

TP Structure

Dakar février 2011

Objectif

L'objectif de ce TP est de faire fonctionner le logiciel [structure](#) pour mettre en évidence la structuration génétique de populations de poissons. Comme exemple, on utilisera un jeu de données de l'IRD que nous a fait parvenir Jean-Dominique Durand. L'espèce étudiée est un poisson marin, le mulot *Mugil cephalus*. C'est le poisson que l'on utilise pour faire la [poutargue](#), une spécialité marseillaise dont nous parlera Etienne.

Déroulement

Fichier de données

Décompresser l'archive *struct.zip*. Dans l'archive se trouve le jeu de données *mugil.txt* contenant les marqueurs moléculaires, ici des microsatellites, pour les 273 poissons échantillonnés. Il y a 10 loci (i.e. 10 différents marqueurs moléculaires) et comme les individus sont diploïdes (les individus diploïdes ont 2 copies de chaque chromosome), il y a 20 données associées à chaque individu. Par exemple, la deuxième ligne du fichier *mugil.txt* se présente sous la forme suivante

```
RU1 18 105 115 187 187 180 200 139 145 137 137 161 161 166 170 191 199  
180 182 118 118
```

RU1 est l'identifiant de l'individu, 18 est l'identifiant de la population, 105 et 115 sont les deux allèles (versions) du premier microsatellite et ainsi de suite.

Structure

Après avoir ouvert le logiciel structure, créez le projet mugil et chargez le jeu de données *mugil.txt*. Lors du démarrage vous aurez besoin des informations suivantes

Number of indiv 713

Ploidy of data 2

Number of loci 10

Missing data value 0

Check box

Row of marker names

Special format : data file stores for individuals in a single line

Individual ID for each individual

Putative population origin for each individual

Vous pouvez maintenant définir les différents paramètres du logiciel

Length of burnin period 10000

Number of MCMC reps after burnin 20000

Ancestry model Use admixture model
Allele frequency model : allele frequencies independent

Et lancer des runs de structure pour un nombre total de populations K allant de 2 à 6

Visualisation des résultats

Nous allons utiliser le logiciel R pour visualiser les résultats. Utiliser le répertoire où se trouve les sorties de *structure* comme répertoire courant dans R. Déplacer le fichier *thenames2* dans ce répertoire. Pour visualiser le résultat pour K=2, utiliser la fonction de R ci-dessous

```
display_barplot<-function(K=2)
{
  file<-paste("q",K, ".indivq",sep="")
  myq<-read.table(file,header=F)
  ncol2<-dim(myq)[2]
  ncol1<-ncol2-(K-1)
  tbl<-myq [,ncol1: ncol2]
  xx<-read.table("thenames2",header=F)
  xx<-as.character(xx[,1])
  xx[myq[,4]]->thenames
  barplot(t(as.matrix(tbl)), col=rainbow(3),xlab="Individuals",
  ylab="Ancestry",border=NA,names.arg=thenames,cex.names=.4)
}
```

Le code R utilise le fichier q2.indivq, fichier qui contient les coefficients de métissage pour les différents individus pour K=2

Proportion of membership of each pre-defined population in each of the 2 clusters

<i>Given Pop</i>	<i>Inferred Clusters</i>		<i>Number of Individuals</i>
	<i>1</i>	<i>2</i>	
<i>1:</i>	<i>0.111</i>	<i>0.889</i>	<i>50</i>
<i>2:</i>	<i>0.045</i>	<i>0.955</i>	<i>48</i>

Vous pouvez visualiser la structuration génétique du mulet dans le fichier q2.ps. Répéter l'opération pour K =3...5.

Description/Interprétation des résultats

Quels sont les clusters principaux qui émergent dans l'analyse avec *structure* ?

A partir de quelle valeur de K, le logiciel n'apporte plus d'information ?