TABLE DES MATIÈRES

INTRODUCTION

L'Homme moderne, comme toute autre espèce, a évolué au cours du temps en étroite relation avec son environnement. Sa plasticité, à la fois biologique et culturelle, lui a permis de coloniser la majorité de la planète. Durant cette conquête, les populations humaines ont rencontré des milieux climatiques extrêmes, se sont exposées à de nouveaux pathogènes et ont appris à tirer leur énergie de sources alimentaires variées. Elles ont également subi une forte croissance démographique et se sont structurées en unités culturelles distinctes, évoluant alors avec différents niveaux d'interactions entre elles. Des barrières à la reproduction complexes, de nature géographique, linguistique, religieuse, ou encore politique, se sont établies, à la fois à l'intérieur des populations et entre elles. Diverses évolutions culturelles, comme par exemple la sédentarisation et la pratique de l'agriculture, ont permis de s'affranchir de certaines pressions environnementales, mais ont également été sources de nouvelles contraintes (notamment alimentaires et infectieuses). Ces changements se sont déroulés sur une courte période aux yeux de l'évolution, l'Homme moderne ayant migré hors d'Afrique pour coloniser les autres continents il y a seulement 60 000 - 40 000 ans (pour une revue, voir Trinkaus, 2005). L'influence de ces évènements sur la diversité biologique humaine (notamment la diversité génétique, morphologique et physiologique) suscite de multiples interrogations.

De nombreuses disciplines s'intéressent ainsi depuis longtemps à l'histoire évolutive humaine, dont notamment l'anthropologie qui, au sens large, étudie les variations biologiques et culturelles entre groupes humains. L'anthropologie fait elle-même appel à de nombreuses disciplines, dont la paléontologie (étude des restes fossiles), l'archéologie (étude des vestiges matériels), l'ethnologie (étude des caractères sociaux et culturels), la médecine, la linguistique, ou encore la primatologie. Plus récemment, de nouvelles disciplines sont venues enrichir nos connaissances, comme la génétique évolutive, et plus précisément la génétique des populations (étude des forces qui influencent la diversité génétique des populations). Ces nouvelles données, qu'il s'agisse initialement de données de polymorphismes génétiques acquises dans les années 1960 (Cavalli-Sforza, 1966), ou de projets à grande envergure d'étude de la diversité génétique mondiale comme le projet HapMap (International HapMap Consortium, 2003, International HapMap Consortium, 2005, International HapMap Consortium *et al.*, 2007), en passant par la publication de la séquence complète du génome humain dans les années 2000 (Lander *et al.*, 2001, Venter *et al.*, 2001), ont permis d'apporter

un nouveau point de vue sur l'histoire évolutive de l'Homme moderne (Cavalli-Sforza *et al.*, 1988, Harpending & Rogers, 2000, Goldstein & Chikhi, 2002, Cavalli-Sforza & Feldman, 2003, Beaumont, 2004, Rosenberg *et al.*, 2005, Li *et al.*, 2008). En effet, la diversité génétique est fortement influencée par l'histoire démographique des espèces (fluctuation de la taille des populations, intensité des migrations entre populations, etc.), ainsi que par les contraintes du milieu (avantage de certaines mutations dans certains environnements). Elle offre donc une source d'informations précieuses pour étudier le passé évolutif des populations humaines. De plus, depuis peu, cette discipline a rassemblé suffisamment d'outils moléculaires, statistiques et informatiques pour pouvoir réellement avancer dans la compréhension de l'histoire évolutive humaine, qui suscite toujours autant d'intérêt, de la part des scientifiques comme des non scientifiques. S'interroger sur ces questions de recherche présente finalement non seulement un intérêt d'ordre théorique, pour comprendre quelles forces majeures gouvernent l'évolution humaine en général, mais également d'ordre médical, puisque les sciences de l'évolution permettent d'appréhender la question des différences de problèmes de santé entre populations (médecine évolutive ou darwinienne, voir Nesse & Williams, 1996, Stearns & Koella, 2007), ainsi que d'ordre sociétal, car la connaissance des points communs entre tous les Hommes, et de la diversité qui fait de chacun de nous un être unique suscite le plus souvent un vif intérêt de la part du grand public.

I. La génétique des populations comme outil pour appréhender le passé

L'objectif de la génétique des populations est de comprendre et d'interpréter le rôle relatif des forces évolutives qui façonnent notre diversité génétique. Cette discipline s'intéresse donc exclusivement à la part du génome qui est variable entre les individus. Les différentes forces évolutives sont la mutation, la recombinaison, la migration, la dérive et la sélection, et chacune laisse des signatures particulières dans notre génome. De toutes les forces évolutives, la mutation est la seule qui crée de la diversité. La recombinaison, comme son nom l'indique, ne crée pas de mutations mais de nouvelles combinaisons de mutations. La dérive génétique est la conséquence des effets d'échantillonnage liés à la reproduction des individus dans une population de taille finie ; elle se traduit par une variation stochastique des fréquences alléliques d'une génération à l'autre et par une perte de diversité. Cette perte de

diversité est d'autant plus forte que la taille efficace de la population N_e [1] est faible (Wright, 1951). Enfin, une migration limitée favorise la différenciation génétique des populations, tandis qu'une forte migration permet leur mélange.

La migration et la dérive sont fréquemment regroupées sous le terme de forces démographiques et ont pour point commun qu'elles influencent, en espérance, tout le génome de la même manière (bien qu'il existe une certaine variance de leurs effets à différents marqueurs). Opposée aux précédentes forces, car elle agit sur un ou plusieurs marqueurs en particulier, intervient la sélection. Cette force favorise l'augmentation (sélection positive) ou la diminution (sélection négative) de la fréquence de certaines mutations en fonction de leurs effets sur la reproduction ou la survie des individus (c'est-à-dire en fonction de leur *valeur sélective*). La sélection change donc, localement dans le génome, le niveau de diversité génétique. Bien que les forces sélectives et démographiques ont des effets confondants sur le niveau de variation à un marqueur donné, il est possible de les distinguer de par cette particularité : les forces démographiques affectent tous les marqueurs de la même façon en espérance, tandis que la sélection laisse une signature singulière seulement à un sous-ensemble de marqueurs (à l'exception de certaines formes de sélection, comme la sélection d'arrière plan : voir Charlesworth et al., 1995). Ainsi, pour détecter les signatures laissées par la sélection, il faut au préalable avoir décrit la diversité génétique neutre des populations.

Deux types d'approches complémentaires sont possibles en génétique des populations :

- d'un côté, l'étude de marqueurs génétiques éloignés des parties codantes du génome (les gènes), et considérés comme *a priori* « neutres », c'est-à-dire n'influençant pas *a priori* la valeur sélective des individus et n'étant pas entraînés génétiquement par des gènes sélectionnés. Ces marqueurs nous informent sur l'histoire démographique des populations (flux migratoires entre populations, changements de taille des populations, barrières à la reproduction). Nous utiliserons cette approche dans la première partie de cette thèse.

- d'un autre côté, l'étude de marqueurs à l'intérieur ou à proximité de gènes, qui influencent potentiellement la valeur sélective des individus. Ces marqueurs nous permettent d'accéder à l'histoire adaptative des populations (les réponses aux contraintes

[1] La taille efficace est la taille d'une population idéale de Wright-Fisher (de taille constante, où tous les individus ont la même probabilité de se reproduire et où la sélection n'agit pas) qui dériverait de la même manière que la population étudiée

environnementales, nutritives ou infectieuses). Nous utiliserons cette approche dans la deuxième partie de cette thèse.

Apport de la génétique des populations à l'histoire évolutive humaine

Les premières études en génétique des populations ont permis de calculer que chez l'Homme moderne, la vaste majorité de la diversité génétique (à hauteur de 86%) se situe à l'intérieur des populations (Lewontin, 1972). Ainsi, deux individus venant de continents différents ne présentent pas beaucoup plus de différences génétiques que deux individus habitant au même endroit. Cette information a eu un fort retentissement car elle a répandu l'idée que les populations humaines n'étaient pas structurées génétiquement et qu'il ne pouvait donc pas y avoir de fortes différences biologiques entre populations. Cette répartition de la diversité génétique peut s'expliquer par le fait que l'Homme moderne est une espèce relativement jeune (apparue en Afrique il y a environ 200 000 ans, Trinkaus, 2005), qui a subi une croissance démographique très récente (à la fin du Paléolithique, il y a environ 12 000 ans, la population humaine ne comptait guère plus d'un million d'individus, Harris, 1996), et pour qui finalement les migrations entre populations jouent certainement un rôle important.

Cependant, bien que nous soyons une espèce jeune et peu structurée génétiquement, certaines études récentes basées sur un grand nombre de marqueurs ont révélé que les individus peuvent être répartis dans des groupes distincts, correspondant approximativement aux différents continents (Cavalli-Sforza *et al.*, 1988, Rosenberg *et al.*, 2002, Cavalli-Sforza & Feldman, 2003, Bamshad *et al.*, 2004, Rosenberg *et al.*, 2005, Li *et al.*, 2008). Bien que la présence d'océans, de déserts ou de montagnes engendre indiscutablement de fortes barrières géographiques, de par la difficulté de les traverser, il est intéressant de se demander dans quelle mesure nous pouvons parler de groupes humains différents et dans ce cas, affilier chaque population à un de ces groupes, ou s'il ne faut pas plutôt considérer qu'il existe des variations continues de diversité (Rosenberg *et al.*, 2005). Il est également remarquable que, plus l'on augmente le nombre de marqueurs étudiés, plus l'on réussit à distinguer les populations entre elles. Plusieurs études en Europe ont ainsi montré que l'on pouvait inférer l'origine géographique des individus, à partir de leur génotype à plus de 500 000 marqueurs, à quelques 700 km près (Lao *et al.*, 2008, Novembre *et al.*, 2008). Ces résultats sont basés uniquement sur des individus ayant les quatre grands-parents de même origine, mais ils montrent bien que l'étude d'un nombre suffisamment important de marqueurs génétiques permet de retracer l'origine géographique des individus.

Répartition de la diversité génétique : le rôle de la géographie et de la linguistique

La géographie est donc un facteur majeur pour expliquer la répartition de la diversité humaine. La probabilité d'appariement plus forte entre individus géographiquement proches a donné lieu au modèle démographique d'isolement par la distance (Malécot, 1973). Ce modèle considère que la migration des individus est localisée (limitée dans l'espace), que les populations sont à l'équilibre démographique (c'est-à-dire sont stables dans le temps), et qu'il n'y a pas de sélection. Il a été montré dans ces modèles que, lorsque que l'on considère un espace à deux dimensions, la distance génétique augmente linéairement avec le logarithme de la distance géographique (Rousset, 1997).

Un autre facteur a également été proposé comme responsable de barrières majeures à la reproduction : la linguistique. Cavalli-Sforza a été le premier à attirer l'attention sur une éventuelle coévolution entre gènes et langues (Cavalli-Sforza *et al.*, 1988). Les matrices de distances linguistiques entre populations sont en effet assez bien corrélées avec celles des distances génétiques, ce qui suggère que les populations échangent des migrants préférentiellement quand il y a intercompréhension linguistique (Sokal, 1988, Poloni *et al.*, 1997, Lum *et al.*, 1998). Cependant, Rosser (2000) a remarqué que, si l'on prend en compte les distances géographiques, la corrélation entre distances linguistiques et génétiques disparaît, suggérant ainsi que la géographie est la seule responsable des différences génétiques. Depuis, d'autres études ont montré, d'après des données génomiques, qu'une corrélation faible mais significative existe bien entre gènes et langues, même si l'on prend en compte la géographie (Belle & Barbujani, 2007, Lansing *et al.*, 2007). La question du rôle respectif des distances linguistiques et géographiques pour favoriser ou empêcher le brassage des populations reste donc ouverte.

Répartition de la diversité génétique : le rôle de l'organisation sociale

D'autres facteurs culturels, comme par exemple l'organisation sociale, se sont également avérés avoir une importance considérable dans la formation de barrières à la reproduction entre populations. En effet, certaines études ont montré que des populations aux organisations sociales contrastées (particulièrement en ce qui concerne les pratiques matrimoniales ou les règles d'héritage) ont une répartition différente de leur diversité génétique, notamment entre hommes et femmes (Wilkins & Marlowe, 2006, Kumar *et al.*, 2006, Chaix *et al.*, 2007). La connaissance de cette histoire sexe-spécifique est permise par

l'étude de marqueurs « uni-parentaux » : le chromosome Y transmis par les pères à leurs fils et l'ADN mitochondrial transmis par les mères à leurs enfants, permettant d'accéder respectivement aux histoires des lignées paternelles et maternelles. Ces études ont montré qu'en général, dans les sociétés patrilocales (dans lesquelles la femme part s'installer dans le village de son mari après le mariage, (Burton *et al.*, 1996), le chromosome Y présente une plus forte structure génétique que l'ADN mitochondrial, c'est à dire plus d'homogénéité à l'intérieur des populations mais plus de variations entre populations (Salem *et al.*, 1996, Seielstad *et al.*, 1998, Perez-Lezaun *et al.*, 1999, Kayser *et al.*, 2003, Malyarchuk *et al.*, 2004, Nasidze *et al.*, 2004, Nasidze *et al.*, 2005, Wilkins & Marlowe, 2006, Chaix *et al.*, 2007). Dans les sociétés matrilocales, où c'est l'homme qui se déplace préférentiellement, l'inverse est également vérifié (Oota *et al.*, 2001, Hamilton *et al.*, 2005, Destro-Bisol *et al.*, 2004).

Les différences de structure génétique entre hommes et femmes peuvent donc refléter les contraintes imposées par chaque société sur les choix de conjoints et les migrations des individus. Cependant, de nombreuses études donnent des résultats contradictoires sur la contribution de ces phénomènes sexe-spécifiques à échelle globale (Seielstad *et al.*, 1998, Dupanloup *et al.*, 2003, Wilder *et al.*, 2004a, Wilder *et al.*, 2004b, Ramachandran *et al.*, 2004). Une meilleure compréhension des différences de structure génétique entre hommes et femmes, de leurs causes et de leurs impacts sur l'évolution des populations humaines sera précisément l'objet de notre étude dans la première partie de cette thèse.

Le rôle des adaptations dans la différenciation génétique

D'autres questions se posent également sur la mesure dans laquelle les populations humaines ont eu le temps de se différencier d'un point de vue biologique, par adaptation locale à leur environnement. Nous avons en effet vu qu'un certain pourcentage des différences génétiques s'explique par des variations entre populations. Mais ces différences sont-elles liées à différentes contraintes sur la morphologie ou la physiologie des individus ? Les populations humaines ont en effet dû s'adapter à des environnements variés, et il semblerait logique qu'il y ait eu des modifications génétiques locales causées par différentes pressions climatiques, alimentaires ou infectieuses. Les écarts observés de prévalence de certaines maladies entre populations humaines, ainsi que les différences de tolérance aux médicaments ont en tout cas fortement stimulé les recherches dans ce sens. Cependant, le rôle de l'adaptation génétique dans l'évolution de l'Homme n'est pas encore clairement élucidé

(Coop *et al.*, 2009, Hofer *et al.*, 2009). La recherche d'adaptations génétiques locales à des environnements différents sera précisément l'objet de la deuxième partie de cette thèse.

La génétique des populations a donc finalement soulevé autant de nouvelles questions sur l'histoire évolutive de l'Homme qu'elle n'a aidé à en résoudre. Nous avons vu qu'une certaine part de la variabilité génétique humaine peut s'expliquer par des différences entre populations. Il est maintenant crucial de comprendre les facteurs responsables de ces différences génétiques, aussi peu nombreuses soient-elles, et leurs conséquences sur les populations humaines. Cette thèse a justement pour ambition générale de contribuer à répondre à cette question, en étudiant l'évolution des populations humaines à travers leur diversité génétique, dans un cadre particulier : celui des populations d'Asie Centrale.

II. Notre terrain d'étude : l'Asie Centrale

Eloignée de toutes les mers pour la majeure partie de son territoire, l'Asie Centrale est une zone aux conditions naturelles hautement variables, présentant à la fois des déserts, des steppes et des montagnes. Elle est délimitée au nord par la Taïga sibérienne, à l'est par les chaînes montagneuses du Pamir et du Tien-Shan (allant jusqu'à 7 000 mètres d'altitude), au sud par les déserts iraniens et les montagnes afghanes, et à l'ouest par la mer Caspienne. L'idée de l'Asie Centrale en tant que région distincte dans le monde a été introduite par Alexander von Humboldt, naturaliste et explorateur allemand, en 1843. Les frontières de cette région sont depuis sujettes à de multiples définitions et aucune n'est universellement acceptée. Malgré l'incertitude liée à la définition de ses frontières, le plus souvent, l'Asie Centrale est définie par les cinq ex-républiques d'URSS (Union des républiques socialistes soviétiques) : le Kirghizistan, l'Ouzbékistan, le Tadjikistan, le Turkménistan et le Kazakhstan. Dans un sens plus large et prenant en compte une continuité historique et culturelle, l'Asie Centrale inclut également d'autres régions telles que le sud de la Sibérie, la Mongolie, l'ouest de la Chine, le nord de l'Inde, le Pakistan, l'Afghanistan et le nord-est de l'Iran (voir figure 1 ci-dessous). Cette deuxième définition est souvent traduite par « Inner Asia » en anglais.

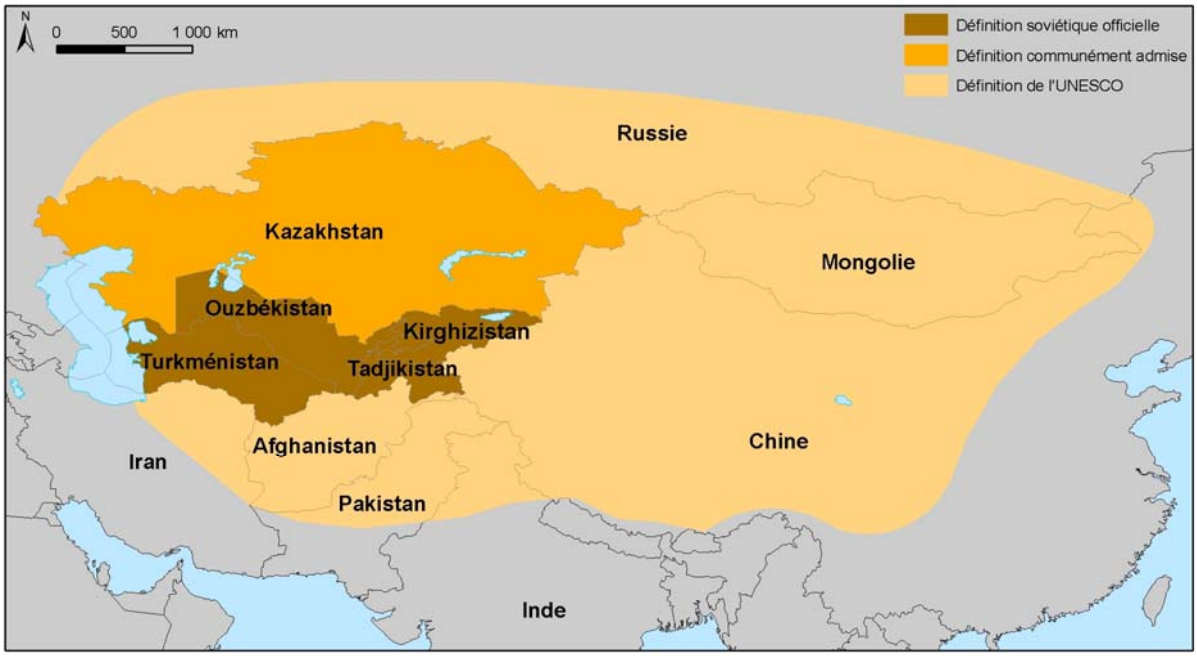


Figure 1 : Différentes définitions de l'Asie Centrale. Source : Sylvain Théry

Pour notre étude, les missions de terrain ont été effectuées au Kirghizistan, en Ouzbékistan et au Tadjikistan, nous donnant accès aux ethnies suivantes : les Kirghiz, les Ouzbeks, les Kazakhs, les Karakalpaks, les Turkmènes et les Tadjiks. Nous ne parlerons donc pas ici d'autres ethnies d'Asie Centrale plus à l'est comme les Ouïgoures ou plus au sud comme les Pashtounes et autres peuples du Pamir.

Diversité linguistique

Originellement habitée par des peuples sédentaires de langues iraniennes (Sogdiens, Chorasmiens) ou semi-nomades (Scythes, Alains), qui ont certainement laissé place aux actuels peuples iraniens dont les Tadjiks, la région a par la suite subi de nombreuses expansions de peuples de langues turques[2] venant de l'est, dont les actuels Ouzbeks, les Kazakhs, les Kirghiz, les Karakalpaks et les Turkmènes seraient entre autre les représentants. Ce véritable carrefour de migration et de diversité est donc une zone de contact entre deux familles linguistiques, avec les langues indo-iraniennes appartenant à la famille indo-européenne et les langues turques appartenant à la famille altaïque (voir figure 2 ci-dessous).

[2] Notons qu'ici, le terme « turque » fait référence aux langues de la sous-famille altaïque, à ne pas confondre avec le turc moderne de Turquie

Les populations de langue turque seront par la suite appelées populations turco-mongoles, pour rendre compte du haut degré de mélange entre ces deux familles de peuples et de langues au départ voisines.

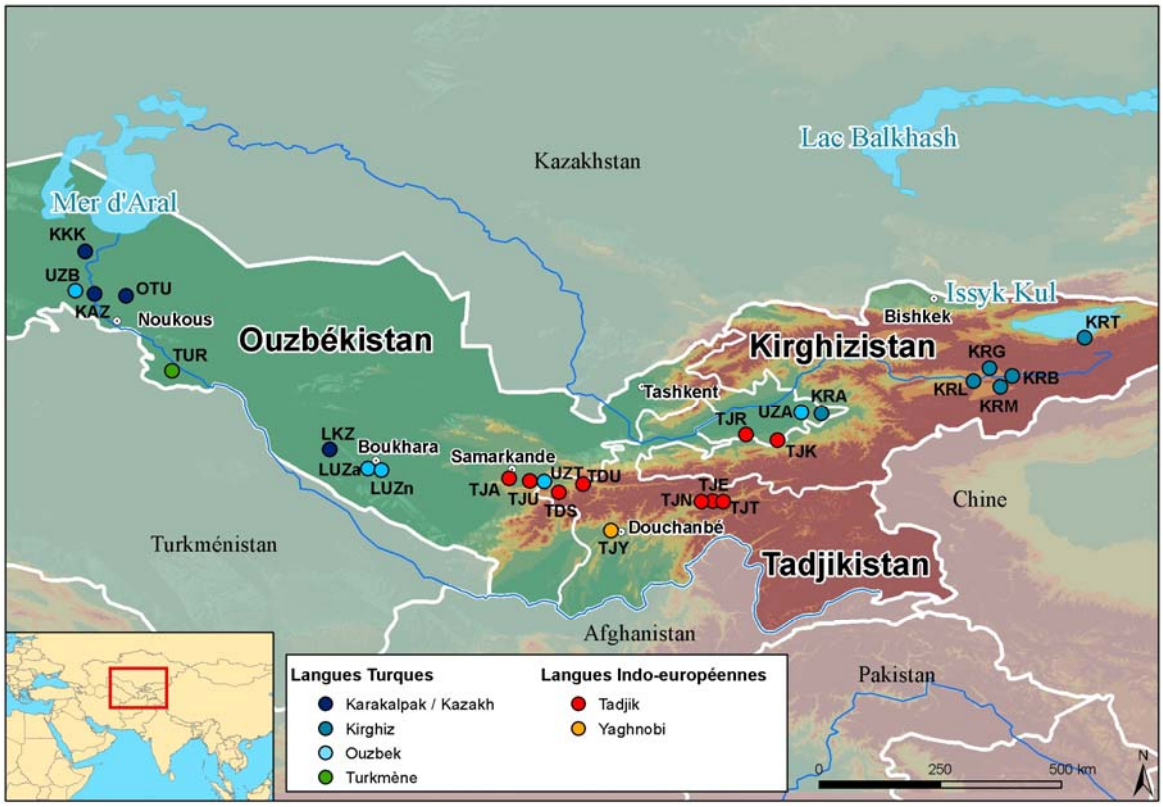


Figure 2 : Zone de rencontre entre familles linguistiques en Asie Centrale.

Source : Sylvain Théry

Diversité des organisations sociales

Ces deux groupes linguistiques correspondent également à des différences marquées d'organisation sociale. Les populations turco-mongoles (Kirghiz, Kazakhs, Turkmènes, Karakalpaks et Ouzbeks) sont organisées selon un mode de filiation patrilinéaire, c'est-à-dire que ces populations sont structurées en groupes de parenté : lignages, clans et tribus, dont l'appartenance est transmise par le père. Ce système est dynamique, dans la mesure où il est redéfini à chaque génération, avec un décompte précis des générations. Cette organisation patrilinéaire se retrouve dans de nombreuses populations mais est particulièrement bien conservée en Asie Centrale (Krader, 1963). Elle est permise par une connaissance incroyablement précise des ancêtres paternels de chacun, connaissance originellement transmise par la tradition orale. Ces populations patrilinéaires pratiquent également une

exogamie lignagère ou clanique, c'est-à-dire que les conjoints sont choisis en dehors du lignage et du clan, mais le plus souvent dans la même tribu (il y a alors endogamie de tribu). A l'inverse des turco-mongols, les indo-iraniens (Tadjiks) sont organisées selon un mode de filiation cognatique, c'est-à-dire que les individus se définissent d'après leur ascendance paternelle et maternelle, et l'unité de parenté fondamentale est la famille, nucléaire ou élargie. De plus, ces populations cognatiques établissent préférentiellement des alliances endogames, c'est-à-dire entre cousins croisés ou parallèles (Jacquesson, 2002).

Diversité des modes de subsistance

A ces différences de langue et d'organisation sociale, s'ajoutent des différences de mode de subsistance, les turco-mongols étant traditionnellement des éleveurs nomades et les indo-iraniens des agriculteurs sédentaires. Parmi les populations turco-mongoles, certaines sont des éleveurs de steppes, comme les Kazakhs, avec un bétail majoritairement constitué de moutons et de vaches, tandis que d'autres sont des éleveurs de montagnes, comme les Kirghiz, avec principalement des moutons et des chevaux. Les Turkmènes, quant à eux, ont un mode de subsistance plus complexe. En effet, traditionnellement, les sociétés turkmènes sont divisées en éleveurs et agriculteurs. Cette division se retrouve à l'intérieur de chaque population, parfois même de chaque famille, et les individus alternent constamment entre ces deux modes de vie (Wood, 2002). L'élevage est principalement basé sur les chameaux, mais les moutons ont pris une importance considérable depuis le 19ème siècle. Ces populations Turkmènes seront classées avec les autres éleveurs turco-mongols du fait de l'importance de l'élevage dans leur régime alimentaire, et de leur affiliation linguistique. Cependant, les résultats pour cette ethnie seront interprétés avec précaution.

Aujourd'hui, les populations traditionnelles d'éleveurs nomades sont toutes sédentarisées, dès le 16ème siècle (ou avant) pour les Ouzbeks, et autour du 19ème siècle pour les autres (sédentarisation de force par les Russes). Ces événements de sédentarisation se sont accompagnés de transitions vers l'agriculture et de modifications plus ou moins prononcées du mode de vie. Par exemple, les Kirghiz de l'est de l'Afghanistan, dans les montagnes du Pamir, étaient, avant 1949, caractérisés par un intense pastoralisme et une forte mobilité, et pouvaient se déplacer jusqu'à 150 km par an pour changer de campement (par exemple entre l'hiver et l'été), leur permettant ainsi de rencontrer des populations dans des vallées éloignées. Cependant, l'ethnographie de Shahrani (1979) a montré que les Kirghiz

effectuaient à cette période plus récente des migrations moins importantes et non uniformes, avec une distance maximale parcourue par an comprise entre 15 et 35 km. Ainsi, les changements politiques ont fortement altéré les modes de migration de ces populations nomades. Shahrani observe également que malgré ces profonds changements, une continuité historique forte existe dans l'identité sociale de ces ethnies.

III. Objectifs de la thèse

L'Asie Centrale a donc la particularité d'être peuplée par deux groupes ethniques différents, les Turco-mongols et les Indo-iraniens, qui, bien que cohabitant sur un même territoire, parlent des langues distinctes, ont des organisations sociales contrastées et ont adopté des modes de subsistance différents. Ceci nous permet de poser une question originale quant à l'évolution des populations humaines : quelle influence a le mode de vie sur la diversité génétique humaine ?

Ce terrain d'étude nous fournit donc un bon modèle pour comprendre comment les facteurs culturels peuvent créer et/ou maintenir des différences génétiques entre populations. Cette thèse, à travers la comparaison de la diversité génétique d'ethnies aux modes de vie contrastés, suit deux axes majeurs : l'étude de l'histoire démographique et de l'histoire adaptative des populations humaines.

1) Tout d'abord, nous allons décrire la répartition générale de la diversité génétique neutre en Asie Centrale, et nous porterons une attention particulière à l'histoire du peuplement de l'Asie Centrale, en relation avec les populations présentes dans les aires géographiques voisines. Ensuite, notre intérêt majeur sera de comprendre les différences de structure génétique entre hommes et femmes, dans des populations ayant adopté des organisations sociales différentes (pratiques matrimoniales, règles d'héritage, etc.). Il s'agira d'abord de comprendre dans quelle mesure la structure génétique observée sur différents systèmes génétiques (autosomes, chromosomes X et Y, ADN mitochondrial) dans les populations humaines est liée à des différences d'histoire démographique entre hommes et femmes (migration et/ou dérive), et ensuite d'identifier les facteurs culturels responsables d'une partie au moins de ces différences, en comparant des ethnies aux organisations sociales distinctes.

2) L'objectif de la seconde partie est de comprendre si les populations d'Asie Centrale ont subi différentes pressions de sélection suite aux changements alimentaires ayant accompagné le Néolithique, transition culturelle majeure d'un mode de vie de prédateur (chasseur, cueilleur et/ou pêcheur) vers un mode de vie de producteur (agriculteur et/ou éleveur). Pour cela, nous allons étudier des gènes candidats sous l'influence potentielle de pressions de sélection liées à l'alimentation, pour voir s'il existe des adaptations génétiques différentes entre populations n'ayant pas le même mode de subsistance (éleveurs semi-nomades, agriculteurs sédentaires). Cette question est d'autant plus intéressante qu'elle permet de mieux comprendre pourquoi les populations actuelles n'ont pas les mêmes susceptibilités aux maladies ou les mêmes tolérances aux médicaments. Comme nous l'avons vu, la recherche de contraintes sélectives sur des endroits particuliers du génome nécessite une bonne connaissance d'ensemble de la diversité génétique des populations étudiées. Ainsi, le premier chapitre de la thèse représente également un préalable indispensable au deuxième. Cette étude permettra finalement de tester si la sélection naturelle est une force prédominante dans l'évolution de l'Homme, ou si l'évolution culturelle a permis aux populations de s'adapter sans modifier leur biologie.

L'étude comparative des ethnies d'Asie Centrale, fil conducteur de cette thèse, a donc pour ambition de mettre en lumière les différentes forces évolutives qui ont façonné la diversité génétique de ces populations et ainsi d'inférer leur histoire passée. De par sa position stratégique, l'Asie Centrale a certainement eu un rôle majeur dans les interactions passées et présentes entre populations ; elle s'avère donc cruciale pour la compréhension globale de l'histoire du peuplement de notre espèce. Malgré cela, ce terrain d'étude a été assez peu étudié au préalable d'un point de vue génétique, notamment à cause de la main mise soviétique sur la région jusqu'en 1991.

Finalement, cette thèse s'inscrit dans un cadre plus large de travaux effectués en Asie Centrale par l'équipe d'Evelyne Heyer, qui a mené de nombreuses campagnes d'échantillonnage depuis 2001 et mis en place des collaborations durables et fructueuses au Tadjikistan avec Firuza Nasyzova, en Ouzbékistan avec l'équipe de Tatiana Hegay et au Kirghizistan avec l'équipe d'Almaz Aldashev. J'ai ainsi eu la chance de participer à quatre missions de terrain : en mai-juin 2007 au Kirghizistan, en octobre-novembre 2008 en Ouzbékistan, et en avril-mai et juillet-aout 2009 au Kirghizistan. Plusieurs questions / problèmes bioéthiques se soulèvent lors d'un échantillonnage sur le terrain, d'autant plus

quand il s'agit de génétique et de populations humaines. Les réflexions autour de ces pratiques de terrain sont présentées dans l'annexe 1. Cette discussion est le fruit de réflexions collectives, entre autres lors d'une réunion avec Anne Cambon-Thomsen (équipe « Génomique, santé, société », Inserm U558, Toulouse). Finalement, cette thèse s'alimente non seulement de mes propres travaux de recherche, mais également de résultats apportés par d'autres études faites sur ce même terrain, et de réflexions communes avec les autres chercheurs du laboratoire qui m'ont permis d'enrichir la discussion.

CHAPITRE I . HISTOIRE DÉMOGRAPHIQUE : PEUPLEMENT ET ORGANISATION SOCIALE EN ASIE CENTRALE

- **Annexe 3** : Martinez-Cruz B.[*], Vitalis R.[*], **Ségurel L.**, Austerlitz F., Georges M., Théry S., Quintana-Murci L., Hegay T., Aldashev A., Nazyrova F. & Heyer E. In the heartland of Eurasia : the multi-locus genetic landscape of Central Asian populations. A soumettre à *European Journal of Human Genetics*

- **Annexe 4** : **Ségurel L.**, Martinez-Cruz B., Quintana-Murci L., Balaresque P., Georges M., Hegay T., Aldashev A., Nazyrova F., Jobling M.A., Heyer E. & Vitalis R. Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. PLoS Genet 2008 Sep 26; 4(9):e 1000200.

[*] Ces auteurs ont également contribué à ce travail

I. Etat de l'art

Peuplement ancien de l'Asie Centrale

En Asie Centrale, les premiers restes humains sont attestés dès 800 000 BP[3] à Kul'dara, au sud du Tadjikistan (Ranov *et al.*, 1995). Les ressources en eau semblent, dans cette région aride à semi-aride, avoir été un facteur déterminant de la répartition des Hommes sur ces territoires (Masson, 1992). Des restes archéologiques ont également été trouvés pendant le Paléolithique moyen (de 300 000 à 30 000 BP), dont notamment des restes de Néandertal à Teshik-Tash (Okladnikov, 1940) datant de 150 000 BP (Grun & Stringer, 2000), définissant ainsi l'aire de répartition la plus orientale de Néandertal, longtemps considéré comme exclusivement européen. Depuis, d'autres restes humains trouvés dans l'Altaï, en Sibérie, ont agrandi l'aire de répartition des Néandertaliens (Krause *et al.*, 2007).

Après cette première colonisation de l'Eurasie par des Hommes archaïques, une autre vague, cette fois-ci d'Hommes modernes, a de nouveau permis de coloniser toute la planète. Ces Hommes modernes ont d'abord été trouvés en Ethiopie, datant d'environ 196 000 BP (McDougall *et al.*, 2005). Mais c'est seulement à partir de 40 000 BP que les Hommes modernes apparaissent sur une aire géographique plus large, en Afrique du Nord et en Eurasie (Trinkaus, 2005). Les données génétiques suggèrent également que l'Homme moderne n'a pas migré hors d'Afrique avant 60 000 - 40 000 ans (voir la figure 3 ci-dessous,Cavalli-Sforza & Feldman, 2003). En dehors des zones insulaires du sud-est, les premières évidences d'Hommes modernes en Asie concernent des restes en Chine, en Corée et au Japon, entre 35 000 et 25 000 BP. Ces données permettent donc d'avancer l'hypothèse d'une première sortie assez précoce le long des côtes asiatiques jusqu'en Océanie (Quintana-Murci *et al.*, 1999, Forster & Matsumura, 2005), tandis que les autres migrations continentales auraient eu lieu plus tardivement.

De par sa localisation, l'Asie Centrale a dû jouer un rôle central dans ces évènements de colonisation, et pourtant son peuplement est mal connu (Nei & Roychoudhury, 1993, Cavalli-Sforza *et al.*, 1994, Comas *et al.*, 1998, Karafet *et al.*, 2001, Wells *et al.*, 2001, Cordaux *et al.*, 2004, Macaulay *et al.*, 2005). Nous ne savons en effet pas si cette région correspond à une zone « source » d'où sont ensuite parties les populations vers l'Europe et l'Asie de l'Est ou s'il s'agit d'une zone « puit», c'est-à-dire de rencontre plus tardive de

[3] Les dates présentées en 'BP' (pour 'Before Present'), obtenues par datation au carbone 14, prennent pour référence l'année 1950.

populations venant d'Europe et d'Asie de l'Est. D'autres éléments permettent de formuler une troisième hypothèse, celle de peuplements d'est en ouest, qui définirait donc une vague de migration venant d'Asie de l'Est, traversant l'Asie Centrale, pour arriver en Europe (Chaix *et al.*, 2008).

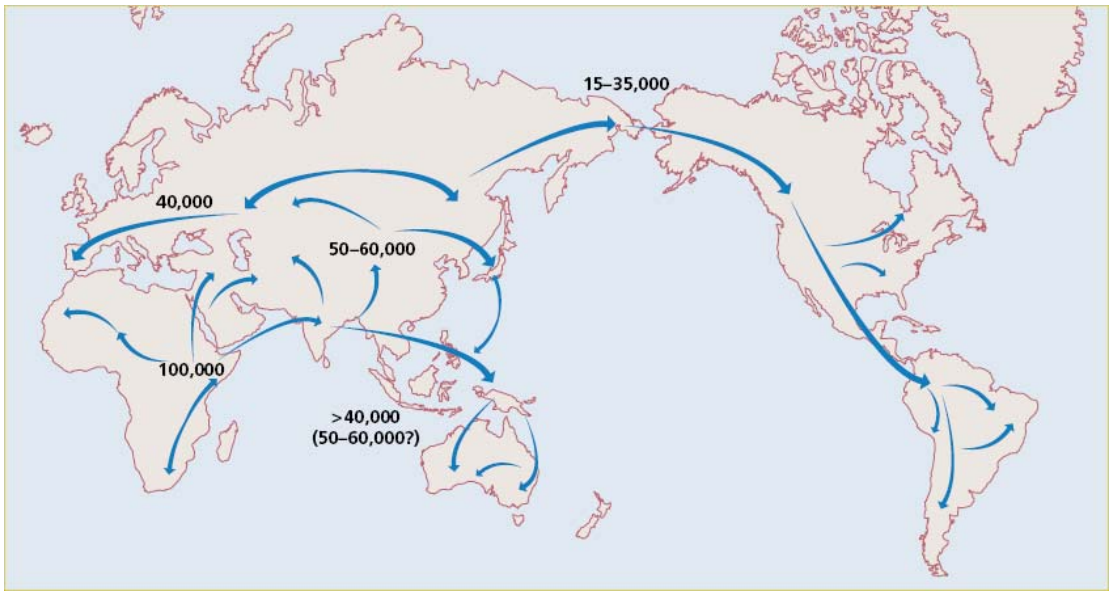


Figure 3 : Scénario pour l'origine de l'Homme moderne et ses voies de colonisation d'après des données génétiques (Cavalli-Sforza & Feldman, 2003)

Mouvements de populations pendant la période historique

Les premières traces historiques sur l'Asie Centrale décrivent les invasions des Ariens de l'actuel Turkménistan vers l'Iran autour de 1 300 BCE[4], pour former l'Empire Perse autour de 700 BCE. Cet ensemble de peuples nomades ensuite appelés les Scythes (ou Saka, ou Saces), ont été décrits comme ayant des traits morphologiques européens et parlant des langues indo-iraniennes. Ils vivaient *a priori* dans les steppes eurasiatiques, depuis l'Ukraine jusqu'à l'Altaï. Par la suite, de nombreuses vagues d'expansions venant de l'est sont arrivées en Asie Centrale, bien qu'il soit difficile de situer ces migrations dans le temps. Il semble qu'autour de 200 BCE, des invasions liées au Huns aient apporté le phénotype asiatique en Asie Centrale. A la même période, l'Asie Centrale est clairement devenue une région dynamique d'échanges de personnes, de biens et d'idées entre l'Europe et l'Asie de l'Est, à travers notamment la « route de la soie », établie par les Chinois autour de 200 BCE, et qui reliera le bassin Méditerranéen à l'Asie de l'Est pendant plus de 16 siècles. A partir du 3ème et

[4] Les dates présentées en 'BCE' (pour 'Before Christ Era') correspondent en français à 'av. J.-C.' ; celles en 'CE' correspondent à 'ap. J.-C.'

4ème siècle, de nombreuses vagues d'expansions turques entraînent des changements linguistiques en Asie Centrale, qui ne cessera par la suite d'être tiraillée entre des expansions successives, parfois d'empires venant de l'ouest s'étendant beaucoup plus à l'est, (comme les invasions musulmanes venant de Perse entre le 7ème et 10ème siècle), ou inversement d'empires venant de l'est et s'étendant beaucoup plus à l'ouest, comme l'empire mongol de Gengis Khan autour du 13ème siècle, l'empire le plus vaste de tous les temps, allant de la Chine jusqu'en Europe de l'est. Gengis Khan unit d'ailleurs plusieurs tribus nomades de l'Asie de l'Est et de l'Asie Centrale sous une nouvelle identité commune, en tant que « mongoles ».

De tous ces mouvements de populations, résulte une diversité ethnique considérable en Asie Centrale, avec actuellement des populations de langue indo-iranienne supposées être les descendants des premiers peuples iraniens décrits dès 1 300 BCE, mais peut-être présents dès le Néolithique, et des populations de langues turques installées dans la région suite aux invasions venant de l'est autour du 4ème siècle.

Notons que l'appellation « Ouzbek » regroupe à l'origine deux types de populations : certaines populations présentes environ dès le 4ème siècle en Asie Centrale, et mélangées depuis aux peuples iraniens locaux, et d'autres populations turco-mongoles arrivées plus tardivement, vers le 16ème siècle (Soucek, 2000).

Apport des données génétiques

Les études génétiques ont montré que la diversité génétique en Asie Centrale fait partie des plus fortes diversités en dehors d'Afrique (Comas *et al.*, 1998, Wells *et al.*, 2001, Hammer *et al.*, 2001). Cette forte diversité provient certainement de la place centrale de cette région, qui permet un brassage potentiel avec des populations venant à la fois d'Europe, du Moyen-Orient et d'Asie de l'Est. Concernant le peuplement de la région, les études basées sur le chromosome Y ont montré que les vagues de migration à l'origine du peuplement de l'Eurasie sont originaires d'Asie Centrale (Comas *et al.*, 1998, Hammer *et al.*, 2001, Zerjal *et al.*, 2002), tandis que les études basées sur l'ADN mitochondrial proposent le schéma inverse : une origine métissée des populations d'Asie Centrale, à partir de populations plus ancestrales en Eurasie (Comas *et al.*, 1998, Comas *et al.*, 2004, Lalueza-Fox *et al.*, 2004, Perez-Lezaun *et al.*, 1999). Les marqueurs uni-parentaux nous livrent donc des résultats différents, ce qui peut entre autres s'expliquer par des histoires démographiques différentes

entre hommes et femmes, elles-mêmes causées par l'organisation sociale des populations (Chaix *et al.*, 2004, 2007).

Objectifs de l'étude

Etant donné que chaque marqueur uni-parental ne représente qu'un unique marqueur non recombinant, des effets d'entraînement sélectif ne sont pas à exclure et peuvent donc, entre autres, conduire à des résultats biaisés (Cummins, 2001, Ballard & Whitlock, 2004, Bazin *et al.*, 2006, Pakendorf & Stoneking, 2005, Harpending, 2006, Balloux, 2009). Ainsi, il semble maintenant important d'obtenir des données multi-locus à partir des autosomes, afin de caractériser l'origine génétique des populations d'Asie Centrale et de les comparer aux populations avoisinantes pour mieux comprendre les mouvements passés des populations dans la région. Nous allons donc dans cette partie commencer par décrire la diversité génétique neutre des populations d'Asie Centrale, à partir des autosomes, afin de comprendre la structure générale de ces populations. Cette analyse nous permettra d'inférer les processus historiques en jeu dans ces populations, ce qui est indispensable pour interpréter les résultats obtenus dans le reste de la thèse. Ces premiers résultats, présentés en annexe 3, sont issus d'analyses majoritairement réalisées par Begoña Martinez-Cruz, post-doctorante au laboratoire, sur des données génétiques que nous avons conjointement produites.

Dans un second temps, nous analyserons ces données multi-locus autosomales conjointement avec les données sur le chromosome X, pour des populations où nous avons également séquencé une partie de l'ADN mitochondrial et génotypé des marqueurs du chromosome Y. Notre objectif est d'étudier la structure génétique sexe-spécifique, pour mieux comprendre les éventuelles différences d'histoire démographique entre hommes et femmes. Nous pourrons également tirer avantage de la diversité des organisations sociales en Asie Centrale, pour tester directement l'influence causale de différents traits de l'organisation sociale sur cette structure sexe-spécifique.

.

II. Matériel et méthodes

A. Populations échantillonnées

1075 individus ont été échantillonnés entre 2001 et 2007 au Kirghizistan, en Ouzbékistan et au Tadjikistan par l'équipe d'Evelyne Heyer (voir la figure 4 et le tableau 1 ci-dessous). Ils sont répartis dans 26 populations : six populations kirghizes (KRA, KRB, KRG, KRL, KRM et KRT), deux populations kazakhes (KAZ et LKZ), deux populations karakalpakes (KKK et OTU), une population turkmène (TUR), cinq populations ouzbèkes (LUZa, LUZn, UZA, UZB et UZT) et dix populations tadjikes (TDU, TDE, TDS, TJA, TJE, TJN, TJR, TJT, TJU et TJY).

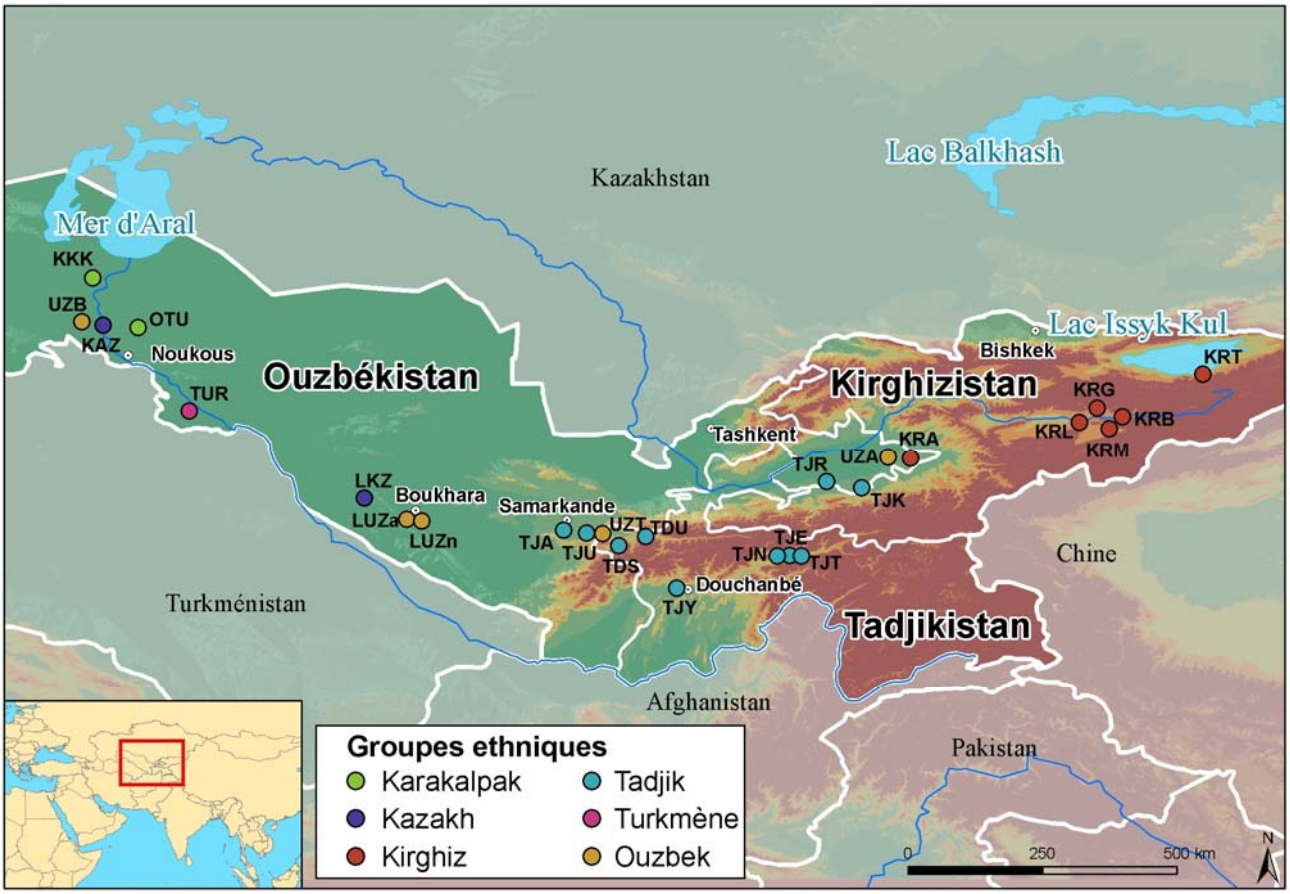


Figure 4 : Répartition géographique des 26 populations échantillonnées répartis dans 6 groupes ethniques. Source : Sylvain Théry

Ethnie	Acronyme	Lieu d'échantillonnage	Long.	Lat.	n_X	n_A	n_Y	n_{mt}
Tadjik	TJA	Frontière Ouzbékistan / Tadjikistan	**39.54**	**66.89**	26	31	32	32
Tadjik	TJU	Frontière Ouzbékistan / Tadjikistan	**39.50**	**67.27**	27	29	29	29
Tadjik	TJR	Frontière Ouzbékistan / Kirghizistan	**40.36**	**71.28**	30	29	29	29
Tadjik	TJK	Frontière Ouzbékistan / Kirghizistan	**40.25**	**71.87**	26	26	35	40
Tadjik	TJE	Nord du Tadjikistan	**39.12**	**70.67**	29	25	27	31
Tadjik	TJN	Nord du Tadjikistan	**38.09**	**68.81**	33	24	30	35
Tadjik	TJT	Nord du Tadjikistan	**39.11**	**70.86**	31	25	30	32
Tadjik	TDS	Frontière Ouzbékistan / Tadjikistan	**39.28**	**67.81**	30	25	31	31
Tadjik	TDU	Frontière Ouzbékistan / Tadjikistan	**39.44**	**68.26**	40	25	31	40
Tadjik	TJY	Ouest du Tadjikistan	**38.57**	**68.78**	39	25	36	40
Tadjiko-Ouzbek	LUZa	Centre de l'Ouzbékistan	**39.73**	**64.27**	14	20	13	16
Tadjiko-Ouzbek	LUZn	Centre de l'Ouzbékistan	**39.70**	**64.38**	12	20	11	15
Karakalpak	KKK	Ouest de l'Ouzbékistan	**43.77**	**59.02**	56	45	54	55
Karakalpak	OTU	Ouest de l'Ouzbékistan	**42.94**	**59.78**	49	45	54	53
Kazakh	KAZ	Ouest de l'Ouzbékistan	**43.04**	**58.84**	47	49	50	50
Kazakh	LKZ	Centre de l'Ouzbékistan	**40.08**	**63.56**	20	25	20	31
Kirghiz	KRA	Frontière Ouzbékistan / Kirghizistan	**40.77**	**72.31**	31	45	46	48
Kirghiz	KRG	Est du Kirghizistan	**41.60**	**75.80**	20	18	20	20
Kirghiz	KRM	Est du Kirghizistan	**41.45**	**76.22**	21	21	22	26
Kirghiz	KRL	Est du Kirghizistan	**41.36**	**75.5**	36	22	40	23
Kirghiz	KRB	Est du Kirghizistan	**41.25**	**76.00**	31	24	31	30
Kirghiz	KRT	Est du Kirghizistan	**42.16**	**77.57**	33	37	37	29
Turkmène	TUR	Ouest de l'Ouzbékistan	**41.55**	**60.63**	42	47	51	51
Ouzbek	UZA	Frontière Ouzbékistan / Kirghizistan	**40.77**	**72.31**	39	25	36	36
Ouzbek	UZB	Ouest de l'Ouzbékistan	**43.04**	**58.84**	50	35	49	40
Ouzbek	UZT	Frontière Ouzbékistan / Tadjikistan	**39.49**	**67.54**	37	25	35	39

Tableau 1 : Informations sur les 26 populations échantillonnées. Les populations indo-iraniennes sont en blanc (Tadjiks et Tadjiko-Ouzbeks) ; Les populations turco-mongoles sont en gris (Kazakhs, Karakalpaks, Kirghiz, Turkmènes et Ouzbeks). Long., longitude; Lat., latitude. n_X , n_A, n_Y et n_{mt} : nombre d'individus analysés pour le chromosome X, les autosomes, le chromosome Y et l'ADN mitochondrial, respectivement.

Des données généalogiques ont été collectées avant l'échantillonnage pour ne retenir que des individus non apparentés à la deuxième génération. Les données linguistiques ont été récoltées par Philippe Mennecier, et les données ethnologiques ont été récoltées par divers ethnologues locaux dans chaque pays, avec en partie l'aide de Svetlana Jacquesson et de Nicolas Lescureux. Tous ces individus ont signé un consentement éclairé pour participer à cette étude.

Dans deux populations Ouzbeks, échantillonnées autour de Boukhara (LUZa et LUZn), une enquête linguistique poussée a permis de conclure que la langue parlée « à la maison » était le tadjik (bien qu'ils sachent également parler ouzbek), et que leur auto-appellation d' « Ouzbek » venait surtout du fait qu'ils habitaient en Ouzbékistan. La question de l'ethnicité n'est clairement pas évidente dans ces pays où les frontières et les identités ont souvent été imposées. Ici, ces deux populations seront appelées Tadjiko-Ouzbeks et classées dans les populations indo-iraniennes, du fait de leur appartenance linguistique (voir tableau 1).

B. Analyses moléculaires

L'ADN génomique a été extrait à partir d'échantillons sanguins ou salivaires, par une procédure d'extraction standard au phénol chloroforme (Maniatis *et al.*, 1982). Le génotypage de 11 STRs (*Short Tandem Repeats*) sur le chromosome Y, pour 886 individus, et le séquençage du premier segment hypervariable de la région de contrôle de l'ADN mitochondrial, HVS-1 (*Hyper Variable Segment 1*), pour 916 individus, a été le fruit d'un travail collaboratif entre l'équipe de Lluis Quintana-Murci (Unité de Génétique Evolutive Humaine, CNRS URA 3012, Institut Pasteur, France), celle de Mark Jobling (Département de Génétique, Université de Leicester, UK) et celle d'Evelyne Heyer. Pour les 27 STRs autosomaux, le génotypage, pour un sous-groupe de 767 individus, a été effectué conjointement par Begoña Martinez-Cruz et moi-même, avec l'aide précieuse de Myriam Georges. Pour le chromosome X, le génotypage de 9 STRs, pour 697 individus, a été effectué par moi-même.

III. Diversité génétique neutre en Asie Centrale

Dans cette étude, les données de 27 marqueurs microsatellites autosomaux sur 767 individus répartis dans les 26 populations d'Asie Centrale ont été comparées aux données disponibles pour ces mêmes marqueurs dans 25 populations mondiales de la base de données HGDP-CEPH (Human Genetic Diversity Panel – Centre d'Etude du Polymorphisme Humain).

Répartition de la diversité génétique

Les populations d'Asie Centrale présentent une diversité génétique intra-population (H_e = 0.795) comparable à celle des populations Européennes, du Pakistan, ou du Moyen-Orient (H_e = 0.775, 0.819 et 0.826, respectivement), mais plus forte que celle d'Asie de l'Est (H_e = 0.706).

La répartition de la diversité génétique en Asie Centrale est significativement expliquée par l'affiliation linguistique (AMOVA, F_{CT} = 0.008, p < 0.0001), avec une forte différenciation entre deux groupes de populations : les turco-mongols et les indo-iraniens. Ainsi, le fait de ne pas parler la même langue semble être une barrière importante aux migrations entre groupes. Cependant, les groupements linguistiques correspondent également à des organisations sociales et des modes de subsistance contrastés, et nous ne pouvons pas exclure le rôle déterminant du mode de vie pour créer des barrières de reproduction entre ces populations, respectivement d'éleveurs patrilinéaires et d'agriculteurs cognatiques. L'affiliation ethnique explique d'ailleurs également une part significative des distances génétiques (F_{CT} = 0.005, p = 0.002). Le rôle des ces affiliations culturelles comme facteur de différenciation génétique est particulièrement bien décrit dans une étude récente de Heyer *et al* sur l'émergence des groupes ethniques en Asie Centrale (Heyer *et al.*, 2009). Le fort niveau de différenciation génétique pourrait également être expliqué par un flux de gène limité entre ces groupes, conséquence de la migration récente des Turco-mongols dans la région, qui n'aurait alors pas laissé le temps à un brassage important.

De manière étonnante, les distances génétiques entre populations, liées aux distances linguistiques et ethniques, ne sont ici pas significativement corrélées aux distances géographiques (test de Mantel, p = 0.25). Ce manque de corrélation est également retrouvé au sein des Indo-iraniens et au sein des Turco-mongols (p = 0.92 et 0.45, respectivement). Pour les Turco-mongols, ceci peut être expliqué par la migration récente de ces populations dans la

région, ou bien par les bouleversements de leur mode de vie engendré par les Russes, qui ont mis un terme aux migrations nomades et ont abouti à une sédentarisation quelque peu aléatoire des populations au cours du 19ème siècle. Pour les Indo-iraniens, il est plus difficile de comprendre pourquoi ces populations n'ont pas établi de patron d'isolement par la distance, pourtant typiquement trouvé dans les populations humaines (Bosch *et al.*, 2006, Manica *et al.*, 2005, Prugnolle *et al.*, 2005).

L'Asie Centrale dans l'Eurasie

Ces groupes de populations cohabitant sur un même territoire sont également caractérisés par des affiliations géographiques assez différentes (voir la figure 5 ci-dessous).

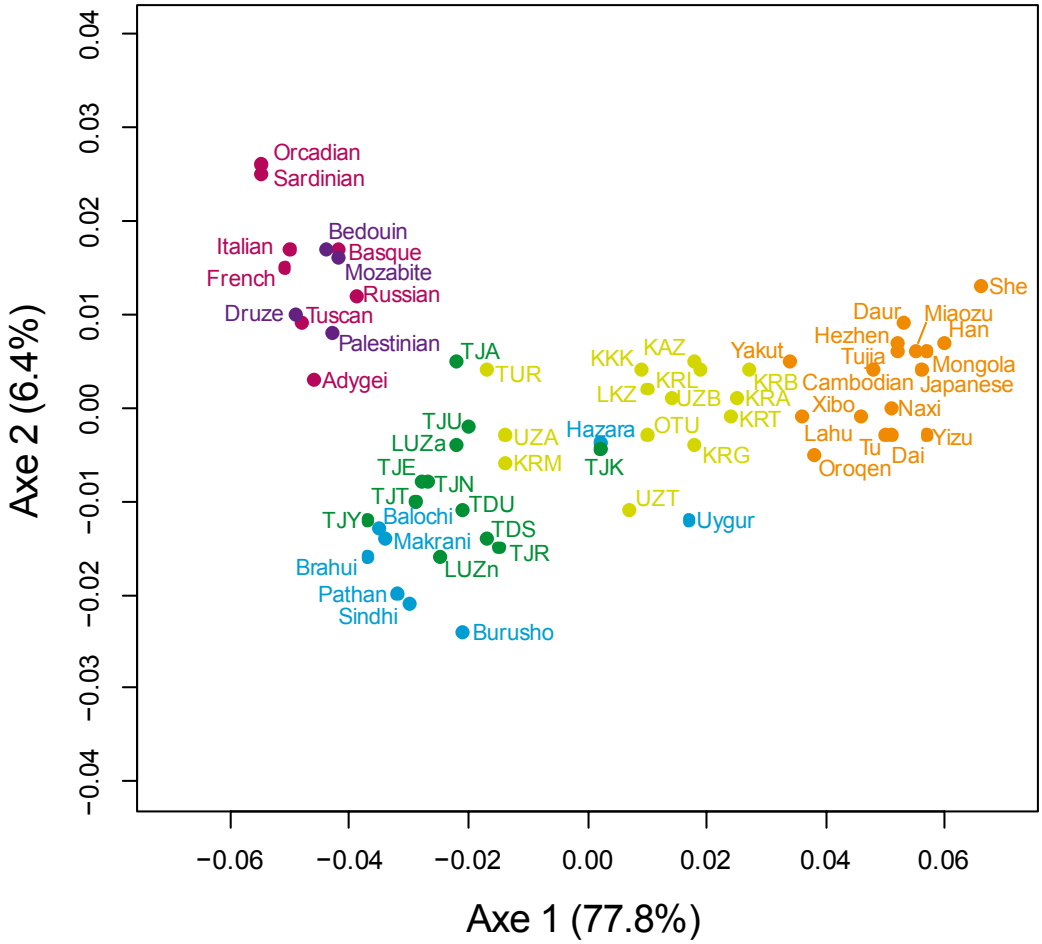


Figure 5 : Analyse en composante principale de la structure génétique de 51 populations mondiales sur 27 marqueurs autosomaux.

Les Turco-mongols (en vert clair) montrent une forte proximité avec les populations est-asiatiques (en orange), tandis que les Indo-iraniens (en vert foncé) se rapprochent plutôt des

populations de l'ouest (les populations du Pakistan sont en bleu, les européennes en rose, et celles du Moyen-Orient en violet). Certaines populations indo-iraniennes sont extrêmement proches des populations du Pakistan, tandis que d'autres se rapprochent plus de l'Europe et du Moyen-Orient.

Ces résultats confirment l'hypothèse d'une origine récente des Turco-mongols venant de l'est, également supportée par le faible niveau de différenciation entre ces populations. Ces résultats montrent également que l'invasion de l'Asie Centrale par des peuples venant de l'est tout au long des deux derniers millénaires (comme les invasions des Huns ou de Gengis Khan) n'a pas clairement abouti à un remplacement des populations Indo-iraniennes locales (Sengupta *et al.*, 2006). La faible proximité entre les populations du Moyen-Orient et les populations indo-iraniennes exclue a priori une origine de ces dernières via le Croissant Fertile. Nous avons également trouvé une importante différenciation génétique au sein des populations indo-iraniennes, ce qui pourrait être du à une origine locale ancienne avec des groupes isolés, ou bien à des origines diverses de ces populations depuis l'ouest et le sud de l'Eurasie. Les données archéologiques supportent plutôt la première hypothèse (Brunet, 1999).

Finalement, les populations Ouzbeks sont réparties entre ces affiliations géographiques différentes. Ceci est certainement du au fait que l'appellation « Ouzbek » regroupe des populations originellement différentes (certaines mélangées aux peuples iraniens, d'autres non). Ainsi, les Ouzbeks seront exclus de la suite de nos analyses, car ils ne constituent pas un groupe ethnique homogène et ont subi des changements culturels qui ont bouleversé leur structure génétique (Chaix *et al.*, 2007).

IV. Influence de l'organisation sociale sur la diversité génétique

A. Problématique

Les études précédentes nous ont permis de nous interroger sur l'histoire générale des populations d'Asie Centrale à partir des autosomes, et leur relation par rapport au reste du continent Eurasiatique, mais l'étude respective de la démographie des hommes et des femmes peut nous apporter de nouvelles informations sur l'histoire de ces populations.

Nous avons vu en effet que dans la majorité des populations humaines, le chromosome Y présente une plus forte structure génétique que l'ADN mitochondrial, ce qui a été majoritairement interprété comme une migration plus forte des femmes, conséquence de la patrilocalité. Cependant, d'autres facteurs démographiques peuvent avoir un effet confondant sur ces résultats. En effet, la structure génétique (exprimée par le paramètre F_{ST}) est influencée par le produit $N_e m$ (Wright, 1931), où N_e est la taille efficace de chaque population et m le taux d'immigration dans chaque population. La structure génétique est donc influencée de la même manière par des différences de migration ou de dérive entre sexes. Ainsi, une plus forte dérive chez les hommes, par exemple du fait de la polygynie (qui traduit le fait que les hommes peuvent en général avoir plusieurs partenaires, mais non les femmes), pourrait engendrer de telles structures sexe-spécifiques. Cependant, la prise en compte d'un niveau réaliste de polygynie, tel que pratiqué dans les populations humaines, ne permet pas d'expliquer les écarts importants observés de structure génétique entre marqueurs (Seielstad *et al.*, 1998). Ainsi, l'hypothèse d'une taille efficace plus importante chez les femmes a été majoritairement négligée par la suite (mais voir Salem *et al.*, 1996, Kayser *et al.*, 2003, Destro-Bisol *et al.*, 2004, Dupanloup *et al.*, 2003, Wilder *et al.*, 2004b, Wilder & Hammer, 2007).

Le chromosome Y et l'ADN mitochondrial ont également des caractéristiques particulières qui compliquent l'interprétation de leur structure génétique, notamment l'hypermutabilité de l'ADN mitochondrial et l'absence de recombinaison sur ces deux marqueurs qui rend impossible la distinction entre les effets démographiques et sélectifs sur ces marqueurs (Cummins, 2001). Les véritables facteurs responsables des différences de structure génétique entre chromosome Y et ADN mitochondrial restent donc pour l'instant mal démêlés, certainement à cause de la limitation méthodologique des marqueurs uni-

parentaux (Ballard & Whitlock, 2004, Bazin *et al.*, 2006, Pakendorf & Stoneking, 2005, Harpending, 2006, Balloux, 2009).

Objectifs de cette étude

D'autres méthodes sont donc nécessaires pour distinguer l'influence respective de ces facteurs et comprendre quel est le véritable lien entre l'organisation sociale des populations humaines (dont la patrilocalité), et leur structure génétique sexe-spécifique. C'est pourquoi nous proposons une nouvelle approche, la comparaison des autosomes et du chromosome X, qui semble particulièrement appropriée pour comprendre la démographie sexe-spécifique (Balaresque & Jobling, 2007). Ces systèmes génétiques subissent en effet tous deux la recombinaison (et donc sont moins susceptibles de subir les effets confondants de la sélection), mais présentent tout de même des différences d'héritabilité (donc des différences sont attendues si la démographie des hommes et des femmes n'est pas la même). De plus, ces systèmes génétiques permettent d'avoir une approche « multi-locus », où plusieurs marqueurs indépendamment influencés par la même histoire sont étudiés, ce qui nous permet notamment d'obtenir des intervalles de confiance sur nos estimations.

Cette étude a été conduite sur 21 populations d'Asie Centrale, toutes patrilocales, c'est-à-dire que la femme migre pour habiter dans le village de son mari. Parmi ces populations, 11 sont turco-mongoles patrilinéaires exogames. Dans ce cas, les individus se définissent selon leur affiliation paternelle, et les choix de conjoints se font préférentiellement en dehors de l'unité sociale (ici le clan). Les 10 autres populations sont indo-iraniennes cognatiques et endogames, donc ici l'affiliation est à la fois paternelle et maternelle, et les conjoints sont préférentiellement choisis à l'intérieur de l'unité sociale. Du fait que ces deux groupes de populations sont patrilocaux, nous pouvons nous attendre à observer une migration plus importante des femmes par rapport aux hommes. Cependant, les règles de choix de conjoints définies par l'exogamie / endogamie ont une forte influence sur ces attendus puisque si la population est complètement endogame, alors le village du mari est le même que celui de la femme, et donc les femmes ne migrent pas plus que les hommes (Kumar *et al.*, 2006).

B. Résultats

Nous avons trouvé sur les marqueurs uni-parentaux obtenus pour 18 populations, que la structure génétique est plus importante sur le chromosome Y que sur l'ADN mitochondrial, à la fois pour les 8 populations patrilinéaires exogames (F_{ST} = 0.177 et 0.010, respectivement) et les 10 populations cognatiques endogames (F_{ST} = 0.069 et 0.034, respectivement). En prenant en compte un modèle de migration en île (Wright, 1931), cet écart de structure génétique peut s'expliquer par une migration et/ou par une taille efficace plus importante chez les femmes. Ainsi, les populations patrilinéaires exogames auraient un nombre efficace de migrants environ 22 fois plus importants chez les femmes, tandis que cet écart serait réduit à un facteur deux chez les populations cognatiques endogames. Nous voyons donc bien que les écarts de structure génétique entre sexes sont très différents selon les populations considérées. Cependant, comme énoncé plus haut, les seuls marqueurs uni-parentaux ne permettent pas de discriminer parmi les différents facteurs responsables de ces différences.

Si l'on considère maintenant les résultats obtenus à partir des marqueurs multi-locus sur 21 populations, nous avons trouvé que chez les 11 populations patrilinéaires, les autosomes ont une structure génétique plus marquée que le chromosome X (F_{ST} = 0.008 et 0.003, respectivement ; test de Wilcoxon p = 0.02), tandis que ce n'est pas le cas pour les 10 populations cognatiques (F_{ST} = 0.014 et 0.013, respectivement; test de Wilcoxon p = 0.36). Or, en théorie, si l'on considère que les hommes et les femmes ont des taux comparables de migration (respectivement m_m, m_f) et de dérive (respectivement N_m et N_f), nous nous attendons à trouver une plus forte différenciation sur le chromosome X, du fait de son nombre moins grand de copies (3/4 de celui des autosomes) et donc de sa plus forte dérive. Pour obtenir un niveau de différenciation génétique plus fort sur les autosomes, nous avons montré, par résolution des expressions de F_{ST} dans un modèle en île, qu'il faut nécessairement considérer une taille efficace plus importante chez les femmes, pour n'importe quel taux de migration considéré (voir figure 6 ci-dessous).

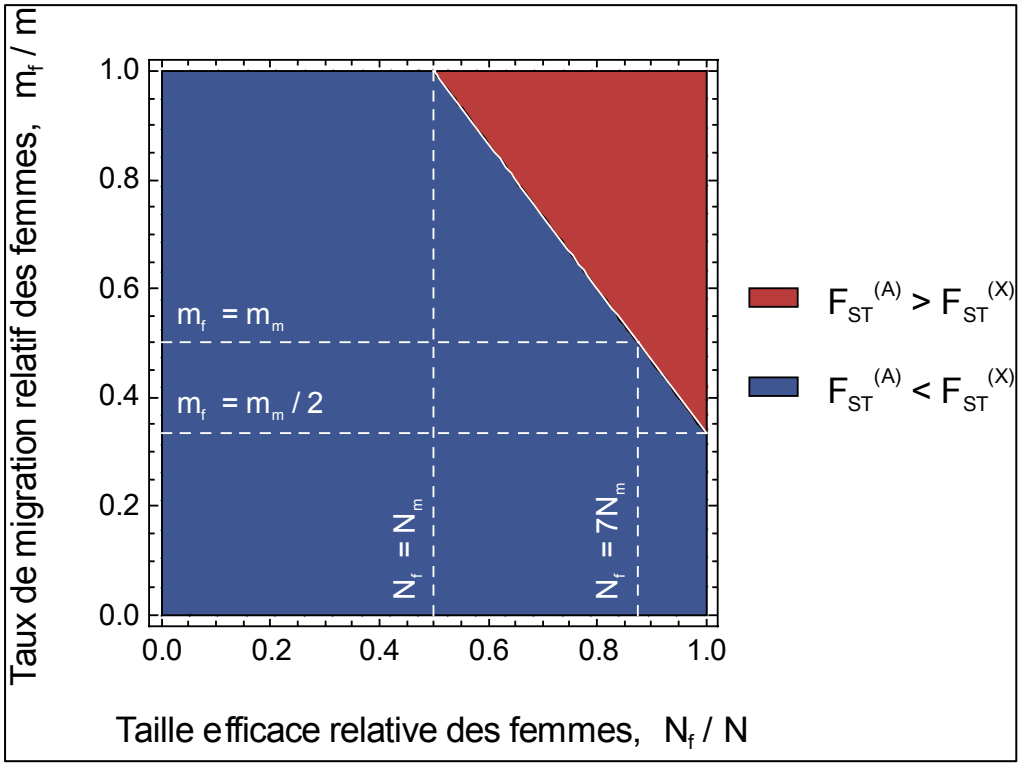


Figure 6 : Diagramme représentant les valeurs attendues de différenciation génétique relative sur les autosomes et le chromosome X, en fonction des paramètres démographiques sexe-spécifiques. Dans la partie rouge, le F_{ST} est plus grand sur les autosomes par rapport au chromosome X, dans quel cas N_f est nécessairement plus grand que N_m (c'est-à-dire $N_f / N >$ 0.5, avec N la taille efficace totale de la population). Dans la partie bleue, le F_{ST} est plus petit sur les autosomes par rapport au chromosome X. La ligne blanche pleine correspond à la situation où les deux valeurs de F_{ST} sont égales.

Ainsi, nous pouvons déduire de ce graphique que dans les populations patrilinéaires, où le F_{ST} des autosomes est significativement plus fort que celui du chromosome X, les femmes ont une taille efficace plus importante que celle des hommes.

Nous avons ensuite voulu comprendre plus précisément quelles combinaisons de paramètres démographiques sexe-spécifiques (N_f, N_m, m_f et m_m) étaient compatibles avec nos données. Ainsi, nous avons calculé, dans un modèle de migration en île, quels étaient les F_{ST} attendus par locus sur le chromosome X, à partir des F_{ST} par locus sur les autosomes, pour chaque combinaison de paramètres démographiques sexe-spécifiques. Nous avons comparé ces valeurs attendues avec les valeurs observées par locus sur le chromosome X. Si les distributions de valeurs étaient significativement différentes entre elles (test de Wilcoxon

significatif), alors les paramètres démographiques utilisés pouvaient être significativement rejetés (voir les zones bleues sur la figure 7, sous-figures A et B, ci-dessous). Par contre, si les distributions n'étaient pas significativement différentes entre elles (test de Wilcoxon non significatif), alors les paramètres démographiques ne pouvaient pas être rejetés (voir les zones rouges sur la figure 7, sous-figures A et B, ci-dessous).

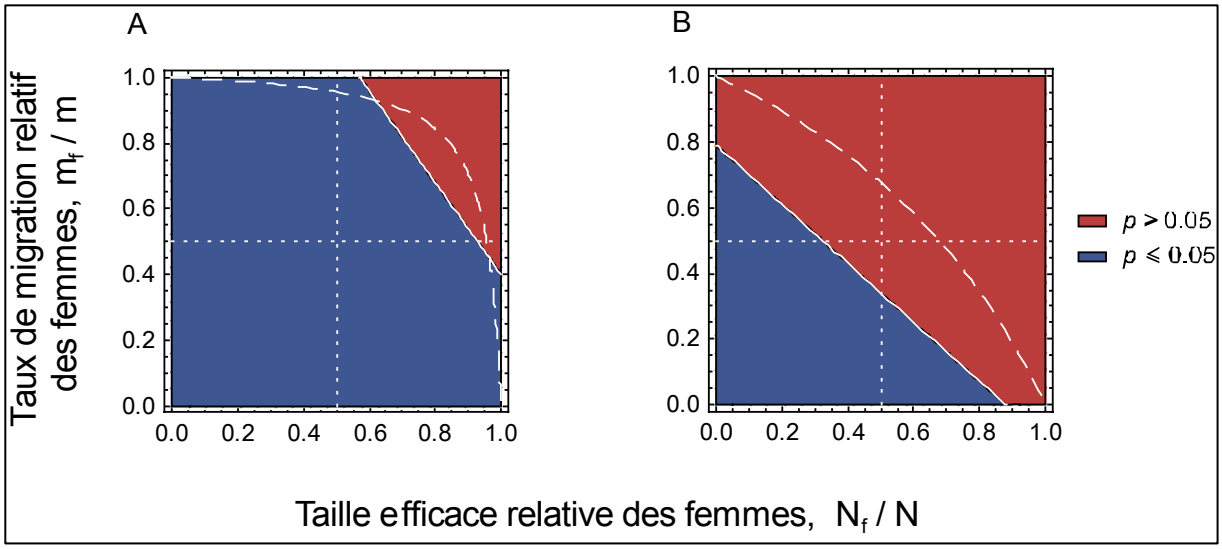


Figure 7 : Résultats des tests de Wilcoxon qui déterminent si chaque combinaison de paramètres démographiques est ou non significativement rejetée. A : Populations patrilinéaires ; B : Populations cognatiques ; les courbes blanches en pointillé correspondent aux combinaisons de paramètres possibles d'après les marqueurs du chromosome Y et de l'ADN mitochondrial seuls

Si l'on regarde la figure 7-A pour les populations patrilinéaires, nous voyons que la zone de paramètres non significativement rejetés (selon les tests de Wilcoxon, zone rouge) correspond exclusivement à des tailles efficaces plus fortes chez les femmes que chez les hommes ($N_f / N > 0.5$). D'après cette même figure, il parait également fortement probable que les taux de migration soient plus importants chez les femmes. Pour les populations cognatiques cependant (figure 7-B), toutes les combinaisons (plus forte migration et/ou taille efficace des femmes) sont envisageables.

Nous voyons également que les résultats basés sur les marqueurs uni-parentaux seuls (ligne pointillée sur les figures 7-A et 7-B) sont cohérents avec ceux basés sur les marqueurs multi-locus (une partie au moins de la ligne pointillée appartient à l'espace défini par la zone

$p > 0.05$). Cependant, nous pouvons également voir qu'avec ces marqueurs, nous ne pouvions pas trancher parmi différentes combinaisons possibles de paramètres démographiques pour les populations patrilinéaires. Ainsi, l'analyse des marqueurs multi-locus nous a permis un gain considérable d'information.

Grâce à nos analyses, nous voyons donc que les tailles efficaces, et très certainement les taux de migration, sont plus importants pour les femmes que pour les hommes, dans les populations patrilinéaires mais non dans les populations cognatiques. Nous avons donc pu mettre en évidence des différences de structure génétique entre les deux groupes de populations ayant des organisations sociales contrastées. Les différences de migration sexe-spécifique entre les populations patrilinéaires et cognatiques étaient attendues, du fait que les premières sont patrilocales exogames, et que les dernières sont plutôt patrilocales endogames. Cependant, les différences de taille efficace sexe-spécifique entre ces populations constituent un résultat original, puisque ce facteur a souvent été négligé dans les précédentes études.

Certains facteurs confondants et/ou limites méthodologiques peuvent être identifiés dans notre étude et sont discutés dans l'annexe 4 (différences de taux de mutation ou de pressions de sélection entre marqueurs, utilisation du modèle en île), mais il ne semble pas que nos conclusions puissent être remises en cause. Les différences de taille efficace entre hommes et femmes observées dans les populations patrilinéaires peuvent être expliquées par plusieurs hypothèses non exclusives (polygynie, plus forte mortalité chez les hommes, différences de temps de génération entre sexes, etc.), mais nous ne discuterons ici que d'un facteur, celui qui semble le plus probable : l'organisation sociale patrilinéaire.

Ce type d'organisation sociale s'accompagne en effet souvent d'une dynamique interne de fusion-fission, particulièrement bien décrite par Neel chez les Yanomama (Neel & Ward, 1970) : quand un village devient trop important, il se scinde en deux villages fils (fission) et quand au contraire un village devient trop petit, il fusionne avec un autre village (fusion). Mais cette fusion ne se fait pas aléatoirement, les villages réunis regroupent des personnes plus apparentées qu'attendus du fait du hasard, apparentement qui suit l'affiliation paternelle. Ainsi, Chaix *et al* (2004, 2007) ont montré qu'en Asie Centrale, dans les populations patrilinéaires, le nombre d'individus portant le même haplotype du chromosome Y est plus important que dans les populations cognatiques, conséquence directe de la dynamique interne de ces populations. De plus, dans les populations patrilinéaires, la transmission culturelle du

succès reproducteur est probablement transmise majoritairement par voie paternelle, comme proposé pour les descendants de Gengis Khan (Zerjal *et al.*, 2003) ou pour les populations d'Amérique du Sud (Neel & Ward, 1970). Ce phénomène de transmission culturelle de la fertilité a également été associé à une forte réduction de la taille efficace (Heyer *et al.*, 2005). Il apparaît finalement très probable que, dans les populations patrilinéaires, le regroupement des hommes apparentés entre eux, ainsi que la transmission paternelle du succès reproducteur, puissent réduire drastiquement la taille efficace des hommes par rapport à celle des femmes.

V. Conclusions et perspectives

Nous avons finalement vu que les proximités entre populations d'Asie Centrale sont liées aux affiliations linguistiques et ethniques, et non pas à leur répartition géographique. Deux groupes de populations assez bien différenciés co-existent : les populations de langues indo-iraniennes et les populations de langues turques. Les indo-iraniens semblent être les descendants des premiers peuples iraniens installés dans la région au moins depuis le Néolithique, et sont caractérisés par une proximité marquée avec les populations de l'ouest de l'Eurasie. Ces populations sont assez divergentes entre elles malgré leur proximité géographique. A l'inverse, les turco-mongols présentent de fortes proximités avec les populations d'Asie de l'est, zone probable de leur origine avant leur migration récente au 4ème siècle. Malgré une répartition assez large sur le territoire, ces populations sont génétiquement extrêmement homogènes.

Ces deux groupes de population ont également adopté des organisations sociales sensiblement différentes. Notamment, les Turco-mongols sont des populations patrilocales patrilinéaires, avec une tendance à l'exogamie. Les Indo-iraniens sont à l'inverse des populations patrilocales cognatiques, avec une certaine tendance à l'endogamie. L'étude conjointe de marqueurs sur les autosomes et le chromosome X nous a permis de mieux comprendre les différences démographiques entre sexes engendrées par ces organisations sociales. Nous avons ainsi pu déduire de notre étude que la patrilocalité, bien qu'elle crée des taux de migration différents entre hommes et femmes, n'est pas le seul facteur impliqué dans la structure génétique sexe-spécifique. Nous avons en effet pu montrer que les tailles efficaces des hommes sont fortement réduites dans les populations patrilinéaires, certainement à travers l'effet de l'organisation sociale sur la ressemblance entre individus au sein des populations.

Les données génétiques nous ont donc finalement permis d'avancer sur la compréhension de l'histoire des populations d'Asie Centrale, puisque nous avons montré qu'il est important de considérer l'effet de la taille efficace pour comprendre les différences sexe-spécifiques. Cependant, de nombreuses questions se posent encore. De combien la taille efficace des hommes est réduite par rapport à celle des femmes dans les populations patrilinéaires ? Cette réduction est-elle variable au sein des populations patrilinéaires, ou est-elle relativement constante ? Nous ne savons pas non plus quels sont les écarts de taux de migration entre hommes et femmes dans ces deux groupes de populations patrilocaux. Ces

questions sont plutôt d'ordre quantitatif et nécessitent donc une estimation précise des paramètres démographiques pour chaque population.

Une de nos perspectives est donc maintenant d'utiliser une approche ABC (Approximate Bayesian Computation, Beaumont *et al.*, 2002), pour estimer ces paramètres démographiques sexe-spécifiques. Cette approche consiste à simuler un grand nombre de jeux de données sous un modèle démographique donné, en tirant les valeurs des paramètres de ce modèle dans des distributions *a priori*. Les données génétiques simulées sont résumées par un certain nombre de « statistiques résumées », également calculées sur les données observées. Les distributions *a posteriori* des paramètres d'intérêt du modèle sont obtenues à partir du sous-ensemble des valeurs des paramètres initiaux permettant de retrouver des statistiques résumées proches de celles observées. La manière de calculer cette proximité et celle d'estimer les paramètres à partir de ces simulations criblées fait actuellement l'objet de débats et d'améliorations méthodologiques (Blum & François, 2009). Nous nous intéresserons pour nos simulations à un modèle spatialisé, avec de la migration (sexe spécifique) limitée dans l'espace. Ce modèle d'isolement par la distance (Malécot, 1973, Rousset, 1997) est plus réaliste et plus robuste que le modèle de migration en île que nous avons considéré précédemment, et qui est englobé dans le modèle d'isolement par la distance (dans la limite où la forme de la courbe de dispersion tend vers une distribution uniforme).

D'autres questions se posent également sur la mesure dans laquelle ces différences de taille efficace sont retrouvées dans d'autres populations humaines. D'après une analyse à échelle globale des autosomes et du chromosome X sur les populations du HGDP-CEPH (Ramachandran *et al.*, 2004), aucune différence de migration ou de taille efficace n'est nécessaire pour expliquer les données génétiques à cette échelle. Cependant, le fait de grouper des populations aux organisations sociales différentes peut parfaitement masquer des patrons existants à une échelle locale. Ainsi, même si la tendance pousse souvent à la généralisation, dans le cas de l'étude de l'interaction entre génétique et organisation sociale, il est important de se placer à une échelle locale, pour bien définir les populations étudiées. En analysant les populations du HGDP-CEPH à une échelle plus locale, nous avons en effet trouvé que pour un sous-groupe de 5 populations du Pakistan présumées patrilinéaires, la structure génétique est également plus importante sur les autosomes que sur le chromosome X, bien que non significativement. Ainsi, nous pensons que les études comparatives entre populations aux organisations sociales contrastées, notamment entre populations patrilinéaires et

matrilinéaires, seront dans l'avenir du plus grand intérêt, notamment si elles sont basées sur des analyses multi-locus de plusieurs systèmes génétiques.

CHAPITRE II . HISTOIRE ADAPTATIVE : ALIMENTATION ET MODE DE VIE EN ASIE CENTRALE

I. Introduction

Il s'agit dans cette partie de se demander si certaines caractéristiques génétiques, liées au régime alimentaire, ont été différemment sélectionnées dans les populations pastorales et agricultrices actuelles d'Asie Centrale, et depuis quand. Nous allons nous intéresser uniquement aux adaptations génétiques directement liées aux changements alimentaires, bien que d'autres fonctions biologiques ont également pu être affectées par ces changements, comme les mécanismes de défense face aux pathogènes. L'étude des gènes liés à l'alimentation présente de nombreux intérêts, car l'alimentation est une source importante et permanente d'interaction avec l'environnement, et donc une cible préférentielle de la sélection. De plus, l'alimentation est à l'interface entre culture, biologie et environnement, et constitue donc un bon modèle pour tester si le mode de vie a engendré de fortes différences d'adaptations entre populations. Dans un premier temps, nous présenterons les grandes lignes de l'évolution des modes de subsistance, avec notamment un intérêt particulier pour comprendre comment le Néolithique est apparu en Asie Centrale. Ensuite, nous regarderons quelles différences de régime alimentaire sont attendues entre éleveurs et agriculteurs et quels gènes ont déjà été identifiés comme sous sélection adaptative dans la littérature, afin de choisir nos gènes candidats.

Emergence de la domestication

A la fin du Paléolithique, aux environs de 12 000 BCE, les sociétés humaines étaient spécialisées dans différents modes de prédation : la chasse, la cueillette et la pêche, ou une combinaison des trois. Puis arrive le Néolithique, transition culturelle majeure définie par la sédentarisation de certaines populations, avec l'apparition indépendante de plusieurs foyers de domestication d'animaux et de végétaux, entre 8 500 et 2 500 BCE. Les foyers primaires de domestication, c'est-à-dire les régions ayant indépendamment adopté la domestication, sont au minimum au nombre de cinq : le Croissant Fertile, la Chine, la Méso-Amérique, la région des Andes / de l'Amazonie et l'Est des Etats-Unis (comme représenté sur la figure 8). Mais quatre foyers plus tardifs peuvent également être inclus : au niveau du Sahel, de l'Afrique tropicale de l'Ouest, de l'Ethiopie et de la Nouvelle-Guinée (pour une revue, voir Diamond, 2002).

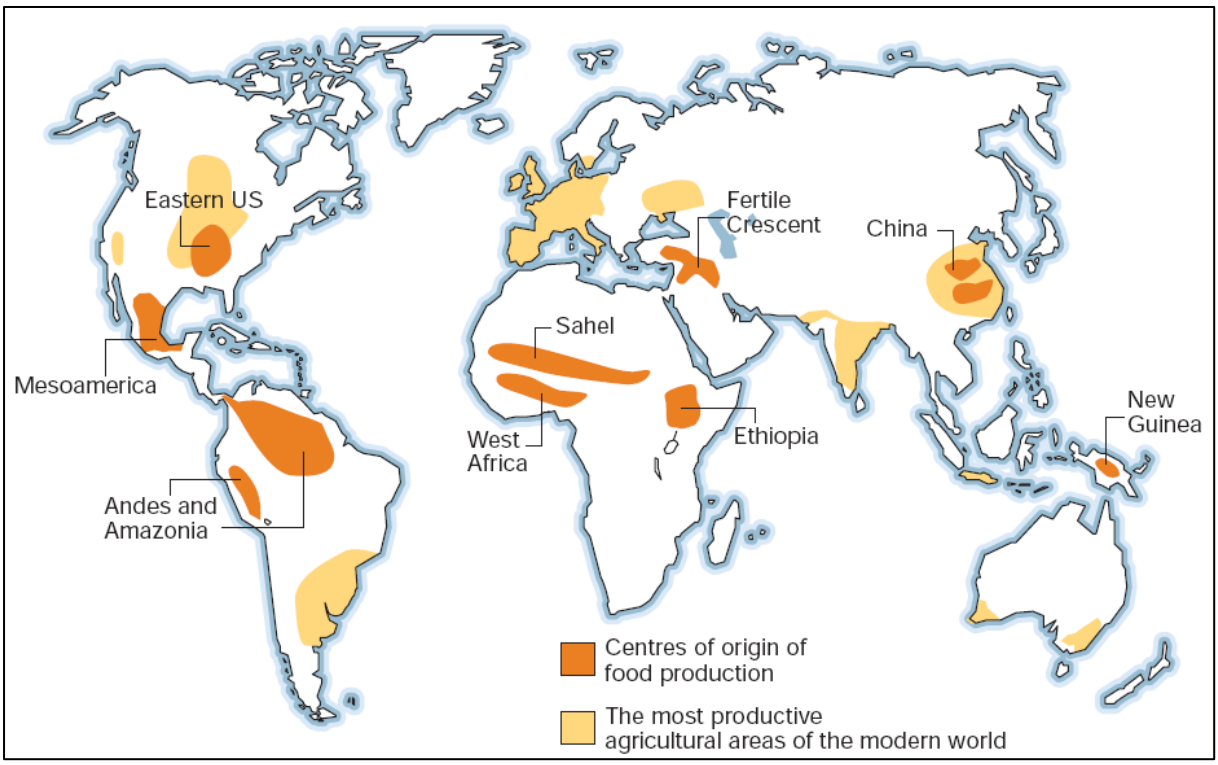


Figure 8 : Différents foyers de domestication (d'après Diamond 2002)

Vu la manière remarquablement convergente avec laquelle ces foyers, relativement bien répartis sur tout le globe, sont apparus à des intervalles de temps proches, se pose la question du contexte qui a engendré ces changements de mode de vie. Il semble que la sédentarisation, l'augmentation en nombre des populations humaines, et la diminution probable de la densité du gros gibier liée à des changements climatiques soient autant de facteurs qui ont conjointement entraîné la diminution des habitats exploitables pour chaque population, poussant ainsi à trouver de nouvelles sources de nourritures ou de nouveaux moyens d'en produire. La sédentarité a également certainement favorisé d'importants développements technologiques, grâce au relâchement des contraintes migratoires. Il est en tout cas certain que cette transition de mode de vie n'est pas due à un seul facteur, mais résulte plutôt de l'interaction de facteurs écologiques, démographiques, technologiques et culturels pour l'instant assez mal démêlés.

Diffusion des foyers de domestication

Ces foyers primaires agropastoraux ont ensuite diffusé progressivement aux alentours, soit par la colonisation d'individus vers de nouveaux territoires (diffusion démique) soit par

l'adoption de l'agriculture par des populations locales au contact de populations d'agriculteurs (diffusion culturelle) (Ammerman & Cavalli-Sforza, 1973, 1984, Cavalli-Sforza, 1996). Ce nouveau mode de vie agropastoral s'est en tout cas certainement propagé en lien avec le milieu rencontré : tandis que diverses formes de culture végétale ont pu être facilement pratiquées dans des écosystèmes fermés ou boisés, les écosystèmes ouverts et herbeux (déserts, steppes, zones montagneuses) ont plutôt favorisé le développement des élevages pastoraux. Enfin, dans certains milieux peu propices à l'agriculture ou l'élevage (forêts denses, aires polaires, milieux insulaires), le mode de vie basé sur la chasse et la cueillette a été maintenu (Mazoyer & Roudart, 1997).

Le Néolithique en Eurasie

La première trace attestée d'agriculture et de pastoralisme vient du « Croissant Fertile », c'est-à-dire dans les actuels pays du Moyen-Orient (voir figure 8). Il semble dans cette région que les conditions climatiques ont été responsables d'une baisse de ressources naturelles entre 9 500 et 8 600 BCE (Harris, 1996). Ont été trouvés dans cette région des restes de plantes domestiquées : orge, blé, pois, lentilles, lin, datant de 8 000 BCE (Harris, 1996). Vers 7 000 BCE, l'agriculture est apparue à d'autres sites au Levant, ainsi que plus à l'est dans les montagnes du Zagros (Iran), et au nord en Anatolie (Turquie). Plus d'incertitude entoure la première domestication animale, de chèvres et moutons, et peu de preuves, voire aucune, n'atteste de leur présence avant 7 000 BCE. L'élevage n'intègre de manière importante l'économie qu'à partir de 6 500 BCE, définissant ainsi un mode de production agropastoral. Plusieurs vagues de diffusion de l'agriculture partent vers l'Eurasie en provenance du Moyen-Orient : une vers le sud-est atteignant le Pakistan vers 7 000 BCE, une vers le nord-est atteignant le Turkménistan vers 6 000 BCE, et une troisième vers le nord atteignant l'Europe entre 5 000 et 3 000 BCE (Harris, 1996). Notons qu'en parallèle, en Asie de l'Est, un deuxième foyer de domestication voit le jour vers 6 500 BCE, associant millet, riz, cochon et poulet, soit séparément en Chine du Nord et du Sud, soit en tant qu'un seul et même évènement.

Premières traces d'agropastoralisme en Asie Centrale

L'agriculture basée sur les céréales (blé et orge), ainsi que le pastoralisme basé sur les chèvres et les moutons part donc du Moyen-Orient vers le nord-est et s'établit à l'ouest de l'Asie Centrale, à la frontière entre l'Iran et le Turkménistan, dès 6 000 BCE (Harris, 1996).

Bien que la présence d'une économie locale antérieure soit possible, la culture trouvée en Asie Centrale, dite « Jeitun », montre de nombreuses similitudes avec celle venant du Levant et d'Anatolie et laisse donc penser à une diffusion des espèces domestiquées de l'ouest vers l'est. La question de la diffusion démique ou culturelle n'est ici pas résolue, mais l'arrivée rapide d'une économie Néolithique développée laisse penser que cette diffusion correspond à une expansion de populations. Une autre culture agropastorale, plus à l'est (au nord de l'Afghanistan et au sud-est du Tadjikistan), connue sous le nom de la culture Hissar, est également retrouvée dès 6 000 BCE, mais les restes ne ressemblent pas à la culture Jeitun dans leur culture matérielle. Cette culture est moins bien documentée et ces sites ont pu être occupés par des pasteurs ou des agropasteurs. Ils pourraient correspondre à une zone de domestication locale (Harris, 1996). En conclusion, il apparaît que l'agriculture peut avoir dans la région une origine exogène, comme le suggère la similitude de la culture Jeitun avec l'ouest, mais aussi endogène, comme le montrent l'apparition de la culture Hissar.

Emergence du pastoralisme spécialisé en Eurasie

Le pastoralisme spécialisé est-il un produit dérivé de l'agropastoralisme ou est-il apparu directement à partir d'un mode vie basé sur la chasse ? La réponse à cette question est essentielle pour savoir dans quelle mesure les pasteurs et les agriculteurs ont un passé alimentaire commun, et donc à quel point ils ont subi des contraintes différentes. Nous pouvons dégager trois hypothèses, non exclusives, qui décrivent différents scénarios pour l'émergence d'un pastoralisme spécialisé en Eurasie.

1) Tout d'abord, Renfrew (1996) propose que le pastoralisme spécialisé émerge à partir de l'agropastoralisme de la culture Jeitun, autour du Turkménistan, donc nécessairement après 6 000 BCE. Cette première hypothèse se base sur des données linguistiques et est totalement spéculative.

2) Le pastoralisme peut également être apparu dans la région de la Volga (en Ukraine), où les premières évidences d'utilisation des chevaux pour la monte dans des populations agropastorales sont datées autour de 5 000 BCE (culture kourgane, Anthony & Brown, 1991). La culture kourgane est caractérisée par des tumulus, c'est-à-dire des monticules de terre et de pierre recouvrant une ou plusieurs tombes, qui ne sont autres que des chambres funéraires réservées aux élites de la société nomade, comprenant souvent des chevaux et des

chariots. Cette culture est donc également bien connue pour être responsable de l'invention des véhicules à roue entre 3 000 et 2 200 BCE (Anthony & Vinogradov, 1995). La culture kourgane a ensuite diffusé vers le nord et l'ouest de l'Europe, entre 4 300 et 2 800 BCE (Gimbutas, 1991), et a fini par s'étendre des montagnes de l'Altaï jusqu'en Bulgarie. Selon cette deuxième hypothèse, le pastoralisme spécialisé serait donc également apparu à partir d'agropastoralisme plus ancien, mais cette fois-ci dans la région de la Volga.

3) Une troisième forme de transition vers un pastoralisme spécialisé est également envisageable. En effet, au Kazakhstan, ont été trouvés dès 3 500 BCE des restes de chevaux domestiques, ainsi que des preuves d'utilisation de brides pour les chevaux (Outram *et al.*, 2009). Or dans la région, les cultures énéolithiques (entre le néolithique et l'Age de Bronze), comme la culture Botai des steppes centrales (3 700-3 100 BC), laissent peu de trace d'une économie basée sur l'agriculture (Kislenko & Tatarintseva, 1999). Les sociétés d'Asie intérieure comme les Atbasar ou les Afanas'ev (~3 000 BC) au Nord Est des steppes étaient principalement des chasseurs-pécheurs, avec un élevage bovin limité (Khlobystina, 1973, Shilov, 1975, Vadetskaya, 1986). De même, dans la région de l'Altaï Mongol, des excavations récentes ont permis de démontrer que, tandis que les représentations artistiques humaines avant l'Age de Bronze font référence à des scènes de chasse (Tseveendorj *et al.*, 2005), la culture semi-nomade Pazyryk, rattachée à la culture Kourgan évoquée plus tôt, est bien représentée pendant l'Age de Bronze (Jacobson-Tepfer, 2008). Les représentations artistiques montrent en outre qu'à cette période, l'utilisation des chevaux est liée à la chasse et non à l'élevage ou le transport, tandis que les yaks remplissent ce dernier rôle (Jacobson, 2001). Etant donné qu'il n'y a pas eu d'agriculture dans cette région à cette période, il semble bien que les populations de chasseurs-cueilleurs et pêcheurs se sont directement spécialisés dans le pastoralisme, bien que ce mode de vie a surement été adopté par diffusion à partir d'autres populations déjà pastorales à l'ouest de la Sibérie (Vadetskaya, 1986, Anthony, 1998, Alekseev *et al.*, 2001). Ce troisième modèle diverge donc fortement des deux précédents modèles proposés puisqu'il témoigne d'une transition directe de la chasse / pêche / cueillette vers le pastoralisme spécialisé dès 3 500 BCE, avec une possible adoption par diffusion culturelle ou démique à partir d'autres cultures pastorales comme la culture Kourgan, qui s'est étendue jusqu'à l'Altaï.

Evolution du pastoralisme en Asie Centrale

Bien que les premières traces de domestication semblent dater de 3 500 BCE au Kazakhstan, la faune de la région montre que le pastoralisme basé sur l'exploitation de chevaux, de vaches et de moutons, est devenu prédominant seulement autour de 2 000 BCE (Tsalkin, 1964).

Dans la région de l'Altaï Russe (à la frontière du Kazakhstan, de la Chine et de la Mongolie), des restes funéraires témoignent de la continuité du pastoralisme entre l'Age de bronze récent et moyen (entre 1 500 et 1 000 BCE, Khazanov, 1994). Ces résultats incluent l'apparition régulière d'os de moutons, de vaches et de chevaux, ainsi que l'apparition de récipients de céramique dont la forme et la substance indiquent le stockage de viande et de produits laitiers. Autour d'environ 1 000 BCE, la sépulture Arzhan I (restes de 300 chevaux et de brides en bronze), ainsi que d'autres célèbres sépultures associées à la culture Pazyryk (600-300 BCE), attestent également des transformations du mode de vie (Gryaznov, 1980). D'autres objets suggèrent également l'importance de l'élevage des ovins dans cette économie, ainsi qu'une dépendance continue à la chasse.

A la frontière entre le Kazakhstan et le Kirghizistan, des excavations datant de l'Age du Bronze, entre 2 500 et 1 000 BCE, indiquent que les populations avaient un mode de vie pastoral de type transhumant vertical, basé majoritairement sur l'élevage de moutons et de vaches (Frachetti, 2008), donc sur différentes espèces que dans les cultures précédemment décrites. Ces résultats sont basés sur le pourcentage écrasant de restes de faune domestique comparée aux restes d'animaux sauvages. Il n'y a pas dans cette zone de trace de stratégie alternative de subsistance, comme la culture de plantes. Les sols de cette vallée sont en effet pauvres et peu propice à la culture de végétaux. Cependant, plus tardivement, entre 1 000 BCE et 500 CE, les restes de 3 sites datant de l'Age de Fer et occupés par des populations nomades nous livrent d'autres résultats. Les analyses archéo-zoologiques et archéo-botaniques indiquent en effet que ces populations pratiquaient à la fois l'élevage (vache, mouton, chèvre, chevaux et peut-être chameaux) et l'agriculture (millet, blé, orge et peut-être riz), couplés à une transhumance verticale de courte distance (Chang *et al.*, 2003). A un des trois sites, la culture du riz augmente entre 415 BCE et 75 CE tandis qu'à un autre site, celle du millet s'intensifie entre 775 et 40 BCE. Enfin, au dernier site, très peu de céréales ont été retrouvées, indiquant que l'élevage a ici été prédominant sur l'agriculture.

Ce patron « multi ressources » est également retrouvé actuellement chez les Yomut Turkmènes ou encore chez les Baloutches. Il y est interprété comme une stratégie hautement adaptative, de par une meilleure flexibilité face aux contraintes extérieures (Iron, 1974). Les Yomut Turkmènes sont en effet composés à la fois d'agriculteurs nomades et d'éleveurs nomades, mais qui adhèrent à une organisation sociale commune et dont les deux modes de production peuvent être flexibles. Ainsi, nous pouvons imaginer que certaines fluctuations des conditions du milieu ont favorisé un changement de ressources entre la culture de céréales et l'élevage chez les habitants du Sud Est du Kazakhstan à l'Age de Fer.

En résumé, il apparaît que le pastoralisme spécialisé a pu apparaitre soit directement à partir de stratégies mixtes agropastorales existantes (comme à partir de la culture Jeitun aux environs du Turkménistan ou dans la région de la Volga, avec la culture Kourgan vers 4 000 BCE), soit directement à partir de la chasse et de la pêche (comme dans les steppes Nord Asiatiques avec la culture Botai, vers 3 500 BCE). Finalement, la possibilité de variations temporelles de stratégies de subsistance suggère une flexibilité importante qui a pu créer des stratégies 'multi-ressources'. Ainsi plusieurs types de transition vers le pastoralisme ont existé, et même une fois adoptés, ces modes de vie ont assurément changé de manière complexe au cours du temps.

Conséquences de la révolution Néolithique sur les régimes alimentaires

La domestication a profondément changé le régime alimentaire des populations humaines. En effet, chez les agriculteurs, la diminution observée du taux de croissance des individus, la réduction de la stature adulte et de la taille des dents sont autant de marqueurs du statut nutritionnel qui témoigne d'importances carences alimentaires dans ces populations, qui ont constitué de fortes pressions de sélection (Larsen, 1995).

Au Paléolithique, l'alimentation était majoritairement basée sur un régime riche en protéines (viande), en fibres (fruits et légumes) et pauvre en glucides (Cordain *et al.*, 2005). Les contraintes alimentaires étaient surtout les quantités limitées de nourriture obtenues par prédation, ainsi que les variations saisonnières ou annuelles entre périodes d'abondance et de famine. Chez les nouveaux éleveurs du Néolithique, le régime alimentaire n'a pas été radicalement transformé, puisqu'ils ont continué à baser une importante part de leur régime

alimentaire sur les protéines animales, et certainement à être limités par les quantités de nourriture, avec toujours peu de glucides. La domestication animale a tout de même amené à la consommation d'un nouvel aliment : le lait. Par contre, une réelle transition a eu lieu plus tard pour ces populations : la révolution industrielle, au 19ème siècle, qui a amené de nouveaux types d'aliments, notamment denses en calories, et surtout en abondance.

Au contraire, pour les premiers agriculteurs du Néolithique, la transition vers leur nouveau mode de vie n'a pas été sans contraintes. En effet, leur alimentation s'est beaucoup spécialisée, entraînant des carences liées à un manque de diversité alimentaire. Leur nouvelle dépendance forte sur les produits de l'agriculture, notamment sur les céréales, a eu pour conséquence une certaine période de maladaptation, et donc de changements physiologiques importants. La forte proportion de glucide dans leur alimentation (à travers la consommation de céréales), au détriment des protéines, a complètement modifié le type de régime alimentaire et les contraintes associées. L'arrivée de nouveaux aliments comme les céréales a également créé de véritables défis physiologiques, notamment pour réussir à digérer ces sucres complexes et à en tirer de l'énergie. Ainsi, bien que la révolution industrielle ait également par la suite fortement affecté ces populations, à travers la richesse calorique des aliments et leur abondance, cette transition récente a peut-être été moins brutale que pour les éleveurs. Notons également qu'il n'est pas rare que les populations pastorales, malgré leur mode de vie nomade qui les empêche de pouvoir directement produire leurs propres ressources agricoles, procèdent à des échanges de nourriture avec les agriculteurs.

Objectifs de cette étude : adaptations génétiques liées à l'alimentation

Les gènes liés au métabolisme ont donc pu être l'objet de conflits évolutifs majeurs, liés au fait qu'ils ont été sélectionnés sur une période de temps suffisamment longue pour être adaptés à un régime alimentaire donné (le mode de vie basé sur la chasse et la cueillette constitue 95% du passé de l'Homme moderne), et qu'ils se sont retrouvés par la suite dans des régimes alimentaires totalement nouveaux. La conséquence de ces changements culturels rapides est, entre autres, l'émergence d'un ensemble de maladies nouvelles, appelées « maladies des civilisations », qui inclue notamment le diabète de type II, l'obésité, les maladies cardio-vasculaires et l'hypertension. . La détection de pressions de sélection sur des gènes liés au métabolisme nous permet de comprendre quels gènes ont été importants pour

notre survie dans le passé, et par conséquent lesquels sont probablement liés à des maladies actuelles, et peut donc, de fait, avoir des retombées intéressantes pour les sciences médicales. Il s'agit donc ici d'explorer les adaptations locales qui peuvent avoir différencié les éleveurs et les agriculteurs dans le passé, pour mieux comprendre la variabilité présente en termes de prévalence de maladies.

Approches méthodologiques

Deux grandes catégories de méthodes permettent de détecter des traces de sélection dans le génome humain : l'approche de « scans génomiques » et l'approche de « gènes candidats ». Les scans génomiques ont pour principe de chercher des traces de sélection dans le génome, sans a priori sur les endroits qui vont préférentiellement être soumis à de la sélection (Storz, 2005). Il s'agit donc de trouver des marqueurs génétiques où les niveaux de diversité contrastent avec le reste du génome, notamment des zones de très forte différenciation génétique entre populations, si l'on cherche à détecter de l'adaptation locale. Ces approches ont pour avantage de s'affranchir de toute connaissance préalable sur les fonctions biologiques des marqueurs génétiques, mais elles ne permettent toutefois pas de comprendre précisément les mécanismes évolutifs qui sont à l'œuvre dans les régions identifiées comme étant sous sélection, puisque ces régions ne sont pas forcément caractérisées d'un point de vue fonctionnel. De plus, cette approche peut engendrer un important taux de faux positifs (Akey, 2009).

L'approche gènes candidats est plus directe, puisqu'elle teste explicitement certains gènes considérés comme de bonnes cibles potentielles de la sélection naturelle, et, en ayant des a priori sur les résultats attendus sous différentes hypothèses, elle permet de faire des inférences sur les processus évolutifs qui ont abouti à la diversité génétique observée. Elle est cependant également plus limitée, puisqu'elle ne peut choisir que des cibles préalablement identifiées par des études fonctionnelles. Cependant, les données fonctionnelles sont nombreuses chez l'Homme, et les candidats ne manquent pas. Notre but est également, si l'on confirme un candidat, de pouvoir dater l'événement qui a différencié les populations et donc d'obtenir des informations sur la transition Néolithique en Asie Centrale.

Choix des gènes candidats

Nous pouvons identifier trois composants alimentaires qui différencient fortement les agriculteurs des éleveurs : le lait, la viande et les céréales. Nous allons donc séparer cette étude en trois parties, reflétant ces différences de régime alimentaire.

Tout d'abord, nous allons étudier le cas d'école de la lactase, l'enzyme responsable de la digestion du lactose, dont les fréquences alléliques semblent être corrélées au mode de vie (Holden & Mace, 1997), et qui a été identifié comme étant fortement sous l'action de la sélection naturelle (International Hap Map Consortium, 2005, Bersaglieri *et al.*, 2004, Tishkoff *et al.*, 2007). Nous allons ensuite tester une mutation sur un gène lié à la digestion de la viande (*AGXT*), mutation également identifiée comme potentiellement sélectionnée en lien avec le mode de vie (Caldwell *et al.*, 2004). Finalement, nous allons étudier plusieurs gènes associés au diabète de type II (excès de sucre dans le sang), donc liés à l'assimilation de glucides. Cette maladie ne présente en effet pas les mêmes prévalences dans toutes les populations et certaines hypothèses évolutives ont proposé un rôle de ces gènes dans l'adaptation aux conditions alimentaires passées, en lien avec le mode de vie (Neel, 1962, Neel, 1992, Brand Miller & Colagiuri, 1994, Colagiuri & Brand Miller, 2002).

Le gène de l'amylase (digestion de l'amidon, stock des glucides dans les végétaux) a également été identifié comme lié au mode de vie (Perry *et al.*, 2007). Nous avons obtenu pour ce gène les premières données en Asie Centrale, mais l'interprétation de ces résultats nécessite l'obtention de données supplémentaires venant d'autres populations. Ainsi, les premiers résultats sur ce gène ne seront pas présentés dans cette thèse.

II. Adaptation à la consommation de lait ?

A. Problématique

La lactase, enzyme codée chez l'Homme par le gène *LCT* sur le chromosome 2 et produite par l'intestin grêle, a pour fonction de digérer le lactose, composant principal du lait, c'est-à-dire de le transformer en glucose et galactose qui sont assimilables par notre organisme. Chez les mammifères, l'activité de la lactase est maximale en début de vie, puis sa production diminue à la fin du sevrage pour enfin disparaître, à environ 5-7 ans chez l'Homme. Ainsi, en général, la consommation de lactose à l'âge adulte non seulement n'apporte aucun bénéfice nutritif, mais en plus peut provoquer des troubles intestinaux. Cependant, il a été montré que chez l'Homme, certains individus présentent une persistance da la lactase, c'est-à-dire que leur lactase ne décroît pas en activité et qu'ils peuvent digérer le lactose toute leur vie en bénéficiant de ses apports nutritifs (Durham, 1991). Ces individus représenteraient environ 45% de la population mondiale (Ingram *et al.*, 2009).

La fréquence de ces individus n'est cependant pas la même dans toutes les populations humaines : elle est très forte en Europe du Nord (90-100% en Suède, Ross *et al.*, 2006) et en Irlande (96%, voir Holden & Mace 1997), ponctuellement en Europe de l'Ouest et Centrale (87% en République Tchèque, voir Holden & Mace, , 1997), dans les populations pastorales autour du Sahara (~80-90% chez les Touaregs, les Bédouins, les Beja et les Peuls, voir Holden & Mace, , 1997), et plus au Sud-est de l'Afrique (90% chez les Tutsi, voir Holden & Mace, , 1997). Cette fréquence décroît ensuite progressivement autour de 30-60% dans de nombreuses populations Africaines et Européennes. En Asie, ces fréquences sont moins élevées, avec les plus fortes valeurs autour de 50%-60% dans certaines populations du Pakistan (Brâhuî, Baloutche, Makrani, Sindhi, Pathan), 44% chez les Punjabi (en Inde et au Pakistan), et 30% au Sri Lanka (Bersaglieri *et al.*, 2004, Holden & Mace, 1997). Dans le reste de l'Asie, les fréquences sont plutôt très faibles (0-8% en Chine, Wang *et al.*, 1984, Bersaglieri *et al.*, 2004, 2% en Thaïlande, Holden & Mace, 1997), tout comme en Amérique (0% chez les Apaches, et les Hopi, 5% chez les Pima Holden & Mace, 1997).

Signatures de la sélection sur le gène de la lactase

Les mutations responsables de ce changement d'expression de la lactase ont été bien identifiées : pour les populations européennes, il s'agit de la mutation dominante C/T -13.910 (rs4988235), dans le gène MCM6 en amont du gène de la lactase (Enattah *et al.*, 2002); pour les populations africaines, il s'agit de la mutation G/T -14.010, située à 100 paires de bases de la mutation européenne (Tishkoff *et al.*, 2007).

En utilisant la diversité génétique de marqueurs autour de ces mutations, il a été estimé que ces mutations sont fortement sous sélection (voir tableau 2 ci-dessous). Pour la population du CEPH (résidents de l'Utah originaires de l'Europe du nord et de l'ouest, dont l'abréviation est CEU), les coefficients de sélection sont compris entre 0.01 et 0.15 et l'apparition de la mutation date d'il y a 2 200 à 20 700 ans[5] (Bersaglieri *et al.*, 2004). Pour la population scandinave, l'expansion de la mutation date d'il y a 1 600 à 3 200 ans, avec des coefficients de sélection compris entre 0.09 et 0.19 (Bersaglieri *et al.*, 2004), et pour la population portugaise et italienne, elle est datée d'il y a 5 900 à 28 400 et 3 400 à 42 200 ans, respectivement (Coelho *et al.*, 2005).

En Afrique, les coefficients de sélection ont été estimés pour 6 groupes de populations (Tishkoff *et al.*, 2007). Ils sont majoritairement compris entre 0.02 et 0.14, avec des dates variables entre populations, comprises entre 1 200 et 6 000 ans pour les plus jeunes, et entre 2 200 et 18 500 ans pour les plus anciennes (voir tableau 2 ci-dessous). Ces signatures moléculaires de sélection sont parmi les plus fortes trouvées sur tout le génome humain (International Hap Map Consortium, 2005).

[5] Il faut noter que les estimations de début d'expansion des mutations se réfèrent au présent. Donc par exemple, une expansion datée entre 2 200 et 20 700 ans correspond à entre 18 700 et 200 BCE

Population	**Coefficients de sélection***	**Age d'expansion** (intervalles)*	**Référence**
CEU	(0,014-0,150)	(2 188-20 650)	Bersaglieri *et al*, (2004)
CEU	**0.039** (0,012-0,107)	**9 323** (2 232-19 228)	Tishkoff *et al*, (2007)[1]
Scandinavie	(0,090-0,190)	(1 625- 3 188)	Bersaglieri *et al*, (2004)
Portugal	-	**9 370** (5 940-17 190)	Coelho *et al*, (2005)[2]
Italie	-	**8 315** (3 440-27 690)	Coelho *et al*, (2005)[2]
Kenya-Afro-Asiatic	**0.070** (0,022-0,142)	**2 966** (1 215- 6 827)	Tishkoff *et al*, (2007)[1]
Kenya-Nilo-Saharan	**0.035** (0,008-0,080)	**6 925** (2 232-18 496)	Tishkoff *et al*, (2007)[1]
Tanzania-Afro-Asiatic	**0.053** (0,018-0,130)	**5 956** (1 575-13 054)	Tishkoff *et al*, (2007)[1]
Tanzania-Nilo-Saharan	**0.070** (0,023-0,143)	**3 757** (1 344- 9 087)	Tishkoff *et al*, (2007)[1]
Tanzania-Niger-Kordofanian	**0.077** (0,026-0,142)	**2 778** (1 219- 6 049)	Tishkoff *et al*, (2007)[1]
Tanzania-Sandawe	**0.043** (0,005-0,132)	**5 717** (1 296-17 971)	Tishkoff *et al*, (2007)[1]

* Pour Tishkoff *et al* (2007) et Coelho *et al* (2005), les intervalles de confiance à 95% sont donnés entre parenthèses. Coelho *et al* (2005) considèrent une croissance rapide *a priori* et n'estiment donc pas de coefficients de sélection. Pour Bersaglieri *et al* (2004), les données entre parenthèses correspondent aux estimations pour des tailles efficaces de population entre 500 et 5 000 ;

[1] Selon l'estimation par modèle dominant ; [2] Selon les estimations avec recombinaison et avec des mesures directes de taux de mutation des microsatellites

Tableau 2: Différentes estimations de coefficients de sélection et de dates d'expansions sur les mutations européennes et africaines associées à la persistance de la lactase

L'hypothèse « historico-culturelle »

Les premières évidences d'utilisation de lait (traces résiduelles de lait sur les céramiques) datent de 6 500 BCE au Nord de l'actuelle Turquie, puis autour de 6 000-5 000 BCE en Europe de l'Est et vers 4 000 BCE en Bretagne (Evershed *et al.*, 2008). Ces traces anciennes montrent que le lait n'était pas stocké sous forme crue mais sous forme transformée. En Afrique, nous n'avons pas de telles preuves, mais nous savons que la vache a été domestiquée en Egypte entre 9 000 et 7 700 BCE (Tishkoff *et al.*, 2007), et que les premières indications d'un mode de vie pastoral plus au Sud du Sahara datent de 4 500-3 300 BCE (Gifford-Gonzales, 2005, Ambrose, 1998). Il a donc été proposé que les mutations permettant de digérer le lait à l'âge adulte ont été sélectionnées positivement au Néolithique, quand le lait d'autres espèces animales est devenu une source importante de l'alimentation : c'est l'hypothèse « historico-culturelle » (Simoons, 1969, Simoons, 1970, McCracken, 1971).

Il s'agirait d'un cas exemplaire de co-évolution bio-culturelle : un changement culturel (la consommation de lait frais à l'âge adulte) aurait engendré une forte pression de sélection du phénotype PL (persistance de la lactase), aboutissant ainsi à une modification biologique. Sous cette hypothèse, les porteurs de la mutation seraient avantagés de par leur capacité à tirer un bénéfice nutritionnel de la consommation du lait. Les non-porteurs de cette mutation pourraient également être contre-sélectionnés à cause des troubles digestifs associés à la mauvaise digestion du lait dont les diarrhées, connues comme source importante de mortalité chez les enfants, mais cette hypothèse est plus controversée (Scrimshaw & Murray, 1988).

Au final, les populations utilisant le plus de lait frais d'animaux domestiques, notamment les populations pastorales, devraient présenter les plus fortes fréquences de ces mutations. Cependant, la majorité des populations d'agriculteurs sont en fait agropastorales et présentent une dépendance entre 10% et 40% sur le pastoralisme (Murdock, 1967 , Murdock & White, 2006). De plus, certaines populations agropastorales sont dépendantes d'animaux dont le lait n'est pas utilisé (comme les cochons), tandis que d'autres sont bien connues pour transformer leur lait par fermentation (comme les cultures méditerranéennes par exemple, par rapport à l'Europe du Nord qui consomme plutôt du lait frais). Il n'est donc pas forcément pertinent de différencier les populations de manière binaire agriculteurs / éleveurs, mais plutôt de les caractériser par leur pourcentage de dépendance au produits laitiers, comme l'ont fait

Holden et Mace (1997), d'après les données de Murdock et White (2006). De telles données ne sont cependant pas disponibles pour toutes les populations.

Des exceptions ?

Bien que l'attendu sous cette hypothèse évolutive soit assez clair, la situation, elle, ne l'est pas toujours, et de nombreuses exceptions sont observées, dans les deux sens. En Afrique notamment, certains groupes ethniques pastoraux ayant un régime alimentaire fortement basé sur les produits laitiers n'ont pas une fréquence élevée de lactase persistance. C'est par exemple le cas des Dinka et Nuer au Soudan (respectivement 22% et 25% de PL, Bayoumi *et al.*, 1982)) et des Somaliens en Ethiopie (24% de PL, Ingram, 2008). De la même manière, les Samis, populations pastorales d'éleveurs de renne en Europe du Nord, ont par exemple entre 40 et 75% de PL, selon les différents groupes (Kozlov & Lisitsyn, 1997), ce qui est bien moins élevé que dans la population générale Suédoise (91% de PL, Ross *et al.*, 2006). Ainsi, les fréquences du phénotype PL ne sont pas toujours bien corrélées au mode de vie pastoral.

Dans le cas des Samis cependant, nous savons que le lait de renne est assez pauvre en lactose (2.4%) et que les Samis ne l'ont sûrement utilisé que de manière limitée (Haglin, 1991), au cours des 1 000 dernières années (Kozlov & Lisitsyn, 1997). Mais la question est ici de savoir pourquoi nous observons une fréquence du phénotype PL si élevée chez les autres populations européennes qui sont seulement partiellement pastorales ? Nous avons vu que dans les populations scandinaves, en effet, les coefficients de sélection estimés (s = 0.09-0.19) suggèrent une très forte pression de sélection assez récente (il y a 1 600-3 200 ans). Ce n'est donc pas vraiment compatible avec l'hypothèse historico-culturelle. Pour tenter d'expliquer les nombreux cas de non-concordance entre mode de vie et fréquence de PL, des hypothèses alternatives ont vu le jour.

Hypothèses évolutives alternatives

Une autre hypothèse a été formulée par Flatz & Rotthauwe (1973). Selon cette hypothèse, dans les environnements situés à haute latitude, où il y a peu d'ensoleillement, la consommation de lait pourrait constituer un apport en calcium important, dans des populations souffrant souvent de déficiences en vitamine D et de rachitisme. Cette hypothèse est confortée par les fortes fréquences de PL trouvées en Europe du Nord. Cependant, dans ces environnements à haute latitude, il y a également un climat plus froid, ce qui pourrait

faciliter une consommation de lait frais plutôt que fermenté, et donc constituer un facteur confondant.

Finalement, une troisième hypothèse a été proposée, celle de l'absorption en eau et en électrolytes que permet la digestion du lait, engendrant ainsi une nouvelle source d'hydratation sans risque de contamination, dans des environnements secs et arides (Cook & al-Torki, 1975, Cook, 1978). Cette hypothèse se base donc plutôt sur la forte contre-sélection des individus n'ayant pas la mutation favorable, qui présentent des diarrhées importantes s'ils consomment du lait. Cette hypothèse se justifie par les importantes fréquences trouvées en Afrique dans les populations vivant autour du Sahara dans des zones désertiques. Cependant, le pastoralisme est souvent une adaptation aux milieux semi-arides où l'agriculture ne peut pas facilement être mise en place à cause des épisodes de sécheresse, et sa distribution recouvre donc au moins partiellement celle des zones arides.

Il est intéressant de noter que même si ces deux hypothèses peuvent aider à comprendre pourquoi certaines populations ne pratiquant pas de pastoralisme présentent des fréquences fortes de PL, elles n'expliquent en rien pourquoi certaines populations pastorales ont une fréquence faible de PL.

Holden & Mace (1997) ont justement testé la capacité de ces trois hypothèses à expliquer le patron global de PL, en étudiant les fréquences du phénotype PL dans 62 populations, pour lesquelles les niveaux de pastoralisme, d'ensoleillement et de sécheresse étaient disponibles. Ces auteurs ont testé la contribution respective de ces facteurs en tenant compte de la non-indépendance des populations du fait de leur histoire partiellement commune, facteur quantifié par la ressemblance génétique sur le reste du génome. Cette étude a montré que, malgré la non-concordance entre mode de vie et fréquence de PL dans certaines populations (notamment les Mongols, les Hazara-Tajiki, les Dinka, les Nuer et les Herero), l'hypothèse historico-culturelle explique de manière satisfaisante les fréquences de PL, sans qu'aucune des deux autres hypothèses ne permette d'améliorer le modèle correspondant à la première hypothèse seule. Itan *et al* (2009) sont également arrivés à la même conclusion avec une étude récente sur les populations européennes.

Persistance de la lactase en Asie ?

C'est dans tous les cas un bel exemple d'évolution convergente, où deux mutations ayant la même fonction ont été sélectionnées indépendamment, en Europe et en Afrique. Mais qu'en est-il en Asie, où le pastoralisme est tout de même extrêmement bien représenté par les populations des steppes, comme les populations mongoles, sibériennes (Evenk, Kizhi) et d'Asie Centrale, typiquement pastorales (voir figure 9 ci-dessous) ? Les récents travaux d'Outram (2009) montrent en effet qu'au Kazakhstan, les chevaux sont utilisés depuis au moins 3 500 BCE pour leur lait. De plus, le lait de jument est de loin un des plus concentré en lactose, notamment de moitié plus que celui des vaches (70g/L chez la jument pour 50 g/L chez la vache, Sharp, 1938). Dans cette région de pastoralisme par excellence, il est donc particulièrement intéressant de se demander quelle a été l'évolution génétique et phénotypique de la persistance de la lactase.

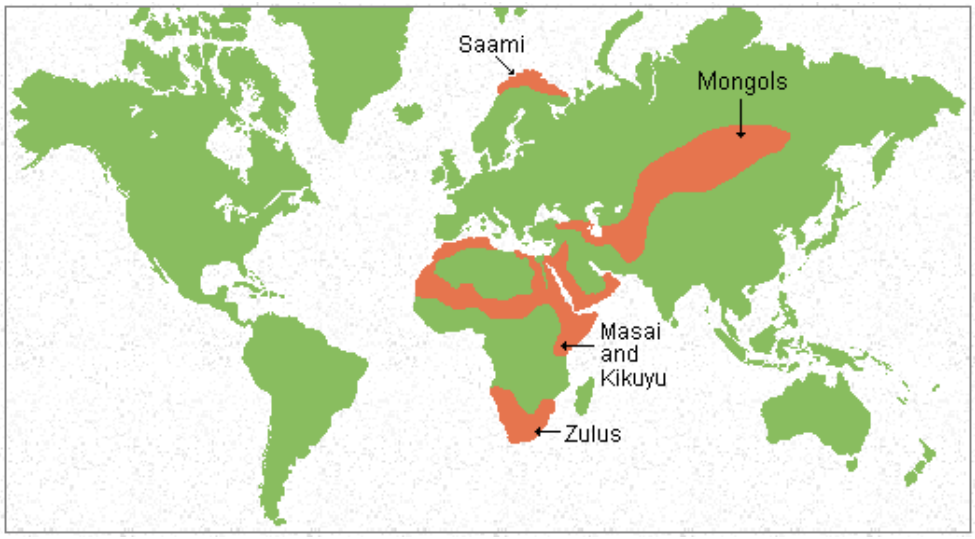


Figure 9: Carte mondiale montrant les régions traditionnellement pastorales en orange (produite par Dennis O'Neil, http://anthro.palomar.edu/subsistence/sub_3.htm)

La seule étude phénotypique effectuée dans la région, à l'Est de l'Asie Centrale (Wang *et al.*, 1984) montre qu'étonnamment, deux populations traditionnellement consommatrices de lait (des Mongols de Mongolie intérieure et des Kazakhs du Nord-est de la Chine), ne présentent pas de fortes fréquence de PL (respectivement 12.1% chez les Mongols et 23.6% chez les Kazakhs), ne les différenciant que peu de l'échantillon d'agriculteurs chinois (7.7%). Cependant, cette étude ne permet pas de savoir si ces populations présentent la mutation

causale européenne ou africaine et en quelle fréquence, ni si cette mutation est ou non sous sélection et depuis quand.

Mission de terrain en Asie Centrale

Ainsi, Evelyne Heyer et Patrick Pasquet, en collaboration avec l'équipe de « Génétique des populations, structuration, sélection, spéciation » dirigée par Miche Veuille (UMR 7205 CNRS-MNHN), ont effectué une mission en Asie Centrale dans l'aire de Boukhara, où de nombreux restes Néolithiques ont été trouvés (Brunet, 1999). Cette mission a eu pour objectif de récolter des données phénotypiques dans deux populations d'Asie Centrale : des Kazakhs traditionnellement éleveurs (n = 80), et des Tadjiko-Ouzbeks agriculteurs (n = 100) afin de mesurer le niveau de PL et de le corréler aux polymorphismes génétiques dans le gène de la lactase pour les mêmes individus. Pour l'étude phénotypique, trois différents critères ont été récoltés pour déterminer le statut de PL, après ingestion d'une dose de 50g de lactose : l'augmentation d'hydrogène dans l'air expiré, l'augmentation de glucose dans le sang, et l'augmentation de syndromes gastro-intestinaux ressentis par les individus. Cette étude montre tout d'abord que la fréquence de PL n'est pas très élevée dans les deux populations, et légèrement supérieure chez les éleveurs, bien que non significativement (respectivement 31.3% chez les Kazakhs et 21.8% chez les Tadjiko-Ouzbeks, p = 0.16). La fréquence de la mutation dominante -13.910 préalablement identifiée dans les populations européennes est de 15.7% et 10%, respectivement chez les Kazakhs et les Tadjiko-Ouzbeks. Le phénotype PL est de plus significativement corrélé à cette mutation (test de Fisher, $p < 10\text{-}8$ pour les deux ethnies), puisque 92% des Tadjiko-Ouzbeks et 94% des Kazakhs sont de phénotype PL avec le génotype CT ou TT, ou de phénotype non PL avec le génotype CC. Au total dans les deux ethnies, seuls neuf individus sont de phénotype PL alors qu'ils ont le génotype CC. Six d'entre eux ont été séquencé et ne présentent pas la mutation -14.010 africaine.

Objectifs de cette étude

Ces premiers résultats suggèrent qu'en Asie Centrale, l'étude du génotype de la mutation -13.910 nous informe sur le phénotype de persistance de la lactase de ces populations. Il nous a donc semblé intéressant de connaître l'hétérogénéité de la fréquence de cette mutation dans différentes populations d'éleveurs et d'agriculteurs en Asie Centrale, pour voir si sa faible fréquence est retrouvée à une échelle ethnique plus large. Ceci nous permettrait également de tester si la fréquence de la mutation varie entre différents types

d'éleveurs nomades d'Asie Centrale ayant potentiellement des modes de consommation de lait variés, par exemple entre les éleveurs des steppes comme les Kazakhs, les éleveurs des montagnes comme les Kirghiz et les Turkmènes mi-éleveurs mi-agriculteurs.

De plus, par rapport aux autres populations asiatiques qui ont une fréquence très faible de PL, la fréquence de 31.3% chez les Kazakhs est relativement élevée et il semble important d'estimer l'intensité de la sélection sur cette mutation et de dater le début de son expansion, pour mieux comprendre quelle pression de sélection a constitué la consommation de lait en Asie Centrale, et depuis quand.

B. Matériel et méthodes

1) Populations échantillonnées

En incluant les 180 individus déjà décrits plus haut, nous avons estimé la fréquence de la mutation -13.910 pour 617 individus au total, répartis dans 18 populations d'Asie Centrale (parmi les 26 populations précédemment décrites dans le tableau 1 et représentées sur la figure 10 ci-dessous selon leur mode de subsistance). Il s'agit de 392 individus dans 11 populations d'éleveurs traditionnellement nomades : 135 individus kirghiz (dans cinq populations), 128 individus kazakhs (dans deux populations), 45 individus karakalpaks (dans deux populations), 46 individus turkmènes (dans une population), 38 individus ouzbeks (dans une population), ainsi que 225 individus dans sept populations d'agriculteurs sédentaires : 125 individus tadjiks (dans six populations) et 100 individus tadjiko-ouzbeks (dans une population).

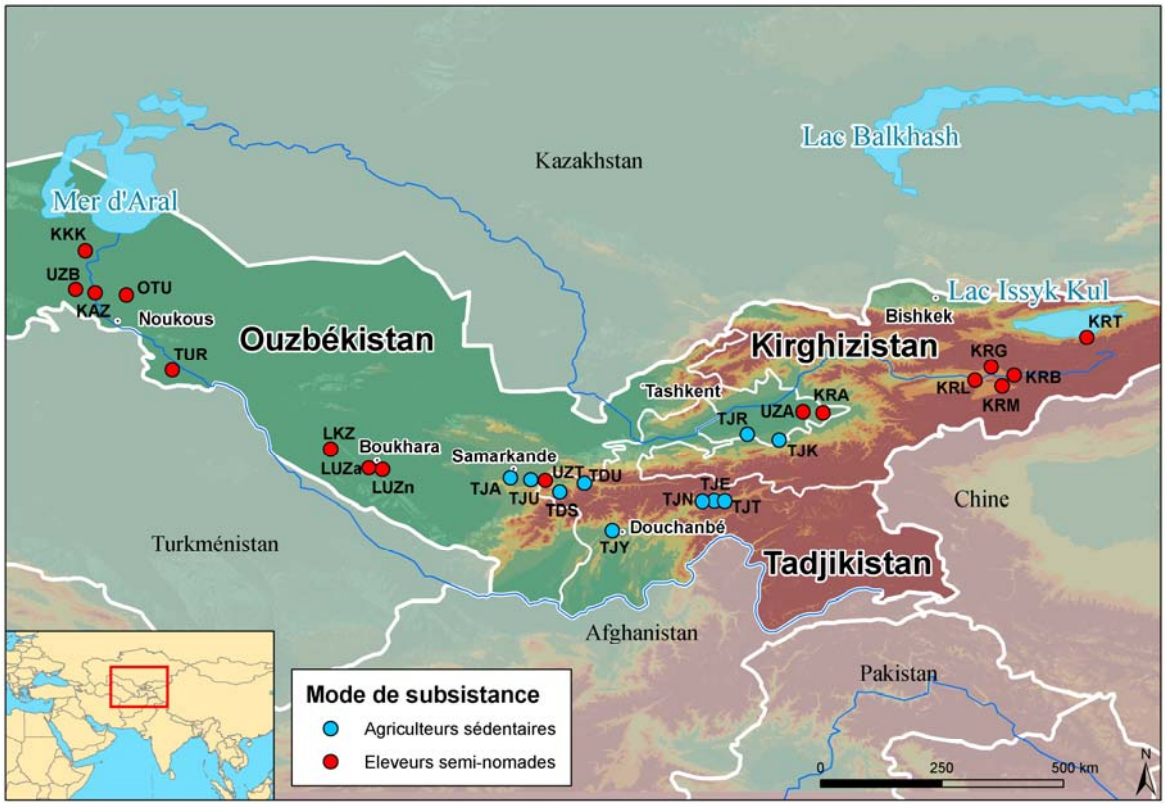


Figure 10 : Répartition des 26 populations échantillonnées en fonction de leur mode de subsistance. Source : Sylvain Théry.

2) Biologie moléculaire

L'ADN a été extrait à partir de sang ou de salive par une méthode standard d'extraction au phénol-chloroforme (Maniatis *et al.*, 1982). Les réactions de PCR (*Polymerase Chain Reaction*) ont été faites dans un volume final de 10µL composé de tampon Eppendorf 1X, de 125µM de chaque dNTP, 0.25U de Taq polymérase Ependorf, 160nM de chaque primer et 10ng d'ADN. Nous avons utilisé les primers Lac-C-L2 CTGCTTTGGTTGAAGCGAAGAT et Lac-C-M-U GCTGGCAATACAGATAAGATAATGGA introduisant un site de restriction, comme décrit précédemment (Enattah *et al.*, 2002, Mulcare *et al.*, 2004). Les PCR ont été effectuées dans le Master Cycler Eppendorf, avec une phase initiale de dénaturation de 5 min à 95°C; suivie par 30 cycles de 1 min à 95°C, 1 min à 59°C et 1 min à 72°C ; et une phase finale d'extension de 5 min à 72°C.

La détection de la mutation -13.910 a ensuite été effectuée par la méthode de RFLP (*Restriction Fragment Length Polymorphism*), qui consiste à digérer un fragment d'ADN pré-amplifié par des enzymes de restriction qui reconnaissent ou non les fragments selon la présence/absence de la mutation d'intérêt. Ainsi la longueur du fragment obtenu après digestion nous indique si le fragment contient ou non la mutation, et son état (homozygote ou hétérozygote). Nous avons ici effectué la digestion en ajoutant 15µL de produit de digestion (2.10^{-3}U de l'enzyme HinfI et 2X de tampon de digestion) aux 10µL de produits de PCR, pendant toute une nuit, à 37°C. Nous avons ainsi obtenu des bandes de tailles différentes révélées sur gel d'agarose 2% : 201pb pour un fragment non digéré (c'est-à-dire sans l'allèle T soit le génotype CC), 177 et 24pb avec l'allèle T (donc pour le génotype TT) et les trois fragments (201, 177 et 24pb) pour les hétérozygotes CT. Le fragment à 24pb n'étant pas détectable, le génotypage s'est fait selon l'absence / présence des bandes à 201 et 177pb. Ce génotypage a été effectué par moi-même, avec l'aide de Myriam Georges.

Le génotypage de 9 SNPs (*Single Nucleotide Polymorphism*) recouvrant environ 463kb autour de la mutation d'intérêt (voir tableau 3 ci-dessous), a également été effectué pour les individus pour lesquels nous avions les données phénotypiques et qui étaient de génotype CT ou TT pour la mutation -13.910. Ces données ont été obtenues par Lionel Brazier de l'équipe de Michel Veuille. La distance génétique (en cM) entre chaque SNP a été obtenue grâce aux données phasées des trios de HapMapIII (voir tableau 3 ci-dessous) :

	-254kb	-210kb	-163kb	-109kb	C/T-13.910
Nom du SNP	rs4953953	rs12649365	rs313528	rs1438307	rs4988235
Position (pb)	136 070 674	136 114 644	136 162 339	136 215 636	136 325 116
Position (cM)	150.3015129	150.3206827	150.3226422	150.3302184	150.3518052
	+25kb	**+59kb**	**+107kb**	**+157kb**	**+209kb**
Nom du SNP	rs1057031	rs309165	rs309142	rs309137	rs2011946
Position (pb)	136 350 432	136 383 771	136 431 794	136 482 421	136 534 086
Position (cM)	150.3538675	150.361383	150.3620502	150.3657048	150.4716136

Tableau 3 : Distances physiques (base de données UCSC : *University of California Santa Cruz*) et génétiques (base de données HapMapIII) pour les 10 SNP génotypés sur les populations d'Asie Centrale. En rouge la mutation -13.910 d'intérêt.

3) Analyses statistiques

Les haplotypes ont été phasés grâce au logiciel Phase utilisant une méthode Bayésienne (Stephens & Donnelly, 2003). Ces données haplotypiques nous ont permis d'estimer conjointement la date et l'intensité de l'expansion de la mutation, selon la méthode d'Austerlitz *et al* (2003). Cette méthode considère que l'haplotype ancestral est celui dont la fréquence est la plus forte dans la population. Chaque autre haplotype présent dans la population (avec l'allèle d'intérêt) est ensuite caractérisé en fonction de son état de recombinaison : nombre de SNPs de chaque côté sans recombinaison par rapport à l'haplotype ancestral et taux de recombinaison observés entre chaque SNPs, selon la base de données phasées HapMapIII. Finalement, cette méthode part de la fréquence moyenne de l'allèle d'intérêt dans la population, la structure des haplotypes comme décrite plus haut, et la taille initiale supposée de la population (seule donnée a priori), pour estimer conjointement le temps écoulé depuis le début de l'expansion de la mutation et son taux de croissance.

C. Résultats

1) Fréquence de la mutation -13.910 en fonction du mode de subsistance en Asie Centrale

La répartition des génotypes obtenus et la fréquence de la mutation d'intérêt dans chaque population, ainsi que par ethnie, sont présentées dans le tableau 4 et la figure 11 ci-dessous :

Mode de subsistance	Groupe ethnique	Population	CT	CC	TT	total	Fréquence de l'allèle T	Fréquence de PL prédite*	Fréquence de PL observée
Eleveurs nomades	Karakalpak	KAK	1	34	0	35	0,014	0.029	-
		OTU	2	8	0	10	0,100	0.200	-
	Kirghiz	KRB	2	28	0	30	0,033	0.067	-
		KRL	2	22	0	24	0,042	0.083	-
		KRM	3	24	0	27	0,056	0.111	-
		KRA	4	30	0	34	0,059	0.118	-
		KRG	4	16	0	20	0,100	0.200	-
	Kazakh	KAZ	3	41	1	45	0,056	0.089	-
		LKZ	24	58	1	83	0,157	0.301	0.318
	Turkmène	TUR	10	36	0	46	0,109	0.217	-
	Ouzbek	UZB	10	28	0	38	0,132	0.263	-
	Eleveurs		65	325	2	392	**0,088**	**0.152**	-
Agriculteurs sédentaires	Tadjik	TDU	12	22	0	23	0,022	0.043	-
		TJR	2	14	0	16	0,063	0.125	-
		TJE	3	19	0	22	0,068	0.136	-
		TJK	3	16	0	19	0,079	0.158	-
		TJY	6	20	0	26	0,115	0.231	-
		TJA	5	14	0	19	0,132	0.263	-
	Tadjiko-Ouzbek	LUZ	16	82	2	100	0,100	0.180	0.213
	Agriculteurs		36	187	2	225	**0,089**	**0.162**	

* Fréquence du phénotype de persistance de la lactase, prédite d'après la fréquence *p* de la mutation dominante -13.910 (p^2+2pq)

Tableau 4 : Fréquence de la mutation -13.910 et du phénotype de persistance de la lactase (PL) prédite d'après les données génétiques et observée, pour 18 populations d'Asie Centrale

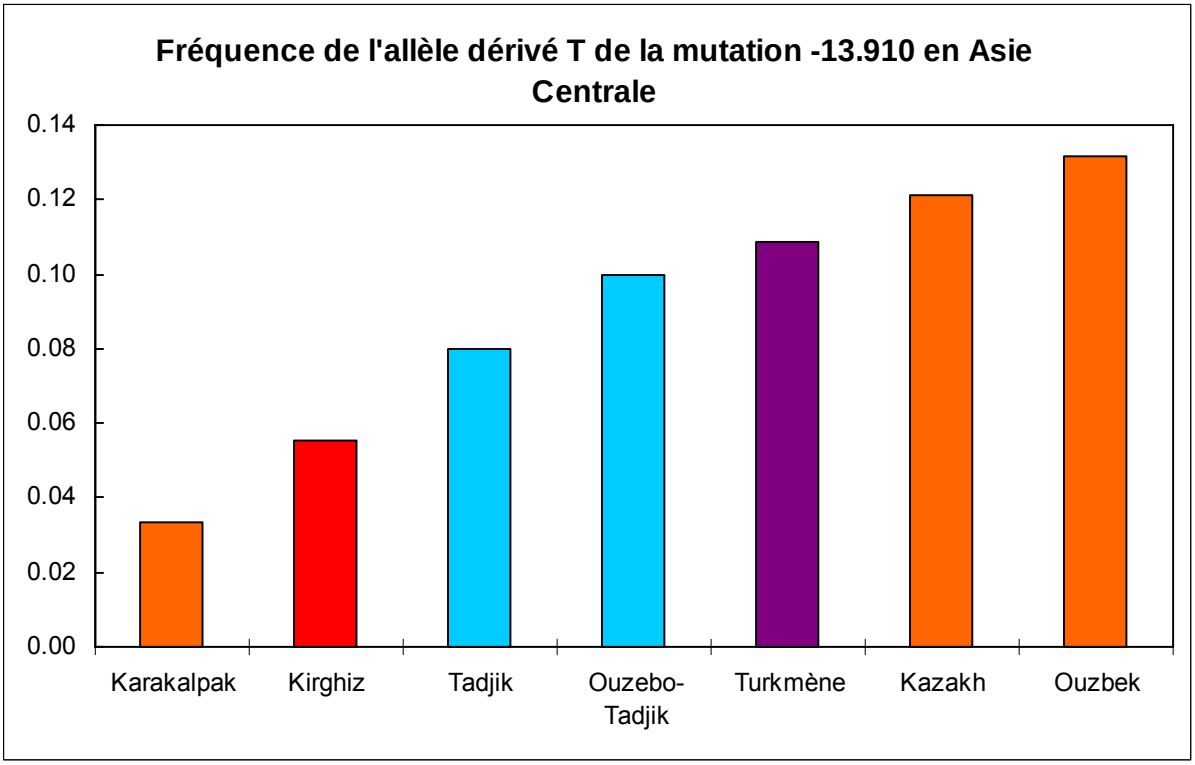


Figure 11 : Fréquence de l'allèle T dérivé de la mutation -13.910, dans les différentes ethnies étudiées. En rouge, éleveurs des montagnes ; en orange, éleveurs des plaines ; en violet, mi-éleveurs, mi-agriculteurs ; en bleu, agriculteurs.

Nous voyons qu'il y a une forte hétérogénéité des fréquences de la mutation -13.910 : entre 1.4% et 15.7% en fonction des populations. Cette hétérogénéité n'est cependant pas liée à l'affiliation ethnique puisque nous observons également une forte variabilité à l'intérieur des ethnies : notamment entre 3.3% et 10% au sein des Kirghiz, et entre 2.2% et 13.2% au sein des Tadjiks. De plus, les fréquences des éleveurs nomades (entre 1.4% et 15.7%, moyenne de 8.8%) recouvrent complètement celles des agriculteurs sédentaires (entre 2.2% et 13.2%, moyenne de 8.9%). Ainsi, le F_{ST} sur l'ensemble des populations est faible (bien que significatif : F_{ST} = 0.0095, p = 0.045), et la variation génétique n'est pas significativement expliquée par les différences de mode subsistance (F_{CT} = -0.00382, p = 0.998).

Il semble donc qu'en Asie Centrale, la fréquence de la mutation -13.910 n'est pas expliquée par le mode de subsistance des populations, avec dans une même ethnie de fortes variations de fréquences.

2) Estimation conjointe du temps et de l'intensité de l'expansion en Asie Centrale

Pour les individus tadjiko-ouzbeks (LUZa) et kazakhs (LKZ) ayant le génotype CT et TT pour la mutation -13.910, nous avons obtenu les données génétiques de 10 SNPs répartis sur 463kb (voir tableau 3), ce qui nous a permis de reconstruire les haplotypes phasés. Nous avons observé au total 6 haplotypes, (3 communs aux deux populations, un unique aux Tadjiko-Ouzbeks et deux uniques aux Kazakhs, voir tableau 5) :

Haplotype	**LUZ**	**LKZ**
TCGTTCTTTA	14	18
TCGTTCTTTC	5	3
TCGTTCTTCC	1	0
TCGTTCCTTA	0	1
TCGTTTTTTA	1	1
TTATTCCTTA	0	2
Total	21	25

Tableau 5 : Répartition des haplotypes trouvés d'après 9 SNPs situés autour de la mutation d'intérêt (en rouge)

Nous avons ensuite appliqué la méthode d'Austerlitz *et al* (2003) sur ces haplotypes phasés pour estimer conjointement le temps et l'intensité de l'expansion de cette mutation. Cette méthode suppose que la taille efficace actuelle de la population est connue. Nous avons donc testé l'influence de ce paramètre a priori en faisant varier cette taille efficace actuelle (10 000, 100 000 ou 1 000 000). Les estimations de la date d'expansion de la mutation sont obtenues en nombre de génération puis converties en années en prenant un temps de génération moyen de 25 ans (Reich & Goldstein, 1998). Les résultats obtenus pour les deux populations (LUZa et LKZ) sont présentés ci-dessous :

Population	Fréquence de l'allèle T	Taille de la population	Taux de croissance	Intervalle de confiance à 95%	Age (générations)	Intervalle de confiance à 95%	Age (années)	Intervalle de confiance à 95%
Tadjiko-Ouzbek	0.10	10.000	**1,0234**	1,0193 - 1,0353	**361,1**	307,4 - 484,4	**9 027,5**	7 685 -12 110
Tadjiko-Ouzbek	0.10	100.000	**1,0238**	1,0195 - 1,0357	**359,5**	304,2 - 481,5	**8 987,5**	7 605 -12 037,5
Tadjiko-Ouzbek	0.10	1.000.000	**1,0265**	1,0209 - 1,0385	**348,7**	287,5 - 461,2	**8 717,5**	7 187,5 - 11 530
Kazakh	0.15	10.000	**1,0295**	1,0244 - 1,0441	**296**	252,8 - 394,8	**7 400**	6 320 - 9 870
Kazakh	0.15	100.000	**1,0301**	1,0247 - 1,0448	**294**	249 - 391,4	**7 350**	6 225 - 9 785
Kazakh	0.15	1.000.000	**1,0345**	1,0274 - 1,0493	**282,7**	234 - 370,1	**7 067,5**	5 850 - 9 252,5

Tableau 6 : Estimation conjointe des dates et de l'intensité d'expansion de la mutation -13.910 dans deux populations (respectivement LUZ, agriculteurs tadjiko-ouzbeks et LKZ, éleveurs kazakhs) d'après la méthode d'Austerlitz *et al* (2003).

Nous voyons tout d'abord que les estimations de l'âge et l'intensité de l'expansion sont assez peu dépendantes de la taille efficace actuelle supposée. L'âge de l'expansion de la mutation est estimé entre 7 200 et 12 100 ans (5 200-10 100 BCE) pour la population tadjiko-ouzbèke et entre 5 800 et 9 900 ans (3 800-7 900 BCE) pour la population kazakh, avec en moyenne 1 500 ans d'écart entre les deux populations.

Les taux de croissance, qui cumulent la valeur de la croissance démographique de la population et le coefficient de sélection de la mutation, ont été estimés entre 1.019 et 1.038 pour la population tadjiko-ouzbèke et entre 1.024 et 1.049 pour les populations kazakhs. Des taux de croissance ont également été estimés avec la même méthode par Magalon *et al* (2008) sur des SNPs intergéniques, entre autres pour des populations tadjikes et kazakhes. Ces estimations donnent des valeurs comprises entre 1.006 et 1.019 pour les Tadjiks, et entre 1.009 et 1.024 pour les populations kazakhes. Ces intervalles ne se recoupent donc pas, respectivement pour chacune des populations, ce qui suggère que le taux de croissance de cette mutation est supérieur au taux de croissance démographique de la population.

Ainsi, la mutation permettant de digérer du lait à l'âge adulte aurait augmenté en fréquence sous l'action de la sélection naturelle, tout d'abord chez les agriculteurs d'Asie Centrale (5 200-10 100 BCE), puis plus tard chez les éleveurs (3 800-7 900 BCE), mais avec une intensité plus importante, engendrant une fréquence actuelle légèrement plus forte chez ces derniers.

3) Données HapMapIII

Afin de pouvoir directement comparer les résultats pour les populations d'Asie Centrale avec d'autres populations, nous avons cherché à estimer conjointement l'âge et l'intensité de l'expansion selon la même méthode (Austerlitz *et al.*, 2003) sur les populations de HapMap phase III (données disponibles sur http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/). Il s'agit de données publiques phasées à l'aide de trios (père / mère / enfant) pour 1184 individus dans 11 populations. Sur ces 11 populations, 4 populations sont des populations de diverses origines habitant aux Etats-Unis. Il est donc fort probable que ces populations se soient mélangées au cours du temps avec d'autres populations, ce qui complique l'interprétation de leur histoire évolutive. Nous ne les avons donc pas retenues pour notre analyse. Sur les sept restantes, les trois populations africaines ainsi que les deux populations asiatiques ne présentent aucun allèle T de la mutation -13.910, en accord avec la faible fréquence de cette mutation en Asie de l'Est et dans certaines populations africaines.

Nous n'avons donc pu effectuer ces analyses que sur les deux populations européennes : la population CEU (individus d'origine européenne, échantillonnés aux Etats-Unis) et la population TSI (Toscans d'Italie). Nous pouvons donc comparer directement les estimations obtenues en Europe et Asie Centrale.

Cependant, les SNPs obtenus sur les populations d'Asie Centrale n'étaient pas tous disponibles dans la base de données HapMap phase III. Ainsi nous avons remplacé les deux premiers SNPs par des SNPs proches (en bleu, voir tableau 7 ci-dessous) et nous avons enlevé un SNP (en gris, voir tableau 7 ci-dessous).

	-247kb	**-212kb**	**-163kb**	**-109kb**	**C/T-13.910**
Nom du SNP	Rs7608702	Rs2278731	Rs313528	Rs1438307	Rs4988235
Position (pb)	136078030	136112770	136162339	136215636	136325116
Position (cM)	150.3015328	150.3202848	150.3226422	150.3302184	150.3518052
	+25kb	**+59kb**	**+107kb**	**+157kb**	**+209kb**
Nom du SNP	Rs1057031	Rs309165	Rs309142	Rs309137	Rs2011946
Position (pb)	136350432	136383771	136431794	136482421	136534086
Position (cM)	150.3538675	150.361383	150.3620502	150.3657048	150.4716136

Tableau 7 : Distances physiques (base de données UCSC) et génétiques (base de données de HapMapIII) pour les 9 SNP génotypés sur les populations d'HapMapIII

Les résultats obtenus pour l'intensité et l'âge de l'expansion de la mutation, pour chacune des deux populations, sont présentés dans le tableau 8 ci-dessous.

Population	Fréquence de l'allèle T	Taille de la population	Taux de croissance	Intervalle de confiance à 95%		Age (générations)	Intervalle de confiance à 95%		Age (années)	Intervalle de confiance à 95%	
CEU	0,74418	10 000	**1,0542**	1,0400	1,0807	**132**	109	164.6	**3 300**	2 725	4 115
CEU	0,74418	100 000	**1,0726**	1,0578	1,1010	**126,8**	107,5	155,4	**3 170**	2 688	3 885
CEU	0,74418	1 000 000	**1,1081**	1,0909	1,1419	**110,8**	96,5	133,9	**2 770**	2 413	3 348
TSI	0,10228	10 000	**1,0317**	1,0152	1,0632	**129,5**	93,1	186,9	**3 237**	2 327	4 672
TSI	0,10228	100 000	**1,0614**	1,0431	1,0968	**111,5**	86,8	152,5	**2 787**	2 170	3 812
TSI	0,10228	1 000 000	**1,0946**	1,0756	1,1325	**102,3**	85,4	132,4	**2 557**	2 135	3 310

Tableau 8 : Estimation conjointe des dates et de l'intensité d'expansion de la mutation -13.910 dans deux populations de HapMap (respectivement CEU, d'origine Nord-Européenne et TSI, Toscans d'Italie) d'après la méthode d'Austerlitz *et al* (2003).

Pour la population CEU, les taux de croissance sont très forts (compris entre 1.06 et 1.14), en concordance avec les précédents résultats (coefficients de sélection entre 0.01 et 0.15). Les intervalles de confiance sont cependant moins grands et centrés sur les plus fortes valeurs. Les dates de début d'expansion sont également beaucoup plus précises : elles sont comprises entre 2 400 et 3 900 ans, ce qui exclue donc une expansion plus ancienne, contrairement aux précédentes estimations (dates comprises entre 2 200 et 21 000 ans). Ces dates sont également plus proches de celles trouvées pour la population Scandinave (comprises entre 1 600 et 3 200 ans). Pour la population italienne, nous retrouvons des dates d'expansion similaires (comprises entre 2 100 et 4 700 ans), mais associées à des coefficients de sélection légèrement plus faibles (comprises entre 1.02 et 1.13). La fréquence de la mutation dans cette population (10%) est en effet largement plus faible que celle dans la population CEU (74%).

D. Discussion

Tout d'abord, nous avons trouvé que la fréquence de l'allèle T de la mutation -13.910 est faible en Asie Centrale, entre 1.4% et 15.7% selon les populations, et n'est pas corrélée au mode de vie (8.8% en moyenne chez les éleveurs et 8.9% en moyenne chez les agriculteurs), ce qui n'est pas attendu sous l'hypothèse historico-culturelle. Ces fréquences génétiques permettent de prédire des fréquences phénotypiques entre 3.3% et 30.1% dans les populations d'éleveurs en Asie Centrale. Ainsi, la proportion d'individus pouvant digérer le lactose à l'âge adulte est plus faible que celles observées dans d'autres populations consommatrices de lait en Europe et en Afrique, mais est en accord avec les mesures effectuées chez les Kazakhs du Nord de la Chine (23.6% de PL, Wang *et al.*, 1984), ou dans d'autres populations pastorales asiatiques comme les Mongols (12.1-19% selon les sous-groupes, Wang *et al.*, 1984, Bersaglieri *et al.*, 2004), les Hazara Tajiki (15.4-18% selon les sous-groupes, Holden & Mace, 1997, Bersaglieri *et al.*, 2004) ou encore les Yakut (17.3%, Bersaglieri *et al.*, 2004).

Grâce à la méthode d'Austerlitz *et al* (2003), nous avons également pu estimer conjointement le temps et l'intensité de l'expansion de cette mutation. De manière générale, nos estimations ont des intervalles de confiance beaucoup moins grands que ceux trouvés dans la littérature. Nous avons vu que l'intensité de sélection de cette mutation en Asie Centrale (taux de croissance de 1.02-1.05) est plus importante que celle de mutations intergéniques dans les mêmes populations (Magalon *et al.*, 2008), mais elle est moins importante que celle estimée dans les populations Européennes par la même méthode (1.06-1.14 pour la population CEU). De plus, les deux populations utilisées pour ces estimations présentent une fréquence de persistance de la lactase assez élevée par rapport aux autres populations génotypées (Tadjiko-Ouzbek : 18% par rapport à la moyenne des agriculteurs : 16.2% et Kazakh : 30.1% par rapport à la moyenne des éleveurs 15.2%), et elles ne sont donc pas forcément représentatives de la situation générale en Asie Centrale.

Les dates trouvées en Asie Centrale pour l'expansion de la mutation (il y a 5 800-9 900 ans) sont plus anciennes que celles trouvées en Europe avec la même méthode (il y a 2 100-4 700 ans), ou en Afrique avec une autre méthode (il y a 1 200-18 500 ans, avec des moyennes entre 2 800 et 6 900 ans). Ces résultats contrastent avec les données archéologiques qui montrent que les traces les plus précoces d'utilisation du lait sont au Moyen-Orient (6 500

BCE en actuelle Turquie), puis en Europe (6 000-4 000 BCE), et finalement plus tard en Asie Centrale (3 500 BCE).

Remise en cause de l'hypothèse historico-culturelle

En définitive, le résultat qui nous semble le plus intéressant est le net manque de concordance entre la fréquence de la mutation -13.910 et le mode de vie en Asie Centrale, résultat pour lequel plusieurs explications sont possibles :

- Tout d'abord, pour concilier les fréquences proches entre populations aux modes de subsistance différents, nous pouvons imaginer que les agriculteurs d'Asie Centrale, au contact des populations nomades ou par leur propre culture, boivent suffisamment de lait frais pour créer une pression de sélection équivalente à celle des éleveurs nomades. Les agriculteurs adoptent en effet assez souvent des stratégies mixtes agropastorales. Cette hypothèse n'explique cependant pas pourquoi des fréquences plus fortes ne sont pas atteintes dans les deux groupes de populations.

- Il est possible, comme suggéré pour les populations africaines pastorales n'ayant pas de fréquences élevées de PL, que les modes de vies observés actuellement ne reflètent pas les modes de vie ancestraux, et certaines populations auraient donc subi des changements récents de mode de subsistance (Ingram *et al.*, 2009). Cette hypothèse peut expliquer des incohérences ponctuelles pour des populations à l'histoire méconnue, mais devient d'autant moins probable que l'on trouve de plus en plus de populations non concordantes (comme dans cette étude), et n'est de plus pas du tout confortée par les données archéologiques en Asie Centrale, qui montrent que le lait est utilisé depuis au moins 3 500 BCE (Outram *et al.*, 2009).

- Etant donnée que l'allèle T de la mutation -13.910 est quasi inexistant dans beaucoup de populations asiatiques, il se pourrait également qu'il n'ait pas été présente dans la variabilité génétique au moment de son expansion dans d'autres populations, et n'aurait donc commencé à augmenter en fréquence que très récemment quand il a été introduit secondairement. Cependant, nous avons trouvé des dates d'expansion de cette mutation plus vieilles en Asie Centrale qu'en Europe, allant donc dans le sens inverse de cette hypothèse. De plus, certaines populations asiatiques, bien que regroupées au sud-ouest de l'Asie,

montrent tout de même des fréquences intermédiaires de l'allèle T : 51-59% chez les Sindhi, Brâhuî, Baloutche, Makrani, et Pathan au Pakistan (Holden & Mace, 1997). De manière intéressante, aucune de ces populations ne dépend fortement du pastoralisme (environ 30% d'après l'atlas de Murdock et White, , 2006). Ainsi, au Pakistan la population la plus pastorale (les Hazara, 50% de pastoralisme) présente la fréquence la plus faible de la région (15-18%, Holden & Mace, 1997 , Bersaglieri *et al.*, 2004).

- Il se peut tout de même que les faibles fréquences de la mutation -13.910 chez les éleveurs d'Asie Centrale (ainsi que de Mongolie et de Sibérie) reflètent des brassages récents réguliers avec d'autres populations d'Asie de l'Est qui, elles, ne présentent presque pas de traces de cette mutation, limitant ainsi l'adaptation locale (Lenormand, 2002). Dans cette hypothèse, nous devrions retrouver une corrélation positive entre le niveau de différenciation neutre entre chaque population d'Asie Centrale et l'Asie de l'Est, et la fréquence de la mutation -13.910, ce qui est bien ce que l'on observe si l'on prend les Han comme population de référence pour l'Asie de l'Est (voir figure 12 ci-dessous), bien que ce ne soit pas significatif (test de Pearson, $p = 0.58$).

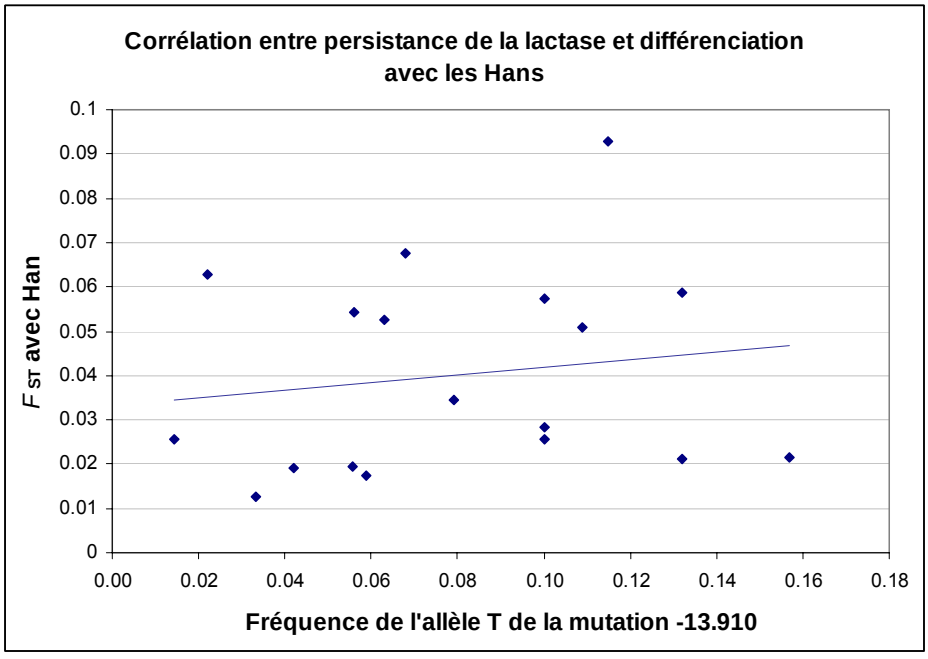


Figure 12 : Corrélation entre la fréquence de la mutation -13.910 de la lactase et le niveau de différenciation à l'Asie de l'Est (Hans)

- Il est également possible qu'en Asie, le mode de consommation du lait diffère de celui en Afrique ou en Europe, et n'engendre pas une pression de sélection aussi forte, malgré un contenu en lactose plus important. Une adaptation culturelle (par fermentation du lait par exemple) a pu rendre possible la consommation du lait sans changement physiologique. En effet, quand le lait est fermenté, sous forme de yaourt, de fromage, ou simplement de boisson fermentée à partir de lait (comme le « koumis », lait de jument fermenté dans une outre et très prisé en été par les Kirghiz), cela signifie que les bactéries ont digéré le lactose par fermentation, et donc les bénéfices nutritifs sont disponible pour l'Homme. Il est également possible que dans ces régions, la consommation de lait ne soit que saisonnière.

- Chez les Somali d'Ethiopie, il a été montré que les individus qui ne présentent pas de persistance de la lactase consomment de grandes quantités de lait frais sans ressentir de syndromes apparents (Ingram *et al.*, 2009). Ainsi, nous pouvons imaginer que dans certaines populations, l'adaptation à la consommation de lait frais se fait au niveau des bactéries de la flore intestinale, et ne nécessite pas d'adaptation génétique. En Asie Centrale, respectivement pour les Tadjiko-Ouzbeks et les Kazakhs, 37% et 22.8% des individus homozygotes CC pour la mutation -13.910 (donc génétiquement non PL) ne décrivent aucun syndrome intestinal après ingestion de lactose et respectivement 28.8% et 11.5% ne présentent pas d'augmentation d'hydrogène dans l'air expiré. Il y a donc tout de même une majorité d'individus qui montrent des signes physiologiques d'indigestion du lactose, surtout chez les éleveurs.

- Enfin, nous ne pouvons pas exclure que la pression de sélection qui explique la diversité génétique du phénotype de persistance de la lactase ne soit pas exclusivement liée à la consommation de lait mais soit associée à un autre facteur non identifié qui permettrait notamment de mieux expliquer la variabilité qu'il existe en Asie Centrale, entre les populations ayant une fréquence de persistance de la lactase autour de 5% et celles autour de 25%.

Conclusion

Finalement en Asie Centrale, il n'y a pas une bonne corrélation entre fréquence de PL et pastoralisme. Cette déviation par rapport à l'attendu sous l'hypothèse historico-culturelle n'est cependant pas spécifique à cette région du monde. Nous avons en effet déjà vu qu'en

Europe, de nombreuses populations présentent des fréquences importantes de persistance de la lactase (entre 85% et 96%), pour une dépendance intermédiaire au pastoralisme (entre 30.5% et 40.5%). Inversement, en Afrique, des populations fortement dépendantes du pastoralisme (entre 60.5% et 93%) présentent une fréquence faible de persistance de la lactase (entre 3% et 33%). En Asie, les populations qui présentent les fréquences de PL les plus fortes sur le continent (40-60%, au Pakistan) sont faiblement pastorales (entre 20 et 30%), tandis que les populations ayant 50% ou 80% de dépendance sur le pastoralisme (Hazara, Mongols et Yakut) ont des fréquences de PL ne dépassant pas les 26%.

Parmi les facteurs qui ont limité l'augmentation en fréquence du phénotype de PL dans les populations pastorales d'Asie Centrale, nous favorisons l'hypothèse d'une consommation de lait exclusivement saisonnière, ou d'une fermentation systématiquement ayant amoindri les pressions de sélection en Asie Centrale. Ainsi, ces populations ne seraient pas dépendantes du lait (c'est-à-dire que leur régime alimentaire ne serait pas majoritairement basé sur du lait frais, et la digestion asymptomatique du lait ne serait pas vitale pour les individus). Des études poussées d'anthropologie nutritionnelle comparative permettraient certainement de mieux avancer sur cette question des pratiques autour de cet aliment controversé qu'est le lait.

III. Adaptation à la consommation de viande ?

Cette partie repose sur l'article suivant présenté en annexe 5 : **Ségurel L.**, Lafosse S., Heyer E. & Vitalis R. Frequency of the AGT Pro11Leu polymorphism in humans: does diet matter? *Ann. Hum. Genet.* 2010 Jan; 64:1.

A. Problématique

L'AGT (alanine : glyoxylate aminotransferase), enzyme exprimée dans le foie, est normalement responsable de la conversion du glyoxylate en glycine (Danpure & Jennings, 1986). Si cette conversion n'est pas assurée, le glyoxylate s'accumule et des sels de calcium insolubles se forment, entraînant alors progressivement une insuffisance rénale. Cette déficience d'AGT et les symptômes associés sont caractéristiques de l'hyperoxalurie primaire de type I (PH1, MIM 259900[6]). Il a été estimé qu'avant l'introduction des thérapies modernes, cette maladie entraînait le décès de 80% des patients atteints avant l'âge de 20 ans (Williams & Smith, 1983).

Dans un tiers des cas, cette maladie n'est pas liée à l'absence d'AGT, mais plutôt à sa mauvaise localisation (Danpure *et al.*, 1990). L'enzyme doit en effet être présente au même site que les précurseurs du glyoxylate pour être efficace. Chez les mammifères, cette localisation dépend du régime alimentaire (Danpure *et al.*, 1990) : chez les herbivores, le précurseur du glyoxylate (le glycolate) ainsi que l'AGT sont présents dans les peroxysomes (Noguchi, 1987); chez les carnivores, au contraire, le précurseur de glyoxylate (l'hydroxyproline) ainsi que l'AGT sont présents dans les mitochondries (Takayama *et al.*, 2003). Chez l'Homme, l'AGT est normalement exprimée au niveau des peroxysomes (Cooper *et al.*, 1988), reflet du régime alimentaire herbivore de nos ancêtres (Holbrook *et al.*, 2000).

Cependant, l'allèle dérivé T de la mutation non synonyme Pro11Leu sur le gène *AGXT* codant l'AGT, permet de changer la localisation de cette protéine. A l'état homozygote, l'allèle T entraîne la relocalisation de 5% de l'AGT dans les mitochondries (Purdue *et al.*, 1990). Dans ce cas, les 95% restant aux peroxysomes suffisent pour convertir correctement le glyoxylate. Par contre, quand cet allèle est associé à une autre substitution, Gly170Arg, plus

[6] Le numéro MIM d'une maladie (pour « *Mendelian Inheritance in Man*») correspond à sa référence dans la base de données NCBI (*National Center for Biotechnology Information*)

de 90% de l'AGT est délocalisée aux mitochondries, entraînant alors l'hyperoxalurie primaire de type I (Danpure *et al.*, 1989).

L'allèle T de la mutation Pro11Leu est donc associé à une maladie létale, et pourtant il est trouvé en forte fréquence dans certaines populations humaines (5-20% chez les Caucasiens, Danpure *et al.*, 1994b, 1994a). C'est pourquoi il a été proposé que cet allèle, quand il n'est pas associé à Gly170Arg, doit être avantageux dans des populations ayant une grande proportion de viande dans leur alimentation, de par la relocalisation d'une petite partie de l'AGT vers les mitochondries (Danpure, 1997). Ainsi, l'attendu sous cette hypothèse est de trouver des fréquences de l'allèle T de la mutation Pro11Leu plus importantes dans les populations d'éleveurs que dans celles d'agriculteurs.

Données mondiales sur la fréquence de Pro11Leu en lien avec le mode de vie

Caldwell *et al* (2004) ont testé l'hypothèse d'un avantage sélectif de la mutation Pro11Leu dans les populations ayant une alimentation riche en viande, en déterminant la fréquence de cette mutation dans diverses populations humaines, et en comparant le niveau de différenciation génétique pour cette mutation à celui de marqueurs neutres. Leur étude montre que les Samis, population d'éleveurs de rennes consommant de grandes quantités de viande (Haglin, 1991, 1999), ont la fréquence la plus forte de la mutation Pro11Leu (27.9%) tandis que les populations ayant les fréquences les plus faibles sont des populations ayant un régime plus mixte (2.3% en Chine), ou clairement végétarien (3% en Inde). Ces résultats suggèrent que la fréquence de Pro11Leu est corrélée au régime alimentaire ancestral des populations humaines. Cependant, la faible fréquence de cette mutation chez les Mongols (6.7%), population d'éleveurs consommant de grandes quantités de viande (Hruschka & Brandon, 2004), n'est pas en accord avec cette hypothèse, et n'est pas discutée dans cette étude.

Caldwell *et al* (2004) montrent de plus que le niveau de différenciation génétique pour certaines paires de populations est plus fort pour la mutation Pro11Leu que pour le reste du génome, bien que non significativement ($p = 0.074$ et $p = 0.266$ pour les comparaisons des Sami par rapport aux Chinois et des Sami par rapport aux Nigérians, respectivement). Ce résultat est interprété dans cette étude comme un argument en faveur de l'influence de la sélection naturelle sur cette mutation, au moins dans certaines populations.

Objectifs de l'étude

L'hypothèse d'un avantage sélectif de cette mutation, différent en fonction du passé alimentaire des populations, est particulièrement intéressante mais les données présentées dans l'étude de Caldwell *et al* (2004) ne permettent toutefois pas de trancher dans un sens ou un autre. Afin de tester la robustesse de cette hypothèse, nous avons déterminé la fréquence de la mutation Pro11Leu dans sept populations d'éleveurs et quatre populations d'agriculteurs en Asie Centrale. L'objectif de notre étude est 1) de décrire la fréquence de cette mutation dans différentes populations d'Asie Centrale aux régimes alimentaires ancestraux différents, 2) de détecter si, de manière générale, la mutation est sous l'effet de la sélection en comparant la différenciation génétique pour cette mutation à celle de marqueurs neutres, enfin 3) de tester si la différenciation génétique pour cette mutation est compatible avec de l'adaptation locale en lien avec le régime alimentaire.

B. Résultats

Le génotypage de la mutation Pro11Leu dans les populations d'Asie Centrale a été effectué conjointement par Sophie Lafosse et moi-même.

Nous avons trouvé des fréquences hétérogènes de la mutation Pro11Leu en Asie Centrale (entre 1.7% et 26.9%, voir tableau 9 ci-dessous), avec en moyenne une fréquence de l'allèle dérivé T plus importante chez les agriculteurs que chez les éleveurs (22.6% et 10.3%, respectivement ; test de Wilcoxon, $p = 0.003$).

Ethnie	**Acronyme**	***n***	**CC**	**CT**	**TT**	***p* (%)**
Tadjik	TJR	17	11	6	0	**17.6**
Tadjik	TJE	23	15	7	1	**19.6**
Tadjik	TDU	24	16	6	2	**20.8**
Tadjik	TJY	26	12	14	0	**26.9**
Agriculteurs		**73**	**43**	**27**	**3**	**22,6**
Karakalpak	KKK	23	17	5	1	**15.2**
Kazakh	KAZ	30	29	1	0	**1.7**
Kazakh	LKZ	49	40	9	0	**9.2**
Kirghiz	KRA	32	24	7	1	**14.1**
Kirghiz	KRG	20	17	3	0	**7.5**
Kirghiz	KRB	26	20	6	0	**11.5**
Turkmène	TUR	34	26	7	1	**13.2**
Eleveurs		**214**	**173**	**38**	**3**	**10,3**

n : nombre d'individus analysés ; *p* : fréquence de l'allèle dérivé T de la mutation Pro11Leu dans le gène *AGXT*.

Tableau 9 : Génotypes de la mutation Pro11Leu dans différentes populations d'Asie Centrale et fréquence de l'allèle T dérivé. Le lieu d'échantillonnage de chaque population est disponible dans le tableau 1.

Ce résultat contredit donc l'hypothèse d'une plus forte fréquence de la mutation Pro11Leu (et donc d'un avantage sélectif) dans les populations qui consomment le plus de viande, mais est en accord avec les faibles fréquences trouvées par Caldwell *et al* (2004) chez les Mongols.

Nous avons ensuite cherché à comprendre si la mutation Pro11Leu était de manière générale sous l'action de la sélection naturelle, en comparant son niveau de diversité à celui

d'autres régions du génome *a priori* neutres, c'est-à-dire loin des gènes connus. Nous avons pour la mutation Pro11Leu utilisé conjointement nos données (11 populations d'Asie Centrale) et celles de Caldwell *et al* (2004) pour 5 populations mondiales (Chinois, Mongols, Arméniens, Gallois et Nigériens). Pour les régions neutres, nous avons utilisé 26 marqueurs microsatellites autosomaux (ceux génotypés en Asie Centrale et présentés dans la première partie de la thèse). Les données pour ces mêmes marqueurs étaient également disponibles pour la base de données HGDP-CEPH (*Human Genetic Diversity Panel - Centre d'Etude du Polymorphisme Humain*), d'où nous avons récupéré les données pour 5 populations proches de celles inclues dans l'étude de Caldwell (Chinois, Mongols, Adygeis, Orcadiens et Yorubas).

Conditionnellement aux valeurs observées de F_{ST} sur les marqueurs neutres (F_{ST} = 0.019), nous avons simulé par un processus de coalescence des données sous un modèle en île (Wright, 1951), pour obtenir la distribution attendue sous neutralité de F_{ST} en fonction de l'hétérozygotie (Beaumont & Nichols, 1996). Comme présenté dans la figure 13 ci-dessous, nous voyons que le niveau de différenciation entre populations n'est pas significativement plus fort pour Pro11Leu (F_{ST} = 0.025) que pour les régions neutres (F_{ST} = 0.019, p = 0.214).

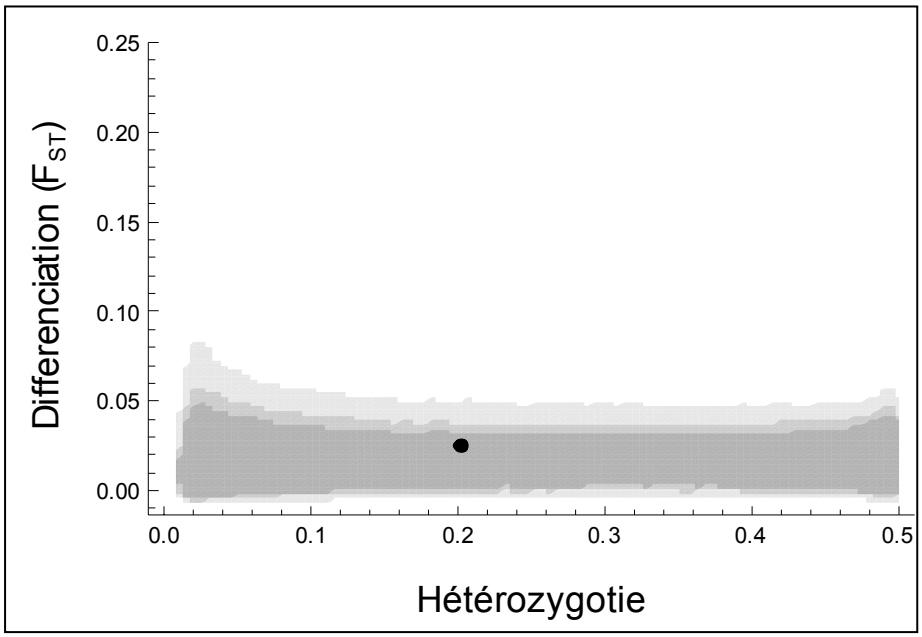


Figure 13 : Comparaison du niveau de différenciation génétique (F_{ST}) mesuré sur la mutation Pro11Leu (représentée par le point noir sur le graphique), et celui attendu sous l'hypothèse nulle de neutralité sur les marqueurs neutres autosomaux (les différents niveaux de gris représentent les intervalles de confiance à 90%, 95% et 99%).

Ce résultat nous permet de voir qu'il n'y a pas de signe d'adaptation locale entre ces populations pour la mutation Pro11Leu. Cependant, dans cette approche, nous regardons si, en moyenne, les valeurs de F_{ST} sont plus fortes qu'attendues ou non, et donc nous ne mettons pas d'*a priori* sur les groupes de populations étant sous adaptation locale. Afin de tester directement quelle est l'influence du régime alimentaire sur la fréquence de cette mutation, nous avons donc calculé les valeurs de F_{CT} (part de la variation due au regroupement des populations en fonction de leur régime alimentaire) sur les données simulées, pour les comparer à la valeur obtenue pour la mutation Pro11Leu. Les F_{CT} ont été estimés par une analyse moléculaire de variance (AMOVA, Excoffier *et al.*, 1992). Les résultats sont présentés dans la figure 14 ci-dessous :

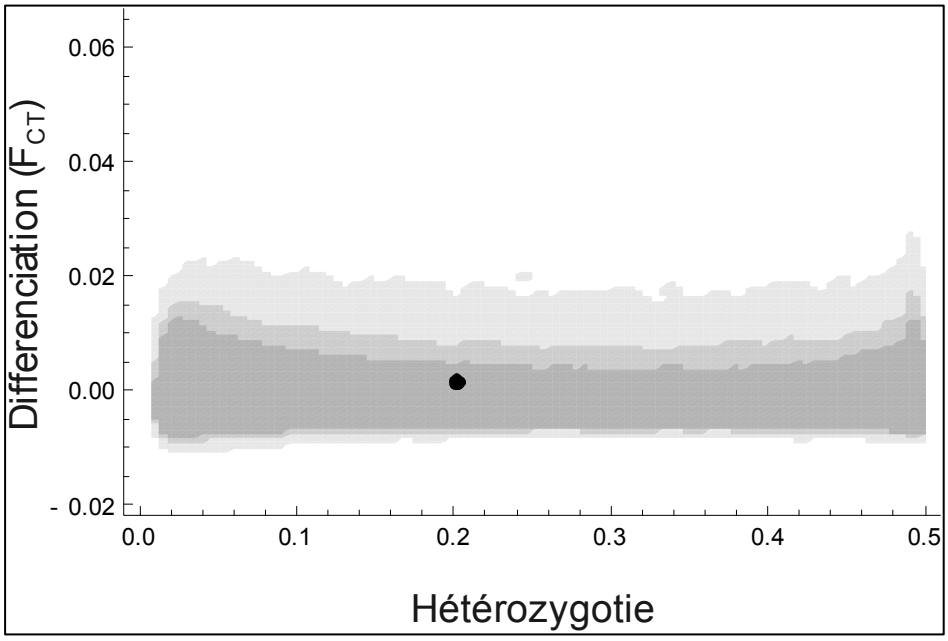


Figure 14 : Comparaison du niveau de différenciation génétique dû au regroupement des populations d'après leurs régimes alimentaires (F_{CT}), sur la mutation Pro11Leu (représentée par le point noir sur le graphique) et celui attendu sous l'hypothèse nulle de neutralité, étant donné la différenciation mesurée sur les marqueurs neutres autosomaux (les différents niveaux de gris représentent les intervalles de confiance à 90%, 95% et 99%).

Ces résultats montrent que la répartition de la diversité génétique sur la mutation Pro11Leu en tenant compte du régime alimentaire des populations (F_{CT} = 0.002), n'est pas significativement différente de celle attendue sous neutralité (p = 0.187).

Bien que nos résultats montrent que la diversité génétique de la mutation Pro11Leu du gène *AGXT* ne semble pas sous l'influence de la sélection naturelle, mais est au contraire en accord avec une influence de la démographie seule, nous ne pouvons pas exclure d'autres hypothèses alternatives impliquant l'action de la sélection naturelle.

Notamment, il est possible qu'il soit difficile de détecter une signature de sélection adaptative sur cette mutation, car elle est à la fois délétère, quand elle est associée à Gly170Arg, et avantageuse, quand elle n'est associée à aucune autre allèle à risque. Cette hypothèse ne permet cependant pas de concilier les écarts importants de fréquence de cette mutation entre différentes populations.

Les fortes fréquences de la mutation Pro11Leu dans certaines populations d'agriculteurs pourraient également être expliquées par le fait que la transition Néolithique vers l'agriculture soit trop récente dans ces populations, ou bien que leur régime alimentaire soit trop riche en viande. A l'inverse, les faibles fréquences de la mutation Pro11Leu chez les éleveurs pourraient être expliquées par des différences de fréquence de la mutation Gly170Arg entre Asie et Europe (avec notamment des fréquences plus importantes en Asie) ou la présence d'une autre mutation permettant de convertir le glyoxylate présent dans la viande en Asie (adaptation convergente). Cette hypothèse est cependant peu probable, étant donné que le gène *AGXT* a déjà été largement étudié du fait de son implication dans la maladie PH1 (Danpure, 2004).

En conclusion, nos résultats ne supportent pas l'hypothèse que la diversité génétique de la mutation Pro11Leu diffère de celle attendue sous neutralité. Cependant, nous ne pouvons pas rejeter des scénarios plus complexes combinant plusieurs hypothèses alternatives, qui permettraient alors de concilier à la fois les fortes fréquences de la mutation Pro11Leu dans certaines populations d'agriculteurs, et les faibles fréquences dans certaines populations d'éleveurs. L'analyse de données de séquençage constituerait certainement une approche plus puissante pour trancher parmi ces différents scénarios, grâce à l'utilisation de tests de neutralité séparément pour chaque population. Cependant, notre étude avait pour objectif de tester si la diversité génétique mondiale de la mutation Pro11Leu était mieux expliquée par de la démographie ou de l'adaptation en fonction du mode de vie, et notre conclusion (qu'il n'y a pas pour cette mutation d'écart à la neutralité) supporte l'idée que des importantes différences de fréquences alléliques entre populations sont souvent compatibles avec des processus

démographiques ((Hofer *et al.*, 2009, Coop *et al.*, 2009)). Nous ne chercherons donc pas à dater la transition des régimes alimentaires en Asie Centrale à partir de cette mutation.

IV. Quand l'adaptation devient maladaptation : le cas du diabète de type II

A. Problématique

Le glucose, molécule énergétique indispensable à notre organisme, nous est fourni grâce à notre alimentation. L'augmentation du taux de glucose dans le sang (après un repas, par exemple) stimule la sécrétion d'insuline par le pancréas, enzyme responsable de l'absorption du glucose dans la plupart des cellules pour le convertir en énergie. L'insuline est chargée de convertir le glucose en glycogène (molécule de stockage) au niveau des cellules musculaire et hépatiques, et favorise également le stockage de graisse dans le tissu adipeux. Cependant, cet équilibre peut être perturbé, comme dans le cas du diabète, caractérisé entre autre par un taux élevé de glucose dans le sang. Cet excès de glucose peut avoir de sévères conséquences à long terme (angiopathie, cardiomyopathie, néphropathie, rétinopathie, neuropathie, etc.).

Parmi les différentes formes de diabète, le diabète de type II (MIM 125853[7]) est de loin le plus fréquent, et devient souvent symptomatique aux alentours de 40 ans. Il est caractérisé par une résistance à l'insuline, combinée avec une réduction (relative ou absolue) de la sécrétion de l'insuline. La résistance à l'insuline est liée à des défauts idiopathiques de récepteur à l'insuline, qui engendrent une inefficacité de l'absorption du glucose et à terme, son accumulation dans le sang.

Le diabète de type II semble largement multifactoriel. D'un côté, les risques relatifs sont de 4 à 6 fois plus importants chez les frères et sœurs (Florez *et al.*, 2003), et les taux de concordances de la maladie sont plus élevés chez les jumeaux monozygotes que chez les dizygotes (Kaprio *et al.*, 1992, Poulsen *et al.*, 1999), suggérant ainsi en partie une étiologie génétique. D'un autre côté, 55% des patients diagnostiqués avec un diabète de type II présentent de l'obésité (Anderson *et al.*, 2003), montrant l'influence de facteurs environnementaux comme l'alimentation et l'activité physique. L'obésité centrale (concentration de la graisse au niveau de la taille, dans les organes abdominaux) est un facteur de risque particulièrement important pour la résistance à l'insuline. L'environnement intra-utérin a également une influence non négligeable : des enfants de 15-19 ans ont une

[7] Le numéro MIM d'une maladie (pour « *Mendelian Inheritance in Man*») correspond à sa référence dans la base de données NCBI (*National Center for Biotechnology Information*)

prévalence de diabète de 1.4% s'ils sont nés de mères non diabétiques qui ont développé un diabète plus tard, et de 33% si la mère a développé un diabète avant la naissance (Pettitt *et al.*, 1981).

Prévalence du diabète de type II dans les populations humaines

Le diabète de type II, comme d'autres maladies liées au mode de vie et regroupées sous le nom de « maladies des civilisations » (obésité, hypertension, maladies cardio-vasculaires, etc.) présente des prévalences très élevées dans certaines populations, et est considéré comme une des principales menaces pour la santé humaine au niveau mondial (Zimmet *et al.*, 2001). Sa répartition entre populations humaines est assez inégale (voir King & Rewers, 1993, et la figure 15 ci-dessous).

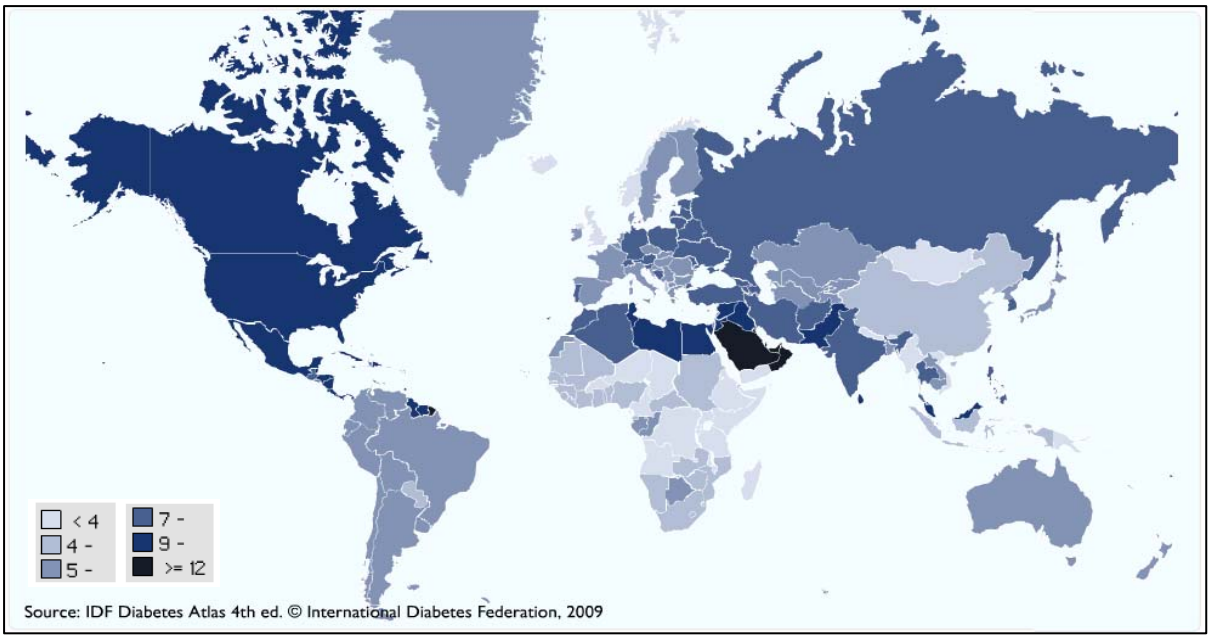


Figure 15 : Prévalence (en %) du diabète de type II en 2009 parmi les 20-79 ans selon les estimations de la quatrième édition de l'atlas du diabète (http://www.diabetesatlas.org/map)

Nous voyons par exemple sur cette carte que la péninsule arabique, quelques pays du Moyen Orient, l'Egypte, l'Algérie, le Pakistan, la Malaisie, ainsi que l'Amérique du Nord et la Guyane sont particulièrement à risque. Cependant, nous ne voyons pas sur cette carte les nombreuses îles qui présentent les prévalences parmi les plus hautes au niveau mondial comme déjà noté par Zimmer *et al* (Zimmet *et al.*, 1990 , 1991 , 1992) : notamment de 13 à 31% en Polynésie et Micronésie (selon les îles), de 16.2% à l'île Maurice, et de 10 à 12 %

dans les îles des Caraïbes (selon l'atlas du diabète, http://www.diabetesatlas.org/downloads). Cette carte ne permet pas non plus de bien visualiser les différences entre ethnies au sein des pays. Par exemple, en Australie, les descendants de populations Européennes ont systématiquement des prévalences moins élevées que les Aborigènes d'Australie dans les mêmes conditions environnementales (O'Dea, 1991). De la même manière, aux Etats-Unis, l'incidence[8] du diabète de type II est 19 fois plus forte chez les Pima (Amérindiens) que chez les descendants de populations européennes. Les Pima ont par ailleurs la prévalence la plus forte au monde (50% chez les individus de plus de 35 ans, Knowler *et al.*, 1978, 1983). Ces grandes différences entre ethnies suggèrent la répartition inégale des facteurs de risque génétiques.

Ces données montrent finalement que le diabète de type II a une prévalence considérable dans certaines populations. Etant données les conséquences de cette maladie sur la valeur sélective des individus, les mutations associées à cette maladie devraient pourtant être contre sélectionnées (sélection négative). Pour expliquer ce paradoxe, plusieurs hypothèses ont proposé l'existence d'un avantage sélectif de ces mutations dans le passé, qui les auraient maintenues en fréquence dans la population.

Hypothèse du « génotype économe »

L'hypothèse du « génotype économe », proposée par Neel dès 1962 ((1962)) et modifiée depuis (Neel, 1976 , 1982 , 1992), propose que dans les sociétés humaines où la disponibilité en nourriture est cyclique, avec des périodes de relative abondance et d'autres de famine, il peut être bénéfique de mettre en place des réserves de graisse quand l'environnement le permet, pour pouvoir survivre pendant les périodes de disette à venir. Il faut en effet pouvoir subvenir à tout moment aux besoins des organes fortement consommateurs tel que le cerveau. Un stockage accru de graisse peut être entre autres favorisé par une « résistance à l'insuline » au niveau des muscles ou du foie, c'est-à-dire que l'insuline serait moins efficace pour stocker le glucose au niveau de ces organes, mais favoriseraient toujours un stockage de graisse au niveau du tissu adipeux.

Comme 95% du passé de l'Homme moderne correspond à un mode de vie de chasseur-cueilleur (c'est à dire depuis environ 200 000 BCE jusqu'à 10 000 BCE), avec une

[8] L'incidence étant le nombre de nouveaux cas de la maladie sur une période donnée, classiquement sur un an.

disponibilité de nourriture variable, il semble assez probable que des mutations permettant d'économiser les molécules énergétiques aient été sélectionnés. Mais actuellement, dans les milieux où la nourriture est abondante et riche en calories, et où l'activité physique est également fortement réduite, comme dans les sociétés industrialisées, et particulièrement en milieu urbain, ces mêmes allèles entraîneraient une stimulation constante du pancréas pour produire de l'insuline, qui finirait à terme par rendre l'organisme insensible à cette enzyme. Ainsi, les allèles sélectionnés seraient devenus un lourd fardeau génétique et favoriseraient notamment le développement du diabète de type II. Selon cette hypothèse, plus les populations humaines ont subi de famines et de fluctuations de disponibilité de nourriture dans le passé, plus elles ont sélectionné le « génotype économe » et plus leur prévalence de résistance à l'insuline doit être élevée.

Cependant, les attendus des différences entre éleveurs et agriculteurs sous cette hypothèse ne sont pas si clairs. En effet, contrairement à la vision classique où les chasseurs-cueilleurs et les éleveurs sont des populations particulièrement sujets à des périodes d'insécurité alimentaire, certaines données ont permis de montrer que les agriculteurs subissent autant de variations saisonnières de disponibilité en nourriture que les autres (Benyshek & Watson, 2006, Prentice *et al.*, 2008). Wendorf *et al* (1989) ont quant à eux suggéré que les variations sont d'autant plus importantes que l'on s'éloigne de l'équateur vers les pôles, en raison notamment de la plus grande spécialisation des modes de subsistance à haute latitude.

Hypothèse de la « piste carnivore »

Une autre hypothèse, celle de la « piste carnivore », a été émise par Brand-Miller & Colagiuri en 1994 et en 2002 (Brand Miller & Colagiuri, 1994, Colagiuri & Brand Miller, 2002). Alors que l'hypothèse du « génotype économe » décrit une contrainte imposée par des cycles d'abondance et de famine, celle de la « piste carnivore » s'appuie sur la quantité limitante de glucose dans l'alimentation des chasseurs-cueilleurs, du fait de leur régime alimentaire carnivore. En effet, il y a environ 3 millions d'années, nos ancêtres étaient principalement végétariens (Gaulin & Konner, 1977) et leur physiologie a évolué en accord avec le glucose comme principale source d'énergie (Sokoloff *et al.*, 1977) ;(Frienkel, 1980). Mais il y a environ 2 millions d'années, au moment des premières périodes glaciaires, la végétation a été fortement réduite (Fagan, 1992), et nos ancêtres ont commencé à consommer

de plus grandes quantités d'aliments d'origine animale (Garn & Leonard, 1989). Ainsi, la molécule énergétique de base, le glucose, n'était plus que faiblement disponible. Afin d'assurer au cerveau un influx constant de 1g/L en glucose (taux nécessaire à son fonctionnement), de fortes pressions de sélection sont apparues, avec notamment un avantage important à développer de la résistance à l'insuline.

Depuis, avec la fin des périodes glaciaires et l'arrivée du Néolithique il y a environ 10 000 ans, les agriculteurs ont retrouvé une part importante de glucides dans leur alimentation, notamment grâce aux céréales (Eaton & Konner, 1985). Cette pression de sélection pourrait donc avoir eu le temps de s'être relâchée, et le génotype favorisant la résistance à l'insuline de diminuer en fréquence. Au contraire, dans les populations d'éleveurs et de chasseurs-cueilleurs, le régime alimentaire est resté majoritairement basé sur les protéines et les lipides, et ce génotype serait resté avantageux. Ainsi, l'attendu sous cette hypothèse est plus clair que précédemment : la prévalence de résistance à l'insuline devrait être inversement liée au temps depuis lequel les populations ont eu de nouveau des quantités satisfaisantes de glucose dans leur alimentation. Brand-Miller & Colagiuri (1994) ont cependant également proposé que la dérive génétique ait pu accentuer ces écarts entre populations, notamment pour les Pima et les Nauru (en Micronésie).

Objectifs de l'étude

Finalement, selon ces deux hypothèses (résumées dans la figure 16 ci-dessous), la résistance à l'insuline est le phénotype asymptomatique qui aurait été avantageux et sélectionné dans le passé, tandis que le diabète de type II correspond à l'expression de ce phénotype mal adapté dans notre environnement actuel. Les mutations associées à la résistance à l'insuline devraient donc avoir été sous sélection pendant la majeure partie du passé de l'Homme moderne, bien que dans certaines populations, cette pression de sélection ait pu se relâcher depuis le Néolithique. Cette prédiction a l'avantage de pouvoir être facilement testée, en regardant 1) si les gènes liés à la résistance à l'insuline présentent des signatures de la sélection naturelle et 2) si des populations aux modes de vie ancestraux contrastés ont des prévalences différentes de résistance à l'insuline et/ou des pressions de sélection différentes sur ces gènes.

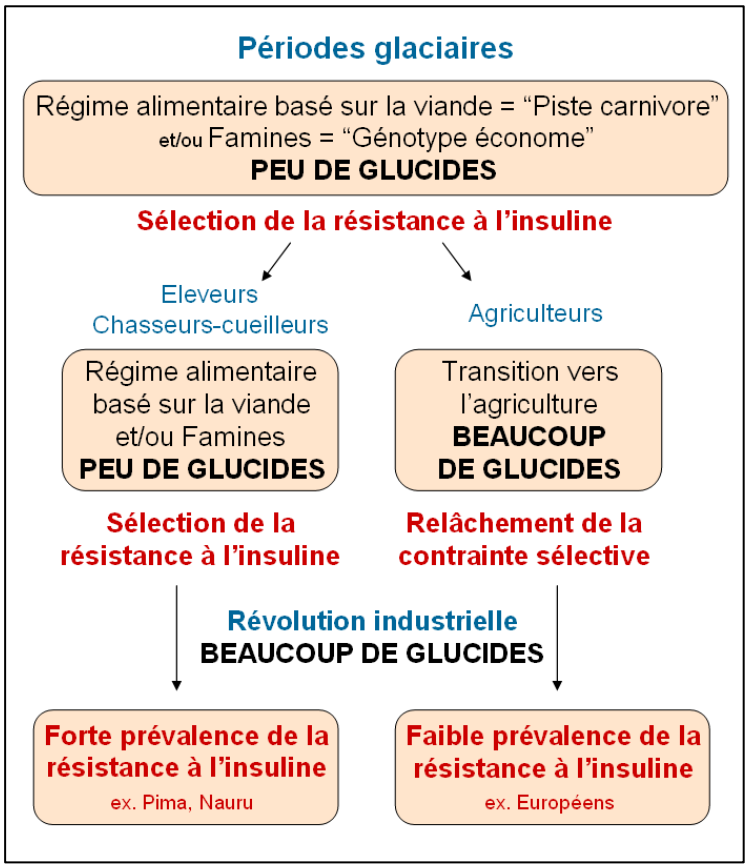


Figure 16 : Hypothèses combinées de la « piste carnivore » et du « génotype économe ». Figure inspirée de Brand-Miller & Colagiuri (1994).

Cependant, les études génétiques basées sur la comparaison entre le niveau de différenciation génétique de marqueurs *a priori* neutres et celui de marqueurs associés au diabète de type II n'ont pas permis d'infirmer ou confirmer ces hypothèses. Certaines études ont trouvé des signes de sélection positive locale sur certains gènes, sans pour autant savoir dans quelles populations ni sur quels allèles la sélection a agit (Fullerton *et al.*, 2002, Myles *et al.*, 2007, Pickrell *et al.*, 2009). D'autres études plus poussées ont également trouvé des signes de sélection positive, mais sur des allèles non associés au diabète de type II (Vander Molen *et al.*, 2005, Helgason *et al.*, 2007, Cheng *et al.*, 2009, Pickrell *et al.*, 2009). Finalement, certaines études n'ont trouvé aucun signe de sélection (Ruiz-Narvaez, 2005).

Notre objectif est donc ici de tester l'hypothèse du « génotype économe » et de la « piste carnivore », en regardant si des adaptations alimentaires différentes ont eu lieu entre des populations aux modes de subsistance différents. Nous cherchons à savoir si les éleveurs et les agriculteurs d'Asie Centrale ont différentes prévalences de résistance à l'insuline, phénotype d'intérêt car potentiellement sélectionné dans le passé, du fait de leur composante

génétique. Pour cela, nous avons effectué quatre missions sur le terrain en Ouzbékistan et au Kirghizistan, afin d'obtenir des données physiologiques nous permettant d'établir le statut de résistance à l'insuline (caractérisée par un découplage physiologique entre les niveaux de glucose et d'insuline dans le sang). Etant donné que la résistance à l'insuline est largement multifactorielle, avec une part environnementale et une part génétique, nous devons prendre en compte l'influence des co-variables environnementales sur ce phénotype, pour avoir accès uniquement aux facteurs génétiques (ici reflétés par l'ethnicité). Ainsi, nous avons effectué des enquêtes sur l'alimentation, l'activité physique et le contexte médical des individus, ainsi que des mesures anthropométriques, pour pouvoir quantifier et comparer certaines co-variables environnementales entre ethnies. Finalement, nous avons également récupéré des échantillons d'ADN, pour tester directement si certains gènes en lien avec la résistance à l'insuline présentent des signes de sélection dans ces populations, et si ces pressions sélectives sont différentes en fonction du mode de subsistance ancestral. Nous avons pour cela séquencé 20 régions *a priori* neutres du génome et 11 régions centrées sur des mutations préalablement associées au diabète de type II.

B. Description du mode de vie et de l'état de santé

1) Matériel et méthodes

a) Campagnes d'échantillonnage

Une première mission préalable de trois semaines de reconnaissance de terrain et de mise au point des questionnaires a été effectuée au Kirghizistan en mai-juin 2007 dans la vallée de Naryn (villages de Kulanak et d'Ak-Muz), ainsi qu'au bord du lac Issyk-Koul dans les villages réunis de Tamga et Barskoom (avec l'aide de Begoña Martinez-Cruz, Philippe Mennecier, Nicolas Lescureux, et de médecins et personnels hospitaliers locaux). Trois campagnes d'échantillonnage ont ensuite été effectuées pour récolter toutes les données nécessaires à notre étude, à savoir : des échantillons ADN, des échantillons sanguins congelés sous forme de plasma et de sérum (pour les dosages physiologiques), des données anthropométriques, des données sur l'alimentation, l'activité physique, le contexte médical, et enfin des données ethniques, linguistiques et démographiques sur l'individu et da famille.

Ces données ont été récoltées pour les Tadjiks (agriculteurs) sur trois semaines (octobre-novembre 2007) en Ouzbékistan : 120 individus ont été échantillonnés à Boukhara (milieu urbain, échantillon appelé Ta_Urb) et 100 individus à Muchaus (milieu rural, Ta_Rur). Pour cet échantillon, le recrutement s'est fait grâce aux médecins et infirmières locaux, qui ont demandé aux gens de participer à une étude sur le diabète de type II. Cette méthode de recrutement a favorisé la venue de gens se sachant atteint de diabète ou le suspectant, d'après des examens antérieurs ou des données familiales. Ainsi, 27% des Ta_Urb et 10% des Ta_Rur se sont déclarés diabétiques ou pensaient avoir un excès de glucose dans le sang.

L'échantillon de Kirghiz (éleveurs) a été constitué au cours de deux missions, une première de trois semaines en avril-mai 2008 à Bichkek (162 individus, milieu urbain, échantillon appelé Ki_Urb) et une deuxième de trois semaines à At-Bashy en juillet-août 2008 (194 individus, milieu rural, Ki_Rur). Notre projet a ici fait partie d'une recherche médicale plus large menée par l'équipe d'Almaz Aldashev sur la prévalence de diverses maladies (cardiaque, pulmonaire, etc.) dans la population kirghize. Ceci a permis de recruter nos échantillons avec moins de biais en faveur des diabétiques (12% de Ki_Urb et 13% de Ki_Rur). Pour corriger les différences de recrutement entre éleveurs (Kirghiz) et agriculteurs (Tadjiks), nous avons exclus de l'étude les individus se sachant diabétiques ou pensant avoir

un excès de glucose dans le sang, d'après nos questionnaires médicaux. Ainsi, sur les 576 individus échantillonnés (120 Ta_Urb, 100 Ta_Rur, 162 Ki_Urb et 194 Ki_Rur), 87 individus ont été exclus, et les analyses ont été effectuées sur les 489 individus restants : 88 Ta_Urb, 90 Ta_Rur, 143 Ki_Urb et 168 Ki_Rur.

La collecte de ces données, sous la direction d'Evelyne Heyer, est le résultat de collaborations internationales fructueuses et n'aurait pas été possible sans, entre autres, l'aide précieuse de Tatiana Hegay et son équipe en Ouzbékistan, l'équipe médicale de l'hôpital de Boukhara, Almaz Aldashev et son équipe au Kirghizistan, l'équipe médicale d'At-Bashy, ainsi que Stanislas Ashouraliev. Il va sans dire que rien n'aurait non plus été envisageable sans les autres membres de la mission (Evelyne Heyer, Patrick Pasquet, et Philippe Mennecier). Les moyens financiers ont été assurés par l'Agence Nationale de la Recherche, à travers le projet NUTGENEVOL (2007-2011).

b) Etablissement des questionnaires

Les questionnaires ethno-linguistiques ont été remplis par Philippe Mennecier, et les autres questionnaires ont été remplis par des médecins ou infirmières locaux.

Le questionnaire ethno-linguistique consiste à récolter des données sur la langue parlée, l'appartenance aux groupes de parenté (pour les populations patrilinéaires), le lieu et la date de naissance et le nombre d'enfants, tout ceci pour l'individu et sa famille (sa femme, ses parents, ses grands parents et ses beaux parents). Ce questionnaire a été rédigé à partir des modèles précédemment utilisés en Asie Centrale (Chaix *et al.*, 2004).

Le questionnaire alimentaire qualitatif appelé « questionnaire de fréquence », et présenté en annexe 2, a été mis au point à partir de modèles fournis par Patrick Pasquet précédemment utilisés sur d'autres terrains en Afrique Centrale (Said-Mohamed *et al.*, 2009). Ce questionnaire consiste à demander à chaque individu la fréquence à laquelle chaque item est consommé, selon un indice de fréquence allant de un (jamais) à six (tous les jours). Ce questionnaire permet donc de récolter des informations sur la diversité alimentaire, mais ne permet pas de connaître les quantités consommées pour chaque aliment. Pour adapter ces questionnaires à l'Asie Centrale, c'est-à-dire lister les items alimentaires consommés au Kirghizistan et en Ouzbékistan, j'ai reçu au préalable l'aide expérimentée de Nicolas

Lescureux et de Bernard Dupaigne, pour finalement obtenir une liste de 41 aliments (ou groupes d'aliments). Dans le but d'obtenir des données quantitatives, des enquêtes alimentaires de 24h, ainsi que des suivis de préparation de plats, ont également été effectués lors de la première mission, mais n'ont pas été reconduits lors des missions suivantes en raison du temps nécessaire pour mener à bien ces enquêtes. Un questionnaire demandant aux individus les quantités de nourriture achetées au marché et/ou produites par semaine pour 11 aliments majeurs nous a tout de même permis d'accéder à des données plus quantitatives (voir annexe 2).

Un autre questionnaire, mis au point avec Patrick Pasquet, a eu pour objectif d'identifier les maladies actuelles, passées et familiales des individus, ainsi que les traitements en cours, pour connaître l'état de santé des individus et mieux interpréter les résultats des dosages physiologiques. Ce questionnaire a été accompagné d'une enquête sur l'activité physique (voir annexe 2). Pour les deux dernières missions au Kirghizistan, une version modifiée du questionnaire IPAQ (questionnaire international sur l'activité physique disponible sur http://www.ipaq.ki.se/downloads.htm) a été utilisée. Ce questionnaire a été internationalement validé (dans sa version française), puis il a été traduit en russe par Philippe Mennecier.

c) Mesures anthropométriques

Les mesures anthropométriques ont été effectuées par Patrick Pasquet et moi-même. Nous avons relevé le poids (mesuré à 100gr près à l'aide d'un pèse personne électronique type Seca), la taille (au mm près à l'aide d'une toise Siber hegner), le tour de taille et de hanche (relevés au cm près à l'aide d'un maître ruban) et le pli sous-cutané tricipital (à l'aide d'une pince à plis Harpenden). Nous avons également récolté des données de bio-impédancemétrie, pour comprendre comment le poids se répartit en masse grasse et masse maigre. L'impédancemétrie est basée sur les différences de propriétés électriques de la masse grasse et de la masse maigre. Cette technique consiste à mesurer la résistivité du corps et utilise le fait que la résistance à un courant alternatif (impédance), généré à travers le corps, est liée au volume corporel (le conducteur) et au carré de la stature (longueur du conducteur). L'eau étant le milieu conducteur par excellence, l'impédancemétrie consiste, avant tout, à mesurer le volume d'eau corporelle totale. Le pourcentage de masse grasse peut ensuite être déduit du volume d'eau corporelle totale, à partir de la formule de Segal (Segal *et al.*, 1988).

d) Dosages physiologiques

Quatre échantillons sanguins ont été prélevés par individu : 5mL pour les analyses génétiques, 2mL pour le dosage du glucose sur plasma (surnageant du sang n'ayant pas coagulé grâce à la présence d'EDTA dans les tubes), et 2x5mL pour les autres dosages sur sérum (surnageant du sang ayant coagulé). Les tubes pour les dosages ont donc été centrifugés et les surnageants ont été récupérés dans des tubes de 2mL. Pour être conservés, les échantillons ont ensuite été stockés sous forme congelée dans des bonbonnes à azote liquide, à une température de -160°C. Une fois en France, les échantillons ont été dosés à l'hôpital Saint Vincent de Paul par l'équipe d'Hormonologie Pédiatrique et Maladies Métaboliques, avec la collaboration de Najiba Lalhou et Marc Roger. Nous avons ainsi obtenu les résultats de glycémie et d'insulinémie, nous permettant de calculer la résistance à l'insuline : (Insuline x Glycémie / 22.5), avec les individus ayant ce rapport supérieur ou égal à 2.6 définis comme « résistants à l'insuline ». Nous avons également effectué des dosages de leptine, d'HDL-cholestérol, de cholestérol total, ainsi que de triglycérides, notamment pour pouvoir obtenir le phénotype du syndrome métabolique (présence chez un individu, d'un ensemble de caractéristiques physiologiques) mais nous ne détaillerons pas ces résultats ici.

2) Données sur l'alimentation

a) Questionnaire alimentaire de fréquence

Parmi les 489 individus de l'étude, 446 individus ont rempli des questionnaires de fréquence (86 Ta_Urb, 69 Ta_Rur, 142 Ki_Urb et 149 Ki_Rur). Ces données ont ensuite été transformées en données de fréquence par mois et les items (dont certains sont groupés ensemble) ont été classés dans 5 catégories, afin de distinguer des tendances générales :

1) les **amylacées** : riz, pâtes, pain, céréales et pommes de terre

2) les **protéines animales** : mouton, chèvre, cheval, lapin, bœuf, dinde, poulet, porc, œuf, et poisson

3) les **produits laitiers** : lait frais, crème de lait, kéfir (boisson issue de la fermentation du lait et légèrement alcoolisée), fromage, fromage blanc et beurre

4) les **fruits et légumes** : fruits secs, fruits de saison, amandes, noix, arachides, oignons, tomates / concombres / salades, épinards, betteraves / navets / choux, aubergines / poivrons, radis, carottes, haricots / pois chiches

5) les **sucreries et boissons** : limonade, autres boissons pétillantes sucrées, bozo / jarma (boissons locales alcoolisées obtenues à partir de fermentation de céréales ou d'orge, respectivement), bière, vodka, gâteau, confiture, miel, bonbon et sucre.

Les résultats, pour chaque item séparément et pour les grandes catégories, ainsi que l'analyse de la comparaison entre ethnies (Tadjiks vs Kirghiz) et entre milieux (Urbains vs Ruraux) sont présentés ci-dessous. Les items alimentaires ne présentant aucune différence significative entre ethnies ou entre milieux (pâtes, pain, lapin, arachide, bière et miel) ne sont pas présentés dans ce tableau.

	Moyenne de fréquence par mois (écart-type)				Comparaison inter ethnie		Comparaison inter milieu	
	Ta_Urb	Ta_Rur	Ki_Urb	Ki_Rur	RT/K*	*p*-val [1]	RU/R*	*p*-val [1]
Amylacées	**14.9** (2.7)	**14.8** (1.4)	**13.2** (3.2)	**12.3** (3.3)	**1.2**	**4E-14**	**1.1**	**0.002**
Riz	8.3	6.6	6.6	4.9	1.3	1E-06	1.3	1E-05
Céréales	3.8	1.0	5.4	3.0	0.6	3E-07	2.0	1E-11
Pommes de terre	27.3	29.8	17.1	16.2	1.7	9E-16	1.0	
Protéines Animales	**5.7** (1.6)	**4.7** (1.5)	**4.0** (1.7)	**2.7** (1.7)	**1.6**	**9E-16**	**1.4**	**4E-11**
Mouton	2.3	1.2	6.7	9.9	0.2	9E-16	0.7	6E-07
Chèvre	0.0	1.0	0.4	1.9	0.4		0.2	2E-15
Cheval	0.2	0.0	1.4	1.4	0.1	9E-16	0.9	
Bœuf	30.0	27.2	17.9	3.5	2.7	9E-16	2.0	2E-15
Dinde	0.6	0.0	0.5	0.2	1.0	0.003	3.6	7E-05
Poulet	4.5	2.8	4.1	2.3	1.2	0.003	1.7	6E-07
Porc	1.0	0.0	0.1	0.0	15.5	0.02	23.7	3E-04
Œuf	15.7	13.0	5.9	6.6	2.3	9E-16	1.1	
Poisson	3.1	1.5	2.1	1.5	1.3	2E-04	1.7	1E-06
Fruits & Légumes	**11.5** (3.0)	**12.0** (5.2)	**9.3** (4.3)	**6.1** (2.7)	**1.6**	**9E-16**	**1.3**	**3E-08**
Fruits secs	9.7	5.8	16.6	9.0	0.6	3E-07	1.7	9E-05
Fruits de saison	19.3	23.8	6.5	1.8	5.2	9E-16	1.3	6E-08
Amandes	4.2	2.0	2.2	0.5	2.5	8E-09	3.1	6E-08
Noix	4.3	2.7	4.2	1.6	1.3		2.2	0.001
Oignons	30.0	29.7	16.9	11.4	2.1	9E-16	1.3	2E-04
Tomates / Concombres / Salades	24.1	19.4	0.7	0.2	48.1	9E-16	1.6	2E-04
Epinards	8.4	4.2	6.6	4.7	1.2	1E-06	1.6	
Betteraves / Navets / Choux	6.5	12.0	7.6	4.7	1.5	2E-09	1.0	0.03
Aubergines / Poivrons	6.5	14.7	9.4	4.3	1.6	4E-06	1.1	0.01
Radis	6.4	12.1	18.9	13.0	0.6	2E-09	1.1	
Carottes	23.9	15.4	1.2	0.2	28.7	9E-16	1.9	8E-10
Haricots / Pois chiches	4.9	10.7	25.9	26.7	0.3	9E-16	0.8	0.002
Produits Laitiers	**8.7** (2.7)	**11.6** (1.4)	**9.1** (3.2)	**14.0** (3.3)	**0.9**	**0.01**	**0.7**	8E-14
Lait frais	12.5	20.5	10.7	25.3	0.9		0.5	2E-15
Crème de lait	5.2	12.2	3.7	18.7	0.7		0.3	2E-15
Kéfir	10.7	17.6	10.2	19.2	0.9		0.6	5E-12
Fromage	7.0	2.1	6.5	3.7	1.0		2.1	1E-08
Fromage blanc	3.5	4.3	4.0	5.7	0.8	0.04	0.7	0.022
Beurre	20.4	11.6	19.2	11.1	1.1		1.7	2E-11

	Moyenne de fréquence par mois (écart-type)				Comparaison inter ethnie		Comparaison inter milieu	
	Ta_Urb	Ta_Rur	Ki_Urb	Ki_Rur	$R_{T/K}$*	p [1]	$R_{U/R}$*	p [1]
Sucreries & boissons	**10.0** (3.9)	**10.2** (5.1)	**8.0** (3.4)	**7.6** (3.3)	**1.3**	**3E-07**	**1.0**	
Limonade	2.2	7.3	4.3	3.2	1.2	0.004	0.8	4E-04
Boisson pétillante sucrée	6.9	10.2	2.1	2.3	3.8	9E-16	0.8	0.01
Bozo / Jarma	1.7	1.5	2.4	1.1	0.9		1.7	2E-05
Vodka	2.5	1.1	1.4	0.9	1.6		2.0	0.003
Gâteau	11.3	9.8	1.4	20.9	0.9	0.03	0.3	2E-15
Confiture	19.7	18.4	20.0	7.7	1.4	7E-05	1.8	4E-11
Bonbon	20.7	18.9	12.8	4.2	2.4	9E-16	1.8	2E-09
Sucre	21.6	19.7	22.3	26.1	0.9	0.003	0.9	

*$R_{A/B}$: Rapport de la moyenne A/B (avec T : Tadjiks, K : Kirghiz, U : Urbains et R : Ruraux) ;
[1] p : p-valeurs (test de Wilcoxon). Seules les valeurs significatives sont reportées ici.

Tableau 10 : Résultat des questionnaires alimentaires de fréquence et comparaisons entre ethnies et entre milieux. Les rapports de moyenne entre groupes ($R_{A/B}$) significativement supérieurs ou égaux à 2 sont écrits en rouge.

Parmi les dix items les plus fréquemment consommés, quatre sont communs à toutes les populations : le pain, les oignons, les pommes de terre et le sucre. La catégorie la plus fréquemment consommée est celle des amylacées, sauf pour les Kirghiz en campagne (Ki_Rur) où ce sont les produits laitiers. Pour les quatre échantillons, la catégorie la moins fréquemment consommée correspond à celle des protéines animales.

Différences entre ethnies

En ce qui concerne les grandes catégories, les Tadjiks consomment significativement plus souvent des amylacées ($p = 4.10^{-14}$), des protéines animales ($p = 9.10^{-16}$), des fruits et légumes ($p = 9.10^{-16}$) et des sucreries et boissons ($p = 3.10^{-7}$). D'un autre côté, les Kirghiz consomment significativement plus souvent des produits laitiers ($p = 0.01$).

Différences entre milieux

Les urbains consomment significativement plus souvent des amylacées ($p = 0.002$), des protéines animales ($p = 4.10^{-11}$) et des fruits et légumes ($p = 3.10^{-8}$), tandis que les ruraux consomment plus souvent des produits laitiers ($p = 8.10^{-14}$). Il n'y a pas de différence significative pour la catégorie des sucreries et boissons.

De manière générale, à part pour quelques aliments décrits plus hauts, les écarts de moyennes, qu'ils soient entre ethnies ou entre milieux ne sont pas très importants : pour les grandes catégories, RT/K (rapport de la moyenne entre Tadjiks et Kirghiz) varie entre 0.9 et 1.6 et RU/R (rapport de la moyenne entre urbains et ruraux) varie entre 0.7 et 1.4. Ainsi, d'après les questionnaires de fréquence, nous n'observons pas un profil de consommation très différent entre populations.

b) Questionnaire alimentaire quantitatif

Sur les 489 individus de l'étude, nous avons récupéré des données quantitatives pour 441 individus (59 Ta_Urb, 90 Ta_Rur, 141 Ki_Urb et 151 Ki_Rur). Il s'agissait de demander aux individus d'indiquer la quantité de différents aliments achetés au marché ou produits (farine, riz, pommes de terre, pâtes, viande, oeuf, lait, sucre et huile), et de préciser le nombre de personnes consommatrices dans le foyer. Ces données sont plus précises que les questionnaires qualitatifs précédents, puisqu'il s'agit de quantités réellement consommées. Les données sont ici présentées en litre d'huile achetés, en kilogramme de viande / poisson / sucre / farine / riz / pommes de terre / pâtes achetés, en litre de lait achetés et/ou produits et en nombre d'œufs achetés, par semaine et par personne (voir tableau 11 ci-dessous). Comme précédemment, les items ne présentant pas de différences significatives entre catégories ne sont pas présentés.

	Moyenne quantitative par semaine et par personne*				Comparaison entre ethnie		Comparaison entre milieu	
	Ta_Urb	**Ta_Rur**	**Ki_Urb**	**Ki_Rur**	RT/K [1]	*p*-val [2]	RU/R*	*p*-val [2]
Amylacées	**3.23** (1.5)	**6.18** (2.8)	**3.27** (1.9)	**5.86** (2.1)	**1.1**		**0.6**	**7E-16**
Farine	1.02	3.87	1.07	3.28	1.2		0.3	7E-16
Riz	0.57	0.69	0.60	0.48	1.2	0.002	1.1	
Pâtes	0.27	0.42	0.48	0.72	0.6	5E-08	0.7	1E-06
Protéines animales [3]	**1.04** (0.5)	**0.63** (0.4)	**1.24** (0.8)	**0.99** (0.6)	**0.7**	**1E-08**	**1.4**	**4E-09**
Viande	0.76	0.42	0.98	0.78	0.6	5E-12	1.4	2E-07
Œufs	4.45	3.06	4.14	2.38	1.1	0.03	1.6	4E-09
Lait	2.17	1.84	1.31	5.05	0.6	0.02	0.4	5E-07
Sucre	0.27	0.39	0.30	0.57	0.8	0.002	0.6	7E-16
Huile	0.29	0.44	0.26	0.30	1.4	7E-08	0.8	1E-06

* Données en litre (pour l'huile et le lait), en unités (pour les œufs) ou en kilogramme (pour les autres items) ; [1] RA/B : Ratio de la moyenne A/B (avec T : Tadjiks, K : Kirghiz, U : Urbains et R : Ruraux) ; [2] *p* significatives du test de Wilcoxon ; [3] Pour cette moyenne, les données sur les œufs ont été converties en poids (65g par œuf).

Tableau 11 : Résultat des questionnaires alimentaires quantitatifs et comparaisons entre ethnies et entre milieux. Les rapports de moyenne entre groupes (RA/B) significativement supérieurs ou égaux à 2 sont écrits en rouge.

Le nombre moyen d'individus par foyer n'est pas significativement différent entre les quatre populations (entre 4.1 et 4.4).

Différences entre ethnies

La consommation d'amylacées n'est pas significativement différente entre Kirghiz et Tadjiks. Par contre, les Kirghiz consomment significativement plus de protéines animales ($p = 1.10^{-8}$), et en fait surtout 1.7 fois plus de viande ($p = 5.10^{-12}$). Ce résultat va à l'inverse du questionnaire précédent où nous trouvions que les Tadjiks consomment 1.6 fois plus souvent de protéines animales. Cette différence entre questionnaires pourrait refléter une mauvaise estimation de la consommation dans l'un ou l'autre des questionnaires, mais elle pourrait également refléter des différences dans le mode de consommation des populations. Ainsi, les Kirghiz pourraient manger moins fréquemment mais en quantité plus importante de la viande, à travers, par exemple, la consommation de plats exclusivement à base de viande comme le « besh-barmak », plat national du Kirghizistan. Nous voyons également que les Kirghiz ont

une consommation significativement plus importante de lait ($p = 0.02$), en cohérence avec le questionnaire de fréquence, ainsi que plus de sucre ($p = 0.002$), surtout en campagne. Les Tadjiks, eux, consomment significativement plus d'huile ($p = 7.10^{-8}$).

Différences entre milieux

Les ruraux consomment significativement plus d'amylacées ($p = 7.10^{-16}$), de lait, de sucre et d'huile. A l'inverse, les urbains consomment significativement plus de viande ($p = 2.10^{-7}$).

Chez les Kirghiz, l'alimentation moins riche en huile et plus riche en lait et en viande fait bien penser à une alimentation traditionnellement d'éleveurs, tandis que celle des Tadjiks (plus riche en huile et moins riche en lait et en viande), est plus proche d'une alimentation d'agriculteurs. Cependant, nous voyons que la soviétisation a amené en grande quantité des amylacées dans la nourriture kirghize, notamment des pâtes et du sucre, aliments *a priori* non traditionnels.

c) Conclusion

En conclusion, pour les données alimentaires, nos questionnaires ne mettent pas en évidence de différences importantes qualitatives entre échantillons. Nous voyons cependant une tendance vers plus de diversité alimentaire chez les Tadjiks et en ville, sauf pour les produits laitiers, où la tendance est inverse. D'un point de vue quantitatif, différents types de régimes alimentaires peuvent être identifiés, avec plus de viande et de lait chez les Kirghiz, et plus d'huile chez les Tadjiks. En milieu urbain, nous voyons une consommation accrue de viande, mais les autres aliments (amylacées, lait, sucre et huile) sont trouvés en plus grande quantité en campagne. Une investigation plus poussée d'anthropologie nutritionnelle permettrait certainement de mieux apprécier ces écarts entre populations, mais l'objectif de cette étude est simplement de déterminer si les régimes alimentaires sont dans l'ensemble différents entre ethnies, afin de comprendre si les prévalences de résistance à l'insuline dans ces populations peuvent être directement liées à des différences d'alimentation.

3) Données sur l'activité physique

a) Questionnaire d'auto-estimation

Nous avons récolté des données sur l'activité physique, d'après un questionnaire d'auto-estimation (voir annexe 2), pour 460 individus (86 Ta_Urb, 90 Ta_Rur, 143 Ki_Urb et 141 Ki_Rur) (voir figure 17 ci-dessous).

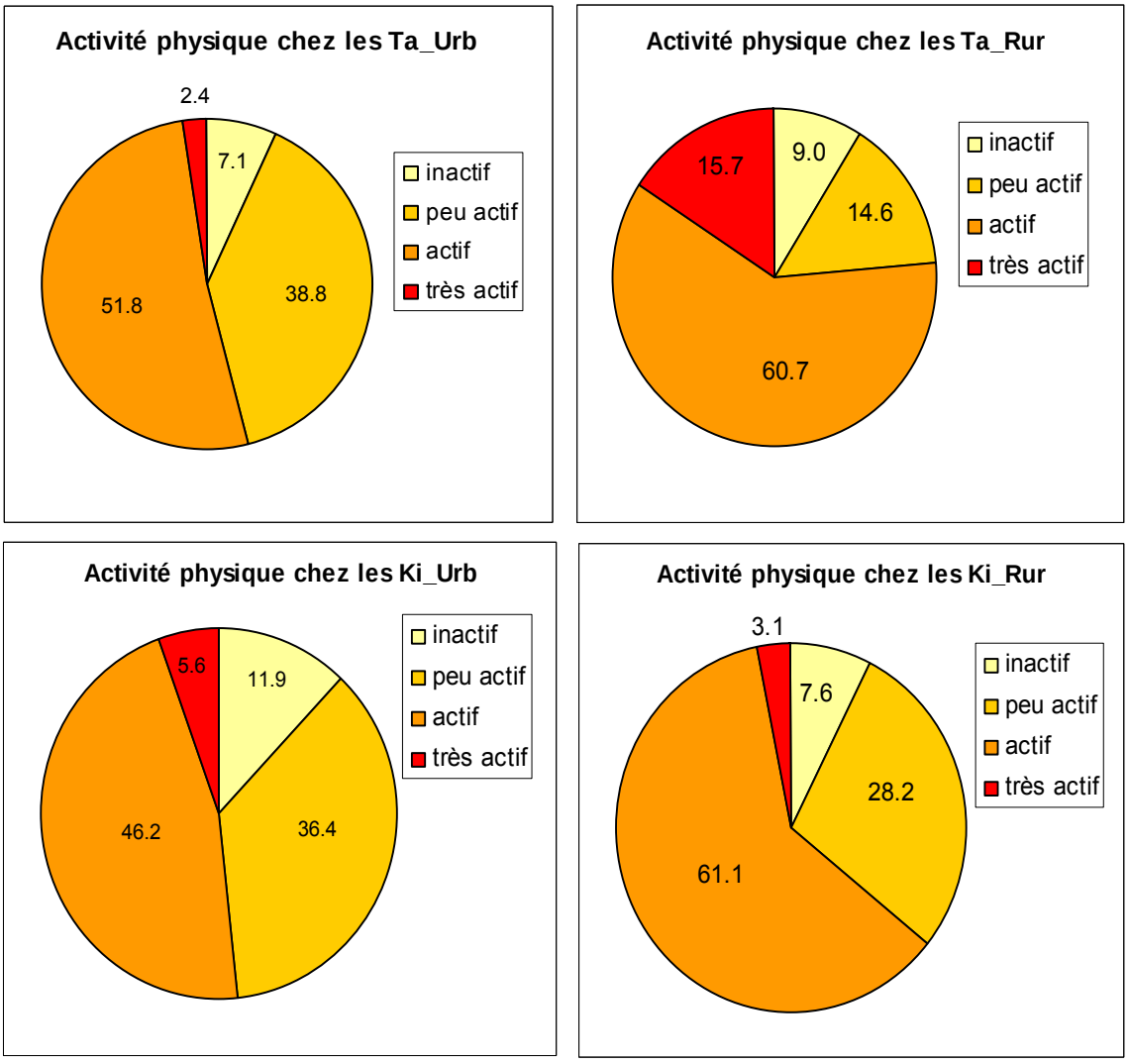


Figure 17 : Activité physique des populations échantillonnées d'après le questionnaire d'auto-estimation

Chez les Tadjiks en ville, nous observé une majorité d'individus « peu actifs » (38.8%) et « actifs » (51.8%), avec seulement 2.4% de « très actifs », tandis qu'en campagne, 15.7% s'estiment « très actifs ». Les indices moyens (2.49 et 2.83 chez les Ta_Urb et les Ta_Rur,

respectivement) sont significativement différents ($p = 2.10^{-3}$), malgré un âge moyen plus élevé chez les Ta_Rur que chez que les Ta_Urb (48.0 vs 40.8 ans).

Chez les Kirghiz, l'activité physique n'est pas significativement différente entre ville et campagne (indice moyen de 2.45 et 2.60, respectivement, et $p = 0.12$) et leur activité physique n'est pas significativement différente de celle des Tadjiks en ville ($p = 0.72$ et 0.24, respectivement). Le seul écart significatif d'activité physique entre ces populations est donc celui entre les Ta_Rur, particulièrement actifs, et les trois autres ($p = 2.10^{-3}$, 9.10^{-4} et 0.01, respectivement pour Ta_Urb, Ki_Urb et Ki_Rur). Nous pouvons noter que, bien que les Ta_Urb ont la même activité physique que les deux autres populations kirghizes, ils ont une moyenne d'âge de 10 ans plus jeune.

Si l'on considère séparément les hommes et les femmes, nous trouvons que chez les Tadjiks, la forte différence d'activité physique entre ville et campagne est majoritairement due à un écart entre les femmes (2.87 vs 2.38), par rapport à l'écart entre hommes (2.76 vs 2.65). Chez les Kirghiz, la différence (plus faible) entre ville et campagne est mieux expliquée par les écarts entre hommes (2.61 vs 2.40) qu'entre femmes (2.58 vs 2.51).

b) Métiers exercés

La plus forte dépense physique des Ta_Rur par rapport aux autres populations peut s'expliquer par leur mode de vie plus traditionnel, comprenant des travaux liés à l'agriculture qui constituent une importante dépense énergétique. En effet, d'après les données sur les professions exercées, 28.6% des Ta_Rur sont agriculteurs contre 2.4% chez les Ta_Urb, 1.4% chez les Ki_Urb (en cumulant également les éleveurs) et 8.8% chez les Ki_Rur (en cumulant également les éleveurs). Les Ta_Rur exercent donc une activité professionnelle plus physique.

c) Mode de transport et activité sportive

Nous avons également obtenu grâce au questionnaire sur l'activité physique le pourcentage d'individus effectuant une dépense énergétique pour aller au travail : en vélo ou à pied (ou à cheval chez les Kirghiz), par rapport à ceux qui prennent la voiture, le bus, ou ne se déplacent pas pour leur travail. Ainsi, 33.8% des Ta_Urb effectuent une activité physique pour aller au travail, par rapport à 91.8% pour les Ta_Rur. Chez les Kirghiz, les urbains sont

23.2% à se dépenser physiquement pour ce trajet, pour 50.4% en milieu rural. Nous retrouvons bien encore un écart entre ville et campagne, avec surtout un écart important entre les Ta_Rur et les trois autres populations. Le pourcentage d'individus pratiquant une activité sportive est de 24.7% chez les Ta_Urb, 0% chez les Ta_Rur (avec seulement environ un quart des individus ayant répondu à la question), 18.6% chez les Ki_Urb et 5.2% chez les Ki_Rur. L'activité sportive « ludique » est donc plus importante en milieu urbain.

d) Questionnaires IPAQ

Nous avons également fait passer des questionnaires IPAQ (*International Physical Activity Questionnaire*) que nous avons adaptés, mais seulement lors des deux dernières missions, pour les Ki_Urb (126 individus) et les Ki_Rur (108 individus). Il consiste en plusieurs questions simples sur le temps passé à différents types d'efforts (nombre d'heures par semaine d'activité intense et intermédiaire, de marche et de temps passé assis), dont les résultats sont ensuite transformés en un indice d'activité physique, faible (1), intermédiaire (2) ou fort (3) selon les critères définis par le protocole de l'IPAQ (également disponible sur le site http://www.ipaq.ki.se/downloads.htm). Les résultats sont présentés ci-dessous :

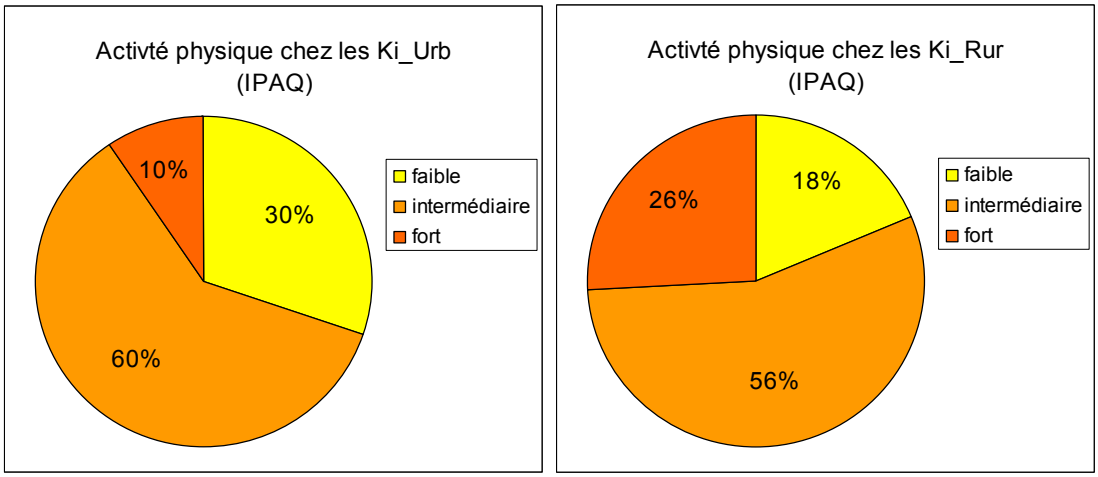


Figure 18 : Activité physique des Kirghiz selon le questionnaire IPAQ

Nous voyons ici une différence plus marquée entre campagne et ville chez les Kirghiz que trouvée précédemment avec les indices d'auto-estimation, avec 26% des Kirghiz en campagne effectuant une activité physique intense, par rapport à 10% en ville. Les indices moyens sont ici significativement différents (1.79 vs 2.07, $p = 0.001$).

J'ai comparé la quantité exacte d'activité physique obtenue d'après le questionnaire IPAQ (information brute avant la transformation en indice) avec les indices d'auto-estimation de la première enquête. Les résultats sont présentés ci-dessous :

Auto-estimation	**Nombre d'heure / jour de sport intense et intermédiaire, ou de marche (selon l'IPAQ)**	
Non actif	Ki_Urb	1.0
	Ki_Rur	2.0
Peu actif	Ki_Urb	2.2
	Ki_Rur	2.9
Actif	Ki_Urb	3.5
	Ki_Rur	6.3
Très actif	Ki_Urb	8.5
	Ki_Rur	10.2

Tableau 12: Quantité de sport ou de marche estimée avec le questionnaire IPAQ, en moyenne pour chaque indice d'auto-estimation, chez les Kirghiz

Ainsi, nous voyons que quand un individu urbain s'estime actif, il effectue en moyenne 3.5 heures de sport par jour, tandis qu'un individu rural s'estimant actif effectue 6.3 heures d'activité physique par jour. Ces résultats nous permettent de penser que la différence entre ville et campagne trouvée dans notre première étude d'auto-estimation est largement sous-estimée. La différence entre les deux enquêtes d'activité physique (figure 17 et 18) s'explique aisément par le fait que l'indice d'auto-estimation correspond à une perception de l'individu, donc certainement par rapport à une référence uniquement propre à sa population. Ainsi, si une population fait plus d'activité physique qu'une autre, il est possible que ce résultat soit sous-estimé par les indices d'auto-estimation, mais mieux révélé par l'enquête IPAQ plus objective.

En résumé, l'activité physique est plus importante en campagne qu'en ville dans les deux populations, avec cependant un contraste plus marqué chez les Tadjiks. Ces écarts sont certainement liés aux travaux physiques intenses liés à l'agriculture en milieu rural. De plus, en ville, les urbains se déplacent majoritairement en voiture, ou en bus, et certains compensent la perte d'activité physique par la pratique d'activités sportives, quasi inexistante en campagne.

4) Données anthropométriques

a) Description de l'échantillon

Nous avons récolté des données anthropométriques pour 484 individus. Les moyennes d'âge dans chaque population et par sexe sont présentées ci-dessous. Toutes les données entre parenthèses, ici et dans la suite, correspondent aux écarts-types. De même, toutes les comparaisons entre groupes se basent sur des tests de Wilcoxon avec des *p*-valeurs corrigées pour les tests multiples d'après la correction FDR (pour *False Discovery Rate*).

	Femmes		**Hommes**		Moyenne d'age (hommes et femmes)
	Proportion dans l'échantillon	**Age moyen**	**Proportion dans l'échantillon**	**Age moyen**	
Ta_Urb	56%	40.1 (11.9)	44%	41.7 (10.1)	40.8 (12.0)
Ta_Rur	67%	45.1 (10.6)	33%	53.7 (14.0)	48.0 (12.9)
Ki_Urb	52%	49.5 (8.0)	48%	51.3 (8.3)	50.3 (8.2)
Ki_Rur	60%	48.9 (10.4)	40%	54.3 (10.6)	51.0 (10.7)

Tableau 13 : Répartition des individus échantillonnés selon l'âge et le sexe

Pour les Kirghiz, l'âge n'est pas significativement différent entre échantillons (*p* compris entre 0.22 et 0.32), sauf pour la comparaison entre les hommes et femmes en milieu rural ($p = 0.004$). Pour les Tadjiks, au contraire, tous les échantillons sont significativement différents (*p* compris entre 0.001 et 0.005), sauf pour la comparaison entre les hommes et femmes en milieu urbain ($p = 0.37$). Les femmes tadjikes urbaines sont de cinq ans plus jeunes que les rurales en moyenne, et les hommes urbains de douze ans plus jeunes que les ruraux en moyenne. Avant le recrutement, nous avons défini des limites d'âge pour les participants à l'étude comprises entre 35 et 70 ans, mais nous n'avons pas pu contrôler au fur et à mesure les âges des individus venant pour notre étude. Ces différences sont importantes à considérer pour l'interprétation des résultats. De plus, l'échantillon tadjik a un effectif deux fois plus faible que l'échantillon kirghiz, ce qui peut diminuer la puissance de nos tests. Ainsi, surtout pour les estimateurs de santé, les comparaisons entre environnement urbain et rural chez les Tadjiks, que ce soit pour les hommes ou les femmes, ne seront que peu discutées. Une dernière mission d'échantillonnage de Tadjiks en Ouzbékistan, initialement prévue en avril 2009, a dû être reportée à cause de changements législatifs sur l'exportation de matériel biologique. Cette dernière mission aura finalement lieu en février 2010 et aura pour but d'augmenter les effectifs et de rétablir ces écarts d'âge dans nos échantillons.

b) Variables de poids

Nous avons enlevé 700g au poids des Ta_Urb, Ta_Rur et Ki_Urb et 500g au poids des Ki_Rur, ce qui correspond à l'estimation du poids des vêtements portés. De la même manière, 1 cm a été retiré au tour de taille et 2 cm au tour de hanche, pour compenser le port des vêtements lors des mesures. Les mesures anthropométriques par sexe et par population sont présentées dans le tableau ci-dessous.

		Poids*	Taille	Tour Taille*	Tour hanche*	Pli	Pgras [1]	IMC [2]
Ta_Urb	**F**	69.8	159.7	87	105.9	25	36.7	27.4
Ta_Rur		68.1	156.5	87.5	103.3	23	37.3	27.8
Ta_Urb	**H**	80	170.3	96	104.5	12	29.5	27.6
Ta_Rur		72.7	164.3	93.4	100.5	10	30.0	27.2
Ki_Urb	**F**	70.2	157.4	88.1	104.2	24	38.3	28.3
Ki_Rur		65.3	154.5	86.7	101.7	20	37.2	27.3
Ki_Urb	**H**	81.6	170.6	98.2	103.8	12	29.4	28
Ki_Rur		69.1	167.2	89	97.3	8.6	27.5	24.7

* Valeur corrigée pour le port de vêtement ; [1] Pgras : pourcentage de masse grasse d'après la bio-impédancemétrie ; [2] IMC : Indice de masse corporelle (Poids / Taille ²)

Tableau 14 : Mesures anthropométriques par population et par sexe

Chez les Tadjiks, il n'y a aucune différence significative entre ville et campagne chez les femmes, et seul le poids ($p = 0.04$) et le tour de hanche ($p = 0.048$) sont significativement différents chez les hommes entre milieu rural et urbain. Chez les Kirghiz, le poids ($p = 0.049$) et le pli ($p = 0.01$) sont significativement différents entre ville et campagne pour les femmes, et toutes les variables de poids sont significativement différentes pour les hommes (poids : $p = 5.10^{-7}$; tour de taille : $p = 5.10^{-6}$; tour de hanche : $p = 5.10^{-6}$; pli : $p = 1.10^{-5}$; pgras : $p = 0.01$; IMC : $p = 5.10^{-6}$), avec des moyennes plus élevées en ville.

Le fait que les Kirghiz, et non les Tadjiks, présentent de forts écarts de poids entre milieu rural et urbain peut paraître étonnant puisque les niveaux d'activité physique, influençant fortement les variables de poids, présentent plus de différences chez les Tadjiks que chez les Kirghiz. Plusieurs interprétations, non exclusives, peuvent être formulées : 1) la

différence de poids entre milieu rural et urbain chez les Kirghiz n'est pas majoritairement liée aux changements d'activité physique mais plutôt à d'autres facteurs, dont éventuellement des différences d'alimentation non mises en évidence ici ; 2) les Kirghiz sont beaucoup plus sensibles à des changements d'activité physique que les Tadjiks ; 3) l'âge en moyenne plus élevé chez les Tadjiks en campagne masque les différences ; 4) la plus petite taille d'échantillon chez les Tadjiks diminue la puissance de nos tests.

Les facteurs de risque liés aux variables de poids peuvent être définis à partir de l'indice de masse corporelle (IMC), qui est le poids en kg sur le carré de la taille en m. Si cet indice est supérieur ou égal à 25, l'individu est considéré en surpoids (par opposition aux normo-pondéraux dont l'IMC est compris entre 18 et 24.9). Si le poids est spécifiquement entre 25 et 29.9, l'individu est considéré comme pré-obèse. Si le poids est au-dessus de 30, l'individu est considéré comme obèse, avec une obésité de type I entre 30 et 34.9 ; de type II entre 35 et 39.9 ; et de type III au-delà de 40. Les résultats pour ces facteurs de risque sont présentés dans le tableau 15 ci-dessous :

		Pré-obésité	**Obésité Type I**	**Obésité Type II**	**Obésité Type III**	**TOTAL (Surpoids)**
Ta_Urb	F	31.3%	12.5%	14.6%	2.1%	60.4%
Ta_Rur	F	31.7%	21.7%	11.7%	1.7%	66.7%
Ta_Urb	H	46.2%	20.5%	2.6%	0.0%	69.2%
Ta_Rur	H	43.3%	6.7%	3.3%	3.3%	56.7%
Ki_Urb	F	36.0%	20.0%	12.0%	2.7%	70.7%
Ki_Rur	F	40.0%	20.0%	4.0%	2.0%	66.0%
Ki_Urb	H	50.0%	25.0%	2.9%	0.0%	77.9%
Ki_Rur	H	45.3%	4.7%	0.0%	0.0%	50.0%

Tableau 15 : Facteurs de risques liés au poids pour chaque population, et par sexe

Nous voyons que ces populations sont en général fortement concernées par des problèmes de poids, avec entre 56.7% et 77.9% de chaque échantillon ayant du surpoids. Les hommes ont en moyenne plus de surpoids que les femmes en milieu urbain, mais les femmes en ont plus en milieu rural. De plus, les femmes sont en général plus touchées par de l'obésité type II et III, tandis que les hommes sont plus touchés par de la pré-obésité.

Les facteurs de risque peuvent également être définis par le tour de taille seul, ou le rapport entre tour de taille et tour de hanche (on parle alors d'obésité centrale). Pour le tour de taille, nous avons vu que seuls les hommes kirghiz ont une différence significative pour cette variable entre milieu urbain et rural ($p = 5.10^{-6}$). Concernant le rapport entre tour de taille et tour de hanche, aucune différence significative n'a pu être mise en évidence.

c) Taille

Nous trouvons une différence de taille entre ville et campagne, avec des moyennes plus élevées en ville pour tous les échantillons : pour les femmes, augmentation de 3.2cm chez les Tadjiks (bien que seulement marginalement significative $p = 0.06$), et de 2.9cm chez les Kirghiz ($p = 0.002$) ; pour les hommes, augmentation de 6 cm chez les Tadjiks ($p = 0.01$) et de 3.4 cm chez les Kirghiz ($p = 0.001$). Cet écart de taille peut s'expliquer par de meilleures conditions de vie en ville : meilleure alimentation, moins de contraintes infectieuses, ou plus d'accès aux soins.

d) Hypertension

Comme les variables de poids, la pression artérielle n'est significativement différente qu'entre ville et campagne chez les hommes kirghiz ($p = 0.002$ et 0.0002 pour les pressions artérielles systoliques et diastoliques, respectivement).

	Pression artérielle systolique	Pression artérielle diastolique	% hyper-tension
TAB	134.6	82.8	36.4
TAM	142.4	85.4	46.7
KIB	137.6	88.4	51.7
KIM	130.4	83.1	31.5

Tableau II-E16 : Pression artérielle selon les populations et le sexe

5) Données physiologiques obtenues à partir des dosages sanguins

Nous avons obtenu des données physiologiques pour 450 individus. Les résultats de glycémie et d'insulinémie, ainsi que les valeurs de résistance à l'insuline (RI), sont présentés dans le tableau 17 ci-dessous.

		Glycémie mmol/l	Insuline µIU/ml	RI^1	RI^2
Ta_Urb	F	5.43	7.3	1.82	16.7%
Ta_Rur		5.51	6.0	1.48	8.9%
Ta_Urb	H	5.53	9.2	2.42	26.5%
Ta_Rur		5.60	5.5	1.41	14.3%
Ki_Urb	F	5.46	8.6	2.15	25.3%
Ki_Rur		5.50	9.4	2.38	27.9%
Ki_Urb	H	5.63	9.6	2.55	31.3%
Ki_Rur		5.59	7.8	1.99	23.3%
RI^1 Résistance à l'insuline = (insuline * glucose) / 22.5 RI^2: % d'individus résistants à l'insuline (avec $IR^1 > 2.6$)					

Tableau 17 : Résultats des dosages physiologiques et de la résistance à l'insuline

L'insuline des hommes kirghiz et tadjiks est significativement plus importante en milieu urbain qu'en milieu rural (p = 0.01 chez les deux ethnies). Aucune autre différence significative n'a été trouvée pour les dosages de glycémie et d'insuline.

Au niveau des prévalences de résistance à l'insuline (voir également figure 19 ci-dessous), nous voyons que les individus en milieu urbain sont plus résistants à l'insuline qu'en milieu rural (sauf les femmes kirghizes), mais les seules différences significatives sont pour les hommes (p = 0.03 pour les Kirghiz et 0.02 pour les Tadjiks). Ces écarts sont certainement dus à la baisse d'activité physique en milieu urbain, comme montré précédemment, ou bien à une dégradation de l'hygiène alimentaire (non mise en évidence ici). Les Kirghiz ont également en moyenne des prévalences de résistance à l'insuline plus fortes que les Tadjiks, bien que seulement la comparaison des femmes en milieu rural soit significative (p = 0.001). Ces écarts peuvent être dus à la composante génétique de chaque ethnie.

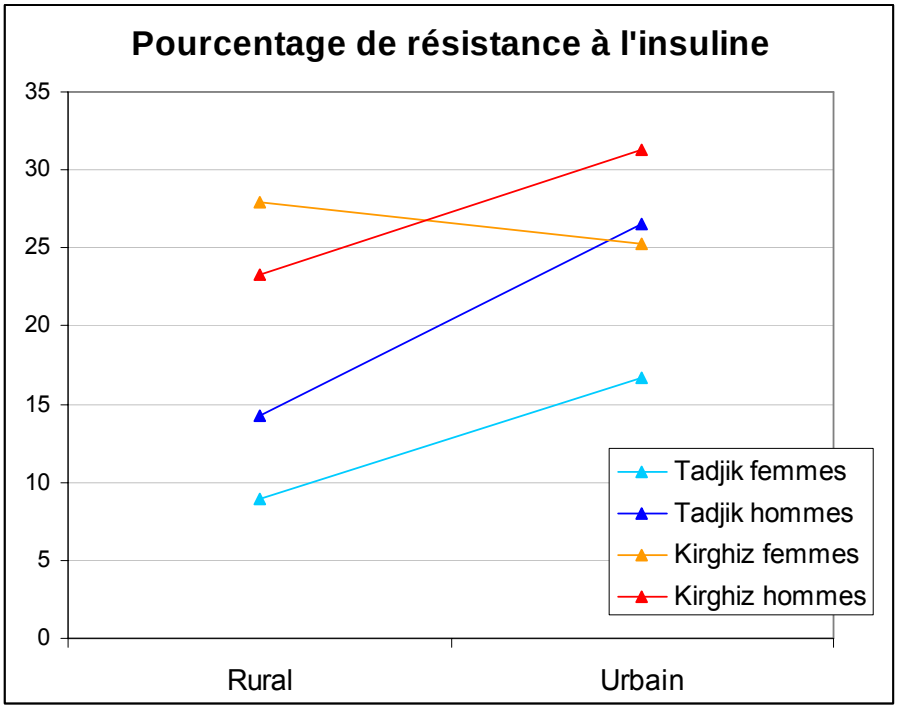


Figure 19 : Pourcentage de résistance à l'insuline en fonction des sexes, des ethnies, et de l'environnement (milieu rural vs urbain)

Ces résultats suggèrent donc une tendance pour une résistance à l'insuline plus élevée en milieu urbain (influence de l'environnement) et chez les Kirghiz (influence de facteurs populationnels). Mais il est important de souligner que ces échantillons sont hétérogènes en ce qui concerne l'âge, le poids, et l'activité physique. Afin de voir s'il existe une différence significative de résistance à l'insuline entre ces populations, il nous faut corriger pour ces co-variables confondantes.

6) Régressions

Nous cherchons à savoir si la variable « ethnie », ici reflet des composantes génétiques propres à chaque mode de subsistance ancestral, est un facteur de risque pour la résistance à l'insuline, en corrigeant pour d'autres co-variables explicatives confondantes, comme l'âge, le sexe ratio, le milieu (urbain vs rural), le poids et l'activité physique. Pour cela, nous avons effectué des régressions linéaires pour 453 individus sur les données de résistance à l'insuline (variable d'intérêt), en utilisant conjointement les variables précédentes comme variables explicatives. Les régressions sur les données continues sont des régressions linéaires (méthode des moindres carrés), tandis que sur les données discrètes, il s'agit de régressions logistiques. Nous n'avons pas corrigé pour des éventuelles différences d'alimentation car le nombre important de données manquantes pour l'alimentation réduisait trop fortement notre échantillon. De plus, nous n'avons pas trouvé de différences importantes d'alimentation entre ethnies.

L'âge est utilisé en variable continue, le sexe, le milieu (urbain vs rural) et l'activité physique en variables discrètes. Pour l'activité physique, les individus ont été classés d'après les questionnaires d'auto-estimation en inactifs vs actifs (cumulant les peu actifs, actifs et très actifs). Il faut noter que le poids est à la fois une co-variable confondante (puisqu'elle reflète notamment des différences environnementales d'alimentation) et une variable d'intérêt (puisque l'obésité est associée à la résistance à l'insuline). Si le surpoids est significativement expliqué par la variable « ethnie », alors nous ne pourrons pas savoir dans quelle mesure cette variable « ethnie » a une influence directe ou indirecte sur la résistance à l'insuline. Nous allons donc commencer par effectuer des régressions avec le poids comme variable d'intérêt.

a) Régressions sur le poids

Plusieurs variables de poids peuvent être utilisées. Nous allons ici tester trois variables : l'IMC, le pli sous-cutané, et le pourcentage de masse grasse (le rapport tour de taille sur tour de hanche n'étant pas significativement entre catégories). L'IMC permet de corriger le poids pour la taille des individus, mais ne permet pas de savoir si l'excès de poids est dû à de la masse maigre (muscles, os et eau corporelle) ou de la masse grasse. Au contraire, les mesures de pli sous-cutané sont censées être plus sensibles à la graisse qu'aux muscles, et les données de bioimpédance nous permettent d'obtenir le pourcentage prédit de masse grasse.

IMC

Pour l'IMC, les sujets normo-pondéraux (137 individus, IMC inférieur à 25) sont comparés à ceux en surpoids (261 individus, IMC supérieur à 25). Les résultats sont présentés ci-dessous :

Paramètre	Facteur de risque (écart-type)	Intervalle de confiance à 95%	*p*
Kirghiz (vs Tadjik)	0.811 (0.192)	0.51 - 1.288	0.374
Urbain (vs Rural)	**1.707 (0.384)**	**1.098 - 2.652**	**0.017**
Femmes (vs Hommes)	1.365 (0.306)	0.880 - 2.118	0.164
Age	**1.050 (0.013)**	**1.025 - 1.075**	**0.000**
Inactif (vs Actifs)	0.707 (0.262)	0.342 - 1.461	0.349

Tableau 18 : Régression logistique sur l'IMC (individus normo-pondéraux vs en surpoids)

Nous voyons donc que l'IMC n'est pas significativement différent entre Kirghiz et Tadjiks ($p = 0.374$), avec cependant une tendance pour un IMC plus faible chez les Kirghiz (Facteur de risque : FR = 0.811). Par contre, l'IMC est significativement plus élevé en milieu urbain ($p = 0.017$) et est positivement corrélé à l'âge ($p = 0.000$). Si nous effectuons une régression linéaire sur les données continues d'IMC, nous trouvons un effet significatif du milieu (urbain vs rural, $p = 0.002$), du sexe ($p = 0.009$) et de l'âge ($p = 0.001$). L'ethnie n'est donc jamais un facteur significativement explicatif de l'IMC.

Pli sous-cutané et pourcentage de graisse

Pour le pli sous-cutané et le pourcentage de graisse, nous n'avons pas de valeur de seuil, donc nous avons uniquement effectué une régression linéaire sur les données continues. Nous trouvons également que ces variables de poids sont significativement expliquées par le milieu (urbain vs rural), le sexe et l'âge. De plus, nous trouvons, à la fois pour le pli sous-cutané et le pourcentage de masse grasse, que les Kirghiz ont significativement moins de « graisse » que les Tadjiks ($p = 0.002$ et 0.001, respectivement). Nous ne trouvions pourtant pas de différences significatives de moyenne entre ethnies pour ces valeurs.

Au final, l'IMC n'est pas corrélé à l'ethnie et donc le fait de corriger pour cette variable dans les régressions ne change pas *a priori* les résultats trouvés pour le facteur « ethnie ». Par contre, le pli sous-cutané et le pourcentage de masse grasse sont significativement différents entre ethnies, avec les Kirghiz ayant moins de masse grasse que les Tadjiks.

b) Régressions sur la résistance à l'insuline

IMC comme variable de poids explicative

Les régressions logistiques sur la résistance à l'insuline, en prenant comme variable de poids l'IMC (en données continues) sont présentées ci-dessous :

Paramètre	**Facteur de risque (écart-type)**	**Intervalle de confiance à 95%**	***p***
Kirghiz (vs Tadjik)	**1.805 (0.540)**	**1.005 - 3.243**	**0.048**
Urbain (vs Rural)	1.290 (0.363)	0.744 - 2.239	0.364
Femmes (vs Hommes)	0.670 (0.185)	0.390 - 1.152	0.147
Age	1.000 (0.014)	0.972 - 1.028	0.997
Pré-obèses (vs Normo-pondéraux)	**3.010 (1.141)**	**1.432 - 6.327**	**0.004**
Obèses (vs Normo-pondéraux)	**8.900 (3.517)**	**4.103 - 19.307**	**0.000**
Inactif (vs Actifs)	**2.306 (0.940)**	**1.037 - 5.124**	**0.040**

Tableau 19 : Régression logistique sur la résistance à l'insuline, avec l'IMC comme variable de poids

Nous voyons ici que les Kirghiz ont significativement 1.8 fois plus de risque que les Tadjiks de présenter de la résistance à l'insuline (p = 0.048). Les individus pré-obèses et obèses ont respectivement 3 et 8.9 fois plus de risque d'être résistants à l'insuline (p = 0.004 et 0.000, respectivement), par rapport aux normo-pondéraux. De même, le fait de ne pas être actif confère un risque accru de 2.3 fois (p = 0.040). Les différences de milieu, de sexe et d'âge n'ont pas d'effet sur la résistance à l'insuline. Si l'on considère la résistance à l'insuline en données continues (régression linéaire), les résultats sont identiques avec cependant l'effet « milieu » qui devient significatif (p = 0.034).

Pli et pourcentage de graisse comme variables de poids explicatives

Si l'on considère le pli comme variable de poids (données continues) dans une régression logistique sur la résistance à l'insuline, l'ethnie est un facteur de risque significatif (FR = 2.040, p = 0.017), ainsi que le sexe (avec un facteur de risque plus fort pour les hommes : FR = 3.300, p = 0.002), l'activité physique (FR = 2.227, p = 0.034) et le pli (FR = 1.081, p = 0.000). Si l'on considère la résistance à l'insuline en données continues (régression linéaire), l'ethnie, le sexe, l'activité physique et le pli sous-cutané sont également des facteurs de risque significatifs (p = 0.000, 0.000, 0.012 et 0.000, respectivement).

Si l'on considère le pourcentage de graisse comme variable de poids (données continues) sur la résistance à l'insuline en données discrètes (régression logistique), l'ethnie est un facteur de risque significatif (FR = 2.259, p = 0.007), ainsi que le sexe (avec un facteur de risque plus fort pour les hommes : FR = 5.319, p = 0.000), l'âge (FR = 0.971, p = 0.049) et le pourcentage de graisse (FR = 1.165, p = 0.000). Si l'on considère la résistance à l'insuline en données continues (régression linéaires), l'ethnie, le sexe, l'âge, l'activité physique et le pourcentage de graisse sont des facteurs de risque significatifs (p = 0.000, 0.000, 0.002, 0.027 et 0.000, respectivement).

Les facteurs explicatifs les plus importants pour comprendre les variations de la résistance à l'insuline sont donc les variables de poids (avec l'IMC étant la meilleure variable prédictive vu les forts facteurs de risque associés), l'activité physique, et l'ethnie. Quelle que soit la manière dont nous analysons les données, ces variables ont toujours un effet significatif (sauf dans un cas pour l'activité physique), avec un risque accru d'être insulino-résistant pour les individus ayant un poids plus important, n'effectuant pas d'activité physique, et appartenant à l'ethnie kirghize.

Le sexe est également un facteur de risque systématiquement significatif, et très fort, lorsque l'on considère comme variable de poids le pli sous-cutané et le pourcentage de masse grasse (FR = 3.3 et 5.3, respectivement), mais pas lorsque l'on considère l'IMC. Les autres co-variables (« milieu » et « âge ») n'ont d'effet significatif qu'une fois (pour le milieu) et deux fois (pour l'âge) sur six analyses, ce qui montre qu'elles ne sont pas associées de façon robuste à la résistance à l'insuline.

Analyse par sexe avec l'IMC

Si l'on refait ces analyses par sexe, l'effet « ethnie » est systématiquement trouvé significatif, sauf pour les hommes, quand l'on considère la résistance à l'insuline en données discrètes. Ce résultat est très certainement la conséquence d'un manque de puissance statistique, lié au fait que les hommes représentent environ seulement 40% de l'échantillon total.

Ainsi, de manière générale, nous retrouvons une influence des facteurs populationnels sur la résistance à l'insuline, avec des facteurs de risque significatifs compris entre 1.8 et 2.3 (sauf pour l'analyse des hommes en données discrètes). Nous observons également une

influence forte du surpoids : facteurs de risque significatifs de 3 pour la pré-obésité et de 8.9 pour l'obésité, bien que le pli sous-cutané et le pourcentage de masse graisse représentent des facteurs de risque faibles (mais significatifs) de 1.1 et 1.2, respectivement. Nous trouvons également une influence considérable de l'activité physique, avec des facteurs de risque significatifs compris entre 2.2 et 2.3, sauf pour la régression logistique avec le pourcentage de graisse en variable de poids.

7) Conclusions

En conclusion, nous avons vu que les enquêtes alimentaires ne montrent pas de différences de régime alimentaire très importantes entre ethnies ou entre milieux (urbain vs rural). Cependant, certaines tendances en accord avec les régimes alimentaires traditionnels de ces populations ressortent, comme une consommation accrue d'huile, ainsi que de fruits et légumes chez les Tadjiks, et de viande et de lait chez les Kirghiz. L'alimentation kirghize a cependant clairement intégré des aliments nouveaux, non traditionnels, dans son régime alimentaire, comme les pommes de terre, le pain, les oignons, les pâtes et le sucre. Ceci peut être attribué à la soviétisation de la région tout au long du 20ème siècle, qui a fortement normalisé les conditions et modes de vie entre ville et campagne, d'une part, et entre ethnies, d'autre part. Nous avons également montré que les urbains consomment plus de viande et de fruits et légumes, tandis que les ruraux consomment plus d'amylacées, de lait, de sucre et d'huile.

Au niveau de l'activité physique, nous avons mis en évidence de forts contrastes entre milieux urbain et rural, particulièrement chez les Tadjiks. Nous n'avons pas trouvé de différences significatives entre ethnies en milieu urbain, bien que les Tadjiks pratiquent une activité physique équivalente aux Kirghiz qui sont plus âgés de 10 ans en moyenne dans notre échantillon. La dernière mission prévue en Ouzbékistan permettra d'obtenir des données sur l'activité physique des Tadjiks d'après les questionnaires IPAQ, qui fournissent des indices moins subjectifs et plus précis que le questionnaire d'auto-estimation.

Au niveau des variables anthropométriques, les populations ne sont pas parfaitement équilibrées en termes de taille d'échantillon, d'âge et de sexe-ratio. Ces différences compliquent l'interprétation des résultats bruts, et justifient l'organisation d'une dernière mission en Ouzbékistan pour compléter et rééquilibrer nos jeux de données. Les variables de poids ne montrent aucune différence significative entre milieu urbain et rural chez les Tadjiks. Par contre, il y a clairement un poids plus important en milieu urbain chez les Kirghiz, surtout chez les hommes, ce qui contraste avec les données d'activité physique. En moyenne, aucune différence de poids n'a été trouvée entre ethnies, sauf avec les régressions linéaires où les Kirghiz semblent avoir un pli sous-cutané et un pourcentage de masse grasse moins importants que les Tadjiks. En ce qui concerne la taille, toutes les catégories sont concernées par une augmentation de quelques centimètres en milieu urbain.

Si nous regardons les données de prévalence de résistance à l'insuline, nous trouvons toujours une valeur plus forte chez les Kirghiz par rapport aux Tadjiks, pour sexes et milieux comparables. De même, il existe toujours une augmentation de cette prévalence entre milieux rural et urbain, sauf chez les femmes kirghizes. Ce résultat peut être dû à une prévalence particulièrement faible dans l'échantillon des femmes kirghizes en milieu urbain, auquel cas les différences entre Kirghiz et Tadjiks pourraient être sous-estimées. Mais cela peut également être dû à une surestimation de la prévalence dans l'échantillon des femmes kirghizes en milieu rural, auquel cas la résistance à l'insuline pourrait être corrélée à l'ethnie et au milieu dans une moindre mesure.

Finalement, en contrôlant pour l'âge, le sexe, l'activité physique, le poids et le milieu (urbain vs rural), nous avons trouvé un niveau de résistance à l'insuline toujours plus important chez les Kirghiz par rapport aux Tadjiks, avec un facteur de risque compris entre 1.8 et 2.3 (selon la variable de poids considérée). Ces différences de résistance à l'insuline entre ethnies peuvent être attribuées au fond génétique de chacune, bien que nous ne pouvons pas strictement exclure que d'autres différences environnementales ou culturelles puissent en être responsables. Cependant, alors que les prévalences de résistance à l'insuline sont en moyenne plus fortes en milieu urbain par rapport aux milieux ruraux, nous ne retrouvons pas, en corrigeant pour les autres co-variables, de facteur de risque significativement associé au milieu (urbain vs rural). Ceci peut signifier que nos corrections pour le poids et l'activité physique ont correctement capturé la part de variations liées aux différences environnementales. Ainsi, nos données supportent l'idée que les facteurs populationnels (donc génétiques) expliquent les écarts entre Kirghiz et Tadjiks.

Pour tester cette hypothèse, nous allons tout d'abord regarder la fréquence de 10 mutations préalablement associées au diabète de type II. Si les facteurs de risque génétiques sont cohérents avec les données phénotypiques, nous nous attendons à observer des allèles à risque en plus forte fréquence chez les Kirghiz. Ensuite, nous chercherons à comprendre si les éventuelles différences de fréquences alléliques entre ethnies sont dues à l'histoire démographique ou à l'histoire adaptative des populations, par une approche de génétique des populations. Pour cela, nous allons tester directement si certains gènes candidats associés à ce phénotype sont sous l'influence de la sélection, en comparant la distribution de leur diversité génétique avec celle de régions *a priori* neutres.

C. Analyses génétiques

1) Matériel et méthodes

a) Choix des régions

Séquences neutres

Les 20 régions autosomales non codantes ont été choisies par Patin (2009) selon les critères suivants : (i) être au moins à 200kb de tout gène connu ou prédit (distance déterminée grâce à la base de données UCSC, version hg18); (ii) de ne pas être en déséquilibre de liaison avec aucun gène connu ou prédit (déterminé grâce à la base de données HapMap, version 16); (iii) de ne pas être en déséquilibre de liaison entre elles et (iv) d'être dans une région d'homologie avec le génome du chimpanzé.

Séquences dans les gènes candidats

De manière générale, une très grande majorité des études d'association a été faite dans des populations d'origine européenne. Pour le diabète de type II, une vingtaine de gènes environ ont été identifiés comme étant associés à cette maladie dans ces populations (Barroso, 2005, Freeman & Cox, 2006, Sladek *et al.*, 2007, Frayling, 2007, Zeggini *et al.*, 2008, Prokopenko *et al.*, 2008). Cependant, jusqu'à assez récemment, peu d'études ont confirmés ces résultats dans d'autres ethnies. Certaines études d'association ont tout de même été effectuées dans des populations japonaises, chinoises, indiennes ou coréennes (Omori *et al.*, 2008, Liu *et al.*, 2008, Sanghera *et al.*, 2008, Unoki *et al.*, 2008, Yasuda *et al.*, 2008, Cho *et al.*, 2009, Zhou *et al.*, 2009b, Zhou *et al.*, 2009a, Hu *et al.*, 2009). Ces études ont majoritairement confirmé l'association de cinq gènes avec le diabète de type II : *IGF2BP2*, *HHEX*, *SLC30A8*, *CDKAL1* et *CDKN2A/B*, et deux gènes supplémentaires ont été identifiés : *KCNJ11* et *KCNQ1*. Tous ces gènes ont été choisis pour notre étude sauf *CDKN2A/B*, dont la séquence d'intérêt était située dans une région difficilement amplifiable à cause de délétions.

Nous avons de plus ajoutés à cette liste le gène *TCF7L2* qui présente les signes les plus forts d'association dans les populations européennes (Weedon, 2007, Saxena *et al.*, 2006, Scott *et al.*, 2007, Sladek *et al.*, 2007, Zeggini *et al.*, 2008), ainsi que des signes de sélection positive (Helgason *et al.*, 2007). Nous avons également ajouté les gènes *FABP2*, *LEPR* et *PON1,* trouvés significativement associés au climat (lui-même potentiellement lié au mode de vie) (Hancock *et al.*, 2008), et également connus pour leur association à la résistance à

l'insuline ou au diabète de type II (FABP2 : Baier *et al.*, 1995, , LEPR :Wauters *et al.*, 2001 , Dulloo *et al.*, 2002 , Park *et al.*, 2006 , Qu *et al.*, 2008a, , et PON1 : Qu *et al.*, 2008b).

Les 11 régions choisies se trouvent donc dans 11 gènes différents, et sont centrées sur des mutations répliquées dans le plus d'ethnies possibles et/ou avec les plus forts signaux d'association trouvés.

b) Séquençage

Les extractions d'ADN ont été faites sur place (par l'équipe de Tatiana Hegay) ou en France (par Myriam Georges et Sophie Lafosse) selon le protocole classique de Maniatis (1982).

Séquences neutres

Les PCR ont été effectuées par moi-même avec l'aide de Myriam Georges dans les conditions suivantes : pour les données neutres, les PCR ont été faites selon deux protocoles différents (voir tableau 20 ci-dessous).

Nom de Séquence	Protocole de PCR	Température d'hybridation	Nom de Séquence	Protocole de PCR	Température d'hybridation
N1	1	56	N11	2	56
N2	1	59	N12	1	56
N3	1	56	N13	1	56
N4	1	59	N14	1	56
N5	1	59	N15	1	59
N6	2	56	N16	2	56
N7	1	59	N17	1	56
N8	2	56	N18	1	56
N9	1	59	N19	2	59
N10	1	59	N20	1	60

Tableau 20 : Conditions de PCR pour les séquences neutres

Les réactions de PCR pour le protocole un (ou pour le protocole deux) ont été faites dans un volume final de 20µL composé de tampon Eppendorf 1X, de 412.5µM de chaque dNTP, 1.25mM (ou 2.5mM) de $MgCl_2$, 0.5U (ou 0.25U) de Taq polymérase Eppendorf, 350nM de chaque amorce et 30ng d'ADN. Les amorces utilisées sont présentées dans le tableau 21 ci-dessous. Les PCR ont été effectuées dans un thermocycleur Master Cycler

Eppendorf, avec une phase initiale de dénaturation de 5 min à 94°C ; suivie par 45 (ou 50 cycles) de 30 sec à 94°C, 30 sec à température d'hybridation (voir tableau 20) et 2 (ou 3) min à 72°C ; et une phase finale d'extension de 10 min à 72°C.

Nom	Chrom.	PCR		Taille du fragment (pb)	Séquençage	
		Amorce F	Amorce R		Amorce F	Amorce R
N1	1p	GGCTTCCAGATAAGCCTTCC	TCAATGGGTGGTTTTCTTCC	1418	TGCCAAGTGAGCATGTATTC	AACGCATCAGAGTAAACCAG
N2	2p	ATGGGCACCTTTGACTGAAC	TTCACATTGCCCAAGACTGA	1409	GAACCATGTTTCTTGCTCGG	TGACTCTTAATTGATGGTAG
N3	2q	TGGCAACGTTACACCAACAT	GGCTGCTCCCAATATAGCAG	1437	AGAAATGGTACCCAGAGGAC	GGATGAAGGCTGGTATCTTC
N4	3p	GGCTGCAGCTAATTGAAAGC	GAAAGTGAAGGGCAAAGTGC	1452	CATGGCTGAGAAATGTAGGC	AGGGAGATTTCAAGCACAGC
N5	3q	CCCCCTTAGAACATGGATGA	ATAGGCAGACGGAGCAAGAA	1448	CTCAAGCATGCCAATATCAC	GAAGAAAACAAGGATCACTG
N6	4p	AGGGAGAGGTCGCAAAGATT	TGCACATTCACACGAACTGA	1454	GCAAGGCTTATATAATGGTG	GATTTCAACGAGTCTGTAAG
N7	4q	TTGCTTGTCATGAGGACTGC	CATTGCCTTCATCCCTTTGT	1405	AGTAGGGAGAATCGGTGTAC	GTATTAACTATGTGGCAGTC
N8	6q	TCGAAAAATGACCAGGAAGG	TTCTGCCCCCTGTGAAATAC	1456	GAGTGATCCAGCAGTCCTAC	CCCAAGCACTATGCAAGCAC
N9	7p	CAGAGTTGGGCTCTGCTACC	TCACCCAATGCAGAAACTCA	1487	AACATCAGCCTGAGACTGAC	TAATTGCTTTTCATGCATCC
N10	7q	CCTTCTGACTCCAGGCAATC	GATTGCTGGCCTTTTAGCTG	1485	TCTACTTTCTGTCTCCATTC	TATCTATTCAGGTCATTTGC
N11	8p	CAGATGGGAAGATACTGAGC	GTGACAAATGCACTCTGGAC	1459	GAATGCTAAGACCCTTAGGC	TGTGGTGGTGCCTGTAATAC
N12	8q	CTTCCTGAGAGATGGGCAAG	AAACCAGGAATTTGCAGTGG	1453	GCTAAATCATATTGAGGATG	TGCAGAAGAACATTTAGCAG
N13	9p	AATGCACCTGGGAACTGAAC	GCCCCAATGCACTTTTACAT	1473	GTTACACATCACTGAGCTGG	AAATTTTTCAGGCTGCATAG
N14	11q	TGAGGGTGATCTCGCTAACA	GTCAGCAGAGCCAGGATTTC	1426	AAGTCTCAATCACACCTTGC	GTTGTCTTATGTTGTACAGC
N15	11q	GGTTTGCTTCTTGCTGCTTC	CCCCACATAGCTGTCCTTGT	1442	ACTGTAGCTCCTGTCACATC	TAAGGCAAGACTCACGTGTC
N16	14q	GGACTCTGGCTAAGCAGCAT	AGCGGTCAGCAACTGAAAAT	1425	TACTCCATATTCTCCTCCTG	ACTCCGTACGTAGTGAACTG
N17	15q	TGGTGGTACCCTGATGGAAT	TTGATGCCCCCTATCAATGT	1402	AAAACAGGTTTGAAATGAGC	ATTCTAGCAGATTTGCCTTC
N18	16q	TTGGTGGTTGGGAGATAAGC	TTGGAAGGATTCAAGGGAAC	1441	ATGCCATCTATAGTATTGTC	AGGCTTTCAGAAGGAGTCAC
N19	18q	ATGTTGGCTATCTGGATGCC	TCCTACCAGGTCATCCATGC	1446	TCATCTGGTTGCTTTGTTGC	CTAATACAATTTGTCTGTGC
N20	20q	TCCCAGACCATGTTCAGAAG	ACATGTGGCATTGGTGAGTC	1500	CTTTCTGGGGAAAAACACAG	CATTTTCTCAAGTGCACAGC

Tableau 21 : Amorces de PCR et de séquençage pour les 20 séquences neutres autosomales. Chrom. : bande chromosomique

Séquences dans les gènes candidats

Les réactions de PCR ont été faites dans un volume final de 20µL composé de tampon Eppendorf 1X, de 125µM de chaque dNTP, 0.5U de Taq polymérase Eppendorf, 125nM de chaque amorce et 20ng d'ADN. Les mêmes amorces ont été utilisées pour les PCR et le séquençage. Elles sont présentées dans le tableau 22 ci-dessous. Les PCR ont été effectuées dans un thermocycleur Master Cycler Eppendorf, avec une phase initiale de dénaturation de 5 min à 94°C ; suivie par 40 cycles de 30 sec à 94°C, 1 min aux températures d'hybridation indiquées dans le tableau 22, et 30 sec à 72°C ; et une phase finale d'extension de 10 min à 72°C. Les températures d'hybridation ont été choisies après avoir effectué des tests en gradient de température entre 55 et 65°C.

Gène		Chrom.	Mutation d'intérêt	Localisation	Amorce F	Amorce R	Taille du fragment (pb)	Temp.
G1	*FABP2*	4q	rs1799883	exon + intron	F1 : TGTGCTTGTTTTAATTTGTGTGA F2 : AACTCTAAAGGCTTCTTTTGCTGT	R1 : TGTCAGAAAGAAATGCCACA R2 : TGAGCAATGCATTCTTGATTTT	1292	59
G2	*TCF7L2*	10q	rs7903146	intron	TTTCTGTTTGAACAAAATTGGAA	CCTCCCAGAAAGCAAATTGA	1253	59
G3	*PPARG*	3p	rs1801282	exon + intron	TCATTTTGGGCTTCACAAATC	GAGCACCACTTGGTCCTACA	1263	59
G4	*LEPR*	1p	rs1137100	exon + intron	TGGGCATTTCCCATAAGAAG	R1 : TCAGAGATATTCCTTGCCTGAA R2 : AATTTGGTGGCATGCAAGA	1353	59
G5	*KCNJ11*	11p	rs5215 rs5219	exon + intron	CGAGGGTCAGAGCTTCCAGT	GGAAGAGTCTGGTGGGGAGT	1248	61
G6	*SLC30A8*	8q	rs13266634	exon + intron	AATGGCAGTGAGGGTTGCT	TGACTATTTGTCCCTTCCACTG	998	60.5
G7	*HHEX*	10q	rs1111875	région intergénique	GCATGTTCCACAAGGGACTC	CGCCTATGATTTGTGAGCTTT	1351	61.5
G8	*IGF2BP2*	3q	rs4402960	intron	GAGGAATAAGGCTGCCACAC	GGATGTGCATGGACACTTGA	1248	60
G9	*CDKAL1*	6p	rs10946398 rs7754840	intron	GAAGCCTTTGACGCTCTGAC	CTGCTCACTGGCATACATCA	1237	63
G10	*KCNQ1*	11p	rs2237892	intron	CAAATGCTCCACATTTGCAT	ACTGATGCCTGGAGCATAGG	1274	60
G11	*PON1*	7q	rs3917498	exon + intron	TTGCTGCAATAGAAGCTAACG	TTAAGGCTCAGGAGGCTAGATCA	1053	59

Tableau 22 : Amorces utilisées pour les PCR et le séquençage des 11 gènes candidats associés au diabète de type II. Chrom. : bande chromosomique ; Temp. : température d'hybridation. Pour FABP2, 4 amorces ont été utilisées car cette séquence présente une insertion de 11 paires de base par rapport à la séquence de référence de UCSC (insertion de TTAAATTGCTA dans l'intron, sur le brin +), et un polyA de 11 paires de base. Pour LEPR, 3 amorces ont été utilisées car cette séquence présente une délétion de 3 paires de base par rapport à la séquence de référence de UCSC (délétion de TCA dans l'intron, sur le brin +).

Les séquences ont été alignées et les SNPs détectés avec le logiciel Genalys v.3.3b. Tous les singletons ont été confirmés par réamplification et reséquencage.

c) Tests statistiques

Les tests de Wilcoxon et les corrections pour les tests multiples FDR (*False Discovery Rate*) ont été effectués grâce au logiciel R (R Development Core Team, 2007). Les régressions (linéaires et logistiques) ont été faites à l'aide du logiciel Systat version 13 (SYSTAT Software Inc.).

Les tests de neutralité ont été effectués avec DNAsp version 5 (Librado & Rozas, 2009) et Arlequin version 3.1 (Excoffier *et al.*, 2005). La significativité des tests a été obtenue avec le logiciel DNAsp, en effectuant 10 000 simulations par coalescence, d'après les valeurs de π observée sur chaque séquence. Les tests de neutralité sélective cherchent en majorité à détecter des déviations par rapport au spectre de fréquences alléliques attendu dans un modèle démographique de taille constante. Quand un excès de mutations à fréquences faibles (ou de jeunes mutations) est trouvé, cela correspond à de la croissance démographique ou de la sélection positive. Au contraire, si un excès de mutations à fréquences intermédiaires (ou mutations plus anciennes) est trouvé, cela reflète une contraction démographique ou de la sélection balancée. Le principe de ces tests est donné dans l'encadré 1 ci-dessous.

Les mesures de différenciation entre populations (F_{ST}), ainsi que les tests exacts de différenciation, ont été réalisés avec le logiciel Genepop v.4.7 (Rousset, 2008). La recherche de signatures de sélection basée sur la différenciation entre populations (en comparant les F_{ST} des gènes candidats aux F_{ST} neutres) a été faite grâce au logiciel Fdist2 (Beaumont & Nichols, 1996).

Le principe général des tests basés sur les spectres de fréquences alléliques est de comparer des estimateurs de diversité génétique qui diffèrent par l'importance donnée aux variants rares et de fréquence intermédiaire (sans groupe externe) et aux variants anciens et plus jeunes (avec groupe externe) : π, le nombre moyen de différences par paires de séquences ; S, le nombre de sites polymorphes ; η, le nombre total de mutations ; η_e, le nombre total de mutations sur les branches externes de la généalogie ; η_s, le nombre de singletons ; H, un estimateur basé sur la fréquence des variants dérivés et k, le nombre d'allèles de l'échantillon.

***D* de Tajima** : ce test est basé sur la différence entre l'estimateur Watterson de θ (S) (Watterson, 1975) et l'estimateur Tajima de θ (π) (Tajima, 1989). Ici, chaque variant rare ajoute un site polymorphe (S) mais change peu la diversité nucléotidique (π). Un test positif signale un excès de variants intermédiaires, tandis qu'un test négatif signale un excès de variants rares.

***D* de Fu et Li** : ce test est basé sur la différence entre η_e et η (Fu & Li, 1993, équation 32). Un test positif montre un excès de mutations anciennes, tandis qu'un test négatif montre un excès de mutations jeunes.

***D** de Fu et Li** : ce test est basé sur la différence entre η_s et η (Fu & Li, 1993, p. 700). Un test positif montre un déficit de variants rares, tandis qu'un test négatif montre un excès de variants rares.

***F* de Fu et Li** : ce test est basé sur la différence entre η_e et π (Fu & Li, 1993, p.702). Un test positif montre un excès de mutations anciennes, tandis qu'un test négatif montre un excès de mutations jeunes.

***F** de Fu et Li** : ce test est basé sur la différence entre π et η_s (Fu & Li, 1993, p. 702). Un test positif montre un déficit de variants rares tandis qu'un test négatif montre un excès de variants rares.

***H* de Fay et Wu** : ce test généralise celui de Tajima en incluant dans le calcul non seulement les singletons mais aussi les variants présents deux fois, trois fois, etc. Ce test est basé sur la comparaison entre π et H (Fay & Wu, 2000, équations 1-3).

***Fs* de Fu** : ce test compare π et k (Fu, 1997). Ce test est positif si il existe un excès d'allèles intermédiaires et négatif s'il existe un excès d'allèles rares (récentes).

Encadré 1 : Tests de neutralité sélective basés sur la déviation par rapport au spectre de fréquences alléliques

D'autres tests indépendants de la démographie se basent sur la comparaison du nombre de mutations synonymes par rapport aux mutations non synonymes. Ces tests sont présentés dans l'encadré 2 ci-dessous.

Les tests basés sur le rapport entre mutations synonymes et non synonymes sont puissants car ils n'interfèrent pas avec la démographie, mais ils nécessitent en contre partie que la sélection soit répétée sur toute la séquence. De plus, ils ne détectent que la sélection sur des régions codantes (non régulatrices).

dN/dS : ce test compare le dN, nombre de mutations non synonymes observées par rapport au nombre maximal de mutations non synonymes observables, au dS, nombre de mutations synonymes observées par rapport au nombre maximal de mutations synonymes observables. Ainsi, le rapport dN/dS nous renseigne sur la déviation par rapport à un modèle où les mutations synonymes et non synonymes sont affectées par les mêmes forces évolutives.

Test de Mc Donald-Kreitman : ce test (McDonald & Kreitman, 1991) équivaut au calcul du rapport dN/dS, mais prend en compte un groupe externe.

Encadré 2 : Tests basés sur le rapport entre mutations synonymes et non synonymes

2) Diversité génétique neutre sur ces populations

Nous voulons ici comparer la diversité génétique des Kirghiz et des Tadjiks, en n'analysant qu'une partie des individus échantillonnés pour les données anthropométriques (489 individus). Nous avons donc choisi aléatoirement un sous-échantillon de 40 Ki_Urb et 40 Ta_Urb pour nos analyses génétiques. Afin de vérifier que les échantillons urbains et ruraux sont homogènes génétiquement, et aussi qu'ils sont représentatifs des groupes précédemment décrits (éleveurs turco-mongols et agriculteurs indo-iraniens, respectivement), nous avons d'abord génotypé un sous-groupe de 158 individus (48 Ki_Urb, 32 Ki_Rur, 46 Ta_Urb et 32 Ta_Rur) aux mêmes 27 marqueurs microsatellites que ceux précédemment analysés dans le chapitre 1. L'analyse en composante principale sur l'ensemble de ces 30 populations d'Asie Centrale est présentée ci-dessous :

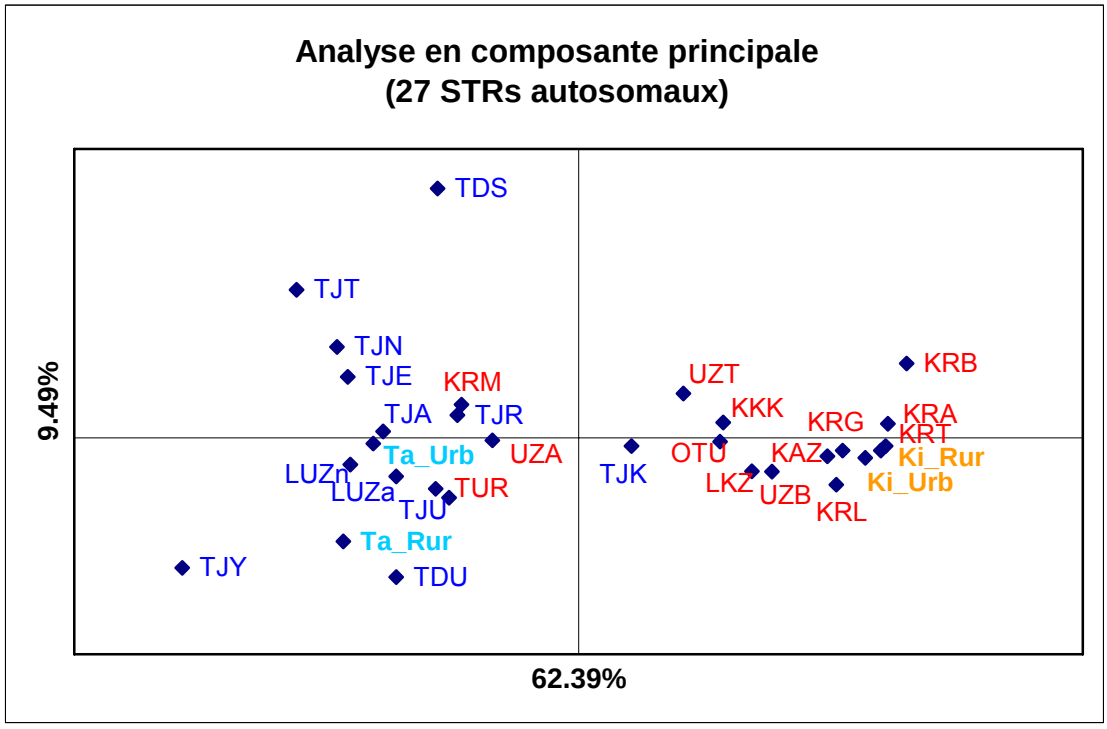


Figure 20 : Analyse en composante principale basée sur la diversité génétique de 27 marqueurs microsatellites autosomaux, pour 30 populations d'Asie Centrale. En rouge les Turco-mongols dont en orange les Kirghiz utilisés dans cette étude ; en bleu foncé les Indo-iraniens dont en bleu clair les Tadjiks utilisés dans cette étude.

Nous voyons donc sur ce graphique, où les deux premiers axes représentent respectivement 62.4% et 9.5% de la variation, que les populations urbaines et rurales de

chaque ethnie sont génétiquement très proches entre elles. Elles ne sont d'ailleurs pas significativement différenciées (F_{ST} = 0.0002, p = 0.10 pour les Kirghiz ; F_{ST} = 0.002, p = 0.06 pour les Tadjiks). Ces populations ont été échantillonnées à 30km d'écart, tandis que les populations kirghizes ont été échantillonnées à environ 200 km. Ces populations sont également bien représentatives des autres populations de leurs groupes, celui des éleveurs turco-mongols et celui des agriculteurs indo-iraniens, respectivement.

3) Fréquences des mutations associées au diabète de type II

Le séquençage de la région du gène *IBF2BP2* n'a pas pu être finalisé à cause de difficultés de lecture dues à des insertions / délétions. Cette séquence n'a donc pas été intégrée aux analyses qui suivent. Les 10 séquences géniques présentées sont alors respectivement dans les gènes: *FABP2*, *TCF7L2*, *PPARG*, *LEPR*, *KCNJ11*, *SLC30A8*, *HHEX*, *CDKAL1*, *KCNQ1* et *PON1*.

Nous allons tout d'abord regarder comment se répartissent les facteurs de risque génétiques entre ethnies (c'est-à-dire les allèles associés au diabète de type II), avant de tester, à l'aide de tests de sélection, quelles forces évolutives sont responsables des fréquences alléliques observées. Les fréquences des allèles à risque pour le diabète de type II sont présentées ci-dessous :

Gène	Mutation	Allèle à risque	Fréquence chez les Kirghiz	Fréquence chez les Tadjiks
FABP2	rs1799883	(T)	38.5%	31.4%
TCF7L2	rs7903146	(T)	10.0%	17.1%
PPARG	rs1801282	(C)	90.0%	90.8%
*LEPR**	rs1137100*	(A)	41.3%	66.2%
KCNJ11	rs5215 rs5219	(C) (T)	36.3%	36.5%
SLC30A8	rs13266634	(C)	61.5%	79.5%
HHEX	rs1111875	(C)	39.5%	45.9%
CDKAL1	rs10946398 rs7754840	(C) (C)	30.0%	19.2%
KCNQ1	rs2237892	(C)	78.9%	97.4%
PON1	rs3917542	(C)	56.3%	75.0%
* Différence significative de fréquence alléliques entre ethnies				

Tableau 23 : Fréquence des allèles à risque associés au diabète de type II chez les Kirghiz et les Tadjiks.

Bien qu'aucune différence significative n'ait été détectée entre fréquences alléliques (à part pour LEPR), nous voyons qu'il y a une tendance vers plus d'allèles à risque chez les Tadjiks. En effet, seulement deux mutations ont des fréquences plus fortes chez les Kirghiz (sur les gènes *FABP2* et *CDKAL1*), versus six chez les Tadjiks (*TCF7L2*, *LEPR*, *SLC30A8*, *HHEX*, *KCNQ1* et *PON1*), les deux dernières étant considérées comme identiques entre populations (*PPARG*, *KCNJ11*).

Ces mutations ne permettent donc pas bien d'expliquer nos données phénotypiques. Ce résultat est certainement dû au biais de détection des allèles à risque, puisque les études d'association sont majoritairement effectuées dans les populations européennes. Ainsi, les mutations identifiées sont en plus forte fréquence dans ces populations. Comme les Tadjiks ont une composante génétique plus proche des populations européennes que les Kirghiz (voir résultats du chapitre I partie III), nous nous attendons à retrouver ce biais de fréquence dans nos échantillons. Nous allons cependant maintenant essayer de comprendre quelles forces évolutives sont responsables de la distribution de la diversité génétique dans ces régions géniques, en étudiant si, et comment, la sélection agit sur ces gènes. Pour cela, nous allons comparer la distribution de la diversité génétique des dix régions géniques candidates à celle des 20 régions *a priori* neutres.

4) Description de la diversité génétique intra-population

Tout d'abord, aucune séquence ne présente de déviation au test d'Hardy-Weinberg, après correction pour les tests multiples (données non présentées).

La description des statistiques qui résument au mieux la diversité génétique intra-population dans chaque population, pour les séquences *a priori* neutres et pour les séquences géniques, est présentée dans les tableaux 24 et 25, respectivement.

Séquence	Taille (pb)	Population	n	S	η_S	π	k	H	H_{div}	Div.	Population	n	S	η_S	π	k	H	H_{div}	Div.
N1	1334	**Ki_Urb**	78	1	0	0,00017	0,226	2	0,226	0,0083	**Ta_Urb**	76	2	1	0,00031	0,419	3	0,412	0,0085
N2	1271	**Ki_Urb**	80	10	1	0,00181	2,298	6	0,729	0,0122	**Ta_Urb**	78	9	1	0,00239	3,032	5	0,735	0,0125
N3	1348	**Ki_Urb**	76	7	2	0,00173	2,331	6	0,726	0,0153	**Ta_Urb**	76	7	2	0,00173	2,331	7	0,706	0,0154
N4	1328	**Ki_Urb**	80	5	1	0,00085	1,132	6	0,730	0,0164	**Ta_Urb**	78	9	5	0,00099	1,311	10	0,775	0,0165
N5	1363	**Ki_Urb**	78	5	2	0,00081	1,101	6	0,764	0,0108	**Ta_Urb**	78	5	1	0,00084	1,141	6	0,771	0,0108
N6	1256	**Ki_Urb**	80	4	0	0,00065	0,822	4	0,386	0,0155	**Ta_Urb**	78	6	0	0,00086	1,083	6	0,464	0,0156
N7	1310	**Ki_Urb**	80	10	2	0,00116	1,517	6	0,424	0,0100	**Ta_Urb**	78	10	2	0,00106	1,390	6	0,478	0,0101
N8	1365	**Ki_Urb**	80	2	1	0,00031	0,429	3	0,422	0,0167	**Ta_Urb**	74	1	0	0,00034	0,470	2	0,470	0,0166
N9	1382	**Ki_Urb**	80	12	1	0,00217	3,002	11	0,810	0,0162	**Ta_Urb**	78	17	7	0,00185	2,559	14	0,768	0,0160
N10	1351	**Ki_Urb**	80	3	0	0,00032	0,437	4	0,402	0,0186	**Ta_Urb**	76	6	2	0,00052	0,696	7	0,548	0,0188
N11	777	**Ki_Urb**	80	6	0	0,00147	1,145	5	0,472	0,0197	**Ta_Urb**	78	6	1	0,00137	1,063	5	0,507	0,0197
N12	1282	**Ki_Urb**	78	5	0	0,00133	1,707	6	0,73	0,0197	**Ta_Urb**	78	4	0	0,00124	1,585	5	0,740	0,0196
N13	1363	**Ki_Urb**	80	4	1	0,00037	0,503	4	0,361	0,0155	**Ta_Urb**	78	5	2	0,00046	0,626	5	0,372	0,0155
N14	1334	**Ki_Urb**	80	8	2	0,00051	0,679	7	0,388	0,0174	**Ta_Urb**	78	7	2	0,00039	0,517	5	0,348	0,0174
N15	1314	**Ki_Urb**	80	4	0	0,00024	0,314	4	0,210	0,0153	**Ta_Urb**	78	2	1	0,00008	0,101	3	0,100	0,0153
N16	1321	**Ki_Urb**	78	7	1	0,00116	1,530	7	0,725	0,0084	**Ta_Urb**	78	6	1	0,00101	1,338	6	0,637	0,0084
N17	1301	**Ki_Urb**	78	5	1	0,00114	1,488	5	0,591	0,0156	**Ta_Urb**	76	4	1	0,00118	1,538	4	0,588	0,0157
N18	1347	**Ki_Urb**	78	10	2	0,00265	3,556	7	0,762	0,0133	**Ta_Urb**	78	10	2	0,00261	3,521	7	0,749	0,0133
N19	1280	**Ki_Urb**	72	4	0	0,00129	1,657	4	0,639	0,0045	**Ta_Urb**	74	6	1	0,00141	1,807	6	0,638	0,0047
N20	1373	**Ki_Urb**	78	4	2	0,00013	0,177	4	0,100	0,0102	**Ta_Urb**	78	3	0	0,00046	0,627	3	0,313	0,0102
Moyenne	**1300**		**79**	**5,8**	**1**	**0,00101**	**1,303**	**5,4**	**0,53**	**0,014**		**77**	**6,3**	**1,6**	**0,00106**	**1,358**	**5,8**	**0,556**	**0,014**

n : nombre de chromosomes ; S : nombre de sites polymorphes ; η_S : nombre de singletons ; π : diversité nucléotidique (nombre moyen de différences par paires de séquences) ; k : nombre moyen de différences nucléotidiques ; H : nombre d'haplotypes ; H_{div} : diversité haplotypique ; Div. : nombre moyen de substitutions par site par rapport au chimpanzé

Tableau 24 : Statistiques résumées décrivant la diversité génétique des séquences neutres par population

Séquence		*n*	Taille (pb)	*S*	η_S	π	*k*	*H*	H_{div}	Div.
FABP2	Ki_Urb	78	1229	12	0	0,00405	4,981	5	0,703	0,0122
	Ta_Urb	66	1229	13	1	0,00414	5,092	7	0,751	0,0122
TCF7L2	Ki_Urb	80	1119	3	1	0,00031	0,348	4	0,188	0,0143
	Ta_Urb	76	1119	2	0	0,00048	0,538	3	0,295	0,0143
PPARG	Ki_Urb	80	1114	1	0	0,00016	0,182	2	0,182	0,0019
	Ta_Urb	76	1114	2	1	0,00018	0,196	3	0,172	0,0019
LEPR	Ki_Urb	80	1252	4	0	0,00098	1,229	4	0,605	0,007
	Ta_Urb	76	1252	5	1	0,00099	1,244	5	0,723	0,0068
KCNJ11	Ki_Urb	80	1147	4	1	0,00190	1,365	4	0,679	0,0113
	Ta_Urb	74	1147	6	3	0,00121	1,394	6	0,690	0,0114
SLC30A8	Ki_Urb	78	923	8	4	0,00189	1,740	8	0,724	0,0124
	Ta_Urb	78	923	8	4	0,00191	1,759	7	0,697	0,0115
HHEX	Ki_Urb	76	1261	2	0	0,00076	0,956	2	0,478	0,0054
	Ta_Urb	74	1261	2	0	0,00080	1,011	2	0,505	0,0055
CDKAL1	Ki_Urb	80	1140	10	1	0,00310	3,532	5	0,537	0,0092
	Ta_Urb	78	1140	9	0	0,00244	2,779	4	0,411	0,0094
KCNQ1	Ki_Urb	76	1173	4	0	0,00062	0,731	5	0,522	0,0089
	Ta_Urb	76	1173	4	0	0,00048	0,567	4	0,282	0,0088
PON1	Ki_Urb	80	959	5	0	0,00153	1,466	4	0,685	0,0021
	Ta_Urb	76	959	6	1	0,00189	1,817	5	0,687	0,0023
Moyenne		**77**	**1132**	**5,5**	**0,9**	**0,00149**	**1,646**	**4,5**	**0,526**	**0,0082**

n : nombre de chromosomes ; *S* : nombre de sites polymorphes ; η_S : nombre de singletons ; π : diversité nucléotidique (nombre moyen de différences par paires de séquences) ; *k* : nombre moyen de différences nucléotidiques ; *H* : nombre d'haplotypes ; H_{div} : diversité haplotypique ; Div. : nombre moyen de substitutions par site par rapport au chimpanzé

Tableau 25 : Statistiques résumées décrivant la diversité génétique des séquences géniques.

a) Indices de diversité

Les 20 séquences neutres ont en moyenne 5.8 sites polymorphes, variant de 1 à 17 (dont 1.3 singletons en moyenne), pour une longueur moyenne de 1300 pb. Les 10 séquences géniques sont significativement plus courtes en moyenne (1132 pb, $p = 9.8\ 10^{-5}$) et présentent un nombre de polymorphismes comparable (5.5, variant de 1 à 13 dont 0.9 singletons en moyenne, p = 0.64 et 0.48 pour les Tadjiks et les Kirghiz, respectivement).

La diversité nucléotidique ($\boldsymbol{\pi}$) varie entre $0.08\ 10^{-3}$ et $4.14\ 10^{-3}$ (voir tableaux 24 et 25) selon la distribution suivante :

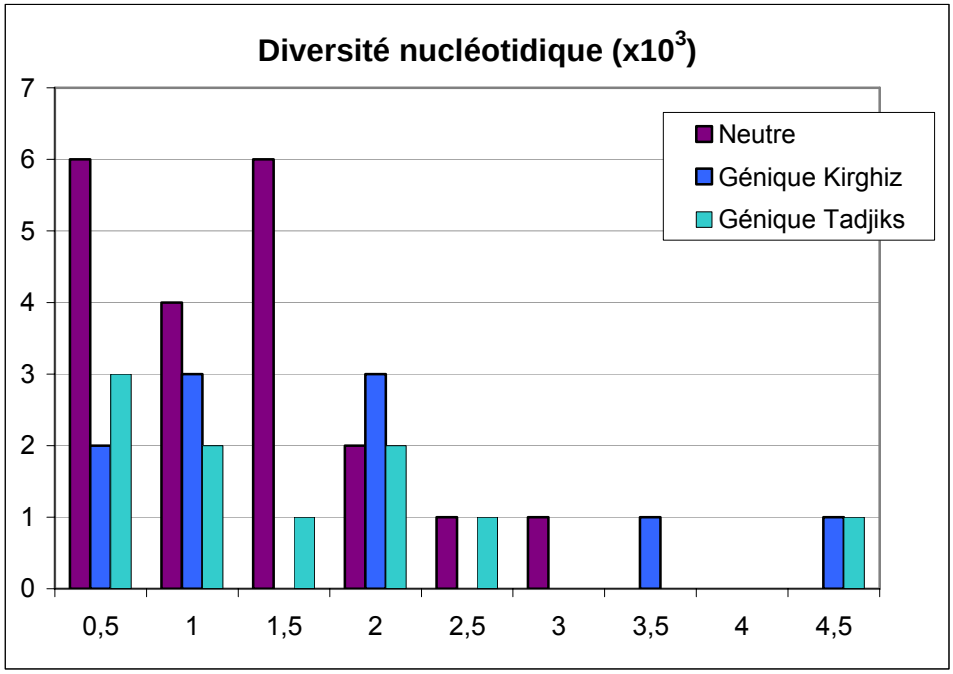


Figure 21 : Diversité nucléotidique ($x10^3$) par séquence

Les séquences neutres chez les Kirghiz et les Tadjiks présentent exactement la même distribution de diversité nucléotidique. Ainsi, elles sont réunies sous une seule et même distribution (appelée neutre dans la figure 21 ci-dessus). Nous voyons sur cette figure qu'une séquence génique est relativement en marge du reste de la distribution, à la fois chez les Kirghiz et les Tadjiks (π = 4.05 10^{-3} et 4.14 10^{-3}, respectivement). Il s'agit de la séquence sur le gène *FABP2*. Ce gène présente donc un très grand nombre de différences nucléotidiques au sein de chaque population. En théorie, ce résultat peut être dû à de la sous-structure cachée de population, de la décroissance démographique de la population (contraction), de la sélection balancée, ou de la sélection positive partielle. Les deux premières hypothèses (démographiques) doivent cependant influencer tout le génome de la même manière en espérance ((Nielsen, 2005)), ce qui n'est pas le résultat observé.

La diversité haplotypique neutre varie quant à elle entre 0.10 et 0.81 (voir tableaux 24 et 25), avec de 2 à 14 haplotypes par population (5.6 en moyenne), ce qui recouvre une large gamme de valeurs, dont notamment celles sur les séquences géniques : entre 0.17 et 0.75, avec de 2 à 8 haplotypes par population (4.5 en moyenne). Nous voyons donc ici un nombre d'haplotypes légèrement plus faible sur les séquences géniques, mais non significativement (test de Wilcoxon, p = 0.27 et 0.14 pour les Tadjiks et les Kirghiz, respectivement). De manière intéressante, le gène *FABP2*, pour lequel la diversité nucléotidique est

particulièrement élevée, présente également une des plus fortes diversités haplotypiques (0.70 et 0.75 chez les Kirghiz et les Tadjiks, respectivement).

- Divergence par rapport au chimpanzé

La divergence de chaque séquence par rapport au chimpanzé se distribue comme suit :

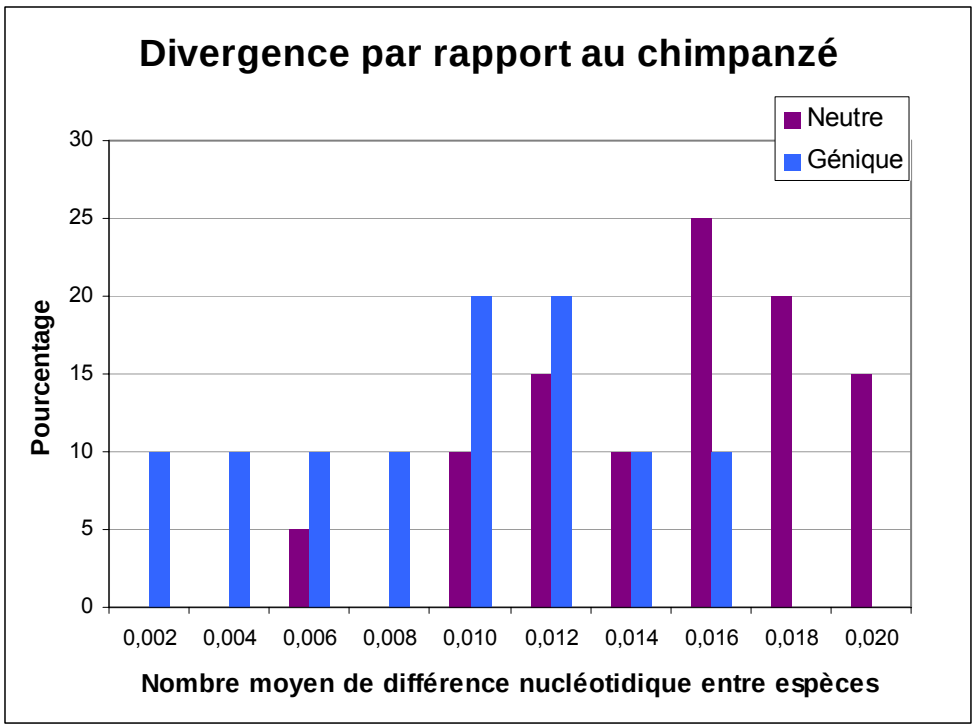


Figure 22 : Divergence nucléotidique par séquence entre homme et chimpanzé

Nous voyons que la divergence entre les séquences des populations d'Asie Centrale et celle du chimpanzé est en moyenne presque deux fois plus forte pour les séquences neutres (0.014) que pour les séquences géniques (0.008), malgré un certain recouvrement des deux distributions (voir tableaux 24 et 25). Cette différence significative (p = 0.003 et 0.005 chez les Tadjiks et les Kirghiz, respectivement) peut s'expliquer par le fait que les séquences géniques sont plus conservées que les séquences neutres, sous l'effet d'une sélection négative. Les séquences géniques les plus conservées sont celles sur *PPARG* (0.0019) et *PON1* (0.0022).

5) Spectre de fréquences alléliques et tests de neutralité sélective

Sur des données de séquence, nous pouvons effectuer des tests de neutralité basés sur les différences de pression de sélection entre mutations synonymes et non-synonymes, ainsi que des tests de neutralité basés sur les écarts de spectres de fréquences alléliques par rapport à l'attendu sous neutralité dans un modèle de taille constante. Le principe de ces tests est rappelé dans la partie « Matériel et Méthodes » de cette section.

- Test de dN/dS (et Mc Donald-Kreitman)

Sur les 10 séquences étudiées, seulement six comprennent des exons, les autres étant entièrement introniques, sauf *HHEX* qui est intergénique (voir tableau 22). Ainsi, ces tests n'ont pu être faits que pour *FABP2*, *PPARG*, *LEPR*, *KCNJ11*, *SLC30A8* et *PON1*. Cependant, la plupart de ces séquences (*PPARG*, *LEPR*, *SLC30A8* et *PON1*) ne possèdent pas suffisamment de mutations dans chaque catégorie pour que les tests puissent être réalisés. Ainsi, nous ne pouvons avoir les résultats que pour deux séquences (voir tableau 26 ci-dessous).

	FABP2	***PPARG***	***LEPR***	***KCNJ11***	***SLC30A8***	***PON1***
S_{MAX}	39,5	19,33	74,32	244,27	28,94	13,56
S_{OBS}	1	0	0	2	0	0
N_{MAX}	131,5	61,67	252,68	790,73	112,06	40,44
N_{OBS}	1	1	1	3	2	1
dN/dS	0,30	nd*	nd*	0,46	nd*	nd*

S_{MAX} : nombre maximal de mutations synonymes observables ; **S_{OBS}** : nombre observé de mutations synonymes ; **N_{MAX}** : nombre maximal de mutations non synonymes observables ; **N_{OBS}** : nombre observé de mutations non synonymes ;
*nd : Rapport dN/dS non disponible car il n'y a aucune mutation synonyme observée.

Tableau 26 : Nombre de mutations synonymes (S) et de mutations non synonymes (N) par séquence, et leur rapport. La proportion de mutations non synonymes dN (observées sur observables) est comparée à la proportion de mutations synonymes dS (observées sur observables).

Nous voyons que pour les séquences sur *FABP2* et *KCNJ11*, les rapports dN/dS sont inférieurs à 1, signe que les mutations synonymes sont proportionnellement en plus grand nombre que les mutations non synonymes. Ce résultat reflète l'action de la sélection négative

qui élimine les mutations entraînant un changement d'acides aminés. Le test de Mc Donald-Kreitman n'est significatif pour aucune de ces deux séquences, suggérant que nous n'avons pas assez de puissance pour ce test avec ces faibles nombres de mutations observées.

- Spectre de fréquences alléliques

Le spectre de fréquences alléliques cumulées pour toutes les séquences neutres et orientées par rapport au chimpanzé est présenté ci-dessous :

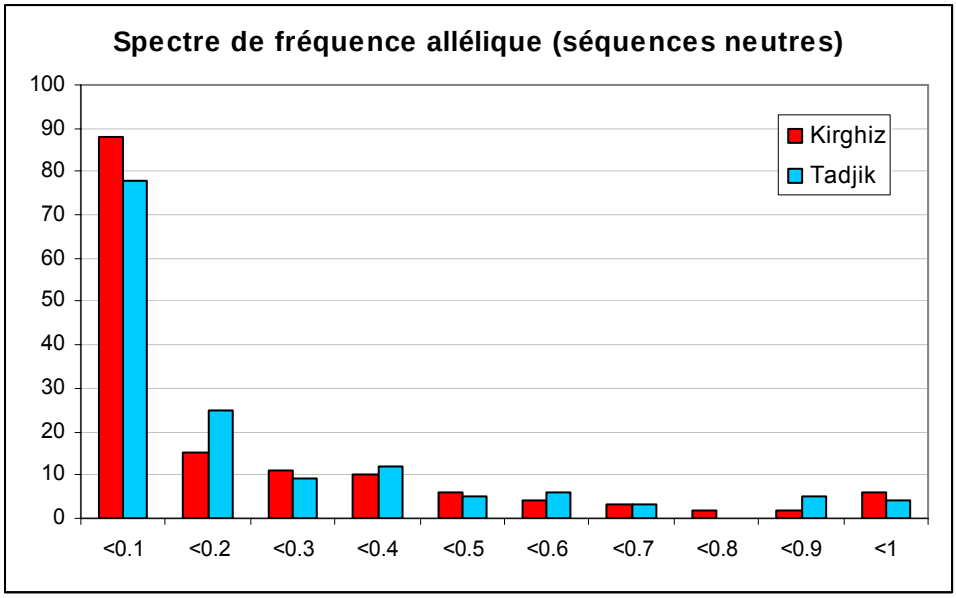


Figure 23: Spectre de fréquences alléliques pour les séquences neutres, orienté par rapport au chimpanzé, pour les Kirghiz et les Tadjiks, respectivement

Nous voyons donc sur les séquences neutres qu'il existe une forte proportion de mutations peu fréquentes, assez peu de mutations à fréquence intermédiaire, et une légère augmentation en fréquence pour les mutations dérivées presque fixées.

Le spectre de fréquences alléliques pour les séquences géniques, cumulées pour les différents gènes mais en gardant l'information de la contribution de chaque gène, et orientées par rapport au chimpanzé est présenté ci-dessous.

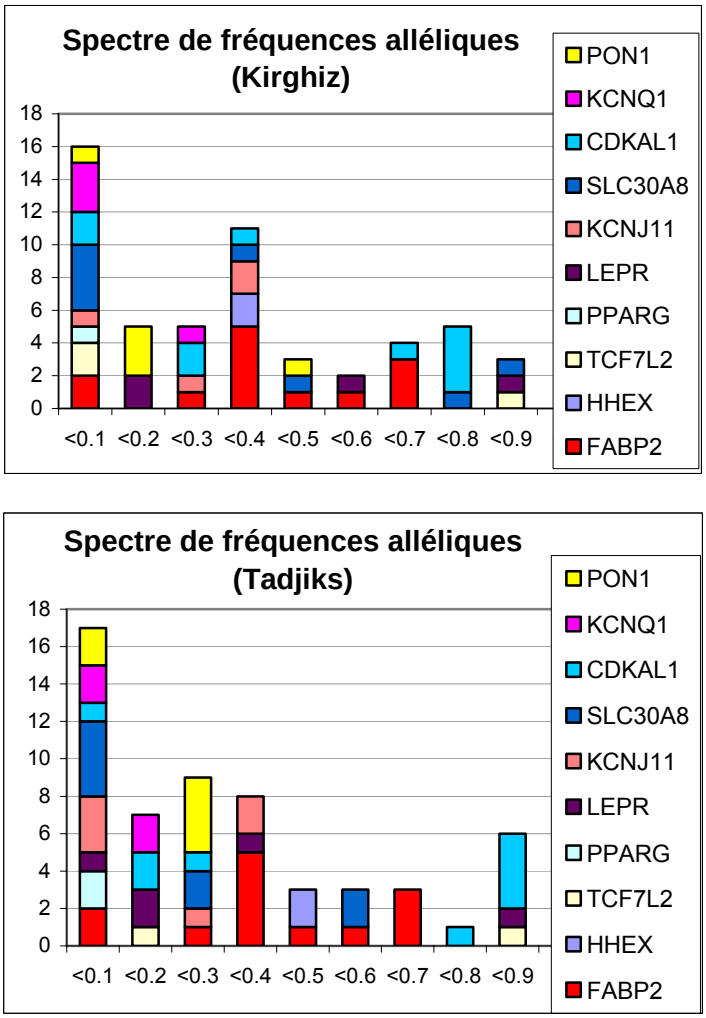


Figure 24 : Spectre de fréquence allélique des séquences géniques, avec la contribution de chaque séquence mise en évidence, pour les Kirghiz et les Tadjiks, respectivement

Sur les séquences géniques, il semble exister un excès de mutations à fréquence intermédiaire par rapport aux séquences neutres, surtout des fréquences entre 30 et 40% (16.5% sur les séquences géniques en moyenne par rapport à 3.9% sur les séquences neutres), au détriment des mutations à faible fréquence (25% de mutation à moins de 10% en fréquence pour les séquences géniques en moyenne par rapport à 44.3% pour les séquences neutres).

Nous avons représenté séparément les gènes *FABP2* et *CDKAL1* qui présentent tous deux des profils particuliers (voir figure 25) : *FABP2* présente un nombre considérable de mutations en fréquences intermédiaires ; *CDKAL1* présente un nombre important de

mutations à forte fréquence (dérivées). Pour tester la significativité de ces écarts au spectre de fréquences alléliques, nous allons effectuer des tests de neutralité.

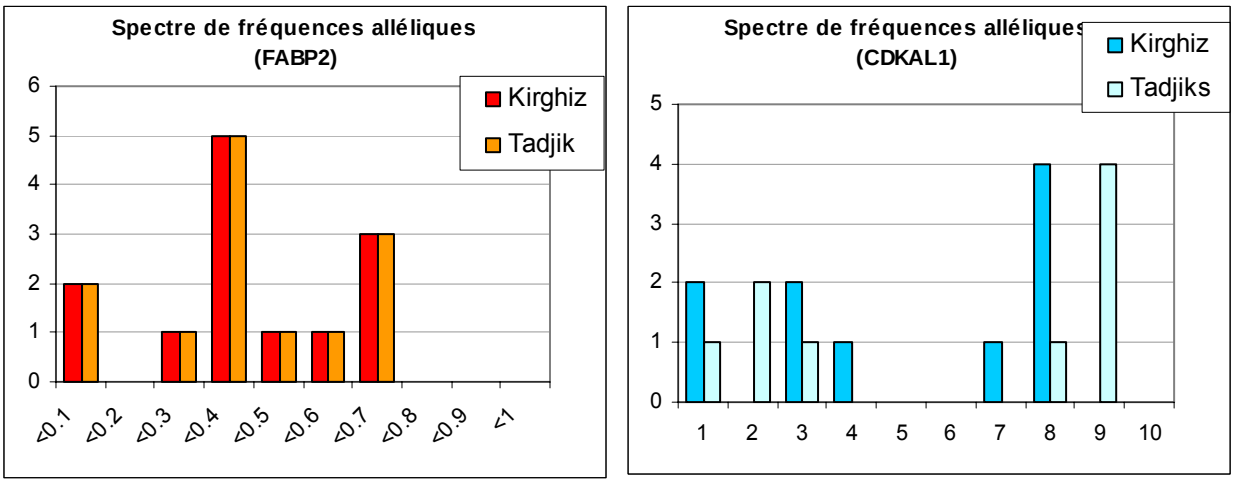


Figure 25 : Spectre de fréquences alléliques pour FABP2 et CDKAL1, pour chaque population

- Tests de neutralité basés sur le spectre de fréquences alléliques

Tout d'abord, le test d'homozygotie d'Ewens - Watterson (Ewens, 1972, Watterson, 1978), qui mesure directement l'écart entre le spectre de fréquences alléliques attendu et observé, n'est significatif pour aucune séquence, après correction pour les tests multiples (*p* entre 0.49 et 0.91). Ce résultat peut paraître étonnant vu l'écart observé entre le spectre des fréquences alléliques sur les séquences neutres et celui sur les séquences géniques. Cependant, ici chaque séquence est testée indépendamment, et ce test peut manquer de puissance.

Les résultats des autres tests de neutralité sont présentés, séparément pour les séquences neutres et géniques, et pour chaque population, dans les tableaux 27 à 30 ci-dessous.

Séquences neutres

Séquence	*D* de Tajima	*D* de Fu et Li	*D** de Fu et Li	*F* de Fu et Li	*F** de Fu et Li	*Fs* de Fu	*H* de Fay et Wu
N1	0,14	0,50	0,51	0,46	0,46	0,66	0,19
N2	0,37	0,72	0,71	0,71	0,70	2,40	-0,39
N3	1,55	-0,73	-0,46	-0,11	0,24	2,38	0,31
N4	0,27	0,01	0,02	0,11	0,12	-0,19	0,69
N5	0,19	-1,04	-1,00	-0,76	-0,73	-0,30	0,80
N6	0,04	0,85	0,96	0,82	0,78	0,72	0,59
N7	-0,66	0,01	0,03	-0,26	-0,24	0,76	-1,32
N8	0,10	-1,04	-1,02	-0,81	-0,79	0,31	-0,64
N9	0,66	0,89	0,87	0,97	0,95	-0,08	1,27
N10	-0,52	0,70	0,85	0,39	0,49	-0,85	0,25
N11	-0,13	1,15	1,14	0,86	0,85	0,71	-0,32
N12	1,52	1,07	1,06	1,43	1,42	1,15	0,84
N13	-0,78	-0,23	-0,21	-0,48	-0,46	-0,52	-1,32
N14	-1,46	-0,31	-0,28	-0,83	-0,80	-2,82	-1,00
N15	-1,26	0,97	0,96	0,30	0,30	-1,62	0,30
N16	0,19	0,37	0,37	0,36	0,36	-0,02	0,72
N17	1,04	0,02	0,03	0,41	0,42	1,55	0,33
N18	2,00	0,02	0,04	0,85	0,84	3,69	1,14
N19	2,12	0,98	0,97	1,58	1,56	2,98	-0,32
N20	-1,62	-1,42	-1,38	-1,75	-1,71	-3,02	-1,70
Moyenne	*0.19*	*0.17*	*0.21*	*0.21*	*0.24*	*0.39*	*0.02*

Tableau 27 : Résultats des tests de neutralité pour les séquences neutres chez les Kirghiz.
En orange : les tests significatifs à 5% avant correction pour les tests multiples ; en rouge, les tests significatifs à 5% après correction pour les tests multiples (FDR)

Chez les Kirghiz, la séquence N20 présente un *D* de Tajima significativement négatif (après correction pour les tests multiples), ainsi qu'un *Fs* de Fu et un *H* de Fay et Wu significativement négatifs. Cette séquence montre donc potentiellement un excès de variants jeunes, donc compatibles avec de la croissance démographique. Cependant, les moyennes des tests de neutralité sont chez les Kirghiz tous positifs, montrant ainsi plutôt une tendance vers la contraction. Les séquences N18 et N19 présentent un *D* de Tajima significativement positifs, mais ce résultat n'est trouvé pour aucun autre test.

Ces mêmes tests sur les séquences neutres des Tadjiks sont présentés dans le tableau ci-dessous :

Séquence	*D* de Tajima	*D* de Fu et Li	*D** de Fu et Li	*F* de Fu et Li	*F** de Fu et Li	*Fs* de Fu	*H* de Fay et Wu
N1	0,04	-1,02	-1,00	-0,82	-0,80	0,236	0,2786
N2	1,71	0,62	0,61	1,18	1,17	5,057	0,19447
N3	1,55	-0,73	-0,46	-0,12	0,24	1,505	0,20772
N4	-0,73	-2,37	-2,27	-2,15	-2,07	-3,042	0,85648
N5	0,28	0,02	0,03	0,12	0,13	-0,188	0,80186
N6	-0,26	1,07	1,14	0,81	0,80	-0,349	0,78055
N7	-0,84	0,02	0,04	-0,33	-0,31	0,441	-1,60906
N8	1,58	0,51	0,51	0,96	0,96	2,017	-0,34802
N9	-0,75	-1,27	-1,65	-1,30	-1,58	-2,851	-1,19347
N10	-1,01	-1,03	-0,70	-1,18	-0,94	-2,784	0,47088
N11	-0,30	0,21	0,22	0,05	0,06	0,461	-0,54878
N12	1,97	0,97	0,96	1,52	1,51	1,782	0,85781
N13	-0,85	-1,04	-1,00	-1,15	-1,12	-1,025	-0,95638
N14	-1,55	-0,50	-0,47	-1,01	-0,98	-1,546	-1,45854
N15	-1,21	-1,03	-1,01	-1,27	-1,25	-2,42	0,09724
N16	0,23	0,21	0,22	0,25	0,26	0,316	0,49417
N17	1,84	-0,22	-0,20	0,51	0,52	2,735	0,22246
N18	1,95	0,02	0,04	0,83	0,82	3,631	1,09357
N19	1,11	0,22	0,23	0,60	0,61	1,293	-0,46797
N20	0,06	0,85	0,85	0,71	0,70	1,156	-0,77256
Moyenne	*0.24*	*-0.23*	*-0.20*	*-0.09*	*-0.06*	*0.32*	*-0.05*

Tableau 28 : Résultats des tests de neutralité pour les séquences neutres chez les Tadjiks.
En orange : les tests significatifs à 5% avant correction pour les tests multiples (FDR)

Chez les Tadjiks, la séquence N4 présente un *D* de Fu et Li significativement négatif et les séquences N14 et N15 un *D* de Tajima significativement négatif, mais ces résultats ne sont pas retrouvés pur d'autres tests de neutralité. Les moyennes des tests de neutralité sont ici plutôt négatives, montrant ainsi une tendance vers la croissance démographique.

Séquences géniques

Les tests de neutralité sur les séquences géniques chez les Kirghiz sont présentés ci-dessous :

Séquence	*D* de Tajima	*D* de Fu et Li	*D** de Fu et Li	*F* de Fu et Li	*F** de Fu et Li	*Fs* de Fu	*H* de Fay et Wu
FABP2	2,88	1,52	1,47	2,40	2,33	9,32	-0,09
TCF7L2	-0,80	-0,55	-0,53	-0,74	-0,72	-1,38	-1,30
PARG	-0,12	0,50	0,51	0,37	0,37	0,34	0,16
LEPR	1,08	0,97	0,96	1,18	1,17	1,95	-1,01
KCNQ1	1,42	-0,23	-0,21	0,35	0,36	2,33	0,68
SLC30A8	0,18	-1,91	-1,83	-1,42	-1,36	-0,30	-1,40
HHEX	2,18	0,71	0,71	1,35	1,35	4,14	0,37
CDKAL1	1,98	0,72	0,71	1,38	1,36	6,24	-2,02
KCNQ1	-0,22	0,97	0,96	0,70	0,69	-0,62	0,61
PON1	1,00	1,07	1,06	1,23	1,22	2,60	0,91
Moyenne	*0.96*	*0.38*	*0.38*	*0.68*	*0.68*	*2.46*	*-0.31*

Tableau 29 : Résultats des tests de neutralité pour les séquences géniques chez les Kirghiz.
En orange : les tests significatifs à 5% avant correction pour les tests multiples ; en rouge, les tests significatifs à 5% après correction pour les tests multiples (FDR)

Nous voyons ici que la séquence sur *FABP2* présente plusieurs tests significativement positifs ; *D* de Tajima, *D**, *F* et *F** de Fu et Li, et *Fs* de Fu, bien que le *D* de Tajima soit le seul encore significatif après correction pour les tests multiples. Ce manque de puissance statistique peut être lié à la trop petite taille des séquences (1229pb pour *FABP2*) et du nombre limité de mutations qui en découle. Cependant, le fait de présenter plusieurs tests significatifs avant correction pour les tests multiples nous laisse penser que cette séquence présente bien un excès significatif de variants à fréquence intermédiaire, ou anciens. Ainsi, *FABP2* semble présenter une déviation à la neutralité et être sous l'action de la sélection balancée, ou de la sélection positive partielle (augmentation en fréquence d'un haplotype différent des autres mais qui n'a pas encore atteint une fréquence forte et est donc en excès à fréquence intermédiaire). Deux autres séquences présentent également un *D* de Tajima significativement positif avant correction pour les tests multiples (*HHEX* et *CDKAL1*), mais ces séquences ne présentent pas d'autres tests significatifs.

Les tests de neutralité sur les séquences géniques chez les Tadjiks sont présentés ci-dessous :

Séquence	*D* de Tajima	*D* de Fu et Li	*D** de Fu et Li	*F* de Fu et Li	*F** de Fu et Li	*Fs* de Fu	*H* de Fay et Wu
FABP2	2,48	1,00	0,97	1,82	1,77	5,77	0,15
TCF7L2	0,52	0,71	0,71	0,76	0,76	0,78	-0,90
PARG	-0,84	-1,02	-1,00	-1,13	-1,11	-1,22	0,18
LEPR	0,49	0,02	0,03	0,20	0,21	0,92	-0,50
KCNQ1	0,31	-1,67	-1,61	-1,20	-1,15	0,39	0,70
SLC30A8	0,21	-1,91	-1,83	-1,41	-1,35	0,46	0,30
HHEX	2,39	0,71	0,71	1,42	1,42	4,33	0,10
CDKAL1	1,35	1,37	1,33	1,62	1,58	5,99	-3,95
KCNQ1	-0,64	0,97	0,96	0,54	0,54	-0,26	0,49
PON1	1,14	0,21	0,22	0,61	0,61	2,29	1,15
Moyenne	*0.74*	*0.04*	*0.05*	*0.32*	*0.33*	*1.95*	*-0.23*

Tableau 30 : Résultats des tests de neutralité sur les séquences géniques chez les Tadjiks. En orange : les tests significatifs à 5% avant correction pour les tests multiples ; en rouge, les tests significatifs à 5% après correction pour les tests multiples (FDR)

La séquence *FABP2* présente ici aussi plusieurs tests significativement positifs (*D* de Tajima, *F* et *F** de Fu et Li), bien qu'aucun ne soit significatif après correction pour les tests multiples. *FABP2* semble donc également être compatible avec de la sélection balancée, ou de la sélection positive partielle, mais avec moins de robustesse que chez les Kirghiz. Nous retrouvons également un *D* de Tajima significativement positif pour *HHEX*, pour qui aucun autre test n'est cependant significatif. Ce gène présente donc éventuellement une tendance vers un excès de fréquences intermédiaires, mais nous ne pouvons rien conclure avec ces résultats seuls.

6) Différenciation entre population

Sous neutralité, les fréquences alléliques varient entre populations du fait notamment de leur histoire démographique, mais tous les locus présentent en espérance le même niveau de différenciation génétique. Cet attendu peut être utilisé pour détecter la sélection naturelle qui, contrairement à la démographie, affecte toujours certaines régions du génome plus que d'autres (Cavalli-Sforza, 1966). Certaines approches, basées sur l'estimation de la distribution du polymorphisme neutre d'après les données observées, permettent ainsi d'identifier les régions du génome sous sélection positive (Beaumont & Nichols, 1996, Vitalis *et al.*, 2001, Beaumont & Balding, 2004, Beaumont, 2005, Pollinger *et al.*, 2005).

Tandis que les premières approches précédemment présentées basées sur des données de séquence nous permettent d'analyser la variation au niveau intra-populationnel, ces approches basées sur le niveau de différenciation permettent de détecter des pressions de sélection différentielles entre populations (par exemple de l'adaptation locale). L'idée sous-jacente est que si la sélection favorise un ou plusieurs allèles dans une population (et pas dans l'autre), alors la différenciation observée à ce locus devrait être plus élevée que celle observée aux locus neutres.

Reprenant la méthodologie utilisée dans l'annexe 5, nous avons comparé le niveau de différenciation génétique attendu sous neutralité (obtenu par simulation dans un modèle en île à partir de paramètres démographiques déduits des 20 séquences neutres) à celui pour chaque mutation génique. Nous utilisons l'approche de Beaumont et Nichols (Beaumont & Nichols, 1996), c'est-à-dire que nous regardons le niveau de différenciation (F_{ST}) en fonction de l'hétérozygotie totale (H_e) pour détecter les mutations en marge de la distribution neutre. Les résultats sont présentés dans la figure 26 ci-dessous :

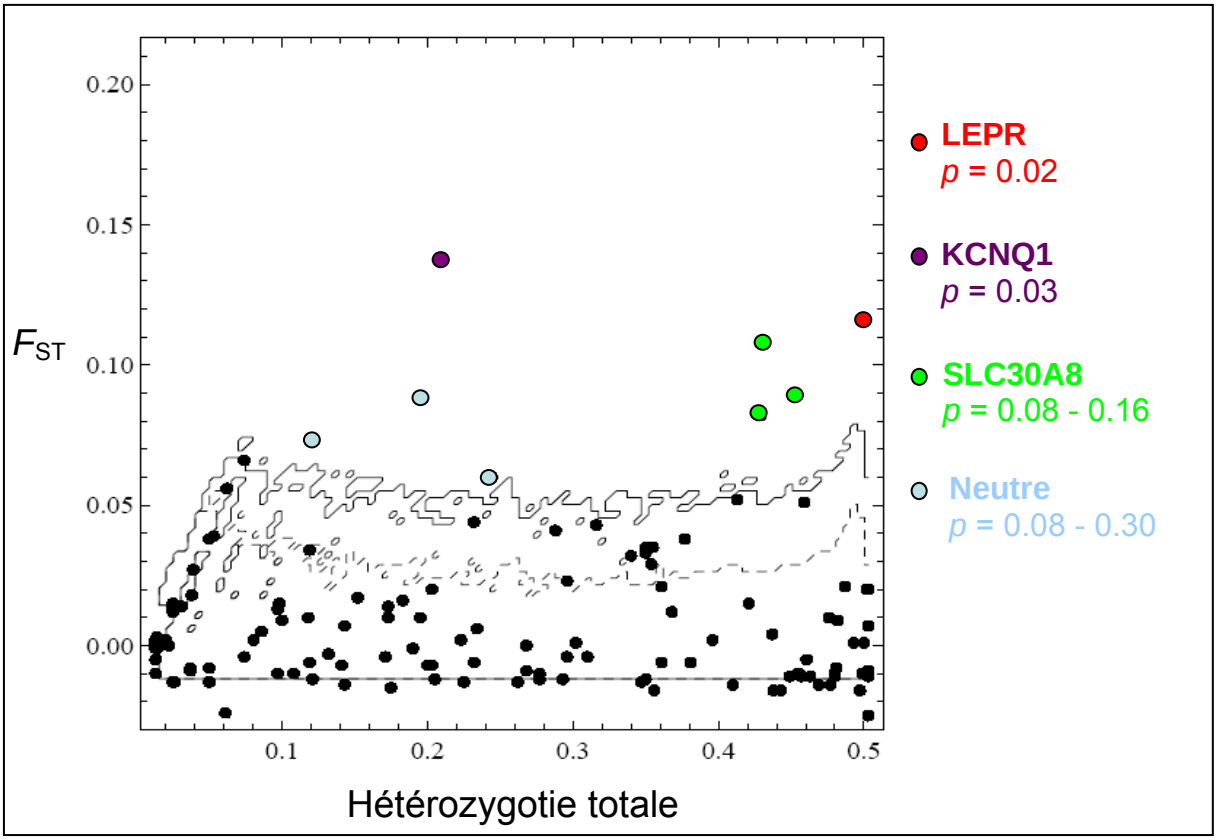


Figure 26: Niveau de différenciation mesuré par le F_{ST} en fonction de l'hétérozygotie totale, à la fois pour les mutations sur les 20 séquences neutres et celles sur les 10 séquences géniques. Les mutations qui ne sortent pas de l'enveloppe à 99% sont en noir tandis que les mutations significativement en dehors de l'enveloppe sont représentées par des points de diverses couleurs. Les *p*-valeurs corrigées pour les tests multiples sont présentées dans la légende.

Nous voyons donc que 5 mutations géniques et 3 mutations neutres sortent de l'enveloppe à 99% bien que seulement deux mutations présentent une *p*-valeur significative après correction pour les tests multiples (FDR). Il s'agit des mutations sur les gènes *KCNQ1* (F_{ST} = 0.138, p = 0.03) et *LEPR* (F_{ST} = 0.116, p = 0.02). Ces mutations géniques présentent donc des niveaux de différenciation particulièrement forts par rapport aux séquences neutres, signature de la sélection positive divergente (adaptation locale), c'est-à-dire favorisant des allèles différents entre populations.

7) Conclusion

Nous avons tout d'abord vu que les facteurs de risque génétiques pour le diabète de type II ne sont pas en plus forte fréquence chez les Kirghiz par rapport aux Tadjiks, bien qu'ils ont 1.8 fois plus de risque de présenter de la résistance à l'insuline. Ainsi, les mutations étudiées ne sont pas forcément celles qui sont le plus fortement associées au phénotype dans les populations d'Asie Centrale.

D'après les données de divergence par rapport au chimpanzé, nous pouvons voir que les régions géniques sont en moyenne plus conservées que les régions neutres, donc *a priori* sous l'action de la sélection négative. Cette hypothèse est également appuyée par les faibles rapports entre mutations synonymes et non synonymes observés pour *FABP2* et *KCNJ11*.

Les tests de neutralité basés sur les spectres de fréquences alléliques nous montrent de plus que *FABP2* est *a priori* sous sélection balancée, avec des valeurs significativement positives chez les Kirghiz pour le *D* de Tajima, le *D**, *F* et *F** de Fu et Li, et le *Fs* de Fu, et chez les Tadjiks pour le *D* de Tajima et le *F* de Fu et Li. Ce gène montre également une diversité haplotypique et nucléotidique forte (H_{div} = 0.70 et 0.75 pour les Kirghiz et les Tadjiks, respectivement, et π = 0.004 pour les deux ethnies), un F_{ST} faible (F_{ST} = -0.011), et un spectre de fréquences alléliques biaisé pour les fréquences intermédiaires (mais non significatif selon le test d'Ewens-Watterson, p = 0.12 et 0.25 pour les Kirghiz et les Tadjiks, respectivement). Ce gène présente donc a priori des forts signes de sélection dans les deux populations, avec peut être une intensité de sélection plus forte chez les Kirghiz. Il va s'agir maintenant d'étudier les arbres de coalescence et la structure des haplotypes pour comprendre quelles pressions sélectives ont opéré sur ces gènes (sélection balancée ou sélection positive partielle). Le séquençage d'une région plus longue du gène FABP2 pourrait également être du plus grand intérêt pour gagner en puissance statistique pour les tests de neutralité basés sur spectre de fréquence allélique et pour essayer d'identifier les SNPs les mieux associés à la pression de sélection (en regardant par exemple l'évolution du *D* de Tajima le long de la séquence). Nous aimerions également caractériser l'intensité de sélection dans chaque population pour voir si les deux ethnies présentent des histoires adaptatives différentes, et dater le début de l'expansion de cette mutation pour comprendre à quelle période correspond cette pression de sélection.

Les tests de différenciation génétique nous montrent de plus que les gènes *KCNQ1* et *LEPR* ont une différenciation plus forte qu'attendue sous neutralité (F_{ST} = 0.138 et 0.116, respectivement). Ainsi, ces gènes présentent une signature caractéristique de sélection positive différentielle entre les populations. Cependant, il est assez étonnant que, malgré ce signal fort d'adaptation locale, nous ne trouvions aucune déviation au spectre de fréquence allélique pour ces gènes, d'après les tests de neutralité.

Nous aimerions donc maintenant savoir dans quelle population la sélection a agit (ou est la plus forte) et quel est l'allèle sélectionné (allèle à risque ou protecteur). Il serait également intéressant de dater l'expansion de la mutation sous pression de sélection pour comprendre à quel moment ces populations ont subi des pressions de sélection différentes.

D. Discussion

Les hypothèses du génotype économe et de la piste carnivore que nous cherchons à tester peuvent être résumées par plusieurs prédictions que nous pouvons confronter à nos données : i) les prévalences de résistance à l'insuline sont différentes entre ethnies pour un milieu équivalent ; ii) ces différences de prévalence s'expliquent par des différences de susceptibilité génétique ; iii) ces différences génétiques sont dues à l'action de la sélection naturelle qui favorise des allèles à risque pour le diabète de type II dans certaines populations ; et iv) ces pressions de sélection sont différentes entre ethnies aux modes de vie contrastés, notamment plus importantes dans les ethnies ayant dans le passé subi le plus de famines (hypothèse du génotype économe) et / ou ayant un régime alimentaire ancestral pauvre en glucides (hypothèse de la piste carnivore).

Différences phénotypiques entre ethnies

Nos résultats suggèrent, comme attendu, que les prévalences observées de résistance à l'insuline sont fortement influencées par le poids (facteur de risque de 3 et de 8.9, respectivement pour les individus pré-obèses et obèses) et l'activité physique (facteur de risque de 2.3 pour les inactifs). Mais nous trouvons également une influence importante des effets populationnels, avec un facteur de risque de 1.8 pour les Kirghiz (éleveurs) comparés aux Tadjiks (agriculteurs), en contrôlant pour les co-variables confondantes. Ces résultats sont donc concordants avec la première prédiction des hypothèses testées. Le fait que les éleveurs aient une prévalence plus importante que les agriculteurs est de plus spécifiquement compatible avec l'influence de la quantité de glucides dans le régime alimentaire, comme attendu sous l'hypothèse de la piste carnivore. Etant donné que nous avons également effectué des dosages physiologiques de triglycérides et de cholestérol dans le sang, ainsi que mesuré la tension, nous pourrons refaire les mêmes analyses phénotypiques à partir du syndrome métabolique, combinaison de déséquilibres métaboliques (résistance à l'insuline, obésité, hypertriglycéridémie, hypercholestérolémie et hypertension) qui augmentent fortement le risque de développer des maladies cardiovasculaires et du diabète. Nos premières analyses montrent que l'étude de ce phénotype aboutit exactement aux mêmes conclusions qu'avec la résistance à l'insuline.

Facteurs de risque génétiques ?

La deuxième prédiction des hypothèses testées est que ces effets populationnels sont liés à des différences génétiques. Ainsi, les facteurs de risque génétiques associés au diabète de type II devraient être inégalement répartis entre ethnies, avec notamment un facteur de risque génétique plus élevé chez les Kirghiz que chez les Tadjiks.

Avant de tester cette hypothèse, nous pouvons comparer le facteur de risque lié à l'effet populationnel estimé dans cette étude (1.8) aux facteurs de risque mis en évidence par les études génétiques. Les études d'association révèlent généralement des facteurs de risque compris entre 1.15 et 1.25 par mutation associée au diabète de type II (pour une revue, voir Prokopenko *et al.*, 2008). Il existe cependant quelques exceptions, dont notamment *TCF7L2* (facteur de risque de 1.64 chez les Sikhs, Sanghera *et al.*, 2008, , et d'environ 1.4 dans d'autres études, Prokopenko *et al.*, 2008). Les études prenant en compte des facteurs de risque cumulés pour une combinaison de mutations associées au diabète de type II montrent qu'un facteur de risque de 1.8 (comme l'effet estimé dans nos analyses) correspond à environ de 8 à 12 allèles à risque cumulés (Miyake *et al.*, 2009, Cauchi *et al.*, 2008). Donc l'effet populationnel obtenu dans notre étude pourrait correspondre à une dizaine d'allèles à risque présent uniquement chez les Kirghiz et non chez les Tadjiks.

En mesurant la fréquence de 10 allèles à risque dans 10 gènes différents en Asie Centrale, nous avons cependant trouvé qu'il y a une tendance vers un plus grand nombre d'allèles à risque chez les Tadjiks. Ce résultat peut être dû à un biais de détection des mutations en plus forte fréquence dans les populations européennes, où sont majoritairement effectuées les études d'association. De manière importante, ce manque de corrélation entre les données phénotypiques et génétiques peut signifier que les mutations que nous avons choisies n'ont pas d'effets phénotypiques forts dans les populations d'Asie Centrale, particulièrement chez les Kirghiz. Ainsi, il est possible que nous sous-estimions les intensités de sélection pouvant exister sur les mutations directement associées au diabète de type II dans ces populations.

Il est également possible que, prises individuellement (comme dans les études d'association), ces mutations aient bien des effets sur le phénotype en Asie Centrale, mais que d'autres facteurs soient en fait majoritairement responsables des différences phénotypiques

observées. Notamment, quelques mutations à fort effet (et à faible fréquence) inégalement réparties entre Kirghiz et Tadjiks pourraient expliquer une majorité de la variation phénotypique. Ces mutations ne seraient pas détectées par les études d'association car i) les études d'association testent des mutations préalablement cataloguées dans les bases de données qui ne recensent souvent que des mutations à fréquence supérieure à 5% et ii) les mutations à faible fréquence ont très peu de chances d'être significativement associées à un phénotype du fait d'un manque de puissance statistique.

Certains auteurs ont également proposé que les mutations expliquant les différences phénotypiques à l'intérieur des populations (donc détectées par études d'association) ne soient pas les mêmes que celles responsables des différences de prévalences entre populations, comme avec le gène *MC1R* responsable de la pigmentation de la peau (Myles *et al.*, 2007).

Egalement, nous pouvons imaginer que d'autres facteurs soient responsables des différences phénotypiques de diabète de type II observées, comme des CNV (*Copy Number Variation*), souvent associés à des maladies chez l'Homme (Orozco *et al.*, 2009, Conrad *et al.*, 2009, Frazer *et al.*, 2009), ou encore des facteurs non génétiques responsables de plasticité phénotypique (variations épigénétiques ou autres adaptations phénotypiques, Kaput *et al.*, 2007, Tobi *et al.*, 2009).

Finalement, pour départager ces hypothèses, le plus informatif serait d'effectuer des études d'association au sein des populations d'Asie Centrale, pour pouvoir identifier et directement tester la présence de sélection sur d'autres mutations qui seraient fortement associées au phénotype de résistance à l'insuline. Cependant, même s'il existe d'autres mutations plus fortement associées au phénotype, nous pouvons tout de même, par une approche de génétique des populations, tester avec nos données les dernières prédictions des hypothèses initiales : iii) si certains allèles à risque pour le diabète de type II présentent des signes de sélection naturelle, et iv) s'ils sont différents entre populations aux modes de subsistance contrastés.

Sélection d'allèles à risque liés au diabète de type II ?

Sur dix gènes candidats associés au diabète de type II, nous avons trouvé des signes de déviation à la neutralité pour au moins trois gènes (*FABP2*, *KCNQ1* et *LEPR*). Pour le premier gène, la pression de sélection est commune aux deux populations, mais son intensité est peut-être plus importante chez les Kirghiz (éleveurs). Bien que l'allèle à risque présente une tendance vers une plus forte fréquence chez les Kirghiz, nous ne savons pas si c'est cet allèle qui est sélectionné chez les Kirghiz, ou l'allèle protecteur qui est sélectionné chez les Tadjiks. De plus, nous ne savons pas quand ont lieu ces pressions de sélection dans le temps, ce qui est fondamental pour savoir si les hypothèses du génotype économe et de la piste carnivore sont supportées par ces données. En effet, ces hypothèses sont compatibles avec une sélection ancienne de l'allèle à risque ou une sélection récente de l'allèle protecteur, mais ne sont pas compatibles avec une sélection ancienne de l'allèle protecteur ou récente de l'allèle à risque. Finalement, nous ne savons pas si le gène *FABP2* présente des signes de sélection balancée ou plutôt de sélection positive partielle.

Pour les deux derniers gènes, les pressions de sélection mises en évidence entraînent une différenciation importante entre ethnies aux modes de vie contrastés, en accord avec l'idée d'une adaptation locale à l'alimentation. Cependant, les allèles à risque sont en plus fortes fréquences chez les Tadjiks. Il est donc probable i) que ces allèles à risque soient sous plus forte sélection chez les Tadjiks, ou ii) que les allèles protecteurs soient sous plus forte sélection chez les Kirghiz. Dans le premier cas, nous retrouvons l'idée de sélection d'allèles à risque, mais dont les causes seraient différentes de celles proposées par les hypothèses du génotype économe et de la piste carnivore (puisque ayant lieu chez des agriculteurs). Dans le deuxième cas, la sélection favoriserait les allèles protecteurs, ce qui n'est pas compatible avec les hypothèses testées.

Finalement, les approches utilisées ici (tests de neutralité basés sur le spectre de fréquence allélique et tests de différenciation) ne nous permettent pas de conclure sur la contribution des données génétiques au débat toujours en cours sur les hypothèses du génotype économe et de la piste carnivore. En effet, seule l'identification de l'allèle sous sélection (à risque ou protecteur), l'estimation des intensités de sélection dans chaque population et la datation de ces pressions de sélection nous permettront de comprendre les processus évolutifs en jeu. Nous aimerions donc maintenant utiliser d'autres approches basées

sur des données haplotypiques, qui permettraient justement d'obtenir les informations complémentaires à notre étude. En effet, deux tests basés sur la longueur d'haplotypes (Voight *et al.*, 2006, Sabeti *et al.*, 2007) permettent de détecter la sélection positive grâce à la présence de déséquilibre de liaison autour de la mutation sous sélection. Ce déséquilibre de liaison est comparé entre les haplotypes portant la mutation d'intérêt et les autres, pour corriger pour les variations de taux de recombinaison entre séquences. Ainsi ces approches permettent de tester quels sont les haplotypes sous sélection, et dans quelle population. La méthode développée par Austerlitz *et al* (2003) précédemment utilisée dans la deuxième partie du chapitre I sur la lactase permettrait également à partir de données haplotypiques d'estimer l'intensité de la sélection dans chaque population, et surtout le temps depuis lequel ces mutations sont en expansion. C'est pourquoi nous avons initié le génotypage d'un grand nombre de marqueurs génétiques pour ces mêmes 80 individus, grâce à la technologie de puce à ADN. Nous avons ici utilisé la puce Illumina 660W Quad, qui comprend 660 000 SNPs et CNVs. Ce génotypage appelé « à haut débit » a été effectué en collaboration avec la plateforme de séquençage de la Pitié Salpetrière dirigée par Wassila Carpentier. Les analyses statistiques sont en cours et sont effectuées en collaboration avec Christine Lonjou de cette même équipe.

L'étude des données haplotypiques obtenues par génotypage par puce à ADN nous permettra donc d'effectuer de nouveaux tests de sélection pour avancer dans la compréhension des processus évolutifs en jeu. Ces données nous permettront également de tester un champ plus large de mutations associées au diabète et donc éventuellement d'éviter le biais du choix des mutations en faveur des plus fréquentes chez les Tadjiks. Finalement, les taux de croissance démographique de chaque population pourront être précisément inférés sur des régions neutres du génome, ce qui nous permettra de distinguer au niveau des mutations d'intérêt la part de la croissance due à la démographie et celle due à la sélection.

Données de la littérature

En attendant de compléter nos résultats par de nouvelles données, nous pouvons dans un premier temps nous appuyer sur les données de la littérature, et notamment sur les données dans les populations de HapMap ou du HGDP-CEPH, pour nous aider à interpréter nos résultats.

- Pour le gène *LEPR*, les F_{ST} sont particulièrement forts entre l'Asie et l'Europe et entre l'Asie et l'Afrique (Voight *et al.*, 2006), et les tests basés sur la longueur d'haplotypes partagés révèlent des signes forts de sélection positive de l'allèle protecteur du diabète de type II, exclusivement en Asie de l'Est (Cheng *et al.*, 2009, Pickrell *et al.*, 2009). Sachant que les populations d'HapMap et du CEPH sont majoritairement constituées d'agriculteurs, nous pouvons imaginer que l'allèle protecteur est sous sélection récente dans ces populations (c'est-à-dire depuis le Néolithique). Ainsi, ces résultats peuvent être réconciliés avec l'hypothèse du génotype économe et de la piste carnivore. Cependant, si ces allèles protecteurs sont sous sélection depuis plus longtemps, ces résultats ne sont plus en accord avec ces hypothèses. Il serait donc intéressant d'appliquer la méthode d'Austerlitz *et al* (2003) sur ces données.

- Les autres gènes que nous avons identifiés sous sélection (*FABP2* et *KCNQ1*) ne présentent pas de signatures caractéristiques de la sélection naturelle dans les populations du CEPH et d'HapMap, et n'ont pas non plus été révélés par d'autres études. Nous pouvons noter que *KCNQ1* est un des seuls gènes à avoir été identifié en premier lieu chez les asiatiques, et non chez les européens.

- Le gène *TCF7L2*, qui ne présente pas de déviation à la neutralité dans notre étude, est pourtant un des gènes où la signature de la sélection est la plus souvent retrouvée. L'étude d'Helgason *et al* (2007) sur les populations de HapMap et celle de Pickrell *et al* (2009) sur les populations du CEPH ont trouvé de forts F_{ST} sur ce gène (0.306 pour un haplotype dans les populations de HapMap), majoritairement dus à l'écart de fréquences entre les populations d'Asie de l'Est et les deux autres. Les données d'Helgason *et al* (2007) montrent de plus une signature de sélection positive particulièrement forte en Asie de l'Est, sur un haplotype portant un allèle protecteur. Les estimations du début de l'expansion de cet haplotype sont de 11 933, 8 401 et 4 051 ans pour les populations européennes, est-asiatiques et africaines, respectivement. Ces résultats sont donc compatibles avec un changement de pression de sélection sur les gènes liés au diabète de type II au moment du Néolithique, et notamment une diminution de la fréquence des allèles à risque à cette période.

Conclusion

En résumé, notre étude a permis de montrer que les prévalences phénotypiques de résistance à l'insuline sont presque deux fois plus importantes chez les Kirghiz par rapport aux Tadjiks. De plus, nous avons identifié trois gènes associés au diabète de type II sous sélection, un sous sélection balancée (ou positive partielle) commune aux deux ethnies (*FABP2*), et deux autres sous sélection positive locale (*LEPR* et *KCNQ1*). Ainsi, ces derniers gènes sont différemment sélectionnés entre ethnies aux modes de subsistance contrastés. De plus, les gènes associés au diabète de type II et identifiés comme sous sélection dans la littérature montrent plutôt des signes de sélection pour les allèles protecteurs (notamment *LEPR* et *TCF7L2*, Cheng *et al.*, 2009, Helgason *et al.*, 2007, Pickrell *et al.*, 2009). Ces résultats peuvent être conciliés avec les hypothèses du génotype économe et de la piste carnivore si les allèles à risque de ces gènes ont d'abord été fortement sélectionnés dans le passé, puis que les allèles protecteurs ont été favorisés plus récemment, au moment de la révolution Néolithique. Ces résultats peuvent être confirmés i) en datant les pressions de sélection par la méthode d'Austerlitz sur nos données pour *LEPR* et *KCNQ1,* et sur les données de la littérature pour *LEPR* et *TCF7L2* et ii) en comparant les intensités de pressions de sélection chez les éleveurs et chez les agriculteurs d'Asie Centrale pour tous ces gènes, pour tester l'influence des modes de subsistance passés sur ces pressions de sélection.

Pour *FABP2*, bien que nous ne puissions pas pour l'instant savoir quel allèle est sélectionné, ni à quelle période, nous pouvons imaginer que la signature observée d'une sélection balancée, ou d'une sélection positive partielle, soit justement le reflet d'un changement de pressions de sélection, depuis un avantage pour l'allèle à risque dans le passé vers un avantage pour l'allèle protecteur au Néolithique. Dans tout les cas, les tests haplotypiques nous permettront enfin de trancher parmi toutes les alternatives possibles.

V. CONCLUSIONS ET PERSPECTIVES

Nous avons vu que plusieurs approches sont possibles pour détecter de la sélection : des approches intra-population basées sur la déviation du spectre de fréquences alléliques attendu sous neutralité (détection de sélection positive et balancée, Watterson, 1975, Fu & Li, 1993, Fu, 1997, Tajima, 1989), des approches inter-population basées sur la comparaison du niveau de différenciation de marqueurs génétiques entre marqueurs neutres et marqueurs d'intérêt (détection de sélection locale, (détection de sélection locale, Beaumont & Nichols, 1996, Beaumont, 2005), et des approches basées sur les longueurs d'haplotypes partagés, permettant de connaître la population et l'allèle sous sélection, ainsi que la date de début d'expansion (Austerlitz *et al.*, 2003, Voight *et al.*, 2006, Sabeti *et al.*, 2007). Ces approches peuvent être utilisées sur des gènes candidats, comme nous l'avons fait dans cette étude, ou indifféremment sur tout le génome (approche de « scan génomique », Storz, 2005), pour identifier les classes de gènes les plus touchées par la sélection.

Approches utilisées dans cette étude

Dans cette étude, nous avons utilisé sur des gènes candidats la plupart des approches méthodologiques précédemment citées, en partant de données moléculaires variées : SNPs ponctuels obtenus par RFLP (*AGXT* et lactase), SNPs ponctuels obtenus par génotypage (lactase), CNV (*Copy Number Variation*, pour l'étude en cours sur l'amylase), séquençage (pour les gènes du diabète de type II) et SNPs obtenus par génotypage « dense », c'est-à-dire permettant de reconstruire des haplotypes (pour la lactase et pour l'étude en cours des données issues de la puce à ADN). Chaque type de données moléculaires représente un compromis entre le temps et le coût nécessaire à l'obtention des données, et l'apport d'information par rapport à la question posée. Les données de SNPs ponctuels et de CNV sont assez rapides à obtenir mais ne peuvent être interprétées que par des approches basées sur le niveau de différenciation inter-population. Les données de séquençage sont plus longues à obtenir mais permettent, en plus des tests basés sur la différenciation, d'adopter des approches intra-population et donc de comprendre dans quelle population la sélection agit, grâce à des tests basés sur les spectres de fréquences alléliques. Les données haplotypiques, si l'on considère la méthode de puce à ADN, sont assez rapides à obtenir mais plus chères, et elles permettent, en plus des tests de différenciation, d'effectuer des tests intra-population basés sur les longueurs d'haplotypes partagés, les seuls permettant de dater l'expansion de la

mutation sélectionnée. Chaque approche comporte donc des avantages et inconvénients, et souvent ces méthodes sont complémentaires.

Adaptation génétique locale aux modes de subsistance ?

Nous avons donc utilisé toutes ces approches sur des populations d'éleveurs et d'agriculteurs en Asie Centrale pour comprendre s'il existe des adaptations génétiques locales au mode de subsistance. Tout d'abord, le gène de la lactase, bien qu'*a priori* sous sélection dans ces populations, ne présente pas de différence significative de pressions de sélection entre éleveurs et agriculteurs dans cette région du monde, certainement à cause de pratiques culturelles de l'utilisation du lait qui amoindrissent l'avantage de digérer le lactose à l'âge adulte chez les éleveurs. Ensuite, le gène *AGXT* ne présente pas de déviation significative à l'attendu sous neutralité en Asie Centrale et plus largement dans un échantillon mondial, contredisant l'hypothèse d'une adaptation de ce gène à la consommation de viande chez les éleveurs. Finalement, la diversité génétique de certains gènes liés au diabète de type II s'écarte significativement de l'attendu sous neutralité en Asie Centrale (*LEPR*, *KCNQ1* et *FABP2*), avec certaines pressions de sélection clairement différentes entre ethnies aux modes de vie contrastés (*KCNQ1* et *FABP2*). Cependant, d'autres tests basés sur des données haplotypiques sont nécessaires pour comprendre les processus évolutifs en jeu. Dans la littérature, les pressions de sélection identifiées pour *LEPR* et *TCF7L2* semblent témoigner de pressions de sélection en faveur des allèles protecteurs pour le diabète de type II. Ces résultats, bien que non prédits directement dans les hypothèses du génotype économe et de la piste carnivore, pourraient tout de même être compatibles avec ces dernières si ces pressions de sélection sont récentes. Allen & Cher (1996) avaient d'ailleurs remarqué qu'il fallait mieux se demander pourquoi le diabète n'est pas en forte fréquence dans toutes les populations humaines, plutôt que de se demander pourquoi certaines populations présentent des fréquences si élevées. Cependant, ces auteurs favorisaient la dérive comme seule force responsable de ces différences de phénotype.

Etant donné que les gènes liés à l'alimentation ont des fonctions métaboliques, il parait raisonnable de penser que ces gènes sont de manière générale également sous l'influence d'autres contraintes environnementales telles que le climat, la température ou l'altitude. Il est donc difficile d'exclure un rôle important de ces variables pour expliquer les pressions de sélection observées. Nous avons en effet vu avec le gène de la lactase que les facteurs « culturels » liés au mode de subsistance et les facteurs « environnementaux » liés à

l'ensoleillement et l'aridité peuvent avoir des effets confondants. Cependant, l'étude majeure de Holden & Mace (1997), ainsi que l'étude plus récente de Itan *et al* (2009), ont montré que seuls les facteurs culturels expliquaient convenablement la distribution de la diversité génétique de la lactase. Ces études nécessitent en tout cas d'avoir des bases de données importantes pour réussir à distinguer parmi les hypothèses alternatives.

Concernant les gènes liés au diabète de type II, l'étude de Hancock *et al* (2008) a également montré que la distribution de la diversité de gènes liés aux fonctions métaboliques peut être significativement corrélée à différentes variables environnementales comme la latitude, les températures hivernales, l'humidité ou l'ensoleillement. Ainsi, en analysant des mutations associées au diabète de type II (les mêmes que celles que nous avons étudié sauf pour *PON1*), cette étude trouve que *FABP2* et *PPARG* sont significativement associés à la latitude, *LEPR* aux températures hivernales et *TCF7L2* et *PON1* à une variable résumant l'humidité, la pluie et les rayons UV. Cette étude ne tente cependant pas de distinguer ces effets de ceux liés au mode de subsistance. En effet, les modes de subsistance sont fortement contraints par l'environnement, avec une majorité de populations pastorales dans des climats extrêmes où l'agriculture est peu praticable : hautes latitudes, milieux froids ou en altitude, milieux secs et arides, et il n'est clairement pas évident de distinguer l'influence causale de tous ces facteurs. Etant donné que les gènes métaboliques sont au final liés à différentes fonctions, nous pouvons également imaginer que la sélection d'allèles protecteurs du diabète de type II puisse être en fait liée à un avantage de es allèles sur d'autres phénotypes comme la régulation corporelle de la température ou encore la gestion de l'énergie. Ainsi, des conflits évolutifs pourraient exister sur ces gènes fortement pléiotropiques.

En Asie Centrale, les Kirghiz sont des populations vivant en montagne, donc dans des conditions climatiques particulières, par rapport aux Tadjiks. Pour vérifier le rôle causal du mode de subsistance sur les pressions de sélection identifiées, nous pourrions obtenir des données sur d'autres populations comme par exemple les Ouigours (majoritairement en Chine), qui sont des agriculteurs de langue Turco-mongole. Cette population pourrait nous éclairer sur le rôle respectif de l'environnement et du mode de vie sur les pressions de sélection identifiées.

Dans les populations du CEPH (majoritairement agricultrices), les résultats pour *LEPR* et *TFC7L2* montrent que les pressions de sélection sont exclusivement trouvées en Asie de

l'Est, différenciant alors fortement des populations ayant un même mode de subsistance mais vivant dans des environnements différents (comme des agriculteurs d'Europe et d'Asie). Ces résultats peuvent suggérer que ces pressions de sélection ne sont pas liées aux modes de subsistance. Nous pouvons également émettre une hypothèse alternative : les pressions de sélection ont agi sur les mêmes phénotypes mais sur des mutations différentes (évolution convergente comme dans le cas de la lactase). Cette hypothèse est particulièrement probable du fait que ces pressions de sélection sont récentes et s'exercent alors sur une variabilité génétique neutre déjà différenciée (Przeworski *et al.*, 2005). De plus, les gènes liés au métabolisme sont organisés en réseaux complexes, et nous pouvons donc supposer qu'un même changement de phénotype soit obtenu par des mutations différentes sur des gènes aux fonctions redondantes.

Perspectives

Finalement, si les adaptations génétiques sur les gènes du diabète de type II sont bien liées à l'alimentation, nous pourrons alors atteindre un de nos objectifs initiaux de dater le début de l'expansion des mutations d'intérêt dans chaque ethnie, pour mieux comprendre quand ont eu lieu les transitions alimentaires en Asie Centrale. Nous avons déjà obtenu les dates de début d'expansion pour la mutation sur la lactase, mais il n'est pas évident d'interpréter ces résultats puisque i) les pressions de sélection ne sont pas différentes entre éleveurs et agriculteurs et ii) ces dates (comprises entre 10 000 et 4 000 BCE) sont plus anciennes que les premières traces connues d'utilisation du lait dans la région (3 500 BCE).

Afin de comprendre quand les populations d'Asie Centrale ont subi des changements majeurs de modes de vie, nous pourrions également, grâce aux données de puce à ADN, effectuer des tests de sélection sur des gènes liés à d'autres fonctions affectées par la révolution Néolithique, notamment comme des gènes liés à la réponse immunitaire (Barreiro *et al.*, 2009). Certains de ces gènes ont en effet été préalablement identifiés comme sous adaptation locale en fonction du mode de vie (Quintana-Murci, communication personnelle). D'autres gènes, liés à la détoxification de substances xénobiotiques comme *NAT2*, ont également montré des signes de sélection associés au mode de subsistance dans un échantillon mondial et précisément en Asie Centrale (Patin *et al.*, 2006, Magalon *et al.*, 2008). Ces informations pourront donc être croisées pour nous donner une bonne vision d'ensemble de l'histoire passée de ces populations.

Bien que nous ayons pour l'instant uniquement adopté une approche de gènes candidats, nous allons également pouvoir utiliser une stratégie de « scan génomique », grâce aux données obtenues par la puce à ADN. Ainsi, nous pourrons tester l'intensité de la sélection le long du génome et ensuite voir quelles classes de gènes sont le plus susceptibles d'être sous sélection. Il sera également intéressant d'étudier le déséquilibre de liaison dû à l'épistasie, pour voir si la résistance à l'insuline n'est pas liée à des combinaisons d'allèles à différents locus (Latta, 1998).

Finalement, le génotypage d'un grand nombre de marqueurs génétiques grâce aux puces à ADN pourra nous permettre d'obtenir le niveau de différenciation génétique observé sur des régions *a priori* neutres du génome. Cette enveloppe de différenciation observée sera utile pour tester la significativité de l'écart des mutations géniques à la neutralité, en s'affranchissant de tout modèle démographique *a priori* (comme ici le modèle en île). Elle pourra également nous servir à comparer l'enveloppe de différenciation observée de celle obtenue par simulation sous un modèle de migration en île, comme nous avons utilisé, ou sous d'autres modèles à tester, afin de tester l'influence des hypothèses du modèle sur les résultats de détection de mutations sous sélection.

CONCLUSION GÉNÉRALE

De manière générale, notre étude a eu pour objectif de comparer la diversité génétique de populations ayant adopté des modes de vies différents en Asie Centrale. Un premier groupe de populations est constitué de peuples parlant des langues indo-iraniennes, définis comme patrilocaux (les femmes partent s'installer dans le village du mari après le mariage), cognatiques (les individus se définissent d'après leurs ascendances paternelles et maternelles), et plutôt endogames (les conjoints sont choisis majoritairement à l'intérieur du groupe social). Ces peuples sont de plus des agriculteurs sédentaires installés *a priori* depuis longtemps en Asie Centrale. D'un autre côté, d'autres populations parlent des langues turques, sont patrilocales, constituées en groupes de parenté patrilinéaires (les individus se définissent uniquement d'après leur ascendance paternelle) et plutôt exogames (les conjoints sont choisis majoritairement en dehors du clan, mais à l'intérieur de la tribu). Ces populations sont de plus traditionnellement des éleveurs nomades. Ces deux groupes de populations sont donc culturellement assez différenciés, et nous posons la question de l'influence de la diversité culturelle sur la diversité génétique.

Le rôle de la géographie

Nous trouvons que la diversité génétique neutre est clairement différente entre ces populations, les Indo-iraniens étant plus proches des populations européennes et plus différenciés entre eux. Les Turco-mongols sont au contraire plus proches des autres populations d'Asie de l'Est et sont très homogènes entre eux. Ainsi, bien que les niveaux de différenciation soient faibles en Asie Centrale (1.4%), nous pouvons conclure que ces populations, qui se répartissent sur un territoire commun, sont clairement distinctes. Nous avons de plus trouvé que la géographie n'explique pas significativement la distribution de la diversité génétique en Asie Centrale. Ce résultat peut paraitre étonnant car au niveau mondial, nous retrouvons classiquement une bonne corrélation entre génétique et géographie (Manica *et al.*, 2005, Bosch *et al.*, 2006). Nous avons ainsi mis en évidence un rare exemple d'une absence de corrélation entre la distance génétique et géographique, quels que soit les marqueurs considérés ou le groupement de populations pris en compte. Il se peut donc que la corrélation au niveau mondial soit en fait la somme d'histoires différentes, qui au final donnent une tendance mondiale, mais que cette tendance ne soit pas forcément bien représentative de la réalité dans chaque région. Contrairement à la géographie, les facteurs culturels sont fortement responsables de cette différenciation génétique : l'affiliation

linguistique et ethnique expliquent significativement la répartition de la diversité génétique. Ainsi, ce qui gouverne les choix de conjoints et les échanges entre populations n'est pas tant leur répartition géographique mais plutôt leurs affiliations culturelles, au moins pour les populations d'Asie Centrale.

Si l'on regarde séparément l'histoire des hommes et des femmes, ces populations révèlent aussi des comportements distincts. Notre approche « multi-locus » basée sur des marqueurs du chromosome X et des autosomes nous a permis de conclure que, chez les Turco-mongols, le taux de migration et la taille efficace sont plus importants chez les femmes par rapport aux hommes, ce qui engendre de fortes différences de structure génétique entre eux. Chez les Indo-iraniens, de tels écarts ne sont pas mis en évidence. Ainsi, les caractéristiques culturelles de chaque population ont clairement une influence sur leur diversité génétique. Ces corrélations ne semblent cependant pas pouvoir être mises en évidence si l'on étudie les populations humaines à une échelle globale. En effet, tandis que les études locales ont souvent trouvé des différences de structure génétique entre hommes et femmes, différences qui varient en fonction de l'organisation sociale des populations (Salem *et al.*, 1996, Perez-Lezaun *et al.*, 1999, Oota *et al.*, 2001, Kayser *et al.*, 2003, Malyarchuk *et al.*, 2004, Destro-Bisol *et al.*, 2004, Nasidze *et al.*, 2004, Nasidze *et al.*, 2005, Hamilton *et al.*, 2005, Kumar *et al.*, 2006, Wilkins & Marlowe, 2006, Chaix *et al.*, 2007), les études à échelle globale ont donné des résultats contradictoires (Seielstad *et al.*, 1998, Dupanloup *et al.*, 2003, Wilder *et al.*, 2004a, Wilder *et al.*, 2004b, Ramachandran *et al.*, 2004).

L'exemple de l'adaptation à l'alimentation

Au niveau de l'étude des adaptations locales, nous avons trouvé que les deux premiers gènes candidats (la lactase et l'*AGXT*) ne présentent pas de différences entre les Turco-mongols et les Indo-iraniens. Cependant, pour les gènes liés au diabète de type II, nous avons trouvé des signes d'adaptation locale sur deux gènes, et nous avons mis en évidence une prévalence de résistance à l'insuline significativement plus importante chez les Kirghiz par rapport aux Tadjiks. Bien que d'autres analyses soient nécessaires pour comprendre les processus évolutifs en jeu, ces résultats suggèrent l'existence d'adaptations locales entre populations aux modes de subsistance contrastés. En parallèle des études locales, la comparaison de nombreuses populations à échelle globale peut permettre de gagner en

puissance et de tester l'influence de différents facteurs corrélés, comme l'ont montré Holden & Mace (1997) pour la lactase.

Cependant, afin de mettre en évidence quelle a été la contribution de chaque aire géographique à la corrélation trouvée entre pastoralisme et persistance de la lactase dans cette étude, j'ai réanalysé les données de Holden & Mace {, 1997 #394} selon le regroupement suivant : Asie / Océanie, Afrique, Europe, et Amérique, et en ajoutant également les données de fréquence de persistance de la lactase pour les populations d'Asie Centrale. Selon l'atlas ethnographique de Murdock (Murdock, 1967, Murdock & White, 2006), les populations Kazakhs, comme les populations Mongoles dépendent à 80.5% de ressources pastorales. Nous pouvons donc extrapoler qu'il en est de même pour les autres populations d'Asie Centrale (Kirghiz, Karakalpaks, Turkmènes et Ouzbeks) qui ne sont pas inclues dans cette base de données. Les résultats sont présentés ci-dessous:

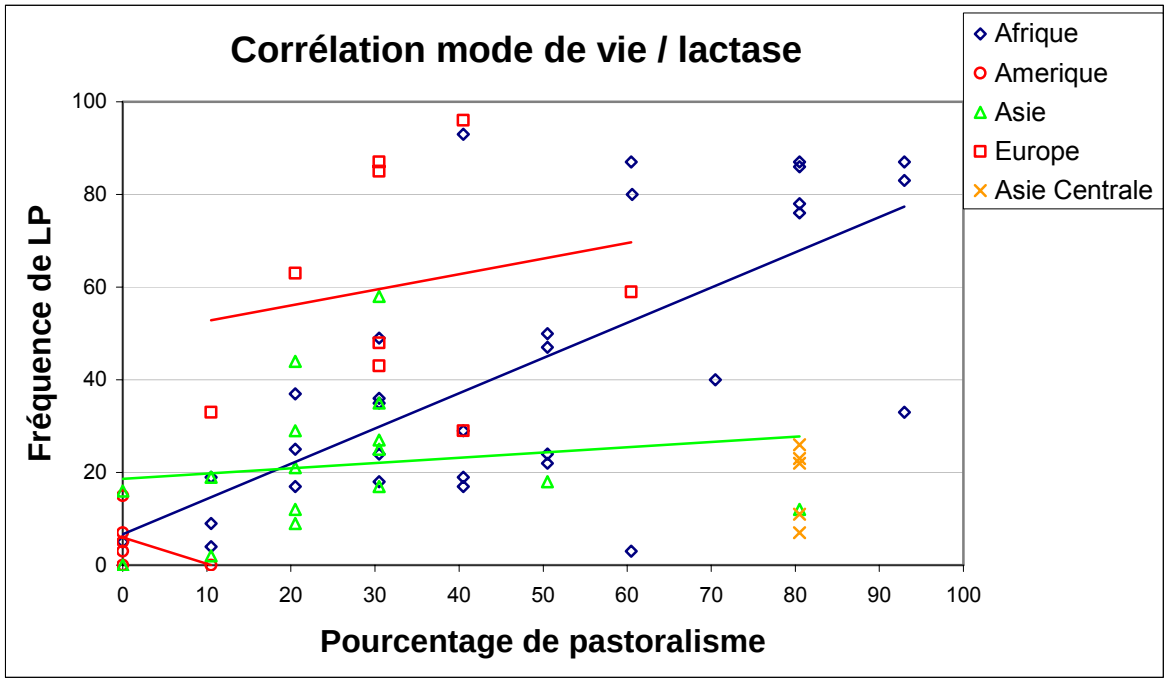


Figure 27 : Corrélation entre pastoralisme et fréquence de tolérance au lactose, d'après les données d'Holden & Mace, {, 1997 #394}.

Nous voyons donc bien, d'après ce graphique, que les populations africaines contribuent fortement à la corrélation observée entre pastoralisme et persistance de la lactase, par rapport aux populations européennes ou asiatiques. Il serait donc intéressant de réappliquer leur

approche sur chaque continent séparément, pour voir si leur conclusion est robuste à ces regroupements géographiques.

Nous avons donc souligné l'importance d'étudier les populations humaines à une échelle locale, surtout quand l'on s'interroge sur l'influence des facteurs culturels sur la diversité génétique. Le problème avec les études sur les populations humaines est que l'on cherche la plupart du temps à tirer des tendances générales applicables à toutes les populations humaines, mais au final, ces tendances masquent souvent des réalités plus complexes. A l'inverse, en étudiant seulement des populations particulières, nous pouvons faire ressortir des liens intéressants entre diversité génétique et culturelle. Il semble ainsi de manière générale que la réponse à la question de l'influence du mode de vie sur la diversité génétique dépende de l'échelle géographique à laquelle nous nous plaçons.

BIBLIOGRAPHIE

Akey, J.M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res,* 19**,** 711-22.

Alekseev, A.Y., Bokovenko, N.A., Boltrik, Y., Chungunov, K.A., Cook, G., Dergachev, V.A., Kovalyukh, N., Possnert, G., Van Der Plicht, J., Scott, E.M., Sementsov, A., Skripkin, V., Vasiliev, S. & Zaitseva, G. (2001) A chronology of the Scythian antiquities of Eurasia based on new archaeological and ^{14}C data. *Radiocarbon,* 43**,** 1085-1107.

Allen, J.S. & Cheer, S.M. (1996) 'Civilisation' and the thrifty genotype. *Asia Pacific J Clin Nutr,* 4**,** 341-342.

Ambrose, S. (1998) Chronology of the later Stone Age and food production in East Africa. *J Arch Sci,* 25**,** 377-391.

Ammerman, A. & Cavalli-Sforza, L.L. (1973) A population model for the diffusion of early farming in Europe. In: *The explanation of culture change: models in prehistory* C. Renfrew (ed.) London: Duckworth.

Ammerman, A. & Cavalli-Sforza, L.L. (1984) *The Neolithic transition and the genetics of populations in Europe.* Princeton.

Anderson, J.W., Kendall, C.W. & Jenkins, D.J. (2003) Importance of weight management in type 2 diabetes: review with meta-analysis of clinical studies. *J Am Coll Nutr,* 22**,** 331-9.

Anthony, D.W. (1998) The opening of the Eurasian steppe at 2000 BCE. In: *The Bronze Age and early Iron Age peoples of Central Asia* V.H. Mair (ed.) Washington: Institute for the study of Man.

Anthony, D.W. & Brown, D.R. (1991) The origins of horseback riding. *Antiquity,* 65**,** 22-38.

Anthony, D.W. & Vinogradov, N.B. (1995) Birth if the chariot. *Archaeology,* 28**,** 36-41.

Austerlitz, F., Kalaydjieva, L. & Heyer, E. (2003) Detecting population growth, selection and inherited fertility from haplotypic data in humans. *Genetics,* 165**,** 1579-86.

Baier, L.J., Sacchettini, J.C., Knowler, W.C., Eads, J., Paolisso, G., Tataranni, P.A., Mochizuki, H., Bennett, P.H., Bogardus, C. & Prochazka, M. (1995) An amino acid substitution in the human intestinal fatty acid binding protein is associated with increased fatty acid binding, increased fat oxidation, and insulin resistance. *J Clin Invest,* 95**,** 1281-7.

Balaresque, P. & Jobling, M.A. (2007) Human populations: houses for spouses. *Curr Biol,* 17**,** R14-6.

Ballard, J.W. & Whitlock, M.C. (2004) The incomplete natural history of mitochondria. *Mol Ecol,* 13**,** 729-44.

Balloux, F. (2009) The worm in the fruit of the mitochondrial DNA tree. *Heredity*.

Bamshad, M., Wooding, S., Salisbury, B.A. & Stephens, J.C. (2004) Deconstructing the relationship between genetics and race. *Nat Rev Genet,* 5**,** 598-609.

Barreiro, L.B., Ben-Ali, M., Quach, H., Laval, G., Patin, E., Pickrell, J.K., Bouchier, C., Tichit, M., Neyrolles, O., Gicquel, B., Kidd, J.R., Kidd, K.K., Alcais, A., Ragimbeau, J., Pellegrini, S., Abel, L., Casanova, J.L. & Quintana-Murci, L. (2009) Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet,* 5**,** e1000562.

Barroso, I. (2005) Genetics of Type 2 diabetes. *Diabet Med,* 22**,** 517-35.

Bayoumi, R.A., Flatz, S.D., Kuhnau, W. & Flatz, G. (1982) Beja and Nilotes: nomadic pastoralist groups in the Sudan with opposite distributions of the adult lactase phenotypes. *Am J Phys Anthropol,* 58**,** 173-8.

Bazin, E., Glemin, S. & Galtier, N. (2006) Population size does not influence mitochondrial genetic diversity in animals. *Science,* 312**,** 570-2.

Beaumont, M. & Nichols, R.A. (1996) Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc Lond,* 263**,** 1619-1626.

Beaumont, M.A. (2004) Recent developments in genetic data analysis: what can they tell us about human demographic history? *Heredity,* 92**,** 365-79.

Beaumont, M.A. (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends Ecol Evol,* 20**,** 435-440.

Beaumont, M.A. & Balding, D.J. (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol,* 13**,** 969-980.

Beaumont, M.A., Zhang, W. & Balding, D.J. (2002) Approximate Bayesian computation in population genetics. *Genetics,* 162**,** 2025-35.

Belle, E.M. & Barbujani, G. (2007) Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. *Am J Phys Anthropol,* 133**,** 1137-46.

Benyshek, D.C. & Watson, J.T. (2006) Exploring the thrifty genotype's food-shortage assumptions: a cross-cultural comparison of ethnographic accounts of food security among foraging and agricultural societies. *Am J Phys Anthropol,* 131**,** 120-6.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E. & Hirschhorn, J.N. (2004) Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet,* 74**,** 1111-20.

Blum, M.G. & François, O. (2009) Non-linear regression models for approximate bayesian computation. *Stat Comput,* in press.

Bosch, E., Calafell, F., Gonzalez-Neira, A., Flaiz, C., Mateu, E., Scheil, H.G., Huckenbeck, W., Efremovska, L., Mikerezi, I., Xirotiris, N., Grasa, C., Schmidt, H. & Comas, D. (2006) Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Ann Hum Genet,* 70**,** 459-87.

Brand Miller, J.C. & Colagiuri, S. (1994) The carnivore connection: dietary carbohydrate in the evolution of NIDDM. *Diabetologia,* 37**,** 1280-6.

Brunet, F. (1999) La Néolithisation en Asie Centrale: un état de la question. *Paléorient,* 24**,** 27-48.

Burton, M.L., Moore, C.C., Whiting, J.W.M. & Romney, A.K. (1996) Regions based on social structure. *Curr Anthro,* 37**,** 87-123.

Caldwell, E.F., Mayor, L.R., Thomas, M.G. & Danpure, C.J. (2004) Diet and the frequency of the alanine:glyoxylate aminotransferase Pro11Leu polymorphism in different human populations. *Hum Genet,* 115**,** 504-509.

Cauchi, S., Meyre, D., Durand, E., Proenca, C., Marre, M., Hadjadj, S., Choquet, H., De Graeve, F., Gaget, S., Allegaert, F., Delplanque, J., Permutt, M.A., Wasson, J., Blech, I., Charpentier, G., Balkau, B., Vergnaud, A.C., Czernichow, S., Patsch, W., Chikri, M., Glaser, B., Sladek, R. & Froguel, P. (2008) Post genome-wide association studies of novel genes associated with type 2 diabetes show gene-gene interaction and high predictive value. *PLoS One,* 3**,** e2031.

Cavalli-Sforza, L.L. (1966) Population structure and human evolution. *Proc R Soc Lond B Biol Sci,* 164**,** 362-79.

Cavalli-Sforza, L.L. (1996) The spread of agriculture and nomadic pastoralism: insights from genetics, linguistics and archaeology. In: *The origins and spread of agriculture and pastoralism in Eurasia* D.R. Harris (ed.) University College London: UCL Press.

Cavalli-Sforza, L.L. & Feldman, M.W. (2003) The application of molecular genetic approaches to the study of human evolution. *Nat Genet,* 33 Suppl**,** 266-75.

Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. (1994) *The history and geography of human genes*. Princeton.

Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. & Mountain, J. (1988) Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proc Natl Acad Sci U S A,* 85**,** 6002-6006.

Chaix, R., Austerlitz, F., Hegay, T., Quintana-Murci, L. & Heyer, E. (2008) Genetic traces of east-to-west human expansion waves in Eurasia. *Am J Phys Anthropol,* 136**,** 309-17.

Chaix, R., Austerlitz, F., Khegay, T., Jacquesson, S., Hammer, M.F., Heyer, E. & Quintana-Murci, L. (2004) The genetic or mythical ancestry of descent groups: lessons from the Y chromosome. *Am J Hum Genet,* 75**,** 1113-6.

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. *Curr Biol,* 17**,** 43-8.

Chang, C., Benecke, N., Grigoriev, F.P., Rosen, A.M. & Tourtellotte, P.A. (2003) Iron Age society and chronology in South-East Kazakhstan. *Antiquity,* 73**,** 298-312.

Charlesworth, D., Charlesworth, B. & Morgan, M.T. (1995) The pattern of neutral molecular variation under the background selection model. *Genetics,* 141**,** 1619-32.

Cheng, F., Chen, W., Richards, E., Deng, L. & Zeng, C. (2009) SNP@Evolution: a hierarchical database of positive selection on the human genome. *BMC Evol Biol,* 9**,** 221.

Cho, Y.M., Kim, T.H., Lim, S., Choi, S.H., Shin, H.D., Lee, H.K., Park, K.S. & Jang, H.C. (2009) Type 2 diabetes-associated genetic variants discovered in the recent genome-wide association studies are related to gestational diabetes mellitus in the Korean population. *Diabetologia,* 52**,** 253-61.

Coelho, M., Luiselli, D., Bertorelle, G., Lopes, A.I., Seixas, S., Destro-Bisol, G. & Rocha, J. (2005) Microsatellite variation and evolution of human lactase persistence. *Hum Genet,* 117**,** 329-39.

Colagiuri, S. & Brand Miller, J. (2002) The 'carnivore connection'--evolutionary aspects of insulin resistance. *Eur J Clin Nutr,* 56 Suppl 1**,** S30-5.

Comas, D., Calafell, F., Mateu, E., Perez-Lezaun, A., Bosch, E., Martinez-Arias, R., Clarimon, J., Facchini, F., Fiori, G., Luiselli, D., Pettener, D. & Bertranpetit, J. (1998) Trading genes along the Silk Road: mtDNA sequences and the origin of Central Asian populations. *Am J Hum Genet,* 63**,** 1824-38.

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. *Eur J Hum Genet,* 12**,** 495-504.

Conrad, D.F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T.D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C.H., Kristiansson, K., Macarthur, D.G., Macdonald, J.R., Onyiah, I., Pang, A.W., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N.P., Lee, C., Scherer, S.W. & Hurles, M.E. (2009) Origins and functional impact of copy number variation in the human genome. *Nature*.

Cook, G.C. (1978) Did persistence of intestinal lactase into adult life originate on the Arabian peninsula? *Man,* 13**,** 418-427.

Cook, G.C. & Al-Torki, M.T. (1975) High intestinal lactase concentrations in adult Arabs in Saudi Arabia. *Br Med J,* 3**,** 135-6.

Coop, G., Pickrell, J.K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R.M., Cavalli-Sforza, L.L., Feldman, M.W. & Pritchard, J.K. (2009) The role of geography in human adaptation. *PLoS Genet.,* 5**,** e1000500.

Cooper, P.J., Danpure, C.J., Wise, P.J. & Guttridge, K.M. (1988) Immunocytochemical localization of human hepatic alanine: glyoxylate aminotransferase in control subjects and patients with primary hyperoxaluria type 1. *J. Histochem. Cytochem.,* 36**,** 1285-1294.

Cordain, L., Eaton, S.B., Sebastian, A., Mann, N., Lindeberg, S., Watkins, B.A., O'keefe, J.H. & Brand-Miller, J. (2005) Origins and evolution of the Western diet: health implications for the 21st century. *Am J Clin Nutr,* 81**,** 341-54.

Cordaux, R., Deepa, E., Vishwanathan, H. & Stoneking, M. (2004) Genetic evidence for the demic diffusion of agriculture to India. *Science,* 304**,** 1125.

Cummins, J. (2001) Mitochondrial DNA and the Y chromosome: parallels and paradoxes. *Reprod Fertil Dev,* 13**,** 533-42.

Danpure, C.J. (1997) Variable peroxisomal and mitochondrial targeting of alanine: glyoxylate aminotransferase in mammalian evolution and disease. *Bioessays,* 19**,** 317-326.

Danpure, C.J. (2004) Molecular aetiology of primary hyperoxaluria type 1. *Nephron Exp Nephrol,* 98**,** e39-e44.

Danpure, C.J., Birdsey, G.M., Rumsby, G., Lumb, M.J., Purdue, P.E. & Allsop, J. (1994b) Molecular characterization and clinical use of a polymorphic tandem repeat in an intron of the human alanine:glyoxylate aminotransferase gene. *Hum Genet,* 94**,** 55-64.

Danpure, C.J., Cooper, P.J., Wise, P.J. & Jennings, P.R. (1989) An enzyme trafficking defect in two patients with primary hyperoxaluria type 1: peroxisomal alanine/glyoxylate aminotransferase rerouted to mitochondria. *J Cell Biol,* 108**,** 1345-1352.

Danpure, C.J., Guttridge, K.M., Fryer, P., Jennings, P.R., Allsop, J. & Purdue, P.E. (1990) Subcellular distribution of hepatic alanine:glyoxylate aminotransferase in various mammalian species. *J Cell Sci,* 97**,** 669-678.

Danpure, C.J. & Jennings, P.R. (1986) Peroxisomal alanine:glyoxylate aminotransferase deficiency in primary hyperoxaluria type 1. *FEBS Letters,* 201**,** 20-24.

Danpure, C.J., Jennings, P.R., Fryer, P., Purdue, P.E. & Allsop, J. (1994a) Primary hyperoxaluria type 1: genotypic and phenotypic heterogeneity. *J Inherit Metab Dis,* 17**,** 487-499.

Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglia, A., Tofanelli, S., Spedini, G. & Capelli, C. (2004) Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Mol Biol Evol,* 21**,** 1673-82.

Diamond, J. (2002) Evolution, consequences and future of plant and animal domestication. *Nature,* 418**,** 700-7.

Dulloo, A.G., Stock, M.J., Solinas, G., Boss, O., Montani, J.P. & Seydoux, J. (2002) Leptin directly stimulates thermogenesis in skeletal muscle. *FEBS Lett,* 515**,** 109-13.

Dupanloup, I., Pereira, L., Bertorelle, G., Calafell, F., Prata, M.J., Amorim, A. & Barbujani, G. (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *J Mol Evol,* 57**,** 85-97.

Durham, W. (1991) *Coevolution: genes, culture and human diversity.* Standford.

Eaton, S.B. & Konner, M. (1985) Palaeolithic nutrition. A consideration of its nature and current implications. *N Engl J Med,* 312**,** 283-9.

Enattah, N.S., Sahi, T., Savilahti, E., Terwilliger, J.D., Peltonen, L. & Jarvela, I. (2002) Identification of a variant associated with adult-type hypolactasia. *Nat Genet,* 30**,** 233-7.

Evershed, R.P., Payne, S., Sherratt, A.G., Copley, M.S., Coolidge, J., Urem-Kotsu, D., Kotsakis, K., Ozdogan, M., Ozdogan, A.E., Nieuwenhuyse, O., Akkermans, P.M., Bailey, D., Andeescu, R.R., Campbell, S., Farid, S., Hodder, I., Yalman, N., Ozbasaran, M., Bicakci, E., Garfinkel, Y., Levy, T. & Burton, M.M. (2008) Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature,* 455**,** 528-31.

Ewens, W.J. (1972) The sampling theory of selectively neutral alleles. *Theor Popul Biol,* 3**,** 87-112.

Excoffier, L., Laval, G. & Schneider, S. (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol Bioinfo Online,* 1**,** 47-50.

Excoffier, L., Smouse, P.E. & Quattro, J.M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics,* 131**,** 479-491.

Fagan, B.M. (1992) *People of the earth.* New York.

Fay, J.C. & Wu, C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics,* 155**,** 1405-13.

Flatz, G. & Rotthauwe, H.W. (1973) Lactose nutrition and natural selection. *Lancet,* 2**,** 76-7.

Florez, J.C., Hirschhorn, J. & Altshuler, D. (2003) The inherited basis of diabetes mellitus: implications for the genetic analysis of complex traits. *Annu Rev Genomics Hum Genet,* 4**,** 257-91.

Forster, P. & Matsumura, S. (2005) Evolution. Did early humans go north or south? *Science,* 308**,** 965-6.

Frachetti, M. (2008) The variability and dynamic landscapes of mobile pastoralism in ethnography and prehistory. In: *The archaeology of mobility* H.B.A.W. Wendrich (ed.) Los Angeles: Costen Institute of Archaeology.

Frayling, T.M. (2007) Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nat Rev Genet,* 8**,** 657-62.

Frazer, K.A., Murray, S.S., Schork, N.J. & Topol, E.J. (2009) Human genetic variation and its contribution to complex traits. *Nat Rev Genet,* 10**,** 241-51.

Freeman, H. & Cox, R.D. (2006) Type-2 diabetes: a cocktail of genetic discovery. *Hum Mol Genet,* 15 Spec No 2**,** R202-9.

Frienkel, N. (1980) Of pregnancy and progeny. *Diabetes,* 29**,** 1023-1034.

Fu, Y.X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics,* 147**,** 915-25.

Fu, Y.X. & Li, W.H. (1993) Statistical tests of neutrality of mutations. *Genetics,* 133**,** 693-709.

Fullerton, S.M., Bartoszewicz, A., Ybazeta, G., Horikawa, Y., Bell, G.I., Kidd, K.K., Cox, N.J., Hudson, R.R. & Di Rienzo, A. (2002) Geographic and haplotype structure of candidate type 2 diabetes susceptibility variants at the calpain-10 locus. *Am J Hum Genet,* 70**,** 1096-106.

Garn, S.M. & Leonard, W.R. (1989) What did our ancestors eat? *Nutrition Reviews,* 47**,** 337-345.

Gaulin, S.J.C. & Konner, M. (1977) On the natural diet of primates, includind humans. In: *Nutrition and the brain* R.J. Wurtman & J.J. Wurtman (eds.) New York: Raven Press.

Gifford-Gonzales, D. (2005) *African Archeology.* London: Blackwell.

Gimbutas, M. (1991) *The civilization of the goddess: the world of old Europe.* San Fransisco.

Goldstein, D.B. & Chikhi, L. (2002) Human migrations and population structure: what we know and why it matters. *Annu Rev Genomics Hum Genet,* 3**,** 129-52.

Grun, R. & Stringer, C. (2000) Tabun revisited: revised ESR chronology and new ESR and U-series analyses of dental material from Tabun C1. *J Hum Evol,* 39**,** 601-612.

Gryaznov, M.P. (1980) *Arzhan: Tsarskii kurgan ranneskiskogo vremeni.* Leningrad: Nauka.

Haglin, L. (1991) Nutrient intake among Saami people today compared with an old, traditional Saami diet. *Arctic Med. Res.,* Suppl**,** 741-746.

Haglin, L. (1999) The nutrient density of present-day and traditional diets and their health aspects: the Sami- and lumberjack families living in rural areas of Northern Sweden. *Int. J. Circ. Health,* 58**,** 30-43.

Hamilton, G., Stoneking, M. & Excoffier, L. (2005) Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc Natl Acad Sci U S A,* 102**,** 7476-80.

Hammer, M.F., Karafet, T.M., Redd, A.J., Jarjanazi, H., Santachiara-Benerecetti, S., Soodyall, H. & Zegura, S.L. (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol,* 18**,** 1189-203.

Hancock, A.M., Witonsky, D.B., Gordon, A.S., Eshel, G., Pritchard, J.K., Coop, G. & Di Rienzo, A. (2008) Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genet,* 4**,** e32.

Harpending, H. (2006) Anthropological genetics: present and future. In: *Anthropological genetics. Theory, methods and applications* M.H. Crawford (ed.) Cambridge: Cambridge University Press.

Harpending, H. & Rogers, A. (2000) Genetic perspectives on human origins and differentiation. *Annu Rev Genomics Hum Genet,* 1**,** 361-85.

Harris, D.R. (1996) The origins and spread of agriculture and pastoralism in Eurasia: an overview. In: *The origins and spread of agriculture and pastoralism in Eurasia* D.R. Harris (ed.) University College London: UCL Press.

Helgason, A., Palsson, S., Thorleifsson, G., Grant, S.F., Emilsson, V., Gunnarsdottir, S., Adeyemo, A., Chen, Y., Chen, G., Reynisdottir, I., Benediktsson, R., Hinney, A., Hansen, T., Andersen, G., Borch-Johnsen, K., Jorgensen, T., Schafer, H., Faruque, M.,

Doumatey, A., Zhou, J., Wilensky, R.L., Reilly, M.P., Rader, D.J., Bagger, Y., Christiansen, C., Sigurdsson, G., Hebebrand, J., Pedersen, O., Thorsteinsdottir, U., Gulcher, J.R., Kong, A., Rotimi, C. & Stefansson, K. (2007) Refining the impact of TCF7L2 gene variants on type 2 diabetes and adaptive evolution. *Nat Genet,* 39**,** 218-25.

Heyer, E. & Quintana-Murci, L. (2009) Evolutionary genetics as a tool to target genes involved in phenotypes of medical relevance. *Evolutionary Applications,* 2**,** 71-80.

Heyer, E., Balaresque, P., Jobling, M.A., Quintana-Murci, L., Chaix, R., Segurel, L., Aldashev, A. & Hegay, T. (2009) Genetic diversity and the emergence of ethnic groups in Central Asia. *BMC Genet,* 10**,** 49.

Heyer, E., Sibert, A. & Austerlitz, F. (2005) Cultural transmission of fitness: genes take the fast lane. *Trends Genet,* 21**,** 234-9.

Hofer, T., Ray, N., Wegmann, D. & Excoffier, L. (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann Hum Genet,* 73**,** 95-108.

Holbrook, J.D., Birdsey, G.M., Yang, Z., Bruford, M.W. & Danpure, C.J. (2000) Molecular adaptation of alanine:glyoxylate aminotransferase targeting in primates. *Mol Biol Evol,* 17**,** 387-400.

Holden, C. & Mace, R. (1997) Phylogenetic analysis of the evolution of lactose digestion in adults. *Hum Biol,* 69**,** 605-28.

Hruschka, D.J. & Brandon, A.K. (2004) Mongolia. In: *Encyclopedia of medical anthropology. Volume II: Cultures* C.R. Ember & M. Ember (eds.) New York: Springer.

Hu, D., Sun, L., Fu, P., Xie, J., Lu, J., Zhou, J., Yu, D., Whelton, P.K., He, J. & Gu, D. (2009) Prevalence and risk factors for type 2 diabetes mellitus in the Chinese adult population: the InterASIA Study. *Diabetes Res Clin Pract,* 84**,** 288-95.

Ingram, C.J., Mulcare, C.A., Itan, Y., Thomas, M.G. & Swallow, D.M. (2009) Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet,* 124**,** 579-91.

Ingram, C.J.E. (2008) The evolutionary genetics of lactase persistence in Africa and the Middle East. In: *University of London*.

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature,* 437**,** 1299-320.

International HapMap Consortium (2003) The International HapMap Project. *Nature,* 426**,** 789-96.

International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., Pasternak, S., Wheeler, D.A., Willis, T.D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhou, J., Gabriel, S.B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R.C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Sun, W., Wang, H., Wang, Y., Xiong, X., Xu, L., Waye, M.M., Tsui, S.K., Xue, H., Wong, J.T., Galver, L.M., Fan, J.B., Gunderson, K., Murray, S.S., Oliphant, A.R., Chee, M.S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.F., Phillips, M.S., Roumy, S., Sallee, C., Verner, A., Hudson, T.J., Kwok, P.Y., Cai, D., Koboldt, D.C., Miller, R.D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.C., Mak, W., Song, Y.Q., Tam, P.K., Nakamura, Y., Kawaguchi, T., Kitamoto, T., Morizono, T., Nagashima, A., Ohnishi, Y., Sekine, A., Tanaka, T., et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature,* 449**,** 851-61.

Iron, W. (1974) Nomadism as a political adaptation: the case of the Yomut Turkmen. *American Ethnologist,* 1**,** 635-658.

Itan, Y., Powell, A., Beaumont, M.A., Burger, J. & Thomas, M.G. (2009) The origins of lactase persistence in Europe. *PLoS Comput Biol,* 5**,** e1000491.

Jacobson-Tepfer, E. (2008) The emergence of cultures of mobility in the Altai mountains of Mongolia. In: *The archaeology of mobility* H.B.A.W. Wendrich (ed.) Los Angeles: Costen Institute of Archaeology.

Jacobson, E. (2001) Cultural riddles: stylized deer and deer stones of the Mongolian Altai. *Bulletin of the Asia Institute. New series,* 15**,** 31-56.

Jacquesson, S. (2002) Parcours ethnographiques dans l'histoire des deltas. In: *Karakalpaks et autres gens de l'Aral: entre rivages et déserts* S. Jacquesson & V. Fourniau (eds.) Tachkent - Aix-en-Provence: Edisud.

Kaprio, J., Tuomilehto, J., Koskenvuo, M., Romanov, K., Reunanen, A., Eriksson, J., Stengard, J. & Kesaniemi, Y.A. (1992) Concordance for type 1 (insulin-dependent) and type 2 (non-insulin-dependent) diabetes mellitus in a population-based cohort of twins in Finland. *Diabetologia,* 35**,** 1060-7.

Kaput, J., Noble, J., Hatipoglu, B., Kohrs, K., Dawson, K. & Bartholomew, A. (2007) Application of nutrigenomic concepts to Type 2 diabetes mellitus. *Nutr Metab Cardiovasc Dis,* 17**,** 89-103.

Karafet, T., Xu, L., Du, R., Wang, W., Feng, S., Wells, R.S., Redd, A.J., Zegura, S.L. & Hammer, M.F. (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am J Hum Genet,* 69**,** 615-28.

Kayser, M., Brauer, S., Weiss, G., Schiefenhovel, W., Underhill, P., Shen, P., Oefner, P., Tommaseo-Ponzetta, M. & Stoneking, M. (2003) Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am J Hum Genet,* 72**,** 281-302.

Khazanov, A.M. (1994) *Nomads and the outside world.* Madison: University of Wisconsin Press.

Khlobystina, M.D. (1973) Origins and devlopment of the first Bronze Age civilization in South Siberia. *Sovetskaya Arkeologiya,* 1**,** 24-38.

King, H. & Rewers, M. (1993) Global estimates for prevalence of diabetes mellitus and impaired glucose tolerance in adults. WHO Ad Hoc Diabetes Reporting Group. *Diabetes Care,* 16**,** 157-77.

Kislenko, A. & Tatarintseva, N. (1999) The eastern Ural steppe at the end of the Stone Age. In: *Late prehistoric exploitation of the Eurasian steppe* M. Levine, Y. Rassamakin, A. Kislenko & T.N. Kislenko (eds.) *Late prehistoric exploitation of the Eurasian steppe.* Cambridge: McDonald Institute for Archeological Research

Knowler, W.C., Bennett, P.H., Hamman, R.F. & Miller, M. (1978) Diabetes incidence and prevalence in Pima Indians: a 19-fold greater incidence than in Rochester, Minnesota. *Am J Epidemiol,* 108**,** 497-505.

Knowler, W.C., Pettitt, D.J., Bennett, P.H. & Williams, R.C. (1983) Diabetes mellitus in the Pima Indians: genetic and evolutionary considerations. *Am J Phys Anthropol,* 62**,** 107-14.

Kozlov, A. & Lisitsyn, D. (1997) Hypolactasia in Saami subpopulations of Russia and Finland. *Anthropol Anz,* 55**,** 281-7.

Krader, L. (1963) *Peoples of Central Asia.* Bloomington.

Krause, J., Orlando, L., Serre, D., Viola, B., Prüfer, K., Richards, M., Hublin, J.J., Hänni, C., Derevianko, A.P. & Pääbo, S. (2007) Neanderthals in central Asia and Siberia. *Nature,* 449**,** 755-927.

Kumar, V., Langstieh, B.T., Madhavi, K.V., Naidu, V.M., Singh, H.P., Biswas, S., Thangaraj, K., Singh, L. & Reddy, B.M. (2006) Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. *PLoS Genet,* 2**,** e53.

Lalueza-Fox, C., Sampietro, M.L., Gilbert, M.T., Castri, L., Facchini, F., Pettener, D. & Bertranpetit, J. (2004) Unravelling migrations in the steppe: mitochondrial DNA sequences from ancient central Asians. *Proc Biol Sci,* 271**,** 941-7.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitzhugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., Levine, R., Mcewan, P., Mckernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, N., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., Mcmurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., Mcpherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., et al. (2001) Initial sequencing and analysis of the human genome. *Nature,* 409**,** 860-921.

Lansing, J.S., Cox, M.P., Downey, S.S., Gabler, B.M., Hallmark, B., Karafet, T.M., Norquest, P., Schoenfelder, J.W., Sudoyo, H., Watkins, J.C. & Hammer, M.F. (2007) Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc Natl Acad Sci U S A,* 104**,** 16022-6.

Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., Comas, D., Holmlund, G., Kouvatsi, A., Macek, M., Mollet, I., Parson, W., Palo, J., Ploski, R., Sajantila, A., Tagliabracci, A., Gether, U., Werge, T., Rivadeneira, F., Hofman, A., Uitterlinden, A.G., Gieger, C., Wichmann, H.E., Ruther, A., Schreiber, S., Becker, C., Nurnberg, P., Nelson, M.R., Krawczak, M.

& Kayser, M. (2008) Correlation between genetic and geographic structure in Europe. *Curr Biol,* 18**,** 1241-8.

Larsen, C.S. (1995) Biological changes in human populations with agriculture. *Ann Rev Anthropo,* 24**,** 185-213.

Latta, R.G. (1998) Differentiation of allelic frequencies at quantitative trait loci affecting locally adaptive traits. *Am Nat,* 151**,** 283-92.

Lenormand, T. (2002) Gene flow and the limits of natural selection. *Trends in Ecology & Evolution,* 17**,** 183-189.

Lewontin, R.C. (1972) The apportionment of human diversity. *Evol Biol,* 6**,** 391-398.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. & Myers, R.M. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. *Science,* 319**,** 1100-4.

Librado, P. & Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics,* 25**,** 1451-2.

Liu, Y., Yu, L., Zhang, D., Chen, Z., Zhou, D.Z., Zhao, T., Li, S., Wang, T., Hu, X., Feng, G.Y., Zhang, Z.F., He, L. & Xu, H. (2008) Positive association between variations in CDKAL1 and type 2 diabetes in Han Chinese individuals. *Diabetologia,* 51**,** 2134-7.

Lum, J.K., Cann, R.L., Martinson, J.J. & Jorde, L.B. (1998) Mitochondrial and nuclear genetic relationships among Pacific Island and Asian populations. *Am J Hum Genet,* 63**,** 613-24.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., Taha, A., Shaari, N.K., Raja, J.M., Ismail, P., Zainuddin, Z., Goodwin, W., Bulbeck, D., Bandelt, H.J., Oppenheimer, S., Torroni, A. & Richards, M. (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science,* 308**,** 1034-6.

Magalon, H., Patin, E., Austerlitz, F., Hegay, T., Aldashev, A., Quintana-Murci, L. & Heyer, E. (2008) Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia. *Eur J Hum Genet,* 16**,** 243-51.

Malécot, G. (1973) Isolation by distance. In: *Genetic structure of populations* N.E. Morton (ed.) Honolulu: University of Hawaï Press.

Malyarchuk, B., Derenko, M., Grzybowski, T., Lunkina, A., Czarny, J., Rychkov, S., Morozova, I., Denisova, G. & Miscicka-Sliwka, D. (2004) Differentiation of

mitochondrial DNA and Y chromosomes in Russian populations. *Hum Biol,* 76**,** 877-900.

Maniatis, T., Fritsh, E.F. & J., S. (1982) Molecular clonning. A laboratory manual. *New York: Cold Spring Laboratory*.

Manica, A., Prugnolle, F. & Balloux, F. (2005) Geography is a better determinant of human genetic differentiation than ethnicity. *Hum Genet,* 118**,** 366-71.

Masson, V.M. (1992) The environment. In: *History of civilizations of Central Asia* A.H. Dani & V.M. Masson (eds.) Paris: UNESCO.

Mazoyer, M. & Roudart, F. (1997) *Histoire des agricultures du monde. Du néolithique à la crise contemporaine.*

McCracken, R.D. (1971) Lactase deficiency: example of dietary evolution. *Curr Anthropol,* 12**,** 479.

McDonald, J.H. & Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature,* 351**,** 652-4.

Mcdougall, I., Brown, F.H. & Fleagle, J.G. (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature,* 433**,** 733-6.

Miyake, K., Yang, W., Hara, K., Yasuda, K., Horikawa, Y., Osawa, H., Furuta, H., Ng, M.C., Hirota, Y., Mori, H., Ido, K., Yamagata, K., Hinokio, Y., Oka, Y., Iwasaki, N., Iwamoto, Y., Yamada, Y., Seino, Y., Maegawa, H., Kashiwagi, A., Wang, H.Y., Tanahashi, T., Nakamura, N., Takeda, J., Maeda, E., Yamamoto, K., Tokunaga, K., Ma, R.C., So, W.Y., Chan, J.C., Kamatani, N., Makino, H., Nanjo, K., Kadowaki, T. & Kasuga, M. (2009) Construction of a prediction model for type 2 diabetes mellitus in the Japanese population based on 11 genes with strong evidence of the association. *J Hum Genet,* 54**,** 236-41.

Mulcare, C.A., Weale, M.E., Jones, A.L., Connell, B., Zeitlyn, D., Tarekegn, A., Swallow, D.M., Bradman, N. & Thomas, M.G. (2004) The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am J Hum Genet,* 74**,** 1102-10.

Murdock, G.P. (1967) *Ethnographic Atlas.* Pittsburgh: University of Pittsburgh Press.

Murdock, G.P. & White, D.R. (2006) Standard cross-cultural sample. On-line edition. *Ethnology,* 9**,** 329-369.

Myles, S., Hradetzky, E., Engelken, J., Lao, O., Nurnberg, P., Trent, R.J., Wang, X., Kayser, M. & Stoneking, M. (2007) Identification of a candidate genetic variant for the high prevalence of type II diabetes in Polynesians. *Eur J Hum Genet,* 15**,** 584-9.

Nasidze, I., Ling, E.Y., Quinque, D., Dupanloup, I., Cordaux, R., Rychkov, S., Naumova, O., Zhukova, O., Sarraf-Zadegan, N., Naderi, G.A., Asgary, S., Sardas, S., Farhud, D.D., Sarkisian, T., Asadov, C., Kerimov, A. & Stoneking, M. (2004) Mitochondrial DNA and Y-chromosome variation in the caucasus. *Ann Hum Genet,* 68**,** 205-21.

Nasidze, I., Quinque, D., Ozturk, M., Bendukidze, N. & Stoneking, M. (2005) MtDNA and Y-chromosome variation in Kurdish groups. *Ann Hum Genet,* 69**,** 401-12.

Neel, J.V. (1962) Diabetes mellitus: a "thrifty" genotype rendered detrimental by "progress"? *Am J Hum Genet,* 14**,** 353-362.

Neel, J.V. (1976) Towards a better understanding of the genetic basis of diabetes mellitus. In: *The genetics of diabetes mellitus* C. W., J. Kobberling & J.V. Neel (eds.) Berlin: Springer Verlag.

Neel, J.V. (1982) The thrifty genotype revisited. In: *The genetics of diabetes mellitus* J. Kobberling & R.B. Tattersall (eds.) London: Academic Press.

Neel, J.V. & Ward, R.H. (1970) Village and tribal genetic distances among American Indians, and the possible implications for human evolution. *Proc Natl Acad Sci U S A,* 65**,** 323-30.

Nei, M. & Roychoudhury, A.K. (1993) Evolutionary relationships of human populations on a global scale. *Mol Biol Evol,* 10**,** 927-43.

Nesse, R.M. & Williams, G.C. (1996) *Why we get sick. The new science of Darwinian Medicine.*

Nielsen, R. (2005) Molecular signatures of natural selection. *Annu Rev Genet,* 39**,** 197-218.

Noguchi, T. (1987) Aromatic-amino-acid aminotransferase from small intestine. *Methods Enzymol,* 142**,** 267-273.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., Stephens, M. & Bustamante, C.D. (2008) Genes mirror geography within Europe. *Nature,* 456**,** 98-101.

O'dea, K. (1991) Westernisation, insulin resistance and diabetes in Australian aborigines. *Med J Aust,* 155**,** 258-64.

Okladnikov, A.P. (1940) Neanderthal man and his culture in Central Asia. *Asia,* 40**,** 357-361.

Omori, S., Tanaka, Y., Takahashi, A., Hirose, H., Kashiwagi, A., Kaku, K., Kawamori, R., Nakamura, Y. & Maeda, S. (2008) Association of CDKAL1, IGF2BP2, CDKN2A/B,

HHEX, SLC30A8, and KCNJ11 with susceptibility to type 2 diabetes in a Japanese population. *Diabetes,* 57**,** 791-5.

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. *Nat Genet,* 29**,** 20-1.

Orozco, L.D., Cokus, S.J., Ghazalpour, A., Ingram-Drake, L., Wang, S., Van Nas, A., Che, N., Araujo, J.A., Pellegrini, M. & Lusis, A.J. (2009) Copy number variation influences gene expression and metabolic traits in mice. *Hum Mol Genet,* 18**,** 4118-29.

Outram, A.K., Stear, N.A., Bendrey, R., Olsen, S., Kasparov, A., Zaibert, V., Thorpe, N. & Evershed, R.P. (2009) The earliest horse harnessing and milking. *Science,* 323**,** 1332-5.

Pakendorf, B. & Stoneking, M. (2005) Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet,* 6**,** 165-83.

Park, K.S., Shin, H.D., Park, B.L., Cheong, H.S., Cho, Y.M., Lee, H.K., Lee, J.Y., Lee, J.K., Oh, B. & Kimm, K. (2006) Polymorphisms in the leptin receptor (LEPR)--putative association with obesity and T2DM. *J Hum Genet,* 51**,** 85-91.

Patin, E., Barreiro, L.B., Sabeti, P.C., Austerlitz, F., Luca, F., Sajantila, A., Behar, D.M., Semino, O., Sakuntabhai, A., Guiso, N., Gicquel, B., Mcelreavey, K., Harding, R.M., Heyer, E. & Quintana-Murci, L. (2006) Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. *Am J Hum Genet,* 78**,** 423-36.

Patin, E., Laval, G., Barreiro, L.B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K.K., Kidd, J.R., Van Der Veen, L., Hombert, J.M., Gessain, A., Froment, A., Bahuchet, S., Heyer, E. & Quintana-Murci, L. (2009) Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS Genet,* 5**,** e1000448.

Perez-Lezaun, A., Calafell, F., Comas, D., Mateu, E., Bosch, E., Martinez-Arias, R., Clarimon, J., Fiori, G., Luiselli, D., Facchini, F., Pettener, D. & Bertranpetit, J. (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet,* 65**,** 208-19.

Perry, G.H., Dominy, N.J., Claw, K.G., Lee, A.S., Fiegler, H., Redon, R., Werner, J., Villanea, F.A., Mountain, J.L., Misra, R., Carter, N.P., Lee, C. & Stone, A.C. (2007) Diet and the evolution of human amylase gene copy number variation. *Nat Genet,* 39**,** 1256-60.

Pettitt, D.J., Baird, H.R., Aleck, K.A. & Knowler, W.C. (1981) Obesity in children following maternal diabetes mellitus during gestation. *Am J Epidemiol,* 114**,** 437.

Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W. & Pritchard, J.K. (2009) Signals of recent positive selection in a worldwide sample of human populations. *Genome Res,* 19**,** 826-37.

Pollinger, J.P., Bustamante, C.D., Fledel-Alon, A., Schmutz, S., Gray, M.M. & Wayne, R.K. (2005) Selective sweep mapping of genes with large phenotypic effects. *Genome Res,* 15**,** 1809-19.

Poloni, E.S., Semino, O., Passarino, G., Santachiara-Benerecetti, A.S., Dupanloup, I., Langaney, A. & Excoffier, L. (1997) Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *Am J Hum Genet,* 61**,** 1015-35.

Poulsen, P., Kyvik, K.O., Vaag, A. & Beck-Nielsen, H. (1999) Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance--a population-based twin study. *Diabetologia,* 42**,** 139-45.

Prentice, A.M., Hennig, B.J. & Fulford, A.J. (2008) Evolutionary origins of the obesity epidemic: natural selection of thrifty genes or genetic drift following predation release? *Int J Obes (Lond),* 32**,** 1607-10.

Prokopenko, I., Mccarthy, M.I. & Lindgren, C.M. (2008) Type 2 diabetes: new genes, new understanding. *Trends Genet,* 24**,** 613-21.

Prugnolle, F., Manica, A. & Balloux, F. (2005) Geography predicts neutral genetic diversity of human populations. *Curr Biol,* 15**,** R159-60.

Przeworski, M., Coop, G. & Wall, J.D. (2005) The signature of positive selection on standing genetic variation. *Evolution,* 59**,** 2312-23.

Purdue, P.E., Takada, Y. & Danpure, C.J. (1990) Identification of mutations associated with peroxisome-to-mitochondrion mistargeting of alanine/glyoxylate aminotransferase in primary hyperoxaluria type 1. *J Cell Biol,* 111**,** 2341-2351.

Qu, Y., Yang, Z., Jin, F., Sun, L., Feng, J., Tang, L., Zhang, C., Zhu, X., Shi, X., Sun, H., Wang, B. & Wang, L. (2008a) The haplotype identified in LEPR gene is associated with type 2 diabetes mellitus in Northern Chinese. *Diabetes Res Clin Pract,* 81**,** 33-7.

Qu, Y., Yang, Z., Jin, F., Sun, L., Zhang, C., Ji, L., Sun, H., Wang, B. & Wang, L. (2008b) The Ser311Cys variation in the paraoxonase 2 gene increases the risk of type 2 diabetes in northern Chinese. *J Genet,* 87**,** 165-9.

Quintana-Murci, L., Semino, O., Bandelt, H.J., Passarino, G., Mcelreavey, K. & Santachiara-Benerecetti, A.S. (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat Genet,* 23**,** 437-41.

R Development Core Team (2007) R: A Language and Environment for Statistical Computing.). Vienna: R Foundation for Statistical Computing.

Ramachandran, S., Rosenberg, N.A., Zhivotovsky, L.A. & Feldman, M.W. (2004) Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. *Hum Genomics,* 1**,** 87-97.

Ranov, V.A., Carbonell, E. & Rodriguez, X.P. (1995) Kuldara: Earliest human occupation in Central Asia in its Afro-Asian context. *Curr Anthropo,* 36**,** 337-346.

Reich, D.E. & Goldstein, D.B. (1998) Genetic evidence for a Paleolithic human population expansion in Africa. *Proc Natl Acad Sci U S A,* 95**,** 8119-23.

Renfrew, C. (1996) Language families and the spread of farming. In: *The origins and spread of africulture and pastoralism in Eurasia* D.R. Harris (ed.) University College London: UCL Press.

Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K. & Feldman, M.W. (2005) Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet,* 1**,** e70.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. & Feldman, M.W. (2002) Genetic structure of human populations. *Science,* 298**,** 2381-2385.

Ross, A.B., Johansson, A., Ingman, M. & Gyllensten, U. (2006) Lifestyle, genetics, and disease in Sami. *Croat Med J,* 47**,** 553-65.

Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., Beckman, G., Beckman, L., Bertranpetit, J., Bosch, E., Bradley, D.G., Brede, G., Cooper, G., Corte-Real, H.B., De Knijff, P., Decorte, R., Dubrova, Y.E., Evgrafov, O., Gilissen, A., Glisic, S., Golge, M., Hill, E.W., Jeziorowska, A., Kalaydjieva, L., Kayser, M., Kivisild, T., Kravchenko, S.A., Krumina, A., Kucinskas, V., Lavinha, J., Livshits, L.A., Malaspina, P., Maria, S., Mcelreavey, K., Meitinger, T.A., Mikelsaar, A.V., Mitchell, R.J., Nafa, K., Nicholson, J., Norby, S., Pandya, A., Parik, J., Patsalis, P.C., Pereira, L., Peterlin, B., Pielberg, G., Prata, M.J., Previdere, C., Roewer, L., Rootsi, S., Rubinsztein, D.C., Saillard, J., Santos, F.R., Stefanescu, G., Sykes, B.C., Tolun, A., Villems, R., Tyler-Smith, C. &

Jobling, M.A. (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet,* 67**,** 1526-43.

Rousset, F. (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics,* 145**,** 1219-28.

Rousset, F. (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Res,* 8**,** 103-106.

Ruiz-Narvaez, E. (2005) Is the Ala12 variant of the PPARG gene an "unthrifty allele"? *J Med Genet,* 42**,** 547-50.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., Mccarroll, S.A., Gaudet, R., Schaffner, S.F., Lander, E.S., Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M., Pasternak, S., Wheeler, D.A., Willis, T.D., Yu, F., Yang, H., Zeng, C., Gao, Y., Hu, H., Hu, W., Li, C., Lin, W., Liu, S., Pan, H., Tang, X., Wang, J., Wang, W., Yu, J., Zhang, B., Zhang, Q., Zhao, H., Zhou, J., Gabriel, S.B., Barry, R., Blumenstiel, B., Camargo, A., Defelice, M., Faggart, M., Goyette, M., Gupta, S., Moore, J., Nguyen, H., Onofrio, R.C., Parkin, M., Roy, J., Stahl, E., Winchester, E., Ziaugra, L., Altshuler, D., Shen, Y., Yao, Z., Huang, W., Chu, X., He, Y., Jin, L., Liu, Y., Sun, W., Wang, H., Wang, Y., Xiong, X., Xu, L., Waye, M.M., Tsui, S.K., Xue, H., Wong, J.T., Galver, L.M., Fan, J.B., Gunderson, K., Murray, S.S., Oliphant, A.R., Chee, M.S., Montpetit, A., Chagnon, F., Ferretti, V., Leboeuf, M., Olivier, J.F., Phillips, M.S., Roumy, S., Sallee, C., Verner, A., Hudson, T.J., Kwok, P.Y., Cai, D., Koboldt, D.C., Miller, R.D., Pawlikowska, L., Taillon-Miller, P., Xiao, M., Tsui, L.C., et al. (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature,* 449**,** 913-8.

Said-Mohamed, R., Allirot, X., Sobgui, M. & Pasquet, P. (2009) Determinants of overweight associated with stunting in preschool children of Yaounde, Cameroon. *Ann Hum Biol,* 36**,** 146-61.

Salem, A.H., Badr, F.M., Gaballah, M.F. & Paabo, S. (1996) The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. *Am J Hum Genet,* 59**,** 741-3.

Sanghera, D.K., Ortega, L., Han, S., Singh, J., Ralhan, S.K., Wander, G.S., Mehra, N.K., Mulvihill, J.J., Ferrell, R.E., Nath, S.K. & Kamboh, M.I. (2008) Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: PPARG2

(Pro12Ala), IGF2BP2, TCF7L2 and FTO variants confer a significant risk. *BMC Med Genet,* 9**,** 59.

Saxena, R., Gianniny, L., Burtt, N.P., Lyssenko, V., Giuducci, C., Sjogren, M., Florez, J.C., Almgren, P., Isomaa, B., Orho-Melander, M., Lindblad, U., Daly, M.J., Tuomi, T., Hirschhorn, J.N., Ardlie, K.G., Groop, L.C. & Altshuler, D. (2006) Common single nucleotide polymorphisms in TCF7L2 are reproducibly associated with type 2 diabetes and reduce the insulin response to glucose in nondiabetic individuals. *Diabetes,* 55**,** 2890-5.

Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U., Prokunina-Olsson, L., Ding, C.J., Swift, A.J., Narisu, N., Hu, T., Pruim, R., Xiao, R., Li, X.Y., Conneely, K.N., Riebow, N.L., Sprau, A.G., Tong, M., White, P.P., Hetrick, K.N., Barnhart, M.W., Bark, C.W., Goldstein, J.L., Watkins, L., Xiang, F., Saramies, J., Buchanan, T.A., Watanabe, R.M., Valle, T.T., Kinnunen, L., Abecasis, G.R., Pugh, E.W., Doheny, K.F., Bergman, R.N., Tuomilehto, J., Collins, F.S. & Boehnke, M. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science,* 316**,** 1341-5.

Scrimshaw, N.S. & Murray, E.B. (1988) The acceptability of milk and milk products in populations with a high prevalence of lactose intolerance. *Am J Clin Nutr,* 48**,** 1079-159.

Segal, K.R., Van Loan, M., Fitzgerald, P.I., Hodgdon, J.A. & Van Itallie, T.B. (1988) Lean body mass estimation by bioelectrical impedance analysis: a four-site cross-validation study. *Am J Clin Nutr,* 47**,** 7-14.

Seielstad, M.T., Minch, E. & Cavalli-Sforza, L.L. (1998) Genetic evidence for a higher female migration rate in humans. *Nat Genet,* 20**,** 278-80.

Sengupta, S., Zhivotovsky, L.A., King, R., Mehdi, S.Q., Edmonds, C.A., Chow, C.E., Lin, A.A., Mitra, M., Sil, S.K., Ramesh, A., Usha Rani, M.V., Thakur, C.M., Cavalli-Sforza, L.L., Majumder, P.P. & Underhill, P.A. (2006) Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. *Am J Hum Genet,* 78**,** 202-21.

Shahrani, M.N. (1979) *The Kirghiz and Wakhi of Afghanistan: Adaptation to closed frontiers and wars.*

Sharp, P.F. (1938) Relation between lactose and ash content of the milk of different mammals. *J Dairy Sci,* 21**,** 127-128.

Shilov, V.P. (1975) Models of pastoral economies in the steppe regions of Eurasia in the Eneolithic and early Bronze Ages. *Sovetskaya Arkeologiya,* 1**,** 5-16.

Simoons, F.J. (1969) Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. I. Review of the medical research. *Am J Dig Dis,* 14**,** 819-36.

Simoons, F.J. (1970) Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. II. A culture historical hypothesis. *Am J Dig Dis,* 15**,** 695-710.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S., Balkau, B., Heude, B., Charpentier, G., Hudson, T.J., Montpetit, A., Pshezhetsky, A.V., Prentki, M., Posner, B.I., Balding, D.J., Meyre, D., Polychronakos, C. & Froguel, P. (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*.

Sokal, R.R. (1988) Genetic, geographic, and linguistic distances in Europe. *Proc Natl Acad Sci U S A,* 85**,** 1722-6.

Sokoloff, L., Fitzgerald, G.G. & Kaufman, E.E. (1977) Cerebral nutrition and energy metabolism. In: *Nutrition and the brain* R.J. Wurtman & J.J. Wurtman (eds.) New York: Raven Press.

Soucek, S. (2000) *A history of inner Asia.* Cambridge.

Stearns, S.C. & Koella, J.C. (2007) *Evolution in health and disease.*

Stephens, M. & Donnelly, P. (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet,* 73**,** 1162-9.

Storz, J.F. (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Mol Ecol,* 14**,** 671-88.

Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123**,** 585-595.

Takayama, T., Fujita, K., Suzuki, K., Sakaguchi, M., Fujie, M., Nagai, E., Watanabe, S., Ichiyama, A. & Ogawa, Y. (2003) Control of oxalate formation from L-hydroxyproline in liver mitochondria. *J Am Soc Nephrol,* 14**,** 939-946.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., Ibrahim, M., Omar, S.A., Lema, G., Nyambo, T.B., Ghori, J., Bumpstead, S., Pritchard, J.K., Wray, G.A. & Deloukas, P.

(2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet,* 39**,** 31-40.

Tobi, E.W., Lumey, L.H., Talens, R.P., Kremer, D., Putter, H., Stein, A.D., Slagboom, P.E. & Heijmans, B.T. (2009) DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. *Hum Mol Genet,* 18**,** 4046-53.

Trinkaus, E. (2005) Early modern humans. *Ann Rev Anthropo,* 34**,** 207-230.

Tsalkin, V.I. (1964) Nekotorye itogi izucheniia kostnykh ostatkov zhivotnykh iz rskopok archeologicheskikh pamiatnikov pozdnego Bronzovogo veka.

Tseveendorj, D., Kubarev, V.D. & E., Y. (2005) *Aral Tolgoin khadny zurag.* Ulaanbaatar: Institute of Archaeology, Mongolian Academy of Sciences.

Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D.P., Holmkvist, J., Borch-Johnsen, K., Jorgensen, T., Sandbaek, A., Lauritzen, T., Hansen, T., Nurbaya, S., Tsunoda, T., Kubo, M., Babazono, T., Hirose, H., Hayashi, M., Iwamoto, Y., Kashiwagi, A., Kaku, K., Kawamori, R., Tai, E.S., Pedersen, O., Kamatani, N., Kadowaki, T., Kikkawa, R., Nakamura, Y. & Maeda, S. (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat Genet,* 40**,** 1098-102.

Vadetskaya, E.B. (1986) *Arkheologicheskie pamyatniki v stepyak srednego Yenisey.* Leningrad: Nauka.

Vander Molen, J., Frisse, L.M., Fullerton, S.M., Qian, Y., Del Bosque-Plata, L., Hudson, R.R. & Di Rienzo, A. (2005) Population genetics of CAPN10 and GPR35: implications for the evolution of type 2 diabetes variants. *Am J Hum Genet,* 76**,** 548-60.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., Gocayne, J.D., Amanatides, P., Ballew, R.M., Huson, D.H., Wortman, J.R., Zhang, Q., Kodira, C.D., Zheng, X.H., Chen, L., Skupski, M., Subramanian, G., Thomas, P.D., Zhang, J., Gabor Miklos, G.L., Nelson, C., Broder, S., Clark, A.G., Nadeau, J., Mckusick, V.A., Zinder, N., Levine, A.J., Roberts, R.J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A.E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T.J., Higgins, M.E., Ji, R.R., Ke, Z., Ketchum, K.A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin,

X., Lu, F., Merkulov, G.V., Milshina, N., Moore, H.M., Naik, A.K., Narayan, V.A., Neelam, B., Nusskern, D., Rusch, D.B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., et al. (2001) The sequence of the human genome. *Science,* 291**,** 1304-51.

Vitalis, R., Dawson, K. & Boursot, P. (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics,* 158**,** 1811-1823.

Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. (2006) A map of recent positive selection in the human genome. *PLoS Biol,* 4**,** e72.

Wang, Y.G., Yan, Y.S., Xu, J.J., Du, R.F., Flatz, S.D., Kuhnau, W. & Flatz, G. (1984) Prevalence of primary adult lactose malabsorption in three populations of northern China. *Hum Genet,* 67**,** 103-6.

Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol,* 7**,** 256-76.

Watterson, G.A. (1978) The Homozygosity Test of Neutrality. *Genetics,* 88**,** 405-417.

Wauters, M., Mertens, I., Rankinen, T., Chagnon, M., Bouchard, C. & Van Gaal, L. (2001) Leptin receptor gene polymorphisms are associated with insulin in obese women with impaired glucose tolerance. *J Clin Endocrinol Metab,* 86**,** 3227-32.

Weedon, M.N. (2007) The importance of TCF7L2. *Diabet Med,* 24**,** 1062-6.

Wells, R.S., Yuldasheva, N., Ruzibakiev, R., Underhill, P.A., Evseeva, I., Blue-Smith, J., Jin, L., Su, B., Pitchappan, R., Shanmugalakshmi, S., Balakrishnan, K., Read, M., Pearson, N.M., Zerjal, T., Webster, M.T., Zholoshvili, I., Jamarjashvili, E., Gambarov, S., Nikbin, B., Dostiev, A., Aknazarov, O., Zalloua, P., Tsoy, I., Kitaev, M., Mirrakhimov, M., Chariev, A. & Bodmer, W.F. (2001) The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci U S A,* 98**,** 10244-9.

Wendorf, M. (1989) Diabetes, the ice free corridor, and the Paleoindian settlement of North America. *Am J Phys Anthropol,* 79**,** 503-20.

Wilder, J.A. & Hammer, M.F. (2007) *Extraordinary population structure among the Baining of New Britain.*

Wilder, J.A., Kingan, S.B., Mobasher, Z., Pilkington, M.M. & Hammer, M.F. (2004a) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. *Nat Genet,* 36**,** 1122-5.

Wilder, J.A., Mobasher, Z. & Hammer, M.F. (2004b) Genetic evidence for unequal effective population sizes of human females and males. *Mol Biol Evol,* 21**,** 2047-57.

Wilkins, J.F. & Marlowe, F.W. (2006) Sex-biased migration in humans: what should we expect from genetic data? *Bioessays,* 28**,** 290-300.

Williams, H.E. & Smith, L.H. (1983) Primary Hyperoxaluria. In: *The Metabolic Basis of Inherited Disease* J.B. Stanbury, J.B. Wyngaarden, D.S. Frederickson, J.L. Goldstein & M.S. Brown (eds.) New York: McGraw-Hill.

Wood, W.A. (2002) Culture Summary: Turkmens.). New Haven, Conn.: HRAF, 2002.

Wright, S. (1931) Evolution in mendelian populations. *Genetics,* 16**,** 97-159.

Wright, S. (1951) The genetical structure of populations. *Ann Eugen,* 15**,** 323-354.

Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H., Hirota, Y., Mori, H., Jonsson, A., Sato, Y., Yamagata, K., Hinokio, Y., Wang, H.Y., Tanahashi, T., Nakamura, N., Oka, Y., Iwasaki, N., Iwamoto, Y., Yamada, Y., Seino, Y., Maegawa, H., Kashiwagi, A., Takeda, J., Maeda, E., Shin, H.D., Cho, Y.M., Park, K.S., Lee, H.K., Ng, M.C., Ma, R.C., So, W.Y., Chan, J.C., Lyssenko, V., Tuomi, T., Nilsson, P., Groop, L., Kamatani, N., Sekine, A., Nakamura, Y., Yamamoto, K., Yoshida, T., Tokunaga, K., Itakura, M., Makino, H., Nanjo, K., Kadowaki, T. & Kasuga, M. (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat Genet,* 40**,** 1092-7.

Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., De Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G., Ardlie, K., Bostrom, K.B., Bergman, R.N., Bonnycastle, L.L., Borch-Johnsen, K., Burtt, N.P., Chen, H., Chines, P.S., Daly, M.J., Deodhar, P., Ding, C.J., Doney, A.S., Duren, W.L., Elliott, K.S., Erdos, M.R., Frayling, T.M., Freathy, R.M., Gianniny, L., Grallert, H., Grarup, N., Groves, C.J., Guiducci, C., Hansen, T., Herder, C., Hitman, G.A., Hughes, T.E., Isomaa, B., Jackson, A.U., Jorgensen, T., Kong, A., Kubalanza, K., Kuruvilla, F.G., Kuusisto, J., Langenberg, C., Lango, H., Lauritzen, T., Li, Y., Lindgren, C.M., Lyssenko, V., Marvelle, A.F., Meisinger, C., Midthjell, K., Mohlke, K.L., Morken, M.A., Morris, A.D., Narisu, N., Nilsson, P., Owen, K.R., Palmer, C.N., Payne, F., Perry, J.R., Pettersen, E., Platou, C., Prokopenko, I., Qi, L., Qin, L., Rayner, N.W., Rees, M., Roix, J.J., Sandbaek, A., Shields, B., Sjogren, M., Steinthorsdottir, V., Stringham, H.M., Swift, A.J., Thorleifsson, G., Thorsteinsdottir, U., Timpson, N.J., Tuomi, T., Tuomilehto, J., Walker, M., Watanabe, R.M., Weedon, M.N., Willer, C.J., Illig, T., Hveem, K., Hu, F.B., Laakso, M., Stefansson, K., Pedersen, O., Wareham, N.J., Barroso, I., Hattersley, A.T., Collins, F.S., Groop, L., Mccarthy, M.I., Boehnke, M. & Altshuler, D. (2008) Meta-analysis of genome-wide association data and large-scale

replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet,* 40**,** 638-45.

Zerjal, T., Wells, R.S., Yuldasheva, N., Ruzibakiev, R. & Tyler-Smith, C. (2002) A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia. *Am J Hum Genet,* 71**,** 466-82.

Zerjal, T., Xue, Y., Bertorelle, G., Wells, R.S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., Fu, S., Li, P., Yuldasheva, N., Ruzibakiev, R., Xu, J., Shu, Q., Du, R., Yang, H., Hurles, M.E., Robinson, E., Gerelsaikhan, T., Dashnyam, B., Mehdi, S.Q. & Tyler-Smith, C. (2003) The genetic legacy of the Mongols. *Am J Hum Genet,* 72**,** 717-21.

Zhou, D., Zhang, D., Liu, Y., Zhao, T., Chen, Z., Liu, Z., Yu, L., Zhang, Z., Xu, H. & He, L. (2009a) The E23K variation in the KCNJ11 gene is associated with type 2 diabetes in Chinese and East Asian population. *J Hum Genet,* 54**,** 433-5.

Zhou, Q., Zhang, K., Li, W., Liu, J.T., Hong, J., Qin, S.W., Ping, F., Sun, M.L. & Nie, M. (2009b) Association of KCNQ1 gene polymorphism with gestational diabetes mellitus in a Chinese population. *Diabetologia*.

Zimmet, P., Alberti, K.G. & Shaw, J. (2001) Global and societal implications of the diabetes epidemic. *Nature,* 414**,** 782-7.

Zimmet, P., Dowse, G., Finch, C., Serjeantson, S. & King, H. (1990) The epidemiology and natural history of NIDDM--lessons from the South Pacific. *Diabetes Metab Rev,* 6**,** 91-124.

Zimmet, P., Serjeatson, S. & Dowes, G. (1991) Diabetes mellitus and cardiovascular disease in developing populations: hunter-gatherers in the fast lane. In: *Sugars in nutrition* N. Kretchmer & E. Rossi (eds.) New York: Nestlé Ltd.

Zimmet, P.Z. (1992) Kelly West Lecture 1991. Challenges in diabetes epidemiology--from West to the rest. *Diabetes Care,* 15**,** 232-52.

ANNEXES

- **Annexe 1 : Gène-éthique**

La génétique est une discipline imbibée d'histoire, et cette histoire diffère selon les sociétés humaines. En 1997, l'UNESCO (Organisation des Nations Unies pour l'Education, la Science et la Culture) fait une première déclaration universelle sur le génome humain et les droits de l'Homme (téléchargeable sur http://portal.unesco.org/fr/ev.php-URL_ID=13177&URL_DO=DO_TOPIC&URL_SECTION=201.html). Cette déclaration considère le génome humain comme patrimoine mondial de l'humanité, mais le reconnaît également comme caractéristique unique de chaque individu. C'est justement toute la dualité du matériel génétique : il constitue la partie la plus privée et la plus commune de notre biologie. De la même manière, il est à la fois une partie du corps humain, et une information sur l'individu (ou le groupe), deux composantes normalement bien différenciées dans la législation. La difficulté est alors de nouer ces deux caractéristiques pour en faire une problématique unique. C'est pourquoi, en 2003, l'UNESCO a fait une nouvelle déclaration universelle (http://portal.unesco.org/fr/ev.php-URL_ID=17720&URL_DO=DO_TOPIC&URL_SECTION=201.html), considérant cette fois-ci les données génétiques humaines dans leur ensemble, ce qui n'est pas le cas des lois nationales. En effet, en France, quand on met en place un terrain de recherche, il faut déclarer d'une part les données collectées au CNIL (Comité National sur l'Information et la Liberté), pour la législation concernant les conditions de stockage et d'utilisation des données « identifiantes », et d'autre part auprès de l'organisme de recherche en tant que collections d'ADN (si il y a stockage). Si la recherche se fait à l'étranger, il faut en plus obtenir une autorisation d'import / export des échantillons biologiques. Il faut également faire appel à un comité d'éthique local, qui est l'organisme central pour la prise de décision du protocole à suivre. Cependant, tous les pays n'ont pas de comité d'éthique propre (ou certains ne sont pas libres et indépendants) et le comité d'éthique français, consultatif, ne peut pas donner son avis sur un protocole à l'étranger, puisqu'il n'est pas considéré comme compétent dans d'autres pays. C'est notamment le cas au Kirghizistan et en Ouzbékistan, où, respectivement, le Ministère de la Recherche et l'Académie des Sciences ont fait office d'organismes de référence locaux pour la validation des projets de recherche. Dans ces cas où il n'y pas de comité d'éthique locaux, il faut faire référence à la déclaration d'Helsinki (http://www.wma.net/fr/30publications/10policies/b3/17c_fr.pdf), que chaque chercheur, dans son contrat, s'engage à respecter. Cette déclaration s'adresse principalement aux médecins,

mais également aux autres acteurs de la recherche, surtout quand leurs études impliquent des êtres humains.

Déclaration d'Helsinki

Nous pouvons retirer de cette déclaration plusieurs principes généraux : le devoir du médecin de protéger la dignité et l'intégrité des personnes impliquées (article 11) ; la soumission du projet à un comité d'éthique indépendant, prenant en compte les législations propres à chaque pays ainsi que les normes internationales si elles n'excluent pas les protections ici garanties (article 15) ; la protection de la confidentialité des informations et de la vie privée des personnes (article 23) ; la nécessité d'un libre consentement de la personne pour participer à l'étude (article 22), consentement qui doit être éclairé (article 24), indépendant et obtenu sans contraintes (article 26), et explicitant l'analyse, le stockage et la réutilisation de matériels d'origine humaine (article 25) ; enfin la mise à disposition des résultats au public concerné (article 30). Notons qu'il est également écrit que « Dans la pratique médicale et la recherche médicale, la plupart des interventions comprennent des risques et des inconvénients » (article 8) et « Une recherche médicale impliquant des êtres humains ne peut être conduite que si l'importance de l'objectif dépasse les risques et inconvénients inhérents pour les personnes impliquées dans la recherche» (article 21).

Ainsi, aucune consigne rigide n'est donnée, mais des problématiques sont soulevées, auxquelles chacun, dans son propre terrain, est amené à réfléchir. Cette souplesse semble être le reflet de la forte variabilité des contraintes et des perceptions dans les différentes sociétés humaines. De plus, comme le signale le dernier article évoqué, certaines valeurs éthiques sont toujours en conflit avec d'autres, et les décisions à prendre ne sont jamais triviales. C'est donc à la morale et la responsabilité individuelle que cette déclaration fait appel, en demandant de hiérarchiser les valeurs entre elles, sans pour autant leur enlever de l'importance.

Au final, les questions majeures qui se sont posées lors de ces missions sont : comment explique-t-on la problématique de recherche aux individus échantillonnés et à quel niveau de compréhension prétend-on ? Quel est le niveau de détail d'informations à fournir pour le consentement écrit ? Quel suivi doit-on assurer après la détection de problèmes de santé ? Quelle compensation doit-on proposer ? A qui ? Comment rapporter les résultats obtenus à posteriori ?

Notre approche

Les réponses à ces questions dépendent de beaucoup de facteurs, et l'important est bel et bien de se les poser avant toute mise au point de protocoles de recherche. En tout cas, concernant notre protocole, l'échantillonnage a été précédé d'une longue phase d'explication orale du projet aux médecins et infirmiers locaux avec qui nous travaillions, qui nous posaient de nombreuses questions pour bien comprendre les enjeux de recherche de ce projet. Ces médecins et infirmiers locaux ont ensuite été chargés d'expliquer l'étude aux autres individus pour les recruter. Nous leur avons ainsi laissé la responsabilité du niveau de détail expliqué et de la manière d'aborder nos problématiques de recherche. Nous avons ainsi privilégié une explication orale du projet de recherche aux individus échantillonnés. Ensuite, pour chaque individu venant participer à notre projet, nous avons fait signer un consentement écrit en lui expliquant de nouveau plus brièvement le projet. Le consentement explicite de plus que les échantillons sanguins seront conservés et potentiellement réutilisés pour d'autres projets de recherche. Nous avons bien précisé aux individus participant que leurs facteurs de risque du diabète de type II ne leur serait pas communiqués sous forme individuelle, mais plutôt sous forme populationnelle. Concernant la détection de problèmes de santé, nous nous sommes une fois de plus reposés sur les médecins locaux pour définir les démarches à suivre pour chaque individu. Pour les compensations, nous avons eu la chance d'avoir un protocole de recherche qui lui-même constitue une sorte de visite chez le médecin, donc très prisé par les populations locales. Chaque individu échantillonné est en effet reparti avec des données sur ses facteurs de risque à diverses maladies (obésité, hyperglycémie, hypertension). Les personnes étaient donc naturellement motivées pour participer à notre étude. Nous avons parfois fourni du matériel aux hôpitaux en guise de compensation, selon leurs nécessités, ainsi qu'un salaire aux médecins et infirmières effectuant les enquêtes avec nous. Les médecins ouzbèks et kirghiz avec qui nous avons travaillé ont eu par la suite accès à toutes nos données saisies électroniquement, ainsi qu'aux résultats des dosages physiologiques récupérés par la suite. Nous leur avons laissé la responsabilité de faire remonter ces informations aux personnes appropriées.

Finalement, bien que je ne sache pas vraiment dans quelle mesure les individus échantillonnés ont réellement compris l'essence de notre problématique de recherche, je suis persuadée que leur participation était libre et indépendante, et que leur dignité et vie privée ont été respectées.

- **<u>Annexe 2</u> : Questionnaires**

<u>Questionnaire alimentaire qualitatif « de fréquence » :</u>

Jamais	- d'1 fois par mois	1-3 fois par mois	1-2 fois par semaine	3-5 fois par semaine	Presque chaque jour
1	2	3	4	5	6

ALIMENTS		**ALIMENTS**	
Riz			
Pates		Fruits secs	
Pain		Amandes	
Gâteaux		Noix	
Céréales		Arachide	
Pommes de terre		Tomate / Concombre / Salade	
		Epinard	
Mouton		Betterave / Navet / Chou	
Chèvre		Aubergine / Poivrons	
Cheval		Radis	
Lapin		Carottes	
Bœuf		Haricot / Pois	
Dinde		Oignons	
Poulet		Ail	
Porc			
Œufs		Limonades	
Poisson		Coca / Fanta / Pepsi	
Lait frais		Bière	
Crème de lait		Vodka	
Kéfir / Aïran		Sucre	
Yaourt		Sel	
Fromage			
Fromage blanc			
Beurre		Cigarette	
Autres ?		Tabac à chiquer	

Questionnaire alimentaire quantitatif :

Nom, prénom, patronyme sexe âge
N°
Lieu d'enquête :

QUESTIONNAIRE ALIMENTAIRE :

Combien de fois par semaine mangez-vous à l'extérieur ?

Faites-vous la cuisine à la maison ? Oui/Non Si oui, quoi ?

Combien de personnes mangent chaque jour à la maison ? Combien d'adultes ?

d'enfants ?

Combien de fois (par semaine, par mois) allez-vous au marché ?

Combien consommez-vous de litres d'huile par semaine ?

Combien de kg de viande achetez-vous par semaine (bœuf, mouton, poulet…)

Combien de fois mangez-vous du poisson par semaine?

Combien de kg de sucre achetez-vous par semaine ?

Combien de litres de lait et de produits laitiers (katyk, kaïmok ou tchakka) achetez-vous par semaine ?

Combien de litres de lait et de produits laitiers (katyk, kaïmok ou tchakka) produisez-vous à la maison par semaine ?

Combien achetez-vous de farine par semaine ou par mois ?

Combien achetez-vous de riz par semaine ou par mois ?

Combien achetez-vous de pomme de terre par semaine ?

Combien achetez-vous de pâtes par semaine ?

Combien achetez-vous d'œufs par semaine ?

Combien achetez-vous de pain par semaine ?

Questionnaire médical et sur l'activité physique :

ENQUETE MEDICALE :

Avez-vous des problèmes de santé :

- Tension : **oui/non** traitement ?
- Niveau élevé de sucre/diabète : **oui/non** traitement ?
- Cholestérol : **oui/non** traitement ?
- Maladies cardio-vasculaires : **oui/non** traitement ?

Autres maladies : **oui/non** traitement ?

Avez-vous été malade depuis 1 mois : **oui/non** de quoi ? traitement ?

Avez-vous un problème médical actuellement : **oui/non** traitement ?

Y a-t-il des maladies chroniques dans votre famille (diabète, cardio-vasculaire) (enfants, parents) :

oui/non Qui ? Quelle maladie ?

QUESTIONS SUR LA CONDITION PHYSIQUE :

Quelle est votre profession ?

Votre niveau d'activité physique, votre mode de vie est-il à votre avis :
inactif peu actif actif très actif

Comment vous rendez-vous à votre travail (à pied, à vélo, en autobus…) ?

Faites-vous du sport ? Lequel ? Combien d'heures par semaine ?

- **<u>Annexe 3</u>** : Martinez-Cruz B.[*], Vitalis R.[*], **Ségurel L.**, Austerlitz F., Georges M., Théry S., Quintana-Murci L., Hegay T., Aldashev A., Nazyrova F. & Heyer E. In the heartland of Eurasia : the multi-locus genetic landscape of Central Asian populations. A soumettre à European Journal of Human Genetics

[*] Ces auteurs ont également contribué à ce travail

In the heartland of Eurasia: the multilocus genetic landscape of Central Asian populations

Begoña Martínez-Cruz[1,2,9], Renaud Vitalis[1,3,9], Laure Ségurel[1], Frédéric Austerlitz[4], Myriam Georges[1], Sylvain Théry[1], Lluis Quintana-Murci[5], Tatyana Hegay[6], Almaz Aldashev[7], Firuza Nasyrova[8], Evelyne Heyer[*,1]

[1]Muséum National d'Histoire Naturelle – Centre National de la Recherche Scientifique Université Paris 7, UMR 7206, « Éco-Anthropologie et Ethnobiologie », CP 139, 57 rue Cuvier, 75231 Paris Cedex 05, France

[2]Current address: Evolutionary Biology Institute, Pompeu Fabra University – CSIC – PRBB, Dr. Aiguader 88, 08003 Barcelona, Spain

[3]Current address: Institut National de la Recherche Agronomique, UMR CBGP (INRA – IRD – CIRAD – Montpellier SupAgro), Campus International de Baillarguet, CS 30016, 34988 Montferrier-sur-Lez, France

[4]Université Paris Sud, CNRS UMR 8079, Laboratoire Écologie, Systématique et Évolution, 91405 Orsay, France

[5]Human Evolutionary Genetics, CNRS URA3012, Institut Pasteur, 75015 Paris, France

[6]Uzbek Academy of Sciences, Institute of Immunology, Tashkent 700060, Uzbekistan

[7]National Center of Cardiology and Internal Medicine, Bishkek 720040, Kyrgyzstan

[8]Tajik Academy of Sciences, Institute of Plant Physiology and Genetics, Dushanbe 734063, Tajikistan

[9]These authors contributed equally to the present study

[*]Correspondence:

Evelyne Heyer

e-mail: heyer@mnhn.fr

Phone:

Fax:

Running title: Multilocus genetic landscape in Central Asia

Keywords: admixture; Central Asia; F_{ST}; microsatellites; isolation by language;

Abstract

Located in the Eurasian heartland, Central Asia has played a major role in both the early spread of modern humans out of Africa and the more recent settlement of differentiated populations. A detailed knowledge of the peopling in this vast region would therefore greatly improve our understanding of range expansions, colonizations, and recurrent migrations, including the impact of the expansion of eastern nomadic groups that occurred in Central Asia. However, despite its presumable importance, little is known about the level and the distribution of genetic variation in this region. We report here the first multilocus genetic survey of Central Asian populations. We genotyped 26 Indo-Iranian- and Turkic-speaking populations, belonging to six different ethnic groups, at 27 autosomal microsatellite loci. We found high levels of genetic diversity and our results show that genetic differentiation across populations is mainly shaped by linguistic and ethnic affiliation, rather than by geography. The analysis of genetic variation clearly puts Central Asia in an intermediate position between Europe and Middle East, Central-South Asia, and East Asia. Turkic-speaking populations are more closely related to East Asian populations, while Indo-Iranian speakers cluster with western Eurasians. Uzbeks are scattered across Turkic- and Indo-Iranian speaking populations, which may reflect their origins from the union of different tribes. This mixed ancestry of Central Asian populations combined with the level of genetic differentiation among the two groups, let us propose the following scenario for the population history in CA: a long term settled group of what is nowadays Tajik populations and a more recent arrival of eastern populations comprising the Turkic group. Our results also show that the expansion of eastern nomadic groups, contrary to what is generally thought, did not result in the complete replacements of the local populations but rather to a process of "soft invasion" that includes admixture.

Introduction

The evolutionary history of modern humans has been characterized by range expansions, colonizations, and recurrent migrations over the last 100,000 years.{Cavalli-Sforza, 1994 #1} Some regions of the world that have served as natural corridors between landmasses are of particular importance in the history of human migrations. Central Asia is probably one of such migration routes.{Cavalli-Sforza, 1994 #1;Nei, 1993 #50;Cavalli-Sforza, 1994 #1;Nei, 1993 #50}.Located in the Eurasian heartland, it encompasses a vast territory, limited to the east by the Pamir and Tien-Shan mountains, to the west by the Caspian Sea, to the north by the Russian taiga and to the south by the Iranian deserts and Afghan mountains. Despite its presumable importance, though, the precise role of Central Asia in both the early spread of modern humans out of Africa and the more recent settlement of differentiated populations{Comas, 1998 #14;Comas, 1998 #14} remains unclear.{Cordaux, 2004 #17;Karafet, 2001 #23;Wells, 2001 #22;Cordaux, 2004 #17;Karafet, 2001 #23;Wells, 2001 #22}

Central Asia entered the historical records about 1300 BC when Arian tribes invaded Iran territory from what is nowadays Turkmenistan and established the Persian Empire in the 7th BC. Around the 7th Century B.C these people, called the Scythians were described in ancient Chinese texts and in Herodotus' *Histories* as having European morphological traits and speaking Indo-Iranian languages. Further, multiple waves of Turco-Mongol invaders arrived in Central Asia, although it is difficult to know precisely when these migrations from the East started. Russian sources {Гумилев Л. Н. / Древние тюрки /АН СССР. Ин-т народов Азии. - М.: Наука, #74}document that in the Second-First century BC, "Hun" brought the east-Asian anthropological phenotype to Central Asia. At the same period, the Chinese established a trade route (the Silk Road), which connected the Mediterranean Basin and Eastern Asia for more than 16 centuries.

Later on, the Turco-Mongol Empire became the largest empire of all time, from Mongolia to the Black Sea, following Genghis Khan's invasions in the 13th century A.D. All these movements of populations resulted in a considerable ethnic diversity in Central Asia, including Indo-Iranian speakers living as sedentary agriculturalists and Turkic speakers living as traditionally nomadic herders.

This intricate demographic history has shaped patterns of genetic variability in a complex manner, yet little is known about the genetic diversity in this region. Most previous studies, which were based on classical markers{Cavalli-Sforza, 1994 #1;Cavalli-Sforza, 1994 #1}, mitochondrial DNA (mtDNA){Chaix, 2004 #6;Chaix, 2007 #4;Comas, 1998 #14;Comas, 2004 #7;Lalueza-Fox, 2004 #18;Perez-Lezaun, 1999 #13;Chaix, 2004 #6;Chaix, 2007 #4;Comas, 1998 #14;Comas, 2004 #7;Lalueza-Fox, 2004 #18;Perez-Lezaun, 1999 #13}, or the non-recombining portion of the Y-chromosome (NRY){Chaix, 2004 #6;Chaix, 2007 #4;Hammer, 2001 #24;Wells, 2001 #22;Zerjal, 2002 #20;Chaix, 2004 #6;Chaix, 2007 #4;Hammer, 2001 #24;Wells, 2001 #22;Zerjal, 2002 #20}, have shown that genetic diversity in Central Asia is among the highest in Eurasia.{Comas, 1998 #14;Hammer, 2001 #24;Wells, 2001 #22;Comas, 1998 #14;Hammer, 2001 #24;Wells, 2001 #22} NRY-based studies suggest that various colonization waves of Eurasia originated in Central Asia[6], while mtDNA-based studies point to an admixed origin of Central Asian populations[10] from previously differentiated Eastern and Western Eurasian populations. Although these uniparentally inherited markers provide valuable information for understanding the role of Central Asia in the early colonization of the world, evaluating the impact of more recent population movements, including the expansion of eastern nomadic groups, require additional, fast-evolving, molecular markers. Furthermore, each of the NRY and mtDNA is a single non-recombining genetic locus, which contains several linked genes that may have been shaped by selection.{Ballard, 2004 #71;Balloux, 2009 #69;Bazin, 2006 #15;Pakendorf, 2005 #72} Using multiple unlinked markers is therefore required to infer with good accuracy the human demographic history in this region.

Here we report on the first multilocus autosomal genetic survey of Central Asian populations. Twenty-six populations belonging to six different ethnic groups were genotyped at 27 autosomal microsatellite markers. We aimed to characterize the genetic origins of Central Asian populations, and to investigate the extent to which historical invasions from the east, have shaped the Central Asian genetic landscape.

Material and methods

DNA samples

We sampled 767 adult males from 26 populations from western Uzbekistan to eastern Kyrgyzstan (Table 1 and Figure 1). The individuals sampled were representative of the ethnological diversity in Central Asia: Tajiks, which are Indo-Iranian speakers (a branch of the Indo-European language family) and Kazakhs, Turkmen, Karakalpaks, Kyrgyz and Uzbeks, which are Turkic speakers (a branch of the Altaic language family). In two Uzbek populations from the Bukhara area (LUZa and LUZn), an extensive linguistic survey showed that individuals were bilingual. Since their home language was Tajik (an Indo-Iranian language), we further classified these two populations into the Indo-Iranian group for subsequent analyses. We collected ethnologic data prior to sampling, including the recent genealogy of the participants. Using this information, we retained only individuals that were unrelated for at least two generations back in time. All individuals gave informed consent for participation in this study. Total genomic DNA was isolated from blood samples by a standard phenol-chloroform extraction.{Maniatis, 1982 #52}

Genotyping

We used 27 microsatellite markers{Segurel, 2008 #61} from the set of 377 markers used in the worldwide study by Rosenberg *et al.*{Rosenberg, 2002 #10;Rosenberg, 2002 #10} The description of markers, PCR and electrophoresis conditions are given in Ségurel *et al.*{Segurel, 2008 #61} Apart from the 767 individuals sampled in Central Asia, we further genotyped 20 individuals from three populations of the HGDP-CEPH Human Genome Diversity Cell Line Panel{Cann, 2002 #21;Rosenberg, 2002 #10;Zhivotovsky, 2003 #62} at the 27 microsatellite loci. We then compared the individual genotypes obtained with the H952 subset of the original HGDP-CEPH Human Genome Diversity Cell Line Panel.{Rosenberg, 2006 #5} For most loci, we detected systematic shifts in allele calling from 1 to 8 bp. We then corrected the whole Central Asian dataset for these systematic shifts, in order to standardize the Central Asian data with the worldwide HGDP-CEPH Human Genome Diversity Cell Line Panel data.

Data analyses

Genetic diversity

In each population, we calculated the allelic richness (*AR*){ElMousadik, 1996 #28} with the software package FSTAT{Goudet, 1995 #29} and unbiased estimates of expected heterozygosity (H_e){Nei, 1978 #32} with GENETIX.{Belkhir, 1996-2004 #53} We tested heterogeneity in both *AR* and H_e across populations using the Kruskal-Wallis test. Allelic richness and expected heterozygosity were also estimated for populations pooled into Indo-Iranian- and Turkic-speaking groups. We tested between-group differences in both *AR* and H_e using Wilcoxon's signed-rank test. We further estimated *AR* and H_e over the pooled data from Central Asia, and over the pooled data for Europe, the Middle-East and East Asia from the HGDP-CEPH Human Genome Diversity Cell Line Panel. We tested heterogeneity in both *AR* and H_e across the four groups of Eurasian populations using the Kruskal-Wallis test. When necessary, a posteriori test (Tukey test) is conducted. All statistical analyses were performed with the software package JMP5.1 (SAS Institute Inc.).{Inc., 2003 #63}

Genetic structure

Population differentiation was calculated overall and between pairs of Central Asian populations, using Weir and Cockerham's estimator of the parameter F_{ST}.{Weir, 1984 #30} Exact tests of differentiation were performed with FSTAT.{Goudet, 1995 #29;Goudet, 1995 #29} A principal component analysis (PCA) based on pairwise F_{ST} estimates was performed, using the software package GenAlex.{Peakall, 2006 #16;Peakall, 2006 #16} The population structure was also inferred by means of a hierarchical analysis of molecular variance (AMOVA{Excoffier, 1992 #64}), with populations pooled into ethnic or linguistic groups. For ethnic grouping, populations were pooled as Tajiks (TJA, TDS, TJT, TJK, TJR, TJN, TDU, TJE, TJY and TJU), Karakalpaks (KKK and OTU), Kazakhs (KAZ and LKZ), Kyrgyz (KRA, KRG, KRL, KRB, KRT and KRM), Uzbeks (UZA, UZB, LUZa, LUZn and UZT) and Turkmen (TUR). For linguistic grouping, populations were pooled as Indo-Iranian speakers (Tajiks and the two Uzbek populations LUZa and LUZn) and Turkic speakers (all other populations). These analyses were performed with ARLEQUIN 3.11.{Excoffier, 2005 #54;Excoffier, 2005 #54} Isolation by distance (IBD) was examined by testing the correlation between pairwise multilocus estimates of F_{ST} / (1 - F_{ST}) against the natural logarithm of the geographical distances between them.{Rousset, 1997 #27} We used the Mantel permutation procedure{Mantel, 1967 #55}, as implemented in GENEPOP 4.0{Rousset, 2008 #65}, to test the null hypothesis that the regression slope is zero.

Clustering analyses

We performed a series of clustering analyses with the software package STRUCTURE{Pritchard, 2000 #25;Pritchard, 2000 #25} on (*i*) the Central Asian populations only, and (*ii*) the Central Asian populations and all the Eurasian populations from the HGDP-CEPH Human Genome Diversity Cell Line Panel. All analyses were performed with the corrected H952 dataset.{Rosenberg, 2006 #5;Rosenberg, 2006 #3} STRUCTURE performs a Bayesian analysis, assigning individuals to clusters, within which both Hardy-Weinberg disequilibrium and linkage disequilibrium across loci are minimized. Each Markov chain was run for 10^6 steps, after a 10^5-step burn-in period. In each case, the results were checked to ensure consistency over ten independent runs. As this analysis requires a hypothetical number *K* of clusters to be selected from the outset, we varied the number (*K*) of *a priori* groups from 1 to 5 for analyses of Central Asian populations only, and from 1 to 5 if all Eurasian populations were included. All chains were run using the *F* model for correlations of allele frequencies across clusters{Falush, 2003 #42;Falush, 2003 #42}, accounting for potential admixture events in the past. We performed 10 runs for every K assayed as for the analyses of Central Asia alone and for the analyses of the Eurasian populations.

Admixture analysis

Admixture analyses can be used to estimate the relative contribution of putative parental populations to an admixed population. The Central Asian genetic pool may be more than just the result of admixture from Eurasian populations, but we were nonetheless interested in investigating the potential origins of Central Asian populations among all Eurasian populations. We used the software package LEADMIX{Wang, 2003 #43;Wang, 2003 #43} to calculate maximum likelihood estimates (MLE) of the admixture proportions for each Central Asian population. We ran the program independently for each of them, considering four putative parental groups from populations from the HGDP-CEPH Human Genome Diversity Cell Line Panel: Central South Asia, East Asia, Europe and Middle East. For the Central South Asian group, we chose a pool of Balochi (n = 25) and Brahui (n = 25) individuals, both populations being non-significantly differentiated (F_{ST} = 0.001; exact test p = 0.097). We chose the Han Chinese (n = 44) to be representative of East Asia. We further considered a pool of French (n = 28) and Russian (n = 25) individuals as representative of Europe, both populations being non-significantly differentiated (F_{ST} = 0.007; p = 0.081).

Last, the Palestinians (n = 46) were considered as representative of the Middle East.{Belle, 2006 #66}

Results

Genetic diversity

Allelic richness and expected heterozygosity for each of the 26 Central Asian populations and for the average values across regions are given in Table 2. We found a significant difference in allelic richness (Kruskal-Wallis test, χ^2 = 67.38, d.f. = 25, p < 0.0001) and in expected heterozygosity (Kruskal-Wallis test, χ^2 = 99.87, d.f. = 25, p < 0.0001) across populations. We found no significant difference in allelic richness between Indo-Iranian (*AR* = 13.8) and Turkic speakers *(AR* = 13.8, Wilcoxon signed rank test, Z = -0.67, p = 0.50), although the expected heterozygosity significantly differed between the two groups (H_e = 0.816 and H_e = 0.789, respectively, Wilcoxon signed rank test, Z = -4.61, p < 0.0001). However, we found a significant difference in allelic richness across Central Asia, Europe, Central-South Asia, Middle East and East Asia (Kruskal-Wallis test, χ^2 =51.06, d.f. = 4, p < 0.0001), as well as in expected heterozygosity (Kruskal-Wallis test, χ^2 = 56.57, d.f. = 4, p < 0.0001). These differences were due to a lower diversity in East Asia (Tukey HDS test, p < 0.0001 for both *AR* and H_e).At this point, Central Asia shows neither higher nor lower diversity than the rest of Eurasia (except for the case of East Asia).

Population differentiation

The 26 Central Asian populations were slightly but significantly differentiated (F_{ST} = 0.014, $CI_{99\%}$ = [0.011-0.017], p < 0.01). Pairwise F_{ST} estimates ranged from -0.004 to 0.053, with 196 out of 325 pairs of populations (i.e., 60.3%) being significantly differentiated (see Table 3). We found no evidence of IBD (Mantel test, p = 0.25), and no obvious relationship between genetic differentiation and geography. The apportionment of genetic variation among linguistic or ethnic groups of populations (Table 4) showed that more than 98% of the total variation laid within populations (p < 0.0001). Yet, both ethnicity and linguistic affiliation accounted significantly for the observed variation (F_{CT} = 0.005, p = 0.0023 and F_{CT} = 0.008, p < 0.0001, respectively).

The principal component analysis (PCA) based on pairwise F_{ST} estimates between Central Asian populations separated Turkic- and Indo-Iranian-speaking populations on the first axis (Figure 2a). The first two axes explained 68.8 % of the total variance. There were some exceptions, though: two Turkic-speaking populations, KRM and TUR, were clearly clustered with Indo-Iranian-speaking populations, and the Indo-Iranian-speaking population

TJK was found to be closer to Turkic-speaking than to other Indo-Iranian-speaking populations. Interestingly, the Uzbek populations (LUZa, LUZn, UZA and UZT) showed a scattered pattern on the PCA, which overlapped the Turkic-speaking and the Indo-Iranian-speaking groups of populations. The PCA based on pairwise F_{ST} estimates between all Eurasian populations placed Central Asian populations in an intermediate position between a cluster of European and Middle Eastern populations, a group of Central-South Asian populations, and a group of East Asian populations (Figure 2b). The first two axes explained 84.2 % of the total variance. Turkic- and Indo-Iranian-speaking populations were also separated on the first axis, with Turkic-speaking populations being closer to East Asian populations, and Indo-Iranian-speaking populations being closer to Central-South Asian populations and the group of European and Middle Eastern populations. It is noteworthy that Central Asian populations were more scattered than any other group of populations in Eurasia (Figure 2b). Interestingly, the Hazaras from Pakistan, who claim to be direct male-line descendants of Genghis Khan{Qamar, 2002 #44;Zerjal, 2003 #19;Qamar, 2002 #44;Zerjal, 2003 #19}, clustered together with the Turkic-speaking populations of Central Asia. Note that we excluded the highly differentiated Kalash population from the analysis, since it explained almost all the variance on the second axis.

Cluster analysis

We found no evidence of population clustering in Central Asia with STRUCTURE (not shown), which is consistent with the low, although significant, differentiation observed (see Table 3) and the limited number of markers used. Indeed, none of the 50 STRUCTURE runs assuming more than one putative group (with K varying from 2 to 6) produced higher posterior probability of the data than the 10 runs with a single putative group (K = 1). Analyzing Eurasian populations altogether, we found that the highest average posterior probability of the data, across 10 runs, was obtained for K = 2 putative clusters. Central Asia seemed to be intermediate between one cluster made of European, Middle Eastern and Central-South Asian populations on the one hand, and one cluster of East Asian populations on the other hand (Figure 3), which is consistent with the PCA (Figure 2b). All individuals from Central Asia belonged to these two main clusters for K = 2, with no individuals assigned completely to a single cluster, with Turkic-speaking individuals having a higher membership coefficient in the East Asian cluster, and Indo-Iranian-speaking individuals having a higher membership coefficient in the cluster made of Europe, Middle East and Central-South Asia.

Whether this mixed ancestry of Central Asian populations is due to recent admixture or to shared ancestry with easternmost and westernmost regions of Eurasia can hardly been deciphered, though{Li, 2008 #67}. Increasing the number of putative clusters to $K = 3$ separated the Kalash population as a distinct cluster, while further increasing K did not challenge the clustering observed at $K = 3$, but rather introduced some noise.

Admixture analysis

The MLE of admixture proportions obtained with LEADMIX for each Central Asian population assuming four putative parental populations (Central-South Asia, East-Asia, Europe and Middle-East) are given in Figure 1 and Table 5. Most Turkic-speaking populations had a large East Asian ancestral contribution, which represented, in general, 50% or more of the total contribution. There were three notable exceptions, though, with KRM, TUR and UZA showing evenly distributed admixture proportions across the putative parental populations. Indo-Iranian-speaking populations had a large western Eurasian contribution (Central-South Asia, Europe and Middle-East), which represented more than 70% of the total contribution (with the notable exception of TJK). The Uzbek populations LUZa and LUZn showed the same pattern, with a low ancestral contribution from East Asian populations (similar in magnitude to most of the Indo-Iranian-speaking populations), which gives further support to their classification as Indo-Iranians.

It is noteworthy that, in general, many geographically close populations showed contrasted admixture proportions (see, e.g., UZT and TJU), which supports the absence of a simple relationship between genetics and geography. The observed differences in the ancestral contributions between Indo-Iranian- and Turkic-speakers (Table 5) further support the idea that language, like ethnicity, is a major determinant of population differentiation in Central Asia.

Discussion

We studied 26 populations representative of the ethnological and linguistic diversity of Central Asia. We characterized the level and distribution of genetic diversity across these populations, using a set of 27 autosomal microsatellite loci. We also combined our data with a subset of the HGDP-CEPH Human Genome Diversity Cell Line Panel, in order to compare the genetic diversity in Central Asia with that observed in Eurasia, and to infer the genetic structure of Central Asian populations within Eurasia. This study undoubtedly fills a gap, as this is the first time that Central Asia has been incorporated into a large autosomal survey.{Rosenberg, 2002 #10;Li, 2008 #67}

Central Asia in the heartland of Eurasia:

We found a high level of genetic diversity in Central Asia, which is consistent with the data available for various genetic systems.{Comas, 1998 #14;Zerjal, 2002 #20;Comas, 1998 #14;Zerjal, 2002 #20} Population differentiation among Central Asian populations was similar, or even stronger, than that measured among populations from other regions in Eurasia (not shown). This pattern is apparent in the PCA (Figure 2b), where Central Asian populations are more scattered than each of the Central South Asian, East Asian, European and Middle-Eastern groups, which suggests a higher diversification in Central Asia. This observed diversity was mainly due to the differentiation between Indo-Iranian-speaking and Turkic-speaking populations: Indo-Iranian-speaking populations are genetically closer to Western Eurasian populations, whereas Turkic-speaking populations are more similar to Eastern Asian populations (Figure 2b). This pattern was also observed with clustering analysis performed with the software package STRUCTURE on all Eurasian populations (Figure 3). We also found significant pairwise F_{ST} estimates between almost all pairwise comparisons between an Indo-Iranian-speaking population (Table 3) and a Turkic-speaking population, which indicates a significant differentiation between Indo-Iranian speakers and Turkic speakers.

Although several studies have shown that geography is, in general, a better determinant of genetic differentiation than ethnicity and linguistics,{Bosch, 2006 #33;Manica, 2005 #34;Bosch, 2006 #33;Manica, 2005 #34} we found no obvious geographical pattern of genetic diversity in Central Asia. This absence of correlation between

geography and genetics in Central Asia, which makes it an exception amongst other regions in the world, might be explained by several non-mutually exclusive factors: first, the arrival of the Turkic-speaking populations in the region may be recent, and thus the genetic signature of isolation by distance did not have time to build up. Second, the nomadism of the Turkic-speaking populations may frequently reshuffle any correlation between genetics and geography. In particular, the compulsory sedentarisation of Turkic-speakers during the Soviet era may have canceled any geographic pattern of genetic variation. Last, mate choice may not depend much on geography. Even more striking is the fact that no geographic pattern of genetic variation was found among sedentary Indo-Iranian speakers either.

Putative origins of Indo-Iranian- and Turkic-speaking populations

We found significant pairwise F_{ST} estimates between almost all pairs of Indo-Iranian-speaking populations (Table 3). Furthermore, admixture analyses suggested diverse putative origins for the Indo-Iranian-speaking populations (Table 5), most of which being very close to the Central South Asian group of populations in the PCA (Figure 2b). This high level of diversity among Tajik populations and the variable level of admixture from the paternal populations are consistent with the hypothesis that Indo-Iranian speakers are the long term settled populations in the area. This hypothesis is strongly supported by archaeological evidence.{Brunet, 1999 #56;Brunet, 1999 #56}

Conversely we found a lower genetic differentiation among Turkic-speaking populations despite their wide geographic distribution (Figure 1), which suggests a more recent common origin of these populations, as compared to Indo-Iranian-speaking populations. The results from admixture and cluster analyses further show that Turkic-speaking populations are closely related to Eastern Asian populations. This is consistent with a recent DNA study of prehistoric skeletal remains from the Bronze and Iron Ages in Kazakhstan,{Lalueza-Fox, 2004 #18} which dated the presence of individuals with eastern genetic characteristics around the 7th century B.C. The authors argued that the distribution of mtDNA haplogroups in their temporal sample could correspond to westward migrations of Asian nomads, presumably from Siberia and Mongolia.{Lalueza-Fox, 2004 #18}

The westernized view of the eastern invasions usually underlies the extreme violence of the hordes coming from the East. For example one of the most famous Hun who has later

invaded Europe, Attila the Hun (A.D. 406-453) is described as making it his boast that the grass never grew again where his horse's hoofs had trod. Later, during the expansion of the Mongolian empire led by Genghis Khan, our western history school books tell us that people that did not submit to the Khan were killed.

However, our results suggest an alternative scenario: a long term settled group of what are nowadays Tajik populations received an arrival of eastern populations (the Turco-Mongol) that did not lead to the replacement of the local settled populations but rather to admixture. Admixture analyses show that, although a large proportion of the gene pool is shared with East Asians, all Turkic-speaking populations in central Asia are also clearly admixed with other populations; conversely the gene pool of the Tajik populations show low but significant level of admixture of eastern origin. Even if western historical records suggest that the local populations were destroyed by these eastern invasions, our genetic data points to a softer model of invasion.

We found that the Uzbek populations were scattered across Turkic- and Indo-Iranian speaking populations (Figure 2). Some Uzbek populations (LUZa, LUZn, UZA) were closer to Tajik populations, while other populations (UZB, UZT) clearly clustered with Turkic-speaking populations. This is consistent with the fact that “Uzbeks” include the 17th century Uzbeks, which were nomadic herders before they sedentarized around the 16th Century,{Chaix, 2007 #4} and former Chagatai Turk groups who were already settled in Uzbekistan.{Soucek, 2000 #68} Uzbeks therefore result from the union of different tribes, some of recent origin clustering with Turkic-speaking populations, and some tracing back to Chagatai Turks who were strongly admixed with Iranian dwellers of Central Asia.

The genetic legacy of the Mongols

We found that the Central Asian populations share little or no ancestry with the Mongols (Figure 2b), which suggests that the Mongol invasions led by Genghis Khan left no major genetic imprint in Central Asia. Our results therefore challenge the conclusions of a Y-chromosome survey that reported evidence of a genetic signature of Mongol invasions in several populations of the region.{Zerjal, 2003 #19} However, more recent analyses of Y-chromosome haplogroups across nine populations, which included four Mongolian ethnic groups, also concluded to the absence of strong genetic affinities between Mongolian and

Central Asian populations. The only exception to this pattern concerned the Khoton Mongolians{Katoh, 2005 #37} who presumably descended from a nomadic tribe of Turkish origin, which migrated into Mongolia from Central Asia in the 17th century.{Katoh, 2005 #36}

Our study also contradicts the claim that these invasions resulted in founder effects.{Zerjal, 2002 #20;Zerjal, 2002 #20} The high level of autosomal diversity observed in all Turkic-speaking populations (Table 2) contrasts indeed with the low level of Y-chromosome diversity found in some populations of the region.{Chaix, 2007 #4;Zerjal, 2002 #20;Chaix, 2007 #4;Zerjal, 2002 #20} Because population bottlenecks would affect all genetic systems similarly, this suggests that the social organization of pastoral populations, which are based on patrilineal descent groups{Chaix, 2007 #4;Chaix, 2007 #4}, is more likely to account for the observed reduction of Y-chromosome diversity.

Evidence for linguistic replacements

We found two cases of linguistic replacements in Central Asia. Two Turkic-speaking populations, TUR (Turkmen) and KRM (Kyrgyz), were found to cluster together with Indo-Iranian-speaking populations (Figure 2b). The genetic similarity between Turkmen and Tajiks (see also Table 5) is consistent with the hypothesis that Turkmen, together with Tajiks, may be the present-day descendants of populations established over long periods of time. The indigenous cultural history of the Turkmen in Turkmenistan can indeed be dated back to 10,000 years B.C. and similarities between the cultures and technologies found in the archaeological record suggest that this region has been continually occupied since 6,000 B.C. A recent linguistic replacement in the TUR population would then explain the observed pattern of a Turkic-speaking population clustering with Indo-Iranian speakers. The Kyrgyz population KRM provides another example of linguistic shift: the Turkic-speaking KRM indeed clusters with the Indo-Iranian speakers (Figure 2 and Table 5). Moreover, despite their geographic proximity, KRM is clearly differentiated from the other Kyrgyz populations. For both TUR and KRM, invasions from eastern Eurasia may have been followed by the replacement of a native Indo-Iranian language with a Turkic language, with relatively little genetic admixture. This may have happened through an elite dominance-language shift, as it

was invoked for other nomad migrations.{Quintana-Murci, 2001 #45;Quintana-Murci, 2001 #45}

A Central Asian origin of the Hazaras?

Our study sheds light on the debated origin of the Hazaras from Pakistan. The Hazaras, who claim to be the direct male-line descendants of Genghis Khan, are assumed to have Mongolian ancestry.{Qamar, 2002 #44;Qamar, 2002 #44} A Y-chromosome genetic survey provided indirect support to these claims.{Zerjal, 2003 #19;Zerjal, 2003 #19} Furthermore, the high frequency of Eastern Eurasian mtDNA lineages found in the Hazaras, but not in the neighboring populations, also suggest that women of East Asian ancestry accompanied the male descendants of Genghis Khan, or other Mongols.{Quintana-Murci, 2004 #40;Quintana-Murci, 2004 #40} The strong pairwise differentiation of Y-linked markers between the Hazaras and other populations from Pakistan has been interpreted as the consequence of drift occurring in a small and isolated population.{Qamar, 2002 #44;Qamar, 2002 #44} Our results show that the history of the Hazaras might be more parsimoniously explained by a Central Asian origin of this population, which clearly clusters with the Turkic-speakers of Central Asia that share a high proportion of polymorphisms with East Asians coupled with the consequence of strong drift.

Conclusion

This study shows that Central Asia is genetically differentiated into two main groups of populations: on the one hand, Indo-Iranian-speaking populations, which include Tajiks and two Uzbek populations, are genetically closer to populations from Western Eurasia; on the other hand, Turkic-speaking populations, which include Karakalpaks, Kazakhs, Kyrgyz, and other Uzbek populations, are closer to Eastern Asian populations (with the exception of the Turkmen and one Kyrgyz population). This mixed ancestry of Central Asian populations combined with the level of genetic differentiation among the two groups, let us propose the following scenario for the population history in CA: a long term settled group of what is nowadays Tajik populations and a more recent arrival of eastern populations comprising the Turkic group. Further analyses based, e.g., on measures of linkage disequilibrium (see Xu *et al.*'s{Xu, 2009 #70} study on Uyghurs) are required to test this scenario. Our results also

show that the expansion of eastern nomadic groups, contrary to what is generally thought, did not result in the complete replacements of the local populations but rather to a process of “soft invasion” that includes admixture.

Acknowledgements

We are indebted to everyone who volunteered to participate to this study. We also thank R. Leblois and P. Verdu for insightful discussions on previous versions of this paper, H. Cann for providing CEPH samples, the *Service de Systématique Moléculaire* (SSM) at the *Museum National d'Histoire Naturelle* (MNHN) for making facilities available, and J.A. Godoy for technical assistance. We are very grateful to CESGA (Supercomputational Centre of Galicia) and to the Computational Biology Service Unit from the Museum National d'Histoire Naturelle (MNHN – CNRS UMS 2700) where the computational analyses were performed, This work was supported by the *Centre National de la Recherche Scientifique* (CNRS) ATIP program (to E.H.), by the CNRS interdisciplinary program "*Origines de l'Homme du Langage et des Langues"* (OHLL) and by the European Science Foundation (ESF) EUROCORES program "The Origin of Man, Language and Languages" (OMLL).

References

1 Cavalli-Sforza LL, Menozzi P, Piazza A: The History and Geography of Human Genes. Princeton, University Press, 1994.

2 Nei M, Roychoudhury AK: Evolutionary relationships of human populations on a global scale. *Molecular Biology and Evolution* 1993; **10**: 927-943.

3 Comas D, Calafell F, Mateu E *et al*: Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. *American Journal of Human Genetics* 1998; **63**: 1824-1838.

4 Cordaux R, Deepa E, Vishwanathan H, Stoneking M: Genetic evidence for the demic diffusion of agriculture to India. *Science* 2004; **304**: 1125-1125.

5 Karafet T, Xu LP, Du RF *et al*: Paternal population history of east Asia: Sources, patterns, and microevolutionary processes. *American Journal of Human Genetics* 2001; **69**: 615-628.

6 Wells RS, Yuldasheva N, Ruzibakiev R *et al*: The Eurasian Heartland: A continental perspective on Y-chromosome diversity. *Proceedings of the National Academy of Sciences of the United States of America* 2001; **98**: 10244-10249.

7 Гумилев Л. Н. / Древние тюрки /АН СССР. Ин-т народов Азии. - М.: Наука, 1967. - 504 с.. с карт. - 4800.

8 Chaix R, Austerlitz F, Khegay T *et al*: The genetic or mythical ancestry of descent groups: Lessons from the Y chromosome. *American Journal of Human Genetics* 2004; **75**: 1113-1116.

9 Chaix R, Quintana-Murci L, Hegay T *et al*: From social to genetic structures in central Asia. *Current Biology* 2007; **17**: 43-48.

10 Comas D, Plaza S, Wells RS *et al*: Admixture, migrations, and dispersals in Central Asia: evidence from maternal DNA lineages. *European Journal of Human Genetics* 2004; **12**: 495-504.

11 Lalueza-Fox C, Sampietro ML, Gilbert MTP *et al*: Unravelling migrations in the steppe: mitochondrial DNA sequences from ancient Central Asians. *Proceedings of the Royal Society of London Series B-Biological Sciences* 2004; **271**: 941-947.

12 Perez-Lezaun A, Calafell F, Comas D *et al*: Sex-specific migration patterns in central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *American Journal of Human Genetics* 1999; **65**: 208-219.

13 Hammer MF, Karafet TM, Redd AJ *et al*: Hierarchical patterns of global human Y-chromosome diversity. *Molecular Biology and Evolution* 2001; **18**: 1189-1203.

14 Zerjal T, Wells RS, Yuldasheva N, Ruzibakiev R, Tyler-Smith C: A genetic landscape reshaped by recent events: Y-chromosomal insights into Central Asia. *American Journal of Human Genetics* 2002; **71**: 466-482.

15 Ballard JWO, Whitlock MC: The incomplete natural history of mitochondria. *Molecular Ecology* 2004; **13**: 729-744.

16 Balloux F: The worm in the fruit of the mitochondrial DNA tree. *Heredity* 2009.

17 Bazin E, Glemin S, Galtier N: Population size does not influence mitochondrial genetic diversity in animals. *Science* 2006; **312**: 570-572.

18 Pakendorf B, Stoneking M: Mitochondrial DNA and human evolution. *Annual Review of Genomics and Human Genetics* 2005; **6**: 165-183.

19 Maniatis T, Fritsch EF, Sambrook J: Molecular Cloning. A Laboratory Manual. New York, Cold Spring Harbor, 1982.

20 Segurel L, Martinez-Cruz B, Quintana-Murci L *et al*: Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. *PLoS Genet* 2008; **4**: e1000200.

21 Rosenberg NA, Pritchard JK, Weber JL *et al*: Genetic structure of human populations. *Science* 2002; **298**: 2381-2385.

22 Cann HM, de Toma C, Cazes L *et al*: A human genome diversity cell line panel. *Science* 2002; **296**: 261-262.

23 Zhivotovsky LA, Rosenberg NA, Feldman MW: Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers. *American Journal of Human Genetics* 2003; **72**: 1171-1186.

24 Rosenberg NA: Standardized subsets of the HGDP-CEPH human genome diversity cell line panel, accounting for atypical and duplicated samples and pairs of close relatives. *Annals of Human Genetics* 2006; **70**: 841-847.

25 ElMousadik A, Petit RJ: High level of genetic differentiation for allelic richness among populations of the argan tree Argania spinosa (L) Skeels endemic to Morocco. *Theoretical and Applied Genetics* 1996; **92**: 832-839.

26 Goudet J: FSTAT (Version 1.2): A computer program to calculate F-statistics. *Journal of Heredity* 1995; **86**: 485-486.

27 Nei M: Estimation of Average Heterozygosity and Genetic Distance from a Small Number of Individuals. *Genetics* 1978; **89**: 583-590.

28 Belkhir K, Borsa P, Chikhi L, Raufaste N, Bonhomme F: GENETIX 4.05, logiciel sous Windows TM pour la génétique des populations. Laboratoire Génome, Populations, Interactions, CRNS UMS 5171, Université de Montpellier II, Montpellier (France) 1996-2004.

29 Inc. SI: JMP Statistics and Graphics Guide, Version 5.1. Cary, NC: SAS Institute Inc. 2003.

30 Weir BS, Cockerham CC: Estimating F-Statistics for the Analysis of Population-Structure. *Evolution* 1984; **38**: 1358-1370.

31 Peakall R, Smouse PE: GENALEX 6: genetic analysis in Excel. Population genetic software for teaching and research. *Molecular Ecology Notes* 2006; **6**: 288-295.

32 Excoffier L, Smouse PE, Quattro JM: ANALYSIS OF MOLECULAR VARIANCE INFERRED FROM METRIC DISTANCES AMONG DNA HAPLOTYPES - APPLICATION TO HUMAN MITOCHONDRIAL-DNA RESTRICTION DATA. *Genetics* 1992; **131**: 479-491.

33 Excoffier L, Laval LG, Schneider S: Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 2005; **1**: 47-50.

34 Rousset F: Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. *Genetics* 1997; **145**: 1219-1228.

35 Mantel N: The detection of disease clustering and a generalized regression approach. *Cancer Research* 1967; **27**: 209-220.

36 Rousset F: GENEPOP ' 007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources* 2008; **8**: 103-106.

37 Pritchard JK, Stephens M, Donnelly P: Inference of population structure using multilocus genotype data. *Genetics* 2000; **155**: 945-959.

38 Rosenberg NA, Mahajan S, Gonzalez-Quevedo C *et al*: Low levels of genetic divergence across geographically and linguistically diverse populations from India. *Plos Genetics* 2006; **2**: 2052-2061.

39 Falush D, Stephens M, Pritchard JK: Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. *Genetics* 2003; **164**: 1567-1587.

40 Wang JL: Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics* 2003; **164**: 747-765.

41 Belle EMS, Landry PA, Barbujani G: Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. *Proceedings of the Royal Society B-Biological Sciences* 2006; **273**: 1595-1602.

42 Qamar R, Ayub Q, Mohyuddin A *et al*: Y-chromosomal DNA variation in Pakistan. *American Journal of Human Genetics* 2002; **70**: 1107-1124.

43 Zerjal T, Xue YL, Bertorelle G *et al*: The genetic legacy of the mongols. *American Journal of Human Genetics* 2003; **72**: 717-721.

44 Li JZ, Absher DM, Tang H *et al*: Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008; **319**: 1100-1104.

45 Bosch E, Calafell F, Gonzalez-Neira A *et al*: Paternal and maternal lineages in the Balkans show a homogeneous landscape over linguistic barriers, except for the isolated Aromuns. *Annals of Human Genetics* 2006; **70**: 459-487.

46 Manica A, Prugnolle F, Balloux F: Geography is a better determinant of human genetic differentiation than ethnicity. *Human Genetics* 2005; **118**: 366-371.

47 Brunet F: La Néolithisation en Asie Centrale: un état de la question. *Paléorient* 1999; **24**: 27-48.

48 Soucek S: A History of Inner Asia. Cambridge, Cambridge University Press, 2000.

49 Katoh T, Munkhbat B, Tounai K *et al*: Genetic features of Mongolian ethnic groups revealed by Y-chromosomal analysis. *Gene* 2005; **346**: 63-70.

50 Katoh T, Mano S, Munkhbat B *et al*: Genetic features of Khoton Mongolians revealed by SNP analysis of the X chromosome. *Gene* 2005; **357**: 95-102.

51 Quintana-Murci L, Krausz C, Zerjal T *et al*: Y-chromosome lineages trace diffusion of people and languages in southwestern Asia. *American Journal of Human Genetics* 2001; **68**: 537-542.

52 Quintana-Murci L, Chaix R, Wells RS *et al*: Where west meets east: The complex mtDNA landscape of the southwest and Central Asian corridor. *American Journal of Human Genetics* 2004; **74**: 827-845.

53 Xu S, Jin W, Jin L: Haplotype-sharing analysis showing Uyghurs are unlikely genetic donors. *Mol Biol Evol* 2009; **26**: 2197-2206.

Table 1 Description of the 26 Central Asian studied populations

Sampled populations (area)	Acronym	Location	Language family	Long.	Lat.	*n*
Tajiks (Samarkand)	TJA	Uzbekistan / Tajikistan border	Western Iranian	**39.54**	**66.89**	31
Tajiks (Samarkand)	TJU	Uzbekistan / Tajikistan border	Western Iranian	**39.50**	**67.27**	29
Tajiks (Ferghana)	TJR	Tajikistan / Kyrgyzstan border	Western Iranian	**40.36**	**71.28**	29
Tajiks (Ferghana)	TJK	Tajikistan / Kyrgyzstan border	Western Iranian	**40.25**	**71.87**	26
Tajiks (Gharm)	TJE	Northern Tajikistan	Western Iranian	**39.12**	**70.67**	25
Tajiks (Gharm)	TJN	Northern Tajikistan	Western Iranian	**38.09**	**68.81**	24
Tajiks (Gharm)	TJT	Northern Tajikistan	Western Iranian	**39.11**	**70.86**	25
Tajiks (Penjikent)	TDS	Uzbekistan / Tajikistan border	Western Iranian	**39.28**	**67.81**	25
Tajiks (Penjikent)	TDU	Uzbekistan / Tajikistan border	Western Iranian	**39.44**	**68.26**	25
Tajiks (Yagnobs from Dushanbe)	TJY	Western Tajikistan	Eastern Iranian	**38.57**	**68.78**	25
Uzbeks (Ferghana)	UZA	Uzbekistan / Kyrgyzstan border	Turkic	**40.77**	**72.31**	25
Uzbeks (Penjikent)	UZT	Northern Tajikistan	Turkic	**39.49**	**67.54**	25
Uzbeks (Bukhara)	LUZn	Central Uzbekistan	Bilinguism	**39.70**	**64.38**	20
Uzbeks (Bukhara)	LUZa	Central Uzbekistan	Bilinguism	**39.73**	**64.27**	20
Uzbeks (Karakalpakia)	UZB	Western Uzbekistan	Turkic	**43.04**	**58.84**	35
Karakalpaks (Qongrat from Karakalpakia)	KKK	Western Uzbekistan	Turkic	**43.77**	**59.02**	45
Karakalpaks (On Tört Uruw from Karakalpakia)	OTU	Western Uzbekistan	Turkic	**42.94**	**59.78**	45
Kazaks (Karakalpakia)	KAZ	Western Uzbekistan	Turkic	**43.04**	**58.84**	49
Kazaks (Bukhara)	LKZ	Central Uzbekistan	Turkic	**40.08**	**63.56**	25

Kyrgyz (Andijan)	KRA	Uzbekistan / Kyrgyzstan border	Turkic	**40.77**	**72.31**	45
Kyrgyz (Narin)	KRG	Eastern Kyrgyzstan	Turkic	**41.60**	**75.80**	18
Kyrgyz (Narin)	KRM	Eastern Kyrgyzstan	Turkic	**41.45**	**76.22**	21
Kyrgyz (Narin)	KRL	Eastern Kyrgyzstan	Turkic	**41.36**	**75.50**	22
Kyrgyz (Narin)	KRB	Eastern Kyrgyzstan	Turkic	**41.25**	**76.00**	24
Kyrgyz (Issyk Kul)	KRT	Eastern Kyrgyzstan	Turkic	**42.16**	**77.57**	37
Turkmen (Karakalpakia)	TUR	Western Uzbekistan	Turkic	**41.55**	**60.63**	47

Long., longitude; Lat., latitude. *n*, sample size.

Table 2 Genetic diversity in the studied populations and in Eurasia

World Area	Population	*AR*	H_e
CENTRAL ASIA	KAZ	7.8	0.784
	KKK	7.7	0.782
	KRA	7.4	0.762
	KRB	7.1	0.757
	KRG	7.6	0.779
	KRL	7.6	0.778
	KRM	7.9	0.814
	KRT	7.4	0.761
	LKZ	7.5	0.778
	LUZa	7.9	0.817
	LUZn	8.2	0.821
	OTU	7.6	0.784
	TDS	7.2	0.784
	TDU	7.4	0.805
	TJA	7.5	0.806
	TJE	7.8	0.814
	TJK	8.0	0.803
	TJN	7.7	0.811
	TJR	7.9	0.812
	TJT	7.8	0.812
	TJU	7.7	0.811
	TJY	7.1	0.799
	TUR	7.7	0.812
	UZA	8.1	0.817
	UZB	7.6	0.774
	UZT	7.5	0.795
CENTRAL ASIA		3.0	0.795
CENTRAL SOUTH ASIA		3.0	0.819
EAST ASIA		2.7	0.706
EUROPE		3.1	0.775
MIDDLE EAST		3.2	0.826

AR, allelic richness; H_e, expected heterozygosity.

Table 3 Differentiation among the 26 Central Asian populations studied

	KAZ	KKK	KRA	KRB	KRG	KRL	KRM	KRT	LKZ	LUZa	LUZn	OTU	TDS	TDU	TJA	TJE	TJK	TJN	TJR	TJT	TJU	TJY	TUR	UZA	UZB	UZT
KAZ		***	***	*	NS	NS	***	**	NS	***	***	*	***	***	***	***	***	***	***	***	***	***	***	*	***
KKK	0,0052		***	**	***	NS	***	***	NS	***	***	NS	***	***	***	***	***	***	***	***	***	***	***	*	*	***
KRA	0,0045	0,0068		*	***	**	***	***	**	***	***	***	***	***	***	***	NS	***	***	***	***	***	***	***	***	***
KRB	0,0025	0,0038	0,0020		**	NS	***	**	NS	***	***	**	***	***	***	***	***	***	***	***	***	***	***	***	*	***
KRG	0,0035	0,0100	0,0107	0,0078		NS	***	***	*	***	***	***	***	***	***	***	***	***	***	***	***	***	***	**	***	***
KRL	0,0000	0,0037	0,0035	0,0011	0,0000		***	NS	NS	***	***	NS	***	***	***	***	***	***	***	***	***	***	***	NS	NS	NS
KRM	0,0198	0,0153	0,0203	0,0236	0,0181	0,0181		***	***	***	**	***	***	***	***	***	NS	***	***	***	***	***	***	*	***	***
KRT	0,0071	0,0060	0,0052	0,0034	0,0080	0,0000	0,0210		NS	***	***	*	***	***	***	***	***	***	***	***	***	***	***	***	*	***
LKZ	0,0010	0,0004	0,0036	0,0021	0,0037	0,0000	0,0119	0,0000		***	***	NS	***	***	***	***	***	***	***	***	***	***	***	NS	NS	**
LUZa	0,0172	0,0118	0,0218	0,0198	0,0221	0,0125	0,0119	0,0207	0,0069		NS	***	***	**	*	**	**	NS	***	*	***	***	*	NS	***	***
LUZn	0,0240	0,0125	0,0277	0,0288	0,0278	0,0181	0,0103	0,0215	0,0124	0,0000		***	***	**	***	**	***	*	***	*	***	***	***	NS	***	**
OTU	0,0032	0,0015	0,0059	0,0067	0,0071	0,0010	0,0084	0,0034	0,0000	0,0105	0,0146		***	***	***	***	***	***	***	***	***	***	***	NS	NS	***
TDS	0,0254	0,0181	0,0291	0,0252	0,0285	0,0222	0,0174	0,0298	0,0191	0,0145	0,0132	0,0201		***	***	***	***	***	***	***	***	***	***	NS	***	***
TDU	0,0225	0,0207	0,0305	0,0294	0,0253	0,0181	0,0175	0,0292	0,0163	0,0047	0,0063	0,0168	0,0212		***	***	***	***	***	***	***	***	***	**	**	***
TJA	0,0187	0,0154	0,0256	0,0244	0,0245	0,0172	0,0103	0,0275	0,0111	0,0019	0,0111	0,0121	0,0185	0,0113		***	***	*	***	**	***	***	***	*	***	***
TJE	0,0221	0,0197	0,0311	0,0295	0,0265	0,0222	0,0097	0,0285	0,0163	0,0075	0,0098	0,0159	0,0177	0,0159	0,0077		***	NS	**	*	***	***	***	*	**	***
TJK	0,0065	0,0076	0,0030	0,0123	0,0117	0,0046	0,0053	0,0096	0,0040	0,0081	0,0138	0,0045	0,0184	0,0165	0,0096	0,0134		**	**	***	***	***	***	***	***	***
TJN	0,0223	0,0151	0,0263	0,0253	0,0272	0,0211	0,0094	0,0253	0,0131	0,0040	0,0071	0,0133	0,0120	0,0121	0,0019	0,0024	0,0108		***	NS	***	***	***	NS	**	***
TJR	0,0174	0,0135	0,0206	0,0222	0,0177	0,0158	0,0087	0,0197	0,0101	0,0073	0,0063	0,0105	0,0160	0,0119	0,0126	0,0087	0,0075	0,0074		***	***	***	***	NS	***	***
TJT	0,0277	0,0197	0,0323	0,0295	0,0301	0,0280	0,0124	0,0320	0,0195	0,0048	0,0070	0,0189	0,0115	0,0170	0,0066	0,0112	0,0174	0,0023	0,0107		***	***	***	*	**	***
TJU	0,0167	0,0130	0,0233	0,0246	0,0202	0,0161	0,0104	0,0230	0,0111	0,0078	0,0107	0,0103	0,0191	0,0106	0,0079	0,0080	0,0073	0,0073	0,0102	0,0134		***	***	*	***	***
TJY	0,0425	0,0329	0,0483	0,0531	0,0479	0,0423	0,0216	0,0514	0,0329	0,0184	0,0152	0,0345	0,0292	0,0216	0,0198	0,0218	0,0293	0,0168	0,0231	0,0201	0,0223		***	***	**	***
TUR	0,0158	0,0119	0,0204	0,0214	0,0209	0,0125	0,0121	0,0192	0,0102	0,0008	0,0069	0,0119	0,0170	0,0075	0,0090	0,0111	0,0093	0,0071	0,0107	0,0104	0,0063	0,0235		*	***	***
UZA	0,0118	0,0074	0,0196	0,0166	0,0093	0,0042	0,0068	0,0148	0,0014	0,0000	0,0072	0,0038	0,0090	0,0068	0,0050	0,0074	0,0090	0,0035	0,0047	0,0100	0,0052	0,0217	0,0037		NS	**
UZB	0,0025	0,0033	0,0048	0,0017	0,0080	0,0018	0,0146	0,0036	0,0000	0,0141	0,0176	0,0012	0,0211	0,0170	0,0171	0,0216	0,0047	0,0183	0,0130	0,0237	0,0139	0,0347	0,0110	0,0076		**
UZT	0,0073	0,0088	0,0107	0,0087	0,0093	0,0045	0,0146	0,0102	0,0042	0,0094	0,0116	0,0050	0,0152	0,0163	0,0176	0,0195	0,0062	0,0156	0,0112	0,0187	0,0135	0,0332	0,0143	0,0079	0,0051	

Below the diagonal, pairwise F_{ST} estimates; above the diagonal, exact test of differentiation.
*$p < 0.01$, **$p < 0.001$, ***$p < 0.0001$, NS: non-significant

Table 4 AMOVA of the 26 Central Asian studied populations

Grouping	Source of variation	Percentage of variation	F_{ST}	F_{SC}	F_{CT}
Linguistic affiliation	Among groups	0.82			0.008***
	Among populations within groups	0.89		0.009***	
	Within populations	98.3	0.017***		
Ethnicity	Among groups	0.51			0.005*
	Among populations within groups	1.0		0.01	
	Within populations	98.51	0.015		

*$p < 0.01$, **$p < 0.001$, ***$p < 0.0001$.

Table 5 Maximum-likelihood estimates of admixture proportions in the 26 Central Asian populations

		Putative parental group			
				Central South	
Population	Ethnic group	Europe	Middle East	Asia	East Asia
KAZ	**Kazakh**	0.100	0.151	0.251	0.499
LKZ	**Kazakh**	0.250	0.004	0.250	0.496
KKK	**Karakalpak**	0.126	0.126	0.251	0.497
OUT	**Karakalpak**	0.250	0.127	0.128	0.495
KRA	**Kyrgyz**	0.197	0.053	0.250	0.499
KRB	**Kyrgyz**	0.125	0.125	0.125	0.625
KRG	**Kyrgyz**	0.006	0.249	0.253	0.492
KRL	**Kyrgyz**	0.074	0.250	0.176	0.500
KRM	**Kyrgyz**	0.235	0.200	0.263	0.302
KRT	**Kyrgyz**	0.250	0.056	0.194	0.500
TUR	**Turkmen**	0.255	0.249	0.251	0.245
UZA	**Uzbek**	0.253	0.183	0.257	0.307
UZB	**Uzbek**	0.124	0.127	0.251	0.498
UZT	**Uzbek**	0.126	0.126	0.253	0.496
LUZa	**Uzbek**	0.298	0.286	0.250	0.166
LUZn	**Uzbek**	0.230	0.274	0.308	0.188
TDS	**Tajik**	0.375	0.126	0.250	0.249
TDU	**Tajik**	0.306	0.195	0.251	0.248
TJA	**Tajik**	0.361	0.249	0.255	0.134
TJE	**Tajik**	0.252	0.323	0.257	0.168
TJK	**Tajik**	0.293	0.132	0.133	0.442
TJN	**Tajik**	0.493	0.138	0.250	0.119
TJR	**Tajik**	0.250	0.308	0.163	0.279
TJT	**Tajik**	0.336	0.229	0.294	0.141
TJU	**Tajik**	0.294	0.371	0.153	0.182
TJY	**Tajik**	0.298	0.284	0.293	0.125

Shaded cells correspond to Turkic-speaking populations, and non-shaded cells to Indo-European-speakers.

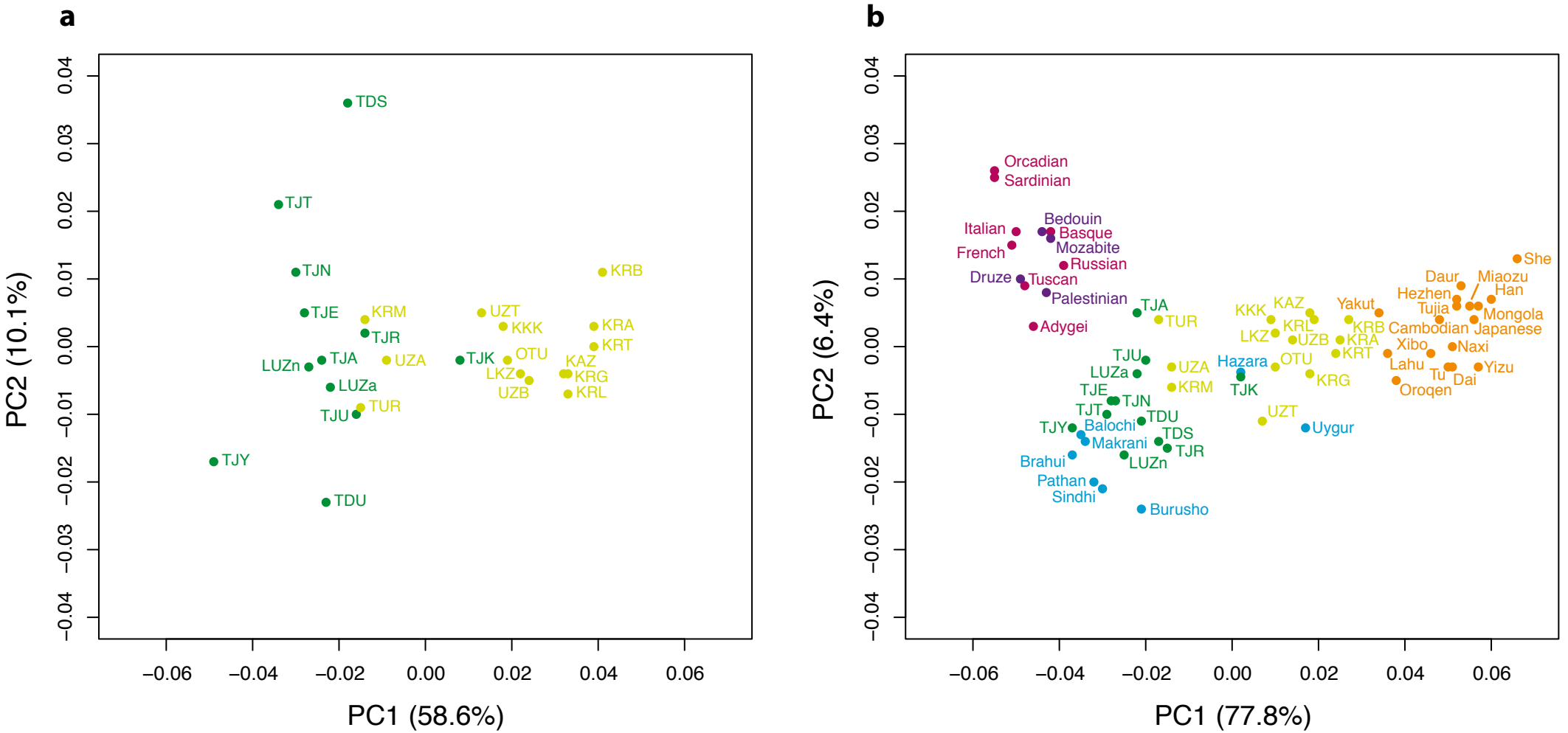
Figure legends

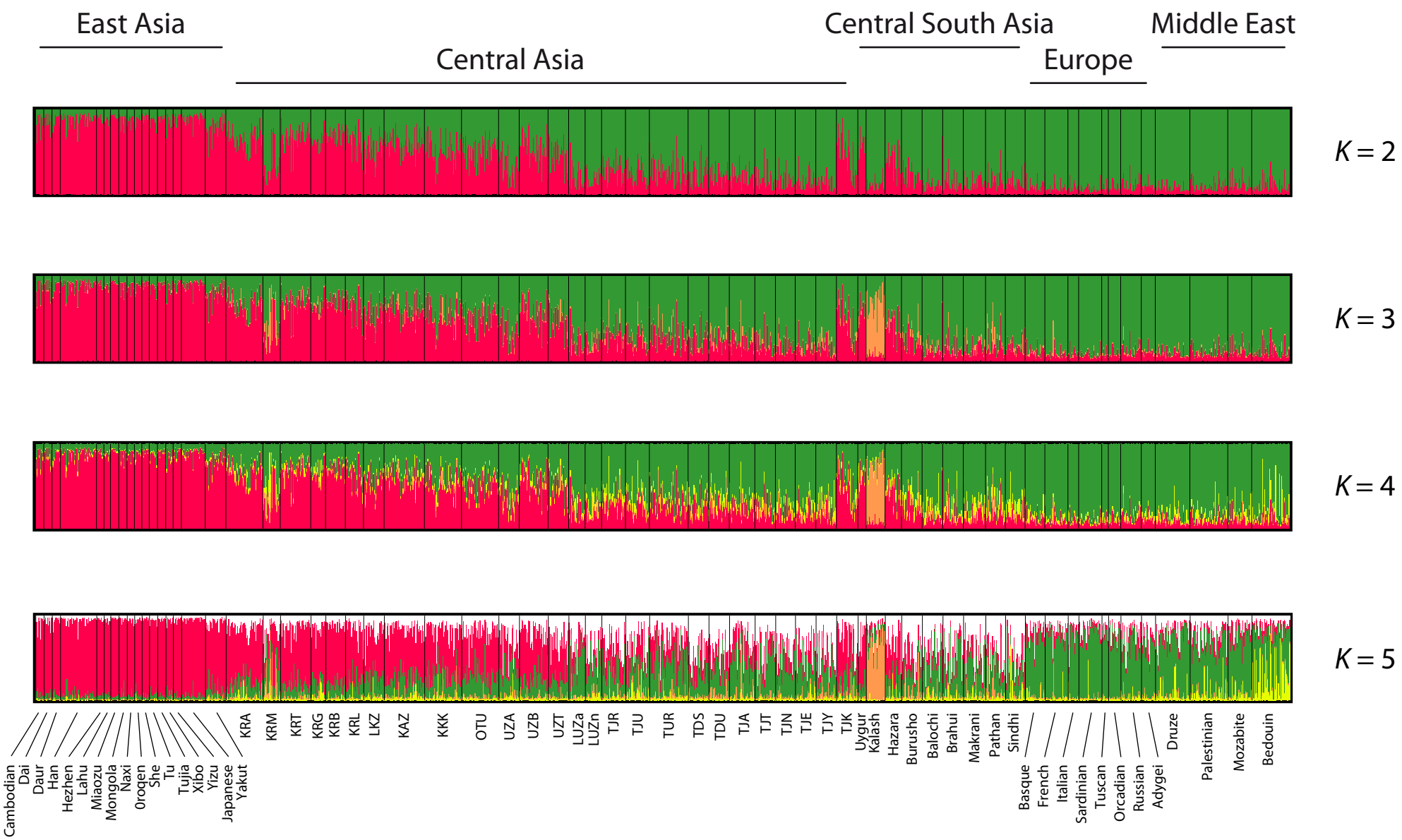
Figure 1 Geographic location of the 26 Central Asian populations sampled. Linguistic affiliation, as well as admixture proportions from putative parental origins (Central South Asia, East Asia, Europe and Middle East) are also indicated. See Table 1 for acronyms.

Figure 2 Principal Component Analysis (PCA) based on pairwise F_{ST} estimates between populations in Central Asia (**a**). Colors indicate language affiliation; dark green: Indo-Iranian speakers; light green, Turkic speakers. PCA based on pairwise F_{ST} estimates between Eurasian populations (excluding the Kalash population, see text) (**b**). Colors represent major geographic regions; magenta: Europe; purple: Middle East; blue: Central-South Asia; dark green: Central Asian, Indo-Iranian speakers; light green: Central Asia, Turkic speakers; orange: East Asia.

Figure 3 Population structure inferred from microsatellite data using the software package STRUCTURE. Each individual is represented by a vertical line, divided into up to *K* colored segments, each of which represents the individual's estimated membership fraction to that cluster. The data consist in 767 individuals from 26 Central Asian populations genotyped at 27 microsatellite loci, plus 755 individuals from 38 Eurasian populations from the HGDP-CEPH Human Genome Diversity Cell Line Panel. *K* represents the number of putative clusters. See Table 1 for acronyms.


Aral Sea
Lake Balkhash
Kazakhstan
Lake Issyk Kul
Kyrgyzstan
Uzbekistan
Turkmenistan
Tajikistan
China
Afghanistan
Pakistan
KKK
OTU
KAZ
UZB
TUR
LKZ
LUZn
LUZa
TJA
TJU
UZT
TDU
TDS
TJY
TJE
TJR
KRA
UZA
TJK
TJN
TJT
KRG
KRB
KRT
KRL
KRM
Language family
Indo-Iranian
Turkic
Admixture proportions
East-Asia
Europe
Central-South Asia
Middle East
0
250
500 km
N


a
PC2 (10.1%)
PC1 (58.6%)
TDS
TJT
TJN
TJE
KRM
TJR
LUZn
TJA
UZA
LUZa
TJU
TUR
TJY
TDU
UZT
KKK
TJK
OTU
LKZ
UZB
KRB
KRA
KRT
KAZ
KRG
KRL
b
PC2 (6.4%)
PC1 (77.8%)
Orcadian
Sardinian
Italian
Bedouin
Basque
Mozabite
French
Russian
Druze
Tuscan
Palestinian
Adygei
TJA
TUR
KKK
KAZ
Yakut
KRL
LKZ
UZB
KRA
KRT
KRB
She
Daur
Miaozu
Hezhen
Han
Tujia
Mongola
Cambodian
Japanese
Xibo
Naxi
Lahu
Tu
Dai
Yizu
Oroqen
TJU
LUZa
UZA
KRM
TJE
TJN
TJT
TDU
TJY
Balochi
TDS
Makrani
TJR
Brahui
LUZn
Pathan
Sindhi
Burusho
Hazara
OTU
TJK
KRG
UZT
Uygur


East Asia
Central Asia
Central South Asia
Europe
Middle East
K = 2
K = 3
K = 4
K = 5
Cambodian
Dai
Daur
Han
Hezhen
Lahu
Miaozu
Mongola
Naxi
Orogen
She
Tu
Tujia
Xibo
Yizu
Japanese
Yakut
KRA
KRM
KRT
KRG
KRB
KRL
LKZ
KAZ
KKK
OTU
UZA
UZB
UZT
LUZa
LUZn
TJR
TJU
TUR
TDS
TDU
TJA
TJT
TJN
TJE
TJY
TJK
Uygur
Kalash
Hazara
Burusho
Balochi
Brahui
Makrani
Pathan
Sindhi
Basque
French
Italian
Sardinian
Tuscan
Orcadian
Russian
Adygei
Druze
Palestinian
Mozabite
Bedouin

- **Annexe 4** : **Ségurel L.**, Martinez-Cruz B., Quintana-Murci L., Balaresque P., Georges M., Hegay T., Aldashev A., Nazyrova F., Jobling M.A., Heyer E. & Vitalis R. Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. *PLoS Genet* 2008 Sep 26; 4(9):e 1000200.

OPEN ACCESS Freely available online

PLoS GENETICS

Sex-Specific Genetic Structure and Social Organization in Central Asia: Insights from a Multi-Locus Study

Laure Ségurel[1]*, Begoña Martínez-Cruz[1¤], Lluis Quintana-Murci[2], Patricia Balaresque[3], Myriam Georges[1], Tatiana Hegay[4], Almaz Aldashev[5], Firuza Nasyrova[6], Mark A. Jobling[3], Evelyne Heyer[1], Renaud Vitalis[1]

1 Muséum National d'Histoire Naturelle – Centre National de la Recherche Scientifique UMR 5145 – Université Paris 7, Éco-Anthropologie et Ethnobiologie, Musée de l'Homme, Paris, France, **2** Human Evolutionary Genetics Unit, CNRS URA3012, Institut Pasteur, Paris, France, **3** Department of Genetics, University of Leicester, Leicester, United Kingdom, **4** Uzbek Academy of Sciences, Institute of Immunology, Tashkent, Uzbekistan, **5** Institute of Molecular Biology and Medicine, National Center of Cardiology and Internal Medicine, Bishkek, Kyrgyzstan, **6** Tajik Academy of Sciences, Institute of Plant Physiology and Genetics, Dushanbe, Tajikistan

Abstract

In the last two decades, mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY) have been extensively used in order to measure the maternally and paternally inherited genetic structure of human populations, and to infer sex-specific demography and history. Most studies converge towards the notion that among populations, women are genetically less structured than men. This has been mainly explained by a higher migration rate of women, due to patrilocality, a tendency for men to stay in their birthplace while women move to their husband's house. Yet, since population differentiation depends upon the product of the effective number of individuals within each deme and the migration rate among demes, differences in male and female effective numbers and sex-biased dispersal have confounding effects on the comparison of genetic structure as measured by uniparentally inherited markers. In this study, we develop a new multi-locus approach to analyze jointly autosomal and X-linked markers in order to aid the understanding of sex-specific contributions to population differentiation. We show that in patrilineal herder groups of Central Asia, in contrast to bilineal agriculturalists, the effective number of women is higher than that of men. We interpret this result, which could not be obtained by the analysis of mtDNA and NRY alone, as the consequence of the social organization of patrilineal populations, in which genetically related men (but not women) tend to cluster together. This study suggests that differences in sex-specific migration rates may not be the only cause of contrasting male and female differentiation in humans, and that differences in effective numbers do matter.

Citation: Ségurel L, Martínez-Cruz B, Quintana-Murci L, Balaresque P, Georges M, et al. (2008) Sex-Specific Genetic Structure and Social Organization in Central Asia: Insights from a Multi-Locus Study. PLoS Genet 4(9): e1000200. doi:10.1371/journal.pgen.1000200

Editor: Molly Przeworski, University of Chicago, United States of America

Received April 7, 2008; **Accepted** August 18, 2008; **Published** September 26, 2008

Funding: This work was supported by the Centre National de la Recherche Scientifique (CNRS) ATIP programme (to EH), by the CNRS interdisciplinary programme "Origines de l'Homme du Langage et des Langues" (OHLL) and by the European Science Foundation (ESF) EUROCORES programme "The Origin of Man, Language and Languages" (OMLL). We also thank the "Fondation pour la Recherche Médicale" (FRM) for financial support. LS is financed by the French Ministry of Higher Education and Research. MAJ is supported by a Wellcome Trust Senior Fellowship in Basic Biomedical Science (grant number 057559).

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: lsegurel@mnhn.fr

¤ Current address: Unidad de Biología Evolutiva, Departamento de Ciencias Experimentales y de la Salud, Universidad Pompeu Fabra, Barcelona, Spain

Introduction

Understanding the extent to which sex-specific processes shape human genetic diversity has long been a matter of great interest for human population geneticists [1,2]. To date, as detailed in Table 1, the focus has mainly been on the analysis of uniparentally inherited markers: mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY). A large number of studies have found that the level of differentiation was greater for the Y chromosome than for mtDNA, both at a global [3] and a local scale [4–11], for a review see [12]. This result has mainly been explained by patrilocality, a widespread tendency for men to stay in their birthplace while women move to their husband's house [13] (see Table 1 for more detailed interpretations). This hypothesis of a higher migration rate of women has been especially strengthened by the comparison of patrilocal and matrilocal populations at a local scale [14–17]. These studies have shown that in patrilocal populations, genetic differentiation is stronger among men than among women, while the reverse is observed in matrilocal populations. It is also noteworthy that the absolute difference between male and female genetic structure is more pronounced in patrilocal than in matrilocal populations [16]. Interestingly, while social practices seem to consistently influence the sex-specific demography at a local scale, the robustness of a sex-specific genetic structure at a global scale is still a challenging issue (see Table 1). A recent analysis of mtDNA and NRY variation at a global scale, which used the same panel of populations for both categories of markers (an omission that was criticized in Seielstad et al.'s [3] study [18]) showed no difference between the male and female genetic structure [19]. Consistent with this result, an analysis of the autosomal and X-linked microsatellite markers in the HGDP-CEPH Human Genome Diversity Cell Line Panel showed no major differences between the demographic history of men and women [20]. The apparent paradox between local and global trends can be resolved though, since the geographical clustering of populations with potentially

Author Summary

Human evolutionary history has been investigated mainly through the prism of genetic variation of the Y chromosome and mitochondrial DNA. These two uniparentally inherited markers reflect the demographic history of males and females, respectively. Their contrasting patterns of genetic differentiation reveal that women are more mobile than men among populations, which might be due to specific marriage rules. However, these two markers provide only a limited understanding of the underlying demographic processes. To obtain an independent picture of sex-specific demography, we developed a new multi-locus approach based on the analysis of markers from the autosomal and X-chromosomal compartments. We applied our method to 21 human populations sampled in Central Asia, with contrasting social organizations and lifestyles. We found that, in patrilineal populations, not only the migration rate but also the number of reproductive individuals is likely to be higher for women. This result does not hold for bilineal populations, for which both the migration rate and the number of reproductive individuals can be equal for both sexes. The social organization of patrilineal populations is the likely cause of this pattern. This study suggests that differences in sex-specific migration rates may not be the only cause of contrasting male and female differentiation in humans, and that differences in effective numbers do matter.

different lifestyles may minimize the differences in sex-specific demography at a global scale [21,22]. It may also be that the global structure reflects more ancient, pre-agricultural, social patterns, as patrilocality may only have increased in human societies only with the recent transition to agriculture [12].

The higher differentiation level found on the NRY as compared to mtDNA at a local scale could also be the consequence of a higher effective number of women, for example through the practice of polygyny, a tendency for men (but not for women) to have multiple mates [4,7,15,23–25], and/or through the paternal transmission of reproductive success [11]. However, the influence of such processes on genetic structure has often been considered as negligible, since realistic rates of polygyny cannot create large differences in male and female genetic structure [3,5,14]. Hence, until now, the effect of local social processes on male and female effective numbers has not been investigated directly, possibly because current methods fail to unravel the relative contribution of effective number and migration rate on the differentiation level [26]. The consequence is that the vast majority of studies fail to show whether the observed differentiation arises from sex-specific differences in migration rate, effective numbers, or both (see Table 1). New methods need therefore to be developed in order to appreciate the relative influence of sex-biased dispersal and differences in effective numbers on genetic structure.

Another limitation to the use of uniparentally inherited markers stems from the fact that each of them is, in effect, a single genetic locus. For that reason, we cannot test for the robustness of the sex-specific genetic structure on these markers. We cannot either rule out the possibility that mtDNA and NRY, which contain multiple linked genes, may be shaped by selection [27,28]. This raises the question of whether results based on uniparentally inherited markers simply reflect stochastic variation, or real differences in sex-specific demography. To answer this question, we propose a novel approach based on the joint analysis of autosomal and X-linked markers. This multi-locus analysis has the potential of providing more robust information, as these markers give an independent picture of sex-specific demography. This approach also aims to disentangle the effects of sex-biased dispersal and effective numbers on genetic structure.

In order to recognize the impact of social organization on these differences, we investigate sex-specific genetic structure in human populations of Central Asia (Figure 1), where various ethnic groups, characterized by different languages, lifestyles and social organizations, co-exist. Although all groups share a patrilocal organization, Tajiks (sedentary agriculturalists) are bilineal, i.e. they are organized into nuclear or extended families where blood links and rights of inheritance through both male and female ancestors are of equal importance, and they preferentially establish endogamous marriages with cousins. By contrast, Kazaks, Karakalpaks, Kyrgyz and Turkmen (traditionally nomadic herders) are patrilineal, i.e. they are organized into paternal descent groups (tribes, clans, lineages), and they practice exogamous marriages, in which a man chooses a bride from a different clan.

Results/Discussion

Uniparentally-Inherited Markers

We sampled 780 healthy adult men from 10 populations of bilineal agriculturalists and 11 populations of patrilineal herders from West Uzbekistan to East Kyrgyzstan, representing 5 ethnic groups (Tajiks, Kyrgyz, Karakalpaks, Kazaks, and Turkmen) (see Figure 1 and Table 2). We genotyped all bilineal populations, and 8 out of 11 patrilineal populations at the HVS-I locus of mtDNA, and at 11 microsatellite markers on the NRY (for more details on the markers used, see Table 3). The overall genetic differentiation was higher for NRY, as compared to mtDNA, both among the 10 bilineal agriculturalist populations $\left(F_{\mathrm{ST}}^{(Y)}=0.069 \text{ vs. } F_{\mathrm{ST}}^{(mtDNA)}=0.034\right)$, and among the subset of 8 patrilineal herder populations $\left(F_{\mathrm{ST}}^{(Y)}=0.177 \text{ vs. } F_{\mathrm{ST}}^{(mtDNA)}=0.010\right)$. Assuming an island model of population structure, this implies that female migration rate (m_{f}), and/or the effective number of females (N_{f}), is higher than of the corresponding parameters for males (m_{m} and N_{m}). These results also suggest that the differences in sex-specific genetic structure are much more pronounced in the patrilineal herders than in the bilineal agriculturalists. From the above F_{ST} estimates, we obtained the female-to-male ratio of the effective number of migrants per generation (see the Methods section for details): $N_{\mathrm{f}}m_{\mathrm{f}}/N_{\mathrm{m}}m_{\mathrm{m}}\approx 2.1$ for bilineal populations and $N_{\mathrm{f}}m_{\mathrm{f}}/N_{\mathrm{m}}m_{\mathrm{m}}\approx 21.6$ for patrilineal populations. The ratio in patrilineal populations is thus one order of magnitude higher than in bilineal populations. However, since each of these markers is a single genetic locus, we cannot test for the robustness of the sex-specific genetic structure on these markers. We therefore examined the amount of information contained in multi-locus data on autosomal and X-linked markers, both of which average over male and female histories.

A New Multi-Locus Approach

In the infinite island model of population structure with two classes of individuals (males and females), we obtained the following expressions of F_{ST} (see the Methods section for details):

$$F_{\mathrm{ST}}^{(A)} \approx \frac{1}{1+4\frac{4N_{\mathrm{f}}N_{\mathrm{m}}}{N_{\mathrm{f}}+N_{\mathrm{m}}}\frac{m_{\mathrm{f}}+m_{\mathrm{m}}}{2}}, \quad (1)$$

for autosomal genes, and

$$F_{\mathrm{ST}}^{(X)} \approx \frac{1}{1+4\frac{9N_{\mathrm{f}}N_{\mathrm{m}}}{2N_{\mathrm{f}}+4N_{\mathrm{m}}}\frac{2m_{\mathrm{f}}+m_{\mathrm{m}}}{3}}, \quad (2)$$

Table 1. Human sex-specific demography inferred from genetic data.

Region	Markers	Method	Social organization[a]	Differences in demographic parameters between males and females[b]			References
				Sex-biased migration		Skewed effective population size	
GLOBAL	mtDNA, NRY SNPs[c]	Genetic structure (AMOVA[d])	NA[e]	None			[19]
GLOBAL	Autosomal STRs[f], X-linked STRs	Genetic structure (AMOVA)	NA	None			[20]
GLOBAL	mtDNA, NRY SNPs[c]	Coalescent-based (TMRCA[g] estimates)	NA	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (polygyny)	[24]
GLOBAL[h]	mtDNA, NRY STRs+SNPs, Autosomal STRs+SNPs	Genetic structure (F_{ST})	NA	$m_f > m_m$ (patrilocality)		Considered as negligible[i]	[3]
GLOBAL[h]	NRY SNPs	Coalescent-based (mismatch distributions)	NA	Not considered[j]		$N_f > N_m$ (polygyny)	[23]
India	mtDNA	Genetic structure (R_{ST}, haplotype sharing)	Endogamy, patrilocality	None			[21]
	NRY STRs		Endogamy, matrilocality	None			
Sinai peninsula	mtDNA, NRY	Genetic diversity	Endogamy and rare patrilocal exogamy, polygyny	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (polygyny)	[4]
West New Guinea	mtDNA, NRY STRs+SNPs	Genetic structure and diversity (F_{ST}, R_{ST}, haplotype diversity)	Exogamy, patrilocality, patrilineality, polygyny	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (polygyny, warfare)	[7]
Sub-Saharan Africa	mtDNA, NRY STRs+SNPs	Genetic structure (AMOVA)	FPP[k]: patrilocality, high polygyny	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (polygyny)	[15]
			HGP[l]: moderate patrilocality, low polygyny	$m_f < m_m$ (multilocality)	and/or	$N_f < N_m$	
Thailand	mtDNA, NRY STRs	Coalescent-based (Approximate Bayesian Computation)	Patrilocality	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (patrilocality)	[16]
			Matrilocality	$m_f < m_m$ (matrilocality)	and/or	$N_f < N_m$ (matrilocality)	
Eastern North America	mtDNA, NRY STRs+SNPs	Genetic structure (AMOVA), coalescent-based (MIGRATE[m])	Patrilocality, patrilineality	$m_f > m_m$ (patrilocality)	and/or	$N_f > N_m$ (patrilocality)	[17]
			Matrilocality, matriliny	$m_f < m_m$ (matrilocality)	and/or	$N_f < N_m$ (matrilocality)	
Central Asia (pastoral populations)	mtDNA, NRY STRs	Genetic structure and diversity (AMOVA, R_{ST})	Exogamy, patrilineality	$m_f > m_m$ (patrilineality, exogamy)	and/or	$N_f > N_m$ (patrilineality,VRS[n])	[11]
New Britain	mtDNA, NRY SNPs, X-linked loci	Coalescent-based (θ^o and TMRCA estimates)	No strong endogamy, ambilocality, polygyny	$m_f < m_m$	and	$N_f > N_m$ (polygyny)	[25]
Central Asia	mtDNA, NRY STRs	Genetic structure (AMOVA)	Exogamy, patrilocality, polygyny	$m_f > m_m$ (patrilocality)		Considered as negligible	[5]
Thailand	mtDNA, NRY STRs	Genetic structure and diversity (haplotype diversity, R_{ST})	Patrilocality	$m_f > m_m$ (patrilocality)		Considered as negligible	[14]
			Matrilocality	$m_f < m_m$ (matrilocality)		Considered as negligible	
Sub-Saharan Africa[h]	mtDNA, NRY SNPs	Genetic structure and diversity (haplotype diversity, AMOVA)	NA[c]	$m_f < m_m$		Not considered	[22]
Continental Asia[h]	mtDNA, NRY SNPs	Genetic structure (F_{ST})	NA[c]	$m_f > m_m$ (patrilocality)		Not considered	[6]
Russia	mtDNA, NRY SNPs	Genetic structure (F_{ST})	Patrilocality, patrilineality	$m_f > m_m$ (patrilocality)		Not considered	[8]
Caucasus	mtDNA, NRY SNPs	Genetic structure (AMOVA)	NA	$m_f > m_m$ (patrilocality)		Not considered	[9]
Turkey	mtDNA, NRY STRs+SNPs	Genetic structure (AMOVA)	NA	$m_f > m_m$ (patrilocality)		Not considered	[10]

Table 1. cont.

This table summarizes the observed patterns of sex-specific differences in demographic parameters reported in a number of recent studies. The first column lists the location of the sampled populations, or indicates whether the study is conducted at a global scale. The second column gives the markers used, and the third column indicates the statistical methods employed. The fourth column provides indications on social organization, available a priori for the populations under study. In the fifth and sixth columns, the authors' interpretations of sex-specific differences in demographic parameters are given, with respect to skewed gene flow and/or effective numbers.

[a]Indications on social organization, marriage rules, etc., as provided by the authors.

[b]The differences in demographic parameters between males and females, as inferred by the authors, are given in terms of sex-biased gene flow, and skewed effective numbers; the authors' interpretation to the observed pattern is given in parentheses, when available.

[c]Single nucleotide polymorphisms.

[d]Analysis of molecular variance [69].

[e]Not available (no detailed information given by the authors concerning social organization, marriage rules, etc.).

[f]Short tandem repeats.

[g]Time to the most recent common ancestor.

[h]mtDNA and NRY were not sampled in the same individuals or populations.

[i]The authors discussed a possible difference in demographic parameters between males and females, but considered it as negligible.

[j]The authors did not consider this pattern.

[k]Food-producer populations.

[l]Hunter-gatherer populations.

[m]Monte Carlo Markov chain method to estimate population sizes and migration rates [70].

[n]Variance in Reproductive Success.

[o]population-mutation parameter.

doi:10.1371/journal.pgen.1000200.t001

for X-linked genes. A special case of interest occurs when $F_{ST}^{(X)} = F_{ST}^{(A)}$, i.e. when the differentiation of X-linked genes exactly equals that of autosomal genes. Combining eqs (1) and (2), we find that this occurs for $\frac{m_f}{m} = \left(5-4\frac{N_f}{N}\right)/3$, with $N = N_f + N_m$ and $m = m_f + m_m$. Furthermore, as shown in Figure 2, if we observe a lower genetic differentiation of autosomal markers, as compared to X-linked markers (blue zone in Figure 2), this suggests that $\frac{m_f}{m} < \left(5-4\frac{N_f}{N}\right)/3$. This may happen, e.g., for $N_f = N_m$ and $m_f = m_m$, i.e. for equal effective numbers of males and females and unbiased dispersal. But if autosomal markers are more differentiated than X-linked markers ($F_{ST}^{(A)} > F_{ST}^{(X)}$, see the red upper-right triangle in Figure 2), this implies that $\frac{m_f}{m} > \left(5-4\frac{N_f}{N}\right)/3$. In this case, since m_f/m and N_f/N are ratios varying between 0 and 1, the effective number of females must be higher than that of males ($N_f > N_m$), and the female migration rate must be higher than half the male migration rate ($m_f > m_m/2$). Hence, a prediction from this model is that when $F_{ST}^{(A)} > F_{ST}^{(X)}$, the effective number of females is higher than that of males, whatever the pattern of sex-specific dispersal. This suggests that it is indeed possible to test for differences in effective numbers between males and females from the joint analysis of autosomal and X-linked data. We note however that when $F_{ST}^{(X)} > F_{ST}^{(A)}$, we cannot conclude on the relative male and female effective numbers and migration rates.

We tested the above prediction in the 10 bilineal agriculturalist populations and 11 patrilineal herder populations sampled in Central Asia by comparing the genetic structure estimated from 27 unlinked polymorphic autosomal microsatellite markers ($AR = 16.2$, $H_e = 0.803$ on average) to that from 9 unlinked polymorphic X-linked microsatellite markers ($AR = 12.6$, $H_e = 0.752$ on average) (for more details on the markers used, see Table 4). Overall heterozygosity was not significantly different between X-linked and autosomal markers, neither in the pooled sample (two-tailed Wilcoxon sum rank test; $p = 0.09$), nor in the bilineal agriculturalists ($p = 0.13$) or the patrilineal herders ($p = 0.12$). The overall population structure was significantly higher for autosomal as compared to X-linked markers among patrilineal herders: $F_{ST}^{(A)} = 0.008$ $[0.006-0.010]$ and $F_{ST}^{(X)} = 0.003$ $[0.001-0.006]$ (one-tailed Wilcoxon sum rank test; $H_0 : F_{ST}^{(A)} = F_{ST}^{(X)}$; $H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}$; $p = 0.02$). Among bilineal agriculturalists, the result was not significant: $F_{ST}^{(A)} = 0.014$ $[0.012-0.016]$ and $F_{ST}^{(X)} = 0.013$ $[0.008-0.018]$ ($p = 0.36$). From these results, and following our model predictions, we conclude that in patrilineal herders (where $F_{ST}^{(A)} > F_{ST}^{(X)}$), the effective number of females is higher than that of males. This conclusion does not hold for the bilineal agriculturalists.

From our model, it is possible to get more precise indications on the sets of $(N_f/N,\ m_f/m)$ values that are compatible with our data. Rearranging eqs (1–2), we get:

$$\frac{1-1\Big/F_{ST}^{(X)}}{1-1\Big/F_{ST}^{(A)}} = \frac{3}{4}\frac{(1+m_f/m)}{(2-N_f/N)}, \tag{3}$$

i.e.:

$$F_{ST}^{(X)} = \frac{4F_{ST}^{(A)}}{4F_{ST}^{(A)} - 3\left(F_{ST}^{(A)}-1\right)\left(\frac{1+m_f/m}{2-N_f/N}\right)}. \tag{4}$$

For any given set of $(N_f/N,\ m_f/m)$ values, we can therefore calculate from eq. (4) the expected value of $F_{ST}^{(X)}$ for each $F_{ST}^{(A)}$

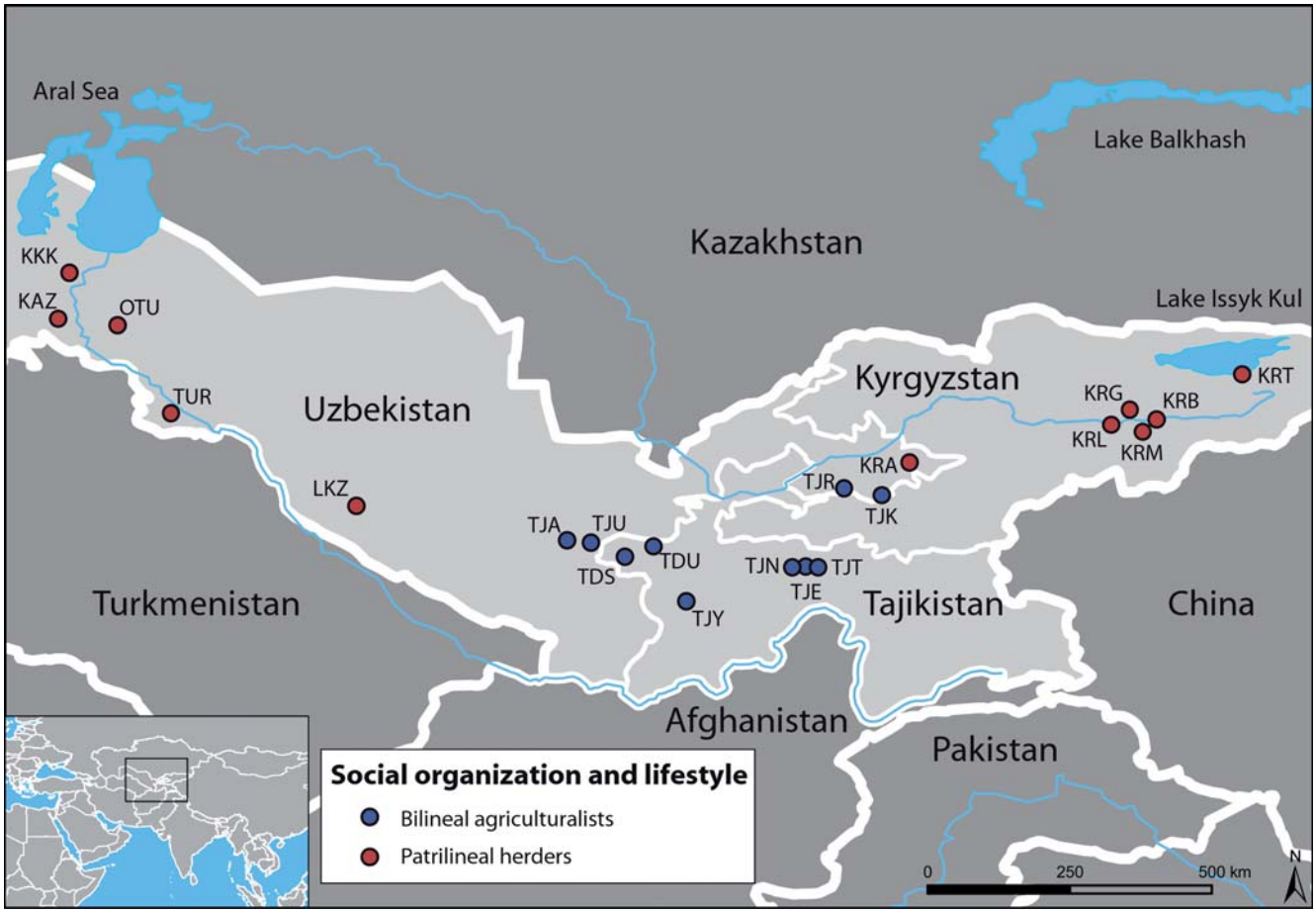


Figure 1. Geographic map of the sampled area, with the 21 populations studied. Bilineal agriculturalist populations are in blue (Tajiks); Patrilineal herders with a semi-nomadic lifestyle are in red (Kazaks, Karakalpaks, Kyrgyz and Turkmen).
doi:10.1371/journal.pgen.1000200.g001

estimate in the dataset. We can then test the null hypothesis $H_0 : F_{ST}^{(X)} = 4F_{ST}^{(A)} \Big/ \left[4F_{ST}^{(A)} - 3\left(F_{ST}^{(A)} - 1\right)\left(\frac{1+m_f/m}{2-N_f/N}\right)\right]$ by comparing the distribution of observed and expected $F_{ST}^{(X)}$ values. If the hypothesis can be rejected at the $\alpha = 0.05$ level, then the corresponding set of $(N_f/N,\ m_f/m)$ values can also be rejected. Following Ramachandran et al. [20], we varied the values of the ratios N_f/N and m_f/m (respectively, the female fraction of effective number, and the female fraction of the total migration rate) from 0 to 1, with an interval of 0.01 between consecutive values. For each set of $(N_f/N,\ m_f/m)$ values, we applied the transformation in eq. (4) to each of the 27 locus-specific $F_{ST}^{(A)}$ values observed. Thus, for each set of $(N_f/N,\ m_f/m)$ values, we obtained 27 expected values of $F_{ST}^{(X)}$, given our data. These expected values of $F_{ST}^{(X)}$ were then compared to the 9 observed locus-specific $F_{ST}^{(X)}$ in our dataset, and we calculated the p-value for a two-sided Wilcoxon sum rank test between the list of 27 expected $F_{ST}^{(X)}$ values and the 9 $F_{ST}^{(X)}$ observed in the dataset. The results are depicted in Figure 3. Significant p-values ($p \leq 0.05$) correspond to a significant difference between the observed and expected values, thus to sets of $(N_f/N,\ m_f/m)$ values that are rejected, given our data (see the blue region in Figure 3). Conversely, non-significant p-values ($p > 0.05$) correspond to sets of $(N_f/N,\ m_f/m)$ values that cannot be rejected (see the red region in Figure 3).

For the patrilineal herder populations (Figures 3A–3B), most sets of $(N_f/N,\ m_f/m)$ values are rejected, except those corresponding to larger effective numbers for females (from Figures 3A–3B: $N_f/N > 0.55$, i.e. $N_f > 1.27N_m$) and $m_f > 0.67m_m$. Because the multi-locus estimate of $F_{ST}^{(A)}$ is significantly higher than the estimate of $F_{ST}^{(X)}$, we expected to find such patterns of non-significant values (see Figure 2). For the bilineal agriculturalist populations, we could not reject the hypothesis that the effective numbers and migration rates are equal across males and females or even lower in females (see Figures 3C–3D). This is also reflected by the fact that the estimates of $F_{ST}^{(A)}$ were not significantly higher than the estimates of $F_{ST}^{(X)}$ in those populations.

Finally, we have shown that the effective number of women is higher than that of men among patrilineal herders, but not necessarily among bilineal agriculturalists. Furthermore, a close inspection of the results depicted in Figures 3A and 3B reveals that, among herders, we reject all the sets of $(N_f/N,\ m_f/m)$ values for which $m_f < m_m$ at the $\alpha = 0.10$ level. This is not true for agriculturalists. This suggests that the migration rates are also likely to be higher for women than for men in patrilineal populations, as compared to bilineal populations (compare Figures 3B and 3D). Although both groups are patrilocal, such a difference in sex-specific migration patterns might be expected, since patrilineal herders are exogamous (among clans) and bilineal agriculturalists are preferentially endogamous. For example, it was observed that in patrilocal and matrilocal Indian populations, where migrations are strictly confined within endogamous groups, sex-specific patterns were not influenced by post-marital residence [21].

Table 2. Sample description.

Sampled populations (area)	Acronym	Location	Long.	Lat.	n_X	n_A	n_Y	n_{mt}
Bilineal agriculturalists								
Tajiks (Samarkand)	TJA	Uzbekistan/Tajikistan border	**39.54**	**66.89**	26	31	32	32
Tajiks (Samarkand)	TJU	Uzbekistan/Tajikistan border	**39.5**	**67.27**	27	29	29	29
Tajiks (Ferghana)	TJR	Tajikistan/Kyrgyzstan border	**40.36**	**71.28**	30	29	29	29
Tajiks (Ferghana)	TJK	Tajikistan/Kyrgyzstan border	**40.25**	**71.87**	26	26	35	40
Tajiks (Gharm)	TJE	Northern Tajikistan	**39.12**	**70.67**	29	25	27	31
Tajiks (Gharm)	TJN	Western Tajikistan	**38.09**	**68.81**	33	24	30	35
Tajiks (Gharm)	TJT	Northern Tajikistan	**39.11**	**70.86**	31	25	30	32
Tajiks (Penjinkent)	TDS	Uzbekistan/Tajikistan border	**39.28**	**67.81**	30	25	31	31
Tajiks (Penjinkent)	TDU	Uzbekistan/Tajikistan border	**39.44**	**68.26**	40	25	31	40
Tajiks (Yagnobs from Douchambe)	TJY	Western Tajikistan	**38.57**	**68.78**	39	25	36	40
Patrilineal herders with a semi-nomadic lifestyle								
Karakalpaks (Qongrat from Karakalpakia)	KKK	Western Uzbekistan	**43.77**	**59.02**	56	45	54	55
Karakalpaks (On Tört Uruw from Karakalpakia)	OTU	Western Uzbekistan	**42.94**	**59.78**	49	45	54	53
Kazaks (Karakalpakia)	KAZ	Western Uzbekistan	**43.04**	**58.84**	47	49	50	50
Kazaks (Bukara)	LKZ	Southern Uzbekistan	**40.08**	**63.56**	20	25	20	31
Kyrgyz (Andijan)	KRA	Tajikistan/Kyrgyzstan border	**40.77**	**72.31**	31	45	46	48
Kyrgyz (Narin)	KRG	Middle Kyrgyzstan	**41.6**	**75.8**	20	18	20	20
Kyrgyz (Narin)	KRM	Middle Kyrgyzstan	**41.45**	**76.22**	21	21	22	26
Kyrgyz (Narin)	KRL	Middle Kyrgyzstan	**41.36**	**75.5**	36	22	-	-
Kyrgyz (Narin)	KRB	Middle Kyrgyzstan	**41.25**	**76**	31	24	-	-
Kyrgyz (Issyk Kul)	KRT	Eastern Kyrgyzstan	**42.16**	**77.57**	33	37	-	-
Turkmen (Karakalpakia)	TUR	Western Uzbekistan	**41.55**	**60.63**	42	47	51	51

Long., longitude; Lat., latitude. n_X, n_A, n_Y and n_{mt}: sample size for X-linked, autosomal, Y-linked and mitochondrial markers, respectively.
doi:10.1371/journal.pgen.1000200.t002

Table 3. Level of diversity and differentiation for NRY markers and mtDNA.

NRY markers			F_{ST}	
Locus name	**Allelic richness (AR)**	H_e	**Herders**	**Agriculturalists**
DYS426	4	0.500	0.3326	0.0068
DYS393	8	0.492	0.1095	0.0517
DYS390	8	0.739	0.1229	0.1253
DYS385 a/b	15	0.858	0.1414	0.0278
DYS388	9	0.531	0.3003	0.0736
DYS19	7	0.743	0.1081	0.1310
DYS392	10	0.516	0.1345	0.0701
DYS391	7	0.495	0.2533	0.0686
DYS389I	6	0.541	0.1537	0.1395
DYS439	7	0.725	0.1638	0.0291
DYS389II	8	0.763	0.1556	0.0395
mtDNA			F_{ST}	
Locus name	**Polymorphic sites**	H_e	**Herders**	**Agriculturalists**
HVS-I	121	0.0156	0.0098	0.0343

We calculated the total allelic richness (*AR*) (over all populations) and the expected heterozygosity H_e [55] using Arlequin version 3.1 [56]. Genetic differentiation among populations was measured both per locus and overall loci, using Weir and Cockerham's F_{ST} estimator [57], as calculated in GENEPOP 4.0 [58]. We calculated the total number of polymorphic sites, the unbiased estimate of expected heterozygosity H_e [55], and F_{ST} using Arlequin version 3.1 [56].
doi:10.1371/journal.pgen.1000200.t003

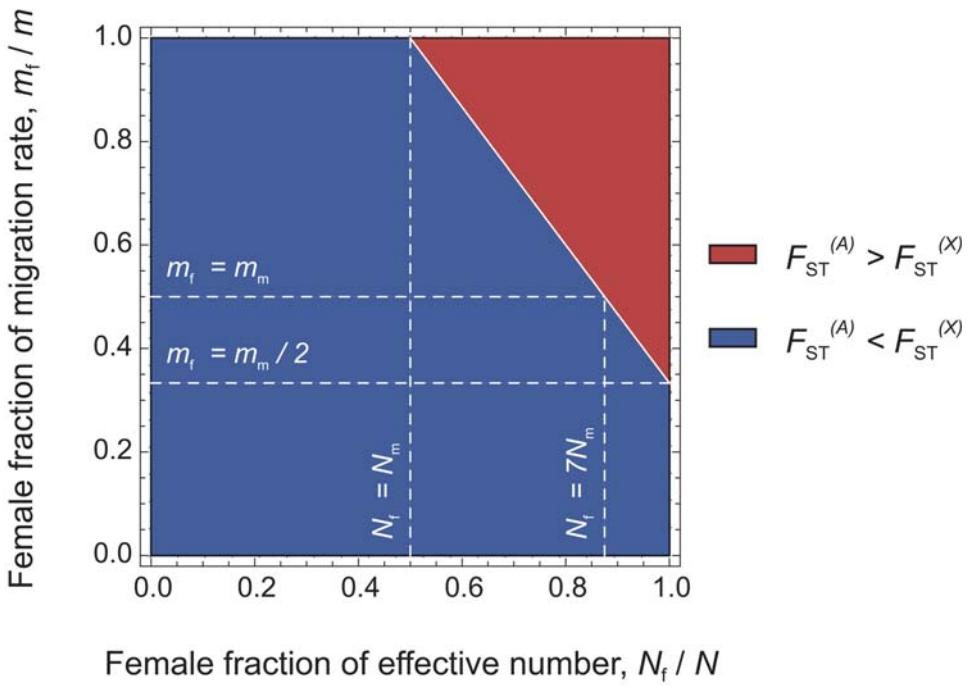


Figure 2. Diagram representing the relative values of expected genetic differentiation for autosomal markers $\left(F_{ST}^{(A)}\right)$ and for X-linked markers $\left(F_{ST}^{(X)}\right)$. In the red upper right triangle, the F_{ST} estimates for autosomal markers are higher than for X-linked markers. In this case, N_f/N is necessarily larger than 0.5. In the blue region of the figure, the F_{ST} estimates for autosomal markers are lower than for X-linked markers. The white plain line, at which $\frac{m_f}{m} = \left(5 - 4\frac{N_f}{N}\right)/3$, represents the set of (N_f/N, m_f/m) values where the autosomal and X-linked F_{ST} estimates are equal. In this case $\left(F_{ST}^{(X)} = F_{ST}^{(A)}\right)$, if $N_f = N_m$, then the lower effective size of X-linked markers (which would be three-quarters that of autosomal markers) can only be balanced by a complete female-bias in dispersal ($m_f/m = 1$). Conversely, if $m_f = m_m$, the large female fraction of effective numbers compensates exactly the low effective size of X-linked markers only for $N_f = 7N_m$. Last, if $m_f = m_m/2$, then the autosomal and X-linked F_{ST} estimates can only be equal as the number of males tends towards zero.
doi:10.1371/journal.pgen.1000200.g002

What Could Explain a Larger Effective Number of Females?

While an influence of post-marital residence on the migration rate of women and men has already been widely proposed [14–17] (see also Table 1), the factors that may locally affect the effective number of women, relatively to that of men, are not well recognized. As seen in Table 1, although a number of studies have compared matrilocal and patrilocal populations, few have compared contrasting groups of populations with respect to other factors as, e.g., the tendency for polygyny [15]. Furthermore, a number of these studies lack ethnological information a priori, concerning social organization, marriage rules, etc., which makes interpretation somewhat difficult (see Table 1). Here, we compared two groups of patrilocal populations with contrasting social organizations, and at least five non-mutually exclusive interpretations for a larger effective number of females can be invoked:

(*i*) *Social organization*, i.e. the way children are affiliated to their parents, can deeply affect sex-specific genetic variation. In Central Asia, herder populations are organized in patrilineal descent groups (tribes, clans, lineages). This implies that children are systematically affiliated with the descent groups of the father. Chaix et al. [11] showed that the average number of individuals carrying the same Y chromosome haplotype was much higher in patrilineal herder populations than in bilineal agriculturalist populations (where children are affiliated both to the mother and the father). These "identity cores" would be the direct consequence of the internal dynamics of their patrilineal organization. Indeed, the descent groups are not formed randomly and related men tend to cluster together, e.g. through the recurrent lineal fission of one population into new groups. This particular dynamics increases relatedness among men, and may therefore reduce the effective number of men, as compared to women.

(*ii*) Indirectly, the social organization can also deflate the effective number of men through *the transmission of reproductive success* [29] if this success is culturally transmitted exclusively from fathers to sons. Because herders are patrilineal (so that inheritance is organized along paternal descent groups), social behaviors are more likely to be inherited through the paternal line of descent only. It has recently been argued that the rapid spread of Genghis Khan's patrilineal descendants throughout Central Asia was explained by this social selection phenomenon [30]. The correlation of fertility through the patriline has also been described in patrilineal tribes in South America [31]. By contrast, in bilineal societies such as the agriculturalists of Central Asia, social behaviors that influence reproductive success are more likely to be transmitted by both sexes. Furthermore, differences of cultural transmission of fitness between hunter-gatherers and agriculturalists have already been reported [32]. Interestingly, a slightly higher matrilineal intergenerational correlation in offspring number has been observed in the Icelandic population, which suggests that in some populations, reproductive behaviors can be maternally-inherited [33].

Table 4. Level of diversity and differentiation for X-linked and autosomal markers.

			F_{ST}	
Locus name	**Allelic richness (AR)**	H_e	**Herders**	**Agriculturalists**
X-linked markers				
CTAT014	19	0.746	0.0018	0.0225
GATA124E07	15	0.847	0.0024	0.0136
GATA31D10	8	0.697	0.0069	0.0007
ATA28C05	7	0.722	0.0086	0.0179
AFM150xf10	14	0.832	−0.0021	0.0152
GATA100G03	14	0.734	−0.0019	0.0084
AGAT121P	15	0.593	−0.0016	0.0048
ATCT003	10	0.797	0.0095	0.0261
GATA31F01	11	0.804	0.0069	0.0053
Autosomal markers				
AFM249XC5	19	0.848	0.0080	0.0081
ATA10H11	13	0.680	0.0128	0.0193
AFM254VE1	14	0.837	0.0105	0.0086
AFMA218YB5	14	0.852	0.0030	0.0151
GGAA7G08	22	0.896	0.0096	0.0138
GATA11H10	16	0.776	0.0017	0.0056
GATA12A07	16	0.857	0.0001	0.0163
GATA193A07	15	0.825	0.0064	0.0087
AFMB002ZF1	11	0.820	0.0028	0.0169
AFMB303ZG9	16	0.858	0.0090	0.0148
ATA34G06	12	0.675	0.0088	0.0132
GATA72G09	18	0.884	−0.0023	0.0131
GATA22F11	21	0.897	0.0152	0.0144
GGAA6D03	13	0.831	0.0048	0.0176
GATA88H02	17	0.892	0.0063	0.0056
SE30	15	0.762	0.0084	0.0103
GATA43C11	16	0.870	0.0028	0.0093
AFM203YG9	14	0.753	0.0105	0.0084
AFM157XG3	13	0.753	0.0147	0.0196
UT2095	16	0.738	0.0032	0.0112
GATA28D01	25	0.896	0.0156	0.0139
GGAA4B09	19	0.707	0.0034	0.0208
ATA3A07	12	0.746	0.0078	0.0070
AFM193XH4	11	0.716	0.0164	0.0129
GATA11B12	26	0.896	0.0104	0.0265
AFM165XC11	13	0.785	0.0058	0.0185
AFM248VC5	20	0.620	0.0246	0.0145

We calculated the allelic richness (*AR*) and unbiased estimates of expected heterozygosity H_e [55], obtained both by locus and on average with Arlequin version 3.1 [56]. Genetic differentiation among populations was measured both per locus and overall loci, using Weir and Cockerham's F_{ST} estimator [57] as calculated in GENEPOP 4.0 [58].

(*iii*) *Polygyny*, in which the husband may have multiple wives, has often been invoked as a factor that could reduce the effective number of men [4,7,15,23–25]. While we could not find any evidence of polygyny in present-day Central Asian populations, this custom was traditionally practiced in the nomadic herder Kazak populations, although limited to the top 10 percent of men from the highest social rank [5,34]. Hence, even though we lack ethnological data to determine to what extent herders are or were practicing polygyny in a recent past, the practice of polygyny among herders in Central Asia might have influenced (at least partially) the observed differences in men and women effective numbers.

(*iv*) *Recurrent bottlenecks in men* due to a higher pre-reproductive mortality could also severely reduce the effective numbers of men. From the study of several groups in West Papua and Papua New Guinea [7,35], it appears that warfare may indeed lead to the quasi-extinction of adult men in some communities, while the mass killing of adult women is far more rarely reported. However, this differential mortality could also be balanced by potentially high death rates of women during childbirth. In any case, a differential mortality is equally likely to arise in herder and agriculturalist populations. It may therefore not be relevant in explaining why we detect higher effective numbers of women (as compared to men) in patrilineal herders and not in bilineal agriculturalists.

(*v*) Since our approach implicitly assumes equal male and female generation time, the observed higher effective number of women, relatively to that of men, could result from a *shorter generation time for women*, due to the tendency of women to reproduce earlier in life than men and the ability of men to reproduce at a later age than women. This has indeed been described in a number of populations with different lifestyles, from complete genealogical records or mean-age-at-first-marriage databases [33,36,37]. It has even been proposed to be a nearly universal trait in humans, although its magnitude varies across regions and cultures [37]. Tang et al. [38] suggested that accounting for longer generation time in males could minimize the difference between maternal and paternal demography. However, the differences in sex-specific generation times that have been reported (e.g., 28 years for the matrilines and 31 years for the patrilines in Iceland [33], 29 years for the matrilines and 35 years for the patrilines in Quebec [36]) are unlikely to explain the observed differences in male and female effective numbers [24].

Limits of the Approach

There might also be non-biological explanations of our results, however, as they are based on the simplifying assumptions of Wright's infinite island model of population structure [39]. This model assumes (*i*) that there is no selection and that mutation is negligible, (*ii*) that each population has the same size, and sends and receives a constant fraction of its individuals to or from a common migrant pool each generation (so that geographical structure is absent), and (*iii*) that equilibrium is reached between migration, mutation and drift. On the first point, we did not find any evidence of selection, for any marker, based on Beaumont and Nichols' method [40] for detecting selected markers from the analysis of the null distribution generated by a coalescent-based simulation model (data not shown). As for the second point, we tested for the significance of the correlation between the pairwise $F_{ST}/(1-F_{ST})$ estimates and the natural logarithm of their geographical distances [41]. We found no evidence for isolation by distance, either for X-linked markers ($p = 0.47$ for agriculturalists, $p = 0.24$ for herders), or for autosomal markers ($p = 0.92$ for agriculturalists, $p = 0.45$ for herders). As for the third point, the X-to-autosomes (X/A) effective size ratio can significantly deviate from the expected three-quarters (assuming equal effective numbers of men and women) following a bottleneck or an

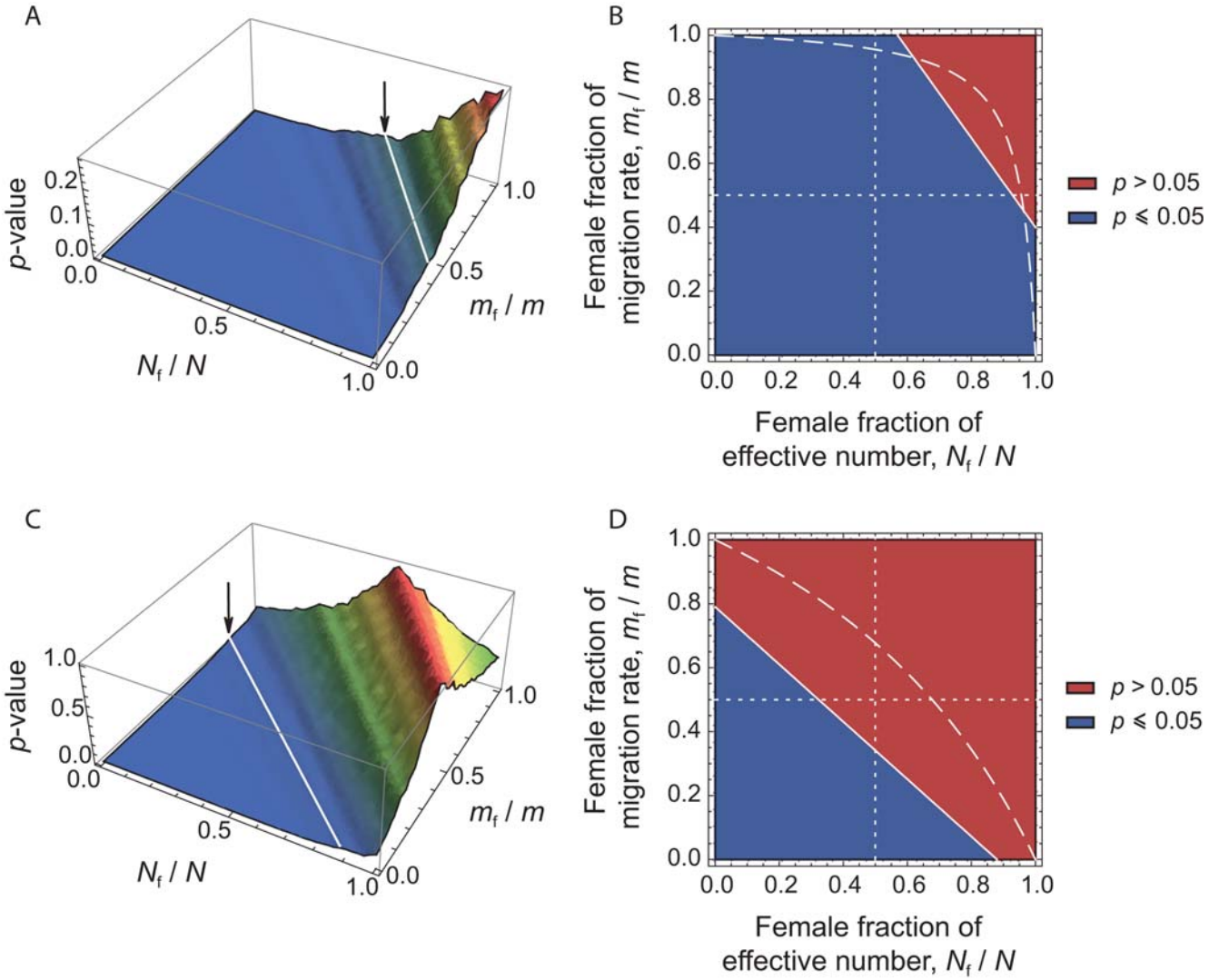


Figure 3. *p*-values of Wilcoxon tests plotted in the (N_f/N, m_f/m) parameter space. For each set of (N_f/N, m_f/m) values, we applied the transformation in eq. (4), and tested whether our data on autosomal and X-linked markers were consistent, given the hypothesis defined by the set of (N_f/N, m_f/m) values. (A) Surface plot of the *p*-values, as a function of the female fraction of effective number and the female fraction of migration rate, for the herders (11 populations). The arrow indicates the line that separates the region where $p \leq 0.05$ from that where $p > 0.05$. Non-significant *p*-values ($p > 0.05$) correspond to the values of (N_f/N, m_f/m) that could not be rejected, given our data. (B) Contour plots, for the same data. The dashed line indicates the range of (N_f/N, m_f/m) values inferred from the ratio of NRY and mtDNA population structure, as obtained from the relationship: $N_f m_f / N_m m_m = \left(1 - 1/F_{ST}^{(mtDNA)}\right) / \left(1 - 1/F_{ST}^{(Y)}\right)$. The dotted lines correspond to the cases where $N_f = N_m$ (vertical line) and $m_f = m_m$ (horizontal line). (C) and (D) as (A) and (B), respectively, for the agriculturalists (10 populations).
doi:10.1371/journal.pgen.1000200.g003

expansion [42]. This is because X-linked genes have a smaller effective size, and hence reach equilibrium more rapidly. After a reduction of population size, the X/A diversity ratio is lower than expected, while after an expansion, the diversity of X-linked genes recovers faster than on the autosomes, and the X/A diversity ratio is then closer to unity. In the latter case, $F_{ST}^{(X)}$ would be reduced and could then tend towards $F_{ST}^{(A)}$. However, neither reduction nor expansion should lead to $F_{ST}^{(X)} < F_{ST}^{(A)}$, as we found in herder populations of Central Asia. Therefore, we do not expect the limits of Wright's island model to undermine our approach.

Evaluation by Means of Stochastic Simulations

We aimed to investigate to what extent the approach proposed here is able to detect differences in male and female effective numbers. To do this, we performed coalescent simulations in a finite island model, for a wide range of (N_f/N, m_f/m) values. The simulation parameters were set to match those of our dataset: 11 sampled demes, 30 males genotyped at 27 autosomal and 9 X-linked markers per deme (for further details concerning the simulations, see the Methods section). We used 1421 sets of (N_f/N, m_f/m) values, covering the whole parameter space (represented as white dots in Figure 4B). For each set of (N_f/N, m_f/m) parameter values, we simulated 100 independent datasets. For each dataset, we calculated the estimates of $F_{ST}^{(A)}$ and $F_{ST}^{(X)}$ at all loci, and we calculated the *p*-value for a one-sided Wilcoxon sum rank test for the list of 27 $F_{ST}^{(X)}$ and 9 $F_{ST}^{(X)}$ estimates $\left(H_0 : F_{ST}^{(A)} = F_{ST}^{(X)};\ H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}\right)$. Hence, for each set of (N_f/N, m_f/m) parameter values, we could calculate the proportion of significant tests at the $\alpha = 0.05$ level, among the 100

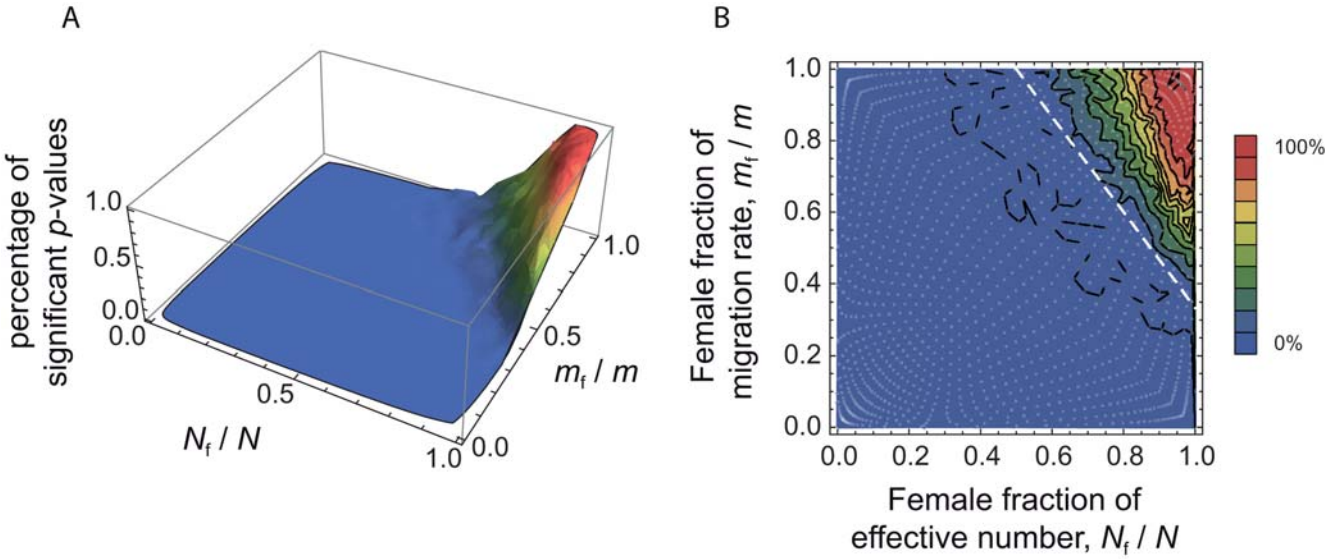


Figure 4. Percentage of significant tests in the (N_f/N, m_f/m) parameter space, for simulated data. We chose a range of 49 ($N_f m_f / N_m m_m$) ratios, varying from 0.0004 to 2401, and for each of these ratios we chose 29 sets of (N_f/N, m_f/m) values. By doing this, we obtained 1421 sets of (N_f/N, m_f/m) values, represented as white dots in the right-hand side panel B, covering the whole parameter space. For each set, we simulated 100 independent datasets using a coalescent-based algorithm, and taking the same number of individuals and the same number of loci for each genetic system as in the observed data. For each dataset, we calculated the *p*-value for a one-sided Wilcoxon sum rank test $\left(H_0 : F_{ST}^{(A)} = F_{ST}^{(X)};\ H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}\right)$, and for each set of ($N_f/N$, m_f/m) values we calculated the percentage of significant *p*-values (at the $\alpha = 0.05$ level). A. Surface plot of the proportion of significant *p*-values (at the $\alpha = 0.05$ level), as a function of the female fraction of effective number and the female fraction of migration rate. B. Contour plot, for the same data. The dotted line, at which $\frac{m_f}{m} = \left(5 - 4\frac{N_f}{N}\right)/3$, represents the set of (N_f/N, m_f/m) values where the autosomal and X-linked F_{ST}'s are equal. The theory predicts that we should only find $F_{ST}^{(A)} > F_{ST}^{(X)}$ in the upper-right triangle defined by the dotted line. Hence, the proportion of significant *p*-values for any set of (N_f/N, m_f/m) values in this upper right triangle gives an indication of the power of the method.
doi:10.1371/journal.pgen.1000200.g004

independent datasets. Figure 4 shows the distribution of the percentage of significant tests in the (N_f/N, m_f/m) parameter space. Theory predicts that in the upper-right triangle where $\frac{m_f}{m} > \left(5 - 4\frac{N_f}{N}\right)/3$, we should have $F_{ST}^{(A)} > F_{ST}^{(X)}$. One can see from Figure 4 that, given the simulation parameters used, the method is conservative: the proportion of significant tests at the $\alpha = 0.05$ level is null outside of the upper-right triangle. However, we find a fairly large proportion of significant tests for large N_f/N and m_f/m ratios which indicates (*i*) that the method presented here has the potential to detect differences in male and female effective numbers, but (*ii*) that only strong differences might be detected, for similarly sized datasets as the one considered here.

Robustness to the Sampling Scheme

We also aimed to investigate whether the results obtained here were robust to our sampling scheme, and that our results were not biased by the inclusion of particular populations. To do this, we re-analyzed both the bilineal agriculturalists and the patrilineal herders datasets, removing one population at a time in each group. For each of these jackknifed datasets, we calculated the *p*-value of a one-sided Wilcoxon sum rank test $\left(H_0 : F_{ST}^{(A)} = F_{ST}^{(X)};\ H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}\right)$, as done on the full datasets. The results are given in Table 5. We found no significant test for any of the bilineal agriculturalist groupings ($p > 0.109$), which supports the idea that, in those populations, both the migration rate and the number of reproductive individuals can be equal for both sexes. In patrilineal herders, the tests were significant at the $\alpha = 0.05$ level for 8 out of 11 population groupings. For the 3 other groupings, the *p*-values were 0.068, 0.078 and 0.073 (see Table 5). Overall, the ratio of $F_{ST}^{(A)}$ over $F_{ST}^{(X)}$ multi-locus estimates ranged from 1.7 to 3.5 in patrilineal herders (and from 0.9 to 1.2 in bilineal agriculturalists). Although in some particular groupings of patrilineal herder populations, the difference in the distributions of $F_{ST}^{(A)}$ and $F_{ST}^{(X)}$ may not be strong enough to be significant, we can clearly distinguish the pattern of differentiation for autosomal and X-linked markers in patrilineal and bilineal groups. Results from coalescent simulations (see above) suggest that this lack of statistical power might be expected for $F_{ST}^{(A)} / F_{ST}^{(X)}$ ratios close to unity. Indeed, we found that the tests were more likely to be significant for fairly large N_f/N and m_f/m ratios (the upper-right red region in Figure 4) which would correspond to $F_{ST}^{(A)} / F_{ST}^{(X)}$ ratios much greater than one.

Comparison with Uniparentally-Inherited Markers

Importantly, our results on X-linked and autosomal markers are consistent with those obtained from NRY and mtDNA (see Figures 3B–3D): in these figures, the dashed line gives all the sets of (N_f/N, m_f/m) values that are compatible with the observed $F_{ST}^{(Y)}$ and $F_{ST}^{(mtDNA)}$ estimates. These are the sets of values that satisfy $\left(\frac{N_f/N}{1 - N_f/N}\right) = 2.1\left(\frac{1 - m_f/m}{m_f/m}\right)$ for the bilineal populations, and $\left(\frac{N_f/N}{1 - N_f/N}\right) = 21.6\left(\frac{1 - m_f/m}{m_f/m}\right)$ for the patrilineal populations, since we inferred $N_f m_f / N_m m_m \approx 2.1$ and $N_f m_f / N_m m_m \approx 21.6$, respectively, for the two groups. For the bilineal agriculturalists (Figure 3D), the set of (N_f/N, m_f/m) values inferred from the $F_{ST}^{(Y)}$ and $F_{ST}^{(mtDNA)}$ estimates fall within the range that was not rejected, given our data on X-linked and autosomal markers. For the patrilineal herders (Figure 3B), the overlap is only partial: from the NRY and mtDNA data only, low N_f/N ratios associated with high m_f/m ratios are as likely as high N_f/N ratios associated with low m_f/m ratios. Yet, it is clear from this figure that a large set of (N_f/N, m_f/m) values inferred

Table 5. Autosomal and X-linked differentiation on jackknifed samples.

Sample removed	$F_{ST}^{(A)}$	$F_{ST}^{(X)}$	*p*-value	$F_{ST}^{(A)}/F_{ST}^{(X)}$
Patrilineal groups				
KAZ	0.0084	0.0050	0.068	1.7
KKK	0.0085	0.0050	0.078	1.7
KRA	0.0078	0.0027	0.022	2.9
KRB	0.0080	0.0030	0.028	2.7
KRG	0.0078	0.0035	0.037	2.2
KRL	0.0086	0.0038	0.018	2.3
KRM	0.0069	0.0023	0.018	3.0
KRT	0.0081	0.0044	0.047	1.8
LKZ	0.0088	0.0025	0.002	3.5
OTU	0.0089	0.0038	0.022	2.3
TUR	0.0054	0.0025	0.073	2.2
Bilineal groups				
TDS	0.0125	0.0109	0.443	1.1
TDU	0.0132	0.0153	0.705	0.9
TJA	0.0144	0.0123	0.109	1.2
TJE	0.0140	0.0133	0.148	1.1
TJK	0.0134	0.0131	0.457	1.0
TJN	0.0148	0.0144	0.387	1.0
TJR	0.0140	0.0141	0.401	1.0
TJT	0.0139	0.0121	0.225	1.1
TJU	0.0139	0.0127	0.283	1.1
TJY	0.0139	0.0116	0.259	1.2

For each group, we removed one sample in turn and calculated the differentiation on autosomal and X-linked markers. The *p*-value gives the result of a one-sided Wilcoxon sum rank test $\left(H_0 : F_{ST}^{(A)} = F_{ST}^{(X)};\ H_1 : F_{ST}^{(A)} > F_{ST}^{(X)}\right)$, as performed on the full dataset.

doi:10.1371/journal.pgen.1000200.t005

from the single-locus estimates $F_{ST}^{(Y)}$ and $F_{ST}^{(mtDNA)}$ can be rejected, given the observed differentiation on X-linked and autosomal markers. All genetic systems (mtDNA, NRY, X-linked and autosomal markers) converge toward the notion that patrilineal herders, in contrast to bilineal agriculturalists, have a strong sex-specific genetic structure. Yet, the information brought by X-linked and autosomal markers is substantial, since we show that this is likely due to both higher migration rates and larger effective numbers for women than for men.

Comparison with Other Studies

Our results, based on the X chromosome and the autosomes, also confirm previous analyses based on the mtDNA and the NRY, showing that men are genetically more structured than women in other patrilocal populations [3–10,14–17] (see also Table 1). A handful of studies have also shown a reduced effective number of men compared to that of women, based on coalescent methods [23,24], but none have considered the influence of social organization on this dissimilarity (see Table 1).

In some respects, our results contrast with those of Wilder and Hammer [25], who studied sex-specific population genetic structure among the Baining of New Britain, using mtDNA, NRY, and X-linked markers. Interestingly, they found that $N_f > N_m$, but $m_f < m_m$, and claimed that a similar result, although left unexplored by the authors, was to be found in a recent study by Hamilton et al. [16]. This raises the interesting point that sex-specific proportions of migrants (m) are likely to be shaped by factors that may only partially overlap with those that affect the sex-specific effective numbers (N). Further studies of human populations with contrasted social organizations, as well as further theoretical developments, are needed to appreciate this point.

In order to ask to what extent our results generalize to other human populations, we investigated sex-specific patterns in the 51 worldwide populations represented in the HGDP-CEPH Human Genome Diversity Cell Line Panel dataset [43], for which the data on the differentiation of 784 autosomal microsatellites and 36 X-linked microsatellites are available (data not shown). By doing this, we found a larger differentiation for X-linked than for autosomal markers $\left(F_{ST}^{(X)} > F_{ST}^{(A)}\right)$. Therefore, we confirmed Ramachandran et al.'s [20] result that no major differences in demographic parameters between males and females are required to explain the X-chromosomal and autosomal results in this worldwide sample. Ramachandran et al.'s approach [20] is based upon a pure divergence model from a single ancestral population, which is very different from the migration-drift equilibrium model considered here. In real populations, however, genetic differentiation almost certainly arises both through divergence and limited dispersal, which places these two models at two ends of a continuum. Yet, importantly, if we apply Ramachandran et al.'s [20] model to the Central Asian data, our conclusions are left unchanged. In their model, the differentiation among populations is $F_{ST} \approx 1 - e^{-t/(2N_e)}$, where t is the time since divergence from an ancestral population and N_e the effective size of the populations (see, e.g., [44]). Hence, we get $F_{ST}^{(A)} \approx 1 - e^{-t/\left(2N_e^{(A)}\right)}$ and $F_{ST}^{(X)} \approx 1 - e^{-t/\left(2N_e^{(X)}\right)}$ for autosomal and X-linked markers, respectively. Therefore, our observation that $F_{ST}^{(A)} > F_{ST}^{(X)}$ implies that $N_e^{(X)} > N_e^{(A)}$, which requires that $N_f > 7N_m$ since $N_e^{(A)} = 8N_f N_m/(N_f + N_m)$ and $N_e^{(X)} = 9N_f N_m/(N_f + 2N_m)$ (see, e.g., [45]). In this case, the female fraction of effective number is larger than that of males, which is consistent with our findings in a model with migration.

The HGDP-CEPH dataset does not provide any detailed ethnic information for the sampled groups, and we can therefore not distinguish populations with different lifestyles. However, at a more local scale in Pakistan, we were able to analyze a subset of 5 populations (Brahui, Balochi, Makrani, Sindhi and Pathan), which are presumed to be patrilineal [46]. For this subset, we found a higher differentiation for autosomal $\left(F_{ST}^{(A)} = 0.003\right)$ than for X-linked markers $\left(F_{ST}^{(X)} = 0.002\right)$, although non-significantly ($p = 0.12$). This result seems to suggest that other patrilineal populations may behave like the Central Asian sample presented here. Therefore, because the geographical clustering of populations with potentially different lifestyles may minimize the differences in sex-specific demography at a global scale [21,22], and/or because the global structure may reflect ancient (pre-agricultural) marital residence patterns with less pronounced patrilocality [12], we emphasize the point that large-scale studies may not be relevant to detect sex-specific patterns, which supports a claim made by many authors.

Conclusion

In conclusion, we have shown here that the joint analysis of autosomal and X-linked polymorphic markers provides an efficient tool to infer sex-specific demography and history in human populations, as suggested recently [12,47]. This new multilocus approach is, to our knowledge, the first attempt to combine the information contained in mtDNA, NRY, X-linked and autosomal markers (see Table 1), which allowed us to test for

the robustness of a sex-specific genetic structure at a local scale. Unraveling the respective influence of migration and drift upon neutral genetic structure is a long-standing quest in population genetics [48,49]. Here, our analysis allowed us to show that differences in sex-specific migration rates may not be the only cause of contrasted male and female differentiation in humans and that, contrary to the conclusion of a number of studies (see Table 1), differences in effective numbers may also play an important role. Indeed, we have demonstrated that sex-specific differences in population structure in patrilineal herders may be the consequence of both higher female effective numbers and female effective dispersal. Our results also illustrate the importance of analyzing human populations at a local scale, rather than global or even continental scale [2,19,21]. The originality of our approach lies in the comparison of identified ethnic groups that differ in well-known social structures and lifestyles. In that respect, our study is among the very few which compare patrilineal vs. bilineal or matrilineal groups (see Table 1), and we believe that it contributes to the growing body of evidence showing that social organization and lifestyle have a strong impact on the distribution of genetic variation in human populations. Moreover, our approach could also be applied on a wide range of animal species with contrasted social organizations. Therefore, we expect our results to stimulate research on the comparison of X-linked and autosomal data to disentangle sex-specific demography.

Methods

DNA Samples

We sampled 10 populations of bilineal agriculturalists and 11 populations of patrilineal herders from West Uzbekistan to East Kyrgyzstan, representing 780 healthy adult men from 5 ethnic groups (Tajiks, Kyrgyz, Karakalpaks, Kazaks, and Turkmen) (see Table 2). The geographic distribution of the samples and information about lifestyle is provided in Figure 1. Also living in Central Asia, Uzbeks are traditionally patrilineal herders too, but they have recently lost their traditional social organization [11], and we therefore chose not to include any sample from this ethnic group for the purpose of this study. We collected ethnologic data prior to sampling, including the recent genealogy of the participants. Using this information, we retained only those individuals that were unrelated for at least two generations back in time. All individuals gave their informed consent for participation in this study. Total genomic DNA was isolated from blood samples by a standard phenol-chloroform extraction [50].

Uniparentally Inherited Markers

The mtDNA first hypervariable segment of the mtDNA control region (HVS-I) was amplified using primers L15987 (5′TCAAATGGGCCTGTCCTTGTA) and H580 (5′TTGAGGAGGTAAGCTACATA) in 18 populations out of 21 (674 individuals, see Table 2). The amplification products were subsequently purified with the EXOSAP standard procedure. The sequence reaction was performed using primers L15925 (5′TAATACACCAGTCTTGTAAAC) and HH23 (5′AATAGGGTGATAGACCTGTG). Sequences from positions 16 024–16 391 were obtained. Eleven Y-linked microsatellite markers (see Table 3) were genotyped in the same individuals, following the protocol described by Parkin et al. [51].

Multi-Locus Markers

27 autosomal and 9 X-linked microsatellite markers (see Table 4) were genotyped in the same individuals. We used the informativeness for assignment index I_n [52] to select subsets of microsatellite markers on the X chromosome and the autosomes from the set of markers used in Rosenberg et al.'s worldwide study [43]. This statistic measures the amount of information that multiallelic markers provide about individual ancestry [52]. This index was calculated among a subset of 14 populations, chosen from the Rosenberg et al.'s dataset [43] to be genetically the closest to the Central Asian populations (Balochi, Brahui, Burusho, Hazara, Pathan, Shindi, Uygur, Han, Mongola, Yakut, Adygei, Russian, Druze and Palestinian). The rationale was to infer the information provided by individual loci about ancestry from this subset of populations, and to extrapolate the results to the populations studied here. For the X chromosome data, we pooled the 'Screening Set10' and 'Screening Set52' from the HGDP-CEPH Human Genome Diversity Cell Line Panel [53] analyzed by Rosenberg et al. [43] which represented a total of 36 microsatellites. We chose 9 markers among the 11 with the highest I_n. For autosomal data, we used the 'Screening Set10', which represented a total of 377 microsatellites, and chose 27 markers among the 30 with the highest I_n. All markers were chosen at a minimum of 2 cM apart from each others [54]. PCR amplifications were performed in a 20 µl final volume composed of 1× Eppendorf buffer, 125 µM each dNTP, 0.5U Eppendorf Taq polymerase, 125 nM of each primer, and 10 ng DNA. The reactions were performed in a Eppendorf Mastercycler with an initial denaturation step at 94°C for 5 min; followed by 36 cycles at 94°C for 30 s, 55°C for 30 s, 72°C for 20 s, and 72°C for 10 min as final extension. Forward primers were fluorescently labeled and reactions were further analyzed by capillary electrophoresis (ABI 310, Applied Biosystems). We used the software package Genemarker (SoftGenetics LLC) to obtain allele sizes from the analysis of PCR products (allele calling).

Statistical Analyses

We calculated the total allelic richness (*AR*) (over all populations), the unbiased estimate of expected heterozygosity H_e [55], the total number of polymorphic sites and F_{ST} for mtDNA using Arlequin version 3.1. [56]. Genetic differentiation among populations for the autosomes, the X and the Y chromosome was measured both per locus and overall loci using Weir and Cockerham's F_{ST} estimator [57], as calculated in GENEPOP 4.0. [58]. The 95% confidence intervals were obtained by bootstrapping over loci [58], using the approximate bootstrap confidence intervals (ABC) method described by DiCiccio and Efron [59]. Isolation by distance (i.e. the correlation between the genetic and the geographic distances) was analyzed by computing the regression of pairwise $F_{ST}/(1-F_{ST})$ estimates between pairs of populations to the natural logarithm of their geographical distances, and rank correlations were tested using the Mantel permutation procedure [60], as implemented in GENEPOP 4.0. [58]. All other statistical tests were performed using the software package R v. 2.2.1 [61].

Sex-Biased Dispersal in the Island Model

Let us consider an infinite island model of population structure [62], with two classes of individuals (males and females), which describes a infinite set of populations with constant and equal sizes that are connected by gene flow. Then the expected values of F_{ST} for uniparentally inherited markers depend on the effective number N_m (resp. N_f) of adult males (resp. females) per population and the migration rate m_m (resp. m_f) of males (resp. females) per generation, as: $F_{ST}^{(mtDNA)} \approx 1/(1+2N_f m_f)$ and $F_{ST}^{(Y)} \approx 1/(1+2N_m m_m)$ (see, e.g., [63]). We can therefore calculate the female-to-male ratio of the effective number of migrants per generation as: $N_f m_f / N_m m_m = \left(1-1/F_{ST}^{(mtDNA)}\right)/\left(1-1/F_{ST}^{(Y)}\right)$.

In this model, we can also compute for the autosomes and the X chromosome the reproductive values for each class (sex), which are interpreted here as the probability that an ancestral gene lineage was in a given class in a distant past [64]. From these, we can obtain the well-known expressions of effective size N_e for autosomal and X-linked genes: $N_e^{(A)} = 8N_f N_m/(N_f + N_m)$ and $N_e^{(X)} = 9N_f N_m/(N_f + 2N_m)$, respectively [45]. Note that N_e is expressed here as a number of gene copies (i.e., twice the effective number of diploid individuals for autosomes). Likewise, the effective migration rate, i.e. the average dispersal rate of an ancestral gene lineage, is given by $m_e^{(A)} = (m_f + m_m)/2$ for autosomal genes, and $m_e^{(X)} = (2m_f + m_m)/3$ for X-linked genes, respectively. Substituting these expressions into the well-known equation: $F_{ST} \approx 1/(1+2N_e m_e)$ [64], we get:

$$F_{ST}^{(A)} \approx \frac{1}{1 + 4\frac{4N_f N_m}{N_f + N_m}\frac{m_f + m_m}{2}}, \quad (5)$$

for autosomal genes, and

$$F_{ST}^{(X)} \approx \frac{1}{1 + 4\frac{9N_f N_m}{2N_f + 4N_m}\frac{2m_f + m_m}{3}}, \quad (6)$$

for X-linked genes.

Evaluation of the Approach through Stochastic Simulations

We performed coalescent simulations, using an algorithm in which coalescence and migration events are considered generation-by-generation until the common ancestor of the whole sample has been reached (see [65]). We simulated a finite island model with 50 demes, each made of $N = N_f + N_m = 500$ diploid individuals, with a migration parameter $m = m_f + m_m = 0.2$. Using these total values for N and m, we then varied the sex-specific parameters to cover the $(N_f/N,\ m_f/m)$ parameter space evenly. Note that the parameter m is the total migration rate, which corresponds to twice the effective migration rate for autosomal markers. Hence, for each set of $(N_f/N,\ m_f/m)$ values, the total number of individuals is 500 (although the number of females may vary from 1 to 499) and the effective migration rate for autosomal markers is $m_e^{(A)} = (m_f + m_m)/2 = 0.1$. We chose these total values for N and m such that, for a ratio $N_f m_f / N_m m_m = 21.6$ (as observed for the herder populations), the distribution of F_{ST} estimates on uniparentally-inherited markers in the simulations were close to the observations: for mtDNA, the 95% highest posterior density interval (see [66], pp. 38–39) for the distribution of F_{ST} estimates in the simulations was [0.007; 0.033] with a mode at 0.014 (estimated value from the real dataset: $F_{ST}^{(mtDNA)} = 0.010$ among the herders) while for the NRY, the 95% highest posterior density interval was [0.088; 0.374] with a mode at 0.187 (estimated value from the real dataset: $F_{ST}^{(Y)} = 0.177$).

Each simulated sample consisted in 330 sampled males from 11 populations (30 males per population), genotyped at 27 autosomal, 9 X-linked markers as well as 10 Y-linked markers and a single mtDNA locus. Each locus was assumed to follow a Generalized Stepwise Model (GSM) [67] with a possible range of 40 contiguous allelic states, except the mtDNA, which was assumed to follow an infinite allele model of mutation. The average mutation rate was 5.10^{-3}, and the mean parameter of the geometric distribution of the mutation step lengths for microsatellites was set to 0.2 [67,68].

Acknowledgments

We thank all the people who volunteered to participate in this study, or who helped us in the field. We are grateful to Sylvain Théry for valuable help in handling geographic data, to Hélène Fréville and Nicolas Perrin for helpful comments on previous versions of this manuscript, as well as to three anonymous reviewers for insightful and constructive comments. We acknowledge the "Service de Systématique Moléculaire" (SSM) at the Museum National d'Histoire Naturelle (MNHN) and the Biological Resource Center of the Foundation Jean Dausset-CEPH for genotyping facilities. Part of this work was carried out by using the resources of the Computational Biology Service Unit from the Museum National d'Histoire Naturelle (MNHN) which was partially funded by Saint Gobain.

Author Contributions

Conceived and designed the experiments: EH RV. Performed the experiments: LS BMC LQM PB MG. Analyzed the data: LS RV. Contributed reagents/materials/analysis tools: BMC TH AA FN MJ EH. Wrote the paper: LS RV. Collected the samples: LS, BMC, EH.

References

1. Disotell TR (1999) Human evolution: sex-specific contributions to genome variation. Curr Biol 9: R29–31.
2. Wilkins JF (2006) Unraveling male and female histories from human genetic data. Curr Opin Genet Dev 16: 611–617.
3. Seielstad MT, Minch E, Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. Nat Genet 20: 278–280.
4. Salem AH, Badr FM, Gaballah MF, Pääbo S (1996) The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. Am J Hum Genet 59: 741–743.
5. Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, et al. (1999) 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.
6. Oota H, Kitano T, Jin F, Yuasa I, Wang L, et al. (2002) Extreme mtDNA homogeneity in continental Asian populations. Am J Phys Anthropol 118: 146–153.
7. Kayser M, Brauer S, Weiss G, Schiefenhovel W, Underhill P, et al. (2003) Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am J Hum Genet 72: 281–302.
8. Malyarchuk B, Derenko M, Grzybowski T, Lunkina A, Czarny J, et al. (2004) Differentiation of mitochondrial DNA and Y chromosomes in Russian populations. Hum Biol 76: 877–900.
9. Nasidze I, Ling EY, Quinque D, Dupanloup I, Cordaux R, et al. (2004) Mitochondrial DNA and Y-chromosome variation in the Caucasus. Ann Hum Genet 68: 205–221.
10. Nasidze I, Quinque D, Ozturk M, Bendukidze N, Stoneking M (2005) MtDNA and Y-chromosome variation in Kurdish groups. Ann Hum Genet 69: 401–412.
11. Chaix R, Quintana-Murci L, Hegay T, Hammer MF, Mobasher Z, et al. (2007) From social to genetic structures in central Asia. Curr Biol 17: 43–48.
12. Wilkins JF, Marlowe FW (2006) Sex-biased migration in humans: what should we expect from genetic data? Bioessays 28: 290–300.
13. Burton ML, Moore CC, Whiting JWM, Romney AK (1996) Regions based on social structure. Curr Anthro 37: 87–123.
14. 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. Nat Genet 29: 20–21.
15. Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, et al. (2004) Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol 21: 1673–1682.
16. Hamilton G, Stoneking M, Excoffier L (2005) Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. Proc Natl Acad Sci U S A 102: 7476–7480.
17. Bolnick DA, Bolnick DI, Smith DG (2006) Asymmetric male and female genetic histories among Native Americans from Eastern North America. Mol Biol Evol 23: 2161–2174.
18. Stoneking M (1998) Women on the move. Nat Genet 20: 219–220.
19. Wilder JA, Kingan SB, Mobasher Z, Pilkington MM, Hammer MF (2004) Global patterns of human mitochondrial DNA and Y-chromosome structure are not influenced by higher migration rates of females versus males. Nat Genet 36: 1122–1125.
20. Ramachandran S, Rosenberg NA, Zhivotovsky LA, Feldman MW (2004) Robustness of the inference of human population structure: a comparison of X-chromosomal and autosomal microsatellites. Hum Genomics 1: 87–97.

21. Kumar V, Langstieh BT, Madhavi KV, Naidu VM, Singh HP, et al. (2006) Global patterns in human mitochondrial DNA and Y-chromosome variation caused by spatial instability of the local cultural processes. PLoS Genet 2: e53.
22. Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, et al. (2001) Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol 18: 1189–1203.
23. Dupanloup I, Pereira L, Bertorelle G, Calafell F, Prata MJ, et al. (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. J Mol Evol 57: 85–97.
24. Wilder JA, Mobasher Z, Hammer MF (2004) Genetic evidence for unequal effective population sizes of human females and males. Mol Biol Evol 21: 2047–2057.
25. Wilder JA, Hammer MF (2007) Extraordinary population structure among the Baining of New Britain. In: Friedlaender JS, ed. Genes, Language, and Culture History in the Southwest Pacific. Oxford, UK: Oxford University Press. pp 199–207.
26. Seielstad M (2000) Asymmetries in the maternal and paternal genetic histories of Colombian populations. Am J Hum Genet 67: 1062–1066.
27. Langergraber KE, Siedel H, Mitani JC, Wrangham RW, Reynolds V, et al. (2007) The genetic signature of sex-biased migration in patrilocal chimpanzees and humans. PLoS ONE 2: e973.
28. Bazin E, Glemin S, Galtier N (2006) Population size does not influence mitochondrial genetic diversity in animals. Science 312: 570–572.
29. Heyer E, Sibert A, Austerlitz F (2005) Cultural transmission of fitness: genes take the fast lane. Trends Genet 21: 234–239.
30. Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, et al. (2003) The genetic legacy of the Mongols. Am J Hum Genet 72: 717–721.
31. Neel JV (1970) Lessons from a "primitive" people. Science 170: 815–822.
32. Blum MG, Heyer E, Francois O, Austerlitz F (2006) Matrilineal fertility inheritance detected in hunter-gatherer populations using the imbalance of gene genealogies. PLoS Genet 2: e122.
33. Helgason A, Hrafnkelsson B, Gulcher JR, Ward R, Stefansson K (2003) A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. Am J Hum Genet 72: 1370–1388.
34. White DR (1988) Rethinking polygyny: co-wives, codes, and cultural systems. Curr Anthro 29: 529–558.
35. Heider KG (1997) Grand valley Dani: peaceful warriors. In: GS, LS, eds. Case studies in cultural anthropology. Forth Worth, Texas: Harcourt Brace College Publishers.
36. Tremblay M, Vezina H (2000) New estimates of intergenerational time intervals for the calculation of age and origins of mutations. Am J Hum Genet 66: 651–658.
37. Fenner JN (2005) Cross-cultural estimation of the human generation interval for use in genetics-based population divergence studies. Am J Phys Anthropol 128: 415–423.
38. Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW (2002) Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. Genetics 161: 447–459.
39. Whitlock MC, McCauley DE (1999) Indirect measures of gene flow and migration: $F_{ST} \neq 1/(4Nm+1)$. Heredity 82: 117–125.
40. Beaumont M, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond 263: 1619–1626.
41. Rousset F (1997) Genetic differentiation and estimation of gene flow from F-statistics under isolation by distance. Genetics 145: 1219–1228.
42. Pool JE, Nielsen R (2007) Population size changes reshape genomic patterns of diversity. Evolution 61: 3001–3006.
43. Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, et al. (2002) Genetic structure of human populations. Science 298: 2381–2385.
44. Reynolds J, Weir BS, Cockerham CC (1983) Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. Genetics 105: 767–779.
45. Wright S (1939) Statistical genetics in relation to evolution. Actualités scientifiques et industrielles 802 Exposés de Biométrie et de Statistique Biologique XIII. Paris: Hermann et Cie.
46. Tamisier JC (1998) Dictionnaire des peuples. Sociétés d'Afrique, d'Amérique, d'Asie et d'Océanie. Paris: Larousse-Bordas.
47. Balaresque P, Jobling MA (2007) Human populations: houses for spouses. Curr Biol 17: R14–16.
48. Lawson-Handley LJ, Perrin N (2007) Advance in our understanding of mammalian sex-biased dispersal. Molecular Ecology 16: 1559–1578.
49. Hurles ME, Jobling MA (2001) Haploid chromosomes in molecular ecology: lessons from the human Y. Mol Ecol 10: 1599–1613.
50. Maniatis T, Fritsh EF, S J (1982) Molecular cloning. A laboratory manual. New York: Cold Spring Laboratory.
51. Parkin EJ, Kraayenbrink T, GLvD, Tshering K, de Knijff P, et al. (2006) 26-Locus Y-STR typing in a Bhutanese population sample. Forensic Science International 161: 1–7.
52. Rosenberg NA, Li LM, Ward R, Pritchard JK (2003) Informativeness of genetic markers for inference of ancestry. Am J Hum Genet 73: 1402–1422.
53. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296: 261–262.
54. Wilson JF, Goldstein DB (2000) Consistent long-range linkage disequilibrium generated by admixture in a Bantu-Semitic hybrid population. Am J Hum Genet 67: 926–935.
55. Nei M (1978) Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics 89: 583–590.
56. Excoffier L, Laval LG, Schneider S (2005) Arlequin ver. 3.0: An integrated software package for population genetics data analysis. Evol Bioinfo Online 1: 47–50.
57. Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. Evolution 38: 1358–1370.
58. Rousset F (2008) Genepop'007: a complete re-implementation of the genepop software for Windows and Linux. Mol Ecol Res 8: 103–106.
59. DiCiccio TJ, Efron B (1996) Bootstrap confidence intervals. Statistical Science 11: 189–228.
60. Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27: 209–220.
61. R Development Core Team (2007) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing.
62. Wright S (1931) Evolution in mendelian populations. Genetics 16: 97–159.
63. Hedrick PW (2007) Sex: differences in mutation, recombination, selection, gene flow, and genetic drift. Evolution 61: 2750–2771.
64. Rousset F (2004) Genetic Structure and Selection in Subdivided Populations. Princeton, New Jersey: Princeton University Press.
65. Leblois R, Estoup A, Rousset F (2003) Influence of mutational and sampling factors on the estimation of demographic parameters in a "continuous" population under isolation by distance. Mol Biol Evol 20: 491–502.
66. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian Data Analysis. Second Edition. New York: Chapman & Hall/CRC.
67. Estoup A, Jarne P, Cornuet JM (2002) Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol Ecol 11: 1591–1604.
68. Dib C, Faure S, Fizames C, Samson D, Drouot N, et al. (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380: 152–154.
69. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131: 479–491.
70. Beerli P, Felsenstein J (2001) Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc Natl Acad Sci U S A 98: 4563–4568.

- **Annexe 5** : **Ségurel L**., Lafosse S., Heyer E. & Vitalis R. Frequency of the AGT Pro11Leu polymorphism in humans: does diet matter? *Ann. Hum. Genet.* 2010 Jan; 74:1, 57-64

aptara | AHG | ahg_549 | Dispatch: September 17, 2009 | CE:
Journal | MSP No. | No. of pages: 8 | PE:

doi: 10.1111/j.1469-1809.2009.00549.x

Frequency of the AGT Pro11Leu Polymorphism in Humans: does Diet Matter?

Laure Ségurel*, Sophie Lafosse, Evelyne Heyer and Renaud Vitalis‡

Muséum National d'Histoire Naturelle–Centre National de la Recherche Scientifique–Université Paris Diderot-Paris 7, UMR 7206 Éco-Anthropologie et Ethnobiologie, CP 139, 57 rue Cuvier, 75231 Paris Cedex 05, France

Summary

The Pro11Leu substitution in the *AGXT* gene, which causes primary hyperoxaluria type 1, is found with high frequency in some human populations (e.g., 5–20% in Caucasians). It has been suggested that this detrimental mutation could have been positively selected in populations with a meat-rich diet. In order to test this hypothesis, we investigated the occurrence of Pro11Leu in both herder and agriculturalist populations from Central Asia. We found a lower frequency of this detrimental mutation in herders, whose diet is more meat-rich, as compared to agriculturalists, which therefore challenges the universality of the previous claim. Furthermore, when combining our original data with previously published results, we could show that the worldwide genetic differentiation measured at the Pro11Leu polymorphism does not depart from neutrality. Hence, the distribution of the variation observed in the *AGXT* gene could be due to demographic history, rather than local adaptation to diet.

Keywords: adaptation, *AGXT*, AMOVA, diet, *F*-statistics, primary hyperoxaluria type 1

Introduction

Primary hyperoxaluria type 1 (PH1, MIM 259900) is a lethal autosomal recessive disease caused by the deficiency of the alanine:glyoxylate aminotransferase (AGT), a liver-specific enzyme usually targeted in the peroxysomes in humans (Danpure & Jennings, 1986). The AGT deficiency leads to the formation of insoluble calcium salts, resulting in progressive renal failure. It has been estimated that, before the introduction of modern therapies, 80% of PH1 patients died from renal failure before the age of 20 years (Williams & Smith, 1983). In more than a third of the patients, the disease is not due to the absence of AGT, but rather to its de-localisation from the peroxysomes to the mitochondria (Danpure et al., 1990). This mistargeting is due to the synergistic effect of two non-synonymous recessive mutations in the human *AGXT* gene that encodes AGT: the Pro11Leu and Gly170Arg substitutions (Purdue et al., 1990). Pro11Leu alone, in a homozygous state, is responsible for the de-localisation of 5% of the protein to the mitochondria (Purdue et al., 1990), although this has not been associated with any clinical sequela. On the contrary, the association of Pro11Leu with the most common PH1 mutation Gly170Arg results in the mistargeting of at least 90% of the AGT to the mitochondria, resulting in primary hyperoxaluria type 1 (Danpure et al., 1989). From analysis of mutant constructs in *Escherichia coli*, it seems likely that Gly170Arg alone has no pathological phenotype (Lumb & Danpure, 2000), yet it has never been found so far in humans without the Pro11Leu substitution. Pro11Leu also acts synergistically with other *AGXT* mutations which, like Gly170Arg, seem innocuous alone (Lumb & Danpure, 2000).

Although the Pro11Leu substitution is therefore likely to be detrimental, by sensitising AGT to the effect of a wide range of PH1 mutations, it is found at relatively high frequency in some populations, e.g., 5–20% in Caucasians (Danpure et al., 1994a, 1994b). How can this paradox be solved? It has been shown that, in mammals, the intracellular location of AGT depends upon natural diet (Danpure et al., 1990). AGT normally converts glyoxylate to glycine (Danpure & Jennings, 1986). However, the sites where the conversion occurs, as well as the precursors of glyoxylate, are both related to natural diet

*Corresponding author: Laure Ségurel, Muséum National d'Histoire Naturelle–Centre National de la Recherche Scientifique– Université Paris 7, UMR 7206 Éco-Anthropologie et Ethnobiologie, CP 139, 57 rue Cuvier, 75231 Paris Cedex 05, France. Tel: +33 (0)1 40 79 81 65; Fax: +33 (0)1 40 79 38 91; E-mail: lsegurel@mnhn.fr
‡Present address: Institut National de la Recherche Agronomique (INRA), UMR CBGP (INRA–IRD–CIRAD–Montpellier SupAgro), Campus International de Baillarguet, CS 30016, 34988 Montferrier-sur-Lez, France

(Danpure, 1997). In herbivores, the main precursor of glyoxylate is thought to be glycolate, which is converted to glyoxylate in the peroxysomes (Noguchi, 1987). In carnivores, the major precursor is hydroxyproline, which is converted to glyoxylate in the mitochondria (Takayama et al., 2003). Therefore, the fact that in herbivores, AGT is principally found in peroxysomes, while in carnivores it is found in mitochondria, might reflect a more effective glyoxylate detoxification by AGT at the site of glyoxylate synthesis (Danpure, 1997). In humans, AGT is usually targeted in the peroxysomes (Cooper et al., 1988), which might reflect the ancestral herbivorous diet of hominoids (Holbrook et al., 2000). Therefore, it might well be that the redirection of a small proportion of AGT from peroxysomes to mitochondria in humans leads to a sub-cellular distribution of AGT that is more compatible with an omnivorous (or carnivorous), rather than herbivorous, diet (Danpure, 1997).

Following this line of ideas, it has been suggested that the redirection of 5% of the AGT activity from peroxysomes to mitochondria, due to the Pro11Leu substitution, might give a selective advantage to individuals who have a larger proportion of meat in their diet (Danpure, 1997), although the precise link between the functional consequences of the human non-synonymous substitution Pro11Leu in *AGXT* and diet remains hypothetical. Interestingly, if this mutation is advantageous in populations with a meat-rich diet, it should have been selected for during the ice age periods, when our hunter-gatherer ancestors were highly dependent on meat-based resources. Hence, we should now observe a decrease of selective constraints in populations with a lower proportion of meat in their diet, like agriculturalists, as compared to hunter-gatherers and herders.

The prediction that the Pro11Leu detrimental substitution has been, and/or still is, advantageous for individuals who have a meat-rich diet has been tested by Caldwell et al. (2004). To do so, they first determined the frequency of the Pro11Leu substitution C → T in 11 human populations. They found the highest frequency of the derived T allele in the Saami (27.9%, see also Kozlov et al., 2008), a population of herders known to have a meat-rich ancestral diet (Haglin, 1991, 1999), and the lowest frequency in the Chinese (2.3%) and Indian populations (3.0%), which are supposed to have, respectively, mixed and vegetarian ancestral diets. They concluded that the frequency of the T allele in human populations has been shaped by dietary selection pressure. Yet, this apparent correlation between the Pro11Leu T allele frequency and the ancestral diet is not robust to the inclusion of, e.g., the Mongol sample, which shows a low frequency of the T allele (6.9%), although the diet of these traditionally nomadic herders mainly consists of meat, milk and dairy products (see Hruschka & Brandon, 2004). Caldwell et al. (2004) also compared the genetic differentiation (as measured by the parameter F_{ST}) at the Pro11Leu polymorphism for Nigerians, Chinese and Saami to that observed at 33,487 presumably neutral SNPs typed in related populations (African Americans, East Asians and European Americans). The rationale was to test whether the genetic differentiation at the Pro11Leu polymorphism exceeded that of the rest of the genome, which could be interpreted as a signature of divergent selection. They found that the F_{ST} estimate at the Pro11Leu polymorphism lay in the tail of the distribution of genome-wide F_{ST} estimates, although not significantly ($p = 0.074$ and $p = 0.266$, respectively, for the Saami vs. Chinese and Saami vs. Nigerians comparisons).

Caldwell et al.'s (2004) results are therefore tentative, and require further investigation. For example, it would be interesting to test whether the proportion of genetic variance explained by diet at the Pro11Leu polymorphism is significantly different from that observed in the rest of the genome. To address this question, hierarchical analysis of molecular variance (AMOVA, Excoffier et al., 1992) can be used to apportion the total observed genetic variation among nested hierarchical levels, with populations pooled into groups according to their ancestral diet. In this study, we aim at testing Danpure's (1997) hypothesis that the Pro11Leu replacement in *AGXT* is beneficial for individuals who have a predominantly meat-based diet. To that end, we provide new data collected from seven traditionally nomadic herder populations and four sedentary agriculturalist populations from Central Asia, that differ by their lifestyle and ancestral diet. In the following, as in Caldwell et al. (2004), we use the current lifestyle of populations as a proxy to the ancestral lifestyle. We do not necessarily assume a strict continuity in the subsistence pattern from the ice-age hunter-gatherer populations to the extant herder populations. We rather assume that, since the Neolithic transition, herder populations relied predominantly on meat and dairy products, as opposed to agriculturalist populations who had a mixed or more balanced diet. Analyzing these original data together with published data sets, we specifically ask (*i*) whether the Pro11Leu allele frequency correlates with ancestral diet, (*ii*) whether the genetic differentiation at the *AGXT* gene departs from the genome wide differentiation as a consequence of selection, and (*iii*) whether the proportion of genetic variance explained by diet at this gene is larger than that of the rest of the genome.

Materials and Methods

DNA Samples

We sampled seven populations of traditionally nomadic herders and four populations of sedentary agriculturalists from West Uzbekistan to East Kyrgyzstan, representing respectively 214 and 90 healthy adult men from five ethnic groups (Kyrgyz,

Table 1 Samples Description

Sampled populations (area)	Acronym	Location	Long.	Lat.	*n*	*p* (%)
Tajiks (Ferghana)	TJR	Tajikistan / Kyrgyzstan border	40.36	71.28	17	17.6
Tajiks (Gharm)	TJE	Northern Tajikistan	39.12	70.67	23	19.6
Tajiks (Penjinkent)	TDU	Uzbekistan / Tajikistan border	39.44	68.26	24	20.8
Tajiks (Yagnobs from Douchambe)	TJY	Western Tajikistan	38.57	68.78	26	26.9
Karakalpaks (Qongrat from Karakalpakia)	KKK	Western Uzbekistan	43.77	59.02	23	15.2
Kazaks (Karakalpakia)	KAZ	Western Uzbekistan	43.04	58.84	30	1.7
Kazaks (Bukara)	LKZ	Southern Uzbekistan	40.08	63.56	49	9.2
Kyrgyz (Andijan)	KRA	Tajikistan / Kyrgyzstan border	40.77	72.31	32	14.1
Kyrgyz (Narin)	KRG	Middle Kyrgyzstan	41.6	75.8	20	7.5
Kyrgyz (Narin)	KRB	Middle Kyrgyzstan	41.25	76	26	11.5
Turkmen (Karakalpakia)	TUR	Western Uzbekistan	41.55	60.63	34	13.2
Sichuan Chinese	*CHI*				*86*	*2.3*
Armenians	*ARM*				*73*	*19.2*
North Welsh	*NOR*				*82*	*14.6*
Nigerians	*NIG*				*62*	*8.9*
Mongols	*MON*				*80*	*6.9*

Sedentary agriculturalist populations are in white; traditionally nomadic herders are in gray. The five latter populations (in italics) correspond to the populations studied by Caldwell et al. (2004) for the Pro11Leu polymorphism, included in the present analysis. Long., longitude; Lat., latitude; *n*, sample size; *p*, frequency of the derived T allele of the Pro11Leu polymorphism in *AGXT*.

Karakalpaks, Kazaks, Turkmen and Tajiks; see Table 1). We collected ethnologic data prior to sampling, including the recent genealogy of the participants. Using this information, we retained only those individuals that were unrelated for at least two generations back in time. All individuals gave their informed consent for participation in this study. Total genomic DNA was isolated from blood samples or from saliva using a standard phenol-chloroform extraction protocol (Maniatis et al., 1982). We also used the data from 5 populations genotyped by Caldwell et al. (2004) at the Pro11Leu polymorphism: Sichuan Chinese, Mongols, Armenians, North Welsh and Nigerians, for which neutral data were also available on closely related populations (Han Chinese, Mongols, Adygeis, Orcadians and Yorubas, respectively).

Genotyping

The Pro11Leu substitution (nucleotide C in chimpanzees, C or T in humans) was detected by PCR-RFLP. PCR amplifications were performed in a 10 μL final volume containing 1X Eppendorf buffer, 125 μM of each dNTP, 0.5 U of Eppendorf Taq polymerase, 125 nM of forward (5′-GCACAGATAAGCTTCAGGGA-3′) and reverse (5′-CTTGAAGGATGGATCCAGGG-3′) primers, and 10 ng of DNA. The reactions were performed in an Eppendorf Mastercycler with an initial denaturation step at 94°C for 5 min, followed by 40 cycles at 94°C for 30 s, 59°C for 1 min and 72°C for 30 s, with a final extension step at 72°C for 10 min. PCR products were then incubated overnight at 37°C, with 10 units of the restriction endonuclease *Sty*I, 0.01 μg/μL acetylated BSA and NEB Buffer 3, in a 15 μL final volume. Digestion products were then electrophoresed at 4°C overnight on a 2% agarose gel stained with ethidium bromide.

There are two main forms of *AGXT* across human populations, the "major" allele (Ma) and the "minor" allele (Mi). Among other differences between the two alleles, the Mi allele presents a 74-bp duplication in intron 1 of the gene. In Caucasian populations, the Pro11Leu polymorphism is in strong linkage disequilibrium with the Mi allele, and it has been previously suggested that the presence of the 74-bp duplication would be indicative of the Pro11Leu polymorphism (Purdue et al., 1991). The Pro11Leu substitution C → T induces a *Sty*I restriction site loss. Therefore, after digestion of the PCR products with *Sty*I, one expects to find a 512-bp fragment for the Ma allele (which contains neither polymorphism), and a 619-bp fragment for the Mi allele (which contains both the 74-bp duplication and the Pro11Leu substitution) (Danpure & Rumsby, 1996). However, Coulter-Mackie et al. (2003) also found a 586-bp fragment in African populations, which they referred to as the Mi^A allele. This fragment contains the 74-bp duplication but not the Pro11Leu substitution. Importantly, Coulter-Mackie et al. (2003) never found Pro11Leu on the background of the Ma allele, i.e. without the duplication. Therefore, after digestion with *Sty*I, we expected to observe the three kinds of fragments, potentially: the Ma allele without Pro11Leu (512 bp), the Mi allele with Pro11Leu (619 bp), or the Mi^A allele (586 bp) without Pro11Leu.

Data Analysis

We aimed to evaluate whether the genetic differentiation observed at the Pro11Leu polymorphism departed from the genome-wide differentiation expected at neutrality among 16 populations whose ancestral diet was known: the 16 populations consisted of the 11 Central Asian populations described

above, and 5 populations genotyped by Caldwell et al. (2004) at the Pro11Leu polymorphism: Sichuan Chinese, Mongols, Armenians, North Welsh and Nigerians. In order to estimate the expected neutral differentiation, we used the genotypic data at 27 autosomal short tandem repeat (STR) markers from Ségurel et al. (2008) for the 11 Central Asian populations, as well as data from the HGDP-CEPH Human Genome Diversity Cell Line Panel dataset (Rosenberg et al., 2002) genotyped at the same STR markers as were used for the Han Chinese, the Mongols, the Adygeis, the Orcadians and the Yorubas. Although these latter populations are not strictly those for which the Pro11Leu frequency data were available, they are probably very closely related.

To identify the signatures of natural selection we used a modified version of the software package Dfdist (Beaumont & Nichols, 1996). This method is based on the principle that genetic differentiation among populations is expected to be higher for loci under divergent selection than for the rest of the genome. The rationale here was to compute F_{ST} and the overall heterozygosity of the pooled sample for the Pro11Leu substitution, and then to compare this value to a neutral distribution of F_{ST} conditional on heterozygosity, generated by means of coalescent simulations in a symmetrical island migration model at migration-drift equilibrium (Wright, 1951), given the observed level of differentiation measured at the 27 autosomal STRs ($F_{ST} = 0.019$). Since the software package Dfdist was specifically designed for the analysis of bi-allelic, dominant markers (see, e.g., Bonin et al., 2006), we modified it in order to simulate co-dominant, bi-allelic data (code available upon request). 500,000 coalescent simulations were performed with a 50-demes island model and $\theta = 2nN\mu = 0.2$ (where $n = 50$ is the number of demes of size N, and μ is the mutation rate). This particular θ value would correspond, e.g., to $N = 100$ and $\mu = 2 \times 10^{-5}$. Because this choice of parameter value might seem somewhat arbitrary, we checked that the distribution of F_{ST} conditional on heterozygosity was robust to alternative values ($\theta = 0.02$ and $\theta = 2.0$), particularly in the range of heterozygosity observed at *AXGT* ($H_e = 0.207$) (data not shown).

The maximum frequency of the most common allele allowed was set to 0.99, for the simulated datasets. To obtain a close approximation of the expected joint distribution of F_{ST} and heterozygosity, we followed the algorithm detailed in the Appendix of Vitalis et al. (2001). This algorithm takes the pairs of (F_{ST}, H_e) estimates from the simulations to generate a 2D histogram, which is a close approximation to the bivariate probability distribution. We further used the Averaged Shifted Histogram (ASH) algorithm (Scott, 2002) to smooth the probability distribution, and to provide a continuous "high probability region" that contains 90%, 95%, or 99% of the total probability distribution. Finally, we derived an empirical *p*-value for the joint (F_{ST}, H_e) estimate at the Pro11Leu polymorphism observed in the real dataset. To do so, we calculated the relative frequency of this observation over the full set of simulated data, i.e. the height of the 2D histogram's cell that corresponds to the Pro11Leu (F_{ST}, H_e) estimate. The empirical *p*-value was then calculated as the proportion of the bivariate probability distribution of the full set of simulations, which is less probable than the Pro11Leu (F_{ST}, H_e) estimate. A so-obtained *p*-value equal to, say, 0.05 indicates that 95% of the simulated data are more probable than the observed Pro11Leu (F_{ST}, H_e) estimate (or, conversely, that only 5% of the simulated data show as "extreme" (F_{ST}, H_e) estimates as at Pro11Leu). In that case, the Pro11Leu (F_{ST}, H_e) estimate would lie at the limit of the 95% "high probability region".

Because we also aimed to test whether the proportion of genetic variance explained by diet was larger than that of the rest of the genome, hierarchical analyses of molecular variance (AMOVAs) were performed, using diet as an explaining factor. Sedentary agriculturalist populations from Central Asia (see Table 1), along with the Sichuan Chinese, the Armenians, the North Welsh and the Nigerians were considered as having a meat-poor diet, and the traditionally nomadic herders from Central Asia (see Table 1), along with the Mongols, as having a meat-rich diet. We used exactly the same simulated data as for the Dfdist analysis previously described, except that we considered diet as a fixed factor in the AMOVA. Hierarchical AMOVAs were all performed with the statistical software R (R Development Core Team, 2007) using the ade4 package (Dray & Dufour, 2007) to calculate the genetic variance components. For each simulated dataset, we calculated the parameter F_{CT}, which measures the part of the variation accounting for differences between diet groups. We then estimated the joint distribution of F_{CT} conditional upon heterozygosity, and proceeded in much the same way as for the joint distribution of F_{ST} and H_e.

Results and Discussion

AGXT Genetic Diversity

We found only two alleles at the Pro11Leu polymorphism in *AGXT* in the Central Asian samples: the major allele without Pro11Leu (Ma, 512 bp) and the minor allele with Pro11Leu (Mi, 619 bp), which differ by the presence of a 74-bp duplication in intron 1 (see the Materials and Methods section). We did not observe a 586-bp fragment, which would correspond to an allele with the 74-bp duplication but without Pro11Leu (Mi^A allele). We can therefore conclude that the Pro11Leu substitution is strongly associated to the 74-bp duplication in Central Asian populations, as it was found in Caucasian populations (Purdue et al., 1991), but not in South African populations (Coulter-Mackie et al., 2003).

Does the Pro11Leu Allele Frequency Correlate with Life-Style?

Despite within-group heterogeneity (see Table 1), the frequency of the derived T allele of the Pro11Leu polymorphism is significantly higher among agriculturalists (17.6–26.9%) than among herders (1.7 – 15.2%) in Central Asia

(Wilcoxon signed-rank test, p-value = 0.003). This contradicts Danpure's (1997) well accepted hypothesis, which claims that this presumably detrimental mutation in *AGXT*, involved in primary hyperoxaluria type 1 disease, might be positively selected in human populations with a more carnivorous diet, such as herders. Yet our results are not the only ones to challenge Danpure's (1997) hypothesis. It is indeed worth noting that Caldwell et al. (2004) also found a low frequency of the T allele (6.9%) in their Mongol sample, although the diet of this population is known to be strongly meat-based (Hruschka & Brandon, 2004), and a high frequency of the T allele (19.7%) in their Norwegian sample, although Norway is a farming population with a mixed ancestral diet.

Is Genetic Differentiation Stronger at AGXT as Compared to the Rest of the Genome?

We used the overall F_{ST} estimate among 16 worldwide populations, calculated from the 27 presumably neutral STR loci (F_{ST} = 0.019), to perform the coalescent simulations and generate the distribution of F_{ST} conditional upon heterozygosity expected for neutral, bi-allelic co-dominant markers. As depicted in Figure 1a, we found that the Pro11Leu polymorphism (F_{ST} = 0.025) does not depart significantly from the neutral expectation, given the STR data (p-value = 0.214). This suggests that demography mainly shapes the observed variation at the Pro11Leu polymorphism. It might well be that the present sampling scheme better fits a hierarchical island model (Slatkin & Voelm, 1991, Vigouroux & Couvet, 2000), with some populations exchanging more migrants within ethnic groups than between ethnic groups. This is the case, e.g., of Central Asian herder populations, which are less differentiated on average, as compared to agriculturalist or worldwide populations (see Table 2). Yet, Excoffier et al. (2009) have shown that taking the hierarchical structure of populations into account results in a much wider distribution of F_{ST} vs. H_e than expected with a strict island model. Hence, ignoring the hierarchical structure, when it exists, tends to generate an excess of false positive outlier loci. Here, since the Pro11Leu polymorphism does not depart from neutrality when we assume a simple island model, we also expect Pro11Leu to lie within the F_{ST} vs. H_e distribution, when accounting for a more complex hierarchical structure.

Interestingly, this result somehow confirms that of Caldwell et al. (2004) who also found that the genetic differentiation measured at Pro11Leu was not significantly different from the genome-wide (presumably neutral) differentiation. Although they claimed that the genetic differentiation at Pro11Leu lay in the tail of the presumably neutral distribution, they based most of their analyses on the pairwise comparisons between the Saami population and both the Chinese and the Nigerian populations. Yet, it is worth stressing that the Saami population is commonly considered as a genetic isolate in Europe (Tambets et al., 2004) and is therefore not easily comparable to the HGDP-CEPH sample. This is why we discarded this sample from the present study. As it is very likely that the Saami population underwent strong genetic drift, demography could indeed account for the high frequency of the Pro11Leu T allele observed in this population.

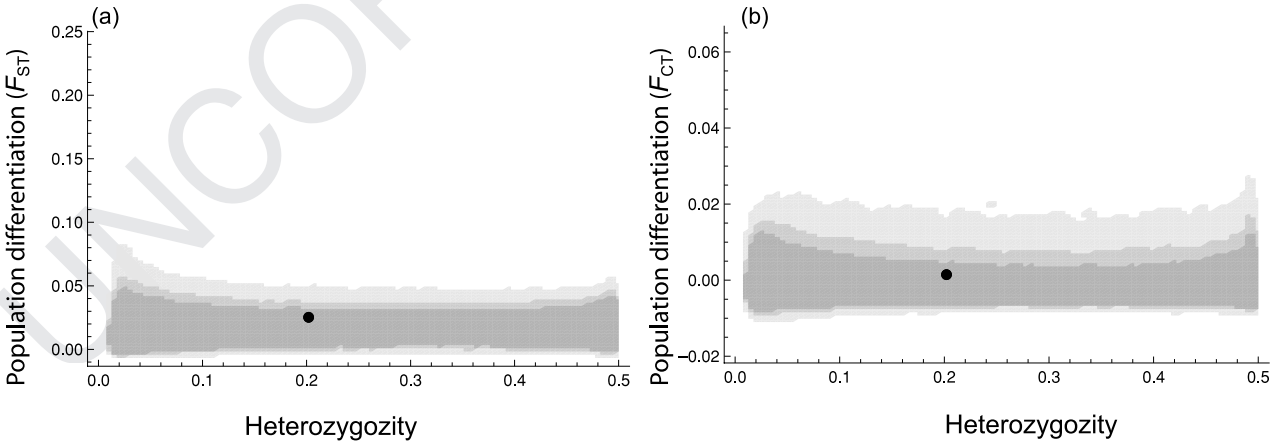


Figure 1 Joint distribution of F_{ST} (a) and F_{CT} (b) vs. the overall heterozygosity of the pooled sample (H_e) expected under neutrality, across 16 worldwide populations. The 90%, 95% and 99% confidence regions of the null distribution are shown from darker to lighter grey, respectively. The smoothed density was obtained using the Average Shifted Histogram (ASH) algorithm (Scott, 2002) with smoothing parameter $m = 2$. The black dots represent the observed measures of differentiation and heterozygosity at the Pro11Leu polymorphism.

Table 2 Pairwise F_{ST} between the studied populations, for 27 autosomal STRs

	Han	Mongola	Orcadian	Adygei	Yoruba	KAZ	KKK	KRA	KRB	KRG	LKZ	TUR	TDU	TJE	TJR	TJY
Han		-	***	***	***	***	***	***	***	***	***	***	***	***	***	***
Mongola	−0.004		***	***	***	*	***	***	-	***	*	***	***	***	***	***
Orcadian	0.115	0.100		-	***	***	***	***	***	***	***	***	***	***	***	***
Adygei	0.090	0.081	0.009		***	***	***	***	***	***	***	***	***	***	***	***
Yoruba	0.132	0.119	0.065	0.059		***	***	***	***	***	***	***	***	***	***	***
KAZ	0.020	0.010	0.054	0.037	0.078		***	***	**	-	-	***	***	***	***	***
KKK	0.026	0.019	0.043	0.024	0.077	0.005		***	**	***	-	***	***	***	***	***
KRA	0.018	0.012	0.062	0.044	0.085	0.004	0.007		**	***	***	***	***	***	***	***
KRB	0.012	0.005	0.063	0.041	0.087	0.002	0.004	0.002		***	-	***	***	***	***	***
KRG	0.028	0.021	0.058	0.041	0.083	0.003	0.010	0.011	0.008		*	***	***	***	***	***
LKZ	0.022	0.013	0.043	0.027	0.075	0.001	0.000	0.004	0.002	0.004		***	***	***	***	***
TUR	0.051	0.042	0.022	0.012	0.057	0.016	0.012	0.020	0.021	0.021	0.010		***	***	***	***
TDU	0.063	0.054	0.029	0.012	0.064	0.022	0.020	0.030	0.029	0.025	0.016	0.008		***	***	***
TJE	0.068	0.058	0.025	0.011	0.060	0.022	0.020	0.031	0.030	0.027	0.017	0.011	0.016		***	***
TJR	0.053	0.046	0.031	0.012	0.062	0.017	0.013	0.021	0.022	0.018	0.010	0.011	0.012	0.009		***
TJY	0.093	0.082	0.029	0.014	0.060	0.043	0.033	0.048	0.053	0.048	0.033	0.024	0.022	0.022	0.023	

Refer to Table 1 for population acronyms. F_{ST} estimates (below the diagonal) were calculated with the software package Genepop v. 4.0 (Rousset, 2008). Results from exact tests of differentiation are given above the diagonal. -: non significant, *: $p < 0.05$, **: $p < 0.01$ and ***: $p < 0.001$.

Is Genetic Variance Due to Diet Larger for AGXT as Compared to the Rest of the Genome?

To answer this question, we calculated from each simulated dataset the parameter F_{CT}, which measures the part of the variation accounting for differences between diet groups. The results are depicted in Figure 1b: the part of the variation accounting for differences between diet groups, measured at the Pro11Leu polymorphism ($F_{CT} = 0.002$), does not depart significantly from neutral expectation, given the STR data (p-value = 0.187).

Alternative Evolutionary Hypotheses

Even though our results show that the differentiation at the Pro11Leu polymorphism does not significantly depart from neutral expectation, we cannot rule out that the observed pattern has been shaped by two opposite selective forces. On the one hand, the Pro11Leu mutation has a presumably deleterious effect, due to the PH1 disease, while on the other hand this mutation is argued to be advantageous in association with a meat-rich diet. Both forces could therefore compensate each other and wipe out a signature of selection, at least among herder populations. This, however, barely explains the extreme differences in allele frequency observed between, e.g., the Saami and the Mongol populations (Caldwell et al., 2004), which both share a meat-rich diet.

As the Neolithic transition to agriculture took place 12,000–9,000 BC in Asia and 7,000–5,000 BC in Europe (and probably even later in Northern Europe), the change in diet may be so recent that the genetic signature of an adaptation to an ancestral predominantly meat-based diet still remains. This may explain why the Pro11Leu is found at such high frequency in, e.g., the Norwegian population. It might also be that in this latter agriculturalist population, individuals have more meat in their diet, as compared to other agriculturalist populations. Yet, neither hypothesis provides a satisfactory interpretation of the low frequency of Pro11Leu observed in Mongolian and Central Asian herders.

One alternative interpretation could be that the associated causative mutation (Gly170Arg) is not present at the same frequency in European and Asian populations, therefore allowing the Pro11Leu mutation to reach higher frequencies in Europe. It would therefore be interesting to collect worldwide frequency data for Gly170Arg. It might also be that another, yet unknown, advantageous mutation in *AGXT* in Mongolian and Central Asian populations provides alternative means to detoxify glyoxylate from meat, in the absence of the Pro11Leu substitution. Such a convergent adaptation has been described for lactase persistency, with two different mutations in the *LCT* gene providing the same phenotype in Africa and Europe (Tishkoff et al., 2007). Yet it is not clear how this hypothesis would account for the high frequency of the Pro11Leu T allele observed in the agriculturalist populations of Central Asia. Furthermore, because of its implication in PH1 disease, the *AGXT* gene has been extensively investigated (Danpure, 2004), and it seems unlikely that causative amino acid substitutions have been overlooked.

In conclusion, our results show that the differentiation at the Pro11Leu polymorphism does not depart from neutral expectation, and do not support the general idea that positive selection has been, or still is, shaping the genetic variation at the *AGXT* locus. Yet, importantly, we cannot exclude some complex scenarios that would involve a combination of the previous hypotheses: for example, the presence of another advantageous allele in Mongolian and Central Asian herder populations (distinct from Pro11Leu), together with a remnant signature of adaptation of to a meat-rich diet among agriculturalist populations in Europe, might explain the pattern observed. There is no doubt that sequence data in the *AGXT* locus would provide powerful tools to test such complex scenarios, by allowing the use of statistical tests of neutrality within populations, e.g., based on the frequency spectrum of DNA polymorphic sites. However, our approach, which aimed to evaluate whether worldwide genetic variation at the Pro11Leu substitution was better explained by demographic history or by adaptation to diet, has been shown to be robust to the vagaries of demographic history (Beaumont, 2005), even in the case of partial selection (see Beaumont & Balding, 2004). Finally, our conclusion that the observed allele frequency differences at the Pro11Leu substitution may result from demographic history rather than selection, supports the recent claim that the observed large allele frequency differences at a number of genetic polymorphisms result from neutral processes rather than from local adaptation (Hofer et al., 2009, Coop et al., 2009).

Acknowledgements

We thank Frédéric Austerlitz for helpful comments on this manuscript. This work was supported by the Centre National de la Recherche Scientifique (CNRS) ATIP programme (to EH), by the CNRS interdisciplinary programme ''Origines de l'Homme du Langage et des Langues'' (OHLL), by the European Science Foundation (ESF) EUROCORES programme ''The Origin of Man, Language and Languages'' (OMLL), by the MNHN department ''Homme, Nature, Société'' and by the ANR-funded programme NUTGENEVOL. We also thank the ''Fondation pour la Recherche Médicale'' (FRM) for financial support. LS was financed by the French Ministry of Higher Education and Research.

References

Beaumont, M. A. (2005) Adaptation and speciation: what can F_{ST} tell us? *Trends Ecol Evol* **20**, 435–440.

Beaumont, M. A. & Balding, D. J. (2004) Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* **13**, 969–980.

Beaumont, M. A. & Nichols, R. A. (1996) Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc Lon* **263**, 1619–1626.

Bonin, A., Taberlet, P., Miaud, C. & Pompanon, F. (2006) Explorative genome scan to detect loci for adaptation along a gradient of altitude in the common frog (*Rana temporaria*). *Mol Biol Evol* **23**, 773–783.

Caldwell, E. F., Mayor, L. R., Thomas, M. G. & Danpure, C. J. (2004) Diet and the frequency of the alanine:glyoxylate aminotransferase Pro11Leu polymorphism in different human populations. *Hum Genet* **115**, 504–509.

Coop, G., Pickrell, J. K., Novembre, J., Kudaravalli, S., Li, J., Absher, D., Myers, R. M., Cavalli-Sforza, L. L., Feldman, M. W. & Pritchard, J. K. (2009) The role of geography in human adaptation. *PLoS Genet* **5**, e1000500.

Cooper, P. J., Danpure, C. J., Wise, P. J. & Guttridge, K. M. (1988) Immunocytochemical localization of human hepatic alanine: glyoxylate aminotransferase in control subjects and patients with primary hyperoxaluria type 1. *J Histochem Cytochem* **36**, 1285–1294.

Coulter-Mackie, M. B., Tung, A., Henderson, H. E., Toone, J. R. & Applegarth, D. A. (2003) The AGT gene in Africa: a distinctive minor allele haplotype, a polymorphism (V326I), and a novel PH1 mutation (A112D) in Black Africans. *Mol Genet Metab* **78**, 44–50.

Danpure, C. J. (1997) Variable peroxisomal and mitochondrial targeting of alanine: glyoxylate aminotransferase in mammalian evolution and disease. *Bioessays* **19**, 317–326.

Danpure, C. J. (2004) Molecular aetiology of primary hyperoxaluria type 1. *Nephron Exp Nephrol* **98**, e39-e44.

Danpure, C. J., Birdsey, G. M., Rumsby, G., Lumb, M. J., Purdue, P. E. & Allsop, J. (1994a) Molecular characterization and clinical use of a polymorphic tandem repeat in an intron of the human alanine:glyoxylate aminotransferase gene. *Hum Genet* **94**, 55–64.

Danpure, C. J., Cooper, P. J., Wise, P. J. & Jennings, P. R. (1989) An enzyme trafficking defect in two patients with primary hyperoxaluria type 1: peroxisomal alanine/glyoxylate aminotransferase rerouted to mitochondria. *J Cell Biol* **108**, 1345–1352.

Danpure, C. J., Guttridge, K. M., Fryer, P., Jennings, P. R., Allsop, J. & Purdue, P. E. (1990) Subcellular distribution of hepatic alanine:glyoxylate aminotransferase in various mammalian species. *J Cell.Sci* **97**, 669–678.

Danpure, C. J. & Jennings, P. R. (1986) Peroxisomal alanine:glyoxylate aminotransferase deficiency in primary hyperoxaluria type 1. *FEBS Letters* **201**, 20–24.

Danpure, C. J., Jennings, P. R., Fryer, P., Purdue, P. E. & Allsop, J. (1994b) Primary hyperoxaluria type 1: genotypic and phenotypic heterogeneity. *J Inherit Metab Dis* **17**, 487–499.

Danpure, C. J. & Rumsby, G. (1996) Strategies for the prenatal diagnosis of primary hyperoxaluria type 1. *Prenat Diagn* **16**, 587–598.

Dray, S. & Dufour, A. B. (2007) The ade4 package: implementing the duality diagram for ecologists. *J Stat Soft* **22**, 1–20.

Excoffier, L., Hofer, T. & Foll, M. (2009) Detecting loci under selection in a hierarchically structured population. *Heredity advance online publication*, doi:10.1038/hdy.2009.74.

Excoffier, L., Smouse, P. E. & Quattro, J. M. (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479–491.

Haglin, L. (1991) Nutrient intake among Saami people today compared with an old, traditional Saami diet. *Arctic Med Res* (Suppl), 741–746.

Haglin, L. (1999) The nutrient density of present-day and traditional diets and their health aspects: the Sami- and lumberjack families living in rural areas of Northern Sweden. *Int J Circ Health* **58**, 30–43.

Hofer, T., Ray, N., Wegmann, D. & Excoffier, L. (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. *Ann Hum Genet* **73**, 95–108.

Holbrook, J. D., Birdsey, G. M., Yang, Z., Bruford, M. W. & Danpure, C. J. (2000) Molecular adaptation of alanine:glyoxylate aminotransferase targeting in primates. *Mol Biol Evol* **17**, 387–400.

Hruschka, D. J. & Brandon, A. K. (2004) Mongolia. In: *Encyclopedia of medical anthropology. Volume II: Cultures* (eds. C. R. Ember & M. Ember), pp. 850–862. New York: Springer.

Kozlov, A., Borinskaya, S., Vershubsky, G., Vasilyev, E., Popov, V., Sokolova, M., Sanina, E., Kaljina, N., Rebrikov, D., Lisitsyn, D. & Yankovsky, N. (2008) Genes related to the metabolism of nutrients in the Kola Sami population. *Int J Circ Health* **67**, 56–66.

Lumb, M. J. & Danpure, C. J. (2000) Functional synergism between the most common polymorphism in human alanine:glyoxylate aminotransferase and four of the most common disease-causing mutations. *J Biol Chem* **275**, 36415–36422.

Maniatis, T., Fritsh, E. F. & Sambrook, J. (1982) *Molecular cloning. A laboratory manual*. New York: Cold Spring Harbor Laboratory.

Noguchi, T. (1987) Aromatic-amino-acid aminotransferase from small intestine. *Methods Enzymol* **142**, 267–273.

Purdue, P. E., Lumb, M. J., Allsop, J. & Danpure, C. J. (1991) An intronic duplication in the alanine: glyoxylate aminotransferase gene facilitates identification of mutations in compound heterozygote patients with primary hyperoxaluria type 1. *Hum Genet* **87**, 394–396.

Purdue, P. E., Takada, Y. & Danpure, C. J. (1990) Identification of mutations associated with peroxisome-to-mitochondrion mistargeting of alanine/glyoxylate aminotransferase in primary hyperoxaluria type 1. *J Cell Biol* **111**, 2341–2351.

R Development Core Team (2007) R: A Language and Environment for Statistical Computing. Vienna: R Foundation for Statistical Computing, URL http://www.R-project.org.

Rosenberg, N. A., Pritchard, J. K., Weber, J. L., Cann, H. M., Kidd, K. K., Zhivotovsky, L. A. & Feldman, M. W. (2002) Genetic structure of human populations. *Science* **298**, 2381–2385.

Rousset, F. (2008) GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Mol Ecol Res* **8**, 103–106.

Scott, D. W. (2002) Multivariate Density Estimation: Theory, Practice, and Visualization. New York: Wiley.

Ségurel, L., Martinez-Cruz, B., Quintana-Murci, L., Balaresque, P., Georges, M., Hegay, T., Aldashev, A., Nasyrova, F., Jobling, M. A., Heyer, E. & Vitalis, R. (2008) Sex-specific genetic structure and social organization in Central Asia: insights from a multi-locus study. *PLoS Genet* **4**, e1000200.

Slatkin, M. & Voelm, L. (1991) F_{ST} in a hierarchical island model. *Genetics* **127**, 627–629.

Takayama, T., Fujita, K., Suzuki, K., Sakaguchi, M., Fujie, M., Nagai, E., Watanabe, S., Ichiyama, A. & Ogawa, Y. (2003) Control of oxalate formation from L-hydroxyproline in liver mitochondria. *J Am Soc Nephrol* **14**, 939–946.

Tambets, K., Rootsi, S., Kivisild, T., Help, H., Serk, P., Loogvali, E. L., Tolk, H. V., Reidla, M., Metspalu, E., Pliss, L., Balanovsky, O., Pshenichnov, A., Balanovska, E., Gubina, M., Zhadanov, S., Osipova, L., Damba, L., Voevoda, M., Kutuev, I., Bermisheva, M., Khusnutdinova, E., Gusar, V., Grechanina, E., Parik, J., Pennarun, E., Richard, C., Chaventre, A., Moisan, J. P., Barac, L., Pericic, M., Rudan, P., Terzic, R., Mikerezi, I., Krumina, A., Baumanis, V., Koziel, S., Rickards, O., De Stefano, G. F., Anagnou, N., Pappa, K. I., Michalodimitrakis, E., Ferak, V., Furedi, S., Komel, R., Beckman, L. & Villems, R. (2004) The western and eastern roots of the Saami–the story of genetic "outliers" told by mitochondrial DNA and Y chromosomes. *Am J Hum Genet* **74**, 661–682.

Tishkoff, S. A., Reed, F. A., Ranciaro, A., Voight, B. F., Babbitt, C. C., Silverman, J. S., Powell, K., Mortensen, H. M., Hirbo, J. B., Osman, M., Ibrahim, M., Omar, S. A., Lema, G., Nyambo, T. B., Ghori, J., Bumpstead, S., Pritchard, J. K., Wray, G. A. & Deloukas, P. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* **39**, 31–40.

Vigouroux, Y. & Couvet, D. (2000) The hierarchical island model revisited. *Genet Sel Evol* **32**, 395–402.

Vitalis, R., Dawson, K. & Boursot, P. (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics* **158**, 1811–1823.

Williams, H. E. & Smith, L. H. (1983) Primary Hyperoxaluria. In: *The Metabolic Basis of Inherited Disease* (eds. J. B. Stanbury, J. B. Wyngaarden, D. S. Frederickson, J. L. Goldstein & M. S. Brown), pp. 204–228. New York: McGraw-Hill.

Wright, S. (1951) The genetical structure of populations. *Ann Eugen* **15**, 323–354.

Received: 19 May 2009
Accepted: 7 September 2009