UNIVERSITÉ DU QUÉBEC

MÉMOIRE PRÉSENTÉ À
L'UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

COMME EXIGENCE PARTIELLE
DE LA MAÎTRISE EN BIOLOGIE CELLULAIRE ET MOLÉCULAIRE

PAR
ALEXANDRA DOYON

DYNAMIQUE DES MARQUEURS GÉNÉTIQUES LIÉS AU SEXE DANS LA POPULATION CANADIENNE-FRANÇAISE POUR L'INTERPRÉTATION DES TRACES D'ADN EN GÉNÉTIQUE FORENSIQUE

OCTOBRE 2018

"Until recently, the Y chromosome seemed to fulfill the role of juvenile delinquent among human chromosomes — rich in junk, poor in useful attributes, reluctant to socialize with its neighbors and with an inescapable tendency to degenerate."

Mark Jobling & Chris Tyler-Smith (2003)

REMERCIEMENTS

Tout d'abord, je tiens à remercier mon directeur de recherche, Dr Emmanuel Milot, pour m'avoir donné l'opportunité de réaliser un projet de recherche dans un domaine qui me passionne depuis toujours. Je suis reconnaissante de son support, de sa confiance, de ses conseils et de toutes les opportunités de présenter mes résultats dont au congrès Haploid Markers à Berlin. Son aide pour la relecture de mon mémoire fut aussi grandement appréciée. Je remercie également mes collègues du laboratoire de génétique des populations pour leur soutien et le partage de connaissances qui m'auront été utiles tout au long de ma maîtrise. C'est une belle équipe que j'ai aimé côtoyer.

De plus, j'aimerais exprimer toute ma gratitude envers nos collaborateurs Claudia Moreau et Damian Labuda ainsi que Jean-François Lefebvre du CHU Sainte-Justine. Ils ont su répondre à tous mes questionnements et apporter de nouvelles idées à approfondir. Leur aide dans l'élaboration du modèle statistique développé dans mon projet de recherche fut grandement appréciée. Je remercie également l'équipe de BALSAC pour le partage des données et leur constant soutien. Je suis également reconnaissante envers le Laboratoire de sciences judiciaires et de médecine légale pour la génération des profils pour le chromosome Y, le partage de leur base de données, leur accueil ainsi que leurs suggestions de pistes de recherche. Ce projet de recherche n'aurait pas été possible sans la contribution de tous ces partenaires.

Je suis reconnaissante envers le Fonds de recherche du Québec–Nature et technologies, le Conseil de recherches en sciences naturelles et en génie du Canada, le Centre international de criminologie comparée, ainsi que l'Université du Québec à Trois-Rivières pour leur soutien financier. Cette recherche s'inscrit également dans l'axe de recherche sur le renseignement scientifique du Laboratoire de recherche en criminalistique.

Je remercie également ma famille, mes amis proches ainsi que mon conjoint pour leur encouragement et leur soutien tout au long de ma maîtrise. Je serai éternellement reconnaissante pour leur patience, leur écoute et leur aide dans les moments plus ardus.

Je tiens à souligner l'aide apportée par les employés de Calcul Québec ainsi que celle du Dr Maarten Larmuseau.

AVANT-PROPOS

La preuve d'ADN joue un rôle particulièrement important en science forensique. Imaginons une trace d'ADN retrouvée sur une scène de crime et que nous la comparons à l'ADN d'un suspect. Face à une concordance entre les deux profils d'ADN, il est primordial d'évaluer la valeur probante de cette concordance afin d'aider le décideur de fait (p. ex. jury) à déterminer l'importance de cette preuve d'ADN dans un cas particulier. Différents marqueurs génétiques peuvent être analysés, mais nous nous sommes concentrés plus particulièrement sur ceux situés sur l'ADN mitochondrial et le chromosome Y. L'ADN mitochondrial est surtout utilisé dans l'analyse des restes humains (p. ex. ossements, dents, cheveux, etc.), alors que le chromosome Y est principalement utilisé dans les cas d'agressions sexuelles. Pour ces marqueurs, l'évaluation du poids statistique d'une concordance est plus complexe à effectuer, ce qui limite leur utilisation en science judiciaire. Par une connaissance plus fine de la dynamique de ces marqueurs dans la population d'intérêt, il sera possible de mieux estimer la valeur probante d'une concordance impliquant ce type de marqueurs.

Mon projet de maîtrise consistait d'abord à étudier la variation spatio-temporelle des fréquences des variantes (haplotypes) de l'ADN mitochondrial et du chromosome Y dans la population canadienne-française du Québec. Pour ce faire, nous avons combiné des données généalogiques à des données moléculaires de gens connectés à la généalogie. Ceci nous a permis d'avoir une couverture populationnelle beaucoup plus grande qu'avec un échantillon de référence classique et d'étudier les fréquences à un pas de temps donné (p. ex. pour une décennie ou une génération) ou dans une région particulière. Deux prémisses sont faites, implicitement ou explicitement, par les laboratoires judiciaires lorsqu'ils recherchent un profil dans une base de données de référence pour en déterminer la fréquence. Ils supposent que les fréquences des haplotypes sont stables dans le temps et qu'elles sont homogènes dans l'espace pour une même population d'intérêt. Une deuxième partie de ma recherche était donc de tester ces prémisses. Finalement, la

dernière partie consistait à quantifier l'impact du non-respect éventuel de ces prémisses sur le calcul de la valeur probante d'une trace d'ADN.

Une grande partie de mon projet a consisté à développer un modèle permettant d'imputer un profil pour le chromosome Y à travers la généalogie à partir d'individus contemporains dont l'ADN avait été analysé. Lorsque j'ai commencé ma maîtrise, le laboratoire du Dr Emmanuel Milot en était à ses tout premiers débuts. De plus, le type de modèle statistique que nous voulions développer n'avait, à ma connaissance, jamais été fait auparavant. Mon projet de maîtrise représentait donc un grand défi à relever et a nécessité beaucoup de débroussaillage afin de comprendre les données disponibles et la programmation informatique avec le logiciel R. C'est donc à la suite de plusieurs essais que j'en suis finalement arrivée à un modèle satisfaisant.

Ce mémoire débute en introduisant les différents marqueurs génétiques, les méthodes d'interprétation actuellement utilisées et leurs limites ainsi qu'un bilan des recherches effectuées sur la population canadienne-française utilisée pour mon projet. Le principe de l'imputation de profils génétiques à travers la généalogie, le modèle développé de même que les résultats obtenus sont présentés dans le deuxième chapitre. Des précisions supplémentaires sur certaines étapes de la méthodologie utilisée sont apportées dans le troisième chapitre. Enfin, une discussion des résultats et des prochaines étapes à envisager sont abordées dans le dernier chapitre.

RÉSUMÉ

La connaissance fine de la dynamique des marqueurs liés au sexe (chromosomes X et Y, ADN mitochondrial) dans une population est essentielle à plusieurs domaines de la génétique, dont la génétique forensique. Un modèle d'interprétation fiable est nécessaire pour estimer les fréquences des variantes (haplotypes) pour ces marqueurs génétiques, puisqu'actuellement les modèles utilisés ne prennent pas suffisamment en compte la complexité génétique réelle des populations. Ainsi, des questions se posent sur la fiabilité des échantillons de référence utilisés pour estimer des fréquences dans une population. Les bases de données (régionales, nationales, voire internationales) sont-elles réellement représentatives de la population d'intérêt dans le temps et l'espace ?

Premièrement, nous avons étudié la variation spatio-temporelle des fréquences des haplotypes pour l'ADN mitochondrial et le chromosome Y pour la population canadienne-française du Québec entre 1621 et 1960 (objectif 1). Pour ce faire, un modèle combinant des données généalogiques et moléculaires a été utilisé. Les données généalogiques proviennent du registre de population BALSAC, alors que les données moléculaires pour l'ADN mitochondrial et le chromosome Y proviennent respectivement de 970 et 275 individus connectés à la généalogie. Ces modèles ont permis de démultiplier un échantillon restreint de profils génétiques pour obtenir une couverture beaucoup plus grande de la population. À partir de ces données, il était possible de tester les prémisses faites par les laboratoires judiciaires selon lesquelles les fréquences des haplotypes sont stables dans le temps et homogènes dans la population d'intérêt (objectif 2). Finalement, nous voulions quantifier l'impact forensique du non-respect éventuel de ces prémisses (objectif 3).

La diversité génétique calculée pour chaque période de 32 ans (temps moyen d'une génération) et la comparaison des probabilités de concordance fortuite entre périodes de 20 ans (âge moyen des bases de données nationales) suggèrent une stabilité temporelle des fréquences pour les deux types de marqueurs, du moins pour la plupart des haplotypes. La diversité génétique a aussi été calculée pour diverses régions et localités et nos résultats montrent une distribution non homogène des haplotypes sur le territoire qui a eu un impact non négligeable sur le calcul de la probabilité de concordance fortuite. La comparaison avec les résultats obtenus à partir des bases de données internationales indique que cette pratique n'est pas conseillée pour la population étudiée. Ainsi, nos résultats remettent en question l'utilisation d'échantillons composés de centaines ou même de milliers d'individus visant à représenter une vaste zone (p. ex. un pays) pour calculer la valeur probante d'une trace analysée avec ces marqueurs. Finalement, le modèle développé pourrait être utilisé en soutien à l'identification de restes humains via l'analyse de l'ADN mitochondrial et du chromosome Y. Les connaissances pourraient aussi être utiles en épidémiologie et en biologie évolutive.

Mots-clés : Canadiens-français, généalogie, ADN mitochondrial, chromosome Y, diversité génétique, structure génétique, probabilité de concordance fortuite

TABLE DES MATIÈRES

REMERCIEMENTS ..... **iii**

AVANT-PROPOS ..... **v**

RÉSUMÉ ..... **vii**

LISTE DES TABLEAUX ..... **xii**

LISTE DES FIGURES ..... **xiv**

LISTE DES ÉQUATIONS ..... **xvi**

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES ..... **xvii**

CHAPITRE I
INTRODUCTION **1**

1.1 Problématique 1

1.2 Génétique forensique 3

1.3 ADN humain 6

1.4 Marqueurs autosomaux 7

1.5 Marqueurs haploïdes 8

1.5.1 ADN mitochondrial 13

1.5.2 Chromosome Y 18

1.6 Interprétation des profils d'ADN 23

1.6.1 Marqueurs autosomaux 23

1.6.1.1 Probabilité de concordance fortuite 24

1.6.1.2 Rapport de vraisemblance 28

1.6.1.3 Probabilité de concordance fortuite ou rapport de vraisemblance 30

1.6.2 Marqueurs haploïdes 31

1.6.2.1 Probabilité de concordance fortuite 32

1.6.2.2 Rapport de vraisemblance 36

1.6.2.3 Bases de données pour les marqueurs haploïdes 37

1.7 Bilan des recherches généalogiques et génétiques au Québec 38

1.8 Objectifs de la recherche 41

CHAPITRE II
VARIATION SPATIOTEMPORELLE DES FRÉQUENCES HAPLOTYPIQUES POUR L'ADN MITOCHONDRIAL ET LE CHROMOSOME Y DANS UNE POPULATION CANADIENNE-FRANÇAISE ET IMPACT SUR LES PROBABILITÉS DE CONCORDANCE FORTUITE 43

2.1 Contribution des auteurs 43

2.2 Résumé de l'article 44

2.3 Article 45

Abstract 45

Keywords 46

Introduction 46

Materials and methods 50

Study population 50

Molecular data 51

Genealogical data 54

Estimation of genealogical error rates 55

Combining genealogical and molecular data to impute haplotypes 56

Non-probabilistic approach 57

Probabilistic approach 58

Molecular data coverage by periods and regions 60

Haplotype frequency, genetic diversity and random match probability 61

Results 64

Molecular data 64

Genealogical error rates for maternal and paternal lineages 65

Molecular data coverage 66

Molecular coverage by periods, regions and localities 67

Genetic diversity in time and space 69

Random match probability in time and space 72

Comparison with international databases 79

Discussion 81

Molecular data coverage and molecular mismatches in lineages 82

Genetic diversity in time and space 83

Random match probability in time and space 85
Limitations of the study 88
Conclusion 89
References 91
Supplementary materials 101
Supplementary materials and methods 101
Technical details on the estimation of haplotype probabilities and frequencies in a population 101
Supplementary tables 103
Supplementary figures 108

CHAPITRE III
MÉTHODOLOGIE—INFORMATIONS SUPPLÉMENTAIRES 114

3.1 Démarche avec le logiciel R 114
3.2 Identification des lignées maternelles et paternelles 115
3.3 Calcul de la distance génétique entre individus génotypés 115
3.4 Calcul du taux de non-concordance 117
3.5 Développement du modèle probabiliste pour l'imputation du chromosome Y dans la généalogie 117

CHAPITRE IV
DISCUSSION ET PERSPECTIVES 121

4.1 Discussion 123
4.2 Perspectives 132

RÉFÉRENCES BIBLIOGRAPHIQUES 137

ANNEXE A
Calcul de la distance génétique pour chaque paire d'individus génotypés pour l'ADN mitochondrial 156

ANNEXE B
Calcul du taux de non-concordances global et du taux de non-concordances dues aux erreurs généalogiques pour le chromosome Y 161

ANNEXE C
Ensemble de fonctions servant à faire l'imputation probabiliste du chromosome Y dans la généalogie 173

ANNEXE D
Fréquences des haplotypes de l'ADN mitochondrial dans les différentes régions analysées 192

ANNEXE E
Fréquences des haplotypes du chromosome Y observés avec 17 STR-Y dans les différentes régions analysées..... 205

ANNEXE F
Fréquences des haplotypes du chromosome Y observés avec 20 STR-Y dans les différentes régions analysées..... 224

LISTE DES TABLEAUX

Tableau		Page
1.1	Résumé des applications pour l'ADN mitochondrial et le chromosome Y...	12
1.2	Exemple d'un profil pur obtenu par l'analyse de quatre STR autosomaux accompagné des fréquences alléliques pour chaque marqueur.....	26
2.1	Molecular coverage for mitochondrial DNA and Y chromosome for each 32-year cohort	69
2.2	Differences in random match probabilities (LOD scores and odd ratios) among regions and localities for the mitochondrial DNA and the Y chromosome.	77
2.3	Comparison of random match probabilities obtained using French-Canadian frequencies, as provided by our genealogico-molecular model, vs. those from the EMPOP (mtDNA) and the YHRD (Ychr) databases for the five most common haplotypes in the Charlevoix region	80
2.S1	Y-STR haplotype probabilities after imputation within a fictive population of five individuals	101
2.S2	Geographical distribution and number of Y-STRs analyzed for the 429 men used to estimate the genealogical error rate	103
2.S3	Distribution of typed and untyped mtDNA lineages according to the number of individuals per lineage	104
2.S4	Distribution of typed and untyped 20 Y-STR lineages according to the number of individuals per lineage	104
2.S5	Distribution of typed and untyped 17 Y-STR lineages according to the number of individuals per lineage	105
2.S6	Molecular coverage for mitochondrial DNA and Y chromosome in the 24 Québec regions defined in the BALSAC register for the period 1941-1960.	106
2.S7	Molecular coverage range for mitochondrial DNA and Y chromosome in 1,188 Québec localities grouped by region for the period 1941-1960	107
D.1	Fréquences des haplotypes mitochondriaux dans la population canadienne-française calculées pour 15 régions québécoises d'après le modèle généalogico-moléculaire développé dans ce projet	193

E.1 Fréquences des haplotypes du chromosome Y observés avec 17 STR-Y dans la population canadienne-française calculées pour 11 régions québécoises d'après le modèle généalogico-moléculaire développé dans ce projet 206

F.1 Fréquences des haplotypes du chromosome Y observés avec 20 STR-Y dans la population canadienne-française calculées pour 5 régions québécoises d'après le modèle généalogico-moléculaire développé dans ce projet 225

LISTE DES FIGURES

Figure		Page
1.1	Représentation d'une cellule humaine contenant de l'ADN nucléaire et de l'ADN mitochondrial (Reproduit de Schanfield *et al.* 2014a)	6
1.2	Représentation de l'ADN mitochondrial (Reproduit de Butler 2012, p. 405-456)	14
1.3	Schéma de la transmission du chromosome Y et de l'ADN mitochondrial dans un exemple de généalogie	15
1.4	Comparaison entre deux séquences d'ADNmt et la séquence de référence rCRS (Reproduit de Butler 2012, p. 405-456).	16
1.5	Représentation des chromosomes X et Y (Reproduit de Butler 2012, p. 371-403)	19
1.6	Exemple de profil ADN obtenu par l'analyse de 17 STR-Y compris dans la trousse commerciale AmpFℓSTR® Yfiler™ (Life Technologies) (Reproduit de Kayser et Ballantyne 2014)	22
2.1	Southern Québec subdivided into 23 BALSAC Register regions	52
2.2	Schematic view of a 'genealogico-molecular model'	57
2.3	Schematic view of the probabilistic model developed for Ychr haplotypes imputation across the genealogy	59
2.4	Number of typed and untyped individuals based on their inclusion in a maternal (mtDNA) or paternal (Ychr) lineage	66
2.5	Genetic diversity at mitochondrial DNA and Y chromosome in the French-Canadian population of Québec, between 1621 and 1960	70
2.6	Genetic diversity in Québec regions between 1941 and 1960 for the mitochondrial DNA (A) and the Y chromosome (B)	71
2.7	Genetic diversity in Québec localities between 1941 and 1960 for the mitochondrial DNA (A) and the Y chromosome (B)	72
2.8	Between cohorts differences in random match probabilities (measured with LOD scores) for the mitochondrial DNA (A) and the Y chromosome with 20 Y-STRs (B) and 17 Y-STRs (C)	74

2.S1 Histogram of the count of mtDNA lineages as a function of the total number of individuals and the number of typed individuals per lineage..... 108

2.S2 Histogram of the count of mtDNA lineages as a function of the total number of individuals and the number of different haplotypes per lineage .. 108

2.S3 Histogram of the count of mtDNA haplotypes as a function of the number of lineages in which each was observed 109

2.S4 Histogram of the count of 20 Y-STRs lineages as a function of the total number of individuals and the number of typed individuals per lineage..... 109

2.S5 Histogram of the count of 20 Y-STRs lineages as a function of the total number of individuals and the number of different haplotypes per lineage .. 110

2.S6 Histogram of the count of 17 Y-STRs lineages as a function of the total number of individuals and the number of typed individuals per lineage..... 110

2.S7 Histogram of the count of 17 Y-STRs lineages as a function of the total number of individuals and the number of different haplotypes per lineage .. 111

2.S8 Distribution of the number of meioses separating two modern individuals, for pairs sharing a common ancestor for the mtDNA (n=2,958 pairs) and the Ychr (n=274 pairs) 111

2.S9 Genetic diversity in Québec regions between 1941 and 1960 for the mitochondrial DNA (A) and the Y chromosome (B) as a function of longitude 112

2.S10 Genetic diversity in Québec localities between 1941 and 1960 for the mitochondrial DNA (A) and the Y chromosome (B) as a function of longitude 113

3.1 Représentation d'une paire d'hommes ayant un ancêtre commun via des liens mère-fils..... 116

3.2 Représentation schématique de la première version du modèle d'imputation pour le chromosome Y. 119

4.1 Nombre de chaque haplotype mitochondrial de l'haplogroupe H dans la population canadienne-française entre 1717 et 1940, d'après le modèle généalogico-moléculaire développé dans ce projet 127

4.2 Exemple d'une carte de répartition possible des fréquences pour un haplotype mitochondrial au Québec 136

LISTE DES ÉQUATIONS

Équation		Page
1.1	$\hat{P}_{ii} = \hat{p}_i^2$	25
1.2	$\hat{P}_{ij} = 2\hat{p}_i\hat{p}_j$	25
1.3	$P_G = \prod_{l=1}^{L} \phi\hat{p}_{l,1}\hat{p}_{l,2} = RMP$	25
1.4	$\hat{P}_{ii} = \hat{p}_i^2 + \hat{p}_i(1-\hat{p}_i)\theta$	27
1.5	$\hat{P}_{ij} = 2\hat{p}_i\hat{p}_j(1-\theta)$	27
1.6	$P(A_iA_i\|A_iA_i) = \frac{[2\theta+(1-\theta)\hat{p}_i][3\theta+(1-\theta)\hat{p}_i]}{(1+\theta)(1+2\theta)}$	28
1.7	$P(A_iA_j\|A_iA_j) = \frac{[2\theta+(1-\theta)\hat{p}_i][\theta+(1-\theta)\hat{p}_j]}{(1+\theta)(1+2\theta)}$	28
1.8	$RV = \frac{\Pr(E\|H_p,I)}{\Pr(E\|H_d,I)}$	29
1.9	$\hat{p} = \frac{x}{N}$	32
1.10	$\hat{p} = \frac{x+1}{N+1}$	32
1.11	$\hat{p} = \frac{x+2}{N+2}$	33
1.12	$LIMSUP\ IC(95\%) = +1{,}96\sqrt{\frac{(\hat{p})(1-\hat{p})}{N}}$	33
1.13	$\sum_{k=0}^{x}\binom{N}{k}p_0^k(1-p_0)^{N-k} = \alpha$	34
1.14	$p_0 = 1-\alpha^{1/N}$	34
1.15	$P(A\|A) = \theta + (1-\theta)\hat{p}_A$	34
1.16	$RV = \frac{N}{1-\kappa}$	36

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ADN	Acide désoxyribonucléique
ADNnu	ADN nucléàire
ADNmt	ADN mitochondrial
ARN	Acide ribonucléique
ChrY	Chromosome Y
CRS	Séquence ADNmt de référence de Cambridge
EMPOP	European DNA Profiling Group Mitochondrial DNA Population Database
HV	Région hypervariable de l'ADN mitochondrial
LSJML	Laboratoire de sciences judiciaires et de médecine légale
Mini-STR	Microsatellite (*short tandem repeat*) court (<270 paires de bases)
NRC	National Research Council
NRY	Partie non recombinante du chromosome Y (*nonrecombining portion of the Y chromosome*)
PCR	Amplification en chaîne par polymérase (*polymerase chain reaction*)
rCRS	Séquence de référence de Cambridge révisée
RFLP	Polymorphisme de longueur des fragments de restriction (*restriction fragment length polymorphism*)
RMP	Probabilité de concordance fortuite (*random match probability*)
RM STR-Y	Microsatellite à mutation rapide (*rapidly mutating Y-chromosomal short tandem repeat*)
RV	Rapport de vraisemblance
SNP	Polymorphisme nucléotidique (*single-nucleotide polymorphism*)
STR	Microsatellite (*short tandem repeat*)
SWGDAM	Scientific Working Group on DNA Analysis Methods

VNTR	Répétition en tandem polymorphe (*variable number of tandem repeat*)
YHRD	Y-STR Haplotype Reference Database

CHAPITRE I

INTRODUCTION

1.1 Problématique

La preuve d'ADN est généralement considérée comme la reine des preuves par les tribunaux (Bader, 2016; Coquoz *et al.*, 2013, p. 145-200). À des fins d'illustration, le nombre d'expertises réalisées, entre 2013 et 2015, en biologie au Laboratoire de sciences judiciaires et de médecine légale (LSJML) à Montréal représente environ 50 % de toutes les expertises effectuées à ce laboratoire (Ministère de la Sécurité publique, 2015) et il y aurait plus de 300 000 analyses d'ADN effectuées chaque année aux États-Unis (Schanfield *et al.*, 2014a). L'expertise d'ADN est perçue comme étant plus objective et infaillible que les autres disciplines en science forensique (Flood, 2016; Lynch, 2003). Toutefois, la collecte de l'ADN, son analyse et son interprétation font aussi appel à une intervention humaine qui peut parfois être biaisée, ce qui n'est pas toujours considéré pleinement par les tribunaux (Flood, 2016; Lieberman *et al.*, 2008). La recherche présentée dans ce mémoire abordera plus spécifiquement les questionnements en lien avec l'interprétation des résultats, et ce, pour des marqueurs génétiques particuliers.

Typiquement, l'analyse de l'ADN par les laboratoires judiciaires porte sur les marqueurs autosomaux (situés sur les chromosomes non sexuels) puisqu'ils ont un grand pouvoir d'individualisation (Coquoz *et al.*, 2013, p. 75-144). Toutefois, ce type de marqueurs performe moins bien dans des situations particulières, telles qu'avec les traces dégradées et les mélanges d'ADN homme/femme, pour lesquelles les marqueurs haploïdes situés sur l'ADN mitochondrial (ADNmt) et le chromosome Y (chrY) ont montré leur utilité particulière (Butler, 2005, p. 145-179). D'ailleurs, les agressions sexuelles représentaient environ 16 à 20 % des dossiers traités en biologie au LSJML durant les quinze dernières années et le chrY a été utilisé dans environ 22 % des dossiers d'agression sexuelle en 2011 (Mélanie Primeau et Dominic Granger, communication

personnelle). Ces marqueurs haploïdes ont aussi été employés pour plusieurs fins telles que l'identification de soldats disparus lors de diverses guerres (Gojanović et Sutlović, 2007; Palo, Jukka U *et al.*, 2007) ou pour identifier des personnes décédées lors de certains événements tels que la tragédie de Lac-Mégantic en 2013 (Roy, 2013), le tsunami survenu en Asie du Sud en 2004 (Ladika, 2005) ou les attentats du 11 septembre 2001 à New York (Biesecker *et al.*, 2005). De plus, ils ont permis d'identifier des restes humains anciens, tels que ceux retrouvés dans de vieux cimetières (Japon (Kurosaki *et al.*, 1993), Québec (Lumbroso, 2016), Europe (Fu *et al.*, 2016)) ou ceux de la famille impériale russe (Romanov) (Coble *et al.*, 2009).

En présence d'une concordance entre le profil retrouvé dans une trace d'ADN et celui d'une personne connue, il est important d'estimer la rareté de ce profil dans la population pour évaluer la probabilité qu'une personne prise au hasard dans cette population ait le même profil que la trace (probabilité de concordance fortuite). Toutefois, cette évaluation est plus complexe pour les marqueurs haploïdes qu'autosomaux. Différents modèles statistiques ont été proposés, mais aucun n'est accepté et appliqué de manière consensuelle (Kayser, 2017). Avec une meilleure connaissance de la dynamique des marqueurs haploïdes dans la population d'intérêt, il serait possible d'obtenir des estimations plus fiables des fréquences des variantes pour ces marqueurs, et par le fait même, de la valeur probante d'une concordance entre deux profils d'ADN. C'est donc dans cette thématique que s'inscrit la recherche présentée dans ce mémoire.

D'abord, une revue de la littérature s'impose afin d'introduire la génétique forensique et aussi, de décrire l'état des connaissances sur les différents types de marqueurs utilisés dans cette discipline. Cela permettra de mieux comprendre les défis reliés à l'interprétation de profils d'ADN obtenus avec les marqueurs haploïdes liés au sexe.

1.2 Génétique forensique

La science forensique est apparue vers la fin du XIX^{e} et le début du XX^{e} siècle dans un contexte de développement des techniques et des sciences (Ribaux et Margot, Dictionnaire de Criminologie en ligne). Elle regroupe les connaissances et les techniques d'une multitude de domaines scientifiques répartis en trois groupes : chimiques (p. ex. toxicologie, fibres, peintures, verres, incendies), biologiques (p. ex. pathologie, anthropologie, odontologie, patrons de taches de sang, analyse d'ADN) et physiques (p. ex. traces digitales, documents, armes à feu, traces d'outils, traces numériques) (Houck et Siegel, 2015, p. 3-22). Ribaux et Margot (Dictionnaire de Criminologie en ligne) la définissent comme suit :

> *[La science forensique] applique une démarche scientifique et des méthodes techniques dans l'étude des traces qui prennent leur origine dans une activité criminelle, ou litigieuse en matière civile, réglementaire ou administrative. Elle aide la justice à se déterminer sur les causes et les circonstances de cette activité.*

Néanmoins, il serait inadéquat de réduire la science forensique à une simple application de techniques ayant pour seul but d'assister le juge ou le jury. Cette science étudie des traces[1] et utilise toutes autres informations pertinentes pour reconstruire un événement passé d'intérêt sécuritaire en suivant une logique indiciaire (Crispino et Houck, 2015; Fraser, 2010). Cela permet de répondre à six questions primordiales aux investigations : qui, quoi, où, quand, comment et pourquoi (Fraser, 2010). Traditionnellement, chaque trace est analysée séparément pour répondre à des besoins tactiques, mais l'utilisation de toute l'information disponible et son analyse logique permet aussi de produire du renseignement afin de soutenir les investigations, mieux orienter l'attribution des ressources policières, détecter des crimes sériels ou permettre de prioriser les prélèvements à faire sur une scène de crime (Milne, 2013, p. 43-61; Ribaux *et al.*, 2006).

---

[1] « Marque, signal ou objet, la trace est un signe apparent (pas toujours visible à l'œil nu). Elle est le vestige d'une présence et/ou d'une action à l'endroit de cette dernière » (Margot, 2011)

En 1900, Karl Landsteiner a découvert les polymorphismes protéiques (antigènes à la surface des globules rouges) des groupes sanguins ABO qui représentent une source de variation génétique entre les individus (Dumache *et al.*, 2016). Entre 1900 et 1950, environ 15 autres systèmes d'antigènes sur les globules rouges ont été découverts tels que les systèmes Rhésus, MNS, Lewis, Kell, Duffy, Kidd et Lutheran (Carracedo, 2015; Houck et Siegel, 2015, p. 3-22). Par la suite, d'autres protéines et enzymes dans le sérum ou sur les globules rouges et blancs ont été utilisées en génétique forensique, surtout pour les tests de paternité (Carracedo, 2015). Dès le début des années 1960, les antigènes des leucocytes humains ont permis une grande avancée dans les tests de paternité grâce à leur plus grand polymorphisme, mais ne permettaient pas l'analyse efficace de traces dégradées, en petite quantité ou composées d'un autre fluide biologique que le sang (Carracedo, 2015). C'est dans ce contexte historique qu'il faut situer trois importants développements technologiques concernant les marqueurs génétiques (Gill et Buckleton, 2005).

Premièrement, dans quatre articles phares publiés dans la revue *Nature* en 1985, Alec Jeffreys a introduit la technique de *DNA fingerprinting* donnant un réel élan au génotypage d'ADN dans l'identification humaine (Gill *et al.*, 1985; Jeffreys *et al.*, 1985a, 1985b, 1985c). Cette nouvelle technique impliquait l'analyse des polymorphismes de longueur des fragments de restriction (*restriction fragment length polymorphism*, RFLP) aussi appelés minisatellites (Gill et Buckleton, 2005; Goldstein et Schlötterer, 1999), soit des séquences d'ADN se répétant en tandem par motifs de 14 à 100 paires de bases (Goldstein et Schlötterer, 1999). Initialement, l'analyse se faisait avec des sondes multilocus s'hybridant sur plusieurs minisatellites simultanément (Graham, 2016).

Par la suite, la technique a été améliorée par l'utilisation de sondes plus spécifiques s'hybridant à un seul minisatellite à la fois (technique aussi connue sous le nom de VNTR pour *Variable Number of Tandem Repeats*), ce qui facilitait l'interprétation des résultats, l'évaluation du poids statistique et permettait la standardisation (Carracedo, 2015; Graham, 2016). Gill *et al.* (1985) ont démontré le potentiel d'utilisation du génotypage

d'ADN sur des traces, tandis que le premier rapport sur l'emploi de cette technique dans une affaire criminelle remonte en 1987 (Gill et Buckleton, 2005; Graham, 2016).

En même temps que le développement du génotypage d'ADN, la réaction de polymérisation en chaîne (*polymerase chain reaction*, PCR) fut développée par Kary Mullis dans les années 1980, une avancée technologique majeure en analyse d'ADN (Dumache *et al.*, 2016; Houck et Siegel, 2015, p. 3-22). Dans la foulée, les premiers marqueurs microsatellites (*short tandem repeats*, STR) ont été décrits en 1989 par Weber et May et ce système est toujours utilisé pour l'identification humaine (Dumache *et al.*, 2016; Gill et Buckleton, 2005).

La génétique forensique est une branche de la science forensique qui porte sur l'analyse de traces biologiques via le génotypage d'ADN. Elle permet entre autres d'identifier la source d'une trace d'ADN (victime/suspect), de tester des liens de parenté, d'aider à innocenter un suspect ou d'identifier des auteurs présumés de crimes en série (Dumache *et al.*, 2016; Goldstein et Schlötterer, 1999). Bien que les développements techniques et les premières utilisations aient concerné des traces d'ADN humain, la génétique forensique ne se limite pas à des applications humaines. L'ADN d'animaux (p. ex. pour investiguer sur des crimes contre la faune tels que le braconnage (Linacre et Tobe, 2013)), de plantes (p. ex. pour identifier des plants de marijuana où la production, la vente ou l'importation est interdite (Hall et Byrd, 2012)) et de bactéries (p. ex. pour localiser des tombes clandestines ou analyser des traces de salive (Leake, 2013; Metcalf *et al.*, 2017)) peut aussi être utilisé dans un contexte forensique. La génétique forensique peut ainsi se définir comme étant « la branche de la science forensique qui utilise la variation génétique présente dans les populations pour effectuer des inférences sur un événement ou un phénomène particulier à partir de traces matérielles » (Milot, Laboratoire de génétique des populations - Science forensique).

1.3 ADN humain

Les cellules humaines se composent de deux types d'ADN (Schanfield *et al.*, 2014a) dont la distinction est importante pour saisir l'objet de mes recherches. Premièrement, le noyau contient l'ADN nucléaire (ADNnu) sous forme de chromosomes totalisant 3,2 milliards de paires de bases (**Figure 1.1**) (Butler, 2010, p. 19-41; Schanfield *et al.*, 2014a). Il y a 22 paires de chromosomes non sexuels (autosomes) et une paire de chromosomes sexuels (X et Y). Les cellules somatiques sont diploïdes, c'est-à-dire qu'elles contiennent deux copies de chaque chromosome provenant de chacun des parents, alors que les cellules germinales et les gamètes (spermatozoïdes et ovules) ne contiennent qu'une seule copie de chaque chromosome et sont donc haploïdes (Schanfield *et al.*, 2014a). Deuxièmement, les cellules humaines comportent des mitochondries dans leur cytoplasme, responsables de la respiration cellulaire et de la production d'énergie nécessaire au fonctionnement des cellules, qui ont leur propre ADN, l'ADN mitochondrial (**Figure 1.1**) (Melton, 2016).

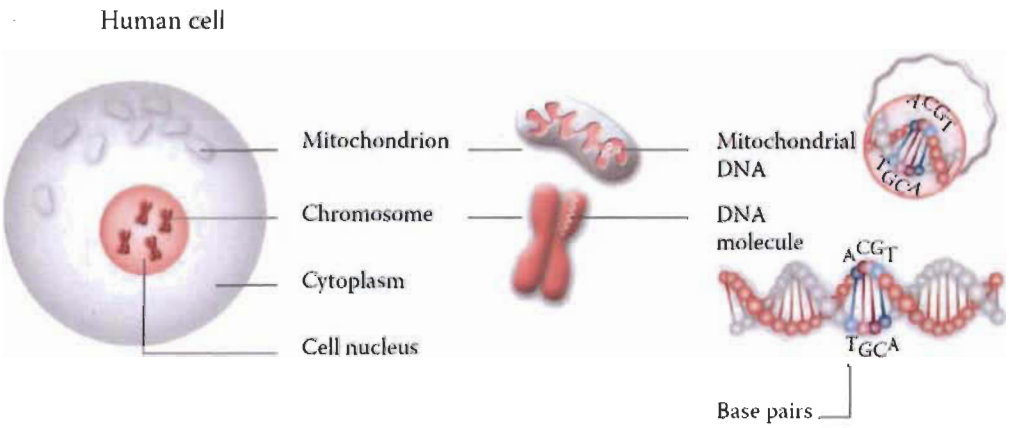

Figure 1.1 Représentation d'une cellule humaine contenant de l'ADN nucléaire et de l'ADN mitochondrial (Reproduit de Schanfield *et al.* 2014a).

En moyenne, seulement 0,3 à 0,5 % du génome humain diffère entre deux individus et cette variation provient de la diversité génétique présente dans la population résultant de mutations et de la recombinaison (Butler, 2010, p. 19-41; 2012, p. 99-139). La recombinaison se produit durant la méiose, c'est-à-dire lors de la production des spermatozoïdes et des ovules, de sorte qu'à chaque génération, les chromosomes des

enfants diffèrent de ceux de leur parent (Walsh, S. J., 2016). Les mutations surviennent quant à elles lorsque la séquence de l'ADN est modifiée soit par un changement de l'identité d'une base ou par l'ajout ou la perte d'une ou plusieurs bases (Coquoz *et al.*, 2013, p. 19-33).

Pour identifier la source d'une trace d'ADN, les marqueurs utilisés doivent nécessairement se situer dans les régions où il existe de la variation interindividuelle (Butler, 2012, p. 99-139). De plus, l'analyse de l'ADN pour fins d'identification se fait typiquement dans les régions non codantes, représentant ~95 % du génome, puisque les mutations s'y accumulent davantage, car ces régions sont moins soumises à l'action de la sélection naturelle (Walsh, S. J., 2016). Les marqueurs principalement utilisés en génétique forensique sont ceux diploïdes sur les autosomes, et de façon plus rare, ceux haploïdes situés sur l'ADN mitochondrial (ADNmt) ou le chromosome Y (chrY) (Butler, 2010, p. 19-41). Deux types de polymorphismes alléliques sont analysés, soit le polymorphisme de séquence (*single-nucleotide polymorphism*, SNP) et le polymorphisme de longueur (de type microsatellites). Les SNP sont le type de variation le plus fréquent chez l'humain, un SNP correspondant à la variation d'une paire de bases à une position précise du génome (Hameed *et al.*, 2014). Les microsatellites (*short tandem repeat*, STR) sont des séquences d'ADN contenant un nombre variable de répétitions en tandem et dont le motif répété contient entre une et huit paires de bases (Butler, 2012, p. 99-139; Graham, 2016; Walsh, S. J., 2016). Pour les raisons expliquées dans la prochaine section, les profils génétiques générés de routine dans les analyses judiciaires sont obtenus par l'analyse des STR situés sur les autosomes (Kayser et de Knijff, 2011).

1.4 Marqueurs autosomaux

L'identification humaine s'effectue le plus souvent par l'analyse des STR autosomaux accompagnée de la détermination du sexe en utilisant un marqueur sur les chromosomes X et Y (Houck et Siegel, 2015, p. 261-290). Les STR autosomaux sont hérités par les deux parents, alors le génotype à un locus peut être homozygote (deux copies du même allèle) ou hétérozygote (deux allèles différents) (Coquoz *et al.*, 2013, p.

75-144). En raison de l'indépendance de la transmission parentale des différents marqueurs (situés sur des chromosomes différents ou très loin les uns des autres sur un même chromosome), chacun fournit une information indépendante. Par conséquent, le pouvoir de discrimination s'accroît rapidement avec l'augmentation du nombre de marqueurs analysés (Coquoz *et al.*, 2013, p. 75-144). Le pouvoir de discrimination correspond à la probabilité de tirer au hasard deux individus dans la population ayant un génotype différent (Coquoz *et al.*, 2013, p. 75-144). Au fil des années, un grand nombre de STR ont été développés pour l'identification humaine, mais le besoin de mettre sur pied des bases de données et de partager l'information entre laboratoires a incité l'établissement d'un consensus sur un nombre déterminé de STR communs (Butler, 2012, p. 99-139; Butler et Hill, 2013). Typiquement entre 12 et 20 STR autosomaux sont analysés pour établir un profil ADN (Butler et Hill, 2013). Des compagnies comme Promega, Life Technologies et Qiagen ont développé des trousses commerciales permettant l'analyse de plusieurs marqueurs STR dans une même réaction PCR offrant un très grand pouvoir discriminant pour l'identification humaine (Butler et Hill, 2013; Carracedo, 2015).

1.5 Marqueurs haploïdes

Certains défis se présentent dans l'analyse des marqueurs autosomaux, la rendant difficile ou tout simplement impossible, comme la présence d'inhibiteurs de la réaction PCR ou d'ADN en petite quantité, les traces d'ADN dégradées et les mélanges d'ADN, qui ont déjà été bien décrits dans la littérature (Butler, 2005, p. 145-179). Les deux derniers éléments sont brièvement abordés ci-dessous, car ils ont un intérêt particulier pour ce mémoire.

Les traces d'ADN peuvent être dégradées à cause de facteurs environnementaux (p. ex. rayons ultraviolets, humidité, conditions très acides, présence d'agents oxydants ou de nucléases), du temps écoulé entre le dépôt d'une trace et sa découverte et des circonstances de l'événement (Butler, 2005, p. 145-179; Houck et Siegel, 2015, p. 261-290; Zietkiewicz *et al.*, 2012). La dégradation cause la rupture des brins d'ADN en plus petits fragments,

ce qui peut empêcher l'amplification par PCR d'un ou plusieurs marqueurs et donner un résultat non concluant. Le développement de nouveaux marqueurs comme les mini-STR, des STR plus courts que ceux traditionnels, a permis d'obtenir des résultats exploitables dans l'analyse de ce type de traces. Toutefois, il n'est pas possible d'analyser un grand nombre de mini-STR dans une même réaction PCR puisque les produits sont de tailles similaires (Butler, 2005, p. 145-179; Wiegand et Kleiber, 2001).

Un mélange d'ADN signifie que l'ADN d'une trace provient de plus d'un contributeur (Butler, 2015, p. 159-182). Il peut s'agir d'un mélange de sang, de salive, de sperme, de cellules épithéliales, etc., ou de n'importe quelle combinaison de tissus biologiques (Harbison, 2016; Schanfield *et al.*, 2014b). La quantité d'ADN et les allèles provenant de chaque personne, de même que la quantité totale d'ADN et le nombre de STR autosomaux analysés, influencent la détection du mélange et son interprétation (Butler, 2005, p. 145-179). De plus, les différents contributeurs peuvent avoir des allèles communs rendant difficile la détermination des génotypes de chacun et pouvant résulter en un profil partiel (Jamieson, 2016). L'interprétation est facilitée lorsqu'un des contributeurs potentiels est connu ou lorsque l'un d'entre eux a laissé davantage d'ADN que les autres (Jamieson, 2016). Un cas particulier de mélange d'ADN est celui impliquant une femme et un ou plusieurs hommes comme dans le cas de prélèvements intimes suite à une agression sexuelle sur une victime féminine. Généralement, le traitement de ce type de mélanges implique une extraction différentielle permettant de séparer les cellules épithéliales (féminines et potentiellement masculines) des spermatozoïdes provenant du sperme, suivi de l'analyse des STR autosomaux dans chaque fraction obtenue tel que décrit dans la **Section 1.4** (Harbison, 2016). Toutefois, cette technique ne permet pas toujours d'obtenir un profil d'ADN masculin valide pour comparaison, par exemple lorsque le sperme ne contient pas de spermatozoïdes (azoospermie, vasectomie), soit le type de cellules le plus abondant dans ce fluide biologique, ou bien quand la trace est âgée ou qu'elle contient des inhibiteurs. Dans le premier cas, le profil masculin peut parfois être généré à partir de la fraction épithéliale obtenue suite à l'extraction différentielle. Néanmoins, cette fraction contient souvent de l'ADN féminin en proportion beaucoup

plus importante que l'ADN masculin, masquant ainsi les allèles masculins (Butler, 2012, p. 371-403; Coquoz *et al.*, 2013, p. 145-200; Harbison, 2016).

Face à ces limitations des marqueurs autosomaux, il a fallu développer des méthodes alternatives pour obtenir des résultats exploitables dans ces situations. C'est dans ce contexte que les marqueurs haploïdes situés sur l'ADN mitochondrial (ADNmt) et le chromosome Y (chrY) ont été introduits au milieu des années 1990 et qu'ils ont depuis montré leur utilité pour ces cas plus problématiques (Carracedo, 2015). D'abord, l'ADNmt est présent dans toutes les mitochondries, donc en de multiples copies dans une seule cellule. Pour cette raison, il offre un avantage par rapport à l'ADNnu lorsque la trace est dégradée ou lorsqu'elle ne contient que très peu de matériel génétique, car on a plus de chance d'en retrouver quelques copies intactes (Parson, 2015). De plus, sa structure circulaire et son emplacement dans les mitochondries participent à sa plus grande résistance (Butler, 2012, p. 405-456; Melton, 2016). Des études ont également démontré que l'analyse de l'ADNmt était réussie dans plus de 92 % des cas impliquant des cheveux et dans 84 % des cas où l'ADNnu ne fournissait pas ou peu d'information (Parson, 2015). Le chrY, pour sa part, permet de fournir un profil masculin, ce qui est très utile en présence d'un mélange d'ADN homme/femme, une situation très fréquente avec les prélèvements obtenus dans le cas d'agressions sexuelles (Butler, 2012, p. 371-403; Coquoz *et al.*, 2013, p. 145-200). Tel qu'expliqué précédemment, la méthode traditionnelle d'extraction différentielle n'est pas toujours possible ni utile. Le chrY permet donc d'analyser ce type de mélanges, peu importe les liquides biologiques impliqués et sans avoir à faire une extraction différentielle, puisque la méthode repose sur l'amplification de STR situés sur le chrY (Ballantyne, J. et Hanson, 2016) (toutefois, son pouvoir d'individualisation est moindre, comme nous le verrons à la **Section 1.6**). En effet, dans une étude d'Olofsson *et al.* (2011), l'analyse des STR-Y a permis d'identifier la présence d'ADN masculin dans 45 % des prélèvements intimes pour lesquels aucun sperme n'était détecté et parmi ces cas, un profil Y complet ou partiel a pu être obtenu dans 29 % des prélèvements. Aussi, le chrY peut aider à déterminer le nombre de contributeurs masculins dans un mélange puisque chacun contribue généralement par un seul allèle à chaque STR (Ballantyne, J. et Hanson, 2016).

L'ADNmt et le chrY, en plus d'être utilisés en génétique forensique, ont des applications dans d'autres domaines, résumées dans le **Tableau 1.1**. Il est important de mentionner que pour plusieurs de ces applications, il n'est pas toujours possible de faire une comparaison directe pour identifier une personne (p. ex. aucun échantillon de référence disponible). Les marqueurs haploïdes permettent donc à des parents rapprochés ou éloignés de servir de référence pour l'identification puisque tous les individus d'une même lignée parentale partagent généralement la même variante génétique tel qu'expliqué ci-dessous (Butler, 2012, p. 405-456).

Tableau 1.1

Résumé des applications pour l'ADN mitochondrial et le chromosome Y

Applications	ADN mitochondrial	Chromosome Y	Références
Analyser des traces dégradées ou présentes en petite quantité (p. ex. cheveux, poils, dents, ossements)	X	X	(Allen *et al.*, 1998; Butler *et al.*, 2003)
Déterminer le sexe d'une personne		X	(Chang *et al.*, 2003; Steinlechner *et al.*, 2002)
Tester la paternité même dans des cas où le profil du père est inconnu		X	(Kayser et Sajantila, 2001; Rolf *et al.*, 2001)
Identifier des restes humains dont ceux à caractère historique	X	X	(Ambers *et al.*, 2014; Anđelinović *et al.*, 2005; Foster *et al.*, 1998; Gerstenberger *et al.*, 1999; Gill *et al.*, 1994; Just *et al.*, 2011; Sullivan *et al.*, 1992)
Identifier des victimes de désastres de masse, de guerres, d'attaques terroristes ou d'accidents	X	X	(Alonso *et al.*, 2005; Anjos *et al.*, 2004; Daoudi *et al.*, 1998; Holland *et al.*, 1993)
Étudier l'histoire des populations humaines (évolution, migration, etc.)	X	X	(Lippold *et al.*, 2014; Roewer *et al.*, 2005; Sajantila *et al.*, 1996; Underhill et Kivisild, 2007)
Étudier des maladies dues à des mutations	X	X	(Ambulkar *et al.*, 2015; Rosenberg *et al.*, 2016; Wallace *et al.*, 1999)
Déterminer l'origine ethnique d'une personne	X	X	(Underhill et Kivisild, 2007; Wetton *et al.*, 2005)
Analyser des mélanges d'ADN (p. ex. homme/femme)		X	(Betz *et al.*, 2001; Dekairelle et Hoste, 2001; Martín *et al.*, 2000)
Déterminer le nombre de contributeurs masculins dans un mélange d'ADN		X	(Daniels *et al.*, 2004; Hanson et Ballantyne, 2004; Hanson *et al.*, 2006)

Ainsi, pour toutes les raisons mentionnées dans cette sous-section, les marqueurs haploïdes représentent un réel intérêt tant pour la génétique forensique que pour d'autres domaines de la génétique. Ci-après, la nature de ces marqueurs ainsi que leur mode de transmission sont abordés avant d'introduire les problématiques pouvant être rencontrées dans l'utilisation de ceux-ci.

1.5.1 ADN mitochondrial

L'ADNmt est circulaire, double brin et, chez l'humain, composé de 16 569 paires de bases (**Figure 1.2**) (Coquoz *et al.*, 2013, p. 145-200). Une cellule peut contenir entre 1 000 et 10 000 copies d'ADNmt (~10 ADNmt/mitochondrie et ~100-1 000 mitochondrie/cellule) (Coquoz *et al.*, 2013, p. 145-200). Cet ADN peut être divisé en deux régions : régions codante et non codante. La première contient l'information génétique de 37 gènes codant pour des protéines impliquées dans la phosphorylation oxydative ou pour des acides ribonucléiques (ARN) (22 ARN de transfert, 2 ARN ribosomaux et 13 protéines) (Melton, 2016). La région non codante, appelée aussi région de contrôle, contient 1 122 paires de bases et est dépourvue de gènes (Butler, 2012, p. 405-456; Melton, 2016). Elle comprend des séquences cis impliquées dans la transcription ainsi que le site d'origine de la réplication de l'un des deux brins de l'ADNmt (Butler, 2012, p. 405-456; Holt et Reyes, 2012).

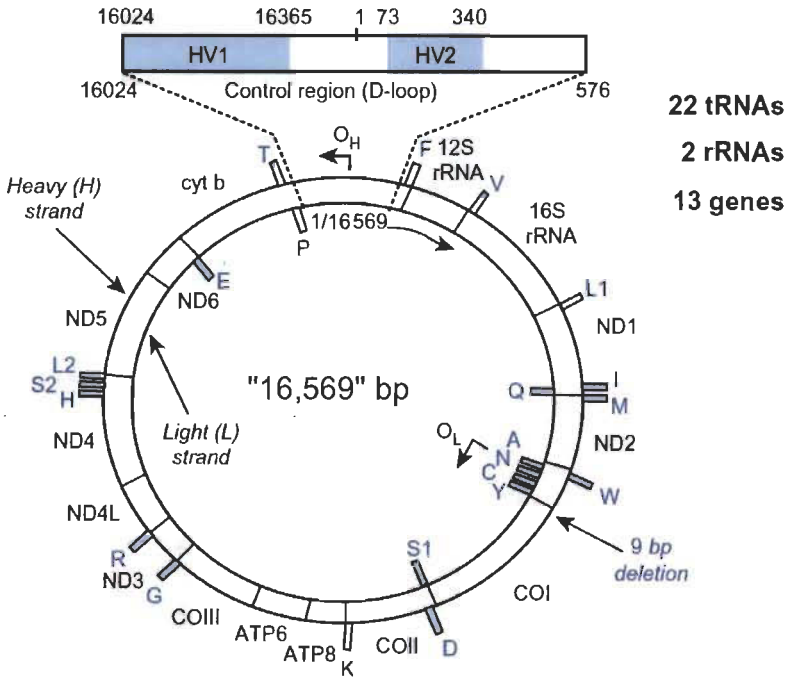

Figure 1.2 Représentation de l'ADN mitochondrial (Reproduit de Butler 2012, p. 405-456).

L'ADNmt est transmis par la mère autant aux garçons qu'aux filles (**Figure 1.3**) (Coquoz *et al.*, 2013, p. 145-200). Les hommes pourraient également le transmettre à leurs enfants dans de rares cas (Cree *et al.*, 2009; Schwartz et Vissing, 2002). L'ADNmt ne subit pas de recombinaison; il est plutôt hérité en un seul bloc. Ainsi, la séquence reste inchangée entre une mère et ses enfants, sauf si une mutation survient dans la cellule germinale, et les différentes variantes observées dans la population sont appelées haplotypes (Coquoz *et al.*, 2013, p. 145-200; Schanfield *et al.*, 2014a). De cette façon, tous les individus faisant partie d'une même lignée maternelle (**Figure 1.3**) présentent le même haplotype mitochondrial (Houck et Siegel, 2015, p. 261-290). Un haplotype n'est donc pas unique à une personne, contrairement à un génotype pour l'ADNnu (sauf rares exceptions), ce qui limite d'emblée le pouvoir discriminant et le poids statistique qu'aurait une concordance entre deux profils pour l'ADNmt (Butler, 2012, p. 405-456) (voir la **Section 1.4** pour la définition du pouvoir de discrimination).

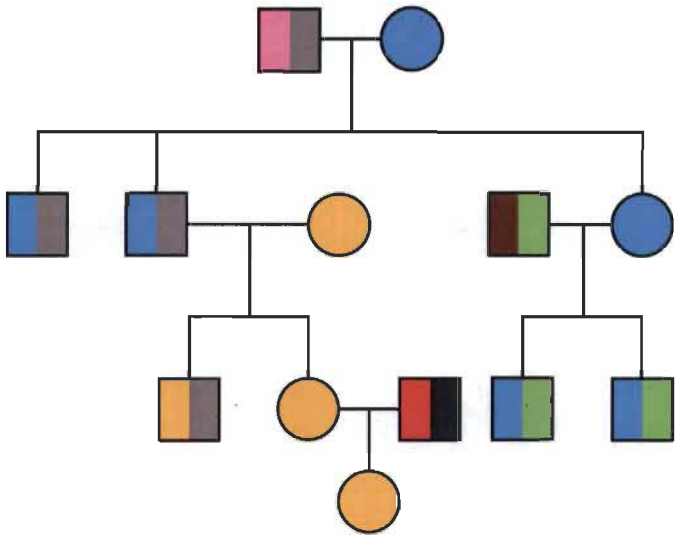

Figure 1.3 **Schéma de la transmission du chromosome Y et de l'ADN mitochondrial dans un exemple de généalogie.**
Les carrés représentent les hommes et les cercles, les femmes. Chaque couleur représente un haplotype différent du chrY ou de l'ADNmt transmis le long des lignées paternelles ou maternelles respectivement. Les femmes ont une seule couleur puisqu'elle ne possède que de l'ADNmt. Les hommes ont deux couleurs, soit pour l'ADNmt (à gauche) et pour le chrY (à droite).

En absence de recombinaison, les mutations représentent la seule source de diversité génétique pour l'ADNmt (Schanfield *et al.*, 2014a). Le taux de mutations de l'ADNmt est de cinq à dix fois plus grand que celui de l'ADNnu puisqu'il y a moins de mécanismes de réparation de l'ADN, notamment aucun mécanisme assurant que la bonne base soit ajoutée lors de la réplication (Butler, 2012, p. 405-456; Holland et Lauc, 2014). La région de contrôle (non codante) de l'ADNmt se caractérise par un plus grand polymorphisme entre les individus qu'ailleurs sur le génome mitochondrial dû à une pression de sélection plus faible sur les mutations (Holland et Lauc, 2014; Parson, 2015). Deux portions sont reconnues pour avoir la plus grande diversité génétique, soit les régions hypervariables I (HVI : positions 16 024-16 365) et II (HVII : positions 73-340) (Coquoz *et al.*, 2013, p. 145-200). Le taux de mutations des régions HVI et HVII est d'environ 0,0043 par transmission mère-enfant signifiant qu'en moyenne 4 enfants sur 1 000 présenteront une différence de séquence à au moins un nucléotide avec leur mère (Sigurðardóttir *et al.*, 2000). Une troisième région hypervariable (HVIII : positions 438-574) est parfois

analysée, ce qui permet d'avoir plus de sites polymorphiques afin de distinguer des échantillons pouvant être identiques aux régions HVI et HVII (Bini *et al.*, 2003; Lutz *et al.*, 2000).

L'ADNmt est dépourvu de séquences répétitives de type STR, alors son analyse en génétique forensique porte sur les polymorphismes de séquence (SNP) principalement situés dans les régions HVI et HVII (Coquoz *et al.*, 2013, p. 145-200; Parson, 2015). Le premier séquençage de l'ADNmt complet a été effectué par Anderson *et al.*, en 1981, qui ont établi la séquence de référence dite de Cambridge (CRS), révisée par la suite en 1999 (rCRS) (Anderson *et al.*, 1981; Andrews *et al.*, 1999). Suite au séquençage d'un ADNmt, il est commun de rapporter les différences entre les séquences obtenues et la rCRS en précisant la position et la nature des substitutions (**Figure 1.4**) (Melton, 2016). Ceci facilite les comparaisons entre plusieurs traces ou entre une trace et un échantillon de référence ainsi que la recherche dans des bases de données. Dans l'exemple de la **Figure 1.4**, l'haplotype de la trace et celui de l'échantillon d'une personne connue sont identiques et diffèrent de la rCRS à deux positions : 16 093 et 16 129. L'interprétation d'une telle concordance trace-référence sera abordée à la **Section 1.6**.

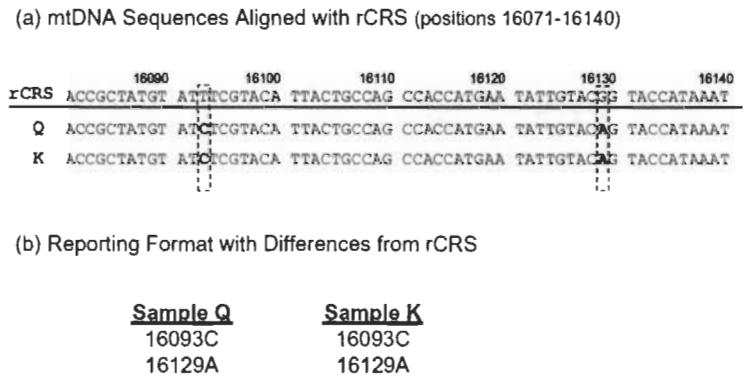

Figure 1.4 **Comparaison entre deux séquences d'ADNmt et la séquence de référence rCRS (Reproduit de Butler 2012, p. 405-456).**
Deux ADNmt ont été séquencés, soit celui provenant de la trace (Q) et celui provenant d'une personne connue (K). En A, les séquences Q et K comprises entre les positions 16 071 et 16 140 sont comparées à la rCRS et les différences sont encadrées. En B, les différences entre les séquences Q et K et la rCRS sont rapportées sous une forme succincte en précisant la position et la nature de la substitution.

Généralement, une personne présente un haplotype mitochondrial dominant (i.e. celui qui est concrètement observé par séquençage), ce que l'on appelle homoplasmie. Toutefois, il est connu que les cellules peuvent également contenir plusieurs autres ADNmt présents à de très faibles fréquences (Coquoz *et al.*, 2013, p. 145-200). Un ou plusieurs de ces ADNmt peuvent se multiplier suffisamment pour être détectés au séquençage conjointement avec la version dominante. Ce phénomène, appelé hétéroplasmie, s'observe souvent dans les cellules d'individus porteurs de mutations associées à des maladies puisque les cellules vont avoir l'ADNmt sauvage et la version mutante (Wonnapinij *et al.*, 2008). Des études chez la vache et l'humain ont montré que les versions mutantes pouvaient rapidement devenir fixées dans les prochaines générations (individu devient homoplasmique) probablement à cause d'un effet de goulot d'étranglement (Chinnery *et al.*, 2000). Un goulot d'étranglement fait en sorte que seules quelques molécules d'ADNmt de la mère dans l'ovocyte (cellule sexuelle) se multiplient durant la maturation de l'ovocyte en ovule mature prêt à être fécondé. Toutefois, l'hétéroplasmie peut aussi persister d'une génération à l'autre (Chinnery *et al.*, 2000). Elle est souvent considérée comme un état temporaire puisque l'ADNmt « mutant » peut devenir l'allèle dominant dans un tissu et être le seul observable au séquençage ou il peut complètement disparaître (Coquoz *et al.*, 2013, p. 145-200). Deux types d'hétéroplasmie existent. L'hétéroplasmie de longueur (variation dans la longueur de la séquence) survient dans ~50 % des individus, et ce, principalement au niveau des régions riches en cytosines (*C-stretch*) (Irwin *et al.*, 2009). Ceci est probablement dû au fait que la polymérase soit plus encline à faire des erreurs de réplication par glissement (*slippage*) dans ces régions d'ADN. D'un autre côté, l'hétéroplasmie de séquence (variation dans l'identité d'une base à une position donnée) survient dans 1 à 15 % des analyses (~1 % pour le sang, 10-15 % pour les cheveux et 1-15 % pour les ossements) (Butler, 2012, p. 405-456; Melton, 2016). Il est aussi important de noter que ces différents haplotypes peuvent être transmis à la descendance (Chinnery *et al.*, 2000; Holland et Lauc, 2014). L'interprétation des résultats devient plus complexe en présence d'hétéroplasmie, mais, d'un autre côté, avec des méthodes de détection appropriées, elle augmente le pouvoir de discrimination pour l'identification humaine (Coquoz *et al.*, 2013, p. 145-200; Schanfield *et al.*, 2014a). À ce

jour, il n'y a pas de pratique universelle dans les laboratoires judiciaires pour rapporter ce type de résultat et en faire l'interprétation (Parson *et al.*, 2014).

Finalement, certaines parties du génome mitochondrial se sont dupliquées dans l'ADNnu sous forme de pseudogènes (Butler, 2012, p. 405-456). Comme les séquences de l'ADNmt original et des copies nucléaires sont semblables, elles pourraient être amplifiées simultanément par PCR. Généralement, les pseudogènes amplifiés devraient être présents en petite quantité par rapport aux amplicons de l'ADNmt (Ramos *et al.*, 2009). Toutefois, des auteurs ont rapporté que certaines conditions peuvent favoriser l'amplification des pseudogènes telles que la présence de plusieurs copies d'un même pseudogène, la région de l'ADNmt étudiée, l'utilisation d'amorces insuffisamment spécifiques à l'ADNmt et le nombre de cycles de PCR utilisé (Butler, 2012, p. 405-456; Parr *et al.*, 2006; Ramos *et al.*, 2009). D'ailleurs, Yao *et al.* (2008) ont montré que des séquences d'ADN provenant de pseudogènes ont souvent été faussement interprétées comme de nouvelles variantes de l'ADNmt dans des études portant sur des maladies mitochondriales.

1.5.2 Chromosome Y

Le chromosome Y (chrY), spécifique aux hommes, est le deuxième plus petit chromosome avec ~60 millions de nucléotides (Kayser et Ballantyne, 2014). Il est divisé en trois régions : euchromatine, hétérochromatine et régions pseudoautosomales (**Figure 1.5**). L'euchromatine est composée de gènes, alors que l'hétérochromatine en est dépourvue et est plutôt riche en séquences répétitives (Ballantyne, J. et Hanson, 2016). Ces deux régions combinées forment la région non recombinante (NRY) qui représente 95 % du chrY (**Figure 1.5**) (Ballantyne, J. et Hanson, 2016; Kayser et Ballantyne, 2014). Comme pour l'ADNmt, les mutations représentent la seule source de diversité génétique dans la région NRY (Butler, 2012, p. 371-403). C'est sur cette région présentant un grand polymorphisme que se concentrent les analyses en génétique forensique. Les régions pseudoautosomales PAR1 et PAR2, situées aux extrémités, représentent un peu plus de

2,8 millions de nucléotides et se recombinent avec des régions homologues sur le chromosome X (Graves *et al.*, 1998).

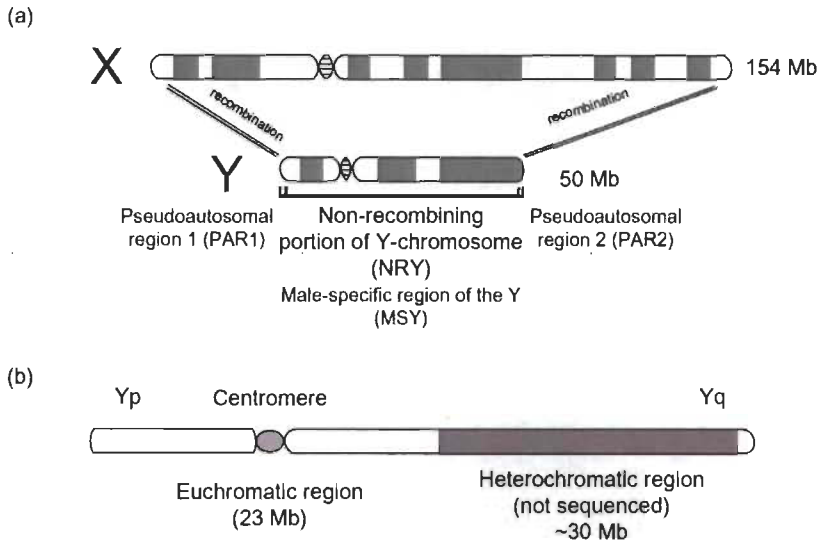

Figure 1.5 Représentation des chromosomes X et Y (Reproduit de Butler 2012, p. 371-403).

Le chrY est transmis uniquement de père en fils. En l'absence de recombinaison, un fils hérite ainsi du même haplotype que son père (Butler, 2012, p. 371-403). Tous les individus faisant partie d'une même lignée paternelle partagent donc le même haplotype sauf en cas de mutation (**Figure 1.3**) (Coquoz *et al.*, 2013, p. 145-200). Deux types de polymorphismes sont analysés sur le chrY, soit les séquences répétitives de type STR (STR-Y) et les polymorphismes de séquence (SNP-Y). Le taux de mutations des STR-Y couramment utilisés est d'environ 10^{-4} à 10^{-3} par marqueur par transmission, ce qui est du même ordre de grandeur que les STR autosomaux (Ballantyne, K. N. *et al.*, 2010; Willuweit et Roewer, 2018a). Cela signifie que pour un marqueur donné, il y aura une mutation à tous les 1 000 à 10 000 hommes d'une même lignée. Pour les SNP-Y, ce taux est d'environ 10^{-9} à 10^{-8} par transmission (Butler, 2012, p. 371-403; Xue *et al.*, 2009). Ballantyne *et al.* (2010) ont découvert 13 nouveaux STR-Y à mutation rapide (RM STR-Y), dont le taux de mutations est environ 10 fois plus grand que pour les STR-Y traditionnels (~10^{-2} par transmission). Dans deux études subséquentes, ces auteurs ont obtenu un pouvoir de discrimination global et un pouvoir de discrimination pour individus apparentés de 90,4 et 5,5 % pour un jeu de 17 STR-Y et de 98,3 et 29 % pour les 13 RM

STR-Y (Ballantyne *et al.*, 2012, 2014), montrant le potentiel de ces derniers pour discriminer des hommes de la même lignée paternelle (voir la **Section 1.4** pour la définition du pouvoir de discrimination).

Une mutation typique dans un STR (autosomal ou Y) consiste en l'ajout ou le retrait d'une unité répétitive complète (parfois plus) (Butler, 2015, p. 403-444). Ballantyne *et al.* (2010) et Goedbloed *et al.* (2009) ont étudié respectivement 186 et 17 STR-Y et conclu que la proportion de gains de répétitions par mutation est sensiblement la même que celle des pertes. De plus, le gain ou la perte de plusieurs unités répétitives à la fois est plus rare.

De manière générale, l'analyse de plusieurs marqueurs sur le chrY, dont la transmission est non-indépendante, revient statistiquement à ne regarder qu'un seul marqueur. Un ensemble de STR-Y donné définit un haplotype, alors que les SNP-Y permettent d'identifier l'haplogroupe auquel appartient un haplotype (Butler, 2012, p. 371-403). Un haplogroupe est un ensemble d'haplotypes similaires ayant un ancêtre évolutif commun. En génétique forensique, les STR-Y sont davantage utilisés que les SNP-Y puisqu'ils mutent plus rapidement et que cela permet d'avoir un meilleur pouvoir de discrimination (Kayser et Ballantyne, 2014). Les SNP-Y peuvent servir, entre autres, à étudier les migrations des populations humaines ou déterminer l'origine ethnique d'une personne, mais ils commencent également à être intégrés dans les trousses commerciales forensiques (voir plus bas).

Comme pour les marqueurs autosomaux, des ensembles de STR-Y communs permettent de comparer des profils générés par différents laboratoires. En 1997, le premier ensemble comprenant 9 STR-Y, nommé *minimal haplotype*, a été établi (Kayser *et al.*, 1997). En 2003, le Scientific Working Group on DNA Analysis Methods (SWGDAM) a défini un nouvel ensemble comprenant les neuf STR-Y précédents et deux nouveaux STR-Y (« SWGDAM core loci ») (Ballantyne, J. et Hanson, 2016; SWGDAM Y-STR Subcommittee, 2007). Différentes trousses commerciales permettant d'analyser plusieurs STR-Y dans une même réaction PCR ont été développées et comprennent les marqueurs mentionnés ci-haut et quelques autres (p. ex. les 17 marqueurs du AmpFℓSTR® Yfiler™

de Life Technologies et les 12 du PowerPlex® Y de Promega) (Ballantyne, J. et Hanson, 2016; Kayser et Ballantyne, 2014). À ce jour, environ 4 500 STR ont été répertoriés sur le chrY (Balanovsky, 2017), incitant l'élaboration de nouvelles trousses plus discriminantes, tels que PowerPlex® Y23 (23 STR-Y) de Promega et Yfiler® Plus (27 STR-Y) de Life Technologies qui incluent respectivement 2 et 6 RM STR-Y (Kayser, 2017; Willuweit et Roewer, 2015). La compagnie Illumina propose la trousse ForenSeq™ permettant d'analyser simultanément 58 STR (autosomaux, X et Y) et jusqu'à 172 SNP (Churchill *et al.*, 2016; Just *et al.*, 2017).

À la suite d'une PCR ciblant des STR-Y, les différents amplicons sont séparés par électrophorèse en fonction de leur poids moléculaire, qui dépend directement du nombre de répétitions de chaque allèle présent (Kayser et Ballantyne, 2014). Une échelle allélique, migrant dans les mêmes conditions électrophorétiques, permet de convertir la longueur des fragments, mesurée en paires de bases, en allèles identifiés selon le nombre de répétitions qu'ils comportent. L'ensemble des allèles des différents marqueurs pris ensemble forme un haplotype Y. Dans l'exemple de la **Figure 1.6**, l'haplotype à 17 STR-Y est rapporté comme suit, où le nombre entre parenthèses indique l'identité de l'allèle (nombre de répétitions) : DYS456 (16), DYS389I (13), DYS390 (24), DYS389II (29), DYS458 (17), DYS19 (14), DYS385 (11,14), DYS393 (13), DYS391 (11), DYS439 (11), DYS635 (23), DYS392 (13), Y_GATA_H4 (12), DYS437 (15), DYS438 (12), DYS448 (19). Cet haplotype peut ensuite être comparé à celui d'une trace d'ADN ou d'une personne connue ou être ajouté/recherché dans une base de données.

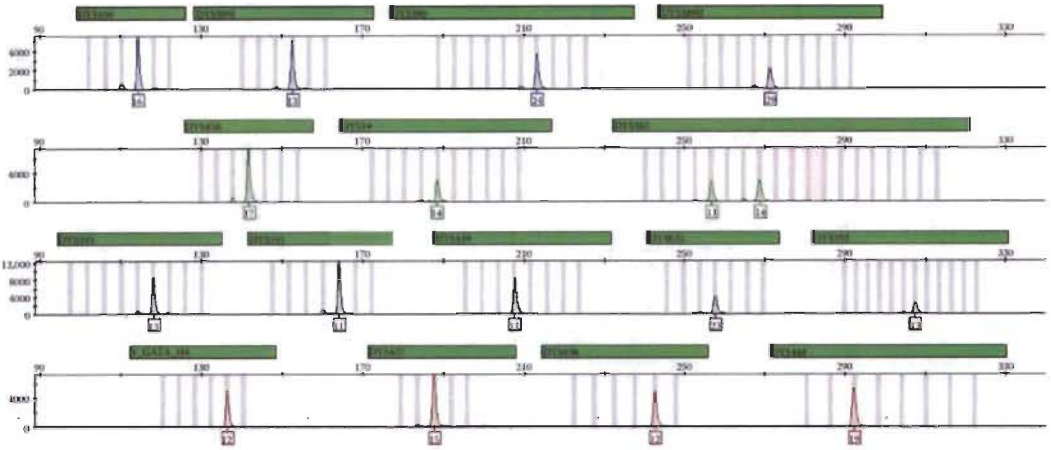

Figure 1.6 Exemple de profil ADN obtenu par l'analyse de 17 STR-Y compris dans la trousse commerciale AmpFℓSTR® Yfiler™ (Life Technologies) (Reproduit de Kayser et Ballantyne 2014).

Certaines analyses misent sur le fait que les individus d'une même lignée paternelle possèdent le même haplotype Y, comme les tests de paternité, l'identification de restes humains en comparant le profil génétique des disparus avec celui des hommes de leur présumée famille, et la prédiction du patronyme à partir d'une trace d'ADN laissée sur une scène de crime afin d'identifier la victime ou un groupe de suspects potentiels (Calafell et Larmuseau, 2016; Coquoz *et al.*, 2013, p. 145-200). Toutefois, les paternités extra-conjugales, les adoptions (aussi vrai pour l'ADNmt) et les fécondations in vitro ou assistées peuvent compliquer l'interprétation des résultats (Calafell et Larmuseau, 2016). En effet, un garçon pourrait présenter un haplotype différent de celui de son père social. Il faut distinguer ces cas de ceux où des différences sont dues à des mutations. Par ailleurs, dans le cas où le vrai père biologique n'est pas le père présumé, mais lui est apparenté, l'enfant pourrait avoir le même haplotype que le père présumé.

La majorité des STR-Y présentent un allèle unique par individu. Toutefois, quelques-uns ont plus d'une copie du locus et donc parfois plus d'un allèle (Kayser et Ballantyne, 2014). Cela peut être causé par une duplication ou une triplication du STR-Y sur le même chromosome, ces copies peuvent être de longueurs différentes et dans ce cas, être détectées séparément (Butler, 2015, p. 403-444). Le scientifique doit alors déterminer si cet événement est courant pour le(s) marqueur(s) analysé(s) ou s'il se retrouve plutôt en présence d'un mélange d'ADN de plusieurs hommes, sachant que dans ce dernier cas,

plusieurs marqueurs montreraient plus d'un allèle. D'un autre côté, des délétions de plusieurs nucléotides, comme dans les régions entourant un STR-Y, pouvant aller jusqu'à plus d'un millier de nucléotides, peuvent se produire et faire échouer l'amplification d'un STR-Y, et même entraîner la perte de l'amélogénine Y servant à déterminer le sexe masculin (Butler, 2012, p. 371-403). Ces duplications et délétions peuvent être transmises de père en fils (Butler, 2012, p. 371-403; 2015, p. 403-444).

1.6 Interprétation des profils d'ADN

Dans cette section, l'interprétation des profils d'ADN en génétique forensique est décrite, en faisant d'abord un bref retour sur les marqueurs autosomaux pour lesquels les outils statistiques sont reconnus et appliqués de manière consensuelle par la communauté scientifique. Par la suite, les méthodes utilisées pour les marqueurs haploïdes ainsi que les différentes bases de données disponibles pour ces marqueurs seront abordées. Finalement, une description des défis que représente ce type de marqueurs pour l'interprétation est présentée.

1.6.1 Marqueurs autosomaux

De manière générale, le profil génétique obtenu à partir d'une trace d'ADN peut être comparé à celui d'autres traces ou de personnes connues, ou être recherché dans une base de données. Il y a trois conclusions possibles selon le résultat de cette comparaison : exclusion, inclusion, non-concluant (SWGDAM, 2009, 2013, 2017). Une exclusion survient lorsque des profils purs diffèrent, ou lorsqu'un profil n'est pas compatible avec une combinaison de profils, et qu'ils ne peuvent donc pas, en théorie, provenir de la même source. Une inclusion se produit lorsque les deux profils comparés sont identiques ou qu'ils ne peuvent être exclus comme pouvant faire partie d'un mélange. Finalement, la comparaison est non-concluante si l'information disponible est insuffisante ou trop ambiguë pour trancher en faveur d'une exclusion ou d'une inclusion. Cela peut arriver, par exemple, lorsqu'on soupçonne que des allèles n'ont pas été détectés (*dropouts*).

Lorsqu'une concordance entre un profil obtenu d'une trace d'ADN et celui d'une personne connue sera utilisée comme élément de preuve à des fins judiciaires, il est nécessaire d'en évaluer la valeur probante avec une approche probabiliste. Cela nécessite d'évaluer la rareté du profil observé sur la trace dans la population d'intérêt (Butler, 2015, p. 281-308), soit la population des individus susceptibles d'avoir laissé leur ADN, à cet endroit, sans considération pour leur origine géographique ou ethnique (Butler, 2015, p. 281-308; Coquoz *et al.*, 2013, p. 303-413; Parson *et al.*, 2014; Szabolcsi *et al.*, 2015). La rareté d'un profil est obtenue en estimant sa fréquence dans la population d'intérêt en se servant des bases de données de référence sur la fréquence des allèles de chaque marqueur analysé (Butler, 2015, p. 281-308). Deux méthodes sont principalement utilisées pour rapporter ce résultat qui ont été bien décrites dans Butler (2015, p. 281-308): la probabilité de concordance fortuite (RMP, ou *Random Match Probability*) et le rapport de vraisemblance (RV, ou *likelihood ratio*).

1.6.1.1 Probabilité de concordance fortuite

Avant d'introduire les méthodes de calcul de la probabilité de concordance fortuite, il est primordial de faire la distinction entre deux concepts pouvant être confondus de par leur similarité. Premièrement, la probabilité de correspondance (P_C) est la probabilité que deux individus tirés au hasard dans la population aient le même profil sans égard à l'identité du profil (Coquoz *et al.*, 2013, p. 303-413). Cela correspond à « 1 moins le pouvoir de discrimination », ce dernier étant abordé dans la **Section 1.4**. La littérature anglophone réfère parfois à cette probabilité sous l'appellation de *match probability* (*Pm*) (Jobling et Gill, 2004; Kayser et de Knijff, 2011). Ensuite, il existe une certaine confusion quant à la définition de la RMP dans la littérature puisque différents auteurs n'expriment pas les mêmes idées derrière ce calcul même si, au final, le résultat mathématique est le même (voir par exemple Butler 2015, p. 281-308 et Weir 2001). Pour ce mémoire, la définition retenue est que la RMP représente la probabilité de tirer une personne au hasard dans la population ayant le même profil que celui observé sur la trace (Coquoz *et al.*, 2013, p. 303-413). Pour la RMP, l'identité du profil est donc importante contrairement à la P_C.

Elle implique aussi que celui-ci soit observé deux fois sous l'hypothèse que le suspect ne soit pas la source de la trace.

D'abord, la fréquence d'un génotype à un marqueur génétique donné est calculée selon les **Équations 1.1** et **1.2** à partir des fréquences des allèles (Coquoz *et al.*, 2013, p. 303-413).

$$\hat{P}_{ii} = \hat{p}_i^2 \tag{1.1}$$

$$\hat{P}_{ij} = 2\hat{p}_i\hat{p}_j \tag{1.2}$$

Où :
$\hat{p}_i$: estimateur de la fréquence de l'allèle *i* dans la population
$\hat{p}_j$: estimateur de la fréquence de l'allèle *j* dans la population
$\hat{P}_{ii}$: estimateur de la fréquence du génotype homozygote *ii*
$\hat{P}_{ij}$: estimateur de la fréquence du génotype hétérozygote *ij*

La fréquence d'un profil complet (pour l'ensemble des marqueurs) est obtenue en multipliant la fréquence des génotypes de chacun des marqueurs (règle du produit), une opération permise par la prémisse d'indépendance de la ségrégation mendélienne des marqueurs (**Équation 1.3**). La fréquence d'un profil ainsi calculée correspond à la RMP dans le cas précis où le fait d'avoir observé le profil une première fois ne modifie pas la connaissance sur la probabilité de le tirer une seconde fois. Cela est vrai dans une population suivant la loi de Hardy-Weinberg (voir plus loin).

$$P_G = \prod_{l=1}^{L} \phi\hat{p}_{l,1}\hat{p}_{l,2} = RMP \tag{1.3}$$

Où :
P_G : probabilité d'un profil génétique composé de L loci
L : nombre de loci
ϕ : variable indicatrice d'homozygotie (1 pour un homozygote et 2 pour un hétérozygote)
$\hat{p}_{l,1}$: estimateur de la fréquence du premier allèle au locus l
$\hat{p}_{l,2}$: estimateur de la fréquence du deuxième allèle au locus l
RMP : probabilité de concordance fortuite

Prenons en exemple un profil pur (non mélangé) obtenu d'une trace d'ADN et qui concorde avec celui d'un suspect (les fréquences alléliques pour la population caucasienne aux États-Unis sont tirées de Butler (2015, p. 497-518)) (**Tableau 1.2**).

Tableau 1.2

Exemple d'un profil pur obtenu par l'analyse de quatre STR autosomaux accompagné des fréquences alléliques pour chaque marqueur

STR	Génotype de la trace et du suspect	Fréquence du premier allèle	Fréquence du deuxième allèle
VWA	14/14	0,093	0,093
D21S11	29/30	0,202	0,283
TH01	7/9	0,194	0,119
FGA	20/21	0,123	0,179

La RMP serait donc calculée avec l'**Équation 1.3** comme suit :

$$RMP = \prod_{l=1}^{L} \phi \hat{p}_{l,1} \hat{p}_{l,2}$$

$$RMP = (0{,}093 * 0{,}093) * (2 * 0{,}202 * 0{,}283) * (2 * 0{,}194 * 0{,}119) * (2 * 0{,}123 * 0{,}179)$$

$$RMP = 2{,}00 \text{ X } 10^{-6}$$

Le résultat est rapporté en énonçant que la probabilité de tirer au hasard une personne dans la population possédant le même profil que celui observé sur la trace est d'une sur 2 millions.

Les **Équations 1.1** à **1.3** se basent sur la loi de Hardy-Weinberg reposant sur les prémisses suivantes (Butler, 2015, p. 281-308). D'abord, les organismes sont diploïdes et ont une reproduction sexuée. L'appariement des gamètes lors de la reproduction est aléatoire par rapport à l'identité de leurs allèles et les allèles à différents gènes sont transmis de manière indépendante à la génération suivante, résultant dans l'absence de corrélation entre 1) les deux copies d'un gène au même locus et 2) les gamètes transmis à

différents loci. Il n'y a pas de mutations au locus considéré et celui-ci n'est pas sous sélection naturelle. Finalement, la population est très grande et il n'y a pas de migration.

Aussi, la bonne population doit être sélectionnée pour estimer la fréquence d'un profil génétique et elle ne doit pas présenter de structure génétique (variation des fréquences alléliques entre des sous-populations) (Butler, 2015, p. 281-308). Dans le cas contraire, la règle du produit pourrait sous-estimer la RMP (Foreman et Evett, 2001). Wright (1951) a introduit le concept de F_{ST} repris par la suite, entre autres, par Weir et Cockerham (1984) et Weir et Hill (2002), qui mesure la probabilité que deux allèles soient identiques en raison de leur transmission par un ancêtre commun dans la même sous-population. Il peut aussi être considéré comme une mesure de la similarité des allèles dans une sous-population en comparaison de celle entre allèles provenant de sous-populations différentes (Buckleton *et al.*, 2016). Une valeur de F_{ST} élevée indique une différence importante entre sous-populations, alors qu'une faible valeur de F_{ST} indique plutôt une similarité entre elles (Butler, 2015, p. 239-279). Le National Research Council (NRC) a proposé le terme de correction θ, équivalent au F_{ST} (National Research Council, 1996, p. 125-165), et c'est ce terme qui est utilisé par le SWGDAM dans ses recommandations sur l'interprétation (SWGDAM, 2017). Dans le cas où la population ne suit pas la loi de Hardy-Weinberg, les **Équations 1.1** et **1.2** sont modifiées comme suit afin d'intégrer la correction θ (**Équations 1.4** et **1.5**) (Butler, 2015, p. 239-279; National Research Council, 1996, p. 90-121):

$$\hat{P}_{ii} = \hat{p}_i^2 + \hat{p}_i(1 - \hat{p}_i)\theta \qquad \textbf{(1.4)}$$

$$\hat{P}_{ij} = 2\hat{p}_i\hat{p}_j(1 - \theta) \qquad \textbf{(1.5)}$$

Où θ représente le facteur de correction pour tenir compte de la structure génétique.

Balding et Nichols (1994) ont proposé deux autres équations pour calculer la RMP à chaque locus qui ont été reprises par le NRC (**Équations 1.6** et **1.7**) (les équations ont été modifiées afin de conserver le même style de notation).

$$P(A_iA_i|A_iA_i) = \frac{[2\theta + (1-\theta)\hat{p}_i][3\theta + (1-\theta)\hat{p}_i]}{(1+\theta)(1+2\theta)} \quad \textbf{(1.6)}$$

$$P(A_iA_j|A_iA_j) = \frac{2[\theta + (1-\theta)\hat{p}_i][\theta + (1-\theta)\hat{p}_j]}{(1+\theta)(1+2\theta)} \quad \textbf{(1.7)}$$

Où :
$P(A_iA_i|A_iA_i)$: probabilité d'observer le génotype homozygote (noté A_iA_i) sachant qu'il a déjà été observé une fois
$P(A_iA_j|A_iA_j)$: probabilité d'observer le génotype hétérozygote (noté A_iA_j) sachant qu'il a déjà été observé une fois

Ces deux équations prennent en compte le fait que d'observer un profil une première fois dans une sous-population augmente la chance de l'observer une seconde fois dans cette même sous-population, ce que ne permettent pas les **Équations 1.1**, **1.2**, **1.4** et **1.5**. Il est à noter que lorsque $\theta = 0$, les **Équations 1.4** à **1.7** sont équivalentes aux **Équations 1.1** et **1.2**. De plus, lorsque $\theta > 0$, la RMP calculée avec les **Équations 1.6** et **1.7** donne un résultat supérieur à celle calculée avec les **Équations 1.1**, **1.2**, **1.4** et **1.5** (Buckleton *et al.*, 2011). Le terme θ utilisé dans les **Équations 1.4** à **1.7** doit être évalué empiriquement pour la population d'intérêt et l'ensemble de STR utilisés (Butler, 2015, p. 239-279). De manière générale, une valeur de 0,01, pour de grandes populations, ou de 0,03, pour des populations plus isolées, petites ou endogames est utilisée (National Research Council, 1996, p. 90-121).

1.6.1.2 Rapport de vraisemblance

Le rapport de vraisemblance (RV), basé sur le théorème de Bayes, peut aussi être utilisé pour exprimer la valeur probante d'une preuve d'ADN. Le RV permet de confronter deux (ou plus) hypothèses en évaluant la vraisemblance des observations (*E*, de l'anglais *evidence*) sous l'une ou l'autre de ces hypothèses, souvent celles de la poursuite (H_p) et de la défense (H_d) (Butler, 2015, p. 281-308; Coquoz *et al.*, 2013, p. 303-413). En

génétique forensique, cela permet plus spécifiquement de comparer sous différentes hypothèses, la probabilité d'observer les profils génétiques, par exemple de la trace d'ADN et du suspect qui concordent (**Équation 1.8**). Une valeur de RV supérieure à 1 signifie que l'hypothèse de la poursuite est davantage soutenue par la trace d'ADN qui, au contraire, soutient davantage l'hypothèse de la défense quand le RV est inférieur à 1. Dans une situation où le RV est de 1, la trace ne soutient aucune hypothèse plus que l'autre (Coquoz *et al.*, 2013, p. 303-413).

$$RV = \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)} \qquad \textbf{(1.8)}$$

Où *I* comprend les autres informations (non génétiques) disponibles sur le cas, comme les informations circonstancielles.

Prenons en exemple le profil du **Tableau 1.2** obtenu d'une trace d'ADN et qui concorde avec le profil du suspect X. Le RV est calculé en utilisant l'**Équation 1.8** comme suit :

H_p : X est la source de l'ADN de la trace

H_d : Une personne inconnue est la source de l'ADN de la trace

$$RV = \frac{\Pr(E|H_p, I)}{\Pr(E|H_d, I)}$$

$$RV = \frac{1}{RMP} = \frac{1}{2{,}00\ X\ 10^{-6}} = 500\ 000$$

Dans cet exemple, $\Pr(E|H_p,I)$ peut être simplifiée à 1 puisque si le suspect est la source de la trace, alors il est certain que les deux profils seront identiques (excluant certaines complications comme les erreurs de génotypage). De plus, $\Pr(E|H_d,I)$ équivaut à la RMP, ce qui représente l'application la plus simple du RV. Le RV a une valeur de 500 000, signifiant que H_p est plus soutenue par la trace d'ADN que H_d. Le résultat se rapporte en énonçant que l'observation (concordance entre le profil de la trace et celui du

suspect X) est 500 000 fois plus probable si l'hypothèse H_p est vraie que si l'hypothèse H_d est vraie (Butler, 2015, p. 281-308).

1.6.1.3 Probabilité de concordance fortuite ou rapport de vraisemblance

Le calcul de la RMP est simple et s'applique bien dans la majorité des situations rencontrées, ce pourquoi la RMP est une méthode répandue pour exprimer la valeur probante d'une concordance entre profils d'ADN, surtout en Amérique du Nord. Toutefois, dans un contexte de preuve judiciaire, il est plus difficile d'y intégrer des informations telles que la possibilité que le suspect et le vrai auteur du crime soient apparentés (Butler, 2015, p. 281-308). Son interprétation est aussi limitée en présence d'une trace d'ADN provenant de plusieurs contributeurs (Centre of Forensic Sciences, 2017). De plus, il survient fréquemment que le décideur de fait (juge/jury) commette une erreur d'interprétation de la RMP appelée la transposition de la conditionnelle (en anglais, *prosecutor's fallacy*) (Coquoz *et al.*, 2013, p. 303-413). Avec la RMP de $2{,}00 \times 10^{-6}$ calculée dans la **Section 1.6.1.1**, une telle erreur reviendrait à dire qu'il y a 1 chance sur 2 millions que le suspect ne soit pas la source de la trace d'ADN. De son côté, le piège du défenseur (en anglais, *attorney's fallacy*) consiste plutôt à supposer que tous les individus avec le même profil que celui observé sur la trace dans la population ont la même chance que le suspect d'être la source de cette trace d'ADN (Butler, 2015, p. 281-308). Avec la RMP calculée dans la **Section 1.6.1.1** et une population fictive de 4 millions d'individus, le défenseur énoncerait qu'en réalité la probabilité que le suspect soit la source de la trace est de 1 sur 8 ($2{,}00 \times 10^{-6} \times 4\,000\,000$). Cet exemple n'est présenté qu'à titre informatif puisqu'une telle erreur d'interprétation ne serait pas en faveur du défenseur dans ce cas-ci. Cohen (2017) a d'ailleurs montré que les erreurs d'interprétation sont très fréquentes dans les tribunaux en Ontario.

Le RV s'avère souvent plus approprié dans un contexte judiciaire. Il permet de confronter deux ou plus hypothèses, ce qui se rapproche davantage du contexte dans lequel la poursuite et la défense cherchent à évaluer la valeur probante d'une trace d'ADN. Dans un cas unique analysé par Cohen (2017), le RV a été bien compris par le décideur de fait

qui n'a pas commis d'erreur d'interprétation. Au Canada, en particulier en Ontario et au Québec, le passage de la RMP vers l'utilisation préférentielle du RV commence à se faire notamment grâce à l'outil informatique STRmix^{TM}, un logiciel de génotypage probabiliste pour la résolution des mélanges d'ADN (Centre of Forensic Sciences, 2017). L'avantage de l'utilisation du RV est qu'il permet d'être davantage polyvalent, cohérent et objectif (Centre of Forensic Sciences, 2017; Coquoz *et al.*, 2013, p. 303-413). Il permet de prendre en compte (dans l'argument *I*) d'autres informations spécifiques au cas étudié (p. ex. informations circonstancielles, apparentement, risque de contamination, hétéroplasmie pour l'ADNmt) (Coquoz *et al.*, 2013, p. 303-413).

1.6.2 Marqueurs haploïdes

Le SWGDAM et l'*International Society for Forensic Genetics* (ISFG) ont émis des recommandations quant à l'utilisation de l'ADNmt (Carracedo *et al.*, 2000; Parson *et al.*, 2014; SWGDAM, 2003, 2013; Tully *et al.*, 2001) et du chrY (Gill *et al.*, 2001; Gusmão *et al.*, 2006; SWGDAM, 2009, 2014) en génétique forensique. En ce qui concerne l'interprétation des résultats, les conclusions possibles suite à la comparaison de profils d'ADN sont les mêmes que pour les marqueurs autosomaux, soit une exclusion, une inclusion (ou concordance) ou un résultat non concluant (Butler, 2012, p. 371-403; Melton, 2016). Pour l'ADNmt, des règles de décisions sont proposées afin de trancher entre une inclusion, une exclusion et un résultat non concluant, alors que chaque laboratoire doit établir lui-même celles relatives au chrY (SWGDAM, 2013, 2014). Il est aussi important de noter que dans le cas d'une concordance, la formulation de cette conclusion diffère de celle pour les marqueurs autosomaux et doit plutôt prendre la forme suivante :

> La personne X, et tous les membres appartenant à la même lignée maternelle/paternelle, ne peuvent être exclus comme pouvant être la source de la trace d'ADN.

De la même façon que pour les marqueurs autosomaux, la valeur probante d'une concordance est évaluée en estimant la rareté du profil de la trace dans la population d'intérêt. Toutefois, les **Équations 1.1** à **1.3** ne peuvent s'appliquer aux marqueurs haploïdes. En effet, ces marqueurs ne sont pas transmis à la prochaine génération de manière indépendante. En raison de leur localisation sur un même chromosome ou plasmide et de leur haploïdie, la loi de Hardy-Weinberg ne s'applique pas dans leur cas. Les principales approches proposées pour calculer la probabilité de concordance fortuite et le rapport de vraisemblance sont décrites dans les **Sections 1.6.2.1 et 1.6.2.2**.

1.6.2.1 Probabilité de concordance fortuite

Il existe différentes expressions pour évaluer la RMP dans le cas d'un haplotype, mais qui donnent en général toutes des valeurs du même ordre de grandeur (Parson *et al.*, 2014). La méthode la plus courante consiste à estimer la fréquence de l'haplotype complet dans une base de données pertinente avec la méthode de comptage (en anglais, *counting method*) (**Équation 1.9**) (SWGDAM, 2013, 2014).

$$\hat{p} = \frac{x}{N} \qquad \textbf{(1.9)}$$

Où :
$\hat{p}$: estimateur de la fréquence de l'haplotype dans la population
x : nombre de fois que l'haplotype a été observé dans une base de données
N : nombre d'individus contenus dans cette base de données

Lorsque l'haplotype est rare ou qu'il n'a jamais été observé auparavant, la RMP peut être calculée de deux façons. La première consiste à ajouter une unité à la valeur de x et de N pour simuler l'ajout de l'haplotype de la trace d'ADN comme nouvel échantillon dans la base de données (**Équation 1.10**) (Egeland et Salas, 2008; Parson *et al.*, 2014).

$$\hat{p} = \frac{x+1}{N+1} \qquad \textbf{(1.10)}$$

La deuxième, plus conservatrice, appelée en anglais *pseudo-count* ou *maximum likelihood estimator*, consiste à ajouter deux unités à la valeur de x et de N puisque l'haplotype a été observé dans la trace d'ADN et la personne connue (p. ex. suspect) (**Équation 1.11**) (Balding et Nichols, 1994).

$$\hat{p} = \frac{x+2}{N+2} \tag{1.11}$$

À l'estimation de la fréquence $\hat{p}$ s'ajoute normalement un intervalle de confiance (à 95 %) permettant de prendre en compte l'incertitude dans la fréquence due à l'échantillonnage et de corriger pour la taille de la base de données, ce qui ajoute au conservatisme. Anciennement, la limite supérieure de l'intervalle de confiance était calculée selon une distribution normale (**Équation 1.12**), mais cela s'est avéré inexact lorsque la taille d'échantillonnage était petite ou que l'haplotype était rarement observé, dévaluant ainsi la valeur probante de la trace d'ADN (Butler, 2012, p. 371-403; 2015, p. 403-444).

$$LIMSUP\ IC(95\%) = +1{,}96\sqrt{\frac{(\hat{p})(1-\hat{p})}{N}} \tag{1.12}$$

Où LIMSUP IC représente la limite supérieure de l'intervalle de confiance.

Clopper et Pearson (1934) ont proposé de plutôt calculer la limite supérieure de l'intervalle de confiance en se basant sur une distribution binomiale (**Équation 1.13**). L'Équation **1.14** constitue un cas particulier de l'Équation **1.13** et est utilisée dans les cas où l'haplotype n'a jamais été observé dans la base de données.

$$\sum_{k=0}^{x} \binom{N}{k} p_0^k (1-p_0)^{N-k} = \alpha \quad \textbf{(1.13)}$$

$$p_0 = 1 - \alpha^{1/N} \quad \textbf{(1.14)}$$

Où :
k : 0,1,2,3…x observations
α : niveau de confiance, généralement 0,05 pour un intervalle de confiance à 95 %
p_0 : limite supérieure de l'intervalle de confiance pour laquelle la probabilité cumulative des k observations de l'haplotype égale α

Quoi qu'il en soit, dans un cas où l'haplotype n'a jamais été observé dans la base de données, la fréquence calculée avec les Équations **1.10, 1.11 et 1.14** ne dépend que de la taille de la base de données. Cela implique donc que tous les haplotypes auront la même estimation de fréquence, et ce, peu importe la méthode de calcul utilisée, dans ces cas précis.

L'utilisation des **Équations 1.9** à **1.11** ainsi que **1.12** à **1.14** est critiquée notamment par Roewer *et al.* (2000), Andersen *et al.* (2013) et Krawczak (2001) comme étant trop conservatrice et dévaluant la valeur probante de la trace d'ADN. De plus, elles ne permettent pas de prendre en compte la structure génétique des populations, tel que soulevé par Buckleton *et al.* (2011). À cet effet, le SWGDAM recommande d'utiliser la même approche que pour les marqueurs autosomaux (SWGDAM, 2013, 2014) (voir **Section 1.6.1.1, Équations 1.6 et 1.7**). L'Équation pour les marqueurs haploïdes découle donc de celle pour les marqueurs autosomaux (**Équation 1.15**) :

$$P(A|A) = \theta + (1-\theta)\hat{p}_A \quad \textbf{(1.15)}$$

Où A est l'haplotype d'intérêt et $\hat{p}_A$, sa fréquence estimée, par exemple, avec la méthode de comptage (**Équation 1.9**) à laquelle est ajoutée la limite supérieure d'un intervalle de confiance (**Équation 1.13**).

La valeur de θ doit être calculée empiriquement pour la population d'intérêt et l'ensemble des STR-Y ou la séquence de l'ADNmt utilisé. Le SWGDAM (2013) propose

des valeurs de θ pouvant être utilisées pour le chrY, alors que pour l'ADNmt aucun consensus n'a été établi sur la manière d'estimer ce paramètre.

Des auteurs ont proposé d'autres approches pour estimer la RMP. En 2000, Roewer *et al.* ont développé la méthode du *haplotype surveying* pour estimer la fréquence d'un haplotype pour le chrY à partir des haplotypes observés dans une base de données. Cette méthode fait appel aux distances génétiques entre un haplotype donné et tous les autres observés dans la base de données ainsi que la fréquence de ceux-ci. Pour justifier leur approche, ces auteurs ont montré que la fréquence d'un haplotype donné est corrélée à celles des haplotypes semblables (du point de vue de la distance génétique). Krawczak (2001) a par la suite suggéré des modifications à cette méthode en justifiant que celle-ci n'intégrait pas bien la possibilité que l'haplotype d'intérêt n'ait pas été observé dans la base de données, mais qu'il ait été observé chez une personne connue (p. ex. un suspect). Au final, Brenner (2010) a critiqué ce type d'approche en rappelant que les haplotypes d'intérêt sont souvent très rares et que leur fréquence est plus influencée par la dérive génétique (fluctuation aléatoire des fréquences d'une génération à l'autre) que par les mutations.

Par la suite, Andersen *et al.* (2013) ont proposé la méthode de *Discrete Laplace* pour estimer la fréquence d'un haplotype pour le chrY. Elle prend en compte la distribution des allèles des différents STR dans une population donnée. Plus récemment, Andersen et Balding (2017) ont développé une approche utilisant la généalogie simulée d'une population pour parvenir à estimer le nombre d'hommes portant un haplotype Y particulier plutôt que d'estimer la RMP. Cette méthode sera abordée davantage dans la discussion.

Finalement, les laboratoires rapportent souvent le nombre de fois que l'haplotype de la trace a été observé dans une base de données plutôt que sa fréquence, en particulier pour les haplotypes rares (Holland et Lauc, 2014). Cette méthode est cependant à proscrire, car elle risque d'être interprétée incorrectement comme une valeur probante par le décideur de fait. Plusieurs haplotypes rares ne sont pas observés dans les bases de

données faisant en sorte que le nombre rapporté est de zéro dans de tels cas. Cela pourrait laisser croire que ces haplotypes sont très rares dans la population générale, alors que certains pourraient être fréquents et que c'est seulement le hasard de l'échantillonnage qui a fait en sorte qu'ils ne soient pas représentés dans la base de données. Il est donc difficile de savoir comment bien interpréter une valeur de 0.

1.6.2.2 Rapport de vraisemblance

Tout comme pour les marqueurs autosomaux, les résultats relatifs à l'ADNmt et au chrY peuvent être exprimés sous forme de RV et l'**Équation 1.8** reste applicable pour les marqueurs haploïdes (Butler, 2015, p. 403-444). Ainsi, le résultat se rapporterait en énonçant que l'observation (concordance entre le profil de la trace et celui du suspect X) est *RV* fois plus probable si le suspect X est la source de la trace que si la source de la trace est une personne inconnue (SWGDAM, 2014). Toutefois, il est important de préciser au décideur de fait que dans le cas des marqueurs haploïdes, tous les individus faisant partie d'une même lignée parentale ont le même haplotype (sauf exception).

Dans un cas où l'haplotype n'a jamais été observé dans la base de données, Brenner (2010) a proposé une autre façon d'évaluer le RV, soit le modèle Kappa, qui prend en compte le nombre d'haplotypes vus une seule fois (singletons) (**Équation 1.16**).

$$RV = \frac{N}{1 - \kappa} \qquad \textbf{(1.16)}$$

Où κ représente la proportion de singletons dans une base de données.

Par exemple, si κ est de 0,90 et que la taille de la base de données N est de 1 000 individus, le RV aura une valeur de 1 000/(1-0,90) = 10 000, signifiant que la valeur probante de la trace d'ADN est beaucoup plus grande quand la concordance implique un singleton qu'un haplotype plus fréquent (Butler, 2015, p. 403-444). Cette méthode permet d'accorder un poids plus réaliste à une concordance impliquant des haplotypes jamais observés (Butler, 2015, p. 403-444). Il est important de noter qu'avec cette approche, tous

les haplotypes jamais observés, peu importe leur « identité », mènent à la même valeur de RV. À cet effet, Buckleton *et al.* (2011) mentionnent que cela entraîne une perte d'information sur le lien entre les haplotypes au niveau évolutif (contrairement à la méthode du *haplotype surveying*).

1.6.2.3 Bases de données pour les marqueurs haploïdes

Il existe différentes bases de données permettant d'estimer la fréquence d'un haplotype pour les marqueurs haploïdes. Pour le chrY, la principale utilisée en génétique forensique est celle du Y-STR Haplotype Reference Database (YHRD) accessible en ligne depuis 2000 (Coquoz *et al.*, 2013, p. 145-200; Willuweit et Roewer, 2015). Au moment de la rédaction de ce mémoire, la base YHRD comprenait 255 811 individus génotypés avec au moins 9 STR-Y (ensemble minimal) provenant de 1 221 populations (Willuweit et Roewer, 2018b). Parmi ces individus, 40 070 et 48 028 ont été génotypés avec les trousses Yfiler Plus et PowerPlex Y23 respectivement. Il est à noter que, lors d'une recherche dans cette base de données, les approches utilisées pour exprimer la fréquence de l'haplotype sont : *Discrete Laplace*, $x+1/N+1$ avec un intervalle de confiance et le modèle Kappa. Les États-Unis ont également leur propre base de données depuis 2007, soit la US Y-STR (Butler, 2015, p. 403-444). Lors de la rédaction, elle comprenait 35 660 individus génotypés avec au moins 11 STR-Y (ensemble du SWGDAM) et parmi ceux-ci, les profils de 2 094 et 5 305 individus ont été obtenus avec les trousses Yfiler Plus et PowerPlex Y23 respectivement (s.a. US Y-STR, 2018). Il y a également les bases de données des développeurs de trousses commerciales comme Applied Biosystems (Coquoz *et al.*, 2013, p. 145-200).

Pour l'ADNmt, la base de données principalement utilisée est celle du European DNA Profiling Group mitochondrial DNA population database (EMPOP) accessible en ligne depuis 2006 (Butler, 2015, p. 403-444; Coquoz *et al.*, 2013, p. 145-200; Parson et Dür, 2007). En 2013, l'ajout de nouveaux échantillons a permis d'atteindre 34 617 individus séquencés pour l'ADNmt (s.a. EMPOP mtDNA database, v3/R11). Comme pour la base YHRD, différentes approches sont utilisées pour exprimer la fréquence de

l'haplotype, soit la méthode de comptage avec un intervalle de confiance, $x+1/N+1$, $x+2/N+2$ et la limite supérieure de l'intervalle de Clopper-Pearson lorsque l'haplotype n'a jamais été observé. La base de données du SWGDAM (SWGDAM Mitochondrial DNA Population Database) est également disponible pour ceux utilisant le système CODIS (Butler, 2015, p. 403-444). De plus, il y a celle de l'Institut de médecine légale de Münster, celle du FBI (CODIS^{mt} database) et les bases de données accessibles en ligne www.mitomap.org et http://mtmanager.yonsei.ac.kr/ (Butler, 2012, p. 405-456; Coquoz *et al.*, 2013, p. 145-200).

Une autre ressource existe pour avoir une estimation des fréquences des haplotypes. Les recherches en généalogie combinées avec les informations moléculaires ont permis la création de bases de données pouvant apporter des informations utiles dans certains cas, mais elles ne sont généralement pas utilisées en génétique forensique puisque l'information moléculaire est reliée à des informations nominatives sur les individus ayant fourni leur ADN (Butler, 2015, p. 403-444).

1.7 Bilan des recherches généalogiques et génétiques au Québec

Afin de bien situer la recherche décrite dans ce mémoire, un bilan des études effectuées sur l'ADNmt et le chrY au Québec est présenté dans cette section.

L'étude de la génétique ou de l'histoire d'une population peut se faire selon différentes stratégies comme l'utilisation de données généalogiques ou de données moléculaires provenant d'un échantillon de personnes. Le Québec, plus précisément la population canadienne-française de cette province, a été étudié selon ces deux stratégies depuis les années 1990. Au départ, ceci avait pour principal objectif de mieux comprendre la répartition géographique de maladies héréditaires récessives dont la fréquence est grande dans certaines régions, alors qu'elles peuvent être absentes ailleurs (Bouchard et de Braekeleer, 1991). Scriver (2001) résume l'histoire du peuplement, de la migration et de la démographie du Québec ainsi que les diverses études portant sur la population canadienne-française, surtout du point de vue des maladies mendéliennes.

Premièrement, le Québec possède une connaissance unique de sa généalogie depuis l'époque de sa fondation au 17e siècle jusque dans les années 1960, ce qui en fait un modèle idéal pour étudier la structure et la dynamique génétique de certains gènes. La consanguinité et le degré d'apparentement ont été étudiés dans la région du Saguenay–Lac-Saint-Jean pour mieux comprendre la prévalence de maladies comme l'Alzheimer (Vézina *et al.*, 1999) ou l'hypertension (Pausova *et al.*, 2002). Vézina *et al.* (2004) ont observé que le niveau des liens d'apparentement varie d'une région à l'autre au Québec, avec les plus hautes valeurs dans les régions de l'est, et qu'il y avait des liens d'apparentement entre tous les sujets étudiés ou presque dans les 26 régions, le Saguenay–Lac-Saint-Jean ne se distinguant pas des autres régions à cet égard. Mourali-Chebil et Heyer (2006) ont montré que la consanguinité globale moyenne au Saguenay–Lac-Saint-Jean a augmenté entre 1630 et le début du 20e siècle avec l'accroissement de l'apparentement entre individus, avant de diminuer avec l'augmentation des migrations devenant de plus en plus importantes. Aussi, la taille efficace a augmenté jusque dans les années 1780 pour ensuite diminuer, alors que la population était en croissance, probablement à cause du phénomène de transmission intergénérationnelle de la taille efficace des familles. Plusieurs auteurs se sont intéressés à l'étude de la contribution des ancêtres au pool génique actuel des Québécois. Parmi ceux-ci, Heyer et Tremblay (1995) ont étudié la contribution des fondateurs français dans la région du Saguenay–Lac-Saint-Jean, Bilodeau (2002) s'est intéressée à la population de l'Abitibi-Témiscamingue, alors que Bergeron (2005) a étudié la contribution des Acadiens au pool génique québécois. Tremblay et Vézina (2000, 2010) ont analysé les lignées maternelles et paternelles dans la généalogie du Québec pour étudier le nombre, l'origine et la contribution génétique des fondateurs ainsi que les intervalles de temps générationnels. Gagnon et Heyer (2001) et Bherer *et al.* (2011) ont étudié la structure génétique de la population du Québec, et leurs études seront abordées davantage dans la discussion. Moreau *et al.* (2011a) ont étudié la dynamique d'expansion de la population du Saguenay-Charlevoix depuis sa fondation et ont déterminé qu'il y avait un avantage évolutif (en termes de descendance) à se retrouver dans le front d'expansion ou à proximité. Finalement, Milot *et al.* (2011) ont montré que la sélection naturelle devançait l'âge à la première reproduction chez les femmes vivant à

l'Île-aux-Coudres à la période préindustrielle, appuyant l'idée qu'il est toujours possible de détecter des traces d'évolution chez l'humain.

Ensuite, quelques recherches ont porté sur l'ADNmt et/ou le chrY au Québec sans utilisation de données généalogiques, ou du moins, aucune donnée étendue au-delà de petites généalogies familiales. Moreau *et al.* (2007) ont étudié la diversité génétique du Québec à l'ADNmt et les chr X et Y en comparaison avec des populations européennes. Laberge *et al.* (2005) ont pu identifier l'existence d'une fondatrice commune aux Canadiens-français porteurs d'une mutation de l'ADNmt causant la neuropathie optique héréditaire de Leber.

Puis, quelques recherches ont utilisé des données généalogiques et moléculaires séparément pour comparer les résultats obtenus selon chaque type de données. Moreau *et al.* (2011b) ont étudié différentes populations de la Gaspésie selon l'ADNmt et la généalogie, ce qui a permis de comprendre que les lignées maternelles amérindiennes dans cette région sont originaires d'Acadie et non du Québec comme le laissait croire l'ADNmt seul. Moreau *et al.* (2013) ont également étudié le degré de métissage avec les Amérindiens dans différentes populations québécoises selon des marqueurs autosomaux et la généalogie. Avec les mêmes types de données, Roy-Gagnon *et al.* (2011) ont également montré la présence d'une structure de population et d'apparentement chez les Canadiens-français issus de sept régions québécoises. Ainsi, il en ressort que les données moléculaires analysées simultanément à des données généalogiques permettent d'avoir un portrait plus complet de l'histoire d'une population.

Très peu d'études combinent toutefois les données généalogiques et moléculaires dans un même modèle d'analyse. Heyer (1997) a estimé le taux de mutations de 9 STR-Y et de la région de contrôle (HVI et HVII) de l'ADNmt (2001) au Québec. Milot *et al.* (2017) ont étudié la dynamique populationnelle de la mutation causant la neuropathie optique héréditaire de Leber, rapporté précédemment, chez les Canadiens-français. En jumelant information moléculaire et généalogique, ils ont pu mesurer précisément sa fréquence dans la population entre 1670 et 1960. Helgason *et al.* (2001; 2003; 2005) et

Kong *et al.* (2018) ont aussi combiné l'étude des généalogies et celle de l'ADNmt et du chrY pour étudier la population d'Islande du point de vue de la contribution génétique des ancêtres, des intervalles générationnels et de la structure de population. Aussi, Larmuseau *et al.* (2012b) ont étudié la population de Flandre en Europe grâce aux généalogies, au chrY et aux patronymes qui se sont avérés d'intérêt pour étudier des événements passés comme la migration ou le métissage. Leurs recherches ont montré une certaine stabilité du taux de paternité extra-conjugale durant les 400 dernières années (Larmuseau *et al.*, 2013). Ils ont également observé une variation temporelle dans la distribution spatiale des haplotypes Y dans la région de Brabant (Larmuseau *et al.*, 2012a).

La population canadienne-française du Québec a donc été étudiée par sa généalogie avec des marqueurs autosomaux et haploïdes. Cependant, aucune étude n'a porté sur la variation spatio-temporelle des fréquences des haplotypes pour l'ADNmt et le chrY à l'échelle de toute la population. Grâce à la connaissance fine de la généalogie, il est possible d'étudier la structure et la dynamique génétique à fine échelle pour ces marqueurs.

1.8 Objectifs de la recherche

Les marqueurs haploïdes situés sur l'ADNmt et le chrY sont très utiles dans des cas particuliers en génétique forensique pour lesquels les marqueurs traditionnels sont moins performants. Lorsqu'une concordance entre deux profils doit être utilisée comme élément de preuve devant les tribunaux, il faut idéalement accompagner l'énoncé de cette concordance avec sa valeur probante. Celle-ci repose sur l'estimation de la fréquence du profil de la trace dans la population d'intérêt. Dans la **Section 1.6.2**, les difficultés reliées à l'estimation de la valeur probante pour ce type de marqueurs ont été décrites. Il est donc nécessaire de mieux comprendre la dynamique de l'ADNmt et du chrY dans la population afin d'avoir des estimations plus fiables de la fréquence des haplotypes.

Le premier objectif de ce mémoire était d'étudier la variation spatio-temporelle des fréquences des haplotypes à fine échelle dans la population canadienne-française du

Québec. Pour ce faire, des données généalogiques étendues sur la population étudiée et des données moléculaires d'individus connectés à cette généalogie ont été combinées puis analysées. Le principe de ce modèle généalogico-moléculaire est décrit au Chapitre II de ce mémoire.

Deux prémisses sont faites, implicitement ou explicitement, par les laboratoires judiciaires lorsqu'ils recherchent un profil dans des bases de données de référence pour estimer la fréquence d'un haplotype (Butler, 2015, p. 403-444; Szabolcsi *et al.*, 2015). Premièrement, ils supposent que les fréquences des haplotypes sont stables dans le temps, c'est-à-dire qu'elles n'ont pas changé de manière significative depuis la collecte des échantillons constituant la base de données. Ensuite, ils supposent que les différents haplotypes sont répartis de manière homogène sur le territoire de la population d'intérêt et donc que les fréquences sont les mêmes d'une sous-population à une autre. Il est reconnu que les marqueurs haploïdes sont plus sensibles à la dérive génétique et aux effets liés à la fondation d'une population en raison de leur haploïdie et de l'absence de recombinaison, ce qui accentue les fluctuations des fréquences dans le temps et dans l'espace (Templeton, 2006). Le deuxième objectif de ce mémoire était donc de tester ces prémisses à partir des connaissances générées par le modèle généalogico-moléculaire mentionné ci-haut.

Finalement, le troisième objectif était de quantifier l'impact forensique du non-respect éventuel de ces prémisses. Pour ce faire, la probabilité de concordance fortuite a été calculée pour tous les haplotypes et les valeurs obtenues ont été comparées entre différents pas de temps ainsi qu'entre différentes régions et localités du Québec. Une comparaison a également été effectuée avec les valeurs obtenues à partir de bases de données internationales.

CHAPITRE II

VARIATION SPATIOTEMPORELLE DES FRÉQUENCES HAPLOTYPIQUES POUR L'ADN MITOCHONDRIAL ET LE CHROMOSOME Y DANS UNE POPULATION CANADIENNE-FRANÇAISE ET IMPACT SUR LES PROBABILITÉS DE CONCORDANCE FORTUITE

Alexandra Doyon[a], Claudia Moreau[b], Damian Labuda[b,c] et Emmanuel Milot[a]

[a] Département de chimie, biochimie et physique, Université du Québec à Trois-Rivières

[b] Centre de recherche du Centre hospitalier universitaire Sainte-Justine

[c] Département de pédiatrie, Université de Montréal

Le contenu de ce chapitre est en préparation en vue d'être soumis à une revue scientifique avec révision par les pairs.

2.1 Contribution des auteurs

En tant qu'auteure principale, j'ai veillé à la mise en place des méthodes utilisées, produit l'ensemble des résultats présentés et effectué la rédaction de l'article. L'ensemble des données moléculaires provient de Claudia Moreau et Damian Labuda du Centre hospitalier universitaire Sainte-Justine. Ceux-ci avaient préalablement combiné les données moléculaires et généalogiques pour l'ADN mitochondrial et ces données ont été utilisées pour en faire l'analyse. Ils ont également contribué au projet par l'apport d'idées et de recommandations. Dr Emmanuel Milot agit comme directeur de recherche et a mis sur pied ce projet de recherche. Il a participé à la conception des méthodes utilisées et à l'interprétation des résultats.

2.2 Résumé de l'article

L'ADN mitochondrial et le chromosome Y sont couramment utilisés dans plusieurs domaines de la génétique comme l'épidémiologie, l'étude de la démographie, l'étude sur l'évolution de traits d'histoire de vie, etc. En génétique forensique, ces marqueurs se sont montrés particulièrement utiles dans l'analyse de l'ADN dégradé, dans le cas de l'ADN mitochondrial, et des mélanges d'ADN homme/femme, dans le cas du chromosome Y. Lorsque deux profils génétiques (haplotypes) concordent, par exemple celui retrouvé sur une trace d'ADN et celui d'une personne connue, il faut généralement estimer la fréquence de l'haplotype de la trace dans la population d'intérêt. Plusieurs approches ont été proposées pour ce type de marqueurs, mais aucune d'elles n'est reconnue ni appliquée de manière consensuelle. Face à ces difficultés d'interprétation, une meilleure connaissance de la dynamique de ces marqueurs s'avère importante afin d'avoir des estimations fiables des fréquences dans une population. Pour ce faire, nous avons étudié la variation spatio-temporelle des fréquences des haplotypes dans la population canadienne-française du Québec en développant un modèle combinant des données généalogiques et moléculaires. Cette population a une connaissance étendue de sa généalogie depuis sa fondation au 17^{e} siècle jusqu'en 1960, ce qui en fait un modèle idéal. Lorsque la concordance entre deux profils doit être présentée devant un tribunal, il est souvent nécessaire de calculer la valeur probante de cette concordance. À cet égard, la plupart des laboratoires judiciaires font implicitement ou explicitement deux prémisses, soit que les fréquences des haplotypes dans la population sont stables dans le temps et qu'elles sont homogènes dans l'espace. Le modèle développé pour ce projet nous a permis de tester ces deux prémisses. Enfin, nous voulions quantifier l'impact du non-respect éventuel des prémisses sur le calcul de la probabilité de concordance fortuite. Nos résultats montrent que la diversité en haplotypes a été relativement stable dans le temps et qu'il existe une structure génétique dans la population étudiée. Cette structuration a eu un impact non négligeable sur le calcul des probabilités de concordance fortuite et remet en question l'utilisation de bases de données nationales ou internationales à des fins de comparaison. Finalement, le modèle développé pour ce projet pourrait également être utile en épidémiologie génétique, en biologie évolutive et aussi, pour soutenir l'identification de restes humains.

2.3 Article

Spatiotemporal variation of mitochondrial DNA and Y chromosome haplotype frequencies in a French-Canadian population and its impact on random match probabilities

Abstract

Mitochondrial DNA and the Y chromosome are commonly studied in many fields of genetics, including forensic genetics where these lineage markers have proved useful in the analysis of degraded DNA, mixtures of female/male DNA and familial searching. In the presence of a match between two haplotypes, an estimate of this haplotype's frequency must generally be determined to weigh the evidence. However, no approach available to do so is fully satisfactory. Typically, it is presumed implicitly or explicitly that haplotype frequencies in the population of interest are stable in time and that they are homogeneous in space. In order to interpret correctly haploid DNA matches, it is important to have a sufficient knowledge of the spatiotemporal dynamics of these markers. Thus, we studied haplotype frequencies in the French-Canadian population of Québec for which the full genealogy of married people is known since the foundation of the population in the early 17^{th} century up to 1960. Mitochondrial DNA sequences (HVI and HVII regions) were obtained for 970 individuals and for the Y chromosome, 275 men were genotyped with at least 17 Y-STRs. We developed a probabilistic model to impute haplotypes of genotyped individuals across the genealogy which accounted for genealogical errors and mutations. We calculated genetic diversity in diverse cohorts as well as the random match probability and our results indicate that haplotypic diversity in the French-Canadian population was relatively stable in time. However, the comparison of genetic diversity among regions and localities uncovered important spatial variations in haplotype frequencies, which had a non-negligible impact on the calculation of random match probabilities, questioning the use of national or international reference databases for forensic comparative purposes.

Keywords

Lineage markers, genealogy, French-Canadian population, genetic diversity, population structure, random match probability

Introduction

Mitochondrial DNA (mtDNA), transmitted from mother to child, and Y chromosome (Ychr), transmitted from father to son, are commonly used in studies of historical demography and evolutionary biology (Aimé *et al.*, 2015; Moreau *et al.*, 2011; Wilson Sayres *et al.*, 2014), epidemiology (Ambulkar *et al.*, 2015; Schaefer *et al.*, 2004) and human identification (Butler, 2010, p. 363-396; Larmuseau *et al.*, 2016a). Since the beginning of the 1990s, the importance of these lineage markers in forensic genetics has been increasingly recognized (Gusmão *et al.*, 2006; Holland and Parsons, 1999). MtDNA is used when nuclear DNA is degraded as with human remains and Ychr is useful in the presence of female/male DNA mixtures, which often occurs on intimate swabs in sexual assault cases (Coquoz *et al.*, 2013, p. 145-200). Both types of markers can also be used in familial searching, namely to find close or distant relatives of a person, to test for biological relatedness and to identify missing people (Anjos *et al.*, 2004; Kayser, 2017).

Lineage markers are not subject to recombination and are not considered independent as for autosomal markers, so a mtDNA or Ychr haplotype is passed unchanged, barring mutations (Cockerton *et al.*, 2012). Thus, everyone in a maternal (mtDNA) or paternal (Ychr) lineage generally share a common haplotype, except when analyzing rapidly mutating (RM) Y-STRs in which cases, related men can have different haplotypes (Ballantyne *et al.*, 2014). As a consequence, the Hardy-Weinberg model cannot be used and the estimation of haplotype frequencies will be solely based on observations of the complete haplotype (Butler, 2015a, p. 403-444; Templeton, 2006).

A common approach to estimate haplotype rarity is the counting method, expressed as x/N, where x is the number of times a haplotype is observed in a database and N the size of the database, with a confidence interval accounting for database size and sampling

variation (SWGDAM, 2013, 2014). Haplotype frequencies can also be estimated as $(x+1)/(N+1)$ or $(x+2)/(N+2)$ (Parson *et al.*, 2014), but these can overestimate the true frequency especially for rare haplotypes (Egeland and Salas, 2008). Other methods have also been proposed (Andersen *et al.*, 2013; Brenner, 2010; Buckleton *et al.*, 2011; Roewer, 2009; Roewer *et al.*, 2000; Tully *et al.*, 2001), but ultimately the common practice is to report only the number of times a haplotype occurs in a database or the upper bound of the confidence interval (Budowle *et al.*, 2003; Holland and Lauc, 2014).

All above methods aim to deal with sampling issues in what is presumed to be a homogeneous randomly mating population. However, none of them address the question of weight-of-evidence satisfactorily, as recently shown by Andersen and Balding (2017). When evaluating weight-of-evidence for lineage markers, two implicit assumptions are that 1) haplotype frequencies in the population have been stable since the time when the reference sample was collected and 2) haplotypes are homogeneously distributed in space (Butler, 2015a, p. 403-444; Szabolcsi *et al.*, 2015), so that the population targeted by the reference sample database matches the population of interest in particular caseworks. However, mtDNA and Ychr are susceptible to not comply with above assumptions due to 1) the genealogical structure of the population of interest resulting in a spatiotemporal variation of haplotype frequencies, 2) the great diversity of haplotypes, and 3) the lack of representativeness of databases with regard to the population of interest due to the challenge to clearly define this population. These issues are discussed in turn below.

First, due to their lower effective population size and the absence of recombination, lineage markers are more affected by genetic drift, gene flow and founder effects than autosomal markers, which can cause haplotypes to be less homogeneously dispersed in space, as well as variation in frequencies over time (Cockerton *et al.*, 2012; Kayser and Ballantyne, 2014; Templeton, 2006). Physical, cultural or social barriers also contribute to spatial structuring (see for example Heyer *et al.* 2015 and Zerjal *et al.* 2001). Indeed, relatives are often geographically clustered, which increases haplotype frequencies locally (Roewer, 2009). The International Society for Forensic Genetics (ISFG) and the Scientific Working Group on DNA Analysis Methods (SWGDAM) recommend taking into account

population structure in the estimation of haplotype frequencies (Gusmão *et al.*, 2006; Parson *et al.*, 2014; SWGDAM, 2013, 2014). In order to do so, the SWGDAM provides θ value estimates for the Ychr for some ethnic groups, but indicates that no consensus exists yet over the method to be used to estimate θ values for the mtDNA owing to variation in the mtDNA segment targeted by different primer sets (SWGDAM, 2013,2014). Moreover, a single F_{ST} metric used to correct for population structure is unlikely to be the appropriate model to account for the spatial variation in frequencies, hence it may be insufficient to adjust frequencies estimated from large-scale databases (e.g., whole country) to deal with local populations of interest (Gómez-Carballa *et al.*, 2016; Toscanini *et al.*, 2012). Therefore, it is not recommendable to apply frequency data from a sample for a large population without considering population genetic aspects of lineage markers especially when dealing with rare haplotypes.

Second, mtDNA and Ychr haplotypes are more diversified than any independent STR marker (Kayser and Ballantyne, 2014). Butler *et al.* (2007) showed that 95 % of the 17 Y-STR haplotypes were found only once in a sample of 656 males from the United States. Similarly, in a sample of 175 men typed at 20 Y-STRs for this study, 92 % of the haplotypes had a frequency of less than 2/175. Parsons and Coble (2001) determined that ~84 % of mtDNA haplotypes were found once in the United States Caucasian population in the MitoSearch database (http://www.mitosearch.org/). As a consequence, every haplotype uncovered in a forensic casework but never observed in a reference database will have the same estimated frequency with the counting method or methods derived from it, whether or not they correct for database size and regardless of the context of the case (Kaestle et al., 2006). The problem should become even more important with the increasing number of Y-STRs in commercial kits, which correlates positively with haplotypic diversity (Kayser, 2017).

Third, due to the two difficulties aforementioned, databases for lineage markers are usually not quite representative of the population of interest. Their size is generally too small to provide a reliable frequency estimate for rare haplotypes (Holland and Lauc, 2014), and it often happens that a haplotype found at a crime scene is rare in the population

of interest. Solutions proposed to evaluate the evidentiary value of these never observed haplotype have their drawbacks (Andersen and Balding, 2017). Another issue pertaining to database validity is the method used to create them. They are often built with DNA profiles from laboratory staff, volunteers, blood banks, law-enforcement officers, caseworks or offenders databases, none of them being random samples (Buckleton, 2005; Cockerton *et al.*, 2012; National Research Council, 1996, p. 125-165). In the case of autosomal markers, recombination limits such sampling biases, which is obviously impossible for lineage markers. Finally, Buckleton (2005) point out that there is no clear methodological guidelines about how to validate a database. One ideal solution is to know the complete genealogy of the population of interest as illustrated in this study.

Another important factor impacting the representativeness of databases is the concept of population of interest. One of the most crucial issues is to calculate haplotype frequencies using data from the appropriate population as underscored by the SWGDAM and ISFG (Gusmão *et al.*, 2006; Parson *et al.*, 2014; SWGDAM, 2013, 2014). However, the "population of interest" is an ill-defined concept (Buckleton, 2005), making the interpretation of a match complex (Andersen and Balding, 2017). Even the concept of biological population lacks a common definition among researchers. This could be due to criteria that are generally more qualitative than quantitative and depend on the context of the study (Waples and Gaggiotti, 2006). Some authors define the population of interest as the pool of individuals susceptible to have left their DNA on a scene, regardless of their geographical origin, determined by the context of the casework and the information collected by investigators (Coquoz *et al.*, 2013; Parson *et al.*, 2014, p. 303-413). According to Buckleton (2005), one can hardly be sure that a database is representative of the population of interest because "we do not always know what group we are trying to represent." It is usually easier to define the appropriate reference database for genetically homogeneous populations. For admixed populations, the ISFG proposes to calculate match probabilities from multiple subpopulations or ethnic groups (Parson *et al.*, 2014). However, this only makes the result more (too?) conservative while not addressing properly (if at all) the population of interest issue, and notwithstanding potential biases in

the self-declaration of origin by people providing DNA samples (Buckleton, 2005; Kaestle *et al.*, 2006).

In this study we used an exceptional dataset to document the spatiotemporal variation of lineage markers in a large population. We were able to compare frequencies and random match probabilities obtained from typical samples or international databases to much more precise spatially fine-tuned measurements. We studied the French-Canadian population of Québec, whose genealogy is known from its foundation in the 17th century up to 1960, making it an ideal model to study the genetic structure and dynamic of lineage markers at a fine scale. Genetic structure in Québec had previously been studied from the genealogy or genetic markers from modern individuals in separate analyses (Bhérer *et al.*, 2011; Gagnon and Heyer, 2001; Moreau *et al.*, 2007; Roy-Gagnon *et al.*, 2011). Here we developed a model to link these two types of data, impute molecular information to individuals in the genealogy and thereby obtain a much higher coverage of the population from a standard molecular sample.

Our results indicate a temporal stability in genetic diversity for both mtDNA and Ychr. Random match probabilities were also not greatly different among cohorts for most haplotypes. Comparison of the genetic diversity among regions or localities showed a non-negligible spatial variation in mtDNA and Ychr haplotype frequencies, impacting on the estimation of the random match probability. In addition, international databases such as EMPOP and YHRD tended to underestimate haplotype frequencies and hence, overestimate the weight-of-evidence compared to values obtained in our study population.

Materials and methods

Study population

The studied population is composed of French-Canadian individuals, who represent ~80 % of Québec population (Gagnon and Heyer, 2001). Overall, 1,065 individuals (486 men and 579 women) from various regions were recruited to create a reference sample for studies in genetics (s.a. Échantillon de référence québécois, 2010). From this sample, we

used a subset composed of 442 men and 533 women (n=975 individuals) for whom we had either information on both HV regions or on at least 7 Y-STRs. The participants gave their consent to provide genealogical information and either blood or saliva as a source of DNA. Genealogical information consisted in date and place of birth, marriage and death of their ancestors over the last three generations to connect them to the BALSAC population register (see below). DNA samples were coded and analyzed at the Centre hospitalier universitaire Sainte-Justine (CHUSJ) Research Center and the Laboratoire de sciences judiciaires et de médecine légale (Québec Ministry of Public Security) in Montréal. This research was approved by the CHUSJ and Université du Québec à Trois-Rivières ethic committees.

Molecular data

DNA samples from the 975 individuals had been previously sequenced for the hypervariable regions (HV) I (positions 16,069-16,383) and II (positions 58-370) of the mtDNA (see Heyer *et al.* 2001, Moreau *et al.* 2011, Moreau *et al.* 2007, Moreau *et al.* 2009 and Vézina *et al.* 2012 for details). The total 628 bp segment of both HV regions combined was successfully sequenced for 970 individuals from the following regions (**Figure 2.1**): Abitibi (n=18), Beauce (n=52), Côte-Nord (n=78), Gaspésie (n=393), Lanaudière (n=29), Montréal (n=159), Outaouais (n=15), Québec city area (n=52) and Saguenay–Lac-Saint-Jean (n=174). Haplotypes were defined with respect to the revised Cambridge Reference Sequence (rCRS) (Andrews *et al.*, 1999). A haplogroup is defined as a group of haplotypes sharing a most recent common ancestor and specific SNPs (Torroni *et al.*, 2006). Haplogroups vary in their biochemical properties and their geographic distribution is thought to reflect adaptation of human populations to local environmental conditions (Wallace, 2015). The haplogroups to which belong mtDNA haplotypes observed in this study were determined using HV sequences and SNPs on positions 7,028, 14,766 and 12,308 (Moreau *et al.*, 2009).

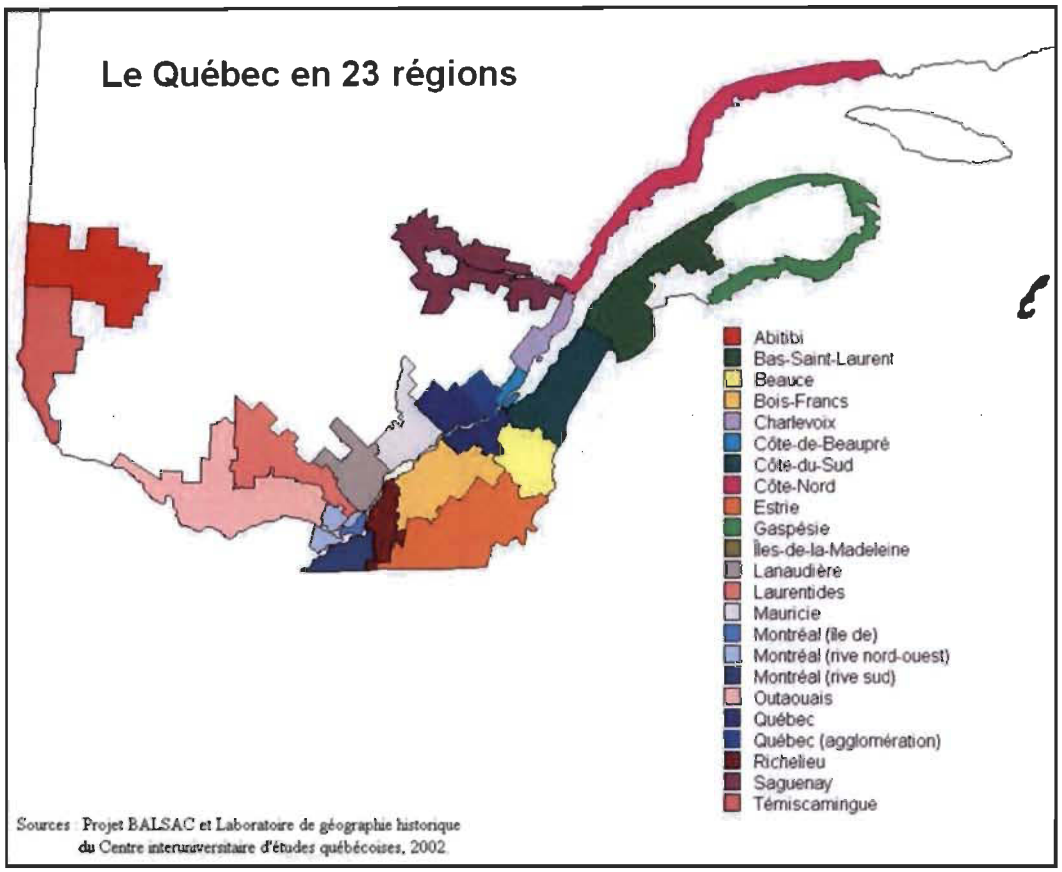

Figure 2.1. Southern Québec subdivided into 23 BALSAC Register regions. These regions were defined based on geography and peopling history (Reprinted from BALSAC - Fichier de population).

The Ychr of the 442 men had been previously genotyped using different Y-STR kits for other projects (Moreau *et al.*, 2007; Moreau *et al.*, 2009; Vézina *et al.*, 2012). For this study, we used a sample of 100 men typed with the 17 Y-STRs of the AmpFℓSTR® Yfiler™ kit from Life Technologies (DYS389I, DYS389II, DYS19, DYS391, DYS390, DYS392, DYS393, DYS438, DYS437, DYS385a/b, DYS439, DYS635, DYS458, Y GATA H4, DYS448 and DYS456). These men were from the following regions: Abitibi (n=7), Beauce (n=19), Côte-Nord (n=14), Lanaudière (n=3), Montréal (n=32), Outaouais (n=5), Québec (n=20) (**Figure 2.1**). To this sample, we also added 175 men, from the Gaspésie region, typed with the 27 Y-STRs of the Yfiler® Plus kit from Life Technologies (17 markers of the AmpFℓSTR® Yfiler™ kit + DYS576, DYS627, DYS460, DYS518, DYS570, DYS449, DYS481, DYF387S1 and DYS533). These two sets of markers are routinely used for forensic purposes (Ballantyne *et al.*, 2010; Kayser *et al.*, 2004; White

et al., 1999). For all analyses presented here, we used 20 of the 27 Y-STRs of the Yfiler® Plus kit. To facilitate haplotype probabilistic imputation (see below), we excluded seven rapidly mutating (RM) Y-STRs included in this kit, at the cost of losing some molecular information. Since Y-STRs are linked on the same haploid chromosome, the set of markers typed for an individual defined his Y chromosome haplotype. The haplogroups to which belong Ychr haplotypes could not be determined from the Y-STRs markers analyzed in this study.

Molecular data was used for two purposes. First, we estimated genealogical error rates based on the premise that two typed individuals ascending to a same lineage founder through the maternal or paternal line should have the same haplotype (barring mutations) if their pedigree is correctly reconstructed. For the mtDNA, five of the 970 individuals were excluded from this analysis because there was an ambiguity at one of their nucleotides, which could possibly be attributed to heteroplasmy. For the Ychr, we used the sample of 275 men described above, to which we added 154 men that were typed for previous projects with either 7 Y-STRs (DYS389I, DYS389II, DYS19, DYS391, DYS390, DYS392 and DYS393) (de Knijff *et al.*, 1997; Kayser *et al.*, 1997), 12 Y-STRs of the PowerPlex® Y from Promega (the 7 former markers + DYS438, DYS437, DYS385a/b and DYS439) (Ayub *et al.*, 2000; Schneider *et al.*, 1998), or 17 Y-STRs of the AmpFℓSTR® Yfiler™ kit (these were not included in the main sample because they had some missing alleles). These 154 men were from the following regions: Beauce (n=12), Côte-Nord (n=17), Gaspésie (n=1), Lanaudière (n=4), Montréal (n=24), Québec (n=4) and Saguenay–Lac-Saint-Jean (n=92) (**Figure 2.1**). In total, 429 men were thus included to estimate genealogical error rates even though they were not all genotyped with the same number of Y-STRs, because even partial profiles can provide information about genealogical errors (**Table 2.S2**). Hence, we used all the molecular information available. Second, we imputed haplotypes along the BALSAC genealogy from modern-day genotyped individuals. For the mtDNA, both HV regions of our 970 individuals were included in this analysis. For the Ychr, we imputed 20 Y-STR haplotypes using 175 men typed at 27 Y-STRs (again excluding RM Y-STRs). We also performed the imputation using 275 men typed at 17 Y-STRs either with the Yfiler® or Yfiler® Plus kit. We used

these two sets in parallel for comparative purposes. Hence, we had more genotyped men at 17 than 20 markers, allowing us to impute a haplotype to 137,627 more individuals in the genealogy, but at the expanse of losing three markers.

Genealogical data

Genealogical data was obtained from the BALSAC population register (BALSAC - Fichier de population) built from Catholic parish and civil marriage registers spanning the period from 1621 up to 1965. New France was founded in 1608, so BALSAC contains information on married people almost since population foundation. The Research Program in Historical Demography (Université de Montréal) participated in the enrichment of BALSAC by providing marriage acts prior to 1800 and contained in the Early Québec Population Register (Desjardins, 1998; Dillon *et al.*, 2018).

Two types of errors can occur when linking acts to reconstruct genealogies: erroneous genealogical links (over-pairing) or missing links between related individuals (under-pairing). Common explanations include partial or erroneous information on marriage acts (e.g., transcription errors by the officiant of the marriage), homonymy, and the use of marriage records from non-Catholic parishes or from outside Québec. When detected, errors are corrected in the BALSAC register. Moreover, the social genealogy, reconstructed from acts, may differ from the "biological" genealogy known from DNA. Parents mentioned in an act are not necessarily the biological parents of a child due to extra-pair paternity or (unreported) adoption.

For this study, we used data from the BALSAC register as of 2015, which includes 3,371,740 individuals married between 1621 and 1960 with a male to female ratio of 0.95:1. The maximal genealogical depth is 19 generations. In this genealogy, maternal or paternal "lineages" composed of < 3 individuals typically correspond to unlinked marriage acts, since for each spouse typically two individuals of a lineage appear on the marriage act (him- or herself and his/her father or mother), except in cases of remarriage addressed below. These floating acts can be unlinked to other marriage acts for various reasons. First, a couple may have no children or they may have emigrated/immigrated

after their marriage and there is therefore no trace of their ascendants or descendants. Second, the parents of one or both spouses can be unknown either because they are not mentioned on the act, their marriage act could not be found, their marriage took place outside Québec, or because they were non-Catholic. Third, some acts could be incompletely filled by the officiant or destroyed, the former being more common. Fourth, in cases of remarriage, the former spouse is often mentioned instead of the parents who may remain unknown. Therefore, only lineages composed of at least three individuals along maternal or paternal lines (e.g., corresponding to > 1 marriage act) were kept.

Estimation of genealogical error rates

First, we estimated the overall genealogical error rate (ε). We obtained two separate estimates of ε, one from mtDNA data and the other from Ychr data. We expected the second to be higher due to extra-pair paternities (EPP). For the Ychr, we applied the iterative method of Larmuseau *et al.* (2013) developed to assess EPP rates from Ychr data. Briefly, we identified 263 pairs of genotyped men sharing a common ancestor and separated by at least seven meioses. Since meioses were partly shared among dyads, we randomly selected between 62 and 65 independent dyads (i.e. sharing no meiosis) and the procedure was repeated 1,000 times. For each of these samples, the raw mismatch rate was estimated as the proportion of dyads showing at least one molecular difference between the two individuals. Mismatches between the two haplotypes involving more than 18 % differences were considered as due to genealogical errors (either from EPP or an error in the linkage of marriage acts) (Larmuseau *et al.*, 2013). Following Larmuseau *et al.* (2013), we assumed that a difference at fewer than 18 % of Y-STRs compared resulted from mutations and not a genealogical error. Thus, for each sample of independent dyads, a total of 10,000 iterations were done to estimate ε by maximum likelihood (Larmuseau *et al.*, 2013) such that $\mathbf{E} \sim Binomial(n, \varepsilon)$, where **E** is a vector indicating whether each putative genealogical link represented or not an error and n, the number of meiosis separating individuals in a dyad, as well as a 95 % confidence interval. We then calculated average values obtained with all samples. For this analysis, we used all available information, including individuals with smaller numbers (<17) of Y-STRs typed, each time using only markers available for both individuals of the pair compared. Therefore,

43 pairs were compared at 7-11 Y-STRs, 124 at 12 Y-STRs, and 96 at 13-20 Y-STRs. For maternal lineages, we identified 2,879 pairs of genotyped individuals sharing a common ancestor and separated by at least seven meioses. As for the Ychr, we randomly selected between 225 and 230 independent dyads to estimate ε. Also, we assumed that mismatches in nucleotide sequences involving more than one difference were due to genealogical errors (Parsons *et al.*, 1997). Finally, for both types of markers, we also estimated mismatch rates due to mutations + genealogical errors, hereafter referred to as global mismatch rates, to assess the importance of the former compared to the latter. The same procedure was used, but this time by accepting no difference between haplotypes in a dyad.

Combining genealogical and molecular data to impute haplotypes

We combined genealogical and molecular data to develop a 'genealogico-molecular model' (GM) (see for example Larmuseau *et al.* 2012a and Milot *et al.* 2017). The molecular information of genotyped individuals was imputed to individuals connected to the ascending genealogy since all individuals in the same maternal lineage share a common mtDNA haplotype and likewise for individuals in a given paternal lineage who share a common Ychr haplotype (barring mutations and errors, see below; **Figure 2.2**). Considering the availability of a (nearly) complete genealogy for our study population, GM models were used to overcome the limited sizes typical of population reference samples and to obtain more accurate estimations of haplotype frequencies. After reaching the lineage founders, we then imputed haplotypes to other individuals along the genealogy descending from these founders (**Figure 2.2**). Two approaches can be used to assign a haplotype across the genealogy and are described in the next sections.

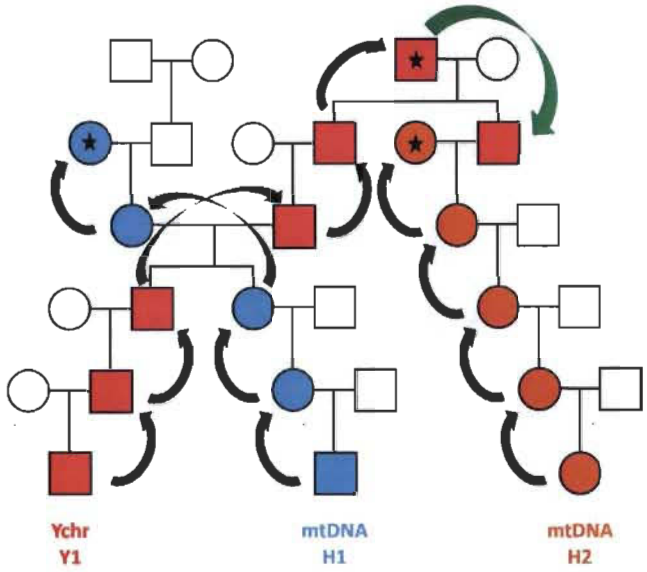

Figure 2.2. Schematic view of a 'genealogico-molecular model'. Men are represented by squares and women by circles. Genotyped individuals are at the bottom along with their mtDNA or Ychr haplotype. Founders are indicated with a star. Arrows indicate the imputation procedure along the ascending (black) and descending (green) genealogy.

Non-probabilistic approach

Under this approach, the haplotype observed in a genotyped individual is simply imputed to all other individuals belonging to the same lineage. It is necessary to verify that only one haplotype is imputed to each founder of a lineage when more than one genotyped individual belong to this lineage, otherwise there is a mismatch. This could be due to mutations, genealogical errors, or genotyping errors. Typically, it is impossible to determine precisely the genealogical link where the error or mutation responsible for the mismatch occurred. A solution can be to impute more than one haplotype to the individuals concerned. The complete lineage could also be excluded as done by Larmuseau *et al.* (2012a; 2012b; 2015). However, errors cannot be fully excluded even in lineages with no mismatches, although they should be less likely. The non-probabilistic approach often results in the removal of lineages from subsequent analyses, hence a loss of information. MtDNA haplotypes were first imputed across the BALSAC genealogy

using a non-probabilistic approach[2] (as in Milot *et al.* 2017). Individuals for whom more than one haplotype was imputed were considered as having no haplotype.

Probabilistic approach

We developed a probabilistic approach for the imputation of Ychr haplotypes taking into account mutation and genealogical error rates in an iterative process (**Figure 2.3**). In each simulation, genealogical errors, leading to the removal of a given genealogical link, were generated randomly based on estimated genealogical error rates (ε). We considered each meiosis in the genealogy as a Bernoulli trial with two possible outcomes: "success" (a genealogical error occurred, with probability ε) and "fail" (no genealogical error occurred, with probability 1-ε). Thus, we generated a vector **E** such that $\mathbf{E} \sim Binomial(n, \varepsilon)$ as before, where n is the number of meiosis (i.e. the number of individuals in the lineage minus 1). The link between an individual and his father was removed for elements of **E** equal to 1. Then, mutations were similarly generated on the remaining links by generating a vector $\mathbf{M} \sim Binomial(n, \bar{\mu})$, where $\bar{\mu}$ is the average mutation rate of Y-STR markers. Mutation rates (μ) for Ychr were obtained from the YHRD website (Willuweit and Roewer, 2018a) and from Ballantyne *et al.* (2010) for the DYS385a/b marker.

[2] For mtDNA, I used the non-probabilistic approach for the purpose of my dissertation due to developing and computer time constraints. Imputation with the probabilistic approach will be done in future work.

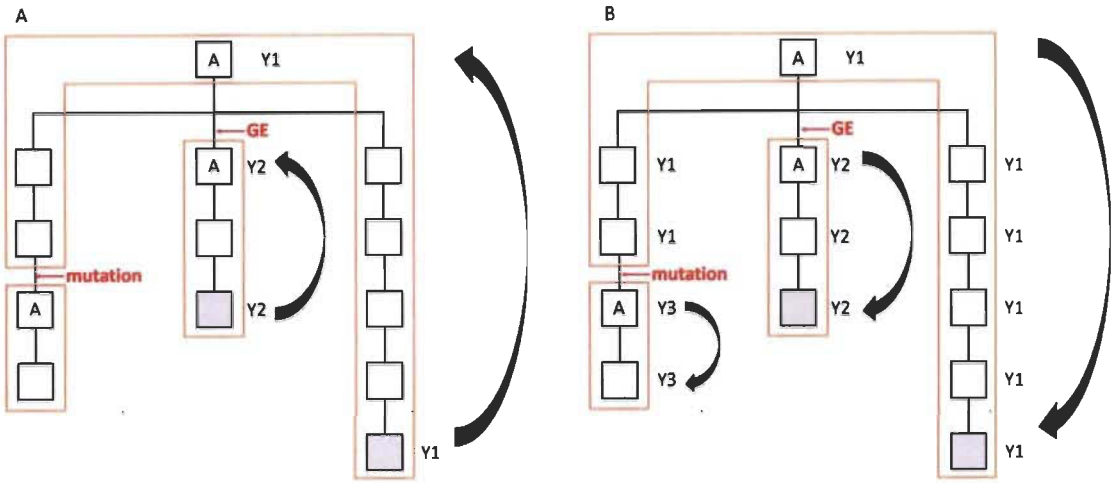

Figure 2.3. Schematic view of the probabilistic model developed for Ychr haplotypes imputation across the genealogy. The pedigree is simplified to show only father-son links (grey squares: genotyped men; open squares: untyped men). Haplotypes are indicated by the letter Y and a number. Panel A: genealogical errors (GE) and mutations are generated randomly in the pedigree, thereby defining sub-pedigrees (outlined in orange). Then, haplotypes of genotyped individuals are imputed to ancestors (A) of sub-pedigrees. Panel B: ancestors' haplotypes are imputed to all individuals of the same sub-pedigree. In case of a mutation, the haplotype of the individual above the link is modified according to a mutation model and this new haplotype is imputed to individuals under him.

Simulated genealogical errors and mutations defined sub-pedigrees (**Figure 2.3**). The ancestor of a sub-pedigree is either a lineage founder or the individual below the link where a mutation or a genealogical error occurred. Haplotypes were imputed from genotyped individuals to sub-pedigree ancestors (**Figure 2.3A**). Next, haplotypes were imputed descending the genealogy starting from sub-pedigree ancestors (**Figure 2.3B**). When several haplotypes had been imputed to the same ancestor, one of them was randomly chosen according to haplotype frequencies in genotyped individuals belonging to the same sub-pedigree. In case of mutations, the haplotype of the sub-pedigree ancestor's father was mutated according to a mutation model and this new haplotype was imputed to the sub-pedigree ancestor and his descendants (**Figure 2.3B**). To mutate haplotypes at one or more markers, we used published Y-STR mutation rates, including gain and loss of STR repeats (Ballantyne *et al.*, 2010; Goedbloed *et al.*, 2009, Decker *et al.*, 2008). The proportion of gain/loss involving one and two repetitive units were respectively 98.8 % and 1.2 %.

The whole procedure was repeated multiple times and we kept only simulations for which the haplotypes imputed to the genotyped individuals were identical to those observed, discarding other simulations. We attempted to have at least 1,000 simulations kept (i.e. haplotype imputations) for each individual in the genealogy, although this was not always possible due to pedigree structure and the distribution of molecular information across individuals. Imputations served to assess the probabilities for each haplotype for each individual in the genealogy, by dividing the number of times an individual had a given haplotype by the number of simulations kept. Haplotypes present in less than one hundredth of individuals in a lineage were grouped together in a category named "others". This was necessary because most of these haplotypes were the result of simulated mutations that appeared once or a few times only (e.g., more than a million different Y-STR haplotypes were generated in simulations). We imputed 20 Y-STR haplotypes as well as 17 Y-STR haplotypes across the BALSAC genealogy using this probabilistic model.

A fictive example of simulation results is provided in the supplementary material and **Table 2.S1**. For each individual i, we calculated his probability of having haplotype h, $\Pr_i(h)$, with $1 = \sum_h \Pr_i(h)$. We did not impute the haplotype with the highest probability to an individual i, which would have resulted in keeping one haplotype per man, because we wanted to keep the information on all haplotypes imputed to an individual, to propagate the uncertainty of imputation to subsequent analyses.

Molecular data coverage by periods and regions

We divided the population history (1621-1960) into 32-year periods, which corresponds to the average generation time for maternal and paternal lineages in this population (Tremblay and Vézina, 2000), to create 11 marriage cohorts (the last cohort covering only 20 years due to data ending in 1960). For each cohort, we calculated the proportion of individuals for whom a haplotype was imputed. We did a spatial analysis for the most recent cohort (1941-1960), for which population size and molecular data coverage were the greatest. This cohort was also the closest to the present-day population (i.e. ~2 generations apart with several individuals still alive). Only individuals for whom

the parish of marriage was known and located in Québec were kept for this analysis. We calculated the molecular coverage in 23 regions (**Figure 2.1**) as well as that in the remaining territory not covered by these regions. We did among-region/locality comparisons for regions/localities meeting the following criteria. For mtDNA, the minimal coverage, namely the proportion of married individuals typed or with an imputed haplotype, had to be 50 % for regions and 60 % for localities. For Ychr, the minimal coverage (men only) for regions and localities were set respectively to 10 and 25 % for 20 Y-STR profiles and 20 and 35 % for 17 Y-STR profiles. Moreover, as some regions or localities had small population sizes, we only compared those with at least 40 individuals.

We also did among-period comparisons by creating three cohorts of 20 years spanning from 1901 to 1960 because an INTERPOL survey (2009) showed that 54 member countries created their own national DNA database between 1995 and 2008. Hence, most databases have between 10 and 23 years old. Moreover, population size and molecular coverage were the greatest in these years. In each of these cohorts, we calculated the molecular coverage in the same regions as above and kept for analyses those with the highest coverage for all three cohorts (excluding the well-covered Îles-de-la-Madeleine region because it is an isolated archipelago with a small population size). Thus, for the mtDNA, the best-covered region was Charlevoix with a coverage ranging from 79.6 to 83.3 % and for the Ychr, it was that of the Gaspésie with a coverage ranging from 29.1 to 31.5% (20 Y-STRs) or from 34.5 to 36.2% (17 Y-STRs).

Haplotype frequency, genetic diversity and random match probability

To estimate the frequency of a haplotype in the entire population, we had to take into account lineages that are not covered by mtDNA or Ychr data. Thus, we assumed that every untyped lineage had a different haplotype. This assumption is more realistic with Ychr than mtDNA data because ~25 % (73/291) of mtDNA haplotypes were detected in more than one lineage (43 in 2 lineages, 17 in 3 lineages, 8 in 4 lineages, 4 in 5 lineages, and 1 in 37 lineages), while ~1.4-3.0 % of Y-STR haplotypes were in the same situation (20 Y-STRs set: 2/148 haplotypes in 2 lineages; 17 Y-STRs set: 5/231 haplotypes present in 2 lineages and 2 in 3 lineages). Moreover, most lineages had an intra-lineage diversity

of zero (mtDNA: 98.5 % of lineages; 20 Y-STRs: 85.8 %; 17 Y-STRs: 91.2 %), meaning that most of the time, only one haplotype was observed in a lineage (we imputed the haplotype with the highest probability for each individual to estimate the intra-lineage diversity). Later we discuss the potential consequences of this assumption (see *Discussion*). For a given cohort, the frequency of haplotype h in the whole population/region/locality was estimated as $p_h = \frac{1}{N}\sum_i \Pr_i(h)$, where N is the population size. Thus, the estimated number of carriers (n_h) of h in the whole population or a given region/locality was $n_h = Np_h$, and could thus be <1 since $\sum_i \Pr_i(h)$ can be <1 (**Table 2.S1**).

Genetic diversity ($\hat{H}$), namely the probability of randomly drawing two different haplotypes from the population, was calculated based on Nei (1987) for each 32-year cohorts and regions/localities selected according to our coverage criteria:

$$\hat{H} = \frac{N}{N-1}\left(1 - \sum\nolimits_h p_h^2\right)$$

where N is the number of individuals and p_h, the frequency of the h^{th} haplotype in the sample.

The random match probability (RMP) was calculated for each 20-year cohorts in the Charlevoix (mtDNA) or Gaspésie (Ychr) region as well as for the regions/localities kept for analyses based on our coverage criteria. It was initially calculated based on Parson *et al.* (2014) with $(n_h+1)/(N+1)$ (notation was changed to correspond with ours). However, adding one in the numerator would sometimes highly overestimate the RMP of rare haplotypes in a specific cohort, region, or locality, especially for the Ychr. This is because adding 1 would drastically inflate the calculation when $n_h < 1$. Thus, instead of 1 we added the minimum value of n_h over all haplotypes, which can be approximately viewed as the minimal number (or fraction) of copies of any given haplotype in the population considered (a cohort, region or locality). The RMP was therefore calculated as $(n_h+n_{h\min})/(N+n_{hmin})$. A worked example is shown in **Table 2.S1**. We compared RMP

values obtained in the 1941-1960 cohort ('reference' cohort) to those obtained in either two preceding 20-year cohorts by calculating pairwise log-of-odds scores as $LOD = \log_{10}(RMP_{cohort}) - \log_{10}(RMP_{reference})$, where RMP_{cohort} stands for RMP values calculated from either the 1901-1920 or the 1921-1940 cohort. This allowed us to keep information on which cohort had the highest value. RMP values obtained for the reference cohort were also compared with those obtained by pooling cohorts, which can be seen as a reference sample with profiles cumulated over the period 1901-1960. For the comparison of RMPs among regions or localities, we calculated LOD scores as $LOD = |\log_{10}(RMP_i) - \log_{10}(RMP_j)|$, where *i* and *j* indices stand respectively for the first and second RMP value for the pair of regions/localities compared. Here, we took the absolute value to facilitate comparisons, since the order of the two RMP values (i.e. *i* or *j*) in a score is irrelevant. RMPs obtained in regions or localities were also compared with those calculated from the pool of regions, which can be considered as a reference sample in which one would search a given profile to estimate its frequency. For all comparisons, LOD scores were calculated for all haplotypes, except for the "others" category, mostly composed of haplotypes created through simulations at very small frequencies. Moreover, we compared RMP values obtained at the regional scale using our data with those estimated by searching the same haplotypes in international databases, as suggested by the SWGDAM (2013, 2014). For the latter analysis, we limited our comparisons to the five most frequent haplotypes (excepting virtual haplotypes imputed to untyped lineages) in Charlevoix, one of the best-covered region for mtDNA and Ychr. For the mtDNA, comparisons were done with the EMPOP database (s.a. EMPOP mtDNA database, v3/R11), for which the sequence range was fixed to 16024-16356 and 73-340, and using default values for other parameters. To increase the chance of observing our haplotypes in the reference database, we chose the worldwide and the European populations, the latter being genetically closest to our study population (Charbonneau *et al.*, 2000). For the Ychr, comparisons of 17 Y-STR haplotypes were done with the YHRD database (Willuweit and Roewer, 2018b). The worldwide, Eurasian-Caucasian and France populations were chosen for the same reasons, the latter being the source of French-Canadian founders (Charbonneau *et al.*, 2000). All analyses were performed in R v.3.3.2 (2016) using the RStudio v.1.0.136 (2015) interface (R scripts available upon request).

Results

Molecular data

A lineage was defined as a group of at least three connected individuals (i.e. representing at least one connection between two marriage acts). A typed lineage is a lineage for which a haplotype could be imputed by connecting a contemporary typed individual to his/her maternal and/or paternal line. Of the 970 individuals typed at mtDNA, 951 were connected to a total of 407 lineages, but haplotypes could be successfully imputed from only 932 individuals covering 404 lineages (i.e. three lineages were considered as untyped because of haplotype mismatches). These individuals carried a total of 291 different haplotypes. The number of individuals per lineage ranged from three to 51,121 (see below for more details; **Table 2.S3**). A total of 145 lineages were linked to more than one typed individual, ranging from two (n=64 lineages) to 34 individuals (n=1 lineage; **Figure 2.S1**). Of these 145 lineages, 31 included individuals carrying different haplotypes, with 24, five and two lineages associated with two, three and four haplotypes, respectively (**Figure 2.S2**). Even with the lineage comprising 34 typed individuals and having a total of 49,399 individuals, only four different haplotypes were observed (31 individuals carrying the same haplotype and three with different haplotypes) with a number of nucleotide changes necessary to pass from one to another ranging from two to seven. In addition, 73 haplotypes were observed in more than one lineage, ranging from two (n=43 haplotypes) to 37 lineages (n=1 haplotype; **Figure 2.S3**). We could also impute a haplotype to 24 unconnected individuals from 14 of the 970 typed individuals. These carried 14 different haplotypes of which nine were present in one to 37 lineages, through other typed individuals.

Of the 175 men typed at 20 Y-STRs, 169 were connected to a total of 127 paternal lineages and carried a total of 148 different haplotypes. The number of individuals per lineage ranged from three to 10,651 (see below for more details; **Table 2.S4**). Twenty-seven of the 127 lineages were linked to more than one typed individual, ranging from two (n=19 lineages) to seven individuals (n=1 lineage; **Figure 2.S4**). Within these 27 lineages, 18 included individuals with different haplotypes (13 and five lineages with two

and three haplotypes, respectively; **Figure 2.S5**). Only two of the 148 haplotypes were observed in two lineages. Similarly, with the set of 275 men typed at 17 Y-STRs, 269 of them were distributed in 215 lineages and carried 231 different haplotypes. The number of individuals per lineage ranged from three to 14,258 (**Table 2.S5**). Thirty-four of these 215 lineages included more than one typed individual, ranging from two (n=22 lineages) to seven (n=1 lineage; **Figure 2.S6**). We observed 19 lineages in which individuals had different haplotypes (13 and six lineages with two and three haplotypes, respectively; **Figure 2.S7**). Seven haplotypes were present in more than one lineage (five in two lineages and two in three). Finally, for both sets of markers, we could also impute a haplotype to 12 unconnected individuals from six of the 175 (20 Y-STRs) or 275 (17 Y-STRs) typed men. These carried six different haplotypes, only one of which present in a single lineage through other typed individuals.

Genealogical error rates for maternal and paternal lineages

To apply our probabilistic model, we first calculated the genetic distance between each pair of typed individuals. A total of 2,958 and 274 pairs shared a common ancestor through maternal and paternal lineages, respectively, and the genetic distances varied from one to 28 meioses for maternal lineages and from four to 23 for paternal lineages (**Figure 2.S8**). As explained in Methods, only pairs with a common ancestor and a genetic distance of at least seven meioses were kept. The number of pairs meeting the above criteria was 2,879 for maternal lineages and 263 for paternal lineages. The number of differences between haplotypes in a pair of individuals ranged from zero (n=2,365) to 14 differences (n=8) for maternal lineages and from zero (n=131) to 17 differences (n=1) for paternal lineages.

Global mismatch rates were estimated to 0.009 (95 % CI: 0.007–0.013) and 0.052 (95 % CI: 0.039–0.069) per generation for mtDNA and Ychr respectively. These were more than twice higher than mismatch rates due to genealogical errors only, for both types of markers (mtDNA: 0.004 [95 % CI: 0.002–0.006]; Ychr : 0.008 [95 % CI: 0.003–0.016]), which suggests that mutations alone caused more mismatches than genealogical errors. Also, mismatch rates were approximately two to six times higher for Ychr than

mtDNA (depending on the type). Finally, assuming that mismatch rates due to genealogical errors are the same for maternal and paternal lineages led to an estimated average rate of extra-pair paternity of 0.4 % (95 % CI: 0.1–1 %) for the French-Canadian population of Québec across the last centuries.

Molecular data coverage

A total of 269 mtDNA haplotypes were imputed to 404 of 50,196 maternal lineages, representing a total of 1,634,944 individuals and 54.1 % of the whole population married between 1621 and 1960. Individuals included in a lineage represented almost 100 % of individuals with imputed haplotypes (only 24 of the latter were not associated with a lineage) and 83.9 % of individuals without imputed haplotypes (**Figure 2.4**). The number of individuals per typed lineage ranged from three to 51,121 and from three to 16,270 for untyped lineages. A majority (70.0 %) of typed lineages had more than 100 individuals and they represented 99.8 % of all individuals with imputed haplotypes (**Table 2.S3**). In contrast, only 1.6 % of untyped lineages had more than 100 individuals, representing 71.5 % of individuals without imputed haplotypes, with 76.4% of these lineages having fewer than ten individuals (**Table 2.S3**).

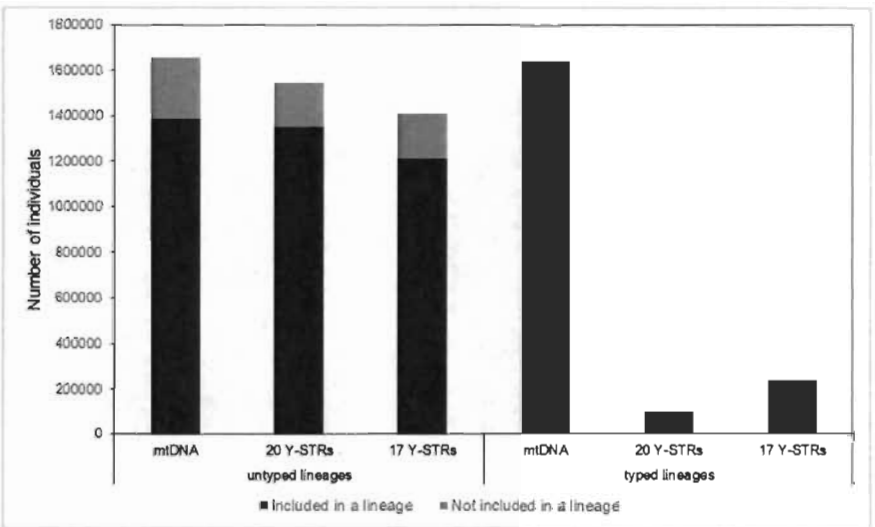

Figure 2.4. Number of typed and untyped individuals based on their inclusion in a maternal (mtDNA) or paternal (Ychr) lineage. Respectively 24 and 12 individuals with imputed mtDNA and Y-STR haplotypes were not connected to a lineage and their proportion is too small to be visible on the figure.

We imputed 20 Y-STR haplotypes using 169 of the 175 typed men who could be connected to a paternal lineage, and similarly, we used 269 of the 275 typed men to impute 17 Y-STR haplotypes. This gave a coverage of 127 lineages with 20 Y-STRs and 215 with 17 Y-STRs, out of 31,819 paternal lineages, representing respectively 6.7 % (97,414 men) and 16.2 % (235,041 men) of the whole male population married between 1621 and 1960. Individuals included in a lineage represented almost 100 % of individuals with imputed haplotypes for both sets of Y-STRs (only 12 men with imputed haplotypes were not connected to a lineage in both cases), and 87.3 % (20 Y-STRs) or 86.1 % (17 Y-STRs) of individuals without imputed haplotypes (**Figure 2.4**). The number of individuals per lineage typed at 20 Y-STRs ranged from three to 10,651 and from three to 14,258 for untyped lineages (**Table 2.S4**). With 17 Y-STRs, the number of individuals per typed lineage varied from three to 14,258 and from three to 7,099 for untyped lineages (**Table 2.S5**). A proportion of 44.1 % (20 Y-STRs) or 64.7 % (17 Y-STRs) of typed lineages comprised more than 100 individuals and they represented 98.6 % (20 Y-STRs) or 99.4 % (17 Y-STRs) of all individuals with imputed haplotypes (**Tables 2.S4** and **2.S5**). In contrast, only 6.1 % (20 Y-STRs) or 5.8 % (17 Y-STRs) of untyped lineages had more than 100 individuals, representing 82.5 % (20 Y-STRs) or 80.5 % (17 Y-STRs) of individuals without imputed haplotypes, with 78.3 % (20 Y-STRs) or 78.6 % (17 Y-STRs) of these lineages having fewer than ten individuals (**Tables 2.S4** and **2.S5**).

The imputation with 20 Y-STRs resulted in a total of 1,661,451 different haplotypes including the 148 true haplotypes observed in typed individuals. For subsequent analyses, we kept these 148 observed haplotypes as well as 897 virtual haplotypes present in at least a hundredth of individuals in a lineage, the remaining haplotypes being pooled in the category "others". These numbers were slightly different with 17 Y-STRs. A total of 1,857 haplotypes were kept for analyses (1,626 virtual haplotypes and 231 observed haplotypes) and 1,712,703 haplotypes were pooled together.

Molecular coverage by periods, regions and localities

We analyzed the molecular coverage (i.e. percent of individuals in the married population with an imputed haplotype) for each 32-year cohort (**Table 2.1**). The coverage

was generally higher for mtDNA (25.2-54.9 %) than for the Ychr (2.7-17.5 %). Also, the larger number of individuals genotyped at 17 Y-STRs explains the twofold difference in coverage compared to the set of 20 Y-STRs. We also analyzed the molecular coverage for the 24 regions defined by the BALSAC register in the last cohort (1941-1960; **Table 2.S6**). It varied from 35.59 % (Outaouais) to 81.00 % (Îles-de-la-Madeleine) for the mtDNA. For the Ychr, it varied from 3.11 % (Montréal (rive sud)) to 36.44 % (Îles-de-la-Madeleine) with 20 Y-STRs and from 8.24 % (Outaouais) to 38.14 % (Îles-de-la-Madeleine) with 17 Y-STRs. The molecular coverage was also assessed for 1,188 localities dispersed across Québec (**Table 2.S7**). The per-locality coverage varied widely among the 24 regions. For the mtDNA, it was lowest in the Montréal region (range=10.00-56.09 %) and highest in the Rest of Québec region (49.01-100.00 %). For the Ychr, the Montréal (rive nord-ouest) and Outaouais regions were the least covered with 20 Y-STRs (0.00-7.32 %) and Gaspésie, the best covered (0.00-61.49 %). With 17 Y-STRs, the lowest coverage was found in the Montréal (rive sud) region (0.00-16.67 %) and the highest, in the Îles-de-la-Madeleine region (33.33-66.67 %). Finally, a total of 15 regions and 507 localities met our criteria for molecular coverage with mtDNA data, representing respectively 426,843 and 303,043 individuals for the 1941-1960 period. For the 20 Y-STR dataset, five regions (48,210 men) and 52 localities (8,774 men) were kept. With 17 Y-STRs, 11 regions (127,594 men) and 108 localities (21,330 men) were kept.

Table 2.1. Molecular coverage for mitochondrial DNA and Y chromosome for each 32-year cohort.

Cohort	Mitochondrial DNA[a]		Y chromosome[b]		
	Total number of individuals	Coverage (%)	Total number of men	Coverage 20 Y-STRs (%)	Coverage 17 Y-STRs (%)
1621-1652	1,036	25.2	1,170	2.7	7.7
1653-1684	2,403	35.1	2,356	2.7	6.4
1685-1716	8,130	42.7	4,877	3.9	9.2
1717-1748	20,422	49.1	11,070	4.7	10.7
1749-1780	43,639	52.6	21,472	5.2	12.5
1781-1812	87,787	54.9	42,534	5.8	14.0
1813-1844	196,457	54.6	94,928	6.1	14.3
1845-1876	371,586	54.5	177,297	6.5	15.3
1877-1908	561,980	54.8	266,672	6.6	16.1
1909-1940	785,822	53.4	374,670	6.7	16.5
1941-1960	944,690	54.2	451,460	7.2	17.5

[a] mtDNA haplotype imputation done from 970 typed individuals.
[b] Y chromosome haplotype imputation done from 175 (20 Y-STRs) and 275 (17 Y-STRs) typed men.

Genetic diversity in time and space

Genetic diversity was first analyzed for each 32-year cohort to evaluate the temporal variation in mtDNA and Ychr haplotype frequencies (**Figure 2.5**). Genetic diversity was relatively stable in time for both haploid markers and fluctuated between 0.9933 and 0.9988 for the mtDNA, 0.9987 and >0.9999 with 20 Y-STRs and 0.9986 and >0.9999 with 17 Y-STRs. Genetic diversity was nearly identical with both sets of Y-STRs. Moreover, excluding untyped lineages from the analysis of the mtDNA led to a genetic diversity of ~1 % smaller only (not shown).

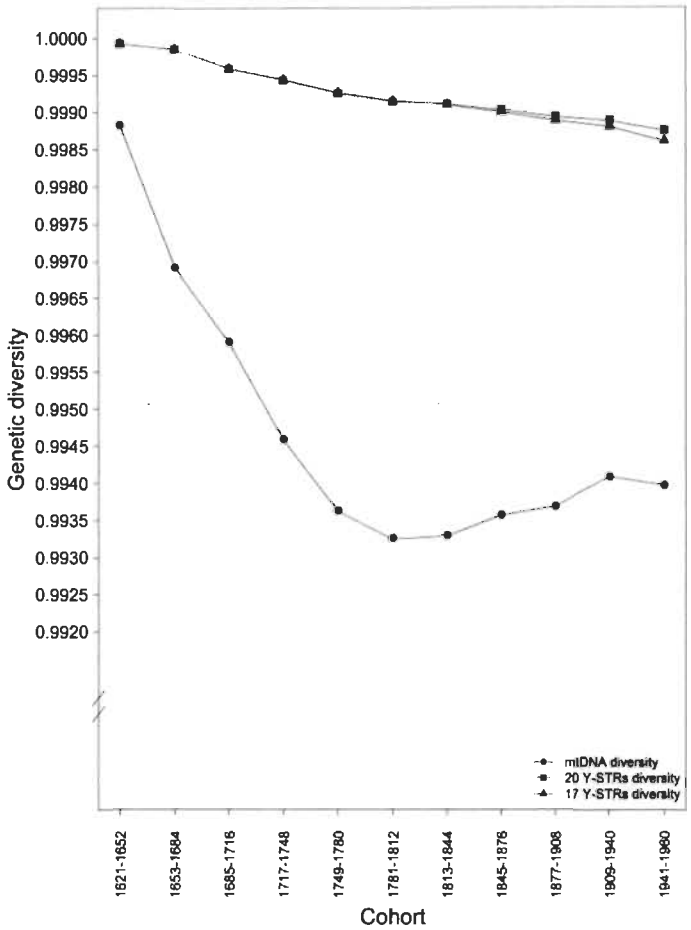

Figure 2.5. Genetic diversity at mitochondrial DNA and Y chromosome in the French-Canadian population of Québec, between 1621 and 1960. Black lines report Nei's (1987) index of diversity. The calculation is based on all married individuals in the population (only men for Ychr), accounting for untyped lineages (assuming a different haplotype for each untyped lineage). Note the scale of the Y-axis implies fluctuations of at most 0.006 between the most and least diverse cohorts.

Based on selection criteria described in Methods, mtDNA diversity was also calculated for 15 regions and Ychr diversity for five (20 Y-STRs) or 11 regions (17 Y-STRs). MtDNA diversity ranged from 0.910 (Îles-de-la-Madeleine) to 0.994 (Estrie) and was 0.991 for all 15 regions pooled together (**Figure 2.6A**). Ychr diversity varied from 0.980 (Îles-de-la-Madeleine) to 0.996 (Gaspésie) with 20 Y-STRs, and from 0.979 (Îles-de-la-Madeleine) to 0.997 (Québec (agglomération)) with 17 Y-STRs, and it was ~0.996 for all regions pooled together in both cases (**Figure 2.6B**), with regions analyzed with both Y-STR sets showing similar diversities. For five regions analyzed with both mtDNA

and Ychr data, diversities were slightly higher with the latter. Finally, two regions, Charlevoix and Îles-de-la-Madeleine, were the least diversified at both types of markers. Interestingly, these regions are among the least peopled but the best covered (coverage ~80 % for mtDNA and ~35-38 % for 17 Y-STRs).

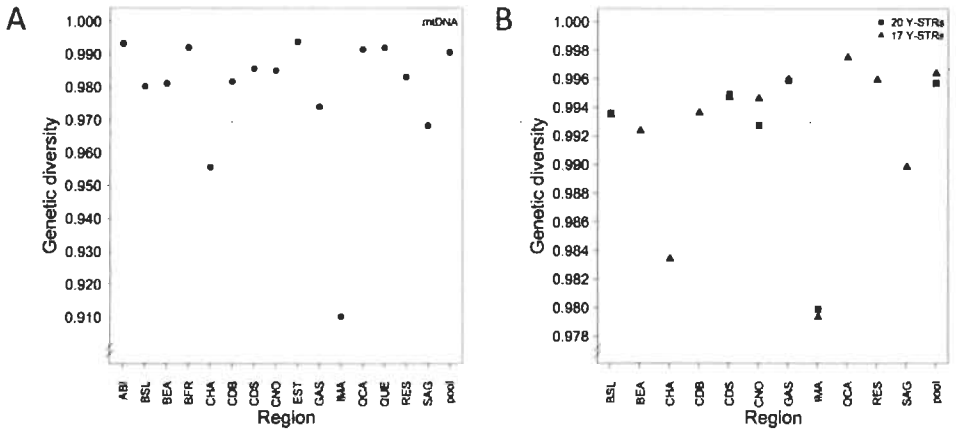

Figure 2.6. Genetic diversity in Québec regions between 1941 and 1960 for the mitochondrial DNA (A) and the Y chromosome (B). Black dots (mtDNA), squares (20 Y-STRs) and triangles (17 Y-STRs) report Nei's (1987) index of diversity. BEA: Beauce, BSL: Bas-Saint-Laurent, CDB: Côte-de-Beaupré, CDS: Côte-du-Sud, CHA: Charlevoix, CNO: Côte-Nord, GAS: Gaspésie, IMA: Îles-de-la-Madeleine, pool : all regions grouped together, RES: Rest of Québec, SAG: Saguenay–Lac-Saint-Jean. Note that the Y-axis scale differs in the two panels.

Genetic diversity was then calculated for 507, 52 and 108 localities respectively for mtDNA, 20 Y-STRs and 17 Y-STRs datasets, again in accordance with our coverage criteria. MtDNA diversity ranged from 0.747 for a locality in the Côte-Nord region to 0.991 for two localities in the Québec (agglomération) and Abitibi regions (**Figure 2.7A**), compared to the value of 0.991 for all regions pooled. Among the five localities with the lowest diversities, one is in Charlevoix and one in Îles-de-la-Madeleine, namely the two regions showing the lowest mtDNA diversity (**Figure 2.6A**). Ychr diversity ranged from 0.946 (20 Y-STRs) or 0.928 (17 Y-STRs) to 0.995 for both marker sets (**Figure 2.7B**), compared to the value of 0.996 for the pool of regions. Among the five localities with the lowest diversities with 20 Y-STRs, three are again in Îles-de-la-Madeleine. With 17 Y-STRs, localities with the lowest diversities are from Îles-de-la-Madeleine, Charlevoix,

Saguenay–Lac-Saint-Jean, Gaspésie and Beauce. Finally, localities analyzed with both Y-STR kits had similar diversities and these were slightly higher than for mtDNA (**Figure 2.7**).

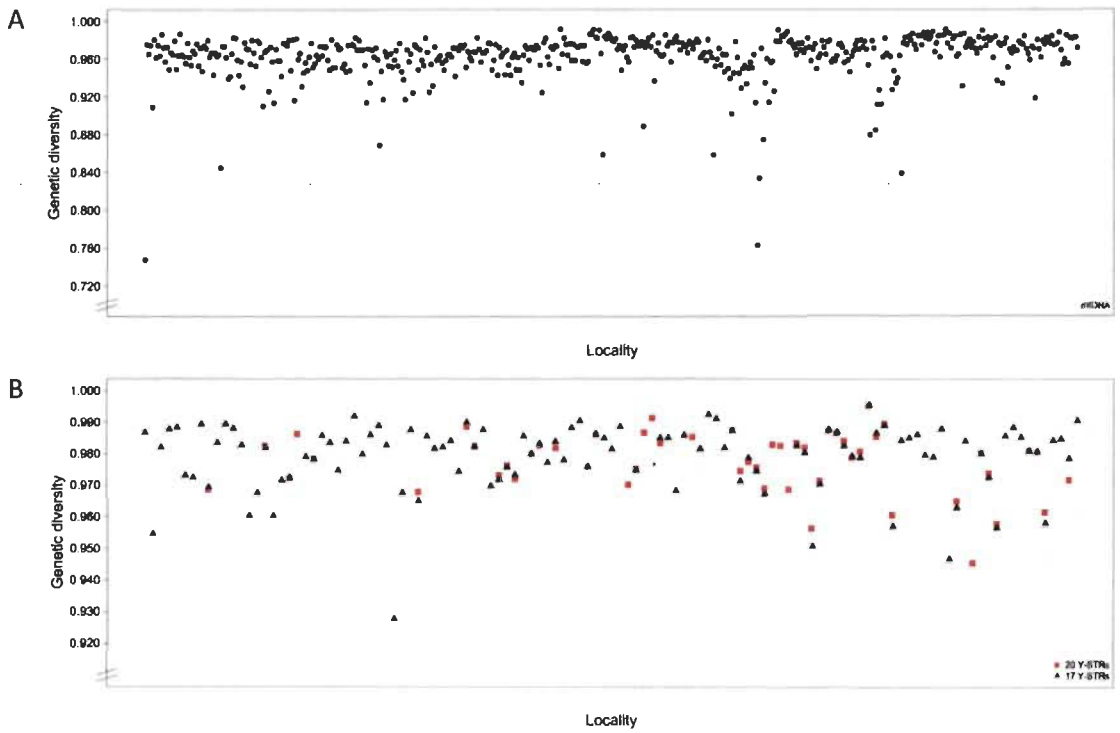

Figure 2.7. Genetic diversity in Québec localities between 1941 and 1960 for the mitochondrial DNA (A) and the Y chromosome (B). Black dots (mtDNA), red squares (20 Y-STRs) and black triangles (17 Y-STRs) report Nei's (1987) index of diversity. Note that the Y-axis scale differs in the two panels. The identity of localities is given in **Table 2.S7**.

Random match probability in time and space

We compared the RMP for the 1941-1960 cohort (hereafter called the 'reference' cohort) to that for either two preceding 20-year cohorts (hereafter called 'alternative' cohorts), by measuring pairwise log-of-odds scores (LOD) in the Charlevoix region for the mtDNA. Distribution of haplotypic LOD scores were similar for both comparisons and LOD ranged from -1.25 to 0.75 (**Figure 2.8A**). The lower bound corresponds to an odd ratio of 0.06 (calculated as 10^{LOD}), meaning that the RMP was 0.06 times smaller in an alternative cohort than in the reference, while a value of 0.75 indicates that the RMP

was 5.6 times higher in an alternative cohort than in the reference. Most LOD scores (~70 %) laid between -0.50 and zero (odd ratio ≈1), indicating that RMPs were essentially identical for the two cohorts compared. The distribution was also similar when comparing the reference to the three cohorts pooled, though differences were slightly more negative in this case (**Figure 2.8A**), suggesting that cumulating profiles over 60 years without removing older ones would have a limited impact on RMP values in this population. The same pattern was observed with mean LOD score (**Figure 2.8D**). All pairwise comparisons resulted in mean values ≈0.3, corresponding to an odd ratio ≈2.

For the Ychr, RMP values were calculated for the Gaspésie region. LOD scores were distributed similarly for both sets of markers ranging from -5.50 to 4.50 with 20 Y-STRs (**Figure 2.8B**) and from -6.00 to 3.00 with 17 Y-STRs (**Figure 2.8C**). Hence, RMP values could be up to 10^{-6} times smaller in an alternative cohort compared to the reference or up to ~30,000 times higher in the former than in the latter. However, even if the variation was greater than with the mtDNA, LOD scores were again mostly grouped around zero, i.e. an odd ratio of 1 (20 Y-STRs: ~70 % of LOD scores were between -0.50 and 0.50; 17 Y-STRs: ~75 % of them were between -1.00 and 1.00). Comparisons between the reference and the pooled cohorts resulted in LOD scores located almost exclusively around zero, which was not the case for mtDNA. Overall, calculating RMP from data taken at different time periods did not yield great differences for the majority of haplotypes.

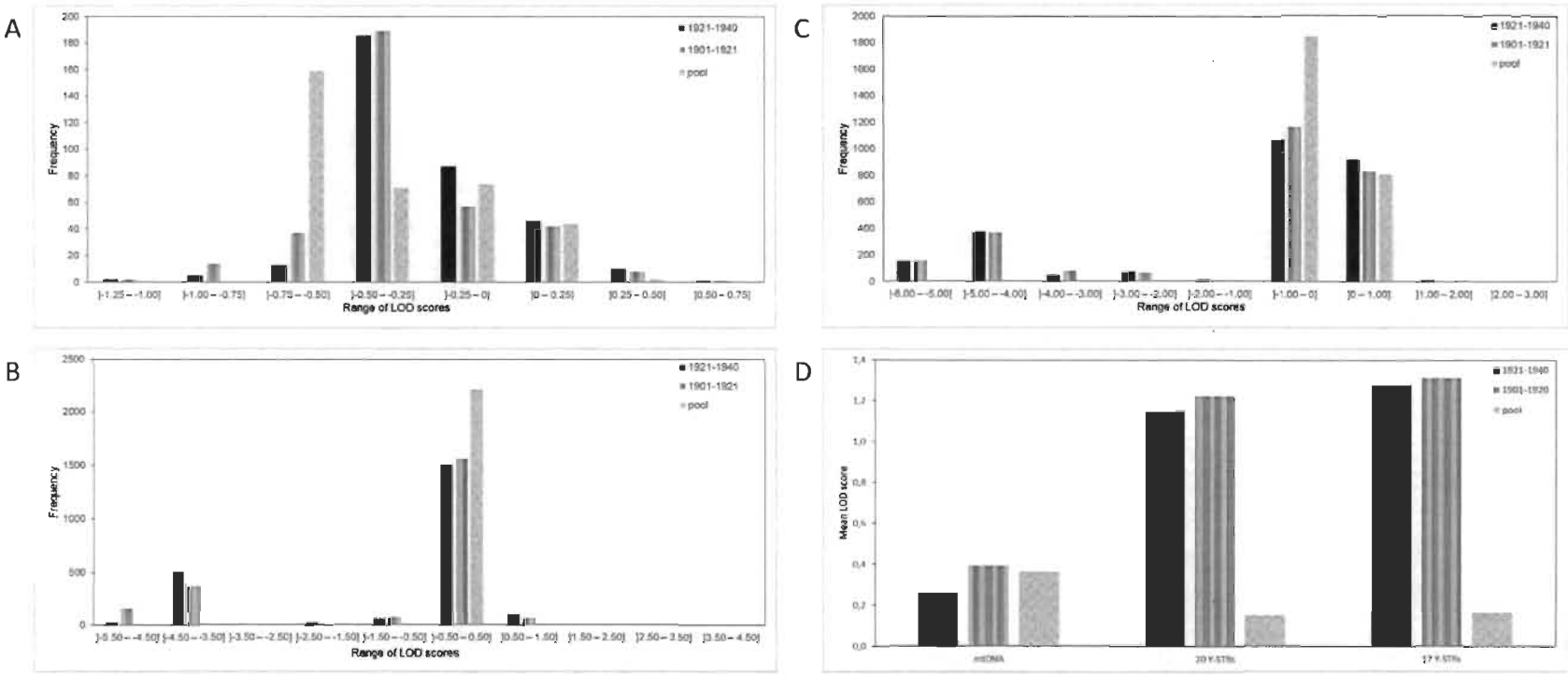

Figure 2.8. Between cohorts differences in random match probabilities (measured with LOD scores) for the mitochondrial DNA (A) and the Y chromosome with 20 Y-STRs (B) and 17 Y-STRs (C). The 1941-1960 cohort was used as the reference to which 'alternative' cohorts were compared by subtracting log(RMP) values for each haplotype present in the reference. Panels A, B and C show histograms of haplotypic LOD scores and panel D the mean (absolute) value for each pairwise comparison.

We also compared the RMP among 15 regions and 507 localities for the mtDNA. LOD scores ranged from <0.001 to 3.6 (between Estrie and Îles-de-la-Madeleine regions; **Table 2.2**). The latter value corresponds to an odd ratio of 3,734, meaning that the RMP was 3,734 times higher in one region than the other, while a value of <0.001 indicates that the RMP was essentially the same between the two regions (odd ratio ≈1). Maximal LOD scores were on average smaller when comparing individual regions with the sample of pooled regions (**Table 2.2**). This suggests that if the population of interest was defined at the regional scale, a pool of regional samples could make a better reference database for haplotype frequencies compared to selecting a wrong region. A LOD range of similar magnitude was measured at the scale of localities, going from 0 to 3.8 (between a locality from Côte-Nord and one from Québec (agglomération); **Table 2.2**). Contrasting with inter-regional comparisons, maximal LOD scores were on average higher when comparing individual localities with the sample of pooled regions, the largest value being 4.4 (between a locality from Laurentides and the sample of pooled regions) compared to maxima of 3.8 and 3.6 between localities and regions, respectively (**Table 2.2**). Thus, if the population of interest was defined at the local scale, a pool of regional samples may not be an adequate reference database for some localities characterized by a genetic composition far from the average.

For the Ychr, RMP values were calculated for 5 regions and 52 localities (20 Y-STRs) or for 11 regions and 108 localities (17 Y-STRs; **Table 2.2**). LOD scores for pairwise comparisons among regions ranged between <0.001 and 7.4 (Gaspésie vs. Îles-de-la-Madeleine) with 20 Y-STRs and between <0.001 and 7.8 (Québec (agglomération) vs. Îles-de-la-Madeleine) with 17 Y-STRs, the highest scores corresponding to an odd ratio of magnitude 10^7. As for the mtDNA, LOD scores were more important in pairwise region comparisons than when comparing individual regions to the sample of pooled regions, with a maximum value of 6.5 (Côte-du-Sud with 20 Y-STRs) or 6.4 (Bas-Saint-Laurent with 17 Y-STRs; **Table 2.2**). At the scale of localities, LOD scores for 20 Y-STRs ranged between <0.001 and 6.9 (between a locality from Gaspésie and one from Bas-Saint-Laurent), and those for 17 Y-STRs ranged between <0.001 and 7.4 (between a locality from Saguenay–Lac-Saint-Jean and one from Beauce; **Table 2.2**). Again, as for

the mtDNA, maximal LOD scores were at least ten times higher when comparing individual localities to pooled regions than in pairwise locality comparisons. The maximal value was 8.1 (with 20 Y-STRs) or 8.6 (with 17 Y-STRs) between a locality from Estrie and the pooled regions, compared to maximal values of 6.9 and 7.4 (with 20 Y-STRs) or 7.4 and 7.8 (with 17 Y-STRs), respectively in pairwise locality and region comparisons (**Table 2.2**).

Table 2.2. Differences in random match probabilities (LOD scores and odd ratios) among regions and localities for the mitochondrial DNA and the Y chromosome.

	Pairwise comparisons among regions	Odd ratio	Comparison of a region to all regions pooled	Odd ratio	Pairwise comparisons among localities	Odd ratio	Comparison of a locality to all regions pooled	Odd ratio
MtDNA								
Average minimum	0.005	1.0	<0.001	1.0	0.03	1.1	0.005	1.0
Average maximum	2.3	198.6	1.7	56.2	1.5	33.6	3.1	1,305.7
Minimum	<0.001	1.0	<0.001	1.0	0	1.0	<0.001	1.0
Maximum	3.6	3,734.1	2.5	344.5	3.8	6,736.5	4.4	2.8×10^{4}
20 Y-STRs								
Average minimum	0.01	1.0	0.002	1.0	0.05	1.1	0.006	1.0
Average maximum	6.9	8.8×10^{6}	5.9	8.4×10^{5}	5.8	5.8×10^{5}	5.1	1.2×10^{5}
Minimum	<0.001	1.0	<0.001	1.0	<0.001	1.0	<0.001	1.0
Maximum	7.4	2.8×10^{7}	6.5	2.9×10^{6}	6.9	8.3×10^{6}	8.1	1.2×10^{8}

Table 2.2 (continued)

	Pairwise comparisons among regions	Odd ratio	Comparison of a region to all regions pooled	Odd ratio	Pairwise comparisons among localities	Odd ratio	Comparison of a locality to all regions pooled	Odd ratio
17 Y-STRs								
Average minimum	0.006	1.0	0.001	1.0	0.02	1.0	0.003	1.0
Average maximum	6.9	8.7×10^{6}	6.1	1.3×10^{6}	6.1	1.2×10^{6}	5.3	1.8×10^{5}
Minimum	<0.001	1.0	<0.001	1.0	<0.001	1.0	<0.001	1.0
Maximum	7.8	6.7×10^{7}	6.4	2.4×10^{6}	7.4	2.6×10^{7}	8.6	3.6×10^{8}

Note: odd ratios correspond to the number of times the largest of the two RMPs compared is greater than the smallest, and is calculated as 10^{LOD}, rounded to the first digit.

Comparison with international databases

We compared RMP values obtained in the Charlevoix region, for the five most common haplotypes in this region, using frequencies estimated from the French-Canadian population data (i.e. this study) with those estimated by searching the same haplotypes in international databases. For the mtDNA, none of the five haplotypes were observed in the EMPOP database no matter which population was used (**Table 2.3**). Moreover, RMP values obtained with the worldwide population were 1,300 to 3,300 times smaller than values obtained in the Charlevoix population, while RMP obtained in the European population were 350 to 900 times smaller (**Table 2.3**). When comparing the smallest and largest RMP values obtained across all localities kept for spatial analyses (n=507 localities) for the same five haplotypes, these were 40 to 10^4 times higher than values obtained in the worldwide population.

For the Ychr, the comparison with the YHRD database was done with the five most frequent 17 Y-STR haplotypes (excepting virtual haplotypes, see Methods) from the Charlevoix region (**Table 2.3**). Only the first two haplotypes were observed in the worldwide population (Y1 was observed twice and Y2, 36 times). Due to the large number of individuals contained in the worldwide population (n=165,259), RMP values for these two haplotypes were of the same magnitude as for the other three unobserved haplotypes. This is because using n_h/N (for observed haplotypes) or $(n_h+1)/(N+1)$ (for unobserved ones) gives similar values when $n_h << N$. Compared to the Charlevoix region, RMP values obtained with the worldwide population were 150 to 10^4 times smaller. None of the five haplotypes were observed in the Eurasian-Caucasian and French populations. Nevertheless, RMP values for Eurasians-Caucasians were of the same magnitude than in Charlevoix and essentially determined by the small number of individuals in the Eurasian-Caucasian sample (n=19). Recall that when a haplotype is not observed in an international database, the RMP value is totally determined by N. RMP values were 5 to 30 times lower in the French population, the source of French-Canadian founders (Charbonneau *et al.*, 2000), due to the larger sample (n=557) for this population. Finally, when comparing the smallest RMP values obtained across all localities kept for spatial analyses (n=108 localities) for the same five haplotypes, these were 200 to 5,000 times smaller than those

obtained in the worldwide population, while the highest RMP values in our localities were 400 to 3×10^4 times higher.

Table 2.3. Comparison of random match probabilities obtained using French-Canadian frequencies, as provided by our genealogico-molecular model, vs. those from the EMPOP (mtDNA) and the YHRD (Ychr) databases for the five most common haplotypes in the Charlevoix region.

Haplotype[a]	Charlevoix region[b]	Worldwide[c,d]	Europe[c] (*n*=8,941)	Eurasian-Caucasian[c] (*n*=19)	France[c] (*n*=557)
MtDNA					
H1 (16234T, 263G)	0.05	<0.0001	0.0001	-	-
H2 (16298C, 72C, 263G)	0.09	<0.0001	0.0001	-	-
H3 (16126C, 16294T, 16296T, 16304C, 73G, 151T, 263G)	0.07	<0.0001	0.0001	-	-
H4 (16129A, 16223T, 73G, 152C, 199C, 204C, 207A, 250C, 263G)	0.1	<0.0001	0.0001	-	-
H5 (16093C, 16224C, 16311C, 73G, 195C, 263G)	0.04	<0.0001	0.0001	-	-
Ychr					
Y1 13_23_29_17_14_12_19_10_16_24_11_13_15_11_14_13_12	0.01	<0.0001	-	0.05	0.002
Y2 13_23_29_17_14_12_19_11_16_24_12_13_15_11_14_13_11	0.03	0.0002	-	0.05	0.002

Table 2.3 (continued)

Haplotype[a]	Charlevoix region[b]	Worldwide[c,d]	Europe[c] (*n*=8,941)	Eurasian-Caucasian[c] (*n*=19)	France[c] (*n*=557)
Y3 13_24_29_17_13_12_19_11_16_25_12_13_15_11_12_13_12	0.04	<0.0001	-	0.05	0.002
Y4 13_24_29_17_14_12_19_10_16_24_11_13_15_11_14_13_12	0.06	<0.0001	-	0.05	0.002
Y5 9_23_25_18_14_12_19_10_16_23_12_13_15_11_16_13_13	0.02	<0.0001	-	0.05	0.002

[a] 17 Y-STRs alleles are given in the following order : DYS389I, DYS635, DYS389II, DYS458, DYS19, YGATAH4, DYS448, DYS391, DYS456, DYS390, DYS438, DYS392, DYS437, DYS385a, DYS385b, DYS393 and DYS439
[b] $n_h+n_{hmin}/N+n_{hmin}$
[c] n_h/N when the haplotype was observed otherwise $n_h+1/N+1$ was used
[d] n=33,691 for mtDNA and n=165,259 for Ychr

Discussion

In spite of the utility of mtDNA and Ychr for forensic genetics, there is actually no satisfying method to evaluate the weight-of-evidence for these lineage markers. Also, the definition of the population of interest is a complex, yet often overlooked issue. Most of the time, laboratories will use a sample that they will consider representative of a large population (e.g., national or international), but is it relevant? Here, we combined genealogical and molecular data to obtain fine-scale measurements of mtDNA and Ychr haplotype frequencies in time and space in the French-Canadian population of Québec. We then compared the RMP among 20-year cohorts, regions and localities for every haplotype sampled in the population, as well as with international databases for the most common haplotypes. This allowed us to assess the impact of population genetic structuring at these markers and of the time elapsed since the reference sample was collected on the evaluation of the weight-of-evidence.

Molecular data coverage and molecular mismatches in lineages

To our knowledge, this is the first study to link genealogical and molecular data to estimate haplotype frequencies at the whole-population scale. This was possible owing to the exceptional quality and coverage of the genealogical data in the Québec population registers, and to the specific history of the French-Canadian population, known from its foundation. Thus, there is a large variation in the individual genetic contribution of 17th and 18th century founders to the gene pool of the modern population, with some founders having very large contributions. Tremblay and Vézina (2010) showed that 64 % of male founders and 53 % of female founders contributed to only one person's genome among 2,221 subjects married between 1945 and 1965, while 0.17 % of male founders and 1 % of female founders contributed to the genome of more than 20 subjects. Two female founders contributed to the genome of almost 5 % of the 2,221 subjects and five male founders to that of 4.5 % of subjects. Taking advantage of this genealogical structure, we imputed a haplotype to several hundreds of thousands (Ychr) or millions (mtDNA) of individuals who lived in Québec between 1621 and 1960, from just a few hundred modern individuals typed. The coverage was generally higher for mtDNA than for the Ychr for the different cohorts and regions, consistent with the fact that there were more typed lineages for the mtDNA. In addition, the imputation from the 275 men genotyped with 17 Y-STRs increased the coverage by at least twofold compared to the imputation from the 175 men genotyped with 20 Y-STRs. For both types of markers, typed lineages (i.e. with molecular information imputed) included a higher average number of individuals than untyped lineages, the latter mostly composed of fewer than 1,000 individuals.

We could estimate mismatch rates due to mutations by subtracting the mismatch rate due to genealogical errors from the global mismatch rate. Thus, we obtained a value of 0.4 % for maternal lineages and 4.4 % for paternal lineages, which is consistent with Ychr haplotype mutation rates (5 % per haplotype per generation with 17 Y-STRs and 6 % with 20 Y-STRs calculated from data in Ballantyne *et al.* 2010 and Willuweit and Roewer 2018a) and mtDNA mutation rate (0.43% per generation (Sigurðardóttir *et al.*, 2000)). Moreover, our results testify the high quality of genealogical data since the majority of mismatches are due to mutations. Jomphe (2011) estimated over-pairing rates

(a type of genealogical errors) of 0.55 % in maternal lineages and 0.82 % in paternal lineages, which is similar to our mismatch rates due to genealogical errors (0.4 % for maternal lineages and 0.8 % for paternal lineages). Moreover, Milot *et al.* (2017) estimated an error rate of 0.38 % for maternal lineages using a simpler model that, unlike ours, does not account for the possibility that a mismatch is caused by more than one genealogical error. The similarity of their estimate to ours suggests that multiple genealogical errors along a given lineage are rare. This is not surprising considering that the mean number of meioses separating two typed individuals is 20 for maternal lineages and 15 for paternal lineages.

Our analyses allowed to estimate the historical extra-pair paternity rate in the French-Canadian population, which was 0.4 % (95 % CI: 0.1–1 %). This estimate is of a similar magnitude, albeit slightly lower to that (~1.20 %) reported by Heyer *et al.* (1997) for the same population based on 9 Y-STRs and 12 small ascending genealogies, or those (~1-2%) published for most contemporary human populations (Larmuseau *et al.*, 2016b). Note that this EPP rate estimation assumes that the difference between global mismatch rates at mtDNA and Ychr are due entirely to differences either in mutation rates or EPPs (i.e. it assumes identical rates of genealogical errors for maternal and paternal lineages).

Genetic diversity in time and space

Genetic diversity was relatively stable in time for both types of markers. Thus, the probability of randomly drawing two different haplotypes in the population was >99.3 % and <99.999 % at all times (except for the more ancient periods for which coverage was low). Hence, the probability of randomly drawing two individuals with the same haplotype (notwithstanding the identity of the haplotype) was in the order of 10^{-3} to 10^{-2}. Clearly, mtDNA and Ychr are less discriminating than current sets of autosomal STR kits, for which this probability is generally $<10^{-10}$ (Butler *et al.*, 2012). Using only molecular data, Moreau *et al.* (2007) obtained smaller diversity estimates with HVI (0.96) and 7 Y-STRs (0.98) data for three regions of Québec (Saguenay–Lac-Saint-Jean, Montréal and Gaspésie). Piercy *et al.* (1993) and Dubut *et al.* (2003) measured a diversity of 0.996 in a population of British individuals and of 0.97 from a sample of French individuals,

respectively. Tremblay and Vézina (2010) showed that the effective population size was greater for the Ychr than for the mtDNA in the French-Canadian population, contributing to the higher diversity observed in the former.

The difference in diversity between 17 and 20 Y-STR sets was negligible, suggesting that the gain in discriminating power to be expected with new STR kits offering more markers will greatly depend on the genealogical structure of the population. In our case, however, fewer men were typed at 20 than 17 Y-STRs. Although our model compensates for this lower coverage by assigning virtual haplotypes to untyped lineages, at this moment we cannot exclude that part of the discrepancy arose from this difference in coverage. Still, in a study of 12 Finnish subpopulations, Palo *et al.* (2008) observed an increase in the genetic diversity (i.e. discriminating power) from 0.965 with the PowerPlex® Y kit (12 Y-STRs) to 0.992 with the AmpFℓSTR® Yfiler™ kit (17 Y-STRs), as well as a decrease in the interregional variation detected. Moreover, mutation rates of Y-STR markers should correlate with the discriminating power so the latter would likely be higher if we considered the RM Y-STRs in our analyses (however at the cost of making imputation much more complicated).

The temporal stability of haplotypic diversity is largely explained by the fact that several haplotypes were very common from one period to another, hence new haplotypes entering the population at different periods had a limited contribution to the total gene pool. As an example, one mitochondrial haplotype had a frequency of ~5 % from 1717 until 1960, compared to others introduced later (between 1845 and 1960) and having a frequency of $\sim 10^{-6}$. This is in line with the disproportionate contribution of certain founders to the French-Canadian gene pool uncovered by Tremblay and Vézina (2010). Vézina *et al.* (2005), and Bherer *et al.* (2011) showed that late immigrants had a limited genetic impact on the contemporary population compared to first settlers. This, of course, holds for the French-Canadian population and may not be generalized to cosmopolitan areas, such as Montréal, that receive a growing influx of immigrants. The stability of the genetic diversity for the entire French-Canadian population also suggests that haplotype frequencies were relatively stable in time, at least for haplotypes not too rare, supporting

the implicit assumption made by laboratories of a temporal stability of frequencies (but see below the section *Random match probability in time and space*).

Previous studies documented regional and local genetic diversity based on genealogies only (Bhérer *et al.*, 2011; Gagnon and Heyer, 2001), molecular information only (Moreau *et al.*, 2007) or genealogies and molecular information used separately (Moreau *et al.*, 2013; Moreau *et al.*, 2009; Roy-Gagnon *et al.*, 2011), and uncovered a genetic structuring of Québec populations. Gagnon *et al.* (2001) showed that eastern regions in Québec were six times more genetically homogeneous than western regions in the 17th and 18th centuries, according to the genetic contribution of founders. Here we found the lowest genetic diversity in two eastern regions, Charlevoix and Îles-de-la-Madeleine (**Figures 2.S9** and **2.S10**), the latter being an isolated archipelago in the far East (Gulf of St. Lawrence). While other authors have documented genetic structuring in the French-Canadian population (e.g., Bhérer *et al.*, 2011; Roy-Gagnon *et al.*, 2011), genetic stratification at the scale of localities was never studied before. In our study, variation was higher among localities than among regions for both mtDNA and Ychr, indicating a greater genetic stratification of populations at the local scale. Many haplotypes in a region or a locality were not observed in another place.

Random match probability in time and space

Most reference databases are still young (< 20 years) and little is known about what should be their lifespan, i.e. for how many years they will remain representative of the population of interest (Butler, 2015b; Ge *et al.*, 2014). In our population, genetic diversity was stable in time, yet this does not mean that RMPs for specific haplotypes are. Thus, when we compared RMPs between a cohort representing the "current generation" (1941-1960) and two earlier cohorts, using the log of the odd ratio (LOD score), we found that they can differ by several order of magnitude, especially for the Ychr. Moreover, LOD score distributions were much wider for the Ychr than the mtDNA. While most LOD scores were close to zero, hence probably not "forensically significant", about 25 % of Ychr haplotypes showed important RMP differences among cohorts. For the Ychr, pooling the three cohorts resulted in LOD scores on average eight times smaller than when

not pooling them, which is partly explained by the fact that the reference (1941-1960) cohort accounted for nearly 45 % of individuals in the pooled sample, compared to ~30 % for the mtDNA. Thus, coherent with the genetic diversity stability, the frequency of most haplotypes did not change enough over 20 or 40 years to modify substantially the random match probability, although this was not true of some other haplotypes. Recall that this variation is observed in a relatively homogeneous population, i.e. from mainly one ethnic origin with most founders coming from the same country (France). We would thus expect temporal RMP differences to be exacerbated in populations where immigration from various ethnic groups is important and fluctuating (e.g., in big cities), an issue that needs to be explored in future studies.

Our objective was also to determine whether the RMP for a given haplotype would change noticeably among regions and localities, in view of the genetic stratification found here. For both mtDNA and Ychr, differences in RMPs among regions and among localities were generally of the same order of magnitude and implied odd ratios up to 10^7. This means that the RMP varies substantially for the same haplotype in the same large population, depending on the specific sub-population of interest. We also compared RMP values obtained in individual regions and in the sample of pooled regions to simulate a reference sample composed of individuals from diverse regions. We found that if the population of interest would be defined at the regional scale, a pool of regional samples could make a better reference database for haplotype frequencies compared to using a wrong sub-population (e.g., the wrong region). This could occur, for instance, if circumstantial information gives no clue about the regional origin of the source of a DNA stain. The somewhat different pattern was observed at the local scale. Average RMP differences were ~5 to 15 times greater when comparing individual localities to the sample of pooled regions than pairwise differences among localities or regions. Therefore, even if we filtered localities to keep in our analyses only those having a better coverage, pooling them did not temper RMP differences. Thus, if the population of interest would be defined at the local scale, a pool of regional samples may not be an adequate reference database for some localities exhibiting a genetic composition highly differing from the average.

It is important to note that 23 out of 52 localities for the analysis with 20 Y-STRs were located in the region of Gaspésie and 9 in Bas-Saint-Laurent while others were distributed in four regions. With 17 Y-STRs, 28 out of 108 localities were located in Saguenay–Lac-Saint-Jean and 12 in Beauce while others were distributed in nine regions. These regions were also involved in the highest RMP differences observed in pairwise locality comparisons, since their greater coverage resulted in a wider range in local RMP values. For the mtDNA, 23 out of 507 localities were from Côte-Nord, 7 were from Québec (agglomération) and the remaining from 18 other regions.

Several authors have discussed why national or international databases for lineage markers may not provide reliable estimates of haplotype frequencies for a given population of interest (Andersen and Balding, 2017; Holland and Lauc, 2014; Kaestle *et al.*, 2006), and that most haplotypes were observed only once (Butler *et al.*, 2007; Parsons and Coble, 2001), hence the observation of a new haplotype is expected to be a common thing in forensic caseworks (Tully *et al.*, 2001). In agreement with this, none of the most frequent mtDNA haplotypes in our study population were found when searched against international databases, as were three of the five most common Ychr haplotypes. We can easily imagine a case where a much rarer haplotype in our population would be searched against the same databases, not found, and thus be assigned the same RMP value than these common haplotypes, greatly underestimating the weight-of-evidence.

SWGDAM guidelines for the interpretation of lineage markers (2013, 2014) strongly suggest choosing the relevant population to obtain frequency estimates. Our results effectively show that RMP values obtained from the worldwide population would cause an important underestimation of the RMP when compared with appropriate local values measured from our extensive, fine-scale coverage data in the French-Canadian population of Québec. The discrepancy was less pronounced with the European (available for mtDNA) and French (available for Ychr) populations, which share a closer ancestry with our study population, but RMPs in these populations were still one to two orders of magnitude lower than our values, while the Eurasian-Caucasian population gave very close RMPs to ours for Ychr haplotypes.

Limitations of the study

To estimate haplotypic diversity and the RMP, we attributed a different virtual (unobserved) haplotype to each untyped lineage. This was to overcome the possibility that many lineages can be absent from a reference database, as mentioned by Andersen and Balding (2017). Our results showed that most lineages had an intra-lineage diversity of zero for both mtDNA and Ychr. A proportion of 25 % of mtDNA haplotypes were present in more than one lineage while this proportion was of 1.4 % with 20 Y-STRs and 3 % with 17 Y-STRs. This is in agreement with a study by Helgason *et al.* (2003) showing that one mtDNA haplotype (HVI region) appeared in 34 lineages and one Ychr haplotype (10 loci) was present in 26 lineages while most haplotypes for both types of markers were present in one or fewer than 15 lineages (mtDNA: 1 haplotype in 1 lineage, 16 in 2 to 5 lineages, 8 in 6 to 10 lineages and 4 in 11 to 14 lineages; Ychr: 12 haplotypes in 1 or 2 lineages, 6 in 3 to 5 lineages and 2 in 11 or 12 lineages; Helgason *et al.*, 2003). As we did, these authors found a higher proportion of mtDNA than Ychr haplotypes that were common to many lineages. Therefore, our attribution of a different haplotype to each untyped lineage may be more realistic for the Ychr.

Another limitation of our model is the imputation, to the founder of a lineage (full or sub-pedigrees), of haplotypes observed in her/his modern descendants. Thus, we did not account for the possibility that the founder had an unobserved haplotype that later mutated to result in one or more observed haplotypes in his/her descendants. However, doing so would result in the rejection of a vast majority (perhaps >99.9%) of simulations, thereby multiplying computation time to a prohibitive extent with the current model, without providing much additional information. Moreover, our model is to some extent buffered against biases that could result from this limitation because situations where the true haplotype of a founder would be unobserved for the specific lineage and lead to the two, three or four different haplotypes observed in modern descendants are expected to be much less likely than a situation where the founder carry one of the observed haplotypes.

Our model could be refined to deal with both limitations outlined above. For example, we could estimate the probability distribution of haplotype diversity and

frequencies among the founders of both typed and untyped French-Canadian lineages using a Bayesian framework incorporating molecular and genealogical information. With the development of such a model, which was beyond the scope of this study, it would be in theory possible to obtain marginal distributions of RMPs after conditioning on probability distribution for haplotype identity and frequencies in the founder population, as well as on mutation and genealogical error rates. It would also be easier to take into account rapidly mutating Y-STRs that are of great interest in forensic science due to the extra discriminating power they provide. However, their integration in the probabilistic model developed here was complicated by the fact that these markers would have resulted in a higher rate of mismatches, thus increasing much simulation rejection rates and the already important computation time.

Finally, the mutation model used for the Ychr could be improved to consider alleles with a non-integer number of repeats (e.g., 11.1). We did not account for the possibility of mutations causing such incomplete alleles because if we had done so, a lot more haplotypes would have been created by simulations, thus again resulting in a higher rate of mismatches.

Conclusion

In conclusion, the genealogico-molecular model developed here allowed us to increase much the molecular coverage of a population compared to the limited samples usually available for mtDNA and Ychr markers. We were also able to study the spatio-temporal variation in haplotype frequencies at a fine scale since the foundation of the population (albeit with a low coverage for the early period and for some regions). This knowledge should be useful for many disciplines, as no study to our knowledge has examined the dynamic of lineage markers at the scale of a whole population with such an extensive spatio-temporal coverage.

Our results on genetic diversity and random match probabilities support the assumption of stability in haplotype frequencies, at least for the majority of haplotypes. The genetic stratification at fine scale in Québec had a non-negligible impact on the

random match probabilities, questioning the use of standard samples comprising hundreds or even thousands of individuals aiming to represent a large territory (e.g., a country) to calculate the weight-of-evidence for lineage markers. Our results also showed that comparisons with large international databases for this specific purpose are dubious and should be avoided. This brings us back to the ill-defined concept of a population of interest from both the genetic ((Waples and Gaggiotti, 2006)) and investigative perspective ((Butler, 2015a, p. 281-308; Coquoz *et al.*, 2013, p. 303-413; Parson *et al.*, 2014; Szabolcsi *et al.*, 2015)).

This study was made possible by the extended genealogical knowledge available for the French-Canadian population of Québec, and our methodological approach would therefore not necessarily be applicable to other populations. Nevertheless, our findings suggest that other approaches than those currently used to assess haplotype frequencies are required to correctly interpret a DNA match involving lineage markers. In that matter, Andersen and Balding (2017) recently developed a simulation model to estimate the number of men carrying a given Ychr haplotype in a population instead of estimating a RMP that can be used without the need to have extended genealogical information. Finally, the model developed here could help to identify human remains using mtDNA and Ychr markers, to predict frequencies of genetic diseases in genetic epidemiology (Ambulkar *et al.*, 2015; Rosenberg *et al.*, 2016; Wallace *et al.*, 1999) or to study natural selection on functional variants involved in fitness-related traits or diseases (Milot *et al.*, 2017).

References

Aimé, C., Heyer, E., and Austerlitz, F. (2015). Inference of sex-specific expansion patterns in human populations from Y-chromosome polymorphism. *American Journal of Physical Anthropology, 157*(2), 217-225.

Ambulkar, P., Chuadhary, A., Waghmare, J., Tarnekar, A., and Pal, A. (2015). Prevalence of Y chromosome microdeletions in idiopathic azoospermia cases in Central Indian men. *Journal of clinical and diagnostic research, 9*(9), GC01-GC04.

Andersen, M.M., and Balding, D.J. (2017). How convincing is a matching Y-chromosome profile? *PLoS Genetics, 13*(11), e1007028.

Andersen, M.M., Caliebe, A., Jochens, A., Willuweit, S., and Krawczak, M. (2013). Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics, 7*(2), 264-271.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics, 23*(2), 147-147.

Anjos, M.J., Carvalho, M., Andrade, L., Lopes, V., Serra, A., Batista, L., Oliveira, C., Tavares, C., Balsa, F., *et al.* (2004). Individual genetic identification of biological samples: a case of an aircraft accident. *Forensic Science International, 146*, S115-S117.

Ayub, Q., Mohyuddin, A., Qamar, R., Mazhar, K., Zerjal, T., Mehdi, S.Q., and Tyler-Smith, C. (2000). Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. *Nucleic acids research, 28*(2), e8.

Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y., van Duijn, K., Vermeulen, M., *et al.* (2010). Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *The American Journal of Human Genetics, 87*(3), 341-353.

Ballantyne, K.N., Ralf, A., Aboukhalid, R., Achakzai, N.M., Anjos, M.J., Ayub, Q., Balažic, J., Ballantyne, J., Ballard, D.J., *et al.* (2014). Toward male individualization with rapidly mutating Y-chromosomal short tandem repeats. *Human mutation, 35*(8), 1021-1032.

Bhérer, C., Labuda, D., Roy-Gagnon, M.H., Houde, L., Tremblay, M., and Vézina, H. (2011). Admixed ancestry and stratification of Quebec regional populations. *American Journal of Physical Anthropology, 144*(3), 432-441.

Brenner, C.H. (2010). Fundamental problem of forensic mathematics—The evidential value of a rare haplotype. *Forensic Science International: Genetics, 4*(5), 281-291.

Buckleton, J.S. (2005). Validating Databases. In Buckleton, J.S., Triggs, C.M. and Walsh, S.J. (éd.), *Forensic DNA Evidence Interpretation*. Boca Raton: CRC Press.

Buckleton, J.S., Krawczak, M., and Weir, B.S. (2011). The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics, 5*(2), 78-83.

Budowle, B., Allard, M.W., Wilson, M.R., and Chakraborty, R. (2003). Forensics and mitochondrial DNA: applications, debates, and foundations. *Annual review of genomics and human genetics, 4*(1), 119-141.

Butler, J.M. (2010). *Fundamentals of Forensic DNA Typing*. Amsterdam, Boston: Academic Press/Elsevier.

Butler, J.M. (2015a). *Advanced Topics in Forensic DNA Typing: Interpretation*. San Diego, CA: Academic Press/Elsevier.

Butler, J.M. (2015b). The future of forensic DNA analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences, 370*(1674), 20140252.

Butler, J.M., Hill, C.R., and Coble, M.D. (2012). Variability of new STR loci and kits in US population groups. http://www.promega.ca/resources/profiles-in-dna/2012/variability-of-new-str-loci-and-kits-in-us-population-groups/.

Butler, J.M., Hill, C.R., Decker, A.E., Kline, M.C., Reid, T.M., and Vallone, P.M. (2007). *New autosomal and Y-chromosome STR loci: characterization and potential uses.* Proceedings of the Eighteenth International Symposium on Human Identification.

Byrnes, J.K., Myres, N.M., and Underhill, P.A. (2014). Genetic Genealogy in the Genomic Era. In Primorac, D. and Schanfield, M.S. (éd.), *Forensic DNA Applications: an Interdisciplinary Perspective* (p. 483-506): CRC Press.

Charbonneau, H., Desjardins, B., Légaré, J., and Denis, H. (2000). The population of the St-Lawrence Valley, 1608–1760. In Haines, M.R. and Steckel, R.H. (éd.), *A population history of North America* (p. 99-142). Cambridge: Cambridge University Press.

Cockerton, S., McManus, K., and Buckleton, J.S. (2012). Interpreting lineage markers in view of subpopulation effects. *Forensic Science International: Genetics, 6*(3), 393-397.

Coquoz, R., Comte, J., Hall, D., Hicks, T., and Taroni, F. (2013). *Preuve par l'ADN : la génétique au service de la justice*. Lausanne: Presses polytechniques et universitaires romandes.

de Knijff, P., Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., *et al.* (1997). Chromosome Y microsatellites: population genetic and evolutionary aspects. *International Journal of Legal Medicine, 110*(3), 134-140.

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*(3), e31-e35.

Desjardins, B. (1998). Le Registre de la population du Québec ancien. In *Annales de démographie historique* (Vol. 2, p. 215-226).

Dillon, L., Amorevieta-Gentil, M., Caron, M., Lewis, C., Guay-Giroux, A., Desjardins, B., and Gagnon, A. (2018). The *Programme de recherche en démographie historique*: past, present and future developments in family reconstitution. *The History of the Family, 23*(1), 20-53.

Dubut, V., Chollet, L., Murail, P., Cartault, F., Béraud-Colomb, E., Serre, M., and Mogentale-Profizi, N. (2003). mtDNA polymorphisms in five French groups: importance of regional sampling. *European Journal of Human Genetics, 12*(4), 293.

Egeland, T., and Salas, A. (2008). Estimating haplotype frequency and coverage of databases. *PLoS ONE, 3*(12), e3988.

Gagnon, A., and Heyer, E. (2001). Fragmentation of the Quebec population genetic pool (Canada): evidence from the genetic contribution of founders per region in the 17th and 18th centuries. *American Journal of Physical Anthropology, 114*(1), 30.

Ge, J., Sun, H., Li, H., Liu, C., Yan, J., and Budowle, B. (2014). Future directions of forensic DNA databases. *Croatian medical journal, 55*(2), 163-166.

Goedbloed, M., Vermeulen, M., Fang, R., Lembring, M., Wollstein, A., Ballantyne, K.N., Lao, O., Brauer, S., Krüger, C., *et al.* (2009). Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpF*l*STR® Yfiler® PCR amplification kit. *International Journal of Legal Medicine, 123*(6), 471-482.

Gómez-Carballa, A., Moreno, F., Álvarez-Iglesias, V., Martinón-Torres, F., García-Magariños, M., Pantoja-Astudillo, J.A., Aguirre-Morales, E., Bustos, P., and Salas, A. (2016). Revealing latitudinal patterns of mitochondrial DNA diversity in Chileans. *Forensic Science International: Genetics, 20*, 81.

Gusmão, L., Butler, J.M., Carracedo, Á., Gill, P., Kayser, M., Mayr, W.R., Morling, N., Prinz, M., Roewer, L., *et al.* (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. *Forensic Science International, 157*(2), 187-197.

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. *The American Journal of Human Genetics, 72*(6), 1370-1388.

Heyer, E., Brandenburg, J.-T., Leonardi, M., Toupance, B., Balaresque, P., Hegay, T., Aldashev, A., and Austerlitz, F. (2015). Patrilineal populations show more male transmission of reproductive success than cognatic populations in Central Asia, which reduces their genetic diversity. *American Journal of Physical Anthropology, 157*(4), 537-543.

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*(5), 799.

Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J., and Labuda, D. (2001). Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *The American Journal of Human Genetics, 69*(5), 1113-1126.

Holland, M.M., and Lauc, G. (2014). Forensic Aspects of mtDNA Analysis. In Primorac, D. and Schanfield, M.S. (éd.), *Forensic DNA Applications: an Interdisciplinary Perspective* (p. 85-104). Boca Raton, Floride: CRC Press.

Holland, M.M., and Parsons, T.J. (1999). Mitochondrial DNA sequence analysis-validation and use for forensic casework. *Forensic Science Review, 11*(1), 21-50.

INTERPOL DNA Unit. (2009). *INTERPOL global DNA profiling survey - Results and analysis*. http://www.dnaresource.com/documents/2008INTERPOLGLOBALDNASURVEYREPORTV2.pdf.

Jomphe, M. (2011). *Validation des généalogies reconstituées à BALSAC à partir de données génétiques*. Rapport interne.

Kaestle, F.A., Kittles, R.A., Roth, A.L., and Ungvarsky, E.J. (2006). Database limitations on the evidentiary value of forensic mitochondrial DNA evidence. *American Criminal Law Review, 43*(1), 53-88.

Kayser, M. (2017). Forensic use of Y-chromosome DNA: a general overview. *Human Genetics*, 1-15.

Kayser, M., and Ballantyne, K.N. (2014). Y Chromosome in Forensic Science. In Primorac, D. and Schanfield, M.S. (éd.), *Forensic DNA Applications: an Interdisciplinary Perspective* (p. 105-134). Boca Raton, Floride: CRC Press.

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., *et al.* (1997). Evaluation of Y-chromosomal STRs: a multicenter study. *International Journal of Legal Medicine, 110*(3), 125-133.

Kayser, M., Kittler, R., Erler, A., Hedman, M., Lee, A.C., Mohyuddin, A., Mehdi, S.Q., Rosser, Z., Stoneking, M., *et al.* (2004). A comprehensive survey of human Y-chromosomal microsatellites. *The American Journal of Human Genetics, 74*(6), 1183-1197.

Larmuseau, M.H.D., Bekaert, B., Baumers, M., Wenseleers, T., Deforce, D., Borry, P., and Decorte, R. (2016a). Biohistorical materials and contemporary privacy concerns-the forensic case of King Albert I. *Forensic Science International: Genetics, 24*, 202-210.

Larmuseau, M.H.D., Boon, N., Vanderheyden, N., Van Geystelen, A., Larmuseau, H.F.M., Matthys, K., De Clercq, W., and Decorte, R. (2015). High Y-chromosomal diversity and low relatedness between paternal lineages on a communal scale in the Western European Low Countries during the surname establishment. *Heredity, 115*(1), 3-12.

Larmuseau, M.H.D., Matthijs, K., and Wenseleers, T. (2016b). Cuckolded fathers rare in human populations. *Trends in ecology & evolution, 31*(5), 327-329.

Larmuseau, M.H.D., Ottoni, C., Raeymaekers, J.A.M., Vanderheyden, N., Larmuseau, H.F.M., and Decorte, R. (2012a). Temporal differentiation across a West-European Y-chromosomal cline: genealogy as a tool in human population genetics. *European Journal of Human Genetics, 20*(4), 434-440.

Larmuseau, M.H.D., Vanoverbeke, J., Gielis, G., Vanderheyden, N., Larmuseau, H.F.M., and Decorte, R. (2012b). In the name of the migrant father—Analysis of surname origins identifies genetic admixture events undetectable from genealogical records. *Heredity, 109*(2), 90-95.

Larmuseau, M.H.D., Vanoverbeke, J., Van Geystelen, A., Defraene, G., Vanderheyden, N., Matthys, K., Wenseleers, T., and Decorte, R. (2013). Low historical rates of cuckoldry in a Western European human population traced by Y-chromosome and genealogical data. *Proceedings of the Royal Society B: Biological Sciences, 280*(1772), 20132400.

Milot, E., Moreau, C., Gagnon, A., Cohen, A.A., Brais, B., and Labuda, D. (2017). Mother's curse neutralizes natural selection against a human genetic disease over three centuries. *Nature Ecology & Evolution, 1*(9), 1400.

Moreau, C., Lefebvre, J.-F., Jomphe, M., Bhérer, C., Ruiz-Linares, A., Vézina, H., Roy-Gagnon, M.H., and Labuda, D. (2013). Native American Admixture in the Quebec Founder Population. *PLoS ONE, 8*(6), 1-9.

Moreau, C., Vézina, H., Jomphe, M., Lavoie, È.-M., Roy-Gagnon, M.H., and Labuda, D. (2011). When genetics and genealogies tell different stories-Maternal lineages in Gaspesia. *Annals of Human Genetics, 75*(2), 247-254.

Moreau, C., Vézina, H., and Labuda, D. (2007). Effets fondateurs et variabilité génétique au Québec. *médecine/sciences, 23*(11), 1008-1013.

Moreau, C., Vézina, H., Yotova, V., Hamon, R., de Knijff, P., Sinnett, D., and Labuda, D. (2009). Genetic heterogeneity in regional populations of Quebec—Parental lineages in the Gaspe Peninsula. *American Journal of Physical Anthropology, 139*(4), 512-522.

National Research Council. (1996). *The Evaluation of Forensic DNA Evidence.* Washington, D.C.: National Academy Press.

Nei, M. (1987). *Molecular Evolutionary Genetics*. New York, NY, USA: Columbia university press.

Palo, J.U., Pirttimaa, M., Bengs, A., Johnsson, V., Ulmanen, I., Lukka, M., Udd, B., and Sajantila, A. (2008). The effect of number of loci on geographical structuring and forensic applicability of Y-STR data in Finland. *International Journal of Legal Medicine, 122*(6), 449-456.

Parson, W., Gusmão, L., Hares, D., Irwin, J.A., Mayr, W.R., Morling, N., Pokorak, E., Prinz, M., Salas, A., *et al.* (2014). DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. *Forensic Science International: Genetics, 13*, 134-142.

Parsons, T.J., and Coble, M.D. (2001). Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. *Croatian medical journal, 42*(3), 304-309.

Parsons, T.J., Muniec, D.S., Sullivan, K., Woodyatt, N., Alliston-Greiner, R., Wilson, M.R., Berry, D.L., Holland, K.A., Weedn, V.W., *et al.* (1997). A high observed substitution rate in the human mitochondrial DNA control region. *Nature genetics, 15*(4), 363-368.

Piercy, R., Sullivan, K., Benson, N., and Gill, P. (1993). The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *International Journal of Legal Medicine, 106*(2), 85-90.

Prost, S., and Anderson, C.N.K. (2011). TempNet: a method to display statistical parsimony networks for heterochronous DNA sequence data. *Methods in Ecology and Evolution, 2*(6), 663-667.

R Development Core Team. (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Roewer, L. (2009). Y chromosome STR typing in crime casework. *Forensic Science, Medicine, and Pathology, 5*(2), 77-84.

Roewer, L., Kayser, M., de Knijff, P., Anslinger, K., Betz, A., Caglia, A., Corach, D., Füredi, S., Henke, L., *et al.* (2000). A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International, 114*(1), 31-43.

Rosenberg, T., Nørby, S., Schwartz, M., Saillard, J., Magalhaes, P.J., Leroy, D., Kann, E.C., and Duno, M. (2016). Prevalence and genetics of Leber hereditary optic neuropathy in the Danish population. *Investigative ophthalmology & visual science, 57*(3), 1370-1375.

Roy-Gagnon, M.H., Moreau, C., Bhérer, C., St-Onge, P., Sinnett, D., Laprise, C., Vézina, H., and Labuda, D. (2011). Genomic and genealogical investigation of the French Canadian founder population structure. *Human Genetics, 129*(5), 521-531.

RStudio Team. (2015) RStudio: integrated development for R. RStudio, Inc, Boston, MA. http://www.rstudio.com/.

s.a. BALSAC - Fichier de population. http://balsac.uqac.ca/. Accessed on March 18, 2018.

s.a. EMPOP mtDNA database, v3/R11. https://empop.online. Accessed on February 10, 2018.

s.a. (2010). Échantillon de référence québécois - Épidémiologie génétique et génétique des populations du Québec. http://www.quebecgenpop.ca/description.html. Accessed on December 5, 2016.

Schaefer, A.M., Taylor, R.W., Turnbull, D.M., and Chinnery, P.F. (2004). The epidemiology of mitochondrial disorders—Past, present and future. *Biochimica et Biophysica Acta (BBA)-Bioenergetics, 1659*(2), 115-120.

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*(1), 61-70.

Sigurðardóttir, S., Helgason, A., Gulcher, J.R., Stefánsson, K., and Donnelly, P. (2000). The mutation rate in the human mtDNA control region. *The American Journal of Human Genetics, 66*(5), 1599-1609.

SWGDAM. (2013). *Interpretation guidelines for mitochondrial DNA analysis by forensic DNA testing laboratories*. http://media.wix.com/ugd/4344b0_c5e20877c02f403c9ba16770e8d41937.pdf.

SWGDAM. (2014). *Interpretation guidelines for Y-chromosome STR typing*. http://media.wix.com/ugd/4344b0_c5e20877c02f403c9ba16770e8d41937.pdf.

Szabolcsi, Z., Farkas, Z., Borbély, A., Bárány, G., Varga, D., Heinrich, A., Völgyi, A., and Pamjav, H. (2015). Statistical and population genetics issues of two Hungarian datasets from the aspect of DNA evidence interpretation. *Forensic Science International: Genetics, 19*, 18-21.

Templeton, A.R. (2006). *Population Genetics and Microevolutionary Theory*. Hoboken, N.J.: Wiley-Liss.

Torroni, A., Achilli, A., Macaulay, V., Richards, M., and Bandelt, H.-J. (2006). Harvesting the fruit of the human mtDNA tree. *Trends in Genetics, 22*(6), 339-345.

Toscanini, U., García-Magariños, M., Berardi, G., Egeland, T., Raimondi, E., and Salas, A. (2012). Evaluating methods to correct for population stratification when estimating paternity indexes. *PLoS ONE, 7*(11), e49832.

Tremblay, M., and Vézina, H. (2000). New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *The American Journal of Human Genetics, 66*(2), 651-658.

Tremblay, M., and Vézina, H. (2010). Genealogical analysis of maternal and paternal lineages in the Quebec population. *Human Biology, 82*(2), 179-198.

Tully, G., Bär, W., Brinkmann, B., Carracedo, Á., Gill, P., Morling, N., Parson, W., and Schneider, P.M. (2001). Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. *Forensic Science International, 124*(1), 83-91.

Vézina, H., Jomphe, M., Lavoie, È.-M., Moreau, C., and Labuda, D. (2012). L'apport des données génétiques à la mesure généalogique des origines amérindiennes des Canadiens français. *Cahiers québécois de démographie, 41*(1), 87-105.

Vézina, H., Tremblay, M., Desjardins, B., and Houde, L. (2005). Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. *Cahiers québécois de démographie, 34*(2), 235-258.

Wallace, D.C. (2015). Mitochondrial DNA variation in human radiation and disease. *Cell, 163*(1), 33-38.

Wallace, D.C., Brown, M.D., and Lott, M.T. (1999). Mitochondrial DNA variation in human evolution and disease. *Gene, 238*(1), 211-230.

Waples, R.S., and Gaggiotti, O. (2006). What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology, 15*(6), 1419-1439.

White, P.S., Tatum, O.L., Deaven, L.L., and Longmire, J.L. (1999). New, male-specific microsatellite markers from the human Y chromosome. *Genomics, 57*(3), 433-437.

Willuweit, S., and Roewer, L. (2018a). Y-chromosome STR haplotype reference database, R56. https://yhrd.org/. Accessed on February 10, 2018.

Willuweit, S., and Roewer, L. (2018b). Y-chromosome STR haplotype reference database, R57. https://yhrd.org/. Accessed on July 20, 2018.

Wilson Sayres, M.A., Lohmueller, K.E., and Nielsen, R. (2014). Natural selection reduced diversity on human Y chromosomes. *PLoS Genetics, 10*(1), 1-12.

Zerjal, T., Beckman, L., Beckman, G., Mikelsaar, A.-V., Krumina, A., Kucinskas, V., Hurles, M.E., and Tyler-Smith, C. (2001). Geographical, linguistic, and cultural influences on genetic diversity: Y-chromosomal distribution in Northern European populations. *Molecular Biology and Evolution, 18*(6), 1077-1087.

Supplementary materials

Supplementary materials and methods

Technical details on the estimation of haplotype probabilities and frequencies in a population

A fictive example of simulation results for a population of five individuals from different lineages is shown in **Table 2.S1**. The probability that an individual i carries haplotype h [$\Pr_i(h)$] was estimated by dividing the number of simulations where i was imputed h by the total number of simulations kept. Thus, $\sum_h \Pr_i(h) = 1$ and $\sum_i \sum_h \Pr_i(h) = N$, where N is the population size.

Table 2.S1. Y-STR haplotype probabilities after imputation within a fictive population of five individuals. In this example, three haplotypes (Y1, Y2, Y3) were observed in at least one hundredth of individuals in a lineage and others were lumped in the "others" category (see main text for explanations).

Individual	Y1	Y2	Y3	others
1	0.6	0.2	0.1	0.1
2	0.6	0.2	0.1	0.1
3	0.1	0.3	0.3	0.3
4	0.1	0.3	0.3	0.3
5	0.4	0.5	0.05	0.05
n_h	1.8	1.5	0.85	0.85
p_h	0.36	0.30	0.17	0.17
RMP	0.45	0.40	0.29	0.29

The number of carriers (n_h) of haplotype h in a population was estimated as $n_h = \sum_i Pr_i(h)$, and the frequency of h (p_h) was given by $p_h = \frac{n_h}{N}$. Note that n_h can be <1 using our model, like in the case of haplotype Y3 in **Table 2.S1**. In the example shown, the number of haplotype Y1 carriers is 0.6 + 0.6 + 0.1 + 0.1 + 0.4 = 1.8, and the frequency of Y1 in this population of 5 individuals is 1.8/5 = 0.36.

Then, estimation of the genetic diversity ($\widehat{H}$) in a population was based on Nei (1987):

$$\widehat{H} = \frac{N}{N-1}\left(1 - \sum\nolimits_h p_h^2\right)$$

In the example above, the genetic diversity in the population is equal to

$$\widehat{H} = \frac{N}{N-1}(1 - \sum_h p_h^2) = \frac{5}{5-1}(1 - (0.36^2 + 0.30^2 + 0.17^2 + 0.17^2)) = 0.90.$$

Finally, the RMP of haplotypes in a cohort, region or locality was estimated as n_h+$n_{h\min}$/N+n_{hmin}, where $n_{h\min}$ is the minimum value of n_h over all haplotypes in the population (see main text for explanations). Therefore, the RMP for the haplotype Y1 in **Table 2.S1** is

$$RMP = \frac{n_h + n_{hmin}}{N + n_{hmin}} = \frac{1.8 + 0.85}{5 + 0.85} = 0.45.$$

Supplementary tables

Table 2.S2. Geographical distribution and number of Y-STRs analyzed for the 429 men used to estimate the genealogical error rate.

Region code	Number of men	Number of Y-STRs
ABI	7	17
BEA	31	17
CNO	31	7 (n=1) and 17 (n=30)
GAS	176	12 (n=1) and 27 (n=175)
MTL	56	7 (n=3) and 17 (n=53)
LAN	7	17
OUT	5	17
QUE	24	17
SAG	92	12

Table 2.S3. Distribution of typed and untyped mtDNA lineages according to the number of individuals per lineage.

	Typed			Untyped		
Number of individuals per lineage	Number of lineages	Total number of individuals	(%)	Number of lineages	Total number of individuals	(%)
<10	28	148	0.01	38,050	177,514	12.78
10-100	93	3,469	0.21	10,966	218,194	15.71
101-1000	53	22,490	1.38	477	172,501	12.42
1001-10000	188	846,705	51.79	294	759,767	54.70
10001-100000	42	762,132	46.62	5	61,032	4.39

Table 2.S4. Distribution of typed and untyped 20 Y-STR lineages according to the number of individuals per lineage.

	Typed			Untyped		
Number of individuals per lineage	Number of lineages	Total number of individuals	(%)	Number of lineages	Total number of individuals	(%)
<10	31	168	0.17	24,830	110,205	8.16
10-100	40	1,179	1.21	4,938	126,614	9.37
101-1000	29	11,512	11.82	1,652	570,420	42.22
1001-10000	26	73,904	75.87	271	529,595	39.20
10001-100000	1	10,651	10.93	1	14,258	1.06

Table 2.S5. Distribution of typed and untyped 17 Y-STR lineages according to the number of individuals per lineage.

	Typed			Untyped		
Number of individuals per lineage	Number of lineages	Total number of individuals	(%)	Number of lineages	Total number of individuals	(%)
<10	32	172	0.07	24,829	110,201	9.08
10-100	44	1,300	0.55	4,934	126,493	10.42
101-1000	72	33,432	14.22	1,609	548,500	45.20
1001-10000	65	175,228	74.55	232	428,271	35.29
10001-100000	2	24,909	10.60	0	0	0.00

Table 2.S6. Molecular coverage for mitochondrial DNA and Y chromosome in the 24 Québec regions defined in the BALSAC register for the period 1941-1960.

Region name	Region code	Mitochondrial DNA[a]		Y chromosome[b]		
		Total number of individuals	Coverage (%)	Total number of men	Coverage 20 Y-STRs (%)	Coverage 17 Y-STRs (%)
Abitibi	ABI	17,533	56.90	8,508	7.73	19.02
Bas-Saint-Laurent	BSL	44,907	73.60	21,792	17.71	28.07
Beauce	BEA	23,355	76.48	11,606	6.82	30.54
Bois-Francs	BFR	51,519	53.88	25,213	6.36	18.58
Charlevoix	CHA	4,214	79.57	2,096	6.92	35.26
Côte-de-Beaupré	CDB	5,651	72.08	2,767	9.79	25.51
Côte-du-Sud	CDS	25,763	72.41	12,546	16.75	27.18
Côte-Nord	CNO	9,115	69.30	4,385	14.85	32.02
Estrie	EST	66,440	52.41	32,178	6.30	17.44
Gaspésie	GAS	17,958	73.37	8,724	29.08	34.53
Montréal (île de)	MTL	293,060	44.08	136,024	5.02	12.19
Îles-de-la-Madeleine	IMA	1,547	81.00	763	36.44	38.14
Lanaudière	LAN	24,157	48.26	11,824	3.85	9.18
Laurentides	LAU	18,334	49.37	8,886	3.50	9.49
Mauricie	MAU	56,809	49.43	27,494	5.10	13.78
Outaouais	OUT	22,866	35.59	10,113	3.21	8.24
Québec (agglomération)	QCA	72,637	63.01	34,917	7.88	20.74
Québec	QUE	29,705	63.00	14,552	6.16	18.27
Rest of Québec	RES	1,238	69.22	602	8.80	25.42
Richelieu	RIC	51,688	45.47	24,873	4.54	11.20
Montréal (rive nord-ouest)	MRN	17,713	45.68	8,470	3.13	9.32
Montréal (rive sud)	MRS	21,938	41.62	10,503	3.11	9.07
Saguenay–Lac-Saint-Jean	SAG	55,261	79.75	27,396	9.98	34.03
Témiscamingue	TEM	10,928	44.89	5,050	6.00	15.07

[a] mtDNA haplotype imputation done from 970 typed individuals.
[b] Y chromosome haplotype imputation done from 175 (20 Y-STRs) and 275 (17 Y-STRs) typed men.

Table 2.S7. Molecular coverage range for mitochondrial DNA and Y chromosome in 1,188 Québec localities grouped by region for the period 1941-1960.

Region code	Number of localities	Mitochondrial DNA[a]		Y chromosome[b]		
		Total number of individuals	Coverage range (%)	Total number of men	Coverage range 20 Y-STRs (%)	Coverage range 17 Y-STRs (%)
ABI	58	17,533	0.00-74.58	8,508	0.00-18.88	0.00-30.89
BSL	101	44,907	41.84-86.86	21,792	9.09-32.17	15.29-45.05
BEA	48	23,355	62.88-90.54	11,606	1.83-15.49	16.67-41.67
BFR	96	51,519	26.47-84.00	25,213	0.00-13.89	7.30-30.86
CHA	19	4,214	62.96-98.47	2,096	0.00-16.67	11.43-61.38
CDB	15	5,651	53.64-83.69	2,767	4.94-14.22	18.06-31.66
CDS	64	25,763	59.84-90.97	12,546	2.81-37.29	5.77-44.07
CNO	37 or 36[c]	9,115	0.00-89.74	4,385	0.00-40.00	0.00-53.76
EST	131	66,440	25.00-81.06	32,178	0.00-50.00	6.45-50.00
GAS	52	17,958	36.89-100.00	8,724	0.00-61.49	0.00-62.64
MTL	23	293,060	10.00-56.09	136,024	0.00-7.58	0.00-21.21
IMA	7	1,547	76.15-92.19	763	28.85-43.48	33.33-66.67
LAN	57	24,157	37.91-65.63	11,824	0.00-9.24	1.74-34.21
LAU	50	18,334	30.86-63.33	8,886	0.00-13.51	0.00-18.92
MAU	55	56,809	0.00-76.42	27,494	0.00-10.00	0.00-29.07
OUT	67	22,866	2.63-58.57	10,113	0.00-7.32	0.00-20.62
QCA	15	72,637	48.86-66.48	34,917	2.76-13.79	7.41-31.03
QUE	56	29,705	45.09-100.00	14,552	0.00-15.00	0.00-37.00
RES	8	1,238	49.01-100.00	602	0.00-15.15	0.00-39.39
RIC	66	51,688	24.74-64.74	24,873	0.00-12.50	0.00-19.70
MRN	31	17,713	34.01-58.82	8,470	0.00-7.32	0.00-20.00
MRS	41	21,938	0.79-100.00	10,503	0.00-11.54	0.00-16.67
SAG	62	55,261	57.14-95.21	27,396	0.00-21.38	0.00-54.09
TEM	29	10,928	24.10-61.36	5,050	1.04-13.39	7.94-24.62

[a] mtDNA haplotype imputation done from 970 typed individuals.
[b] Y chromosome haplotype imputation done from 175 (20 Y-STRs) and 275 (17 Y-STRs) typed men.
[c] mtDNA: 37 localities, Y-STRs: 36 localities.

Supplementary figures

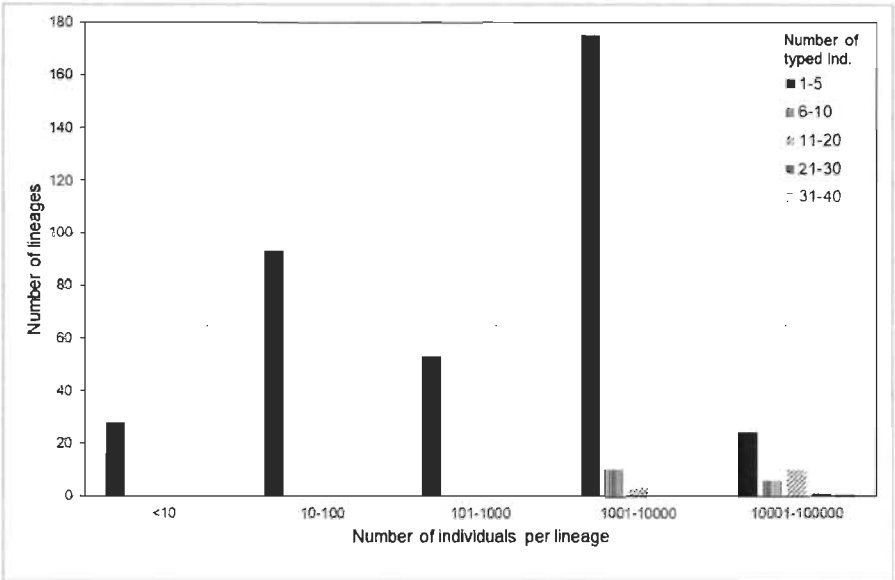

Figure 2.S1. Histogram of the count of mtDNA lineages as a function of the total number of individuals and the number of typed individuals per lineage. Color shades show lineage counts separately according to the number of modern individuals typed per lineage.

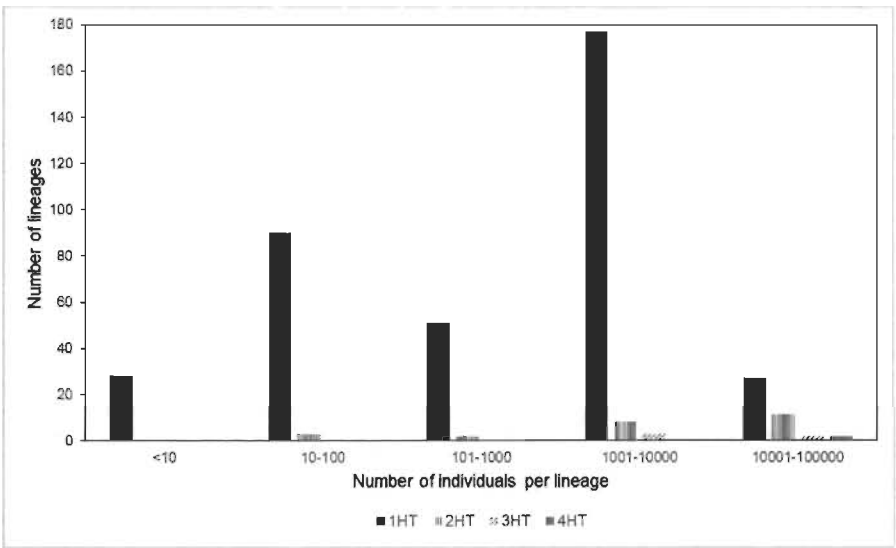

Figure 2.S2. Histogram of the count of mtDNA lineages as a function of the total number of individuals and the number of different haplotypes per lineage. Color shades show lineage counts separately according to the number of different haplotypes ("HT") observed in typed individuals per lineage.

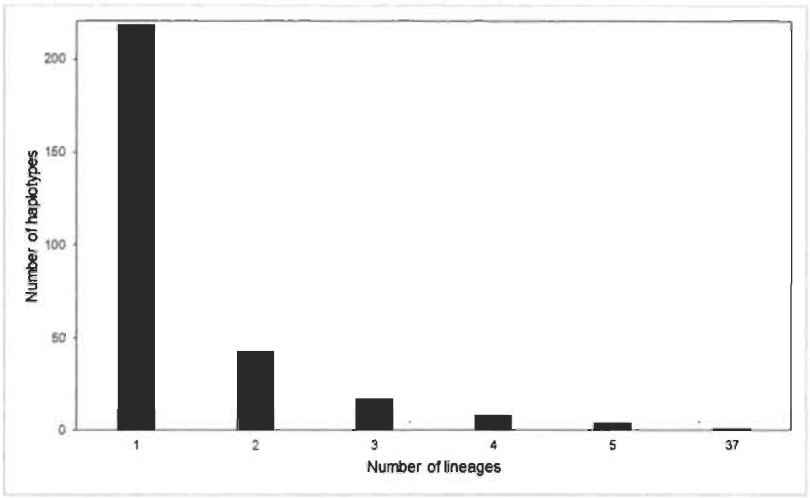

Figure 2.S3. Histogram of the count of mtDNA haplotypes as a function of the number of lineages in which each was observed.

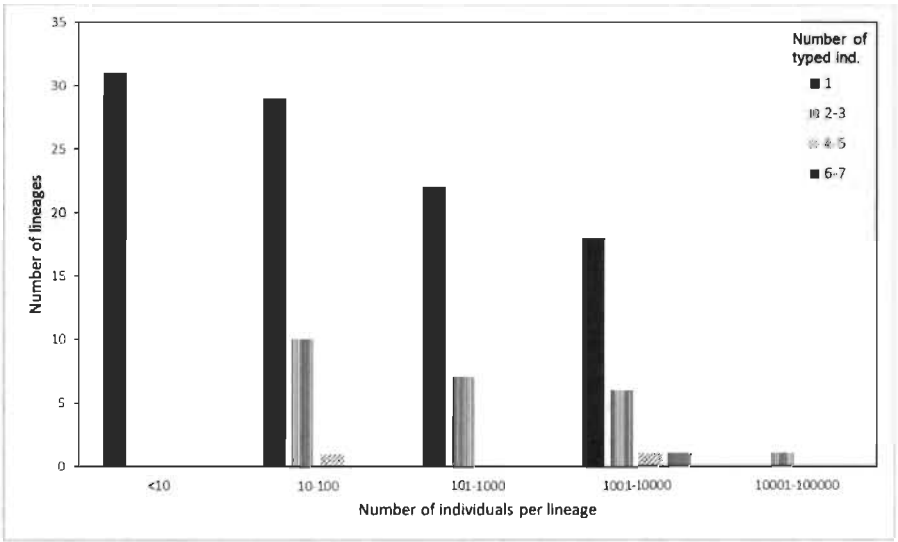

Figure 2.S4. Histogram of the count of 20 Y-STRs lineages as a function of the total number of individuals and the number of typed individuals per lineage. Color shades show lineage counts separately according to the number of modern individuals typed per lineage.

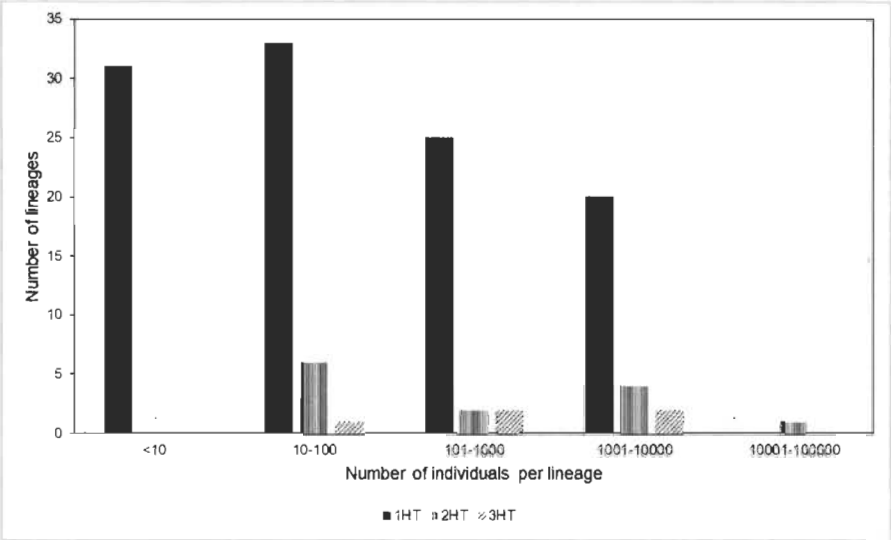

Figure 2.S5. Histogram of the count of 20 Y-STRs lineages as a function of the total number of individuals and the number of different haplotypes per lineage. Color shades show lineage counts separately according to the number of different haplotypes ("HT") observed in typed individuals per lineage.

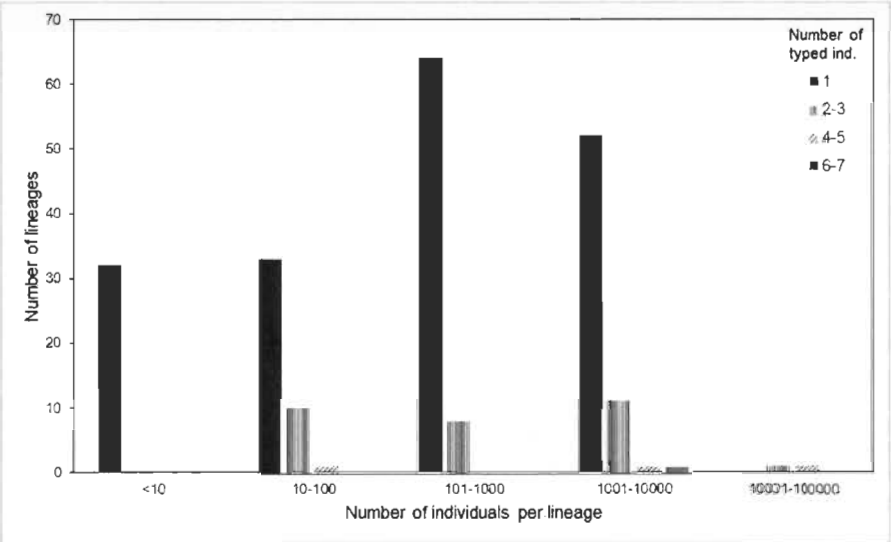

Figure 2.S6. Histogram of the count of 17 Y-STRs lineages as a function of the total number of individuals and the number of typed individuals per lineage. Color shades show lineage counts separately according to the number of modern individuals typed per lineage.

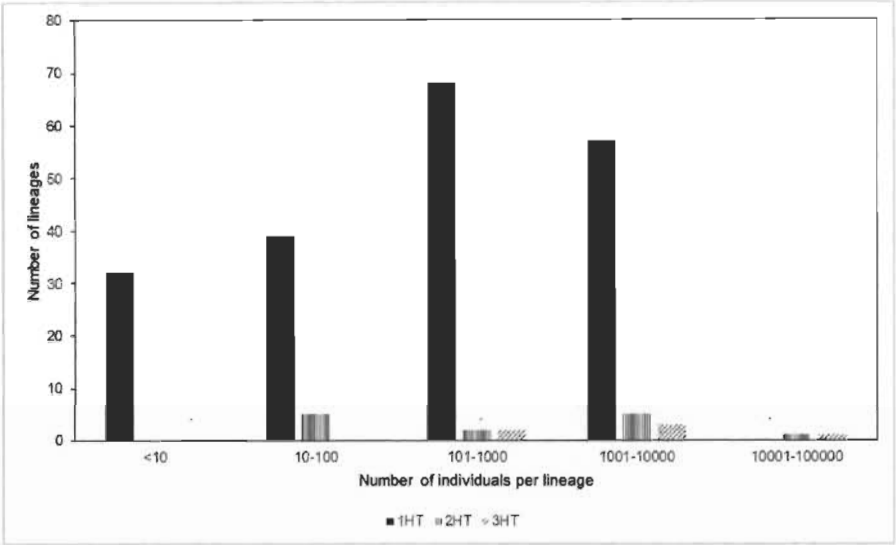

Figure 2.S7. Histogram of the count of 17 Y-STRs lineages as a function of the total number of individuals and the number of different haplotypes per lineage. Color shades show lineage counts separately according to the number of different haplotypes ("HT") observed in typed individuals per lineage.

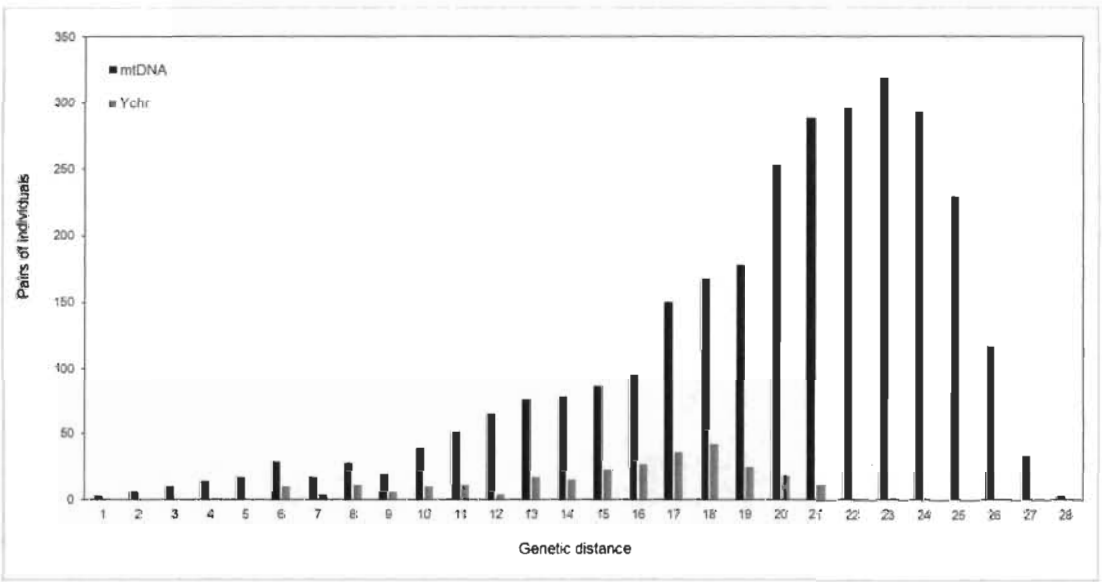

Figure 2.S8. Distribution of the number of meioses separating two modern individuals, for pairs sharing a common ancestor for the mtDNA (n=2,958 pairs) and the Ychr (n=274 pairs).

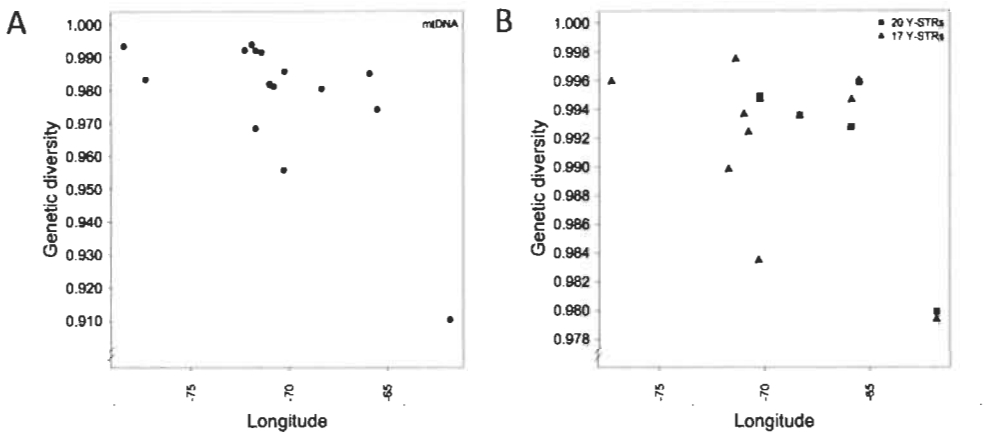

Figure 2.S9. Genetic diversity in Québec regions between 1941 and 1960 for the mitochondrial DNA (A) and the Y chromosome (B) as a function of longitude. Black dots (mtDNA), squares (20 Y-STRs) and triangles (17 Y-STRs) report Nei's (1987) index of diversity. The longitude was obtained by taking the average value for localities of a given region.

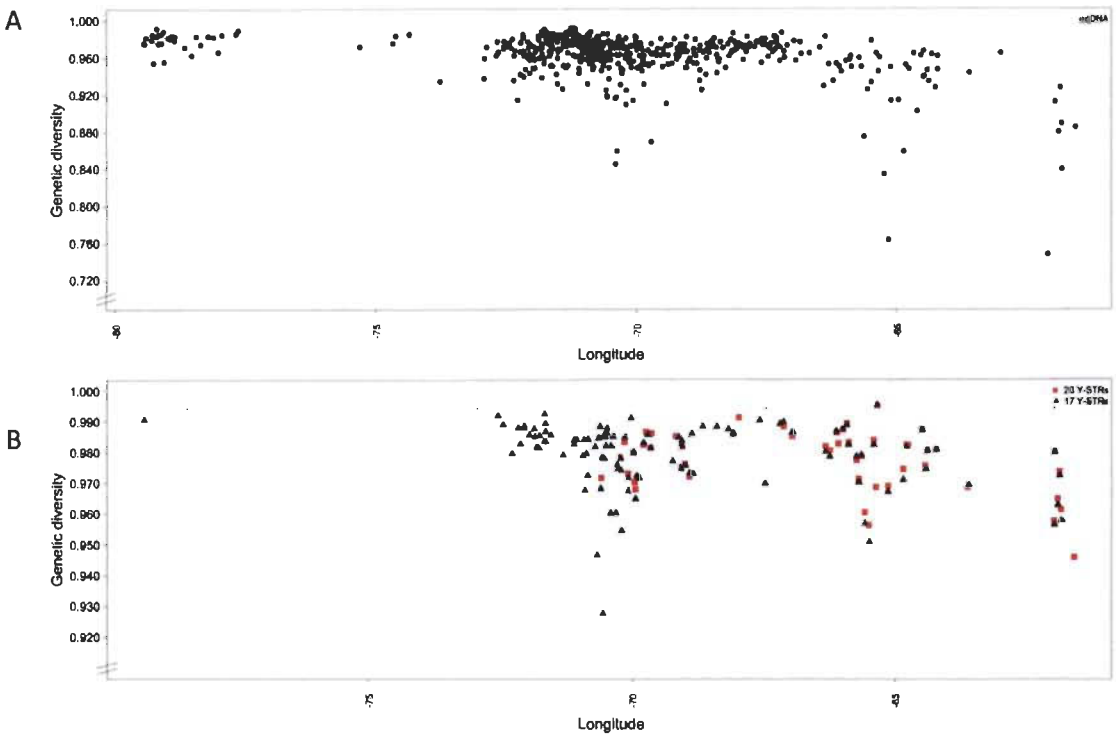

Figure 2.S10. Genetic diversity in Québec localities between 1941 and 1960 for the mitochondrial DNA (A) and the Y chromosome (B) as a function of longitude. Black dots (mtDNA), red squares (20 Y-STRs) and black triangles (17 Y-STRs) report Nei's (1987) index of diversity.

CHAPITRE III

MÉTHODOLOGIE—INFORMATIONS SUPPLÉMENTAIRES

Le présent chapitre introduit brièvement le logiciel utilisé pour réaliser les analyses. Ensuite, une description plus détaillée de certaines étapes de la méthodologie utilisée dans l'article au Chapitre II est présentée. Finalement, ce chapitre contient le détail des essais réalisés avant de parvenir au modèle probabiliste final décrit dans l'article.

3.1 Démarche avec le logiciel R

Le traitement des données ainsi que leur analyse ont été effectués avec le logiciel R v.3.3.2 (R Development Core Team, 2016). Il s'agit d'un logiciel de programmation et d'analyses statistiques utilisé par une grande communauté de chercheurs. Plusieurs modules comprenant des ensembles de fonctions complémentaires peuvent être ajoutés selon les besoins spécifiques d'analyse. Une partie substantielle de la recherche présentée dans ce mémoire a donc été consacrée à l'apprentissage du langage de programmation R. Parmi les connaissances que j'ai développées, il y a notamment la mise en forme des données pour utiliser des fonctions existantes, l'indiçage pour rechercher et manipuler les grands jeux de données, la gestion des messages d'erreurs et l'illustration graphique des résultats. Il était également nécessaire de comprendre et de modifier, au besoin, des fonctions existantes ou des modules pour les adapter aux besoins du projet présenté dans ce mémoire. De plus, cette recherche m'a menée à créer de nouvelles fonctions et à les optimiser en vue de développer un modèle généalogico-moléculaire (voir plus bas). À cet effet, différentes sources d'information ont été utilisées (guide d'utilisation, aide intégrée ou en ligne, forums de soutien technique, etc.). L'aide des techniciens de Calcul Québec a également été nécessaire à certaines étapes clés, en particulier pour l'optimisation. Le code R que nous avons développé pour les fonctions les plus importantes se retrouvent aux Annexes A, B et C.

3.2 Identification des lignées maternelles et paternelles

Une première étape nécessaire à la réalisation de ce projet a été de regrouper les individus du registre BALSAC en lignées maternelles et paternelles. Dans ce projet, une lignée maternelle est définie comme l'ensemble des liens généalogiques passant par les femmes uniquement, et descendant depuis une même fondatrice. Une lignée paternelle passe quant à elle par les hommes depuis un fondateur. Les fondateurs sont définis comme des « immigrants ou individus au-delà desquels il n'est pas possible de poursuivre la généalogie » (Vézina *et al.*, 2012). Pour les lignées paternelles, l'identifiant BALSAC des fondateurs a d'abord été trouvé. Ensuite, un numéro de lignée unique a été donné à chaque fondateur ainsi qu'à tous ses descendants masculins. Pour les lignées maternelles, le principe était le même, sauf que tous les enfants de la fondatrice, garçons et filles, de même que ceux des descendantes, ont été inclus dans la lignée. Le module GENLIB (Gauvin *et al.*, 2015) comprend une fonction nommée *gen.lineages* servant à créer un fichier contenant les lignées maternelles ou paternelles à partir d'une généalogie. Toutefois, le résultat n'étant pas dans le format requis pour la suite des analyses, soit un tableau à deux colonnes (individu, numéro de lignée), nous avons créé une nouvelle fonction pour identifier les lignées.

3.3 Calcul de la distance génétique entre individus génotypés

La distance génétique entre deux individus correspond au nombre minimal de méioses les séparant dans la généalogie (voir Annexe A pour le script utilisé pour l'ADNmt). Pour l'ADNmt, 965 des 970 individus génotypés ont été utilisés et ils totalisaient 465 130 paires à comparer (nombre de paires = $n(n+1)/2 - n$). De même, les 429 individus génotypés pour le chrY totalisaient 91 806 paires. Pour toutes ces paires, la distance génétique a été calculée à l'aide de fonctions du module GENLIB (Gauvin *et al.*, 2015), soit la fonction *gen.find.Min.Distance.MRCA*, permettant de calculer le nombre minimal de méioses séparant deux individus, et la fonction *gen.findMRCA*, pour trouver l'identifiant des plus récents ancêtres communs ainsi que les distances séparant ces ancêtres de chaque individu de la paire. La fonction *gen.findMRCA* fournissant autant les

ancêtres masculins que féminins, le script R a été adapté pour que seuls les hommes soient conservés dans le cas des lignées paternelles, et les femmes dans le cas des lignées maternelles. Toutefois, même lorsque l'ancêtre commun à une paire était du bon sexe, cela ne garantissait pas que seuls les liens père-fils ou mère-enfants étaient retenus pour calculer la distance génétique. La **Figure 3.1** illustre un exemple de cas devant être éliminé : une paire d'hommes partageant un même ancêtre masculin, mais via certains liens mère-fils dans la généalogie, alors que seuls les liens père-fils peuvent être utilisés pour définir une lignée paternelle. Pour éliminer ce type de cas, la fonction *gen.lineages* a été utilisée pour ne conserver que les liens père-fils connectant les individus génotypés au fondateur de leur lignée paternelle. De la même façon, seuls les liens mère-enfants connectant les individus génotypés à la fondatrice de leur lignée maternelle étaient conservés.

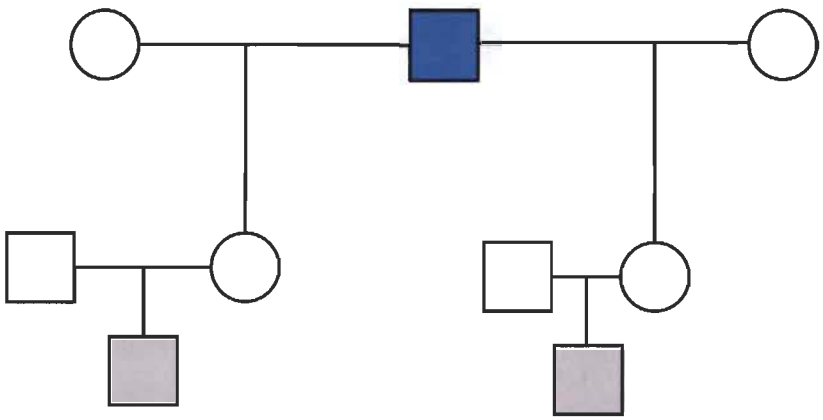

Figure 3.1 **Représentation d'une paire d'hommes ayant un ancêtre commun via des liens mère-fils.**
Les carrés représentent les hommes et les cercles, les femmes. Les individus formant une paire à comparer sont en gris et l'ancêtre commun est en bleu. Le calcul de la distance génétique pour les lignées paternelles devant prendre en compte uniquement les liens père-fils, à partir d'un ancêtre masculin commun aux deux individus, la généalogie illustrée ici n'était pas pertinente pour mesurer une telle distance.

Le temps requis pour calculer la distance génétique pour une paire d'individus sur un ordinateur de bureau standard était d'environ 1 h, ce qui aurait impliqué un temps total de calcul prohibitif pour comparer l'ensemble des > 500 000 paires. Les ressources informatiques de Calcul Québec ont permis d'accélérer substantiellement cette

performance en utilisant le parallélisme, c'est-à-dire l'utilisation de nombreux cœurs informatiques pour effectuer plusieurs calculs simultanément. De plus, nous avons intégré des modifications suggérées par l'équipe technique de Calcul Québec à la fonction *gen.findMRCA* et à ses sous-fonctions, ce qui a permis de réduire davantage le temps de calcul. Une modification du script impliquant la fonction *tryCatch* du module de base a aussi été nécessaire pour éviter que les paires d'individus pour lesquelles aucun ancêtre commun n'existait ne provoquent une erreur de fonctionnement. Une série de tests a été effectuée afin de valider le code R, d'optimiser les différents paramètres de calcul et d'estimer le temps requis pour obtenir les résultats. Avec tous ces ajustements, le temps de calcul de la distance génétique pour une paire d'individus a diminué à une vingtaine de secondes, ce qui était plus acceptable.

3.4 Calcul du taux de non-concordance

À partir des informations sur la distance génétique entre chaque paire d'individus, du nombre de différences entre les haplotypes des individus d'une même paire (une concordance d'haplotypes correspondant à l'observation d'aucune différence entre eux), il était possible d'appliquer la méthode de Larmuseau *et al.* (2013) pour estimer, par une approche de maximum de vraisemblance, la probabilité d'observer une non-concordance. Leur méthode ayant été élaborée pour le logiciel MATLAB (Mathworks, Natick, MA), elle a été adaptée au langage de programmation utilisé dans R (voir Annexe B pour le script).

3.5 Développement du modèle probabiliste pour l'imputation du chromosome Y dans la généalogie

Dans une première version du modèle, les haplotypes des individus génotypés étaient d'abord imputés de façon ascendante à tous les ancêtres faisant partie de la même lignée paternelle en remontant la généalogie jusqu'au fondateur (**Figure 3.2A**). Dans le cas où plusieurs personnes génotypées faisaient partie de la même lignée, plusieurs haplotypes pouvaient donc être attribués à un ancêtre. En exemple, supposons qu'un

ancêtre se voyait imputer deux fois l'haplotype Y1 et une fois l'haplotype Y2, nous considérions que cet ancêtre avait plus de chance d'avoir porté l'haplotype Y1 et celui-ci était donc conservé, l'autre étant rejeté. Lorsque les différents haplotypes avaient le même nombre d'assignations à un ancêtre, ils étaient tous conservés. Ensuite, un haplotype nommé « autre » était ajouté à l'ensemble des haplotypes possibles pour le fondateur de la lignée afin de prendre en compte la possibilité qu'il ait eu un haplotype différent de ceux observés chez ses descendants actuels. Après cette étape d'imputation ascendante, une seconde étape consistait à redescendre toutes les branches de la lignée à partir du fondateur, incluant celles ne menant pas à des individus actuels génotypés (**Figure 3.2B**). Pour ce faire, des simulations étaient effectuées dans lesquelles un des possibles haplotypes du fondateur était choisi aléatoirement à chaque itération. Lors d'une itération donnée, à chaque lien père-fils, une mutation pouvait survenir avec une probabilité correspondant au taux de mutations moyen de l'ensemble des marqueurs. En présence d'une mutation, l'haplotype muté était sélectionné parmi la liste des haplotypes attribués au fondateur. Finalement, l'itération était conservée si les haplotypes imputés par simulation aux individus réellement génotypés correspondaient aux vrais haplotypes observés.

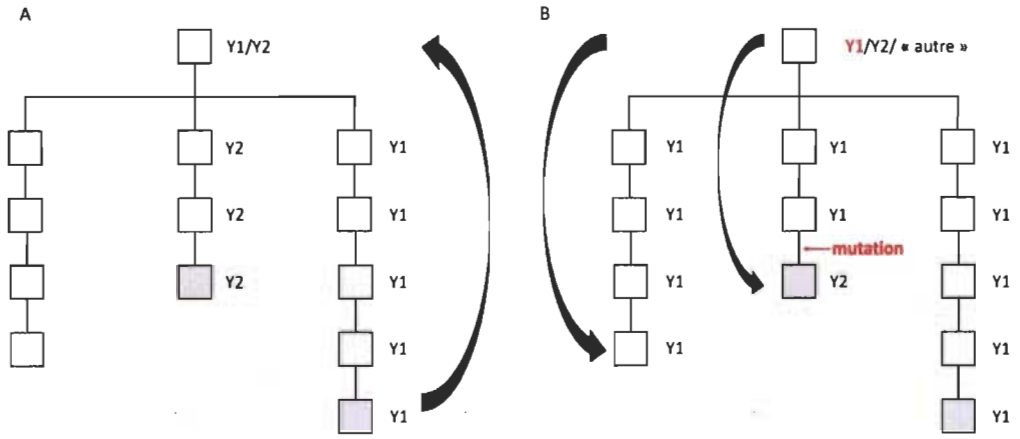

Figure 3.2 Représentation schématique de la première version du modèle d'imputation pour le chromosome Y.
Le pédigrée est simplifié de manière à ne représenter que les liens père-fils. Les hommes génotypés pour le chromosome Y sont en gris pâle. Les différents haplotypes sont indiqués par la combinaison de la lettre Y et d'un chiffre. En A, les haplotypes des individus génotypés sont imputés à leurs ancêtres jusqu'au fondateur. En B, l'haplotype « autre » s'ajoute aux possibilités d'haplotypes du fondateur. À chaque itération, un de ces haplotypes (p. ex. Y1) est tiré au hasard et imputé aux descendants à travers la généalogie. En présence d'une mutation, l'haplotype devient l'un des autres haplotypes imputés au fondateur (Y2 ou « autre » dans ce cas).

Cette première version du modèle comportait plusieurs lacunes. Premièrement, l'haplotype « autre » regroupait certes toutes les possibilités d'haplotypes non observés, mais il n'était pas logique de présumer que celui-ci pouvait immédiatement muter en un des haplotypes observés chez les individus génotypés (p. ex. Y1) sans tenir compte du nombre de mutations nécessaires à cette transition d'un haplotype à un autre. Il a été envisagé d'utiliser une matrice contenant la probabilité de mutation de chaque haplotype vers tous les autres haplotypes, de façon à sélectionner de manière plus réaliste un nouvel haplotype lorsque survenait une mutation dans les simulations. Cette solution n'a pas été conservée dans le modèle final puisqu'il y avait trop de possibilités d'haplotypes et, au final, presque aucune itération n'aurait abouti à un résultat valable (c.-à-d. où les individus réellement génotypés se voyaient imputés le bon haplotype). De plus, cette version du modèle n'intégrait pas la possibilité d'erreurs généalogiques. Des modifications ont donc été apportées au modèle afin de corriger ces lacunes et parvenir au modèle décrit dans l'article au Chapitre II (voir Annexe C pour le script).

Premièrement, le modèle final intègre le taux d'erreurs généalogiques et celui de mutations permettant, à chaque simulation, de générer aléatoirement des erreurs ou des mutations dans la généalogie avant de procéder à l'imputation ascendante des haplotypes à partir des individus génotypés. Le taux de non-concordances dues aux erreurs généalogiques calculé selon la méthode de Larmuseau *et al.* (**Section 3.4**) a été utilisé comme mesure du taux d'erreurs généalogiques spécifique à la population étudiée (et au type de lignée étudié, soit maternelle ou paternelle). Ensuite, l'ensemble des haplotypes imputés à un ancêtre sont conservés et l'un d'entre eux est choisi au hasard selon leur fréquence chez les individus génotypés afin d'imputer un haplotype aux descendants de cet ancêtre. Un ajout important par rapport à la première version du modèle a été d'inclure un modèle de mutation. Ainsi, lorsqu'un fils était identifié comme portant un haplotype muté par rapport à son père (en fonction du taux de mutations moyen des marqueurs), l'haplotype du père était modifié en prenant en compte plusieurs facteurs : taux de mutations de chaque marqueur en prenant en compte la proportion de gains ou de pertes d'unités répétitives (pour déterminer dans un premier temps si une mutation ajoute ou enlève des unités) et de la proportion de cas impliquant un gain ou une perte d'une seule unité répétitive versus deux unités répétitives (pour déterminer si la mutation ajoute/enlève une ou deux unités). Ce modèle de mutation permettait donc de prendre en compte la possibilité que des individus d'une lignée puissent avoir un haplotype différent de ceux observés chez les individus génotypés.

CHAPITRE IV

DISCUSSION ET PERSPECTIVES

L'analyse de l'ADN est utilisée dans plusieurs domaines de la génétique (p. ex. : épidémiologie, démographie, étude sur l'évolution de traits d'histoire de vie, etc.) et notamment en science forensique. Elle constitue une part très importante des analyses effectuées couramment dans les laboratoires judiciaires. L'identification humaine repose principalement sur l'analyse des marqueurs autosomaux puisqu'ils ont un grand pouvoir de discrimination. Toutefois, ceux-ci performent moins bien en présence de traces âgées ou dégradées et de certains mélanges d'ADN de plusieurs personnes, notamment dans les mélanges homme/femme retrouvés dans les prélèvements intimes lors d'agressions sexuelles, et où le profil masculin peut être masqué par le profil féminin lors d'une PCR. Dans ces cas particuliers, les marqueurs haploïdes situés sur l'ADNmt et le chrY peuvent être d'un grand secours. Brièvement, l'ADNmt est présent en de multiples copies dans une seule cellule, faisant en sorte que nous avons plus de chance d'en retrouver des copies intactes dans les traces dégradées ou celles contenant peu de matériel génétique (p. ex. des cheveux). Le chrY permet quant à lui de fournir un profil masculin en présence d'un mélange d'ADN homme/femme, et ce, peu importe le ou les fluides biologiques composant le mélange (sang, salive, sperme).

L'ADN obtenu d'une trace est généralement comparé à celui d'une source connue afin de faire des inférences sur la source de cette trace. Lorsque les deux profils concordent, et que la valeur probante de cette concordance doit être évaluée, il est important de connaître la rareté du profil de la trace dans la population d'intérêt. Il s'agit notamment d'évaluer s'il ne pourrait pas s'agir d'une concordance fortuite entre deux personnes ayant un profil génétique identique. Pour les marqueurs autosomaux, le modèle de Hardy-Weinberg (ou sa version prenant en compte la structure génétique de populations avec le paramètre θ) permet de prédire la fréquence d'un génotype à partir des fréquences alléliques prises dans une base de données de référence. Toutefois, ce

modèle ne s'applique pas pour les marqueurs haploïdes à cause de leur mode de transmission uniparentale. Pour ces marqueurs, l'interprétation repose généralement sur la fréquence de l'haplotype de la trace estimée à partir du nombre de fois que celui-ci a été observé dans un échantillon de référence (parfois aussi appelé « étude de population ») ou une base de données. Cette approche a été critiquée par plusieurs auteurs qui estiment qu'elle est trop conservatrice, dévaluant la valeur probante d'une trace d'ADN. De plus, elle ne prend pas suffisamment en compte la complexité génétique réelle des populations, tel qu'illustré par les résultats de ce mémoire. D'autres auteurs ont proposé des approches pour estimer la rareté d'un haplotype qui n'aurait jamais été observé dans un échantillon de référence ou une base de données, mais au final la même fréquence est attribuée à tous les haplotypes dans ce cas, peu importe leurs caractéristiques moléculaires ou leur fréquence réelle. Nous nous sommes donc questionnés sur la validité des différentes approches utilisées pour estimer la rareté d'un profil pour l'ADNmt et le chrY et plus particulièrement, sur la pertinence de rechercher un profil dans une base de données.

Ainsi, une meilleure connaissance de la dynamique spatio-temporelle de l'ADNmt et du chrY permettrait d'avoir des estimations plus fiables des fréquences des haplotypes. Le premier objectif était donc d'étudier la variation spatio-temporelle de ces fréquences dans la population canadienne-française.

Lorsque les laboratoires judiciaires évaluent une concordance ADN en utilisant les fréquences estimées à partir d'un échantillon de référence ou d'une base de données, ils font implicitement ou explicitement deux prémisses. Premièrement, ils supposent que les fréquences des haplotypes sont stables dans le temps, c'est-à-dire qu'elles n'ont pas varié de manière significative depuis la création de la base de données ou de l'échantillon de référence utilisé. Ensuite, ils considèrent que les fréquences ne diffèrent pas significativement d'un endroit à un autre dans la population d'intérêt. Le deuxième objectif de mon projet consistait alors à tester ces prémisses empiriquement. Finalement, le dernier objectif était de quantifier l'impact du non-respect éventuel de ces prémisses sur le calcul de la probabilité de concordance fortuite.

Quelques études ont montré que la population canadienne-française présentait une structure génétique spatiale en étudiant la généalogie seulement (Bhérer *et al.*, 2011; Gagnon et Heyer, 2001). D'autres portant sur des données moléculaires d'ADNmt et du chrY ont aussi montré l'existence d'une structure génétique entre les régions québécoises (Montréal, Saguenay–Lac-Saint-Jean, Gaspésie) (Moreau *et al.*, 2007; Moreau *et al.*, 2009). Cependant, à notre connaissance, aucune étude n'avait encore calculé les fréquences pour l'ADNmt et le chrY à une échelle aussi fine qu'ici. Pour ce faire, nous avons développé un modèle combinant des données généalogiques et moléculaires de gens connectés à la généalogie canadienne-française. Cela a été rendu possible grâce à la connaissance étendue de la généalogie des Canadien-français mariés entre 1621 et 1960, fournie par le registre de population BALSAC, remontant quasiment à la fondation de la Nouvelle-France en 1608 (à noter que 1621 est l'année la plus ancienne pour laquelle des actes de mariage ont été enregistrés dans les registres catholiques).

4.1 Discussion

À partir d'un échantillon moléculaire de quelques centaines de participants volontaires, nous avons pu imputer un haplotype à 54,1 % de la population canadienne-française mariée au Québec entre 1621 et 1960 pour l'ADNmt et à 6,7 % (avec 20 STR-Y) et 16,2 % (avec 17 STR-Y) de la population masculine pour le chrY. Ainsi, le jumelage de données moléculaires et généalogiques a permis de démultiplier l'échantillon par environ un facteur 1 000. Le nombre de participants génotypés avec 17 STR-Y était presque deux fois plus grand que celui pour les individus génotypés à 20 STR-Y, ce qui explique l'obtention d'une plus grande couverture après imputation avec 17 STR-Y. La couverture correspond à la proportion de la population totale des individus mariés pour lesquels un haplotype a pu être imputé. Celle-ci était aussi plus importante pour l'ADNmt que le chrY, tant au niveau temporel que spatial, en raison du plus grand nombre de participants génotypés pour l'ADNmt (non restreint aux hommes).

La couverture moléculaire était plus petite dans les premières cohortes (avant 1700) que celles plus récentes, ce qui peut s'expliquer entre autres par l'éloignement temporel. En effet, plusieurs lignées anciennes ne se sont pas rendues jusqu'à aujourd'hui et n'ont

donc pas pu être typées via des descendants actuels. Aussi, la complétude des informations généalogiques, c.-à-d. la proportion d'ancêtres connus à une génération donnée, diminue de manière importante lorsqu'on retourne vers le début du 17^{e} siècle. À cette époque, il s'agissait des premiers arrivants en Nouvelle-France et l'information sur ceux-ci est moins complète (Tremblay et Vézina, 2010). Entre régions et localités, la différence de couverture dépend essentiellement de l'échantillonnage des lignées effectué au moment du recrutement des participants, une région avec un grand nombre de participants ayant plus de chance d'être mieux couverte, et de la répartition géographique des lignées plus ou moins populeuses.

Anderson et Balding (2017) font une critique sévère des modèles d'interprétation actuels proposés pour les marqueurs haploïdes. Ces auteurs mentionnent que ceux-ci ne prennent pas en compte le fait qu'une concordance entre deux profils pour le chrY survient plus souvent entre hommes apparentés, mais que ce lien entre hommes est parfois trop éloigné pour être connu d'eux ou de leurs proches. Ils précisent aussi que des lignées paternelles pourraient être absentes de la base de données de référence. D'ailleurs, nos résultats montrent effectivement que nos échantillons moléculaires comptant près de 200 à 1 000 participants selon les marqueurs, couvrent au maximum 404 lignées sur des dizaines de milliers. La majorité des lignées n'étaient donc pas représentées dans nos données moléculaires. Toutefois, dans le cas de la population canadienne-française, ces 404 lignées représentaient près de 55 % des individus. De plus, nous avons observé une grande variation dans le nombre d'individus inclus dans les lignées non-typées, ce nombre pouvant aller de 3 à environ 17 000 individus. Cette disparité devrait être prise en compte dans le calcul de la rareté d'un haplotype, tel que proposé par Andersen et Balding (2017), surtout considérant que les individus apparentés peuvent vivre dans une même région (Andersen et Balding, 2017; Gill *et al.*, 2001; Kayser et Ballantyne, 2014). Imaginons un cas où un haplotype jamais observé auparavant est recherché dans une base de données contenant 1 000 profils. La fréquence calculée avec l'**Équation 1.10** à partir de la base de données considérée comme représentative d'une population hypothétique de 1 million d'individus serait de ~0,001. Toutefois, si la personne ayant cet haplotype fait partie d'une lignée ayant 10 ou 1 000 individus vivants aujourd'hui, la fréquence réelle serait de

0,00001 dans le premier cas et 0,001 dans le deuxième. L'utilisation de l'**Équation 1.10** mène donc à une surestimation de la fréquence dans le premier cas. De plus, la recherche d'un haplotype dans un échantillon de référence ou une base de données repose sur la prémisse que les individus d'une lignée sont répartis de manière homogène sur le territoire de la population d'intérêt, ce que les résultats de cette recherche contredisent.

Dans ce projet, nous avons combiné les données généalogiques aux données moléculaires pour étudier la variation spatio-temporelle des fréquences des haplotypes. Nos résultats montrent que la diversité génétique est restée relativement stable dans le temps, à environ 99 %, autant pour l'ADNmt que le chrY. Cette diversité correspond à la probabilité de tirer au hasard deux haplotypes différents dans la population et donne une indication du pouvoir de discrimination de ces marqueurs. Ce pouvoir est inférieur à celui des marqueurs autosomaux (typiquement >99,9999 %) (Butler, 2010, p. 363-396). Nos valeurs sont aussi en accord avec celles publiées sur la base de données moléculaires uniquement, tant au Québec (Moreau *et al.*, 2007) que dans des populations britanniques (Piercy *et al.*, 1993) ou françaises (Dubut *et al.*, 2003). La littérature rapporte également que le pouvoir discriminant des STR-Y est généralement supérieur à celui de l'ADNmt, ce que nous avons aussi observé (Andersen et Balding, 2017; Butler, 2012, p. 371-403). Cet aspect sera abordé davantage plus loin.

Butler (2012, p. 371-403) a rapporté une augmentation du pouvoir de discrimination en passant de 17 à 37 STR-Y. Palo *et al.* (2008) ont quantifié cette augmentation à 2,8 % en utilisant 17 plutôt que 12 STR-Y. Nos données montrent plutôt un pouvoir quasi identique avec 17 et 20 STR-Y, toutefois il faut dire que la différence entre le nombre de marqueurs utilisés était plus grande dans leurs études que la nôtre. Cela pourrait néanmoins s'expliquer en partie par le fait que les résultats de Butler ou de Palo *et al.* ont été obtenus à partir d'un échantillon d'au maximum ~900 personnes, alors que les nôtres proviennent d'une couverture très grande de la population et d'une connaissance précise de la taille des lignées paternelles. Il faut aussi noter que l'imputation des haplotypes s'est faite à partir d'un plus grand échantillon à 17 STR-Y qu'à 20 STR-Y, ce qui pourrait avoir influencé nos résultats. Il serait aussi pertinent de sélectionner des lignées paternelles

ayant été imputées avec les deux ensembles de marqueurs afin de comparer plus directement le pouvoir de discrimination à 17 et 20 STR-Y. Le taux de mutations des STR-Y est d'ailleurs un autre aspect important à considérer car, plus il est grand, plus le pouvoir de discrimination devrait augmenter. Dans une étude de Ballantyne *et al.* (2014), ce pouvoir a augmenté de 0,99995 (avec 17 STR-Y) à 0,999997 (avec 13 STR-Y à mutation rapide). Le nombre de marqueurs analysés mais également leur taux de mutations sont donc tous les deux des facteurs importants influant sur la diversité génétique. Il serait donc intéressant de prendre en compte, dans notre modèle, les marqueurs à mutation rapide (RM Y-STR), qui n'ont pas été inclus ici pour des raisons déjà évoquées, afin de comparer la diversité génétique obtenue en variant le nombre de ces marqueurs.

Nos résultats montrent que la diversité génétique est demeurée relativement stable dans le temps. Il faut souligner que les premiers arrivants ont eu un avantage démographique par rapport aux immigrants des périodes plus récentes (Bhérer *et al.*, 2011; Vézina *et al.*, 2005), faisant en sorte que leurs haplotypes soient davantage fréquents dans la population contemporaine. D'ailleurs, cela est très bien illustré par le cas des haplotypes mitochondriaux faisant partie de l'haplogroupe H, le plus fréquent dans la population (**Figure 4.1**). Ainsi, les haplotypes les plus communs un siècle après la fondation de la population deviennent très communs les siècles suivants, tandis que leur fréquence ne varie pas beaucoup, même avec l'arrivée de nouveaux haplotypes. Finalement, nos résultats sur la diversité génétique suggèrent que les fréquences des haplotypes sont aussi demeurées relativement stables dans le temps.

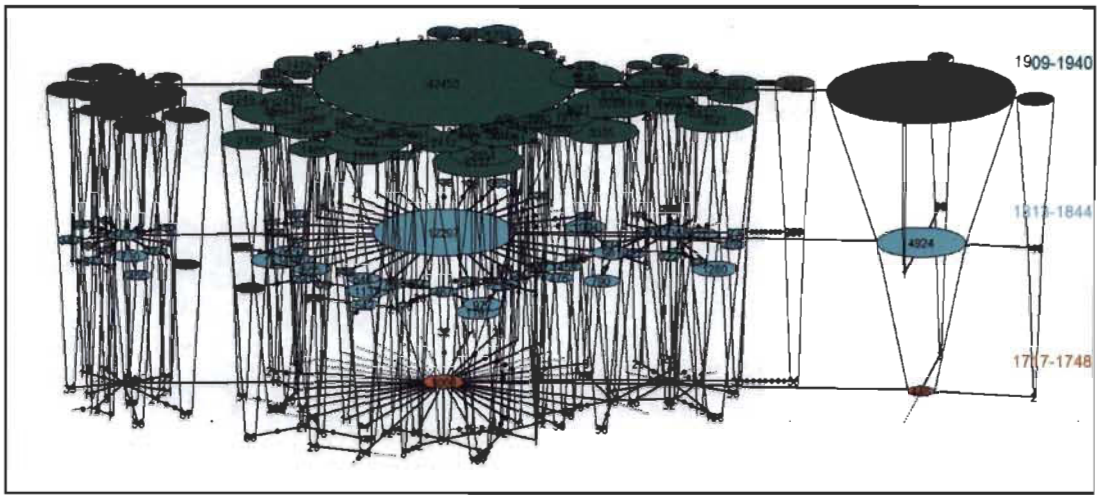

Figure 4.1 Nombre de chaque haplotype mitochondrial de l'haplogroupe H dans la population canadienne-française entre 1717 et 1940, d'après le modèle généalogico-moléculaire développé dans ce projet.
Un ovale représente un haplotype et la taille de celui-ci est proportionnelle au nombre d'individus ayant cet haplotype. Les lignes verticales reliant deux ovales signifient que le même haplotype est présent dans les deux périodes. Cette image a été obtenue par l'analyse des régions HVI et II à l'aide du script R TempNet (Prost et Anderson, 2011). Seuls les haplotypes de l'haplogroupe H sont représentés par souci de clarté.

Selon un sondage mené par INTERPOL (2009), la plupart des bases de données de référence n'ont pas plus d'une vingtaine d'années d'existence. Celles-ci se composent généralement de profils qui s'accumulent depuis le moment de leur création, la prémisse étant que les fréquences sont stables dans le temps. Il serait important de vérifier si l'échantillon de référence demeure représentatif de la population d'intérêt même après plusieurs années, et ce, plus particulièrement pour les marqueurs haploïdes. La littérature actuelle n'en fait aucune mention, mais ce sera certainement un enjeu futur. Nos résultats montrent une certaine stabilité du pouvoir discriminant dans la population canadienne-française pour l'ADNmt et le chrY. Toutefois, cela n'implique pas nécessairement que la fréquence d'un haplotype donné demeure stable dans le temps. Nous avons donc comparé les probabilités de concordance fortuite (RMP) entre une cohorte de référence (1941-1960) et deux autres cohortes de 20 ans (1901-1920 et 1921-1940). Les valeurs de RMP variaient davantage pour le chrY (jusqu'à 10^6) que pour l'ADNmt (jusqu'à un peu plus de 10). Quoi qu'il en soit, les RMP calculées à partir des deux cohortes alternatives étaient

très similaires à celles calculées avec la cohorte de référence pour la grande majorité des haplotypes, et ce, pour les deux types de marqueurs. La cohorte de référence a aussi été comparée à un échantillon formé par le regroupement des trois cohortes et les différences étaient de nouveau petites. De manière générale, nos résultats soutiennent la prémisse de stabilité temporelle des fréquences des haplotypes et indirectement, que la probabilité de concordance fortuite est aussi stable, du moins pour la plupart des haplotypes, à l'échelle de quelques décennies. Le nombre et la taille des lignées maternelles ou paternelles pourraient influencer les différences observées entre cohortes et il serait intéressant de mesurer leur corrélation avec les différences de RMP obtenues à partir de deux références séparées temporellement. En effet, la disparition ou l'apparition de lignées entre les temps t et t+1 pourrait expliquer en bonne partie les différences de RMP obtenues entre deux échantillons de référence récoltés à des temps distincts. Il serait aussi pertinent d'étudier les RMP dans une population moins homogène que celle utilisée dans cette recherche, puisque les variations entre cohortes devraient être plus importantes dans une telle population. Imaginons le cas d'une ville cosmopolite comme Montréal. Le portrait génétique de la population varie probablement plus rapidement que dans la population canadienne-française, rendant plus aigu le problème de l'utilisation des échantillons de référence sur de grandes périodes de temps. En ce qui concerne les marqueurs autosomaux, comme les fréquences des profils sont calculées à partir de celles de chacun des allèles, et que ceux-ci présentent moins de variation qu'un haplotype mitochondrial ou du chrY, la RMP ne devrait pas varier autant dans le temps que pour les marqueurs haploïdes. Toutefois, cela nécessiterait d'être étudié davantage.

Plusieurs études portant sur la population québécoise ont montré la présence d'une structure génétique à partir de données généalogiques (Bhérer *et al.*, 2011; Gagnon et Heyer, 2001), de données moléculaires sur l'ADNmt et le chrY (Moreau *et al.*, 2007; Moreau *et al.*, 2009) et de données généalogiques et moléculaires (SNP autosomaux) utilisées séparément (Moreau *et al.*, 2013; Roy-Gagnon *et al.*, 2011). Nos résultats ont aussi montré une variation de diversité génétique entre régions et une variation plus importante encore entre localités. Cette variation était plus prononcée pour l'ADNmt que le chrY, toutefois il faut rappeler que davantage de régions et localités étaient comparées

dans le cas de l'ADNmt. Aussi, les critères utilisés pour la sélection des régions et localités n'étaient pas les mêmes pour les deux types de marqueurs, la plus grande couverture obtenue avec l'ADNmt nous a permis d'adopter des critères davantage sélectifs. Comme pour l'analyse de la diversité dans le temps, le chrY avait des diversités plus grandes que celles pour l'ADNmt. Cela est principalement dû au taux de mutations plus élevé pour les STR-Y (10^{-4} à 10^{-3} par marqueur par transmission) que pour des changements de nucléotides sur la séquence de l'ADNmt ($\sim 10^{-6}$ par site par génération) faisant en sorte que les haplotypes Y soient plus variables que ceux pour l'ADNmt (Butler, 2015, p. 403-444; Sigurðardóttir *et al.*, 2000). Andersen *et al.* (2017) mentionnent d'ailleurs que le nombre d'individus ayant le même haplotype tend à être plus grand pour l'ADNmt que le chrY notamment à cause du plus petit taux de mutations de l'ADNmt. Tremblay et Vézina (2010) ont montré que la taille de population efficace et la diversité génétique sont plus grandes pour le chrY que l'ADNmt dans la population canadienne-française, ce qui peut aussi avoir contribué à maintenir une plus grande diversité génétique chez le premier. De plus, Moreau *et al.* (2007) ont montré que les populations du Saguenay–Lac-Saint-Jean et de la Gaspésie se distinguaient de celle de Montréal selon l'ADNmt et le chrY. D'autres auteurs (Bhérer *et al.*, 2011; Gagnon et Heyer, 2001) ont également obtenu des diversités plus faibles dans les régions de l'est du Québec (p. ex. : Gaspésie, Charlevoix). Nos résultats pour l'ADNmt et le chrY concordent en partie avec ces études puisque les régions de Charlevoix et des Îles-de-la-Madeleine, situées dans l'est, étaient les moins diversifiées. Les localités les moins diversifiées se situaient aussi souvent dans ces deux régions. Ces dernières étaient les moins peuplées, mais aussi parmi les mieux couvertes de celles analysées dans ce projet. D'autres régions, comme celle représentant « le reste du Québec » (c.-à-d. l'ensemble des régions ne faisant pas partie des 23 régions définies par BALSAC représentées à la **Figure 2.1** de l'article), avaient une couverture et un nombre d'individus similaires, mais une diversité beaucoup plus grande. Cela reflète en partie les lignées présentes dans ces différentes régions. En effet, deux régions ayant une couverture similaire, dont une qui contiendrait deux lignées volumineuses et l'autre, une dizaine de petites lignées n'auraient pas la même diversité. À cet égard, nos résultats montrent clairement que les fréquences des haplotypes ne sont pas homogènes dans

l'espace, ce qui invalide la deuxième prémisse faite par les laboratoires judiciaires[1]. Cela implique que les fréquences retrouvées dans une sous-population ne sont pas nécessairement les mêmes que celles dans une autre sous-population. Concrètement, imaginons qu'un crime ait eu lieu à Montréal et qu'un laboratoire ait accès à deux bases de données, soit l'une composée uniquement de profils d'individus provenant de la région de Montréal et l'autre, de la région de Québec, voire du Québec entier. À la lumière de nos résultats, la seconde base de données risque d'être inadéquate pour évaluer la rareté d'un profil.

Tel que mentionné auparavant, l'utilisation de la méthode de comptage ou de méthodes similaires pour calculer la fréquence d'un haplotype pour le chromosome Y à partir d'une base de données a été critiquée par Andersen *et al.* (2017), d'une part parce que le nombre d'haplotypes différents possible est très grand et d'autre part parce que ces méthodes ne prennent pas en compte la possibilité que des individus apparentés (donc ayant un même haplotype) vivent dans une même zone géographique. Nos résultats portant sur la comparaison des RMP entre régions ou localités soutiennent cette assertion. En effet, les valeurs de RMP variaient considérablement d'une région ou d'une localité à une autre. Pour un haplotype mitochondrial donné, la RMP pouvait être 7 000 fois plus grande dans une population qu'une autre, alors que pour un haplotype Y, la différence de RMP pouvait aller jusqu'à 10^7. Concrètement, cela signifie que la probabilité de tirer au hasard une personne ayant le même haplotype Y que celui observé sur la trace, par exemple, peut être de 10^{-1} dans une population et de 10^{-8} dans une autre, ce qui risque de changer l'appréciation, voire la conclusion du décideur de fait. Étant donné la variation dans la diversité haplotypique, plus grande à l'échelle locale que régionale, nous nous attendions à ce que la RMP varie plus d'une localité à une autre que d'une région à une autre. Les différences étaient en effet deux fois plus grandes au niveau local qu'au niveau régional pour l'ADNmt. Toutefois, pour le chrY, elles étaient dix fois plus petites au niveau local que régional avec 20 STR-Y et du même ordre de grandeur dans les deux cas avec 17 STR-Y. Une raison pouvant expliquer ce résultat est la sélection des localités pour

[1] Les fréquences des haplotypes pour l'ADN mitochondrial et le chromosome Y dans les différentes régions analysées dans ce projet sont présentées aux tableaux **D.1** (Annexe D), **E.1** (Annexe E) et **F.1** (Annexe F).

cette analyse. Près de 45 % des localités comparées avec 20 STR-Y, mais seulement 25 % de celles comparées avec 17 STR-Y, étaient situées dans la même région.

Les valeurs de RMP ont aussi été comparées en prenant chaque région versus l'ensemble des régions retenues selon les critères de couverture (une couverture d'au moins 50 % pour l'ADNmt, 10 % pour 20 STR-Y et 20 % pour 17 STR-Y et une population d'au moins 40 individus), ce dernier pouvant être considéré comme un échantillon de référence composé d'individus provenant de régions diverses. Les résultats suggèrent que le fait de regrouper les régions ensemble homogénéise d'une certaine façon les fréquences des haplotypes pour les deux types de marqueurs. Ce n'était toutefois pas le cas pour les comparaisons entre chaque localité (retenues sur la base d'une couverture d'au moins 60 % pour l'ADNmt, 25 % pour 20 STR-Y et 35 % pour 17 STR-Y et d'une population d'au moins 40 individus) et l'ensemble des régions retenues. Avec ce dernier, les différences observées étaient de 5 à 15 fois plus grandes qu'en comparant les régions ou les localités entre elles. Ainsi, contrairement au niveau régional, dans un cas où la population d'intérêt serait définie au niveau local, un échantillon composé d'individus provenant de diverses régions ne constituerait pas une référence adéquate pour calculer la RMP puisque certaines localités se distinguent de façon importante par rapport à la moyenne. Il serait plus approprié de choisir la bonne localité pour estimer la fréquence d'un profil. Tous ces résultats suggèrent donc que le choix de la population a une incidence importante sur le calcul de la probabilité de concordance fortuite.

Les bases de données pour les marqueurs haploïdes sont considérées par plusieurs comme étant trop petites pour fournir des estimations fiables des fréquences, en particulier pour les haplotypes rares (voir **Sections 1.6.2.1 et 1.6.2.2**) (Andersen et Balding, 2017; Holland et Lauc, 2014). Elles contiennent beaucoup d'haplotypes observés une seule fois (Butler *et al.*, 2007; Parsons et Coble, 2001). Nos résultats sur la comparaison des RMP pour des haplotypes fréquents d'une population québécoise avec les valeurs retrouvées dans des bases de données internationales montrent effectivement que ces haplotypes étaient très rarement observés dans ces bases de données. Déjà en 1992, Krane *et al.* avaient calculé la RMP en analysant quatre marqueurs nucléaires de type VNTR et

comparé les valeurs obtenues à partir d'échantillons de référence de différentes ethnies. Ils ont montré qu'en utilisant un mauvais choix d'ethnie pour comparaison, cela entraînait le plus souvent une sous-estimation des RMP allant de 10 à 100 fois. L'utilisation d'un échantillon mixte entraînait aussi très souvent une sous-estimation de la RMP. De façon similaire et plus prononcée, nous avons effectivement observé que l'utilisation de la population mondiale pour l'ADNmt et le chrY entraînait une sous-estimation importante de la RMP. Même en prenant comme référence la population française, génétiquement très proche des Canadiens-français, la RMP était sous-estimée. Une base de données a fourni des valeurs de RMP du même ordre de grandeur que celles obtenues avec les données utilisées ici, soit celle sur les Eurasiens-caucasiens, curieusement. Toutefois, cela s'explique par le très petit nombre de profils dans cet échantillon (n=19) et à l'utilisation de la formule x+1/N+1. En bref, ces résultats soulèvent un doute important sur la pertinence d'utiliser de grandes bases de données, considérées comme représentatives de plusieurs populations, pour fins d'évaluation de la valeur probante d'une concordance ADN impliquant des marqueurs STR-Y ou l'ADNmt.

4.2 Perspectives

En vue d'améliorer le modèle d'imputation présenté ici pour des développements futurs, différents aspects de ce modèle devront être validés par des études plus poussées qui n'entraient pas dans le cadre de ce mémoire. Premièrement, le modèle imputait au fondateur d'une lignée un haplotype observé parmi les descendants génotypés, alors que ce fondateur aurait pu avoir un tout autre haplotype. Nous pourrions peut-être ainsi imputer au fondateur un haplotype non observé parmi ses descendants actuels, mais cela risque le plus souvent de mener au rejet de la simulation, en particulier pour les lignées plus courtes pour lesquelles la probabilité d'observer une mutation ou une erreur généalogique est très faible.

Avec le modèle probabiliste, nous obtenions aussi la proportion d'itérations pour lesquelles aucun haplotype n'avait pu être imputé à un individu, par exemple en raison d'une erreur généalogique scindant la lignée. Cela pouvait aussi être le résultat d'une itération rejetée, car elle n'aboutissait pas à une imputation des bons haplotypes aux

individus en bout de lignées correspondant aux participants recrutés et dont le vrai haplotype était connu. Cette proportion d'itérations avec des « non-imputations » est intéressante puisqu'elle donne une mesure de l'incertitude sur l'imputation des haplotypes aux individus. Cette information pourrait être intégrée d'une façon ou d'une autre dans les analyses effectuées après imputation. Par exemple, il serait intéressant d'utiliser cette incertitude comme critère pour conserver, dans les analyses de fréquences post-simulations, uniquement les individus dont l'imputation est plus certaine. Différents seuils pourraient être testés afin de voir si cela influence les résultats obtenus par rapport à ceux présentés dans ce mémoire.

Ensuite, afin d'inclure les lignées non-typées pour les analyses spatio-temporelles et le calcul de la RMP, nous avions attribué un haplotype virtuel différent à chaque lignée non-typée. Il sera important de tester cette prémisse du modèle, surtout que la couverture moléculaire dans les différentes régions et localités varie. Pour ce faire, différentes approches sont possibles. Premièrement, le succès reproducteur et les années de mariage des individus pourraient être comparés entre les lignées non-typées et celles typées. Dans le cas où ces paramètres seraient comparables, il serait possible d'utiliser les distributions de diversité et de fréquences haplotypiques chez les fondateurs des lignées typées et utiliser les mêmes pour attribuer des haplotypes virtuels aux lignées non-typées en respectant ces paramètres. Une autre approche pourrait être de calculer la contribution génétique des fondateurs à chaque génération pour ensuite calculer un indice de diversité avec la méthode utilisée par Tremblay et Vézina (2010). À partir de ces informations, il serait possible de faire des simulations d'imputation d'haplotypes en vérifiant que la diversité obtenue correspond à ce qui a été calculé.

L'imputation des haplotypes pour l'ADNmt a été effectuée avec un modèle non-probabiliste pour les besoins de ce mémoire. Une imputation avec un modèle probabiliste tel que développé pour le chrY se fera prochainement au laboratoire du Dr Emmanuel Milot. De cette façon, cela permettra de comparer sur une même base les résultats obtenus pour l'ADNmt et le chrY. Le modèle utilisé pour le chrY pourrait être modifié pour tenir

compte de la variation dans la probabilité de mutation entre les sites nucléotidiques des régions HVI et HVII.

En ce qui concerne l'imputation du chrY, les STR-Y à mutation rapide (RM STR-Y) ont été exclus des analyses principalement pour deux raisons. Premièrement, cela aurait nécessité un modèle de mutation plus poussé que celui utilisé puisqu'il aurait été inadéquat d'utiliser la moyenne du taux de mutations de tous les marqueurs pour attribuer au hasard des mutations à travers la généalogie (le taux de mutations est environ 10 fois plus grand pour les RM STR-Y que les autres STR-Y). Deuxièmement, comme ces marqueurs mutent plus rapidement, il y aurait eu beaucoup plus de non-concordances entre les individus génotypés faisant partie d'une même lignée, entraînant ainsi un rejet de la plupart des simulations et une augmentation énorme du temps de calcul. Ballantyne *et al.* (2014) ont montré qu'il était possible de différencier une paire d'individus appartenant à une même lignée paternelle dans 29 % des cas avec 13 RM STR-Y contrairement à seulement 5,5 % des cas avec 17 STR-Y. Ensuite, il est connu que les haplotypes au sein de certains haplogroupes Y se ressemblent beaucoup. C'est le cas notamment de l'haplogroupe R-M269 qui est très fréquent dans l'ouest de l'Europe (Balaresque *et al.*, 2010) et probablement aussi dans la population canadienne-française (les marqueurs Y analysés ne permettant pas de déterminer l'haplogroupe auquel appartient un haplotype donné). Larmuseau *et al.* (2014) ont observé que pour 31 % des paires d'individus ayant un haplotype identique avec 17 STR-Y, les individus n'appartenaient pas à un même sous-haplogroupe du R-M269. Cette proportion montait à 69 % lorsqu'une seule différence était observée entre les deux haplotypes. Avec 38 STR-Y, toutes les paires affichant moins de trois différences étaient composées d'individus du même sous-haplogroupe, alors que c'était le cas de seulement 58 % de celles présentant trois différences ou plus. Il serait donc pertinent d'augmenter le nombre de marqueurs Y utilisés pour révéler d'éventuelles différences non détectées ici. Présentement, nous ne disposons que d'un petit échantillon génotypé à 27 STR-Y (qui inclus des RM STR-Y), mais avec le recrutement de participants effectués par notre laboratoire depuis 2016, ce nombre augmentera. Finalement, il sera important de valider les seuils utilisés dans le calcul du taux d'erreurs généalogiques avec une approche plus quantitative et probabiliste. À cet effet, Walsh

(2001) a développé une méthode pour déterminer si le nombre de mutations observées entre individus est logiquement possible étant donné le nombre de générations entre ceux-ci et leur plus récent ancêtre commun.

Nos résultats ont montré que ~15 % des individus non-typés étaient non connectés à des lignées, autant pour l'ADNmt que le chrY. De plus, l'imputation des marqueurs dans la généalogie à partir d'individus génotypés ne peut pas se faire pour les individus des lignées éteintes (aucun descendant vivant aujourd'hui). Il y a également plusieurs lignées qui ne sont pas représentées dans l'échantillon moléculaire utilisé pour ce mémoire. Malgré tout, une couverture moléculaire importante de la population a été atteinte avec cet échantillon en comparaison avec des échantillons de référence typiques de quelques centaines ou milliers d'individus. De plus, les individus recrutés par notre laboratoire, une fois jumelés à la généalogie de BALSAC, pourront être ajoutés aux données utilisées, ce qui pourrait permettre de mieux couvrir certaines régions.

Les analyses de diversité génétiques pour évaluer le pouvoir discriminant pourraient être poussées un peu plus loin. Larmuseau *et al.* (2012a) ont, par exemple, utilisé les F_{ST} qui comparent la composition génétique entre paires de populations pour évaluer si deux populations qui ont des diversités similaires ont une composition génétique complètement différente. Cela apporterait une perspective complémentaire à la comparaison faite ici des RMP entre cohortes, régions ou localités.

Les résultats de l'imputation pourraient aussi permettre d'établir des cartes de répartition géographique des fréquences de chaque haplotype. Celles-ci pourraient être utilisées autant en épidémiologie génétique pour comprendre la prévalence de maladies associées à certains haplotypes que pour déterminer la rareté d'un profil en génétique forensique (**Figure 4.2**).

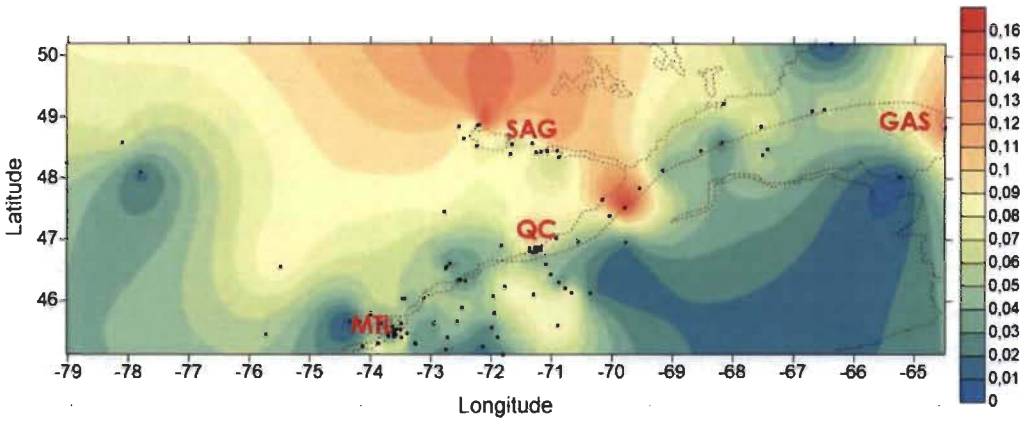

Figure 4.2 **Exemple d'une carte de répartition possible des fréquences pour un haplotype mitochondrial au Québec.**
L'échelle à droite correspond à la fréquence de l'haplotype. Les villes de Montréal (MTL), Québec (QC), Saguenay–Lac-Saint-Jean (SAG) et Gaspésie (GAS) sont indiquées en rouge afin de les situer sur la carte du Québec tracée en pointillés. Les points noirs représentent les localités conservées pour cette analyse.

Finalement, le modèle développé pour ce mémoire est déjà utilisé par des collègues du laboratoire du Dr Emmanuel Milot. Un des projets consiste à identifier des restes humains anciens (19^{e} siècle) en comparant l'ADNmt et le chrY d'ossements aux lignées maternelles et paternelles de BALSAC. Il pourra aussi servir à tester empiriquement la méthode proposée récemment par Andersen *et al.* (2017) qui proposent d'estimer le nombre d'individus portant le même haplotype dans une population, dans leur cas pour le chrY, au lieu d'estimer des fréquences à partir de bases de données de référence. L'avantage de ce modèle est qu'il peut être utilisé pour des populations qui n'ont pas une connaissance aussi étendue de leur généalogie que celle du Québec. Enfin, d'autres applications sont possibles en biologie évolutive, pour l'étude de la sélection naturelle ou de l'association entre l'environnement et la fréquence de certains types d'ADN.

RÉFÉRENCES BIBLIOGRAPHIQUES

Allen, M., Engström, A.-S., Meyers, S., Handt, O., Saldeen, T., von Haeseler, A., Pääbo, S., et Gyllensten, U. (1998). Mitochondrial DNA sequencing of shed hairs and saliva on robbery caps: Sensitivity and matching probabilities. *Journal of Forensic Sciences, 43*(3), 453-464.

Alonso, A., Martín, P., Albarrán, C., Garcí, P., Simón, D., Fernández, L., Iturralde, M.J., Fernández-Rodríguez, A., Atienza, I., *et al.* (2005). Challenges of DNA profiling in mass disaster investigations. *Croatian medical journal, 46*(4), 540-548.

Ambers, A., Gill-King, H., Dirkmaat, D., Benjamin, R., King, J.L., et Budowle, B. (2014). Autosomal and Y-STR analysis of degraded DNA from the 120-year-old skeletal remains of Ezekiel Harper. *Forensic Science International: Genetics, 9*, 33-41.

Ambulkar, P., Chuadhary, A., Waghmare, J., Tarnekar, A., et Pal, A. (2015). Prevalence of Y chromosome microdeletions in idiopathic azoospermia cases in Central Indian men. *Journal of clinical and diagnostic research, 9*(9), GC01-GC04.

Anđelinović, Š., Sutlović, D., Erceg, I.I., Škaro, V., Ivkošić, A., Paić, F., Režić, B., Definis-Gojanović, M., et Primorac, D. (2005). Twelve-year experience in identification of skeletal remains from mass graves. *Croatian medical journal, 46*(4), 530-539.

Andersen, M.M., et Balding, D.J. (2017). How convincing is a matching Y-chromosome profile? *PLoS Genetics, 13*(11), e1007028.

Andersen, M.M., Eriksen, P.S., et Morling, N. (2013). The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies. *Journal of Theoretical Biology, 329*, 39-51.

Anderson, S., Bankier, A.T., Barrell, B.G., de Bruijn, M.H.L., Coulson, A.R., Drouin, J., Eperon, I.C., Nierlich, D.P., Roe, B.A., *et al.* (1981). Sequence and organization of the human mitochondrial genome. *Nature, 290*(5806), 457-465.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., et Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature genetics, 23*(2), 147-147.

Anjos, M.J., Carvalho, M., Andrade, L., Lopes, V., Serra, A., Batista, L., Oliveira, C., Tavares, C., Balsa, F., *et al.* (2004). Individual genetic identification of biological

samples: a case of an aircraft accident. *Forensic Science International, 146*, S115-S117.

Bader, S. (2016). Introduction to Forensic Genetics. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling* (p. 3-8). Royaume-Uni: Wiley.

Balanovsky, O. (2017). Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. *Human Genetics*, 1-16.

Balaresque, P., Bowden, G.R., Adams, S.M., Leung, H.Y., King, T.E., Rosser, Z.H., Goodwin, J., Moisan, J.P., Richard, C., *et al.* (2010). A predominantly neolithic origin for European paternal lineages. *PLoS biology, 8*(1), e1000285.

Balding, D.J., et Nichols, R.A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International, 64*(2), 125-140.

Ballantyne, J., et Hanson, E.K. (2016). Y-Chromosome Short Tandem Repeats. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling* (p. 149-154). Royaume-Uni: Wiley.

Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y., van Duijn, K., Vermeulen, M., *et al.* (2010). Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *The American Journal of Human Genetics, 87*(3), 341-353.

Ballantyne, K.N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S.B., Ralf, A., Vermeulen, M., de Knijff, P., et Kayser, M. (2012). A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Science International: Genetics, 6*(2), 208-218.

Ballantyne, K.N., Ralf, A., Aboukhalid, R., Achakzai, N.M., Anjos, M.J., Ayub, Q., Balažic, J., Ballantyne, J., Ballard, D.J., *et al.* (2014). Toward male individualization with rapidly mutating Y-chromosomal short tandem repeats. *Human mutation, 35*(8), 1021-1032.

Bergeron, J. (2005). *Contribution différentielle des ancêtres d'origine acadienne au bassin génétique des populations régionales du Québec.* Mémoire de maitrise en médecine expérimentale, Université du Québec à Chicoutimi, Québec.

Betz, A., Bäßler, G., Dietl, G., Steil, X., Weyermann, G., et Pflug, W. (2001). DYS STR analysis with epithelial cells in a rape case. *Forensic Science International, 118*(2-3), 126-130.

Bhérer, C., Labuda, D., Roy-Gagnon, M.H., Houde, L., Tremblay, M., et Vézina, H. (2011). Admixed ancestry and stratification of Quebec regional populations. *American Journal of Physical Anthropology, 144*(3), 432-441.

Biesecker, L.G., Bailey-Wilson, J.E., Ballantyne, J., Baum, H., Bieber, F.R., Brenner, C.H., Budowle, B., Butler, J.M., Carmody, G., *et al.* (2005). DNA identifications after the 9/11 World Trade Center attack. *Science, 310*(5751), 1122-1123.

Bilodeau, M. (2002). *Caractéristiques démogénétiques des populations de l'Abitibi et du Témiscamingue.* Mémoire de maitrise en médecine expérimentale, Université du Québec à Chicoutimi, Québec.

Bini, C., Ceccardi, S., Luiselli, D., Ferri, G., Pelotti, S., Colalongo, C., Falconi, M., et Pappalardo, G. (2003). Different informativeness of the three hypervariable mitochondrial DNA regions in the population of Bologna (Italy). *Forensic Science International, 135*(1), 48-52.

Bouchard, G., et de Braekeleer, M. (1991). *Histoire d'un génôme: population et génétique dans l'est du Québec*. Québec: Presses de l'Université du Québec.

Brenner, C.H. (2010). Fundamental problem of forensic mathematics—The evidential value of a rare haplotype. *Forensic Science International: Genetics, 4*(5), 281-291.

Buckleton, J.S., Krawczak, M., et Weir, B.S. (2011). The interpretation of lineage markers in forensic DNA testing. *Forensic Science International: Genetics, 5*(2), 78-83.

Buckleton, J.S., Taylor, D., Curran, J.M., et Bright, J.-A. (2016). Population Genetic Models. Dans Buckleton, J.S., Bright, J.-A. et Taylor, D. (sous la direction de), *Forensic DNA Evidence Interpretation* (2e éd., p. 87-117). Boca Raton, Florida: CRC Press.

Butler, J.M. (2005). *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers* (2e éd.). Amsterdam: Academic Press.

Butler, J.M. (2010). *Fundamentals of Forensic DNA Typing*. Amsterdam, Boston: Academic Press/Elsevier.

Butler, J.M. (2012). *Advanced Topics in Forensic DNA Typing: Methodology*. San Diego: Academic Press.

Butler, J.M. (2015). *Advanced Topics in Forensic DNA Typing: Interpretation*. San Diego, CA: Academic Press/Elsevier.

Butler, J.M., et Hill, C.R. (2013). Biology and Genetics of New Autosomal STR Loci Useful for Forensic DNA Analysis. Dans Shewale, J.G. et Liu, R.H. (sous la direction de), *Forensic DNA Analysis - Current Pratices and Emerging Technologies* (p. 181-198). Boca Raton, Floride: CRC Press.

Butler, J.M., Hill, C.R., Decker, A.E., Kline, M.C., Reid, T.M., et Vallone, P.M. (2007). *New autosomal and Y-chromosome STR loci: characterization and potential uses.* Communication présentée au 18e symposium international sur l'identification humaine.

Butler, J.M., Shen, Y., et McCord, B.R. (2003). The development of reduced size STR amplicons as tools for analysis of degraded DNA. *Journal of Forensic Sciences, 48*(5), 1054-1064.

Calafell, F., et Larmuseau, M.H.D. (2016). The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Human Genetics*, 1-15.

Carracedo, Á. (2015). Forensic Genetics: History. Dans Houck, M.M. (sous la direction de), *Forensic Biology* (p. 19-22). San Diego, CA: Academic Press.

Carracedo, Á., Bär, W., Lincoln, P.J., Mayr, W.R., Morling, N., Olaisen, B., Schneider, P.M., Budowle, B., Brinkmann, B., *et al.* (2000). DNA Commission of the International Society for Forensic Genetics: guidelines for mitochondrial DNA typing. *Forensic Science International, 110*, 79-85.

Centre of Forensic Sciences. (2017). *Improving the interpretation of complex DNA mixtures with probabilistic genotyping - A guide to STRmix™ for clients.* Accessible à l'adresse <https://www.mcscs.jus.gov.on.ca/sites/default/files/content/mcscs/docs/STRmix%20Client%20Guide%20-%20March%202017.pdf>.

Chang, Y.M., Burgoyne, L.A., et Both, K. (2003). Higher failures of amelogenin sex test in an Indian population group. *Journal of Forensic Sciences, 48*(6), 1309-1313.

Chinnery, P.F., Thorburn, D.R., Samuels, D.C., White, S.L., Dahl, H.-H.M., Turnbull, D.M., Lightowlers, R.N., et Howell, N. (2000). The inheritance of mitochondrial DNA heteroplasmy: random drift, selection or both? *Trends in Genetics, 16*(11), 500-505.

Churchill, J.D., Schmedes, S.E., King, J.L., et Budowle, B. (2016). Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling. *Forensic Science International: Genetics, 20*, 20-29.

Clopper, C.J., et Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika, 26*(4), 404-413.

Coble, M.D., Loreille, O.M., Wadhams, M.J., Edson, S.M., Maynard, K., Meyer, C.E., Niederstätter, H., Berger, C., Berger, B., *et al.* (2009). Mystery solved: the identification of the two missing Romanov children using DNA analysis. *PLoS ONE, 4*(3), e4838.

Cohen, D. (2017). *Assessing the use of DNA expert evidence, by justice system participants, in Ontario criminal courts.* Mémoire de maitrise en science, University of Ontario Institute of Technology, Ontario.

Coquoz, R., Comte, J., Hall, D., Hicks, T., et Taroni, F. (2013). *Preuve par l'ADN : la génétique au service de la justice*. Lausanne: Presses polytechniques et universitaires romandes.

Cree, L.M., Samuels, D.C., et Chinnery, P.F. (2009). The inheritance of pathogenic mitochondrial DNA mutations. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1792*(12), 1097-1102.

Crispino, F., et Houck, M.M. (2015). Principles of Forensic Science. Dans Houck, M.M. (sous la direction de), *Forensic Biology* (p. 1-5). San Diego, CA: Academic Press.

Daniels, D.L., Hall, A.M., et Ballantyne, J. (2004). SWGDAM developmental validation of a 19-locus Y-STR system for forensic casework. *Forensic Science International, 49*(4), 668-683.

Daoudi, Y., Morgan, M., Diefenbach, C., Ryan, J., Johnson, T., Conklin, G., Duncan, K., Smigielski, K., Huffine, E., *et al.* (1998). *Identification of the Vietnam tomb of the Unknown Soldier: the many roles of mitochondrial DNA.* Communication présentée au 9e symposium international sur l'identification humaine, Scottsdale.

Dekairelle, A., et Hoste, B. (2001). Application of a Y-STR-pentaplex PCR (DYS19, DYS389I and II, DYS390 and DYS393) to sexual assault cases. *Forensic Science International, 118*(2-3), 122-125.

Dubut, V., Chollet, L., Murail, P., Cartault, F., Béraud-Colomb, E., Serre, M., et Mogentale-Profizi, N. (2003). mtDNA polymorphisms in five French groups: importance of regional sampling. *European Journal of Human Genetics, 12*(4), 293.

Dumache, R., Ciocan, V., Muresan, C., et Enache, A. (2016). Molecular Genetics and its Applications in Forensic Sciences. Dans *Forensic Analysis-From Death to Justice* (p. 87-96): InTech.

Egeland, T., et Salas, A. (2008). Estimating haplotype frequency and coverage of databases. *PLoS ONE, 3*(12), e3988.

Flood, C.A. (2016). Legal Issues with Forensic DNA in the USA. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling* (p. 355-368). Royaume-Uni: Wiley.

Foreman, L.A., et Evett, I. (2001). Statistical analyses to support forensic interpretation for a new ten-locus STR profiling system. *International Journal of Legal Medicine, 114*(3), 147-155.

Foster, E.A., Jobling, M.A., Taylor, P.G., Donnelly, P., de Kniff, P., Miermet, R., Zerjal, T., et Tyler-Smith, C. (1998). Jefferson fathered slave's last child. *Nature, 396*, 27-28.

Fraser, J.C. (2010). What is forensic science? Dans *Forensic Science : A Very Short Introduction* (p. 1-6). Oxford: Oxford University Press.

Fu, Q., Posth, C., Hajdinjak, M., Petr, M., Mallick, S., Fernandes, D., Furtwängler, A., Haak, W., Meyer, M., *et al.* (2016). The genetic history of Ice Age Europe. *Nature, 534*(7606), 200-205.

Gagnon, A., et Heyer, E. (2001). Fragmentation of the Quebec population genetic pool (Canada): evidence from the genetic contribution of founders per region in the 17th and 18th centuries. *American Journal of Physical Anthropology, 114*(1), 30.

Gauvin, H., Lefebvre, J.-F., Moreau, C., Lavoie, È.-M., Labuda, D., Vézina, H., et Roy-Gagnon, M.H. (2015). GENLIB: an R package for the analysis of genealogical data. *BMC Bioinformatics, 16*(1), 160.

Gerstenberger, J., Hummel, S., Schultes, T., Häck, B., et 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*(4), 469-477.

Gill, P., Brenner, C.H., Brinkmann, B., Budowle, B., Carracedo, Á., Jobling, M.A., de Knijff, P., Kayser, M., Krawczak, M., *et al.* (2001). DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. *Forensic Science International, 124*, 5-10.

Gill, P., et Buckleton, J.S. (2005). Biological Basis for DNA Evidence. Dans Buckleton, J.S., Triggs, C.M. et Walsh, S.J. (sous la direction de), *Forensic DNA Evidence Interpretation*. Boca Raton: CRC Press.

Gill, P., Ivanov, P.L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I., Hagelberg, E., et Sullivan, K. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nature genetics, 6*(2), 130-135.

Gill, P., Jeffreys, A.J., et Werrett, D.J. (1985). Forensic application of DNA 'fingerprints'. *Nature, 318*, 577-579.

Goedbloed, M., Vermeulen, M., Fang, R., Lembring, M., Wollstein, A., Ballantyne, K.N., Lao, O., Brauer, S., Krüger, C., *et al.* (2009). Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpF*l*STR® Yfiler® PCR amplification kit. *International Journal of Legal Medicine, 123*(6), 471-482.

Gojanović, M.D., et Sutlović, D. (2007). Skeletal Remains from World War II Mass Grave: from Discovery to Identification. *Croatian medical journal, 48*(4), 520-527.

Goldstein, D.B., et Schlötterer, C. (1999). *Microsatellites - Evolution and Applications*. Oxford: Oxford University Press.

Graham, E.A.M. (2016). DNA: An Overview. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling* (p. 9-27). Royaume-Uni: Wiley.

Graves, J.A., Wakefield, M.J., et Toder, R. (1998). The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Human molecular genetics, 7*(13), 1991-1996.

Gusmão, L., Butler, J.M., Carracedo, Á., Gill, P., Kayser, M., Mayr, W.R., Morling, N., Prinz, M., Roewer, L., *et al.* (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. *Forensic Science International, 157*(2), 187-197.

Hall, D.W., et Byrd, J.H. (2012). *Forensic Botany : A Practical Guide*. Chichester, West Sussex, Angleterre: Wiley-Blackwell.

Hameed, I.H., Jebor, M.A., Ommer, A.J., Yoke, C., Zaidian, H.K., Al-Saadi, A.H., et Abdulazeez, M.A. (2014). Genetic variation and DNA markers in forensic analysis. *African Journal of Biotechnology, 13*(31).

Hanson, E.K., et Ballantyne, J. (2004). A highly discriminating 21 locus Y-STR "megaplex" system designed to augment the minimal haplotype loci for forensic casework. *Journal of Forensic Sciences, 49*(1), 40-51.

Hanson, E.K., Berdos, P.N., et Ballantyne, J. (2006). Testing and evaluation of 43 "noncore" Y chromosome markers for forensic casework applications. *Journal of Forensic Sciences, 51*(6), 1298-1314.

Harbison, S.-A. (2016). Sources of DNA. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling* (p. 59-68). Royaume-Uni: Wiley.

Helgason, A. (2001). *The ancestry and genetic history of the Icelanders: an analysis of mtDNA, Y chromosome haplotypes and genealogies.* Thèse de doctorat, University of Oxford, Angleterre.

Helgason, A., Hrafnkelsson, B., Gulcher, J.R., Ward, R., et 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. *The American Journal of Human Genetics, 72*(6), 1370-1388.

Helgason, A., Yngvadóttir, B., Hrafnkelsson, B., Gulcher, J.R., et Stefánsson, K. (2005). An Icelandic example of the impact of population structure on association studies. *Nature genetics, 37*(1), 90-95.

Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., et de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Human molecular genetics, 6*(5), 799.

Heyer, E., et Tremblay, M. (1995). Variability of the genetic contribution of Quebec population founders associated to some deleterious genes. *The American Journal of Human Genetics, 56*(4), 970-978.

Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J., et Labuda, D. (2001). Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *The American Journal of Human Genetics, 69*(5), 1113-1126.

Holland, M.M., Fisher, D.L., Mitchell, L.G., Rodriquez, W.C., Canik, J.J., Merril, C.R., et Weedn, V.W. (1993). Mitochondrial DNA sequence analysis of human skeletal remains: identification of remains from the Vietnam War. *Journal of Forensic Sciences, 38*(3), 542-553.

Holland, M.M., et Lauc, G. (2014). Forensic Aspects of mtDNA Analysis. Dans Primorac, D. and Schanfield, M.S. (sous la direction de), *Forensic DNA Applications: An Interdisciplinary Perspective* (p. 85-104). Boca Raton, Floride: CRC Press.

Holt, I.J., et Reyes, A. (2012). Human mitochondrial DNA replication. *Cold Spring Harbour Perspectives in Biology, 4*(12).

Houck, M.M., et Siegel, J.A. (2015). *Fundamentals of Forensic Science* (3e éd.). San Diego, CA: Academic Press.

INTERPOL DNA Unit. (2009). *INTERPOL global DNA profiling survey - Results and analysis*. Accessible à l'adresse < http://www.dnaresource.com/documents/2008INTERPOLGLOBALDNASURVEYREPORTV2.pdf>.

Irwin, J.A., Saunier, J.L., Niederstätter, H., Strouss, K.M., Sturk, K.A., Diegoli, T.M., Brandstätter, A., Parson, W., et Parsons, T.J. (2009). Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples. *Journal of molecular evolution, 68*(5), 516-527.

Jamieson, A. (2016). Interpretation of Mixtures; Graphical. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling* (p. 119-131). Royaume-Uni: Wiley.

Jeffreys, A.J., Brookfield, J.F.Y., et Semeonoff, R. (1985a). Positive identification of an immigration test-case using human DNA fingerprints. *Nature, 317*, 818-819.

Jeffreys, A.J., Wilson, V., et Thein, S.L. (1985b). Hypervariable 'minisatellite' regions in human DNA. *Nature, 314*, 67-73.

Jeffreys, A.J., Wilson, V., et Thein, S.L. (1985c). Individual-specific 'fingerprints' of human DNA. *Nature, 316*(6023), 76-79.

Jobling, M.A., et Gill, P. (2004). Encoded evidence: DNA in forensic analysis. *Nature Reviews Genetics, 5*(10), 739-752.

Just, R.S., Loreille, O.M., Molto, J.E., Merriwether, D.A., Woodward, S.R., Matheson, C., Creed, J., McGrath, S.E., Sturk-Andreaggi, K., *et al.* (2011). *Titanic*'s unknown child: The critical role of the mitochondrial DNA coding region in a re-identification effort. *Forensic Science International: Genetics, 5*(3), 231-235.

Just, R.S., Moreno, L.I., Smerick, J.B., et Irwin, J.A. (2017). Performance and concordance of the ForenSeq™ system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens. *Forensic Science International: Genetics, 28*, 1-9.

Kayser, M. (2017). Forensic use of Y-chromosome DNA: a general overview. *Human Genetics*, 1-15.

Kayser, M., et Ballantyne, K.N. (2014). Y Chromosome in Forensic Science. Dans Primorac, D. and Schanfield, M.S. (sous la direction de), *Forensic DNA Applications: An Interdisciplinary Perspective* (p. 105-134). Boca Raton, Floride: CRC Press.

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., *et al.* (1997). Evaluation of Y-chromosomal STRs: a multicenter study. *International Journal of Legal Medicine, 110*(3), 125-133.

Kayser, M., et de Knijff, P. (2011). Improving human forensics through advances in genetics, genomics and molecular biology. *Nature Reviews Genetics, 12*(3), 179-192.

Kayser, M., et Sajantila, A. (2001). Mutations at Y-STR loci: implications for paternity testing and forensic analysis. *Forensic Science International, 118*(2), 116-121.

Kong, A., Thorleifsson, G., Frigge, M.L., Vilhjalmsson, B.J., Young, A.I., Thorgeirsson, T.E., Benonisdottir, S., Oddsson, A., Halldorsson, B.V., *et al.* (2018). The nature of nurture: effects of parental genotypes. *Science, 359*(6374), 424-428.

Krane, D.E., Allen, R.W., Sawyer, S.A., Petrov, D.A., et Hartl, D.L. (1992). Genetic differences at four DNA typing loci in Finnish, Italian, and mixed Caucasian populations. *Proceedings of the National Academy of Sciences, 89*(22), 10583-10587.

Krawczak, M. (2001). Forensic evaluation of Y-STR haplotype matches: a comment. *Forensic Science International, 118*(2), 114-115.

Kurosaki, K., Matsushita, T., et Ueda, S. (1993). Individual DNA identification from ancient human remains. *The American Journal of Human Genetics, 53*(3), 638-643.

Laberge, A.-M., Jomphe, M., Houde, L., Vézina, H., Tremblay, M., Desjardins, B., Labuda, D., St-Hilaire, M., Macmillan, C., *et al.* (2005). A "Fille du Roy" introduced the T14484C Leber hereditary optic neuropathy mutation in French Canadians. *The American Journal of Human Genetics, 77*(2), 313-317.

Ladika, S. (2005). DNA helps identify missing in the tsunami zone. *Science, 307,* 504.

Larmuseau, M.H.D., Ottoni, C., Raeymaekers, J.A.M., Vanderheyden, N., Larmuseau, H.F.M., et Decorte, R. (2012a). Temporal differentiation across a West-European Y-chromosomal cline: genealogy as a tool in human population genetics. *European Journal of Human Genetics, 20*(4), 434-440.

Larmuseau, M.H.D., Vanderheyden, N., Van Geystelen, A., van Oven, M., de Knijff, P., et Decorte, R. (2014). Recent radiation within Y-chromosomal haplogroup R-M269 resulted in high Y-STR haplotype resemblance. *Annals of Human Genetics, 78*(2), 92-103.

Larmuseau, M.H.D., Vanoverbeke, J., Gielis, G., Vanderheyden, N., Larmuseau, H.F.M., et Decorte, R. (2012b). In the name of the migrant father—Analysis of surname origins identifies genetic admixture events undetectable from genealogical records. *Heredity, 109*(2), 90-95.

Larmuseau, M.H.D., Vanoverbeke, J., Van Geystelen, A., Defraene, G., Vanderheyden, N., Matthys, K., Wenseleers, T., et Decorte, R. (2013). Low historical rates of cuckoldry in a Western European human population traced by Y-chromosome and genealogical data. *Proceedings of the Royal Society B: Biological Sciences, 280*(1772), 20132400.

Leake, S.L. (2013). Is human DNA enough? - potential for bacterial DNA. *Frontiers in Genetics, 4*.

Lieberman, J.D., Carrell, C.A., Miethe, T.D., et Krauss, D.A. (2008). Gold versus platinum: do jurors recognize the superiority and limitations of DNA evidence compared to other types of forensic evidence? *Psychology, Public Policy, and Law, 14*(1), 27-62.

Linacre, A., et Tobe, S.S. (2013). *Wildlife DNA Analysis : Applications in Forensic Science*. Chichester, West Sussex, UK: John Wiley & Sons Inc.

Lippold, S., Xu, H., Ko, A., Li, M., Renaud, G., Butthof, A., Schröder, R., et Stoneking, M. (2014). Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investigative genetics, 5*, 13-28.

Lumbroso, S. (2016). Effet fondateur : l'héritage des pionniers (7/8). *Québec Science*. Consulté le 4 avril 2018.

Lutz, S., Wittig, H., Weisser, H.-J., Heizmann, J., Junge, A., Dimo-Simonin, N., Parson, W., Edelmann, J., Anslinger, K., *et al.* (2000). Is it possible to differentiate mtDNA by means of HVIII in samples that cannot be distinguished by sequencing the HVI and HVII regions? *Forensic Science International, 113*(1-3), 97-101.

Lynch, M. (2003). God's signature: DNA profiling, the new gold standard in forensic science. *Endeavour, 27*(2), 93-97.

Margot, P. (2011). La trace comme vecteur fondamental de la police scientifique. Dans Ricordel, Y. (sous la direction de), *L'expertise en police scientifique*. SA, Paris: Xavier Montauban.

Martín, P., Albarrán, C., Garcia, O., Garcia, P., Sancho, M., et Alonso, A. (2000). Application of Y-STR analysis to rape cases that cannot be solved by autosomal STR analysis. *Progress in Forensic Genetics, 8*, 526-528.

Melton, T. (2016). Mitochondrial DNA: Profiling. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling* (p. 245-250). Royaume-Uni: Wiley.

Metcalf, J.L., Xu, Z.Z., Bouslimani, A., Dorrestein, P., Carter, D.O., et Knight, R. (2017). Microbiome tools for forensic science. *Trends in Biotechnology, 35*(9), 814-823.

Milne, R. (2013). *Forensic Intelligence*. Boca Raton, Floride: CRC Press.

Milot, E. Laboratoire de génétique des populations - Science forensique. Accessible à l'adresse < https://oraprdnt.uqtr.uquebec.ca/pls/public/gscw031?owa_no_site=4214&owa_no_fiche=8>. Consulté le 18 janvier 2018.

Milot, E., Mayer, F.M., Nussey, D.H., Boisvert, M., Pelletier, F., et Réale, D. (2011). Evidence for evolution in response to natural selection in a contemporary human population. *Proceedings of the National Academy of Sciences, 108*(41), 17040-17045.

Milot, E., Moreau, C., Gagnon, A., Cohen, A.A., Brais, B., et Labuda, D. (2017). Mother's curse neutralizes natural selection against a human genetic disease over three centuries. *Nature Ecology & Evolution, 1*(9), 1400.

Ministère de la Sécurité publique. (2015). *Rapport annuel 2014-2015*. Accessible à l'adresse < https://www.securitepublique.gouv.qc.ca/fileadmin/Documents/laboratoire/rapport_annuel/2014-2015.pdf>.

Moreau, C., Bhérer, C., Vézina, H., Jomphe, M., Labuda, D., et Excoffier, L. (2011a). Deep human genealogies reveal a selective advantage to be on an expanding wave front. *Science, 334*(6059), 1148-1150.

Moreau, C., Lefebvre, J.-F., Jomphe, M., Bhérer, C., Ruiz-Linares, A., Vézina, H., Roy-Gagnon, M.H., et Labuda, D. (2013). Native american admixture in the Quebec founder population. *PLoS ONE, 8*(6), 1-9.

Moreau, C., Vézina, H., Jomphe, M., Lavoie, È.-M., Roy-Gagnon, M.H., et Labuda, D. (2011b). When genetics and genealogies tell different stories-Maternal lineages in Gaspesia. *Annals of Human Genetics, 75*(2), 247-254.

Moreau, C., Vézina, H., et Labuda, D. (2007). Effets fondateurs et variabilité génétique au Québec. *médecine/sciences, 23*(11), 1008-1013.

Moreau, C., Vézina, H., Yotova, V., Hamon, R., de Knijff, P., Sinnett, D., et Labuda, D. (2009). Genetic heterogeneity in regional populations of Quebec—Parental lineages in the Gaspe Peninsula. *American Journal of Physical Anthropology, 139*(4), 512-522.

Mourali-Chebil, S., et Heyer, E. (2006). Evolution of inbreeding coefficients and effective size in the population of Saguenay Lac-St.-Jean (Québec). *Human Biology, 78*(4), 495-508.

National Research Council. (1996). *The Evaluation of Forensic DNA Evidence*. Washington, D.C.: National Academy Press.

Olofsson, J., Mogensen, H.S., Hjort, B.B., et Morling, N. (2011). Evaluation of Y-STR analyses of sperm cell negative vaginal samples. *Forensic Science International: Genetics Supplement Series, 3*(1), e141-e142.

Palo, J.U., Hedman, M., Söderholm, N., et Sajantila, A. (2007). Repatriation and identification of Finnish World War II soldiers. *Croatian medical journal, 48*(4), 528-535.

Palo, J.U., Pirttimaa, M., Bengs, A., Johnsson, V., Ulmanen, I., Lukka, M., Udd, B., et Sajantila, A. (2008). The effect of number of loci on geographical structuring and forensic applicability of Y-STR data in Finland. *International Journal of Legal Medicine, 122*(6), 449-456.

Parr, R.L., Maki, J., Reguly, B., Dakubo, G.D., Aguirre, A., Wittock, R., Robinson, K., Jakupciak, J.P., et Thayer, R.E. (2006). The pseudo-mitochondrial genome influences mistakes in heteroplasmy interpretation. *BMC Genomics, 7*(1), 185.

Parson, W. (2015). Mitochondrial DNA. Dans Houck, M.M., (sous la direction de), *Forensic Biology* (p. 117-125). San Diego, CA: Academic Press.

Parson, W., et Dür, A. (2007). EMPOP—A forensic mtDNA database. *Forensic Science International: Genetics, 1*(2), 88-92.

Parson, W., Gusmão, L., Hares, D., Irwin, J.A., Mayr, W.R., Morling, N., Pokorak, E., Prinz, M., Salas, A., *et al.* (2014). DNA Commission of the International Society for Forensic Genetics: revised and extended guidelines for mitochondrial DNA typing. *Forensic Science International: Genetics, 13*, 134-142.

Parsons, T.J., et Coble, M.D. (2001). Increasing the forensic discrimination of mitochondrial DNA testing through analysis of the entire mitochondrial DNA genome. *Croatian medical journal, 42*(3), 304-309.

Pausova, Z., Jomphe, M., Houde, L., Vézina, H., Orlov, S.N., Gossard, F., Gaudet, D., Tremblay, J., Kotchen, T.A., *et al.* (2002). A genealogical study of essential hypertension with and without obesity in French Canadians. *Obesity, 10*(6), 463-470.

Piercy, R., Sullivan, K., Benson, N., et Gill, P. (1993). The application of mitochondrial DNA typing to the study of white Caucasian genetic identification. *International Journal of Legal Medicine, 106*(2), 85-90.

Prost, S., et Anderson, C.N.K. (2011). TempNet: a method to display statistical parsimony networks for heterochronous DNA sequence data. *Methods in Ecology and Evolution, 2*(6), 663-667.

R Development Core Team. (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Accessible à l'adresse < http://www.R-project.org>.

Ramos, A., Santos, C., Alvarez, L., Nogués, R., et Aluja, M.P. (2009). Human mitochondrial DNA complete amplification and sequencing: A new validated primer set that prevents nuclear DNA sequences of mitochondrial origin co-amplification. *Electrophoresis, 30*(9), 1587-1593.

Ribaux, O., et Margot, P. Dictionnaire de Criminologie en ligne. Accessible à l'adresse < http://www.criminologie.com/article/science-forensique>.

Ribaux, O., Walsh, S.J., et Margot, P. (2006). The contribution of forensic science to crime analysis and investigation: forensic intelligence. *Forensic Science International, 156*(2-3), 171-181.

Roewer, L., Croucher, P.J., Willuweit, S., Lu, T.T., Kayser, M., Lessig, R., de Knijff, P., Jobling, M.A., Tyler-Smith, C., *et al.* (2005). Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. *Human Genetics, 116*(4), 279-291.

Roewer, L., Kayser, M., de Knijff, P., Anslinger, K., Betz, A., Caglia, A., Corach, D., Füredi, S., Henke, L., *et al.* (2000). A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Science International, 114*(1), 31-43.

Rolf, B., Keil, W., Brinkmann, B., Roewer, L., et Fimmers, R. (2001). Paternity testing using Y-STR haplotypes: assigning a probability for paternity in cases of mutations. *International Journal of Legal Medicine, 115*(1), 12-15.

Rosenberg, T., Nørby, S., Schwartz, M., Saillard, J., Magalhaes, P.J., Leroy, D., Kann, E.C., et Duno, M. (2016). Prevalence and genetics of Leber hereditary optic neuropathy in the Danish population. *Investigative ophthalmology & visual science, 57*(3), 1370-1375.

Roy-Gagnon, M.H., Moreau, C., Bhérer, C., St-Onge, P., Sinnett, D., Laprise, C., Vézina, H., et Labuda, D. (2011). Genomic and genealogical investigation of the French Canadian founder population structure. *Human Genetics, 129*(5), 521-531.

Roy, J. (2013). Lac-Mégantic - Technologie avancée pour identifier les huit dernières victimes. *Le Journal de Québec*. Accessible à l'adresse <http://www.journaldequebec.com/2013/10/28/technologie-avancee-pour-identifier-les-huit-dernieres-victimes>. Consulté le 4 avril 2018.

s.a. EMPOP mtDNA database, v3/R11. Accessible à l'adresse < https://empop.online>. Consulté le 10 février 2018.

s.a. (2018). US Y-STR. Accessible à l'adresse < https://www.usystrdatabase.org/>. Consulté le 10 février 2018.

Sajantila, A., Salem, A.-H., Savolainen, P., Bauer, K., Gierig, C., et Paabo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proceedings of the National Academy of Sciences, 93*(21), 12035-12039.

Schanfield, M.S., Primorac, D., et Marjanovic, D. (2014a). Basic Genetics and Human Genetic Variation. Dans Primorac, D. and Schanfield, M.S. (sous la direction de), *Forensic DNA Applications: An Interdisciplinary Perspective* (p. 3-54). Boca Raton, Floride: CRC Press.

Schanfield, M.S., Primorac, D., et Marjanovic, D. (2014b). Forensic DNA Analysis and Statistics. Dans Primorac, D. and Schanfield, M.S. (sous la direction de), *Forensic DNA Applications: An Interdisciplinary Perspective* (p. 55-84). Boca Raton, Floride: CRC Press.

Schwartz, M., et Vissing, J. (2002). Paternal inheritance of mitochondrial DNA. *New England Journal of Medicine, 347*(8), 576-580.

Scriver, C.R. (2001). Human genetics: lessons from Quebec populations. *Annual review of genomics and human genetics, 2*(1), 69-101.

Sigurðardóttir, S., Helgason, A., Gulcher, J.R., Stefánsson, K., et Donnelly, P. (2000). The mutation rate in the human mtDNA control region. *The American Journal of Human Genetics, 66*(5), 1599-1609.

Steinlechner, M., Berger, B., Niederstätter, H., et Parson, W. (2002). Rare failures in the amelogenin sex test. *International Journal of Legal Medicine, 116*(2), 117-120.

Sullivan, K., Hopgood, R., et Gill, P. (1992). Identification of human remains by amplification and automated sequencing of mitochondrial DNA. *International Journal of Legal Medicine, 105*(2), 83-86.

SWGDAM. (2003). Guidelines for mitochondrial DNA (mtDNA) nucleotide sequence interpretation. *Forensic Science Communications, 5*(2).

SWGDAM. (2009). Y-chromosome short tandem repeat (Y-STR) interpretation guidelines. *Forensic Science Communications, 11*(1).

SWGDAM. (2013). *Interpretation guidelines for mitochondrial DNA analysis by forensic dna testing laboratories*. Accessible à l'adresse <http://media.wix.com/ugd/4344b0_c5e20877c02f403c9ba16770e8d41937.pdf>.

SWGDAM. (2014). *Interpretation guidelines for Y-chromosome STR typing*. Accessible à l'adresse <http://media.wix.com/ugd/4344b0_c5e20877c02f403c9ba16770e8d41937.pdf>.

SWGDAM. (2017). *Interpretation guidelines for autosomal STR typing by forensic dna testing laboratories*. Accessible à l'adresse <https://docs.wixstatic.com/ugd/4344b0_50e2749756a242528e6285a5bb478f4c.pdf>.

SWGDAM Y-STR Subcommittee. (2007). Report on the current activities of the Scientific Working Group on DNA Analysis Methods Y-STR Subcommittee. *Forensic Science Communications, 6*, 1-2.

Szabolcsi, Z., Farkas, Z., Borbély, A., Bárány, G., Varga, D., Heinrich, A., Völgyi, A., et Pamjav, H. (2015). Statistical and population genetics issues of two Hungarian

datasets from the aspect of DNA evidence interpretation. *Forensic Science International: Genetics, 19*, 18-21.

Templeton, A.R. (2006). *Population Genetics and Microevolutionary Theory*. Hoboken, N.J.: Wiley-Liss.

Tremblay, M., et Vézina, H. (2000). New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *The American Journal of Human Genetics, 66*(2), 651-658.

Tremblay, M., et Vézina, H. (2010). Genealogical analysis of maternal and paternal lineages in the Quebec population. *Human Biology, 82*(2), 179-198.

Tully, G., Bär, W., Brinkmann, B., Carracedo, Á., Gill, P., Morling, N., Parson, W., et Schneider, P.M. (2001). Considerations by the European DNA profiling (EDNAP) group on the working practices, nomenclature and interpretation of mitochondrial DNA profiles. *Forensic Science International, 124*(1), 83-91.

Underhill, P.A., et Kivisild, T. (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics, 41*, 539-564.

Vézina, H., Heyer, E., Fortier, I., Ouellette, G., Robitaille, Y., et Gauvreau, D. (1999). A genealogical study of alzheimer disease in the Saguenay region of Quebec. *Genetic epidemiology, 16*(4), 412-425.

Vézina, H., Jomphe, M., Lavoie, È.-M., Moreau, C., et Labuda, D. (2012). L'apport des données génétiques à la mesure généalogique des origines amérindiennes des Canadiens français. *Cahiers québécois de démographie, 41*(1), 87-105.

Vézina, H., Tremblay, M., Desjardins, B., et Houde, L. (2005). Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. *Cahiers québécois de démographie, 34*(2), 235-258.

Vézina, H., Tremblay, M., et Houde, L. (2004). Mesures de l'apparentement biologique au Saguenay-Lac-St-Jean (Québec, Canada) à partir de reconstitutions généalogiques. *Annales de démographie historique*(2), 67-83.

Wallace, D.C., Brown, M.D., et Lott, M.T. (1999). Mitochondrial DNA variation in human evolution and disease. *Gene, 238*(1), 211-230.

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*(2), 897-912.

Walsh, S.J. (2016). DNA. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling* (p. 30-36). Royaume-Uni: Wiley.

Weber, J.L., et May, P.E. (1989). Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. *The American Journal of Human Genetics, 44*(3), 388-396.

Weir, B.S. (2001). DNA match and profile probabilities: comment on Budowle et al. (2000) and Fung and Hu (2000). *Forensic Science Communications, 3*(1).

Weir, B.S., et Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. *Evolution, 38*(6), 1358-1370.

Weir, B.S., et Hill, W.G. (2002). Estimating F-statistics. *Annual Review of Genetics*, 721.

Wetton, J.H., Tsang, K.W., et Khan, H. (2005). Inferring the population of origin of DNA evidence within the UK by allele-specific hybridization of Y-SNPs. *Forensic Science International, 152*(1), 45-53.

Wiegand, P., et Kleiber, M. (2001). Less is more–length reduction of STR amplicons using redesigned primers. *International Journal of Legal Medicine, 114*(4-5), 285-287.

Willuweit, S., et Roewer, L. (2015). The new Y chromosome haplotype reference database. *Forensic Science International: Genetics, 15*, 43-48.

Willuweit, S., et Roewer, L. (2018a). Y-Chromosome STR haplotype reference database, R56. Accessible à l'adresse < https://yhrd.org/>. Consulté le 10 février 2018.

Willuweit, S., et Roewer, L. (2018b). Y-chromosome STR haplotype reference database, R58. Accessible à l'adresse < https://yhrd.org/>. Consulté le 26 septembre 2018.

Wonnapinij, P., Chinnery, P.F., et Samuels, D.C. (2008). The distribution of mitochondrial DNA heteroplasmy due to random genetic drift. *The American Journal of Human Genetics, 83*(5), 582-593.

Wright, S. (1951). The genetical structure of populations. *Annals of eugenics, 15*(1), 323-354.

Xue, Y., Wang, Q., Long, Q., Ng, B.L., Swerdlow, H., Burton, J., Skuce, C., Taylor, R., Abdellah, Z., *et al.* (2009). Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology, 19*(17), 1453-1457.

Yao, Y.-G., Kong, Q.-P., Salas, A., et Bandelt, H.-J. (2008). Pseudomitochondrial genome haunts disease studies. *Journal of medical genetics, 45*(12), 769-772.

Zietkiewicz, E., Witt, M., Daca, P., Żebracka-Gala, J., Goniewicz, M., Jarząb, B., et Witt, M. (2012). Current genetic methodologies in the identification of disaster victims and in forensic analysis. *Journal of applied genetics, 53*(1), 41-60.

ANNEXE A

CALCUL DE LA DISTANCE GÉNÉTIQUE POUR CHAQUE PAIRE D'INDIVIDUS GÉNOTYPÉS POUR L'ADN MITOCHONDRIAL

La procédure pour l'ADN mitochondrial et le chromosome Y est très similaire. Les endroits où des modifications doivent être apportées pour le chromosome Y sont indiqués.

```
###### 1. Preparation of R ######
# Determine the working directory
setwd("folder path")

# Import data
#Genealogies from BALSAC
#Genealogies should be imported as a data frame in the
variable "Balsac"
#Columns should be in order: individual, father, mother,
sex, marriage year of parents, place of marriage of
parents, marriage year of the individual and place of
marriage of the individual

#Molecular data
#mtDNA data should be imported as a data frame in the
variable "mtDNA_HT"
#Columns should be in order: individual, DNA sample code,
lineage number and haplotype

#We keep only genotyped individuals
mtDNA_HT <- mtDNA_HT[which(nchar(mtDNA_HT$mtDNA)!=0),]

###### 2. Prepare list of pairs of individuals ######
#IDs
ind_ID <- mtDNA_HT[,"ind"]
save(ind_ID, file="ind_ID_mito.RData")
#To have all possible combinations of 2 individuals in a
data frame
pairs_ind <- combn(ind_ID,2)
```

```
pairs_ind <- as.data.frame(t(pairs_ind))
colnames(pairs_ind) <- c("ind1", "ind2")

#Save list of pairs of individuals
list_pairs_mito <- vector(mode="list")
for(i in 1:nrow(pairs_ind)){
  list_pairs_mito[i] <-
list(c(pairs_ind[i,"ind1"],pairs_ind[i,"ind2"]))
  names(list_pairs_mito)[i] <-
paste("pair",pairs_ind[i,"ind1"],pairs_ind[i,"ind2"],sep =
"_")
}
save(list_pairs_mito, file="list_pairs_mito.RData")

###### 3. Prepare genealogy with maternal lineages ######
library("GENLIB")

#I have to create a date frame with ind ID, father, mother,
sex
#to create a GLgen object
pedigree <- Balsac[,c("ind","father","mother","sex")]
ped_mito <- gen.lineages(pedigree,pro=ind_ID,
maternal=TRUE)
#Maternal = FALSE for paternal lineages for Y chromosome
save(ped_mito, file="ped_mito.RData")

###### 4. Function to find the most recent common ancestor
######
#Originally from GENLIB but modified by Éric Giguère
(Calcul Québec) to optimize the speed

gen.findMRCA_fast <- function (gen, individuals, NbProcess
= parallel::detectCores() -1)
{
  retour = tryCatch({
    NbProcess = min(length(individuals), NbProcess)
    cluster = makePSOCKcluster(NbProcess)
    registerDoParallel(cluster)
    clusterExport(cluster, c("gen.branching"))
    n = length(individuals)
    branch = foreach(i = 1:n) %dopar% {
```

```
      gen.branching(gen, individuals[i])
    }
    clusterExport(cluster, c("gen.getAncestorsPAR_branch"))
    x = foreach(i = 1:n) %dopar% {
      gen.getAncestorsPAR_branch(gen, branch[[i]])
    }
    indInter <- c()
    if (length(x) > 1) {
      indInter <- intersect(x[[1]], x[[2]])
      if (length(x) > 2)
        z <- lapply(c(3:length(x)), function(i) {
          indInter <<- intersect(indInter, x[[i]])
        })
    }
    else indInter <- x[[1]]
    b_1 <- gen.genout(gen.branching(gen, individuals[1]))
    inter <- subset(b_1, b_1$ind %in% indInter, T)
    inter <- subset(inter, !(inter$ind %in% inter$father |
             inter$ind %in% inter$mother), T)
    print(paste(dim(inter)[1], "MRCA"))
    clusterExport(cluster, c("gen.climbPAR_branch"))
    y <- foreach(i = 1:length(inter$ind)) %dopar%
gen.climbPAR_branch(gen = gen,
branch = branch, n = n, founder = inter$ind[i])
    x <- matrix(nrow = length(individuals), ncol =
length(inter$ind))
    for (i in 1:length(inter$ind)) {
      message <- y[[i]]
      f <- message$founder
      d <- message$distance
      x[, i] <- d
    }
    colnames(x) <- inter$ind
    rownames(x) <- individuals
    x
  }, warning = function(w) {
    print("warning")
    message(w)
    return(NULL)
  }, error = function(e) {
    print("error")
```

```
    message(e)
    return(NULL)
  }, finally = {
    stopCluster(cluster)
  })
  return(retour)
}

###### 5. Function to calculate the minimum genetic
distance ######
#example with a list of pairs of individuals named
pairs_mito_T1
library(stringr)
library(parallel)
no_cores <- 12
cl <- makeCluster(no_cores, outfile="")

a <- vector(mode = "list")
aa <- vector(mode = "list")
clusterExport(cl,
c("a","aa","pairs_mito_T1","ped_mito","Balsac"))
clusterEvalQ(cl, c(library(GENLIB), library(stringr)))

find_MRCA <- function(data){
  reponse <- tryCatch({
  mat <<- GENLIB::gen.findMRCA_fast(ped_mito, data,
NbProcess=2)
  },
  error=function(e){

write(paste(data,e[[1]]),file="messages_erreurs_fast12x2.tx
t",append=T)
  mat <<- NULL
  })

  if(is.null(mat) == FALSE){ #in cases where there's no
MRCA, I want the function to continue even if there's an
"error"
    MRCAs <- as.numeric(colnames(mat))
    MRCA_woman <- Balsac[Balsac$ind %in% MRCAs & Balsac$sex
== 2,1]
```

```
#for paternal lineages, put MRCA_man <- Balsac[Balsac$ind
%in% MRCAs & #Balsac$sex == 1,1]

    mat_meiosis <- GENLIB::gen.find.Min.Distance.MRCA(mat)
    mat_meiosis <- as.data.frame(mat_meiosis)
    mat_meiosis_dist <- mat_meiosis[mat_meiosis$founder
%in% MRCA_woman,4]
#for paternal lineages, put mat_meiosis_dist <-
#mat_meiosis[mat_meiosis$founder %in% MRCA_man,4]
    mat_meiosis_dist <- sort(unique(mat_meiosis_dist))
    dist <- stringr::str_c(mat_meiosis_dist, collapse=",")

    return(dist)}

  else{NA}
}

##to apply this function to pairs of individuals with
parlapply
#pairs were separated in multiple files to do multiple
tasks simultaneously

for (i in 1:ceiling(length(pairs_mito_T1)/2000)){
  n.lines <- (1:2000) + 2000*(i-1)
  n.lines <- n.lines[n.lines<=length(pairs_mito_T1)]
  list_subset <- pairs_mito_T1[n.lines]
  aa <- parLapply(cl, list_subset,find_MRCA)
  a <- c(a,aa)
  filename <- paste("output_mito_T1_fast12x2","_boucle",
i,".RData", sep="")
  save(a, file = filename)
}

stopCluster(cl)

###### 6. Compile results ######
#Check each case where there is an error message
#Transform into a data frame with columns individual 1,
individual 2 and genetic distance
    # Combine all tasks in a single data frame
```

ANNEXE B

CALCUL DU TAUX DE NON-CONCORDANCES GLOBAL ET DU TAUX DE NON-CONCORDANCES DUES AUX ERREURS GÉNÉALOGIQUES POUR LE CHROMOSOME Y

La procédure pour l'ADN mitochondrial et le chromosome Y est très similaire. Les endroits où des modifications doivent être apportées pour l'ADN mitochondrial sont indiqués.

```
###### 1. Preparation of R ######
# Determine the working directory
setwd("folder path")

# Import data
#Genealogies should be imported as a data frame in the
variable "yped"
#Columns should be in order: individual, father, mother,
sex and lineage number

#Genetic distance between each pair of individuals should
be imported in the variable "gen_dist" for Y chromosome and
mitochondrial DNA
#Columns should be in order: individual 1, individual 2,
"individual 1_individual 2" and genetic distance

#Those without common ancestor are not needed
#We keep only pairs separated by at least 7 meiosis
gen_dist2 <- gen_dist[!is.na(gen_dist$Gen_Dist),]
gen_dist2 <- gen_dist2[which(gen_dist2$Gen_Dist>6),]
rownames(gen_dist2)<-NULL

###### 2. List of meiosis ######
#To have the list of individuals of interest
ind.list <- unique(unlist(gen_dist2[,c(1,2)]))
```

```
#Function to have the list of ancestors and a vector of all
meiosis
meiosis <- character()

list.meiosis <- function(pedigree, lineage.type = "father",
ind){
    while (!pedigree[pedigree[,1]==ind,lineage.type] ==0) {
    ancestor <- pedigree[pedigree[,1]==ind,lineage.type]
    a <- paste(ind,ancestor,sep="_")
    meiosis <- c(meiosis,a)
    ind=ancestor
  }

  paste(meiosis,collapse=";")
}

#Function to have the meiosis separating individuals from a
pair

list.meiosis.pairs <-
function(pedigree,lineage.type="father", individuals){
  inds <- as.numeric(unlist(strsplit(individuals,
split="_")))

  a <- list.meiosis(pedigree, ind=inds[1])
  a <- unlist(strsplit(a, split=";"))
  b <- list.meiosis(pedigree, ind=inds[2])
  b <- unlist(strsplit(b, split=";"))

  common.meiosis <- intersect(a,b)

    if(length(common.meiosis)>0){
      final.meiosis <- common.meiosis[1]

      aa <- a[-c(which(a==final.meiosis):length(a))]
      bb <- b[-c(which(b==final.meiosis):length(b))]
    } else {
      aa <- a
      bb <- b
    }
```

```
  list.meiosis <- paste(c(aa,bb), collapse=";")
  list.meiosis
}

#Add the list of meiosis to gen_dist2 table
gen_dist2$meiosis <- NA
meiosis <- character()

for(i in 1:nrow(gen_dist2)){
  individuals <- gen_dist2[i,"IND1_IND2"]

#Below, lineage.type must be specified as "mother" for
mitochondrial DNA
  gen_dist2[i,"meiosis"] <- list.meiosis.pairs(yped,
individuals=individuals)
}

save(gen_dist2, file="gendist_meiosis_table.RData")
#For the mitochondrial DNA, pairs with degenerated
nucleotides were removed from "gen_dist2"

###### 3. Function to select independant meiosis for EPP
analysis ######

select.pairs <- function(pair.table=gen.dist2){

pairs<-sample(pair.table$IND1_IND2,size=1,replace=F)
p1.meiosis <- pair.table[pair.table$IND1_IND2 == pairs,
"meiosis"]
p1.meiosis <- unlist(strsplit(p1.meiosis,split=";"))

for(i in 1:nrow(pair.table)){

  p2.meiosis <- pair.table[i, "meiosis"]
  p2.meiosis <- unlist(strsplit(p2.meiosis,split=";"))

  other.meiosis <- pair.table[pair.table$IND1_IND2 %in%
pairs, "meiosis"]
  other.meiosis <-
unlist(strsplit(other.meiosis,split=";"))
```

```
  all.meiosis <- c(other.meiosis,p2.meiosis)

  duplic <- sum(duplicated(all.meiosis))

  if(duplic==0){
    pairs <- c(pairs, pair.table[i, "IND1_IND2"])
  }
}

pairs
}

###### 4. Functions to select pairs and prepare data ######
#Preparation of molecular data
#Molecular data should be imported as a data frame in the
variable "Ychrgen"
#Columns should be in order: individual and the list of Y-
STR considered

##Function to select the Y-STR kit desired
find.YSTR.kit <- function(dat,kit=c(7,11,17,21,27)){
  if(kit==7){
    Ychr_HT <<- dat[ , c(1,3,5,9,12,14,16,23,30)]
  }
  if(kit==11){
    Ychr_HT <<- dat[ ,
c(1,3,5,9,12,14,15,16,19,20,21,23,24,30)]
  }
  if(kit==17){
    Ychr_HT <<- dat[ ,
c(1,3,4,5,8,9,10,11,12,13,14,15,16,19,20,21,23,24,30)]
  }
  if(kit==21){ #without RM-STR
    Ychr_HT <<- dat[ ,
c(1,3,4,5,7,8,9,10,11,12,13,14,15,16,19,20,21,23,24,25,29,3
0)]
  }
  if(kit==27){ #with RM-STR
    Ychr_HT <<- dat
  }
```

```
}


#We decided to take 20 Y-STR by removing rapidly mutating
Y-STR
find.YSTR.kit(Ychrgen,kit=21)

#Add one columns with haplotype as a string of character
Ychr_HT$HT <- NA
library("stringr")

for(i in 1:nrow(Ychr_HT)){
  Ychr_HT[i,ncol(Ychr_HT)] <-
paste(Ychr_HT[i,2:(ncol(Ychr_HT)-2)], collapse="_")}

#For the mitochondrial DNA
#Molecular data should be imported as a data frame in the
variable "mitoHVII_HT_Seq"
#Columns should be in order: individual, lineage number,
haplotype with haplogroup, haplotype without haplogroup and
DNA sequence

#Function to prepare data frame for the next function to
calculate EPP rate
#For the Y chromosome only
prep_data <- function(data,HTtable){
   #Add haplotypes to the table gen_dist with genetic
distances and check if both individuals have the same
  gen_dist_HT <- data
  gen_dist_HT$HT_IND1 <- NA
  gen_dist_HT$HT_IND2 <- NA
  gen_dist_HT$same_HT <- NA

  for(i in 1:nrow(gen_dist_HT)){
    gen_dist_HT[i,"HT_IND1"] <-
HTtable[which(HTtable[,1]==gen_dist_HT[i,1]),ncol(HTtable)]
    gen_dist_HT[i,"HT_IND2"] <-
HTtable[which(HTtable[,1]==gen_dist_HT[i,2]),ncol(HTtable)]
    gen_dist_HT[i,"same_HT"] <-
gen_dist_HT[i,"HT_IND1"]==gen_dist_HT[i,"HT_IND2"]
  }
```

```
  #Calculate the number of differences and set EPP to 1 if
more than maxdiff

  gen_dist_HT$NumSTR <- NA
  gen_dist_HT$Numdiff <- NA
  gen_dist_HT$NumEPP <- NA

  for(i in 1:nrow(gen_dist_HT)){
    if(gen_dist_HT[i,"same_HT"]==FALSE){
      a <- gen_dist_HT[i,"HT_IND1"]
      a <- str_split(a, pattern="_")
      c <- which(a[[1]]=="0"|a[[1]]=="NA") #we don't
compare alleles if there is 0 or NA
      b <- gen_dist_HT[i,"HT_IND2"]
      b <- str_split(b, pattern="_")
      d <- which(b[[1]]=="0"|b[[1]]=="NA")

      e <- c(c,d)
      e <- unique(e)

      if(length(e)!=0){ #to have only markers that can be
compared
        a <- a[[1]][-e]
        b <- b[[1]][-e]}

      if(length(e)==0){
        a <- a[[1]]
        b <- b[[1]]}

      gen_dist_HT[i,"NumSTR"] <- length(a)
      gen_dist_HT[i,"Numdiff"] <- length(a)-sum(a==b)
#length(a) is the number of markers compared without NAs or
0s

      maxdiff <- round(7*gen_dist_HT[i,"NumSTR"]/38)
#adjust number of differences accepted according to the
number of STR really compared based on Larmuseau et al.
(2013)
      #maxdiff is put to 0 in order to estimate global
mismatch rate
```

```
      if(gen_dist_HT[i,"Numdiff"]>maxdiff){
        gen_dist_HT[i,"NumEPP"] <- 1
      }
      if(gen_dist_HT[i,"Numdiff"]<=maxdiff){
        gen_dist_HT[i,"NumEPP"] <- 0
      }
    }

    if(gen_dist_HT[i,"same_HT"]==TRUE){
      a <- gen_dist_HT[i,"HT_IND1"]
      a <- str_split(a, pattern="_")
      c <- which(a[[1]]=="0"|a[[1]]=="NA")
      b <- gen_dist_HT[i,"HT_IND2"]
      b <- str_split(b, pattern="_")
      d <- which(b[[1]]=="0"|b[[1]]=="NA")

      e <- c(c,d)
      e <- unique(e)

      if(length(e)!=0){ #to have only markers that can be
compared
        a <- a[[1]][-e]
        b <- b[[1]][-e]}

      if(length(e)==0){
        a <- a[[1]]
        b <- b[[1]]}

      gen_dist_HT[i,"NumSTR"] <- length(a) #number of STR
compared

      gen_dist_HT[i,"Numdiff"] <- 0
      gen_dist_HT[i,"NumEPP"] <- 0
    }

  }

  #Keep some columns (pair ID, genetic distance and the
presence or absence of EPP)
```

```
  gen_dist_HT <<-
gen_dist_HT[,c("IND1_IND2","Gen_Dist","NumEPP")]
  rownames(gen_dist_HT) <<- NULL
}

#Function to prepare data frame for the next function to
calculate EPP rate
#For the mitochondrial DNA only
prep_data <- function(data,HTtable){

  #Add haplotypes to the table gen_dist with genetic
distances and check if same HT
  gen_dist_HT <- data
  gen_dist_HT$HT_IND1 <- NA
  gen_dist_HT$HT_IND2 <- NA
  gen_dist_HT$same_HT <- NA

  for(i in 1:nrow(gen_dist_HT)){
    gen_dist_HT[i,"HT_IND1"] <-
HTtable[which(HTtable[,"ind"]==gen_dist_HT[i,
"IND1"]),"mtDNA"]
    gen_dist_HT[i,"HT_IND2"] <-
HTtable[which(HTtable[,"ind"]==gen_dist_HT[i,
"IND2"]),"mtDNA"]
    gen_dist_HT[i,"same_HT"] <-
gen_dist_HT[i,"HT_IND1"]==gen_dist_HT[i,"HT_IND2"]
  }

  #Add sequences to the table
  gen_dist_HT$Seq_IND1 <- NA
  gen_dist_HT$Seq_IND2 <- NA
  gen_dist_HT$same_Seq <- NA

  for(i in 1:nrow(gen_dist_HT)){
    gen_dist_HT[i,"Seq_IND1"] <-
HTtable[which(HTtable[,"ind"]==gen_dist_HT[i,
"IND1"]),"Sequence"]
    gen_dist_HT[i,"Seq_IND2"] <-
HTtable[which(HTtable[,"ind"]==gen_dist_HT[i,
"IND2"]),"Sequence"]
```

```
    gen_dist_HT[i,"same_Seq"] <-
gen_dist_HT[i,"Seq_IND1"]==gen_dist_HT[i,"Seq_IND2"]
  }
#Calculate the number of differences and set EPP to 1 if
more than maxdiff

  gen_dist_HT$Numdiff <- NA
  gen_dist_HT$NumEPP <- NA

  for(i in 1:nrow(gen_dist_HT)){
    if(gen_dist_HT[i,"same_HT"]==FALSE &
gen_dist_HT[i,"same_Seq"]==FALSE){
      a <- gen_dist_HT[i,"Seq_IND1"]
      a <- str_split(a, pattern="")
      a <- a[[1]]
      b <- gen_dist_HT[i,"Seq_IND2"]
      b <- str_split(b, pattern="")
      b <- b[[1]]

      gen_dist_HT[i,"Numdiff"] <- length(a)-sum(a==b)}
#length(a) is the number of nucleotides compared

    if(gen_dist_HT[i,"same_HT"]==FALSE &
gen_dist_HT[i,"same_Seq"]==TRUE){ #if same haplotype but
not same haplogroup
      gen_dist_HT[i,"Numdiff"] <- 1} #To verify if really
just one difference

    if(gen_dist_HT[i,"same_HT"]==TRUE){
      gen_dist_HT[i,"Numdiff"] <- 0}

    maxdiff <- 1
#maxdiff is put to 0 to estimate global mismatch rates

    if(gen_dist_HT[i,"Numdiff"]>maxdiff){
      gen_dist_HT[i,"NumEPP"] <- 1
    }
    if(gen_dist_HT[i,"Numdiff"]<=maxdiff){
      gen_dist_HT[i,"NumEPP"] <- 0
    }
  }
```

```
  #Keep some columns (pair, num meiosis and num EPP)
  gen_dist_HT <<-
gen_dist_HT[,c("IND1_IND2","Gen_Dist","NumEPP")]
  rownames(gen_dist_HT) <<- NULL
}

###### 5. Function to estimate mismatch rates ######
#Function based on Larmuseau for the mismatch rate
estimation
library("binom")

estimateP <- function(dat,niters,tol){

  #Find pairs where BCA is not GCA
  indice <- which(dat$NumEPP>0)
  sizes <- as.matrix(dat[indice, "Gen_Dist"])
  nsizes <- dim(sizes)

  #Search iteratively for maximum likelihood estimate of p

  for (i in 1:niters){
    estim_p <- binom.exact(sum(dat[,"numEPP"]),
sum(dat[,"Gen_Dist"]), 0.5, conf.level=tol)
    minint <- estim_p$lower
    maxint <- estim_p$upper
    pestim <- estim_p$mean

    for (j in 1:nsizes[1]){
      N <- sizes[j]
      summ <- 0

      for (k in 1:N){
        summ <- summ + dbinom(k, N, pestim)/(1-dbinom(0, N,
pestim))*k
      }

      dat[indice[j], "numEPP"] <- summ
    }
  }
```

```
  p <- pestim
  c(p,c(minint,maxint))

}

###### 6. Set of functions to apply to calculate EPP ######

EPP_calcul <- function(pair.table, HTtable, niters=100000,
tol=0.95){
lines<-select.pairs(pair.table)
sub.gendist <- pair.table[which(pair.table$IND1_IND2 %in%
lines),]

prep_data(sub.gendist,HTtable)
estimateP(gen_dist_HT,niters,tol)
}

list.EPP <- list()

#Repeat the sampling a 1000 times
for(v in 1:1000){
  print(v)
 list.EPP[[v]] <- EPP_calcul(gen_dist2,Ychr_HT,
niters=10000)
}
#Use "mitoHVII_HT_Seq" instead of "Ychr_HT" for
mitochondrial DNA

save(list.EPP, file="EPPwithmutation_indepmeiosis.RData")

###### 6. Analysis of results #######

EPP <- numeric()
for(i in 1:length(list.EPP)){
  EPP[[i]] <- list.EPP[[i]][1]
  EPP <- unlist(EPP)
}

mean(EPP)

minEPP <- numeric()
```

```
for(i in 1:length(list.EPP)){
  minEPP[[i]] <- list.EPP[[i]][2]
  minEPP <- unlist(minEPP)
}

mean(minEPP)

maxEPP <- numeric()
for(i in 1:length(list.EPP)){
  maxEPP[[i]] <- list.EPP[[i]][3]
  maxEPP <- unlist(maxEPP)
}

mean(maxEPP)
```

ANNEXE C

ENSEMBLE DE FONCTIONS SERVANT À FAIRE L'IMPUTATION PROBABILISTE DU CHROMOSOME Y DANS LA GÉNÉALOGIE

Le script ci-dessous a été réalisé pour l'attribution avec 17 STR-Y du chromosome Y.

```
###### 1. Preparation of data ######
# Determine the working directory
setwd("folder path")

###### 2. Load data ######
#Genealogies should be imported as a data frame in the
variable "Balsac"
#Columns should be in order: individual, father, mother,
sex, marriage year of parents, place of marriage of
parents, marriage year of individual and place of marriage
of individual
#Load data on paternal lineages
#Lineages should be imported as a data frame in the
variable "pat_lin_linnumber"
#Columns should be in order: individual and paternal
lineage number

###### 3. Load data on haplotypes ######
#Molecular data should be imported as a data frame in the
variable "Ychrgen"
#Columns should be in order: individual and the list of Y-
STR considered

###### 4. Keep only individuals with 17 YSTR ######
#The columns corresponding to the Y-STRs wanted are put in
the vector "colystr"
colystr <- c(3,4,5,8,9,10,11,12,13,14,15,16,19,20,21,23,24)

Y_ind_17STR <- data.frame()
```

```
#We keep individuals genotyped with at least the Y-STR in
"colystr"
for(i in 1:nrow(Ychrgen)){
  if(sum(!is.na(Ychrgen[i,colystr]))>=17){
    Y_ind_17STR <- rbind(Y_ind_17STR, Ychrgen[i,])
  }
}

###### 5. Function to select the Y-STR kit ######
find.YSTR.kit <- function(dat,kit=c(7,12,17,20,27)){
  if(kit==7){
    Ychr_HT <- dat[ , c(1,3,5,9,12,14,16,23,30)]
  }
  if(kit==12){
    Ychr_HT <- dat[ ,
c(1,3,5,9,12,14,15,16,19,20,21,23,24,30)]
  }
  if(kit==17){
    Ychr_HT <- dat[ ,
c(1,3,4,5,8,9,10,11,12,13,14,15,16,19,20,21,23,24,30)]
  }
  if(kit==20){ #without RM-STR
    Ychr_HT <- dat[ ,
c(1,3,4,5,7,8,9,10,11,12,13,14,15,16,19,20,21,23,24,25,29,3
0)]
  }
  if(kit==27){ #with RM-STR
    Ychr_HT <- dat
  }
  Ychr_HT
}
#To keep only columns corresponding to chosen Y-STRs
Y_ind_17STR_woRM <- find.YSTR.kit(Y_ind_17STR,kit=17)
#without RM-STR


##Remove individuals for which there is 0 as allele
Y_ind_17STR_fin <- data.frame()
for(i in 1:nrow(Y_ind_17STR_woRM)){
  if(sum(Y_ind_17STR_woRM[i,2:18] !=0)>=17){
```

```
    Y_ind_17STR_fin <- rbind(Y_ind_17STR_fin,
Y_ind_17STR_woRM[i,])
  }
}

###### 6. Add haplotype in the table pat_lin_linnumber
######
#Add one column with haplotype as a string of characters
Y_ind_17STR_fin$HT <- NA
library("stringr")

for(i in 1:nrow(Y_ind_17STR_fin)){
  Y_ind_17STR_fin[i,ncol(Y_ind_17STR_fin)] <-
paste(Y_ind_17STR_fin[i,2:(ncol(Y_ind_17STR_fin)-2)],
collapse="_")
}

#Add the haplotype for the genotyped people in a data frame
with columns individual, lineage number and haplotype
pat_lin_HT <- merge(pat_lin_linnumber,
Y_ind_17STR_fin[,c(1,ncol(Y_ind_17STR_fin))], by = "ind",
all=TRUE)
pat_lin_HT[is.na(pat_lin_HT$HT),3] <- ""

save(pat_lin_HT, file="pat_lin_17str.RData")

#To have a data frame with individual, father, mother, sex,
lineage number and haplotype
yped <- merge(Balsac[Balsac$sex==1, c("ind", "father",
"mother", "sex")], pat_lin_HT, by="ind", all=TRUE)
yped$Hn <- yped$HT

save(yped, file="yped_17str_032018.RData")

###### 7. Create a list with paternal lineages which have a
genotyped person ######
ind_Y <- yped[yped$HT != "", "ind"]
lineage_list <- unique(pat_lin_HT[which(pat_lin_HT$ind %in%
ind_Y), "lineageNum"])

save(ind_Y, file="ind_Y_17str.RData")
```

```
save(lineage_list, file="lineage_list17Ystr.RData")

###### 8. Data on mutation rates for Y chromosome ######
#Data on mutation rates should be imported as a data frame
in the variable "YSTR_mut_rate"
#Columns should be in order Y-STR, mutation rate per marker
per generation, proportion of repetitive unit gain,
proportion of repetitive unit loss, proportion of gain or
loss of one repetitive unit, proportion of gain or loss of
2 repetitive units
#Identification of Y-STR names for the 17 Y-STR kit
YSTR17 <- c("DYS456",    "DYS389I",     "DYS390",
     "DYS389II",    "DYS458", "DYS19",  "DYS385a",
     "DYS385b",     "DYS393", "DYS391", "DYS439", "DYS635",
     "DYS392", "YGATAH4",     "DYS437", "DYS438", "DYS448")
save(YSTR17, file="YSTR17_list.RData")

#Function to calculate mean mutation rate for the Y-STR kit
find.mean.mutrate <- function(dat,kit=c(7,12,17,20,27)){
  if(kit==7){
    Ychr_mutrate <- mean(dat[dat[,"YSTR"] %in%
YSTR7,"mut_rate"])
  }
  if(kit==12){
    Ychr_mutrate <- mean(dat[dat[,"YSTR"] %in%
YSTR12,"mut_rate"])
  }
  if(kit==17){
    Ychr_mutrate <- mean(dat[dat[,"YSTR"] %in%
YSTR17,"mut_rate"])
  }
  if(kit==20){
    Ychr_mutrate <- mean(dat[dat[,"YSTR"] %in%
YSTR20,"mut_rate"])
  }
  if(kit==27){
    Ychr_mutrate <- mean(dat[dat[,"YSTR"] %in%
YSTR27,"mut_rate"])
  }

  Ychr_mutrate <<- Ychr_mutrate
```

```
}

find.mean.mutrate(YSTR_mut_rate,kit=17) #gives the number
of mutations per marker per generation in the variable
"Ychr_mutrate"

save(Ychr_mutrate, file="Ychr_meanmutrate_17str.RData")

###### 9. Function to randomly assign genealogical errors
######
ped.error <- function(pedigree, lineage.type = "father",
ped.break=-9){
  if(lineage.type == "father") {err.gen = 0.008}
  if(lineage.type == "mother") {err.gen = 0.004}


  err <- rbinom(sum(pedigree[,lineage.type]!=0),1,err.gen)
  pedigree2 <- pedigree

pedigree2[pedigree2[,lineage.type]!=0,][err==1,lineage.type
] <- ped.break
  pedigree2
}

###### 10. Function to create sub-pedigrees ######
sub.ped.sets <- function(pedigree, lineage.type = "father",
DNA.typed){
  sets <- list()
  for (i in 1:length(DNA.typed)){
    ancestor <- find.ancestor(pedigree, lineage.type =
lineage.type, ind=DNA.typed[i], ped.break=-9)
    sets[[i]] <- sub.ped(pedigree, lineage.type =
lineage.type, ancestor)
  }
  unique(sets)
}

###### 11. Function to identify the ancestor in a sub-
pedigree #######
find.ancestor <- function(pedigree, lineage.type =
"father", ind, ped.break=-9){
```

```
  if (!pedigree[pedigree[,"ind"]==ind,lineage.type] %in%
c(0,ped.break)) {
    ancestor <-
pedigree[pedigree[,"ind"]==ind,lineage.type]
    ancestor <-
find.ancestor(pedigree=pedigree,lineage.type=lineage.type,
ind=ancestor,ped.break=ped.break)
  } else {
    ancestor <- ind
  }
  ancestor
}

###### 12. Function to identify all descendants of an
ancestor #######
sub.ped <- function(pedigree, lineage.type = "father",
ancestor){
  descendants <- pedigree[pedigree[,lineage.type] %in%
ancestor,"ind"]
  if (length(descendants) > 0) {
    descendants <- sub.ped(pedigree=pedigree,
lineage.type=lineage.type, ancestor=descendants)
  }
  c(ancestor, descendants)
}

###### 13. Function to randomly assign mutations in the
genealogy #######
ped.mut <- function(pedigree, lineage.type = "father",
mut.rate, numSTR=17){
  if(lineage.type == "father") {mut.rate=mut.rate}

  mut <- rbinom(sum(!pedigree[,lineage.type] %in% c(0,-
9)),numSTR,mut.rate)
  new.pedigree2 <- pedigree
  new.pedigree2$mutation <- 0
  new.pedigree2[!new.pedigree2[,lineage.type] %in% c(0,-9),
"mutation"] <- mut
  new.pedigree2
}
```

```
###### 14. Function to create sub-sub-pedigrees #######
sub.ped.sets.mut <- function(pedigree, lineage.type =
"father", DNA.typed, ind.list){
  sets <- list()
  for (i in 1:length(DNA.typed)){
    ancestor.mut <- find.ancestor.mut(pedigree,
lineage.type = lineage.type, ind=DNA.typed[i], ped.break=-
9)
    sets[[i]] <- sub.ped.mut(pedigree, lineage.type =
lineage.type, ancestor.mut)
  }

  ind.toAdd <- pedigree[which(pedigree$mutation > 0),
"ind"]
  if(length(ind.toAdd) >0){
    ind.toAdd <- ind.toAdd[which(ind.toAdd %in% ind.list)]

    if(length(ind.toAdd) >0){
    for (i in 1:length(ind.toAdd)){
      ancestor.mut <- ind.toAdd[i]
      sets[[(length(sets)+1)]] <- sub.ped.mut(pedigree,
lineage.type = lineage.type, ancestor.mut)
    }
    }
  }

  for (i in 1:length(sets)){
    sets[[i]] <- as.numeric(sets[[i]])
  }
   unique(sets)
}

###### 15. Function to identify the ancestor in a sub-sub-
pedigree #######
find.ancestor.mut <- function(pedigree, lineage.type =
"father", ind, ped.break=-9){
  if (!pedigree[pedigree[,"ind"]==ind,lineage.type] %in%
c(0,ped.break) &
!pedigree[pedigree[,"ind"]==ind,"mutation"] >0) {
    ancestor <-
pedigree[pedigree[,"ind"]==ind,lineage.type]
```

```
    ancestor <-
find.ancestor.mut(pedigree=pedigree,lineage.type=lineage.ty
pe, ind=ancestor,ped.break=ped.break)
  } else {
    ancestor <- ind
  }
  ancestor
}

###### 16. Function to identify all descendants of an
ancestor in a sub-sub-pedigree #######
sub.ped.mut <- function(pedigree, lineage.type = "father",
ancestor){
    descendants <- pedigree[pedigree[,lineage.type] %in%
ancestor,"ind"]
    if (length(descendants) > 0) {
    descendants <-
descendants[which(pedigree[pedigree[,"ind"] %in%
descendants,"mutation"] ==0)]
    descendants <- sub.ped.mut(pedigree=pedigree,
lineage.type=lineage.type, ancestor=descendants)
  }
  c(ancestor, descendants)
}

###### 17. Function to order sub-sub-pedigrees according to
mutations #######
find.level <- function(pedigree, lineage.type = "father",
ind, ped.break=-9){
  level <- numeric()

  if(pedigree[pedigree[,"ind"]==ind,"mutation"] == 0){ #if
this individual have not mutated compared to his father
    aa <- 0
    a <- c(a,aa)

  } else { #if this individual has mutated compared to his
father
    a <- TRUE #we count this individual's mutation

    son <- ind
```

```
    cond <- !pedigree[pedigree[,1]==son,lineage.type] %in%
c(0,ped.break) #if this individual has a known father
  while (cond==T) { #check if ancestors of this individual
have mutated
      ancestor <- pedigree[pedigree[,1]==son,lineage.type]

      aa <- pedigree[pedigree[,1]==ancestor,"mutation"] !=0
      a <- c(a,aa)
      son <- ancestor
      cond <- !pedigree[pedigree[,1]==son,lineage.type]
%in% c(0,ped.break)
    }
  }
  level<-sum(a)
  level
}

###### 18. Function to attribute to ancestors haplotypes
observed in genotyped individuals #######
attr.Y.toAncestor <- function(pedigree, sub.pedigree,
DNA.typed){
  new.pedigree <- pedigree
  for(i in 1:length(sub.pedigree)){
    if(length(sub.pedigree[[i]]) > 1){
      ind.typed <- sub.pedigree[[i]]
[which(sub.pedigree[[i]] %in% DNA.typed)]
      HT.typed <- c(new.pedigree[new.pedigree$ind %in%
ind.typed, "HT"])

      ancestor <- sub.pedigree[[i]][1]
      new.pedigree[new.pedigree$ind %in% ancestor, "Hn"] <-
paste(c(HT.typed), collapse=";")
    }
  }
  new.pedigree
}

###### 19. Function to attribute haplotypes across all the
genealogy #######
```

```
attr.Y.down.mut <- function(pedigree, sub.pedigree,
YSTR.table, YSTR.list, lineage.type = "father", individuals
= ind_Y){
  #Create a vector for mutation probabilities
  mut_prob <- YSTR_mut_rate[match(YSTR.list,
YSTR_mut_rate$YSTR), "mut_rate"]

  #Prepare objects for "for loops"
  tbl_downY_sub <- pedigree
  tbl_downY_sub$Hsim <- NA

  tbl_final <- pedigree[,c("ind", "lineageNum")]

  var.subped <- 1:length(sub.pedigree)
  num.repeat <- 0

  repeat{
    num.repeat <- num.repeat+1 #to limit the number of
repeat if really impossible to have a result

    #For each sub-sub-pedigree, give haplotype to people
    for(i in var.subped){
      #Find the ancestor's ID
      ancestor <- sub.pedigree[[i]][1]
      num.mut <- pedigree[pedigree$ind == ancestor,
"mutation"]

      #Get ID of genotyped individuals
      genotyped_ind <-
tbl_downY_sub[which(tbl_downY_sub$ind %in% individuals),
"ind"]

      if(num.mut ==0){
        #Get haplotype attributed to ancestor for
probability calculations
        hap_anc <- pedigree[which(pedigree$ind %in%
ancestor),"Hn"]
        hap_anc <- unlist(strsplit(hap_anc, split = ";"))

        #Do a table of frequencies to have probability for
each haplotype
```

```
        hap_anc_table <- as.data.frame(table(hap_anc),
stringsAsFactors = F)
        hap_anc_table$frequency <-
hap_anc_table[,2]/sum(hap_anc_table[,2])
        colnames(hap_anc_table) <- c("hap",
"number","frequency")

        #Choose a random haplotype for ancestor
        hap_chosen <- sample(hap_anc_table$hap, 1, prob =
hap_anc_table$frequency) #select one haplotype for the
ancestor
        tbl_downY_sub[which(tbl_downY_sub$ind %in%
sub.pedigree[[i]]),"Hsim"] <- hap_chosen #Give to ancestor
the haplotype chosen before (2nd for loop)
      }

      if(num.mut !=0){
        if(sum(genotyped_ind %in% sub.pedigree[[i]]) ==
length(genotyped_ind)){ #if all the genotyped ind are in
this subped

      #Get haplotype attributed to ancestor and give it to
everyone in this sub-sub-pedigree
          hap_anc <- pedigree[which(pedigree$ind %in%
ancestor),"Hn"]
          hap_anc <- unlist(strsplit(hap_anc, split = ";"))

          #Do a table  of frequencies to have probability
for each haplotype
          hap_anc_table <- as.data.frame(table(hap_anc),
stringsAsFactors = F)
          hap_anc_table$frequency <-
hap_anc_table[,2]/sum(hap_anc_table[,2])
          colnames(hap_anc_table) <- c("hap",
"number","frequency")

          #Choose a random haplotype for ancestor
          hap_chosen <- sample(hap_anc_table$hap, 1, prob =
hap_anc_table$frequency) #select one HT for the ancestor
          tbl_downY_sub[which(tbl_downY_sub$ind %in%
sub.pedigree[[i]]),"Hsim"] <- hap_chosen
```

```
        } else {

        #Get haplotype from the father of ancestor
        ID_father <- pedigree[which(pedigree$ind ==
ancestor), lineage.type]
        hap_father <- tbl_downY_sub[which(tbl_downY_sub$ind
== ID_father), "Hsim"]

        if(is.na(hap_father)==T){
          new.hap <- NA}

        if(is.na(hap_father)==F){

        #Determine which marker is mutated
        YSTR.mut <- sample(YSTR.list, size = num.mut,
replace = F, prob = mut_prob)

        new.hap <- hap_father
        new.hap <- as.numeric(unlist(strsplit(new.hap,
split="_")))

        for(j in 1:length(YSTR.mut)){
          type.mut <- sample(c("gain","loss"), size = 1,
prob = as.numeric(YSTR.table[which(YSTR.table$YSTR ==
YSTR.mut[j]), c("prop_gain", "prop_loss")]))
          diff.mut <- sample(c(1,2,3,4,5), size = 1, prob =
as.numeric(YSTR.table[which(YSTR.table$YSTR ==
YSTR.mut[j]), c("prop_single", "prop_2", "prop_3",
"prop_4", "prop_5")]))

         pos.mut <- which(YSTR.list == YSTR.mut[j])

          if(type.mut == "gain"){
            new.hap[pos.mut] <- new.hap[pos.mut]+diff.mut}

          if(type.mut == "loss"){
            new.hap[pos.mut] <- new.hap[pos.mut]-diff.mut}
        }

        new.hap <- paste(new.hap, collapse = "_")
        }
```

```
        tbl_downY_sub[which(tbl_downY_sub$ind %in%
sub.pedigree[[i]]),"Hsim"] <- new.hap #Give to everyone the
new haplotype

        }
      }
    }

    #Test if it gets to what we observed
    a <- vector(mode="logical")
    aa <- vector(mode="logical")

    for (p in 1:length(genotyped_ind)){
      true_hap <- tbl_downY_sub[which(tbl_downY_sub$ind
%in% genotyped_ind[p]),"HT"] #Search haplotype observed for
a genotyped individual

      a <- identical(true_hap,
tbl_downY_sub[tbl_downY_sub$ind %in%
genotyped_ind[p],"Hsim"]) #Compare observed haplotype to
the one attributed
      aa <- c(aa,a) #Compile results for each genotyped
individual
    }

    if(sum(aa)==length(genotyped_ind)){#Keep this
simulation if attribution gave exactly what we observed
(new column)
      break
    }

    #Find sub-sub-pedigree to repeat because it did not
gave what we observed
    Fal.pos <- which(aa == FALSE)
    ind.rep <- genotyped_ind[Fal.pos]

    sub.rep <- list()
    for(n in 1:length(ind.rep)){
      sub.rep[[n]] <- which(lapply(sub.pedigree,
function(x) grep(ind.rep[n], x))>0)
```

```
    }
    sub.rep <- unlist(sub.rep)
    var.subped <- unique(sub.rep)

    if(num.repeat==100){
      break
    }
  } #close repeat

  ##Save result
  a <- vector(mode="logical")
  aa <- vector(mode="logical")

  for (p in 1:length(genotyped_ind)){
    true_hap <- tbl_downY_sub[which(tbl_downY_sub$ind %in%
genotyped_ind[p]),"HT"] #Search haplotype observed for a
genotyped individual

    a <- identical(true_hap,
tbl_downY_sub[tbl_downY_sub$ind %in%
genotyped_ind[p],"Hsim"]) #Compare observed haplotype to
the one attributed
    aa <- c(aa,a) #Compile results for each genotyped
individual
  }

  if(sum(aa)==length(genotyped_ind)){


    #Assign NA to everyone who are not in any sub-sub-
pedigree
    all_ind_ped <- pedigree$ind
    all_ind_subped <- unlist(sub.pedigree)

    if(length(all_ind_ped) != length(all_ind_subped)){
      absent_ind <- all_ind_ped[which(!(all_ind_ped %in%
all_ind_subped))]
      tbl_downY_sub[which(tbl_downY_sub$ind %in%
absent_ind),"Hsim"] <- NA #people for whom we cannot
attribute a haplotype
    }
```

```
    #Save result in a new table
    tbl_final$Hsim <- tbl_downY_sub$Hsim

  } else {
    tbl_final <- tbl_final
  }
  #Prepare the table with simulations
  if(ncol(tbl_final)>2){
    ind_HTattributed <- tbl_final[,c("ind","Hsim")]
    rnames <- ind_HTattributed[,1]
    ind_HTattributed <- ind_HTattributed[,-1, drop=FALSE]
    rownames(ind_HTattributed) <- rnames
    ind_HTattributed <-
as.data.frame((t(ind_HTattributed)), stringsAsFactors = F)
    rownames(ind_HTattributed) <- NULL
    ind_HTattributed
  }
}
###### 20. Function to calculate the occurrence obtained
for each haplotype for each individual #######
calcul.prob <- function(HT_simul, pedigree=gen){
  #To know the probability of each haplotype for each
individual
  ind_prob_HT <- pedigree[,c("ind", "lineageNum")] #prepare
the table
  hap_final_tbl <- as.character(unique(unlist(HT_simul)))

  if(sum(is.na(hap_final_tbl))>0){
    hap_final_tbl[which(is.na(hap_final_tbl))] <- "NA"
  } else {
    hap_final_tbl <- unique(c(hap_final_tbl, "NA"))
  }

  ind_prob_HT[hap_final_tbl] <- 0

  for (i in 1:ncol(HT_simul)){ #for each individual
(column) do a table of haplotype occurrences and put
results in ind_prob_HT

    HT_attributed <- unlist(HT_simul[,i])
```

```
    tbl_prob <- as.data.frame(table(HT_attributed,
useNA="always"), stringsAsFactors = F)
    colnames(tbl_prob) <- c("HT", "Number") #we want the
number of times a haplotype was observed and not the
frequency

    #To change NA into "NA" for the next step
    HT.names <- tbl_prob[,1]
    if(sum(is.na(HT.names)) >= 1){
      HT.names[which(is.na(HT.names))] <- "NA"
      tbl_prob[,1] <- HT.names
    }

    #Complete the table with all the counts
    for (j in 1:nrow(tbl_prob)){
      ind_prob_HT[which(ind_prob_HT$ind ==
as.numeric(colnames(HT_simul[i]))),
which(colnames(ind_prob_HT) == tbl_prob[j,"HT"])] <-
tbl_prob[j,"Number"]
    }
  }
  ind_prob_HT
}

###### 21. Group of functions to assign genealogical errors
and mutation and to attribute the Y chromosome across the
genealogy #######
simul_attr <- function(gen, lineage.type = "father",
ped.break=-9, individuals, mut.rate, YSTR.table,
YSTR.list){
  #Generate random genealogical errors
  gen2 <- ped.error(gen, lineage.type=lineage.type,
ped.break=ped.break)

  #Create sub-pedigrees from genotyped individuals
  indHT.gen <- gen[!(gen$HT == ""),"ind"]
  if (length(indHT.gen) > 0) subped <- sub.ped.sets(gen2,
lineage.type=lineage.type, DNA.typed = indHT.gen)

  #Generate random mutations
```

```
  gen.mut <- ped.mut(gen2, lineage.type=lineage.type,
mut.rate = mut.rate, numSTR=17)

  #Create sub-sub-pedigrees
  ind.subped.list <- unlist(subped)

  subped.mut <- sub.ped.sets.mut(gen.mut,
lineage.type=lineage.type, DNA.typed = indHT.gen, ind.list
= ind.subped.list)

  #Give a level to the different list of this subped.mut
  ind.subset.mut <- numeric()
  for(i in 1:length(subped.mut)){
    ind.subset.mut[[i]] <- subped.mut[[i]][1]
  }

  ind.subset.mut <- unlist(ind.subset.mut)

  level.ind <- numeric()
  a<<-logical()
  aa<-logical()

  for(i in 1:length(ind.subset.mut)){
    bb <- find.level(gen.mut, ind=ind.subset.mut[i])
    level.ind <- c(level.ind, bb)
  }

  names(subped.mut) <- paste("L", level.ind, sep = "")
  subped.mut <- subped.mut[order(names(subped.mut))] #to
replace list in order of level for mutation

  #Attribute haplotypes to ancestors
  ped.HTtoanc <- attr.Y.toAncestor(gen.mut, subped.mut,
indHT.gen)

  #Attribute haplotypes to ancestors' descendants
  ped.simul.HTdown <- attr.Y.down.mut(pedigree=ped.HTtoanc,
sub.pedigree=subped.mut, YSTR.table=YSTR.table,
YSTR.list=YSTR.list, lineage.type=lineage.type,
individuals=individuals)
```

```
  ped.simul.HTdown
}

###### 22. Function to do the attribution of the Y
chromosome for a given paternal lineage #######
#The variable "yped" was created in section 6
#The variables "ind_Y" and "lineage_list" were created in
section 7
#The variables "YSTR_mut_rate", "YSTR17" and "Ychr_mutrate"
were described in section 8

find.prob.HT.par <- function(dat,iter=8000){
  function.list <-
c("find.ancestor","sub.ped","sub.ped.sets","ped.error","ped
.mut","find.ancestor.mut",
"sub.ped.mut","sub.ped.sets.mut","find.level","attr.Y.toAnc
estor","attr.Y.down.mut",
"calcul.prob","simul_attr")

gen <<- yped[yped$lineageNum %in% dat,]

  cl<-makeCluster(detectCores() - 1)
  clusterExport(cl, varlist=c("gen", "ind_Y",
"Ychr_mutrate", "YSTR_mut_rate", "YSTR17", function.list))

  ped.allsimul.HT <- parSapply(cl,1:iter, function(x)
simul_attr(gen, lineage.type = "father", ped.break=-9,
individuals=ind_Y, mut.rate=Ychr_mutrate,
YSTR.table=YSTR_mut_rate, YSTR.list=YSTR17))
  stopCluster(cl)

  num.null <- sum(sapply(ped.allsimul.HT, is.null))

  if(num.null >0){
    ped.allsimul.HT <-
ped.allsimul.HT[!sapply(ped.allsimul.HT, is.null)]
    ped.allsimul.HT <-
data.table::rbindlist(ped.allsimul.HT)
    ped.allsimul.HT <- as.data.frame(ped.allsimul.HT,
stringsAsFactors = F)
```

```
  }

  if(num.null ==0){
    ped.allsimul.HT <- as.data.frame(t(ped.allsimul.HT))
  }

  #To calculate counts after all simulations
  if(nrow(ped.allsimul.HT)>0){
    ind.HT.prob <- calcul.prob(ped.allsimul.HT,
pedigree=gen)
    filename <- paste("ind.HT.prob.17str",dat, ".RData",
sep="")
    save(ind.HT.prob, file=filename)
  }
}

library(parallel)
lapply(lineage_list, function(x)
find.prob.HT.par(dat=x,iter=125000))
#"dat" is the lineage number
#"iter" is the number of simulations
```

ANNEXE D

FRÉQUENCES DES HAPLOTYPES DE L'ADN MITOCHONDRIAL DANS LES DIFFÉRENTES RÉGIONS ANALYSÉES

Les haplotypes mitochondriaux ont été définis par l'identité et la position des nucléotides différant par rapport à la séquence de référence (rCRS) entre les positions 16 069 et 16 383 (HVI) ainsi qu'entre 58 et 370 (HVII). Lorsque l'haplotype était identique à la référence, il a été défini comme étant l'haplotype « rCRS ». La fréquence de chaque haplotype est donnée pour les 15 régions retenues pour analyses (voir CHAPITRE II pour les détails) : Abitibi (ABI), Bas-Saint-Laurent (BSL), Beauce (BEA), Bois-Francs (BFR), Charlevoix (CHA), Côte-de-Beaupré (CDB), Côte-du-Sud (CDS), Côte-Nord (CNO), Estrie (EST), Gaspésie (GAS), Îles-de-la-Madeleine (IMA), Québec (agglomération) (QCA), Québec (QUE), Reste of Québec (RES) et Saguenay–Lac-Saint-Jean (SAG).

Tableau D.1. Fréquences des haplotypes mitochondriaux dans la population canadienne-française calculées pour 15 régions québécoises d'après le modèle généalogico-moléculaire développé dans ce projet.

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16304C.16311C.106DELG.107DELG.108DELA.109DELG.110DELC.111DELA.263G	1.483E-03	3.786E-04	2.141E-04	4.581E-03	2.373E-04	3.539E-04	6.210E-04	0	2.288E-03	5.569E-05	0	1.432E-03	3.232E-03	1.616E-03	1.212E-03
16183C.16189C.16249C.16362C.107A.257G.263G	4.563E-04	4.008E-04	8.135E-04	1.553E-04	2.373E-04	1.770E-04	7.763E-05	7.680E-04	2.559E-04	5.624E-03	0	1.308E-03	1.683E-04	5.654E-03	2.895E-04
111G.263G	2.053E-03	1.180E-03	3.639E-03	2.465E-03	9.492E-04	1.770E-03	3.765E-03	5.485E-04	2.498E-03	8.743E-03	0	4.048E-03	7.978E-03	0	1.122E-03
16189C.131C.195C.263G	3.080E-03	1.165E-02	5.566E-04	5.474E-03	6.170E-03	2.654E-03	9.587E-03	3.401E-03	2.784E-03	2.227E-03	0	3.180E-03	3.265E-03	3.231E-03	2.624E-03
143A.152C.263G	2.795E-03	1.537E-03	9.034E-03	4.503E-03	2.373E-04	3.716E-03	2.834E-03	6.253E-03	3.793E-03	0	0	2.189E-03	2.491E-03	0	5.248E-04
143A.195C.263G	2.224E-03	6.680E-05	1.285E-04	6.600E-04	0	0	1.553E-04	0	1.114E-03	2.784E-04	0	1.377E-03	3.232E-03	8.078E-04	2.714E-04
16304C.146C.195C.263G	0	0	0	1.941E-05	0	0	0	0	0	0	0	1.377E-05	0	0	0
146C.195C.263G	3.992E-03	2.227E-05	4.453E-03	9.123E-04	0	1.770E-04	1.941E-04	1.097E-04	4.124E-03	1.058E-03	0	1.253E-03	3.366E-04	2.423E-03	1.140E-03
146C.200G.263G	3.707E-03	0	2.697E-03	9.123E-04	2.373E-04	8.848E-04	1.164E-04	2.194E-04	1.806E-03	5.569E-05	0	1.597E-03	9.426E-03	0	1.086E-04
16270T.146C.263G	2.281E-04	1.781E-04	1.285E-04	4.270E-04	0	0	7.763E-05	1.426E-03	2.709E-04	6.070E-03	0	1.143E-03	1.246E-03	0	9.048E-05
16189C.146C.263G	1.711E-03	1.113E-04	1.285E-04	4.270E-04	2.373E-04	0	7.763E-05	1.097E-04	1.445E-03	0	0	3.992E-04	7.406E-04	3.231E-03	1.810E-04
16264T.146C.263G	0	0	0	0	0	0	0	0	0	2.227E-04	0	0	0	0	0
16129A.146C.263G	2.852E-04	4.454E-05	4.282E-05	0	0	0	0	4.388E-04	0	1.114E-03	0	5.507E-05	0	0	0
16093C.16213A.146C.263G	5.704E-05	1.113E-04	4.282E-05	2.523E-04	0	0	3.882E-05	2.194E-04	4.515E-04	0	0	1.927E-04	3.366E-04	0	1.810E-05
16189C.16193.1C.16220G.150T.195C.257G.263G	1.369E-03	8.217E-03	4.282E-05	3.882E-05	9.492E-04	3.539E-04	8.928E-04	1.536E-03	3.010E-04	6.682E-04	0	1.487E-03	1.212E-03	8.078E-04	5.067E-04
16216G.16311C.150T.263G	0	1.113E-04	0	0	0	0	0	2.194E-04	0	2.840E-03	0	0	0	0	0
150T.263G	0	0	0	0	0	0	0	0	0	5.569E-05	0	0	0	0	0
16093C.16274A.16311C.151T.152C.263G	8.555E-04	6.680E-05	0	1.495E-03	0	0	7.763E-05	0	1.023E-03	5.569E-05	0	9.637E-05	3.366E-05	0	7.238E-05
16311C.151T.263G	3.194E-03	5.589E-03	3.811E-03	2.329E-03	7.119E-04	1.947E-03	7.220E-03	3.620E-03	2.002E-03	5.569E-03	1.939E-03	3.056E-03	3.770E-03	1.616E-03	3.348E-03
152C.194T.263G	2.966E-03	3.763E-03	1.584E-03	1.339E-03	9.018E-03	5.486E-03	3.144E-03	1.646E-03	3.387E-03	1.671E-04	0	2.987E-03	4.444E-03	3.231E-03	6.315E-03
152C.199C.263G	1.369E-03	3.340E-04	3.297E-03	7.764E-04	2.373E-03	1.770E-03	3.882E-04	6.583E-04	8.128E-04	5.569E-05	0	1.280E-03	1.145E-03	1.616E-03	1.466E-03
152C.263G	4.392E-03	1.899E-02	4.710E-04	5.513E-03	2.136E-03	5.309E-04	7.957E-03	3.730E-03	3.808E-03	4.455E-03	6.464E-04	3.332E-03	2.760E-03	6.462E-03	3.909E-03
16249C.152C.263G	7.985E-04	1.002E-03	8.563E-04	1.320E-03	1.424E-03	8.848E-04	1.359E-03	1.097E-04	7.827E-04	3.397E-03	0	1.556E-03	1.178E-03	1.616E-03	2.280E-03
16248T.16354T.152C.263G	0	0	0	0	0	0	0	0	0	5.569E-04	0	0	0	0	0
16360T.152C.263G	2.909E-03	8.239E-04	5.481E-03	2.271E-03	7.119E-04	5.309E-03	2.251E-03	1.975E-03	4.515E-03	5.012E-04	0	1.886E-03	3.198E-03	2.423E-03	6.315E-03
16239G.16256T.16311C.152C.263G	6.274E-04	4.008E-04	2.141E-04	3.882E-04	0	0	2.290E-03	9.874E-04	2.258E-04	5.624E-03	0	4.543E-04	1.010E-04	0	1.086E-04
16221T.16291T.152C.263G	2.281E-04	8.907E-05	4.282E-05	1.165E-04	0	0	3.882E-05	2.194E-04	7.225E-04	6.460E-03	0	1.377E-04	6.733E-05	0	3.257E-04

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16186T.16311C.152C.263G	0	0	0	0	0	0	0	0	0	5.569E-05	0	0	0	0	0
16304C.152C.263G	1.711E-04	0	0	1.941E-05	7.119E-04	0	3.882E-05	0	0	0	0	3.304E-04	0	0	7.238E-05
16093C.16129A.16316G.152C.263G	2.053E-03	3.118E-04	6.637E-03	2.853E-03	2.136E-03	2.654E-03	2.018E-03	1.207E-03	2.378E-03	5.569E-05	0	5.700E-03	4.174E-03	3.231E-03	8.505E-03
16304C.152Y.263G	6.274E-04	8.907E-05	1.713E-03	5.047E-03	2.373E-04	1.770E-04	5.822E-04	0	2.137E-03	0	0	7.847E-04	1.986E-03	8.078E-04	3.438E-04
16302G.183G.263G	5.247E-03	3.140E-03	2.102E-02	8.346E-03	1.187E-03	7.078E-03	5.861E-03	1.097E-03	8.955E-03	5.680E-03	0	1.101E-02	1.007E-02	3.231E-03	2.678E-03
16189C.16260T.189G.191G.263G	1.711E-04	5.300E-03	8.563E-05	1.359E-04	0	1.770E-04	1.786E-03	2.194E-04	7.526E-05	3.341E-04	0	5.507E-04	4.713E-04	0	5.067E-04
16232A.16260T.189G.263G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.267E-04
16354T.194T.263G	0	3.340E-04	0	0	0	0	0	1.097E-04	0	1.504E-03	0	8.260E-05	0	0	3.619E-05
16293G.16311C.195C.207A.263G	5.704E-05	4.008E-04	5.994E-04	1.941E-05	0	0	1.048E-03	1.097E-04	7.526E-05	1.782E-03	0	3.855E-04	8.416E-04	0	1.810E-05
16129A.16264T.16316G.195C.263G	3.821E-03	2.450E-03	2.972E-02	4.950E-03	2.373E-04	7.255E-03	3.066E-03	4.388E-04	8.669E-03	1.002E-03	0	5.287E-03	7.440E-03	2.423E-03	3.257E-04
16238.1T.16293G.16311C.195C.263G	5.704E-05	4.454E-05	0	4.464E-04	0	0	0	0	2.408E-04	0	0	5.507E-05	0	8.078E-04	1.810E-05
16092C.16140C.16293G.16311C.195C.263G	0	0	0	0	0	0	0	0	0	5.569E-05	0	0	0	0	0
16176T.16184T.195C.263G	5.133E-03	3.941E-03	1.674E-02	5.687E-03	6.882E-03	1.203E-02	1.126E-02	8.777E-04	9.121E-03	1.504E-03	0	1.094E-02	1.336E-02	8.885E-03	3.131E-03
16311C.195C.263G	6.274E-04	0	0	3.494E-04	0	0	3.882E-05	0	6.020E-05	0	0	1.377E-05	3.366E-05	0	1.810E-05
195C.263G	7.358E-03	5.500E-03	9.720E-03	6.056E-03	3.417E-02	7.963E-03	7.996E-03	3.840E-03	8.940E-03	1.002E-03	6.464E-04	1.188E-02	9.628E-03	2.181E-02	3.465E-02
16189C.195C.263G	5.133E-04	2.227E-05	4.282E-05	1.941E-05	0	0	0	1.097E-04	7.526E-05	0	0	1.377E-05	0	0	1.810E-05
16286G.195C.263G	0	0	0	0	0	0	0	0	0	7.239E-04	0	0	0	0	0
16172C.16354T.204C.263G	7.415E-04	1.781E-04	0	3.339E-03	0	0	3.882E-05	0	9.482E-04	0	0	4.956E-04	3.030E-04	0	0
16362C.239C.263G	3.023E-03	1.292E-03	3.425E-04	3.688E-04	1.661E-03	1.239E-03	1.941E-04	9.874E-04	1.701E-03	2.840E-03	0	2.120E-03	8.046E-03	8.078E-04	4.886E-04
16291T.16362C.239C.263G	3.080E-03	3.786E-04	8.692E-03	3.144E-03	4.746E-04	4.955E-03	2.329E-03	3.291E-04	5.117E-03	0	0	3.827E-03	1.040E-02	8.078E-04	4.524E-04
16249C.16362C.239C.263G	0	0	0	0	0	0	0	0	0	1.114E-04	0	0	0	0	0
263G.291.1A	8.840E-03	3.369E-02	1.841E-03	4.717E-03	4.746E-04	2.831E-03	4.332E-02	1.492E-02	5.132E-03	2.072E-02	0	8.894E-03	3.434E-03	1.131E-02	6.116E-03
16295T.263G.292C	1.939E-03	0	0	3.882E-05	0	1.770E-04	1.164E-04	0	2.107E-04	5.569E-05	0	1.927E-04	1.010E-04	5.654E-03	1.810E-04
16362C.263G	5.875E-03	2.209E-02	4.796E-03	2.795E-03	2.610E-03	7.078E-04	1.727E-02	4.608E-03	2.709E-03	1.147E-02	0	4.103E-03	2.020E-03	5.654E-03	6.171E-03
16129A.263G	1.084E-03	5.278E-03	8.992E-04	9.705E-04	1.329E-02	5.663E-03	1.902E-03	4.827E-03	7.676E-04	2.784E-03	0	3.992E-03	9.426E-04	8.078E-04	1.661E-02
16311C.263G	1.996E-03	9.575E-04	7.707E-04	2.310E-03	1.898E-03	1.416E-03	5.046E-04	3.401E-03	2.438E-03	1.114E-04	0	3.566E-03	2.457E-03	1.616E-03	9.157E-03
263G	4.021E-02	2.594E-02	5.057E-02	5.891E-02	1.732E-02	3.787E-02	3.346E-02	2.940E-02	4.675E-02	8.002E-02	5.301E-02	3.473E-02	3.535E-02	2.342E-02	2.296E-02

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16093C.16221T.263G	5.646E-03	1.871E-03	8.949E-03	4.678E-03	2.610E-03	4.778E-03	3.260E-03	8.777E-04	5.524E-03	1.002E-03	0	5.823E-03	6.531E-03	4.039E-03	4.958E-03
16302G.263G	2.053E-03	2.227E-04	1.319E-02	1.359E-03	2.373E-04	8.848E-04	5.434E-04	2.194E-04	1.701E-03	1.559E-03	0	1.597E-03	1.481E-03	0	1.991E-04
16287T.263G	6.274E-04	0	4.282E-05	3.494E-04	0	0	3.882E-05	0	5.268E-04	0	0	8.260E-05	6.733E-05	8.078E-04	3.619E-05
16261T.263G	2.338E-03	5.322E-03	2.997E-04	1.029E-03	0	0	8.345E-03	2.084E-03	1.114E-03	5.680E-03	0	2.464E-03	3.030E-04	0	1.176E-03
16235G.263G	0	0	0	0	0	0	0	0	0	5.569E-05	0	0	0	0	0
16235G.16291T.263G	5.704E-05	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16235G.16291T.16293G.263G	0	2.004E-03	4.282E-05	0	0	0	0	3.291E-03	6.020E-05	1.225E-03	0	6.884E-05	0	0	1.086E-04
16234T.263G	3.530E-02	3.812E-02	5.322E-02	2.762E-02	5.458E-02	3.911E-02	5.267E-02	4.301E-02	2.491E-02	4.850E-02	3.232E-03	3.812E-02	3.427E-02	7.270E-02	6.972E-02
16209C.16218T.263G	1.141E-04	8.907E-05	2.141E-04	7.764E-05	2.373E-04	1.239E-03	2.562E-03	1.097E-04	2.709E-04	0	0	9.086E-04	6.396E-04	0	6.153E-04
16129A.16311C.16316G.263G	7.415E-04	2.895E-04	7.493E-03	1.184E-03	0	0	7.763E-04	0	1.520E-03	1.114E-04	0	6.746E-04	6.733E-04	0	1.086E-04
16239T.263G	0	0	0	0	0	0	0	0	0	2.227E-04	0	0	0	0	0
16183C.16189C.263G	4.563E-04	0	0	0	0	0	1.941E-04	0	0	2.005E-03	0	1.652E-04	2.659E-03	0	0
16319A.263G	1.939E-03	4.543E-03	3.854E-04	2.543E-03	2.848E-03	1.770E-04	1.541E-02	7.680E-04	1.761E-03	2.060E-03	0	1.762E-03	1.246E-03	3.231E-03	5.972E-04
16176T.16311C.263G	6.844E-04	5.344E-04	0	2.135E-04	2.373E-03	5.309E-04	3.882E-04	8.009E-03	6.020E-05	1.392E-03	2.715E-02	8.260E-04	2.020E-04	5.654E-03	9.211E-03
16304C.263G	3.593E-03	3.563E-04	6.637E-03	2.426E-03	2.373E-04	1.079E-02	2.057E-03	1.097E-04	2.438E-03	2.729E-03	0	5.259E-03	2.222E-03	8.078E-04	8.505E-04
16093C.16304C.16352C.263G	7.529E-03	1.405E-02	6.209E-03	5.804E-03	1.020E-02	9.733E-03	1.196E-02	1.317E-02	1.237E-02	1.008E-02	0	1.096E-02	5.285E-03	8.078E-03	9.989E-03
16086C.16147Y.16311C.263G	1.711E-04	4.053E-03	1.713E-04	1.262E-03	0	1.770E-04	2.911E-03	7.680E-04	5.569E-04	1.893E-03	0	8.949E-04	3.703E-04	8.078E-04	2.226E-03
16093C.16129A.263G	1.597E-03	3.162E-03	5.138E-04	9.511E-04	2.515E-02	1.451E-02	4.270E-04	1.262E-02	9.934E-04	3.341E-04	0	3.125E-03	1.178E-03	1.373E-02	1.413E-02
16129A.16256T.16311C.263G	2.338E-03	4.454E-05	1.285E-04	8.152E-04	0	1.239E-03	1.164E-04	0	5.268E-04	0	0	4.158E-03	4.713E-03	2.423E-03	2.714E-04
16235G.16291T.16311C.263G	3.935E-03	3.207E-03	7.450E-03	9.899E-04	9.492E-04	8.494E-03	2.290E-03	3.950E-03	1.144E-03	5.569E-04	0	3.318E-03	1.751E-03	0	3.257E-03
16189C.263G	3.821E-03	8.685E-04	3.618E-02	3.630E-03	2.373E-04	1.947E-03	1.591E-03	3.291E-04	8.052E-03	4.288E-03	0	3.304E-03	2.558E-03	6.462E-03	1.050E-03
16189C.16356C.263G	1.255E-03	8.017E-04	1.833E-02	1.514E-03	2.848E-03	5.840E-03	1.708E-03	1.097E-04	3.191E-03	2.227E-04	0	2.451E-03	3.299E-03	4.039E-03	1.647E-03
16213A.263G	2.852E-04	0	0	1.708E-03	0	0	1.553E-04	0	3.763E-04	0	0	4.130E-05	0	0	3.619E-05
16235T.16286T.263G	3.992E-04	0	0	1.048E-03	0	0	0	0	1.054E-04	0	0	0	0	0	1.810E-05
16093C.16311C.263G	1.768E-03	1.670E-03	1.627E-03	5.804E-03	0	5.309E-04	2.445E-03	1.042E-02	2.920E-03	1.871E-02	3.038E-02	3.345E-03	1.414E-03	0	3.058E-03
16179T.16189C.16356C.16362C.263G	1.141E-04	0	0	1.359E-03	0	0	0	0	3.311E-04	0	0	1.101E-04	6.733E-05	0	7.238E-05
16093C.16223T.16311C.263G	2.053E-03	1.982E-03	2.141E-04	3.242E-03	0	1.239E-03	2.717E-04	9.874E-04	1.806E-03	5.457E-03	0	1.060E-03	2.054E-03	0	1.375E-03
16093C.16223T.263G	7.985E-04	2.761E-03	4.282E-05	2.892E-03	2.373E-04	3.539E-04	1.980E-03	3.620E-03	1.430E-03	1.181E-02	0	8.123E-04	2.693E-04	8.078E-04	9.229E-04

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16291T.263G	0	0	0	0	0	0	0	0	0	5.569E-05	0	0	0	0	0
16223T.263G	5.704E-05	0	0	5.629E-04	0	0	0	0	7.977E-04	0	0	0	3.366E-05	0	1.810E-05
16188G.263G	3.593E-03	1.113E-04	1.541E-02	2.582E-03	2.373E-04	3.539E-04	5.434E-04	0	5.915E-03	1.671E-04	0	2.203E-03	3.366E-03	4.039E-03	5.248E-04
16170G.16234T.263G	1.768E-03	2.672E-04	0	1.165E-04	0	1.770E-04	5.434E-03	1.097E-04	7.225E-04	1.169E-03	0	1.046E-03	2.020E-04	0	8.686E-04
16234T.16301T.263G	2.624E-03	2.182E-03	8.135E-04	3.300E-04	2.373E-04	0	1.281E-02	4.388E-04	1.490E-03	1.225E-03	0	2.161E-03	2.155E-03	3.231E-03	1.538E-03
16269G.263G	3.992E-04	1.781E-04	8.563E-05	3.494E-04	0	2.300E-03	2.329E-04	0	1.957E-04	0	1.293E-03	9.224E-04	1.616E-03	6.462E-03	2.968E-03
16267T.263G	1.939E-03	7.438E-03	3.211E-03	2.679E-03	4.746E-04	1.593E-03	4.231E-03	1.975E-03	3.025E-03	7.573E-03	0	2.726E-03	1.751E-03	0	1.719E-03
16189C.16261T.16356C.16362C.263G	0	0	0	0	0	0	0	0	0	1.281E-03	0	0	0	0	3.619E-05
16148T.263G	0	0	0	0	0	0	0	0	0	5.569E-05	0	0	0	0	0
16304C.16362C.263G	1.711E-04	8.907E-05	0	5.823E-05	0	0	7.763E-05	9.874E-04	1.054E-04	7.685E-03	0	8.123E-04	1.010E-04	0	1.810E-04
16293G	0	0	0	1.941E-04	0	0	0	0	5.117E-04	0	0	0	0	0	0
rCRS	5.532E-03	3.216E-02	3.254E-03	2.019E-03	3.085E-03	5.309E-03	1.339E-02	7.570E-03	2.604E-03	5.624E-03	0	5.810E-03	3.265E-03	1.131E-02	5.338E-03
16195C	0	0	0	0	0	0	0	0	0	1.671E-04	0	0	0	0	0
16126C.64T.152C.263G	1.768E-03	3.786E-04	0	3.106E-04	0	1.770E-04	7.763E-04	5.485E-04	9.031E-04	3.007E-03	0	4.543E-04	6.396E-04	2.423E-03	1.991E-04
16111T.16192T.16223T.16290T.16319A.16362C.64T.73G.146C.153G.235G.263G	5.704E-05	4.454E-05	0	3.300E-04	0	0	0	1.097E-04	1.054E-03	0	0	0	3.366E-05	0	2.533E-04
16111T.16223T.16290T.16319A.16325C.16362C.64T.73G.94A.146C.153G.235G.263G	9.696E-04	2.316E-03	0	1.747E-04	2.373E-04	1.770E-04	3.882E-04	5.156E-03	4.515E-04	8.130E-03	0	9.362E-04	1.683E-04	0	1.484E-03
16298C.16311C.72C.195C.263G	9.297E-03	2.450E-04	2.141E-04	1.209E-02	7.119E-04	0	5.434E-04	8.777E-04	2.815E-03	0	0	1.143E-03	4.073E-03	6.462E-03	9.048E-04
16298C.16311C.72C.263G	1.597E-03	8.952E-03	2.141E-04	1.378E-03	4.746E-04	1.593E-03	1.188E-02	4.717E-03	1.972E-03	5.736E-03	0	2.726E-03	8.753E-04	4.039E-03	2.461E-03
16298C.72C.263G	1.951E-02	1.421E-02	4.744E-02	1.190E-02	8.709E-02	4.990E-02	2.736E-02	1.788E-02	1.924E-02	8.631E-03	3.232E-03	3.037E-02	2.562E-02	4.685E-02	5.217E-02
16298C.16301T.72C.263G	0	0	0	0	2.373E-04	0	0	0	0	5.569E-05	0	0	0	0	0
16162G.16298C.16311C.72C.263G	3.308E-03	6.680E-05	4.282E-05	1.359E-03	2.373E-04	0	3.882E-05	2.194E-04	1.400E-03	5.569E-05	0	1.239E-03	8.652E-03	8.078E-04	3.438E-04
16218T.16298C.72C.263G	6.844E-04	7.794E-04	3.854E-04	7.570E-04	0	1.770E-04	5.822E-04	1.097E-03	6.773E-04	5.569E-04	4.783E-02	1.611E-03	4.713E-04	0	1.991E-04
16126C.16294T.16325C.72C.73G.263G	4.620E-03	1.904E-02	8.692E-03	4.620E-03	6.645E-03	8.848E-03	1.700E-02	5.924E-03	4.892E-03	9.634E-03	0	6.471E-03	5.016E-03	8.078E-04	2.552E-03
72G.152C.263G	2.624E-03	5.567E-04	2.997E-04	3.669E-03	2.373E-04	5.309E-04	3.493E-04	4.388E-04	2.363E-03	4.455E-04	0	1.776E-03	7.070E-04	1.616E-03	6.695E-04
16224C.16234T.16311C.73G.114T.263G	0	0	0	1.941E-05	0	0	0	0	1.505E-05	0	0	8.260E-05	0	0	1.810E-05
16209C.16223T.16255A.16292T.73G.119C.189G.195C.204C.207A.263G	1.654E-03	1.781E-04	4.282E-05	1.108E-02	0	3.539E-04	3.882E-05	0	3.417E-03	5.569E-05	0	4.405E-04	8.079E-04	8.078E-04	1.086E-04

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16223T,16292T,73G,119C,189G,195C,204C,207A,263G	5.133E-04	2.450E-04	1.285E-04	7.764E-05	2.373E-04	3.539E-04	1.164E-04	1.251E-02	3.010E-05	1.336E-03	1.681E-01	7.159E-04	6.733E-05	0	8.686E-04
16166DELA,16183C,16189C,16249C,16311C,73G,143A,152C,263G,285T	2.852E-04	0	0	1.281E-03	2.373E-04	0	0	0	0	0	0	0	0	0	0
16183C,16189C,16234T,16270T,73G,146C,150T,263G	3.080E-03	5.233E-03	3.725E-03	3.164E-03	2.373E-04	1.787E-02	3.299E-03	5.595E-03	4.485E-03	1.565E-02	0	6.539E-03	2.188E-03	3.231E-03	3.076E-03
16224C,16234T,16311C,73G,146C,150T,263G	3.137E-03	1.559E-04	8.563E-05	1.786E-03	2.373E-04	1.770E-04	1.553E-04	1.097E-04	4.034E-03	2.227E-04	0	3.992E-04	2.357E-04	8.078E-04	4.524E-04
16224C,16311C,73G,146C,152C,185A,263G	2.795E-03	9.353E-04	1.541E-03	4.833E-03	2.373E-04	1.416E-03	4.503E-03	3.291E-04	3.733E-03	0	0	4.488E-03	3.501E-03	1.616E-03	1.231E-03
16356C,73G,146C,152C,195C,263G	2.567E-03	2.227E-04	8.563E-05	1.262E-03	0	1.770E-04	0	1.097E-04	9.031E-04	1.671E-04	0	3.304E-04	3.366E-05	0	5.429E-05
16092C,16129C,16182C,16183C,16189C,16362C,73G,146C,152C,217C,263G	0	0	0	3.882E-05	0	0	0	0	3.010E-05	2.506E-03	0	1.514E-04	0	0	0
16224C,16311C,73G,146C,152C,263G	6.901E-03	2.846E-02	5.952E-03	3.669E-03	1.139E-02	1.274E-02	1.032E-02	9.106E-03	5.223E-03	6.515E-03	0	1.234E-02	8.551E-03	1.696E-02	1.764E-02
16224C,16311C,16320T,16362C,73G,146C,152C,263G	6.274E-04	0	4.282E-05	5.047E-04	0	0	0	0	4.064E-04	0	0	5.507E-05	3.366E-05	0	9.048E-05
16189C,16224C,16311C,16320T,73G,146C,152C,263G	6.844E-04	7.349E-04	3.854E-04	8.346E-04	7.119E-04	5.309E-04	6.288E-03	1.097E-04	1.716E-03	1.448E-03	0	1.239E-03	2.020E-04	8.078E-04	2.352E-04
16187T,16189C,16223T,16255A,16278T,16319A,73G,146C,153G,195C,225A,226C,263G	4.848E-03	8.017E-04	7.921E-03	3.824E-03	7.119E-04	4.955E-03	1.304E-02	1.865E-03	6.020E-03	8.074E-03	0	7.021E-03	5.555E-03	3.231E-03	1.393E-03
16069T,16126C,16145A,16172C,16222T,16261T,73G,146C,183G,242T,263G,295T	1.711E-04	2.004E-04	0	1.359E-04	0	0	3.105E-04	0	0	5.569E-05	0	2.203E-04	1.010E-04	4.039E-03	0
16069T,16092C,16126C,16261T,73G,146C,185A,228A,263G,295T	3.992E-04	2.227E-05	0	1.398E-03	0	0	0	0	6.773E-04	5.569E-05	0	1.514E-04	0	0	1.810E-05
16069T,16126C,16145A,16172C,16261T,73G,146C,189G,242T,263G,295T	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1.991E-04
16224C,16311C,73G,146C,195C,263G	2.338E-03	3.897E-03	2.098E-03	3.882E-03	3.536E-02	4.601E-03	1.475E-03	8.886E-03	2.830E-03	1.114E-04	0	5.287E-03	8.652E-03	7.270E-03	9.627E-03
16356C,73G,146C,195C,263G	3.365E-03	6.012E-04	1.293E-02	2.213E-03	3.560E-03	3.716E-03	1.553E-03	6.583E-04	1.851E-03	5.569E-05	0	1.776E-03	2.760E-03	3.231E-03	1.900E-03
16069T,16126C,16145A,16179T,16231C,16261T,73G,150T,152C,195C,215G,263G,295T	3.080E-03	2.182E-03	8.992E-03	6.134E-03	2.373E-04	7.432E-03	7.103E-03	4.388E-04	5.087E-03	1.615E-03	0	5.576E-03	5.117E-03	2.423E-03	1.828E-03
16069T,16126C,16134T,16145A,16179T,16231C,16261T,73G,150T,152C,195C,215G,263G,295T	7.985E-04	1.113E-04	3.854E-04	6.794E-04	0	1.770E-03	9.316E-04	0	4.967E-04	1.114E-04	0	2.189E-03	2.626E-03	0	1.629E-04

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16069T,16126C,16145A,16231C,16261T,73G,150T,152C,195C,215G,263G,295T	3.821E-03	1.937E-03	5.010E-03	4.076E-03	1.187E-03	1.416E-03	7.802E-03	1.865E-03	4.154E-03	1.114E-04	0	4.695E-03	8.113E-03	4.039E-03	8.867E-04
16223T,16298C,16325C,16327T,73G,150T,152C,249DELA,263G,290DELA,291DELA	2.909E-03	2.004E-03	9.420E-04	6.017E-04	7.119E-04	1.239E-03	2.989E-03	2.940E-02	9.482E-04	1.074E-01	0	3.290E-03	1.178E-03	0	2.172E-03
16069T,16126C,16193T,16278T,73G,150T,152C,263G,295T	1.945E-02	4.022E-02	1.965E-02	1.174E-02	9.018E-03	2.265E-02	3.656E-02	2.030E-02	1.200E-02	1.136E-02	6.464E-04	2.095E-02	1.387E-02	1.777E-02	1.954E-02
16168T,16343G,73G,150T,152C,263G	4.506E-03	6.680E-05	1.713E-04	5.745E-03	0	0	2.329E-04	0	4.892E-03	0	0	3.717E-04	2.020E-04	2.423E-03	2.714E-04
16189C,16325C,73G,150T,152C,263G	0	0	0	0	0	0	0	1.097E-04	0	8.910E-04	0	4.130E-05	0	0	0
16069T,16126C,16213A,16261T,73G,150T,185A,228A,263G,295T	0	0	0	0	0	0	0	0	0	1.169E-03	0	0	0	0	0
16176T,16270T,73G,150T,185A,263G	0	0	0	0	0	0	0	0	0	7.796E-04	0	0	0	0	1.810E-05
16189C,16270T,73G,150T,185A,263G	1.312E-03	2.227E-05	0	0	0	0	3.882E-05	0	1.505E-05	0	0	1.377E-05	0	0	1.810E-05
16179T,16356C,73G,150T,195C,263G	1.312E-03	2.160E-03	1.413E-03	3.727E-03	2.610E-03	4.247E-03	1.553E-03	3.291E-04	1.174E-03	0	0	2.712E-03	2.659E-03	0	1.267E-04
16093C,16224C,16311C,16362C,73G,150T,263G	1.312E-03	2.004E-04	8.563E-05	5.920E-03	2.373E-04	3.539E-04	7.763E-05	3.291E-04	2.152E-03	1.671E-04	0	1.101E-04	5.386E-04	0	3.619E-05
16126C,16153A,16207T,16294T,73G,150T,263G	0	0	0	0	0	0	0	0	0	1.114E-04	0	0	0	0	0
16126C,16294T,16296T,16304C,73G,150T,263G	0	2.227E-05	0	0	0	0	0	2.194E-04	0	2.227E-04	0	0	0	0	1.810E-05
16343G,73G,150T,263G	0	0	0	0	0	0	0	0	0	1.114E-04	0	0	0	0	0
16192T,16270T,73G,150T,263G	0	4.454E-05	0	0	0	0	0	1.097E-04	0	3.341E-04	0	2.753E-05	0	0	3.619E-05
16192T,16270T,16319A,73G,150T,263G	4.392E-03	4.899E-04	2.783E-03	2.679E-03	5.933E-03	1.398E-02	3.726E-03	7.680E-04	1.987E-03	1.392E-03	0	7.861E-03	1.168E-02	8.078E-04	3.312E-03
16189C,16270T,73G,150T,263G	1.027E-03	7.794E-04	1.285E-04	3.882E-04	2.373E-04	8.848E-04	2.329E-04	3.620E-03	4.064E-04	4.511E-03	0	3.153E-03	1.515E-03	0	5.429E-04
16189C,16270T,16311C,73G,150T,263G	3.479E-03	1.441E-02	9.420E-04	1.378E-03	7.119E-04	3.539E-04	1.203E-02	1.755E-03	1.626E-03	1.225E-03	0	2.877E-03	2.188E-03	4.847E-03	1.411E-03
16126C,16153A,16294T,73G,150T,263G	1.084E-03	5.567E-04	8.135E-03	6.211E-04	1.661E-03	5.132E-03	7.375E-04	3.620E-03	1.550E-03	3.898E-04	0	1.143E-03	6.396E-04	0	2.533E-03
16224C,16265C,16311C,73G,150T,263G	9.126E-04	6.680E-05	2.569E-04	3.882E-05	3.085E-03	2.831E-03	1.553E-04	0	6.020E-05	0	0	3.442E-04	7.743E-04	4.039E-03	9.048E-04
16126C,16163G,16186T,16189C,16294T,16311C,73G,150T,263G	6.844E-04	2.227E-05	8.563E-05	4.464E-04	0	0	1.164E-04	0	9.031E-04	5.569E-05	0	1.927E-04	1.010E-04	0	1.086E-04
16114A,16192T,16270T,73G,150T,263G	1.711E-03	4.454E-05	4.282E-05	7.376E-03	2.373E-04	0	0	1.097E-04	1.294E-03	1.838E-03	0	2.478E-04	3.030E-04	8.078E-04	1.810E-04
16256T,16270T,73G,150T,263G	1.768E-03	1.113E-04	1.028E-03	6.405E-04	9.492E-04	9.202E-03	5.434E-04	4.388E-04	7.526E-04	5.569E-05	0	6.884E-03	4.343E-03	7.270E-03	4.216E-03
16126C,16294T,16296T,16304C,73G,151T,152C,263G	5.418E-03	1.960E-03	3.254E-03	8.463E-03	1.187E-03	5.309E-03	3.804E-03	8.777E-04	7.586E-03	3.898E-04	0	5.231E-03	1.188E-02	0	1.719E-03

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16126C,16294T,16296T,16304C,73G,151T,263G	2.008E-02	4.908E-02	5.091E-02	1.429E-02	6.502E-02	6.070E-02	2.946E-02	5.442E-02	1.946E-02	3.714E-02	0	3.065E-02	1.959E-02	3.393E-02	5.233E-02
16129C,16183C,16189C,16256T,16362C,73G,152C,183G,217C,263G	1.312E-03	5.567E-03	8.135E-04	6.522E-03	1.424E-03	2.831E-03	1.824E-03	1.097E-03	4.064E-03	1.114E-04	0	4.598E-03	7.978E-03	0	1.303E-03
16069T,16126C,16186T,16189C,73G,152C,185A,188G,228A,263G,295T	5.704E-05	6.680E-05	0	3.882E-05	0	0	0	0	1.054E-04	0	0	1.377E-05	3.366E-05	8.078E-04	0
16126C,16163G,16186T,16189C,16294T,73G,152C,195C,263G	3.422E-04	0	0	6.405E-04	0	0	0	0	9.031E-04	2.227E-04	0	5.507E-05	6.733E-05	0	1.810E-05
16126C,16163G,16186T,16189C,16294T,16354T,73G,152C,195C,263G	3.023E-03	7.705E-03	3.854E-04	1.611E-03	1.661E-03	1.593E-03	3.377E-03	6.583E-03	9.482E-04	8.798E-03	3.232E-03	2.065E-03	1.481E-03	4.847E-03	5.881E-03
16126C,16163G,16186T,16189C,16234T,16294T,73G,152C,195C,263G	0	2.227E-05	0	0	0	0	0	0	0	5.569E-05	0	0	0	0	0
16129A,16172C,16223T,16311C,73G,152C,199C,203A,204C,250C,263G	0	4.454E-05	1.285E-04	1.747E-04	4.509E-03	0	1.553E-04	6.583E-04	4.515E-05	7.796E-04	0	2.616E-04	2.693E-04	0	3.203E-03
16086C,16129A,16223T,16319A,73G,152C,199C,204C,207A,239C,250C,263G	0	0	0	0	0	0	0	0	0	1.114E-04	0	0	0	0	0
16129A,16223T,73G,152C,199C,204C,207A,250C,263G	1.084E-02	1.423E-02	1.400E-02	5.571E-03	1.331E-01	5.256E-02	4.891E-03	3.445E-02	6.939E-03	2.394E-03	0	1.448E-02	7.608E-03	4.847E-02	1.117E-01
16129A,16223T,16248T,73G,152C,199C,204C,207A,250C,263G	1.711E-04	0	2.997E-04	1.747E-04	2.373E-04	0	3.882E-04	0	3.010E-05	1.782E-03	0	2.891E-04	3.703E-04	8.078E-04	0
16082T,16183C,16189C,16224C,16234T,16294T,73G,152C,199C	1.426E-03	1.113E-04	2.569E-04	4.270E-03	0	8.848E-04	1.203E-03	0	1.505E-03	0	0	1.294E-03	1.010E-03	0	8.324E-04
16126C,16163G,16186T,16189C,73G,152C,200G,263G	1.084E-03	6.859E-03	9.420E-04	7.764E-04	1.898E-03	1.062E-03	6.987E-03	1.755E-03	5.569E-04	2.840E-03	0	3.428E-03	5.386E-04	0	1.140E-03
16129C,16183C,16189C,16362C,73G,152C,217C,263G	0	2.227E-05	4.282E-05	0	0	0	0	0	0	6.125E-04	0	0	0	0	0
16084A,16129C,16183C,16189C,16362C,73G,152C,217C,263G	7.985E-04	6.859E-03	1.713E-04	1.165E-04	2.373E-04	5.309E-04	6.599E-04	5.815E-03	4.515E-04	4.065E-03	0	6.333E-04	1.683E-04	1.616E-03	1.086E-04
16224C,16311C,16319A,73G,152C,263G	8.555E-04	4.454E-05	0	2.329E-04	2.373E-04	0	0	0	9.482E-04	0	0	2.753E-05	6.733E-05	3.231E-03	5.429E-05
16145A,16176G,16223T,73G,152C,263G	5.133E-03	1.777E-02	1.028E-03	2.582E-03	3.085E-03	2.124E-03	1.720E-02	7.241E-03	3.085E-03	1.164E-02	0	3.690E-03	1.515E-03	4.847E-03	3.366E-03
16104T,16126C,16294T,16304C,73G,152C,263G	7.871E-03	2.160E-03	7.065E-03	2.155E-03	2.231E-02	1.380E-02	1.025E-02	4.937E-03	6.788E-03	3.898E-04	0	8.191E-03	5.656E-03	6.462E-03	1.721E-02
16126C,16189C,16294T,16296T,73G,152C,263G	0	0	0	0	0	0	0	0	0	2.227E-04	0	0	0	0	0

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16126C.16278T.16294T.16296T.16304C.16362C.73G.152C.263G	0	2.227E-05	0	0	0	0	0	0	0	5.012E-04	0	0	0	0	0
16126C.16163G.16186T.16189C.16294T.73G.152C.263G	1.141E-03	2.895E-04	0	7.570E-04	0	8.848E-04	0	2.194E-04	2.017E-03	5.569E-05	0	1.652E-04	6.733E-05	1.616E-03	1.267E-04
16162G.16209C.73G.152C.263G	0	0	0	0	0	0	0	0	0	2.227E-04	0	0	0	0	0
16256T.16270T.73G.152C.263G	0	0	0	0	0	0	0	1.097E-04	0	1.671E-04	0	0	0	0	0
16172C.16219G.16278T.73G.152C.263G	6.958E-03	3.496E-03	1.541E-03	8.249E-03	1.898E-03	2.831E-03	8.151E-04	3.083E-02	2.845E-03	1.370E-02	9.632E-02	2.272E-03	3.871E-03	0	3.836E-03
16183C.16189C.16213A.16223T.16254C.16278T.73G.153G.195C.200G.225A.263G	0	0	0	0	0	0	0	0	0	0	0	0	3.366E-05	0	0
16129A.16189C.16223T.16278T.73G.153G.195C.225A.226C.263G	0	0	0	0	0	0	0	0	1.505E-05	0	0	0	0	0	0
16189C.16223T.16278T.73G.153G.195C.225A.226C.263G	1.255E-03	1.559E-04	4.282E-05	1.553E-04	0	0	7.763E-05	1.097E-04	6.020E-05	2.784E-04	0	9.499E-04	7.406E-04	1.616E-03	6.876E-04
16145A.16189C.16223T.16278T.16301T.73G.153G.195C.225A.226C.263G	1.312E-03	2.115E-03	2.355E-03	3.999E-03	4.746E-04	7.078E-04	6.210E-04	2.315E-02	2.453E-03	1.119E-02	1.151E-01	9.362E-04	1.145E-03	8.078E-04	1.882E-03
16189C.16223T.16278T.73G.153G.195C.225A.263G	1.027E-03	8.907E-05	4.282E-05	5.435E-04	0	0	7.763E-05	3.291E-04	1.355E-03	1.114E-04	0	1.652E-04	1.347E-04	3.231E-03	3.619E-05
16069T.16126C.16366T.73G.185A.188G.200G.228A.263G.295T	0	0	0	0	0	0	0	0	0	0	0	2.753E-05	0	0	3.619E-05
16069T.16126C.16266T.73G.185A.188G.228A.263G.295T	5.704E-05	0	0	0	0	0	0	1.097E-04	1.505E-05	8.353E-04	0	1.377E-05	0	0	0
16069T.16126C.16163G.73G.185A.188G.228A.263G.295T	0	0	0	0	0	0	0	0	0	5.569E-05	0	0	0	0	0
16223T.16292T.16311C.16344T.73G.185A.189G.195C.204C.207A.263G	4.791E-03	3.941E-03	2.573E-02	4.076E-03	2.136E-03	6.017E-03	6.521E-03	6.583E-04	5.629E-03	2.227E-04	0	8.756E-03	1.040E-02	6.462E-03	1.430E-03
16129A.16223T.16292T.16311C.16344T.73G.185A.189G.195C.204C.207A.263G	2.453E-03	1.247E-02	8.563E-05	9.511E-04	2.373E-03	2.831E-03	6.055E-03	2.194E-03	4.365E-04	1.002E-03	0	1.900E-03	1.145E-03	0	3.655E-03
16069T.16126C.73G.185A.228A.263G.295T	5.704E-04	2.360E-03	4.282E-05	1.165E-04	0	1.770E-04	1.281E-03	1.317E-03	1.505E-04	5.012E-04	7.757E-03	2.891E-04	3.030E-04	2.423E-03	3.619E-04
16069T.16126C.16325C.73G.185A.228A.263G.295T	0	0	0	0	0	0	0	0	1.054E-04	0	0	0	0	0	0
16069T.16126C.73G.185A.263G.295T	2.852E-04	2.227E-04	6.423E-04	5.823E-05	0	0	3.882E-05	7.131E-03	1.355E-04	3.453E-03	7.628E-02	3.717E-04	1.010E-04	0	1.176E-03
16069T.16126C.16259T.73G.185A.263G.295T	2.852E-04	1.559E-04	0	1.941E-05	0	0	0	9.984E-03	6.020E-05	1.114E-04	1.164E-02	2.753E-04	1.347E-04	0	3.619E-05

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16069T,16126C,16311C,73G,185A,263G,295T	0	0	4,282E-05	0	0	0	0	0	0	0	0	0	0	0	0
16069T,16126C,16147T,73G,185A,263G,295T	2,395E-03	8,907E-04	1,713E-04	2,407E-03	0	1,770E-04	7,763E-05	2,194E-04	1,520E-03	1,671E-04	0	4,543E-04	7,406E-04	1,616E-03	7,419E-04
16093C,16224C,16311C,73G,188G,263G,264T	1,825E-03	8,596E-03	2,098E-03	7,570E-04	0	3,539E-04	5,822E-03	1,536E-03	8,880E-04	2,172E-03	0	1,556E-03	2,996E-03	8,078E-04	1,918E-03
16223T,16292T,16362C,73G,189G,194T,195C,204C,207A,263G	5,704E-05	2,227E-05	0	1,941E-05	0	0	0	0	0	0	0	0	0	0	0
16223T,16292T,73G,189G,195C,199C,204C,263G	2,281E-04	2,895E-04	0	5,629E-04	0	3,539E-04	2,717E-04	0	7,977E-04	5,569E-05	0	2,340E-04	6,733E-05	0	5,429E-05
16192T,16256T,16270T,73G,194T,263G	1,711E-04	2,227E-05	0	1,165E-04	0	0	0	0	7,526E-05	0	0	1,377E-05	0	0	5,067E-04
16223T,16292T,73G,195C,204C,207A,263G	2,281E-04	2,316E-03	3,854E-04	6,405E-04	1,851E-02	4,424E-03	3,882E-05	5,705E-03	1,385E-03	1,336E-03	0	2,161E-03	1,414E-03	0	7,673E-03
16189C,16223T,16278T,73G,195C,225A,226C,263G	3,821E-03	2,160E-03	1,049E-02	2,504E-03	1,187E-03	9,910E-03	4,192E-03	1,097E-04	4,846E-03	1,392E-03	0	4,488E-03	4,107E-03	8,078E-04	1,846E-03
16093C,16224C,16311C,73G,195C,263G	1,192E-02	7,927E-03	1,276E-02	6,114E-03	4,248E-02	2,159E-02	1,681E-02	2,370E-02	9,798E-03	4,232E-03	0	1,569E-02	1,589E-02	3,150E-02	3,956E-02
16356C,73G,195C,263G	9,696E-04	1,826E-03	5,994E-04	1,087E-03	9,492E-03	1,663E-02	3,920E-03	1,646E-03	1,475E-03	2,005E-03	0	5,466E-03	3,501E-03	4,847E-03	1,015E-02
16183C,16189C,16221T,16234T,16290T,16324Y,73G,195C,263G	4,791E-03	4,320E-03	8,135E-04	1,611E-03	1,187E-03	2,124E-03	3,532E-03	1,865E-03	2,875E-03	5,569E-04	0	4,708E-03	7,675E-03	8,078E-03	1,249E-03
16093C,16126C,16224C,16311C,73G,195C,263G	0	0	0	0	0	0	0	0	1,505E-05	6,125E-04	0	0	0	0	0
16093C,16224C,16291T,16311C,73G,195C,263G	5,704E-05	0	0	0	0	0	0	0	1,505E-05	1,114E-04	0	0	0	0	0
16187T,16224C,16311C,73G,195C,263G	0	0	4,282E-05	0	0	0	0	0	0	5,569E-04	0	0	0	0	0
16179T,16278T,16356C,73G,195C,263G	1,141E-04	2,227E-05	1,285E-04	1,339E-03	0	1,770E-04	1,941E-04	0	7,225E-04	0	0	3,855E-04	3,838E-03	0	0
16261T,16356C,73G,195C,263G	0	0	0	0	0	0	0	0	0	0	0	0	0	0	5,429E-05
16129A,16172C,16223T,16311C,73G,199C,203A,204C,250C,263G	3,821E-03	2,004E-04	2,569E-04	1,456E-03	2,373E-04	3,539E-04	6,987E-04	4,388E-04	2,258E-03	6,125E-04	0	1,900E-03	6,430E-03	3,231E-03	8,686E-04
16129A,16223T,16362C,73G,199C,204C,250C,263G	2,852E-04	0	4,282E-05	3,882E-05	2,373E-04	0	0	0	1,656E-04	0	0	2,299E-03	1,010E-03	0	4,524E-04
16223T,16325C,16362C,73G,213C,240G,263G	1,711E-04	8,685E-04	1,285E-04	3,882E-05	7,119E-04	1,593E-03	3,882E-04	9,106E-03	1,204E-04	1,749E-02	0	1,046E-03	3,030E-04	4,847E-03	4,524E-04
16069T,16126C,73G,228A,263G,295T	2,167E-03	1,113E-04	2,569E-04	1,126E-03	7,119E-04	1,416E-03	1,553E-04	2,194E-04	5,719E-04	0	0	1,638E-03	1,003E-02	1,616E-03	4,162E-04

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16223T,16227C,16290T,16311C,16319A,73G,235G,263G	1.483E-03	2.895E-04	8.563E-05	4.639E-03	0	0	8.151E-04	1.097E-04	5.614E-03	3.341E-04	0	4.543E-04	3.030E-04	0	3.076E-04
16069T,16126C,16145A,16172C,16192T,16222T,16261T,73G,242T,263G,295T	0	0	0	0	0	0	0	0	0	0	0	0	2.020E-04	0	0
16069T,16126C,16145A,16172C,16222T,16261T,73G,242T,263G,295T	8.783E-03	3.741E-03	7.236E-03	7.842E-03	1.210E-02	1.557E-02	1.747E-03	1.755E-03	5.102E-03	1.504E-03	0	8.494E-03	1.973E-02	1.131E-02	1.183E-02
16069T,16126C,73G,263G,295T	0	0	0	0	0	0	0	0	0	3.341E-04	0	6.884E-05	0	0	0
16069T,16126C,16145A,16222T,16261T,73G,263G,295T	2.966E-03	4.899E-04	2.269E-03	8.094E-03	2.373E-04	0	4.658E-04	5.485E-04	4.591E-03	1.726E-03	0	6.746E-04	5.386E-04	4.847E-03	2.352E-04
16224C,16311C,73G,263G	1.894E-02	7.651E-02	2.924E-02	1.361E-02	8.543E-03	1.681E-02	3.109E-02	2.743E-02	1.714E-02	1.838E-02	0	1.993E-02	1.205E-02	1.050E-02	2.850E-02
16126C,16183C,16189C,16294T,16296T,73G,263G	1.483E-03	1.982E-03	5.994E-04	2.232E-03	4.746E-04	1.770E-03	1.591E-03	4.388E-04	1.355E-03	5.569E-05	0	2.643E-03	2.794E-03	8.078E-04	4.705E-04
16126C,16294T,16296T,16304C,73G,263G	7.985E-04	1.002E-03	0	9.317E-04	2.373E-04	0	1.164E-04	5.485E-04	1.520E-03	1.893E-03	0	3.029E-04	8.079E-04	0	5.429E-05
16126C,16239T,16294T,16296T,16304C,73G,263G	1.010E-02	9.419E-03	2.783E-03	1.733E-02	9.492E-04	1.947E-03	9.355E-03	2.973E-02	5.358E-03	1.414E-02	1.558E-01	4.860E-03	4.915E-03	4.039E-03	3.836E-03
16154C,16192T,16256T,16270T,16311C,73G,263G	3.080E-03	5.567E-04	5.523E-03	1.805E-03	2.373E-04	0	1.902E-03	0	2.800E-03	0	0	1.721E-03	9.089E-04	0	3.981E-04
16192T,16256T,16270T,16291T,16294T,73G,263G	1.255E-03	1.336E-04	8.563E-05	2.426E-03	0	3.539E-04	4.270E-04	0	9.633E-04	5.569E-05	0	2.203E-04	6.733E-05	0	9.048E-05
73G,263G	9.867E-03	3.118E-04	1.070E-03	2.970E-03	4.746E-04	7.078E-04	6.599E-04	4.388E-04	1.054E-03	5.569E-05	0	2.313E-03	5.252E-03	2.425E-03	1.249E-03
16093C,16224C,16256T,16311C,73G,263G	1.141E-03	1.113E-04	4.282E-05	1.747E-04	2.373E-04	0	7.763E-05	2.194E-04	1.806E-04	1.671E-04	0	5.094E-04	2.828E-03	8.078E-04	7.238E-04
16126C,16294T,73G,263G	6.844E-04	2.227E-05	1.541E-03	2.155E-03	0	0	9.704E-04	4.388E-04	3.612E-04	1.782E-03	0	6.884E-04	7.743E-04	2.423E-03	3.800E-04
16126C,16163G,16186T,16189C,16221T,16294T,73G,263G	3.080E-03	1.069E-03	5.138E-04	1.495E-03	3.749E-02	9.733E-03	5.822E-04	1.360E-02	1.475E-03	8.910E-04	0	3.855E-03	1.582E-03	2.019E-02	2.352E-02
16126C,16163G,16172C,16186T,16189C,16294T,16298C,73G,263G	1.483E-03	7.126E-04	4.282E-04	6.017E-04	0	1.770E-04	9.316E-04	3.401E-03	1.791E-03	2.227E-04	0	5.094E-04	8.416E-04	5.654E-03	2.787E-03
16126C,16294T,16296T,73G,263G	1.084E-03	1.537E-03	1.627E-03	9.511E-04	2.373E-04	2.831E-03	1.359E-03	1.097E-04	1.234E-03	1.281E-03	0	1.459E-03	8.079E-04	0	8.686E-04
16192T,16256T,16270T,73G,263G	1.141E-04	0	0	0	0	0	0	0	0	1.615E-03	0	0	0	0	1.810E-05
16256T,16270T,73G,263G	4.563E-04	2.227E-05	4.282E-05	0	0	0	0	2.084E-03	1.505E-05	6.125E-04	0	5.507E-05	3.366E-05	0	1.448E-04
16182C,16183C,16189C,16217C,16305T,73G,263G	2.281E-04	0	3.854E-04	2.912E-04	0	0	6.210E-04	0	6.623E-04	0	0	3.304E-04	5.050E-04	0	1.810E-05
16311C,73G,263G	0	0	0	0	0	0	0	0	0	5.569E-05	0	0	0	0	0

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16093C,16224C,16311C,73G,263G	0	0	0	0	0	0	0	0	0	7,796E-04	0	0	0	0	5,429E-05
16224C,16265C,16311C,73G,263G	2,453E-03	6,747E-03	4,796E-03	5,474E-03	2,373E-04	1,380E-02	6,172E-03	2,194E-03	4,786E-03	2,784E-03	0	5,231E-03	5,016E-03	1,616E-03	2,190E-03
16093C,16224C,16301T,16311C,73G,263G	2,110E-03	1,292E-03	4,025E-03	1,572E-03	7,119E-04	8,848E-03	5,201E-03	5,485E-04	1,415E-03	2,227E-04	0	3,483E-03	2,390E-03	4,039E-03	1,357E-03
16126C,16256T,16294T,16296T,16362C,73G,263G	2,567E-03	1,336E-04	1,606E-02	2,155E-03	1,661E-03	1,593E-03	2,251E-03	7,680E-04	3,251E-03	2,227E-04	0	2,822E-03	2,962E-03	0	3,511E-03
16126C,16147T,16179T,16294T,16296T,16297C,16304C,73G,263G	2,053E-03	3,340E-04	2,997E-04	1,941E-04	0	3,539E-04	3,882E-05	0	2,709E-04	0	0	7,847E-04	7,574E-03	1,616E-03	9,953E-04
16126C,16172C,16294T,16296T,16304C,73G,263G	2,224E-03	2,316E-03	4,924E-03	1,514E-03	1,970E-02	1,239E-03	5,124E-03	9,106E-03	3,823E-03	8,910E-04	0	3,098E-03	2,256E-03	5,654E-03	8,161E-03
16126C,16294T,16296T,16304C,16362C,73G,263G	1,255E-03	4,454E-05	4,282E-05	1,029E-03	2,373E-04	1,062E-03	1,164E-04	0	1,370E-03	0	0	2,120E-03	4,747E-03	8,078E-04	1,629E-04
16162G,73G,263G	4,563E-04	8,017E-04	0	7,958E-04	2,373E-04	1,770E-04	3,882E-05	0	7,526E-04	5,569E-05	0	4,543E-04	6,060E-04	0	2,352E-04
16192T,16256T,16270T,16291T,73G,263G	4,962E-03	6,101E-03	1,713E-03	2,077E-03	2,373E-04	1,239E-03	9,665E-03	6,583E-03	2,514E-03	5,958E-03	0	4,516E-03	8,854E-03	4,039E-03	8,324E-04
16093C,16192T,16256T,16270T,16291T,73G,263G	1,939E-03	2,227E-05	0	2,329E-04	2,373E-04	0	3,882E-05	1,097E-04	3,913E-04	0	0	1,101E-04	1,347E-04	5,654E-03	3,619E-05
16256T,16270T,16293G,16311C,73G,263G	0	6,680E-05	8,992E-04	5,241E-04	0	7,078E-04	6,987E-04	0	1,927E-03	5,569E-05	0	3,166E-04	1,044E-03	0	2,895E-04
16183C,16189C,16270T,16362C,73G,93G,150T,263G	0	0	0	0	0	0	0	0	0	0	0	1,377E-05	1,010E-04	0	5,429E-04
16162G,73G,93G,263G	5,133E-04	8,907E-05	0	1,068E-03	0	0	0	2,194E-04	1,114E-03	5,569E-05	0	2,203E-04	2,357E-04	0	5,429E-05
16181G,93G,200G,263G	0	0	0	0	0	0	0	1,097E-04	0	0	2,586E-03	0	0	0	3,619E-05
93G,263G	2,852E-04	4,454E-05	0	3,688E-04	0	8,848E-04	0	0	5,870E-04	6,125E-04	0	1,652E-04	0	0	2,172E-04
16184T,16284G,93G,263G	0	0	0	0	0	0	0	0	0	2,784E-04	0	0	0	0	0
16069T,16126C,16311C,73G,146C,185A,188G,228A,263G,295T	1,255E-03	1,113E-04	8,050E-03	1,883E-03	2,373E-04	0	4,270E-04	1,097E-04	3,025E-03	5,569E-05	0	7,297E-04	1,582E-03	0	1,104E-03
16183C,16189C,16270T,73G,150T,185A,189G,263G	6,844E-04	4,454E-05	2,569E-03	1,611E-03	0	0	2,329E-04	3,291E-04	4,365E-04	0	0	7,572E-04	1,616E-03	0	7,238E-05
16129A,16224C,16311C,73G,150T,199C,263G	0	0	8,563E-05	0	0	0	0	0	0	0	0	0	0	0	0
16256T,16270T,16287T,16327T,73G,263G	5,704E-05	4,454E-05	7,707E-04	2,523E-04	2,373E-04	3,539E-04	1,863E-03	0	3,010E-04	1,114E-04	0	1,638E-03	9,089E-04	1,616E-03	9,048E-04
16256T,16263G,16270T,73G,263G	1,198E-03	1,470E-03	1,670E-03	1,048E-03	4,746E-04	7,255E-03	1,164E-03	2,194E-04	1,249E-03	2,784E-04	0	2,382E-03	1,717E-03	0	5,972E-04
16074G,16302G,263G	5,133E-04	2,227E-05	1,370E-03	9,123E-04	2,373E-04	1,947E-03	4,658E-04	3,291E-04	8,128E-04	0	3,232E-03	3,043E-03	3,063E-03	0	6,695E-04

Tableau D.1 (suite)

Haplotype ADNmt (SNP)	ABI	BSL	BEA	BFR	CHA	CDB	CDS	CNO	EST	GAS	IMA	QCA	QUE	RES	SAG
16069T,16145A,16231C,73G,150T,152C,189G,195C,215G,263G,295T	1.540E-03	6.680E-05	1.199E-03	5.823E-04	2.373E-04	1.239E-03	6.987E-04	1.097E-04	3.763E-04	5.569E-05	0	7.820E-03	4.881E-03	8.078E-04	9.591E-04
16129A,16278T,93G,146C,263G	1.654E-03	1.514E-03	6.851E-04	2.407E-03	4.746E-04	3.539E-04	2.484E-03	2.633E-03	1.219E-03	1.225E-03	0	1.569E-03	3.636E-03	8.078E-04	4.886E-04
16129A,16223T,73G,199C,204C,250C,263G	2.281E-04	9.130E-04	2.141E-04	2.135E-04	0	0	7.763E-05	4.388E-04	1.054E-04	3.341E-04	0	3.442E-04	0	0	3.619E-05
16192T,16298C,72C,195C,263G	1.996E-03	4.454E-05	5.695E-03	1.960E-03	1.424E-03	6.548E-03	3.183E-03	0	2.288E-03	4.455E-04	0	5.231E-03	5.487E-03	0	3.257E-04

ANNEXE E

FRÉQUENCES DES HAPLOTYPES DU CHROMOSOME Y OBSERVÉS AVEC 17 STR-Y DANS LES DIFFÉRENTES RÉGIONS ANALYSÉES

Les haplotypes Y ont été définis par l'ensemble des allèles obtenus pour 17 STR-Y dont l'ordre est le suivant : DYS389I, DYS635, DYS389II, DYS458, DYS19, YGATAH4, DYS448, DYS391, DYS456, DYS390, DYS438, DYS392, DYS437, DYS385a, DYS385b, DYS393 and DYS439. Les allèles sont séparés par le symbole « _ ». La fréquence des haplotypes observés chez les personnes génotypées seulement (c.-à-d. excluant ceux générés par les simulations lors de l'imputation) est donnée pour les 11 régions retenues pour analyses (voir CHAPITRE II pour les détails) : Bas-Saint-Laurent (BSL), Beauce (BEA), Charlevoix (CHA), Côte-de-Beaupré (CDB), Côte-du-Sud (CDS), Côte-Nord (CNO), Gaspésie (GAS), Îles-de-la-Madeleine (IMA), Québec (agglomération) (QCA), Reste of Québec (RES) et Saguenay–Lac-Saint-Jean (SAG).

Tableau E.1. Fréquences des haplotypes du chromosome Y observés avec 17 STR-Y dans la population canadienne-française calculées pour 11 régions québécoises d'après le modèle généalogico-moléculaire développé dans ce projet.

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
11_23_28_16_15_13_20_11_16_25_11_11_14_11_14_13_10	0	0	0	0	0	0	1,042E-04	0	0	0	0
12_19_29_15_16_12_21_10_14_22_10_12_15_14_17_15_10	6,128E-05	6,148E-05	0	0	5,682E-05	0	8,187E-05	0	5,853E-04	0	3,650E-05
12_20_29_18_16_12_21_10_16_21_10_11_16_13_15_14_10	1,592E-03	7,851E-03	9,251E-04	3,293E-03	4,928E-03	3,033E-04	5,110E-03	0	3,250E-03	0	1,853E-03
12_21_28_15_14_11_20_10_13_22_10_11_16_13_14_13_11	1,035E-08	0	0	4,540E-09	1,295E-09	5,922E-09	0	0	4,098E-09	0	8,906E-10
12_21_28_15_14_11_20_10_14_22_10_11_16_13_14_13_13	6,141E-07	9,765E-08	1,142E-07	9,764E-07	1,760E-07	4,366E-07	1,148E-04	0	1,974E-07	0	4,486E-07
12_21_28_15_14_11_20_10_14_22_10_11_16_13_15_14_11	3,129E-05	7,195E-04	0	1,244E-03	3,906E-04	0	9,027E-05	0	4,766E-03	4,747E-03	6,951E-04
12_21_28_15_14_11_20_10_14_23_10_11_16_14_14_13_11	6,847E-03	2,494E-03	1,320E-03	3,016E-03	1,069E-02	1,975E-03	2,572E-03	0	2,915E-03	4,785E-07	2,152E-03
12_21_28_15_14_11_20_10_14_23_10_11_16_15_16_13_12	5,771E-04	3,942E-04	6,052E-04	1,175E-03	1,033E-03	1,334E-03	2,310E-04	0	2,096E-03	3,328E-03	1,982E-03
12_21_28_15_15_11_19_10_14_24_8_11_14_16_18_12_12	6,822E-04	2,624E-03	0	0	5,653E-03	2,959E-03	3,688E-03	0	1,728E-03	0	6,285E-04
12_21_28_15_15_11_20_10_14_22_10_11_16_13_14_13_10	9,526E-03	1,290E-04	0	1,276E-03	2,811E-03	2,080E-03	2,357E-03	0	1,573E-03	1,130E-03	9,612E-04
12_21_28_15_15_11_20_10_14_22_10_11_16_13_14_13_12	1,044E-03	2,734E-04	7,466E-04	9,445E-04	4,492E-04	7,749E-04	2,713E-04	0	9,905E-04	9,111E-04	7,487E-04
12_21_28_15_16_11_20_10_14_22_10_11_16_14_15_13_11	0	1,790E-04	0	0	1,633E-04	0	0	0	1,198E-04	1,184E-03	9,359E-05
12_21_28_16_14_11_19_10_14_22_10_11_16_13_14_13_11	2,235E-04	2,991E-04	3,088E-04	7,751E-04	1,714E-04	1,317E-07	3,153E-04	0	1,588E-03	3,504E-03	8,692E-04
12_21_28_16_14_11_19_10_14_23_10_11_16_14_14_13_11	2,478E-03	1,143E-03	9,270E-08	6,918E-04	2,553E-03	7,762E-04	8,576E-04	0	3,596E-04	1,027E-03	1,949E-04

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
11_23_28_16_15_13_20_11_16_25_11_11_14_11_14_13_10	0	0	0	0	0	0	1.042E-04	0	0	0	0
12_19_29_15_16_12_21_10_14_22_10_12_15_14_17_15_10	6.128E-05	6.148E-05	0	0	5,682E-05	0	8,187E-05	0	5,853E-04	0	3.650E-05
12_20_29_18_16_12_21_10_16_21_10_11_16_13_15_14_10	1.592E-03	7.851E-03	9.251E-04	3.293E-03	4,928E-03	3,033E-04	5,110E-03	0	3,250E-03	0	1.853E-03
12_21_28_15_14_11_20_10_13_22_10_11_16_13_14_13_11	1.035E-08	0	0	4,540E-09	1,295E-09	5,922E-09	0	0	4,098E-09	0	8,906E-10
12_21_28_15_14_11_20_10_14_22_10_11_16_13_14_13_13	6,141E-07	9,765E-08	1,142E-07	9,764E-07	1,760E-07	4,366E-07	1,148E-04	0	1.974E-07	0	4.486E-07
12_21_28_15_14_11_20_10_14_22_10_11_16_13_15_14_11	3.129E-05	7.195E-04	0	1,244E-03	3,906E-04	0	9,027E-05	0	4.766E-03	4,747E-03	6.951E-04
12_21_28_15_14_11_20_10_14_23_10_11_16_14_14_13_11	6.847E-03	2.494E-03	1.320E-03	3,016E-03	1.069E-02	1,975E-03	2.572E-03	0	2,915E-03	4,785E-07	2,152E-03
12_21_28_15_14_11_20_10_14_23_10_11_16_15_16_13_12	5.771E-04	3.942E-04	6,052E-04	1,175E-03	1,033E-03	1.334E-03	2.310E-04	0	2,096E-03	3,328E-03	1,982E-03
12_21_28_15_15_11_19_10_14_24_8_11_14_16_18_12_12	6.822E-04	2.624E-03	0	0	5,653E-03	2,959E-03	3.688E-03	0	1,728E-03	0	6,285E-04
12_21_28_15_15_11_20_10_14_22_10_11_16_13_14_13_10	9.526E-03	1,290E-04	0	1,276E-03	2,811E-03	2,080E-03	2,357E-03	0	1,573E-03	1,130E-03	9,612E-04
12_21_28_15_15_11_20_10_14_22_10_11_16_13_14_13_12	1,044E-03	2.734E-04	7,466E-04	9,445E-04	4,492E-04	7,749E-04	2.713E-04	0	9,905E-04	9,111E-04	7,487E-04
12_21_28_15_16_11_20_10_14_22_10_11_16_14_15_13_11	0	1,790E-04	0	0	1,633E-04	0	0	0	1,198E-04	1,184E-03	9,359E-05
12_21_28_16_14_11_19_10_14_22_10_11_16_13_14_13_11	2,235E-04	2.991E-04	3,088E-04	7,751E-04	1,714E-04	1.317E-07	3.153E-04	0	1,588E-03	3,504E-03	8,692E-04
12_21_28_16_14_11_19_10_14_23_10_11_16_14_14_13_11	2,478E-03	1.143E-03	9.270E-08	6,918E-04	2,553E-03	7.762E-04	8,576E-04	0	3,596E-04	1,027E-03	1,949E-04

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
12_21_28_16_14_11_20_10_14_22_10_11_16_13_14_13_12	1.704E-04	2.242E-04	1.817E-07	7.378E-03	5.197E-05	1.220E-07	1.541E-07	0	4.106E-04	3.406E-07	8.263E-05
12_21_28_16_14_11_20_10_15_23_10_11_15_14_14_13_11	0	0	0	0	0	0	1.146E-04	0	0	0	0
12_21_28_16_15_11_19_10_13_24_9_11_16_14_17_12_11	3.134E-05	0	3.733E-04	0	5.956E-05	0	0	0	8.365E-05	0	5.730E-05
12_21_28_17_13_10_20_10_15_24_10_11_14_12_19_12_12	2.475E-03	1.312E-02	1.007E-03	5.165E-04	3.895E-03	8.429E-04	1.694E-03	0	1.822E-03	1.129E-03	9.900E-04
12_21_28_17_16_10_21_10_14_25_10_11_15_13_16_13_12	0	6.151E-05	3.579E-04	0	0	0	0	0	0	0	7.581E-05
12_21_28_18_15_11_21_10_15_22_10_11_16_12_14_14_11	0	0	0	0	0	0	0	0	0	0	0
12_21_29_15_14_11_20_10_14_23_10_11_16_13_14_13_12	0	1.819E-04	0	0	0	0	0	0	6.252E-05	0	5.600E-05
12_21_29_16_15_12_21_10_15_22_10_11_16_14_14_14_11	3.777E-05	0	0	0	0	1.524E-03	3.591E-03	0	1.204E-04	0	3.012E-05
12_21_29_17_14_10_18_10_14_24_10_11_14_11_11_15_13	4.410E-04	6.158E-05	0	0	1.140E-04	0	5.054E-03	0	1.771E-04	0	7.832E-05
12_21_29_17_15_12_21_10_15_22_10_11_16_13_15_14_12	0	0	0	0	0	0	1.146E-04	0	0	0	0
12_21_30_17_15_11_22_10_15_23_10_11_16_13_14_13_11	0	0	0	0	0	0	0	0	2.244E-05	0	5.083E-05
12_22_28_14_14_11_20_11_15_22_10_11_16_13_14_12_13	3.601E-05	3.000E-03	0	0	0	0	0	0	9.005E-05	0	0
12_22_28_15_14_11_20_10_14_23_10_11_16_14_16_13_12	1.450E-07	1.105E-07	1.435E-07	3.145E-07	2.131E-07	3.079E-07	4.071E-08	0	4.455E-07	6.602E-07	4.943E-07
12_22_28_16_14_11_20_10_14_22_10_11_16_13_14_14_12	1.916E-03	1.002E-03	4.128E-04	7.988E-04	1.949E-03	3.591E-04	1.598E-03	0	2.099E-03	0	7.103E-04

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
12_22_28_16_14_11_21_10_15_22_9_11_14_11_12_12_11	0	0	0	0	0	0	0	0	2.601E-05	0	0
12_22_28_16_15_11_19_10_13_24_9_11_16_10_18_12_12	7,148E-04	2.850E-03	0	0	2,528E-03	0	8,571E-05	0	1,089E-03	0	2,763E-04
12_22_28_16_15_13_19_10_14_22_10_11_16_13_15_13_11	8,142E-05	0	2,065E-03	0	0	2,113E-03	0	0	0	0	2,455E-04
12_22_29_15_14_11_19_10_15_22_10_11_16_13_13_13_11	0	0	0	0	0	0	1,146E-04	0	0	0	0
12_23_28_14_14_11_20_10_14_23_10_11_16_14_15_13_11	5,927E-04	2,176E-03	3,247E-04	3,531E-03	5,942E-03	9,955E-04	4,576E-03	0	2,056E-03	2,212E-03	8,762E-04
12_23_28_15_14_11_20_10_14_22_10_11_15_13_14_13_12	2,018E-03	7,124E-05	7,691E-04	0	0	1,141E-03	9,917E-05	0	1,600E-04	0	1,123E-03
12_23_28_15_14_11_20_10_14_22_10_11_16_13_14_13_11	0	0	0	0	0	0	1,146E-04	0	0	0	0
12_23_28_15_14_12_19_11_16_24_12_13_15_11_14_14_12	2,494E-04	3,214E-05	9,760E-04	3,636E-04	2,815E-05	2,626E-03	1,282E-03	6,188E-03	1,651E-04	0	7,806E-04
12_23_28_15_14_12_19_11_16_24_12_13_15_11_14_15_12	6,797E-04	1,348E-04	4,134E-03	1,178E-03	1,204E-04	1,088E-02	4,302E-03	2,601E-02	6,794E-04	0	3,305E-03
12_23_28_15_15_11_20_11_14_23_10_11_16_13_14_13_11	1,490E-09	9,826E-04	0	7,068E-04	6,895E-05	0	2,343E-04	0	2,019E-04	0	2,726E-05
12_23_28_16_14_11_20_10_15_22_10_11_16_13_14_13_11	0	0	0	0	0	0	3,001E-04	0	0	0	0
12_23_28_17_14_11_18_10_16_23_12_13_14_12_15_14_13	0	5,612E-05	0	0	6,269E-05	0	0	0	1,866E-05	0	3,000E-05
12_23_28_17_14_12_19_11_16_23_12_12_16_11_13_13_13	2,730E-04	0	0	0	0	6,265E-03	9,925E-05	1,140E-02	2,250E-05	0	3,944E-04
12_23_28_17_14_12_19_11_16_25_12_13_15_10_11_13_12	2,045E-02	2,908E-04	2,555E-03	7,639E-04	1,174E-02	9,175E-03	7,681E-03	0	2,093E-03	2,318E-03	3,345E-03

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
12_23_28_18_14_11_19_10_16_24_12_13_15_12_16_13_12	3,269E-04	6,765E-05	3,571E-04	2,710E-04	3,083E-03	0	9,928E-05	0	2,027E-04	1,185E-03	2,601E-05
12_23_29_17_14_11_19_11_15_24_12_13_15_11_14_13_12	2,259E-08	6,280E-07	6,987E-08	6,453E-08	4,604E-08	3,255E-04	1,614E-03	0	8,389E-04	1,114E-07	8,222E-05
12_23_29_17_16_12_20_11_15_24_11_11_14_11_14_14_10	2,524E-04	4,778E-04	0	0	0	3,795E-03	8,398E-04	3,097E-02	6,540E-05	0	8,686E-04
12_23_30_18_14_11_19_10_15_23_12_13_15_11_12_13_12	0	0	0	0	0	8,249E-04	3,604E-04	0	2,437E-05	0	4,838E-05
12_24_28_17_14_12_19_11_16_23_12_13_15_11_14_13_12	0	0	0	0	0	0	1,146E-04	0	0	0	0
12_24_28_17_15_11_20_10_16_23_10_12_15_12_14_13_11	0	0	0	0	0	0	3,739E-04	0	0	0	0
12_24_30_15_16_13_20_11_15_24_11_11_14_11_14_13_10	0	0	3,239E-04	0	0	1,627E-04	0	0	3,884E-05	0	0
13_20_30_17.2_15_11_20_10_15_23_10_11_14_13_17_12_12	1,663E-04	8,606E-05	3,313E-04	0	5,279E-05	4,705E-03	1,467E-03	1,004E-02	1,886E-04	0	3,854E-04
13_21_26_15_14_11_20_10_15_22_9_11_15_13_13_12_12	0	8,216E-05	0	0	6,886E-05	2,071E-04	1,820E-03	0	0	4,174E-03	1,265E-04
13_21_29_14_14_11_20_11_14_22_10_11_16_13_14_13_11	0	0	0	0	0	0	1,092E-04	0	0	0	0
13_21_29_14_14_12_20_10_14_23_10_11_16_13_14_13_11	0	0	4,505E-04	0	0	0	1,435E-03	0	0	0	0
13_21_29_16_16_12_20_9_14_23_11_11_14_12_12_13_11	3,779E-04	0	0	0	6,264E-05	1,779E-04	3,437E-03	0	1,376E-04	0	1,469E-04
13_21_29_16_17_12_20_10_14_23_10_12_14_16_16_15_12	7,201E-05	0	0	1,031E-03	2,249E-04	2,556E-03	6,527E-03	0	5,221E-04	0	7,808E-05
13_21_29_17_13_12_20_9_16_24_10_11_14_13_14_13_10	3,618E-05	0	0	0	0	5,386E-04	1,374E-03	0	4,508E-05	0	5,621E-05

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_21_29_17_14_11_19_10_15_23_9_13_14_14_16_13_12	0	0	0	0	0	0	7,557E-04	0	0	0	0
13_21_29_18_14_12_18_11_16_24_12_13_14_11_14_13_12	2.640E-04	5,218E-03	2,958E-04	1,801E-03	3,475E-04	1,792E-04	5,462E-04	6,778E-08	2,746E-03	4,472E-03	2,913E-04
13_21_30_15_13_12_21_11_18_25_10_11_14_16_19_13_12	1.428E-04	8.304E-04	0	8,512E-04	5,136E-04	6,290E-04	5,155E-04	0	6,977E-04	0	1,814E-04
13_21_30_16_14_11_20_11_14_22_10_12_16_14_15_13_11	3,613E-05	0	0	0	0	1.875E-03	1,146E-04	0	0	0	0
13_21_30_17_14_11_19_10_16_23_9_13_14_14_16_13_11	0	0	0	0	0	0	0	0	0	0	0
13_21_32_20_13_12_20_10_16_23_9_11_14_16_18_13_11	3,282E-05	7,584E-03	9,743E-04	7,433E-03	6,896E-04	0	0	0	3,301E-03	0	3,302E-04
13_22_29_16_14_11_20_10_15_22_10_11_16_13_14_13_12	0	0	0	0	0	0	0	0	0	0	0
13_22_30_15_14_12_20_10_14_23_10_11_16_13_14_13_11	0	0	0	0	0	0	0	0	0	0	0
13_22_30_16_13_12_20_10_17_24_10_11_14_15_19_13_13	0	0	0	0	0	0	1,092E-04	0	0	0	0
13_22_30_17_14_11_20_10_14_23_10_12_14_14_16_14_11	0	0	0	0	0	9.364E-05	2.685E-04	0	1.128E-05	0	4,496E-05
13_22_30_18_14_11_20_10_14_23_10_12_14_14_16_14_11	0	0	0	0	0	9.606E-05	2.986E-04	0	1,150E-05	0	4,601E-05
13_22_31_17_15_12_20_10_15_22_11_11_16_14_14_13_11	0	0	0	0	0	0	0	0	0	0	0
13_23_28_15_14_12_19_11_16_24_12_14_15_11_14_12_13	1,051E-03	6.179E-05	3,414E-04	0	2.877E-03	3,110E-04	4,815E-03	0	4,732E-04	0	2,399E-04
13_23_28_16_14_12_19_11_15_23_12_13_15_12_14_13_11	0	0	0	0	0	0	0	0	0	0	0

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_23_28_16_14_12_19_11_16_23_12_13_15_11_15_13_11	4,370E-05	0	0	0	0	4,350E-04	1,093E-04	0	0	0	0
13_23_28_17_14_12_19_10_17_22_12_13_15_11_14_13_13	0	0	0	0	0	0	1,093E-04	0	0	0	0
13_23_28_17_14_12_19_11_16_24_12_13_15_11_16_13_12	4,719E-07	5,461E-06	0	0	7,061E-06	1,312E-05	2,303E-06	0	4,656E-06	4,377E-08	1,525E-06
13_23_29_12_14_12_20_11_16_23_12_13_15_11_14_13_12	5,814E-05	6,801E-05	0	6,074E-05	6,510E-06	5,022E-04	1,021E-03	4,476E-03	6,222E-05	0	5,420E-05
13_23_29_13_14_12_20_11_16_23_12_13_15_11_14_13_12	2,988E-04	3,553E-04	0	2,943E-04	3,329E-05	2,582E-03	5,377E-03	2,273E-02	3,133E-04	0	2,813E-04
13_23_29_13_15_12_20_11_16_23_12_13_15_11_14_13_12	1,408E-04	1,774E-04	0	1,320E-04	1,470E-05	1,238E-03	2,725E-03	9,974E-03	1,479E-04	0	1,165E-04
13_23_29_15_14_12_19_11_16_24_12_13_15_11_14_13_12	9,068E-09	1,507E-03	3,580E-04	0	1,167E-04	6,921E-08	3,791E-07	0	1,904E-04	2,702E-07	5,651E-05
13_23_29_15_14_12_20_11_15_24_12_13_15_11_14_13_13	5,053E-03	7,880E-04	0	2,227E-03	2,542E-04	2,013E-03	1,380E-03	0	8,898E-04	0	9,131E-04
13_23_29_15_14_13_19_11_16_23_12_13_16_12_14_13_11	0	0	0	0	0	0	1,146E-04	0	0	0	0
13_23_29_16_13_12_19_11_16_24_12_14_15_11_14_13_11	1,005E-03	2,428E-03	6,505E-04	0	2,998E-03	7,868E-04	3,606E-03	0	1,423E-03	0	4,756E-04
13_23_29_16_14_11_18_11_18_24_12_13_14_11_13_13_11	0	5,867E-05	0	0	1,826E-04	0	0	0	2,040E-05	0	2,727E-05
13_23_29_16_14_11_19_11_15_24_12_13_15_12_15_13_12	5,967E-05	1,496E-03	3,094E-04	2,349E-04	1,177E-04	0	0	0	8,809E-04	0	1,093E-04
13_23_29_16_14_11_19_11_17_23_12_14_14_11_12_13_12	2,577E-04	3,158E-03	0	1,040E-03	5,186E-04	1,522E-04	5,720E-04	9,430E-04	2,552E-03	1,069E-03	3,556E-04
13_23_29_16_14_12_19_11_16_24_12_13_15_11_13_13_12	0	1,002E-09	0	0	0	0	6,129E-03	0	2,248E-05	0	0

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_23_29_16_14_12_20_11_15_24_12_13_14_11_15_13_12	0	0	0	0	6,049E-05	0	1,146E-03	0	7,519E-06	0	0
13_23_29_16_14_12_20_11_15_24_12_13_15_11_15_13_12	1,004E-07	6,011E-07	5,725E-08	5,636E-07	9,640E-06	1,992E-08	4,948E-04	0	1,989E-05	3,501E-07	4,389E-08
13_23_29_16_14_12_20_11_16_25_12_13_15_11_14_12_12	7,736E-04	4,373E-04	0	0	1,075E-04	4,517E-03	1,308E-02	4,271E-02	9,269E-04	2,490E-08	9,132E-04
13_23_29_16_14_13_19_11_14_24_12_13_15_11_15_13_12	1,537E-04	7,115E-05	3,932E-04	1,888E-03	1,129E-03	5,396E-04	5,463E-04	0	4,718E-04	0	2,831E-03
13_23_29_17_13_12_19_11_16_24_12_13_14_11_14_13_13	5,190E-03	8,967E-04	1,303E-03	2,349E-04	8,152E-03	1,187E-03	1,464E-03	0	1,106E-03	1,127E-03	5,114E-04
13_23_29_17_13_13_19_11_15_24_12_13_15_11_15_13_13	0	3,049E-04	0	1,008E-03	5,424E-05	0	0	0	6,210E-04	0	0
13_23_29_17_14_11_18_10_15_24_12_13_14_11_11_13_12	5,290E-03	1,402E-03	6,431E-04	5,085E-04	4,037E-03	1,630E-03	8,228E-03	0	1,576E-03	2,230E-03	9,790E-04
13_23_29_17_14_11_19_11_15_23_12_13_15_11_14_13_12	2,564E-04	5,253E-03	6,822E-04	5,049E-04	6,955E-04	1,795E-04	3,277E-04	0	3,098E-03	1,131E-03	1,010E-03
13_23_29_17_14_11_19_11_16_24_12_13_15_11_14_12_12	1,198E-02	6,138E-05	3,240E-04	8,154E-04	3,481E-04	1,496E-03	3,656E-03	0	9,611E-04	1,310E-03	6,752E-04
13_23_29_17_14_11_19_11_16_24_12_13_15_11_14_13_12	1,144E-04	8,580E-07	3,579E-06	1,469E-06	5,361E-07	7,667E-03	3,059E-07	3,396E-08	1,065E-06	2,552E-06	2,705E-04
13_23_29_17_14_11_19_11_17_23_12_13_15_11_14_13_12	1,729E-08	5,002E-07	8,263E-08	3,453E-08	3,934E-08	4,814E-08	3,279E-04	0	2,028E-07	1,114E-07	5,827E-08
13_23_29_17_14_12_18_11_16_24_12_13_14_11_14_13_12	1,774E-03	2,382E-04	3,253E-04	1,100E-03	4,291E-04	2,641E-03	1,082E-02	3,541E-02	7,310E-04	1,081E-03	1,072E-03
13_23_29_17_14_12_18_11_16_25_12_14_15_11_13_13_11	0	0	0	0	0	0	0	0	0	0	0
13_23_29_17_14_12_18_11_17_25_12_14_15_11_13_14_12	0	0	0	0	0	0	0	0	0	0	0

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IVA	QCA	RES	SAG
13_23_29_17_14_12_19_10_15_23_12_13_15_11_14_14_12	0	0	0	0	0	0	3,181E-04	0	0	0	0
13_23_29_17_14_12_19_10_15_24_12_13_14_11_16_13_12	3,118E-05	5,843E-05	0	0	0	0	0	0	1,171E-04	0	5,216E-05
13_23_29_17_14_12_19_10_15_24_12_13_15_11_14_13_10	3,724E-04	4,396E-03	3,097E-04	7,508E-04	3,406E-03	3,033E-04	2,420E-04	0	1,797E-03	3,501E-03	1,763E-04
13_23_29_17_14_12_19_10_16_24_11_13_15_11_14_13_12	1.700E-03	8,307E-03	1,339E-02	4,811E-03	5,064E-03	2,415E-03	5,978E-03	0	3,818E-03	8,228E-03	8,287E-03
13_23_29_17_14_12_19_10_17_24_12_12_15_11_14_13_11	0	0	0	0	0	0	1,146E-04	0	0	0	0
13_23_29_17_14_12_19_11_15_23_12_13_14_11_14_13_12	1.329E-03	1,362E-04	0	0	1,938E-03	1,625E-04	2,505E-04	5,255E-08	4,301E-04	2,554E-03	6,454E-04
13_23_29_17_14_12_19_11_15_23_12_13_15_11_11_13_11	3,275E-05	0	0	0	0	5,050E-04	2,571E-03	0	2,143E-05	0	5,224E-05
13_23_29_17_14_12_19_11_15_24_12_13_15_11_14_13_11	5.396E-05	8,319E-06	1,075E-04	3,724E-05	2,811E-05	8,222E-05	1,024E-03	0	3,022E-05	1,023E-04	1.250E-04
13_23_29_17_14_12_19_11_15_24_12_13_15_11_15_13_12	3.609E-05	1,570E-07	1,272E-08	2,583E-04	1,965E-07	2,991E-09	7,594E-08	0	5,939E-04	2,491E-03	5.333E-05
13_23_29_17_14_12_19_11_15_24_12_13_15_15_15_13_12	7.048E-04	2,431E-03	3,258E-04	5.163E-03	2,135E-03	3,262E-04	1,407E-03	0	1,376E-03	1,135E-03	8,390E-04
13_23_29_17_14_12_19_11_15_25_12_13_14_11_15_13_12	1,839E-09	0	0	0	2,100E-09	0	1,519E-03	0	3,741E-10	0	1,451E-09
13_23_29_17_14_12_19_11_16_24_12_13_15_11_14_13_11	3,405E-03	4,743E-04	3,460E-02	8.736E-03	1,366E-03	1,740E-02	1,322E-03	0	4,048E-03	1,004E-02	2,778E-02
13_23_29_17_14_12_19_11_16_24_12_13_15_11_16_13_12	3,288E-05	3,718E-04	0	5,068E-09	5,268E-04	2,853E-03	1,643E-04	0	3,291E-04	1,970E-07	1,391E-04
13_23_29_17_14_12_19_11_16_24_12_14_15_11_14_14_11	0	0	0	0	0	0	1,036E-04	0	0	0	0

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_23_29_17_14_12_19_11_16_25_12_13_15_11_14_13_12	1,120E-03	5,232E-03	5,691E-06	9,844E-04	3,485E-03	6,289E-04	2,787E-03	1,051E-07	3,327E-03	3,289E-03	1,676E-03
13_23_29_17_14_12_19_11_17_24_12_13_14_11_15_13_11	0	0	0	0	0	0	0	0	0	0	0
13_23_29_17_14_12_19_11_17_25_12_13_15_11_11_13_12	1,117E-04	0	0	0	0	1,879E-04	3,115E-03	0	1,429E-04	0	3,005E-05
13_23_29_17_14_12_19_12_16_24_12_13_15_11_14_13_11	1,189E-04	1,649E-05	1,081E-03	2,946E-04	4,657E-05	6,032E-04	1,179E-03	0	1,334E-04	3,599E-04	9,112E-04
13_23_29_17_14_12_20_11_15_23_12_13_15_11_14_13_12	4,217E-06	2,288E-05	2,596E-06	2,081E-05	3,671E-05	1,932E-06	2,970E-06	0	1,129E-05	1,559E-07	1,600E-06
13_23_29_17_14_12_20_11_15_24_12_13_15_11_14_13_12	5,400E-04	3,644E-03	3,258E-04	3,087E-03	4,950E-03	3,501E-04	3,977E-04	0	1,855E-03	0	2,108E-04
13_23_29_17_14_12_20_11_16_25_12_13_15_11_14_12_12	8,233E-05	5,251E-05	0	4,238E-08	1,157E-05	5,057E-04	7,339E-04	6,326E-03	9,602E-05	1,444E-07	1,123E-04
13_23_29_17_14_13_19_11_15_25_12_14_14_12_14_13_12	0	0	0	0	0	0	2,239E-04	0	9,921E-05	0	0
13_23_29_17_14_13_19_11_16_24_12_13_15_11_14_13_11	3,332E-05	1,790E-03	3,142E-04	7,834E-05	2,675E-04	1,715E-04	1,226E-05	0	1,551E-04	1,040E-04	2,675E-04
13_23_29_17_15_11_18_11_15_24_12_13_14_11_13_13_12	1,494E-03	6,154E-05	9,325E-03	1,475E-03	2,409E-04	2,323E-03	8,208E-05	0	8,843E-04	1,132E-03	6,298E-03
13_23_29_17_15_12_18_11_15_23_11_13_15_11_13_13_13	0	0	0	0	0	0	0	0	2,046E-05	0	2,487E-05
13_23_29_17_15_12_18_11_16_23_12_14_15_11_15_13_12	6,343E-04	4,389E-03	0	5,293E-04	3,623E-04	4,813E-04	7,787E-05	0	4,758E-04	0	2,201E-04
13_23_29_17_15_12_19_10_16_24_12_13_15_11_14_13_12	6,588E-07	2,934E-06	6,268E-06	1,952E-06	1,801E-06	1,190E-06	2,178E-06	0	1,466E-06	3,365E-06	3,579E-06
13_23_29_17_15_12_19_11_16_24_12_14_14_11_14_13_13	1,464E-04	7,107E-05	3,401E-04	0	5,009E-04	3,036E-09	1,694E-03	0	1,012E-04	0	0

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_23_29_17_15_12_19_11_16_25_12_13_15_12_13_14_12	3,271E-05	4,827E-04	0	2,572E-04	0	0	7,780E-05	0	4,433E-04	1,131E-03	1,293E-04
13_23_29_18_14_12_18_10_15_22_12_13_15_11_14_13_13	0	0	0	0	0	0	0	0	0	0	4,014E-10
13_23_29_18_14_12_19_10_15_24_12_13_15_11_14_13_13	2,041E-03	4,076E-05	3,937E-04	5,297E-06	3,454E-06	3,487E-03	1,803E-04	5,225E-08	2,093E-04	1,305E-05	1,865E-03
13_23_29_18_14_12_19_10_15_24_13_13_15_11_14_13_11	3,111E-04	1,761E-04	0	3,614E-04	3,366E-09	3,106E-04	7,833E-05	0	1,063E-04	2,214E-08	1,242E-04
13_23_29_18_14_12_19_10_16_24_12_13_15_12_13_13_12	1,135E-04	4,458E-03	0	1,165E-03	6,554E-05	0	9,019E-05	0	3,501E-04	0	3,007E-05
13_23_29_18_14_12_19_11_15_24_12_13_15_11_14_13_11	2,308E-03	2,436E-04	6,504E-07	5,235E-04	1,189E-03	4,962E-04	2,903E-03	0	8,289E-04	3,563E-03	1,180E-03
13_23_29_18_14_12_19_11_15_24_12_13_15_11_15_14_12	0	0	0	0	0	0	2,042E-03	0	4,525E-05	0	5,899E-05
13_23_29_18_14_12_19_11_16_23_12_13_15_11_13_13_12	0	4,713E-09	0	0	0	0	4,339E-09	0	1,957E-05	0	0
13_23_29_18_14_12_19_11_16_24_12_13_15_11_14_13_12	1,313E-04	3,920E-03	1,354E-03	7,128E-03	9,017E-04	1,422E-04	3,370E-05	1,226E-07	1,543E-03	2,640E-06	5,369E-04
13_23_29_18_14_12_19_11_16_24_12_13_15_11_14_13_13	2,390E-04	3,223E-02	1,319E-03	3,874E-03	1,981E-03	1,585E-06	8,231E-05	0	2,534E-03	4,705E-08	6,486E-04
13_23_29_18_14_12_19_11_16_24_12_14_15_11_14_14_11	0	0	0	0	0	0	1,060E-05	0	0	0	0
13_23_29_18_14_12_20_10_15_24_12_13_15_11_14_13_13	4,858E-05	7,353E-03	2,108E-06	1,034E-03	6,712E-04	1,638E-05	1,311E-06	0	9,184E-04	2,214E-03	2,589E-04
13_23_29_18_14_12_20_11_16_24_12_13_15_11_14_13_11	4,619E-04	4,371E-03	1,677E-03	9,215E-03	1,377E-03	4,688E-04	4,648E-04	0	2,998E-03	1,670E-06	5,575E-04
13_23_29_18_15_11_19_12_16_24_12_13_15_11_14_13_11	3,688E-04	9,555E-04	1,549E-03	3,462E-04	3,344E-04	4,555E-04	1,171E-04	0	2,943E-04	2,229E-03	2,759E-03

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_23_29_18_15_12_19_11_15_23_12_13_14_10_15_13_12	0	0	0	0	0	0	0	0	0	0	0
13_23_29_18_15_12_19_11_15_23_12_13_15_11_14_13_12	2,654E-04	2,022E-04	0	9,362E-09	4,006E-09	3,596E-04	1,462E-03	0	6,977E-05	0	2,883E-05
13_23_29_19_13_12_19_11_16_24_12_13_15_11_14_13_11	1,943E-03	6,784E-04	1,439E-07	2,336E-04	3,919E-03	1,544E-03	2,246E-03	0	1,615E-03	1,127E-03	9,597E-04
13_23_29_19_14_11_19_11_16_25_12_13_14_11_15_13_12	3,970E-05	0	0	0	5,402E-05	5,236E-04	7,555E-04	2,949E-03	9,471E-05	0	2,737E-05
13_23_29_19_14_11_19_12_16_24_12_13_15_11_14_13_11	4,335E-03	8,647E-03	1,035E-02	6,650E-03	6,608E-03	2,690E-03	2,946E-03	0	3,332E-03	7,708E-03	1,230E-02
13_23_29_19_14_12_19_11_15_24_12_13_15_11_14_13_11	8,820E-05	7,030E-04	0	2,503E-04	2,102E-03	4,767E-06	1,779E-04	0	4,647E-04	3,329E-05	3,671E-05
13_23_29_19_14_12_19_11_15_24_12_13_15_11_14_13_12	8,918E-07	3,624E-03	0	1,230E-03	1,409E-04	5,140E-08	2,053E-06	0	8,257E-04	1,547E-07	1,075E-04
13_23_29_19_15_11_18_11_16_24_12_12_14_11_14_14_12	3,131E-05	1,298E-04	0	0	0	0	0	0	1,054E-04	0	2,614E-05
13_23_29_20_14_12_19_11_17_24_12_13_15_11_14_13_12	0	3,826E-09	0	0	1,130E-09	0	0	0	2,843E-09	0	0
13_23_30_14_16_12_21_10_15_23_9_11_15_13_16_12_11	5,502E-03	2,103E-03	4,035E-03	5,569E-03	2,242E-03	2,522E-03	6,996E-03	0	4,178E-03	8,532E-03	6,591E-03
13_23_30_15_15_12_20_10_16_25_11_11_14_11_14_13_10	0	0	0	0	0	0	0	0	0	0	0
13_23_30_16_12_12_19_10_15_23_12_13_15_11_15_13_13	0	6,446E-05	0	1,109E-03	5,015E-04	0	1,831E-03	0	5,607E-04	0	0
13_23_30_16_14_11_20_11_16_24_12_13_15_11_14_13_12	0	0	0	0	0	1,792E-04	2,577E-04	0	0	0	1,101E-09
13_23_30_17_14_11_18_10_16_24_12_14_14_13_14_13_10	7,440E-04	7,134E-04	3,099E-04	9,833E-04	4,593E-04	1,553E-04	0	0	2,304E-03	1,303E-03	5,045E-04

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_23_30_17_14_11_19_10_16_23_12_13_15_11_14_13_12	3,292E-05	0	0	0	0	0	8,185E-05	0	6,146E-05	0	0
13_23_30_17_14_12_18_11_17_24_12_13_16_11_14_13_12	1,573E-04	0	0	0	0	2,174E-04	6,014E-04	0	0	0	0
13_23_30_17_14_12_19_11_15_25_12_12_14_11_14_13_12	3,782E-05	0	0	0	0	0	2,038E-04	0	0	0	0
13_23_30_17_14_12_20_10_15_25_12_13_15_9_14_12_12	1,469E-04	0	3,578E-04	0	0	3,865E-03	2,244E-03	0	1,362E-04	0	1,369E-04
13_23_30_17_14_13_19_11_16_24_12_13_15_11_15_13_12	2,442E-05	5,333E-04	4,078E-04	2,008E-03	2,787E-04	5,536E-05	2,998E-05	0	5,136E-04	0	1,689E-04
13_23_30_17_15.2_12_19_11_15_24_13_14_15_12_13_13_12	0	0	0	0	0	0	0	0	0	0	0
13_23_30_18_14_12_19_10_14_23_12_13_15_12_14_13_12	0	0	0	0	0	0	0	0	0	0	0
13_23_30_18_14_12_19_10_15_25_12_13_15_11_14_13_12	1,399E-03	5,611E-05	0	0	2,861E-04	3,065E-03	1,557E-02	4,743E-02	5,660E-04	1,176E-03	7,590E-04
13_23_30_19_14_11_19_11_15_24_12_13_15_11_14_13_11	0	1,070E-03	0	0	0	0	0	0	0	0	0
13_23_31_16_15_12_20_11_17_24_11_11_14_10_14_13_10	4,168E-05	0	0	0	7,233E-05	0	7,342E-04	0	0	0	1,043E-04
13_23_31_17_14_12_19_11_15_23_12_13_15_12_13_13_12	0	0	4,771E-04	0	0	0	0	0	0	0	0
13_23_31_17_14_13_19_11_16_24_12_13_15_11_15_13_12	1,329E-04	2,697E-03	2,167E-03	1,089E-02	1,420E-03	2,775E-04	5,399E-04	0	2,630E-03	0	9,019E-04
13_23_31_18_14_12_19_10_15_23_12_13_15_11_12_11_12	0	0	0	0	0	6,219E-04	7,524E-04	0	1,741E-05	0	3,433E-05
13_23_31_18_14_12_19_10_15_23_12_13_15_11_12_13_12	0	0	0	0	0	9,562E-04	3,165E-04	0	2,571E-05	0	3,255E-05

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_23_31_18_15_11_19_10_14_24_9_11_15_14_15_12_12	4,130E-04	1,999E-04	1,023E-03	0	5,449E-05	1,276E-03	3,350E-03	0	8,208E-05	0	2,732E-05
13_24_29_16_14_10_19_11_16_24_12_13_15_11_14_13_12	1,003E-04	5,641E-03	0	7,884E-04	3,506E-04	0	0	0	1,321E-03	0	2,491E-05
13_24_29_16_14_12_19_10_15_24_12_14_15_11_14_13_12	0	3,807E-03	0	0	0	0	8,059E-04	0	3,012E-04	0	5,922E-05
13_24_29_16_14_12_19_10_15_24_12_14_16_11_15_13_11	0	0	0	0	0	0	2,237E-04	0	0	0	0
13_24_29_16_14_12_19_11_17_24_12_14_15_12_15_13_12	0	0	0	0	0	0	0	0	0	0	0
13_24_29_16_14_12_20_10_15_22_12_13_15_11_14_13_11	1,533E-04	3,730E-03	1,466E-04	2,281E-04	2,128E-04	0	3,752E-05	0	3,499E-04	0	2,530E-04
13_24_29_16_15_13_19_11_15_24_12_14_15_11_13_13_12	0	0	0	0	0	0	9,970E-04	0	0	0	3,153E-05
13_24_29_17_13_12_19_11_16_25_12_13_15_11_12_13_12	3,051E-03	9,128E-04	4,103E-02	8,998E-03	7,508E-04	1,249E-02	8,855E-04	0	3,619E-03	1,508E-02	2,174E-02
13_24_29_17_14_12_19_10_15_24_12_13_15_11_15_13_10	2,149E-08	2,994E-07	9,505E-08	7,321E-08	3,648E-07	3,220E-04	1,458E-05	0	2,350E-07	1,665E-07	1,914E-08
13_24_29_17_14_12_19_10_16_24_11_13_15_11_14_13_12	5,070E-03	7,655E-04	5,708E-02	1,101E-02	1,895E-03	2,623E-02	1,984E-03	0	5,937E-03	2,283E-02	4,986E-02
13_24_29_17_14_12_19_11_15_24_12_13_15_11_14_13_12	5,683E-08	6,813E-05	1,113E-07	1,074E-07	6,287E-05	6,950E-08	4,378E-08	0	3,770E-04	0	2,255E-07
13_24_29_17_14_12_19_11_15_25_12_13_15_11_14_13_13	2,604E-08	3,333E-07	0	1,381E-07	8,111E-08	0	1,041E-04	0	5,261E-05	0	9,230E-08
13_24_29_17_15_12_19_10_15_22_12_13_15_11_14_13_11	1,820E-04	5,597E-03	1,728E-04	2,740E-04	2,040E-04	0	4,406E-05	0	4,251E-04	0	2,713E-04
13_24_29_17_15_12_19_11_15_24_12_13_14_11_16_13_11	4,375E-05	5,960E-04	0	0	8,091E-04	0	0	0	2,726E-04	0	0

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_24_29_18_14_12_19_10_16_24_12_13_15_11_14_13_12	4.968E-03	1.654E-03	1,127E-02	4,505E-03	3,321E-03	9,481E-03	6,209E-04	0	2,109E-03	2,273E-03	9.451E-03
13_24_29_18_14_12_19_11_15_24_12_13_15_11_14_13_13	6,275E-05	1,125E-04	9,195E-08	4,933E-04	1,088E-04	2,984E-07	3,296E-08	0	1,855E-03	0	2,256E-04
13_24_29_18_14_12_19_11_15_25_12_13_15_11_14_13_12	3,210E-04	1,541E-03	0	5,116E-09	3,006E-04	2,468E-08	1,794E-08	3,153E-07	4,789E-04	2,211E-08	2.688E-04
13_24_29_18_14_12_19_11_16_24_12_13_14_11_14_13_12	4,522E-03	2.160E-04	8,962E-04	1,887E-03	3,901E-03	5,650E-03	6,540E-03	8,109E-04	1,603E-03	1,026E-03	1,743E-03
13_24_29_18_15_11_21_9_15_23_10_11_14_15_16_12_11	0	0	0	0	0	0	0	0	0	0	0
13_24_30_14_16_12_21_10_15_23_9_11_15_13_16_12_11	6.108E-03	1,622E-03	3.861E-03	4,790E-03	1,662E-03	2,640E-03	2,324E-03	0	3,517E-03	7.806E-03	6,078E-03
13_24_30_15_15_12_21_11_17_23_9_11_15_12_14_12_11	1.653E-02	7,214E-04	3,121E-03	8,128E-04	1,090E-02	4.312E-03	3,581E-03	0	2.964E-03	7.066E-03	5,552E-03
13_24_30_16_14_11_19_11_16_24_12_14_15_12_14_13_12	6,547E-05	0	3,402E-04	0	0	1,180E-03	1,637E-04	0	2,142E-05	0	1,303E-04
13_24_30_17_13_13_19_11_15_25_12_13_15_11_14_12_13	0	1,455E-04	0	0	0	0	0	0	2,366E-05	0	0
13_24_30_19_14_13_19_10_14_25_12_13_15_11_14_13_12	0	0	0	0	0	0	1,469E-03	0	4,506E-05	0	9,949E-05
13_25_29_17_14_12_19_10_16_23_12_13_15_11_16_11_12	0	0	0	0	0	0	0	0	0	0	0
13_25_29_18_14_12_18_11_16_24_11_13_15_11_15_13_12	0	0	0	0	0	0	2,238E-04	0	2,478E-05	0	3,156E-05
13_25_29_18_14_12_19_11_16_23_12_13_14_11_15_13_12	2,216E-09	1,391E-09	0	0	1,289E-09	3,686E-09	7,058E-04	0	2,781E-09	0	0
13_25_29_18_15_12_19_11_16_23_12_13_14_11_15_13_12	0	0	0	0	0	0	5,180E-04	0	0	0	0

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
13_25_29_19_14_11_19_11_12_24_12_13_15_11_14_13_13	0	0	3,423E-04	0	0	0	0	0	0	0	2,607E-05
13_25_30_16_15_11_20_10_15_25_12_11_14_11_14_13_11	0	0	0	0	0	0	0	0	0	0	0
13_26_28_17_14_13_20_11_16_23_12_13_15_11_14_13_12	0	0	0	0	0	0	1,146E-04	0	0	0	0
14_20_27_17_17_12_21_11_14_23_10_11_15_12_13_13_12	3,133E-05	0	0	0	0	0	0	0	2,474E-04	0	1,879E-04
14_21_29_15_15_13_21_10_14_22_10_11_14_12_13_14_12	0	0	0	0	0	0	1,146E-04	0	0	0	0
14_21_31_17_15_11_20_11_14_24_10_11_16_11_14_14_11	0	0	0	0	0	0	1,146E-04	0	0	0	0
14_22_31_14_15_12_21_9_16_23_9_11_14_13_15_12_11	3,463E-04	8,300E-04	0	2,053E-03	0	4,981E-04	1,642E-04	0	7,681E-04	1,185E-03	4,499E-04
14_22_31_16_13_11_20_10_17_24_10_11_14_17_18_13_11	0	0	0	0	0	0	1,092E-04	0	5,462E-05	0	0
14_22_32_15_15_11_20_10_15_23_10_12_14_15_15_15_11	0	0	0	0	0	4,280E-04	1,744E-03	0	0	0	0
14_23_29_17_14_12_19_11_16_24_12_13_15_11_13_13_11	0	0	0	0	0	0	0	0	2,039E-05	0	0
14_23_30_16_14_11_19_11_15_22_12_13_15_11_14_13_12	7,485E-09	0	0	0	3,106E-08	0	9,443E-05	0	3,858E-09	0	3,279E-09
14_23_30_16_14_11_21_11_15_25_9_11_15_14_17_12_12	0	0	0	1,235E-03	0	1,709E-04	0	0	8,246E-04	0	2,488E-05
14_23_30_16_14_12_18_11_16_25_12_14_15_11_13_13_12	0	0	0	0	0	0	1,146E-04	0	0	0	0
14_23_30_16_14_12_19_11_15_24_12_13_15_11_14_13_12	1,208E-02	4,006E-04	1,094E-03	1,441E-03	1,980E-02	2,158E-03	5,756E-03	0	2,976E-03	1,137E-03	2,441E-03

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
14_23_30_16_14_12_19_11_15_24_12_13_15_11_14_13_13	1,458E-04	5,112E-06	8,950E-06	1,784E-05	2,411E-04	2,610E-05	3,970E-04	0	3,592E-05	1,233E-05	2,714E-05
14_23_30_16_14_12_19_11_16_24_12_13_15_12_14_14_12	1,796E-04	8,748E-05	3,505E-04	0	6,882E-05	4,789E-03	1,697E-03	1,026E-02	2,153E-04	0	4,018E-04
14_23_30_16_14_12_20_11_15_24_12_13_15_11_15_13_12	1,460E-06	5,572E-08	7,269E-08	1,736E-07	2,247E-06	2,398E-07	6,122E-07	0	3,158E-07	6,653E-08	2,334E-07
14_23_30_17_13_12_19_11_15_24_12_13_15_11_14_13_12	3,575E-04	2,191E-04	2,387E-04	4,610E-07	4,791E-04	6,090E-07	3,502E-03	0	9,680E-05	2,218E-07	2,266E-05
14_23_30_17_13_12_19_11_15_24_12_13_15_11_15_13_12	9,584E-05	3,782E-05	4,007E-05	0	8,902E-05	6,931E-09	7,783E-04	0	1,518E-05	0	3,275E-06
14_23_30_17_14_11_18_10_15_24_12_13_14_12_14_13_14	8,814E-05	0	1,516E-04	0	0	2,715E-04	4,088E-03	0	8,199E-05	0	3,923E-05
14_23_30_17_14_11_18_10_15_24_12_13_14_12_14_14_13	8,940E-05	0	2,024E-04	0	0	2,656E-04	4,456E-03	0	7,340E-05	0	4,472E-05
14_23_30_17_14_13_19_11_15_24_12_13_15_11_16_13_11	0	0	0	0	0	0	2,093E-04	0	0	0	0
14_23_30_17_15_12_19_10_16_24_12_13_15_11_15_13_13	2,017E-04	2,815E-03	0	5,291E-04	3,214E-03	6,837E-04	4,031E-03	0	1,140E-03	2,376E-03	9,541E-04
14_23_30_18_14_11_17_9_15_24_12_13_14_11_14_13_11	0	0	0	0	0	0	2,135E-04	0	0	0	0
14_23_30_18_14_12_19_10_16_23_12_13_15_12_13_13_12	0	0	0	0	0	0	0	0	0	0	0
14_23_30_19_14_12_19_11_15_24_12_13_15_11_14_14_12	0	0	0	0	0	0	3,275E-04	0	0	0	0
14_23_30_20_14_12_19_10_17_23_12_13_15_11_11_13_12	0	0	0	0	0	1,976E-04	9,368E-04	0	0	0	0
14_23_31_14_15_13_20_10_15_26_11_11_14_11_14_13_10	0	0	0	0	0	2,172E-04	3,226E-04	0	2,596E-05	0	0

Tableau E.1 (suite)

Haplotype 17 STR-Y	BSL	BEA	CHA	CDB	CDS	CNO	GAS	IMA	QCA	RES	SAG
14_23_31_17_14_12_19_11_15_24_12_13_15_12_15_13_11	0	0	0	0	0	0	0	0	0	0	0
14_23_31_17_14_9_20_10_17_23_9_11_15_12_15_12_11	4,303E-04	1,066E-03	0	2,840E-04	5,958E-05	2,071E-04	5,010E-04	0	1,719E-04	0	5,206E-05
14_23_32_16_14_12_18_11_17_23_12_13_15_11_15_13_12	4,692E-06	0	0	0	0	0	2,394E-04	0	4,159E-05	0	1,664E-05
14_23_32_17_14_11_18_11_15_24_12_13_14_11_14_13_12	1,833E-04	5,575E-05	0	0	1,557E-04	4,502E-04	0	0	2,902E-04	1,078E-03	4,746E-05
14_23_33_16_14_12_18_11_17_23_12_13_15_11_15_13_12	2,686E-05	0	0	0	0	0	1,527E-03	0	2,403E-04	0	9,812E-05
14_24_30_14_14_11_22_9_16_22_9_11_14_13_19_12_13	3,442E-05	0	3,239E-04	2,345E-04	0	0	0	0	3,756E-04	1,185E-03	1,370E-04
14_24_30_17_15_12_19_10_15_24_12_13_15_11_14_13_12	6,106E-05	1,602E-05	0	2,402E-04	1,642E-05	9,570E-05	1,248E-04	0	3,096E-04	0	1,329E-04
14_24_30_17_15_12_19_11_16_24_12_13_15_11_14_13_12	3,927E-05	6,563E-06	0	7,389E-05	1,425E-05	5,524E-05	3,274E-04	0	1,222E-04	0	5,438E-05
14_24_31_17_15_12_19_11_16_24_12_13_15_11_14_13_12	2,793E-04	3,854E-05	0	4,184E-04	8,666E-05	3,170E-04	1,040E-03	0	7,425E-04	0	3,165E-04
15_23_31_17_13_12_19_11_16_24_12_13_15_11_15_13_12	5,844E-05	3,766E-05	4,552E-05	0	8,550E-05	0	6,639E-04	0	1,409E-05	0	3,603E-06
9_23_25_18_14_12_19_10_16_23_12_13_15_11_16_13_13	5,716E-03	5,399E-03	1,566E-02	9,788E-03	6,465E-03	1,077E-02	2,610E-03	0	4,922E-03	1,116E-02	1,974E-02
9_23_25_18_14_12_19_11_16_23_12_13_15_11_16_13_13	1,225E-02	2,059E-03	1,145E-03	5,768E-04	2,441E-03	4,720E-03	2,013E-03	8,918E-04	2,204E-03	2,537E-03	1,670E-03

ANNEXE F

FRÉQUENCES DES HAPLOTYPES DU CHROMOSOME Y OBSERVÉS AVEC 20 STR-Y DANS LES DIFFÉRENTES RÉGIONS ANALYSÉES

Les haplotypes Y ont été définis par l'ensemble des allèles obtenus pour 20 STR-Y dont l'ordre est le suivant : DYS389I, DYS635, DYS389II, DYS460, DYS458, DYS19, YGATAH4, DYS448, DYS391, DYS456, DYS390, DYS438, DYS392, DYS437, DYS385a, DYS385b, DYS393, DYS439, DYS481, DYS533. Les allèles sont séparés par le symbole « _ ». La fréquence des haplotypes observés chez les personnes génotypées seulement (c.-à-d. excluant ceux générés par les simulations lors de l'imputation) est donnée pour les cinq régions retenues pour analyses (voir CHAPITRE II pour les détails) : Bas-Saint-Laurent (BSL), Côte-du-Sud (CDS), Côte-Nord (CNO), Gaspésie (GAS) et Îles-de-la-Madeleine (IMA).

Tableau F.1. Fréquences des haplotypes du chromosome Y observés avec 20 STR-Y dans la population canadienne-française calculées pour 5 régions québécoises d'après le modèle généalogico-moléculaire développé dans ce projet.

Haplotype 20 STR-Y	BSL	CDS	CNO	GAS	IMA
11_23_28_10_16_15_13_20_11_16_25_11_11_14_11_14_13_10_23_12	0	0	0	1,012E-04	0
12_21_28_10_15_14_11_20_10_14_22_10_11_16_13_14_13_13_26_11	0	0	0	1,146E-04	0
12_21_28_10_15_14_11_20_10_14_23_10_11_16_14_14_13_11_26_10	6.010E-03	9,464E-03	1,749E-03	2,323E-03	0
12_21_28_10_18_15_11_21_10_15_22_10_11_16_12_14_14_11_21_9	0	0	0	0	0
12_21_28_11_16_14_11_19_10_14_23_10_11_16_14_14_13_11_25_11	2.164E-03	2.264E-03	6,917E-04	7,663E-04	0
12_21_28_11_16_14_11_20_10_15_23_10_11_15_14_14_13_11_23_11	0	0	0	1,146E-04	0
12_21_29_10_16_15_12_21_10_15_22_10_11_16_14_14_14_11_21_9	3,558E-05	0	1,441E-03	3,389E-03	0
12_21_29_11_17_14_10_18_10_14_24_10_11_14_11_11_15_13_25_11	4.026E-04	1.031E-04	0	4.679E-03	0
12_21_29_11_17_15_12_21_10_15_22_10_11_16_13_15_14_12_23_10	0	0	0	1,146E-04	0
12_22_28_10_15_14_11_20_10_14_23_10_11_16_14_16_13_12_24_11	0	1,192E-09	0	0	0
12_22_28_10_16_14_11_20_10_14_22_10_11_16_13_14_14_12_24_11	1,782E-03	1,797E-03	3,346E-04	1,513E-03	0
12_22_29_10_15_14_11_19_10_15_22_10_11_16_13_13_13_11_25_10	0	0	0	1,146E-04	0
12_23_28_10_14_14_11_20_10_14_23_10_11_16_14_15_13_11_25_11	5,369E-04	5,314E-03	9,056E-04	4,110E-03	0
12_23_28_10_15_14_11_20_10_14_22_10_11_16_13_14_13_11_25_11	0	0	0	1,146E-04	0
12_23_28_11_15_14_12_19_11_16_24_12_13_15_11_14_14_12_23_12	4,982E-04	7,156E-05	6,675E-03	2,607E-03	1,590E-02
12_23_28_11_15_14_12_19_11_16_24_12_13_15_11_14_15_12_23_12	3,535E-04	5,778E-05	5,416E-03	2,657E-03	1,304E-02
12_23_28_11_16_14_11_20_10_15_22_10_11_16_13_14_13_11_25_11	0	0	0	2,883E-04	0
12_23_28_11_17_14_12_19_11_16_25_12_13_15_10_11_13_12_22_11	1,827E-02	1,047E-02	8,222E-03	6,943E-03	0
12_23_29_11_17_14_11_19_11_15_24_12_13_15_11_14_13_12_22_13	1,380E-09	0	2,943E-04	1,554E-03	0
12_23_29_11_17_16_12_20_11_15_24_11_11_14_11_14_14_10_22_12	2,350E-04	0	3,542E-03	7,912E-04	2,888E-02

Tableau F.1 (suite)

Haplotype 20 STR-Y	BSL	CDS	CNO	GAS	IMA
12_23_30_11_18_14_11_19_10_15_23_12_13_15_11_12_13_12_22_12	0	0	7,248E-04	3,550E-04	0
12_24_28_10_17_15_11_20_10_16_23_10_12_15_12_14_13_11_26_12.1	0	0	0	3,517E-04	0
12_24_28_11_17_14_12_19_11_16_23_12_13_15_11_14_13_12_22_12	0	0	0	1,146E-04	0
13_20_30_11_17.2_15_11_20_10_15_23_10_11_14_13_17_12_12_25_11	1,489E-04	4,815E-05	4,109E-03	1,357E-03	8,863E-03
13_21_26_10_15_14_11_20_10_15_22_9_11_15_13_13_12_12_23_11	0	6,609E-05	2,011E-04	1,755E-03	0
13_21_29_10_14_14_11_20_11_14_22_10_11_16_13_14_13_11_26_11	0	0	0	1,076E-04	0
13_21_29_10_14_14_12_20_10_14_23_10_11_16_13_14_13_11_25_11	0	0	0	1,188E-03	0
13_21_29_10_14_14_12_20_10_14_23_10_11_16_13_14_13_11_26_11	0	0	0	2,309E-04	0
13_21_29_11_16_16_12_20_9_14_23_11_11_14_12_12_13_11_22_12	3,550E-04	5,829E-05	1,650E-04	3,222E-03	0
13_21_29_11_17_13_12_20_9_16_24_10_11_14_13_14_13_10_25_11	3,357E-05	0	5,004E-04	1,285E-03	0
13_21_29_11_17_14_11_19_10_15_23_9_13_14_14_16_13_12_23_12	0	0	0	7,148E-04	0
13_21_29_12_16_17_12_20_10_14_23_10_12_14_16_16_15_12_25_13	6,731E-05	2,027E-04	2,344E-03	6,092E-03	0
13_21_30_11_16_14_11_20_11_14_22_10_12_16_14_15_13_11_26_11	3,364E-05	0	1,761E-03	1,146E-04	0
13_21_30_11_17_14_11_19_10_16_23_9_13_14_14_16_13_11_23_13	0	0	0	0	0
13_21_30_9_15_13_12_21_11_18_25_10_11_14_16_19_13_12_22_12	1,324E-04	4,637E-04	5,606E-04	4,996E-04	0
13_22_29_11_16_14_11_20_10_15_22_10_11_16_13_14_13_12_25_11	0	0	0	0	0
13_22_30_10_15_14_12_20_10_14_23_10_11_16_13_14_13_11_26_11	0	0	0	0	0
13_22_30_11_17_14_11_20_10_14_23_10_12_14_14_16_14_11_23_12	0	0	8,956E-05	2,565E-04	0
13_22_30_11_18_14_11_20_10_14_23_10_12_14_14_16_14_11_23_12	0	0	8,844E-05	2,792E-04	0
13_22_30_9_16_13_12_20_10_17_24_10_11_14_15_19_13_13_23_12	0	0	0	1,078E-04	0

Tableau F.1 (suite)

Haplotype 20 STR-Y	BSL	CDS	CNO	GAS	IMA
13_22_31_10_17_15_12_20_10_15_22_11_11_16_14_14_13_11_21_9	0	0	0	0	0
13_23_28_10_17_14_12_19_10_17_22_12_13_15_11_14_13_13_22_13	0	0	0	1,077E-04	0
13_23_28_11_16_14_12_19_11_15_23_12_13_15_12_14_13_11_23_12	0	0	0	0	0
13_23_28_11_16_14_12_19_11_16_23_12_13_15_11_15_13_11_22_12	4,313E-05	0	4,285E-04	1,076E-04	0
13_23_28_11_17_14_12_19_11_16_24_12_13_15_11_16_13_12_22_14	0	0	0	0	0
13_23_28_12_15_14_12_19_11_16_24_12_14_15_11_14_12_13_22_13	9,518E-04	2,610E-03	2,777E-04	4,472E-03	0
13_23_29_10_16_13_12_19_11_16_24_12_14_15_11_14_13_11_22_12	8,890E-04	2,669E-03	7,014E-04	3,291E-03	0
13_23_29_10_16_14_12_19_11_16_24_12_13_15_11_13_13_12_22_12	0	0	0	5,808E-03	0
13_23_29_10_16_14_13_19_11_14_24_12_13_15_11_15_13_12_22_13	1,463E-04	1,001E-03	5,010E-04	5,381E-04	0
13_23_29_10_17_13_12_19_11_16_24_12_13_14_11_14_13_13_23_13	4,675E-03	7,404E-03	1,088E-03	1,344E-03	0
13_23_29_10_17_14_12_19_11_15_23_12_13_15_11_11_13_11_22_12	2,969E-05	0	4,617E-04	2,386E-03	0
13_23_29_10_17_14_13_19_11_15_25_12_14_14_12_14_13_12_22_12	0	0	0	2,223E-04	0
13_23_29_10_18_14_12_19_11_16_24_12_13_15_11_14_13_12_23_12	8,688E-10	0	0	0	0
13_23_29_10_18_15_12_19_11_15_23_12_13_15_11_14_13_12_23_12	2,494E-04	0	3,330E-04	1,392E-03	0
13_23_29_10_19_13_12_19_11_16_24_12_13_15_11_14_13_11_22_12	1,778E-03	3,510E-03	1,472E-03	2,040E-03	0
13_23_29_11_12_14_12_20_11_16_23_12_13_15_11_14_13_12_23_12	5,441E-05	5,890E-06	4,769E-04	1,000E-03	4,225E-03
13_23_29_11_13_14_12_20_11_16_23_12_13_15_11_14_13_12_23_12	2,805E-04	2,999E-05	2,423E-03	5,199E-03	2,126E-02
13_23_29_11_13_15_12_20_11_16_23_12_13_15_11_14_13_12_23_12	1,111E-04	1,035E-05	9,747E-04	2,345E-03	7,546E-03
13_23_29_11_15_14_13_19_11_16_23_12_13_16_12_14_13_11_23_12	0	0	0	1,146E-04	0
13_23_29_11_16_14_12_20_11_15_24_12_13_14_11_15_13_12_22_13	0	5,963E-05	0	1,145E-03	0

Tableau F.1 (suite)

Haplotype 20 STR-Y	BSL	CDS	CNO	GAS	IMA
13_23_29_11_16_14_12_20_11_15_24_12_13_15_11_15_13_12_22_13	0	6.069E-06	0	4.460E-04	0
13_23_29_11_16_14_12_20_11_16_25_12_13_15_11_14_12_12_22_12	6.830E-04	9.753E-05	4.030E-03	1.183E-02	3.752E-02
13_23_29_11_17_14_11_18_10_15_24_12_13_14_11_11_13_12_19_12	4.731E-03	3.601E-03	1.471E-03	7.626E-03	0
13_23_29_11_17_14_11_19_11_16_24_12_13_15_11_14_12_12_22_13	1.089E-02	3.160E-04	1.352E-03	3.292E-03	0
13_23_29_11_17_14_11_19_11_17_23_12_13_15_11_14_13_12_22_12	0	0	0	3.227E-04	0
13_23_29_11_17_14_12_18_11_16_24_12_13_14_11_14_13_12_22_12	1.587E-03	3.816E-04	2.393E-03	9.976E-03	3.165E-02
13_23_29_11_17_14_12_18_11_17_25_12_14_15_11_13_14_12_25_12	0	0	0	0	0
13_23_29_11_17_14_12_19_10_15_23_12_13_15_11_14_14_12_26_12	0	0	0	3.116E-04	0
13_23_29_11_17_14_12_19_11_15_24_12_13_15_11_14_13_11_23_13	0	0	0	9.319E-04	0
13_23_29_11_17_14_12_19_11_15_24_12_13_15_15_15_13_12_23_13	6.511E-04	1.952E-03	2.948E-04	1.327E-03	0
13_23_29_11_17_14_12_19_11_15_25_12_13_14_11_15_13_12_23_12	0	0	0	1.466E-03	0
13_23_29_11_17_14_12_19_11_16_24_12_13_15_11_14_13_11_22_12	8.722E-09	7.206E-09	5.336E-05	9.734E-05	9.844E-08
13_23_29_11_17_14_12_19_11_16_24_12_14_15_11_14_14_11_23_13	0	0	0	1.041E-04	0
13_23_29_11_17_14_12_19_11_17_24_12_13_14_11_15_13_11_22_12	0	0	0	0	0
13_23_29_11_17_14_12_19_11_17_25_12_13_15_11_11_13_12_22_12	1.050E-04	0	1.770E-04	2.971E-03	0
13_23_29_11_17_14_12_19_12_16_24_12_13_15_11_14_13_11_22_12	7.477E-10	0	5.131E-07	1.101E-03	0
13_23_29_11_17_14_12_20_11_16_25_12_13_15_11_14_12_12_22_12	8.526E-05	1.197E-05	5.846E-04	8.894E-04	6.772E-03
13_23_29_11_17_15_12_19_10_16_24_12_13_15_11_14_13_12_23_11	0	0	0	0	0
13_23_29_11_18_14_12_18_10_15_22_12_13_15_11_14_13_13_22_12	0	0	0	0	0
13_23_29_11_18_14_12_19_11_15_24_12_13_15_11_14_13_11_21_12	2.101E-03	1.037E-03	4.508E-04	2.675E-03	0

Tableau F.1 (suite)

Haplotype 20 STR-Y	BSL	CDS	CNO	GAS	IMA
13_23_29_11_18_14_12_19_11_15_24_12_13_15_11_15_14_12_22_12	0	0	0	1,947E-03	0
13_23_29_11_18_14_12_19_11_16_24_12_14_15_11_14_14_11_23_13	0	0	0	1,006E-05	0
13_23_29_11_19_14_11_19_11_16_25_12_13_14_11_15_13_12_21_12	3,796E-05	4,858E-05	4,834E-04	7,134E-04	2,690E-03
13_23_29_11_19_14_11_19_12_16_24_12_13_15_11_14_13_11_22_11	4,227E-03	6,183E-03	2,803E-03	2,818E-03	0
13_23_29_12_16_14_12_20_11_16_25_12_13_15_11_14_12_12_22_12	1,108E-05	1,255E-06	5,260E-05	4,697E-04	4,991E-04
13_23_29_12_17_14_12_18_11_16_25_12_14_15_11_13_13_11_25_13	0	0	0	0	0
13_23_29_12_17_14_12_19_10_17_24_12_12_15_11_14_13_11_21_13	0	0	0	1,146E-04	0
13_23_29_9_17_15_12_19_11_16_24_12_14_14_11_14_13_13_22_12	1,368E-04	4,503E-04	0	1,594E-03	0
13_23_30_10_17_14_12_20_10_15_25_12_13_15_9_14_12_12_22_12	1,375E-04	0	3,630E-03	2,115E-03	0
13_23_30_11_16_12_12_19_10_15_23_12_13_15_11_15_13_13_22_12	0	4,671E-04	0	1,768E-03	0
13_23_30_11_16_14_11_20_11_16_24_12_13_15_11_14_13_12_22_13	0	0	1,669E-04	2,362E-04	0
13_23_30_11_17_14_12_19_11_15_25_12_12_14_11_14_13_12_22_12	3,567E-05	0	0	1,967E-04	0
13_23_30_11_17_15.2_12_19_11_15_24_13_14_15_12_13_13_12_22_14	0	0	0	0	0
13_23_30_11_18_14_12_19_10_14_23_12_13_15_12_14_13_12_21_12	0	0	0	0	0
13_23_30_11_18_14_12_19_10_15_25_12_13_15_11_14_13_12_20_12	4,724E-05	1,143E-06	2,373E-05	5,419E-04	6,612E-04
13_23_30_11_18_14_12_19_10_15_25_12_13_15_11_14_13_12_22_12	1,204E-03	2,546E-04	2,710E-03	1,358E-02	4,078E-02
13_23_30_12_14_16_12_21_10_15_23_9_11_15_13_16_12_11_24_12	4,362E-03	1,539E-03	1,970E-03	3,726E-03	0
13_23_30_12_14_16_12_21_10_15_23_9_11_15_13_16_12_11_24_13	5,830E-03	2,045E-03	2,633E-03	5,035E-03	0
13_23_30_12_15_15_12_20_10_16_25_11_11_14_11_14_13_10_23_12	0	0	0	0	0
13_23_30_12_17_14_12_18_11_17_24_12_13_16_11_14_13_12_22_12	1,503E-04	0	2,143E-04	5,776E-04	0

Tableau F.1 (suite)

Haplotype 20 STR-Y	BSL	CDS	CNO	GAS	IMA
13_23_31_10_16_15_12_20_11_17_24_11_11_14_10_14_13_10_25_12	4,059E-05	7,027E-05	0	7,151E-04	0
13_23_31_11_17_14_12_19_11_15_23_12_13_15_12_13_13_12_22_12	0	0	0	0	0
13_23_31_11_17_14_13_19_11_16_24_12_13_15_11_15_13_12_24_11	2,124E-04	2,159E-03	4,066E-04	5,892E-04	0
13_23_31_11_18_14_12_19_10_15_23_12_13_15_11_12_11_12_22_12	0	0	6,191E-04	6,874E-04	0
13_23_31_11_18_14_12_19_10_15_23_12_13_15_11_12_13_12_22_12	0	0	9,640E-04	3,419E-04	0
13_23_31_11_18_15_11_19_10_14_24_9_11_15_14_15_12_12_24_12	3,707E-04	4,849E-05	1,146E-03	3,052E-03	0
13_24_29_10_16_14_12_19_11_17_24_12_14_15_12_15_13_12_25_13	0	0	0	0	0
13_24_29_10_17_14_12_19_11_15_25_12_13_15_11_14_13_13_24_13	0	0	0	1,010E-04	0
13_24_29_11_16_14_12_19_10_15_24_12_14_15_11_14_13_12_22_11	0	0	0	7,770E-04	0
13_24_29_11_16_14_12_19_10_15_24_12_14_16_11_15_13_11_22_13	0	0	0	2,222E-04	0
13_24_29_11_16_15_13_19_11_15_24_12_14_15_11_13_13_12_22_12	0	0	0	9,566E-04	0
13_24_29_11_18_14_12_19_11_16_24_12_13_14_11_14_13_12_23_13	3,969E-03	3,425E-03	4,944E-03	6,105E-03	7,018E-04
13_24_29_11_18_15_11_21_9_15_23_10_11_14_15_16_12_11_28_12	0	0	0	0	0
13_24_29_12_17_14_12_19_10_15_24_12_13_15_11_15_13_10_22_13	0	0	1,516E-04	7,395E-06	0
13_24_29_12_17_14_12_19_10_15_24_12_13_15_11_15_13_10_23_13	0	0	1,544E-04	7,888E-06	0
13_24_30_10_15_15_12_21_11_17_23_9_11_15_12_14_12_11_22_12	1,474E-02	9,779E-03	3,836E-03	3,219E-03	0
13_24_30_11_19_14_13_19_10_14_25_12_13_15_11_14_13_12_24_12	0	0	0	1,403E-03	0
13_25_29_11_17_14_12_19_10_16_23_12_13_15_11_16_11_12_28_13	0	0	0	0	0
13_25_29_11_18_14_12_18_11_16_24_11_13_15_11_15_13_12_22_12	0	0	0	2,223E-04	0
13_25_29_11_18_14_12_19_11_16_23_12_13_14_11_15_13_12_23_12	0	0	0	7,004E-04	0

Tableau F.1 (suite)

Haplotype 20 STR-Y	BSL	CDS	CNO	GAS	IMA
13_25_29_11_18_15_12_19_11_16_23_12_13_14_11_15_13_12_23_12	0	0	0	4.727E-04	0
13_25_30_11_16_15_11_20_10_15_25_12_11_14_11_14_13_11_22_12	0	0	0	0	0
13_26_28_11_17_14_13_20_11_16_23_12_13_15_11_14_13_12_23_12	0	0	0	1.146E-04	0
14_21_29_11_15_15_13_21_10_14_22_10_11_14_12_13_14_12_23_12	0	0	0	1.146E-04	0
14_21_31_10_17_15_11_20_11_14_24_10_11_16_11_14_14_11_24_12	0	0	0	1.146E-04	0
14_22_31_9_16_13_11_20_10_17_24_10_11_14_17_18_13_11_23_12	0	0	0	1.076E-04	0
14_22_32_11_15_15_11_20_10_15_23_10_12_14_15_15_15_11_27_12	0	0	4.181E-04	1.682E-03	0
14_23_29_11_17_14_12_19_11_16_24_12_13_15_11_13_13_11_22_12	0	0	0	0	0
14_23_30_10_17_14_11_18_10_15_24_12_13_14_12_14_13_14_22_11	8.653E-05	0	2.650E-04	4.018E-03	0
14_23_30_10_17_14_11_18_10_15_24_12_13_14_12_14_14_13_22_11	7.842E-05	0	2.345E-04	3.899E-03	0
14_23_30_10_18_14_11_17_9_15_24_12_13_14_11_14_13_11_22_12	0	0	0	2.088E-04	0
14_23_30_11_16_14_11_19_11_15_22_12_13_15_11_14_13_12_22_12	9.220E-08	9.329E-08	1.195E-08	8.954E-05	0
14_23_30_11_16_14_12_18_11_16_25_12_14_15_11_13_13_12_23_12	0	0	0	1.146E-04	0
14_23_30_11_16_14_12_19_11_15_24_12_13_15_11_14_13_12_22_12	1.076E-02	1.761E-02	1.919E-03	5.124E-03	0
14_23_30_11_16_14_12_19_11_15_24_12_13_15_11_14_13_13_22_12	1.557E-04	2.591E-04	2.798E-05	3.963E-04	0
14_23_30_11_16_14_12_19_11_16_24_12_13_15_12_14_14_12_22_13	1.636E-04	6.413E-05	4.312E-03	1.563E-03	9.267E-03
14_23_30_11_16_14_12_20_11_15_24_12_13_15_11_15_13_12_22_13	8.936E-08	9.758E-08	2.380E-08	3.566E-08	0
14_23_30_11_17_13_12_19_11_15_24_12_13_15_11_14_13_12_23_12	3.365E-04	4.302E-04	7.968E-09	3.384E-03	0
14_23_30_11_17_13_12_19_11_15_24_12_13_15_11_15_13_12_23_12	8.886E-05	7.702E-05	0	7.096E-04	0
14_23_30_11_17_14_13_19_11_15_24_12_13_15_11_16_13_11_21_13	0	0	0	2.036E-04	0

Tableau F.1 (suite)

Haplotype 20 STR-Y	BSL	CDS	CNO	GAS	IMA
14_23_30_11_17_15_12_19_10_16_24_12_13_15_11_15_13_13_22_12	1,837E-04	2,905E-03	6,270E-04	3,774E-03	0
14_23_30_11_18_14_12_19_10_16_23_12_13_15_12_13_13_12_22_12	0	0	0	0	0
14_23_30_11_19_14_12_19_11_15_24_12_13_15_11_14_14_12_22_12	0	0	0	3,227E-04	0
14_23_30_11_20_14_12_19_10_17_23_12_13_15_11_11_13_12_22_12	0	0	8,866E-05	5,408E-04	0
14_23_30_12_20_14_12_19_10_17_23_12_13_15_11_11_13_12_22_12	0	0	1,018E-04	3,737E-04	0
14_23_31_11_17_14_12_19_11_15_24_12_13_15_12_15_13_11_21_12	0	0	0	0	0
14_23_31_12_14_15_13_20_10_15_26_11_11_14_11_14_13_10_23_12	0	0	2,142E-04	3,163E-04	0
14_23_31_9_17_14_9_20_10_17_23_9_11_15_12_15_12_11_22_11	3,914E-04	5,472E-05	2,010E-04	4,813E-04	0
14_23_32_10_16_14_12_18_11_17_23_12_13_15_11_15_13_12_23_11	1,370E-05	0	0	7,318E-04	0
14_23_33_10_16_14_12_18_11_17_23_12_13_15_11_15_13_12_23_11	1,459E-05	0	0	9,259E-04	0
14_24_30_11_17_15_12_19_11_16_24_12_13_15_11_14_13_12_22_12	1,508E-04	5,371E-05	2,074E-04	7,390E-04	0
14_24_31_11_17_15_12_19_11_16_24_12_13_15_11_14_13_12_22_12	2,097E-04	5,755E-05	2,222E-04	7,035E-04	0
15_23_31_11_17_13_12_19_11_16_24_12_13_15_11_15_13_12_23_12	4,900E-05	7,554E-05	0	5,657E-04	0