TP RobPower

Voir Notice du Logiciel R (S. Teyssèdre)

Nous allons utiliser le logiciel pour illustrer les résultats de ce matin, à savoir les variations des risques de première espèce vrais pour un risque supposé en fonction de la méthode utilisée et de l'héritabilité, puis pour calculer la puissance de l'expérience simulée que nous allons utiliser durant tout ce cours.

1. Illustration vrai risque de première espèce selon les méthodes GWAS

Notre échantillon d'étude est un échantillon qui a été simulé par Hélène avec le logiciel LDSO (voir demain). Il servira de fil conducteur pour tous les TP du cours.

Cet échantillon comprend 781 animaux phénotypés et génotypés sur lesquels nous allons réaliser ensuite les études d'association. La population complète comprend 17000 animaux, dont 1609 ont été génotypés.

Ouvrir une fenêtre de terminal

Copier le répertoire du Cours dans votre propre espace avec le nom de votre choix :

```
cp -r /travail/COURS_RT/aricard   /chezvous/TParicard
```

Le TP RobPower est dans le répertoire `RobPower`. Se placer dans ce répertoire.

(l'application RobPower est téléchargeable sur QGSP

https://qgsp.jouy.inra.fr/index.php?option=com_content&view=article&id=54&Itemid=59, mais c'est déjà fait ici)

Se placer dans le bon répertoire (`cd /chezvous/TParicard/RobPower`)

Le fichier `MainTP.r` comporte toutes les commandes R utilisées.

Lancer R, tapez : `R`

Charger le programme

```
source(file="fonctions/RobPower.r")
```

La matrice de parenté entre les 781 animaux phénotypés et génotypés a été calculée en avance.

Elle s'appelle `Matrice_A.txt`

Dans RobPower, il y a la possibilité de n'extraire qu'une partie de la matrice de parenté totale. Comme ici, cette extraction est déjà faite (on n'a pas calculé la matrice de parenté avec les 17000

animaux), le fichier fileped n'est qu'un fichier qui donne dans l'ordre les chevaux à récupérer donc les 781 animaux, le fichier a été créé il s'appelle Choix_A.txt. Il n'y a donc qu'à charger la matrice de parenté

Charger la matrice de parenté :

```
source(file="fonctions/recupA.r")

Afull = read.table(file="Matrice_A.txt",header=FALSE)

creation = recupA(fileped="Choix_A.txt",Afull)

A = creation[[1]]

D = creation[[2]]
```

Nous sommes prêts à lancer RobPower pour notre échantillon de test.

Nous allons d'abord nous concentrer sur les variations des risques de première espèce. Le choix des options va donc être :

1. The A matrix -> *celle qu'on a obtenu : A*
2. The D matrix -> *celle qu'on a obtenu : D*
3. A value of heritability (can be a single value or a vector)-> *vecteur de 0.1 à 0.9*
4. A threshold -> *0.05*
5. The phenotypic variance explained by the QTL -> *1% (non utile dans cet exemple)*
6. The methods to compute -> *vecteur "reg" et, "grammar" (les plus démonstriatifs*).
7. Genomic Control (GC). FALSE
8. Linkage disequilibrium (r2) between tested marker and QTL.: *1*

```
RobPower(A,D,h2=seq(0.1,0.9,0.1),threshold=0.05,var.QTL=c(0.0
1),methods=c("reg","grammar"),GC=FALSE,r2=1)
```

Les résultats s'affichent à l'écran (ne pas s'inquiéter du message d'avis, cela n'est pas grave)

Par méthode, on a la Robustness (=risque de première espèce vrai) en fonction de l'héritabilité puis la puissance (Power) en fonction de l'héritabilité (lignes) et de l'importance du QTL (colonnes)

Pour faire des illustrations, récupérer ce résultat dans un objet :

```
Ex<-
RobPower(A,D,h2=seq(0.1,0.9,0.1),threshold=0.05,var.QTL=c(0.0
1),methods=c("reg","grammar"),GC=FALSE,r2=1)
```

Nous allons faire un graphique du risque de première espèce en fonction de l'héritabilité et de la méthode utilisée.

```
hh<-ex$regression$Robustness[1:9]

type_1_errorr<-ex$regression$Robustness[10:18]
```

```
type_1_errorg<-ex$grammar$Robustness[10:18]
fdr<-matrix(c(type_1_errorr,type_1_errorg),nrow=9)
matplot(hh,fdr,type="l")
abline(h=0.05,col="pink")
```

Sur le graphique en abscisse l'héritabilité, en ordonnée le risque de première espèce vrai. La ligne rose est le risque de première espèce supposé, en noir le vrai risque avec la régression, en rouge avec grammar. Et voila !(par ex, avec la régression, sur notre échantillon, avec une forte héritabilité, on multiplie par plus de 2 le vrai risque de première espèce par rapport à celui supposé)

2. Puissance de notre dispositif

Calculons maintenant la puissance de notre dispositif. La méthode que nous allons ensuite utiliser pour le GWAS est un modèle mixte SNP par SNP en supposant l'héritabilité connue, soit FASTA. L'héritabilité de la performance dans notre échantillon est 0.86 (no question).

On va choisir comme risque de première espèce, des risques « usuels »10%,5%, 1% quand on réalise 1 test. Mais comme on l'a vu ce matin, comme on va pendant le GWAS réaliser des milliers de tests, on va aussi utiliser un risque corrigé BONFERRONI. L'échantillon simulé comporte 2 chromosomes de 5000 SNP chacun dont 3842 et 3754 informatifs (non monomorphes). Le risque corrigé BONFERRONI est donc $2.6.10^{-5}$ (pour 10%) $1.3.10^{-5}$ (pour 5%) et $2.6.10^{-6}$ (pour 1%) (par chromosome)

Pour la valeur de l'effet du QTL, une petite révision. Avec le modèle :

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{x}\beta + \mathbf{Zu} + \mathbf{e}$$

L'effet d'un des allèles sur l'échelle brute est β , donc les 3 géntoypes ont pour effet $0 \to 0,1 \to \beta,2 \to 2\beta$

L'effet d'un allèle en écart type phénotypique est β / σ_y

La variance expliquée par le SNP est $\left(2p(1-p)\beta^2\right) / \sigma_y^2$ (avec *p l*a fréquence d'un des allèles)

Si le SNP n'est pas le QTL et est en déséquilibre de liaison avec le QTL mesurée par r^2 , l'effet capturé par le SNP est régréssé par rapport à l'effet du i QTL soit : $\beta_{SNP} = r\beta_{QTL}$

Donc on va choisir 1 « gros » QTL qui explique 5% de la variance phénotypique (il peut avoir un effet sur l'échelle phénotypique de 0.53 et une fréquence de 0.1 ou un effet de 0.32 et une fréquence de 0.5) et un « petit » QTL qui explique 1% de la variance phénotypique (effet 0.24 avec fréquence 0.1 ou effet 0.14 fréquence 0.5). Un effet d'un ½ écart type phénotypique (le 5% avec p=0.1) veut dire que l'écart entre les 2 homozygotes sera d'un écart type phénotypique, ce qui commence à être une différence très importante. Puis on va d'abord considérer que le SNP est le QTL ($r^2 = 1$), et ensuite on va supposer que le QTL ne fait pas parti de nos SNP et qu'il est en DL fort avec un SNP ($r^2 = 0.8$) ou faible ($r^2 = 0.2$).

Choix des options :

1. The A matrix -> *obtained previously*
2. The D matrix ->*obtained previously*
3. A value of heritability (can be a single value or a vector)-> *0.86*
4. A threshold -> *0.0000026, 000013, 0.000026, 0.01, 0.05, 0.10 (successivement)*
5. The phenotypic variance explained by the QTL -> *0.01 ,0.05*
6. The methods to compute -> *"fasta"*.
7. Genomic Control (GC). FALSE
8. Linkage disequilibrium (r2) between tested marker and QTL.: *1 (puis 0.8 puis 0.2)*

Pour les paramètres threshold et linkage desequilibrium, on est obligé de faire des boucles car le logiciel n'a pas prévu de vecteurs.

```
seuil<-c(0.0000026,0.000013, 0.000026,0.01,0.05,0.10)

dl<-c(0.2,0.8,1.0)

result<-matrix(0,nrow=length(seuil),ncol=2*length(dl))

for (i in 1:6) {

for (j in 1:3) {

ex<-
RobPower(A,D,h2=0.85),threshold=seuil[i],var.QTL=c(0.01,0.05)
,methods="fasta",GC=FALSE,r2=dl[j])

result[i,c(j*2-1,j*2)]<- ex$fasta$power[1:2]

}}

result
```

La matrice `result` contient les puissances avec en ligne les 6 seuls (0.0000026 0.000013 0.000026 0.01 0.05 0.10) et en colonne dans l'ordre le QTL 1% 5% avec dl=0.2 puis 1% 5% avec dl 0.8 puis 1% 5% avec dl de 1.

On constate qu'on n'a aucune chance de trouver un petit QTL (1%), même si il figure parmi les SNP (puissance 5.5% pour α=5%/3842) avec un seuil corrigé BONFERRONI. Pour trouver ce petit QTL il faut prendre les seuils « classiques 5% (puissance=79%) mais dans ce cas il sera noyé parmi 5%*3842=192 faux positifs. Un gros QTL (5%) ne sera trouvé que si il existe un SNP en dl suffisant (0.8) : puissance 80% avec α=1%/3842 mais puissance=3% si dl=0.2. Bref, notre dispositif avec 781 animaux phénotypés ne permet de trouver que les gros QTL. (essayer avec QTL 2% 3% ?)

Les tableaux sont dans /Resultats/Tableau

TP MULLER

OBJECTIF

Nous allons faire un GWAS sur 100 SNP de la population d'exemple qui servira toute la semaine. Puis nous calculerons les seuils avec la méthode de Muller et al. Nous illustrerons ces résultats par un Manhattan plot avec les seuils retenus.

Le GWAS est réalisé avec un modèle mixte simple pour chaque SNP

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{x}b + \mathbf{Zu} + \mathbf{e}$$

Avec **y** le vecteur des performances, b l'effet de l'allèle codé 2 au SNP, **u** le vecteur des valeurs polygénique, **e** la residuelle. **x** est un vecteur des génotypes (=0/1/2), **Z** est la matrice d'incidence des performances sur les valeurs polygéniques. On a $V(\mathbf{u}) = \mathbf{A}\sigma_u^2$ avec A la matrice de parenté. (Il n'y a pas d'autres effets fixes). L'héritabilité est supposée connue, elle est de 0.85.

Dans cet exemple, 1609 animaux sont génotypés et 781 sont phénotypés.

La généalogie comporte 17000 chevaux

Il y a 9995 SNP répartis équitablement sur 2 chromosomes

MULLER

Copier le répertoire /travail/COURS_RT/aricard chez vous (si ce n'est déjà fait)

Le fichier des genotypes est `genotypes1.uga`

Le fichier des généalogies est `genealogie`

Le fichier des performances est `sim_ped_perf_mean`

Le script qui permet de lancer les GWAS est `chaine_RT.sh`

Ouvrir le script `chaine_RT.sh`

Les premières lignes donnent les paramètres

```
#_____________________________________________________________
#
#      PARAMETRES
#_____________________________________________________________


REPERTOIRE=/prodanr/aricard/Cours      # repertoire local ou sont
actuellement les fichiers et programmes
PERFO=sim_ped_perf_mean             # fichier performance
```

```
GENOTYPE=genotypes1.uga                 # fichier genotype ou Repertoire
pour phases
GENEALOGIE=genealogie                   # Fichier genealogie
VY=33.2113080                           # variance phenotypique
H2=0.85                                 # heritabilite
MANQUANT=5                              # code genotype manquant
SNPMIN=1                                # numero du premier SNP traite
SNPMAX=100                              # numero du dernier SNP traite
SOLUTION=solution                       # Nom du fichiers de solution
REPERTOIREW=$REPERTOIRE/work                    #repertoire de travail
temporaire
ASREML=snp                              # nom temporaire des fichiers de
donnees
GENOTYPEB=${GENOTYPE}bis                  # nom du fichiers des genotpes
exploitables (pour Muller)
```

Changer dans le script le répertoire dans lequel vous travaillez : /prodanr/aricard/Cours **par le votre**

Les autres paramètres doivent être justes. Nous allons commencer par analyser les 100 premiers SNP. (SNPMAX=100)

Lancer le script en tapant `./chaine_RT.sh`

Il écrit ce qu'il est en train de faire, soit :

Lire les fichiers initiaux

Créer un répertoire de travail (work) dans lequel il écrit autant de fichier de données, fichiers de paramètres que de SNP, ainsi qu'un fichier de commande qui va lancer en parallèle la résolution des modèles mixtes grâce au logiciel de Misztal BLUPF90, et écrire les fichiers de solutions.

Lire les solutions (tests et solutions) et les compiler dans un seul fichier qu'il va écrire dans votre répertoire : `solution`

(attention, l'executable efface les fichiers paramètres parmp parml et le répertoire work si ils existent déjà)

Vous devez avoir un réperoire `work`, un fichier de solution `solution`, un fichier des génotypes utiles `genotypes1.ugabis`

Le fichier solution comporte une ligne par SNP avec le n° du SNP (dans l'ordre), son numéro de traitement (sans les SNP monomorphes non informatifs), le nombre d'animaux utilisés, la valeur du test F, la valeur de la p_value, la solution pour b, μ et leurs écarts types.

Le répertoire work comporte tous les fichiers intermédiaires. Ceux-ci seront utilisés pour le programme muller et pourront être effacés ensuite.

Le fichier genotypes1.ugabis comporte tous les génotypes utiles (nécessaire pour faire tourner le programme de muller). Par exemple pour les 100 premiers SNP seuls 73 sont utilisables.

Verifiez que l'executable : `exec.sh` pour le programme de muller correspond bien

Voir notice du programme de muller. Attention, le nombre de SNP utilisé n'est pas 100 mais 73

Taper `./exec.sh` pour lancer le programme de muller.

(il ecrit des warning mais c'est pas grave)

A l'issu du programme il imprime en fonction du taux d'erreur de première espèce choisi (1%, 5%, 10% ; 20%) le seuil de la p_value correspondant après correction pour les tests multiples. (notez ces valeurs, elles peuvent différer pour chacun car elles sont issues de simulations)

Nous allons illustrer ces seuils en réalisant un Manhattan plot dans R

Lancer R : R

Les différentes lignes de commandes utiles ont été regroupées dans `post_gwas.r`

Lire les données : `snps=read.table("solution",header=T)`

Nous allons d'abord faire un QQplot pour vérifier la bonne distribution de notre test

Créer un vecteur des –log10(p_value) triées : `yy <- sort(-log10(snps$pvalue))`

Sous H0 ces p_values sont distribuées uniformément, nous allons donc créer un vecteur avec des pvalues uniformes

Créer un vecteur des pvalues uniformes triées : `xx <- sort(-log10((1:73)/73))`

Le QQplot est la figure des p_values observées en fonction des théoriques

Figure :

```
plot (xx,yy)
abline (a=0,b=1,col="red")
```

La ligne rouge illustre la distribution attendue : bissectrice

Que faut il en penser ? La plupart des SNP n'ont pas d'effet, les pvalues obtenues correspondent donc bien à H0 (sauf pour les plus faibles, donc log plus élevé), si le choix de la loi du test est la bonne, on doit avoir pour toutes ces valeurs la même chose que les valeurs théoriques donc suivre la bissectrice… ce qui est le cas.

Nous allons maintenant réaliser le Manhattan plot : graphique qui donne les pvalue en fonction de la position du SNP sur le chromosome

Manhattan plot :

```
col=rep(c("blue","green"),each=3842)

plot(snps$snp,-log10(snps$pvalue),col=col,ylim=c(0,5))
```

Utilisons maintenant des seuils pour déclarer des effets significatifs

Tracer le seuil 10% : `abline(h=-log10(0.10),col="pink")`

Tracer le seuil 10% bonferroni : `abline(h=-log10(0.10/73),col="pink")`

Tracer le seuil 10% muller : `abline(h=-log10(3.82e-3),col="red")`

Qu'en pensez vous. Avec 100 SNP on ne voit pas grand-chose, on visualise bien les seuils : le « brut » très lache, le BONFERRONI très sévére et le MULLER plus adéquat.

Je vous ai calculé en avance les 10000 solutions pour tous les SNP (en fait 3842 sur le premier chromosome et 3754 sur le deuxième soit au total 7596 SNP après exclusion des monomorphes) dans le fichier `solution_tot`.

Faire les graphiques avec toutes ces solutions sachant que les seuils Muller par chromosome sont :

SEUIL	MULLER	BONFERRONI	-log10(MULLER)	-log10(BONFERRONI)	Nbr SNP "exclusifs"
0.01	4.9E-06	2.6E-06	5.31	5.58	2055
0.05	2.8E-05	1.3E-05	4.56	4.89	1797
0.1	6.0E-05	2.6E-05	4.22	4.58	1671
0.2	1.3E-04	5.2E-05	3.88	4.28	1514

La place des vrais QTL est donné par les verticales rouges.

Toutes les commandes pour le ficher `solution_tot` sont dans post_gwas.r

Sachant que la part de variance du premier QTL est 0.4%, le deuxième 1.5% le troisième 2.6%, et le quatrième est tri-allélique mais représente environ 0.3%, sommes nous chanceux ? (à10%, selon le TP précédent ou nous avions calculé la puissance, nous avions 38% de chance de trouver un QTL qui fait 2% de la variance en dl de 1 avec un risque première espèce 10% BONFERRONI et 72% pour un QTL qui fait 3%.

La valeur des solutions des effets des SNP vous semble-t elle cohérente (pour le SNP 6000 : -1.89 pour un ecart type phénotypique de 5.7… on l'a sans doute sur-estimé)