Stratification de population

dans les études d'association de génome entier

Contexte

Étude de la composante génétique des maladies humaines

- Les méthodes utilisées avec succès pour l'étude des maladies mendéliennes (analyse de liaison) n'ont pas donné de bons résultats pour les maladies complexes (pathologies dysimmunitaires, psychiatriques, etc) ou les traits quantitatifs (taux de cholestérol, tension artérielle, etc)

- Apparition d'une technologie de génotypage massive, le SNP array (puce de génotypage)
  1998 : 1494 SNPs, 2003 : 10 000 SNPs, 2006 : 500 000 SNPs...

☞ Étude d'association avec le génome entier, ou GWAS : Genome Wide Association Studies

Contexte

Genome Wide Association Studies

- Les SNPs sont des polymorphismes d'un nucléotide

```
...CGCGCGCGTTACA...
...CGCGCGCGTTGCA...
```

  ☞ SNP A/G ; trois génotypes possibles AA, AG, GG

- Adoption d'une méthodologie très simple : tester l'association entre la maladie étudiée et tous les SNPs, un à un
  - Les trois génotypes sont recodés 0, 1, 2 et la variable est traitée comme quantitative (« effet dose »)
  - On ne suppose pas que les polymorphismes causaux font partie des SNPs génotypés, mais qu'ils sont corrélés (en « déséquilibre de liaison ») avec certains de ceux-ci

☞ Tests multiples ; le standard de publication est $p < 5 \cdot 10^{-8}$

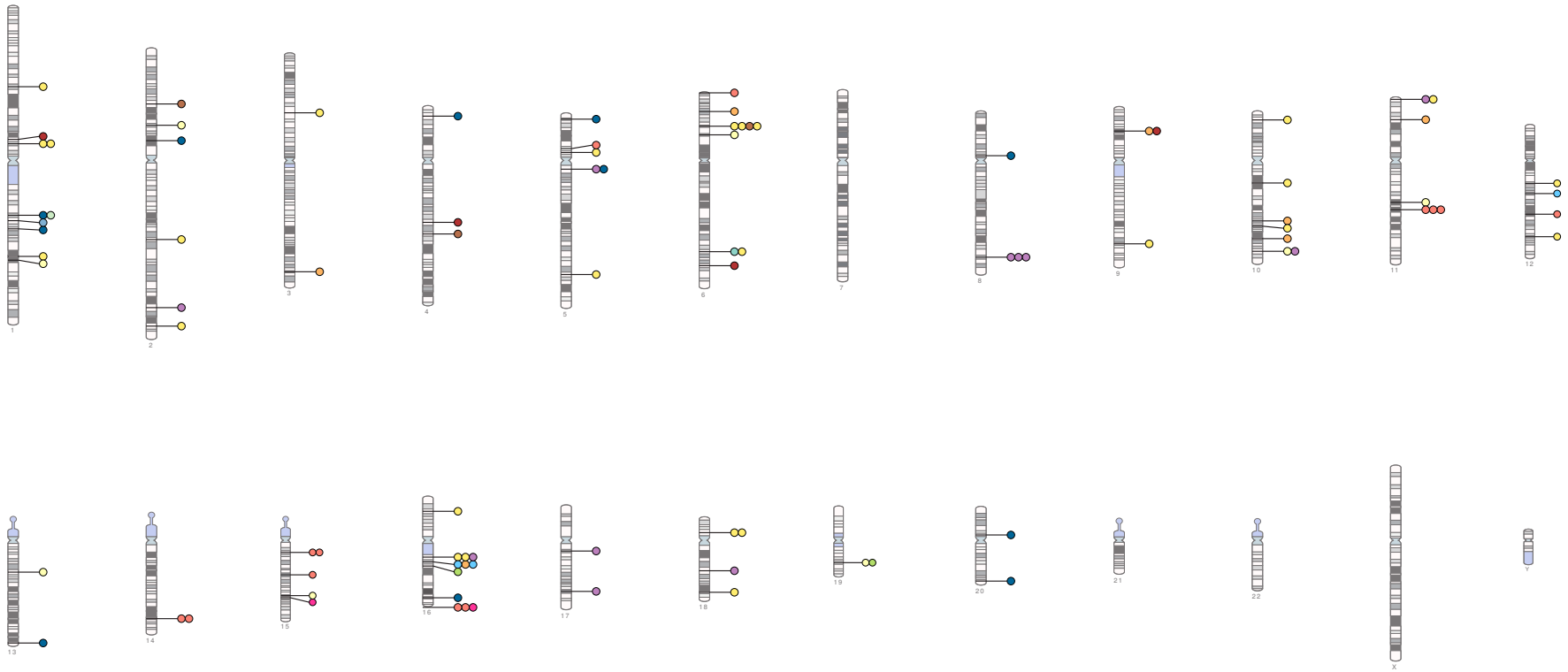
Résultats des études GWAS (fin 2005)

Klein et al, Science (2010)

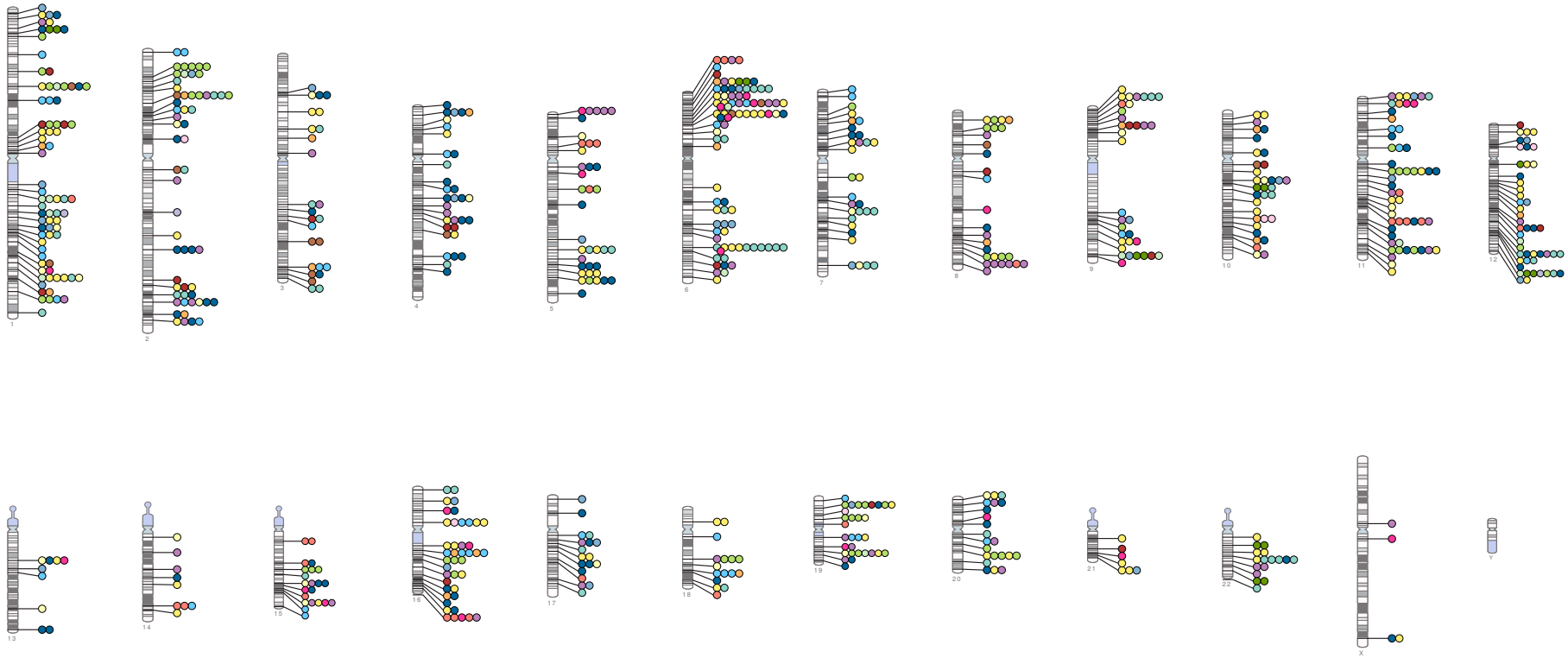
GWAS sur la DMLA, 96 cas et 50 contrôles, puce de 100K

☞ un SNP avec un OR de 7.4 (CFH)

Résultats des études GWAS (fin 2007)

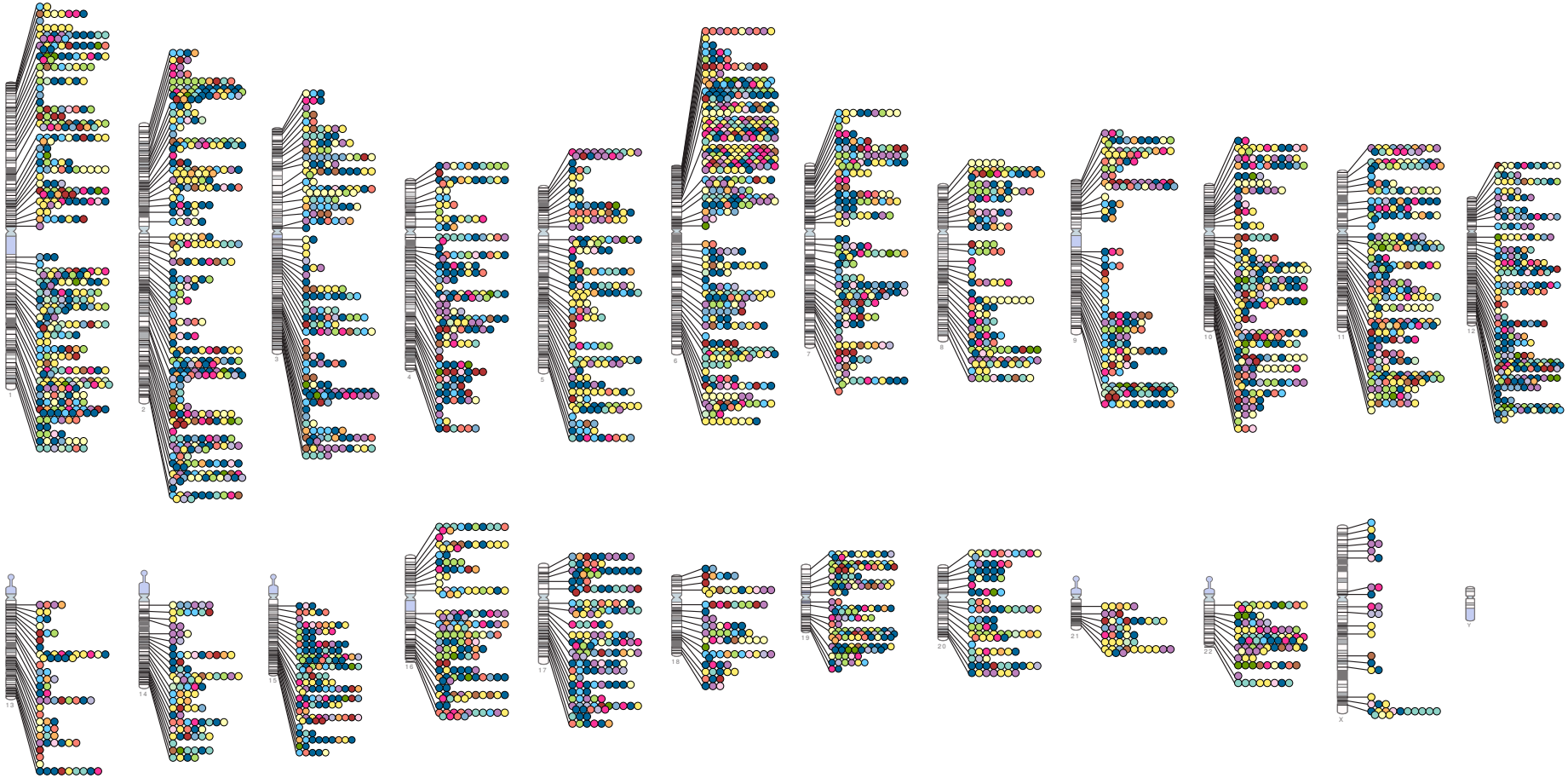


Résultats des études GWAS (fin 2009)



Résultats des études GWAS (fin 2011)

Résultats des études GWAS (fin 2013)



Résultats des études GWAS (juin 2017)

Stratification de population

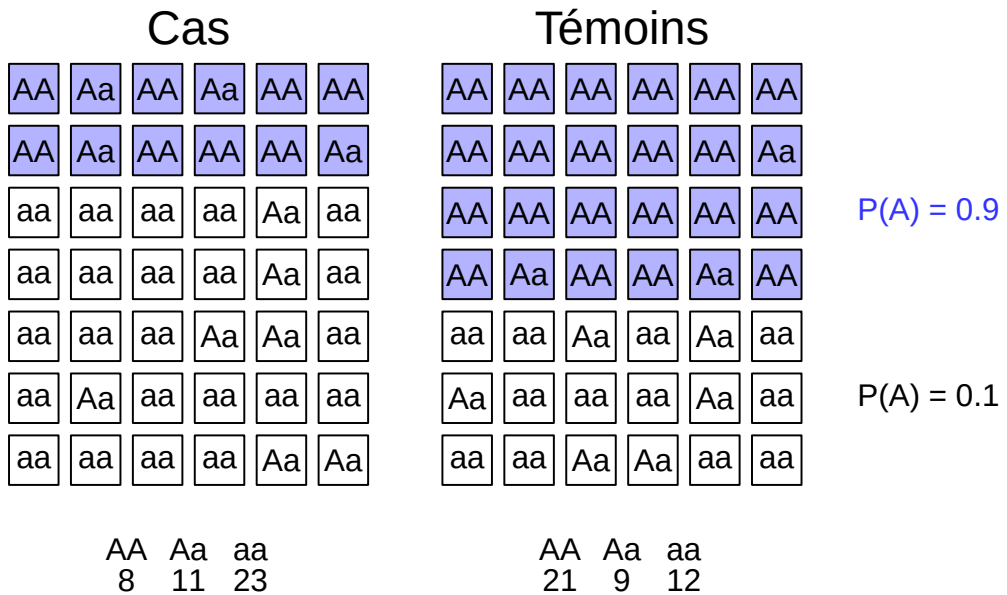
Pour identifier des SNPs avec des effets faibles, les GWAS ont besoin de grandes tailles d'échantillon

☞ recrutement d'un grand nombre de patients dans des études polycentriques, couvrant souvent sur plusieurs pays ;

☞ la population de recrutement n'est pas homogène ; l'existence d'une structure de population, ou stratification de population, biaise les tests d'association

Stratification de population

Étude cas/témoins où le recrutement a lieu dans une population stratifiée :



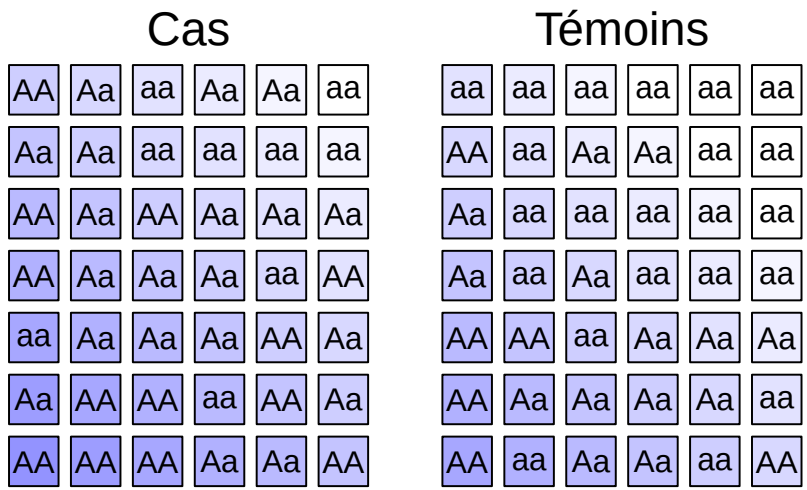
☞ la sous-population est une variable de confusion

Un contrôle qualité soigné peut limiter le problème, mais ça n'est pas suffisant

Étude cas/témoins et stratification de population

La solution évidente est de prendre en compte la strate dans l'analyse. Mais...

- la stratification peut être « invisible » (ou « cryptique »)
- le mélange de population peut être « continu »...



Possibilité d'avoir un facteur environnemental impliqué dans la maladie et corrélation gène-environnement (e.g. gradient nord/sud).

☞ Que faire ?

Que faire ?

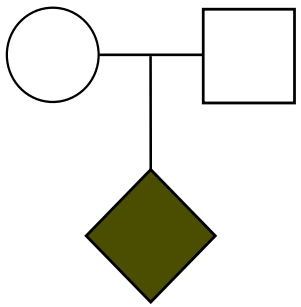
Deux grandes voies sont possibles :

- Test d'association sur données familiales (TDT / FBAT)
- Estimer la structure de la population à partir des données et la prendre en compte dans l'analyse
  - ☞ Analyse en composantes principales des données génomiques
  - ☞ Régression sur les composantes principales / Modèle mixte

Le Transmission Disequilibrium Test (TDT)

Le Transmission Disequilibrium Test

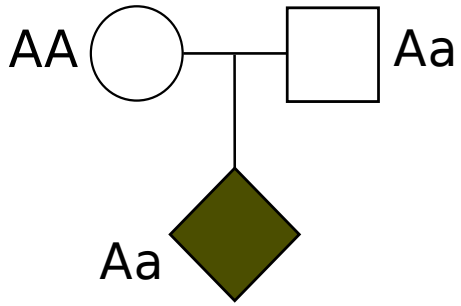
Le TDT traite des familles trio :



Deux parents et un enfant atteint.

Le Transmission Disequilibrium Test

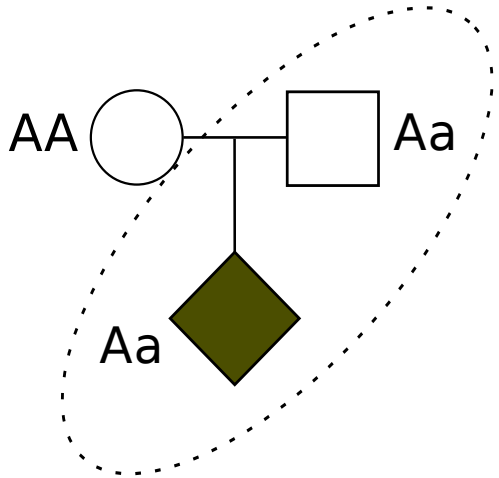
Le TDT traite des familles trio :



Deux parents et un enfant atteint. Leurs génotypes à un marqueur di-allélique, d'allèles A et a.

Le Transmission Disequilibrium Test

Le TDT traite des familles trio :



Deux parents et un enfant atteint. Leurs génotypes à un marqueur di-allélique, d'allèles A et a.

L'événement informatif est la transmission de A ou a par un parent hétérozygote.

Sous l'hypothèse nulle, A est transmis aussi souvent que a.

Le Transmission Disequilibrium Test

- Les parents homozygotes ne fournissent aucune information ;
- on compte le nombre de transmissions des allèles A et a par un parent hétérozygote.

génotypes			transmis	
père	mère	enfant	A	a
AA	Aa	AA	1	0
Aa	AA	Aa	0	1
AA	AA	AA	0	0
Aa	Aa	Aa	1	1
Aa	Aa	AA	2	0
⋮				
		Total	n_A	n_a

Le Transmission Disequilibrium Test

On compare ensuite le nombre de transmissions de A et a, respectivement n_A et n_a, au nombre attendu sous H_0 : « probabilité pour un parent Aa de transmettre A $= \frac{1}{2}$ ».

☞ test du χ^2

	A	a	Total
Observés	n_A	n_a	$n = n_A + n_a$
Attendus	$\frac{1}{2}n$	$\frac{1}{2}n$	n

$$\chi^2(1) = \frac{\left(n_A - \frac{1}{2}n\right)^2}{\frac{1}{2}n} + \frac{\left(n_a - \frac{1}{2}n\right)^2}{\frac{1}{2}n} = \frac{(n_A - n_a)^2}{n_A + n_a}.$$

Le TDT résiste à la stratification de population

Si dans la « population bleue » on a $\tau = \frac{1}{2}$ (la proba de transmettre A) ; si c'est également le cas dans la « population blanche » ; alors c'est toujours vrai dans le mélange des deux populations !

- dans une famille, il n'y a distorsion de la transmission que si le marqueur est lié à un locus maladie
- la distorsion n'est au profit du même allèle A (ou a) dans toutes les familles que si le marqueur, en plus d'être lié au locus maladie, est en déséquilibre de liaison avec lui.

☞ le TDT teste à la fois la liaison et l'association.

Remarques finales

Si on suppose que les risques relatifs sont multiplicatifs :

$$\mathbb{P}(\text{atteint}|Aa) = r \times \mathbb{P}(\text{atteint}|AA)$$
$$\mathbb{P}(\text{atteint}|aa) = r^2 \times \mathbb{P}(\text{atteint}|AA)$$

alors on peut estimer r par $\widehat{r} = \frac{n_a}{n_A}$.

Le cadre formel de la *régression logistique conditionnelle* permet de considérer d'autres modèles (dominant, récessif, etc) ou de réaliser des tests haplotypiques, etc.

FBAT permet de traiter des familles nucléaires plus générales.

La faiblesse du TDT est la nécessité de recruter les parents. Le coût du génotypage est multiplié par 1,5 par rapport à un design cas/contrôles équilibré ; seuls les parents hétérozygotes sont informatifs !

Analyse en Composantes Principales (ACP)

Réduction de dimension, contrôle qualité

ACP – réduction de dimension

- On considère une population de n individus, génotypés en p SNPs (par exemple $n = 1000$, $p = 500\,000\ldots$)
- Les génotypes sont codés $0, 1, 2$
- Les génotypes de l'individu i forment un vecteur de longueur p

$$a_i = (a_{i1}, a_{i2}, \ldots,) = (0, 1, 0, 0, 1, 2, 2, 0, \ldots)$$

☞ Matrice $A = [a_{ij}] \in \mathbb{R}^{n \times p}$ des génotypes (une ligne = un individu, une colonne = un SNP).

Ou encore : chaque individu est un point dans un espace de dimension $n = 500\,000\ldots$

ACP – réduction de dimension

- On transforme la matrice des génotypes en « standardisant » chaque colonne :
  si q_j est la fréquence de l'allèle alternatif du SNP j, on remplace a_{ij} par

$$g_{ij} = \frac{a_{ij} - 2q_j}{\sqrt{2q_j(1-q_j)}}$$

  (génotype centré réduit).

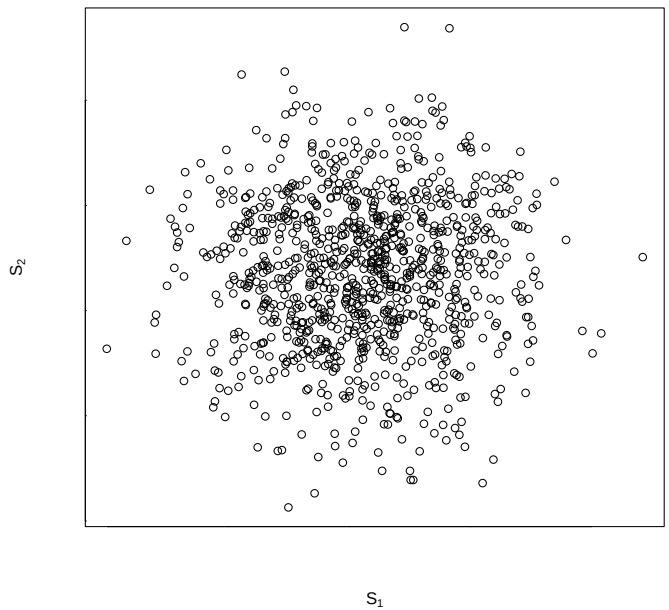
☞ Matrice $G = [g_{ij}] \in \mathbb{R}^{n \times p}$ des génotypes standardisés

ACP – réduction de dimension

- But : représenter les individus dans un espace de dimension plus petite (par exemple dimension 2)
- On associe à chaque individu $i = 1, \ldots, N$ deux scores de la forme

$$S_i = \ell_1 g_{i1} + \ell_2 g_{i2} + \cdots + \ell_p g_{ip}$$

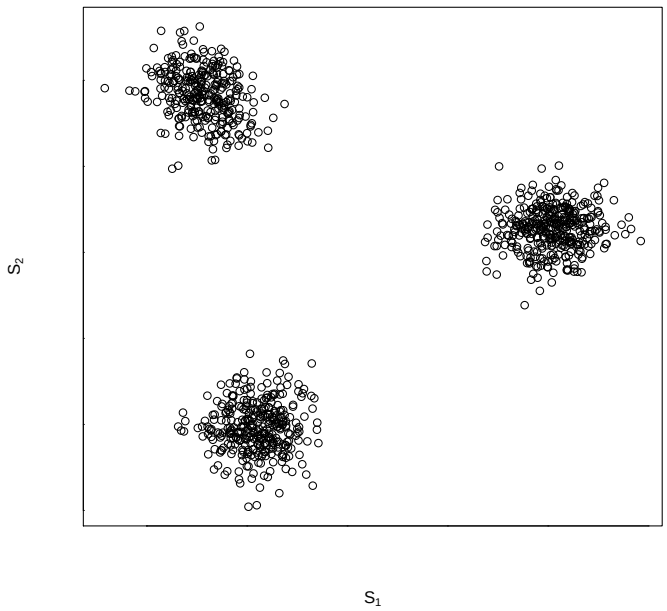
- On calcule deux tels scores S_{i1} et S_{i2}, et on fait le nuage des points (S_{i1}, S_{i2}).
  Si on choisit les ℓ_i au hasard, on obtient quelque chose comme ça :



ACP – réduction de dimension

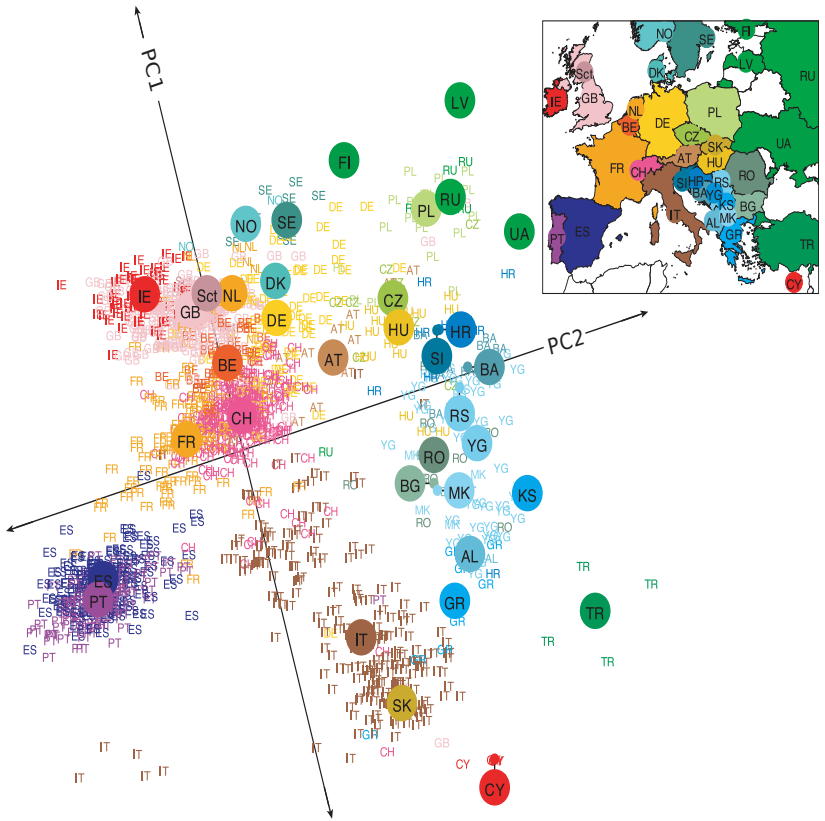
- Choisissons pour $\ell_1, \ldots, \ell_p$ les valeurs qui maximisent la dispersion des points, c'est-à-dire la variance des scores

  tout en imposant $\ell_1^2 + \cdots + \ell_p^2 = 1$
- Les scores s'appellent alors les *composantes principales* de la matrice des génotypes, et les ℓ_j sont les « *loadings* » (des SNPs).
- Notre population se révèle être un mélange de trois sous-populations distinctes.



ACP – réduction de dimension

Et voici le résultat sur un échantillon d'Européens...



Novembre 2008, Genes mirror geography within Europe

ACP – un peu de technique

La *Genetic Relationship Matrix (GRM)* est une matrice de corrélation génomique entre individus :

$$K = \frac{1}{p-1} GG'.$$

C'est une matrice symétrique positive. Ses coefficients $k_{ii'}$ sont des estimateurs de $2\Phi_{ii'}$, avec $\Phi_{ii'}$ le coefficient d'apparentement entre i et i'.

OK pour apparentés proches (premier, deuxième degré) et si la population est homogène.

La matrice K a des valeurs et vecteurs propres $\lambda_1 \geq \cdots \geq \lambda_n \geq 0 \in \mathbb{R}$ et $v_1, \ldots, v_n \in \mathbb{R}^n$, les v_i étant de norme 1 et orthogonaux 2 à 2.

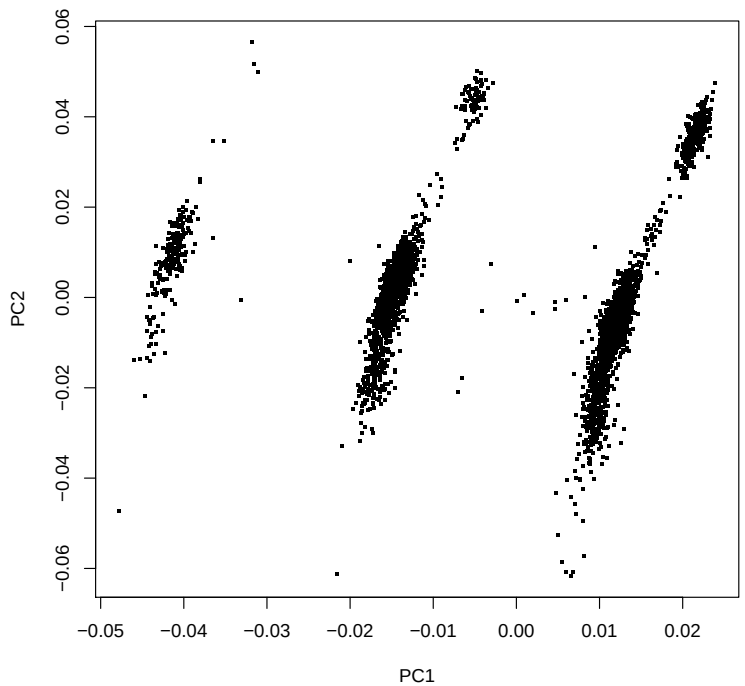
Les composantes principales sont les vecteurs $PC_i = \sqrt{\lambda_i} v_i$.

Les loadings peuvent se calculer par $L_i = \frac{1}{\sqrt{(p-1)\lambda_i}} G' v_i$.

ACP – Quelques écueils : LD thinning

Il ne faut pas utiliser la totalité du génome, mais de ne prendre que des SNPs en faible déséquilibre de liaison (*LD thinning* ou *LD prunning*)

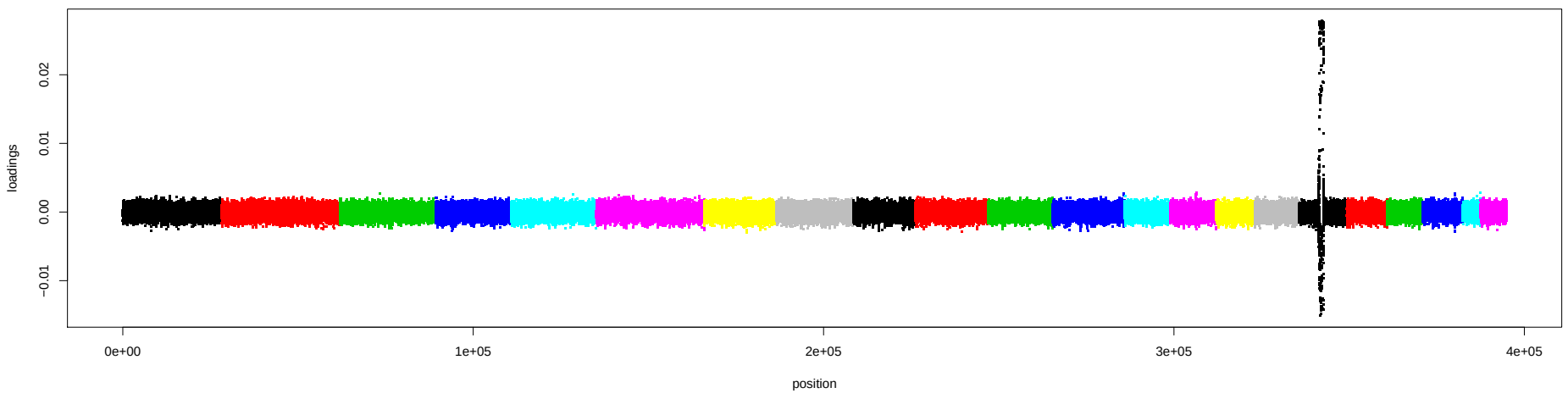
Ceci afin d'éviter que les premières composantes ne soient principalement déterminées par une région génomique où il y a un LD important.



ACP – Quelques écueils : LD thinning

Il ne faut pas utiliser la totalité du génome, mais de ne prendre que des SNPs en faible déséquilibre de liaison (*LD thinning* ou *LD prunning*)

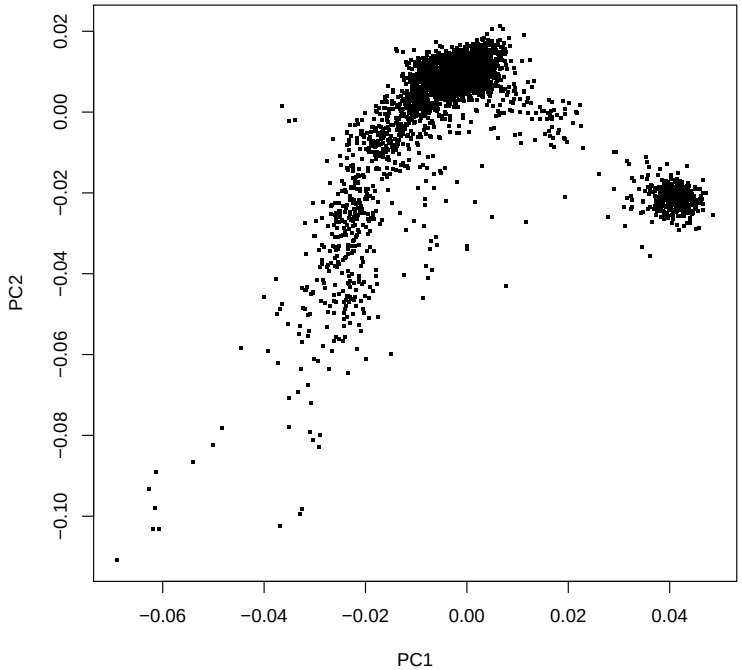
Ceci afin d'éviter que les premières composantes ne soient principalement déterminées par une région génomique où il y a un LD important.

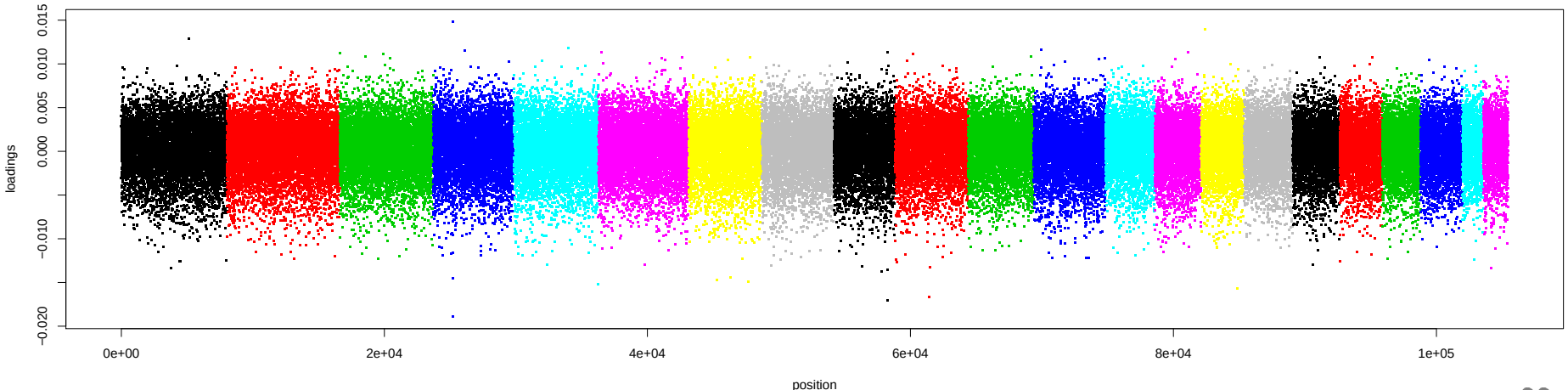


(Inversion commune sur le chromosome 17)

ACP – Quelques écueils : LD thinning

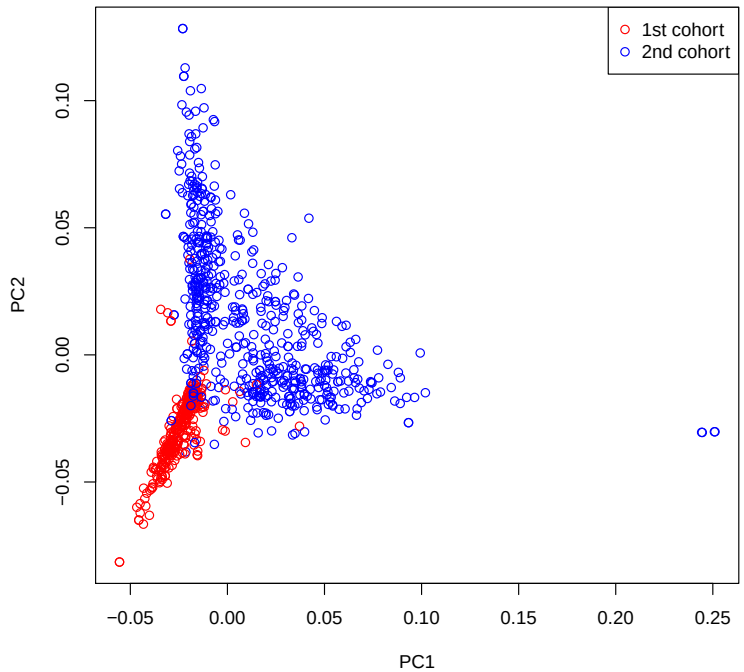
Après LD thinning :

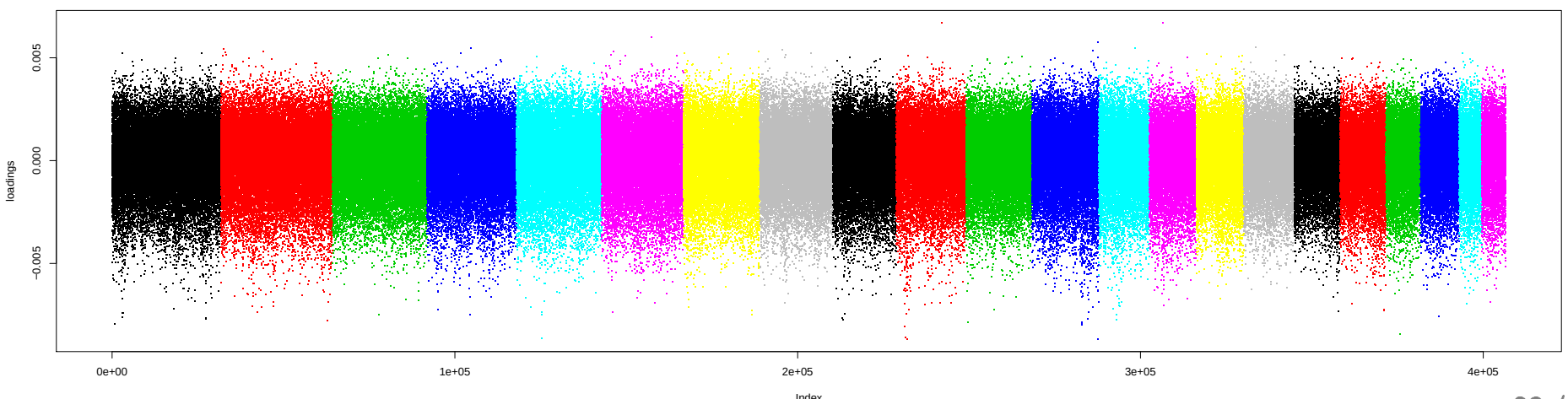




ACP – Quelques écueils : individus apparentés ou dupliqués

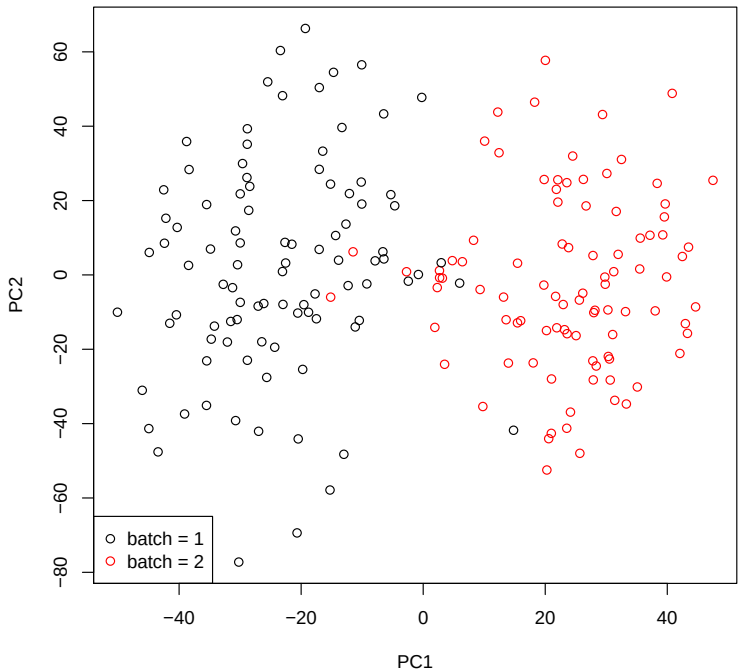
Les points à droite correspondent à deux frères tous les deux dupliqués

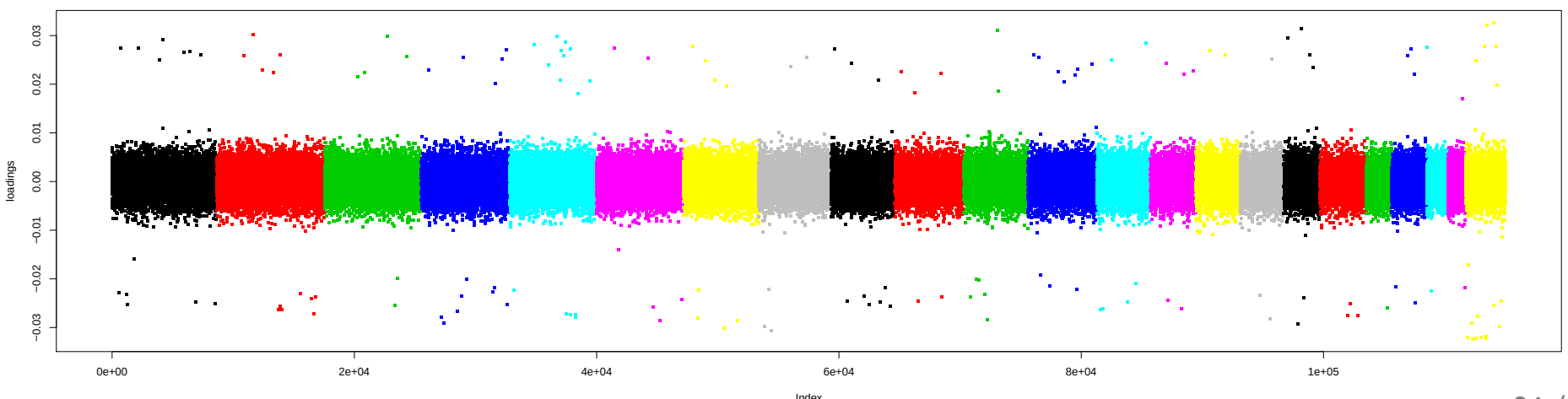




ACP – Quelques écueils : effet batch

Une poignée de SNPs « mal génotypés » sont déterminants dans l'ACP





Régression sur les Composantes Principales

Ajustement des études d'association

On peut inclure les premières composantes principales comme covariables dans une régression linéaire :

$$Y = \alpha_0 + \alpha_1 PC_1 + \cdots + \alpha_m PC_m + \beta X + \gamma SNP + \varepsilon$$

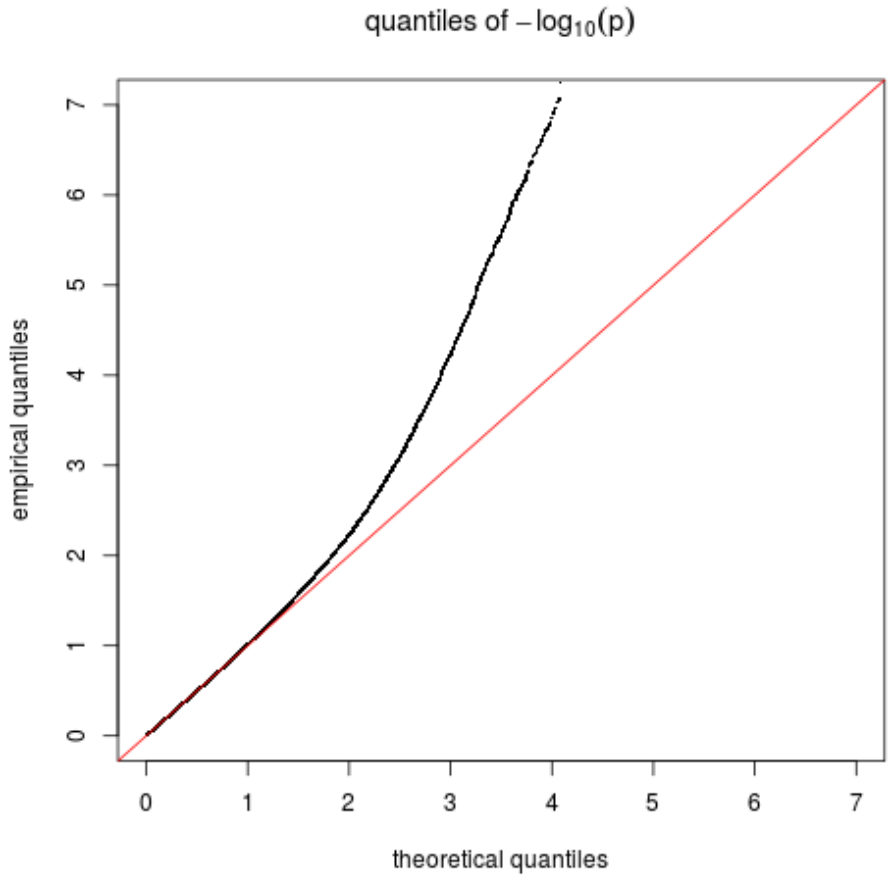
ou dans une régression logistique :

$$\text{logit}\, \mathbb{P}(Y = 1) = \alpha_0 + \alpha_1 PC_1 + \cdots + \alpha_m PC_m + \beta X + \gamma SNP.$$

(On prendra par exemple $m = 10$).

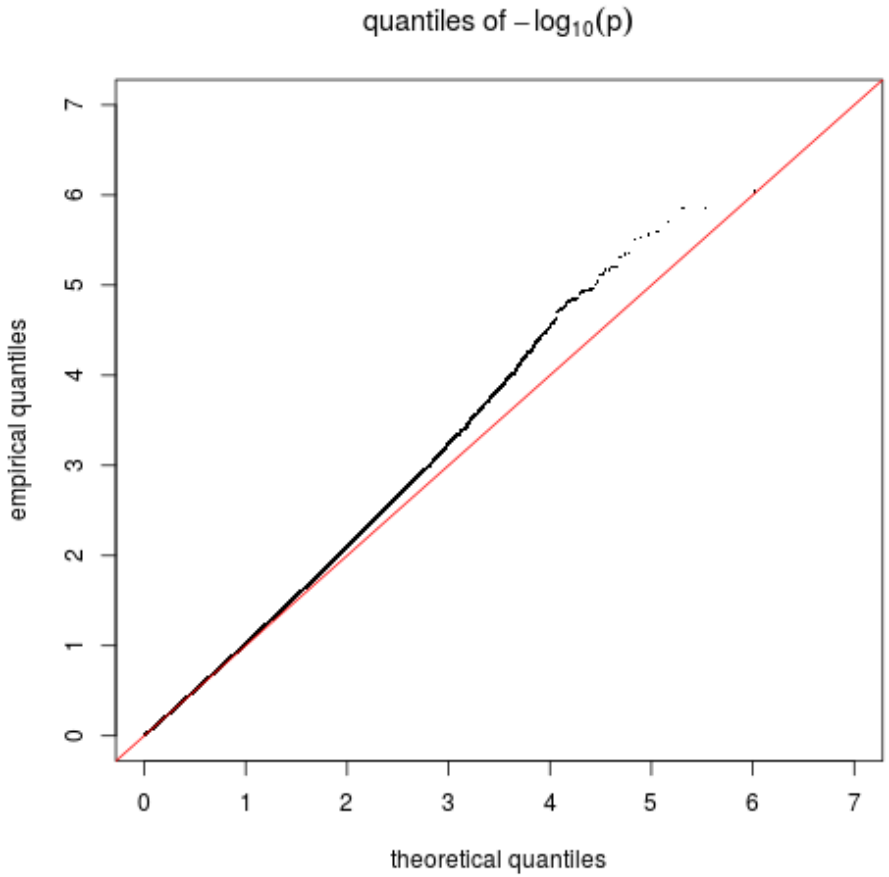
Examen visuel du graphique quantiles-quantiles

On ne prend pas de composantes principales en compte



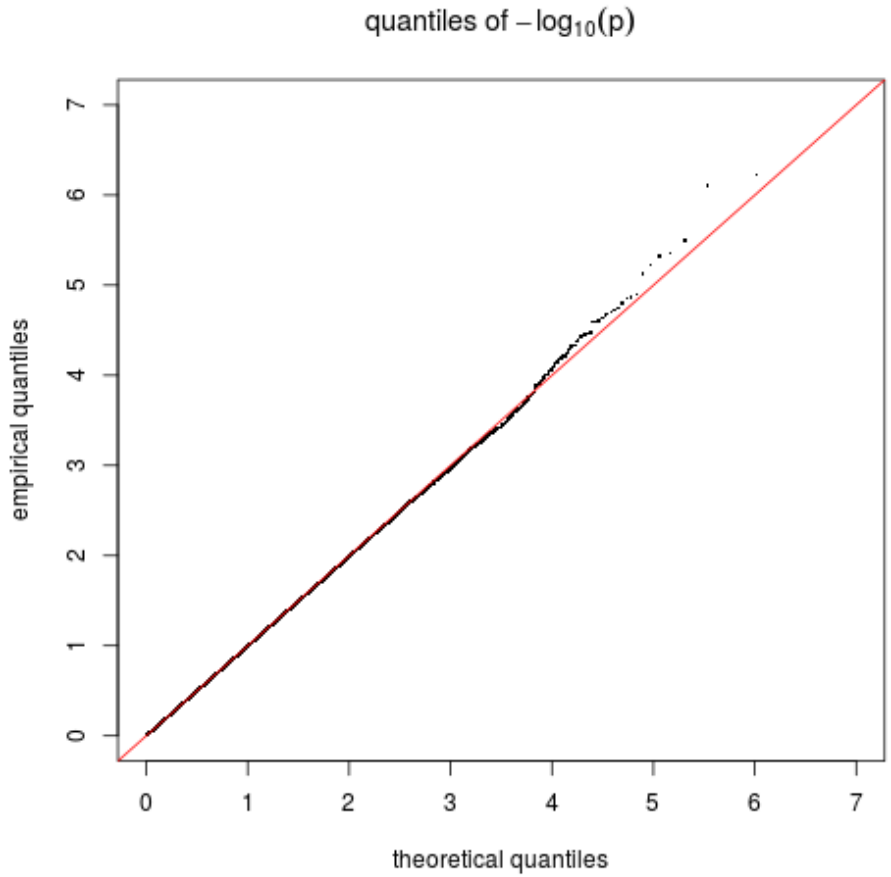
Examen visuel du graphique quantiles-quantiles

Trois composantes principales incluses



Examen visuel du graphique quantiles-quantiles

Cinq composantes principales incluses



Il peut être nécessaire d'inclure davantage de composantes principales...

Généralisation : le modèle mixte

Inclusion du génome dans le modèle

- On peut interpréter la régression sur les composantes principales comme une façon d'inclure un résumé du génome en covariable.
- Une autre façon de faire cela est d'inclure l'effet du reste du génome dans la partie aléatoire du modèle :

$$Y_i = \alpha + \beta X_i + u_i + \varepsilon_i \quad (i = 1, \ldots, n)$$

  - X_i est le génotype au SNP considéré
  - u_i est l'effet du reste du genome
  - ε_i est un terme d'erreur
  - On peut aussi inclure des covariables d'ajustement (sexe, etc)
- Forme matricielle

$$Y = \alpha + \beta X + u + \varepsilon$$

Inclusion du génome dans le modèle

$$Y = \alpha + \beta X + u + \varepsilon$$

- On suppose que les u_i sont normaux, mais ils ne sont pas indépendants : la covariance $\mathrm{cov}(u_i, u_j)$ sera proportionnelle à la covariance entre les génotypes des individus i et j, estimée sur tout le génome ; au niveau vectoriel

  $$\mathrm{var}(u) = \tau K$$

  où K est la GRM.

- les $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ soint indépendants entre eux et indépendants des u_i

Inclusion du génome dans le modèle

- Une formulation équivalente est

$$Y = \alpha + \beta X + v_1 PC_1 + \cdots + v_n PC_n + \varepsilon.$$

  Les effets $v_1, \ldots, v_n$ sont aléatoires, indépendants, de même loi $\mathcal{N}(0, \tau)$. On inclut ainsi la totalité des composantes principales ; le fait que les effets sont aléatoires permet de le faire sans avoir de sur-ajustement.

- Le même modèle est utilisé pour les estimations d'héritabilité

- Logiciels pour mettre ceci en œuvre : GEMMA / GCTA / package R gaston

Pou pou pidou