A la croisée de l'anthropologie et de la biologie évolutive : diversité génétique et comportements migratoires en Asie intérieure

Nina Marchi

Muséum National d'Histoire Naturelle
École Doctorale « Sciences de la Nature et de l'Homme » – ED227

Année 2017

N° attribué par la bibliothèque

THÈSE

Pour obtenir le grade de

Docteur du Muséum National d'Histoire Naturelle

Spécialité : Anthropologie Génétique

Présentée et soutenue publiquement par

Nina Marchi

Le 2 novembre 2017

(sous réserve d'autorisation des rapporteurs)

A la croisée de l'anthropologie et de la biologie évolutive : diversité génétique et comportements migratoires en Asie intérieure

---

Sous la direction de Madame Évelyne Heyer, Professeure – MNHN,
et la co-direction de Madame Laure Ségurel, Chargée de Recherche – CNRS

Jury :

Mme. Emmanuelle Génin	Directrice de Recherche – CNRS	Rapportrice
M. Lluis Quintana-Murci	Directeur de Recherche – CNRS	Rapporteur
M. Laurent Excoffier	Professeur – Université de Berne	Examinateur
M. Stéphane Mazières	Chargé de Recherche – CNRS	Examinateur

Remerciements

Je voudrais remercier toutes les personnes ayant participé, de plus ou moins près à ces travaux de thèse et à la production de ce manuscrit, et je m'excuse d'avance pour les oublis et la longueur de cette section.

En premier lieu, je voudrais remercier du fond du cœur celles sans qui rien de tout cela n'aurait été possible : Évelyne Heyer et Laure Ségurel. Directrices, accompagnantes et mentors, elles m'ont guidée à travers les "tortuosités" de cette thèse, tout en me laissant mûrir. Un grand merci à Évelyne pour m'avoir éclairée au fil de ces années de ses lumières scientifiques et de sa chaleur humaine, et de m'avoir emmenée dans ses valises en Ouzbékistan ! Mes sincères remerciements à Laure pour sa patience et son dévouement dans chaque étape et pour m'avoir poussée à m'améliorer, de sa main de fer dans un gant de velours.

Dans un second temps, je tiens à remercier tous mes collègues de l'équipe d'Anthropologie Évolutive, pour l'ambiance de travail et la convivialité qu'ils génèrent. En particulier, un grand merci à Frédéric pour ses conseils en sciences et sur la vie en général, à Paul pour ses éclairages "minute" précieux en génétique des populations, sémantique ou ichtyologie, à Romain pour ses, "euh comment", coups de main que je ne compte plus, qui tombent toujours à pic et qui m'ont évitée d'avoir une vie m*******, à Raphaëlle pour ses nombreuses suggestions de lecture et discussions passionnantes et pour m'avoir donné envie de rejoindre cette équipe incroyable, à Bruno pour son soutien empathique et sympathique en période de rédaction, à Samuel pour son infinie gentillesse, à Céline en partie pour son dernier fait d'armes héroïque : le détricotage de l'histoire de l'Asie intérieure, à Aline pour nous avoir certifié que certaines positions de yoga étaient anatomiquement impossibles provoquant le rire si chaleureux et communicatif de Marie-France, à Priscille pour être la bibliothécaire la plus sympathique que je connaisse, à Franz pour m'avoir appelée par mon prénom (et ce plusieurs fois !), à Philippe M. pour sa sagesse digne d'un aksakal, à Pierre D. et Gilles pour l'intérêt qu'ils ont porté à mon sujet et leurs suggestions bibliographiques, à Claire et Laure G. pour leurs conseils sur l'avenir professionnel et l'exemple qu'elles montrent, sans oublier les "petits" nouveaux Marie-Claude, Jean-Marc, Philippe C. et Pascal.
Je voudrais aussi remercier Myriam pour m'avoir patiemment et pédagogiquement accompagnée dans mes premiers pas à la paillasse, ainsi que Sophie et Françoise pour avoir pris la suite et m'avoir apporté une aide précieuse. Merci à Flora P. pour son aide avec QGIS et les rasters.
Mes salutations amicales vont au petit peuple des éphémères qui habite les couloirs du Musée de l'Homme ou ont habité ceux de la rue Buffon. Merci à Bérénice pour avoir égayé mes journées entre autres avec sa comédie musicale, à Goki pour sa zen-attitude communicative (dans une certaine mesure) et pour m'avoir rappelé l'heure des repas, à Valentin pour ses moments de science infuse au centre de ressources mais pas uniquement, à Christophe pour m'avoir appris comment prononcer le nom "Yuezhi", à notre ethno-musicologue masquée et douce empoisonneuse Camille, à l'homme-rhinocéros Guilhem qui a plus d'une corne à son arc, à Lou pour m'avoir expliqué comment boire dans un crâne, à Mahkameh pour mon premier porte-stylo, à Pierre S. pour son stoïcisme au cœur du bureau 203, à Farrokh pour les kuruts (surtout les ronds !), à Aurore pour avoir gentiment relu le History of Civilizations of Central Asia Vol I, à Clément, Anouck et la bande des ethno-éco pour les moments d'échange. J'ai aussi une pensée émue pour Élise et Julie dont la présence bienveillante a éclairé mes pauses et mes débuts de thèse (et aussi ma fin depuis le Danemark), me montrant la marche à suivre. Je n'oublie pas non plus les "anciens" Friso, Carla, Agnès, Jérémy, Kisito, Flora J., Victor, Marie, Aude et Noémie pour les bons moments passés.
Merci à Taouès pour m'avoir facilité, voire sauvé, la vie (sans exagération), et à Florence et Sylvie pour m'avoir permis de partir à la découverte, scientifique évidemment, de l'Angleterre, de la Croatie, de

l'Australie et de l'Ouzbékistan. Je remercie aussi Serge pour m'avoir accueillie au sein de l'UMR, ainsi que mes collègues de préhistoire, de primatologie, des collections et d'ethnologie pour les échanges interdisciplinaires ayant aiguisé ma curiosité : Antoine, Aurélien, Florent, Shelly, Cécile, Audrey, Liliana, Véro pour ne citer qu'eux. Sans oublier les "collègues" du yoga.
Merci aussi à Émilie Detouillon, Émeline Parent, Audrey Bonnemort, Camille Noize, Charlène Selva pour m'avoir accompagnée dans mes premiers pas en médiation scientifique qui a été un à-côté passionnant de cette thèse.

Des pans de cette thèse n'ont pu émergé qu'avec l'aide de collaborateurs que j'aimerais saluer : Christine Harmant, Josie Lambourdière, Laure Bagaït. Merci à Mark Jobling et à son équipe de m'avoir accueillie à Leicester. Merci à Tatyana et Nargisa de m'avoir fait découvrir l'Ouzbékistan et de m'avoir donné envie d'y retourner.

Merci également à mon comité de thèse de m'avoir gardée dans les rails et guidée : Anne-Louise Leutenegger, Ophélie Ronce, Raphaëlle Chaix et Sylvie Lebomin.
Merci à mes nombreux relecteurs pour l'énergie qu'ils ont consacrée à l'amélioration de ce manuscrit : je vous dois une reconnaissance éternelle.
Je voudrais aussi remercier les membres du jury d'avoir accepté de lire ce manuscrit, et plus particulièrement Emmanuelle Génin et Lluis Quintana-Murci d'avoir accepté d'en faire le rapport.

Parce que la thèse ne se fait pas uniquement derrière un ordinateur, je voudrais remercier mon entourage, famille et amis, pour m'avoir apporté un soutien sans faille et sans commune mesure.

Merci donc aux chimistes pour les rendez-vous hebdomadaires dans la tanière de l'ours, les voyages scolaires et les discussions à refaire le monde. Merci aux Marseillais pour m'avoir accueillie à bras ouverts à chaque occasion et pour avoir cédé à mes caprices cet été. Merci aux Lyonnais-es, en particulier à Félix pour m'avoir montré et remontré le Tao et à HD, mon ami du phare ouest. Une pensée pour mon amie de toujours, Mathilde, qui me doit un repas au Trocadéro ! À ma belette baroudeuse, pour son intérêt pour la linguistique et ses visites inopinées générant des passages au Comptoir moderne ou des soirées burger, bref pour ses nombreuses attentions si chères à mes yeux : aligatô !

Tel le vent bleu, ma famille m'a soutenue (et supportée) au cours de ces années parisiennes, comme je sais qu'elle le fera toujours, et pour cela je ne la remercierai jamais assez. Merci donc pour les visites, les lettres, les SMS d'encouragement, les parties de pétanque gagnées, les petits plats, les panisses, les questions sur les dinosaures, toutes ces attentions qui ont su toucher mon estomac et mon cœur quand mon cerveau était (pré-)occupé. *Une citation des Fatals Picards s'est cachée dans cette phrase, sauras-tu la trouver ?* En particulier, merci à mes parents et mon frère d'avoir fait face à mes doutes aux cours de ces derniers mois avec stoïcisme et philosophie : "ça va aller parce que ça doit aller". Merci aussi à ma belle-famille pour le gite et parfois le couvert.
"Éventuellement", mes remerciements vont à Rémi, mon partenaire essentiel, inclassable et indétrônable. Preux chevalier de l'ombre, il a été présent tout au long de cette thèse, m'a montré la voie, m'a apporté le soutien scientifique, logistique et sentimental dont j'avais besoin et a fait preuve d'une patience sans limite (que je vais continuer d'explorer). Pour tout cela et bien plus encore, merci mon amour.

Table des matières

Abréviations

ACP Analyse en Composantes Principales ; PCA en anglais

ADN Acide DésoxyriboNucléique

AEC Avant l'Ère Commune (en remplacement de "av. J.-C.") ; BCE en anglais

AMOVA *Analysis of Molecular Variance* en anglais, pour "Analyse de VAriance MOléculaire"

ASD *Allele-Sharing Dissimilarities* en anglais

CCAS Comportement Culturel Asymétrique entre les Sexes

cM centimorgan

EC Ère Commune (en remplacement de "apr. J.-C.) ; CE en anglais

EHG & WHG *Eastern & Western European Hunter-Gatherers* pour parler de populations européennes de chasseurs-cueilleurs du Méso/Néolithique

FBD *Father's Brother's Daughter* en anglais pour parler d'un type de cousine paternelle

HBD *Homozygous-By-Descent* en anglais, ou "Homozygote par Ascendance" en français

HVS1,2 Séquence hypervariable 1 ou 2 de l'ADN mitochondrial

IBD *Identical-by-descent* en anglais, "identiques par descendance ou ascendance" en français

IBS *Identical-by-state* en anglais, simplement "identiques" en français

kb, Mb Kilo-base ou Méga-base

MBD *Mother's Brother's Daughter* en anglais pour parler d'un type de cousine maternelle

MDS *MultiDimensional Scaling* en anglais, traduit par "positionnement multidimensionnel" en français

NRY *Non-Recombining Region of the Y Chromosome*

ROH *Run Of Homozygosity* en anglais, pouvant être traduit par "segment homozygote" en français

SNP *Single Nucleotide Polymorphism* en anglais, pouvant être traduit par "polymorphisme d'une paire de base" en français

STR *Short Tandem Repeats* en anglais, pouvant être traduit par "microsatellite" en français

1C, 2C, 2x1C, AV, OUT Codes respectivement utilisés pour parler des unions entre cousins germains, issus de cousins germains, doubles cousins, avunculaires ou moins apparentés appelés *Outbred*

Introduction générale

Nous sommes actuellement plus de 7,5 milliards d'*Homo sapiens* sur Terre [1]. En tant que membres de la même espèce, nous partageons de nombreux traits communs mais nous présentons aussi quelques différences visibles à l'œil nu, comme une large palette de couleurs de peau ou de cheveux. Au-delà de cette variation phénotypique évidente, notre espèce recèle des différences plus discrètes dont certaines nichées au sein de notre ADN. Ainsi, les génomes de deux humains choisis au hasard sur la planète diffèrent, en moyenne, par 4 à 5 millions de paires de bases, soit moins de 0,1% de différences génétiques (The 1000 Genomes Project Consortium 2015). Cette faible diversité génétique est la conséquence de l'origine récente de notre espèce. En effet, les données archéologiques situent les plus anciennes traces d'*Homo sapiens* en Afrique, il y a seulement 200 à 300 000 ans (Trinkaus 2005 ; Hublin *et al.* 2017), et ce n'est qu'il y a 50 à 100 000 ans que quelques représentants de notre espèce quittèrent le continent africain pour peupler le reste de la planète (Figure 1).
En outre, la grande majorité de la variation génétique humaine est observée entre des individus d'une même population (environ 95%) et seulement 5% entre des individus originaires de populations différentes (Rosenberg *et al.* 2002). L'étude et la compréhension de la dynamique évolutive de cette diversité génétique sont au cœur de la recherche en génétique des populations (Serre 2006) et de cette thèse centrée sur l'Asie intérieure.

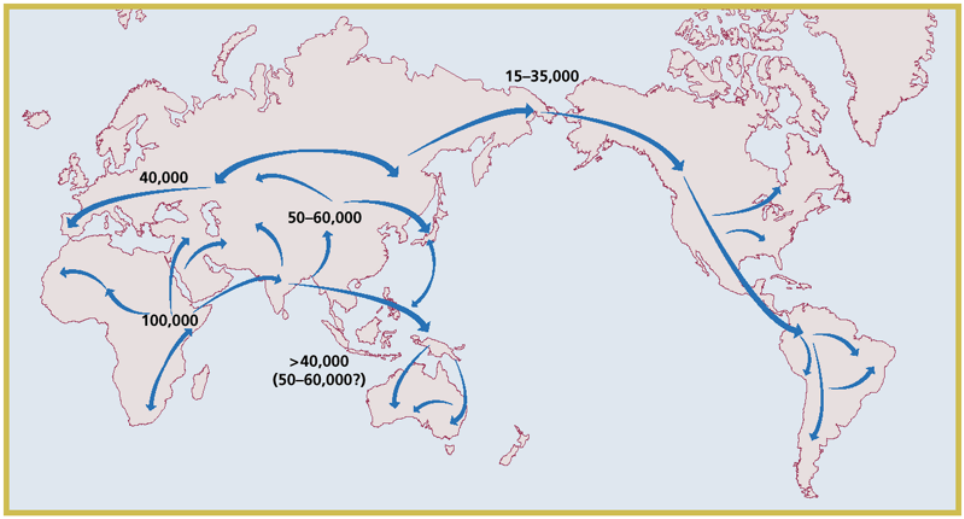


Figure 1 – Migrations d'*Homo sapiens* hors d'Afrique. *Figure 3 tirée de Cavalli-Sforza et Feldman (2003).*

A l'échelle des populations, la diversité génétique dépend directement des mutations survenues chez les individus et de la recombinaison à l'origine de la variation, mais aussi de trois forces évolutives qui changent les fréquences de ces mutations au sein des populations, en lien avec leur histoire démographique (Wright 1931) :

1. **La sélection :** dans un environnement donné, certains individus présentent un caractère avantageux qui leur assure une descendance plus nombreuse par rapport à celle d'individus ne possédant pas ce caractère. Si les conditions environnementales sont stables, la fréquence de ce caractère augmente dans la population, génération après génération, jusqu'à sa fixation, à l'image de la forme du bec des célèbres pinsons de Darwin (Darwin 1889 ; Grant et Grant 2011). Au contraire, les caractères délétères sont perdus, d'autant plus vite que leurs effets sont néfastes.
2. **La dérive :** les fréquences alléliques d'une population varient aléatoirement, par dérive, du fait

1. 7 515 285 sur la base d'une estimation de l'Institut National d'Études Démographiques pour l'année 2017.

de la stochasticité des combinaisons gamétiques survenues lors des événements reproducteurs. La dérive conduit à la fixation de certains allèles et à la perte des autres, et donc à une réduction de la diversité génétique (Kimura et Ohta 1969). Il est impossible de prédire quel allèle sera fixé, la dérive étant un processus aléatoire. L'intensité de la dérive est inversement proportionnelle au nombre d'individus de la population participant au processus reproductif. Cet effectif est la taille efficace de la population (Wright 1931 ; Charlesworth 2009), notée Ne et correspond à l'effectif d'une population idéale de Wright-Fisher dont le degré de dérive génétique est équivalent à celui mesuré pour la population étudiée.

3. **La migration :** des échanges génétiques sont réalisés entre des populations au moyen de migrations d'individus (Fix 1999). Ces échanges favorisent l'homogénéisation génétique des groupes. La façon dont les individus migrent, en groupe ou seuls, la distance qu'ils parcourent, la fréquence des migrations, ou encore le sexe des migrants influencent la diversité génétique des populations.

L'étude des forces évolutives peut se faire par des approches

- théoriques et mathématiques par modélisation de leurs effets au cours du temps,
- expérimentales comme la célèbre expérience de Buri qui a étudié la fixation par dérive d'une mutation chez 107 populations de drosophiles pendant 20 générations, et a montré le caractère aléatoire de la dérive (Buri 1956),
- empiriques en s'appuyant sur l'étude de populations naturelles.

Cette dernière approche a motivé, chez l'Homme, la constitution de jeux de données tels que le *Human Genetic Diversity Panel* (HGDP-CEPH), pour décrire la diversité génétique humaine à l'échelle mondiale ou régionale. L'étude de ce jeu de données a montré que la diversité génétique des populations humaines diminue en fonction de la distance à l'Afrique (Figure 2), ce qui constitue un argument supplémentaire en faveur d'une origine africaine de notre espèce. Cette observation est compatible avec un modèle de peuplement par des effets fondateurs en série, dans lequel chaque vague de migration emporte vers la nouvelle colonie une partie seulement de la diversité génétique de la population source (Li *et al.* 2008). Migration après migration, la diversité génétique est ainsi supposée décroître, si d'autres forces ne compensent pas l'effet des migrations en série.

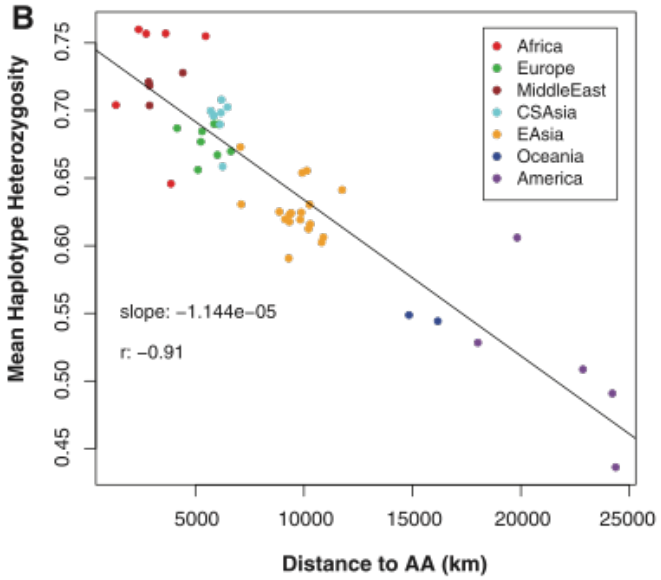


Figure 2 – Hétérozygotie des populations du jeu de données HGDP-CEPH en fonction de la distance à Addis-Abeba. *Figure 3 tirée de Li* et al. *(2008).*

Outre des processus biologiques conditionnant l'histoire évolutive de l'Homme, la diversité génétique humaine est modifiée par certaines de nos pratiques culturelles qui font varier l'intensité des forces évolutives présentées (Cavalli-Sforza *et al.* 1994). Par exemple, les pratiques agro-alimentaires, sanitaires ou vestimentaires nous permettent de nous soustraire à certaines pressions de sélection, tout en en créant des nouvelles (Hünemeier *et al.* 2012 ; Lachance et Tishkoff 2013). Des règles de mariage au sein d'un groupe, comme le système de caste en Inde, peuvent le diviser en sous-groupes de plus petits effectifs, ce qui augmente l'intensité de la dérive (Bittles 2005). Également, des barrières culturelles, par exemple linguistiques (Cavalli-Sforza *et al.* 1994), peuvent limiter les flux géniques entre différentes populations (Fix 2004). En sus, les normes culturelles suivies par la population peuvent autoriser les migrations d'individus d'un sexe mais pas de l'autre, comme la règle de résidence patri ou matrilocale (Heyer *et al.* 2012).

Pour comprendre l'histoire évolutive de notre espèce, il est donc crucial d'intégrer ces comportements culturels aux études de génétique des populations. Cette thèse a pour objectif de documenter l'influence de comportements culturels, pouvant causer des migrations, sur la diversité génétique dans un cadre particulier : celui des populations d'Asie intérieure.

Objet de l'étude : les populations humaines d'Asie intérieure

L'Asie intérieure est peuplée par deux groupes culturels, les Turco-Mongols et les Indo-Iraniens qui, bien que cohabitant sur un même territoire, se distinguent notamment par leur langue, leur organisation sociale et leur mode de subsistance. Ce terrain d'étude constitue donc un cadre exceptionnel pour explorer l'effet de comportements culturels sur la diversité génétique et l'évolution génétique des populations humaines, mais aussi pour tenter d'inférer et de dater la mise en place de ces différences culturelles. Nous dressons ici un portrait géographique, historique et ethnologique de cette région.

L'Asie intérieure - Repères géographiques -

Une façon de définir l'Asie intérieure est d'utiliser comme frontières la taïga russe au nord, la mer Caspienne à l'ouest, les déserts iraniens et les montagnes afghanes au sud, les contreforts des montagnes du Pamir et du Tian Shan au sud-est, et le lac Baïkal au nord-est (Figure 3). L'Asie intérieure ainsi définie s'étend sur environ 3 000 km d'est en ouest, et sur 2 000 km du nord au sud.

Cette région, parfois appelée Asie centrale au sens large, inclut deux aires géographiques souvent étudiées séparément mais partageant une histoire et une culture communes :

- l'Asie centrale (au sens strict), aussi appelée Turkestan russe ou occidental, qui correspond aux républiques du Kazakhstan, Kirghizstan, Tadjikistan, Turkménistan et Ouzbékistan ;
- l'Asie du nord qui inclut la Mongolie occidentale, une partie de la Russie - la Sibérie méridionale dont les républiques d'Altaï et Tuva proches de la frontière mongole - et parfois la région la plus occidentale de Chine - la république autonome ouïghoure du Xinjiang - .

Le paysage de cette région se compose principalement de steppes et de déserts ponctués d'oasis - en particulier le désert du Kara-Koum s'étendant sur 80% du territoire du Turkménistan et celui du Kizil-Koum en plein cœur de l'Asie centrale renfermant l'oasis de Boukhara.

Figure 3 – Géographie de l'Asie intérieure. Les noms des pays sont inscrits en noir, ceux des déserts en orange, des lacs en bleu et des chaînes de montagnes en marron. Les républiques russes d'Altaï et Tuva, et celle chinoise du Xinjiang sont indiquées en italique.

Le paysage steppique aurait favorisé le développement du nomadisme dans la région, qui consiste en une transhumance des troupeaux et des populations humaines entre des zones de pâturage saisonnières selon un rythme cyclique (Stépanoff *et al.* 2013). Le nomadisme, très courant en Asie intérieure ainsi que la célèbre Route de la Soie qui traversait la région (Figure 4), en auraient fait un haut lieu d'échanges culturels, démiques [2] et commerciaux. L'Asie intérieure a aussi connu de nombreux épisodes de migrations, de métissages génétiques et culturels, et a été sous l'influence de cultures variées au cours de son histoire.

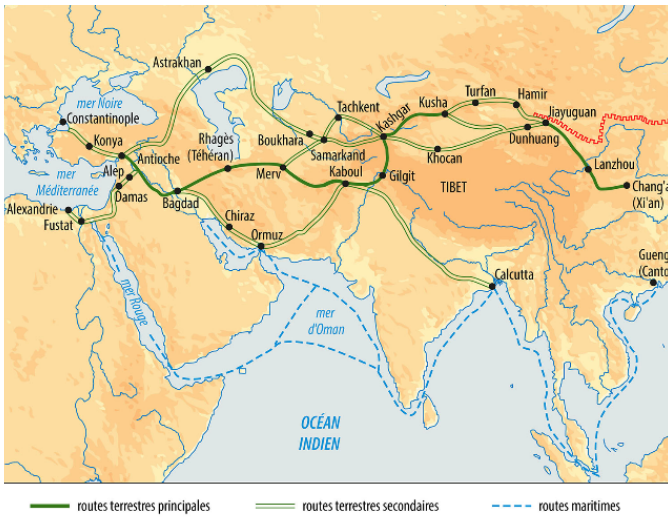


Figure 4 – Itinéraires principaux de la Route de la Soie. Encyclopedia Universalis - Route de la Soie.

2. Terme employé initialement par Cavalli-Sforza *et al.* (1993) pour parler d'une diffusion liée aux migrations des individus et non uniquement à des échanges d'idées ou de biens.

Histoire de la région

L'occupation de l'Asie intérieure par l'Homme est ancienne, remontant au Paléolithique (Figure 5), avec des restes fossiles témoignant de la présence de :

- l'Homme de Denisova. Les seuls restes retrouvés à ce jour pour cet Homme archaïque proviennent d'Asie intérieure, de la grotte éponyme dans l'Altaï et datent de 40 000 ans (Reich *et al.* 2010) ;
- l'Homme de Néandertal, dont les restes les plus orientaux ont été retrouvés en Asie intérieure, à Teshik-Tash en Ouzbékistan, et en Altaï dans la grotte de Denisova et dans la grotte voisine d'Okladnikov, et datent de 30 à 50 000 ans (Krause *et al.* 2007 ; Prüfer *et al.* 2013) ;
- l'Homme anatomiquement moderne dont les plus anciens restes retrouvés en Asie intérieure sont localisés à Pokrovka dans l'Altaï et datent d'il y a 27 000 ans (Akimova *et al.* 2010).
  Le climat de la région a permis la conservation de l'ADN dans les fossiles, et ainsi l'obtention de génomes humains préhistoriques dont celui de Mal'ta en Sibérie (24 000 ans) (Raghavan *et al.* 2014). À ce jour, le plus ancien génome humain, dit d'Ust'-Ishim (45 000 ans) a été retrouvé dans une région sibérienne proche de l'Asie intérieure (Fu *et al.* 2014).

Figure 5 – Restes fossiles d'Hommes paléolithiques pour lesquels de l'ADN a été conservé. Les Hommes anatomiquement modernes sont en bleu, et les Hommes de Denisova et de Néandertal en rouge. L'homme de Ust'-Ishim aurait vécu il y a 45 000 ans (Fu *et al.* 2014), l'enfant de Mal'ta il y a 24 000 ans (Raghavan *et al.* 2014), les Néandertaliens de l'Altaï il y a 30 à 50 000 ans (Krause *et al.* 2007 ; Prüfer *et al.* 2013) ; aucune date n'a été obtenue pour les restes de l'enfant néandertalien de Teshik-Tash.

Des traces plus fournies témoignent d'une occupation continue de l'Asie intérieure à compter du Néolithique (V^{e} millénaire **A**vant l'**È**re **C**ommune), par diverses cultures.

Du Néolithique à l'âge du Bronze

Au Néolithique, l'Asie intérieure était divisée en trois aires d'influence, correspondant à trois processus de néolithisation indépendants (c'est-à-dire de domestication des animaux et végétaux ayant conduit à une économie de production), et n'étant entrées en contact qu'à l'âge du Bronze :

- l'Asie centrale des oasis, sur les territoires du Turkménistan et de l'Ouzbékistan actuels. Dès le début du Néolithique, l'agriculture et l'élevage sont pratiqués par des populations sédentaires, qui ont développé des cultures proto-urbaines ou urbaines (comme celle de Djeitun, VII^{e} au VI^{e} millénaire AEC) (Francfort et Grenet 2017 ; Leroi-Gourhan 1994). Cela a débouché à l'âge du Bronze sur un urbanisme particulièrement développé associé à la culture de l'Oxus ou "Complexe

Archéologique Bactro-Margien" (BMAC, 2300 - 1700 ans AEC) (Dani et Masson 1994). Cette culture était en contact avec des cultures d'Asie du sud, d'Iran, de Mésopotamie ou de la vallée de l'Indus.

- l'Asie centrale des steppes, sur le territoire du Kazakhstan. L'apparition de l'agriculture et de la sédentarité est tardive dans cette région : au début du Néolithique, ses habitants vivaient plutôt de chasse et de pêche, et résidaient dans des campements saisonniers ; puis, les membres de la culture Botaï (IV^{e} millénaire AEC) domestiquèrent des chevaux (Outram *et al.* 2009). À l'âge du Bronze, la présence de chevaux est également attestée, sans être associée à du nomadisme, dans des sites rattachés à la culture d'Afanasievo (3300 - 2400 ans AEC) à l'est, et à celle de Sintashta (2100 - 1800 ans AEC) qui aurait introduit les chars à roues dans les steppes kazakhes (Mallory et Adams 1997 ; Anthony 2010) (Figure 6). Cette région aurait été en contact avec l'Europe de l'est, notamment avec les steppes ukrainiennes dont les cultures, telles que celle de Srubna présentent des ressemblances avec celles d'Asie centrale.
- la Sibérie et la Mongolie dont la période proto-historique est encore assez mal connue. On peut cependant évoquer des cultures néolithiques axées sur une forte activité de pêche, comme la culture Kitoî de Sibérie (Weber *et al.* 2011), ou celle de Tamsagbulag en Mongolie, qui constitue un pôle de néolithisation indépendant (Séfériadès 2003). Au II^{e} millénaire AEC, ces cultures sont remplacées par la culture de Glaskovo autour du lac Baïkal, et à l'ouest par la culture d'Afanasievo, venue des steppes d'Asie centrale (Marchand *et al.* 2017).

À l'âge du Bronze tardif (1800 - 1200 ans AEC), la culture d'Andronovo s'étendait sur l'ensemble de l'Asie intérieure, remplaçant les entités locales et créant un fonds culturel commun, même si les archéologues distinguent plusieurs courants. On lui attribue des sépultures en tumulus, appelées kourganes, qui la rapprochent de la "culture des tombes à kourganes" d'Europe de l'est, supposée à l'origine des langues indo-européennes (Leroi-Gourhan 1994 ; Mallory et Adams 1997). D'autres cultures, dites du Karassouk et d'Okunevo, la remplacèrent tardivement en Sibérie (Keyser *et al.* 2009).

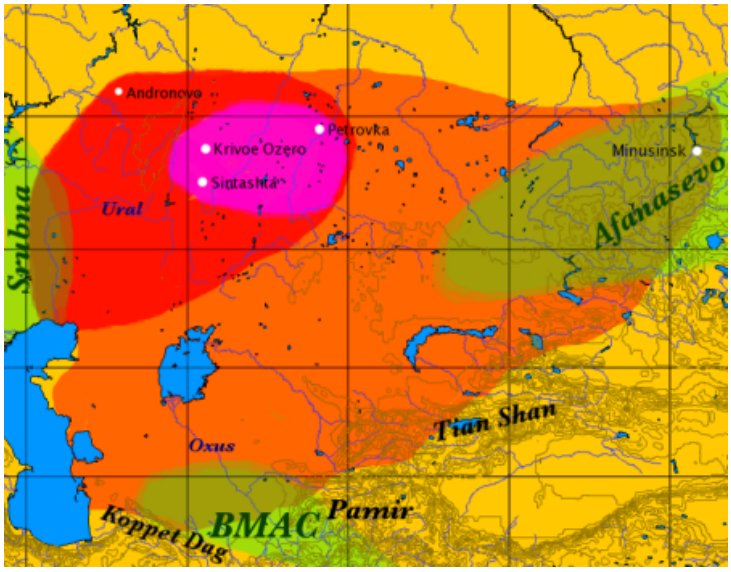


Figure 6 – Les cultures de l'âge du Bronze en Asie intérieure. D'ouest en est : la culture de Sinthasta en rouge, d'Andronovo en orange, d'Afanasievo et de l'Oxus (ou BMAC) en vert. Les premières traces de chars à roues ont été trouvées dans la zone rose. Wikipédia - Culture d'Afanasievo.

À l'âge du Fer

Bien que les cultures de l'âge du Bronze disposaient de chars, leurs membres vivaient de façon plutôt sédentaire dans des villages et d'une économie essentiellement agro-pastorale. Au VIII^{e} siècle AEC, les occupants de l'Asie intérieure abandonnèrent la sédentarité au profit d'un nomadisme pastoral. Ces tribus nomades retrouvées dans toute l'Asie intérieure, et même jusqu'en Ukraine, sont appelées Scythes, Saces ou encore Sakas (Dani et Masson 1994). Ils étaient décrits par des auteurs chinois et Hérodote comme physiquement européens et parlant des langues indo-européennes.
En particulier, la branche scythe la plus méridionale d'Asie intérieure, les Sogdiens, était présente dans une région appelée la Sogdiane (correspondant à l'Ouzbékistan et au Tadjikistan actuels). Ils fondèrent notamment la cité de Samarcande. La branche scythe qui était présente hors d'Asie intérieure, dans la région entre la mer noire et la mer Caspienne est nommée "sarmate" (Davis-Kimball *et al.* 1995).
La culture scythe est reconnaissable sur l'ensemble de son aire de répartition par un trait culturel original : la représentation selon une stylisation et des conventions spécifiques d'animaux sur les objets mobiliers et sous forme de tatouages (Unterländer *et al.* 2017) (Figure 7). On trouve une variation régionale discrète à travers les animaux représentés de manière préférentielle et la présence essentiellement en Mongolie et autour du lac Baïkal de mégalithes gravés, appelés "pierres à cerf" (Magail 2005). La discontinuité culturelle entre la culture scythe et les cultures antérieures, en particulier à travers la pratique du nomadisme, et la grande homogénéité culturelle scythe soulèvent des interrogations quant à l'origine de ces tribus de l'âge du Fer.

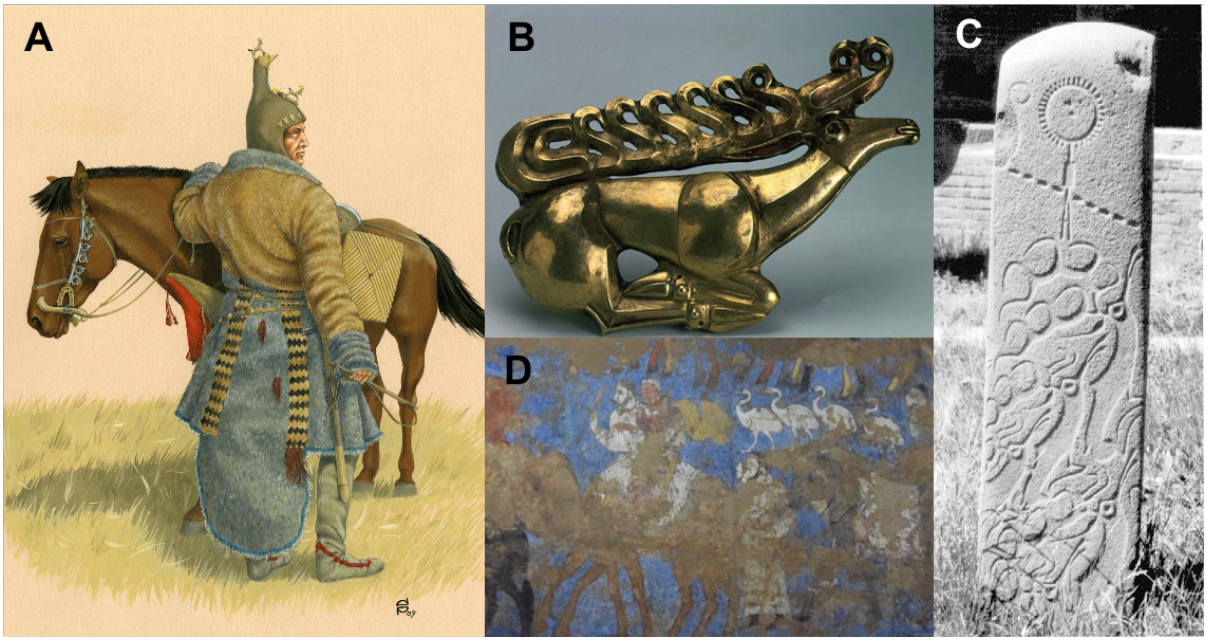


Figure 7 – Artefacts de la culture scythe : cavalier (A), art animalier (B), pierre à cerfs (C), peinture des ambassadeurs sur le site d'Afrasiab à Samarcande (D). Les images ABC sont tirées de Unterländer *et al.* (2017), Encyclopedia Universalis - Scythes, Magail (2005) et la D est une photographie personnelle réalisée à Samarcande.

L'Antiquité et le Moyen-Âge, le temps des migrations

À compter du premier millénaire ACE, de l'âge du Fer jusqu'au Moyen-Âge, l'Asie intérieure fut prise en tenaille par de nombreuses vagues de migration venues du sud-ouest et de l'est (Figure 8) (Harmatta *et al.* 1994 ; Litvinskii *et al.* 1996) :

- en 550 AEC, les Perses établirent leur premier empire, dit des Achéménides, du Proche Orient jusqu'en Asie centrale. Un siècle et demi plus tard, les Achéménides furent renversés par les troupes d'Alexandre le Grand lors de la bataille de Gaugamèles ;
- consécutivement, au III^{e} siècle AEC, les Grecs implantèrent un royaume gréco-bactrien dans les régions de la Bactriane et de la Sodgiane, sur les territoires actuels ouzbek, tadjik et afghan. Ce royaume était voisin de l'empire parthe, fondé autour de 247 AEC par des Scythes sur le territoire iranien, à la frontière du Turkménistan ;
- au I^{e} siècle AEC, les Yuezhis, des Indo-Européens repoussés hors de Chine par les Xiongnu venus de Mongolie, gagnèrent l'Asie centrale. Ils y rencontrèrent les Scythes et défirent l'empire gréco-bactrien, fondant en remplacement l'empire kouchan qui s'étendait jusqu'en Inde ;
- entre le III et IV^{e} siècles EC, des peuples de l'Altaï se répandirent en Mongolie. De là, à l'image des Xiongnu et des Huns, ils se dirigèrent vers l'Asie centrale où ils se métissèrent et/ou remplacèrent les populations indo-iraniennes locales, marquant la fin de la suprématie scythe. Pendant ce temps-là, plus au sud, un nouvel empire iranien, celui des Sassanides, supplanta l'empire des Parthes ;
- au VIII^{e} siècle EC, les Arabes vainquirent l'empire iranien sassanide, puis prirent le contrôle du sud de l'Asie centrale (Ouzbékistan et Kirghizstan actuels) après leur victoire sur la dynastie chinoise Tang lors de la bataille de Talas, dans la vallée du Ferghana, au Kirghizstan. Cette dynastie s'était imposée brièvement en Asie centrale autour de 740, avant d'être repoussée vers l'est en 751. Les Arabes abbassides créèrent différents califats et répandirent l'islam en Asie intérieure, conduisant à l'affaiblissement du zoroastrisme et du bouddhisme ;
- aux IX^{e}-X^{e} siècles EC, une vague venue d'Iran mit en place l'empire des Samanides et introduisit en Asie centrale une langue ouest-iranienne, remplaçant les langues est-iraniennes précédemment parlées ;
- au XIII^{e} siècle EC, le mongol Genghis Khan créa le plus grand empire jamais connu, s'étendant de la Mongolie à la mer noire, englobant toute l'Asie intérieure. Cet empire fût par la suite divisé en royaumes, les khanats.

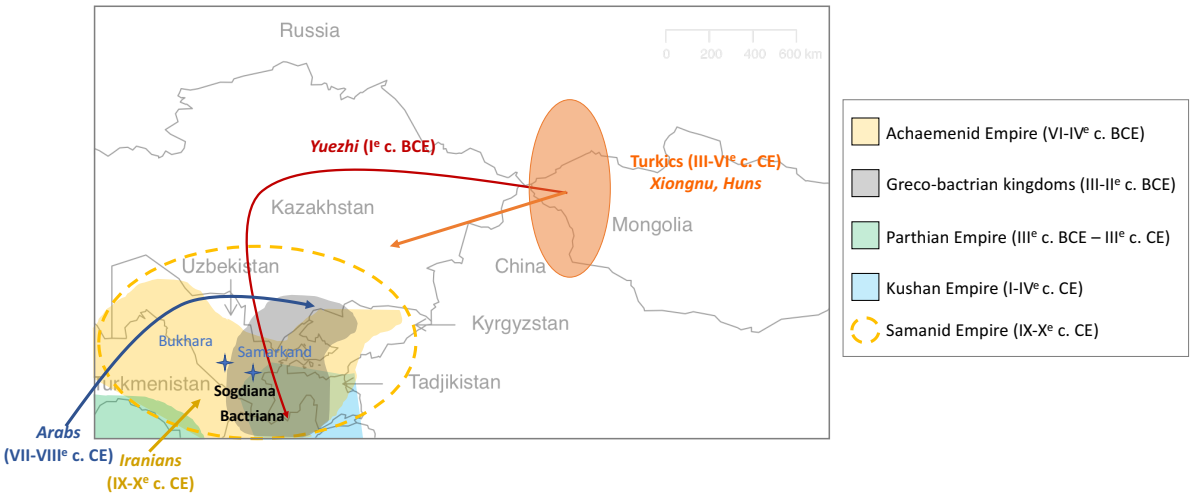


Figure 8 – Empires et migrations en Asie intérieure, de l'âge du Fer au Moyen-Âge. Les territoires des empires sont représentés au moment de leur apogée. Les régions de la Sogdiane et de la Bactriane sont indiquées en lettrage noir, et les villes de Samarcande et Boukhara en bleu.

Le temps des khanats

L'Asie centrale passa successivement sous la domination du khanat des Chagataï au XIV^{e} siècle, des Timourides au XV^{e} siècle, et des Chaybanides au XVI^{e} siècle, tous descendant de Genghis Khan. Du XVII^{e} siècle et jusqu'à la création de l'URSS, des khanats plus petits comme celui de Boukhara, fleurirent dans toute l'Asie intérieure et l'autorité politique fût morcelée (Adle *et al.* 2005). Ces différents khanats furent un temps sous l'autorité chinoise de la dynastie Qing au XVIII^{e} siècle, puis à partir du XIX^{e} siècle sous l'autorité russe, tout d'abord impériale, puis soviétique (Adle *et al.* 2003). La Mongolie fût disputée par la Chine, la Russie et le Japon. Les républiques d'Asie intérieure ne gagnèrent leur autonomie totale qu'au début des années 1990, à la fin de l'union soviétique.

Notons que les échanges entre l'Asie du nord et centrale se poursuivirent durant toute cette période, façonnant le visage de l'Asie intérieure contemporaine et donnant naissance aux groupes ethniques actuels.

Groupes ethniques

L'Asie intérieure est peuplée par une grande diversité de groupes ethniques. De façon générale, un "groupe ethnique" est une entité rassemblant des personnes qui considèrent partager une ascendance commune, une histoire commune et/ou une culture commune (Barth 1969 ; Kottak 2008). En Asie intérieure, les groupes ethniques étudiés incluent des groupes d'individus de même langue, se présentant sous un même nom, et sont donc formés sur une définition émique [3].

Plusieurs courants de pensée se sont interrogés sur cette idée d'ascendance commune, pour déterminer si elle est biologique ou mythologique (Streiff-Fénart et Poutignat 1995) :

- est-elle, comme postulé par les théories primordialistes, naturelle et intrinsèque à l'Homme ?
- est-ce l'extension des liens de parenté comme envisagé dans les théories socio-biologiques ?
- est-elle de nature sociale et définie
  - par une forte ressemblance culturelle entre ses membres selon les théories substantialistes,
  - par des interactions entre individus, même en l'absence de traits communs pour les théories des interactions non-substantielles,
  - par des motivations pragmatiques ou idéologiques pour les théories non-substantialistes instrumentalistes et marxistes,
  - comme un symbole permettant aux individus de se situer dans un ordre social plus large, dans les théories néo-culturalistes ?

Dans nos études, nous ne supposons pas *a priori* d'ascendance biologique commune aux membres du groupe ethnique, mais nous avons intégré cette thématique à nos questions de recherche.

Tadjik

De nos jours, le groupe ethnique tadjik est le plus représenté au Tadjikistan par 85% des 8 millions d'habitants [4], mais la plupart des membres de ce groupe résident hors du Tadjikistan : ils sont plus de 16 millions en Afghanistan, où l'ethnie tadjike est la seconde ethnie la plus nombreuse, et près d'un million en Ouzbékistan, essentiellement dans les régions de Samarcande et Boukhara. Il est le groupe ethnique majoritaire de langue indo-iranienne d'Asie intérieure et n'est présent qu'en Asie centrale.

3. qui correspond au point de vue des autochtones (Kottak 2008).
4. Les effectifs nationaux proviennent de *populationdata.net* et les proportions ethniques proviennent des recencements, les plus récents par pays, du DHS Program.

La présence de ce groupe ethnique est la plus ancienne attestée en Asie centrale, datant d'avant les migrations mongoles du XIII^{e} siècle (Minahan 2014). Son nom signifie "arabe" en langue turcique et faisait référence aux Indo-Européens d'Asie centrale convertis précocement à l'Islam au Moyen-Âge.

Au sein de ce groupe, on distingue la branche des Tadjiks parlant une langue tadjike et celle des Yagnobs, vivant dans les montagnes du Tadjikistan et parlant le yagnobi, une langue indo-iranienne se rapprochant de langues persanes médiévales (Soucek 2000 ; Heyer et Mennecier 2009).

Turkmène

78% des 5,6 millions d'habitants du Turkménistan fait partie de l'ethnie turkmène. Ce groupe, présent uniquement en Asie centrale, aurait été formé avant les grandes invasions mongoles du XIII^{e} siècle par la coalition de tribus d'Asie centrale et du nord (Minahan 2014 ; Grousset 1970).

Ouzbek

De nos jours, le groupe ethnique ouzbek représente 86% des 32 millions d'habitants de l'Ouzbékistan et est présent en Asie centrale exclusivement.

L'histoire des Ouzbeks en tant que groupe remonte à l'année 1429 quand différentes tribus furent confédérées par le khan Abu'I-Khayr de la dynastie des Chaybanides, descendante de Genghis Khan (DeWeese 2010). La confédération incluait notamment la tribu des Chagatai, d'autres descendants de Genghis Khan qui résidaient en Ouzbékistan à cette période (Soucek 2000) et qui se seraient largement mélangés avec des peuples iraniens autochtones (Heyer *et al.* 2015).
Ce groupe domina l'Ouzbékistan à compter de 1507, à la chute des Timourides, dynastie fondée par Timour/Tamerlan, l'un des grands conquérants originaires d'Asie intérieure (Adle *et al.* 2003). Au cours du XVI^{e} siècle EC, la majorité du groupe ouzbek abandonna le nomadisme pastoral pour un mode de vie sédentaire et l'agriculture (Soucek 2000). Certaines tribus auraient quitté la confédération et migré plus au nord pour participer à la formation du groupe kazakh.

Kazakh

Ce groupe représente 53% des 17,5 millions d'habitants du Kazakhstan et est présent en Asie centrale et du nord. Il se serait formé au sud de l'actuel Kazakhstan, à la fin du XV^{e} siècle, à la chute de l'empire timouride, sous la forme d'une confédération de groupes variés, dont des membres de l'ancien empire mongol comme les tribus ouzbèkes (Dulik *et al.* 2011 ; Adle *et al.* 2003). La première entité politique kazakhe date de 1470, sous la forme d'un khanat qui exerça son autorité sur le Kazakhstan jusqu'au XIX^{e} siècle. Cependant, l'unification du groupe n'eut vraiment lieu qu'au cours du XVI^{e} siècle.
Les Kazakhs sibériens modernes seraient les descendants de tribus, ou hordes, kazakhes originaires d'Asie centrale (Minahan 2014).

Karakalpak

Ce groupe est retrouvé exclusivement en Asie centrale, principalement autour de la mer d'Aral, en Ouzbékistan (dont il représente 2,5% de l'effectif national). Leur nom signifie "chapeau noir" et fait référence à leur coiffe traditionnelle en pelisse de mouton (Minahan 2014).
La première référence écrite aux Karakalpaks remonte à l'an 1598, faisant état de tribus nomades vivant dans la région de Sygnaq au sud de l'actuel Kazakhstan (Jacquesson 2002). Le groupe aurait été formé au XV-XVI^{e} siècle par la coalition de tribus, dont certaines kazakhes, ce qui expliquerait la proximité

linguistique de la langue karakalpak avec la langue kazakhe (Minahan 2014). À l'heure actuelle, le groupe karakalpak est divisé en deux branches : les Qongïrat et les On Tört Uruw, qui descendent hypothétiquement de tribus différentes rassemblées au sein de la coalition karakalpak (Jacquesson 2002).

Kirghize

70% des 6,1 millions d'habitants du Kirghizstan fait partie du groupe ethnique kirghize, que l'on retrouve en Asie centrale et du nord. Historiquement, des écrits attestent de la présence de ce groupe sous sa forme actuelle en Asie centrale à compter de 1503, mais il aurait été présent précédemment en Asie du nord, dès le premier siècle AEC, avant d'être incorporé dans l'empire mongol. Une branche kirghize aurait quitté le groupe ancestral sibérien et se serait installée dans le Tian Shan au XIII^{e} siècle EC, avant de migrer vers l'Asie centrale, au XV^{e} siècle EC (Adle *et al.* 2003).

Khakasse

Ce groupe ethnique de Sibérie occupe actuellement le territoire de l'ancien khanat kirghize et pourrait être apparenté au groupe ethnique éponyme (Adle *et al.* 2003). Alternativement, ce groupe pourrait avoir émergé durant l'époque soviétique par coalition de divers groupes ethniques présents en Khakassie.
Ce groupe pratique traditionnellement la chasse, la pêche et la cueillette comme moyen de subsistance [5].

Ensemble mongol

Le terme mongol fait référence à plusieurs groupes ethniques d'Asie du nord, dont le groupe majoritaire des Khalkhas, communément appelé mongol, qui serait apparu au XV^{e} siècle (Minahan 2014). Jusqu'au XX^{e} siècle, ce groupe était sous l'autorité de khans de la lignée de Genghis Khan, connus sous le nom de Borjigin.

Un autre groupe ethnique mongol est celui des Bouryates, localisé autour du lac Baikal (Minahan 2014). Ses deux branches principales sont définies par des critères géographiques : au nord et à l'ouest du lac *versus* au sud et à l'est. Actuellement, c'est le groupe autochtone de Sibérie le plus représenté comptabilisant 500 000 personnes [6].

Ensemble altaïen

Les populations du sud de la Sibérie sont divisées en deux grandes branches, les Altaïens du nord et ceux du sud, incluant elles-mêmes différents groupes ethniques (Minahan 2014).
Parmi les Altaïens du sud, on dénombre en particulier le groupe des Altaï-Kizhi (comptabilisant environ 68 000 personnes) et des Télengits (4 000 personnes).
Le groupe d'Altaïens du nord inclut notamment les Chors (14 000 personnes) qui seraient apparus autour du XVII-XVIII^{e} siècle et les Tubalars (2 000 personnes). Leur mode de subsistance traditionnel est basé sur la chasse, la pêche et la cueillette.

Ensemble tuvain

À l'image de l'ensemble altaïen, plusieurs groupes ethniques dont les Irgits, Monguches et Ondars sont rassemblés sous le terme Tuvains, peuplant la République de Tuva en Sibérie du sud (Minahan 2014).

5. D'après des observations de terrain.
6. Informations tirées du *UNESCO red book on endangered languages: Northeast Asia.*

Groupes linguistiques

Tous ces groupes ethniques d'Asie intérieure peuvent être regroupés au sein de deux groupes linguistiques :

1. la branche indo-iranienne de la famille linguistique indo-européenne ;
2. des branches turcique et mongole de la famille linguistique altaïque (Soucek 2000).

Les langues indo-iraniennes d'Asie intérieure sont parlées par le groupe ethnique tadjik et par le groupe des Juifs de Boukhara (parlant le bukhori ou judéo-tadjik). Dans quelques "poches de résistance", dans le Pamir et la vallée du Yagnob au Tadjikistan, certains Tadjiks des montagnes, continuent de parler le yagnobi, l'une des langues iraniennes orientales d'Asie intérieure qui étaient celles des Sogdiens avant la migration iranienne du X^{e} siècle (Heyer et Mennecier 2009). Suite à cette migration, les langues iraniennes occidentales, dont le tadjik, devinrent prédominantes en Asie centrale avant d'être supplantées par les langues altaïques *a priori* au moment de l'apparition des différents groupes ethniques turco-mongols dans la région.
En effet, les langues altaïques (appelées turco-mongoles dans cette thèse) sont parlées par un nombre important de groupes ethniques, comme les Kirghizes, Kazakhs, Ouzbeks ou Mongols, dont la présence en Asie intérieure serait plus récente que celle des Indo-Iraniens (Johanson 1998). Au sein de ce groupe linguistique, la branche turcique, parlée en Asie centrale et du nord, est elle-meme divisée en sous-groupes coïncidant avec la répartition géographique des locuteurs : karluk au sud-est, kiptchak au nord-ouest, oghouze au sud-ouest, sibérienne au nord-est (Consortium Multitree 2014) (Figure 9). En particulier, la langue turkmène est la seule d'Asie intérieure à faire partie de la branche oghouze, qui regroupe des locuteurs de l'ouest de l'Eurasie, azhéris et turcs de Turquie.
L'autre branche, mongole, est parlée uniquement en Asie du nord, et ce en dépit des grandes migrations mongoles du II^{e} millénaire EC.

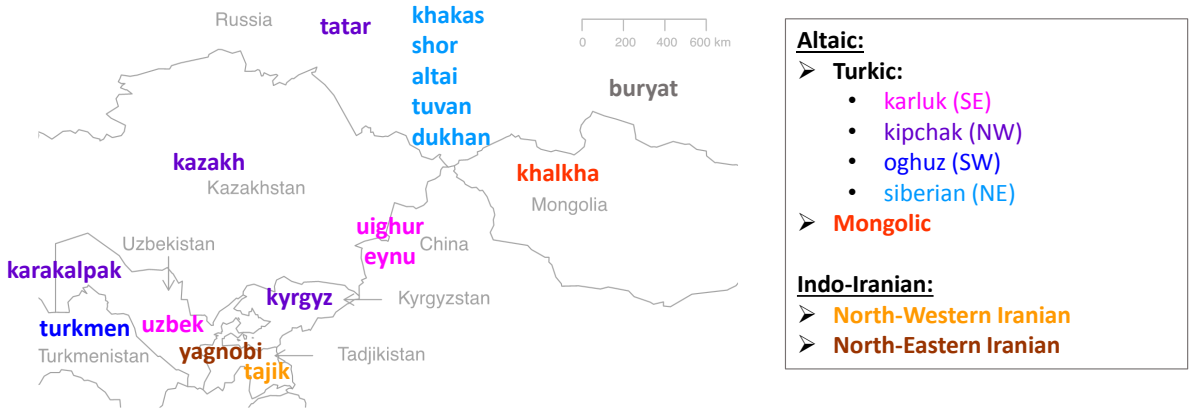


Figure 9 – Répartition géographique des groupes de langues altaïques et indo-iraniennes en Asie intérieure. Établie partir de http://glottolog.org/resource/languoid/.

Organisation sociale et mode de subsistance

En Asie intérieure, les deux groupes linguistiques correspondent chacun à une organisation sociale particulière et à un mode de subsistance qui les distinguent de leurs voisins ne parlant pas la même langue. Par la suite, lorsque nous parlons des Turco-Mongols et Indo-Iraniens, nous mettons donc en contraste deux groupes ayant des langues, des organisations sociales et des modes de subsistance différents.

Groupe linguistique	**Organisation sociale**[1]			**Subsistance**
	Résidence	Filiation	Alliance	
Indo-Iranien	Patrilocalité	Cognatisme (Familles étendues)	Endogamie de village Polygynie légère	Agriculteurs Sédentaires
Turco-Mongol	Patrilocalité	Patrilinéarité (Groupes de descendance)	Exogamie de lignage et clan ; Endogamie de tribu ; Polygynie légère	Éleveurs[2] Semi-nomades[3]

1 - L'organisation sociale se rapporte à la structure d'une société en instances, entretenant des liens réglementés. En particulier, sa compréhension passe par l'étude de la parenté, à travers les règles de résidence, de filiation et d'alliance, dans une conception structuro-fonctionnaliste des sociétés humaines (Ghasarian 1996). Ces règles sont présentées en détail dans le Chapitre II.

• La norme de résidence des couples en Asie intérieure est la patrilocalité, c'est-à-dire une résidence dans le village du mari.

• En Asie intérieure, la filiation ou la transmission de la parenté sociale peut être patrilinéaire si l'appartenance à des groupes de descendance (lignages, clans et/ou tribus) est transmise par le père, ou cognatique si la descendance reconnaît les membres de ses branches généalogiques maternelle et paternelle comme des parents sociaux.

• Les règles d'alliance définissent si les mariages doivent être réalisés au sein d'un même groupe (endogamie géographique, sociale...) ou à l'extérieur (exogamie).

2 - Certains groupes turco-mongols de Sibérie pratiquent des économies traditionnelles comme la pêche, la chasse et la cueillette, plutôt que l'élevage ; les membres du groupe ethnique ouzbek se seraient sédentarisés et pratiqueraient l'agriculture depuis le XVI^{e} siècle (Soucek 2000).

3 - Le semi-nomadisme se réfère à des transhumances entre des campements saisonniers. Pendant la période soviétique, cette pratique traditionnelle a été largement abandonnée, mais connaît un nouveau souffle de nos jours.

Jeu de données analysé

Les populations d'Asie intérieure ont été étudiées depuis 2001 par l'équipe d'Anthropologie Évolutive de l'UMR 7206, dirigée par Évelyne Heyer. Cette équipe a mené plusieurs campagnes d'échantillonnage en Sibérie, Mongolie, Ouzbékistan, Tadjikistan et Kirghizstan, constituant un jeu de données génétiques considérable d'environ 2 100 individus, à la base de nombreux travaux incluant ceux menés au cours de ma thèse.

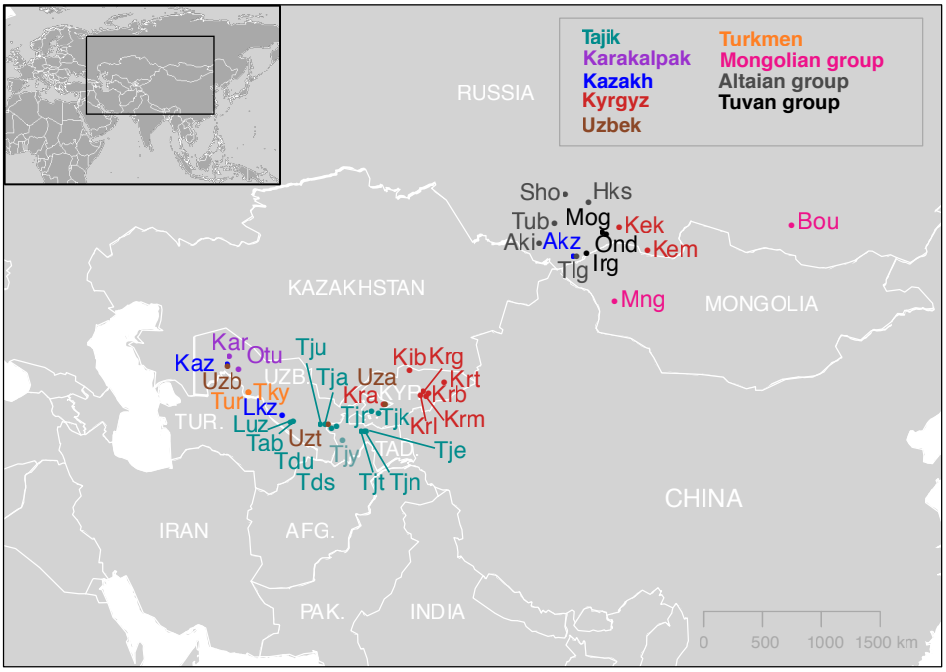


Figure 10 – Répartition géographique des populations étudiées au cours de ma thèse. Chaque trigramme correspond à une population, et la couleur renseigne sur le groupe ethnique, ou l'ensemble, dont elle fait partie. En général, plusieurs populations ont été échantillonnées par groupe ethnique.

L'une des richesses de ce jeu de données est d'associer de l'ADN extrait de prélèvements sanguins ou salivaires et des informations ethno-démographiques. En effet, chaque personne incluse dans la collecte de matériel biologique a aussi répondu à des questions portant, entre autres, sur son lieu de naissance, la composition de sa fratrie, sa langue maternelle, son appartenance lignagère. Le/la volontaire a également répondu pour son conjoint-e-, ses parents et ses beaux-parents (voir le questionnaire ethno-démographique type en Annexe).

En complément, en vue de réaliser des analyses linguistiques, pour chaque lieu d'échantillonnage, entre un et sept individus (moyenne = 3,4) ont répondu à un questionnaire linguistique fondé sur la liste de mots swadesh (Mennecier *et al.* 2015). Cet outil est utilisé en linguistique comparée pour établir quels sont les mots utilisés par le locuteur pour parler d'environ 200 concepts jugés universels ("femme", "homme", parties du corps, "voir", "parler"...), en vue de quantifier des différences linguistiques entre les locuteurs.

Également, au cours de plusieurs missions, l'accent a été mis sur l'évolution de certains traits phénotypiques, tels que le diabète de type II ou la perception des goûts (Ségurel *et al.* 2013 ; Sjöstrand 2015), et, à ces fins, des données phénotypiques, médicales et nutritives ont été collectées.

Les volontaires sont généralement des personnes d'une cinquantaine d'années, ayant *a priori* fini leur vie reproductive matrimoniale (voir l'âge des participants parmi les informations ethno-démographiques produites en Annexe).

Historiquement, les premières missions de l'équipe d'Anthropologie Évolutive se sont concentrées sur les hommes d'Asie centrale, mais par la suite l'ADN des femmes a été aussi collecté, et l'aire d'échantillonnage a été étendue à l'Asie du nord (voir le nombre et sexe des participants en Annexe). Ainsi, ce jeu de données nous permet de considérer l'Asie intérieure dans son ensemble, ce qui est particulièrement intéressant du fait d'échanges humains incessants à l'œuvre dans cette région depuis l'âge du Bronze. Les 2100 individus échantillonnés sont regroupés en 48 populations, dont 41 utilisées dans mes travaux de thèse, et qui représentent 18 groupes ethniques (Figure 10). Nous définissons le groupe ethnique de chaque individu sur la base de la langue maternelle qu'il a indiquée dans le questionnaire ethno-démographique. Nous appelons population un groupe d'individus parlant la même langue et échantillonnés dans la même aire géographique (village ou groupe de villages voisins).
Lors de l'échantillonnage, nous n'avons inclus que des individus non-apparentés sur la base des données ethno-démographiques. Cela a été vérifié par la suite au moyen d'informations génétiques autosomales et du logiciel Relpair (Michael Boehnke 1997 ; Epstein *et al.* 2000).

Les informations génétiques incluses dans ce jeu de données correspondent à plusieurs types de marqueurs :

- **SNP,** pour *Single-Nucleotide Polymorphism*, qui est un polymorphisme d'une seule base de l'ADN. Nous avons obtenu ce type de données par détection Taqman pour des loci d'intérêt du chromosome Y (Balaresque *et al.* 2015 ; Marchi *et al.* 2017), par séquençage de l'ADN mitochondrial complet ou de HVS1 (Ségurel *et al.* 2008 ; Marchi *et al.* 2017), et par séquençage de régions d'intérêt et génotypage sur puce à ADN pour des données autosomales (Ségurel *et al.* 2013). Plusieurs puces à ADN ont été utilisées pour génotyper les différentes populations étudiées. Lors de cette thèse, j'ai fusionné les informations de cinq de ces puces en un set de 253 532 SNPs, après avoir réalisé pour chacune des puces un contrôle-qualité des marqueurs et des individus. Ce processus est détaillé dans le Chapitre III *(Supplementary Data* de l'article).
- **STR,** pour Short-Tandem Repeat, qui est un polymorphisme de répétitions d'une courte séquence d'ADN, d'une longueur de 2 à 5 paires de bases. L'information utilisée est le nombre de répétitions du motif. Nous avons obtenu des STRs pour les autosomes, les chromosomes X et Y (Ségurel *et al.* 2008 ; Marchi *et al.* 2017). Notons que ce type de marqueur mute plus rapidement que la plupart des SNPs, du fait d'erreurs de réplication fréquentes de la part de la polymérase (Fan et Chu 2007).

Les travaux présentés dans ce manuscrit portent en particulier sur des SNPs autosomaux (Chapitres I et III) et des marqueurs du chromosome Y et de l'ADN mitochondrial (Chapitre II). Nous précisons, au fil des chapitres, les marqueurs et les populations que nous avons utilisés.

Le jeu de données de l'équipe d'Évelyne Heyer, unique pour la région, nous a amenées à collaborer avec l'équipe d'Eske Willerslev, du GeoGenetics Center de Copenhague. Cette équipe danoise menait un projet appelé *A population genomic history of the steppe*, visant à donner une image génétique et évolutive des populations présentes dans les steppes eurasiatiques depuis l'âge du Bronze jusqu'à nos jours. Pour cela, 137 génomes anciens, datant de 2 500 ans AEC à 1 500 ans EC, ont été séquencés. En complément, nous avons fourni 331 des 502 génotypes modernes produits pour cette étude, auxquels s'ajoutent 320 génotypes extraits de la littérature, que nous avons contribué à analyser. Ces individus modernes sont répartis en 53 populations et 16 groupes ethniques, génotypés pour un total de 242,406 SNPs autosomaux. Ces données anciennes et modernes constituent le jeu de données *Steppes* auquel nous nous référons dans le premier chapitre de ce manuscrit.

Objectifs de la thèse

Située au cœur de l'Asie, l'Asie intérieure a été, et reste, une zone de contacts entre des peuples de cultures et de langues différentes. Leurs rencontres ont généré des échanges de biens, d'idées et de gènes, façonnant la diversité culturelle et biologique de cette région et lui confèrent un intérêt anthropologique. À l'heure actuelle, de nombreux groupes ethniques coexistent en Asie intérieure, héritiers de ce passé complexe et constituent un matériel précieux pour la recherche en anthropologie génétique.

Ce manuscrit rend compte des travaux réalisés au cours de ma thèse : il s'organise en trois chapitres, correspondant à trois sujets de recherche s'intéressant à l'influence de la culture sur la diversité génétique.

Le **Chapitre I** porte sur l'histoire du peuplement humain de l'Asie intérieure dans le contexte de l'Eurasie, à l'origine de la diversité génétique actuelle. Il débute par une description de la diversité autosomale actuelle recensée en Asie intérieure, en particulier en étudiant les ressemblances et dissimilarités génétiques entre les groupes linguistiques et ethniques. Puis, il présente une tentative d'inférence de l'histoire des habitants de cette région depuis l'âge du Bronze jusqu'à nos jours, en couplant des données d'ADN ancien et moderne dans le cadre du projet collaboratif *Steppes*.

Le **Chapitre II** s'intéresse à l'impact des comportements culturels asymétriques entre sexes, comme des migrations sexe-spécifiques, sur la diversité génétique des populations. Nous avons mesuré cet effet en contrastant la diversité génétique observée, pour des marqueurs sexe-spécifiques, au sein des populations et entre les populations turco-mongoles et indo-iraniennes. Ces analyses ciblent en particulier des comportements liés aux règles de résidence et de filiation. Nous avons également exploré le processus d'ethnogénèse du point de vue des lignées masculines à partir d'informations génétiques du chromosome Y.

Le **Chapitre III** traite de la consanguinité, manifestation génétique d'unions entre apparentés pouvant être motivées par des choix culturels et qui correspond à une réduction de la diversité génétique individuelle. Les travaux menés dans ce chapitre s'intéressent au lien entre la consanguinité et les migrations matrimoniales, imposées par des règles culturelles [7]. En effet, sous l'hypothèse que des individus éloignés géographiquement le sont aussi génétiquement, les migrations seraient un moyen d'éviter de se reproduire avec ses parents. Pour tester cette hypothèse, nous avons suivi une approche pluri-disciplinaire visant à confronter, à l'échelle de l'individu et de la population, une distance de dispersion calculée à partir de données ethnologiques et la consanguinité estimée à partir de données *genome-wide*.

Ces travaux ont donné lieu à une publication en tant que premier auteur dans le journal AJPA, et deux autres manuscrits sont actuellement en révision et re-considération par les journaux Scientific Reports et Nature (en premier et second auteur respectivement). Chaque chapitre est organisé autour de l'une de ces publications et inclut un développement bibliographique autour de la thématique ciblée, une présentation des résultats principaux obtenus ainsi que la publication complète, puis un paragraphe de réflexions personnelles sur le travail effectué et ses perspectives.
Afin de fluidifier la lecture de ce manuscrit, j'ai choisi de présenter les méthodes utilisées en Annexes.

7. Les règles d'alliance définissent si les mariages doivent être réalisés au sein d'un même groupe, à l'échelle du village dans notre cas (endogamie géographique), ou à l'extérieur (exogamie).

Ma thèse s'inscrit dans l'une des thématiques de recherche développées par l'équipe d'Anthropologie Évolutive de l'UMR7206, qui étudie comment être un "animal social" a pu jouer sur notre évolution biologique. Cette approche pluri-disciplinaire nuance la tendance du "tout-génétique" ainsi que l'opposition classique entre biologie et culture, et tente de rendre compte de la nature mixte de l'Homme [8].

J'ai réalisé mes travaux à partir de données génétiques et ethnologiques collectées avant mon arrivée dans l'équipe, que j'ai traitées et analysées selon les modalités décrites dans ce manuscrit. Ma contribution est originale dans le sens où j'ai été la première à analyser les données récoltées en Asie du nord, alors que les travaux préalables se concentraient sur l'Asie centrale. De plus, bien qu'une partie de mes travaux s'inscrive dans la continuité de ceux menés par l'équipe (Chaix *et al.* 2007 ; Heyer *et al.* 2009 ; Martínez-Cruz *et al.* 2011), la thématique de la consanguinité n'avait jamais été abordée à partir de ce jeu de données. Enfin, ce manuscrit présente la première utilisation couplée d'ADN ancien et des données modernes autosomales collectées par l'équipe d'Anthropologie Évolutive.

8. "Tout est fabriqué et tout est naturel chez l'homme." (Merleau-Ponty - Phénoménologie de la perception - p.221)

Chapitre I

Histoire du peuplement de l'Asie intérieure

Sommaire

I.1 Diversité génétique actuelle

Résultats préalablement obtenus en Asie centrale par l'équipe d'Anthropologie Évolutive

Préalablement à ma thèse, les travaux réalisés par l'équipe d'Anthropologie Évolutive se concentraient sur des populations d'Asie centrale, dans la région occidentale de l'Asie intérieure. 26 populations ont notamment été génotypées pour 27 STRs autosomaux et ont été analysées en contexte avec d'autres populations d'Eurasie (Martínez-Cruz *et al.* 2011). Les auteurs ont montré que, génétiquement, les populations d'Asie centrale se situent le long du gradient formé entre les populations est-asiatiques et les populations européennes (Figure I.1). Plus en détail, les populations d'Asie centrale apparaissent sur le graphique en deux groupes qui correspondent aux groupes linguistiques (et donc d'organisation sociale et de mode de subsistance). En effet, la distance sur le graphique et donc génétique entre des populations du même groupe linguistique semble plus faible que la distance mesurée entre des populations indo-iraniennes et turco-mongoles. Une AMOVA réalisée en séparant les populations selon ce critère linguistique confirme une différence génétique significative entre ces groupes (p-value<0,01) (Heyer et Mennecier 2009). Enfin, les auteurs ont observé que les populations indo-iraniennes étaient proches du pôle européen, et que les populations turco-mongoles étaient proches du pôle asiatique. En outre, au sein de chaque groupe linguistique, les distances entre populations calculées à partir de données génétiques autosomales et de données linguistiques sont corrélées (test de Mantel, p-value<0,05) : plus les populations sont linguistiquement proches, plus elles le sont aussi génétiquement. Cette corrélation ne s'explique pas par des distances géographiques réduites (corrélation de Spearman non significative entre les distances génétiques et géographiques, p-value = 0,18). Ainsi, ces résultats suggèrent des échanges génétiques entre populations sur la base d'une proximité linguistique et non géographique : la langue (et/ou l'organisation sociale, et le mode de subsistance) est une barrière aux flux géniques en Asie centrale.

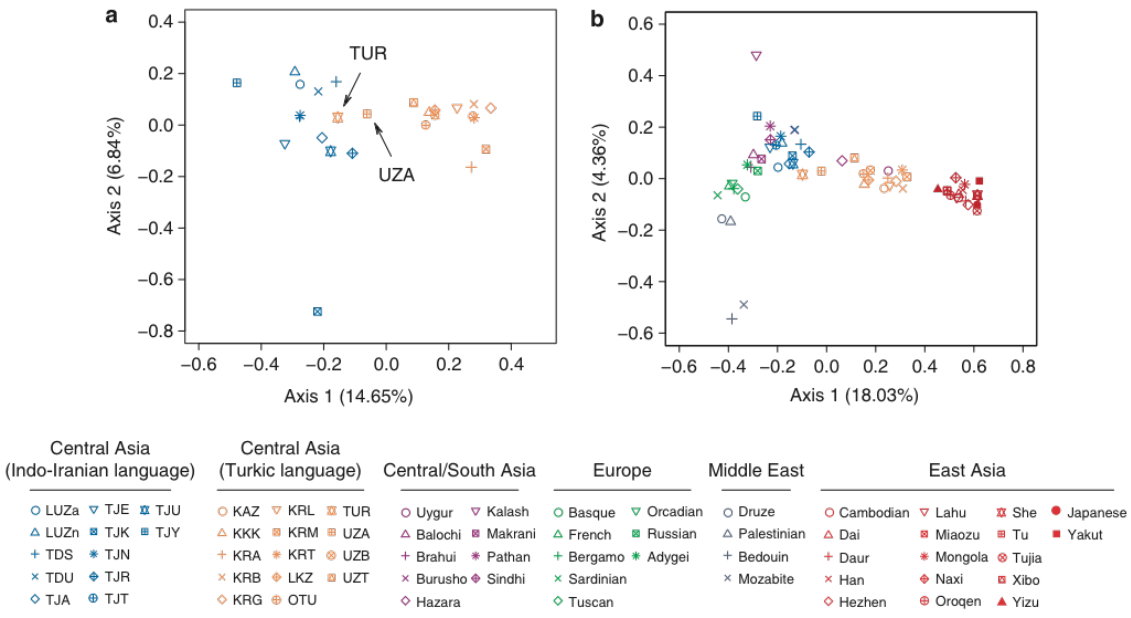


Figure I.1 – Analyse factorielle des correspondances réalisée pour 26 populations d'Asie centrale (a), et en contexte avec d'autres populations d'Eurasie (b). La couleur bleue est utilisée pour représenter les populations indo-iraniennes et la couleur orange, les populations turco-mongoles. *Figure 2 de Martínez-Cruz* et al. *(2011).*

Cependant, certaines populations de langue turco-mongole, comme les Turkmènes (Tur), sont génétiquement plus proches des populations indo-iraniennes. Cette observation pourrait être due à un changement linguistique d'une langue ancestrale *a priori* indo-iranienne vers une langue de prestige turco-mongole parlée de nos jours par ce groupe ethnique, sans que cela soit associé à un métissage génétique du fonds indo-iranien avec les locuteurs turco-mongols. Cette hypothèse a aussi été avancée pour décrire le processus de diffusion de la langue turcique à l'échelle de toute l'Eurasie, depuis la Sibérie, entre le IX^{e} et le XVII^{e} siècle (Yunusbayev *et al.* 2015).

Parmi les 26 populations d'Asie centrale, on compte trois populations ouzbèkes, dont l'une (UZA) est retrouvée parmi le groupe génétique indo-iranien, tandis que les deux autres (UZB et UZT) sont dans le groupe turco-mongol (Figure I.1). L'analyse linguistique accompagnant cette analyse génétique révèle une hétérogénéité au sein du groupe ouzbek : la population génétiquement indo-iranienne parle une langue karluk (ce qui est classique pour les Ouzbeks), tandis que les deux autres populations parlent une langue kiptchak de la même branche que les langues kirghize et kazakhe. Ces hétérogénéités linguistiques et génétiques pourraient remonter à la création du groupe par un processus de coalition entre de tribus parlant des langues différentes, et ayant des fonds génétiques différents (Soucek 2000).

Plus généralement, la place occupée par l'Asie centrale sur le gradient eurasiatique soulève des interrogations sur le rôle de cette région dans l'histoire de l'Eurasie : est-ce une zone de métissage génétique entre l'Europe et l'Asie, ou bien est-ce une source ayant contribué au peuplement de l'Europe et du reste de l'Asie ? Le modèle retenu combine ces deux hypothèses : à certaines périodes, l'Asie intérieure aurait été la source de migrations vers d'autres régions d'Eurasie, tandis qu'à d'autres elle aurait reçu des migrations. Notamment, une étude menée sur des marqueurs uniparentaux pour des populations modernes originaires de toute l'Eurasie a détecté une expansion démographique datant d'il y a 60 000 ans en Asie de l'est qui se serait propagée, *via* des migrations humaines, en Asie intérieure et jusque dans l'ouest de l'Eurasie (Chaix *et al.* 2008). Plus récemment, l'Asie centrale aurait été la source d'une vague de migrations en direction de l'Asie du sud il y a moins de 10 000 ans, d'après des informations tirées des haplogroupes du chromosome Y (Quintana-Murci *et al.* 2001). En sus, l'Asie intérieure aurait joué un rôle clé dans le peuplement du continent américain que nous n'abordons pas au cours de cette thèse (Mazières 2011 ; Reich *et al.* 2012).

La coexistence des différents groupes linguistiques et ethniques en Asie intérieure soulèvent des interrogations majeures quant à leur constitution : quand se sont-ils formés et à partir de quelles sources ? Les différences culturelles observées sont-elles liées à des histoires évolutives variées ?
Des éléments de réponse ont été apportés par l'utilisation des données génétiques dans une approche ABC à partir d'un scénario assez simple : les populations actuelles d'Asie centrale résulteraient d'un métissage génétique entre des Européens et des Asiatiques (Palstra *et al.* 2015). Ce scénario a été testé pour deux populations : les TAB, choisis comme représentants des Tadjiks et des Indo-Iraniens, et les KIB, pris comme représentants des Kirghizes et des Turco-Mongols.
Le modèle qui convient le mieux à la population tadjike est un métissage entre des populations européennes et d'autres venues d'Asie de l'est, il y a 8 275 ans [2 350 - 24 675] pour un temps de génération de 25 ans, formant un groupe proto-Tadjik dont on ne connaît pas les traits culturels mais qui aurait été de langue indo-iranienne. Pour la population kirghize, le métissage aurait impliqué ces proto-Tadjiks et des populations asiatiques, peut-être venues d'Asie du nord, il y a 2 350 ans [350 - 7 600].
Une autre approche basée sur un modèle d'*isolation-with-migration* (Hey 2010) donne un scénario cohérent avec les inférences ABC et ajoute une information supplémentaire sur la divergence entre les

populations kirghizes et asiatiques il y a 6 917 ans [2 454 – 17 180], et entre les populations tadjikes et européennes actuelles il y a 21 643 ans [11 379 – 35 923]. De plus, cette méthode montre un flux de gènes important depuis l'Asie en direction des Kirghizes, mais aussi en direction des Tadjiks et des Européens à une moindre intensité.
Ce scénario de métissages génétiques successifs est cohérent avec la différenciation moins prononcée observée entre les populations turco-mongoles d'Asie centrale qu'entre les populations indo-iraniennes (Martínez-Cruz *et al.* 2011), mais cette différence de diversité génétique pourrait également avoir été causée par d'autres facteurs (voir Chapitre II).

Outre ces origines différentes, ces deux populations d'Asie centrale, KIB et TAB, vivent selon des modes de subsistance très différents, à savoir respectivement le pastoralisme semi-nomade et l'agriculture sédentaire. Du fait de ces différences de mode de vie, on pourrait s'attendre à des différences de croissance démographique : dans le reste de l'Eurasie et en Afrique, le nomadisme est associé à une croissance démographique réduite, voire nulle, comparée à celle des sédentaires. Cependant, en Asie centrale, les populations KIB, TAB et 28 autres populations agricultrices ou pastorales ont connu une expansion de même intensité quelque soit le mode de vie, pour des données autosomales ou mitochondriales (Aimé *et al.* 2013, 2014). Cette similarité entre nomades et sédentaires peut s'expliquer par deux hypothèses non-exclusives : l'aridité ayant régné par le passé en Asie centrale n'était pas optimale pour le développement de l'agriculture et aurait pu restreindre la croissance des populations agricultrices ; les populations pastorales étaient agricultrices avant de se convertir au pastoralisme et ont donc connu une croissance démographique "agricultrice" avant leur transition.

Nouvelles analyses génétiques à l'échelle de l'Asie intérieure

En sus des populations étudiées en Asie centrale, de nouvelles populations ont été échantillonnées en 2011-2012 en Asie du nord. Ainsi, au cours de ma thèse, j'ai pu explorer l'histoire démographique de 17 populations à une échelle géographique large : celle de l'Asie intérieure, à partir de données *genome-wide* représentant 253 532 SNPs.

Dans un premier temps, j'ai étudié les *allele-sharing dissimilarities* (ASD) pour toutes les paires d'individus et les ai représentées à partir d'une approche *Multidimensional scaling* (MDS) en deux dimensions (Figure I.2). Pour s'affranchir de la réduction de dimension inhérente à la MDS, j'ai également réalisé un arbre de neighbour-joining (Gascuel 1997) basé sur des distances F_{ST} entre populations.
Ces deux approches donnent des résultats cohérents entre eux, et avec ce qui avait été précédemment observé en Asie centrale : on retrouve deux groupes de populations génétiquement proches qui correspondent globalement aux groupes linguistiques. En effet, la première dimension de la MDS sépare les populations indo-iraniennes (sur la droite du graphique), des populations turco-mongoles. Cette observation est également réalisée à partir de l'arbre de *neighbour-joining* avec une longue branche indo-iranienne d'un côté et un regroupement de branches turco-mongoles plus courtes de l'autre. On retrouve le cas exceptionnel des Turkmènes (Tur) qui, bien que de langue turco-mongole, se rapprochent, génétiquement, des Indo-Iraniens. Quantitativement, les valeurs moyennes de différenciation génétique entre des populations du même groupe linguistique sont significativement trois fois plus faibles que celles entre populations de groupes différents (F_{ST} entre populations turco-mongoles = 0,013 ; entre populations indo-iraniennes = 0,012 ; turco-mongoles *vs* indo-iraniennes = 0,031 ; p-values<0.003 pour des tests de Mann-Whitney bilatéraux). Ces résultats indiquent une certaine unité génétique au sein du groupe turco-mongol, en dépit de leur grande aire de répartition, à la fois en Asie du nord et centrale. Plus en détails, on observe une tendance pour les populations turco-mongoles d'Asie centrale à se situer plutôt au centre

du graphique issu de la MDS quand celles d'Asie du nord sont sur la gauche. Cependant, cette distinction n'est pas systématique : la population kazakhe d'Asie du nord (Akz) se retrouve superposée au nuage turco-mongol d'Asie centrale, et ressemble fortement génétiquement aux Kazakhs d'Asie centrale (Kaz), avec un $F_{ST} = 0,002$, p-value $<10^{-3}$. Le groupe ethnique kirghize est lui aussi présent à la fois en Asie centrale (Kib) et du nord (Kel), mais ces populations sont moins proches sur les graphiques et leur F_{ST} est plus élevé ($0,007$, p-value $< 10^{-3}$).
En outre, la seconde dimension de la MDS permet de distinguer trois populations d'Asie du nord du reste du nuage turco-mongol : les Khakasses (Hks), les Chors (Sho) et les Tubalars (Tub), qui ont la particularité de vivre selon un mode de subsistance différent de celui des autres Turco-Mongols, à savoir de chasse, de pêche et de cueillette plutôt que d'élevage.

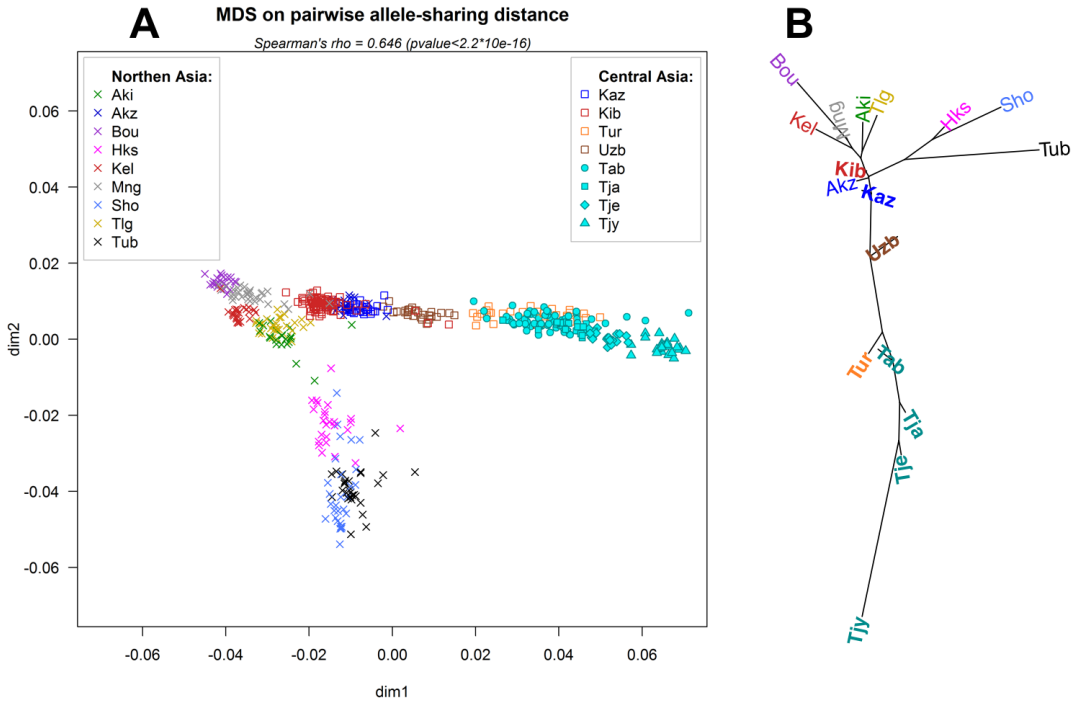


Figure I.2 – Représentation des *allele-sharing dissimilarities* entre individus analysées par *Multidimensional scaling* (A) et *neighbour-joining* basé sur les distances F_{ST} entre populations (B), calculées à partir de SNPs indépendants. Les populations indo-iraniennes sont représentées en turquoise ; sur la figure A) les populations d'Asie centrale sont représentées par des cercles ou des carrés, et les populations d'Asie du nord par des croix ; sur la figure B) le nom des populations d'Asie centrale est indiqué en gras. Le ρ de Spearman indique la corrélation entre la matrice d'ASD et la matrice des distances réduites par la MDS.

Afin d'acquérir une meilleure définition des proximités génétiques entre individus ou populations, nous avons réalisé des analyses d'ADMIXTURE, pour un nombre de composantes K compris entre 2 et 5, avec 30 répétitions à chaque fois.
À K=2, nous avons trouvé un seul mode pour les 30 répétitions. Les deux composantes respectivement orange et jaune sont définies par la population tadjike Tjy et par la population bouryate Bou. Nous avons observé une plus grande part de composante orange chez les populations indo-iraniennes (Tja, Tje, Tjy) et chez les Turkmènes (Tur). Hormis ces derniers, la composante jaune est la plus importante chez les autres populations turco-mongoles. Nous avons observé que les populations kirghizes d'Asie du nord (Kek et Kem) diffèrent génétiquement de la population kirghize d'Asie centrale (Kib), tandis que les populations kazakhes (Kaz et Akz) d'Asie centrale et du nord se ressemblent fortement.

À K=3, les 30 répétitions montrent une composante bleue émergeant chez trois populations turco-mongoles de Sibérie (Hks, Sho et Tub). Nous avons déjà observé que ces populations diffèrent génétiquement des autres dans la Figure I.2.
À K=4, nous avons détecté deux modes : le premier, trouvé pour 19 répétitions, montre une composante rose définie par des membres de la population Tjy, chez qui elle est retrouvée en proportions importantes et de manière quasi-exclusive étant faiblement retrouvée chez les autres populations). Le second mode, observé pour 11 répétitions, montre une composante verte principalement retrouvée chez deux populations sibériennes (Hks et Sho), chez qui elle remplace la composante bleue trouvée à K=3. Les Khakasses, parfois décrits comme héritiers des anciens Kirghizes sibériens, présentent un profil génétique différant de celui des Kirghizes modernes, suggérant qu'ils ne partagent pas de lien de parenté récent.
À K=5, les 30 répétitions donnent un consensus des modes trouvés à K=4, avec la population Tjy fortement représentée par la composante rose, et Hks et Sho par la composante verte.

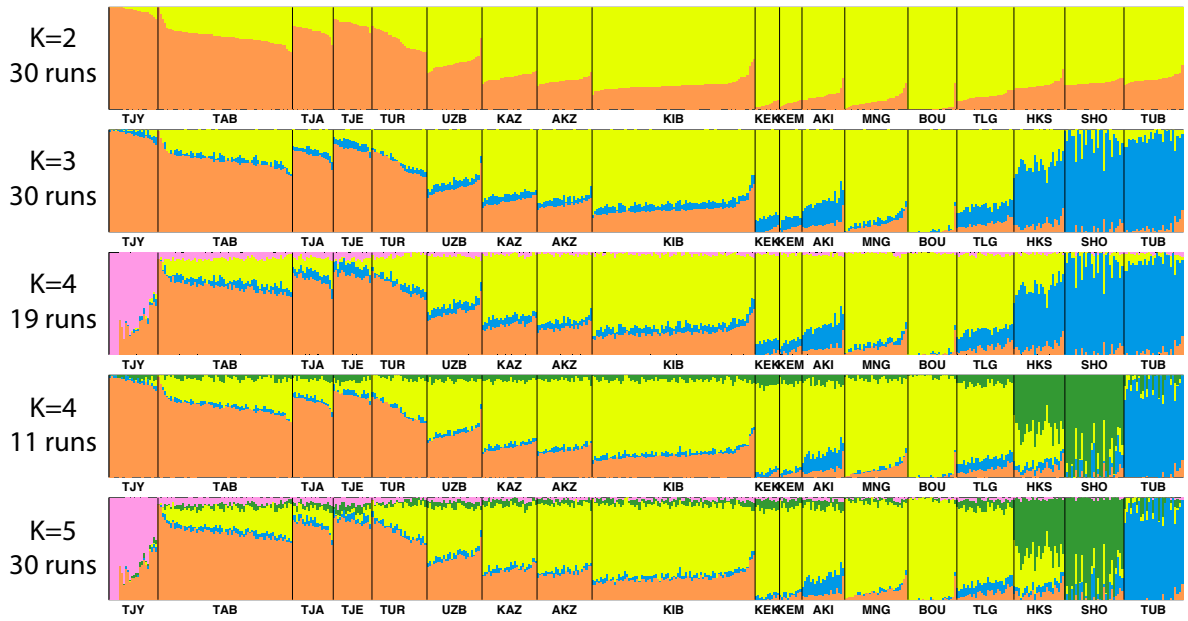


Figure I.3 – Graphique d'ADMIXTURE. Le nombre de composantes explorées K est indiqué dans la marge, ainsi que le nombre de répétitions pour lequel chaque mode a été observé. Chaque composante est représentée par une couleur et chaque individu est représenté par une ligne verticale divisée en segments de couleur, dont les hauteurs sont proportionnelles à la part de chaque composante dans le génome de l'individu.

Le panorama génétique produit dans cette thèse est le plus complet à avoir été réalisé à ce jour pour la région de l'Asie intérieure, en incluant 17 populations, 12 groupes ethniques, deux familles linguistiques et des données autosomales *genome-wide*. Nous montrons en particulier que la diversité génétique observée est principalement organisée autour des deux groupes linguistiques et culturels. Nos travaux soulignent également que les Turco-Mongols d'Asie centrale et du nord partagent un fonds génétique commun, mais que l'on trouve néanmoins des dissimilarités entre ces deux régions.
À partir de ces seules analyses, nous ne savons cependant pas ce qui a pu causer les dissimilarités génétiques observées entre les groupes turco-mongol et indo-iranien. Elles pourraient découler d'origines différentes pour ces deux groupes en lien avec le peuplement de cette région, ou bien avoir été causées par un phénomène de dérive à partir d'un fonds génétique commun. Pour départager ces scénarios, il nous faut donc aller au-delà de la seule description de la diversité génétique et retracer l'histoire évolutive des populations de cette région. Pour ce faire, nous avons recours à des données paléogénétiques.

I.2 Inférer l'histoire au moyen de la paléogénétique

Les résultats et données présentés dans cette partie couvrent la période de l'âge du Bronze au Moyen-Âge et sont tirés de travaux publiés ainsi que de ceux réalisés au cours du projet *Steppes* auquel nous avons collaboré.

Éclairages paléogénétiques et bibliographiques sur le peuplement de l'Asie intérieure

À l'âge du Bronze

Plusieurs auteurs se sont intéressés à l'origine des populations d'Asie intérieure de l'âge du Bronze (pour mémoire, ce sont les cultures d'Afanasievo, BMAC, Sintashta, Andronovo, Karassouk et Okunevo). Des analyses paléogénétiques ont pointé une forte composante européenne chez les populations d'Asie intérieure de l'âge du Bronze et ont tenté de l'associer à une culture archéologique en particulier (Allentoft *et al.* 2015). Une candidate sérieuse serait la culture Yamnaya peuplant les steppes européennes à l'âge du Bronze ancien, et vraisemblablement à l'origine de la diffusion en Eurasie des langues indo-européennes et de certains phénotypes qualifiés d'européens (yeux et peau clairs, cheveux parfois roux) (Keyser *et al.* 2009 ; Hollard *et al.* 2014).
Cette culture s'est avérée être à l'origine de migrations à la fois en direction de l'Europe de l'ouest (Anthony 2010) et vers l'est jusqu'en Asie intérieure orientale (Ning *et al.* 2015 ; Hollard *et al.* 2014 ; González-Ruiz *et al.* 2012 ; Li *et al.* 2010) (Figure I.4). En ce qui concerne l'Asie intérieure, la composante européenne retrouvée pour les populations de la culture d'Afanasievo du Bronze ancien serait effectivement héritée des Yamnaya (Allentoft *et al.* 2015), mais les populations du Bronze tardif, de la culture Sintashta ou d'Andronovo, découleraient plutôt d'une autre culture européenne du Bronze ancien : celle de la céramique cordée, retrouvée plus à l'ouest de l'Europe que les Yamnaya.

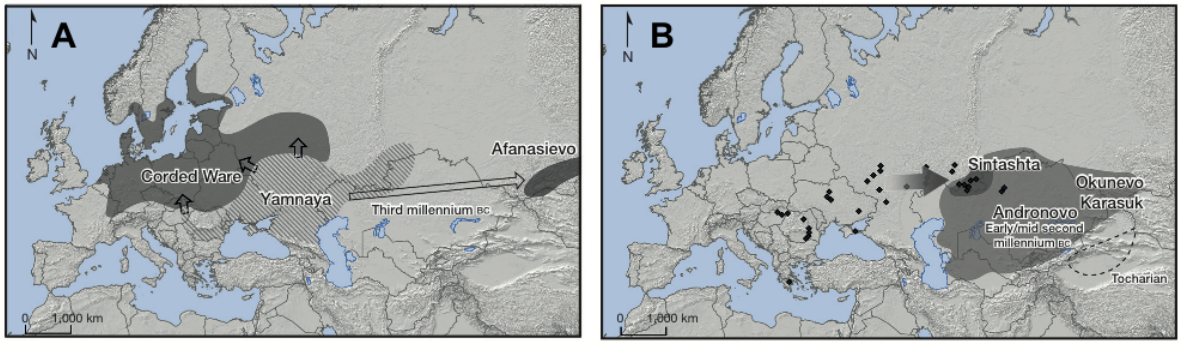


Figure I.4 – Répartition géographique des cultures de l'âge du Bronze (A) ancien : Yamnaya, Cordée et Afanasievo ; (B) tardif : Sintashta, Andronovo, Okunevo, et Karassouk. Les flèches indiquent les migrations associées aux Yamnaya. Figure 1 (Allentoft *et al.* 2015).

À l'âge du Fer

À l'âge du Fer, le nomadisme réapparut en Asie intérieure *via* la culture scythe qui dominait l'Asie intérieure à cette période. Une étude réalisée à partir de 150 séquences mitochondriales HVS1 provenant de l'extrême est de l'Europe et de l'est de l'Asie intérieure (d'Altaï et de Khakassie), a inféré une origine multirégionale au groupe scythe par une approche ABC : les Scythes de l'ouest et de l'est ne partagaient pas d'ancêtre commun récent sur la base de données mitochondriales (Unterländer *et al.* 2017).

De plus, à partir de huit génomes, il semblerait que les Scythes soient issus d'un mélange génétique entre des Yamnaya et des Asiatiques de l'est ancestraux non identifiés, donc sans continuité avec les cultures d'Asie centrale du Bronze tardif.

D'autre part, une analyse réalisée pour la séquence HVS1 de 36 squelettes trouvés au Kazakhstan (1 500 AEC - 1 500 EC) a documenté l'arrivée d'une vague venue de l'est entre le VII^{e} et le III^{e} siècles AEC : les squelettes datant de cette époque portent des haplogroupes assignés à l'est de l'Asie, alors que précédemment seules des lignées européennes étaient détectées (Lalueza-Fox *et al.* 2004). La culture associée à ces migrants asiatiques n'est pas identifiée, et pourrait coïncider avec celle à l'origine des Scythes étant contemporaines.

Apports du projet *Steppes*

Hormis ces quelques références, peu d'études et de données paléogénétiques sont disponibles pour la période post-néolithique en Asie intérieure. Le projet *Steppes* mené par Eske Willerslev (GeoGenetics Center, Copenhague) et auquel nous avons collaboré vient compléter ces lacunes, en séquençant *de novo* 137 génomes anciens (2500 ans AEC - 1500 ans EC) auxquels s'ajoutent des génomes précédemment publiés[1−5]. Ce jeu de données est le plus conséquent à avoir été produit pour une étude de paléogénétique (Table I.1). À partir de ces données anciennes, il est alors possible de réaliser un suivi temporel depuis l'âge du Bronze jusqu'au Moyen Âge des populations de cette région et plus largement de toute la steppe eurasiatique.

Afin de décrire les proximités génétiques entre différentes populations, ces génomes sont analysés au moyen d'Analyses en Composantes Principales (Patterson *et al.* 2006) et d'ADMIXTURE (Alexander *et al.* 2009). De plus, des flux de gènes potentiels, survenus entre les populations sont détectés par des D-statistiques (Green *et al.* 2010), et les contributions ancestrales à l'origine de populations génétiquement métissées sont inférées par une approche qpAdm (Haak *et al.* 2015).

Les résultats obtenus sont succinctement présentés dans ce chapitre et le manuscrit en cours de considération dans le journal Nature est rendu disponible en Annexes.

1. (Rasmussen *et al.* 2014)
2. (Raghavan *et al.* 2014)
3. (Haak *et al.* 2015 ; Mathieson *et al.* 2015)
4. (Olalde *et al.* 2014 ; Lazaridis *et al.* 2014 ; Gamba *et al.* 2014)
5. (Lazaridis *et al.* 2016)
6. (Allentoft *et al.* 2015 ; Mathieson *et al.* 2015)
7. (Lazaridis *et al.* 2014 ; Mathieson *et al.* 2015 ; Olalde *et al.* 2015)

Table I.1 – Descriptif des données d'ADN ancien du projet *Steppes* : les génomes sont regroupés en population selon la culture matérielle à laquelle ils sont associés, leur période et leur lieu d'échantillonnage.

Population Label	Geo. Range	Period	Language	Subsistence	N
New samples					
Lchashen Metsamor	Caucasus	Iron Age	Indo-European ?	Nomadic / pastoral	2
Alan	Caucasus & Europe	Late Iron Age to Medieval	Indo-European	Nomadic / pastoral	6
Hallstatt-Bylany	Europe	Iron Age	Indo-European	Agriculturist	2
Hungarian Scythian	Europe	Iron Age	Indo-European	Nomadic / pastoral	4
Sarmatian	Europe	Iron Age	Indo-European	Nomadic / pastoral	10
North Lithuania	Europe	Late Iron Age	Indo-European	Agriculturist	1
Poprad	Europe	Late Iron Age	Indo-European	Agriculturist	1
Medieval Hungarian	Europe	Medieval	Indo-European	Unknown	1
Saltovo-Mayaki	Europe	Medieval	Mixed	Nomadic / pastoral	3
Andronovo	Central Asia	Late Bronze Age	Indo-European	Nomadic / pastoral	1
Central Saka	Central Asia	Iron Age	Indo-European	Nomadic / pastoral	8
Iron Age Nomad	Central Asia	Iron Age	Indo-European	Nomadic / pastoral	4
Tagar	Central Asia	Iron Age	Indo-European	Nomadic / pastoral	8
Tian Shan Saka	Central Asia	Iron Age	Indo-European	Nomadic / pastoral	12
Hun-Sarmatian	Central Asia	Late Iron Age	Unknown	Nomadic / pastoral	2
Tian Shan Hun	Central Asia	Late Iron Age	Altaic	Nomadic / pastoral	23
Hun Period Nomad	Central Asia	Late Iron Age	Altaic	Nomadic / pastoral	2
Wusun	Central Asia	Late Iron Age	Indo-European	Nomadic / pastoral	4
Kangju	Central Asia	Late Iron Age to Medieval	Indo-European	Nomadic / pastoral	6
Golden Horde	Central Asia	Medieval	Altaic	Nomadic / pastoral	2
Kara-khanid	Central Asia	Medieval	Altaic	Nomadic / pastoral	3
Karluk	Central Asia	Medieval	Altaic	Nomadic / pastoral	2
Kimak	Central Asia	Medieval	Altaic	Nomadic / pastoral	1
Kipchak	Central Asia	Medieval	Altaic	Nomadic / pastoral	2
Medieval Nomad	Central Asia	Medieval	Altaic	Nomadic / pastoral	10
Turk	Central Asia	Medieval	Altaic	Nomadic / pastoral	3
Historical Kazakh	Central Asia	Historical	Altaic	Nomadic / pastoral	2
Historical Nomad	Central Asia	Historical	Altaic	Nomadic / pastoral	1
Glazkovo	East Asia	Early Bronze Age	Unknown	Hunter-Gatherer	4
Xiongnu	East Asia	Late Iron Age	Altaic	Nomadic / pastoral	3
Western Xiongnu	East Asia	Late Iron Age	Altaic	Nomadic / pastoral	2
Previously published					
Clovis[1]	America				1
Mal'ta (MA1)[2]	Siberia	Paleolithic			1
EHG[3]	Europe (East)	Meso-Neolithic	Indo-European	Hunter-Gatherer	3
WHG[4]	Europe (West)	Meso-Neolithic	Indo-European	Hunter-Gatherer	3
Iran Neo[5]	Levant	Neolithic	Indo-European	Early farmer	3
Natufian[5]	Levant & CA	Mesolithic	Unknown	Hunter-Gatherer	6
Steppe_EMBA[6]	Europe (East)/Altai	Early Bronze Age (Afanasievo, Yamnaya...)	Indo-European	Nomadic / pastoral	28
Steppe_MLBA[6]	Europe (East)	Middle Bronze Age (Andronovo, Sintashta...)	Indo-European	Nomadic / pastoral	22
Europe_EN[7]	Europe	Neolithic	Indo-European	Agriculturist	29

Qui sont les Scythes ?

Une question centrale du projet *Steppes* est de trouver l'origine des Scythes et d'en identifier les ascendants européens et asiatiques, en complément des travaux de Unterländer *et al.* (2017). Cette question se heurte à une première difficulté : les Scythes sont présents sur une aire géographique particulièrement vaste et leurs traces les plus anciennes sont retrouvées aussi bien à l'ouest qu'à l'est de l'Asie intérieure. On peut donc légitimement se demander si, en dépit d'une grande homogénéité culturelle, toutes les populations scythes partagent une origine commune.
Le jeu de données *Steppes* inclut quatre groupes géographiques scythes : des Européens des plaines hongroises et des Scythes d'Asie intérieure, à savoir des Sakas du Kazakhstan et du Tian Shan et des Tagar de Sibérie. Les données génomiques autosomales montrent une séparation entre la branche scythe d'Europe et celle d'Asie sur la base d'une ACP et de distances F_{ST} plus élevées entre les populations occidentales et orientales qu'au sein de chaque région (entre 0,24 et 0,3 ; et 0,15 et 0,2, respectivement). Ce résultat conforte l'idée d'une confédération de peuples scythes observée à partir d'ADN mitochondrial par Unterländer *et al.* (2017).

Comme dans les travaux de Unterländer *et al.* (2017), les branches scythes asiatiques auraient été formées par un métissage génétique entre des Européens et Asiatiques : d'après des analyses qpAdm et ACP, la source dite européenne a été rattachée à une population d'Asie intérieure de chasseurs-cueilleurs de l'âge du Bronze tardif, soit celle des Srubnaya du Caucase ou des Andronovo, mais pas des Yamnaya, contrairement à ce qui avait été trouvé (Figure I.5). La source asiatique la plus probable est celle des chasseurs-cueilleurs du lac Baïkal de la culture de Glazkovo. Les analyses menées n'ont pas trouvé de composante asiatique chez Scythes européens.
En outre, pour trois populations scythes, une composante régionale est également détectée : chez les Scythes d'Europe c'est une contribution néolithique européenne qui remplace la composante est-asiatique (Europe_EN), chez les Scythes de Sibérie en sus des composantes chasseurs-cueilleurs européens et asiatiques, s'ajoute une composante sibérienne de type Mal'ta (MA1), et chez ceux du Tian Shan c'est une composante néolithique iranienne (Iran_Neo). Les Scythes du Kazakhstan serait issus des seules populations de chasseurs-cueilleurs européens (56%) et est-asiatiques (44%) avec une contribution est-asiatique serait essentiellement masculine, d'après des estimations basées sur le chromosome X et les autosomes.

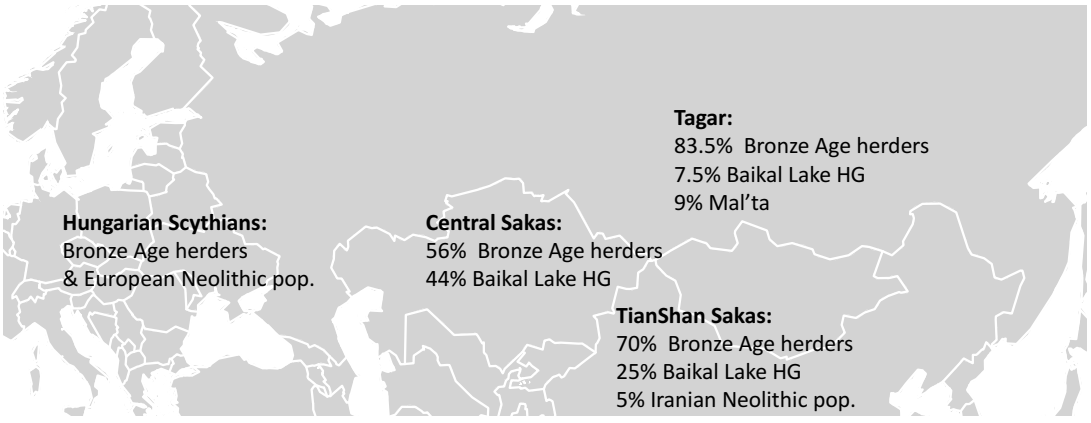


Figure I.5 – Composantes ancestrales inférées par qpAdm chez les quatre groupes scythes étudiés dans les steppes d'Eurasie.

Le projet *Steppes* inclut aussi des génomes sarmates, d'une branche de la culture scythe qui était présente en Ukraine. Bien que voisins et alliés des Scythes européens, les Sarmates diffèrent génétiquement

de ceux-ci d'après des observations réalisées d'après l'ACP, en présentant notamment une contribution est-asiatique absente chez les Scythes européens.

Origine des peuples d'Asie du nord-est de l'Antiquité et du Moyen-Âge

Le projet *Steppes* s'intéresse également à l'origine des peuples d'Asie du nord-est ayant à de nombreuses reprises migré vers l'ouest, au cours de l'Antiquité et du Moyen-Âge, comme le groupe des Xiongnu. Sur la base de cinq génomes, il semblerait que le groupe Xiongnu soit génétiquement hétérogène avec des individus principalement asiatiques et d'autres présentant une contribution occidentale notable dont ils auraient hérité à 95% des Sakas du Kazakhstan.

Les Huns qui leur succédèrent en Mongolie pourraient être leurs descendants, mais les analyses réalisées montrent qu'ils descendraient majoritairement de Sakas du Tian Shan (90%), et plus légèrement des Xiongnu (10%). La composante est-asiatique Xiongnu n'étant pas retrouvée à partir des données du chromosome X : seuls les hommes en auraient été vecteurs.

Plus tard, au Moyen-Âge, des peuples originaires de l'est de l'Asie intérieure, comme les hordes mongoles et en particulier la célèbre Horde d'or menée par le fils ainé de Genghis Khan, se répandirent dans toute l'Eurasie. On retrouve à travers les génomes des steppes asiatiques associés à ces hordes une forte composante est-asiatique associée à une composante européenne en proportions variables, suggérant que ces hordes étaient des confédérations de peuples aux origines variées.

Par le passé, l'Asie intérieure semble avoir connu de nombreux événements de métissages génétiques, impliquant des composantes européennes et asiatiques. En couplant les informations d'ADN ancien du projet *Steppes* avec des données modernes, nous avons essayé d'inférer les proportions des différentes composantes ancestrales au sein des génomes actuels d'Asie intérieure.

Quelles origines pour les groupes ethniques actuels ?

Dans le cadre du projet *Steppes*, j'ai participé à la création d'un jeu de données "modernes" et à son analyse (Figure I.6). Les résultats obtenus sont présentés dans le *Supplementary Materials - Section 4* de la publication produite en Annexe.
Le jeu de données inclut 822 individus dont :

- 320 individus issus de populations européennes, du Caucase et asiatiques publiées par Yunusbayev *et al.* (2015) ;
- 171 individus génotypés *de novo* pour le projet *Steppes*. Ils ont été échantillonnés par l'équipe de Rana Dajani de l'université Hashemite en Jordanie, auprès de communautés tchétchènes et circassiennes vivant en Jordanie ;
- 502 individus échantillonnés par l'équipe d'Évelyne Heyer en Asie centrale et du nord ;

Certaines populations échantillonnées par nos soins ont été combinées à celles de l'étude de Yunusbayev *et al.* quand elles appartenaient au même groupe ethnique, ce qui conduit à un total de 46 populations eurasiennes (voir Table I.2).

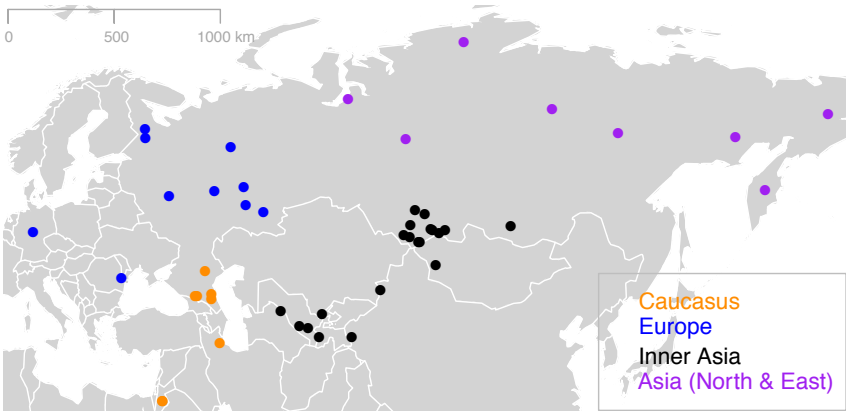


Figure I.6 – Populations modernes incluses dans le jeu de données *Steppes*. En orange les populations du Caucase, en bleu les Européennes, en noir pour celles d'Asie intérieure et en violet pour les Asiatiques d'Asie du nord et de l'est.

Table I.2 – Populations modernes incluses dans le projet *Steppes*. Les astérisques indiquent les populations indo-iraniennes d'Asie intérieure. En gras, les populations d'Asie intérieure incluant plus de 10 individus.

Population	Region	N	Source
Azeris-Dagestan	Caucasus	5	Published data (azd)
Azeris-Iran	Caucasus	18	Published data (az)
Balkars	Caucasus	3	Published data (bl)
Circassian	Caucasus	76	Newly genotyped data
Kabardins	Caucasus	3	Published data (kb)
Kalmyks	Caucasus	14	Published data (kl)
Kumyks	Caucasus	3	Published data (ku)
Tjentjen	Caucasus	95	Newly genotyped data
Bashkirs	Europe	23	Published data (ba)
Chuvashes	Europe	2	Published data (ch)
Gagauzes	Europe	12	Published data (gg)
Germans	Europe	13	Published data (ge)
Karelians	Europe	15	Published data (kr)
Komis	Europe	16	Published data (km)
Russians	Europe	33	Published data (ru)
Tatars	Europe	23	Published data (tt)
Udmurts	Europe	16	Published data (ud)
Vepsas	Europe	11	Published data (vp)
Karakalpaks	Inner Asia (Central)	10	Published data (kk)
Turkmens	Inner Asia (Central)	8 + 27	Published data (tn) + Our data (Tur)
Uygurs	Inner Asia (Central)	1	Published data (ug)
Uzbeks	Inner Asia (Central)	27	Our data (Uzb)
Tajiks*	Inner Asia (Central)	4 + 27	Published data (tj) + Our data (Tab)
Tajiks-Pamir*	Inner Asia (Central)	29	Published data (tjk)
Yaghnobi*	Inner Asia (Central)	6	Published data (yg)
Altaï-Kizhis	Inner Asia (Siberia)	21	Our data (Aki)
Altaïans	Inner Asia (Siberia)	2	Published data (al)
Buryats	Inner Asia (Siberia)	3 + 23	Published data (br) + Our data (Bou)
Kazakhs	Inner Asia (Siberia)	2 + 27	Published data (kz) + Our data (Akz)
Khakas	Inner Asia (Siberia)	24	Our data (Hks)
Kyrgyz	Inner Asia (Siberia)	9 + 23	Published data (ki) + Our data (Kek & Kem)
Mongush	Inner Asia (Siberia)	14	Our data (Mog)
Mongolians	Inner Asia (Siberia)	2 + 31	Published data (mo) + Our data (Mng)
Ondar	Inner Asia (Siberia)	2	Our data (Ond)
Shores	Inner Asia (Siberia)	28	Our data (Sho)
Telengits	Inner Asia (Siberia)	28	Our data (Tlg)
Tubalars	Inner Asia (Siberia)	29	Our data (Tub)
Tuvans	Inner Asia (Siberia)	3	Published data (tv)
Chukchis	Asia (North & East)	2	Published data (chk)
Evenks	Asia (North & East)	3	Published data (ev)
Evens	Asia (North & East)	3	Published data (en)
Kets	Asia (North & East)	2	Published data (kt)
Koryaks	Asia (North & East)	2	Published data (ko)
Nenets	Asia (North & East)	15	Published data (nnf & nnt)
Nganasans	Asia (North & East)	2	Published data (ng)
Yakuts	Asia (North & East)	2	Published data (ya)

À partir de ces données, nous avons complété le panorama génétique de l'Eurasie actuelle, au moyen d'une Analyse en Composante Principale en incluant à la fois des populations européennes, du Caucase et d'Asie (Figure I.7) dans la continuité des travaux réalisés par Di Cristofaro *et al.* (2013) qui s'intéressaient particulièrement à des populations d'Afghanistan.
Les deux dimensions de l'Analyse en Composantes Principales considérées révèlent quatre groupes principaux de populations : les clusters "Européen", "Caucasien", "Asiatique (intérieure et de l'est)" et "nord-Asiatique", incluant principalement des populations originaires de la région éponyme. Les Tchéchènes et Circassiens ne sont pas représentés sur cette figure, étant trop différents génétiquement des autres populations occidentales.

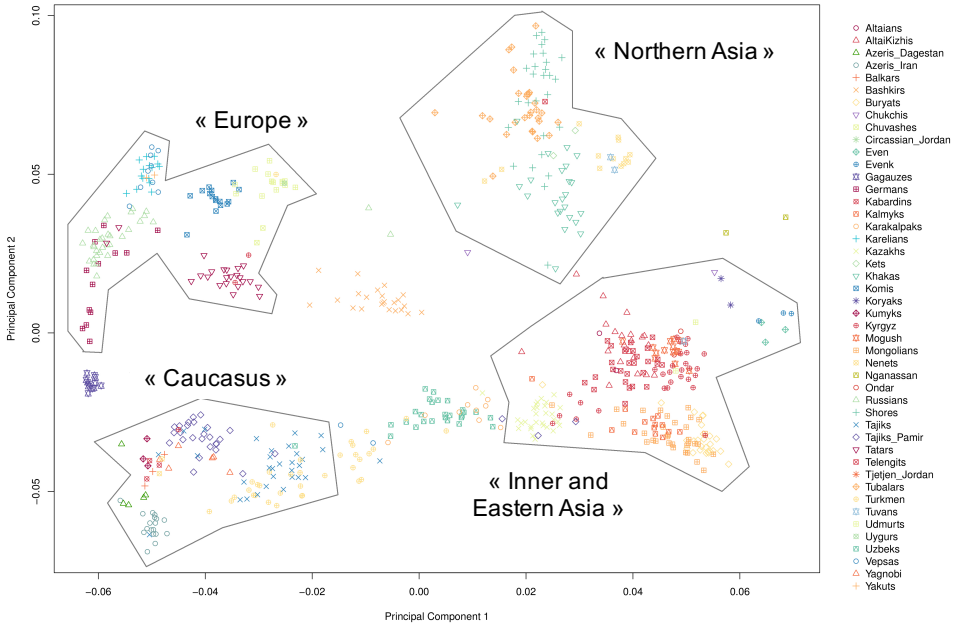


Figure I.7 – Analyse en Composantes Principales à partir des génotypes inclus dans le jeu de données moderne du projet *Steppes*. Les Tchéchènes et Circassiens sont exclus car semblaient trop différents des autres populations européennes et du Caucase. Le premier axe explique 3,5% de la variation génétique et le deuxième axe 0,5%.

Table I.3 – Populations incluses dans les clusters de l'ACP. En bleu les populations européennes, en orange du Caucase, en violet d'Asie du nord et de l'est, en noir d'Asie intérieure, avec en particulier celles de langue indo-iranienne indiquées par *.

"Europe"	**"Caucase"**	**"Asie du nord"**	**"Asie intérieure et de l'est"**	Hors cluster
Vepsas	Tajiks-Pamir*	Shores	AltaiKizhis	Uzbeks
Karelians	Tajiks*	Tubalars	Telengits	Nganassan
Russians	Kumyks	Kets	Kyrgyz	Gagauzes
Germans	Azeris-Dagestan	Nenets	Ondar	Bashkirs
Komis	Yagnobi*	Khakas	Buryats	Chukchis
Udmurts	Kabardins	Tuvans	Mongolians	Koryaks
Chuvashes	Azeris-Iran		Kazakhs	Karakalpaks
Tatars	Turkmen		Even	
Yakuts	Balkars		Evenks	
Uygur			Kalmyks	
			Mogush	
			Altaians	

La première dimension, expliquant 3,5% de la variance, sépare d'un côté les clusters européen et caucasien, et de l'autre les deux clusters asiatiques. Cette dimension nous permet de distinguer les populations d'Asie intérieure : les Indo-Iraniens (représentés ici par les *Tajiks, Tajiks-Pamir* et *Yagnobi*), ainsi que les Turco-Mongols *Turkmens*, font partie du cluster du Caucase ; les autres Turco-Mongols d'Asie centrale, à savoir les *Uzbeks* et les *Karakalpaks*, sont entre le cluster d'Asie intérieure et de l'est et celui du Caucase ; les populations turco-mongoles de Sibérie et Mongolie se situent dans les clusters d'Asie interieure et de l'est ou du nord. Le cluster nord-asiatique inclut précisément les *Shores, Khakhas, Tubalars* et *Tuvans*. Ces Turco-Mongols de Sibérie ont tendance à avoir un mode de subsistance de type chasseur-pêcheur, et sont génétiquement proches de populations nord-sibériennes, comme les *Nenets* et les *Kets*, deux groupes d'éleveurs de rennes du cercle arctique. Cependant et en dépit d'une proximité géographique, les autres populations turco-mongoles de Sibérie sont présentes dans un cluster différent, celui d'Asie intérieure et de l'est, ce qui suggère que les populations originaires de Sibérie sont génétiquement hétérogènes.

Des D-statistiques ont permis d'associer la variance génétique observée sur la première dimension de l'ACP à un excès de composante de type est-asiatique chez les populations des clusters asiatiques par rapport aux populations des clusters européen et caucasien. Pour la seconde dimension, ce serait un excès de composante nord-eurasienne chez les Européens et Asiatiques du nord. Ces composantes sont respectivement représentées dans les tests par le génome BHG d'un chasseur-cueilleur du lac Baïkal de la culture Glazkovo et par le génome paléolithique de Mal'ta MA1 (Raghavan *et al.* 2014), dont la composante est présente chez de nombreuses populations du nord de l'Eurasie.

Nous avons ensuite modélisé les composantes ancestrales des populations actuelles à partir de trois populations sources (Figure I.8) choisies car elles étaient les populations les plus extrêmes sur l'ACP réalisée sur les données anciennes et modernes (Figure 2 de l'article). Il s'agit de :

- la population mésolithique natufienne du Proche-Orient, datant d'environ 12 000 ans AEC, (Lazaridis *et al.* 2016) ;
- la population apparentée aux chasseurs-cueilleurs d'Europe de l'est, datant du Méso-Néolithique autour de 6 000 ans AEC et appelée *Eastern Hunter-Gatherers* ou EHG (Mathieson *et al.* 2015) ;
- la population de chasseurs-cueilleurs du lac Baïkal, d'Eurasie du nord-est, produite *de novo*.

Pour les populations modernes les plus occidentales, une autre composante d'Europe de l'ouest a dû être ajoutée : celle des chasseurs-cueilleurs mésolithiques européens de la Brana, Loschbour et Motala, appelée *Western Hunter-Gatherers* ou WHG (Olalde *et al.* 2014 ; Lazaridis *et al.* 2014). Pour la population actuelle chukchi de l'extrême est de l'Eurasie, une composante paléo-américaine de la culture Clovis est requise (Rasmussen *et al.* 2014). Ces populations ne sont pas nécessairement celles à la source des groupes ethniques actuels mais permettent d'estimer des grandes tendances à l'origine des fonds génétiques d'Asie intérieure, à savoir européennes de l'est ou de l'ouest, caucasiennes ou sibériennes.

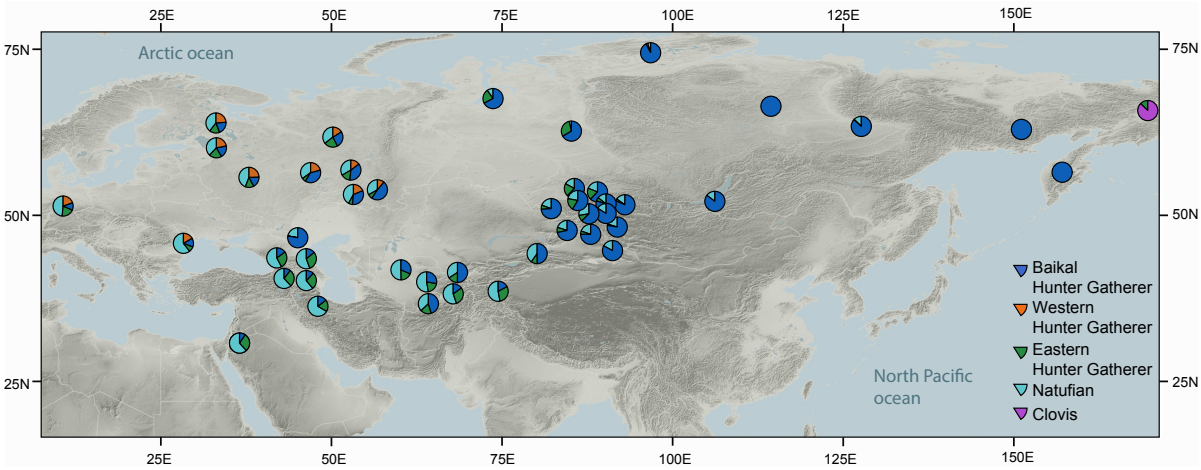


Figure I.8 – Représentation des composantes ancestrales retrouvées chez les populations modernes du projet "Steppes"

Ainsi nous avons trouvé que la composante asiatique de chasseurs-cueilleurs du lac Baïkal est très marquée chez les populations modernes vivant en Asie (67%) mais moins chez celles résidant en Europe et dans le Caucase (27% et 20%), et qu'elle augmente avec la longitude (ρ de Spearman = 0,88 ; p-value=$1 * 10^{-14}$). Au contraire, celles des chasseurs-cueilleurs d'Europe de l'est et des Natufiens diminuent avec la longitude (ρ de Spearman = -0,82 et -0,54 ; p-values<10^{-3}). Au total, ces deux composantes contribuent à 80% pour les populations du Caucase, 55% pour celles d'Europe, 35% pour celles d'Asie intérieure. La composante de chasseurs-cueilleurs d'Europe de l'ouest n'est détectée que chez les populations européennes (19%), ce qui les distingue des populations du Caucase. Celles-ci ont en contrepartie un excès de composante natufienne par rapport aux Européennes (56% contre 41%).

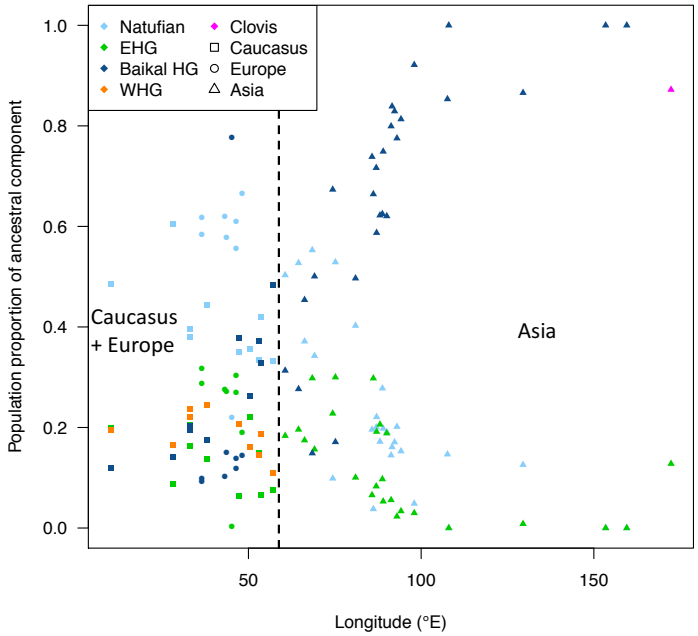


Figure I.9 – Proportion par population des différentes composantes ancestrales en fonction de la longitude. Les populations d'Europe sont représentées par des cercles, celles du Caucase par les carrés et celles d'Asie par des triangles. Les couleurs correspondent aux différentes composantes ancestrales, et chaque population apparaît donc plusieurs fois sur le graphique (une fois pour chacune de ses composantes).

Dans cette thèse, j'ai voulu analyser plus en détail les différences de composantes ancestrales retrouvées chez les Indo-Iraniens et Turco-Mongols d'Asie intérieure. Pour cela, j'ai retenu 15 populations modernes du projet *Steppes* dont les effectifs étaient supérieurs à 10 individus (en gras dans la Table I.2) :

- deux populations indo-iraniennes d'Asie centrale ;
- trois populations turco-mongoles d'Asie centrale ;
- dix populations turco-mongoles de Sibérie et Mongolie.

La modélisation de ces populations ne requiert que trois composantes ancestrales : natufienne, européenne de l'est (EHG) et asiatique du lac Baïkal. Nous avons observé des proportions de ces composantes ancestrales différentes selon les groupes géographiques et linguistiques, sans que les observations ne soient étayées par des tests statistiques significatifs, du fait du faible nombre de populations retenues. Nous avons trouvé que la composante du lac Baïkal est plus forte chez les Turco-Mongols que chez les Indo-Iraniens, alors que les composantes natufienne et européenne de l'est sont plus présentes chez les Indo-Iraniens (Figure I.10). Nous avons également observé une tendance pour les Turco-Mongols d'Asie du nord à avoir un profil plus asiatique que les Turco-Mongols d'Asie centrale, avec une plus forte composante asiatique et moins de composantes occidentales natufienne et EHG. Les deux groupes ethniques mongols, *Buryats* et *Mongolians*, étudiés ici ont la particularité de ne pas présenter de composante EHG.

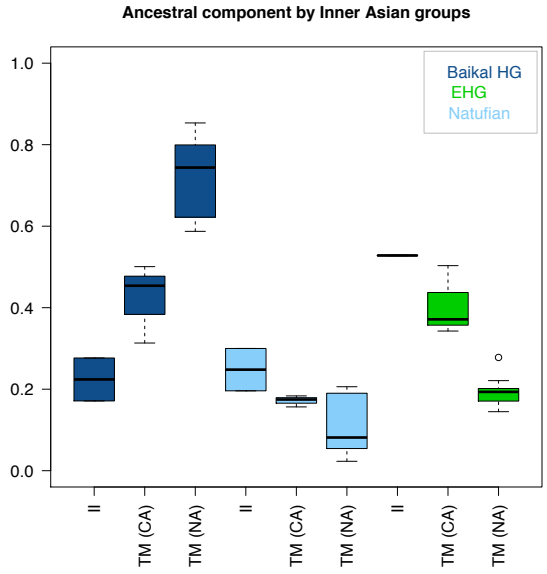


Figure I.10 – Proportion de trois composantes ancestrales retrouvées chez des populations modernes d'Asie intérieure. II : Indo-Iraniens d'Asie centrale ; TM (CA) : Turco-Mongols d'Asie centrale ; TM (NA) : Turco-Mongols d'Asie du nord. En bleu marine, la contribution du lac Baïkal, en bleu clair celle des Natufiens, en vert celle des Européens de l'est (EHG).

Ces derniers résultats sont particulièrement intéressants pour expliquer l'origine de la diversité génétique actuelle observée en Asie intérieure. En effet, ceux-ci montrent que les différences génétiques entre les groupes turco-mongol et indo-iranien sont en partie causées par des fonds génétiques différents, et non uniquement sous l'effet de la dérive génétique à partir d'un même patrimoine génétique.
Ainsi, ces groupes culturels et génétiques auraient été formés par des ancêtres différents et, d'après des inférences sur les données modernes, lors de différents épisodes du peuplement de la région (Palstra *et al.* 2015). Certaines des composantes ancestrales à l'origine des populations d'Asie intérieure ont aussi contribué au fonds génétique des populations d'Europe et d'Asie de l'est, ce qui expliquerait les proximités génétiques observées en Eurasie à partir des données modernes.

I.3 Discussion

Ce chapitre de ma thèse illustre la complémentarité des analyses combinant ADN ancien et moderne dans l'inférence de l'histoir du peuplement d'une région, dans le passé et jusqu'à l'époque actuelle.
Pour ce faire, l'ADN moderne est évidemment nécessaire permettant d'accéder à la diversité génétique actuelle mais aussi de faire des inférences sur l'histoire démographique des populations (Chaix *et al.* 2008 ; Aimé *et al.* 2015 ; Palstra *et al.* 2015). D'autre part, la paléogénétique permet de documenter la diversité génétique à un instant donné, l'évolution des contributions génétiques entre deux périodes et d'estimer des contributions ancestrales au fonds génétique actuel et donc de reconstruire des pans de l'histoire démographique des populations.

Bien que très utile et fascinant, l'ADN ancien présente un certain nombre de limites dont il faut avoir conscience au moment d'interpréter les observations. Notamment, les données anciennes auxquelles nous avons accès ne peuvent représenter qu'une part de la population ancestrale car elles ne sont produites que pour les fossiles qui ont été retrouvés et dont l'ADN a été préservé. Cela nous pousse donc à relativiser la "discontinuité" génétique parfois observée entre des populations modernes et anciennes souvent représentées par quelques individus (par exemple celle observée entre des populations de l'âge du Fer au Kazakhstan (Lalueza-Fox *et al.* 2004)).
De plus, en ADN ancien, l'attribution des fossiles et génomes à une culture peut être problématique : les individus génotypés dans le projet *Steppes* et associés à la Horde d'or présentent deux profils génétiques différents, l'un principalement asiatique, et l'autre avec une forte composante européenne. Ce signal pourrait indiquer une coalition d'individus au sein de la horde, mais pourrait aussi évoquer des esclaves, satellites à la horde sans y être intégrés. Les données d'ADN ancien nécessitent donc un travail détaillé d'archéologie afin de se prémunir d'interprétations n'ayant pas de sens en archéologie. Ce type de difficultés est également retrouvé en génétique des populations modernes : la collecte d'échantillons doit être associée à un travail ethnologique rigoureux, afin de savoir à quelle structure sociale sont associés les individus échantillonnés et ce dont ils sont représentatifs (village, clan, tribu...) et ainsi de pouvoir restreindre les interprétations réalisées aux seules plausibles dans le contexte ethnologique étudié.

Dans ce projet *Steppes*, nous avons pu mettre en évidence des différences de composantes ancestrales chez les populations d'Asie intérieure à partir de données *genome-wide*. Cela avait été précédemment indiqué par Unterländer *et al.* (2017) qui modélisaient les Indo-Iraniens comme des descendants des Scythes européens et les Turco-Mongols des Scythes asiatiques à partir de données mitochondriales.
Cependant, les méthodes utilisées dans le projet *Steppes* ne nous permettent pas d'estimer des composantes ancestrales plus récentes que celles datant du Méso-Néolithique. En effet, ces méthodes (qpAdm, D-statistiques, ACP ou ADMIXTURE) sont fondées sur l'analyse de fréquences alléliques et donc sensibles aux événements de métissage génétique : elles "saturent" lors d'événements de métissages récurrents (Verdu et Rosenberg 2011). Or, depuis les Scythes de l'âge du Fer issus d'un métissage entre des sources exclusivement européenne et asiatique, toutes les populations d'Asie intérieure sont un mélange de populations déjà porteuses de ces deux composantes. Ces méthodes, en recourant à des populations distinctes génétiquement, permettent donc d'estimer des composantes assez vagues (d'Europe de l'est ou de l'ouest, du Caucase...) mais ne sont pas suffisantes pour inférer des sources plus récentes et plus précises qui sont elles-mêmes mélangées sur un plan génétique. Ces méthodes sont tout de même très utiles pour décrire des fluctuations des composantes asiatique ou européenne entre deux périodes ou entre deux populations (par exemple, la comparaison Xiongnu *versus* Huns), pour inférer si une population était structurée rassemblant des individus plutôt asiatiques et d'autres plutôt européens, comme cela a été fait pour la

Horde d'or ou les Xiongnu, et même dans l'absolu pour détecter l'introgression d'autres composantes dans les génomes, par exemple sud-asiatiques.

Pour tenter de reconstruire de manière plus satisfaisante l'origine des populations de cette région, on pourrait se fonder sur des inférences bayésiennes directement à partir de données modernes comme cela a été fait par Palstra *et al.* (2015) pour rechercher le scénario le plus compatible possible avec la diversité génétique moderne observée. Ces méthodes sont très intéressantes mais sont néanmoins limitées par le choix du scénario parmi ceux testés et celui des populations modernes retenues pour mimer les populations sources. De plus, elles requièrent des temps de calcul informatique considérables.
À l'heure actuelle, ces inférences n'ont été réalisées que pour deux populations vivant actuellement en Asie intérieure et représentant deux groupes ethniques, les Tadjiks et Kirghizes. Il serait intéressant de réaliser ces inférences pour les autres groupes ethniques mais aussi pour plusieurs populations par groupe ethnique afin d'accéder à une histoire démographique complète de chaque groupe.

Enfin, nous souhaitons poursuivre l'analyse des relations d'ascendance entre les populations actuelles d'Asie intérieure et des populations anciennes déjà métissées mais plus récentes comme celles de l'âge du Fer ou Moyen-Âge. Une manière de procéder serait d'avoir recours à une méthode empirique fondée sur la base de comparaison de fragments IBD partagés entre les génomes, par exemple au moyen du logiciel ALDER (Loh *et al.* 2013) comme fait par Yunusbayev *et al.* (2015).

Chapitre II

Diversité génétique et comportements culturels asymétriques entre hommes et femmes

Sommaire

II.1 Comportements culturels asymétriques entre sexes

L'organisation et le fonctionnement des sociétés humaines reposent, en partie, sur une dualité et une complémentarité des sexes (Héritier 1996), dont découlent des fonctions et comportements asymétriques entre hommes et femmes (Godelier 2004). Alors que ces distinctions étaient traditionnellement justifiées par la "nature", on leur reconnaît désormais une dimension culturelle (Bereni *et al.* 2012).
Ces comportements culturels asymétriques entre sexes, auxquels nous nous référons sous le sigle CCAS, ont été étudiés par d'innombrables travaux en ethnologie et plus récemment en génétique. En effet, comme de nombreux autres comportements culturels les CCAS façonnent la diversité génétique humaine (Creanza et Feldman 2016) et en particulier, ils sont susceptibles de laisser une empreinte sur la diversité génétique sexe-spécifique que nous explicitions dans cette première partie.

Polygamie, temps de génération et mortalité asymétriques

La polygamie, définie comme l'union d'un individu avec plusieurs personnes du sexe opposé, est courante dans notre espèce : 83% des sociétés sont polygames contre 17% de monogames (Marlowe 2000 ; Murdock 1967). Dans la plupart des cas, la polygamie est en fait de la polygynie, c'est-à-dire qu'un homme épouse plusieurs femmes, tandis que la polyandrie n'est pratiquée que dans 1% des sociétés humaines. On distingue les sociétés légèrement polygynes, où seuls certains hommes "exceptionnels" accèdent à la polygynie (52% des sociétés humaines), des sociétés largement polygynes dans lesquelles la plupart des hommes accèdent à la polygynie (31% de nos sociétés).
Si le sex-ratio est équilibré, la polygynie entraîne nécessairement une variance du succès reproducteur masculin : certains hommes ont plusieurs femmes et donc plus d'enfants que ceux qui ne se marient pas et n'ont (en théorie) pas de descendants. En parallèle, la quasi-totalité des femmes se marie, conduisant à une moindre variance reproductrice féminine parfois sous l'influence légère de la polygynie : dans certaines sociétés polygynes, les femmes des couples monogames peuvent être 1,5 plus fertiles que les femmes de couples polygames (Heyer *et al.* 2012). De manière générale, la polygynie agit plus fortement sur la variance reproductive des hommes que des femmes, comme chez les Pygmées Aka faiblement polygynes où les variances sont de 5,20 chez les femmes et 8,64 chez les hommes (Hewlett 1987) (Figure II.1). Cette asymétrie de variance reproductrice entraîne une diminution de la diversité génétique et de la taille efficace masculine par rapport à celles des femmes, mais les différences sont faibles entre les sexes, d'un facteur 2 pour la taille efficace dans des cas extrêmes de polygynie (Heyer *et al.* 2012).

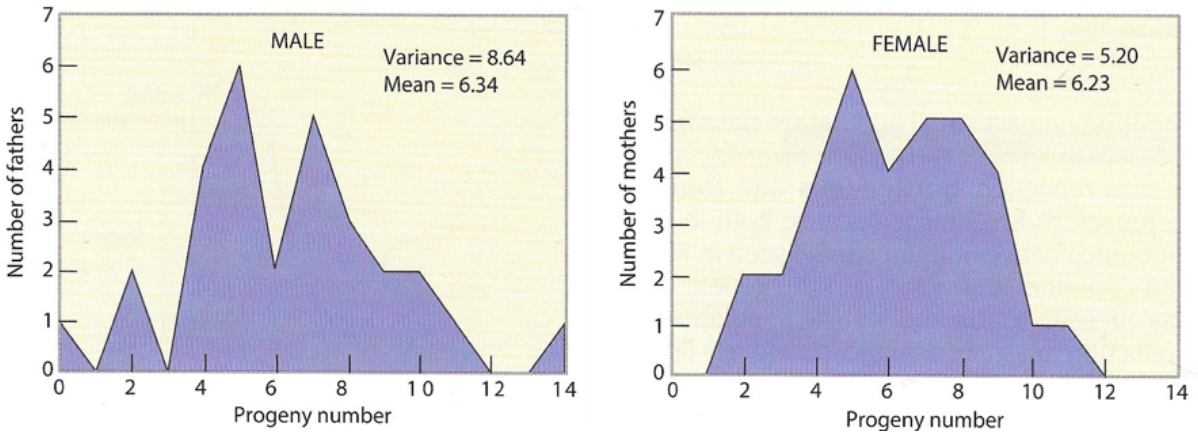


Figure II.1 – Succès reproducteur masculin et féminin chez les Pygmées Aka, dont environ 15% des unions sont polygynes. *Figure 5.9 de Jobling* et al. *(2013), à partir de Hewlett (1987).*

Dans les populations en croissance, les hommes ont tendance à épouser des femmes plus jeunes, car plus nombreuses, permettant au grand nombre d'être polygyne (Pison 1986). Ce type d'unions cause un écart de temps de génération entre les sexes, c'est-à-dire l'âge moyen auquel les hommes et les femmes ont des enfants. Dans la plupart des sociétés humaines, polygynes ou non, le temps de génération est plus court pour les femmes que pour les hommes : l'écart est généralement d'environ 10%, avec des âges de 28 et 30 ans respectivement en Islande (Helgason *et al.* 2003), 29,5 et 33,9 ans au Québec (Tremblay et Vézina 2000), ou 27,2 et 40,2 ans chez les Peuls du Sénégal (Pison 1986).
Ainsi, le nombre de générations féminines survenues pendant un laps de temps est généralement plus élevé que le nombre de générations masculines. Par dérive, un certain nombre d'haplotypes est perdu à chaque génération et la diversité mitochondriale diminue plus fortement que celle du chromosome Y (Helgason *et al.* 2003). Mais d'autre part, l'ADN mitochondrial est susceptible de présenter plus de mutations que le chromosome Y, survenues au moment des divisions cellulaires plus nombreuses, et donc une diversité génétique plus grande. De manière générale, la dérive et les mutations ont des effets opposés sur la diversité génétique, et leur prévalence respective n'est pas clairement établie dans ce contexte de temps de génération asymétriques.

La polygamie peut aussi exister de manière sérielle : un individu peut avoir au cours de sa vie plusieurs époux/ses mais pas de façon contemporaine, comme dans le cas d'un remariage après un veuvage ou un divorce. En effet, sous le prisme de la biologie, ces remariages correspondent à une forme de polygamie s'ils engendrent des enfants. En particulier, ces secondes noces dont naissent des enfants sont généralement le fait d'hommes remariés à des femmes plus jeunes, ce qui conduit aussi à un écart du temps de génération.

La polygynie, officielle ou officieuse, peut aussi survenir lorsque le sex-ratio est déséquilibré et que les femmes en âge de se reproduire sont plus nombreuses que les hommes. Cette situation survient notamment lorsque la mortalité masculine est plus forte que la mortalité des femmes, par exemple lors de guerres. En effet, la guerre agit drastiquement sur la mortalité des hommes jeunes (à la fin de la première guerre mondiale, seule la moitié des hommes nés 25 ans plus tôt en France était encore en vie (Héran 2014)), mais moins sur celle des femmes généralement moins exposées. Pour illustration, en Nouvelle-Guinée, au cours de la seconde moitié du XX^{e} siècle, lors des fréquents affrontements entre des sociétés papoues, il était permis de tuer ses ennemis masculins mais il était immoral de tuer des femmes (Kayser *et al.* 2003). Dans la communauté Dani, 28% des hommes auraient été tués contre 2% des femmes.
De nos jours en Europe, d'autres facteurs créent des asymétries de mortalité et sont responsables d'une espérance de vie masculine plus courte que celle des femmes avec un écart de 7 ans en France en 2003 et de 13 ans en Russie en 2000 (Meslé 2004). Cet écart se creuse dès l'âge de 15 ans notamment du fait de morts violentes plus fréquentes chez les hommes jeunes que chez les femmes et affecte le sex-ratio pendant la période reproductive.
Du point de vue génétique, cette asymétrie de mortalité cause une diminution de l'effectif masculin à la génération concernée. Sans polygynie, la même réduction d'effectif reproducteur est alors induite chez les femmes, puisque certaines femmes ne peuvent pas accéder à la reproduction faute d'hommes disponibles et vivants en âge de se reproduire.

Transmission du succès reproducteur

Comme nous venons de le souligner avec la polygynie, des différences de succès reproducteur peuvent exister entre individus de même sexe. Ces différences reproductrices peuvent être transmises d'une génération à la suivante : le nombre de descendants d'un individu est parfois corrélé au nombre de descendants de ses parents - soit à la taille de sa fratrie - (Austerlitz et Heyer 1998 ; Pluzhnikov *et al.* 2007).
Cela s'explique par la transmission, d'une génération à l'autre, de différents facteurs responsables de la variance du succès reproducteur :

- biologiques conditionnant le succès reproducteur en agissant sur la survie, l'accès à la reproduction et/ou la production de descendants pour les parents et leurs descendants (Heyer et Cazes 1999) ;
- culturels agissant sur la fitness des parents et de leurs descendants comme la possession matérielle, la position sociale, la présence dans la population d'apparentés avec qui coopérer (Gagnon et Heyer 2001).

Du point de vue génétique, la transmission du succès reproducteur entraîne une réduction de la taille efficace de la population : au Québec, dans la population de Saguenay-Lac-St-Jean, la transmission du succès reproducteur aurait entraîné une réduction de la taille efficace d'un facteur 10 à 20 par rapport au modèle classique de Wright-Fisher, ce qui aurait causé une augmentation en fréquence de certaines maladies héréditaires (Austerlitz et Heyer 1998 ; Heyer *et al.* 2005 ; Sibert *et al.* 2002).

La transmission des facteurs, biologiques ou culturels, peut être influencée par le sexe du descendant. En effet, certains allèles sont bénéfiques aux descendants mâles et peuvent nuire au succès reproducteur de leurs sœurs et inversement (Innocenti et Morrow 2010) ; ou encore, l'héritage culturel ou matériel transmis aux fils et aux filles n'est pas nécessairement identique, ce qui peut favoriser le succès reproducteur des descendants d'un sexe en particulier. Ainsi, une transmission plus marquée en direction des fils a été observée chez les Schmiedeleut Huttérites (Pluzhnikov *et al.* 2007) et en Asie centrale (Heyer *et al.* 2015). On observe aussi des différences d'intensité de cette transmission entre les populations humaines : la transmission en direction des filles est plus marquée chez les chasseurs-cueilleurs africains que chez leurs voisins agriculteurs (Blum *et al.* 2006).
La conséquence génétique de cette asymétrie de transmission est une diversité génétique et une taille efficace réduites pour le sexe qui présente la plus forte transmission.

Organisation sociale

Les sociétés humaines s'organisent autour d'une dualité homme/femme qui se retrouve au sein même de la structure de la société en instances : l'organisation sociale est définie par un ensemble de règles d'alliance, de résidence et de filiation souvent asymétriques entre les sexes.

Règles d'alliance

Pour le courant structuraliste et notamment Lévi-Strauss, le mariage est un pont entre des groupes d'individus et l'échange de conjoints est un facteur essentiel de l'organisation sociale, un ciment définissant un système d'alliance, de coopération et d'entraide (Lévi-Strauss 1949 ; Ghasarian 1996).
Afin de réaliser des alliances profitables, des règles prescrivant ou proscrivant certains types d'unions sont édictées, reposant notamment sur les notions d'endogamie, qui consiste en un mariage au sein de son groupe (géographique, social, religieux, linguistique...), et d'exogamie, qui est le fait de se marier à l'extérieur de son groupe. Ainsi, ces notions permettent de préciser les groupes d'individus avec lesquels on peut, ou ne peut, pas se marier. Par exemple, chez les populations indo-iraniennes d'Asie intérieure, la

règle d'alliance est l'endogamie de village ; au contraire, pour les Turco-Mongols, les mariages se font de manière exogame hors du clan et du lignage, ce qui correspond généralement à de l'exogamie de village (Krader 1966).

Les règles d'alliance débouchent souvent sur des asymétries entre sexes, notamment en autorisant des échanges entre groupes pour un sexe mais pas pour l'autre, comme pour la règle d'hypergamie dans le système indien de castes qui autorise les femmes à se marier hors de leur caste, dans des castes plus hautes, mais l'interdit aux hommes. Du point de vue génétique, on s'attend à observer une homogénéité des castes à travers les ADN des femmes, mais une structuration de l'ADN des hommes par caste (Bamshad *et al.* 1998).

Le système d'alliance introduit des interdictions matrimoniales mais des barrières peuvent également s'ériger sans avoir été formellement édictées, découlant de facilités géographiques ou sociales, comme par exemple la barrière linguistique (Cavalli-Sforza 1997) ou ethniques (Barth 1969 ; Henrich et Boyd 1998). On peut alors imaginer que ces barrières sont plus poreuses pour un sexe que pour l'autre, que les individus de ce sexe changent plus facilement de groupe, ce qui conduit à l'homogénéisation génétique des groupes pour le marqueur uniparental associé au sexe migrant.

Règle de résidence

Vikings, conquistadors, hordes mongoles... Ces quelques exemples historiques illustrent des cas de migrations ponctuelles exclusivement masculines. Cependant, loin de l'image d'Épinal de l'homme conquérant, la grande majorité des migrations asymétriques correspondent à des migrations de femmes, de manière moins spectaculaire, à des échelles géographiques réduites, mais survenant de manière récurrente (Stoneking 1998). Ces migrations sont principalement motivées par le mariage et sont régies par des règles de résidence qui définissent le lieu d'habitation d'un couple nouvellement formé. Elles précisent ainsi quel est le sexe migrant et quel est le sexe philopatrique dans le cadre d'unions exogames (Burton *et al.* 1996 ; Murdock 1981) :

- dans le cadre d'une résidence patrilocale, la femme migre en direction du lieu d'origine de son époux (70% des populations humaines) ;
- dans le cadre d'une résidence matrilocale, l'homme migre en direction du lieu d'origine de son épouse (20% des populations humaines) ;
- dans un système de néolocalité, les deux conjoints migrent vers un nouveau lieu qu'ils ont choisi comme lieu de résidence ;
- dans un système de multilocalité, les époux oscillent entre des résidences alternées (10% des populations humaines sont néo ou multilocales).

Du point de vue génétique, la circulation des femmes entre des populations patrilocales devrait causer une différenciation des populations plus marquée pour les marqueurs sexe-spécifiques masculins que féminins. Dans le cas de populations matrilocales où ce sont les hommes qui circulent, on s'attend à observer le schéma inverse.

Règles de filiation

Les règles de filiation régissent la transmission de la parenté sociale et définissent le groupe de parenté auquel appartient chaque individu, ainsi que sa relation aux autres. La transmission de la parenté peut se faire (Godelier 2004 ; Ghasarian 1996) (Figure II.2) :

- unilatéralement, par le père (filiation patrilinéaire, dans 45% des sociétés humaines) ou par la mère (filiation matrilinéaire, dans 12% des sociétés humaines). L'individu appartient asymétriquement

au groupe de parenté issu de la branche paternelle ou maternelle de sa généalogie ;

- de manière indifférenciée (cognatisme ou bilinéarité), l'individu reconnaît l'ensemble des branches de sa généalogie comme ses parents sociaux (41% des sociétés humaines). Dans le cas de la bilinéarité, pour quelques traits culturels, l'individu est considéré comme l'héritier de son père ou de sa mère.

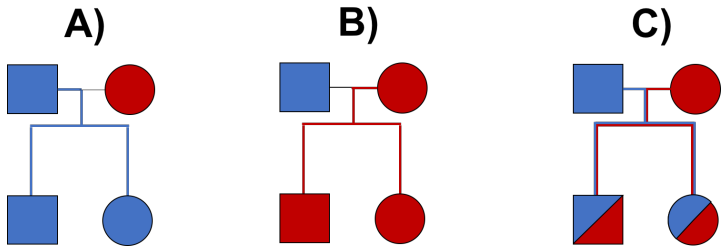


Figure II.2 – Différentes règles de filiation : patrilinéarité (A), matrilinéarité (B), indifférenciée (bilinéarité et cognatisme) (C). Chaque couleur représente un groupe de filiation.

Les règles de filiation dans les sociétés unilinéaires entraînent la formation à différents niveaux hiérarchiques de groupes de filiation appelés lignages, clans et tribus. Chaque niveau hiérarchique annonce descendre d'un ancêtre commun social. Une étude menée en Asie centrale sur des populations patrilinéaires a révélé une filiation biologique superposée à la filiation sociale à l'échelle du lignage et du clan (Chaix *et al.* 2004). Les auteurs montrent que les hommes issus des mêmes lignages et des mêmes clans sont plus apparentés par voie paternelle que des hommes issus de groupes différents, et que l'ancêtre commun aux individus des lignages aurait vécu il y a entre 14 et 20 générations et celui des clans il y a 20 à 50 générations. Par contre, chez ces populations, l'apparentement entre hommes originaires de la même tribu ou de tribus différentes est équivalent ; l'ancêtre de la tribu est donc social mais non biologique.

Du point de vue génétique, des dynamiques internes aux groupes de filiation, ici patrilinéaires, seraient responsables d'une réduction et d'une structuration de la diversité génétique masculine (Chaix *et al.* 2007) :

- quand un groupe devient trop conséquent, il se sépare en sous-groupes, qui ne sont pas constitués au hasard : les hommes les plus apparentés demeurent ensemble, ce qui réduit la variation des chromosomes Y dans les sous-groupes par rapport au groupe ancestral. Ce phénomène de segmentation lignagère a été notamment documenté ethnologiquement chez les Yanomama en Amérique du Sud (Smouse *et al.* 1981) ;
- du fait du fort apparentement par la branche paternelle, la taille efficace masculine de ces groupes devrait être faible, ce qui augmenterait la dérive et la fixation de certains haplotypes du chromosome Y dans la population au détriment d'autres, et réduirait la diversité des haplotypes observés.

Théoriquement, on devrait observer le schéma inverse pour la diversité féminine dans des populations matrilinéaires.

II.2 Outils de détection des différences génétiques sexe-spécifiques

Les CCAS présentés dans la première partie de ce chapitre sont susceptibles de générer des différences génétiques sexe-spécifiques dont l'observation se fonde généralement sur des marqueurs uniparentaux, permettant un accès facile aux lignées féminines et masculines. Bien que leur utilisation soit moins intuitive, le chromosome X et les autosomes permettent également d'étudier des CCAS du point de vue génétique.

Marqueurs uniparentaux

Chromosome Y

Le chromosome Y, d'une taille de 57 Mb environ, est l'un des plus petits chromosomes humains et celui présentant la plus faible densité génique : il ne contient que 78 gènes codant pour des protéines, soit une densité de 1,3 gènes/Mb (Skaletsky *et al.* 2003) contre 7,1 sur le chromosome X (Ross et al. 2005). Il est composé à 5% de régions pseudo-autosomales télomériques recombinant avec le chromosome X, et à 95% d'une région non-recombinante, appelée NRY pour *non-recombining region of the Y chromosome*. Cette région est transmise à l'identique d'une génération à l'autre, de père en fils, aux seules mutations *de novo* près, ce qui permet de retrouver facilement des liens généalogiques entre hommes, mais aussi de retracer l'histoire des lignées masculines (Jobling et Tyler-Smith 2003).

ADN mitochondrial

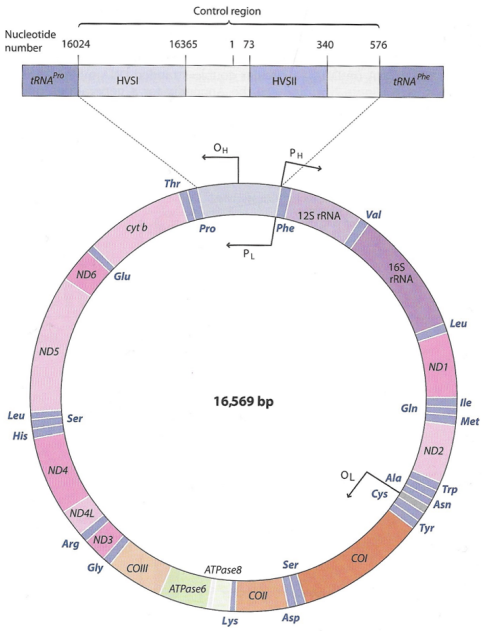


Figure II.3 – Structure de l'ADN mitochondrial. En détail, la région non-codante dite de contrôle, contenant les séquences hypervariables HVS1 et HVS2. *Figure 2.21 du livre de Jobling* et al. *(2013)*.

Cette molécule circulaire de 16,5 kb est présente en grand nombre dans chaque cellule humaine (entre 100 et 1000 copies). Elle contient 37 gènes codant pour des protéines, essentiellement impliquées dans la production d'énergie cellulaire ou participant à la structure de la mitochondrie.
Le génome mitochondrial contient une région non-codante dite de contrôle ou D-loop, au sein de laquelle

se trouvent deux séquences hypervariables (HVS1 entre les positions 16024 et 16365 ; HVS2 entre les positions 73 et 340), dont le taux de mutation est élevé atteignant $1,17 * 10^{-5}$ par paire de base et par génération dans HVS1 (Heyer *et al.* 2001).

En symétrique du chromosome Y, l'ADN mitochondrial est transmis presque exclusivement par la mère à ses enfants (Giles *et al.* 1980 ; Schwartz et Vissing 2002), sans recombinaison, à l'identique aux mutations *de novo* près, et permet d'accéder facilement à des informations sur les lignées féminines. La taille efficace respective des chromosomes uniparentaux dépend uniquement de la taille efficace masculine (Ne_M) et féminine (Ne_F) (Hedrick 2007) : $Ne_Y = Ne_M$ et $Ne_{Mt} = Ne_F$.

Haplogroupes

L'absence de recombinaison des ADN uniparentaux permet de regrouper les différents haplotypes en clades monophylétiques hierarchisés, appelés haplogroupes, définis sur la base de marqueurs à mutation lente (The Y Chromosome Consortium 2002 ; Karafet *et al.* 2008 ; van Oven et Kayser 2009). La répartition de ces haplogroupes a la particularité d'être très structurée dans l'espace, suivant les mouvements associés au peuplement de la Terre par notre espèce (voir en Annexe les Figures 31 & 32). Leur structuration géographique a notamment été utilisée lors d'études phylogéniques pour détecter des événements de migration et de métissage génétique, du point de vue des femmes et des hommes (Destro-Bisol *et al.* 2010 ; Vigilant *et al.* 1991 ; Voskarides *et al.* 2016). En particulier, plusieurs études rapportent des asymétries entre sexes lors de tels événements (Wilkins 2006) : au Groenland, des chromosomes Y européens ont été retrouvés parmi un fond génétique très largement inuit (Pereira *et al.* 2015), alors qu'aucun ADN mitochondrial européen n'a été retrouvé (Bosch *et al.* 2003), ce qui suggère un métissage génétique entre des hommes européens et des femmes inuits locales. Un autre exemple célèbre est celui de la conquête des Amériques par des hommes européens, dont on retrouve les chromosomes Y associés à des ADN mitochondriaux amérindiens, mais à aucun ADN mitochondrial européen (Nuñez *et al.* 2010 ; Adhikari *et al.* 2016). En Papouasie-Nouvelle-Guinée, les chromosomes Y sont principalement asiatiques alors que les ADN mitochondriaux sont mélanésiens (Kayser *et al.* 2008) (Figure II.4). À une échelle plus locale, des métissages asymétriques entre sexes ont aussi été détectés dans cette région : les migrations féminines ont lieu exclusivement au sein des hautes terres tandis que les hommes migrent plutôt entre la côte et les montagnes (Kayser *et al.* 2003).

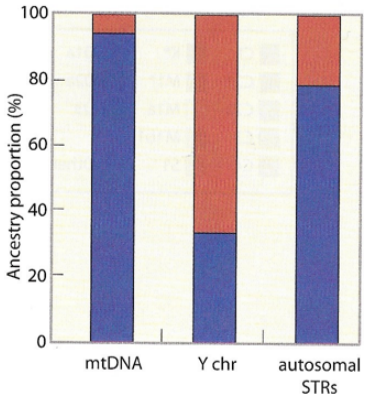


Figure II.4 – Ascendances mélanésienne (en rouge) et asiatique (en bleu) mesurées pour différents marqueurs dans les îles de l'Amirauté en Papouasie Nouvelle-Guinée. *Figure 13-26 de Jobling* et al. *(2013), à partir de Kayser* et al. *(2008).*

Utilisation couplée des chromosomes autosomaux et X

Outre les chromosomes uniparentaux, le chromosome X, au cœur du système de détermination sexuelle humain, peut être utilisé pour étudier des CCAS car il présente des asymétries entre les sexes. En effet, il est diploïde chez les femmes mais haploïde chez les hommes, le sexe hétérogamétique, dont l'unique chromosome X est hérité de leur mère. Ce mode de transmission, différent de celui des autosomes, leur confère une taille efficace valant 3/4 de celle des autosomes si le sex-ratio est équilibré, mais cette valeur dépend du sex-ratio de la population et donc des CCAS ayant cours. En effet, la taille efficace du chromosome X, contrairement à celle des autosomes, est sensible aux différences d'effectifs efficaces entre les hommes et les femmes présents dans la population et soumis aux effets des CCAS (Hedrick 2007 ; Schaffner 2004) :

$$Ne^X = \frac{9Ne_F * Ne_M}{2Ne_F + 4Ne_M} \tag{II.1}$$

$$Ne^A = \frac{4Ne_F * Ne_M}{Ne_F + Ne_M} \tag{II.2}$$

$$\frac{Ne^X}{Ne^A} = \frac{9}{8(1 + Ne_M/(Ne_F + Ne_M))} \tag{II.3}$$

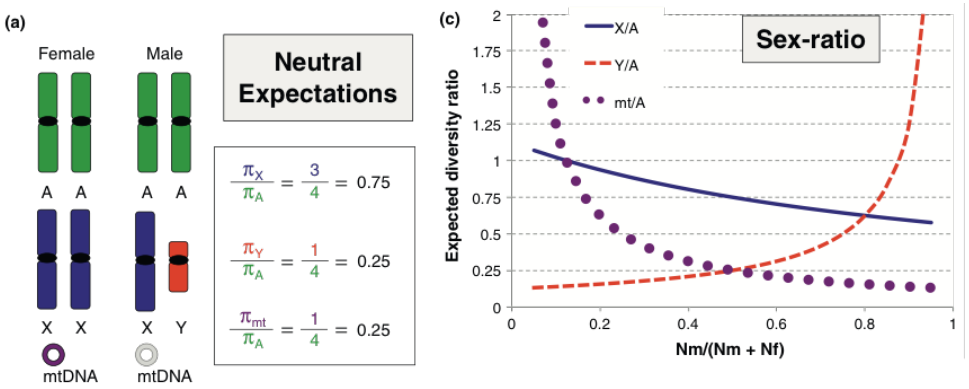


Figure II.5 – Diversité génétique des chromosomes X, Y et mitochondrial rapportée à celle des autosomes. La diversité est calculée comme $\pi = 2Ne\mu$, où le taux de mutation est noté μ et Ne est la taille efficace de chaque chromosome. Le taux de mutation étant supposé constant sur l'ensemble du génome, le ratio des diversités correspond au ratio des tailles efficaces des chromosomes, et varie en fonction du sex-ratio de la population. Les *Neutral expectations* sont données pour un sex-ratio équilibré. *Figure 1 a & c de Webster et Wilson Sayres (2016).*

Dans la pratique, les tailles efficaces sont accessibles à travers des analyses assez poussées et pour étudier les CCAS on peut mesurer de manière plus directe la différenciation génétique entre deux populations indépendamment pour le chromosome X et pour les autosomes. Le ratio des F_{ST} permet alors d'inférer conjointement les tailles efficaces masculine et féminine de la métapopulation et les taux de migration sexe-spécifiques (Ségurel *et al.* 2008) :

$$\frac{1-1/F_{ST}^{X}}{1-1/F_{ST}^{A}} = \frac{3(1+m_F/m)}{4(2-Ne_F/Ne)} \tag{II.4}$$

$$\text{soit } F_{ST}^{X} = \frac{4F_{ST}^{A}}{4F_{ST}^{A} - 3(F_{ST}^{A}-1)(\frac{1+m_F/m}{2-Ne_F/Ne})} \tag{II.5}$$

où m_F/m représente la proportion de femmes migrantes parmi le nombre total de migrants par génération et Ne_F/Ne la proportion féminine de la taille efficace de la métapopulation.

Le cas $F_{ST}^{A} > F_{ST}^{X}$ est particulièrement informatif car il permet de découpler l'information de taille efficace et de migration : il n'est observé que si la taille efficace féminine est plus grande que celle des hommes, pour n'importe quel taux de migration (Figure II.6). L'autre cas de figure, $F_{ST}^{A} < F_{ST}^{X}$, ne permet pas de conclure à des différences de taux de migration et de tailles efficaces entre sexes, ni de découpler leur estimation, et on ne pas tirer de conclusion en terme de CCAS de cette observation.

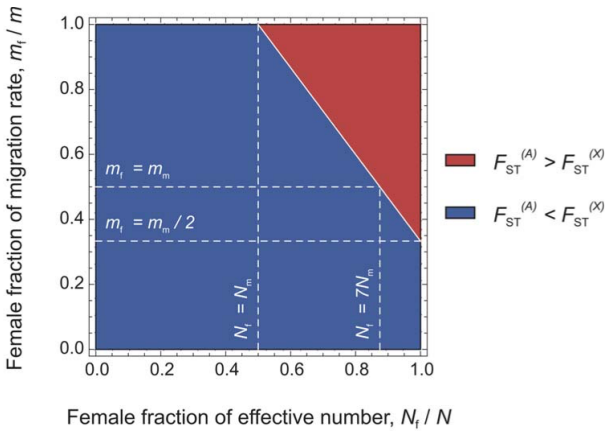


Figure II.6 – Comparaison des F_{ST} obtenus pour les chromosomes autosomaux et X en fonction de la taille efficace et du taux de migration des femmes, rapportés à ceux de la métapopulation. En particulier, le triangle rouge illustre le cas où $F_{ST}^{A} > F_{ST}^{X}$ est causé par une taille efficace féminine plus grande que celle masculine. *Figure 2 de Ségurel* et al. *(2008)*.

Cette approche comparative entre le chromosome X et les autosomes peut également être appliquée à des données d'ADN ancien afin d'estimer si les contributions d'une populations source à une population métissée étaient symétriques ou asymétriques entre les sexes. En effet, le ratio des F_{ST} mesurés entre la population "source" et la population "métissée" pour le chromosome X et les autosomes renseigne sur le sex-ratio de cet événement de métissage (Goldberg *et al.* 2017).

Données autosomales

Les données génétiques autosomales transmises par les deux parents à leurs descendants sont *a priori* peu informatives dans l'étude des CCAS. Cependant, en contrastant des mesures de diversité génétique obtenues séparément chez des hommes et des femmes, on peut inférer des différences entre sexes, imputables à des CCAS. Dans le cadre de migrations sexe-spécifiques où les individus d'un sexe migrent entre différentes communautés tandis que ceux de l'autre sexe demeurent à l'endroit où ils sont nés, c'est-à-dire sont philopatriques, on s'attend à une différenciation entre les communautés plus faible pour le sexe migrant que pour le sexe philopatrique. Cela va de pair avec un apparentement entre individus d'une même communauté plus fort pour le sexe philopatrique que pour le sexe migrant (Prugnolle et

de Meeus 2002).
De plus, les données autosomales apportent, en théorie, des informations sur le déséquilibre de sex-ratio survenu lors d'événements de métissages génétiques ponctuels (Goldberg *et al.* 2014) : les niveaux de métissage retrouvés chez les individus de la population métissée sont plus homogènes si les nombres d'hommes et de femmes d'une même population source ayant contribué au métissage sont très différents. En effet, dans le cas d'un sex-ratio complètement déséquilibré où des femmes d'une population source A se reproduisent exclusivement avec des hommes d'une autre population source B, la variance des niveaux de métissage est nulle et tous les individus métissés sont A à 50% et B à 50 %. Mais si des femmes A s'unissent à des hommes originaires de populations A et B, certains de leurs descendants sont uniquement A et d'autres A & B à 50%, ce qui crée de la variance dans les métissages observés.

Avantages et limites de chaque marqueur

Du fait de leur transmission et absence de recombinaison, les marqueurs uniparentaux sont une manière simple d'accéder à l'histoire des lignées masculines et féminines et à des informations sexe-spécifiques, au contraire du chromosome X qui n'est informatif que lorsqu'il est comparé aux autosomes. Par leur transmission asymétrique entre sexes, les chromosomes X, Y et mitochondriaux capturent des différences de diversités génétiques sur plusieurs générations alors que les données autosomales ne rendent compte que de différences survenues à la génération étudiée, puisque le signal est perdu à chaque nouvelle génération par le mélange des chromosomes des pères et des mères.
L'absence de recombinaison des ADN uniparentaux peut cependant être un inconvénient : tous leurs marqueurs sont liés, sont donc soumis à la même pression de sélection et ne témoignent de l'histoire que d'un seul ancêtre. Au contraire, le chromosome X est porteur de plusieurs loci indépendants, ce qui permet de moyenner les effets de la sélection, mais aussi d'accéder à des informations héritées de plusieurs ancêtres.

Le fait de contraster la diversité de chromosomes différents peut également être problématique. Tout d'abord, ces différents chromosomes, voire les régions de ces chromosomes, ont des taux de mutation différents et certains accumulent donc plus de mutations que d'autres pendant le même laps de temps. En particulier, l'utilisation de HVS1 dont le taux de mutation est très élevé conduirait à sous-estimer la différenciation entre populations, par rapport à ce qui est observé avec le chromosome Y (Wilder *et al.* 2004). Un moyen de dépasser ce biais serait d'utiliser des marqueurs portés par les segments homologues des chromosomes X et Y qui ont des taux de mutation comparables (Balaresque *et al.* 2006).
De plus, la sélection n'agit pas de manière comparable pour tous ces chromosomes : les mutations récessives délétères portées par le chromosome X sont systématiquement exposées chez les hommes et donc soumettent ce chromosome à des pressions de sélection plus fortes que les autosomes. Cela limite le maintien de nouvelles mutations et entraîne une diminution de la diversité génétique de ce chromosome. De plus, par sa taille efficace, le chromosome X est aussi soumis à plus de dérive que les autosomes et est donc moins diversifié. Cette petite taille modifie aussi sa réaction lors de changements démographiques : lors d'une réduction de taille de la population, la dérive de ce chromosome est majorée et sa diversité est encore plus faible relativement à celle des autosomes. Au contraire, dans le cas d'une expansion, la diversité du chromosome X retrouve plus rapidement son équilibre que celle des autosomes (Pool et Nielsen 2007) ce qui pourrait expliquer certains cas de figures observés dans la littérature, par exemple par Yang *et al.* (2010).
Pour les marqueurs uniparentaux, on peut également s'attendre à des différences de sélection.

II.3 Quelques observations de différences génétiques sexe-spécifiques

Différences de diversité génétique et CCAS

De nombreuses études de génétique des populations fondées sur l'analyse parallèle des marqueurs uniparentaux rapportent des différences de diversité génétique pour les lignées masculines et féminines (Destro-Bisol *et al.* 2004 ; Nasidze *et al.* 2004).
Ces différences sont probablement causées par des CCAS mais, dans la plupart de ces études, le CCAS responsable n'est pas formellement identifié. En effet, associer les différences génétiques observées à un comportement culturel en particulier n'est pas toujours chose aisée, notamment car cela suppose de connaître les populations sur un plan ethno-démographique. D'autre part, plusieurs CCAS peuvent avoir cours dans les populations étudiées : en Papouasie-Nouvelle-Guinée, des différences génétiques sexe-spécifiques sont observées chez des populations à la fois patrilocales, patrilinéaires et pouvant présenter des asymétries entre sexes de transmission du succès reproducteur ou de mortalité du fait d'affrontements entre sociétés (Kayser *et al.* 2003). D'autres études ne prennent pas de telles précautions et imputent les différences observées à un CCAS particulier, sans envisager un effet combiné de plusieurs CCAS : chez des populations de la péninsule du Sinai, la polygynie est tenue pour responsable la diversité génétique du chromosome Y 14 fois inférieure à celle de l'ADN mitochondrial (Salem *et al.* 1996). Cela semble être un effet trop fort pour la seule polygynie : dans des cas de polygynie extrême, les tailles efficaces féminines ne valent que le double des tailles masculines (Heyer *et al.* 2012).

Afin d'identifier l'effet d'un CCAS précis sur la diversité génétique, certaines études choisissent d'étudier conjointement des populations différant pour ce CCAS, sous l'hypothèse qu'elles sont identiques toute chose égale par ailleurs. Cela suppose donc une connaissance ethnologique détaillée des populations mais aussi de travailler à une échelle géographique locale pour avoir des conditions environnementales et de peuplement comparables. Notons que des études menées à des échelles plus larges ont aussi observé des différences génétiques sexe-spécifiques sur lesquelles nous revenons à la fin de ce chapitre.

Ainsi, l'étude de la règle de résidence a été particulièrement étudiée en Thaïlande, où des populations matri et patrilocales cohabitent. La comparaison de leur diversité génétique confirme les attendus théoriques : les populations matrilocales sont plus diverses pour le chromosome Y que pour l'ADN mitochondrial, tandis que l'observation inverse est réalisée pour les populations patrilocales, et que des différences de diversités sexe-spécifiques sont observées entre les groupes (Oota *et al.* 2001 ; Besaggio *et al.* 2007) (Figure II.7).

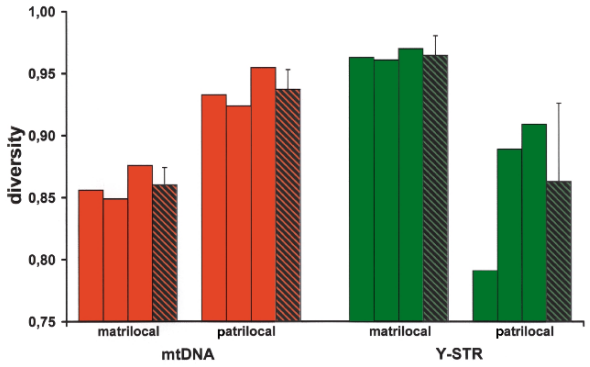


Figure II.7 – Diversité génétique de populations matrilocales et patrilocales de Thaïlande. Les populations matrilocales présentent une diversité mitochondriale réduite par rapport à la diversité du chromosome Y, et le schéma opposé est observé chez les populations patrilocales. *Figure 1 de Oota* et al. *(2001).*

Cependant, des analyses Bayésiennes ont montré que la matrilocalité et la patrilocalité ne seraient pas des CCAS exactement symétriques : les migrations sont plus asymétriques entre les sexes pour les sociétés thaïlandaises patrilocales que pour les matrilocales. Les femmes sont légèrement moins nombreuses (d'un facteur 0,79) que les hommes à migrer dans les sociétés matrilocales, alors que dans les populations patrilocales, les hommes migrants sont 15 fois moins nombreux que les femmes migrantes (Hamilton *et al.* 2005) (Figure II.8).

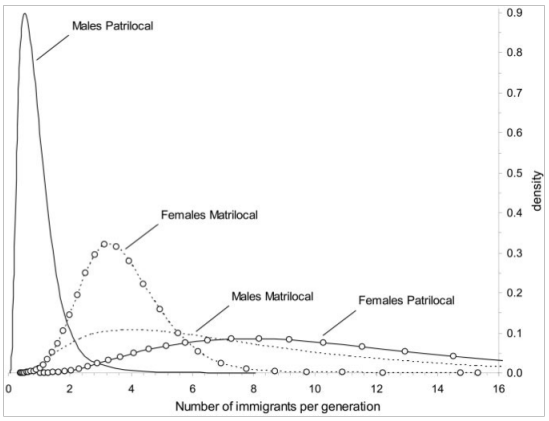


Figure II.8 – Nombre d'hommes et de femmes migrant par génération entre des sociétés patrilocales ou matrilocales de Thaïlande. *Figure 1 de Hamilton* et al. *(2005).*

Selon le même principe que pour les règles de résidence et en s'appuyant sur des données ethnologiques, il est possible d'estimer l'impact de certaines barrières sur la diversité génétique et éventuellement détecter des asymétries entre sexes, en mesurant la différenciation entre des populations de part et d'autre de cette barrière. Pour la barrière de la langue, une étude s'intéresse à des populations africaines de quatre groupes linguistiques majeurs, et trouve une corrélation entre les distances génétiques mesurées pour le chromosome Y et les distances linguistiques, sans voir de structuration spatiale (Wood *et al.* 2005). Pour l'ADN mitochondrial, par contre, la langue et la génétique ne sont plus corrélées, suggérant des migrations de femmes entre des groupes de langues différentes. De même, dans l'est de l'Inde, les familles linguistiques sont différenciées pour le chromosome Y mais pas pour l'ADN mitochondrial (Sahoo et Kashyap 2006).

Outre ces barrières, d'autres facteurs peuvent intervenir à la suite des mariages compliquant les estimations réalisées au moyen des données génétiques. Un exemple passionnant a été documenté dans l'ouest de l'Afrique centrale où des événements asymétriques de métissage génétique sont recensés entre des populations agricultrices non-pygmées et pygmées, en lien avec des barrières culturelles. Ethnologiquement, des unions entre des femmes pygmées et des hommes non-pygmées ont été observées, associées à un départ de ces femmes dans le groupe de leur époux, tandis que les unions entre des hommes pygmées et des femmes non-pygmées semblent proscrites.
Génétiquement, comme attendu, aucun chromosome Y pygmée n'est retrouvé chez les non-Pygmées mais de façon plus surprenante, une introgression de chromosomes Y non-pygmées est observée chez les Pygmées et aucun échange d'ADN mitochondrial n'est détecté entre les groupes (Berniell-Lee *et al.* 2009 ; Marks *et al.* 2015 ; Quintana-Murci *et al.* 2008). Ces observations, quelque peu surprenantes d'après la règle de patrilocalité et le type d'unions observées, seraient expliquées par une discrimination sociale contre les Pygmées (Verdu *et al.* 2013) : les mariages mixtes pygmée/non-pygmée conduisent souvent à un retour de la femme pygmée dans son groupe d'origine avec les enfants nés de cette union mixte, en

cas de séparation ou de décès de l'époux. Cette arrivée des enfants génétiquement mélangés dans la population pygmée, en particulier les petits garçons, expliquerait la présence dans la population maternelle de marqueurs non-pygmées hérités de leur père.

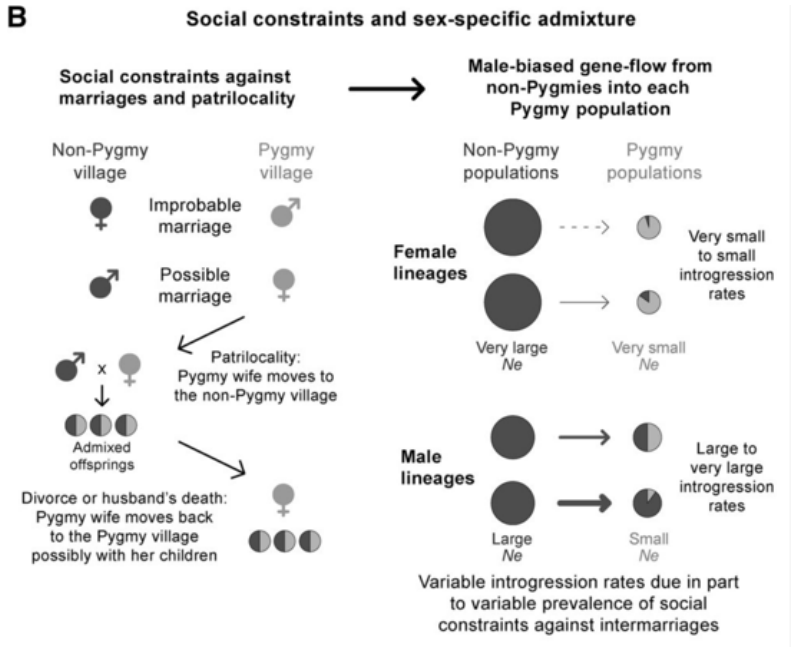


Figure II.9 – Modèle de métissages asymétriques entre des populations pygmées et non pygmées voisines. *Figure 7 b (Verdu* et al. *2013).*

Transmission du succès reproducteur

L'effet de migrations entre populations sur les diversités génétiques a été étudié à de nombreuses reprises mais ne serait cependant pas suffisant pour expliquer certaines différences sexe-spécifiques observées, comme les expansions ciblées du chromosome Y sur tous les continents à des périodes différentes (Poznik *et al.* 2016). On peut expliquer ces expansions de quelques lignées par le succès reproducteur de quelques hommes ou femmes d'une population pendant plusieurs générations, c'est-à-dire une transmission du succès reproducteur assimilée à un CCAS (Balaresque *et al.* 2015).

Deux exemples asiatiques pour le chromosome Y ont particulièrement retenu l'attention : l'une de ces deux lignées (haplogroupe C3c - motif dit "Mandchou") est assez fréquente à l'heure actuelle en Mongolie et dans le nord de la Chine, incluant 3% des hommes d'Asie de l'est. Elle aurait été propagée par les descendants de Giocangga, fondateur de la dynastie Qing, il y a 590 ans (± 340) (Xue *et al.* 2005 ; Balaresque *et al.* 2015). L'autre lignée (anciennement dite C3*, actuellement notée C2-M217) atteint une fréquence de près de 30% chez certaines ethnies d'Asie intérieure comme les Kazakhs, Mongols et chez les Hazaras du Pakistan. Elle aurait émergé en Mongolie il y a 1 000 ans et est attribuée à Genghis Khan ou plus généralement au clan Keraït auquel il aurait été rattaché par sa lignée paternelle (Abilev *et al.* 2012 ; Zerjal *et al.* 2003). En Europe, actuellement, 64% des haplogroupes du chromosome Y appartiennent à trois lignées (I1, R1a, R1b), apparues il y a 7 500 à 3 500 ans et dont l'expansion serait survenue à l'âge du Bronze, il y a 4 200 à 2 100 ans (Batini *et al.* 2015).

Un cas a aussi été documenté pour les haplogroupes mitochondriaux : l'haplogroupe H s'est répandu en Europe, remplaçant l'haplogroupe "paléolithique" U, au moment de la transition néolithique et est porté par 40% des Européens actuels (Barbujani 2012).

II.4 La diversité génétique sexe-spécifique en Asie intérieure

État de l'art

Travaux antérieurs de l'équipe d'Anthropologie Évolutive sur les CCAS

L'Asie intérieure est une région d'intérêt dans l'étude des CCAS, avec plusieurs CCAS décrits sur la base d'observations ethnologiques, tels que la patrilocalité, une faible pratique de la polygynie et, pour certaines populations, la patrilinéarité (Krader 1963).

L'impact de certains de ces CCAS sur la diversité génétique a été étudié à partir des chromosomes Y et mitochondriaux, mais aussi du chromosome X contrasté avec les autosomes.
Des différences génétiques ont été observées entre des populations turco-mongoles patrilinéaires d'Asie centrale et leurs voisines indo-iraniennes cognatiques, et ont donc été imputées à la patrilinéarité (Ségurel *et al.* 2008) : les populations patrilinéaires présentent des tailles efficaces féminines plus grandes que les tailles masculines, tandis que les populations cognatiques ne présentent pas d'asymétrie. De plus, la différenciation entre populations patrilinéaires est plus forte qu'entre populations cognatiques pour le chromosome Y (F_{ST}=0,177 contre 0,069), mais similaire pour l'ADN mitochondrial (F_{ST}=0,010 contre 0,034), illustrant le fait que la patrilinéarité n'affecte que la diversité génétique spécifique masculine. Une structuration particulière de la diversité du chromosome Y a été décelée chez les populations patrilinéaires : des haplotypes du chromosome Y ont été observés à haute fréquence dans ces populations, appartenant à des hommes apparentés au sein du même lignage paternel et donc porteurs des mêmes versions du chromosome Y (Chaix *et al.* 2007). Cette structure est appelée "noyaux d'identité" (Figure II.10). En plus de ces noyaux, de plus rares haplotypes ressemblant fortement à l'haplotype du noyau et donc apparus par mutation depuis l'haplotype ancestral sont observés, ainsi que d'autres haplotypes à faible fréquence, plus éloignés génétiquement car appartenant à quelques individus issus de lignées masculines distinctes. Les noyaux d'identité pourraient être le résultat sur le plan génétique des dynamiques démographiques de fission et de dérive, précédemment décrites, agissant au sein des groupes de filiation patrilinéaires.

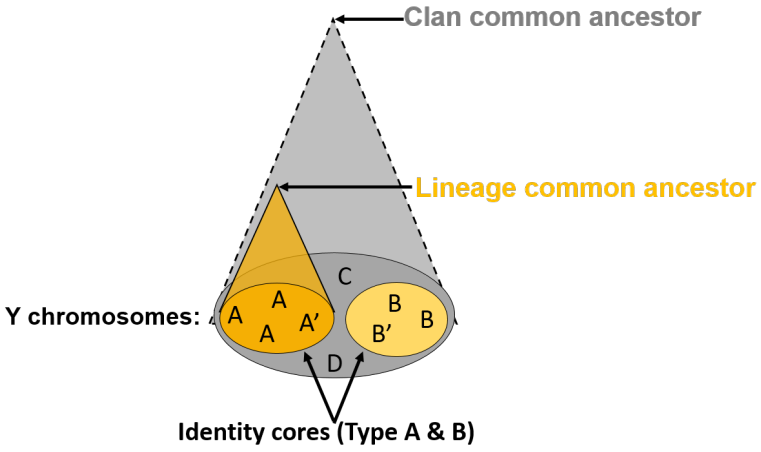


Figure II.10 – Noyaux d'identité résultant de la structuration masculine en groupe de descendance, formés à partir d'ancêtres biologiques. Plusieurs hommes de la même population, du fait de leur ascendance commune par la lignée paternelle, portent le même chromosome Y (ici de type A ou B), ce qui cause une faible diversité d'haplotypes du chromosome Y dans la population. On observe également quelques haplotypes hors des noyaux d'identité (C et D), et des haplotypes dérivés de l'haplotype ancestral du noyau d'identité (A' et B').

L'influence de la patrilinéarité a également été observée à travers une transmission paternelle du succès reproducteur plus marquée pour les populations patrilinéaires que pour les populations cognatiques (Heyer *et al.* 2015), causant des clusters de chromosomes Y fortement apparentés et représentant 38% des chromosomes asiatiques actuels (Balaresque *et al.* 2015).
La patrilinéarité pourrait également être responsable de l'absence de signal d'expansion des populations patrilinéaires mesurée en utilisant le chromosome Y, quand les populations cognatiques connaissent elles une expansion (Aimé *et al.* 2015), et que toutes les populations sont en expansion en se fondant sur l'ADN mitochondrial (Aimé *et al.* 2013).

En outre, la patrilocalité a également étudiée et serait responsable d'une différenciation plus marquée pour le chromosome Y que pour l'ADN mitochondrial chez des populations kazakhes (Tarlykov *et al.* 2013) et kirghizes (Pérez-Lezaun *et al.* 1999) d'Asie centrale. En sus, les travaux de l'équipe d'Évelyne Heyer montrent que des échanges d'ADN mitochondrial sont réalisés entre différentes populations, outrepassant les barrières linguistiques et ethniques (Heyer *et al.* 2009). En particulier, la différenciation légèrement plus forte observée pour l'ADN mitochondrial entre les populations cognatiques (F_{ST}=0,034 contre 0,010 chez les patrilinéaires) pourrait s'expliquer par des mouvements de femmes plus fréquents entre les populations patrilinéaires qu'entre les populations cognatiques, en lien avec des règles d'alliance exogame différentes entre les groupes linguistiques.

Inférence sur l'histoire du peuplement et des groupes ethniques

Outre l'étude des CCAS, les marqueurs uniparentaux ont été utilisés pour inférer l'histoire démographique de l'Asie intérieure du point de vue unilatéral des hommes et des femmes. Ainsi, sur la base des fréquences d'haplogroupes, on retrouve le patron de proximités génétiques décrit précédemment au moyen de données autosomales. En effet, un gradient d'haplogroupes européens et asiatiques est retrouvé pour l'ADN mitochondrial : en Asie centrale, ils représentent respectivement 50% des haplogroupes observés, avec une contribution discrète de l'Asie du sud (Comas *et al.* 2004) alors qu'en Asie du nord, la composante est-asiatique atteint 81% et la composante européenne n'est que de 17% (Derenko *et al.* 2003). En particulier, le groupe indo-iranien tadjik présente une plus forte contribution européenne qu'asiatique (63% contre 26%) (Comas *et al.* 2004). Pour le chromosome Y, la distribution des différents haplogroupes suit également un gradient est-ouest révélant des différences entre l'Asie centrale et du nord, mais certaines populations d'Asie du nord, comme les Altai-Kizhi, sont plus proches de populations d'Asie centrale que d'autres populations d'Asie du nord (Dulik *et al.* 2012). Malgré ce gradient, les populations d'Asie intérieure partagent toutes certaines haplogroupes comme C et R1a1 (Derenko *et al.* 2006) - l'haplogroupe C inclut notamment les lignages attribués à Genghis Khan (Zerjal *et al.* 2003) et Gioccanga (Xue *et al.* 2005) et l'haplogroupe R1a1a ferait partie des haplogroupes apparus en Asie intérieure (Wells *et al.* 2001) - . Par ailleurs, les haplogroupes du chromosome Y sont extrêmement variés en Asie intérieure (Chiaroni *et al.* 2009) et sont partagés avec d'autres régions du monde comme le Proche-Orient et le Caucase (haplogroupe T ou J), l'Europe et l'Asie de l'est (N et R1a), suggérant des mouvements d'hommes entre l'Asie intérieure et ces régions.

Les marqueurs uniparentaux permettent également de s'intéresser plus en détail à l'histoire des groupes ethniques, en choisissant le point de vue des hommes ou des femmes. En particulier, certains groupes ethniques d'Asie intérieure sont aussi présents dans des régions voisines comme l'Afghanistan, l'Iran ou le Xinjiang chinois. Sur la base des distributions d'haplogroupes, les populations regroupées au

sein d'un même groupe ethnique partageraient bel et bien une origine commune malgré les frontières nationales (Malyarchuk *et al.* 2013 ; Di Cristofaro *et al.* 2013 ; Shan *et al.* 2014). Au sein de l'Asie intérieure, le groupe kazakh est présent à la fois en Asie centrale et du nord ; l'analyse de l'ADN mitochondrial révèle une forte affinité entre les Kazakhs d'Asie du nord et centrale, avec un apport nord-asiatique chez ceux du nord (Derenko *et al.* 2012) tandis que les données du chromosome Y montrent une homogénéité des Kazakhs, néanmoins associée des différences entre les Kazakhs d'Asie centrale et de Sibérie sans qu'elles soient imputables à un apport génétique sibérien et peut-être causées par de la dérive (Dulik *et al.* 2011).

L'une des études menées par l'équipe d'Évelyne Heyer sur des données collectées sur le terrain ou dans la littérature porte sur l'inférence de l'âge des groupes ethniques de langue turco-mongole d'Asie centrale (Heyer *et al.* 2009). L'appartenance au groupe étant plus marquée pour les hommes que pour les femmes, l'étude s'intéresse à l'âge de l'ethnie du point de vue des hommes, en utilisant des données STR du chromosome Y. Les âges estimés sont plus vieux que ceux des sources historiques, ce qui suggère que l'ethnie est un regroupement de populations sur une base culturelle plutôt que biologique. La même conclusion avait été rendue pour le niveau précédant d'organisation sociale : la tribu (Chaix *et al.* 2004). En revanche, les éléments emboîtés de la tribu, à savoir les clans et lignages, étaient eux des groupes "biologiques", au sein desquels les apparentements des hommes étaient plus forts qu'à l'extérieur de ces structures.

Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations **- Résumé -**

Dans la continuité des travaux menés en Asie centrale sur l'effet des CCAS et le peuplement, la région étudiée a été élargie au cours de cette thèse à l'ensemble de l'Asie intérieure, en incluant des populations d'Asie du nord, patrilocales et patrilinéaires. Cet échantillonnage nous a permis :

1. de tester l'existence de structure en noyaux d'identité du chromosome Y chez des populations d'Asie du nord décrites comme étant patrilinéaires et de détecter d'éventuelles différences dans la pratique de la patrilinéarité entre l'Asie centrale et du nord, à partir de données génétiques ;
2. de chercher des effets de la patrilocalité en Asie intérieure et d'évaluer l'impact de la distance sur les échanges entre populations ;
3. de dater la formation des groupes ethniques en incluant des populations provenant de toute leur aire de répartition dans le cas des Kazakhs et Kirghizes ;
4. d'observer les proximités génétiques des populations d'Asie intérieure avec des populations européennes et est-asiatiques, en combinant en une seule étude l'Asie centrale et du nord et des données d'haplogroupes des chromosomes Y et mitochondriaux. En particulier, l'utilisation de données d'haplogroupes nous a permis d'observer des processus plus anciens que ceux détectés à partir des haplotypes STRs ou des séquences HVS1, en lien avec le peuplement de la région.

Matériel

Pour cette étude, nous avons étudié 39 populations des deux groupes linguistiques majeurs d'Asie intérieure (turco-mongol et indo-iranien), réparties en 16 groupes ethniques dont

- un groupe indo-iranien d'Asie centrale : les Tadjiks ;
- trois groupes turco-mongols d'Asie centrale : les Karakalpaks, Turkmènes et Ouzbeks ;
- dix groupes turco-mongols d'Asie du nord : les Altai-Kizhi, Bouryates, Irgits, Khakasses, Mongols, Monguches, Ondars, Chors, Telengits et Tubalars ;
- deux groupes turco-mongols présents à la fois en Asie centrale et du nord : les Kazakhs et Kirghizes.

Cet échantillonnage inclut 1 499 participants dont 1 231 hommes et 268 femmes. Nous avons publié dans cet article les séquences HVS1 de 1 428 ADN mitochondriaux et 57 mitogénomes complets. En parallèle, pour le chromosome Y, nous disposons de données pour un consensus de huit STRs, pour 1 152 hommes. Les données ont été nouvellement produites pour 12 populations (soit 217 hommes) et celles des 27 autres populations avaient déjà été publiées (Balaresque *et al.* 2015). Pour nos analyses portant sur les données STR, nous avons ajouté six populations de la littérature (typées avec le même kit) afin de disposer de populations kazakhes d'Asie du nord (Dulik *et al.* 2011, 2012).
Les haplogroupes mitochondriaux ont été définis à partir de positions variables séquencées dans HVS1, plus des positions hors HVS1 en cas d'ambiguïté, en utilisant le logiciel Haplogrep (Kloss-Brandstätter *et al.* 2011), et ceux du chromosome Y à partir de marqueurs diagnostic typés par réaction TaqMan et inclus dans l'arbre ISOGG Y-ADN 2014 version 9.01. Pour contextualiser nos données d'Asie intérieure, nous avons choisi dans la littérature trois populations d'Europe et d'Asie de l'est dont les données d'haplogroupes mitochondriaux et du chromosome Y étaient disponibles (Santos *et al.* 2014 ; Wang *et al.* 2014).

Méthodes

L'analyse des haplogroupes a consisté en une Analyse en Composantes Principales réalisée à partir des fréquences des haplogroupes présents chez les populations d'Asie intérieure, d'Europe et d'Asie de l'est, pour le chromosome Y et l'ADN mitochondrial indépendamment. L'intérêt de cette approche est de ne pas avoir à attribuer une origine géographique *a priori* aux haplogroupes dans cette région au cœur de l'Eurasie, mais de se baser sur de simples ressemblances en terme de distribution de fréquences d'haplogroupes.

Pour les données haplotypiques, nos analyses ont cherché à mettre en évidence des asymétries de diversité génétique entre le chromosome Y et l'ADN mitochondrial. Elles s'intéressaient en particulier :

1. aux distances F_{ST} entre populations, visualisées par une MDS en deux dimensions, pour étudier les proximités génétiques entre populations ;
2. à la corrélation entre les distances géographiques et génétiques mesurées entre populations, *via* des tests de Mantel (Oksanen *et al.* 2016), pour estimer l'impact des distances géographiques sur les échanges entre populations selon les sexes ;
3. à la répartition de la variance génétique sexe-spécifique (AMOVA (Excoffier *et al.* 1992) sur le chromosome Y ou ADN mitochondrial) pour des groupes formés selon des critères géographiques (populations d'Asie centrale *versus* d'Asie du nord), ethniques (16 groupes ethniques pour toute l'Asie intérieure, 6 en ne se concentrant que sur l'Asie centrale ou 12 pour l'Asie du nord) ;
4. la présence de noyaux d'identité au sein de la diversité intra-populationnelle du chromosome Y à partir du nombre moyen d'individus partageant le même haplotype (C), de la proportion moyenne d'haplotypes présents une seule fois dans chaque population (Ps) et de l'hétérozygotie haplotypique (H) (Chaix *et al.* 2007). Dans ce manuscrit, nous produisons également des MDS proportionnelles réalisées à partir de matrices de distances *Average Square* de Goldstein et Pollock (1997) calculées entre les haplotypes du chromosome Y présents dans chaque population. Le but de ces MDS est de représenter les distances entre les haplotypes mais aussi le nombre d'individus porteurs de ces haplotypes (voir (Chaix *et al.* 2007) et en Annexe).
5. à dater la formation des six groupes ethniques pour lesquels nous disposons de plusieurs populations, *via* BATWING (Wilson *et al.* 2003). Nous nous sommes concentrées sur le point de vue de la lignée paternelle car l'appartenance au groupe ethnique semble un élément structural plus marquant pour la diversité génétique masculine que féminine.

Résultats

Histoire du peuplement de l'Asie intérieure

L'analyse des fréquences d'haplogroupes, aussi bien pour le chromosome Y que pour l'ADN mitochondrial, a montré un gradient eurasien avec à ses extrémités les populations européennes et est-asiatiques (Figure 2 de Marchi *et al.* (2017)). Les populations d'Asie intérieure se répartissent le long de ce gradient avec, comme attendu, les populations indo-iraniennes plus proches du pôle européen et les populations d'Asie du nord plus proches du pôle est-asiatique. Cette analyse nous a permis d'établir un portrait complet de la région à partir des haplogroupes ce qui n'avait jamais été fait à notre connaissance. De plus, nous avons trouvé que les processus anciens ayant créé la diversité des haplogroupes seraient symétriques entre sexes, puisque les mêmes proximités génétiques avec l'Europe et l'Asie de l'est ont été retrouvées pour les deux types d'haplogroupes, avec une forte corrélation entre les positions des populations sur la

première composante principale trouvée pour l'ADN mitochondrial et celle pour le chromosome Y (ρ de Spearman = 0,70, p-value$<2 * 10^{-6}$).

Relations entre les populations

L'analyse des différenciations F_{ST} entre populations, pour les haplotypes STR du chromosome Y et les séquences mitochondriales, a révélé des asymétries entre les deux marqueurs (Figure 3 de Marchi *et al.* (2017)) : le F_{ST} estimé pour toutes les populations d'Asie intérieure est beaucoup plus élevé pour le chromosome Y que pour l'ADN mitochondrial (0,156 contre 0,045, p-values<0,001). Les proximités entre populations pour l'ADN mitochondrial ressemblent à celles "historiques" observées à partir des haplogroupes, mais les proximités trouvées pour les haplotypes Y reflètent d'autres processus. En effet, les populations indo-iraniennes sont resserrées au centre du graphique (avec une faible distance entre populations) tandis que les populations turco-mongoles sont dispersées à la périphérie du graphique, montrant une faible différenciation entre populations indo-iraniennes et une forte différenciation génétique entre les populations turco-mongoles. Cette asymétrie entre les marqueurs et la différence observée entre les populations turco-mongoles, patrilinéaires, pourraient refléter un effet de la patrilinéarité sur la diversité génétique masculine.

Influence de l'organisation sociale

En comparant différents estimateurs de la diversité génétique du chromosome Y entre les populations patrilinéaires et cognatiques, nous avons trouvé une hétérozygotie intra-populationnelle plus faible pour les populations patrilinéaires ainsi qu'une tendance pour les populations patrilinéaires à posséder un nombre moyen d'individus partageant le même haplotype (C) plus élevé et une proportion d'haplotypes présents chez un seul individu de la population (Ps) plus faible que les cognatiques, signe d'une structure génétique en noyaux d'identité (Chaix *et al.* 2007) (Figure II.11). Pour compléter l'analyse de cette structure particulière, nous avons représenté à partir d'une MDS proportionnelle la fréquence des différents haplotypes et leurs différences génétiques pour chacune des populations (en Annexe). De manière purement descriptive, nous distinguons effectivement des noyaux chez la plupart des populations patrilinéaires : chez les Kirghizes, les Turkmènes, deux des trois populations kazakhes, une des deux populations karakalpaks ainsi que certaines populations d'Asie du nord comme les Tubalars, les Altai-Kizhis, les Bouryates. Le faible effectif des autres populations de cette région rend l'interprétation difficile, mais les plus grands effectifs collectés par Dulik pour des populations similaires montrent une structure en noyaux. En parallèle, ces noyaux ne sont pas visibles chez neuf des onze populations tadjikes cognatiques mais le sont chez les deux populations restantes. Le groupe ouzbek retient l'attention car il est supposé patrilinéaire mais deux des trois populations de ce groupe ne présentent pas de noyaux. Dans l'ensemble, la présence de noyaux semble corréler assez bien avec les règles de filiation décrites pour les différents groupes ethniques et certaines des incohérences pourraient être expliquées par des particularités ethnologiques connues (comme le passage à un mode de vie agriculteur sédentaire chez les Ouzbeks (Soucek 2000)), mais pour d'autres, il nous faudra documenter de manière plus précise les règles de filiation des populations concernées.

En outre, nous avons exploré plus en détails la diversité des populations patrilinéaires d'Asie centrale et du nord : nous avons observé un léger excès de C et une réduction de Ps chez les populations d'Asie centrale, suggérant une patrilinéarité discrètement plus marquée en Asie centrale qu'en Asie du nord.

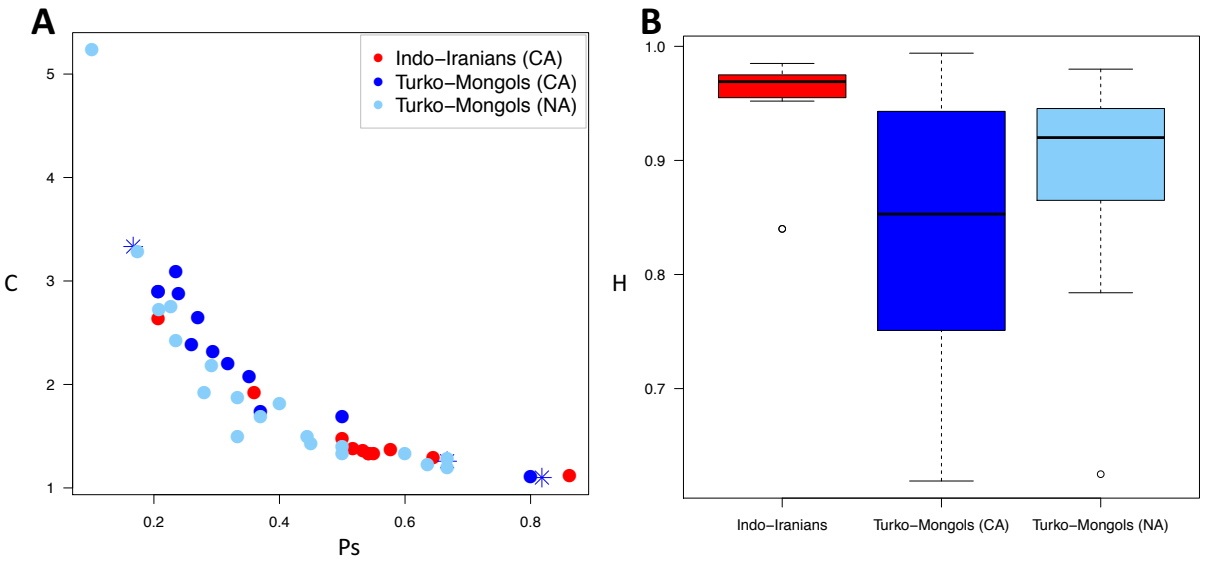


Figure II.11 – Éléments constitutifs des noyaux d'identité du chromosome Y : A) une faible proportion d'haplotypes présents chez un seul individu de la population (Ps) et un nombre moyen d'individus partageant le même haplotype (C) élevé ; B) une hétérozygotie populationnelle (H) réduite. Ces estimateurs sont mesurés pour les chromosomes Y de 45 populations d'Asie intérieure : indo-iraniennes et cognatiques (en rouge), turco-mongoles et patrilinéaires d'Asie centrale (bleu marine) et d'Asie du nord (bleu clair). Les trois populations ouzbèkes sont représentées avec des étoiles car elles auraient abandonné leur organisation sociale traditionnelle (Soucek 2000).

En outre, cette structuration particulière de la diversité génétique pourrait notamment être tenue pour responsable, par dérive, de la plus grande différenciation pour le chromosome Y des populations patrilinéaires comparées aux populations cognatiques et à leur fort niveau de différenciation au sein des groupes ethniques, comparé à celui mesuré pour l'ADN mitochondrial (Table 2 de Marchi *et al.* (2017)).

Influence de la géographie et de l'appartenance ethnique

La variance génétique expliquée par les groupes ethniques est plus élevée pour le chromosome Y que pour l'ADN mitochondrial (5,18% *versus* 2,66%, Table 1 de Marchi *et al.* (2017)). Ce résultat suggère que les barrières entre groupes ethniques sont plus étanches pour les hommes que pour les femmes, que les hommes migrent moins souvent hors de leur groupe ethnique que les femmes.

Si l'on sépare les populations selon leur lieu d'origine, soit en Asie centrale ou du nord, la part de la variance génétique expliquée par la géographie est plus forte pour l'ADN mitochondrial que pour le chromosome Y. De plus, nous avons trouvé, pour l'ADN mitochondrial, des F_{ST} plus forts mesurés entre des populations venant de régions différentes que pour des populations voisines (0,07 *versus* 0,04). Pour le chromosome Y, les F_{ST} sont forts, peu importe l'échelle géographique considérée (0,17 *versus* 0,15). Les tests de Mantel que nous avons réalisés pour l'ADN mitochondrial sont significatifs à l'échelle de l'Asie intérieure et non significatifs en Asie centrale et du nord. Par contre, pour le chromosome Y, les tests de Mantel sont significatifs à toutes les échelles. Ces résultats suggèrent des migrations féminines structurées par la géographie à l'échelle de l'Asie intérieure, mais pas d'effet de la géographie à l'intérieur des aires régionales. En outre, la structuration plus forte de la diversité du chromosome Y, entre les aires mais aussi au sein de chaque aire, par rapport à la diversité mitochondriale, pourrait être causée par la patrilocalité.

Ethnogénèse

Nous avons étudié l'émergence de six groupes ethniques turco-mongols, du point de vue des hommes, à partir des données haplotypiques du chromosome Y collectées par nos soins. Nos résultats confirment les observations de Heyer *et al.* (2009) sur d'autres populations, à savoir que l'âge génétique de l'ethnie est plus vieux que l'âge historique (Figure 4A de Marchi *et al.* (2017)). L'ethnie serait donc une entité sociale ne reposant pas sur des ascendances biologiques récentes. La formation de l'ethnie serait alors due à un mécanisme de fusion de populations, plus ou moins apparentées, en un groupe, sur la base de motivations sociales et/ou culturelles selon la théorie de Barth (1969) (Figure II.12).

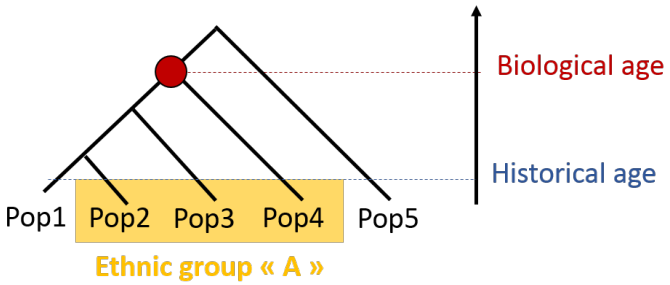


Figure II.12 – Ethnogènese par fusion selon le modèle de F. Barth. L'âge historique du groupe ethnique "A" correspond au moment de la fusion des populations, formant un nouveau groupe ethnique. L'âge biologique du groupe nouvellement formé est celui de la séparation des populations qui la constituent et remonte à des temps plus lointains que l'âge historique.

Comme nous disposions de populations kirghizes et kazakhes résidant dans toute l'Asie intérieure, nous nous sommes intéressées aux liens entre leurs branches d'Asie centrale et du nord (Figure 4B de Marchi *et al.* (2017)). Nous avons trouvé deux scénarios différents : les populations kazakhes sont regroupées par région en deux branches distinctes ce qui rend compte de la forte différenciation intra-groupe observée, tandis que les Kirghizes d'Asie du nord forment une branche incluse au sein de la diversité d'Asie centrale. Ce dernier résultat suggère une séparation assez récente des populations kirghizes d'Asie centrale et du nord, responsable de leur faible différenciation intra-groupe, mais soulève des interrogations : est-ce dû à un départ d'Asie centrale vers l'Asie du nord de quelques populations il y a 700 ans (ce qui est contredit par les sources historiques - voir Introduction), ou à la migration de certaines populations du nord en direction de l'Asie centrale il y a 700 ans et à l'inclusion au sein du groupe kirghize de populations d'Asie intérieure séparées des Kirghizes d'Asie du Nord il y a 1 200 ans ?

Conclusions

En suivant une approche pluri-disciplinaire, nous avons documenté l'impact de plusieurs CCAS sur la diversité sexe-spécifique de populations d'Asie intérieure, dans la continuité des études menées par Chaix *et al.* (2007) ; Heyer *et al.* (2009) en Asie centrale.
Mes travaux de thèse apportent néanmoins des éclairages nouveaux sur cette thématique de recherche. Notamment, l'utilisation couplée de données d'haplogroupes et d'haplotypes a montré que les différences sexe-spécifiques observées sont apparues après la constitution du patrimoine génétique des populations d'Asie intérieure, du fait de processsus plus récents dont l'effet n'est visible qu'à travers les données haplotypiques du chromosome Y. Nous avons par la suite imputé ces différences à de la patrilinéarité, un CCAS pratiqué par l'un des deux groupes culturels d'Asie intérieure. Lors des travaux préalables, la patrilinéarité avait été associée à une structure de la diversité génétique du chromosome Y, dite en noyaux d'identité. Au cours de cette thèse, nous avons retrouvé cette structure chez des populations patrilinéaires

nouvellement échantillonnées en Asie du nord, validant dans une seconde aire géographique le lien entre ce CCAS et la manifestation génétique précedemment décrite.
Cet échantillonnage étendu à l'Asie du nord nous a également permis d'étudier les effets de la patrilocalité à l'échelle de toute l'Asie intérieure. Nous avons alors mis en évidence des différences dans les schémas migratoires selon que l'on travaillait à l'échelle de toute la région ou au sein des deux aires qui la composent.
Nous avons aussi pu apporter un éclairage nouveau à l'histoire des groupes ethniques kazakh et kirghize en incluant des populations originaires de régions différentes dans la reconstruction de leur ethnogénèse. Les résultats obtenus pointent une ethnogénèse par fusion, illustrant l'importance de la culture dans l'organisation des sociétés humaines et dans la mise en place de barrières aux flux de gènes entre des populations voisines mais de groupes différents.

Received: 27 April 2016 | Revised: 30 November 2016 | Accepted: 2 December 2016

DOI 10.1002/ajpa.23151

RESEARCH ARTICLE

WILEY American Journal of PHYSICAL ANTHROPOLOGY

Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations

Nina Marchi[1] | Tatyana Hegay[2] | Philippe Mennecier[1] | Myriam Georges[1] | Romain Laurent[1] | Mark Whitten[3] | Philipp Endicott[1] | Almaz Aldashev[4] | Choduraa Dorzhu[5] | Firuza Nasyrova[6] | Boris Chichlo[1] | Laure Ségurel[1]* | Evelyne Heyer[1]*

[1]Eco-anthropologie et Ethnobiologie, UMR 7206 CNRS, MNHN, Univ Paris Diderot, Sorbonne Paris Cité, F-75016, Paris, France

[2]Uzbek Academy of Sciences, Institute of Immunology, Tashkent, Uzbekistan

[3]MPRG on Comparative Population Linguistics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[4]Institute molecular biology and medicine, Bishkek 720040, Kyrgyzstan

[5]State University of Tuva Republic, Kyzyl, Russia

[6]Laboratory of Plant Genetics, Institute of Botany, Plant Physiology and Genetics, TAS, Dushanbe 734063, Tajikistan

Correspondence
Nina Marchi, Musée de l'Homme, 17 place du Trocadéro, 75016 Paris, France.
Email: nina.marchi@mnhn.fr

Funding information
ANR NUTGENEVOL, Grant Number: 07-BLAN-0064; ANR Altérité culturelle, Grant Number: 10-ESVS-0010; CNRS Programme international de collaboration scientifique

Abstract

Objectives: Sex-specific genetic structures have been previously documented worldwide in humans, even though causal factors have not always clearly been identified. In this study, we investigated the impact of ethnicity, geography and social organization on the sex-specific genetic structure in Inner Asia. Furthermore, we explored the process of ethnogenesis in multiple ethnic groups.

Methods: We sampled DNA in Central and Northern Asia from 39 populations of Indo-Iranian and Turkic-Mongolic native speakers. We focused on genetic data of the Y chromosome and mitochondrial DNA. First, we compared the frequencies of haplogroups to South European and East Asian populations. Then, we investigated the genetic differentiation for eight Y-STRs and the HVS1 region, and tested for the effect of geography and ethnicity on such patterns. Finally, we reconstructed the male demographic history, inferred split times and effective population sizes of different ethnic groups.

Results: Based on the haplogroup data, we observed that the Indo-Iranian- and Turkic-Mongolic-speaking populations have distinct genetic backgrounds. However, each population showed consistent mtDNA and Y chromosome haplogroups patterns. As expected in patrilocal populations, we found that the Y-STRs were more structured than the HVS1. While ethnicity strongly influenced the genetic diversity on the Y chromosome, geography better explained that of the mtDNA. Furthermore, when looking at various ethnic groups, we systematically found a genetic split time older than historical records, suggesting a cultural rather than biological process of ethnogenesis.

Conclusions: This study highlights that, in Inner Asia, specific cultural behaviors, especially patrilineality and patrilocality, leave a detectable signature on the sex-specific genetic structure.

KEYWORDS
Central Asia, demographic history, mitochondrial DNA, patrilineality, Y chromosome

1 | INTRODUCTION

Sex-specific patterns of genetic diversity have already been largely documented in humans, using the male-inherited Y chromosome and the female-inherited mitochondrial DNA (mtDNA) (Jorde et al., 2000), and sex-specific migrations of men and women are thought to be

*Cosupervised the work.
Current address: Myriam Georges, 2LM2E-UMR6197, Laboratoire de Microbiologie des Environnements Extrêmes, Institut Universitaire Européen de la Mer Technopôle Brest-Iroise, Plouzane 29280, France

mostly responsible for these patterns. While human migrations are often pictured as being male-biased at a large geographical scale (Stoneking, 1998), notably during events of settlement (Bosch et al., 2003; Moreno-Estrada et al., 2013; Nuñez et al., 2010), such cases remain marginal and only a few have been demonstrated. On the contrary, most observations support a higher migration of women at a fine geographical scale (Lippold et al., 2014; Seielstad, Minch, & Cavalli-Sforza, 1998; Wilkins, 2006). A key determinant of the sex-biased migrations of men and women, and therefore sex-specific genetic patterns, is the postmarital residence system (Destro Bisol, Capocasa, & Anagnostou, 2012), defined by the place of living of mates after the wedding. Notably, most human populations (~70%) are patrilocal: women move to their husband's natal domicile (Burton et al., 1996; Murdock, 1981). Furthermore, patrilocality seems to exist since a long time (see Heyer, Chaix, Pavard, & Austerlitz, 2012 for a review). On the contrary, in matrilocal populations, men move to their wife's natal domicile. These residence rules have strongly impacted the human genetic diversity at a local scale, with patrilocal populations having a greater level of genetic differentiation between populations and a lower degree of diversity within population on the Y chromosome as compared to mtDNA (Chaix et al., 2007; Lippold et al., 2014; Wilkins, 2006), and the opposite in matrilocal populations (Besaggio et al., 2007; Oota, Settheetham-Ishida, Tiwawech, Ishida, & Stoneking, 2001). These patterns have also been observed when comparing the genetic diversity of the X chromosome to that of the Y chromosome or the autosomes, respectively (Balaresque, Manni, Dugoujon, Crousau-Roy, & Heyer, 2006; Bustamante & Ramachandran, 2009; Keinan, Mullikin, Patterson, & Reich, 2009; Ségurel et al., 2008; Verdu et al., 2013).

Sex-biased migration is not the only factor that can contribute to such sex-specific genetic patterns. Indeed, various cultural features such as polygyny, increased male mortality, shorter female generation time, variance of male transmission of reproductive success and patrilineality (structuration of populations in male descent groups) have been proposed to reduce the male effective population size as compared to that of female, and thus to result in sex-specific genetic patterns (Balaresque et al., 2015; Chaix et al., 2007; Hammer, Mendez, Cox, Woerner, & Wall, 2008; Heyer et al., 2012, 2015; Kayser et al., 2003; Ségurel et al., 2008). However, variations of the migration rate and the effective population size have confounding effects on genetic diversity; therefore, their respective impact is difficult to distinguish.

In this context, Inner Asia is an interesting area to study given that multiple populations, associated to two linguistic groups, Indo-Iranian from the Indo-European linguistic family and Turkic-Mongolic included in the Altaic linguistic family, coexist. These two linguistic groups are patrilocal but are perfectly correlated with differences in social organization: the Indo-Iranian-speaking populations practice endogamous marriages, while their Turkic-Mongolic-speaking neighbors practice exogamous marriages (Krader, 1966). They also have diverse descent rules: Turko-Mongol populations are patrilineal, with individuals belonging to their father's lineage, while Indo-Iranians are cognatic, with individuals belonging to both their mother's and father's lineage. Previous work in this region has already shown that the Y chromosome and the autosomes are more differentiated than the mtDNA and X chromosome, respectively, in both groups (Chaix et al., 2007; Pérez-Lezaun et al., 1999; Ségurel et al., 2008), as expected in patrilocal populations. Contrasting the X and autosomal differentiation patterns, it has further been shown that there is an additional reduction of the male effective population size only in patrilineal populations (Ségurel et al., 2008). This likely reflects the fact that, due to patrilineality, men are more related to one another because they descend from the same recent male ancestor, as compared to women who are outsiders because of the exogamy and patrilocality.

Concerning the history of the people belonging to these two linguistic families, it has been estimated that the Indo-Iranian presence in the region traces back to 8,500 years ago (Palstra, Heyer, & Austerlitz, 2015), *i.e.*, at Neolithic times, even if their exact location of origin remains under discussion (Allentoft et al., 2015). The Turko-Mongol presence in the western Inner Asia was estimated to be more recent, around 2,300 years ago, with people coming from the East (Palstra et al., 2015). This linguistic family is thought to have emerged in Northern Asia, in the Altai area (Yunusbayev et al., 2014), even if this hypothesis has to be confirmed with ancient DNA studies.

Almost all populations in Inner Asia speak languages belonging to these two families, with the exception of more recent migrants due to the reshuffling of people during USSR times. Speakers of Indo-Iranian linguistic family can be found only in the western part of Inner Asia, *i.e.*, in Central Asia. Speakers of the Turkic-Mongolic linguistic family can be found both in Central Asia and in Northern Asia, with some ethnic groups encompassing both areas.

Interestingly, while previous studies focused on a local scale, we reassessed the interplay between cultural and genetic factors at a larger geographical scale. Indeed, we included populations not only from Central Asia (Uzbekistan, Kyrgyzstan, and Tajikistan) but also from Northern Asia (West Mongolia and South Siberia: Buryatia, Altai, and Tuva Republics). Thus, we were able to explore their sex-specific genetic structure in their whole geographical area of settlement. This also enabled us to test for the distinct influence of geography and ethnicity on the genetic differentiation between these populations. Furthermore, because differences in genetic structure between unilateral markers can be due to sex-specific social behaviors and/or to different demographic histories during the settlement of the area and consequent movement of populations, we also investigated broader questions about the origins of Inner Asian populations. Notably, using haplogroup data for both mitochondrial DNA and Y chromosome, we assessed the relationship of each ethnic group with European and Asian populations. Finally, we explored the demographic history of various ethnic groups and their construction as a cultural entity, focusing on Kazakh and Kyrgyz, two ethnic groups who are currently found both in Central and Northern Asia.

2 | MATERIALS AND METHODS

2.1 | DNA samples

For this study, we collected blood or saliva during several field expeditions in Central Asia conducted between 2001 and 2009. Additional

samples were obtained from Northern Asia, including South Siberia and West Mongolia, in 2011 and 2012. Informed consent was obtained for all participants. We collected ethnological questionnaires prior to DNA sampling and excluded individuals more related than first cousins, based on ethnological information. The ethnicity of each participant was determined based on self-reported native spoken language. Our dataset includes three Turko-Mongol ethnic groups from Central Asia (Karakalpak, Turkmen, and Uzbek), ten from Northern Asia (Altai-Kizhi, Bouryat, Irgit, Khakhas, Mongol, Mongush, Ondar, Shor, Telengit, and Tubalar), two present in both Central and Northern Asia (Kazakh and Kyrgyz), and an Indo-Iranian ethnic group from Central Asia (Tajik). For some Turko-Mongol ethnic groups in Northern Asia, only one population was sampled per ethnic group. We defined populations as groups of individuals living in a similar village, and each population was associated to an ethnic group. In total, we obtained DNA from 1,499 participants including 1,231 men and 268 women, from 39 populations (26 Central and 13 Northern Asian populations) (Figure 1 and Table S1).

Additionally, for some analyses based on the Y-STR dataset, we included six South Siberian populations from the literature (Dulik, Osipova, & Schurr, 2011; Dulik et al., 2012): TubD (Tubalar), ChkD (Chelkan), KumD (Kumandin), AkiD (Altai-Kizhi), SekD, and SwkD (Kazakh). We also collected haplogroup data from the literature from two geographically distinct Eurasian areas: South Europe (Santos et al., 2014) and East Asia (Wang et al., 2014). We chose three populations from each area that included more than 15 individuals per population: Val (Valencia), Gal (Galicia), and Alm (Almeria) in Spain; Yj (Hekou Town), Db (Danba), and Bm (Bamei Town) in County of Sichuan, China.

2.2 | Molecular methods

DNA was extracted from blood and saliva samples using standard protocols (Maniatis, Fritsch, & Sambrook, 1982; Quinque, Kittler, Kayser, Stoneking, & Nasidze, 2006).

We genotyped 1,154 men for a set of eight short tandem repeats (STRs) of the non-recombining region of the Y chromosome (NRY): DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, and DYS439. This dataset includes already published data from 27 populations (see references in Table S1) and newly produced data for 12 populations. We produced these new data using AmpFl-STRYFiler PCR Amplification kit (Applied Biosystems Inc, Foster City, CA) and the products of amplification were analyzed on an ABI3130 Genetic Analyzer by the "Service de Systématique Moléculaire" (UMS2700-CNRS) or on an ABI3730 Genetic Analyzer by GenoScreen (Lille, France). Electrophoresis results were analyzed using GeneMapper ID v4.0 software. For individuals where DYS19 was duplicated, we only kept the smaller number of repetitions for the analyses.

We defined Y chromosomal haplogroups thanks to hierarchical TaqMan-assays (Applied Biosystems Inc) of diagnostic biallelic markers selected from the worldwide Y chromosome genealogy. A total of 50 SNPs were included in the assays. Samples were typed for relevant SNPs in a hierarchical procedure, following the Y chromosomal haplogroup tree (more details provided in Table S2). In order to compare our Y haplogroup data with the literature, we grouped haplotypes in 16 categories: Y\CR; D; E; C; F\G, H1, I, J, K; G; H1; I; J; K\L, N1, P1*, Q, R; L; N1; P1*; Q; R; and R\R1 (Table S3A).

We sequenced the first hypervariable segment (HVS1) of the mtDNA control region in 1,428 samples. Variable positions were determined from position 16,024 to 16,383, as previously described

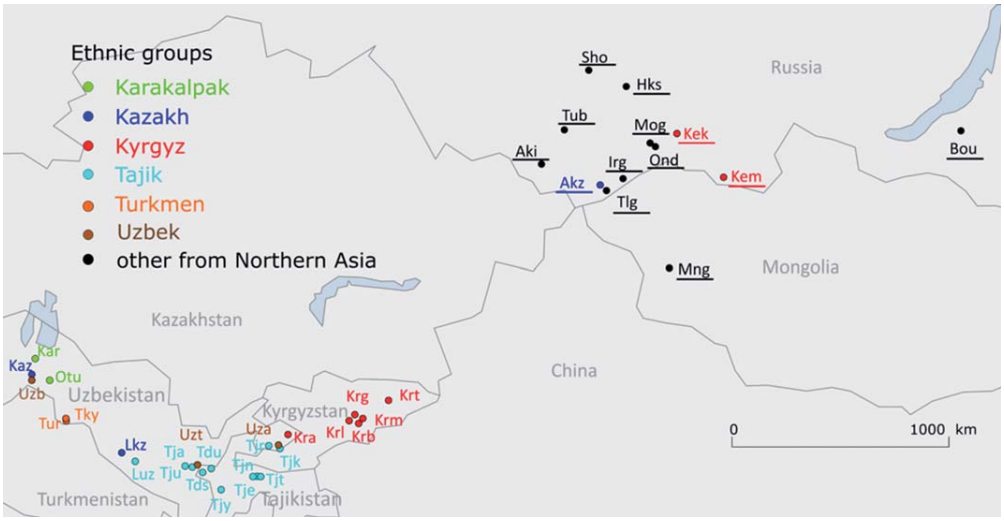


FIGURE 1 Geographic map of the studied populations in Central and Northern Asia. Populations are colored based on their ethnic group affiliation. For some ethnic groups in Northern Asia, only one population was sampled per ethnic group; such populations are shown in black. Outside of the 39 populations sampled by our team, six populations from Northern Asia were sampled by Dulik et al. (2011, 2012). Populations from Northern Asia are underlined. See Table S1 for the population code and ethnic group affiliation. In particular, the Tajik populations belong to the Indo-Iranian linguistic family and are cognatic, while the other populations belong to Turkic-Mongolic linguistic family and are patrilineal

(Quintana-Murci et al., 2004). We excluded from the analysis the C-tract length variation from position 16,179 to 16,194.

We obtained the mitochondrial haplogroup of each individual with the algorithm implemented in Haplogrep (Kloss-Brandstätter et al., 2011) which refers to Phylotree build 16 (van Oven & Kayser, 2009). For 460 samples, information provided by the HVS1 was not enough to determine haplogroups; therefore, we genotyped by Snapshot additional variable positions distributed in three different multiplexes. As it was still not enough for 34 samples, we added information provided by the HVS2 (from position 0 to 574). As for Y-haplogroups, not all samples were typed for the same positions (details in Table S2). Additionally, for two Kyrgyz populations (Krb and Krl, 57 samples), complete mitochondrial sequences were produced and used in the haplogroup determination. These complete mitochondrial genomes were sequenced on an Illumina Genome Analyzer IIx using a method described in (Barbieri et al., 2013). Due to diverse levels of haplogroup resolution between studies, we grouped haplotypes in five categories: L\M, N; M\D; D; N\R, and R; and 18 samples remained undetermined (Table S3A).

All raw genetic data are reported in Table S2. We released in GenBank the 1,428 HVS1 sequences (accession numbers from KX101238 to KX102614) and the 57 complete mtDNA sequences (accession numbers from KX675270 to KX675326).

2.3 | Statistical analysis

We performed a principal component analysis (PCA) on the haplogroup frequencies per population (Table S3A) using the R *ade4* package (Dray & Dufour, 2007; R Core Team, 2014). For both mtDNA and Y chromosome haplogroup frequencies, the two first Principal Components are presented as Supporting Information 1A and B. However, we analyzed only the first principal component (Table S3B), based on a Scree test (Cattell, 1966) (Supporting Information 2). Additionally, in a more traditional way, we performed Correspondence Analysis (ca package in R; Nenadic & Greenacre, 2007) on the haplogroup frequencies for both markers (Supporting Information 1C and D).

We calculated F_{ST} (Wright, 1950) and R_{ST} (Slatkin, 1995) distances between populations with Arlequin v3.5 (Excoffier, Laval, & Schneider, 2007), using Y-STR haplotypes from all the 45 populations. Only the results obtained for F_{ST} are presented in the Main Text, as F_{ST} is expected to have a lower variance at a large geographical scale (Hardy, Charbonnel, Fréville, & Heuertz, 2003). We made R_{ST} results available in the Supporting Information, and the correlation between F_{ST} and R_{ST} in Supporting Information 3B. For the HVS1 sequences from the 39 populations, F_{ST} values were computed with the pairwise differences model and the Kimura 2-parameters model, using a transition/transversion ratio of 10 and an alpha (gamma shape parameter) of .26 (Heyer et al., 2009). Only the latter model is presented in the Main Text, as both estimators were highly correlated (see Supporting Information 3A). F_{ST} values were visualized using a classical MultiDimensional Scaling (MDS) analysis with R in two dimensions, and to estimate how accurate this transformation was, we computed the Spearman's rank-sum correlation *rho* between the Euclidian distances calculated for pairs of populations in the MDS plot and their F_{ST} distance. We compared the Euclidian distances between Kyrgyz populations to the ones between Tajik populations using a one-tailed Mann–Whitney's *U* test. Moreover, to test the presence of clusters of populations, we ran two-tailed and one-tailed Mann–Whitney's *U* tests over the population coordinates on the MDS first dimension.

For the 39 populations sampled in our laboratory, we performed correlations between geographical and genetic F_{ST} distances, computed over Y-STRs or HVS1, using Mantel's tests implemented in the R *vegan* package (Oksanen et al., 2016). To estimate the effect of ethnic group affiliation on the correlation, partial Mantel's tests from the same package were performed, correcting for ethnic group affiliation. We converted the geographical distances in a log scale (adding 1 km to all distances). The genetic distances were expressed as $F_{ST}/(1 - F_{ST})$. We performed 10,000 permutations and estimated Spearman's correlations, being aware of the potential overestimation of the p-values obtained from these tests (Guillot & Rousset, 2013). In order to determine how the overall genetic diversity is distributed within and between populations, we performed various analyses of molecular variance (AMOVA) (Excoffier, Smouse, & Quattro, 1992) on F_{ST} values using Arlequin v3.5. For the effect of geography, we considered two groups: central *versus* northern populations. For the effect of ethnicity, we considered the 16 groups listed in Table S1.

For each of the six ethnic groups composed by more than one population sampled in our laboratory, we performed an AMOVA in order to assess the intragroup differentiation level (therefore, with two levels: within populations and among populations within the ethnic group). Supplementary AMOVAs were performed for Kazakh and Kyrgyz, distinguishing Central and Northern Asian populations.

We estimated the split times between populations and the male population effective sizes for the five Turko-Mongol Inner Asian ethnic groups on the Y-STR haplotypes using a coalescent and MCMC-based approach with BATWING (Wilson, Weale, & Balding, 2003). We included the two Kazakh populations from Dulik. For the Tajik ethnic group, we only estimated the effective population size, and not the split times as we do not know the historical age of this group. The program assumes that the populations under study have diverged from an ancestral population, had the same growth rate, and had not exchanged migrants after their split. We used a generation time of 30 years (Tremblay & Vézina, 2000), and a fixed but different mutation rate for each STR (Willems, Gymrek, Poznik, Tyler-Smith, & Erlich, 2016). We also computed estimations with a fixed mutation rate of .0028 for all STRs, as it is the mean of the eight individual rates. We chose a broad uniform distribution between 100 and 100,000 for the initial population size for the ethnic group (N) and computed 110,000 MCMC samples while the first 10,000 were discarded. We allowed 200 changes in model parameters between sampling occasions (Nbetsamp), and 100 changes to the tree attempted between changes in the model parameters (treebetN). We estimated the effective population size for the ethnic group (Ne) as the mean of the various N computed over the runs. Within ethnic group, we determined the age of first split as the average age over the 100,000 runs of the oldest merging event. All the BATWING results are presented in Table S4.

Using Python3.0 (https://www.python.org/download/releases/3.0/), we computed the following summary statistics on the Y-STRs dataset including the additional populations from the literature, after having removed haplotypes with missing data: the haplotypic heterozygosity (H) (Nei, 1978), the mean number of individuals sharing the same combination of 8 Y-STRs, called "haplotype" (C), and the proportion of haplotypes observed only once in the population (proportion of singletons Ps). When we compared the estimators of genetic diversity between cognatic and patrilineal populations, we used a one-tailed Mann–Whitney's *U* test, as we expected patrilineality to decrease genetic diversity. When we compared patrilineal populations from Central *versus* Northern Asia, we used a two-tailed Mann–Whitney's *U* test, as we had no prior expectations. The genetic diversity within population was calculated as the mean number of pairwise differences π with Arlequin v3.5, integrating Y-STRs haplotypes loci with less than 5% of missing data (Excoffier et al., 2007) and was compared with two-tailed Mann–Whitney's *U* tests.

3 | RESULTS

3.1 | Inner Asian populations in a broad geographical context

We first used the haplogroup data to investigate the genetic background of the studied Inner Asian populations and their proximity to reference populations from South Europe and East Asia.

We performed a PCA on the haplogroup frequencies for the Y chromosome and mtDNA, respectively, and plotted the first principal component (PC1) for each sex-specific marker (Figure 2). Note that PC1 explains more variation for the mtDNA (49.9%) than for the Y chromosome (18.4%). For both markers, the populations from South Europe and East Asia cluster together, respectively, at two opposite poles, with the studied Inner Asian populations all situated in between these two poles. At this broad geographical scale, populations from Inner Asia show an overall good and significant correlation of their genetic relationship to reference populations between the Y chromosome and mtDNA data (Spearman's *rho* = .70, *p*-value = 2×10^{-6}). This suggests no major large-scale sex-specific events during the original settlement of this area or subsequent migrations of populations.

As for the general relationship between populations, we observed that Tajik and Uzbek populations appear closer to South European populations, while Kyrgyz, Kazakh, and Northern Asian populations (except the Shor ethnic group) tend to be closer to East Asian populations (Figure 2). The two Karakalpak and one Turkmen population have an intermediate position, but as only two populations were sampled per ethnic group, it is difficult to make a general conclusion. The first dimension of the complementary correspondence analysis based on haplogroup frequencies confirmed the proximities observed on the PCA-PC1 plots for both the Y chromosome and mitochondrial DNA (Supporting Information 1C and D). However, this analysis distinguished more clearly the East Asian populations from the rest of the dataset: the Southern European and all the Inner Asian populations.

We then looked more closely at particular haplogroups of interest. Notably, two of the Y haplogroups were associated to high reproductive success of some male lineages in our area of interest (Balaresque et al., 2015). The C-M217 subhaplogroup, previously found to be frequent in Kazakh and Mongol, was for example attributed to Genghis Khan's and Giocangga's lineage (Abilev et al., 2012; Xue et al., 2005). In our study, the C haplogroup represents the major haplogroup in seven populations (frequency $\geq$50% in the Mongol, Buryat, all Kazakh and some Kyrgyz populations); however, we were not able to determine whether these individuals carried the C-M217 subhaplogroup. We also looked at the R1 haplogroup, which is widespread in Inner Asia and is of interest for the settlement of the Americas; indeed, the most common Native American haplogroup is Q, the sister lineage of R1. In our study, R1 has an average frequency of 33% per population, with a maximum of 81% found in a Kyrgyz population from Central Asia.

On the mtDNA, we found a large diversity of haplogroups detailed in Table S2. Most of them are supposed to have originated in Europe, East Asia or South Asia (A, B, C, D, F, G, Y, M, HV, JT, UK, I, W, and N) (Underhill & Kivisild, 2007). We even observed 21 samples with haplogroups affiliated to the Americas (A2, B2, and C1 haplogroups), and 16 within the African L haplogroups. We also observed two haplogroups thought to have originated in Central Asia: D4c and G2a (Comas et al., 2004). However, these are carried by only 12 and 36 individuals from 9 and 16 populations, respectively, among the 1,410 individuals from this study. Another interesting case is the U haplogroup that includes a subhaplogroup carried by the 24,000-old Upper Paleolithic man found in Siberia (Raghavan et al., 2014) but out of any known modern branch of the haplogroup. Moreover, the U haplogroup is documented to be rare or absent in modern Eastern Siberia, where the Paleolithic man originated, and more frequent in Central Asia and West Eurasia. Consistently, we found a low proportion of individuals carrying this haplogroup among Northern Asian populations (less than 5.5% in the South Siberian populations) while almost 13% of individuals carried it in the Mongolian population, 13% within Central Asian Turko-Mongol, and 18% within Tajik (with a maximum of 33% reached for one particular Tajik population).

Because Y chromosome and mtDNA haplogroups are defined by diagnostic markers and therefore represent a somehow biased representation of the genetic diversity, we further chose to study eight Y-STRs for the Y chromosome and the HVS1 region for the mtDNA in order to examine more closely the influence of sex-specific behaviors of genetic diversity.

3.2 | Relationships among Inner Asian populations

As previously described in the literature for other geographical areas (Lippold et al., 2014; Wilkins, 2006), Inner Asian populations are overall more differentiated on the Y chromosome (45 populations; F_{ST} = .156; *p*-value < .001) than on the mtDNA (39 populations; F_{ST} = .045; *p*-value < .001). To explore the genetic relationships between populations, we performed a MDS analysis on F_{ST} genetic distances between populations (Figure 3). The MDS analysis based on R_{ST} genetic distance for the Y chromosome is available in Supporting Information 4. For the

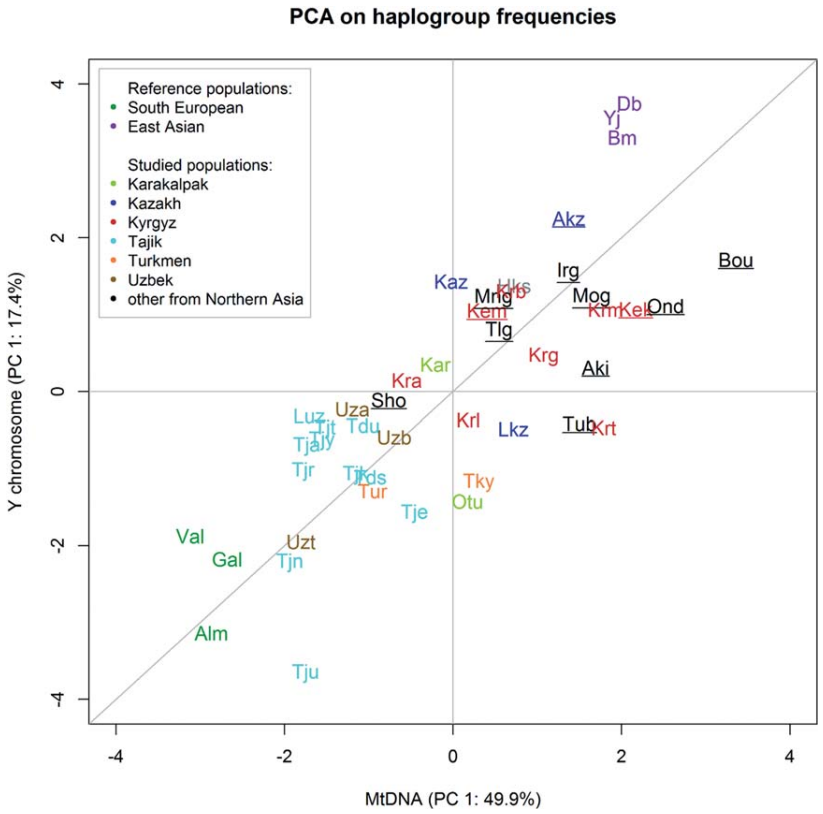


FIGURE 2 Genetic proximities of the studied Inner Asian populations to reference populations from South Europe and East Asia, based on Y chromosome and mtDNA haplogroups frequencies. Shown is the first principal component obtained from the Y chromosome and mtDNA, with the percentage of variance explained by each axis written in parenthesis. The $x = 0$, $y = 0$, and $x = y$ lines are represented in gray. The reference populations are from Santos et al. (2014), Val (Valencia), Gal (Galicia), Alm (Almeria) in Spain, and from Wang et al. (2014), Yj (Hekou Town), Db (Danba), Bm (Bamei Town) in County of Sichuan, China. Populations from Northern Asia are underlined

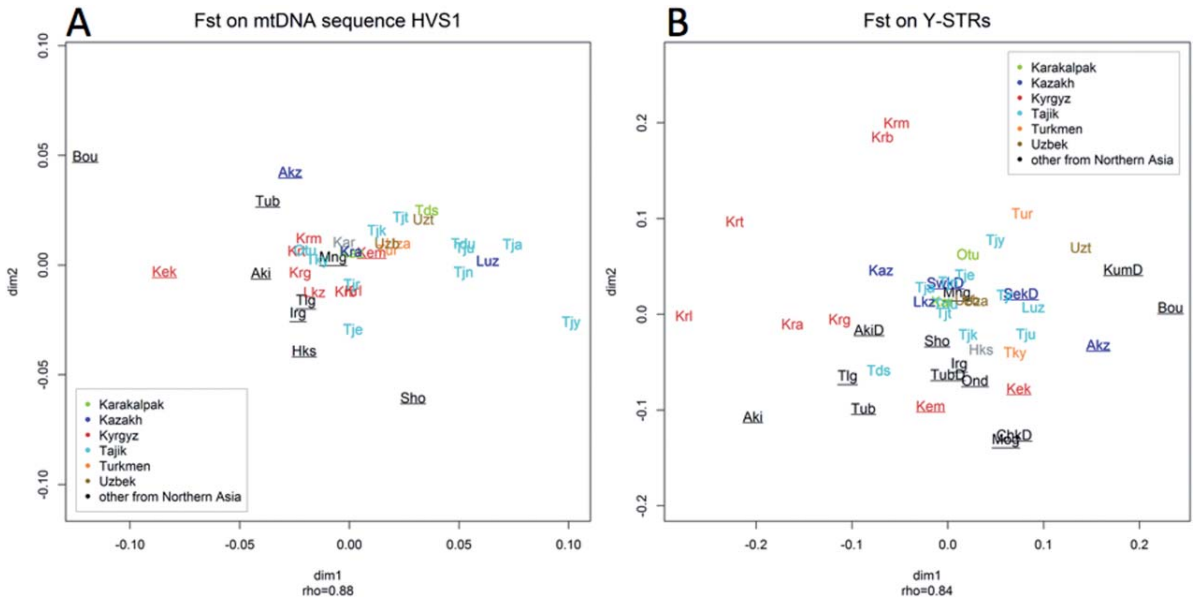


FIGURE 3 Multidimensional scaling (MDS) analysis based on pairwise F_{ST} between populations. Genetic distances were computed over mitochondrial HVS1 sequences for 39 populations (A) and over Y-STRs for 45 populations (B). Spearman's coefficient of correlation (*rho*) was calculated between the matrix of pairwise F_{ST} values and the scaled matrix. See Table S1 for the population code and ethnic group affiliation. Populations from Northern Asia are underlined

TABLE 1 Hierarchical AMOVA and fixation indexes based on ethnic group affiliation and/or geography

Grouping	No. groups	No. population	Source of variation	Percentage of variation		Fixation indexes		
				Y chr	*mtDNA*	*Y chr*	*mtDNA*	
Geography	2	39	Among groups	1.58	2.04	0.0158	0.0204	F_{CT}
			Among populations within groups	15.09	3.56	0.1533	0.0363	F_{SC}
			Within populations	83.33	94.40	0.1667	0.0560	F_{ST}
Geography *Kyrgyz and Kazakh excluded*	2	28	Among groups	1.99	2.90	0.0200	0.0290	F_{CT}
			Among populations within groups	11.34	3.81	0.1157	0.0392	F_{SC}
			Within populations	86.67	93.30	0.1333	0.0670	F_{ST}
Ethnicity *Inner Asia*	16	39	Among groups	5.18	2.66	0.0518	0.0289	F_{CT}
			Among populations within groups	11.18	2.11	0.1179	0.0193	F_{SC}
			Within populations	83.64	95.22	0.1636	0.0477	F_{ST}
Ethnicity *Central Asia*	6	26	Among groups	5.78	0.97	0.0578	0.0097	F_{CT}
			Among populations within groups	9.69	1.86	0.1028	0.0188	F_{SC}
			Within populations	84.53	97.17	0.1547	0.0283	F_{ST}
Ethnicity *Northern Asia*	12	13	Among groups	19.67	N.S.	0.1966	N.S.	F_{CT}
			Among populations within groups	N.S.	6.25	N.S.	0.0620	F_{SC}
			Within populations	81.40	94.45	0.1861	0.0555	F_{ST}

N.S. is nonsignificant *p*-value (>.001). The numbers of considered groups and populations are indicated.

mtDNA, we excluded the Ondar population (Ond) as it was a clear outlier (Supporting Information 5). For the mtDNA F_{ST} distances, as for the haplogroup data, we observed a separation between the Indo-Iranian cognatic and the Turko-Mongol patrilineal populations on the first dimension. Indeed, the Tajik populations (the only cognatic from our study) cluster on the right side of the plot while the other populations, all patrilineal, are grouped together in the middle and left side of the plot (Figure 3A) (one-tailed Mann–Whitney's *U* test performed between the first dimension coordinates of Tajik and patrilineal populations, *p*-value $= 3 \times 10^{-6}$). We also observed a tendency of Northern Asian populations to be further away from cognatic populations than the Central Asia ones (one-tailed Mann–Whitney's *U* test on the first dimension coordinates, Northern Asian having smaller values than Inner Asian patrilineal populations: *p*-value $= 4 \times 10^{-3}$). For the Y chromosome F_{ST} distances (Figure 3B), we observed another pattern: Tajik populations do not cluster in one side of the plot, when compared with patrilineal populations (two-tailed Mann–Whitney's *U* test on the first dimension coordinates revealed a non-significant *p*-value = .457). On contrary, Tajik populations cluster at the core of the plot, with a mean Euclidian distance between two Tajik populations reaching .076, whereas Turko-Mongol populations from Central and Northern Asia are peripherally scattered, with Kyrgyz populations particularly widespread (mean distance = .208). This disparity between clustered Tajik populations and scattered Kyrgyz populations was significantly supported by a one-tailed Mann–Whitney's *U* test on the distances computed within these two groups (*p*-value $= 7 \times 10^{-10}$).

Because the sampled populations differ for several parameters such as their geographical location, their ethnic group affiliation or their social organization, each of these features or a combination of them could impact the pattern of genetic differentiation by limiting recent gene flow between some populations and favoring exchanges between others. Consequently, we specifically explored the influence of each of these three features on the genetic differentiation between populations.

3.3 | Impact of geography and ethnicity on population differentiation

Geography is a main feature shaping the genetic differentiation of human populations (Li et al., 2008; Ramachandran et al., 2005). To assess the role of geography in shaping the genetic structure in this area, we performed an AMOVA, grouping the 39 populations in two main geographical groups: Central *versus* Northern Asia. We found that geography explains slightly more the patterns of genetic differentiation for the mtDNA (2.04%, *p*-value < .001) than for the Y chromosome (1.58%, *p*-value < .002) (Table 1). Consistently, geographical and genetic distances (F_{ST}) between populations, while being both significant, are more correlated for the mtDNA (Spearman's *rho* = .35,

Mantel's test p-value $< .0001$) than for the Y chromosome (Spearman's $rho = .17$; Mantel's test p-value $< .009$) (Supporting Information 6; Table S5A and B for Y chromosome R_{ST} analysis). When testing this correlation separately within region, we found that for mtDNA, distances are not correlated within Central Asia or Northern Asia (Mantel's test p-value $= .79$ and .45, respectively). On the contrary for the Y chromosome, genetic and geographical distances are correlated both within Central Asia (Spearman's $rho = .17$; Mantel's test p-value $= .037$) and within Northern Asia (Spearman's $rho = .45$; Mantel's test p-value $= .034$). The impact of geography on the genetic diversity is therefore stronger for maternal than for paternal markers at the whole geographical scale, but within Central Asia and Northern Asia, this relationship holds true only for paternal markers.

Because populations from the same ethnic group tend to be geographically closer than populations from different ethnic groups (we calculated an average of 483 km between populations from the same ethnic group against 1,376 km for populations from two different ethnic groups), we further assessed the genetic-geographical correlation correcting for ethnic group affiliation using a partial Mantel's test. The correlations for the mtDNA and the Y chromosome both decrease but are still significant (Spearman's $rho = .32$, Mantel's test p-value $< .0002$ and Spearman's $rho = .12$, Mantel's test p-value $< .04$, respectively). This suggests a direct relationship between geography and genetic diversity at the whole region level, despite the ethnic group structuration.

We then performed an AMOVA based on ethnic group affiliation alone. Differences among ethnic groups significantly explain part of the genetic variance both on the Y chromosome and the mtDNA, but with a higher fraction explained for the Y chromosome as compared to the mtDNA (5.18 *vs* 2.66%, p-value $< .001$) (Table 1). The same pattern is observed within Central Asia (5.78 *vs* .97%, p-value $< .001$) and is even stronger in Northern Asia (19.67%, p-value $< .001$ *vs* −.69%, non-significant p-value $= .56$). The impact of ethnicity on the genetic diversity is therefore consistently more important for paternal than for maternal markers.

3.4 | Population relationships within ethnic groups

Additionally, we explored the genetic differentiation within each ethnic group, regardless of geography, both for the Y chromosome and the mtDNA, using AMOVAs. Overall, these differentiation levels are always found to be higher on the Y chromosome as compared to the mtDNA, even though the ratio largely varied among ethnic groups (Table 2 and S5C for Y chromosome R_{ST} analysis). Indeed, Tajik populations present the higher value for mtDNA differentiation (3.16%, p-value $< .001$) but the lower value for the Y chromosome (7.42%, p-value $< .001$), whereas the five other ethnic groups show an average mitochondrial DNA intra-group differentiation of 1.78% and an average Y chromosomal value of 15.18%, with a lower standard deviation found for the mtDNA (.49% against 6.81%). The variance of these estimates between populations appears to be larger for the Y chromosome than for the mtDNA.

TABLE 2 Percentage of F_{ST} variance explained by differences within each ethnic group, for both Y chromosome and mtDNA

		Percentage of intragroup differentiation	
Ethnic group	***N***	***Y chr***	***mtDNA***
Karakalpak	2	7.45	N.S.
Kazakh	3	19.38	2.35
Central Asian Kazakh	2	N.S.	N.S.
Northern Asian Kazakh	1	–	–
Kyrgyz	8	18.69	2.01
Central Asian Kyrgyz	6	17.02	N.S.
Northern Asian Kyrgyz	2	N.S.	10
Tajik	11	7.42	3.16
Turkmen	2	22.11	1.52
Uzbek	3	8.25	1.25

N.S. is nonsignificant p-value ($>.001$). The number of populations by group is indicated.

For Kazakh and Kyrgyz, the ethnic groups found both in Central and Northern Asia, we explored further the relationship between the two geographical areas. For the mtDNA, both ethnic groups present an overall low intragroup differentiation (2.35 and 2.01%, respectively, p-value $< .001$), suggesting that, in both cases, populations from Central and Northern Asia are not strongly differentiated. For the Y chromosome, on the contrary, both ethnic groups present a high level of intragroup differentiation (19.38 and 18.64%, respectively, p-value $< .001$). In Kazakh, the two Central Asian populations present a nonsignificant differentiation, and there is only one sampled population in Northern Asian. Consequently, the major part of the Kazakh differentiation seems to be due to differences between the two geographical groups. In Kyrgyz, however, the differentiation is high within Central Asian populations (17.02%, p-value $< .001$) but nonsignificant within Northern Asia. This suggests that, for the Y chromosome, most, if not all, Kyrgyz intra-group variation is due to differentiation among the six Central Asian populations.

3.5 | Ethnogenesis: case study from the Y chromosome

Given that, in the studied populations, ethnic group affiliation significantly explains the Y chromosome genetic structure, and that ethnic group affiliation is transmitted from father to sons in such patrilocal populations, we investigated the process of ethnogenesis using STR data on the Y chromosome (Table S4). First, we estimated the male effective population size (Ne) of each ethnic group and found that Tajik ethnic group have by far the higher Ne, around 11,000 individuals, while the other ethnic groups have effective population sizes estimated between 1,300 and 3,800 individuals (Figure 4A). Secondly, for Turko-Mongol groups, we calculated the split time between pairs of populations from the same ethnic group and focused on the oldest event of split that occurred within ethnic group. We found that the first split

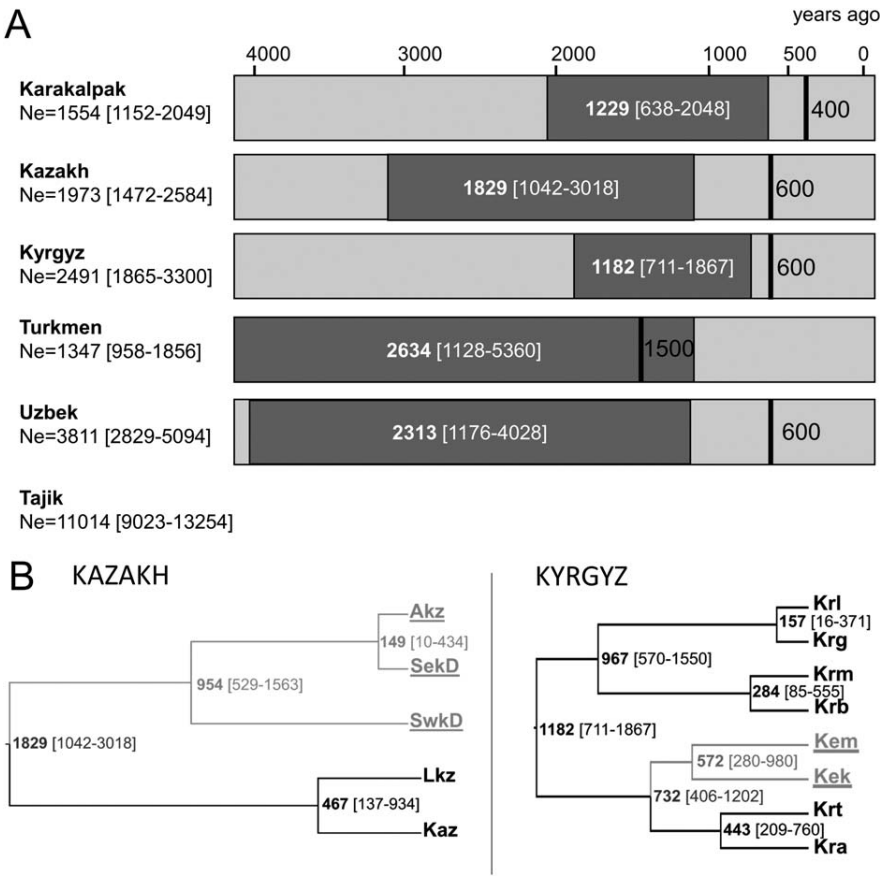


FIGURE 4 Split times for each ethnic group and phylogenetic trees for Kazakh and Kyrgyz ethnic groups. (A) Below each ethnic group name is written the estimated effective population size (Ne) and its confidence interval. The box plots present the estimated time of the first split between populations within each ethnic group, based on genetic data, and its confidence interval (in gray), as well as the time of origin of each ethnic group based on historical records (in black). The historical estimate for the Kyrgyz ethnic group correspond to groups living in Kyrgyzstan. (B) Phylogenetic tree of the Kazakh and Kyrgyz ethnic group (on the left and right panel, respectively). The trees were reconstructed node by node from split events between pairs of populations. Times of split on each node were estimated in generations but are presented in years, given a generation time of 30 years. The confidence interval represents the values including 95% of the distribution. The populations from Central Asia are colored in black and populations from Northern Asia are in light gray

was always older than 1,000 years ago, revealing a relatively deep ethnic group history. Kyrgyz and Karakalpak seem to have the more recent history (around 1,200 years ago), while for other groups, the value is always older than 1,800 years ago, with Turkmen and Uzbek ages overpassing 2,300 years ago. However, the confidence interval for these split times is often quite large.

Furthermore, we aimed to assess the putative northern (Altaic) origin of some Central Asian Turko-Mongol populations, focusing on Kazakh and Kyrgyz (Yunusbayev et al., 2014). The male phylogenetic tree for Kazakh revealed that central and northern populations respectively cluster together (Figure 4B), which prevents us from inferring where this ethnic group originated. For Kyrgyz, we found that the northern populations form a sub-branch that split from other central populations around 700 years ago (Figure 4B).

3.6 | Impact of social organization on genetic diversity

To explore the impact of social organization on the Y chromosome genetic diversity, we compared, through four estimators, the genetic diversity of patrilineal populations to that of cognatic populations (Table S6). Patrilineal populations appear to have less haplotypic heterozygosity (H = .87 *vs* .95, one-tailed Mann–Whitney's *U* test *p*-value = .008), more individuals sharing the same haplotype (C = 2.11 *vs* 1.51, one-tailed Mann–Whitney's *U* test *p*-value = .015), and a smaller proportion of haplotype singletons (Ps = .39 *vs* .53, one-tailed Mann–Whitney's *U* test *p*-value = .012) than cognatic populations. On the contrary, no significant differences are observed on π values (two-tailed Mann–Whitney's *U* test, *p*-value = .38). When dividing patrilineal populations into Central Asia and Northern Asia populations, the same pattern is observed for each region, and the two groups do not differ from each other for any estimator (two-tailed Mann–Whitney's *U* test p-value > .07).

Because Uzbek populations are thought to have shifted from being nomadic patrilineal herders to a more sedentary cognatic agriculturist lifestyle at the 16th century (Soucek, 2000), a less important genetic signature of a patrilineal organization is expected in these populations. In this context, we compared the genetic diversity of patrilineal populations with that of cognatic populations, excluding the three Uzbek

populations. At the whole Inner Asia and Central Asia level, the observed tendencies did not change. However, we found less genetic diversity in the Central Asian patrilineals without Uzbeks than Northern Asian patrilineal populations (one-tailed Mann–Whitney's *U* test *p*-value = .02; on average, .83 *vs* .90), more pairwise differences for the Y chromosome (one-tailed Mann–Whitney's *U* test *p*-value = .01; on average, 3.30 *vs* 4.26), and less individuals sharing the same Y-haplotype (one-tailed Mann–Whitney's *U* test *p*-value = .04; on average, 2.33 *vs* 2.01).

4 | DISCUSSION

4.1 | Haplogroup data

Based on haplogroup data, we can distinguish two genetic profiles of the studied populations, consistent with known historical data. Indeed, Indo-Iranian-speaking populations (*i.e.*, Tajik), are genetically closer to South European populations, while most of the Turkic-Mongolic-speaking ethnic groups (Kyrgyz, Kazakh, South Siberian, and Mongolian) are closer to East Asian populations. These proximities were already described with autosomal data (Martínez-Cruz et al., 2011). Note that these differences in linguistic affiliation and genetic proximities correlate perfectly with differences in social organization: Indo-Iranian-speaking populations are cognatic while Turkic-Mongolic-speaking populations are patrilineal. Interestingly, the Uzbek populations, which are thought to have shifted from being nomadic patrilineal herders to a more sedentary cognatic agriculturist lifestyle at the 16th century (Soucek, 2000) are closer to South-European populations than to East Asian. This is consistent with the fact that Uzbek are thought to be a conglomerate of Turko-Mongol populations originated from East and of a Turkic group already present in Uzbekistan that was highly admixed with Iranians (Soucek, 2000). In turn, Turkmen populations, documented as being originally Indo-Iranian-speaking populations who later experimented a linguistic shift to a Turkic-Mongolic language (Heyer & Mennecier, 2009) are also found in an intermediate position. Such a linguistic shift could have resulted from the movements of eastern nomadic Turkic-Mongolic-speaking groups westwards and their subsequent incorporation of local Turkmen populations. Under this scenario, we could imagine that men predominantly brought their genes (and their language), while women would have kept their local Indo-Iranian background. However, based on haplogroup data, we see no trace of such a Turko-Mongol replacement or introgression of neither Y chromosomes nor mtDNA into Turkmen populations. As we only sampled two populations, it would be interesting to collect more data on Turkmen.

If the original settlement of Inner Asia had been strongly sex-specific, or that subsequent sex-specific movements of populations had occurred, we would expect contrasted genetic backgrounds between the Y chromosome and the mtDNA. Yet, based on haplogroup data, all populations showed consistent genetic relationships between unilateral markers, suggesting they did not have vastly contrasted population sources for their paternal and maternal ancestors. However, the genetic proximities of populations were not consistent when comparing the MDS plots based on Y-STRs *versus* the HVS1 region (Figure 3). This difference can result from the nature of the markers: indeed Y-STRs and the HVS1 region have a higher mutation rate (2.8×10^{-3}/generation for Y-STRs (Kayser & Sajantila, 2001) and 1.17×10^{-5}/generation for HVS1 (Heyer et al., 2001) than SNPs (10^{-8}/generation (Altshuler et al., 2010). Therefore, we expect both HVS1 and Y-STRs to be informative about more recent time scales than SNPs-based haplogroups, which could explain the observed discrepancy.

4.2 | Patrilocality and patterns of migration

Overall, based on Y-STRs and HVS1, we find that Inner Asian populations are about three times more differentiated on the Y chromosome than on the mtDNA (F_{ST} = .156 and .045, respectively), suggesting that there are more exchanges of women than men between populations. This homogenization of the mtDNA diversity relative to the Y chromosome is consistent with the fact that these populations are patrilocal, *i.e.*, women migrate to their husbands' place of origin after the wedding.

When partitioning these genetic differences thanks to hierarchical AMOVAs, we found that geography is a main determinant of the mtDNA differentiation (significantly explaining 2.04% of the genetic variance). This pattern is mainly driven by differences between central and northern populations, as the genetic-geographical correlation is highly significant overall (*p*-value $< 10^{-4}$) but not within each region (*p*-value $\geq$ .45). This suggests that women are mostly migrating at a local scale, and not often between Central and Northern Asia, even though some ethnic groups are found in both areas. While the results based on Mantel's test may be biased (Guillot & Rousset, 2013), this pattern of small range migrations is also consistent with the geographical gradient observed on the MDS. Moreover, the same migration pattern was observed in Lesotho, where women were documented to migrate at shorter distances than men (Marks, Levy, Martinez-Cadenas, Montinaro, & Capelli, 2012).

For the Y chromosome, the genetic differentiation is high (F_{ST} = .156) and geography significantly explains only 1.58% of the genetic variance, suggesting there are limited exchanges of individuals between populations. The genetic-geographical correlation is significant, but three times lower than for the mtDNA. This low correlation may be due to genetic proximities within ethnic group, independently of their geographical distribution, and therefore could reflect a shared and quite recent common ancestry of individuals from the same ethnic group, even when they are geographically distant. Alternatively, it could be due to preferential migrations of men within ethnic groups, even between geographical zones. Throughout our analyses, we found that ethnicity is the main determinant of the Y chromosome variation (explaining 5.18% of the genetic variance). Thus, ethnicity clearly acts as a barrier for movement of men between populations from different ethnic groups, more strongly so in Northern Asia than in Central Asia (19.67 and 5.78% of the variance explained by ethnicity, respectively).

4.3 | Ethnicity and ethnogenesis process

We then estimated the first genetic split time occurring in each Turko-Mongol ethnic group based on Y-STRs. We observed a time lapse

between the historical records of ethnogenesis and the first genetic split time for all six ethnic groups. This is in concordance with the process of ethnogenesis by fusion proposed by Barth (Barth, 1969). In this model, a new ethnic group is formed by the fusion of populations that are not close genealogical relatives, for putative social and cultural reasons. Therefore, the social age of the ethnic group is necessarily more recent than the estimated biological age of the conglomerated populations. This ethnogenesis process contrasts with the idea of a biological origin of each ethnic group, in which populations split from a unique ancestral population. A limitation of our approach is that we used a model without migration nor expansion, leading us to underestimate the split times between populations. Therefore, the time lapse between biological and historical ages could even be larger, and our results are conservative.

For Kazakh and Kyrgyz that are found in both Central and Northern Asia, the shape of the reconstructed genealogical trees can inform us on the history of their ethnic group. For Kazakh, their tree suggests an old separation between the Central and Northern Asia populations, consistent with the intra-ethnic group differentiation mainly driven by differences between the two geographical areas. On the contrary, Kyrgyz seems to have experienced a recent split between the geographically separated populations, in agreement with the small part of the intra-ethnic group differentiation found between areas. Therefore, we can speculate that Kyrgyz originated in Central Asia, and subsequently migrated to Northern Asia. However, another hypothesis, supported by historical records (Yunusbayev et al., 2014) could be that Kyrgyz originated in Northern Asia and then migrated to Central Asia around 700 years ago, where some of them incorporated local populations present in Central Asia before their arrival. Therefore, the Central Asian Kyrgyz actually sampled would present various degree of admixture between Central Asian and Northern Asian Kyrgyz. However, as we only sampled two northern *versus* six central Kyrgyz populations, we cannot exclude that we are missing some genetic variation in Northern Asia and therefore prefer not to conclude on a geographical location of origin for the Kyrgyz ethnic group.

4.4 | Patrilineality

We searched for a genetic signature of patrilineality by contrasting patterns of genetic diversity in patrilineal populations to those observed in the Tajik cognatic populations. We found a reduced Y-haplotypic heterozygosity, a reduced number of Y-haplotype singletons and an increased number of individuals sharing the same Y haplotype in patrilineal when compared to cognatic populations. However, we did not find any difference in terms of mean number of pairwise differences between groups. This genetic pattern might reflect a structuration of these populations in paternal descent groups due to their patrilineal social organization (Chaix et al., 2007), resulting in reduced male effective population size. A descent group is defined as all the descendants of an ancestor, meaning all the individuals carrying the same Y haplotype. At the population scale, the impact of such a structuration is the rapid fixation of a different haplotype by group, leading to a high degree of polymorphism within population since each population is made of several groups of ancestry. However, this polymorphism can be reduced by the extinction of some of groups, driven by differential reproductive success. Thus, we expect a decrease in haplotypic heterozygosity, but not necessarily in the number of pairwise differences. We could also observe such a pattern if there was different demographic histories between patrilineal or cognatic populations. Yet, this was ruled out by a previous study (Aimé et al., 2014) showing no growth differences between groups. Note that if there was a difference in the initial male population size between cognatic and patrilineal populations, we would also observe a difference in terms of π, which is not the case here.

Even though the signature of patrilineality was found throughout Inner Asia, it seemed stronger in Central Asia compared to Northern Asia. Moreover, the Uzbeks present an empirical pattern similar to the cognatic populations, consistent with their shift to a sedentary lifestyle in the 16th century (Soucek, 2000), and their likely adoption of cognatic practices at that time. However, this hypothesis has to be further tested with simulations.

5 | CONCLUSIONS

At the scale of the whole Inner Asia, we have investigated the relationships of ethnic groups with European and Asian populations. We confirmed the proximity of Indo-Iranian populations with Europeans and of Turko-Mongol populations with East Asians. Interestingly, we observed the same pattern for both mitochondrial DNA and the Y chromosome, suggesting an absence of strong sex-specific settlement. However, at a more recent time scale, we found the genetic diversity to be strongly impacted by descent rules: patrilineal populations show a much higher level of Y chromosome differentiation than the cognatics. For the Uzbek, a supposedly patrilineal group, we found a cognatic-like imprint, suggesting that the social organization shifted during the last 500 years. We also detected a higher effect of geography for women than for men, whereas ethnicity influenced more the Y chromosome diversity than the mitochondrial one. Finally, we explored the ethnogenesis of the Turko-Mongol ethnic groups and revealed that it is a cultural rather than biological process, with the fusion of populations that are not close genealogical relatives during the formation of a new ethnic group. These results provide insights into the demographic history of Inner Asian ethnic groups. In conclusion, our study highlights the impact of recent sex-specific cultural traits on the genetic diversity and illustrates the importance of jointly studying both unilateral markers.

ACKNOWLEDGMENTS

We wish to thank Pr. Mark Jobling for his help with the haplogroup analyses, Jon Wetton and Patricia Balaresque for their assistance in Y haplogroup determination, Alena Kushniarevich for advices on mtDNA haplogroup determination, Laure Bagait for her help on GeneMapper, Christine Harmant for technical help on Y SNP genotyping, Angélique Rocha and Livia Camargo for producing part of the mtDNA HVS1 data, and the anonymous reviewers for helpful

discussions and comments on the manuscript. The BATWING analyses were run on the Linux cluster of the "Muséum National d'Histoire Naturelle", Paris, France (administrated by Julio Pedraza from the Computational Biology Service Unit).

AUTHOR CONTRIBUTIONS

N.M. performed the analyses. N.M., L.S., and E.H. interpreted the data and wrote the article. T.H., P.M., A.A., C.S., F.N., B.C., and E.H. participated to the sampling of populations. M.G. extracted DNA. N. M., M.W., and P.E. produced the DNA data. R.L. calculated the geographical distances between populations. E.H. and L.S. designed the study.

REFERENCES

Abilev, S., Malyarchuk, B., Derenko, M., Wozniak, M., Grzybowski, T., & Zakharov, I. (2012). The Y-chromosome C3* star-cluster attributed to Genghis Khan's descendants is present at high frequency in the Kerey clan from Kazakhstan. *Human Biology*, *84*, 79–89. http://doi.org/10.3378/027.084.0106

Aimé, C., Verdu, P., Ségurel, L., Martinez-Cruz, B., Hegay, T., Heyer, E., & Austerlitz, F. (2014). Microsatellite data show recent demographic expansions in sedentary but not in nomadic human populations in Africa and Eurasia. *European Journal of Human Genetics*, *22*, 1201–1207. http://doi.org/10.1038/ejhg.2014.2.

Allentoft, M. E., Sikora, M., Sjögren, K. G., Rasmussen, S., Rasmussen, M., Stenderup, J., … Willerslev, E. (2015). Population genomics of Bronze Age Eurasia. *Nature*, *522*, 167–172. http://doi.org/10.1038/nature14507

Altshuler, D. M., Gibbs, R. A., Peltonen, L., Dermitzakis, E., Schaffner, S. F., Yu, F., … McEwen, J. E. (2010). Integrating common and rare genetic variation in diverse human populations. *Nature*, *467*, 52–58. http://doi.org/10.1038/nature09298

Balaresque, P., Manni, F., Dugoujon, J. M., Crousau-Roy, B., & Heyer, E. (2006). Estimating sex-specific processes in human populations: Are XY-homologous markers an effective tool? *Heredity*, *96*, 214–221. http://doi.org/10.1038/sj.hdy.6800779

Balaresque, P., Poulet, N., Cussat-Blanc, S., Gerard, P., Quintana-Murci, L., Heyer, E., & Jobling, M. A. (2015). Y-chromosome descent clusters and male differential reproductive success: Young lineage expansions dominate Asian pastoral nomadic populations. *European Journal of Human Genetics*, *23*, 1413–1422. http://doi.org/10.1038/ejhg.2014.285.

Barbieri, C., Vicente, M., Rocha, J., Mpoloka, S. W., Stoneking, M., & Pakendorf, B. (2013). Ancient substructure in early mtDNA lineages of Southern Africa. *American Journal of Human Genetics*, *92*, 285–292. http://doi.org/10.1016/j.ajhg.2012.12.010

Barth, F. (1969). *Ethnic groups and boundaries. The social organization of culture difference (results of a symposium held at the University of Bergen, 23rd to 26th February 1967)*. Bergen/London: Universitetsforlaget/Allen & Unwin.

Besaggio, D., Fuselli, S., Srikummool, M., Kampuansai, J., Castrì, L., Tyler-Smith, C., … Bertorelle, G. (2007). Genetic variation in Northern Thailand Hill Tribes: Origins and relationships with social structure and linguistic differences. *BMC Evolutionary Biology*, *7*(Suppl 2), S12. http://doi.org/10.1186/1471-2148-7-S2-S12

Bosch, E., Calafell, F., Rosser, Z. H., Nørby, S., Lynnerup, N., Hurles, M. E., & Jobling, M. A. (2003). High level of male-biased Scandinavian admixture in Greenlandic Inuit shown by Y-chromosomal analysis. *Human Genetics*, *112*, 353–363. http://doi.org/10.1007/s00439-003-0913-9

Burton, M. L., Moore, C. C., Romney, A. K., Aberle, D. F., Barcelo, J. A., Dow, M. M., … Linnekin, J. (1996). Regions based on social structure. *Current Anthropology*, *37*, 87–123. http://doi.org/10.1086/204474

Bustamante, C. D., & Ramachandran, S. (2009). Evaluating signatures of sex-specific processes in the human genome. *Nature Genetics*, *41*, 8–10. http://doi.org/10.1038/ng0109-8

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245–276. http://doi.org/10.1207/s15327906mbr0102_10

Chaix, R., Quintana-Murci, L., Hegay, T., Hammer, M. F., Mobasher, Z., Austerlitz, F., & Heyer, E. (2007). From social to genetic structures in Central Asia. *Current Biology*, *17*, 43–48. http://doi.org/10.1016/j.cub.2006.10.058

Comas, D., Plaza, S., Wells, R. S., Yuldaseva, N., Lao, O., Calafell, F., & Bertranpetit, J. (2004). Admixture, migrations, and dispersals in Central Asia: Evidence from maternal DNA lineages. *European Journal of Human Genetics*, *12*, 495–504. http://doi.org/10.1038/sj.ejhg.5201160

Destro Bisol, G., Capocasa, M., & Anagnostou, P. (2012). When gender matters: New insights into the relationships between social systems and the genetic structure of human populations. *Molecular Ecology*, *21*, 4917–4920. http://doi.org/10.1111/mec.12001

Dray, S., & Dufour, A. B. (2007). The ade4 package: Implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*, 1–20. http://doi.org/10.1.1.177.8850

Dulik, M. C., Osipova, L. P., & Schurr, T. G. (2011). Y-chromosome variation in Altaian Kazakhs reveals a common paternal gene pool for Kazakhs and the influence of Mongolian expansions. *PLoS ONE*, *6*, e17548. http://doi.org/10.1371/journal.pone.0017548

Dulik, M. C., Zhadanov, S. I., Osipova, L. P., Askapuli, A., Gau, L., Gokcumen, O., … Schurr, T. G. (2012). Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. *American Journal of Human Genetics*, *90*, 229–246. http://doi.org/10.1016/j.ajhg.2011.12.014

Excoffier, L., Laval, G., & Schneider, S. (2007). Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, *1*, 47–50. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2658868&tool=pmcentrez&rendertype=abstract.

Excoffier, L., Smouse, P., & Quattro, J. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, *491*, 479-491. Retrieved from http://www.genetics.org/content/131/2/479.short.

Guillot, G., & Rousset, F. (2013). Dismantling the Mantel tests. *Methods in Ecology and Evolution*, *4*, 336-344. http://doi.org/10.1111/2041-210x.12018

Hammer, M. F., Mendez, F. L., Cox, M. P., Woerner, A. E., & Wall, J. D. (2008). Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genetics*, *4*, e1000202. http://doi.org/10.1371/journal.pgen.1000202

Hardy, O. J., Charbonnel, N., Fréville, H., & Heuertz, M. (2003). Microsatellite allele sizes: A simple test to assess their significance on genetic differentiation. *Genetics*, *163*, 1467–1482. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1462522&tool=pmcentrez&rendertype=abstract.

Heyer, E., Balaresque, P., Jobling, M. A., Quintana-Murci, L., Chaix, R., Segurel, L., … Hegay, T. (2009). Genetic diversity and the emergence

of ethnic groups in Central Asia. *BMC Genetics*, *10*, 49. http://doi.org/10.1186/1471-2156-10-49.

Heyer, E., Brandenburg, J. T., Leonardi, M., Toupance, B., Balaresque, P., Hegay, T., … Austerlitz, F. (2015). Patrilineal populations show more male transmission of reproductive success than cognatic populations in Central Asia, which reduces their genetic diversity. *American Journal of Physical Anthropology*, *157*, 537-543. http://doi.org/10.1002/ajpa.22739.

Heyer, E., Chaix, R., Pavard, S., & Austerlitz, F. (2012). Sex-specific demographic behaviours that shape human genomic variation. *Molecular Ecology*, *21*, 597-612. http://doi.org/10.1111/j.1365-294X.2011.05406.x

Heyer, E., & Mennecier, P. (2009). In F. d'Errico & J.-M. Hombert (Eds.) 6, 163–180, *Becoming eloquent*. Amsterdam: John Benjamins Publishing Company. http://doi.org/10.1075/z.152.

Heyer, E., Zietkiewicz, E., Rochowski, A., Yotova, V., Puymirat, J., & Labuda, D. (2001). Phylogenetic and familial estimates of mitochondrial substitution rates: Study of control region mutations in deep-rooting pedigrees. *American Journal of Human Genetics*, *69*, 1113–1126. http://doi.org/10.1086/324024

Jorde, L. B., Watkins, W. S., Bamshad, M. J., Dixon, M. E., Ricker, C. E., Seielstad, M. T., & Batzer, M. A. (2000). The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosome data. *American Journal of Human Genetics*, *66*, 979–988. http://doi.org/10.1086/302825

Kayser, M., Brauer, S., Weiss, G., Schiefenhövel, W., Underhill, P., Shen, P., … Stoneking, M. (2003). Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *The American Journal of Human Genetics*, *72*, 281–302. http://doi.org/10.1086/346065.

Kayser, M., & Sajantila, A. (2001). Mutations at Y-STR loci: Implications for paternity testing and forensic analysis. *Forensic Science International*, *118*, 116–121. http://doi.org/10.1016/S0379-0738(00)00480-1.

Keinan, A., Mullikin, J. C., Patterson, N., & Reich, D. (2009). Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nature Genetics*, *41*, 66–70. http://doi.org/10.1038/ng.303.

Kloss-Brandstätter, A., Pacher, D., Schönherr, S., Weissensteiner, H., Binna, R., Specht, G., & Kronenberg, F. (2011). HaploGrep: A fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human Mutation*, *32*, 25–32. http://doi.org/10.1002/humu.21382.

Krader, L. (1966). *People of Central Asia* (2nd ed., p. 26). Bloomington: Indiana University Publications.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., … Myers, R. M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science*, *319*, 1100–1104. http://doi.org/10.1126/science.1153717

Lippold, S., Xu, H., Ko, A., Li, M., Renaud, G., Butthof, A., … Stoneking, M. (2014). Human paternal and maternal demographic histories: Insights from high-resolution Y chromosome and mtDNA sequences. *Investigate Genetics*, *5*, 13. http://doi.org/10.1101/001792.

Maniatis, T., Fritsch, E. F., & Sambrook, J. (1982). *Molecular cloning: A laboratory manual* (Vol. *1*). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory.

Marks, S. J., Levy, H., Martinez-Cadenas, C., Montinaro, F., & Capelli, C. (2012). Migration distance rather than migration rate explains genetic diversity in human patrilocal groups. *Molecular Ecology*, *21*, 4958–4969. http://doi.org/10.1111/j.1365-294X.2012.05689.x

Martínez-Cruz, B., Vitalis, R., Ségurel, L., Austerlitz, F., Georges, M., Théry, S., … Heyer, E. (2011). In the heartland of Eurasia: The multilocus genetic landscape of Central Asian populations. *European Journal of Human Genetics*, *19*, 216–223. http://doi.org/10.1038/ejhg.2010.153

Moreno-Estrada, A., Gravel, S., Zakharia, F., McCauley, J. L., Byrnes, J. K., Gignoux, C. R., … Bustamante, C. D. (2013). Reconstructing the population genetic history of the Caribbean. *PLoS Genetics*, *9*, e1003925. http://doi.org/10.1371/journal.pgen.1003925

Murdock, G. P. (1981). *Atlas of world cultures*. London/Pittsburgh: Feffer and Simons/University of Pittsburgh. Retrieved from http://digital.library.pitt.edu/cgi-bin/t/text/text-idx?idno=31735057895496;view=toc;c=pittpress.

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*, *89*, 583–590. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17248844

Nenadic, O., & Greenacre, M. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: The ca package. *Journal of Statistical Software*, *20*, 1–13. http://doi.org/10.18637/jss.v020.i03.

Nuñez, C., Baeta, M., Sosa, C., Casalod, Y., Ge, J., Budowle, B., & Martínez-Jarreta, B. (2010). Reconstructing the population history of Nicaragua by means of mtDNA, Y-chromosome STRs, and autosomal STR markers. *American Journal of Physical Anthropology*, *143*, 591–600. http://doi.org/10.1002/ajpa.21355

Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., … Wagner, H. (2016). *vegan: Community ecology package*. Retrieved from https://cran.r-project.org/package=vegan.

Oota, H., Settheetham-Ishida, W., Tiwawech, D., Ishida, T., & Stoneking, M. (2001). Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics*, *29*, 20–21. http://doi.org/10.1038/ng711

Palstra, F., Heyer, E., & Austerlitz, F. (2015). Statistical inference on genetic data reveals the complex demographic history of human populations in Central Asia. *Molecular Biology and Evolution*, *32*, 1411–1424. http://doi.org/10.1093/molbev/msv030

Pérez-Lezaun, A., Calafell, F., Comas, D., Mateu, E., Bosch, E., Martínez-Arias, R., … Bertranpetit, J. (1999). Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *American Journal of Human Genetics*, *65*, 208–219. http://doi.org/10.1086/302451

Quinque, D., Kittler, R., Kayser, M., Stoneking, M., & Nasidze, I. (2006). Evaluation of saliva as a source of human DNA for population and association studies. *Analytical Biochemistry*, *353*, 272–277. http://doi.org/10.1016/j.ab.2006.03.021

Quintana-Murci, L., Chaix, R., Wells, R. S., Behar, D. M., Sayar, H., Scozzari, R., … McElreavey, K. (2004). Where west meets east: The complex mtDNA landscape of the southwest and Central Asian corridor. *American Journal of Human Genetics*, *74*, 827–845. http://doi.org/10.1086/383236

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from http://www.r-project.org/.

Raghavan, M., Skoglund, P., Graf, K. E., Metspalu, M., Albrechtsen, A., Moltke, I., … Willerslev, E. (2014). Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, *505*, 87–91. http://doi.org/10.1038/nature12736

Ramachandran, S., Deshpande, O., Roseman, C. C., Rosenberg, N. A., Feldman, M. W., & Cavalli-Sforza, L. L. (2005). Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 15942–15947. http://doi.org/10.1073/pnas.0507611102

Santos, C., Fregel, R., Cabrera, V. M., Álvarez, L., Larruga, J. M., Ramos, A., … González, A. M. (2014). Mitochondrial DNA and Y-

chromosome structure at the Mediterranean and Atlantic façades of the Iberian peninsula. *American Journal of Human Biology*, *26*, 130–141. http://doi.org/10.1002/ajhb.22497

Ségurel, L., Martínez-Cruz, B., Quintana-Murci, L., Balaresque, P., Georges, M., Hegay, T., … Vitalis, R. (2008). Sex-specific genetic structure and social organization in Central Asia: Insights from a multi-locus study. *PLoS Genetics*, *4*, e1000200. http://doi.org/10.1371/journal.pgen.1000200

Seielstad, M. T., Minch, E., & Cavalli-Sforza, L. L. (1998). Genetic evidence for a higher female migration rate in humans. *Nature Genetics*, *20*, 278–280. http://doi.org/10.1038/3088

Slatkin, M. (1995). A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, *139*, 457–462. Retrieved from http://www.genetics.org/content/139/1/457.short

Soucek, S. (2000). *A history of Inner Asia*. Cambridge University Press. 389.

Stoneking, M. (1998). Women on the move. *Nature Genetics*, *20*, 219–220. http://doi.org/10.1038/3012

Tremblay, M., & Vézina, H. (2000). New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *American Journal of Human Genetics*, *66*, 651–658. http://doi.org/10.1086/302770

Underhill, P. A., & Kivisild, T. (2007). Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual Review of Genetics*, *41*, 539–564. http://doi.org/10.1146/annurev.genet.41.110306.130407

van Oven, M., & Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, *30*, E386–E394. http://doi.org/10.1002/humu.20921

Verdu, P., Becker, N. S. A., Froment, A., Georges, M., Grugni, V., Quintana-Murci, L., … Austerlitz, F. (2013). Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Molecular Biology and Evolution*, *30*, 918–937. http://doi.org/10.1093/molbev/mss328

Wang, C. C., Wang, L. X., Shrestha, R., Zhang, M., Huang, X. Y., Hu, K., … Li, H. (2014). Genetic structure of Qiangic populations residing in the Western Sichuan Corridor. *PLoS ONE*, *9*, e103772. http://doi.org/10.1371/journal.pone.0103772

Wilkins, J. F. (2006). Unraveling male and female histories from human genetic data. *Current Opinion in Genetics & Development*, *16*, 611–617. http://doi.org/10.1016/j.gde.2006.10.004

Willems, T., Gymrek, M., Poznik, G. D., Tyler-Smith, C., & Erlich, Y. (2016). Population-scale sequencing data enable precise estimates of Y-STR mutation rates. *American Journal of Human Genetics*, *98*, 919–933. http://doi.org/10.1016/j.ajhg.2016.04.001

Wilson, I., Weale, M., & Balding, D. J. (2003). Inferences from DNA data: Population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society*, *166*, 155–201. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/1467-985X.00264/full

Wright, S. (1950). Genetical structure of populations. *Nature*, *166*, 247–249. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/15439261

Xue, Y., Zerjal, T., Bao, W., Zhu, S., Lim, S., Shu, Q., … Tyler-Smith, C. (2005). Recent spread of a Y-chromosomal lineage in Northern China and Mongolia. *The American Journal of Human Genetics*, *77*, 1112–1116. http://doi.org/10.1086/498583.

Yunusbayev, B., Metspalu, M., Metspalu, E., Valeev, A., Litvinov, S., Valiev, R., … Villems, R. (2014). The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genetics*, *11*, e1005068. http://doi.org/10.1101/005850

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

How to cite this article: Marchi N, Hegay T, Mennecier P, Georges M, Laurent R, Whitten M, Endicott P, Aldashev A, Dorzhu C, Nasyrova F, Chichlo B, Ségurel L, and Heyer E. Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations. *Am J Phys Anthropol*. 2016;00:1–14. doi:10.1002/ajpa.23151.

II.5 Discussion

Dans le développement bibliographique, seules des diversités génétiques sexe-spécifiques mesurées à une échelle géographique réduite ont été détaillées. Ce choix reposait sur une volonté de ne présenter que des différences sexe-spécifiques clairement imputées à un CCAS identifié. D'autres travaux s'intéressent à la diversité génétique sexe-spécifique à l'échelle mondiale mais donnent des résultats et des interprétations contradictoires en terme de CCAS : pour certains, la différenciation entre populations est 3,5 fois plus élevée pour le chromosome Y que pour l'ADN mitochondrial (F_{ST}=0,645 contre 0,186) interprétée comme de la patrilocalité (Seielstad *et al.* 1998), tandis que d'autres ne trouvent pas de différence (F_{ST}=0,334 pour le chromosome Y et 0,382 pour l'ADN mitochondrial) et concluent à des migrations symétriques entre les sexes (Wilder *et al.* 2004).
Ces deux études diffèrent par les marqueurs génétiques qu'elles utilisent (SNPs de NRY et sites de restriction de l'ADN mitochondrial *versus* STRs et séquence du gène mitochondrial MTCO3). En outre, des travaux récents menés sur des données de séquençage du chromosome Y et mitochondrial, *a priori* moins biaisées, trouvent également des différences de diversité sexe-spécifique pour les populations du jeu de données HGDP-CEPH (Lippold *et al.* 2014), et de taille efficace chez 110 populations de différents jeux de données (Karmin *et al.* 2015).
Au-delà de ces différences de marqueurs, ces études diffèrent également par les populations choisies et par la distance entre les populations, ce qui pourrait expliquer l'incohérence des résultats obtenus.

En effet, pour observer l'effet d'un CCAS sur la diversité génétique de populations, certains choix méthodologiques doivent être faits pour s'adapter à l'histoire de ce CCAS, par exemple en tenant compte du temps depuis lequel le CCAS a été mis en place dans la population étudiée et de son intensité. En effet, ces paramètres semblent cruciaux car plusieurs études ciblant la patrilocalité n'ont pas trouvé pas de traces génétiques de ce CCAS. À Sumatra, une plus faible diversité mitochondriale est trouvée, comme attendu, pour les sociétés matrilocales que pour les sociétés patrilocales, mais aucune différence n'est observée pour le chromosome Y (Gunnarsdóttir *et al.* 2011). Les auteurs donnent plusieurs pistes pour expliquer ce résultat : i) la séquence mitochondriale complète permettrait de détecter un effet de la matrilocalité sur la diversité génétique mais les haplotypes du chromosome Y incluant seulement quelques marqueurs ne permettraient pas d'observer un effet de la patrilocalité ; ii) la matrilocalité serait mieux appliquée que la patrilocalité (ce qui va à l'encontre d'observations réalisées en Thaïlande (Hamilton *et al.* 2005)) et laisserait donc plus d'empreintes ; iii) la patrilocalité n'aurait été appliquée que trop récemment et n'aurait pas encore eu le temps de laisser une marque sur la diversité génétique. On estime que l'apparition de la patrilocalité chez l'Homme est récente, remontant au Néolithique. Auparavant, la règle ancestrale aurait été la matrilocalité, d'après des phylogénie des règles de résidence de populations de différents continents (Holden et Mace 2003 ; Jordan *et al.* 2009), et l'observation, au moyen de données génétiques, de la matrilocalité chez les chasseurs-cueilleurs pygmées actuels (Verdu *et al.* 2013) ou Khoi-San (Behar *et al.* 2008), qui vivraient selon un mode de vie ressemblant à celui d'avant la révolution néolithique.

Les changements de règles de résidénce, par exemple la mise en place de la patrilocalité, correspondent à un changement de flux migratoire entre des populations d'un réseau matrimonial dont les effets sont visibles sur la différenciation génétique des populations au bout d'un temps inversement proportionnel à la distance géographique entre les populations liée à l'intensité de leur échanges migratoires sous un modèle d'isolement par la distance (Wilkins et Marlowe 2006). Ainsi, un changement récent du taux de migration s'observe en étudiant des populations proches, tandis que des populations éloignées conservent les traces

des migrations anciennes. En connaissant la date approximative du changement, et sous l'hypothèse d'un habitat continu, on peut alors calculer la distance critique entre deux populations, en-dessous de laquelle leur différenciation génétique est imputable au nouveau taux de migration (Wilkins et Marlowe 2006) :

$$x^* \approx \sqrt{2\sigma_2^2 \tau} \tag{II.6}$$

où x^* est la distance critique, τ l'âge du processus en générations, et σ_2 le taux de migration entre populations, sous la forme d'une variance de dispersion (en km/génération).

En ce qui concerne la patrilocalité, si elle est apparue il y a environ 10 000 ans en Eurasie, en comptant 25 ans par génération soit τ=400, et en prenant un σ_2=30 km/génération, on s'attend à voir un patron de différenciation génétique sexe-spécifique attendu sous un régime de patrilocalité pour des populations distantes de moins de 1000 km (Wilkins et Marlowe 2006).
Pour revenir au cas controversé des différences de diversité sexe-spécifique à l'échelle mondiale, les auteurs interprètent les différences, ou leur absence, en terme de patrilocalité. Or, cette controverse pourrait simplement s'expliquer par une hétérogénéité d'échantillonnage des différentes études : Seielstad *et al.* (1998), Lippold *et al.* (2014) et Karmin *et al.* (2015) utilisent un échantillonnage assez dense, avec de courtes distances entre populations, qui pourrait capter le signal génétique récent de patrilocalité, tandis que l'échantillonnage plus dispersé de Wilder *et al.* (2004) ne verrait que des processus plus anciens, comme ceux liés au peuplement et *a priori* symétriques entre les sexes.

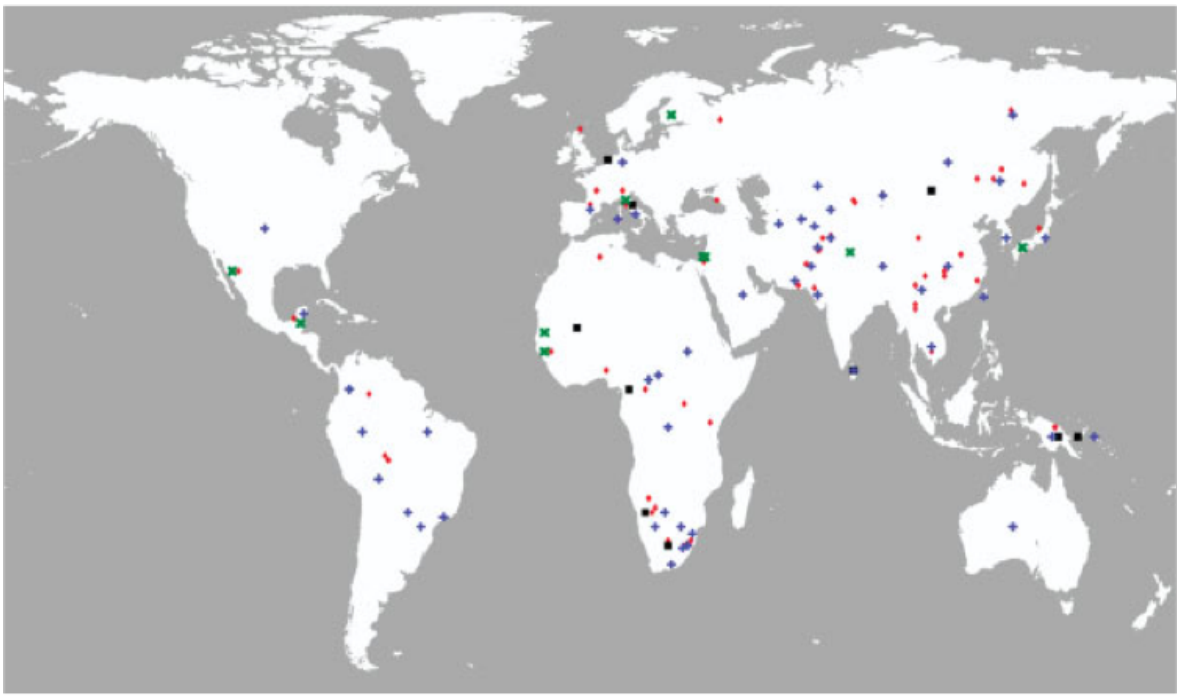


Figure II.13 – Distribution spatiale des populations étudiées par Seielstad *et al.* (1998) en vert pour les données mitochondriales et en bleu pour les données NRY étudiées ; en noir les données de Wilder *et al.* (2004) ; en rouge les données HGDP-CEPH utilisées par Lippold *et al.* (2014). *Figure 4 de Wilkins et Marlowe (2006).*

Dans l'étude des CCAS, la distance intervient également dans la définition de l'asymétrie même d'un

CCAS : une étude réalisée sur des populations patrilocales bantoues du Basotho et espagnoles d'Espagne trouve des taux de migration féminins plus forts pour les deux groupes à de faibles distances (<150 km). En revanche, pour des distances plus grandes, les taux de migration masculin et féminin sont équivalents donc aucune asymétrie ne serait visible. Cela implique que pour détecter des différences sexe-spécifiques dues à de la patrilocalité, il faut comparer des populations situées en-deça du seuil pour lequel l'asymétrie disparaît (Marks *et al.* 2012).

Enfin, si l'on s'intéresse aux migrations asymétriques, il faut veiller à échantillonner des populations susceptibles d'échanger suffisamment de migrants pour pouvoir détecter un effet sur les données génétiques, donc éviter au possible des populations séparées par des barrières culturelles ou géographiques (Destro-Bisol *et al.* 2012). Notamment, les auteurs d'une étude sur des populations humaines de l'est de l'Inde, ne trouvant pas de différences génétiques entre sexes, proposent comme explication une exogamie trop faible, et donc des migrations asymétriques trop marginales pour laisser une trace (Kumar *et al.* 2006). Une étude comparant les diversités génétiques sexe-spécifique chez le chimpanzé et l'Homme (Langergraber *et al.* 2007) trouve des différences génétiques entre sexes plus fortes pour le chimpanzé que pour l'Homme, signe selon eux d'une patrilocalité plus prononcée chez le chimpanzé. Ils soulignent cependant que les populations humaines étudiées proviennent de tribus différentes, alors que les migrations matrimoniales se font plutôt entre villages d'une même tribu. Ainsi, les migrations réalisées ne seraient pas assez importantes, ou pas assez asymétriques, pour impacter fortement la diversité génétique sexe-spécifique.
D'un point de vue méthodologique, l'étude de l'impact d'un CCAS sur la diversité génétique nécessite donc un échantillonnage couvrant des distances géographiques compatibles avec l'âge du CCAS, et incluant des populations susceptibles d'échanger des migrants.

Le jeu de données sur lequel se fonde mes travaux inclut plusieurs populations par groupe ethnique. Or, nous savons que la majorité des unions en Asie intérieure se font au sein des groupes ethniques (Heyer *et al.* 2009), et principalement entre des membres de villages différents chez les Turco-Mongols (Soucek 2000), ce qui nous laisse penser que cet échantillonnage est adapté à l'étude de CCAS créant des migrations matrimoniales asymétriques. Cependant, nous n'affirmons pas que les populations échantillonnées échangent directement des migrants entre elles, mais plutôt qu'elles font partie d'un réseau de migrations au sein du groupe ethnique, sans que tous les villages impliqués n'aient été échantillonnés. Il serait d'ailleurs intéressant de retracer le réseau d'échanges matrimoniaux à l'échelle de l'Asie intérieure, en recensant les lieux de naissance des migrants à partir des questionnaires ethno-démographiques dont nous disposons. En Asie intérieure, par contre, la plupart des groupes ethniques d'Altaï et de Tuva ne sont représentés que par une seule population. Idéalement, il faudrait échantillonner plus de populations dans cette région-là pour estimer la fréquence des migrations féminines et masculines au sein des différents groupes ethniques. Notons tout de même que des échanges entre groupes ethniques sont réalisés dans cette région en quantité non négligeable et devraient donc laisser une empreinte visible sur la diversité génétique.

Les distances moyennes entre les populations échantillonnées (620 km environ entre des populations d'Asie centrale et 480 km entre des populations d'Asie du nord, tous groupes ethniques confondus) sont compatibles avec la distance théorique de 1000 km sous laquelle les effets de la patrilocalité sont visibles sur la diversité génétique en supposant que le σ^2 en Asie intérieure n'est que de 30 km/génération. Par ailleurs, comme nous trouvons une structuration géographique de l'ADN mitochondrial entre l'Asie centrale et du nord, et que les populations d'Asie centrale et du nord sont en moyenne distantes de 2150 km, on peut se demander si cette structuration est causée par des événements antérieurs à la mise en

place de la patrilocalité, en lien par exemple avec le peuplement. Cette hypothèse suppose que les effets de la patrilocalité soient visibles entre ces régions en dépit de la présence de groupes ethniques différents pour le plus grand nombre dans ces deux aires. Alternativement, des barrières culturelles pourraient également causer cette différence génétique entre les régions si les échanges entre les différents groupes ethniques de ces deux aires sont inexistants ou faibles. Ou encore, les lieux d'échantillonnage ne formant pas un gradient continu - aucun échantillon n'ayant été collecté dans l'est du Kazakhstan ou en Chine - la différence génétique pourrait être due à un biais d'échantillonnage. Ces hypothèses ne changent pas l'interprétation que nous avons faite à ce résultat, à savoir que les déterminants des migrations de femmes ne sont pas les mêmes au sein de chaque région ou à l'échelle de l'Asie intérieure, mais seraient intéressantes à explorer.

Une des particularités des travaux menés dans cette thèse et préalablement dans l'équipe d'Anthropologie Évolutive est d'étudier conjointement des populations cognatiques et patrilinéaires (Ségurel *et al.* 2008 ; Chaix *et al.* 2007 ; Heyer *et al.* 2009). Comme ces populations ont la même règle de résidence patrilocale, leurs différences de diversité génétique seraient causées par leurs règles de filiation différentes. À ma connaissance, aucune autre étude ne distingue l'effet des règles de résidence de celui des règles de filiation, la matrilinéarité allant souvent de pair avec la matrilocalité, et la patrilinéarité avec la patrilocalité. Nous verrons cependant dans le chapitre suivant que des différences de patrilocalité sont attendues entre les groupes linguistiques aux vues de leurs règles d'alliance.

Dans mes travaux, nous disposons donc de différents groupes ethniques originaires d'une grande aire géographique (Asie centrale et du nord) dont certains, les Turco-Mongols, partagent la même règle de filiation patrilinéaire. Cela nous autorise donc des comparaisons portant sur la pratique de la patrilinéarité, entre groupes et régions, et nous permet aussi de tester l'omniprésence des noyaux d'identité chez les patrilinéaires d'Asie intérieure. Pour l'instant, cette structuration n'a fait l'objet d'étude qu'en Asie centrale, et du nord depuis ma thèse, mais il serait intéressant de l'étendre à d'autres régions du monde. Enfin, les différences génétiques observées entre les populations patrilinéaires et cognatiques suggèrent que ces règles de filiation ont été mises en place depuis un temps suffisant pour avoir agi sur la diversité génétique, mais il serait intéressant d'inférer leur âge.

Chapitre III

La consanguinité et son évitement par des migrations matrimoniales

Sommaire

III.1 La consanguinité

L'étude de la consanguinité se situe à la croisée de l'anthropologie et de la biologie évolutive, nourrissant une vaste littérature dans le domaine de la santé, des mathématiques, de l'écologie, des sciences humaines ou encore de la génétique des populations. Nous essayons de rendre compte de l'aspect pluridisciplinaire de cette thématique et de donner des éléments d'explication à l'intérêt qu'elle suscite, y compris dans le cadre de l'étude des populations d'Asie intérieure.

Définition et estimation

Au sens littéral du terme, un individu *consanguin* est issu "du même sang", c'est-à-dire chez l'Homme que ses parents partagent au moins un ancêtre commun une à trois générations auparavant (Alvarez *et al.* 2011). Consensuellement, on retient comme consanguines les unions d'un individu avec ses ascendants et ses descendants sur deux générations, les membres de sa fratrie (frères, sœurs et demi-frères, demi-sœurs), de celles de ses parents (oncles, tantes, demi-oncles, demi-tantes, soit ses avunculaires), la descendance de ses frères et sœurs, et ses cousins de divers types (dont doubles ou simples germains, et issus de germains) (Figure III.1).
Chez les animaux, les unions consanguines généralement décrites sont celles entre des apparentés au premier degré, à savoir parents et enfants, frères et sœurs, voire au second degré avec des relations entre demi-frères et demi-sœurs. Quelques études sur les primates et les souris s'intéressent aussi aux relations avunculaires (Pusey 1990) ou entre cousins (Pusey et Wolf 1996).

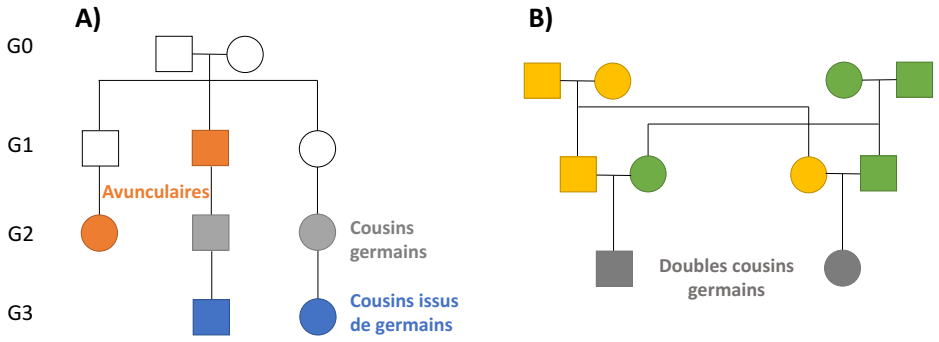


Figure III.1 – Illustration de quelques liens généalogiques tels qu'avunculaires, cousins germains et issus de germains (A) ou doubles cousins germains (B).

Du point de vue de la génétique, tout individu, issu d'une union consanguine ou non, a une probabilité non nulle de posséder, pour un locus donné, deux copies alléliques identiques, dites ***I**dentical **B**y **S**tate* ou homozygotes. Ceci peut arriver par chance mais aussi du fait d'une union consanguine. Les allèles sont alors hérités de l'ancêtre récent et commun des parents (Figure III.2) et sont homozygotes par descendance (***I**dentical **B**y **D**escent*) ou d'autozygotes.
Cette probabilité est appelée coefficient de consanguinité, usuellement noté F, et par définition compris entre 0 et 1 (Malécot 1948 ; Bengtsson et Jacquard 1976). Le coefficient est d'autant plus élevé que les parents sont de proches apparentés car ils partagent un plus grand nombre de loci hérités du même ancêtre. Son estimation peut se faire selon différentes approches.

La première est généalogique : elle repose sur l'identification des liens de parenté entre les partenaires au moyen de questionnaires, registres civils ou paroissiaux pour l'Homme, ou de données démographiques observées pour d'autres espèces. Le coefficient de consanguinité est établi à partir de la méthode des

chemins (Wright 1922, 1943) :

$$F = \sum 0.5^{n_M+n_P+1}(1+F_A) \tag{III.1}$$

où F_A est le coefficient de consanguinité de l'ancêtre commun A,

n_M le nombre de générations séparant l'ancêtre A de la mère, n_P du père

et où cette reconstruction est faite pour chaque ancêtre commun à la mère et au père.

Puis, au début de l'avènement de la génétique, il a été possible d'utiliser quelques marqueurs ADN (allozymes ou microsatellites) pour identifier les parents d'un individu, pour ensuite calculer un coefficient de consanguinité généalogique (Charpentier *et al.* 2007).

Depuis, les avancées dans le domaine du génotypage ont permis la production de données sur l'ensemble du génome et le calcul d'un coefficient de consanguinité directement à partir des données génétiques. Ce coefficient peut notamment être estimé comme la proportion de loci autozygotes dans un génome (Ritland 1996), par une approche simple-point en considérant indépendamment l'information de chaque locus (Gazal *et al.* 2014b). Cependant, cette approche se heurte à la difficulté à distinguer les loci autozygotes, causés par la consanguinité, de ceux IBS par chance (Powell *et al.* 2010). Pour ce faire, on peut calculer le niveau d'homozygotie d'un individu et le corriger par l'homozygotie attendue du fait du hasard dans la population :

$$F = \frac{Homozygotie_{Ind} - Homozygotie_{Attendue}}{1 - Homozygotie_{Attendue}} \tag{III.2}$$

De fait, le coefficient de consanguinité mesure un excès ou un déficit d'homozygotie, supposée IBD, chez des individus, par rapport à l'attendu dans la population, soit un écart à la panmixie qui n'est pas que rarement calculé à partir des données généalogiques.
Une des limites de cette approche simple-point est de devoir recourir à une population de référence pour calculer l'attendu.

D'autres manières d'estimer la consanguinité génétique sans avoir recours à une population de référence ont été développées en considérant des segments chromosomiques autozygotes, ce qui est une approche multi-points (Figure III.2). De façon générale, au cours de la méiose, la recombinaison crée des chromosomes hybrides incluant des segments des chromosomes parentaux. L'un de ces chromosomes est ensuite transmis à la descendance par la reproduction sexuée, porteur d'une combinaison de segments chromosomiques provenant de chacun des futurs grands-parents. (Broman et Weber 1999). Dans le cas particulier d'une union consanguine, certains segments chromosomiques peuvent être à l'état autozygote dans le génome des descendants, s'ils proviennent du même chromosome ancestral. La taille et le nombre de ces segments dépendent du nombre d'événements de recombinaison survenus depuis cet ancêtre : plus le nombre de générations écoulées est élevé, plus la taille moyenne des segments est petite et plus leur nombre est grand (Table III.1) (Thomas *et al.* 1994 ; Gazal 2014) :

- pour un nombre de méioses d, la taille moyenne des segments attendue est de $\frac{100}{d}cM$, soit 16,67 cM pour une descendance de cousins germains ;
- le nombre de segments attendu (Nseg) est de $\frac{b(rd+c)}{2^{d-1}}$ pour un nombre b ancêtres, c chromosomes (22 dans la Table III.1), r la longueur du génome (en morgan, 35.3 dans la Table III.1) ;
- la probabilité de n'observer aucun de ces segments dans le génome considéré après d méioses est de e^{-Nseg}.

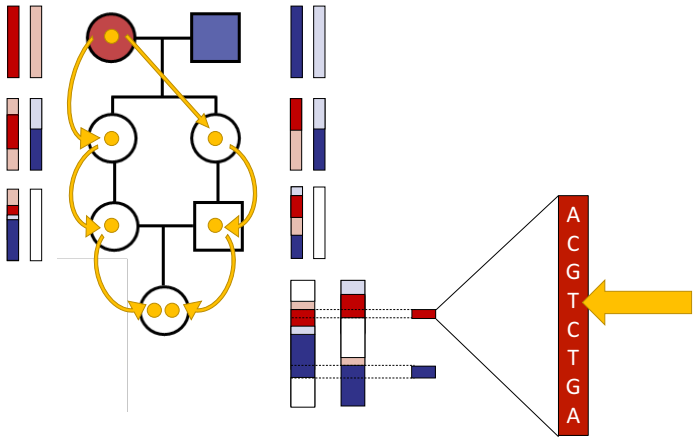


Figure III.2 – Consanguinité d'un enfant de cousins germains pour une paire de chromosomes. Le point jaune figure la transmission par descendance d'un allèle à un locus précis. Ce locus est inclus dans l'un des deux segments autozygotes retrouvés dans le génome du descendant. À chaque génération, la recombinaison combine des segments issu des chromosomes parentaux homologues en de nouveaux chromosomes "hybrides". De fait, la taille des segments du chromosome ancestral diminue de génération en générations. *Inspirée de McQuillan* et al. *(2008)*.

Table III.1 – Consanguinité selon l'apparentement généalogique des parents : coefficient de consanguinité généalogique (F) pour un F_A nul, nombre d'ancêtres en commun (b), nombre de méioses depuis l'ancêtre (d), nombre de segments autozygotes (Nseg), taille moyenne de ces segments en cM (TM) et probabilité de ne pas en observer dans le génome des descendants (Non-obs.). *Largement inspiré de Gazal (2014).*

Apparentement	**F**	b	d	**Segments autozygotes**		
Exemples				***Nseg***	***TM (cM)***	***Non-obs.***
Identité	1/2					
Auto-fécondation		1	2	46,30	50	$7{,}80 * 10^{-22}$
Jumeaux monozygotes		1	2	46,30	50	$7{,}80 * 10^{-22}$
Premier degré	1/4					
Parents/Enfants		1	3	31,97	33,33	$1{,}30 * 10^{-15}$
Frères/Sœurs		2	4	40,80	25	$1{,}91 * 10^{-19}$
Second degré	1/8					
Demi-frères/Sœurs		1	4	20,40	25	$1{,}38 * 10^{-10}$
Grands-parents/Petits-enfants		1	4	20,40	25	$1{,}38 * 10^{-10}$
Avunculaires (Oncle/Nièce, Tante/Neveu)		2	5	24,81	20	$1{,}67 * 10^{-12}$
Doubles cousins germains		4	6	29,22	16,67	$2{,}03 * 10^{-14}$
Troisième degré	1/16					
Cousins germains		2	6	14,61	16,67	$4{,}51 * 10^{-8}$
Cinquième degré	1/64					
Cousins issus de germains		2	8	4,76	12,50	0,01

Dans la pratique, plusieurs méthodes permettent d'identifier dans le génome des segments homozygotes, appelés ROHs pour *Runs Of Homozygosity* (Purcell *et al.* 2007 ; Browning et Browning 2010 ; Han et Abney 2013). Il faut ensuite distinguer les ROHs autozygotes (*Homozygotes-by-descent*, HBD) et ceux IBS par chance. Pour cela, on peut utiliser un critère de longueur de segments (discuté dans l'article III.2 et Gazal *et al.* (2014b)) : plus un ROH est long, moins il est probable que les nombreux loci qui le

constituent soient tous homozygotes par chance, et donc ce ROH est probablement HBD.
Ainsi, la proportion génomique des segments HBD correspond à elle seule à une mesure de la consanguinité (McQuillan *et al.* 2008 ; Verweij *et al.* 2014). Notons que ce coefficient est défini en fonction du seuil de taille retenu pour inférer le statut HBD des ROHs. Il est aussi possible d'inférer le type d'union consanguine à l'origine de l'autozygotie observée chez un individu en se basant sur la taille des ROHs et leur nombre.

D'autres méthodes d'estimation de la consanguinité utilisent les ROHs mais sans tenir compte de leur taille : elles infèrent le statut d'autozygotie de chaque locus du génome en se fondant sur leur probabilité d'être autozygote en sachant que les loci voisins sont autozygotes ou non, et en connaissant les fréquences alléliques au locus étudié (Leutenegger *et al.* 2003). Ces méthodes estiment donc un coefficient de consanguinité génétique en combinant les avantages des approches simple et multi-points.

Outre les individus, le coefficient de consanguinité peut également être calculé pour des populations, comme la moyenne des coefficients des individus la composant, souvent noté α dit de Bernstein (Jacquard 1968b).

Conséquences biologiques à l'échelle individuelle

Les unions consanguines augmentent le nombre de sites homozygotes au sein du génome des descendants par rapport à des unions moins consanguines. De fait, les individus consanguins ont un risque accru d'être homozygotes pour des mutations récessives délétères (Charlesworth et Willis 2009), ou à des loci sous avantage hétérozygote (Balloux *et al.* 2004 ; Mead 2003 ; Quintana-Murci et Barreiro 2010), ce qui devrait diminuer leur fitness.
Cette diminution de fitness, aussi appelée dépression de consanguinité, a notamment été étudiée, chez les végétaux et les animaux, par des comparaisons de fitness entre des descendances non-consanguines et celles issues de croisements consanguins, expérimentaux ou empiriques. Chez l'Homme, la comparaison de fitness entre des non-consanguins et consanguins est possible lorsque l'on a accès à la généalogie des individus. C'est notamment le cas pour quelques populations humaines devenues des cas d'école, telles que :

- la population du Québec. Le Registre de la Population du Québec Ancien renseigne sur la généalogie et la démographie de près de 700 000 personnes, depuis la fondation de cette colonie en 1608, jusqu'en 1800.
- la population québécoise de Saguenay-Lac-Saint-Jean. La base de données BALSAC recense la généalogie et la démographie, mais aussi la prévalence de certaines maladies dans cette colonie fondée au cours de la seconde moitié du 19^{e} siècle (Heyer et Tremblay 1995 ; Heyer *et al.* 1997).
- certaines communautés religieuses d'Amérique du Nord, comme les Amish ou les Mormons. Ces colonies sont souvent sujettes à la consanguinité : la secte anabaptiste des Huttérites, installée en Amérique depuis les années 1870 sous la forme de trois branches endogames (les Schmiedeleuts, Dariusleuts et Lahrerleuts) (Hostetler 1985) est la population la plus consanguine d'Amérique du Nord (en moyenne, les époux sont autant apparentés que des cousins issus de germains). De plus, leur mode de vie très sain (des aliments Bio, une activité physique quotidienne, et pas de consommation de tabac ni d'alcool) fait que les maladies observées dans ce groupe seraient principalement d'origine génétique (Gao *et al.* 2015).
- la population islandaise. Depuis le peuplement de l'île il y a près de 1000 ans, la généalogie islandaise est inscrite dans le registre Íslendingabók. Plus récemment, les données génétiques et

phénotypiques d'un tiers des 300 000 Islandais ont été intégrées à la base de données DeCODE. De plus, la population islandaise est assez homogène sur un plan socio-économique, et on s'attend donc à voir peu d'effets des paramètres socio-économiques sur la fitness des Islandais.

La dépression de consanguinité est étudiée à travers différentes composantes de la fitness comme la mortalité, la fertilité, et à travers des paramètres tels que la prévalence d'anomalies congénitales variées ou certaines mensurations.

La mortalité

Chez les animaux et les végétaux, une augmentation de la consanguinité entraîne généralement une réduction de la viabilité, aux stades immatures et adultes (Keller et Waller 2002). Chez l'Homme, on recense également une plus forte mortalité des individus consanguins : dans une méta-analyse de 69 études comprenant 2,4 millions de naissances, la mortalité chez les descendants de cousins germains est accrue de 3,5% par rapport aux descendants de couples moins apparentés (Bittles et Black 2010). Au Moyen-Orient, cela représente 15/1000 morts supplémentaires pour les enfants de moins de 5 ans descendant de cousins germains (Pedersen 2002).
Un statut socio-économique plus faible pour les couples consanguins pourrait expliquer partiellement les différences de mortalité observées (Hussain et Bittles 1998). Mais à statut socio-économique comparable, les différences persistent : au sein de la communauté mormone de l'Utah, des frères ont généralement le même niveau de vie, mais la mortalité de leur descendance augmente s'ils sont mariés avec des cousines plutôt qu'à des non-apparentées (Figure III.3) (Jorde 2001).

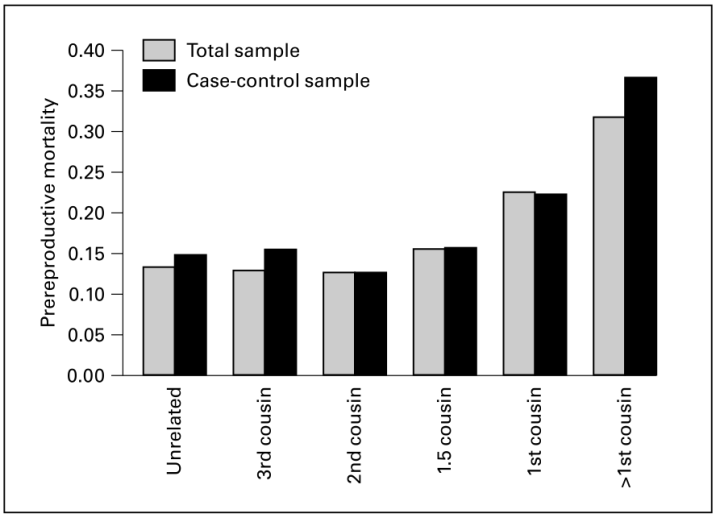


Figure III.3 – Mortalité accrue avant l'âge de 16 ans chez les descendants de cousins germains (1st cousin et 1.5 cousin) ou plus proche (>1st cousin), dans la population mormone de l'Utah entre 1847 et 1945. Cette comparaison est réalisée pour tous les couples de la base de données (barres grises) ou pour un sous-échantillon de cas-contrôles : des hommes impliqués dans des unions consanguines (barres noires - 3rd cousins jusqu'à >1st cousin) dont les frères sont mariés à des non-apparentées (barre noire - unrelated). *Figure 2 de Jorde (2001).*

Les loci responsables de cette augmentation de la mortalité ne sont pas identifiés dans ce type d'étude et peuvent être rattachés à diverses anomalies congénitales détaillées par la suite. Cependant, il a été démontré que la consanguinité diminuait la résistance à certains pathogènes tels que les virus des hépatites ou *Mycobacterium tuberculosis*, entrainant de fait une augmentation de la mortalité (Alvarez *et al.* 2011).

La fertilité

Darwin reporta le premier une fertilité moindre chez les plantes pratiquant l'auto-fécondation par comparaison à des plantes dont la pollinisation était croisée entre individus : la production de graines était diminuée de 41% chez les autogames comparées aux allogames (Charlesworth et Willis 2009 ; Darwin 1876). Chez l'Homme, on pourrait également s'attendre à ce que les couples consanguins aient moins d'enfants sachant que de nombreux allèles sont létaux à l'état homozygote, comme certains allèles de différents loci HLA (A, B, DR, DQ, G) (Chapitre 7 de Bittles (2012)), et que la consanguinité augmente leur risque d'être présents à l'état homozygote. Au contraire, certaines études trouvent une plus grande fertilité chez des couples consanguins (par exemple au nord de l'Inde (Fareed *et al.* 2016) ou parmi 30 populations d'Asie, Afrique et Moyen-Orient comptabilisant 550 000 naissances (Bittles *et al.* 2002)). Cependant, cette tendance pourrait s'expliquer par des différences socio-économiques :

- d'âge au mariage. Les couples consanguins sont en général mariés plus jeunes (au Pakistan, 18,8 ans en moyenne pour des cousins germains contre 19,7 ans pour des non-apparentés ; Chapitre 7 de Bittles (2012)) et ont des enfants plus rapidement ;
- d'accès à la contraception, souvent associé à des modes de vie moins traditionnels et est donc moins répandu chez les couples consanguins (Bittles 2001) ;
- de compensation reproductive : après la perte d'un enfant, les couples peuvent choisir de le "remplacer", d'avoir un autre enfant (Overall *et al.* 2002). Ce mécanisme pourrait être compenser la mortalité infantile plus forte chez les couples consanguins.

Bien que l'on n'observe pas de différence de fertilité au niveau des couples consanguins ou non, la consanguinité semble avoir un effet notable sur la fertilité de leurs descendants : à Saguenay-Lac-Saint-Jean, la fertilité des hommes consanguins diminue relativement aux non-consanguins au cours de leur seconde moitié de vie reproductive (Robert *et al.* 2009) ; en Islande, des partenaires plus apparentés que des second-cousins ont moins de petits-enfants que des non-apparentés (Helgason *et al.* 2008). La cause de cette baisse de fertilité chez les individus consanguins n'est pas clairement identifiée chez l'Homme mais elle pourrait être associée à une diminution de la motilité des spermatozoïdes et à une augmentation du nombre de spermatozoïdes morphologiquement anormaux, comme cela a été observé chez les poneys Shetland (Van Eldik *et al.* 2006).

La taille

La taille, à la naissance et à l'âge adulte, serait une approximation de la fitness d'un individu. Or, les nouveaux-nés consanguins ont de plus petites mensurations que les non-consanguins, et à l'âge adulte, les descendants de cousins germains, originaires du sud-Tyrol, de l'archipel écossais des Orcades ou d'Irlande, sont plus petits, de 3 cm en moyenne, que les individus non-consanguins de leur population (McQuillan *et al.* 2012). Une fois encore, l'impact de facteurs socio-économiques serait à prendre en compte (Chapitre 8 de Bittles (2012), (Verweij *et al.* 2014 ; Fareed et Afzal 2014)).

Anomalies congénitales

Des anomalies diminuant la fitness ont clairement été reliées à la consanguinité (Bittles 2008). Le premier cas historique remonte à 1902 quand sur les 18 personnes atteintes d'alcaptonurie recensées en Europe du Nord et Amérique, 12 étaient en fait issues d'unions entre cousins germains (Garrod 1902). Outre cette première étude, plusieurs groupes humains, notamment au Proche-Orient (Saad *et al.* 2014 ; Al-Gazali *et al.* 2006), au sud de l'Inde et au Pakistan (Chapitre 10 de Bittles (2012)), combinent un degré

élevé de consanguinité et des anomalies fréquentes. Cette relation peut s'expliquer par l'augmentation du nombre d'invidus homozygotes récessifs, exprimant le caractère anormal, en fonction de la consanguinité (Jacquard et Reynès 1968) :

$$f(aa) = p^2 + Fp(1-p), \tag{III.3}$$

où *f(aa)* est la fréquence des individus porteurs de l'allèle récessif à l'état homozygote,

p est la fréquence de l'allèle délétère récessif a et F le coefficient de consanguinité.

Cette approche statistique a contribué au développement de l'*homozygosity mapping* (Lander et Botstein 1987 ; Génin et Todorov 2006) qui tire profit de segments autozygotes retrouvés chez les individus présentant l'anomalie pour identifier les gènes candidats à l'origine de l'anomalie en question, comme dans cette étude du syndrome de Taybi-Linder (Leutenegger *et al.* 2006).
Les effets néfastes de la consanguinité sont également visibles à travers l'accumulation d'anomalies au sein de lignées concluant des unions consanguines de façon récurrente. Cela a notamment été le cas pour la lignée royale des Hasbourg et en particulier la branche espagnole dont le dernier membre, Charles II, présentait un coefficient de consanguinité de 0,245 (Figure III.4) (Alvarez *et al.* 2009, 2011). Charles II, dont le surnom était l'Ensorcelé, cristallisait les stigmates de la consanguinité : il était chétif avec une tête disproportionnée, avait des difficultés de locution et locomotion en lien avec une hypotonie, souffrait de troubles intestinaux, d'œdèmes, d'hématurie, de convulsions et d'hallucinations et mourut à 39 ans, sans descendance. De manière spéculative et *a posteriori*, ces anomalies cliniques ont été attribuées à deux pathologies combinées (une déficience en hormones pituitaires et une acidose tubulaire rénale), rares dans la population générale et qui auraient été causées par la forte consanguinité de la lignée. Ironiquement, la consanguinité de cette lignée est souvent illustrée par leur prognathisme mandibulaire dit "la lippe Hasbourgeoise" (Galippe 1905), alors que ce caractère suit une transmission dominante, sans lien avec la consanguinité (Figure III.4) (Wolff *et al.* 1993 ; Thompson et Winter 1988).

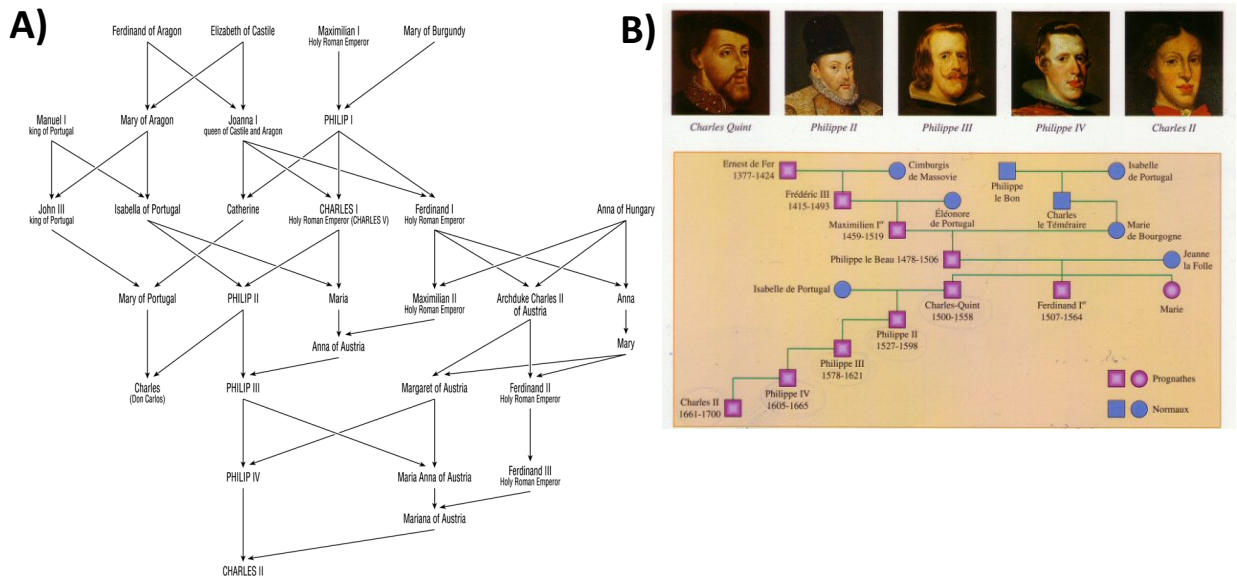


Figure III.4 – Arbre généalogique des Hasbourgs d'Espagne (A) et transmission du prognathisme mandibulaire dans cette lignée (B). *Figure 1 de Alvarez* et al. *(2011) et sujet SVT niveau collège : étude d'un arbre généalogique, le prognathisme dans la famille de Habsbourg.*

Au sein des familles consanguines, la survenue de ces anomalies reste en général inattendue, les deux parents étant souvent porteurs sains. De plus, du fait de la transmission chromosomique stochastique,

les enfants nés de mêmes parents ne sont pas tous touchés, ou peuvent l'être à différents degrés, voire pour des anomalies différentes. Le cas d'une famille du Penjab illustre parfaitement cette stochascticité de la consanguinité (Knight *et al.* 2008) : sur les six enfants nés de parents sains et cousins germains, l'un est sain et les cinq autres sont sujets à l'épilepsie et/ou atteints de surdité (Figure III.5). Ces deux anomalies ont été rattachées à deux régions génomiques distinctes, portées par deux chromosomes différents et transmises aux enfants de manière aléatoire.

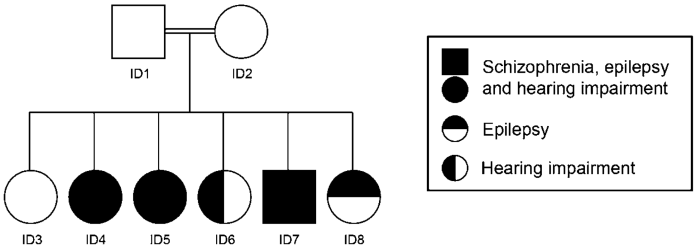


Figure III.5 – Anomalies congénitales au sein d'une famille de six enfants. *Figure 1 de Knight* et al. *(2008).*

Conséquences biologiques à l'échelle des populations

Au-delà de la seule fitness des individus, la consanguinité agit aussi sur la qualité biologique des populations : une augmentation de 10% de consanguinité à l'échelle de la population équivaut à une perte de 5 à 10% de sa fitness (Keller et Waller 2002).

On exprime la qualité biologique d'une population à travers sa charge génétique (*genetic load*) mesurant la diminution de sa fitness du fait d'effets génétiques, par rapport à une population d'individus de référence (aux génotypes optimaux, ou simplement quelques générations plus tôt). On peut la décomposer en trois facteurs : la charge de dérive (*drift load*), la charge mutationelle (*mutation load*) et la charge de consanguinité (*inbreeding load*) (Hedrick et Garcia-Dorado 2016).

La charge de dérive correspond à la diminution de fitness imputable à la fixation par dérive d'allèles légèrement délétères dont le coefficient de sélection rapporté à la taille efficace de la population vaut moins de 1 (Jacquard 1971b ; Hedrick et Garcia-Dorado 2016). Les mutations plus fortement délétères sont perdues du fait de la sélection. La dérive génétique est particulièrement intense dans des populations isolées et de petite taille : dans la communauté Rom en Slovaquie, la prévalence du glaucome congénital est de 1/1250 contre 1/22000 dans la population générale ; sur l'île de Pingelap en Micronésie, dévastée par un typhon en 1775 et repeuplée par les quelques survivants, l'achromatopsie est devenue extrêmement fréquente (entre 5 à 10% de personnes atteintes, 30% de porteurs sains) contre 1 cas sur 50 000 dans le reste du monde (Sacks 1997).

La charge mutationnelle correspond à la fitness d'individus non-consanguins (notée w_0) et se calcule comme la somme, à tous les loci présentant un allèle délétère, du terme : $2hsq(1-q) + sq^2$ (où q est la fréquence dans la population de l'allèle délétère, $1-s$ est la fitness des homozygotes pour cet allèle et $1-hs$ celle des hétérozygotes) (Charlesworth et Willis 2009 ; Hedrick et Garcia-Dorado 2016).

Connaissant w_0 et ayant mesuré la fitness d'individus présentant différents degrés de consanguinité, il est possible d'estimer la charge de consanguinité B à partir de l'équation (O'Grady *et al.* 2006) :

$$w_F = w_0 e^{-BF} \tag{III.4}$$

où w_F est la fitness des individus ayant un degré F de consanguinité

$$\text{et donc } B = -\frac{ln(w_F/w_o)}{F} \tag{III.5}$$

Cependant, la charge consanguine peut être réduite par de la purge : la consanguinité expose à la sélection les allèles délétères récessifs, ce qui fait baisser leur fréquence dans la population. Ce phénomène a notamment été observé chez une population captive de gazelles (Moreno *et al.* 2015). Cette purge est particulièrement efficace pour des allèles létaux, ou avec des effets délétères très marqués, et agit sur ceux ayant un plus faible effet si l'augmentation de la consanguinité n'est pas trop rapide (équilibre dérive/purge) (Cazes 1980). La dynamique de la purge se mesure pour différentes générations (t) par :

$$B = -\frac{ln(w_t - w_0)}{g_t} \tag{III.6}$$

où g_t dépend de F et de l'efficacité de la purge notée $d = \frac{s}{2} - hs$.

On décompose cette dynamique en deux temps (Figure III.6) (Hedrick et Garcia-Dorado 2016) :

1. diminution de la fitness au cours des premières générations en lien avec la charge de dérive. Cette phase est plus marquée et plus courte pour des populations de petite taille efficace.
2. purge des allèles délétères et donc augmentation de la fitness (rebond) parfois jusqu'à dépasser la fitness de la population de référence n'ayant pas purgé. Des tailles efficaces réduites accélèrent la venue du rebond.

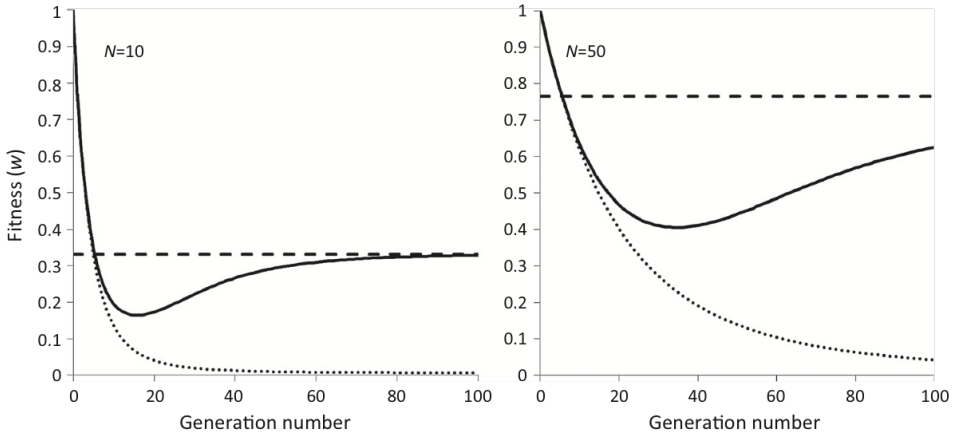


Figure III.6 – Effet de la purge sur la fitness. Deux tailles efficaces de populations ont été simulées (N_e=10 et 50). La charge de consanguinité et mutationnelle sont fixées par : B = 5, s = 0,25 et h = 0,2. La ligne pleine correspond à l'équation (III.6) tandis que la ligne pointillée illustre l'équation sans purge (III.4). *Box 1b de Charlesworth et Willis (2009).*

En dehors des mutations délétères, les populations sont menacées par la diminution de la diversité génétique causée par la consanguinité : les unions entre apparentés créent moins de combinaisons génétiques et la purge aggrave cette perte de diversité par un phénomène d'auto-stop génétique. En laboratoire, cette perte de diversité génétique a été associée à un risque d'extinction accru des populations (Frankham 2005). Chez des populations naturelles de papillons, une diminution de la survie des larves, de la longévité des adultes et de leur niveau de ponte a été observée lorsque l'hétérozygotie de la population

diminue, soit un risque d'extinction pour les populations accru proportionnellement en fonction de leur consanguinité(Saccheri *et al.* 1998). Une diversité réduite équivaut aussi à une adaptabilité moindre et donc à une réponse moins modulable face à la prédation, aux stress environnementaux ou aux pathogènes (Keller et Waller 2002). Ainsi, chez la drosophile, la résistance à des pathogènes et toxines est reliée linéairement à la consanguinité : $R = 0,488 - 0,416F^2$ (Spielman *et al.* 2004) .

Pratiques de la consanguinité et génétique des populations

En dépit des risques biologiques généralement connus, 10% de la population mondiale actuelle descend de parents au moins autant apparentés que des cousins issus de germains (Bittles et Black 2010), et il existe de grandes fluctuations dans la prévalence de la consanguinité selon les populations considérées. Pour illustrer ces variations, nous avons choisi d'utiliser la base de données *consang.net*, disponible en ligne, qui recense des informations portant sur la consanguinité de 563 communautés (Figure III.7A) à partir de données de recensements, de registres civils ou paroissiaux (Figure III.7B).

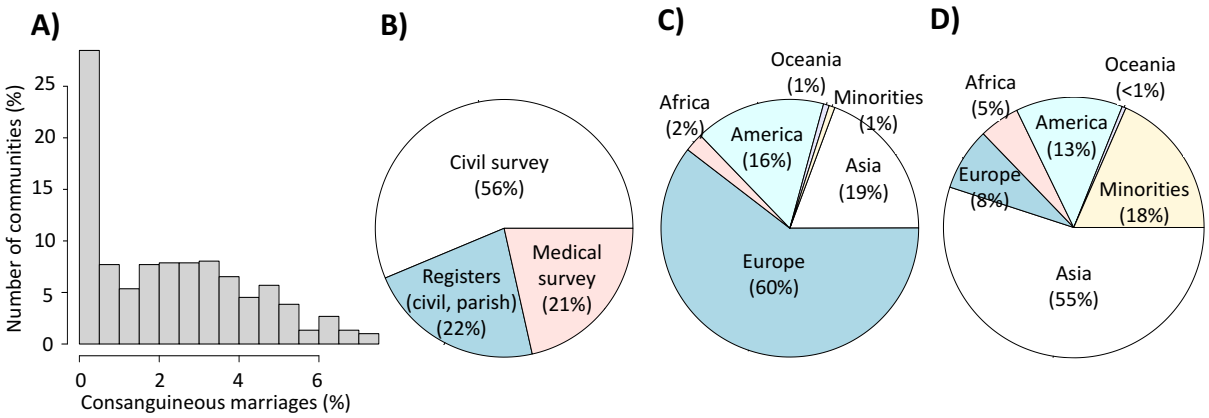


Figure III.7 – Informations obtenues auprès de 563 communautés recensées : A) Pourcentage d'unions entre des cousins issus de germains ou plus apparentés par communauté, B) types de données récoltées, C) nombre d'individus étudiés par continent, D) nombre de communautés étudiées par continent.

À partir de ces informations, la consanguinité a été représentée par pays (Figure III.8) :

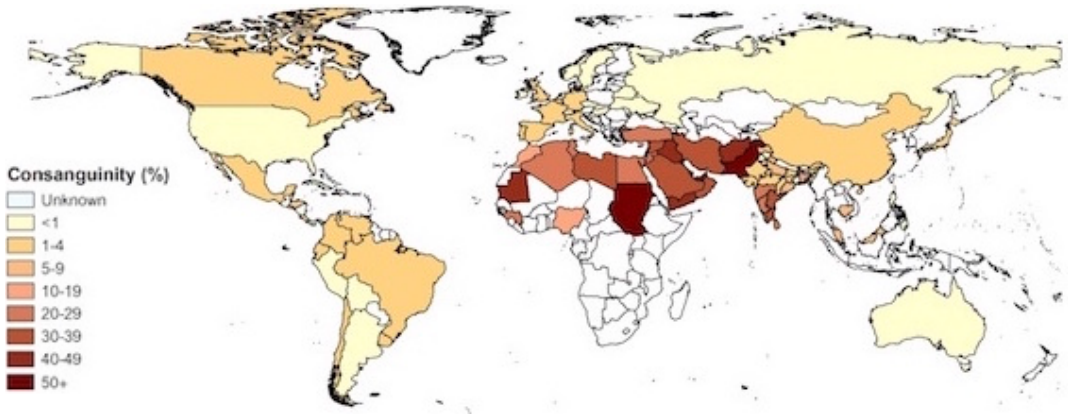


Figure III.8 – Consanguinité moyenne par pays ou région. *Source : consang.net/index.php/Global_prevalence.*

De notre côté, nous avons représenté par continent les distributions des pourcentages de consanguinité des différentes communautés référencées dans la base de données (Figure III.9). Malgré une disparité du nombre de communautés représentées par continent (Figures III.7 C,D), on observe deux dynamiques : les communautés américaines, européennes et océaniennes ne dépassent pas 20% de consanguinité, avec des valeurs médianes de 1,5, 0,8 et 0,5% respectivement, tandis que la plupart des communautés africaines et asiatiques atteint plus de 20% de consanguinité, certaines dépassant 50%, ce qui se manifeste par des médianes atteignant 34 et 28%.

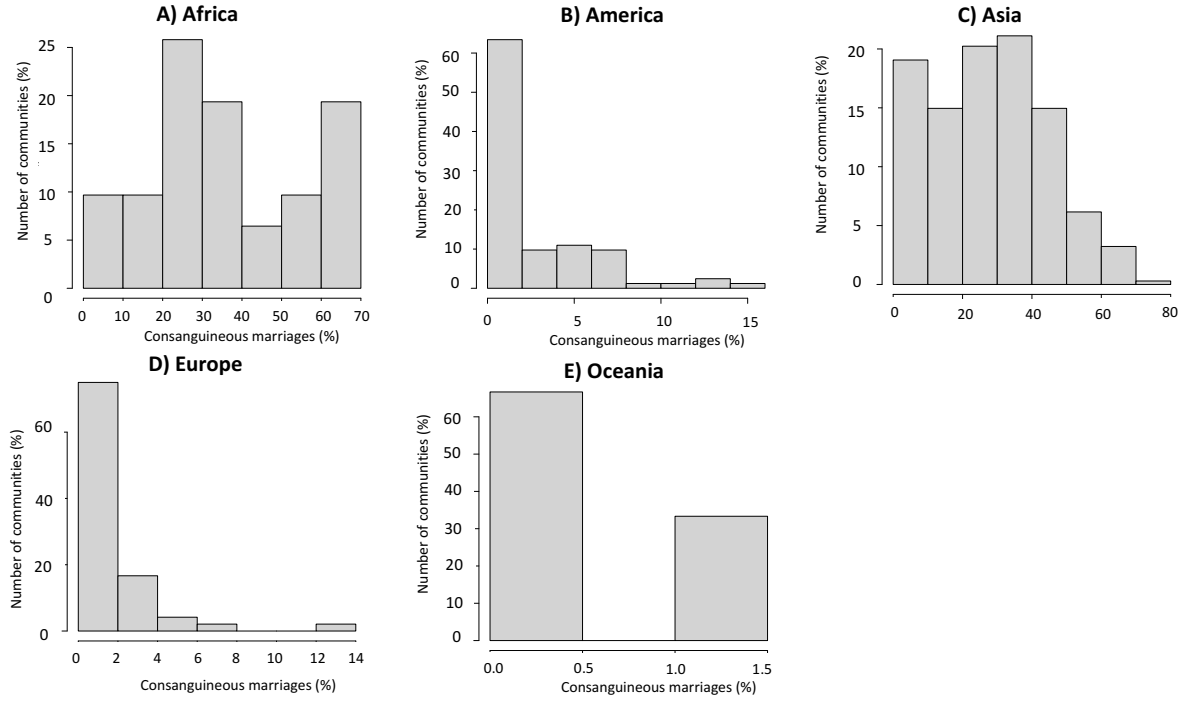


Figure III.9 – Distribution des pourcentages de consanguinité par continent.

Motivations socio-économiques

Pour expliquer ces écarts dans la pratique de la consanguinité, il faut se pencher sur des raisons

- d'ordre religieux ou coutumier poussant les partenaires à se conformer à des schémas matrimoniaux traditionnels ;
- sociales dans le but de renforcer des liens familiaux ;
- économiques en facilitant les négociations pré-nuptiales et baissant le prix de la dot ou pour conserver un patrimoine au sein de l'entité familiale.

En dépit de la dépression biologique de consanguinité, les individus mariés à des apparentés bénéficient de certains avantages socio-économiques. En particulier, les femmes mariées à un apparenté font déjà partie de leur belle-famille avant même leur mariage et connaissent leur époux, et ainsi entretiennent de meilleurs liens avec leurs beaux-parents et époux, ce qui contribue à diminuer le taux de divorce et répudiation et à améliorer le statut des femmes (Bittles 2012 ; Bittles et Black 2010).

La validité de ces motivations socio-économiques, leur prévalence et les facteurs les justifiant sont intrinsèques à chaque communauté humaine, ce qui peut expliquer les grandes variations quant à la pratique de la consanguinité entre différentes communautés. Par exemple, au Japon, la consanguinité varie en fonction des groupes religieux étudiés : les bouddhistes sont plus consanguins que les catholiques

(Chapitre 5 de Bittles (2012)).
Parmi les normes des populations, des types de mariages consanguins préférentiels peuvent être spécifiés : mariages avunculaires chez les Wolofs (Ghasarian 1996), entre frères et soeurs dans la ville d'Arsinoé en Égypte ptolémaïque (Ager 2005) ou entre cousins dans de nombreuses sociétés. Ce dernier type matrimonial est le plus fréquent dans notre espèce (Chapitre 5 - (Bittles 2012)), représentant jusqu'à 50% des mariages au Pakistan, et ce, sans déclin depuis les années 1950 (Chapitre 2 par Bittles dans (Wolf 2004)). Dans le détail, on distingue plusieurs types de mariages entre germains dont la prévalence change selon les régions (Pison 1986) : dans le "mariage arabe", les hommes épousent préférentiellement leurs cousines parallèles FBD (pour *Father's Brother's daughters*, ou bint' ammi), tandis qu'au sud de l'Inde, ou chez les Hans ou les Touaregs, ce sont les cousines croisées MBD (*Mother's Brother's daughters*) (Chapitre 2 par Bittles dans (Wolf 2004)).
Cependant, des travaux menés chez les Dogons de Tabi du Mali montrent que dans la réalité les mariages entre cousins répondent à des contraintes démographiques (entre des partenaires appartenant à la même classe d'âge) plutôt qu'à un choix préférentiel d'un type d'union (Cazes 1981).

Consanguinité de dérive

Choisir d'épouser un apparenté (*mating choice inbreeding* ou *parental relatedness* (Pemberton *et al.* 2012)) constitue un écart à la panmixie, généralement plus marqué dans les campagnes qu'en ville. Cette tendance pourrait s'expliquer par des motivations socio-économiques plus traditionnelles dans les campagnes (Ben M'Rad et Chalbi 2006) mais aussi par un choix des partenaires limité. En effet, certains villages fonctionnent comme des entités matrimoniales closes, n'échangeant que peu d'individus avec l'extérieur. Par exemple, dans un village des Alpes suisses, pendant 300 ans, seulement 129 personnes ont immigré et sont peu intervenues dans les mariages recensés (75% des 900 mariages n'impliquaient pas d'immigrants ou de descendants d'immigrants) (Hagaman *et al.* 1978)). Un tel repli peut causer une pénurie de partenaires non apparentés du même âge et/ou de même condition sociale, ce qui était un motif valable aux yeux de l'Église catholique romaine pour obtenir une dispense de mariage consanguin (pour cause de "petitesse"). La consanguinité ainsi provoquée est appelée consanguinité de dérive (*drift or background relatedness inbreeding* (Pemberton *et al.* 2012)) et survient au bout d'un temps proportionnel à l'effectif initial de la population et à son degré d'isolement (Roberts 1976).
Ainsi, la consanguinité mesurée d'une population, dite totale, peut se décomposer en deux termes dont l'un est la consanguinité de dérive (Jacquard 1968b) :

$$(1_\alpha) = (1 - D_\alpha)(1 - \delta) \tag{III.7}$$

où α est la consaguinité totale de la population, D_α est la consanguinité de dérive,

et δ est le coefficient d'écart à la panmixie, soit une estimation de l'influence du choix du conjoint.

Plus en détail, on peut distinguer la consanguinité de dérive proche, lorsque le manque de choix de partenaires pousse des individus récemment apparentés à se marier (par exemple à Tristan da Cunha en 1856 où des cousins n'ont eu d'autres choix que de se marier alors que la consanguinité avait été auparavant évitée (Roberts 1976)), et la consanguinité de dérive éloignée lorsque les époux sont des individus non apparentés sur les dernières générations mais qui partagent de nombreux ancêtres lointains (Mourali-Chebil et Heyer 2006).
Cette consanguinité de dérive éloignée est généralement la cause de l'augmentation de la consanguinité observée dans les populations humaines : au fil du temps, la consanguinité totale d'une population isolée

augmente, tandis que le nombre de mariages consanguins entre des individus apparentés à quatre générations se maintient voire diminue (Figure III.10) (Bideau *et al.* 1994). C'est alors le nombre d'ancêtres communs entre les conjoints à plus de quatre générations qui augmente (O'Brien *et al.* 1988 ; Mourali-Chebil et Heyer 2006). A. Jacquard a calculé que dans une population de 100 individus, se mariant entre cousins germains ou doubles cousins germains pendant 40 générations, le coefficient de consanguinité final d'un descendant équivalait à celui d'un descendant de frère/sœur classique (Jacquard 1968a).
Le poids de la consanguinité de dérive éloignée est d'autant plus important pour des populations isolées, comme celles de villages des Pyrénées (Serre *et al.* 1985) ou du Québec où 63% de la consanguinité totale est due à de la consanguinité de dérive éloignée (Bideau *et al.* 1994). Au contraire, dans des villages moins isolés, comme le village breton de Plozevet (Jakobi et Jacquard 1971) ou la vallée de la Valserine (Bideau *et al.* 1994), la part de la consanguinité de dérive dans la consanguinité totale est plus faible.

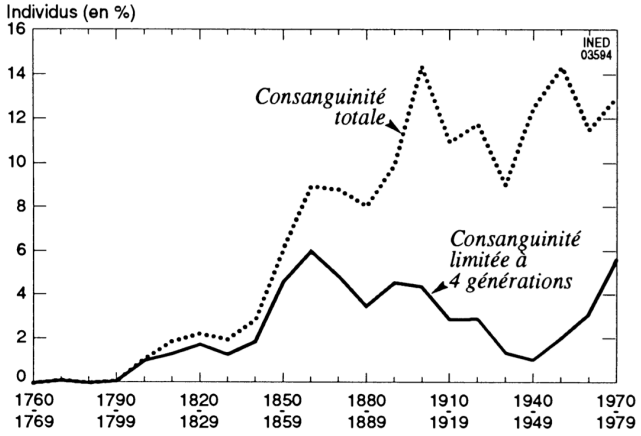


Figure III.10 – Évolution de la consanguinité totale et de la part de la consanguinité proche dans la population de la vallée de la Valserine, dans le Jura, du XVIII^{e} au XX^{e} siècle. La différence entre les deux courbes est la part de la consanguinité de dérive éloignée. *Figure 4 de Bideau* et al. *(1994).*

Approche de génétique des populations

Les données génétiques ont donné un nouveau souffle à l'étude de la consanguinité des populations, permettant de s'affranchir d'une collecte chronophage de données généalogiques à partir des registres paroissiaux ou civils.

L'intérêt est surtout de pouvoir accéder à différents paramètres de la consanguinité, comme

- les écarts à la panmixie à partir d'approches simple-points (Gazal *et al.* 2014b) ;
- la consanguinité proche au moyen de ROHs longs : une étude menée sur la population orcadienne du jeu de données ORCADES a montré que les niveaux de consanguinité proche estimés à partir de la généalogie et la proportion du génome contenue dans des ROHs d'une taille supérieure à 1,5 Mb sont fortement corrélées (McQuillan *et al.* 2008) ;
- la consanguinité de dérive éloignée au moyen de ROHs de taille intermédiaire (Pemberton *et al.* 2012).

Les données génétiques ont facilité la comparaison des niveaux de consanguinité proche et de dérive entre les populations de diverses régions sans avoir besoin de données ethnologiques. En particulier, des travaux menés sur les populations des jeux de données HGDP-CEPH, HapMap et 1000Genomes ont montré l'omniprésence des unions consanguines (Leutenegger *et al.* 2011 ; Pemberton *et al.* 2012 ; Gazal *et al.* 2015). À l'échelle de la planète, c'est près d'un quart des 2500 individus séquencés par le projet 1000 Genomes qui est au moins autant consanguin que des descendants de cousins issus de germains

(Gazal *et al.* 2015).
Néanmoins, les auteurs observent des différences de fréquences et des types d'unions consanguines selon les populations et les régions : les populations les plus consanguines sont échantillonnées au Proche-Orient, en Asie du sud où les unions préférentielles se font entre cousins, et en Amérique (chez des populations amérindiennes, où les unions se font plutôt entre avunculaires) (Figure III.11).

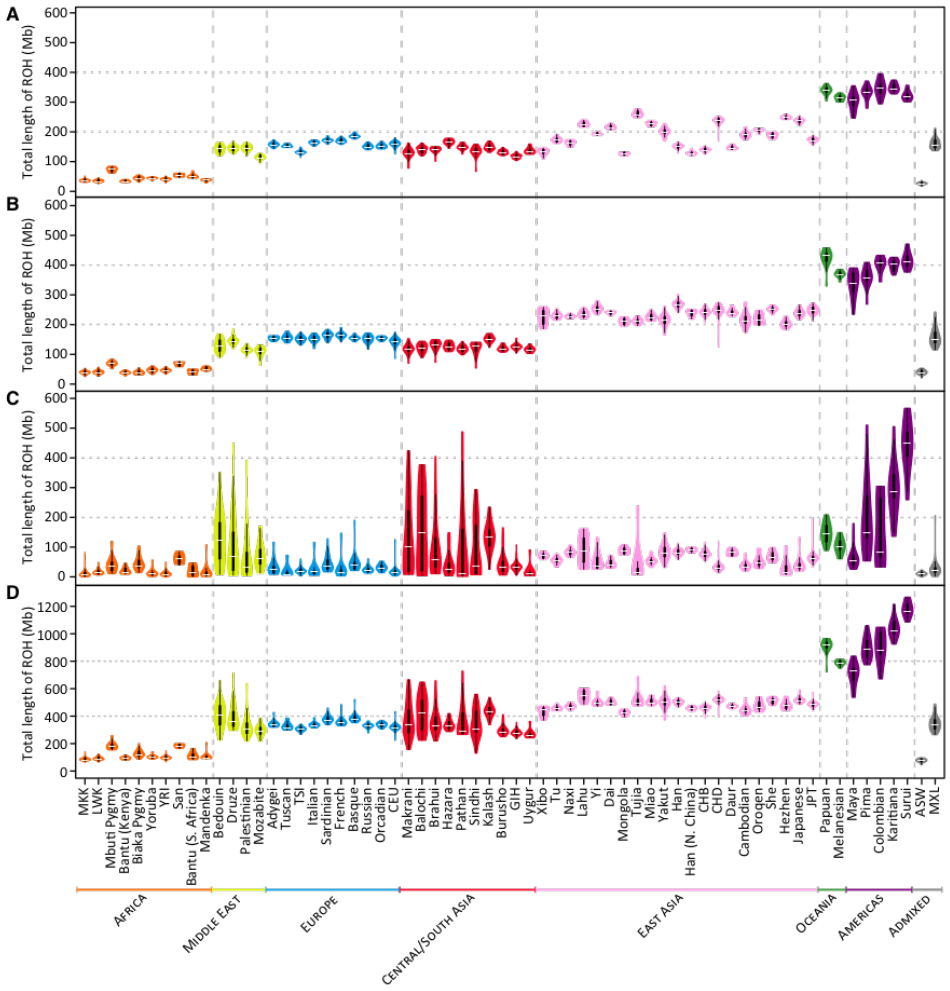


Figure III.11 – Distribution des tailles totales des ROHs trouvés au sein du génome de chaque individu, courts (A), de taille intermédiaire (B) ou longs (C) vraisemblablement causés par de la consanguinité récente. En D), tous les ROHs sont regroupés. *Figure 2 de Pemberton* et al. *(2012).*

Les données génétiques permettent aussi d'obtenir des informations sur la consanguinité éloignée, qui n'est en général pas accessible par une approche généalogique sauf pour des populations dont on connait toute la généalogie et la démographie. Ainsi, en comparant des communautés avec une longue histoire de consanguinité (comme c'est le cas au Pakistan ou au Proche-Orient) et des communautés avec une consanguinité plus récente ("Travellers" irlandais), les descendants de cousins germains dans les communautés avec une consanguinité prolongée présentent 5% de sites homozygotes en plus que des "Travellers" issus de cousins germains (Woods *et al.* 2006).
Ces loci autozygotes dus à de la consanguinité de dérive sont intégrés à des ROHs de taille intermédiaire, dont le nombre et la longueur varient selon les populations selon le degré d'endogamie (Pemberton *et al.* 2012) (Figure III.11 B). Cependant, la variation observée pour ces ROHs au sein d'une population est

plus faible que celle des longs ROHs car la consanguinité de dérive affecte toute la population. Cependant si la population n'est pas homogène et inclut des individus issus d'unions endogames (dont les quatre grands-parents étaient nés sur la même île des Orcades appelés *Endogamous Orcadians*) et d'autres issus d'unions entre des autochtones et des migrants plus ou moins proches (les *Mixed Orcadians* ont au moins trois de leurs grand-parents originaires des Orcades mais d'îles différentes, et *Half Orcadians* ont deux de leurs grand-parents originaires des Orcades et les deux autres de l'île principale de l'Écosse), on peut voir des différences de niveau de consanguinité de dérive dans les génomes (Figure III.12) (McQuillan *et al.* 2008). En effet, les descendants d'endogames ont beaucoup plus de ROHs intermédiaires (plus courts que 50 Mb) que les descendants de migrants. Cet article démontre élégamment que l'endogamie est la cause de l'augmentation de la consanguinité de dérive et que l'exogamie hors des Orcades, ou juste hors de la même île, suffirait à l'éviter.

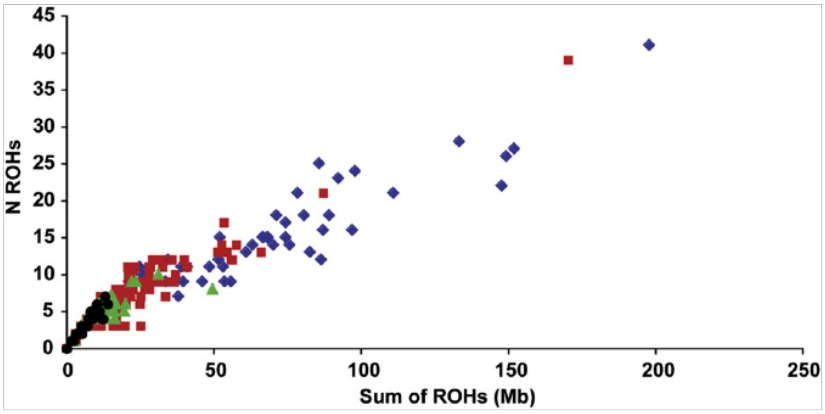


Figure III.12 – Effet du métissage génétique ou de l'endogamie géographique sur les ROHs (nombre et taille). Les ROHs considérés dans cet article sont ceux plus longs que 500 kb. Les Orcadiens endogames consanguins (issus de cousins germains ou au second degré) sont représentés en bleu et les non-consanguins en rouge ; les *mixed Orcadians* sont en vert et les *half Orcadians* en noir. *Figure 4 de McQuillan* et al. *(2008).*

Ainsi, en couplant des ROHs de taille intermédiaire et longs, on peut dresser le profil de consanguinité des populations étudiées : par exemple pour le jeu de données HGDP-CEPH, les populations d'Asie du sud et du Proche-Orient présentent principalement de la consanguinité due à des unions entre apparentés proches ; la consanguinité de dérive prédomine chez les Européens, chez les populations d'Asie de l'est, chez les agriculteurs sub-sahariens et chez les populations océaniennes ; les populations amérindiennes ont à fois de la consanguinité de dérive et récente, toutes deux particulièrement élevées, rendant compte de leur isolement (Kirin *et al.* 2010).

Éviter la consanguinité

Puisque la dépression de consanguinité apparaît comme une menace pour les individus et les populations, on peut s'attendre à ce que des comportements d'évitement de la consanguinité aient émergé au cours de l'évolution. Effectivement, les animaux en captivité font de moins bons géniteurs quand ils se reproduisent entre apparentés qu'entre non-apparentés, avec un temps nécessaire à la reproduction allongé et une descendance moins nombreuse (Blouin et Blouin 1988 ; Muniz *et al.* 2006). Outre des mécanismes post-copulatoires, des mécanismes pré-copulatoires d'évitement de la consanguinité existent et peuvent être répartis en deux grandes classes (Blouin et Blouin 1988 ; Archie *et al.* 2007) :

- ceux évitant la reproduction entre des apparentés au sein d'une même population. Cela repose sur une capacité à reconnaître ses apparentés, puis à se comporter différemment avec eux qu'avec les non-apparentés (Hamilton et Vonk 2015 ; Pusey et Wolf 1996), notamment (mais pas systématiquement) à travers des temps d'approche plus longs, une moindre fidélité, de l'évitement physique en période reproductive ou encore une agressivité plus appuyée (Simmons 1989) ;
- ceux réduisant le contact entre les apparentés de sexe opposé, principalement par de la séparation géographique et temporelle.

Évitement intra-populationnel

Reconnaissance phénotypique des apparentés

Chez les animaux mais aussi chez l'Homme, plusieurs indices phénotypiques renseignent sur la parenté (*Box 2* de Hauber et Sherman (2001)). Ces indices doivent être héritables, et nécessitent un apprentissage et une reconnaissance du phénotype des parents ou de son propre phénotype pour créer un référentiel.
Ces indices peuvent être faciaux chez l'Homme (DeBruine *et al.* 2008 ; Alvergne *et al.* 2014) mais aussi chez d'autres primates, ce qui permet notamment aux macaques rhésus d'identifier des demi-frères paternels sans jamais les avoir rencontrés auparavant (Pfefferle *et al.* 2014).
Ils peuvent également être auditifs, comme les vocalisations chez les macaques rhésus et les mandrills (Pfefferle *et al.* 2013 ; Levréro *et al.* 2015)).

La reconnaissance phénotypique peut aussi se faire par des messages chimiques, *a priori* olfactifs, comme chez le hamster dont les individus reconnaissent leurs apparentés sans indice visuel ni auditif et sans jamais les avoir rencontrés auparavant (Mateo et Johnston 2000). Ces messages chimiques pourraient être également impliqués dans la reconnaissance de l'odeur maternelle chez le chiot (Hamilton et Vonk 2015). De manière inattendue, les oiseaux aussi semblent utiliser les odeurs pour identifier leurs apparentés (Coffin *et al.* 2011). Chez l'Homme, des expériences ont aussi montré la capacité des parents à reconnaître olfactivement leurs descendants biologiques (Alvergne *et al.* 2009) et celle des enfants à identifier leurs frères et sœurs (Weisfeld *et al.* 2003).
Des complexes de gènes polymorphes seraient à l'origine de ces signaux chimiques : le MHC humain (Complexe d'Histocompatibilité Majeur) qui aiderait à l'évitement de la consanguinité chez les Européens américains du jeu de données HapMap (Laurent et Chaix 2012) ou chez les mandrills (Setchell *et al.* 2013), et le MUP (Protéine Urinaire Murine) chez les souris qui permet aux femelles d'identifier leurs sœurs pour collaborer (Green *et al.* 2015).

Évolutivement, l'identification des apparentés n'intervient pas exclusivement dans l'évitement de la consanguinité mais aussi dans d'autres fonctions telles que (Perrin et Lehmann 2001)

- l'altruisme : dans la formule énoncée par Hamilton (1987), un individu développe d'autant plus un comportement altruiste envers un autre individu qu'il lui est apparenté ;

- la coopération, pouvant prendre une forme d'entraide mais aussi de comportements d'évitement pour réduire la compétition avec certains individus pour la nourriture ou la reproduction ;
- le soin parental en identifiant sa propre descendance, à l'image de l'autruche dominante qui malgré un nid communal couve en priorité ses propres œufs ;
- la reconnaissance de ses parents et donc l'accès aux soins parentaux.

La prévalence de ces différentes fonctions varient selon le sexe des acteurs et/ou le moment de leur cycle de vie : les guppy se regroupent avec leurs frères hypothétiquement pour collaborer quand ils sont juvéniles, mais à l'âge adulte ils se séparent et forment des bancs avec des non-apparentés, peut-être pour diminuer la compétition entre apparentés (Hain et Neff 2007) ; les mâles souris grises, sentant un signal d'œstrus, privilégient le contact avec des femelles non-apparentées, alors qu'habituellement, ils passent autant de temps avec leurs sœurs qu'avec leurs non-apparentées (Krackow et Matuschak 1991).

Il existe aussi des moyens indirects d'éviter ses apparentés (Hauber et Sherman 2001) :

Copulation hors du groupe

Certaines espèces, comme la chauve-souris *Plecotus auritus* (Burland *et al.* 2001) ou les macaques rhésus (Widdig *et al.* 2017), ont développé des stratégies de copulation discrète hors du groupe, avec des géniteurs *a priori* moins apparentés qu'au sein de leur groupe.

Familiarité

L'association pendant l'enfance, appelée familiarité, supprimerait la volonté de se reproduire avec ses familiaux qu'ils soient ou non apparentés (Blouin et Blouin 1988) : ce phénomène est décrit dans certaines espèces animales, comme les écureuils (Holmes et Sherman 1982) mais également chez l'Homme, par le Dr. Westermarck qui note "une remarquable absence de sentiments érotiques entre personnes ayant vécu étroitement ensemble depuis l'enfance" (Westermarck 1921).

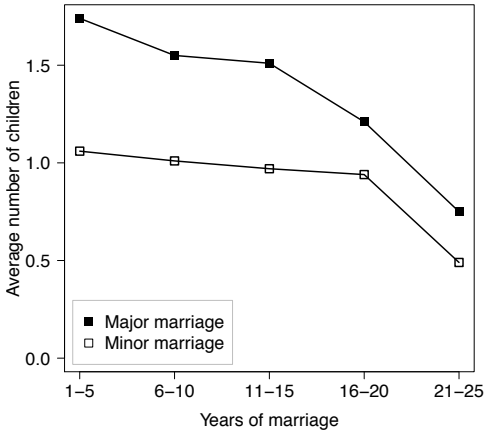


Figure III.13 – Différence du nombre d'enfants par famille pour des partenaires *Minor* élevés ensemble durant leur enfance et *Major* ne se connaissant que depuis l'age adulte. Étude menée entre 1900 et 1925 sur 303 couples dont 132 Minor dans un village chinois de Taiwan. *Chiffres tirés de Wolf (1970) - Table 7.*

L'effet Westermarck a été documenté dans le cas des kibboutz : les enfants non apparentés élevés dès leur plus jeune âge de façon communautaire ne se marient pas ensemble (Talmon 1964). Cela a aussi été observé au sein d'un village chinois de Taïwan où deux types de mariages coexistent : des mariages classiques , dits *Major* survenant entre des personnes adultes et s'étant peu cotoyées avant le jour de la

noce, et ceux *Minor*, entre adultes ayant passé leur enfance ensemble au sein de la même famille (dans ce type de mariage, la future mariée est adoptée très jeune et élevée par sa belle-famille). L'effet décrit par Westermarck s'observe par une diminution du nombre d'enfants par famille (Figure III.13), une augmentation des divorces, des séparations (24% contre 1%) et d'actes adultères pour les couples *Minor* par comparaison aux *Major* sans que cela soit imputable à des différences socio-économiques (Wolf 1970).

Reconnaissance anthropisée

L'Homme a développé des stratégies d'évitement qui lui sont propres, basées notamment sur la connaissance des liens de parenté. Ces liens sont utiles au fonctionnement des sociétés et sont à la base des rapports sociaux d'alliance, de résidence, de transmission des biens, des terrains ou des statuts (Ghasarian 1996). Pour l'ethnologue ou l'anthropologue de la parenté, il est nécessaire d'identifier la structure de parenté pour comprendre l'organisation sociale du groupe. Notamment, l'étude du vocabulaire de la parenté est cruciale : les termes utilisés reflètent les liens entre les parents sociaux et donc la structure sociale, et sont différents selon les sociétés notamment en distinguant ou non les consanguins des alliés (en français, le mot "tante" se réfère aussi bien à la femme de l'oncle qu'à la sœur des parents).
De cette organisation de la parenté, découle la notion d'inceste social qui prohibe les relations sexuelles entre certains apparentés (Godelier 2004). La liste des apparentés à éviter varie selon les sociétés mais le concept est universel. Notons que l'inceste ne concerne pas exclusivement des apparentements biologiques : par exemple en France, le code civil interdit de se marier avec les enfants d'un beau-parent, ou avec un frère adoptif. Au contraire, dans certaines sociétés souvent unilatérales, des parents biologiques peuvent ne pas être considérés comme des apparentés sociaux notamment si les contacts sont rares, par exemple en cas de résidence dans des villages différents, et ne sont donc pas concernés par les interdits de mariage.
La connaissance des liens de parenté peut reposer sur la base de l'oral mais aussi sur des traces écrites. Ainsi, la généalogie complète islandaise est inscrite dans un livre et a récemment été mise en ligne (https://www.islendingabok.is/English.jsp), permettant ainsi à deux personnes de connaître facilement leur degré d'apparentement au moyen d'un ordinateur et même de leurs smartphones (par le biais de l'application Islendinga), et de fait d'éviter la consanguinité due au hasard des rencontres.

À travers le conseil génétique pré-nuptial ou anté-natal, l'Homme a aussi développé des méthodes génétiques permettant d'estimer le degré d'apparentement des conjoints potentiels et/ou les risques pour leur descendance d'être homozygote pour certaines mutations récessives délétères (Hamamy 2012).

Évitement par séparation spatiale des apparentés de sexe opposé

La seconde grande catégorie de comportements d'évitement de la consanguinité est la séparation des apparentés de sexe opposé dans l'espace par la dispersion.

Dispersion

La dispersion correspond au départ permanent d'un individu hors d'un groupe et donc à une reproduction hors de son groupe de naissance, dans un seul (*natal dispersion*) ou plusieurs nouveaux groupes (*breeding dispersion*) (Ronce 2007).
Par espèce, en général un seul sexe disperse majoritairement : les femelles chez les oiseaux, les wombats (Banks *et al.* 2002) et les chimpanzés ; les mâles chez la plupart des mammifères et primates polygynes (Figure III.14) (Greenwood 1980 ; Pusey et Wolf 1996). Notons que la dispersion n'est pas nécessairement complète pour le sexe dispersant : certains individus de ce sexe peuvent rester dans leur lieu natal, et il arrive qu'une moindre proportion de l'autre sexe disperse aussi (Pusey 1990 ; Pusey et Packer 1987).

La dispersion et l'installation au sein d'un nouveau groupe sont des prises de risque qui augmentent la mortalité pendant le trajet, le risque d'être agressé dans le nouveau groupe, et supprime l'entraide que l'on aurait connu auprès de ses apparentés. Ces désavantages pourraient expliquer la dispersion d'un seul sexe, permettant à l'autre de bénéficier des avantages de la philopatrie (Pusey et Packer 1987 ; Banks *et al.* 2002). Bien que dispersant, certains individus peuvent rechercher une coopération en migrant dans des groupes contenant certains de leurs apparentés de même sexe (Arandjelovic *et al.* 2014) : 40% des femelles *Gorilla beringei* ont au moins une apparentée dans le groupe où elles se sont installées (Roy *et al.* 2014).

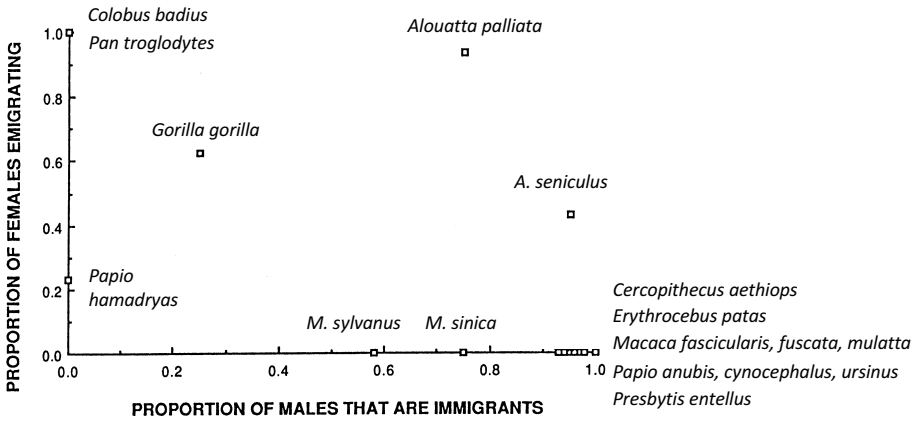


Figure III.14 – Dispersion de femelles et mâles pour différentes espèces de primates. Dans chaque groupe d'animaux, la proportion de femelles émigrantes observée ainsi que celle des mâles immigrants *Figure 7.1 de Pusey (1990).*

D'après plusieurs études, la dispersion pourrait être une stratégie d'évitement de la consanguinité, mais d'autres thèses la considèrent plutôt comme un moyen d'éviter de la compétition entre apparentés de même sexe pour les ressources et la reproduction. La prévalence des deux mécanismes, sans qu'ils soient nécessairement exclusifs, dépendrait de l'espèce, du sexe dispersant et du moment de la vie des individus (Perrin et Mazalov 2000 ; Zhao *et al.* 2008). Plusieurs exemples vont dans le sens de l'hypothèse d'évitement de consanguinité : les femelles babouins *cynocephalus* restent dans leur groupe natal quelque soit le niveau de compétition entre elles mais les mâles quittent leur groupe même si aucun autre mâle n'y est installé (Pusey 1990) ; chez les chevaux sauvages, les juments émigrent même si la densité de population est faible dans leur groupe natal et que les ressources sont abondantes (Clutton-Brock et Lukas 2012). Un autre argument en faveur de l'hypothèse d'évitement de la consanguinité est le moment choisi par les animaux pour disperser : dans certaines espèces, les parents migrent quand leurs enfants de sexe opposé atteignent la maturité (Pusey et Packer 1987), et les femelles chimpanzés émigrent quand elles réalisent leur premier cycle d'œstrus complet et s'éloignent donc de leurs frères (Pusey 1980).

Reproduction supprimée en présence d'apparentés de sexe opposé

Chez plusieurs espèces animales, les femelles connaissent une inhibition de leur œstrus, comme chez les chiens de prairie (Hoogland 1982), ou une maturité sexuelle retardée chez les tamarins (Blouin et Blouin 1988) tant que leurs frères ou pères sont présents dans la population. Cette solution est donc temporaire, dans l'attente d'une séparation des apparentés de sexe opposé.

L'exogamie géographique humaine

La dispersion chez l'Homme peut prendre la forme d'unions exogames, avec un partenaire originaire d'un autre groupe, qu'il soit généalogique, social, linguistique, religieux ou géographique. En particulier, l'exogamie géographique dépend de plusieurs facteurs comme l'éloignement géographique entre villages (Lathrop et Pison 1982), la connaissance que l'on a des villages voisins (Boyce *et al.* 1967), et les règles d'alliance ayant cours (Segalen 1986).
En terme de consanguinité, ce comportement migratoire est considéré comme un moyen d'éviter les unions entre "gens de même sang" (Morgan 1871), et certains groupes très endogames comme les berbères Kel Kummer (Jacquard 1972) ou les Indiens Jicaques du Honduras (Jacquard 1974) ont été étudiés car supposés très consanguins. En ce qui concerne la France au 20^{e} siècle, l'essor de l'exogamie de village serait responsable de la diminution drastique de la consanguinité (de $84 * 10^{-5}$ en 1926 à $22 * 10^{-5}$ en 1956 (Jacquard et Reynès 1968)). Cette ouverture aurait notamment fait fortement baisser le nombre de mariages entre cousins germains (Figure III.15 - (Jacquard et Reynès 1968)). Ce phénomène a eu lieu à l'échelle nationale et à l'échelle départementale (Sutter et Tabah 1955), et a gommé les disparités de consanguinité entre départements observées avant 1945 (Figure III.15) (Tabah et Sutter 1950).

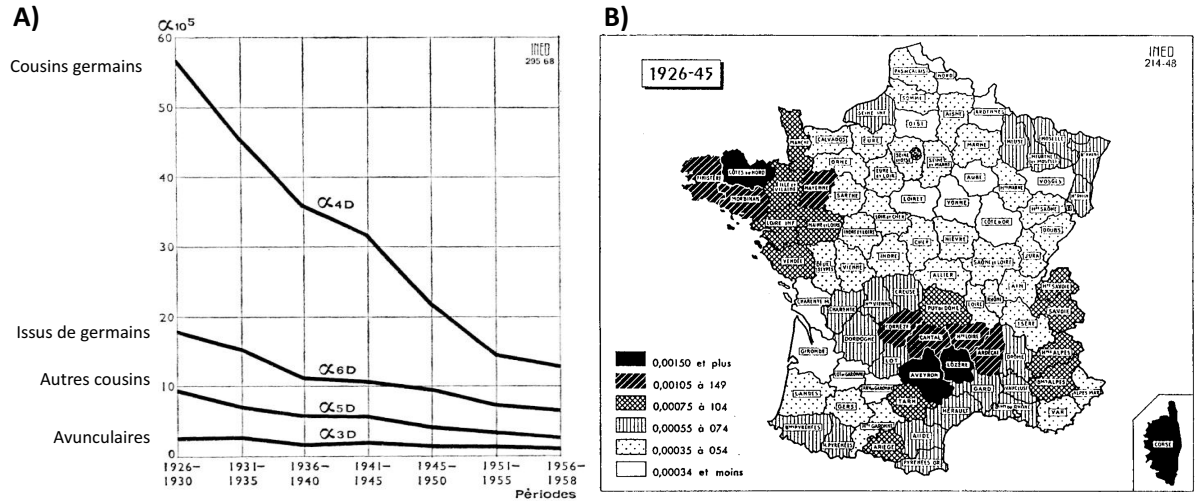


Figure III.15 – Diminution de la consanguinité en France entre 1926 et 1958, détaillée par type d'unions consanguines, et coefficient de consanguinité par département français entre 1926 et 1945. La consanguinité mesurée ici est basée sur la proportion de mariages consanguins recensés. *Graphique 2 de Jacquard et Reynès (1968) et Graphique 3 de Tabah et Sutter (1950).*

Cependant, certaines motivations socio-économiques intervenant chez l'Homme, comme la création ou le renchaînement d'alliances (Brudner et White 1997 ; Segalen 1985) pourraient changer le lien supposé entre consanguinité et exogamie : les mariages exogames réalisés pourraient avoir lieu de manière récurrente et privilégiée entre les mêmes groupes conduisant à des proximités génétiques en dépit d'un éloignement géographique. Notre travail mené en Asie intérieure confronte les pratiques culturelles, à travers les unions conclues et la pratique de l'exogamie, et les attendus de la biologie évolutive pour la consanguinité.

III.2 Consanguinité et exogamie en Asie intérieure

Close inbreeding and low genetic diversity despite geographical exogamy in Inner Asian human populations - Résumé -

En Asie intérieure, les deux groupes linguistiques et culturels différent notamment par leurs règles d'alliance : traditionnellement et majoritairement, les Turco-Mongols se marient de manière exogame, hors de leur village natal, tandis que les Indo-Iraniens se marient de manière endogame au sein du même village. Ces règles d'alliance devraient donc conduire à des migrations matrimoniales plus nombreuses chez les Turco-Mongols que chez les Indo-Iraniens.

En partant de l'hypothèse que des individus nés dans des villages différents sont moins apparentés que ceux nés au sein des mêmes villages, la dispersion (ici sous la forme d'exogamie géographique) devrait permettre d'éviter les unions consanguines et augmenter la diversité génétique des populations recevant des migrants. Afin d'explorer cette hypothèse, généralisée à tout le règne animal mais à notre connaissance jamais testée chez aucune espèce, nous estimons :

1. le niveau d'exogamie des populations ;
2. la consanguinité des individus composant la population ;
3. le niveau de diversité des populations ;
4. la relation entre exogamie et consanguinité, à l'échelle des individus et des populations.

Matériel et Méthodes

Notre étude combine des données de dispersion et des données génétiques pour 16 populations d'Asie intérieure (Figure 1 de l'article) : 604 couples ont répondu au questionnaire ethno-démographique parmi lesquels 503 individus ont été génotypés sur des puces à ADN.

Estimer l'exogamie géographique

Dans le questionnaire ethno-démographique, les individus nous renseignent sur leur lieu de naissance, celui de leur conjoint et ceux de leurs parents et beaux-parents. Dans un premier temps, nous avons retrouvé les coordonnées géographiques de la plupart des lieux recensés. Puis, à partir de ces coordonnées, nous avons calculé une distance géographique pour chaque couple, à la génération échantillonnée et à la précédente (celle des parents et beaux-parents), en appliquant la loi des cosinus pour une sphère de 6367,445 km de rayon.

Nous avons ensuite comparé les distributions de ces distances pour les groupes turco-mongol et indo-iranien, à chaque génération (Figure 2 et S2 de l'article). Pour manipuler plus aisément ces informations de distance, nous les avons converties en informations binaires : les couples sont endogames ou exogames si les lieux de naissance entre les partenaires sont respectivement en-dessous ou au-dessus d'un seuil d'exogamie fixé. Nous avons choisi arbitrairement différents seuils d'exogamie (à 10, 20, 30, 40 ou 50 km) et déterminé un seuil de 4 km d'après l'observation de nos données. Nous avons ainsi pu calculer un pourcentage de couples exogames pour chaque population (Figure S1 de l'article).

Mesurer la consanguinité des individus

Pour les 503 individus génotypés sur des puces à ADN, nous avons obtenu des informations *genome-wide* pour 242 406 SNPs autosomaux, dont 105 858 SNPs indépendants (pour un $r^2<0.5$).

Ces données génétiques nous permettent de calculer différents estimateurs de la consanguinité individuelle :

- la proportion de sites homozygotes par génome grâce au logiciel Plink (Purcell *et al.* 2007) ;
- le coefficient simple-point de Plink qui mesure un excès ou déficit d'homozygotie par rapport à des fréquences alléliques de référence. Dans ces travaux, nous avons choisi d'utiliser les fréquences calculées pour chaque population ;
- le nombre de ROHs et leur longueur génomique pour chaque individu. Nous nous sommes intéressées à deux types de ROHs : les "longs", vraisemblablement causés par de la consanguinité proche et les "intermédiaires" associés à de la consanguinité éloignée (Pemberton *et al.* 2012). Pour discerner ces deux classes de ROHs, nous avons utilisé (i) des seuils fixes (intermédiaires : entre 500 et 1500 kb, longs : au-dessus de 1500 kb), et (ii) des seuils estimés pour chaque population d'après la méthode publiée par Pemberton *et al.* (2012) ;
- le coefficient F-Median calculé au moyen du logiciel FSuite (Leutenegger *et al.* 2003 ; Gazal *et al.* 2014b), qui infère le statut d'autozygotie des loci en utilisant des informations multi-points et les fréquences alléliques ;
- une inférence du type d'union dont est issu l'individu : union avunculaire, entre doubles cousins, cousins germains ou issus de germains, ou d'une union moins apparentée que cela, notée "outbred".

Mesurer la diversité des populations

Par population, nous avons calculé le nombre médian de dissimilarités génomiques entre individus (c'est-à-dire la médiane des ASD mesurées entre les individus d'une même population (Szpiech 2011)). Nous avons aussi calculé l'hétérozygotie haplotypique de chaque population en nous inspirant de Verdu *et al.* (2014), afin de limiter l'effet de biais d'*ascertainment* par une approche multi-points. Ces estimateurs devaient nous renseigner sur la diversité des populations, et éventuellement sur leur niveau de consanguinité éloignée.

Résultats

À propos de l'exogamie géographique

L'analyse des distances géographiques entre les lieux de naissance des conjoints a confirmé les migrations matrimoniales décrites en Asie intérieure : les lieux de naissance de partenaires turco-mongols sont plus éloignés géographiquement que ceux de conjoints indo-iraniens (Figure 2 de l'article). En choisissant un seuil d'exogamie à 4 km, les niveaux d'exogamie par population sont statistiquement plus élevés pour les populations turco-mongoles qu'indo-iraniennes (test de Mann-Whitney unilatéral : p-value=0,021), avec une certaine variation au sein des groupes (Figure S1 de l'article). Cette différence entre les groupes culturels se retrouve également à la génération précédente, avec des taux d'exogamie fortement corrélés d'une génération à l'autre (ρ de Spearman=0,81). Nous avons également retrouvé cette tendance pour les autres seuils d'exogamie, choisis arbitrairement cette fois-ci : 10, 20, 30, 40 et 50 km.

À propos de la diversité génétique des populations

Conformément à ce qui avait été observé en Asie intérieure dans la littérature et au cours de ce manuscrit, les populations turco-mongoles, pourtant plus exogames, présentent moins de diversité génétique que les populations indo-iraniennes. Cela se traduit par des pourcentages de sites homozygotes globalement plus élevés pour les populations turco-mongoles que pour les populations indo-iraniennes (Figure III.16), et des hétérozygoties haplotypiques (Figure 3A de l'article) et ASD médianes (Figure S3 de l'article) plus faibles.

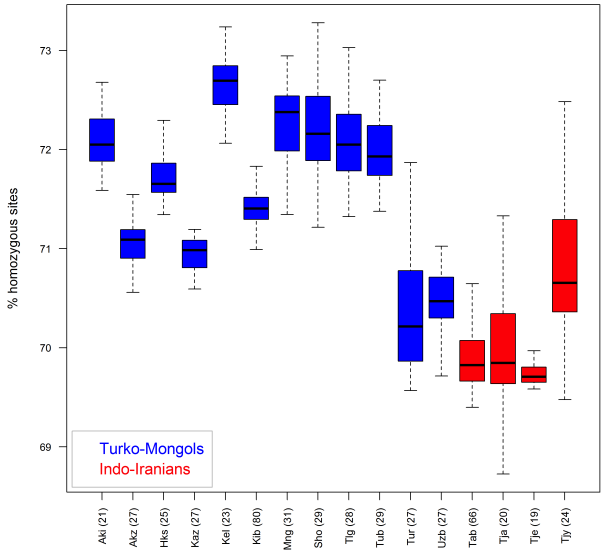


Figure III.16 – Pourcentage de sites homozygotes par génome. Les individus sont regroupés par population, en bleu les Turco-Mongoles, en rouge les Indo-Iraniennes. Les individus turco-mongols ont statistiquement plus de loci homozygotes dans leur génome que les Indo-Iraniens (Test de Mann-Whitney bilatéral : p-value $<10^{-46}$).

Cette faible diversité génétique chez les Turco-Mongols est associée à une consanguinité de dérive éloignée plus marquée, mesurée par des ROHs de taille intermédiaire statistiquement plus nombreux et représentant une plus grande part du génome (Figure S5 de l'article).
Cependant, en comparant le niveau d'exogamie géographique et ces mesures de diversité intra-populationnelle nous n'avons pas observé de corrélation significative au sein de chaque groupe linguistique : l'exogamie ne serait donc pas un facteur accroissant la diversité génétique à l'échelle des populations (Figure III.17).

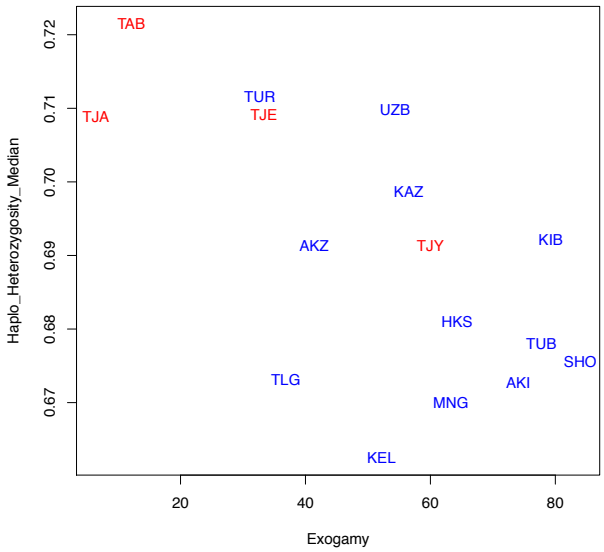


Figure III.17 – Hétérozygotie haplotypique des populations en fonction de leur pourcentage d'exogamie. En bleu les populations turco-mongoles, en rouge les Indo-Iraniennes. Les pourcentages d'exogamie correspondent aux proportions de couples exogames (>4 km) par population.

À propos de la consanguinité proche individuelle

À l'échelle des individus, nous avons observé des niveaux de consanguinité similaires entre les populations turco-mongoles et indo-iraniennes, à partir des longs ROHs ou mesurés par le coefficient de consanguinité de FSuite (Figure 3b et Figure S5 de l'article). Ainsi, la plus forte exogamie des populations turco-mongoles n'est pas corrélée à une consanguinité plus faible, ce qui est un résultat assez inattendu.
Nous avons décelé dans chaque population la présence d'individus issus d'unions entre cousins germains, entre issus de cousins germains et d'unions moins apparentées dites *outbred* (Table S1 de l'article). La répartition de ces trois types d'unions est variable selon les populations étudiées, sans trouver de différence entre les groupes linguistiques. Ainsi, en dépit de niveaux d'exogamie plus élevés, les populations turco-mongoles incluent autant d'individus consanguins que les populations indo-iraniennes. Encore une fois l'exogamie ne corrèle pas avec le niveau de consanguinité des populations.

Pourtant, sur la base du coefficient de consanguinité simple-point de Plink, nous avons trouvé dans toutes les populations que la plupart des individus était moins homozygote qu'attendu dans les populations (Figure 3c de l'article). Cela pourrait suggérer que toutes les populations évitent la consanguinité.

Consanguinité individuelle et exogamie des parents

En nous intéressant de plus près aux types d'union dont sont issus les individus étudiés, nous avons trouvé que de nombreux individus non-consanguins étaient issus de couples endogames, illustrant des mécanismes d'évitement de la consanguinité sans dispersion (Figure III.18). Au contraire, nous avons trouvé que certains individus consanguins descendaient de couples exogames.

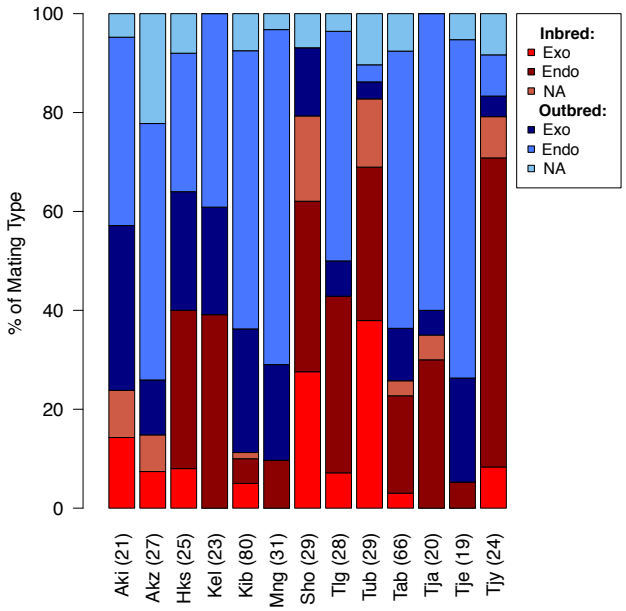


Figure III.18 – Inférence à partir de données génétiques d'unions consanguines (1C, 2C ou 2x1C) ou non-consanguines en fonction de l'exogamie des parents. En rouge les unions génétiquement consanguines, en bleu les non-consanguines. L'exogamie est définie pour des distances >4 km, l'endogamie ≤ 4 km et NA signifie que nous n'avons pas d'information sur l'exogamie du couple parental. Les unions non-consanguines et endogames sont particulièrement intéressantes (rouge vif), ainsi que leur pendant, les unions consanguines exogames (bleu roi). Pour illustration, dans la population Tjy, de haut en bas, on compte 8,3% d'individus consanguins issus d'exogames, 62,5% de consanguins issus d'endogames, 8,3% de consanguins pour lesquels on ne connaît pas le statut des parents ; 4% de non-consanguins exogames, 8,5% de non-consanguins endogames et 8,5% de non-consanguins pour lesquels on ne connaît pas le statut des parents.

Ce résultat nous a interpelées et nous avons décidé d'explorer la relation entre la distance entre les lieux de naissance des parents et la consanguinité de leur enfant. Pour cela, nous n'avons étudié que des individus turco-mongols (dont le nombre d'exogames nous paraissait suffisamment). Nous avons trouvé que des conjoints nés à plus de 4 km mais moins de 40 km avaient statistiquement des enfants plus consanguins que les couples endogames (sur la base du coefficient de FSuite ou du nombre et de la taille totale des longs ROHs ; Figure 4 et S7 de l'article). Au-delà de 40 km, les descendants des couples exogames ont des niveaux de consanguinité comparables, voire inférieurs, à ceux des enfants d'endogames. Ainsi, la distance géographique matrimoniale, à moins de 40 km, affecte la consanguinité des individus dans le sens contraire de celui attendu sous l'hypothèse d'évitement de la consanguinité par la dispersion. Cela suggère que des individus épousent des apparentés nés de villages différents, soit une exogamie géographique couplée à de l'endogamie généalogique.

Conclusions

Ce travail a apporté une confirmation quantitative aux migrations matrimoniales traditionnellement décrites en Asie intérieure, à savoir que les Turco-Mongols pratiquent des alliances exogames hors des villages et les Indo-Iraniens des unions endogames au sein des villages.
Du point de vue de la génétique, nous avons calculé, de manière inédite car à partir de données *genome-wide*, un large éventail d'estimateurs de la diversité intra-populationnelle, ce qui nous donne accès certains pans de l'histoire évolutive de cette région et montre des différences entre les populations du groupe turco-mongol et indo-iranien.
En outre, nos travaux sur la consanguinité nous ont permis de décrire la diversité génétique de cette région sous un angle nouveau : nous nous sommes intéressées à la diversité génétique des individus au moyen de différents estimateurs de la consanguinité proche, notamment les longs ROHs.
En suivant l'axe directeur de cette thèse, ce chapitre a porté sur l'étude du lien entre la perte de diversité causée par des unions consanguines et les migrations matrimoniales qui résultent dans une certaines mesure de choix culturels. De façon surprenante, nous n'avons pas trouvé de corrélation entre ces deux paramètres, à l'échelle des individus et des populations. Ainsi, les résultats obtenus illustrent le fait que certains choix culturels réalisés par l'Homme vont à l'encontre des attendus de la biologie évolutive. Dans notre cas, une exogamie géographique peut aller de pair avec une endogamie généalogique, remettant en cause l'hypothèse d'évitement de la consanguinité par la dispersion.

Ces derniers résultats soulèvent de nombreuses interrogations portant sur le lien biologique entre apparentés exogames : s'agit-il de mariages entre des cousins par la branche maternelle ? Pour tester cette hypothèse, nous aimerions compléter notre étude par l'analyse d'informations génétiques du chromosome X ou mitochondriales, dont nous disposons sans les avoir intégrées à ce projet. Une autre question cruciale porte sur le lien entre consanguinité de dérive dans la population et exogamie : un individu très apparenté à sa population a-t-il tendance à se marier avec un partenaire extérieur à cette population ? Et de manière plus large, quelles sont les motivations sous-jacentes à l'exogamie ? Nous nous proposons de poursuivre ces travaux par des analyses portant sur l'apparentement génétique en lien avec l'exogamie géographique dont nous présentons quelques résultats préliminaires à la fin de ce chapitre.

Article soumis dans Scientific Reports

Close inbreeding and low genetic diversity despite geographical exogamy in Inner Asian human populations

Short Title : Exogamy and inbreeding in Inner Asian populations

Authors : Nina Marchi[a], Philippe Mennecier[a], Myriam Georges[a,b], Sophie Lafosse[a], Tatyana Hegay[c], Tchodouraa Dorjou[d], Boris Chichlo[a], Laure Ségurel[a,*], Evelyne Heyer[a,*]
**Co-supervised this work*
Author Affiliations : [a] Eco-anthropologie et Ethnobiologie, UMR 7206 CNRS, MNHN, Univ Paris Diderot,Sorbonne Paris Cité, Sorbonne Universités, F-75016, Paris, France
[b] 2LM2E-UMR6197, Laboratoire de Microbiologie des Environnements Extrêmes, Institut Universitaire Européen de la Mer, Technopôle Brest-Iroise, Plouzane 29280, France
[c] Republican Scientific Center of Immunology, Ministry of Public Health, Tashkent, Uzbekistan 100060
[d] Department of biology and ecology, Tuvan State University, Kyzyl, Russia, 667000

Corresponding Author : Nina Marchi. Musée de l'Homme, 17 place du Trocadéro, 75016 Paris. 0033.6.23.44.28.26. nina.marchi@mnhn.fr.

Classification : Biological Sciences (Major) ; Genetics (Minor)
Keywords : Dispersal, Mating choice, Consanguinity

Author Contributions : E.H., N.M., and L.S. interpreted the results and wrote the manuscript. P.M., T.H., T.D., B.C., L.S., and E.H. participated to the sampling. M.G. and S.L. handled the DNA samples. N.M. analyzed the data. E.H. and L.S designed the study.

ABSTRACT *(250 words max)*

When related individuals mate, their offspring present an increased level of homozygosity, which is often associated with a lower fitness. Geographical exogamy, by favouring the mating between distant (supposedly less related) individuals, is thought to be a mechanism of inbreeding avoidance; however, no data has clearly tested this prediction. Thanks to the diversity of matrimonial systems in humans, we investigated this question in Inner Asia and explored the impact of geographical exogamy on genetic diversity and inbreeding patterns. We sampled a large number of individuals in 16 populations from two cultural groups (previously described as mostly endogamous *versus* mostly exogamous), for which we collected ethnographic questionnaires and genotyped genome-wide SNPs. We estimated the rate of exogamy in each population and confirmed dispersal differences between groups, *i.e.,* Turko-Mongols are geographically more exogamous than Indo-Iranians. Although populations included in average 36% of inbred individuals, they were all found to be less inbred than expected under random mating. Despite their contrasted exogamy rates, we did not find major differences of inbreeding patterns between Turko-Mongols and Indo-Iranians. Furthermore, across populations within each group, exogamy correlated neither with the proportion of inbred individuals, nor with their genetic diversity. Among Turko-Mongols, descendants from exogamous couples were more inbred than descendants from endogamous couples, except for large distances (>40 km). These results illustrate that, in Inner Asia, geographical exogamy is not always synonym of outbreeding and is therefore not efficient for increasing genetic diversity, nor for avoiding inbreeding (at least at distances below 40 km).

SIGNIFICANCE STATEMENT (*120 words max.*)

Animals should avoid mating between relatives, as inbreeding causes potential health defects. One mechanism for inbreeding avoidance is geographic exogamy, which favours mating between distant, supposedly less related, individuals. To investigate how efficient this mechanism is in humans, we focused on Inner Asia, where populations present different rate of exogamy. Combining ethnological and genomic data, we showed that exogamy is not associated with inbreeding at the population scale, challenging the common knowledge that exogamy is generally synonym of outbreeding. At the individual scale, exogamy was found to decrease inbreeding at distances above 40 km but to increase it for smaller distances. To our knowledge, this is the first time that such data are combined, whether in humans or other species.

Introduction

Around 10% of humans descent from parents that are more related than second cousins, based on registers, as well as civil and medical surveys (1). Such consanguineous mating events can happen for different reasons: it can be by choice, as emblematically described for European royal families (2) (mating choice inbreeding), or by chance, for example if there are too many related individuals in the population, as in isolated groups (drift inbreeding). At the individual scale, the offspring of such close relatives are prone to carry a genetic burden (3): their genomes indeed contain multiple chromosomal segments that are identical by descent (4), leading to a high proportion of homozygous sites (5). Therefore, inbreeding increases the risk for deleterious recessive mutations to be phenotypically expressed (6), and reduces fitness at

loci under heterozygote advantage (7). Thus, in the long term, inbreeding is thought to be responsible for reduced fertility (8, 9) and viability (10, 11), and is in general associated with a wide range of genetic disorders (12). At the population level, inbreeding decreases effective population size, and thus population genetic variability (13). A lack of genetic diversity is often associated with a decrease in the population's adaptive response (14) and could lead to its extinction (15, 16). Moreover, in small populations undergoing strong genetic drift, inbreeding can increase the frequency (or even fix) mildly deleterious mutations (17, 18). However, inbreeding can also enable the purge of severely deleterious recessive mutations (19, 20) and, eventually, improve the population fitness (21).

How do species cope with such inbreeding issues? Several pre- and post-copulatory behaviours are thought to result in reducing inbreeding in the animal kingdom (9, 22). Among pre-copulatory behaviours, most involve kin recognition, which evolved both for avoidance and for cooperation (23, 24). These can rely on the learning of familiarity called Westermarck's effect in humans (25, 26), on physical and acoustic cues (27–30) (e.g. human facial kin recognition (31)), or on the expression of some genes, such as the human MHC (32) or mice MUP (33). Additionally, in some species, there is a delayed sexual maturation in the presence of relatives from the opposite sex (9, 34). Finally, a more straightforward mechanism, even though also costly (35), is to avoid relatives by geographically dispersing (9, 36, 37). Yet, it is unclear if this dispersal aims originally at reducing intra-sexual competition or at avoiding inbreeding (38). In western gorillas, females appear to specifically avoid related males when dispersing (39), giving support to the latter hypothesis. In humans, geographical exogamy, *i.e.,* choosing spouses among geographically distant individuals, has been interpreted as a mechanism to avoid marriages between related individuals (40). Though isolation by distance patterns have indeed been described in numerous species including humans (41, 42), it has not clearly been tested whether observed exogamous partners are really less related than those born in the same area. We are thus lacking data, whether in humans or in other species, to better understand the general impact of geographical dispersal on inbreeding patterns and population genetic diversity. This is likely because dispersal data are hard to obtain for most species and combining such data with genetic information is even more challenging (43).

Interestingly, in Inner Asia, there are different matrimonial systems: Indo-Iranian-speaking populations practice mainly geographically endogamous marriages, while Turkic- and Mongolic-speaking populations (referred to later as Turko-Mongol populations) practice mostly geographically exogamous marriages (44). This allowed us to explore the impact of geographical exogamy on the genetic diversity, using both ethnological, geographical and genetic data. We also aimed to untangle drift inbreeding, due to small population sizes, from mating choice inbreeding, due to matrimonial preferences. On one hand, we collected for 644 couples, ethno-demographic questionnaires including spatial information and measured the exogamy rate for 16 Inner Asian populations (four Indo-Iranian and 12 Turko-Mongol populations from 10 distinct ethnic groups). On the other hand, we genotyped 503 individuals from these 16 populations, for genome-wide autosomal markers. To our knowledge, this is the first time that such quantitative data are combined, in humans or in any other species.

Material and Methods

Population samples: we obtained geographical information for 644 couples from Inner Asian 16 populations (Figure 1, Table S1) and DNA from 503 of the participants. These 503 unrelated individuals were genotyped for 253,532 SNPs, including 105,858 independent SNPs (SI Materials and Methods).

Geographical exogamy was calculated from geographical distance between the places of birth of spouses (SI Materials and Methods). We defined exogamous couples as those with spouses born at more than 4 km based on Figure 2, and we tested for other definitions of exogamy with arbitrary thresholds at 10, 20, 30, 40, and 50 km.

Genetic diversity was estimated from populations pairwise F_{ST}, and allele-sharing dissimilarity (ASD) distances between individuals. We also calculated the haplotypic heterozygosity of each population based on (45) (SI Materials and Methods).

Inbreeding coefficients: we estimated for each individual the inbreeding coefficient (F-Median) (46) with FSuite v1.0.3 (47) and the mating type of their parents couple: avunculars (AV), double first-cousins (2x1C), first-cousins (1C), second-cousins (2C), or unrelated individuals (OUT; defined as less related than second-cousins). We also calculated the genomic excess of homozygosity (48) relative to an expected baseline of homozygosity for each of the Inner Asian populations and for the 11 populations from the HapMap3 worldwide dataset (SI Materials and Methods). Furthermore, we identified runs of homozygosity, called ROHs (49) and computed the number of ROHs observed within each individual genome and their total length.

Results

Dispersal behaviours in Inner Asia

To quantify the amount of dispersal in human populations from Inner Asia, both in Indo-Iranians and in Turko-Mongols (see sampled populations in Figure 1), we calculated the distance between the birth places of spouses for each of the 644 couples (Figure 2). We found that this distance ranged between 0 km for spouses born in the same village (strictly endogamous couples) to 1,474 km, with a median of 5.6 km. Comparing Turko-Mongol and Indo-Iranian couples, we found, as expected based on ethnographic data (44), that Turko-Mongols choose spouses born at larger distances than Indo-Iranians (median of 17 km and 0 km, respectively; one-tailed Mann-Whitney's U, or MWU, test *p-value*=10^{-15}). Based on the local minimum of distance densities, we set the limit for geographical exogamy at 4 km, and found the percentage of exogamous couples to be significantly higher in Turko-Mongols (60% in average, from 33% to 84% per population) than in Indo-Iranians (in average 28%, from 6% to 60% per population) (Figure S1A, Table S1, one-tailed MWU test *p-value*=0.021). When defining exogamy with other thresholds, we still found significant differences between groups at 10, 20 and 50 km (*p-value*<0.034), and a similar trend for 30 and 40 km (*p-value*=0.051 and 0.057, respectively). Focusing on the exogamous couples at 4 km, we found a tendency for larger distances for Turko-Mongol couples than for Indo-Iranians (60 km and 42 km, respectively, one-tailed MWU test *p-value*=0.023). Therefore, Turko-Mongol couples are geographically more exogamous than Indo-Iranians, and they are composed of spouses born at slightly larger distances.

In addition to the distance within the sampled couples, we also calculated the distance between the places of birth of the parents and the parents-in-law, respectively, of the sampled individuals, *i.e.,* from the previous generation (Figure S2). As found for the current generation, Turko-Mongol parental couples also have higher exogamy rates than Indo-Iranians (in average, 34% and 11%, respectively; one-tailed MWU test *p-value*=0.001) (Figure S1B). Moreover, these parental distances are significantly larger for Turko-Mongols than for Indo-Iranians (one-tailed MWU test *p-value*=10^{-10}), suggesting this trend is stable over few generations. The exogamy rates of the previous generation are indeed significantly correlated to those of the current one (Spearman's *rho*=0.81; *p-value*=0.001), even though we observed lower exogamy rates for the parental generation.

Genetic diversity in Turko-Mongol and Indo-Iranian populations

To explore differences in genetic diversity between Turko-Mongols and Indo-Iranians, we computed ASD distances between pairs of individuals from the same population (Figure S3). We found that Indo-Iranians have higher median ASD distance than Turko-Mongols (average per population of 0.280 and 0.269, respectively; two-tailed MWU test *p-value*=0.029), *i.e.,* Indo-Iranians have a higher genetic diversity. Because DNA arrays are known to be enriched for Europeans SNPs (50), and because Indo-Iranian populations are genetically closer to Europeans than Turko-Mongol populations (51, 52), the lower diversity observed within Turko-Mongol populations could be an artefact. To correct for this ascertainment bias, we used another measure of population diversity, the haplotypic heterozygosity, which is less sensitive to ascertainment bias than site-by-site measures (53, 54). For this measure also, we found significantly higher values for the Indo-Iranian populations than for the Turko-Mongol ones (respectively, on average, 0.71 against 0.69, s.d.=0.012 and 0.016 across 22 autosomal chromosomes; one-tailed MWU test *p-value*=0.029) (Figure 3A, Table S1). Therefore, Turko-Mongol populations have overall a lower genetic diversity in comparison to Indo-Iranians, likely reflecting their smaller effective population sizes. These two cultural groups indeed represent two distinct genetic groups having contrasted demographic histories (51, 52). Here, using a genome-wide autosomal dataset including a large set of populations from Inner Asia, we confirm these genetic relationships (Figure S4A): Indo-Iranians and Turko-Mongols cluster in two distinct areas on the first dimension of a MDS plot based on ASD distances between individuals (except for the Turkmen and Uzbeks clustering with Indo-Iranians, as expected based on their particular histories (52)). Population pairwise F_{ST} distances give the same pattern, with similar genetic proximities between populations (Figure S4B, Table S2). Also, pairs of populations within Turko-Mongols and within Indo-Iranians were genetically significantly more similar than pairs between groups (averaged F_{ST}=0.013, 0.012, *versus* 0.031, respectively; one-tailed MWU test *p-value*=10^{-10} and $6*10^{-4}$).

Inbreeding differences between exogamous and endogamous populations

In order to investigate inbreeding differences between populations, we estimated a site-by-site inbreeding coefficient with FSuite and obtained the most likely parental mating type for each of the 503 genotyped individuals (Figure 3B). Across populations, 39% of individuals have positive inbreeding coefficients (from 10 to 90% of each population), with a similar average among Turko-Mongols and Indo-Iranians (respectively 39% and 38%, two-tailed MWU test

p-value=0.86). Based on likelihood ratio tests performed by FSuite, we infer that 36% are inbred across populations, with a similar average between Indo-Iranian and Turko-Mongol populations (on average, 36% and 35%, respectively; two-tailed MWU test *p-value*=0.86). The majority of these individuals have parents that are second-cousins: they represent between 5% and 76% in each population (Table S1), while individuals with parents being first-cousins are detected in only nine out of the 16 studied populations and represent between 1% and 21% of these nine populations, without any statistical differences between the two cultural groups (two-tailed MWU test *p-value*=0.86 and 0.41, respectively, including all of the 16 populations). None of the parents have avuncular relationships, but one case of double-first-cousins was inferred. Overall, the Turko-Mongol and Indo-Iranian groups have similar distribution of these mating types categories: respectively, 32% and 35% of inbred types, including 25% and 27% of second-cousins, 7% of first-cousins for both, and 0% *versus* less than 1% of double-first-cousins (non-significant chi^2 test *p-value* with Yates's correction= 0.83).

We also used runs of homozygosity (ROHs), which sizes and numbers are informative for the type of inbreeding (55). Indeed, ROHs of intermediate size probably result from matings between individuals sharing distant ancestry, and are therefore mostly due to drift, while long ROHs are likely derived from mating between close relatives, likely due to matrimonial preferences (56). First, we found that Turko-Mongol individuals have statistically more intermediate ROHs (500-1,500 kb) than Indo-Iranians (52.2 *versus* 36.6; two-tailed MWU test *p-value*<10^{-38}; Figure S5A), and these ROHs represent a larger portion of their genomes (44,097 kb *versus* 30,613 kb; two-tailed MWU test *p-value*<10^{-36}; Figure S5B). This is consistent with our previous observation of their overall lower genetic diversity due to drift. Then, for long ROHs (>1,500 kb), we observed that the Turko-Mongol individuals have more ROHs than Indo-Iranian (10.9 *versus* 9.0; two-tailed MWU test *p-value*=10^{-6}) (Figure S5C), suggesting more matings between closely related individuals in Turko-Mongols. This is not consistent with our results from FSuite where we did not detect any significant difference of mating choice inbreeding between groups. This difference could be due to an incorrect definition of ROH classes, as they are based on arbitrary thresholds (*i.e.,* between 500 and 1,500 kb, and over 1,500 kb, respectively), despite their utility for the comparison with data from the literature (56–58). When instead defining boundaries per population as done in (54), we obtained different values: on average, class B (intermediate) ROHs ranged between 885 and 2,647 kb, while class C (long) ROHs were above 2,647 kb (Table S1). Using these new boundaries, we did not detect differences either for the number or for the total length of class C-ROHs between Turko-Mongols and Indo-Iranians (two-tailed MWU test *p-value*=0.14 and 0.24, respectively, Figure S5CD), but we still found more and longer class B-ROHs in Turko-Mongols (28.3 *versus* 20.1; two-tailed MWU test *p-value*<10^{-22}; and 36,340 *versus* 26,389 kb; two-tailed MWU test *p-value*=10^{-17}, Figure S5AB). The fact that the chosen definition for class C-ROHs gives contrasted results suggests that the arbitrary cut-off of 1,500 kb, originally defined in European populations (56), might be inappropriate for the studied populations.

Inbreeding avoidance by geographical dispersal?

To investigate inbreeding avoidance, we compared the observed homozygous proportion of each individual genome to the expectation under random mating (48). In each population,

most individuals are less homozygous than expected; overall, the median differences between observed and expected homozygosity are negative, comprised between -0.032 and -0.004, on average -0.019, without any statistical difference between Indo-Iranians and Turko-Mongols (two-tailed MWU test *p-value*=0.31; Figure 3C). This deficit of homozygote sites suggests that some mechanism results in inbreeding avoidance (whether pre- or post-copulatory, voluntary or unconscious) at the population scale. To test whether this result was restricted to Inner Asia, we analysed 11 worldwide populations from HapMap3, and found that they also have negative median values, except the Gujarati Indians from Houston (GIH) that have a median of $1.6*10^{-6}$; Figure S6). This pattern is therefore widespread in humans and could for example be explained by the avoidance of brother-sister marriages (59).

Geographical dispersal is thought to be one mechanism of inbreeding avoidance. Under that hypothesis, we expected populations that practice exogamy to present less inbreeding than endogamous populations. However, surprisingly, in our dataset, the rate of exogamy per population (whether defined with a limit of 4 to 50 km) was not significantly correlated with the percentage of inbred individuals, nor with the proportion of first-cousins' descendants (Spearman's correlation test *p-value*>0.1, for all populations, only for Turko-Mongols or only for Indo-Iranians). We then tested whether, at the individual scale, endogamous couples have descendants that are more inbred than exogamous couples. This was only done in Turko-Mongols, as there were too little exogamous couples in Indo-Iranians. With exogamy defined at 4 km, we detected no statistical differences for the number and total length of long (class C) ROHs between these categories (two-tailed MWU test *p-values*=0.45 and 0.35 respectively). Based on FSuite estimations, we further found descendants from endogamous and exogamous couples to have similar inbreeding coefficients (on average, F-Median=0.005 and 0.008, respectively; two-tailed Mann-Whitney's test *p-value*=0.18). These descendants are also in similar proportion outbred (69.0% and 62.8%, respectively), inbred with parents being second-cousins (25.7% and 30.2%) or first-cousins (5.3% and 7%), suggesting there are no differences of relatedness between geographically exogamous and endogamous couples (chi^2 test *p-value* with Yates's correction=0.64). While these results hold true for exogamy based on 10, 20 and 30 km (*p-values*>0.1), we observed significantly lower proportions of inbred mating-types, lower F-Median values, as well as lower total length and number of class C-ROHs in descendants from exogamous couples at 40 and 50 km (*p-values*<0.03; Figure 4), suggesting an effect of geographical distance on the relatedness between parents. Consistently, focusing on the Turko-Mongol exogamous parental couples (>4 km), we found that the distance between spouses (in log scale) had an effect on their descendant's inbreeding coefficient (Spearman's *rho*=-0.34, *p-value*=10^{-4}, Figure S7). We further detected a significant negative correlation between parental couple distances and the number and total length of long (class C) ROHs: Spearman's *rho*=-0.32 and -0.36, *p-value*s=$3*10^{-4}$ and $7*10^{-5}$ respectively). Altogether, these results suggest that exogamy is associated with a decrease in inbreeding at geographical distances larger than 40 km. Further, we found that descendants from exogamous couples born between 4 and 20 km, and between 20 and 40 km apart, had higher inbreeding levels (Figure 4, two-tailed MWU test *p-values*<0.05) than descendants from couples born in the same location (≤4 km). However, descendants whom parental distances were above 40 km had similar levels of inbreeding than the descendants of endogamous (two-tailed MWU test *p-values*>0.05 for class C-ROH total length and F-

Median), and were even less inbred when looking at the number of class C-ROHs (two-tailed MWU test *p-value*=0.03). For illustration, inbred descendants represent 31% of the individuals with parental distance below 4 km, 54% for those 4-20 km apart, 56% for those 20-40 km apart, but only 27% for those above 40 km.

Discussion

Matrimonial preferences in Inner Asia

Two main cultural groups coexist in Inner Asia: Turko-Mongols that are described in the literature as mainly exogamous, and Indo-Iranians that are thought to be mainly endogamous [57]. However, it is not always clear from the ethnological data if exogamy refers to clan (ethnic) or village (geographical) exogamy. Here, using an empirical threshold of 4 km for exogamy, we measured the amount of geographic exogamy per population and confirmed matrimonial behaviours described in the literature: Turko-Mongols are geographically mostly exogamous while Indo-Iranians are mostly endogamous, even if we found some exceptions: 3/12 Turko-Mongol populations have less than 50% of exogamy, while 1/4 Indo-Iranian populations have more than 50% of exogamy. The exception in Indo-Iranians is the Tjy population that underwent population displacement in the 1970s (based on fieldwork observations), and that show much lower values in the parental generation (60% versus 9%, respectively). In general, though, exogamous rates measured for the parental and current generations are highly correlated, suggesting a certain constancy over time at the population level, despite a trend toward higher values in the current generation. This possibly reflects changes in matrimonial behaviours through time (e.g., linked to urbanisation), or some recall bias for information about parents and parents-in-law couples.

We used these contrasted dispersal behaviours to investigate the level of inbreeding in these populations and to test whether exogamy is a mechanism of inbreeding avoidance. First, we showed that inbreeding resulting from mating between relatives that were at least second-cousins indeed occurred in Inner Asia: on average, 36% of the individuals are inbred. This estimation exceeded the worldwide mean of 10.4% [1], but was still within the range of estimations made for neighbouring Chinese, South Asian and Near East regions (on average, for Xinjiang Turko-Mongol and Tajik populations: 28.6%; in South Asian area: 35.7% including Afghanistan: 48.7%; Iran: 34%; in Western Asian area: 30% including Syria 33.7%) (http://consang.net/index.php/Global_prevalence). Interestingly, marriages between first-cousins prescribed in some Muslim societies (but rarely exceeding 30% (60)), only represent a small proportion of the Inner Asian marriages, while most of the inbred alliances are between second-cousins. Moreover, the avuncular inbred relationship, rare in South and Central Asia according to (55), was not observed in our dataset. As inbreeding information are lacking for Inner Asia, our data complete this global picture of inbreeding prevalence. Interestingly, among the 16 studied populations, we found a great variability of the inbreeding level (from 10 to 80%) and of mating type proportions. As currently, no factor seems clearly associated with this variability, further ethnological fieldwork would be needed to explore this question.

Inbreeding avoidance in humans?

A first result of our study is that the observed inbreeding levels in all Inner Asian populations deviate from the expectation under random mating: given their allelic frequencies, each

population is composed of less inbred offspring than expected by chance. We found the same result in nearly all populations from the HapMap3 worldwide dataset, suggesting this is a general pattern. Such a deficit of inbreeding could be due to the avoidance of close kin matings (*i.e.* brother-sister) that are proscribed in all societies (59), or to other pre/post-copulatory mechanisms. Second, we aimed to test whether inbreeding patterns were influenced by geographical exogamy. Surprisingly, though, our results show that exogamous populations have similar patterns of mating choice inbreeding than endogamous populations, both in terms of the proportion of inbred individuals and the distribution of long (class C) ROHs. This suggests that exogamy is not efficient to avoid inbreeding at the population scale. At the individual scale, focusing on Turko-Mongols, we found that inbreeding levels are similar between descendants from exogamous and endogamous couples for thresholds of 4, 10, 20 and 30 km. These results illustrate that most of endogamous couples are indeed able to avoid inbreeding, while more surprisingly that geographical exogamy is not always synonym of outbreeding. Indeed, in Turko-Mongols, 37% of descendants from exogamous couples (>4 km) are inbred. However, at larger distances (40 and 50 km), descendants from exogamous couples become significantly less inbred than endogamous ones, and there is a significant effect of geographical distance (>4 km) on inbreeding patterns. This suggests that inbreeding patterns are somehow influenced by large geographical distances.

Altogether, these observations can be explained by two scenarios. First, spouses could be chosen at quite large distances but still within a genetically quite homogeneous unit, depending on neighbourhood knowledge and dispersive abilities (61, 62). Under this scenario, we do not expect genetic differentiation between close-by populations that recurrently exchange spouses over generations (e.g., matrimonial relinking (63)). A second scenario is that geographical exogamous marriages preferentially involve spouses from the same family (kinship endogamy). In the context of patrilineal societies like Turko-Mongol ones, where individuals cannot mate with related spouses from the paternal line, this means that spouses would have to be chosen amongst relatives from the maternal side. This hypothesis is supported by the excess of inbreeding found for descendants from exogamous couples born 4-40 km apart as compared to descendants from couples born in the same location (Figure 4). However, to properly examine both hypotheses, we would need to sample, for each exogamous couple, the two populations of origin of the spouses, which is technically unrealistic. These results also challenge the intuition that exogamy necessarily increases the genetic diversity within population, and therefore reduce drift inbreeding. Because of the different genetic background and effective population size between the Turko-Mongol and Indo-Iranian groups (64), we could not disentangle the effect of exogamy from that of demographic history by looking at the genetic diversity of both groups together. Focusing on each group separately, we found that population exogamy rates (whatever the definition of exogamy) are not associated with genetic diversity, as measured by the mean haplotypic heterozygosities (Spearman's correlation test: $p\text{-}value>0.29$ for Turko-Mongols and $p\text{-}value=0.75$ for Indo-Iranians). These results corroborate our conclusion that exogamy is not efficient for increasing the population genetic diversity.

Overall, we provide new insights into the relationship between dispersal and inbreeding in humans, demonstrating that geographical exogamy is not always associated with outbreeding. This suggests, contrarily to the common situation in many animals, that humans who practise

exogamy at small geographical scale might be rather focusing on alliances strategies that result in kinship endogamy.

Acknowledgments

We wish to thank W.Carpentier (Plate-forme Post-Génomique de la Pitié-Salpétrière), B.Regnault and L.Lemée (Institut Pasteur – Genopole) for the genotyping processing; L.Berman for developing the code to compute geographical distances; A-L.Leutenegger and S.Gazal for precious help with FSuite; F.Austerlitz, R.Chaix, R.Laurent, G.Ly, and P.Verdu for constructive discussions.

Funders: ANR NUTGENEVOL (07- BLAN-0064); ANR Altérité culturelle (10-ESVS-0010); CNRS Programme international de collaboration scientifique

References

1. Bittles A (2008) Consanguinity and its relevance to clinical genetics. *Clin Genet* 60(2):89–98. doi:10.1034/j.1399-0004.2001.600201.x.
2. Alvarez G, Ceballos FC, Quinteiro C (2009) The Role of Inbreeding in the Extinction of a European Royal Dynasty. *PLoS One* 4(4):e5174. doi:10.1371/journal.pone.0005174.
3. Woods CG, et al. (2006) Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *Am J Hum Genet* 78(5):889–896. doi:10.1086/503875.
4. Li L-H, et al. (2006) Long contiguous stretches of homozygosity in the human genome. *Hum Mutat* 27(11):1115–1121. doi:10.1002/humu.20399.
5. Mitton JB (1993) Theory and Data Pertinent to the Relationship between Heterozygosity and Fitness. *The Natural History of Inbreeding and Outbreeding: Theoretical and Empirical Perspectives* (University of Chicago Press), pp 17–41.
6. Gao Z, Waggoner D, Stephens M, Ober C, Przeworski M (2015) An Estimate of the Average Number of Recessive Lethal Mutations Carried by Humans. *Genetics* 199(4):1243–1254. doi:10.1534/genetics.114.173351.
7. Charlesworth D, Willis JH (2009) The genetics of inbreeding depression. *Nat Rev Genet* 10(11):783–796. doi:10.1038/nrg2664.
8. Robert A, Toupance B, Tremblay M, Heyer E (2009) Impact of inbreeding on fertility in a pre-industrial population. *Eur J Hum Genet* 17(5):673–81. doi:10.1038/ejhg.2008.237.
9. Pusey A, Wolf M (1996) Inbreeding avoidance in animals. *Trends Ecol Evol* 11(5):201–206. doi:10.1016/0169-5347(96)10028-8.
10. Fareed M, Ahmad MK, Azeem Anwar M, Afzal M (2016) Impact of consanguineous marriages and degrees of inbreeding on fertility, child mortality, secondary sex ratio, selection intensity and genetic load: a cross-sectional study from Northern India. *Pediatr Res*. doi:10.1038/pr.2016.177.
11. Pedersen J (2002) The influence of consanguineous marriage on infant and child mortality among palestinians in the West Bank and Gaza, Jordan, Lebanon and Syria. *Community Genet* 5(3):178–181. doi:10.1159/000066333.
12. Alvarez G, Quinteiro C, Ceballos F (2011) Inbreeding and Genetic Disorder. *InTechOpen*, pp 21–41.
13. Charlesworth D (2003) Effects of inbreeding on the genetic diversity of populations. *Philos Trans R Soc London B Biol Sci* 358(1434):1051–1070. doi:10.1098/rstb.2003.1296.

14. Spielman D, Brook BW, Briscoe D a, Frankham R (2004) Does inbreeding and loss of genetic diversity reduce disease resistance? *Conserv Genet* 5:439–448. doi:10.1023/B:COGE.0000041030.76598.cd.
15. Frankham R (2005) Genetics and extinction. *Biol Conserv* 126(2):131–140. doi:10.1016/j.biocon.2005.05.002.
16. Saccheri I, et al. (1998) Inbreeding and extinction in a butterfly metapopulation. *Nature* 392(6675):491–494. doi:10.1038/33136.
17. Sacks O (1997) *The Island of the Colorblind* (A.A. Knopf).
18. Ubbink GJ, van de Broek J, Hazewinkel H a, Rothuizen J (1998) Cluster analysis of the genetic heterogeneity and disease distributions in purebred dog populations. *Vet Rec* 142(9):209–13. doi:10.1136/vr.142.9.209.
19. Keller L, Waller DM (2002) Inbreeding effects in wild populations. *Trends Ecol Evol* 17(5):230–241. doi:10.1016/S0169-5347(02)02489-8.
20. Xue Y, et al. (2015) Mountain gorilla genomes reveal the impact of long-term population decline and inbreeding. *Science (80-)* 348(6231):242–245. doi:10.1126/science.aaa3952.
21. Moreno E, Pérez-González J, Carranza J, Moya-Laraño J (2015) Better fitness in captive Cuvier's gazelle despite inbreeding increase: Evidence of purging? *PLoS One* 10(12):1–15. doi:10.1371/journal.pone.0145111.
22. Firman RC, Simmons LW (2015) Gametic interactions promote inbreeding avoidance in house mice. *Ecol Lett* 18(9):937–943. doi:10.1111/ele.12471.
23. Holmes WG, Sherman PW (1982) The Ontogeny of Kin Recognition in Two Species of Ground Squirrels. *Am Zool* 22(3):491–517. doi:10.1093/icb/22.3.491.
24. Hepper PG (1986) Kin Recognition: Functions and Mechanisms a Review. *Biol Rev* 61(1):63–93. doi:10.1111/j.1469-185X.1986.tb00427.x.
25. Westermarck E (1921) *The history of human marriage* (Macmillan).
26. Wolf AP (1970) Childhood association and sexual attraction - a further test of Westermarck hypothesis. *Am Anthr* 72(3):503–515. doi:10.1525/aa.1970.72.3.02a00010.
27. Hauber ME, Sherman PW (2001) Self-referent phenotype matching: Theoretical considerations and empirical evidence. *Trends Neurosci* 24(10):609–616. doi:10.1016/S0166-2236(00)01916-0.
28. Pfefferle D, Kazem AJN, Brockhausen RR, Ruiz-Lambides A V., Widdig A (2014) Monkeys Spontaneously Discriminate Their Unfamiliar Paternal Kin under Natural Conditions Using Facial Cues. *Curr Biol* 24(15):1806–1810. doi:10.1016/j.cub.2014.06.058.
29. Levréro F, et al. (2015) Social shaping of voices does not impair phenotype matching of kinship in mandrills. *Nat Commun* 6(May):7609. doi:10.1038/ncomms8609.
30. Crepy MA, Casal JJ (2016) Kin recognition by self-referent phenotype matching in plants. *New Phytol* 209(1):15–16. doi:10.1111/nph.13638.
31. DeBruine LM, Jones BC, Little AC, Perrett DI (2008) Social perception of facial resemblance in humans. *Arch Sex Behav* 37(1):64–77. doi:10.1007/s10508-007-9266-0.
32. Laurent R, Chaix R (2012) MHC-dependent mate choice in humans: Why genomic patterns from the HapMap European American dataset support the hypothesis. *BioEssays* 34(4):267–271. doi:10.1002/bies.201100150.
33. Green JP, et al. (2015) The Genetic Basis of Kin Recognition in a Cooperatively Breeding Mammal. *Curr Biol* 25(20):2631–2641. doi:10.1016/j.cub.2015.08.045.
34. Hoogland JL (1982) Prairie Dogs Avoid Extreme Inbreeding. *Science (80-)* 215(4540):1639–1641. doi:10.1126/science.215.4540.1639.

35. Ronce O (2007) How does it feel to be like a rolling stone? Ten questions about dispersal evolution. *Annu Rev Ecol Evol Syst* 38:231–253. doi:10.1146/annurev.ecolsys.38.091206.095611.
36. Lambin X (1994) Natal Philopatry, Competition for Resources, and Inbreeding Avoidance in Townsend's Voles (Microtus Townsendii). *Ecology* 75(1):224–235. doi:10.2307/1939396.
37. Pusey AE, Packer C (1987) The Evolution of Sex-Biased Dispersal in Lions. *Behaviour* 101(4):275–310. doi:10.1163/156853987X00026.
38. Pusey A (1990) Mechanisms of Inbreeding Avoidance in Nonhuman Primates. *Pedophilia* (Springer New York, New York, NY), pp 201–220. doi:10.1007/978-1-4613-9682-6_8.
39. Bradley BJ, Doran-Sheehy DM, Vigilant L (2007) Potential for female kin associations in wild western gorillas despite female dispersal. *Proc Biol Sci* 274(1622):2179–85. doi:10.1098/rspb.2007.0407.
40. Fix AG (1999) *Migration and colonization in human microevolution*.
41. Ramachandran S, et al. (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* 102(44):15942–7. doi:10.1073/pnas.0507611102.
42. Relethford J (2004) Global Patterns of Isolation by Distance Based on Genetic and Morphological Data. *Hum Biol* 76(4):499–513. doi:10.1353/hub.2004.0060.
43. Ims R a, Yoccoz NG (1997) Studying Transfer Processes in Metapopulations. *Metapopulation Biology* (Elsevier), pp 247–265. doi:10.1016/B978-012323445-2/50015-8.
44. Krader L (1966) Peoples of Central Asia, 2nde edition. *Indiana Univ Publ Bloom* 26.
45. Verdu P, et al. (2014) Patterns of Admixture and Population Structure in Native Populations of Northwest North America. *PLoS Genet* 10(8):e1004530. doi:10.1371/journal.pgen.1004530.
46. Leutenegger A-L, et al. (2003) Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* 73(3):516–23. doi:10.1086/378207.
47. Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L (2014) FSuite: exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics* 30(13):1940–1941. doi:10.1093/bioinformatics/btu149.
48. Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–75. doi:10.1086/519795.
49. Pemberton TJ, et al. (2012) Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* 91(2):275–292. doi:10.1016/j.ajhg.2012.06.014.
50. Stoneking M (2001) From the evolutionary past. *Nature* 409(February):821–822. doi:10.1038/35057279.
51. Martínez-Cruz B, et al. (2011) In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *Eur J Hum Genet* 19(2):216–23. doi:10.1038/ejhg.2010.153.
52. Marchi N, et al. (2017) Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations. *Am J Phys Anthropol* 162(4):627–640. doi:10.1002/ajpa.23151.
53. Conrad DF, et al. (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nat Genet* 38(11):1251–1260. doi:10.1038/ng1911.
54. Pemberton TJ, Rosenberg NA (2014) Population-Genetic Influences on Genomic Estimates of the Inbreeding Coefficient: A Global Perspective. *Hum Hered* 77(1–

4):37–48. doi:10.1159/000362878.
55. Leutenegger A-L, Sahbatou M, Gazal S, Cann HM, Génin E (2011) Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? *Eur J Hum Genet* 19(5):583–7. doi:10.1038/ejhg.2010.205.
56. McQuillan R, et al. (2008) Runs of Homozygosity in European Populations. *Am J Hum Genet* 83(3):359–372. doi:10.1016/j.ajhg.2008.08.007.
57. Joshi PK, et al. (2015) Directional dominance on stature and cognition in diverse human populations. *Nature* 523(7561):459–62. doi:10.1038/nature14618.
58. Kirin M, et al. (2010) Genomic runs of homozygosity record population history and consanguinity. *PLoS One* 5(11):e13996. doi:10.1371/journal.pone.0013996.
59. Thompson EA, Roberts DF (1980) Kinship structure and heterozygosity on Tristan da Cunha. *Am J Hum Genet* 32(3):445–52.
60. Cuisenier J (1962) Endogamie et exogamie dans le mariage arabe. *L'Homme* 2(2):80–105.
61. Boyce AJ, Küchemann CF, Harrison GA (1967) Neighbourhood knowledge and the distribution of marriage distances. *Ann Hum Genet* 30(4):335–338. doi:10.1111/j.1469-1809.1967.tb00035.x.
62. Wright S (1946) Isolation by distance under diverse systems of mating. *Genetics* 31(1):39–59.
63. Segalen M (1986) *Historical anthropology of the family* (Cambridge University Press).
64. Aimé C, et al. (2014) Microsatellite data show recent demographic expansions in sedentary but not in nomadic human populations in Africa and Eurasia. *Eur J Hum Genet* 22(10):1201–1207. doi:10.1038/ejhg.2014.2.

Figure 1 – Geographical location of the 16 Inner Asian populations sampled for this study. Populations are colored based on their linguistic affiliation, which also correlates with their matrimonial system. Note that the Kel population is composed of Northern Asian Kyrgyz from two different locations, and that Kaz and Uzb, as well as Akz and Tlg, were sampled at the same location. UZB. : Uzbekistan, KYR. : Kyrgyzstan, TUR. : Turkmenistan, TAD. : Tadjikistan, AFG. : Afghanistan, PAK. : Pakistan.

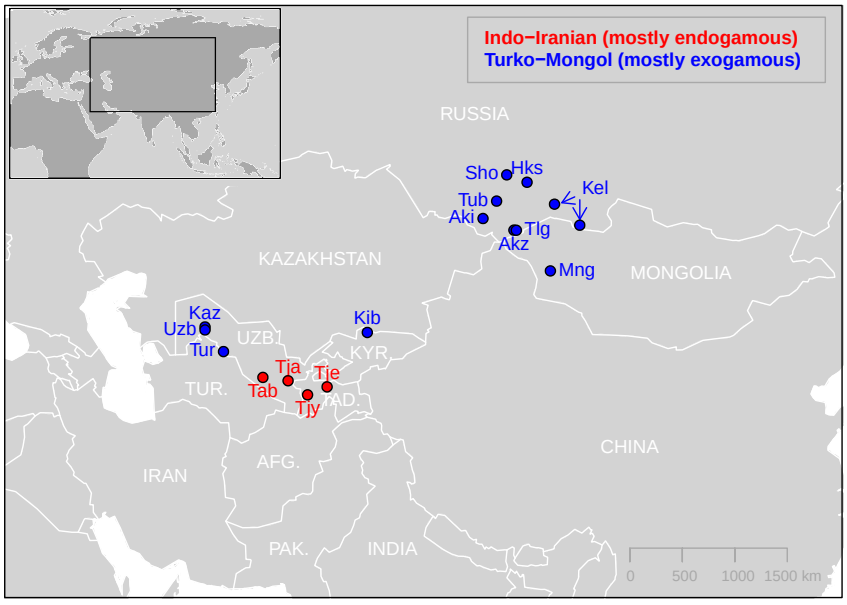


Figure 2 – Geographical distances between the places of birth of couples from Turco-Mongol and Indo-Iranian populations. The distances are plotted in log scale (km).

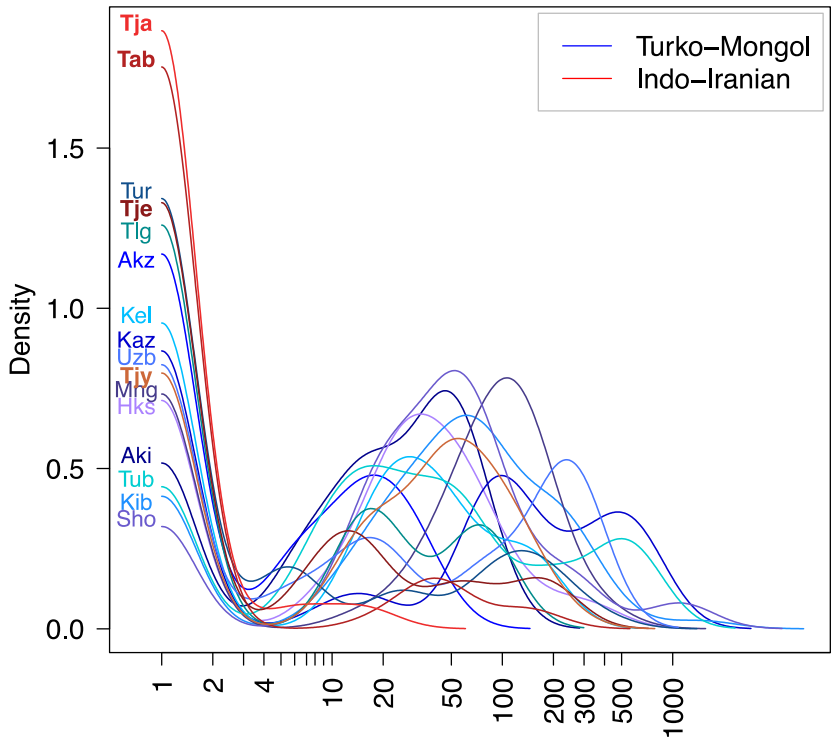


Figure 3 – Genetic diversity and inbreeding pattern within populations. A) Population haplotypic heterozygosity across autosomes. B) F-Median inbreeding coefficient. Grey lines represent inbreeding values corresponding to 2C and 1C. C) Differences between the observed homozygosity within individual genomes and the population baseline, expected under panmixia.

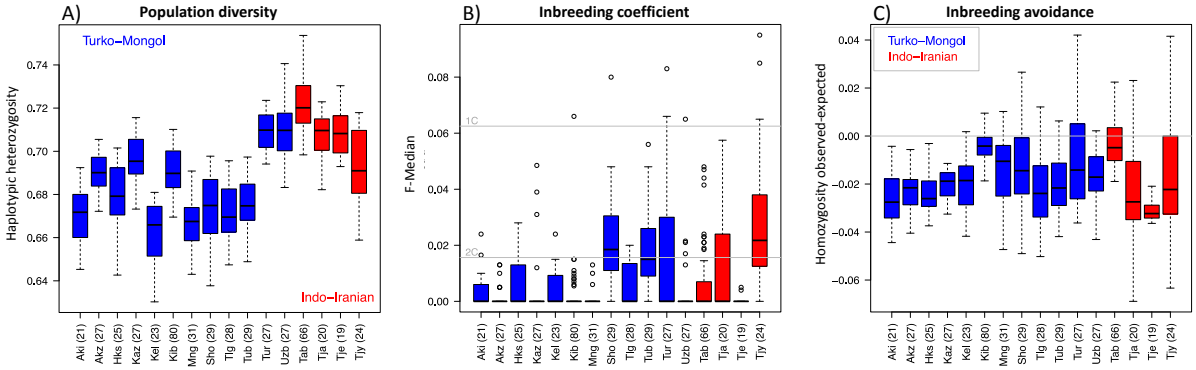
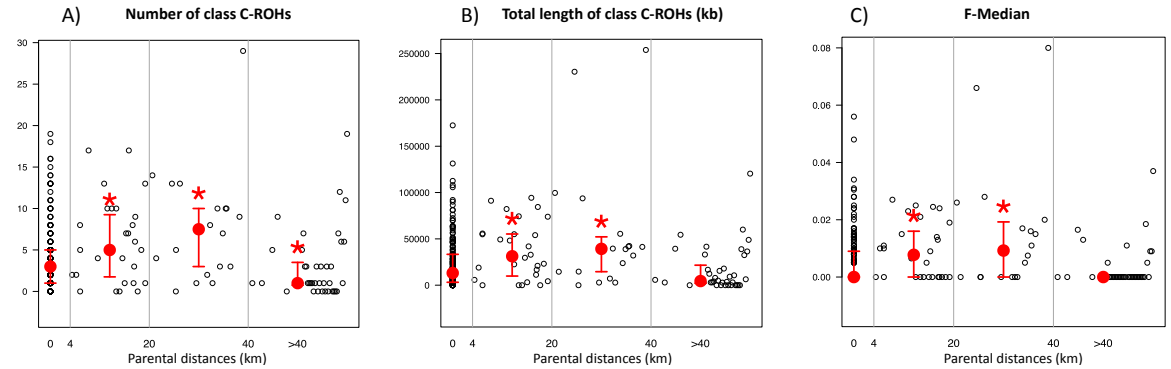


Figure 4 – Inbreeding and parental couple distance for the offspring of Turco-Mongol couples. Three estimators of inbreeding are represented for classes of parental distances (number of long ROHs, their total length and F-Median inbreeding coefficient). Their median, 1st and 3rd quantiles are plotted in red. The distribution of inbreeding estimator for each class is compared with the distribution observed for the class 0-4 km, using a two-tailed MWU test (* for p-value<0.05).



III.3 Résultats préliminaires sur l'apparentement génétique en lien avec l'exogamie géographique

Dans le cadre théorique d'un évitement de la consanguinité par dispersion, on s'attendrait à ce que des individus fortement apparentés au reste de la population soient plus susceptibles d'émigrer ou de conclure des alliances exogames. En complément aux analyses portant sur la consanguinité et pour tester cette hypothèse, nous avons analysé le niveau d'apparentement d'individus endogames et exogames par rapport à leur population d'origine et quand cela était possible le niveau d'apparentement des conjoints par comparaison avec les autres partenaires possibles.
Pour ce faire, nous nous sommes concentrées sur le groupe turco-mongol dont le niveau d'exogamie est assez fort et le nombre de couples exogames correct. Nous avons utilisé des coefficients d'apparentement calculés de manière simple à partir du logiciel Plink (Purcell *et al.* 2007) (Détails méthodologiques donnés en Annexe).

Les individus apparentés à leur population natale sont-ils plus exogames ?

A partir des données ethno-démographiques, nous avons tout d'abord identifié pour chaque population un groupe d'individus philopatriques (leur lieu de naissance et de résidence sont distants de moins de 4 km), constituant un "cœur de population" de référence mimant la population avant l'arrivée des partenaires exogames migrants. Nous avons calculé les coefficients d'apparentement pour toutes les paires d'individus de ce cœur ; la moyenne de ces coefficients constitue l'apparentement moyen de la population étudiée. Cet apparentement moyen est précieux car il nous a permis de normaliser les apparentements calculés au sein des différentes populations turco-mongoles avant de les combiner en une seule analyse. En effet, les niveaux d'apparentement varient fortement entre les populations, à l'image de leur hétérozygotie (Figure III.19).

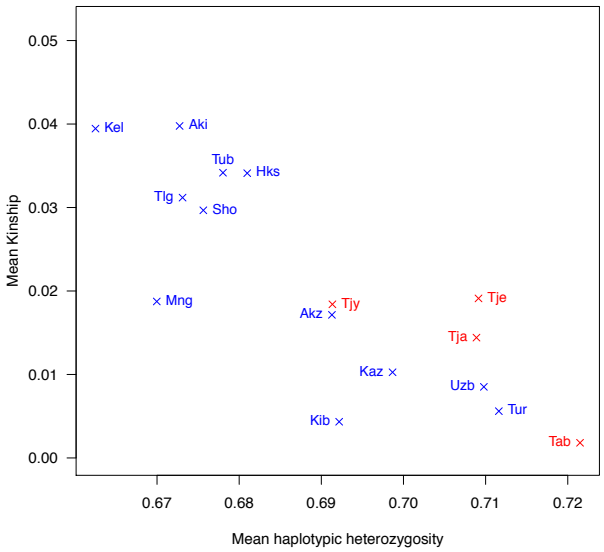


Figure III.19 – Apparentement moyen calculé pour le cœur de population, en fonction de l'hétérozygotie haplotypique moyenne de la population (totale). Nous avons trouvé un ρ de Spearman=-0,81 (p-value=$2*10^{-5}$).

Toujours à partir des données ethno-démographiques, nous avons distingué plusieurs catégories d'individus :

- les endogames philopatriques, dont le lieu de naissance est distant de moins de 4 km de celui de leur conjoint
- les exogames philopatriques, nés à plus de 4 km de distance de leur conjoint mais qui ne se sont pas déplacés depuis leur naissance (moins de 4 km entre leur lieu de naissance et de résidence) ;
- les exogames migrants, nés à plus de 4 km de distance de leur conjoint et qui se sont déplacés depuis leur naissance (plus de 4 km entre leur lieu de naissance et de résidence).

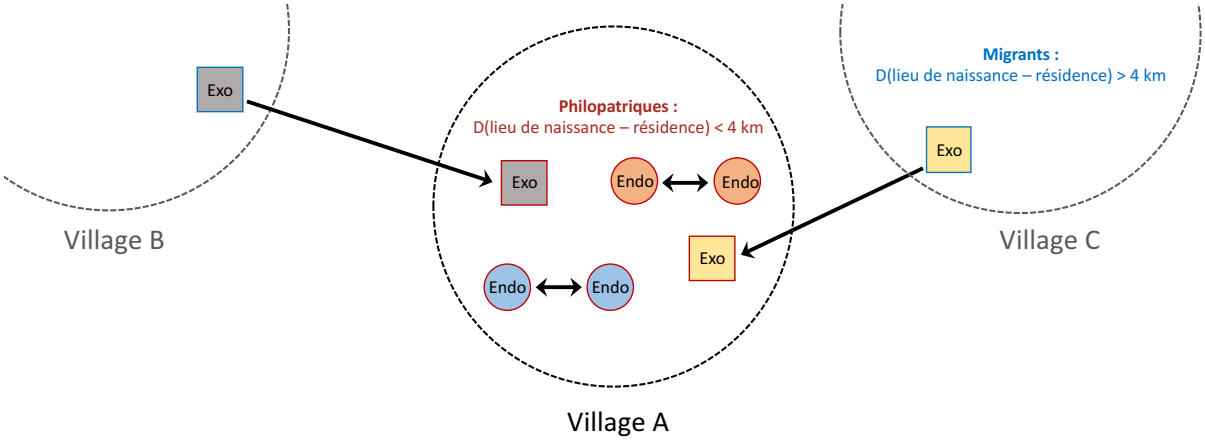


Figure III.20 – Les différents types d'individus considérés dans cette étude : en prenant le village A comme référentiel, les philopatriques y sont nés et y résident ; les partenaires endogames sont tous deux philopatriques alors que pour les couples exogames l'un des conjoints est philopatrique mais l'autre, dit migrant, est originaire d'un autre village, ici B ou C.

Table III.2 – Effectifs

	NTotal	NPhilopatriques =*Cœur*	NEndogames Philopatriques	NExogames Philopatriques	NExogames Migrants
Aki	18	12	5	7	6
Akz	17	13	7	6	4
Hks	21	12	9	3	9
Kaz	20	10	5	5	10
Kel	22	14	12	2	8
Kib	60	12	2	10	48
Mng	22	10	2	8	12
Sho	22	8	2	5	14
Tlg	25	16	12	4	9
Tub	22	13	5	8	9
Tur	20	10	7	3	10
Uzb	22	18	5	10	4
TOTAL	203	148	73	71	55

Pour chaque population, nous avons calculé l'apparentement de chaque individu endogame ou exogame philopatrique avec les individus du cœur de population. Nous avons ensuite corrigé ces valeurs par l'apparentement moyen de la population :

$$K_{corrigé} = \frac{K - KmoyenPop}{1 - KmoyenPop}$$

Puis nous avons regroupé les valeurs obtenues au sein des différentes populations turco-mongoles pour les 73 individus philopatriques endogames et pour les 71 philopatriques exogames (Figure III.21). Nous

avons comparé ces distributions sans voir de différence significative (test de Mann-Whitney bilatéral : p-value=0,94). Nous avons aussi réalisé un test exact de Fisher sur les catégories (Kcorrigé > 0 / Kcorrigé <=0 ; Endogame philopatrique / Exogame philopatrique) pour différents seuils d'exogamie (à 4, 10, 20, 30, 40 et 50 km), sans qu'il soit significatif (p-values>0,22), suggérant une indépendance de ces paramètres.

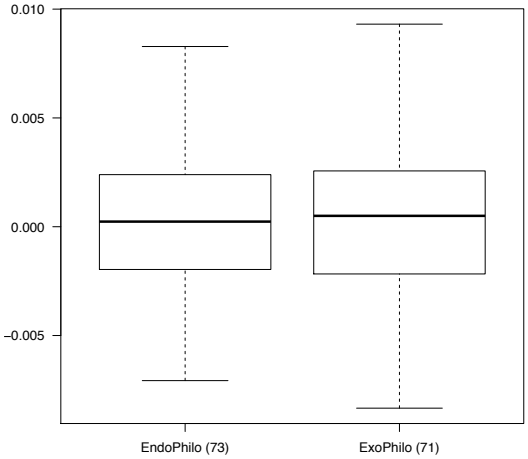


Figure III.21 – Apparentement corrigé calculé pour des individus turco-mongols philopatriques, endogames ou exogames.

Par ailleurs, nous n'avons pas trouvé de corrélation entre le coefficient d'apparentement et les distances entre les lieux de naissance de couples exogames (Figure III.22, corrélation de Spearman : p-value = 0,72).

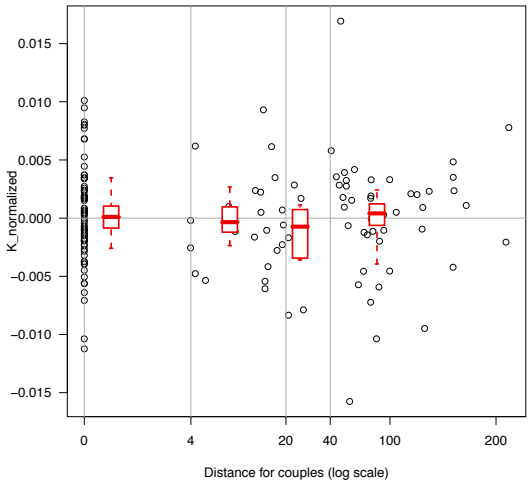


Figure III.22 – Apparentement normalisé calculé pour des individus philopatriques en fonction de la distance entre leur lieu de naissance et celui de leur conjoint, en échelle de log. Les boîtes à moustache représentent les distributions des coefficients d'apparentement corrigé pour des individus philopatriques endogames ($\leq$ 4km), ou exogames (entre 4 et 20 km, 20 et 40, ou >40).

Ainsi, chez quelques populations d'Asie intérieure, l'apparentement n'est pas une motivation nette à la pratique de l'exogamie. Une des limites de cette approche est, encore une fois, de devoir avoir recours à une population de référence. À l'heure actuelle, il n'y a pas de coefficient d'apparentement équivalent au coefficient de consanguinité calculé par inférence par FSuite. Dans notre situation, le nombre d'individus philopatriques est assez faible, et ces résultats devront être confirmés après un échantillonnage plus important.

Les exogames philopatriques sont-ils plus apparentés à leur conjoint, migrant, qu'au reste de leur population ?

Dans quelques rares cas, les deux conjoints d'un même couple turco-mongol ont été génotypés (soit 12 couples endogames pour lequel les deux conjoints étaient philopatriques, et 35 couples exogames dont l'un des conjoints est philopatrique). À partir de cet effectif réduit, nous nous sommes demandées dans quelle mesure le conjoint migrant d'un exogame philopatrique lui était plus ou moins apparenté que les autres individus de sa population natale. Pour ce faire, nous avons calculé l'apparentement entre les conjoints et l'avons comparé à celui calculé entre l'exogame philopatrique et les individus du cœur de population : $\frac{K_{conjoints}}{K_{philopatriqueVScoeur}}$

Dans deux tiers des cas environ, les conjoints sont moins apparentés entre eux que ne l'est le conjoint philopatrique au cœur de population. Ce résultat irait en faveur d'un évitement des apparentés par l'exogamie. Cependant, dans le tiers des cas restants, un individu peut choisir un partenaire qui lui est plus apparenté que les individus de son village natal. Ce résultat fait écho à la consanguinité observée chez des descendants de couples exogames et va à l'encontre de l'attendu de la biologie évolutive.

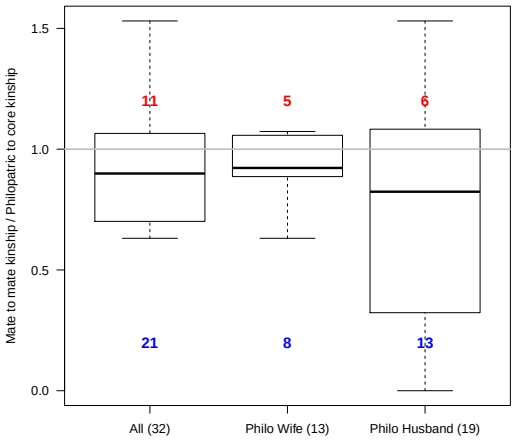


Figure III.23 – Coefficient d'apparentement des couples rapporté à celui du conjoint philopatrique par rapport au cœur de population. Le choix du conjoint pouvant être soumis à des différences entre sexes, nous avons distingué les cas où l'épouse est philopatrique de ceux où l'époux est philopatrique. "All" regroupe ces deux cas. En rouge, le nombre de cas où le ratio était supérieur à un, donc l'apparentement du couple était plus fort que celui du conjoint philopatrique envers les membres du cœur de population. En bleu, le nombre de cas où les conjoints étaient moins apparentés.

En particulier, ces résultats ont été observés que l'individu philopatrique soit un homme ou une femme. Ce choix culturel ne semble pas asymétrique entre les sexes, mais cette observation ne repose que sur un très faible effectif et devra donc être confirmé par des données supplémentaires. De plus, nous avons considéré tous les individus du cœur comme des partenaires possibles, sans tenir compte de leur sexe ou de leur âge, ce qui est une approximation dont on pourrait se départir avec un échantillonnage plus conséquent.

Quant au partenaire exogame migrant, on ne sait pas s'il a été choisi dans une population dont tous les membres étaient hautement apparentés au conjoint philopatrique en conséquence d'échanges récurrents entre les villages dont sont originaires les conjoints ou s'il faisait partie des quelques apparentés présents dans la population et qu'il a été choisi pour sa relation de parenté avec le conjoint philopatrique. Cette

incertitude sera difficile à évacuer car supposerait que l'on ait accès à la population source de l'individu migrant et demande donc un échantillonnage fastidieux.

L'exogamie est-elle transmise d'une génération à la suivante ?

Des observations réalisées sur la base d'informations démographiques auprès de plusieurs populations, notamment à Saguenay-Lac-Saint-Jean (Tremblay *et al.* 2000) ou dans les Alpes suisses (Hagaman *et al.* 1978), ont montré que l'endogamie pouvait être transmise d'une génération à la suivante. Nous avons décidé d'interroger sous cet angle les données ethnologiques dont nous disposons, ce qui représente environ 210 couples, de 8 populations étant donné que nous n'avons pas d'information parentale pour les populations Kaz, Tur et Uzb et que nous avons choisi d'exclure la population urbaine Kib.

Nous avons distingué les cas de figure suivants :

1. les parents et leurs descendants sont endogames (transmission) ;
2. les parents et leurs descendants sont exogames (transmission) ;
3. les parents sont endogames, mais leurs descendants sont exogames ;
4. les parents sont exogames, mais leurs descendants sont endogames ;

De plus, pour chaque couple de descendants, nous disposons de deux informations parentales : l'une pour le couple des parents de l'épouse et l'autre pour ceux de l'époux. Nous avons étudié séparément ces informations car nous envisageons une transmission du comportement matrimonial asymétrique entre les sexes, comme observé par Gagnon *et al.* (2006) au Canada et en Allemagne : les migrations des hommes sont conditionnées par celles de leur père mais celles des femmes ne sont pas corrélées à celles de leur mère.

Nous avons trouvé une situation de transmission dans 56% des cas du côté de l'épouse et 48% du côté du mari. Ce léger écart cause pourtant une différence notable pour les tests exacts de Fisher qui sont significatifs pour les femmes (p-value=0,008) mais pas pour les hommes (p-value=0,37), sous le format (parents endogames / parents exogames ; descendants endogames / descendants exogames). Ainsi, les statuts matrimoniaux des parents sont liés à ceux de leurs filles mais pas à ceux de leurs fils, ce qui va dans le sens inverse des observations de Gagnon *et al.* (2006) réalisées dans des sociétés occidentales.

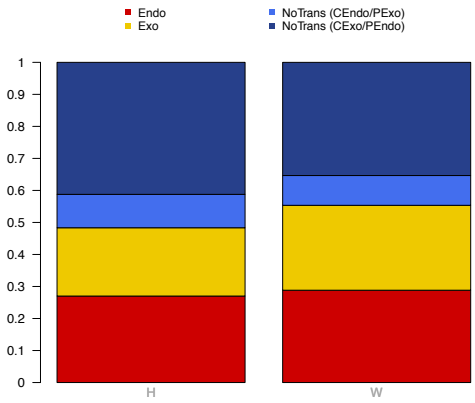


Figure III.24 – Comportement migratoire de deux générations successives, pour un seuil d'exogamie à 4 km. Nous distinguons une transmission entre les parents et leurs filles (W) ou fils (H).

Parmi les cas de non-transmission, la classe où les parents sont endogames et leurs enfants exogames est particulièrement représentée : 41% pour les hommes, 35% pour les femmes. Nous avions déjà observé

une augmentation du nombre d'exogames entre la génération des parents et l'actuelle lors de nos travaux sur la consanguinité. Cependant, nous n'avons pas d'explication à cette différence, à part un hypothétique effet de génération ou un biais des données récoltées pour les parents indirectement auprès des enfants. Dans le cas où l'exogamie est transmise des parents aux enfants, nous avons trouvé une corrélation significative entre les distances mesurées entre les lieux de naissance des parents et celles de leurs descendants, hommes ou femmes : ρ de Spearman 0,4 et 0,48 (p-value$<10^{-3}$). Ainsi, des parents exogames se mariant loin ont en général des enfants se mariant loin eux aussi.

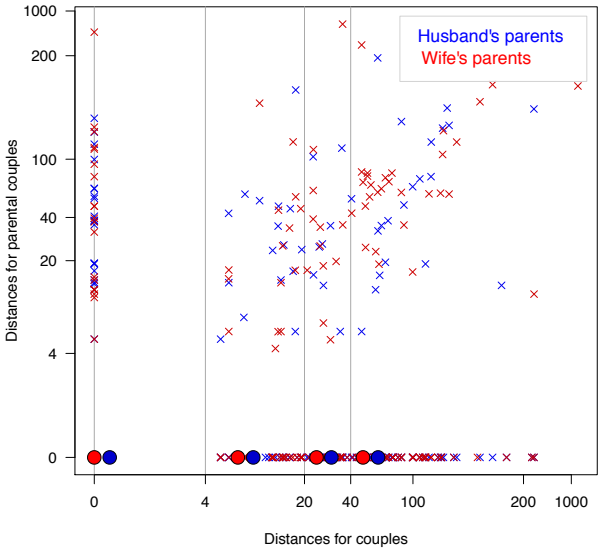


Figure III.25 – Distances entre les lieux de naissance des conjoints des couples échantillonnés et parentaux. En échelle de log.

Nous nous sommes aussi interrogées sur le niveau de consanguinité des individus en fonction de leur statut migratoire matrimonial et de celui de leurs parents : l'endogamie transmise sur plusieurs générations augmente-t-elle la consanguinité des individus ? Pour cela, nous avons sommairement analysé la distribution de trois estimateurs de la consanguinité proche (taille totale et nombre de longs ROHs au sein des génomes ou coefficient FMedian) chez les individus des quatre catégories précédemment décrites (Figure III.26).

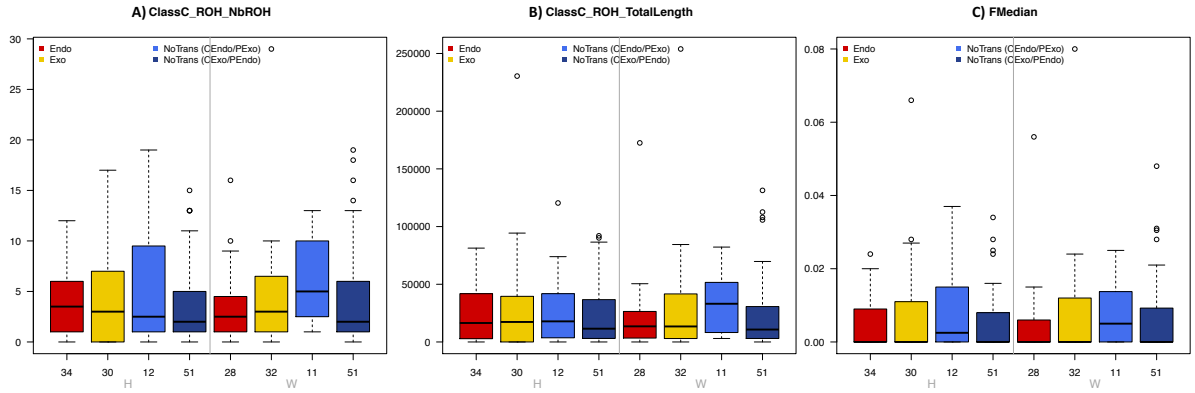


Figure III.26 – Différents estimateurs de la consanguinité des descendants en fonction de leur statut migratoire matrimonial et de celui de leurs parents. En rouge, le cas de transmission de l'endogamie, en jaune de l'exogamie, en bleu les cas de non transmission du statut. Nous avons distingué la consanguinité des hommes et des femmes dans cette analyse.

Les distributions obtenues pour les quatre catégories sont assez homogènes et nous n'avons pas trouvé de différences significatives pour les trois estimateurs entre les catégories sur la base de test de Mann-Whitney bilatéraux (p-values>0,05). Cependant, au sein de chacune des catégories, nous avons observé de la variance entre individus.
De manière inattendue, les individus endogames dont les parents sont exogames (bleu clair) semblent légèrement plus consanguins que les autres, avec une médiane de FMedian différente de 0 et des valeurs médianes de taille et nombre de longs ROHs légérement supérieures à celles des autres catégories, surtout pour les femmes. Ces observations reposent cependant sur de très faibles effectifs et pourraient être biaisées.
D'après ces résultats préliminaires, nous trouvons une légère transmission du comportement migratoire matrimonial entre les parents et leurs filles et des distances géographiques chez les exogames. La transmission du statut n'est pas significative pour les hommes et nous ne trouvons pas d'effet notable sur la consanguinité des individus turco-mongols étudiés.

III.4 Discussion

L'un des objectifs de cette thèse était d'apporter des éléments de réponse à la question très large "Pourquoi les Hommes migrent-ils ?". En particulier, nous nous sommes concentrées uniquement sur des migrations matrimoniales et non sur celles motivées par des raisons économiques ou géo-politiques. La problématique est donc centrée uniquement sur les motivations de l'exogamie.
Dans le Chapitre II, nous avons suspecté des migrations de femmes plus importantes que pour les hommes sur la base d'une plus grande différenciation entre populations pour le chromosome Y que pour l'ADN mitochondrial. À partir des distances calculées entre les lieux de naissance et de résidence des conjoints dans le Chapitre III, nous avons obtenu le chiffre de 348 femmes migrantes contre 263 hommes migrants sur l'ensemble des 16 populations étudiées. Une grande partie de ces effectifs migrants est observée dans des cas de néolocalité : dans 39% des unions exogames, les conjoints sont donc nés dans des villages différents et résident dans une localité qui n'est ni le village d'origine de l'époux ou de l'épouse (Figure 29) ; dans 27% des unions endogames, les époux sont par définition nés dans le même village mais le couple n'y réside plus (Figure 30). Cette néolocalité pourrait être une règle de résidence traditionnelle mais pourrait aussi être le fait de migrations économiques récentes. Outre ces cas de migration des deux conjoints, la plupart des couples exogames (43%) sont patrilocaux, c'est-à-dire que c'est l'épouse qui vient habiter dans le village de naissance de son époux, contre 18% de couples matrilocaux avec une migration masculine en direction du lieu de naissance de l'épouse. En restreignant l'étude aux couples exogames patri ou matrilocaux, ce sont 135 femmes qui ont migré lors de leur mariage contre 55 hommes. Les populations étudiées seraient ainsi, dans l'ensemble et à la génération étudiée, plutôt patrilocales même si deux populations Tub et Tur seraient plutôt matrilocales et d'autres néolocales. La problématique pourrait donc être reformulée sous la forme "Pourquoi les femmes migrent-elles ?".

Au cours de cette thèse, nous avons exploré une première motivation hypothétique, inspirée par la biologie évolutive : celle de l'évitement de la consanguinité motivé par la dépression de consanguinité. Dans le cas des populations d'Asie intérieure, nos travaux montrent que cette explication n'est pas satisfaisante : nous ne trouvons pas de corrélation entre les niveaux d'exogamie et d'apparentement ou de diversité génétique, à l'échelle des individus ou des populations. Des cas contredisant cette hypothèse sont même observés : des individus consanguins peuvent être issus d'unions exogames. De plus, dans ces populations, des couples endogames évitent les unions consanguines, illustrant que la dispersion n'est pas l'unique parade à la consanguinité en Asie intérieure.

Une autre hypothèse serait celle envisagée par Levi-Strauss (Lévi-Strauss 1949) selon laquelle la migration matrimoniale serait une des nombreuses manifestations de comportements culturels et/ou socio-économiques et ne serait pas motivée par la biologie, avec une transmission d'une génération à la suivante comme illustré par Tremblay *et al.* (2000) ; Gagnon *et al.* (2006) ; Heyer (1993) en Occident. Certains de nos résultats préliminaires obtenus en Asie intérieure vont également dans ce sens-là mais un échantillonnage plus conséquent est nécessaire avant d'affirmer que cette transmission est pratiquée en Asie intérieure.
D'un point de vue méthodologique, il est important de prendre connaissance d'une telle transmission des comportements migratoires car cela débouche sur une structuration des populations humaines dont nous devons tenir compte lors des échantillonnages (Hagaman *et al.* 1978 ; Heyer 1995, 1993). En effet, les individus endogames depuis plusieurs générations forment un noyau stable (constituant un cœur) tandis que ceux qui pratiquent l'exogamie seraient inclus dans un réseau d'alliances différent et leur présence serait moins pérenne dans la population (formant une frange). De plus, comme toute structuration,

celle-ci réduirait la taille efficace de la population accentuant ainsi les effets connus de la dérive.

En particulier, l'un de nos résultats majeurs est le fait que des mariages exogames unissent des apparentés proches, parfois des cousins germains, ce qui va à l'encontre de l'hypothèse évolutive et pourrait aller dans le sens de la seconde hypothèse. Il serait donc intéressant de documenter les motivations culturelles d'un tel choix. En particulier, on peut se demander si ces unions respectent la règle d'exogamie clanique décrite sur le plan ethnologique chez les Turco-Mongols. On peut aussi se demander si ces unions correspondent à des alliances préférentielles, et tenter d'identifier le type. Pour ce faire, il faudrait compléter nos travaux par une étude d'anthropologie de la parenté poussée pour comprendre les liens de parenté sociale unissant les conjoints exogames, mais aussi par l'utilisation de marqueurs sexe-spécifiques, comme l'ADN mitochondrial ou le chromosome X, afin d'identifier un apparentement par voie maternelle.

Au cours de cette thèse, nous avons observé des unions consanguines chez toutes les populations étudiées en Asie intérieure. Outre l'hypothèse culturelle, la consanguinité pourrait aussi exister parce qu'elle ne produit pas de désavantages biologiques majeurs. En effet, la dépression de consanguinité est souvent étudiée pour des apparentements proches notamment entre cousins germains. Or dans le cas de l'Asie intérieure, nous avons observé majoritairement des mariages entre cousins issus de germains (dits 2C). Les effets de la consanguinité ressentis à ce niveau d'apparentement ne seraient peut-être pas suffisants pour justifier l'abandon des avantages socio-économiques conférés par les unions consanguines. Au-delà de ça, les unions consanguines de type 2C pourraient être avantageuses par rapport à des unions moins consanguines. Ce phénomène appelé "dépression hybride" a été documenté chez l'Homme en Islande : l'optimum de fitness est atteint pour des couples de cousins au troisième et quatrième degré, donc pour des consanguins assez éloignés et non pour des individus non-consanguins (Helgason *et al.* 2008). Cet optimum pourrait être expliqué biologiquement par certains loci sous avantage homozygote, voire létaux à l'état hétérozygote. Le cas d'école de l'incompatibilité foeto-maternelle pour le système sanguin Rhésus illustre cette idée : le risque de fausse couche pour une mère homozygote récessive (Rh-) augmente si son enfant est hétérozygote (donc Rh+).
De plus, la petite taille efficace des populations turco-mongoles pourrait entraîner une purge assez efficace, autorisant les unions consanguines car peu délétères.

Ces travaux traitant de la consanguinité m'ont amenée à m'interroger sur la notion de taille critique d'une population humaine, et en particulier sur l'avenir de populations isolées d'Asie intérieure pratiquant couramment la consanguinité (comme les Tadjiks de la branche yagnob (Tjy), les Tubalars (Tub) ou les Chors (Sho) - Figure III.18). Cependant, notre espèce elle-même est peu diverse génétiquement, avec 0,1% de différences génétiques entre deux humains (The 1000 Genomes Project Consortium 2015) et une grande homogénéité des populations (Rosenberg *et al.* 2002). Nous partageons tous un certain nombre d'ancêtres communs dont le plus récent aurait vécu il y a seulement 2 000 à 5 000 ans (Rohde 2003). L'existence de ce dernier ancêtre commun, il y a environ 70-170 générations, symbolise le fait que nous sommes tous apparentés les uns aux autres, et nos unions sont d'une certaine manière consanguines. Si l'on estime que notre espèce a plutôt bien réussi à éviter les unions consanguines, avec seulement 10% d'individus dans le monde plus consanguins que des descendants de cousins issus de germains, l'Homme de Néandertal, lui, aurait été menacé par des taux de consanguinité élevé (Prüfer *et al.* 2013).
Cependant, le cas extrême de la communauté d'Indiens Jicaque du Honduras, fondée par seulement huit personnes en 1870, montre qu'une population isolée et de petite taille peut gérer la consanguinité de dérive. Grâce à un évitement de la consanguinité au sein de la population et à quelques migrations, un

effectif de 300 personnes a été atteint un siècle après la fondation (Jacquard 1974).

Enfin, ces travaux sur le lien entre consanguinité et exogamie présentent plusieurs points de faiblesse qui pourraient être améliorés par des échantillonnages ajustés. Les noms de lieux de naissance sont parfois imprécis, à l'échelle du pays si le conjoint est étranger, ou trop précis, à l'échelle du quartier, ce qui rend difficile leur conversion en des coordonnées géographiques et leur exploitation analytique. Dans mes analyses, j'ai pris le parti de mettre en données manquantes les cas que j'avais identifiés comme imprécis et avec l'aide de Philippe Mennecier et sa grande connaissance de l'Asie intérieure, nous avons pu retrouver les coordonnées des lieux trop détaillés. A titre personnel, je n'ai que peu de notions géographiques sur les différentes zones étudiées et ne connais pas, par exemple, la distance moyenne entre deux villages (qui est *a priori* de plus de 4 km). Si nous poursuivons les travaux réalisés au cours de cette thèse par des modélisations, il nous faudra alors fournir un travail de cartographie plus poussé.
Nous avons aussi observé une augmentation du nombre de couples exogames à la génération actuelle, qui a répondu aux questionnaires. On peut légitimement s'interroger sur la "réalité" de cette observation : est-elle due à un changement culturel survenu entre la génération étudiée et celle des parents ? Ou bien est-elle due à un biais de mémoire des enfants, qui pensent que leurs parents sont nés dans le même village alors qu'ils étaient peut-être exogames. Pour éliminer ce doute, on pourrait envisager de collecter les informations directement auprès des parents, s'ils sont encore en vie. Ou bien de s'appuyer sur des registres civils si l'on y a accès.
Enfin, comme souligné dans la section sur l'apparentement, un échantillonnage plus conséquent, notamment de couples est souhaitable afin d'explorer les questions laissées en suspens sur la relation de parenté existant entre les conjoints et sur le choix des conjoints motivé par des facteurs culturels ou la seule stochasticité démographique.

Conclusion générale

Ce travail de thèse a pour objectif d'évaluer l'influence de comportements culturels sur la diversité génétique de populations d'Asie intérieure qui, bien que voisines, diffèrent pour un certain nombre de traits culturels.
Les résultats obtenus ont effectivement montré des effets des comportements culturels sur la diversité génétique à différentes échelles spatiales et temporelles.

Dans le **Chapitre I**, nous nous sommes placées à une échelle géographique large, celle de l'Eurasie. Sur la base de données modernes, nous avons montré que les deux groupes culturels d'Asie intérieure correspondaient à deux entités génétiques distinctes, l'une proche des populations européennes et l'autre des populations est-asiatiques. Pour expliquer l'origine de ces différences et proximités, nous avons eu recours à des données d'ADN ancien. Nous avons alors inféré que la diversité génétique actuelle serait la conséquence d'événements de peuplement propres à chacun des deux groupes culturels.

Dans le **Chapitre II**, nous avons étudié comment des comportements culturels asymétriques entre hommes et femmes ayant cours pendant plusieurs générations modifiaient les diversités sexe-spécifiques. Nous nous sommes concentrées sur des comportements en rapport avec les règles de résidence et de filiation, en utilisant les différences d'organisation sociale et d'origine géographique (Asie du nord *versus* Asie centrale) entre les populations échantillonnées. Nous avons ainsi documenté une structuration particulière en noyaux d'identité du chromosome Y chez la plupart des patrilinéaires d'Asie centrale et du nord par opposition aux cognatiques, ainsi qu'une diversité génétique moins structurée pour l'ADN mitochondrial que pour le chromosome Y, résultat associé à de la patrilocalité. Nous avons aussi retracé l'histoire des groupes ethniques de la région, dont certains étaient présents à la fois en Asie centrale et en Asie du nord, et nous avons souligné l'importance de la culture dans le processus d'ethnogénèse.

Dans le **Chapitre III**, nous avons travaillé à l'échelle réduite des migrations matrimoniales, en nous concentrant sur la diversité génétique des individus. Cette diversité est directement façonnée par les choix matrimoniaux des parents qui se marient ou non avec des apparentés conduisant à de la consanguinité. Ces travaux font partie des rares à replacer la consanguinité dans un contexte de migrations, et dans notre cas, nous disposons d'informations quantitatives sur les distances entre les lieux de naissance des conjoints, ce qui est un matériau inédit dans ce type d'études. Les résultats obtenus ne montrent pas de lien entre les niveaux de consanguinité observés chez les individus d'une population ou la diversité intra-populationnelle et son niveau d'exogamie mais révèlent des différences démographiques entre les groupes linguistiques. À l'échelle des individus, nos analyses mettent en évidence que qu'un certain nombre de couples endogames évitent les unions entre apparentés, quand certains couples exogames se mariaient de façon consanguine surtout dans un rayon de moins de 40 km. Ces résultats suggèrent donc que l'exogamie géographique ne serait pas un moyen efficace pour éviter la consanguinité en Asie intérieure, voire au contraire serait motivée par des unions consanguines, et que l'endogamie géographique serait associée à des moyens d'éviter la consanguinité en intra-population dans une grande majorité des cas.

La problématique de cette thèse l'inscrit nécessairement dans un cadre pluri-disciplinaire. En particulier, l'étude de la consanguinité et des migrations matrimoniales dans le Chapitre III n'a pu être réalisée que parce que nous disposions d'un jeu de données incluant à la fois des données génétiques et ethno-démographiques. Nous avons aussi montré la nécessité de pouvoir contraster des populations pour un comportement culturel asymétrique entre sexes en particulier afin d'en identifier les effets, comme cela a été fait dans le Chapitre II à partir d'informations ethnologiques sur les règles de filiation tirées de la littérature ou par des mesures directes de l'exogamie dans le Chapitre III. De même, les conclusions rendues sur l'ethnogénèse reposent sur une connaissance historique de la région en sus de l'étude génétique.

L'intérêt de cette démarche pluri-disciplinaire est double. D'une part, nous contribuons à améliorer les connaissances de l'histoire de cette région à partir de la génétique *via* l'ADN ancien dans le Chapitre I, et des groupes ethniques dans le Chapitre II à partir de données du chromosome Y. La description de la diversité génétique actuelle dans le Chapitre I est particulièrement intéressante car elle est réalisée à la lumière de connaissances ethno-linguistiques et permet ainsi d'accéder à des pans de l'histoire des différents groupes ethniques ou linguistiques de la région. D'un autre coté, en contrastant les populations pour certains traits culturels, nous pouvons mesurer l'impact de ces traits sur la diversité génétique et donc sur l'histoire évolutive des populations, comme cela a été fait pour les règles de résidence dans le Chapitre II ou pour l'exogamie dans le Chapitre III. Une perspective de cette thèse pourrait être de mettre au point des outils permettant de reconstruire toute l'organisation sociale (règles de résidence, de filiation, d'alliance et leur date d'entrée en vigueur) d'une population à partir de données génétiques. La structuration en noyaux d'identité du chromosome Y pourrait notamment être un outil de diagnostic d'une règle de filiation patrilinéaire.

En somme, ces travaux de thèse ont permis de documenter comment des pratiques culturelles en lien avec l'organisation sociale façonnent la diversité génétique autosomale et sexe-spécifique des populations et des individus, et comment la diversité des groupes culturels a été mise en place lors du peuplement de la région.
Ce projet de thèse est le fruit de multiples collaborations, notamment dans la constitution du jeu de données, dont résulte la variété des données utilisées : ethno-démographiques et génétiques, ADN ancien et moderne, autosomales et uniparentales. La complémentarité de ces données a rendu possible le travail aux différents échelles spatiales et temporelles précédemment citées.
Parmi les résultats obtenus dans ces travaux, certains confirment des observations ethnologiques et des analyses de génétique des populations préalablement réalisées (en ce qui concerne les migrations matrimoniales, la patrilinéarité ou l'ethnogénèse). D'autres résultats sont inédits, comme ceux sur la consanguinité, et ouvrent la voie à de nouvelles problématiques à développer dans cette région. En outre, pour les travaux menés dans le Chapitre III, nous avons combiné différentes analyses de la consanguinité (écart à la panmixie, différents ROHs, estimation du type d'union et du coefficient de consanguinité individuel par FSuite) dont l'utilisation couplée serait intéressante à systématiser puisqu'ils permettent d'avoir une image globale de ce phénomène.

Dans la continuité de ce projet de thèse, un certain nombre de perspectives peuvent être envisagées pour améliorer notre connaissance de l'histoire évolutive de cette région mais aussi à une échelle géographique plus large. Les résultats présentés dans ce manuscrit reposent uniquement sur une approche descriptive et soulèvent des questions auxquelles la seule analyse empirique des données ne permet *a priori* pas de répondre. Le recours à une approche fondée sur des modèles pourrait apporter des éléments de réponse à ces pistes de recherche. Ainsi, par des approches bayésiennes, les scénarios de l'origine des différentes populations actuelles d'Asie intérieure pourraient être inférés. Classiquement, cette modélisation se fait au moyen de données modernes mimant les populations sources ancestrales ; comme nous disposons désormais de populations anciennes dans cette région, il serait intéressant de les utiliser directement comme sources potentielles dans l'inférence. De plus, si les populations anciennes incluses dans l'inférence ne sont pas exactement celles à l'origine des populations, elles pourraient servir à dresser un portrait robot génétique des populations ancestrales sources. D'autre part, nous avons associé les différences génétiques observées à partir des marqueurs uniparentaux à des comportements culturels asymétriques entre sexes, parcimonieusement sur la base de différences culturelles observées chez les différentes populations de la

région. Cependant, il serait judicieux de confirmer cette association en modélisant les effets des comportements culturels asymétriques entre sexes sur des données génétiques. Les effets de la patrilinéarité sur la diversité des populations et des métapopulations seraient particulièrement intéressants à décrire, notamment pour évaluer l'intensité de cette pratique dans les populations. Afin de compléter l'étude de l'impact de la patrilocalité, nous pourrions quantifier le nombre de femmes migrantes *versus* d'hommes migrants dans les populations étudiées. Nous avons partiellement réalisé cette estimation sur la base de données ethnologiques pour la génération actuelle et cela pourrait être modélisé sur plusieurs générations par une approche bayésienne à partir des données génétiques comme cela a été fait en Thaïlande par Hamilton *et al.* (2005). Ce type d'approche permettrait de co-estimer l'âge et l'intensité des comportements culturels asymétriques entre sexes comme cela a été fait pour l'étude de la mono/polygamie dans le logiciel SMARTPOP (Guillot et Cox 2014) et nous informerait sur la date de mise en place de ces comportements, non documentée par les documents historiques.

Enfin, il serait très intéressant de modéliser les effets de l'exogamie sur la diversité génétique des populations en fonction de l'intensité des im- et émigrations, de la taille des populations et en intégrant le fait que certaines unions exogames puissent être consanguines. Cela permettrait de reconstruire l'évolution des unions matrimoniales dans cette région au fil des générations et également d'estimer la part de la consanguinité de dérive agissant au sein de populations présentant différents niveaux d'exogamie. Dans les travaux réalisés au cours de cette thèse, nous avons observé des écarts négatifs à la panmixie que nous avons interprétés comme de l'évitement de la consanguinité. A. Jacquard montrait que le fait qu'une population évite les unions entre frères et sœurs entraine déjà un écart à la panmixie (Jacquard 1971a). Dans les faits, nous aimerions estimer l'impact de l'évitement d'unions consanguines plus plausibles telles qu'entre cousins ou issus de germains sur le niveau d'homozygotie des individus par rapport à l'attendu de la population. Il nous faudrait mettre en place un modèle mathématique centré sur ce type d'évitement et non sur les unions proscrites comme celles entre frères et sœurs, et tenant compte de l'apparentement entre les individus de la population pour estimer si un évitement existe ou si les unions observées sont réalisées au hasard. Cela pourrait demander un échantillonnage ethno-démographique plus poussé pour établir les motivations aux unions et aussi de collecter un plus grand effectif d'individus philopatriques par population pour tester l'écart à la panmixie. Enfin, dans le cas des unions consanguines et exogames, nous tentons à l'heure actuelle d'identifier le lien généalogique entre les partenaires pour mieux documenter ces unions inattendues et intrigantes.

Certains des travaux que nous avons menés auprès de populations d'Asie intérieure pourraient être aisément étendus à d'autres régions du monde, permettant ainsi de généraliser ou non les inférences réalisées au cours de cette thèse. Ainsi, la description relativement complète que nous avons faite de la consanguinité des individus et des populations apporterait de nouveaux éléments de comparaison d'un point de vue génétique sur cette pratique dans différentes régions. Cela permettrait également de tester la structuration de la diversité du chromosome Y en noyaux d'identité chez des patrilinéaires originaires d'autres régions du monde, et pourquoi pas mettre au point des outils génétiques pour décrire l'organisation sociale des populations tant présentes que passées.
Plusieurs pistes de recherche sont donc à envisager aussi bien en Asie intérieure que dans d'autres régions afin d'améliorer la connaissance que nous avons de l'évolution génétique de notre espèce en lien avec des pratiques culturelles mais également de notre évolution culturelle à partir de données génétiques.

Annexes

Sommaire

A Méthodes et statistiques utilisées

Contrôle qualité et mise en forme des données génétiques autosomales

Afin de réaliser les analyses exclusivement sur des individus non apparentés, nous avons recherché des liens d'apparentement au moyen du logiciel Relpair (Michael Boehnke 1997 ; Epstein *et al.* 2000). Pour chaque population, nous avons généré trois sous-ensembles non redondants de 9999 SNPs indépendants (r^2<0,2), bialléliques et respectant l'équilibre de Hardy-Weinberg (les détails sont donnés dans les *Supplementary Methods* du Chapitre III), à partir desquels nous avons établi une liste de paires d'individus strictement plus apparentés que des cousins germains. En utilisant le logiciel Plink 1.9 (Chang *et al.* 2015), à partir du jeu de données complet, nous avons :

- exclu l'individu ayant le moins bon *call-rate* au sein de chaque paire d'apparentés (option *–remove*) ;
- exclu les SNPs n'étant pas sous l'équilibre de Hardy-Weinberg (option *–hardy* ; p-value<10^{-5} sur l'ensemble des populations) ;
- généré un sous-ensemble de 105 858 SNPs indépendants (option *–indep 50 5 2*, soit un $r^2 < 0,5$ basé sur des corrélations multiples).

Analyse des données génétiques autosomales

ADMIXTURE

Les analyses d'ADMIXTURE ont été réalisées à partir d'un sous-ensemble de SNPs indépendants, plus stringent ($r^2 < 0,1$ entre des paires de SNPs) comme recommandé par les développeurs (Alexander *et al.* 2009). Nous avons fixé le nombre de composantes K entre 2 et 17 à observer chez 527 individus d'Asie intérieure, en répétant les analyses 30 fois par K. Puis, pour chaque K, les modes d'ADMIXTURE ont été identifiés à partir du logiciel CLUMPP (Jakobsson et Rosenberg 2007) selon les mêmes paramètres que dans Verdu *et al.* (2014). Ensuite, nous avons calculé la part de chacune des composantes au sein du génome de chaque individu pour chaque mode et les avons représentées à l'aide du logiciel DISTRUCT (Rosenberg 2004). Nous avons choisi de ne présenter dans ce manuscrit que les modes obtenus jusqu'à K=5.

ASD

Pour mesurer la distance génétique entre les 527 individus, nous avons calculé leurs *allele-sharing dissimilarities* (Gao et Martin 2009), à partir du logiciel *asd* (Szpiech 2011), et pour le jeu de SNPs indépendants ($r^2 < 0,5$).

```
./asd-master/src/asd --tped ....tped --tfam ....tfam --out ... --full
```

$\mathbf{F}_{ST}$

Nous avons calculé les distances génétiques entre toutes populations à partir du programme ARLSUMSTAT sous Windows (Excoffier et Lischer 2010), en gardant l'option par défaut de 1000 permutations.

Hétérozygotie haplotypique

Nous avons calculé cet estimateur de la diversité de la population, pour chaque chromosome, à partir d'un code Python3 que j'ai développé en m'inspirant des travaux de Verdu *et al.* (2014). Dans un premier

temps, nous avons reconstruit des blocs non-chevauchants de 5 à 15 SNPs en déséquilibre de liaison : le taux de recombinaison entre deux SNPs voisins est inférieur à 0,5 cM/Mb, en se fondant sur la carte de recombinaison de l'assemblage GRCh37. Nous avons ainsi généré 2 267 blocs, soit en moyenne 103 par chromosome, contenant 13 340 SNPs, soit 5,3% de ceux inclus dans le jeu de données. Puis nous avons phasé ces blocs (option *–hap* de Plink 1.7) et estimé leurs fréquences par population (option *–hap-freq*). Enfin nous avons calculé l'hétérozygotie de chaque bloc à partir de la formule de Nei (1987) et l'avons moyennée pour tous les blocs trouvés dans chacun des 22 autosomes :
$H_{chromosomeK} = \frac{2N_{pop}/(2N_{pop}-1)\sum_{j=1}^{Nblocs}(1-\sum_{i=1}^{Nallèles} f_{i,j}^2)}{Nblocs}$, où j correspond à un bloc et i à une version haplotypique de ce bloc.

Coefficient d'apparentement

Nous présentons le coefficient d'apparentement obtenu à partir de l'option *–genome* de Plink 1.9 (Chang *et al.* 2015) : nous avons calculé la valeur PI_HAT pour l'ensemble des paires d'individus d'une population, qui est la proportion de sites IBD entre deux individus et que nous avons divisée par deux pour obtenir le coefficient d'apparentement.

Coefficient de consanguinité simple-point

A partir de l'option *–het* de Plink 1.9, nous avons calculé la proportion de loci homozygotes dans le génome des individus (O(HOM)) et l'excès (ou déficit) d'homozygotie génomique par rapport à l'attendu de chaque population calculé comme :
F=(Nombre de sites homozygotes observés - attendu) / (Nombre de sites total - attendu).
Le nombre de sites attendu est calculé comme $\sum_{locus=1}^{N}[1-2p_{locus}(1-p_{locus})]$ où p_{locus} est l'une des fréquences alléliques observées au locus étudié.

FSuite

La pipeline FSuite (Gazal *et al.* 2014a) intègre des fonctions préalablement développées dans FEstim (Leutenegger *et al.* 2003) : elle permet notamment de calculer par un maximum de vraisemblance un coefficient de consanguinité individuel à partir de données génomiques, à travers la modélisation de l'autozygotie par des chaînes de Markov cachées. Pour ce faire, 100 sous-ensembles de SNPs indépendants sont générés, avec un marqueur tous les 0,5 cM, et les calculs sont réalisés pour chaque sous-ensemble. Le coefficient de consanguinité final est la valeur médiane des 100 calculées (FMedian). FSuite calcule un autre paramètre noté A, qui découle de l'estimation de la longueur attendue des segments autozygotes dans le génome d'un individu : $A = 1/(L(1-F))$. La valeur médiane de A, couplée à celle de FMedian estimée pour chaque individu, permet de calculer la probabilité que cet individu soit consanguin, et en particulier qu'il descende d'une union entre avunculaires, cousins germains, issus de germains ou doubles-cousins (Leutenegger *et al.* 2011).

Détection des ROHs

Nous avons utilisé l'option *–homozyg* de Plink 1.9 pour détecter des ROHs au sein du génome des individus. Nous avons chercher à détecter tous les ROHs plus longs que 20 kb et contenant au moins 50 SNPs (Joshi *et al.* 2015). Les autres options sont celles par défaut, à savoir qu'il faut observer dans chaque ROH au moins 1 SNP tous les 50 kb, moins de 1000 kb entre deux SNPs consécutifs, et dans chaque fenêtre de 50 kb on tolère au maximum un locus hétérozygote et 5 loci ayant des données manquantes.

```
./plink1.9 --file ... --homozyg --homozyg-kb 20 --homozyg-snp 50
```

A partir de tous les ROHs détectés, nous avons établi des filtres : nous avons gardé les ROHs de taille comprise entre 500 et 1500 kb ou au-dessus de 1500 kb. Nous avons également utilisé la méthode de Pemberton *et al.* (2012) pour identifier des ROHs de trois types : A (courts), B (intermédiaires) et C (longs). Pour ce faire, nous avons cherché à diviser la distribution des longueurs de ROHs observée par population en trois distributions Gaussiennes, au moyen du package R *mclust*. A partir des valeurs extrêmes trouvées pour ces distributions, nous avons calculé le seuil nous permettant de discrimer les ROHs de la classe A, B et C : limite A/B=$\frac{Amax+Bmin}{2}$ et limite B/C=$\frac{Bmax+Cmin}{2}$.

Analyses menées dans le cadre du projet *Steppes*

ACP

L'Analyse en Composantes Principales (Patterson *et al.* 2006) a été réalisée à partir des génotypes modernes du jeu de données *Steppes*, à l'aide du logiciel Plink 1.9 (Chang *et al.* 2015).

D-statistiques

Ces estimateurs permettent d'inférer la contribution génétique d'une population ancestrale (X), par métissage, à une population plus récente (Y) par comparaison avec une seconde population moderne (Z) (Green *et al.* 2010). La statistique est appliquée au groupe (Outgroup ; X ; Y ; Z). Si D>0 alors X et Y sont génétiquement plus proches que X et Z (et inversement). En particulier, les D-statistiques dont la valeur absolue est trois fois supérieure à celle de leur erreur standard sont considérées comme significatives.

qpAdm

Cette statistique, dérivée du ratio F4, permet d'inférer une combinaison linéaire des différentes contributions (ω) de chacune des populations ancestrales supposées être une source (S) de la population métissée génétiquement (T) : $T = \sum_i \omega_i S_i$ (Haak *et al.* 2015).

Analyse des données génétiques uniparentales

ACP

Nous avons réalisé une Analyse en Composantes Principales (Patterson *et al.* 2006), à partir des fréquences des haplogroupes, en utilisant la fonction *dudi.pca* du package *ade4* de R (Dray et Dufour 2007). Nous n'avons retenu que la première composante dans nos analyses, sur la base de l'éboulis des valeurs propres après la première valeur (Cattell 1966).

```
dudi.pca(..., center = TRUE, scale = TRUE, scannf = TRUE)
```

AMOVA

Nous avons évalué la distribution de la variation génétique au sein de populations ou de groupes de populations à partir d'une Analyse de la VAriance MOléculaire (Excoffier *et al.* 1992). Nous avons formé des groupes de populations sur la base de critères géographiques (Asie centrale *versus* Asie du nord), d'appartenance à des groupes ethniques ou linguistiques (Turco-Mongol *versus* Indo-Iranien). En

particulier, nous avons réalisé une AMOVA par groupe ethnique, afin d'étudier la différenciation entre les différentes populations au sein du même groupe ethnique.

BATWING

Ce programme, basé sur le modèle du coalescent, nous permet de retracer l'histoire démographique des groupes ethniques (Wilson *et al.* 2003). En particulier, nous avons daté l'âge de séparation des populations au sein d'un même groupe, et la taille efficace ancestrale de chaque population et du groupe , à partir de données STR du chromosome Y. Le modèle implémenté dans BATWING suppose une séparation instantanée entre les populations au moment de la fission et un arrêt des migrations entre populations, ce qui n'est pas forcément très réaliste dans notre situation. Nous avons utilisé un taux de mutation par STR, estimé dans Willems *et al.* (2016), un temps de génération de 30 ans, et un modèle simple sans croissance démographique. Dans le détail, nous avons choisi une distribution uniforme pour la taille efficace, comprise entre 100 et 100 000. Puis nous avons lancé 110 000 boucles de chaînes de Markov cachées dont les 10 000 premières ont été écartées. A la fin du processus, nous avons calculé la taille efficace et l'âge de la première séparation comme les moyennes des valeurs obtenues pour les 100 000 boucles.

$\mathbf{F}_{ST}$ & $\mathbf{R}_{ST}$

Nous avons calculé des distances F_{ST} (Wright 1949) entre populations à partir des séquences HVS1 ou d'haplotypes STR du chromosome Y, à partir du logiciel Arlequin v3.5 (Excoffier *et al.* 2007). Pour le chromosome Y, nous avons également calculé des distances R_{ST} (Slatkin 1995). Pour l'ADN mitochondrial, nous avons choisi de présenter les résultats du modèle de Kimura à 2 paramètres, avec un ratio transition/transversion=10 et un paramètre α=0,26 (Heyer *et al.* 2009), bien que nous ayions aussi calculé le F_{ST} du modèle par différences entre paires.

Indices de diversité génétique

En m'inspirant des travaux de Chaix *et al.* (2007), j'ai écrit un code en Python3 afin de calculer différents estimateurs de la diversité génétique du chromosome Y, au sein de chacune des populations :

- l'hétérozygotie haplotypique (H), en version haploïde (Nei 1987) ;
- le nombre moyen d'individus partageant le même haplotype de 8 STRs (C). Les haplotypes contenant des données manquantes ne sont pas pris en compte dans le calcul ;
- la proportion d'haplotypes observés une seule fois dans la population, dits singletons (Ps pour proportion de singletons) ;

À partir des haplotypes STR du chromosome Y, nous avons calculé des distances *Average Square* de Goldstein et Pollock (1997), implémentées dans le logiciel Populations, 1.2.30 Copyright (C) 1999, Olivier Langella, CNRS UPR9034. Nous avons ensuite analysé ces distances pour chaque population au moyen d'une MDS proportionnelle afin de rendre compte des fréquences des différents haplotypes dans la population et ainsi visualiser des structures en noyaux d'identité.
Nous avons aussi calculé le nombre moyen de différences entre haplotypes du chromosome Y ou de HVS1, avec le logiciel Arlequin (Excoffier *et al.* 2007) : $\pi = \sum_{i=1}^{NombreHaplotypes} \sum_{j<1} p_i * p_j * d_{i,j}$ où $d_{i,j}$ est le nombre de différences entre les haplotypes i et j, et p la fréquence des haplotypes.

Représentation des distances

MDS

Nous avons représenté les distances génétiques entre individus (ASD) ou entre populations (F_{ST}) au moyen d'une analyse *MultiDimensional Scaling* (MDS ; option *cmdscale* de R). Ce mode de représentation permet d'analyser les données dans un nombre de dimensions réduites (ici 2). Nous avons calculé à quel point les distances "réelles" que l'on avait calculées étaient corrélées à celles obtenues par la MDS, en estimant le coefficient de Spearman entre ces deux matrices.
Nous avons également réalisé des MDS proportionnelles : sur ces représentations, chaque figuré correspond à un haplotype STR du chromosome Y, la distance entre deux points est la projection en deux dimensions de la somme des différences au carré entre deux haplotypes, et la taille du figuré indique la fréquence de l'haplotype dans la population (voir Figures "Noyaux d'identité du chromosome Y" en Annexes du Chapitre II).

Neighbour-Joining Tree

Pour représenter les distances F_{ST}, nous avons aussi opté pour une approche de Neighbour-Joining, implémentée dans la fontion *bionj* du package *ape* de R (Gascuel 1997).

Statistiques

Toutes les statistiques ont été calculées dans le logiciel GNU R. À savoir,

- des corrélations de Spearman :
  `cor.test(..., ..., method= "spearman",exact=FALSE)`
- des tests non-paramétriques de Mann-Whitney, bilatéraux ou unilatéraux, pour tester si deux distributions de valeurs sont similaires :
  `wilcox.test(...,..., alternative=c("two.sided", "less", "greater"))`
- des tests de Mantel entre des matrices de distances génétiques, par exemple F_{ST} *versus* R_{ST}, ou des matrices de distances génétiques *versus* géographiques. Dans ce cas là, nous avons utilisé la forme $\frac{F_{ST}}{1-F_{ST}}$ et une échelle logarithmique pour les distances géographiques. Le test utilisé est celui du package *vegan* (Oksanen *et al.* 2016) :
  `mantel(..., ..., permutations = 10000, method="spearman")`
  De plus, nous avons utilisé le test partiel de Mantel afin de prendre en compte le fait la plus grande proximité géographique des populations d'un même groupe ethnique dans le calcul de la corrélation entre les distances géographiques et génétiques :
  `mantel.partial(MGeo,MGenet, MEthnie, method = "spearman", permutations = 10000)`
- des tests exacts de Fisher (fonction *fisher.test*) pour tester la contingence de deux paramètres, comme l'apparentement et la consanguinité ou le statut migratoire des couples parentaux et de ceux de leurs descendants.

B Informations complémentaires sur les données collectées

Questionnaire ethno-démographique

№ ..... № .....

Inspection site ..... **Date** ...../08/2012

M	**F**
Name.....	**Name**.....
Date Of Birth… Place of Birth	**DOB ……** Place of Birth
Child order in the family	Child order in the family .....
Son order	Girl order .....

Your spouse

Alive: ..... Yes / No First marriage : ..... Yes / No (Который : ....)
Are you related to your spouse ? ..... Yes / No

M	**Location**	**F**
Place of Birth		Place of Birth .....
Domicile ..... , when ?		Domicile ..... , when ? .....
Domicile of father		Domicile of father .....
Paternal grandfather		Paternal grandfather .....
Paternal grandmother		Paternal grandmother .....
Domicile of mother		Domicile of mother .....
Maternal grandfather		Maternal grandfather .....
Maternal grandmother		Maternal grandmother .....

M	**Language**	**F**
Your language		Your language .....
Father's language		Father's language .....
Paternal grandfather		Paternal grandfather .....
Paternal grandmother		Paternal grandmother .....
Mother's language		Mother's language .....
Maternal grandfather		Maternal grandfather .....
Maternal grandmother		Maternal grandmother .....

M	**Tribe**	**F**
Your tribe		Your tribe .....
Father's tribe		Father's tribe .....
Paternal grandfather		Paternal grandfather .....
Paternal grandmother		Paternal grandmother .....
Mother's tribe		Mother's tribe .....
Maternal grandfather		Maternal grandfather .....
Maternal grandmother		Maternal grandmother .....

.../...

Married sons and daughters

Total amount of children (including died)	Amount of daughters (including died)	Amount of sons (including died) .....
Total amount of married sons and daughters	Amount of married daughters	Amount of married sons .....
Is last child married ? Yes / No		Age of last child .....

Eldest child :	Sex: M / F	DOB	Domicile .....
Spouse :	Date of birth	Language	Tribe .....
Are the spouse related ? Yes / No			

Second child :	Sex: M / F	DOB	Domicile .....
Spouse :	Date of birth	Language	Tribe .....
Are the spouse related ? Yes / No			

Third child :	Sex: M / F	DOB	Domicile .....
Spouse :	Date of birth	Language	Tribe .....
Are the spouse related ? Yes / No			

Fourth child :	Sex: M / F	DOB	Domicile .....
Spouse :	Date of birth	Language	Tribe .....
Are the spouse related ? Yes / No			

Fifth child :	Sex: M / F	DOB	Domicile .....
Spouse :	Date of birth	Language	Tribe .....
Are the spouse related ? Yes / No			

Sixth child :	Sex: M / F	DOB	Domicile .....
Spouse :	Date of birth	Language	Tribe .....
Are the spouse related ? Yes / No			

Seventh child :	Sex: M / F	DOB	Domicile .....
Spouse :	Date of birth	Language	Tribe .....
Are the spouse related ? Yes / No			

Eighth child :	Sex: M / F	DOB	Domicile .....
Spouse :	Date of birth	Language	Tribe .....
Are the spouse related ? Yes / No			

Ninth child :	Sex: M / F	DOB	Domicile .....
Spouse :	Date of birth	Language	Tribe .....
Are the spouse related ? Yes / No			

.../...

Autorisations de collecte

République de Tuva - Russie

CONVENTION
de collaboration entre l'Université d'État de la République Touva (Fédération Russe) et le Muséum national d'histoire naturelle (France)

Considérant la nécessité d'intégrer les sciences de l'enseignement supérieur et de la recherche et la conservation du milieu naturel pour le développement des connaissances et des activités des équipes de recherche sous l'égide de l'Institut public fédéral de l'enseignement supérieur, l'Université d'État de la République de Touva (TuvGU) en la personne de son recteur M. Sergeï Oktiaïévtich Ondar et le Muséum national d'histoire naturelle en la personne de son directeur M. Thomas Grenon, appelés ci-après « les Parties », ont établi la Convention suivante :

1. Objet de la convention

Les Parties, afin de réaliser les tâches fixées en commun :

1.1. Mettre en œuvre de recherches communes pour étudier de l'histoire de la formation des peuples d'Asie Centrale, de Sibérie et de Mongolie sur le projet « Analyse de l'évolution complexe des gènes dans l'adaptation de l'homme au régime alimentaire ».
1.2. Dans le cadre de ce projet, étudier les spécificités phéno- et génotypiques des populations locales en fonction de leurs traditions alimentaires et de leur mode de vie.
1.3. Procéder à des enquêtes ethnologies, linguistiques et génétiques, et prélever de la salive (au moins 150 échantillons) dans les différentes populations pour réaliser des recherches ethnogénétiques au Muséum national d'histoire naturelle.
1.4. Enquêter dans chaque population auprès de 25-30 couples âgés de 20 à 45 ans.
1.5. Organiser une expédition commune du 4 au 31 août 2012. La mission comprend 5 personnes, dont Tchodouraa Mikhaïlovna Dorjou, membre de l'Université d'Etat de Touva, Tatiana Hegaï, membre de l'Institut d'immunologie de l'Académie des sciences de la République d'Ouzbékistan et le Professeur Evelyne Heyer, le Dr Philippe Mennecier et Eric Verzele, membre du CNRS de France.
1.6. Favoriser l'accès des chercheurs et des enseignants aux ressources scientifiques des parties.
1.7. Publier des travaux scientifiques communs avec les chercheurs, les enseignants, les aspirants et les étudiants.

2. Volume et modalités du financement

2.1. Chaque partie est responsable des financements concernant ses activités propres. Aucun flux financier n'intervient entre les parties.

3. Droits et obligations des parties

3.1. Les parties ont le droit de dénoncer unilatéralement la convention si l'une des parties ne remplit pas les conditions de la Convention.

4. Responsabilité des parties

4.1. Si les obligations de la présente Convention ne sont pas assurées, en totalité ou en partie, le TuvGU et le MNHN sont mutuellement responsables conformément au Règlement en vigueur.

5. Conditions complémentaires

5.1. Toute modification ou ajout à la Convention fait l'objet d'un protocole supplémentaire

6. Cas de force majeure

6.1. Sont considérées comme cas de force majeure empêchant l'exécution des obligations citées dans la présente Convention les circonstances prises comme telles par la Législation de la Fédération Russe et de la République de Touva. Dans ce cas, l'action de la Convention s'arrête pour la durée de validité de ces circonstances sans information spéciale des parties.

7. Durée de validité de la Convention

7.1. La Convention entre en vigueur à la signature des ses parties et reste valide 5 ans.

8. Coordonnées des parties

ФГБОУ ВПО «Тувинский государственный университет»
667000, г. Кызыл
ул. Ленина, 36
Республика Тыва, Россия
Тел/факс: +7 (39422)2-19-69, 3-84-52
e-mail: tgu@tuvsu.ru;

6-8-2012

Ректор_____

М.П.

Национальный музей естественной истории Франции
57, rue Cuvier
75005 Paris, France

Tel/fax: +33 1 40 79 30 00
http://www.mnhn.fr/

Директор_ Serge Bahuchet

Serge BAHUCHET
Professeur d'ethnobiologie
Département Hommes, Natures, Sociétés
Muséum National d'Histoire Naturelle

МП

3 août 2012

République d'Altai - Russie

РЕСПУБЛИКА АЛТАЙ
МИНИСТЕРСТВО ЗДРАВООХРАНЕНИЯ

649000, г.Горно-Алтайск,
Коммунистический пр-т, д.54
Тел. 8-388-22-2 20 78, факс 2 73 42
E-mail: minzdravra@narod.ru

АЛТАЙ РЕСПУБЛИКАНЫН
СУ-КАДЫК КОРЫЫР МИНИСТЕРСТВОЗЫ

649000, Горно-Алтайск кала,
Коммунистический пр-т, д.54
Тел.. 8-388-22-2 20 78, факс 2 73 42
E-mail: minzdravra@narod.ru

22.08.2011 № 3996
На №

Главному врачу
Муниципального учреждения
здравоохранения
Республики Алтай

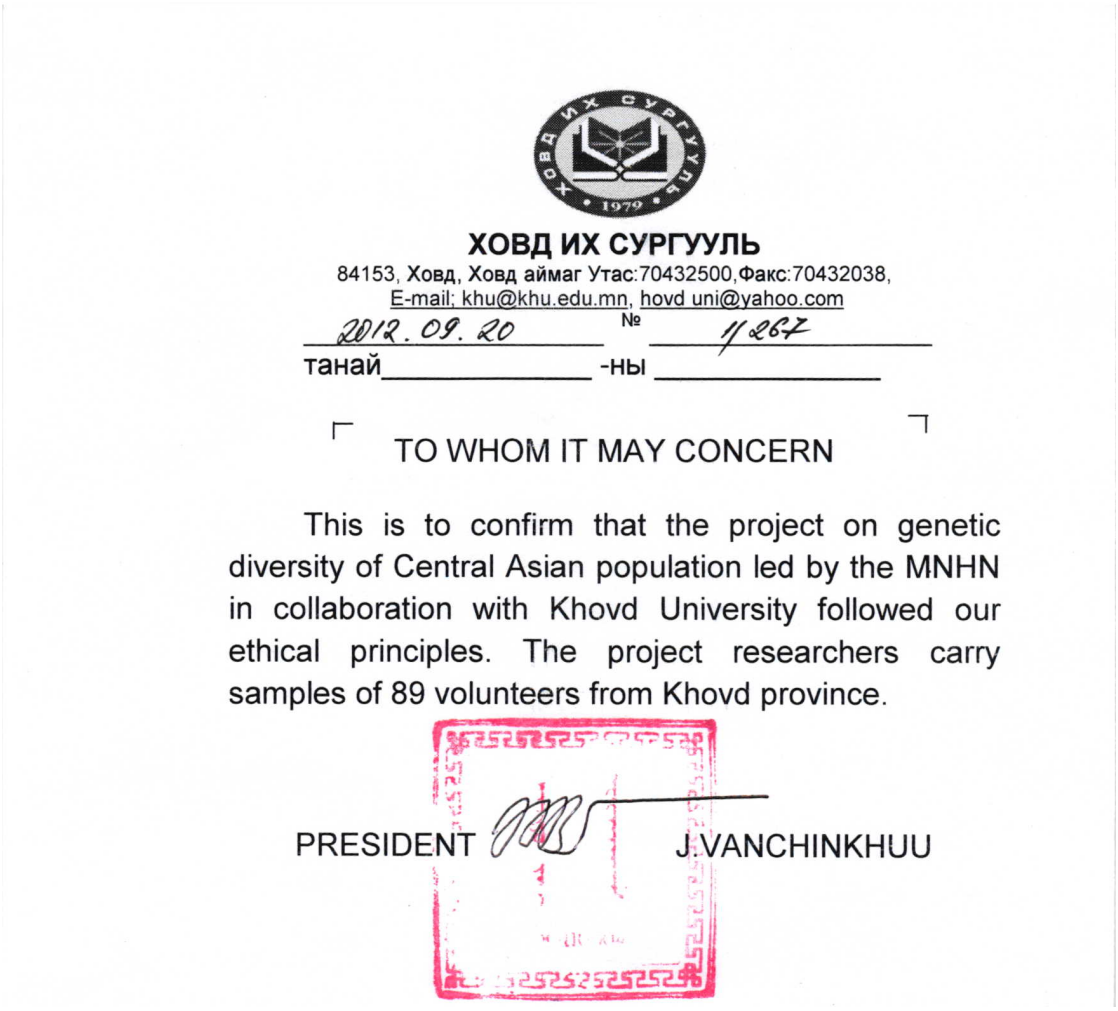
О содействии

Министерство здравоохранения Республики Алтай и Национальный Музей Естественной Истории проводят уникальные совместные франко-российские антропогенетические исследования народов Республики Алтай. Главная цель проекта является изучить древние эволюционные процессов формирования народов Республики Алтай и генетические аспекты алиментарной адаптации, и их влияние на заболевания сердечнососудистой системы в регионе. Результаты этого проекта предоставят новые возможности в улучшении профилактических мер в борьбе с этими заболеваниями в Республики Алтай.

В связи с этим руководство Министерство здравоохранения Республики Алтай убедительно просит Вас оказать содействие в реализации задач проекта.

Министр И.Э.Яимов

Mongolie



ХОВД ИХ СУРГУУЛЬ
84153, Ховд, Ховд аймаг Утас:70432500,Факс:70432038,
E-mail; khu@khu.edu.mn, hovd uni@yahoo.com
2012.09.20 № 1/267
танай_____ -ны _____

TO WHOM IT MAY CONCERN

This is to confirm that the project on genetic diversity of Central Asian population led by the MNHN in collaboration with Khovd University followed our ethical principles. The project researchers carry samples of 89 volunteers from Khovd province.

PRESIDENT J.VANCHINKHUU

Ouzbékistan

Muséum National d'Histoire Naturelle
Département *Homme, natures, sociétés.*
Unité d'Eco-Anthropologie – Equipe « Anthropologie Génétique »
ASSOCIÉ AU CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE ET A L'UNIVERSITE PARIS 7

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
UMR 7206

Paris, 16 february 2010

Declaration of collaboration

Research Project

One of the greatest challenges in human evolutionary genetics is to elucidate the evolution of biological multigenic traits.

The nutritional transitions that occurred during human evolution represent major changes in our environment. One of these major transitions is the emergence of agriculture in human societies, for the first time 10,000 yrs ago, when some populations shifted their predominantly meat diet to a predominantly cereal diet. Before this major transition, genes favouring insulin resistance and gluconeogenesis were selected for, in order to constantly maintain sufficient level of glucose in the blood. These genes may now be detrimental in present societies, because under the present high carbohydrate diet, insulin resistance and gluconeogenesis may lead to metabolic disorders such as diabetes, obesity, and hypertension.

In order to understand how genes involved in alimentary processes and associated with diabetes have evolved in response to different selective pressures, we will comparatively study different societies that have lived for thousands of years under contrasted diets. This case study will be conducted in Central Asia where pastoral societies and agriculturist societies still co-exist. Pastoralists are though to have a higher input of meat and diary products compared to agriculturists whose alimentation is mainly based on cereals.

Methods

We will take phenotypic measurements (nutritional anthropometry, blood measurements of fasting glucose and insulin, high density lipoprotein cholesterol and triglycerides) for different ethnic groups with different food intake. We will perform alimentation surveys and we will sample blood for DNA analyses. This will allow us to study the sequence polymorphism for 10 major genes involved in insulin resistance/sensitivity. We will first apply classical selection tests to these sequences. Then we will develop new multi-locus selection tests that may be more powerful in this context. In parallel, a simulation study will enable us to better understand the evolution of this complex trait under different nutritional conditions and assess the efficiency of the classical and newly developed tests of selection in this context.

Schedule

This project is a four years project from 2008 to 2012. Several field studies are planned in Uzbekistan. In 2010 we will sample in Tashkent.

Publication of results

All published scientific results will be co-authored by French and Uzbek researchers

Courrier/Mail : Pr Evelyne HEYER - UMR 7206 Eco-anthropologie - Equipe "Anthropologie génétique"
CP 139 - 57 rue Cuvier - 75231 Paris Cedex 05 - FRANCE
Téléphone : (33 -1) 40 79 81 58 - Fax : 33 1 40 79 32 31 - e-mail : heyer@mnhn.fr

Collaboration between

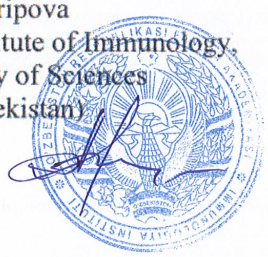
Professor Evelyne Heyer and her team
Human population Genetics
UMR 7206 Eco-anthropologie
CP 139 - 57 rue Cuvier
75231 Paris Cedex 05
France

Prof Tamara Aripova and Dr Tatyana Hegay
Institute for Immunology,
Uzbek Academy of Sciences
74 Ya. Gulamov Street
Tashkent, 100060
Uzbekistan

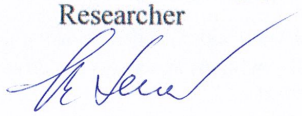
Professor Evelyne Heyer

Head of Human population Genetics team

Prof. Tamara Aripova
Director of Institute of Immunology,
Uzbek Academy of Sciences
(Tashkent, Uzbekistan)



Dr. Tatyana Hegay
Researcher



Informations ethno-démographiques des populations étudiées pour les données autosomales *genome-wide*

Âge des participants

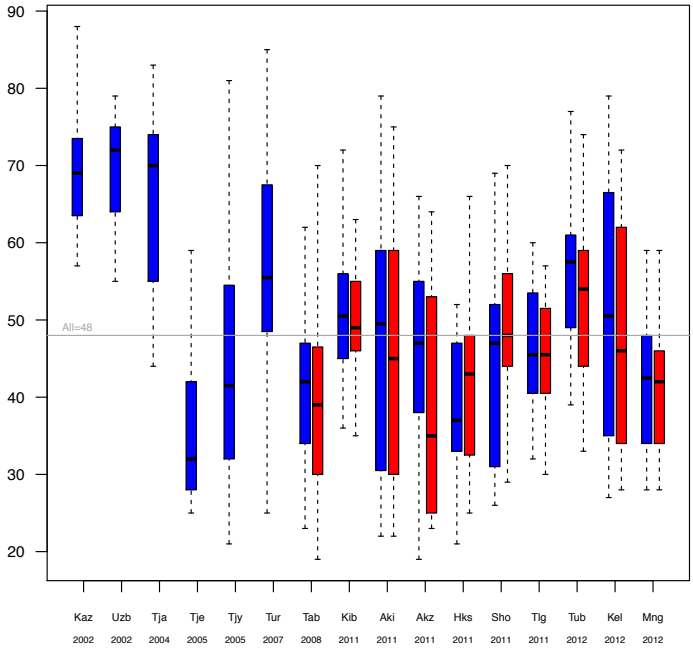


Figure 27 – Âge des participants au moment de l'échantillonnage. L'année de collecte est indiquée sous le trigramme de la population.

Sex-ratio

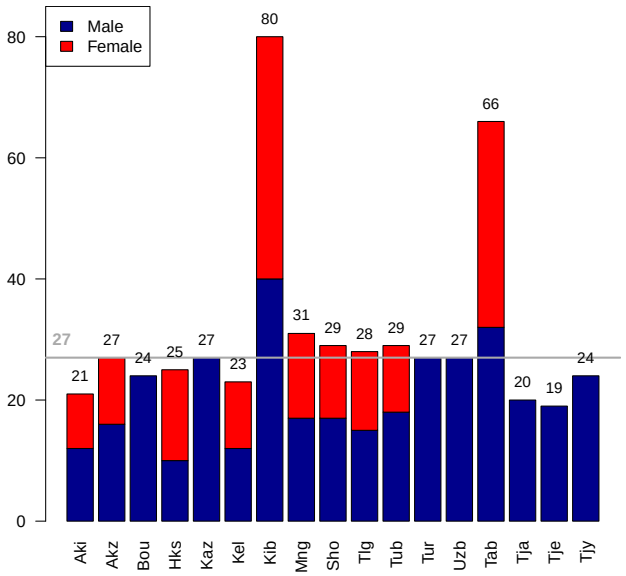


Figure 28 – Nombre d'hommes et de femmes échantillonnés par population.

Règle de résidence observée chez les couples exogames et endogames

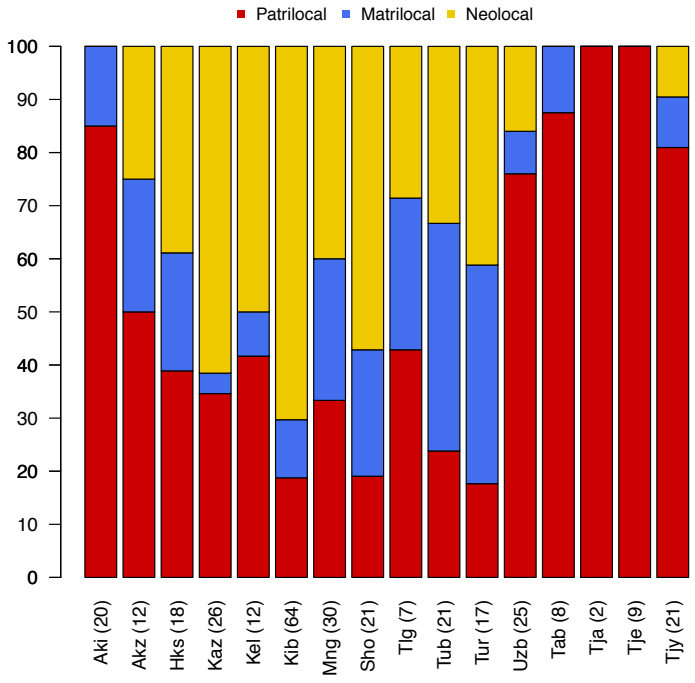


Figure 29 – Règle de résidence des couples exogames à 4 km : patrilocalité (en rouge), matrilocalité (en bleu) ou néolocalité (en jaune). Au total, 43% des couples exogames sont patrilocaux, 18% matrilocaux et 39% néolocaux.

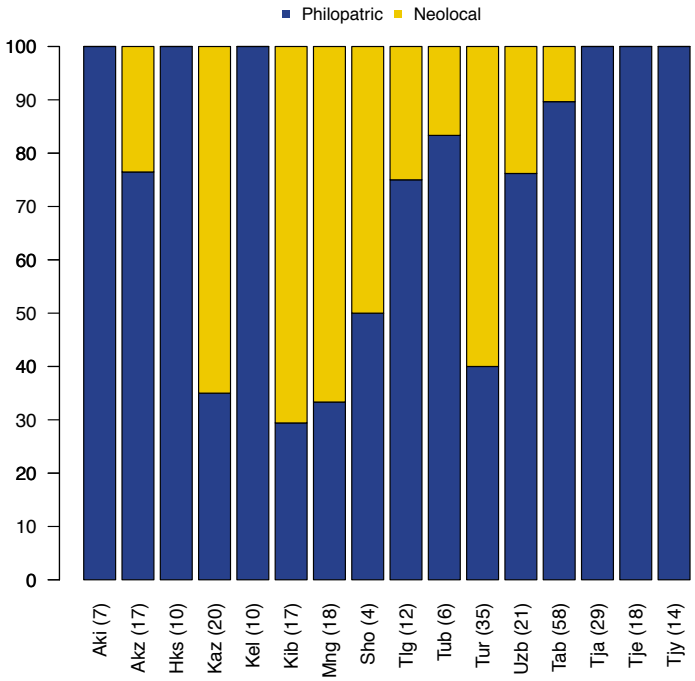


Figure 30 – Règle de résidence des couples endogames à 4 km : philopatrie des deux conjoints (en bleu) ou néolocalité (en jaune).

C Annexes du Chapitre I

Article soumis dans Nature - Version non définitive

A population genomic history of the Eurasian steppe

Peter de Barros Damgaard, Nina Marchi, Simon Rasmussen, Michaël Peyrot, Gabriel Renaud, Thorfinn Korneliussen1, Jose Victor Moreno-Mayar, Mikkel Winther Pedersen, Amy Goldberg, Emma Usmanova, Nurbol Baimukhanov, Valeriy Loman, Lotte Hedeager, Anders Gorm Pedersen, Kasper Nielsen, Gennady Afanasiev, K. Akmatov, Almaz Aldashev[1], Ashyk Alpaslan[2], Gabit Baimbetov, Vladimir I. Bazaliiskii, Arman Beisenov, Bazartseren Boldbaatar[1], Bazartseren Boldgiv[2], Choduraa Dorzhu, Sturla Ellingvag, Diimaajav Erdenebaatar, Rana Dajani, Evgeniy Dmitriev, Valeriy Evdokimov, Karin Frei, Andrey Gromov, Alexander Goryachev, Hakon Hakonarson, Tatyana Hegay, Zaruhi Khachatryan, Ruslan Khashkhanov, Egor Kitov, Alina Kolbina, Tabaldiev Kubatbek[2], Alexey Kukushkin, Igor Kukushkin, Nina Lau, Ashot Margaryan, Inga Merkyte, Ilya V. Mertz, Viktor K. Mertz, Enkhbayar Mijiddorj, Vyacheslav Moiyesev, Gulmira Mukhtarova, Bekmukhanbet Nurmukhanbetov, Zh. Orozbekova, Irina Panyushkina, Karol Pieta, Václav Smrčka, Irina Shevnina, Andrey Logvin, Karl-Göran Sjögren, Tereza Štolcová, Kadicha Tashbaeva, Alexander Tkachev, Turaly Tulegenov, Dmitriy Voyakin, Levon Yepiskoposian, Sainbileg Undrakhbold, Victor Varfolomeev, Andrzej Weber, Nikolay Kradin, Morten E. Allentoft, Ludovic Orlando, Rasmus Nielsen, Martin Sikora, Evelyne Heyer, Kristian Kristiansen, Eske Willerslev*

Abstract

The Eurasian steppe stretching about 8000 kilometres from Hungary and Romania in the west, to Mongolia and western China in the east, is culturally among the most dynamic areas in the world. In the past four millennia, it has been variously dominated by Iranian-, Turkic- and Mongolic-speaking groups, and its temperate grasslands have been a crossroad for extensive movements of peoples, goods, and ideas between Europe, Siberia, South and East Asia. In order to understand the genetic history of the Eurasian steppe populations, we have sequenced the genomes of 137 ancient humans (~1X average coverage) spanning a 4000 year period. Additionally we genotyped 502 individuals from 16 contemporary self-reported ethnic groups. We find evidence of a highly dynamic population history: the Iranian-speaking Scythians that dominated the Eurasian steppe throughout the Iron Age (~1st millennium BCE to common era) emerged following admixture between Late Bronze Age herders of western Eurasian descent and East Asian hunter-gatherers. The steppe nomads later further admixed with Turkic- and Mongolic-speaking groups of East Asian ancestry that spread westward across the steppe in multiple waves: firstly, the Xiongnu confederations that emerged in Mongolia around the 3nd/2nd century BC; secondly, the Huns (4-5th century CE), infected with plague basal to the Justinian *Y. pestis* strain that destabilized the eastern Roman Empire in the 6th century CE; and thirdly during various short term dynasties, including the Mongol Empire of Genghis Khan and his descendants. These recent historical events transformed the Eurasian steppe populations from being Indo-European speakers of largely western Eurasian ancestry to the present-day Turkic-speaking groups, primarily of East Asian ancestry.

Introduction

The demographic processes that shaped the Eurasian steppe populations are among the most complex recorded and have been inferred from a wealth of scientific approaches, including those of archaeology[1–5], linguistics[6–8], and physical anthropology[2,9–11]. Genetic studies based on uniparental genetic markers[12–26], and to a lesser extent genome-wide markers of contemporary and ancient populations[26–31], have also been carried out. These studies all indicate that the history of the Eurasian pastoral nomads has been incredibly dynamic during the past 5000 years. However, the genetic history of the populations that formed the many cultures roaming the Eurasian steppe remains tentative and may best be understood through analyses of an

extensive genomic dataset of steppe nomads and surrounding groups, sampled through the entire chronology of these nomad cultures.

Recently, cross-geographical and temporal genomic datasets have revealed two processes of major importance to human mobility, influencing the human gene pool on the steppe: first the domestication of the horse during the late 4th millennium BCE, followed by pastoralist expansion from the Pontic-Caspian steppe both towards north-eastern Europe (Yamnaya) and the Altai (Afanasievo)[29,30,32]; secondly, the invention of horse-driven spoke-wheel chariots during the early 2nd millennium BCE amongst the Sintashta horizon in the trans Urals, superseding previous disc-wheel chariots pulled by cattle. Spoke-wheel chariotry was later associated with population movement of pastoralist groups, collectively known as Andronovo and Srubnaya, genetically admixed with early European farmers, and spreading into the central steppe in present-day Kazakhstan, and as far south-east as the Tian Shan mountains and the Tarim basin in present-day China[30,31]. These events were chronologically associated with the oldest documented plague victims[33,34]. However, with genome-wide data published from only nine Iron Age steppe individuals, discontinuous in space and time[26,31], the demographic events impacting the Eurasian steppe populations after the Bronze Age remain largely unexplored.

These periods, spanning from the end of the Bronze Age around 900 BCE until today, saw two major cultural transformations. The first covered the transition from the Bronze Age to the Iron Age when mounted warfare evolved. Mounted nomadism fully developed from 850 BCE, and transformed the increasingly sedentary Late Bronze Age herding populations into a powerful mobile military force[1]. It gained further momentum by the development of the new iron technology, which introduced lighter and more flexible mail chain/leather body-armour, as well as the introduction of the short-recurved bow that could be used from horse back, and new socketed iron arrows[35]. The invention of mounted archery infused a new military dynamic into the existing steppe societies, with implications for territorial organization and dominance and hence population mobility. This is reflected in the vast territorial organizations, first of the Iranian-speaking Scythians and their characteristic culture from Hungary in the west to Altai in the east[36], and later the Hunnic expansion into Europe in the 4th-5th century CE[37,38], the Turk Khaganates[39–41], Kimak and Cuman-Kipchak Khanates[42] and eventually the Mongol Empire[42–45] that engulfed most of the Eurasian steppe belt under unified nomad warrior confederations.

The second cultural transformation, which resulted in profound linguistic changes, started during the 3rd and 2nd centuries BCE in the eastern steppe in Mongolia, with the formation of the Xiongnu confederation, which resisted attempts of the Chinese Han dynasty to take control over the nomadic steppe territory[46]. These powerful East Asian nomads, likely in part speaking an early form of Turkic[7,47], established a communication network between Europe and East Asia[44,48], progressively spreading the Xiongnu culture throughout Central Asia. This cultural spread continued until it eventually encompassed most of the steppe region during the westward expansion of the Huns, who shared notable cultural elements with the Xiongnu and ruled during the 3rd and 4th centuries CE. Like the Scythians, the Huns shared a material culture, notably cauldrons for feasting[49], and formed short-lasting confederations, such as the empire of Attila with its royal seat in the great Hungarian plain in the Roman Pannonia province[50]. During their western expansion and conquests, the Huns triggered extensive mobility among Goths and Langobards in Europe, and they briefly penetrated deeper into Europe with military allies that threatened the Roman Empire[37,38,51,52]. The steppe populations

had predominantly been Iranian-speaking but became, from this point in time, increasingly Turkic-speaking (see Supplementary Information Section 1 and 2).

These two major transformations during the last three millennia entailed a complex mix of technological/military, social-demographic, and linguistic changes. However, the gene pools of these groups is largely unexplored, because the Eurasian steppe populations has been underrepresented in both modern and ancient genomic studies. Of ancient genomic data, only the Bronze Age is covered[29–31] together with a few Iron Age low coverage genomes from southern Siberia and the Altai[30], and genome-wide sequence data from nine Iron Age individuals from South Siberia and Altai and Pontic steppe[26,31]. To elucidate the population histories relating to the genetic and linguistic formation of present day steppe populations, we generated a comprehensive genomic dataset of ancient and modern Eurasian steppe nomads covering all areas of the steppe, and of surrounding groups.

The ancient and modern datasets

Multiple geographical definitions of the Eurasian steppe exist. In this study, we adopt a broad definition encompassing the regions that encircle the Kazakhstan steppe (central steppe: CS), i.e. the Altai mountains, the Mongolian steppe (eastern steppe: ES), Tian Shan, Hindu Kush, and the Uzbek basins until the Caspian Sea where the western steppe (Pontic-Caspian steppe: PS) stretches into Europe including the Hungarian plains (Figure 1). To screen the skeletal material for DNA preservation we used highly optimized aDNA extraction methods[30,53–55], combined with 'shotgun' sequencing, on ancient teeth and petrous bones of approximately 200 individuals from the region (Figure 1). Of these, we obtained 137 individual genomes sequenced to ~1X average coverage from DNA extracts with an average of 40% human DNA content (see Extended Table 1). We radiocarbon-dated 83 individuals by Accelerator Mass Spectrometry (AMS) at the 14Chrono Centre, Belfast, to obtain an improved chronology (see Extended Table 2). This effort resulted in a dataset covering ~4,000 years (~2500 BCE - ~1500 CE) across the Eurasian steppe (Figure 1), largely completing the chronological transect of the major pastoral societies commenced in previous studies[29–31] here analyzed together. In order to facilitate reading, we created a table (Extended Table 3) that briefly describes each population label mentioned in this article, providing a fast key to contextualize the results presented here. We underline that this table is not intended to be a precise and nuanced description of the diversity of past cultural horizons; for which information can be found in the Supplementary Section 1. We analyzed the ancient data in the context of SNP genotypes from modern populations[56], previously published low coverage ancient genomes[30], and ancient 'captured' genome-wide data[56] on 1,233,553 genomic variable positions. Furthermore, we genotyped 502 individuals of 16 self-reported ethnicities from across Central Asia, Altai, Siberia and the Caucasus (presented in Supplementary Section 4) and merged with published Eurasian data[28] for a final overlap of 242,406 genomic positions. In order to illustrate gene flow across time and geographical regions between two groups in question, all possible proxies were inserted in the D-statistic of the form D(Test, Mbuti; Group 1, Group 2), denoted according to the nomenclature in[63] indicating gene flow between Test and Group 1 when positive and between Test and Group 2 when negative. All statistics significantly deviating from 0 were plotted. Admixture models were constructed using qpAdm [29] with a set of outgroups designed to distinguish between ancestral components of major relevant gene pools discussed herein. Lastly, in order to improve our understanding of DNA molecule preservation in ancient skeletal material, we also conducted an investigation on differential DNA preservation in the mineral matrix of the teeth as compared to DNA preserved in the organic fibrils, presented in Supplementary Section 5.

The genetic composition of the Scythian and Sarmatian confederations

Starting at around 800 BCE, the Eurasian steppe became dominated by the Iranian-speaking Scythians. This period of Scythian domination lasted throughout most of the Iron Age, i.e. until 200 BCE. The Scythians were divided into geographically distinct groups, but were united by similarities in cultural expression, including opulent tombs, art style, and mounted warfare[36]. However, the origins and population structure of the Scythians are controversial due to conflicting interpretations of the archaeological and genetic records. The existing views can be grouped into the following two models: in the first (A), the Scythians are viewed as product of a conquest from a single source originating in the northern Caucasus/steppe region[25,57] while in the second model (B) they are viewed as a product of multiple transitions taking place locally involving social and cultural borrowing in combination with gradual, small scale human movements[3,26].

The archaeological chronology of the central steppe suggests a major population turnover associated with the onset of the Scythian period, where the increasing aridity at the end of the Bronze Age from 1200- 800 BCE led to a gradual exodus of many groups thus depleting the central steppe from human occupation[58]. According to the multiple transitions model (B), the rapid increase in precipitation from 800 BCE[59] improved grazing productivity over large stretches of the central steppe, attracting nomads from the Khakassian and Minusinsk Basins. These migrations formed or were integrated into the Scythian's cultural complex in an area including also Tuva, where early Scythian royal tombs were found[59,60]. According to this model (B), these processes were independent of populations emerging from the Pontic Steppe.

In contrast, the single source conquest model (A) posits an origin for the Scythians in the northern Caucasus[61] and a later expansion into eastern Europe, in a series of conquest migrations, as described by Herodotus[35]. In this model, the Scythians would represent a homogenous group with a single origin showing cultural similarities. In the multiple transition model, European Scythians and Inner Asian Scythians, called Sakas, were all genetically distinct populations, and cultural differences between them indicate a confederal organization of separate tribes.

An additional complication is that the Scythians were not the only force occupying the Eurasian steppe during the Iron Age. In the western steppe area, the Sarmatians formed a large Iranian-speaking group adjacent to and partnering in war alliances with some European Scythians. Their genetic relationship with the Hungarian Scythians is unknown.

Genetic studies on uniparental markers of Inner Asian Sakas have argued that Scythians arose following admixture between Bronze Age herders of west Eurasian genomic ancestry and Inner Asian hunter-gatherers[18,19,21–24]. These studies made a full genome assessment of this putative East Asian admixture into the Iron Age steppe necessary, and recently, this East Asian contribution to Scythians was confirmed through genome-scale SNP-data from six Inner Asian Sakas[26]. However, the exact sources of the Bronze Age herders and Asian hunter-gatherers giving rise to the Scythians are still debated, as is the level of gene flow between the various Scythian groups. Recently, a mtDNA study argued for continuity between western Scythians and Late Bronze Age Srubnaya populations, rather than Andronovo-related herders[25]. However, based on genome-scale data it has recently been claimed that the source of west Eurasian ancestry in Iron Age Scythians was neither Srubnaya nor Andronovo herders and their descendents, but rather nomads genetically associated to early Bronze Age

Yamnaya/Afanasievo populations[26]. Based on mtDNA evidence, the same study also suggested substantial gene flow between Inner Asian Sakas and western Scythians.

Our genome dataset allowed us to reconcile all these studies within a single consistent scenario. Using PCA and admixture analyses (Figure 2 and 3), we observe a clear separation between two groups of Iron Age Scythians: the Hungarian Scythians and the Inner Asian Sakas. Furthermore, we find fine-scaled structure within the Inner Asian Sakas separating: 1) the 'Tagar' of southern Siberia, 2) the 'Central Sakas' of the Central steppe, of which most have been described as belonging to the Tasmola culture (see Supplementary Section 3), and 3) the 'Tian Shan Sakas' of the Tian Shan mountain range (see Figure 1 for geographical distribution and Figure 3 for qualitative assessment of clusters). This reflects the confederal nature of the Scythian organization, and supports the multiple transition model (B) with gradual, small scale movement[3] rather than the conquest model (A) with its single source in the northern Caucasus/Pontic steppe region[57].

Based on mtDNA it has been claimed that western Scythians were direct descendants of populations belonging to the Srubnaya culture[25], while Andronovo formed a cluster with Bronze Age Siberians and Catacomb individuals from present-day Ukraine. Using genome-wide data, we find no D-statistics of the form D('Iron Age Scythians', Mbuti; Srubnaya, Andronovo) deviating significantly from 0. We caution that the Srubnaya genomes and the southern Siberian Andronovo genomes analyzed here are extremely similar (see Extended Figure 1 'Transversions only', and Extended Figure 2 for pairwise plot of all possible values of f4(Source1, Outgroup1 ; Outgroup2, Outgroup3) used in qpAdm modelling[29] based on a set of 7 outgroups: Mbuti, Ust'Ishim, Clovis, Kostenki14, Switzerland_HG, Natufian and MA1). The results may also be affected by low sampling density: the Andronovo horizon has been interpreted as being structured between the more southwestern Alakul and the southern Siberian Fedorovo. Moreover, the chronologically last phases of the Sargary-Alekseev culture[62] is still unrepresented in ancient genomics studies. We suspect that their relationship to western Scythians may be more complex than currently discernible and we therefore abstain from a conclusive interpretation until greater sampling density of the Late Bronze Age steppe is achieved.

In turn, our data do not support the recent contention that Early Bronze Age Yamnaya/Afanasievo represent a direct source population to the Iron Age Scythians[26]. Unlike the Yamnaya/Afanasievo cluster, the Late Bronze Age herder populations display increased European farmer ancestry relative to lower proportions of Eastern Hunter-Gatherer-like ancestry[29–31]. After reproducing the qpAdm models used to claim Yamnaya/Afanasievo ancestry in Scythians[26], we used a new set of outgroups that enabled us to distinguish between early and Late Bronze Age steppe ancestry: Mbuti, Ust'Ishim, Clovis, Kostenki14, Switzerland_HG, Natufian and MA1 (see Extended Figure 3 for pairwise plots of all possible values of f4(Source1, Outgroup1 ; Outgroup2, Outgroup 3)). We find that the Late Bronze Age herders are a better genetic source for the west Eurasian ancestry in Scythians than are Early Bronze Age Yamnaya/Afanasievo, the key difference being their European farmer ancestry (Extended Table 4). Using ADMIXTURE models we also illustrate the shared ancestry between Neolithic farmers (from Anatolia or Europe), Late Bronze Age herders, and Iron Age steppe nomads that is not shared with Yamnaya herders (Extended Figure 4 and Supplementary Section 4). Lastly, we find that the genomes from Bronze Age hunter-gatherers from the Lake Baikal area (of the Glazkovo culture denoted BHG_BA for Baikal Hunter-Gatherer Bronze Age) are a better fit for the East Asian ancestry in Scythians than the recently suggested Han or Nganasan[26] (Extended Table 4).

Using D-statistics, we illustrate that the four distinct Scythian genetic clusters likely emerged through gene flow from neighbouring groups into Late Bronze Age herder communities. First, we show that Hungarian Scythians had relatively increased European farmer ancestry (D(Europe_EN, Mbuti; Hungarian Scythian, Andronovo)=0.02; Z= 7.66) (Extended Figure 5) and show no signs of East Asian gene flow (D(BHG_BA, Mbuti; Hungarian Scythian, Andronovo) = 0). Inner Asian Sakas show relatively increased East Asian ancestry with the strongest gene flow observed into the Central Sakas (D(BHG_BA, Mbuti; Central Saka, Andronovo)=0.05; Z=18.5, followed by Tian Shan Sakas: D(BHG_BA, Mbuti; Tian Shan Saka, Andronovo)=0.03; Z=12.7 and lastly the Tagar of Southern Siberia: D(BHG_BA, Mbuti; Tagar, Andronovo)=0.03; Z=10.). This East Asian admixture is also reflected in the negative 'admixture f3s'[63] (Extended Figure 6). Next we illustrate the relevant differences between the four Scythian clusters: in particular the increase in East Asian ancestry in Central Sakas relative to the three other groups through D(BHG_BA, Mbuti; Central Saka, Tagar) = 0.02; Z = 11.76; D(BHG_BA, Mbuti; Central Saka, Tian Shan Saka) = 0.02; Z= 10.58; D(BHG_BA, Mbuti; Central Saka, Tian Shan Saka) = 0.06; Z = 27.34. The increase in Neolithic Iranian ancestry in the Tian Shan Sakas is significant when compared to Central Sakas: D(Iran_N, Mbuti; Tian Shan Saka, Central Saka) = 0.01; Z = 3.2. Tagar displays increased EHG ancestry as compared to all other Scythians: D(EHG, Mbuti; Tagar, Hungarian Scythians) = 0.01; Z=3.81; D(EHG, Mbuti; Tagar, Central Sakas) = 0.01; Z=5.31; D(EHG, Mbuti; Tagar, Tian Shan Sakas) = 0.01; Z=6.11. The increase in European farmer ancestry in Hungarian Scythians relative to all other Sakas is highly significant: D(Neolithic Europe, Mbuti; Hungarian Scythians, Tian Shan Sakas) = 0.04; Z = 24.44 ; D(Neolithic Europe, Mbuti; Hungarian Scythians, Tagar) = 0.03; Z = 18.63 ; D(Neolithic Europe, Mbuti; Hungarian Scythians, Central Sakas) = 0.05; Z = 25.12. Lastly, the high genetic differentiation between western and eastern Scythians is emphasized by observing higher F_{st} values between Hungarian Scythians and all Inner Asian Sakas (F_{st} ranges from 0.24 to 0.3) than observed among the different Inner Asian Sakas groups (F_{st} ranges from 0.15 to 0.2) (see Extended Table 5).

The qpAdm modelling[29] of this ancient genomic dataset is consistent with these findings. The Central Sakas can be modelled as a simple two-way mixture of Late Bronze Age pastoralists and East Asian hunter-gatherers (BHG_BA), with almost equal proportions of Bronze Age herder (56 %) and East Asian hunter-gatherer ancestry (44 %). Intriguingly, contrasting the differential ancestral contributions on the X-chromosomes and the autosomes, we observe an increase in male East Asian hunter-gatherer and steppe pastoralist female contribution in this area (Extended Figure 7). The Southern Siberian Tagar show unequal ancestry contributions from Bronze Age herders (83.5 %) and East Asian hunter-gatherers (7.5 %), but also an additional contribution of Mal'ta (MA1) like ancestry (9 %), likely indicating that the BHG_BA are a poor proxy for admixture in the Tagar, and that the true source was geographically Siberian or Altaian in origin where present-day distribution of MA1-like ancestry is highest[101]. We note that in order to estimate this ancestral proportion we removed MA1 as an outgroup for this particular admixture model and replaced it with EHG. Similarly, the Saka population of the Tian Shan mountain display a high proportion of Late Bronze Age steppe herder ancestry (70 %) followed by East Asian hunter-gatherer ancestry (25 %), as well as an additional 5 % ancestry from a source related to a Neolithic population from Iran. Taken together, our data do not support the recent mtDNA-based claim of extensive gene flow between the different Scythian groups[26], but indicate instead local admixture between populations of Late Bronze Age herder descent and various local groups, in agreement with the multiple origins model (B).

The Sarmatians were another Iranian-speaking force that occupied the western fringe of the steppe, distinguished from their neighbors in particular by the prominent position of warrior women in their society[2]. We therefore investigated the genetic affinities between Scythians and the Sarmatians, both representing groups of Iron Age nomads. Principal Component Analysis, admixture analysis (Figure 2, Extended Figure 8), and D-statistics (Extended Figure 9) revealed that the Sarmatians of the Late Iron Age are genetically shifted towards present day East Asians and Altaians as compared to the Hungarian Scythians. The relevant D-statistic is D(BHG_BA, Mbuti; Sarmatians, Hungarian Scythians)=0.03; Z= 11.09. Thus, the neighbouring Hungarian Scythians and Sarmatians were clearly genetically distinct.

Altogether, our data show that the culturally similar Scythians and Sarmatians represented genetically structured groups within the Eurasian steppe. In particular, the Siberian Tagar, Central Sakas, and the Tian Shan Sakas were Scythian groups that arose through admixture between Late Bronze Age pastoral groups and Inner Asian hunter-gatherers, in contrast to the Hungarian Scythians that received gene flow from farming groups within Europe, thus showing less East Asian ancestry than the neighbouring Sarmatians. The additional gene flow from a source related to the Neolithic Iranians detected in the Tian Shan Sakas suggests that southern steppe nomads interacted with the civilization of the Bactria-Margiana Archaeological Complex (BMAC) of present-day southern Uzbekistan and eastern Turkmenistan[6]. These findings rule out recently proposed population models where Scythians arose from remnant Early Bronze Age Yamnaya/Afanasievo populations[26].

Xiongnu and the Hunnic expansions

Turkic language elements arguably first emerged among the Xiongnu nomads[65]. This confederation of several nomadic tribes occupied the eastern steppe from the 3rd century BCE and are believed to be of East Asian ancestry[19,20] although ancient Y-chromosomal data have indicated a possibly heterogeneous population admixed with central steppe nomads[66]. Ancient Chinese sources described the Xiongnu and the Huns as being similar in terms of culture and customs, society structure, and political institutions. For these reasons, Huns have been argued to derive directly from the Xiongnu[52] while others claim that there is no evidence connecting Huns (3rd-5th century CE) with Xiongnu[67]. It is commonly believed that the Huns spread westward, disseminating Turkic languages throughout Central Asia at the cost of Iranian languages. Today, Turkic-speaking populations encompass more than 20 diverse ethnicities that are widely distributed across Eurasia. This interpretation has recently gained some support from genome-wide marker analyses of contemporary populations, indicating that Turkic languages spread through elite dominance across the steppe, during a time approximately corresponding to the Hun period, except for particular populations that were impacted later during the Mongol Period (13th-14th century CE)[28]. However, both the Xiongnu and the Huns were military confederations and may, like the Scythians, have consisted of genetically differentiated groups. Therefore, the relationship between the Xiongnu and the Huns and the spread of the Turkic languages may be more complex. For example, linguistically, there is evidence of influence of pre-Proto-Mongolic on early Turkic, probably during the Xianbei period when these defeated the Xiongnu (2nd-4th Century CE)[47]. Finally, it is known that the expansion of the Xiongnu nomads impacted the movements of other cultural groups from the south-eastern side of the Tian Shan Mountains, such as the Wusun and Kangju for whom the genetic ancestries and linguistic affiliations have so far remained unknown. Based on the archaeological record, it has tentatively been suggested that they belonged to the Iranian-speaking branch of Indo-European speakers[78].

Principal Component analyses and D-statistics indicate that the Xiongnu individuals fall in two distinct groups, one being of East Asian origin, and the other presenting considerable admixture levels with West Eurasian sources (Figure 2, Extended Figure 10 and 11). This suggests population movements and admixture into the Mongolian steppe, possibly i) during the Early Bronze Age by Afanasievo-like populations, or ii) in the Late Bronze Age through Andronovo-like populations, or lastly iii) during the Iron Age through Saka-like populations. We find that Central Sakas are accepted as source in a single-wave model. In line with this finding, no East Asian geneflow is detected compared to Central Sakas as these form a clade with respect to Xiongnu (ie. D(Xiongnu, Mbuti; Central Sakas, Xiongnu_WE) = ~0), and cluster together in the PC Analysis (Figure 2), indicating a close genetic contact between the central steppe and the eastern steppe during the Iron Age. We henceforth use Xiongnu as a label for the East Asian individuals and Xiongnu_WE for the admixed individuals.

We then used D-statistics of the form D(Test, Mbuti; Xiongnu, Tian Shan Huns) to investigate the genetic relationship between Iron Age nomads, the East Asian Xiongnu, and the early Huns of the Tian Shan. We find that the Huns have increased shared drift with West Eurasians as compared to the Xiongnu, and with all Iron Age nomads tested (Extended Figure 12). Furthermore, qpAdm modelling indicates that the Hunnic individuals from the Tian Shan received most of their ancestry from Iron Age Sakas (as ~90%), and not the east Asian Xiongnu which contributed to only ~10% of their ancestry. However, the East Asian contribution could not be detected on the X-chromosome, suggesting that the ~10% ancestry described above resulted from a male-driven expansion. We note that these X-chromosomal analyses do not allow us to confidently assess the correct East Asian proportion on the X-chromosome, but the striking difference between the autosomal and the X-chromosomal ancestral proportions likely reflects that the Huns of the Tian Shan emerged through a minor male-driven East Asian gene flow into the existing local Saka population. We also note the presence of genetic outliers of East Asian descent in the steppe during this period, with the earliest burial of an individual clustering at the extreme of East Asian variation in the PCA, dated to 200 BCE (Figure 2).

Finally, we tested for patterns of shared drift between the Xiongnu and the Wusun, the preceding Sakas, and the slightly later Huns (2nd century CE). We find that both the earlier Sakas and the later Huns have more East Asian ancestry than the Wusun (D(Xiongnu, Mbuti; Tian Shan Saka, Wusun) = 0.02 ; Z = 8.87 and D(Xiongnu, Mbuti; Tian Shan Hun, Wusun) = 0.03 ; Z = 14.03). This is also apparent from model-based clustering and PCA (Extended Figure 13). Similar results are seen with the contemporaneous and later Kangju groups that, like the Wusun, re-emerged into the steppe from south-east of the Tian Shan mountains. Both groups require a Neolithic Iranian-related source (Iran_N) for modelling ancestral proportions in the qpAdm framework (Extended Table 6), in addition to Bronze Age pastoralists (Steppe_MLBA) and the East Asian hunter-gatherers (BHG_BA). Thus, we suspect that the Wusun and Kangju groups represent descendants of Bronze Age pastoralists that interacted with the BMAC civilization in southern Uzbekistan/eastern Turkmenistan, implying these harbored Neolithic Iranian ancestry, but remained much less admixed with East Asians when compared to Sakas.

Over all, our data show that the Xiongnu confederation was genetically heterogeneous. They also reveal that the Huns emerged following minor male-driven East Asian gene flow into the preceding Sakas that they invaded. As such our results confirm the contention that the disappearance of the Scythians/Sakas around the beginning of the Common Era was a cultural

transition that coincided with the westward migration of the Xiongnu. The Xiongnu invasion also led to the displacement of isolated remnant groups related to Late Bronze Age pastoralists living in the south-eastern side of the Tian Shan mountains.

Repeated conquests and waves of East Asian impact

In the 6th Century CE the Hunnic Empire had been broken up and dispersed, as a new unified Khaganate, the Turk Khaganate, overtook the military and political domination of the steppe[39,41]. Khanates/Khaganates are steppe nomad political organizations that can vary in size, and became dominant during this period, in contrast with previous stateless organization during the Iron Age[46]. Written Chinese sources trace the Turk Khaganate origins back to a blacksmith working force of the Rouran Khaganate in the eastern steppe of Mongolia, 4-6th Century CE[69]. The Turks then sought independent rule by moving south and west, where the Hunnic Empire collapsed, and they gained control over the Silk Road while expanding into Eastern Europe, reaching Crimea in 576 CE[70]. The Turk Khaganate was eventually replaced by a number of short lived steppe-cultures[40]. These included the Kipchak and the Tungusic Kimak populations spreading southwards towards the Tian Shan and westward towards the Ural Mountains forming the Kimak Khaganate of the central steppe during the 8th to 11th Century CE[42,71]. Towards the 11th Century, the Kimak Khaganate was overthrown by local Kipchak groups, who in turn allied themselves with the Cuman of western Eurasia. Eventually, the short-lived Khaganates were overtaken by the Mongol Empire that emerged through unification of East Mongolian and Transbaikalian tribes. It expanded heavily during Genghis Khans' rule in the 13th Century CE, lastly including most of the Eurasian steppe[42–45]. Mongolic influence in later Turkic languages is discussed in Supplementary Section 2.

We find evidence that elite soldiers associated with the Turk Khaganate are genetically closer to East Asians than are the preceding Huns of the Tian Shan mountains (D(Xiongnu, Mbuti; Turk, Tian Shan Hun)= 0.02; Z=6.53). One Turk period nomad was a genetic outlier with pronounced European ancestries indicating the presence of ongoing migrations onto the central steppe. Only one sample here represents Kimak nomads, and it does not show elevated East Asian ancestry, as the D-statistics of the form D(Xiongnu, Mbuti; Kimak, Turk) does not deviate from 0. During the Kipchak period in the 11th Century CE, the rule was allegedly taken over by another group originating from the geographical area of Tuva. We have genomic data from two individuals from this period, one of which shows East Asian ancestry, while the other has pronounced European ancestry (Supplementary Material Section 3, outgroup-f3 DA23 and DA179). These individuals date back to the Cuman-Kipchak alliance, which englobed both the western and eastern steppe. When the region became incorporated into the Karakhanid Khanate that encompassed present-day regions of Uzbekistan, Tajikistan, Kazakhstan and Kyrgyzstan, D-statistics of the form D(Xiongnu, Mbuti; Karakhanid, Turk) do not identify an influx of East Asian ancestry as compared to the earlier Turk period. However, we do note that the two genetically unrelated nomads from the Karakhanid period are shifted towards East Asians when compared to Turks in the PCA plot (Figure 2 and Extended Figure 14). Additionally, we analyzed 10 culturally unaffiliated Medieval nomads, most showing pronounced East Asian ancestry, albeit at very different proportions (Extended Figure 14). We also find the presence of an individual of western European descent buried together with members of Jochi Khan's Golden Horde army from the Ulytau mountains (Supplementary Section 3 – Individuals DA28 is 'East Asian' and DA29 is 'European'). This could suggest assimilation of distinct groups into the Golden Horde, but may also represent a slave or a servant of west Eurasian descent.

Our data suggest that Turk cultural customs were imposed by an East Asian minority elite onto central steppe nomad populations, resulting in a small detectable increase in East Asian ancestry. However, we also find that steppe nomad ancestry in this period was extremely heterogeneous with several individuals being genetically distributed at the extremes of the first principal component (Figure 2) separating Eastern and Western descent. Based on this notable heterogeneity, revealing a gradient of East Asian ancestry across these periods, we interpret that during Medieval times, the steppe populations were exposed to gradual admixture from the East, and the strong variation is a direct window into ongoing admixture processes. As such, our data show that present-day levels of East Asian ancestry in central steppe nomad populations (as evident in the Kazakhs and Kyrgyz) were only reached at homogenous levels after these Medieval times (Figure 4).

Origins and spread of the Justinian plague

Recent findings suggest that plague epidemics could have been a driver in the Bronze Age migrations across Europe and Asia, and perhaps were diffused through these migrations[33]. A few decades after the period of Hunnic driven mobility across the Eurasian steppe, large areas of Europe were depopulated due to the Justinian plague pandemic[72]. While the first reports of the pandemic point to an outbreak in Egypt from where it is thought to have spread into Europe[73], the primordial origins of the Justinian plague remain unknown. However, the most basal strains of present-day plague (0.PE7 clade) have been found in Qinghai, south-east of the Tian Shan mountains[74], and the clade basal to the Justinian plague (0.ANT1) was found in Xinjiang in China, pointing to a possible Inner Asian origin of the Justinian plague.

We screened for genetic traces of *Yersinia pestis,* the agent of plague, in the non-human DNA sequences obtained from the dataset passing first round of human contamination assessment (139 ancient individuals), using an approach previously outlined[33]. We found two individuals to show detectable levels of *Y. pestis* DNA, compatible with the characterization of the full genome sequence at 8.7X and 0.24X coverage, respectively. The first individual was a Hun from the Tian Shan Mountains (DA101), dating to 182 CE, while the second individual was a member of the Alanic culture from the Rostov region (DA147), dating several centuries later (6^{th}-9^{th} century CE). The genome from the *Y. pestis* strain DA101 (Supplementary Table 6.1 and Supplementary Table 6.2 in Supplementary Section 6), which we name 0.ANT5, branches off from the main plague lineage just basal to the Justinian plague strain 0.ANT4, identified from an individual in Aschheim, Germany and dated to 533 CE[73] (Supplementary Figure 6.3 in Supplementary Section 6). The limited bootstrap support obtained for the node (44/100) is, however, compatible with other branching patterns but supports the close genetic proximity of the two strains. Investigating high quality SNVs between the Tian Shan and the Aschheim strain, we identify 92 variants specific for the Aschheim strain and only 5 variants specific for the Tian Shan strain. As expected the Tian Shan strain contained the *ymt* gene reported to be missing in the more ancestral Bronze Age plague strains[33]. The strain also displayed the loss of function mutations in *pde2*, *pde3*, *rcsA* and *ureD* that are required for flea transmission in the traditional blocked flea model (Supplementary Figure 6.4 in Supplementary Section 6)[70–72]. This, coupled with a fully functional plasminogen activator gene, indicates that the 'Hunnic' plague strain had full bubonic capability and flea transmissibility[75].

The high number of strain-specific variants in the Aschheim strain is in line with the difference in sampling time (182 CE *vs.* 533 CE) and the potentially multiple replication cycles associated with pandemics[74]. To investigate this, we used the MCMC samples generated during the BEAST[76] analysis reported in[33], and compared the substitution rates on the Aschheim branch to the substitution rates on all other branches in the *Y. pestis* phylogeny.

The data strongly supports a substantially higher rate for the Aschheim strain: specifically, we find that the substitution rate on the branch leading to the Aschheim strain was higher than 97.7% of all substitution rates in the tree (posterior mean, 95% HPD interval: 0.94-0.99), and that the posterior probability for this rate being higher than any other rate in the tree was 27% (Supplementary Figure 6.5 in Supplementary Section 6). Mutation rates in pathogens have been hypothesized to be affected by epidemics; not only because of natural selection, but simply due to an increase in replication rate[74]. Therefore, our observation of an accelerated mutation rate is in agreement with this hypothesis and that the Ascheim strain was responsible for a major outbreak, namely the Justinian plague.

Given that the most basal strains of present-day plague (0.PE7 clade) originate from Qinghai[74], and the clade basal to the Justinian plague (0.ANT1) is from Xinjiang, China, two areas close to the Tian Shan, we find provisional support for a hypothesis that the pandemic was brought to Europe towards the end of the Hunnic period through the Silk Road along the southern fringes of the steppe.

Discussion

Our study supports the following population history for the central Eurasian steppe over the past 5000 years: i) Early Bronze Age Yamnaya herders first admixed with European farmers, giving rise to populations that later migrated eastwards, forming the Middle to Late Bronze Age Sintashta/Andronovo communities (~2000-1000 BCE), ii) Late Bronze Age nomads then admixed with East Asian hunter-gatherers (BHG_BA), that were themselves an admixture between a Paleosiberian Mal'ta like source and an ancient East Asian source, forming the Scythians (900-200 BCE), iii) Local nomads gradually admixed with multiple expanding East Asian Turkic-speaking sources, including the Xiongnu (200 BCE – 200 CE), second the Huns (200-500 CE), and finally several short-lived Khaganates such as the Turk Khaganate, Karakhanid Khanate, and Mongol Empire (500-1300 CE). The population history that formed the genetic composition of present-day steppe populations is illustrated in Figure 4, where we model the entire known ancient and present-day diversity of Central Asia using the key ancestral groups: EHG (mesolithic hunter-gatherers from Eastern Europe), WHG (mesolithic hunter-gatherers from Western Europe), East Asian hunter-gatherers (BHG_BA from Lake Baikal), and the late Upper Paleolithic hunter-gatherers from the Near East (Natufian). We illustrate that our set of outgroups is adequate using pairwise f4-statistics of the form D(Source1, Outgroup1; Outgroup2, Outgroup3) with a set of five outgroups differentially related to the sources (Mbuti, Ust'Ishim, Clovis, Kostenki14 and Switzerland HG) in Extended Figure 15.

Our findings fit well with the current insights in the historical linguistics of this region (see Supplementary Section 2). The steppe was likely largely Iranian-speaking in the 2nd and 1st millennia BCE. This is supported by the split of the Indo-Iranian linguistic branch into Iranian and Indian[77–79], the distribution of the Iranian languages, and the preservation of Old Iranian loanwords in Tocharian[80,81]. The wide distribution of the Turkic languages from Northwest China, Mongolia and Siberia in the east to Turkey, Bulgaria, Romania and Lithuania in the west implies large-scale migrations out of the homeland in Mongolia since the beginning of the Common Era[7,82]. The diversification within the Turkic languages suggests that several waves of migrations occurred[83], and on the basis of the impact of local languages gradual assimilation to local populations must be assumed[84]. The East Asian migration starting with the Xiongnu complies well with the hypothesis that early Turkic was their major language[7,47]. Further migrations of East Asians westwards find a good linguistic correlate in the influence

of Mongolian on Turkic and Iranian in the last millennium[8] and the recent spread of the Mongolic Kalmyck language to west of the Caspian Sea[85]. No obvious evidence is, however, available for a linguistic interpretation of the East Asian pulse detected in the Asian Sakas in the 1st millennium BCE.

Climatic changes have been a suggested as key drivers of human migrations across the steppes[86]. The onset of a dry period after 1200 BCE has been argued as being a primary reason for Late Bronze Age Andronovo's descendant populations going south and leaving the steppe deserted[2]. Later, the onset of a wet period from 850 BCE allowed new areas to become populated such as Tuva in south Siberia. It has even been suggested that Saka tradition in the central steppe could not have had local roots because the area was previously depopulated[58]. We can reject this scenario, as we find clear evidence that Iron Age Saka received a large proportion of their genetic ancestry from the Bronze Age herders. We agree that climatic changes most likely together with diseases such as the plague, played an important role for human migrations, at the onset of the Saka period, where a severe drought followed by increased humidity has been registered[59]. In contrast, the existence of remnant Yamnaya/Afanasievo-like population is incompatible with the archaeological and genomic evidence. It has also been hypothesized that the harsh winters in the Mongolian steppe, resulting in loss of livestock, may have been a major driver of the west and southward raiding by the Xiongnu and later Turks[86] at the end of the Iron Age.

With regards to the population dynamics following the Bronze Age, the Eurasian steppe history appears in strong contrast to human demographic changes in Europe where most of the European ancestral components were already established by this time[29,30]. In contrast, the steppe received continued gene flow from neighbouring groups. Bioarchaeological research of Scythians suggest that young males had a higher death rate than other sedentary Iron Age populations, despite good health conditions, leading to a decrease in population size due to extensive warfare[87]. Hence, engaging with newcomers might have helped steppe nomads maintain population sizes. This is consistent with our findings of multi-ethnical cultural assemblages and the presence of gender-biased migrations.

We find it reasonable to speculate that mounted warrior nomad culture facilitated a constant recruitment of new groups through confederal alliances. Taking control over the steppe and the Silk Road might also have been an incentive for continued warfare, leading to the rise and fall of the many groups entering and assimilating into steppe nomads from the Iron Age to Medieval times. In any event, the repeated impacts of migratory waves from the Iron Age to Medieval times lead to a predominance of East Asian ancestry in contemporary steppe populations. As such, the genomic history of the Eurasian steppe is the story of a gradual transition from Bronze Age pastoralists of western Eurasian ancestry, towards mounted warriors of increased East Asian ancestry – a process that continued into historical times.

References

1. Koryakova, L. & Epimakhov, A. V. *The Urals and western Siberia in the Bronze and Iron ages*. (Cambridge University Press, 2014).

2. Davis-Kimball, J., Bashilov, V. A. & Yablonsky, L. T. *Nomads of the Eurasian steppes in the early Iron Age*. (Citeseer, 1995).
3. Bashilov, V. A. & Yablonsky, L. T. Some current problems concerning the history of Early Iron Age Eurasian Steppe nomadic societies. *Kurgans Ritual Sites Settl. Eurasian Bronze Iron Age Ed. J Davis-Kimball EM Murphy Koryakova LT Yablonksy Br. Archaeol. Rep. Int Ser. S* **890,** 9–12 (2000).
4. Bokovenko, N. The emergence of the Tagar culture. *Antiquity* **80,** 860–879 (2006).
5. Legrand, S. The emergence of the Scythians: bronze age to iron age in south Siberia. *Antiquity* **80,** 843–859 (2006).
6. Lubotsky, A. The Indo-Iranian substratum. *Early contacts between Uralic and Indo-European.* Carpelan ea 301–317 (2001).
7. Janhunen, J. *Manchuria: an ethnic history*. (Finno-Ugrian Society, 1996).
8. Doerfer, G. 1963-1975. *Türk. Mongolische Elem. Im Neupersischen* 1–4
9. Kozintsev, A. G. The 'Mediterraneans' of Southern Siberia and Kazakhstan, Indo-European migrations, and the origin of the Scythians: A multivariate craniometric analysis. *Archaeol. Ethnol. Anthropol. Eurasia* **36,** 140–144 (2008).
10. Schmidt, R. W. & Evteev, A. A. Iron Age nomads of southern Siberia in craniofacial perspective. *Anthropol. Sci.* **122,** 137–148 (2014).
11. Movsesian, A. A. & Bakholdina, V. Y. Nonmetric cranial trait variation and the origins of the Scythians. *Am. J. Phys. Anthropol.* **162,** 589–599 (2017).
12. Chaix, R., Austerlitz, F., Hegay, T., Quintana-Murci, L. & Heyer, E. Genetic traces of east-to-west human expansion waves in Eurasia. *Am. J. Phys. Anthropol.* **136,** 309–317 (2008).
13. Aimé, C., Heyer, E. & Austerlitz, F. Inference of sex-specific expansion patterns in human populations from Y-chromosome polymorphism. *Am. J. Phys. Anthropol.* **157,** 217–225 (2015).
14. Balaresque, P. *et al*. Y-chromosome descent clusters and male differential reproductive success: young lineage expansions dominate Asian pastoral nomadic populations. *Eur. J. Hum. Genet.* **23,** 1413–1422 (2015).
15. Pérez-Lezaun, A. *et al.* Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am. J. Hum. Genet.* **65,** 208–219 (1999).

16. Lalueza-Fox, C. *et al.* Unravelling migrations in the steppe: mitochondrial DNA sequences from ancient Central Asians. *Proc. R. Soc. Lond.-B* **271,** 941–948 (2004).
17. Comas, D. *et al.* Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *Eur. J. Hum. Genet.* **12,** 495–504 (2004).
18. Ricaut, F.-X., Keyser-Tracqui, C., Bourgeois, J., Crubézy, E. & Ludes, B. Genetic analysis of a Scytho-Siberian skeleton and its implications for ancient Central Asian migrations. *Hum. Biol.* 109–125 (2004).
19. Keyser-Tracqui, C., Crubezy, E. & Ludes, B. Nuclear and mitochondrial DNA analysis of a 2,000-year-old necropolis in the Egyin Gol Valley of Mongolia. *Am. J. Hum. Genet.* **73,** 247–260 (2003).
20. Keyser-Tracqui, C., Crubézy, E., Pamzsav, H., Varga, T. & Ludes, B. Population origins in Mongolia: genetic structure analysis of ancient and modern DNA. *Am. J. Phys. Anthropol.* **131,** 272–281 (2006).
21. González-Ruiz, M. *et al*. Tracing the origin of the east-west population admixture in the Altai region (Central Asia). *PLoS One* **7,** e48904 (2012).
22. Pilipenko, A. S., Molodin, V. I., Trapezov, R. O., Cherdantsev, S. V. & Zhuravlev, A. A. A Genetic Analysis of Human Remains from the Bronze Age (2nd Millennium BC) Cemetery Bertek-56 in the Altai Mountains. Anthropology and Paleogenetics. (2016).
23. Pilipenko, A. S., Trapezov, R. O. & Polosmak, N. V. A paleogenetic study of Pazyryk people buried at Ak-Alakha-1, the Altai Mountains. *Archaeol. Ethnol. Anthropol. Eurasia* **43,** 144–150 (2015).
24. Gubina, M. A. *et al.* The dynamics of the composition of mtDNA haplotypes of the ancient population of the Altai Mountains from the early bronze age (3rd millennium BC) to the iron age (2nd–1st centuries BC). *Russ. J. Genet.* **52,** 93–106 (2016).
25. Juras, A. *et al.* Diverse origin of mitochondrial lineages in Iron Age Black Sea Scythians. *Sci. Rep.* **7,** (2017).
26. Unterländer, M. *et al.* Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. *Nat. Commun.* **8,** (2017).
27. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
28. Yunusbayev, B. *et al.* The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet* **11,** e1005068 (2015).

29. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (2015).
30. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522,** 167–172 (2015).
31. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528,** 499–503 (2015).
32. Anthony, D. W. & Ringe, D. The Indo-European homeland from linguistic and archaeological perspectives. *Annu Rev Linguist* **1,** 199–219 (2015).
33. Rasmussen, S. *et al.* Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. *Cell* **163,** 571–582 (2015).
34. Valtuena, A. A. *et al*. The Stone Age Plague: 1000 years of Persistence in Eurasia. *bioRxiv* 094243 (2016).
35. Kristiansen, K. *Europe before history*. (Cambridge University Press, 2000).
36. Chlenova, N. L. On the Degree of Similarity between M aterial Culture Compo—nents within the 'Scythian World'. *Archaeol. Steppes Methods Strateg. Napoli* 520–521 (1994).
37. Bóna, I. *Das Hunnenreich*. (Corvina, 1991).
38. Thompson, E. A. *The Huns (Revised and with an afterword by P. HEATHER)*. (Oxford: Blackwell, 1996).
39. Sinor, D. The establishment and dissolution of the Türk empire. *Camb. Hist. Early Inn. Asia* 285–316 (1990).
40. Golden, P. B. An introduction to the history of the Turkic peoples. *Ethnogenesis State-Form. Mediev. Early Mod. Eurasia Middle East* 127–36 (1992).
41. Findley, C. V. *The Turks in world history*. (Oxford University Press, 2004).
42. Hildinger, E. *Warriors of the Steppe: A Military History of Central Asia, 500 BC to 1700 AD*. (Da Capo Press, 1997).
43. Kradin, N. N. & Skrynnikova, T. D. Imperiya Chingis-Khana [The Genghis Khan Empire]. *Mosc. Vostochnaya Lit. RAN* (2006).
44. Biran, M. The Mongol Empire in World History: The State of the Field. *Hist. Compass* **11,** 1021–1033 (2013).
45. Fitzhugh, W. W., Rossabi, M. & Honeychurch, W. *Genghis Khan and the Mongol Empire*. (Genghis Khan Exhibits, 2009).

46. Kradin, N. Stateless empire: The structure of the Xiongnu nomadic super-complex chiefdom. *Xiongnu Archaeol. Multidiscip. Perspect. First Steppe Emp. Inn. Asia* 77–96 (2011).
47. Schönig, C. Turko-Mongolic relations. *The Mongolic languages*, Ed. J. Janhunen. London, 403-419 (2003).
48. Brosseder, U. Xiongnu Empire. *Encycl. Emp.* (2016).
49. Érdy, M. Archaeological Continuity between the Xiongnu and the Huns. *DSCA J.* 11 (2008).
50. Vingo, P. Shifting populations in Late Antiquity. Germanic populations, nomads and the transformation of the Pannonian limes. *Acta Archaeol.* **61,** 261–282 (2010).
51. Heather, P. The fall of Rome. *New Hist.* (2006).
52. Pohl, W. Goths and Huns. *Companion Ethn. Anc. Mediterr.* 555–568 (2014).
53. Damgaard, P. B. *et al.* Improving access to endogenous DNA in ancient bones and teeth. *Sci. Rep.* **5,** (2015).
54. Pinhasi, R. *et al.* Optimal ancient DNA yields from the inner ear part of the human petrous bone. *PloS One* **10,** e0129102 (2015).
55. Hansen, H. B. *et al.* Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum. *PLoS One* **12,** e0170940 (2017).
56. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536,** 419–424 (2016).
57. Grakov, B. N., Yelagina, N. G. & Yatsenko, I. V. The Early Iron Age. *Mosc. Mosc. State Univ. Russ.* (1977).
58. Chlenova, N. L. Predystorija torgovogo puti Gerodota. *Sov. Arxeologija* **1,** (1983).
59. van Geel, Bas, et al. "Climate change and the expansion of the Scythian culture after 850 BC: a hypothesis." *Journal of Archaeological Science* 31.12 (2004): 1735-1742.
60. Armbruster, B. Gold technology of the ancient Scythians–gold from the kurgan Arzhan 2, Tuva. *ArchéoSciences* 187–193 (2009).
61. Petrenko, V. G. *Scythian culture in the North Caucasus*. (Citeseer, 1995).
62. Frachetti, M. D. *Pastoralist landscapes and social interaction in Bronze Age Eurasia*. (Univ of California Press, 2009).
63. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192,** 1065–1093 (2012).
64. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–1664 (2009).

65. Dybo, A. V. Linguistic contacts of early Turks. *Lex. Fund Pre-Turk. Period Moscow«Vostochnaya Lit. Publ.* (2007).

66. Kim, K. *et al.* A western Eurasian male is found in 2000-year-old elite Xiongnu cemetery in Northeast Mongolia. *Am. J. Phys. Anthropol.* **142,** 429–440 (2010).

67. De la Vaissière, É. Huns et Xiongnu. *Cent. Asiat. J.* **49,** 3–26 (2005).

68. Schönig, C. & Johanson, L. *Discoveries on the Turkic linguistic map. Swedish Research institute in Istanbul*. (JSTOR, 2003).

69. Kradin, N. N. From tribal confederation to empire: The evolution of the Rouran society. *Acta Orient.* **58,** 149–169 (2005).

70. Beckwith, C. I. *Empires of the silk road: A history of central Eurasia from the Bronze Age to the present*. (Princeton University Press, 2009).

71. Kumekov, B. E. & Sulejmenov, B. *Gosudarstvo kimakov X-XI vv. po arabskim istočnikam*. (1972).

72. Little, L. K. *Plague and the end of antiquity: the pandemic of 541-750*. (Cambridge University Press, 2007).

73. Wagner, D. M. *et al.* Yersinia pestis and the Plague of Justinian 541–543 AD: a genomic analysis. *Lancet Infect. Dis.* **14,** 319–326 (2014).

74. Cui, Y. *et al.* Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. *Proc. Natl. Acad. Sci.* **110,** 577–582 (2013).

75. Zimbler, D. L., Schroeder, J. A., Eddy, J. L. & Lathem, W. W. Early emergence of Yersinia pestis as a severe respiratory pathogen. *Nat. Commun.* **6,** (2015).

76. Bouckaert, R. *et al.* BEAST 2: a software platform for Bayesian evolutionary analysis. *PLoS Comput Biol* **10,** e1003537 (2014).

77. Kuz'mina, E. E. *The origin of the Indo-Iranians*. (Brill, 2007).

78. Mallory, J. P. In Search of the Indo-Europeans/Language, Archaeology and Myth. *Praehistorische Z.* **67,** 132–137 (1992).

79. Parpola, A. *The roots of Hinduism: the early Aryans and the Indus civilization*. (Oxford University Press, USA, 2015).

80. Tremblay, X. Irano-Tocharica et Tocharo-Iranica. *Bull. Sch. Orient. Afr. Stud.* **68,** 421–449 (2005).

81. Peyrot, M. Tocharian language. In: E. Yarshater (ed.), *Encyclopedia Iranica.* www.iranicaonline.org/articlestocharian-language (2015)

82. Nichols, J. Forerunners to globalization: The Eurasian steppe and its periphery. Language contact in times of globalization. Ed. C. Hasselblatt et al. 177–195 (2011).
83. Johanson, L. The history of Turkic. *Turk. Lang.* 81–125 (1998).
84. Johanson, L. Turkic language contacts. *Handb. Lang. Contact* 652–672 (2010).
85. Bläsing, U. Kalmuck. The Mongolic languages, Ed. J. Janhunen. London, 229-47 (2003).
86. Kradin, N. N. Nomadic Empires of Inner Asia. In J. Bemmann & M. Schmauder (eds.): *Complexity of Interaction along the Eurasian Steppe Zone in the First Millennium CE:* 11-48. Bonn. (2015)
87. Lukasik, S. *et al.* Warriors die young: increased mortality in early adulthood of Scythians from Glinoe, Moldova, 4th–2nd c. BC. *J. Anthropol. Res.* (2017).
88. Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nat. Protoc.* **2,** 1756–1762 (2007).
89. Dabney, J. & Meyer, M. Length and GC-biases during sequencing library amplification: a comparison of various polymerase-buffer systems with ancient and modern DNA sequencing libraries. *Biotechniques* **52,** 87–94 (2012).
90. Fu, Q. *et al.* A revised timescale for human evolution based on ancient mitochondrial genomes. *Curr. Biol.* **23,** 553–559 (2013).
91. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16,** 224 (2015).
92. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9,** 88 (2016).
93. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
94. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
95. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).
96. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).
97. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. in *Proc. R. Soc. B* **279,** 4724–4733 (The Royal Society, 2012).

98. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. & Orlando, L. mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* btt193 (2013).

99. Skoglund, P., Storaa, J., Götherström, A. & Jakobsson, M. Accurate sex identification of ancient human remains using DNA shotgun sequencing. *J. Archaeol. Sci.* **40,** 4477–4482 (2013).

100. Korneliussen, T. S. & Moltke, I. NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* btv509 (2015).

101. Wong, Emily HM, et al. "Reconstructing genetic history of Siberian and Northeastern European populations." *Genome research* 27.1 (2017): 1-14.

Acknowledgments

We thank Kim Magnussen, Lillian Petersen, Cecilie Mortensen and Andaine Seguin-Orlando at the Danish National Sequencing Centre for producing the analyzed sequences. We thank Paula Reimer and Stephen Hoper at the 14Chrono Center Belfast for providing AMS datings. We thank Susanne Hackenbeck for discussing paleodietary reconstructions. We thank Helene Elisabeth Heyerdahl, the Explico Foundation team, Ditte Christiansen Appelt, Jainagul Isakova, Batyrzhan Daulet, Aidyn Tairov, Nurlan Abduov, Bakhtishat Tudiyarov, Vladimir Volkov, Maksum Akchurin, Ilyas Baimukhan, Nikolay Namdakov, Yuldash Yusupov, Erlan Ramankulov, Arman Nurgaziyev, Abdul Kusaev for important assistance in fieldwork. We thank Jesper Stenderup, Pernille V. Olsen and Tina Brand for technical assistance in the laboratory. We thank all involved archaeologists, historians and geographers from Kazakhstan: Alexander Suslov, Irina Erofeeva, Erzhan Nurmaganbetov, Baktyar Kozhakhmetov, Nadezhda Loman, Yuri Parshin, Sergey Ladunskiy, Marina Bedelbaeva, Antónia Marcsik, Oliver Gábor, Marek Půlpán, Yerkin Kubeev, Rymbek Zhumashev, Khyshish Omarov, Serik Kasymov, Umut Akimbayeva. We thank Paul Rodzianko for creating the initial contact between PBD, SE and EU. We thank Sten Jacobsen and Jonh O'Brien for translating and proof-reading Russian contributions. The project was funded by the Danish National Research Foundation (EW), the Lundbeck Foundation (EW), and KU2016 (EW).

Author contributions:

EW: initiated and led study
PBD, EW, EU, EH: designed study
PBD: produced the data
PBD, NM, SR, MS, GR, TK[1], AG, MWP, AGP, KN: analyzed or assisted in analysis of data
PBD, EW, KK: interpreted results with considerable input from MS, RN, MP, NK, SR, LO, MEA, JVMM
PBD, EW, KK, MP, SR: wrote the manuscript with considerable input from NK, LH, MS, RN, MEA, LO, JVMM and contributions from all authors
PBD, MEA, LO, EU, NB, VL, GA, KA, AA[1], AA[2], GB, VIB, AB, BB[1], BB[2], CD, SE, DE, RD, ED, VE, KF, AG, AG, HH, TH, ZK, RK, EK, AK, TK[2], AK, IK, NL, AM, VKM, IVM,

IM, EM, VM, GM, BN, ZO, IP, KP, VS, IS, AL, KGS, TS, KT, AT, TT, DV, LY, SU, VV, AW excavated, curated, sampled and/or described analyzed skeletons

All authors contributed to final interpretation of data.

Competing financial interests.

The authors declare no competing financial interests.

Figure Captions

Figure 1. Cultural and geographical presentation of the ancient samples presented in this study. a, The geographical distribution of all samples. Symbols corresponds to samples of a specific age; circle: Bronze age, square: Iron Age, diamond: Hun period, triangle upwards: Turk period, triangle downwards: Medieval times. b, Each symbol have been sorted according to geographical region highlighted on map (panel a), and are given in the grey boxes in panel b. Abbreviations corresponds C = Caucasus, CAS = Caspian steppe, CS = Central steppe, ES = Eastern steppe, HP = Hungarian plains, PS = Pontic steppe, STE = Siberia, Tungus & Eastern Steppe. The barplot presents the known cultural time interval.

Figure 2. Principal Component Analyses. The Principal Components 1 and 2 were plotted for the ancient data analyzed together with the modern data (no projection bias). A) PCA plot highlighting new ancient data and new modern data. See Supplemenyary Section 4 for description of modern data. Dimension 1 explaining 3% of the variance represents a gradient stretching from Europe to East Asia, defining most Central Asian diversity. Dimension 2 explaining 0.6 % of the variance represents a gradient mainly represented by ancient DNA starting from 'basal-rich' cluster of Natufian hunter-gatherers and ending with Eastern Hunter-Gatherer (EHG), a gradient that defines most west Eurasian diversity. B) Main genomic 'trajectory' across time for central steppe nomads is highlighted.

Figure 3. Analyses of Iron Age clusters. A) PCA of Iron Age nomads and ancestral sources that explain the diversity between them. B) PCA of Iron Age nomads alone. C) Model-based clustering at K=7 illustrating differences in ancestral proportions – individuals labelled (A) Andronovo, (B) Neolithic European (Europe_EN), (C) East Asian Hunter-Gatherer (BHG_BA), (D) Neolithic Iranian (Iran_N). Here we illustrate the admixture analyses with K=7 as it approximately identifies the major component of relevance (Anatolian/European farmer component, Caucasian ancestry, EHG related ancestry and East Asian ancestry).

Figure 4. QpAdm results depicts the changes in ancestry across time in Central Asia. The changes reflect a gradual increase in East Asian ancestry in the central steppe nomads coupled to a decrease in ancestry associated to Eastern-Hunter Gatherers, starting high in Yamnaya and finishing low in present-day Kazakhs/Kyrgyz. The set of outgroups used is: Mbuti, Ust'Ishim, Clovis, Kostenki14 and Switzerland HG.

Figure 5. Summary review map of the main migratory events associated to the genomic history of the steppe pastoralists 5kyr ago until today.

Methods

Sample selection

Approximately 200 relevant skeletons were first screened for their human DNA content and contamination proportions. Sample substrates were either tooth cementum, sampled as in[53], or petrous bone sampled as in[54,55]. All pre-PCR sample processing was undertaken in the dedicated clean-laboratory facilities of the Centre for GeoGenetics, Natural History Museum, University of Copenhagen according to the following steps:

1) Samples were pre-digested for 30 minutes at 50C with a proteinase K and EDTA buffer, modified from[53] by excluding the N-lauryl sarcosyl, because surfactants have no effect on digestion of bone or tooth[88]. Samples were finally digested in 4.9 mL EDTA and proteinase K at either 24 hours or 48 hours in an incubator at 50C.

2) DNA was extracted using a binding buffer optimized at binding ultra-short DNA fragments to silica particles. The buffer was used in combination with a silica-in-solution approach which has been shown to outperform spin-column purification[30]. The buffer consist of 500 ml Qiagen buffer PB with 9 ml sodium acetate (5M), and 2.5 ml sodium chloride (5M).

3) Next-generation sequencing libraries were then built according to a modified NEBNext DNA Sample Prep Master Mix Set 2 (E6070) incorporating P5 and P7 adaptors as previously described[30]. All libraries were amplified with Kapa U+ which has been shown to be an optimal enzyme for amplifying ancient human DNA due its low GC-bias[53,89]. Libraries were amplified in a two-round amplification set-up as in[30], with a total of 18 to 22 PCR cycles.

For the screening phase, approximately 10-15 million sequences were generated on each sample. A total of 155 individuals qualified for full genome sequencing based on three criteria of which only the third was an absolute disqualifier whenever applied: 1) by containing more than 20 % human DNA, or 2) being of particular relevance to the research questions, and 3) not displaying more than 5% contamination on the upper CI of contamination estimated with the contamMix programme[90]. Through these criteria, sample size was reduced to 145 DNA extracts averaging 40% human DNA content. These resulted in 145 genomes sequenced to an overall average coverage of ~1X, and the authentication step was then extended using the approaches incorporated in Schmutzi[91] resulting in 137 genomes passing contamination criterion – see "DNA authentication" below for outline.

Processing of read data

All libraries were sequenced single-read to 80 bp, on an Illumina HiSeq 2500 at the Danish National High-Throughput Sequencing Centre. The sequences were basecalled using the Illumina software CASAVA CASAVA-1.8.2 and de-multiplexed using a full match of the 6 nucleotide index incorporated during library amplification. The reads were trimmed using AdapterRemoval-2.1.3[92] for adapter sequences and leading/trailing stretches of Ns. Additionally bases with quality of 2 or less were removed by trimming from the 3'. Reads of at least 30 bp were mapped to GRCh37 using bwa-0.7.10[93] with the seed disabled. Alignments were processed using samtools-1.3.1[94] removing reads with a mapping quality lower than 30 and merged to libraries. Hereafter duplicates were removed using picard-1.127 MarkDuplicates (https://broadinstitute.github.io/picard/), libraries merged to sample level and realigned using GATK-3.3.0[95] with Mills and 1000G gold standard indels. Finally, realigned

bams had the md-tag updated and extended BAQs calculated using samtools calmd. Read depth and coverage were determined using pysam (http://code.google.com/p/pysam/) and BEDtools[96]. Statistics of the read data processing is shown in Extended Table 1.

Damage assessment

Several damage parameters were estimated from the data in order to characterize the fragmentation and damage of the ancient human DNA. First, the fragmentation was computed by fitting an exponential model to the decaying part of the sequence length distribution according to the decay model outlined in[97]. Next, position specific mismatches were estimated using mapDamage2.0[98], as well as damage parameters of the Bayesian model implemented in this programme of which we report the δs parameter ie. the probability of deamination within single stranded overhangs characteristic for ancient DNA, in Extended Table 8.

DNA authentication

Contamination of the samples was estimated with two approaches.

First, mapping affinities of sequences that mapped to the mitochondrial genomes using the standard mapping approaches described above were compared to their own reference and to a global dataset of potential contaminant sources, using the contamMix approach[90].

For further authentication, all trimmed reads were aligned to a reference excluding the autosomal chromosomes – the rCRS only, using SHRiMP version 2.2.3. The resulting aligned DNA fragments were used as input by schmutzi, a Bayesian algorithm aiming at co-estimating levels of present-day human contamination, while also inferring the mitochondrial sequence of the endogenous material. Contamination estimates were repeated by disabling the prediction of the contaminant mitogenome (without the option --notusepredC). The endogenous consensus was called with a quality cutoff of 30 on the PHRED scale on the predicted base thus ensuring a probability of error of less than 1/1000. Haplogrep2.0 [nar.oxfordjournals.org/content/early/2016/04/15/nar.gkw233.full] was then used on those mitochondrial sequences to assign haplogroups. Samples with > 10X coverage showing contamination levels > 10 % in the CI from either the contamMix or the schmutzi contamination estimate obtained without predicting the contaminant, were removed from the study. Various criteria were used to filter out low quality mitochondrial sequences namely: 1) Average coverage of more than 25X, 2) Contamination rates less than 10% and 3) Stable Haplogroup assignment when stricter quality cutoffs was used for the prediction endogenous mitochondrial genome. Finally, mitochondrial sequences for which we had more than 1% undefined sites were removed from downstream mitogenome analyses.

Finally, X-chromosome based contamination estimates for the male individuals were performed as an additional confirmation that the data is not affected by contamimation. We used the two approaches implemented in angsd (http://www.popgen.dk/angsd/index.php/ANGSD) using HapMap variable sites exactly as undertaken in the original publication it was used in[102]. The first approach is based on total read count while the other approach is based on sampling random reads. Neither methods detected any potential contamination in the samples.

Genetic sex determination

Gender was determined using the sex chromosome ratio Rγ approach described in[99].

Relatedness analyses

We estimated relatedness using a two-step approach. First we computed all the outgroup-f3 statistics of the form f3(Individual X, Individual Y; Mbuti) using randomsly selected alleles, and flagged all pairs of individuals with excess shared ancestry by setting a threshold of 0.3 for the outgroup f3-statistic. Then, we used ngsrelate[100] to estimate relatedness from these individuals, using a background panel of 1200 Eurasian individuals at the Affymetrix positions from the Human Origins dataset[56] for background frequencies. A cut-off at first degree cousins (ie. relatedness coefficient of 0.0625) was applied. When relatives were identified, the individual with highest coverage was selected for downstream analyses. This resulted in a final selection of 132 unrelated individuals.

Datasets

By merging random alleles from the ancient samples with previously published data, we generated two datasets for the population genomic analyses. We excluded reads with a BAQ score <30 (from bam files that were previously filtered for reads with mapping quality under 30). One panel included the newly generated unrelated 132 samples, random alleles from previously published genomes of relevance[30] and previously published variants including ancient and modern populations[56]. The second dataset was generated by merging random alleles from the newly generated genomes, relevant genomes[30], and a dataset containing overlapping positions between previously published data[28] and new Central Asian and Siberian data generated in this project, see Supplementary Section 4. Related individuals were estimated using plink1.9 and removed from the dataset.

Data analyses

Principal Component Analyses were conducted using the PLINK1.9 software including the ancient samples in the calculations of the components. The model-based clustering was similarly calculated including the ancient samples in the analyses, using the ADMIXTURE software, defining K clusters from K=2 to K=20 and running 20 replicates for each run. The data was converted to eigenstrat format and the D- and f-statistics were calculated using the admixtools package (https://github.com/DReichLab/AdmixTools).

Figure 1

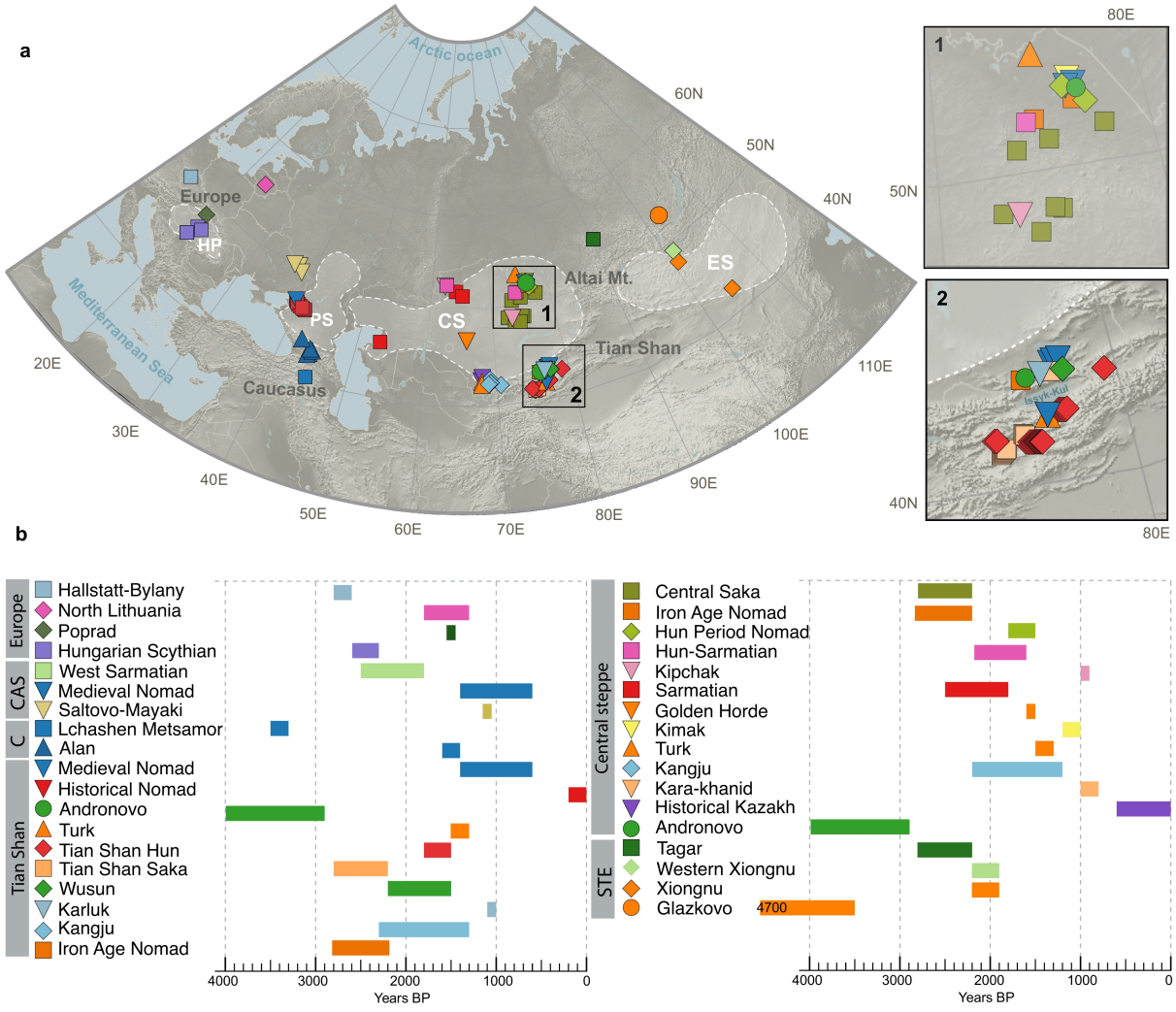

a
Arctic ocean
Europe
HP
Mediterranean Sea
PS
Caucasus
CS
Altai Mt.
1
Tian Shan
2
ES
60N
50N
40N
20E
30E
40E
50E
60E
70E
80E
90E
100E
110E
80E
50N
40N
80E
b
Europe
Hallstatt-Bylany
North Lithuania
Poprad
Hungarian Scythian
CAS
West Sarmatian
Medieval Nomad
Saltovo-Mayaki
C
Lchashen Metsamor
Alan
Tian Shan
Medieval Nomad
Historical Nomad
Andronovo
Turk
Tian Shan Hun
Tian Shan Saka
Wusun
Karluk
Kangju
Iron Age Nomad
Central steppe
Central Saka
Iron Age Nomad
Hun Period Nomad
Hun-Sarmatian
Kipchak
Sarmatian
Golden Horde
Kimak
Turk
Kangju
Kara-khanid
Historical Kazakh
Andronovo
STE
Tagar
Western Xiongnu
Xiongnu
Glazkovo
4700
4000
3000
2000
1000
0
Years BP
4000
3000
2000
1000
0
Years BP

Figure 2

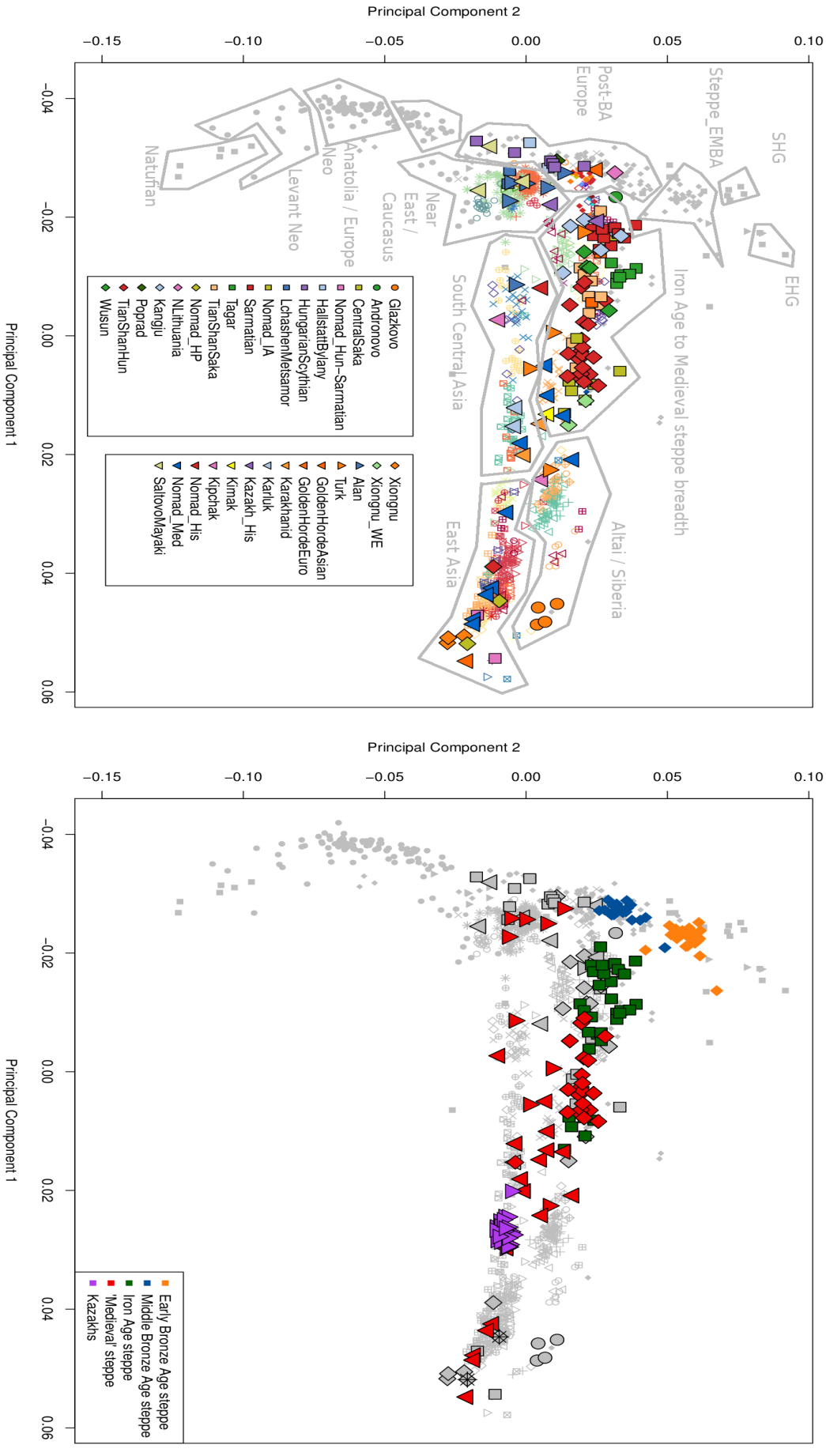

Principal Component 2
−0.15
−0.10
−0.05
0.00
0.05
0.10
Principal Component 1
−0.04
−0.02
0.00
0.02
0.04
0.06
Natufian
Levant Neo
Anatolia / Europe Neo
Near East / Caucasus
Post-BA Europe
Steppe_EMBA
SHG
EHG
South Central Asia
Iron Age to Medieval steppe breadth
East Asia
Altai / Siberia
Glazkovo
Andronovo
CentralSaka
Nomad_Hun−Sarmatian
HallstattBylany
HungarianScythian
LchashenMetsamor
Nomad_IA
Sarmatian
Tagar
TianShanSaka
Nomad_HP
NLithuania
Kangju
Poprad
TianShanHun
Wusun
Xiongnu
Xiongnu_WE
Alan
Turk
GoldenHordeAsian
GoldenHordeEuro
Karakhanid
Karluk
Kazakh_His
Kimak
Kipchak
Nomad_His
Nomad_Med
SaltovoMayaki
Principal Component 2
−0.15
−0.10
−0.05
0.00
0.05
0.10
Principal Component 1
−0.04
−0.02
0.00
0.02
0.04
0.06
Early Bronze Age steppe
Middle Bronze Age steppe
Iron Age steppe
'Medieval' steppe
Kazakhs

Figure 3

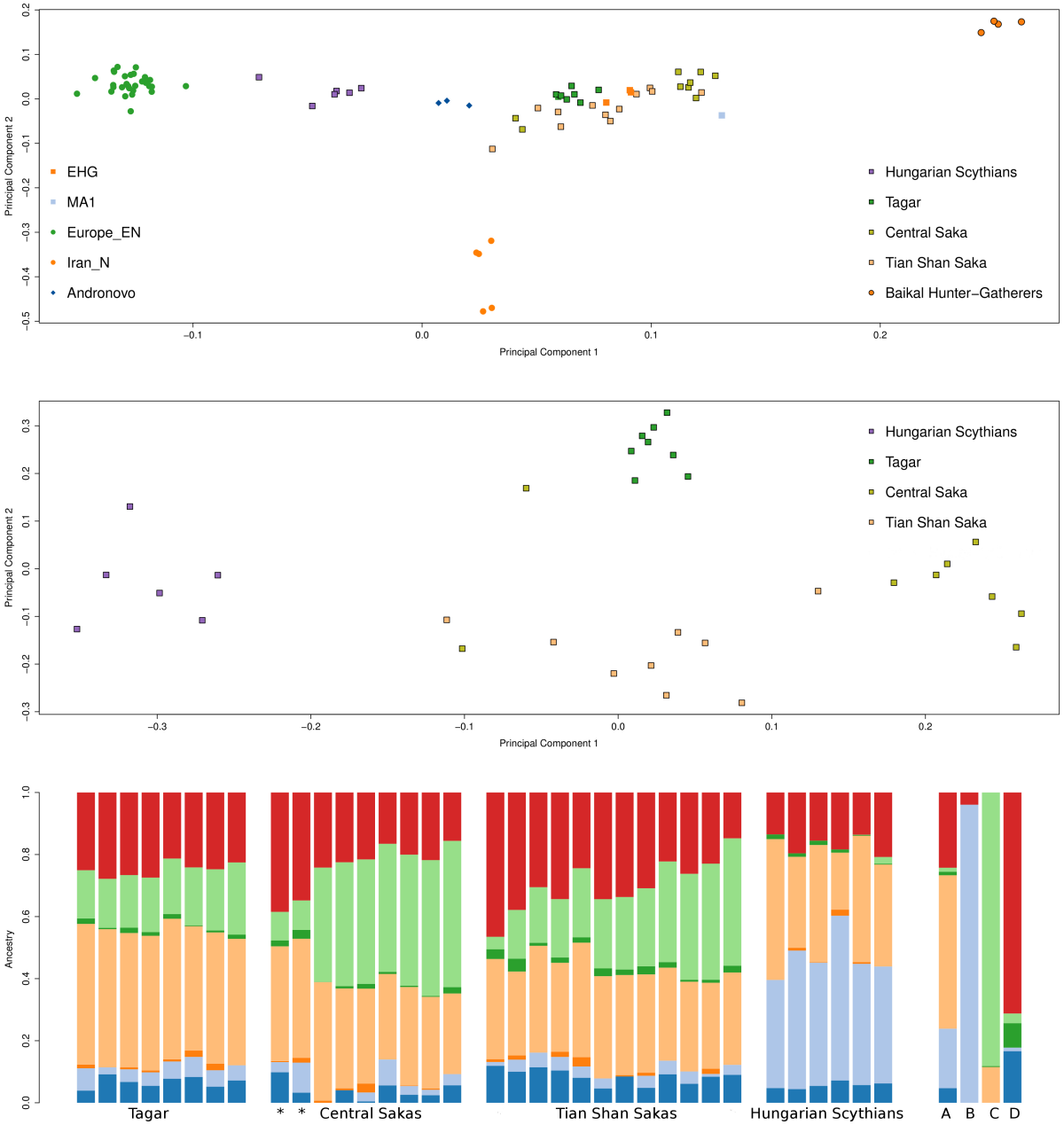

Principal Component 2
Principal Component 1
EHG
MA1
Europe_EN
Iran_N
Andronovo
Hungarian Scythians
Tagar
Central Saka
Tian Shan Saka
Baikal Hunter−Gatherers
Principal Component 2
Principal Component 1
Hungarian Scythians
Tagar
Central Saka
Tian Shan Saka
Ancestry
Tagar
* * Central Sakas
Tian Shan Sakas
Hungarian Scythians
A B C D

Figure 4

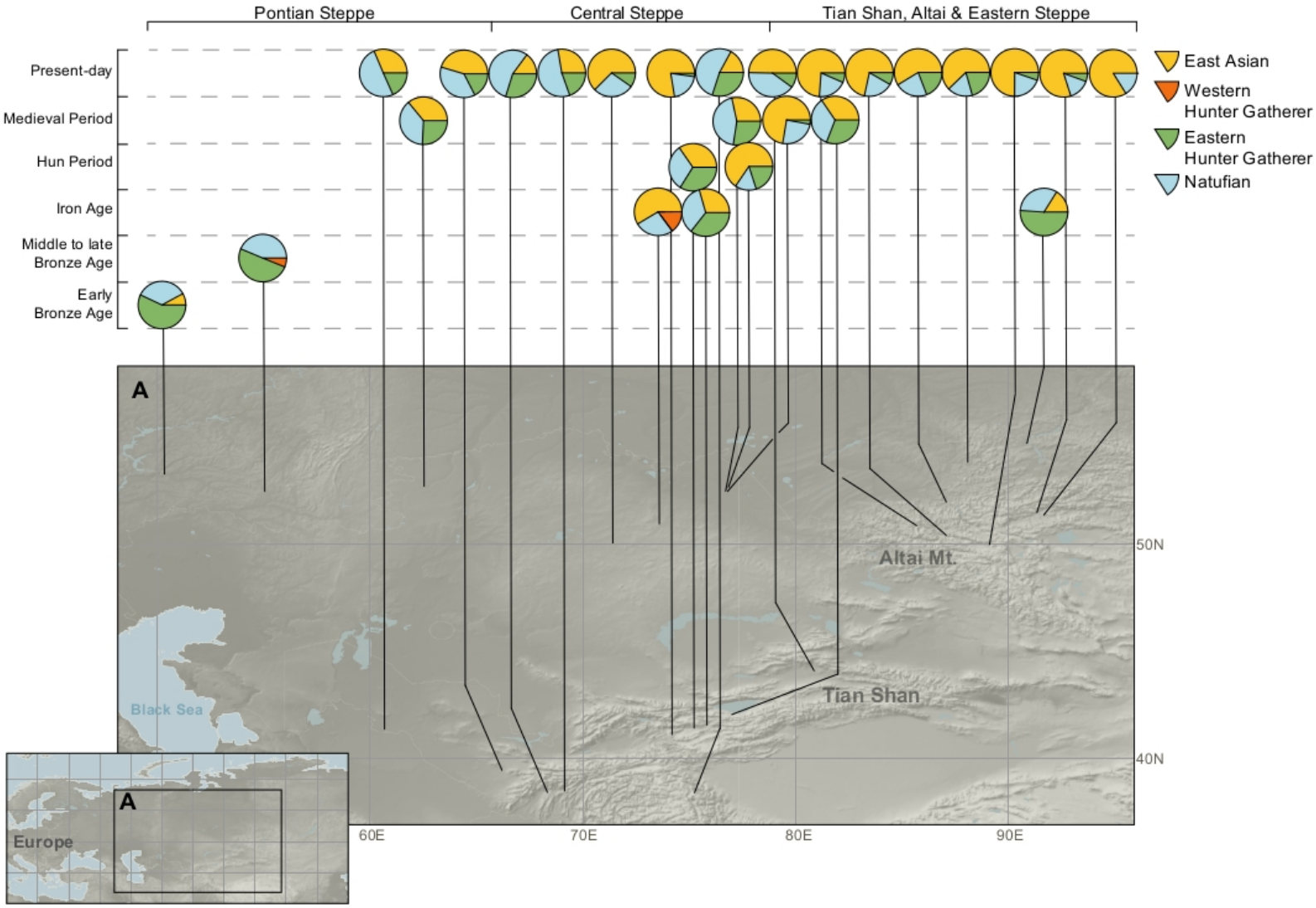
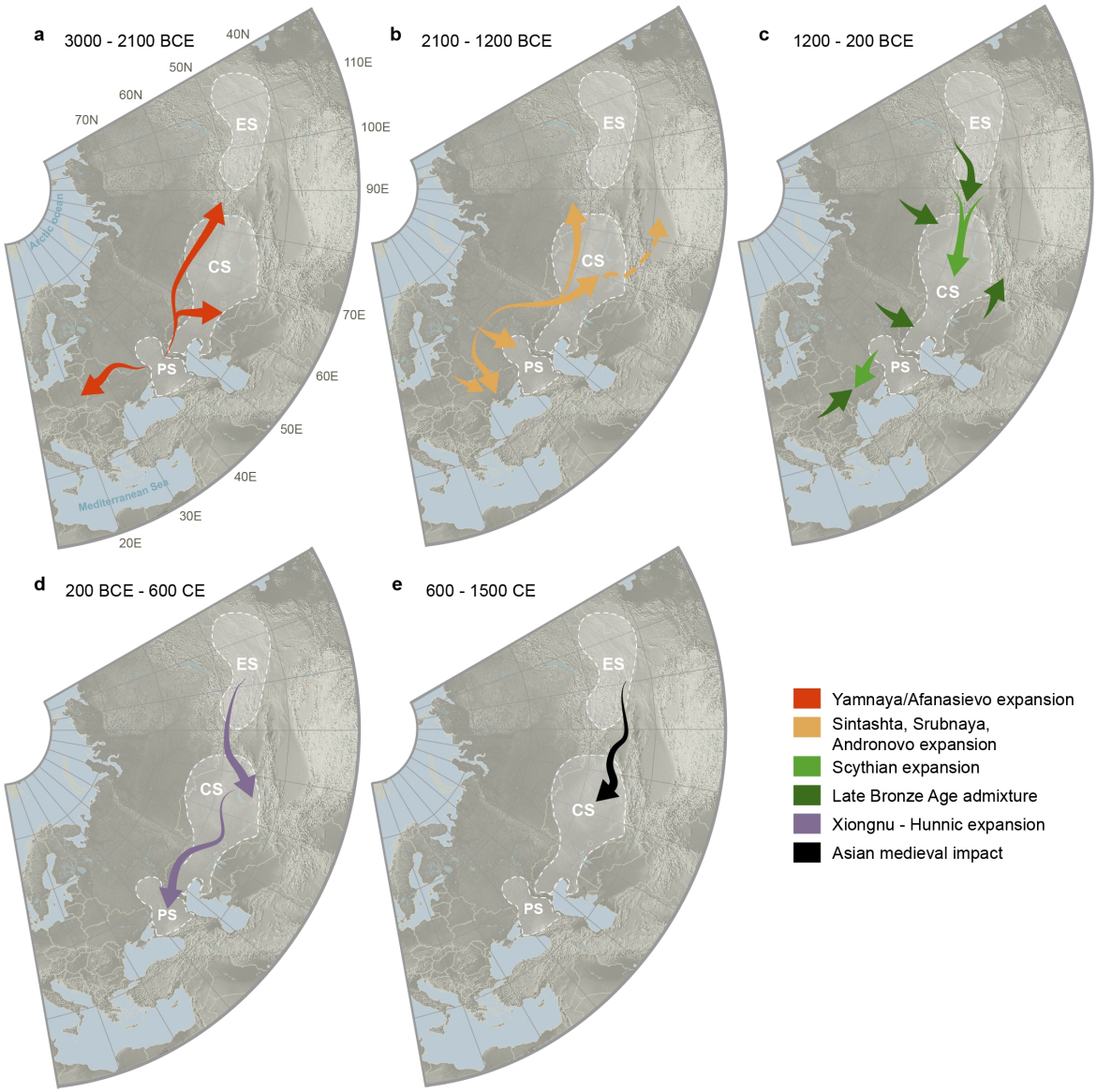

Pontian Steppe
Central Steppe
Tian Shan, Altai & Eastern Steppe
Present-day
Medieval Period
Hun Period
Iron Age
Middle to late Bronze Age
Early Bronze Age
East Asian
Western Hunter Gatherer
Eastern Hunter Gatherer
Natufian
A
Altai Mt.
Tian Shan
Black Sea
50N
40N
60E
70E
80E
90E
Europe
A

Figure 5


a
3000 - 2100 BCE
b
2100 - 1200 BCE
c
1200 - 200 BCE
d
200 BCE - 600 CE
e
600 - 1500 CE
ES
CS
PS
40N
50N
60N
70N
110E
100E
90E
70E
60E
50E
40E
30E
20E
Arctic Ocean
Mediterranean Sea
Yamnaya/Afanasievo expansion
Sintashta, Srubnaya, Andronovo expansion
Scythian expansion
Late Bronze Age admixture
Xiongnu - Hunnic expansion
Asian medieval impact

Supplementary Materials

A population genomic history of the steppe

Contents:

Section 4: Modern dataset

By Peter de Barros Damgaard*, Nina Marchi*, Tatyana Hegay, Choduraa Dorzhu, Mikkel Winther, Hakon Hakonarson, Ludovic Orlando, Rasmus Nielsen, Thorfinn Korneliusen, Martin Sikora, Rana Dajani, Evelyne Heyer", Eske Willerslev"
* contributed equally
" contributed equally and corresponding authors

More than just a large geographical area, the Eurasian steppe is currently inhabited by numerous human populations, belonging to several ethnic groups. These ethnic groups have both historically and currently inhabited distinct geographical regions: some are settled in Central Asia, in the central steppe and Uzbek basins (e.g. Tajik, Turkmen, Uzbek), others are present in Northern Asia, in Siberia, in the Altai mountains, and the Mongolian steppe (e.g. Altai-Kizhi, Mogush, Ondar, Telengit, Tubalars, Buryats, Khakas, Shores or Mongolians), while others reside over the whole Eurasian steppe (e.g. Kyrgyz or Kazakh). In most cases, these ethnic groups are present over the limits of eponymous former USSR republics. Culturally, these ethnic groups belong to two main groups: the Indo-Iranian-speaking sphere, *versus* the Turkic- and the Mongolic-speaking spheres. Besides being contrasted by this linguistic criterion[1], these groups are also associated with two different lifestyles (sedentary *versus* semi-nomadic)[2], different matrimonial systems (respectively, in majority endogamous *versus* exogamous), and subsidence modes (agriculturist *versus* herders or fisher-hunters). Some ethnic groups vary a little from this pattern, such as the Uzbeks who have shifted to a sedentary-agriculturist lifestyle since the sixteenth century[3]. As such, the present-day distribution of mostly sedentary Indo-European speaking groups and mostly nomadic Turkic and Mongolic speaking groups reflects the major changes that the populations of the central steppe area underwent during the transition from Iron Age to Medieval times.
This high level of cultural diversity motivates us to investigate the histories of the living ethnic groups, their origins and past population dynamics. In this study, we newly sampled and genotyped 502 individuals from 16 distinct ethnicities (Extended Table 9) and analyzed them in the context of the ancient genomic dataset, in order to discern patterns of genomic ancestry.

Sampling & Genotyping

During several field expeditions Heyer’s lab and collaborators collected DNA from Central and Northern Asian volunteers who provided informed consent and ethno-linguistical informations. Ethnicities were defined based on self-reported spoken native language. Participants were assigned to populations, defined as groups of individuals living in a similar area and belonging to the same ethnic group. DNA was extracted from blood and saliva, then genotyped on an Illumina genotyping array (either Omni1-Quad, or Omni2.5) by “Institut Pasteur – Genopole (Génotypage des Eucaryotes)”, in Paris, France. Secondly, samples from populations belonging to the Altaian groups Mogush and Ondar were genotyped on the OmniExpress-Exome v1.2 at Aros Applied Biotechnology A/S, Aarhus, Denmark.
Chechen and Circassian diaspora, currently settled in Jordan in the Middle East, were sampled by Rana Dajani's group. All donors provided signed informed consent for the analyses. For these two groups, genomic DNA was isolated from whole blood sample using a phenol-chloroform protocol, and the samples were then SNP genotyped using the GeInfiniumII OMNI-Express BeadChip technology (Illumina), at the Center for Applied Genomics at The Children’s Hospital of Philadelphia (CHOP), USA.
After a Quality-Control process performed independently on each array, data was merged into a single

modern dataset, with all positions flipped to Hg19+. We controlled for call-rates and relatedness by removing individuals with more than 5% missing data, and excluded relatives up to first-degree cousins (included). Relatedness was estimated using all final positions with the software package plink 1.9. The final dataset included 502 individuals. To contextualize this data, we added to the dataset previously published genotypes of populations from Siberia, Central Asia, Caucasus, and Eastern Europe[4] and retained overlapping positions. Thus, the final dataset included 502 new individuals and 612 previously published individuals for a total of 242,406 autosomal SNPs.

Genetic proximities among modern populations from the Eurasian steppe

A Principal Component Analysis (PCA) was performed on the modern dataset alone. As the 171 Chechen and Circassian genotypes are outliers on the second Principal Component driving most of the variance, we remove them from this analysis (Figure 1).
We qualitatively assess four clusters in this PCA (Table 1). In details, the first Principal Component 1, explaining 3.5 % of the variance, corresponds to a gradient from Caucasus (cluster Y) and Europe (cluster X), to East Asia (at the extreme right of cluster W). Next, on the second Principal Component, that represents 0.5% of the variance, we distinguish European from Caucasian populations (X and Y clusters respectively). This PC also separates the Inner Asian populations in clusters W and Z.
Indeed, three main pools emerge for Inner Asia on the plot: first, cluster Y, the “Indo-Iranian” one, including Yagnobis, Tajiks from Pamir and from Uzbekistan, that are genetically close to populations from Caucasus, such as Azeris and Kumyks.
The second pool (cluster W) englobes variability amongst Turkic and Mongolic-speaking groups (notably Kyrgyz, Telengits, Kazakhs, Mongols, Buryats) from Northern Asia, clustering with populations from East Asia (Even, Evenks, Chukchis).
The third pool (cluster Z) includes some Siberian populations (namely the Shores, Tubalars, Telengits and Khakas), that are traditionally fisher-hunters, and who cluster with the Nenets and Kets, reindeer herders from the Russian Artic Circle. We can note that, despite close neighbouring, Siberian Turkic groups are genetically heterogeneous, being in both cluster W and Z.
Three Turkic-speaking populations from Central Asia are included to this study: the Uzbeks, the Karakalpaks and the Turkmens. While most of the Turkmen samples are included in cluster Y, the other Turkmens, the Uzbeks and the Karakalpaks are found between cluster Y and W, indicating that they are genetically related to both Indo-Iranian and Turkic-speaking populations of Northern Asia. This intermediate genetic component has already been documented for uniparental markers[5] and autosomal STRs[6]; and these populations are documented to have experienced interesting cultural histories: both Turkmens and Uzbeks likely descend from Indo-Iranian speaking tribes assimilated by Turkic-speaking invaders.

Cluster X	Cluster Y	Cluster Z	Cluster W	Unclustered
Vepsas	Tajiks_Pamir	Shores	AltaiKizhis	Uzbeks
Karelians	Tajiks	Tubalars	Telengits	Nganassan
Russians	Kumyks	Kets	Kyrgyz	Gagauzes
Germans	Azeris_Daghestan	Nenets	Ondar	Bashkirs
Komis	Yagnobi	Khakas	Buryats	Chukchis
Udmurts	Kabardins	Tuvans	Mongolians	Koryaks
Chuvashes	Azeris_Iran		Kazakhs	Karakalpaks
Tatars	Turkmen		Even	
Yakuts	Balkars		Evenks	
Uygur			Kalmyks	
			Mongush	
			Altaians	

Table 1: Qualitative assessment of four clusters in the Principal Component Analysis. In blue the European populations, in orange the Caucasians, in black the Inner Asians, in purple the Asians from outside Inner Asia (Northern Siberia and East Asia).

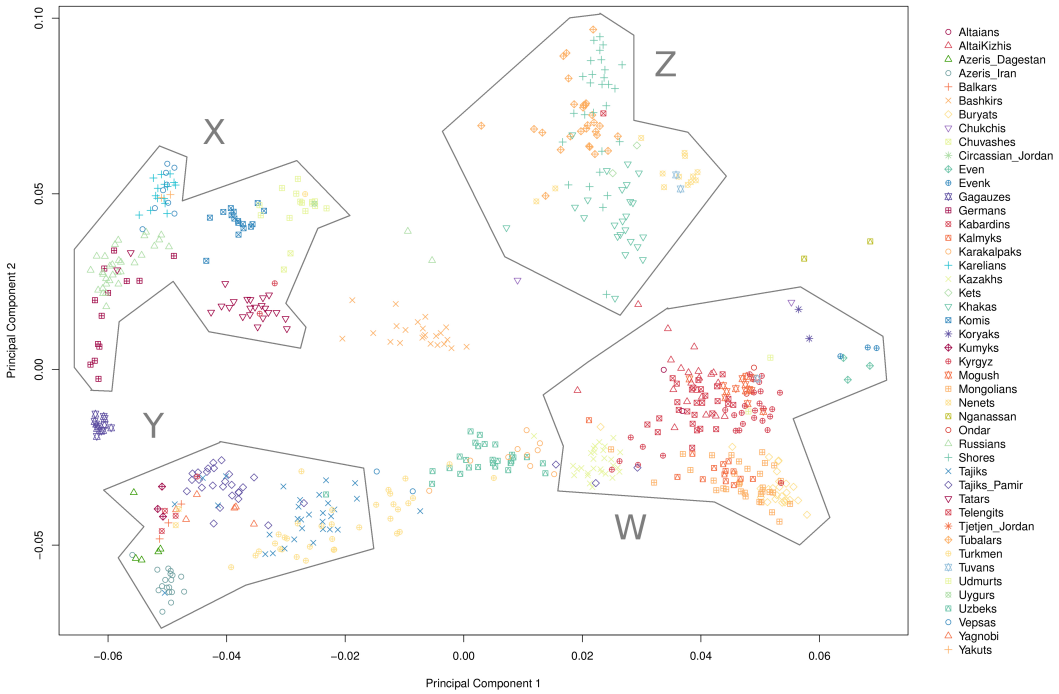


Figure 1: Principal Component Analysis of the modern dataset, excluding Chechens and Circassians

Interestingly, we discern a pattern of East Eurasian ancestry (Baikal Hunter-Gather like, from Glazkovo culture) driving PC1 and, Ancient North Eurasian MA1-like ancestry driving PC2. In order to confirm and illustrate that these ancestries are driving the clines representing the two Principal Components, we then calculate D-statistics for each populations of the clusters individually of the following forms: 1) for illustrating gradient of Principal Component 1: D(BHG, Mbuti; Cluster Z, Cluster X); D(BHG, Mbuti; Cluster W, Cluster Z); 2) for illustrating gradient of Principal Component 2: D(MA1, Mbuti; Cluster X, Cluster Y), D(MA1, Mbuti; Cluster Z, Cluster W). All of these D-statistics are highly significant towards the hypothesized value, i.e., reflecting increased shared ancestry between Baikal Hunter-Gatherers and cluster Z and W compared to X and Y respectively, and increased shared ancestry between MA1 and clusters X and Z compared to Y and W respectively (see Extended Table 9). Interestingly, in this dataset, we include three European populations Vepsas, Udmurts and Komis of Finno-Ugrian origin, and representing a recently described pole in the modern European genetic diversity[7]. These groups Vepsas and Komis are at the MA1-rich extremity of PC2 indicating some MA1-like gene flow coming in through the arctic corridor of Siberia, independent of Bronze Age steppe impact in Europe, and therefore a population genomic history to be explored in future studies to remedy the absence of relevant ancient genomes.

Formation of the Central Asian gene pool

We identified the ancestral representatives composing the extremities of the first two dimensions on the Principal Component Analysis with the ancient samples (main Figure 2). They are samples from previously published data (Natufian from the Near East[11], and Eastern Hunter-Gatherers, EHG, from Eastern Europe[9,10]) and new East Asian hunter-gatherer sequences obtained in this study (Lake Baikal Hunter-Gatherers, BHG). We find that almost all the steppe populations, across both time and space, can be modelled using qpAdm[9] as a mixture of these three ancestral hunter-gatherer groups, while some require additional specific. In particular, broadly across West Eurasia, an additional Mesolithic hunter-gatherer source was required (Western Hunter-Gatherers, WHG, from Europe[12–14]), and finally in the Far-East, Chukchis required Native American ancestries (here Clovis, a Paleoamerican [15]).
In total, we trace back the genetic history of modern populations by exploring their ancestries relatively to these five hunter-gatherer populations. We infer the proportion of each ancestry with the qpAdm approach[9] (Extended Table 9). Our results show an increase in Baikal HG ancestry eastwards (on average, 67% in Inner Asia, against 20% in Caucasus and 27% in Europe; Spearman's correlation between longitude and Baikal HG ancestry part: *rho*=0.82, p-value=$1*10^{-15}$), while westwards the Natufian and EHG ancestries increase (in total, 80% in Caucasus; 55% in Europe; 35% in Asia; Spearman's correlation between longitude and respective ancestry part: *rho*=-0.82 and -0.54, p-values=$4*10^{-12}$ and $2*10^{-4}$). This confirms two obvious trends: East Asian ancestries increase at the expense of Natufian and EHG ancestries along an East/West geographical cline (Figure 2).

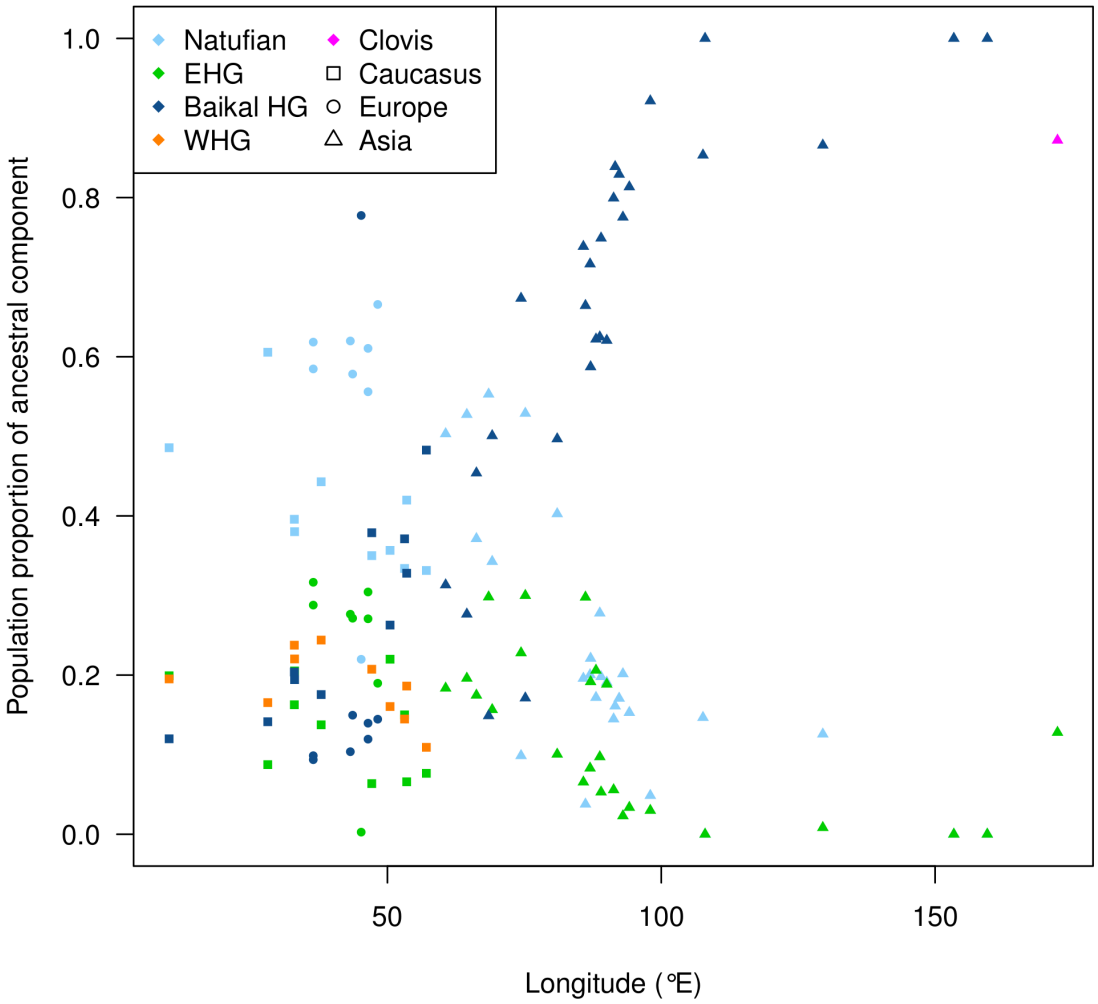


Figure 3: Hunter-Gatherer ancestries for each population correlated with its longitudinal coordinate.

European and Caucasian modern populations are differentiated by WHG ancestry in modern European populations (19%), and an excess of Natufian in Caucasians (56% against 41% in Europeans). We also correlated the Natufian ancestry with latitude, revealing a negative association (Spearman's correlation between latitude and Natufian ancestry part: *rho*=-0.66, p-value=$7*10^{-7}$), coupled to the requirement of WHG ancestry (Figure 3).

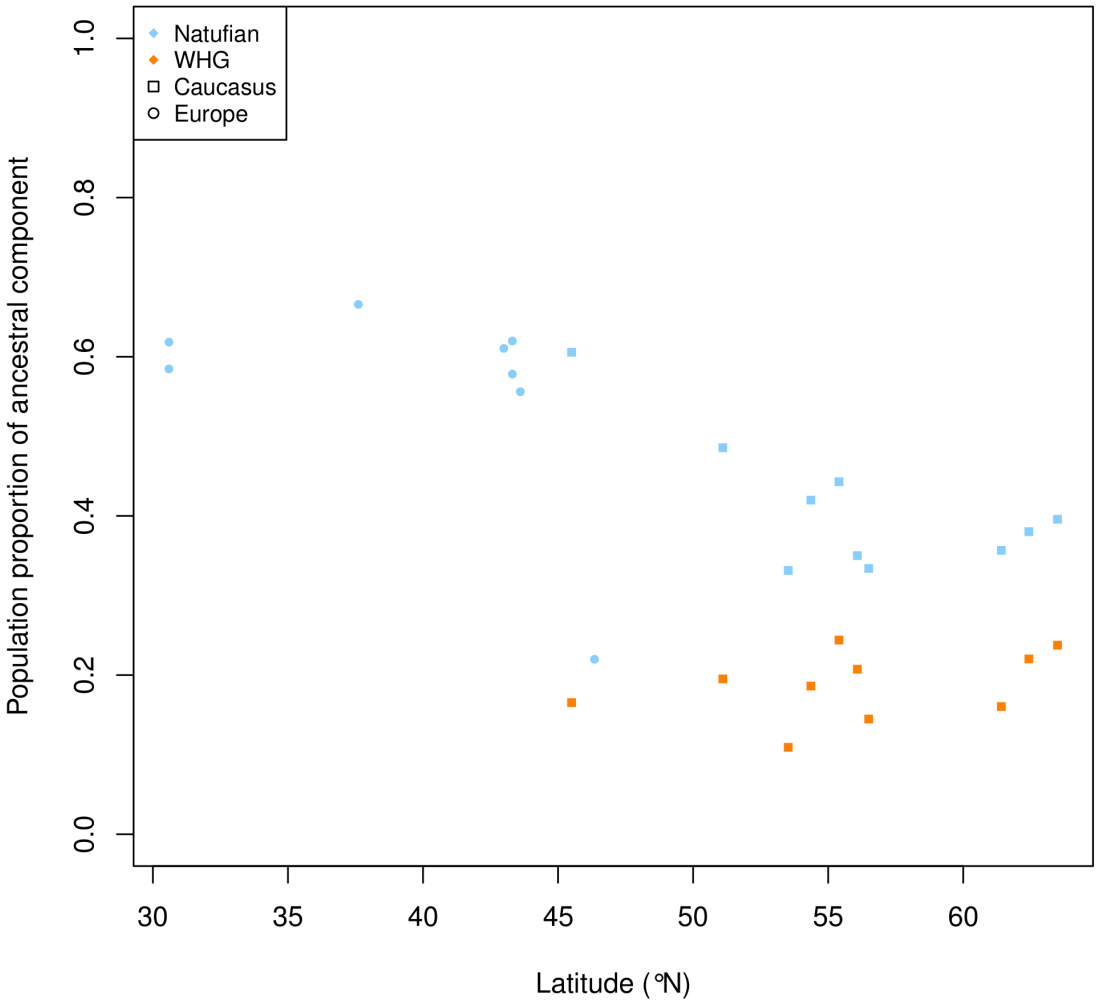


Figure 3: Natufian and WHG ancestral proportions in Caucasian and European populations according to latitude.

All the results are illustrated in Figure 4. Importantly, we note that the 'Native American' ancestry in Chuckchis is highly inflated (87%) as Clovis ancestry then replaces East Asian ancestry in the model.

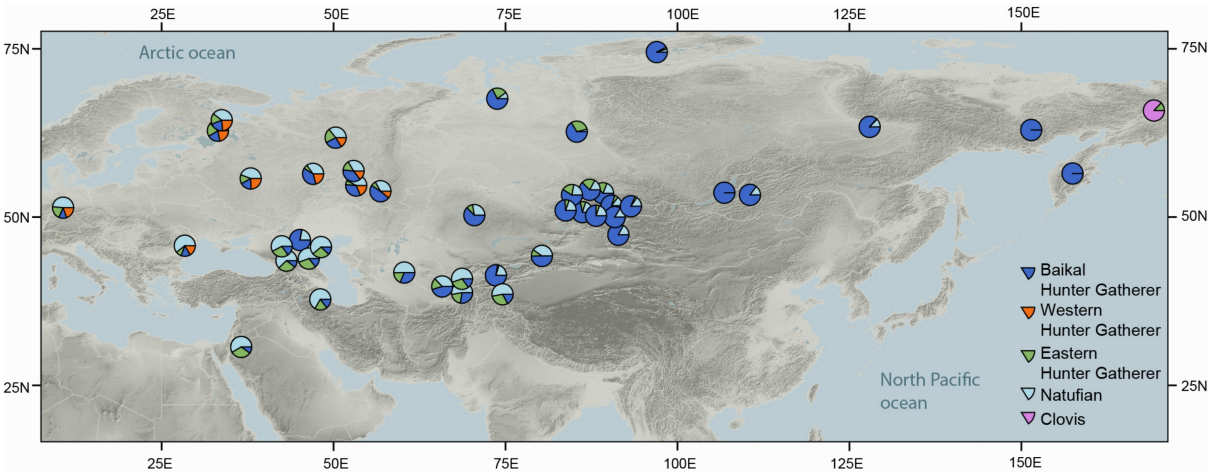
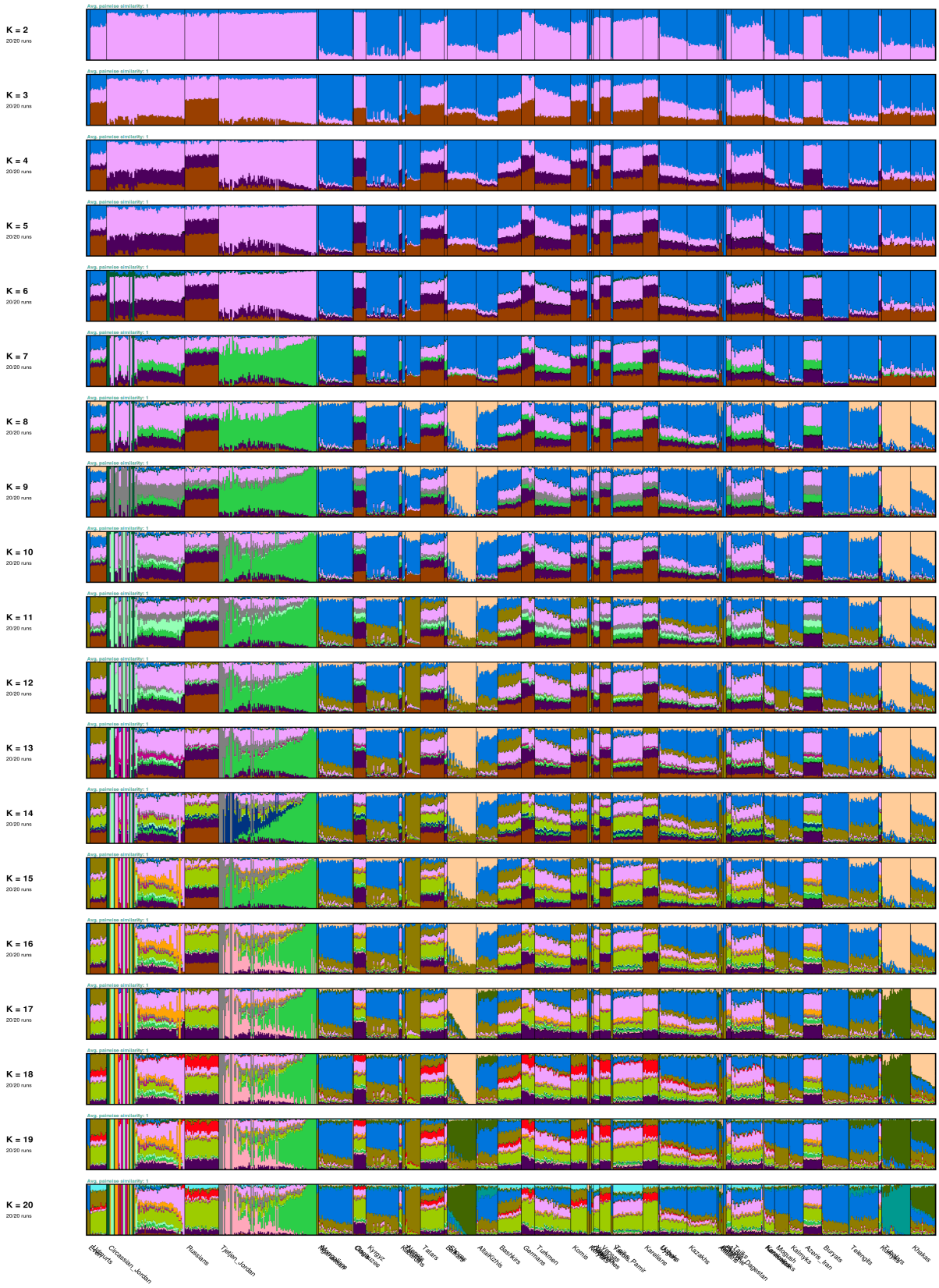
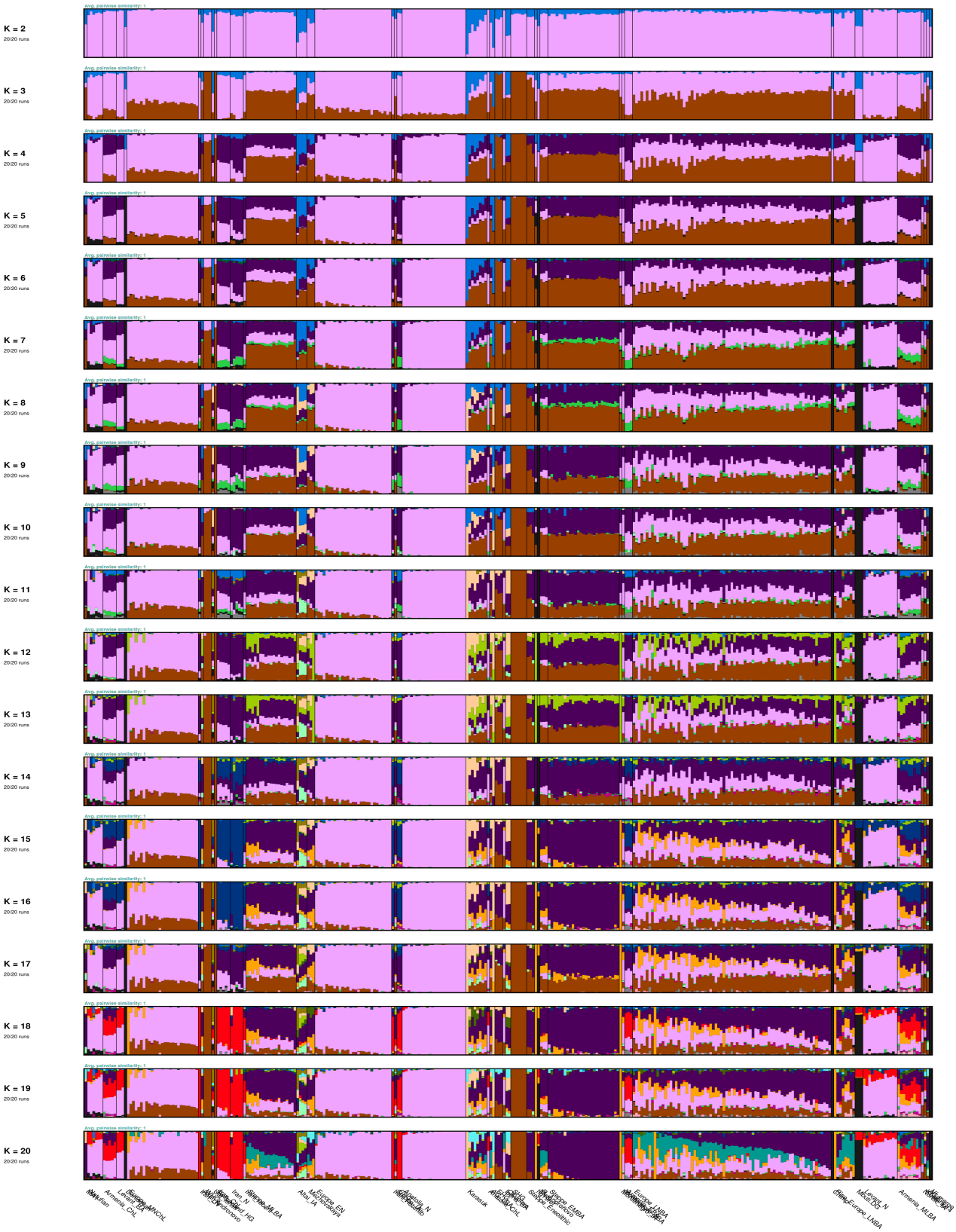
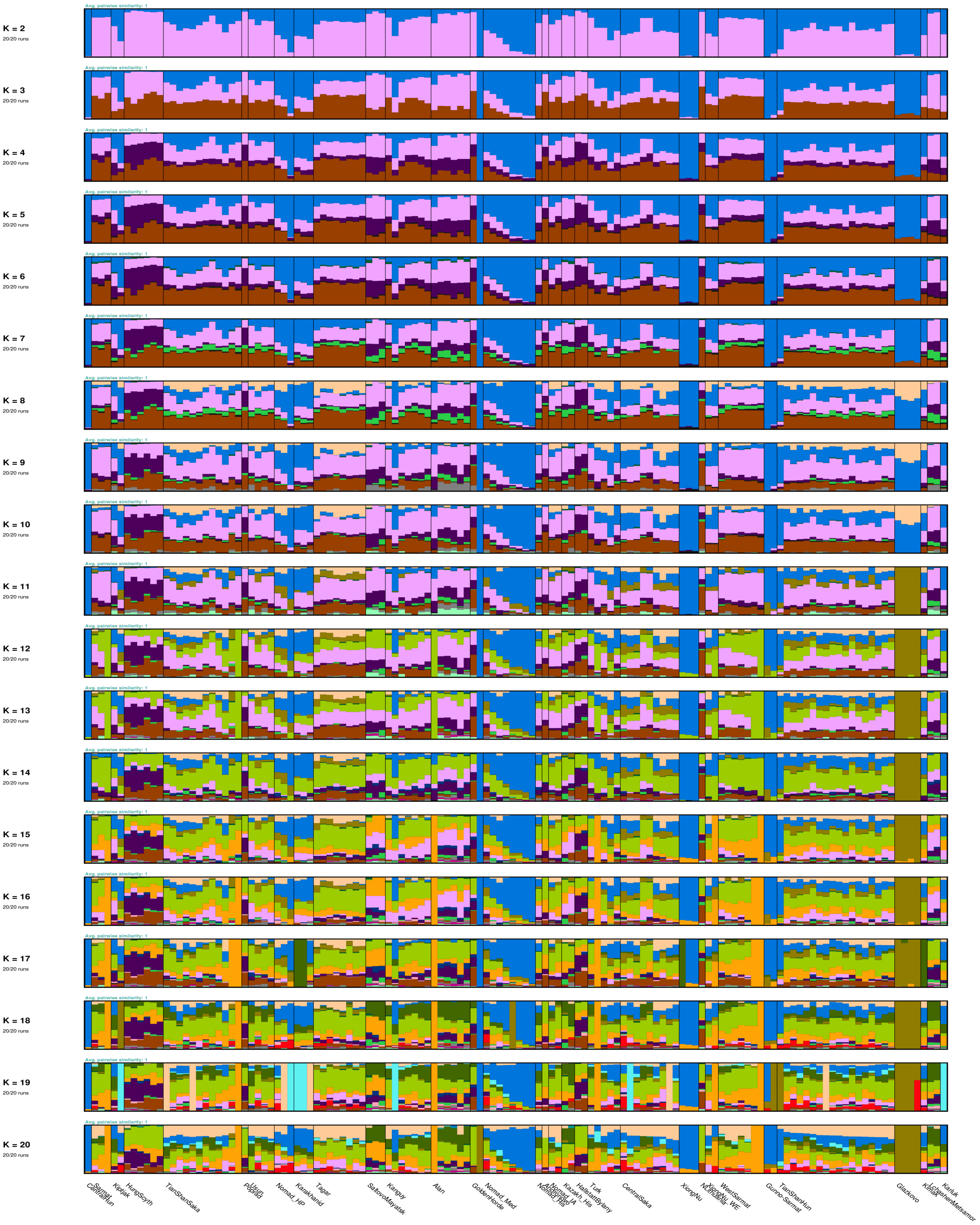


Figure 5: A depiction of hunter-gatherer ancestries in present-day populations. The East/West gradient of Baikal and Eastern Hunter-Gatherer ancestry is evident (blue) while the distribution of Natufian ancestry is notable in the Near East (light blue), the southern fringes of the steppe and Europe where Western Hunter-Gatherer ancestry is evident (orange). The Clovis ancestry (pink) is only required for Chuckchis from the extreme east of Asia in these models.

Finally, we computed ADMIXTURE runs with K=2 to K=20 (Figure 5-7) in a total of 20 replicates per run, with the full dataset containing ancient and modern genomes. The results are visualized in three plots: “Previously published ancient genomes”, “Modern SNP data”, ”New ancient genomes”.


K = 2
20/20 runs
K = 3
20/20 runs
K = 4
20/20 runs
K = 5
20/20 runs
K = 6
20/20 runs
K = 7
20/20 runs
K = 8
20/20 runs
K = 9
20/20 runs
K = 10
20/20 runs
K = 11
20/20 runs
K = 12
20/20 runs
K = 13
20/20 runs
K = 14
20/20 runs
K = 15
20/20 runs
K = 16
20/20 runs
K = 17
20/20 runs
K = 18
20/20 runs
K = 19
20/20 runs
K = 20
20/20 runs
Circassian_Jordan
Russians
Tajiks_Jordan
Kyrgyz
Tatars
Altaians
Bashkirs
Germans
Turkmen
Komis
Karelians
Kazakhs
Mansis
Kalmyks
Azeris_Iran
Buryats
Tuvinians
Khakas


K = 2
20/20 runs
K = 3
20/20 runs
K = 4
20/20 runs
K = 5
20/20 runs
K = 6
20/20 runs
K = 7
20/20 runs
K = 8
20/20 runs
K = 9
20/20 runs
K = 10
20/20 runs
K = 11
20/20 runs
K = 12
20/20 runs
K = 13
20/20 runs
K = 14
20/20 runs
K = 15
20/20 runs
K = 16
20/20 runs
K = 17
20/20 runs
K = 18
20/20 runs
K = 19
20/20 runs
K = 20
20/20 runs


K = 2
K = 3
K = 4
K = 5
K = 6
K = 7
K = 8
K = 9
K = 10
K = 11
K = 12
K = 13
K = 14
K = 15
K = 16
K = 17
K = 18
K = 19
K = 20
20/20 runs
Avg. pairwise similarity: 1
CentralHun
Sarmat
Kiptjak
HungScyth
TianShanSaka
Poprad
Usun
Nomad_HP
Karakhanid
Tagar
SaltovoMayatsk
Kanguy
Alan
GoldenHorde
Nomad_Med
Nomad_His
Andronovo
Nomad_IA
Kazakh_His
HallstattBylany
Turk
CentralSaka
XiongNu
NLithuania
XiongNu_WE
WestSarmat
Gunno-Sarmat
TianShanHun
Glazkovo
Kimak
LchashenMetsamor
Karluk

Without surprise, these analyses revealed gradients of ancestry across the Eurasian steppe, explained by archaeological and proto-historical genomes. The historical transect (Figure 4 from main text) portrays the genetic composition of more recent ancient cultures (e.g. Scythians, Sarmatians, Xiongnu, Turkic-speaking Khanates) revealing that these groups can, just like the modern populations, all be modeled as mixtures of the same five Eurasian hunter-gatherer groups. This illustrates that the vast majority of present-day variation is explained by i) genetic drift occurred in a period of human history characterized by smaller hunter-gatherer populations, and ii) by admixture between groups with varying proportions of these hunter-gatherer ancestries during the last 5,000 years. Thus, the genomic transect of the last 5,000 years of steppe nomads depict a complex history of gradual admixture, in particular between groups with increased genomic ancestries traced back to East Asian hunter-gatherers (here represented by 4 individuals from the early Bronze Age Lake Baikal).
In this study, we pictured the genetic landscape of the whole Eurasian steppe, for a considerable dataset of SNPs. We highlighted genetic differences between geographical and cultural groups, such as Eastern Europe, Caucasus, East and Central Asia, and Siberia. Finally, we found that these differences could largely be explained by increased Near Eastern ancestries in populations living south of the steppe belt, and increased East Asian ancestries in central and eastern steppe nomads, as well as particular Altaian groups.

References

1. d'Errico, F. & Hombert, J.-M. *Becoming eloquent: advances in the emergence of language, human cognition, and modern cultures*. (John Benjamins Publishing, 2009).
2. Krader, L. *Peoples of Central Asia*. **26,** (Indiana university, 1971).
3. Bregel, Y. An historical atlas of Central Asia. *J. Asiat.* **291,** 295–300 (2003).
4. Yunusbayev, B. *et al.* The genetic legacy of the expansion of Turkic-speaking nomads across Eurasia. *PLoS Genet* **11,** e1005068 (2015).
5. Marchi, N. *et al.* Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations. *Am. J. Phys. Anthropol.* **162,** 627–640 (2017).
6. Martínez-Cruz, B. *et al.* In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations. *Eur. J. Hum. Genet.* **19,** 216–223 (2011).
7. Khrunin, A. V. *et al.* A genome-wide analysis of populations from European Russia reveals a new pole of genetic diversity in northern Europe. *PloS One* **8,** e58552 (2013).
8. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522,** 167–172 (2015).
9. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522,** 207–211 (2015).
10. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528,** 499–503 (2015).
11. Lazaridis, I. *et al.* Genomic insights into the origin of farming in the ancient Near East. *Nature* **536,** 419–424 (2016).
12. Olalde, I. *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* **507,** 225–228 (2014).
13. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513,** 409–413 (2014).
14. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5,** (2014).
15. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506,** 225–229 (2014).

D Annexes du Chapitre II

Haplogroupes

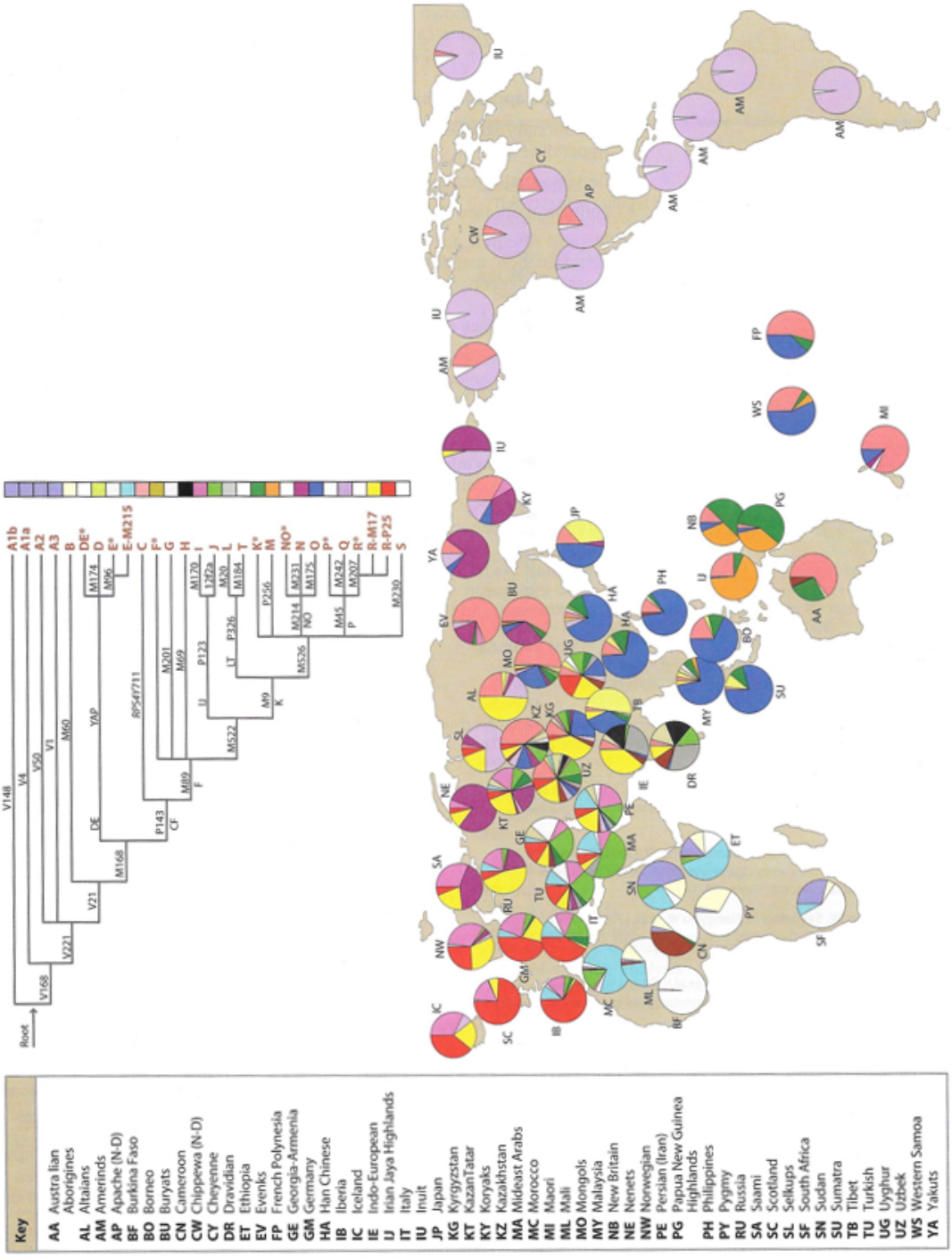


Figure 31 – Phylogénie des haplogroupes du chromosome Y. *Appendix Figure 3 du livre Jobling* et al. *(2013).*

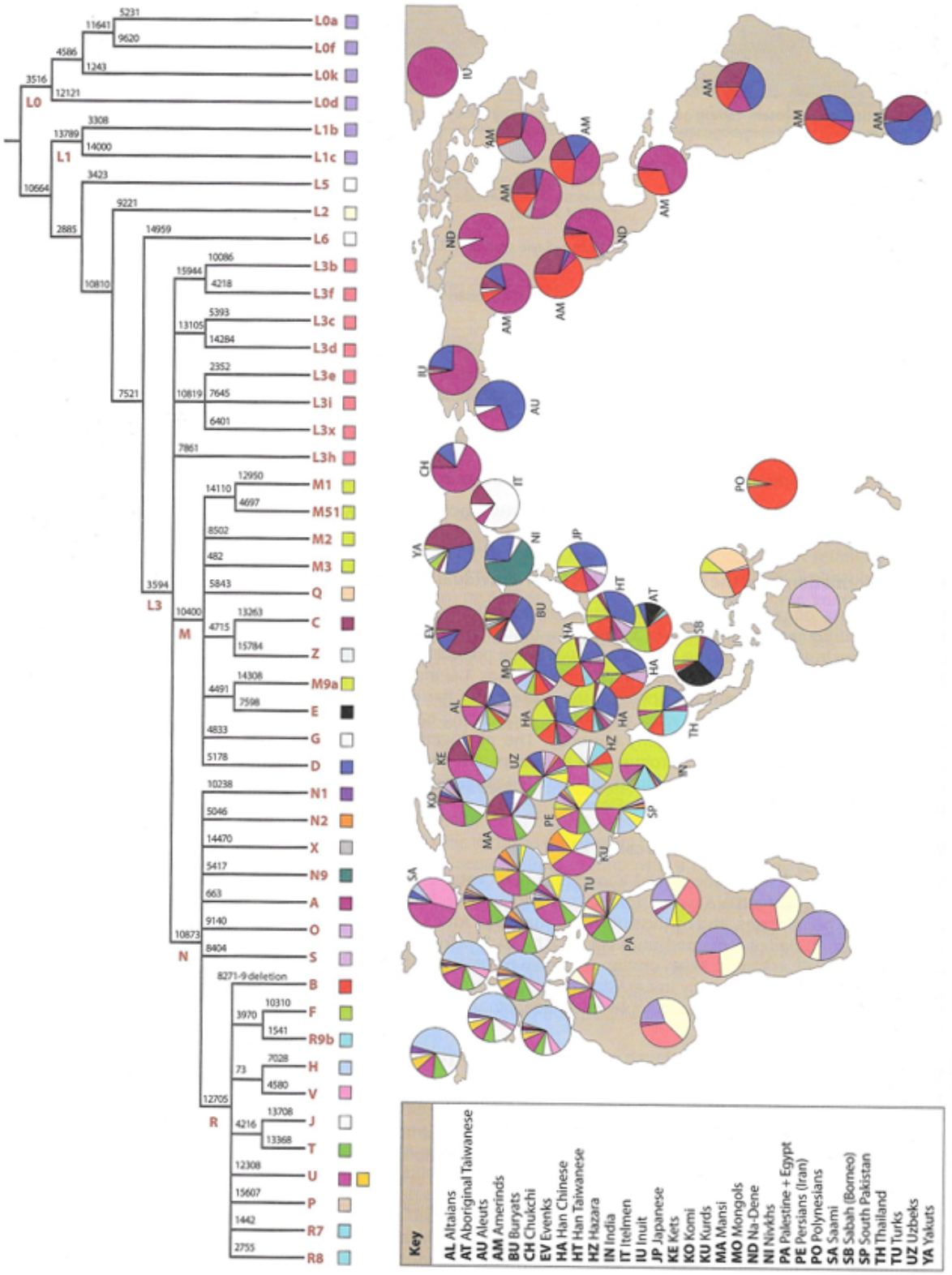


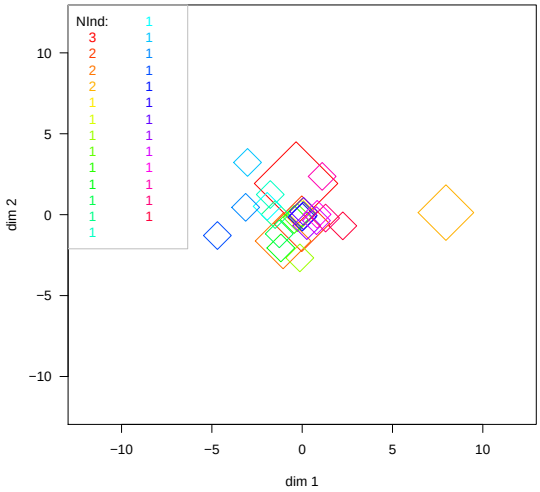
Figure 32 – Phylogénie des haplogroupes mitochondriaux. *Appendix Figure 2 du livre Jobling* et al. *(2013).*

Noyaux d'identité du chromosome Y

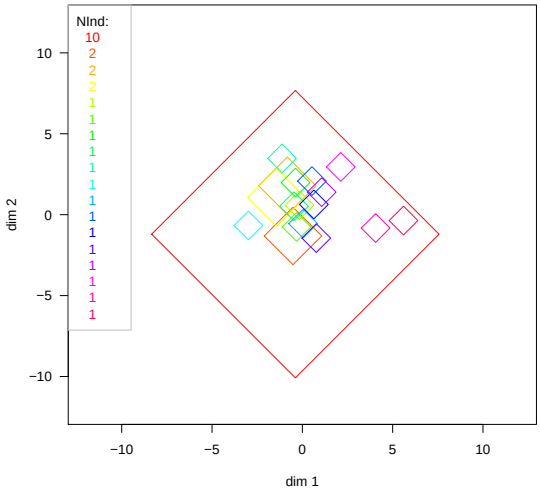
Dans ces MDS proportionnelles, chaque losange représente un haplotype et sa taille est proportionnelle au nombre d'individus porteurs de cet haplotype indiqué en légende.

Pop. (Nb haplotypes différents/Effectif). TAJIK:

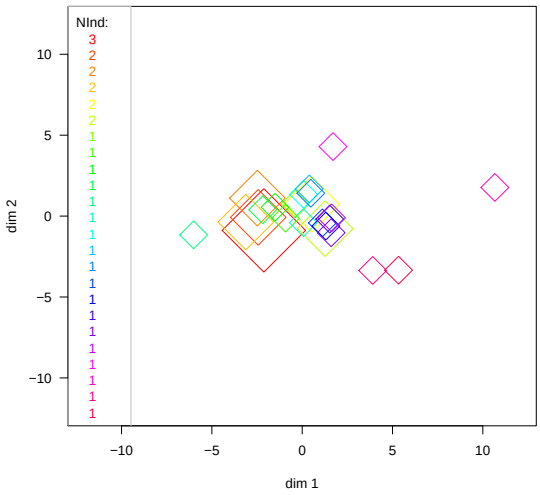
Luz (26/31)



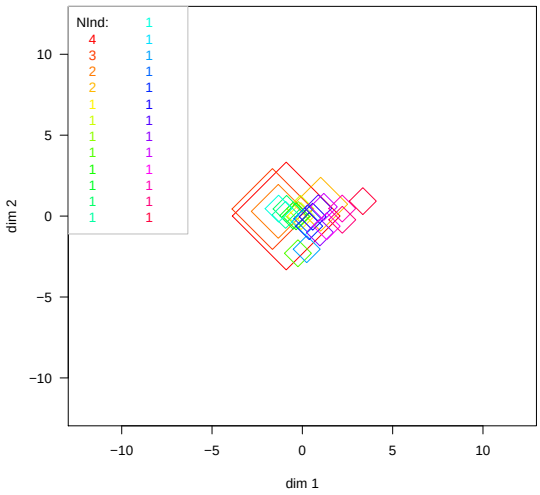
Tds (18/30)



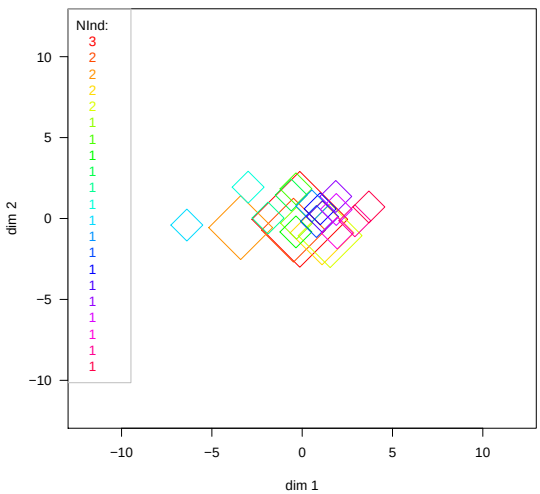
Tdu (24/31)



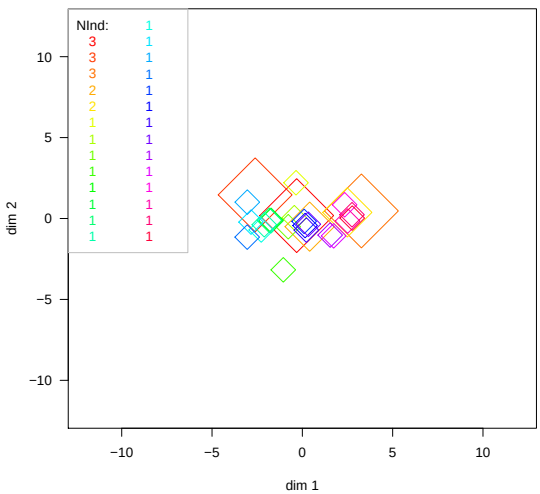
Tja (25/32)



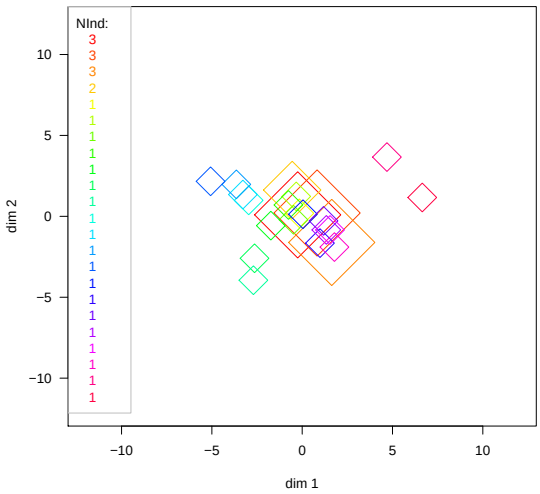
Tje (21/27)



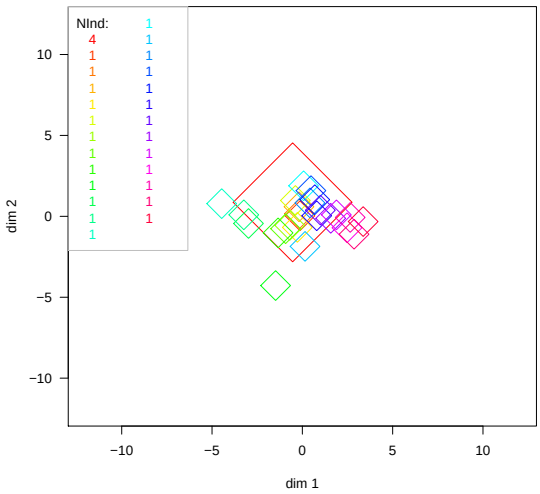
Tjk (27/35)



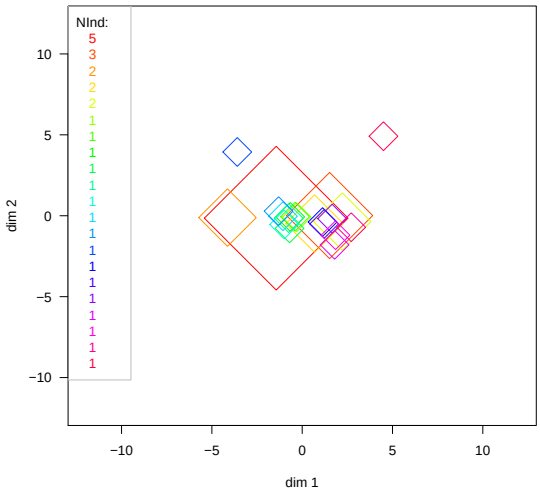
Tjn (23/30)



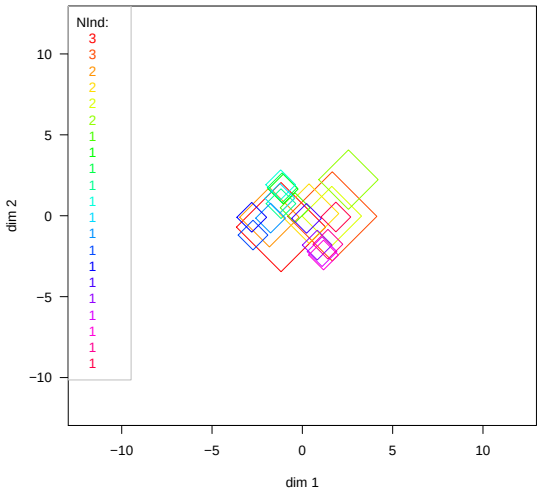
Tjr (26/29)



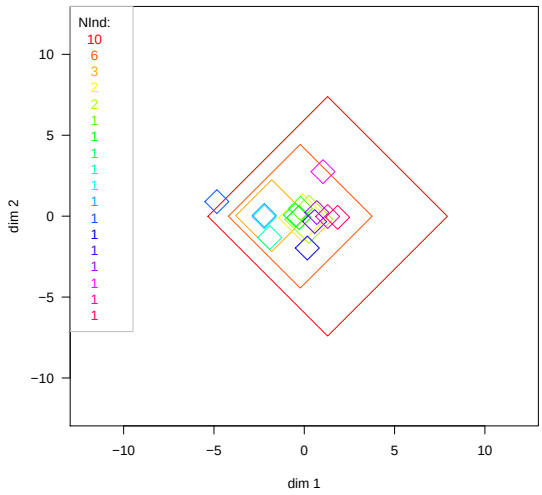
Tjt (21/30)



Tju (21/29)

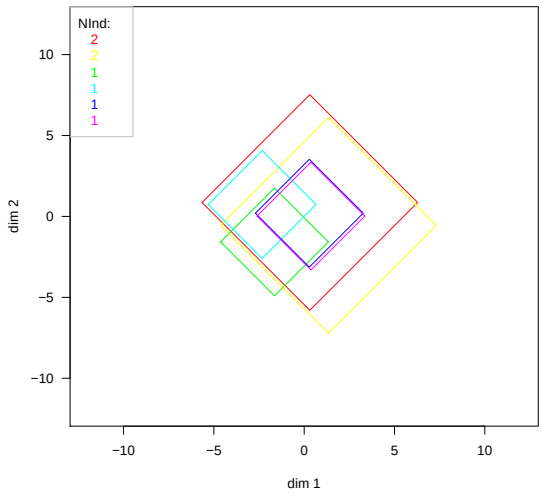


Tjy (18/36)

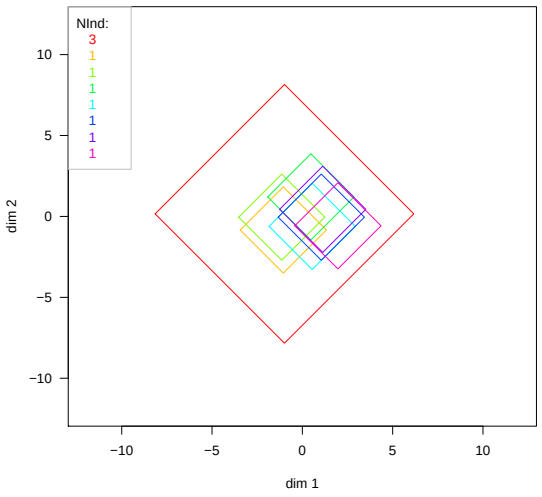


Kyrgyz:

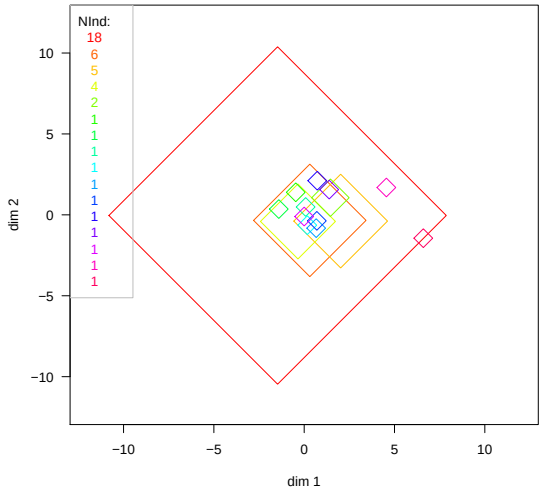
Kek (6/8)



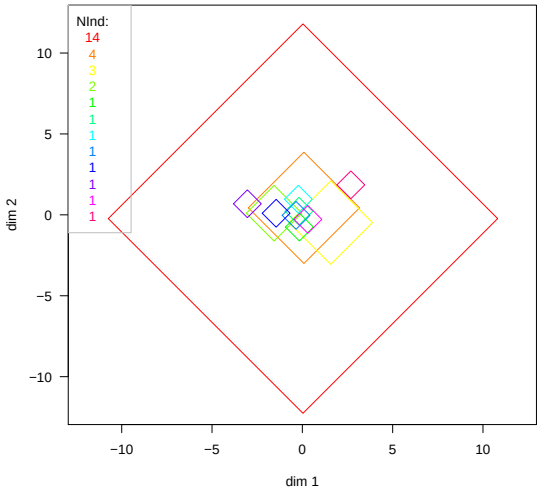
Kem (8/10)



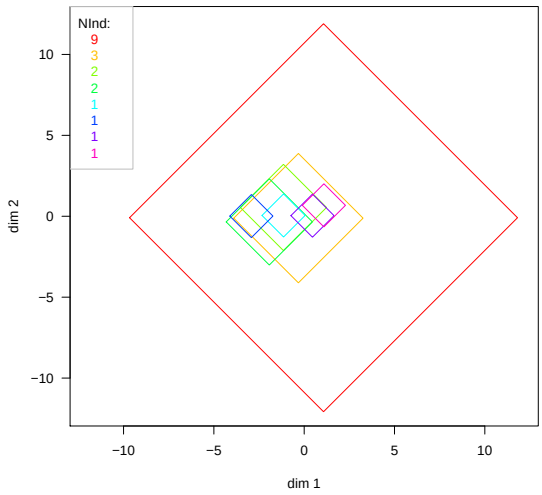
Kra (16/46)



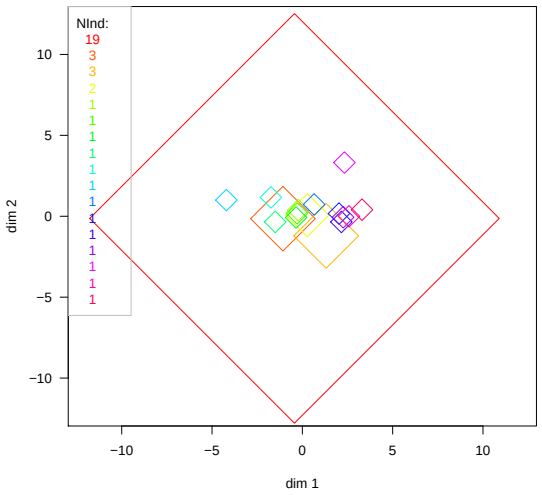
Krb (12/31)

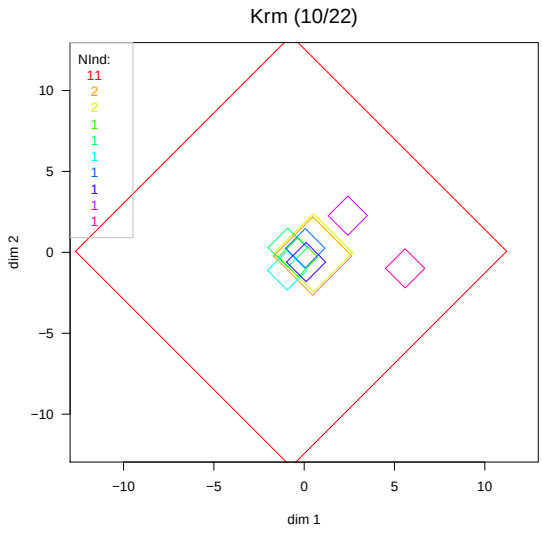


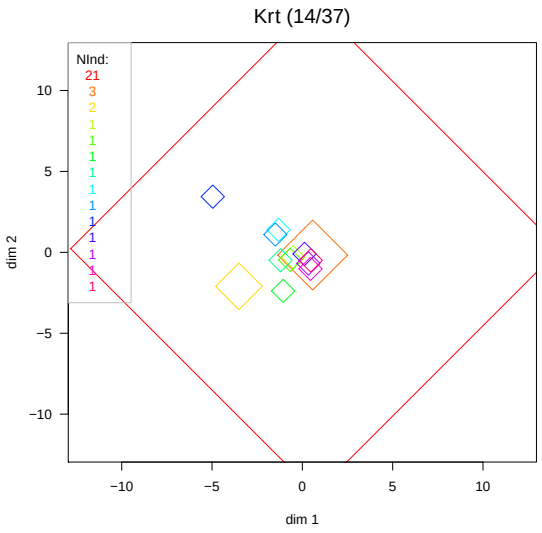
Krg (8/20)



Krl (17/40)

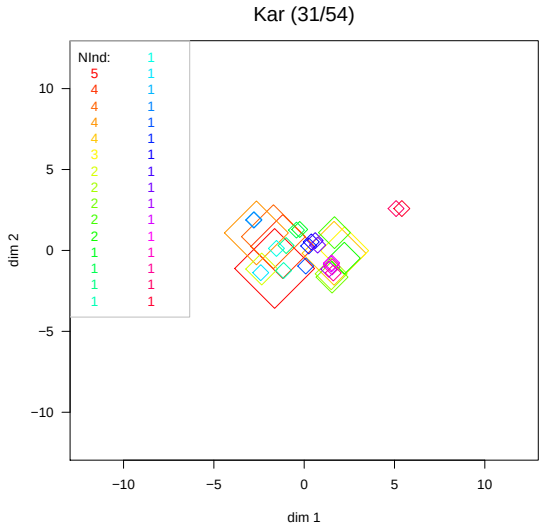


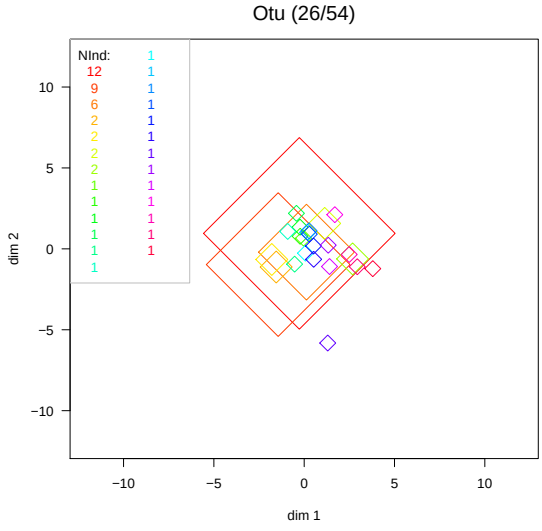


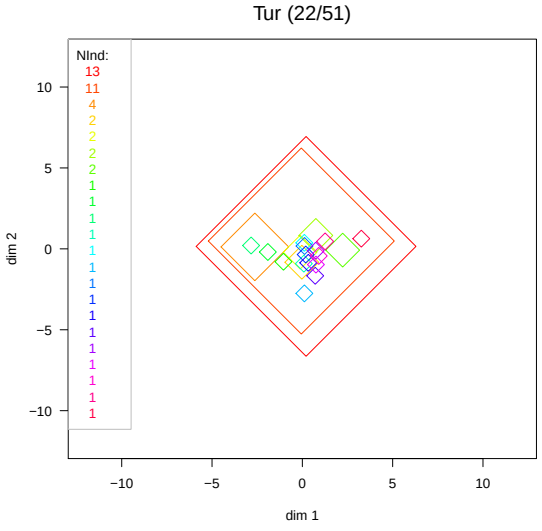


KARAKALPAK:

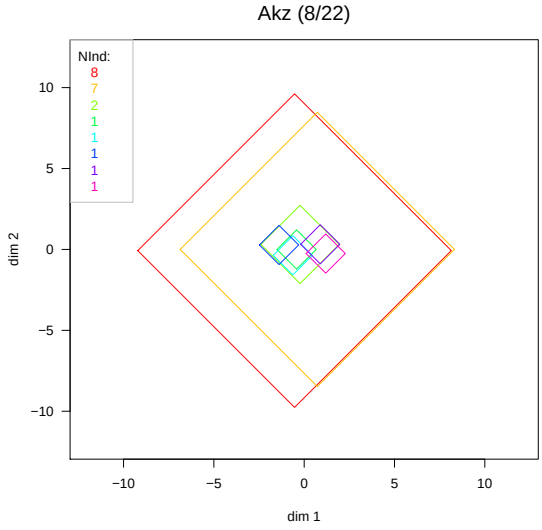
TURKMEN:

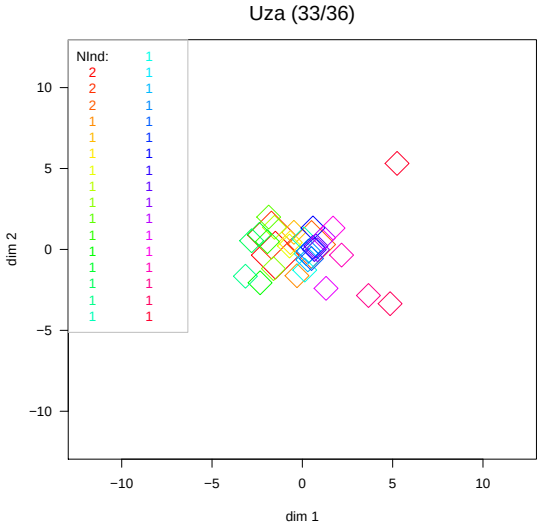


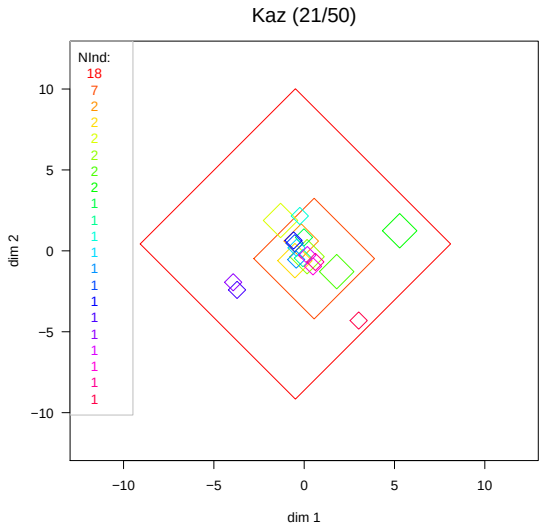


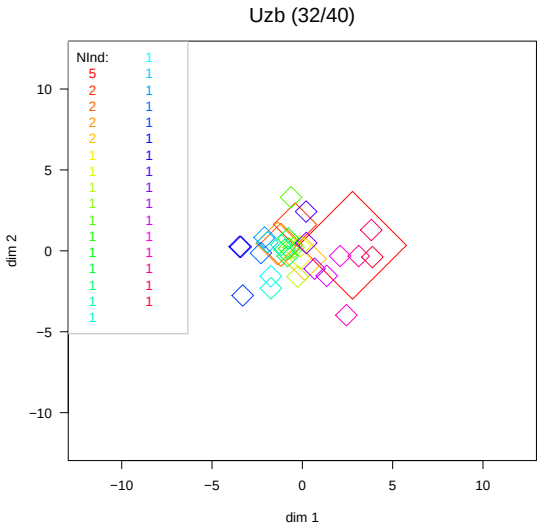


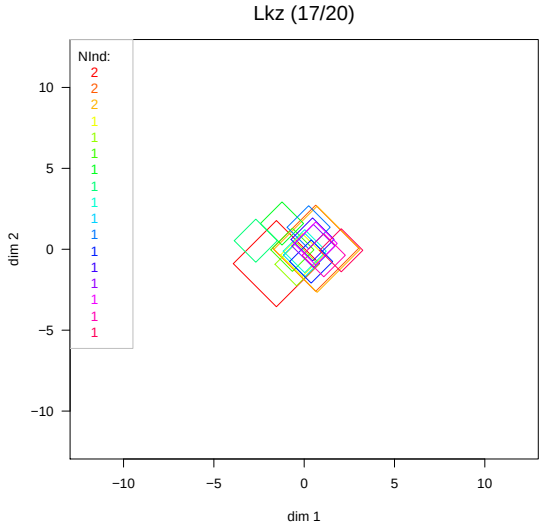
KAZAKH: UZBEK:

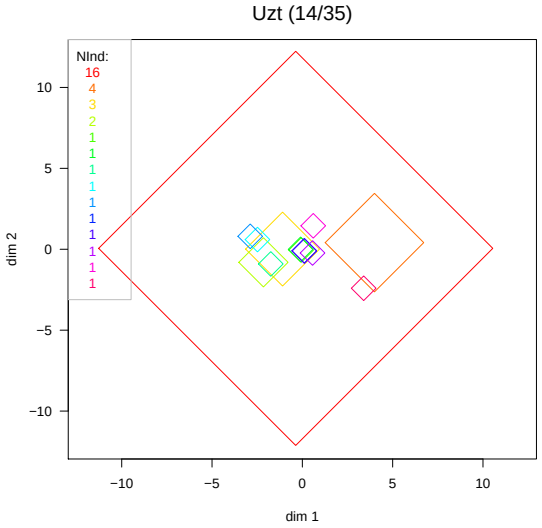






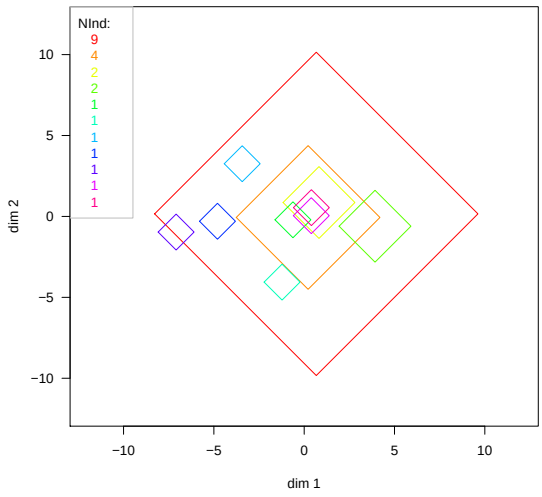




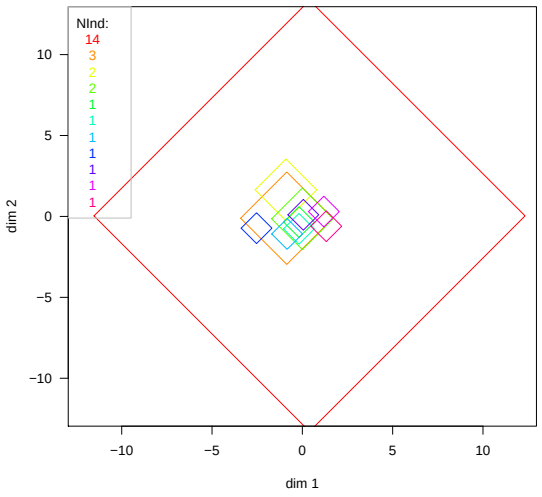


Northern Asia:

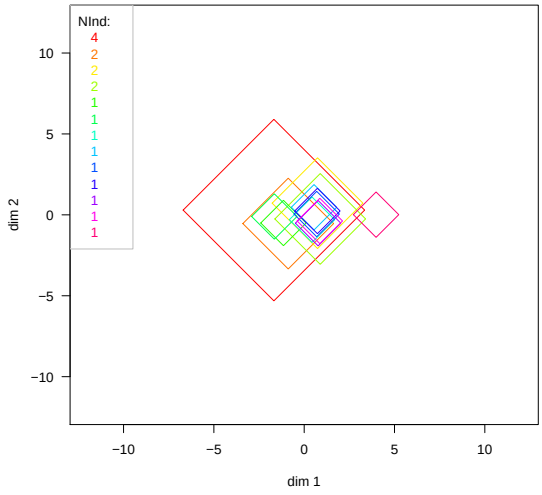
Aki (11/24)



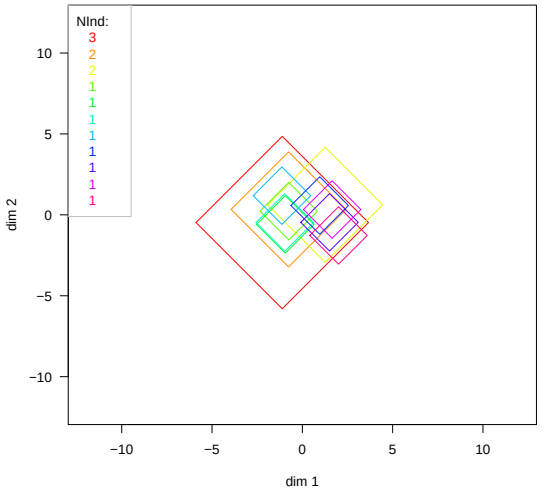
Bou (11/28)



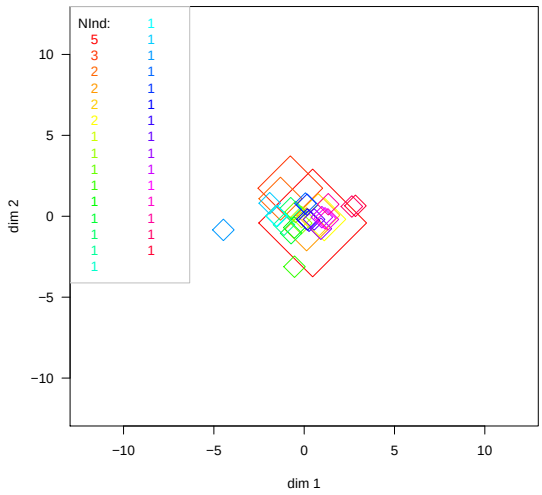
Hks (13/19)



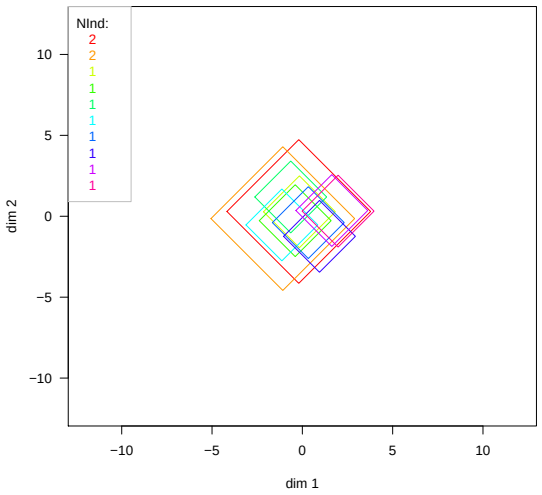
Irg (11/15)

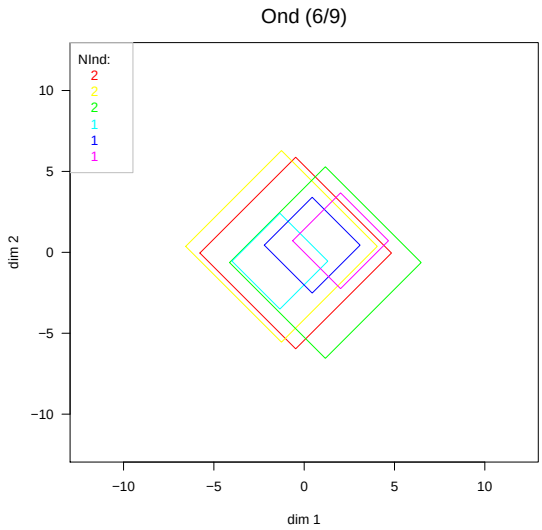


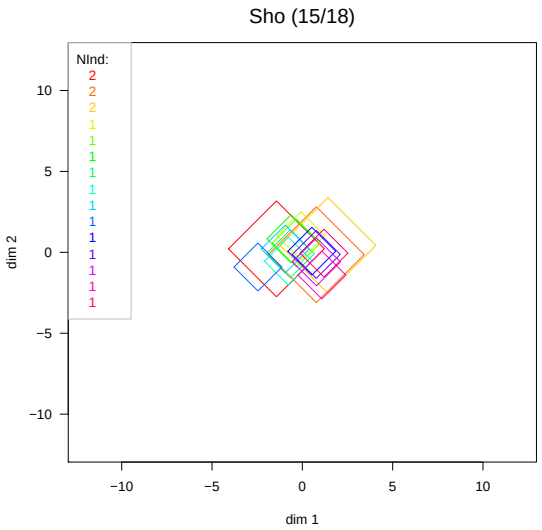
Mng (30/40)

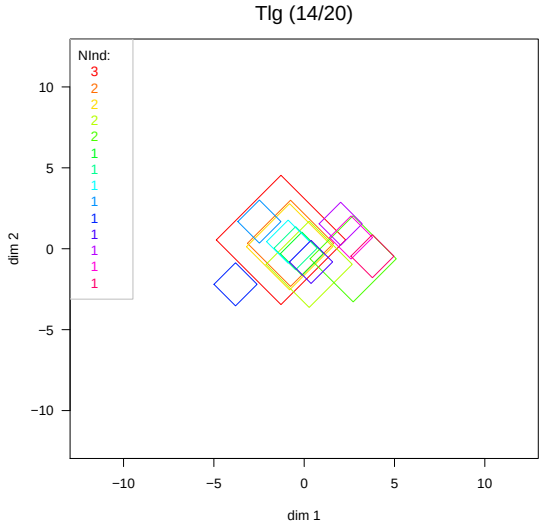


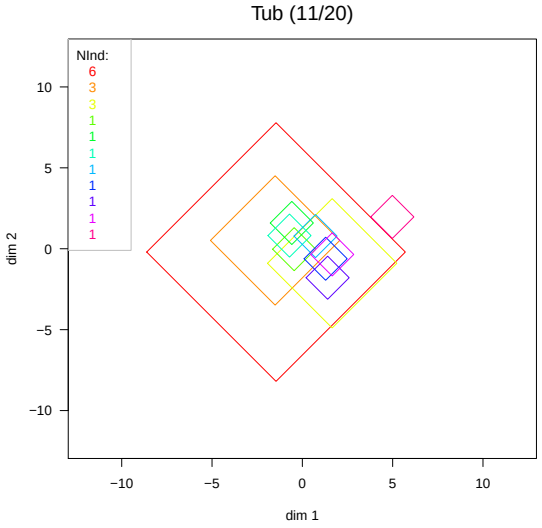
Mog (10/12)



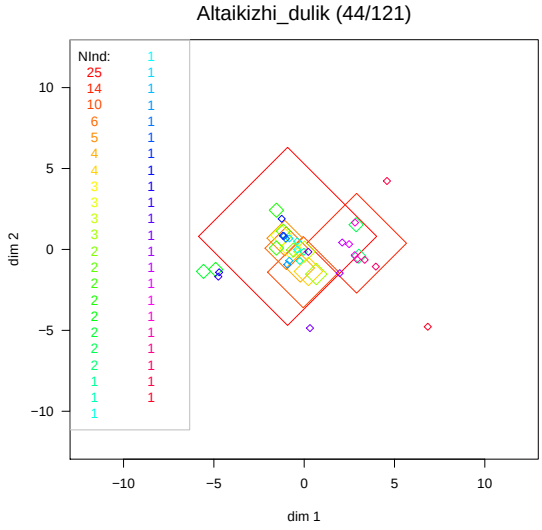

Ond (6/9)
NInd:
2
2
2
1
1
1
dim 2
dim 1
10
5
0
−5
−10
−10
−5
0
5
10

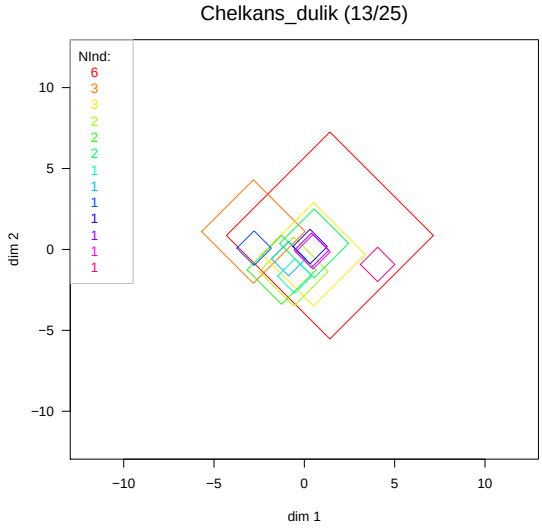

Sho (15/18)
NInd:
2
2
2
1
1
1
1
1
1
1
1
1
1
1
dim 2
dim 1
10
5
0
−5
−10
−10
−5
0
5
10

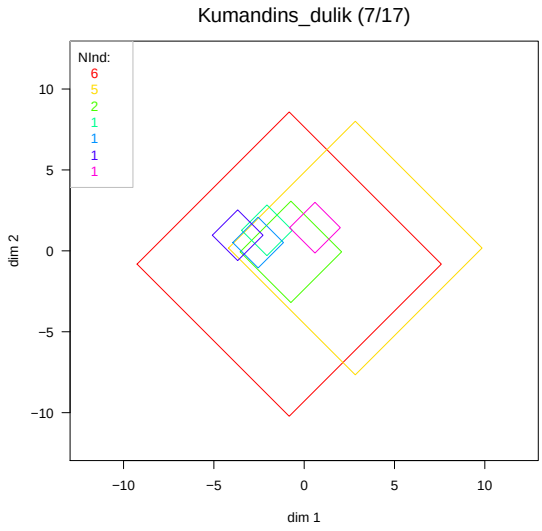

Tlg (14/20)
NInd:
3
2
2
2
2
1
1
1
1
1
1
1
1
1
dim 2
dim 1
10
5
0
−5
−10
−10
−5
0
5
10

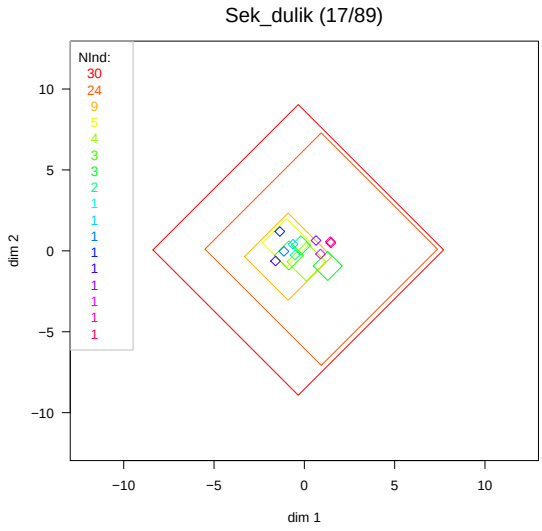

Tub (11/20)
NInd:
6
3
3
1
1
1
1
1
1
1
1
dim 2
dim 1
10
5
0
−5
−10
−10
−5
0
5
10

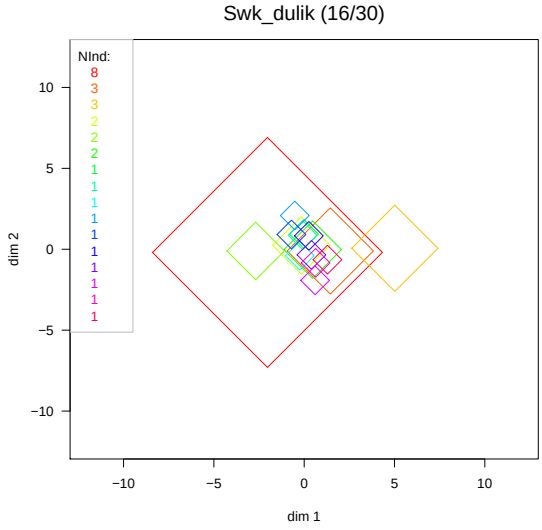
From Dulik:

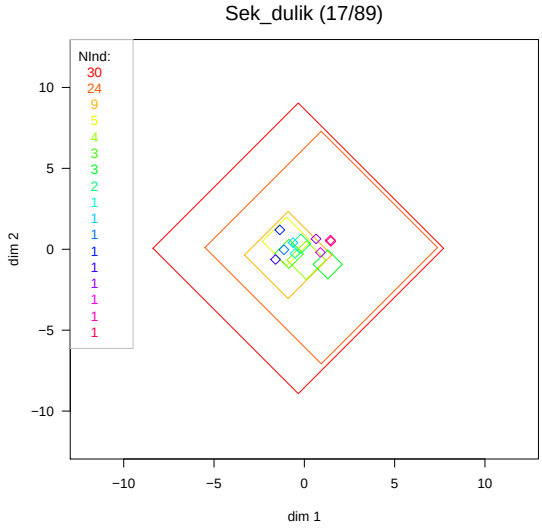

Altaikizhi_dulik (44/121)
NInd:
25
14
10
6
5
4
4
3
3
3
3
2
2
2
2
2
2
2
2
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
1
dim 1
dim 2


Chelkans_dulik (13/25)
NInd:
6
3
3
2
2
2
1
1
1
1
1
1
1
dim 1
dim 2


Kumandins_dulik (7/17)
NInd:
6
5
2
1
1
1
1
dim 1
dim 2


Sek_dulik (17/89)
NInd:
30
24
9
5
4
3
3
2
1
1
1
1
1
1
1
1
dim 1
dim 2


Swk_dulik (16/30)
NInd:
8
3
3
2
2
2
1
1
1
1
1
1
1
1
1
1
dim 1
dim 2

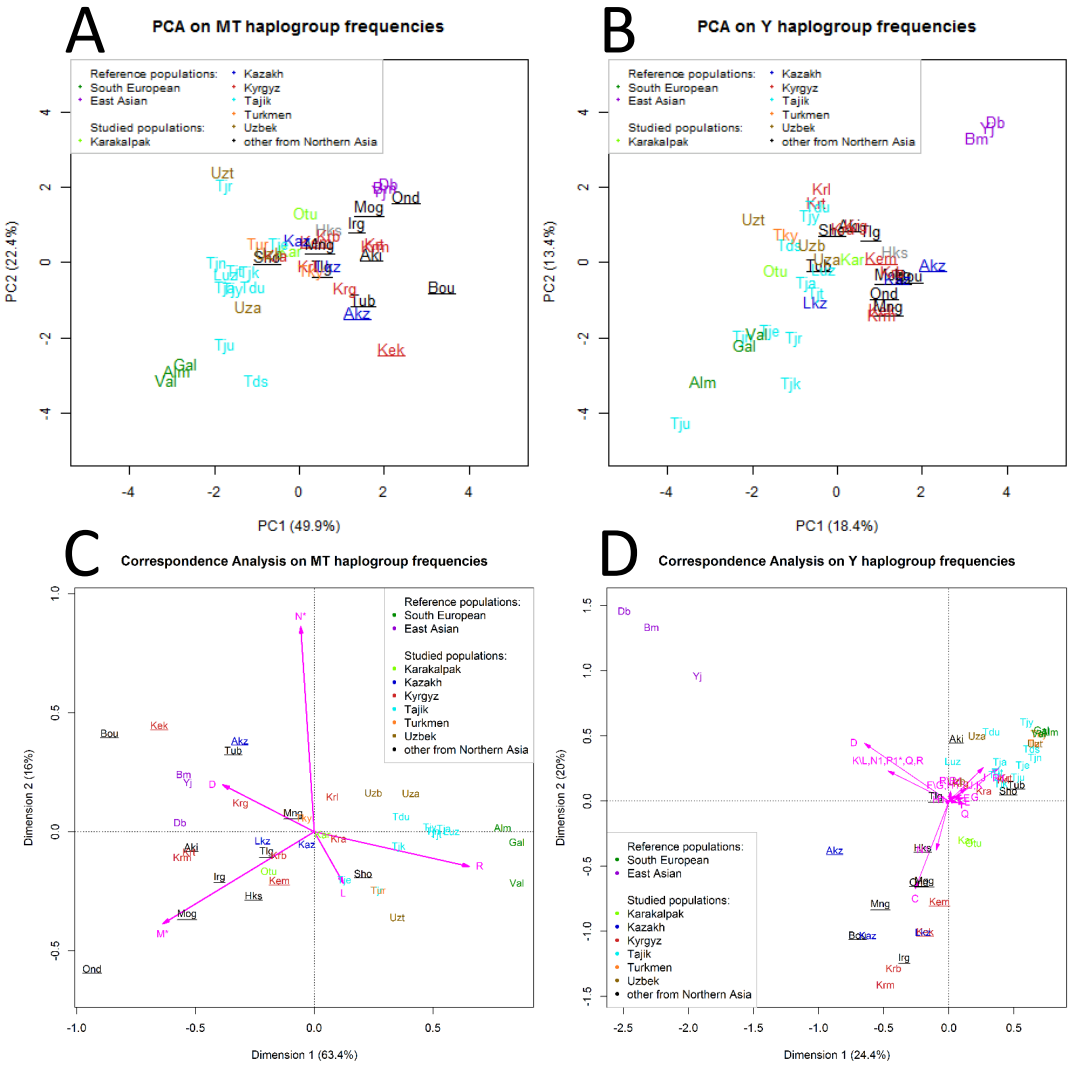

Sek_dulik (17/89)
NInd:
30
24
9
5
4
3
3
2
1
1
1
1
1
1
1
1
dim 1
dim 2

Supplementary Information - *Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations.*

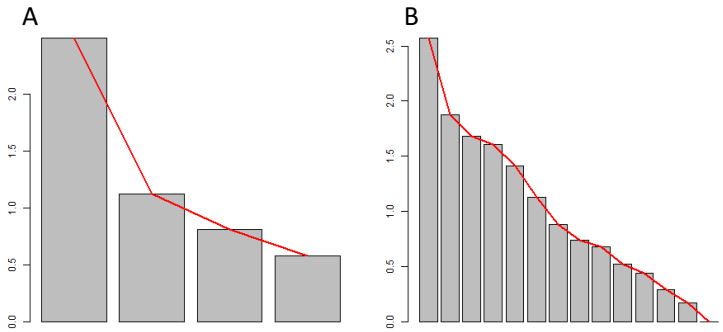
Disponibles aux liens suivants :

- Supplementary Figures
- Supplementary Table 1
- Supplementary Table 2
- Supplementary Table 3
- Supplementary Table 4
- Supplementary Table 5
- Supplementary Table 6

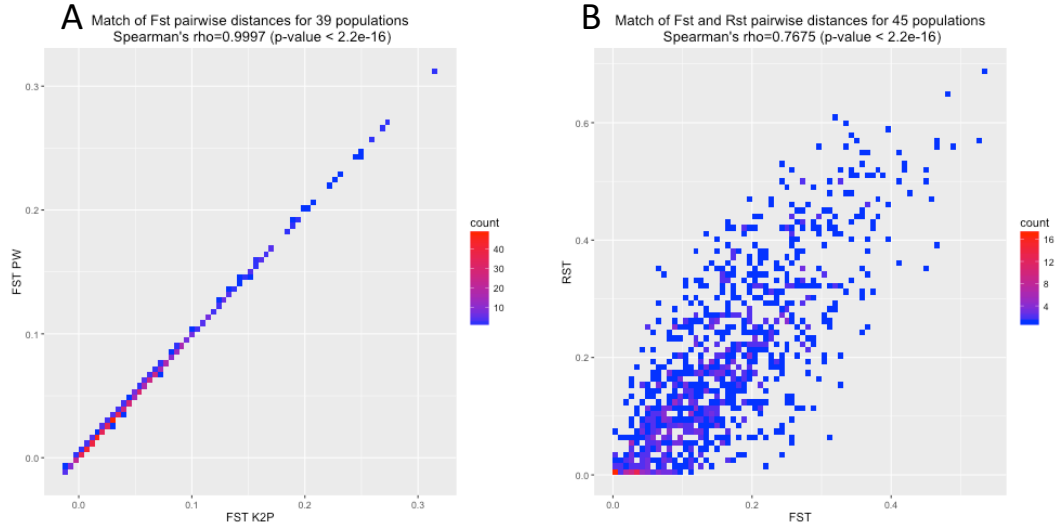
Supplementary Data 1 – Analyses computed over haplogroup frequencies. (A & B) The two first dimensions of the Principal Component Analyses led on mitochondrial haplogroup distribution (A) or Y chromosome (B) are represented. (C & D) The two first dimensions of the Correspondence Analyses led on mitochondrial haplogroup distribution (C) or Y chromosome (D) are represented in this asymmetrical plot. The pink arrows indicate the contribution of haplogroup categories to the variance.



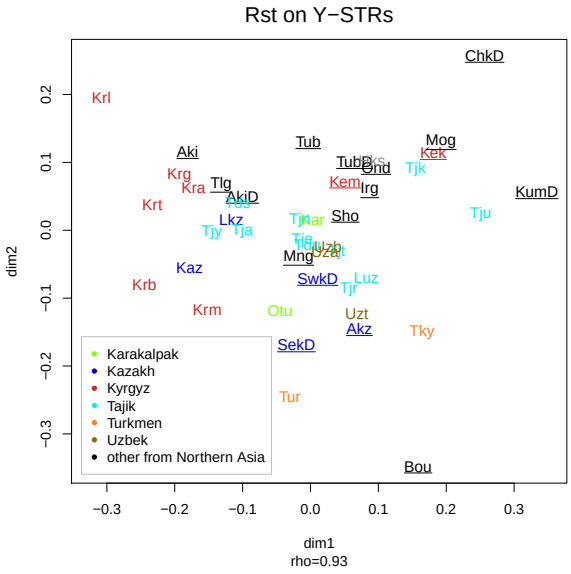
Supplementary Data 2 – Representation of the eigenvalues for successive factors of PCA over haplogroup frequencies. This scree plot is used to graphically determine the optimal number of factors to retain for mitochondrial (A) or Y chromosome (B) analysis. The red line pictures the break after the first factor.



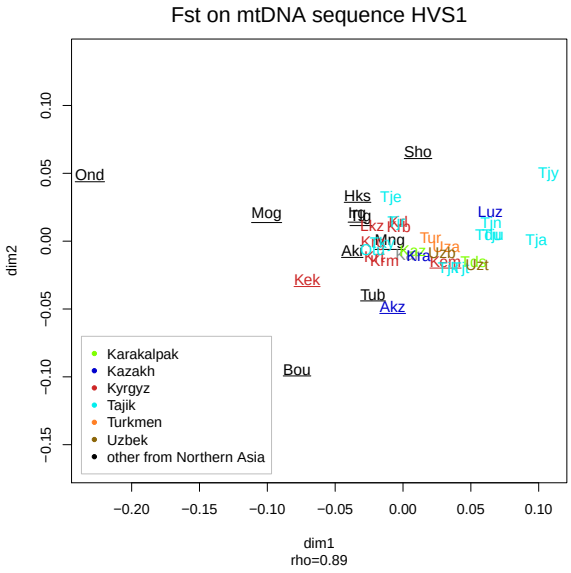
Supplementary Data 3 – Comparison between population pairwise genetic distances. (A) The correlation between FST distances obtained with "Kimura-2-Parameters" (K2P) model and "Pairwise differences" (PW) model over mitochondrial HV1 sequence was of 0.9997 (p-value$<2.2*10^{-16}$). (B) The correlation between FST and RST distances obtained over Y-STRs was of 0.7675 (p-value$=10^{-14}$).



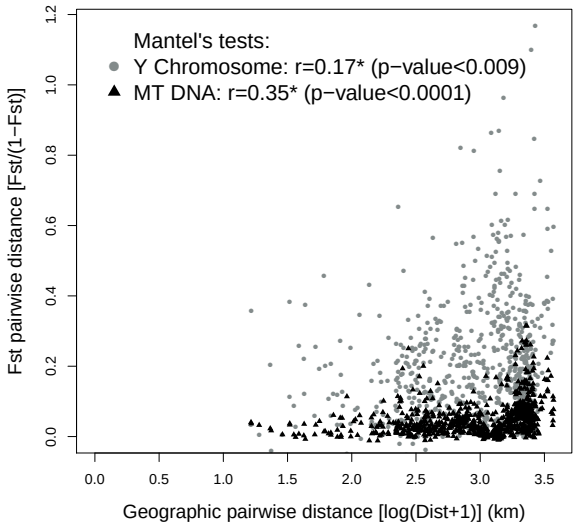
Supplementary Data 4 – Multidimensional Scaling (MDS) analysis based on pairwise RST between populations over Y-STRs. Spearman's coefficient of correlation (rho) is calculated between the matrix of pairwise RST values and the scaled matrix. See Table 1 for the population code and ethnic group affiliation.



Supplementary Data 5 – Multidimensional Scaling (MDS) analysis based on pairwise FST between populations over mitochondrial HVS1. Spearman's coefficient of correlation (rho) is calculated between the matrix of pairwise FST values and the scaled matrix. See Table 1 for the population code and ethnic group affiliation. This plot includes the Ondar population Ond, outlier excluded in Figure 3.



Supplementary Data 5 – Representation of the genetic distances as a function of the geographic distances. Genetic distances are FST population pairwise distances expressed in FST/(1-FST). Population pairwise geographic distances are expressed in log. In grey, values for Y-STR haplotypes ; in black, values for mitochondrial HVSI. In the legend, Spearman's correlation between the geographical and genetic distances computed with Mantel's tests.



Supplementary Table 1 - Sample Descriptions and Estimators of Y-Chromosome Genetic Diversity. Samples are described with their code acronym, ethnic group membership, social organisation, way of life, geographic location in Central or Northern Asia, population size for mitochondrial DNA and Y chromosome (N). The patrilineal populations are in the white area ; the cognatic populations are in the dark grey area ; the Uzbek populations are in the light grey area as they were historically nomadic patrilineal herders but have made a sedentary agriculturist transition since the 16th century. The estimators of genetic diversity are π, the mean number of pairwise differences ; H, the haplotypic heterozygosity ; Ps, the proportion of singletons ; C, the mean number of individuals sharing the same Y STR haplotype.

Population informations							MT DNA	Y Chromosome				
Code	Reference	Linguistic family*	Ethnic group	Social Organisation	Way of life	Location (Asia)	Nb	Nb	H	Ps	C	π
Luz	(Martínez-Cruz et al.. 2011	I-I	Tajik	Cognatic	Agriculturist	Central	46	31	0.968	0.550	1.333	3.077
Tds	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	31	30	0.840	0.360	1.923	3.664
Tdu	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	40	31	0.952	0.533	1.364	2.226
Tja	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	32	32	0.976	0.645	1.292	4.964
Tje	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	31	27	0.975	0.542	1.333	4.613
Tjk	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	40	35	0.975	0.542	1.333	5.123
Tjn	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	35	30	0.969	0.577	1.368	2.823
Tjr	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	29	29	0.985	0.862	1.115	5.377
Tjt	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	32	30	0.958	0.500	1.474	5.131
Tju	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	29	29	0.975	0.517	1.381	5.037
Tjy	(Ségurel et al.. 2008)	I-I	Tajik	Cognatic	Agriculturist	Central	40	36	0.840	0.207	2.636	3.200
Uza	(Martínez-Cruz et al.. 2011	T-M	Uzbek	Patrilineal	Pastoralist/ Agriculturist	Central	36	36	0.994	0.818	1.100	5.056
Uzb	(Chaix et al.. 2004)	T-M	Uzbek	Patrilineal	Pastoralist/ Agriculturist	Central	40	40	0.981	0.667	1.258	5.223
Uzt	(Martínez-Cruz et al.. 2011	T-M	Uzbek	Patrilineal	Pastoralist/ Agriculturist	Central	39	35	0.701	0.167	3.333	3.479
Tky	Present study	T-M	Turkmen	Patrilineal	Pastoralist	Central	35	31	0.883	0.500	1.692	2.204
Tur	(Chaix et al.. 2004)	T-M	Turkmen	Patrilineal	Pastoralist	Central	51	51	0.888	0.294	2.318	3.450
Kar	(Chaix et al.. 2004)	T-M	Karakalpak	Patrilineal	Pastoralist	Central	55	54	0.971	0.370	1.742	5.044
Otu	(Chaix et al....	T-M	Karakalpak	Patrilineal	Pastoralist	Central	53	54	0.915	0.352	2.077	4.790
Akz	Present study	T-M	Kazakh	Patrilineal	Pastoralist	Northern	41	22	0.784	0.227	2.750	4.004
Kaz	(Chaix et al.. 2004)	T-M	Kazakh	Patrilineal	Pastoralist	Central	50	50	0.853	0.260	2.381	3.578
SekD	(Dulik et al.. 2011)	T-M	Kazakh	Patrilineal	Pastoralist	Northern		89	0.803	0.101	5.235	4.304
SwkD	(Dulik et al.. 2011)	T-M	Kazakh	Patrilineal	Pastoralist	Northern		30	0.915	0.333	1.875	4.747
Lkz	(Ségurel et al.. 2008)	T-M	Kazakh	Patrilineal	Pastoralist	Central	31	20	0.978	0.800	1.111	2.763
Kek	Present study	T-M	Kyrgyz	Patrilineal	Pastoralist	Northern	19	8	0.929	0.500	1.333	4.964
Kem	Present study	T-M	Kyrgyz	Patrilineal	Pastoralist	Northern	25	10	0.917	0.667	1.286	3.933
Kra	(Ségurel et al.. 2008)	T-M	Kyrgyz	Patrilineal	Pastoralist	Central	48	46	0.821	0.239	2.875	3.843
Krb	(Ségurel et al.. 2008)	T-M	Kyrgyz	Patrilineal	Pastoralist	Central	30	31	0.751	0.207	2.900	2.574
Krg	(Ségurel et al.. 2008)	T-M	Kyrgyz	Patrilineal	Pastoralist	Central	20	20	0.751	0.207	2.900	4.347
Krl	(Ségurel et al.. 2008)	T-M	Kyrgyz	Patrilineal	Pastoralist	Central	24	40	0.733	0.270	2.643	2.201
Krm	(Ségurel et al.. 2008)	T-M	Kyrgyz	Patrilineal	Pastoralist	Central	36	22	0.753	0.318	2.200	2.593
Krt	(Ségurel et al.. 2008)	T-M	Kyrgyz	Patrilineal	Pastoralist	Central	29	37	0.619	0.235	3.091	2.111
Aki	Present study	T-M	Altai Kizhi	Patrilineal	Pastoralist	Northern	41	24	0.841	0.292	2.182	3.580
Bou	Present study	T-M	Bouryat	Patrilineal	Pastoralist	Northern	28	28	0.625	0.174	3.286	2.757
Irg	Present study	T-M	Irgit	Patrilineal	Pastoralist	Northern	40	15	0.945	0.500	1.400	4.162
Hks	Present study	T-M	Khakas	Patrilineal	Agriculturist	Northern	39	19	0.941	0.444	1.500	4.175
Mng	Present study	T-M	Mongol	Patrilineal	Pastoralist	Northern	76	40	0.978	0.600	1.333	4.895
Mog	Present study	T-M	Mongush	Patrilineal	Pastoralist	Northern	26	12	0.964	0.636	1.222	4.424
Ond	Present study	T-M	Ondar	Patrilineal	Pastoralist	Northern	16	9	0.917	0.333	1.500	4.722
Sho	Present study	T-M	Shor	Patrilineal	Hunter	Northern	39	18	0.980	0.667	1.200	4.405
Tlg	Present study	T-M	Telengit	Patrilineal	Pastoralist	Northern	40	20	0.963	0.450	1.429	4.258
Tub	Present study	T-M	Tubalar	Patrilineal	Hunter	Northern	35	20	0.889	0.400	1.818	4.068
AkiD	(Dulik et al.. 2012)	T-M	Altai Kizhi	Patrilineal	Pastoralist	Northern		120	0.933	0.208	2.727	4.525
ChkD	(Dulik et al.. 2012)	T-M	Chelkan	Patrilineal	Pastoralist	Northern		25	0.920	0.280	1.923	4.403
KumD	(Dulik et al.. 2012)	T-M	Kumandin	Patrilineal	Pastoralist	Northern		17	0.809	0.235	2.429	3.676
TubD	(Dulik et al.. 2012)	T-M	Tubalar	Patrilineal	Hunter	Northern		27	0.946	0.370	1.688	4.897

** I-I: Indo-Iranian: T-M: Turko-Mongolic*

Supplementary Table 2 – Raw genetic data. The table provides all genetic data used in this study and produced by our team. For Y chromosome, STRs, SNPs and haplogroup based on Phylotree are available. For mitochondrial DNA, typed SNPs used in the determination of the haplogroup are listed. For the sampled that are resequenced for HVS1 (and possibly HVS2), all the positions in the given range are tested and only the alleles that differ from rCRS were listed. The detailed haplogroup is the one determined with Haplogrep, with a certain quality in %. Simplified haplogroup correspond to the categories used in the frequencies study. GenBank accession number are given for the resequenced data. *Cette table est disponible sous le format excel en ligne Supp Table 2.*

Supplementary Table 3 – Haplogroup distributions and results from the analysis that followed. Table A provides haplogroup frequencies by population for Y chromosome and mitochondrial DNA. Table B presents the eigenvalues on the first Principal Component from the 1D-PCA plot realised on the sampled populations and the 3 reference populations from South Europe and East Asia, for mtDNA and Y chromosome respectively.

Table A :

Y Chromosome																		
Population	Population size	Y(xCR)	D	E	C	F\G,H1,I,J,K	G	H1	I	J	K\L,N1,P1*,Q,R	L	N1	P1*	Q	R1	R\R1
Aki	24	0.00	0.08	0.00	0.04	0.00	0.00	0.00	0.00	0.04	0.08	0.00	0.04	0.00	0.04	0.67	0.00
Bou	28	0.00	0.00	0.04	0.75	0.00	0.00	0.00	0.00	0.00	0.18	0.00	0.04	0.00	0.00	0.00	0.00
Hks	19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.16	0.00	0.53	0.00	0.00	0.32	0.00
Irg	14	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.57	0.00	0.00	0.07	0.00
Mng	39	0.00	0.03	0.00	0.51	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.18	0.00	0.00	0.10	0.03
Mog	11	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.09	0.09	0.00	0.55	0.00	0.09	0.09	0.00
Ond	9	0.00	0.00	0.00	0.22	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.33	0.00	0.22	0.11	0.00
Sho	18	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.22	0.00	0.06	0.67	0.00
Tlg	20	0.00	0.05	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.15	0.00	0.00	0.55	0.00
Tub	20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.10	0.00	0.30	0.60	0.00
Akz	22	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.09	0.41	0.00	0.00	0.00	0.00	0.00	0.00
Kaz	50	0.00	0.04	0.02	0.78	0.00	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.04	0.06	0.00
Lkz	20	0.00	0.00	0.00	0.65	0.00	0.15	0.00	0.05	0.00	0.00	0.05	0.05	0.00	0.00	0.05	0.00
Kek	8	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.25	0.00	0.00
Kem	10	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.30	0.00
Kra	46	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.02	0.07	0.00	0.02	0.00	0.04	0.67	0.00
Krb	31	0.00	0.00	0.00	0.84	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.13	0.00
Krg	20	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.10	0.15	0.00	0.00	0.00	0.00	0.55	0.00
Krl	40	0.00	0.00	0.00	0.15	0.03	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.00	0.03	0.75	0.00
Krm	22	0.00	0.00	0.00	0.91	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.05	0.00	0.00
Krt	37	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.03	0.00	0.03	0.00	0.00	0.00	0.00	0.81	0.00
Tky	22	0.00	0.00	0.05	0.05	0.00	0.00	0.00	0.00	0.55	0.00	0.00	0.00	0.00	0.09	0.27	0.00
Tur	51	0.00	0.00	0.00	0.00	0.06	0.02	0.00	0.02	0.14	0.00	0.00	0.04	0.00	0.00	0.73	0.00
Kar	54	0.00	0.00	0.00	0.19	0.00	0.02	0.00	0.00	0.02	0.04	0.00	0.20	0.00	0.11	0.43	0.00
Otu	54	0.00	0.00	0.00	0.31	0.00	0.26	0.00	0.06	0.13	0.02	0.00	0.07	0.00	0.02	0.13	0.00
Uza	36	0.00	0.03	0.00	0.08	0.00	0.03	0.00	0.00	0.36	0.11	0.00	0.00	0.00	0.03	0.31	0.06
Uzb	40	0.00	0.05	0.03	0.13	0.03	0.03	0.00	0.03	0.05	0.08	0.00	0.05	0.00	0.23	0.30	0.03
Uzt	35	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.09	0.69	0.00	0.00	0.00	0.00	0.14	0.00	0.00
Luz	31	0.00	0.00	0.06	0.13	0.00	0.13	0.00	0.00	0.26	0.23	0.00	0.00	0.00	0.06	0.13	0.00
Tds	30	0.00	0.00	0.00	0.03	0.00	0.13	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.67	0.03
Tdu	31	0.00	0.00	0.00	0.00	0.03	0.03	0.00	0.00	0.13	0.16	0.03	0.00	0.00	0.00	0.61	0.00
Tja	31	0.00	0.00	0.00	0.10	0.00	0.03	0.00	0.00	0.16	0.06	0.06	0.00	0.00	0.00	0.52	0.06
Tje	27	0.00	0.00	0.00	0.07	0.00	0.07	0.00	0.00	0.26	0.00	0.11	0.00	0.00	0.11	0.33	0.04
Tjk	35	0.00	0.00	0.00	0.03	0.00	0.00	0.00	0.00	0.03	0.03	0.09	0.03	0.00	0.40	0.29	0.11
Tjn	30	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.13	0.00	0.13	0.00	0.00	0.07	0.47	0.00
Tjr	29	0.00	0.00	0.00	0.07	0.00	0.14	0.00	0.00	0.10	0.03	0.03	0.07	0.00	0.03	0.34	0.17
Tjt	30	0.00	0.00	0.00	0.07	0.00	0.03	0.00	0.00	0.17	0.07	0.00	0.00	0.00	0.33	0.30	0.03
Tju	29	0.00	0.00	0.21	0.00	0.00	0.10	0.03	0.00	0.07	0.03	0.07	0.10	0.00	0.24	0.14	0.00
Tjy	36	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.42	0.06	0.00	0.00	0.00	0.03	0.50	0.00
Gal	44	0.00	0.00	0.18	0.00	0.00	0.02	0.00	0.05	0.14	0.02	0.00	0.00	0.00	0.00	0.59	0.00
Val	59	0.00	0.00	0.12	0.00	0.00	0.03	0.00	0.05	0.12	0.02	0.00	0.00	0.00	0.00	0.66	0.00
Alm	36	0.00	0.00	0.19	0.00	0.00	0.06	0.00	0.11	0.14	0.00	0.00	0.00	0.00	0.00	0.50	0.00
Bm	16	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.13	0.00	0.00	0.00	0.06
Db	18	0.00	0.50	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.44	0.00	0.00	0.00	0.00	0.00	0.00
Yj	47	0.00	0.26	0.00	0.11	0.00	0.00	0.00	0.00	0.02	0.62	0.00	0.00	0.00	0.00	0.00	0.00

MT DNA						
Population	Population size	D	L\M,N	M\D	N\R	R
Aki	44	0.14	0.00	0.49	0.09	0.28
Bou	41	0.33	0.00	0.41	0.19	0.07
Hks	27	0.10	0.00	0.44	0.03	0.44
Irg	39	0.18	0.00	0.42	0.03	0.37
Mng	38	0.22	0.01	0.22	0.08	0.47
Mog	55	0.16	0.00	0.53	0.00	0.32
Ond	49	0.06	0.00	0.81	0.00	0.13
Sho	18	0.03	0.00	0.28	0.05	0.64
Tlg	24	0.05	0.00	0.43	0.10	0.43
Tub	30	0.21	0.00	0.29	0.18	0.32
Akz	48	0.18	0.00	0.30	0.20	0.33
Kaz	31	0.12	0.02	0.29	0.08	0.49
Lkz	20	0.10	0.00	0.39	0.10	0.42
Kek	37	0.06	0.00	0.53	0.29	0.12
Kem	26	0.09	0.00	0.39	0.04	0.48
Kra	28	0.06	0.02	0.27	0.10	0.54
Krb	31	0.19	0.00	0.29	0.03	0.48
Krg	48	0.10	0.00	0.40	0.15	0.35
Krl	79	0.22	0.00	0.14	0.08	0.57
Krm	19	0.15	0.00	0.50	0.08	0.27
Krt	16	0.18	0.00	0.46	0.07	0.29
Tky	53	0.13	0.00	0.28	0.10	0.50
Tur	39	0.04	0.00	0.25	0.02	0.69
Kar	24	0.11	0.02	0.26	0.09	0.52
Otu	31	0.11	0.04	0.38	0.08	0.40
Uza	40	0.03	0.00	0.14	0.14	0.69
Uzb	32	0.18	0.03	0.08	0.11	0.61
Uzt	31	0.08	0.06	0.19	0.00	0.67
Luz	40	0.06	0.00	0.06	0.06	0.81
Tds	35	0.00	0.00	0.10	0.29	0.61
Tdu	29	0.05	0.00	0.15	0.10	0.69
Tja	32	0.00	0.00	0.13	0.09	0.78
Tje	29	0.06	0.00	0.29	0.03	0.61
Tjk	40	0.03	0.00	0.20	0.08	0.70
Tjn	40	0.00	0.03	0.14	0.11	0.71
Tjr	40	0.04	0.07	0.25	0.07	0.57
Tjt	35	0.06	0.00	0.09	0.06	0.78
Tju	51	0.00	0.00	0.03	0.21	0.76
Tjy	36	0.00	0.00	0.15	0.10	0.75
Gal	56	0.00	0.00	0.00	0.05	0.95
Val	132	0.00	0.03	0.02	0.02	0.92
Alm	89	0.00	0.03	0.01	0.10	0.85
Bm	40	0.18	0.00	0.43	0.18	0.23
Db	46	0.24	0.00	0.41	0.09	0.26
Yj	192	0.15	0.00	0.44	0.17	0.23

Table B :

Populations	Y Chromosome	MtDNA
Aki	0.302	1.690
Bou	1.704	3.357
Hks	1.377	0.728
Irg	1.549	1.366
Mng	1.211	0.482
Mog	1.220	1.644
Ond	1.105	2.523
Sho	-0.117	-0.763
Tlg	0.780	0.545
Tub	-0.424	1.501
Akz	2.242	1.373
Kaz	1.433	-0.026
Lkz	-0.482	0.717
Kek	1.063	2.169
Kem	1.044	0.404
Kra	0.145	-0.548
Krb	1.310	0.689
Krg	0.454	1.074
Krl	-0.363	0.184
Krm	1.066	1.812
Krt	-0.469	1.790
Tky	-1.178	0.306
Tur	-1.288	-0.956
Kar	0.352	-0.208
Otu	-1.425	0.172
Uza	-0.228	-1.195
Uzb	-0.589	-0.697
Uzt	-1.949	-1.804
Luz	-0.315	-1.708
Tds	-1.107	-0.982
Tdu	-0.438	-1.067
Tja	-0.706	-1.739
Tje	-1.581	-0.458
Tjk	-1.080	-1.152
Tjn	-2.219	-1.937
Tjr	-1.032	-1.779
Tjt	-0.483	-1.519
Tju	-3.661	-1.751
Tjy	-0.633	-1.558
Gal	-2.173	-2.682
Val	-1.875	-3.118
Alm	-3.138	-2.866
Bm	3.306	2.009
Db	3.748	2.092
Yj	3.539	1.888

Means	Y Chromosome	MtDNA
Europeans	-2.395	-2.888
East Asians	3.531	1.996
Karakalpaks	-0.536	-0.018
Kazakh	1.064	0.688
Kyrgyz	0.531	0.947
Uzbeks	-0.922	-1.232
Tajiks	-1.205	-1.423
Turkmens	-1.233	-0.325
others from Northern Asia	0.871	1.307

Supplementary Table 4 – BATWING priors and results. Ethnic group demographic parameters (merging times, effective population and subpopulation sizes) were estimated with two approaches : a fixed mutation rate for all the STRs (0.0028) or a fixed mutation rate for each STR, based on the literature. In any case, 110,000 runs were simulated and 10,000 were burn as a warmup. Moreover, the used model was with no migration and a constant population size. The population size was estimated with a prior between 100 and 1,000,000, and the effective population size Ne is the averaged estimation. Two priors for the tree shape (Nbetsamp and treebetN) were fixed at 200 and 100 respectively. We chose to use estimations made over the whole 100,000 runs for the oldest merging event (columns in light grey). The correspondence between populations name and number used into BATWING are available in each table. Merging events were determined based on the formula given into BATWING user guide. For each merging event, we presented the combination that obtained the highest number of runs.
Cette table est disponible sous le format excel en ligne Supp Table 4.

Supplementary Table 5 – AMOVAs and Mantel's tests based on Y-STR RST distances. As for FST distances computed on Y-STRs, we analysed RST distances for hierarchical AMOVA and fixation indexes based on ethnic group affiliation and/or geography (A), correlation with geographic distances using Mantel's tests (B) or intra ethnic group differentiation (C).

Table A :

Grouping	N groups	N populations	Source of variation	Percentage of variation		Fixation indexes		
				Y chr	*mtDNA*	*Y chr*	*mtDNA*	
Geography	2	39	Among groups	N.S.	2.04	N.S.	0.0204	F_{CT}
			Among populations within groups	18.97	3.56	0.19	0.0363	F_{SC}
			Within populations	80.84	94.4	0.1917	0.056	F_{ST}
Geography *Kyrgyz and Kazakh excluded*	2	28	Among groups	N.S.	2.9	N.S.	0.029	F_{CT}
			Among populations within groups	15.68	3.81	0.1576	0.0392	F_{SC}
			Within populations	83.85	93.3	0.1615	0.067	F_{ST}
Ethnicity *Inner Asia*	16	39	Among groups	7.43	2.66	0.0743	0.0289	F_{CT}
			Among populations within groups	12.38	2.11	0.1338	0.0193	F_{SC}
			Within populations	80.18	95.22	0.1982	0.0477	F_{ST}
Ethnicity *Central Asia*	6	26	Among groups	8.69	0.97	0.0869	0.0097	F_{CT}
			Among populations within groups	10.73	1.86	0.1175	0.0188	F_{SC}
			Within populations	80.59	97.17	0.1941	0.0283	F_{ST}
Ethnicity *Northern Asia*	12	13	Among groups	24.05	N.S.	0.2405	N.S.	F_{CT}
			Among populations within groups	N.S.	6.25	N.S.	0.062	F_{SC}
			Within populations	76.33	94.45	0.2367	0.0555	F_{ST}

Hierarchical AMOVA and fixation indexes based on ethnic group affiliation and/or geography.
N.S. for non-significant p-value (>0.05). The numbers of considered groups and populations are indicated.

Table B :

	Region	Spearman's r	p-value
Mantel Tests	Inner Asia	0.15	<0.01
	Central Asia	0.1177	0.087
	Northern Asia	0.4919	0.025
Partial Mantel Test	Inner Asia	0.1166	0.034

Table C :

Ethnic group	N	Percentage of intra-group differentiation	
		Y chr	*mtDNA*
Karakalpak	2	11.77	N.S.
Kazakh	3	16.36	2.35
Central Asian Kazakhs	*2*	*5.05*	*N.S.*
Northern Asian Kazakhs	*1*	*/*	*/*
Kyrgyz	8	18.8	2.01
Central Asian Kyrgyz	*6*	*14.29*	*N.S.*
Northern Asian Kyrgyz	*2*	*N.S.*	*10*
Tajik	11	9.31	3.16
Turkmen	2	13.9	1.52
Uzbek	3	4.45	1.25

Percentage of R_{ST} variance explained by differences within each ethnic group, for both Y chromosome and mtDNA. N.S. for non-significant p-value (>0.05). The number of populations by group is indicated.

Supplementary Table 6 – Mann-Whitney's U tests realised on Y chromosome genetic diversity estimators. Two-tailed U tests are realised to test equality of the repartitions of the estimator values obtained for populations in groups. One-tailed U tests are realised to test hypothesis of a group with more or less high values than the other. Yellow cells indicate significant p-values at 5For groups, P is for Patrilineal, CP : Patrilineal from Central Asia, NP : Patrilineal from Northern Asia, CC : Central Asian Cognatic.

	Mann-Whitney's test	**Estimator**	**Uzbeks included to Central Asian patrilineal group**		**Uzbeks excluded from Central Asian patrilineal group**	
			Value	*p-value*	*Value*	*p-value*
Patrilineals (Central / Northern Asia)	Bilateral (CP=NP)	Y π	195.0	0.071	177.0	0.010
	Unilateral (CP<NP)	H	182.0	0.088	164.0	0.022
		Ps	158.0	0.301	139.0	0.160
	Unilateral (CP>NP)	C	107.0	0.112	70.0	0.039
Patrilineals (Central and Northern Asia) / Cognatics (Central Asia)	Bilateral (CC=P)	Y π	153.0	0.382	131.0	0.269
	Unilateral (P<CC)	H	95.0	0.008	74.0	0.003
		Ps	100.5	0.012	80.5	0.005
	Unilateral (P>CC)	C	269.5	0.015	257.5	0.007
Patrilineals (Central Asia) / Cognatics (Central Asia)	Bilateral (CP=CC)	Y π	57.0	0.198	35.0	0.059
	Unilateral (CP<CC)	H	45.0	0.027	24.0	0.005
		Ps	42.5	0.020	22.5	0.004
	Unilateral (CP>CC)	C	125.0	0.015	113.0	0.002
Patrilineals (Northern Asia) / Cognatics (Central Asia)	Bilateral (NP=CC)	Y π	96.0	0.735		
	Unilateral (NP<CC)	H	50.0	0.010		
		Ps	58.0	0.024		
	Unilateral (NP>CC)	C	144.5	0.044		

E Annexes du Chapitre III

Supplementary Information - *Close inbreeding and low genetic diversity despite geographical exogamy in Inner Asian human populations.*

Figure S1 – Population exogamy rate. Percentage of exogamous couples with a distance between the partners' place of birth >4 km for the sampled couples (A), or their parents (B)

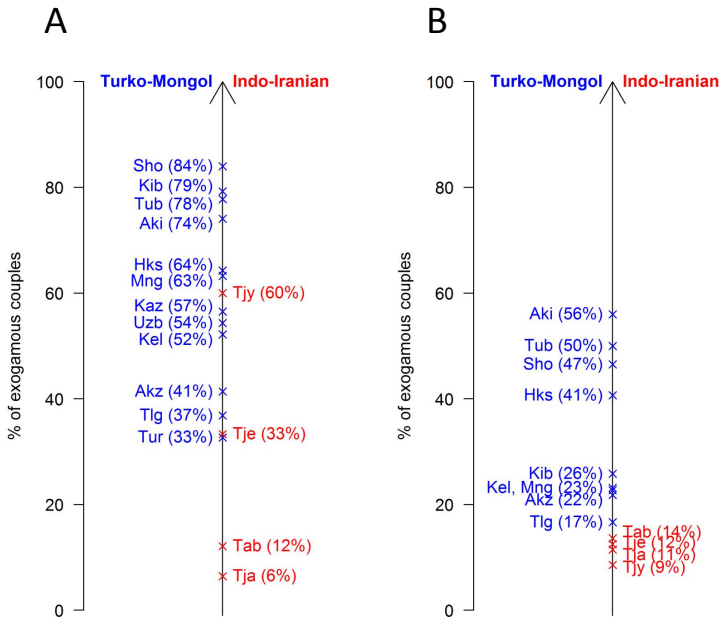


Figure S2 – Geographical distances between parents' places of birth. The distances are plotted in log scale (km).

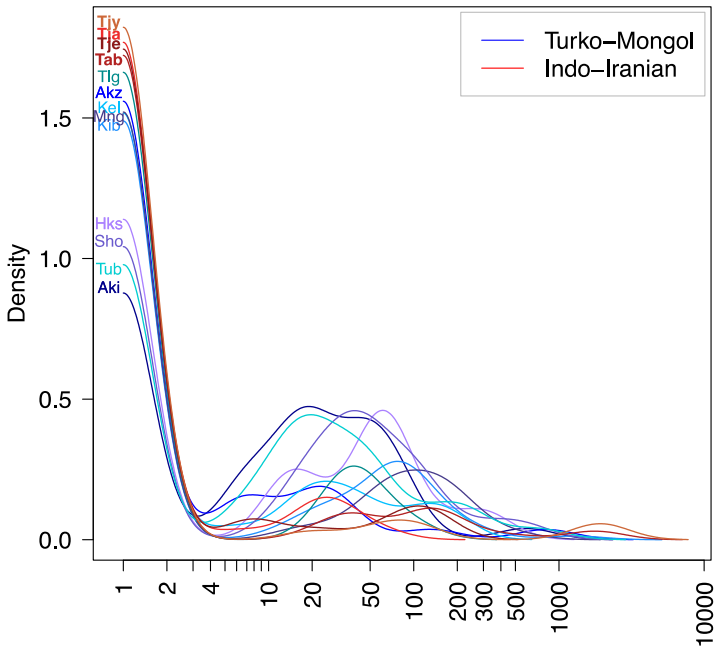


Figure S3 – ASD distances computed for pairs of individuals within population. The number of samples by population is indicated after their name.

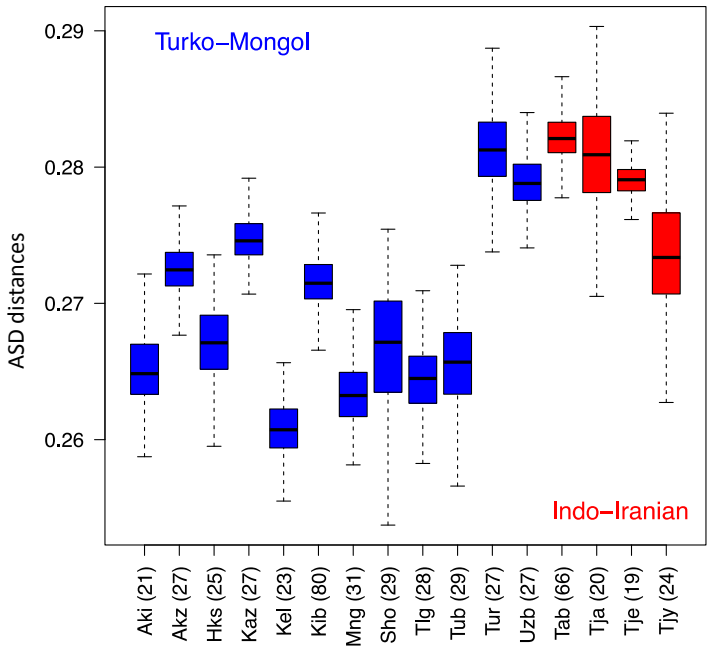


Figure S4 – Multidimensional Scaling (MDS) plot over individuals pairwise ASD distances (A), and populations pairwise FST distances (B). Only the two first dimensions are represented. Each Spearman's coefficient of correlation (rho) was calculated between the MDS matrix and the pairwise genetic distances matrix.

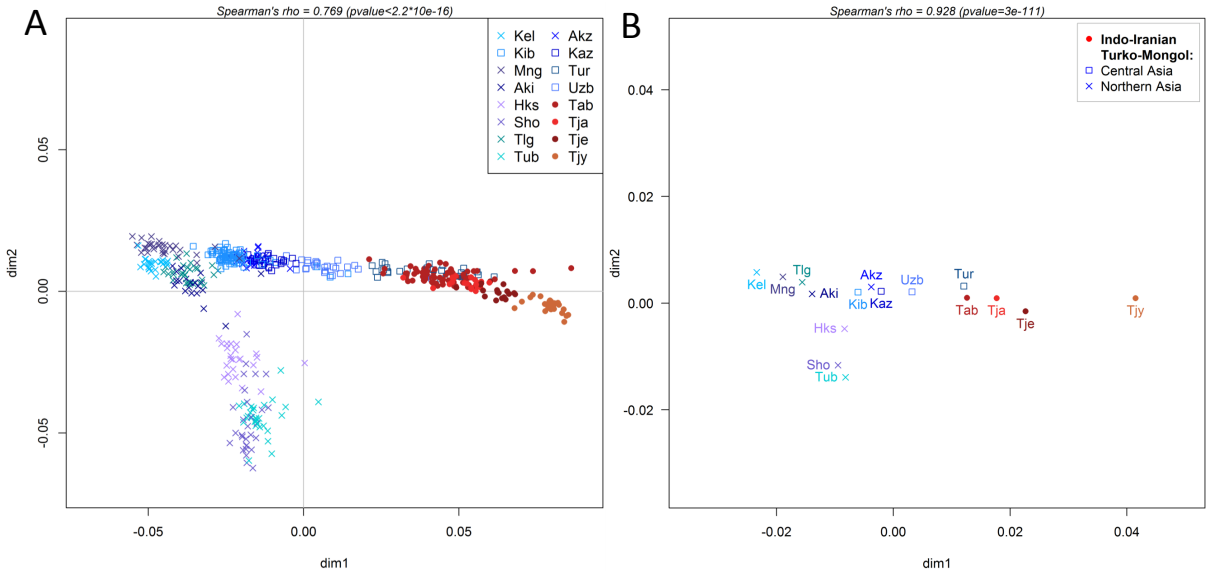


Figure S5 – Populations pattern of intermediate sized ROHs (length between 500 and 1500 kb, or class B) (top) and long sized ROHs (larger than 1500 kb or class C) (bottom). Information is summarized by the number of segments per individual genome (A and C), or their total length by genome (B and D, in kb), and represented for each population. The number of genomes by population is indicated after their name.

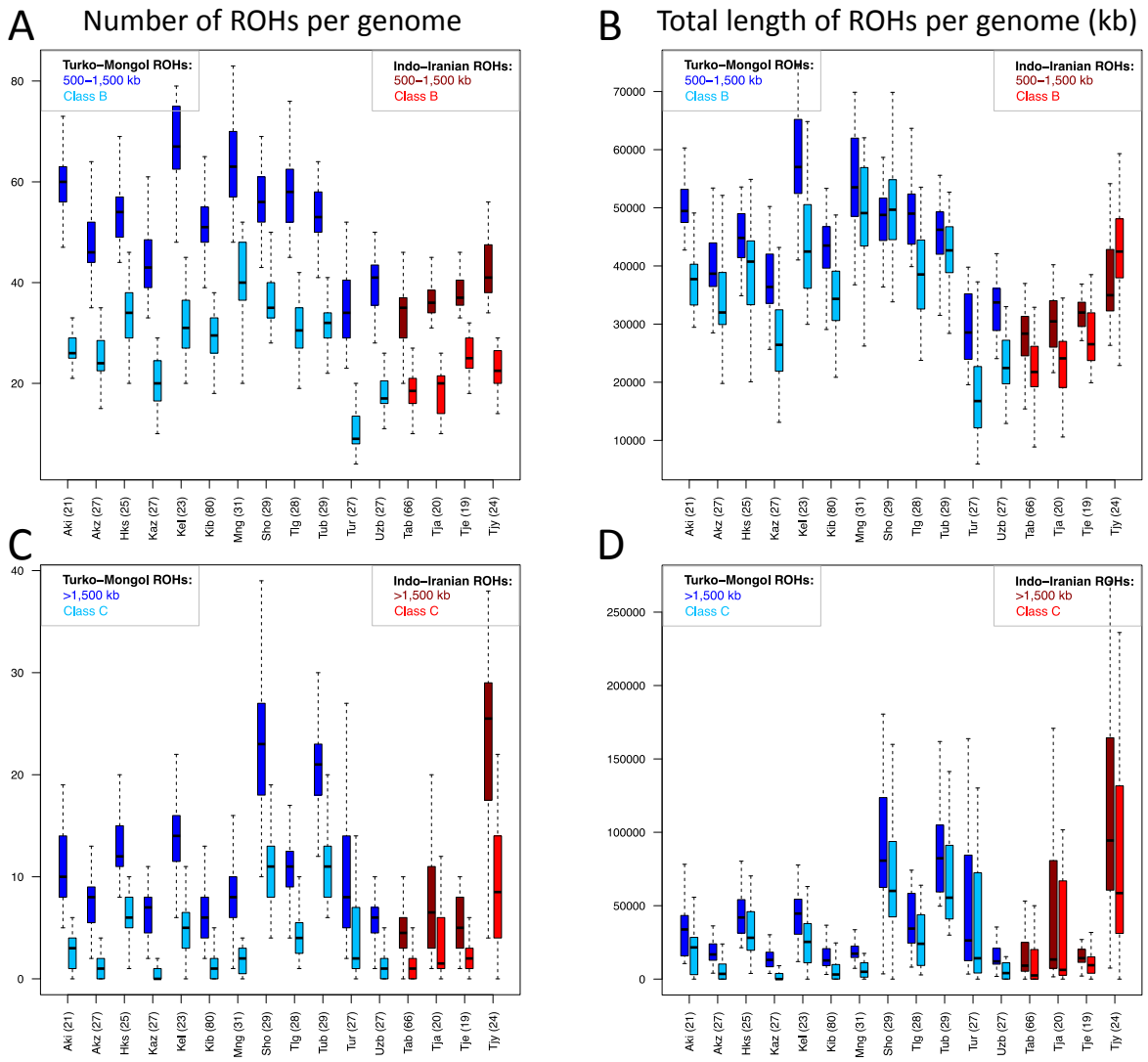


Figure S6 – Inbreeding avoidance within HapMap3 populations. Differences between the observed homozygosity within individual genomes and the population baseline, expected under panmixia. Number of samples by population is indicated after their name.

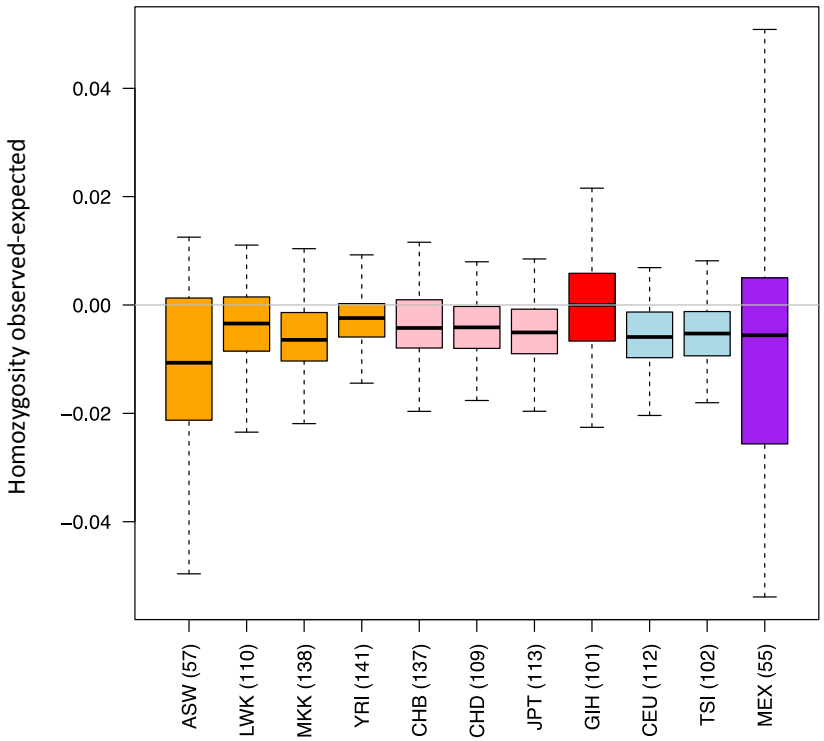


Figure S7 – Inbreeding and parental couple distance for the offspring of exogamous Turko-Mongol couples. Three estimators of inbreeding are plotted against the distances between parent birth places of exogamous couples. These two variables were correlated based on Spearman's correlation test (illustrated as a regression line on the plot).

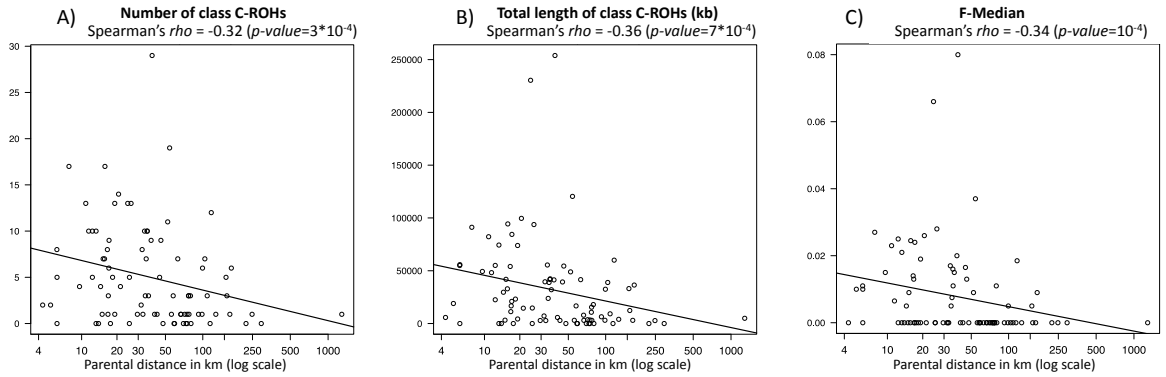


Table S1 - Population description, genetic diversity, inbreeding and exogamy pattern.

Population:		Aki	Akz	Hks	Kaz	Kel	Kib	Mng	Sho	Tlg	Tub	Tur	Uzb	Tab	Tja	Tje	Tjy
Description	Cultural group	Turko-Mongol	Turko-Mongol	Turko-Mongol	Turko-Mongol	Turko-Mongol	Turko-Mongol	Turko-Mongol	Turko-Mongol	Turko-Mongol	Turko-Mongol	Turko-Mongol	Turko-Mongol	Indo-Iranian	Indo-Iranian	Indo-Iranian	Indo-Iranian
	Ethnic Group	Altai-Kizhi	Kazakh	Khakas	Kazakh	Kyrgyz	Kyrgyz	Mongol	Shor	Telengit	Tubalar	Turkmen	Uzbek	Tajik	Tajik	Tajik	Tajik-Yagnob
	Region	NA	NA	NA	CA	CA	CA	NA	CA	NA	NA	CA	CA	CA	CA	CA	CA
Genetic dataset sample size (N)	Females	9	11	15	0	11	40	15	12	13	11	0	0	35	0	0	0
	Males	12	16	10	27	12	40	16	17	15	18	27	27	31	20	19	24
	Couples	4	4	2	0	8	38	14	5	12	4	0	0	24	0	0	0
ASD between individuals	Mean distance	0.265	0.272	0.267	0.275	0.260	0.272	0.263	0.266	0.264	0.265	0.281	0.279	0.282	0.280	0.278	0.272
	s.d.	*0.006*	*0.003*	*0.003*	*0.002*	*0.005*	*0.002*	*0.004*	*0.005*	*0.004*	*0.005*	*0.004*	*0.002*	*0.002*	*0.006*	*0.006*	*0.008*
Haplotypic heterozygosity	Mean	0.673	0.691	0.681	0.699	0.663	0.692	0.670	0.676	0.673	0.678	0.712	0.710	0.721	0.709	0.709	0.691
	s.d.	*0.017*	*0.013*	*0.018*	*0.015*	*0.019*	*0.015*	*0.016*	*0.020*	*0.018*	*0.019*	*0.013*	*0.013*	*0.014*	*0.014*	*0.011*	*0.017*
ROH limits	Limit A-B (kb)	931	863	791	946	928	808	773	859	847	870	1107	889	813	858	709	1117
	Limit B-C (kb)	2748	2768	2239	2896	2569	2596	2410	2641	2461	2546	3243	2604	2400	2490	2119	3642
Individual F-Median	Median	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.019	0.000	0.015	0.000	0.000	0.000	0.000	0.000	0.022
	s.d.	*0.0066*	*0.0040*	*0.0089*	*0.0129*	*0.0067*	*0.0080*	*0.0030*	*0.0178*	*0.0073*	*0.0134*	*0.0232*	*0.0137*	*0.0122*	*0.0171*	*0.0014*	*0.0263*
	Positive F-Median (%)	33%	22%	48%	15%	43%	14%	10%	83%	46%	90%	48%	19%	29%	35%	11%	79%
Type of parental relatedness	Outbred (%)	76	85	60	85	61	89	90	21	57	17	52	81	74	65	95	21
	2C-type (%)	24	15	40	7	39	10	10	69	43	76	30	15	20	30	5	54
	1C-type (%)	0	0	0	7	0	1	0	10	0	7	19	4	6	5	0	21
	2x1C-type (%)	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
Individuals with a deficit in homozygosity		100% (21/21)	100% (27/27)	96% (24/25)	89% (24/27)	91% (21/23)	76% (61/80)	90% (28/31)	79% (23/29)	93% (26/28)	90% (26/29)	70% (19/27)	89% (24/27)	68% (45/66)	85% (17/20)	100% (19/19)	75% (18/24)
Distances between partners' birth places	Couples in the ethnological dataset (N)	27	29	28	46	23	82	49	25	19	27	52	46	66	31	27	35
	Median (km)	15	0	22	40	13	48	58	46	0	23	0	12	0	0	0	18
	0-95% (km)	0-67	0-26	0-123	0-607	0-149	0-261	0-180	0-252	0-74	0-576	0-163	0-264	0-43	0-3	0-137	0-111
	Exogamous couples (%)	74	41	64	57	52	79	63	84	37	78	33	54	12	6	33	60

CA: Central Asia; NA: Northern Asia (Siberia, Mongolia)

Table S2 - F_{ST} pairwise differences

	Aki	Akz	Bou	Hks	Kaz	Kel	Kib	Mng	Sho	Tab	Tja	Tje	Tjy	Tlg	Tub	Tur	Uzb
Aki	0	0.00784	0.0141	0.01318	0.00759	0.00832	0.00581	0.00694	0.01977	0.0297	0.03801	0.04463	0.06432	0.00674	0.01975	0.02978	0.01233
Akz	0.00784	0	0.01418	0.01083	0.00194	0.01096	0.00194	0.00609	0.01739	0.01716	0.02459	0.02993	0.04887	0.00876	0.01941	0.01772	0.00428
Bou	0.0141	0.01418	0	0.02251	0.01487	0.01218	0.01095	0.00784	0.03033	0.04484	0.05491	0.06336	0.0837	0.01456	0.03369	0.04465	0.02233
Hks	0.01318	0.01083	0.02251	0	0.01016	0.01591	0.00986	0.01458	0.00971	0.02488	0.03199	0.03702	0.05602	0.01436	0.01964	0.02582	0.01225
Kaz	0.00759	0.00194	0.01487	0.01016	0	0.01094	0.00151	0.00639	0.01625	0.01426	0.02113	0.02637	0.04483	0.0087	0.0179	0.01499	0.00221
Kel	0.00832	0.01096	0.01218	0.01591	0.01094	0	0.00767	0.00559	0.02376	0.03876	0.04828	0.0559	0.07631	0.00803	0.0272	0.03847	0.0179
Kib	0.00581	0.00194	0.01095	0.00986	0.00151	0.00767	0	0.00327	0.01628	0.01979	0.02728	0.03317	0.05199	0.00648	0.0184	0.02008	0.00487
Mng	0.00694	0.00609	0.00784	0.01458	0.00639	0.00559	0.00327	0	0.0221	0.03462	0.04399	0.05191	0.07216	0.00739	0.02504	0.0343	0.0134
Sho	0.01977	0.01739	0.03033	0.00971	0.01625	0.02376	0.01628	0.0221	0	0.02865	0.03565	0.04011	0.05865	0.02127	0.02091	0.02996	0.01759
Tab	0.0297	0.01716	0.04484	0.02488	0.01426	0.03876	0.01979	0.03462	0.02865	0	0.00429	0.00472	0.01918	0.03131	0.02869	0.00383	0.00651
Tja	0.03801	0.02459	0.05491	0.03199	0.02113	0.04828	0.02728	0.04399	0.03565	0.00429	0	0.00676	0.02183	0.04004	0.03537	0.00817	0.01245
Tje	0.04463	0.02993	0.06336	0.03702	0.02637	0.0559	0.03317	0.05191	0.04011	0.00472	0.00676	0	0.01975	0.04686	0.03994	0.00932	0.01615
Tjy	0.06432	0.04887	0.0837	0.05602	0.04483	0.07631	0.05199	0.07216	0.05865	0.01918	0.02183	0.01975	0	0.06654	0.05766	0.02475	0.03336
Tlg	0.00674	0.00876	0.01456	0.01436	0.0087	0.00803	0.00648	0.00739	0.02127	0.03131	0.04004	0.04686	0.06654	0	0.02297	0.03143	0.01354
Tub	0.01975	0.01941	0.03369	0.01964	0.0179	0.0272	0.0184	0.02504	0.02091	0.02869	0.03537	0.03994	0.05766	0.02297	0	0.02998	0.01869
Tur	0.02978	0.01772	0.04465	0.02582	0.01499	0.03847	0.02008	0.0343	0.02996	0.00383	0.00817	0.00932	0.02475	0.03143	0.02998	0	0.00801
Uzb	0.01233	0.00428	0.02233	0.01225	0.00221	0.0179	0.00487	0.0134	0.01759	0.00651	0.01245	0.01615	0.03336	0.01354	0.01869	0.00801	0

All pairwise distances were significant with *p-values* = 0.00000 (+-0.0000) for 1023 permutations.

Material and Methods

Population samples

During several field expeditions conducted in Inner Asia between 2001 and 2012, we collected ethno-geographical information from volunteers, who provided information about the place of birth (village name) and home language for themselves, their spouse, their parents, and their parents-in-law. Participants were assigned to populations, defined as groups of individuals living in a similar area and belonging to the same ethnic group, based on the self-reported spoken native language. Note that the Kel population is composed of Northern Asian Kyrgyz from two different close locations. In total, we obtained geographical information for 644 couples from 16 populations (Figure 1).

In addition, DNA was extracted from blood and saliva for 503 of the participants including 332 males (66%) and 171 females (34%) (Table S1). Written informed consents were obtained for all participants.

Geographical exogamy

To infer the geographical distance between the places of birth of spouses, we obtained the geographical coordinates of each location and calculated the pairwise great-circle distances with an own designed python3 code using the Spherical Law of Cosines formula and an Earth radius of 6,367.445 km. We represented the average distances within couples per population using a Kernel's density estimate implemented in R with a smoothing bandwidth of 0.2 (Figure 2). Based on the population local minima on this density plot (comprised between 2 km and 12 km, on average 4 km while Tja, a population that did not have a local minimum, was excluded), we defined exogamous couples as those with spouses born at more than 4 km.

As for the current generation, we distinguished exogamous couples from endogamous couples for the parental generation (*i.e.*, for each sampled individual, the distance between the birth places of his/her parents and his/her parents-in-law, respectively). Note that for three out of 16 populations (Kaz, Tur and Uzb), we had no information for the parental generation. Furthermore, we tested for other definitions of exogamy with arbitrary thresholds at 10, 20, 30, 40, and 50 km.

SNP genotyping and quality control

All individuals were genotyped on an Illumina genotyping array (either 660W-Quad, OmniExpress, Omni1-Quad, Omni2.5, or Omni5Exome). Genotyping was performed by the "Plate-forme Post-Génomique de la Pitié-Salpétrière (P3S)" or by the "Institut Pasteur – Genopole (Génotypage des Eucaryotes)", in Paris, France. For two of the arrays (Illumina Omni1-Quad and Omni2.5), we performed our own genotype-calling quality control whereas for the others, the quality control was done by the sequencing platform. Indeed, for each array separately, we performed a three-stage quality control procedure inspired from (1), restricting our analyses to autosomal SNPs (*i.e.*, excluding CNVs and non-autosomal markers). Briefly, we excluded SNPs with ambiguous genomic position, SNPs that failed genotyping in the sample set, SNPs with a call rate below 90% or a cluster separation below 0.2, SNPs without ID in rs nomenclature, SNPs found to be duplicates, indels and monomorphic SNPs. We removed samples with more than 15% of missing data. Then, we performed a population-genetic quality control following (1). We first

identified pairs of samples more related than first cousins using Relpair 2.0.1 software (2, 3). We performed this analysis on three non-overlapping subsets of SNPs in Hardy-Weinberg equilibrium (*p-value*>10^{-5} for each population and >10^{-2} for each pair of populations) and bimorphic in every population. For pairs of individuals closer than first cousins, we removed the individual with the lowest genotyping-call or the individual involved in the highest number of relations, resulting in 52 individuals being excluded. Finally, we removed the monomorphic SNPs within each array. At the end, for the five arrays, we obtained data for 526,823 to 2,609,107 SNPs genotyped for the 519 samples of interest.

To generate a joint dataset, we manually merged the five datasets based on SNPs rs. To control that alleles are coded on the same strand between arrays, we removed A-T and C-G polymorphisms, and SNPs that were triallelic on the two arrays, resulting in a dataset of 253,606 SNPs. We found three pairs of relatives within the merged dataset, and as previously described, we excluded three samples, in addition to 13 samples with a call-rate < 95%, leading to a total of 503 individuals. 292 of these individuals genotyped on the Omni1 and Omni2.5 arrays have already been published in (4), while the 211 other individuals were newly published in our present study. Finally, we excluded SNPs in Hardy-Weinberg disequilibrium (plink1.9 function --hardy, *p-value*>10^{-5}), leading to a final dataset of 253,532 SNPs. Within this dataset, we generated a subset of 105,858 independent SNPs with r^2<0.5 (plink1.9 function --indep 50 5 2).

Genetic diversity

We computed pairwise F_{ST} distances for each pair of populations on the independent SNPs dataset using ArlSumStat (5) (Table S2). We present only these F_{ST} distances as they are highly correlated to those based on the whole SNP dataset (Spearman's *rho*=0.997, *p-value*<$2.2*10^{-16}$). We also computed the matrix of allele-sharing dissimilarity between all the 503 individuals for the independent SNPs dataset, using the software *asd* (6). We represented both the F_{ST} and ASD pairwise distance matrices as the two first dimensions of MultiDimensional Scaling (MDS) performed with R function *cmdscale* (7). To evaluate whether the first two MDS dimensions represented accurately the whole matrix, we calculated the Spearman's rank-sum correlation *rho* between the MDS distances on dimensions 1 and 2, and their corresponding F_{ST} and ASD distances, respectively.

We calculated the haplotypic heterozygosity of each population based on (1). We first reconstructed non-overlapping blocks of SNPs in high linkage disequilibrium (LD) using the whole SNPs dataset. We defined blocks of LD as intervals of 5-15 SNPs where the recombination rate is below 0.5cM/Mb between each pair of contiguous SNPs (using the GRCh37 recombination map). We generated 2,267 such blocks of LD, with on average 103 blocks per chromosome (from 24 on chromosome 21 to 222 on chromosome 1). These blocks include 13,340 SNPs, *i.e.*, 5.3% of the total dataset.

Then we phased these blocks with plink 1.7 --hap function, that uses an expectation-maximization phasing algorithm. For each population, we estimated the frequencies of the haplotypes (plink 1.7 --hap-freq function), calculated the heterozygosity of each haplotype based on Nei's formula and computed the mean haplotypic heterozygosity along each chromosome: $H_{k\text{-chromosome}} = \frac{\frac{2Npop}{2Npop-1} * \sum_{j=1}^{Nblocks} (1 - \sum_{i=1}^{Nalleles} {f_{i,j}}^2)}{Nblocks}$.

Inbreeding coefficients and parental mating types

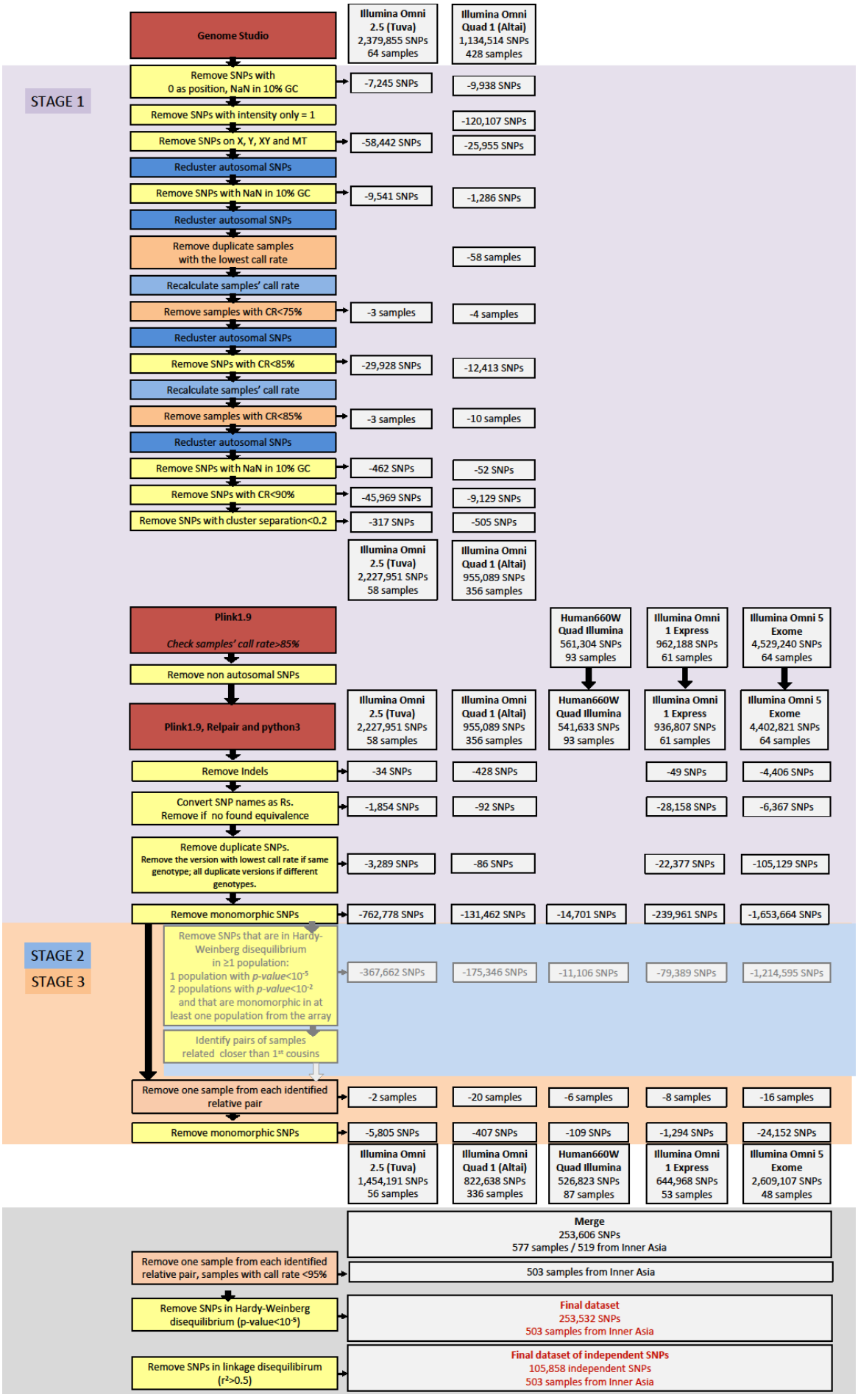
We estimated each individual's inbreeding coefficient (F-Median) (8) using FSuite v1.0.3 (9), by generating 100 submaps of independent SNPs at a threshold of 0.5 cM (10), and computed the median of the 100 coefficients. A co-parameter, called A-Median, was also estimated with FSuite. Moreover, a likelihood ratio test was performed to assign each individual as inbred or outbred. For each individual, based on both F-Median and A-Median, it was possible to estimate the probability to be the descendant of avunculars (AV), double first-cousins (2x1C), first-cousins (1C), second-cousins (2C), or unrelated individuals (OUT; defined as less related than second-cousins). We kept the most likely inbred mating type for each individual. We chose to use FSuite, as it is efficient to infer distant relationships (1C and 2C type) (11), even in small-sized samples (9).

In parallel, as inbreeding leads to a genomic excess of homozygosity, we calculated the proportion of homozygous sites within each individual genome, relative to an expected baseline of homozygosity per population, using plink1.9 --*het* function on the independent SNPs dataset (12). For comparison, we also computed it for the 11 populations from the HapMap3 worldwide dataset. First, we removed 236 first-degree relatives, and SNPs that were non-autosomal, monomorphic or deviating from Hardy-Weinberg equilibrium (*p-value*$<10^{-5}$ in at least one population, and $<10^{-2}$ in at least two populations) (13), to eventually generate a dataset of 195,874 independent SNPs ($r^2<0.5$, plink1.9 function --indep 50 5 2).

Furthermore, we identified runs of homozygosity, called ROHs (14), within each individual genome based on the whole SNPs dataset, using the option --homozyg-snp 50 as in (15). We aimed at distinguishing intermediate ROHs, probably resulting from matings between individuals sharing distant ancestry, from long ROHs likely derived from mating between close relatives (16). First, to be able to compare with other studies from the literature (16, 15, 17), we categorized ROHs based on classical thresholds: between 500 kb and 1,500 kb for intermediate ROHs, *versus* longer than 1,500 kb for long ROHs. However, as Pemberton *et al.* (14) showed that these ROH thresholds can be variable between populations, we also used population-specific categories defined from their ROH size classification method, resulting in what we call intermediate (class B-ROHs) and long ROHs (class C-ROHs). For each of the four ROH classes, we computed the number of ROHs observed within an individual genome and their total length.

Pipeline for quality control and datasets merging.

The pipeline includes three stages of analysis. Briefly, the first step cleans SNPs and samples with a low quality or in duplicate; the second step estimates the genetic relatedness between samples and the third steps remove relatives closer than 1st cousins as well as monomorphic SNPs. Then the five arrays were merged, quality and relatedness were controlled one more time and other individuals included in these arrays but not from Inner Asia were removed. Moreover, we added a Hardy-Weinberg equilibrium test on the SNPs dataset. An additional step generates a subset of independent SNPs.


STAGE 1
STAGE 2
STAGE 3
Genome Studio
Illumina Omni 2.5 (Tuva) 2,379,855 SNPs 64 samples
Illumina Omni Quad 1 (Altai) 1,134,514 SNPs 428 samples
Remove SNPs with 0 as position, NaN in 10% GC
-7,245 SNPs
-9,938 SNPs
Remove SNPs with intensity only = 1
-120,107 SNPs
Remove SNPs on X, Y, XY and MT
-58,442 SNPs
-25,955 SNPs
Recluster autosomal SNPs
Remove SNPs with NaN in 10% GC
-9,541 SNPs
-1,286 SNPs
Recluster autosomal SNPs
Remove duplicate samples with the lowest call rate
-58 samples
Recalculate samples' call rate
Remove samples with CR<75%
-3 samples
-4 samples
Recluster autosomal SNPs
Remove SNPs with CR<85%
-29,928 SNPs
-12,413 SNPs
Recalculate samples' call rate
Remove samples with CR<85%
-3 samples
-10 samples
Recluster autosomal SNPs
Remove SNPs with NaN in 10% GC
-462 SNPs
-52 SNPs
Remove SNPs with CR<90%
-45,969 SNPs
-9,129 SNPs
Remove SNPs with cluster separation<0.2
-317 SNPs
-505 SNPs
Illumina Omni 2.5 (Tuva) 2,227,951 SNPs 58 samples
Illumina Omni Quad 1 (Altai) 955,089 SNPs 356 samples
Plink1.9
Check samples' call rate>85%
Human660W Quad Illumina 561,304 SNPs 93 samples
Illumina Omni 1 Express 962,188 SNPs 61 samples
Illumina Omni 5 Exome 4,529,240 SNPs 64 samples
Remove non autosomal SNPs
Plink1.9, Relpair and python3
Illumina Omni 2.5 (Tuva) 2,227,951 SNPs 58 samples
Illumina Omni Quad 1 (Altai) 955,089 SNPs 356 samples
Human660W Quad Illumina 541,633 SNPs 93 samples
Illumina Omni 1 Express 936,807 SNPs 61 samples
Illumina Omni 5 Exome 4,402,821 SNPs 64 samples
Remove Indels
-34 SNPs
-428 SNPs
-49 SNPs
-4,406 SNPs
Convert SNP names as Rs. Remove if no found equivalence
-1,854 SNPs
-92 SNPs
-28,158 SNPs
-6,367 SNPs
Remove duplicate SNPs. Remove the version with lowest call rate if same genotype; all duplicate versions if different genotypes.
-3,289 SNPs
-86 SNPs
-22,377 SNPs
-105,129 SNPs
Remove monomorphic SNPs
-762,778 SNPs
-131,462 SNPs
-14,701 SNPs
-239,961 SNPs
-1,653,664 SNPs
Remove SNPs that are in Hardy-Weinberg disequilibrium in ≥1 population: 1 population with p-value<10-5 2 populations with p-value<10-2 and that are monomorphic in at least one population from the array
-367,662 SNPs
-175,346 SNPs
-11,106 SNPs
-79,389 SNPs
-1,214,595 SNPs
Identify pairs of samples related closer than 1st cousins
Remove one sample from each identified relative pair
-2 samples
-20 samples
-6 samples
-8 samples
-16 samples
Remove monomorphic SNPs
-5,805 SNPs
-407 SNPs
-109 SNPs
-1,294 SNPs
-24,152 SNPs
Illumina Omni 2.5 (Tuva) 1,454,191 SNPs 56 samples
Illumina Omni Quad 1 (Altai) 822,638 SNPs 336 samples
Human660W Quad Illumina 526,823 SNPs 87 samples
Illumina Omni 1 Express 644,968 SNPs 53 samples
Illumina Omni 5 Exome 2,609,107 SNPs 48 samples
Merge
253,606 SNPs
577 samples / 519 from Inner Asia
Remove one sample from each identified relative pair, samples with call rate <95%
503 samples from Inner Asia
Remove SNPs in Hardy-Weinberg disequilibrium (p-value<10-5)
Final dataset
253,532 SNPs
503 samples from Inner Asia
Remove SNPs in linkage disequilibrium (r2>0.5)
Final dataset of independent SNPs
105,858 independent SNPs
503 samples from Inner Asia

1. Verdu P, et al. (2014) Patterns of Admixture and Population Structure in Native Populations of Northwest North America. *PLoS Genet* 10(8):e1004530. doi:10.1371/journal.pgen.1004530.
2. Michael Boehnke NJC (1997) Accute Inference of Relationships in Sib-Pair Linkage Studies. *Am J Hum Genet* 61:423–429.
3. Epstein MP, Duren WL, Boehnke M (2000) Improved inference of relationship for pairs of individuals. *Am J Hum Genet* 67(5):1219–31. doi:10.1016/S0002-9297(07)62952-8.
4. Damgaard P, et al. Population genomic history of the Eurasian steppe. *Under soumission*.
5. Excoffier L, Lischer HEL (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Resour* 10(3):564–567. doi:10.1111/j.1755-0998.2010.02847.x.
6. Szpiech ZA (2011) asd.
7. R Core Team (2014) R: A Language and Environment for Statistical Computing.
8. Leutenegger A-L, et al. (2003) Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* 73(3):516–23. doi:10.1086/378207.
9. Gazal S, Sahbatou M, Babron M-C, Génin E, Leutenegger A-L (2014) FSuite: exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics* 30(13):1940–1941. doi:10.1093/bioinformatics/btu149.
10. Leutenegger A-L, Sahbatou M, Gazal S, Cann HM, Génin E (2011) Consanguinity around the world: what do the genomic data of the HGDP-CEPH diversity panel tell us? *Eur J Hum Genet* 19(5):583–7. doi:10.1038/ejhg.2010.205.
11. Gazal S, Génin E, Leutenegger A-L (2015) Relationship inference from the genetic data on parents or offspring: A comparative study. *Theor Popul Biol*. doi:10.1016/j.tpb.2015.09.002.
12. Purcell S, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* 81(3):559–75. doi:10.1086/519795.
13. Pemberton TJ, Wang C, Li JZ, Rosenberg N a. (2010) Inference of unexpected genetic relatedness among individuals in HapMap phase III. *Am J Hum Genet* 87(4):457–464. doi:10.1016/j.ajhg.2010.08.014.
14. Pemberton TJ, et al. (2012) Genomic patterns of homozygosity in worldwide human populations. *Am J Hum Genet* 91(2):275–292. doi:10.1016/j.ajhg.2012.06.014.
15. Joshi PK, et al. (2015) Directional dominance on stature and cognition in diverse human populations. *Nature* 523(7561):459–62. doi:10.1038/nature14618.
16. McQuillan R, et al. (2008) Runs of Homozygosity in European Populations. *Am J Hum Genet* 83(3):359–372. doi:10.1016/j.ajhg.2008.08.007.
17. Kirin M, et al. (2010) Genomic runs of homozygosity record population history and consanguinity. *PLoS One* 5(11):e13996. doi:10.1371/journal.pone.0013996.

Références bibliographiques

Abilev S., Malyarchuk B., Derenko M., Wozniak M., Grzybowski T., Zakharov I. (2012) The Y-chromosome C3* star-cluster attributed to Genghis Khan's descendants is present at high frequency in the Kerey clan from Kazakhstan. *Human biology*, **84**, 79–89.

Adhikari K., Mendoza-Revilla J., Chacón-Duque J.C., Fuentes-Guajardo M., Ruiz-Linares A. (2016) Admixture in Latin America. *Current Opinion in Genetics & Development*, **41**, 106–114.

Adle C., Habib I., Baipakov K.M. (2003) *History of Civilizations of Central Asia, Volume V : Development in contrast : from the sixteenth to the mid-nineteenth century.* UNESCO.

Adle C., Palat M.K., Tabyshalieva A. (2005) *History of civilizations of Central Asia, Volume VI : Towards contemporary civilization : from the mid-nineteenth century to the present time.* UNESCO.

Ager S.L. (2005) Familiarity breeds : incest and the Ptolemaic dynasty. *The Journal of Hellenic Studies*, **125**, 1–34.

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

Aimé C., Laval G., Patin E. *et al.* (2013) Human genetic data reveal contrasting demographic patterns between sedentary and nomadic populations that predate the emergence of farming. *Molecular biology and evolution*, **30**, 2629–44.

Aimé C., Verdu P., Ségurel L. *et al.* (2014) Microsatellite data show recent demographic expansions in sedentary but not in nomadic human populations in Africa and Eurasia. *European Journal of Human Genetics*, **22**, 1201–1207.

Akimova E., Higham T., Stasyuk I., Buzhilova A., Dobrovolskaya M., Mednikova M. (2010) A new direct radiocarbon ams date for an upper palaeolithic human bone from siberia. *Archaeometry*, **52**, 1122–1130.

Al-Gazali L., Hamamy H., Al-Arrayad S. (2006) Genetic disorders in the Arab world. *BMJ (Clinical research ed.)*, **333**, 831–834.

Alexander D.H., Novembre J., Lange K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, **19**, 1655–64.

Allentoft M.E., Sikora M., Sjögren K.G. *et al.* (2015) Population genomics of Bronze Age Eurasia. *Nature*, **522**, 167–172.

Alvarez G., Ceballos F.C., Quinteiro C. (2009) The Role of Inbreeding in the Extinction of a European Royal Dynasty. *PLoS ONE*, **4**, 1–7.

Alvarez G., Quinteiro C., Ceballos F. (2011) Inbreeding and Genetic Disorder. In *Advances in the Study of Genetic Disorders*, chap. 2, pp. 21–41. InTech.

Alvergne A., Faurie C., Raymond M. (2009) Father-offspring resemblance predicts paternal investment in humans. *Animal Behaviour*, **78**, 61–69.

Alvergne A., Perreau F., Mazur A., Mueller U., Raymond M. (2014) Identification of visual paternity cues in humans. *Biology Letters*, **10**, 20140063–20140067.

Anthony D.W. (2010) *The horse, the wheel, and language : how Bronze-Age riders from the Eurasian steppes shaped the modern world*. Princeton University Press.

Arandjelovic M., Head J., Boesch C., Robbins M.M., Vigilant L. (2014) Genetic inference of group dynamics and female kin structure in a western lowland gorilla population (Gorilla gorilla gorilla). *Primate Biology*, **1**, 29–38.

Archie E.A., Hollister-Smith J.A., Poole J.H. *et al.* (2007) Behavioural inbreeding avoidance in wild African elephants. *Molecular Ecology*, **16**, 4138–4148.

Austerlitz F., Heyer E. (1998) Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. *Proceedings of the National Academy of Sciences of the United States of America*, **95**, 15140–15144.

Balaresque P., Manni F., Dugoujon J.M., Crousau-Roy B., Heyer E. (2006) Estimating sex-specific processes in human populations : Are XY-homologous markers an effective tool ? *Heredity*, **96**, 214–21.

Balaresque P., Poulet N., Cussat-Blanc S. *et al.* (2015) Y-chromosome descent clusters and male differential reproductive success : young lineage expansions dominate Asian pastoral nomadic populations. *European Journal of Human Genetics*, **23**, 1413–1422.

Balloux F., Amos W., Coulson T. (2004) Does heterozygosity estimate inbreeding in real populations ? *Molecular Ecology*, **13**, 3021–3031.

Bamshad M.J., Watkins W.S., Dixon M.E. *et al.* (1998) Female gene flow stratifies Hindu castes. *Nature*, **395**, 651–652.

Banks S.C., Skerratt L.F., Taylor A.C. (2002) Female dispersal and relatedness structure in common wombats (Vombatus ursinus). *Journal of Zoology*, **256**, 389–399.

Barbujani G. (2012) Human Genetics : Message from the Mesolithic. *Current Biology*, **22**, R631–R633.

Barth F. (1969) Ethnic groups and boundaries. The social organization of culture difference. (Results of a symposium held at the University of Bergen, 23rd to 26th February 1967). *Bergen, London : Universitetsforlaget ;Allen & Unwin*.

Batini C., Hallast P., Zadik D. *et al.* (2015) Large-scale recent expansion of European patrilineages shown by population resequencing. *Nature Communications*, **6**, 1–8.

Behar D.M., Villems R., Soodyall H. *et al.* (2008) The Dawn of Human Matrilineal Diversity. *American Journal of Human Genetics*, **82**, 1130–1140.

Ben M'Rad L., Chalbi N. (2006) Milieu de résidence origine des conjoints et consanguinité en Tunisie. *Antropo*, pp. 63–71.

Bengtsson B.O., Jacquard A. (1976) Loi de distribution des homozygotes identiques dans une population. *Population (french edition)*, pp. 63–71.

Bereni L., Chauvin S., Revillard A., Jaunait A. (2012) *Introduction aux études sur le genre*, vol. 2. De Boeck Université.

Berniell-Lee G., Calafell F., Bosch E. *et al.* (2009) Genetic and Demographic Implications of the Bantu Expansion : Insights from Human Paternal Lineages. *Molecular Biology and Evolution*, **26**, 1581–1589.

Besaggio D., Fuselli S., Srikummool M. *et al.* (2007) Genetic variation in Northern Thailand Hill Tribes : origins and relationships with social structure and linguistic differences. *BMC evolutionary biology*, **7 Suppl 2**, S12 (1–10).

Bideau A., Brunet G., Heyer E., Plauchu H. (1994) La consanguinité, révélateur de la structure de la population. L'exemple de la vallée de la Valserine du XVIIIe siècle à nos jours. *Population (french edition)*, **49**, 145–160.

Bittles A. (2001) A background summary of consanguineous marriage. *Center for Human Genetics, Edith Cowan University, Perth*, **1**, 1–10.

Bittles A. (2008) Consanguinity and its relevance to clinical genetics. *Clinical Genetics*, **60**, 89–98.

Bittles A.H. (2005) Endogamy, consanguinity and community disease profiles. *Community Genetics*, **8**, 17–20.

Bittles A.H. (2012) *Consanguinity in context*, vol. 63. Cambridge University Press.

Bittles A.H., Black M.L. (2010) Consanguinity, human evolution, and complex diseases. *Proceedings of the National Academy of Sciences*, **107**, 1779–1786.

Bittles A.H., Grant J.C., Sullivan S.G., Hussain R. (2002) Does inbreeding lead to decreased human fertility ? *Annals of Human Biology*, **29**, 111–130.

Blouin S.F., Blouin M. (1988) Inbreeding avoidance behaviors. *Trends in Ecology and Evolution*, **3**, 230–233.

Blum M.G.B., Heyer E., François O., Austerlitz F. (2006) Matrilineal Fertility Inheritance Detected in Hunter–Gatherer Populations Using the Imbalance of Gene Genealogies. *PLoS Genetics*, **2**, 1138–1146.

Bosch E., Calafell F., Rosser Z.H. *et al.* (2003) High level of male-biased Scandinavian admixture in Greenlandic Inuit shown by Y-chromosomal analysis. *Human genetics*, **112**, 353–363.

Boyce A.J., Küchemann C.F., Harrison G.A. (1967) Neighbourhood knowledge and the distribution of marriage distances. *Annals of Human Genetics*, **30**, 335–338.

Broman K.W., Weber J.L. (1999) Long Homozygous Chromosomal Segments in Reference Families from the Centre d'Étude du Polymorphisme Humain. *The American Journal of Human Genetics*, **65**, 1493–1500.

Browning S.R., Browning B.L. (2010) High-Resolution Detection of Identity by Descent in Unrelated Individuals. *American Journal of Human Genetics*, **86**, 526–539.

Brudner L.A., White D.R. (1997) Class, property, and structural endogamy : Visualizing networked histories. *Theory and Society*, **26**, 161–208.

Buri P. (1956) Gene Frequency in Small Populations of Mutant Drosophila. *Society for the Study of Evolution*, **10**, 367–402.

Burland T.M., Barratt E.M., Nichols R.A., Racey P.A. (2001) Mating patterns, relatedness and the basis of natal philopatry in the brown long-eared bat, Plecotus auritus. *Molecular Ecology*, **10**, 1309–1321.

Burton M.L., Moore C.C., Romney A.K. *et al.* (1996) Regions Based on Social Structure. *Current Anthropology*, **37**, 87–123.

Cattell R.B. (1966) The Scree Test For The Number Of Factors. *Multivariate Behavioral Research*, **1**, 245–276.

Cavalli-Sforza L.L. (1997) Genes, peoples, and languages. *Proceedings of the National Academy of Sciences*, **94**, 7719–7724.

Cavalli-Sforza L.L., Feldman M.W. (2003) The application of molecular genetic approaches to the study of human evolution. *Nature Genetics*, **33**, 266–275.

Cavalli-Sforza L.L., Menozzi P., Piazza A. (1993) Demic expansions and human evolution. *Science (New York, N.Y.)*, **259**, 639–46.

Cavalli-Sforza L.L., Menozzi P., Piazza A. (1994) *The history and geography of human genes*. Princeton university press.

Cazes M.H. (1980) Hasard et sélection dans les populations d'effectif limité. *Population (french edition)*, pp. 417–435.

Cazes M.H. (1981) Les echanges matrimoniaux chez les Dogons de Tabi. Absence d'effet statistique global des unions dites "préférentielles". *Population (French Edition)*, **36**, 1069.

Chaix R., Austerlitz F., Hegay T., Quintana-Murci L., Heyer E. (2008) Genetic traces of east-to-west human expansion waves in Eurasia. *American journal of physical anthropology*, **136**, 309–17.

Chaix R., Austerlitz F., Khegay T. *et al.* (2004) The Genetic or Mythical Ancestry of Descent Groups : Lessons from the Y Chromosome. *The American Journal of Human Genetics*, **75**, 1113–1116.

Chaix R., Quintana-Murci L., Hegay T. *et al.* (2007) From Social to Genetic Structures in Central Asia. *Current Biology*, **17**, 43–48.

Chang C.C., Chow C.C., Tellier L.C., Vattikuti S., Purcell S.M., Lee J.J. (2015) Second-generation PLINK : rising to the challenge of larger and richer datasets. *GigaScience*, **4**, 7.

Charlesworth B. (2009) Fundamental concepts in genetics : Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, **10**, 195–205.

Charlesworth D., Willis J.H. (2009) The genetics of inbreeding depression. *Nature Reviews Genetics*, **10**, 783–796.

Charpentier M.J.E., Widdig A., Alberts S.C. (2007) Inbreeding depression in non-human primates : A historical review of methods used and empirical data. *American Journal of Primatology*, **69**, 1370–1386.

Chiaroni J., Underhill P.A., Cavalli-Sforza L.L. (2009) Y chromosome diversity, human expansion, drift, and cultural evolution. *Proceedings of the National Academy of Sciences*, **106**, 20174–20179.

Clutton-Brock T.H., Lukas D. (2012) The evolution of social philopatry and dispersal in female mammals. *Molecular Ecology*, **21**, 472–492.

Coffin H.R., Watters J.V., Mateo J.M. (2011) Odor-based recognition of familiar and related conspecifics : A first test conducted on captive Humboldt penguins (Spheniscus humboldti). *PLoS ONE*, **6**, 10–13.

Comas D., Plaza S., Wells R.S. *et al.* (2004) Admixture, migrations, and dispersals in Central Asia : evidence from maternal DNA lineages. *European Journal of Human Genetics*, **12**, 495–504.

Consortium Multitree (2014) MultiTree : A digital library of language relationships.

Creanza N., Feldman M.W. (2016) Worldwide genetic and cultural change in human evolution. *Current Opinion in Genetics & Development*, **41**, 85–92.

Dani A.H., Masson V.M. (1994) *History of Civilizations of Central Asia, Volume I : The dawn of civilization : earliest times to 700 B.C.*

Darwin C. (1876) *The effects of cross and self fertilisation in the vegetable kingdom.* John Murray.

Darwin C. (1889) *Journal of Researches Into the Natural History and Geology of the Countries Visited During the Voyage of HMS" Beagle" Round the World, Under the Command of Capt. Fitz Roy*, vol. 2. Ward, Lock and Company.

Davis-Kimball J., Bashilov V.A., Yablonsky L.T. (1995) *Nomads of the Eurasian Steppes in the Early Iron Age.* Zinat Press Berkeley, CA.

DeBruine L.M., Jones B.C., Little A.C., Perrett D.I. (2008) Social perception of facial resemblance in humans. *Archives of Sexual Behavior*, **37**, 64–77.

Derenko M., Malyarchuk B., Denisova G. *et al.* (2012) Complete Mitochondrial DNA Analysis of Eastern Eurasian Haplogroups Rarely Found in Populations of Northern Asia and Eastern Europe. *PLoS ONE*, **7**, e32179 (1–12).

Derenko M., Malyarchuk B., Denisova G.a. *et al.* (2006) Contrasting patterns of Y-chromosome variation in South Siberian populations from Baikal and Altai-Sayan regions. *Human genetics*, **118**, 591–604.

Derenko M.V., Grzybowski T., Malyarchuk B.a. *et al.* (2003) Diversity of mitochondrial DNA lineages in South Siberia. *Annals of human genetics*, **67**, 391–411.

Destro-Bisol G., Capocasa M., Anagnostou P. (2012) When gender matters : new insights into the relationships between social systems and the genetic structure of human populations. *Molecular ecology*, **21**, 4917–4920.

Destro-Bisol G., Donati F., Coia V. *et al.* (2004) Variation of female and male lineages in sub-Saharan populations : The importance of sociocultural factors. *Molecular Biology and Evolution*, **21**, 1673–1682.

Destro-Bisol G., Jobling M.A., Rocha J. *et al.* (2010) Molecular anthropology in the genomic era. *Journal of Anthropological Sciences*, **88**, 93–112.

DeWeese D. (2010) *Islamization and Native Religion in the Golden Horde : Baba Tjkles and Conversion to Islam in Historical and Epic Tradition.* Pennsylvania State University Press.

Di Cristofaro J., Pennarun E., Mazières S. *et al.* (2013) Afghan Hindu Kush : Where Eurasian Sub-Continent Gene Flows Converge. *PLoS ONE*, **8**, e76748 (1–12).

Dray S., Dufour A.B. (2007) The ade4 Package : Implementing the Duality Diagram for Ecologists. *Journal of Statistical Software*, **22**, 1 – 20.

Dulik M.C., Osipova L.P., Schurr T.G. (2011) Y-Chromosome Variation in Altaian Kazakhs Reveals a Common Paternal Gene Pool for Kazakhs and the Influence of Mongolian Expansions. *PLoS ONE*, **6**, e17548 (1–12).

Dulik M.C., Zhadanov S.I., Osipova L.P. *et al.* (2012) Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. *American journal of human genetics*, **90**, 229–46.

Epstein M.P., Duren W.L., Boehnke M. (2000) Improved inference of relationship for pairs of individuals. *American journal of human genetics*, **67**, 1219–31.

Excoffier L., Laval G., Schneider S. (2007) Arlequin (version 3.0) : an integrated software package for population genetics data analysis. *Evolutionary bioinformatics online*, **1**, 47–50.

Excoffier L., Lischer H.E.L. (2010) Arlequin suite ver 3.5 : a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular Ecology Resources*, **10**, 564–567.

Excoffier L., Smouse P., Quattro J. (1992) Analysis of Molecular Variance Inferred From Metric Distances Among DNA Haplotypes : Application to Human Mitochondrial DNA Restriction Data. *Genetics*, **491**, 479–491.

Fan H., Chu J.Y. (2007) A Brief Review of Short Tandem Repeat Mutation. *Genomics, Proteomics & Bioinformatics*, **5**, 7–14.

Fareed M., Afzal M. (2014) Evidence of inbreeding depression on height, weight, and body mass index : A population-based child cohort study. *American Journal of Human Biology*, **26**, 784–795.

Fareed M., Ahmad M.K., Azeem Anwar M., Afzal M. (2016) Impact of consanguineous marriages and degrees of inbreeding on fertility, child mortality, secondary sex ratio, selection intensity and genetic load : a cross-sectional study from Northern India. *Pediatric Research.*

Fix A.G. (1999) *Migration and colonization in human microevolution.* Cambridge University Press.

Fix A.G. (2004) Kin-structured migration : Causes and consequences. *American Journal of Human Biology*, **16**, 387–394.

Francfort H.P., Grenet F. (2017) Asie centrale (in Encyclopædia Universalis [en ligne]).

Frankham R. (2005) Genetics and extinction. *Biological Conservation*, **126**, 131–140.

Fu Q., Li H., Moorjani P. *et al.* (2014) Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, pp. 8–13.

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

Gagnon A., Toupance B., Tremblay M., Beise J., Heyer E. (2006) Transmission of migration propensity increases genetic divergence between populations. *American Journal of Physical Anthropology*, **129**, 630–636.

Galippe V. (1905) *L'hérédité des stigmates de dégénérescence et les familles souveraines*. Masson.

Gamba C., Jones E.R., Teasdale M.D. *et al.* (2014) Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, **5**, 5257.

Gao X., Martin E.R. (2009) Using allele sharing distance for detecting human population stratification. *Human Heredity*, **68**, 182–191.

Gao Z., Waggoner D., Stephens M., Ober C., Przeworski M. (2015) An Estimate of the Average Number of Recessive Lethal Mutations Carried by Humans. *Genetics*, **199**, 1243–1254.

Garrod A.E. (1902) The incidence of alkaptonuria : a study in chemical individuality. *The Lancet*, **160**, 1616–1620.

Gascuel O. (1997) BIONJ : an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, **14**, 685–695.

Gazal S. (2014) *Consanguinity in the High-Throughput Genome Era : Estimations and Applications*. Theses, Université Paris Sud - Paris XI.

Gazal S., Sahbatou M., Babron M.C., Génin E., Leutenegger A.L. (2014a) FSuite : exploiting inbreeding in dense SNP chip and exome data. *Bioinformatics*, **30**, 1940–1941.

Gazal S., Sahbatou M., Babron M.C., Génin E., Leutenegger A.L. (2015) High level of inbreeding in final phase of 1000 Genomes Project. *Scientific Reports*, **5**, 1–7.

Gazal S., Sahbatou M., Perdry H., Letort S., Genin E., Leutenegger A.L. (2014b) Inbreeding coefficient estimation with dense SNP data : comparison of strategies and application to HapMap III. *Human heredity*, **77**, 49–62.

Génin E., Todorov A.A. (2006) Homozygosity Mapping. In *Encyclopedia of Life Sciences*, pp. 1–5. John Wiley & Sons, Ltd, Chichester.

Ghasarian C. (1996) *Introduction à l'étude de la parenté*. Éditions du Seuil.

Giles R.E., Blanc H., Cann H.M., Wallace D.C. (1980) Maternal inheritance of human mitochondrial DNA. *Proceedings of the National Academy of Sciences of the United States of America*, **77**, 6715–9.

Godelier M. (2004) *Métamorphoses de la parenté*, vol. 682. Fayard Paris.

Goldberg A., Günther T., Rosenberg N.A., Jakobsson M. (2017) Ancient X chromosomes reveal contrasting sex bias in Neolithic and Bronze Age Eurasian migrations. *Proceedings of the National Academy of Sciences*, **114**, 2657–2662.

Goldberg A., Verdu P., Rosenberg N.A. (2014) Autosomal Admixture Levels Are Informative About Sex Bias in Admixed Populations. *Genetics*, **198**, 1209–1229.

Goldstein D.B., Pollock D.D. (1997) Launching Microsatellites : A Review of Mutation Processes and Methods of Phylogenetic Inference. *Journal of Heredity*, **88**, 335–342.

González-Ruiz M., Santos C., Jordana X. *et al.* (2012) Tracing the Origin of the East-West Population Admixture in the Altai Region (Central Asia). *PLoS ONE*, **7**, e48904 (1–11).

Grant P.R., Grant B.R. (2011) *How and why species multiply : the radiation of Darwin's finches.* Princeton University Press.

Green J.P., Holmes A.M., Davidson A.J. *et al.* (2015) The Genetic Basis of Kin Recognition in a Cooperatively Breeding Mammal. *Current Biology*, **25**, 2631–2641.

Green R.E., Krause J., Briggs A.W. *et al.* (2010) A Draft Sequence of the Neandertal Genome. *Science*, **328**, 710–722.

Greenwood P.J. (1980) Mating systems, philopatry and dispersal in birds and mammals. *Animal Behaviour*, **28**, 1140–1162.

Grousset R. (1970) *The empire of the steppes : a history of Central Asia.* Rutgers University Press.

Guillot E.G., Cox M.P. (2014) SMARTPOP : inferring the impact of social dynamics on genetic diversity through high speed simulations. *BMC Bioinformatics*, **15**, 175.

Gunnarsdóttir E.D., Nandineni M.R., Li M. *et al.* (2011) Larger mitochondrial DNA than Y-chromosome differences between matrilocal and patrilocal groups from Sumatra. *Nature Communications*, **2**, 1–6.

Haak W., Lazaridis I., Patterson N. *et al.* (2015) Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, **522**, 207–211.

Hagaman R.M., Elias W.S., Netting R.M. (1978) The genetic and demographic impact of in-migrants in a largely endogamous community. *Annals of human biology*, **5**, 505–515.

Hain T.J.A., Neff B.D. (2007) Multiple paternity and kin recognition mechanisms in a guppy population. *Molecular Ecology*, **16**, 3938–3946.

Hamamy H. (2012) Consanguineous marriages preconception consultation in primary health care settings. *Journal of Community Genetics*, **3**, 185–192.

Hamilton G., Stoneking M., Excoffier L. (2005) Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 7476–7480.

Hamilton J., Vonk J. (2015) Do dogs (Canis lupus familiaris) prefer family ? *Behavioural Processes*, **119**, 123–134.

Hamilton W.D. (1987) Kinship, recognition, disease, and intelligence : constraints of social evolution. *Animal societies : theories and facts*, pp. 81–102.

Han L., Abney M. (2013) Using identity by descent estimation with dense genotype data to detect positive selection. *European Journal of Human Genetics*, **21**, 205–211.

Harmatta J., Puri B.N., Etemadi G.F. (1994) *History of Civilizations of Central Asia, Volume II : The Development of Sedentary and Nomadic Civilizations, 700 B. C. to A.*

Hauber M.E., Sherman P.W. (2001) Self-referent phenotype matching : Theoretical considerations and empirical evidence. *Trends in Neurosciences*, **24**, 609–616.

Hedrick P.W. (2007) Sex : Differences in mutation, recombination, selection, gene flow, and genetic drift. *Evolution*, **61**, 2750–2771.

Hedrick P.W., Garcia-Dorado A. (2016) Understanding Inbreeding Depression, Purging, and Genetic Rescue. *Trends in Ecology and Evolution*, **31**, 940–952.

Helgason A., Hrafnkelsson B., Gulcher J.R., Ward R., Stefánsson K. (2003) A Populationwide Coalescent Analysis of Icelandic Matrilineal and Patrilineal Genealogies : Evidence for a Faster Evolutionary Rate of mtDNA Lineages than Y Chromosomes. *The American Journal of Human Genetics*, **72**, 1370–1388.

Helgason A., Palsson S., Guthbjartsson D.F., Kristjansson t., Stefansson K. (2008) An Association Between the Kinship and Fertility of Human Couples. *Science*, **319**, 813–816.

Henrich J., Boyd R. (1998) The Evolution of Conformist Transmission and the Emergence of Between-Group Differences. *Evolution and Human Behavior*, **19**, 215–241.

Héran F. (2014) Générations sacrifiées : le bilan démographique de la Grande Guerre. *Population & Sociétés (INED)*, pp. 1–4.

Héritier F. (1996) Masculin/Féminin. La pensée de la différence. *Paris, Odile Jacob*, **1**, 329.

Hewlett B.S. (1987) Sexual selection and paternal investment among Aka pygmies. *Human Reproductive Behaviour*, pp. 263–276.

Hey J. (2010) The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. *Molecular Biology and Evolution*, **27**, 921–933.

Heyer E. (1993) Population structure and immigration ; a study of the Valserine Valley (French Jura) from the 17th century until the present. *American journal of human genetics*, **20**, 565–573.

Heyer E. (1995) Mitochondrial and nuclear genetic contribution of female founders to a contemporary population in northeast Quebec. *American journal of human genetics*, pp. 1450–1455.

Heyer E., Balaresque P., Jobling M.A. *et al.* (2009) Genetic diversity and the emergence of ethnic groups in Central Asia. *BMC genetics*, **10**, 1–8.

Heyer E., Brandenburg J.T., Leonardi M. *et al.* (2015) Patrilineal populations show more male transmission of reproductive success than cognatic populations in Central Asia, which reduces their genetic diversity. *American Journal of Physical Anthropology*, **157**, 537–543.

Heyer E., Cazes M.H. (1999) Les "enfants utiles" : Une mesure démographique pour la génétique des populations. *Population (french edition)*, **54**, 677–691.

Heyer E., Chaix R., Pavard S., Austerlitz F. (2012) Sex-specific demographic behaviours that shape human genomic variation. *Molecular ecology*, **21**, 597–612.

Heyer E., Mennecier P. (2009) Genetic and linguistic diversity in Central Asia. In *Becoming Eloquent* (edited by F D'Errico, JM Hombert), pp. 163–180. John Benjamins Publishing Company, Amsterdam.

Heyer E., Sibert A., Austerlitz F. (2005) Cultural transmission of fitness : genes take the fast lane. *Trends in Genetics*, **21**, 234–239.

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

Heyer E., Tremblay M., Desjardins B. (1997) Seventeenth-century European origins of hereditary diseases in the Saguenay population (Quebec, Canada). *Human biology*, **69**, 209–225.

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

Holden C.J., Mace R. (2003) Spread of cattle led to the loss of matrilineal descent in Africa : A coevolutionary analysis. *Proceedings of the Royal Society B : Biological Sciences*, **270**, 2425–2433.

Hollard C., Keyser C., Giscard P.H. *et al.* (2014) Strong genetic admixture in the Altai at the Middle Bronze Age revealed by uniparental and ancestry informative markers. *Forensic Science International : Genetics*, **12**, 199–207.

Holmes W.G., Sherman P.W. (1982) The Ontogeny of Kin Recognition in Two Species of Ground Squirrels. *American Zoologist*, **22**, 491–517.

Hoogland J.L. (1982) Prairie Dogs Avoid Extreme Inbreeding. *Science*, **215**, 1639–1641.

Hostetler J.A. (1985) History and relevance of the Hutterite population for genetic studies. *American journal of medical genetics*, **22**, 453–462.

Hublin J.J., Ben-Ncer A., Bailey S.E. *et al.* (2017) New fossils from Jebel Irhoud, Morocco and the pan-African origin of Homo sapiens. *Nature*, **546**, 289–292.

Hünemeier T., Amorim C.E.G., Azevedo S. *et al.* (2012) Evolutionary Responses to a Constructed Niche : Ancient Mesoamericans as a Model of Gene-Culture Coevolution. *PLoS ONE*, **7**, e38862 (1–10).

Hussain R., Bittles A.H. (1998) The prevalence and demographic characteristics of consanguineous marriages in Pakistan. *Journal of biosocial science*, **30**, 261–275.

Innocenti P., Morrow E.H. (2010) The Sexually Antagonistic Genes of Drosophila melanogaster. *PLoS Biology*, **8**, e1000335 (1–10).

Jacquard A. (1968a) Évolution des populations d'effectif limité. *Population (french edition)*, **23**, 279–300.

Jacquard A. (1968b) Panmixie et consanguinité. Quelques précisions de langage. *Population (french edition)*, **23**, 1065–1090.

Jacquard A. (1971a) Effect of exclusion of sib-mating on genetic drift. *Theoretical Population Biology*, **2**, 91–99.

Jacquard A. (1971b) Effet de la subdivision d'une population sur les variances et covariances des fréquences géniques des sous-populations. *Génétique et populations*, p. 141.

Jacquard A. (1972) Évolution du patrimoine génétique des Kel Kummer. *Population*, **4**, 784–800.

Jacquard A. (1974) Some Studies of Human Populations. In *The Genetic Structure of Populations* (edited by Springer-Verlag), chap. 16, pp. 494–532. Springer Berlin Heidelberg, Berlin, Heidelberg, biomathema edn..

Jacquard A., Reynès F. (1968) Mesure démographique du fardeau génétique. *Population (french edition)*, **23**, 625–648.

Jacquesson S. (2002) Parcours ethnographiques dans l'histoire des deltas. *Cahiers d'Asie centrale*, pp. 51–92.

Jakobi L., Jacquard A. (1971) Consanguinité proche, consanguinité éloignée. Essai de mesure dans un village breton. *Cahier INED*, **60**, 263–268.

Jakobsson M., Rosenberg N.A. (2007) CLUMPP : A cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics*, **23**, 1801–1806.

Jobling M., Hollox E., Kivisild T., Tyler-Smith C. (2013) *Human Evolutionary Genetics*. 2nd edn..

Jobling M.a., Tyler-Smith C. (2003) The human Y chromosome : an evolutionary marker comes of age. *Nature Reviews Genetics*, **4**, 598–612.

Johanson L. (1998) The history of Turkic. *The Turkic Languages*, pp. 81–125.

Jordan F.M., Gray R.D., Greenhill S.J., Mace R. (2009) Matrilocal residence is ancestral in Austronesian societies. *Proceedings of the Royal Society B : Biological Sciences*, **276**, 1957–1964.

Jorde L.B. (2001) Consanguinity and prereproductive mortality in the Utah Mormon population. *Human Heredity*, **52**, 61–65.

Joshi P.K., Esko T., Mattsson H. *et al.* (2015) Directional dominance on stature and cognition in diverse human populations. *Nature*, **523**, 459–62.

Karafet T.M., Mendez F.L., Meilerman M.B., Underhill P.A., Zegura S.L., Hammer M.F. (2008) New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Research*, **18**, 830–838.

Karmin M., Saag L., Vicente M. *et al.* (2015) A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Research*, **25**, 459–466.

Kayser M., Brauer S., Weiss G. *et al.* (2003) Reduced Y-Chromosome, but Not Mitochondrial DNA, Diversity in Human Populations from West New Guinea. *The American Journal of Human Genetics*, **72**, 281–302.

Kayser M., Choi Y., Van Oven M. *et al.* (2008) The impact of the Austronesian expansion : Evidence from mtDNA and Y chromosome diversity in the Admiralty Islands of melanesia. *Molecular Biology and Evolution*, **25**, 1362–1374.

Keller L., Waller D.M. (2002) Inbreeding effects in wild populations. *Trends in Ecology & Evolution*, **17**, 230–241.

Keyser C., Bouakaze C., Crubézy E. *et al.* (2009) Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Human Genetics*, **126**, 395–410.

Kimura M., Ohta T. (1969) The average number of generations until fixation of a mutant gene in a finite population. *Genetics*, **61**, 763–771.

Kirin M., McQuillan R., Franklin C.S., Campbell H., McKeigue P.M., Wilson J.F. (2010) Genomic Runs of Homozygosity Record Population History and Consanguinity. *PLoS ONE*, **5**, e13996 (1–7).

Kloss-Brandstätter A., Pacher D., Schönherr S. *et al.* (2011) HaploGrep : a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Human mutation*, **32**, 25–32.

Knight H.M., Maclean A., Irfan M. *et al.* (2008) Homozygosity mapping in a family presenting with schizophrenia, epilepsy and hearing impairment. *European Journal of Human Genetics*, **16**, 750–758.

Kottak C. (2008) *Mirror for Humanity : A Concise Introduction to Cultural Anthropology.* McGraw-Hill, 7th edn..

Krackow S., Matuschak B. (1991) Mate choice for non-siblings in wild house mice : evidence from a choice test and a reproductive test. *Ethology*, **88**, 99–108.

Krader L. (1963) Social organization of the Mongol-Turkic pastoral nomads. *Indiana University Publications, Bloomington*, **20**.

Krader L. (1966) Peoples of Central Asia, 2nde edition. *Indiana University Publications, Bloomington*, **26**.

Krause J., Orlando L., Serre D. *et al.* (2007) Neanderthals in central Asia and Siberia. *Nature*, **449**, 902–904.

Kumar V., Langstieh B.T., Madhavi K.V. *et al.* (2006) Global Patterns in Human Mitochondrial DNA and Y-Chromosome Variation Caused by Spatial Instability of the Local Cultural Processes. *PLoS Genetics*, **2**, e53.

Lachance J., Tishkoff S.A. (2013) Population Genomics of Human Adaptation. *Annual Review of Ecology, Evolution, and Systematics*, **44**, 123–143.

Lalueza-Fox C., Sampietro M.L., Gilbert M.T.P. *et al.* (2004) Unravelling migrations in the steppe : mitochondrial DNA sequences from ancient Central Asians. *Proceedings of the Royal Society B : Biological Sciences*, **271**, 941–947.

Lander E., Botstein D. (1987) Homozygosity mapping : a way to map human recessive traits with the DNA of inbred children. *Science*, **236**, 1567–1570.

Langergraber K.E., Siedel H., Mitani J.C. *et al.* (2007) The Genetic Signature of Sex-Biased Migration in Patrilocal Chimpanzees and Humans. *PLoS ONE*, **2**, e973 (1–7).

Lathrop M., Pison G. (1982) Méthode statistique d'étude de l'endogamie. Application à l'étude du choix du conjoint chez les Peul Bandé. *Population (french edition)*, pp. 513–541.

Laurent R., Chaix R. (2012) MHC-dependent mate choice in humans : Why genomic patterns from the HapMap European American dataset support the hypothesis. *BioEssays*, **34**, 267–271.

Lazaridis I., Nadel D., Rollefson G. *et al.* (2016) Genomic insights into the origin of farming in the ancient Near East. *Nature*, **536**, 419–424.

Lazaridis I., Patterson N., Mittnik A. *et al.* (2014) Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, **513**, 409–413.

Leroi-Gourhan A. (1994) *Dictionnaire de la Préhistoire.* Presses universitaires de France.

Leutenegger A.L., Labalme A., Génin E. *et al.* (2006) Using Genomic Inbreeding Coefficient Estimates for Homozygosity Mapping of Rare Recessive Traits : Application to Taybi-Linder Syndrome. *The American Journal of Human Genetics*, **79**, 62–66.

Leutenegger A.L., Prum B., Génin E. *et al.* (2003) Estimation of the Inbreeding Coefficient through Use of Genomic Data. *The American Journal of Human Genetics*, **73**, 516–523.

Leutenegger A.L., Sahbatou M., Gazal S., Cann H., Génin E. (2011) Consanguinity around the world : what do the genomic data of the HGDP-CEPH diversity panel tell us ? *European Journal of Human Genetics*, **19**, 583–587.

Lévi-Strauss C. (1949) *Les structures élémentaires de la parenté.* Mouton de Gruyter.

Levréro F., Carrete-Vega G., Herbert A. *et al.* (2015) Social shaping of voices does not impair phenotype matching of kinship in mandrills. *Nature Communications*, **6**, 1–7.

Li C., Li H., Cui Y. *et al.* (2010) Evidence that a West-East admixed population lived in the Tarim Basin as early as the early Bronze Age. *BMC Biology*, **8**, 1–12.

Li J.Z., Absher D.M., Tang H. *et al.* (2008) Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science*, **319**, 1100–1104.

Lippold S., Xu H., Ko A. *et al.* (2014) Human paternal and maternal demographic histories : insights from high-resolution Y chromosome and mtDNA sequences. *Investigative genetics*, **5**, 1–17.

Litvinskii B.A., Guang-da Z., Shabani Samghabadi R. (1996) *A History of the Civilizations of Central Asia, Volume III : The Crossroads of Civilization : A.D. 250 to 750.*

Loh P.R., Lipson M., Patterson N. *et al.* (2013) Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics*, **193**, 1233–1254.

Magail J. (2005) L'art des «pierres à cerfs» de Mongolie. *Arts asiatiques*, **60**, 172–180.

Malécot G. (1948) Mathématiques de l'Hérédité.

Mallory J.P., Adams D.Q. (1997) *Encyclopedia of Indo-European Culture.* Taylor & Francis.

Malyarchuk B., Derenko M., Wozniak M., Grzybowski T. (2013) Y-chromosome variation in Tajiks and Iranians. *Annals of human biology*, **40**, 48–54.

Marchand P., Elisseeff V., Mennessier G. (2017) Sibérie (in Encyclopædia Universalis [en ligne]).

Marchi N., Hegay T., Mennecier P. *et al.* (2017) Sex-specific genetic diversity is shaped by cultural factors in Inner Asian human populations. *American Journal of Physical Anthropology*, **162**, 627–640.

Marks S.J., Levy H., Martinez-Cadenas C., Montinaro F., Capelli C. (2012) Migration distance rather than migration rate explains genetic diversity in human patrilocal groups. *Molecular Ecology*, **21**, 4958–4969.

Marks S.J., Montinaro F., Levy H. *et al.* (2015) Static and Moving Frontiers : The Genetic Landscape of Southern African Bantu-Speaking Populations. *Molecular Biology and Evolution*, **32**, 29–43.

Marlowe F. (2000) Paternal investment and the human mating system. *Behavioural Processes*, **51**, 45–61.

Martínez-Cruz B., Vitalis R., Ségurel L. *et al.* (2011) In the heartland of Eurasia : the multilocus genetic landscape of Central Asian populations. *European Journal of Human Genetics*, **19**, 216–223.

Mateo J.M., Johnston R.E. (2000) Kin recognition and the 'armpit effect' : evidence of self-referent phenotype matching. *Proceedings of the Royal Society B : Biological Sciences*, **267**, 695–700.

Mathieson I., Lazaridis I., Rohland N. *et al.* (2015) Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, pp. 1–16.

Mazières S. (2011) Towards a reconciling model about the initial peopling of America. *Comptes Rendus Biologies*, **334**, 497–504.

McQuillan R., Eklund N., Pirastu N. *et al.* (2012) Evidence of Inbreeding Depression on Human Height. *PLoS Genetics*, **8**, e1002655 (1–14).

McQuillan R., Leutenegger A.L., Abdel-Rahman R. *et al.* (2008) Runs of Homozygosity in European Populations. *The American Journal of Human Genetics*, **83**, 359–372.

Mead S. (2003) Balancing Selection at the Prion Protein Gene Consistent with Prehistoric Kurulike Epidemics. *Science*, **300**, 640–643.

Mennecier P., Nerbonne J., Heyer E., Manni F. (2015) A Central Asian language survey : Collecting data , measuring relatedness and detecting loans. **1**, 1–40.

Meslé F. (2004) Espérance de vie : un avantage féminin menacé ? *Population & Sociétés*, pp. 1–4.

Michael Boehnke N.J. (1997) Accute Inference of Relationships in Sib-Pair Linkage Studies. *Am J Hum Genet*, **61**, 423–429.

Minahan J.B. (2014) *Ethnic Groups of North, East, and Central Asia : An Encyclopedia.* ABC-CLIO.

Moreno E., Pérez-González J., Carranza J., Moya-Laraño J. (2015) Better Fitness in Captive Cuvier's Gazelle despite Inbreeding Increase : Evidence of Purging ? *PLOS ONE*, **10**, e0145111 (1–15).

Morgan L.H. (1871) *Systems of consanguinity and affinity of the human family*, vol. 218. Smithsonian institution.

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

Muniz L., Perry S., Manson J.H., Gilkenson H., Gros-Louis J., Vigilant L. (2006) Father–daughter inbreeding avoidance in a wild primate population. *Current Biology*, **16**, R156–R157.

Murdock G.P. (1967) Ethnographic Atlas : A Summary. *Ethnology*, **6**, 109–236.

Murdock G.P. (1981) *Atlas of world cultures*. UnUniversity of Pittsburgh Press.

Nasidze I., Ling E.Y.S., Quinque D. *et al.* (2004) Mitochondrial DNA and Y-chromosome variation in the Caucasus. *Annals of Human Genetics*, **68**, 205–221.

Nei M. (1987) *Molecular Evolutionary Genetics*, vol. 17. Columbia University Press.

Ning C., Gao S., Deng B. *et al.* (2015) Ancient mitochondrial genome reveals trace of prehistoric migration in the east Pamir by pastoralists. *Journal of human genetics*, **61**, 103–108.

Nuñez C., Baeta M., Sosa C. *et al.* (2010) Reconstructing the population history of Nicaragua by means of mtDNA, Y-chromosome STRs, and autosomal STR markers. *American Journal of Physical Anthropology*, **143**, 591–600.

O'Brien E., Jorde L.B., Rönnlöf B., Fellman J.O., Eriksson A.W. (1988) Inbreeding and genetic disease in Sottunga, Finland. *American Journal of Physical Anthropology*, **75**, 477–486.

O'Grady J.J., Brook B.W., Reed D.H., Ballou J.D., Tonkyn D.W., Frankham R. (2006) Realistic levels of inbreeding depression strongly affect extinction risk in wild populations. *Biological Conservation*, **133**, 42–51.

Oksanen J., Blanchet F.G., Kindt R. *et al.* (2016) vegan : Community Ecology Package.

Olalde I., Allentoft M.E., Sánchez-Quinto F. *et al.* (2014) Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*, **507**, 225–228.

Olalde I., Schroeder H., Sandoval-Velasco M. *et al.* (2015) A common genetic origin for early farmers from mediterranean cardial and central european LBK cultures. *Molecular Biology and Evolution*, **32**, 3132–3142.

Oota H., Settheetham-Ishida W., Tiwawech D., Ishida T., Stoneking M. (2001) Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics*, **29**, 20–21.

Outram A.K., Stear N.a., Bendrey R. *et al.* (2009) The Earliest Horse Harnessing and Milking. *Science*, **323**, 1332–1335.

van Oven M., Kayser M. (2009) Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation*, **30**, E386–E394.

Overall A.D.J., Ahmad M., Nichols R.A. (2002) The effect of reproductive compensation on recessive disorders within consanguineous human populations. *Heredity*, **88**, 474–479.

Palstra F.P., Heyer E., Austerlitz F. (2015) Statistical Inference on Genetic Data Reveals the Complex Demographic History of Human Populations in Central Asia. *Molecular Biology and Evolution*, **32**, 1411–1424.

Patterson N., Price A.L., Reich D. (2006) Population Structure and Eigenanalysis. *PLoS Genetics*, **2**, e190 (2074–2093).

Pedersen J. (2002) The influence of consanguineous marriage on infant and child mortality among palestinians in the West Bank and Gaza, Jordan, Lebanon and Syria. *Community Genetics*, **5**, 178–181.

Pemberton T.J., Absher D., Feldman M.W., Myers R.M., Rosenberg N.a., Li J.Z. (2012) Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics*, **91**, 275–292.

Pereira V., Tomas C., Sanchez J.J. *et al.* (2015) The peopling of Greenland : further insights from the analysis of genetic diversity using autosomal and X-chromosomal markers. *European Journal of Human Genetics*, **23**, 245–251.

Pérez-Lezaun A., Calafell F., Comas D. *et al.* (1999) Sex-Specific Migration Patterns in Central Asian Populations, Revealed by Analysis of Y-Chromosome Short Tandem Repeats and mtDNA. *The American Journal of Human Genetics*, **65**, 208–219.

Perrin N., Lehmann L. (2001) Is sociality driven by the costs of dispersal or the benefits of philopatry ? A role for kin-discrimination mechanisms. *The American naturalist*, **158**, 471–483.

Perrin N., Mazalov V. (2000) Local Competition, Inbreeding, and the Evolution of Sex-Biased Dispersal. *The American Naturalist*, **155**, 116–127.

Pfefferle D., Kazem A.J., Brockhausen R.R., Ruiz-Lambides A.V., Widdig A. (2014) Monkeys Spontaneously Discriminate Their Unfamiliar Paternal Kin under Natural Conditions Using Facial Cues. *Current Biology*, **24**, 1806–1810.

Pfefferle D., Ruiz-Lambides A.V., Widdig A. (2013) Female rhesus macaques discriminate unfamiliar paternal sisters in playback experiments : support for acoustic phenotype matching. *Proceedings of the Royal Society B : Biological Sciences*, **281**, 20131628–20131628.

Pison G. (1986) La démographie de la polygamie. *Population (french edition)*, pp. 93–122.

Pluzhnikov A., Nolan D.K., Tan Z., McPeek M.S., Ober C. (2007) Correlation of Intergenerational Family Sizes Suggests a Genetic Component of Reproductive Fitness. *The American Journal of Human Genetics*, **81**, 165–169.

Pool J.E., Nielsen R. (2007) Population size changes reshape genomic patterns of diversity. *Evolution*, **61**, 3001–3006.

Powell J.E., Visscher P.M., Goddard M.E. (2010) Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics*, **11**, 800–805.

Poznik G.D., Xue Y., Mendez F.L. *et al.* (2016) Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences. *Nature Genetics*, **48**, 593–599.

Prüfer K., Racimo F., Patterson N. *et al.* (2013) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43–49.

Prugnolle F., de Meeus T. (2002) Inferring sex-biased dispersal from population genetic tools : a review. *Heredity*, **88**, 161–165.

Purcell S., Neale B., Todd-Brown K. *et al.* (2007) PLINK : A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, **81**, 559–575.

Pusey A. (1990) Mechanisms of Inbreeding Avoidance in Nonhuman Primates. In *Pedophilia*, chap. 7, pp. 201–220. Springer New York, New York, NY.

Pusey A., Wolf M. (1996) Inbreeding avoidance in animals. *Trends in Ecology & Evolution*, **11**, 201–206.

Pusey A.E. (1980) Inbreeding avoidance in chimpanzees. *Animal Behaviour*, **28**, 543–552.

Pusey A.E., Packer C. (1987) The Evolution of Sex-Biased Dispersal in Lions. *Behaviour*, **101**, 275–310.

Quintana-Murci L., Barreiro L.B. (2010) The role played by natural selection on Mendelian traits in humans. *Annals of the New York Academy of Sciences*, **1214**, 1–17.

Quintana-Murci L., Krausz C., Zerjal T. *et al.* (2001) Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *American journal of human genetics*, **68**, 537–542.

Quintana-Murci L., Quach H., Harmant C. *et al.* (2008) Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences*, **105**, 1596–1601.

Raghavan M., Skoglund P., Graf K.E. *et al.* (2014) Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature*, **505**, 87–91.

Rasmussen M., Anzick S.L., Waters M.R. *et al.* (2014) The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature*, **506**, 225–229.

Reich D., Green R.E., Kircher M. *et al.* (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, **468**, 1053–1060.

Reich D., Patterson N., Campbell D. (2012) Reconstructing native American population history. *Nature*, **488**, 370–374.

Ritland K. (1996) A Marker-Based Method for Inferences About Quantitative Inheritance in Natural Populations. *Evolution*, **50**, 1062–1073.

Robert A., Toupance B., Tremblay M., Heyer E. (2009) Impact of inbreeding on fertility in a pre-industrial population. *European Journal of Human Genetics*, **17**, 673–681.

Roberts D.F. (1976) Les concepts d'isolats. *Jacquard A. L'étude des isolats. Paris : INED*, pp. 75–92.

Rohde D.L.T. (2003) On the Common Ancestors of All Living Humans. pp. 1–30.

Ronce O. (2007) How does it feel to be like a rolling stone ? Ten questions about dispersal evolution. *Annual Review of Ecology, Evolution, and Systematics*, **38**, 231–253.

Rosenberg N.A. (2004) DISTRUCT : A program for the graphical display of population structure. *Molecular Ecology Notes*, **4**, 137–138.

Rosenberg N.a., Pritchard J.K., Weber J.L. *et al.* (2002) Genetic structure of human populations. *Science (New York, N.Y.)*, **298**, 2381–5.

Ross M.T., al. E. (2005) The DNA sequence of the human X chromosome. *Nature*, **434**, 325–337.

Roy J., Gray M., Stoinski T., Robbins M.M., Vigilant L. (2014) Fine-scale genetic structure analyses suggest further male than female dispersal in mountain gorillas. *BMC Ecology*, **14**, 1–16.

Saad H.A., Elbedour S., Hallaq E., Merrick J., Tenenbaum A. (2014) Consanguineous Marriage and Intellectual and Developmental Disabilities among Arab Bedouins Children of the Negev Region in Southern Israel : A Pilot Study. *Frontiers in public health*, **2**, 1–3.

Saccheri I., Kuussaari M., Kankare M., Vikman P., Fortelius W., Hanski I. (1998) Inbreeding and extinction in a butterfly metapopulation. *Nature*, **392**, 491–494.

Sacks O. (1997) *The Island of the Colorblind*. A.A. Knopf.

Sahoo S., Kashyap V.K. (2006) Phylogeography of mitochondrial DNA and Y-chromosome haplogroups reveal asymmetric gene flow in populations of Eastern India. *American Journal of Physical Anthropology*, **131**, 84–97.

Salem A.H., Badr F.M., Gaballah M.F., Pääbo S. (1996) The genetics of traditional living : Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. *American journal of human genetics*, **59**, 741–743.

Santos C., Fregel R., Cabrera V.M. *et al.* (2014) Mitochondrial DNA and Y-chromosome structure at the mediterranean and atlantic façades of the iberian peninsula. *American Journal of Human Biology*, **26**, 130–141.

Schaffner S.F. (2004) The X chromosome in population genetics. *Nature Reviews Genetics*, **5**, 43–51.

Schwartz M., Vissing J. (2002) Paternal Inheritance of Mitochondrial DNA. *New England Journal of Medicine*, **347**, 576–580.

Séfériadès M. (2003) An Aspect of Neolithisation in Mongolia : the Mesolithic-Neolithic Site of Tamsagbulag (Dornod District). *Documenta Praehistorica*, **XXXI**, 139–149.

Segalen M. (1985) Quinze générations de Bas-Bretons : parenté et société dans le pays bigouden-sud, 1720-1980. p. 404.

Segalen M. (1986) *Historical anthropology of the family*. Cambridge University Press.

Ségurel L., Austerlitz F., Toupance B. *et al.* (2013) Positive selection of protective variants for type 2 diabetes from the Neolithic onward : a case study in Central Asia. *European Journal of Human Genetics*, **21**, 1146–1151.

Ségurel L., Martínez-Cruz B., Quintana-Murci L. *et al.* (2008) Sex-Specific Genetic Structure and Social Organization in Central Asia : Insights from a Multi-Locus Study. *PLoS Genetics*, **4**, e1000200 (1–14).

Seielstad M.T., Minch E., Cavalli-Sforza L.L. (1998) Genetic evidence for a higher female migration rate in humans. *Nature Genetics*, **20**, 278–280.

Serre J.L. (2006) *Génétique des populations*. Dunod.

Serre J.L., Jakobi L., Babron M.c. (1985) A genetic isolate in the French Pyrenees : probabilities of origin of genes and inbreeding. *Journal of Biosocial Science*, **17**, 405–414.

Setchell J.M., Abbott K.M., Gonzalez J.P., Knapp L.A. (2013) Testing for post-copulatory selection for major histocompatibility complex genotype in a semi-free-ranging primate population. *American Journal of Primatology*, **75**, 1021–1031.

Shan W., Ablimit A., Zhou W., Zhang F., Ma Z., Zheng X. (2014) Genetic polymorphism of 17 y chromosomal STRs in Kazakh and Uighur populations from Xinjiang, China. *International Journal of Legal Medicine*, **128**, 743–744.

Sibert A., Austerlitz F., Heyer É. (2002) Wright–Fisher Revisited : The Case of Fertility Correlation. *Theoretical Population Biology*, **62**, 181–197.

Simmons L. (1989) Kin recognition and its influence on mating preferences of the field cricket, Gryllus bimaculatus (de Geer). *Animal Behaviour*, **38**, 68–77.

Sjöstrand A.E. (2015) *Origins and Adaptation in Humans : A Case Study of Taste and Lifestyle*. Ph.D. thesis, Uppsala University, Evolutionary Biology.

Skaletsky H., Kuroda-Kawaguchi T., Minx P.J. *et al.* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, **423**, 825–837.

Slatkin M. (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**, 457–462.

Smouse P.E., Vitzthum V.J., Neel J.V. (1981) The impact of random and lineal fission on the genetic divergence of small human groups : A case study among the Yanomama. *Genetics*, **98**, 179–197.

Soucek S. (2000) *A History of Inner Asia*. Cambridge University Press.

Spielman D., Brook B.W., Briscoe D.a., Frankham R. (2004) Does inbreeding and loss of genetic diversity reduce disease resistance ? *Conservation Genetics*, **5**, 439–448.

Stépanoff C., Ferret C., Lacaze G., Thorez J. (2013) *Nomadismes d'Asie centrale et septentrionale*. Armand Colin.

Stoneking M. (1998) Women on the move. *Nature Genetics*, **20**, 219–220.

Streiff-Fénart J., Poutignat P. (1995) *Théories de l'ethnicité*. Le sociologue. Presses Universitaires de France.

Sutter J., Tabah L. (1955) L'évolution des isolats de deux départements français : Loir-et-Cher, Finistère. *Population (french edition)*, pp. 645–674.

Szpiech Z.A. (2011) asd computer program.

Tabah L., Sutter J. (1950) La mesure de la consanguinité : Perspectives d'application à la démographie. *Population (french edition)*, pp. 689–712.

Talmon Y. (1964) Mate Selection in Collective Settlements. *American Sociological Review*, **29**, 491–508.

Tarlykov P.V., Zholdybayeva E.V., Akilzhanova A.R. *et al.* (2013) Mitochondrial and Y-chromosomal profile of the Kazakh population from East Kazakhstan. *Croatian Medical Journal*, **54**, 17–24.

The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

The Y Chromosome Consortium (2002) A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups. *Genome Research*, **12**, 339–348.

Thomas A., Skolnick M.H., Lewis C.M. (1994) Genomic mismatch scanning in pedigrees. *Mathematical Medicine and Biology*, **11**, 1–16.

Thompson E.M., Winter R.M. (1988) Another family with the 'Habsburg jaw'. *Journal of Medical Genetics*, **25**, 838–842.

Tremblay M., Heyer É., St-Hilaire M. (2000) Comparaisons intergénérationnelles de l'endogamie à partir des lieux de mariage et de résidence . L'exemple de la population du Saguenay. *Cahiers québécois de démographie*, **29**, 119–146.

Tremblay M., Vézina H. (2000) New Estimates of Intergenerational Time Intervals for the Calculation of Age and Origins of Mutations. *The American Journal of Human Genetics*, **66**, 651–658.

Trinkaus E. (2005) Early Modern Humans. *Annual Review Of Anthropology*, **34**, 207–230.

Unterländer M., Palstra F., Lazaridis I. *et al.* (2017) Ancestry and demography and descendants of Iron Age nomads of the Eurasian Steppe. *Nature Communications*, **8**, 1–10.

Van Eldik P., Van Der Waaij E.H., Ducro B., Kooper A.W., Stout T.A.E., Colenbrander B. (2006) Possible negative effects of inbreeding on semen quality in Shetland pony stallions. *Theriogenology*, **65**, 1159–1170.

Verdu P., Becker N.S.A., Froment A. *et al.* (2013) Sociocultural Behavior, Sex-Biased Admixture, and Effective Population Sizes in Central African Pygmies and Non-Pygmies. *Molecular Biology and Evolution*, **30**, 918–937.

Verdu P., Pemberton T.J., Laurent R. *et al.* (2014) Patterns of Admixture and Population Structure in Native Populations of Northwest North America. *PLoS Genetics*, **10**, e1004530 (1–17).

Verdu P., Rosenberg N.a. (2011) A general mechanistic model for admixture histories of hybrid populations. *Genetics*, **189**, 1413–1426.

Verweij K.J.H., Abdellaoui A., Veijola J. *et al.* (2014) The Association of Genotype-Based Inbreeding Coefficient with a Range of Physical and Psychological Human Traits. *PLoS ONE*, **9**, e103102 (1–6).

Vigilant L., Stoneking M., Harpending H., Hawkes K., Wilson A.C. (1991) African Populations and the Evolution of Human Mitochondrial DNA. *Science*, **253**, 1503–1507.

Voskarides K., Mazières S., Hadjipanagi D. *et al.* (2016) Y-chromosome phylogeographic analysis of the Greek-Cypriot population reveals elements consistent with Neolithic and Bronze Age settlements. *Investigative Genetics*, **7**, 1–14.

Wang C.C., Wang L.X., Shrestha R. *et al.* (2014) Genetic Structure of Qiangic Populations Residing in the Western Sichuan Corridor. *PLoS ONE*, **9**, e103772 (1–14).

Weber A., Katzenberg M.A., Schurr T.G. (2011) *Prehistoric hunter-gatherers of the Baikal region, Siberia : bioarchaeological studies of past life ways.* University of Pennsylvania Press.

Webster T.H., Wilson Sayres M.A. (2016) Genomic signatures of sex-biased demography : progress and prospects. *Current Opinion in Genetics and Development*, **41**, 62–71.

Weisfeld G.E., Czilli T., Phillips K.A., Gall J.A., Lichtman C.M. (2003) Possible olfaction-based mechanisms in human kin recognition and inbreeding avoidance. *Journal of Experimental Child Psychology*, **85**, 279–295.

Wells R.S., Yuldasheva N., Ruzibakiev R. *et al.* (2001) The Eurasian Heartland : A continental perspective on Y-chromosome diversity. *Proceedings of the National Academy of Sciences*, **98**, 10244–10249.

Westermarck E. (1921) *The history of human marriage*. Allerton Book Company.

Widdig A., Muniz L., Minkner M. *et al.* (2017) Low incidence of inbreeding in a long-lived primate population isolated for 75 years. *Behavioral Ecology and Sociobiology*, **71**, 18.

Wilder J.a., Kingan S.B., Mobasher Z., Pilkington M.M., Hammer M.F. (2004) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nature Genetics*, **36**, 1122–1125.

Wilkins J.F. (2006) Unraveling male and female histories from human genetic data. *Current Opinion in Genetics & Development*, **16**, 611–617.

Wilkins J.F., Marlowe F.W. (2006) Sex-biased migration in humans : What should we expect from genetic data ? *BioEssays*, **28**, 290–300.

Willems T., Gymrek M., Poznik G.D., Tyler-Smith C., Erlich Y. (2016) Population-Scale Sequencing Data Enable Precise Estimates of Y-STR Mutation Rates. *American Journal of Human Genetics*, **98**, 919–933.

Wilson I., Weale M., Balding D.J. (2003) Inferences from DNA data : population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society*, **166**, 155–201.

Wolf A.P. (1970) Childhood Association and Sexual Attraction : A Further Test of the Westermarck Hypothesis. *American Anthropologist*, **72**, 503–515.

Wolf A.P. (2004) *Inbreeding, incest, and the incest taboo : The state of knowledge at the turn of the century*. Stanford University Press.

Wolff G., Wienker T.F., Sander H. (1993) On the genetics of mandibular prognathism : analysis of large European noble families. *Journal of medical genetics*, **30**, 112–116.

Wood E.T., Stover D.a., Ehret C. *et al.* (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa : evidence for sex-biased demographic processes. *European Journal of Human Genetics*, **13**, 867–876.

Woods C.G., Cox J., Springell K. *et al.* (2006) Quantification of homozygosity in consanguineous individuals with autosomal recessive disease. *American journal of human genetics*, **78**, 889–896.

Wright S. (1922) Coefficients of inbreeding and relationship. *The American Naturalist*, **56**, 330–338.

Wright S. (1931) Evolution in Mendelian populations. *Genetics*, **16**, 97–159.

Wright S. (1943) Isolation by Distance. *Genetics*, **28**, 114–138.

Wright S. (1949) The Genetical Structure Of Populations. *Annals of Eugenics*, **15**, 323–354.

Xue Y., Zerjal T., Bao W. *et al.* (2005) Recent Spread of a Y-Chromosomal Lineage in Northern China and Mongolia. *The American Journal of Human Genetics*, **77**, 1112–1116.

Yang N.N., Mazières S., Bravi C. *et al.* (2010) Contrasting Patterns of Nuclear and mtDNA Diversity in Native American Populations. *Annals of Human Genetics*, **74**, 525–538.

Yunusbayev B., Metspalu M., Metspalu E. *et al.* (2015) The Genetic Legacy of the Expansion of Turkic-Speaking Nomads across Eurasia. *PLOS Genetics*, **11**, e1005068 (1–24).

Zerjal T., Xue Y., Bertorelle G. *et al.* (2003) The Genetic Legacy of the Mongols. *The American Journal of Human Genetics*, **72**, 717–721.

Zhao D., Ji W., Li B., Watanabe K. (2008) Mate competition and reproductive correlates of female dispersal in a polygynous primate species (Rhinopithecus roxellana). *Behavioural Processes*, **79**, 165–170.

ABSTRACT

My PhD thesis is about the influence of cultural behaviours on the neutral genetic diversity of human populations from Inner Asia. Notably, I investigated how specific behaviours may affect the demographic history of populations, by acting on the intensity of migration and genetic drift. To do so, I combined genetic and ethnological data, collected in present-day Inner Asian populations that belong to two major cultural and linguistic groups and have different social organisations.

The first part of this work aims at understanding how Inner Asia was peopled, from the Bronze Age to nowadays. This was done in the framework of an international collaboration, through the study of both ancient and modern genomic data. The results obtained showed that modern populations are divided in two distinct genetic groups, mirroring the two cultural groups, and exhibiting contrasted ancestral components.

I was then interested in exploring the influence of cultural behaviours on the sex-specific genetic structure of present-day populations from Inner Asia. By studying the genetic diversity of uniparental markers, namely mitochondrial DNA and the Y chromosome, I was able to characterize sex-specific genetic differences, such as a reduced population differentiation for mitochondrial DNA as compared to the Y chromosome. This maternal genetic homogeneity between populations may be explained by patrilocality, a residence rule shared by all the studied populations and generating mostly female migrations between populations. On the other hand, I showed there were some significant differences in genetic diversity between the two cultural groups for the Y chromosome. This observation may be related to the different filiation rules of these two groups. Indeed, one is patrilineal : the social filiation is inherited from the father, while the other is cognatic : the transmission is undifferentiated between the parents. It could then be that patrilineality leads to the formation of cores of related men within the population, who share the same Y chromosome. This population structuration would result in a reduced genetic diversity for the Y chromosome in patrilineal populations, compared to cognatics. As expected, the mitochondrial diversity is comparable between patrilineal and cognatic group, comforting the idea that patrilineality affects only the male genetic diversity. Finally, to investigate the ethnogenesis process, I calculated the genetic age of patrilineal ethnic groups from STR markers of the Y chromosome. I showed that this biological age is older than the one from historical sources, which suggests that, at least for Turko-Mongolic from Inner Asia, the ethnic group is partly a social construct, rather than an actual biological entity.

In the third part, I focused on whether dispersal can be an inbreeding avoidance mechanisms by dispersal. Notably, I tested the hypothesis that exogamous unions, between spouses born in different villages, would lead to less inbreeding than endogamous unions. Despite a strong variation of the exogamous rate between the populations of the studied dataset, no significant difference was found for inbreeding, which was estimated from a genome-wide dataset. At the individual scale, I showed that some of the descendants of exogamous unions are inbred. This is especially true for spouses born less than 40 km away, in which case their descendants are statistically more inbred than those from endogamous unions. This shows that, in human populations, specific matrimonial behaviours, driven by culture, may contradict the results expected by evolutionary biology.

In conclusion, my work shows several cases, at different time and geographic scales, where cultural behaviours left a footprint into the genetic diversity of Inner Asian populations.

KEY-WORDS : genetic diversity ; inbreeding ; sex-specific ; settlement ; ethnogenesis ; ethnic groups ; patrilineality ; patrilocality ; exogamy ; social organisation ; culture ; migration

RÉSUMÉ

Ma thèse s'intéresse à l'influence des comportements culturels sur la diversité génétique neutre des populations humaines, en particulier les populations d'Asie intérieure. Notamment, ces travaux explorent comment certains comportements affectent l'histoire démographique des populations, en agissant sur l'intensité des migrations et de la dérive génétique. Pour ce faire, j'ai étudié des données génétiques, au regard de données ethnologiques, collectées dans des populations habitant actuellement en Asie intérieure, qui diffèrent, entre autres, par leur organisation sociale.

La première partie de cette thèse cherche à retracer l'histoire du peuplement de l'Asie intérieure, de l'âge du Bronze jusqu'à nos jours. Pour ce faire, et dans le cadre d'une collaboration internationale, des données génomiques d'ADN moderne et ancien ont été obtenues et analysées conjointement. Les résultats montrent que les populations actuelles forment deux groupes génétiques distincts, correspondant à deux groupes linguistiques, et reflétant des composantes ancestrales contrastées.

En étudiant la diversité génétique de marqueurs uniparentaux, à savoir l'ADN mitochondrial et le chromosome Y, j'ai pu montrer des différences génétiques sexe-spécifiques, telles qu'une différenciation des populations réduite pour l'ADN mitochondrial par rapport à celle du chromosome Y. Cette homogénéité génétique des populations pourrait être causée par de la patrilocalité, une règle de résidence commune à toutes les populations étudiées et entrainant principalement des migrations féminines entre populations. D'autre part, j'ai observé des différences de diversité génétique entre les groupes d'Asie intérieure pour le chromosome Y. J'ai interprété cette observation à la lumière des différences de règles de filiation suivies par ces deux groupes : l'un des groupes est patrilinéaire, c'est-à-dire que la filiation sociale est héritée du père ; l'autre groupe est cognatique, et la transmission est indifférenciée entre les parents. La patrilinéarité conduirait à la formation de noyaux d'hommes apparentés par la lignée masculine dans la population et donc partageant le même chromosome Y. Cette structuration en noyaux conduirait à une diminution de la diversité génétique du chromosome Y des populations patrilinéaires, comparées aux cognatiques. La diversité mitochondriale est, par contre, similaire entre patrilinéaires et cognatiques, illustrant le fait que seule la diversité génétique masculine est affectée par la patrilinéarité. Finalement, pour étudier le processus d'ethnogénèse, j'ai calculé l'âge génétique des groupes ethniques patrilinéaires à partir de marqueurs STRs du chromosome Y, et j'ai montré que cet âge biologique est plus ancien que les âges historiques, suggérant que l'ethnie, du moins chez les Turco-Mongols d'Asie intérieure, est une construction en partie sociale, plutôt qu'une entité entièrement biologique.

Dans la troisième partie, je me suis intéressée aux mécanismes d'évitement de la consanguinité, que j'ai estimée au moyen de données génomiques. J'ai notamment testé l'hypothèse selon laquelle des unions exogames, entre conjoints nés dans des villages différents, permettraient de réduire la consanguinité. Malgré une importante variabilité du taux d'exogamie entre populations et entre groupes linguistiques dans notre jeu de données, je n'ai trouvé aucune différence significative de consanguinité. A l'échelle des individus, j'ai pu mettre en évidence le fait que certains descendants de couples exogames sont néanmoins consanguins. Cette situation est particulièrement répandue pour des conjoints nés à moins de 40 km l'un de l'autre, à tel point que leurs descendants sont statistiquement plus consanguins que les descendants de couples endogames. Ces résultats illustrent que, chez l'Homme, des comportements culturels d'alliance peuvent s'opposer aux attendus de la biologie évolutive.

Ainsi, mes travaux illustrent plusieurs cas de figure, à des échelles géographiques et temporelles différentes, où des comportements culturels ont modifié et laissé une signature génétique particulière sur la diversité des populations humaines d'Asie intérieure.

MOTS-CLÉS : diversité génétique ; consanguinité ; sexe-spécifique ; peuplement ; ethnogénèse ; groupes ethniques ; patrilinéarité ; patrilocalité ; exogamie ; organisation sociale ; culture ; migration