

Création des matrices $M_{qq'}$ pour l'algorithme proposé par JME pour l'application de la méthode de Muller et al., 2012

Par rapport au document « Contrôle de l'erreur de première espèce. Approche de Muller et al. V2 », on remplace les étapes 1, 2 et 3 par BLUPF90 et on se branche à la sortie de celui-ci pour construire $M_{qq'}$. Puis on reprend l'algorithme à l'étape 4.

Définitions :

N le nombre de SNP concernés (50 000 pour une puce classique, 500 000 pour haute densité)

N_{ch} le nombre de SNP utilisés simultanément dans la matrice M, donc par exemple par chromosome (3000 pour la puce classique)

n_{α_q} le nombre de niveau pour l'effet « SNP » (1 si modèle régression allélique 1 SNP, 10/20 si haplotypes et dans ce cas variable par SNP)

n_{β} le nombre de niveaux d'effets fixes autres (de qqz dizaines à qqz centaines)

n_u le nombre d'animaux genotypés (qqz milliers)

Pour le $q^{ième}$ SNP parmi les N, le modèle est

$$y = X\beta + W_q\alpha_q + u + e$$

Avec y le vecteur des performances, β le vecteur des effets fixes, α_q le vecteur des différents haplotypes ou génotypes ou allèles du SNP, u l'effet polygénique, e la résiduelle. Les matrices X, W_q sont des matrices d'incidences. Les différents effets sont estimés avec BLUPF90 (Miszta).

Pour faire tourner BLUPF90, N fichiers paramètres sont écrits qui décrivent le modèle ci-dessus. Dans le fichier paramètre, on donne le nombre d'effets (nombre d'effets fixes + 1 (pour l'effet du SNP) + 1 (pour l'effet polygénique)) et le nombre de niveaux par effet. Les effets sont listés sous le mot clé EFFECTS :...Selon leur position dans cette liste ils sont numérotés de 1 à nbr d'effets. Par convention, la modification de BLUPF90 que j'ai réalisé pour sortir les matrices suppose que l'effet du SNP est l'avant dernier et l'effet polygénique le dernier. Ainsi avec un modèle qui comprend comme dans mon exemple pour effets fixes : un effet moyenne (1 niveau), un effet sexe (3 niveaux) et un effet région (11 niveaux), donc au total 5 effets (3+2), l'effet numéroté 4 est l'effet du SNP et l'effet numéroté 5 est l'effet polygénique.

L'objectif est de calculer

$$\text{cov}(\hat{\alpha}_q, \hat{\alpha}_{q'}) = \sigma_e^2 \left(C_{\alpha_q \beta} X' W_q C_{\alpha_{q'} \alpha_{q'}} + C_{\alpha_q \alpha_q} W_q' W_q C_{\alpha_{q'} \alpha_{q'}} + C_{\alpha_q u} W_q C_{\alpha_{q'} \alpha_{q'}} \right) = M_{qq'}$$

Et

$$V(\hat{\alpha}_q) = \sigma_e^2 C_{\alpha_q \alpha_q} = M_{qq}$$

Avec pour définition des éléments de cette formules :

Les équations du modèle mixte résolue par BLUPF90 sont :

$$\begin{bmatrix} X'X & X'W_q & X' \\ W_q'X & W_q'W_q & W_q' \\ X & W_q & I + \lambda A^{-1} \end{bmatrix} \begin{bmatrix} \hat{\beta} \\ \hat{\alpha}_q \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'y \\ W_q'y \\ y \end{bmatrix} \text{ Avec } \lambda = \frac{\sigma_e^2}{\sigma_a^2} \text{ et}$$

On va utiliser

$$\begin{bmatrix} C_{\beta\beta} & C_{\beta\alpha_q} & C_{\beta u} \\ C_{\alpha_q\beta} & C_{\alpha_q\alpha_q} & C_{\alpha_q u} \\ C_{u\beta} & C_{u\alpha_q} & C_{uu} \end{bmatrix} \begin{bmatrix} X'X & X'W_q & X' \\ W_q'X & W_q'W_q & W_q' \\ X & W_q & I + \lambda A^{-1} \end{bmatrix} = \begin{bmatrix} I & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & I \end{bmatrix}$$

Création des fichiers utiles au calcul des matrices $M_{qq'}$

La sortie de BLUPF90 donne un fichier en format libre avec 4 colonnes séparées par des espaces.

Le nom du fichier est mat_XXX avec XXX=nom du fichier de données utilisé en entrée. Nous avons N fichiers de ce type.

Chaque ligne correspond à 1 élément d'une matrice X soit $x(i, j)$. Le premier indice est toujours celui du niveau de l'effet SNP, le deuxième est celui alternativement du niveau de l'effet SNP, du niveau des effets fixes, du niveau des effets polygéniques.

Colonne 1 : identifiant de la matrice selon le code (entier) :

$$0 = \frac{1}{\sigma_e^2} W_q' X \text{ (matrice } n_{\alpha_q} \times n_{\beta} \text{), notons la } A_q$$

$$1 = \sigma_e^2 C_{\alpha_q \alpha_q} \text{ (matrice } n_{\alpha_q} \times n_{\alpha_q} \text{), notons la } B_q$$

$$2 = \sigma_e^2 C_{\alpha_q \beta} \text{ (matrice } n_{\alpha_q} \times n_{\beta} \text{), notons la } C_q$$

$$3 = \sigma_e^2 C_{\alpha_q u} \text{ (matrice } n_{\alpha_q} \times n_u \text{) notons la } D_q$$

Colonne 2 : indice i de l'élément $x(i, j)$ (entier)

Colonne 3 : indice j de l'élément $x(i, j)$ (entier)

Colonne 4 : valeur de l'élément $x(i, j)$ (réel)

Comment se repérer dans les indices ?

- Pour l'effet du SNP donc i , ils sont numérotés de 1 à n_{α_q}
- Pour les effets fixes, ils sont numérotés de 1 au nombre total de niveaux pour tous les effets fixes. Dans mon exemple, il y a 3 effets fixes avec au total 15 niveaux (effet moyenne (1 niveau), un effet sexe (3 niveaux) et un effet région (11 niveaux)) donc ils sont numérotés de 1 à 15.
- Pour les effets polygéniques, ils sont numérotés de 1 à nombre d'animaux *dans le fichier pedigree* (éventuellement supérieur au nombre d'animaux génotypés). *Seuls sont écrits les éléments des chevaux génotypés* (ceux qui nous intéressent). Cet identifiant est le même que celui qui est donné dans le fichier des génotypes et de généalogies. Pour faire les calculs il sera sûrement plus intéressant de les renuméroter en interne de 1 à nombre de chevaux génotypés (sinon ça fait plein de zéros inutiles). Dans mon exemple il y a 9481 chevaux dans le fichier pedigree mais seulement 783 sont génotypés et par exemple le premier cheval génotypé a le numéro 1397, c'est donc le premier élément écrit pour la matrice D_q (code 3)

Calcul des matrices $M_{qq'}$

Pour ce calcul il va falloir utiliser le fichier des génotypes ou des haplotypes selon le modèle de l'effet des SNP.

Si le modèle est un modèle effet allélique du SNP, les génotypes sont codés 0/1/2 selon le nombre d'allèles de référence que le cheval possède à ce SNP. Traditionnellement, j'utilise comme fichier de génotype un fichier avec une ligne par animal et sur les 10 premiers caractères le n° de l'animal, 1 blanc, puis sur les N positions suivantes les génotypes à chaque SNP sans espace (du genre 001202221102...). Il n'y a donc qu'une seule ligne pour les 50000 génotypes, et un seul fichier.

Si le modèle est un modèle effet haplotypique du SNP, les haplotypes seront codés de 1 à n_{α_q} pour l'haplotype paternel et l'haplotype maternel. Traditionnellement les fichiers des haplotypes sont séparés par chromosome et chaque fichier est rangé dans un répertoire identifié par le nom du chromosome (style ECA12 pour le chromosome 12 de Equus Caballus). Le fichier issu de DAGPHASE est sous la forme de deux lignes par animal (l'haplotype paternel et l'haplotype maternel). Chaque ligne est constituée sur les 6 premières positions du numero de l'animal, puis sur 3 positions le n° de l'haplotype (1=paternel, 2=maternel), puis toutes les 4 positions le n° de l'haplotype au SNP pour tous les SNP du chromosome.

Il faut récupérer la variance résiduelle σ_e^2

A partir de ces fichiers et du fichier des génotypes, il faut réaliser $\frac{N_{ch}(N_{ch}-1)}{2}$ calculs de $M_{qq'}$ pour chaque chromosome. Donc une triple boucle :

DO chromo=1 TO nbr de chromosome

DO q=1 TO N_{ch}

DO q'=q+1 TO N_{ch} (termes $q \neq q'$, pour $q=q'$ voir en bas)

Il faut constituer 2 matrices supplémentaires

- W_q , une matrice notée E_q , de dimension $\langle n_u \times n_{\alpha_q} \rangle$. Pour le modèle allélique, le terme $i,1$ (car $n_{\alpha_q} = 1$) est :

$w_{q',i1} = 1$ si le génotype du cheval i est 0 pour le SNP q'

$w_{q',i1} = 2$ si le génotype du cheval i est 1 pour le SNP q'

$w_{q',i1} = 3$ si le génotype du cheval i est 2 pour le SNP q'

Pour le modèle haplotypique, le terme i,j est :

$w_{q',ij} = 2$ si les deux haplotypes du cheval i sont j pour le SNP q'

$w_{q',ij} = 1$ si l'un des haplotypes du cheval i est j pour le SNP q' (il doit dans ce cas y avoir 2 termes en j pour ce i avec $w_{q',ij} = 1$)
 $w_{q',ij} = 0$ pour tous les autres cas

- $W_q'W_q$, une matrice notée F_{qq} , de dimension $\langle n_{\alpha_q} \times n_{\alpha_q} \rangle$ dont un terme i,j est :

$$\omega_{ij} = \sum_{k=1}^{n_y} w_{q,ik} w_{q',jk}$$

Avec le même codage pour les éléments $w_{q',ik}$ que précédemment.

A partir de ces fichiers et des matrices supplémentaires il faut calculer les produits matriciels suivants :

- $\sigma_e^2 C_{\alpha_q \beta} X' W_q C_{\alpha_q \alpha_q}$, donc avec nos notations $C_q A_q' B_q$, donc un produit de dimension $\langle n_{\alpha} \times n_{\beta} \rangle \cdot \langle n_{\beta} \times n_{\alpha'} \rangle \cdot \langle n_{\alpha'} \times n_{\alpha'} \rangle$
- $\sigma_e^2 C_{\alpha_q \alpha_q} W_q' W_q C_{\alpha_q \alpha_q}$, donc avec nos notations $\frac{1}{\sigma_e^2} B_q F_{qq} B_q$, donc un produit de dimension $\langle n_{\alpha_q} \times n_{\alpha_q} \rangle \cdot \langle n_{\alpha_q} \times n_{\alpha_q} \rangle \cdot \langle n_{\alpha_q} \times n_{\alpha_q} \rangle$
- $\sigma_e^2 C_{\alpha_q u} W_q C_{\alpha_q \alpha_q}$, donc avec nos notations $\frac{1}{\sigma_e^2} D_q E_q B_q$, donc un produit de dimension $\langle n_{\alpha_q} \times n_u \rangle \cdot \langle n_u \times n_{\alpha_q} \rangle \cdot \langle n_{\alpha_q} \times n_{\alpha_q} \rangle$

On fait la somme de ces 3 matrices et on obtient M_{qq} .

Pour le terme diagonal M_{qq} il est tout simplement égale à B_q

Exemple

Sous dga11, dans /travail/aricard (transferable ailleurs où vous auriez les droits....) il y a un exemple avec un modèle allélique :

- Un répertoire **work_23M1** dans lequel il y a les fichiers correspondants au passage de BLUPF90 et aux fichiers des matrices utiles pour réaliser ce calcul pour les 10 premiers SNP de données de génotypage du projet GENENDURANCE
 - o Fichiers de données par SNP : snpregx.dat (x=1 à 10)
 - o Fichiers paramètres pour BLUPF90 : parmx (x=1 à 10)
 - o Fichier des généalogies util_labo (pour faire tourner BUPF90)
 - o Fichier de sortie de BLUPF90 pour les matrices : mat_snpregx.dat (x=1 à 10)
 - o (les fichiers test_snpregx.dat et solu_snpregx.dat sont les fichiers des tests et des solutions du GWAS mais non utilisé ici)

- Le programme BLUPF90 utilisé est celui de misztal modifié pour créer les sorties utiles. Ce programme est une succession de programme dans le répertoire **f90-05** (l'original est dans le répertoire f90-05_copie). Le programme de base est dans le répertoire f90-05/blup (c'est blupf90.f90). Pour recompiler il suffit de se placer dans f90-05 et taper make all.
- Le fichier initial des génotypes utilisé est /travail/aricard/genotypeb
- Pour s'y repérer parmi les SNP du fichier des genotypes pour savoir où il se situe dans une carte, j'ai le fichier freq_coresSNPb mais il est assez compliqué. A savoir en colonne 4 le n° d'ordre du SNP dans le fichier genotypeb (0 si le SNP n'est pas dans le fichier genotypeb) , colonne 5 le n° du chromosome, en colonne 7 la position en paires de base.