Présentation des Logiciels

RobPower

Calcul du risque de première espèce et de la puissance d'un protocole avec des animaux apparentés pour différentes méthodes d'analyse d'association

Muller

Calcul des seuils pour les tests multiples

Muller : Objectif

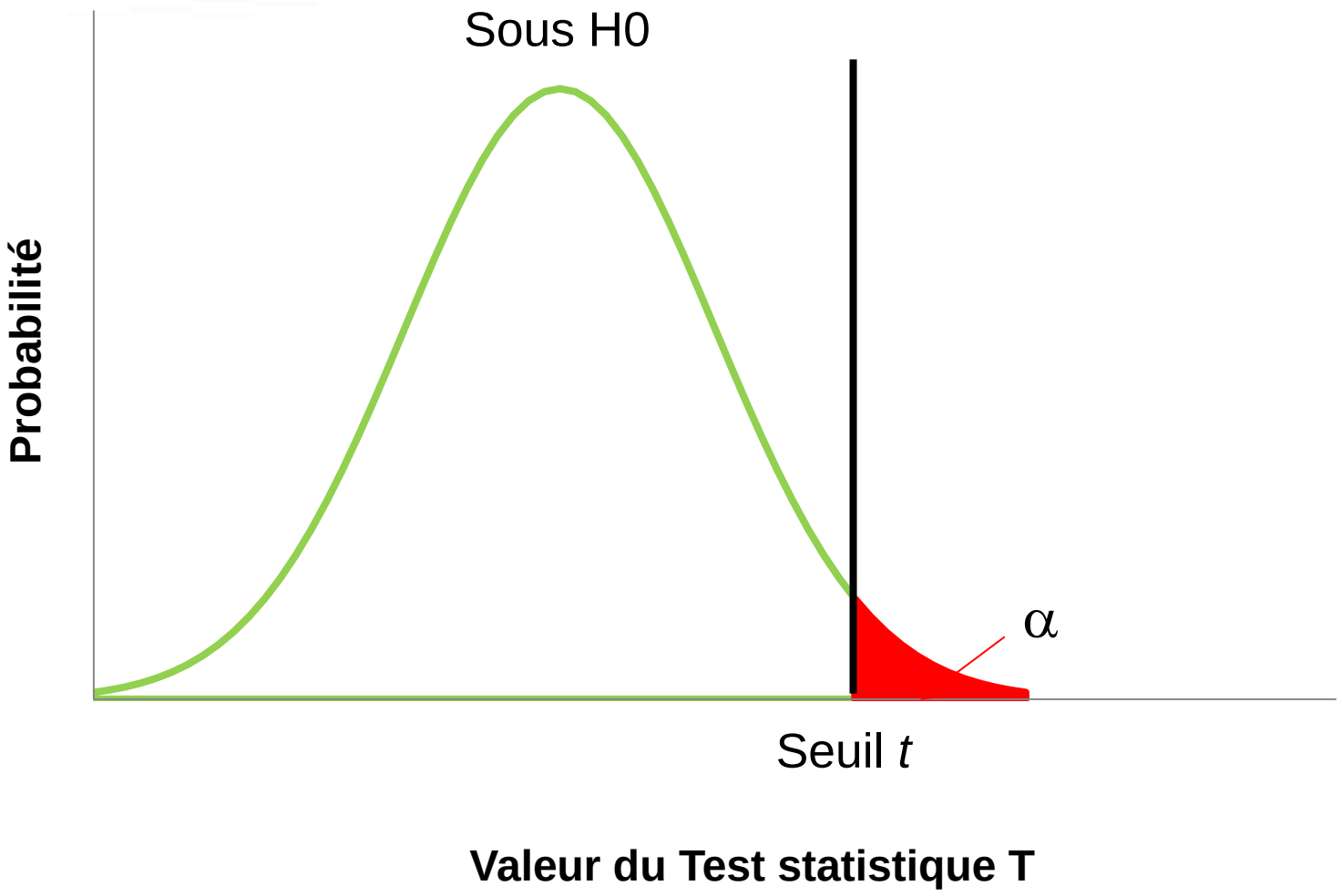
- Quand on réalise un GWAS SNP par SNP, on effectue plus de 50 000 tests (puce 50K)
- Le risque de première espèce est la probabilité pour que la valeur du test statistique soit supérieur au seuil choisi en l'absence d'effet (H0)
- Avec un risque de première espèce à 1% on obtient donc en espérance 500 tests significatifs par hasard.
- Ce qu'on cherche ce n'est pas d'avoir 1% de chance pour chaque SNP de croire qu'il a un effet alors qu'il n'en n'a pas, c'est d'avoir 1% de chance que l'expérience conclut à l'existence d'un effet alors qu'il n'y en a pas.
- L'idée est donc de contrôler ce risque sur l'ensemble de l'expérience (tests multiples) et non sur chaque test.

Muller BU, Stich, B., Piepho HP., A general method for controlling the genome-wide type I error rate in linkage and association maping experiments in plants. Heredity, 106, 825-831.

Correction Bonferroni

- Supposons que nous choisissions un risque de première espèce α pour un test statistique T avec un seuil t

Risque de première espèce α



Correction Bonferroni

- Supposons que nous choisissions un risque de première espèce α pour un test statistique T avec un seuil t

$$P(T > t) = \alpha$$

Correction Bonferroni

- Supposons que nous choisissions un risque de première espèce α pour un test statistique T avec un seuil t

$$P(T > t) = \alpha$$

- Ce test est répété pour *N* SNP

Correction Bonferroni

- Supposons que nous choisissions un risque de première espèce α pour un test statistique T avec un seuil t

$$P(T > t) = \alpha$$

- Ce test est répété pour *N* SNP
- La probabilité qu'on ait au moins 1 test significatif sous H0 est

$$P(T_1 < t \cup T_2 < t .... \cup T_N < t)$$

$$P(T_1 > t \cup T_2 > t .... T_N > t) \leq NP(T > t) \leq N\alpha$$

Correction Bonferroni

- Supposons que nous choisissions un risque de première espèce α pour un test statistique T avec un seuil t

$$P(T > t) = \alpha$$

- Ce test est répété pour *N* SNP
- La probabilité qu'on ait au moins 1 test significatif sous H0 est

$$P(T_1 < t \cup T_2 < t .... \cup T_N < t)$$

$$P(T_1 > t \cup T_2 > t .... T_N > t) \leq NP(T > t) \leq N\alpha$$

- Pour que cette probabilité soit inférieure à un risque choisi α'

$$P(T_1 > t \cup T_2 > t .... T_N > t) \leq NP(T > t) \leq N\alpha = \alpha'$$

- Il suffit de choisir

$$\alpha = \frac{\alpha'}{N}$$

Correction Bonferroni

- Pour avoir 1% de chance que notre expérience de *N* tests donne sous H0 au moins un test qui dépasse le seuil *t*, il faut choisir un risque de première espèce pour chaque test de

$$\frac{1\%}{N}$$

  - soit pour 50 000 tests , 1% -> 2.10^{-7}
  - Ou pour 1 chromosome de 2000 marqueurs, 1% -> 5.10^{-6}

- Cependant cette correction est très sévère, surtout dans le cas de tests non indépendants (comme c'est le cas avec des SNP en déséquilibre de liaison)
- Muller et al. ont cherché à contrôler le risque de première espèce dans le cas de tests multiples en tenant compte des relations entre les tests.

Muller et al. : le principe

- On réalise *N* tests avec le modèle mixte simple pour chaque SNP *i*

$$\mathbf{y} = \mathbf{1}\mu_i + \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{e} \qquad \mathbf{V} = V(\mathbf{e}) = \mathbf{A}\sigma_u^2 + \mathbf{I}\sigma_e^2$$

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{X}_i{}'\mathbf{V}^{-1}\mathbf{X}_i)\mathbf{X}_i{}'\mathbf{V}^{-1}\mathbf{y}$$

- Les estimations des effets des différents SNP sont corrélées entre elles

$$\mathrm{cov}(\hat{\boldsymbol{\beta}}_i, \hat{\boldsymbol{\beta}}_j) = (\mathbf{X}_i{}'\mathbf{V}^{-1}\mathbf{X}_i)\mathbf{X}_i{}'\mathbf{V}^{-1}\mathbf{X}_j(\mathbf{X}_j{}'\mathbf{V}^{-1}\mathbf{X}_j)$$

- On peut donc construire une matrice de variance-covariance entre les estimations des *N* effets des SNP
  - Ne nécessite que les génotypes et la matrice de parenté

Muller et al. : le principe

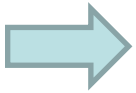
- Calcule la matrice de variance-covariance entre les estimations des effets
  - *Astuce : utiliser les équations du modèle mixte plutôt que la formulation initiale du BLUE*
- Connaissant cette matrice, on va simuler ces *N* estimations sous H0
  - *Astuce : décomposition de la matrice en valeur singulières, simulation rapide de N normales indépendantes*.
- Calculer les *N* tests correspondants

→ Cette simulation constitue la réalisation d'***une*** expérience

→ Pour cette expérience, le fait de considérer qu'on a un test significatif va dépendre de la valeur du test la plus grande

Muller et al. : le principe

- On conserve la plus grande valeur du test
- On va répéter l'expérience (donc la simulation des *N* tests) un nombre très grand nombre de fois ($\rightarrow \propto$)



On obtient la distribution de nos plus grands tests

- Pour obtenir le seuil adéquat pour n'avoir qu'1% des expériences qui donnent sous H0 un test significatif, il suffit de trier ces valeurs et de choisir celle qui correspond au quantile des 1% (idem pour 5%, 10%...)

De quoi dépendent ces corrélations ?

- Largement les corrélations entre les tests dépendent de la corrélation entre les SNP (DL)
- La répartition des génotypes / valeurs polygéniques et aux effets fixes, jouent aussi

Exemples de résultats

Projet Genendurance, les premiers SNP du chromosome 1, 100 000 simulations

CORRECTION BONFERRONI

	Nombre de SNP		
Seuil choisi	10	100	1000
1%	1.0E-03	1.0E-04	1.0E-05
5%	5.0E-03	5.0E-04	5.0E-05
10%	1.0E-02	1.0E-03	1.0E-04

CORRECTION MULLER

	Nombre de SNP		
Seuil choisi	10	100	1000
1%	1.3E-03	1.3E-04	1.3E-05
5%	7.1E-03	7.3E-04	7.2E-05
10%	1.5E-02	1.6E-03	1.6E-04

Exemples de résultats

Projet Genendurance, les premiers SNP du chromosome 1, 100 000 simulations

CORRECTION BONFERRONI

	Nombre de SNP		
Seuil choisi	10	100	1000
1%	1.0E-03	1.0E-04	1.0E-05
5%	5.0E-03	5.0E-04	5.0E-05
10%	1.0E-02	1.0E-03	1.0E-04

CORRECTION MULLER : nombre de tests « réels »

	Nombre de SNP		
Seuil choisi	10	100	1000
1%	8	76	749
5%	7	68	692
10%	7	63	640