UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

CONCORDANCE ADN DANS LES DOSSIERS CRIMINELS : CONNECTER L'ÉVALUATION DE LA PREUVE À LA BONNE POPULATION D'INTÉRÊT

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

PAR
JESSIE BEAUCHEMIN

JANVIER 2023

Our own genomes carry the story of evolution, written in DNA, the language of molecular genetics, and the narrative is unmistakable.

Kenneth R. Miller

REMERCIEMENTS

D’abord, mes premiers remerciements sont sans aucun doute réservés à ma famille, mes parents Denis et Hélène, ainsi que ma sœur Kathleen. Merci d’avoir toujours cru en moi et de m’avoir toujours donné, jour après jour la motivation, la confiance et les encouragements nécessaires. Vous avez amplement su être à la hauteur et je vous en serai éternellement reconnaissante. Votre amour et votre soutien n’auraient pu être plus forts. Ma réussite est la vôtre.

Je tiens également à remercier mon directeur de recherche, Dr Emmanuel Milot, qui a joué une grande part dans la réalisation de ma maîtrise. Ta confiance, ta disponibilité ainsi que ton support tout au long des dernières années auront fait de moi une scientifique accomplie. Merci de m’avoir donné l’opportunité de travailler dans un domaine qui me passionne depuis très longtemps, celui des sciences forensiques. Je suis choyée d’avoir travaillé sur un projet des plus intéressants, bien que rempli de défis. Merci également à mon codirecteur, Dr Frank Crispino, d’avoir participé à la réalisation de ce projet de maîtrise. Vos conseils et vos commentaires auront su perfectionner et rehausser la qualité de ce projet de maîtrise, ce qui fut très bénéfique pour ce mémoire et pour ma formation.

Je ne peux passer sous silence l’aide inestimable de mon ami Étienne Beaulac, ressource première pour régler tous mes problèmes depuis l’année 2012. Merci pour ta patience intarissable et tes précieux conseils tout au long de mon initiation à l’univers de la programmation (qui l’aurait cru!). Ces quelques lignes ne sont pas suffisantes pour t’exprimer toute ma reconnaissance. Je te souhaite le meilleur pour la suite : aucun défi ni aucun projet n’est trop grand pour toi.

Merci à Roxane Landry, d’avoir été mon mentor lors de mon premier stage au sein du laboratoire de génétique des populations. Tu as su me rassurer dans mes moments de panique et tu as toujours su être présente pour moi. Merci de m’avoir guidé tout au long de mon parcours (et ce, même à environ 5 500 km de moi au Danemark lorsque que tu

répondais à mes appels de détresse). Encore à ce jour, je continue de prendre exemple sur toi et j'ai énormément d'admiration pour toi. Merci d'être la magnifique personne que tu es. Merci de même à mes collègues du laboratoire pour votre aide précieuse, votre dévouement et vos partages de connaissance. Vous avez contribué à faire de moi une personne rigoureuse, travaillante et débrouillarde, qualités qui m'ont grandement servi dans la réalisation de mon projet de maîtrise.

Finalement, merci à mes chers amis, Lydia, Maxence, Élizabeth et Amélie. Vous occupez tous une place immense et inestimable dans mon cœur. Vous êtes incontestablement des personnes uniques et irremplaçables. Non seulement je me considère sincèrement chanceuse, mais je suis également infiniment reconnaissante de vous avoir dans ma vie, pour encore un grand nombre d'années je l'espère. Pour finir, merci à toutes ces personnes trop nombreuses qui ont croisé mon chemin. Vous avez su me marquer à votre manière et, par le fait même, rendre mon parcours universitaire des plus inoubliables.

AVANT-PROPOS

En science forensique, les traces ADN retrouvées sur une scène de crime constituent un élément de preuve déterminant, permettant de corroborer ou d'invalider l'hypothèse d'une enquête criminelle. Il est possible d'extraire de ces supports une infime quantité d'ADN provenant d'une source biologique telle que du sang, de la salive ou du sperme, ce qui permettra de générer un profil génétique. La norme est d'évaluer la rareté de ce profil génétique (probabilité de concordance fortuite) en le comparant à un échantillon génétique de référence provenant d'un groupe ethnique jugé pertinent. Il s'agit donc d'une estimation de la fréquence d'un profil ADN particulier susceptible d'être tiré au hasard dans une population donnée, donc par pure coïncidence.

Afin d'évaluer le poids de la preuve ADN, ces statistiques doivent être basées sur une base de données composée des fréquences alléliques portées par une population d'intérêt appropriée au regard des hypothèses d'enquête. Les laboratoires d'expertises disposent d'échantillons d'ADN séparés pour des groupes ethniques différents : Caucasiens, Asiatiques, Noirs, etc. Ces bases de données regroupent l'ADN de donneurs volontaires et permettent la comparaison directe de profils génétiques. Dans le cas où l'ethnie de l'auteur d'un crime est connue, la base de données constituant la population de référence ne reflète pas la composition génétique réellement présente au sein d'une population d'intérêt, ce qui implique que les calculs de poids de concordance manquent d'exactitude et peuvent ainsi être biaisés. Les bases de données ne sont donc pas nécessairement représentatives de la population d'intérêt pour un cas criminel particulier.

Mon projet de maîtrise a pour cible la population d'intérêt liée à l'enquête (l'échantillon de référence), dans l'objectif d'attribuer de façon plus précise la rareté d'un profil génétique. Ce projet vise à remanier l'approche classique quant à l'utilisation et la composition des banques de données génétiques actuelles. Il propose à la communauté scientifique un nouveau modèle robuste basé sur les traces ADN trouvées dans

l'environnement pertinent à l'enquête, intégrant potentiellement une grande diversité génétique.

En effet, la ville de Montréal est idéale pour démontrer l'impact de l'aspect multiethnique d'une métropole sur les calculs de valeurs probantes utilisés en sciences forensiques. Selon le recensement de 2016 du profil sociodémographique, la superficie de l'agglomération de Montréal est répartie sur 499 km^2, accueillant pas moins de 1,9 million d'habitants. Parmi ces habitants, 645 000 sont immigrants, c'est-à-dire originaires d'un autre pays. Bien que ce projet de maîtrise fût concentré en grande majorité dans l'Est de la ville, il n'en demeure pas moins qu'une analyse complète de l'agglomération de Montréal sera nécessaire pour générer un échantillon complet couvrant la totalité du territoire, permettant donc de représenter au mieux la population d'intérêt dans une population caractérisée d'hétérogène.

Ce projet tient son originalité de la prise en compte de l'émergence de la multiethnicité de la ville de Montréal au cours des dernières décennies. Ce projet se veut donc innovateur, puisque personne ne s'est encore attardé à redéfinir le concept de population d'intérêt à l'enquête au moyen de l'ADN environnant.

RÉSUMÉ

Sur une scène de crime, l'ADN retrouvé sur un élément de preuve est possiblement analysé, permettant de comparer le profil génétique de la trace à celui du suspect. Si concordance il y a, il est essentiel d'évaluer la possibilité ou le risque qu'elle soit fortuite ou impertinente. La probabilité de concordance fortuite consiste donc en la probabilité de tirer au hasard un individu (autre que le suspect) dans une population d'intérêt donnée, dont le génotype correspond à celui déjà observé sur la trace. Puisque cette probabilité est basée sur les fréquences alléliques présentes au sein d'une population, elle permet d'évaluer la rareté d'un profil ADN. Ainsi, les bases de données utilisées par les laboratoires judiciaires servent d'échantillon de référence et sont divisées par ethnicité (Caucasiens, Asiatiques, Noirs, etc.), dont on doit tenir compte lors de l'évaluation du poids de la preuve. Cette approche n'est valable que si l'on dispose d'informations concernant l'ethnie de l'auteur des faits (et non du suspect). Dans le cas contraire, le profil de la trace est comparé aux différentes bases disponibles et la valeur probante la plus conservatrice (favorable au suspect) est alors avancée.

Cette méthode peine à intégrer la situation où la population d'intérêt s'avère multiethnique, d'où l'importance de s'intéresser aux populations se composant de nombreux groupes ethniques mélangés. L'objectif de l'étude est de tester un échantillon généré à partir de traces ADN prélevées aléatoirement dans l'environnement et laissées par les activités humaines, dans le but d'intégrer davantage la population d'intérêt lié à l'enquête et, de mieux évaluer la rareté d'un profil génétique au sein de cette population. Une première étape permettant de tester cet objectif était de comparer un échantillon de recrutement, servant de contrôle positif, à un échantillon de référence (base de données caucasienne), utilisé par le Laboratoire de sciences judiciaire et de médecine légale. Notre hypothèse est que l'occurrence de fréquences alléliques d'un échantillon contrôle multiethnique est indifférenciable d'un échantillon de référence ethnicisé, remettant donc en question la validité de la structure par division ethnique des bases de données et l'utilisation de celles-ci pour en calculer la valeur probante d'une preuve ADN.

Les résultats obtenus soutiennent cette hypothèse, puisqu'aucune différence significative au niveau de la composition génétique (fréquences alléliques) ou de l'indice de différentiation n'a été détectée. L'échantillon de référence ou de recrutement ne semblait pas avoir un impact majeur sur la probabilité d'évaluer une concordance ADN. Il a également été démontré qu'aucune structure de population n'a été détectée, suggérant des échantillons relativement homogènes.

Mots-clés : Probabilité de concordance fortuite, Base de données ADN, profil génétique, fréquence allélique, population d'intérêt.

TABLE DES MATIÈRES

LISTE DES TABLEAUX

Tableau **Page**

Table

LISTE DES FIGURES

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ADN	Acide désoxyribonucléique
BNDG	Banque nationale de données génétiques du Canada
CODIS	*Combined DNA INDEX System*
COI	Index des condamnées (*Convicted Offenders Index*)
CSI	Index des scènes de crime (*Crime Scene Index*)
EM	Expectation maximization
FBI	Bureau fédéral d'enquête (*Federal Bureau of investigation*)
GPS	Système mondial de positionnement (*Global Positioning System*)
HWE	Équilibre Hardy-Weinberg (*Hardy-Weinberg equilibrum*)
HWLE	Équilibre Hardy-Weinberg et de liaison (*Hardy-Weinberg and linkage equilibra*)
LE	Équilibre de liaison (*Linkage equilibrum*)
LSJML	Laboratoire de sciences judiciaires et de médecine légale
MCMC	*Markov Chain Monte Carlo*
PCR	Réaction en chaîne par polymérase (*Polymerase chain reaction*)
PDQ	Poste de quartier (*Local police station*)
QGIS	Système d'Information Géographique Libre (*Open-source Geographic Information System*)
RMP	Probabilité de concordance fortuite (*Random match probability*)
SNP	Polymorphismes de nucléotide unique (*Single-nucleotide polymorphism*)
SPVM	Service de Police de la Ville de Montréal
STR	Répétition en tandem court (*Short tandem repeat*)
WGS	Système géodésique Mondial (*World Geodetic System*)

GLOSSAIRE

Échantillon de référence :

L'échantillon de référence consiste en un échantillon de 276 individus recrutés volontairement, correspondant à la base de données caucasienne de référence utilisée, encore à ce jour, par le Laboratoire de sciences judiciaires et de médecine légale (LSJML).

Échantillon de recrutement :

L'échantillon de recrutement consiste en un échantillon de 126 individus recrutés volontairement et récemment collectés à l'UQTR, Sherbrooke et Dolbeau-Mistassini.

Échantillon de l'environnement :

L'échantillon de l'environnement consiste en un échantillon de 693 spécimens collectés en fonction de délimitations géographiques (Postes de Quartiers, PDQ) de la Ville de Montréal.

CHAPITRE I

INTRODUCTION

1.1 Problématique

En science forensique, les traces d'ADN sont des indices pouvant être déterminants pour la résolution d'une enquête, notamment lorsque le profil ADN d'un suspect concorde avec celui retrouvé sur une scène de crime. Toutefois, il est nécessaire pour l'enquêteur ou le décideur de fait (juriste, magistrat, jury) d'évaluer le poids de cette concordance, c'est-à-dire l'hypothèse qu'elle soit purement fortuite. Une concordance fortuite survient lorsque la trace est laissée par un inconnu qui possède le même profil génétique que le suspect. La probabilité de concordance fortuite dépend des fréquences (rareté) des diverses variantes (allèles) d'un gène existant dans la population d'intérêt pour l'enquête (p. ex., la population montréalaise si on soupçonne que l'auteur d'un crime vient de cette ville). Un échantillon génétique de référence, collecté dans la population d'intérêt, sert de banque d'ADN pour évaluer la rareté d'un profil génétique.

De façon générale, les laboratoires d'expertise disposent d'échantillons d'ADN provenant de dons volontaires et séparés en fonction des différents groupes ethniques, étant donné que ces ethnicités sont caractérisées par des variations génétiques considérables : Caucasiens, Asiatiques, Noirs, etc. (Budowle et al., 2001). Cette façon de constituer une banque pose trois problèmes. D'abord, elle est (présumément) valable seulement lorsqu'on a affaire à des populations ethniquement homogènes. À ce jour, l'aspect multiethnique de beaucoup de populations urbaines n'est pas pris en compte dans le cas où l'ethnie de l'auteur du crime n'est pas connue. Dans ce cas, il est donc essentiel de s'intéresser à la composante de la diversité ethnique d'une population, dans l'optique de générer une base de données représentative des traces pouvant être retrouvées dans l'environnement. Ensuite, ces banques ne répondent pas vraiment à la question pour lesquelles elles sont réellement constituées : quelle est la composition génétique des traces

ADN (et non des individus) laissées sans cesse dans l'environnement par les activités humaines quotidiennes. Enfin, les bases de données par divisions ethniques ne sont pertinentes (valables) que lorsque l'ethnie de l'auteur du crime est à priori connue par les circonstances de l'enquête, dans le but de confronter les traces ADN à celles appartenant au bon groupe ethnique.

Ainsi, en l'absence d'hypothèse liée à l'enquête (c'est-à-dire concernant l'auteur du crime), l'approche actuelle limite l'évaluation d'une trace ADN en raison de la structure définie par division ethnique des bases de données consultables. Notre approche propose donc de confronter une trace ADN à celles prélevées aléatoirement dans l'environnement pour en évaluer sa rareté, permettant de dresser géographiquement le portrait génétique (occurrence des fréquences alléliques) d'une population hétérogène, sans égard à la composition de nombreuses ethnicités.

1.2 Contexte historique

1.2.1 Génétique forensique

En 1901, les recherches de Landsteiner sur les différents groupes sanguins ABO constitue une découverte majeure et primitive au concept d'identification forensique (Landsteiner, 1901). Après plusieurs décennies, le Professeur Jeffreys et ses collègues s'intéresse aux variations génétiques entre les humains (Dumache, Ciocan, Muresan, et Enache, 2016). C'est en étudiant le gène de la myoglobine et en comparant plusieurs séquences que le généticien Jeffreys découvre des séquences répétitives de l'ADN (marqueurs génétiques), hautement variables et informatives. Ainsi, ces signatures de l'ADN sont qualifiées par son équipe comme des empreintes génétiques. Appelé polymorphisme de longueur des fragments de restriction, ce marqueur génétique découvert par Jeffreys consiste en la digestion des molécules d'ADN (grâce à des enzymes spécifiques) permettant la fragmentation à différentes localisations du brin d'ADN. Il en résulte donc la migration de fragments de tailles variables sur un gel électrophorèse, se différenciant d'un individu à l'autre (Chaudhary et Maurya, 2020). Jeffreys, pionnier

des tests d'identification au moyen du profilage ADN, publie pour la première fois en 1985, sur le potentiel des empreintes génétiques et ses applications embryonnaires en forensique (Jeffreys, Wilson, et Thein, 1985a, 1985b; Saad, 2005).

Ce n'est qu'en 1986, à Leicestershire, que l'ADN fût utilisé pour la première fois comme preuve au sein d'une enquête criminelle. Une chasse à l'homme en Grande Bretagne mène au prélèvement d'ADN de masse (spécimens de sang ou de salive), regroupant plus de 5000 individus. Le profilage ADN de seulement 10 % des hommes possédant le même groupe sanguin que le tueur est analysé en profondeur, mais sans succès. Par la suite, des informations divulguées à la police permettent de diriger l'enquête vers Colin Pitchfork. Celui-ci fût fournit un échantillon de sang, qui permettant d'observer une concordance avec l'ADN retrouvé sur les deux victimes. Pitchfork est reconnu coupable sur la base de traces ADN du viol et du meurtre de deux jeunes femmes. C'est ainsi que le profilage ADN est marqué comme révolutionnaire au niveau des enquêtes criminelles et utilisé partout dans le monde (Saad, 2005).

Depuis 1998, la *Loi sur l'identification par les empreintes génétiques* est implantée au Canada, permettant la compilation des profils génétiques au sein d'une base de données à des fins d'identification de criminels (Gouvernement du Canada, 1998). Seulement deux ans plus tard, le 30 juin 2000, la Banque Nationale de données génétiques du Canada (BNDG) est instaurée, et contient à ce jour (mars 2020), plus de 575 682 profils ADN. Dans la même année, la BNDG permet pas moins de 6 202 correspondances entre des profils ADN prélevés sur une scène de crime avec un profil de contrevenant condamné (Gendarmerie Royale du Canada, 2020).

1.3 ADN et utilisation dans enquête criminelle

1.3.1 L'ADN humain

L'ADN génomique comprend des polymorphismes de séquences non-codantes regroupés en trois catégories : les minisatellites, les microsatellites et les polymorphismes

de nucléotide unique (SNP). L'information permettant de créer un profil génétique consiste en l'utilisation de marqueurs génétiques, retrouvés sur les chromosomes autosomaux ou sur le chromosome Y et répartis partout dans le génome (Griffiths Anthony, Sanlaville, et Charmot-Bensimon, 2013, pp. 136-140; Vieira, Santini, Diniz, et Munhoz, 2016). Par définition, ces marqueurs génétiques consistent en un gène ou une séquence polymorphe d'ADN permettant d'étiqueter une portion chromosomique connue. Cette localisation spécifique du gène sur le chromosome est connue sous le nom de locus (Griffiths Anthony et al., 2013). Seulement 0,3 % des molécules d'ADN diffèrent entre les individus de la population. Cette infime fraction représente environ 10 millions de nucléotides et confère aux régions polymorphes la variabilité génétique (Butler, 2005, p. 26).

Les Short Tandem Repeat (STR) sont des microsatellites, dont la séquence est désignée par la longueur de l'unité répétitive (motif de nucléotides de 2 à 8 paires de bases) et le nombre d'unités répétitives contiguës (2 à 20 répétitions), **Figure 1.1**. Les allèles STR diffèrent ainsi les uns des autres et sont nommés de par le nombre de répétitions qu'ils contiennent (Jamieson et Bader, 2016, p. 32). Ces microsatellites impliquent une donc grande variance d'allèles entre les individus de la population (Griffiths Anthony et al., 2013). Bien que cette variabilité au sein des locus STR attribue une valeur informative à l'ADN, celles-ci doivent concéder une certaine stabilité aux allèles, leur permettant d'être ainsi transmis de génération en génération (Butler, 2005, p. 466).

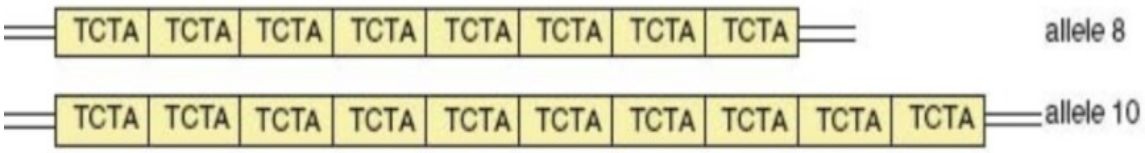


Figure 1.1 Séquence ADN, dont la longueur de l'unité répétitive est de quatre nucléotides (motif TCTA).
Le nombre d'unités contiguës est de 8 et 10 répétitions respectivement pour ces séquences et correspond aux allèles du marqueur D8S1179 (Goodwin, Linacre, et Hadi, 2011, p. 17).

La variabilité allélique confère également à l'ADN un grand pouvoir discriminant. En science forensique, le pouvoir discriminant est la probabilité de séparer deux à deux les individus par la méthode au sein d'une population donnée (Smalldon et Moffat, 1973), et correspond donc à la sélectivité, ou le taux de vrais négatifs.

1.3.2 Préservation des échantillons ADN

L'ADN est un outil de choix dans les dossiers judiciaires puisqu'il permet l'identification potentielle d'auteurs en plus de permettre la discrimination de l'auteur du crime et de la victime. La preuve biologique retrouvée sur la scène de crime, tel que du sang, de la salive ou du sperme ou des cellules épithéliales, est transférée via dépôt direct soit sur les victimes ou sur une surface quelconque (gobelet, paille, cannette, mégot) et adhère à celle-ci, faisant référence à une trace qui contient une infime quantité d'ADN (Lee et Ladd, 2001; van Oorschot, Ballantyne, et Mitchell, 2010). Cet ADN sera collecté en vue de générer un profil génétique qui sera comparé avec un échantillon de référence, en vue de l'interprétation des données (Jamieson et Bader, 2016, p. 107).

À titre de référence, dépendamment du type de spécimen, la quantité d'ADN pouvant être extraite varie grandement, comme l'indique le **Tableau 1.1**. Sur une scène de crime, les conditions (notamment environnementales) peuvent impacter considérablement la quantité et la qualité de l'ADN, étant susceptible de réagir avec plusieurs agents physiques ou réactions chimiques (Alaeddini, Walsh, et Abbas, 2010; Lee et Ladd, 2001). Notamment, lorsqu'exposées aux rayons UV du soleil ou à des environnements chauds et humides, les traces ADN sont beaucoup moins propices à la préservation de celle-ci. Quelques heures d'expositions seraient suffisantes pour créer un dommage à l'ADN et causer sa dégradation (Jamieson et Bader, 2016, p. 141). De plus, des facteurs comme la quantité, la dégradation (causée par la contamination bactérienne) et la pureté d'un spécimen peuvent résulter en l'incapacité de générer un profil génétique complet. Étant donné que sur une scène de crime, les spécimens peuvent être entourés de nombreuses substances diverses telles que des saletés, des graisses ou des fluides, il est important de limiter la zone de cueillette à même la trace ADN, ce qui permet d'éviter la

détérioration de l'ADN et d'affecter la qualité de l'ADN lors du processus d'analyse (Lee et Ladd, 2001).

Tableau 1.1

Types de spécimens pertinents en forensique, permettant d'extraire une certaine quantité d'ADN optimale, en vue de l'amplification et de la détection des profils génétiques (Lee et Ladd, 2001)

Type of sample	Amount of DNA
Liquid blood	20,000-40,000 ng/mL
stain	250-500 ng/cm^2
Liquid semen	150,000-300,000 ng/mL
Postcoital vaginal swab	10-3,000 ng/swab
Hair (with root)	
Plucked	1-750 ng/root
Shed	1-10 ng/root
Liquid saliva	1,000-10,000 ng/mL
Oral swab	100-1500 ng/swab
Urine	1-20 ng/mL
Bone	3-10 ng/mg
Tissue	50-500 ng/mg

[a]Quantity of DNA recovered from evidentiary samples is significantly affected by environmental factors.

1.3.3 Détection d'un profil génétique au moyen de marqueurs STR

Dans l'optique de récolter un maximum d'ADN contenu dans les spécimens, les prélèvements se font selon l'approche la plus appropriée, soit dépendamment de l'état et de la condition de la pièce à conviction (mégot, gomme, gobelet, paille) (Lee et Ladd, 2001). Par exemple, écouvillonner directement lorsqu'il s'agit de gobelet, de paille ou de gommes et découper directement un petit morceau de papier filtre lorsqu'il s'agit d'un mégot. La source d'ADN d'une preuve biologique est rarement visible à l'œil nu et collecter une trace comporte des risques. Il importe de cibler la zone de collecte adéquatement : écouvillonner une zone plus petite risque de ne pas prélever assez d'ADN, tandis qu'écouvillonner une zone plus grande risque de répartir l'ADN sur une plus grande surface, et par le fait même, ne pas en prélever assez (van Oorschot et al., 2010).

Ainsi, l'ADN extrait, permettant de le séparer du reste du contenu cellulaire et d'éliminer tout inhibiteur. De la résine paramagnétique, recouverte de silice est utilisée, de par sa forte affinité, pour capturer les molécules d'ADN en présence d'éléments chaotropiques (Jamieson et Bader, 2016, p. 101). L'ADN extrait doit ensuite être quantifié, étape nécessaire et préalable afin de déterminer la quantité d'ADN optimale pour l'étape d'amplification (seuil minimal de 0,1-0,5 ng requis par les kits d'amplification utilisés par les laboratoires judiciaires) (Goodwin et al., 2011, p. 45; van Oorschot et al., 2010).

Par la suite, l'amplification permet d'accroître la quantité d'ADN de manière exponentielle, en copiant la portion locus d'une région conservative de l'ADN. Des amorces spécifiques sont désignées pour reconnaitre la séquence ADN d'intérêt à amplifier. Utiliser des petites séquences microsatellites (loci STR), plutôt que des minisatellites, a pour avantage de pouvoir amplifier des traces ADN dégradées grâce à la sensibilité de la méthode (Butler, 2015b; Goodwin et al., 2011, p. 53). Cependant, les séquences doivent être suffisamment intactes pour que les amorces spécifiques puissent se fixer aux extrémités de la séquence ADN et permettre l'amplification (Butler, 2015b).

Ainsi, l'ADN migre sous l'effet d'un champ électrique, au travers de petits capillaires remplis d'une solution de polymère enchevêtré. Puisque du colorant fluorescent est attaché aux amorces des produits PCR, les allèles sont détectés via une électrophorèse par capillaire. La longueur des loci peut se chevaucher, permettant ainsi jusqu'à cinq colorants fluorescents d'être utilisés pour détecter plusieurs allèles simultanément (Goodwin et al., 2011, p. 70).

Le processus de séparation des analytes s'effectue en fonction de la mobilité, c'est-à-dire en fonction de la taille de l'ADN (les plus petits fragments d'ADN migrent plus rapidement) (Jamieson et Bader, 2016, pp. 37-38). Les molécules d'ADN sont donc séparées lorsqu'elles passent devant de petites fenêtres de détection, où les colorants fluorescents des amorces sont excités par un laser. Le signal d'intensité relative de la

lumière capté par les pixels de la caméra est converti, via différents algorithmes, en unité relative de fluorescence (RFU), correspondant à la hauteur des pics sur un électrophorégramme (proportionnel à la quantité de produits PCR détectés). Ainsi, le détecteur électronique peut mesurer jusqu'à quatre types de données pouvant être collectés : la position du capillaire, la longueur d'onde émise du spectre de lumière visible, le signal d'intensité relative de la lumière à une certaine longueur d'onde et la période de temps pour chaque cadre de lecture (Butler, 2015a).

Les kits multiplex utilisés par les laboratoires forensiques permettent de performer ces techniques de profilage ADN, c'est-à-dire l'amplification simultanée et parallèle de plusieurs loci STR (Jamieson et Bader, 2016, pp. 115-118). Ainsi, 16 marqueurs STR sont utilisés afin de générer un profil génétique: D21S11, CSF1PO, Vwa, D8S1179, TH01, D18S51, D5S818, D16S539, D3S1358, D2S1338, TPOX, FGA, D7S820, D13S317, D19S433 et Amelogenin (Butler, 2005, p. 98). Pour des applications d'identification humaine en science forensique, ces marqueurs comportent plusieurs caractéristiques critiques. D'abord, ils ont été sélectionnés de manière à éviter les problèmes de liaison entre eux, c'est-à-dire le risque d'observer des génotypes corrélés. Ces marqueurs sont donc situés sur des chromosomes distincts ou suffisamment distancés sur le même chromosome, permettant la ségrégation indépendante durant la méiose (Butler, 2006, 2012b; Chakraborty, Stivers, Su, Zhong, et Budowle, 1999). Ensuite, ces marqueurs choisis ne comportent pas de signification fonctionnelle majeure, impliquant que les génotypes ne peuvent pas être prédits. Enfin, ces marqueurs sont suffisamment polymorphes (offrant une grande gamme de tailles d'allèles) pour générer une diversité génétique considérable, et ce, même dans les populations les plus isolées géographiquement (Chakraborty et al., 1999).

1.3.4 Interprétation des profils ADN

Puisque les séquences polymorphiques se distinguent les unes des autres par le nombre de répétitions qu'elles contiennent, la longueur des produits PCR doit être mesurée précisément, puis comparée à des standards. La désignation des allèles dépend

de l'échelle allélique standardisée pour les marqueurs STR utilisée pour l'analyse, **Figure 1.2** (Goodwin et al., 2011, pp. 72-76; SWGDAM, 2017).

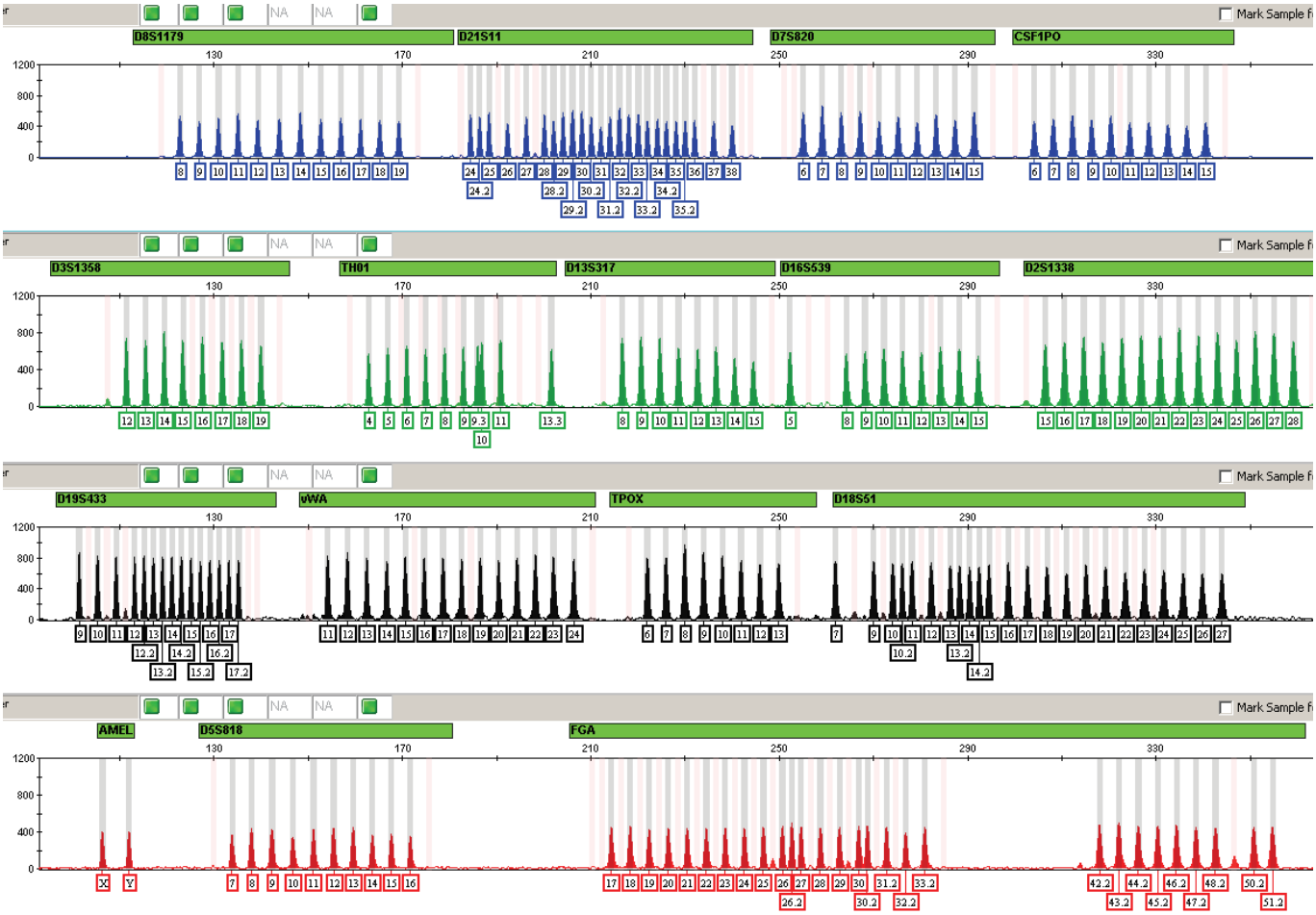


Figure 1.2 Échelle allélique standardisée contenant les allèles communs de 16 marqueurs STR recommandés par le Combined DNA Index System (CODIS) (Hares, 2015).
Différents fluorochromes sont utilisés pour la détection simultanée des produits PCR. Sous chaque pic, le numéro d'allèle est indiqué dans un encadré (Thermo Fisher Scientific, 2018).

Généralement, le profil ADN d'un individu possède à chaque locus un ou deux pics d'allèles, caractérisant respectivement ceux-ci comme un allèle homozygote ou hétérozygote. Lorsque 3 ou 4 allèles sont présents au même locus, et que cette tendance se répète à plusieurs loci au sein du même profil, le spécimen analysé suggère un mélange potentiel de deux individus (Butler, 2012b; Goodwin et al., 2011, pp. 86-87). Idéalement, deux allèles hétérozygotes seront présents dans une proportion de 1:1 (Gill, Sparkes, et

Kimpton, 1997). Toutefois, de légères différences de hauteurs de pics peuvent découler d'une mauvaise efficacité de l'amplification PCR, d'où l'importance de comprendre les limitations de la technologie d'électrophorèse par capillaire au moyen de STR et de l'interprétation des données (SWGDAM, 2017).

Afin de s'assurer de l'interprétation et de la fiabilité des allèles (donnée réelle vs artefacts), il est nécessaire d'analyser les données empiriques obtenues via l'électrophorèse capillaire et de comparer avec précision ces résultats à des valeurs de seuils (Gill et al., 1997; SWGDAM, 2017). **Le seuil analytique** (valeur de 50 RFU) correspond à la valeur minimale permettant de détecter et de distinguer un pic réellement observé par rapport aux valeurs de bruit de fond de l'instrument. **Le seuil stochastique** correspond à la valeur minimale permettant de raisonnablement mettre en hypothèse une perte d'allèles dans un échantillon à profil unique. Les effets stochastiques surviennent lors de l'amplification d'échantillons faiblement concentrés en ADN : les allèles hétérozygotes d'un locus particulier présentent un rapport de hauteur de pic inférieur à 60 % (débalancement considérable) ou bien l'électrophorèse par capillaire ne parvient pas à détecter un allèle. **Les pics de bégaiements (stutters)** doivent habituellement avoir une intensité inférieure à 15 % des allèles associés pour être considérés comme des pics artéfacts.

1.4 Base de données génétiques

1.4.1 Un outil d'enquête criminelle

Les bases de données permettent de tirer profit de la technologie ADN afin d'assister les acteurs de la justice lors d'investigation criminelle (Jamieson et Bader, 2016, p. 177). Les bases de données générées à partir des traces ADN sont composées d'échantillons provenant de dons volontaires ou d'échantillons de diverses sources, telles que les banques de sang, les laboratoires de test de paternité, le personnel de laboratoire, les agents de la force publique et les personnes accusées de crimes (criminels identifiés et connus) (National Research Council, 1996). Les échantillons ADN sont donc prélevés et doivent

être analysés par les laboratoires judiciaires afin d'obtenir les profils génétiques, c'est-à-dire les génotypes STR, qui seront par la suite compilés dans un système de base de données. Éventuellement, des spécimens inconnus prélevés sur des scènes de crime pourront être comparés à la base de données dans l'optique d'obtenir de potentielles concordances ADN (Butler, 2012c).

Ainsi, la Banque Nationale de données génétiques (BNDG) est un outil puissant pour les enquêtes criminelles puisqu'elle regroupe des centaines de milliers de profils ADN. Bien qu'il existe de nombreuses particularités liées à l'attribution d'un profil génétique, les spécialistes se servent des bases de données pour traiter et comparer les profils génétiques, et ce, dans le but d'identifier la source d'une trace biologique, mais également afin d'établir des connexions entre des scènes de crime appartenant à différentes juridictions (Gendarmerie Royale du Canada, 2020). Le BNDG est un répertoire de données basé sur deux index, qui sont d'une importance primordiale pour les enquêtes criminelles. L'ADN collecté des scènes de crime est regroupé dans **l'index des scènes de crime** (Crime Scene Index, CSI), alors que **l'index des condamnées** (Convicted Offenders Index, COI), avec autorisation de la cour, regroupe les profils génétiques de personnes ayant commis divers types de crime (Milot, Lecomte, Germain, et Crispino, 2013). Dans l'espoir d'obtenir une concordance ADN dans une enquête non résolue, les profils génétiques provenant du COI sont comparés à ceux du CSI (Lalonde, 2006). D'ailleurs, le Laboratoire de sciences judiciaires et de médecine légale (LSJML), ayant collaboré pour cette étude, contient des profils génétiques d'enquête non résolue, appartenant à l'index CSI (Lalonde, 2006).

1.4.2 Le système CODIS

Le logiciel de réseau de profils génétiques, le Combined DNA Index System (CODIS), est un programme implanté depuis novembre 1997 et reconnu dans plusieurs laboratoires judiciaires du monde. Cet index conçu par le Federal Bureau of investigation (FBI) et le Département de la Justice des États-Unis rend accessible au BNDG les données génétiques à des fins de comparaisons de profils ADN (Butler, 2006; Gendarmerie Royale

du Canada, 2020). Ce système CODIS regroupe les informations suivantes : (1) un identifiant permettant de retracer la personne ou l'organisme soumettant l'échantillon, (2) un identifiant de l'échantillon, (3) un identifiant de la personne responsable de l'analyse du profil génétique (4) ainsi que le profil génétique en soi (valeur numérique pour chaque allèle analysé pour tous les loci) (FBI - Federal Bureau of Investigation).

Ainsi, l'analyse du profil génétique dans le système CODIS était originalement basée sur 13 loci STR: CSF1PO, FGA, TH01, TPOX, vWA, D3S1358, D5S818, D7S820, D8S1179, D13S317, D16S359, D18S51, et D21S11. Plus tard, le FBI a décidé d'étendre le nombre de loci STR dans le but d'augmenter les chances de compatibilité entre les laboratoires au niveau international, et donc la performance des bases de données. Puisque le choix des marqueurs STR utilisés varie entre les différents pays, augmenter le nombre de loci STR permet nécessairement d'uniformiser et de standardiser davantage le partage de données à des fins de comparaison internationale, principalement pour ce qui a trait aux comparaisons de profils de source unique (Ge, Eisenberg, et Budowle, 2012; Hares, 2012). Cela permet également d'augmenter le pouvoir de discrimination des profils génétiques, ce qui permet de réduire et de minimiser le risque de concordance fortuite. Les 7 loci STR additionnels ayant été implémentés au système de base de données CODIS sont : D1S1656, D2S441, D2S1338, D10S1248, D12S391, D19S433 et D22S1045 (Hares, 2012, 2015; Moretti et al., 2016).

Ces marqueurs STR sont universels pour les laboratoires qui collaborent avec le système CODIS et ont un grand pouvoir de discrimination et d'individualisation (Butler, 2012c). C'est pourquoi ces loci STR sont particulièrement utilisés en forensique, puisqu'ils se caractérisent par leurs séquences hypervariables, qui diminuent les risques de concordance fortuite de par leur grande hétérogénéité. Il importe donc de bien définir et de bien représenter dans son ensemble la composition d'une base de données afin d'évaluer de manière conservative les génotypes ainsi que les fréquences alléliques (Chakraborty, 1992).

1.4.4 Caractériser une population au travers d'une base de données génétiques

Par définition, une population est un groupe d'individus dépendant d'une géographie délimitée et à un temps donné et qui partage un même ancêtre commun (Butler, 2015b; Waples et Gaggiotti, 2006). Dans une optique plus statistique, une population se compose d'individus dont son nombre tend vers l'infini (ou est plutôt difficilement mesurable dans son entièreté). Une sous-population peut être définie comme une extraction représentant un groupe d'individus habitant dans la même région ou le même pays, voir **Figure 1.3**. Puisqu'il existe de nombreuses variations génétiques héréditaires, la génétique des populations permet de quantifier les fréquences d'allèles et de génotypes existant à l'intérieur d'une population ou parmi les populations (Butler, 2005, p. 465; 2015b).

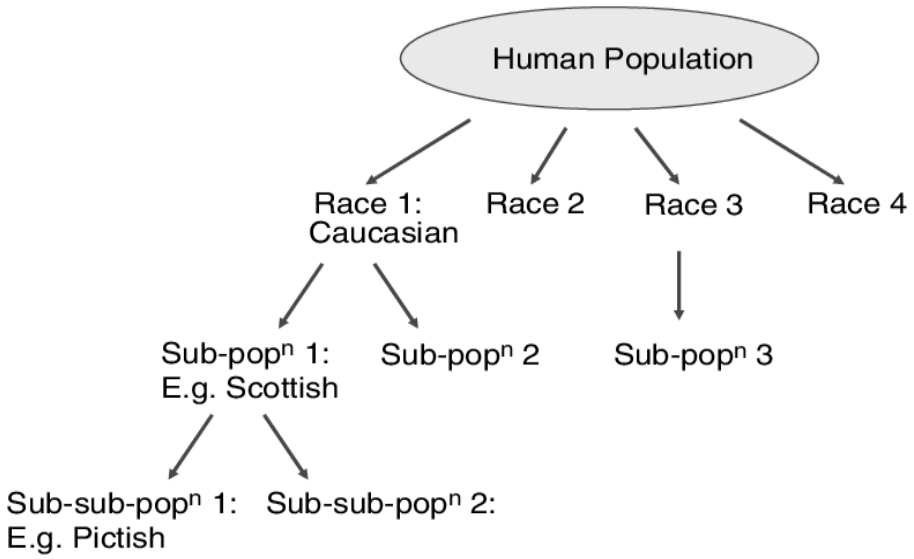


Figure 1.3 Structure de la population humaine et ses ramifications en sous-populations, groupée en fonction de la race (Buckleton, Triggs, et Walsh, 2004, p. 152).

En forensique, une population d'intérêt fait référence à un groupe d'individus étant considérés comme suspects ou pouvant être à l'origine (considérant, par exemple, l'aspect géographique) du crime en question. Puisqu'il est impossible de disposer d'une base représentative de cette population d'intérêt, la population de référence est basée sur l'utilisation des bases de données existantes données (D. J. Balding et Nichols, 1994). Les bases de données, notamment CODIS, disposent donc d'échantillons d'ADN, qui en

plus d'être complètement anonymes, sont divisés en sous-populations en fonction des groupes ethniques/raciaux majeurs, soit Afro-Américains (Noirs), Asiatique, Caucasiens, Hispanique ou Amérindiens ("Appendix 2 – NRC I & NRC II Recommendations," 2015; Budowle, Shea, Niezgoda, et Chakraborty, 2001). Pour générer au mieux une base de données exprimant et représentant toute la complexité de la structure génétique d'une population réelle, la base de données médico-légale est composée de plusieurs sous-populations d'intérêt. En effet, ces sous-populations peuvent s'exprimer à différents niveaux et ne devraient donc pas être complètement séparées ou indépendantes les unes des autres en raison des différentes structures géographiques et sociales qui existent (Curran, Buckleton, et Triggs, 2003).

1.5 Évaluation du poids de la preuve

1.5.1 Évaluer la rareté d'une preuve ADN

Comme les laboratoires judiciaires disposent de bases de données regroupant les allèles communs dans la population considérée (ici toujours divisée par ethnicité), les fréquences d'allèles peuvent être estimées de manière fiable. Il est ainsi possible de comparer un profil génétique à une population d'intérêt, dont les fréquences alléliques sont compilées en fonction du groupe ethnique, afin d'attribuer la rareté de ce profil particulier parmi la population en question (Butler, 2005, p. 473). Toutefois, même si la base de données peut contenir des allèles rares et que très peu présents au sein de la population, les allèles communs doivent être observés plusieurs fois afin de représenter au mieux la population (Butler, 2015b). Afin de ne pas biaiser la base de données et d'estimer les fréquences alléliques d'un locus de manière conservative, un allèle doit au minimum avoir été observé 5 fois afin que sa fréquence observée soit considérée lors des calculs statistiques, le cas échéant, la fréquence allélique minimale observée à ce locus doit être utilisée (Butler, 2005, p. 477; 2015b). Finalement, afin d'assurer de représenter au mieux la population d'intérêt, un minimum de 100 individus est requis pour constituer un échantillon de population. Cet échantillon permet d'estimer de manière conservative les fréquences alléliques, provenant des possibilités de génotypes infinis, en raison de

l'utilisation de marqueurs STR hypervariables (Chakraborty, 1992). En effet, une taille d'échantillon trop faible risque nécessairement de ne pas bien refléter les fréquences alléliques réelles d'une population, et ce, en raison des erreurs d'échantillonnage. Plusieurs conditions sont donc primordiales pour déterminer la fréquence d'un génotype au travers d'une population : la population d'intérêt est clairement identifiée, la taille de l'échantillon doit être suffisamment grande pour représenter avec précision la population, l'échantillon est aléatoirement prélevé et la population homogène suit l'équilibre Hardy-Weinberg (HWE) pour chaque locus, et tous les loci sont en équilibre de liaison (LE) (Lander, 1989).

1.5.2 Équilibre de liaison et équilibre Hardy-Weinberg

D'abord, l'équilibre de liaison et l'équilibre Hardy-Weinberg sont basés sur deux prémisses. La première, la loi de la ségrégation, implique que lors de la méiose, les paires de chromosomes se divisent aléatoirement en deux gamètes haploïdes. La deuxième, la loi de l'assortiment indépendant, implique que la recombinaison génétique, entre les générations, influence la ségrégation des différentes paires de gènes de manière indépendante (Butler, 2015b).

L'équilibre de Hardy-Weinberg nécessite un appariement aléatoire entre les individus de la population et implique que, pour chaque locus particulier, la fréquence des génotypes demeure constante d'une génération à une autre (Stern, 1943). Ainsi, considérant que le génotype attendu pour deux allèles (A et a) à un marqueur génétique particulier est AA, Aa et aa, et que la fréquence de ces deux allèles est respectivement représentée par p et q, il est possible de calculer la fréquence du génotype estimée via un carré de Punnett, soit p^2, $2pq$ et q^2, voir **Figure 1.4** (Butler, 2015b). Toutefois, il importe que l'échantillon de population soit infini, que l'appariement survienne aléatoirement, que la population ne subisse pas de migration, qu'il n'y ait pas de sélection naturelle et qu'il n'y ait pas de mutations. Puisqu'il est pratiquement impossible pour une population humaine de suivre le modèle HWE et ses conditions, les déviations de fréquences de

génotypes peuvent être prédites, puis la sous-structure de la population peut être corrigée et ajustée adéquatement (Goodwin et al., 2011, pp. 95-99).

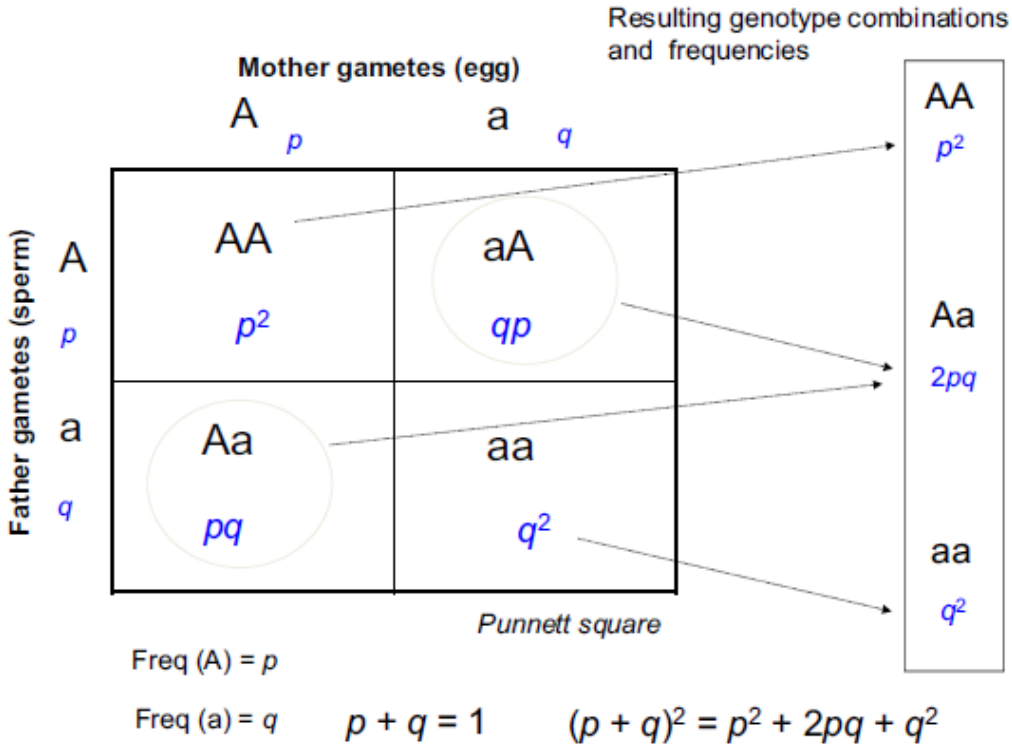


Figure 1.4 L'équilibre de Hardy Weinberg permet d'estimer la fréquence génotypique dans un carré de Punnett en fonction des fréquences alléliques dans une population ou l'appariement est aléatoire (Butler, 2015b).

Ensuite, l'équilibre de liaison implique que les gènes ne sont pas hérités dans un même ensemble. Puisqu'un allèle à un locus particulier est indépendant des autres allèles sur différents loci, l'association des gènes peut se faire de manière aléatoire. On parle alors de loci en équilibre de liaison (Butler, 2015b, 2015c). Ainsi, il est possible d'estimer la fréquence du profil génétique, parce que les loci STR sélectionnés dans les kits forensiques sont situés sur des chromosomes différents (les fréquences des génotypes peuvent donc être multipliées) (Butler, 2015c). Ainsi, HWE et LE reposent tous deux sur un état d'indépendance : le premier, entre les allèles à un locus particulier et le second, entre les allèles à différents loci (Buckleton et al., 2004, p. 70).

1.5.3 Probabilité d'un profil génétique

La probabilité d'un génotype particulier est dépendant du nombre de fois qu'un génotype peut être observé ainsi que la fréquence des allèles (Butler, 2015c). Ainsi, la règle du produit, selon HWE et LE, permet de calculer la probabilité d'un profil génétique estimée : la fréquence d'un génotype à chaque locus (**Équation 1.1**) est considérée puis multipliée pour tous les locus (**Équation 1.2**), étant donné que les fréquences d'allèles sont connues dans une banque de données de la population (Council, 1992; National Research Council, 1996)

L'équation 1.1 permet de calculer la fréquence d'un génotype à un locus particulier.

$$P_G = \emptyset\, p_1 p_2 \quad \textbf{(1.1)}$$

Où

P_G : Probabilité du génotype à un locus particulier

$\emptyset$: Valeur homozygote (1) et hétérozygote (2)

p_1 : Fréquence du premier allèle au travers d'une population

p_2 : Fréquence du deuxième allèle au travers d'une population

L'équation 1.2 permet de calculer la probabilité d'un profil génétique estimée et considérée pour tous les locus, et ce, au travers d'une population.

$$P_G = \prod_{l=1}^{N_L} \emptyset P_{l,1} P_{l,2} \textbf{(1.2)}$$

Où

P_G : Probabilité du génotype pour tous les locus

N_L : Nombre de locus considérés

$\emptyset$: Valeur homozygote (1) et hétérozygote (2)

$p_{l,1}$: Fréquence du premier allèle au locus l

$p_{l,2}$: Fréquence du deuxième allèle au locus l

Ainsi, les allèles sont répertoriés pour tous les loci STR, et ce, dans chaque population (Afro-Américains (Noirs), Asiatique, Caucasiens, Hispanique ou Amérindiens), donnant lieu à des valeurs de fréquence d'allèle connues et compilées dans les bases de données (Butler, 2015c). Toutefois, l'incertitude de ces calculs repose sur l'attribution de la bonne population d'intérêt, compris comme le groupe ethnique correspondant. Chaque groupe ou sous-population de la base de données possède non seulement une taille d'échantillon différente, mais également des fréquences d'allèles différentes, et ce, indépendamment de la taille de l'échantillon. La probabilité d'un profil génétique particulier se retrouve donc influencée (National Research Council, 1996).

En 1996, le National Academy of Science's National Research Council (NRC) a émis des recommandations concernant la valeur probante des preuves ADN (National Research Council, 1996). La recommandation 4.1 suggère de sélectionner la base de données relative au groupe racial particulier de la personne ayant laissé la trace ADN sur la scène de crime, si elle est connue. Dans le cas contraire, tous les groupes raciaux dont l'auteur du crime est susceptible d'appartenir doivent être évalués dans les calculs de probabilités.

1.5.4 Probabilité de concordance fortuite

Une analyse ADN entre une trace prélevée sur une scène de crime et un suspect peut résulter en trois conclusions : non-concordance, non-concluant, ou concordance. Lors d'une concordance, il est possible que (1) le suspect soit la source de la trace ADN, (2) que par hasard, le suspect possède le profil ADN de la trace, bien qu'il ne soit pas la source de l'échantillon (3) ou bien qu'il s'agisse d'un résultat faux positif dû à des erreurs de type humaines (Butler, 2012a; SWGDAM, 2017). Il est donc nécessaire d'évaluer statistiquement l'incertitude de la concordance ADN obtenue, à savoir si cette concordance est purement fortuite (SWGDAM, 2017) (National Research Council, 1996).

La probabilité de concordance fortuite (RMP) consiste donc en la probabilité de tirer au hasard un individu (autre que le suspect) dans une population d'intérêt donnée, dont le

génotype correspond à celui déjà observé chez le suspect (Butler, 2015c; Curran, Walsh, et Buckleton, 2007). Ainsi, la comparaison de profils génétique permettant de conclure une concordance ADN ou non doit se baser sur le/les génotypes particuliers possibles, et ce, pour chaque contributeur, plutôt que d'évaluer la présence ou l'absence d'un allèle à chaque locus (SWGDAM, 2017). Comme le définit le terme « concordance », nécessairement deux profils sont impliqués au sein des calculs de probabilité, cela ne fait donc pas référence à la fréquence d'un profil génétique dans une population donnée. Donc, si le suspect est en réalité innocent, deux copies de ce même profil génétique ont été observées au travers de la base de données, ce qui est faiblement probable (D. J. Balding et Nichols, 1994; D. J. Balding et Nichols, 1995).

Le concept de probabilité fait référence au nombre de fois qu'un événement survient divisé par le nombre de chances que l'événement se produise. Bien qu'il soit intuitif de penser que la probabilité qu'un événement se produise soit : l'événement survient (1) ou l'événement ne survient pas (0), la notion de probabilité est beaucoup plus complexe. En raison de l'incertitude d'un évènement, l'éventail des possibilités se situe entre 0 et 100 % (Butler, 2005, p. 457). En forensique, l'incertitude d'une probabilité, plus précisément au niveau de l'interprétation de la preuve ADN, doit être prise en considération. Non seulement l'incertitude d'une probabilité peut être influencée par des erreurs possibles de manipulation ou contamination de l'échantillon de preuve, mais cela peut également être influencée par l'attribution d'une population de référence sur une échelle géographique, non nécessairement directement corrélée à une ethnie ou un groupe racial spécifique. (D. J. Balding et Steele, 2015, p. 14).

En effet, le calcul de probabilité de concordance fortuite présenté aux acteurs de la justice est une valeur assez robuste en forensique. Il suffit de 10 ou 11 loci STR pour obtenir un très fort pouvoir discriminant (D. J. Balding, 1999). En moyenne, la probabilité de concordance pour 13 loci STR est estimée à 1 sur 1 billion, et ce, même si la population est limitée à une faible variabilité génétique. Cette probabilité est suffisamment faible pour donc justifier l'utilisation de 13 loci STR en science forensique (Chakraborty et al., 1999).

Ainsi, le **Tableau 1.2** présente un exemple de calcul de la probabilité d'un profil génétique attendu, dont les fréquences alléliques sont basées sur les données de la population des États-Unis ("Appendix 1 - STR Allele Frequencies from U.S. Population Data," 2015).

Tableau 1.2

Calcul de probabilité de concordance fortuite d'un profil génétique basé sur 13 locus STR

STR Locus	**Allèle 1, Allèle 2**	**Fréquence d'allèle 1 (*p*)**	**Fréquence d'allèle 2 (*q*)**	**Formule**	**Fréquence du génotype attendu**
D18S51	17, 17	0,1520	0,1520	p^2	0,02310
CSF1PO	11, 12	0,2490	0,2950	$2pq$	0,14700
FGA	21.2, 22	0,00731	0,1990	$2pq$	0,00291
TH01	8, 8	0,1960	0,1960	p^2	0,03840
TPOX	11, 11	0,2160	0,2160	p^2	0,04670
VWA	17, 18	0,2350	0,1490	$2pq$	0,07000
D3S1358	16, 16	0,3190	0,3190	p^2	0,10200
D5S818	11, 12	0,2340	0,3700	$2pq$	0,17300
D7S820	11, 13	0,2030	0,0146	$2pq$	0,00593
D8S1179	13, 15	0,2190	0,1900	$2pq$	0,08320
D13S317	9, 10	0,0336	0,0307	$2pq$	0,00206
D16S539	11, 13	0,3140	0,1230	$2pq$	0,07720
D21S11	28, 32.2	0,2460	0,0614	$2pq$	0,03020
AMEL	X, Y				
				RMP	$4{,}24 \times 10^{-19}$

Les fréquences d'allèles utilisées correspondent à celles retrouvées dans la population caucasienne ("Appendix 1 – STR Allele Frequencies from U.S. Population Data," 2015; Butler, 2015c).

1.5.5 Déviations de l'équilibre Hardy-Weinberg et ajustement de cet impact

La déviation de l'équilibre de Hardy-Weinberg est la conséquence de trois phénomènes (National Research Council, 1996). D'abord, cela peut être causé par un accouplement qui n'est pas complètement aléatoire parmi les individus de la population. En effet, il n'est pas rare que des parents partagent des ancêtres communs (coancesterie). Les parents ont donc tendance à transmettre aux enfants les mêmes allèles, ce qui engendre davantage d'homozygotes au sein de la population (D. J. Balding et Nichols, 1994). Ensuite, la subdivision des populations en fonction des groupes ethniques majeurs implique nécessairement que des individus sont liés et possèdent des ancêtres communs. Finalement, la déviation peut être causée par la sélection naturelle, s'expliquant par les différents taux de survie et de reproduction en fonction des génotypes.

Bien que le modèle de probabilité d'un génotype se dit basé sur l'équilibre de Hardy-Weinberg (HWE) et l'équilibre de liaison (LE), ces hypothèses sont contradictoires en raison des fréquences alléliques qui diffèrent au niveau des sous-populations. Le modèle utilisé expose donc un certain déséquilibre. Ainsi, lorsque les bases de données présentent des variations majeures au niveau de leur composition causées par un déséquilibre de Hardy-Weinberg et une subdivision de la population, les calculs de probabilité de concordance fortuite sont affectés, ce qui engendre des valeurs qui sont sous-estimées (D. J. Balding et Nichols, 1994; Krane, Allen, Sawyer, Petrov, et Hartl, 1992; Lewontin et Hartl, 1991). Pour inférer des probabilités conditionnelles conservatives, il importe non seulement d'attribuer le groupe racial pertinent, modélisé selon la génétique des populations, mais également d'apporter un facteur corrigeant la structure détectée, soit le paramètre θ (Earl et Vonholdt, 2012).

Pour contrer ce déséquilibre, l'ajustement de la structure des sous-populations (paramètre θ) a toutefois établi avoir un impact significativement conservateur, démontrant n'affecter que faiblement la performance du modèle (Buckleton, Curran, et Walsh, 2006). Ainsi, l'ajustement θ correspond à une estimation empirique de la coancesterie reflétant la structure des différentes sous-populations. Ce paramètre vient corriger l'impact de la subdivision de la population et permet de compenser le risque de

sous-estimer les probabilités de concordance fortuite calculées. Puisque les bases de données ne reflètent pas avec précision la population d'intérêt à l'enquête, l'incertitude est ainsi atténuée par θ (Steele, Court, et Balding, 2014). L'ajustement permet de calculer plus adéquatement le profil génétique qui s'y trouve influencé (Butler, 2015b). Cette valeur θ varie de 0,01 pour une population typique et relativement homogène, à 0,03 pour une petite population isolée. Comme il existe de nombreux groupes raciaux avec différents niveaux de stratification, il devient intéressant de les comparer dans un contexte forensique (D. J Balding et Nichols, 1995)

Ainsi, la recommandation 4.2 du NRC indique que si la sous-population de référence pour l'échantillon est inconnue ou les valeurs de fréquences alléliques pour la sous-population ne sont pas disponibles, il suffit de calculer (**Équations 1.3 et 1.4**) pour le groupe racial majeur la fréquence de chaque génotype pour tous les loci, puis de les multiplier ensemble ("Appendix 2 – NRC I & NRC II Recommendations," 2015).

Les équations 1.3 et 1.4, homozygote et hétérozygote respectivement, permettent de calculer la fréquence d'un profil génétique lorsqu'une base de données de la sous-population n'est pas accessible, et ce, avec la correction de la subdivision (D. J. Balding et Nichols, 1994).

$$P(A_iA_i|A_iA_i) = \frac{[2\theta+(1-\theta)p_i][3\theta+(1-\theta)p_i]}{(1+\theta)(1+2\theta)} \textbf{(1.3)}$$

$$P(A_iA_j|A_iA_j) = \frac{2[\theta+(1-\theta)p_i][\theta+(1-\theta)p_j]}{(1+\theta)(1+2\theta)} \textbf{(1.4)}$$

1.6 Objectifs de la recherche

L'objectif général est de tester un échantillon tiré des traces ADN retrouvées dans l'environnement par les activités humaines, à savoir s'il représente davantage la population d'intérêt à l'enquête, plutôt que l'échantillon de référence actuel (base de données génétiques divisée par groupes ethniques).

Pour ce faire, ce projet vise en premier lieu à remettre en question l'utilisation et la validité du modèle de bases de données actuelles, en comparant un échantillon de recrutement à l'échantillon de référence. D'abord, **l'échantillon de recrutement** a été récemment collecté à l'UQTR, Sherbrooke et Dolbeau-Mistassini et sert de contrôle positif dans cette étude. Ensuite, cet échantillon de recrutement a été comparé à **l'échantillon de référence**, correspondant à une base de données caucasienne datant des années 1990, déjà établie et actuellement utilisée par le laboratoire judiciaire LSJML. La composition génétique (fréquence allélique) des échantillons, la mesure de la différenciation significative, la mesure de la différence quantitative de la rareté des profils génétiques (probabilité de concordance fortuite) et la détection de toute structure de population ont ainsi été analysées. Notre hypothèse est que l'occurrence des fréquences alléliques d'un échantillon contrôle multiethnique (échantillon de recrutement) est indifférenciable d'un échantillon de référence ethnicisé typique des laboratoires judiciaires, suggérant que le modèle actuel n'est pas adéquat pour calculer de manière exacte la rareté d'un profil génétique (probabilité de concordance fortuite).

Cette validation d'un échantillon contrôle (recrutement) constituait une première étape pour éventuellement tester l'alternative d'échantillonner des traces ADN aléatoirement dans l'environnement et donc, 1) mieux définir la population d'intérêt à l'enquête et 2) établir une base de données génétiques de référence à partir de traces environnantes. Pour générer un échantillon de l'environnement, une ville cosmopolite, soit Montréal, a été ciblée pour la collecte de données. Préalablement, il importait de bien définir géographiquement les zones d'échantillonnage, délimitées au moyen de postes de quartiers (PDQ), permettant d'optimiser la démarche de confronter les profils génétiques aux traces environnementales, sous l'hypothèse que l'auteur d'un crime (réalisé dans un PDQ particulier) provient (ou fréquente) ce même PDQ. Un total de 700 points d'échantillonnage déterminés aléatoirement ont été positionnés sur le territoire de la Ville de Montréal, tout en notant les coordonnées géographiques pour cartographier les données. À chaque point, un objet délaissé (verres jetables, gommes, mégots de cigarettes, gobelet, etc.) a été récolté comme source d'ADN. Ces objets ont été sélectionnés dans la mesure où ils étaient susceptibles de contenir des traces suffisantes d'ADN à partir de la

salive, malgré la variable des conditions environnementales qui ne pouvait être contrôlée (Chatterjee, 2019; Lee et Ladd, 2001; Verdon, Mitchell, et van Oorschot, 2014). Éventuellement, l'ADN environnemental sera analysé en laboratoire, en collaboration avec le Laboratoire de sciences judiciaires et de médecine légale (LSJML, Ministère de la Sécurité publique du Québec), dans le but d'obtenir les profils génétiques des spécimens et ainsi faire des analyses comparatives avec l'échantillon de référence et de recrutement.

CHAPITRE II

CONCORDANCE IN CRIMINAL INVESTIGATION: CONNECTING THE EVALUATION OF FORENSIC DNA EVIDENCE TO THE CURRENT POPULATION OF INTEREST

Jessie Beauchemin, Emmanuel Milot, Frank Crispino

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

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 effectué la collecte de données, contribué à générer les résultats, en collaboration avec le Laboratoire de sciences judiciaires et de médecine légale, accompli l'ensemble des analyses, en plus d'avoir rédigé cet article. Pr Emmanuel Milot, directeur de recherche, a élaboré le projet de recherche et a veillé à l'avancement et la réalisation de celui-ci. Pr Emmanuel Milot et Pr Frank Crispino ont tous deux participé à la révision du manuscrit de l'article.

2.2 Résumé de l'article (français)

L'évaluation de l'ADN dans les enquêtes criminelles est nécessaire aux acteurs du système de justice pénale pour obtenir des informations susceptibles d'incriminer les auteurs d'infractions. La norme consiste à calculer le poids de la preuve ADN, et ce, lorsqu'il y a une concordance entre le profil génétique de la trace d'ADN prélevée sur une scène de crime et celui du suspect. Il est alors habituel d'évaluer la rareté de ce profil

génétique au sein de la population d'intérêt, en le comparant à un échantillon de référence selon un groupe ethnique jugé pertinent. Les laboratoires judiciaires disposent de bases de données, qui en plus d'être complètement anonymes, sont divisées en fonction des différents groupes ethniques : Afro-Américains (Noirs), Asiatique, Caucasiens, Hispanique ou Amérindiens. Si cette pratique est justifiable lorsqu'il s'agit d'une population relativement homogène, elle est plus difficilement valable lorsque la population s'avère mélangée. L'objectif de notre étude est donc de tester une nouvelle approche à partir de traces ADN collectées aléatoirement à partir de l'environnement pour 1) davantage définir la population d'intérêt à l'enquête lorsqu'elle est composée de nombreux groupes ethniques hétérogènes et 2) générer un échantillon d'ADN de référence à partir de traces prélevées aléatoirement dans cette population. Pour ce faire, un échantillon de recrutement a d'abord été utilisé comme contrôle positif dans le but d'être comparé à l'échantillon de référence (bases de données du laboratoire judiciaire), questionnant ainsi la validité du modèle actuel divisé par groupe ethnique. Des analyses permettant de mesurer la composition (fréquence allélique) et l'indice de différentiation n'ont détecté aucune différence significative entre les échantillons de recrutement et de référence, n'impactant que faiblement la probabilité de concordance fortuite quant à la rareté des profils génétiques. Ultérieurement, l'échantillon de l'environnement, généré à partir de traces ADN humaines présentes dans la ville de Montréal, sera analysé dans le but de définir un modèle davantage représentatif de la population d'intérêt à l'enquête.

2.3 Article complet (anglais)

Abstract

The interpretation of DNA in criminal investigations is necessary for actors of the criminal justice system to assess assumptions regarding perpetrators of crimes. When there is a match between the DNA trace left on a crime scene and the genetic profile of the suspect, the standard is to calculate the random match probability, which assesses the rarity of this particular genetic profile within a population of interest, allowing the possibility that this concordance occurred by chance. Factually, this statistic weight of evidence is based on the genetic composition (allele frequencies) of a reference DNA sample from a relevant ethnic group. While this practice is justifiable when dealing with a relatively homogeneous population of reference, it is more difficult when the population is composed of numerous mixed ethnic groups. To test this assumption against current used databases, we propose to geographically define the relevant population of interest by sampling random human DNA traces present in the environment of Montreal city. Firstly, we aimed question the validity of current reference databases from forensic laboratories. Hence, a recent recruitment sample (from UQTR, Sherbrooke and Dolbeau-Mistassini) used as a control was compared to the reference database (a Caucasian sample) from the Laboratoire de sciences judiciaires et de médecine légale (LSJML, Québec). Regardless of the sample used (reference or recruitment), the genetic composition, i.e., allele frequencies, tends to be similar, both samples did not reflect differentiated populations and they only slightly (almost negligible) affected the rarity of a genetic profile. Finally, no structural difference in genetic variance was detected. Results support our assumptions questioning the validity of databases divided by ethnic groups, since the recruitment sample tend to be equivalent to the reference sample currently used by LSJML. This study represents a first step to testing an environment sample as a genetic database, by truly defining the population of interest through environment DNA with the aim to accurately calculate the statistical weight of a DNA profile.

Keywords: Random Match Probability, DNA Database, Genetic Profile, Allele Frequency, Reference Sample

Introduction

In forensic science, DNA traces are clues that can be decisive in solving a criminal investigation. To obtain genetic profiles from traces left on a crime scene, genetic markers known as short tandem repeats (STRs, microsatellites), found on autosomal and sex chromosomes, are used (Vieira, Santini, Diniz, & Munhoz, 2016). STR alleles differ from each other by their number of repeats of a short motif (typically 4-5 bp). STRs exhibit great allelic variability because of their high mutation rate, providing great discriminatory power measured by the probability that the genetic profiles of two individuals randomly drawn from the population are different at least for one allele. It allows targeting the source of the trace to the exclusion of other possibilities, although it does not guarantee certainty (Jamieson & Bader, 2016, pp. 32-34, 69-70; Smalldon & Moffat, 1973). A match between a crime scene DNA profile and that of a *person of interest* (POI, e.g., a suspect) must be weighted probabilistic (National Research Council, 1992).

When there is a match between two DNA profiles (e.g., from a crime scene trace and a suspect), the weight of this evidence must be assessed. This involves determining the probability that the DNA match is purely fortuitous (D. J. Balding, 1999). In fact, it cannot be excluded that the POI has the same genetic profile as the DNA trace by pure chance, although he is not the source of the trace (Butler, 2015a; SWGDAM, 2017). Thus, the random match probability (RMP) is the probability of randomly drawing an individual (other than the POI) in the population with a genotype corresponding to the one already observed on the DNA trace (D. J. Balding, 1999; Curran, Walsh, & Buckleton, 2007).

Probabilistic inference of DNA weight-of-evidence is based on allele frequencies estimated from a DNA reference database, presumably appropriate for the population of interest in a given criminal case. The size of a DNA database must not only be sufficiently large enough to represent all possible alleles, but also to ensure that the weight of evidence is not biased in disfavour of the defendant (Chakraborty, 1992; R Lempert, 1993). By using a population database, the frequency of any DNA profile can be calculated from estimated allele frequencies, conditional on the assumptions of the population genetic

model used for that purpose, such as the classic “random mating” model, which implies Hardy-Weinberg and linkage equilibria or coancestry model of Balding and Nichols (1994), which assumes that the population genetic structure is correctly described by a Dirichlet distribution with parameter theta (i.e. random mating within local subpopulations and reduced mating between them, with all subpopulations being genetically equidistant from one another). Therefore, the RMP depends on the frequencies (rarity) of the various variants (alleles) of a gene existing in the population of interest (Curran et al., 2007; National Research Council, 1996).

DNA databases are composed of samples from voluntary donations or convenience samples: diverse sources such as blood banks, laboratory personnel, law enforcement officers, etc. (Lapointe, Rogic, Bourgoin, Jolicoeur, & Séguin, 2015; National Research Council, 1996). Typically, forensic laboratories have separate reference DNA samples for different ethnic groups, such as African Americans, Caucasians, Hispanics, East Asians and Native Americans (Budowle, Shea, Niezgoda, & Chakraborty, 2001). Assigning (assuming) ethnicity to the source of a DNA trace allows one to calculate the probability (expected frequency) of a particular genetic profile more accurately (National Research Council, 1996). However, in a criminal investigation, if the identity of the person who left a DNA trace on a criminal scene remains unknown, the ethnicity of this individual is assigned on the basis of information gathered by investigators, whenever possible. Moreover, it is sometimes difficult to classify an individual with mixed ancestry (Lowe, Urquhart, Foreman, & Evett, 2001). In those cases, because the samples compiled in the genetic database are not randomly sampled based on geographic considerations, it cannot accurately target the right population of interest regarding criminal investigation (D. J. Balding & Nichols, 1994).

In a forensic context, the population of interest refers to a group of individuals from which the author of a crime is likely to come from, often limited to a geographical area. Depending on the circumstances of the investigation (e.g., where the crime was committed?), the population of interest then generally consists of individuals frequenting that particular area (D. J. Balding & Nichols, 1994). However, it is possible that the

perpetrator comes from a different area from where the crime was committed. How then to assess the relevant database of people of interest, while ethnic model database fitting the spatial area of interest is questionable?

The current ethnocentric way of setting up a DNA database is simplistic and poses two problems. First, DNA databases used by forensic laboratories have uncertainties in their composition and should represent at best the genetic complexity existing at several levels of substructure (Curran, Buckleton, & Triggs, 2003). Since samples come from voluntary donor's index, the reference sample is unlikely to be always representative of the population of interest (D. J. Balding & Nichols, 1994; Gouvernement du Canada, 1998). Although there can be several ethnic groups living in the same geographical area, strictly speaking forensic databases are valid for estimation of the weight of evidence only when dealing with ethnically homogeneous populations (D. J. Balding & Nichols, 1995; National Research Council, 1996). Consequently, the population of interest is poorly represented by these samples in admixted populations, a problem likely exacerbated in cosmopolitan cities. A genetically heterogeneous local population won't meet the assumptions of the models commonly used in forensic expertise. Little is known about how this impacts the accuracy of the weight of DNA evidence presented in courtrooms and the validity of its interpretations (D. J. Balding & Nichols, 1994).

The second problem with ethnocentric databases is that they do not really answer the question for which they are constituted: what is the genetic composition of the DNA traces constantly left in the environment by daily human activity? In the absence of investigative hypotheses concerning the perpetrator of a crime, it becomes relevant to compare DNA traces from the crime scene to a reference sample of DNA traces collected in the environment within the spatial area occupied by the population of interest.

In the past 20 years, some authors tried to generate more adequate databases, considering ethnicity and subpopulations factors. To counter the problem of non-representativeness in cosmopolitan populations, Gill, Foreman, Buckleton, Triggs, and Allen (2003) sampled 24 different populations and combined them to generate a more

conservative DNA database. The problem of accurately representing ethnicity of the offender databases was raised by Kruijver (2016), whose approach avoids small-scale sampling and possibly unrepresentative population surveys. His proposition implied allelic frequencies of the subpopulations and their proportion can be estimated by the composition of the database itself, based on the expectation maximization (EM) algorithm. He demonstrated that three subpopulations (European, African and Asian) were sufficient to explain the heterogeneity in the database by the major extent of their clustering. He also confirmed that the theta parameter of 0.03 was sufficiently conservative regarding random match probability calculations.

The objective of this project was to test and compare a new approach to 1) better define the population of interest in the survey, whatever the potential mixture of numerous ethnic groups, and to 2) propose a more representative DNA database from traces randomly collected in the environment. To do so, this project aimed to compare different samples. Firstly, a volunteer recruitment sample (collected from UQTR, Sherbrooke and Dolbeau-Mistassini) was used as a control to validate the use of the reference database already established and currently used by the LSJML, and whether it is still a valid and representative sample to date. Secondly, an environment sample from traces naturally present in the environment of Montreal City was collected from several areas. Random geographic coordinates were generated and objects susceptible to contain DNA traces were collected. Subsequently, DNA will be analysed in order to obtain the genetic profile (STR alleles) of unknown individuals. This environmental sample will be later studied, with the aim of comparing the differences in allelic occurrences, the rarity of the genetic profiles and population genetic structure from the reference and the recruitment samples.

Methods

Data used for the study

Reference and recruitment samples

The first one is the **reference database** currently in use at the Laboratoire de sciences judiciaires et de médecine légale (LSJML), which provides forensic expertise for police forces Québec (Gouvernement du Québec, 2013). This database is composed of the genetic profiles at the 15 STR markers of the Identifiler+ kit (Applied Biosystems) for 276 Caucasians, collected in the early 1990s in the cities of Montréal and Chicoutimi (**Figure 2.1)**. The **recruitment sample** (*n*=126) was done by UQTR, and recently collected in three cities of Québec: Trois-Rivières (UQTR), Sherbrooke and Dolbeau-Mistassini. Genetic profiles of the recruitment sample were already provided. It is expected that its composition of mixed Caucasian, Natives and international. The **environment sample** (*n*=693) was collected from objects found in the urban environment of Montréal (see next section). We first compared the UQTR recruitment and LSJML Caucasian reference samples. The rationale is that the latter was collected nearly 30 years ago (~1 human generation). Thus, before comparing it to the contemporary environmental sample, we wanted to verify that the LSJML sample is still representative of the population it aims to characterize.

Figure 2.1 Map of the province of Québec with the corresponding locations for the reference sample (Montréal City [Montréal region] and Chicoutimi, [Saguenay region], blue stars) and the recruitment sample (Trois-Rivières, Sherbrooke and Lac-Saint-Jean region, red stars) (Gouvernement du Québec, 2019).

Environmental sample

The **environment sample** was drawn from Montréal City, which is part of the urban community of municipalities occupying the island of Montréal. This city is the metropolis of Québec and was chosen because of its cosmopolitan and multi-ethnic composition that we sought for this study.

According to sociodemographic data, 1.7 million people lived in Montréal City in 2016, spread over an area of 365 Km^2 (Ville de Montréal, 2016). We used public data on local police stations (*postes de quartiers*, PDQ) and administrative boundaries to generate maps for sampling purposes (https://donnees.montreal.ca/ville-de-montreal/polygones-arrondissements#resource-limite_administrative_de_l'agglom%C3%A9ration_de_montr%C3%A9al). Maps were created using R software (R Foundation for Statistical Computing, version 3.6.3), and shown in see **Figures 2.2** and **2.3**.

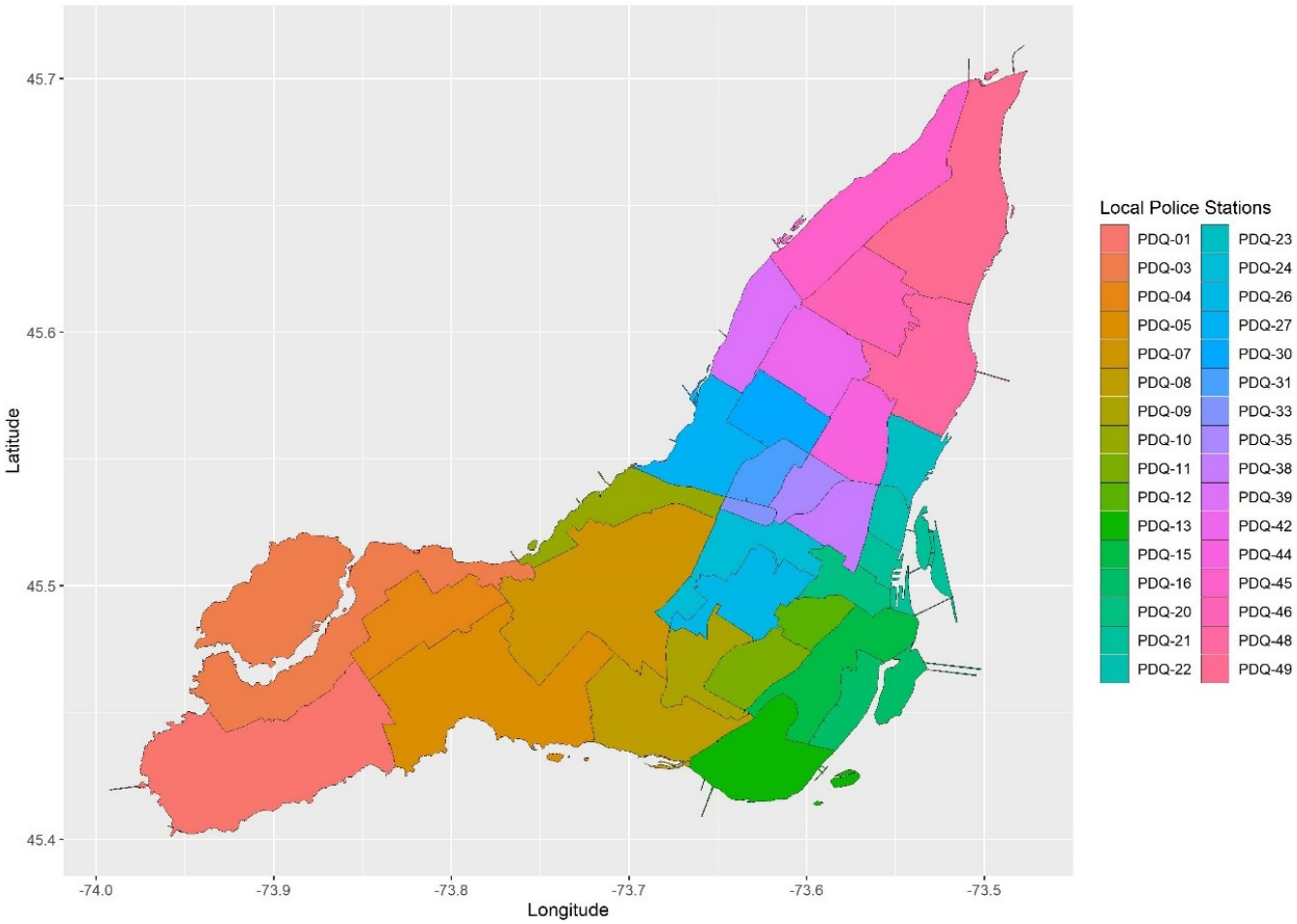


Figure 2.2 Map showing the boundaries of local police station (PDQ) jurisdiction on the island of Montréal.

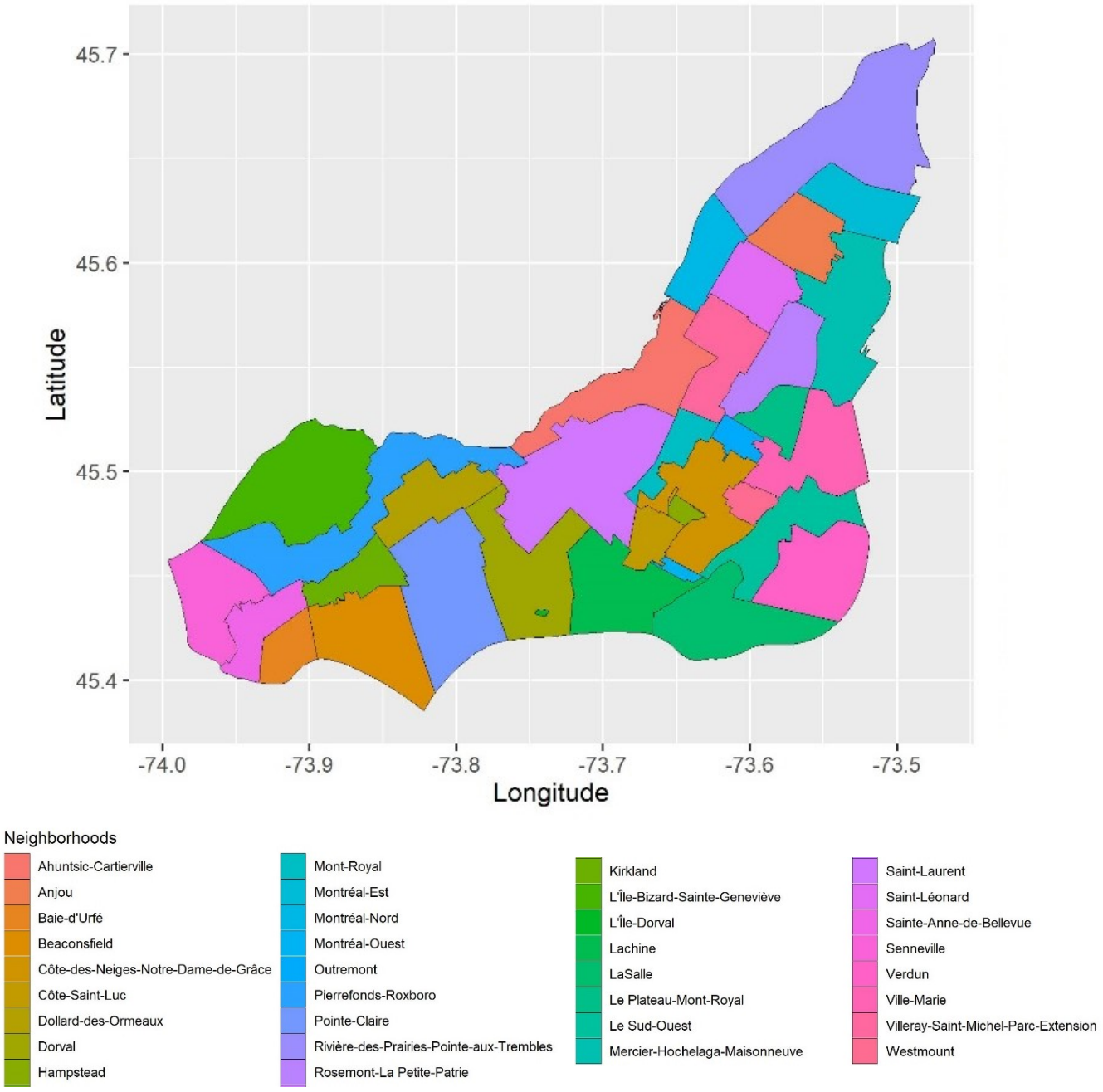


Figure 2.3 Map of Montréal City with administrative boundaries.

The territory of seven PDQs was selected to collect environmental DNA samples, (listed in **Table 2.1)**. Approximately 100 specimens were collected in each PDQ's territory, providing per-PDQ samples of *high-density* coverage (8.3 specimens/Km^2). Note that PDQ 22 and 23 were combined together as they consisted of lower contiguous areas (**Table 2.1**). A lower density sampling (0.7 specimens/Km^2) was also conducted over a broader area, composed of the grouping of all PDQs in the eastern half in Montréal City, including the seven aforementioned, plus 11 others, PDQs 20, 21, 24, 27, 30, 31, 35, 42, 45, 46, 49 (here in designated as PDQALL). The reasons for having two sampling scales were: 1) to have a fine-grained spatial picture of allele frequencies, yet obtained at the cost of substantial local sampling effort; 2) to measure the amplitude of allelic frequency variation at small spatial scale; 3) to assess whether a lower density sampling, which allows to cover a greater area at less cost, can nevertheless provide a good estimation of allele frequencies; and 4) to assess whether a mix of high and low density sampling could be a good trade-off between sampling effort and precision.

Table 2.1 Local police station's (PDQ) territory sampled and the number of specimens collected.

Neighbourhood name	PDQ	Number of specimens
Montréal-Nord	PDQ39	100
Côtes-des-Neiges, Mont-Royal, Outremont	PDQ26	99
Plateau Mont-Royal	PDQ38	98
Rosemont, La Petite-Patrie	PDQ44	99
Centre-Sud	PDQ22	35
Hochelaga, Maisonneuve	PDQ23	66
Boroughs Mercier, Hochelaga, Maisonneuve	PDQ48	99
Montréal Est (surrounding all PDQ of interest)	PDQALL	97

QGIS (version 3.16.3) was used visualize and analyze geospatial data for Montréal City **(Figure 2.4A)**. In the selected PDQs **(Figure 2.4B)**, each sample was collected by randomly drawing 100 geographic coordinates and plotting them on a map **(Figure 2.4C)**. Those coordinates were transferred to a GPS receiver (model eTrex Vista HCx from Garmin) to locate the sampling points on the field. Then, a single specimen was collected within a maximum 100 meters radius of a given set of geographic coordinates generated. The radius was fixed to maximise the chances of finding an appropriate specimen, taking into account that streets and buildings occupy spaces where it is more difficult, when not impossible to search for specimens. If a sampling point fell on a private property or other inaccessible place, so that it was not possible to collect an object within the 100 meters radius, then a new set of coordinates was randomly drawn. To avoid sampling points too close from one another, the minimum distance between their coordinates was determined at 100 meters. The World Geodetic System (WGS 1984, EPSG:4326) was the spatial reference system used in this study, to set sampling locations and for GPS navigation (National Geospatial-Intelligence Agency, 2022).

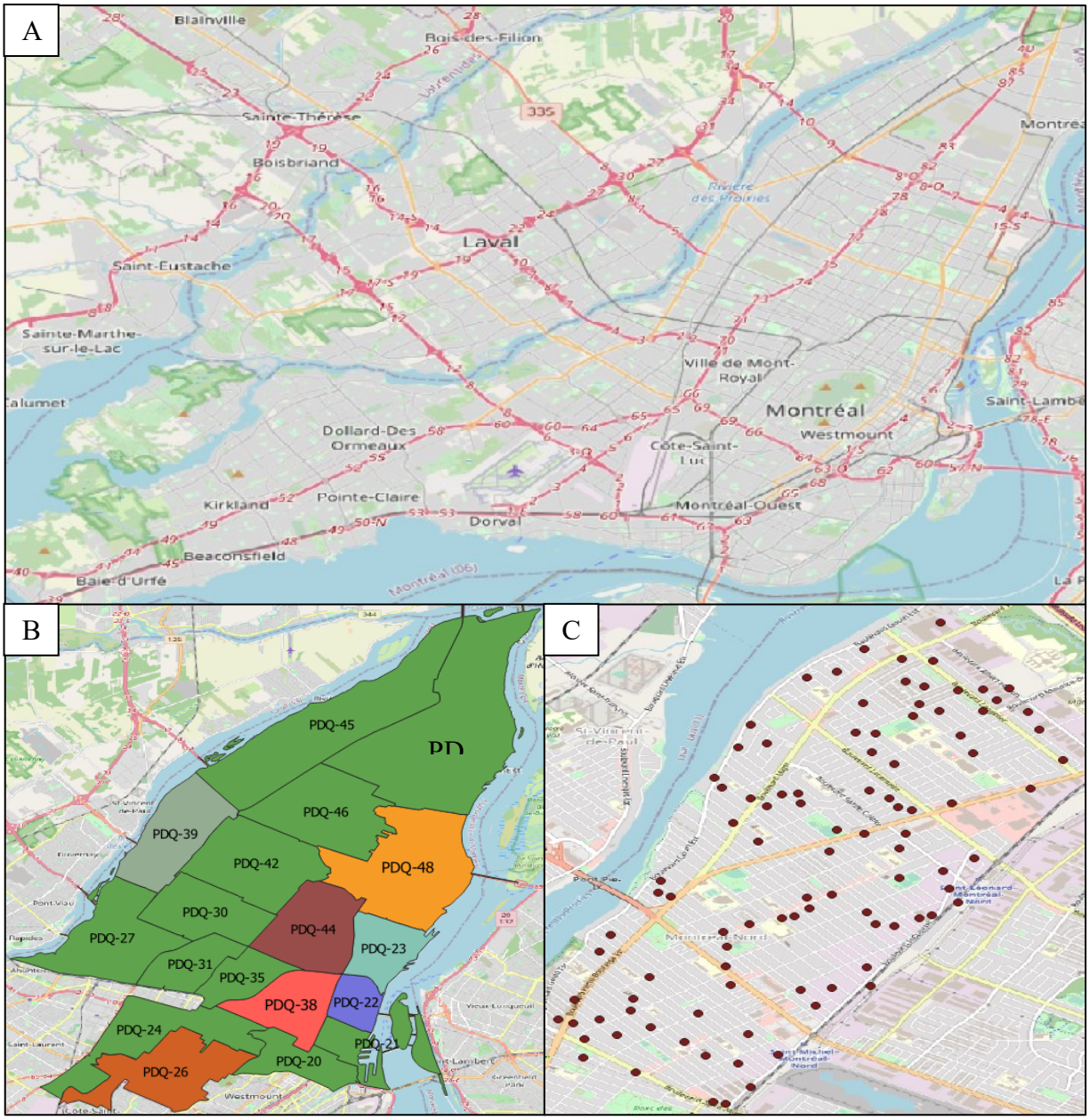


Figure 2.4 Geographical map showing the study area and sampling sites, generated with the WGS 1984, EPSG:4326 system. (A) The island of Montréal. (B) PDQs targeted (green area corresponds to the area designated as PDQALL and regroups 11 others PDQ of lower density). (C) Example of PDQ39 with 100 different geographical coordinates randomly generated by QGIS, distanced by at least 100 m from each other.

Items that were likely to contain good quality DNA from a single (or dominant) donor were prioritized. This is because mixed genetic profiles would imply challenging STR interpretation, due to the uncertainty in the number of contributors (Ge, King, Smuts, & Budowle, 2021). Since the objective was to calculate the number of copies of each allele in profiles from environmental specimen, single source DNA limits uncertainty over the number of allele copies. Thus, by measuring the concentration of the specimens, it was possible to mix DNA profiles by two (with equal concentration) in order to facilitate the interpretation of STRs from balanced mixtures. This was particularly important considering the *in vitro* DNA mixing approach taken to avoid obtaining as end results single-person profiles, for an ethical reason (see below). Types of specimens collected in the field are listed in **Figure 2.5**. Using new sterile nitrile gloves each time, specimens were individually packaged in a paper bag then sealed with tape, to limit contamination (Butler, 2005; Lee, Gaensslen, Bigbee, & Kearney, 1991). Packages were identified with specimen number and type, collection date and time, geographic coordinates and approximate distance in meters from the sampling point center (ranging from 0 to 100 m).

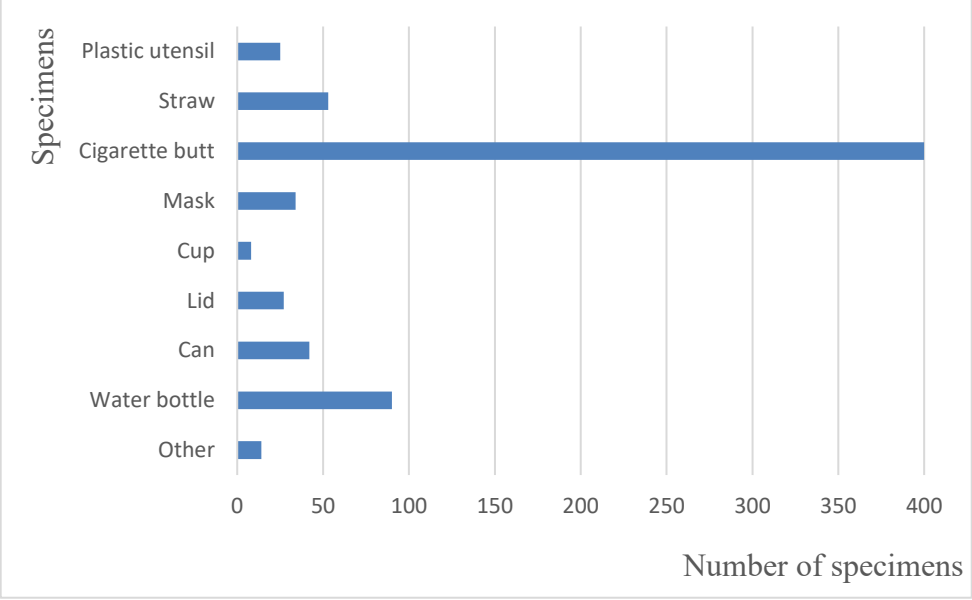


Figure 2.5 Types of specimens collected for the sampling in Montreal City.

Generation of STR profiles

A same method to generate STR profiles, from DNA extraction to allele calling, was applied to all three samples (LSJML reference, UQTR recruitment, environmental). The steps are described in the following sections. Note that LSJML staff had performed all steps for its reference sample and for the UQTR recruitment sample. DNA from environmental specimens was extracted and quantified at UQTR by JB and subsequent steps were done at the LSJML.

DNA extraction from the environmental sample

Trace sampling of each specimen was adapted to the nature of the object. For cigarettes a ~1 x 1 cm piece of the butt paper was cut with bleach-sterilized tools, taking care not to transfer fibers from the filter itself into the vial. Other objects were swabbed with a cotton-tipped applicator slightly moisten with sterile water (Medicom^{TM}) in areas most likely to contain DNA traces. All samples were placed into individual 1.5 mL microtubes. DNA was extracted using the DNA IQ^{TM} kit (Promega Corporation), designed for forensic purposes. Briefly, the extraction consisted of a 2 h digestion at 56 °C with proteinase K (Thermo Fisher Scientific). The DNA molecules were then captured by adding the magnetic beads resin (14 µl) to the samples, then the tubes were placed on a magnetic support to separate the DNA bounded to the resin from the cell lysate. Washing buffer was added to remove contaminants or PCR inhibitors while retaining the purified DNA. The elution buffer was added, and tubes were heated on a heat block for 5 min at 67 °C to release DNA from the resin. The volumes and concentrations of buffers followed the protocol provided with the DNA IQ kit (Promega, 2016).

Quantification of extracted DNA

Quantification by qPCR was used to determine the amount of amplifiable human DNA in extracted samples. This step was necessary to determine the appropriate amount of the DNA sample to be added to the subsequent PCR amplification of STR markers.

The Quantifiler™ Trio kit (Thermo Fisher Scientific) targets human-specific loci: small autosomal for detecting total human genomic DNA, large autosomal as an indicator of DNA degradation, and Y-chromosome for detecting human male genomic DNA. The quality index is measured by the large/small autosomal DNA proportion. The kit uses TaqMan dye-labeled probes that first anneal to the targeted marker and are then chopped by the Taq polymerase during the PCR, thereby causing fluorescence emission. Thus, Quantifiler THP PCR reaction mix and Primer Mix (probes ABY, JUN, VIC and FAM) were added to the samples, and the reaction was performed on the Applied Biosystems QuantStudio™ 5 Real-time PCR machine. The volumes of THP PCR reaction mix, Primer mix and DNA samples used in reactions followed the protocol provided with the Quantifiler™ Trio kit (Thermo Fisher Scientific Inc, 2018b).

Amplification of forensic STR markers from DNA samples

Once the concentration of the samples had been determined, samples were sent to the LSJML. The laboratory staff performed DNA amplification using the AmpFLSTR™ Identifiler™ Plus kit (Thermo Fisher Scientific). Based on the PCR amplification protocol of denaturation, hybridization and elongation (see **Table 2.2**), multiple copies of short tandem repeats (STRs) were generated. This kit amplifies 15 loci with a tetranucleotide repeat, as well as the amelogenin sex-determination marker. The 16 STR markers used are: D21S11, CSF1PO, Vwa, D8S1179, TH01, D18S51, D5S818, D16S539, D3S1358, D2S1338, TPOX, FGA, D7S820, D13S317, D19S433 and Amelogenin. Identifiler Master Mix and direct Primer Mix were added to the samples, which were analysed on the GeneAmp™ PCR System (Thermo Fisher Scientific). The volumes of Identifiler Master Mix and primers followed the protocol provided with the kit (Thermo Fisher Scientific Inc, 2018a). For ethical reasons, we wished to avoid ending up with the pure profile of unknown individuals in our data. Thus, we mixed two or three samples with similar concentrations prior to the PCR step. This way, we expected to have good quality genetic mixture profiles from which the number allele copies could be determined in most cases.

Table 2.2 Thermal cycling conditions of PCR amplification, from the AmpFLSTR™ Identifiler™ Plus user guide (Thermo Fisher Scientific Inc, 2018a).

Number of cycles	Steps	Temperature	Time
1X	Initial incubation	95 °C	11 min
28-30 cycles	Denaturation	94 °C	20 sec
	Hybridation/ Elongation	59 °C	3 min
1X	Final elongation	60 °C	10 min

Detection of PCR amplicons by capillary electrophoresis

PCR products were migrated on a capillary analyser of model 3730 (Applied Biosystems) at the LSJML. STR alleles were then called using the GeneMapper™ ID-X Software (version 1.6). The genetic profiles were then manually checked one by one to confirm the identity of each allele.

Data analysis

Due to problems occurring during the DNA extraction phase for environmental specimens, along with time constraints, genetic profiles for the environment sample could not be included in the results presented in this thesis (but will be in the future). Actually, no or very few alleles were detected for 26 environmental DNA samples tested. The origin of this problem has not yet been identified, despite re-testing the same samples in multiple conditions. Therefore, the present chapter reports the outcome of the comparison made between the UQTR recruitment sample and the LSJML reference sample.

We calculated allele frequencies for the reference and the recruitment samples. We tested whether per-locus genotype frequencies met those expected under Hardy-Weinberg equilibrium (HWE). Departure from HWE could be due to population subdivision or non-random sampling. Then, we compared metrics for the weight of

evidence obtained independently from reference vs. recruitment sample alleles frequencies. Finally, we conducted a clustering analysis to detect unsuspected population genetic structuring that could occur. These analyses are detailed next.

Allele frequencies and fixation index

First, by using the R implementation of the *Genepop* software (version 4.7.5, Raymond & Rousset, 1995b), the genetic composition and the fixation index were analysed through the **Basic info test** and **F_{ST} test** respectively. **Basic info test** provided alleles and genotype frequencies per locus (15 loci total), for each sample (reference and recruitment). These data were plotted in bubble graphs to better visualize and compare the genetic composition of these two samples. The **F_{ST} test** provided a measure of the fixation index, based on the stepwise mutation model appropriate for STR data. According to Weir and Cockerham (1984), it measures the level of gene correlation between individuals from the same subpopulation relative to that expected under the HW model. The fixation index can range from 0 to 1, where 0 means no evidence for population genetic structuring while 1 means no sharing of genetic material between subpopulations (i.e. fixation of each subpopulation for different sets of allele) (Raymond & Rousset, 1995a).

Random match probability

We assessed how the random match probability may be impacted by the population sample (reference or recruitment) selected to estimate allele frequencies. So, R software (Script developed by Emmanuel Milot) allowed us to randomly assign alleles for all 15 STR locus, to generate virtual genetic profiles based on allele frequencies from our samples. In order to measure the quantitative difference in the rarity of genetic profiles, random match probabilities for virtual multilocus STR profiles were calculated assuming a departure from *HWE,* where the F_{ST} parameter (theta of 0.01) corresponded to the conservative value recommended by the National Research Council (1996). The minimum allele frequency allowed to compensate for some alleles (sometime rare alleles) not sufficiently sampled or represented in the database. If an allele observed on a genetic

profile had an estimated frequency smaller than 5/2N, where N corresponds to the number of individuals in the population sample, its frequency was set to the minimal threshold (Bodner et al., 2016; National Research Council, 1996). According to our recruitment and reference sample respectively, the number of individuals were set to 126 and 276, which leads to a minimum allele frequency of 0.0198 and 0.0091.

A total of 1,000 virtual genetic profiles were generated based on allele frequencies estimated from the UQTR recruitment sample. Another 1,000 profiles were created likewise based on allele frequencies estimated from the LSJML reference sample. For each of those 2,000 profiles, the random match probability was then calculated using frequencies for both the recruitment and reference samples. The $\log_{10}$ of the ratio of these two probabilities (represented by d_{rec} and d_{ref}) measures how much more probable a genotype is in the population according to one sample relative to the other (**Equations 2.1 and 2.2**), as initially proposed by Gill et al. (2003) and cited in Hessab, Aranha, Moura--Neto, Balding, and Schrago (2018). Log ratios are presented in **Equations 2.3** and **2.4**.

$$d_{rec} = \log_{10} \frac{Recruitment}{Reference} \quad \textbf{(2.1)}$$

$$d_{ref} = \log_{10} \frac{Reference}{Recruitment} \quad \textbf{(2.2)}$$

$$\text{Log ratio} = \log d_{rec} \quad \textbf{(2.3)}$$

$$\text{Log ratio} = \log d_{ref} \quad \textbf{(2.4)}$$

Genetic structure

The Structure software was used to detect potential unsuspected genetic population structuring (J. K. Pritchard, Stephens, & Donnelly, 2000; Jonathan K Pritchard, Wen, & Falush, 2010). The method uses the profile data for the 15 autosomal STRs and fits a maximum likelihood-based model to infer the most likely number of genetic clusters in HWE represented in the sample (J. K. Pritchard et al., 2000). To that end, individuals are initially randomly assigned to *K* pre-determined clusters (*K* being chosen by the user).

Then, during MCMC (Markov chain Monte Carlo) iterations, individuals are re-assigned to the group to which their STR profile is more likely to be observed according to the transient allele frequencies in the *K* group, until the analysis converges toward the best clustering of the sample (Porras-Hurtado et al., 2013).

All samples (reference and recruitment) were pooled together, totalling 404 individuals. Before running Structure, some prior parameters need to be set (Jonathan K Pritchard et al., 2010). The population **admixture model** was selected because it allows for the existence of individuals with mixed ancestry, which is realistic. Also, we chose the **correlated allele frequency model**, allowing the genetic composition across populations to be similar. This is in line with the population model underlying the calculation of RMP in forensic cases (i.e. incorporating the theta parameter). It enables a better detection of subtle population structures. **Parameter *K*** (number of populations) was set from 1 to 6. Three separate **simulations** were conducted with each *K* values, for a total of 18 structures analyses. The **burnin** corresponds to the number of MCMC iteration done before collecting any data from the posterior distribution and was set to 5,000. The **number of MCMC iterations after burnin** was set to 50,000. These values were set after running several tests and observing consistent results, in order to reach a stationary plateau (approximately), as recommended by Jonathan K Pritchard et al., 2010.

Because sets of allele frequency characterize a particular population, the model denoted **X**; the observed genotypes of the sampled individuals and ***K***; the number of clusters/populations, then, an estimate of the posterior probability on the inference of *K* was given by Pr(X|*K*) (J. K. Pritchard et al., 2000). This posterior probability is reduced to log likelihood of the data, called "LnP(D)", also referring to L(*K*) (Evanno, Regnaut, & Goudet, 2005).

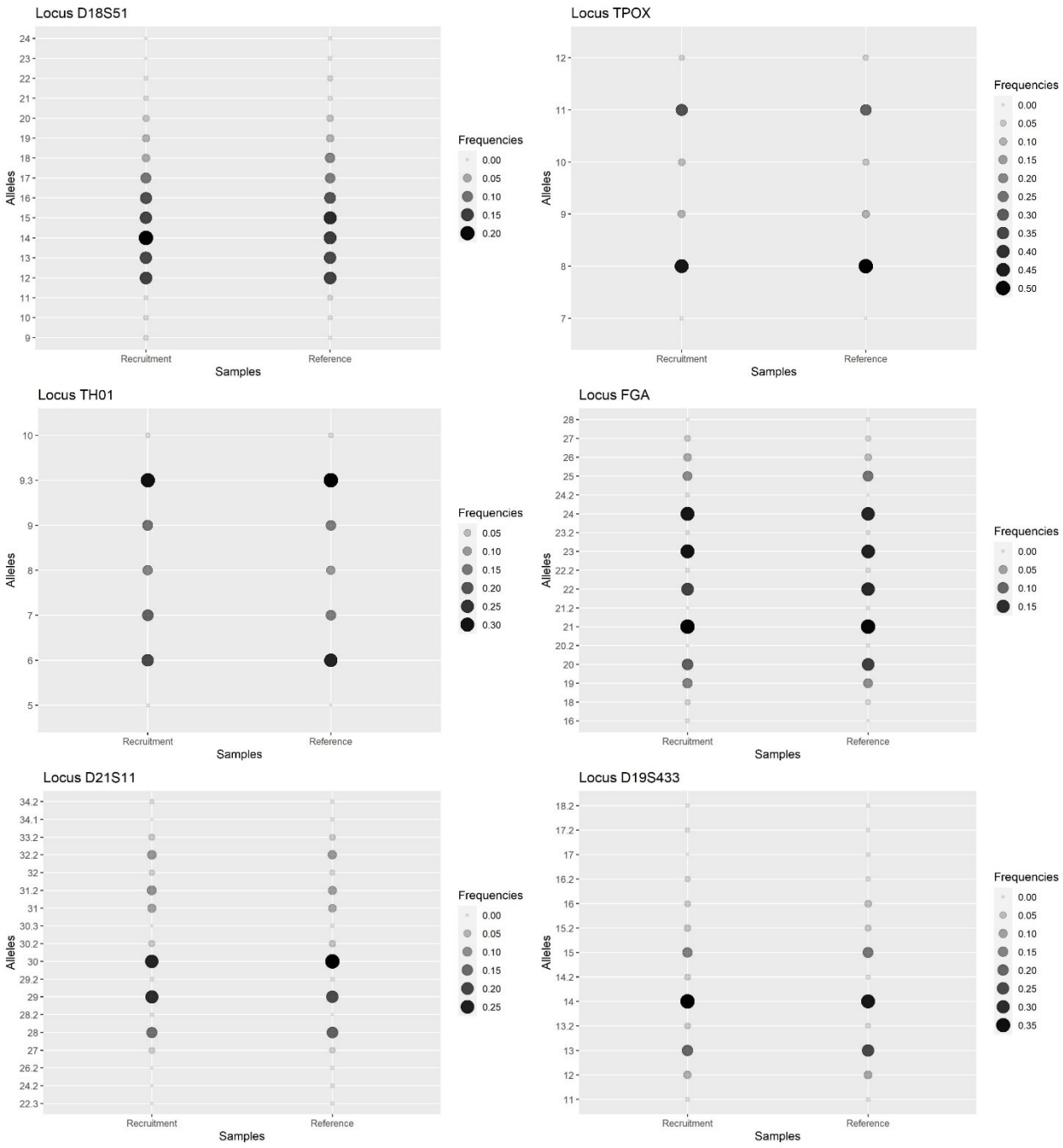
Results

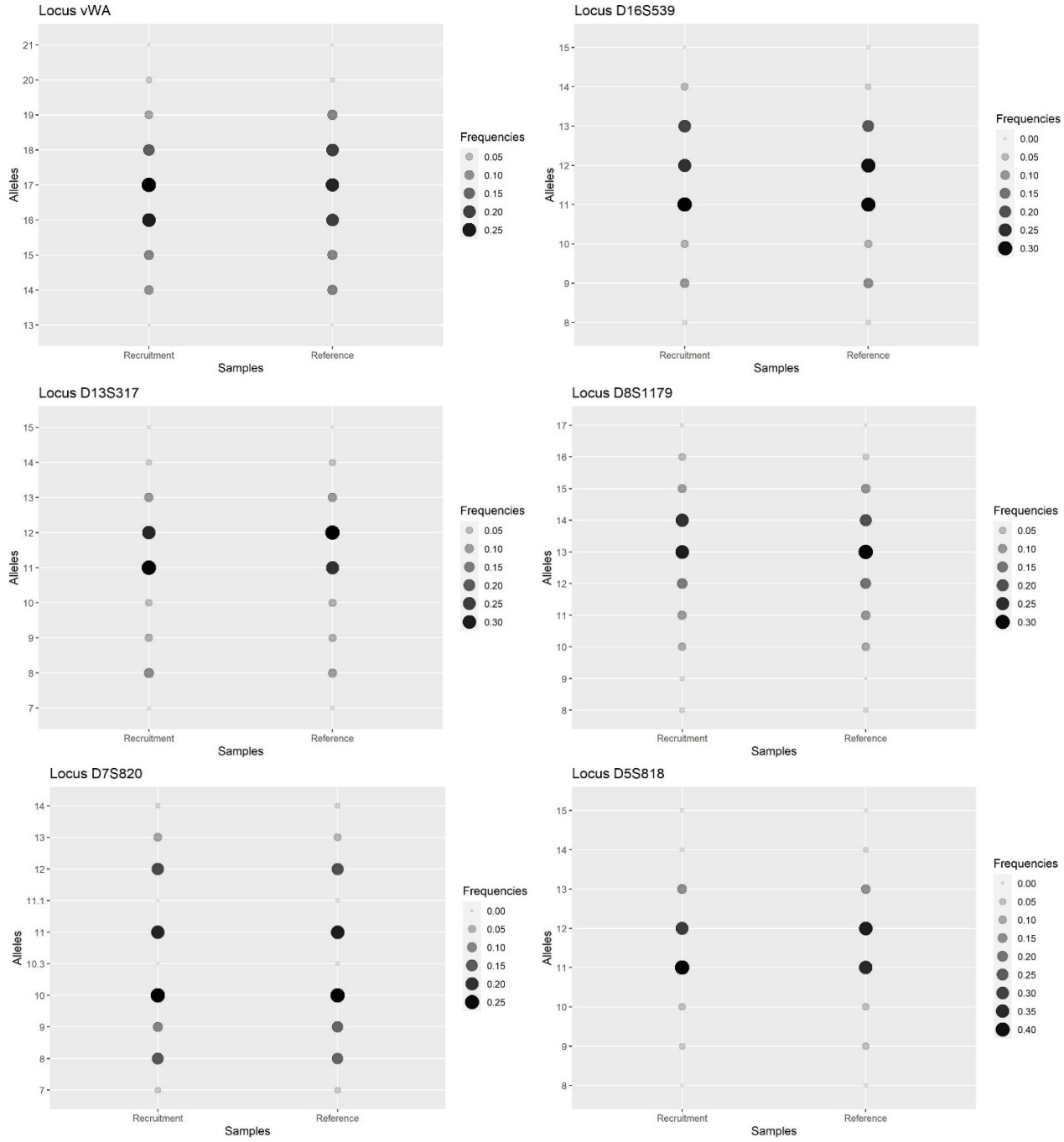
Allele frequencies

Allelic composition of the two population samples (reference and recruitment) exhibits frequency differences (**Figure 2.6**). The maximal difference observed in allele frequencies is 8.3% among the 15 loci (**Figure 2.6**). Loci **D21S11** (allele 30), **TH01** (allele 6), **D16S539** (allele 12), **D19S433** (allele 13), **TPOX** (allele 8) and **D18S51** (allele 14) show significant differences in allele frequencies observed at the level of a single allele, values among these 6 loci ranging from 5% to 5.9% (**Table 2.3**). Also, loci **D3S1358** (allele 15 and 16), **D13S317** (allele 11 and 12) and **D5S818** (allele 11 and 12) show differences in allele frequency observed at two alleles, values among these 3 loci ranging from 6.1% to 8.3% (**Table 2.3**). All frequency data are available as Supplementary Material **Table 2.S1**. Out of 15 loci, 11 have at least one allele absent (frequency of 0) in one of the two samples, either reference or recruitment, while it was detected in the other (**D21S11**, **D7S820**, **CSF1PO**, **D3S1358**, **D16S539**, **D2S1338**, **D19S433**, **TPOX**, **D18S51**, **D5S818** and **FGA**). However, the frequency of these alleles does not exceed 0.7%.

Table 2.3 Loci showing differences in allele frequency greater than 5% (when significant at one or two alleles) between the two population samples (recruitment and reference). Data for all loci and alleles can be found in the Supplementary Material **Table 2.S1**.

Locus	Alleles	Differences in allele frequencies
D8S1179	-	-
D21S11	30	5.0%
D7S820	-	-
CSF1PO	-	-
D3S1358	15 16	6.2% 7.3%
TH01	6	5.6%
D13S317	11 12	7.1% 6.2%
D16S539	12	5.2%
D2S1338	-	-
D19S433	13	5.4%
vWA	-	-
TPOX	8	5.7%
D18S51	14	5.9%
D5S818	11 12	8.3% 6.1%
FGA	-	-


Locus D18S51
Alleles
Samples
Recruitment
Reference
Frequencies
0.00
0.05
0.10
0.15
0.20
Locus TPOX
Alleles
Samples
Recruitment
Reference
Frequencies
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
Locus TH01
Alleles
Samples
Recruitment
Reference
Frequencies
0.05
0.10
0.15
0.20
0.25
0.30
Locus FGA
Alleles
Samples
Recruitment
Reference
Frequencies
0.00
0.05
0.10
0.15
Locus D21S11
Alleles
Samples
Recruitment
Reference
Frequencies
0.00
0.05
0.10
0.15
0.20
0.25
Locus D19S433
Alleles
Samples
Recruitment
Reference
Frequencies
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35


Locus vWA
Alleles
Samples
Recruitment
Reference
Frequencies
0.05
0.10
0.15
0.20
0.25
Locus D16S539
Alleles
Samples
Recruitment
Reference
Frequencies
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Locus D13S317
Alleles
Samples
Recruitment
Reference
Frequencies
0.05
0.10
0.15
0.20
0.25
0.30
Locus D8S1179
Alleles
Samples
Recruitment
Reference
Frequencies
0.05
0.10
0.15
0.20
0.25
0.30
Locus D7S820
Alleles
Samples
Recruitment
Reference
Frequencies
0.00
0.05
0.10
0.15
0.20
0.25
Locus D5S818
Alleles
Samples
Recruitment
Reference
Frequencies
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40

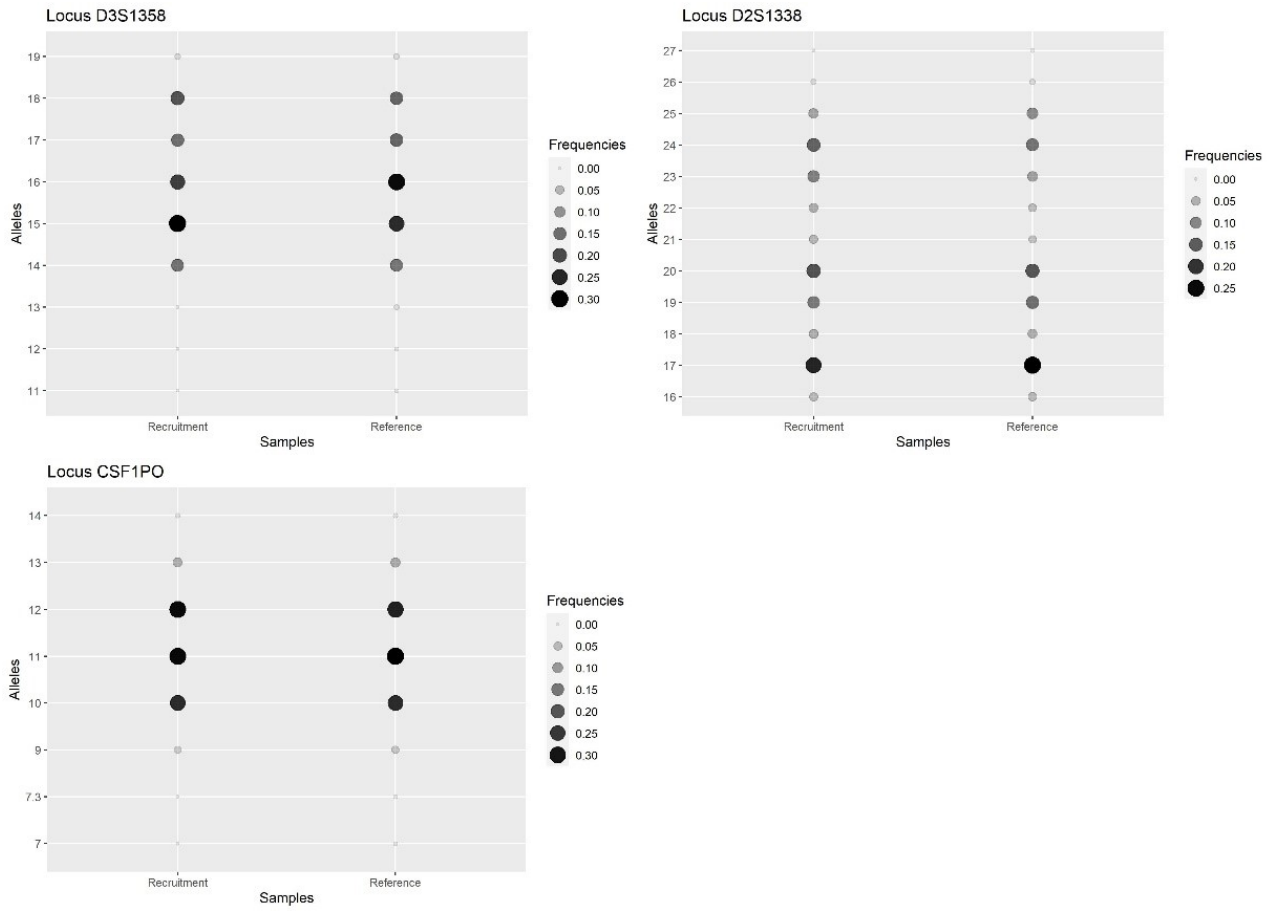


Figure 2.6 Bubble plots of allele frequencies for all observed alleles at 15 STR loci for the reference and recruitment samples. The size of a bubble indicates the frequency of the corresponding allele.

Fixation index (F_{ST})

The fixation index was used as a metric of the genic differentiation between the population samples (**Table 2.4**). The global F_{ST} value for all 15 loci taken together is 0.0009.

Table 2.4 F_{ST} estimates for the 15 STR loci when the reference and recruitment samples are compared.

Locus	F_{ST}
D8S1179	0.0000
D21S11	0.0001
D7S820	-0.0016
CSF1PO	-0.0017
D3S1358	0.0033
TH01	0.0012
D13S317	0.0037
D16S539	0.0001
D2S1338	0.0003
D19S433	0.0013
vWA	0.0021
TPOX	0.0011
D18S51	0.0002
D5S818	0.0047
FGA	-0.0014
All	0.0009

Random match probabilities

Because the RMP corresponds to the probability of randomly drawing a DNA profile from the population of interest that matches the trace's DNA profile and because allele frequencies determine the rarity of the profile, we quantified how the RMP was impacted by the sample used for its calculation (reference or recruitment).

A positive RMP log ratio value means that a genetic profile was more likely to be observed in the sample from which it was randomly generated. For example, if a genetic profile generated from the recruitment sample had a $\log(d_{rec})=1$, it would be 10x more likely to be observed in a population having allele frequencies corresponding to those obtained from the recruitment sample than the reference sample. The reverse is true for negative $\log(d_{rec})$ values. The same logic applies to $\log(d_{ref})$ except that positive values favour the reference sample in this case.

Values of $\log(d_{rec})$ for the genetic profiles generated from the recruitment sample frequencies (**Figure 2.7**) ranged from -1.24 to 2.13. In addition, 80.4% of genetic profiles had positive values (see red dotted box in **Figure 2.7**). Similarly, values of $\log(d_{ref})$ for the genetic profiles generated from the reference sample (**Figure 2.8**) ranged from -1.71 to 1.46. 56.3% of genetic profiles had positive values (see red dotted box in **Figure 2.8**).

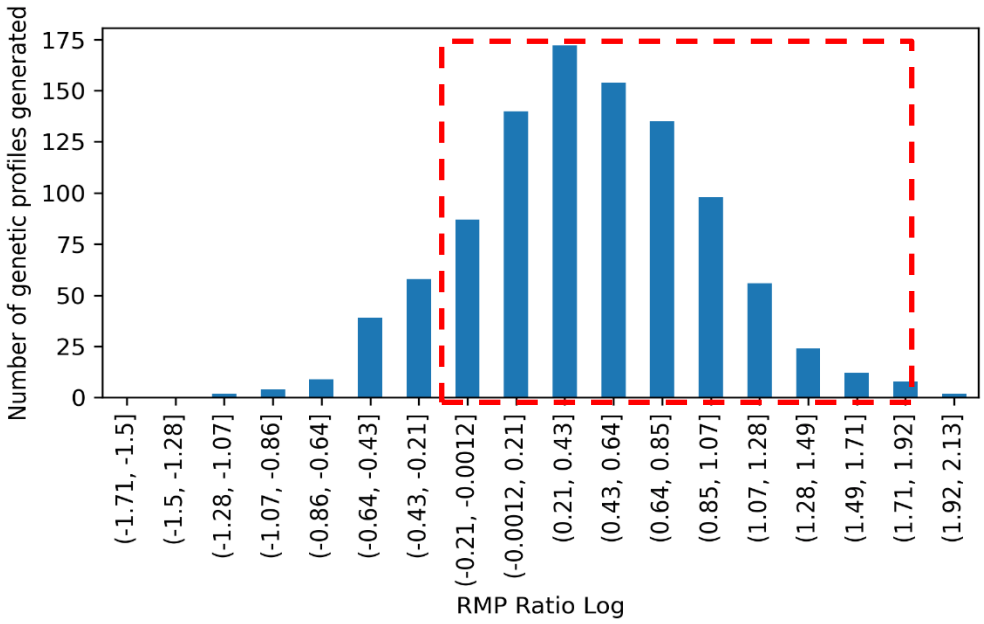


Figure 2.7 Distribution of log(d_{rec}) of the 1,000 genetic profiles generated from the recruitment sample. The red dotted box corresponds to the positive values of RMP ratio log.

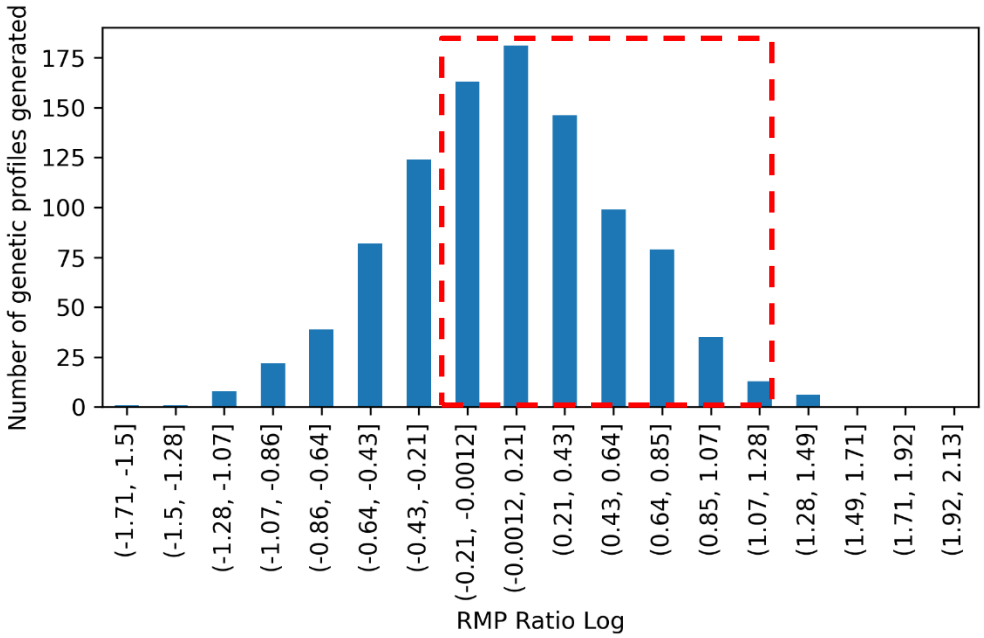


Figure 2.8 Distribution of log(d_{ref}) of the 1,000 genetic profiles generated from the reference sample. The red dotted box corresponds to the positive values of RMP ratio log.

Structure analysis of the samples

We uploaded the output from Structure's analyses into the Structure Harvester (Earl & Vonholdt, 2012, version v0.6.94) to visualize the posterior log-likelihood value L(*K*) for each model run, and identify the number of genetic clusters (*K*) that fits the data best (Earl & Vonholdt, 2012). Evanno et al. (2005) suggest using the maximal L(*K*) value before it reaches a plateau, as the best estimate of the number of populations represented in the genetic data. According to the L(*K*) plot (**Figure 2.9**), the break in the slope of the distribution was not clear enough to identify this best estimate of *K*. The data used to generate this graph can be found in the Evanno Table (**Table 2.5)**.

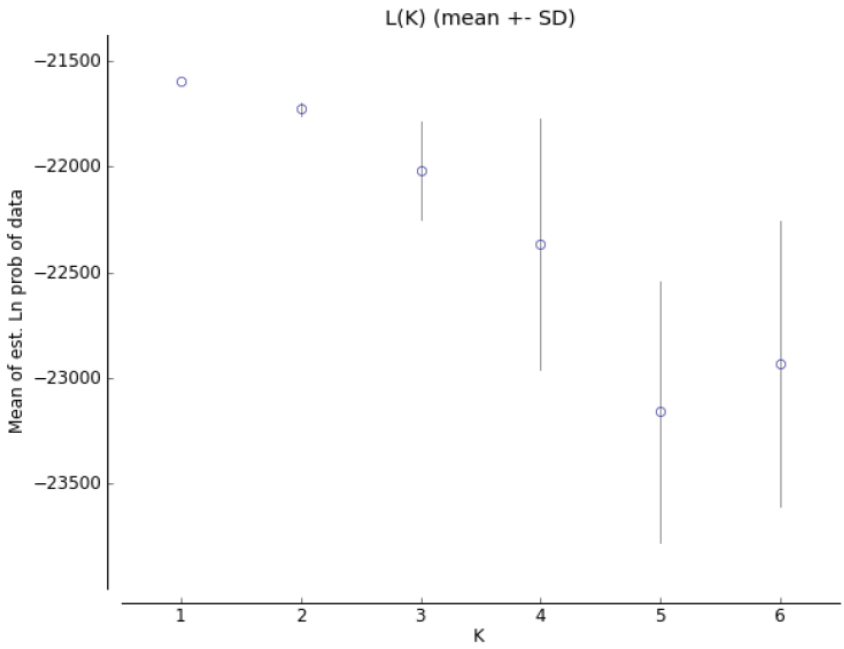


Figure 2.9 Posterior log-likelihood values L(*K*) for different number *K* of clusters. Averages and standard deviations are based on three replicate Structure analyses for each *K*.

To identify the *K* that best fits our data, we also calculated the Δ*K* metric, namely the rate of change in log-likelihood between successive *K* values (Evanno et al., 2005). First, we calculated L′(*K*) = L(*K*) − L(*K* − 1), i.e. the mean difference between successive likelihood values of *K*. Then, by using absolute values of the difference between successive L'(*K*), the second order of change was calculated as |L″(*K*)| = |L′(*K* + 1) −

L′(*K*)|. Finally, Δ*K* was obtained by the mean of the absolute values of the second order of change, divided by the standard deviation of L(*K*), Δ*K* = mean |L″(*K*)| / sd[L(*K*)]. Δ*K* shows a peak value of 5.3 at *K*=2 (**Figure 2.10**). This suggests that *K*=2 best represents the number of populations represented by the combined reference and recruitment samples. The data used to generate this graph can be found in Table (**Table 2.5**).

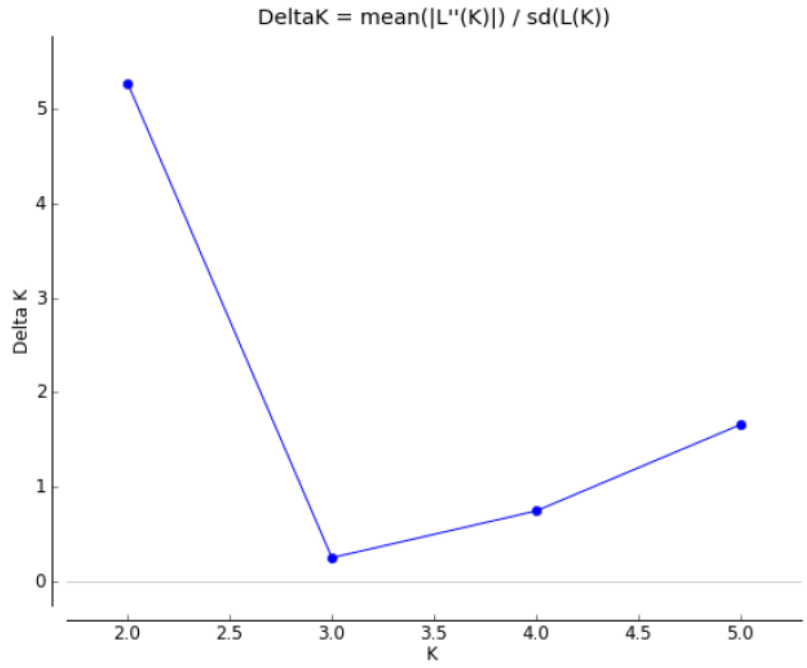


Figure 2.10 Distribution of the Δ*K* values for different numbers *K* of genetic clusters.

Table 2.5 Posterior log-likelihood of *K* and orders of change.

K	Reps	Mean LnP(*K*)	Stdev LnP(*K*)	Ln'(*K*)	|Ln''(*K*)|	Δ*K*
1	3	-21,594.80	0.44	—	—	—
2	3	-21,725.93	30.67	-131.13	161.63	5.27
3	3	-22,018.70	233.74	-292.77	58.03	0.25
4	3	-22,369.50	594.77	-350.80	441.93	0.74
5	3	-23,162.23	618.28	-792.73	1023.70	1.66
5	3	-23,162.23	618.28	-792.73	1023.70	1.66
6	3	-22,931.27	675.46	230.97	—	—

Figure 2.11 shows a bar graph of ancestry from Structure for the 404 individuals when *K*=2 clusters is retained. It is based on the estimated membership coefficient (Q-Matrix) (Pritchard et al., 2000). Thus, a single vertical line (X axis) represents a single individual from the data set, and each individual's bar is colored according to its cluster origin (i.e. from cluster #1 in red, cluster #2 in green, or mixed origin for bars with two colors). Thus, the overall proportion of membership of the combined sample is 50.1% for the first cluster and 49.9% for the second cluster. Respectively for each cluster, coefficients of membership ranged from 38.2% to 59.4%, and 40.6% to 61.8%. On average, each individual seemed to belong to 50% of the first cluster detected and 50% to the second. So, this suggests that the individuals did not seem to distinctly belong to one cluster more than the other.

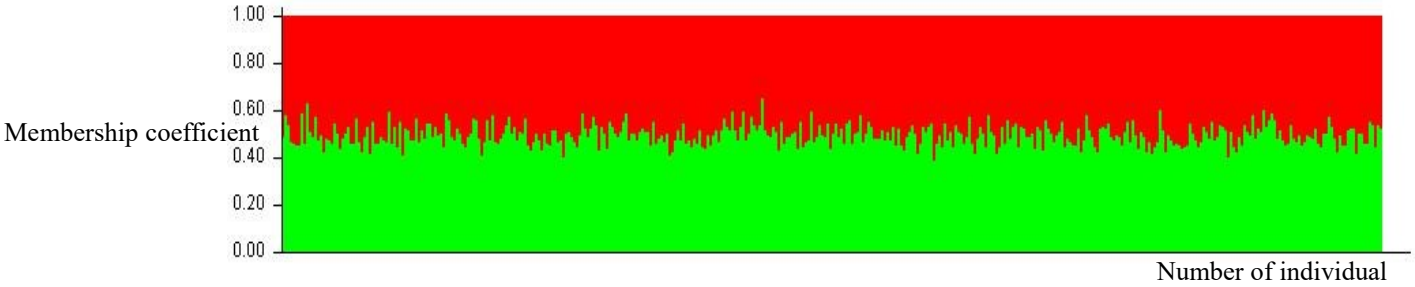


Figure 2.11 Bar graph of ancestry estimates (coefficient membership) of the combined reference and recruitment samples.

Discussion

The objective of this project was to propose a new method to define the population of interest database when composed of numerous mixed ethnic groups, as it is the case of Montreal city. Because criminal investigations are influenced by crime circumstances, the aim is to collect DNA traces from the environment and validate if it provides a more representative population of interest and therefore, adequately assess the weight of DNA evidence.

A first step into defining the population of interest was to compare different population samples: a **recruitment sample**, which corresponds to a positive control recently collected in 3 different regions (UQTR, Sherbrooke and Dolbeau-Mistassini), to the **reference sample**, i.e., the Caucasian population of the LSJML. Results obtained through the comparison analysis of the recruitment and reference samples did not suggest differentiated populations and both samples appeared to be equivalent in genetic composition (allele frequencies). No significant differences were found through the comparison of the variability of the weight of evidence with the recruitment and the reference samples, although the recruitment did not respect the Caucasian condition of the reference sample. Those results lead us to question the conception of the reference database, and therefore, the accurate population of interest, since a random recruitment sample "validated" the reference DNA database from the LSJML (here the Caucasian population) regarding both the genetic composition and the attribution of random match probability. Eventually, the environment sample, randomly collected in the city of Montreal, will be compared to both the recruitment and the reference samples to test this approach.

Allele frequencies

By comparing the occurrence of allele frequencies, both recruitment and reference samples suggest similar genetic composition for all 15 STR loci analysed. More specifically, analyses highlighted that differences observed in the genetic composition are not major. Differences are concentrated in 9 loci out of 15, never exceeding 8.3% within all alleles (**Table 2.3**). Loci with differences in allele frequencies greater than 5% were only observed for one or two alleles. Since the loci are polymorphic and between 6 and 14 alleles were listed, the samples suggested that the variations were not very pronounced across loci. Also, variations regarding low allele frequencies between samples were in fact very concentrated: when an allele frequency was of 0% in one sample, the allele was always detected below 1% frequency in the second sample, which did not suggest major or extreme diversity through allele frequencies.

Among others, studies of Budowle et al. (2001), Krane et al. (1992) and Chakraborty, Stivers, Su, Zhong, and Budowle (1999) all suggest contradictory results regarding the differences in allele frequency variations throughout populations. In comparison with these authors who highlighted differences among ethnicities, the approach of our study focuses otherwise. A random recruitment sample, whose population basin is composed of mixed Caucasian, Amerindians and international, did not reflect major diversity through allele frequencies when compared to a reference sample, i.e., the Caucasian population reference sample from the LSJML.

Fixation index (F_{ST})

With fixation index typically ranging from 0.01 to 0.03, the overall value of our data (0.0009, see **Table 2.4**) was found largely inferior, suggesting that the reference and recruitment samples both shared similar genetic material, without deviating from the expected proportions of Hardy-Weinberg. Because the fixation index reflects the level of differentiation between populations, this suggests the absence of differentiation between the recruitment and reference samples, and thus a major homogeneity, which is consistent with the distribution of allele frequencies.

As demonstrated by Steele, Court, and Balding (2014), F_{ST} estimates strongly vary according to the reference sample. Normally, in genetic population studies, F_{ST} estimates are based on a hypothetical ancestral population. However, F_{ST} estimates generated by a particular reference population allowed widely different results, which could transpose in our problematic into the wrong attribution/selection of the population of interest.

Budowle et al. (2001) studied data on the 13 CODIS core STR loci regrouping several sample populations from each of the following major population groups: African Americans, U.S. Caucasians, Hispanics, Far East Asians, and Native Americans. Variation of fixation index within ethnic groups tends to fluctuate, which overall values of F_{ST} estimates for the several populations suggested low degree of relatedness. Our results, which compared a random recruitment sample to a reference Caucasian

sample, suggested a low degree of differentiation. The global value of our results (0.0009) was similar to the Caucasian F_{ST} value (0.0005) reported by Budowle et al. (2001).

Random match probabilities

By generating random genetic profiles from the reference and recruitment samples, we showed that the choice of the sample of interest slightly (almost neglectable) impacted the RMP calculations, i.e., a small favorable difference depending on the sample used to generate the genetic profiles. The median value of the RMP ratio (**Figure 2.7**) means it was about 2.5 times more likely to observe a particular profile in the recruitment sample, randomly generated from the recruitment sample, when compared to the reference sample. Likewise, when a particular genetic profile is randomly generated from the reference sample, the median value of the RMP ratio (**Figure 2.8**) means it was about 1.2 times more likely to observe it in the reference sample contrary to the recruitment sample.

Random match probabilities (all values combined, **Figures 2.7** and **2.8**) are ranging from 10^{-16} to 10^{-27}. However, it has been determined by Hopwood et al. (2012) that the reasonable limit of the RMP thresholds was assigned to $1/10^{9}$ for unrelated individuals. By setting this threshold, it still supports the robustness of the model regarding calculation of the strength of evidence, without invoking unrealistic assumptions. We can therefore state that regardless of the choice of the population of interest (here the recruitment or the reference samples), results suggest small and neglectable impact on the proband values.

In the scientific literature, many authors demonstrated the unrepresentativeness of current structured databases regarding the population of interest and tried to overcome the ethnic aspect influencing the attribution of the statistical weight. Krane et al. (1992) had raised that by using an inappropriate ethnic group, random match probabilities affected by population subdivision yielded to artificially small estimates and biased values. Malaspinas et al. (2011) showed that not taking into consideration subdivision structure within the population can lead to underestimate the random match probability in a scenario where the two individuals originated from the same subpopulations. Kanthaswamy and

Smith (2014) and Ng, Oldt, and Kanthaswamy (2018), raised that the Native American population database reflected much higher genetic diversity than other major US populations used in databases. Therefore, Native American populations required an appropriate well-characterized subpopulation database in order to calculate statistical analysis (weight of evidence) more accurately.

In contrast to these studies, which have an approach based on ethnic division, our results showed that the random match probabilities normally calculated against a Caucasian reference sample was not impacted when calculated against a random recruitment sample. This suggests the necessity of redefining the population of interest, since an ethnically divided database does not seem to be an appropriate model for calculating the weight of evidence.

Structure analysis

Individuals were pooled together from both recruitment and reference samples, although they are composed of samples collected from distinct regions: Montreal, Chicoutimi, UQTR, Sherbrooke and Dolbeau-Mistassini. No premise nor geographical data were provided to Structure Software to attempt assigning individuals to a population structure model on the basis of their genotypes.

When no structure is detected in the sample by Structure software, typically, the proportion of the sample assigned to each cluster (population) roughly corresponds to $\sim 1/K$ in each population, meaning most individuals could be fairly admixed (Jonathan K Pritchard et al., 2010). This was consistent with our data, see **Figure 2.10**, where two different populations (K=2) seemed to best represent (at first) the genotypic differences between the combined reference and recruitment sample. Thus, the bar graph of ancestry estimates of the combined reference and recruitment samples (**Figure 2.11**) presented membership coefficients of each individual, which were on average 50%. The fact that individuals cannot be clustered into separate groups is coherent with previous data of allele frequencies and fixation index, where only small

allelic frequencies differences between the recruitment and reference samples were found. Presumably, the overall data (*K* parameter and membership coefficient) suggested a single population, which best reflected the structure of the combined reference and recruitment sample, since there was no distinctive difference detected within the structure. However, many authors have reported difficulties to accurately cluster groups based on genetics and inconsistent results, hence the importance of using Structure software with caution (Funk et al., 2020; Kalinowski, 2011) Gilbert et al. (2012). Moreover, because the Evanno method cannot perform comparison analysis when K=1, particular attention must be given.

All combined, our results raised the question of the validity of ethnic defined databases when assessing weight of evidence calculations. Although the recruitment sample did not respect the Caucasian condition of the reference sample, no significant differences were found for the random match probability. Hence, the first step of this study demonstrated that a random sample invalidated the use of the current model of forensic databases divided by ethnic groups and thus, the concept of the attribution of the population of interest.

Next step of this research will be the aim of validating an environment sample, trying to answer the question: How then to assess the relevant database of people of interest, while ethnic model database fitting the spatial area of interest is questionable? An environment sample consists of a promising database and would be more representative of the population of interest, as it is composed of DNA traces constantly left in the environment by human activities, i.e., a mixed basin population from Montreal city. Therefore, DNA traces from a crime scene will be compared (and thus confronted) to DNA traces constantly left by human activities in that very same area, with the aim to better assess the rarity of a genetic profile in a more precise way. Contrary to current databases divided by ethnicities, this approach avoids the need of prior information regarding the offender ethnicity (or the grey area of an individual of mixed ethnicity) to be able to confront DNA traces to the relevant population of interest (i.e., the relevant ethnic group).

Our method abstracts the ideal concept of database model defined by Champod, Evett, and Jackson (2004), which are of three types: the database of traces found on a similar crime, the database of offenders of the type of crime committed, and finally the database of suspects cleared of such a crime. Operational and ethical contingencies make the availability of at least two of them illusory. This study is particularly important in the context of the large-scale mobility of individuals in modern societies, which tend to be increasingly multi-ethnic. At the scientific level, this research will allow the approach of sampling DNA traces rather than volunteer individuals to better target the population of interest. Consequently, DNA traces from the perpetrator of a crime will be confronted to genetic profiles from the environment sample found in that very localized environment. At the legal level, an environment human DNA database would allow to support a much more precise statistical weight of a DNA match in a less biased way, when presented to the court to avoid miscarriage of justice and biased interpretations.

Limitations

For now, the environment sample presents some limitations according to the data collection geographically limited on a few specific PDQs from Montreal city. Eventually, the environment sample will necessarily need to consider all different and defined areas of Montreal (PDQs) to fully exploit this database model. A complete sampling of Montreal would accurately reflect the whole genetic diversity through DNA traces left by human activities, with the goal of serving as an actual reference database for criminal investigation. On the basis of crime circumstances, random match probabilities could be directly calculated through that environment database (considering where the crime was committed), adding more credibility and precision to the weight of evidence.

Moreover, the comparison of the recruitment sample with the Caucasian population of the reference samples involves other limitations. For the purpose of this study, the Caucasian population was hypothetically selected on the premise that, for example, the ethnicity of the author of a crime appeared to be Caucasian regarding information from the criminal investigation. Hence, could a random recruitment sample also invalidate the

use of a database ethnically divided by presenting similar data when compared among other groups, i.e., the African Americans of Native Americans groups from the LSJML?

Conclusion

This study supports that the recruitment sample, recently collected in UQTR, Sherbrooke and Dolbeau-Mistassini served as a positive control and was equivalent to the reference sample to assess the weight of evidence of DNA identification (Caucasian population dating from 1990 and still today used by LSJML). In fact, there were no major differences in the genetic composition (occurrence of allele frequencies). The recruitment sample nor the reference sample seemed to have a more significant impact on the rarity of a genetic profile. Besides, no structural differences were detected between the samples. Although the recruitment sample confirmed and validated the reference sample questioning the relevancy of ethnically defined database, the next step of this study will be to eventually compare the environment sample to both the reference and recruitment samples. In fact, forensic databases should be better defined by the population of interest regarding criminal investigation, taking into consideration the area of interest (be it where the crime was committed, or integrating investigation intelligence regarding the area of origin of the author). The genetic composition of DNA traces constantly left in the environment by daily activities (environment sample) might present a better approach to assess the population of interest of Montreal city. Future work should also consider whether it can be applied to other metropoles (e.g., Toronto, Vancouver) in Canada, or in cities where genetic variation (due to multiethnicity) is independent from territorial boundaries. Furthermore, these environmental data taken at a specific point in time could eventually be compared to future samples to follow the genetic evolution of populations and see if the allele frequencies tend to vary among the population through years and generation.

References

Balding, D. J. (1999). When can a DNA profile be regarded as unique? *Science & Justice, 39*(4), 257-260. doi:10.1016/s1355-0306(99)72057-5

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), 125-140. doi:10.1016/0379-0738(94)90222-4

Balding, D. J., & Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica, 96*(1), 3-12.

Bodner, M., Bastisch, I., Butler, J. M., Fimmers, R., Gill, P., Gusmão, L., . . . Parson, W. (2016). Recommendations of the DNA Commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal Short Tandem Repeat allele frequency databasing (STRidER). *Forensic science international. Genetics, 24*, 97-102. doi:10.1016/j.fsigen.2016.06.008

Budowle, B., Shea, B., Niezgoda, S., & Chakraborty, R. (2001). CODIS STR loci data from 41 sample populations. *Journal of Forensic Sciences, 46*(3), 453-489.

Butler, J. M. (2005). *Forensic DNA typing: biology, technology, and genetics of STR markers*: Elsevier.

Butler, J. M. (2015a). Chapter 1 - Data Interpretation Overview. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Interpretation* (pp. 3-24). San Diego: Academic Press.

Butler, J. M. (2015b). Chapter 11 - DNA Profile Frequency Estimates and Match Probabilities. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Interpretation* (pp. 281-308). San Diego: Academic Press.

Chakraborty, R. (1992). Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. *Human Biology, 64*(2), 141-159.

Chakraborty, R., Stivers, D. N., Su, B., Zhong, Y., & Budowle, B. (1999). The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis, 20*(8), 1682-1696. doi:10.1002/(sici)1522-2683(19990101)20:8<1682::Aid-elps1682>3.0.Co;2-z

Champod, C., Evett, I. W., & Jackson, G. (2004). Establishing the most appropriate databases for addressing source level propositions. *Science & Justice, 44*(3), 153-164. https://doi.org/10.1016/S1355-0306(04)71708-6

Council, N. R. (1992). *DNA Technology in Forensic Science*. Washington, DC: The National Academies Press.

Curran, J. M., Buckleton, J. S., & Triggs, C. M. (2003). What is the magnitude of the subpopulation effect? *Forensic Science International, 135*(1), 1-8. doi:10.1016/s0379-0738(03)00171-3

Curran, J. M., Walsh, S. J., & Buckleton, J. (2007). Empirical testing of estimated DNA frequencies. *Forensic Science International: Genetics, 1*(3-4), 267-272. doi:10.1016/j.fsigen.2007.06.004

Earl, D. A., & Vonholdt, B. (2012). Structure Harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources, 4*.

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Molecular Ecology, 14*(8), 2611-2620. doi:10.1111/j.1365-294X.2005.02553.x

Funk, S. M., Guedaoura, S., Juras, R., Raziq, A., Landolsi, F., Luís, C., . . . Cothran, E. G. (2020). Major inconsistencies of inferred population genetic structure estimated in a large set of domestic horse breeds using microsatellites. *Ecology and Evolution, 10*(10), 4261-4279. https://doi.org/10.1002/ece3.6195

Ge, J., King, J. L., Smuts, A., & Budowle, B. (2021). Precision DNA Mixture Interpretation with Single-Cell Profiling. *Genes (Basel), 12*(11). doi:10.3390/genes12111649

Gilbert, K. J., Andrew, R. L., Bock, D. G., Franklin, M. T., Kane, N. C., Moore, J.-S., . . . Vines, T. H. (2012). Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program structure. *Molecular Ecology, 21*(20), 4925-4930. https://doi.org/10.1111/j.1365-294X.2012.05754.x

Gill, P., Foreman, L., Buckleton, J. S., Triggs, C. M., & Allen, H. (2003). A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations. *Forensic Science International, 131*(2), 184-196. https://doi.org/10.1016/S0379-0738(02)00423-1

Gouvernement du Canada. (1998). *DNA Identification Act (S.S. 1998, c. 37)*. Retrieved from https://laws-lois.justice.gc.ca/eng/acts/D-3.8/FullText.html

Gouvernement du Québec. (2013). *Laboratoire de sciences judiciaires et de médecine légale*. Retrieved from https://www.securitepublique.gouv.qc.ca/laboratoire/a-propos.html

Gouvernement du Québec. (2019). *Partenariat Données Québec, carte interactive*. Retrieved from https://www.donneesquebec.ca/recherche/fr/dataset/cartes-topographiques-a-l-echelle-de-1-100-000/resource/089de803-c851-4088-9b85-55356b2833be

Hessab, T., Aranha, R. S., Moura-Neto, R. S., Balding, D. J., & Schrago, C. G. (2018). Evaluating DNA evidence in a genetically complex population. *Forensic science international. Genetics, 36*, 141-147. doi:10.1016/j.fsigen.2018.06.019

Hopwood, A. J., Puch-Solis, R., Tucker, V. C., Curran, J. M., Skerrett, J., Pope, S., & Tully, G. (2012). Consideration of the probative value of single donor 15-plex STR profiles in UK populations and its presentation in UK courts. *Sci Justice, 52*(3), 185-190. doi:10.1016/j.scijus.2012.05.005

Jamieson, A., & Bader, S. (2016). *A guide to forensic DNA profiling / edited by Allan Jamieson, Scott Bader*. Chichester, West Sussex, England: Wiley.

Kalinowski, S. T. (2011). The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity, 106*(4), 625-632. doi:10.1038/hdy.2010.95

Kanthaswamy, S., & Smith, D. G. (2014). Genetic and ethnohistoric evidence suggest current Native American population datasets in the FBI's CODIS database are not sufficiently representative. *Forensic Science International: Genetics, 13*, e13-15. doi:10.1016/j.fsigen.2014.05.006

Krane, D. E., Allen, R. W., Sawyer, S. A., Petrov, D. A., & Hartl, D. L. (1992). Genetic differences at four DNA typing loci in Finnish, Italian, and mixed Caucasian populations. *Proceedings of the National Academy of Sciences of the United States of America, 89*(22), 10583-10587. doi:10.1073/pnas.89.22.10583

Kruijver, M. (2016). Characterizing the genetic structure of a forensic DNA database using a latent variable approach. *Forensic Science International: Genetics, 23*, 130-149. doi:10.1016/j.fsigen.2016.03.007

Lapointe, M., Rogic, A., Bourgoin, S., Jolicoeur, C., & Séguin, D. (2015). Leading-edge forensic DNA analyses and the necessity of including crime scene investigators, police officers and technicians in a DNA elimination database. *Forensic Science International: Genetics, 19*, 50-55. https://doi.org/10.1016/j.fsigen.2015.06.002

Lee, H. C., Gaensslen, R. E., Bigbee, P. D., & Kearney, J. (1991). Guidelines for the collection and preservation of DNA evidence. *Journal of Forensic Identification, 41*, 344-356.

Lowe, A. L., Urquhart, A., Foreman, L. A., & Evett, I. W. (2001). Inferring ethnic origin by means of an STR profile. *Forensic Science International, 119*(1), 17-22. doi:10.1016/s0379-0738(00)00387-x

Malaspinas, A.-S., Slatkin, M., & Song, Y. S. (2011). Match probabilities in a finite, subdivided population. *Theoretical Population Biology, 79*(3), 55-63. doi:10.1016/j.tpb.2011.01.003

National Geospatial-Intelligence Agency. (2022). World Geodetic System 1984 (WGS 84). Retrieved from https://earth-info.nga.mil/index.php?dir=wgs84&action=wgs84

National Research Council. (1996). *The Evaluation of Forensic DNA Evidence*. Washington, DC: The National Academies Press.

Ng, J., Oldt, R. F., & Kanthaswamy, S. (2018). Assessing the FBI's Native American STR database for random match probability calculations. *Legal Medicine, 30*, 52-55. https://doi.org/10.1016/j.legalmed.2017.10.012

Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, A., & Lareu, M. V. (2013). An overview of STRUCTURE: applications, parameter settings, and supporting software. *Frontiers in genetics, 4*, 98-98. doi:10.3389/fgene.2013.00098

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959. doi:10.1093/genetics/155.2.945

Pritchard, J. K., Wen, W., & Falush, D. (2010). Documentation for STRUCTURE software: Version 2. *University of Chicago, Chicago, IL*.

Promega. (2016). DNA IQ™ System - Database Protocol. Retrieved from https://www.promega.com/-/media/files/resources/protocols/technical-bulletins/101/dna-iq-system-database-protocol.pdf

R Lempert. (1993). DNA, Science and the Law: Two Cheers for the Ceiling Principle,. *Jurimetrics, 34*, 41-57.

Raymond, M., & Rousset, F. (1995a). An exact test for population differentiation. *Evolution, 49*(6), 1280-1283. doi:10.1111/j.1558-5646.1995.tb04456.x

Raymond, M., & Rousset, F. (1995b). GENEPOP (version 1.2): population genetics software for exact tests and ecumenicism. *Journal Heredity, 86*, 248-249.

Smalldon, K. W., & Moffat, A. C. (1973). The calculation of discriminating power for a series of correlated attributes. *J Forensic Sci Soc, 13*(4), 291-295. doi:10.1016/s0015-7368(73)70828-8

SWGDAM. (2017). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. Retrieved from https://www.swgdam.org/publications

Thermo Fisher Scientific Inc. (2018a). AmpFℓSTR™ Identifiler™ Plus PCR Amplification Kit – USER GUIDE. Retrieved from https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4440211_AmpFlSTR_IdentifilerPlus_UG.pdf

Thermo Fisher Scientific Inc. (2018b). Quantifiler™ HP and Trio DNA Quantification Kits - USER GUIDE. Retrieved from https://assets.thermofisher.com/TFS-Assets/LSG/manuals/4485354.pdf

Thermo Fisher Scientific Inc. (2019). GeneMapper™ ID-X Software v1.6, full installation. Retrieved from https://www.thermofisher.com/order/catalog/product/A39975

Vieira, M. L. C., Santini, L., Diniz, A. L., & Munhoz, C. d. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology, 39*(3), 312-328. doi:10.1590/1678-4685-GMB-2016-0027

Ville de Montréal - Données ouvertes. (2020). Limite administrative de l'agglomération de Montréal (Arrondissements et Villes liées). Retrieved from https://donnees.montreal.ca/ville-de-montreal/polygones-arrondissements#resource-limite_administrative_de_l'agglom%C3%A9ration_de_montr%C3%A9al

Ville de Montréal - Données ouvertes. (2022). Limite des secteurs de poste de quartier de police. Retrieved from https://donnees.montreal.ca/ville-de-montreal/limites-pdq-spvm#resource-limites_des_postes_de_quartier

Ville de Montréal. (2016). Profils sociodémographiques, Recensement de 2016. Retrieved from http://ville.montreal.qc.ca/portal/page?_pageid=6897,68055570&_dad=portal&_schema=PORTAL

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution, 38*(6), 1358-1370. doi:10.2307/2408641

Supplementary material

Table 2.S1 Allele frequencies for 15 listed loci from recruitment and reference sample

D8S1179

Alleles	8	9	10	11	12	13	14	15	16	17
Recruitment	0,020	0,012	0,071	0,094	0,138	0,272	0,244	0,091	0,051	0,008
Reference	0,016	0,005	0,074	0,101	0,150	0,314	0,197	0,105	0,032	0,005

D21S11

Alleles	22,3	24,2	26,2	27	28	28,2	29	29,2	30	30,2	30,3	31	31,2
Recruitment	0,000	0,000	0,000	0,024	0,142	0,004	0,232	0,004	0,244	0,028	0,000	0,079	0,098
Reference	0,002	0,004	0,002	0,023	0,161	0,000	0,191	0,004	0,294	0,034	0,002	0,069	0,081
Alleles	32	32,2	33,2	34,1	34,2								
Recruitment	0,020	0,091	0,028	0,000	0,008								
Reference	0,014	0,088	0,027	0,002	0,002								

D7S820

Alleles	7	8	9	10	10,3	11	11,1	12	13	14
Recruitment	0,024	0,156	0,100	0,260	0,000	0,216	0,000	0,172	0,064	0,008
Reference	0,023	0,139	0,135	0,258	0,002	0,224	0,002	0,161	0,043	0,013

CSF1PO

Alleles	7	7,3	9	10	11	12	13	14
Recruitment	0,000	0,000	0,028	0,264	0,319	0,319	0,067	0,004
Reference	0,002	0,002	0,034	0,265	0,334	0,283	0,074	0,005

D3S1358

Alleles	11	12	13	14	15	16	17	18	19
Recruitment	0,000	0,000	0,000	0,143	0,302	0,214	0,147	0,183	0,012
Reference	0,002	0,002	0,007	0,135	0,240	0,287	0,159	0,159	0,009

TH01

Alleles	5	6	7	8	9	9,3	10
Recruitment	0,004	0,217	0,181	0,13	0,154	0,303	0,012
Reference	0,002	0,273	0,141	0,101	0,141	0,327	0,016

D13S317

Alleles	7	8	9	10	11	12	13	14	15
Recruitment	0,004	0,126	0,071	0,047	0,335	0,283	0,102	0,028	0,004
Reference	0,007	0,099	0,069	0,069	0,264	0,345	0,103	0,043	0,002

D16S539

Alleles	8	9	10	11	12	13	14	15
Recruitment	0,008	0,102	0,059	0,319	0,244	0,217	0,051	0,000
Reference	0,011	0,116	0,06	0,305	0,296	0,188	0,023	0,002

D2S1338

Alleles	16	17	18	19	20	21	22	23	24	25	26	27
Recruitment	0,039	0,217	0,051	0,114	0,161	0,043	0,055	0,106	0,142	0,063	0,008	0,000
Reference	0,042	0,262	0,054	0,126	0,157	0,031	0,036	0,069	0,121	0,092	0,009	0,002

D19S433

Alleles	11	12	13	13,2	14	14,2	15	15,2	16	16,2	17	17,2	18,2
Recruitment	0,008	0,075	0,197	0,035	0,378	0,031	0,154	0,047	0,035	0,024	0,000	0,012	0,004
Reference	0,005	0,092	0,251	0,018	0,336	0,011	0,170	0,038	0,054	0,014	0,005	0,004	0,002

vWA

Alleles	13	14	15	16	17	18	19	20	21
Recruitment	0,004	0,094	0,11	0,244	0,283	0,161	0,071	0,028	0,004
Reference	0,004	0,117	0,112	0,202	0,236	0,202	0,112	0,011	0,004

TPOX

Alleles	7	8	9	10	11	12
Recruitment	0,004	0,461	0,102	0,075	0,323	0,035
Reference	0,000	0,518	0,101	0,058	0,283	0,04

D18S51

Alleles	9	10	11	12	13	14	15	16	17	18	19	20	21
Recruitment	0,008	0,008	0,004	0,146	0,138	0,209	0,142	0,13	0,094	0,047	0,039	0,024	0,008
Reference	0,002	0,005	0,011	0,143	0,137	0,15	0,166	0,126	0,087	0,081	0,04	0,027	0,005
Alleles	22	23	24										
Recruitment	0,004	0,000	0,000										
Reference	0,014	0,004	0,002										

D5S818

Alleles	8	9	10	11	12	13	14	15
Recruitment	0,000	0,039	0,055	0,433	0,311	0,150	0,008	0,004
Reference	0,004	0,056	0,054	0,350	0,372	0,144	0,018	0,002

FGA

Alleles	16	18	19	20	20,2	21	21,2	22	22,2	23	23,2	24	24,2
Recruitment	0,004	0,012	0,075	0,106	0,000	0,189	0,000	0,13	0,004	0,173	0,004	0,169	0,004
Reference	0,000	0,009	0,065	0,134	0,002	0,184	0,002	0,152	0,007	0,159	0,004	0,155	0,000
Alleles	25	26	27	28									
Recruitment	0,067	0,043	0,020	0,000									
Reference	0,087	0,029	0,011	0,002									

CHAPITRE III

DISCUSSION ET PERSPECTIVES

3.1 Discussion

En science forensique, lorsqu'il y a une concordance entre le profil génétique de la trace d'ADN prélevée sur une scène de crime et celui du suspect, la norme consiste à évaluer la rareté de ce profil génétique au sein de la population d'intérêt à l'enquête, en le comparant à un échantillon de référence selon un groupe ethnique jugé pertinent. Comme les bases de données de référence des laboratoires judiciaires sont divisées par ethnicités et ne sont pas toujours collectées aléatoirement au sein de la population, il est légitime de se poser la question de leur pertinence vis-à-vis de la population d'intérêt à l'enquête. Ainsi, comment définir la base de données la plus représentative de la population d'intérêt, alors que la méthode actuelle soulève ce type de questionnement. La pertinence locale, géographique des lieux de la commission des faits, ne serait-elle pas plus adéquate?

De fait, Champod, Evett, et Jackson (2004) ont déjà démontré que les bases de données forensiques idéales d'identification génétique sont de trois types : la base des traces relevées sur un crime similaire, celle relative aux délinquants du type de crime commis, et finalement celle des suspects innocentés d'un tel crime. Des contingences opérationnelles et éthiques rendent illusoire la disponibilité d'au moins deux d'entre elles. Pour contourner ces contraintes, les enquêtes criminelles étant influencées par les circonstances du crime, l'idée était de redéfinir le concept de population d'intérêt à l'enquête et de confronter les traces ADN à celles recueillies dans un environnement défini comme environnement d'intérêt, par exemple les lieux (au sens large, géographique) de la commission du crime. Non seulement il était nécessaire de valider si cet échantillon tiré de l'environnement constituait une population d'intérêt davantage représentative, mais également d'évaluer avec précision la rareté d'un profil génétique au sein de cette

population d'intérêt. Ce projet visait donc à définir la population d'intérêt lié à l'enquête, quelle qu'en soit sa composition ethnique, en plus de proposer une base de données ADN à partir de traces collectées au hasard dans l'environnement.

Une première étape menant à la réalisation de cet objectif était de tester et comparer différents échantillons de population, questionnant ainsi la validité du modèle actuel divisé par groupe ethnique. **L'échantillon de recrutement**, qui correspond à un contrôle positif récemment collecté dans trois régions différentes (UQTR, Sherbrooke et Dolbeau-Mistassini), a donc été comparé à **l'échantillon de référence**, soit la population caucasienne du LSJML, collecté dans les années 1990 à Montréal et Chicoutimi. Des analyses permettant de mesurer la composition (fréquence allélique), l'indice de différentiation, la rareté d'un profil génétique (probabilité de concordance fortuite) ainsi que toute structure détectée au sein des échantillons ont été réalisées.

3.1.1 Fréquence allélique

Dans la littérature, il est attendu que seulement 5 % (microsatellites) à 10 % (SNPs) de la diversité génétique humaine expliquent la différentiation génétique majeure au travers des populations humaines en raison de la dérive génétique et de la sélection naturelle (Holsinger et Weir, 2009). Non seulement la diversité génétique est présente entre les populations, mais elle serait également d'autant plus importante au sein des individus composant la même population, comme l'ont démontré Barbujani, Magagni, Minch, et Cavalli-Sforza (1997). De leur côté, Silva, Pereira, Poloni, et Currat (2012) ont démontré que la variation de la diversité génétique parmi les populations était principalement influencée et déterminée par la géographie, ayant pour effet de sous-estimer les niveaux de structure génétique au sein de celles-ci. Dans notre étude, il importait donc de définir les fréquences alléliques de nos échantillons de recrutement et de référence pour ainsi comprendre leur composition génétique.

Ainsi, les fréquences alléliques des échantillons de référence et de recrutement ont d'abord été comparées, et ce, pour les 15 loci répertoriés, suggérant des échantillons

sensiblement similaires (**Chapitre II, Table 2.3**). Sur les 15 loci analysés, 6 loci (**D8S1179**, **D7S820**, **CSF1PO**, **D2S1338**, **vWA** et **FGA**) ne présentaient pas un écart de fréquence allélique supérieur à 5 %. Au contraire, 9 loci sur 15 ont démontré avoir des différences alléliques légèrement plus notables. De ces loci, il est possible de remarquer que certains présentent un écart de fréquence allélique supérieur à 5 % pour seulement un seul allèle : **D21S11** (allèle 30), **TH01** (allèle 6), **D16S539** (allèle 12), **D19S433** (allèle 13), **TPOX** (allèle 8) et **D18S51** (allèle 14). L'étendue des valeurs de différence de fréquence allélique pour ces loci varie de 5 % à 5,9 %. D'autres loci présentent un écart de fréquence allélique observé pour deux allèles, soit **D3S1358** (allèle 15 et 16), **D13S317** (allèle 11 et 12) et **D5S818** (allèle 11 et 12). Légèrement plus élevée, l'étendue des valeurs de différence de fréquence allélique pour ces loci varie de 6,1 % à 8,3 %. Également, il a été observé que lorsqu'un allèle était détecté dans un échantillon, et que la fréquence de ce même allèle était de 0 % dans le second échantillon, l'écart de fréquence allélique entre les échantillons n'excédait jamais plus de 0,7 %.

En général, les variations de fréquence allélique n'étant pas majeures, les échantillons de référence et de recrutement ne démontrent pas de différence significative entre eux au niveau de la composition génétique, suggérant des échantillons plutôt homogènes. Ainsi, un échantillon de recrutement aléatoire, dont le bassin de population est composé d'un mélange de Caucasiens, d'Amérindiens et d'internationaux, n'a pas reflété de diversité majeure à travers les fréquences alléliques lorsqu'il a été comparé à l'échantillon de référence de la population caucasienne du LSJML. D'ailleurs, on constate la même tendance entre les loci des deux échantillons, ce qui implique que même si certains allèles sont détectés dans un échantillon et absents dans l'autre, les allèles sont relativement fréquents dans la même étendue. Des écarts majeurs entre les échantillons ou des variations considérables, pouvant refléter de l'hétérogénéité ou de la diversité génétique, n'ont donc pas été observés.

Plusieurs auteurs ont étudié les différences de distribution quant à la variation de la composition génétique, considérant le facteur de la structure des populations. Notamment, l'étude de Krane et al. (1992) comparait deux sous-populations caucasiennes (Finlandais

et Italiens) à une base de données mixte caucasienne, suggérant des différences prononcées entre les populations. De leur côté, Chakraborty, Stivers, Su, Zhong et Budowle (1999) ont démontré que les loci STR étaient tous hautement polymorphes au sein de sept grands groupes ethniques (Américains d'origine européenne, Italiens, Suisses, Américains d'origine africaine, Jamaïcains, Chinois et Amérindiens d'origine apache), suggérant des variations de fréquence alléliques majoritairement aléatoires entre les populations. Ces études contradictoires ne soutiendraient-elles pas notre approche qui s'affranchit de toute division ethnique ?

3.1.2 Index de fixation (F_{ST})

Bien que l'analyse des échantillons de référence et de recrutement suggérait une composition allélique similaire, la comparaison des échantillons a été approfondie au moyen de l'indice de différentiation (Fst). Comme l'ont démontré Steele, Court et Balding (2014), les estimations de Fst varient fortement, tout dépendamment du choix de l'échantillon de référence, ce qui pourrait se transposer dans notre problématique liée à l'attribution/sélection erronée de la base de données de référence en raison d'une population hétérogène ne reflétant pas la population d'intérêt réelle à l'enquête.

En effet, la mesure du F_{ST} se trouve directement liée à la variation des fréquences alléliques répertoriées entre les populations. L'estimation F_{ST} est donc utilisée pour démontrer la différence significative entre des échantillons de populations, et ce, par rapport à la population entière (Holsinger & Weir, 2009) (Weir & Cockerham, 1984). L'ajustement F_{ST} est particulièrement important en science forensique, puisqu'il vient corriger les incertitudes liées au choix de la base de données, qui n'est pas toujours le plus approprié à la population d'intérêt à l'enquête pour évaluer la rareté d'une trace ADN (Steele et al., 2014). Typiquement, ces valeurs de F_{ST} sont comprises entre 0 (partage complet du matériel génétique) et 1 (absence de partage du matériel génétique) (Raymond et Rousset, 1995). L'estimation F_{ST} a ainsi été générée pour les 15 loci analysés, dont la valeur globale est de 0,0009, **Chapitre II, Table 2.4**.

La valeur globale F_{ST} de nos résultats (0,0009, **Table 2.4**), comparant la différentiation des échantillons de recrutement et de référence, s'avère largement inférieure, soit près de 11 fois inférieure à la valeur conservative (0,01) permettant de corriger l'impact de la déviation de l'équilibre de Hardy-Weinberg et de l'équilibre de liaison au sein d'une population urbaine (National Research Council, 1996). Nos résultats démontrent que les échantillons partagent fortement leur matériel génétique, ne suggérant pas des populations largement différenciées. D'ailleurs, notre valeur F_{ST} mesurée pour 15 loci STR (0,0009) se trouve sensiblement similaire à la valeur globale estimée pour la population caucasienne (0,0005) mesurée pour 13 loci STR par B. Budowle et al. (2001). En effet, leur étude comportait des échantillons de plusieurs groupes de populations majeures: Afro-Américains, Caucasiens américains, Hispaniques, Asiatiques d'Extrême-Orient et Amérindiens. À travers ces différents groupes, la variation de l'indice de fixation au sein des groupes ethniques avait tendance à fluctuer, suggérant un faible degré de parenté entre ces groupes. Similairement, une étude sur 13 loci STR par Bruce Budowle and Chakraborty (2001) a démontré qu'une faible valeur de F_{ST} (0,0028) était obtenue au travers de 11 populations de l'Europe et que cette valeur n'engendrait pratiquement pas d'impact sur l'évaluation de la rareté d'une trace ADN. L'utilisation d'une valeur conservative de F_{ST} (0,01) était d'ailleurs justifiée pour la population européenne (Bruce Budowle et Chakraborty, 2001). Au contraire, John Buckleton et al. (2016) ont suggéré dans leur étude de bonifier légèrement la valeur Fst, considérant que celle-ci avait un impact sur la RMP calculé à partir des jeux de données de fréquences disponibles, n'ayant pas un effet conservateur tel que l'on pensait. Les auteurs ont donc proposé des probabilités de concordance alléliques pour des paires d'allèles répertoriés dans une population, et ce, ne négligeant pas la possibilité que deux individus possèdent des ancêtres communs. Une fois de plus, notre approche ne permettrait-elle pas de s'affranchir de cette contradiction ?

3.1.3 Probabilités de concordance fortuite

Dans la littérature scientifique, de nombreux auteurs ont démontré les impacts de la diversité génétique liée à l'ethnicité par rapport au modèle de base de données actuel sur

les calculs de valeurs probantes, menant à des valeurs biaisées et affectées par cette division/structure des populations (Kanthaswamy et Smith, 2014; Krane et al., 1992; Malaspinas, Slatkin, et Song, 2011; Ng, Oldt, et Kanthaswamy, 2018). Il a donc été jugé pertinent de vérifier l'impact de la composition génétique des échantillons lors de l'évaluation du poids statistique d'une trace ADN, bien que les résultats comparant les échantillons de recrutement et référence ne suggéraient pas une diversité génétique (occurrence des fréquences alléliques) prononcée. En effet, la probabilité de concordance fortuite dépend d'une base de données appropriée, définie et caractérisée pour une population particulière, dans le but d'estimer de manière fiable la fréquence d'un génotype (Chakraborty, 1992).

En générant aléatoirement des milliers de profils génétiques à partir des allèles répertoriés des échantillons de recrutement et de référence, il s'avère que le choix de cet échantillon n'impacte que légèrement les calculs de valeurs probantes, avec des RMP se situant entre 10^{-16} et 10^{-27}, et ce, pour tous échantillons confondus. Lorsque les valeurs RMP calculées pour chacun des deux échantillons étaient comparées, une faible différence favorable en fonction de l'échantillon utilisé lors de la génération des profils génétiques est observée.

Lorsqu'un profil génétique particulier est généré aléatoirement à partir des allèles présents dans l'échantillon de **recrutement**, la valeur médiane du ratio RMP (valeurs allant de -1,24 à 2,13, **Chapitre II, Figure 2.7**), suggère qu'il est environ 2,5 fois plus probable d'observer ce profil dans l'échantillon de recrutement, que dans l'échantillon de référence. La valeur maximale (2,13) suggère qu'il est au maximum 135 fois plus probable d'observer ce profil dans l'échantillon à partir duquel il est généré (ici recrutement).

De même, lorsqu'un profil génétique particulier est généré aléatoirement à partir des allèles présents dans l'échantillon de **référence**, la valeur médiane du ratio RMP (valeurs allant de -1,71 à 1,46, **Chapitre II, Figure 2.8**), suggère qu'il est environ 1,2 fois plus probable de l'observer dans l'échantillon de référence, que dans l'échantillon de recrutement. Dans ce cas-ci, la valeur maximale se trouve beaucoup plus faible (1,46),

suggérant qu'il est au maximum 28 fois plus probable d'observer ce profil dans l'échantillon de référence.

Cependant, Hopwood et al. (2012) ont déterminé qu'un seuil raisonnable pour estimer la valeur d'une probabilité de concordance fortuite devait être limité à $1/10^9$, afin d'éviter des interprétations irréalistes. Nos valeurs dépassant ainsi largement ce seuil (10^{-16} to 10^{-27}), il devient négligeable de dire qu'il est 135 fois plus probable d'observer un profil génétique particulier dans un échantillon plutôt que dans l'autre. Ainsi, ces valeurs soutiennent que notre échantillon de recrutement (récolté aléatoirement sans tenir compte du facteur ethnique) est tout aussi représentatif que l'échantillon de référence (fondé sur ce facteur). On peut dès lors questionner la validité de la méthode classique utilisée pour l'échantillon de référence, nonobstant ses éventuels défis éthiques.

3.1.4 Analyse de structure

Finalement, il a été pertinent d'analyser l'ensemble de nos données de l'échantillon combiné (recrutement et référence, soit 404 individus) au moyen du logiciel Structure. Ce logiciel utilise un modèle d'analyse basé sur les données génotypiques, utilisant des marqueurs non liés (ici 15 loci) et prenant en considération un équilibre Hardy-Weinberg et un équilibre de liaison, dans le but de détecter et d'inférer toute structure de population (J. K. Pritchard, Stephens, & Donnelly, 2000; Jonathan K Pritchard, Wen, & Falush, 2010). Les fréquences alléliques estimées sont utilisées pour évaluer la probabilité qu'un génotype donné soit originaire de chaque population, et ce, en se référant aux valeurs de vraisemblance maximale. L'utilisation de méthodes statistiques standard a permis de déduire le nombre de groupements, c'est-à-dire *K*, le nombre de populations, à partir des observations des tirages aléatoires (J. K. Pritchard et al., 2000).

Le graphique à barres des estimations d'ascendance obtenu par le logiciel Structure correspond aux valeurs des coefficients d'appartenance estimés dans l'échantillon combiné (**Chapitre II, Figure 2.11**) (J. K. Pritchard et al., 2000). Ainsi, la proportion globale d'appartenance de l'échantillon combiné correspondait à 50,1 % pour le premier

groupement et à 49,9 % pour le second groupement, les valeurs variant respectivement de 38,2 % à 59,4 % et de 40,6 % à 61,8 %. Puisqu'en moyenne, chaque individu semblait appartenir à 50 % à chaque population/groupement, les résultats suggèrent que les individus ne semblent pas appartenir distinctement à une population plus qu'à une autre. Les individus n'étant donc pas classés en groupes distincts, ce résultat est cohérent avec les données précédentes (fréquences alléliques et de l'indice de différentiation), où seules de faibles variances génétiques entre les échantillons de recrutement et de référence ont été détectées. Vraisemblablement, les données globales des coefficients d'appartenance suggèrent qu'une seule population homogène reflète le mieux la structure de l'échantillon combiné, puisqu'aucune différence distinctive n'a été détectée au sein de cette structure.

Tang et al. (2005) ont d'ailleurs déjà étudié la division ethnique par groupement, analysant ainsi les données génétiques de 326 marqueurs microsatellites à partir d'un échantillon composé de quatre groupes raciaux/ethnique majeurs auxquels chaque individu s'était préalablement identifié. Ils ont obtenu des résultats de correspondance très précis entre la race/ethnicité auto-identifiée et les groupements générés par le logiciel structure, soutenant la fiabilité du logiciel à détecter ces ethnicités.

Cependant, les logiciels de structure doivent être utilisés avec prudence, puisque plusieurs auteurs ont signalé des difficultés à regrouper avec précision des groupes basés sur leur génétique (Funk et al., 2020; Kalinowski, 2011). Ainsi, Gilbert et al. (2012) ont démontré l'incohérence concernant la reproductibilité de l'inférence du paramètre *K*, c'est-à-dire le nombre de groupements de populations, où environ 30 % de leurs résultats divergeaient des papiers originaux. Pour minimiser ces divergences, il faut s'assurer que les paramètres suivants soient suffisamment élevés : le nombre de répétitions, la durée du burnin et les répétitions après le burnin. De plus, comme la méthode d'Evanno ne peut effectuer d'analyse comparative lorsque K=1, une attention particulière doit être accordée dans l'interprétation de cette valeur.

3.1.5 Modèle basé sur des traces environnementales

L'ensemble de nos résultats soutient que la validité du modèle des bases de données actuelles selon l'ethnicité est questionnable, non seulement opérationnellement, mais aussi éthiquement. Elle relance la question de la définition de la base de données la plus représentative de la population d'intérêt à l'enquête, en l'absence de renseignement concernant l'ethnie de l'auteur de faits (Robertson, Vignaux, et Berger, 2016).

Notre approche définissant un modèle de base de données générée à partir des traces environnementales fait abstraction de toutes hypothèses concernant l'ethnicité de l'auteur du crime. Dans ce modèle, nous décortiquons deux scénarios plausibles concernant l'auteur d'un crime. Le scénario le plus simple considère que l'auteur du crime provient de cette zone spécifique relative à la base de données de l'environnement, ce qui corrobore la pertinence de cette base. La probabilité de concordance fortuite est ainsi calculée en confrontant le profil de l'auteur du crime aux échantillons ayant été récoltés dans ce même environnement.

L'autre scénario considère que l'auteur du crime ne provient pas de cette zone spécifique ayant permis de constituer cette base de données. Cependant, la base de données environnementale demeure pertinente et valide dans cette situation. En effet, considérons qu'une concordance ADN soit établie avec un suspect. Soit celui-ci appartient à la population locale échantillonnée (ayant permis de créer la base environnementale), alors le RMP calculé lui est favorable, car il est maximisé du fait que ce suspect appartient à cette population. Soit celui-ci n'appartient pas à la population locale, et dans ce cas on s'attend à ce que son RMP soit plus faible que le précédent, illustrant le fait qu'il n'y appartient pas. Dans le cas contraire indétectable, le RMP calculé est donc conservatoire pour le suspect, puisqu'il lui attribue une probabilité d'identification plus faible que celle attendue.

3.2 Perspectives et conclusion

Cette étude soutient que la composition génétique, c'est-à-dire les fréquences alléliques, tend à être similaire entre les échantillons de référence et de recrutement. Ceux-ci ne reflètent pas des populations différenciées et n'affectent que légèrement (effet presque négligeable) la rareté d'un profil génétique. Enfin, aucune différence structurelle dans la variance génétique n'a été détectée entre les échantillons de référence et de recrutement. Ces résultats confirment nos hypothèses mettant en doute la validité des bases de données divisées par groupes ethniques dans le cas où l'auteur n'est pas associé à un tel groupe, puisque l'échantillon de recrutement, collecté aléatoirement, tend à être équivalent à la population caucasienne de l'échantillon de référence actuellement utilisé par le LSJML.

La manière de générer les bases de données génétiques actuelles utilisées par les laboratoires judiciaires n'étant pas la plus appropriée lorsque l'ethnie de l'auteur du crime demeure inconnue, nous jugeons que la méthode proposée à partir de traces de l'environnement est prometteuse. La prochaine étape de cette recherche aura pour but de valider un échantillon de l'environnement de la ville de Montréal pour davantage définir la population de référence et par le fait même réellement refléter la population d'intérêt dictée par les circonstances de l'enquête. Contrairement aux bases de données actuelles, non seulement cette approche utilisant un échantillon de l'environnement ne requiert pas d'attribuer l'ethnie de l'auteur du crime pour pouvoir confronter les traces ADN à la population d'intérêt pertinente, mais fait également abstraction de toute zone grise concernant les individus composés d'ethnicités mélangées. Ainsi, lors d'une concordance ADN, le profil de l'auteur du crime pourrait éventuellement être comparé (et donc confronté) aux traces d'ADN constamment laissées par les activités humaines et retrouvées dans ce même environnement, dans le but d'évaluer plus précisément la rareté d'un profil génétique. Ce modèle proposé, basé sur des traces ADN de l'environnement, aura certainement de grandes retombées au niveau de l'évaluation des profils génétiques par le laboratoire judiciaire (LSJML), en raison de la diversité génétique (due à la multiethnicité) qui n'est pas liée à la séparation territoriale.

Des problèmes sont survenus lors de la phase d'extraction de nos spécimens (échantillons de l'environnement) : sur l'ensemble des 26 échantillons testés, la plupart n'ont pas pu détecter d'allèles et seuls quelques échantillons ont pu détecter des profils génétiques partiels ou complets. Les raisons potentielles de cette difficulté à générer des profils génétiques pourraient s'expliquer par les variables non contrôlées, comme la météo, pouvant avoir affecté la qualité et l'intégrité des traces ADN. Les échantillons pouvant être restés de nombreux jours (voire semaines), dans l'environnement ont probablement été impactés par la pluie ou le soleil (rayon UV), soit des facteurs pouvant avoir mené à la dégradation de l'ADN.

Cet échantillon de l'environnement présente également deux limitations quant à la représentation réelle d'une population d'intérêt à l'enquête. D'abord, cet échantillon de l'environnement est limité géographiquement quant aux zones sélectionnées pour la collecte des données dans la ville de Montréal : seulement quelques PDQs spécifiques, principalement concentrés dans l'Est de la ville, ont été ciblés pour recueillir les spécimens. Éventuellement, il sera nécessaire de considérer toutes les zones différentes et définies de Montréal pour exploiter pleinement ce modèle, dans le but de servir de base de données de référence réelle pour les enquêtes criminelles.

Ensuite, il se peut que l'échantillon de l'environnement ne représente pas au mieux l'ensemble de la population. En effet, 58 % de l'ensemble des objets récoltés correspondent à des mégots de cigarettes (**Chapitre II, Figure 2.9**). Les autres objets échantillonnés consistent en des ustensiles de plastique, des pailles, des masques chirurgicaux, des gobelets/couvercles, des cannettes, des bouteilles d'eau et autres. Considérant que plus de la moitié des échantillons constituent des mégots de cigarettes, il se peut que l'échantillonnage soit fortement biaisé en faveur des individus fumant la cigarette, ce qui élimine dès lors une grande proportion des individus habitant l'île de Montréal. L'échantillonnage, ciblant spécifiquement certains individus de la population, n'est donc pas forcément aléatoire.

RÉFÉRENCES BIBLIOGRAPHIQUES

Alaeddini, R., Walsh, S. J., et Abbas, A. (2010). Forensic implications of genetic analyses from degraded DNA—A review. *Forensic Science International: Genetics, 4*(3), 148-157. https://doi.org/10.1016/j.fsigen.2009.09.007

. Appendix 1 - STR Allele Frequencies from U.S. Population Data. (2015). In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Interpretation* (pp. 497-518). San Diego: Academic Press.

. Appendix 2 - NRC I & NRC II Recommendations. (2015). In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Interpretation* (pp. 519-528). San Diego: Academic Press.

Balding, D. J. (1999). When can a DNA profile be regarded as unique? *Science & Justice, 39*(4), 257-260. doi:10.1016/s1355-0306(99)72057-5

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), 125-140. doi:10.1016/0379-0738(94)90222-4

Balding, D. J., & Nichols, R. A. (1995). A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica, 96*(1), 3-12.

Balding, D. J., & Steele, C. D. (2015). *Weight-of-evidence for Forensic DNA Profiles*: John Wiley & Sons.

Barbujani, G., Magagni, A., Minch, E., & Cavalli-Sforza, L. L. (1997). An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences of the United States of America, 94*(9), 4516-4519. doi:10.1073/pnas.94.9.4516

Buckleton, J., Curran, J., Goudet, J., Taylor, D., Thiery, A., & Weir, B. S. (2016). Population-specific FST values for forensic STR markers: A worldwide survey. *Forensic science international. Genetics, 23*, 91-100. doi:10.1016/j.fsigen.2016.03.004

Buckleton, J., Triggs, C., & Walsh, S. (2004). *Forensic DNA evidence interpretation*.

Buckleton, J. S., Curran, J. M., & Walsh, S. J. (2006). How reliable is the sub-population model in DNA testimony? *Forensic Science International, 157*(2-3), 144-148. doi:10.1016/j.forsciint.2005.04.004

Budowle, B., & Chakraborty, R. (2001). Population variation at the CODIS core short tandem repeat loci in Europeans. *Legal Medicine, 3*(1), 29-33. https://doi.org/10.1016/S1344-6223(01)00008-6

Budowle, B., Shea, B., Niezgoda, S., & Chakraborty, R. (2001). CODIS STR loci data from 41 sample populations. *Journal of Forensic Sciences, 46*(3), 453-489.

Butler, J. M. (2005). *Forensic DNA typing: biology, technology, and genetics of STR markers*: Elsevier.

Butler, J. M. (2006). Genetics and genomics of core short tandem repeat loci used in human identity testing. *Journal of Forensic Sciences, 51*(2), 253-265. doi:10.1111/j.1556-4029.2006.00046.x

Butler, J. M. (2012a). Chapter 1 - Sample Collection, Storage, and Characterization. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Methodology* (pp. 1-27). San Diego: Academic Press.

Butler, J. M. (2012b). Chapter 5 - Short Tandem Repeat (STR) Loci and Kits. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Methodology* (pp. 99-139). San Diego: Academic Press.

Butler, J. M. (2012c). Chapter 8 - DNA Databases: Uses and Issues. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Methodology* (pp. 213-270). San Diego: Academic Press.

Butler, J. M. (2015a). Chapter 2 - Data, Models, Thresholds. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Interpretation* (pp. 25-46). San Diego: Academic Press.

Butler, J. M. (2015b). Chapter 10 - STR Population Data Analysis. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Interpretation* (pp. 239-279). San Diego: Academic Press.

Butler, J. M. (2015c). Chapter 11 - DNA Profile Frequency Estimates and Match Probabilities. In J. M. Butler (Ed.), *Advanced Topics in Forensic DNA Typing: Interpretation* (pp. 281-308). San Diego: Academic Press.

Chakraborty, R. (1992). Sample size requirements for addressing the population genetic issues of forensic use of DNA typing. *Human Biology, 64*(2), 141-159.

Chakraborty, R., Stivers, D. N., Su, B., Zhong, Y., & Budowle, B. (1999). The utility of short tandem repeat loci beyond human identification: implications for development of new DNA typing systems. *Electrophoresis, 20*(8), 1682-1696. doi:10.1002/(sici)1522-2683(19990101)20:8<1682::Aid-elps1682>3.0.Co;2-z

Champod, C., Evett, I. W., & Jackson, G. (2004). Establishing the most appropriate databases for addressing source level propositions. *Science & Justice, 44*(3), 153-164. https://doi.org/10.1016/S1355-0306(04)71708-6

Chatterjee, S. (2019). Saliva as a forensic tool. *J Forensic Dent Sci, 11*(1), 1-4. doi:10.4103/jfo.jfds_69_18

Chaudhary, R., & Maurya, G. (2020). Restriction Fragment Length Polymorphism.

Council, N. R. (1992). *DNA Technology in Forensic Science*. Washington, DC: The National Academies Press.

Curran, J. M., Buckleton, J. S., & Triggs, C. M. (2003). What is the magnitude of the subpopulation effect? *Forensic Science International, 135*(1), 1-8. doi:10.1016/s0379-0738(03)00171-3

Curran, J. M., Walsh, S. J., & Buckleton, J. (2007). Empirical testing of estimated DNA frequencies. *Forensic Science International: Genetics, 1*(3-4), 267-272. doi:10.1016/j.fsigen.2007.06.004

Dumache, R., Ciocan, V., Muresan, C., & Enache, A. (2016). Molecular DNA Analysis in Forensic Identification. *Clin Lab, 62*(1-2), 245-248. doi:10.7754/clin.lab.2015.150414

Earl, D. A., & Vonholdt, B. (2012). Structure Harvester: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources, 4*.

FBI - Federal Bureau of Investigation. *CODIS and NDIS Fact Sheet*. Retrieved from https://www.fbi.gov/services/laboratory/biometric-analysis/codis/codis-and-ndis-fact-sheet

Ge, J., Eisenberg, A., & Budowle, B. (2012). Developing criteria and data to determine best options for expanding the core CODIS loci. *Investigative genetics, 3*, 1-1. doi:10.1186/2041-2223-3-1

Gendarmerie Royale du Canada. (2020). *Banque nationale de données génétiques du Canada - Rapport annuel 2019-2020.* Retrieved from https://www.rcmp-grc.gc.ca/fr/banque-nationale-donnees-genetiques-du-canada-rapport-annuel-20192020

Gill, P., Sparkes, R., & Kimpton, C. (1997). Development of guidelines to designate alleles using an STR multiplex system. *Forensic Science International, 89*(3), 185-197. https://doi.org/10.1016/S0379-0738(97)00131-X

Goodwin, W., Linacre, A., & Hadi, S. (2011). *An introduction to forensic genetics* (Vol. 2): John Wiley & Sons.

Gouvernement du Canada. (1998). *DNA Identification Act (S.S. 1998, c. 37)*. Retrieved from https://laws-lois.justice.gc.ca/eng/acts/D-3.8/FullText.html

Griffiths Anthony, J. F., Sanlaville, C., & Charmot-Bensimon, D. (2013). *Introduction à l'analyse génétique / Griffiths, Wessler, Carroll,... [et al.]; traduction de la 10e édition américaine par Chrystelle Sanlaville révision scientifique de Dominique Charmot-Bensimon* (6e édition ed.). Bruxelles: De Boeck.

Hares, D. R. (2012). Expanding the CODIS core loci in the United States. *Forensic Science International: Genetics, 6*(1), e52-e54. https://doi.org/10.1016/j.fsigen.2011.04.012

Hares, D. R. (2015). Selection and implementation of expanded CODIS core loci in the United States. *Forensic Science International: Genetics, 17*, 33-34. https://doi.org/10.1016/j.fsigen.2015.03.006

Holsinger, K. E., & Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nature reviews. Genetics, 10*(9), 639-650. doi:10.1038/nrg2611

Hopwood, A. J., Puch-Solis, R., Tucker, V. C., Curran, J. M., Skerrett, J., Pope, S., & Tully, G. (2012). Consideration of the probative value of single donor 15-plex STR profiles in UK populations and its presentation in UK courts. *Sci Justice, 52*(3), 185-190. doi:10.1016/j.scijus.2012.05.005

Jamieson, A., & Bader, S. (2016). *A guide to forensic DNA profiling / edited by Allan Jamieson, Scott Bader*. Chichester, West Sussex, England: Wiley.

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985a). Hypervariable 'minisatellite' regions in human DNA. *Nature, 314*(6006), 67-73. doi:10.1038/314067a0

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985b). Individual-specific ‘fingerprints’ of human DNA. *Nature, 316*(6023), 76-79. doi:10.1038/316076a0

Kanthaswamy, S., & Smith, D. G. (2014). Genetic and ethnohistoric evidence suggest current Native American population datasets in the FBI's CODIS database are not sufficiently representative. *Forensic Science International: Genetics, 13*, e13-15. doi:10.1016/j.fsigen.2014.05.006

Krane, D. E., Allen, R. W., Sawyer, S. A., Petrov, D. A., & Hartl, D. L. (1992). Genetic differences at four DNA typing loci in Finnish, Italian, and mixed Caucasian populations. *Proceedings of the National Academy of Sciences of the United States of America, 89*(22), 10583-10587. doi:10.1073/pnas.89.22.10583

Lalonde, S. A. (2006). Canada's National DNA Data Bank: A Success Story. *Canadian Society of Forensic Science Journal, 39*(2), 39-46. doi:10.1080/00085030.2006.10757135

Lander, E. S. (1989). Population genetic considerations in the forensic use of DNA typing. *Banbury Report 32: DNA technology and forensic Science*, 143-153.

Landsteiner, K. (1901). Uber agglutination-reehenugen Narmalen Menschlihen hlutes wein. *Klin Wehschr, 14*(1), 132-1134.

Lee, H. C., & Ladd, C. (2001). Preservation and collection of biological evidence. *Croatian Medical Journal, 42*(3), 225-228.

Lewontin, R. C., & Hartl, D. L. (1991). Population genetics in forensic DNA typing. *Science, 254*(5039), 1745-1750. doi:10.1126/science.1845040

Malaspinas, A.-S., Slatkin, M., & Song, Y. S. (2011). Match probabilities in a finite, subdivided population. *Theoretical Population Biology, 79*(3), 55-63. doi:10.1016/j.tpb.2011.01.003

Milot, E., Lecomte, M., Germain, H., & Crispino, F. (2013). The National DNA Data Bank of Canada: a Quebecer perspective. *Frontiers in Genetics, 4*(249). doi:10.3389/fgene.2013.00249

Moretti, T. R., Moreno, L. I., Smerick, J. B., Pignone, M. L., Hizon, R., Buckleton, J. S., . . . Onorato, A. J. (2016). Population data on the expanded CODIS core STR loci for eleven populations of significance for forensic DNA analyses in the United States. *Forensic Science International: Genetics, 25*, 175-181. https://doi.org/10.1016/j.fsigen.2016.07.022

National Research Council. (1996). *The Evaluation of Forensic DNA Evidence*. Washington, DC: The National Academies Press.

Ng, J., Oldt, R. F., & Kanthaswamy, S. (2018). Assessing the FBI's Native American STR database for random match probability calculations. *Legal Medicine, 30*, 52-55. https://doi.org/10.1016/j.legalmed.2017.10.012

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics, 155*(2), 945-959. doi:10.1093/genetics/155.2.945

Pritchard, J. K., Wen, W., & Falush, D. (2010). Documentation for structure software: Version 2. *University of Chicago, Chicago, IL*.

Raymond, M., & Rousset, F. (1995). An exact test for population differentiation. *Evolution, 49*(6), 1280-1283. doi:10.1111/j.1558-5646.1995.tb04456.x

Robertson, B., Vignaux, G. A., & Berger, C. E. (2016). *Interpreting evidence: evaluating forensic science in the courtroom*: John Wiley & Sons.

Saad, R. (2005). Discovery, development, and current applications of DNA identity testing. *Proceedings (Baylor University. Medical Center), 18*(2), 130-133. doi:10.1080/08998280.2005.11928051

Silva, N. M., Pereira, L., Poloni, E. S., & Currat, M. (2012). Human neutral genetic variation and forensic STR data. *PLoS One, 7*(11), e49666. doi:10.1371/journal.pone.0049666

Smalldon, K. W., & Moffat, A. C. (1973). The calculation of discriminating power for a series of correlated attributes. *J Forensic Sci Soc, 13*(4), 291-295. doi:10.1016/s0015-7368(73)70828-8

Steele, C. D., Court, D. S., & Balding, D. J. (2014). Worldwide Estimates Relative to Five Continental-Scale Populations. *Annals of Human Genetics, 78*(6), 468-477. https://doi.org/10.1111/ahg.12081

Stern, C. (1943). The Hardy-Weinberg Law. *Science, 97*(2510), 137-138. Retrieved from http://www.jstor.org/stable/1670409

SWGDAM. (2017). SWGDAM Interpretation Guidelines for Autosomal STR Typing by Forensic DNA Testing Laboratories. Retrieved from https://www.swgdam.org/publications

Tang, H., Quertermous, T., Rodriguez, B., Kardia, S. L. R., Zhu, X., Brown, A., . . . Risch, N. J. (2005). Genetic Structure, Self-Identified Race/Ethnicity, and Confounding in Case-Control Association Studies. *The American Journal of Human Genetics, 76*(2), 268-275. https://doi.org/10.1086/427888

Thermo Fisher Scientific. (2018). User Guide: AmpFlSTR Identifiler Plus PCR Amplification Kit. Retrieved from https://www.thermofisher.com/order/catalog/product/4427368

van Oorschot, R. A., Ballantyne, K. N., & Mitchell, R. J. (2010). Forensic trace DNA: a review. *Investigative genetics, 1*(1), 14-14. doi:10.1186/2041-2223-1-14

Verdon, T. J., Mitchell, R. J., & van Oorschot, R. A. (2014). Swabs as DNA collection devices for sampling different biological materials from different substrates. *Journal of Forensic Sciences, 59*(4), 1080-1089.

Vieira, M. L. C., Santini, L., Diniz, A. L., & Munhoz, C. d. F. (2016). Microsatellite markers: what they mean and why they are so useful. *Genetics and Molecular Biology, 39*(3), 312-328. doi:10.1590/1678-4685-GMB-2016-0027

Waples, R. S., & Gaggiotti, O. (2006). INVITED REVIEW: What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity. *Molecular Ecology, 15*(6), 1419-1439. https://doi.org/10.1111/j.1365-294X.2006.02890.x

Weir, B. S., & Cockerham, C. C. (1984). Estimating F-Statistics for the Analysis of Population Structure. *Evolution, 38*(6), 1358-1370. doi:10.2307/2408641

ANNEXE A

CODE R PERMETTANT DE GÉNÉRER DES CARTOGRAPHIES DE LA VILLE DE MONTRÉAL EN Y INTÉGRANT DES DONNÉES SOCIODÉMOGRAPHIQUES

```
# Based on https://ourcodingclub.github.io/2016/12/11/maps_tutorial.html

library(ggplot2)
library(rgdal)
library(raster)
library(rworldmap)
library(ggsn)
library(rworldxtra)
library(rgeos)
library(readxl)
library(sp)
library(rworldmap)

# Set boundaries (latitude and longitude) of Montreal city for the figure:
West <- -74.1
East <- -73.45
South <- 45.35
North <- 45.75

#####ADMINISTRATIVE BOUNDARIES OF MONTREAL CITY

# Specify the path leading to the shapefiles on your computer
SHPdir_admin <- "R DataBase/Limites-adminMTL"
```

```
# Prepare the background map
world <- getMap(resolution = "high")

# Make a SpatialPolygons object which defines a bounding box inside which to crop the world map polygons:
clipper_Qc <- as(extent(West, East, South, North), "SpatialPolygons
# extent(min_longitude, max_longitude, min_latitude, max_latitude)

# Change the coordinate reference systems of both the bounding box and the world map to be equal:
proj4string(clipper_Qc) <- CRS(proj4string(world))

# Clip world by the area of the bounding box, clipper_Qc:
world_clip <- raster::intersect(world, clipper_Qc
# ? Warning messages: 1: In intersect(world, clipper_Qc) : non identical CRS, 2: In RGEOSBinTopoFunc(spgeom1, spgeom2, byid, id, drop_lower_td, unaryUnion_if_byid_false,  : spgeom1 and spgeom2 have different proj4 strings

# Converts the SpatialPolygonsDataFrame to a normal flat dataframe for use in ggplot()
world_clip_f <- fortify(world_clip)

# Add column for coloring land area outside regions of interest with one color:
world_clip_f$co <- rep('1', nrow(world_clip_f))

# Prepare Montreal neighborhoods to plot on top of background map
MTL <- readOGR(dsn = SHPdir_admin, layer = "LIMADMIN")

# Specify the correct CRS, which in this case is EPSG:WGS84 (+proj=longlat +datum=WGS84):
shpdata_MTL <- spTransform(MTL, CRS("+proj=longlat +datum=WGS84"))
shpdata_MTL_clip <- raster::intersect(shpdata_MTL, clipper_Qc)
```

```
# Transform the spatial object to a dataframe, where each polygon will be given an id of which 'region' it is from:
shpdata_MTL_clip_f <- fortify(shpdata_MTL_clip, shpdata_MTL_clip$NOM)
arron<- as.character(shpdata_MTL_clip$NOM)
shpdata_MTL_clip_f$region <- as.character(rep("0",nrow(shpdata_MTL_clip_f)))
for (i in 1:34) {shpdata_MTL_clip_f[shpdata_MTL_clip_f$id == i,]$region <- arron[i]}

# Correction of the spelling of region names
Encoding(shpdata_MTL_clip_f[["region"]]) <- "UTF-8"

# Plot the map with the neighborhoods
jpeg('MTLneighborhoodsOnbackgroundMapADMIN.jpeg', width = 11, height = 8.5, units="in", res = 300)
(map <- ggplot() +
   geom_polygon(data = world_clip_f,
             aes(x = long, y = lat, group = group, fill = co),
             color = "black", size = 0.3) +
geom_polygon(data = shpdata_MTL_clip_f ,
               aes(x = long, y = lat, group = group, fill = region),
               color = "black", size = 0.1) + labs(fill = "Quartiers", x="Longitude", y= "Latitude") + coord_quickmap())
dev.off()

# Plot the map with the neighborhoods (cut the background for esthetic)
jpeg('MTLneighborhoodsAloneADMIN.jpeg', width = 11, height = 8.5, units="in", res = 300)
(map <- ggplot() +
geom_polygon(data = shpdata_MTL_clip_f ,
             aes(x = long, y = lat, group = group, fill = region),
             color = "black", size = 0.1) + labs(fill = "Quartiers", x="Longitude", y= "Latitude") + coord_quickmap())
dev.off()
```

```
#####PDQ BOUNDARIES OF MONTREAL CITY

# Specify the path leading to the shapefiles on your computer
SHPdir_pdq<- "R DataBase/Limites-pdqMTL"

# Prepare the background map
world <- getMap(resolution = "high")

# Make a SpatialPolygons object which defines a bounding box inside which to crop the
world map polygons:
clipper_Qc <- as(extent(West, East, South, North), "SpatialPolygons")
# Extent(min_longitude, max_longitude, min_latitude, max_latitude)

# Changes the coordinate reference systems of both the bounding box and the world map
to be equal:
proj4string(clipper_Qc) <- CRS(proj4string(world))

# Clip world by the area of the bounding box, clipper_Qc:
world_clip <- raster::intersect(world, clipper_Qc
# ? Warning messages: 1: In intersect(world, clipper_Qc) : non identical CRS,
2: In RGEOSBinTopoFunc(spgeom1, spgeom2, byid, id, drop_lower_td,
unaryUnion_if_byid_false,  : spgeom1 and spgeom2 have different proj4 strings

# Converts the SpatialPolygonsDataFrame to a normal flat dataframe for use in ggplot()
world_clip_f <- fortify(world_clip)

# Add column for coloring land area outside regions of interest with one color:
world_clip_f$co <- rep('1', nrow(world_clip_f))

# Prepare Montreal neighborhoods to plot on top of background map
MTLpdq <- readOGR(dsn = SHPdir_pdq, layer = "Limites_PDQ_2016_Lat_Long")
```

```
# Specify the correct CRS, which in this case is EPSG:WGS84 (+proj=longlat +datum=WGS84) :
shpdata_MTLpdq<-spTransform(MTLpdq, CRS("+proj=longlat +datum=WGS84"))
shpdata_MTLpdq_clip <- raster::intersect(shpdata_MTLpdq, clipper_Qc)

# Transform the spatial object to a dataframe, where each polygon will be given an id of which 'region' it is from:
shpdata_MTLpdq_clip_f<-fortify(shpdata_MTLpdq, shpdata_MTL_clip$Nom_PDQ)
pdq<- as.character(shpdata_MTLpdq$Nom_PDQ)
shpdata_MTLpdq_clip_f$region<-as.character(rep("0",
nrow(shpdata_MTLpdq_clip_f)))
for (i in 1:32) { shpdata_MTLpdq_clip_f[shpdata_MTLpdq_clip_f$id == i-1,]$region <- pdq[i]}

# Plot the map with PDQ
jpeg('MTLneighborhoodsOnbackgroundMapPDQ.jpeg', width = 11, height = 8.5, units="in", res = 300)
(map <- ggplot() +
        geom_polygon(data = world_clip_f,
                     aes(x = long, y = lat, group = group, fill = co),
                     color = "black", size = 0.3) +
        geom_polygon(data = shpdata_MTLpdq_clip_f ,
                     aes(x = long, y = lat, group = group, fill = region),
                     color = "black", size = 0.1) + labs(fill= "Postes de quartier", x="Longitude", y= "Latitude") + coord_quickmap())
dev.off()

# Plot the map with PDQ (cut the background for esthetic)
jpeg('MTLneighborhoodsAlonePDQ.jpeg', width = 11, height = 8.5, units="in", res = 300)
(map <- ggplot() +
   geom_polygon(data = shpdata_MTLpdq_clip_f ,
```

```
        aes(x = long, y = lat, group = group, fill = region),
        color = "black", size = 0.1) + labs(fill = "Postes de quartier", x="Longitude",
y= "Latitude") + coord_quickmap())
dev.off()

#####HEATMAP OF SOCIO-DEMOGRAPHIC AND CRIMINALITY DATA

# Import excel data: sexual assault and criminality data of PDQ MTL and demographic
data of ADMIN MTL
library(readxl)

DataAGpdq<-read_excel("R DataBase/Data_agres_sex_PDQ_MTL.xlsx")
DataDemoadmin<-read_excel("R DataBase/Data_démogr_ADMIN_MTL.xlsx")
DataCRpdq<-read_excel("R DataBase/Data_crimes_PDQ_MTL.xlsx")

# Add column Total AG from DataAGpdq to shpdata_MTLpdq_clip_f with match
function
match(shpdata_MTL_clip_f$region,DataDemoadmin$`Quartier ADMIN`)
DataDemoadmin$`PourcMinorite`[match(shpdata_MTL_clip_f$region,DataDemoadmin
$`Quartier ADMIN`)]
shpdata_MTL_clip_f$PourcMinorite=DataDemoadmin$`PourcMinorite`[match(shpdata
_MTL_clip_f$region,DataDemoadmin$`Quartier ADMIN`)]
#shpdata_MTL_clip_f$PourcMinorite<-NULL

match(shpdata_MTL_clip_f$region,DataDemoadmin$`Quartier ADMIN`)
DataDemoadmin$`PourImmdirectind`[match(shpdata_MTL_clip_f$region,DataDemoad
min$`Quartier ADMIN`)]
shpdata_MTL_clip_f$PourcentagePopImmigrante=DataDemoadmin$`PourImmdirectin
d`[match(shpdata_MTL_clip_f$region,DataDemoadmin$`Quartier ADMIN`)]
#shpdata_MTL_clip_f$PourcentagePopImmigrante<-NULL
```

```
match(shpdata_MTLpdq_clip_f$region,DataAGpdq$PDQ)
DataAGpdq$`Total
AS`[match(shpdata_MTLpdq_clip_f$region,DataAGpdq$PDQ)]
shpdata_MTLpdq_clip_f$AgressionsSexuelles=DataAGpdq$`Total
AS`[match(shpdata_MTLpdq_clip_f$region,DataAGpdq$PDQ)]
#shpdata_MTLpdq_clip_f$AgressionSex<-NULL

match(shpdata_MTLpdq_clip_f$region,DataCRpdq$PDQ)
DataCRpdq$`Total
Crime`[match(shpdata_MTLpdq_clip_f$region,DataCRpdq$PDQ)]
shpdata_MTLpdq_clip_f$Criminalitetotale=DataCRpdq$`Total
Crime`[match(shpdata_MTLpdq_clip_f$region,DataCRpdq$PDQ)]
#shpdata_MTLpdq_clip_f$Criminalitetotale<-NULL

# Incorporate data into heatmap and export into jpeg files
jpeg('GraphMinoADMIN.jpeg', width = 11, height = 8.5, units="in", res = 300)
(map <- ggplot() +
        geom_polygon(data = shpdata_MTL_clip_f ,
          aes(x = long, y = lat, group = group, fill = PourcMinorite),
          color = "black", size = 0.1) + labs(fill = "Pourcentage de
minorité",x="Longitude", y="Latitude") + coord_quickmap())
dev.off()

jpeg('GraphImmADMIN.jpeg', width = 11, height = 8.5, units="in", res = 300)
(map <- ggplot() +
        geom_polygon(data = shpdata_MTL_clip_f ,
          aes(x = long, y = lat, group = group, fill = PourcentagePopImmigrante),
          color = "black", size = 0.1) + labs(fill = "Pourcentage de la population
immigrante", x="Longitude", y= "Latitude") + coord_quickmap())
dev.off()
```

```
jpeg('GraphAGPDQ.jpeg', width = 11, height = 8.5, units="in", res = 300)
(map <- ggplot() +
   geom_polygon(data = shpdata_MTLpdq_clip_f,
            aes(x = long, y = lat, group = group, fill = AgressionsSexuelles),
            color = "black", size = 0.1) + labs(fill = "Pourcentage d'agressions sexuelles",
x="Longitude", y= "Latitude") + coord_quickmap())
dev.off()

jpeg('GraphCRPDQ.jpeg', width = 11, height = 8.5, units="in", res = 300)
(map <- ggplot() +
   geom_polygon(data = shpdata_MTLpdq_clip_f,
            aes(x = long, y = lat, group = group, fill = Criminalite totale),
            color = "black", size = 0.1) + labs(fill = "Pourcentage de la criminalité totale",
x="Longitude", y= "Latitude") + coord_quickmap())
dev.off()
```

ANNEXE B

CODE R PERMETTANT DE GÉNÉRER DES ANALYSES GENEPOP À PARTIR DE DONNÉES GÉNÉTIQUES DES ÉCHANTILLONS

```
# Based on: https://kimura.univ-montp2.fr/~rousset/Genepop.htm

library(sp)
library(genepop)
library(ggplot2)
library(xlsx)
library(dplyr)
library(reshape2)
library(scales)
library(lsjml)
library(xlsx)
library(DNAtools)
library(ggplot2)
library(scales)
library(forcats)

#####GENEPOP DATA FORMAT

#Tests of genic and genotypic differentiation
#(Emplacement fichier mère, direction fichier outputfile), fct cat lire fichier texte dans R
unlink("Donneestestdiff.txt")
test_diff("C:/Users/jessi/Documents/CompilationsdonneesGenepopformat.txt", genic =
TRUE, pairs = FALSE, outputFile="Donneestestdiff.txt")
cat(readLines("Donneestestdiff.txt"), sep="\n")
```

```
#Allele and genotype frequencies per locus and per sample
unlink("Donneesbasicinfo.txt")
basic_info("C:/Users/jessi/Documents/CompilationsdonneesGenepopformat.txt",
outputFile="Donneesbasicinfo.txt")
cat(readLines("Donneesbasicinfo.txt"), sep="\n")

#Hardy Weinberg test, heterozygote deficiency (Fis)
unlink("DonneesHWT.txt")
test_HW("C:/Users/jessi/Documents/CompilationsdonneesGenepopformat.txt",
which="deficit",batches = 500, outputFile ="DonneesHWT.txt")
cat(readLines("DonneesHWT.txt"), sep="\n")

#Fst estimation
unlink("DonneesFst.txt")
Fst("C:/Users/jessi/Documents/CompilationsdonneesGenepopformat.txt", outputFile =
"DonneesFst.txt")
cat(readLines("DonneesFst.txt"), sep="\n")

#Extraire fréquence d'un fichier .txt
lines <- readLines("Donneesbasicinfo.txt")
loci <- list()

for (i in 1:length(lines)) {
  if (grepl("Tables of allelic frequencies for each locus:", lines[i], fixed = TRUE)) {
    debut <- i
    break}}
debutLocus <- debut + 2
stop <- FALSE

while (grepl("Locus", lines[debutLocus], fixed = TRUE)) {
  nomLocus <- substring(lines[debutLocus], 9)
  ligneAlleles <- debutLocus + 4
```

```
  alleles <- as.numeric(scan(text = lines[ligneAlleles], what = ""))
  alleles <- ifelse(alleles < 50, alleles, alleles / 10)
  ligneFrequences <- ligneAlleles + 1
  data.locus = list()
  alleles <- as.character(alleles)
  data.locus[["Allele"]] = factor(alleles, levels = alleles)
  while (lines[ligneFrequences] != "") {
    content <- scan(text = lines[ligneFrequences], what = "")
    nomPop = content[1]
    freqs <- as.numeric(content[2:(length(content) - 1)])
    data.locus[[nomPop]] = c(freqs)
    ligneFrequences <- ligneFrequences + 1
  }
  df <- data.frame(data.locus)
  loci[[nomLocus]] <- df
  debutLocus <- ligneFrequences + 1}

#Regroup lists of dataframe into one big dataframe for RMP calculations
Tableaufreq15STR<-bind_rows(loci, .id="column_label")
#Change pop names
colnames(Tableaufreq15STR)<-c("Locus", "Allele", "Frequency", "Frequency2")

#Extraire + reshape dataframe pour générer plot
bubble_plot <- function(datatable, locus, filename=NULL, height=NA) {
  datatable = datatable %>%
    filter(Locus == locus) %>%
    rename("Recruitment" = Frequency) %>%
    rename("Reference" = Frequency2) %>%
    select(-one_of("Locus")) %>%
    mutate(Allele = as.factor(as.numeric(as.character(Allele)))) %>%
    arrange(Allele)
```

```
  melted <- reshape2::melt(datatable, id.vars = "Allele", variable.name = "Samples",
value.name = "Size")

  if (!is.null(filename)) {
    # jpeg(filename)}

  pop_vector = c(datatable[["Recruitment"]], datatable[["Reference"]])
  g <- ggplot(melted, aes(x=Samples, y= Allele)) +
    geom_point(aes(size = pop_vector, alpha = pop_vector))+
    scale_size_continuous(name = "Frequencies", breaks=seq(from=0, to=0.5, by=0.05))+
    scale_alpha_continuous(name = "Frequencies",breaks=seq(from=0, to=0.5,
by=0.05))+
    labs(x="Samples", y= "Alleles",title = paste("Locus", locus))

  if (!is.null(filename)) {
    ggsave(filename, plot=g, height=height)
    # dev.off()
  } else {
    return(g)}}

#plot de fréquence
bubble_plot(Tableaufreq15STR, "D8S1179", "GGPLOT-D8S1179.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "D21S11", "GGPLOT-D21S11.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "D7S820", "GGPLOT-D7S820.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "CSF1PO", "GGPLOT-CSF1PO.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "D3S1358", "GGPLOT-D3S1358.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "TH01", "GGPLOT-TH01.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "D13S317", "GGPLOT-D13S317.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "D16S539", "GGPLOT-D16S539.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "D2S1338", "GGPLOT-D2S1338.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "D19S433", "GGPLOT-D19S433.jpeg", height=5)
```

```
bubble_plot(Tableaufreq15STR, "vWA", "GGPLOT-vWA.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "TPOX", "GGPLOT-TPOX.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "D18S51", "GGPLOT-D18S51.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "D5S818", "GGPLOT-D5S818.jpeg", height=5)
bubble_plot(Tableaufreq15STR, "FGA", "GGPLOT-FGA.jpeg", height=5)

##### EQUATION FREQ.GENO

freq.geno.11 <- function(freq){
   freq.table <- data.frame(expand.grid(freq[,"Allele"], freq[,"Allele"]),
expand.grid(freq[,"Frequency"], freq[,"Frequency"]))

  colnames(freq.table) <- c("Allele.1", "Allele.2", "Freq.Allele.1", "Freq.Allele.2")
   freq.table <-freq.table[which(freq.table[,1]<=freq.table[,2]),]
   freq.table$Freq.Geno <- freq.table[,3] * freq.table[,4] * (as.numeric(freq.table[,1] !=
freq.table[,2]) + 1)
   freq.table}

##### COMPUTE THE RMP EQ8

Eq8 <- function(DNA.profiles, freq=NULL, freq.min=0, theta=0.01){
   if(!is.null(freq) & class(freq)=="freqdata") freq<-freq@frequencies
   if (class(DNA.profiles) == "array") {
   if (!length(dim(DNA.profiles)) == 3 | !dim(DNA.profiles)[2] == 2) stop("Les
dimensions de ce tableau sont inadequates")
   loci <- dimnames(DNA.profiles)[[3]]
   genotypes <- DNA.profiles}
   if (class(DNA.profiles) == "genedata") {
   loci <- sort(DNA.profiles@loci)
   genotypes <- DNA.profiles@genotypes
   if(is.null(freq)) freq <- DNA.profiles@frequencies@frequencies}
```

```
    if (is.null(freq)) stop("Aucun jeu de frequence d'allele n'est specifie ou contenu dans les donnees")
    L <- length(loci)
    N.loci.ID <- NULL
    freq.DNA <- rep(1, dim(genotypes)[1])
    for (i in 1:dim(genotypes)[1]){
    N.loci.ID[i] = 0
       for (l in 1:L) {
     rows <- freq[freq[, 1] == loci[l], ]
          if(all((!is.na(genotypes[i, , l]))) & nrow(rows)>0){
       N.loci.ID[i] = N.loci.ID[i]+1
       homoz <- as.numeric(genotypes[i, 1, l]!=genotypes[i, 2, l]) + 1
       freq.al1 <- rows[rows[, 2] %in% genotypes[i, 1, l], 3]
       if(length(freq.al1) == 0) freq.al1 <- 0
       if(freq.al1 < freq.min) freq.al1 <- freq.min
       freq.al2 <- rows[rows[, 2] %in% genotypes[i, 2, l], 3]
       if(length(freq.al2) == 0) freq.al2 <- 0
       if(freq.al2 < freq.min) freq.al2 <- freq.min
            freq.DNA[i] <- freq.DNA[i] * homoz * freq.al1 * freq.al2}}}
  if(any(!loci%in%freq$Locus)) {
    cat("Les loci suivants ne sont pas inclus dans les donnees de frequences choisies et ne sont pas pris en compte dans le calcul:\n",loci[!loci%in%freq$Locus],"\n\n")}
  loci.not.used=loci[!loci%in%freq$Locus]
  if(length(loci.not.used)==0) loci.not.used <- "none"
  return(list(loci.not.used=loci.not.used, Nb.loci.for.each.individual=N.loci.ID, Multilocus.genotype.frequencies=freq.DNA))}

#Pop1 = Recrutement
#Pop2 = Référence
```

```
#Split dataframe in 2
Tableaufreq15STR_Population1<- Tableaufreq15STR[ , c("Locus",
"Allele","Frequency")]
Tableaufreq15STR_Population2<- Tableaufreq15STR[ , c("Locus",
"Allele","Frequency2")]

#Générer génotype aléatoire
Tableaufreq15STR_Population1$Allele<-
as.numeric(as.character(Tableaufreq15STR_Population1$Allele))
Tableaufreq15STR_Population2$Allele<-
as.numeric(as.character(Tableaufreq15STR_Population2$Allele))
colnames(Tableaufreq15STR_Population2)<-c("Locus", "Allele", "Frequency")
sapply(Tableaufreq15STR, class)
profils <- genosim(n=1000, freq=Tableaufreq15STR_Population1)
profils2 <- genosim(n=1000, freq=Tableaufreq15STR_Population2)

# Calcul de la probabilité P_G de chaque génotype avec l'équation 8 dans la population
P_G1.1<- Eq8(DNA.profiles = profils, theta = 0.01, freq =
Tableaufreq15STR_Population1, freq.min = 0.0198)
P_G1.2<- Eq8(DNA.profiles = profils, theta = 0.01, freq =
Tableaufreq15STR_Population2, freq.min = 0.0091)
P_G2.2<- Eq8(DNA.profiles = profils2, theta = 0.01, freq =
Tableaufreq15STR_Population2, freq.min = 0.0091)
P_G2.1<- Eq8(DNA.profiles = profils2, theta = 0.01, freq =
Tableaufreq15STR_Population1, freq.min = 0.0198)

# ... visualisation du résultat
P_G1.1
P_G1.2
P_G2.2
P_G2.1
```

```
numeric_to_excel_string <- function(v) {
cat(paste(gsub("\\.", ",", as.character(v)), collapse="\n"))}

numeric_to_excel_string(P_G1.1$Multilocus.genotype.frequencies)
numeric_to_excel_string(P_G1.2$Multilocus.genotype.frequencies)
numeric_to_excel_string(P_G2.2$Multilocus.genotype.frequencies)
numeric_to_excel_string(P_G2.1$Multilocus.genotype.frequencies)
```