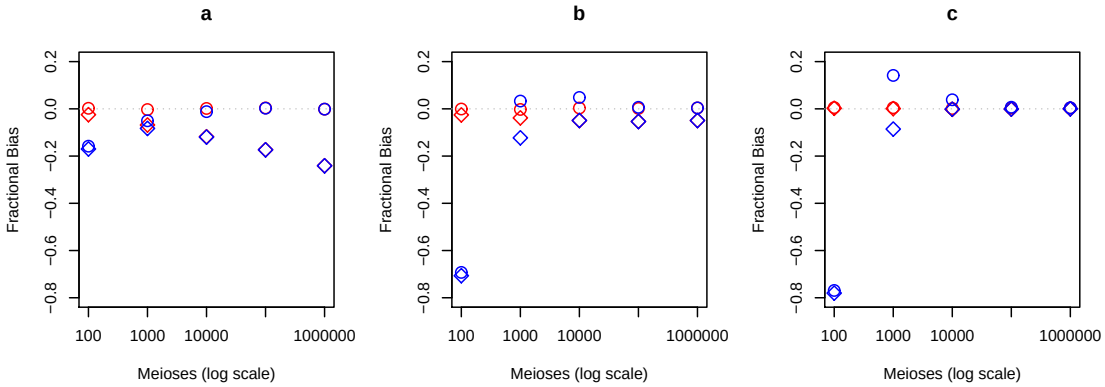

FIGURE 3.33: Fractional bias of $\hat{t}$ vs. number of meioses (log scale) a. mutation rate distribution 2, b. mutation rate distribution 1, c. mutation rate distribution 3 (○ - true rates, ◇ - average of true rates, ○ - empirical rates, ◇ - average of empirical rates)

The corresponding FB is show in figure 3.33. For mutational distribution 2 we find the estimates based on the true rates are unbiased for few meioses whilst those based on the empirical rates produce underestimates of t (fig. 3.33a). All four rates' FB converge before diverging into two groups (the site-specific rates and the average rates) as the number of meioses increase. The latter have increasing FB, whilst the former are unbiased. For mutation rate distribution 1, for 100 meioses, there is a considerable difference in FB between the true rates (○ and ◇), which are approximately unbiased, and the empirical rates which produce negative FB (○ and ◇, fig. 3.33b). The FB for the average rates converge as the number of meioses increase but remain negative. On the other hand, the true site-specific mutation rates' FB (○) is approximately constant and unbiased with increasing number of meioses. The empirical rates' FB (○) is negative for few meioses and thereafter produces positive values, reaching a maximum for 10,000 meioses, before approaching zero. A similar pattern in FB emerges with the empirical rates for mutation rate distribution 3 (fig. 3.33c). However, the maximum positive fractional bias occurs earlier, at 1,000 meioses. Furthermore all four rates' FB converge by 1,000,000.

3.6 Conclusions

Preliminary assessment of real data demonstrated that estimates of TMRCA based on the SMM were larger compared to the IAM, particularly for older TMRCAs.

Both models were implemented in a Bayesian framework which did not incorporate any underlying distribution for the mutation rates.

In order to assess the performance of the models we simulated replicate datasets for a wide range of TMRCA values using only ascertained mutation rates drawn from three different distributions. MLEs of TMRCA based on the SMM and ISM were computed using four sets of mutation rates: true site-specific, true average, empirical site-specific and empirical average.

The ISM performs better than the SMM at low values of TMRCA so long as the distribution from which the mutation rates are drawn has a low variance. Otherwise, the ISM systematically underestimates TMRCA, more and more as the value of TMRCA increases.

Using the SMM, estimates of TMRCA based on site-specific mutation rates are only affected by the variance of the mutation rate distribution for very old TMRCA, i.e. for TMRCAs which predate the surname establishment period. In addition, employing the average of the mutation rates results in less biased estimates of TMRCA than using site-specific rates but crucially only for short TMRCA values, provided the mutation rates are drawn from a distribution with a low variance. Lastly, differences between estimates of TMRCA are negligible using the true or empirical mutation rates again only when the TMRCA is small.

In summary, modelling the mutation rate distribution is necessary since its variance affects the estimates of TMRCA. The SMM, using site-specific mutation rates, adequately accounts for hidden mutations, which are more likely to occur for older MRCAs. In contrast the ISM does not and performs poorly when the variance of the mutation rate distribution is high.

Chapter 4

Modelling Mutational Mechanisms

In this chapter, we cover several aspects of modelling the various processes involving mutations rates. Firstly we consider the unequal stepwise mutation model (USMM) where we alter the SMM (section 2.1) to allow for unequal probabilities for the increase and decrease mutations.

We then develop a model which takes into account:

- the ascertainment of STRs, i.e. the initial discovery of variable loci at the population level,
- the calibration process in which empirical estimates of the mutation rates at each STR are derived typically by considering the number of mismatches between father-son pairs, i.e. a number of meioses, and
- the underlying mutation rate distribution.

This model will then be analysed using real data and thereafter using simulated data from the intermediary mutation rate review (section 2.3). In both we will carry out a sensitivity analysis of the model by misspecifying key parameters.

4.1 Preliminary Analysis

Walsh (2001) outlined the stepwise mutation model in which the probability of an increase and decrease mutation was equal, i.e. 0.5. However, in the mutation rate reviews we carried out, it was evident that real data showed a difference in the proportion of increase versus decrease mutations. We found that there was

a higher proportion of increase mutations at 0.5736 (95% CI: 0.5296-0.6166) in the final mutation rate review. This suggested that when a mutation occurs it is more likely that the mutation would be an increase in the number of repeats. Thus, in this section, we develop the unequal stepwise mutation model (USMM) for estimating the TMRCA (t) which allows for an unequal probability of increase and decrease mutations.

In figure 4.1 we depict the relationship between male 1 and male 2 under the USMM. As in the SMM we consider the direction of the mutations going from male 1 to male 2. Coupling this with the unequal probability of the two kinds of mutations we require the MRCA to be defined. Thus we have four classes of mutations; real increase, real decrease, pseudo increase and pseudo decrease mutations. The pseudo mutations are so named since they were in fact actually the opposite type of mutation generated in the lineage from the MRCA $\rightarrow$ Male 1, which become time-reversed when the flow of change is modelled from male 1 to male 2.

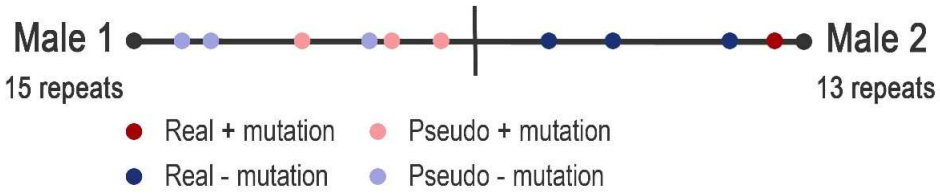

FIGURE 4.1: Schematic of USMM

We introduce the following notation:

- n_+: the number of increase mutations when moving in the direction Male 1 $\rightarrow$ MRCA $\rightarrow$ Male 2, i.e. the sum of the pseudo increase mutations and the real increase mutations;
- n_-: the number of decrease mutations when moving in the direction Male 1 $\rightarrow$ MRCA $\rightarrow$ Male 2, i.e. the sum of the pseudo decrease mutations and the real decrease mutations;
- ϕ_+: the probability of a real increase mutation;
- ϕ_-: the probability of a real decrease mutation;
- μ: the probability of any mutation, i.e. $\phi_- + \phi_+$

Consider the case $n_+ - n_- = 2k$, where k is an integer, i.e. the case where the observed repeat count difference is even. Then it follows that the sum of mutations

is also even, i.e. $n_+ + n_- = 2m$ where m is a positive integer. Let $X = n_+$, so $X \sim Bi(2m, 1/2)$. Then $X = m - k$ and we have two possibilities: $0 \leq k \leq m$ and $-m \leq k \leq 0$. Considering the former, it follows that

$$P(X = m - k | n_+ + n_- = 2m) = \binom{2m}{m-k} \left(\frac{1}{2}\right)^{2m}. \tag{4.1}$$

Now let L = the number of real decrease mutations from MRCA $\rightarrow$ Male 1 in n_+, i.e. the number of pseudo increase mutations. Then

$$L|m, k, \phi_-, \phi_+ \sim Bi(m - k, \nu_-), \text{ where } \nu_- = \frac{\phi_-}{\phi_- + \phi_+}.$$

Also let R = the number of actual increase mutations from MRCA $\rightarrow$ Male 1 in n_-, i.e. the number of pseudo decrease mutations. Then

$$R|m, k, \phi_-, \phi_+ \sim Bi(m + k, \nu_+), \text{ where } \nu_+ = \frac{\phi_+}{\phi_- + \phi_+}.$$

Thus

$$\begin{aligned} P(L = l, R = r | m, k) &= \binom{m-k}{l} \nu_-^l \nu_+^{m-k-l} \binom{m+k}{r} \nu_+^r \nu_-^{m+k-r} \\ &= \binom{m-k}{l} \binom{m+k}{r} \nu_-^{m+k-r+l} \nu_+^{m-k-l+r}. \end{aligned} \tag{4.2}$$

Now let Y = the total number of mutations, $n_+ + n_-$. So then $Y \sim Po\big(2t(\phi_- + \phi_+)\big)$, i.e.

$$P(Y = 2m | t) = e^{-2t(\phi_- + \phi_+)} \frac{[2t(\phi_- + \phi_+)]^{2m}}{(2m)!}. \tag{4.3}$$

Combining (4.1),(4.2) and (4.3) we obtain

$$\begin{aligned}
&P(n_+ - n_- = 2k|t) \\
&= \sum_{\substack{n_+,n_-,l,r \\ \text{s.t. } n_+-n_-=2k}} P(n_+, n_-, l, r|t) \\
&= \sum_{m=k}^{\infty}\sum_{r=0}^{m+k}\sum_{l=0}^{m-k} \frac{(2m)!}{(m+k)!(m-k)!}\left(\frac{1}{2}\right)^{2m}\frac{(m-k)!}{(m-k-l)!l!}\frac{(m+k)!}{(m+k-r)!r!} \\
&\left(\frac{\phi_-}{\phi_-+\phi_+}\right)^{m+k-r+l}\left(\frac{\phi_+}{\phi_-+\phi_+}\right)^{m-k-l+r} e^{-2t(\phi_-+\phi_+)}\frac{[2t(\phi_-+\phi_+)]^{2m}}{(2m)!} \\
&= e^{-2t(\phi_-+\phi_+)}\sum_{m=k}^{\infty}\sum_{r=0}^{m+k}\sum_{l=0}^{m-k} t^{2m}\frac{\phi_-^{m+k-r+l}\phi_+^{m-k+r-l}}{(m-k-l)!l!(m+k-r)!r!}.
\end{aligned}$$

Introducing the change of variable $m' = m - k$, we obtain

$$\begin{aligned}
&P(n_+ - n_- = 2k|t) \\
&= e^{-2t(\phi_-+\phi_+)}\sum_{m'=0}^{\infty}\sum_{r=0}^{m'+2k}\sum_{l=0}^{m'} t^{2m'+2k}\frac{\phi_-^{m'+2k-r+l}\phi_+^{m'+r-l}}{(m'-l)!l!(m'+2k-r)!r!} \\
&= e^{-2t(\phi_-+\phi_+)}\sum_{m'=0}^{\infty} t^{2m'+2k}\underbrace{\sum_{l=0}^{m'}\frac{\phi_+^{m'-l}\phi_-^{l}}{(m'-l)!l!}}_{a}\underbrace{\sum_{r=0}^{m'+2k}\frac{\phi_-^{m'+2k-r}\phi_+^{r}}{(m'+2k-r)!r!}}_{b}.
\end{aligned} \tag{4.4}$$

Using the binomial theorem to simplify a and b, (4.4) reduces to

$$\begin{aligned}
&P(n_+ - n_- = 2k|t) \\
&= e^{-2t(\phi_-+\phi_+)}\sum_{m'=0}^{\infty} t^{2m'+2k}\frac{(\phi_-+\phi_+)^{m'}}{m'!}\frac{(\phi_-+\phi_+)^{m'+2k}}{(m'+2k)!} \\
&= e^{-2t(\phi_-+\phi_+)}\sum_{m'=0}^{\infty}\frac{[(\phi_-+\phi_+)t]^{2m'+2k}}{m'!(m'+2k)!}.
\end{aligned} \tag{4.5}$$

The case with $-m \le k \le 0$ follows similarly. Combining the two cases, we have:

$$P(|n_+ - n_-| = 2k|t) = \begin{cases} e^{-2t(\phi_-+\phi_+)}\sum_{m'=0}^{\infty}\frac{[(\phi_-+\phi_+)t]^{2m'+2k}}{m'!(m'+2k)!} & k = 0; \\ 2e^{-2t(\phi_-+\phi_+)}\sum_{m'=0}^{\infty}\frac{[(\phi_-+\phi_+)t]^{2m'+2k}}{m'!(m'+2k)!} & k \ge 1; \\ 0 & k < 0. \end{cases} \tag{4.6}$$

Note that if $\phi_- + \phi_+$ is replaced by μ we recover (3.7), i.e. the same result as for the SMM. The case where $n_+ - n_-$ is odd also follows in an analogous way.

4.1.1 Discussions

That the model accounting for the directionality of the mutation (4.6) reduced to the SMM (3.7) was a surprise. On reflection, it is evident that the classification of a mutation (as real or pseudo and as increase or decrease) depends entirely on whether it is placed on the left or right of the MRCA when considering mutation process as a random walk (fig. 4.1). Since the ancestral haplotype is unknown there is additional symmetry which necessitates that the labelling of the descendants as Male 1 or Male 2 is completely arbitrary. Hence the simplified result.

Nonetheless, we checked to ensure this was indeed the case by simulating data allowing for unequal probabilities for an increase versus decrease mutation. In addition to the notation already used, we now define:

- ν: the fraction of the total mutation rate which causes an increase in the number of repeats of an STR, i.e. $\frac{\phi_+}{\phi_+ + \phi_-}$;
- n_{1+}: the number of real increase mutations from MRCA $\rightarrow$ Male 1, i.e. the number of pseudo decrease mutations;
- n_{1-}: the number of real decrease mutations from MRCA $\rightarrow$ Male 1, i.e. the number of pseudo increase mutations;
- n_{2+}: the number of real increase mutations from MRCA $\rightarrow$ Male 2;
- n_{2-}: the number of real decrease mutations from MRCA $\rightarrow$ Male 2.

The data simulation was carried out as follows for n loci:

1. Assign mutation rates: μ_i $(i = 1, \ldots, n)$.
2. Generate the total number of mutations at each locus: $n_+ + n_- \sim Po(2t\mu_i)$
3. Generate the number of real increase and decrease mutations in going from MRCA $\rightarrow$ Male 1 and MRCA $\rightarrow$ Male 2: $n_{1+}, n_{1-}, n_{2+}, n_{2-} \sim MN(n_+ + n_-, \frac{\nu}{2}, \frac{1-\nu}{2}, \frac{\nu}{2}, \frac{1-\nu}{2})$.
4. Compute the absolute difference in the number of STR repeats between males 1 and 2: $|n_{1+} - n_{1-} - n_{2+} + n_{2+}|$ at each locus.

We then calculated the simulated absolute difference of the number of STR repeats for two males 10,000 times with the parameters set as follows:

- $t = 50$ generations,
- $n = 50$ loci,
- $\mu_i = 0.1$ per locus per generation,
- $\nu = 0$ and 0.5 the latter corresponding to the SMM.

The choice of a fairly high mutation rate was simply to allow mutation to accumulate. Histograms of the resulting data were then plotted and examined for differences. In order to test formally whether the resultant distributions were indeed drawn from the same distribution, we apply a two-sample Kolmogorov-Smirnov test by using the R function `ks.boot()`. The null hypothesis in this case is that the two sets of data are drawn from the same distribution.

Figure 4.2 shows the distribution of the absolute difference of the number of repeats between two simulated male STR profiles, for $\nu = 0$ and $\nu = 0.5$. Subjectively there appeared to be little difference between the two distributions. The p-value from the Kolmogorov-Smirnov test was 0.92 was obtained. We conclude that there is no statistically significant difference between the two distributions.

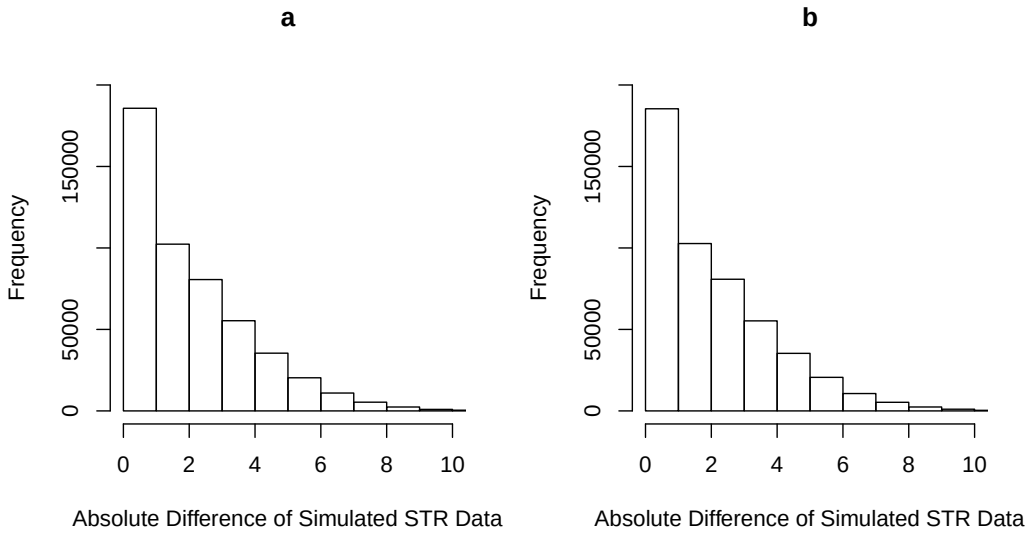

FIGURE 4.2: Histogram of simulated absolute STR differences: a. $\nu = 0$ b. $\nu = 0.5$

Thus it was unnecessary to take into account the asymmetric proportion of increase and decrease mutations.

4.1.2 Conclusions

In this section we derived formulae for the likelihood assuming that the probability of an increase and a decrease mutation was unequal. Recall that this was reflective of real data, both in the intermediary and final mutation rate reviews. Our derivation led us back to the likelihood for the standard SMM, as outlined by Walsh (2001). We carried out a simple simulation, which verified this result.

4.2 Introduction

This section will outline the development of the model that accounts for the mutational process. Following on from the previous chapter, we again assume that mutation rates are drawn from an underlying Gamma distribution. However the model must also take into account that the loci typed when determining the TMRCA are not drawn randomly. Indeed a process of locus ascertainment is explicitly applied to ensure that the loci typed are variable at the population level (Nicholson et al., 2002). Once ascertained, these loci are then calibrated, i.e. the mutation rates estimated by examining the number of mismatches between an often large number of father-son pairs. The method by which the data is simulated is outlined first, followed by the formation of the posterior distribution.

We will use the following notation in this section:

- μ: the, per locus per generation, mutation rate. This will usually be indexed by i referring to the i^{th} locus;
- α: the shape parameter in the Gamma distribution used to describe the underlying mutation rate distribution;
- β: the scale parameter in the Gamma distribution used to describe the underlying mutation rate distribution;
- N_e: the effective size of the population on which the loci were ascertained;
- L: the total branch length, in units of N_e generations, of the tree of Y-chromosomes in the sample from which the loci were ascertained.

4.3 Materials and Methods

4.3.1 Data Simulation

In the data simulation process we aim to simulate mutation rates across n loci such as might be used in a Y-STR study to estimate TMRCA. This requires generating mutation rates, ascertaining whether the rates are variable and thereafter calibrating them.

The steps for this three-stage process are outlined below, where each locus is simulated identically and independently. Firstly:

$$\mu_i \sim Ga(\alpha, \beta) \quad \text{for} \quad i = 1, \ldots, n. \tag{4.7}$$

The variability of a locus is determined by defining the ascertainment sample size, n_{asc}, which is typically $4 - 10$ (Kayser et al., 2004) and assumed fixed across all loci. By the coalescent theory (Wakeley, 2009) we have the following p.d.f. for L

$$f_L(L) \sim \frac{n_{asc} - 1}{2} e^{-L/2} \left(1 - e^{-L/2}\right)^{n_{asc}-2}, \tag{4.8}$$

in the ascertainment sample.

Sampling L directly from this distribution is not possible so we apply the probability integral transformation (DeGroot and Schervish, 2002). We sample from this distribution via its c.d.f.

$$F_L(s) = \int_0^s f_L(s')ds' = \left(1 - e^{-s/2}\right)^{n_{asc}-1},$$

by introducing $u \sim Un[0, 1]$ and evaluating

$$s = F_L^{-1}(u),$$

which is drawn from the distribution.

Hence,

$$s = -2 \ln \left(1 - u^{1/(n_{asc}-1)}\right).$$

The next step involves generating the effective population size:

$$N_e \sim N(\mu_{N_e}, \sigma^2_{N_e}). \tag{4.9}$$

We use estimates of μ_{N_e} and $\sigma^2_{N_e}$ based on the work of Thomson et al. (2000) who examined three genes on the Y-chromosome: SMCY, DBY and DFFRY, covering 64,120 basepairs. They estimated the mutation rate to be 1.24×10^{-9} per site per year. The authors obtain a mean effective population size of 6000 using the software GENETREE to provide a maximum likelihood estimate of $\theta = 2N_e\mu_g$, where μ_g is a per gene per generation mutation rate. This was estimated at 24, with a 95% probability interval of (6.7, 33.9). For our purpose we need the variance of the distribution from which the effective population size is estimated.

Using a conversion of 25 years per generation we obtain a per generation mutation rate of 0.00198772. Based on the point and interval estimate of θ we can estimate the mean, lower and upper limits of the effective population size from $N = \theta/2\mu_g$. So we have:

$$\text{at } \theta = 6.7, \text{ Lower } N_e = 1685.348,$$
$$\text{at } \theta = 24, \text{ Mean } N_e = 6037.068,$$
$$\text{at } \theta = 33.9, \text{ Upper } N_e = 8527.358.$$

Thus we use $\mu_{N_e} = 6037$ and assuming that:

$$\text{Upper}N_e - \text{Lower}N_e = 2 \times 1.96 \times \sigma,$$

we estimate $\sigma_{N_e} = 1745$.

Next we model the ascertainment process. Let $R_i = 1$ if the locus is ascertained (i.e. variable), otherwise let $R_i = 0$ the locus is not ascertained.

$$R_i \sim Bi(1, 1 - e^{-\mu_i L N_e}) \quad \text{for} \quad i = 1, \ldots, n. \tag{4.10}$$

This is because, if we assume that the number of mutations, $Y \sim Po(\mu_i L N_e)$, the probability that a mutation occurs is:

$$1 - Po(Y = 0|L, N_e, \mu_i) = 1 - e^{-\mu_i L N_e}.$$

Suppose k loci are ascertained whilst $n-k$ are not ascertained. For the former k loci, we can calibrate their mutation rates by taking the number of mutations in m_i meioses:

$$r_i \sim Bi(m_i, \mu_i) \quad \text{for} \quad i = 1, \ldots, k. \tag{4.11}$$

We can estimate the parameters of interest at this stage, namely α, β, L, and N_e, based on the simulated data, i.e. r_i at each ascertained locus, $i = 1, \ldots, k$, the number of meioses, m_i and the ascertainment sample size, n_{asc}.

However we can alter the scenario slightly: whilst there may be a large number of loci which are ascertained, it is more realistic that only a subset of them will be calibrated (fig. 4.3).

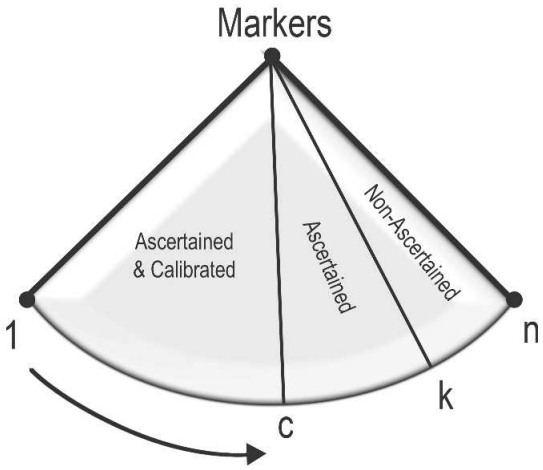

FIGURE 4.3: Categories of markers in the mutational mechanisms model

Hence, suppose c loci are calibrated from within the ascertained loci, we will refer to these as the calibrated loci. Hence $k-c$ ascertained loci will not be calibrated. For these the only information we have will be based on the ascertainment process (4.10), whilst for those calibrated we will have this information as well as the calibration data, i.e. we restrict (4.11) as follows:

$$r_i \sim Bi(m_i, \mu_i) \quad \text{for} \quad i = 1, \ldots, c. \tag{4.12}$$

It is on the basis of model that we estimate the parameters of interest by developing a fully probabilistic model using Bayesian methodology.

4.3.2 Bayesian Modelling

We now begin to derive the posterior distribution for inference of the mutational processes using Bayesian statistics. The full joint density of parameters and data is:

$$\begin{aligned} &P\left(\{\mu_i\}, L, N_e, \alpha, \beta, \{R_i\}, \{r_i\}, n_{asc}, \{m_i\}\right) \\ &= P\left(\{r_i\}|\{m_i\}, \{\mu_i\}, \{R_i\}\right) P\left(\{R_i\}|\{\mu_i\}, L, N_e\right) P\left(L|n_{asc}\right) P\left(\{\mu_i\}|\alpha, \beta\right) P\left(\alpha\right) \\ &\quad \times P\left(\beta\right) P\left(N_e\right), \end{aligned} \tag{4.13}$$

where we have exploited the conditional independence of many of the parameters in our hierarchical model.

Let the sets of indices corresponding to the calibrated, ascertained and non-ascertained loci, respectively be $\mathfrak{C}$, $\mathfrak{A}$, $\mathfrak{N}$. Then this probability is:

$$\begin{aligned} &= P\left(\{r_i : i \in \mathfrak{C}\}|\{m_i\}, \{\mu_i\}\right) P\left(R_i = 1 : i \in \mathfrak{A}|\{\mu_i\}, L, N_e\right) \\ &\quad \times P\left(R_i = 0 : i \in \mathfrak{N}|\{\mu_i\}, L, N_e\right) \\ &\quad \times P\left(\{\mu_i : i \in \mathfrak{C} \cup \mathfrak{A} \cup \mathfrak{N}\}|\alpha, \beta\right) P\left(L|n_{asc}\right) P\left(\alpha\right) P\left(\beta\right) P\left(N_e\right). \end{aligned} \tag{4.14}$$

From (4.10) and (4.12), we have:

$$\begin{aligned} &P(\{r_i : i \in \mathfrak{C}\}|\{m_i\}, \{\mu_i\}) P(\{R_i = 1 : i \in \mathfrak{A}\}|\{\mu_i\}, L, N_e) \\ &\quad \times P(\{R_i = 0 : i \in \mathfrak{N}\}|\{\mu_i\}, L, N_e) \\ &= \prod_{i=1}^{c} P(r_i|m_i, \mu_i) P(R_i = 1|\mu_i, L, N_e) \prod_{i=c+1}^{k} P(R_i = 1|\mu_i, L, N_e) \\ &\quad \times \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L, N_e) \\ &= \prod_{i=1}^{c} \binom{m_i}{r_i} \mu_i^{r_i} \left(1 - \mu_i\right)^{m_i - r_i} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=c+1}^{k} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=k+1}^{n} e^{-\mu_i L N_e}. \end{aligned} \tag{4.15}$$

From (4.7) and (4.8) we also have:

$$P(\{\mu_i\}|\alpha, \beta) = \prod_{i=1}^{n} \frac{\mu_i^{\alpha-1} e^{-\mu_i/\beta}}{\beta^{\alpha} \Gamma(\alpha)} \tag{4.16}$$

$$P(L|n_{asc}) = \frac{n_{asc} - 1}{2} e^{-L/2} \left(1 - e^{-L/2}\right)^{n_{asc}-2}. \tag{4.17}$$

The joint probability of the data and the unknown parameters may also be written as:

$$\begin{aligned}
&P(\{\mu_i : i \in \mathfrak{C} \cup \mathfrak{A} \cup \mathfrak{N}\}, L, N_e, \alpha, \beta, \{R_i\}, \{r_i\}, n_{asc}, \{m_i\}) \\
&= P(\{\mu_i\}, L, N_e, \alpha, \beta, |n_{asc}, \{m_i\}, \{R_i\}, \{r_i\}) P(n_{asc}, \{m_i\}, \{R_i\}, \{r_i\}). \qquad (4.18)
\end{aligned}$$

The last term, the probability of the data is constant (i.e. does not depend on the parameters) so by Bayes' theorem (1.2), we have:

$$\begin{aligned}
&P(\{\mu_i\}, L, N_e, \alpha, \beta, \{R_i\}, \{r_i\} | n_{asc}, \{m_i\}) \\
&\propto P(\{r_i\} | \{m_i\}, \{\mu_i\}, \{R_i\}, L, N_e, \alpha, \beta, n_{asc}) P(\{\mu_i\}, \{R_i\}, L, N_e, \alpha, \beta, n_{asc}) \\
&\propto \prod_{i=1}^{c} \binom{m_i}{r_i} \mu_i^{r_i} (1-\mu_i)^{m_i - r_i} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=c+1}^{k} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=k+1}^{n} e^{-\mu_i L N_e} \\
&\quad \times \prod_{i=1}^{n} \frac{\mu_i^{\alpha-1} e^{\mu_i/\beta}}{\beta^\alpha \Gamma(\alpha)} \\
&\quad \times \frac{n_{asc} - 1}{2} e^{-L/2} \left(1 - e^{-L/2}\right)^{n_{asc}-2} \\
&\quad \times P(\alpha) P(\beta) P(N_e). \qquad (4.19)
\end{aligned}$$

Using exponential priors for α and β and the normal prior for N_e truncated at zero we have the posterior distribution:

$$\begin{aligned}
&P(\{\mu_i\}, L, N_e, \alpha, \beta, \{R_i\}, \{r_i\} | n_{asc}, \{m_i\}) \\
&\propto \prod_{i=1}^{c} \binom{m_i}{r_i} \mu_i^{r_i} (1-\mu_i)^{m_i - r_i} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=c+1}^{k} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=k+1}^{n} e^{-\mu_i L N_e} \\
&\quad \times \prod_{i=1}^{n} \frac{\mu_i^{\alpha-1} e^{-\mu_i/\beta}}{\beta^\alpha \Gamma(\alpha)} \\
&\quad \times \frac{n_{asc} - 1}{2} e^{-L/2} \left(1 - e^{-L/2}\right)^{n_{asc}-2} \\
&\quad \times e^{-\lambda_\alpha \alpha} e^{-\lambda_\beta \beta} \frac{1}{\sigma_{N_e}\sqrt{2\pi}} \exp\left(-\frac{(N_e - \mu_{N_e})^2}{2\sigma_{N_e}^2}\right) \qquad (4.20)
\end{aligned}$$

4.3.3 Markov-Chain Monte-Carlo Sampling

MCMC allows simulation from a distribution of interest by making this distribution the limiting distribution of a Markov chain. Once the chain has reached equilibrium, it produces dependent draws from the target distribution, in this case,

a posterior distribution (Gamerman, 1997). The Markov chain should be appropriately constructed and there are a large number of techniques available to do so (see Gilks et al. (1998) or Gamerman (1997)).

We use an adaptive Metropolis algorithm to achieve the objective of sampling from the posterior distribution of interest, which includes an automated procedure to update the tuning parameters.

4.3.3.1 Metropolis-Hastings Algorithm

The M-H algorithm is a sampling technique introduced in a specific context by Metropolis et al. (1953) and later developed in a general setting by Hastings (1970). The notation used to describe the M-H algorithm is given below:

- x: the vector of parameters,
- $p(x = z) = f(z)$: the posterior density of x,
- $p(x = z^*|x = z) = g(z, z^*)$: - the proposal distribution for proposing a new value of $x = z^*$ given the current value of $x = z$.

The algorithm samples from the posterior distribution as follows:

1. Start with some initial values of the parameters x_0 and initialize the step counter $k = 0$.
2. Depending on the current value, x_k, select a new proposed value, x_{k+1}, from the proposal distribution $p(x_{k+1} = z^*|x_k = z) = g(z, z^*)$.
3. Calculate the Metropolis ratio:

$$R = \frac{f(z^*)g(z^*, z)}{f(z)g(z, z^*)}.$$

   (Note: where the proposal distribution is symmetric, i.e. $g(z, z^*) = g(z^*, z)$ we calculate:

$$R = \frac{f(z^*)}{f(z)}$$

   This is often termed the Metropolis Algorithm.)
4. Generate $S \sim Un[0, 1]$ and compute the acceptance probability:

$$\alpha(z, z^*) = \min\{1, R\}.$$

5. If $S \leq \alpha$ we accept the proposed value, i.e. $x_{k+1} = z^*$, otherwise if $S > \alpha$ we reject it (thus remaining at the current value, i.e. $x_{k+1} = z$).
6. Increment the step counter and repeat from step 2 until convergence is reached.
7. Remove the 'burn-in' steps and thin as required.
8. Compute summary statistics for the vector of parameters.

Several aspects of the algorithm will now be expanded further. In terms of updating vector x, it may be updated parameter by parameter (see section 4.3.3.2) allowing flexibility in the order parameters are updated and how often, although it is possible to update x as a whole.

Burn-in is required to discard the initial steps in the chain before the chain has reached convergence, i.e. to remove the dependency of the chain on the initial values, x_0. Thinning, whilst not mandatory (Raferty and Lewis, 1998), is ideal if we wish to reduce the number of draws stored from a single MCMC chain. This involves saving only every m^{th} step ($m > 1$). It may also be used to reduce the correlation between steps from a chain, although this is often directly related to the proposal distribution used. For example, if a proposal is close to the current value, it is likely to be accepted making the chain highly correlated, whereas a proposal far from the current value may be unlikely and therefore rejected. The latter may mean the chains may not sample from the full space.

The aim is to have a fairly uncorrelated but well-mixed chain and several diagnostics exist to examine the effectiveness of the proposal distribution and/or initial values. Firstly, plotting the accepted value at each step in the chain against the step number may help identify common problems such as bad initial values, slow mixing and insufficient chain length to sample from the distribution of interest. Secondly, the acceptance rate, i.e. the average percentage of steps which are accepted should be computed (Hastings, 1970). It has been suggested that an acceptance rate of 20-50% reflects the optimum balance of acceptance/rejection of proposals (Gamerman, 1997). A further diagnostic is the ergodic average (Gamerman, 1997):

$$\bar{x}_{sim} = \frac{1}{n_{sim}} \sum_{k=1}^{n_{sim}} x_k, \tag{4.21}$$

where n_{sim} is the total number of steps in the chain after burn-in. A plot of the ergodic average against step number may be useful in monitoring the convergence

of the chain. Further retrospective discussion of the methodology is provided in Chapter 7.

4.3.3.2 Adaptive MCMC

Typically, the proposal distribution is a normal centred on the current value with the user specifying the variance of the normal, $g(z^*, z) \sim N(z, \sigma^2)$. Thus the proposal distribution is symmetric, i.e. $g(z^*, z) = g(z, z^*)$. The Metropolis ratio may be straightforwardly calculated (step 3 in section 4.3.3.1).

Now, σ^2 is determined by the user. If it is too large the acceptance rate will be too low. Conversely if σ^2 is too small most proposals will be accepted. Consequently the acceptance rate will be too high and the chain will be highly correlated. Roberts and Rosenthal (2009) propose an adaptive Metropolis algorithm as an automated method of choosing the value σ^2. Define the log standard deviation of the proposal distribution for the ith variable being updated in x as l_i. These l_i are then adapted to ensure the acceptance rate is close to 0.44.

Specifically the algorithm involves:

1. Initialise l_i and the batch counter $b = 0$.
2. Implement a specified number of updates of the Metropolis algorithm (section 4.3.3.1, with each parameter updated individually).
3. Compute the acceptance rate of the batch.
4. Compute
$$\delta(b) = \min\left(0.01, b^{-\frac{1}{2}}\right)$$
5. Update l_i as follows: If the acceptance rate > 0.44,
$$l_i = l_i + \delta(b).$$
If the acceptance rate ≤ 0.44
$$l_i = l_i - \delta(b).$$
Conditioned on l_i lying in $[-M, M]$, where M is an integer
6. Increment the batch counter and repeat steps 2-5 as required.

In order to compare the effectiveness of this adaptive algorithm, the autocorrelation time of the updates in the chain may be calculated using the `acf()` function in R. If the adaptive scheme is effective, it should reduce this time compared to a non-adaptive approach. We also compute the average squared jumping distance of the chains. Assuming there are r updates of each parameter, the average squared jumping distance is computed as:

$$\frac{1}{r-1}\sum_{i=1}^{r-1}(x_{i+1}-x_i)^2,$$

and should be increased by an effective adaptive scheme.

4.3.3.3 Implementation: Modelling Mutational Mechanisms

In this section we implement the adaptive Metropolis algorithm to sample from the posterior distribution of the parameters in the mutational processes model (4.20)

In the algorithm the parameters, the components of

$$\theta = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \\ \mu_{k+1} \\ \vdots \\ \mu_n \\ L \\ N_e \\ \alpha \\ \beta \end{bmatrix}, \tag{4.22}$$

are updated one-by-one.

The proposal distribution for each element in θ is a normal centred at the current value with the variance updated according to the adaptive scheme. As the proposal distribution is symmetric in all cases, the Metropolis ratio is

$$R = \frac{f(\theta^*)}{f(\theta)}, \tag{4.23}$$

where $\theta^* = \theta$ for each component other than the one being updated and $f(\theta)$ is the posterior distribution. Since R may be expensive to compute, we can simplify it according to which parameter is being updated. For example, for L, the total branch length, we have:

$$\begin{aligned} R &= \frac{P(\{\mu_i\}, L^*, N_e, \alpha, \beta, \{R_i\}, \{r_i\}|n_{asc}, \{m_i\})}{P(\{\mu_i\}, L, N_e, \alpha, \beta, \{R_i\}, \{r_i\}|n_{asc}, \{m_i\})} \\ &= \frac{P(\{r_i : i \in \mathfrak{C}\}|\{m_i\}, \{\mu_i\})P(\{R_i = 1 : i \in \mathfrak{A}\}|\{\mu_i\}, L^*, N_e)}{P(\{r_i : i \in \mathfrak{C}\}|\{m_i\}, \{\mu_i\})P(\{R_i = 1 : i \in \mathfrak{A}\}|\{\mu_i\}, L, N_e)} \\ &\quad \times \frac{P(\{R_i = 0 : i \in \mathfrak{N}\}|\{\mu_i\}, L^*, N_e)P(\{\mu_i\}|\alpha, \beta)P(L^*|n_{asc})P(\alpha)P(\beta)P(N_e)}{P(\{R_i = 0 : i \in \mathfrak{N}\}|\{\mu_i\}, L, N_e)P(\{\mu_i\}|\alpha, \beta)P(L|n_{asc})P(\alpha)P(\beta)P(N_e)} \\ &= \frac{P(L^*|n_{asc}) \prod_{i=1}^{k} P(R_i = 1|\mu_i, L^*, N_e) \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L^*, N_e)}{P(L|n_{asc}) \prod_{i=1}^{k} P(R_i = 1|\mu_i, L, N_e) \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L, N_e)}. \end{aligned} \tag{4.24}$$

This R is then used to either accept or reject the proposal, L^*. Similarly, we can use a reduced form of R when updating the other parameters. So, for N_e, we have:

$$R = \frac{P(N_e^*) \prod_{i=1}^{k} P(R_i = 1|\mu_i, L, N_e^*) \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L, N_e^*)}{P(N_e) \prod_{i=1}^{k} P(R_i = 1|\mu_i, L, N_e) \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L, N_e)}; \tag{4.25}$$

for updating α:

$$R = \frac{P(\alpha^*) \prod_{i=1}^{n} P(\{\mu_i\}|\alpha^*, \beta)}{P(\alpha) \prod_{i=1}^{n} P(\{\mu_i\}|\alpha, \beta)}; \tag{4.26}$$

for updating β:

$$R = \frac{P(\beta^*) \prod_{i=1}^{n} P(\{\mu_i\}|\alpha, \beta^*)}{P(\beta) \prod_{i=1}^{n} P(\{\mu_i\}|\alpha, \beta)}; \tag{4.27}$$

for updating the mutation rate at each calibrated locus, i.e. μ_i where $i = 1, \ldots, c$:

$$R = \frac{P(r_i|m_i, \mu_i^*)P(R_i = 1|\mu_i^*, L, N_e)P(\mu_i^*|\alpha, \beta)}{P(r_i|m_i, \mu_i)P(R_i = 1|\mu_i, L, N_e)P(\mu_i|\alpha, \beta)}; \tag{4.28}$$

for updating the mutation rate at each ascertained locus, i.e. μ_i where $i = c + 1, \ldots, k$:

$$R = \frac{P(R_i = 1|\mu_i^*, L, N_e)P(\mu_i^*|\alpha, \beta)}{P(R_i = 1|\mu_i, L, N_e)P(\mu_i|\alpha, \beta)}; \tag{4.29}$$

and for updating the mutation rate at each non-ascertained locus, i.e. μ_i where $i = k + 1, \ldots, n$:

$$R = \frac{P(R_i = 0|\mu_i^*, L, N_e)P(\mu_i^*|\alpha, \beta)}{P(R_i = 0|\mu_i, L, N_e)P(\mu_i|\alpha, \beta)}. \tag{4.30}$$

The proposal will be rejected immediately if it is outwith the acceptable range for the parameter e.g. if $L^* < 0$ then $R = 0$. Otherwise we calculate R and generate $S \sim Un[0, 1]$. Thereafter we accept or reject the proposal as outlined in step 5 of the M-H Algorithm (section 4.3.3.1).

4.3.4 Data Simulation and Analysis Program

The use of the function `mumodel` is detailed below:

Description

`mumodel` allows the user to simulate data from the mutational distribution underlying the model by specifying the model parameters and sample size (see DETAILS below). The data are then analysed using MCMC to provide estimates of the mutation rates across all loci, α, β, N_e and L (in units of N_e and generations) and the mean and variance of the distribution from which the mutation rates are drawn. Several diagnostics are also produced to evaluate the performance of the chain.

Usage

```
mumodel(BatchLen=50, TotBatch=1000, BinBatch= 10, AccRate=0.44,
```

```
maxLSD, Loci, pCAL=0.211, Pnasc=NA, ascsamp=8, meioses=10000, L,
Ne, muNe=6037, sdNe=1745, alpha=1.703311,beta=0.001404, startL=3,
startNe=6037, startalpha=1.7, startbeta=0.00014, startTrue=T,
psdmu=0.002, psdL=0.5, psdNe=500, psdAlpha=0.05, psdBeta=0.06,
lambdaA=0, lambdaB=0, getdata=F, runs, method=2,
graph="graphres.eps")
```

Required Arguments

`BatchLen`: this is the number of runs within each batch before the log standard deviation of the proposal distribution is adapted.

`TotBatch`: the total number of batches to run.

`BinBatch`: the total number of batches to remove as burn-in.

`AccRate`: the optimum acceptance rate.

`MaxLSD`: the boundary for proposed values of the log standard deviation of the proposal distribution.

`Loci`: the number of STRs, n.

`pCAL`: the proportion of the ascertained loci that are calibrated.

`Pnasc`: assigns the percentage of non-ascertained loci in the analysis. It can be set to a different value by choosing a value in the range $(0, (1 - ascCAL/\texttt{Loci}))$ where $ascCAL$ is the number of calibrated loci. When set to `NA` (default), the true percentage of non-ascertained loci is passed to the MCMC sampler.

`ascsamp`: the ascertainment sample size.

`meioses`: the number of meioses used to calibrate the ascertained loci.

`L`: the total branch length in the Y-chromosome sample in which loci are ascertained, in units of N_e generations.

`Ne`: the effective population size.

`muNe`: the prior mean effective Y-chromosome population size in the prior for N_e.

`sdNe`: the standard deviation of the effective population size, in the prior for N_e.

`alpha`: the shape parameter in the Gamma distribution used to sample mutation rates.

`beta`: the scale parameter in the Gamma distribution used to sample mutation rates.

`startL`: the initial value for the MCMC for L.

`startNe`: the initial value for the MCMC for N_e.

`startalpha`: the initial value for the MCMC for α.

`startbeta`: the initial value for the MCMC for β.

`startTrue`: logical, if `TRUE` initialises the MCMC chains to start at the true value for each parameter.

`psdmu`: the initial standard deviation for the proposal distribution of the mutation rates.

`psdL`: the initial standard deviation for the proposal distribution of L.

`psdNe`: the initial standard deviation for the proposal distribution of N_e.

`psdAlpha`: the initial standard deviation for the proposal distribution of α.

`psdBeta`: the initial standard deviation for the proposal distribution of β.

`lambdaA`: the lambda value for α's prior distribution.

`lambdaB`: the lambda value for β's prior distribution.

`getdata`: logical, if `TRUE` the program will try to retrieve the data and parameters stored in a list named 'results'.

`runs`: used to produce a graph of the mutation rate distribution with number of draws equal to runs.

`method`: controls which posterior distribution is to be sampled from; 2 for the distribution in (4.20), 1 for any competing posterior distribution.

`graph`: the name of the file that must end in ".eps" that any graphical diagnostics will be saved to. If `NA`, no graphical diagnostics will be returned or saved.

Side Effects & Returns

The function returns a list of the true parameter values, MCMC mean, standard deviation and credible regions for L (in both generations and N_e units), N_e, α, β, mean of gamma, variance of gamma and L (generations). In addition the list includes the acceptance rates for each parameter. Also included are details of the mutation rates including the true values, MCMC means, standard deviations and credible regions, the number of mutations and meioses for the calibrated rates as well as their calibrated estimates. Additionally, the true proportion and number of calibrated and non-ascertained loci with any misspecified values stated, along with the total number of loci are given. Finally the list includes the mean and standard deviation of the prior for N_e, as well as the number of simulated Y-chromosomes used to determine the ascertained mutation rates.

For `graph=''graphres.eps''`, a 3×6 plot of the chains for the first mutation rate (an ascertained locus), the last mutation rate (a non-ascertained locus), L (units of N_e generations), N_e, α and β are shown, with the true value indicated by a solid grey line (fig. 4.6). The corresponding updates of the log of the standard deviation of each of the parameters' proposal distributions follow. The last row shows the histogram of the MCMC samples, excluding burn-in, for the corresponding parameters.

DETAILS

The program is entirely coded in the language R which simulates data according to the steps outlined in section 4.3.1, creates the appropriate update vector and passes this and other relevant parameters to the adaptive MCMC loop. This generates a new proposal for the parameter being updated and either accepts or rejects it on the basis of the posterior distribution (4.20).

After every `BatchLen` updates of all the parameters, the log standard deviation of the proposal distribution is adapted according to section 4.3.3.2 and the specified optimum acceptance rate. Once the chosen number of batches, i.e. `TotBatch`, has been constructed for each parameter, the R code generates the summary statistics followed by the graphical diagnostics (if specified) and returns the list specified in Side Effects & Returns, above. Where `getdata=T`, the data must have been previously stored in a list named 'results'. This is ideal for comparing different analyses of the same data, for example when altering the amount of data or misspecifying parameters.

4.3.5 Software

The software used to develop the mutational mechanisms model and related analyses was R. The results presented in this chapter were obtained on a Linux platform.

4.3.6 Analysis

The first set of results is based on the intermediary mutation rate review (section 2.3.2). The aim is to estimate the parameters of interest based on the real data so as to ensure the simulated data are generated from a plausible set of model parameters.

We will show the results obtained from the application of the model to estimates of 86 Y-STR mutation rates obtained from the intermediary mutation rate review. This will provide estimates of the mutation rates: the calibrated and non-calibrated loci, as well as non-ascertained loci. Furthermore we can derive estimates of the shape and scale parameters of the gamma distribution from which the mutation rates are assumed to be drawn. Importantly within the framework we have developed we must know the total number of STRs on the Y-chromosome along with the number of ascertained loci and the subset of the loci which are calibrated. In order to provide reasonable values for these necessary parameters we reviewed the work of Kayser et al. (2004).

These authors carried out a survey of STRs on a single Y-chromosome. Now although the authors define an STR or microsatellite to consist of repeats unit of 1-6 bp, the survey involved looking for repeat units of 3-6 bp and also having a minimum of eight perfectly matching repeats. 475 STRs were counted as a result of these criteria. Thereafter, Kayser et al. (2004) attempted to design primers for each STR. However, this was unsuccessful for 149 of them. Additionally 45 were previously known STRs. In order to establish a male-specific protocol, the 281 new STRs' underwent PCR-based optimization procedures in three males and two females. 48 of these loci were not optimized and for an additional 67 STRs it proved difficult to design protocols that did not lead to amplification in females. Consequently, 194 STRs (166 new STRs and 28 previously known) were typed in 8 males, belonging to Y-SNP haplogroups R, I, B, A, J, C, E and K*). 28 loci were found to be non-variable across the eight Y-chromosomes. Thus 85.6% of the typed loci were found to be variable. Applying this percentage to the total number of Y-STRs (475), we obtain the theoretical value of 407 Y-STRs being variable in an ascertainment sample size of 8. In addition, based on the intermediary mutation rate review, 86 STRs comprise the ascertained and calibrated loci.

Based on this data, we applied our MCMC sampler with 5000 batches each of length 50 updates. 10% of the batches were discarded as burn-in and the results of this single run are given for the main parameters of interest, i.e. α, β, L, N_e, as well as a calibrated (and hence ascertained locus) and non-ascertained locus.

In addition we examine the effect of varying the percentage of non-ascertained loci on the parameters paying particular attention to α and β. In figure 4.3, this means we vary the size of the last segment between k and n. In particular we allow the percentage to vary as follows: 0%, 5%, 10%, 14.4% (the value estimated on the

basis of Kayser et al. (2004)), 20%, 40%, 57.9%, 81.9% (since there are 18.1% (86) calibrated loci, the maximum percentage of non-ascertained loci is $100 - 18.1\%$).

The results have been obtained from a single run of the MCMC sampler with 1000 batches of 50 updates with 100 batches discarded as burn-in. Graphs of the estimates of the main parameters (α, β, L in generations, N_e) are plotted against the percentage of non-ascertained loci. We also examine the parameters of the mutation rate distribution, i.e. the mean and variance of the gamma distribution with shape α and scale β. The final summary of the real data is a plot of the mean of the following sets of mutation rates:

- Low: the calibrated rates equal to zero;
- Intermediate: the calibrated rates less than 0.003 but greater than zero;
- High: the calibrated rates greater than 0.003;
- Non-calibrated loci: the ascertained loci which are not calibrated;
- Non-ascertained loci.

We then further examine varying the percentage of non-ascertained loci but on a much narrower range between 10-20%, to explore the sensitivity of the parameters in the mutation rate distribution. For each parameter, we plot the mean MCMC estimate and the 95% credible region (CR).

On the basis of these parameter estimates and assuming there are 475 STRs with 21% of the ascertained loci being calibrated, we simulate data according to section 4.3.1 with the rate of each calibrated locus estimated from 10,000 meioses. We begin by examining the effect on the parameters of varying the amount of information, i.e. varying the proportion of the loci that are calibrated from within the ascertained loci. In figure 4.3 we see this means the first segment is allowed to vary with respect to a fixed overall segment for markers 1 to k. The percentage of calibrated loci will vary as follows: 0%, 10%, 20% 40%, 60%, 80% and 100%.

For each parameter, we plot the MCMC mean and the 95% CR versus the percentage of calibrated loci within the ascertained loci, with the true value superimposed as a straight dashed line.

Next we carry out a brief misspecification assessment by varying the percentage of non-ascertained loci for a single data set. Due to the nature of the simulations, the percentage ascertained and the percentage calibrated will be random variables;

however, the percentage of non-ascertained loci will include 0%, 5%, 10%, true%, 20% 40%, 60%, 80% and 100− percentage of calibrated loci out of total loci.

The 95% CR along with the posterior mean MCMC estimate will be plotted for each parameter, of interest versus the percentage of non-ascertained loci, with the true parameter value superimposed as a dashed straight line. Also the average of the MCMC estimates of the five sets of mutation rates mentioned above will be plotted against the percentage of non-ascertained loci.

4.4 Results

4.4.1 Real Data

To begin with, we compare the empirical value of the mutation rates of the calibrated loci to those obtained from the Bayesian analysis.

In table 4.1, we see that the MCMC means are quite close to the empirical rates, particularly when mutations have occurred and there is a large number of meioses. In addition, for markers where no mutations have occurred, the Bayesian estimate correlates negatively with the number of meioses. The results are shown graphically for the mutation rates that are empirically variable in figure 4.4 whilst those that are empirically non-variable are in figure 4.5.

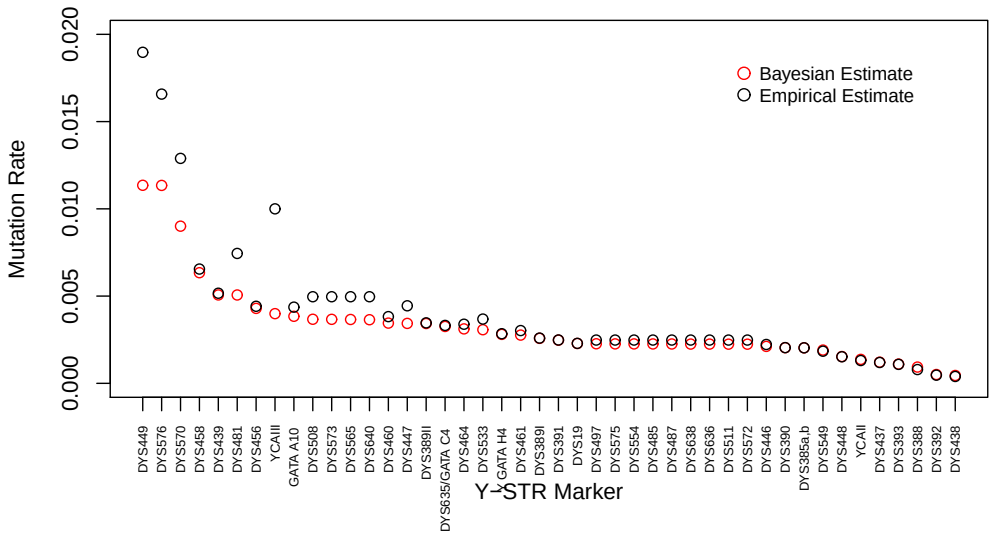

FIGURE 4.4: Empirically variable mutation rates: Bayesian and empirical estimates

TABLE 4.1: Empirical and Bayesian mutation rate estimates

STR Marker	Mutations	Meioses	Empirical Rate	MCMC Mean
DXYS156	0	1027	0.000000	0.000587
DYS19	57	24890	0.002290	0.002285
DYS385a,b	55	27019	0.002037	0.002035
DYS388	2	2533	0.000790	0.000932
DYS389I	36	13876	0.002594	0.002578
DYS389II	48	13847	0.003466	0.003429
DYS390	49	23950	0.002046	0.002044
DYS391	40	16094	0.002485	0.002474
DYS392	7	14955	0.000468	0.000505
DYS393	15	13801	0.001087	0.001105
DYS426	0	139	0.000000	0.001553
DYS435	0	161	0.000000	0.001464
DYS437	12	10050	0.001194	0.001214
DYS438	4	10151	0.000394	0.000451
DYS439	53	10264	0.005164	0.005066
DYS443	0	80	0.000000	0.001771
DYS444	0	80	0.000000	0.001754
DYS446	1	449	0.002227	0.002118
DYS447	2	450	0.004444	0.003435
DYS448	10	6567	0.001523	0.001536
DYS449	7	369	0.018970	0.011350
DYS456	29	6567	0.004416	0.004295
DYS458	43	6565	0.006550	0.006340
DYS460	5	1308	0.003823	0.003452
DYS461	3	992	0.003024	0.002765
DYS464	5	1476	0.003388	0.003120
DYS472	0	403	0.000000	0.001012
DYS476	0	403	0.000000	0.001024
DYS480	0	403	0.000000	0.001020
DYS481	3	403	0.007444	0.005064
DYS485	1	403	0.002481	0.002258
DYS487	1	403	0.002481	0.002256
DYS488	0	403	0.000000	0.001020
DYS490	0	403	0.000000	0.001011
DYS491	0	403	0.000000	0.000996
DYS492	0	403	0.000000	0.001012
DYS494	0	403	0.000000	0.001012
DYS495	0	403	0.000000	0.001013
DYS497	1	403	0.002481	0.002268
DYS505	0	403	0.000000	0.001018
DYS508	2	403	0.004963	0.003673
DYS511	1	403	0.002481	0.002245
DYS520	0	80	0.000000	0.001770
DYS522	0	543	0.000000	0.000879
DYS525	0	403	0.000000	0.001023
DYS530	0	403	0.000000	0.001003
DYS531	0	483	0.000000	0.000912
DYS533	2	543	0.003683	0.003073
DYS537	0	403	0.000000	0.001011
DYS540	0	403	0.000000	0.001013
DYS549	1	543	0.001842	0.001907
DYS554	1	403	0.002481	0.002260
DYS556	0	403	0.000000	0.001005
DYS557	0	80	0.000000	0.001744
DYS565	2	403	0.004963	0.003660
DYS567	0	403	0.000000	0.001027
DYS568	0	403	0.000000	0.001021
DYS569	0	403	0.000000	0.001015
DYS570	7	543	0.012891	0.009006
DYS572	1	403	0.002481	0.002244
DYS573	2	403	0.004963	0.003668
DYS575	1	403	0.002481	0.002261
DYS576	9	543	0.016575	0.011342
DYS578	0	403	0.000000	0.001006
DYS579	0	403	0.000000	0.001031
DYS580	0	403	0.000000	0.001005
DYS583	0	403	0.000000	0.001040
DYS589	0	403	0.000000	0.001008
DYS590	0	403	0.000000	0.001030
DYS594	0	403	0.000000	0.001006
DYS617	0	403	0.000000	0.001045
DYS618	0	403	0.000000	0.001024
DYS622	0	80	0.000000	0.001783
DYS630	0	80	0.000000	0.001764
DYS635/GATA C4	25	7533	0.003319	0.003260
DYS636	1	403	0.002481	0.002250
DYS638	1	403	0.002481	0.002251
DYS640	2	403	0.004963	0.003646
DYS641	0	403	0.000000	0.001023
DYS643	0	543	0.000000	0.000866
DYS709	0	80	0.000000	0.001738
GATA A10	5	1145	0.004367	0.003843
Y GATA H4	22	7738	0.002843	0.002806
YCAI	0	150	0.000000	0.001487
YCAII	3	2296	0.001307	0.001373
YCAIII	1	100	0.010000	0.003994

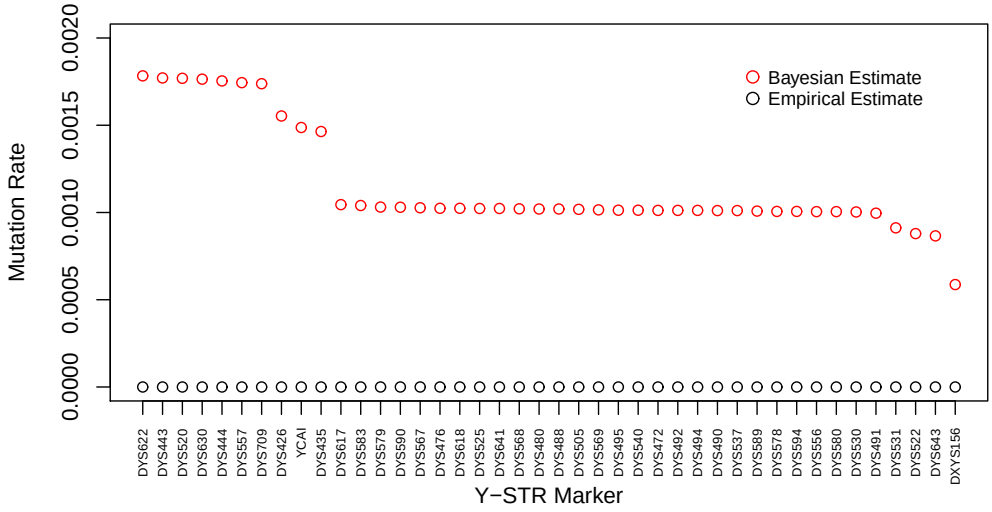

FIGURE 4.5: Empirically non-variable mutation rates: Bayesian and empirical estimates

In figure 4.6 the first column shows the output for the first locus, which is both ascertained and calibrated. The chain for the rate appears to hit off the boundary at zero, though otherwise mixes well. The log standard deviation of the proposal distribution decreases through the run of the chain until it stabilizes. Overall the mean of the calibrated loci rate estimates is 0.00223 which is comparable to the mean for all the ascertained loci at 0.00222. For the non-ascertained loci, the mean of the estimated rate is 7.33×10^{-5}, whilst the mean across all the loci is 0.00191. For the last locus, the second column in figure 4.6, a non-ascertained locus, the MCMC chain appears mix less well, reducing the spread as the run progresses. The last row shows that the posterior distribution is very right-skewed. For

TABLE 4.2: Real data Bayesian parameter estimates

Parameter	Mean	St. Dev.	95% Credible Region	
			Lower	Upper
L	3.01	1.49	1.03	6.75
N_e	4262	1828	939	7943
α	0.5872	0.1711	0.3910	0.9619
β	0.003483	0.001017	0.001796	0.005843

the total branch length L, the posterior mean is 3.01 (table 4.2) with a 95% CR of 1.03-6.75. The MCMC chain (fig. 4.6, third column) mixes well and the log standard deviation of the proposal distribution quickly stabilizes. Its posterior distribution of updates is also right-skewed. Next we examine the output for N_e (fig. 4.6, fourth column), where we see the chain also mixes well. This posterior

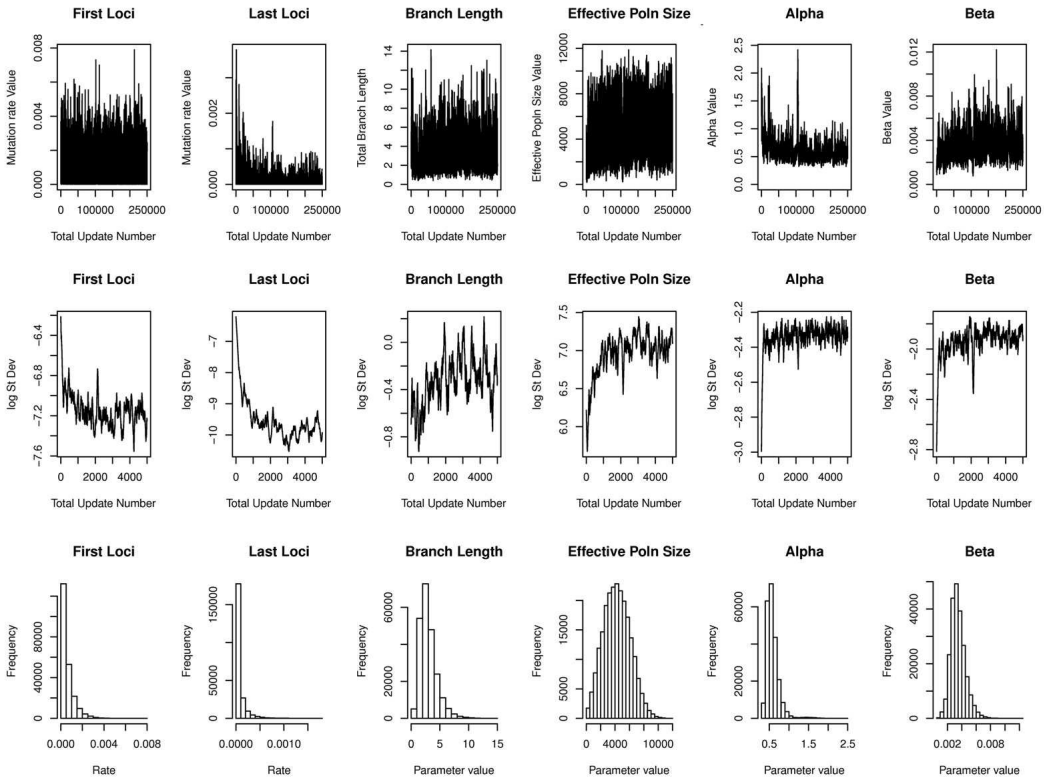

FIGURE 4.6: MCMC graphical output using real data

for N_e shows an almost symmetric distribution with mean value 4262 (95% CR 939-7943, table 4.2).

The diagnostic and posterior summaries for α and β are summarised in figure 4.6 (last two columns) and table 4.2. Again mixing appears satisfactory.

Next we examine the results from varying the percentage of non-ascertained loci. In figure 4.7a, we plot the posterior mean α and the 95% CR across each percentage of non-ascertained loci. We see that, as the latter increases, the mean rapidly decreases, thereafter increasing very slightly till the percentage of non-ascertained loci reaches 81.9%. Conversely, $\hat{\beta}$ increases until the percentage of non-ascertained is 20%, thereafter gradually decreasing (fig. 4.7b). In terms of the mean and variance of the gamma distribution from which the mutation rates are drawn, we see, in figure 4.8a, that the mean appears to decrease at an almost constant rate as the percentage of the non-ascertained loci increases. On the other hand, the variance of the mutation rate distribution reaches its peak at 14.4%, though the CR is widest at this point (fig. 4.8b).

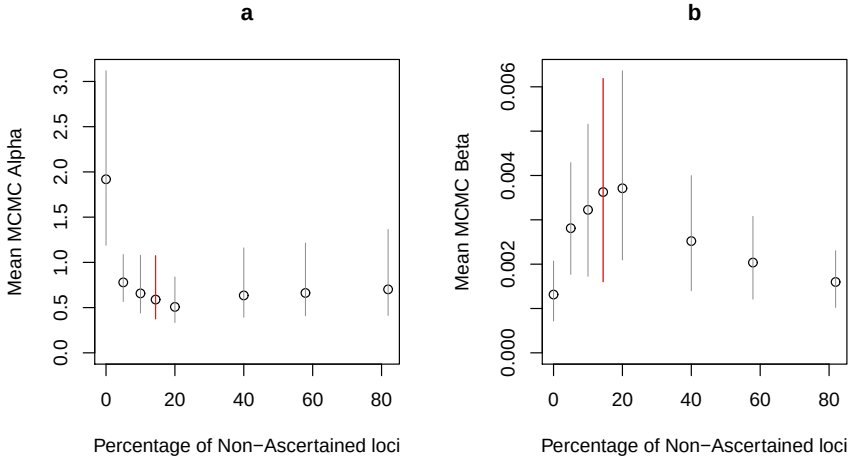

FIGURE 4.7: Posterior mean with credible region vs. percentage of non-ascertained loci: a. α b. β
(Gray line - credible region, red line - credible region for true percentage of non-ascertained loci)

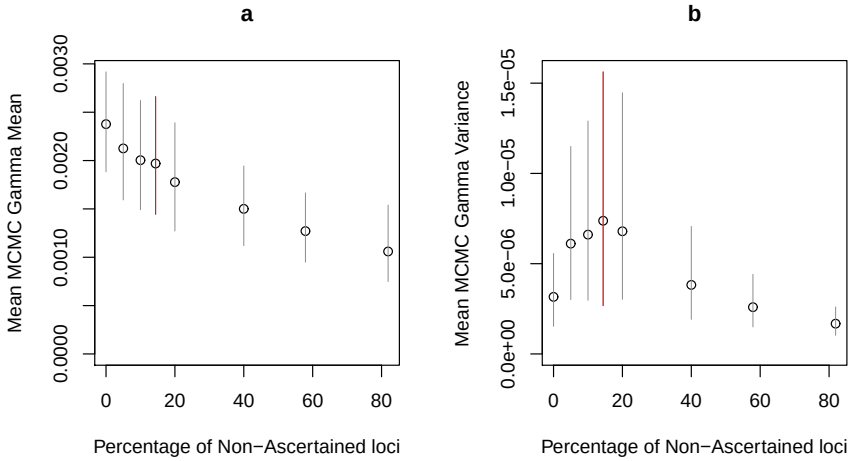

FIGURE 4.8: Posterior mean with credible region vs. percentage of non-ascertained loci: a. gamma mean b. gamma variance
(Gray line - credible region, red line - credible region for true percentage of non-ascertained loci)

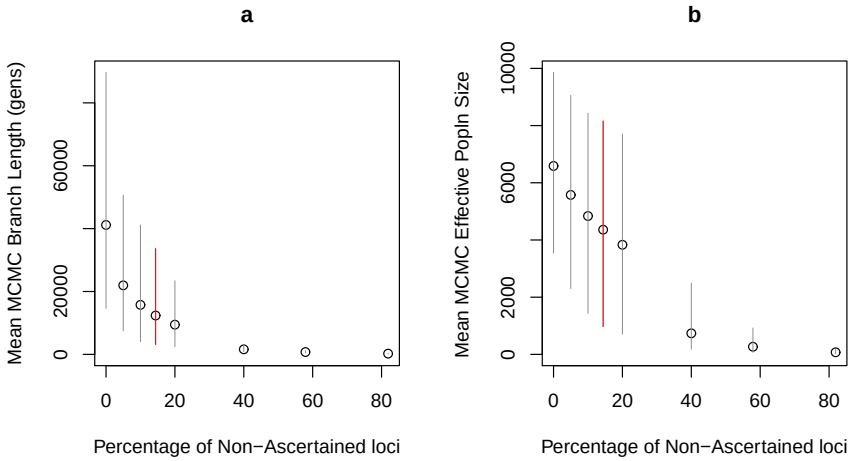

FIGURE 4.9: Posterior mean with credible region vs. percentage of non-ascertained loci: a. L b. N_e
(Gray line - credible region, red line - credible region for true percentage of non-ascertained loci)

For L, the posterior mean decreases quickly before reaching a plateau when the percentage of non-ascertained is 40%. The CR is also short by this point (fig. 4.9a). A similar picture emerges for the effective population size, N_e, which decreases as the percentage increases. In addition, the range of the CR is fairly constant until 20%, thereafter decreasing to very little by 81.9% (fig. 4.9b).

As a summary, we plot the mean of the posterior means across the sets of mutation rates outlined above in figure 4.10.

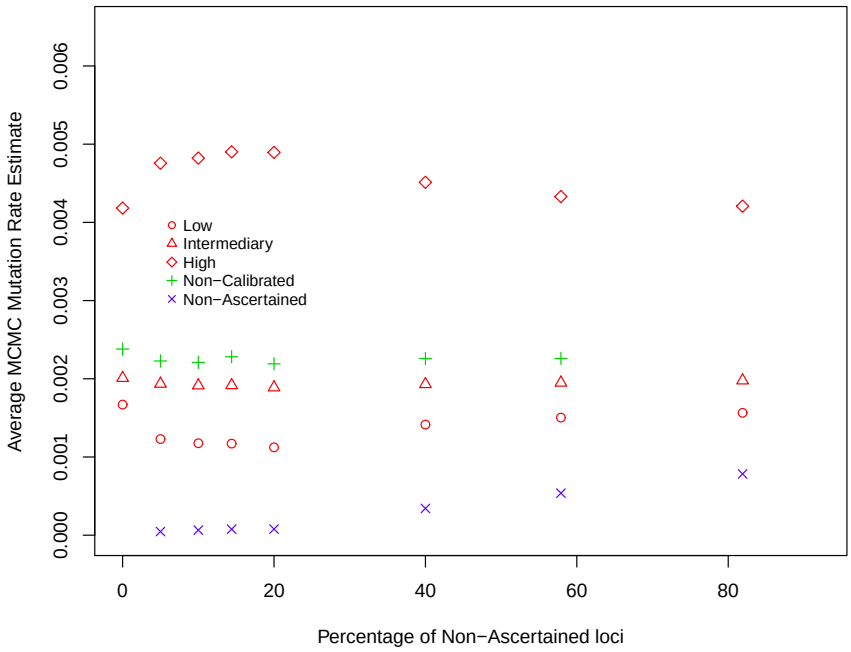

FIGURE 4.10: Average posterior mean mutation rate vs. percentage of non-ascertained loci

For the low calibrated rates (○), we find that the mean of the posterior means decreases from around 0.0017 when the percentage of non-ascertained loci is zero until to ≈ 0.001 at 20% thereafter increasing gradually to just over 0.0015. For the intermediate calibrated mutation rates (△), the mean is roughly constant. For the high calibrated rates (◇), we see that the mean increases quickly from ≈ 0.0041 when there are no non-ascertained loci to just below 0.005. Subsequently, the mean gradually reduces to about 0.004, when the percentage of non-ascertained loci reaches 81.9%. The non-calibrated loci (+), always have a mean rate around 0.0022. Contrast this with the mean of the mutation rates for the non-ascertained loci (×), which are only constant till 20%, thereafter increasing gradually to a rate of ≈ 0.0008.

We now focus on the effect of varying the proportion of non-ascertained loci in the range 10-20%. A summary of this is given in table 4.3.

TABLE 4.3: Regression of posterior mean against percentage of non-ascertained loci

Parameter	Intercept	Slope	P-Value for Slope
α	0.8041	-1.6204	0.0017
β	0.00279	0.00523	0.0038
Mean of Gamma	0.002264	-0.002445	0.0000
Variance of Gamma	0.00000679	0.00000118	0.4950
L (generations)	21464	-59535	0.0021
N_e	5701	-8531	0.0080

For α we have the results in figure 4.11a. There is a negative linear relationship between the mean of α and the percentage of non-ascertained loci. In fact we find that the slope of -1.6204 is found to be statistically significant (table 4.3). For β, in figure 4.11b we see that there is a positive linear relationship between the posterior mean and percentage. Here too we find that the slope is statistically significant (table 4.3). For the mean and the variance of the gamma, we see that, whilst the mean has a negative linear relationship, the variance appears roughly constant (fig. 4.12). For the former, the slope estimated as -0.00214 is significant. On the other hand, the latter has a non-significant. The estimates of the final parameters, L and N_e, both show a negative linear relationship with the percentage of non-ascertained loci (fig. 4.13). In addition the fitted lines both have significant slopes.

4.4.2 Simulated Data Analysis

The results in this section are based on simulating data based on the estimates of the parameters from the real data, i.e.

- the shape parameter, $\hat{\alpha} = 0.5871$;
- the scale parameter, $\hat{\beta} = 0.00348$;
- the effective population size, $\hat{N}_e = 4262$;
- the total branch length, $\hat{L} = 3.01$;
- the number of loci, $n = 475$, and
- the percentage of calibrated loci within the ascertained loci, 21.1%.

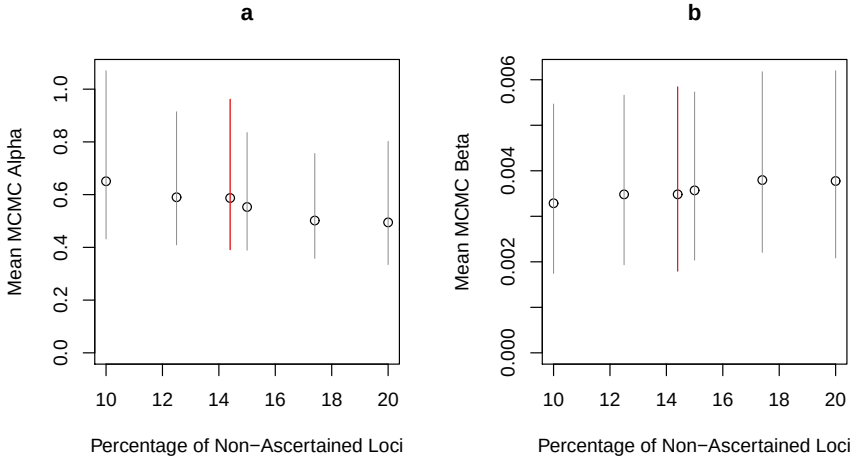

FIGURE 4.11: Posterior mean with credible region vs. percentage of non-ascertained loci: a. α b. β
(Gray line - credible region, red line - credible region for true percentage of non-ascertained loci)

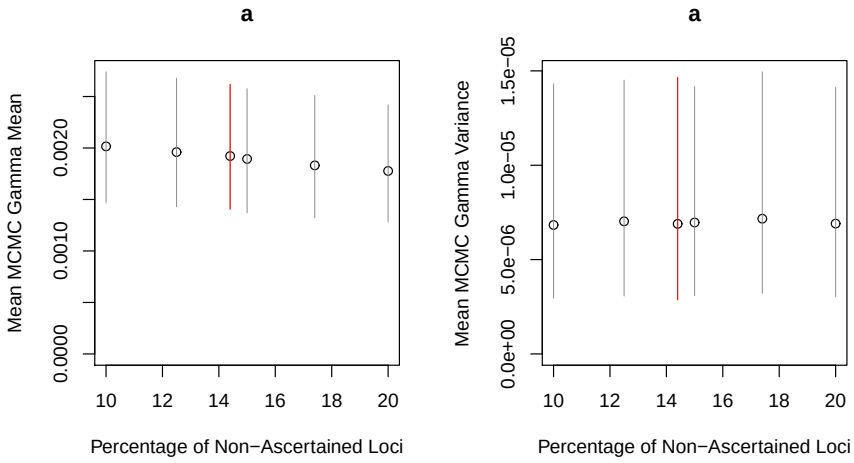

FIGURE 4.12: Posterior mean with credible region vs. percentage of non-ascertained loci: a. gamma mean b. gamma variance
(Gray line - credible region, red line - credible region for true percentage of non-ascertained loci)

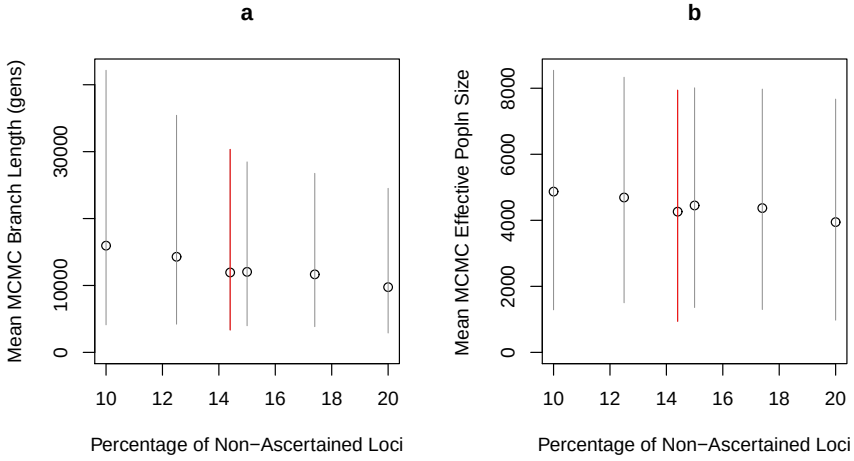

FIGURE 4.13: Posterior mean with credible region vs. percentage of non-ascertained loci: a. L (generations) b. N_e
(Gray line - credible region, red line - credible region for true percentage of non-ascertained loci)

4.4.2.1 Varying the Information

In this section, we examine the effect on the parameters of varying the proportion of the loci that are calibrated from within the ascertained loci. The simulated data set resulted in 414 ascertained loci so the percentage of calibrated loci varied as follows: 0% (0 loci), 10% (41), 20% (83), 40% (166), 60% (248), 80% (331), 100% (414).

The estimates of α are shown in figure 4.14a, where we see that, when there are no calibrated loci, $\hat{\alpha}$ is close to the true value although the CR is widest at this point. As the percentage of calibrated loci increases, the CR narrows but the point estimates lie below the true value. In this case, $\hat{\alpha}$ is underestimating. For the estimates of β, we see a converse result (fig. 4.14b). The initial estimate at 0% is close to the true value. The CR is very wide and unsymmetrical. The remaining estimates all lie above the true value (dashed line), but their CRs appear to just include the true value. In addition, the overestimation decreases as the percentage of calibrated loci increases. Next we check the estimate of the mean of the gamma distribution from which the mutation rates are drawn (fig. 4.15a). When there are no calibrated loci the mean is underestimated albeit with a wide CR. As the percentage of calibrated loci increases, the posterior mean overestimates, though the CRs include the true gamma mean. The range of the CRs decreases with increasing proportion of calibrated loci. A similar result is produced for the estimate of the variance of the gamma distribution, although in this case all the posterior means are overestimates (fig. 4.15b). As the percentage of calibrated loci increases, the estimates lie closer to the true value (dashed line) and the CRs become narrower. The results for L are shown in figure 4.16a. Here the overestimation reduces with the increasing number of calibrated loci. All the CRs include the true value (dashed line). Similarly for $\hat{N}_e$, there is a decrease in the posterior mean as the percentage of calibrated loci increases. The range of the CRs remains fairly constant throughout (fig. 4.16b) and include the true value (dashed line).

4.4.2.2 Misspecification

We now investigate the effect of misspecification of the percentage of non-ascertained loci on the parameters of the mutation rate distribution model. Here the percentage of non-ascertained loci will vary as follows: 0%, 5%, 10%, 10.5% (true%), 20%, 40%, 60% and 81.3% (100- percentage of calibrated out of total loci%).

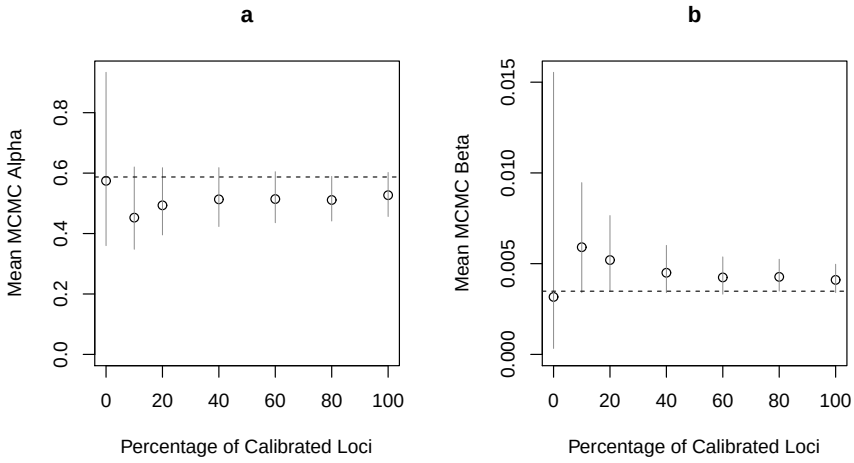

FIGURE 4.14: Posterior mean with credible region (gray line) vs. percentage of calibrated loci: a. α b. β

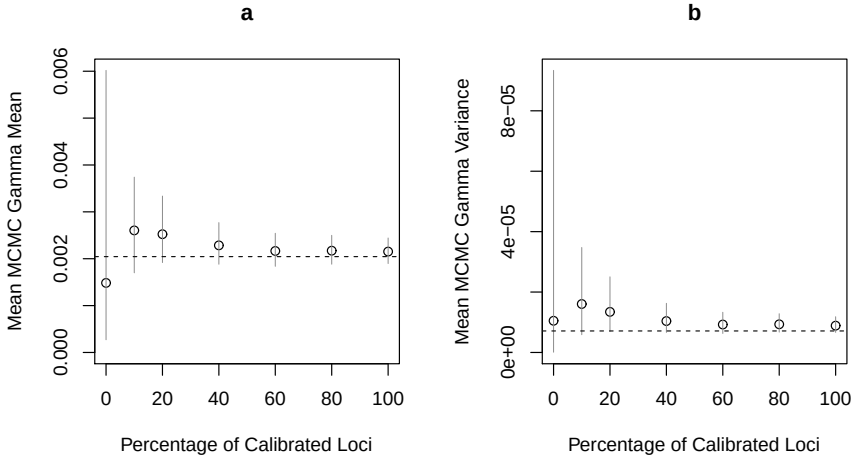

FIGURE 4.15: Posterior mean with credible region (gray line) vs. percentage of calibrated loci: a. gamma mean b. gamma variance

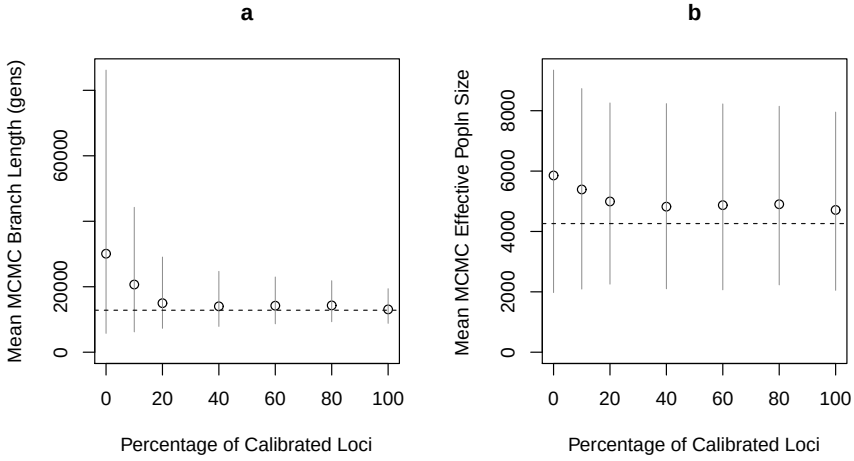

FIGURE 4.16: Posterior mean with credible region (gray line) vs. percentage of calibrated loci: a. L b. N_e

For the mutation rate distribution shape and scale parameters, α and β, respectively, the results are shown in figure 4.17. $\hat{\alpha}$ initially decreases sharply as the percentage of non-ascertained loci increases, thereafter approaching an asymptote. In addition, the range of the CR reduces. Comparing the points to the dashed line indicating the true α shows that near the true percentage of non-ascertained loci α is well estimated. However, at either extreme it is over- or underestimated. Indeed the true α is well contained in the CR for 10% and 10.5% and only just contained for 5% and 20%. For the remaining cases, the CR does not contain the true value. For $\hat{\beta}$ we find that it appears to have a maximum of $\sim$0.004 near 20%. This is slightly higher than the true value. However, the CR does contain this value. The CRs for the 0% and 81.3% estimates do not contain the true value, whilst the others are wide enough to do so. When the percentage of non-ascertained loci is close to the true value(10.5%), the posterior mean is close to the true β. Reparameterising in terms of the mean and variance of the gamma distribution produces the results in figure 4.18. Again the true mean and variance are indicated by the dashed black line. For the mean, we see a reduction in the estimate when the percentage of non-ascertained loci increases. The range of the CRs reduces and for the first five points the CRs include the true mean. However, for the last three points, this is not the case. At these points, the posterior means and CRs all underestimate the true gamma mean (fig. 4.18a). For the gamma variance, we see a parabolic form for the estimates as the percentage of non-ascertained loci increases. The peak lies between 10% and 20%, where the variance is slightly overestimated compared to the true value (dashed line). The remaining points are all underestimates to a varying extent, though the CR for each point until 40% contains the true variance. For 60% and 80% the CRs do not contain the true value (fig. 4.18b). The branch length estimate of L (fig. 4.19a), declines as the percentage of non-ascertained loci increases. It intersects the true L (dashed line) at $\sim$10%. The CR is very wide initially but it decreases to a negligible amount by 40%. The CR includes the true value for the points between 5-20%. The results for N_e are shown in figure 4.19b. It too shows a gradual decrease in the estimates as the percentage of non-ascertained loci increases. Whilst the CRs are very wide initially, the spread decreases to a very small amount by 81.3%. Between 5-40% the CRs contain the true value (dashed line), otherwise they do not.

The last set of results discussed in this section are the average estimated mutation rates in the various classes enumerated above. Superimposed onto the plot is the true average of the various sets of loci as shown in the legend in figure 4.20.

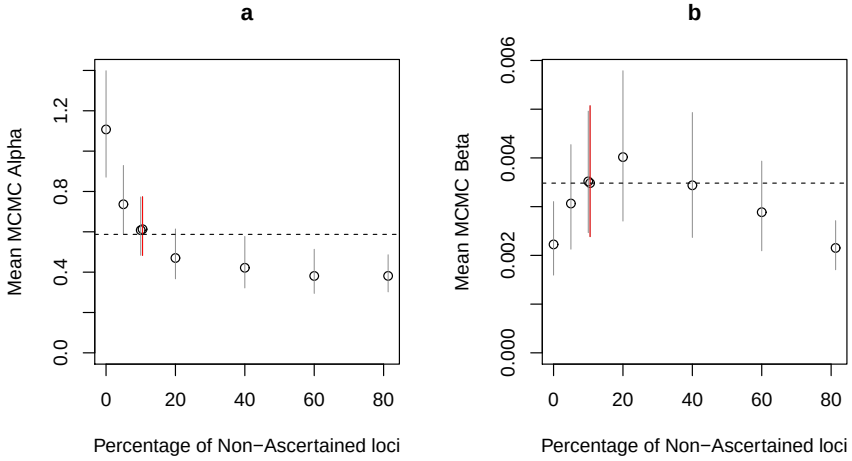

FIGURE 4.17: Posterior mean with credible region vs. percentage of non-ascertained loci: a. α b. β
(Gray line - credible region, red line - credible region for true percentage of non-ascertained loci)

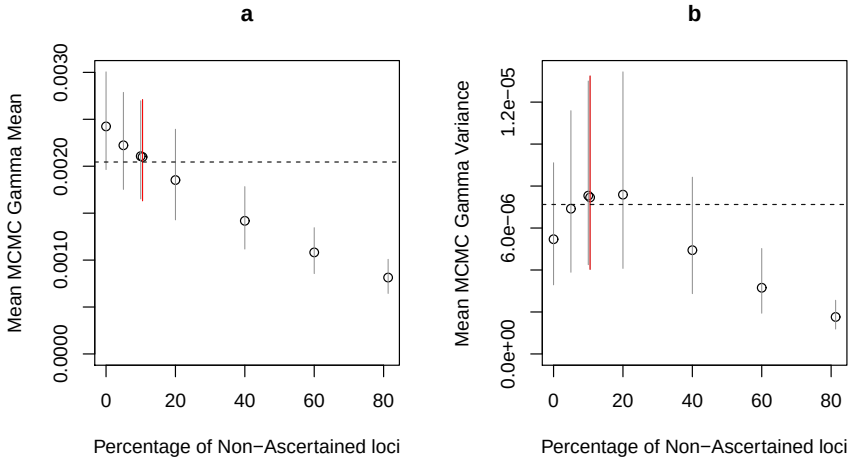

FIGURE 4.18: Posterior mean with credible region vs. percentage of non-ascertained loci: a. gamma mean b. gamma variance
(Gray line - credible region, red line - credible region for true percentage of non-ascertained loci)

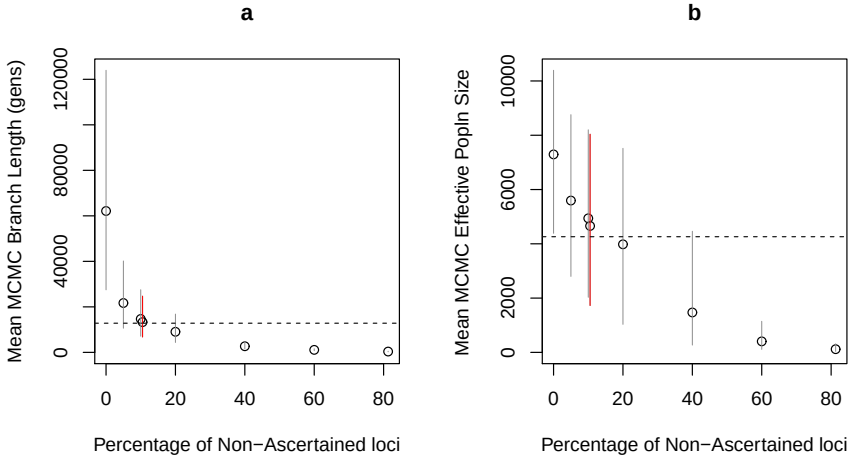

FIGURE 4.19: Posterior mean with credible region vs. percentage of non-ascertained loci: a. L b. N_e
(Gray line - credible region, red line - credible region for true percentage of non-ascertained loci)

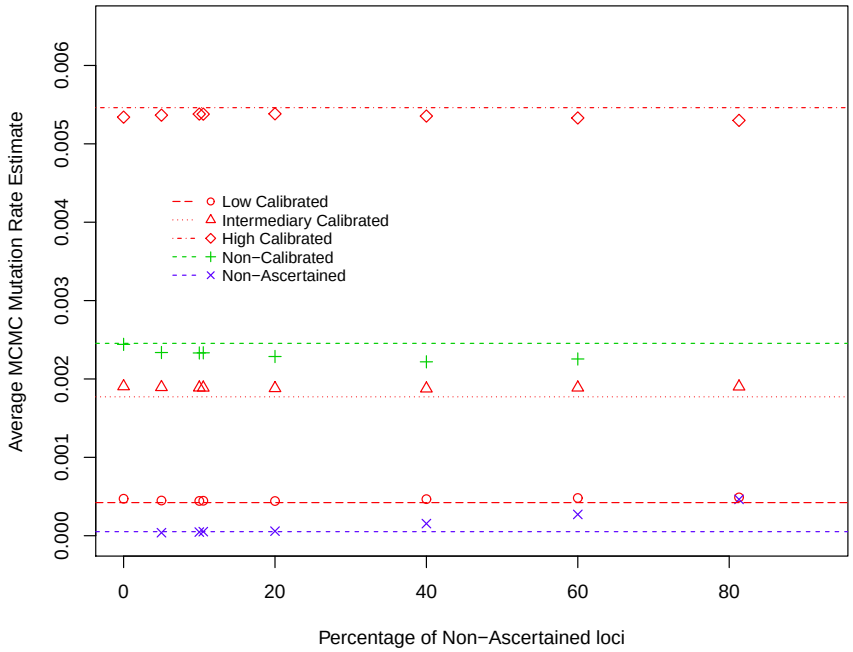

FIGURE 4.20: Simulated data: average posterior mean mutation rate vs. percentage of non-ascertained loci

The low calibrated mutation rates (○) all lie close to the true value (red dashed line). For the intermediate calibrated rates (△), the estimated averages lie parallel to the true value (red dotted line). Hence they overestimate by approximately the same amount. For the high calibrated rates (◇), the average estimates all lie below the red dotted-dashed line, thus slightly underestimating. Next we have the non-calibrated loci (+), which all lie below the green dashed line and thus are underestimated. Conversely the average estimates for the non-ascertained loci (×) all lie above the blue dashed line, which indicates their true average, and they overestimate by an amount which increases as the percentage of non-ascertained loci increases.

4.5 Discussion

One of the key differences between the real and simulated analyses carried out in the previous two sections was the average estimates of the five sets of mutation rates. Comparing the results in figures 4.10 and 4.20 we see that, although the intermediate calibrated (△), the non calibrated (+) and the non-ascertained (×)

have very similar behaviour, this is not so for the high (◇) and low (○) calibrated mutation rates average.

Since the estimates of the mutation rates are directly affected by the number of meioses used to calibrate, we examine this aspect. In the simulated dataset we fixed the number of meioses to be 10,000 across the calibrated loci. Thus being a rather large number implies the rates of low/intermediate and high calibrated loci are rather accurately estimated. However this doesn't correspond well to the number of meioses used to estimate Y-STR rates in the intermediary mutation rate review. In figure 4.21, we see the distribution of the number of meioses per locus in this mutation rate review. The bulk of the loci have fewer than 10,000 meioses. Indeed the mean lies at 2880, though given the skewed distribution of the data the mode of 403 emphasises the discrepancy between the real and simulated data.

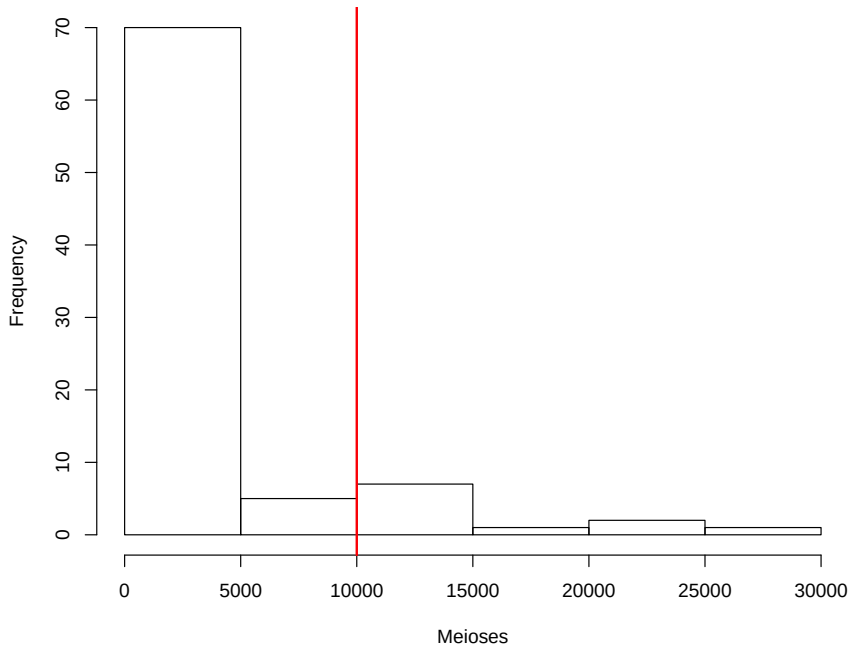

FIGURE 4.21: Intermediary mutation rate review: histogram of the number of meioses per locus
(Red line - 10,000 meioses)

Given this observation, we chose to reanalyse the data by sampling from the real dataset's meiosis numbers when calibrating the loci. That is, the m_i in equation 4.12 are sampled with replacement for the loci that are calibrated in the simulation from those m_i in the real data. As before we simulated a single dataset and

analysed the data varying the percentage of non-ascertained loci, i.e. 0%, 5%, 8.6% (true %), 10%, 20%, 40%, 60% and 91.4% (100 - percentage of calibrated out of total loci%).

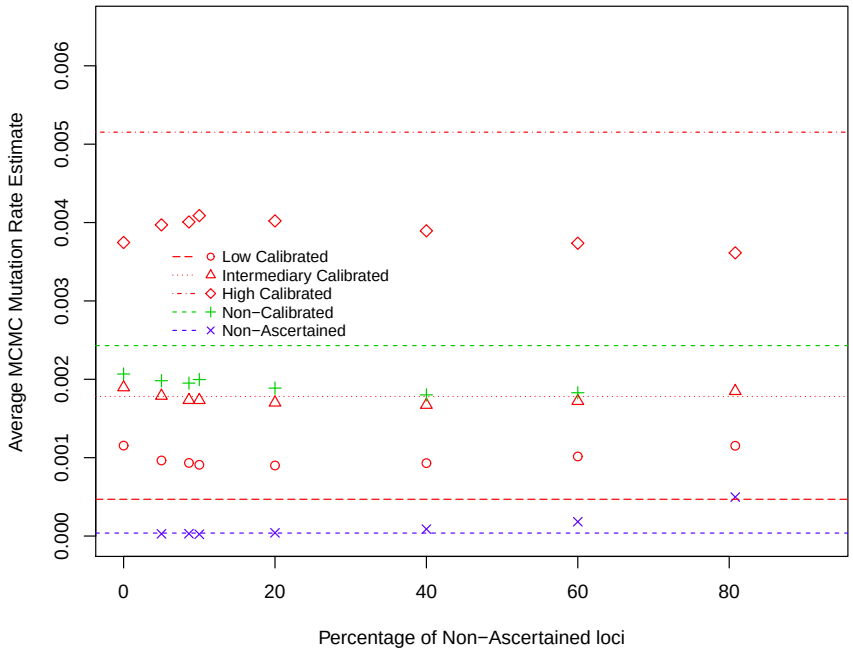

FIGURE 4.22: Simulated data: average posterior mean mutation rate vs. percentage of non-ascertained loci

The results for the average mutation rates in the five categories along with their true averages are shown in figure 4.22. The low calibrated loci (○, red dashed line) average is overestimated by almost 0.005 though the overestimation seems least when the percentage of non-ascertained loci is closest to the true, i.e. 8.6%. For the intermediate calibrated loci (△, red dotted line), the averages lie close to their true value, though there is a slight underestimation from 20-60%. Contrast this with the high calibrated loci (◇, red dotted-dashed line), which are substantially underestimated ∼0.001. Again the underestimation is least when near the true percentage of 8.6%. The non-calibrated mutation rates' averages are also underestimates (+, green dashed line), which increase as the percentage of non-ascertained loci increases. The averages are roughly 0.0005 less than the true value. Theses estimated average of the non-ascertained loci (×) lie close to the true value (blue dashed line) initially but increase as the percentage of non-ascertained loci increases, thus gradually increasing the amount by which it overestimates.

Thus, using simulated data where the calibration of the mutation rates is based on more realistic numbers of meioses produces results much more similar to those produced for the real data. Yet compared to the simulated data, based on 10,000 meioses (fig. 4.20), a drastic underestimation of the high calibrated rates (◇) and the non-calibrated rates as well as overestimation in the case of the low calibrated mutation rates occurred. This was clearly down to the effect of the number of meioses, since the real values extended beyond the 10,000 used for the simulated data.

Since it did not seem natural for any calibrated locus, particularly one with a low true mutation rate, to have a very high number of meioses sampled, we developed a multi-stage process to obtain estimates of the calibrated mutation rates. Hence we initially calibrate selected loci using a few meioses, thereafter calibrating the most variable again but with a greater number of meioses and so on. In fact we based the calibration on the number of times we wished to repeat the process. Suppose we define n_c to be the number of calibration steps, then we specified the number of loci to calibrate according the current step, i.e. at the i^{th} step, the number of loci to be calibrated was the following fraction of the total number specified to be calibrated:

$$\frac{n_c - i + 1}{n_c}.$$

For example, if we wished to apply three calibration steps, the first step would calibrate all the proportion of the ascertained loci chosen to be calibrated. The next step ($i = 2$) would recalibrated only $2/3$ of the fastest loci (based on the mutations and meioses so far). In the final calibration loop ($i = 3$) only $1/3$ would be recalibrated. In addition, the number of meioses varies according to i:

- $i = 1$: meioses=[80-500]
- $i > 1$ and $i < n_c$: meioses=[500-2500]
- $i = n_c$: meioses=[5000-27000]

Based on this process we again simulated data according to the real data estimates of α, β, L, N_e and analysed it for a range of values of the percentage of non-ascertained loci, i.e. 0%, 5%, 9.9% (true %), 10%, 20%, 40%, 60% and 81.1% (100− percentage of calibrated out of total loci%).

A plot of the average of the posterior means for the five classes of mutation rates are shown in figure 4.23. For the low calibrated rates (○, red dashed line), there is a minimum in the averages at 10. However the values are all overestimates. On the other hand, the intermediate calibrated mutation rate average (△, red dotted line) are quite well estimated. The averages of the high calibrated loci (◇, red dotted-dashed line) lie straight line parallel to their true and underestimate by over 0.001. The non calibrated mutation rates averages (+) also have a minimum around the true percentage of the non-ascertained loci which lies just above its true value (green dashed line). The averages of the non-ascertained loci (×, blue dashed line) gradually increase in the amount by which they overestimate.

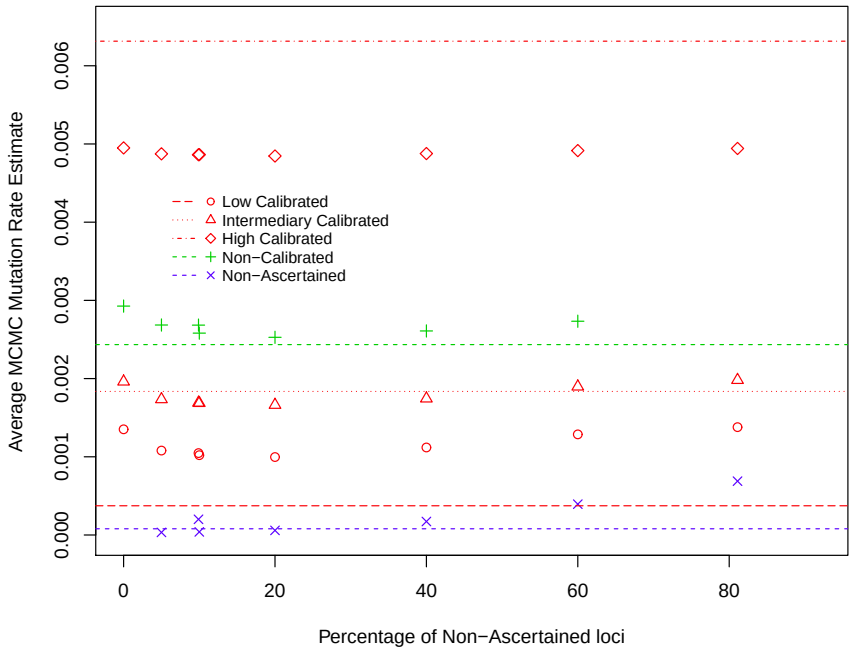

FIGURE 4.23: Simulated data: average posterior mean mutation rate vs. percentage of non-ascertained loci

The pattern in all the sets of mutation rate averages is rather similar to the real data (fig. 4.10), except for the high calibrated mutation rate (◇), which is consistently underestimated in this simulation. Furthermore, there is clearer demarcation between averages of the non-calibrated (+) and intermediary calibrated loci (△) when using the multi-stage calibration than by simply sampling from the realistic number of meioses (fig. 4.22).

4.6 Conclusions

Having established a framework for modelling the mutation mechanisms involved in Y-STRs, we note that accounting for variable proportions of increase and decrease mutations is not necessary.

As such we focussed on a model for the mutation rate distribution which incorporated both ascertainment (through the introduction of the parameters N_e and L) and calibration of Y-STRs markers. Application to the intermediary mutation rate review data provided estimates of the parameters of the gamma mutation rate distribution, along with the total branch length and effective population size. In addition, we found that when the percentage of non-ascertained loci varies in a narrow range (10-20%) the estimates of the parameters are fairly robust. However, varying the percentage of non-ascertained more drastically (0-80%) resulted in considerable variation in the estimates of each parameter.

A simulation study showed that, whilst varying the percentage of calibrated loci did not have much effect on the estimates of the parameters unless the percentage fell below 20, varying the percentage of non-ascertained loci did. In addition, we found that simply using a fixed and rather large number of meioses (10,000) to simulate the calibration of Y-STR mutation rates did not reflect real data. Consequently, we developed a multi-stage calibration process by which mutation rates may be simulated.

Chapter 5

Modelling TMRCA

5.1 Introduction

In this chapter we will incorporate the methodology developed in Chapter 4, with the model of estimating the time to the most recent common ancestor (TMRCA), t, as outlined by Walsh (2001). The primary aim is to produce an estimate, $\hat{t}$, of t for two males who are not known to be closely related, on the basis of the absolute difference in the number of STR repeats across a number of loci (fig. 5.1).

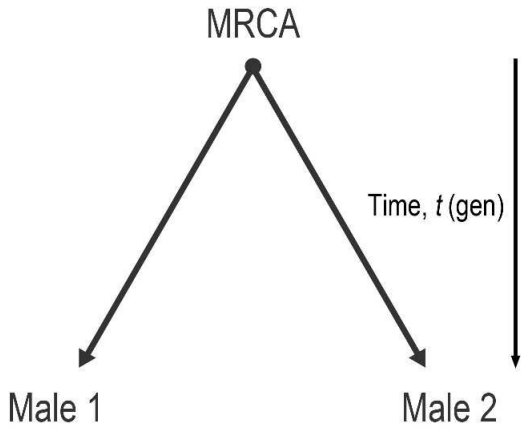

FIGURE 5.1: Estimation of the time to the MRCA

Furthermore, we will assess how factors such as the number of typed STRs and the rate at which the STRs mutate affect $\hat{t}$, as well as the other parameters. This will involve a simulation study.

5.2 Materials and Methods

5.2.1 Simulation Study

In this section, the process by which we simulate STR repeats for two contemporaneous males will be outlined. This process firstly involves simulating appropriate mutation rates as detailed in section 4.3.1. On the basis of these rates, STR repeats from two males will be simulated in accordance with Walsh (2001).

Importantly, we must define various sets of mutation rates for the total number of STRs, due to the nature of the mutation model we developed in chapter 4. In figure 5.2 we see that there is a total of n loci. A subset, of size k, of these are the ascertained loci, i.e. these are markers found to be variable at the population level, whilst $n - k$ are non-ascertained loci. Within the ascertained loci, we have c loci which are calibrated, i.e. for these loci, mutation rates are empirically estimated, typically by counting the number of mismatches between a large number of father-son pairs . From within this set, s loci will be typed in the two males whose TMRCA is being estimated. Thus we have the following constraints: $k \leq n$, $c \leq k$, and $s \leq c$. (In principal we could have the constraint $s \leq k$ instead of the last, since potentially we could use markers that are known to be variable at the population level but which have not been calibrated. However, this will not be the case here.)

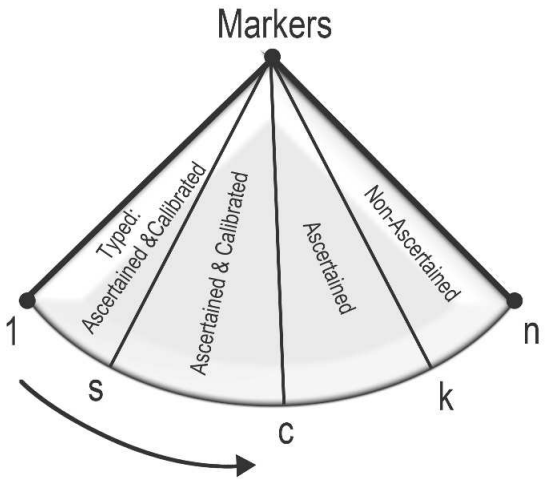

FIGURE 5.2: Categories of markers in the TMRCA model

In addition, we also have the sample size used to ascertain the loci, n_{asc}, as well as the number of meioses, m_i, across each of the calibrated markers $(1 = 1, \ldots, c)$.

The following notation will be used in this chapter:

- μ, the, per locus per generation, mutation rate. This will usually be indexed by i referring to the i^{th} locus;
- α, the shape parameter in the gamma distribution used to describe the underlying mutation rate distribution;
- β, the scale parameter in the gamma distribution used to describe the underlying mutation rate distribution;
- N_e, the effective size of the population in which the loci were ascertained;
- L, the total branch length, in units of N_e generations, of the sample in which the loci were ascertained;
- R, the results from the ascertainment process, where $R = 1$ if ascertained and $R = 0$ otherwise. This will usually be indexed by i;
- r, the per-locus mutation counts in the calibration process, usually indexed bi i;
- n_+, the total number of increase mutations when moving in the direction Male 1 $\rightarrow$ MRCA $\rightarrow$ Male 2, at a locus;
- n_-, the total number of decrease mutations when moving in the direction Male 1 $\rightarrow$ MRCA $\rightarrow$ Male 2, at a locus;
- x_1, x_2, the number of repeats of an STR for male 1 and 2, respectively. These may be indexed $x_{j,i}$ where $j = 1, 2$ refers to which male and i refers to the locus;
- m, the number of meioses used to calibrate the mutation rate; this may also be indexed by i;
- n_c, the number of calibration steps used to provide empirical estimates of the calibrated mutation rates.

A summary of the steps to simulate the data is given below:

1.
$$\mu_i \sim Ga(\alpha, \beta) \quad \text{for} \quad i = 1, \ldots, n. \tag{5.1}$$

2.
$$L \sim \frac{n_{asc} - 1}{2} e^{-L/2} \left(1 - e^{-L/2}\right)^{n_{asc}-2}. \tag{5.2}$$

3.
$$N_e \sim N(\mu_{N_e}, \sigma^2_{N_e}). \tag{5.3}$$

4.
$$R_i \sim Bi(1, 1 - e^{-\mu_i L N_e}) \quad \text{for} \quad i = 1, \ldots, n. \tag{5.4}$$

5. For $j = 1, \ldots, n_c$ we have:

$$r_i \sim Bi(m_i, \mu_i) \quad \text{for} \quad i = 1, \ldots, \left(\tfrac{n_c - j + 1}{n_c}\right) c, \tag{5.5}$$

where $c \leq k$ and the number of meioses is chosen according to j, i.e.

- $j = 1$: $m_i \sim Un(80 - 500)$;
- $j > 1$ and $j < n_c$: $m_i \sim Un(500, 2500)$;
- $j = n_c$: $m_i \sim Un(5000 - 27000)$.

6. Compute the empirical mutation rate for calibrated loci: $\hat{\mu}_i = r_i / m_i$.
7. Generate the total number of mutations: $n_{+,i} + n_{-,i} \sim Po(2t\mu_i)$.
8. Generate the total number of increase mutations: $n_{+,i} \sim Bi(n_{+,i} + n_{-,i}, 0.5)$.
9. Compute the total number of decrease mutations: $n_{-,i} = (n_{+,i} + n_{-,i}) - n_{+,i}$.
10. Standardise male 1 as having zero STR repeats: $x_{1,i} = 0$.
11. Compute the number of STR repeats for male 2 relative to male 1: $x_{2,i} = n_{+,i} - n_{-,i}$.
12. Compute the observed data, i.e. the absolute difference in the number of STR repeats between males 1 and 2: $|x_{1,i} - x_{2,i}|$.

Data generated on the basis of this simulation will be used to estimate the parameters of interest by developing a fully probabilistic model using Bayesian methodology.

5.2.2 Bayesian Modelling

The joint distribution of the model parameters and the data is:

$$\begin{aligned} &P\left(\{\mu_i\}, t, L, N_e, \alpha, \beta, \{R_i\}, \{r_i\}, \{x_{1,i}\}, \{x_{2,i}\}, n_{asc}, \{m_i\}\right) \\ &= P\left(\{x_{1,i}\}, \{x_{2,i}\} | \{\mu_i\}, t\right) P\left(\{r_i\} | \{m_i\}, \{\mu_i\}, \{R_i\}\right) P\left(\{R_i\} | \{\mu_i\}, L, N_e\right) \\ &\quad \times P\left(L | n_{asc}\right) P\left(\mu_i | \alpha, \beta\right) P\left(\alpha\right) P\left(\beta\right) P\left(N_e\right) P\left(t\right), \end{aligned} \tag{5.6}$$

by repeated application of the product rule and exploiting the conditional independence in the model. Here let the sets of indices corresponding to the typed, calibrated, ascertained and non-ascertained loci, respectively, be $\mathfrak{T}$, $\mathfrak{C}$, $\mathfrak{A}$, $\mathfrak{N}$. Then

this probability is:

$$\begin{aligned} &P\left(\{x_{1,i}, x_{2,i} : i \in \mathfrak{T}\}|\{\mu_i\}, t\right) P\left(\{r_i : i \in \mathfrak{C}\}|\{m_i\}, \{\mu_i\}\right) \\ &\quad \times P\left(\{R_i = 1 : i \in \mathfrak{A}\}|\{\mu_i\}, L, N_e\right) P\left(\{R_i = 0 : i \in \mathfrak{N}\}|\{\mu_i\}, L, N_e\right) \\ &\quad \times P\left(\{\mu_i : i \in \mathfrak{T} \cup \mathfrak{C} \cup \mathfrak{A} \cup \mathfrak{N}\}|\alpha, \beta\right) P\left(L|n_{asc}\right) P\left(\alpha\right) P\left(\beta\right) P\left(N_e\right) P\left(t\right). \end{aligned} \tag{5.7}$$

Using the SMM outlined in section 3.1 we have:

$$P\left(\{x_{1,i}, x_{2,i} : i \in \mathfrak{T}\}|\{\mu_i\}, t\right) \propto \prod_{i=1}^{s} e^{-2t\mu_i} I_{|x_{1,i}-x_{2,i}|}(2t\mu_i) \tag{5.8}$$

and using (5.4) and (5.5) we have:

$$\begin{aligned} &P(\{r_i : i \in \mathfrak{C}\}|\{m_i\}, \{\mu_i\}) P(\{R_i = 1 : i \in \mathfrak{A}\}|\{\mu_i\}, L, N_e) \\ &\quad \times P(\{R_i = 0 : i \in \mathfrak{N}\}|\{\mu_i\}, L, N_e) \\ &= \prod_{i=1}^{c} P(r_i|m_i, \mu_i) P(R_i = 1|\mu_i, L, N_e) \prod_{i=c+1}^{k} P(R_i = 1|\mu_i, L, N_e) \\ &\quad \times \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L, N_e) \\ &= \prod_{i=1}^{c} \binom{m_i}{r_i} \mu_i^{r_i} \left(1 - \mu_i\right)^{m_i - r_i} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=c+1}^{k} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=k+1}^{n} e^{-\mu_i L N_e}. \end{aligned} \tag{5.9}$$

From (5.1) and (5.2) we have:

$$P(\{\mu_i : i \in \mathfrak{T} \cup \mathfrak{C} \cup \mathfrak{A} \cup \mathfrak{N}\}|\alpha, \beta) = \sum_{i=1}^{n} \frac{\mu_i^{\alpha-1} e^{-\mu_i/\beta}}{\beta^{\alpha} \Gamma(\alpha)}, \tag{5.10}$$

$$P(L|n_{asc}) = \frac{n_{asc} - 1}{2} e^{-L/2} \left(1 - e^{-L/2}\right)^{n_{asc}-2}. \tag{5.11}$$

An alternative way of writing the joint probability of the data and the unknown parameters is:

$$\begin{aligned} &P\left(\{\mu_i\}, t, L, N_e, \alpha, \beta, \{R_i\}, \{r_i\}, \{x_{1,i}\}, \{x_{2,i}\}, n_{asc}, \{m_i\}\right) \\ &= P\left(\{\mu_i\}, t, L, N_e, \alpha, \beta|n_{asc}, \{m_i\}, \{R_i\}, \{r_i\}, \{x_{1,i}\}, \{x_{2,i}\}\right) \\ &\quad \times P\left(n_{asc}, \{m_i\}, \{R_i\}, \{r_i\}, \{x_{1,i}\}, \{x_{2,i}\}\right). \end{aligned} \tag{5.12}$$

Since the probability of the data is constant, we have:

$$
\begin{aligned}
&P(\{\mu_i\}, t, L, N_e, \alpha, \beta | n_{asc}, \{m_i\}, \{R_i\}, \{r_i\}, \{x_{1,i}\}, \{x_{2,i}\}) \\
&\propto \prod_{i=1}^{s} e^{-2t\mu_i} I_{|x_{1,i}-x_{2,i}|}(2t\mu_i) \binom{m_i}{r_i} \mu_i^{r_i} (1-\mu_i)^{m_i-r_i} \left(1-e^{-\mu_i L N_e}\right) \\
&\quad \times \prod_{i=s+1}^{c} \binom{m_i}{r_i} \mu_i^{r_i} (1-\mu_i)^{m_i-r_i} \left(1-e^{-\mu_i L N_e}\right) \prod_{i=c+1}^{k} \left(1-e^{-\mu_i L N_e}\right) \prod_{i=k+1}^{n} e^{-\mu_i L N_e} \\
&\quad \times \prod_{i=1}^{n} \frac{\mu_i^{\alpha-1} e^{-\mu_i/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \frac{n_{asc}-1}{2} e^{-L/2} \left(1-e^{-L/2}\right)^{n_{asc}-2} P(N_e) P(\alpha) P(\beta) P(t).
\end{aligned}
\tag{5.13}
$$

Using exponential priors for t, α and β and a normal prior for N_e truncated at zero we have the posterior distribution:

$$
\begin{aligned}
&P(\{\mu_i\}, t, L, N_e, \alpha, \beta | n_{asc}, \{m_i\}, \{R_i\}, \{r_i\}, \{x_{1,i}\}, \{x_{2,i}\}) \\
&\propto \prod_{i=1}^{s} e^{-2t\mu_i} I_{|x_{1,i}-x_{2,i}|}(2t\mu_i) \binom{m_i}{r_i} \mu_i^{r_i} (1-\mu_i)^{m_i-r_i} \left(1-e^{-\mu_i L N_e}\right) \\
&\quad \times \prod_{i=s+1}^{c} \binom{m_i}{r_i} \mu_i^{r_i} (1-\mu_i)^{m_i-r_i} \left(1-e^{-\mu_i L N_e}\right) \prod_{i=c+1}^{k} \left(1-e^{-\mu_i L N_e}\right) \prod_{i=k+1}^{n} e^{-\mu_i L N_e} \\
&\quad \times \prod_{i=1}^{n} \frac{\mu_i^{\alpha-1} e^{-\mu_i/\beta}}{\beta^{\alpha}\Gamma(\alpha)} \\
&\quad \times \frac{n_{asc}-1}{2} e^{-L/2} \left(1-e^{-L/2}\right)^{n_{asc}-2} \\
&\quad \times \frac{1}{\sigma_{N_e}\sqrt{2\pi}} \exp\left(-\frac{(N_e-\mu_{N_e})^2}{2\sigma_{N_e}^2}\right) e^{-\lambda_\alpha \alpha} e^{-\lambda_\beta \beta} e^{-\lambda_t t}.
\end{aligned}
\tag{5.14}
$$

5.2.3 Markov-Chain Monte-Carlo Sampling

As in chapter 4, we make use of Markov-chain Monte-Carlo (MCMC) in order to sample from the posterior distribution (5.14). Again we use an adaptive Metropolis algorithm to achieve this objective. This methodology was outlined in section 4.3.3.2.

5.2.3.1 Implementation: TMRCA Model

All the model parameters are collected in a vector θ:

$$\theta = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \\ \mu_{k+1} \\ \vdots \\ \mu_n \\ L \\ N_e \\ \alpha \\ \beta \\ t \end{bmatrix}. \tag{5.15}$$

Each element in θ is updated according to a normal proposal distribution centred at the current value of the element in θ. The variance of the update will be updated according to the adaptive aspect of the algorithm. The proposal distribution is symmetric and the Metropolis ratio is:

$$R = \frac{f(\theta^*)}{f(\theta)}, \tag{5.16}$$

where $\theta^* = \theta$ for each component other than the one being updated and $f(\theta)$ is the posterior distribution. Hence the Metropolis ratio may be simplified, as before, for each parameter. For example for t we have:

$$R = \frac{P(t^*) \prod_{i=1}^{s} P(\{x_{1,i}\}, \{x_{2,i}\} | \{\mu_i\}, t^*)}{P(t) \prod_{i=1}^{s} P(\{x_{1,i}\}, \{x_{2,i}\} | \{\mu_i\}, t)}. \tag{5.17}$$

This form is then used to either accept or reject the proposal, t^*. Similarly we can use a reduced form of R when updating the other parameters. For L we have:

$$R = \frac{P(L^*|n_{asc}) \prod_{i=1}^{k} P(R_i = 1|\mu_i, L^*, N_e) \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L^*, N_e)}{P(L|n_{asc}) \prod_{i=1}^{k} P(R_i = 1|\mu_i, L, N_e) \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L, N_e)}; \tag{5.18}$$

for N_e:

$$R = \frac{P(N_e^*) \prod_{i=1}^{k} P(R_i = 1|\mu_i, L, N_e^*) \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L, N_e^*)}{P(N_e) \prod_{i=1}^{k} P(R_i = 1|\mu_i, L, N_e) \prod_{i=k+1}^{n} P(R_i = 0|\mu_i, L, N_e)}; \tag{5.19}$$

for α:

$$R = \frac{P(\alpha^*) \prod_{i=1}^{n} P(\{\mu_i\}|\alpha^*, \beta)}{P(\alpha) \prod_{i=1}^{n} P(\{\mu_i\}|\alpha, \beta)}; \tag{5.20}$$

for β:

$$R = \frac{P(\beta^*) \prod_{i=1}^{n} P(\{\mu_i\}|\alpha, \beta^*)}{P(\beta) \prod_{i=1}^{n} P(\{\mu_i\}|\alpha, \beta)}; \tag{5.21}$$

for the mutation rate at each typed marker, i.e. μ_i where $i = 1, \ldots, s$ we have:

$$R = \frac{P\left(x_{1,i}, x_{2,i}|\mu_i^*, t\right) P(r_i|m_i, \mu_i^*) P(R_i = 1|\mu_i^*, L, N_e) P(\mu_i^*|\alpha, \beta)}{P\left(x_{1,i}, x_{2,i}|\mu_i, t\right) P(r_i|m_i, \mu_i) P(R_i = 1|\mu_i, L, N_e) P(\mu_i|\alpha, \beta)}; \tag{5.22}$$

for each of the remaining calibrated loci, i.e. μ_i where $i = s + 1, \ldots, c$:

$$R = \frac{P(r_i|m_i, \mu_i^*) P(R_i = 1|\mu_i^*, L, N_e) P(\mu_i^*|\alpha, \beta)}{P(r_i|m_i, \mu_i) P(R_i = 1|\mu_i, L, N_e) P(\mu_i|\alpha, \beta)}; \tag{5.23}$$

for each of the remaining ascertained loci, i.e. μ_i where $i = c + 1, \ldots, k$:

$$R = \frac{P(R_i = 1|\mu_i^*, L, N_e) P(\mu_i^*|\alpha, \beta)}{P(R_i = 1|\mu_i, L, N_e) P(\mu_i|\alpha, \beta)}; \tag{5.24}$$

and for the non-ascertained loci, i.e. μ_i where $i = k+1, \ldots, n$:

$$R = \frac{P(R_i = 0|\mu_i^*, L, N_e)P(\mu_i^*|\alpha, \beta)}{P(R_i = 0|\mu_i, L, N_e)P(\mu_i|\alpha, \beta)}. \tag{5.25}$$

5.2.4 Data Simulation and Analysis Program

The use of the function `tmodel` is detailed below:

Description

`tmodel` allows the user to simulate data from pairs of males by specifying the model parameters and sample size (see Details below). The data is then analysed using MCMC to provide estimates of the mutation rates across all loci, t, α, β, L (coalescent units) and N_e as well as L (generations) and the mean and variance of the distribution from which the mutation rates are drawn, i.e. gamma(α, β). Several diagnostics are also produced to evaluate the performance of the chains.

Usage

```
tmodel(BatchLen=50, TotBatch=1000, BinBatch=10, AccRate=0.44,
maxLSD, Loci, pCAL=0.211, Pnasc=NA, ptloci=0.5, t, ascsamp=8,
meioses=NA, cloops=3, sortCAL=F, L, Ne, muNe=6037, sdNe=1745,
alpha=1.70331, beta=0.001404, startL=3, startNe=6037,
startalpha=1.7, startbeta=0.00014, startTrue=F, psdmu=0.002,
psdL=0.5, psdNe=500, psdAlpha=0.05, psdBeta=0.06, psdT=3,
lambdaA=0, lambdaB=0, lambdaT=0, getdata=F, method=2,
graph="graphres.eps")
```

Required Arguments

`BatchLen`: this is the number of updates within each batch before the log standard deviation of the proposal distribution is adapted.

`TotBatch`: the total number of batches to run.

`BinBatch`: the total number of batches to remove as burn-in.

`AccRate`: the optimum acceptance rate.

`MaxLSD`: the boundary for proposed values of the log standard deviation of the proposal distribution.

`Loci`: the number of STRs, n.

`pCAL`: the proportion of the ascertained loci that are calibrated.

`ptloci`: the proportion of the calibrated loci that are typed in the male pairs.

`Pnasc`: assigns the percentage of non-ascertained loci in the analysis. It can be set to a different value by choosing a value in the range $(0, (1 - ascCAL/loci))$ where $ascCAL$ is the number of calibrated loci. When set to `NA` (default), the true percentage of non-ascertained loci is passed to the MCMC sampler.

`ascsamp`: the ascertainment sample size.

`meioses`: numeric (integer). The number of meioses used to calibrate the ascertained loci to be calibrated. When set to `NA` (default), will use the multi-stage calibration process.

`cloops`: numeric (integer). The number of cycles using the multi-stage calibration process to be carried on those loci to be calibrated when used in conjunction with `meioses=NA`

`t`: the time to the most recent common ancestor for the two males.

`L`: the total branch length in the Y-chromosome sample in which loci are ascertained, in units of N_e generations.

`Ne`: the effective Y-chromosome population size,

`muNe`: the mean effective population size, in the prior for N_e.

`sdNe`: the standard deviation of the effective population size, in the prior for N_e.

`alpha`: the shape parameter in the Gamma distribution used to sample mutation rates.

`beta`: the scale parameter in the Gamma distribution used to sample mutation rates.

`startL`: the initial value for the MCMC for L.

`startNe`: the initial value for the MCMC for N_e.

`startalpha`: the initial value for the MCMC for α.

`startbeta`: the initial value for the MCMC for β.

`startt`: the initial value for the MCMC for t.

`startTrue`: logical, if `TRUE` initialises the MCMC chains to start at the true value for each parameter.

`psdmu`: the initial standard deviation for the proposal distribution of the mutation rates.

`psdL`: the initial standard deviation for the proposal distribution of L.

`psdNe`: the initial standard deviation for the proposal distribution of N_e.

`psdAlpha`: the initial standard deviation for the proposal distribution of α.

`psdBeta`: the initial standard deviation for the proposal distribution of β.

`psdT`: the initial standard deviation for the proposal distribution of t.

`lambdaA`: the lambda value for α's prior distribution.

`lambdaB`: the lambda value for β's prior distribution.

`lambdaT`: the lambda value for t's prior distribution.

`SortCAL`: logical, if `TRUE` sorts the calibrated loci according to calibrated rates, in descending order.

`getdata`: logical, if `TRUE` the program will try to retrieve the data and parameters stored in a list named 'results'.

`method`: controls which posterior distribution is to be sampled from; 2 is detailed in equation 5.14, 1 is any competing posterior distribution.

`graph`: the name of the file that must end in ".eps" that any graphical diagnostics will be saved as. If `NA`, no graphical diagnostics will be returned or saved

Side Effects & Returns

The function returns a list of the true parameter values, MCMC mean, standard deviation and credible regions for t (generations), L (in both generations and N_e units), N_e, α, β, mean of gamma, variance of gamma. In addition the list includes the acceptance rates for each parameter. Also included are details of the mutation rates including the true values, MCMC means, standard deviations and credible regions, the number of mutations and meioses for the calibrated rates as well as their calibrated estimates. Additionally, the true proportion and number of calibrated and non-ascertained loci with any misspecified values stated, along with the total number of loci are given. Finally the list includes the mean and standard deviation of the prior for N_e, as well as the number of simulated Y-chromosomes used to determine the ascertained mutation rates.

For `graph="graphres.eps"`, a 3×6 plot of the chains for the first mutation rate (a typed locus), a calibrated locus, an ascertained locus, the last mutation rate (a non-ascertained locus), L (units of N_e generations), N_e, α, β and t are shown, with the true value indicated by a solid grey line. The corresponding updates of the log of the standard deviation of each of the parameters' proposal distributions follow.

Details

The program is entirely coded in the R language which simulates data according to the steps outlined in section 5.2.1, creates the appropriate update vector and passes this and other relevant parameters to the adaptive MCMC loop. This generates a new proposal for the parameter being updated and either accepts or rejects it on the basis of the posterior distribution.

After every `BatchLen` updates of all the parameters, the log standard deviation of the proposal distribution is adapted according to section 4.3.3.2 and the specified optimum acceptance rate. Once the chosen number of batches, i.e. `TotBatch`, has been constructed for each parameter, the R code generates the summary statistics followed by the graphical diagnostics (if specified) and returns the list specified in Side Effects & Returns, above. Where `getdata=T`, the data must have been previously stored in a list named 'results'. This is ideal for comparing different analyses of the same data, for example when altering the amount of data or misspecifying parameters.

5.2.5 Software

The software used to develop the TMRCA model and related analyses was R. The results presented in this chapter were obtained on a Linux platform.

5.2.6 Analysis

The results are based on the following choice of parameters, which are fixed throughout the analysis: $L = 3.01$, $N_e = 4262$, $\alpha = 0.5872$, $\beta = 0.00348$, $n = 475$ and $n_c = 3$.

For each simulated dataset, we will obtain estimates of the parameters by applying 1000 batches ($TotBatch$) of length 50 ($Batchlen$) of the MCMC sampler with 100 batches ($Binbatch$) discarded as burn-in. The prior parameters; λ_α, λ_β and λ_t, are all zero, corresponding to uniform (improper) priors.

The first analysis involves simulating STR profiles for pairs of males with most recent common ancestor at times t= 5, 10, 15, 20, 25, 50, 100 (generations) ago. For each t, we generate 100 independent datasets, setting the percentage of typed

loci to 100%. This means that the number of STRs typed in the simulated male pairs is equal to the total number, c, of calibrated loci. Thereafter we subsample the following percentages of typed loci from the initial datasets: 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%.

As a consequence, we have a total of 700 independent datasets across t. For each combination of t and the percentage of typed loci, we will compute the MSE of the posterior mean, as estimator of t. To scale out t to allow a direct comparison across t we also compute the fractional squared error (FSE), fractional variance and fractional bias squared, which are obtained by dividing, respectively, the MSE, variance and squared bias by t^2. The fractional bias is simply the bias divided by t. For the other parameters, such as α, β, the mean and variance of the gamma distribution with shape α and scale β, L, N_e, we compute the MSE, variance and bias. For every parameter any plots will have the same scale unless this obscures the results.

Following on from this, we will examine the effect of varying the percentage of the fastest mutation rates on $\hat{t}$. In this case, we will use the datasets generated as described above. However, the calibrated mutation rates will be ordered according to their empirical mutation rates, i.e. based on the simulated meioses and mutations. Thereafter the percentage of the fastest typed loci will be varied as: 10%, 20%, 30%, 40%, 50%, 60%, and 70%.

Lastly we will conduct a misspecification study on the 100 original datasets at each t by varying the percentage of non-ascertained loci away from the true value. Given the random nature of the simulation the percentage of non-ascertained loci will vary from one dataset to another. Nonetheless we examine the effect of misspecifying this as follows: 5%, 10%, 15% and 20%.

5.3 Results

5.3.1 Varying the Proportion of Typed STRs

Figure 5.3 plots the FSE of $\hat{t}$ against the percentage of typed loci for each time, with the fractional variance (FV) superimposed in a solid red line. In general, we find that the FSE is high when the percentage of typed loci is only 10% but quickly decreases as the percentage increases. This is the case across all times.

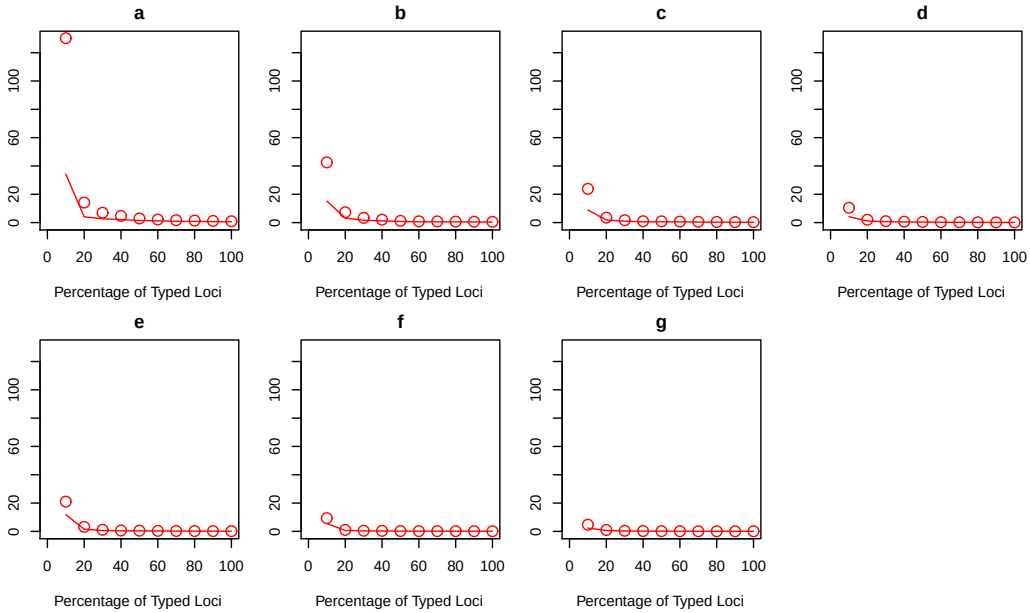

FIGURE 5.3: Fractional squared error and variance of $\hat{t}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - FSE/FV)

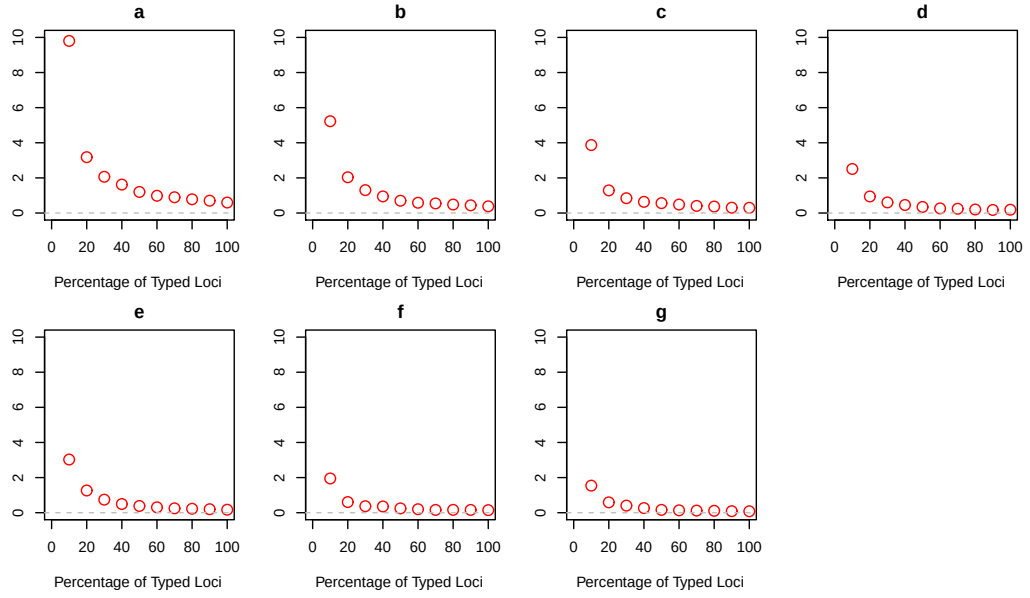

FIGURE 5.4: Fractional bias of $\hat{t}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

In addition, as t increases from 5-100 generations, the FSE for low values of the percentage of typed loci reduces substantially. In figure 5.3a where $t = 5$, the FV is about a quarter of the FSE at 10%. The remaining component is from the fractional bias squared (FBSQ). As the percentage of typed loci increases, the FV component increases, so there is less fractional bias. This is also the case for t at 10, 15, 20 and 25 generations. However, at low percentages of typed loci, at each of these times, the fractional variance component is increasingly greater

(figs. 5.3b-e). At $t = 50$ and 100 generations, the main component of the FSE is the FV across the entire range of the percentage of typed loci.

Next we examine the fractional bias (FB) of $\hat{t}$ (fig. 5.4). For all t, the FB is positive and it quickly decreases approaching an asymptote close to zero as the percentage of typed loci increases. In addition, the rate at which the FB decreases reduces as time increases. Consequently, at 10% for $t = 100$ (fig. 5.4g), the FB is $\sim$2, whereas, for $t = 5$, it is $\sim$10 (fig. 5.4a).

For $\hat{\alpha}$ the MSE against the percentage of typed loci at $t = 5 - 25$ is approximately constant at $\sim$0.006 (figs. 5.5a-e). This is also the case for $t = 50$ although the MSE is slightly higher at $\sim$0.007. For $t = 100$, the MSEs of $\hat{\alpha}$ vary in an inconsistent manner from 0.008-0.012 as the percentage on typed loci increase, possibly due to poor mixing and convergence. The variance (solid red line) contributes just over half of the MSE of $\hat{\alpha}$ and is roughly constant as the percentage of loci increases from $t = 5$ to 25 (fig. 5.5a-e). However the contribution is greater at $t = 50$ and 100. The bias in $\hat{\alpha}$ is negative (fig. 5.6). $\hat{\alpha}$ is underestimating by a roughly constant amount across the percentage of typed loci for all t, but at $t = 100$ the bias is slightly less. The overall amount of bias does not vary across $t = 5 - 25$ in any consistent way, but, for $t = 50$ and 100 the bias is reduced compared to that for other t.

Next we examine the MSE of $\hat{\beta}$, which is roughly constant as the percentage of typed loci increases for each t (fig. 5.7). However, the MSE is greatest when $t = 5$ and reduces until $t = 20$ (figs. 5.7ad). Thereafter, it increases till $t = 50$ (fig. 5.7e) before reducing when $t = 100$((fig. 5.7g). For each t, the variance is approximately constant as the percentage of typed loci increases (solid red line). Consequently, we find that as t increases the contribution the variance component makes to the MSE also increases. The bias in $\hat{\beta}$ versus the percentage of typed loci is shown in figure 5.8. Across each t, the bias is positive, so the method is overestimating $\hat{\beta}$. Also there is a trend of decreasing bias as the percentage of typed loci increases. Furthermore the overall bias for $t = 5 - 20$ reduces relative to lower t (figs. 5.8a-d). At $t = 25$ the bias increases relative to the shorter times, decreasing thereafter (figs. 5.8e-g).

The MSE in the estimate of the mean of the gamma distribution from which the mutation rates are drawn is shown in figure 5.9. The MSE is constant as the percentage of typed loci increases, for each t. Additionally, the MSE is small

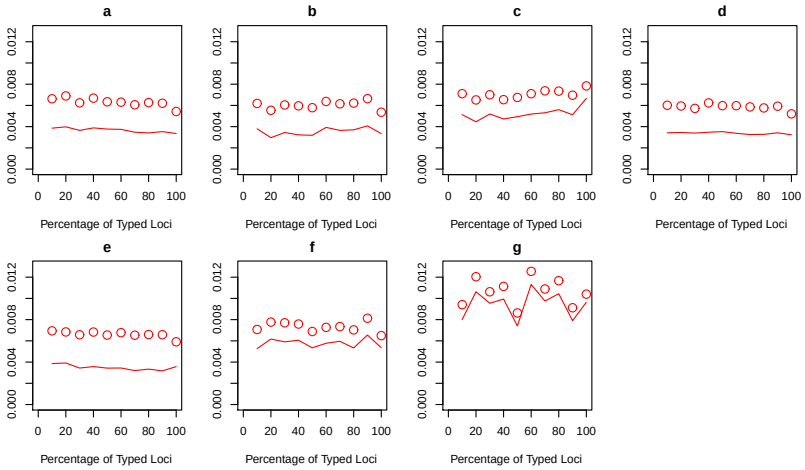

FIGURE 5.5: Mean squared error and variance of $\hat{\alpha}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

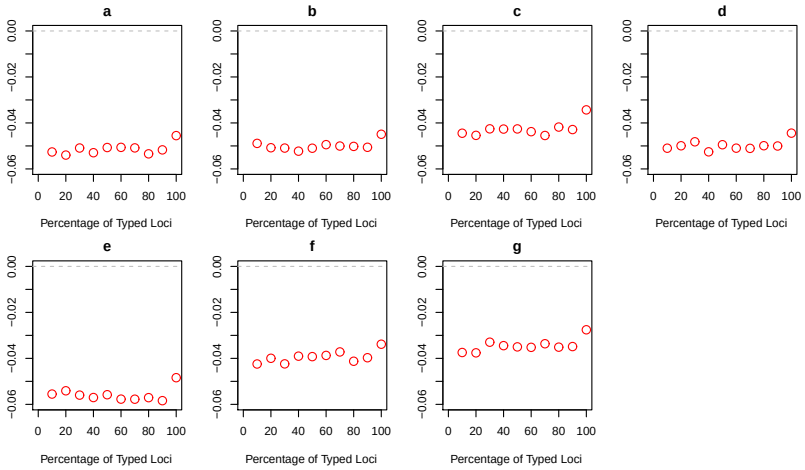

FIGURE 5.6: Bias of $\hat{\alpha}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

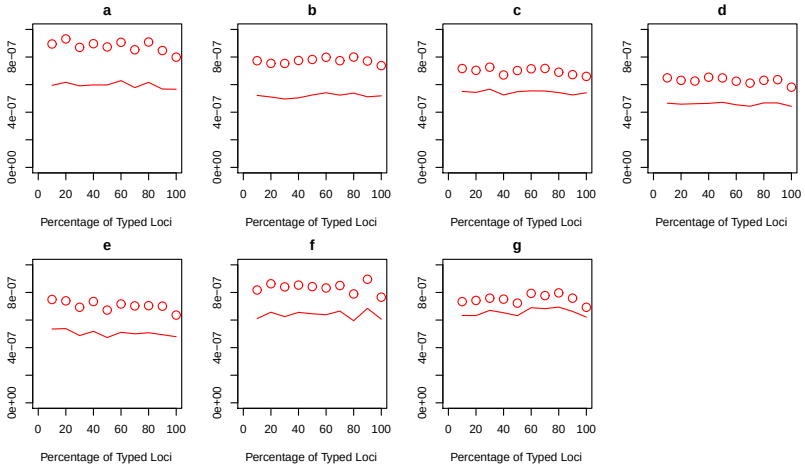

FIGURE 5.7: Mean squared error and variance of $\hat{\beta}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

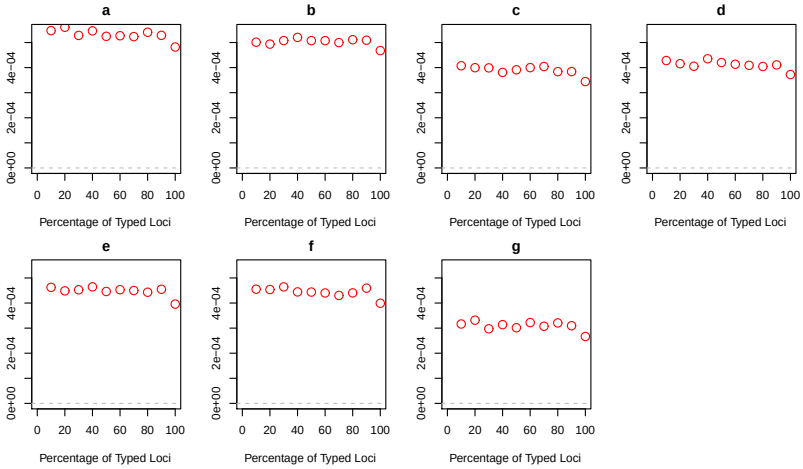

FIGURE 5.8: Bias of $\hat{\beta}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

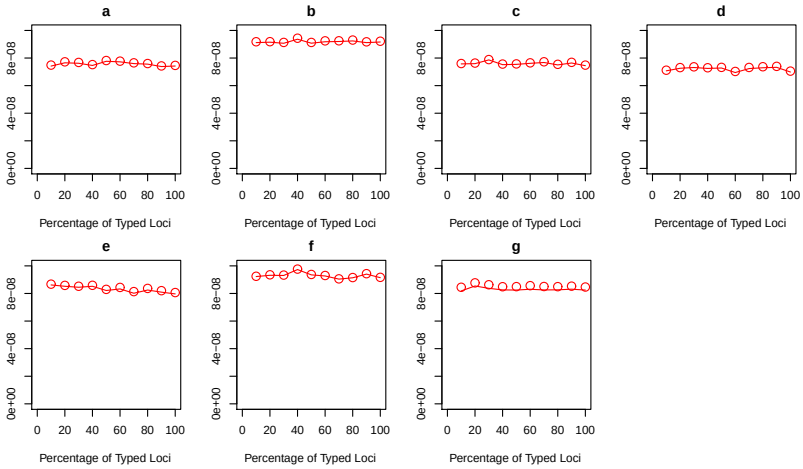

FIGURE 5.9: Mean squared error and variance of gamma mean vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

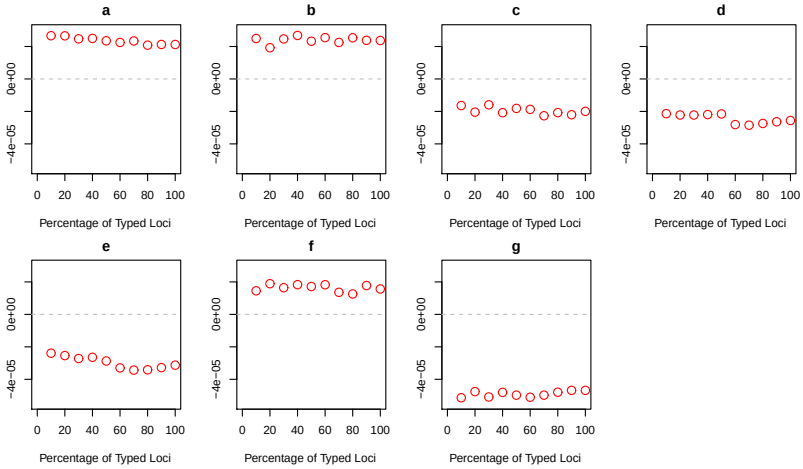

FIGURE 5.10: Bias of gamma mean vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

across t. The variance component in the MSE is shown in a solid red line and forms the majority of the MSE for each t. The bias in this estimator is shown in figure 5.10. Here we see that, for $t = 5, 10, 50$, the bias is positive (figs. 5.8abf) and, for the remaining t, is it negative. Nonetheless, the bias is very small. In addition varying the percentage of typed loci has little effect on the bias.

For the estimates of the variance of the gamma distribution, the MSE, with the variance superimposed, is shown in figure 5.11. We see that there is a weak negative relationship between the MSE and the percentage of typed loci particularly for $t = 5 - 25$ (figs. 5.11a-e). For the other two cases, the MSEs do not appear to vary systematically with increasing percentage of typed loci. In addition, we find that, as t increases from 5 to 20, the MSE reduces overall. In contrast, this is not true for $t = 25 - 100$. As with the MSE, the variance decreases slightly with increasing percentage of typed loci for $t = 5 - 25$ (figs. 5.11a-e). Also, the contribution the variance makes to the MSE increases as t increases. For the remaining cases ($t = 50$ and 100), the variance is roughly constant, yet the contribution the variance makes to the MSE is proportionally greater for $t = 100$ than for $t = 50$ (figs. 5.11gf). The bias in the estimates of the variance of the gamma distribution reduces slightly with increasing percentage of typed loci though remains positive (fig. 5.12). This is the case for each t. Furthermore the overall bias reduces as t increases.

For $\hat{L}$, the MSE versus the varying percentage of typed loci is shown in figure 5.13. For each t, the MSE decreases slightly with the percentage of typed loci. Moreover, the overall MSE reduces as t increases from 5-15 generations before increasing to the overall maximum at $t = 25$ (fig. 5.13e) and then decreasing again. The variance component (solid red line) is about a third of the MSE for each t, with the exception of $t = 100$. As with the MSE for each t, the bias in $\hat{L}$ reduces slightly with the percentage of typed loci. It remains positive throughout (fig. 5.14). In addition, the overall bias decreases when $t = 5 - 15$, thereafter increasing before decreasing when $t = 50$ and 100.

The last parameter to have its MSE examined for this set of analyses is $\hat{N}_e$ (fig. 5.15). Here we also find a weak negative relationship between the MSE and the percentage of typed loci, for each value of t. The overall MSE reduces between $t = 5 - 15$ (figs. 5.15a-c). Thereafter it increases until $t = 25$, before reducing again. The variance component (solid red line) contributes approximately a fifth of the MSE for lower t but slightly more at higher t. As for the bias in $\hat{N}_e$, we find it is positive for each t and as the percentage of typed loci increases the bias

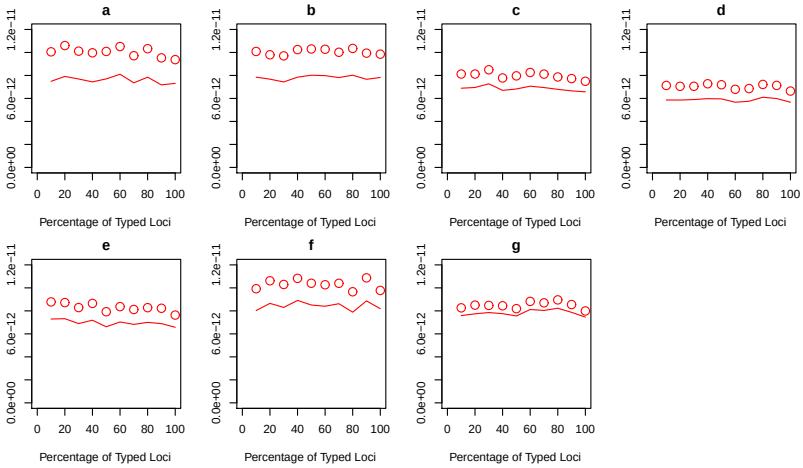

FIGURE 5.11: Mean squared error and variance of gamma variance vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

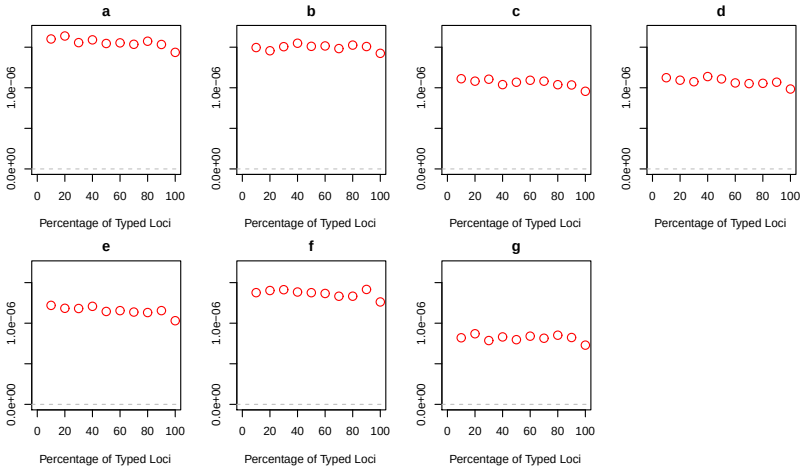

FIGURE 5.12: Bias of gamma variance vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

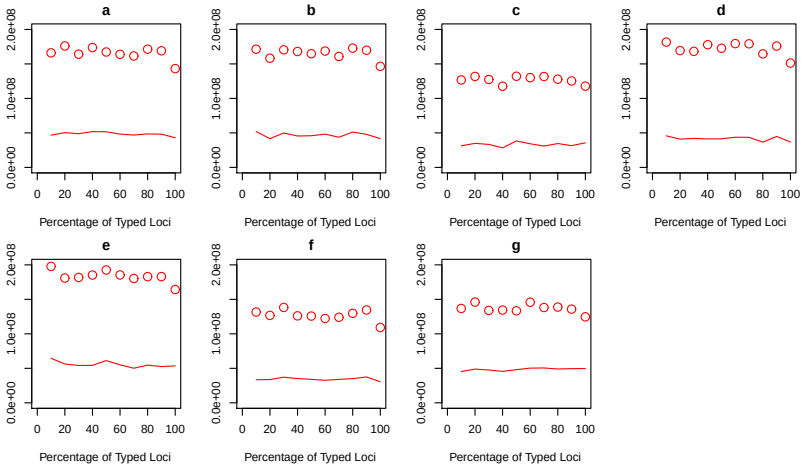

FIGURE 5.13: Mean squared error and variance of $\hat{L}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

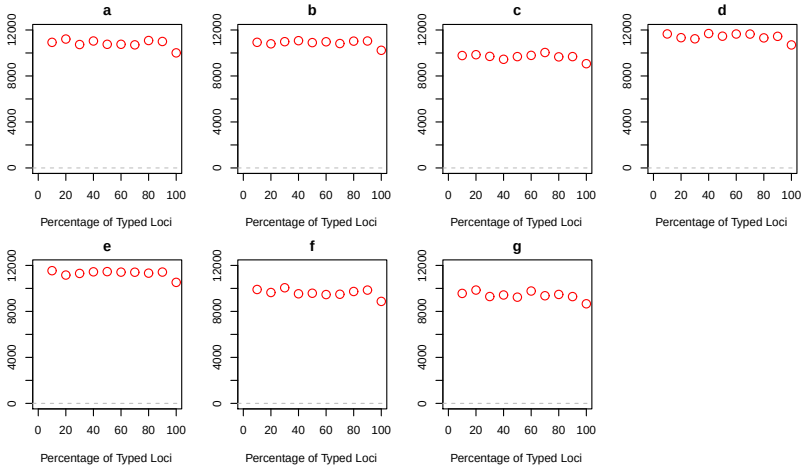

FIGURE 5.14: Bias of $\hat{L}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

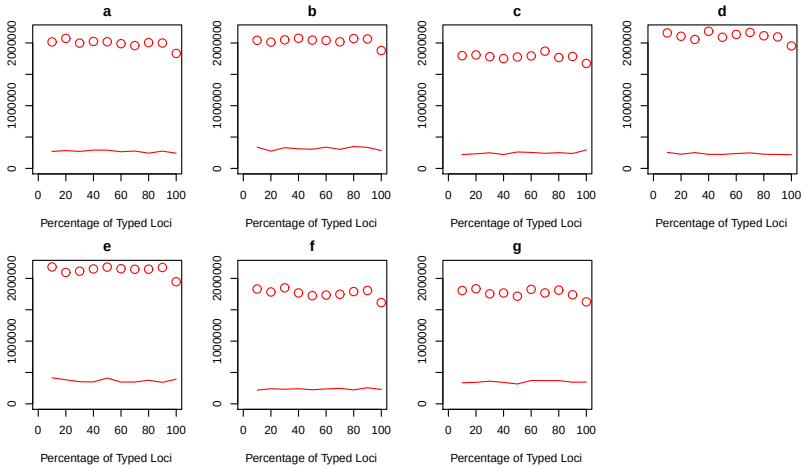

FIGURE 5.15: Mean squared error and variance of $\hat{N}_e$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

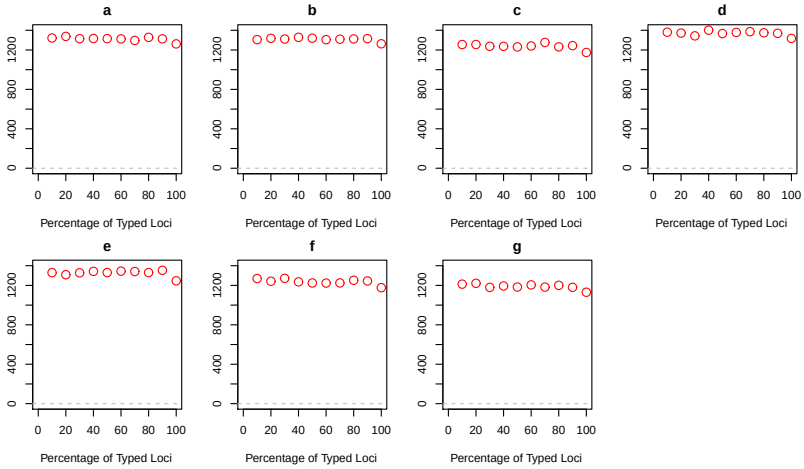

FIGURE 5.16: Bias of $\hat{N}_e$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

drops very slightly (fig. 5.16). The overall bias is greatest at $t = 25$ (fig. 5.16e) though for the other values of t it is not substantially less.

5.3.2 Use of Fast Mutating STRs

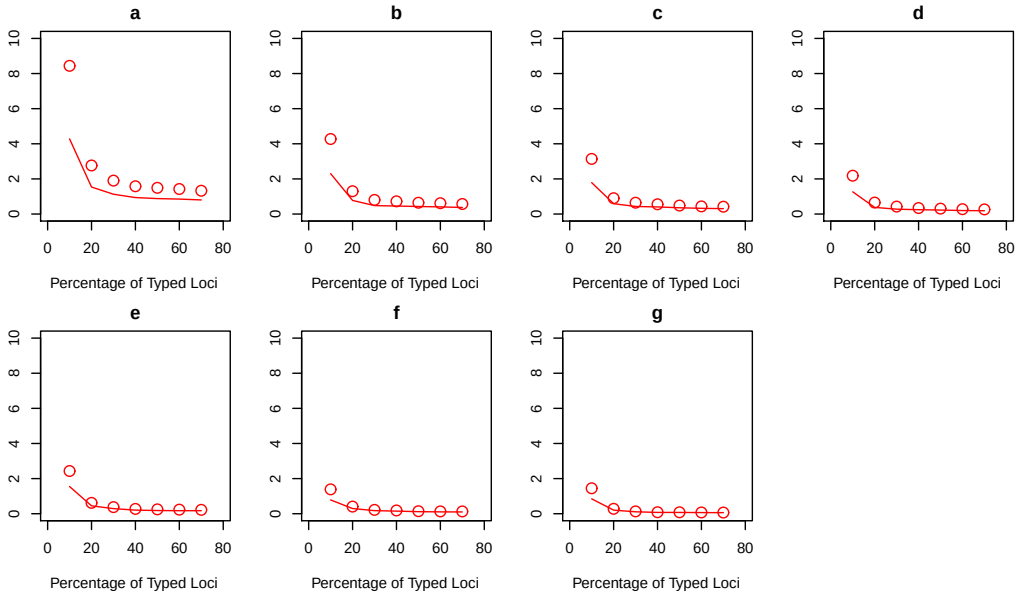

FIGURE 5.17: Fractional squared error and variance of $\hat{t}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - FSE/FV)

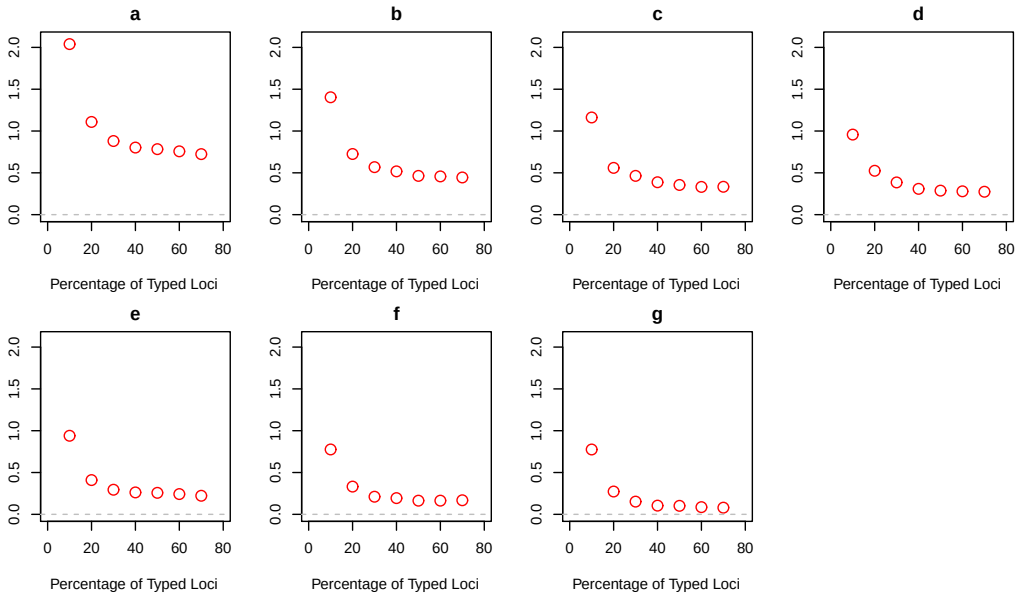

FIGURE 5.18: Fractional bias of $\hat{t}$ vs. percentage of typed loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

Here we will examine the effect on $\hat{t}$ only of using the fastest loci. Figure 5.17 is a plot of the FSE of $\hat{t}$ versus the percentage of typed loci with the fractional variance.

For the fastest loci we find that there is a trend of reducing FSE with increasing percentage of typed loci. This is the case for each t, but the reduction is most apparent when $t = 5$, where the FSE is ~ 9 at 10% and close to 1 by 40%. In addition, its FV component forms the majority of the FSE throughout. The FB of $\hat{t}$ decreases as the percentage of the fast typed loci increases approaching zero (fig. 5.18). Also the overall FB across the percentage of typed loci reduces as t increases.

5.3.3 Misspecification of the Proportion of Non-Ascertained Loci

In this section we examine the effect on the estimates of the main parameters of varying the percentage of non-ascertained loci, starting with $\hat{t}$. The FSE of $\hat{t}$ versus the percentage of non-ascertained loci is shown in figure 5.19, with the FV (solid red line). In general, we find that both the FSE and FV are constant as the percentage varies. This is the case for each t. However, the overall FSE decreases as t increases. Yet the contribution the FV makes to the FSE increases with increasing t. The FV makes up over half of the FSE at $t = 5$ and at 100 FV is the main component of the FSE (figs. 5.19ag). For the FB of $\hat{t}$, we see that there no change in it as the percentage of non-ascertained loci increases at each t (fig. 5.20). In addition, the FB reduces as t increases though it remain positive for each t.

Next, we examine the MSE of $\hat{\alpha}$ versus the percentage of non-ascertained loci (fig. 5.21). Here, at each t, we find that the MSE has a minimum at around 10% of the non-ascertained loci. In addition, for $t = 5-50$, we find that the variance has a negative relationship with the percentage of non-ascertained loci. Combining this with the parabolic form of the MSE means that the variance contributes the most at 10% of non-ascertained loci. For the remaining percentages of non-ascertained loci, the variance forms under half of the MSE. At $t = 100$, the variance has a peak at 10% where it makes up almost all the MSE, but at the other percentages only contributes less than a third to the MSE (fig. 5.21g). The bias in $\hat{\alpha}$ decreases with increasing percentage of non-ascertained loci at each t. In addition, the bias is positive at 5% and negative for the remaining percentages.

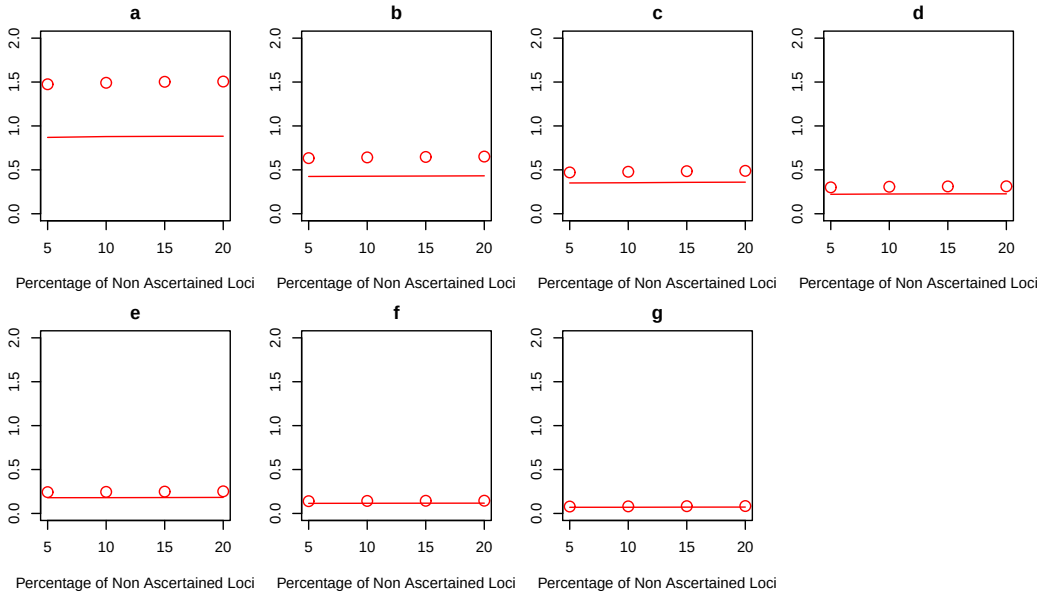

FIGURE 5.19: Fractional squared error and variance of $\hat{t}$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/FV)

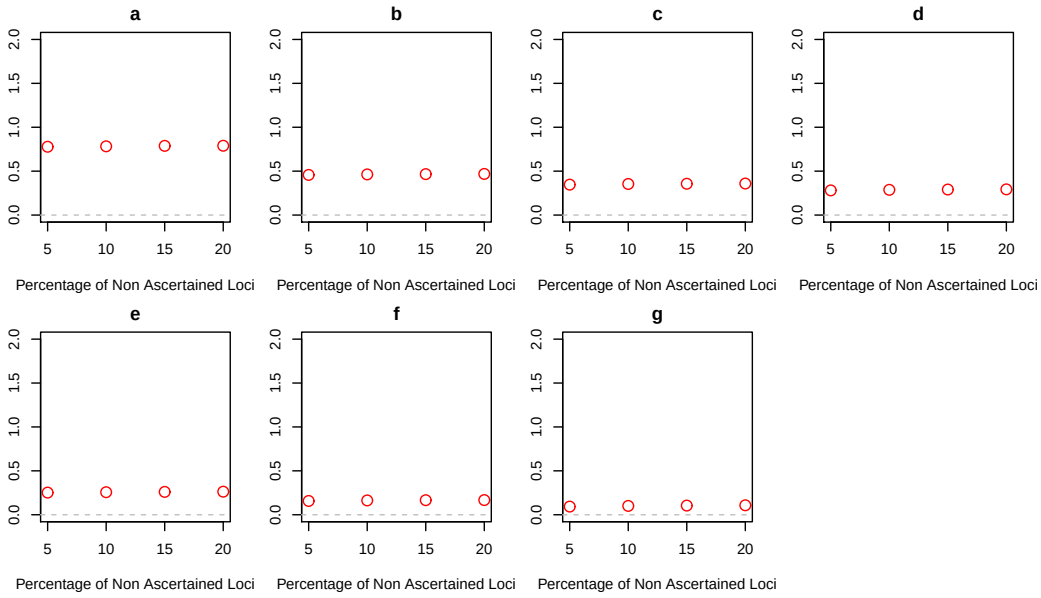

FIGURE 5.20: Fractional bias of $\hat{t}$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

For $\hat{\beta}$, the MSE increases at a rather linearly with increasing percentage of non-ascertained loci. This is the case at each t (fig. 5.23). Similarly, the variance increases but a lower rate (solid red line). At 5% the variance forms almost all the MSE whilst at 20% this is much less at around half. The bias $\hat{\beta}$ is shown in figure 5.24. It has a positive relationship with the percentage of non-ascertained loci at each t. We find that the bias is positive at each percentage of non-ascertained loci except 5%.

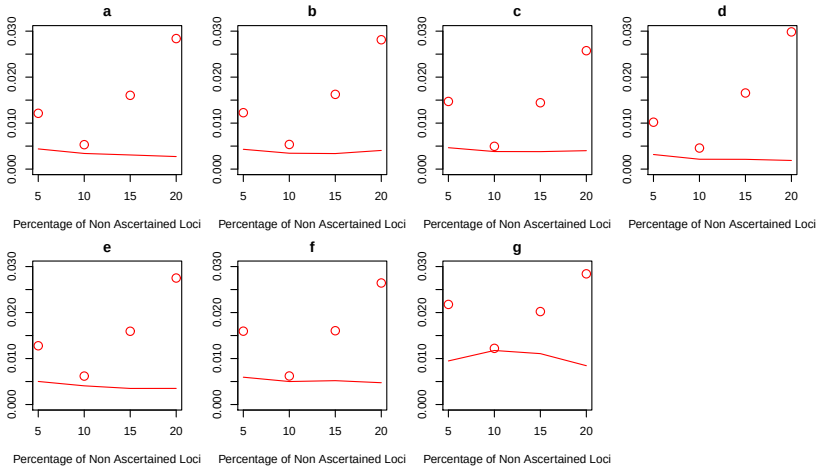

FIGURE 5.21: Mean squared error and variance of $\hat{\alpha}$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

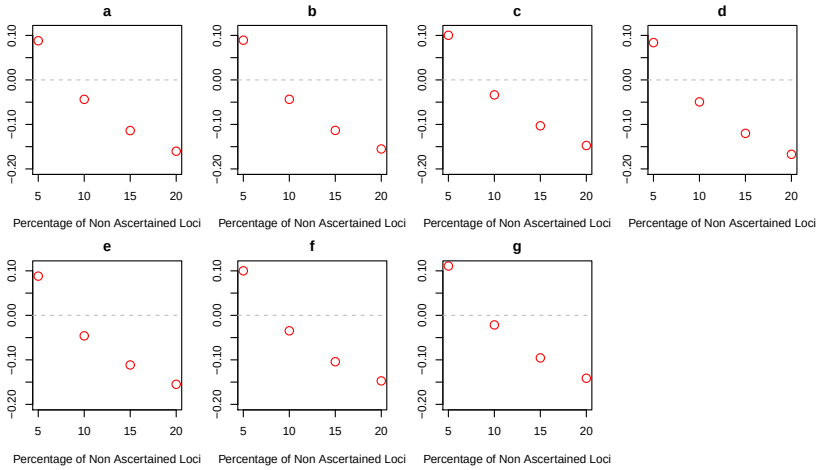

FIGURE 5.22: Bias of $\hat{\alpha}$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

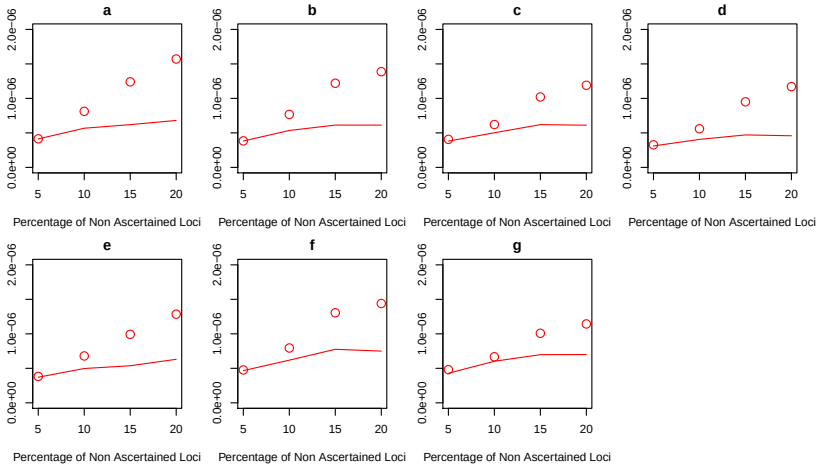

FIGURE 5.23: Mean squared error and variance of $\hat{\beta}$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

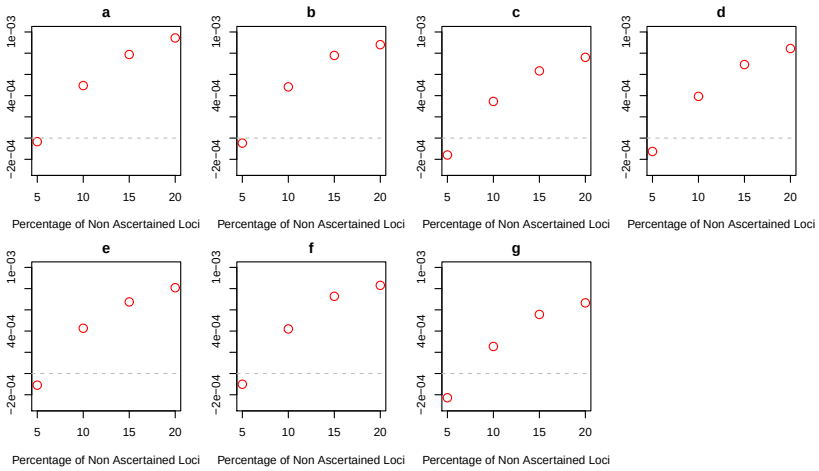

FIGURE 5.24: Bias of $\hat{\beta}$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

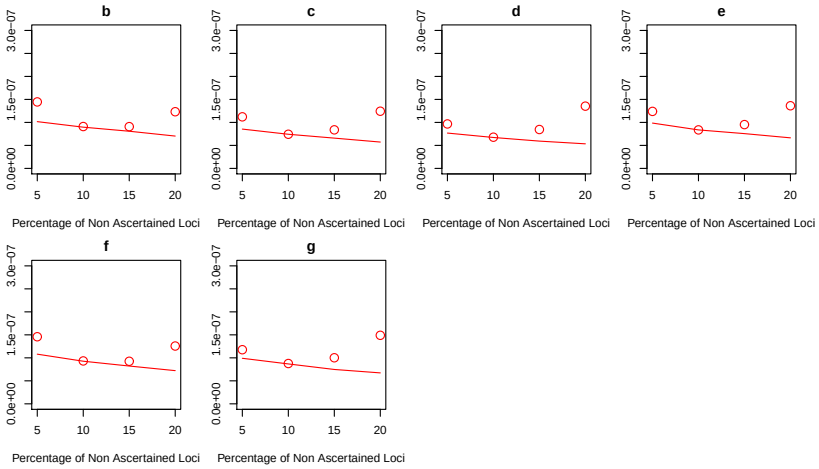

FIGURE 5.25: Mean Squared Error and Variance of Gamma mean vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

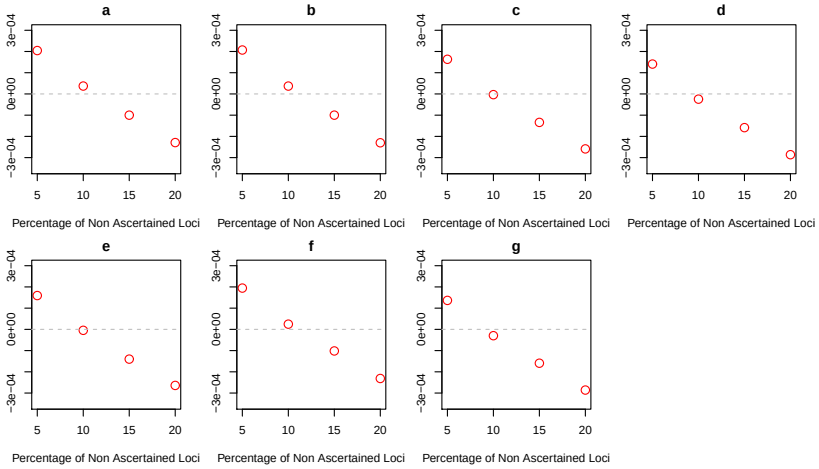

FIGURE 5.26: Bias of Gamma mean vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

Now we examine the MSE of the estimates of the mean of the gamma distribution from which the mutation rates are drawn. The MSEs form a parabola with a minimum at 10% of the non-ascertained loci at each t (fig. 5.25). On the other hand the variance (solid red line) is strictly decreasing with increasing percentage of non-ascertained loci, contributing the bulk of the MSE at 10%. For the other percentages the variance makes up half or more of the MSE. The bias in the estimates of the mean decreases with increasing percentage of non-ascertained loci, crossing zero close to 10% at each t (fig. 5.26).

For the MSE of the estimates of the variance of the gamma distribution, we see a similar pattern as for the mean: the points form a parabola but with a maximum around 10-15% (fig. 5.27). Also the bias in figure 5.28 is positive for all percentages of non-ascertained loci. As t increases, there is no consistent change in the overall bias. However, we find that the highest overall bias occurs at $t = 5$ and the lowest when $t = 100$.

The results for the total branch length, L, are examined next. The MSE has a negative relationship with the percentage of non-ascertained loci for all t (fig. 5.29). The variance component of the MSE decreases when the percentage of non-ascertained loci increases. However, it does so at a much slower rate than the MSE. For example at 5%, across each t, the variance forms only about a sixth of the MSE whilst at 20% it makes up the majority. The bias in $\hat{L}$ decreases with increasing percentage of non-ascertained loci, though it remains positive throughout as shown in figure 5.30. The pattern is consistent as t varies.

Lastly, for $\hat{N}_e$ we see that its MSE decreases when the percentage of non-ascertained loci increases for each t (fig. 5.31). Conversely, the variance component of the MSE (solid red line) increases with the percentage of non-ascertained loci at each t. Thus at 5% the variance contributes very little to the MSE whilst at 20% it forms the bulk of it. The bias in $\hat{N}_e$ reduces with increasing percentage of non-ascertained loci for each t as shown in figure 5.32. Furthermore, the bias remains positive with little difference in the values across t.

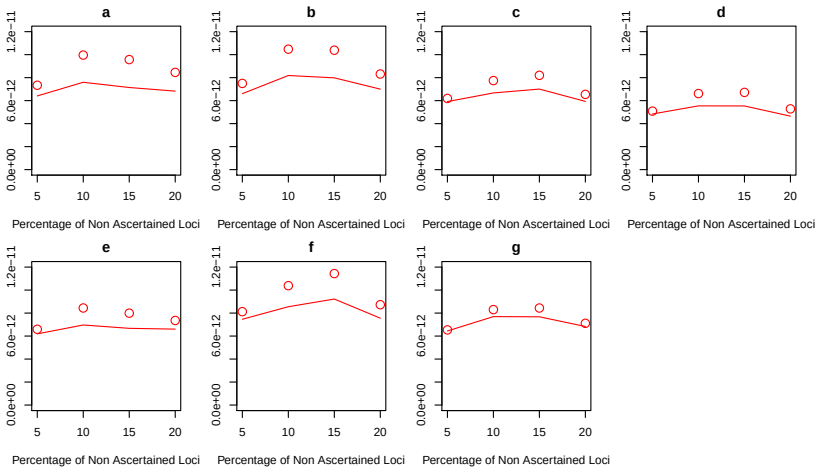

FIGURE 5.27: Mean squared error and variance of gamma variance vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

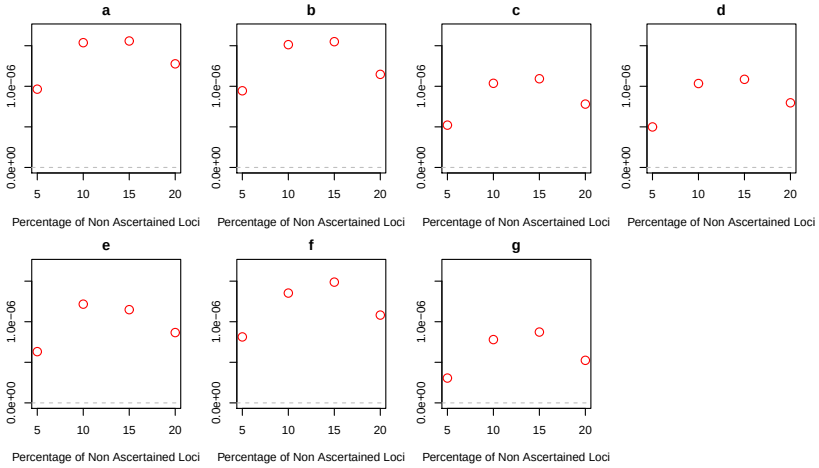

FIGURE 5.28: Bias of gamma variance vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

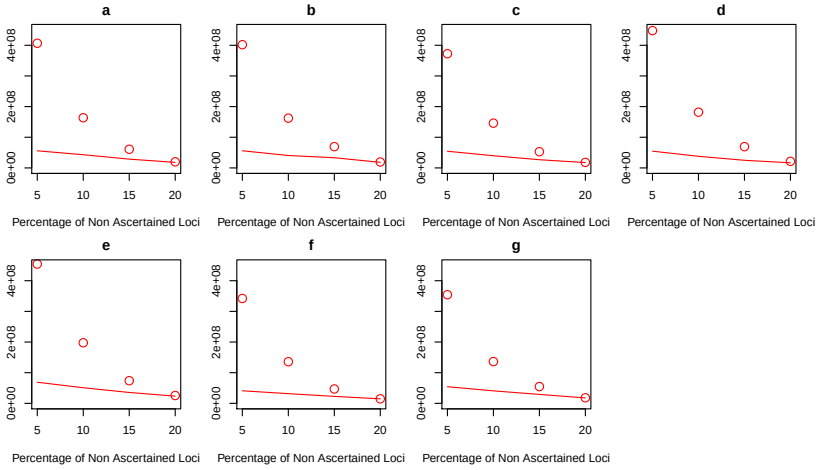

FIGURE 5.29: Mean squared error and variance of $\hat{L}$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

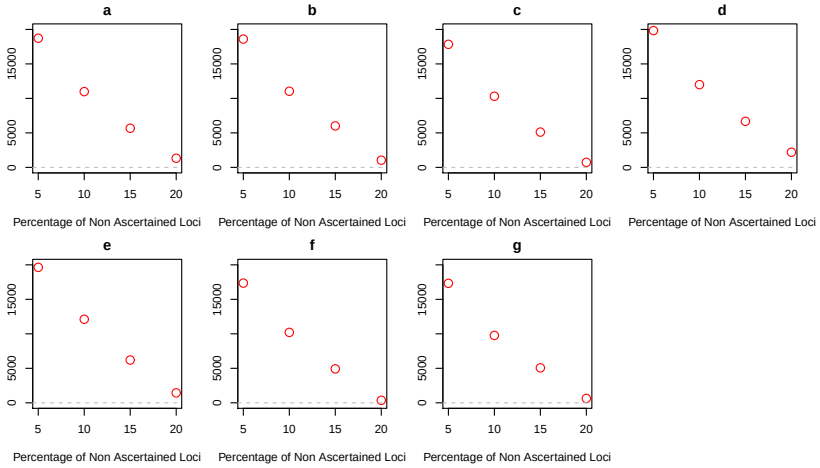

FIGURE 5.30: Bias of $\hat{L}$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

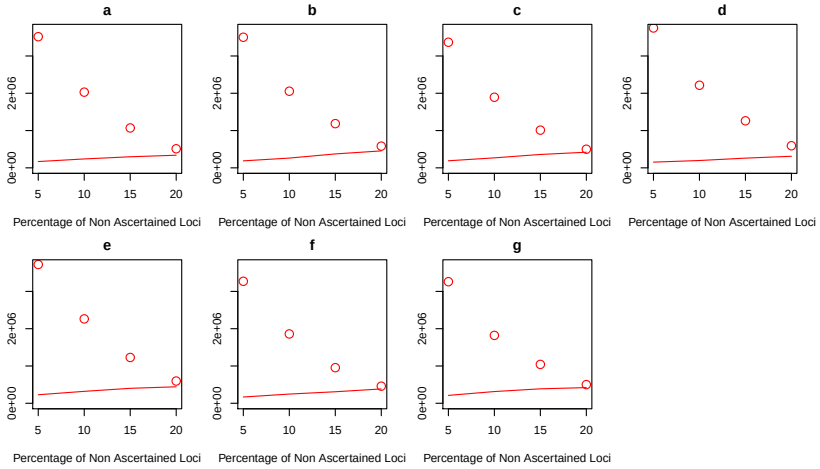

FIGURE 5.31: Mean squared error and variance of $\hat{N}_e$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$
(○/solid red line - MSE/variance)

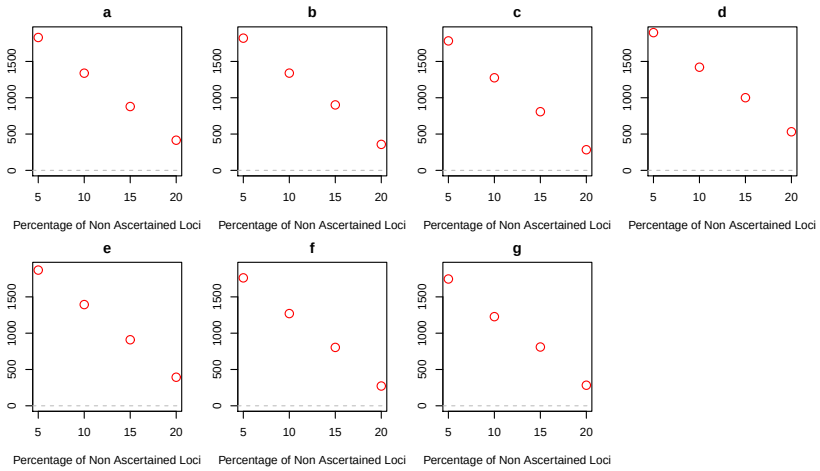

FIGURE 5.32: Bias of $\hat{N}_e$ vs. percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

5.4 Discussion

On the basis of the first set of results, it is clear that varying the percentage of typed loci affects the estimates of t greatly: increasing the amount of data reduces the amount of (positive) bias in $\hat{t}$. Also the estimates for β, the variance of the gamma, L (gens) and N_e are all overestimates whilst for α our method produces an underestimate. For the mean of the gamma the estimates are fairly accurately estimated with a small bias. Yet none of these additional estimators' biases are affected by the percentage of typed loci.

In contrast misspecifying the percentage of non-ascertained loci affects all the parameters estimates with the exception of $\hat{t}$ whose bias is only affected by the value of t. For α, β and the mean of the gamma, the percentage of non-ascertained loci at which the bias is closest to zero is from 5-10%. Plotting histograms of the actual percentage of non-ascertained loci across the 100 independent datasets at each t reveals why this should be the case. In figure 5.33 we see that the range lies between 0.07 and 0.15 for each t with a maximum lying at around 0.10-0.12. The distributions all appear fairly symmetric. The means for each t are, respectively, 0.105, 0.105, 0.107, 0.104, 0.108, 0.105 and 0.108.

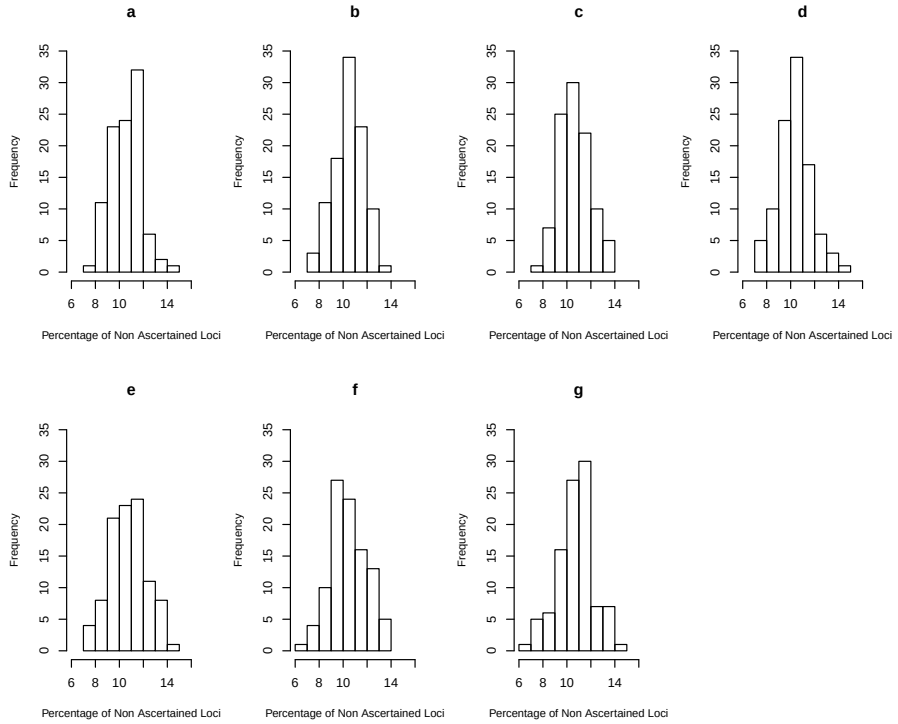

FIGURE 5.33: Histogram of percentage of non-ascertained loci: a. $t = 5$, b. $t = 10$, c. $t = 15$, d. $t = 20$, e. $t = 25$, f. $t = 50$, g. $t = 100$

In particular, recap that the variance of the gamma the amount of bias has an unusual pattern (fig. 5.28) and even with the true percentage of non-ascertained loci the estimator is upwardly biased (fig. 5.12). This is also the case for L and N_e although its bias appears reduced most when the percentage of non-ascertained loci is greatest (figs. 5.30 and 5.32).

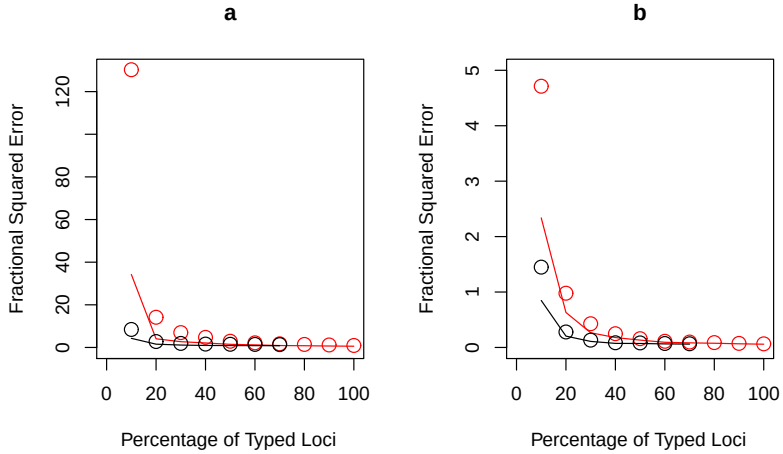

FIGURE 5.34: Fractional squared error and variance of $\hat{t}$ vs. percentage of typed loci: a. t =5, b. t =100
(○/solid red line - FSE/FV of random loci, ○/solid black line - FSE/FV of fast loci)

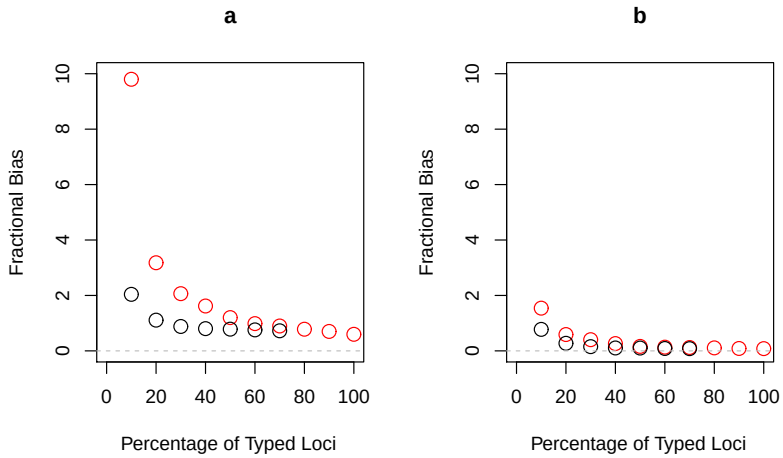

FIGURE 5.35: Fractional bias of $\hat{t}$ vs. percentage of typed loci: a. t =5, b. t =100
(○ - random loci, ○ - fast loci)

Next recap, the FSE of $\hat{t}$ when using the fast loci rather than random loci is considerably lower (figs 5.17 and 5.3). Here, random loci refers to the original order of the calibrated mutation rates whilst fast loci refers to the calibrated mutation rates sorted in descending order according to their empirical rate. For example, in the most extreme case when $t = 5$ we have an FSE of more than 120 for the random markers compared to only ∼10 for the fast loci (fig. 5.34). For the highest $t = 100$, this difference is much less; nonetheless the FSE for the random loci is

still higher. This would suggest that using the fastest markers would produce less biased results for $\hat{t}$.

Indeed, when we examine a plot of the fractional bias in $\hat{t}$ when $t = 5$, the value for the fast loci is only about a fifth of that when employing the random loci when only using 10% of the markers (fig. 5.35a). Indeed at least 20% more of the random typed loci (hence 30% of the loci) must be used to produce the equivalent level of fractional bias in $\hat{t}$. This difference is less stark at $t = 100$ (fig. 5.35b). Here the fractional bias for the random loci (○) at 10% is only about double that when using the fast loci (○). Also less than 20% of the random loci are required to produce the equivalent fractional bias at 10% for the fast loci.

5.5 Conclusions

Having developed a model for estimation of the TMRCA which incorporates the mutation rate model developed in chapter 4, we carried out a simulation study across a range of TMRCAs ($5 - 100$ generations). In addition, the percentage of typed loci out of the total calibrated loci was varied. It was apparent that this percentage affects the estimates of TMRCA. In general, and perhaps unsurprisingly, increasing the number of markers reduces the overestimation, irrespective of the true value of TMRCA. However for the remaining parameters, α, β, the mean and variance of the gamma distribution with shape α and scale β, the total branch length, L and the effective population size, N_e, varying the percentage of typed loci had little effect on their estimates across the seven values of TMRCA. On the other hand, estimates of TMRCA are not affected by misspecifying the percentage of non-ascertained loci, whilst the other parameters are affected to varying degrees.

Our results also indicate that the nature of the typed markers as well as the number of them affect the estimates of TMRCA for pairs of males. Typing fast STRs whose rates have been empirically estimated using a large number of meioses will reduce the amount of positive bias in TMRCA estimates when compared to a random set of STRs.

Chapter 6

TMRCA Estimation: Real Data Applications

6.1 Haplogroup- and Surname-Based Priors

The work of King et al. (2006) established that there is an inherent relationship between TMRCA and some measure of the frequency of a British surname. Specifically, King et al. highlighted that rare or less frequent surnames produced on average lower estimates of TMRCA between pairs of males sharing the same surname. This was in accordance with the historical evidence that rare surnames may have a single founder at the time of surname establishment. It was also clear that sharing the same haplogroup or not would entail a different relationship with the TMRCA. In addition factors such as the surname origin were considered. The aim of this section is to develop priors for TMRCA estimation based on the presence or absence of haplogroup information as well as surname frequency and surname origin information.

6.1.1 Materials and Methods

In order to incorporate a prior on TMRCA based on the surname frequency into our model, we amalgamated data from King et al. (2006) and King and Jobling (2009*a*), excluding in the former any surnames that overlap with the latter. For King and Jobling (2009*a*), we paired the data within each surname using three

sampling approaches, each without replacement. The first allowed random pairing of males within each surname ignoring the haplogroup of the males. The second strategy randomly paired those within the same haplogroup within each surnames. The final pairing involved randomly pairing those in different haplogroups within each surname. In particular for the different haplogroup sampling, sampling was carried out to avoid over-sampling the most frequent haplogroup. This was carried out by sampling within the haplogroup with the fewest samples, next randomly choosing another haplogroup with equal probability, thereafter taking a random sample within the chosen haplogroup. In total, the random haplogroup sampling resulted in 829 pairs, the same haplogroup sampling in 785 pairs and the different haplogroup sampling in 554 pairs.

It was necessary to include the earlier data from King et al. (2006) due to the fact that King and Jobling (2009*a*) largely sampled less frequent surnames and development of a prior based on the frequency of a surname would benefit from data across the full range of surname frequencies. The origins of the surnames in King et al. (2006) were researched and classified according to the categories described in King and Jobling (2009*a*): ambiguous/unknown, locative, nickname, occupational, patronymic/matronymic and topographic (Cottle, 1978; Hanks et al., 1988; Reaney and Wilson, 1997). The ambiguous/unknown category included any surnames with multiple origins of which several had both locative and topographic origins. Where the researched origin conflicted with the origin specified in King and Jobling (2009*a*), the latter origin was used. This was the case for four of the surnames that overlapped with King and Jobling (2009*a*). The frequency of the surnames was consistent with those reported in King et al. (2006). 39 of the surnames in King and Jobling (2009*a*) were also sampled in King et al. (2006), with one of the surnames, Haythornthwaite, including the variant spellings 'Hawthornthwaite' and 'Haythorn'. A total of 40 pairs were excluded from the King et al. (2006) data due to the potential overlap with samples in King and Jobling (2009*a*). The remaining 110 pairs were added to random haplogroup data from King and Jobling (2009*a*). Of these, 56 pairs shared the same haplogroup and were added to the appropriate dataset sampled from King and Jobling (2009*a*). An analogous strategy was employed for the 54 pairs which belonged to different haplogroups. Thus, in total, the random haplogroup class consisted of 939 pairs, the same haplogroup class of 841 pairs and the different haplogroup class of 608 pairs.

Estimation of TMRCA was carried out using the method described in sections 5.2.2

and 5.2.3 for each pair. In particular, $n = 475$, the percentage of non-ascertained loci was 14.4% and the ascertainment sample size was eight in accordance with Kayser et al. (2004). The empirical mutation rates were based on the final mutation rate review in section 2.3.3. We applied 1000 batches of 50 updates with 10% burn-in of our MCMC sampler with uniform priors on the mutation rate distribution parameters and TMRCA to produces the estimates of TMRCA for each pair.

For each data set (random/same/different haplogroup) an initial analysis of covariance (ANCOVA) was carried out for the estimates TMRCA, or some transformation of it (see below), the covariate being the natural logarithm of the surname frequency and the grouping variable surname origin. Both interaction and main effects models were examined. Where appropriate a regression model was used, as well as a one-way analysis of variance (one-way ANOVA). Checking the assumptions of linearity, homoscedasticity (constant variance) and normality were carried out by examining plots of standardized residuals versus fitted values and normal Q-Q plots.

The resulting fitted models only provide point estimates for the TMRCA based on the haplogroup information and/or surname frequency and surname origins. As such we will first consider an exponential distribution with the scale parameter equal to the fitted TMRCA for the prior on TMRCA.

The resulting prior should ideally have a standard deviation similar to the fitted model and sample. To investigate this, random draws from the proposed exponential distribution were taken and compared to 10,000 simulated data points from the fitted model and also of the sample. For the former, the variance of the error term was based on the fitted model's $\hat{\sigma}^2$ and for the latter we computed the standard deviation across variant spelling of surnames with more than one pair of observations. Where the exponential's standard deviation was not adequate, a gamma distribution was also examined. In this situation the mean was the fitted TMRCA and its variance reflected the variance of data simulated from the fitted model. Alternatively, we may fit an exponential or gamma distribution to the raw estimates of TMRCA by the method of moments.

Since King and Jobling (2009*a*) largely analysed surnames which are less common it is possible that there is underrepresentation of more common surnames. Consequently, it may be desirable to thin the data set from King and Jobling (2009*a*)

since there is data on only one pair of males per surname from King et al. (2006). Hence, for each unique surname we randomly sampled one pair of males and fitted an appropriate distribution such as an exponential or gamma. This process was carried out 1,000 times allowing statistics such as the mean and standard deviation of the fitted parameters to be computed.

Where an exponential prior distribution has been fitted, to make the prior less influential, we used a deflated rate, namely the mean rate minus two standard deviations of the 1000 rates. Similarly, when fitting a gamma distribution, we inflated variance. This was achieved by firstly computing the thinned means and thinned variance which are based on the fitted shapes and rates for each thinned data set. Thereafter the mean of the resultant gamma is put equal to the mean of the thinned means whilst the variance is put equal to the mean variance plus two standard deviation of the thinned variances. In addition if another distribution is fitted, its variance will similarly be inflated.

Given the sampling of males within surnames, the assumption of independence for linear models may be violated. As such ANCOVA permutation tests were carried out to determine if any interaction models were indeed significant, based on the concept of removing the functional relationship (Gail et al., 1988; Good, 2000). More specifically, this involved fitting only the main effects and saving the residuals from this model. Supposing that the null hypothesis is true, i.e. the slope parameters are the same across the groups, then there should be no structure left in the residuals. To test this we fit another ANCOVA with the response the saved residuals and the covariate and grouping variable as before. The group labels would then be permuted 10,000 times, each time fitting a new ANCOVA and saving the slope coefficients for each group. These values would then be compared to the slope coefficients from the original main effect residuals to give a p-value for each group assessing whether the slopes are significantly different to zero (Sundaresan et al., 2007). The significance level is adjusted to allow for multiple comparisons using the Bonferroni correction, i.e. for n comparisons, the α_{new} value would be α/n where α is typically 0.05. The choice of using the individual slope coefficients as the test statistic of choice rather than the F statistic as suggested by Edgington (1995) was due to inconsistent results between the two. To validate this we carried out a simulation study whereby we chose the regression lines in table 6.1 for six groups. Note that group A has a different slope from the other groups. In addition all the groups have different intercept parameters. We generated data by simulating

TABLE 6.1: Model for validating ANCOVA permutation test

Group	Intercept	Slope coefficient
A	33.04	0.05
B	42.45	0.22
C	82.32	0.22
D	12.96	0.22
E	72.75	0.22
F	67.81	0.22

random normal noise. We then fitted an interaction ANCOVA, thereafter carried out a permutation ANCOVA with 10,000 permutations saving the associated p-value for each group's slope coefficient as well as the F statistic. This process was repeated 100 times and the p-value of the F statistic was compared to the usual significance level whilst the slope p-values were compared to the Bonferroni corrected significance level of 0.0083. The results were summarised by computing the type I and II error rates where appropriate. Table 6.2 shows that the type

TABLE 6.2: Simulated data error rates

Parameter	Type II Error
Slope A	0.46
F Statistic	0.77

II error rates. For group A the type II error rate was 0.46, but based on the F statistic, it was even higher at 0.77.

Where a simple linear regression model was fitted, permutation tests were used to determine whether a main effects ANCOVA model was significant. Here the residuals from the linear regression model were permuted by group label 10,000 times and fitted to a one-way ANOVA. For each permutation, the F statistic was recorded and compared to that obtained from a one-way ANOVA of the residuals with the observed group labels. A p-value less than 0.05 would indicate that the group effect was significant and hence a main effects ANCOVA more suitable.

In addition, a permutation simple linear regression was carried out to test the null hypothesis that the slope equals zero. The underlying principle is that if the null were true then the responses could have been observed in any order (Anderson, 2001; Legendre and Legendre, 1998). It involves permuting the response variables and computing the Pearson correlation coefficient 10,000 times. The permuted correlation coefficients are then compared to the observed coefficient in order to

produce a p-value. In this case a p-value less than 0.05 would imply that the slope is not equal to zero, so the simple linear regression is more appropriate.

Finally to test whether a one-way ANOVA model was significant, i.e. the null hypothesis being that the groups means are not different, the group labels were permuted and a one-way ANOVA was fitted (Anderson, 2001). The F statistic was computed for 10,000 permutations and compared to the original F statistic to obtain a p-value.

In instances where the raw data did not appear normal, the above analysis was carried out using an appropriate transformation such as the (natural) logarithm or Box-Cox. The latter has the form:

$$\frac{x^{\lambda} - 1}{\lambda} \tag{6.1}$$

for a variable x where λ is referred to as the Box-Cox lambda. Its maximum likelihood estimate was obtained using the R function `boxcox()`.

6.1.2 Same haplogroup

6.1.2.1 Surname Origins and Frequency

For pairs of males sharing the same haplogroup, it was clear from the raw data that to discern a meaningful relationship between the estimated TMRCA in generations and the surname frequency some transformations were necessary. Thus as well as the log of the surname frequency, a Box-Cox transformation of TMRCA with the Box-Cox lambda equal to 0.0606 was applied. The resulting relationship between the variables looks linear although this is somewhat obscured by the overlap of points particularly at lower values of the Box-Cox TMRCA (figure 6.1). Taking the mean value for each unique surname results in figure 6.2. A clear negative linear relationship appears between the Box-Cox TMRCA and the reverse log surname frequency, which is also apparent for some of the surname subtypes. However it is important to note that many surnames were categorised into the ambiguous/unknown category and therefore the relationship between the variables in this case may be unreliable. Also there is a notable absence of surnames of locative origin on the left side of figure 6.2, and there are only six unique surnames of topographical origin, amounting to just over 4% of the same haplogroup data.

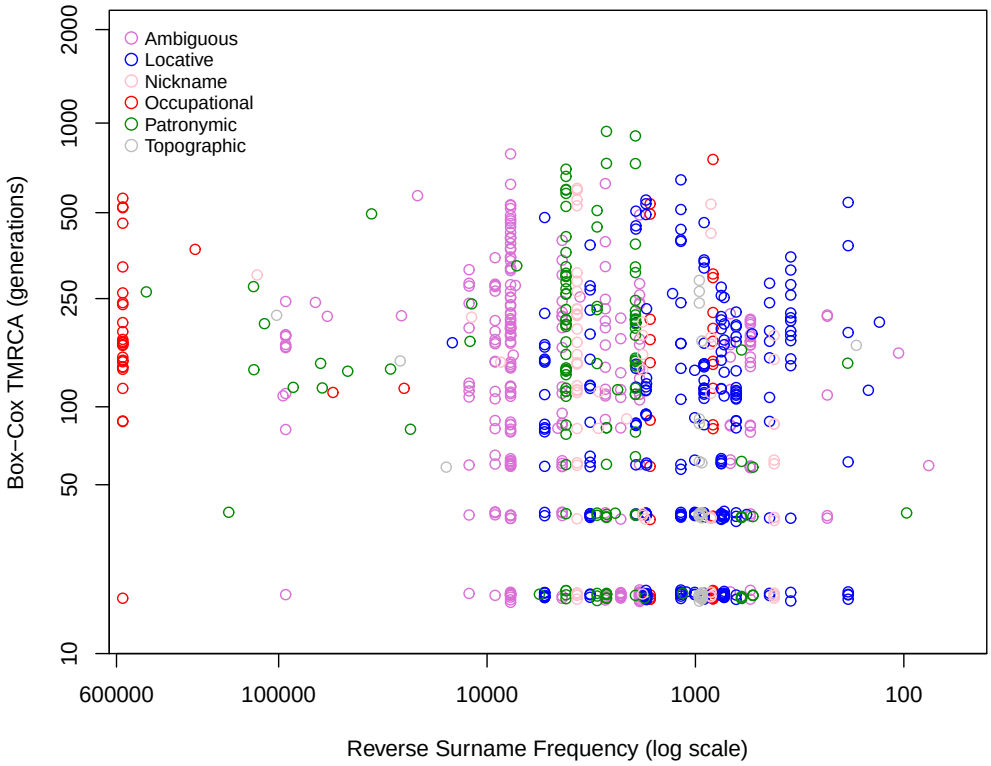

FIGURE 6.1: Same haplogroup: Box-Cox TMRCA vs. log surname frequency

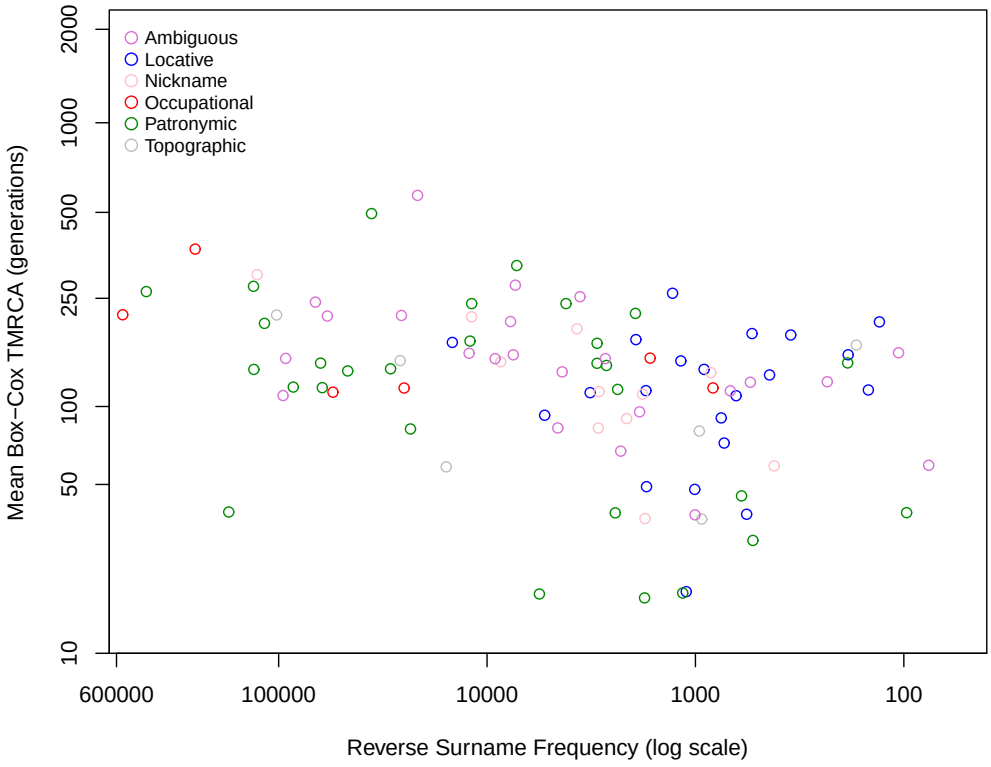

FIGURE 6.2: Same haplogroup: mean Box-Cox TMRCA vs. log surname frequency

Nonetheless a main effects and interaction ANCOVA for the response Box-Cox TMRCA, covariate log surname frequency and grouping variable surname origins was fitted. The interaction term for this model is significant (p-value = 0.00059) and the resulting fitted model is:

$$\hat{t} = \left(0.0606[\alpha_i + \beta_i \log(S_f)] + 1\right)^{\frac{1}{0.0606}}, \tag{6.2}$$

where S_f is the surname frequency and $i = 1, \ldots, 6$ represents the surname origins ambiguous/unknown, locative, nickname, occupational, patronymic and topographic, respectively, with the parameters as shown in table 6.3. For all but

TABLE 6.3: Same haplogroup: surname origin and frequency fitted model parameters

Surname Origin	α	β
Ambiguous/Unknown	3.599	0.211
Locative	6.085	-0.177
Nickname	0.904	0.565
Occupational	3.187	0.225
Patronymic	2.725	0.330
Topographic	1.914	0.326

one surname origin the estimated slope is positive resulting in a negative linear relationship between the response and the reverse log surname frequency. For surnames of locative origin the slope is negative. This appears to be counterintuitive to the notion that rare surnames should have a lower TMRCA to reflect the increased likelihood that those surnames have a single founder and that there has not been enough time for the name to reach high frequency. The lack of high frequency data in the locative surnames may be a factor in influencing this result, which is unsurprising given that most locative surnames tend to be less common (McKinley, 1990).

Furthermore, the assumptions of ANCOVA may be violated. In figure 6.3 we see that constant variance may not hold for the residuals. The banding reflects the nature of the data, i.e. the non-negative integer number of mutational steps between the pairs of males, as well as the discrete values for the raw surname frequency. In addition, the normal Q-Q plot show deviation from normality at the extremes. Given the potential violations of the underlying modelling assumptions, we carried out a permutation ANCOVA to examine if any of the slopes for the different surname origins were indeed significant. The final fitted model is shown in figure 6.4.

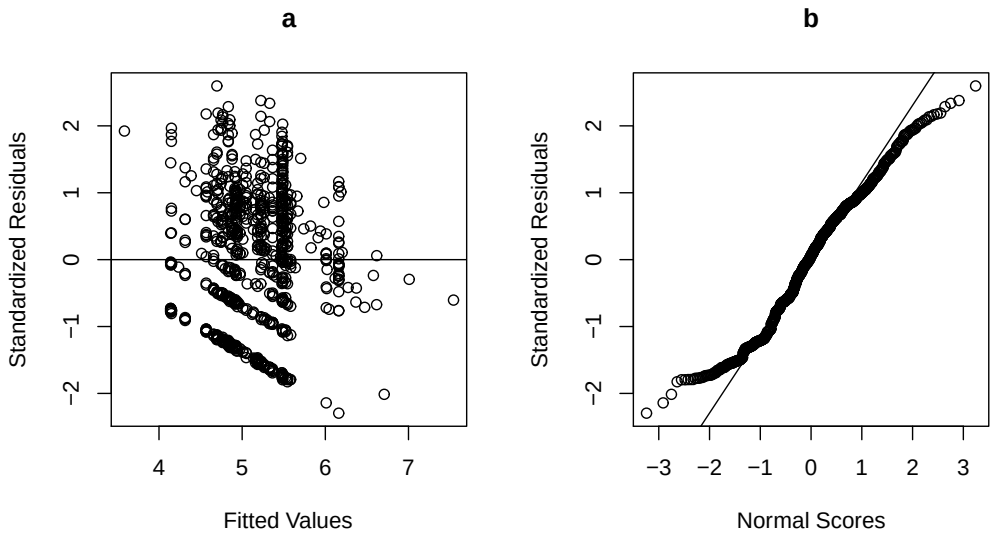

FIGURE 6.3: Same haplogroup: a. residuals vs. fitted values b. normal Q-Q plot of residuals

The results of the permutation test are given in table 6.4 which we compare to the Bonferroni-corrected significance level of 0.0083. We find two surname origins, locative and nickname, have slopes significantly different from zero, indicating that an interaction model is most suitable for describing the same haplogroup data. Next we consider the exponential prior for the TMRCA with its rate parameter

TABLE 6.4: Same haplogroup: surname origin and frequency permutation test p-values

Surname Origin	Permutation p-value
Ambiguous/Unknown	0.9429
Locative	0.0000
Nickname	0.0004
Occupational	0.8463
Patronymic	0.0397
Topographic	0.3584

derived from the fitted TMRCA based on the interaction ANCOVA. The standard deviation appears lower than both the sample and fitted model standard deviations in this case. This is the case across all the surname origins, and is illustrated in figure 6.5 for the patronymic surnames. So instead a gamma prior for the TMRCA was chosen (fig. 6.5). The final prior for pairs of males sharing the same haplogroup where information on both surname frequency and origin are available will have a gamma prior with the shape and rate parameters based on the fitted model as show in equation 6.2 with parameters in table 6.3 and $\sigma^2 = 1.834$, the estimated error variance in the Box-Cox transformed TMRCA.

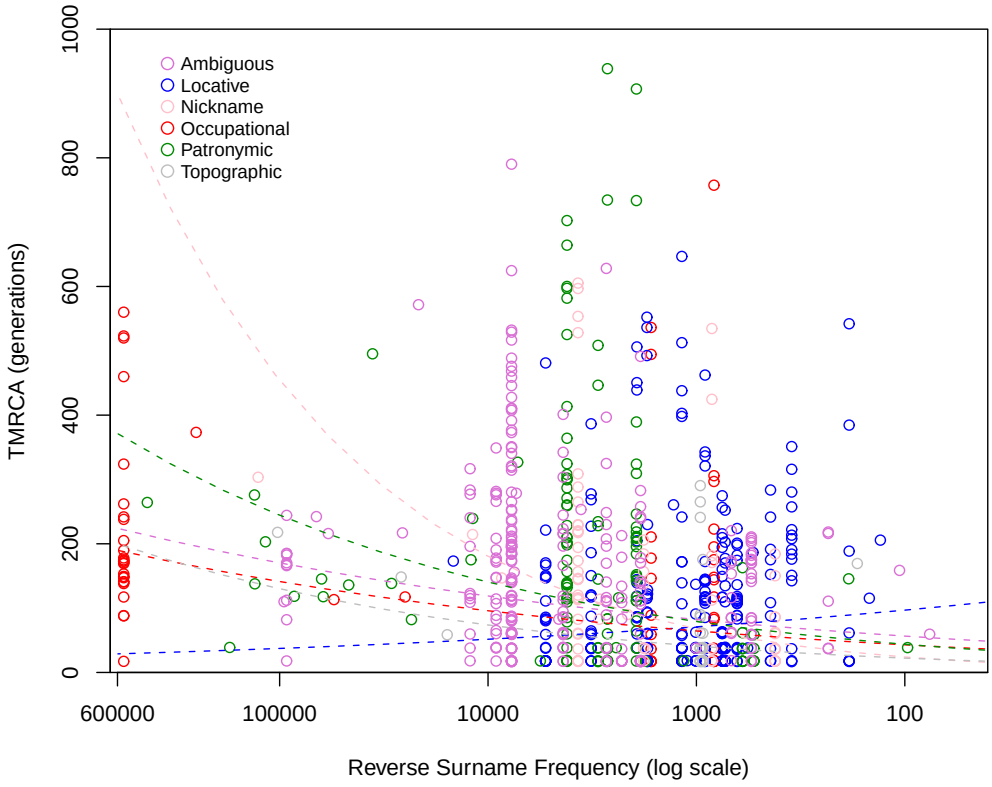

FIGURE 6.4: Same haplogroup: surname origin and frequency fitted model

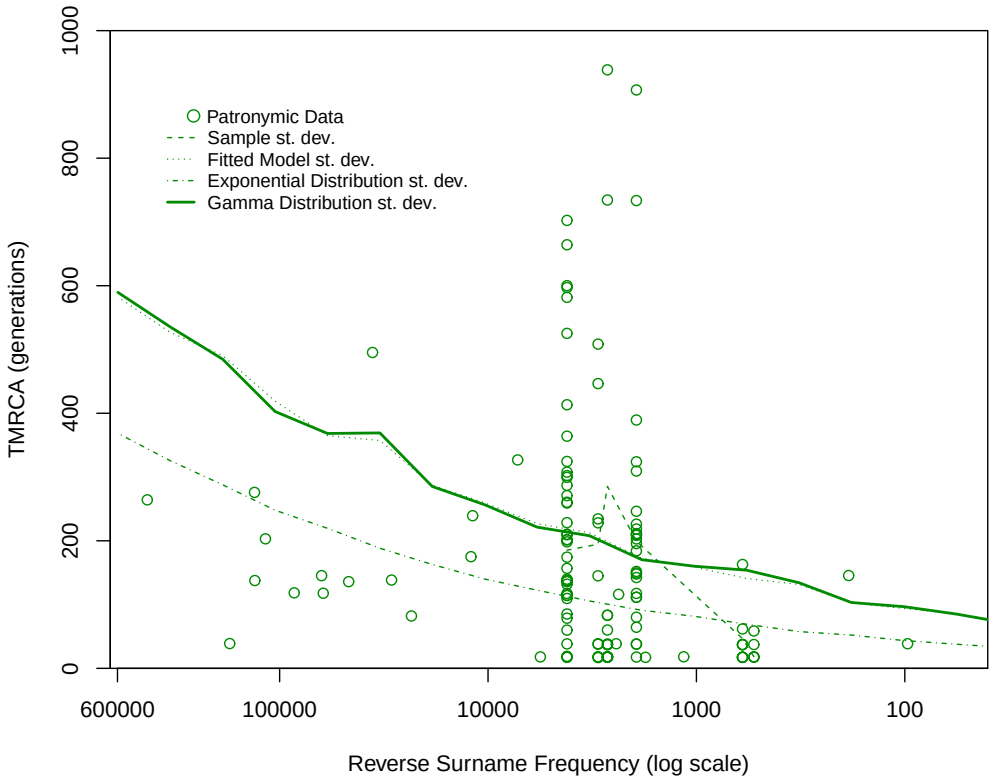

FIGURE 6.5: Same haplogroup: standard deviation for patronymic TMRCA surname origin and frequency priors

6.1.2.2 Surname Frequency

For the same haplogroup data where surname origin information is lacking we used regression to quantify the relationship between the Box-Cox transformed TMRCA and log of surname frequency. The Box-Cox λ was 0.0606 as before and the final fitted model is summarised below:

$$\hat{t} = \left(0.0606[3.316 + 0.231 \log(S_f)] + 1\right)^{\frac{1}{0.0606}}, \tag{6.3}$$

where S_f is the surname frequency.

The explanatory was statistically significant (p-value < 0.001). The assumption of constant variance is doubtful although linearity is fine. However the residuals seem to depart from normality more so than in section 6.1.2.1 (data not shown). The final model is displayed in figure 6.6. A permutation test was also carried out testing the null hypothesis that the slope equals zero. In this case the correlation coefficient permutation p-value was equal to zero, so the slope is significantly different to zero.

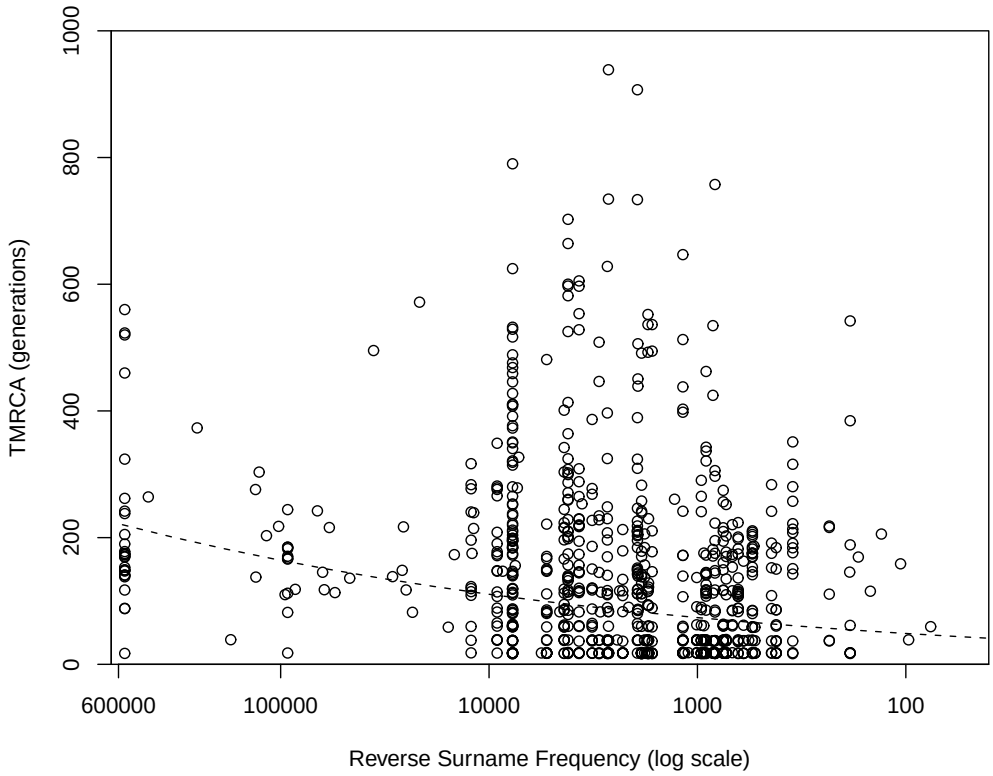

FIGURE 6.6: Same haplogroup: surname frequency fitted model

As with the previous prior, the prior was chosen as a gamma rather than as exponential since the latter did not have sufficient spread (fig. 6.7). The scale and shape parameters for the gamma were based on the fitted TMRCA (6.3) and $\sigma^2 = 1.901$.

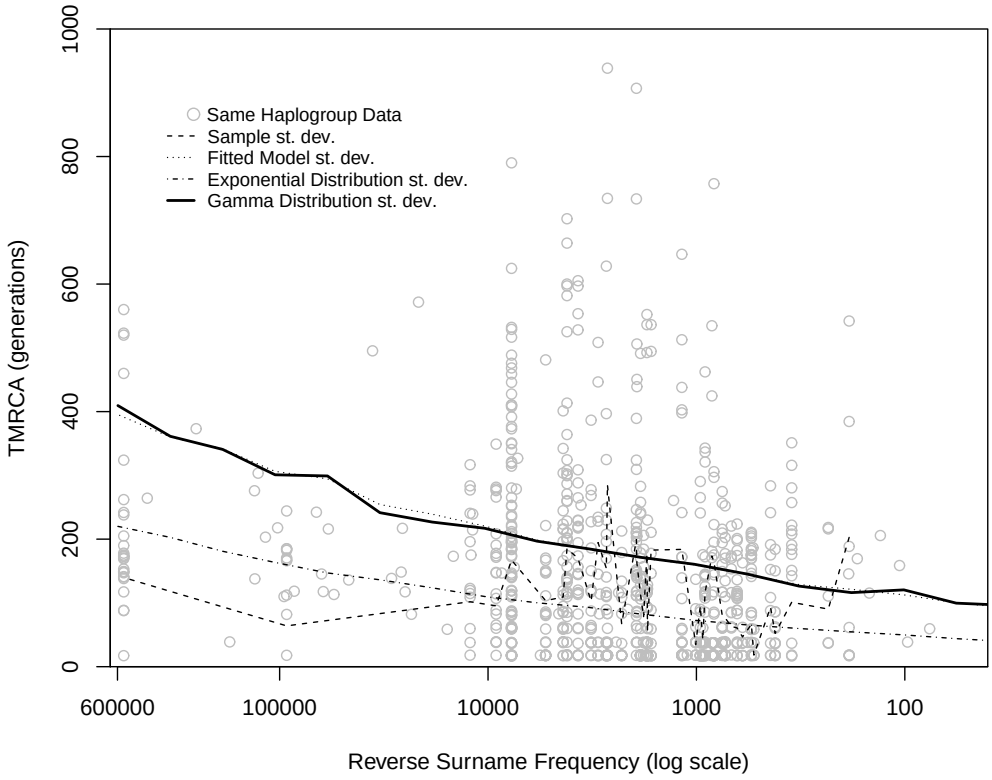

FIGURE 6.7: Same haplogroup: standard deviation for surname frequency prior

6.1.2.3 Surname Origins

In the context where two males share the same surname and haplogroup and we only have access to their surname origin the following prior on TMRCA was developed. We initially examined the relationship of the Box-Cox transformed estimates of TMRCA (Box-Cox $\lambda = 0.0606$) with surname origins (fig. 6.8). Here we note that the data is divided into two groups: the locative and topographic surnames have similarly right-skewed TMRCA estimates unlike the remaining surname origins which appear more symmetric and if anything left-skewed. The topographic surnames have the lowest median Box-Cox TMRCA followed by the locative surnames, with the remaining surnames origins having a similar rather high value.

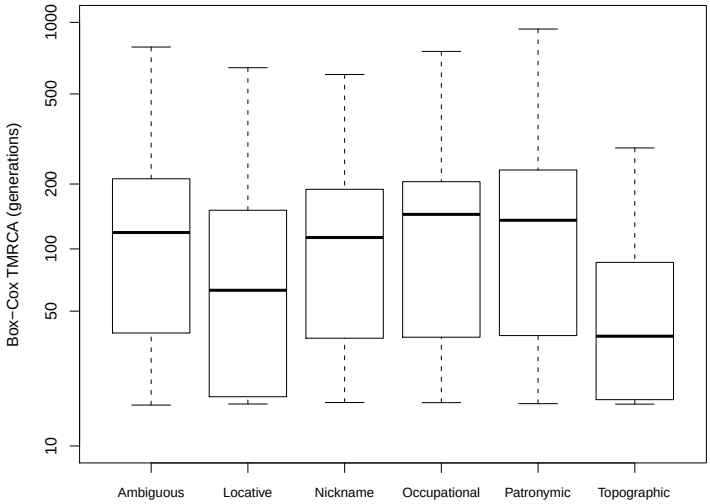

FIGURE 6.8: Same haplogroup: boxplot of Box-Cox TMRCA by surname origin

Carrying out a one-way ANOVA results in a p-value less than 0.001: surname origins are a statistically significant predictor of TMRCA. The assumption of constant variance is reasonable but normality is questionable (data not shown). Tukey multiple comparisons show that the significant differences in mean TMRCA occur between comparisons of either the locative or topographic surnames with the other surname origins only (table 6.5). A permutation test of the surname origins resulted in a significant p-value of 0.

TABLE 6.5: Same haplogroup: significant Tukey multiple Comparisons between surname origin

Surname Origins		95% Confidence Interval of Difference		Adjusted p-value
Group 1	Group 2	Lower	Upper	
Nickname	Topographic	0.087	1.750	0.0207
Ambiguous	Topographic	0.391	1.844	0.0001
Occupational	Topographic	0.295	1.979	0.0017
Patronymic	Topographic	0.400	1.938	0.0002
Ambiguous	Locative	0.167	0.852	0.0003
Patronymic	Locative	0.133	0.991	0.0026

Having established the need to model each surname origin separately, we now develop the prior in this case. We began by fitting an exponential and gamma based on the ANOVA results, both of which were found to be rejected by a Kolmogorov-Smirnov test of goodness of fit to the data across all surnames.

As such for each surname origin we fitted an exponential and gamma to the raw TMRCA estimates using the method of moments. The resulting fitted parameters are shown in tables 6.6 and 6.7 for the exponential and gamma respectively. In

TABLE 6.6: Same haplogroup: exponential surname origin fitted model parameters

Surname Origin	Rate	K-S p-value
Ambiguous/Unknown	0.00650	0.0064
Locative	0.00878	0.0000
Nickname	0.00695	0.2554
Occupational	0.00593	0.2502
Patronymic	0.00551	0.0511
Topographic	0.01420	0.0555

the exponential case, the Kolmogorov-Smirnov null hypothesis is rejected for the ambiguous and locative surnames only, whilst this is extended to include the topographic surnames for the gamma (tables 6.6 and 6.7). As such there is a preference to fit an exponential despite the fact that the fitted exponential has a standard deviation greater than the sample standard deviation in only three surname classes. It is possible that over sampling of rarer, e.g. topographic and locative, may be

TABLE 6.7: Same haplogroup: gamma surname origin fitted model parameters

Surname Origin	Shape	Rate	K-S p-value
Ambiguous/Unknown	1.3386	0.00870	0.0035
Locative	0.8825	0.00775	0.0000
Nickname	0.9628	0.00670	0.2459
Occupational	1.1033	0.00654	0.1383
Patronymic	0.9040	0.00498	0.1104
Topographic	0.7637	0.01084	0.0099

having an effect. As such we thinned the data as described in section 6.1.1 to generate 1000 thinned samples. To each, we fitted the exponential and gamma distributions. Although the distributions of the fitted exponential rates are fairly symmetric, there is a particularly unusual distribution for the topographic surnames (data not shown). Further examination showed that the same haplogroup topographic data consist of only six unique surnames, of which four had only a single observation each throughout the thinning process, and this was the cause of the strange distribution. Since we wish our priors to be conservatively broad the final rate parameter for all the surnames were deflated resulting in the values shown in table 6.8. The number of times the Kolmogorov-Smirnov null hypothesis is rejected across the 1000 thinned data sets is low for all surname origins.

Consequently for males who share the same haplogroup as well as the same surname, we choose an exponential prior to model the TMRCA across each surname origin with the rates shown in table 6.8.

TABLE 6.8: Thinned same haplogroup: exponential surname origin fitted model parameters

Surname Origin	Rate	K-S Null Rejection
Ambiguous/Unknown	0.00492	10
Locative	0.00549	5
Nickname	0.00524	0
Occupational	0.00321	0
Patronymic	0.00560	0
Topographic	0.00656	0

6.1.2.4 No Frequency or Surname Origin

The final prior we develop in this section is in the context where we have no additional information aside from the shared haplogroup status and surname of pairs of males. This involved modelling the TMRCA estimates alone. Initially, we examined a boxplot of the raw TMRCA (fig. 6.9a), finding an extremely right-skewed distribution. Applying a natural log transformation results in a much more symmetric distribution(fig. 6.9b). However, a histogram of the logged TMRCAs shows two modes (fig. 6.9c). As a consequence, we initially fitted both an exponential and a gamma to the raw TMRCA, but both were rejected by a Kolmogorov-Smirnov test.

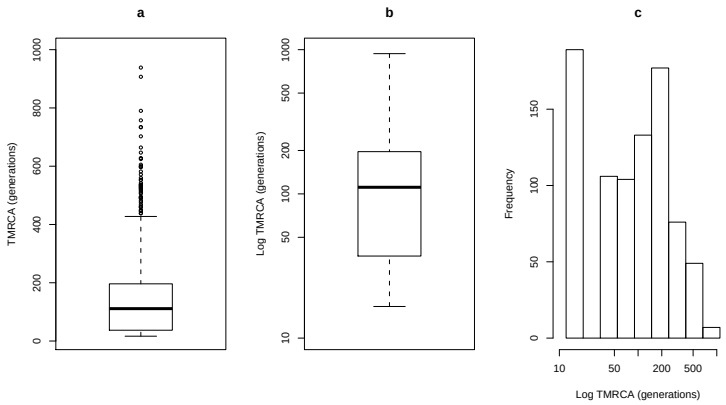

FIGURE 6.9: Same haplogroup: a. boxplot of TMRCA b. boxplot of log TMRCA c. histogram of log TMRCA

Thereafter we fitted an exponential to the TMRCA from 1000 thinned data sets with significant lack of fit for 452 by the Kolmogorov-Smirnov test. We also fitted the gamma distribution to the thinned data sets. In this case, the Kolmogorov-Smirnov goodness of fit null hypothesis was rejected less often (202 times). Thus the gamma was overall a better fit than the exponential. However, the histograms of the thinned data showed a rather half-normal appearance (as can be seen for a subset of the thinned data sets in figure 6.10). As such we fitted a half normal

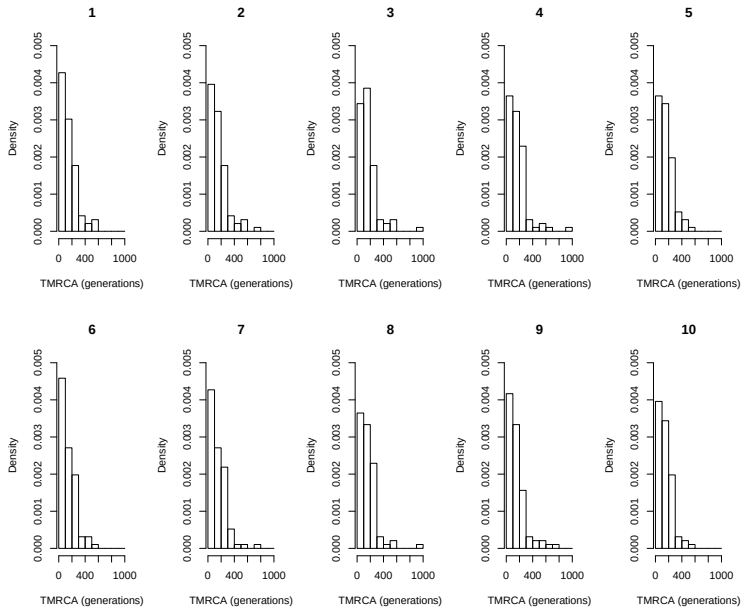

FIGURE 6.10: Ten thinned same haplogroup samples

to the full data which nonetheless rejected the Kolmogorov-Smirnov test (p-value < 0.001). We proceeded to fit the half normal to 1000 thinned samples resulting in a scale parameter of $\theta = 0.005731$. Still, for 227 of the 1000 thinned samples, the fit was rejected. The exponential and gamma distributions were also consid-

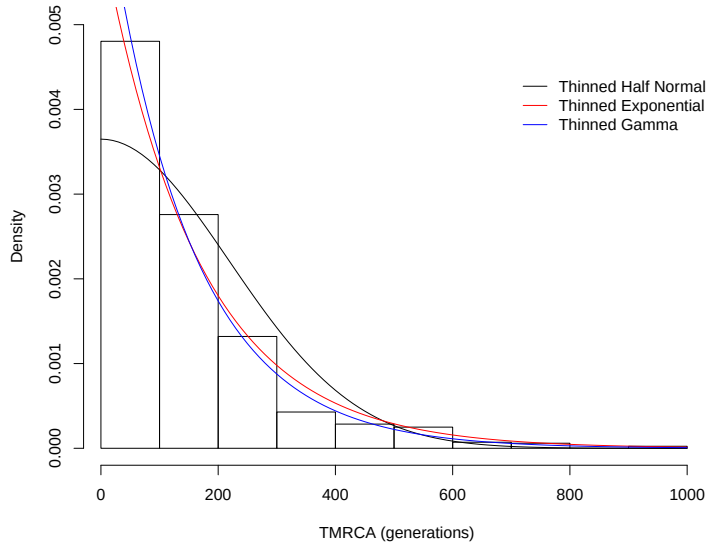

FIGURE 6.11: Same haplogroup: comparison of exponential, gamma and half-normal distributions

ered. Comparisons of the three priors together with a histogram of the TMRCAs is shown in figure 6.11. The exponential and gamma have fairly similar distributions which do not give as much weight to intermediate values of TMRCA. On the other hand, the half normal gives more weight to intermediate values and correspondingly less weight to lower values of TMRCA. Nevertheless, we chose a half

normal distribution to model the TMRCA for pairs of males who share both their surname and haplogroup status and no additional information on their surnames is available. In addition, results were not sensitive to the choice of the exponential and gamma based on limited exploration.

6.1.3 Random/No Haplogroup

Where no haplogroup information is available when estimating the TMRCA for pairs of males sharing the same surname, we have developed a prior based on the data of King and Jobling (2009*a*) and King et al. (2006) by random pairing of males within each surname as described in section 6.1.1. As with the same haplogroup data a Box-Cox transformation of the response was necessary with $\lambda = 0.2626$.

6.1.3.1 Surname Origin and Frequency

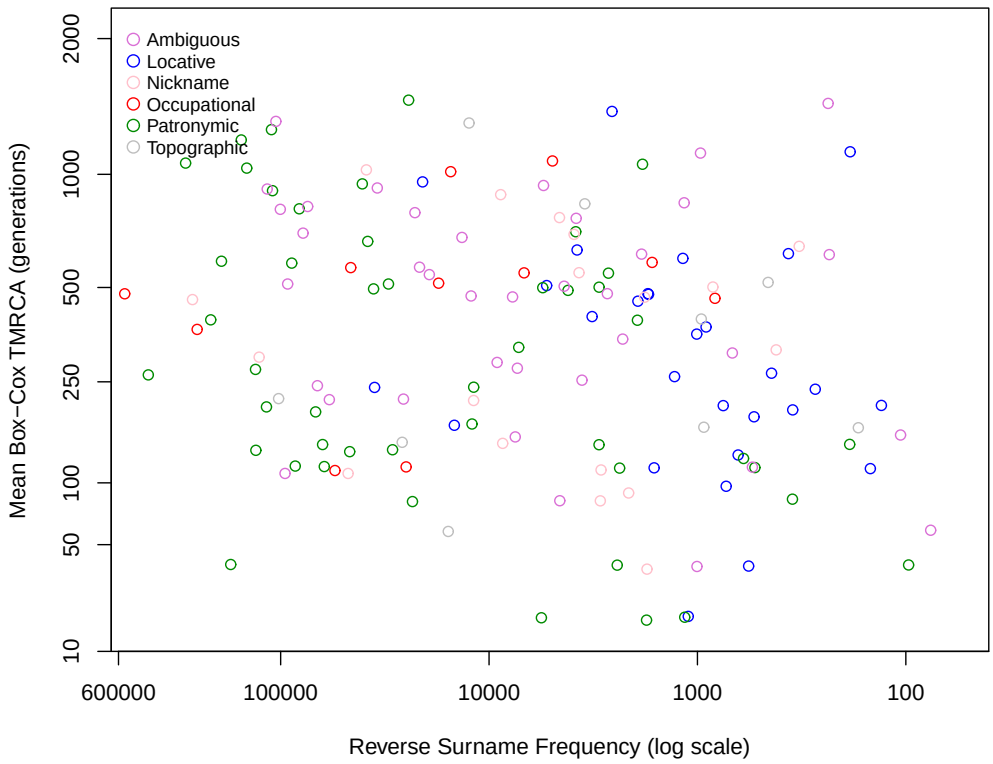

FIGURE 6.12: Random haplogroup: mean Box-Cox estimated TMRCA vs. log surname frequency

Taking the mean estimated TMRCA for each unique surname shows that there is a slightly negative linear relationship between the response and covariate (figure 6.12). There is also considerable overlap between the different surname origins. Again we began by modelling the data by fitting a main effects and interaction ANCOVA with the surname origin as the grouping variable. In this case the interaction term was only just significant (p-value = 0.04259). Constant variance holds as does linearity, but normality is doubtful (data not shown).

For this reason a permutation test was used to determine if the interaction is indeed significant. In this case the resulting p-values for each surname origin are given in table 6.9. All are above the adjusted Bonferroni multiple comparisons significance level of 0.0083 suggesting that the interaction term is not necessary. As such the

TABLE 6.9: Random haplogroup: permutation test p-values

Surname Origin	Permutation p-value
Ambiguous/Unknown	0.1771
Locative	0.0505
Nickname	0.4344
Occupational	0.1187
Patronymic	0.0598
Topographic	0.9478

main effects ANCOVA was fitted and both main effect were statistically significant (p-values < 0.001). The resulting fitted model is:

$$\hat{t} = \left(0.2626[\alpha_i + 0.5375 \log(S_f)] + 1\right)^{\frac{1}{0.2626}}, \tag{6.4}$$

where S_f is the surname frequency and $i = 1, \ldots, 6$ represents the surname origins ambiguous/unknown, locative, nickname, occupational, patronymic and topographic, respectively, with the parameters as shown in table 6.10.

TABLE 6.10: Random haplogroup: fitted surname origin and frequency model parameters

Surname Origin	Intercept
Ambiguous/Unknown	9.162
Locative	9.106
Nickname	8.503
Occupational	8.194
Patronymic	9.173
Topographic	6.958

Figure 6.13 shows the fitted model for the raw TMRCA and this is the prior that is used for t in the context where we have a pair of males of unknown haplogroup but with the same surname and both surname origin and frequency are available.

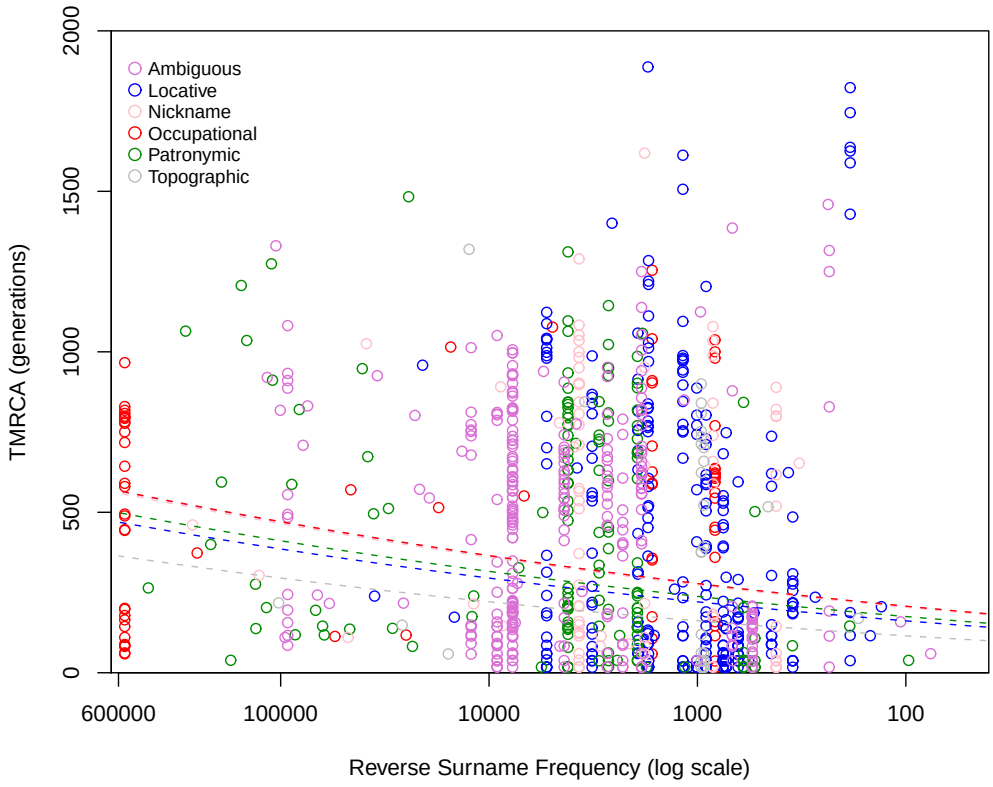

FIGURE 6.13: Random haplogroup: fitted surname origin and frequency main effects model

6.1.3.2 Surname Frequency

Next we wished to develop a prior based solely on the log of surname frequency, where no haplogroup information is available nor is the surname origin. We modelled the TMRCA using regression and found that the predictor was significant (p-value < 0.001). The modelling assumption of normality is however doubtful, whilst constant variance and linearity appear to hold (fig. 6.14). The fitted model is given by:

$$\hat{t} = \left(0.2626[7.7948 + 0.6470\log(S_f)] + 1\right)^{\frac{1}{0.2626}}, \tag{6.5}$$

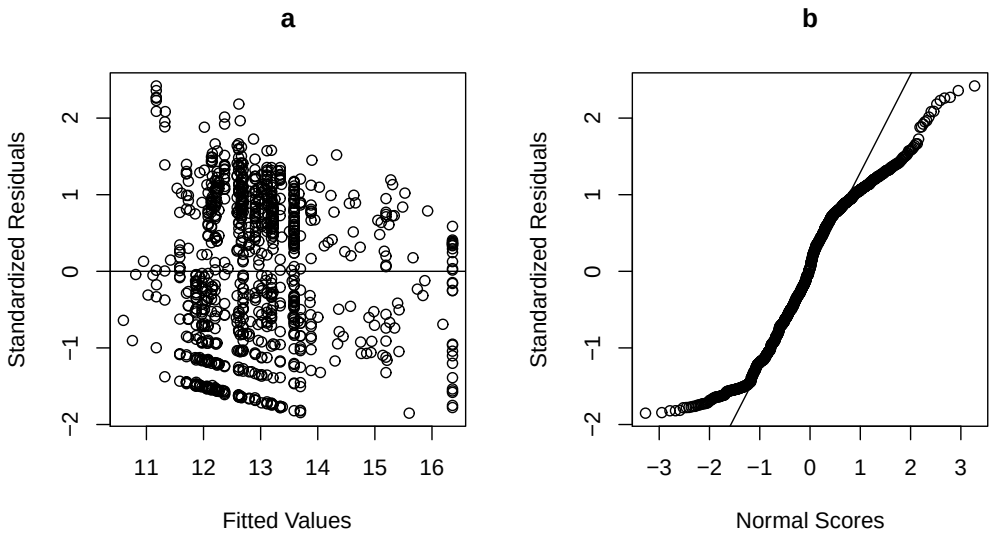

FIGURE 6.14: Random haplogroup: a. residuals vs. fitted values b. normal Q-Q plot of residuals

and shown in figure 6.15. A permutation test was carried out to check that the slope in this case was significant, which turned out to be the case.

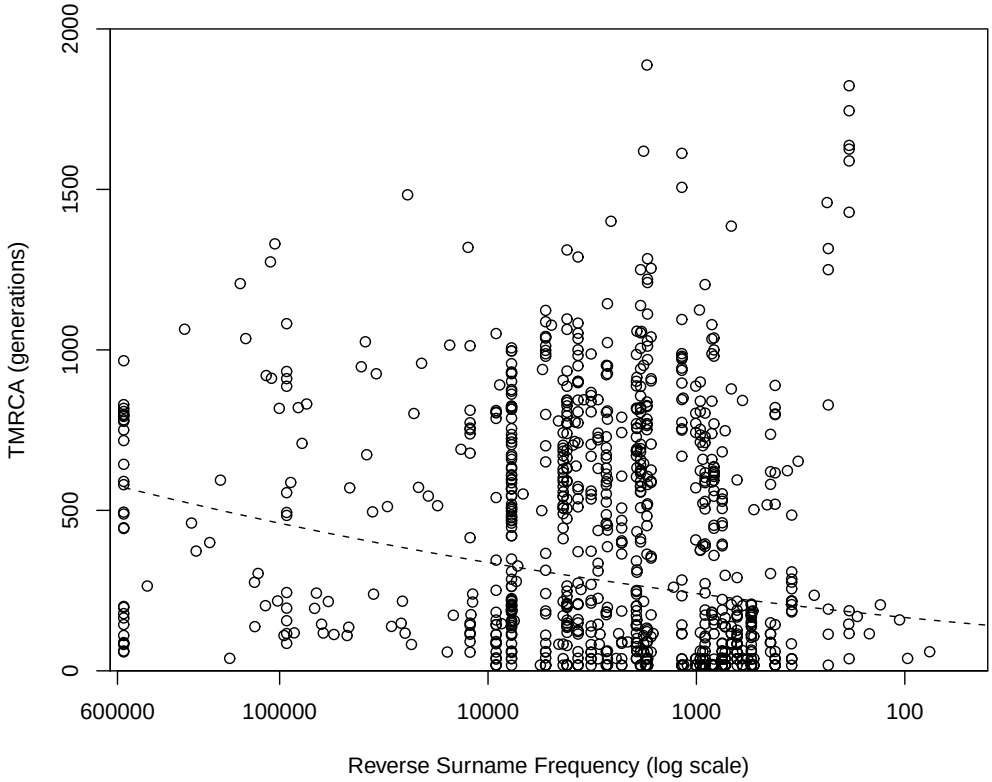

FIGURE 6.15: Random haplogroup: surname frequency fitted model

As such the final prior for when no haplogroup information exists for pairs of surname-sharing males where surname frequency is given is based on the fitted model (6.5) with $\sigma^2 = 26.261$. Although an exponential prior with scale equal

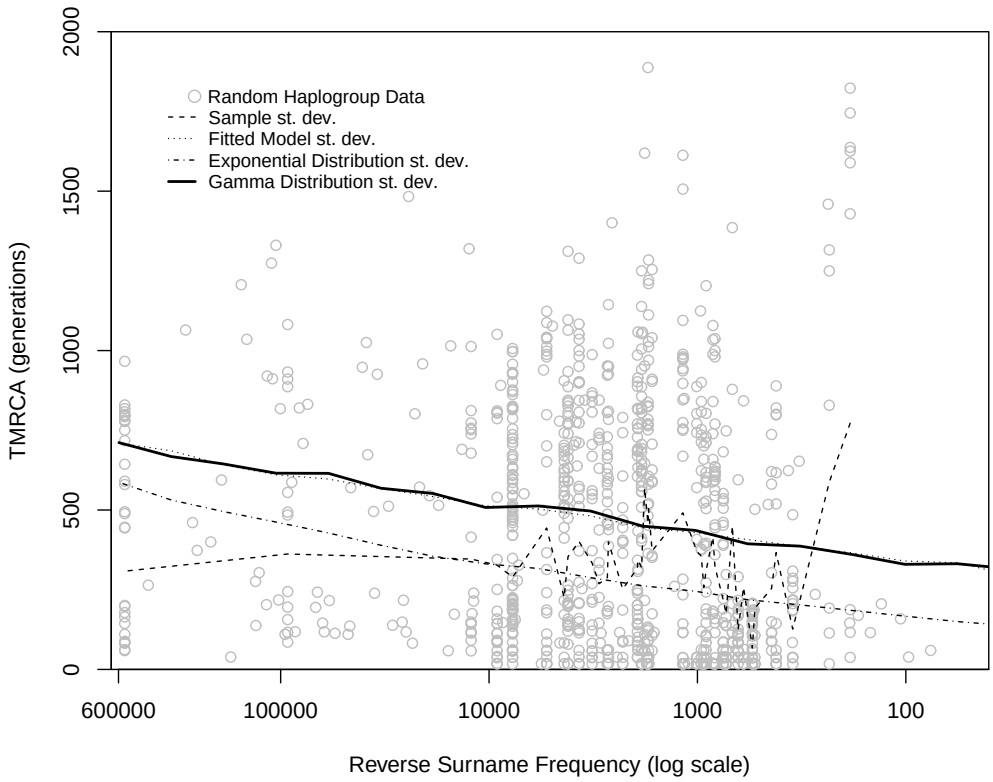

FIGURE 6.16: Random haplogroup: standard deviation for surname frequency prior

to the fitted mean was considered, it did not have enough spread, so a gamma distribution was chosen instead (fig. 6.16).

6.1.3.3 Surname Origins

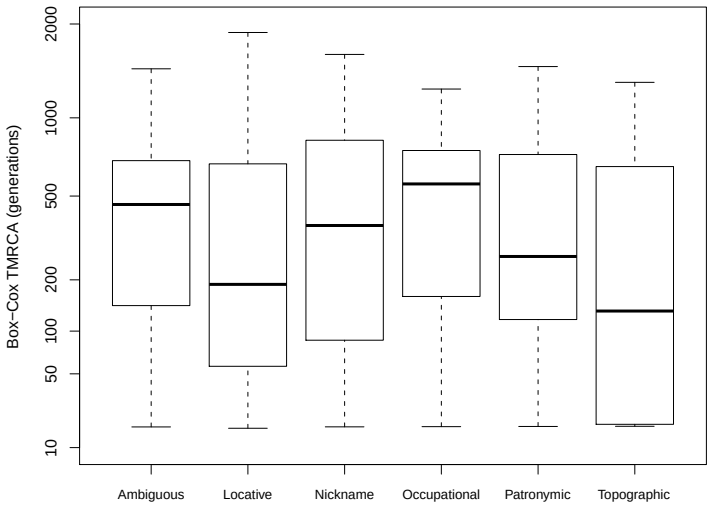

FIGURE 6.17: Random haplogroup: boxplot of Box-Cox TMRCA vs. surname origin

We considered the modelling of TMRCA by surname origins alone for males sharing the same surname with no haplogroup information. The distributions are shown in boxplots of the Box-Cox TMRCA versus surname origins (fig. 6.17). Here the distributions are fairly symmetric except for the topographic surnames which is right-skewed. There is also substantial overlap across the groups.

The surname origin is found to be significant is a one-way ANOVA. In addition, a permutation test also produces a significant result We find that significant differences in mean Box-Cox TMRCA only occur for the ambiguous and occupational surnames when they are compared to either the locative or topographic surnames (table 6.11). The last stage of development of the surname origins based prior in

TABLE 6.11: Random haplogroup: significant Tukey multiple comparisons between surname origin

Surname Origins		95% Confidence Interval of Difference		Adjusted p-value
Group 1	Group 2	Lower	Upper	
Ambiguous	Topographic	0.302	5.380	0.0180
Occupational	Topographic	0.645	6.522	0.0069
Ambiguous	Locative	0.448	2.875	0.0014
Occupational	Locative	0.490	4.318	0.0047

this context involved fitting exponential and gamma distributions by the method of moments to both the full data as well as 1000 thinned samples extracted from the full data. The results are shown in tables 6.12 and 6.13. For the exponential fits, the Kolmogorov-Smirnov goodness of fit test null is rejected for all but the nickname surnames. Also we find that the standard deviation of data simulated from the fitted exponentials is less than the sample standard deviation for the locative and topographic surnames only. For the fitted gamma distribution, the

TABLE 6.12: Random haplogroup: exponential surname origin fitted model parameters

Surname Origin	Rate	K-S p-value
Ambiguous/Unknown	0.00224	0.0000
Locative	0.00259	0.0000
Nickname	0.00213	0.0504
Occupational	0.00200	0.0016
Patronymic	0.002354	0.0167
Topographic	0.00310	0.0016

Kolmogorov-Smirnov p-value is less than 0.05 for all surname origins, although the standard deviation of data generated from the fitted distribution is very close to the sample standard deviation. It is possible that the distributions are not good

TABLE 6.13: Random haplogroup: gamma surname origin fitted model parameters

Surname Origin	Shape	Rate	K-S p-value
Ambiguous/Unknown	1.7984	0.00403	0.0001
Locative	0.8589	0.00222	0.0004
Nickname	1.3303	0.00284	0.0022
Occupational	2.3505	0.00470	0.0023
Patronymic	1.3455	0.00316	0.0011
Topographic	0.8183	0.00254	0.0012

fits to the data due to over representation of the low frequency surnames. So we applied the exponential to 1000 thinned data sets and produced the deflated rate parameters and the associated statistics shown in table 6.14. As in section 6.1.2.3, the topographic surnames rate parameters were unusually distributed. We find that in this case there are nine data points in each thinned sample of which only two vary. The Kolmogorov-Smirnov test p-value is less than 0.05 for 29 of the 1000 ambiguous thinned samples. For the remaining surname origins no thinned samples reject the goodness of fit null hypothesis.

TABLE 6.14: Thinned random haplogroup: exponential surname origin fitted model parameters

Surname Origin	Rate	K-S Null Rejection
Ambiguous/Unknown	0.00165	29
Locative	0.00190	0
Nickname	0.00191	0
Occupational	0.00152	0
Patronymic	0.00222	0
Topographic	0.00185	0

6.1.3.4 No Frequency or Surname Origin

The prior when no haplogroup information is available for two males sharing the same surname is based solely on the TMRCA estimates, a histogram is shown in figure 6.18. There are two modes, one near $t = 0$ and another between 600-700 generations. Oversampling of low frequency surnames may have played a role in this unusual distribution. Recall that the vast majority of the data was from King and Jobling (2009*a*) which focussed on low frequency, surnames whilst King et al. (2006) covered a broad range of surname frequencies but with only one sample from each. We decided to analyse 1000 thinned data sets. Ten realisations are shown in figure 6.19 in black solid lines. In each we may compare the density with that of the full data (red dashed line). Although all the random densities have

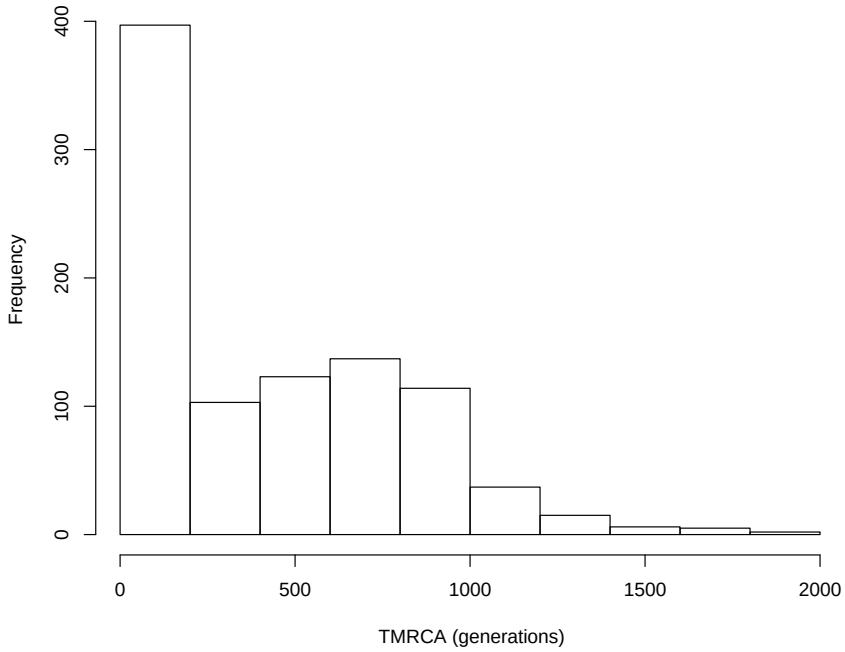

FIGURE 6.18: Random haplogroup: histogram of TMRCA

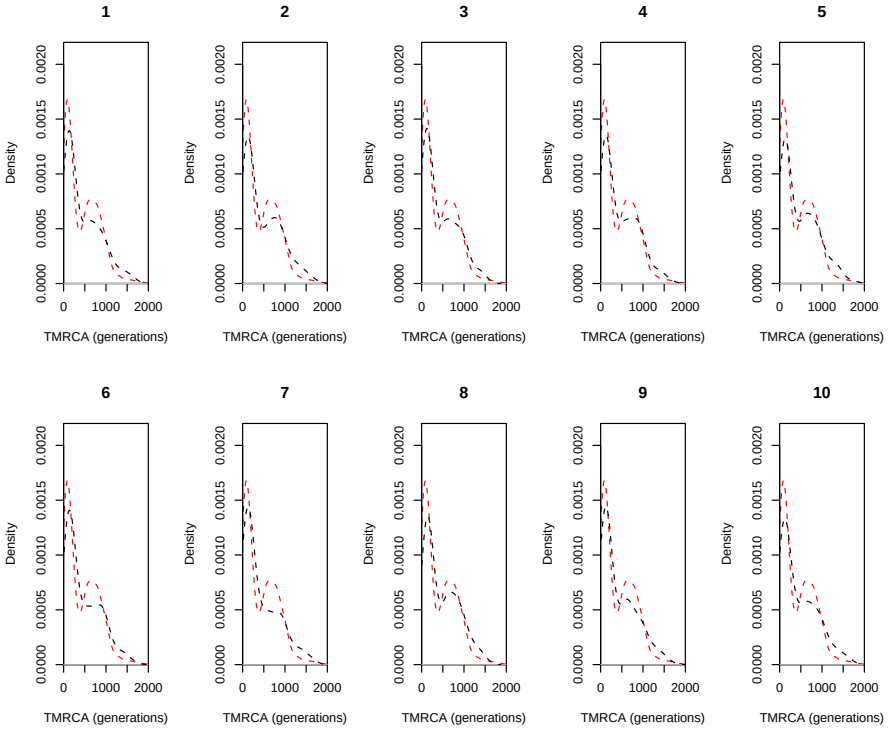

FIGURE 6.19: Random haplogroup: fitted densities of random samples (black dashed line) with full data density (red dashed line)

two peaks, the intensity of the lower peak, in most cases, is markedly less than the corresponding peak in the full data.

We fitted the exponential and gamma distributions to 1000 thinned samples. For the former we reject the Kolmogorov-Smirnov goodness of fit test 65 times whilst

for the latter this is much higher at 806. Thus the prior for TMRCA is an exponential in the context where we have two males who share the same surname but no additional information in terms of their haplogroups or surname frequency or origin is known, with a deflated rate of 0.00207.

6.1.4 Different Haplogroup

6.1.4.1 Surname Origin and Frequency

In order to cover the various possibilities with respect to the haplogroup status that may be encountered when wishing to estimate TMRCA, we end by developing a prior for when the haplogroup differs between pairs of males. In this case no transformation of TMRCA was necessary, although, as before, the logarithm of the surname frequency was taken.

Firstly we wished to confirm that neither the surname origin nor frequency of the surname were useful predictors of TMRCA. Given that both these variables provide information on recent rather than ancient ancestry, we expect that the differing haplogroup status between pairs of males would result in older estimates of TMRCA which predate the period of surname establishment.

To this end, we began by fitting a main effects and interaction ANCOVA model which produced a non-significant interaction term (p-value $= 0.1072$) with the assumptions of normality holding although constant variance was questionable (data not shown). Fitting a main effects model revealed that, whilst the surname origins were significant (p-value < 0.001), log of surname frequency was not significant (p-value $= 0.5339$).

6.1.4.2 Surname Origin

We next modelled TMRCA by the surname origins alone. Boxplots of the data are given in figure 6.20. There does not appear to be much obvious difference in the distributions of TMRCA across the surname origins. Applying a one-way ANOVA shows that surname origin is statistically significant (p-value < 0.001). A permutation test reinforces the one-way ANOVA results.

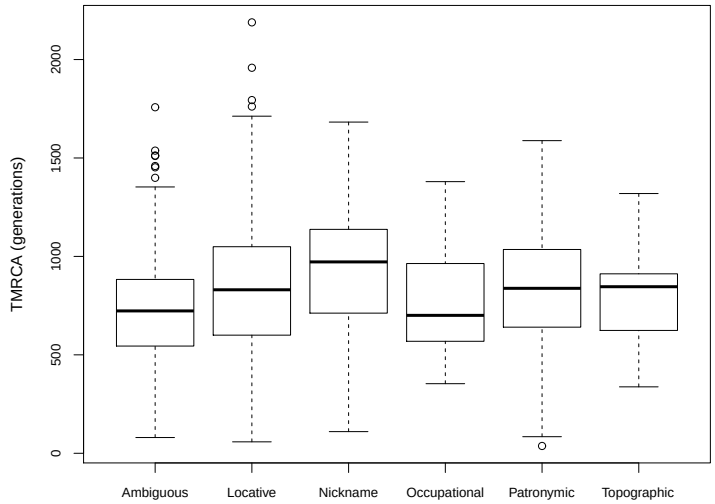

FIGURE 6.20: Different haplogroup: boxplot of TMRCA by surname origin

In order to determine between which groups the differences lie, we produced Tukey multiple comparisons which reveal that there are significant differences in mean TMRCA between three pairs of surname origins: the locative and ambiguous surnames, the nickname and ambiguous surnames and the nickname and occupational surnames (table 6.15). The differences involving the ambiguous group may be due to the ambiguous surnames being ill-defined. This shows that the

TABLE 6.15: Different haplogroup: significant Tukey multiple comparisons between surname origin

Surname Origin		95% Confidence Interval of Difference		Adjusted p-value
Group 1	Group 2	Lower	Upper	
Locative	Ambiguous	3.185	186.856	0.0377
Nickname	Ambiguous	81.628298	328.700	0.0000
Nickname	Occupational	36.615	354.489	0.0062

estimated TMRCA of male-pairs who share the same surname but do not share the same haplogroup would be best modelled by surname origin alone.

6.1.5 Discussion

For the same haplogroup data we modelled the estimated TMRCA using an ANCOVA model with explanatory variables surname frequency and origin. Here we found the interaction significant and a positive slope for each surname origin except the locative surnames. This would imply that as the frequency of the surname

decreases the estimated TMRCA would increase. Given this peculiar result we excluded this surname group and re-examined the same haplogroup data in order to assess how influential the locative surnames are on the fitted model. We began by fitting a main effects and interaction ANCOVA model as before. Interestingly, we find that the interaction term is not significant (p-value= 0.237887). However, log surname frequency and the surname origin, are both significant when fitting the main effects only. This suggests that a main effects model would also apply for the full same haplogroup data, i.e. locative surnames are influencing the results.

Nonetheless, carrying out a permutation test produces a p-value of 0.0001 for the slope for the nickname residuals (table 6.16), which is below the Bonferroni corrected significance level of 0.01, thus indicating an interaction term may be necessary after all even when locative surnames are excluded. In the case where we used

TABLE 6.16: Same haplogroup: permutation test p-values

Surname Origin	Permutation p-value
Ambiguous/Unknown	0.1121
Nickname	0.0001
Occupational	0.6481
Patronymic	0.1719
Topographic	0.5664

the surname origin alone to model estimated TMRCA for the same haplogroup data, we formed an exponential prior based on thinned data (section 6.1.2.3, table 6.8). We also fitted the gamma to the thinned data (table 6.17). The fits resulted in Kolmogorov-Smirnov p-values less than 0.05. Importantly we note that the fit was rejected for only five of the thinned locative data sets. Thus the gamma distribution may also be considered a useful prior on TMRCA. However, for simplicity, we chose the exponential prior.

TABLE 6.17: Thinned same haplogroup: gamma surname origin fitted model parameters

Surname Origin	Shape	Rate	K-S Null Rejection
Ambiguous/Unknown	1.2219	0.00729	0
Locative	0.6509	0.00525	5
Nickname	0.6901	0.00514	0
Occupational	1.2134	0.00603	0
Patronymic	0.6551	0.00360	0
Topographic	1.5437	0.01307	0

For the random haplogroup data set, we found that a main effects model was significant when modelling TMRCA by both surname frequency and origins (section

6.1.3.1). However it was of interest to examine if this model for the random haplogroup data where surname origin data was included was preferred, in a permutation test compared to a model ignoring the surname origin. Thus the surname origin label for the residuals from the surname frequency linear model (section 6.1.3.2) were permutated resulting in a p-value of 0.1082. Thus, despite fitting a main effects ANCOVA including both surname origins and log surname frequency, a regression model, with explanatory log surname frequency only, would explain the variability in estimated TMRCA in this case. As such for pairs of males who share their surname but their haplogroups are unknown, we model the estimated TMRCA by 6.5 even when the surname origin is known.

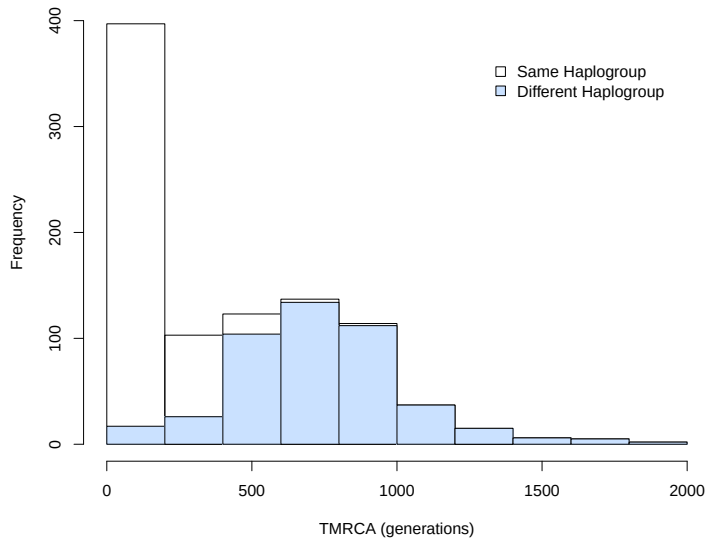

FIGURE 6.21: Distribution of random haplogroup for same and different haplogroup status

Also for the random haplogroup data, there was an unusual pattern in the distribution of the estimated TMRCA both when it was modelled alone and also when it was modelled according to the surname origin. We found two distinct peaks: one close to 100 generations and another at around 700 generations. Although thinned data sets showed a reduced second peak it still remained evident (fig. 6.19). It may be argued that in addition to the oversampling of the low frequency surnames there are far more low frequency surnames sampled overall than higher frequency surnames. However, the data can be segregated according to the haplogroup status, i.e. whether the pairs of males share the same haplogroup or not. Figure 6.21 shows the results of this segregation for the distribution of the estimated TMRCA.

We see that those male-pairs that have different haplogroups have an estimated TMRCA that is centred around ∼700 generations, whilst those with the same

haplogroup lie much closer to 0. Indeed this is also the case when modelling the TMRCA by the surname origin (fig. 6.22). Again the different haplogroup male-pairs have a rather symmetrically distributed estimated TMRCA centred at around ~ 700 generations. Although the haplogroup status would be unknown in this context, it is clear that it affects the distribution of estimated TMRCA.

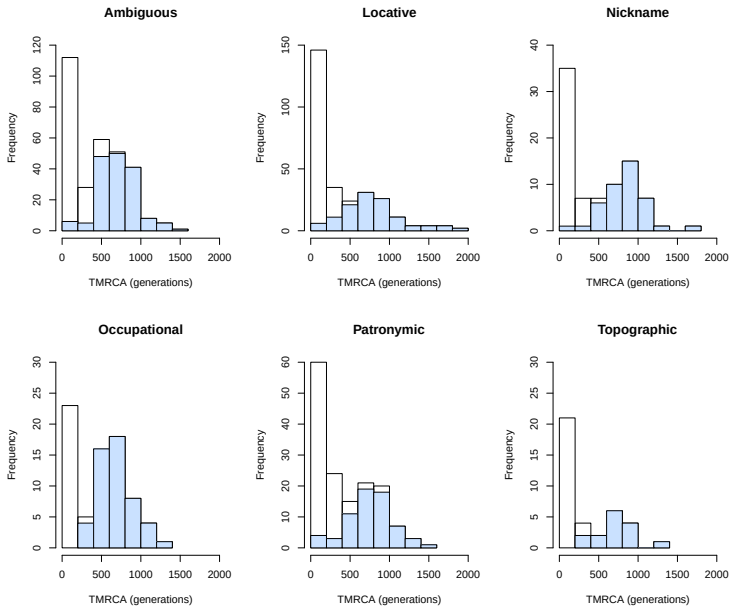

FIGURE 6.22: Distribution of random haplogroup for same and different haplogroup status within surname origin

In section 6.1.3.3 for the random haplogroup data, a thinned exponential was the final choice of prior when the estimates of TMRCA were modelled by surname origin based on table 6.14. However we also fitted the gamma distribution to the thinned samples to produce the results in table 6.18. In addition the Kolmogorov-Smirnov null is only rejected for three of the ambiguous surnames whilst not at all for the remaining surnames. Nonetheless for simplicity we chose the exponential as the prior in this context.

TABLE 6.18: Thinned random haplogroup: gamma surname origin fitted model parameters

Surname Origin	Shape	Rate	K-S Null Rejection
Ambiguous/Unknown	1.5949	0.00296	3
Locative	0.6116	0.00157	0
Nickname	0.9180	0.00221	0
Occupational	5.1328	0.00547	0
Patronymic	1.8096	0.00340	0
Topographic	0.8578	0.00201	0

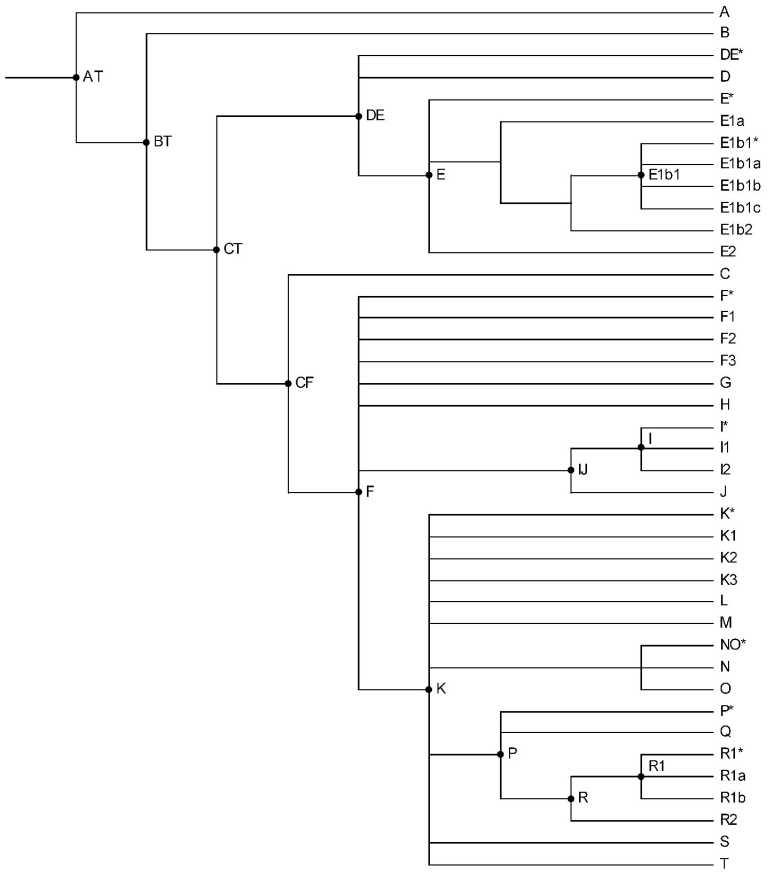

FIGURE 6.23: Y-chromosome SNP tree

For the different haplogroup data, we found that the surname origin was a statistically significant explanatory. However it is intuitive that there may be more information in the haplogroup of the male pairs rather than their surname origin since their ancestry is likely to enormously predate surname origin. Recall the Y-chromosome haplogroup tree (fig. 6.23). Here we have adapted the tree to include labels for some of the internal nodes in accordance with the work of Karafet et al. (2008). For each pair of males that belong to different haplogroups we can identify the internal node from which they diverge. Thus we may model the dependency of TMRCA on the ancestral node (fig. 6.24).

It is clear that those males converging at node AT (the deepest node) have a much older estimated TMRCA than those pairs of males with other ancestral nodes. Applying a one-way ANOVA shows that haplogroup node is a statistically significant predictor of TMRCA (p-value < 0.001).

Consequently we wish to develop a prior for pairs of males who share the same surname but do not share the same haplogroup based on the node on which their

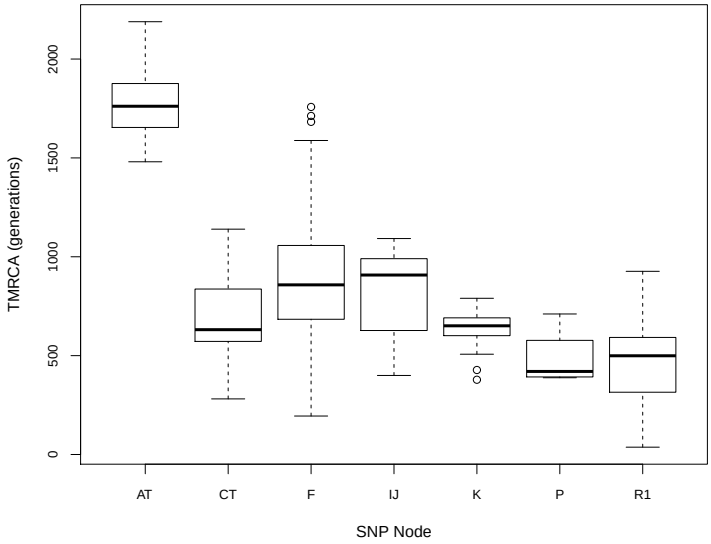

FIGURE 6.24: Different haplogroup: boxplot of TMRCA by haplogroup node

haplogroups converge. Although it is possible to form the prior based on the one-way ANOVA results, crucially this excludes information from SNP data, on which the haplogroup tree is based. Hence incorporating SNP information may provide much more accurate estimates of TMRCA in this context. As such we propose a series of priors based largely on the work of Karafet et al. (2008), which provides selected estimates of TMRCA for males in different haplogroups based solely on SNP mutations. There the authors use a model-free approach in order to estimate the time of various nodes in the Y-chromosome haplogroup tree as shown in figure 6.23.

Karafet et al. (2008) begin by choosing to date the node CT (the calibration node, see figure 6.23) consistent with the out-of-Africa which was 70,000 years BP. The remaining labelled nodes are then estimated by examining the distribution of the counts of mutations along the particular branch relative to the calibration node. Importantly, the accuracy of the calibration node's age is unknown, although the method allows for confidence intervals (CI) to be produced for the other estimates of age conditioned on the age of the calibration node. Additionally, in order to avoid ascertainment bias only uniformly ascertained SNPs were used in the estimation of ages (pers. comm. Mendez) and not all the SNPs identified in the Y tree by Karafet et al. (2008).

The CIs of the ages of the other nodes in Karafet et al. (2008) do not take into account the uncertainty of the age of the calibration node. For our purpose, we propose a standard error (SE) of 10,000 years on the estimate of the age of the

calibration node, CT, which we will then use to compute the SE of the other nodes based on a SE derived from both their CI and the SE of the calibration node.

This was carried out as followed:

- Compute the ratio of the age estimate of the node relative to the age estimate of the calibration node, i.e. $z = \frac{T_{node}}{y}$ where $y = T_{calibration} = 70000$ years.
- Compute the standard error of z on the basis that the length of Karafet et al. CI is approximately 4 SE long, i.e. $z = (T_{upper} - T_{lower})/(2 \times 1.96 \times y)$.
- Assuming independence of the parameters z and y, we may compute the SE of each of the other nodes in Karafet et al. (2008) using the Delta method as outlined in appendix A, i.e.

$$\mathrm{SE}(T_{node}) = \sqrt{y^2\big(\mathrm{SE}(z)\big)^2 + z^2\big(\mathrm{SE}(y)\big)^2}$$

In addition, the age of node AT was needed due to the inclusion of haplogroup A males in King and Jobling (2009*a*). With the addition of the age of node BT, this would cover all major clades. However, this was not feasible based on the Karafet et al. data due to the potential violation of uniform ascertainment for data in these clades (pers. comm. Mendez). This was also the case if we wished to resolve the tree further to include the ages of nodes such as R1a. Crucially any estimates of the age of nodes AT and BT must be compatible with the work of Karafet et al. (2008). Thus they must be greater that the age of the calibration node. Since the accuracy of the age of this node is questionable, we would like to have any estimates of the age of BT of at least as great as 70,000 years.

Three major papers in this area were examined: Thomson et al. (2000), Tang et al. (2002) and Meligkotsidou and Fearnhead (2005). Several estimates based on the literature were inconsistent with this criterion (table 6.19). Meligkotsidou and Fearnhead (2005) is the only paper which provides age estimates for both nodes whilst only one estimate in the paper of Thomson et al. (2000) is compatible with Karafet et al. (2008). On the other hand, Tang et al. (2002) provide various estimates of which several are well above 70,000 years. We therefore proposed the values of 109,000 and 89,500 years for the ages of nodes AT and BT, respectively, after careful consideration of the available literature. The SEs for both is set to 20,000 years reflecting the large uncertainty in the point estimates and assuming that the CI for AT, in Tang et al. (2002), is approximately two standard deviations wide. Given that all the estimated node ages were in years, we converted them

TABLE 6.19: Published estimates of the ages of nodes AT and BT

Author (Publication Date)	Details	Node	Estimates of TMRCA (CI, x10^3 yrs)
Thomson et al. (2000)	Constant Ne	AT	84 (55-149)
	Exponential Ne growth	AT	59 (40-140)
	Model free approach	AT	70 (SE=30)
Tang et al. (2002)	Constant Ne	AT	117
	Exponential Ne growth	AT	75
	30yrs/gen	AT	109 (72-156)
	25yrs/gen	AT	91 (60-130)
Meligkotsidou and Fearnhead (2005)	Reduced data	AT	63 (40-100)
	Reduced data	BT	45 (32-74)
	Full data	AT	56 (39-83)
	Full data	BT	36 (27-50)

into generations using the conversion rate of 31.93 years per generation with a standard deviation of 8.06, based on the work of Helgason et al. (2003). The point estimates could be directly converted by dividing by the rate, whilst the SEs in generations were obtained using the Delta method. Consequently, we have the

TABLE 6.20: Different haplogroup: prior parameters

Node	TMRCA Mean	TMRCA SE	Gamma Parameters	
			Shape	Rate
AT	109000	20000	29.69	0.00870
BT	89500	20000	20.02	0.00714
CT	70000	10000	48.93	0.02232
CF	68900	9935.283	48.02	0.02226
DE	65000	9577.714	45.99	0.02259
E	52500	8340.120	39.56	0.02406
E1b1	47500	7840.892	36.64	0.02463
F	48000	8113.426	34.95	0.02325
IJ	38500	6803.737	31.97	0.02651
I	22200	4911.258	20.39	0.02933
K	47400	7893.312	36.01	0.02425
P	34000	6151.936	30.49	0.02863
R	26800	5305.878	25.46	0.03034
R1	18500	4280.621	18.64	0.03217

results in table 6.20 for the nodes dated in Karafet et al. (2008) as well as nodes AT and BT. The shape and rate parameter of the gamma priors here are obtained by fitting via the method of moments and apply in the context where two males share the same surname but not the same haplogroup status.

Now it is possible to use the gamma distributions above as direct priors on t. However since the node is a maximum bound for the time to the MRCA for male-pairs, it would seem more appropriate to use a uniform distribution with lower bound 0 and upper bound T_{dHG} which in turn is modelled by the gamma

distributions described in table 6.20 or more generally $T_{dHG} \sim Ga(k_T, \theta_T)$. We can include T_{dHG} as a parameter in our Bayesian analysis.

In this context our posterior distribution would be altered slightly: using the uniform prior for t, the gamma prior for T_{dHG}, the exponential priors for α and β and the normal prior for N_e (5.3) we have the posterior distribution:

$$
\begin{aligned}
&P\left(\{\mu_i\}, t, T_{dHG}, L, N_e, \alpha, \beta | n_{asc}, \{m_i\}, \{R_i\}, \{r_i\}, \{x_{1,i}\}, \{x_{2,i}\}\right) \\
&\propto P\left(\{x_{1,i}, x_{2,i} : i \in \mathfrak{T}\} | \{\mu_i\}, t\right) P\left(\{r_i : i \in \mathfrak{C}\} | \{m_i\}, \{\mu_i\}\right) \\
&\quad \times P\left(R_i = 1 : i \in \mathfrak{A} | \{\mu_i\}, L, N_e\right) P\left(R_i = 0 : i \in \mathfrak{N} | \{\mu_i\}, L, N_e\right) \\
&\quad \times P\left(\{\mu_i : i \in \mathfrak{T} \cup \mathfrak{C} \cup \mathfrak{A} \cup \mathfrak{N}\} | \alpha, \beta\right) P\left(L | n_{asc}\right) \\
&\quad \times P\left(\alpha\right) P\left(\beta\right) P\left(N_e\right) P\left(t | T_{dHG}\right) P\left(T_{dHG}\right) \\
&\propto \prod_{i=1}^{s} e^{-2t\mu_i} I_{|x_{1,i}-x_{2,i}|}(2t\mu_i) \binom{m_i}{r_i} \mu_i^{r_i} \left(1-\mu_i\right)^{m_i - r_i} \left(1 - e^{-\mu_i L N_e}\right) \\
&\quad \times \prod_{i=s+1}^{c} \binom{m_i}{r_i} \mu_i^{r_i} \left(1-\mu_i\right)^{m_i - r_i} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=c+1}^{k} \left(1 - e^{-\mu_i L N_e}\right) \prod_{i=k+1}^{n} e^{-\mu_i L N_e} \\
&\quad \times \prod_{i=1}^{n} \frac{\mu_i^{\alpha-1} \exp\left(-\mu_i/\beta\right)}{\beta^{\alpha} \Gamma(\alpha)} \\
&\quad \times \frac{n_{asc} - 1}{2} e^{-L/2} \left(1 - e^{-L/2}\right)^{n_{asc}-2} \\
&\quad \times \frac{1}{\sigma_{N_e}\sqrt{2\pi}} \exp\left(-\frac{(N_e - \mu_{N_e})^2}{\sigma_{N_e}^2}\right) e^{-\lambda_\alpha \alpha} e^{-\lambda_\beta \beta} \frac{1}{T_{dHG}} \frac{T_{dHG}^{k_T - 1} \exp\left(\frac{-T_{dHG}}{\theta_T}\right)}{\theta_T^{k_T} \Gamma(k_T)}.
\end{aligned}
\tag{6.6}
$$

The vector of parameters is augmented by T_{dHG}. The Metropolis ratio does not change for any of the parameters except t, where the prior is now uniformly distributed as described:

$$
R = \frac{\prod_{i=1}^{s} P\left(\{x_{1,i}\}, \{x_{2,i}\} | \{\mu_i\}, t^*\right) P(t^* | T_{dHG})}{\prod_{i=1}^{s} P\left(\{x_{1,i}\}, \{x_{2,i}\} | \{\mu_i\}, t\right) P(t | T_{dHG})}. \tag{6.7}
$$

In addition, the Metropolis ratio for T_{dHG} is:

$$
R = \frac{P(t | T_{dHG}^*) P(T_{dHG}^*)}{P(t | T_{dHG}) P(T_{dHG})}. \tag{6.8}
$$

6.2 Additional Priors

6.2.1 Absence of Surname and Haplogroup Information

We have thus far formed priors for t in all scenarios except the situation where we have typed STRs from two males on whom we have no surname or haplogroup information, i.e. a random pair of males who do not necessarily share the same surname. We also assume the males in question are not known to be recently related. Here we use a prior motivated by the coalescent process (for a review, see Wakeley (2009)), which states that in a population of constant 'effective' size, N_e, the distribution of the TMRCA of two random Y-chromosomes is exponential:

$$P(t|N_e) = \frac{1}{N_e} \exp\left(\frac{-t}{N_e}\right). \tag{6.9}$$

The posterior distribution is amended only in this aspect. This will clearly necessitate a new Metropolis ratio for updating t:

$$R = \frac{\prod_{i=1}^{s} P\left(\{x_{1,i}\}, \{x_{2,i}\}|\{\mu_i\}, t^*\right) P(t^*|N_e)}{\prod_{i=1}^{s} P\left(\{x_{1,i}\}, \{x_{2,i}\}|\{\mu_i\}, t\right) P(t|N_e)}. \tag{6.10}$$

But, in addition, the Metropolis ratio for updating N_e is altered due to the dependency of t on N_e:

$$R = \frac{P(t|N_e^*)P(N_e^*) \prod_{i=1}^{k} P(\{R_i = 1\}|\mu_i, L, N_e^*) \prod_{i=k+1}^{n} P(\{R_i = 0\}|\mu_i, L, N_e^*)}{P(t|N_e)P(N_e) \prod_{i=1}^{k} P(\{R_i = 1\}|\mu_i, L, N_e) \prod_{i=k+1}^{n} P(\{R_i = 0\}|\mu_i, L, N_e)} \tag{6.11}$$

6.2.2 Generation Time in Years

In an academic context expressing times in generations is convenient. However in the more populist genealogical context it would be more beneficial to express the TMRCA is years. This requires a conversion factor. Simply using a point estimate, such as 25 years per generation, would provide us with a crude estimate of t, but without any measure of uncertainty in the conversion.

However if have both the mean and standard error of the conversion factor, it is possible to incorporate the conversion factor into our analysis as a further parameter that is updated in the MCMC. This involves modelling the conversion rate using an appropriate distribution. Though the normal distribution may be one such option, it allows the possibility of negative values. As such, instead we use a gamma distribution.

Helgason et al. (2003) suggest the following conversion factors:

- 31.13 years per generation (SD=7.57);
- 31.93 years per generation (SD=8.06);

based on, respectively, 122,822 and 117,486 unique father-son pairs in the patrilineal coalescent genealogies of 8,275 and 1,859, respectively, Icelandic patrilines. These estimates are close to the estimate of 33.9 years per generation for 2,221 French Canadian patrilines (Tremblay and Vezina, 2010), though both are slightly less than the earlier work of (Tremblay and Vezina, 2000) which put this value at 35 years per generation based on only 87 patrilines. Neither of these works included a clear measure of the uncertainty in their estimates. As such we focussed on the conversion factors of Helgason et al. (2003). For the purpose of this thesis, erring on the side of conservativeness, we will use the second rate of Helgason et al. (2003), as it has a slightly higher SD to form the resulting gamma distribution. We define Y_g to be the years per generation conversion factor and estimate the shape and scale parameters of the gamma using the method of moments, to give:

$$Y_g \sim Ga(k_Y = 15.693, \theta_Y = 2.0346). \tag{6.12}$$

This prior distribution is depicted in figure 6.25.

The vector of parameters is augmented by Y_g. For updating Y_g, the Metropolis ratio is trivially

$$R = \frac{P(Y_g^*)}{P(Y_g)}. \tag{6.13}$$

The time in years can be found by simply multiplying the chains for t with Y_g, update by update and thereafter constructing a chain for time in years, t_y. The mean, standard deviation and credible region of the chain will then form the posterior summaries of t_y.

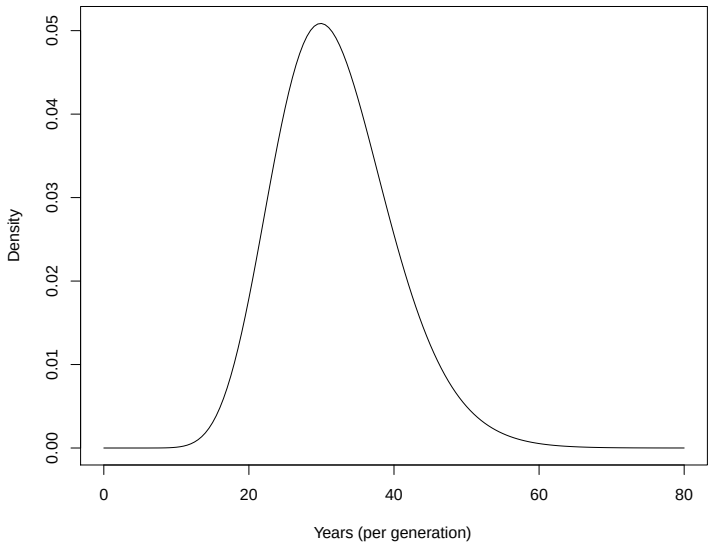

FIGURE 6.25: Years per generation distribution: $Ga(15.70, 2.035)$

6.3 Assessment of Priors on TMRCA

In this section we will assess the effect the various priors on the TMRCA has on our estimates compared to sections 6.1 and 6.2, which is based on the model outlined in chapter 5.

6.3.1 Materials and Methods

A summary of the nine priors for the TMRCA formed on the basis of section 6.1 is given in table 6.21.

For priors 3 and 6, the parameters have the values shown in table 6.22. For priors 4, 7 and 8 the parameters of the gamma distribution are found by the method of moments, i.e. from

$$k\theta = \left(\lambda[\alpha + \beta log(S_f)] + 1\right)^{\frac{1}{\lambda}}$$

and

$$k\theta^2 = \text{Var}\left(\left(\lambda[\alpha + \beta log(S_f) + \varepsilon] + 1\right)^{\frac{1}{\lambda}}\right)$$

where $\varepsilon \sim \text{N}(0, \sigma^2)$ and the values of the parameters are given in table 6.23. It would have been ideal to examine new data with known ancestry (i.e. where the TMRCA was known) to examine the effectiveness for both the original model used to construct the priors as well as with the new priors. However given the lack of access to such data, we chose to examine the data from King et al. (2006) using

TABLE 6.21: TMRCA Priors for pairs of males

Prior	Surname Status	Haplogroup Status	Surname Frequency	Surname Origin	Prior	Details
1	Unknown	NA	NA	NA	$t \sim \text{Ex}(\beta = N_e)$	N_e is the scale parameter
2	Same	Unknown	Unknown	Unknown	$t \sim \text{Ex}(\theta)$	$\theta = 0.002068961$ is the rate parameter
3	Same	Unknown	Unknown	S_o	$t \sim \text{Ex}(\theta_{S_o})$	For θ_{S_o} see table 6.22
4	Same	Unknown	S_f	Unknown/S_o	$t \sim \text{Ga}(k_{S_f}, \theta_{S_f})$	k_{S_f},θ_{S_f} are the shape and rate parameters respectively
5	Same	Same	Unknown	Unknown	$t \sim \text{HN}(\theta)$	$\theta = 0.005731699$ is the scale parameter of the Half Normal
6	Same	Same	Unknown	S_o	$t \sim \text{Ex}(\theta_{S_o})$	For θ_{S_o} see table 6.22
7	Same	Same	S_f	Unknown	$t \sim \text{Ga}(k_{S_f}, \theta_{S_f})$	k_{S_f}, θ_{S_f} are the shape and rate parameters, respectively
8	Same	Same	S_f	S_o	$t \sim \text{Ga}(k_{S_f,S_o}, \theta_{S_f,S_o})$	k_{S_f,S_o}, θ_{S_f,S_o} are the shape and rate parameters respectively
9	Same	Different	Unknown/S_f	Unknown/S_o	$t \sim \text{Un}(0, T_{dHG})$	$T_{dHG} \sim \text{Ga}(k_{T_{dHG}}, \theta_{T_{dHG}})$ and $k_{T_{dHG}}$, $\theta_{T_{dHG}}$ are the shape and rate parameters respectively

TABLE 6.22: θ_{S_o} values for priors 3 and 6

Surname Origin	θ_{S_o}	Prior 3	Prior 6
Ambiguous	θ_{A}	0.00165	0.00492
Locative	θ_{L}	0.00190	0.00549
Nickname	θ_{N}	0.00191	0.00524
Occupational	θ_{O}	0.00152	0.00321
Patronymic	θ_{P}	0.00222	0.005594
Topographic	θ_{T}	0.00185	0.00656

TABLE 6.23: Parameter values for Priors 4, 7 and 8

Parameter	Prior 4	Prior 7	Prior 8
λ	0.2626	0.0606	0.0606
α/α_{S_o}	7.795	3.316	A: 3.599 L: 6.085 N: 0.904 O: 3.187 P: 2.725 T: 1.914
β/β_{S_o}	0.647	0.231	A: 0.211 L: -0.177 N: 0.565 O: 0.225 P: 0.330 T: 0.326
σ	5.124	1.379	1.354

each of the priors, excluding the first prior which is only applicable for two random males excluding any surname information. This is not ideal since the data used for constructing the prior is being used for posterior analysis, but it would give an idea of how the priors were performing. In addition, it was not possible to analyse the complete dataset used to construct priors due to time constraints.

Consequently the STR differences from the 150 pairs of males from King et al. (2006) formed what we will regard as the random haplogroup dataset. Of these 65 pairs shared the same haplogroup, the same haplogroup dataset. The different haplogroup data comprised the 85 remaining pairs.

The random haplogroup data set is analysed using three different priors: priors 2-4 (table 6.21). The same haplogroup data set is analysed using priors 5-8 whilst the different haplogroup data is only analysed once using prior 9. For each analysis, we will examine plots of the mean of the MCMC chain for t in generations against the means produced using the model outlined in chapter 5 referred as the 'standard' prior/model. Superimposing the line of equality will allow a direct comparison of the effect of the alternate priors compared to the standard model.

6.3.2 Same haplogroup

The $\hat{t}$ obtained when applying the four priors to the same haplogroup data for pairs of males sharing the same surname from King et al. (2006) are each compared to the estimates from the standard prior model in figure 6.26, each with the line of equality superimposed.

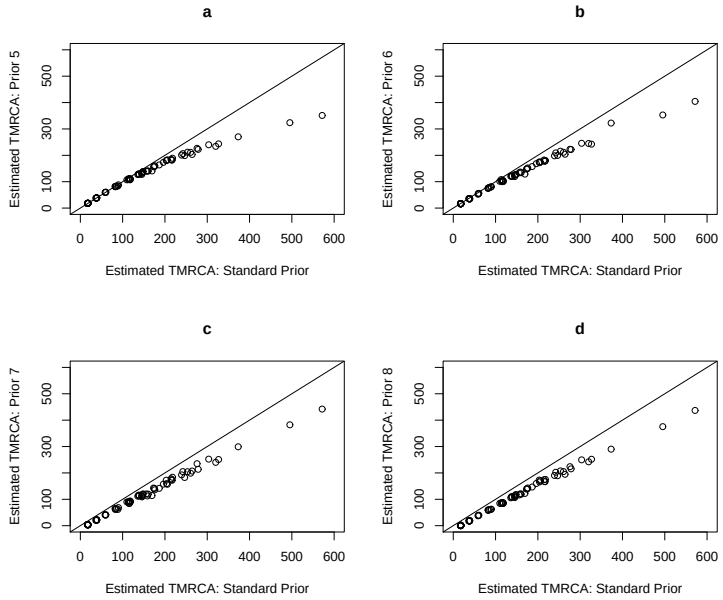

FIGURE 6.26: Same haplogroup TMRCA (new prior) vs. TMRCA (standard prior): a. prior 5 b. prior 6 c. prior 7 d. prior 8

Prior 5 (fig. 6.26a) does not include information on the surname frequency or origin. For low values of TMRCA, the points lie on the line of equality. However, as TMRCA increases, the difference between the two models' estimates increases, so much so that the last point has a $\hat{t} \sim 600$ using the standard prior model and only 330 when using prior 5. These results indicate that prior 5 produces lower estimates of TMRCA than the standard model particularly for older TMRCA.

Incorporating surname origins into the prior produces the results for $\hat{t}$ in figure 6.26b, where we find that for $\hat{t} < 40$ there is little difference in the estimates of TMRCA from prior 6 and the standard prior. However, as the TMRCA increases, the amount by which the points lie below the line of equality increases. This demonstrates that using the prior 6 model produces lower estimates of TMRCA than the standard prior particularly for older values of TMRCA.

Figure 6.26c, shows prior 7 estimates versus those for the standard prior. Here we find that including only surname frequency results in all the points lying below

the line of equality and the amount by which this occurs increases as TMRCA increases. Thus using prior 7 produces lower estimates of TMRCA than the standard model for all TMRCA though the reduction is greater for older times.

When using prior 8 (fig. 6.26d), which includes both surname frequency and origin, the $\hat{t}$s are again consistently less than those produced when using the standard prior. As $\hat{t}$ increases the difference between the estimates of the two models increases.

6.3.3 Random Haplogroup

Figure 6.27 shows the results for $\hat{t}$ when applying the alternative priors. Figure 6.27a compares the standard prior to the prior for pairs of males who share the same surname but their haplogroups are unknown, i.e. they may or may not share haplogroups. With additional information on the surname namely its origin, prior 3 may be employed and figure 6.27b compares the results from this to those of the standard prior. Lastly we present the results for prior 4 compared to the standard prior (fig. 6.27c). Here the alternative prior is applied to surname-sharing male-pairs whose haplogroups are unknown but the surname frequency is given and the surname origin may or may not be known.

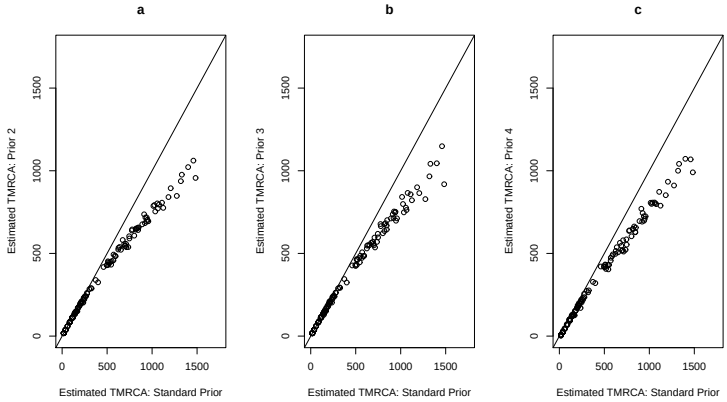

FIGURE 6.27: Random haplogroup TMRCA (new prior) vs. TMRCA (standard prior): a. prior 1 b. prior 2/3 c. prior 4

In all three cases we find that the points generally lie on or below the line of equality suggesting that the alternative priors produce estimates of TMRCA equal to or less than the standard prior's estimate of TMRCA. In fact for estimated TMRCA less than 500 generations there is little difference between the two models. However, for higher values of TMRCA, the difference between the estimates of the two models increases gradually as the TMRCA increases. However, on closer inspection, when

using prior 4, which includes surname frequency, all the points lie below the line of equality, even for the lowest values of TMRCA (fig. 6.28).

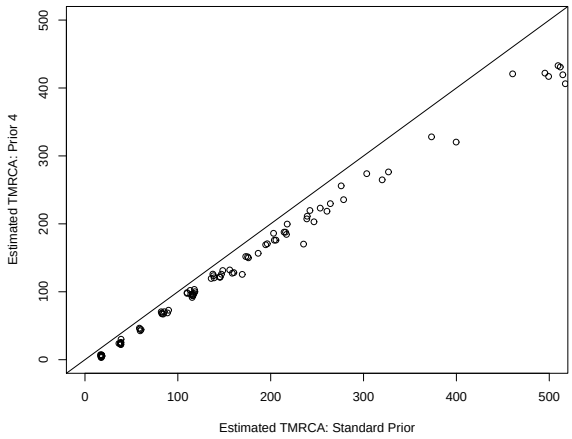

FIGURE 6.28: Random haplogroup TMRCA (prior 4) vs. TMRCA (standard prior): range (0,500))

6.3.4 Different Haplogroup

Lastly the results from prior 9 are described. This prior is applied in the situation where pairs of males share the same surname but do not share the same haplogroup. The estimated TMRCA of the different haplogroup data from King et al. (2006) using this prior is compared to the standard prior (fig. 6.29). In addition, each pair is labelled according to its ancestral node, i.e. the internal node that the haplogroups of the pair of males diverge from. In this case, there are only four ancestral nodes: F, R1, CT and K.

Consider the R1 estimates: these points (○) lie on the line of equality for TMRCA less than 250 generations. For larger values, however, the points diverge to approach an asymptote at around 500 generations on the y-axis. This pattern is also evident for the ancestral node F points (○), however the asymptote on the y-axis is shifted further back in time to around 1,100 generations. On the other hand for the remaining ancestral nodes K and CT this does not appear to be the case. In these cases the points lie on or close to the line of equality suggesting that both priors agree in their estimation of TMRCA. This may in part be due to these points have a relatively lower estimated time based on the standard prior.

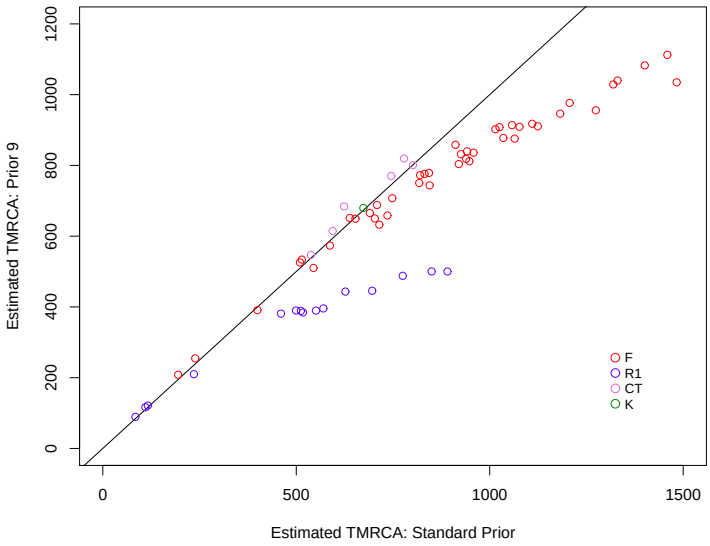

FIGURE 6.29: Different Haplogroup: TMRCA (new prior) vs. TMRCA (standard prior)

6.3.5 Discussion

The different haplogroup prior is conditioned on an upper bound of the age of the ancestral node (T_{dHG}) which may possibly explain the asymptotic trend of the TMRCA estimates based on the different haplogroup prior compared to estimates based on the standard model. Indeed this is found to be the case when we examine a plot of the estimates of the age of the ancestral node versus the estimated of TMRCA based on the different haplogroup model (fig. 6.30). The estimated age of the R1 ancestral node is around 600 generations whilst the estimated age of F is close to 1500.

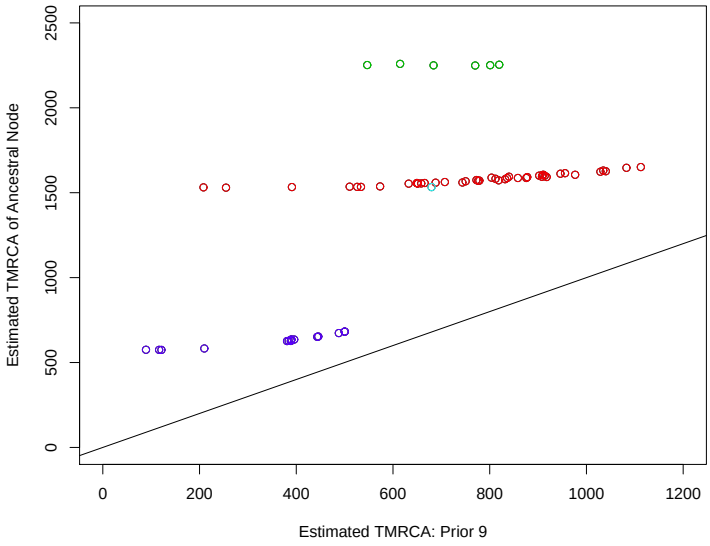

FIGURE 6.30: Different haplogroup: TMRCA ancestral node vs. TMRCA (prior 9)

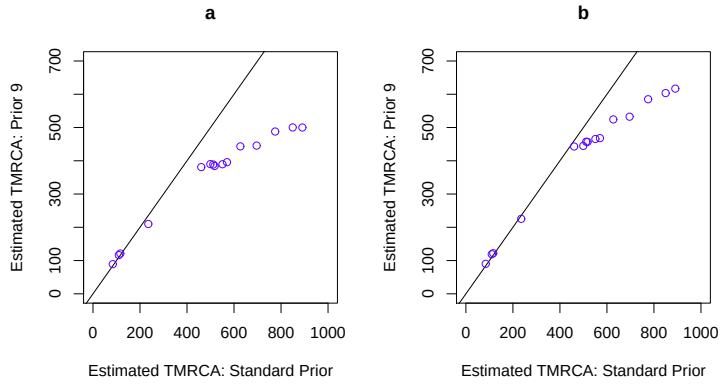

FIGURE 6.31: Different haplogroup TMRCA (new prior) vs. TMRCA (standard prior): a. prior R1 b. prior R

To investigate further if indeed the age of the ancestral node was the reason for the asymptotic trend we examined a subset of the different haplogroup data, i.e. we reanalysed the pairs that had the R1 ancestral node but misspecified the older node R as the ancestral node instead. Figure 6.31 shows the original results for this subset of data next to the results when applying the misspecified ancestral node. Both figures show the estimated TMRCA using the different haplogroup model versus the standard model. It is clear that the points in figure 6.31b lie higher up and closer to the line of equality than the points in figure 6.31a do, particularly for large $\hat{t}$

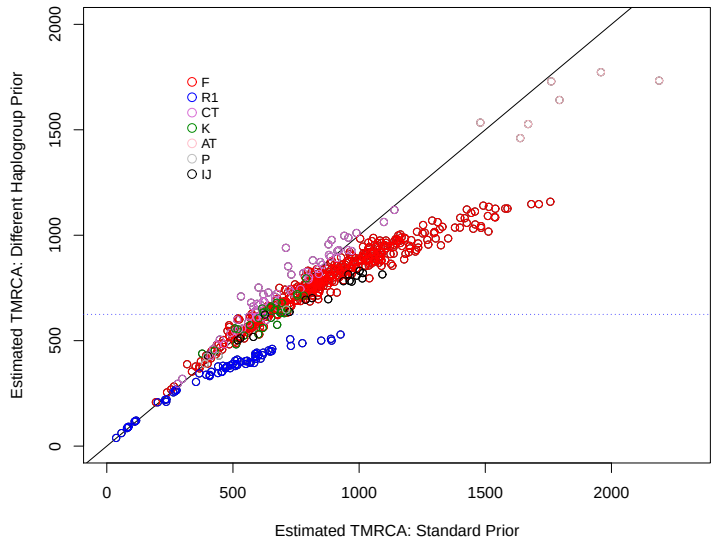

FIGURE 6.32: Different haplogroup (complete data): TMRCA (new prior) vs. TMRCA (standard prior)

Furthermore, it was possible to apply the different haplogroup model to the complete data set, $n = 608$, excluding any potential overlaps between King et al. (2006) and King and Jobling (2009*a*), to examine if the asymptotic relationship between the estimated TMRCA of the two models also applied. This comparison

between the models is shown in figure 6.32 and indeed we find the asymptotic trend. It is most obvious for the R1 ancestral node, which we can compare to the dashed blue line which represents the mean estimated age of the R1 node.

6.4 Conclusions

In section 6.1, we developed three priors for the TMRCA for surname-sharing pairs of males whose haplogroup is unknown, i.e. they may or may not share the same haplogroup. Where no additional surname information is given, an exponential prior is assigned to TMRCA. This is also the case when the origin of the surname is known. In this case, we have a exponentials of different rates according to the surname origin. However surname origin is not necessary in addition to the frequency of the surname, which sufficiently explains the variability of the estimated TMRCA using a gamma distribution with parameters depending solely on the frequency. We have outlined four priors for the TMRCA of surname- and haplogroup-sharing pairs of males. Where only this basic information is known for the pairs, a half-normal distribution forms the prior on TMRCA. This gives more weight to recent times than an exponential distribution. However this latter distribution was an adequate prior on TMRCA when surname origin is included. On the other hand, it is necessary to have a gamma prior on TMRCA when only the frequency of the surname is known in addition to the males sharing the same surname and haplogroup. This distribution also forms the prior when both the surname origin and frequency are included. Finally, for surname-sharing pairs of males who do not share the same haplogroup, we developed a uniform prior on TMRCA with limits 0 to the age of the ancestral node. This is the age of the most recent node in the SNP haplogroup tree to which the haplogroups of the males converge. This age was included as an additional parameter in our model and it was modelled using a gamma distribution.

We then implemented each of these priors on the relevant data from King et al. (2006). Application of the random and same haplogroup priors showed the same trend in their estimates of TMRCA when compared to the standard model in chapter 5: there was little difference between the models in their estimates of lower TMRCA values but, for higher TMRCA,the alternate prior on time produced lower estimates than those using the standard model. We found that, when the frequency of the surname was included into the prior on TMRCA (priors 4, 7 and 8), reduced

estimates of TMRCA were produced across all times, suggesting that the surname frequency was more informative for younger TMRCA than other information. The prior applied in the different haplogroup case was strongly affected by the choice and age of the ancestral node and in this case also, the estimates of TMRCA were lower for older times than those based on the standard model.

However these results would only be determined conclusively by examining STR data from males whose ancestry and, in particular, their TMRCA are known. It may be possible to simulate such data. However, this would require careful modelling of all factors that affect surname-haplotype transmission including the number of founders, non-paternity events, genetic drift as well as the STR mutation process.

Chapter 7

Discussion

This thesis has presented a comprehensive methodology for the estimation of TMRCA in the genealogical context, demonstrating through the use of simulation studies that factors such as the manner in which mutational mechanisms are modelled and the mutation rate employed affect estimates of TMRCA. The conclusions may be compromised if the simulated data were not reflective of real data or the application of the methodology is flawed.

In terms of the latter, the application of the MCMC developed in chapters 4 and 5 relied on visual diagnostics alone to assess the convergence of the chains for each parameter. Importantly, given the large number of parameters it was not feasible to check this for every parameter and it was indeed impractical to do so even for the main parameters in the simulation studies. As such an objective measure of the convergence should ideally be computed, e.g. the potential scale reduction factor (PSRF). This involves running more than two chains with over-dispersed starting values for each parameter in the MCMC, thereafter computing the PSRF based on the between chain and within chain variances (Gelman and Rubin, 1992). Values close to 1 would indicate that the correct posterior is being sampled from, provided that the initialisations are indeed over-dispersed though this is not always the case in practice. This is easily implemented in the absence of an adaptive scheme, which itself may result in destroying convergence in some applications (Roberts and Rosenthal, 2007). The adapted parameters (the log standard deviation for each parameters proposal distribution) may indirectly influence the distribution that is sampled in the MCMC, as such it may be argued that they too must have reached convergence. However, the work of Roberts and Rosenthal (2007) argue

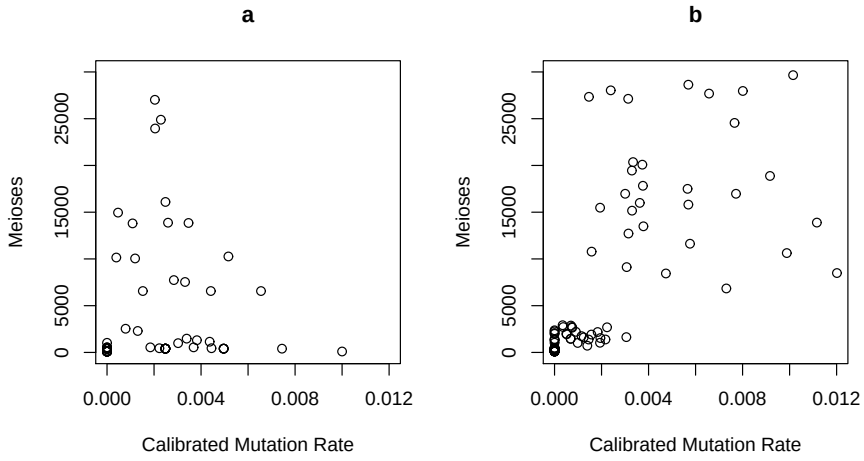

FIGURE 7.1: Number of meioses vs. empirical mutation rate: a. intermediary mutation rate review, b. Simulated Data (three calibration steps)

to the contrary provided the condition of diminishing adaptation holds, which is the case for the scheme implemented. Nonetheless, applying the MCMC using fixed values of the proposal distributions standard deviation based on the results of the adaptive scheme might be a sensible approach.

For the former, it was of particular interest to compare the inferences drawn from estimates of TMRCA based on simulated mutation rates using the multi-stage calibration process to those based on sampling the number of meioses from the real data reported in the intermediate mutation rate review. Recall in chapter 4 that simulated mutation rates were based on a fixed number of meioses (10,000). Categorising the markers as having a low, intermediate or high mutation rate (alongside the non-ascertained and non-calibrated groups) showed that the simulated data did not reflect the distribution of real-life STR mutation rates. Consequently we developed a multi-stage calibration process within our simulation by which the more mutable markers, as determined empirically, were estimated from a larger number of meioses than used to estimate the less mutable markers. We compared the relationship of the number of meioses and the empirical estimate of the mutation rate for the simulated data (section 4.4.2.2) with that for the real data, based on the intermediary mutation rate review (section 4.4.1), in order to examine the effect, if any, of the multi-stage calibration process. This is shown in figure 7.1. Here we can see that, for the real data, the spread of number of meioses is greatest around 0.002 (fig. 7.1a). Contrast this with the simulated data, which has a roughly equal spread of number of meioses between 0.004 and 0.012 (fig. 7.1b).

Now the results in chapter 5 involved simulating data using the layered calibration

procedure to produce the empirical mutation rates. It was found to be advantageous to use the fastest markers to estimate TMRCA of a pair of males rather than a random set of markers. However, the layered process allows a much greater number of meioses for the fastest markers, as determined empirically, which would imply that these are far more accurately measured. Thus the benefit of using the fastest markers may partly be due to their being both fast mutating *and* accurately estimated. This may potentially cause bias in the estimates of TMRCA. For example, markers which are seen as more mutable on the basis of a low number of meioses will not be as accurately estimated as those based on many meioses. There is a notable absence of high mutation rates estimated from a small number of meioses in the simulated data, in contrast to real data (fig. 7.1ba).

In order to examine the effect of such mutation rates on estimates of TMRCA, we again simulate data as outlined in chapter 5 for the case $t = 5$ using random markers to estimate TMRCA for simulated STR profiles for 100 pairs of males. Crucially, instead of using the layered process for deriving empirical estimates of the mutation rates when generating the datasets, we sample from the number of meioses obtained for the 86 STRs in the intermediary mutation rate review. We vary the percentage of typed STRs from 10%-100%. Next we order the typed STRs according to their empirical mutation rate from the most mutating to the least. The markers at the start of the list are referred to as the fast markers. Thereafter we analyse the simulated datasets once again, allowing the percentage of fast markers to vary from 10% to 70%.

The fractional squared error (FSE) against the percentage of typed loci is shown for the random markers (○) as well as the fast markers (○) alongside the same case when using the layered process in chapter 5 (fig. 7.2ab). The use of samples from the real number of meioses increases the FSE of both the random and fast markers. For example, when only 10% of the typed loci are employed to estimate the TMRCA, the FSE using the random markers is around 150 but only 12 for the fast markers (fig. 7.2a). On the other hand, the FSE when simulating data using the layered empirical number of meioses are 130 and 10, respectively (fig. 7.2b). Similarly the fractional variance is slightly greater for both the random and fast markers when using the real number of meioses compared to using the layered approach. Nonetheless the benefit of using the fast markers is still evident.

This is also the case when examining the bias in the estimates of TMRCA across

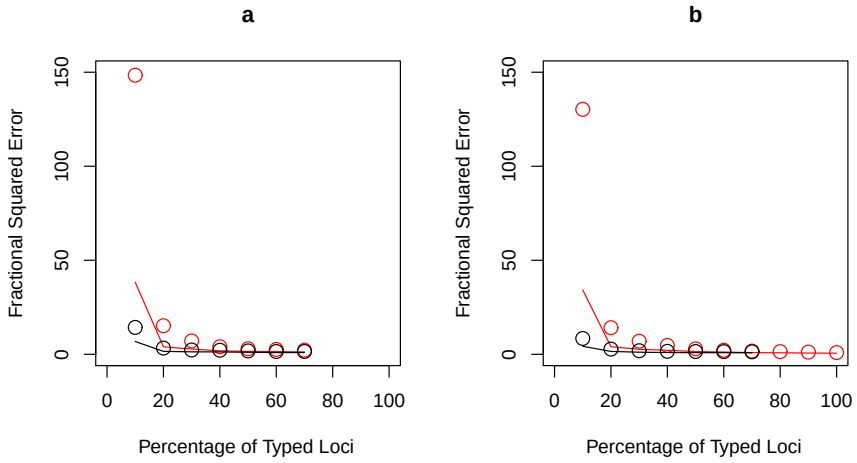

FIGURE 7.2: Fractional squared error and variance of $\hat{t}$ estimates vs. percentage of typed loci: a. real meioses b. multi-stage calibration meioses
(○/solid red line - FSE/FV of random loci, ○/solid black line - FSE/FV of fast loci)

the 100 simulated datasets. In figure 7.3a, we show the real meioses' results alongside the layered meioses' (fig. 7.3b). For the former we see that bias is considerably less for the fast markers (○) particularly at lower percentage of typed loci than the random markers (○). This was also the case for the latter. Comparing the bias in the random markers across the two methodologies shows that using the real number of meioses produces a slightly greater positive bias than the layered approach (○ in fig. 7.3ab). This is also true for the fast markers (○ in fig. 7.3ab). Importantly, we find that altering the manner in which empirical mutation rates are simulated does not change the conclusions we drew.

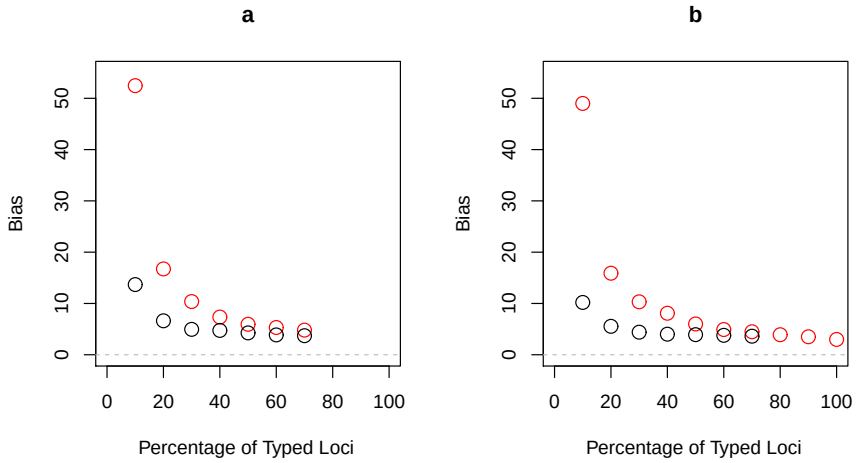

FIGURE 7.3: Bias of $\hat{t}$ estimates vs. percentage of typed loci: a. real meioses, b. multi-stage calibration meioses
(○ - random loci, ○ - fast loci)

Next, recall that we simulated mutation rates from an underlying gamma distribution. In chapter 4 we produced estimates of the shape and scale of this distribution using the data from our intermediary mutation rate review (section 2.3.2). The

resulting distribution is shown in figure 7.4 alongside the three considered in the simulation study in chapter 3. The MLEs of TMRCA suggested that the simpler infinite sites model (ISM) outperformed the stepwise mutation model (SMM) for short times with the caveat that the mutation rates were drawn from a gamma distribution that did not have high variance (chapter 3). It is clear that the estimated distribution (solid black line) is most similar to the distribution with the highest variance (red dashed line). Therefore, it is conceivable that the ISM would perform poorly for real data even if the TMRCA were low, though this model was not directly compared to our final model of TMRCA which was based on the SMM.

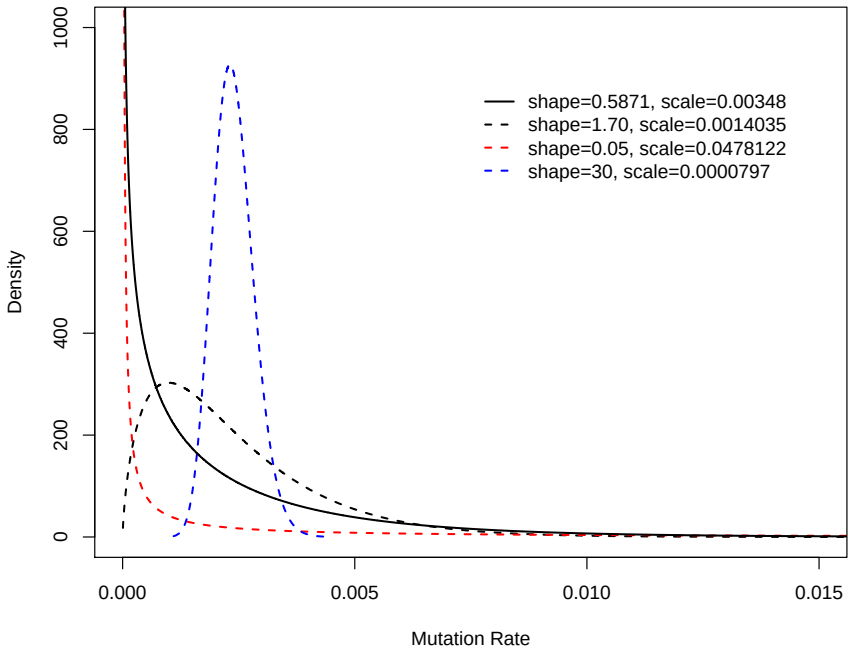

FIGURE 7.4: Mutation rate distributions in chapter 3 and based on the intermediary mutation rate review)

Valdes et al. (1993) argue that the SMM of Ohta and Kimura (1973) was based on the notion that the allele frequencies followed a normal distribution, i.e. that there would be a frequent allele flanked symmetrically by less common alleles. The 1,769 male haplotypes from King and Jobling (2009*a*) were examined in this respect. The distributions of alleles for DYS389I, DYS393, DYS434, DYS435, DYS436, DYS438 and DYS461 are fairly symmetrical. On the other hand, the distributions for the remaining ten markers are not (data not shown). Such deviations from normality may undermine the use of the SMM.

Even so, the final model of TMRCA was based on the SMM and it modelled mutation rates by defining classes of markers, namely, ascertained, non-ascertained, calibrated and typed. Estimates of TMRCA were insensitive to the percentage of non-ascertained markers, though estimates of the remaining parameters were sensitive to varying this. The total number of loci (n) was fixed at 475 on the basis of the work of Kayser et al. (2004). Importantly, of the 45 known STRs ascertained in their study, an additional eight previously known STRs were not ascertained. As such the total number of loci may be greater than 475. Sensitivity of our inferences to this value was not explored, though it is plausible that this may not affect the estimates of TMRCA. This is because the increase in the total number of loci is relatively small ($< 2\%$) and varying the percentage of non-ascertained loci did not affect the estimates of TMRCA. The number of non-ascertained loci will be proportionately increased as a consequence of increasing the total number of loci.

Our final mutation rate review produced an empirical average of 0.00234 mutations per locus per generation. The simulation study in chapter 3 makes it clear that it is not justifiable to use an average mutation rate for the estimation of TMRCA, since it will not be able to correct for recurrent mutations particularly when the true TMRCA is high. Our estimated empirical average is comparable to those estimated in the recent works of Ballantyne et al. (2010); Burgarella and Navascues (2011) (0.00263 and 0.00212, respectively). The former carried out a study of 186 Y-STRs in 1,966 father-son pairs, observing 924 mutations in 352,999 meioses, albeit utilizing a Bayesian approach for estimating site-specific mutation rates. On the other hand, Burgarella and Navascues (2011) carried out a meta-analysis of 29 published studies providing empirical estimates of 80 Y-STRs, which were modelled using binary logistic regression on the length of the STR repeat unit, the STR complexity and the relative genetic diversity. The latter was estimated by examining the variance of the STR repeat counts in the population, also based on published sources. Burgarella and Navascues argued that this variability provides information on the mutation rate. In this way, mutation rate estimates were predicted for 30 Y-STRs with no direct information in a collection of meioses.

There is evidence for a higher mutation rate to be used in the genealogical context (Athey, 2006; Chandler, 2006), although this is based on commonly typed STRs and inferred ancestral haplotypes of individuals of surname-sharing pedigrees (Kerchner, 2009). Indeed, Ballantyne et al. (2010) identified a new set of 13 rapidly

mutating (RM) Y-STRs with mutation rates of the order 0.01 (Ballantyne et al., 2011). These may aide in discrimination of haplotypes, particularly in the forensic setting, but may also be useful genealogically given that this research shows that there is lower bias in estimates of TMRCA when using the most mutable markers.

In contrast, the estimated evolutionary mutation rate of Y-STRs is typically much lower (Bianchi et al., 1998; Forster et al., 2000; Zhivotovsky et al., 2004). For example, Zhivotovsky et al.'s rate of 0.0069 has been criticised for being far too low for application to genealogical data (Athey, 2006; Chandler, 2006). This is supported by extensive published studies of father-son meioses and pedigrees, assuming the absence of selection on the Y-chromosome. However, Tyler-Smith and McVean (2003) argue that in the evolutionary context the Y-chromosome should not be treated as being neutral due to the presence of a deletion in Y-DNA associated with infertility in $\sim 2\%$ of men across the world. Whilst the details of the manner in which this selection operates are unclear, assuming that this mutation is deleterious, it would be fair to argue that this would result in fewer mutations being observed on an evolutionary scale, resulting in comparatively lower estimated mutation rates than in pedigrees where not enough time has passed for natural selection to take effect.

Another, as yet ignored, source of information within pedigrees may come from SNPs. Although SNPs are generally used to characterize ancient ancestry (e.g. to identify haplogroups), there is no reason to suppose that there may not be SNPs informative at the pedigree level. For example, Xue et al. (2009) found a 67 Y-STR match for 2 males separated by 13 generations but four single base substitutions. As whole-genome sequencing becomes increasingly cheap, it does not stretch the imagination to envisage it beginning to play a role in genealogical investigations (Xue and Tyler-Smith, 2010), providing another tool for inferring TMRCA for pairs of males.

A crucial aspect of estimating the TMRCA in a pedigree setting is the choice of the years per generation conversion factor. Recall, we formed our prior on the effective population size based on the work of Thomson et al. (2000) using a conversion factor of 25 years per generation, resulting in the prior $N_e \sim N(\mu_{N_e} = 6037, \sigma_{N_e} = 1745)$ (chapter 4). Employing the conversion rate of 31.93 years per generation of Helgason et al. (2003), we would instead have the prior $N_e \sim N(\mu_{N_e} = 4727, \sigma_{N_e} = 1367)$, i.e. with a lower mean and variance. Indeed, employing the conversion factor used by King et al. (2006), King et al. (2007) and

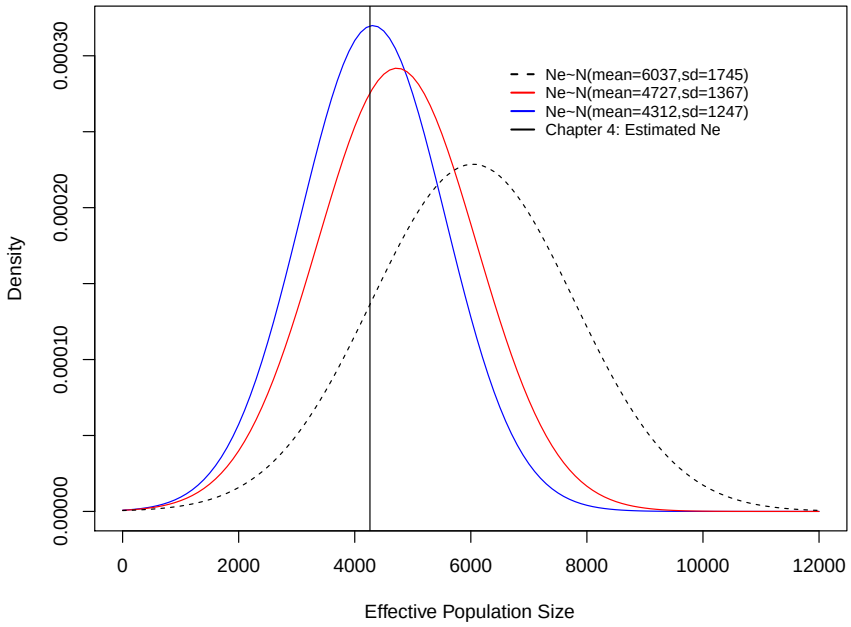

FIGURE 7.5: Comparison of prior distributions on N_e with $\hat{N}_e$ based on chapter 4

King and Jobling (2009*a*) of 35 years per generation, further shifts the distribution left and reduces its standard deviation.

A comparison of these three distributions together with the estimated effective population size, of 4,262, based on the intermediary mutation rate review is shown in figure 7.5. Whilst the estimated N_e (solid black line) lies to the left of the original prior, it lies closer to the centre of the prior as the year per generation is increased. So although a lower year per generation was employed early on in this thesis, our model has estimated N_e consistent with a prior constructed using a higher years per generation conversion.

A brief literature review found that the earliest work of Tremblay and Vezina (2000) estimated the years per generation at 35. This value is much closer to the authors more recent estimate of 33.9 (Tremblay and Vezina, 2010) which is slightly less than the estimates used to form the prior from the works of Helgason et al. (2003) (31.13 and 31.93). However, many of the papers reviewed in chapter 2 estimated the TMRCA for Y-chromosomes utilising the conversion of 25 years per generation (Bedoya et al., 2006; McEvoy and Bradley, 2006; Moore et al., 2006; Thomas et al., 1998). This will clearly have the effect of increasing estimates of the TMRCA in generations compared to using a higher conversion factor. In the genealogical context, the conversion factor is particularly important. For example, potential extension of family trees may be discounted if living descendants are

presumed to have a common ancestor who lived prior to the period of surname establishment, crucially based on an estimated TMRCA in years. In Britain, surnames are thought to have been established around 500-700 years ago. Using 25 years per generation, this would set surname establishment to have occurred 20-28 generation ago. On the other hand, this would be much lower, 14-20 generations, using the 35 years per generation conversion.

As such modelling of the conversion rate of years per generation is necessary when estimating the TMRCA in years, particularly when applied to the genealogical context. In our model, including the years per generation as an additional parameter to be inferred allows estimates of TMRCA to incorporate the uncertainty in this conversion factor. Given the plentiful supply of historical records in the period since surname establishment, one might propose that more effort be made to estimate this vital parameter more precisely. Indeed it is straightforward to specify alternate values in the prior for the years per generation conversion in our model.

In the British context, priors on TMRCA were developed on the basis of the haplogroup status of pairs of males who share the same surname, which could be the same, different or unknown haplogroups. Also in the genealogical framework, the haplogroup of males is often inferred on the basis of their STR haplotype. Methodologies to do so have been developed in both genealogical and academic circles by Athey (2007) and Schlecht et al. (2008), respectively. Application of these different methods to 119 Argentine males typed at 7 Y-STRs found that Schlecht et al.'s classifier produced fewer false positives though $\sim 48\%$ of males remained unclassified (Muzzio et al., 2011). However both methods had a much higher proportion of males classified in the R* haplogroup, reflecting the calibration of their methods on European populations. Indeed this inference of haplogroup based on haplotype has been particularly successful for Irish DNA where haplogroup R1b3 has a frequency of $> 78\%$ (Hill et al., 2000; Moore et al., 2006). However, the use of inferred haplogroups in our TMRCA model may result in applying an inappropriate prior on TMRCA. In such instances it would be advisable to use the unknown haplogroup priors rather than the same or different haplogroup priors. This is justified on the basis of the unknown haplogroup being a mixture of the other two statuses.

For male-pairs of unknown haplogroup or those sharing the same haplogroup we

also incorporate surname frequency in the prior on TMRCA. In order for the surname frequency to be adequately modelled it is necessary to compute this as the sum of the number of bearers across variant spellings of a surname. It is important to consider the alternative spellings carefully to avoid confusing different surnames as being derivatives when in fact they are not (McKinley, 1990). It was also found that pairs sharing the same haplogroup status on average coalesce around 100 generations ago ($\sim 3,500$ years BP) whilst those not sharing haplogroups status coalesced much later at around 500 generations ($\sim 17,500$ years BP). This might be evidence of the divergence of haplogroups following the last glacial maximum or the coalescence of European haplogroups during the early settlement of Europe/Eurasia, or indeed an amalgamation of the two. Hence, that the surname origin retained some explanatory power in this case was rather intriguing and remains unclear. It is worth noting that two of the three significant differences involved the ambiguous surnames (table 6.15). Since this is an amalgamation of the other surnames it is possible for some systematic difference in their TMRCA values. The last significant difference, between the occupational and nickname surnames, may be down to the fact that occupational surnames are much more widely dispersed geographically and thus male pairs in this group will not coalesce to younger nodes on average. Exploration of the nodes that are coalesced to within each surname origin may provide clearer explanation. Now the ages of the nodes in the different haplogroup priors were largely based on the estimates of Karafet et al. (2008), which assumed an out-of-Africa of 70,000 years BP. In the recent work of Cruciani et al. (2011) this node was dated much younger (38,000), through direct SNP sequencing of seven Y-chromosomes, whilst the node AT is potentially older. In light of this, it is important to note it would be straightforward to specify alternative priors (in terms of their means or SEs) in our model for any of the different haplogroup nodes. For random males the coalescent is the prior of choice, which assumes constant effective population size. This is not realistic but was used for simplicity. Indeed it is possible to incorporate alternative demographic models such as exponential growth easily into our model.

On a broader scope, it may be useful to examine if the geographical proximity of the pairs of male typed is a worthwhile explanatory of the TMRCA. In this case a distance measure, based on the longitude and latitudes of the males, might prove to be useful and thus incorporated as an additional prior in the model. In conjunction with historical surname distribution maps, such as those developed

by Steve Archer on the 1881 census returns (Archer, 2012), a complex spatio-temporal model may be developed to improve estimates of TMRCA which takes into account the spread of a surname through time as well as the surname frequency and origin. This necessitates a multi-disciplinary approach combining the input of linguists, historians, genealogists, geneticists and statisticians in order to develop a multi-factorial model of the TMRCA of pairs of males.

In the wider background of genetic genealogy, claims by genetic testing companies as to what they offer are not necessarily realistic often exploiting conventional notions of ancestry and in some cases woven into mythology (Bandelt et al., 2008). Nor are the techniques they use clearly specified along with the scientific constraints of their results (chapter 1). In such a setting, this thesis provides a transparent framework for estimation of TMRCA with clear measures of uncertainty.

Chapter 8

Conclusions

In this thesis we have developed a comprehensive model to allow for the estimation of the time to the most recent common paternal ancestor (TMRCA) for pairs of males.

We began our analysis by comparing the posterior modes of the TMRCA of the pairs from the data of King et al. (2006) using the stepwise mutation model (SMM) and infinite alleles model of Walsh (2001). The former produced larger estimates than the latter, particularly for older TMRCA. Implementing a simulation study comparing the MLE of TMRCA based on the SMM to those obtained from the infinite sites model (ISM) showed the latter performed well particularly when the TMRCA was low provided that the distribution from which the mutation rates were drawn had a low variance. Otherwise this model produced underestimates of TMRCA since it did not take into account recurrent mutation. Increasing the number of markers did not improve the estimates: in fact, in some cases it increased the amount of bias. The converse held using the SMM. Use of average mutation rates is only beneficial in the SMM when the mutation rates are drawn from a distribution with low variance, otherwise the TMRCA is underestimated. The model overestimated TMRCA when it was ≤ 400 generations but produced underestimates for larger times. This model appears to overcorrect for recurrent mutation for low values of TMRCA, when the effect is less likely to have occurred.

Since the variance of the mutation rate distribution was a key factor in influencing the MLE of the TMRCA, we next developed a framework for modelling STR mutational mechanisms in a Bayesian manner. Importantly, modelling the statistically significant difference in the proportion of increase and decrease mutations found

in our mutation rate review, was not necessary. Mutation rates were assumed to be drawn from a gamma distribution and their ascertainment was explicitly modelled using a coalescent argument depending on the total branch length, ascertainment sample size and the effective population size. Using estimates of the parameters based on the intermediate mutation rate review, the manner in which STR haplotypes were simulated was altered so as to reflect more closely real data, by introducing a multi-stage calibration process.

Thereafter we incorporated the SMM into our mutation rate model and carried out an extensive simulation study. Here the percentage of typed markers was allowed to vary for a range of TMRCA (5-100 generations). Whilst estimates of TMRCA were affected by the percentage of typed markers within the calibrated loci, the other parameters were not. Conversely, misspecifying the percentage of non-ascertained loci affected the estimates of the total branch length, the effective population size and the parameters of the mutation rate distribution, but not estimates of TMRCA.

We concluded our analysis by developing priors for TMRCA using British data from King et al. (2006); King and Jobling (2009*a*). This involved estimating the TMRCA of pairs of males within each surname. Eight priors were formed for males who share the same surname according to their haplogroup status: unknown, same or different. Furthermore, for males falling in the first two of these groups, additional information, such as the surname frequency or origin, could be specified as part of the prior on TMRCA. For male-pairs with differing haplogroups, the prior was based on the age of the ancestral SNP node, i.e. the node on the SNP tree from which the haplogroups diverge. The Bayesian estimates of the TMRCA based on these priors were then compared to those from the standard model outlined in chapter 5, using the data from King et al. (2006). The priors all led to reduced estimates of higher TMRCAs compared to the standard model. In addition, for pairs of males with the same or unknown haplogroup, priors based on the surname frequency reduced estimates of lower TMRCA (below the time of surname establishment). Also, in contrast to our earlier simulation study, which showed that the TMRCA model will produce underestimates of larger TMRCA values, the different haplogroup prior reduces the TMRCA estimates compared to the standard model.

From these results it may be argued that our model outlined in chapter 5 'overcorrects' for multiple mutations which are wiped out particularly when the TMRCA

is older. The pattern and its explanation is clear. For large TMRCA (well outside the range of genealogical interest), the information present in Y-STRs becomes weak, as a result of recurrent mutation. As a result the likelihood is rather flat, and posterior estimates become much more sensitive to prior assumptions. In this situation, it is important to use any additional information such as surname frequency or origin, but to use it carefully.

In summary this thesis has detailed a multi-faceted approach to modelling the estimation of the time to the most recent paternal ancestor of pairs of males based on their Y-STR haplotype with specific application to British surnames.

Appendix A

The Delta Method

The Delta method allows the appropriate computation of the variance of functions of random variables whose variance are themselves known through use of Taylor's Theorem.

Suppose we have two random variables, X, and Y and we wish to compute the variance of some function of these two variables, $f(x, y)$.

We begin by expanding f by Taylor's Theorem:

$$f \approx f_0 + x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y}.$$

So the variance of f is:

$$\begin{aligned}\mathrm{Var}(f) &\approx \mathrm{Var}\left(x\frac{\partial f}{\partial x} + y\frac{\partial f}{\partial y}\right) \\ &= \left(\frac{\partial f}{\partial x}\right)^2 \mathrm{Var}(x) + \left(\frac{\partial f}{\partial y}\right)^2 \mathrm{Var}(y) + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\mathrm{Cov}\left(x, y\right).\end{aligned}$$

If x and y are independent then $\mathrm{Cov}\left(x, y\right) = 0$. Hence

$$\mathrm{Var}(f) \approx \left(\frac{\partial f}{\partial x}\right)^2 \mathrm{Var}(x) + \left(\frac{\partial f}{\partial y}\right)^2 \mathrm{Var}(y).$$

For example, if $f(x, y) = xy$, then

$$\mathrm{Var}(f) \approx y^2\mathrm{Var}(x) + x^2\mathrm{Var}(y).$$

Bibliography

Adamov, D. and Karzhavin, S. (2010), 'About the influence of population size on the accuracy of TMRCA estimation, done by standard methods using STR locus complex', *Russian Journal of Genetic Genealogy* **1**, 76–99.

African Ancestry (2010), 'Homepage', Available at: <http://www.africanancestry.com/> Accessed [13 January 2010].

AncestrybyDNA (2010), 'Homepage', Available at: <http://www.ancestrybydna.com/> Accessed [13 January 2010].

Ancestry.com (2010), 'Paternal lineage test', Available at: <http://corporate.ancestry.com> Accessed [17 January 2010].

Anderson, M. J. (2001), 'Permutation tests for univariate or multivariate analysis of variance and regression', *Canadian Journal of Fisheries and Aquatic Sciences* **58**, 626–39.

Archer, S. (2012), 'Archer Software', Available at: <http://www.archersoftware.co.uk/> Accessed [10 May 2012].

Aslan, S. (2009), 'Incoherent state: the controversy over Kurdish naming in Turkey', *European Journal of Turkish Studies* **6**. [Online], Available at: <http://ejts.revues.org/index4142.html> Accessed [19 November 2011].

Athey, W. (2006), 'Haplogroup prediction from Y-STR values using a Bayesian-allele-frequency approach', *Journal of Genetic Genealogy* **3**, 34–9.

Athey, W. (2007), 'Mutation rates - who's got the right values? (editorial)', *Journal of Genetic Genealogy* **3**, i–iii.

Ballantyne, K. N., 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, M. R. and Kayser, M. (2010), 'Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications', *American Journal of Human Genetics* **87**, 341–53.

Ballantyne, K. N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S. B., Ralf, A., Vermeulen, M., de Knijff, P. and Kayser, M. (2011), 'A new future of forensic Y-chromosome analysis: Rapidly mutating Y-STRs for differentiating male relatives and paternal lineages', *Forensic Science International: Genetics* .

Ballard, D. J., Phillips, C., Wright, G., Thacker, C. R., Robson, C., Revoir, A. P. and Court, D. S. (2005), 'A study of mutation rates and the characterisation of intermediate, null and duplicated alleles for 13 Y chromosome STRs', *Forensic Science International* **155**, 65–70.

Bandelt, H. J., Yao, Y. G., Richards, M. B. and Salas, A. (2008), 'The brave new era of human genetic testing', *Bioessays* **30**, 1246–51.

Bedoya, G., Montoya, P., Garcia, J., Soto, I., Bourgeois, S., Carvajal, L., Labuda, D., Alvarez, V., Ospina, J., Hedrick, P. W. and Ruiz-Linares, A. (2006), 'Admixture dynamics in Hispanics: a shift in the nuclear genetic ancestry of a South American population isolate', *Proceedings of the National Academy of Sciences of the United States of America* **103**, 7234–9.

Berger, B., Lindinger, A., Niederstatter, H., Grubwieser, P. and Parson, W. (2005), 'Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay', *International Journal of Legal Medicine* **119**, 241–6.

Bianchi, N. O., Catanesi, C. I., Bailliet, G., Martinez-Marignac, V. L., Bravi, C. M., Vidal-Rioja, L. B., Herrera, R. J. and Lopez-Camelo, J. S. (1998), 'Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations', *American Journal of Human Genetics* **63**, 1862–71.

Bonné-Tamir, B., Korostishevsky, M., Redd, A. J., Pel-Or, Y., Kaplan, M. E. and Hammer, M. F. (2003), 'Maternal and paternal lineages of the Samaritan isolate: mutation rates and time to most recent common male ancestor', *Annals of Human Genetics* **67**, 153–64.

Bowden, G. R., Balaresque, P., King, T. E., Hansen, Z., Lee, A. C., Pergl-Wilson, G., Hurley, E., Roberts, S. J., Waite, P., Jesch, J., Jones, A. L., Thomas,

M. G., Harding, S. E. and Jobling, M. A. (2008), 'Excavating past population structures by surname-based sampling: the genetic legacy of the Vikings in northwest England', *Molecular Biology and Evolution* **25**, 301–9.

Budowle, B., Adamowicz, M., Aranda, X. G., Barna, C., Chakraborty, R., Cheswick, D., Dafoe, B., Eisenberg, A., Frappier, R., Gross, A. M., Ladd, C., Lee, H. S., Milne, S. C., Meyers, C., Prinz, M., Richard, M. L., Saldanha, G., Tierney, A. A., Viculis, L. and Krenke, B. E. (2005), 'Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America', *Forensic Science International* **150**, 1–15.

Burgarella, C. and Navascues, M. (2011), 'Mutation rate estimates for 110 Y-chromosome STRs combining population and father-son pair data', *European Journal of Human Genetics* **19**, 70–5.

Butler, J. M. (2005), *Forensic DNA Typing: Biology. Technology, and Genetics of STR Markers*, 2nd edn, Elsevier Academic Press, Amsterdam.

Butler, J. M. (2009), *Fundamentals of Forensic DNA Typing*, 1st edn, Academic Press, Amsterdam.

Butler, J. M., Buel, E., Crivellente, F. and McCord, B. R. (2004), 'Forensic DNA typing by capillary electrophoresis using the ABI Prism 310 and 3100 genetic analyzers for STR analysis', *Electrophoresis* **25**, 1397–412.

Butler, J. M. and McCord, B. R. (2006), 'Advanced topics in STR DNA analysis: Y-STRs and mtDNA', Available at: <http://www.cstl.nist.gov/strbase/ppt/AAFS2006_7_YSTRs_mtDNA.pps> Accessed [6 April 2011].

Cambridge DNA Services (2010), 'Homepage', Available at: <http://www.cambridgedna.com/> Accessed [13 January 2010].

Capelli, C., Redhead, N., Abernethy, J. K., Gratrix, F., Wilson, J. F., Moen, T., Hervig, T., Richards, M., Stumpf, M. P., Underhill, P. A., Bradshaw, P., Shaha, A., Thomas, M. G., Bradman, N. and Goldstein, D. B. (2003), 'A y chromosome census of the british isles', *Curr Biol* **13**(11), 979–84.

Chandler, J. F. (2006), 'Estimating per-locus mutation rates', *Journal of Genetic Genealogy* **2**, 27–33.

Christian, P. (2009), *The Genealogist's Internet*, 4th edn, National Archives, Kew.

Collins, A., Lonjou, C. and Morton, N. E. (1999), 'Genetic epidemiology of single-nucleotide polymorphisms', *Proceedings of the National Academy of Sciences of the United States of America* **96**, 15173–7.

Compete Inc. (2010), 'ancestry.com', Available at: <https://siteanalytics.compete.com/m/profiles/site/ancestry.com/> Accessed [26 January 2010].

Contexo (2009), 'So how do they count repeats? Anyway?', Available at: <http://www.contexo.info/DNA_Basics/microsatellite_analysis.htm> Accessed [4 March 2010].

Cooke, H. J., Brown, W. R. and Rappold, G. A. (1985), 'Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal', *Nature* **317**, 687–92.

Cottle, B. (1978), *The Penguin Dictionary of Surnames*, 2nd edn, Allen Lane, London.

Cruciani, F., Trombetta, B., Massaia, A., Destro-Bisol, G., Sellitto, D. and Scozzari, R. (2011), 'A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa', *American Journal of Human Genetics* **88**, 814–818.

Cyndi's List (2012), 'DNA, Genetics & Family Health', Available at: <http://cyndislist.com/dna> Accessed [13 May 2012].

de Souza Goes, A. C., de Carvalho, E. F., Gomes, I., da Silva, D. A., Gil, E. H., Amorim, A. and Gusmão, L. (2005), 'Population and mutation analysis of 17 Y-STR loci from Rio de Janeiro (Brazil)', *International Journal of Legal Medicine* **119**, 70–6.

Decker, A. E., Kline, M. C., Redman, J. W., Reid, T. M. and Butler, J. M. (2008), 'Analysis of mutations in father-son pairs with 17 Y-STR loci', *Forensic Science International: Genetics* **2**, e31–5.

DeGroot, M. H. and Schervish, M. J. (2002), *Probability and Statistics*, 3rd edn, Addison-Wesley, Boston.

DNA Heritage (2010), 'Homepage', Available at: <http://www.dnaheritage.com/> Accessed [13 January 2010].

Domingues, P. M., Gusmão, L., da Silva, D. A., Amorim, A., Pereira, R. W. and de Carvalho, E. F. (2007), 'Sub-Saharan Africa descendents in Rio de Janeiro (Brazil): population and mutational data for 12 Y-STR loci', *International Journal of Legal Medicine* **121**, 238–41.

Dupuy, B. M., Andreassen, R., Flones, A. G., Tomassen, K., Egeland, T., Brion, M., Carracedo, A. and Olaisen, B. (2001), 'Y-chromosome variation in a Norwegian population sample', *Forensic Science International* **117**, 163–73.

Dupuy, B. M., Stenersen, M., Egeland, T. and Olaisen, B. (2004), 'Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci', *Human Mutation* **23**, 117–24.

Edgington, E. S. (1995), *Randomization Tests*, 3rd edn, M. Dekker, New York.

Ethnoancestry (2010), 'Welcome to Ethnoancestry!', Available at: <http://www.ethnoancestry.com/> Accessed [13 January 2010].

Excoffier, L., Smouse, P. E. and Quattro, J. M. (1992), 'Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data', *Genetics* **131**, 479–91.

Family Tree DNA (2010), 'Homepage', Available at: <http://www.familytreedna.com/> Accessed [13 January 2010].

Farfán, M. J. and Prieto, V. (2009), 'Mutations at 17 Y-STR loci in father-son pairs from Southern Spain', *Forensic Science International: Genetics Supplement Series* **2**, 425–26.

Forster, P., Röhl, A., Lunnemann, P., Brinkmann, C., Zerjal, T., Tyler-Smith, C. and Brinkmann, B. (2000), 'A short tandem repeat-based phylogeny for the human Y chromosome', *American Journal of Human Genetics* **67**, 182–96.

Foster, E. A., Jobling, M. A., Taylor, P. G., Donnelly, P., de Knijff, P., Mieremet, R., Zerjal, T. and Tyler-Smith, C. (1998), 'Jefferson fathered slave's last child', *Nature* **396**, 27–8.

Fu, Y. X. and Chakraborty, R. (1998), 'Simultaneous estimation of all the parameters of a stepwise mutation model', *Genetics* **150**, 487–97.

Gail, M. H., Tan, W. Y. and Piantadosi, S. (1988), 'Tests for no treatment effect in randomized clinical-trials', *Biometrika* **75**, 57–64.

Gamerman, D. (1997), *Markov Chain Monte Carlo : Stochastic Simulation for Bayesian Inference*, 1st edn, Chapman & Hall, London.

Ge, J., Budowle, B., Aranda, X. G., Planz, J. V., Eisenberg, A. J. and Chakraborty, R. (2009), 'Mutation rates at Y chromosome short tandem repeats in Texas populations', *Forensic Science International: Genetics* **3**, 179–84.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004), *Bayesian Data Analysis*, 2nd edn, Chapman & Hall/CRC, Boca Raton, Florida.

Gelman, A. and Rubin, D. B. (1992), 'Inference from iterative simulation using multiple sequences', *Statistical Science* **7**, 457–472.

Gene Tree DNA Testing Center (2010), 'Homepage', Available at: <http://www.genetree.com/> Accessed [13 January 2010].

Genebase (2010), 'Secure online order form', Available at: <https://www.genebase.com/orderAdvanced.php> Accessed [13 January 2010].

GENUKI (2010), 'British Isles Genealogy on the Internet: Timeline', Available at: <http://homepages.gold.ac.uk/genuki/timeline/> Accessed [23 January 2010].

Gerstenberger, J., Hummel, S., Schultes, T., Hack, B. and Herrmann, B. (1999), 'Reconstruction of a historical genealogy by means of STR analysis and Y-haplotyping of ancient DNA', *European Journal of Human Genetics* **7**, 469–77.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1998), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, Boca Raton, Florida.

Giraldo, A., Martinez, I., Guzman, M. and Silva, E. (1981), 'A family with a satellited Yq chromosome', *Human Genetics* **57**, 99–100.

Gitschier, J. (2009), 'Inferential genotyping of Y chromosomes in Latter-Day Saints founders and comparison to Utah samples in the HapMap project', *American Journal of Human Genetics* **84**, 251–8.

Goedbloed, M., Vermeulen, M., Fang, R. N., Lembring, M., Wollstein, A., Ballantyne, K., Lao, O., Brauer, S., Kruger, C., Roewer, L., Lessig, R., Ploski, R., Dobosz, T., Henke, L., Henke, J., Furtado, M. R. and Kayser, M. (2009), 'Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat

polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit', *International Journal of Legal Medicine* **123**, 471–82.

Good, P. I. (2000), *Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses*, 2nd edn, Springer, New York.

Gunn, A. (2009), *Essential Forensic Biology*, 2nd edn, Wiley-Blackwell, Chichester.

Gusmão, L., Butler, J. M., Carracedo, A., Gill, P., Kayser, M., Mayr, W. R., Morling, N., Prinz, M., Roewer, L., Tyler-Smith, C. and Schneider, P. M. (2006), 'DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis', *International Journal of Legal Medicine* **120**, 191–200.

Gusmão, L. and Carracedo, A. (2003), 'Y chromosome-specific STRs', *Profiles in DNA* **10**, 3–6. [Online], Promega, Available at: <http://www.promega.com/profiles/601/profilesindna_601_03.pdf> Accessed [26 January 2011].

Gusmão, L., Sanchez-Diz, P., Calafell, F., Martin, P., Alonso, C. A., Alvarez-Fernandez, F., Alves, C., Borjas-Fajardo, L., Bozzo, W. R., Bravo, M. L., Builes, J. J., Capilla, J., Carvalho, M., Castillo, C., Catanesi, C. I., Corach, D., Di Lonardo, A. M., Espinheira, R., Fagundes de Carvalho, E., Fárfan, M. J., Figueiredo, H. P., Gomes, I., Lojo, M. M., Marino, M., Pinheiro, M. F., Pontes, M. L., Prieto, V., Ramos-Luis, E., Riancho, J. A., Souza Goes, A. C., Santapa, O. A., Sumita, D. R., Vallejo, G., Vidal Rioja, L., Vide, M. C., Vieira da Silva, C. I., Whittle, M. R., Zabala, W., Zarrabeitia, M. T., Alonso, A., Carracedo, A. and Amorim, A. (2005), 'Mutation rates at Y chromosome specific microsatellites', *Human Mutation* **26**, 520–8.

Hanks, P., Gold, D. L. and Hodges, F. (1988), *A Dictionary of Surnames*, Oxford University Press, Oxford.

Hastings, W. K. (1970), 'Monte-Carlo sampling methods using Markov chains and their applications', *Biometrika* **57**, 97–109.

Hein, J., Schierup, M. H. and Wiuf, C. (2005), *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*, Oxford University Press, Oxford.

Helgason, A., Hrafnkelsson, B., Gulcher, J. R., Ward, R. and Stefánsson, K. (2003), 'A populationwide coalescent analysis of Icelandic matrilineal and patrilineal

genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes', *American Journal of Human Genetics* **72**, 1370–88.

Herbert, M. (2009), 'Y-DNA testing company STR marker comparison chart', Available at: <http://www.gendna.net/ydnacomp.htm> Accessed [29 December 2009].

Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E. and de Knijff, P. (1997), 'Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees', *Human Molecular Genetics* **6**, 799–803.

Hill, E. W., Jobling, M. A. and Bradley, D. G. (2000), 'Y-chromosome variation and Irish origins', *Nature* **404**, 351–2.

Hohoff, C., Dewa, K., Sibbing, U., Hoppe, K., Forster, P. and Brinkmann, B. (2007), 'Y-chromosomal microsatellite mutation rates in a population sample from northwestern Germany', *International Journal of Legal Medicine* **121**, 359–63.

Holtkemper, U., Rolf, B., Hohoff, C., Forster, P. and Brinkmann, B. (2001), 'Mutation rates at two human Y-chromosomal microsatellite loci using small pool PCR techniques', *Human Molecular Genetics* **10**, 629–33.

Hornblower, M., August, M., Aunapu, G., Black, C., Daly, M., Rutherford, M. and Woodbury, R. (1999), 'Genealogy: roots mania', Available at: <http://www.time.com/time/magazine/article/0,9171,990751,00.html> Accessed [16 January 2010].

Howard, W. E. (2009*a*), 'The use of correlation techniques for the analysis of pairs of Y-STR haplotypes, part 1: Rationale, methodology and genealogy time scale', *Journal of Genetic Genealogy* **5**, 256–270.

Howard, W. E. (2009*b*), 'The use of correlation techniques for the analysis of pairs of Y-STR haplotypes, part 2: Application to surname and other haplotype clusters', *Journal of Genetic Genealogy* **5**, 271–88.

iGENEA (2010), 'Homepage', Available at: <http://www.igenea.com/> Accessed [13 January 2010].

Immel, U. D., Krawczak, M., Udolph, J., Richter, A., Rodig, H., Kleiber, M. and Klintschar, M. (2006), 'Y-chromosomal STR haplotype analysis reveals

surname-associated strata in the East-German population', *European Journal of Human Genetics* **14**, 577–82.

Internet World Stats (2010), 'World internet users and population stats', Available at: <http://www.internetworldstats.com/stats.htm> Accessed [13 January 2010].

Jobling, M. A. (2001), 'In the name of the father: surnames and genetics', *Trends in Genetics* **17**, 353–7.

Jobling, M. A., Hurles, M. and Tyler-Smith, C. (2004), *Human Evolutionary Genetics : Origins, Peoples & Disease*, Garland, New York, NY ; London.

Jobling, M. A. and Tyler-Smith, C. (1995), 'Fathers and sons: the Y chromosome and human evolution', *Trends in Genetics* **11**, 449–56.

Jobling, M. A. and Tyler-Smith, C. (2000), 'New uses for new haplotypes: the human Y chromosome, disease and selection', *Trends in Genetics* **16**, 356–62.

Karafet, T. M., Mendez, F. L., Meilerman, M. B., Underhill, P. A., Zegura, S. L. and Hammer, M. F. (2008), 'New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree', *Genome Research* **18**, 830–8.

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, G., Perez-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C. and Roewer, L. (1997), 'Evaluation of Y-chromosomal STRs: a multicenter study', *International Journal of Legal Medicine* **110**, 125–33.

Kayser, M., Kittler, R., Erler, A., Hedman, M., Lee, A. C., Mohyuddin, A., Mehdi, S. Q., Rosser, Z., Stoneking, M., Jobling, M. A., Sajantila, A. and Tyler-Smith, C. (2004), 'A comprehensive survey of human Y-chromosomal microsatellites', *American Journal of Human Genetics* **74**, 1183–97.

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger, C., Krawczak, M., Nagy, M., Dobosz, T., Szibor, R., de Knijff, P., Stoneking, M. and Sajantila, A. (2000), 'Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by

direct observation in father/son pairs', *American Journal of Human Genetics* **66**, 1580–8.

Kayser, M. and Sajantila, A. (2001), 'Mutations at Y-STR loci: implications for paternity testing and forensic analysis', *Forensic Science International* **118**, 116–21.

Kerchner, C. (2009), 'Y-STR haplotype observed mutation rates in surname projects study and log', Available at: <http://www.kerchner.com/cgi-bin/ystrmutationrate.cgi> Accessed [28 November 2011].

Kim, S. H., Kim, N. Y., Kim, K. S., Kim, J. J., Park, J. T., Chung, K. W., Han, M. S. and Kim, W. (2009), 'Population genetics and mutational events at 6 Y-STRs in Korean population', *Forensic Science International: Genetics* **3**, e53–4.

Kimura, M. and Ohta, T. (1978), 'Stepwise mutation model and distribution of allelic frequencies in a finite population', *Proceedings of the National Academy of Sciences of the United States of America* **75**, 2868–72.

King, T. E., Ballereau, S. J., Schurer, K. E. and Jobling, M. A. (2006), 'Genetic signatures of coancestry within surnames', *Current Biology* **16**, 384–8.

King, T. E., Bosch, E., Adams, S. M., Parkin, E. J., Rosser, Z. H. and Jobling, M. A. (2005), 'Inadvertent diagnosis of male infertility through genealogical DNA testing', *Journal of Medical Genetics* **42**, 366–8.

King, T. E. and Jobling, M. A. (2009*a*), 'Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames', *Molecular Biology and Evolution* **26**, 1093–102.

King, T. E. and Jobling, M. A. (2009*b*), 'What's in a name? Y chromosomes, surnames and the genetic genealogy revolution', *Trends in Genetics* **25**, 351–60.

King, T. E., Parkin, E. J., Swinfield, G., Cruciani, F., Scozzari, R., Rosa, A., Lim, S. K., Xue, Y., Tyler-Smith, C. and Jobling, M. A. (2007), 'Africans in Yorkshire? The deepest-rooting clade of the Y phylogeny within an English genealogy', *European Journal of Human Genetics* **15**, 288–93.

Klyosov, A. (2009), 'DNA genealogy, mutation rates, and some historical evidence written in the Y-Chromosome: I. Basic principles and the method', *Journal of Genetic Genealogy* **5**, 186–216.

Kurihara, R., Yamamoto, T., Uchihi, R., Li, S. L., Yoshimoto, T., Ohtaki, H., Kamiyama, K. and Katsumata, Y. (2004), 'Mutations in 14 Y-STR loci among Japanese father-son haplotypes', *International Journal of Legal Medicine* **118**, 125–31.

Lebedev, N. N. and Silverman, R. A. (1972), *Special Functions and their Applications*, rev. english edn, Dover Publications, New York.

Lee, H. Y., Park, M. J., Chung, U., Yang, W. I., Cho, S. H. and Shin, K. J. (2007), 'Haplotypes and mutation analysis of 22 Y-chromosomal STRs in Korean father-son pairs', *International Journal of Legal Medicine* **121**, 128–35.

Legendre, P. and Legendre, L. (1998), *Numerical Ecology*, 2nd english edn, Elsevier, Amsterdam.

Lehmann, E. L. (1986), *Testing Statistical Hypotheses*, 2nd edn, Wiley, New York.

Lessig, R. and Edelmann, J. (1998), 'Y chromosome polymorphisms and haplotypes in west Saxony (Germany)', *International Journal of Legal Medicine* **111**, 215–8.

Lim, S. K., Xue, Y., Parkin, E. J. and Tyler-Smith, C. (2007), 'Variation of 52 new Y-STR loci in the Y Chromosome Consortium worldwide panel of 76 diverse individuals', *International Journal of Legal Medicine* **121**, 124–7.

Liu, H. M., Chen, P. S., Chen, Y. J., Lyou, J. Y., Hu, H. Y., Lin, J. S. and Tzeng, C. H. (2007), 'Y-chromosome short tandem repeats analysis to complement paternal lineage study: a single institutional experience in Taiwan', *Transfusion* **47**, 918–26.

Malaspina, P., Ciminelli, B. M., Viggiano, L., Jodice, C., Cruciani, F., Santolamazza, P., Sellitto, D., Scozzari, R., Terrenato, L., Rocchi, M. and Novelletto, A. (1997), 'Characterization of a small family (CAIII) of microsatellite-containing sequences with X-Y homology', *Journal of Molecular Evolution* **44**, 652–9.

Manni, F., Toupance, B., Sabbagh, A. and Heyer, E. (2005), 'New method for surname studies of ancient patrilineal population structures, and possible application to improvement of Y-chromosome sampling', *American Journal of Physical Anthropology* **126**, 214–28.

Marjanovic, D., Durmic-Pasic, A., Kovacevic, L., Avdic, J., Dzehverovic, M., Haveric, S., Ramic, J., Kalamujic, B., Lukic Bilela, L., Skaro, V., Projic, P., Bajrovic, K., Drobnic, K., Davoren, J. and Primorac, D. (2009), 'Identification of skeletal remains of Communist Armed Forces victims during and after World War II: combined Y-chromosome (STR) and MiniSTR approach', *Croatian Medical Journal* **50**, 296–304.

Martin, C. L. (2003), 'Chromosome abnormalities', Available at: <http://www.subtelomeres.org/ChromosomeAbnormalities_272.html> Accessed [13 January 2010].

McElduff, F., Mateos, P., Wade, A. and Borja, M. C. (2008), 'What's in a name? The frequency and geographic distributions of UK surnames', *Significance* **5**, 189–92.

McEvoy, B. and Bradley, D. G. (2006), 'Y-chromosomes and the extent of patrilineal ancestry in Irish surnames', *Human Genetics* **119**, 212–9.

McEvoy, B., Brady, C., Moore, L. T. and Bradley, D. G. (2006), 'The scale and nature of Viking settlement in Ireland from Y-chromosome admixture analysis', *European Journal of Human Genetics* **14**, 1288–94.

McKinley, R. A. (1990), *A History of British Surnames*, Longman, London.

Meligkotsidou, L. and Fearnhead, P. (2005), 'Maximum-likelihood estimation of coalescence times in genealogical trees', *Genetics* **171**, 2073–84.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953), 'Equation of state calculations by fast computing machines', *Journal of Chemical Physics* **21**, 1087–92.

Moore, L. T., McEvoy, B., Cape, E., Simms, K. and Bradley, D. G. (2006), 'A Y-chromosome signature of hegemony in Gaelic Ireland', *American Journal of Human Genetics* **78**, 334–8.

Motluk, A. (2005), 'Anonymous sperm donor traced on internet', *New Scientist* **6**, 6.

Mulero, J. J., Chang, C. W., Calandro, L. M., Green, R. L., Li, Y., Johnson, C. L. and Hennessy, L. K. (2006), 'Development and validation of the AmpFlSTR Yfiler PCR amplification kit: a male specific, single amplification 17 Y-STR multiplex system', *Journal of Forensic Sciences* **51**, 64–75.

Muzzio, M., Ramallo, V., Motti, J. M., Santos, M. R., Lopez Camelo, J. S. and Bailliet, G. (2011), 'Software for Y-haplogroup predictions: a word of caution', *International Journal of Legal Medicine* **125**, 143–7.

National Genographic Project (2010), 'The Genographic Project', Available at: <http://genographic.nationalgeographic.com/genographic/index.html/> Accessed [13 January 2010].

NBC (2010), 'Who do you think you are?', Available at: http://www.nbc.com/who-do-you-think-you-are/> Accessed [11 March 2011].

Nei, M., Chakraborty, R. and Fuerst, P. A. (1976), 'Infinite allele model with varying mutation rate', *Proceedings of the National Academy of Sciences of the United States of America* **73**, 4164–8.

Nicholson, G., Smith, A. V., Jonsson, F., Gustafsson, O., Stefánsson, K. and Donnelly, P. (2002), 'Assessing population differentiation and isolation from single-nucleotide polymorphism data', *Journal of the Royal Statistical Society Series B, Statistical Methodology* **64**, 695–715.

Nordtvedt, K. (2008), 'More realistic TMRCA calculations', *Journal of Genetic Genealogy* **4**, 96–103.

Ohta, T. and Kimura, M. (1973), 'A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population', *Genetical Research* **22**, 201–4.

Oliveira, S., Guerra-Amorim, C., Godinho, N., Barcelos, R., Gontijo, C., Falcão-Alencar, G., Diniz, M., Ribeiro, G. and Klautau-Guimarães, M. (2008), 'Correlation of surnames and Y-chromosome in Central-Brazil', *Forensic Science International: Genetics Supplement Series* **1**, 228–9.

Olver, F. W. J. and National Institute of Standards and Technology (U.S.) (2010), *NIST Handbook of Mathematical Functions*, Cambridge University Press, Cambridge.

Onofri, V., Buscemi, L. and Tagliabracci, A. (2009), 'Evaluating Y-chromosome STRs mutation rates: A collaborative study of the Ge.F.I.-ISFG Italian Group', *Forensic Science International: Genetics Supplement Series* **2**, 419–20.

Oxford Ancestors (2010), 'Welcome to Oxford Ancestors', Available at: <http://www.oxfordancestors.com/> Accessed [13 January 2010].

Padilla-Gutierrez, J. R., Valle, Y., Quintero-Ramos, A., Hernandez, G., Rodarte, K., Ortiz, R., Olivares, N. and Rivas, F. (2008), 'Population data and mutation rate of nine Y-STRs in a mestizo Mexican population from Guadalajara, Jalisco, Mexico', *Legal Medicine (Tokyo)* **10**, 319–20.

Paternity Experts (2010), 'Homepage', Available at: <http://www.paternityexperts.com/> Accessed [13 January 2010].

Pestoni, C., Cal, M. L., Lareu, M. V., Rodriguez-Calvo, M. S. and Carracedo, A. (1999), 'Y chromosome STR haplotypes: genetic and sequencing data of the Galician population (NW Spain)', *International Journal of Legal Medicine* **112**, 15–21.

Plant, J. (2009), 'Surname studies with genetics', [Online], Available at: <http://cogprints.org/6595/> Accessed [31 August 2009].

Pollin, T. I., McBride, D. J., Agarwala, R., Schaffer, A. A., Shuldiner, A. R., Mitchell, B. D. and O'Connell, J. R. (2008), 'Investigations of the Y chromosome, male founder structure and YSTR mutation rates in the Old Order Amish', *Human Heredity* **65**, 91–104.

Pomery, C. (2007), *Family History in The Genes : Trace Your DNA and Grow Your Family Tree*, National Archives, Kew.

Pontes, M. L., Caine, L., Abrantes, D., Lima, G. and Pinheiro, M. F. (2007), 'Allele frequencies and population data for 17 Y-STR loci (AmpFlSTR Y-filer) in a Northern Portuguese population sample', *Forensic Science International* **170**, 62–7.

Powell, K. (2009), '1911 UK census now online', Available at: <http://genealogy.about.com/b/2009/01/13/1911-uk-census-now-online.htm> Accessed [17 January 2010].

PowerPlex Y Haplotype Database (2010), 'Homepage', Available at: <http://www.promega.com/techserv/tools/pplexy/Default.htm> Accessed [13 January 2010].

Raferty, A. E. and Lewis, S. M. (1998), Implementing MCMC, *in* W. R. Gilks, S. Richardson and D. J. Spiegelhalter, eds, 'Markov Chain Monte Carlo in Practice', Chapman & Hall, Boca Raton, Florida, pp. 115–130.

Reaney, P. H. and Wilson, R. M. (1997), *A Dictionary of English Surnames*, rev. 3rd edn, Oxford University Press, Oxford.

Roberts, G. O. and Rosenthal, J. S. (2007), 'Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms', *Journal of Applied Probability* **44**, 458–475.

Roberts, G. O. and Rosenthal, J. S. (2009), 'Examples of adaptive MCMC', *Journal of Computational and Graphical Statistics* **18**, 349–367.

Robinson, M. and Davidson, G. (2003), *Chambers 21st Century Dictionary*, Chambers Harrap, Edinburgh.

Rodgers, J. (2009), 'Davina doc draws 6.4m', Available at: <http://www.broadcastnow.co.uk/ratings/davina-doc-draws-64m/5003649.article> Accessed [16 January 2010].

Roots for Real (2010), 'Homepage', Available at: <http://www.rootsforreal.com/> Accessed [13 January 2010].

Sanchez-Diz, P., Alves, C., Carvalho, E., Carvalho, M., Espinheira, R., Garcia, O., Pinheiro, M. F., Pontes, L., Porto, M. J., Santapa, O., Silva, C., Sumita, D., Valente, S., Whittle, M., Yurrebaso, I., Carracedo, A., Amorim, A. and Gusmao, L. (2008), 'Population and segregation data on 17 Y-STRs: results of a GEP-ISFG collaborative study', *International Journal of Legal Medicine* **122**, 529–33.

Schlecht, J., Kaplan, M. E., Barnard, K., Karafet, T., Hammer, M. F. and Merchant, N. C. (2008), 'Machine-learning approaches for classifying haplogroup from Y chromosome STR data', *PLoS Computational Biology* **4**, e1000093.

Schmid, M., Haaf, T., Solleder, E., Schempp, W., Leipoldt, M. and Heilbronner, H. (1984), 'Satellited Y chromosomes: structure, origin, and clinical significance', *Human Genetics* **67**, 72–85.

Schneider, P. M., Meuser, S., Waiyawuth, W., Seo, Y. and Rittner, C. (1998), 'Tandem repeat structure of the duplicated Y-chromosomal STR locus DYS385 and frequency studies in the German and three Asian populations', *Forensic Science International* **97**, 61–70.

Sharif, M. (2007), Statistical models of SNP variation as applied to admixture analysis, M.Sc. dissertation, University of Glasgow.

Shi, M. S., Tang, J. P., Bai, R. F., Yu, X. J., Lv, J. Y. and Hu, B. (2007), 'Haplotypes of 20 Y-chromosomal STRs in a population sample from southeast China (Chaoshan area)', *International Journal of Legal Medicine* **121**, 455–62.

Sims, L. M., Garvey, D. and Ballantyne, J. (2009), 'Improved resolution haplogroup G phylogeny in the Y chromosome, revealed by a set of newly characterized SNPs', *PLoS One* **4**, e5792.

SMGF (2010), 'Y-Chromosome Database', Available at: <http://www.smgf.org/pages/ydatabase.jspx> Accessed [13 January 2010].

Sundaresan, S. R., Fischhoff, I. R. and Rubenstein, D. I. (2007), 'Male harassment influences female movements and associations in Grevy's zebra (Equus grevyi)', *Behavioral Ecology* **18**, 860–5.

Sykes, B. and Irven, C. (2000), 'Surnames and the Y chromosome', *American Journal of Human Genetics* **66**, 1417–9.

Tang, H., Siegmund, D. O., Shen, P., Oefner, P. J. and Feldman, M. W. (2002), 'Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition', *Genetics* **161**, 447–59.

Thomas, M. G., Parfitt, T., Weiss, D. A., Skorecki, K., Wilson, J. F., le Roux, M., Bradman, N. and Goldstein, D. B. (2000), 'Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba, the "Black Jews of Southern Africa"', *American Journal of Human Genetics* **66**, 674–86.

Thomas, M. G., Skorecki, K., Ben-Ami, H., Parfitt, T., Bradman, N. and Goldstein, D. B. (1998), 'Origins of Old Testament priests', *Nature* **394**, 138–40.

Thomson, R., Pritchard, J. K., Shen, P., Oefner, P. J. and Feldman, M. W. (2000), 'Recent common ancestry of human Y chromosomes: evidence from DNA sequence data', *Proceedings of the National Academy of Sciences of the United States of America* **97**, 7360–5.

Torres-Rodriguez, M., Martinez-Cortes, G., Paez-Riberos, L. A., Sandoval, L., Munoz-Valle, J. F., Ceballos-Quintal, J. M., Pinto-Escalante, D. and Rangel-Villalobos, H. (2006), 'Forensic potential of the STR DXYS156 in Mexican populations: inference of X-linked allele null', *Legal Medicine (Tokyo)* **8**, 52–4.

Toscanini, U., Gusmão, L., Berardi, G., Amorim, A., Carracedo, A., Salas, A. and Raimondi, E. (2008), 'Y chromosome microsatellite genetic variation in two Native American populations from Argentina: population stratification and mutation data', *Forensic Science International: Genetics* **2**, 274–80.

Tremblay, M. and Vezina, H. (2000), 'New estimates of intergenerational time intervals for the calculation of age and origins of mutations', *American Journal of Human Genetics* **66**, 651–8.

Tremblay, M. and Vezina, H. (2010), 'Genealogical analysis of maternal and paternal lineages in the Quebec population', *Human Biology* **82**, 179–98.

Tsai, L. C., Yuen, T. Y., Hsieh, H. M., Lin, M., Tzeng, C. H., Huang, N. E., Linacre, A. and Lee, J. C. (2002), 'Haplotype frequencies of nine Y-chromosome STR loci in the Taiwanese Han population', *International Journal of Legal Medicine* **116**, 179–83.

Tucker, A. (2010), 'Looking for a brilliant Christmas gift?- make a painting of your own DNA!', Available at: <http://escapethereality.wordpress.com/> Accessed [4 March 2010].

Turrina, S., Atzei, R. and De Leo, D. (2006), 'Y-chromosomal STR haplotypes in a Northeast Italian population sample using 17plex loci PCR assay', *International Journal of Legal Medicine* **120**, 56–9.

Tyler-Smith, C. and McVean, G. (2003), 'The comings and goings of a Y polymorphism', *Nature Genetics* **35**, 201–2.

US National Institute of Standards and Technology (2009), 'Summary list of Y Chromosome STR loci and available fact sheets', Available at: <http://www.cstl.nist.gov/strbase/ystr_fact.htm> Accessed [8 April 2011].

US Y-STR Database (2010), 'US Y-STR Database', Available at: <http://www.usystrdatabase.org> Accessed [13 January 2010].

Valdes, A. M., Slatkin, M. and Freimer, N. B. (1993), 'Allele frequencies at microsatellite loci: the stepwise mutation model revisited', *Genetics* **133**, 737–49.

Vermeulen, M., Wollstein, A., van der Gaag, K., Lao, O., Xue, Y., Wang, Q., Roewer, L., Knoblauch, H., Tyler-Smith, C., de Knijff, P. and Kayser, M. (2009),

'Improving global and regional resolution of male lineage differentiation by simple single-copy Y-chromosomal short tandem repeat polymorphisms', *Forensic Science International: Genetics* **3**, 205–13.

Vieira-Silva, C., Dario, P., Ribeiro, T., Lucas, I., Geada, H. and Espinheira, R. (2009), 'Y-STR mutational rates determination in South Portugal Caucasian population', *Forensic Science International: Genetics Supplement Series* **2**, 60–1.

Wakeley, J. (2009), *Coalescent Theory: An Introduction*, Roberts & Co. Publishers, Greenwood Village, Colorado.

Walsh, B. (2001), 'Estimating the time to the most recent common ancestor for the Y chromosome or mitochondrial DNA for a pair of individuals', *Genetics* **158**, 897–912.

Walsh, P. S., Fildes, N. J. and Reynolds, R. (1996), 'Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA', *Nucleic Acids Research* **24**, 2807–12.

Williams, S. R. (2005), 'Genetic genealogy: the Woodson family's experience', *Culture, Medicine and Psychiatry* **29**, 225–52.

Willis, A. J. (1970), *Genealogy for Beginners*, 2nd edn, Phillimore, London.

Wilson, I. J. and Balding, D. J. (1998), 'Genealogical inference from microsatellite data', *Genetics* **150**, 499–510.

Xu, H., Chakraborty, R. and Fu, Y. X. (2005), 'Mutation rate variation at human dinucleotide microsatellites', *Genetics* **170**, 305–12.

Xue, Y. and Tyler-Smith, C. (2010), 'The hare and the tortoise: one small step for four SNPs, one giant leap for SNP-kind', *Forensic Science International: Genetics* **4**, 59–61.

Xue, Y., Wang, Q., Long, Q., Ng, B. L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., Zhao, Y., MacArthur, D. G., Quail, M. A., Carter, N. P., Yang, H. and Tyler-Smith, C. (2009), 'Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree', *Current Biology* **19**, 1453–7.

Y Chromosome Haplotype Reference Database (2008), 'Loci', Available at: <http://www.smgf.org/pages/ydatabase.jspx> Accessed [25 April 2008].

YBase (2010), 'Homepage', Available at: <http://www.ybase.org> Accessed [13 January 2010].

YFiler (2010), 'Welcome to the YFiler Haplotype Database', Available at: <http://www.appliedbiosystems.com/yfilerdatabase> Accessed [13 January 2010].

YHRD (2010), 'Welcome to YHRD', Available at: <http://www.yhrd.org> Accessed [13 January 2010].

YMatch (2010), 'Homepage', Available at: <http://www.dna-fingerprint.com/modules.php?op=modload&name=ymatch> Accessed [13 January 2010].

YSearch (2010), 'Homepage', Available at: <http://www.ysearch.org> Accessed [13 January 2010].

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., Li, P., Yuldasheva, N., Ruzibakiev, R., Xu, J., Shu, Q., Du, R., Yang, H., Hurles, M. E., Robinson, E., Gerelsaikhan, T., Dashnyam, B., Mehdi, S. Q. and Tyler-Smith, C. (2003), 'The genetic legacy of the Mongols', *American Journal of Human Genetics* **72**, 717–21.

Zhivotovsky, L. A., Underhill, P. A., Cinnioglu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, F., Destro-Bisol, G., Spedini, G., Chambers, G. K., Herrera, R. J., Yong, K. K., Gresham, D., Tournev, I., Feldman, M. W. and Kalaydjieva, L. (2004), 'The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time', *American Journal of Human Genetics* **74**, 50–61.