Y-haplogroup diversity, non-paternity estimates, and Y chromosome mutation rates in Quebec

GERARDO MARTÍNEZ

Department of Human Genetics, Faculty of Medicine and Health Sciences,
McGill University, Montreal, Quebec, Canada

January 21, 2025

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science.

A Pabloma.

No viste la noche crecida, no quisiste el amanecer.

Abstract

The male-specific region of the Y chromosome (MSY) is extensively used to assess relatedness between male individuals. Single-nucleotide polymorphisms in the MSY enable this assessment, as they are mostly identical by descent in individuals due to the MSY's lack of recombination and slow mutation rate. Specific bi-allelic sets of mutations in the MSY (haplogroups) are also utilized to analyze population diversity and migration events. In addition, when genealogical data is available, markers in the MSY can help confirm genealogical relationships. Conversely, genealogical relationships can help us address questions on the biology of the MSY, specifically mutation rates.

In this study, we utilized a large-scale cohort of present-day male individuals from metropolitan areas in Quebec and a deep-rooted genealogy to examine the interplay between the MSY and paternal lineages. We aimed to characterize the haplogroup composition of men in Quebec to uncover patterns of diversity and migration. We sought to estimate the rate of non-paternity events within the genealogy. Finally, leveraging genealogical paternal lineages, we aimed to estimate the MSY point mutation rate and evaluate the effect of paternal age on mutation rates.

The top five more common haplogroups in our cohort were R1b1a1 (62.4%), E1b1b1 (6.4%), R1a1a1 (3.13%), J2a1a1 (2.7%) and E1b1a1 (2.58%). The observed haplogroup diversity reflects multiple waves of migration to Quebec dating back to the 17th century. We identified a non-paternity rate of 0.73% per generation, consistent with previous findings in Quebec and other European populations. Finally, in the MSY, we observed a rate of 2.398×10^{-8} mutations per base pair per generation and 6.47×10^{-10} per base pair per year.

Résumé

La région du chromosome Y spécifique au sexe masculin (MSY) est largement utilisée pour évaluer la parenté entre les individus de sexe masculin. Les polymorphismes nucléotidiques dans la MSY permettent cette évaluation, car ils sont pour la plupart identiques par filiation chez les individus en raison de l'absence de recombinaison et du faible taux de mutation dans la MSY. Des ensembles bi-alléliques spécifiques de mutations dans le MSY (haplogroupes) sont également utilisés pour analyser la diversité des populations et les événements migratoires. En outre, lorsque des données généalogiques sont disponibles, les marqueurs de l'MSY peuvent aider à confirmer les relations généalogiques. Inversement, les relations généalogiques peuvent nous aider à répondre à des questions sur la biologie de l'MSY, en particulier les taux de mutation.

Dans cette étude, nous avons utilisé une cohorte à grande échelle d'individus masculins provenant de régions métropolitaines du Québec et une généalogie profonde pour examiner l'interaction entre le MSY et les lignées paternelles. Nous avons cherché à caractériser la composition des haplogroupes des hommes du Québec afin de découvrir des modèles de diversité et de migration. De plus, nous avons cherché à estimer le taux d'événements de non-paternité dans la généalogie. Enfin, en nous appuyant sur les lignées paternelles dans la généalogie, nous avons cherché à estimer le taux de mutation de l'MSY et à évaluer l'effet de l'âge paternel sur les taux de mutation.

Les cinq haplogroupes les plus fréquents dans notre cohorte étaient R1b1a1 (62,4%), E1b1b1 (6,4%), R1a1a1 (3,13%), J2a1a1 (2,7%) et E1b1a1 (2,58%). La diversité des haplogroupes observée reflète les multiples vagues de migration vers le Québec depuis le 17e siècle. Nous avons identifié un taux de non-paternité de 0,73% par génération, ce qui est cohérent avec des résultats obtenus au Québec et dans d'autres populations européennes. Enfin, nous avons observé, dans le MSY, un taux de 2.398×10^{-8} mutations par paire de base par génération et de 6,5 × 10-8 mutations par paire de base par génération. par paire de base par génération et de 6.47×10^{-10} par paire de base par an.

Table of contents

List of Abbreviations

MSY male-specific region of the Y chromosome

SNP single-nucleotide polymorphism

mtDNA mitochondrial DNA

Ychr Y chromosome

STR short tandem repeat

ISOGG International Society of Genetic Genealogy

Y-SNP SNP in the Y chromosome

Y-STR STR in the Y chromosome

NPE non-paternity event

PGMR per-generation mutation rate

MRCA most recent common ancestor

PYMR per-year mutation rate

Xchr X chromosome

PAR pseudoautosomal regions

XTR X-transposed

XDG X-degenerate

T2T Telomere-to-Telomere

PAL eight palindromic sequences and three inverted repeats

rAMP spacer and tandem repeats

CI confidence interval

DNM *De novo* mutation

NPR non-paternity rate

AD allele depth

ECDF empirical cumulative distribution function

TiTv transition to transversion ratio

List of Figures

List of Tables

Acknowledgments

I want to begin by thanking my supervisor, Simon Gravel. You provided a lot of help and support during these years and showed that not only you are a great scientist but also an empathetic person who cares for and understands his students. I would also like to thank you for the opportunity you gave me to study at such a prestigious institution as McGill.

I want to continue by thanking all my colleagues at Gravel lab. Thank you all for the fruitful scientific discussions and fun unscientific ones.

I want to thank my supervisory committee Claude Bhérer and Emmanuel Milot for their useful feedback on this project.

I also want to thank my family for their unconditional support and trust in my hard-to-follow endeavors. I know you are all proud of me and I appreciate that deeply!

I want to finish by thanking all my friends. Thank you all for the time that you have shared and continue sharing with me. Thank you for the laughs, the debates, and the embraces.

Gracias. Merci. Thank you.

Format of the Thesis

This thesis is presented in a traditional format, in agreement with the guidelines for thesis preparation set out by the Graduate and Postdoctoral Studies Office (GPS) and the Department of Human Genetics of McGill University.

It comprises five chapters and two appendices. Chapter 1 describes the problem and objective of this work and includes a literature review. Chapter 2 describes the data sets and methodology used for this work. Chapter 3 shows the results obtained. Chapter 4 discusses the findings and contrasts them with the existing literature. Chapter 5 provides the conclusions and outlines future directions. Appendix A contains supplementary methods, while Appendix B includes supplementary results. References are formatted in APA style.

Contribution of Authors

The author performed the conceptualization, formal analysis, code implementation, figures, literature review, and writing of this text. Conceptualization and supervision for this project was done by Simon Gravel. The UMAP-based clusters used in Section 2.2 were built by Alex Diaz-Papkovich.

Chapter 1

Introduction

1.1 Y chromosome and the Quebec population

One of the ways to establish relatedness between humans is using haploid markers such as mitochondrial DNA (mtDNA) and the male-specific region of the Y chromosome (MSY) (Calafell & Larmuseau, 2017). mtDNA is non-recombining and is transmitted from mother to offspring, so it can be used to trace maternal lineages. For readers interested in using mtDNA for exploring relatedness and identification in forensic sciences, see Amorim, Fernandes, and Taveira (2019). In contrast, the MSY, the non-recombining region of Y chromosome (Ychr), allows us to reconstruct paternal lineages. This work focuses on the Ychr and patrilineages—groups of male individuals with a common paternal ancestor.

Two main types of markers in the Ychr are commonly analyzed today: slow-mutating single-nucleotide polymorphisms (SNPs) and moderately fast evolving short tandem repeats (STRs) (Calafell & Larmuseau, 2017; Vergani, 2021). The International Society of Genetic Genealogy (ISOGG) has identified over 90,000 bi-allelic SNPs in the MSY. Due to a low mutation rate of 3.07×10^{-8} (Helgason et al., 2015) and paternal inheritance, SNPs in the Y chromosome (Y-SNPs) are considered identical by descent between individuals. Y-SNPs enabled researchers to reconstruct a paternal phylogenetic tree that links all males to their most recent common paternal ancestor, or "Y-Adam" (The Y Chromosome Consortium, 2002). The branches of this phylogenetic tree are called *haplogroups*, and they have been used to analyze geographic patterns of diversity and migration events (Jobling & Tyler-Smith, 2003). On the other hand, STRs in the Y chromosome (Y-STRs) shows an average mutation rate of 3.83×10^{-4} mutations per generation (Willems, Gymrek, Poznik, Tyler-Smith, & Erlich, 2016), higher than that of Y-SNPs. Commercially available Y-STR genotyping kits have a higher mutation rate ranging from 1.7×10^{-3} to 3.99×10^{-3} (Balanovsky et al., 2015). Y-STRs are widely used to assess paternity in father-son pairs and to exclude male

suspects in crime scene investigations (Calafell & Larmuseau, 2017; Vergani, 2021).

When genealogical data is available, Ychr markers can confirm or exclude two individuals as biological relatives. One example of this is seen in the study by King et al. (2014). After identifying the skeletal remains of King Richard III of England, they compared the Y-haplogroups of five living relatives. They found four relatives shared a common haplogroup, while one did not, suggesting that a non-paternity event (NPE) had occurred within the last four generations. Deep-rooted male pedigrees have also been used to understand historical extra-pair paternity in human populations (e.g., Larmuseau et al. (2013)). In broad terms, extra-pair paternity events are identified by contrasting the genealogical ancestor of two men with their biological ancestor, as estimated from Ychr data. Studies across human populations estimate between 0.009 and 0.018 NPEs per generation.

Conversely, genealogies are relevant for studying the biological properties of the MSY. Dividing the number of differences in their MSY sequences of two men by the number of generations that separate them is a direct way of estimating the per-generation mutation rate (PGMR) in the MSY. Additionally, knowing the number of years to the most recent common ancestor (MRCA) for a male pair allows estimation of the per-year mutation rate (PYMR). An essential evolutionary application of knowing the mutation rate in the MSY is estimating the MRCA of all Ychr. According to Helgason et al. (2015), this MRCA likely lived between 188,000 and 296,000 years ago.

In the province of Quebec, we have two data sets that allow us to answer questions about male individuals. The first is the BALSAC database, a deep-rooted genealogy of about 6.5 million individuals. It has been reconstructed from transcribed and digitized parish and civil records (Tarride et al., 2023). BALSAC captures the genealogical relationships of individuals spanning from the 1600s to the 1960s. The second data set is the CARTaGENE cohort, a population-based biobank including approximately 30,000 participants from metropolitan areas in Quebec. Among other biological samples, it includes SNP array data for nearly all participants and whole-genome sequences for thousands.

1.2 Objectives of this work

Using the BALSAC genealogy and the CARTaGENE cohort, we aim to address various questions using Ychr and genealogical data.

1) The first objective is to **obtain the Y-haplogroup composition of male individuals in Quebec**. Moreau et al. (2009) studied 176 individuals from the Gaspé Peninsula in Quebec with a limited number of Y-SNPs. No known study has analyzed the Y-haplogroup composition across the entire province of Quebec with a large sample and a high marker resolution.

2) The second objective is to **estimate the non-paternity rate in Quebec** using Ychr data. Two studies have examined genealogical mismatches in Quebec using Y-STR to detect NPEs (Doyon, 2018; Heyer, Puymirat, Dieltjes, Bakker, & de Knijff, 1997). With the CARTaGENE SNP array data, we aim to assess the effectiveness of Y-SNPs at detecting NPEs.

3) The third objective is to **calculate the SNP mutation rate in the MSY** using the BALSAC genealogy. The most recent estimate was reported by Helgason et al. (2015), where they observed a differential mutation rate in two regions of the MSY. Additionally, concordant to previous studies, they found an impact of the father's age at conception in the SNP mutation rate. In our work, we aim to replicate these results in a population from Quebec.

To provide context and a framework for the analysis, we will present an overview of existing literature that examines the main concepts and findings related to Y-haplogroup, non-paternity rates, and Ychr mutation rates.

1.3 Biology and structure of the human Y chromosome

Both sex chromosomes in humans evolved from a pair of regular autosomes before the divergence of placental and marsupial mammals (Hughes & Page, 2015). Due to large-scale structural variations, likely occurring in the Ychr, recombination between the X and Y chromosomes was largely suppressed. In humans, the Ychr does not undergo recombination with its meiotic partner, the X chromosome (Xchr), except for two regions at both of its ends known as the pseudoautosomal

regions (PAR). Loss of crossing over had two main consequences. First, there is a substantial reduction in size: the Xchr is 154 Mb in total length, while the Ychr is only 62 Mb (Nurk et al., 2022; Rhie et al., 2023). The second consequence is a reduction in the number of genes in their euchromatin regions: the Xchr has 1456 genes (including protein-coding and non-coding RNA), whereas the Ychr has 883 genes, and only 488 are predicted to be protein-coding National Center for Biotechnology Information (2024); Rhie et al. (2023).

1.3.1 Sequence classes

The Ychr can be segmented into distinct classes depending on their homology with the Xchr, sequence type, and gene content (Rhie et al., 2023; Skaletsky et al., 2003). The five most important classes from longest to shortest are the heterochromatin region of the q-arm, the ampliconic, the X-degenerate (XDG), the X-transposed (XTR), and the PAR regions (Figure 1.1. The complete list of classes can be found in Rhie et al. (2023). The set of sequences excluding the PAR regions is known as the MSY. We will now explain in detail the most important regions of the MSY.

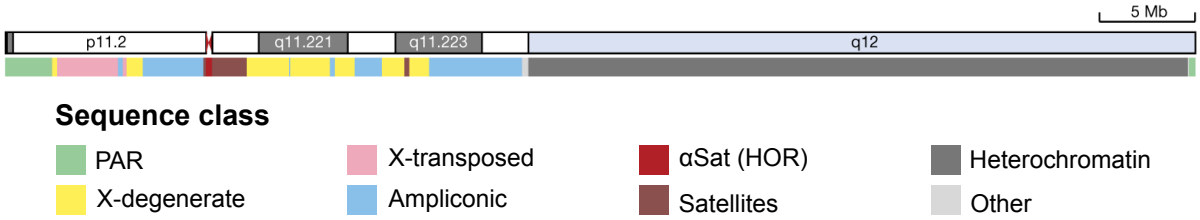

Figure 1.1: Diagram of Y-chromosome sequence classes. HOR: highest order satellites. Modified from Rhie et al. (2023).

With a length ~34 Mb, the heterochromatin region on the distal end of the q-arm comprises 55.43% of the Ychr. Its sequence remained largely unresolved in the GRCh38 assembly due to its highly repetitive structure, which made it difficult to assemble. This region contains two interspersed satellite sequences, DYZ1 and DYZ2, contributing to this complexity (Schmid, Guttenbach, Nanda, Studer, & Epplen, 1990). In 2023, the first complete Ychr sequence was published by the Telomere-to-Telomere (T2T) consortium (Rhie et al., 2023).

The ampliconic regions are a set of seven non-contiguous segments that are composed of sequences that show high similarity – as much as 99.99% identity – to other sequences in the MSY

(Skaletsky et al., 2003). The ampliconic region can be divided into two main components (Helgason et al., 2015; Skaletsky et al., 2003): first, a set of eight palindromic sequences and three inverted repeats (PAL); and second, other spacer and tandem repeats (rAMP). In this region, there is the highest density of genes in the Ychr. Most of these genes are multicopy and expressed predominantly in the testes, suggesting their function is related to spermatogenesis (Hughes & Page, 2015).

The XDG region is a set of eight non-contiguous sequences with various levels of homology with the Xchr. This region includes single-copy and pseudogenes that display between 60% and 99% sequence identity to their X-linked homologs (Skaletsky et al., 2003). In contrast to the ampliconic regions, most genes in the XDG are expressed ubiquitously in the body and are involved in housekeeping activities. Of importance in this region is the sex-determining region responsible for triggering events that lead to the testes development in therian mammals (Wallis, Waters, & Graves, 2008).

The last class is the XTR regions. They comprise two non-contiguous sequences that are the result of an X-to-Y transposition. These sequences are not found in other hominids, which shows that they appeared after the divergence of humans from the rest of the hominids, between 4.3 and 2.56 million years ago (Page, Harper, Love, & Botstein, 1984; Püschel, Bertrand, O'Reilly, Bobe, & Püschel, 2021). The sequences in these two regions show approximately 99% homology with the sequences in Xq21 and carry genes with functional X- and Y- linked homologs. It was widely thought that crossing over was impossible in these regions as it would destroy the integrity of the X and Y chromosomes (Page et al., 1984). However, diversity patterns in the XTR show that the region might still undergo recombination (Cotter, Brotman, & Wilson Sayres, 2016).

1.4 Y haplogroups

Due to its lack of recombination, diversity in the MSY is driven by mutational events. Non-recurrent binary polymorphisms can be used to build a maximum parsimony phylogenetic tree that explains this diversity (The Y Chromosome Consortium, 2002). The Y Chromosome Consortium (2002) defined the first set of known polymorphisms to use to build the tree and a nomenclature for its

branches. The main branches of this tree, the *haplogroups*, are designated with a letter from A to R. Subclades within a major haplogroup are indicated by numbers (e.g., A1, A2, A3, ...), with subsequent divisions denoted by lowercase letters (A1a, A1b, ...). Further divisions alternate between numbers and lowercase letters (e.g. R1a2b). The internal or terminal nodes are called subhaplogroups or simply haplogroups, depending on the authors – when there is no confusion, we will use the term haplogroup to refer to both main branches and terminal branches. Haplogroups can also be named by the mutation that defines them – for example, the haplogroup R1a2b can be named R-YP4132. If multiple mutations define a branch, any of them can be used to name the haplogroup.

Haplogroups follow geographic patterns and have been largely used to understand demographic patterns and history. A review of the geographic distribution of the Y haplogroups can be found in Calafell and Comas (2021). In Quebec, haplogroup diversity has been analyzed in Moreau et al. (2009). In that study, 13 polymorphisms were analyzed from a sample of 176 living in the Gaspe Peninsula, Quebec. They found that 67% of the sample belongs to a R1b subclade. Other haplogroups found in more than 1% of the sample were I (17.6%), J2 (3.9%), F excluding the subclades I, J2, and K (3.9%), R1a1 (3.4%), E (2.27%) and K (1.1%).

1.5 Mutation rate in the Y chromosome

1.5.1 Estimates in the literature

Compared to other chromosomes, the lack of recombination in the MSY simplifies the estimation of its mutational events. Two main mutation rate estimates are reported in the literature: one for SNP and another for STR. This section will focus on SNP mutation rate; a review of the various estimates of the STR mutation rate and the technologies used to obtain them can be found in Balanovsky (2017). Additionally, we will focus on the *genealogical approach* to estimate the SNP mutation rate. Hereafter, whenever we use the term "mutation rate," we will refer specifically to "the SNP mutation rate in the MSY".

The genealogical approach for estimating the mutation rate is based on a known male pedigree (see Section 2.4). The first step in this approach is to count the number of mutations observed in the sample in a predefined genome region. Then, one must obtain either the number of generations

Table 1.1: MSY genealogical mutation rate estimates in the literature. Modified from Balanovsky (2017).

Study	Sequence class	Number of samples	Number of distinct patrilineages	PGMR [$\times 10^{-8}$] (%95 CI)	PYMR [$\times 10^{-10}$] (%95 CI)
Xue et al. (2009)	—	2	1	3.0 (0.89 – 7.0)	10 (3.0 – 25)
Helgason et al. (2015)	PAL	753	274	2.55 (2.21 – 2.93)	7.37 (6.41 – 8.48)
Helgason et al. (2015)	XDG + XTR + rAMP	753	274	3.01 (2.77 – 3.26)	8.71 (8.03 – 9.43)
Balanovsky et al. (2015)	—	9	1	—	7.8 (6.2 – 9.4)

or the number of years from the sample individuals to their MRCA. Finally, the proportion of mutations observed in the genome region is divided by the total number of generations to calculate a PGMR. Similarly, dividing the proportion by the total number of years provides a PYMR. In Table 1.1, we compare four mutation rate estimates found in the literature.

The first study that estimated the genealogical mutation rate using a deep-rooted pedigree is Xue et al. (2009). In that study, the authors analyzed two individuals of the O3a1 haplogroup separated by 13 generations. Candidate mutations found using a standard SNP array were confirmed by re-sequencing the regions spanning the mutations and verifying additional family members. In total, the researchers found four mutations in a sequence of length ~10.15Mb. They obtained a PGMR of 3.0×10^{-8} (95% confidence interval (CI): $0.89 \times 10^{-8} - 7.0 \times 10^{-8}$) and a PYMR of 1×10^{-9} ($3.0 \times 0.3^{-9} - 2.5 \times 10^{-9}$). Despite their mutation rate being consistent with previous estimates obtained comparing human to chimpanzee Ychr sequences (Kuroki et al., 2006), their estimate had a wide confidence interval. The authors hypothesized that a larger sample size would make the estimate more accurate.

In Helgason et al. (2015), the authors utilized a larger and deeper genealogy to obtain a more accurate mutation rate. They analyzed 753 individuals of the haplogroups E1b1, I1, I2, Q1a3a, R1a1, and R1b1a, forming 274 distinct genealogical patrilineages. They obtained 1456 candidate mutations from whole genome sequencing and validated 101 to estimate the false positive rate. Additionally, they explored the hypothesis of differential mutation rate in different sequence classes in the Ychr. They found that the mutation rate in PAL is lower than in the other sequence classes they analyzed. They hypothesized that damaged nucleotides in this region are corrected with paralogous sequences in the opposite arm of the palindromes. They obtained a PGMR of 3.01×10^{-8} (95%

CI: $2.77 \times 10^{-8} - 3.26 \times 10^{-8}$) and a PYMR of 8.71×10^{-10} (95% CI: 8.03×9.43^{-10} for the combined regions of XTR, XDG, and rAMP. They also reported a PGMR of 2.55×10^{-8} (95% CI: $2.21 \times 10^{-8} - 2.93 \times 10^{-8}$) and a PYMR of 7.37×10^{-10} (95% CI: $6.41 \times 10^{-10} - 8.48 \times 10^{-10}$) for the PAL.

Balanovsky et al. (2015) analyzed nine samples from the G1-L1323 haplogroup to give their estimate of the mutation rate. A historical common ancestor of these nine individuals was suspected to have lived approximately 600 years ago. The topology of the phylogenetic tree produced from Ychr sequences fit the genealogical records and thus a PYMR of 7.8×10^{-10} (95% CI: $6.2 \times 10^{-10} - 9.4 \times 10^{-10}$) was obtained. No PGMR was produced in this study.

Balanovsky (2017) compared the three previously described estimates for the mutation rate with a Mann-Whitney U test and found that their differences were insignificant except for the PAL estimate provided in Helgason et al. (2015).

1.5.2 Paternal effect on germline mutations

De novo mutations (DNMs) are mutations that appear in an individual but are not present in their parent's (Kessler et al., 2020). DNMs can arise either during embryogenesis, resulting in *germline mutations*, or post-fertilization, resulting in *somatic mutations* (Veltman & Brunner, 2012). The mutation rate of germline single-nucleotide variants in humans is estimated at approximately 1×10^{-8} per genome per generation, translating to an expected number of 45 to 60 DNMs per individual per generation (Goldmann et al., 2016).

Studies have shown a positive correlation between parental age and the number of DNMs (see K. A. Wood and Goriely (2022) for an exhaustive review on the topic). Father's age at conception increases the number of DNMs in approximately 1.5 mutations per year, whereas mother's age has a smaller effect with approximately 0.4 mutations per year. The primary mechanism underlying the paternal age effect on DNM rate is the process of spermatogenesis (K. A. Wood & Goriely, 2022). While all cell divisions that give rise to oocytes occur before a woman's birth, spermatogonial stem cells divide continuously throughout men's reproductive life. A 20-year-old male is estimated to undergo 150 cell divisions from primordial stem cells, increasing to approximately 610 by age 40 (Crow, 2000).

Helgason et al. (2015) (see above for details on their work) studied the impact of the father's age at conception on the number of mutations in the MSY. They analyzed the correlation of the PGMR and the mean generation interval in genealogical branches. They observed a significant, albeit weaker, correlation compared to findings in studies focused on autosomes. They suggest this reduced effect may be attributable to the shorter length of the MSY relative to autosomes, resulting in fewer detectable DNMs per generation.

1.6 Non-paternity rate

An extra-pair paternity (EPP) event occurs in a socially monogamous species when the social parent is not the biological parent (Larmuseau et al., 2013). Several studies have attempted to estimate the EPP rate in humans to determine the frequency of cuckoldry in human populations (e.g., Anderson (2006); Bellis, Hughes, Hughes, and Ashton (2005)). However, in the context of genealogy, the genealogical ancestor might not be the biological ancestor for reasons other than cuckoldry: undeclared adoptions, errors in historical records, errors in the digitizing process, etc. For this reason, and following other authors (e.g., Greeff and Erasmus (2015)), we will use the term "non-paternity event."

This section will explore four studies that calculated the non-paternity rate (NPR) in human populations using deep male pedigrees and Ychr data. Heyer et al. (1997) was the first study that obtained an estimate of the NPR in a population from Quebec. Larmuseau et al. (2013) developed an iterative algorithm to analyze pairs of male individuals and detect NPE. This technique was subsequently applied to a population from Quebec by Doyon (2018) and refined by Larmuseau et al. (2019) to explain how covariates affect the historical NPR in two populations from the Low Countries. The selection of these four studies is admittedly arbitrary and is not meant to be exhaustive. A complete historical recollection of the studies where the NPR was estimated can be found in Greeff and Erasmus (2015) and Scelza et al. (2020). Table 1.2 compares the estimates obtained in some of these studies.

Table 1.2: Non-paternity rate estimates found in the literature ordered by date of publication.

Study	Population studied	Number of samples	Type of data used	NPR [$\times 10^{-3}$] (%95 CI)
Heyer et al. (1997)	French-Canadian	42	STR	3.89 (0.0985 – 21.5)
Strassmann et al. (2012)	Dogon	1,218	STR	18.0 (%95 CI unknown)
Larmuseau et al. (2013)	Flemish	1071	SNP + STR	9.10 (4.10 – 17.5)
Greeff and Erasmus (2015)	Afrikaner	199	STR	8.64 (4.32 – 15.41)
Boattini et al. (2015)	Northern Italian	149	SNP + STR	12.1 (4.00 – 61.3)
Larmuseau et al. (2017)	Dutch	~3000	SNP + STR	9.6 (4.6 – 17.6)
Doyon (2018)	French-Canadian	429	STR	4.00 (1.00 – 10.0)
Larmuseau et al. (2019)	Flemish + Dutch	236 + 63	SNP + STR	16.0 (12.0 – 21.0)

1.6.1 Estimates in the literature

Heyer et al. (1997) analyzed nine Y-STR loci in 42 men from the Saguenay region in Quebec to investigate the Y-STR mutation rate. They found that one individual had seven haplotype differences compared to the others of the same patrilineage and, consequently, declared that one NPE had occurred in that patrilineage. Based on 257 independent meioses across 12 distinct patrilineages, they estimated a NPR of 3.89×10^{-3} (%95 CI: $(0.0985 \times 10^{-3} - 21.5 \times 10^{-3})$).

Larmuseau et al. (2013) analyzed a sample of 1071 male individuals whose oldest reported paternal ancestor lived in Flanders, Belgium, before 1800. They reconstructed the genealogies of the sample based on parish and civil records and established 60 pairs of individuals with a common genealogical ancestor. To compare the biological ancestors in a pair, they used a set of 38 STRs and a set of SNPs. They defined that a NPE must have occurred between a pair of men if they did not belong to the haplogroup or they belonged to the same haplogroup but had more than seven differences in they STR-haplotypes. They obtained a NPR of 9.10×10^{-3} (%95 CI: $(4.10 \times 10^{-3} - 17.5 \times 10^{-3})$).

Doyon (2018) analyzed 429 men from eight regions in Quebec, Canada. All individuals were connected to the BALSAC genealogy. Samples were genotyped using different STR kits at different resolutions (ranging from 7 to 27 STR) depending on the time of their recruitment. The author

defined that NPE had occurred between two individuals if they differed in more than 18% of the STRs compared. The non-paternity rate obtained was 4.0×10^{-3} (%95 CI: $1.0\times10^{-3}-10.0\times10^{-3}$).

Larmuseau et al. (2019) analyzed 263 Flemish and 63 Dutch individuals to explore if socioeconomic status and population density affect the historical NPR. They selected 513 independent pairs of individuals separated by at least seven meioses. Genealogical relationships between individuals were reconstructed using civil records, parish records, and other secondary documents such as notarial acts. This allowed the authors to reconstruct genealogical patrilineages dating back to 1315. To estimate the NPR, the authors used a panel of 191 SNPs and 38 STRs. Individual's socioeconomic status was inferred using the father's occupation, and population density was obtained from historical records. The authors defined that a NPE had occurred in a pair of individuals if they did not share the same haplogroup or if the number of generations estimated from the number of differences in STRs did not fit the number of generations in the genealogy. The average NPR estimated in this study was 16.0×10^{-3} (%95 CI: $12.0 \times 10^{3} - 21.0 \times 10^{-3}$). Additionally, the authors found that the NPR varied significantly with socioeconomic status and population density. They found that the NPR increased with population density and in low socioeconomic populations ranging from 6×10^{-3} to 59×10^{-3}.

Chapter 2

Materials and Methods

2.1 Sample information

The CARTaGENE cohort includes 29,337 (15,698 female, 13,639 male) biological samples of participants living in metropolitan regions of Quebec (Awadalla et al., 2012). Recruitment of participants took place in six cities: Montreal (70% of the cohort), Quebec City (15.3%), Sherbrooke (4.5%), Saguenay (4.2%), Gatineau (3.7%) and Trois-Rivières (2.3%). Participants were randomly sampled within each population based on health insurance registries.

3320 men from the CARTaGENE cohort have been linked to the BALSAC genealogy. For these individuals, we have information about their parents, the place and time of their marriages, and those of their parents. Eighty percent of these men were born in one of the 17 administrative regions in Quebec (Figure 2.1.a and Section A.1). The known patrilineages in this sample range from 1 to 15 generations, with an average of 11 generations (Figure 2.1.b). Furthermore, the oldest known paternal ancestor of the individuals in the sample got married for the first time between 1560 and 1979. (Figure 2.1.c).

2.1.1 SNP array samples

We had SNP-array data for the whole sample of men connected to BALSAC. Samples were genotyped at five points in time using different SNP-array panels and number of SNPs (Table 2.1). At each phase, variants with more than 0.05 missingness across all samples were removed. Additionally, we removed variants labeled heterozygotic according to PLINK 1.9 (Chang et al., 2015; Purcell & Chang, 2020) and variants from the PAR. A total of 4198 SNPs were kept in the MSY.

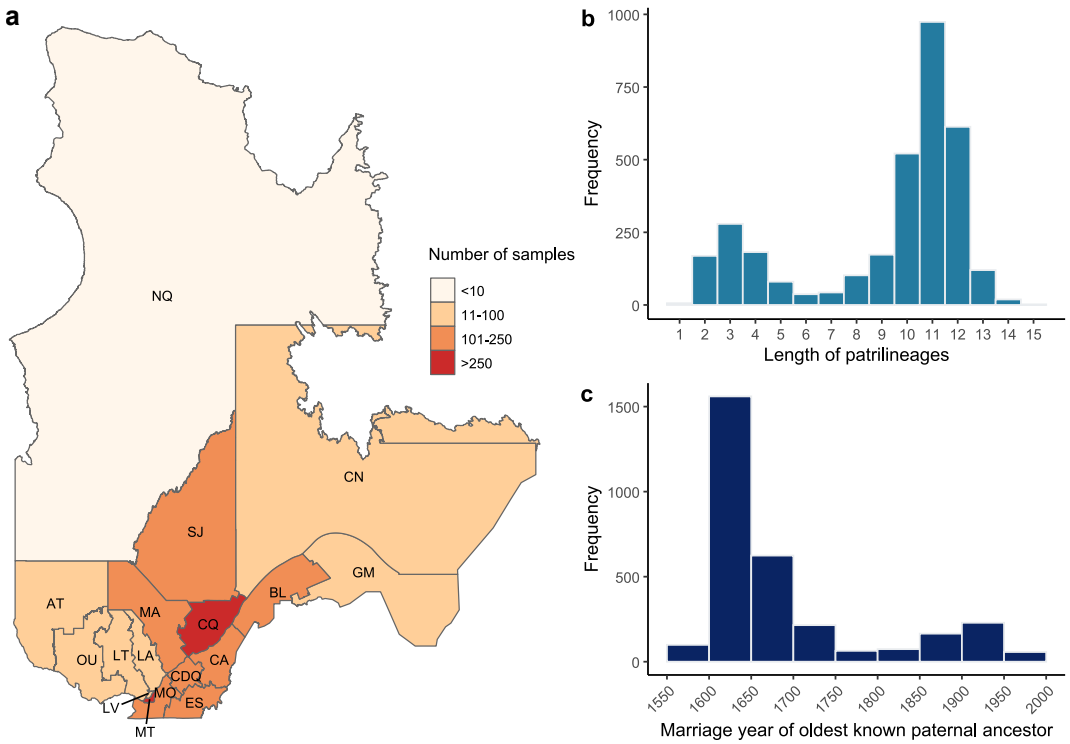

Figure 2.1: Information about the sample of 3320 men connected to the BALSAC genealogy. a) Map of the province of Quebec colored by the number of samples in each region. The divisions correspond to the administrative regions defined by the Ministry of Natural Resources and Forests (Direction générale de l'information géospatiale, 2024). AT: Abitibi-Témiscamingue, BL: Bas-Saint-Laurent, CA: Chaudière-Appalaches, CN: Côte-Nord, CDQ: Centre-du-Québec, CQ: Capitale-Nationale, ES: Estrie, GM: Gaspésie–Iles-de-la-Madeleine, LA: Lanaudière, LT: Laurentides, LV: Laval, MA: Mauricie, MO: Montréal, MT: Montérégie, NQ: Nord-du-Québec, OU: Outaouais, SJ: Saguenay–Lac-Saint-Jean. b) Number of generations in patrilineages. c) Marriage year of the oldest known paternal ancestor of the sample individuals.

2.1.2 Complete Y-chromosome sequences

The CARTaGENE project includes 2,184 full genome sequences, of which 985 are from male individuals. Of these, 449 were linked to the BALSAC genealogy.

Details on the sequencing pipeline will be published shortly. To call variants in the Ychr, we used the standard DRAGEN pipeline (Illumina, 2020). We first removed samples with more than 5% missing data across all chromosomes and variants that included heterozygotic genotypes. Next, we kept only bi-allelic variants and removed those corresponding to indels. Finally, we filtered positions according to the 1000 Genomes accessibility mask (Auton et al., 2015); this mask was

Table 2.1: Information on the SNP array data. The array names correspond to Illumina® genotyping manifest names found in Illumina (2024).

Year	Array type	Number of samples	Number of SNPs
2017	GSAMD-24v1-0_20011747_A1	2155	1195
2018	GSA-24v1-0_A1	352	1271
2018	GSAMD-24v2-0_20024620_A1	2056	5758
2020	GSAMD-24v3-0-EA_20034606_C1	938	3564
2021	GSAMD-24v2-0_20024620_A + specific SNPs included by CARTaGENE	8138	4710

constructed using a filter on coverage and quality of the mapping to the reference genome.

VCF files generated by the DRAGEN pipeline include an FT info field based on genotype quality and depth (Illumina, 2020). We set to missing all genotypes with an FT field that was not equal to PASS.

We extracted the allele depth (AD) of singletons called the alternate allele for all the individuals in the sample (see Figure A.1 for the distribution of AD in our sample). The rationale for focusing on singletons is that they are more prone to be confused with sequencing errors. We kept genotypes with an AD in the interval 6-23, corresponding to the central 90% of the empirical distribution of AD. We set genotypes falling outside this interval to missing. After filtering, 1433 SNPs appear both in the SNP-array and sequence data.

2.2 Y-haplogroup classification

We assigned samples to the deepest haplogroup using the software Y-Lineage Tracker (Chen, Lu, Lu, & Xu, 2021) and the ISOGG 2019-2020 haplogroup tree (International Society of Genetic Genealogy, 2020). The main goal of this software is to find the most likely lineage track leading from a sample to Y-Adam. This software uses key markers commonly used in studies and of declared importance in the ISOGG tree. It checks the key markers that appear in a sample as derived alleles and builds a set of potential terminal haplogroups. Following the criteria outlined by Chen

et al. (2021), it selects the most likely lineage track and outputs both the terminal haplogroup and a lineage to Y-Adam. An important feature of their algorithm is that it can determine a terminal haplogroup even if key mutation markers are missing.

2.3 Non-paternity rate estimation

To calculate the NPR, we utilized the SNP array data set as it presented the bigger sample size. We developed the following method to detect NPEs.

First, we grouped them into genealogical patrilineages according to BALSAC. Hereafter, unless specified, the term *patrilineage* will refer to *genealogical patrilineage*.

Second, we calculated the number of haplotype differences for each pair of individuals in a tree and set a threshold of 20 haplotype differences. We assume that two samples with more than 20 differences did not find their common biological ancestor within the genealogy. We aimed at defining a NPE independently of haplogroup classification.

Third, we built a graph with the set of individuals as vertices. If the number of pairwise differences between i and j was larger than the previously determined threshold, we connected two vertices, i and j.

Then, we iteratively removed individuals from the set of vertices to obtain a completely disconnected graph (i.e., a tree with no NPE). At each step, we checked where the most likely NPE have occurred in the tree (see Section A.3 for an example of the algorithm applied to a tree with seven terminal leaves and the general explanation of the algorithm).

Finally, the number of NPEs equals the number of iterations in the previous step. For each tree k, let us denote by p_k the number of NPEs.

For each tree, we found the MRCA of its leaves and counted the number of meioses m_k that took place in the subtree having the MCRA as root. Denoting by N the total number of patrilineages, we define the per-generation NPR rate as

$$p = \frac{\sum_{k=1}^{N} p_k}{\sum_{k=1}^{N} m_k}, \tag{2.1}$$

i.e., the non-paternity events summed across all trees and divided by the total number of meioses

in the sample.

2.3.1 False negative rate estimate

We used the following approach to determine the performance of our method at detecting NPEs. First, we selected one random individual from a random patrilineage. Then, we selected a second individual from a different patrilineage and swapped it with the first individual. This step creates a virtual NPE in the first lineage as the second individual does not share a known genealogical paternal ancestor within the last ~400 years with the members of the first patrilineage. Finally, we computed the number of NPE in the patrilineage described above. The proportion of times where our algorithm does not detect the NPE is an estimate of the false-negative rate.

2.4 Point mutation rate

2.4.1 Removal of genealogical errors

To compute the mutation rate, we excluded individuals with whole Ychr data indicating genealogical errors. For this, we first computed the number of sequence differences within each pair of individuals in the same patrilineage and the number of generations separating them. Following the rule in Helgason et al. (2015), we set a threshold for the maximum number of differences allowed in pairs of $d_{max}(m) = 5 + 2m$, where m is the number of meioses that separate the pair of individuals (see Section A.4). In patrilineages with only two individuals, we excluded both if their sequence differences exceeded the threshold. In larger patrilineages, we removed individuals with differences above the threshold relative to the rest of their patrilineage.

2.4.2 Per-generation estimation

We propose two approaches to estimating the mutation rate in the Ychr: as a tree-wise mutation rate and as a branch-wise mutation separating between internal and terminal branches of a patrilineage. In both methods, we started by counting the number of polymorphic sites in a genealogical clade. We then removed positions that appeared as polymorphic in more than one tree.

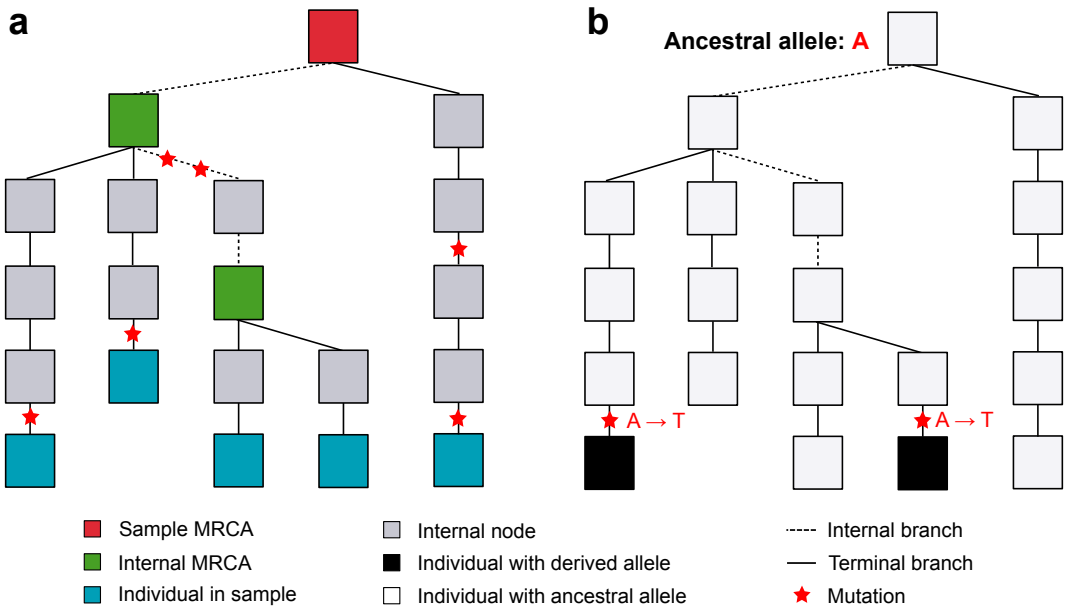

Figure 2.2: Diagram of mutations on a patrilineage and homoplasy. (a) A patrilineage can be split into terminal (a branch connecting a leaf to a MRCA) and internal (a branch connecting two MRCA). Mutation rates can be defined for each branch type (see text). (b) The structure of the patrilineage allows us to determine homoplasies. We removed from the analysis positions that could only be explained through recurrent mutations.

Let π_k and m_k be the number of polymorphic sites and meioses under the MRCA in the k-th patrilineage, respectively (recall Section 2.3). Noting ℓ the length of the analyzed portion of the genome, we estimate the tree-wise PGMR as

$$\mu = \frac{\sum_{k=1}^{N} \pi_k}{\ell \sum_{k=1}^{N} m_k}, \tag{2.2}$$

i.e., the sum of polymorphisms over all trees is divided by the total number of generations multiplied by the length of the analyzed portion of the genome.

To obtain a per-branch mutation rate, we first split a patrilineage into internal and terminal branches. Then, we assigned mutations to individual branches in the tree. To identify where mutations occurred, we retrieved the haplogroup of the leaves and built a consensus sequence of the most recent haplogroup ancestral to the leaves in the tree. For example, if the haplogroup of the tree was R1b1a1b1a1a, we obtained a consensus sequence for the haplogroup R1b1a1b and declared this sequence the ancestral sequence.

Mutations can occur in *terminal* or *internal* branches (see Figure 2.2a). An internal branch is a set of edges in the patrilineage between two internal MRCAs or between an internal MRCA and

the sample MRCA. A terminal branch is a set of edges between a leaf and any MRCA. We use the superscripts (i) and (t) to refer to internal and terminal branches, respectively. In Figure 2.2.a, the number of mutations in internal branches is $\pi^{(i)} = 2$ and in terminal branches it is $\pi^{(t)} = 4$. We define the per-branch mutation rates for internal and terminal branches, respectively, as

$$\mu^{(i)} = \frac{\sum_{k=1}^{N} \pi_k^{(i)}}{\ell \sum_{k=1}^{N} m_k^{(i)}} \quad \text{and} \quad \mu^{(t)} = \frac{\sum_{k=1}^{N} \pi_k^{(t)}}{\ell \sum_{k=1}^{N} m_k^{(t)}}. \tag{2.3}$$

It is clear that $\mu = \mu^{(i)} + \mu^{(t)}$. Additionally, we utilized the tree structure to filter out mutations for which there is no parsimonious explanation given the genealogy, as per Figure 2.2.b.

2.4.3 Assignment of mutations to Ychr sequence classes

We assigned mutations to the discrete classes in the MSY defined in Helgason et al. (2015). In that study, the regions were defined in the GRCh37 Human genome assembly based on the work by (Skaletsky et al., 2003). We applied the liftOver software to annotate the regions in the GRCh38 version of the assembly (Hinrichs et al., 2006). Three regions were not correctly assigned to the new genome assembly and were thus removed. These included a ~0.3 Mb PAL region, a ~0.15 Mb in the rAMP region, and the ~1.8 Mb-long sequence of the heterochromatin of the centromere.

2.4.4 Per-year estimation

The BALSAC genealogy includes the marriage date of each individual and their parents'. We utilized this information as a proxy for the paternal age. More precisely, we estimated an individual's age as the difference between their parents' marriage date and theirs. For the k-the tree, let us denote by a_k the summed age of all the individuals up to the most common recent ancestor of the leaves. Then, we define the PYMR as

$$\mu = \frac{\sum_{k=1}^{N} \pi_k}{\ell \sum_{k=1}^{N} a_k}, \tag{2.4}$$

and give an analogous definition to (2.3) for the per-year branch-wise mutation rate. Individuals who had either a missing marriage date or a missing parental marriage date had their age imputed with the mean paternal age of the sample.

Chapter 3

Results

3.1 Haplogroup classification

3.1.1 General population

We successfully assigned 13,280 men (97% of the sample) to a Y-haplogroup. We excluded an additional 359 samples from haplogroup analysis due to insufficient data on key mutations necessary for accurate haplogroup assignment. We obtained 105 distinct subhaplogroups from all of the 18 main haplogroups. The most common haplogroup found was R1b1a1, comprising 62.4% of the sample. Of the men assigned to this haplogroup, 88% belonged to the R1b1a1b sub-haplogroup (also named R-M269), while the remaining 12% lacked sufficient markers for further classification. The next four most common haplogroups were E1b1b1 (6.4%), R1a1a1 (3.13%), J2a1a1 (2.7%), and E1b1a1 (2.58%). Haplogroups present in over 1% of the sample are listed in Table B.1.

We analyzed the correlation between the haplogroup classification and sample ancestry. We used the clusters defined in Diaz-Papkovich et al. (2023) on the autosomal SNP array data in CARTaGENE as a proxy for genetic ancestry. Clusters in that study were built using HDBSCAN($\hat{\varepsilon}$) applied to UMAP coordinates (Malzer & Baum, 2020). We removed clusters with fewer than 20 samples to maintain anonymity. As expected, haplogroup proportion in clusters correlates with genetic ancestry (Chi-square test: $p < 2.2 \times 10^{-16}$). We can observe several patterns regarding haplogroup composition and country of birth (Figure 3.1). Cluster 0 has the highest proportion of E1b1a1 (64%) of all clusters. This cluster consists primarily of men born in countries from Sub-Saharan Africa or with parents born there. Cluster 1 also displays a clear pattern: with a proportion of 80%, it is the only cluster containing sub-branches of the O haplogroup. 98% of men in this cluster were born in countries from Southeast Asia or have a father born in these countries. Clusters 10, 12, 13, and 14 have the highest proportion of R1b1a1 and consist mainly of individuals born in

Canada, European nations, or former European colonies.

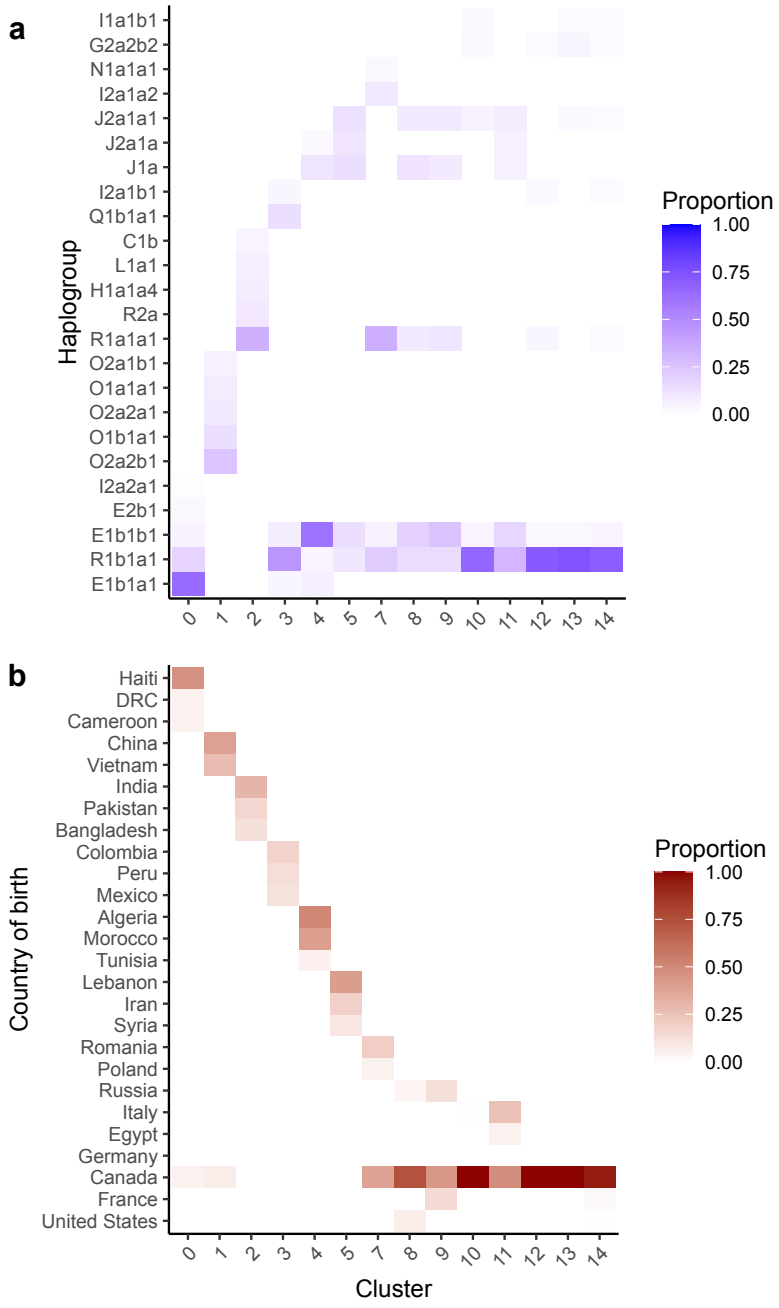

Figure 3.1: Most common haplogroups by ancestry clusters. Clusters were defined in Diaz-Papkovich et al. (2023) using HDBSCAN($\hat{\varepsilon}$) on UMAP coordinates. Table B.2 shows the number of samples in each cluster. We excluded cluster 6 from this study for anonymity reasons. a) The top five haplogroups have the highest proportion by cluster in ascending order. b) Top three most representative countries of birth of members of each cluster in descending order. DRC: Democratic Republic of the Congo.

3.1.2 Historical haplogroup composition

The top seven haplogroups of individuals with genealogical data were R1b1a1 (69.4%), E1b1b1 (4.86%), G2a2b2 (2.45%), I1a1b1 (2.45%), J2a1a1 (2.38%), R1a1a1 (2.31%), and I1a2a1 (2.08%).

We analyzed the haplogroup composition among genealogical founders. First, we found groups of sample individuals with a common genealogical founder. Men labeled as genealogical mismatches due to NPEs were excluded (see Section 3.2). Then, we assigned a consensus haplogroup for each founder based on the most frequent haplogroup observed among their descendants. The top seven haplogroups among 1548 founders spanning approximately 450 years were R1b1a1 (67.0 %), E1b1b1 (4.78%), R1a1a1 (3.29%), I2a1b1 (2.84 %), G2a2b2 (2.33), I1a2a1 (2.07%), and I1a1b1 (1.81%) (Figure 3.2). 85% of the men from the R1b1a1 haplogroup belonged to the R1b1a1b subclade; the rest did not have sufficient markers for a more detailed haplogroup classification. R1b1a1 was consistently the most common haplogroup introduced in Quebec from the 16th century until the 20th century (Figure B.2). All haplogroups with frequencies above 1% were first introduced between 1550 and 1650 and reintroduced across subsequent centuries (Figure B.3).

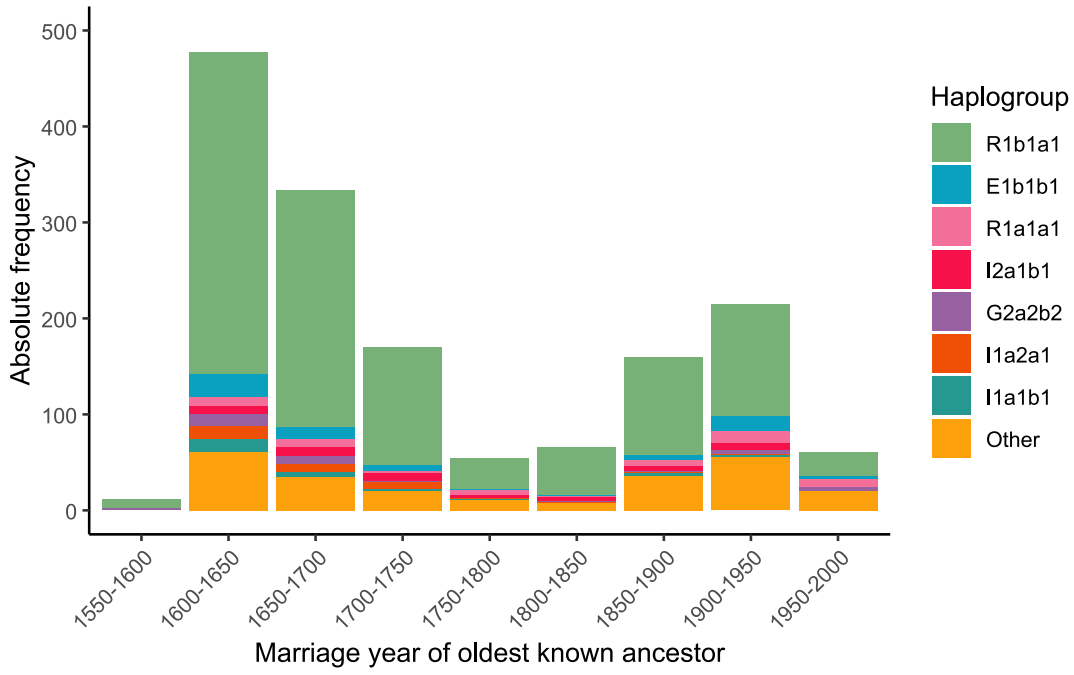

Figure 3.2: Frequency of the top seven most common haplogroups of the male founders by marriage year. Men in the sample were grouped by their oldest known common genealogical ancestor. We estimated the founder's haplogroup as the most common haplogroup of his descendants.

3.2 Non-paternity error rate

Of the 3320 individuals connected to BALSAC, we found 1847 distinct patrilineages. 1340 of these patrilineages included only one individual in the sample, and the rest were patrilineages of two or more individuals (Figure B.1.a). The largest set of individuals sharing a common ancestor had 36 samples. Individuals within the same patrilineage were separated between 1 and 26 generations (Figure B.1.b) with a median of 18 generations.

We calculated the number of haplotype differences between pairs of individuals using SNP array data. A position was accounted for in the comparison only if the genotype of both individuals was known. Figure 3.3 shows the empirical cumulative distribution function (ECDF) of haplotype differences between pairs of individuals. Approximately 95% of pairs of individuals within the same patrilineage have fewer than 20 haplotype differences, the threshold we set to detect NPEs (recall Section 2.3). In turn, approximately 51% of the pairs that did not belong to the same patrilineage also had fewer than 20 haplotype differences.

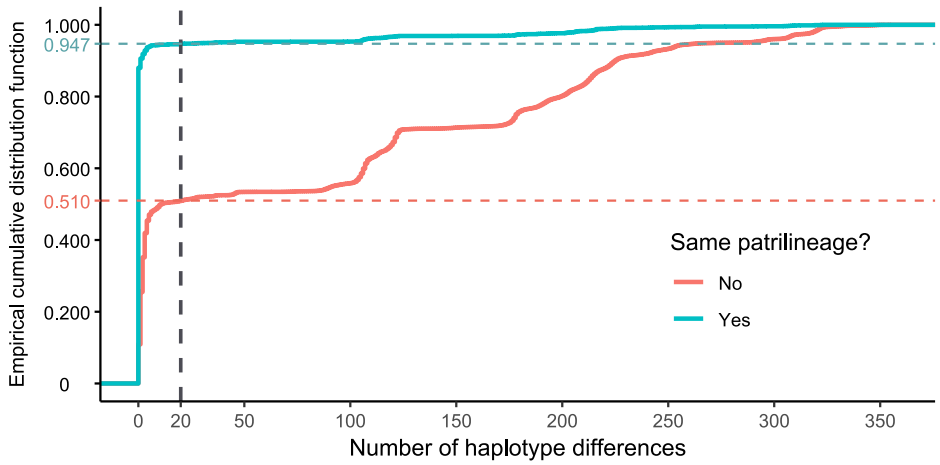

Figure 3.3: ECDF of haplotype differences in SNP array data. Haplotypes of individuals from the same and distinct patrilineages were compared.

We detected 57 NPEs in 507 distinct patrilineages with more than one sample. A total of 14794 meioses summed over all trees gave us an NPR of $\varepsilon = 3.8 \times 10^{-3}$ (Clopper-Pearson exact 95% CI: $2.9 \times 10^{-3} - 4.9 \times 10^{-3}$). We obtained a false negative rate of 0.52 by creating 10^5 virtual NPEs. Accounting for this rate, we report an NPR of 7.3×10^{-3} (95 % CI: $5.5 \times 10^{-3} - 9.4 \times 10^{-3}$).

3.3 Point mutation rate in the Y chromosome

Of 448 individuals with complete MSY sequences, 132 shared a common ancestor with at least one other individual in the genealogy. We identified a total of 52 distinct patrilineages. An additional 16 individuals were filtered out as they were assumed to be genealogical errors (see Section A.4).

We found 223 candidate mutations in the analyzed portion of the genome. We excluded nine candidate mutations from further analysis as we detected them in more than one patrilineage. After assigning mutations to branches, we excluded extra mutation due to its classification as a homoplasy (recall 2.2.b). We tested the distance pairs of candidate mutations to detect potential larger structural variants and found no supporting evidence (see Section B.4)

We assigned mutations to one of the four regions, rAMP, PAL, XDG, and XTR. Table 3.1 presents the treewise mutation rate per generation and per year, along with the transition to transversion ratio (TiTv) ratio for these mutations. 33 mutations could not be classified as they occurred in unannotated regions. We performed a two-sample proportion test using the Benjamini-Hochberg correction for multiple testing (Benjamini & Yekutieli, 2001) and found no differences in the mutation rate in any of the four regions. These findings contrast with those of Helgason et al. (2015), who reported a significant difference in rAMP relative to other regions.

Table 3.1: Mutation counts and rate by sequence region.

	Type of mutations				Mutation rate (95% CI)	
Sequence class	Total	Ti	Tv	TiTv ratio	Per generation ($\times 10^{-8}$)	Per year ($\times 10^{-10}$)
rAMP	51	37	14	2.64	2.19 (1.63 – 2.89)	5.96 (4.44 – 7.84)
PAL	43	26	17	1.52	2.29 (1.65 – 3.08)	6.22 (4.50 – 8.37)
XDG	84	52	32	1.62	2.30 (1.83 – 2.84)	6.24 (4.98 – 7.73)
XTR	2	2	0	—	1.13 (1.37 – 4.11)	3.08 (0.374 – 11.1)
Unclassified	33	19	14	1.35	—	—
All regions included	33	19	14	1.35	2.38 (2.07 – 2.73)	6.47 (5.63 – 7.41)

As there were no significant differences in the mutation rate across the different regions of the Ychr, we combined all regions to estimate the mutation rate. With 1119 meioses in 52 trees and a

7.9Mb-long Ychr region, we found a tree-wise PGMR of 2.398×10^{-8} (95% CI = $2.08 \times 10^{-8} - 2.76 \times 10^{-8}$) and a PYMR of 6.47×10^{-10} (95% CI = $5.63 \times 10^{-10} - 7.41 \times 10^{-10}$). We found no statistically different mutation rates in internal and terminal branches (two-sample proportion test: $p = 0.93$, see Section 2.4.2).

We built an intercept-free linear regression model of the number of mutations with the number of meioses as a predictor (Figure 3.4). We tested the significance of regression coefficients using a permutation test assuming heteroscedastic errors (Helwig, 2023). As expected, the number of meioses was a statistically significant predictor both for the tree-wise ($p = 0.0088$, $R^2 = 0.80$) and the branch-wise number of mutations ($p = 1 \times 10^{-4}$, $R^2 = 0.58$).

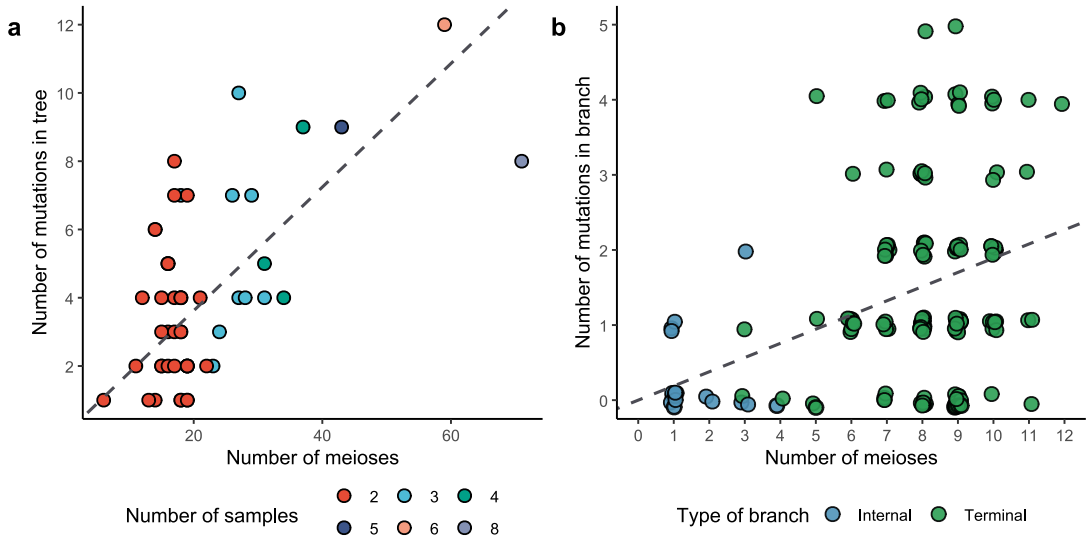

Figure 3.4: Number of mutations in trees and branches by total number of meioses. a) Number of mutations in a patrilineage as a function of the number of meioses down the most recent common ancestor. Colors correspond to the number of samples in the patrilineage. b) Number of mutations in a tree branch as a function of the number of meioses (or generations) found on that branch. An explanation of the type of branches can be found in Section 2.4.2.

We tested whether sequencing errors impacted the results by building a bivariate regression model, adding the number of samples in a patrilineage and the number of meioses as predictors. The rationale behind this is that if sequencing errors were a confounder, patrilineages with more samples would have a higher number of mutations, irrespective of the number of meioses. In the presence of the number of samples, neither the number of meioses nor the number of samples were significant predictors (Huh-Jhun permutation test: $p = 0.80$). The lack of significance can be explained due to colinearity of both predictors (Spearman's correlation test: $p = 5.08 \times 10^{-5}$,

Figure B.6).

Following Helgason et al. (2015), we analyzed the effect of the father's age at conception through a linear regression model of the mutation rate in branches and the mean age of individuals. We removed two branches of length 1 from the analysis as the ages of those individuals were unknown. In contrast to Helgason et al. (2015), the mutation rate in a branch did not correlate with the mean paternal age (Kendall's rank correlation test: $p = 0.7$, Figure B.7). Consequently, the mean paternal age was not a significant linear predictor of the branch mutation rate ($\beta = 0.005$, $p = 0.348$).

Chapter 4

Discussion

4.1 Haplogroup composition

4.1.1 General population in Quebec

We conducted the first large-scale study on the distribution of Y-haplogroups across six metropolitan areas of the province of Quebec. The analysis of the most prevalent haplogroups reflects the migration waves to Quebec since the establishment of New France in the 17th century.

We observed that the haplogroup with the highest frequency in this population is R1b1a1, present in 62.4% of the sample. This result aligns with Moreau et al. (2009), who found R1b as the most common haplogroup in a population from the Gaspé Peninsula. The high frequency of the R1b1a1b sub-haplogroup (found in 55% of the sample) is partially explained by pre-20th-century settlements and migration events of European men in Quebec. For instance, Balaresque et al. (2010) estimated a frequency between 68% and 80% in the present-day regions of Brittany and Pays de la Loire in France, and Ramos-Luis et al. (2014) found it in between 19% and 68% of men across France. High frequencies of R1b1a1b are also reported in Wales (92.3%), Ireland (85%), and the counties of Cornwall and Leicestershire, England (78% and 62%, respectively) (Balaresque et al., 2010). Additionally, 20th-century Latin American migration contributed to this haplogroup, with 44% of Latin American men or their descendants in the sample carrying it.

The haplogroup E1b1b1 in the sample (also known as E-M35) reflects Quebec's diverse migration history. Our genealogical data suggests this haplogroup was introduced in Quebec as early as the 17th century (Figure 3.2). This finding aligns with the evidence of this haplogroup in modern-day French populations (Ramos-Luis et al., 2014). The frequency of E1b1b1 also results from the 20th-century migration from the Maghreb (Azdouz, 2014): 20.9% of the carriers were either born in the Maghreb or had a father born there. Its parental haplogroup E1b1b (also known as E-M215)

is found in frequencies between 50% and 90% in the Berber populations of Morocco and Algeria (Trombetta et al., 2015). Further introductions of E1b1b1 likely arose from late 19th-century migration of the Greeks and Italians to Quebec (Constantinides, 2014; Ramirez, 2014), as E1b1b is also found in low to moderate frequencies in Europe along in the Mediterranean and Balkan regions (Trombetta et al., 2015).

Our study further shows that Y-haplogroup membership correlates with genetic ancestry. Clusters built from autosomal data correlate with the haplogroup composition and individuals' place of birth (Figure 3.1). Cluster 0 is notable for having a high frequency of the E1b1a1 haplogroup, commonly found in Sub-Saharan Africa (E. T. Wood et al., 2005). This cluster's moderate R1b1a1 frequency and members' birthplaces suggest it includes people with Sub-Saharan ancestry or admixture (e.g., from Haiti and the transatlantic slave trade). Other clusters reflect various ancestries: Southeast Asian (cluster 1), South Asian (cluster 2), Latin American (cluster 3), North African (cluster 4), Middle Eastern (cluster 5), Eastern European (cluster 7), and French-Canadian or English-Canadian (clusters 10, 12-14). Clusters 8, 9, and 11 are more challenging to distinguish from Y-haplogroup composition alone. The presence of J subclades suggests these are men with ancestries from Mediterranean or Balkan populations (Italians, Greeks, Bulgarians, Albanians), Middle-Eastern (Syrians, Lebanese) or Jewish populations (both Ashkenazi and Sephardic) (Hammer et al., 2009; Semino et al., 2004).

4.1.2 Genealogical founders

We successfully estimated the haplogroup diversity of 1547 genealogical founders in Quebec. This result enhances our understanding of the haplogroup diversity found in present-day Quebec and offers insights into the potential haplogroup diversity present in the founders' source populations.

Examining present-day haplogroup frequencies allows us to trace the origins of some haplogroups to specific source populations. As previously discussed, the most prevalent haplogroup of the founders, R1b1a1, appears not only in contemporary French populations but also across various European populations, including Great Britain and Ireland (Balaresque et al., 2010; Lall et al., 2021; Ramos-Luis et al., 2014). Until 1765, 88.8% of the genealogical male founders in BALSAC were of French origin, mainly from the regions of Normandie, Poitou, Bretagne, and Ile-de-France

(Vézina, Tremblay, Desjardins, & Houde, 2005). This distribution suggests that the initial influx of R1b1a1 haplogroups was likely of French origin, with later introductions from other European regions. An analogous interpretation applies to E1b1a1, which, while present in France (Ramos-Luis et al., 2014), appears relatively rare in Great Britain—the second most common origin of pre-1765 founders (Capelli et al., 2003).

Conversely, R1a1b1 may have several potential sources. The R1a1b1 haplogroup has been documented in one individual in present-day Paris (Rozhanskii & Klyosov, 2012), and (Ramos-Luis et al., 2014) also reported the finding of individuals in modern-day France carrying R1a haplogroups, though not the R1a1a1 subclade. In contrast, Rozhanskii and Klyosov (2012) reported four individuals from the British Isles with this haplogroup, while Lall et al. (2021) also reported five individuals from Oxford and Dorset, England. According to Vézina et al. (2005), less than 1% of the male founders came from Great Britain between 1600 and 1765. Given the frequency of R1a1a1 in the founders in this period (approximately 2%), this suggests that either R1a1a1 is rare and under-reported in France or most of the founders with this haplogroup were British.

Our haplogroup analysis of founders did not allow us to distinguish historical migration waves from other sources, such as the Irish migration in the 19th century (Jolivet, 2014). As mentioned above, R1b1a1a has been estimated to be highly frequent in Ireland and other European populations. Distinguishing historical migration waves based on haplogroup composition alone would require detailed haplogroup profiling of source populations.

4.2 Non-paternity rate estimate

With a sample of 1980 men from 507 distinct genealogical patrilineages and 4198 Y-SNPs, we report a NPR of 7.3×10^{-3} (95% CI: $5.5 \times 10^{3} - 9.4 \times 10^{3}$) per generation, after accounting for the false-negative rate. This NPR estimate corresponds to observing around 108 approximately 108 NPEs across 14794 meioses. This result is consistent with two previous estimates of the NPR in populations from Quebec: Heyer et al. (1997) (Fisher's exact test: $p = 1$) and Doyon (2018) ($p = 0.71$). Our larger sample size compared to these studies allowed us to obtain smaller confidence and, hence, a more accurate NPR estimate for the Quebec population. In addition, our report also

aligns with the NPRs reported for similarly deep-rooted pedigrees in Flanders by Larmuseau et al. (2013) ($p = 0.69$) and in a Dutch population by Larmuseau et al. (2017) ($p = 0.70$).

In contrast to previous reports focusing on pairs of individuals from independent reproductive events (see Doyon (2018); Larmuseau et al. (2017, 2019, 2013)), our method compares pairs of individuals using underlying tree structure of paternal lineages. The rich genealogical records from BALSAC made this possible, which allowed us to find large groups of men with common genealogical paternal ancestors. Maintaining the tree structure has several advantages. Our method provides the branch where the NPEs could have occurred. Knowing this branch is helpful in genetic studies that rely on the correctness of the genealogy, as individuals below the NPE can be excluded from these studies. Another advantage is identifying multiple NPEs happening in a patrilineage.

A drawback of our method is its dependence on a threshold λ, which sets the maximum allowable number of differences for NPE detection. Consequently, the estimated NPR is also dependent on this threshold (Figure 4.1.a). For comparison, Figure 4.1.a includes the highest estimated NPR in Table 1.2. Low λ values ranging from 1 to 3 yield NPRs that exceed previous reports. Given that the Illumina arrays utilized for the genotyping of the samples have a call rate of 99.5% (Illumina, 2014), 1 to 5 genotype errors are expected in a set of 4198 SNPs. Thus, low thresholds likely overestimate the NPR by generating false positives.

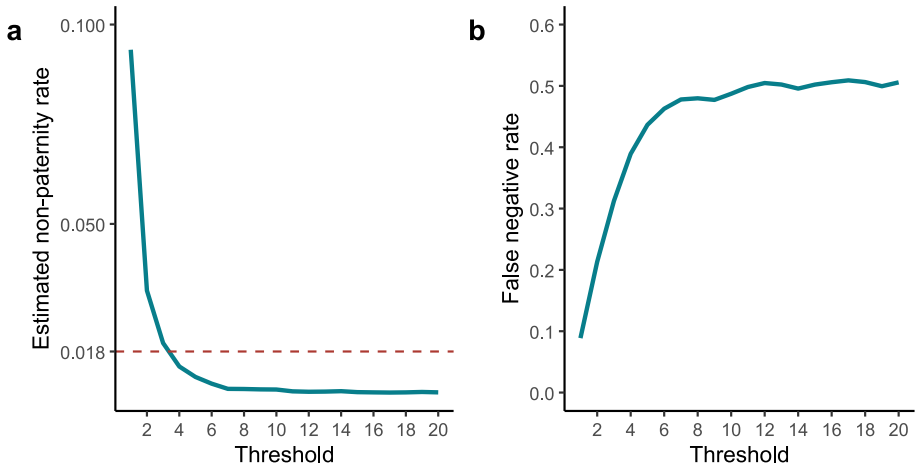

Figure 4.1: Estimated non-paternity rate and false-negative rate as a function of the method's threshold. a) The estimated non-paternity rate after accounting for the false-negative rate. The dotted red line corresponds to the highest NPR estimate in Table 1.2. b) The false-negative rate was estimated according to the method described in Section 3.2.

We produced false-negative rates of our method by introducing an individual of an unrelated

patrilineage to a given patrilineage, thereby simulating virtual NPEs (Figure 4.1.b). For thresholds between 8 and 20, the false-negative rate remains stable at around 0.50. This shows that our chosen threshold of 20 pairwise differences, while conservative to avoid false positives due to genotyping errors, did not substantially inflate the false-negative rate.

We had complete Ychr sequences for 17 men out of the 83 identified as NPE cases through the SNP array data. 4 of the 17 men also found a common paternal ancestor in the genealogy with sequence data. All four individuals were confirmed as mismatches using the threshold outlined in Section 2.4. To accurately estimate the false-positive rate, whole Ychr data would be necessary for all patrilineages with suspected NPEs.

4.3 Mutation rate estimate

Among 132 individuals forming patrilineages of more than one sample, with 212 polymorphisms private to these lineages, we report a PGMR of 2.398×10^{-8} (95% CI: $2.08 \times 10^{-8} - 2.76 \times 10^{-8}$) and a PYMR of 6.47×10^{-10} (95% CI: $5.63 \times 10^{-10} - 7.41 \times 10^{-10}$) in the MSY. This result is consistent with the genealogical mutation rate estimates of Xue et al. (2009) (exact proportion test: $p = 1$). Additionally, the 95% CI of our PYMR intersects that of Balanovsky et al. (2015), suggesting no significant differences.

Unlike Helgason et al. (2015), we did not observe a reduced mutation rate in the PAL region. As a result, our mutation rate estimate falls outside their 95% CI estimation of the mutation rate (both per-generation and per-year) for the combined XDG, rAMP, and XTR regions. Various hypotheses may explain this discrepancy. First, we could not assign 33 mutations (15.5%) of our candidate mutations to any MSY target regions. We extracted the region boundaries defined for the hg37 assembly in Helgason et al. (2015) and updated the coordinates to the hg38 assembly using the liftOver software. Two PAL and rAMP regions did not correspond in the hg38 assembly, and mutations in these regions may account for the discrepancy with Helgason et al. (2015). Second, our threshold on AD may vary across regions in the MSY (see Section 2.1.2). Hence, we could be removing potential polymorphisms differentially across regions. Finally, Helgason et al. (2015), identified 1456 candidate mutations, approximately seven times more than we detected.

Consequently, our power to detect fine-scale differences in distinct regions in the Ychr might be limited.

Despite literature suggesting otherwise, we found no significant effect of paternal age on the mutation rate in branches (see Section 1.5). Several factors can explain this result. Our genealogical data lacks birth years, so we estimated the father's age at conception with the intermarriage interval (recall Section 2.4.4). This results in an underestimation of paternal age as it assumes that each man was born in his father's marriage year. A more accurate approach would involve counting siblings and ordering them according to their respective intermarriage intervals, though unmarried siblings could not be ranked. Thus, our father's age estimates are inherently underestimated in the absence of birth years. Additionally, our stringent filter of genotypes may be limit our ability to detect the paternal age effect by reducing the number of detected mutations. More precisely, we observed both short and long branches (in terms of years) with zero mutations which directly lowers the correlation between parental age and mutation rate (Figures 3.4.b and B.7).

Our genotype filter based on AD aimed to reduce false positives. Nevertheless, some may remain. In their study about mutation rate, Xue et al. (2009) initially found 18 mutations from a SNP array (recall Section 1.5) and, after sequencing, confirmed 12 of them. After genotyping other family members and placing the mutations in the pedigree, they confirmed only 4 of the 12 mutations as occurring *in vivo*. This suggests that in our work, we might be capturing *in vitro* mutations, as most of our mutations are singletons and in patrilineages of two samples (Figure 3.4). However, we found no differences in the mutation rate of internal and terminal branches (Section 2.4). Mutations in internal branches correspond to at least doubletons. They are unlikely to have occurred *in vitro*, which suggests that even if we do have *in vitro* mutations, they do not significantly inflate our estimates.

Chapter 5

Conclusion and Future Directions

In our study, we conducted a large-scale analysis of the Y haplogroup composition in a cohort of 13,280 men from metropolitan areas in Quebec. As expected, the most prevalent haplogroups were also commonly seen in European populations, reflecting Quebec's colonial history. However, we also observed a substantial diversity of haplogroups, indicative of the various ancestral backgrounds stemming from 20th-century migration waves. Additionally, our access to genealogical data allowed us to provide the first known Y-haplogroup description of 1548 founders, 1,025 of whom lived during the colonial era of New France. The analysis of founders offered insights into the haplogroup composition of their modern-day source populations.

Using 4198 Y-SNPs across 3320 individuals, we estimated a historical NPR rate in Quebec to be 0.73%, consistent with previous findings in the same population and other European populations. We developed a method of detecting NPEs based solely on pairwise differences in Y-SNPs. Simulations of NPEs allowed us to estimate the false negative rate, which we found to be moderate.

Our study focused exclusively on paternal lineages and the MSY, without investigating maternal lineages and mtDNA. The CARTaGENE cohort has both SNP-array data and complete sequences of mtDNA for women connected to BALSAC. A previous study by Doyon (2018) examined the maternal NPR in Quebec with a smaller cohort and found it comparable to the paternal NPRs. Investigating a larger cohort like CARTaGENE with a larger number of SNPs in mtDNA would allow us to assess whether the results are consistent in a larger sample. Additionally, using the BALSAC genealogy, we could explore mutation rates in mtDNA.

We intentionally refrained from inferring the sources of the NPEs we observed, which may include extra-pair copulations, undeclared adoptions, errors in the records, etc. Since our method identifies branches likely affected by a NPE, we could collaborate with BALSAC to examine whether these events are related to the fidelity of the records or behavioral factors.

With 212 polymorphisms in 52 patrilineages, we estimated a rate of 2.398×10^{-8} mutations

per generation and 6.47×10^{-10} mutations per year in the MSY. While largely consistent with previous reports, our findings differ in two aspects. Contrary to Helgason et al. (2015), we did not observe a differential mutation rate in four regions or a paternal age effect in our mutation rate. We hypothesize that these differences arise from the stringent allele depth filtering we applied to genotypes. Sequencing additional samples linked to the BALSAC genealogy and calling variants with methods specifically designed for sex chromosomes (e.g., Webster et al. (2018)) could help clarify these discrepancies.

In conclusion, our work affirms the relevance of the Ychr in explaining observed patterns of ancestry within a population of diverse origins, such as Quebec's. We also emphasize, as Calafell and Larmuseau (2017), the benefits of using Ychr to address questions originating outside genetics, such as estimating errors in genealogies. Finally, our work highlights the importance of a rich genealogy such as the one present in Quebec to advance biological understanding—in our case, the mutation rate in the MSY.

References

Amorim, A., Fernandes, T., & Taveira, N. (2019). Mitochondrial DNA in human identification: a review. *PeerJ*, *7*, e7314.

Anderson, K. G. (2006). How well does paternity confidence match actual paternity? evidence from worldwide nonpaternity rates. *Current Anthropology*, *47*(3), 513–520. doi: 10.1086/504167

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., … National Eye Institute, N. I. H. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. doi: 10.1038/nature15393

Awadalla, P., Boileau, C., Payette, Y., Idaghdour, Y., Goulet, J.-P., Knoppers, B., … Laberge, o. b. o. t. C. P., Claude (2012). Cohort profile of the CARTaGENE study: Quebec's population-based biobank for public health and personalized genomics. *International Journal of Epidemiology*, *42*(5), 1285-1299. doi: 10.1093/ije/dys160

Azdouz, R. (2014). Les québécois d'origine maghrébine: entre bricolage, affirmation et reconstruction identitaire. In G. Berthiaume, C. Corbo, & S. Montreuil (Eds.), *Histoires d'immigrations au Québec* (1st ed., pp. 233–249). Presses de l'Université du Québec.

Bagdonaviçius, V., Kruopis, J., & Nikulin, M. (2011). *Non-parametric Tests for Complete Data*. ISTE Ltd and John Wiley & Sons, Inc.

Balanovsky, O. (2017). Toward a consensus on SNP and STR mutation rates on the human Y-chromosome. *Human Genetics*, *136*, 575-590.

Balanovsky, O., Zhabagin, M., Agdzhoyan, A., Chukhryaeva, M., Zaporozhchenko, V., Utevska, O., … Balanovska, E. (2015). Deep phylogenetic analysis of haplogroup G1 provides estimates of SNP and STR mutation rates on the human Y-chromosome and reveals migrations of Iranic speakers. *PLOS ONE*, *10*(4), 1-20. doi: 10.1371/journal.pone.0122968

Balaresque, P., Bowden, G. R., Adams, S. M., Leung, H.-Y., King, T. E., Rosser, Z. H., … Jobling, M. A. (2010). A predominantly neolithic origin for European paternal lineages. *PLoS Biology*, *8*, e1000285.

Bellis, M. A., Hughes, K., Hughes, S., & Ashton, J. R. (2005). Measuring paternal discrepancy

and its public health consequences. *Journal of Epidemiology and Community Health*, *59*, 749-54.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, *29*(4), 1165 – 1188. doi: 10.1214/aos/1013699998

Boattini, A., Sarno, S., Pedrini, P., Medoro, C., Carta, M., Tucci, S., ... Pettener, D. (2015). Traces of medieval migrations in a socially stratified population from Northern Italy. evidence from uniparental markers and deep-rooted pedigrees. *Heredity*, *114*(2), 155–162. doi: 10.1038/hdy.2014.77

Calafell, F., & Comas, D. (2021). The Y Chromosome. In N. Saitou (Ed.), *Evolution of the human genome II: Human evolution viewed from genomes* (pp. 121–136). Tokyo: Springer Japan. doi: 10.1007/978-4-431-56904-6_5

Calafell, F., & Larmuseau, M. H. D. (2017). The Y chromosome as the most popular marker in genetic genealogy benefits interdisciplinary research. *Human Genetics*, *136*, 559-573.

Capelli, C., Redhead, N., Abernethy, J. K., Gratrix, F., Wilson, J. F., Moen, T., ... Goldstein, D. B. (2003). A Y chromosome census of the British Isles. *Current Biology*, *13*(11), 979-984.

Chang, C. C., Chow, C. C., Tellier, L. C. A. M., Vattikuti, S., Purcell, S. M., & Lee, J. J. (2015). Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience*, *4*(1), s13742-015-0047-8. doi: 10.1186/s13742-015-0047-8

Chen, H., Lu, Y., Lu, D., & Xu, S. (2021). Y-LineageTracker: a high-throughput analysis framework for Y-chromosomal next-generation sequencing data. *BMC Bioinformatics*, *22*(1), 114. doi: 10.1186/s12859-021-04057-z

Constantinides, S. (2014). La communauté grecque du Québec: des pionniers du début du XIXe siècle à l'intégration actuelle. In G. Berthiaume, C. Corbo, & S. Montreuil (Eds.), *Histoires d'immigrations au Québec* (1st ed., pp. 111–125). Presses de l'Université du Québec.

Cotter, D. J., Brotman, S. M., & Wilson Sayres, M. A. (2016). Genetic diversity on the human X chromosome does not support a strict pseudoautosomal boundary. *Genetics*, *203*, 485-92.

Crow, J. F. (2000). The origins, patterns and implications of human spontaneous mutation. *Nature Reviews Genetics*, *1*(1), 40–47. doi: 10.1038/35049558

Diaz-Papkovich, A., Zabad, S., Ben-Eghan, C., Anderson-Trocmé, L., Femerling, G., Nathan, V.,

… Gravel, S. (2023). Topological stratification of continuous genetic variation in large biobanks. *bioRxiv*. doi: 10.1101/2023.07.06.548007

Ding, Q., Hu, Y., Koren, A., & Clark, A. G. (2021). Mutation Rate Variability across Human Y-Chromosome Haplogroups. *Molecular Biology and Evolution*, *38*(3), 1000–1005. doi: 10.1093/molbev/msaa268

Direction générale de l'information géospatiale. (2024). *Découpages administratifs.* Retrieved from `https://mrnf.gouv.qc.ca/repertoire-geographique/couches-decoupages-administratifs/`

Doyon, A. (2018). *Dynamique des marqueurs génétiques liés au sexe dans la population canadienne-française pour l'interprétation des traces d'ADN en génétique forensique* (Master's thesis, Université du Québec à Trois-Rivières). Retrieved from `https://depot-e.uqtr.ca/id/eprint/8544/`

Goldmann, J. M., Wong, W. S. W., Pinelli, M., Farrah, T., Bodian, D., Stittrich, A. B., … Niederhuber, J. E. (2016). Parent-of-origin-specific signatures of de novo mutations. *Nature Genetics*, *48*(8), 935–939. doi: 10.1038/ng.3597

Greeff, J. M., & Erasmus, J. C. (2015). Three hundred years of low non-paternity in a human population. *Heredity*, *115*(5), 396–404. doi: 10.1038/hdy.2015.36

Hammer, M. F., Behar, D. M., Karafet, T. M., Mendez, F. L., Hallmark, B., Erez, T., … Skorecki, K. (2009). Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood. *Human Genetics*, *126*, 707-17.

Helgason, A., Einarsson, A. W., Guðmundsdóttir, V. B., Sigurðsson, a., Gunnarsdóttir, E. D., Jagadeesan, A., … Stefánsson, K. (2015). The Y-chromosome point mutation rate in humans. *Nature Genetics*, *47*, 453-7.

Helwig, N. E. (2023). nptest: Nonparametric bootstrap and permutation tests [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=nptest` (R package version 1.1)

Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., & de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Human Molecular Genetics*, *6*(5), 799-803. doi: 10.1093/hmg/6.5.799

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., … Kent, W. J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, *34*, D590-D598.

Hughes, J. F., & Page, D. C. (2015). The biology and evolution of mammalian Y chromosomes [Journal Article]. *Annual Review of Genetics*, *49*(Volume 49, 2015), 507-527. doi: https://doi.org/10.1146/annurev-genet-112414-055311

Illumina. (2014). *Infinium® Genotyping Data Analysis* (Tech. Rep.). Retrieved from `https://www.illumina.com/Documents/products/technotes/technote_infinium_genotyping_data_analysis.pdf`

Illumina. (2020). Illumina DRAGEN Bio-IT Platform v3.5 User Guide [Computer software manual]. Retrieved from `https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/dragen-bio-it/dragen-bio-it-platform-v3.5-user-guide-1000000111887-00.pdf`

Illumina. (2024). *Broad Public Datasets.* Retrieved from `https://42basepairs.com/browse/gs/broad-public-datasets/IlluminaGenotypingArrays/metadata`

International Society of Genetic Genealogy. (2020). *Y-DNA Haplogroup Tree 2019-2020.* Retrieved from `https://isogg.org/tree/`

Jobling, M. A., & Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics*, *4*(8), 598–612. doi: 10.1038/nrg1124

Jolivet, S. (2014). Une histoire des Irlandais et de leur intégration au Québec depuis 1815. In *Histoires d'immigrations au Québec* (1st ed., pp. 25–41). Presses de l'Université du Québec.

Kessler, M. D., Loesch, D. P., Perry, J. A., Heard-Costa, N. L., Taliun, D., Cade, B. E., … Zoellner, S. (2020). De novo mutations across 1,465 diverse genomes reveal mutational insights and reductions in the amish founder population. *Proceedings of the National Academy of Sciences*, *117*(5), 2560-2569. doi: 10.1073/pnas.1902766117

King, T. E., Fortes, G. G., Balaresque, P., Thomas, M. G., Balding, D., Delser, P. M., … Schürer, K. (2014). Identification of the remains of King Richard III. *Nature Communications*, *5*(1), 5631. doi: 10.1038/ncomms6631

Kuroki, Y., Toyoda, A., Noguchi, H., Taylor, T. D., Itoh, T., Kim, D.-S., … Fujiyama, A. (2006).

Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nature Genetics*, *38*(2), 158–167. doi: 10.1038/ng1729

Lall, G. M., Larmuseau, M. H. D., Wetton, J. H., Batini, C., Hallast, P., Huszar, T. I., … Jobling, M. A. (2021). Subdividing Y-chromosome haplogroup R1a1 reveals Norse Viking dispersal lineages in Britain. *European Journal of Human Genetics*, *29*(3), 512–523. doi: 10.1038/s41431-020-00747-z

Larmuseau, M. H. D., Claerhout, S., Gruyters, L., Nivelle, K., Vandenbosch, M., Peeters, A., … Decorte, R. (2017). Genetic-genealogy approach reveals low rate of extrapair paternity in historical Dutch populations. *American Journal of Human Biology*, *29*(6), e23046. doi: 10.1002/ajhb.23046

Larmuseau, M. H. D., van den Berg, P., Claerhout, S., Calafell, F., Boattini, A., Gruyters, L., … Wenseleers, T. (2019). A historical-genetic reconstruction of human extra-pair paternity. *Current Biology*, *29*(23), 4102–4107.e7. doi: 10.1016/j.cub.2019.09.075

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

Malzer, C., & Baum, M. (2020). A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)* (p. 223-228). doi: 10.1109/MFI49285.2020.9235263

Math Exchange User. (2018). *On the difference of two discrete uniform random variables.* Mathematics Stack Exchange. Retrieved from `https://math.stackexchange.com/q/2933243` (URL:https://math.stackexchange.com/q/2933243 (version: 2018-09-27))

Moreau, C., Vézina, H., Yotova, V., Hamon, R., de Knijff, P., Sinnett, D., & Labuda, D. (2009). Genetic heterogeneity in regional populations of Quebec – Parental lineages in the Gaspe Peninsula. *American Journal of Physical Anthropology*, *139*, 512-22.

National Center for Biotechnology Information. (2024). *X-chromosome gene list.* Retrieved from `https://www.ncbi.nlm.nih.gov/gene/?term=X[CHR]+AND+"Homo+sapiens"[Organism]+AND+(("genetype+miscrna"[Properties]+OR+"genetype+`

ncrna"[Properties]+OR+"genetype+rrna"[Properties]+OR+"genetype+
trna"[Properties]+OR+"genetype+scrna"[Properties]+OR+"genetype+
snrna"[Properties]+OR+"genetype+snorna"[Properties])+OR++"genetype+
protein+coding"[Properties]+AND+alive[prop])

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., ... Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44-53. doi: 10 .1126/science.abj6987

Page, D. C., Harper, M. E., Love, J., & Botstein, D. (1984). Occurrence of a transposition from the x-chromosome long arm to the y-chromosome short arm during human evolution. *Nature*, *311*(5982), 119–123. doi: 10.1038/311119a0

Purcell, S., & Chang, C. (2020). *Plink 1.9.* Retrieved from https://www.cog-genomics.org/plink/1.9/

Püschel, H. P., Bertrand, O. C., O'Reilly, J. E., Bobe, R., & Püschel, T. A. (2021). Divergence-time estimates for hominins provide insight into encephalization and body mass trends in human evolution. *Nature Ecology & Evolution*, *5*(6), 808–819. doi: 10.1038/s41559-021-01431-1

Ramirez, B. (2014). Immigrants Italiens dans l'espace social et culturel montréalais: une synthèse historique. In G. Berthiaume, C. Corbo, & S. Montreuil (Eds.), *Histoires d'immigrations au Québec* (1st ed., pp. 43–59). Presses de l'Université du Québec.

Ramos-Luis, E., Blanco-Verea, A., Brión, M., Van Huffel, V., Sánchez-Diz, P., & Carracedo, A. (2014). Y-chromosomal DNA analysis in French male lineages. *Forensic Science International: Genetics*, *9*, 162-168. doi: https://doi.org/10.1016/j.fsigen.2013.12.008

Rhie, A., Nurk, S., Cechova, M., Hoyt, S. J., Taylor, D. J., Altemose, N., ... Phillippy, A. M. (2023). The complete sequence of a human Y chromosome. *Nature*, *621*(7978), 344–354. doi: 10.1038/s41586-023-06457-y

Robino, C., Crobu, F., Di Gaetano, C., Bekada, A., Benhamamouch, S., Cerutti, N., ... Torre, C. (2008). Analysis of Y-chromosomal SNP haplogroups and STR haplotypes in an Algerian population sample. *International Journal of Legal Medicine*, *122*(3), 251–255. doi: 10.1007/s00414-007-0203-5

Rozhanskii, I., & Klyosov, A. (2012). Haplogroup R1a, its subclades and branches in Europe

during the last 9,000 years. *Advances in Anthropology*, *Vol. 2*, 139-156. doi: 10.4236/aa.2012.23017

Scelza, B. A., Prall, S. P., Swinford, N., Gopalan, S., Atkinson, E. G., McElreath, R., … Henn, B. M. (2020). High rate of extrapair paternity in a human population demonstrates diversity in human reproductive strategies. *Science advances*, *6*, eaay6195.

Schmid, M., Guttenbach, M., Nanda, I., Studer, R., & Epplen, J. T. (1990). Organization of DYZ2 repetitive DNA on the human Y chromosome. *Genomics*, *6*, 212-8.

Semino, O., Magri, C., Benuzzi, G., Lin, A. A., Al-Zahery, N., Battaglia, V., … Santachiara-Benerecetti, A. S. (2004). Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *American Journal of Human Genetics*, *74*, 1023-34.

Sin Lo, K. (2023). *Additional information about the CARTaGENE genetic data* (Tech report). CARTaGENE. Retrieved from `https://cartagene.qc.ca/files/documents/other/Info_GeneticData3juillet2023.pdf`

Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., … Page, D. C. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, *423*(6942), 825–837. doi: 10.1038/nature01722

Strassmann, B. I., Kurapati, N. T., Hug, B. F., Burke, E. E., Gillespie, B. W., Karafet, T. M., & Hammer, M. F. (2012). Religion as a means to assure paternity. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, 9781-5.

Tarride, S., Maarand, M., Boillet, M., McGrath, J., Capel, E., Vézina, H., & Kermorvant, C. (2023). Large scale genealogical information extraction from handwritten Quebec parish records. *International Journal on Document Analysis and Recognition*.

The Y Chromosome Consortium. (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Research*, *12*, 339-48.

Trombetta, B., D'Atanasio, E., Massaia, A., Ippoliti, M., Coppa, A., Candilio, F., … Cruciani, F. (2015). Phylogeographic refinement and large scale genotyping of human Y chromosome haplogroup E provide new insights into the dispersal of early pastoralists in the African continent. *Genome Biology and Evolution*, *7*(7), 1940-1950.

Underhill, P. A., Poznik, G. D., Rootsi, S., Järve, M., Lin, A. A., Wang, J., … Villems, R. (2015). The phylogenetic and geographic structure of Y-chromosome haplogroup R1a. *European Journal of Human Genetic*, *23*, 124-31.

Veltman, J. A., & Brunner, H. G. (2012). De novo mutations in human genetic disease. *Nature Reviews Genetics*, *13*(8), 565–575. doi: 10.1038/nrg3241

Vergani, D. (2021). The Y-Chromosomal STRs in Forensic Genetics: Y Chromosome STRs. In E. Pilli & A. Berti (Eds.), *Forensic dna analysis: Technological development and innovative applications* (pp. 77–89). New York: Apple Academic Press. doi: 10.1007/978-4-431-56904-6_5

Vézina, H., & Bournival, J.-S. (2020). An overview of the BALSAC population database. Past developments, current state and future prospects. *Historical Life Course Studies*.

Vézina, H., Tremblay, M., Desjardins, B., & Houde, L. (2005). Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. *Cahiers québécois de démographie*, *34*(2), 235–258. doi: https://doi.org/10.7202/014011ar

Wallis, M. C., Waters, P. D., & Graves, J. A. M. (2008). Sex determination in mammals–before and after the evolution of SRY. *Cellular and Molecular Life Sciences*, *65*, 3182-95.

Webster, T. H., Couse, M., Grande, B. M., Karlins, E., Phung, T. N., Richmond, P. A., … Sayres, M. A. W. (2018). Identifying, understanding, and correcting technical biases on the sex chromosomes in next-generation sequencing data. *bioRxiv*. doi: 10.1101/346940

Willems, T., Gymrek, M., Poznik, G., Tyler-Smith, C., & Erlich, Y. (2016). Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *The American Journal of Human Genetics*, *98*(5), 919-933.

Wood, E. T., Stover, D. A., Ehret, C., Destro-Bisol, G., Spedini, G., McLeod, H., … Hammer, M. F. (2005). Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *European Journal of Human Genetics*, *13*(7), 867–876. doi: 10.1038/sj.ejhg.5201408

Wood, K. A., & Goriely, A. (2022). The impact of paternal age on new mutations and disease in the next generation. *Fertility and Sterility*, *118*, 1001-1012.

Xue, Y., Wang, Q., Long, Q., Ng, B. L., Swerdlow, H., Burton, J., … Tyler-Smith, C. (2009).

Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Current Biology*, *19*, 1453-7.

Appendix A

Supplementary methods

A.1 Distribution of birthplace of the sample

Table A.1: Birthplace of the individuals in the sample. The divisions correspond to the administrative regions defined by the Ministry of Natural Resources and Forests (Direction générale de l'information géospatiale, 2024). An individual's birthplace was inferred using their parents' place of marriage.

Region	Number of samples
Abitibi-Témiscamingue	69
Bas-Saint-Laurent	158
Capitale-Nationale	267
Centre-du-Québec	119
Chaudière-Appalaches	207
Côte-Nord	11
Estrie	186
Gaspésie–Îles-de-la-Madeleine	51
Lanaudière	70
Laurentides	85
Laval	25
Mauricie	167
Montréal	845
Montérégie	198
Nord-du-Québec	1
Outaouais	48
Saguenay–Lac-Saint-Jean	238
Unknown origin or outside Quebec	575
Total	3320

A.2 Distribution of AD for sequence data

Figure A.1 shows the distribution of AD for singletons in the sequence data called the alternate allele. This distribution was obtained after setting to missing those genotypes with an FT field different than PASS (see Section 2.1.2). This explains the low number of genotypes with an AD equal to 0 in Figure A.1.

Genotypes were set to missing if they fell outside the interval 6 to 23. This corresponds to the central 90 % AD distribution of the whole sample (A.1, above). This threshold allowed us to retain approximately 76% of the genotypes of individuals connected to BALSAC (A.1, below).

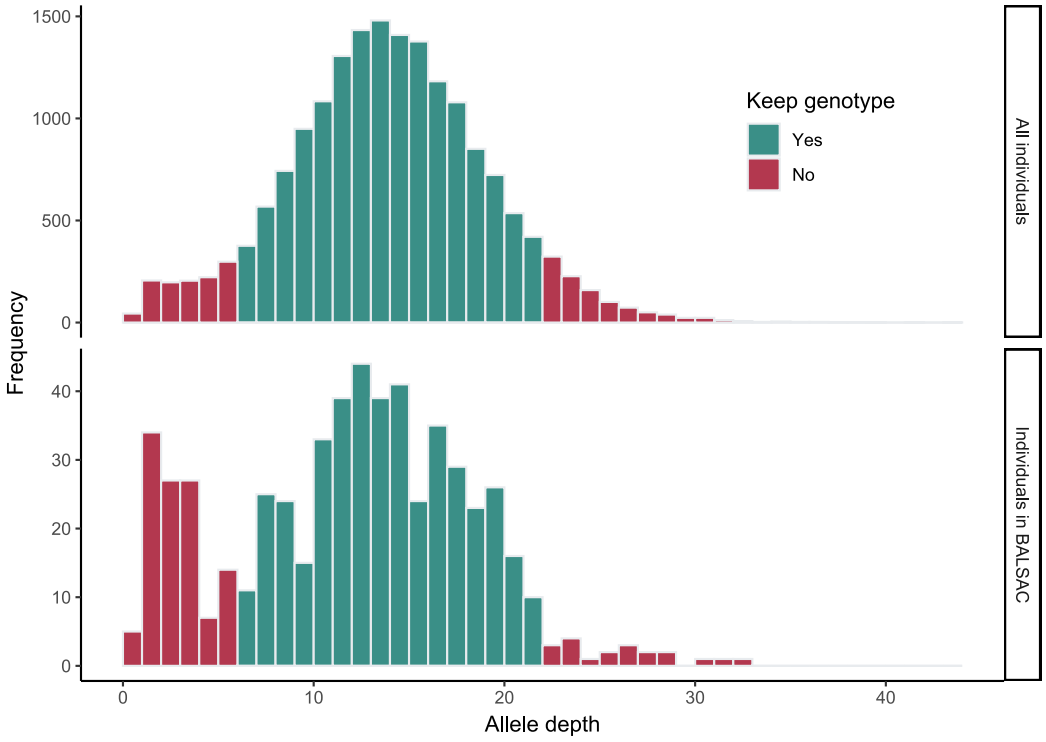

Figure A.1: Histogram of the allele depth of singletons called the alternate allele in the sequence data.

A.3 Calculating errors in a genealogical tree

In Section 2.3, we gave general details on how the algorithm to obtain the number of non-paternity events in a patrilineage works. Here, we show a detailed example of the algorithm applied to a tree with seven leaves.

Let us denote by $\mathcal{L}_0 = \{x_1, \ldots, x_7\}$, the original set of leaves (i.e. the set of sampled individuals). The first step is to calculate the number of pairwise differences in the haplotypes of all leaves. These differences are stored in a 7×7 matrix $\mathbf{D}$. For example, we can have:

$$\mathbf{D} = \begin{array}{c|ccccccc} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ \hline x_1 & 0 & 218 & 223 & 230 & 1 & 217 & 231 \\ x_2 & 218 & 0 & 1 & 1 & 214 & 1 & 1 \\ x_3 & 223 & 1 & 0 & 0 & 235 & 0 & 0 \\ x_4 & 230 & 1 & 0 & 0 & 222 & 0 & 0 \\ x_5 & 1 & 214 & 235 & 222 & 0 & 213 & 223 \\ x_6 & 217 & 1 & 0 & 0 & 213 & 0 & 0 \\ x_7 & 231 & 1 & 0 & 0 & 223 & 0 & 0 \end{array}$$

The second step is to obtain a *thresholding matrix*. Given a threshold λ, we define a new matrix $\mathbf{T}_1^\lambda$ having as entries

$$(\mathbf{T}_1^\lambda)_{ij} = \begin{cases} 1 & \text{if } \mathbf{D}_{ij} \geq t \\ 0 & \text{if } \mathbf{D}_{ij} < t. \end{cases} \tag{A.1}$$

Applying a threshold $\lambda = 20$ to the matrix $\mathbf{D}$ yields

$$\mathbf{T}_1^\lambda = \begin{array}{c|ccccccc} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & x_7 \\ \hline x_1 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ x_2 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ x_3 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ x_4 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ x_5 & 0 & 1 & 1 & 1 & 0 & 1 & 1 \\ x_6 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ x_7 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \end{array}$$

The matrix $\mathbf{T}_1^\lambda$ can be seen as an adjacency matrix of a graph on the set of leaves $\mathcal{L}_0$.

The third step is to calculate the number of discrepancies of an individual with the rest. This is equivalent to analyzing the degree of each sample in the graph. If we sum the entries of $\mathbf{T}_1^\lambda$ row-wise, we can calculate the number of discrepancies between one individual and the

rest. Let us denote by d_1 the number of discrepancies in this algorithm step. In the example, $d_1 = (5, 2, 2, 2, 5, 2, 2)^\intercal$.

The fourth step is to find which individual has the maximum number of discrepancies with the rest. We see that the maximum number of discrepancies is 5 and both x_1 and x_5. This implies that a non-paternity event must have occurred in a branch above the leaves x_1 and x_5. So far, the number of NPEs in this tree is 1.

The fifth step is to remove the samples that include the error. In this case, we will remove x_1 and x_5 from the analysis. We define $\mathcal{L}_1 = \{x_2, x_3, x_4, x_6, x_7\}$ the set of leaves after removing x_1 and x_5. Let us define $\mathbf{T}_1^\lambda$ to be the matrix $\mathbf{T}_0^\lambda$ restricted to the columns in $\mathcal{L}_1$. Thus,

$$\mathbf{T}_2 = \begin{array}{c|ccccc} & x_2 & x_3 & x_4 & x_6 & x_7 \\ \hline x_2 & 0 & 0 & 0 & 0 & 0 \\ x_3 & 0 & 0 & 0 & 0 & 0 \\ x_4 & 0 & 0 & 0 & 0 & 0 \\ x_6 & 0 & 0 & 0 & 0 & 0 \\ x_7 & 0 & 0 & 0 & 0 & 0 \end{array}$$

Since there are no more errors in this matrix, we will stop the algorithm. We conclude that the number of NPEs for this tree is 1 and can be found in a branch above the leaves x_1 and x_5.

The algorithm can be generalized as follows. For a tree with an initial set of leaves $\mathcal{L}_0 = \{x_1, \ldots, x_n\}$:

1) Initialize $p = 0$, the number of errors in the tree.

2) Calculate $\mathbf{D}$ the matrix of pairwise differences.

3) Given a threshold λ, calculate $\mathbf{T}_0^\lambda$ as defined in (A.1).

4) At step k and for a matrix $\mathbf{T}_k^\lambda \in \mathbb{R}^{p \times p}$ with $0 \leq p \leq n$, repeat:

   (a) If $\mathbf{T}_k^\lambda$ is the null matrix, stop.

   (b) Else, define $d^{(k)} \in \mathbb{R}^m$ the row-wise sum of the elements of the matrix $\mathbf{T}_k^\lambda$, i.e. $d_i^{(k)} = \sum_{j=1}^m (\mathbf{T}_k^\lambda)_{ij}$.

(c) Find $\arg\max_{i=1,\ldots,m} d_i^{(k)}$.

(d) Set $p \leftarrow p + 1$.

(e) Define $\mathcal{L}_{k+1} = \mathcal{L}_k \setminus \arg\max_{i=1,\ldots,m} d_i^{(k)}$ the number of leaves after removing the errors.

(f) Define $\mathbf{T}_{k+1}^{\lambda}$ as the restriction of $\mathbf{T}_k^{\lambda}$ to the columns in $\mathcal{L}_{k+1}$.

5) Return p.

A.4 Filtering out sample individuals in sequence data

We compared the haplogroups of pairs of individuals, classifying them based on the comparison results. We defined an *exact match* when both individuals shared the same subhaplogroup. A *partial match* occurred when the haplogroup of one individual was a prefix of the other. *Distinct* haplogroups were those where neither haplogroup was a prefix of the other.

Direct haplogroup comparison of pairs of individuals did not give us a straightforward way of detecting errors in the genealogy due to the lack of resolution of the haplogroup classification obtained (Figure A.2.a). This limited resolution resulted from the genotype filtering described in Section 2.1.2. Instead, we applied the threshold in Helgason et al. (2015) for the maximum number of pairwise differences allowed (Figure A.2.a). The threshold is $d_{\text{max}} = 5 + 2m$, where m is the number of meioses separating two individuals. The rationale behind this threshold is that up to five differences may be expected due to sequencing errors, with an additional $2m$ differences due to random point mutations.

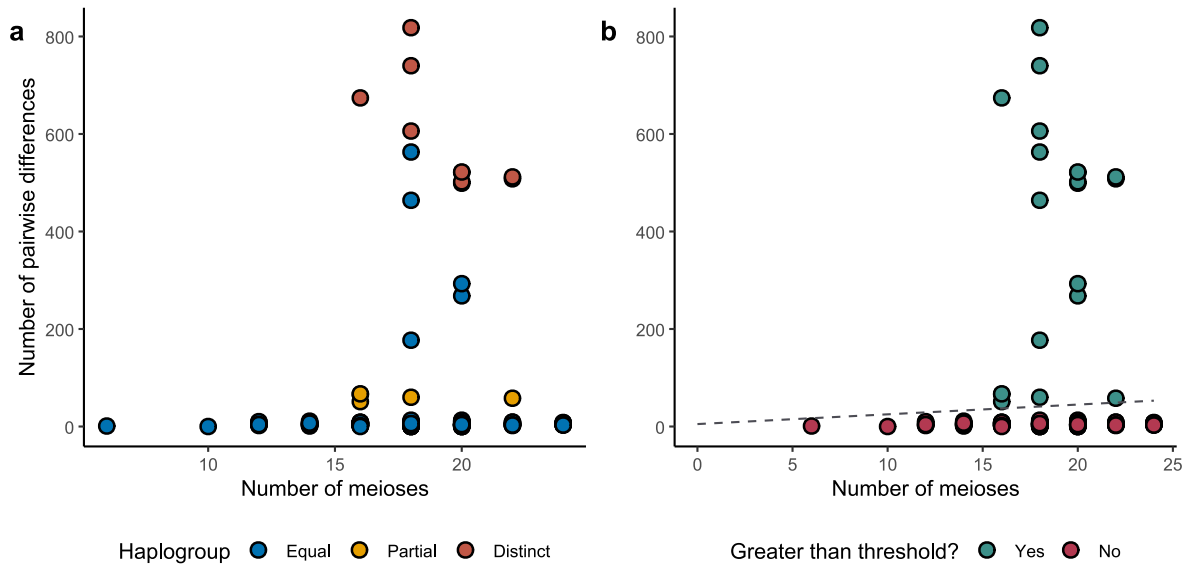

Figure A.2: Number of sequence differences as a function of the number of meioses. Each point corresponds to a pair of individuals in a patrilineage. (a) Points are colored according to the comparison of their haplogroups. A partial match occurs when the haplogroup of one individual is a prefix of the other. (b) Points are colored according to the comparison with the dashed line $d_{\max}(m) = 5 + 2m$, where m is the number of meioses that separate both individuals.

Appendix B

Supplementary results

B.1 Shape of patrilineages in the SNP array data

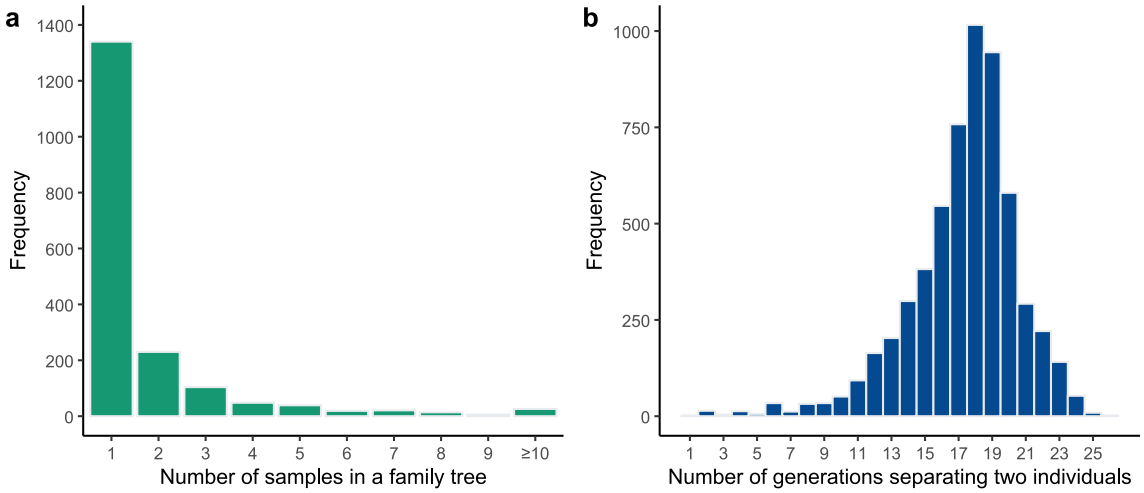

Figure B.1: Shape of patrilineages in the SNP array data. a) Number of samples in a genealogical patrilineage. b) Number of generations separating two individuals within the same patrilineage.

B.2 Haplogroup classification with SNP array data

B.2.1 Most commonly seen haplogroups

Table B.1: Most commonly seen haplogroups found in the CARTaGENE cohort. Only the haplogroups with a frequency higher than 1% in the sample are shown. Haplogroups' names were truncated at most six characters.

Haplogroup	Number of samples	Proportion of sample (%)	Haplogroup	Number of samples	Proportion of sample (%)
R1b1a1	8280	62.4	I1a1b1	260	1.96
E1b1b1	850	6.40	J2a1a	232	1.75
R1a1a1	416	3.13	I1a2a1	216	1.63
J2a1a1	358	2.70	I2a1a1	187	1.41
E1b1a1	342	2.58	J1a	170	1.28
G2a2b2	331	2.49	F	163	1.23
I2a1b1	282	2.12	Other	1191	8.96

B.2.2 Number of samples by cluster

Table B.2: Number of samples by ancestry cluster. Clusters were defined in Diaz-Papkovich et al. (2023).

Cluster	Number of samples	Cluster	Number of samples
0	315	8	102
1	142	9	46
2	100	10	403
3	227	11	466
4	288	12	528
5	185	13	1170
6	13	14	9062
7	233		

B.2.3 Haplogroups of the genealogical founders

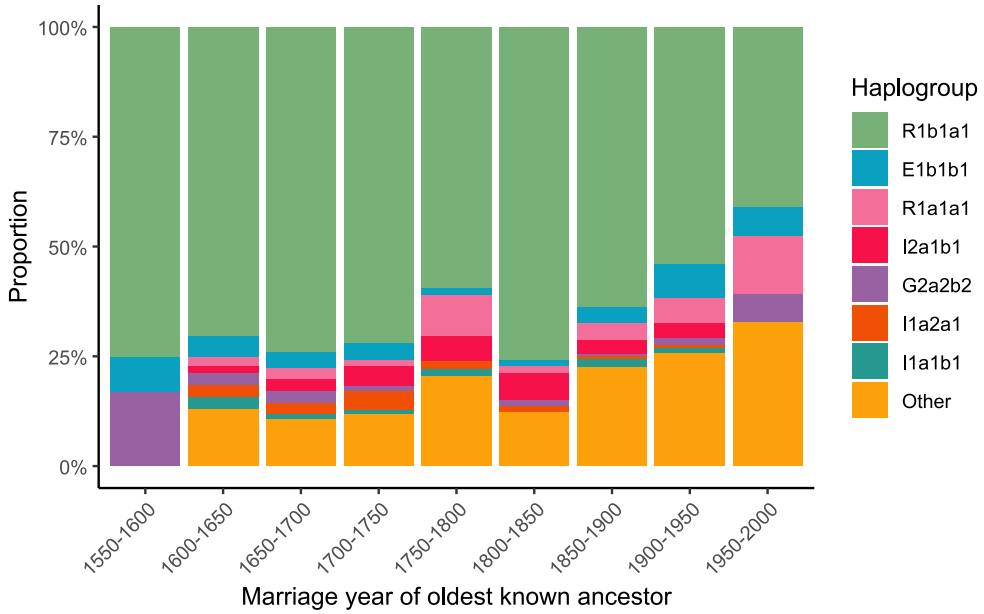

Figure B.2: Proportion of the seven most common haplogroups of the male founders by marriage year. Men in the sample were grouped by their oldest known common genealogical ancestor. The haplogroup of the founder was estimated as the most common haplogroup of his descendants.

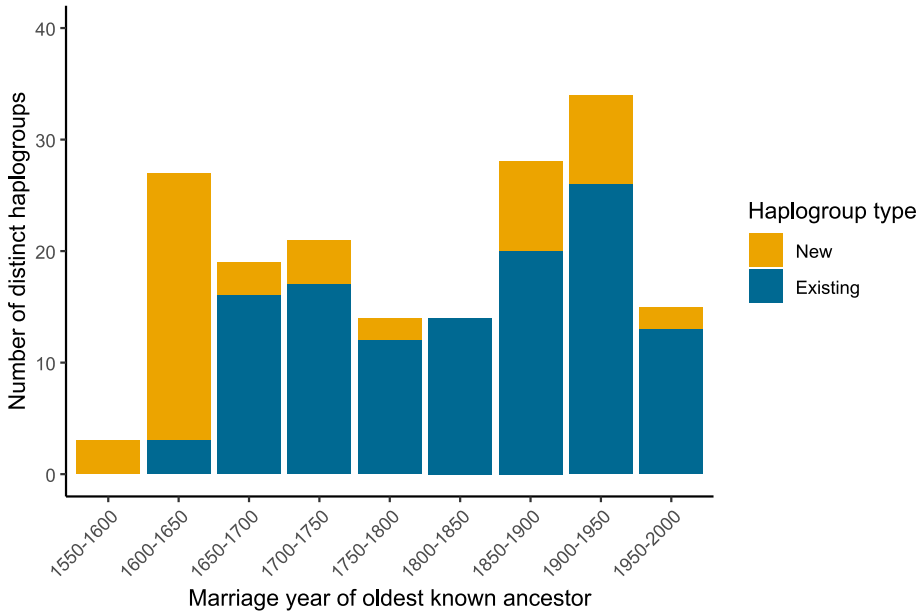

Figure B.3: Number of distinct founder haplogroups by the founder's marriage year. New haplogroups correspond to haplogroups not previously seen in founders before that year. Once they are seen, if reintroduced, they are labeled as *existing*.

B.3 Haplogroup comparison of SNP array data

We expected pairs from the same patrilineage to match exactly or partially. In contrast, we expected distinct haplogroups to signal the existence of a NPEs and correlate with a high number of haplotype differences.

Figure B.4 shows that, as expected, most pairs with a low number of haplotype differences in the SNP array data are either exact or partial matches. However, pairs with distinct haplogroups do not consistently show a high number of haplotype differences. For instance, some pairs share 15 or more digits in their haplogroup names and come from distinct haplogroups yet exhibit very few haplotype differences. Genotyping was done at different phases and with varying sets of SNPs (see Section 2.1.1), which may have led to differences in haplogroup resolution between individuals. Thus, we rejected fine-grain haplogroup classification as a method to detect NPEs, as this approach could cause a high false positive rate.

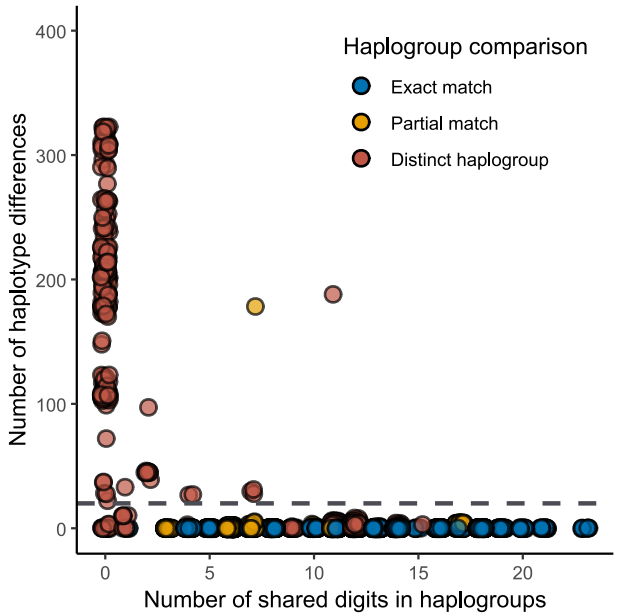

Figure B.4: Number of pairwise differences in the SNP array as a function of the shared haplogroup string. Each point corresponds to a pair of men in the same patrilineage. The dashed line corresponds to 20 haplotype differences, the threshold of differences defined to detect NPEs (see Sections 2.3 and A.3).

B.4 Distribution of distances of mutations

We assume that the place in the genome where SNPs occur follows a discrete uniform distribution in the analyzed region of the portion of the Ychr L. If the observed polymorphisms were truly SNPs and not other structural variants, we would expect the absolute distance of two observed polymorphisms to follow a known distribution (Math Exchange User, 2018). However, because the region consists of a disjoint union of intervals – resulting from applying an accessibility mask (see Section 2.1.2) – this distribution, while still tractable, becomes analytically cumbersome to derive.

In Figure B.5, we compare the observed distribution of absolute differences of polymorphisms and the theoretical distribution. To obtain the observed distribution, we compared the absolute distance of pairs of mutations identified in individuals sharing a common patrilineage. For the theoretical distribution, we simulated 10^7 pairs of variables uniformly distributed in L and computed their absolute difference.

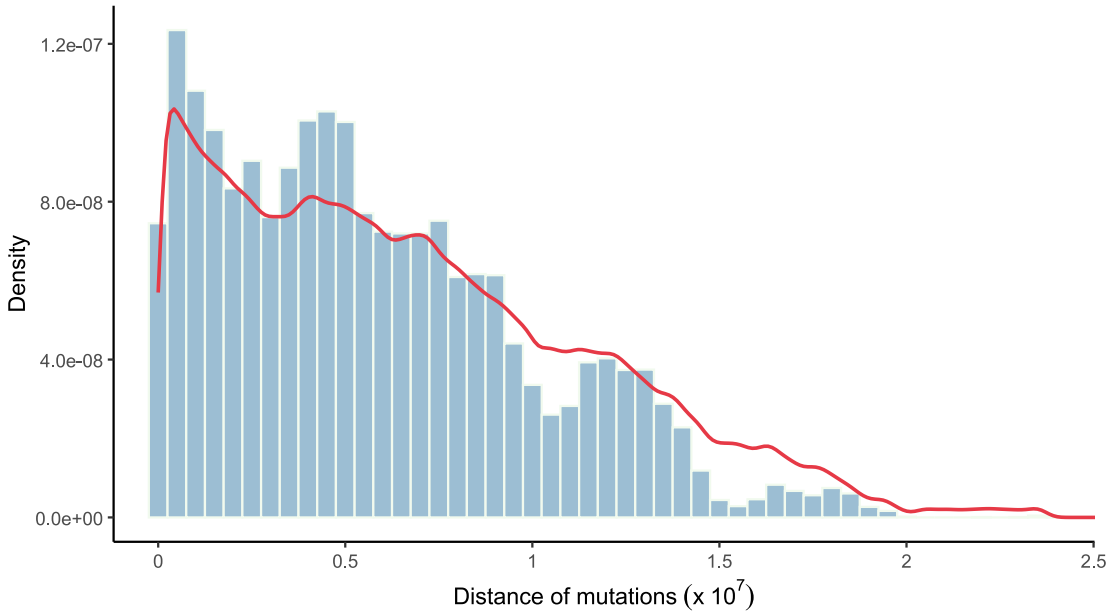

Figure B.5: Histogram of the distance between pairs of mutations. In red, the theoretical simulated distribution assuming mutations follow a discrete uniform distribution in L.

We compared the observed and theoretical distributions using a χ^2 goodness-of-fit test following Bagdonaviçius, Kruopis, and Nikulin (2011). Let $X_1, \ldots, X_n$ be the observed distance of mutations and $Y_1, \ldots, Y_m$ the simulated sample. We divided the range of the observed values into k finite intervals, $\{(a_i, a_{i+1}]\}_{i=1}^{k}$ and defined,

$$O_j = \sum_{i=1}^{n} \mathbf{1}_{(a_j, a_{j+1}]}(X_i) \quad \text{and} \quad E_j = \frac{n}{m} \sum_{i=1}^{m} \mathbf{1}_{(a_j, a_{j+1}]}(Y_i),$$

where $\mathbf{1}_{(a_j, a_{j+1}]}(x)$ is the characteristic function of the interval $(a_j, a_{j+1}]$. We obtained the χ^2 statistic,

$$\chi^2 = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j^2},$$

that is asymptotically χ^2 distributed with $k - 1$ degrees of freedom. Taking $k = 30$, we obtained $\chi^2 = 16.49$ and a p-value of 0.96. Hence, we do not reject the hypothesis that mutations follow a discrete uniform distribution on L.

B.5 Supplementary plots of the mutation rate analysis

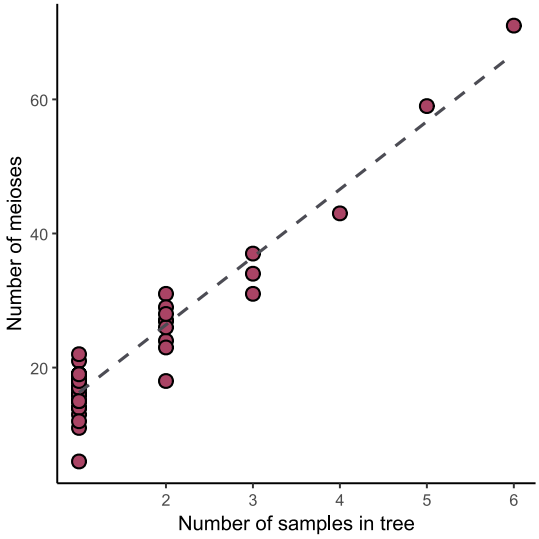

Figure B.6: Number of meioses below the MRCA of a patrilineage as a function of the number of samples in the patrilineage. Each point corresponds to a patrilineage of individuals with complete Ychr sequences. The dotted line corresponds to the best linear model ($R^2 = 0.92$).

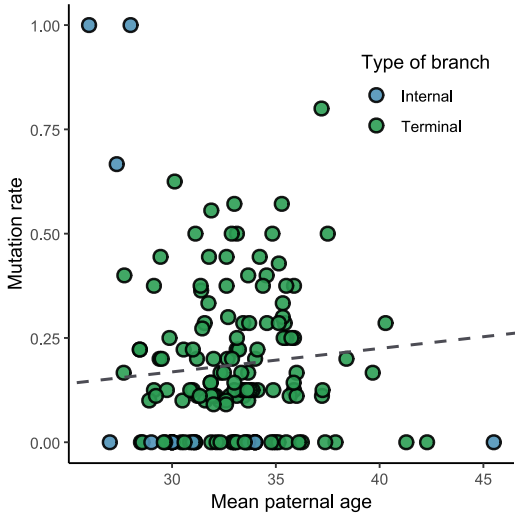

Figure B.7: Mutation rate by mean branch age. Paternal age was estimated as the inter-marriage age of individuals in a branch. *Internal* and *terminal* branches were defined in Section 2.4.2. The dotted line corresponds to the best intercept-free linear model ($\beta = 0.005$, permutation test: $p = 0.348$).