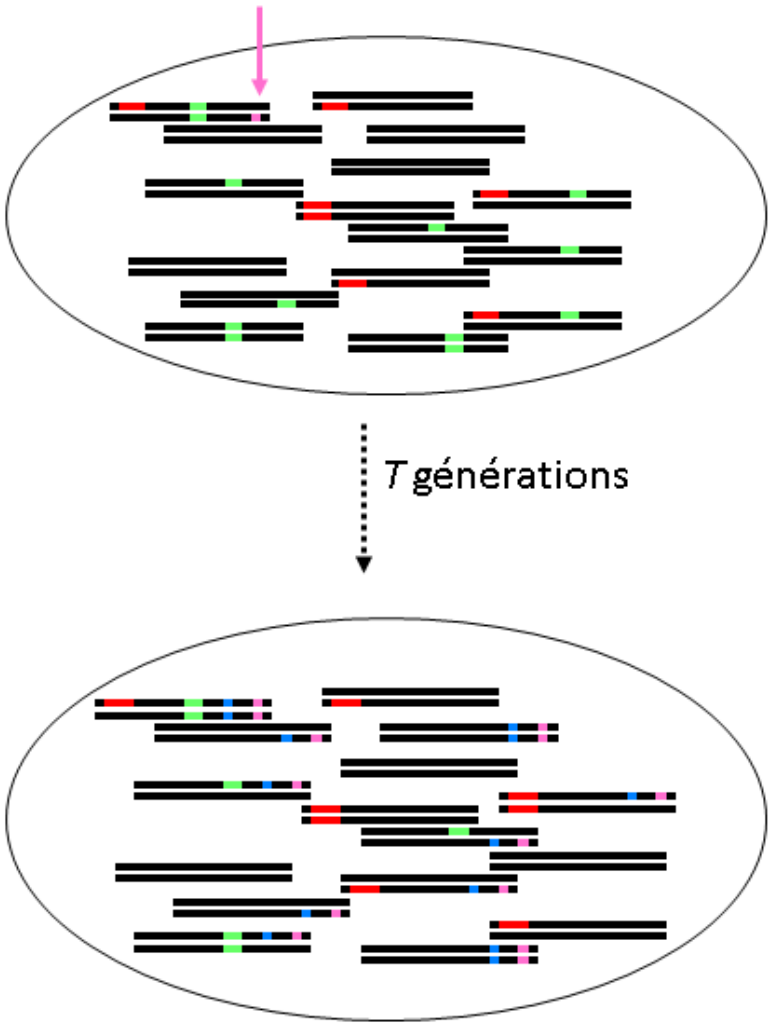
Génétique d'association

Brigitte Mangin, Anne Genissel

Septembre 2011

Plan

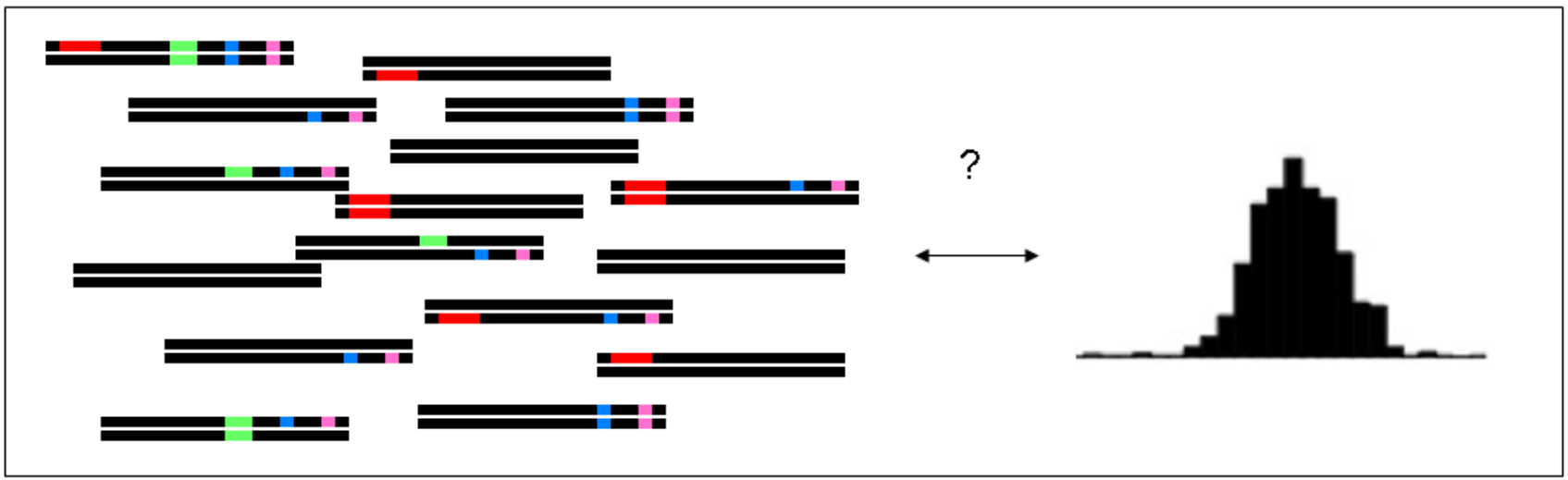
Contexte évolutif



Soit une population a un temps t_0 avec apparition d'une nouvelle mutation

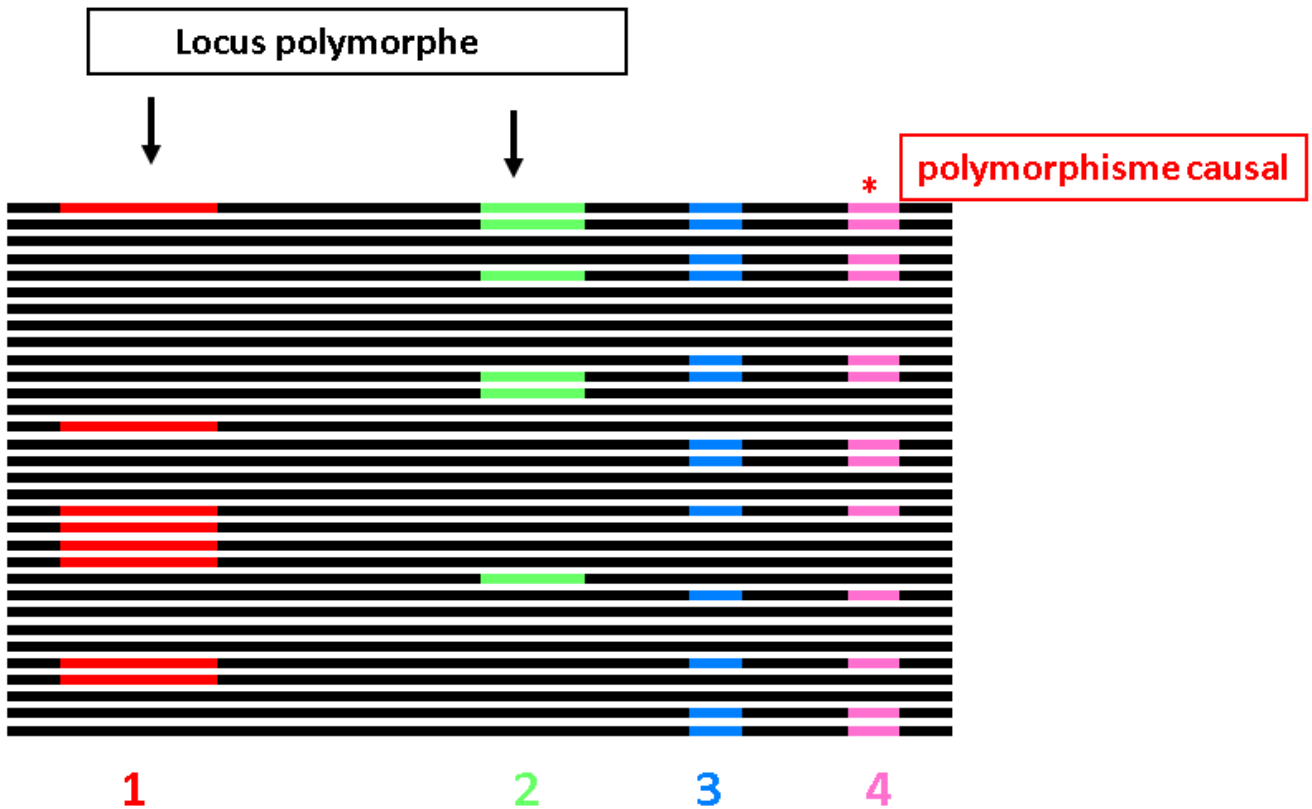
Au sein de cette même population a un temps $(t_0 + T)$ ce site polymorphe a augmenté en fréquence

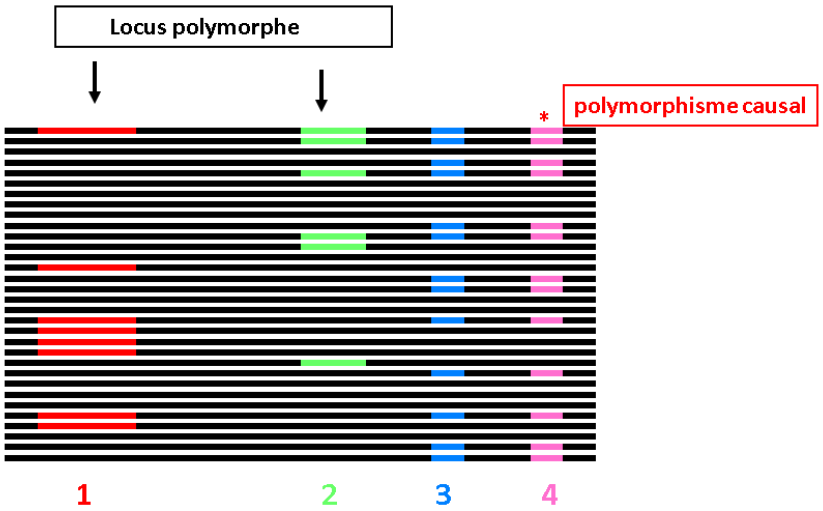
Objectif



Quel est le polymorphisme causal ?

- Tester une différence d'effet de l'allèle mutée par rapport à l'allèle sauvage
- en tout locus polymorphe





Les tests aux loci 3 et 4 sont identiques.

C'est le fait du déséquilibre de liaison (DL), que l'on exploite pour

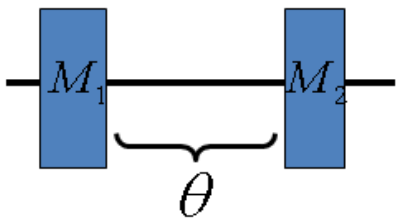
- réduire le génotypage (TAG SNP)
- en conservant la "couverture" du génome (ou de la région génomique)

Définition du déséquilibre de liaison

Déséquilibre gamétique $\pi_{ij} \neq \pi_{i+}\pi_{+j}$

Loci		M_2			Total
	Allèles		i		
M_1					
	j		π_{ij}		π_{+j}
Total			π_{i+}		

Liaison (physique) Taux de recombinaison $\theta < 1/2$



Déséquilibre de liaison : déséquilibre + liaison

Cas biallélique

$$D_{ij} = \pi_{ij} - \pi_{i+}\pi_{+j}$$

$$D \begin{cases} = D_{11} = D_{22} \\ = -D_{12} = -D_{21} \end{cases}$$

$$r^2 = \frac{D^2}{\pi_{1+}\pi_{+1}\pi_{2+}\pi_{+2}}$$

Mesure du χ^2

$$D' = \begin{cases} \dfrac{D}{\min(\pi_{1+}\pi_{+2}, \pi_{2+}\pi_{+1})} & \text{si } D > 0 \\ \dfrac{D}{\min(\pi_{1+}\pi_{+1}, \pi_{2+}\pi_{+2})} & \text{si } D < 0 \end{cases}$$

Mesure comprise entre -1 et +1
Lewontin (1964)

Un regard plus statistique

- $\Delta_{M_1,i}$ la dose d'allèle i au locus M_1
- $\Delta_{M_2,j}$ la dose d'allèle j au locus M_2

$$D_{ij} = Cov(\Delta_{M_1,i}, \Delta_{M_2,j}) \quad r^2_{ij} = Cor^2(\Delta_{M_1,i}, \Delta_{M_2,j})$$

Pour des haplotypes (phase connue)

$\Delta_{M_1,i} = 0$ ou 1 (id $\Delta_{M_2,j}$)

Pour des génotypes (phase inconnue)

$\Delta_{M_1,i} = 0$, 1, 2 (id $\Delta_{M_2,j}$)

Rogers & Huff, Genetics, 2009

Estimation

Soit l'observation des doses alléliques pour un échantillon de taille N

$$\Delta_{M_1,i} = \begin{pmatrix} \delta_{M_1,i,1} \\ \vdots \\ \delta_{M_1,i,n} \\ \vdots \\ \delta_{M_1,i,N} \end{pmatrix} \begin{pmatrix} \delta_{M_2,j,1} \\ \vdots \\ \delta_{M_2,j,n} \\ \vdots \\ \delta_{M_2,j,N} \end{pmatrix} = \Delta_{M_2,j}$$

$$\hat{r}^2_{ij} = \widehat{Cor}^2(\Delta_{M_1,i}, \Delta_{M_2,j})$$

où $\widehat{Cor}$ est la corrélation empirique

$\hat{r}^2$ est biaisé, $Esp(\hat{r}^2) \neq r^2$

- lorsque l'échantillon a une structure
- lorsque les individus sont fortement apparentés et d'apparentements contrastés

Des mesures qui corrigent ces biais r^2_S, r^2_V, r^2_{VS}

Mangin et al., Heredity, 2011

- lorsque la structure est connue (ou estimée) S
- lorsque l'apparentement est connu (ou estimé) V

Comme r^2 (dans le modèle simple), ces nouvelles mesures sont liés à la puissance du test d'association (dans le modèle réaliste)

corrigée de la structure S

$$r^2_{S,ij} = Corr^2(\Delta_{M_1,i}, \Delta_{M_2,j}; S)$$

corrigée de l'apparentement V

$$r^2_{V,ij} = Cor^2(V^{-1/2}\Delta_{M_1,i}, V^{-1/2}\Delta_{M_2,j})$$

corrigée de la structure et de l'apparentement S, V

$$r^2_{VS,ij} = Corr^2(V^{-1/2}\Delta_{M_1,i}, V^{-1/2}\Delta_{M_2,j}; S)$$

où $Corr(X, Y; Z)$ dénote la corrélation partielle de X et Y lorsque Z est constant,
ou encore la corrélation des résidus ϵ_X et ϵ_Y des régressions linéaires $X = S\beta + \epsilon_X$ et $Y = S\beta' + \epsilon_Y$

Les nouvelles mesures

Estimation

Comme pour r^2, la corrélation est estimée par la corrélation empirique.

On utilise la matrice S de structure en K groupes de l'échantillon

$$S = \begin{bmatrix} S_{1,1} & \dots & S_{1,K} \\ \vdots & \vdots & \vdots \\ S_{n,1} & \dots & S_{n,K} \\ \vdots & \vdots & \vdots \\ S_{N,1} & \dots & S_{N,K} \end{bmatrix}$$

Et/ou la matrice V de variance-covariance de l'échantillon

$$V = \begin{bmatrix} V_{1,1} & \dots & V_{1,n} & \dots & V_{1,N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ V_{n,1} & \dots & V_{n,n} & \dots & V_{n,N} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ V_{N,1} & \dots & V_{N,n} & \dots & V_{N,N} \end{bmatrix}$$

Package R : LDcorSV

Les nouvelles mesures

$$\Delta^t_{M_1,1} = \{ \underbrace{1,\ldots,1,}_{\text{50 fois}} \underbrace{0,\ldots,0,}_{\text{50 fois}} \underbrace{1,\ldots,1,}_{\text{90 fois}} \underbrace{0,\ldots,0}_{\text{10 fois}} \}$$

premier groupe — deuxième groupe

$$\Delta^t_{M_2,1} = \{ \underbrace{0,\ldots,0,}_{\text{50 fois}} \underbrace{1,\ldots,1,}_{\text{50 fois}} \underbrace{1,0,\ldots,1,0,}_{\text{40 fois}} \underbrace{0,\ldots,0,}_{\text{10 fois}} \underbrace{1,\ldots,1}_{\text{10 fois}} \}$$

premier groupe — deuxième groupe

$$\hat{r}^2 = 0.4286 \qquad \hat{r}_S = 0$$

Dans le premier groupe : $\hat{r}^2 = 0$

Dans le deuxième groupe : $\hat{r}^2 = 0$

Les nouvelles mesures

$$\Delta^t_{M_1,1} = \{ \underbrace{1,\ldots,1,}_{\text{50 fois}} \underbrace{1,\ldots,1,}_{\text{10 fois}} \underbrace{0,\ldots,0}_{\text{50 fois}} \}$$

clônes

$$\Delta^t_{M_2,1} = \{ \underbrace{1,\ldots,1,}_{\text{50 fois}} \underbrace{1,0,\ldots,1,0,}_{\text{30 fois}} \}$$

clônes

$$\hat{r}^2 = 0.217 \qquad \hat{r}_V = 0.001$$

En ne gardant qu'un seul des clônes : $\hat{r}^2 = 0.001$

Bien estimer le DL, pourquoi ?

Pour limiter le génotypage

Deux SNP en fort DL apportent une information redondante, il n'est donc pas d'un grand intérêt de les génotyper tous les deux.

Pour "couvrir" toute la région génomique d'intérêt

L'objectif est que tous les SNP non génotypés soient "couverts" par au moins un SNP génotypé en fort DL.

Tester l'association

Comme pour la cartographie de gènes par analyse de liaison, deux cas :

les phénotypes de maladie

le dispositif cas-contrôle

les phénotypes continus

- modèle simple
- modèle corrigé des effets de la structure et de l'apparentement

Le dispositif cas-contrôle

Marqueur		M_1		Total
Allèles		i		
Cas malade		N_{di}		N_{d+}
Contrôle sain		N_{si}		N_{s+}
Total		N_{+i}		N

Utilisation de r^2

Le phénotype "maladie" peut être vu comme un marqueur un peu particulier. Au lieu de la dose allélique à ce marqueur

$$\Delta_{I^m} = \begin{cases} 1 & \text{si malade} \\ 0 & \text{si sain} \end{cases}$$

On estimera alors $r_i^2 = Cor^2(\Delta_{I^m}, \Delta_{M_1,i})$ par

$$\hat{r}_i^2 = \widehat{Cor}^2(\Delta_{I^m}, \Delta_{M_1,i})$$

Sous l'hypothèse H_0 : { pas d'association allèle i avec la maladie }

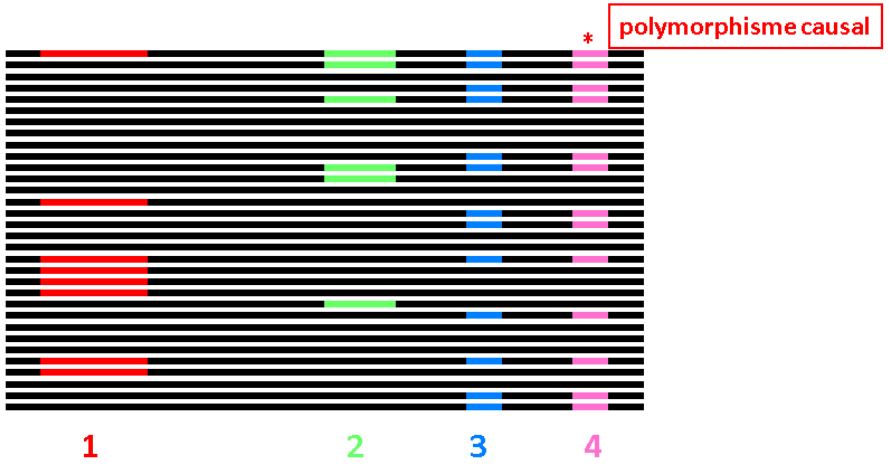
$$\hat{r}_i^2 \sim \chi^2 \text{ à 1 ddl}$$

L'association du marqueur avec la maladie peut aussi être testée en utilisant le test classique du χ^2 dans une table de contingence.

Puissance et mesure r^2

Pritchard & Przeworski, Am. J. Hum.Genet., 2001

Pour des marqueurs biallélliques



r^2_{14} r^2_{24} r^2_{34} Mesure r^2 de DL

N_1 N_2 N_3 N_4 Nbre d'observations

Pour avoir la même puissance, qu'au locus causal

$$N_i = N_4 / r^2_{i4}$$

Modèle linéaire

Effet du SNP fixe

Cas d'individus homozygotes, sans donnée manquante pour le génotype : Y_n

$$\nearrow \quad \text{SNP}_n^l = 1 \quad Y_{1k} = \mu + \theta^l + \epsilon_{1k}$$

$$\searrow \quad \text{SNP}_n^l = 0 \quad Y_{0k} = \mu + \epsilon_{0k}$$

Ce modèle se généralise aux cas de génotypes manquants inférés ou imputés, ainsi qu'aux individus hétérozygotes en choisissant une modélisation additive de l'effet du SNP, ou une modélisation avec effet de dominance.

Les limites du modèle

- postulat du modèle : les observations sont indépendantes
- objectif de l'analyse : rechercher les SNPs qui sont causaux

Modèle linéaire mixte

Effet du SNP fixe + stucture fixe + covariance génétique

Cas d'individus homozygotes, sans donnée manquante pour le génotype : Y_n

$\nearrow$ $\text{SNP}_n^I = 1$ $Y_n = \mu + S_n\beta + \theta^I + G_n + \epsilon_n$

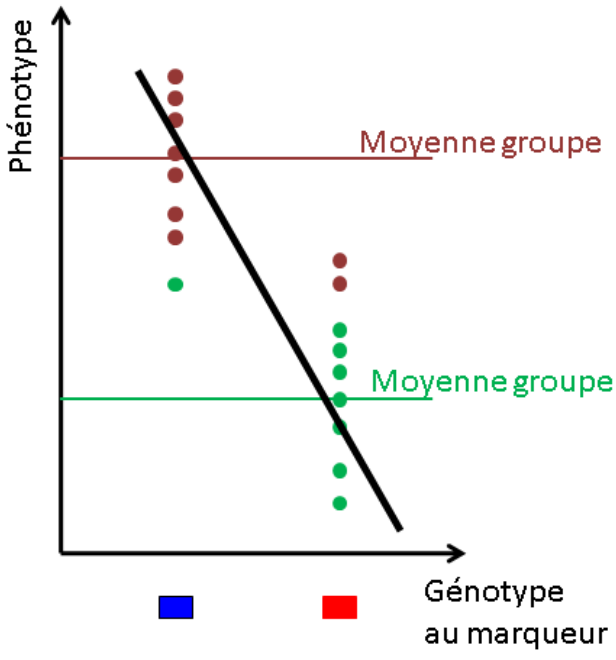
$\searrow$ $\text{SNP}_n^I = 0$ $Y_n = \mu + S_n\beta + G_n + \epsilon_n$

- S_n est la ligne correspondant à l'individu n dans la matrice de structure S
- G_n est un valeur génétique de n. Soit $G^t = (G_1, \ldots, G_n, \ldots, G_N)$, $Var(G) = \sigma_G^2 \Sigma_G$

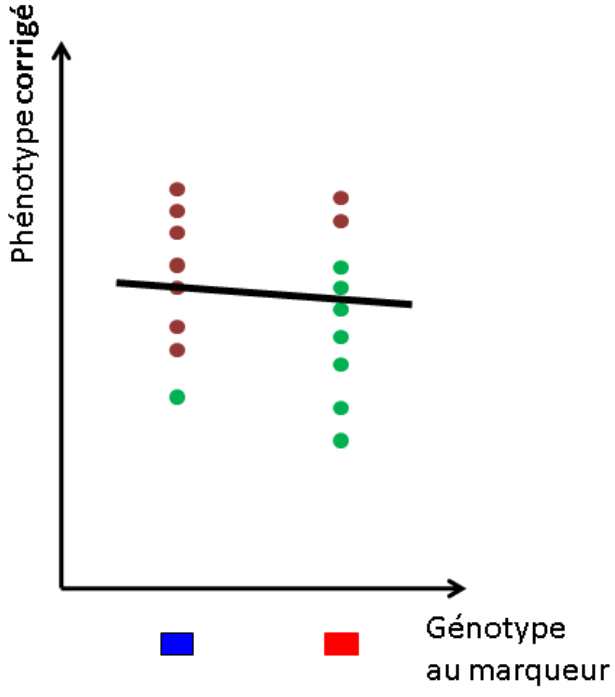
$$Var(Y) = \sigma_G^2 \Sigma_G + \sigma_\epsilon^2 Id$$

mêmes remarques pour la généralisation du modèle que pour le modèle simple

Les phénotypes continus : un modèle plus réaliste



Modèle simple



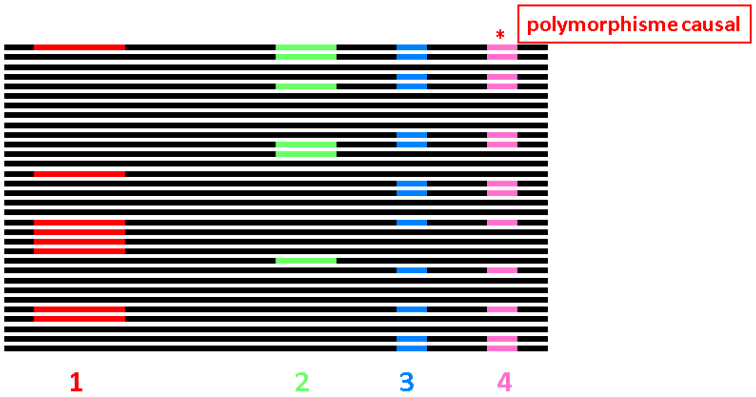
Modèle avec structure en groupe

La valeur des tests sur les marqueurs corrélés à la structure diminue.

Puissance et mesure r^2_{VS}

Mangin et al., Heredity, 2011

Pour des marqueurs biallélliques



Pour avoir la même puissance, qu'au locus causal

$$N_i = N_4 / r^2_{VS_{i4}}$$

avec

$$V = \sigma^2_G \Sigma_G + \sigma^2_\epsilon Id$$

Le modèle réaliste est un modèle linéaire mixte

$$Y_n = \mu + S_n\beta + SNP_n^l\theta^l + G_n + \epsilon_n$$

Deux types de paramètres à estimer

- les paramètres entrant dans la variance de Y (composantes de la variance)
- les paramètres des effets fixes

Les composantes de la variance σ^2_G et σ^2_ϵ

Elles sont estimées par ML (Maximum Likelihood) ou REML (Restricted ML)
Si on parle de vraisemblance, c'est que Y_n a une loi connue.
Cette loi est une Gaussienne $\Rightarrow$ G_n et ϵ_n sont aussi Gaussiens.
C'est un postulat nécessaire pour le modèle mixte.

Les composantes de la variance σ^2_G et σ^2_ϵ

Les estimateurs du maximum de vraisemblance sont des estimateurs biaisés (leur espérance n'est pas égale aux paramètres qu'il estiment).

le REML

Méthode qui consiste à estimer par maximum de vraisemblance mais après avoir projeté Y sur l'espace orthogonal pour V aux effets fixes

La différence entre ML et REML

..... juste une question de dénominateur ....

Exemple : $Y_n = \mu + \epsilon_n$ pour $n = 1, \ldots, N$

$$\hat{\sigma}^{2ML}_\epsilon = \frac{\sum(Y_n - \hat{\mu})^2}{N} \qquad \hat{\sigma}^{2REML}_\epsilon = \frac{\sum(Y_n - \hat{\mu})^2}{N-1}$$

Algorithme de ML et/ou REML

Il n'existe pas de formule analytique pour calculer les estimateurs du ML ou du REML.

Les algorithmes qui résolvent cette question de maximisation, atteignent le maximum par itération successives. Ils prennent beaucoup de temps CPU, en particulier car la matrice de covariance génétique Σ_G doit être inversée. Et ils sont longs à converger.

Une autre approche consiste à ne pas maximiser la vraisemblance mais seulement une approximation de la vraisemblance, plus simple à maximiser. Cette approximation a cependant la propriété d'être équivalente asymptotiquement à la vraisemblance. Elle permet le "passage à l'échelle" c'est-à-dire tester des millions de SNP.

Tester l'effet d'un SNP dans $Y_n = \mu + S_n\beta + SNP_n^l\theta^l + G_n + \epsilon_n$

test de Wald

Le principe est de faire comme si la variance des observations était connue et d'utiliser l'estimateur et sa variance classiquement obtenus par les moindres carrés généralisés
Pour estimer les composantes de la variance on utilise le REML.
La variance "supposée connue" est

$$\widehat{V} = \hat{\sigma}_G^{2^{REML}} \Sigma_G + \hat{\sigma}_\epsilon^{2^{REML}} Id$$

Le test au locus l $\quad \frac{(\hat{\theta}^l - \theta^l)^2}{var(\hat{\theta}^l)}$
suit asymptotiquement une loi de $\chi^2(1)$ sous l'hypothèse H_0 : { pas d'association }

Tester l'effet d'un SNP dans $Y_n = \mu + S_n\beta + SNP_n^l\theta^l + G_n + \epsilon_n$

test du rapport de vraisemblance (ML)

Pour faire ce test on utilise le ML jamais le REML

$$RV = \frac{sup_{\mu,\beta,\theta^l=0,\sigma^2_G,\sigma^2_\epsilon} V(Y;\mu,\theta^l=0,\sigma^2_G,\sigma^2_\epsilon)}{sup_{\mu,\beta,\theta^l,\sigma^2_G,\sigma^2_\epsilon} V(Y;\mu,\theta^l,\sigma^2_G,\sigma^2_\epsilon)}$$

$-2\ln(RV)$ suit asymptotiquement une loi de $\chi^2(1)$ sous l'hypothèse H_0 : { pas d'association }

Des millions de tests ....

Tests multiples non indépendants

Que ce soit pour les phénotypes binaires ou continus, un test d'association est affectué par SNP. Chacun de ces tests sous l'hypothèse H_0 : { pas d'association } est comparé un χ^2 à 1 degré de liberté. Mais d'un SNP à un autre, les tests ne sont pas indépendants.

Bonferroni trop conservateur

$\Longrightarrow$ On ne peux pas utiliser la correction de Bonferroni car elle conduit à un test beaucoup trop conservateur, donc très très peu de puissance, donc pas de détection.

FDR

$\Longrightarrow$ On utilise le FDR pour contrôler le taux de faux positifs.

Lorsque le pedigree est connu

Le coefficient d'apparentement (coancestry)

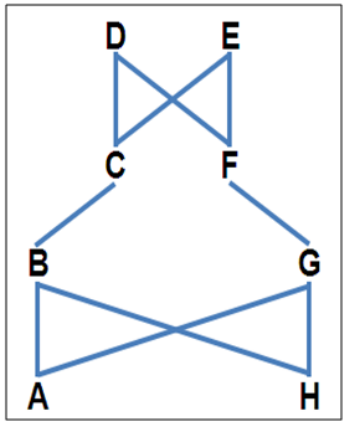
Weir et al., Nature Rev Genet, 2006

$$\rho_{ij} = \sum_a (1 - F_a)(1/2)^{n_a}$$

a ancêtre commun de i et j

n_a le nbre d'individus du pedigree sur le chemin le plus court de i à j en passant par a

F_a coefficient de consanguinité de a



Exemple

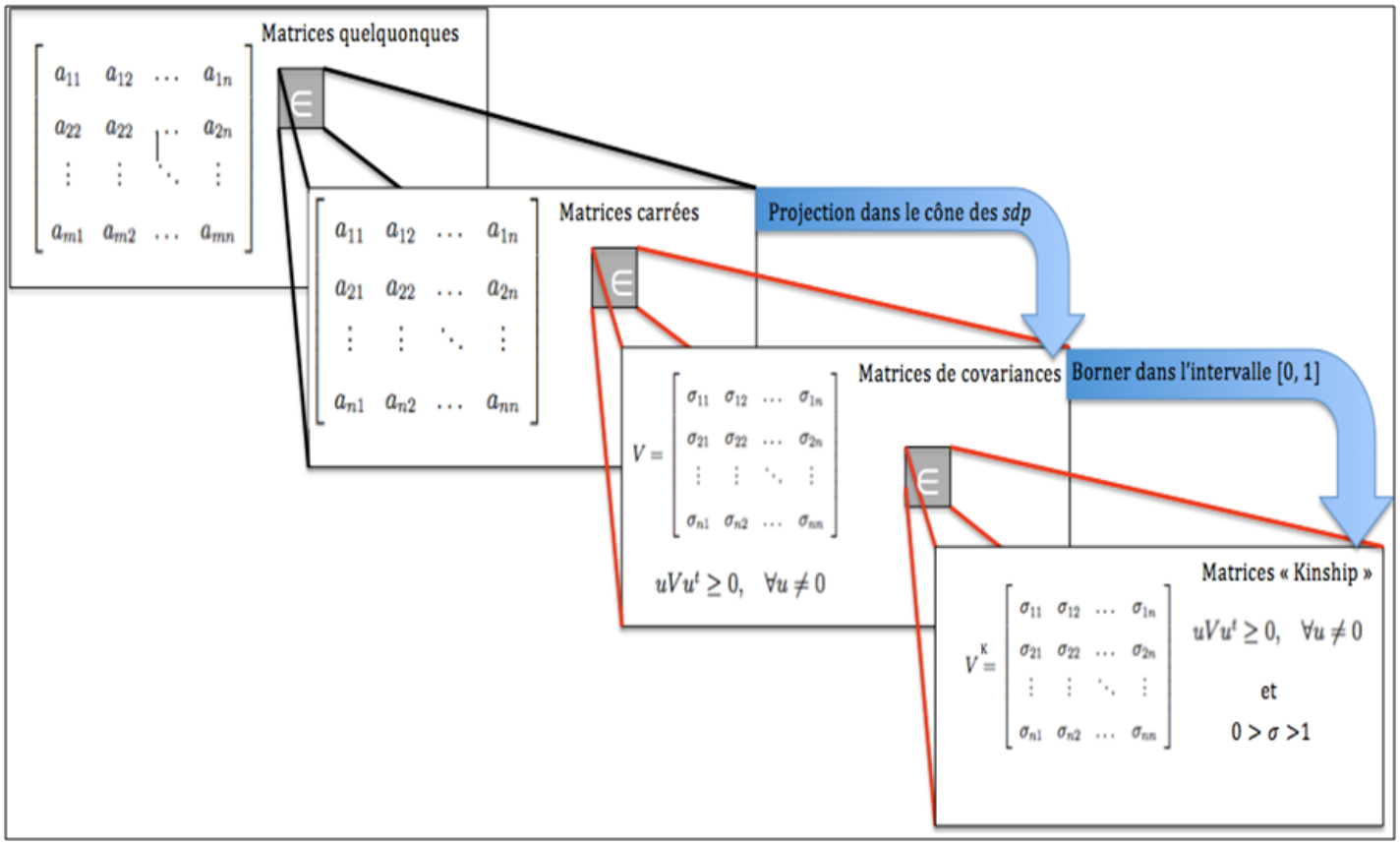
Si D et E non consanguins, $\rho_{AF} = 2(1/2)^5$

Lorsque la covariance génétique est estimée avec les marqueurs

Estimateur	Principe	Plusieurs Populations	Matrice Kinship
AIS	Proba (IBS)	✖	
BNO	Proba(IBD)=Proba (IBS)-correction	✔	
WAIS	Proba(IBD)=Proba (IBS)-correction	✔	
LOI	Proba(IBD)=Corrélation des fréquences alléliques	✖	
MIL	Max de vraisemblance	✖	

Software:CoCoa, Maenhout et al., Bioinformatics, 2009

Matrice d'apparentement, (kinship)



Une matrice *semi-définie positive (sdp)* dont les éléments sont compris entre 0 et 1

La structure

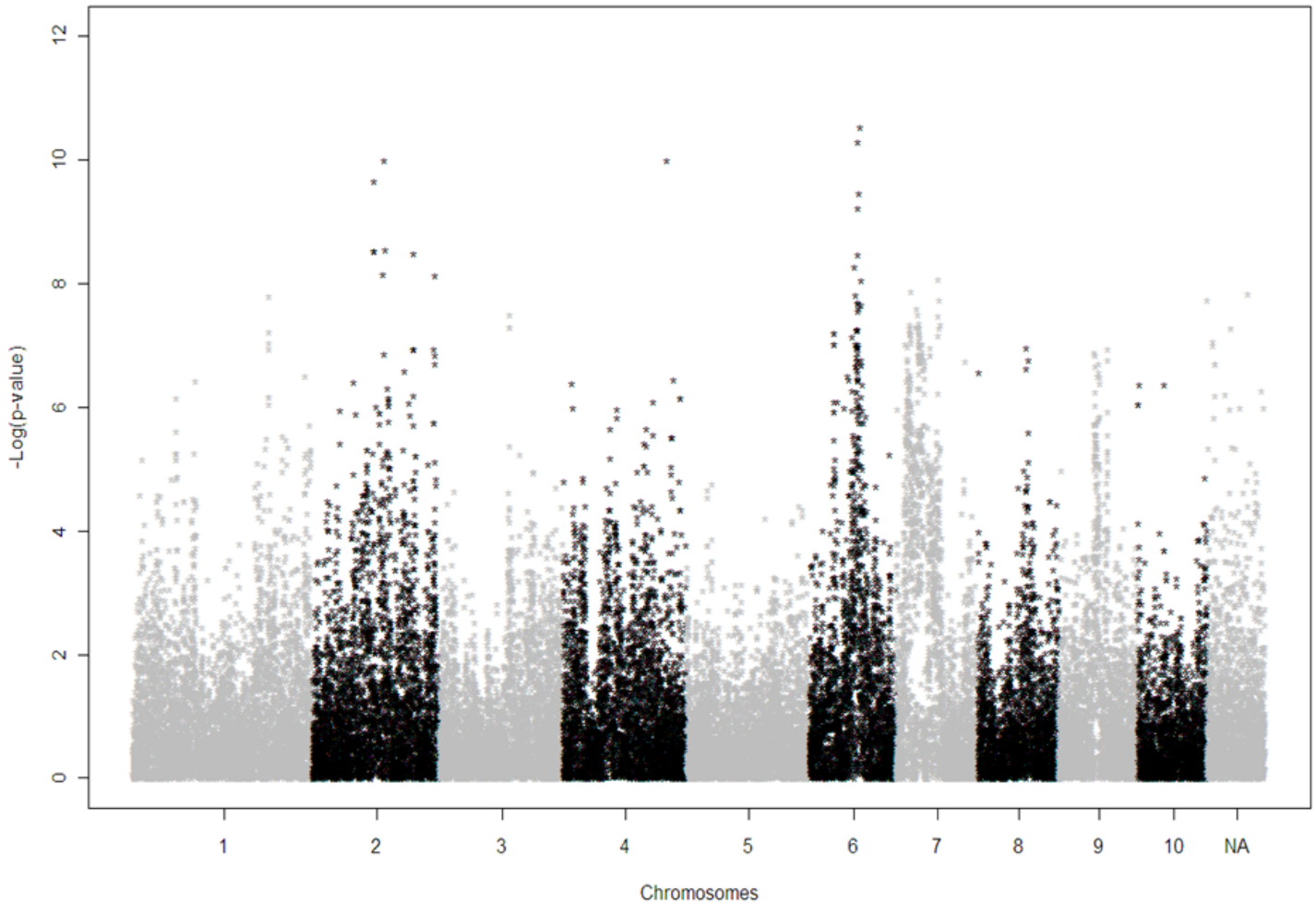
Voir la partie du cours de “génétique des populations”

Les logiciels

- ASREML, générique pour les modèles mixtes, dans R, maximise la vraisemblance, test du rapport de vraisemblance
- EMMA, spécifique de "association mapping", dans R, maximise une approximation de la vraisemblance, test de Wald
- Tassel, spécifique de "association mapping", java, propose les 2 types de maximisations et de tests
- Plink, spécifique de "association mapping", pas de modèle mixte, plutôt spécifique des applications en génétique humaine

Résultats d'un modèle

Manhattan plot



Résultats de plusieurs modèles

Différentes matrices de structure, différentes matrices de covariance génétique