Gaston, un package R pour les données de génome entier

Claire Dandine-Roulland

16 Juin 2017

Package 'gaston'

May 25, 2017

Type Package

Title Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models

Version 1.5

Date 2017-05-24

Encoding UTF-8

Author Hervé Perdry & Claire Dandine-Roulland

Maintainer Hervé Perdry <herve.perdry@u-psud.fr>

Description

Manipulation of genetic data (SNPs), computation of Genetic Relationship Matrix, Linkage Disequilibrium, etc. Efficient algorithms for Linear Mixed Model (AIREML, diagonalization trick).

Sommaire

Single Nucleotide Polymorphism (SNP)

Un polymorphisme génétique est une variation de la séquence génétique. Cette variation se traduit par la présence de plusieurs versions dans la population d'un même locus, appelées allèles.

Les SNPs sont des polymorphismes d'une seule paire de base.

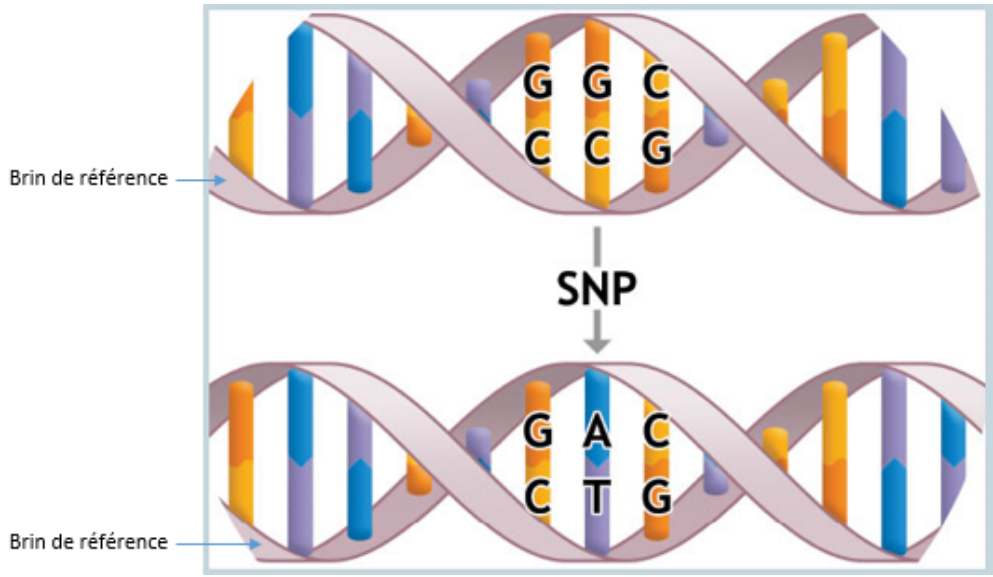


Figure – Single Nucleotide Polymorphism

Dans cet exemple, trois génotypes sont possibles : AA, AG et GG **codés 0, 1 et 2**.

Les données « Genome-Wide »

Pour une étude « Genome-Wide », nous disposons généralement de

Phénotype

$$n \text{ individus} \left\{ \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_n \end{bmatrix} \right.$$

Covariables

$$\underbrace{\begin{bmatrix} 1 & x_{11} & . & . & . & x_{k1} \\ 1 & x_{12} & . & . & . & x_{k2} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ 1 & x_{1n} & . & . & . & x_{kn} \end{bmatrix}}_{k}$$

Génotypes

$$\underbrace{\begin{bmatrix} 0 & 2 & 1 & 1 & . & . & . & 2 \\ 1 & 1 & NA & 2 & . & . & . & 0 \\ . & . & . & . & & & & . \\ . & . & . & . & & & & . \\ . & . & . & . & & & & . \\ 1 & 0 & 0 & 2 & . & . & . & 1 \end{bmatrix}}_{>500\,000}$$

Ainsi que des détails sur les SNPs observés.

Problème : Comment charger cette quantité de données dans R ?

bed.matrix

Gaston permet de charger dans R des données **génotypiques** (pour des SNPs diallélique) à partir de fichiers :

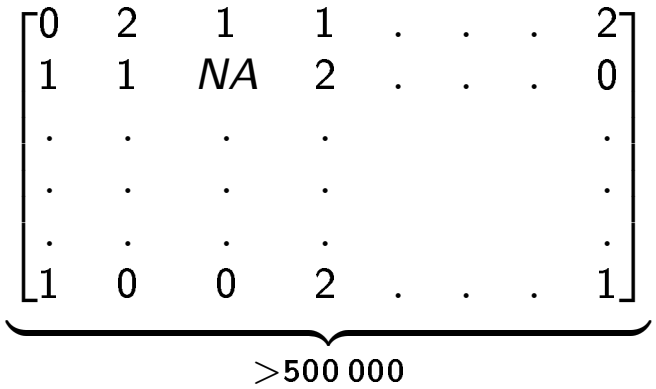
- VCF
- BED, BIM et FAM (fichiers au standard de PLINK).

Les données sont alors présentées sous la forme d'un objet de classe S4 appelé `bed.matrix` contenant un certain nombre de « slots » (accessible avec @) que je vais détailler au fur et à mesure.

Exemple :

```
# Lecture des fichiers TTN.bed, TTN.bim et TTN.fam
x <- read.bed.matrix('TTN')
x

## A bed.matrix with 503 individuals and 733 markers.
## snps stats are set
## ped stats are set

slotNames(x)

## [1] "ped"                  "snps"                 "bed"
## [4] "p"                    "mu"                   "sigma"
## [7] "standardize_p"        "standardize_mu_sigma"
```

Matrices de génotypes, @bed



Exemple

```
x@bed

## <pointer: 0x263ec40>
```

Remarque : le slot `bed` contient en fait un pointeur vers la matrice de génotypes stockée en mémoire. Dans le but d'optimiser la manipulation de jeux de données importants, les données sont stockées de façon compacte (chaque génotype sur 2 bits).

Afin de faciliter la manipulation de cet objet, il est directement possible de :

- sélectionner une « sous `bed.matrix` » en utilisant les `[]`
- au contraire, fusionner deux `bed.matrix` avec les fonctions `cbind` et `rbind`
- convertir une `bed.matrix` en une matrice numérique (`as.matrix`)
- convertir une matrice numérique en une `bed.matrix` (`as.bed.matrix`)
- multiplier une `bed.matrix` par un vecteur ou une matrice (`%*%`).

Statistiques descriptives pour les individus, @ped

```
head(x@ped)
```

```
##     famid      id father mother sex     pheno  N0  N1  N2 NAs N0.x N1.x
## 1 HG00096 HG00096      0      0   0 2.4795014 128  82 523   0    0    0
## 2 HG00097 HG00097      0      0   0 3.9356993 109  81 543   0    0    0
## 3 HG00099 HG00099      0      0   0 0.7413937  75 154 503   1    0    0
## 4 HG00100 HG00100      0      0   0 4.5435488 148  86 499   0    0    0
## 5 HG00101 HG00101      0      0   0 0.5878363  18 394 320   1    0    0
## 6 HG00102 HG00102      0      0   0 3.6305912  50 180 503   0    0    0
##   N2.x NAs.x N0.y N1.y N2.y NAs.y N0.mt N1.mt N2.mt NAs.mt  callrate
## 1    0     0    0    0    0     0     0     0     0      0 1.0000000
## 2    0     0    0    0    0     0     0     0     0      0 1.0000000
## 3    0     0    0    0    0     0     0     0     0      0 0.9986357
## 4    0     0    0    0    0     0     0     0     0      0 1.0000000
## 5    0     0    0    0    0     0     0     0     0      0 0.9986357
## 6    0     0    0    0    0     0     0     0     0      0 1.0000000
##          hz callrate.x hz.x callrate.y hz.y callrate.mt hz.mt
## 1 0.1118690        NaN  NaN        NaN  NaN         NaN   NaN
## 2 0.1105048        NaN  NaN        NaN  NaN         NaN   NaN
## 3 0.2103825        NaN  NaN        NaN  NaN         NaN   NaN
## 4 0.1173261        NaN  NaN        NaN  NaN         NaN   NaN
## 5 0.5382514        NaN  NaN        NaN  NaN         NaN   NaN
## 6 0.2455662        NaN  NaN        NaN  NaN         NaN   NaN
```

Statistiques descriptives pour les SNPS, @snps

```
x <- set.hwe(x, method='exact')
head(x@snps)
```

```
##   chr         id dist       pos A1 A2 N0  N1  N2 NAs N0.f N1.f N2.f NAs.f
## 1   2  rs7571247    0 179200322  C  T  5  88 410   0   NA   NA   NA    NA
## 2   2  rs3813253    0 179200714  G  A 24 187 292   0   NA   NA   NA    NA
## 3   2  rs6760059    0 179200947  T  C 11 139 353   0   NA   NA   NA    NA
## 4   2 rs16866263    0 179201048  T  G  2  53 448   0   NA   NA   NA    NA
## 5   2 rs77946091    0 179201380  A  G  2  53 448   0   NA   NA   NA    NA
## 6   2 rs77711640    0 179201557  A  G  2  54 447   0   NA   NA   NA    NA
##   callrate        maf        hz       hwe
## 1        1 0.09741551 0.1749503 0.8018482
## 2        1 0.23359841 0.3717694 0.4552613
## 3        1 0.16003976 0.2763419 0.6207438
## 4        1 0.05666004 0.1053678 0.6679028
## 5        1 0.05666004 0.1053678 0.6679028
## 6        1 0.05765408 0.1073559 0.6759765
```

La fonction `set.hwe` donne les p-valeurs du test pour l'équilibre d'Hardy-Weinberg.

Sommaire

Contrôle qualité

Toutes les statistiques précédemment montrées peuvent être utilisées pour le contrôle qualité avec les fonctions `select.snps` et `select.inds`.

Exemple

```
select.inds(x, callrate==1)

## A bed.matrix with 316 individuals and 733 markers.
## snps stats are set
##   There are  26  monomorphic SNPs
## ped stats are set

select.snps(x, maf>0.1)

## A bed.matrix with 503 individuals and 501 markers.
## snps stats are set
## ped stats are set
```

Remarque : il est également possible de données un vecteur de booléens à ces fonctions.

Il est aussi possible avec `gaston` de regarder le sexe génomique :

```
x <- set.genomic.sex(x)
select.inds(x, sex==genomic.sex)
```

Matrices des génotypes standardisés, Z

Dans l'analyse de données génétiques avec des modèles mixtes, il est souvent nécessaire d'utiliser la matrice de génotypes centrée et réduite. Chaque génotype G_{ij} (i est l'index de l'individu et j celui du SNP) est remplacé par

$$Z_{ij} = (G_{ij} - \mu_j)/\sigma_j$$

avec

- $\mu_j = 2p_j$ la moyenne des génotypes codés 0, 1 et 2 avec p_j la fréquence de l'allèle alternatif
- σ_j l'écart-type empirique des génotypes ou son espérance $\sqrt{2p_j(1-p_j)}$.

Matrices des génotypes standardisés, Z

Dans l'analyse de données génétiques avec des modèles mixtes, il est souvent nécessaire d'utiliser la matrice de génotypes centrée et réduite. Chaque génotype G_{ij} (i est l'index de l'individu et j celui du SNP) est remplacé par

$$Z_{ij} = (G_{ij} - \mu_j)/\sigma_j$$

avec

- $\mu_j = 2p_j$ la moyenne des génotypes codés 0, 1 et 2 avec p_j la fréquence de l'allèle alternatif
- σ_j l'écart-type empirique des génotypes ou son espérance $\sqrt{2p_j(1-p_j)}$.

Les 5 derniers slots dont je n'ai pas encore parlé permettent de paramétrer la façon de centrer et réduire les données :

- `@p`, les fréquences alléliques pour chaque SNP
- `@mu`, la moyenne des génotypes codés 0, 1 et 2 (qui correspond à $2p$)
- `@sigma`, l'écart-type empirique pour chaque SNP
- `@standardize_p`, booléen indiquant si nous voulons utiliser l'espérance de la variance sous l'équilibre d'Hardy-Weinberg pour réduite les génotypes
- `@standardize_mu_sigma`, booléen indiquant si nous voulons utiliser la variance empirique pour réduite les génotypes

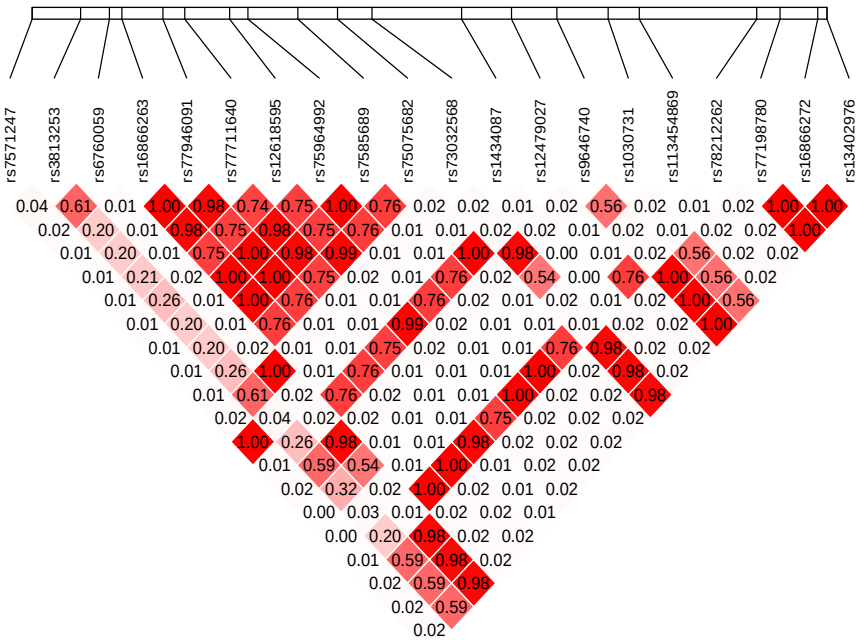
Déséquilibre de liaison

gaston propose également d'estimer le déséquilibre de liaison ($\approx$ corrélation) avec LD par :

$$LD = Z'Z/(n-1).$$

Exemple

```
## Plot
ld.x <- LD(x, c(1,20))
LD.plot( ld.x, snp.positions = x@snps$pos[1:20])
```




```
## Thinning
LD.thin(x, threshold = 0.4,
  max.dist = 500e3, extract=TRUE)

## A bed.matrix with 503 individuals and 70 m
## snps stats are set
## ped stats are set
```

Cette méthode de calcul est aussi utilisée dans la fonction LD.thin qui permet d'extraire un jeu de SNPs en faible déséquilibre de liaison.

Matrice des corrélations génétiques entre individus (GRM)

La matrice de corrélations génétiques (GRM) des individus calculée par la fonction `GRM` s'écrit

$$GRM = \frac{ZZ'}{q-1}.$$

Exemple :

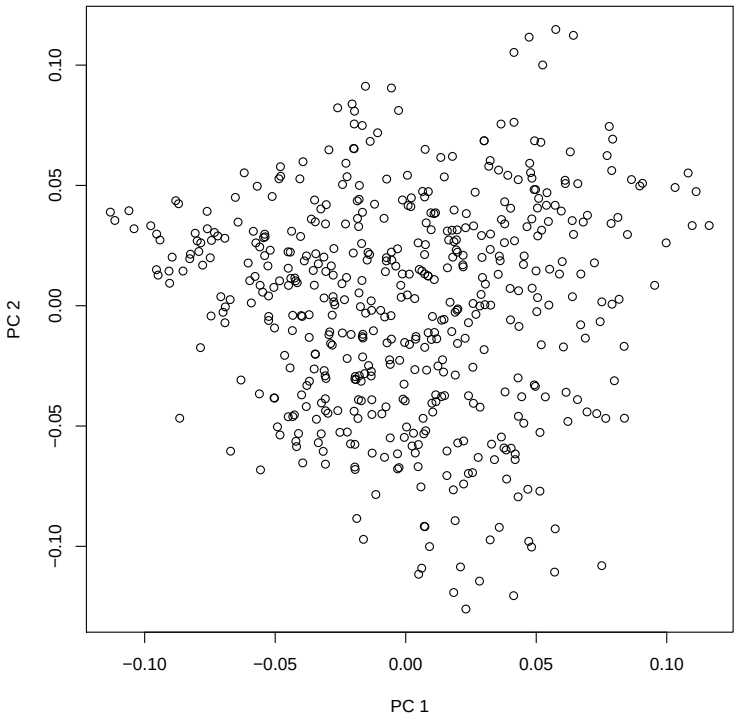
```
prun <- LD.thin(x, 0.2, extract=TRUE)
standardize(prun) <- 'p'
K <- GRM(prun)
K[1:5,1:5]

##             HG00096    HG00097     HG00099     HG00100     HG00101
## HG00096  1.51281333 0.14047940 -0.23943165  0.02294134 -0.06128093
## HG00097  0.14047940 0.79377544  0.08061131  0.11261071  0.04850153
## HG00099 -0.23943165 0.08061131  0.79367918 -0.04554661 -0.07143136
## HG00100  0.02294134 0.11261071 -0.04554661  0.66316450 -0.07108668
## HG00101 -0.06128093 0.04850153 -0.07143136 -0.07108668  0.78965151
```

Analyse en Composantes Principales

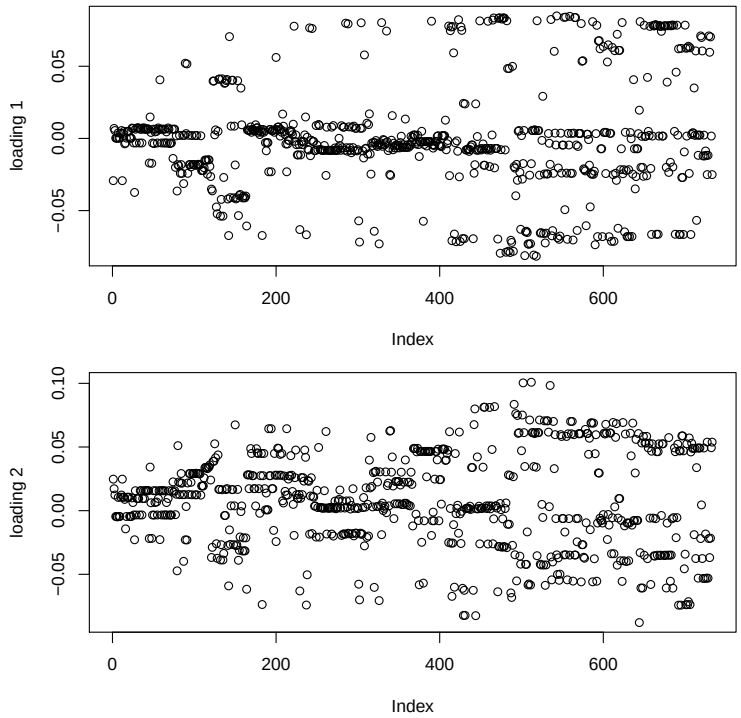
GRM ⇒ matrice de ressemblance génétique ⇒ ACP

Exemple

```
eig <- eigen(K)
par(mar=c(5,4,0.1,0.1))
plot(eig$vectors[,1:2], xlab='PC 1',
     ylab='PC 2')
```




```
L <- bed.loadings(x, eig$vectors[,1:2])
par(mar=c(5,4,0.1,0.1), mfrow=c(2,1))
plot(L[,1], ylab='loading 1')
plot(L[,2], ylab='loading 2')
```



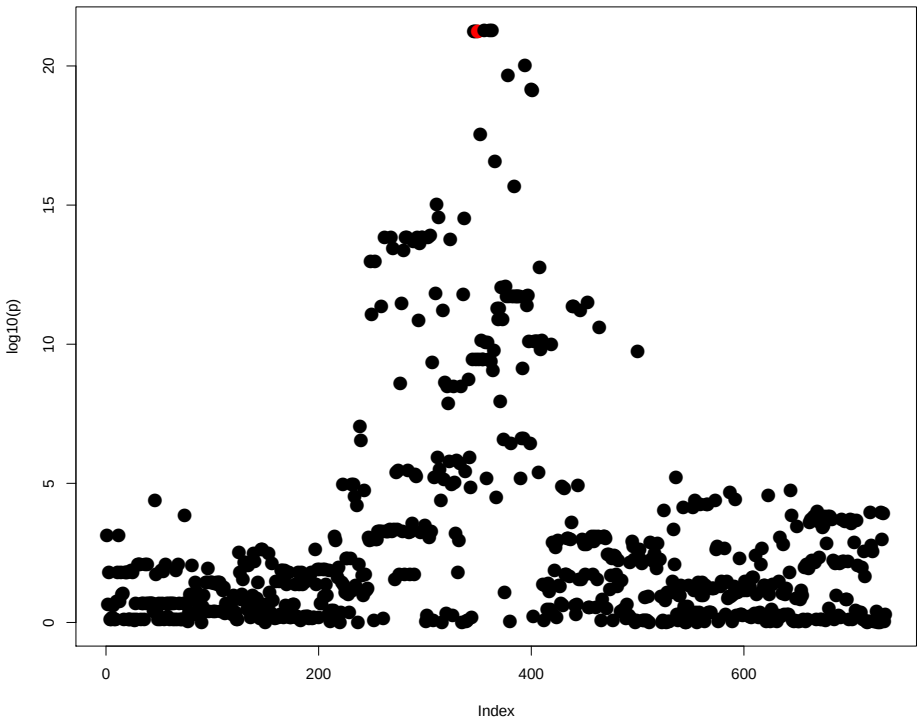
Sommaire

Test d'association classique SNP par SNP (test d'Armitage)

gaston propose de faire le test d'association classique de Wald marqueur par marqueur.

$$Y = \underbrace{X\beta}_{\text{covariables}} + \underbrace{g\gamma}_{\text{marqueur(s) d'intérêt}} + \underbrace{e}_{\sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)}$$

```
t <- association.test(x, x@ped$pheno, method = "lm", test='wald', response="quantitative")
names(t)

## [1] "chr"  "pos"  "id"   "beta" "sd"   "p"
```



Le test de Wald pour les traits binaires est également disponible dans la dernière version de gaston (`response = "binary"`).

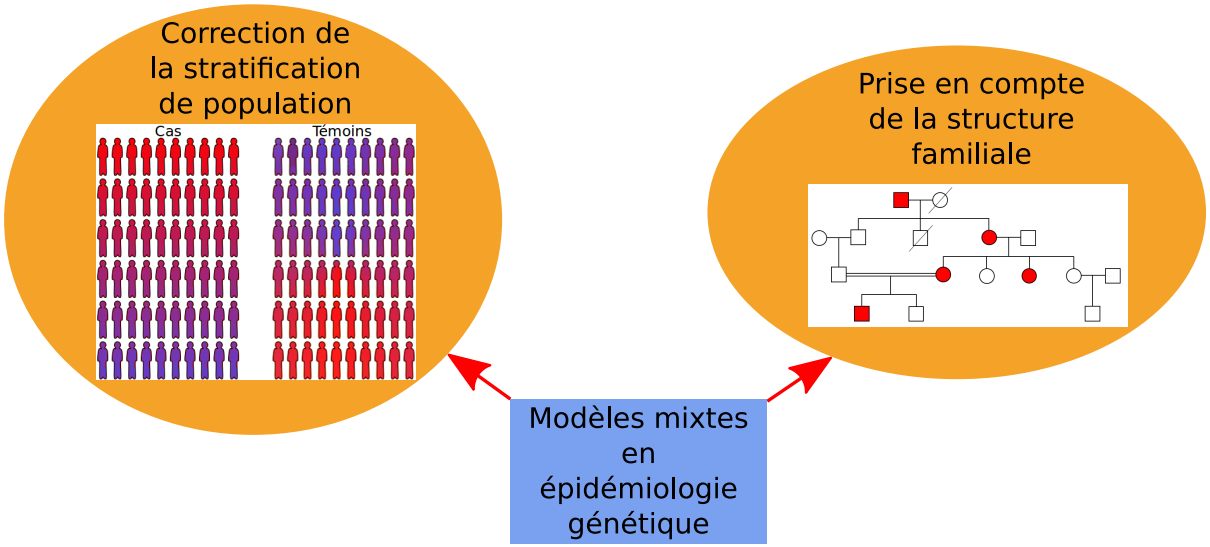
Utilisation des modèles mixtes en épidémiologie génétique

Modèle mixte : Modèle statistique rassemblant des effets fixes et des effets aléatoires

Utilisation des modèles mixtes en épidémiologie génétique

Modèle mixte : Modèle statistique rassemblant des effets fixes et des effets aléatoires

$\Rightarrow$ Corrélation non-nulles entre certaines observations.

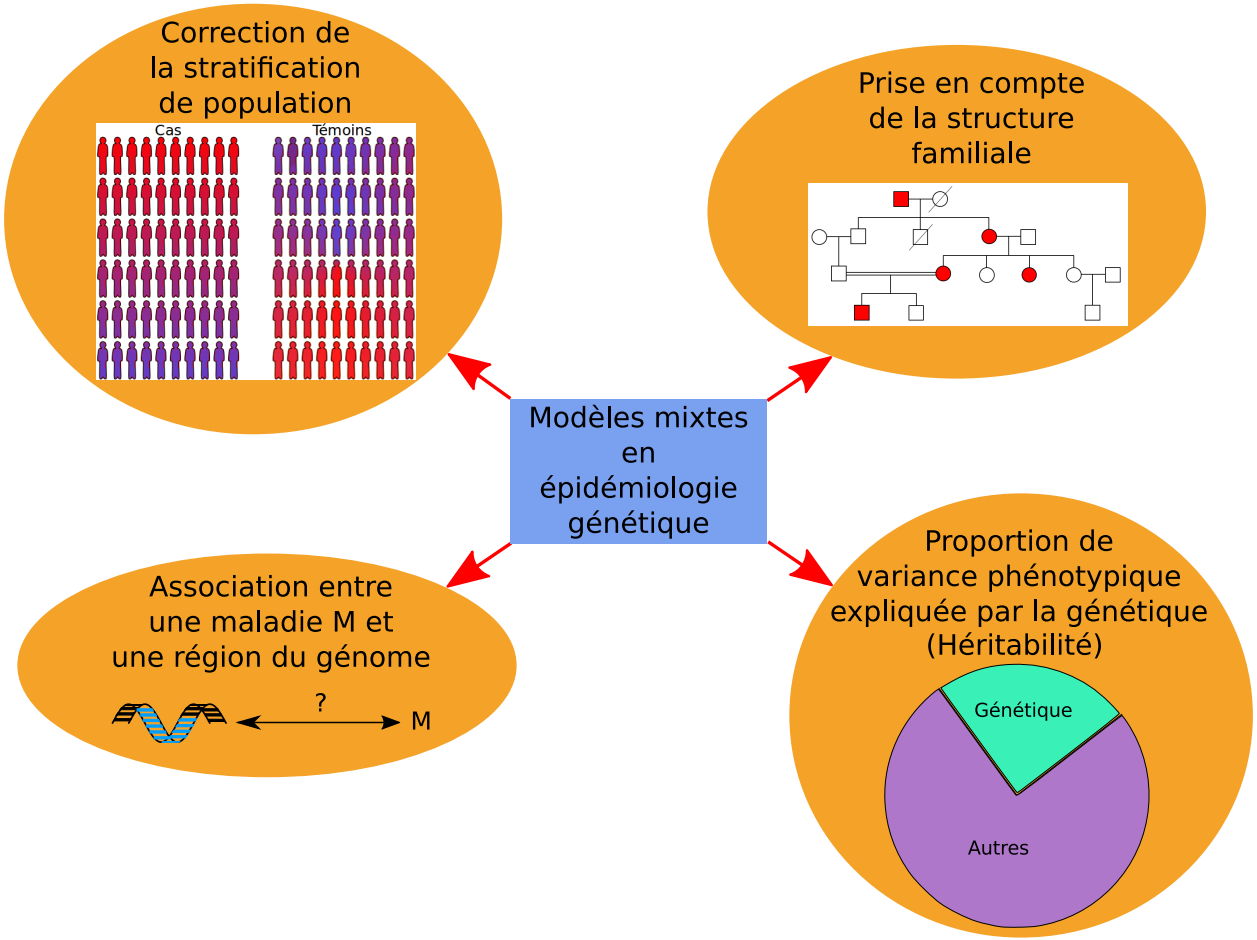


Utilisation des modèles mixtes en épidémiologie génétique

Modèle mixte : Modèle statistique rassemblant des effets fixes et des effets aléatoires

⇒ Corrélation non-nulles entre certaines observations.

⇒ Un seul paramètre de variance pour modéliser plusieurs variables.

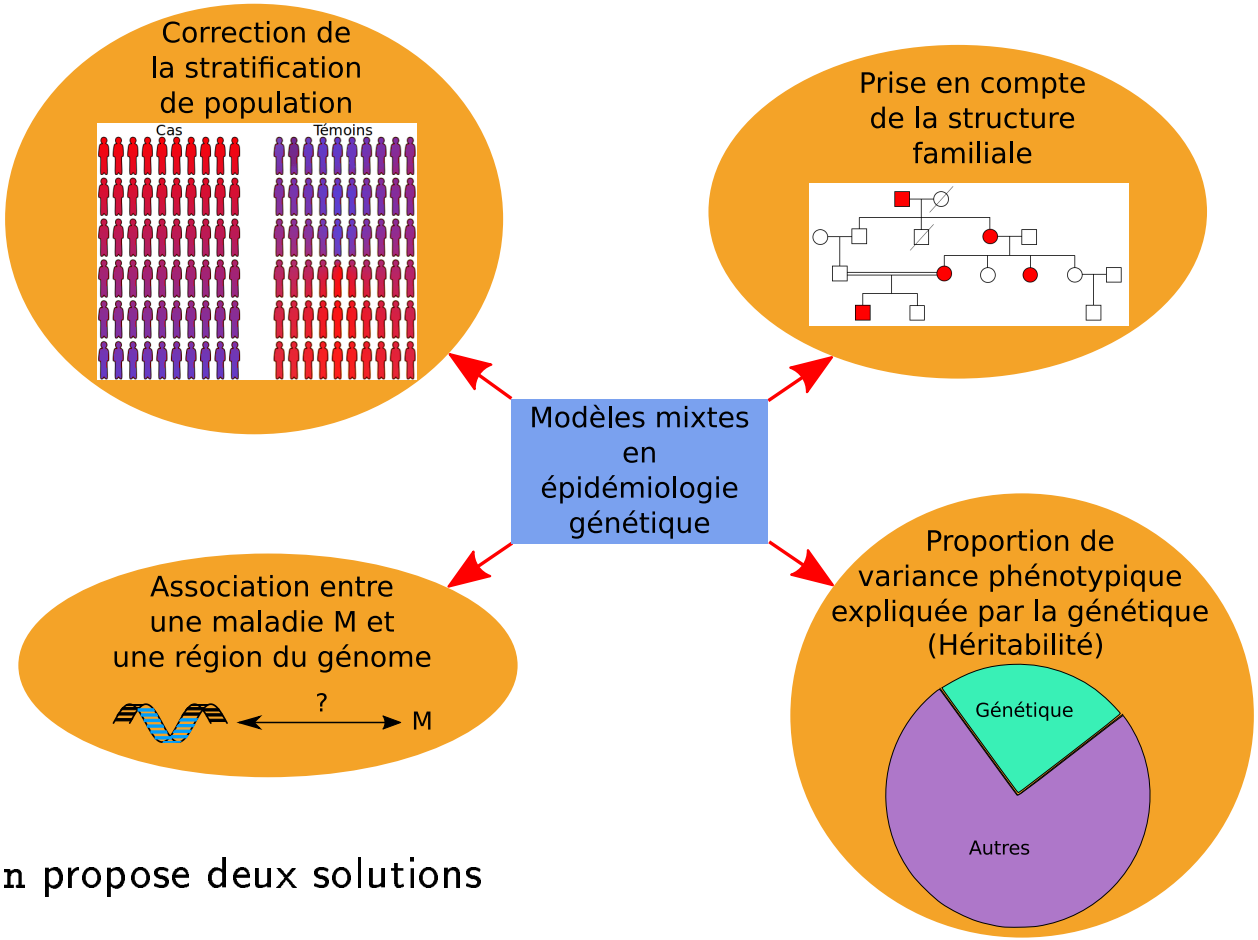


Utilisation des modèles mixtes en épidémiologie génétique

Modèle mixte : Modèle statistique rassemblant des effets fixes et des effets aléatoires

⇒ Corrélation non-nulles entre certaines observations.

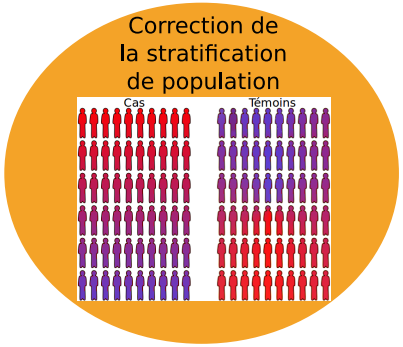
⇒ Un seul paramètre de variance pour modéliser plusieurs variables.

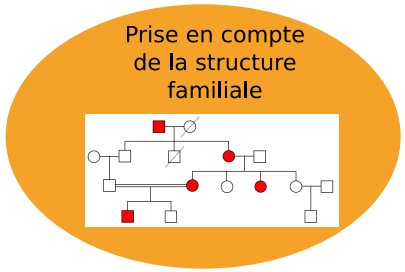


Pour cela, `gaston` propose deux solutions

- AIREML
- « *Diagonal Trick* » (un seul groupe d'effet aléatoire) ⇒ utile pour ajouter des PCs ou regarder la vraisemblance

Test d'association SNP par SNP





`gaston` propose de faire le test d'association marqueur par marqueur en les mettant en effet fixe.

$$Y = \underbrace{X\beta}_{\substack{\text{covariables} \\ + \text{ effets fixes}}} + \underbrace{g\gamma}_{\substack{\text{marqueur(s)} \\ \text{d'intérêt} \\ + \text{ effet(s) fixe(s)}}} + \underbrace{\omega}_{\substack{\text{effet} \\ \text{polygénique ou} \\ \text{stratification}}} + \underbrace{e}_{\sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)}$$

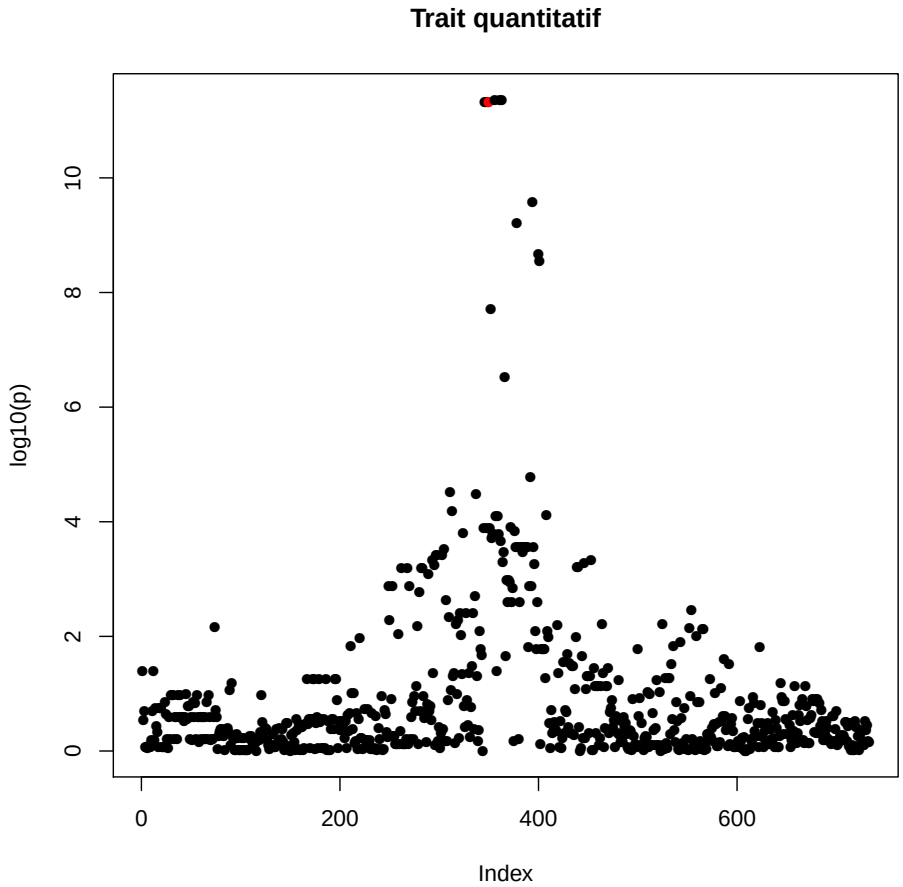
Les tests disponibles sont :

- pour un trait quantitatif : tests de Wald, du rapport de vraisemblance et du score

```
 t <- association.test(x, x@ped$pheno, eigenK = eig, method = "lmm", test="wald"); names(t)
## [1] "chr"  "pos"  "id"   "h2"   "beta" "sd"   "p"
```

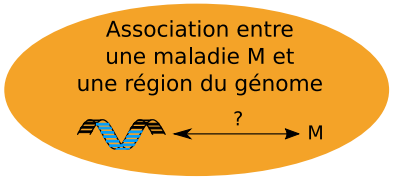
- pour un trait binaire : tests de Wald et du score

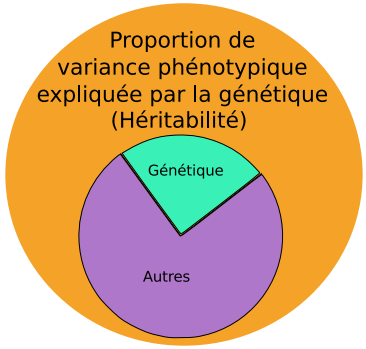
Test d'association SNP par SNP



Remarque : Il est également possible avec `gaston` d'appliquer un test du score sur un ensemble de variables introduites en effets fixes avec les fonctions `score.fixed.linear` et `score.fixed.logistic`.

Test d'association pour une région du génome ou le génome entier





gaston propose de faire le test du score pour regarder l'association avec une région du génome inclue avec des effets aléatoires.

$$Y = \underbrace{X\beta}_{\substack{\text{covariables} \\ + \text{ effets fixes}}} + \underbrace{Gu}_{\substack{\text{marqueurs} \\ \text{d'intérêt} \\ + \text{ effets aléatoires}}} + \underbrace{\omega}_{\substack{\text{effet} \\ \text{polygénique} \\ \text{ou} \\ \text{stratification}}} + \underbrace{e}_{\sim \mathcal{N}_n(0, \sigma^2 \mathbb{I}_n)}$$

Exemple

```
# Estimer le modèle
estimates <- lmm.aireml(x@ped$pheno, K = K)
names(estimates)
```

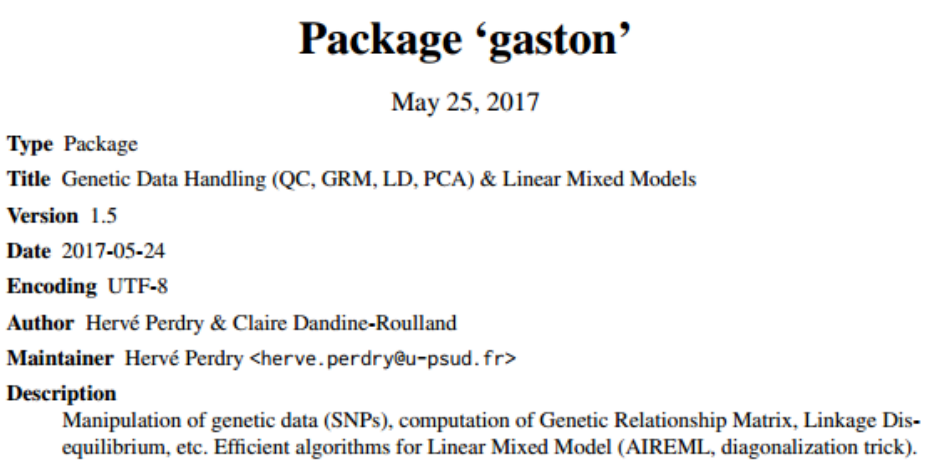
```
##  [1] "sigma2"     "tau"        "logL"
##  [4] "logL0"      "niter"      "norm_grad"
##  [7] "Py"         "BLUP_omega" "BLUP_beta"
## [10] "varbeta"    "varXbeta"
```

```
# Tester la variance
t <- score.variance.linear(K, x@ped$pheno)
str(t)
```

```
## List of 2
##  $ score: num [1, 1] 2089
##  $ p    : num 7.51e-10
```

Pour un trait binaire, les fonctions sont `logistic.mm.aireml` et `score.variance.logistic`.

Conclusions



Package 'gaston'

May 25, 2017

Type Package

Title Genetic Data Handling (QC, GRM, LD, PCA) & Linear Mixed Models

Version 1.5

Date 2017-05-24

Encoding UTF-8

Author Hervé Perdry & Claire Dandine-Roulland

Maintainer Hervé Perdry <herve.perdry@u-psud.fr>

Description

Manipulation of genetic data (SNPs), computation of Genetic Relationship Matrix, Linkage Disequilibrium, etc. Efficient algorithms for Linear Mixed Model (AIREML, diagonalization trick).

Un package R `gaston` permettant :

- la manipulation d'une matrice de génotypes (contrôle qualité, GRM, ...)
- le calcul du déséquilibre de liaison
- l'estimation des composantes du modèle linéaire mixte pour un trait quantitatif ou binaire,
- la prédiction,
- l'analyse d'association avec un effet fixe ou aléatoire.

Les performances de `gaston` sont comparables voir meilleures que celles des logiciels classiques (PLINK, GCTA,...).

Tout commentaire ou retour est le bienvenu...