

Polymorphisme des Populations Humaines en Asie Centrale.

M. Sémon et S. Mousset

Génétique des Populations - Génétique Humaine; Données mitochondriales (séquence d'ADN); Données nucléaires (Microsatellites du Chromosome Y); Mesure du polymorphisme en Génétique des Populations Humaines; Mesure de l'Hétérozygotie; Structuration des Populations; D'après Chaix et al. [2007]

Table des matières

1	Analyse des données mitochondriales	3
1.1	Estimation de la diversité génétique dans chaque population . . .	3
1.2	Estimation de la diversité génétique entre populations	4
1.3	Représentation de la distance entre individus	5
1.4	Histoire des populations	5
2	Polymorphisme des populations humaine : Chromosome Y	6
2.1	Fréquences des Allèles dans les populations	7
2.2	Diversité moyenne (θ_π)	7
2.3	Hétérozygotie H	7
2.4	Structuration des populations	8
3	Bibliographie	9


Ce TP a pour but de présenter des indices très utilisés en génétique des populations humaine, en les appliquant au jeu de données de Chaix et al. [2007].


Le but de l'étude est d'évaluer l'impact de l'organisation sociale sur la diversité génétique dans les populations d'Asie Centrale. Les populations de pasteurs et de cultivateurs ont coexisté en Asie Centrale depuis le 4ème millénaire avant JC. Elles présentent des différences de mode de vie et de modes d'alimentation mais aussi différents types d'organisation sociale :


Les populations pastorales sont organisées en groupes de descendance (tribus, clans et lignées), et pratiquent les mariages exogames, c'est à dire qu'un homme choisit une femme dans un clan ou une lignée différente. En Asie Centrale, ces différents groupes sont patrilinéaires : les enfants sont systématiquement affiliés au groupe de descendance de leur père.

Les populations de cultivateurs, au contraire, sont organisées en familles (étendues ou nucléaires) et souvent établissent des mariages endogames entre cousins.

Les données que nous allons utiliser sont issues de marqueurs microsatellites sur le Y et de séquences de la région de contrôle mitochondriale (HVS-1), pour 741 individus appartenant à une vingtaine de populations d'Asie Centrale.

Les analyses seront réalisées avec le logiciel . Comme les bibliothèques de fonctions spécifiques à ces questions sont encore peu développées, nous utiliserons des fonctions que nous avons écrites pour ce TP.

Remarque : Les librairies  `adegenet` et `pegas` pourraient être utilisées pour l'analyse d'autres jeux de données de génétique des population.

Vous commencerez donc le TP par télécharger sur le bureau le dossier qui est disponible à l'adresse suivante : `ftp://pbil.univ-lyon1.fr/pub/cours/MOUSSET/gdp-gh/`. Vous ouvrirez ensuite une console de travail, et vous irez vous placer dans le dossier que vous venez de télécharger. Vous pouvez ensuite lancer  (taper `R` dans la console). Il faut ensuite vérifier que les librairies `seqinr`, `ape` et `ade4` sont installées (accessoirement, on utilisera dans ce TP la bibliothèque graphique `grDevices`). Pour cela taper :

```
library("ade4")
library("seqinr")
library("ape")
library("grDevices")
```

Si l'une de ces librairies n'est pas installée, il vous faudra l'installer à l'aide de la commande suivante (à adapter selon la librairie manquante) :

```
install.packages("ade4")
```

1 Analyse des données mitochondriales

1.1 Estimation de la diversité génétique dans chaque population

D'après ce que vous avez lu des caractéristiques des populations de pasteurs et d'agriculteurs en Asie centrale, répondez aux questions suivantes :

- ★ Pourquoi contraster les données mitochondriales et les données sur le chromosome Y ?
- ★ Pourquoi contraster les cultivateurs et les pasteurs ?

Lire les données mitochondriales, pour cela taper :

```
options(encoding = "latin1")
x = 4
source(file = "scripts/donneesm.r")
```

Ceci permet le chargement de deux fichiers dans votre espace de travail, et deux objets sont donc disponibles, nommés `popm` et `donneesm`. Ils contiennent respectivement les caractéristiques des populations étudiées et les séquences des individus dans chaque population. Un troisième objet, nommé `lpop` donne la correspondance entre le nom de l'individu et l'identifiant de sa population d'origine.

```
popm
  popid popname      population      status long lat
1      1   KK1      Karakalpaks (Qongirat) pastoral  59  43
2      2   KZ1      Kazaks (Karakalpakia) pastoral  63  44
3      3   OTU1     Karakalpaks (On Tort Uruw) pastoral  60  42
4      4   TK1 Turkmen (Uzbekistan/Turkmenistan border) pastoral  59  42
5      5   UZ1      Uzbeks (North) farmer  60  43
6      6   KZ3      Kazakhs (Kazakhstan) pastoral  80  45
7      7   LKIR3     Kyrgyzs (Talas) pastoral  72  42
8      8   HKIR3     Kyrgyzs (Sary-Tash) pastoral  73  40
9      9   UI3      Uyghurs (Kazakhstan) farmer  82  47
10     10  DUN4      Dungans farmer  78  41
11     11  KK4      Karakalpaks (Uzbekistan) pastoral  58  43
12     12  KZ4      Kazakhs pastoral  68  42
13     13  KUZ4      Uzbeks (Khorezm) farmer  61  42
14     14  KIR4      Kyrgyzs (Kyrgystan) pastoral  74  41
15     15  TK4      Turkmen pastoral  59  40
16     16  UI4      Uyghurs (Kyrgystan) farmer  79  42
17     17  UZ4      Uzbeks (South) farmer  66  40
18     18  TD4      Tajiks (Yagnobi) farmer  71  39
19     19  UZ2      Uzbeks (South,Uzbekistan:Surkhandarya) farmer  67  38
20     20  TK2      Turkmen (Turkmenistan) pastoral  60  39
21     21  KRT2      Kurds farmer  59  39

table(lpop)
lpop
  1 10 11 12 13 14 15 16 17 18 19  2 20 21  3  4  5  6  7  8  9
55 16 20 20 20 20 20 16 20 20 42 50 41 32 53 51 40 55 48 47 55
```

Vous pouvez estimer la diversité génétique intrapopulation dans ces données en tapant dans la console :

```
source("scripts/pim.r")
```

Ceci va créer un vecteur nommé `pim` dans votre environnement de travail. `pim` contient le nombre moyen de différences entre paires de séquences, pour chaque population.

- ★ Les valeurs du vecteur `pim` sont les valeurs d'un estimateur classiquement utilisé en génétique des populations. Comment s'appelle cet estimateur ?
- ★ Faites un graphique représentant la diversité génétique pour chaque population (utiliser la fonction `barplot`)
- ★ La diversité génétique mitochondriale est-elle significativement différente entre populations agricultrices et pastorale ?

Un peu d'aide pour commencer...

```
pimfarmer <- pim[popm[["status"]] == "farmer"]
pimpastoral <- pim[popm[["status"]] == "pastoral"]
```

Vous pouvez aussi estimer la diversité génétique intrapopulation dans ces données en tapant :

```
source("scripts/hm.r")
```

dans la console. Ceci va créer un vecteur nommé `hm` dans votre environnement de travail. `hm` contient l'hétérozygotie moyenne pour chaque population.

- ★ Comment est calculé cet estimateur ?
- ★ Que signifie "hétérozygotie pour ce type de données ?
- ★ Que vous attendez-vous à voir dans le cas des données mitochondriales ?
- ★ Faites un graphique représentant la diversité génétique pour chaque population (utilisez la fonction `barplot`)
- ★ La diversité génétique mitochondriale est-elle significativement différente entre populations agricultrices et pastorale ?
- ★ Conclure sur le lien entre le mode de vie et la diversité intra-population mesurée par des marqueurs mitochondriaux.

1.2 Estimation de la diversité génétique entre populations

Nous allons à présent étudier l'effet de la structuration des populations sur la diversité génétique. Pour cela, nous allons quantifier la part de diversité génétique créée par cette structuration.

- ★ De quelle façon pourriez-vous mesurer la diversité génétique entre populations ?
- ★ Que vous attendez-vous à voir dans le cas des données mitochondriales ?

Vous pouvez calculer les ϕ_{ST} entre paires de populations en tapant :

```
source(file = "scripts/phistm.r")
```

Ceci retourne une matrice de ϕ_{ST} pour chaque paire de population. Elle est stockée dans la matrice `tabphi`.

- ★ Les distances inter-populations (mesurées par les ϕ_{ST}) sont-elles différentes selon que l'on compare des populations agricultrices ou pastorales ?

- ★ Conclure sur le lien entre le mode de vie et la structuration des populations mesurée par des marqueurs mitochondriaux.

Aide : les numéros des populations sont accessibles de cette manière :

```
popfarmer = popm$popid[popm[["status"]] == "farmer"]
poppastoral = popm$popid[popm[["status"]] == "pastoral"]
```

1.3 Représentation de la distance entre individus

- ★ Comment feriez-vous pour représenter graphiquement les distances entre individus ?

En tapant la ligne ci-dessous dans la console, vous obtenez un graphique où chaque chiffre représente un individu et chaque couleur une population.

```
source(file = "scripts/arbrem.r")
```

- ★ Cette représentation confirme-t-elle vos interprétations des données mitochondriales ?
- ★ En général, les données ne sont pas représentées par des arbres en GdP, pourquoi ?

1.4 Histoire des populations

Le D de Tajima [Tajima, 1989] est souvent utilisé en génétique des populations pour caractériser l'histoire démographique des populations.

- ★ Rappelez ce qu'est le D de Tajima, et comment il s'interprète.

Pour calculer le D de Tajima, il est d'abord nécessaire de calculer l'estimateur de Watterson de $\theta = 4N_e\mu$ [Watterson, 1975]. Pour le calcul de cet estimateur (noté $\hat{\theta}_w$), plusieurs nouvelles valeurs seront calculées et stockées dans le tableau `popm` en tapant les commandes suivantes :

```
source(file = "scripts/thetam.r")
source(file = "scripts/Dtajm.r")
```

De nouvelles colonnes sont créées dans le tableau `popm` :

`nmut` est le nombre inféré de mutations dans chaque population

`S` est le nombre de sites ségrégeant dans chaque population

`nsites` est le nombre de sites (sans gap) dans chaque population

`theta.w` est la valeur de $\hat{\theta}_w$

Le D de Tajima dans une population peut être calculé en utilisant la fonction `tajD` à présent définie. Par exemple, pour la population 3, le calcul du D de Tajima s'effectue de la façon suivante :

```
tajD(popm[3, ])
[1] -1.917314
```

- ★ Calculez le D de Tajima pour chaque ligne du tableau `popm` et stockez le dans le tableau `popm` :
- ★ Comment interprétez-vous les valeurs obtenues ?

2 Polymorphisme des populations humaine : Chromosome Y

Nous nous intéressons à présent aux données récoltées à six locus microsatellites du chromosome Y. Les données proviennent de 18 populations. Pour lire les données, vous devez entrer la commande suivante :

```
source(file = "scripts/donneesy.r")
```

Trois nouvelles variables sont créées dans votre environnement \mathbb{R} . Il s'agit de :

`donneesy` : un tableau contenant les données.

`locusy` : les noms des six locus microsatellites.

`popy` : un tableau décrivant les populations (le "code" utilisé dans le tableau `donneesy` est donné dans la colonne `popname` du tableau).

On peut simplement obtenir les effectifs des différents allèles à l'un des locus avec la fonction `table` de \mathbb{R} , comme ci-dessous pour le locus DYS388 :

```
locusy[1]
[1] "DYS388"
table(donneesy[, locusy[1]], donneesy[, "pop"])
      KK1 KZ1 OTU1 TK1 UZ1 DUN2 KIR2 KRT2 KZ2 MG2 TK2 TD2 UZ2 UI2 KZ3 HKIR3 LKIR3 UI3
10    1  0    1  0  1    4    1    0  0  3  0  0  1  2  0  0  0  2
11    0  0    1  1  0    0    0    0  0  0  0  0  0  0  1  0  0  1
12   42 10   25 42 29   11   32  13  11 38 15 18 16 15  5  41  26 25
13    1 35    4  1  2    2    2  0 24 11  1  2  3  2  3  0  7  2
14    9  5   22  2  5    2    4  2  3 12  3  1  3  8 39  2  3  9
15    0  0    0  2  0    2    1  1  0  0  2  1  2  2  0  0  0  0
16    1  0    1  1  2    0    1  4  0  1  0  0  2  3  0  0  0  0
17    0  0    0  2  0    1    0  0  0  0  0  0  1  1  0  0  2  0
18    0  0    0  0  1    0    0  0  0  0  0  0  0  0  1  0  3  0
```

En effectuant les sommes en colonne des effectifs, on obtient les effectifs de l'échantillon qui permettent d'établir les fréquences alléliques.

```
colSums(table(donneesy[, locusy[1]], donneesy[, "pop"]))
      KK1  KZ1  OTU1  TK1  UZ1  DUN2  KIR2  KRT2  KZ2  MG2  TK2  TD2  UZ2  UI2
54    50    54    51    40    22    41    20    38    65    21    22    28    33
      KZ3  HKIR3  LKIR3  UI3
49    43    41    39
```

Par la suite on distinguera les populations agricultrices des populations pastorales. On pourra utiliser les commandes ci-dessous :

```
popfarmer <- popy[popy[["status"]] == "farmer", "popname"]
popfarmer
[1] "UZ1" "DUN2" "KRT2" "TD2" "UZ2" "UI2" "UI3"
poppastoral <- popy[popy[["status"]] == "pastoral", "popname"]
poppastoral
[1] "KK1" "KZ1" "OTU1" "TK1" "KIR2" "KZ2" "MG2" "TK2" "KZ3" "HKIR3"
[11] "LKIR3"
```

2.1 Fréquences des Allèles dans les populations

La commande ci-dessous permet de calculer une matrice `freqal` qui contient les fréquences de chacun des allèles dans les sous-populations :

```
source(file = "scripts/freqypop.r")
```

Donnez une représentation graphique des fréquences alléliques aux six locus. Vous pourrez utiliser une boucle `for` et représenter les six graphiques sur la même fenêtre de sortie en utilisant la commande :

```
par(mfrow = c(2, 3), mar = c(4, 2, 3, 2), las = 2)
```

- ★ Quels mécanismes peuvent expliquer les variations de fréquences alléliques dans les populations ?
- ★ Où vous attendez-vous à trouver les variations de fréquences les plus rapides ?

2.2 Diversité moyenne (θ_π)

On cherche à présent à étudier la diversité génétique de chacune des populations considérées. La commande ci-dessous permet de calculer les deux variables suivantes :

`diffmat` contient la matrice des nombres de différences entre toutes les paires d'haplotypes (de chromosomes Y).

`obspi` contient le vecteur des nombres moyens de différences par paires de séquences, calculés par population.

```
source(file = "scripts/piy.r")
```

- ★ Donnez une représentation graphique des diversités génétiques par population.
- ★ La diversité moyenne (θ_π) diffère-t-elle entre fermiers et pasteurs ?
- ★ Votre conclusion est-elle en accord avec ce que vous savez des populations agricultrices et pastorales ?

2.3 Hétérozygotie H

L'hétérozygotie (notée H), telle que calculée par les auteurs de l'étude, est la probabilité pour une paire de chromosomes d'être porteurs d'haplotypes différents dans chacun des échantillons. Ici, H est donc calculée sur les haplotypes constitués par les 6 locus du chromosome Y. La commande ci-dessous permet de calculer les valeurs observées de H par population. Ces valeurs sont stockées dans le vecteur `Hobs`.

```
source(file = "scripts/hy.r")
```

- ★ Donnez une représentation graphique des hétérozygoties par population.
- ★ L'hétérozygotie moyenne (H) diffère-t-elle entre fermiers et pasteurs ?
- ★ Votre conclusion est-elle en accord avec ce que vous savez des populations agricultrices et pastorales ?
- ★ Que pensez-vous de cette méthode de calcul de H ? (Imaginez que l'on augmente le nombre de marqueurs microsatellites utilisés sur le chromosome Y).


2.4 Structuration des populations

La structuration des populations (“distance génétique” entre populations) peut être calculée à l’aide de l’indice R_{ST} . Cet indice prend en compte le mécanisme mutationnel des microsatellites : la distance entre deux haplotypes est égale à la somme des carrés des différences des allèles de ces haplotypes [Slatkin, 1995]. La commande ci-dessous permet de définir une fonction `Rst` et de calculer les R_{ST} entre paires de populations. Ces distances génétiques sont stockées dans la matrice de distances `genetd` :

```
source(file = "scripts/rst.r")
```

La commande ci-dessous permet de calculer les distances géographiques entre populations (en unités arbitraires) à partir de leurs latitudes et longitudes. Ces distances sont stockées dans la matrice de distances `geod`.

```
source(file = "scripts/geod.r")
```

- ★ Faites une représentation graphique de la distribution des distances génétiques et géographiques entre paires de populations pour les populations agricultrices et pastorales (vous pourrez utiliser la commande `boxplot` de .
- ★ La structuration génétique des populations est-elle la même pour les deux types de populations ?
- ★ Cette différence de structuration génétique est-elle le reflet d’une différence de structuration géographique ?
- ★ Ces résultats sont-ils en accord avec ce que vous savez de ces deux types de populations ?
- ★ Quelle critique pouvez-vous concernant le test statistique effectué ?

En génétique des populations, un phénomène couramment observé est le phénomène d’isolement par la distance (*isolation by distance*). Ce phénomène se caractérise par une corrélation positive entre les distances génétiques et les distances géographiques. La commande suivante permet de représenter la distance génétique (R_{ST}) en fonction de la distance géographique et d’ajuster un modèle (dit “modèle linéaire”) du type

$$R_{ST} = \lambda d_{\text{geo}} + \varepsilon$$

où ε est l’erreur du modèle (dans un modèle linéaire au sens strict, les erreurs ε sont supposées indépendantes et distribuées dans la même loi normale).


```
source(file = "scripts/ibdy.r")
```

- ★ Que pensez-vous du résultat de ces modèles ? On pourra construire et faire l'analyse des modèles linéaires proposés en entrant les commandes suivantes.

```
source(file = "scripts/linearibdy.r")
summary(ibdfarmer)
summary(ibdpastoral)
```

- ★ Quelle critique formuleriez-vous à l'encontre de cette analyse statistique ?

L'analyse statistique ci-dessus est inappropriée (et nous ne vous l'avons montré qu'à titre d'exemple à *ne surtout pas suivre*). La méthode appropriée pour tester la corrélation entre deux matrices de distances est le test de Mantel qui procède à l'aide de rééchantillonnages.

- ★ Pour chacun des deux types de population, refaites les analyses concernant l'isolement par la distance à l'aide d'un test de Mantel (commande `mantel.randtest` de .
- ★ L'hypothèse d'isolement par la distance est-elle supportée par les données génétiques du chromosome Y ?

```
library(ade4)
mantel.randtest(as.dist(genetd[popfarmer, popfarmer]), as.dist(geod[popfarmer,
  popfarmer]), nrepet = 10000)
mantel.randtest(as.dist(genetd[poppastoral, poppastoral]), as.dist(geod[poppastoral,
  poppastoral]), nrepet = 10000)
```

3 Bibliographie

Références

- Raphaëlle Chaix, Lluís Quintana-Murci, Tatyana Hegay, Michael F Hammer, Zahra Mobasher, Frédéric Austerlitz, and Evelyne Heyer. From social to genetic structures in central asia. *Curr Biol*, 17(1) :43–48, Jan 2007. doi : 10.1016/j.cub.2006.10.058. URL <http://dx.doi.org/10.1016/j.cub.2006.10.058>.
- F. Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, 123(3) :585–595, 1989.
- G.A. Watterson. On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.*, 7 :256–276, 1975.
- M. Slatkin. A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1) :457–462, Jan 1995.