PROJET BALSAC

Volet : Infrastructures

Validation des généalogies reconstituées à BALSAC à partir de données génétiques

par

Michèle Jomphe

Décembre 2011

(Document I-C-243)

(adresse postale)

PROJET BALSAC
Université du Québec à Chicoutimi
555 boulevard de l'Université
Chicoutimi (Québec) G7H 2B1
Téléphone: (418) 545-5517
Télécopieur: (418) 545-5518
courriel : balsac@uqac.ca

Introduction

À toutes les étapes de sa construction, le fichier BALSAC est soumis à d'importantes mesures de validation et de correction, que ce soit au moment de la saisie des actes de l'état civil ou au cours des diverses étapes de jumelage qui consistent à reconstituer les familles et les liens généalogiques.

Le fichier BALSAC ayant maintenant atteint un niveau de développement très avancé, il est utilisé comme l'outil principal pour des reconstitutions généalogiques à des fins de projets de recherche, notamment en génétique. La qualité des données étant un facteur essentiel dans ce type de recherches, il apparaît important d'estimer un taux d'erreur global des généalogies reconstituées à BALSAC. À cette fin, les données génétiques obtenues par le séquençage et le génotypage de l'ADN extrait de la salive de participants recrutés dans quatre régions du Québec pour le projet de recherche *Modèles génétiques et histoire de la population du Québec*, sous la responsabilité de Damian Labuda et d'Hélène Vézina, ont été comparées aux généalogies des mêmes individus. Les résultats d'haplogroupes obtenus par l'ADN mitochondrial et le chromosome Y ont été juxtaposés aux lignées généalogiques maternelles et paternelles reconstituées à BALSAC. Le présent document décrit la démarche, les incohérences observées et les taux d'erreur estimé.

L'auteure tient à remercier les participants qui ont bien voulu collaborer au projet de recherche en fournissant un échantillon de salive duquel a été extrait les données génétiques et en remplissant un questionnaire généalogique qui a servi de point de départ aux généalogies reconstituées à BALSAC. Merci également à Claudia Moreau qui a produit les données génétiques à l'Hôpital Ste-Justine et identifié les incohérences, à Ève-Marie Lavoie qui a participé à l'élaboration des schémas des lignées maternelles et paternelles, de même qu'au personnel du Projet BALSAC qui a procédé aux reconstitutions généalogiques. Un merci spécial à M. France Néron qui a examiné en détail toutes les incohérences généalogiques relevées au cours de cet exercice. Mario Bourque a collaboré à l'interprétation des résultats et a révisé le présent document.

1. Données disponibles

Les données qui ont été utilisées pour cet exercice proviennent d'un projet de recherche intitulé « *Modèles génétiques et histoire de la population du Québec* » sous la responsabilité de Damian Labuda et d'Hélène Vézina. Dans le cadre de cette étude, la collaboration de 794 sujets recrutés sur une base volontaire en Gaspésie, au Saguenay, sur la Côte-Nord ou à Montréal a été demandée. Après avoir signé un formulaire de consentement, ceux-ci devaient fournir un échantillon de salive qui a été acheminé pour analyse au laboratoire de Damian Labuda, à l'Hôpital Ste-Justine. Les participants devaient également transmettre au Projet BALSAC un questionnaire généalogique qu'ils avaient rempli de manière à fournir l'information nécessaire pour démarrer la reconstitution de leur généalogie.

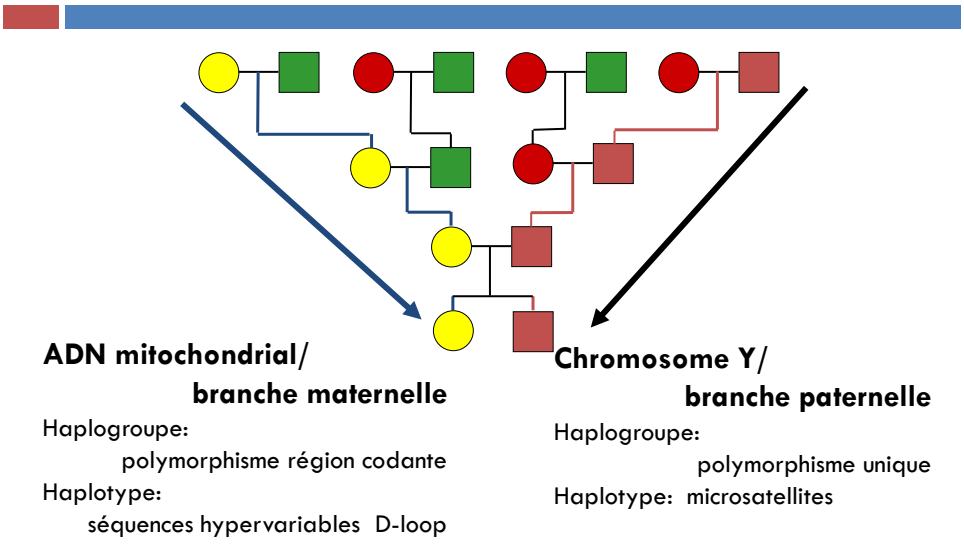
1.1. Données généalogiques

Les généalogies de 794 participants ont été reconstituées avec le module RETRO, le fichier BALSAC et les outils généalogiques complémentaires disponibles. Ces généalogiques comptent plus de 2 800 000 ancêtres dont 91 000 ancêtres distincts. Pour les fins du présent exercice de validation, seules les lignées maternelles et paternelles des participants ont été examinées. Les lignées maternelles comptent 794 points de départ, 8 695 ancêtres dont 6 580 ancêtres distincts. Pour les lignées paternelles, seuls les participants masculins (N=351) sont utiles parce que porteurs d'un chromosome Y. Leurs lignées paternelles comptent 3 161 ancêtres dont 2 829 ancêtres distincts.

1.2. Données génétiques

Les données génétiques ont été obtenues à partir d'un échantillon de salive des participants. L'équipe du laboratoire de Damian Labuda a effectué le séquençage des régions hypervariables I et II (D-loop) et au génotypage de trois positions de la région codante de l'ADN mitochondrial. Des marqueurs du chromosome Y ont également été examinés, tel qu'illustré à la figure 1.

FIGURE 1: Données génétiques



1.3. Comparaison des informations généalogiques et génétiques

Une fois les généalogies reconstituées à BALSAC, les lignées maternelles et paternelles ont été dessinées en utilisant la fonction « plot.lig » de la librairie graphique de GENLIB développée pour le logiciel TIBCO Spotfire S+.

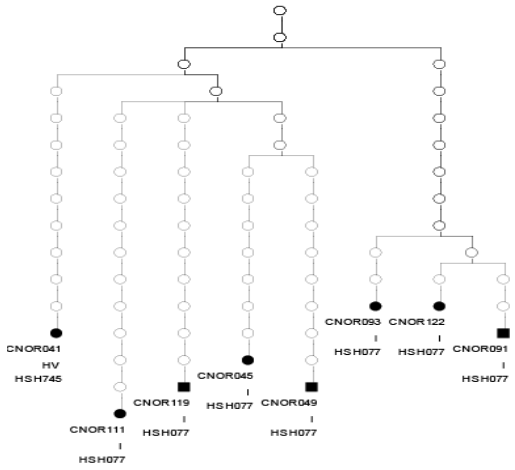
Les informations génétiques sur les participants consistaient en des résultats d'haplotypes et d'haplogroupes tel que décrit à la figure 2.

FIGURE 2: Exemple d'informations génétiques sur les participants

Participant		ADN-mt		ADN-Y	
IND	CODE	Haplotype	Haplogroupe	Haplotype	Haplogroupe
409060	JCL317	HSH712	H1	I	Y7-127
408820	GEL212	HSH770	H1	K(xO,P)	Y7-129K
408211	DS102	HSH771	H1		
408935	JCL350	HSH771	H1		
665100	Sag0860-2200	HSH098	J	R1(xR1a)	Y7-079
409390	GEL401	HSH475	J	R1(xR1a)	Y7-085
43023	Mtl053	HSH476	J		

Les figures 3 et 4 présentent des exemples où les informations généalogiques des lignées généalogiques maternelles et paternelles reconstituées à BALSAC sont combinées avec les données d'haplogroupes et d'haplotypes obtenues au laboratoire de Damian Labuda.

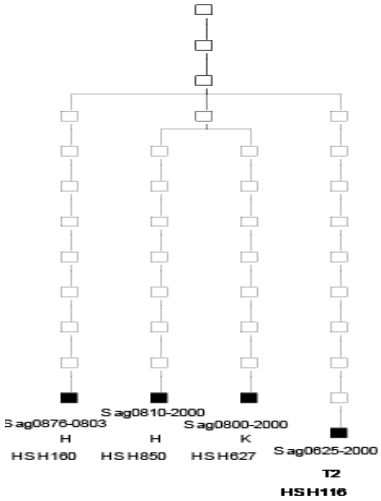
FIGURE 3: Exemple de branches maternelles avec information génétique sur les participants



Branches maternelles:

- 794 points de départ
- 8695 ancêtres
- 6580 ancêtres distincts

FIGURE 4: Exemple de branches paternelles avec information génétique sur les participants



Branches paternelles:

- 794 points de départ
- 7290 ancêtres
- 6051 ancêtres distincts

Branches paternelles génétiquement utiles:

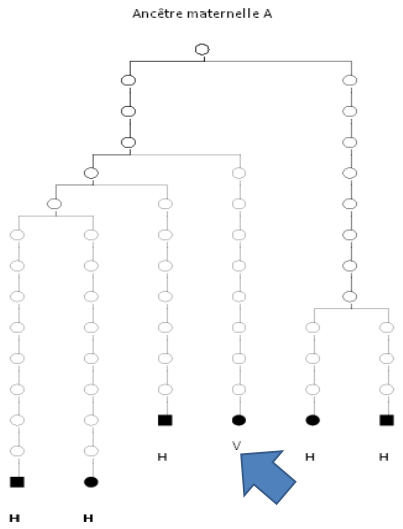
- 351 points de départ masculins
- 3161 ancêtres
- 2829 ancêtres distincts

2. Identification des incohérences généalogiques vs génétiques

En juxtaposant les informations généalogiques et les informations génétiques, il est possible d'identifier des incohérences dans des grappes d'apparentés qui, théoriquement, devraient partager un même haplogroupe s'ils sont issus d'un ancêtre maternel ou paternel commun.

La figure 5 illustre une incohérence observée dans des lignées maternelles. L'individu d'haplogroupe V ne peut être relié généalogiquement à l'ancêtre maternelle A qui a transmis l'haplogroupe H aux autres individus via son ADN mitochondrial.

FIGURE 5: Exemple d'une incohérence observée dans une lignée maternelle



2.1. Erreurs de sur-jumelage

Un examen approfondi de tous les liens généalogiques des lignées maternelles où sont retrouvées des incohérences permet d'identifier des erreurs de reconstitution.

La figure 6a illustre une situation où un individu d'haplogroupe H1 a été associé par erreur à des individus d'haplogroupe V via une ancêtre maternelle commune A. Après vérification de tous les liens généalogique de la branche H1 une erreur sur-jumelage dû à l'homonymie est identifiée. La figure 6b montre la fiche de couple en erreur dans laquelle Marcelline Levesque apparaissait comme enfant du couple Elie Levesque et Marie (Scholastique) Caron marié à l'Islet en 1807. Dans la figure 6c, Marcelline Levesque est replacée dans la famille du couple homonyme Elie Levesque et Marie (Domitilde) Caron marié dans Kamouraska en 1832. Le fichier BALSAC a donc pu être corrigé grâce a cet exercice.

FIGURE 6a: Exemple d'erreur de sur-jumelage

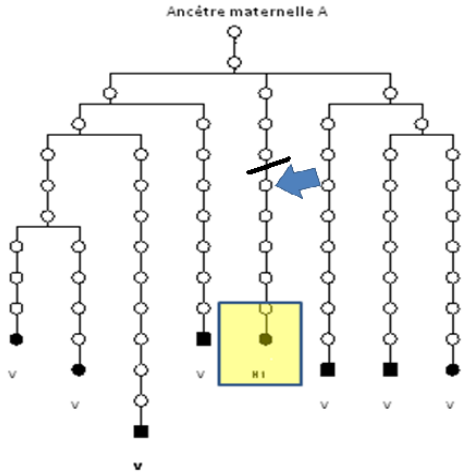


FIGURE 6b: Fiche de couple erronée

15 Juin 2010, 14:00 **Fiche de Couple** **Base de données BALSAC**

Famille 1618201 (11) LEVESKE ELIE KARON ENJELIKE MARIE

Époux		Épouse		Acte	Date Ev	Par	Résid		Professior	E S T Prénom	Stage
LEVEQUE	ELIE	CARON	MARIE ANGELIQU	8746972-1	05-oct.-1807	553	116	999 9	132 9	132 C M M	MAR
			Fils de LEVEQUE		ANTOINE			GAGNON		MADELEINE	
						553	229	999 9	132 9	999 C F M	C=N
			Fille de CARON		LOUIS			BOUCHARD		MARIE ANGELIQUE	
LEVEQUE	ELIE	CARON	MARIE	8574491-3	26-oct.-1830	510	229	220 9	999 9	999 C F M MARIE OSITHE	
Conj:	MICHAUD	FRANCOIS	Fils de MICHAUD		JEAN			LEVASSEUR		APPOLINE	
LEVEQUE	ELIE	CARON	MARIE	8743721-2	09-août-1831	424	229	112 9	999 9	132 C M M EDOUARD	
Conj:	VAILLANCOURT	MONIQUE	Fille de VAILLANCOURT		JEAN BAPTISTE			DUBE		ROSE	
LEVEQUE	HELIE	CARON	MARIE	8746835-3	21-août-1832	553	229	999 9	999 9	999 C F M SOPHIE	
Conj:	GAMACHE	MARCEL	Fils de GAMACHE		JEAN BAPTISTE			COTE		MARGUERITTE	
LEVEQUE	HELI	CARON	MARIE	8743819-3	05-nov.-1840	553	229	999 9	999 9	999 C F M SCHOLASTIQUE	
Conj:	PELLETIER	EOSITHEE	Fils de PELLETIER		CHARLES			THIBAULT		GENEVIEVE	
LEVEQUE	ELIE	CARON	MARIE	8743842-3	16-nov.-1841	424	229	190 9	256 9	999 C F M SERAPHINE	
Conj:	PHILIBERT	FREDERIC	Fils de PHILIBERT		MICHEL			OUELLET		JULIE	
LEVESQUE	HELI	CARON	MARIE	8744318-3	14-nov.-1843	424	999	190 9	999 9	999 C F M MODESTE	
Conj:	BELANGER	LOUIS	Fils de BELANGER		FRANCOIS			DUBE		LOUISE	
LEVEQUE	CLEMENT	CARON	MARIE SHOLASTI	8788823-3	18-juin-1844	536	229	792 9	999 9	999 C F M LUCE	
Conj:	BOULET	GEORGE	Fils de BOULET		JOSEPH			METIVIER		MARIE	
LEVEQUE	HELI	CARON	MARIE	8744880-2	13-janv.-1846	424	999	190 9	999 9	999 C M M FRANCOIS	
Conj:	DESSEIN ST PIERR	SOULANGE	Fille de DESSEIN ST PIERRE		ANTOINE			DUBE		MARIE MAGDELEINE	
LEVEQUE	ELIE	CARON	MARIE SCHOLAST	8577232-3	07-janv.-1851	510	999	220 9	999 9	999 C F M ESTHER	
Conj:	SOUCY	JEAN BAPTISTE	Veuf de GAGNON		PRISCILLE						
LEVEQUE	ELIE	CARON	MARIE	8770663-3	25-nov.-1862	253	129	126 9	999 9	999 C F M MARCELLINE	
Conj:	L'ETOILE L'ITALI	ALEXANDRE	Fils de L'ETOILE L'ITALIEN		LOUIS			BRISSON		CELESTE	

Nombre de lignes: 24 Statut de jumelage: 2 Statuts: H: 0 F: 0 Note: X

Observation

SCHOLASTIQUE= SUBSTITUTION NOMINATIVE

Avant correction

FIGURE 6c: Fiche de couple corrigée

22 Juin 2010, 14:13 — Fiche de Couple — Base de données BALSAC

Famille 1203631 (5) LEVESKE ELIE KARON DOMITILLE MARIE

Époux		Épouse		Acte	Date Ev	Par	Résid		Profession	E S T Prénom	Stage
L'EVESQUE	ELIE	CARON	MARIE DOMITILL	8516733-'	10-janv.-1832	213	116		9	132 ? M M	MAI
		***** Acte incomplet *****									
						213	116	999 9 999 9		999 C F M	C=1
			Fille ce CARON		LOUIS		CHAROIE			MARIE ROSALIE	
LEVEQUE	ELIE	CARON	MARIE DOMITILD	8569993-4	10-oct.-1842	266	129		9	132 V M M	M=M-I
Conj: GAGNE		ANGELE	Fille ce GAGNE		CHARLES		BOUCHARD			JOSEPHTE	
LEVEQUE	ELIE	CARON	LUCIE	8394160-2	28-oct.-1856	385	108	999 9 999 9		132 C M M ELIE	
Conj: DUMAIS		MAGDELAINE	Fille ce DUMAIS		FRANCOIS		CANTIN			BASILISSE	
LEVEQUE	ELIE	CARON	DESANGES	8394810-2	24-janv.-1859	385	108	999 9 999 9		132 C M M HONORE	
Conj: BLANCHET		ARTEMISE	Fille ce BLANCHET		JOSEPH		DAIGON			DIDIANE	
LEVEQUE	ELIE	CARON	MARIE	8774863-3	25-nov.-1862	263	129	126 9 999 9		999 C F M MARCELLINE	
Conj: L'ETOILE L'ITALI		ALEXANDRE	Fils ce L'ETOILE L'ITALIEN		LOUIS		BRISSON			CELESTE	

Nombre de lignes: 12 — Statut de jumelage: 2 Statuts: H: 0 F: 0 Note: X

Observation

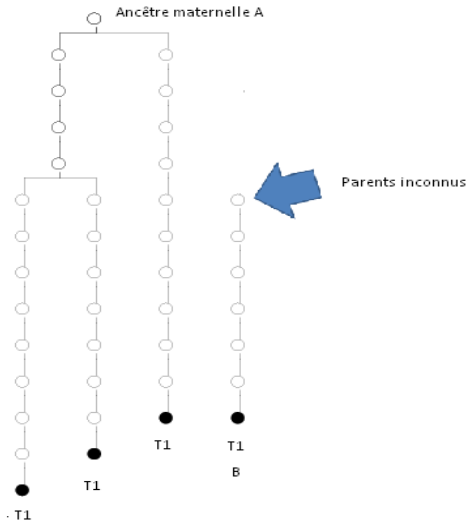
MARIE DOMITHILDE, LUCIE, DESANGES (PRENOMS OBSERVES)

Après correction

2.2. Erreurs de sous-jumelage

La figure 7 illustre une erreur de sous-jumelage. Dans ce cas-ci, les quatre individus d'haplogroupe T1 devraient partager la même ancêtre maternelle A.
Après vérification, les sources généalogiques disponibles à BALSAC indiquent que la branche maternelle de l'individu B se termine par un individu dont les parents sont inconnus. Il n'est donc pas possible poursuivre la généalogie et de recréer le lien permettant de relier les quatre individus.

FIGURE 7: Exemple de sous-jumelage



3. Description des erreurs identifiées dans les lignées maternelles et paternelles et calcul du taux d'erreur.

La figure 8 décrit l'ensemble des incohérences observées dans les lignées maternelles et paternelles. Toutes les incohérences détectées (N=53) ont été minutieusement vérifiées et corrigées lorsque l'erreur généalogique était identifiable.

FIGURE 8: Description des incohérences identifiées dans les lignées maternelles et paternelles

TYPE D'INCOHÉRENCES	LIGNÉES MATERNELLES		LIGNÉES PATERNELLES		CAUSES D'ERREUR
	Erreurs (N)	Taux d'erreur (%)	Erreurs (N)	Taux d'erreur (%)	
Sur-jumelage	4	0,09			Homonymie, changement de patronyme, noms inconnus
	2	0,05			Branches généalogiques anglophones
	5	0,11	1	0,09	Branches généalogiques hors Québec
	13	0,3	8	0,73	Erreur introuvable
TOTAL sur-jumelage	**24**	**0,55**	**9**	**0,82**	
Sous-jumelage	9	NA			Branches généalogiques interrompues (adoption, parents inconnus, enfants naturels)
	6	NA			Branches généalogiques anglophones
	3	NA			Branches généalogiques hors Québec
	2	NA			Erreur introuvable
TOTAL sous-jumelage	**20**	**NA**			

Comme les erreurs de sur-jumelage ne peuvent être identifiées que dans les grappes de plusieurs apparentés partageant un ancêtre commun, le dénominateur pour le calcul du taux d'erreur correspond au nombre d'individus distincts (ou le nombre de liens généalogiques) retrouvés dans les grappes d'apparentés.
Globalement, le taux d'erreur dû à des liens généalogiques erronés (erreurs de sur-jumelage) est estimé à 0,55% dans les lignées maternelles (N=24) et de 0,82% dans les lignées paternelles (N=9).

FIGURE 9: Calcul du taux d'erreur de sur-jumelage dans les lignées maternelles et paternelles

Lignées maternelles

Nombre d'erreurs de sur-jumelage = 24

Dénominateur: 4383 individus distinct compris dans les grappes d'apparentés par les branches maternelles = nombre de liens généalogiques effectués

Taux d'erreur de sur-jumelage dans les lignées maternelles: 24/4383*100=0,55 %

Lignées paternelles

Nombre d'erreurs de sur-jumelage = 9

Dénominateur: 1102 individus distincts compris dans les grappes d'apparentés par les branches paternelles = nombre de liens généalogiques effectués

Taux d'erreur de sur-jumelage dans les lignées paternelles : 9/1102*100=0,82 %

Par ailleurs, 20 liens généalogiques n'ont pu être effectués (cas de sous-jumelage) mais le taux d'erreur ne peut être calculé puisque le dénominateur est indéterminé. Il faut cependant noter qu'une part des erreurs est attribuable à l'utilisation d'autres outils généalogiques que BALSAC, notamment lors de la reconstitution de branches généalogiques hors Québec, ou à la qualité déficiente des sources elles-mêmes, par exemple les actes de l'état civil non catholiques incomplets. Une discussion sur les causes d'erreurs se retrouve à la section suivante du présent document.

Discussion

L'exercice de comparaison des lignées généalogiques maternelles et paternelles produites par BALSAC avec des informations sur les haplogroupes obtenues par l'analyse de l'ADN-mitochondrial et du chromosome Y de participants volontaires a permis d'établir un taux d'erreur de sur-jumelage de 0,55% dans les lignées généalogiques maternelles et de 0,82% dans les lignées généalogiques paternelles. Le calcul du taux jumelage non résolu ne peut être calculé étant donné que le dénominateur (le nombre d'individus qui auraient dû se retrouver dans les grappes d'apparentés) est inconnu.

Parmi les erreurs de sur-jumelage, quatre (4) sont directement attribuables à des défaillances dans la reconstitution des familles par BALSAC et elles représentent un taux d'erreur de 0,09%. Elles sont dues à de l'homonymie ou à des changements de noms et prénoms dans les actes. Elles ont été documentées et corrigées dans le fichier BALSAC.

Les autres erreurs identifiées, bien que représentatives de la qualité générale des lignées maternelles et paternelles reconstituées ne sont cependant pas causées par des défauts de construction du fichier BALSAC lui-même. En effet, une part des erreurs est due à l'utilisation d'autres sources généalogiques, par exemple, une portion des lignées généalogiques acadiennes comprenait des mariages hors Québec qu'il a fallu rechercher dans d'autres sources. Cinq (5) erreurs de sur-jumelage (taux d'erreur de 0,11%) dans les lignées maternelles, une (1) dans les lignées paternelles (taux d'erreur de 0,09%) ainsi que trois (3) cas de sous-jumelages sont attribuables à des mariages hors Québec donc hors BALSAC.

Il est reconnu que les actes de mariage non catholiques sont en général moins riches en information que les actes de mariage catholiques. Très souvent, les parents des époux ne sont pas mentionnés dans l'acte ce qui rend quasi impossible les opérations de reconstitution des familles. Ainsi, certains liens généalogiques comportant des patronymes anglophones n'ont pu être reconstitués qu'en consultant d'autres sources que BALSAC, notamment les recensements nominatifs, les répertoires ou les sites généalogiques disponibles sur Internet. Deux (2) erreurs de sur-jumelage (taux d'erreur de 0,05%) sont attribuables à des lignées généalogiques anglophones. Dans bien des cas, les liens généalogiques sont impossibles à établir et ils sont comptabilisés comme des cas de sous-jumelage (N=6).

Les actes de baptême, de mariage et de sépulture eux-mêmes peuvent contenir des informations erronées qui viennent entacher les reconstitutions généalogiques. En effet, les naissances illégitimes, les cas d'adoption non mentionnés, les erreurs de transcription du célébrant, ... sont des causes d'erreur qui sont soit comptabilisées parmi les 13 erreurs de sur-jumelage introuvables (taux d'erreur de 0,3%) dans les lignées maternelles et les 8 erreurs introuvables (taux d'erreur de 0,73%) dans les lignées paternelles. Certaines erreurs contenues dans les actes peuvent occasionner des sous-jumelages qu'il n'est pas possible de repérer et de corriger avec les outils généalogiques usuels.

Finalement, bien qu'elle ne puisse pas être mesurée ici, la possibilité d'erreurs dans la production des données génétiques ne peut être exclue.