UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

VALIDATION À GRANDE ÉCHELLE D'UN MODÈLE DE SIMULATION POUR DÉTERMINER LA FRÉQUENCE DES HAPLOTYPES Y DANS LA POPULATION CANADIENNE-FRANÇAISE

MÉMOIRE PRÉSENTÉ
COMME EXIGENCE PARTIELLE DE LA
MAÎTRISE EN BIOLOGIE CELLULAIRE ET MOLÉCULAIRE

PAR
ROXANE LANDRY

SEPTEMBRE 2020

« Rien dans la vie n'est à craindre,
tout doit être compris. C'est maintenant
le moment de comprendre davantage,
afin de craindre moins »

Marie Curie

REMERCIEMENTS

Mes premiers remerciements s'adressent à mes parents. Andrée, Gilles, Jacques et Louise, aucun remerciement ne saura exprimer à juste titre la gratitude qui m'habite de vous avoir dans ma vie. Votre amour inconditionnel et votre support tout au long de mes études m'ont permis de persévérer dans les moments les plus difficiles. Vous avez toujours cru en mon potentiel, en mes rêves et ambitions et cette réussite n'aurait assurément pas été possible sans vous.

Je tiens également à remercier mon frère, Patrick, et sa conjointe, Andréane. Malgré la distance, votre amour et votre support étaient présents tout au long de ce processus d'apprentissage. Sans le savoir, votre fierté à me voir persévérer et réussir au cours de ces deux années a été une source de motivation continuelle afin de repousser mes limites et je vous en serai éternellement reconnaissante. Au moment d'écrire ces lignes, vous mettiez au monde une merveilleuse petite fille, la belle Olivia, à qui j'espère un jour pouvoir transmettre ma passion pour les sciences.

Évidemment, l'achèvement de ce projet n'aurait pas été possible sans l'opportunité et le support que m'a offert mon directeur de recherche, Emmanuel Milot. Je serai éternellement reconnaissante de l'aide fournie pendant mon parcours, m'aidant à m'épanouir en tant que scientifique. Je suis sincèrement reconnaissante d'avoir pu développer mes compétences et connaissances scientifiques au sein de son laboratoire, en plus d'avoir eu plusieurs occasions de participer à des congrès provinciaux et internationaux.

Je tiens également à remercier Mikkel Meyer Andersen de l'Université d'Aalborg au Danemark. Je suis reconnaissante pour son accueil pendant les trois mois passés au sein de son laboratoire, mais également pour son aide inestimable pour l'aspect technique et informatique du projet.

Je ne peux passer sous silence l'aide précieuse que j'ai reçue de mes collègues. Le soutien qu'ils m'ont apporté lors des nombreuses difficultés rencontrées tout au long de l'avancement de mon projet de maîtrise a fait de mon parcours une route plus agréable. Certes, leur présence a su rendre mon passage dans le laboratoire beaucoup plus enrichissant et divertissant. Je vous souhaite à tous une bonne continuation en espérant avoir la chance de vous recroiser dans mon chemin de vie et de travailler de nouveau avec vous.

Le support financier durant mon parcours m'a permis de me concentrer entièrement sur mon projet, mais également de participer à un congrès national et un congrès international afin de présenter mes résultats et d'enrichir mes connaissances. Ce remerciement s'adresse donc aux organismes suivants pour leur support financier : le Conseil de recherches en sciences naturelles et en génie du Canada, le Centre international de criminologie comparée, la Société canadienne des sciences judiciaires, le Laboratoire de recherche en criminalistique ainsi que l'Université de Trois-Rivières. Je tiens également à remercier BALSAC pour l'accès aux données généalogiques ainsi que l'équipe de Calcul Québec pour le soutien technique.

Le dernier remerciement, mais non le moindre, s'adresse à tous mes amis qui ont fait partie de cette aventure. Une reconnaissance particulière à ma meilleure amie Marjo qui a été présente dans les hauts et les bas de cette aventure, mais qui a toujours su m'encourager et me soutenir. À ma gang de Sherbrooke, bien que nous ayons tous pris des parcours différents dans des villes différentes, vous avez toujours été présent et de partager cette aventure avec vous a été un réel plaisir. À tous mes amis de Trois-Rivières que cette aventure qu'est la maitrise m'a permis de rencontrer, je tiens à vous remercier. Un merci particulier à Laurence et Paméla avec qui j'ai partagé de nombreux diners accompagnés de longues discussions. Ces moments partagés avec vous font clairement partie des plus mémorables de mon parcours. Un merci particulier à ma première et unique stagiaire Jessie, maintenant devenue une collègue, mais surtout une amie. Bien que nous n'ayons eu que quelques mois à se côtoyer, ta présence au labo m'a permis de connaitre une amie et scientifique incroyable. Partager nos connaissances, nos problèmes et

inquiétudes ensemble a été d'une aide incroyable et une expérience enrichissante. Finalement, à tous les autres trop nombreux pour que je puisse vous nommer un à un, votre présence dans ma vie est tout également importante et vous avez tous eu, à votre façon, un impact positif à différents moments de mon parcours. Je serai éternellement reconnaissante de vous avoir dans ma vie.

AVANT-PROPOS

Depuis près de 30 ans, l'analyse de l'ADN est effectuée dans une grande proportion de dossiers criminels. Dans certains cas spécifiques, comme les dossiers d'agressions sexuelles, il s'avère difficile de distinguer l'ADN d'un contributeur masculin de celui d'une victime féminine. Dans ces cas, les experts en biologie auront parfois recours à l'analyse des marqueurs génétiques situés sur le chromosome Y, puisque celui-ci se retrouve seulement chez les hommes. Cela permet ainsi d'isoler l'ADN d'un potentiel agresseur de celui de la victime. Toutefois, pour diverses raisons génétiques et démographiques qui seront élaborées dans ce mémoire, il est encore difficile d'établir la valeur probante d'une concordance entre le profil ADN obtenu en analysant une trace et celui d'un suspect, malgré les outils statistiques développés dans la dernière décennie. En d'autres mots, il est difficile d'établir la probabilité qu'un autre individu pris au hasard dans la population soit la source du profil ADN à l'origine de la trace retrouvée sur une pièce à conviction plutôt que le suspect lui-même. Ce calcul qui tient compte de la fréquence du profil ADN dans la population permet d'aider les juges et jurys à prendre une décision concernant la culpabilité d'un individu. Les méthodes actuellement existantes ont tous tendance à biaiser les valeurs probantes calculées, faisant en sorte que les biologistes judiciaires doivent être conservateur dans l'interprétation du résultat calculé ce qui sera alors favorable à l'accusé. L'émergence du mouvement #metoo a mené à l'accroissement du nombre de dossiers d'agressions sexuelles soumis pour analyse génétique dans les laboratoires judiciaires. Il est donc actuel et impératif de développer de meilleurs moyens pour calculer la valeur probante, ce qui permettra de mieux outiller les enquêteurs sachant que sans la génétique, la preuve repose principalement sur la parole de la victime contre celle de l'accusé.

Pour contrer ces difficultés, deux chercheurs ont récemment proposé une nouvelle méthode pour attribuer cette valeur probante aux profils ADN obtenus par l'analyse du chromosome Y (haplotypes Y). Leur modèle consiste à simuler par ordinateur une population, puis d'attribuer des haplotypes Y aux individus en spécifiant différents

paramètres génétiques et démographiques. Bien que prometteuse, cette méthode nécessitait d'être testée avec un large jeu de données empiriques.

La population canadienne-française (Québec, CA) est idéale pour valider à grande échelle le modèle proposé ci-haut. Effectivement, fondée par un petit groupe d'individus (~5 000 fondateurs masculins) il y a seulement un peu plus de 400 ans, elle possède une structure démographique retrouvée dans peu de population dû à sa récente fondation. De plus, en raison des églises catholiques qui conservaient les actes de mariage précieusement, nous possédons une fine connaissance de la généalogie du Québec (qui remonte jusqu'à la fondation en 1608), ce qui nous permet de l'utiliser comme outil pour tester le modèle en question comme nulle part ailleurs. L'objectif principal de mon projet de maîtrise consistait donc à tester à grande échelle la nouvelle méthode proposée pour calculer la valeur probante d'une concordance entre deux haplotypes Y en utilisant la généalogie de la population canadienne-française. Une validation de la sorte était nécessaire pour déterminer les applications possibles du modèle, mais aussi les contraintes reliées à son utilisation dans les divers laboratoires judiciaires. Le projet présenté dans ce mémoire propose de nouvelles avenues pour se sortir des carcans des concepts d'interprétations habituels qui ont démontré leurs limites et les résultats qui y sont présentés indiquent qu'il est pertinent et pressant de le faire.

RÉSUMÉ

Depuis la fin des années 90, l'analyse des marqueurs génétiques dans le milieu des sciences judiciaires est en évolution incessante, à un point où la trace d'ADN est parfois considérée comme « la reine des preuves ». Toutefois, l'établissement d'une valeur probante pour présenter la preuve ADN en cours de justice reste problématique dans certain cas, plus particulièrement lors de l'utilisation des marqueurs génétiques situés sur le chromosome Y. Effectivement, le mode de transmission du chromosome Y, qui se fait de père en fils, rend ce dernier sensible à l'effet fondateur ainsi qu'à la dérive génétique. Ces éléments combinés au fait que les bases de données sont constituées d'un nombre limité d'échantillons, qui sont vraisemblablement non représentatifs de la population d'intérêt, rend difficile l'attribution d'une valeur probante à une concordance ADN basée sur un profil obtenu par l'analyse du chromosome Y (haplotype Y).

Pour contrer ce problème, Andersen et Balding ont publié en 2017 une nouvelle méthode consistant à simuler une population afin de déterminer les fréquences d'haplotypes Y sans recours à des bases de données de références. Celle-ci se base sur différents paramètres démographiques et génétiques pour créer une population fictive et attribuer des haplotypes Y permettant d'estimer la distribution d'hommes partageant un même haplotype Y dans une population plutôt que la probabilité de concordance fortuite, telle qu'utilisée présentement. Toutefois, cette méthode prometteuse nécessite une validation empirique avant d'en faire un outil utilisable dans les dossiers judiciaires.

L'objectif principal de mon projet de maîtrise était de tester à grande échelle le modèle de simulation proposé par Andersen et Balding en utilisant les données généalogiques de la population canadienne-française du Québec. Notre étude avait également comme objectif d'étudier l'hétérogénéité spatiale des distributions de fréquences d'haplotypes Y pour déterminer si certains d'entre eux sont plus concentrés dans certaines régions. Les résultats obtenus démontrent qu'un effet fondateur non négligeable est présent dans la population du Québec et que le modèle de simulation n'est pas tout à fait adapté à ce genre de population. De plus, nous avons réussi à démontrer que certains haplotypes sont concentrés à une échelle régionale, ce qui est un élément important à prendre en compte lors de l'attribution de fréquence d'un haplotype Y. Ces résultats remettent en question la définition de la population d'intérêt, principalement au Québec, qui dans certains cas, à l'évidence de nos résultats devraient se définir à une échelle régionale et non à l'ensemble de la province.

Mots-clés : agressions sexuelles, chromosome Y, évaluation de la preuve ADN, généalogie, génétique des populations, microsatellite, science forensique, short-tandem repeats, valeur probante.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AND	Acide désoxyribonucléique
CODIS	Combined DNA Index System
ISFG	International Society for Forensic Genetics
LSJML	Laboratoire de sciences judiciaires et de médecine légale
NRY	Segment non recombinant du chromosome Y
PCR	Réaction en chaîne par polymérase (polymerase chain reaction) RFLP : Polymorphisme de longueur des fragments de restriction
RMP	Probabilité de concordance fortuite (*random match probability*)
RM Y-STR	Répétition en tandem court à mutation rapide situé sur le chromosome Y (*Rapidly Mutating* Y-STR)
RV	Rapport de vraisemblance
STR	Répétition en tandem court (*Short tandem repeat*) SWGDAM : Scientific Working Group on DNA Analysis Method VRS : Variance en succès reproducteur
Y-STR	Répétition en tandem court situé sur le chromosome Y YHRD : Y-STR haplotype reference database

GLOSSAIRE

Décideur de fait[1] :

Personne possédant le pouvoir de décision dans le secteur de la justice (p.ex : juges, jurys).

Allèle[1] :

Chacune des multiples formes que peut prendre un marqueur génétique.

Génétique forensique[2] :

Branche de la science forensique qui utilise les variations génétiques interindividuelles au sein d'une population pour émettre des conclusions sur des événements particuliers à partir de traces matérielles.

Haplotype[1] :

L'ensemble des allèles détectés par l'analyse de l'ADN pour un caractère génétique présent en une seule copie (chromosome Y et ADN mitochondrial).

Locus (pluriel :loci)[1] :

Localisation spécifique d'un segment d'ADN sur un chromosome. Dans le contexte de la génétique forensique, ce terme peut être utilisé comme équivalent de « marqueur génétique ».

Marqueur génétique[1] :

Endroit de l'ADN possédant un polymorphisme pouvant être utilisé pour l'identification des individus. Les marqueurs génétiques peuvent prendre plusieurs formes : VNTR, SNPs, STRs.

---

[1] Coquoz et al., « Preuve par l'ADN : La génétique au service de la justice », 3e éd., 457 pages.

[2] Emmanuel MILOT., « Laboratoire de génétique des populations : Science forensique », [En ligne] https://oraprdnt.uqtr.uquebec.ca/pls/public/gscw031?owa_no_site=4214&owa_no_fiche=8&owa_bottin= (Page consultée le 18 février 2020)

Population d'intérêt[1] :

En statistique, une population correspond à l'ensemble d'éléments à l'étude. En génétique forensique, la population d'intérêt correspond à l'ensemble des individus pouvant être la source de la trace ADN.

Profil génétique[1] :

L'ensemble des allèles détectés chez un même individu suite à l'analyse des différents marqueurs génétiques.

Science forensique[3] :

Science qui étudie la trace en appliquant une démarche et des méthodes scientifiques dans le but de mettre en relation l'origine des activités criminelles avec des individus, des objets et des lieux. Ces éléments viennent en aide aux systèmes de justice afin de statuer sur les causes et circonstances des activités criminelles.

Science judiciaire[4] :

Mise en application des méthodes scientifiques regroupant plusieurs disciplines (chimie, biologie, physique) pour l'étude des pièces à conviction. Les résultats de ces analyses sont interprétés puis transmis aux enquêteurs ou à la cour à des fins d'assistance.

Trace[3] :

Signe, pouvant être visible à l'œil nu ou non, prenant la forme d'une marque ou d'un objet suite à une présence ou une action.

---

[3] Olivier RIBAUX & Pierre MARGOT, « Dictionnaire de la criminologie en ligne : science forensique », [En ligne] http://criminologie.com/article/science-forensique (Page consultée le 12 février 2020)

[4] La société canadienne des sciences judiciaires, « Le scoop sur les vrais sciences judiciaires », [En ligne] https://www.csfs.ca/fr/coin_des_etudiants/que-sont-les-sciences-judiciaires/ (Page consultée le 5 mars 2020)

CHAPITRE I

INTRODUCTION

1.1 Contexte

La trace d'ADN est parfois considérée comme la « reine des preuves » depuis son utilisation dans le domaine de la science forensique. Bien que son introduction aux alentours de 1990 fut une percée majeure pour les sciences judiciaires, elle fut fortement critiquée initialement par certains généticiens et en cours de justice en raison du manque de validation des technologies et de modèles adéquats pour l'interprétation des résultats [1]. Ce qu'on a appelé la « guerre de l'ADN » est depuis longtemps terminée, les méthodes d'analyse et d'interprétation ayant évolué et étant maintenant éprouvées, au point où les demandes d'expertise en ADN ne cessent de croitre [1]. À titre d'exemple, le Laboratoire de sciences judiciaires et de médecine légale (LSJML; Ministère de la Sécurité publique du Québec) qui traite l'ensemble des demandes d'expertises scientifiques touchant aux dossiers criminels (biologie/ADN, toxicologie, chimie, documents, médecine légale, balistique, incendies/explosions) pour le Québec a vu une augmentation de près de 29% de ces demandes entre 2014-2015 et 2018-2019. De plus, le nombre total d'expertises effectuées dans le département de biologie/ADN s'élevait à 5 957 en 2019, représentant 42,7% des expertises effectuées dans ce laboratoire, toutes disciplines confondues [2]. Malgré les avancées, l'interprétation des profils ADN reste un élément matière à critique. Effectivement, cette interprétation faisant appel à l'utilisation de statistiques et de modèles probabilistes, elle peut mener à des difficultés de vulgarisation par le scientifique et à des erreurs de compréhension de la part des acteurs de justice [1, 3]. Le présent mémoire abordera la problématique de l'interprétation des marqueurs génétiques situés sur le chromosome Y.

Avant tout, il importe de bien situer le problème dans son contexte, en faisant l'état des connaissances actuelles et des défis que représentent l'utilisation de l'ADN comme preuve et son interprétation pour la cour de justice.

1.2 Génétique forensique

La science forensique n'est pas un calque de l'anglais *forensic science*, puisque l'expression latine *medicina forensis* apparaissait déjà dans l'Encyclopédie de Dierot et d'Alembert et référait à la médecine légale. Le terme *forensique* dérive du latin *forum* signifiant le lieu public, qui était auparavant le lieu de jugement du peuple [4, 5]. Faisant son apparition à la fin du XIX^{e} siècle et au début du XX^{e} siècle, la science forensique regroupe les connaissances et méthodes de différentes disciplines qui peuvent être subdivisées en trois grands groupes distincts : la physique qui inclut l'analyse de documents, la balistique et les traces d'outils; la chimie qui comprend la toxicologie, l'analyse de fibres, de peinture et de verres; la biologie qui englobe l'analyse de traces ADN, l'analyse des tâches de sang, la pathologie et l'odontologie [5]. Or, la science forensique n'est pas qu'un simple amalgame de disciplines scientifiques mais bien une science à part entière. Elle consiste plus spécifiquement en l'analyse de traces matérielles afin d'établir la nature d'une activité criminelle ainsi que des relations entre des individus, des lieux et des objets en lien avec cette dite activité criminelle. L'ensemble des informations scientifiques récoltées sont interprétées en effectuant des inférences logiques et déductives, puis traduit en informations juridiques afin d'aider les acteurs de la justice à déterminer les circonstances et les causes probables des activités criminelles [4-6].

La génétique forensique est une branche de la science forensique qui découle de la fusion entre la médecine légale et la criminologie. Elle utilise les variations présentes au niveau de l'ADN des individus d'une population pour effectuer des inférences sur des événements spécifiques à partir des traces matérielles [7]. Bien que cette discipline soit largement utilisée pour l'identification humaine dans les dossiers judiciaires, celle-ci possède une plus large application. Effectivement, la génétique forensique s'applique également à l'identification de personnes disparues lors de désastres de masse (p. ex : la

tragédie du Lac-Mégantic, CA), à l'identification d'ossements anciens ou encore au suivi du mouvement des animaux par leur trace ADN, en écologie[2]. L'émergence de la discipline remonte à la découverte des groupes sanguins ABO par Landsteiner en 1900 qui furent utilisés à des fins d'exclusion pendant plusieurs années [8]. Ainsi, les groupes sanguins pouvaient servir à exclure un individu comme source d'une trace ADN, mais étant donné la faible variabilité présente au sein d'une population, cette méthode n'était pas adéquate pour identifier un individu comme étant la source de la trace ADN retrouvée sur une scène de crime, par exemple. Pendant la première moitié du XX^{e} siècle, une quinzaine d'autres groupes sanguins ont été découverts dont les systèmes MNS, Kell, Duffy, Kidd et Lutheran, qui furent utilisés en science forensique. Puis, en 1956, la découverte de protéines polymorphiques dans le sérum, les immunoglobulines, a permis une avancée marquée dans le domaine. Quelques années plus tard, la découverte des antigènes à la surface des lymphocytes, les complexes majeurs d'histocompatibilité (CMH) de classe 1 furent découverts. Les allèles possibles étaient HLA-A, HLA-B et HLA-C et ces marqueurs ont été pendant longtemps utilisés pour les tests de paternité [9, 10].

Plus récemment, des percées en biologie moléculaire ont permis le développement du génotypage de l'ADN. Tout d'abord, la découverte de la structure en double hélice de l'ADN, dans les années 1950, a mené à de nouvelles méthodes d'analyse moléculaire [11]. Parmi celles-ci, le typage des polymorphismes de longueur des fragments de restrictions (RFLP) mena au développement des premières « empreintes d'ADN » individuelles [12, 13]. Cette méthode consiste à digérer le génome entier avec des enzymes de restrictions spécifiques produisant des fragments de tailles variées selon les sites de restrictions (courtes séquences de quelques paires de base) reconnus par les enzymes et dont la présence varie entre les individus. Des sondes sont par la suite utilisées pour détecter la longueur des fragments, créant une sorte de code barre spécifique à chaque individu [14]. Cette technique fut utilisée dans un premier cas d'agression sexuelle suivi d'un meurtre en 1987. En plus de lier deux dossiers entre eux, l'utilisation des RFLP a permis d'exonérer un individu injustement accusé dans ces dossiers [15].

Cependant, une autre avancée scientifique allait peu après remplacer les empreintes RFLP. Effectivement, la réaction en chaîne par polymérase (PCR), développée par Kary

Mullis a permis de caractériser des marqueurs ADN de type *short tandem repeats* (STRs), ou « microsatellites ». La PCR permet d'amplifier l'ADN, c'est-à-dire d'en créer plusieurs copies afin de pouvoir l'analyser par la suite. Son principal avantage est sa grande sensibilité de sorte qu'une très faible quantité d'ADN est suffisante pour obtenir un profil ADN, ce qui n'était pas le cas avec les RFLP. Ainsi, il est possible d'utiliser cette méthode pour amplifier les STRs, puis d'analyser leur taille pour obtenir un profil ADN [16, 17]. Encore à ce jour, ce sont les marqueurs de type STRs qui sont, de loin, les plus utilisés en génétique forensique puisqu'ils offrent un excellent pouvoir discriminant.

1.3 ADN humain

Les cellules de notre organisme sont composées de diverses structures (ex : membrane, mitochondries, cytoplasme, noyau, etc.) qui ont chacun un rôle à jouer dans l'exécution des multiples réactions biochimiques de notre corps. C'est dans le noyau des cellules que l'on retrouve la molécule d'acide désoxyribonucléique (ADN) qui contient le programme génétique des individus. Une exception s'applique, puisque de petites molécules d'ADN (~ 16 500 paires de bases codant pour 37 gènes distincts) sont également retrouvés dans les mitochondries de nos cellules. Or, le présent projet porte sur l'ADN nucléaire et donc l'ADN mitochondrial ne sera pas davantage discuté. Dans le noyau des cellules, l'ADN en double hélice est compacté pour ainsi former des structures que l'on nomme les chromosomes constitués de 3,4 milliards de paires de base [18]. Chez l'humain, on retrouve 22 paires de chromosomes autosomaux (c.-à-d. non-sexuels) ainsi qu'une paire de chromosomes sexuels (X et Y).

Chaque molécule d'ADN contenu dans les différents types de cellules d'un même individu est identique. Toutefois, des variations entre les individus existent dans une faible proportion de l'ADN (~1%). Les régions où des variations interindividuelles existent servent pour l'identification de la source d'une trace ADN [19]. Ces variations sont causées principalement par deux phénomènes distincts : la recombinaison méiotique et les mutations.

La recombinaison méiotique survient lors de la méiose, c'est-à-dire lorsque les cellules diploïdes de la lignée germinale se divisent pour former les gamètes haploïdes

(spermatozoïdes et ovules). Ainsi, à chaque génération, les chromosomes transmis aux enfants diffèreront de ceux de leurs parents [20]. Toutefois, une exception s'applique à cette règle. Le chromosome Y, qui n'est présent que chez les hommes, ne recombine que partiellement aux extrémités (~5% du chromosome) avec le chromosome X lors de la méiose et ainsi chaque fils possédera, à 95%, le même chromosome Y que leur père à l'exception des mutations qui auraient pu survenir.

Les mutations, quant à elles, consistent en la modification d'un ou de plusieurs nucléotides contenant l'information génétique qui peut survenir naturellement de façon aléatoire ou suite à un contact avec un agent mutagène. Les mutations peuvent prendre plusieurs formes (p. ex : altération d'une base azotée, l'insertion ou la délétion d'une ou plusieurs bases azotées), mais auront toutes un même résultat dont l'impact variera: l'altération de la séquence de l'ADN. Lorsque les mutations surviennent dans les cellules germinales, elles seront passées à la descendance ; on parle alors de mutations héréditaires. En comparaison, les mutations qui surviennent dans les cellules somatiques n'auront aucun impact sur la descendance de l'individu.

1.4 Chromosome Y

Le chromosome Y est l'un des plus petit des chromosomes, n'étant composé que de 60 000 paires de bases [18]. Ce dernier, qui détermine le sexe masculin, n'est présent que chez les hommes et est composé de trois régions distinctes : l'hétérochromatine, l'euchromatine et les régions pseudo-autosomales (PARs). Ces dernières, situés aux extrémités du chromosome, composent 5% du chromosome Y. Ce sont ces régions qui se recombineront avec le chromosome X lors de la méiose (**Figure 1.1**). Les 95% restant constitue la portion non recombinante (NRY), composée surtout d'euchromatine, c'est-à-dire de segments d'ADN plus ou moins condensés contenant la majorité des gènes du chromosome Y. La partie non recombinante du chromosome Y restera inchangée lors de la transmission de père en fils, à l'exception des mutations qui pourraient y survenir. Elle contient aussi de l'hétérochromatine, pauvre en gènes codants et en partie composée de séquences répétitives, parmi lesquelles on retrouve les microsatellites utilisés en génétique forensique [21].

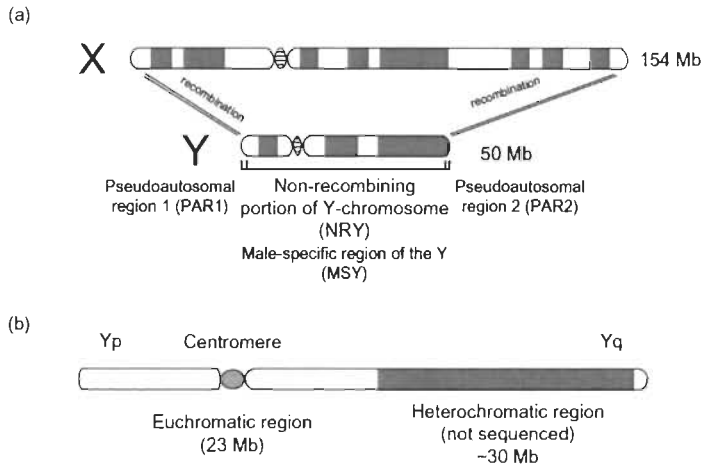

Figure 1.1 Représentation des différentes composantes du chromosome Y (reproduit de Butler 2012, p.375).

En génétique forensique, le chromosome Y est analysé à des fins multiples : la détermination du sexe par l'analyse du gène de l'amélogénine (aussi présent sur le chromosome X), l'étude des relations patrilinéaires dans une même population, les tests de paternité, le soutien à l'identification des victimes lors de désastres de masse, l'identification de suspect dans les dossiers d'agression sexuelle ainsi que la recherche de suspects par l'ADN familial [22-24].

1.5 ADN en génétique forensique

L'ADN est présent dans les différents types de fluides et tissus biologiques tels que le sang, la salive, le sperme, les cheveux ou encore les cellules épithéliales. Lorsque des traces biologiques sont retrouvées sur des scènes de crime, il est alors possible d'en extraire l'ADN pour tenter d'obtenir un profil génétique de la (des) personne(s) à l'origine de la trace. Depuis la découverte de la PCR dans les années 1980, des marqueurs de types *short tandem repeat* (STRs) sont utilisés en génétique forensique pour obtenir des profils ADN. Les STRs consistent en des séries d'unités ou de motifs répétés situés dans des régions non codantes de l'ADN. Il en existe deux types : les minisatellites dont la longueur des motifs répétés varie entre 8 et 100 paires de bases et les microsatellites dont la

longueur varie entre 2 et 7 paires de bases. Ce sont ces derniers qui sont privilégiés en génétique forensique principalement en raison de la petite taille de l'amplicon qui sera produit (<1kb) (**Figure 1.2**). Ils sont donc plus faciles à amplifier et offrent un fort avantage lorsque l'ADN est dégradé, ce qui est souvent le cas avec les traces [25]. Effectivement, la conservation de l'ADN n'est pas toujours optimale : les traces peuvent avoir été exposée à des facteurs environnementaux, comme les rayons UV ou l'humidité, avant d'être récoltées par la police.

Figure 1.2 Exemple d'un short tandem repeat de type microsatellite (reproduit de Butler 2014, p.102).

Ici, l'unité répétitive est formée des paires de bases GATA répétées 11 fois (souligné en bleu dans la figure)

Ces marqueurs génétiques possèdent des polymorphismes de longueur, c'est-à-dire qu'un éventail de nombre de répétitions est possible. Les diverses variantes d'une séquence répétitive donnée forment ce que l'on nomme les allèles. Par exemple, pour le marqueur génétique D8S1179, les allèles possibles sont : 7, 8, 9, 10, 10.2, 11, 12, 12.3,

13, 13.2, 14, 15, 15.1, 15.3, 16, 17, 17.1, 18, 19, 20 [5]. Puisque plusieurs allèles existent pour un même locus et que plusieurs locus sont analysés pour obtenir un profil ADN, ces marqueurs autosomaux sont un bon outil pour l'identification des personnes [26].

Lorsque des pièces à conviction sont acheminées dans un laboratoire judiciaire, elles sont examinées afin de trouver des traces biologiques (salive, sang, sperme, cellules épithéliales ou autres). Des prélèvements sont faits sur ces traces, par exemple par écouvillonnage, découpe d'un morceau de tissu ou sur un mégot de cigarette, ou autres. L'ADN est extrait de ces prélèvements puis amplifié par PCR. Les trousses PCR commerciales comprennent des paires d'amorces permettant d'amplifier plusieurs loci simultanément (jusqu'à ~30); on parle alors de PCR en multiplex. Une fois les amplicons générés, le produit de la PCR est soumis à une migration par électrophorèse dans un capillaire contenant un polymère. Une charge électrique est appliquée, de sorte que les amplicons d'ADN, chargés négativement, migreront vers l'extrémité positive du capillaire, l'anode. La solution de polymère offrant une résistance aux amplicons en migration, la rapidité de cette dernière dépendra de leur taille, les plus petits migrant plus rapidement que les grands, ce qui permet la séparation des différents amplicons en fonction de leur taille. Les amorces de PCR étant couplées à un fluorochrome à leur extrémité 5', ce dernier est excité par un laser et émet une fluorescence au moment précis où l'amplicon passe devant un capteur CCD, permettant ainsi sa détection [27]. Dans les trousses PCR commerciales, au moins quatre fluorochromes distincts sont utilisés. Les amorces (une paire par locus) sont couplées à des fluorochromes différents afin de pouvoir distinguer les loci dont les amplicons ont des tailles chevauchantes. Puisque les amplicons de taille équivalente migreront à la même vitesse et seront donc détectés au même moment lors de l'électrophorèse capillaire, la couleur (longueur d'onde) transmise par le fluorochrome permettra de distinguer un locus d'un autre [28] (**Figure 1.3**).

---

[5] NIST. « STRbase SRD-130», https://strbase.nist.gov/str_D8S1179.htm

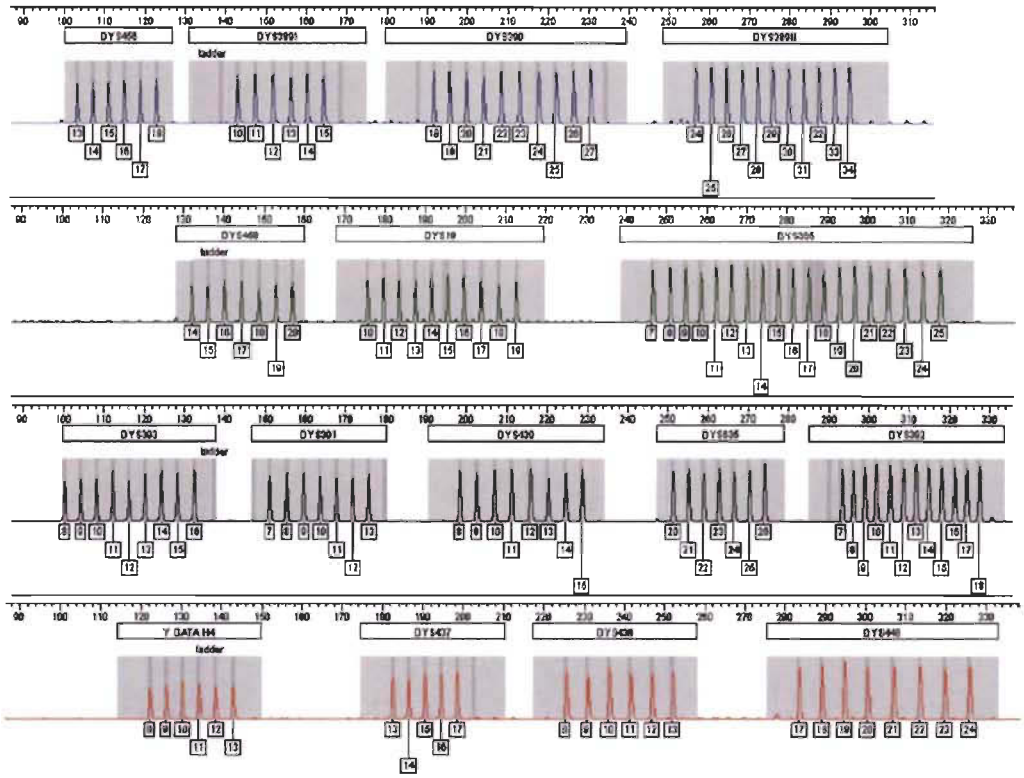

Figure 1.3 Échelle allélique de la trousse commerciale AmpFLSTR™ Yfiler™ démontrant les tailles possibles d'allèles ainsi que la fluorochrome (couleur) associée à chacun des 17 loci inclus dans la trousse (reproduit de Butler 2012, p.381).

1.5.1 *Marqueurs autosomaux*

D'emblée, ce sont les marqueurs STR situés sur les chromosomes autosomaux qui sont utilisés lors d'une expertise ADN pour un dossier criminel. Cela s'explique par le fait qu'étant localisés sur des chromosomes différents ou étant physiquement assez éloignés sur le même chromosome pour être séparés lors de la recombinaison méiotique, on considère que les allèles d'un locus sont transmis des parents à un enfant de manière indépendante. Ainsi, la reproduction sexuée engendre à chaque génération des individus ayant des profils génétiques STR (quasi) uniques. C'est pourquoi les marqueurs STRs ont un grand pouvoir d'individualisation. La première trousse, développée par le Forensic Science Service (FSS) au Royaume-Uni, comprenait quatre loci : vWA, TH01, FES/FPS,

F13A1. Combinés, ils possédaient un pouvoir discriminant de l'ordre de 1 sur 10 000 dans une population caucasienne [29]. Le pouvoir discriminant correspond à la probabilité de tirer au hasard deux individus dans la population possédant des génotypes différents [26]. Ainsi, le pouvoir discriminant s'accroit au fur et à mesure que des marqueurs génétiques sont ajoutés à l'analyse. Quelques années plus tard, une deuxième trousse de six marqueurs a été développée, cette fois offrant un pouvoir discriminant de 1 sur 50 millions. Elle incluait les marqueurs suivants : vWA, TH01, FGA, D8S1179, D18S51, D21S11. En 1996, le FBI a financé le développement d'un projet d'un système de base de données génétiques national connu sous le nom de CODIS (Combined DNA Index System). Le nombre de marqueurs a été élevé à 13 et ce jeu est utilisé depuis dans les laboratoires forensiques d'au moins 28 pays (D8S1179, D21S11, D5S818, CSF1PO, D3S1358, TH01, D13S317, D16S539, TPOX, D18S51, vWA, D7S820, FGA) [30]. Ces 13 marqueurs « de base » sont aujourd'hui inclus dans toutes les trousses PCR vendus dans le commerce. Plus récemment, sept nouveaux marqueurs (D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 et D22S1045) ont été ajoutés au système CODIS qui en comprend donc maintenant 20 [31, 32].

Des mutations dans les cellules germinales de la mère ou du père peuvent survenir dans les loci analysés et donc modifier le génotype qui sera transmis à l'enfant tel qu'expliqué dans la **section 1.3**. Les mutations survenant aux marqueurs STR (autosomaux et Y) consistent principalement en l'ajout ou la délétion d'une ou plusieurs unités répétitives, ce qui entraîne la détection d'un allèle respectivement plus long ou plus court, selon le cas. La probabilité que la mutation provoque l'ajout d'unité(s) répétitive(s) est relativement équivalente à la probabilité qu'elle provoque une (des) délétion(s) [33, 34]. Effectivement, en étudiant 787 mutations, Ballantyne et al. (2010) ont démontré que 423 d'entre elles avaient provoqué la délétion d'unités répétitives alors que 364 mutations avaient provoqué un ajout d'unités répétitives, représentant un ratio de 1,16 :1 [33]. Le taux de mutations des marqueurs autosomaux est de l'ordre de 10^{-3} à 10^{-4} ce qui signifie qu'il survient en moyenne une mutation par locus à toutes les 1 000 ou 10 000 générations [1].

1.5.2 *Marqueurs haploïdes*

Malgré que les marqueurs STR autosomaux (diploïdes) possèdent un grand pouvoir d'individualisation, il est parfois utile ou nécessaire de recourir aux marqueurs haploïdes. Il en existe deux types : ceux sur l'ADN mitochondrial, transmis de la mère à ses enfants, et ceux sur le chromosome Y, transmis de père en fils.

Les marqueurs situés sur le chromosome Y seront discutés davantage dans le présent chapitre, puisqu'ils sont le sujet du présent projet. Leur utilisation sans doute la plus courante en génétique forensique est lorsqu'un mélange ADN homme/femme est obtenu d'une trace. Par exemple, dans les dossiers d'agressions sexuelles, les traces consistent souvent en des prélèvements intimes faits sur une victime féminine [35, 36]. Dans ces cas, il est fréquent qu'un mélange d'ADN masculin/féminin soit obtenu et que l'ADN de la victime, en trop grande proportion (profil majeur), masque l'ADN de l'agresseur qui se retrouve en plus faible concentration (profil mineur). Il devient alors difficile, voire impossible d'isoler le profil génétique autosomal de l'agresseur. En effet, lorsque le ratio ADN homme/femme dans la quantité d'ADN est inférieur à 1/10, les allèles du profil génétique mineur sont généralement peu détectables, étant masqués par les composantes génétiques majoritaires (l'ADN de la victime) [26]. Toutefois, le développement récent de logiciels informatiques permettant d'effectuer du génotypage probabilistique (**section 1.6.2**) rend maintenant possible l'analyse de mélanges où la concentration d'ADN masculin est en faible concentration. Malgré tout, dans plusieurs cas, il est encore difficile voire impossible de détecter le profil ADN masculin dans la trace. C'est alors que le chromosome Y peut être utilisé pour obtenir un profil ADN de l'agresseur, puisque la femme ne possède pas ce chromosome (**Figure 1.4**) [37, 38].

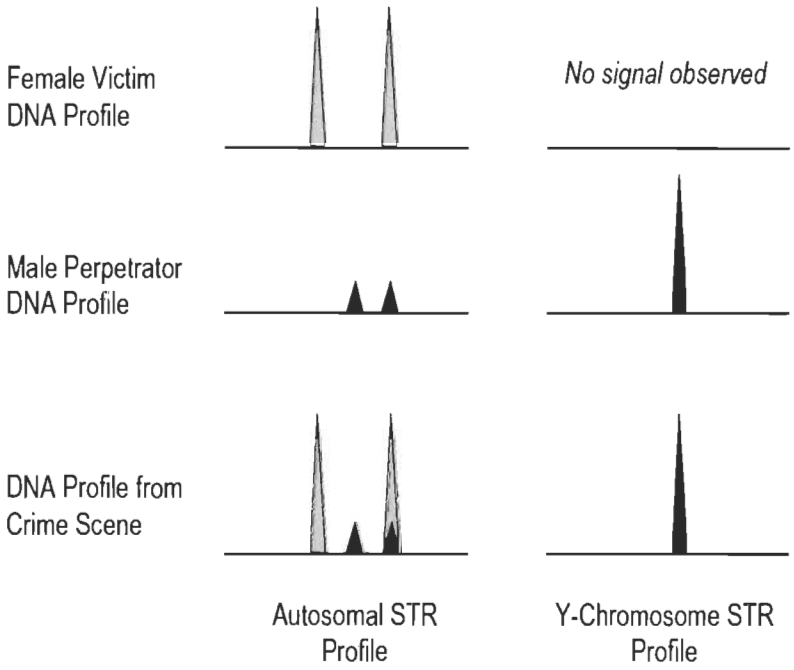

Figure 1.4 Schéma illustrant les allèles qui seraient obtenus d'une trace contenant un mélange de l'ADN de la victime et de celui de l'agresseur pour un marqueur STR autosomal donné et pour un marqueur Y-STR (reproduit de Butler 2012, p.373).

On remarque que l'homme et la femme partagent un allèle autosomal, faisant en sorte que celui de l'homme se trouve masqué sous celui de la femme sur l'électrophorégramme

Étant donné que le chromosome Y n'est présent qu'en une seule copie dans les cellules, les haplotypes obtenus par analyse des marqueurs Y-STRs seront essentiellement composés d'un seul allèle par locus (**Figure 1.5**). Toutefois, des loci dupliqués font exception et peuvent mener à la détection de deux ou trois allèles lors de l'analyse (DYS385 a/b et DYS389 I/II).

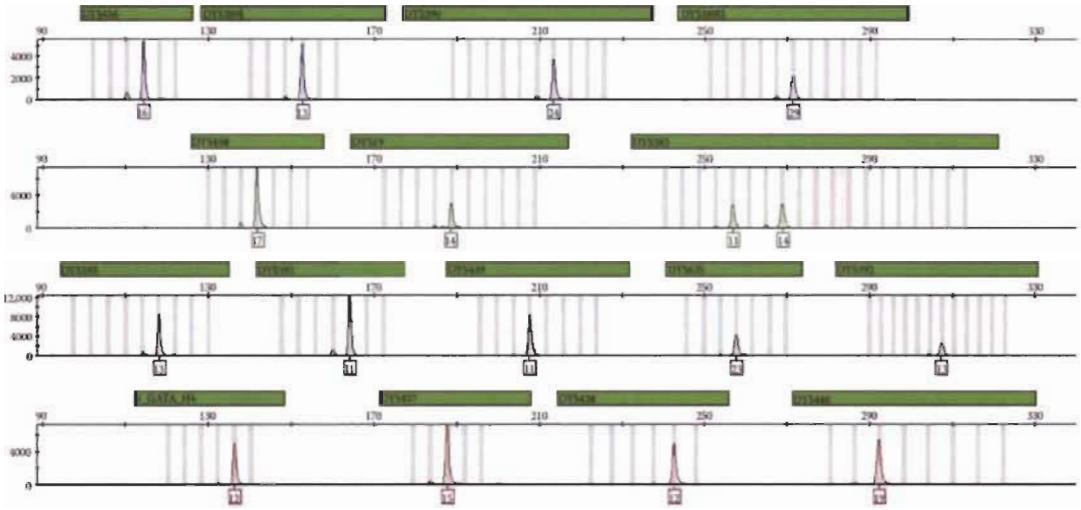

Figure 1.5 Exemple d'haplotype Y obtenu en utilisant les 17 marqueurs Y-STRs inclus dans la trousse AmpFℓSTR® Yfiler™ (Life Technologies) (reproduit de Kayser et Ballantyne, 2014).

Les marqueurs étant situés sur la partie non recombinante du chromosome Y, ils seront transmis en bloc de génération en génération, c'est-à-dire de manière non indépendante. À l'exception des mutations, l'haplotype du fils sera identique à celui de son père, de ses frères, de son grand-père paternel et ainsi de suite. Cela aura un impact considérable sur l'interprétation des profils.

Suite à la découverte, en 1992, des premiers marqueurs STR sur la partie non recombinante du chromosome Y et leur utilisation immédiate dans un premier dossier judiciaire, les efforts se sont multipliés pour découvrir davantage de marqueurs Y-STR de sorte qu'actuellement, jusqu'à 27 marqueurs sont inclus dans les trousses commerciales [39-41]. Suivant les recommandations d'une étude internationale, neuf loci ont initialement servi à établir un haplotype Y en 1997 [42]. Ces neuf loci, référés comme le « minimal haplotype » incluait : DYS19, DYS385a, DYS385b, DYS389I, DYS389II, DYS390, DYS391, DYS392 et DYS393. Puis, en 2003, le Scientific Working Group on DNA Analysis Method (SWGDAM) [43], un regroupement de laboratoires américains, recommandait l'ajout de deux loci supplémentaires [43, 44]. Les trousses commerciales disponibles sur le marché incluent ces 11 loci en plus de combinaisons variées d'autres

Y-STR. Le **Tableau 1.1** présente les marqueurs génétiques inclus dans les trois trousses parmi les plus utilisés aux États-Unis, soit la trousse PowerPlex Y23® de Promega ainsi que les trousses AmpFlSTR® Yfiler™ et Yfiler™ Plus d'Applied Biosystems [45].

Les taux de mutations des marqueurs inclus dans les trousses commerciales sont, en majorité, de l'ordre de 10^{-3} à 10^{-4} ce qui est semblable à des marqueurs autosomaux [33]. Cela signifie que, pour un marqueur donné, qu'une mutation surviendra à tous les 1 000 à 10 000 individus d'une même lignée paternelle. Plus récemment, de nouveaux marqueurs à mutation rapide (RM Y-STRs) ont été caractérisés. Leur taux de mutation est de l'ordre de 10^{-2}. Ceux-ci permettent de distinguer plus souvent deux hommes issus d'une même lignée paternelle, tel que démontré par Ballantyne et al. (2010) dans une étude comparant 103 paires d'individus séparés par une à 20 générations. En utilisant 13 RM Y-STR, ces auteurs ont notamment trouvé que les deux individus d'une paire différaient l'un de l'autre à au moins un marqueur dans 70% des paires père-fils, 56% de celles composées de frères et 67% de celles composées de cousins. En comparaison, en utilisant la trousse Yfiler™ qui ne comprend pas de RM Y-STR parmi ses 17 marqueurs, ces mêmes auteurs n'ont pu distinguer les individus d'aucune des paires père-fils ni aucune de celles composées de deux frères, tandis que seulement 6% des paires de cousins ont pu être différenciées [33].

Tableau 1.1 Marqueurs Y-STR inclus dans trois trousses commerciales fréquemment utilisées.

Loci	PowerPlex Y23®	AmpFlSTR® Yfiler™	Yfiler™ Plus
DYS19*	X	X	X
DYS385a*	X	X	X
DYS385b*	X	X	X
DYS387S1a			X
DYS387S1b			X
DYS389I*	X	X	X
DYS389II*	X	X	X
DYS390*	X	X	X
DYS391*	X	X	X
DYS392*	X	X	X
DYS393*	X	X	X
DYS437	X	X	X
DYS438*	X	X	X
DYS439*	X	X	X
DYS448	X	X	X
DYS449			X
DYS456	X	X	X
DYS458	X	X	X
DYS460			X
DYS481	X		X
DYS518			X
DYS533	X		X
DYS549	X		
DYS570	X		X
DYS576	X		X
DYS627			X
DYS635	X	X	X
DYS643	X		
Y-GATA-H4	X	X	X

*Loci de « bases » dont l'utilisation est recommandée par le SWGDAM entre autres.

1.6 Interprétation des profils

Lorsqu'un profil ADN est obtenu suite à l'analyse d'une trace matérielle, ce dernier sera comparé au profil d'une personne d'intérêt à l'enquête, comme un suspect, ou à ceux contenus dans une base de données. Cette comparaison permet d'établir l'une des conclusions suivantes : l'inclusion du suspect comme parmi les sources possibles de l'ADN, l'exclusion du suspect comme source de l'ADN ou un résultat d'analyse non concluant [43]. Dans l'éventualité qu'une concordance (inclusion) est obtenue entre le profil génétique d'un suspect et celui mis au jour sur une trace, le forensicien voudra attribuer un poids statistique à celle-ci, appelée valeur probante, qui permet de donner une estimation de la rareté de l'observation. Si l'analyse d'ADN est présentée en preuve au tribunal, sa valeur probante sera prise en compte par les décideurs de fait (juge, jurés) ayant à se prononcer sur la culpabilité d'un suspect. Dans le cas des marqueurs autosomaux, entre 15 et 20 loci sont amplifiés de sorte que la probabilité de concordance entre le profil génétique retrouvé sur la trace et celui d'un individu tiré au hasard dans la population est typiquement plus petite que 1 sur 100 milliards [46], si on exclut les cas de jumeaux identiques. Plusieurs concepts sont nécessaires à la bonne compréhension des méthodes d'interprétation d'un profil ADN dans un contexte forensique et sont présentés en profondeur dans les sections suivantes.

1.6.1 *Probabilité de concordance fortuite (P_M ou RMP)*

Le calcul de la probabilité de concordance fortuite se base sur la fréquence des allèles (marqueurs autosomaux) ou des haplotypes (marqueurs de lignées) dans la population pour déterminer la rareté d'un profil ADN. Il permet de répondre à la question suivante : quelle est la probabilité de tirer au hasard un individu dans la population d'intérêt, autre que le suspect, qui aurait le même profil ADN que celui retrouvé sur la trace ? Plus cette probabilité est faible, plus l'observation de la concordance soutient l'hypothèse de la poursuite, à savoir que le suspect, et non un autre individu, est la source de la trace [47]. Tout d'abord, la fréquence des allèles dans la population est estimée par comptage dans un échantillon aléatoire d'individus de la population d'intérêt [26]. On pourrait croire que la fréquence d'un allèle correspond au nombre d'individus portant cet allèle dans la

population, mais cela est faux pour les marqueurs autosomaux. En effet, comme certains individus seront homozygotes et porteront deux copies du même allèle, la fréquence de celui-ci peut différer de la proportion d'individus le portant. Par exemple, selon le **Tableau 1.2**, deux individus sur dix portent l'allèle 12, alors que ce dernier a en réalité une fréquence de 3/20. Ainsi, la fréquence d'un allèle pour les marqueurs autosomaux sera toujours égale ou plus faible que le nombre d'individus le portant [26].

Tableau 1.2 Exemple d'un échantillon fictif de 10 individus servant à estimer la fréquence des allèles pour un locus donné.

En (a), le tableau comptabilise les allèles détectés chez 10 individus pour un locus donné. En (b), le tableau comptabilise la fréquence des allèles pour le locus analysé en (a)

(a)

	Allèles détectés au locus D8S1179	
Individu 1	10	11
Individu 2	12	16
Individu 3	14	18
Individu 4	11	18
Individu 5	10	15
Individu 6	12	12
Individu 7	13	18
Individu 8	16	17
Individu 9	13	14
Individu 10	13	16

(b)

Allèles	Nombre de fois observé	Fréquence
10	2	10% (2/20)
11	2	10% (2/20)
12	3	15% (3/20)
13	3	15% (3/20)
14	2	10% (2/20)
15	1	5% (1/20)
16	3	15% (3/20)
17	1	5% (1/20)
18	3	15% (3/20)
Total	20	100%

Les équations **1.1** et **1.2** permettent de calculer, à partir de la fréquence des allèles, la fréquence d'un génotype dans une population.

$$P_{AA} = p_A^2 \qquad \textbf{(1.1)}$$

$$P_{Aa} = 2p_A p_a \quad \textbf{(1.2)}$$

Où

P_{AA} : fréquence du génotype homozygote AA dans la population

P_{Aa} : fréquence du génotype hétérozygote Aa dans la population

p_A : fréquence de l'allèle A dans la population

p_a : fréquence de l'allèle a dans la population

Ainsi, sous la prémisse que les marqueurs génétiques sont transmis de façon indépendante les uns des autres et que l'appariement des gamètes est aléatoire dans la population, il est possible d'obtenir la probabilité de concordance fortuite en multipliant la fréquence du génotype de chacun des marqueurs génétiques ayant été utilisés pour obtenir le dit profil, selon l'équation **1.3**.

$$P_M = RMP = P_G = \prod_{l=1}^{N_L} \phi \, P_{l,1} P_{l,2} \quad \textbf{(1.3)}$$

où

P_M : probabilité de tirer au hasard un individu dans la population portant le génotype G déjà observé sur la trace

P_G : probabilité de tirer au hasard un individu dans la population portant le génotype G (ou RMP, de l'anglais *random match probability*)

N_L : nombre de locus formant le génotype G

ϕ : indicateur d'homozygotie. Valeur de 1 pour un homozygote, valeur de 2 pour un hétérozygote

$P_{l,1}$: fréquence du premier allèle au locus l

$P_{l,2}$: fréquence du deuxième allèle au locus l

Il faut savoir que les équations **1.1 à 1.3** font la prémisse que la population suit la loi de Hardy-Weinberg. Cette dernière stipule que les fréquences d'allèles restent inchangées d'une génération à une autre car la transmission de celles-ci ne dépend que du hasard de la fécondation et de la méiose. Afin qu'une population satisfasse strictement la loi de Hardy-Weinberg, elle doit avoir une taille infinie, il ne doit pas survenir de mutations aux marqueurs typés, et l'appariement reproducteur doit se faire aléatoirement entre les gamètes (et les individus). De plus, aucune pression de sélection naturelle ni

mouvement de migration ne doit avoir lieu dans la population. Toutefois, il est déraisonnable de penser qu'une population humaine puisse respecter l'ensemble de ces critères. Celui posant sans doute le plus de problème étant l'appariement aléatoire des gamètes, qui contredit la subdivision souvent observée des populations humaines en sous-populations génétiquement différentiées [48, 49]. L'indice de fixation (θ ou F_{ST}) permet de mesurer la structure génétique dans les populations. Initialement introduit en 1951 par Wright, cette variable permet de quantifier la variance dans les fréquences de génotypes dans la population globale qui est expliquée par sa structure en sous-populations [50]. Les valeurs possibles de θ se situent entre 0 et 1. Une valeur élevée signifie une plus forte structure génétique et une valeur faible signifie plutôt une population relativement homogène, tendant vers l'appariement aléatoire. L'indice de fixation est aussi un indicateur du niveau de parenté des individus dans une sous-population par rapport à la population totale [1]. Les équations **1.1** et **1.2** peuvent être modifiées pour estimer la fréquence d'un génotype en intégrant le paramètre θ [19, 47] :

$$P_{AA} = P_A^2 + P_A P_a \theta \quad \textbf{(1.4)}$$

$$P_{Aa} = 2P_A P_a \, (1 - \theta) \quad \textbf{(1.5)}$$

Toutefois dans ce cas, contrairement au précédent (équations **1.1** et **1.2**), la probabilité d'un génotype (P_G) n'équivaut pas à la probabilité de concordance fortuite (P_M ou RMP) qui nous intéresse en science forensique, tel que démontré par Balding et Nichols (1994). Ce auteurs ont établi les équations suivantes afin de calculer la P_M dans une population ne suivant pas la loi de Hardy-Weinberg en introduisant le paramètre θ [51] :

$$P(A_l A_l | A_l A_l) = \frac{(2\theta+(1-\theta)p_l)(3\theta+(1-\theta)p_l)}{(1+\theta)(1+2\theta)} \quad \textbf{(1.6)}$$

$$P(A_l A_i | A_l A_i) = \frac{2(\theta+(1-\theta)P_l)(\theta+(1-\theta)P_i}{(1+\theta)(1+2\theta)} \quad \textbf{(1.7)}$$

où

$P(A_lA_l|A_lA_l)$: probabilité de tirer au hasard dans la population le génotype homozygote A_lA_l sachant qu'il a déjà été observé sur la trace (sous l'hypothèse qu'il s'agit de sources différentes)

$P(A_lA_i|A_lA_i)$: probabilité de tirer au hasard dans la population le génotype hétérozygote A_lA_i sachant qu'il a déjà été observé sur la trace (sous l'hypothèse qu'il s'agit de sources différentes)

1.6.2 *Rapport de vraisemblance (RV)*

Le rapport de vraisemblance (RV) est le ratio de la probabilité des observations génétiques (E, p. ex. une concordance) sous une hypothèse (numérateur) par rapport à une autre (dénominateur). Les deux hypothèses typiquement évaluées sont celles de la poursuite (Hp, p. ex. le suspect est la source de la trace) et de la défense (Hd, p. ex. un autre individu que le suspect est la source de la trace). Le rapport de vraisemblance se calcule de la façon suivante [46]:

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

où

I : autres éléments circonstanciels (non génétiques) associés au cas

Celui-ci est une mesure relative et conditionnelle qui permet, selon l'ampleur de la valeur obtenue, de soutenir une des deux hypothèses plus que l'autre. Dans le cas le plus simple (et le plus courant), l'équation **1.8** peut être réduite à :

$$RV = \frac{1}{RMP} \qquad \textbf{(1.9)}$$

Dans cette équation simplifiée, $Pr(E|H_p,I)$ prend la valeur 1, puisqu'on peut penser que si le suspect est bel et bien à l'origine de la trace, alors les deux profils ADN (le sien et celui de la trace) concorderont assurément. De même, $Pr(E|H_d,I)$ prend la valeur de RMP. Ce calcul est souvent celui privilégié par les laboratoires judiciaires dans les rapports d'analyses transmis aux policiers et à la cour puisqu'il permet d'évaluer deux hypothèses en concurrence, ce qui convient mieux au contexte judiciaire où deux parties défendent des hypothèses différentes (Hp et Hd).

Dans les dossiers judiciaires, il n'est pas rare qu'un mélange complexe d'ADN de deux ou plusieurs personnes soit obtenu par l'analyse de la trace ADN. Il peut alors rapidement devenir fastidieux pour l'expert judiciaire d'effectuer l'interprétation de ces profils ADN et de calculer un rapport de vraisemblance de façon manuelle. Effectivement, une analyse manuelle ne permet pas à l'expert judiciaire de prendre en compte l'ensemble des informations disponibles dans le profil ADN, tel que la hauteur des pics de l'électrophorégramme, les probabilités de drop-outs aux différents loci, qui sont des éléments importants à prendre en compte dans l'évaluation d'une valeur probante. De plus, l'évolution des techniques d'analyse ADN dans les récentes années, combinée à l'augmentation de la sensibilité des kits commerciaux, a mené à la diminution de la quantité d'ADN nécessaire pour l'obtention d'un profil ADN interprétable, un autre élément qui complexifie l'interprétation des profils ADN [52]. Récemment, des outils informatiques ont été développés permettant la modélisation de valeurs probantes pour des mélanges complexes, ce que l'on nomme le génotypage probabilistique. Ces logiciels (ex : STRmix™, TrueAllele®) intègrent des algorithmes prenant en compte des modèles statistiques, des concepts de biologie et des distributions de probabilité afin de calculer des valeurs de rapports de vraisemblance de mélanges complexes [52-54]. Ces logiciels permettent de calculer un rapport de vraisemblance pour des combinaisons de profils ADN comptant jusqu'à cinq, voire six contributeurs.

1.6.3 *Interprétation des profils Y*

Étant donné que les marqueurs STR utilisés pour obtenir un profil Y sont situés sur la partie non recombinante de ce chromosome, la méthode pour calculer une valeur probante

diffère légèrement de celle utilisée pour les marqueurs autosomaux. Effectivement, puisqu'il n'y a pas de recombinaison, les loci analysés seront transmis en bloc de père en fils et donc, la fréquence des génotypes aux différents marqueurs ne peut pas être considérée comme indépendante ; il faut plutôt estimer la fréquence de l'haplotype entier. Pour ce faire, l'haplotype est comparé à une base de données afin de déterminer le nombre de fois où celui-ci est observé dans la base. Dans sa forme la plus rudimentaire, la fréquence d'un haplotype Y est calculée en divisant le nombre de fois que ce dernier est observé dans une base de données de référence par le nombre total d'échantillons qu'elle contient [55].

$$P = \frac{x}{N} \qquad \textbf{(1.10)}$$

où

x : nombre de fois que l'haplotype a été observé dans la base de données

N : nombre de profils inclus dans la base de données

Des variantes de l'équation **1.10** sont utilisées pour calculer la fréquence des haplotypes rares qui ne seraient pas présents dans la base de données en question. La première consiste à ajouter une unité à la valeur de x et de N de sorte que l'haplotype Y retrouvé sur la trace soit considéré comme un échantillon additionnel de la base de données :

$$P = \frac{x+1}{N+1} \qquad \textbf{(1.11)}$$

Les équations **1.10** et **1.11** dépendent entièrement de la disponibilité et de la taille des bases de données. Par conséquent, l'ensemble des haplotypes Y qui ne sont pas présents dans la base de données en question se verront attribuer la même fréquence, valeur qui sera biaisée dans plusieurs cas. Par exemple, supposons qu'un haplotype Y est obtenu suite à l'analyse ADN d'une trace et que ce dernier n'est pas observé dans la base de données de référence d'une taille de 500 échantillons. Alors, en utilisant l'équation **1.11**,

la fréquence de cet haplotype Y sera de 0,002 (1/501). Si en réalité, la fréquence de cet haplotype Y dans la population est de 0,0005 (1/2000), alors la fréquence attribuée sera 40x plus élevée que la fréquence réelle. De plus, l'ensemble des haplotypes Y qui ne sont pas comptabilisés dans la base de données de référence se verront attribuer la même fréquence. Ces éléments démontrent clairement que l'utilisation des équations **1.10 et 1.11** tant à biaiser la fréquence réelle des haplotypes Y.

Lorsque les équations **1.10 et 1.11** sont utilisées, il est important d'y ajouter un intervalle de confiance (généralement à 95%) afin de prendre en compte l'incertitude dans l'évaluation de la valeur probante selon la taille de la base de données utilisée [30]. La limite supérieure de cet intervalle sera calculée de la façon suivante et directement fournie comme valeur probante conservatrice :

$$IC = P + 1.96\sqrt{\frac{(P)(1-P)}{N}} \qquad \textbf{(1.12)}$$

Il est également possible de calculer un rapport de vraisemblance (RV) tel que présenté dans la section 1.6.2 (**équation 1.8**) lorsque les marqueurs situés sur le chromosome Y sont utilisés. Ainsi, le résultat soutiendra une hypothèse plus qu'une autre, par exemple qu'il est plus probable d'observer la concordance si le suspect est la source de la trace (hypothèse de la poursuite) que si un individu inconnu est la source de la trace (hypothèse de la défense). Dans l'éventualité où l'haplotype Y retrouvé sur une trace ne serait jamais observé dans la base de données de référence, il serait alors possible d'utiliser le modèle Kappa proposé par Brenner en 2010 [56] :

$$RV = \frac{N}{1-\kappa} \qquad \textbf{1.13}$$

où

κ : Nombre de singletons (haplotypes retrouvés qu'une seule fois) dans la base de données

1.6.4 *Base de données d'haplotypes Y*

Les équations présentées dans la section précédente ont été critiquées du fait qu'elles dépendent toutes d'un seul et même élément : la disponibilité d'une base de données comprenant des haplotypes Y. L'une des plus grandes bases de données est une banque internationale en ligne, la Y-STR Haplotype Reference Database (YHRD) qui, en 2020, comptait 307 169 haplotypes amplifiés avec 8 Y-STR, 246 821 haplotypes amplifiés avec la trousse Yfiler (17 Y-STR), 73 006 avec la trousse PowerPlex Y23 (23 Y-STR) et 73 810 haplotypes amplifiés avec la trousse Yfiler Plus (27 Y-STR). Malgré une collaboration internationale permettant une croissance rapide du nombre d'haplotype versés dans la banque de données, il est important de noter que plusieurs pays n'y sont pas représentés, dont le Canada [57].

En décembre 2007, les États-Unis ont développé une base de données propre à leur pays, la US Y-STR database, permettant d'estimer la fréquence des haplotypes dans cinq groupes ethniques de leur population (Caucasien, Asiatique, Afro-Américain, Amérindien et Hispanique). En date de 2014, cette base de données comprenait 32 972 haplotypes Y amplifiés avec les 11 Y-STR recommandés par SWGDAM, 23 169 haplotypes amplifiés avec la trousse Yfiler (17 Y-STR) et 4 837 haplotypes avec la trousse PowerPlex Y23 (23 Y-STR) [19]. Toutefois, cette base de données n'existe plus depuis peu, les haplotypes contenus dans celle-ci ayant été versés dans la banque Y-HRD.

Malgré tout, l'utilisation de ces bases de données pour déterminer la fréquence des haplotypes du chromosome Y est critiquée dans la littérature, notamment par Andersen et Balding (2017). D'abord, on soulève qu'elles ne sont pas constituées d'échantillons aléatoires et qu'elles sont de trop petites tailles pour bien estimer la fréquence de la grande majorité des haplotypes Y très rares. Effectivement, pour la plupart des populations, la taille des bases de données disponible dépasse rarement 1 000 haplotypes, ce qui est trop peu pour représenter l'ensemble de la population [58]. À titre d'exemple, Doyon (2018) ont reliés les données moléculaires 450 hommes d'origine canadienne-française à seulement 342 lignées paternelles alors que plusieurs milliers de lignées (>8 000) existent

[59]. Ainsi, l'utilisation des équations **1.10** et **1.11** aura pour effet d'estimer la fréquence de certains haplotypes inadéquatement, principalement ceux qui sont rares dans la population (soit la plupart) [60]. De plus, malgré l'existence de ces équations, le SWGDAM qui émet différentes recommandations quant à l'analyse et l'interprétation des profils ADN aux laboratoires judiciaires, n'a toujours pas atteint de consensus concernant la méthode appropriée à utiliser pour calculer la fréquence d'un haplotype Y [43]. L'ensemble des éléments présentés démontre la pertinence de développer de nouveaux outils afin de permettre l'analyse et l'interprétation des haplotypes Y de façon plus adéquate.

1.7 La généalogie du Québec

Afin de bien situer la recherche effectuée dans le cadre de ce projet de maîtrise, il est important de connaître et comprendre la structure de la généalogie de la population canadienne-française du Québec. La colonisation du territoire par les Français a eu lieu entre 1608 et 1759. On estime que durant cette période, un peu moins que 10 000 colons se sont installés le long du fleuve Saint-Laurent, et que seulement 8 384 d'entre eux auraient contribué génétiquement à la population que l'on connait aujourd'hui [61]. En 1972, l'Université du Québec à Chicoutimi a lancé un projet afin de reconstituer la démographie historique de la population du Saguenay. Pour ce faire, près de 660 000 actes de baptême, de mariage et de sépulture qui avaient été conservés par l'église catholique ont été utilisés pour reconstruire la population du Saguenay, couvrant la période s'étendant entre 1838 et 1971. Sous le nom de fichier BALSAC, le projet s'est élargi au cours des années de sorte que le fichier contient maintenant de l'information pour l'ensemble des régions du Québec (en ce qui concerne les mariages du moins). À partir de 3 millions d'actes, il a été possible d'inclure dans BALSAC plus de 5 millions d'individu couvrant une période de trois siècles et demi [62].

Le Québec ayant grâce à BALSAC, une fine connaissance de sa généalogie depuis la fondation de la Nouvelle France au XVII^{e} siècle, constitue une population d'étude exceptionnelle pour les recherches en génétique et plusieurs autres domaines. Déjà, de

nombreuses recherches ont été effectuées au Québec à partir du fichier BALSAC, tant en génétique, qu'en démographie, en épidémiologie et en histoire. Le but premier du fichier était d'étudier la transmission des maladies héréditaires au Québec, puisque quelques-unes d'entre elles sont beaucoup plus fréquentes dans certaines régions que d'autres. À titre d'exemple, des recherches ont été effectuées sur les maladies récessives telle que l'ataxie spastique, affectant les individus de Charlevoix et du Saguenay (De Braekeleer et al., 1993), les anévrismes intracrâniens ou la dystrophie myotonique dans la population du Saguenay–Lac-St-Jean (De Braekeleer et al., 1996 ; Dao et al., 1993) pour ne nommer que celles-là [63-65]. Par la suite, plusieurs autres études se sont penchées sur l'histoire du Québec pour comprendre sa structure généalogique et génétique. En 2001, Scriver a publié un article décrivant en détails le peuplement du Québec, les vagues de migrations, la démographie et l'impact de ses facteurs sur la génétique et la généalogies actuelles du Québec [66]. Plus récemment, Doyon et al. (soumis) ont développé un modèle généalogico-moléculaire afin d'étudier la variation spatio-temporelle des haplotypes Y et mitochondriaux dans la population du Québec et l'impact que cette dernière peut avoir sur le calcul de la valeur probante dans un contexte forensique [67]. Leurs résultats ont mis en évidence que la fréquence des haplotypes est stable dans le temps, mais qu'une variation existe au niveau régional et même, plus marquée à l'échelle des localités.

1.8 Objectifs

Au Québec, le nombre de dossiers d'agressions sexuelles soumis pour analyse au Laboratoire de sciences judiciaires et de médecine légale a fait un bon de plus de 29% entre 2014-2015 et 2018-2019, dans la foulée du mouvement #metoo. Ainsi, les dossiers d'agressions sexuelles occupent une proportion importante des demandes d'expertises en biologie/ADN au Québec [2]. Considérant, d'une part l'utilité du chromosome Y dans ce type de dossier, et d'autre part, les problèmes entourant les bases de données dans l'interprétation des haplotypes Y, il importe de développer de nouvelles approches pour mieux évaluer le poids des concordances ADN. À cet égard, Andersen et Balding (2017) ont développé une nouvelle méthode très différente des précédentes afin d'attribuer une valeur probante à une concordance entre deux haplotypes Y, reposant sur un modèle de

simulation. Elle permet de contrecarrer les problèmes associés aux bases de données inadéquates ou manquantes. Le modèle sert à simuler des populations fictives selon plusieurs paramètres démographiques et génétiques, puis d'attribuer des haplotypes aux individus virtuels pour ainsi estimer la distribution attendue des hommes portant un même haplotype dans la population. Andersen et Balding pensent que cette façon de rapporter un résultat sera mieux comprise par les acteurs du système de justice (magistrats, avocats, policiers).

Bien que prometteuse, la méthode proposée par Andersen et Balding nécessite une validation empirique. La complétude et la qualité des données généalogiques canadiennes-françaises du Québec offrent l'opportunité parfaite et unique pour tester le modèle à l'échelle d'une population entière. C'était le premier objectif du présent projet. Pour ce faire, j'ai procédé de deux façons. D'abord, j'ai utilisé les paramètres démographiques mesurés sur la population d'origine canadienne-française pour simuler une population totalement virtuelle se rapprochant le plus possible de celle du Québec actuel, en utilisant la méthode originale d'Andersen et Balding. Par la suite, j'ai modifié cette méthode pour simuler la transmission des haplotypes Y en utilisant la vraie généalogie du Québec. J'ai ensuite confronté les résultats de ces deux analyses pour évaluer la performance de la méthode et y apporter des modifications.

Le deuxième objectif du projet consistait à évaluer la variation spatiale dans la répartition des haplotypes Y dans le Québec. En effet, il est plutôt bien rapporté dans la littérature que la population du Québec possède une structure particulière en raison des mouvements migratoires à l'origine de sa fondation. Il est connu que plusieurs petits groupes d'individus issus des fondateurs de la Nouvelle France se sont déplacés pour coloniser, par vagues, les régions du Québec. Les individus établis dans une région se mariaient plus rarement avec des gens d'une autre région, créant une structure génétique entre celles-ci [68]. Ce type de structure complexe, qui n'est que très rarement, voire jamais considérées par les laboratoires judiciaires, pourraient mener à des calculs de valeur probante peu fiables en raison de l'utilisation de base de données inadéquates pour estimer les fréquences d'haplotypes Y. En effet, l'étude menée récemment par Doyon et al. (soumis) démontrait que les valeurs de RMP pouvaient différer d'un ordre de grandeur

allant jusqu'à 10^7 selon si la région ou la localité était considérée [67]. Toutefois, le modèle utilisé ne permettait pas de considérer l'ensemble des lignées paternelles ayant existé, par manque de données moléculaires s'y rapprochant. Le modèle d'Andersen et Balding permet quant à lui d'estimer par région la distribution des hommes portant un même haplotype, en considérant l'ensemble des lignées paternelles disponibles. En associant les résultats de mon projet de maîtrise à ceux de Doyon et al. (soumis), je termine ce mémoire en proposant une nouvelle façon d'interpréter les concordances de profils génétiques Y intégrant le modèle modifié d'Andersen et Balding et des données moléculaires.

CHAPITRE II

TEST EMPIRIQUE À GRANDE ÉCHELLE D'UN MODÈLE DE SIMULATION DE POPULATION POUR QUANTIFIER LE POIDS D'UNE CONCORDANCE ADN POUR DES PROFILS GÉNÉTIQUES DU CHROMOSOME Y

Roxane Landry[1], Mikkel Meyer Andersen[2], Emmanuel Milot[1]

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

[2] Département des sciences mathématiques, Université d'Aalborg, Danemark

Le contenu de ce chapitre est en préparation en vue d'une publication dans une revue scientifique avec comité de révision par les pairs.

2.1 Contribution des auteurs

En tant qu'auteure principale, j'ai contribué à l'élaboration du projet, accompli l'ensemble des analyses présentées dans cet article en plus de rédiger ce dernier. Pr Emmanuel Milot, en tant que directeur de recherche, a élaboré le projet de recherche et m'a apporté son soutien tout au long du projet en plus de réviser le manuscrit de l'article. Pr Mikkel Meyer Andersen, le développeur de la méthode, a apporté un soutien technique important au projet et implémenté dans son module R plusieurs modifications identifiées comme souhaitables en cours de projet.

2.2 Résumé de l'article

Dans un contexte judiciaire, lorsqu'un profil ADN retrouvé sur une pièce à conviction concorde avec celui d'un suspect, une valeur probante est attribuée à cette concordance pour éclairer la cour de justice. L'attribution de la valeur probante peut s'avérer difficile, particulièrement dans les dossiers d'agression sexuelle où des combinaisons d'ADN homme-femme sont régulièrement retrouvées. Les pièces à conviction récoltées dans ce genre de dossiers incluent souvent des écouvillons intimes provenant de la victime, qui comportent un mélange de l'ADN de cette dernière et de l'agresseur. Il n'est pas rare que l'ADN de la victime, en trop grande concentration, limite ou empêche la détection de l'ADN de l'agresseur. Les marqueurs situés sur le chromosome Y, qui n'est présent que chez les hommes, peuvent compenser les limites de ceux situés sur les chromosomes autosomaux, en permettant d'isoler le profil masculin. Or, le chromosome Y, qui se transmet de père en fils, n'est pas soumis à la recombinaison méiotique. En conséquence, la composition génétique d'une population sur le plan du chromosome Y est très influencée par les effets fondateurs. Ce phénomène, ainsi que le manque de bases de données de référence adéquates rendent difficile l'attribution d'une valeur probante à ces concordances de profils Y, de sorte qu'il n'existe actuellement pas de consensus quant à la méthode appropriée à utiliser. Pour contrer ces problèmes, Andersen et Balding ont publié, en 2017, une méthode consistant en un modèle de simulation de population permettant d'estimer la distribution du nombre d'hommes portant un même profil ADN. L'objectif de notre étude était de tester à grande échelle ce modèle en utilisant la généalogie de la population canadienne-française. Pour ce faire, nous avons réalisé des simulations en utilisant les paramètres de la population canadienne-française pour simuler une population virtuelle dont la généalogie se rapproche de celle de la vraie. Par la suite, nous avons réalisé des simulations d'attribution d'haplotypes en utilisant la vraie généalogie, puis confronté les résultats des deux approches. Lorsque la vraie généalogie est utilisée, on observe qu'une minorité d'haplotypes est portée par une grande proportion d'individus, ce qui n'est pas observé lorsque la population est simulée. De plus, nous observons que certains haplotypes ne se retrouvent que dans certaines régions seulement, et non pas dans tout le Québec.

2.3 Article complet en anglais

Large-scale empirical test of a population simulation model to quantify the weight of DNA match evidence for Y chromosome markers

Abstract

Markers located on autosomal chromosomes are routinely analyzed to obtain DNA profiles from traces, but they may be of limited help in cases of male-female DNA mixtures. In sexual assault cases, such mixtures are often recovered from intimate swabs taken on a female victim. The DNA of the victim is typically overwhelming on these swabs so that it can be hard to detect the DNA of a male contributor(s). Y chromosome markers can help to resolve these mixtures by providing an interpretable DNA profile of a male contributor. Nevertheless, the attribution of a probative value to a Y DNA match remains problematic, due to the mode of inheritance of the Y chromosome, the variation in population history, and issues in using reference DNA databases. To counter these problems, Andersen and Balding (2017) published a method that allows the attribution of a probative value for Y haplotype matches without the need to use a reference database, by using simulations to model the distribution of men sharing haplotypes in a population. Here, we report on a large-scale empirical test of this method on the French-Canadian population of Québec, Canada. Using the original Andersen and Balding's method, we estimated that 99% of Y haplotypes were shared by 152 men or less. We then modified the model to make it accurately tailored to the Québec population by integrating population-wide genealogical data and estimated that 99% of Y haplotypes were shared by 295 men or less. Moreover, as a result of founder effects in the French-Canadian population, 1% of the haplotypes represented up to 58% of the male population, depending on the scenario tested. In contrast, that proportion was at most 7% with the original model without genealogical data. We also found a large heterogeneity in the geographic distribution of Y-STR haplotypes. In the four regions analyzed, between 500 and 1,500 haplotypes occurred in only one region, out of 4,000 to 6,000 total haplotypes. Some

modifications to the method should help to extend the range of population histories it may deal with. We also propose a strategy mixing Andersen and Balding's model and molecular data to assign a weight to Y chromosome evidence in forensic casework.

Keywords

Forensic genetics, Genealogy, French-Canadian population, Paternal lineage, Weight of evidence, Short tandem repeats, Y haplotype

Introduction

In forensic genetic, short tandem repeats (STRs) located on autosomal chromosomes are used to obtain a DNA profile from crime scene traces and are quite powerful for individualization purposes (Purps *et al.*, 2015; Andersen and Balding, 2017). However, autosomal STRs have limits to resolve DNA mixtures coming from two or more contributors to DNA traces. While the development of probabilistic genotyping software can help much with the interpretation of mixtures, it may not always be sufficient to identify the profile of when a minor male genetic profile is detected. In sexual assault cases, evidence often includes intimate swabs taken on the victim, from which a mixture of female and male DNA may be recovered (Gusmao *et al.*, 2006; Coquoz *et al.*, 2013). Since the autosomal DNA of the victim is typically in much higher proportion than that of the male offender, it may be hard to isolate his genetic profile. The STRs located on the Y chromosome can then be used to obtain a DNA profile of the offender. However, when a match occurs between a Y haplotype of a suspect and the one recovered from a trace, it is hard to assess a probative value to that match (Gusmao *et al.*, 2006; Butler, 2014).

The Y chromosome is transmitted from a father to his son without barely any change, as only the extremities (~5%) recombine with the X chromosome. Therefore, the Y chromosome is at 95% inherited without any change, leaving mutations as the only source of possible variations (Cockerton *et al.*, 2012). Every man of a paternal lineage will have the same Y chromosome if no mutation occurred along that lineage. As a consequence, the loci analyzed on the Y chromosome are linked to one another so that the frequency of a haplotype cannot be determined using the Hardy-Weinberg law and is rather based on the whole haplotype (Kayser, 2017). To do so, a laboratory needs a Y chromosome reference database relevant to the population of interest (Gusmao *et al.*, 2006) and the frequency is calculated using the counting method:

$$p = \frac{x}{N} \quad \textbf{(1)}$$

where x is the number of times the haplotype is observed in a database of size N. A number of variants of equation (**1**) have been proposed in the literature, the most frequent one being the augmented counting method: $p=(x+1)/(N+1)$ (Egeland and Salas, 2008). A confidence interval is generally applied to account for the size of the database and sampling variation (SWGDAM, 2014). Other methods have been proposed, but the counting method and its variants are the most applied (Roewer *et al.*, 2000; Roewer, 2009; Andersen *et al.*, 2013).

Multiple problems are linked to the use of reference databases. First, most of the databases available (ex: YHRD, Promega, Applied Biosystems) include haplotypes for large regions (e.g., an entire country) and using them as is assumes that no population genetic structure exists (so that haplotypes are geographically randomly distributed), which is rarely the case (Ploski *et al.*, 2002; Kayser *et al.*, 2003; Gusmao *et al.*, 2006). In fact, the Y chromosome, due to its way of inheritance, is more sensitive to genetic drift and founder effects that may result in the spatial heterogeneity of haplotypes in a population (Templeton, 2006; Cockerton *et al.*, 2012). The Scientific Working Group on DNA Analysis Method (SWGDAM) and the DNA Commission of the International Society of Forensic Genetics (ISFG), two groups that issue guidelines on the analysis and the interpretation of DNA for forensic laboratories, suggest to estimate metrics of population structuring (e.g., θ) for several ethnic groups, and integrate them in the counting method (SWGDAM Y-STR Subcommittee, 2007). Second, databases are not necessarily representative of the population of interest, i.e. the group of individuals who may be the source of the DNA recovered on a trace. Most databases are not built randomly, containing DNA profiles from laboratory workers, blood banks samples, as well as police officers, who do not represent randomly selected people in the population (National Research Council (U.S.), 1996; Andersen and Balding, 2017). Consequently, the counting method tends to underestimate the frequency of haplotypes, mostly the rare ones (Egeland and Salas, 2008). Third, men from the same family are more likely to be geographically grouped, creating a spatial heterogeneity of the frequency of Y haplotypes in a population (Roewer, 2009). Because of all those elements, the SWGDAM still has not reached a consensus on the best method to use to determine a probative value for Y haplotype

matches, while the ISFG proposes to use the counting method but warns scientists against the various issues related to it (Gill *et al.*, 2001).

Recently, a new method was proposed by Andersen and Balding (2017) in a form of a population simulation model that allows to determinate the frequency of Y haplotypes without the need of reference databases (Andersen and Balding, 2017). This method uses different parameters to simulate a male population and attribute Y haplotypes to individuals. Instead of calculating haplotype frequencies, this method estimates the distribution of men sharing a same haplotype in the population, which could facilitate the comprehension of probative values by judges and juries (Andersen and Balding, 2017; Andersen and Balding, 2019). However, this promising method requires empirical testing before making it a tool that can be used in forensic cases.

The main objective of our study was to conduct a large-scale empirical test of Andersen and Balding's method using the genealogical data of the French-Canadian population (Québec, Canada). To do so, we compared the distribution of men sharing the same Y haplotype in the population under different simulation scenarios. We simulate a virtual population using Andersen and Balding's method and French-Canadian population parameters. We then modified the model to integrate real genealogical data on millions of men to reconstruct Y haplotype distributions on much more information. Our results show that the method is sensitive to founder effects and spatial heterogeneity, and that some modifications to it should help to extend the range of population histories it may deal with. We also propose a strategy mixing Andersen and Balding's model and molecular data to assign a weight to Y chromosome evidence in forensic casework.

Methods

Genealogical dataset

The genealogical data comes from the BALSAC population register (BALSAC - Fichier des populations), which includes Catholic individuals who are mostly French-Canadian.

French-Canadians represent about 80% of the current Québec population (Gagnon and Heyer, 2001). Built mostly with marriage and death certificates, the register includes data on 4,364,381 individuals (men and women taken together) who lived between 1608 (the foundation of Québec) and 1960. The data were filtered to keep men only. The information included in the register is, for every individual, an identity number, an identity number for his/her parents, the gender, birth date, the location and the date of the marriage and those of the individual's parents. To be able to do simulations with these genealogical data, we assigned an identity number for every paternal lineage and counted the number of men included in each. We also determined the generation number of the individual in his paternal lineage (the founder was assigned the highest generation and the individuals in the last one were assigned generation 0) as well as the total number of generations included in the paternal lineage. Then, paternal lineages of minimally 10 generations were selected for this study as they go back long enough in time to be considered as founder lineages. They totaled up 1,239,775 men spread across 1,306 lineages. **Table 2.1** shows the characteristics of the selected lineages.

Table 2.1 Characteristics of paternal lineages included in the simulations with genealogical data.

Number of individuals per lineage	Number of lineages	Total number of individuals	Number of individuals in the last three generations (considered as the "living" population)
10-100	80	5,427	2,473
101-1,000	877	406,390	193,619
1,001-10,000	345	760,681	298,716
10,001-100,000	4	67,277	29,027
Total	1,306	1,239,775	523,835

Simulation model

Andersen and Balding's model is implemented in the *malan* package for the R software (R Development Core Team, 2019). The functions in *malan* allow simulating a population of men with their Y haplotypes. The user specifies the following parameters as input: the number of individuals to be kept in the last generation, the number of generations to simulate, the variance in male reproductive success, Y-STRs loci used to create Y haplotypes and the mutation rates of every marker used (Andersen and Balding, 2017). To create haplotypes, a random allele for every marker is initially drawn from the pool of possible alleles at the loci used, here the 23 markers from the Promega PowerPlex® Y23 kit. The single founder haplotype is then inherited by his sons, grandsons, and so on. New haplotypes appear by mutation and are likewise inherited by the descendants of the mutant individual. A minimal and maximal boundary, corresponding to the smaller and larger possible alleles, is set for every locus to ensure that mutations do not result in non-existing haplotypes. For simulations integrating the real genealogy, we modified the model to add the genealogical error rate. This is an important parameter as genealogical errors can occur when pairing acts. Also, that parameter takes into account mismatches between the haplotype of a father and that of his putative son caused by extra-pair paternity (Larmuseau *et al.*, 2013). Once haplotypes have been attributed to individuals in the population, the distribution of men sharing the same haplotype is calculated for the "living" population, defined as the last three generations.

Simulations on the French-Canadian population

First, simulations using the demographic parameters of the French-Canadian population were conducted with the method as initially proposed by the authors. To assess the extent to which the distribution of men sharing haplotypes approached the real population, we then realized simulations incorporating the genealogy of the French-Canadians. We also confronted our results to those of a previous study done by Doyon *et al.* (submitted) with the French-Canadian genealogical data. That study used a genealogico-molecular model (genealogy data combined to molecular information, i.e. Y haplotypes of living

individuals) rather than a simulation model (Doyon *et al.*, submitted). This model will be presented further in the discussion.

Simulations without Genealogical Data

For simulations without genealogical data, the size of the population, corresponding to the number of men in the last generation, was set at 1,189,672, as what is observed in the actual population (Québec 2019). Considering a generation time of 32 years, based on a study of the French-Canadian (Tremblay and Vezina, 2000), the number of generations was set at 15 as the foundation of Québec goes back to 1608. The loci in the Promega PowerPlex® Y23 kit were chosen to create fictive Y haplotypes, then the respective mutation rate for each locus was used to implement haplotypes to following generations. The male variance in reproductive success (VRS) was set at 0.2, corresponding to the estimate for the modern American population (**Table 2.2**) (Weeden *et al.*, 2006). When so, with a 95% probability, the paternity will varied between 0.32 and 2.05 among men. A total of five simulations were done with those parameters. As the simulated population was not quite similar to what we were expecting in terms of lineage numbers, five other simulations were made with a different variance in reproductive success. As the male reproductive success is mostly (but not entirely) defined by the number of offspring in a lifetime, using a higher variance in reproductive success will result in a limited number of paternal lineages as some men would have a lot of sons and some others not, which better reflects the real genealogy of the French-Canadian population. Thus, for those other simulations, the value was set at 0.98.

Simulation with Genealogical Data

As mentioned above, some modifications to the model were done to incorporate genealogical data. A function was added to the *malan* package to load those genealogical data. We also added a parameter for genealogical error rate (e), which correspond to the proportion of genealogical links that are erroneous, hence should not appear in the data. During haplotype attribution, a proportion e of links was randomly removed. Where a link

was removed, the putative son was assigned a different haplotype from his father, but one that already exists in the population, i.e. one that had been assigned to one of the founders. We used the estimate $e = 0.008$ reported by Doyon and *al.* (submitted) for the same population. These authors calculated *e* using Larmuseau and *al.'s* (2013) method based on the number of differences observed between two haplotypes from individuals in the same paternal lineage. To do so, they selected pairs of men separated by a minimum of seven meiosis and calculated the mismatch rate by the number of differences observed between the Y haplotypes of a pair. When two haplotypes from a pair had more than 18% of differences, it was considered to be caused by a genealogical error instead of mutations (Doyon *et al.*, submitted).

Simulations were done by keeping paternal lineages of at least 10 generations, as those lineages go back far enough in time to be considered as founding paternal lineages. Unique haplotypes were attributed randomly to founders as described above and then assigned to their descendants, accounting for mutations as before and this time also for genealogical errors. A total of 100 simulations were realized and then the distribution of men sharing haplotypes in the living population was obtained by selecting the last three generations of every lineage.

To better understand the impact of founder effect in French-Canadian population, additional simulations were realized by attributing haplotypes to founders in two other different ways: a starting pool of 100 distinctive haplotypes and a starting pool of 1,000 distinctive haplotypes. For every founder, a random haplotype was picked in the starting pool and then attributed to each of them. **Table 2.2** summarizes the simulations performed.

Rapidly mutating Y-STRs

We did additional simulations to assess the increase in discriminatory power provided by rapidly mutating Y-STRs (RM-STRs). Unique haplotypes were randomly created using markers included in Yfiler™ Plus kit, which contains seven RM-STRs. Simulations were conducted with genealogical data (**Table. 2.2**).

Table 2.2 Parameters settings for the simulations conducted in this study.

Parameters	Simulations without genealogical data		Simulations with genealogical data			
Population size (number of men in the last generation)	1,189,672	1,189,672	Paternal lineages were built from the BALSAC population register. Lineage of at least 10 generations kept.			
Number of generations	15	15				
Variance in male reproductive success	0.2	0.98				
Genealogical error rate	n.a.	n.a.	0.008	0.008	0.008	0.008
Haplotype attributed to founders	A single haplotype created by randomly selecting alleles at the 23 loci of the PowerPlex Y23 kit	A single haplotype created by randomly selecting alleles at the 23 loci of the PowerPlex Y23 kit	From a starting pool of 100 haplotypes created by randomly selecting alleles at the 23 loci of the PowerPlex Y23 kit	From a starting pool of 1,000 haplotypes created by randomly selecting alleles at the 23 loci of the PowerPlex Y23 kit	A unique haplotype attributed to each founder by randomly selecting alleles at the 23 loci of the PowerPlex Y23 kit	A unique haplotype attributed to each founder by randomly selecting alleles at the 27 loci of the Yfiler Plus kit
Mutation rate	Mutation rates of loci used to create haplotypes					

Spatial Analysis

Simulations using French-Canadian genealogical data were used to assess the spatial heterogeneity in the distribution of Y haplotypes in the population. To do so, men sharing a given haplotype were divided according to the regions they got married, for each 23 Québec regions as delimited in the BALSAC register (note that those delimitations based on colonization history and data structure do not entirely map to current administrative regions; **Figure 2.1**). Then, the distribution of men sharing haplotypes in each region was compared to that for the whole Québec province.

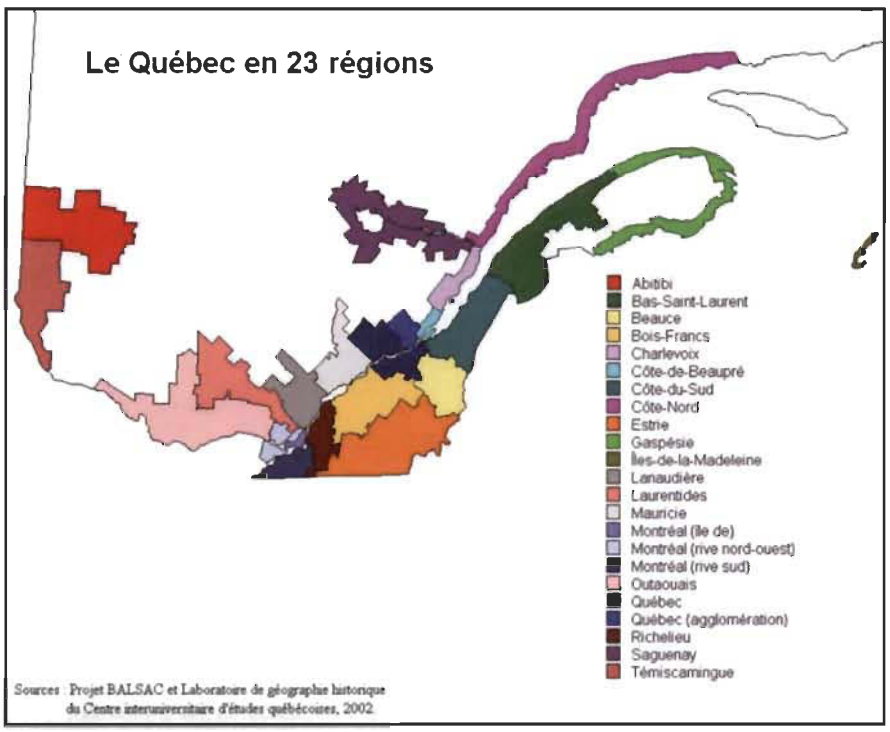

Figure 2.1 The 23 Québec regions delimited in the BALSAC register (reprinted from http://balsac.uqac.ca/fichier-balsac/apercu-des-donnees/).

Results

For every simulation parameter setting, the distribution of men sharing same haplotypes in the population was obtained by keeping the last three generations of each paternal lineage (herein the "living population").

Simulations without genealogical data

After simulations, 2,412,990 men are included in the living population. When the variance in reproductive success was set to 0.20, a high proportion of haplotypes are shared by 50 men or less **(Figure 2.2 A)**, and none by more than 160 men (**Figure 2.S1 A**). In fact, 50% of the haplotypes are shared by four men or less (**Table 2.S1**) and 99% by 48 men or fewer (**Table 2.3**). The most frequent haplotype in the population was shared by 152 men (**Table 2.S2**). The percentage of individuals carrying a haplotype included in the 1% most frequent ones is 6.3% (**Table 2.3**).

Results of simulations with a VRS set to 0.98 are quite similar. A major part of haplotypes is shared by 50 men or less, and none by more than 160 men (**Figure 2.2B**), but again in this case, a small proportion of haplotypes are shared by more than 50 men (**Figure 2.S1B**). The most frequent haplotype is shared by 150 men (**Table 2.S2**) and 99% of haplotypes are shared by 42 men or less. The percentage of individuals carrying a haplotype included in the 1% most frequent ones is 6.7% (**Table 2.3**).

Figure 2.2 Distribution of men sharing the same haplotype for simulations without genealogical data. The variance in reproductive success was set to 0.2 (panel A) or 0.98 (panel B).

Table 2.3 Proportion of the simulated population sharing a Y-STR haplotype more frequent than the 99th percentile.

Variance in reproductive success	Number of individuals in the living population	99th percentile*	Number of men carrying a haplotype more frequent than the 99th percentile	Proportion of men (%)
0.20	2,414,547	48	151,553	6.3
0.98	2,500,223	42	167,416	6.7

*number of men sharing a haplotype corresponding to the 99th percentile, i.e. with only 1% of haplotypes being shared by more men.

Simulations with genealogical data

To assess how well the Andersen and Balding's model fits with the French-Canadian population, we realized simulations keeping paternal lineage of at least 10 generations, so that the living population was composed of 523,835 men. In a first simulation, a unique haplotype was attributed to every founder using the same 23 markers as in simulations without genealogical data. The results show that a high proportion of haplotypes are shared by 50 men or less but, contrasting with the previous simulations, also that a very small proportion of haplotypes are shared by more than 1,000 men (**Figure 2.3 G,H,I**). More precisely, 99% of the haplotypes in the population are shared by 197 men or fewer (**Table 2.4**).

To relax the assumption that each founder carried a unique haplotype, other simulations were done by attributing them haplotypes randomly drawn from a pool of either 100 or 1,000 different ones, created in the same manner (i.e. by randomly selecting alleles for each locus). Again, the majority of haplotypes in the "living population" were shared by 50 men or less (**Figure 2.3 A,D**). In simulations using a starting pool of 100 haplotypes, the 99th percentile value of the distribution of men sharing the same haplotype was 273 individuals (**Figure 2.3 B, Table 2.4**), while that number increased to 295 men with a starting pool of 1,000 haplotypes (**Figure 2.3 E, Table 2.4**). Individuals carrying a haplotype among those 1% most frequent ones represented ~57% of the population with both starting pools (**Table 2.4**). The most frequent haplotype was carried by 7,302 men in the population (**Table 2.S2**).

Table 2.4 Proportion of the population reconstructed from genealogical data sharing a Y-STR haplotype more frequent than 99th percentile.

Starting pool of haplotypes	Number of individuals in the living population	99th percentile	Number of men carrying a haplotype more frequent than 99th percentile	Proportion of men (%)
100	523,835	273	302,359	57.7
1,000		295	297,915	56.9
Unique to each founder		197	198,583	37.9

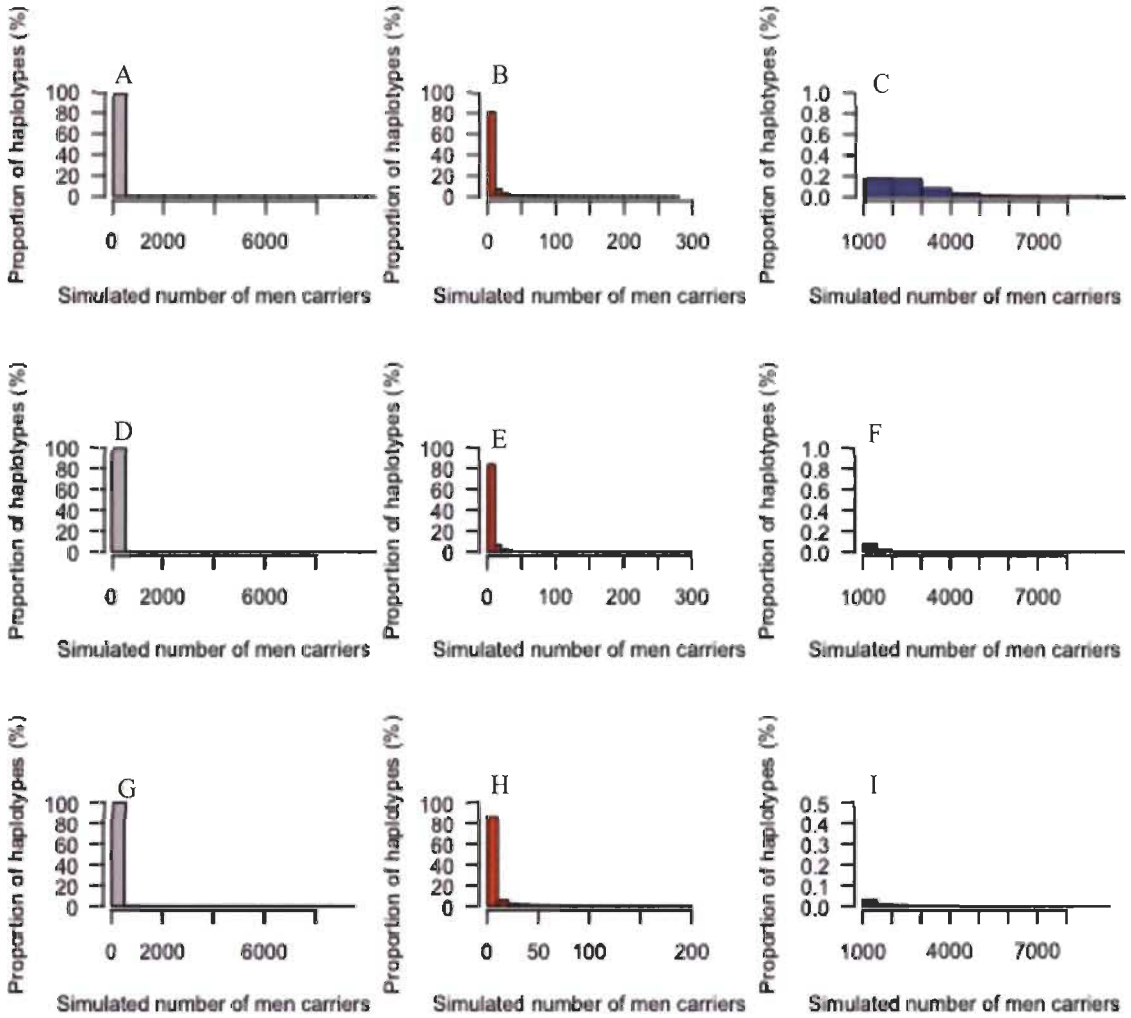

Figure 2.3 Distribution of men sharing the same haplotype for simulations using genealogical data.

Panels A, B and C show the distribution for simulations where haplotypes attributed to founders were randomly drawn from a pool of 100. Panels D, E and F show the distribution for simulations where haplotypes were randomly drawn from a pool of 1,000. Panels G, H and I show the distribution for simulations where a unique haplotype was attributed to every founder. Grey panels (A, D, G) represent all haplotypes, red panels (B, E, H) show a close up on haplotypes shared by fewer men than the 99th percentile value for the corresponding simulation, and blue panels (C, F, I) show haplotypes carried by more than 1,000 men.

Rapidly mutating Y-STRs

Using rapidly mutating markers diminished the number of men sharing a haplotype (**Figure 2.4**). The value of the 99th percentile was 98 compared to 197 without RM-STRs. The proportion of individuals carrying haplotypes among those 1% most frequent dropped to 33.0% while it was higher in other simulations using the genealogical data (37.9% – 57.7%, **Table 2.4**) and the most frequent one was shared by 6,633 men (**Table 2.S2**).

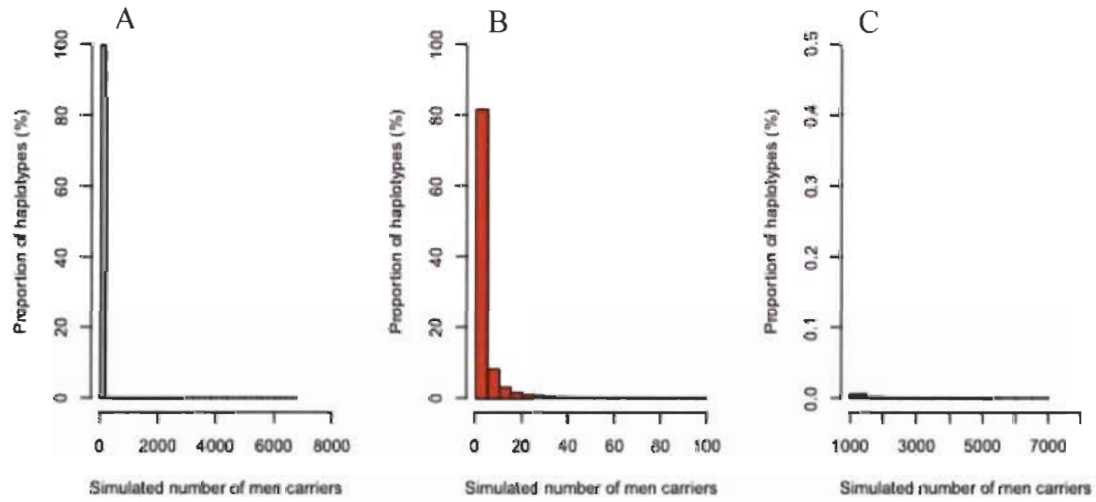

Figure 2.4 Distribution of men sharing the same haplotype for simulation using genealogical data and incorporating rapidly mutating markers.

Panel A shows the proportion of haplotypes in the population as a function of the number of male carriers. Panels B and C show the distribution only for haplotypes shared by 100 or 1,000 men or fewer, respectively.

Spatial analysis

Simulations with genealogical data were used to study the spatial distribution of Y-STR haplotypes in the whole population of Québec, as well as in four regions selected on the basis of their high proportion in male individuals in the living population (**Table 2.S5**): Saguenay, Bas-St-Laurent, Mauricie, Estrie (**Figure 2.1**). The region of marriage was unknown for 976 (0.19%) individuals of the living population (**Table 2.S5**). Each of these four regions had between 675 and 1,327 haplotypes not found in other regions, representing ~ 20% of all haplotypes observed in those regions (**Table 2.5**; **Figure 2.5**).

Table 2.5 Number of haplotypes observed in each of four Québec regions, as obtained from simulations with genealogical data.

Region	Number of haplotypes observed	Number of haplotypes shared by men only living in that region	Proportion of haplotypes shared by men only living in that region (%)
Bas-St-Laurent	3,288	675	20.5
Estrie	6,231	1,327	21.3
Mauricie	5,255	1,270	24.2
Saguenay	2,760	467	17.0

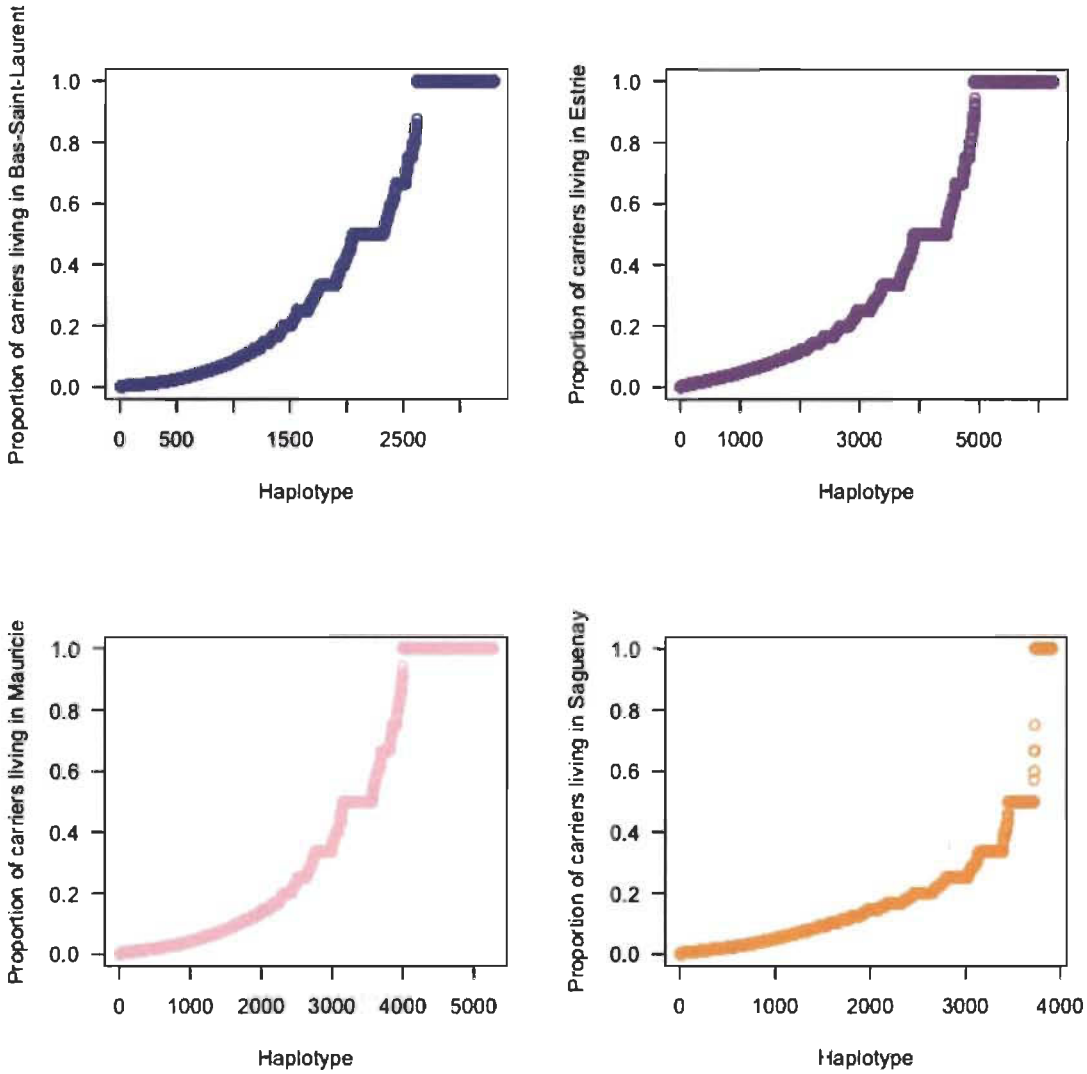

Figure 2.5 Proportion of men from a specific region carrying a given haplotype, for all haplotypes observed in that region, shown by increasing order of proportion. The results are shown for the four regions selected: Saguenay, Mauricie, Estrie and Bas-Saint-Laurent.

As a way to visualize the variation in the spatial distribution of simulated haplotypes, the latter were classified in four different groups: those shared by ≥ 50 and < 100 men, by ≥ 100 and < 500 men, by ≥ 500 and < 1 000 men, and by ≥ 1 000 men. Four haplotypes were randomly drawn in each group and their spatial distribution is illustrated in **Figure 2.6**.

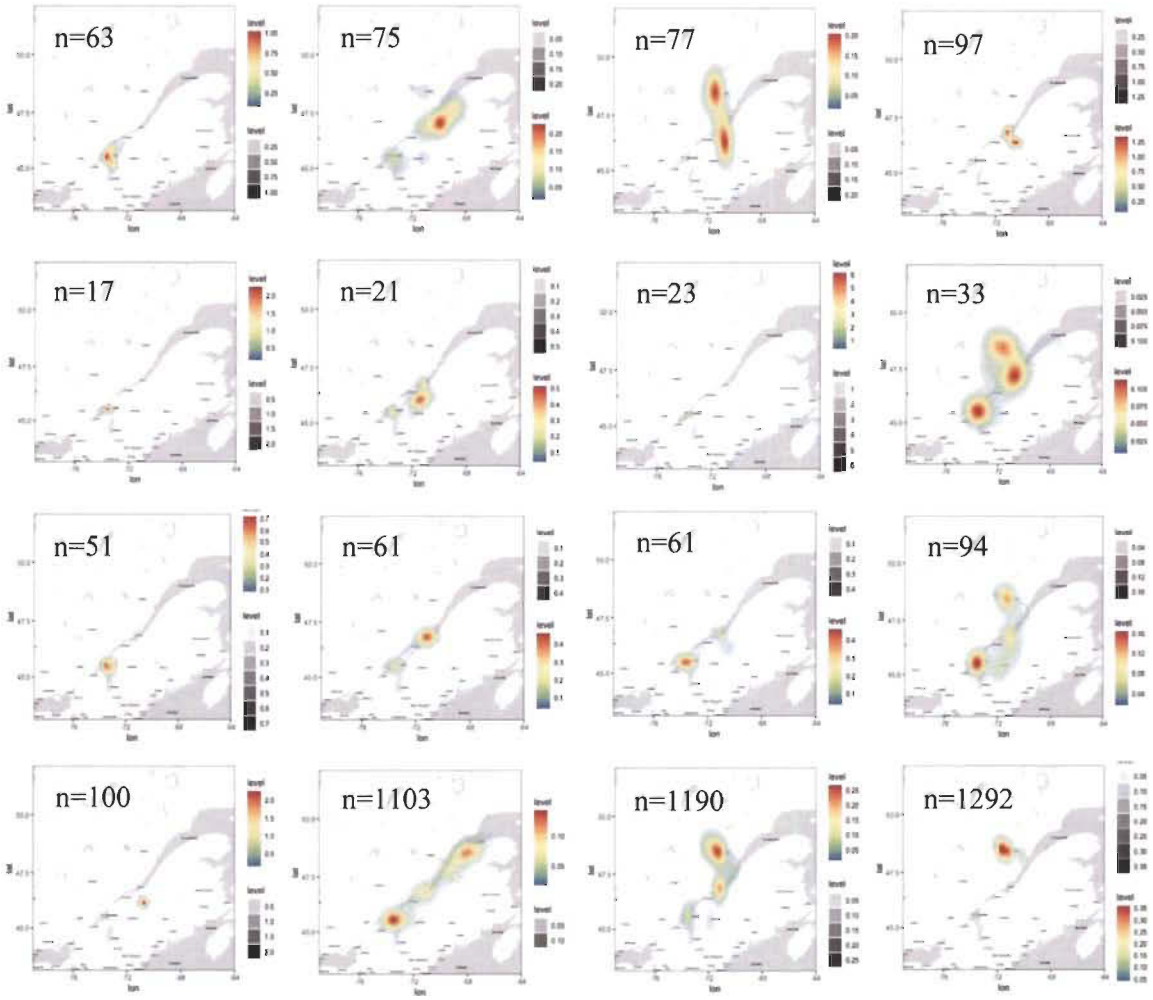

Figure 2.6 Geographical maps representing the spatial distribution of 16 haplotypes randomly drawn in four different groups defined by their number of male carriers. The boundaries of the groups were as follows: haplotypes shared by ≥ 50 and < 100 men (top line of panels), by ≥ 100 and < 500 men (second line), by ≥ 500 and < 1 000 men (third line), and by ≥ 1 000 men (bottom line). The values of n correspond to the number of men sharing the haplotype mapped.

Discussion

Test of Ansersen and Balding's model

Our objective was to test empirically the simulation model proposed by Andersen and Balding (2017) to weight Y-chromosome DNA evidence in criminal casework. The genealogical knowledge of the French-Canadian population since its foundation in 1608 allowed us to achieve that goal to an extent that could not be possible for most other human populations. To do so, we compared distributions of men sharing the same haplotype in two typed of simulations: 1) without the incorporation of genealogical data, as in the original model is intended because such data is unavailable, at least at a large scale, but for a handful of populations; 2) with the incorporation of whole-population scale genealogical data, which is expected to provide results much closer to the reality. Results obtained with different simulation parameter settings differ from one another. The value defining the 99th percentile of the distribution of men sharing haplotypes is approximately four times higher using genealogical data compared to when the population is simulated (197 men vs. 42-48 men, respectively). Moreover, the proportion of the male population carrying a haplotype more frequent than this 99th percentile value was six times higher with than without genealogical data (37.9% vs. 6.3%, respectively). Those results highlight the large discrepancy between the population simulated based on French-Canadian demographic parameters and the real population known from the genealogy. The former is composed of over 100,000 paternal lineages while the latter of the 6,485 lineages are included in the genealogical data (**Table 2.S3**). As a result, the number of individuals included in a paternal lineage is, on average, much smaller in the simulated than the real population (**Table 2.S3**). Considering that a haplotype is mainly shared by men from the same paternal lineage, the distribution of men carrying the same haplotype is thus shifted toward lower values in the simulated population (Andersen and Balding, 2017). Using a higher value for the variance in male reproductive success (VRS of 0.98 instead of 0.20) had a very limited effect on the results. Another point that could explain, in part, the differences between simulations is that we set the size of the simulated "living" population as equal to the size of the current French-Canadian population (data of 2019),

while our genealogical data stops in 1960, when the population was smaller. However, a smaller population should contain shorter lineages with fewer living representatives, hence fewer men sharing a given haplotype, on average (all else being equal). Our results show the opposite.

Founder effects

Founder effects largely explain why simulations done without genealogical data generated a population with a structure quite different from that of the French-Canadian population. The French-Canadian population has a particular genetic structure as it was founded by a ~5,000 French male settlers some 15 generations ago (Vezina *et al.*, 2006). Nowadays, Québec's population counts 8.5 millions of individuals, 80% of them being descendants of these French-Canadian founders (Charbonneau *et al.*, 2000; Scriver, 2001; Moreau *et al.*, 2013). Therefore, men belong to a limited number of paternal lineages compared to what is expected from the simulated population. Consequently, the number of paternal lineages simulated exceeded 100,000, which is largely over the 6,500 lineages observed in the genealogical data, and consistent with the number of founders. Note that the larger number of paternal lineages than male founders in the genealogical data is explained by immigrants of later arrival, as well as to a lack of familial information to connect some apparently shorter lineages to the rest of the genealogy.

It was shown that founder effects were important for the genetic of the Québec population. Scriver (2001) found that 64% of founders contributed to the autosomal genome of only one individual among 2,221 people married between 1945 and 1965, whereas only 0.17% of founders contributed to the genome of ≥ 20 individuals (Tremblay and Vezina, 2000). Such strong founder effects were not observed in the simulation without genealogical data, as the most frequent haplotype was shared by 158 men in the living population and as only 7% of men carried a haplotype among those in the 1% most frequent ones. These results line up with those of Andersen and Balding (2017), who demonstrated that most haplotypes are shared by a few men and rarely by more than a few hundreds. In comparison, when the real genealogy was incorporated in the simulations, the most frequent haplotype was shared by 7,302 men, whereas nearly 40% of men carried

a haplotype among those in the 1% most frequent ones. Moreover, when we compared the values of the 99th percentile between the two types of simulation, the values are four times higher in the simulations using the genealogical data. The comparison between the distributions of men sharing a haplotype in the two models shows that founder effects had non-negligible consequences on the genealogical structure of paternal lineages in the French-Canadian population, which are not reproduced by simulations using the original model of Andersen and Balding. However, every population in the world has its own demographic and genetic structure. They vary in the number of generations since foundation or back to coalescence of Y haplotypes, and founder effects also depend on the history of the settlement. The various parameters tested by Andersen and Balding (2017) suggest that the model could perform better in a population with a longer history than the French-Canadian one and with less pronounced founder effects. In their article, these authors test multiple values of VRS (0;0.20;1), population growth rate (1;1.02), Y-STRs kit (Y filer; PowerPlex Y23; Yfiler Plus) simulating a population size of 10^5 with a depth of 250 generations. Their results show, for the same parameters used for the French-Canadian population (i.e. PowerPlex Y23 and VRS 0.20), that the value of the 99th percentile was of 95, which is twice lower of what we found in simulations with genealogical data, and twice higher than for our simulated population. Nevertheless, considering the goal of the method, namely to measure the weight of Y DNA evidence, in all cases the 99th percentile values (theirs and ours) are comparable in magnitude, suggesting that the vast majority of Y haplotypes in different human populations will be shared by a modest number of men. The striking difference, though, is observed for those 1% more frequent haplotypes, for which Andersen and Balding's model provide results very far from reality and moreover not conservative, due to founder effects in the French-Canadian population.

Rapidly mutating Y-STRs

Rapidly mutating Y-STRs provide more resolution to separate the Y chromosome profile of men belonging to the same paternal lineage (Rakha *et al.*, 2018; Wang *et al.*, 2019). The study of 18 endogamous Pakistani families show that RM-STRs increases

discriminative power by 29.2% and 26.0% compared to the AmpFlSTR® Yfiler™ and PowerPlex® Y23 kits, respectively, which do not include such markers (Rakha *et al.*, 2018). Moreover, among 4,096 pairs of same-lineage men, 35.3% could be separated using 13 RM-STRs compared to 9.6% with the Y-filer kit (Adnan *et al.*, 2019). In our simulation including genealogical data, a shift towards lower values was observed in the distribution of men sharing a haplotype when RM Y-STRs were included. In the living population, 99% of haplotypes were shared by 197 men or fewer with markers from the PowerPlex Y23® kit, in comparison to 98 men or fewer with the Yfiler™ Plus kit that contains seven RM Y-STRs. Similarly, Balding and Andersen (2017) report 99th percentile values of 95 with PowerPlex23 kit and 57 with the Yfiler Plus kit, the latter including RM Y-STRs. The use of these seven RM Y-STRs (added to other non RM markers) decreased the proportion of highly frequent Y haplotypes in the population: the most frequent haplotype was shared by 6,633 men when including RM Y-STRs and by 9,041 men when not. Arguably, these differences are modest and suggest that RM Y-STRs may not always make a forensically significant difference in terms of discriminatory power or weight of evidence, especially to tell apart men sharing a common ancestor several generations ago.

Comparison with a genealogico-molecular model

To better understand how Y haplotypes are distributed in the French-Canadian population of Québec and how it impacts the accuracy of the model proposed, we compared our results with those from a recent model developed by Doyon and *al.* (submitted) to estimate Y haplotypes frequencies in the French-Canadian population and combining molecular and genealogical data. These authors used Y-STR profiles (17 or 20 markers) of real men living in Québec connected to the genealogy so that their paternal lineage is known. From there, Y haplotypes were imputed to their ancestors up to lineage founders and then back to all other descendants, accounting for mutation rates and genealogical errors (Doyon *et al.*, submitted). Comparing haplotype frequencies in five different regions (Bas-Saint-Laurent, Côte-Nord, Côte-du-Sud, Gaspésie, Iles-de-la-Madeleine) relative to the frequency in these five regions pooled, using the $\log_{10}$ value of the ratio of

the larger over the smaller frequency (LOD score). They show that the minimal and minimal LOD scores were respectively < 0.001 and 6.4, meaning that at one extreme some haplotypes were equally frequent in a region than in another ($10^{0.001} \approx 1$), and at the other extreme that some haplotypes are up to $10^{6.4} \approx 2{,}500{,}000$ times more frequent in one region than in the five regions pooled (or the reverse).

We did a similar exercise using the results from our simulations incorporating the genealogy to compare the frequency of haplotypes in the same five regions. LOD scores ranged between < 0.001 and 2.6 (**Table 2.S4**). The maximal value indicates that some haplotypes are up to $10^{2.6} \approx 316$ times more frequent in one region than in the whole Québec (or the reverse). Several factors may explain this difference with Doyon *et al.*'s results. First, these authors did not include all Québec regions in the pooled sample due to limited molecular coverage in several regions, while our model considered all paternal lineages and all regions. Considering more lineages in a given area should result in lower frequency estimates. Second, we used haplotypes at 23 Y-STRs while Doyon and *al.* (submitted) used either 17 or 20 Y-STRs. Consequently, the greater haplotypic diversity in our simulations may limit situations whereby a haplotype is very frequent in one region and much less in another one. Third, molecular data likely better reflect the spatial variation in frequency for the most common haplotypes, suggesting that Andersen and Balding's model is especially sensitive to population history and idiosyncrasies shaping the distribution of these haplotypes. These observations lead us to propose a mixed approach to assign a weight to DNA evidence involving Y-STR profiles, as described in the next section.

A mixed approach to report the weight of Y-STR evidence

Our results suggest that the 99th percentile value is not sufficient, from a forensic point of view, to evaluate the weight of Y-STR evidence in criminal casework, at least for populations with a short history and strong founder effects, such as the French-Canadian one. For example, in a real forensic case, if a match occurred between the Y haplotype of a suspect and the one recovered on the trace, the meaning of the 99th percentile value

would strongly on whether the haplotype is part of the 1% most frequent or of the 99% rarer haplotypes. This could impact the probative value to the detriment of a suspect.

A mixed approach to weigh evidence could resolve this issue. First, since Andersen and Balding's model seems as a good tool to estimate the distribution of men sharing rarer haplotypes, the 99th percentile of this distribution could be used when the haplotype is actually known to be a rare one. Second, the 99th percentile could be used to determine the minimal size a molecular sample should have to allow the identification of those frequent haplotypes in the population, i.e. those among the 1% most frequent. Here is an example. Let's say that a forensic laboratory used *malan* to estimate the distribution of men sharing Y haplotypes and that the results show that 99% of haplotypes are shared by 287 or fewer men. Let's also say that the rarest among the 1% "frequent" haplotypes is expected to be shared by 500 men in a population of one million, thus in 1/2000 men. Therefore, a random sample of minimally 2000 men typed at Y-STRs would be necessary to identify all (or most) these more frequent haplotypes and estimate their frequency in the population. Having this list, it would be possible to determine if the suspect's Y haplotype belong to the set of frequent ones. If not, i.e. if that haplotype had a frequency <1/2000 in the sample or was not observed at all, then the 99th percentile serve to provide a weight to the DNA evidence. Otherwise, the probative value could be based on the populational frequency estimated for this haplotype.

To further demonstrate this approach, we compare our results with those from the genealogico-molecular model of Doyon *et al.* (submitted). In our study, when a unique haplotype is attributed to every founder, the simulations showed that 99% of the haplotypes are shared by 197 men or less and that the most common Y haplotype is shared by 9,041 men in a population of 523,835, so that one man out of 60 possesses that Y haplotype (for the sake of the demonstration here we use the values from simulations with genealogical data, which are larger, hence more conservative from the defense perspective). On the other hand, Doyon *et al.* (submitted) study was based on a sample of 429 men, meaning that we expect a haplotype to be detected if it occurs at least in ~9-10,000 men of a population of approximately million men. Consequently, a sample size of 429 is likely to miss many frequent haplotypes or, in other words, would not allow

telling apart rare and frequent ones. Interestingly, and at odds with the preceding statement, 65 haplotypes are observed more than once in Doyon *et al.*'s data, specifically between two and six times. This likely reflects the spatial heterogeneity in haplotype frequencies that can cause undetected stratification in samples (see Doyon *and al.*, submitted). Consequently, using the 99th percentile of the distribution of men sharing haplotypes may not be appropriate when there is spatial heterogeneity in Y haplotype frequencies and when molecular samples do not account explicitly for this heterogeneity.

Spatial analysis

A few studies have been done on the genetic diversity of the French-Canadian population, combining molecular and genealogy information (Gagnon and Heyer, 2001; Doyon, 2018). They show that some regions are more genetically homogeneous. Genetic homogeneity is about six times higher in region located East of Québec City than in those located to the West (Gagnon and Heyer, 2001). Doyon *and al.* further demonstrated that Y haplotype frequencies vary not only among regions but also at a scale as small as localities (Doyon, 2018). However, this study only considered regions where the coverage of molecular data was sufficient (five regions). The simulation model allowed us to study the spatial distribution of Y haplotypes in the French-Canadian population using all regions and the full genealogy of Québec. Our results show that, for each region studied, a proportion of haplotypes occurred only in that region. The number of haplotypes unique varied among regions, with Mauricie and Estrie showing the highest proportions of unique haplotypes. Actually, the spatial distribution of Y haplotypes is heterogeneous across Québec, as we illustrated by the geographical maps of 16 different haplotypes. Given those results, it's important that forensic laboratories define adequately the boundaries of the population of interest (i.e. of individuals that could be potential suspects). As demonstrate here, some haplotypes will be found in one or a few specific regions, an important element to consider when calculating the weight of DNA evidence, which is currently not the case because forensic labs typically use Y-STR reference samples/databases at the scale of the large regions (country, province, etc.) under their jurisdiction.

Conclusion

Andersen and Balding's simulation model implemented in the *malan* package provides a new and easier-to-interpret way to calculate and present probative value in court for Y-STR haplotypes. The test of the model in the French-Canadian population allowed us to highlight some advantages and limitations related to the interpretation of DNA matches involving Y haplotypes. Our results show that the model is sensitive to founder effects and spatial heterogeneity in Y haplotype distribution. Nevertheless, an approach mixing Andersen and Balding's model and molecular data can help resolve these issues. In addition, the parameters set by the user help to adapt the simulations to different populations. Also, in a forensic context we must recall that the definition of the population of interest is of high importance and, considering the potential population genetic structure at Y chromosome markers, it should not be neglected as it can influence the opinion of judges and juries.

References

Adnan, A., Rakha, A., Nazir, S., Khan, M.F., Hadi, S., and Xuan, J. (2019). "Evaluation of 13 rapidly mutating Y-STRs in endogamous Punjabi and Sindhi ethnic groups from Pakistan." *International Journal of Legal Medicine*, **133**(3): 799-802.

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

Andersen, M.M., and Balding, D.J. (2019). "Y-profile evidence: Close paternal relatives and mixtures." *Forensic Science International: Genetics*, **38**: 48-53.

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

Butler, J. (2014). "Advanced topics in forensic dna typing : Interpretation". Boston, MA: Elsevier.

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

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

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

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." *Mémoire présenté à l'Université du Québec à Trois-Rivières.*

Doyon, A., Moreau, C., Labuda, D., and Milot,E. (2020). "Spatiotemporal variation of mitochondrial DNA and Y chromosome haplotype frequencies in a French-Canadian population and its impact on random match probabilities." *Sumitted.*

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

Gagnon, A. and Heyer, E. (2001). "Fragmentation of the Quebec population genetic pool (Canada): evidence from the genetic contribution of founders per region in the 17th and 18th centuries." *American Journal of Physical Anthropology,* **114**(1): 30-41.

Gagnon, A. and Heyer, E. (2001). "Intergenerational correlation of effective family size in early Quebec (Canada)." *American Journal of Human Biology*, **13**(5): 645-659.

Gill, P., Brenner, C., Brinkmann, B., Budowle, B., Carracedo, A., Jobling, M.A., de Knijff, P., Kayser, M., Krawczak, M., *et al.* (2001). "DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs." *Forensic Science International,* **124**(1): 5-10.

Gusmao, L., Butler, J.M., Carracedo, A., Gill, P., Kayser, M., Mayr, W.R., Morling, N., Prinz, M., Roewer, L., *et al.* (2006). "DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis." *International Journal of Legal Medicine* **120**(4): 191-200.

Kayser, M. (2017). "Forensic use of Y-chromosome DNA: a general overview." *Human Genetics,* **136**(5): 621-635.

Kayser, M., Brauer, S., Schadlich, H., Prinz, M., Batzer, M.A., Zimmerman, P.A., Boatin, B.A., and Stoneking, M. (2003). "Y chromosome STR haplotypes and the genetic structure of U.S. populations of African, European, and Hispanic ancestry." *Genome Research,* **13**(4): 624-634.

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

Moreau, C., Lefebvre, J.-F., Jomphe, M., Bherer, C., Ruiz-Linares, A., Vezina, H., Roy-Gagnon, M.-H., and Labuda, D. (2013). "Native American admixture in the Quebec founder population." *PLoS One*, **8**(6): e65507.

National Research Council (U.S.). (1996) "The evaluation of forensic DNA evidence". Washington, D.C.: National Academy Press.

Ploski, R., Wozniak, M., Pawlowski, R., Monies, D.M., Branicki, W., Kupiec, T., Kloosterman, A., Dobosz, A., Bosch, E., *et al.* (2002). "Homogeneity and distinctiveness of Polish paternal lineages revealed by Y chromosome microsatellite haplotype analysis." *Human Genetics,* **110**(6): 592-600.

Purps, J., Geppert, M., Nagy, M., and Roewer, L. (2015). "Validation of a combined autosomal/Y-chromosomal STR approach for analyzing typical biological stains in sexual-assault cases." *Forensic Science International: Genetics*, **19**: 238-242.

Institut de la statistique du Québec. (2019). "Le bilan démographique du Québec." https://www.stat.gouv.qc.ca/statistiques/population-demographie/bilan2019.pdf

Rakha, A., Oh, Y.N., Lee, H.Y., Hussain, S., Waryah, A.M., Adnan, A., and Shin, K.J. (2018). "Discriminating power of rapidly mutating Y-STRs in deep rooted endogamous pedigrees from Sindhi population of Pakistan." *Legal Medicine (Tokyo),* **34**: 17-20.

R Development Core Team (2019). "R: A language and environment for statistical computing. " *R Foundation for Statistical Computing*, Viennna, Austria. http://www.R-project.org.

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

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

s.a. BALSAC – Fichier de population. http://balsac.uqac.ca/ . Accessed on February 12, 2020.

Scriver, C.R. (2001). "Human genetics: lessons from Quebec populations." *Annual Review of Genomics and Human Genetics,* **2**: 69-101.

SWGADM. (2014). "Interpretation Guidelines for Y-Chromosome STR Typing." 1-20. https://www.thermofisher.com/content/dam/LifeTech/Documents/PDFs/SWGDAM_YSTR_Guidelines_APPROVED_01092014_v_02112014_FINAL.pdf

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

Templeton, A.R. (2006). "*Population genetics and microevolutionary theory*". Hoboken, N.J.:Wiley-Liss.

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

Vezina, H., Tremblay, M., Desjardins, B., and Houde, L. (2006). "Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise." *Cahiers québécois de démographie,* **34**(2): 235-250.

Wang, Q., Jin, B., An, G., Zhong, Q., Chen, M., Luo, X., Li, Z., Jiang, Y., Liang, W., and Zhang, L. (2019). "Rapidly mutating Y-STRs study in Chinese Yi population." *International Journal of Legal Medicine,* **133**(1): 45-50.

Weeden, J., M. J. Abrams, M. C. Green and J. Sabini (2006). "Do high-status people really have fewer children? : Education, income, and fertility in the contemporary U.S." *Human Nature,* **17**(4): 377-392.

Supplementary material

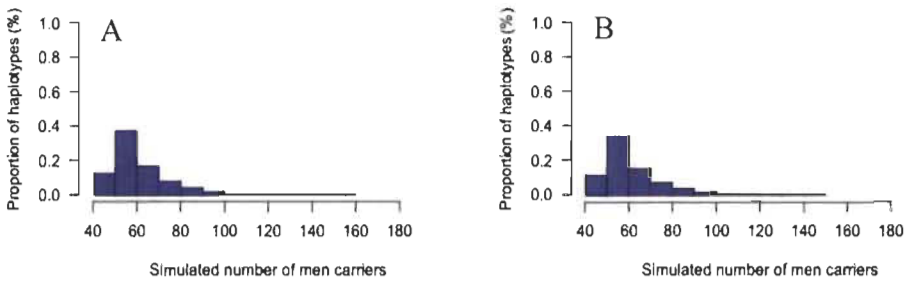

Figure 2.S1 Distribution of men sharing a haplotype, for haplotypes shared by more than 42 (panel A) or 48 (panel B men), which correspond to the values of the 99th percentile for simulation without genealogical dataset. The variance in reproductive success was set to 0.2 (panel A) or 0.98 (panel B).

Table 2.S1 Number of men sharing haplotypes under different parameter settings for simulations with and without genealogical data.

Percentile	No genealogical data		With genealogical data			
	Variance in reproductive success		Starting pool of haplotypes assigned to founders			RM-STRs
	0.20	0.98	100	1,000	Unique to each founder	
50%	4	4	2	2	2	1
99%	48	42	273	295	197	98

Table 2.S2 Number of men carrying the 10 most frequent haplotypes for simulations done with different parameter settings.

	Without genealogy		With genealogy			
	VRS=0.20	VRS=0.98	Starting pool=100	Starting pool=1,000	Unique haplotype	With RM Y-STRS
Haplotype 1	152	150	16,478	12,030	9,041	6,633
Haplotype 2	150	136	16,394	11,981	8,850	6,555
Haplotype 3	139	125	16,066	11,731	8,803	6,436
Haplotype 4	131	124	16,026	11,652	8,656	6,418
Haplotype 5	128	123	15,761	11,646	8,612	6,227
Haplotype 6	127	121	15,582	11,623	8,601	6,013
Haplotype 7	125	117	15,562	11,540	8,529	5,963
Haplotype 8	122	114	15,526	11,362	8,427	5,761
Haplotype 9	120	112	15,420	11,290	8,389	5,629
Haplotype 10	120	110	15,389	11,223	8,364	5,606

Table 2.S3 Characteristics of the population in every type of simulation.

Population	VRS	Number of paternal lineage	Number of men in the living population	Number of haplotypes in the living population	Ratio individual/ paternal lineage
Simulated (no genealogical data)	VRS 0.20	122,619	2,412,990	302,771	19,68
	VRS 0.98	142,399	2,500,223	311,186	17,56
Genealogical data (unique haplotype per founder)		6,485	523,835	32,856	80,78

Table 2.S4 Differences in LOD score of the RMP values between regions against all Quebec

LOD score values	Bas-Saint-Laurent	Côte-du-Sud	Côte-Nord	Gaspésie	Iles-de-la-Madeleine
Minimum	0.00065	0.002606	0.001254	0.001314	0.001196
Maximum	1.560722	1.570808	2.197403	1.912472	2.689

Table 2.S5 Number of men forming the living population and married in each Québec region, according to the BALSAC population register.

Region	Number of men married in the region	Proportion of the living population (%)
ABITIBI	6,712	1.28
BAS SAINT LAURENT	22,599	4.31
BEAUCE	13,761	2.63
BOIS FRANCS	28,876	5.51
CHARLEVOIX	4,163	0.79
COTE DE BEAUPRE	2,970	0.57
COTE DU SUD	14,073	2.69
COTE NORD	3,325	0.63
ESTRIE	31,852	6.08
GASPESIE	6,408	1.22
ILE DE MONTREAL	114,494	21.86
ILES DE LA MADELEINE	1,072	0.20
LANAUDIERE	15,214	2.90
LAURENTIDES	12,883	2.46
MAURICIE	27,206	5.19
OUTAOUAIS	13,315	2.54
QUEBEC (AGGLOMERATION)	30,763	5.87
REGION DE QUEBEC	17,179	3.28
RESTE DU QUEBEC	491	0.09
RICHELIEU	24,951	4.76
RIVE NORD OUEST (MTL)	9,400	1.79
RIVE SUD (MTL)	10,996	2.10
SAGUENAY (LAC ST JEAN)	106,110	20.26
TEMISCAMINGUE	4,046	0.77
NA	976	0.19
TOTAL	523,835	100.00

CHAPITRE III

MÉTHODE – INFORMATIONS SUPPLÉMENTAIRES

3.1 Logiciel R

L'ensemble des analyses présentées dans ce mémoire a été effectué dans le logiciel R v3.6.8 [69]. Ce dernier est utilisé par un grand nombre de chercheurs pour effectuer des analyses statistiques, entre autres. Plusieurs modules de base sont inclus dans le logiciel, contenant des fonctions permettant d'effectuer une multitude d'analyses, alors que d'autres modules offrant des fonctions plus spécifiques sont téléchargeables séparément, depuis le logiciel. Une partie de mon projet de maitrise a été d'apprendre le langage R et de comprendre son fonctionnement, puisque j'étais néophyte dans le domaine. Parmi ces apprentissages, on retrouve la compréhension de fonctions, la résolution de messages d'erreurs, le développement de nouvelles fonctions ainsi que la mise en forme de résultats et de graphiques. Il a également été nécessaire d'apprendre les fonctions contenues dans le module *malan*, qui comporte plusieurs particularités. D'ailleurs, des modifications ont été apportées à *malan* pour adapter les fonctions au besoin du projet, comme la prise en compte de données généalogiques qui sera expliquée en détail dans la section 3.4.

3.2 Lignées paternelles

Les données généalogiques étaient structurées sous forme de tableaux comprenant un numéro d'identité pour les individus ainsi qu'un numéro d'identité pour le père et la mère de l'individu respectivement. Les lignées paternelles ont été extraites, du fondateur jusqu'à la dernière génération connue de sa descendance, en rattachant les individus masculins les uns aux autres par le biais de leur identifiant numérique et celui de leur père. Pour ce faire, le module *pedinf* disponible sur la plateforme GitHub (https://github.com/mikldk/pedinf) a été utilisé. Suite à cette étape, les lignées paternelles ont été identifiées en leur attribuant un numéro unique. Pour chacune d'entre elles, le

nombre d'individus contenus dans la lignée a été répertorié, puis le nombre de générations la constituant a été calculé. Puisque la méthode proposée se base sur la sélection de la population vivante, c'est-à-dire les trois dernières générations, la génération dans laquelle les individus se trouvent dans leur lignée paternelle a été identifiée.

3.3 Module *malan*

Le module *malan* implémente le modèle d'Andersen et Balding (2017) pour simuler une population de lignées paternelles selon différents paramètres : le nombre de générations constituant la population à créer, le nombre d'individus dans la dernière génération ainsi que la variance dans le succès reproducteur des hommes. Une fois la population créée, des haplotypes fictifs sont attribués aux fondateurs composés du nombre de loci Y-STR voulu (23 dans notre cas pour mimer des haplotypes Y qui auraient été obtenus avec la trousse commerciale PowerPlex Y23®). Afin de créer ces haplotypes fictifs, il suffit de spécifier l'ensemble des allèles possibles pour les loci Y-STR désirés ainsi qu'une limite inférieure et supérieure, correspondant respectivement, au plus petit et au plus grand allèle répertorié pour un locus. Ainsi, le logiciel pige au hasard un allèle pour chacun des loci afin de créer un haplotype Y complet à attribuer au fondateur d'une lignée paternelle. Ces haplotypes Y sont ensuite imputés à la descendance en tenant compte du taux de mutations des loci utilisés. En spécifiant une limite inférieure et supérieure, cela empêchera le processus de mutations de créer un allèle inexistant dans la population. Suite à l'attribution des haplotypes, les individus des trois dernières générations sont sélectionnés pour former la population vivante et calculer la distribution des hommes portant un même haplotype. Un exemple simplifié présenté à la **Figure 3.1** permet de mieux comprendre le fonctionnement de cette méthode.

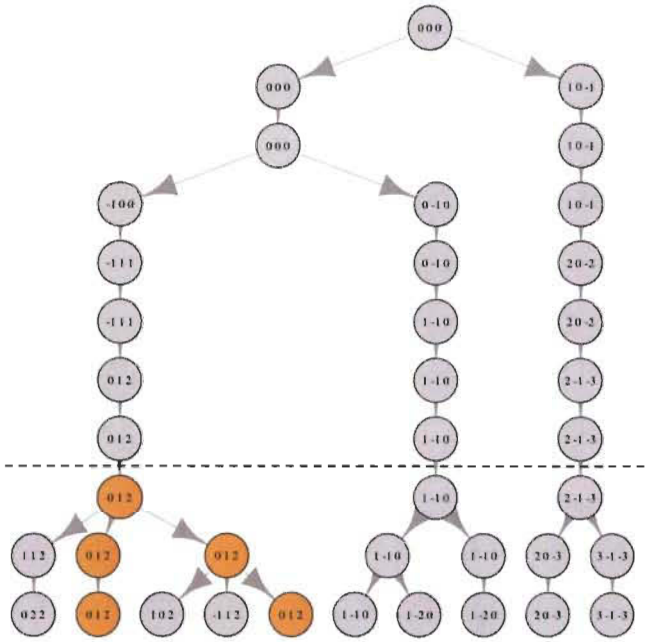

Figure 3.1 Exemple d'une population créée par le modèle implémenté dans *malan.*

Ici, une population de 10 générations a été simulée et des haplotypes à trois loci (fictifs), dont le taux de mutations était de 0.2 pour chacun, ont été attribués au fondateur et à ses descendants. Les individus se trouvant en dessous de la ligne pointillée représentent les hommes des trois dernières générations formant la population dite vivante. Les individus en orange portent tous l'haplotype 0;1;2, démontrant que cinq hommes de la population vivante le partagent. Le même calcul est fait pour chacun des haplotypes afin d'obtenir la distribution des hommes portant un même haplotype.

3.4 Modifications du module *malan*

Deux modifications majeures ont été apportées au module en collaboration avec l'un des développeurs de la méthode, le Pr. Mikkel Meyer Andersen. La première consiste en l'ajout d'une fonction (*load_individuals*) permettant de télécharger les vraies données généalogiques d'une population sous forme de lignées paternelles, plutôt que de simuler la population. La deuxième consiste en l'ajout d'un paramètre, soit la probabilité d'erreurs

généalogiques à l'étape de l'imputation des haplotypes aux descendants. Ce paramètre est spécifié par l'utilisateur dans les simulations faites sur de vraies généalogies (ici celle du Québec). Les erreurs généalogiques consistent en des liens généalogiques erronés en raison d'erreurs lors du jumelage des actes de mariage ou de paternités extra-conjugales [70]. Ce taux a été estimé à 0,008 pour la population canadienne-française, et ce, en utilisant la méthode proposée par Larmuseau et *al.* (2013) qui se base sur le nombre de différences observées entre les haplotypes d'individus d'une même lignée [70]. Ainsi, des erreurs sont générées dans la généalogie en sélectionnant aléatoirement des liens selon une probabilité correspondant au taux d'erreur généalogique. Ce lien est alors considéré comme faux puis retiré de la généalogie. Un haplotype Y est ensuite pigé au hasard parmi ceux ayant été attribués aux fondateurs. Ce dernier est alors attribué au fils du lien père-fils éliminé pour s'assurer qu'il s'agit bien d'un haplotype déjà présent dans la population, mais différent de celui de son père.

3.5 Analyse spatiale

Pour les analyses spatiales, j'ai utilisé les 23 régions (**Figure 2.1**) associées aux individus inclus dans les données généalogiques en utilisant le code du lieu de mariage. Toutefois, le lieu de mariage n'étant pas disponible pour 28% des individus, dans ces cas le lieu de mariage des parents de l'individu a été utilisé.

CHAPITRE IV

DISCUSSION ET PERSPECTIVES

4.1 Discussion

Les méthodes actuelles pour attribuer une valeur probante aux concordances ADN impliquant des haplotypes Y-STR reposent sur l'utilisation des bases de données. Plusieurs auteurs critiquent le fait que ces bases de données ne sont pas construites aléatoirement, mais plutôt à partir de profils provenant de dossiers judiciaires, d'échantillons fournis par des travailleurs de laboratoire et du système de justice ou d'autres donneurs volontaires d'ADN qui ne sont pas nécessairement représentatifs de la population en général [58, 71]. De plus, Andersen et Balding (2017) ont critiqué ces bases de données parce qu'elles fournissent des renseignements difficilement interprétables par la cour. D'abord, en raison du grand nombre d'haplotypes existant dans les populations, et qui sont loin d'être tous représentés dans les bases de données, mais aussi de la possible proximité géographique d'individus appartenant à une même lignée paternelle et portant donc un haplotype Y identique, une réalité qui n'est pas considérée dans les équations disponibles [58]. En proposant un modèle permettant de simuler une population, le problème de l'utilisation des bases de données est évité. Toutefois, ce modèle nécessitait une validation à grande échelle, ce qu'il nous était possible de réaliser avec la généalogie de la population canadienne-française. Cette validation était la première du genre, puisque le modèle n'a, jusqu'à présent, jamais été testé avec de réelles données généalogiques. De plus, ce modèle nous donnait l'occasion d'étudier la répartition spatiale des haplotypes dans la population du Québec.

Nous avons testé le modèle d'Andersen et Balding en comparant les simulations basées sur des valeurs de paramètres démographiques de la population québécoise aux simulations utilisant les vraies données généalogiques. Nos résultats ont démontré quelques divergences au niveau des distributions d'hommes portant un même haplotype.

En comparant la valeur délimitant le 99^{e} percentile, nous remarquons que celle-ci est beaucoup plus basse avec la méthode originale (99% des haplotypes sont portés par 48 hommes ou moins) que celle modifiée utilisant la vraie généalogie (99% des haplotypes sont portés par 197 hommes ou moins). De plus, l'haplotype le plus fréquent n'est jamais porté par plus de 152 hommes dans le premier cas tandis qu'il est porté par 9041 hommes dans le second. Ces différences sont importantes au sens forensique, puisque dans un dossier réel où il y aurait une concordance entre le profil ADN retrouvé sur une trace et celui d'un suspect, l'opinion du décideur de fait pourrait différer selon la valeur probante présentée. Dans ce cas, la valeur obtenue en utilisant le 99^{e} percentile des simulations avec les données généalogiques serait plus favorable à l'accusé puisqu'elle est plus élevée. Lors de l'interprétation des concordances de profils ADN, les experts optent généralement pour la méthode la plus conservatrice, afin d'éviter d'incriminer injustement un innocent. En analysant la structure des populations résultant des différentes simulations, il est possible de constater que le nombre de lignées paternelles présentes dans la population simulée est relativement élevé par rapport au nombre d'individus inclus dans la population vivante (c.-à-d. les trois dernières générations). Effectivement, le nombre de lignées paternelles simulées s'élève à 122 619 (VRS=0,2) et à 142 399 (VRS=0,98) alors que la population vivante est constituée respectivement de 2 412 990 individus et 2 500 223 individus (**Tableau 4.1**). Ainsi, le ratio du nombre de lignée paternelles par rapport au nombre d'individus dans la population vivante est de 1:20 (VRS=0,2) et de 1:18 (VRS=0,98). En comparaison, dans les simulations intégrant les données généalogiques, ce ratio est de 1:81 en conservant les lignées paternelles de 10 générations et plus. Ainsi, avec le modèle original, le nombre de lignées paternelles est beaucoup plus élevé par rapport à ce qui est retrouvé dans les données généalogiques, faisant en sorte que chacune des lignées paternelles créées virtuellement se retrouve avec un nombre plus restreint d'hommes. Puisque nous avons attribué un haplotype unique à chacun des fondateurs de lignées paternelles, il était alors moins probable avec ce modèle de retrouver un haplotype porté par plusieurs centaines d'hommes, contrairement au modèle intégrant les données généalogiques avec lequel la plus grosse lignée paternelle est composée de 28 292 hommes répartis sur seulement 12 générations. En comparaison, la taille des lignées paternelles dans la population simulée variait de 15 à 290 hommes. La méthode

d'Andersen et Balding semble donc assez sensible aux effets fondateurs et devrait mieux fonctionner pour les populations fondées il y a très longtemps (p. ex. en Europe; voir plus bas). Aucune différence notable n'a été observée lorsque la variance dans le succès reproducteur a été augmentée à 0,98 pour tenter de se rapprocher davantage de la population réelle (**Tableau 4.1**).

Tableau 4.1 Informations sur les lignées paternelles et le nombre d'haplotypes observés selon les paramètres d'analyse.

Généalogie	VRS ou nombre de générations	Nombre de lignées paternelles	Nombre d'individus dans la « population vivante »	Nombre d'haplotypes distincts dans la population	Ratio individus/ haplotypes	Ratio individus/ lignées paternelles
simulée	VRS=0,20	122 619	2 412 990	302 771	8,0	19,7
	VRS=0,98	142 399	2 500 223	311 186	8,0	17,6
réelle	≥ 4 générations	8 301	727 294	43 227	16,8	87,6
	≥ 10 générations	6 485	523 835	32 856	15,9	80,8

La population canadienne-française ayant été fondée en 1608 a connu une croissance exponentielle et constitue une population idéale pour tester à grande échelle le modèle proposé par Andersen et Balding (2017), car nous connaissons sa généalogie depuis les premiers mariages célébrés au XVII^{e} siècle. Ces auteurs ont également démontré que les concordances survenaient principalement entre les hommes d'une même lignée paternelle, suggérant que les concordances inter lignées, dues à des mutations accumulées, sont rares. De plus, ils ont démontré de façon théorique que la distance séparant deux individus ayant un même haplotype Y pouvait atteindre dix méioses. La profondeur généalogique de notre jeu de données permettait la prise en compte d'une temporalité de cette ampleur. Des simulations ont été effectuées avec les données généalogiques en ne conservant que les lignées paternelles constituées au minimum par soit quatre générations, soit dix générations. Lorsque les haplotypes ont été attribués aux fondateurs à partir d'un échantillon de départ de 1 000 haplotypes Y, 99% des haplotypes observés dans la population étaient portés par ≤ 336 ou ≤ 295 hommes ou moins pour les

simulations retenant les lignées de ≥ 4 ou ≥ 10 générations, respectivement (**Figure 4.1**). La proportion de la population portant les haplotypes plus fréquents que la valeur délimitant le 99^{e} percentile est de 41,8% comparativement à 41,5% lorsque les lignées d'au minimum dix générations sont utilisées.

Aucune différence notable fut observée dans les résultats selon que les lignées paternelles de ≥ 4 ou ≥ 10 générations furent conservées, alors que nous nous attendions à observer une diminution marquée du nombre d'haplotypes portés par peu d'hommes en sélectionnant uniquement les lignées paternelles de dix générations et plus. En effet, puisque ces dernières remontent plus loin dans le temps, il est raisonnable de penser qu'elles auraient été constituées d'un plus grand nombre d'hommes (donc moins de petites lignées), résultant en des haplotypes qui seraient partagés par davantage d'individus. Or, ce n'est pas ce qui est observé puisque la plus petite lignée est constituée de seulement 20 hommes en conservant les lignées d'au moins dix générations comme démontré dans le **Tableau 4.2**. Ainsi, on peut penser que les lignées paternelles de quatre à neuf générations sont trop peu nombreuses pour réellement impacter les distributions d'hommes portant un même haplotype, et donc la valeur du 99^{e} percentile. D'ailleurs, en calculant la distribution d'hommes portant un même haplotype plutôt que la fréquence de l'haplotype en utilisant une base de données, le modèle *malan* permet de prendre en compte ces divergences entre la grosseur des lignées paternelles, ce qui n'est pas le cas en utilisant la méthode de comptage [58].

Tableau 4.2 Sommaire des lignées paternelles de la population canadienne-française selon le nombre de générations.

Lignées paternelles	Nombre total d'individus	Nombre de lignées paternelles	Taille des lignées paternelles
≥ 4 générations	1 558 205	8 301	5 à 28 292 hommes
≥ 10 générations	1 239 775	6 485	20 à 28 292 hommes

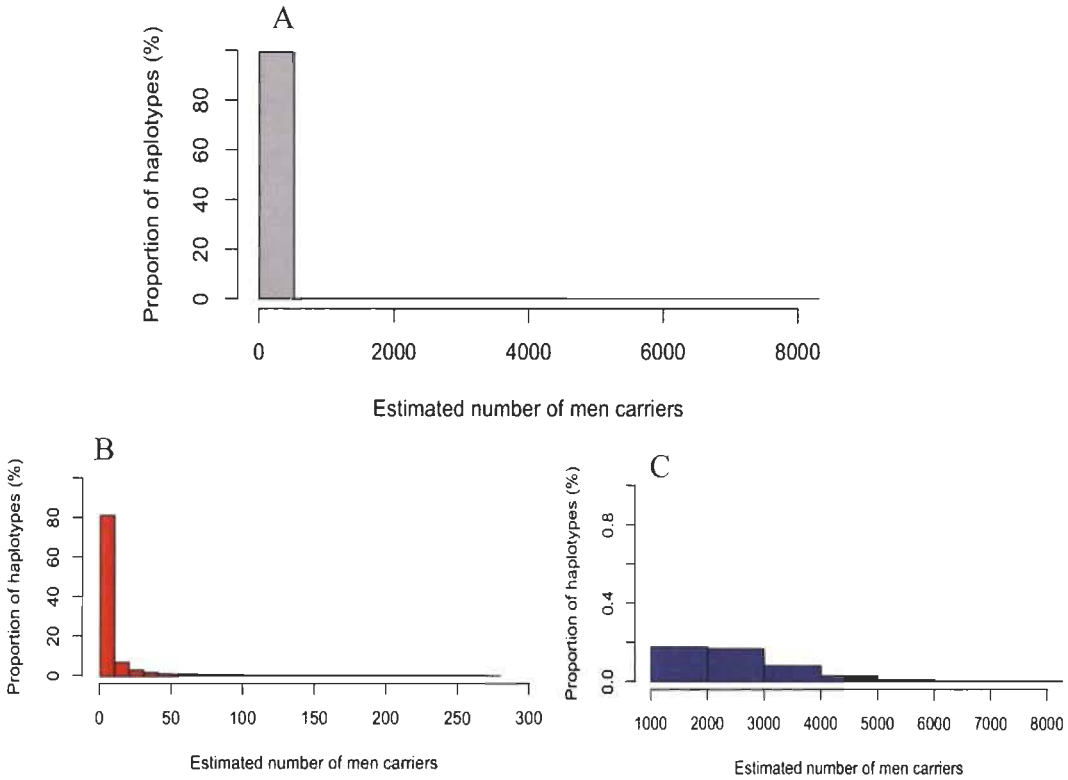

Figure 4.1 Distribution d'hommes portant un même haplotype pour les simulations utilisant les lignées paternelles de quatre générations et plus lorsque les haplotypes sont attribués aux fondateurs à partir d'un échantillon de départ de 1 000.

Les graphiques montrent la proportion d'haplotypes Y dans la population en fonction du nombre d'hommes les portant pour les simulations utilisant la vraie généalogie. A : ensemble des haplotypes. B : haplotypes partagés par 300 hommes et moins. C : haplotypes portés par plus de 1 000 hommes.

Valeur probante et effet fondateur

Tel que mentionné dans l'introduction de ce mémoire, les fondateurs et fondatrices ayant émigré de la France pour s'installer et peupler la Nouvelle France (surtout entre 1608 et 1760) se sont retrouvés isolés géographiquement et socialement de la même patrie [66, 72]. Ainsi, la majorité de la population du Québec moderne descend directement de ces

fondateurs européens [72]. En effet, on estime que ~80% de la population actuelle du Québec (plus de 8 millions d'individus) est d'origine canadienne-française, descendant de quelque 8 500 fondateurs venus s'installer à partir de 1608 [73]. La génétique de cette population est donc caractérisée par des effets fondateurs importants [68]. Pour en tenir compte, nous avons simulé la transmission d'haplotypes Y en utilisant les lignées fondatrices (≥ 10 générations). Andersen et Balding ont mis de l'avant que, typiquement, un haplotype Y retrouvé sur une trace et concordant avec celui d'un suspect était généralement porté par quelques dizaines d'individus et très rarement par plus de quelques centaines d'hommes [74]. Nos résultats se rapprochent de leur prédiction, puisque la valeur du 99^{e} percentile des distributions d'hommes portant le même haplotype est de 197, tandis que seulement 0,99% des haplotypes dans notre population (en moyenne 465 haplotypes) sont portés par plus de 197 hommes. Toutefois, la proportion d'individus qui porte ces quelques haplotypes plus fréquents représente 37,9% de la population vivante, démontrant que certains fondateurs ont un apport génétique beaucoup plus important que d'autres, un aspect déjà démontré par d'autres études sur la population canadienne-française. Par exemple, l'indice de recouvrement, se définissant comme la proportion des lignées généalogiques dans lesquelles un ancêtre donné apparait[6], montre que quatre fondateurs figurent dans 60 à 75% des généalogies du Québec. Gagnon et al. (2001) ont, quant à eux, démontré que seulement 8% des fondateurs pris ensembles comptent pour 50% de l'apport génétique [75]. De plus, ces résultats indiquent que la simple prise en compte du 99^{e} percentile n'est pas suffisante pour des fins forensiques. Effectivement, dans le cas d'un dossier criminel où une concordance surviendrait entre l'haplotype Y d'un suspect et celui retrouvé sur une trace, il ne serait pas nécessairement possible de savoir si cet haplotype est rare ou fait plutôt partie 1% les plus fréquents. Ainsi, l'utilisation de la valeur du 99^{e} percentile de la distribution d'hommes portant un même haplotype s'avèrerait un bon outil pour fournir une valeur probante à la concordance si l'haplotype est rare dans la population. Par contre, s'il est dans le 1% des plus fréquents, alors la valeur du 99^{e} percentile pourrait être très en deçà de la réalité. Nos résultats

---

[6] BALSAC, « Saviez-vous que certains ancêtres apparaissent dans presque toutes les généalogies québécoises? », [En ligne] http://balsac.uqac.ca/blog/2013/11/07/saviez-vous-que-certains-ancetres-apparaissent-dans-presque-toutes-les-genealogies-quebecoises/ (Page consultée le 18 mars 2020)

indiquent en effet que dans la population québécoise, quelques haplotypes sont portés par plusieurs milliers d'hommes, soit un à deux ordres de grandeur au-dessus du 197 cité plus haut.

En plus de l'effet fondateur, nous avons examiné l'impact de l'immigration de plusieurs fondateurs venus d'une même famille ou région de la France et qui ont ainsi pu introduire dans la population canadienne-française le même haplotype Y [76]. Pour ce faire, nous avons attribué les haplotypes de trois façons différentes : à partir d'une base de données contenant 100 haplotypes différents, à partir d'une base de données contenant 1 000 haplotypes différents, ou en attribuant un haplotype unique à chacun des fondateurs. Dans les deux premiers cas, certains haplotypes ont donc été attribués à plus d'un fondateur. Tous les haplotypes étaient formés à partir des 23 marqueurs Y-STR inclus dans la trousse PowerPlex23® de Proméga et nous avons conservé les lignées de dix générations (n=6 485 lignées). Avec 100 haplotypes de départ, nous observons un ratio haplotype/fondateur de 1/65 ce qui veut dire qu'en moyenne 65 fondateurs avaient le même haplotype, contre 1/6,5, soit entre six et sept fondateurs possédant le même haplotype, avec 1 000 haplotypes de départ. Toutefois, un écart plutôt minime est observé dans la proportion de la population portant un haplotype plus fréquent que le 99^{e} percentile, passant de 57,7% des hommes avec 100 haplotypes de départ à 56,9% avec 1 000 haplotypes. En contrepartie, cette différence est plutôt marquée lorsqu'on compare ces pourcentages à celui obtenu (37,9%) lorsque qu'un haplotypes unique fut attribué à chaque fondateur (**Tableau 4.3**). Ainsi, l'effet fondateur est accentué par le fait que certaines familles françaises sont représentées plus d'une fois (p. ex. frères) parmi les fondateurs de la Nouvelle France.

Tableau 4.3 Proportion de la population portant un haplotype plus fréquent que le 99e percentile selon le nombre d'haplotypes utilisés au départ pour l'attribution aux fondateurs.

Nombre d'haplotypes de départ	Valeur du 99e percentile	Nombre d'individus dans la population « vivante »	Nombre d'hommes portant un haplotype plus fréquent que la valeur du 99e percentile	Proportion de la population (%)
100	273	523 835	302 359	57,7
1 000	295		297 915	56,9
Haplotype unique	197		198 583	37,9

Une approche mixte pour attribuer une valeur probante aux concordances Y

La valeur du 99e percentile pourrait malgré tout être un excellent guide servant aux laboratoires forensiques à établir la taille minimale d'un échantillon d'ADN à collecter aléatoirement dans la population masculine, puis à typer aux marqueurs Y ceux-ci, afin de déterminer l'identité de ces fameux haplotypes plus fréquents dans la population. Par exemple, supposons qu'un laboratoire utilise les paramètres démographiques de la population sous sa juridiction pour estimer la distribution d'hommes portant un même haplotype Y, avec la méthode originale d'Andersen et Balding. Supposons également que l'analyse estime que 99% des haplotypes sont portés par moins de 200 hommes, mais que le plus « rare » des fréquents est porté par 5 000 hommes dans une population de deux millions d'individus ; ce dernier serait porté par un homme sur 400 dans la population. Ainsi, un échantillon d'au minimum 400 individus (voir quelques milliers) devrait être récolté dans la population pour avoir de bonnes chances de détecter tous les haplotypes les plus fréquents et d'en estimer la fréquence. En utilisant ensuite une approche mixte, des valeurs probantes plus adéquates pourraient être présentées en cour. En effet, il serait possible de d'abord déterminer si l'haplotype d'un suspect fait partie de ceux qui sont rares ou des 1% plus fréquents. La valeur probante rapportée serait donc la valeur du 99e percentile dans le cas des haplotypes rares. Pour ceux fréquents, le scientifique aurait deux choix : 1) rapporter la fréquence de l'haplotype dans la population, si celle-ci peut être estimée avec une précision raisonnable ; 2) mentionner dans son rapport que l'haplotype

détecté fait partie des plus fréquents et pourrait être porté par des milliers d'hommes, par exemple.

Pour développer davantage ce point, nous avons analysé les haplotypes Y réels de 429 hommes de la population canadienne-française présentés dans l'étude de Doyon et al. (soumis) [67]. Ces derniers ont été typés avec les marqueurs de la trousse AmpFℓSTR® Yfiler™ ou ceux de la trousse Yfiler® Plus. Ces deux trousses commerciales partagent 17 marqueurs communs (DYS389I, DYS389II, DYS19, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS385a/b, DYS439, DYS635, DYS458, Y GATA H4, DYS448 et DYS456) qui ont été retenus ici[7]. Or, selon les simulations présentées dans le présent mémoire, chacun des haplotypes du groupe des 1% les plus fréquents serait partagé par entre ~1 000 et ~7 000 d'hommes. Considérant une population d'environ 4 millions d'hommes (d'origine canadienne-française) au Québec, la fréquence des ces haplotypes fréquents oscillerait donc entre ~1/4 000 et ~1/600. Par conséquent, l'échantillon de 429 profils Y est insuffisant pour dresser la liste des haplotypes les plus fréquents. Parmi les haplotypes détectés dans cet échantillon de 429 hommes, 65 apparaissent plus d'une fois, soit entre deux et six fois. Il est donc probable que ces 65 haplotypes, ou du moins la majorité, appartiennent au groupe des 1% les plus fréquents. Cependant, il serait hasardeux de conclure à partir ce petit échantillon que les haplotypes détectés une seule fois appartient au même groupe. Un échantillon d'au moins 8 000 hommes serait nécessaire si l'on souhaite que le plus rare des haplotypes fréquents soit détecté en moyenne au moins deux fois dans un tel échantillon, pour le distinguer des 99% d'haplotypes rares.

Marqueurs à mutation rapide (RM Y-STR)

L'utilisation des *rapidly mutating* Y-STR (RM Y-STR) s'est accrue dans les dernières années en raison de leur meilleure capacité à distinguer des hommes d'une même lignée paternelle [33, 77, 78]. Dans une étude où 1 966 paires père-fils ont été

[7] À noter que le nombre des marqueurs retenus (17) diffère de celui (23) utilisé dans les simulations. Normalement, ces deux nombres devraient être identiques. Nous associons néanmoins les résultats des deux études pour les besoins de la démonstration.

typées, les taux de mutation se situaient entre 10^{-3} et 10^{-4} pour 173 marqueurs Y-STR « standard » alors qu'ils étaient de l'ordre de 10^{-2} pour 13 autres, catégorisés comme des *rapidly mutating* Y-STR [33]. Ballantyne et al. (2011) ont démontré que ces 13 RM Y-STR permettaient de distinguer 50% des dyades père-fils et 60% des paires de frères, alors que ces valeurs étaient de seulement 8% en utilisant les 17 marqueurs « standard » de la trousse Yfiler™ [78]. De plus, Andersen et Balding (2019) ont démontré avec leur modèle de simulation que la valeur du 99e percentile de la distribution des hommes portant le même haplotype était de 115 avec la trousse PowerPlex® Y23 et diminuait à 63 lorsque la trousse Yfiler™ Plus (contenant sept RM Y-STR) était utilisée, une diminution d'un peu plus de 45% [74]. Dans notre étude, 99% des haplotypes étaient portés par 197 hommes ou moins dans l'analyse excluant les RM Y-STR alors que ce nombre était de 98 en incluant ces marqueurs, soit une diminution de 50%, semblable à celle rapportée par Andersen et Balding. En conséquence, la proportion de la population portant un haplotype plus fréquent que le 99e percentile (>98 hommes) a légèrement baissé, passant de 38,0% à 33,2%, tandis que le nombre d'haplotypes portés par plus de 1 000 hommes a nettement diminué. En utilisant la trousse PowerPlex® Y23, 0,05% des haplotypes dans la population étaient portés par plus de 1 000 hommes alors que cette proportion a diminué de cinq fois en utilisant la trousse Yfiler™ Plus, passant à 0,01%. De plus, l'haplotype le plus fréquent dans la population était porté par 6 633 hommes en incluant les RM Y-STR et par 9 041 hommes en les excluant. Considérant que, dans la généalogie du Québec, plusieurs lignées paternelles sont formées de milliers d'individus pouvant porter un même haplotype Y, il serait intéressant d'étudier l'impact qu'aurait d'inclure davantage de RM Y-STR dans les trousses commerciales.

Analyse spatiale

Lorsque les laboratoires judiciaires attribuent une valeur probante à une concordance entre deux profils ADN, qu'ils proviennent de l'analyse des marqueurs autosomaux ou de lignées (ADNmt et Y), une prémisse est faite implicitement, soit celle que les allèles ou haplotypes sont répartis uniformément dans le territoire géographique occupé par la population d'intérêt [79, 80]. Or, plusieurs études ont démontré qu'il existe des structures

génétiques dans la population du Québec, en contradiction avec cette prémisse. En effet, Gagnon et al. (2001) ont démontré une homogénéité génétique plus élevée dans les régions de l'Est par rapport aux régions de l'Ouest [75]. Tout au long du XIX^{e} siècle, des mouvements migratoires ont eu lieu pour peupler les différentes régions les plus à l'est du territoire, ce qui a mené, en association avec un taux de fécondité élevé, à un effet fondateur plus important dans ces régions [81]. Dans une étude menée précédemment dans notre laboratoire, Doyon *et al.* (soumis) ont démontré, en combinant des données moléculaires aux données généalogiques, qu'il y avait une faible diversité génétique dans certaines régions de l'Est, de même que des variations de fréquences d'haplotypes à une échelle locale. Toutefois, le modèle utilisé dans cette étude n'incluait pas l'ensemble des lignées paternelles [67], contrairement à la présente étude. Nos résultats révèlent que certains haplotypes sont probablement exclusifs à une région spécifique. Notamment, on a estimé qu'au moins 1 000 haplotypes sont uniquement présents dans l'une ou l'autre des régions de la Mauricie et de l'Estrie et environ 500 au Saguenay et au Bas-St-Laurent (**Chapitre II – Figure 2.5**). La cartographie géographique permet de constater en un coup d'œil cette variation spatiale dans les fréquences d'haplotypes. La **Figure 4.2** en illustre un exemple. Ainsi, quatre groupes d'haplotypes ont été créés, soit ceux portés par ≥ 50 hommes et < 100 hommes, par ≥ 100 et < 500 hommes, par ≥ 500 hommes et $< 1\,000$ hommes et par $\geq 1\,000$ hommes. Des haplotypes ont été pigés au hasard dans ces groupes puis la répartition géographique de ceux-ci, extraite de nos simulations, est illustrée sur la carte (**Figure 4.2**).

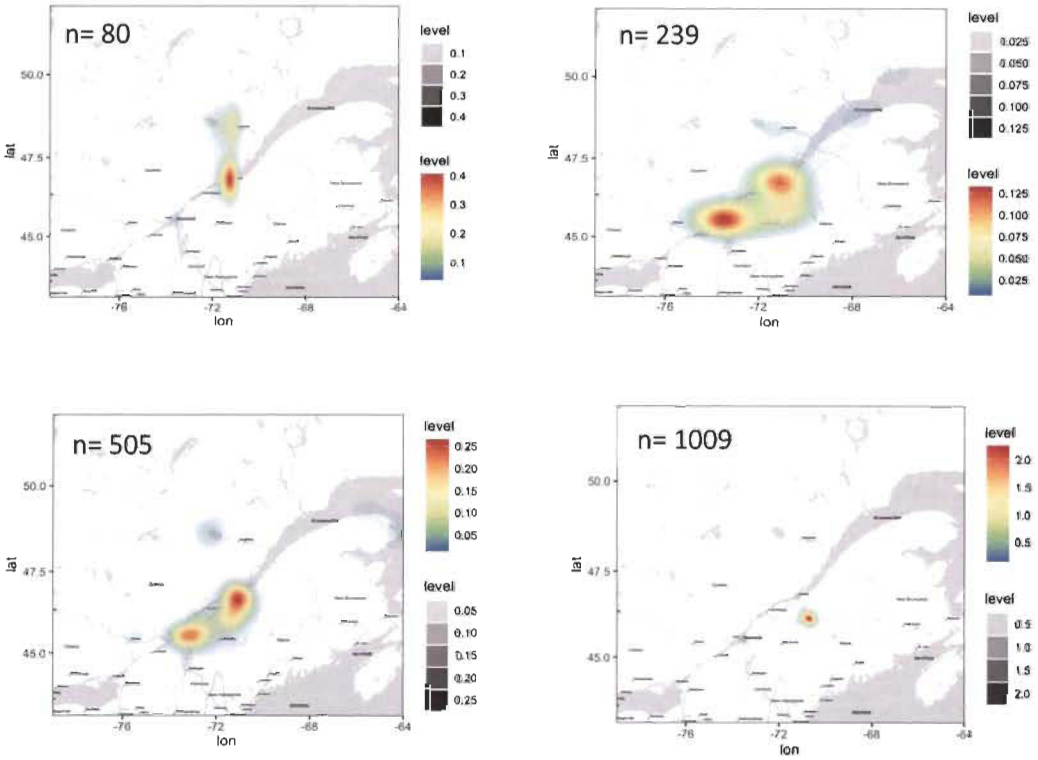

Figure 4.2 Répartition géographique de quatre haplotypes tirés au hasard des simulations. Ils sont portés respectivement par 80, 239, 505 et 1 009 hommes.

Cette analyse montre que certains haplotypes sont concentrés dans une région spécifique, tel que l'haplotype porté par 1 009 hommes de la **Figure 4.2**, qui se retrouve uniquement dans la région de la Beauce, alors que l'haplotype porté par 239 hommes est concentré dans les régions de Montréal et de Québec, mais il est également présent dans quelques autres régions. En conséquence, le choix de la population d'intérêt, c'est-à-dire l'ensemble des individus pouvant *a priori* être à l'origine de la trace ADN retrouvée sur une scène de crime, et donc des potentiels suspects, devrait varier selon le contexte géographique de l'affaire criminelle. Ce choix devrait donc être un élément clé lors du calcul de la valeur probante, puisqu'il pourrait directement influencer la valeur fournie à la cour.

4.2 Perspectives

Les avantages du modèle de simulation sont nombreux pour le domaine de la génétique forensique. En plus de permettre d'éviter les problèmes associés à l'utilisation des bases de données qui sont rarement représentatives des populations d'intérêt, ce modèle permet de fournir une valeur probante à une concordance impliquant un haplotype Y, et ce, indépendamment de l'identité réelle de cet haplotype. Ce modèle estime le nombre d'hommes portant un même haplotype dans la population plutôt qu'une valeur probante sous forme de rapport de vraisemblance, tel qu'actuellement utilisé par certains laboratoires judiciaires. Selon Andersen et Balding (2017) cette façon de faire fournit une valeur probante plus facilement comprise par les acteurs du système de justice (juges, jury, avocats, policiers, etc.) qui n'ont généralement pas de connaissances approfondies en statistique. Il serait alors intéressant de vérifier cette idée en étudiant de quelle manière ces acteurs interprètent cette valeur comparativement à celle fournie par le rapport de vraisemblance. De plus, le modèle de simulation répond au problème de déterminer un poids statistique dans le cas d'haplotypes qui sont retrouvés plus rarement dans la population (valeur estimée du 99^{e} percentile). Il n'offre toutefois pas de solution claire pour les 1% des haplotypes les plus fréquents. Or, il pourrait servir à établir combien de profils Y devraient être aléatoirement échantillonnés pour connaître l'identité de ces haplotypes plus fréquents, et idéalement leur fréquence.

Le modèle d'Andersen et Balding offre une certaine flexibilité en permettant de faire varier les valeurs de certains paramètres démographiques pour s'ajuster à la population d'intérêt. De plus, ces auteurs ont publié un second article en 2019 rapportant une étude de l'efficacité et l'applicabilité de leur méthode pour les traces d'ADN mélangées [74]. Selon les résultats, la valeur probante calculée pour une combinaison d'ADN provenant de deux individus était sensiblement la même que pour un profil pur, alors que les combinaisons à trois ou quatre contributeurs présentaient des valeurs probantes plus faibles. Le modèle s'avère être également efficace pour estimer le nombre d'hommes portant un même haplotype compatible avec une combinaison d'ADN de plusieurs contributeurs, ce qui s'avère un fort avantage puisque ces combinaisons sont

présentes dans une forte proportion des dossiers judiciaires ayant fait l'objet d'analyses génétiques.

Deux améliorations pourraient être apportées au modèle afin de s'approcher davantage de la population réelle. Premièrement, il serait bien de pouvoir spécifier le taux d'accroissement de la population. Il serait possible d'estimer ce taux pour le Québec, à différentes époques, par les données généalogiques de la population canadienne-française. Deuxièmement, avec des ajustements aux fonctions R, il serait possible d'introduire une corrélation génétique entre la valeur du succès reproducteur d'un père et celle de son fils, puisque le succès reproducteur est héritable, c'est-à-dire qu'un individu plus fécond aura tendance à engendrer des enfants eux-mêmes plus féconds [82].

Les données généalogiques utilisées couvrent une période de temps s'échelonnant de la fondation du Québec en 1608 jusqu'en 1960. Or, le fait qu'elles ne couvrent pas une période de temps plus récente pourrait affecter certaines de nos conclusions. Effectivement, dans les 20-30 dernières années, la population a davantage migré vers les villes où l'éducation supérieure y est plus accessible ainsi que les emplois plus nombreux. De ce fait, cela peut avoir un impact sur les résultats de l'analyse spatiale puisque ce sont seulement les trois dernières générations dans les données qui ont été utilisées pour ces analyses, alors que la période de 1960 à aujourd'hui représente environ deux générations de plus. Ainsi, il serait intéressant d'inclure les données généalogiques de 1960 à aujourd'hui si elles devenaient disponibles pour refaire les mêmes analyses spatiales.

Finalement, l'immigration récente a sûrement un impact sur la persistance de l'effet fondateur au Québec. En effet, bon nombre de migrants sont venus s'installer dans la province de Québec apportant donc avec eux leur bagage génétique. Il est plausible que ces nouvelles arrivées aient eu pour impact de réduire la valeur du 99^{e} percentile dans certaines régions, puisque davantage d'haplotypes seraient portés par un moins grand nombre d'individus. Il serait donc intéressant d'étudier cet aspect pour en évaluer l'ampleur.

RÉFÉRENCES BIBLIOGRAPHIQUES

1. Balding, D.J, Steele, C.D. (2015). Weight-of-evidence for forensic DNA profiles. United Kingdom: John Wiley & Sons Ltd (213 p.).

2. Laboratoire de sciences judiciaires et de médecine légale. (2019). Rapport annuel 2018-2019. *Ministère de la sécurité publique*. Accessible à l'adresse <https://www.securitepublique.gouv.qc.ca/fileadmin/Documents/laboratoire/rapport_annuel/rapport_annuel_2018-2019.pdf.>.

3. Gill, P. (2014). Misleading DNA Evidence: Reasons for Miscarriages of Justice. London: Elviser Inc.

4. Ribaux, O., Margot, P. Dictionnaire de Criminologie en ligne. Accessible à l'adresse <http://criminologie.com>.

5. Houck, M.M., Siegel, J.A. (2015). Fundamentals of forensic science (3e éd.). San Diego, CA: Academic Press (703 p.).

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

7. Arenas, M., Pereira, F., Oliveira, M., Pinto, N., Lopes, A.M., Gomes, V., *et al.* (2017). Forensic genetics and genomics: Much more than just a human affair. *PLoS genetics*. 13(9):e1006960.

8. Yamamoto, F., and Hakomori, S. (1990) Sugar-nucleotide donor specificity of histo-blood group A and B transferases is based on amino acid substitutions. *The Journal of biological chemistry*. 265(31):19257-62.

9. Maronas, O., Sochtig, J., Ruiz, Y., Phillips, C., Carracedo, A., Lareu, M.V. (2015). The genetics of skin, hair, and eye color variation and its relevance to forensic pigmentation predictive tests. *Forensic science review*. 27(1):13-40.

10. Siegel, J.A., Saukko, P.J., Houck, M.M. (2013). Encyclopedia of forensic sciences (2nd éd.). London, UK: Elsevier, Academic Press.

11. Reich, D.E., Schaffner, S.F., Daly, M.J., McVean, G., Mullikin, J.C., Higgins, J.M., *et al.* (2002). Human genome sequence variation and the influence of gene history, mutation and recombination. *Nature genetics*. 32(1):135-42.

12. Jeffreys, A.J., Wilson, V., Thein, S.L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature*. 314(6006):67-73.

13. Bar, W., and Hummel, K. (1991). DNA fingerprinting: its application in forensic case work. *Experientia Supplementum*. 58:349-55.

14. Jeffreys, A.J., Wilson, V., Thein, S.L. (1985). Individual-specific 'fingerprints' of human DNA. *Nature*. 316(6023):76-9.

15. Gill, P., Werrett, D.J. (1987) Exclusion of a man charged with murder by DNA fingerprinting. *Forensic science international*. 35(2-3):145-8.

16. Mullis, K., Faloona, F., Scharf, S., Saiki, R., Horn, G., Erlich, H. (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. *Cold Spring Harbor symposia on quantitative biology*. 51:263-73.

17. Mullis, K.B. (1990). The unusual origin of the polymerase chain reaction. *Scientific American*. 262(4):56-61, 4-5.

18. Budowle, B., Planz, J.V., Campbell, R.S., Eisenberg, A.J. (2004). Single Nucleotide Polymorphisms and Microarray Technology in Forensic Genetics - Development and Application to Mitochondrial DNA. *Forensic science review*. 16(1):21-36.

19. Butler, J.M. (2014). Advanced topics in forensic dna typing : interpretation. Boston, MA: Elsevier.

20. Walsh, S.J., (2014). DNA. A guide to forensic DNA profiling. Chichester, West Sussex, United Kingdom: Wiley (p.29-36).

21. Tilford, C.A., Kuroda-Kawaguchi, T., Skaletsky, H., Rozen, S., Brown, L.G., Rosenberg, M., *et al.* (2001). A physical map of the human Y chromosome. *Nature*. 409(6822):943-5.

22. Anjos, M.J., Carvalho, M., Andrade, L., Lopes, V., Serra, A., Batista, L., *et al.* (2004). Individual genetic identification of biological samples: a case of an aircraft accident. *Forensic science international*. 146 Suppl: (S115-7).

23. Jobling, M.A., Pandya, A., Tyler-Smith, C. (1997). The Y chromosome in forensic analysis and paternity testing. *International journal of legal medicine.* 110(3):118-24.

24. Cerri, N., Ricci, U., Sani, I., Verzeletti, A., De Ferrari, F. (2003). Mixed stains from sexual assault cases: autosomal or Y-chromosome short tandem repeats? *Croatian medical journal.* 44(3):289-92.

25. Gill, P., Jeffreys, A.J., Werrett, D.J. (1985) Forensic application of DNA 'fingerprints'. *Nature.* 318(6046):577-9

26. Coquoz, R., Comte, J., Hall, D., Hicks, T., Taroni, F. (2013). Preuve par l'ADN: la génétique au service de la justice (3e éd.). Lausanne : Presses polytechnique et universitaires romandes.

27. Giusti, W.G., Adriano, T. (1993). Synthesis and characterization of 5'-fluorescent-dye-labeled oligonucleotides. *PCR methods and applications.* 2(3):223-7.

28. Butler, J.M. (2012). Advanced topics in forensic DNA typing : methodology. Walthan, MA: Elsevier/Academic Press (680 p.).

29. Lygo, J.E., Johnson, P.E., Holdaway, D.J., Woodroffe, S., Whitaker, J.P., Clayton, T.M., *et al.* (1994). The validation of short tandem repeat (STR) loci for use in forensic casework. *International journal of legal medicine.* 107(2):77-89.

30. Walsh, S.J. (2014). Databases. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling.* Chichester, West Sussex, United Kingdom: Wiley (p.177-84).

31. Hares, D.R., Selection and implementation of expanded CODIS core loci in the United States. *Forensic science international Genetics.* 17:33-4.

32. Karantzali, E., Rosmaraki, P., Kotsakis, A., Le Roux-Le Pajolec, M.G., Fitsialos, G. (2019). The effect of FBI CODIS Core STR Loci expansion on familial DNA database searching. *Forensic science international: Genetic*s. 43:102129.

33. Ballantyne, K.N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., *et al.* (2010). Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *American Journal of Human Genetics.* 87(3):341-53.

34. Goedbloed, M., Vermeulen, M., Fang, R.N., Lembring, M., Wollstein, A., Ballantyne, K., *et al.* (2009). Comprehensive mutation analysis of 17 Y-chromosomal short tandem repeat polymorphisms included in the AmpFlSTR Yfiler PCR amplification kit. *International Journal of Legal Medicine.* 123(6):471-82.

35. Tozzo, P., Ponzano, E., Spigarolo, G., Nespeca, P., Caenazzo, L. (2018). Collecting sexual assault history and forensic evidence from adult women in the emergency department: a retrospective study. *BMC Health Services Research.* 18(1):383.

36. Gingras, F., Paquet, C., Bazinet, M., Granger, D., Marcoux-Legault, K., Fiorillo, M., *et al.* (2009). Biological and DNA evidence in 1000 sexual assault cases. *Forensic Science International.* 2:138-40.

37. Kayser, M. (2017). Forensic use of Y-chromosome DNA: a general overview. *Human Genetics.*136(5):621-35.

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

39. Roewer, L., Arnemann, J., Spurr, N.K., Grzeschik, K.H., Epplen, J.T. (1992). Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Human Genetics.* 89(4):389-94.

40. Roewer, L., Epplen, J.T. (1992). Rapid and sensitive typing of forensic stains by PCR amplification of polymorphic simple repeat sequences in case work. *Forensic Science International.* 53(2):163-71.

41. Gopinath, S., Zhong, C., Nguyen, V., Ge, J., Lagace, R.E., Short, M.L., et al. (2016). Developmental validation of the Yfiler (®) Plus PCR Amplification Kit: An enhanced Y-STR multiplex for casework and database applications. *Forensic Science International: Genetics.* 24:164-75.

42. Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., *et al.* (1997) Evaluation of Y-chromosomal STRs: a multicenter study. *International Journal of Legal Medicine.* 110(3):125-33, 41-9.

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

44. Ballantyne, J., Hanson, K.E. (2014). Y-Chromosome Short Tandem Repeats. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling*. Chichester, West Sussex, United Kingdom: Wiley (p.149-54).

45. Krenke, B.E., Viculis, L., Richard, M.L., Prinz, M., Milne, S.C., Ladd, C., *et al.* (2005). Validation of male-specific, 12-locus fluorescent short tandem repeat (STR) multiplex. *Forensic Science International*. 151(1):111-24.

46. Champod, C. (2014). Identification and Individualization. Dans Jamieson, A. et Bader, S. (sous la direction de), *A Guide to Forensic DNA Profiling*. Chichester, West Sussex, United Kingdom: Wiley (p.69-78).

47. National Research Council (U.S.). (1996). The Evaluation of Forensic DNA Evidence. Washington, DC: National Academy Press.

48. Graffelman, J., Jain, D., Weir, B. (2017). A genome-wide study of Hardy-Weinberg equilibrium with next generation sequence data. *Human Genetics*. 136(6):727-41.

49. Chen, B., Cole, J.W., Grond-Ginsbach, C. (2017). Departure from Hardy Weinberg Equilibrium and Genotyping Error. *Frontiers in Genetics*. 8:167.

50. Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*. 15(4):323-54.

51. Balding, D.J., Nichols, R.A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. *Forensic Science International*. 64(2-3):125-40.

52. Alladio, E., Omedei, M., Cisana, S., D'Amico, G., Caneparo, D., Vincenti, M., *et al.* (2018) DNA mixtures interpretation - A proof-of-concept multi-software comparison highlighting different probabilistic methods' performances on challenging samples. *Forensic Science International: Genetics*. 37:143-50.

53. Noel, S., Noel, J., Granger, D., Lefebvre, J.-F., Seguin, D. (2019). STRmix() put to the test: 300 000 non-contributor profiles compared to four-contributor DNA mixtures and the impact of replicates. *Forensic Science International: Genetics*. 41:24-31.

54. Buckleton, J.S., Bright, J.A., Gittelson, S., Moretti, T.R., Onorato, A.J., Bieber, F.R., *et al.* (2019). The Probabilistic Genotyping Software STRmix: Utility and Evidence for its Validity. *Journal of Forensic Sciences*.64(2):393-405.

55. SWGDAM. (2014). Interpretation Guidelines for Y-Chromosome STR Typing. Accessible à l'adresse <https://www.thermofisher.com/content/dam/LifeTech/Documents/PDFs/SWGDAM_YSTR_Guidelines_APPROVED_01092014_v_02112014_FINAL.pdf>.

56. Brenner, C.H. (2010). Fundamental problem of forensic mathematics-the evidential value of a rare haplotype. *Forensic Science International: Genetics*. 4(5):281-91.

57. Willuweit, S., Roewer, L. (2015). The new Y Chromosome Haplotype Reference Database. *Forensic Science International: Genetics*. 15:43-8.

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

59. 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. *Mémoire présenté à l'Université du Québec à Trois-Rivières*.

60. Brenner, C.H. (2014). Understanding Y haplotype matching probability. *Forensic Science International: Genetics*. 8(1):233-43.

61. De Braekeleer, M., Dao, T.N. (1994) Hereditary disorders in the French Canadian population of Quebec. I. In search of founders. *Human Biology*. 66(2):205-23.

62. s.a. BALSAC – Fichier des populations. Accessible à l'adresse <http://balsac.uqac.ca.>

63. De Braekeleer, M., Giasson, F., Mathieu, J., Roy, M., Bouchard, J.P., Morgan, K. (1993) Genetic epidemiology of autosomal recessive spastic ataxia of Charlevoix-Saguenay in northeastern Quebec. *Genetic Epidemiology*. 10(1):17-25.

64. De Braekeleer, M., Perusse, L., Cantin, L., Bouchard, J.M., Mathieu, J. (1996) A study of inbreeding and kinship in intracranial aneurysms in the Saguenay Lac-Saint-Jean region (Quebec, Canada). *Annals of Human Genetics*. 60(2):99-104.

65. Dao, T.N., Mathieu, J., Bouchard, J.P., De Braekeleer, M. (1993). Infant mortality in myotonic dystrophy in Saguenay-Lac-St-Jean: a historical perspective. *Clinical Genetics*. 43(1):25-7.

66. Scriver, C.R. (2001). Human genetics: lessons from Quebec populations. *Annual Review of Genomics and Human Genetics*. 2:69-101.

67. Doyon, A., Moreau, C., Labuda, D., Milot, E. (Submitted). Spatiotemporal variation of mitochondrial DNA and Y chromosome haplotype frequencies in a French-Canadian population and its impact on random match probabilities.

68. Roy-Gagnon, M.H., Moreau, C., Bherer, C., St-Onge, P., Sinnett, D., Laprise, C., *et al.* (2011). Genomic and genealogical investigation of the French Canadian founder population structure. *Human Genetics*. 129(5):521-31.

69. R Development Core Team. (2019). A language and environment for statistical computing. Viennna, Austria.

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

71. Cockerton, S., McManus, K., Buckleton, J. (2012). Interpreting lineage markers in view of subpopulation effects. *Forensic Science International: Genetics*. 6(3):393-7.

72. Moreau, C., Lefebvre, J.-F., Jomphe, M., Bherer, C., Ruiz-Linares, A., Vezina, H., *et al.* (2013). Native American admixture in the Quebec founder population. *PloS One*. 8(6):e65507.

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

74. Andersen, M.M., Balding, D.J. (2019). Y-profile evidence: Close paternal relatives and mixtures. *Forensic Science International: Genetics*. 38:48-53.

75. Gagnon, A., Heyer, E. (2001). Fragmentation of the Quebec population genetic pool (Canada): evidence from the genetic contribution of founders per region in the 17th and 18th centuries. *American Journal of Physical Anthropology*. 114(1):30-41.

76. Vezina, H., Tremblay, M., Desjardins, B., Houde, L. (2006). Origines et contributions génétiques des fondatrices et des fondateurs de la population québécoise. *Cahiers québécois de démographie*. 34(2):235-50.

77. Turrina, S., Caratti, S., Ferrian, M., De Leo, D. (2016). Are rapidly mutating Y-short tandem repeats useful to resolve a lineage? Expanding mutability data on distant male relationships. *Transfusion*. 56(2):533-8.

78. Ballantyne, K.N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S.B., Ralf, A., *et al.* (2012). A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Science International: Genetics*. 6(2):208-18.

79. Szabolcsi, Z., Farkas, Z., Borbely, A., Barany, G., Varga, D., Heinrich, A., et al. (2015). Statistical and population genetics issues of two Hungarian datasets from the aspect of DNA evidence interpretation. *Forensic Science International: Genetics*. 19:18-21.

80. Templeton, A.R. (2006). Population genetics and microevolutionary theory. Hoboken, N.J.: Wiley-Liss (705 p.).

81. Bherer, C., Labuda, D., Roy-Gagnon, M.H., Houde, L., Tremblay, M., Vezina, H. (2011). Admixed ancestry and stratification of Quebec regional populations. *American Journal of Physical Anthropology*. 144(3):432-41.

82. Heyer, E., Brandenburg, J., Leonardi, M., Toupance, B., Balaresque, P., Hegay, T., et al. (2015). Patrilineal populations show more male transmission of reproductive success than cognatic populations in Central Asia, which reduces their genetic diversity. *American Journal of Physical Anthropology*. 157(4):537-43.

ANNEXE A

CODE R PERMETTANT D'ARRANGER LES DONNÉES GÉNÉALOGIQUES

```
library(tidyverse)

library(pedinf) #install from github (https://github.com/mikldk/pedinf)

gen_quebec <- read.table("~/Documents/Maitrise - UQTR/Donnees/Genealogical_data/BALSAC/BAL4M_sansPRDH.txt", fill = TRUE, header = TRUE, sep = ",")

gen_quebec_males <- subset(gen_quebec, sex == 1)


# gen_quebec: Load in Quebec data

# gen_quebec_males <- ... Only take out males, e.g. gen_quebec_males <- subset(gen_quebec, sexe == 1)


pop_males <- with(gen_quebec_males,

                  load_individuals(pid = ind,

                pid_mom = rep(0L, nrow(gen_quebec_males)), # No mothers

                  pid_dad = father,

                is_male = rep(TRUE, nrow(gen_quebec_males)), # No mothers

                  error_on_pid_not_found = TRUE,progress = TRUE))


peds_males <- build_pedigrees_recursive(pop_males, progress = TRUE) #male pedigree construction
```

```
gen_quebec_males_tt <- gen_quebec_males %>%
  mutate(paternalped_id = get_pedigree_id_from_pid(pop_males, ind), #Attribution of pe
digree identity number
  paternalped_size = get_pedigree_size_from_pid(pop_males, ind))   #Calculation of ped
igree size

gen_quebec_males_tt %>%
  group_by(paternalped_id) %>%
  summarise(n = n()) %>%          #Get the size of every paternal pedigree
  arrange(desc(n))

### Get generation of all individual in genealogy ####

library(malan) #install from github (https://github.com/mikldk/malan)
pop <- load_individuals(pid = gen_quebec_males_tt$ind, pid_dad = gen_quebec_males_
tt$father)
pedigrees <- build_pedigrees(pop)

inds <- get_individuals(population = pop) %>% unlist()
n <- length(inds)
gen_inds <- integer(n)
m <- length(inds)
inds_pid <- integer(m)
```

```
for (j in seq_along(inds)) {
  a <- get_generation(inds[[j]])
  gen_inds[j] <- a
  b <- get_pid(inds[[j]])
  inds_pid[j] <- b
}

gen_pid_ind <- cbind(gen_inds, inds_pid)

qc_ped_gen_bd <- merge(x= gen_quebec_males_tt, y= gen_pid_ind[, c("inds_pid", "gen
_inds")], by.x = "ind", by.y = "inds_pid", all.x = TRUE) %>% as_tibble()

save(qc_ped_gen_bd, file = "qc_ped_gen_db.Rdata")  #Final Files with all information
```

ANNEXE B

CODE R POUR LES SIMULATIONS SANS LES DONNÉES GÉNÉALOGIQUES

```
######################
#Load packages
######################
library(dplyr)
library(malan)
library(tibble)

########################
#Database creation
########################

ystr_profil <- ystr_kits %>%
  filter(Kit == "Yfiler") %>%
  inner_join(ystr_markers, by = "Marker") %>%
  rowwise() %>%
  mutate(IntegerAlleles = list(Alleles[Alleles == round(Alleles)]), MinIntAllele = min(In
tegerAlleles), MaxIntAllele = max(IntegerAlleles)) %>%
  ungroup() %>%
  select(-Kit, -Alleles)
```

```
mu <- ystr_profil %>% pull(MutProb)


min_bound <- ystr_profil$MinIntAllele

max_bound <- ystr_profil$MaxIntAllele


generate_random_haplotype <- function() {

  ystr_profil %>%

    rowwise() %>%

    mutate(Allele = IntegerAlleles[sample.int(length(IntegerAlleles), 1)]) %>%

    pull(Allele) }


###################

#Loop elements

###################


N <- 1189672

dirichlet_alpha <- 1000000  #VRS of 0.98#


num_repeat_pop <- 3

num_repeat_hap <- 2

l=0

num_repeat_tot <- num_repeat_hap*num_repeat_pop
```

```
quant_table_founder_hap <- vector("list", num_repeat_tot)
haps_id_table_founder <- vector("list", num_repeat_tot)
table_founders_livepop <- vector("list", num_repeat_pop)


##############################################
# Loop population and haplotype simulations
##############################################

for (i in seq_len(num_repeat_pop)) {

sim_res <- sample_geneology(population_size = N,
                            generations = 15,
                            enable_gamma_variance_extension = TRUE,
                            gamma_parameter_shape = dirichlet_alpha,
                            gamma_parameter_scale = 1/dirichlet_alpha,
                           progress = FALSE, verbose_result = TRUE)

ped_pop <- build_pedigrees(sim_res$population)

plot(ped_pop)

live_individuals <- sim_res$individuals_generations
```

```
t <- length(live_individuals)

live_ind_pid_na <- sim_res$individual_pids[, 1:3] %>% as.vector()

live_ind_pid <- live_ind_pid_na[!is.na(live_ind_pid_na)] %>% sort()

founder_number <- sim_res$founders

table <- c(founder_number, t)

table_founders_livepop[[i]] <- table

for( j in seq_len(num_repeat_hap)) {

l=l+1

 pedigrees_all_populate_haplotypes_custom_founders(pedigrees = ped_pop,

                                    mutation_rates=mu,                    get_founder_haplotype
        = generate_random_haplotype,

                    prob_genealogical_error = 0.008,

                    progress = FALSE)

hashmap <- build_haplotype_hashmap(live_individuals, progress = FALSE)

hap_ids <- haplotypes_to_hashes(population = sim_res$population, pids = live_ind_pid
)

hap_ids_tab <- table(hap_ids) %>% enframe()

haps_id_table_founder[[l]] <- hap_ids_tab
```

```
q <- quantile(hap_ids_tab$value, c(0.5, 0.99))
quant_df <- data.frame(q)
quant_table_founder_hap[[l]] <- data.frame(q)

}
 }
save(haps_id_table_founder, file= "haps_id_table_founder.Rdata")
save(quant_table_founder_hap, file= "quant_table_founder_hap.Rdata")
save(table_founders_livepop, file = "table_founders_livepop.Rdata")



############################################################
#Graphs of proportion of men carrying a common haplotype
############################################################


load("~/Quebec_malepedigree/Hap_quebec_sim/haps_id_table_founder.Rdata")
library(purrr)

hap_list <- as.integer(as.data.frame(haps_id_table_founder[[1]])[,2])

hap_list_more100 <- keep(hap_list, function(x) x>100)

hist(hap_list, breaks=50, col = "grey", main = NULL, ylim = c(0,200000),
    xlab = "Number of time an haplotype is observed", ylab = "Number of observation")
```

```
hist(hap_list, ylim= c(0,100), xlim = c(100,200), col = "blue", main = NULL,
    xlab = "Number of time an haplotype is observed", ylab = "Number of observation")


hist_tot <- hist(hap_list)


hist_tot$density = hist_tot$counts/sum(hist_tot$counts)*100
plot(hist_tot, freq=FALSE, xlab = "Estimated number of men carriers", ylab = "Proporti
on of haplotypes (%)",
    main = NULL, col = "grey", ylim = c(0,100))


plot(hist_tot, freq=FALSE, xlim = c(50, 200), xlab = "Estimated number of men carriers
", ylab = "Proportion of haplotypes (%)",
    main = NULL, col = "blue", ylim = c(0, 1))
```

ANNEXE C

CODE R POUR LES SIMULATIONS UTILISANT LES DONNÉES GÉNÉALOGIQUES

```
###Subsetting the genealogical data (pedigree with 10 generations or more) ####

library(tibble)
library(malan)
library(dplyr)

load("~/Quebec_malepedigree/qc_ped_gen_db.Rdata")
quebec_ped_gen_more10 <-  qc_ped_gen_bd %>% as_tibble() %>% filter(dist >= 10)

###Load the data and create the pedigree###

#load("~/Quebec_malepedigree/y23_2014-05-23.RData")
pop <- load_individuals(pid = quebec_ped_gen_more10$ind, pid_dad =
quebec_ped_gen_more10$father)

pedigrees <- build_pedigrees(pop)

### Create living population ###

pids <- quebec_ped_gen_more10 %>% filter(gen_inds < 3) %>% pull(ind)
live_pop <- lapply(pids, function(pid) get_individual(pop, pid))
```

```
#mu_db_PP23 <- c(0.0022374145, 0.0029290746, 0.0041229909, 0.0021137586,
0.0024538293, 0.0005188740, 0.0010515984, 0.0003747775,
#            0.0054475439, 0.0012204281, 0.0015216489, 0.0042882833, 0.0063641395,
0.0043338286, 0.0030266344, 0.0147194112, 0.0050205386,
 #           0.0037541061, 0.0036747818, 0.0133475707, 0.0013513514)



################################################
#Attribute a random haplotype to a founder and the lineage and select a random individual
#and the frequencies of its haplotype
##########################################

#### To put bounderies to mutation ######

d_tmp <- ystr_kits %>%
        filter(Kit == "PowerPlex Y23") %>%
       inner_join(ystr_markers, by = "Marker") %>%
        rowwise()  %>%
        mutate(IntegerAlleles = list(Alleles[Alleles == round(Alleles)]), MinIntAllele =
min(IntegerAlleles), MaxIntAllele = max(IntegerAlleles)) %>%
        ungroup() %>%
       select(-Kit, -Alleles)

l_min <- d_tmp %>% pull(MinIntAllele)
l_max <- d_tmp %>% pull(MaxIntAllele)

generate_random_haplotype <- function() {
        d_tmp %>%
          rowwise() %>%
          mutate(Allele = IntegerAlleles[sample.int(length(IntegerAlleles), 1)]) %>%
```

```
        pull(Allele) }

mu_pp23 <- d_tmp %>% pull(MutProb)

###################################
#Simulation
###################################

num_repeat <- 100
quant_table_eachfond_more10gen <- vector("list", num_repeat)
hap_ids_tab_eachfond <- vector("list", num_repeat)

for(i in seq_len(num_repeat)) {

        pedigrees_all_populate_haplotypes_ladder_bounded(
          pedigrees = pedigrees,
          get_founder_haplotype = generate_random_haplotype,
          mutation_rates = mu_pp23, prob_genealogical_error = 0.008,
          ladder_min = l_min, ladder_max = l_max,
          progress = FALSE)

        hashmap <- build_haplotype_hashmap(live_pop, progress = FALSE)
        hap_ids <- haplotypes_to_hashes(population = pop, pids = pids)
        hap_ids_tab <- table(hap_ids) %>% enframe()

        hap_ids_tab_eachfond[[i]] <- hap_ids_tab

        q <- quantile(hap_ids_tab$value, c(0.5, 0.99))
        quant_df <- data.frame(q)
        quant_table_eachfond_more10gen[[i]] <- quant_df
```

```
    delete_haplotypeids_hashmap(hashmap)
  }

quant_table_list_eachfond_more10gen <- quant_table_eachfond_more10gen %>%
bind_rows() %>% as_tibble

mean_quant_eachfond_more10gen <- quant_table_list_eachfond_more10gen %>%
group_by(Var1) %>% summarise(mean(Freq))


save(hap_ids_tab_eachfond, file= "hap_ids_tab_eachfond.Rdata")
save(quant_table_list_eachfond_more10gen,file="quant_table_list_eachfond_more10ge
n.Rdata")

##############################
#Results analysis
##############################
library(purrr)
library(tidyverse)

######################
#Quantile analysis
######################

load("~/Quebec_malepedigree/Hap_freq_10gen/quant_table_list_eachfond_more10gen.
Rdata")

mean_quant_eachfond <- quant_table_list_eachfond_more10gen %>% group_by(Var1)
%>% summarise(mean(Freq))
```

```
############################
#Distribution d'haplotypes
############################

load("~/Quebec_malepedigree/Hap_freq_10gen/hap_ids_tab_eachfond.Rdata")

hapFreq_eachfond <- unlist(lapply(1:length(hap_ids_tab_eachfond), function(i){
        as.integer(as.data.frame(hap_ids_tab_eachfond[[i]])[,2])
       }))



z <- keep(hapFreq_eachfond, function(x) x > 1000)
y <- keep(hapFreq_eachfond, function(x) x < 197)
x <- keep(hapFreq_eachfond, function(x) x == 197)
zz <- keep(hapFreq_eachfond, function(x) x > 197)

hist_tot_eachfond <- hist(hapFreq_eachfond)
hist_tot_eachfond$density                                                    =
hist_tot_eachfond$counts/sum(hist_tot_eachfond$counts)*100

plot(hist_tot_eachfond, freq=FALSE, xlab = "Simulated number of men carriers", ylab =
"Proportion of haplotypes (%)", main = NULL, col = "grey", ylim = c(0,100), xlim = c(0,
8000), las = 1)

hist_tot_quant99 <- hist(z)
hist_tot_quant99$density                                                     =
hist_tot_quant99$counts/sum(hist_tot_eachfond$counts)*100

plot(hist_tot_quant99, freq=FALSE, xlab = "Simulated number of men carriers", ylab =
"Proportion of haplotypes (%)",  main = NULL, col = "blue", ylim = c(0, 0.5), xlim =
c(1000,8000), angle = 90, las = 1)
```

```
hist_tot_quant1 <- hist(y)
hist_tot_quant1$density = hist_tot_quant1$counts/sum(hist_tot_eachfond$counts)*100

plot(hist_tot_quant1, freq=FALSE, xlab = "Simulated number of men carriers", ylab =
"Proportion of haplotypes (%)",  main = NULL, col = "red", ylim = c(0, 100), xlim =
c(0,200), angle = 180, las = 1)
```

ANNEXE D

CODE R POUR L'ANALYSE SPATIALE DES HAPLOTYPES Y

```
######################
#Load packages
######################

library(malan)
library(dplyr)
library(tibble)
library(stringr)
library(readxl)

################################################
#Create a population and living individuals
################################################

load("~/Quebec_malepedigree/Spatial_analysis/data_qc_region.Rdata")
load("~/Quebec_malepedigree/qc_ped_gen_db.Rdata")

quebec_ped_gen_more10 <-  qc_ped_gen_bd %>% as_tibble() %>% filter(dist >= 10)

pop <- load_individuals(pid = quebec_ped_gen_more10$ind, pid_dad =
quebec_ped_gen_more10$father)
pedigrees <- build_pedigrees(pop)
```

```
        pids <- quebec_ped_gen_more10 %>% filter(gen_inds < 3) %>% pull(ind)
        live_pop <- lapply(pids, function(pid) get_individual(pop, pid))

####################################
#Region de marriage de la living pop
####################################

pid_live_pop <- lapply(live_pop, function(live_pop) get_pid(live_pop))

pop_by_region <- vector("double", length(pid_live_pop))

for(i in seq_along(pid_live_pop)){

pop_region <- data_qc_region %>% filter(ind %in% pid_live_pop[[i]]) %>%
select("Région")
pop_by_region[[i]] <- pop_region


        }

living_pop_region <- unlist(pop_by_region) %>% table(useNA = "always") %>%
as.data.frame()

####################################
#Mutations rates and bounderies
####################################

load("~/Quebec_malepedigree/y23_2014-05-23.RData")

mu_db_PP23 <- c( 0.0022374145, 0.0029290746, 0.0041229909, 0.0021137586,
0.0024538293, 0.0005188740, 0.0010515984, 0.0003747775, 0.0054475439,
```

```
0.0012204281, 0.0015216489, 0.0042882833, 0.0063641395, 0.0043338286,
0.0030266344, 0.0147194112, 0.0050205386, 0.0037541061, 0.0036747818,
0.0133475707, 0.0013513514)

min_bound <- c(9,9,9,17,5,6,7,5,5,10,13,9,10,13,6,10,16,5,7,10,4)
max_bound <- c(20,17,36,30,16,20,18,19,19,19,25,23,24,30,20,25,34,16,15,26,17)

###################################
#Attribution of haplotypes
###################################

all_haps <- y23$haplotypes$integer_alleles$db

dbqc_hap <- function() {
        all_haps[sample(nrow(all_haps), 1), ]
       }

pedigrees_all_populate_haplotypes_ladder_bounded(
        pedigrees = pedigrees,
        get_founder_haplotype = dbqc_hap,
        mutation_rates = mu_db_PP23, ladder_min = min_bound, ladder_max =
max_bound,
        prob_genealogical_error = 0.008, progress = FALSE)


###################################
#Analyse spatiale des haplotypes
###################################

long_lat <- read_excel("~/Documents/Maitrise -
UQTR/Donnees/Genealogical_data/BALSAC/URBLonLat_rox.xlsx")
```

```
region_qc <- read_excel("~/Documents/Maitrise - UQTR/Donnees/Genealogical_data/BALSAC/URBlist_rox.xlsx")

data_qc_region <- merge(x = quebec_ped_gen_more10, y=region_qc[,c("No", "Région")] , by.x = "lieum", by.y = "No", all.x = TRUE)

qc_data_reg_long_lat <- merge(x= data_qc_region, y=long_lat, by.x = "lieum", by.y = "Localities", all.x = TRUE)

all_haplotypes <- get_haplotypes_individuals(live_pop)
all_unic_hap <- unique(all_haplotypes)
subset_hap <- split(all_unic_hap, 1:nrow(all_unic_hap))

hashmap <- build_haplotype_hashmap(live_pop, progress = FALSE)
match_pid <- lapply(subset_hap, function(subset_hap) get_matching_pids_from_hashmap(hashmap, subset_hap))

region_match_map <- vector("list", length(match_pid))

for(i in seq_along(match_pid)) {
    ind_matching <- qc_data_reg_long_lat %>% filter(ind %in% match_pid[[i]]) %>% select("Région", "Longitude", "Latitude")
    region_match_map[[i]] <- ind_matching
  }

region_match <- vector("list", length(match_pid))
  for(i in seq_along(match_pid)) {
    ind_match <- qc_data_reg_long_lat %>% filter(ind %in% match_pid[[i]]) %>% select("Région")
    region_match[[i]] <- ind_match
  }
```

```
region_match_table <- vector("list", length(region_match))

for(i in seq_along(region_match)){
        match_region <- table(region_match[[i]], useNA = "always") %>%
as.data.frame()
        region_match_table[[i]] <- match_region
       }

#########################
#Ajout Fréquence totale
#########################

region_freq_tot <- vector("list", length(region_match_table))

for(i in seq_along(region_match_table)) {

        a <- region_match_table[[i]] %>% mutate(Num_hap = sum(Freq)) %>%
mutate(Freq_reg = Freq/sum(Freq)) %>% mutate(Freq_qc = Num_hap/523835)
        region_freq_tot[[i]] <- a
       }

#####################################
#Selection of haplotypes in Saguenay
#####################################

saguenay_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
sag <- region_freq_tot[[i]] %>%  filter( Var1 == "SAGUENAY (LAC ST JEAN)")
```

```
        saguenay_match[[i]] <- data.frame(sag)

        }

pop_saguenay <- living_pop_region %>% filter(. == "SAGUENAY (LAC ST JEAN)")

########################################
#Selection of haplotypes in Bois-Francs
########################################

boisfrancs_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
bf <- region_freq_tot[[i]] %>% filter(Var1 == "BOIS FRANCS")
        boisfrancs_match[[i]] <- data.frame(bf)
         }

pop_boisfrancs <- living_pop_region %>% filter(. == "BOIS FRANCS")

#############################################
#Selection of haplotypes in BAS SAINT LAURENT
#############################################

bsl_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
         bsl <- region_freq_tot[[i]] %>% filter(Var1 == "BAS SAINT LAURENT")
         bsl_match[[i]] <- data.frame(bsl)

        }
```

```
pop_bsl <- living_pop_region %>%  filter(. == "BAS SAINT LAURENT")

###################################
#Selection of haplotype in BEAUCE
###################################

beauce_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        beauce <- region_freq_tot[[i]] %>% filter(Var1 == "BEAUCE")
        beauce_match[[i]] <- data.frame(beauce)
       }

pop_beauce <- living_pop_region %>% filter(. == "BEAUCE")

###################################
#Selection of haplotype in Abiti
###################################

abiti_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        abiti <- region_freq_tot[[i]] %>% filter(Var1 == "ABITI")
        abiti_match[[i]] <- abiti
       }

pop_abiti <- living_pop_region %>% filter(. == "ABITI")

#####################################
#Selection of haplotype in Charlevoix
#####################################
```

```
charlevoix_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        charlevoix <- region_freq_tot[[i]] %>% filter(Var1 == "CHARLEVOIX")
        charlevoix_match[[i]] <- charlevoix
       }

pop_charlevoix <- living_pop_region %>% filter(. == "CHARLEVOIX")

##############################################
#Selection of haplotype in COTE DE BEAUPRE
##############################################

cdb_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        cdb <- region_freq_tot[[i]] %>% filter(Var1 == "COTE DE BEAUPRE")
        cdb_match[[i]] <- cdb
       }

pop_cdb <- living_pop_region %>% filter( . == "COTE DE BEAUPRE")

########################################
#Selection of haplotype in Cote du Sud
########################################

cds_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        cds <- region_freq_tot[[i]] %>% filter(Var1 == "COTE DU SUD")
```

```
        cds_match[[i]] <- cds
    }

pop_cds <- living_pop_region %>%  filter(. == "COTE DU SUD")

###################################
#Selection of haplotype Cote Nord
###################################

cn_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        cn <- region_freq_tot[[i]] %>% filter(Var1 == "COTE NORD")
        cn_match[[i]] <- cn

    }

pop_cn <- living_pop_region %>%  filter(. == "COTE NORD")

###################################
#Selection of haplotype in Estrie
###################################

estrie_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        estrie <- region_freq_tot[[i]] %>% filter(Var1 == "ESTRIE")
        estrie_match[[i]] <- estrie

    }
```

```
pop_estrie <- living_pop_region %>% filter( . == "ESTRIE")


#######################################
#Selection of haplotypes in Gaspesie
#######################################

gaspesie_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        gaspesie <- region_freq_tot[[i]] %>% filter(Var1 == "GASPESIE")
        gaspesie_match[[i]] <- gaspesie

      }

pop_gaspesie <- living_pop_region %>% filter( . == "GASPESIE")

#######################################
#Selection of haplotypes in Lanaudiere
#######################################

lanaudiere_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        lanaudiere <- region_freq_tot[[i]] %>% filter(Var1 == "LANAUDIERE")
        lanaudiere_match[[i]] <- lanaudiere
      }

pop_lanaudiere <- living_pop_region %>% filter(. == "LANAUDIERE")
```

```
##########################################
#Selection of haplotypes in Ile de Montreal
##########################################

mtl_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
        mtl <- region_freq_tot[[i]] %>% filter(Var1 == "ILE DE MONTREAL")
        mtl_match[[i]] <- mtl
       }

pop_mtl <- living_pop_region %>% filter(. == "ILE DE MONTREAL")

########################################
#Selection of haplotypes in Mauricie
########################################

mauricie_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
       mauricie <- region_freq_tot[[i]] %>% filter(Var1 == "MAURICIE")
       mauricie_match[[i]] <- mauricie
       }

       pop_mauricie <- living_pop_region %>% filter(. == "MAURICIE")

########################################
#Selection of haplotypes in Laurentide
########################################

laurentide_match <- vector("list", length(region_freq_tot))
```

```
for(i in seq_along(region_freq_tot)) {
        laurentide <- region_freq_tot[[i]] %>% filter(Var1 == "LAURENTIDE")
        laurentide_match[[i]] <- mauricie
       }

pop_laurentide <- living_pop_region %>% filter(. == "LAURENTIDE")

########################################
#Selection of haplotypes in Outaouais
########################################

outaouais_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
 outaouais <- region_freq_tot[[i]] %>% filter(Var1 == "OUTAOUAIS")
        outaouais_match[[i]] <- outaouais
       }

pop_outaouais <- living_pop_region %>% filter(. == "OUTAOUAIS")

#####################################################
#Selection of haplotypes in Quebec(agglomération)
#####################################################

quebec_match <- vector("list", length(region_freq_tot))

       for(i in seq_along(region_freq_tot)) {
        quebec <- region_freq_tot[[i]] %>% filter(Var1 == "QUEBEC
(AGGLOMERATION)")
        quebec_match[[i]] <- quebec
```

```
}

pop_quebec <- living_pop_region %>% filter(. == "QUEBEC (AGGLOMERAITON)")

#######################################
#Selection of haplotypes in Richelieu
#######################################

richelieu_match <- vector("list", length(region_freq_tot))

for(i in seq_along(region_freq_tot)) {
  richelieu <- region_freq_tot[[i]] %>% filter(Var1 == "RICHELIEU")
  richelieu_match[[i]] <- richelieu
}

pop_richelieu <- living_pop_region %>% filter(. == "RICHELIEU")

##########################
#Maps of haplotype frequencies
##########################

library(ggmap)
library(ggplot2)
library(RColorBrewer)
library(purrr)
match_1000 <- keep(region_match_map, ~nrow(.x) >1000)      #to keep haplotype shared by more than a thousand men

map1 <- match_1000[[15]]
```

```
map_bounds <- c(left= -79, bottom = 43, top= 52, right = -64)

quebec_map <- get_stamenmap(map_bounds, zoom = 7, maptype = "toner-lite")

quebec_map <- ggmap(quebec_map, extent = "device", legend = "none")

quebec_map <- quebec_map + stat_density2d(data=map1, aes(x= Longitude, y=Latitude,
fill = ..level..,  alpha=..level..), geom = "polygon")

quebec_map <- quebec_map + scale_fill_gradientn(colours = rev(brewer.pal(7,
"Spectral")))

quebec_map <- quebec_map+theme_bw()
plot(quebec_map)
ggsave(filename = "quebec_map.png")
```