Techniques statistiques pour la détection de gènes à effets quantitatifs

Charles-Elie Rabier

Institut de Mathématiques de Toulouse (IMT)
Station d'Amélioration Génétique des Animaux, INRA Toulouse

Directeurs de thèse :
Jean-Marc Azaïs (IMT), Jean-Michel Elsen (INRA)

Co-encadrante :
Céline Delmas (INRA)

16 Juin 2010

Qu'est-ce qu'un QTL ?

QTL = Quantitative Trait Locus

Un QTL est un locus à l'origine
de la variation d'un caractère quantitatif

Comment détecter et localiser un QTL ?

On a besoin :

- d'une population en ségrégation (obtenue à l'aide de croisements)
- de marqueurs génétiques positionnés le long du génome
- de valeurs phénotypiques

$\Rightarrow$ les méthodes statistiques vont nous permettrent de détecter et localiser le QTL

Feuille de route

Première partie : Selective Genotyping

- Problème soulevé par les généticiens
- Modèle applicable à d'autres domaines

Deuxième partie : Génome Scan

- Etude d'un modèle propre à la génétique

Première partie :

Selective Genotyping

Le QTL est présent sur un marqueur donné

Modèle en l'absence de censure

- X : variable aléatoire correspondant au génotype au QTL

$$X = \begin{cases} -1 & \text{avec probabilité } 1-p \\ 1 & \text{avec probabilité } p \end{cases}$$

On suppose $p \neq \{0, 1\}$

- Y : variable aléatoire correspondant au phénotype

$$Y = \mu + q\,X + \varepsilon \qquad \text{où} \quad \varepsilon \sim N(0, \sigma^2)$$

Modèle en l'absence de censure

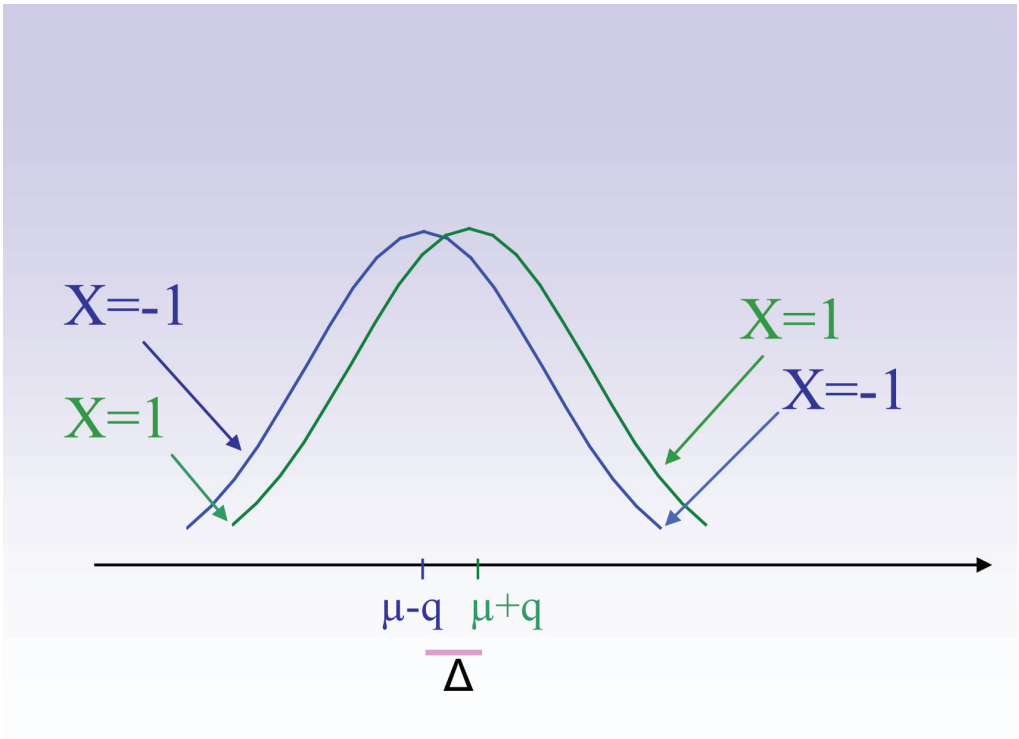


FIG.: Distribution des phénotypes *Y*

Test statistique oracle (μ, q, σ)

- A l'aide de n observations (X_j, Y_j) iid, on souhaite tester :

$$H_0 : \ q = 0 \ \ vs \ \ H_1 : \ q \neq 0$$

On considère une alternative locale $H_a : q = \frac{a}{\sqrt{n}}$

- Test statistique oracle :

$$T = \frac{\sum_{j=1}^{n} \ \frac{1}{p}(Y_j - \overline{Y}) \, 1_{X_j=1} \ - \ \frac{1}{1-p}(Y_j - \overline{Y}) \, 1_{X_j=-1}}{\hat{\sigma} \ \sqrt{\frac{n}{p \ (1-p)}}}$$

$$T \overset{H_0}{\rightarrow} N(0, \ 1) \quad \text{et} \quad T \overset{H_a}{\rightarrow} N\left(\frac{2a \ \sqrt{p \ (1-p)}}{\sigma}, \ 1 \right)$$

Selective Genotyping

Génotyper coûte très cher

$\Rightarrow$ Selective Genotyping : génotypage uniquement des individus présentant des phénotypes Y extrêmes.

Le nombre d'individus génotypés, afin d'obtenir une puissance donnée, est réduit considérablement à condition que le nombre d'individus phénotypés ait été augmenté

Lebowitz et al. (Theoretical and Applied Genetics, 1987)

Questions abordées

- Combien d'individus supplémentaires faut-il phénotyper pour avoir même puissance qu'en situation oracle ?
- Doit-on génotyper uniquement les individus présentant les plus grands phénotypes, ou au contraire ceux présentant les plus petits phénotypes, ou bien un mélange des deux ?
- Doit-on conserver les phénotypes non extrêmes dans l'analyse statistique ?

Modèle correspondant au Selective Genotyping

X disponible uniquement pour les individus présentant un phénotype extrême Y

$\Rightarrow$ On n'observe plus X mais $\overline{X}$:

$$\overline{X} = \begin{cases} X & \text{si } Y \notin [S_- , S_+] \\ 0 & \text{sinon} \end{cases}$$

où S_- et S_+ sont deux réels tels que $S_- \leq S_+$.

Modèle correspondant au Selective Genotyping

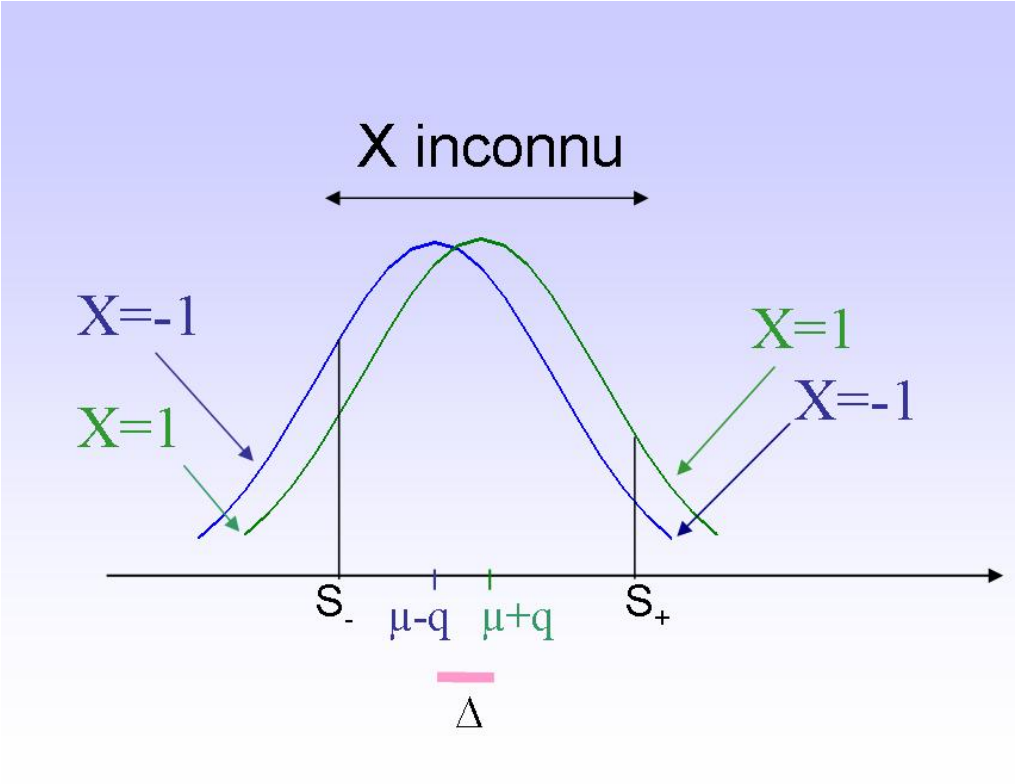


FIG.: Distribution des phénotypes Y

Ecriture de la vraisemblance

- Vraisemblance pour une observation $(Y\ ,\ \overline{X})$

$$
\begin{aligned}
L &= \frac{1-p}{\sigma}\varphi\left(\frac{y-\mu+q}{\sigma}\right)1_{\overline{X}=-1} + \frac{p}{\sigma}\varphi\left(\frac{y-\mu-q}{\sigma}\right)1_{\overline{X}=1} \\
&+ \left\{\frac{1-p}{\sigma}\varphi\left(\frac{y-\mu+q}{\sigma}\right) + \frac{p}{\sigma}\varphi\left(\frac{y-\mu-q}{\sigma}\right)\right\}1_{\overline{X}=0}
\end{aligned}
$$

avec φ densité d'une loi normale standardisée

- Vraisemblance très difficile à maximiser

$\Rightarrow$ Algorithme EM nécessaire afin d'obtenir les EMV $\hat{\mu}$, $\hat{q}$ et $\hat{\sigma}$

Test de Wald (μ, q, σ)

$$H_0 : \ q = 0 \ \ vs \ \ H_1 : \ q \neq 0$$

On considère une alternative locale $H_a : q = \frac{a}{\sqrt{n}}$

- Test de Wald

$$W_1 = \frac{2\sqrt{n}}{\hat{\sigma}^2} \sqrt{\hat{\mathcal{A}}\, p(1-p)}\ \hat{q} \quad , \quad W_1 \overset{H_0}{\rightarrow} N(0,\ 1)$$

$$\text{où} \ \ \mathcal{A} = E_{H_0}\left[(Y-\mu)^2\, 1_{\overline{X}\neq 0}\right] \ \ , \ \ \hat{\mathcal{A}} = \frac{1}{n}\sum_{j=1}^{n}\left(Y_j - \overline{Y}\right)^2\, 1_{\overline{X}_j \neq 0}$$

loi sous l'alternative locale ?

Test de Wald (μ, q, σ)

3ème lemme de Le Cam

Soient P_n et Q_n deux suites de mesures de probabilité sur des espaces $(\Omega_n, \mathcal{B}_n)$ et $T_n : \Omega_n \mapsto \mathbb{R}^d$ une suite de variables aléatoires.

Si

$$\left(T_n, \log\left(\frac{dQ_n}{dP_n}\right)\right)' \xrightarrow{P_n} N_{d+1}\left(\begin{pmatrix} \xi \\ -\frac{1}{2}\nu^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau' & \nu^2 \end{pmatrix}\right)$$

alors ,

$$T_n \xrightarrow{Q_n} N_d(\xi + \tau,\ \Sigma)$$

Théorème

Soient $C_1, ..., C_n$ un échantillon iid provenant d'une distribution P_θ. Supposons que Θ est un ouvert de $\mathbb{R}^d$ et que le modèle $(P_\theta : \theta \in \Theta)$ est régulier. On note $\theta_0 \in \Theta$ et $\hat{\theta}$ l'EMV de θ, alors pour toute séquence convergente de type $h_n \to h$, on a :

$$i) \quad \text{sous } P_{\theta_0} : \quad \sqrt{n}\ (\hat{\theta} - \theta_0) \to N(0,\ I^{-1}(\theta_0))$$

$$ii) \quad \text{sous } P_{\theta_0 + h_n/\sqrt{n}} : \quad \sqrt{n}\ (\hat{\theta} - \theta_0) \to N(h,\ I^{-1}(\theta_0))$$

Test de Wald (μ, q, σ)

$$H_0 : \ q = 0 \ \ vs \ \ H_1 : \ q \neq 0$$

On considère une alternative locale $H_a : q = \frac{a}{\sqrt{n}}$

- Test de Wald

$$W_1 = \frac{2\sqrt{n}}{\hat{\sigma}^2} \sqrt{\hat{\mathcal{A}}\, p(1-p)} \ \hat{q} \quad , \quad W_1 \overset{H_0}{\rightarrow} N(0,\ 1)$$

$$\text{alors} \ \ W_1 \overset{H_a}{\rightarrow} N\left(\frac{2a\ \sqrt{\mathcal{A}\, p(1-p)}}{\sigma^2},\ 1 \right)$$

Efficacité du test de Wald (μ, q, σ)

Comment est affectée la puissance du test de Wald lorsque le nombre d'individus est augmenté tout en conservant le même effet QTL $q = \frac{a}{\sqrt{n}}$?

- $n^\star$ nouveau nombre d'individus
- ratio $\zeta = \frac{n^\star}{n}$

Définition

*On définit **l'efficacité** du test de Wald, $\kappa_1 = \frac{1}{\zeta_{eff}}$, où ζ_{eff} désigne la valeur de ζ pour laquelle la puissance du test de Wald est égale à la puissance du test oracle*

Efficacité du test de Wald (μ, q, σ)

On note $\gamma = \mathbb{P}_{H_0}\left(Y \notin [S_-,\ S_+]\right)$

A la fois sous H_0 et sous H_a, γ correspond asymptotiquement au pourcentage d'individus génotypés

De la même manière, on note :

- $\gamma_+ = \mathbb{P}_{H_0}\left(Y > S_+\right)$
- $\gamma_- = \mathbb{P}_{H_0}\left(Y < S_-\right)$

Bien évidemment : $\gamma = \gamma_+ + \gamma_-$

Efficacité du test de Wald :

$$\forall p,\ \ \kappa_1 = \gamma\ +\ z_{\gamma_+}\varphi(z_{\gamma_+})\ -\ z_{1-\gamma_-}\ \varphi(z_{1-\gamma_-})$$

Optimisation du génotypage

- On souhaite génotyper uniquement un pourcentage γ de la population

$\Rightarrow$ Comment choisir les γ_+ et γ_- optimaux ?

$\forall p$, κ_1 atteint son maximum M pour $\gamma_+ = \gamma_- = \gamma/2$

$$M = \gamma + 2\, z_{\gamma/2}\, \varphi(z_{\gamma/2})$$

$\forall p$, on doit génotyper le même pourcentage d'individus à "droite" qu'à "gauche" !

$$Si\ \gamma = 0.3\ alors\ M = 78\%$$

Illustration graphique

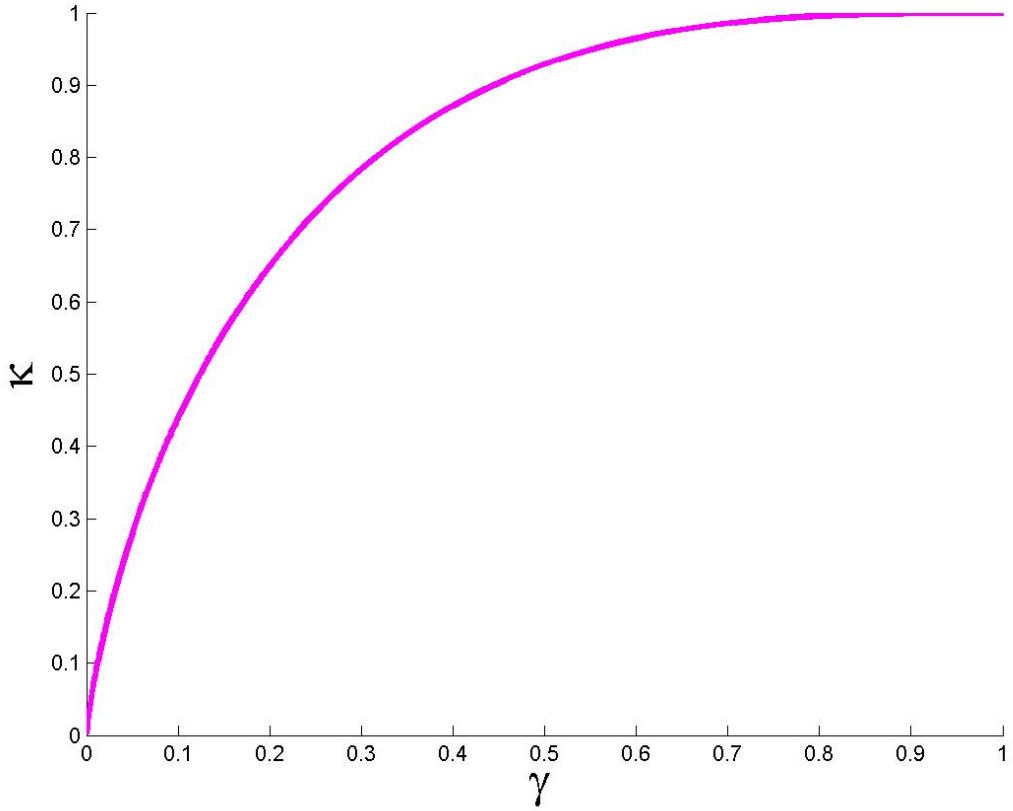


FIG.: Efficacité du test de Wald en fonction de γ ($\gamma_+ = \gamma_- = \gamma/2$)

Une autre question propre au Selective Genotyping

Existe-t-il de l'information dans la "bande" ?

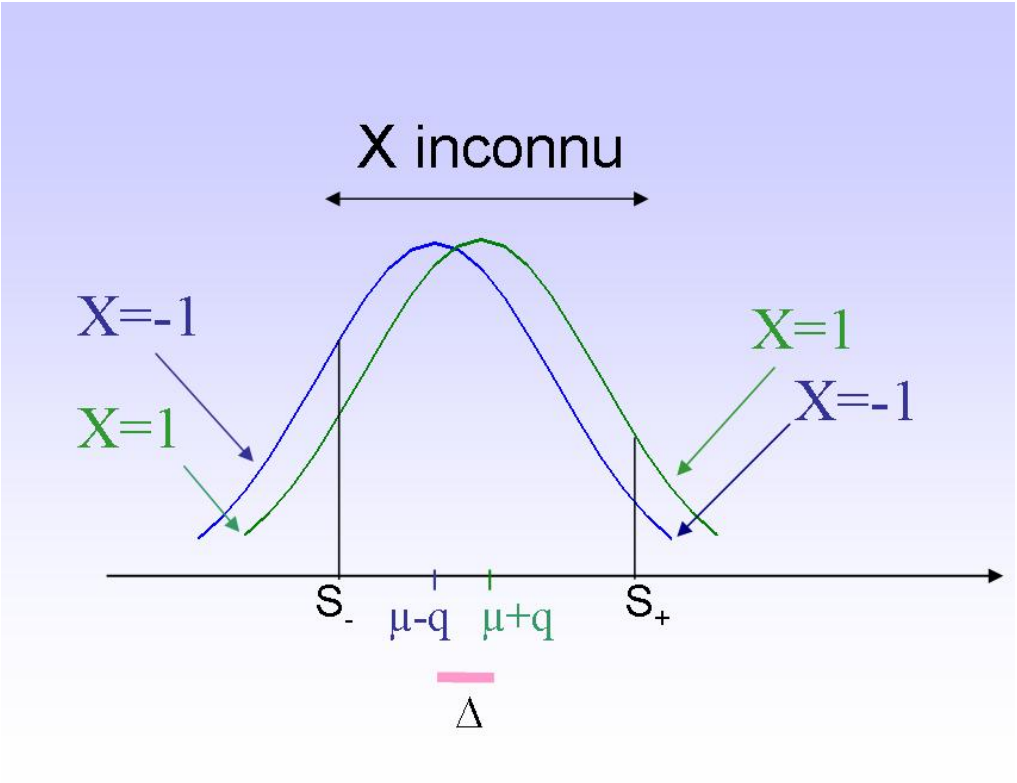
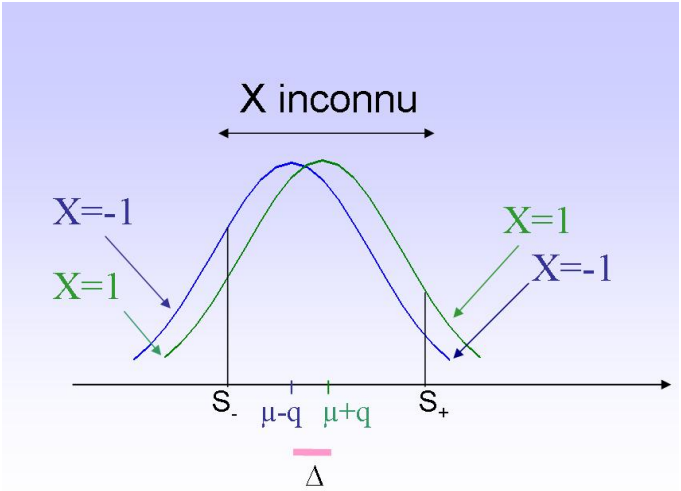


FIG.: Distribution des phénotypes Y

Comparaison des 3 stratégies

3 stratégies pour l'analyse de données en Selective Genotyping :

1. Test de Wald basé sur l'ensemble des phénotypes
2. Comparaison de moyenne basée sur les phénotypes extrêmes
3. Test de Wald basé sur les phénotypes extrêmes



Comparaison des 3 stratégies (μ, q, σ)

Lemme

$$W_1 := \frac{2\sqrt{n}}{\hat{\sigma}^2}\sqrt{\hat{\mathcal{A}}\, p(1-p)}\;\hat{q}_1$$

$$T_2 := \sqrt{p(1-p)}\left\{\frac{\sum_{j=1}^{n}\frac{1}{p}(Y_j-\overline{Y})1_{\overline{X}_j=1} - \frac{1}{1-p}(Y_j-\overline{Y})1_{\overline{X}_j=-1}}{\sqrt{n\,\hat{\mathcal{A}}}}\right\}$$

$$W_3 := \frac{2\sqrt{n}}{\hat{\sigma}^2}\sqrt{\hat{\mathcal{A}}\, p(1-p)}\;\hat{q}_3$$

présentent les mêmes lois asymptotiques sous H_0 et sous H_a, à savoir :

$$N(0,\ 1) \quad \textit{et} \quad N\left(\frac{2a\ \sqrt{\mathcal{A}\, p(1-p)}}{\sigma^2},\ 1\right)$$

où $\hat{q}_1$ et $\hat{q}_3$ sont les EMV de q pour les stratégies une et trois, et où

$$\hat{\mathcal{A}} = \frac{1}{n}\sum_{j=1}^{n}(Y_j-\overline{Y})^2 1_{\overline{X}_j\neq 0} \quad , \quad \overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_j$$

$$\mathcal{A} = \sigma^2\left\{\gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-})\right\} \quad , \quad \hat{\sigma}^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j-\overline{Y})^2$$

Comparaison des 3 stratégies (μ, q, σ)

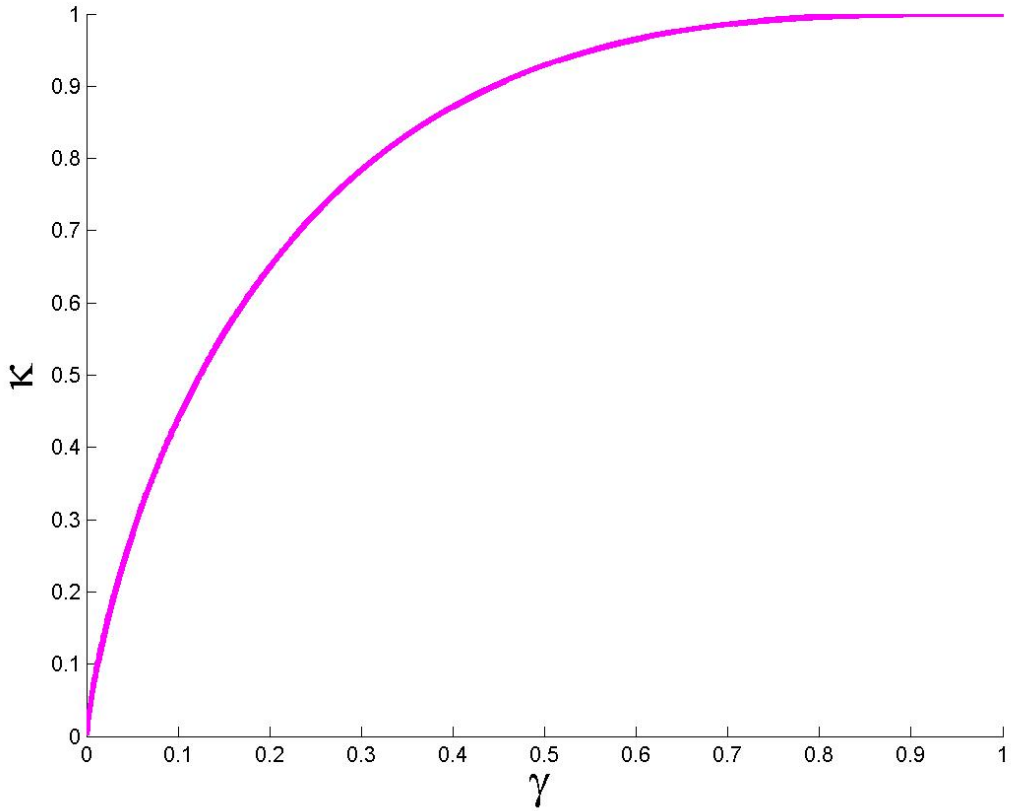


FIG.: Efficacité des tests, correspondant aux différentes stratégies, en fonction de γ ($\gamma_+ = \gamma_- = \gamma/2$)

Convergence vers l'asymptotique (q)

γ	β_{MC}	β_{th}	IC en %
0.1	38.27%	37.45%	[37.32 ; 39.22]
0.2	48.80%	48.61%	[47.82 ; 49.78]
0.3	54.64%	54.77%	[53.66 ; 55.62]
0.4	58.60%	58.58%	[57.63 ; 59.57]
0.5	61.48%	60.93%	[60.53 ; 62.43]
0.6	61.73%	62.33%	[60.78 ; 62.68]
0.7	63.21%	63.13%	[62.26 ; 64.16]
0.8	63.27%	63.52%	[62.33 ; 64.21]
0.9	63.79%	63.68%	[62.85 ; 64.73]
1	63.56%	63.68%	[62.62 ; 64.50]

TAB.: Puissance théorique (β_{th}) et puissance par Monte-Carlo (β_{MC}) en fonction du pourcentage de génotypés γ ($\gamma_+ = \gamma_- = \gamma/2$, $p = 1/2$, $nb_{ech} = 10000$, $n = 30$, $q = \frac{2}{\sqrt{30}} = 0.3651$)

Conclusions sur le Selective Genotyping

- On doit génotyper le même pourcentage d'individus aux deux extrêmes
- Il n'y a pas d'information dans la "bande"
- Mêmes conclusions pour un selective genotyping avec deux caractères corrélés

Deuxième partie :

Génome Scan

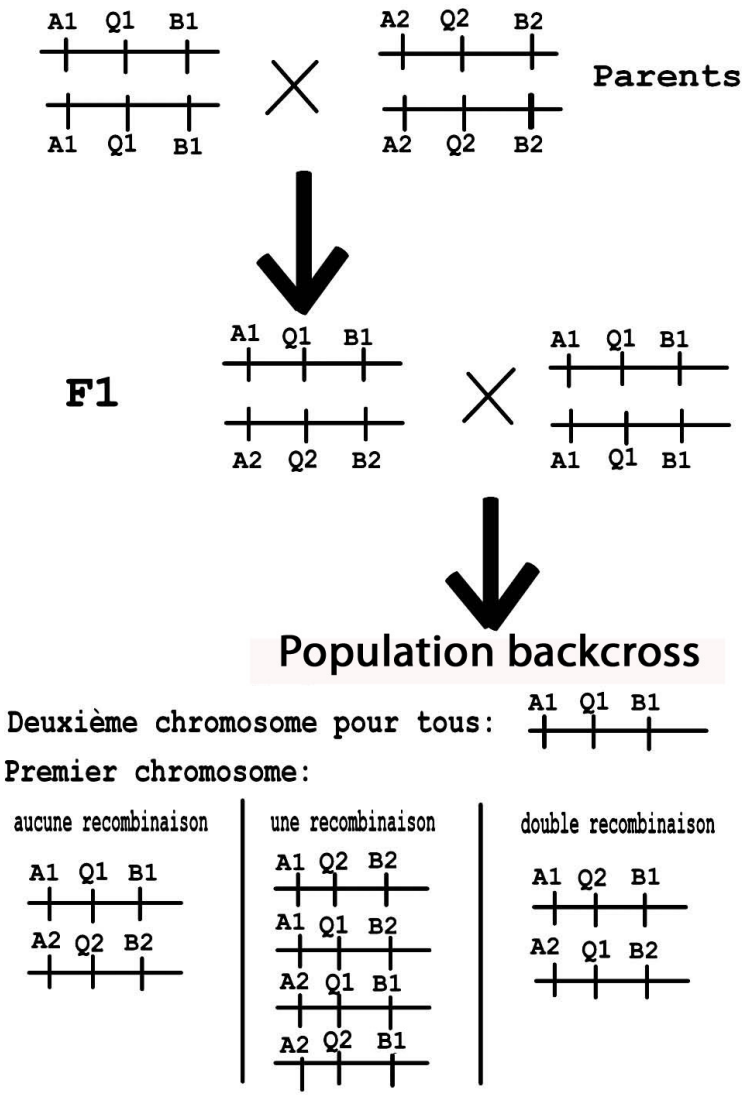
La position du QTL est inconnue

Contexte

- On modélise un chromosome par un segment $[0, T]$
- Considérons tout d'abord seulement deux marqueurs génétiques A et B présents aux extrémités du chromosome
- A et B possédent chacun deux allèles (A_1 et A_2 pour A et B_1, B_2 pour B)
- Existe-t-il un QTL Q (allèles Q_1 et Q_2) sur $[0, T]$? Si oui, à quelle position ?

On s'intéressera ici au backcross, schéma expérimental fondamental chez les végétaux

Un schéma expérimental : le backcross



Modèle

- X : variable aléatoire correspondant au génotype au QTL

On utilise le codage : $\begin{cases} 1 \text{ pour } Q_1Q_1 \\ -1 \text{ pour } Q_1Q_2 \end{cases}$

D'où

$$X = \begin{cases} 1 & \text{avec probabilité } 1/2 \\ -1 & \text{avec probabilité } 1/2 \end{cases}$$

- Y : variable aléatoire correspondant au phénotype

$$Y = \mu + q\,X + \varepsilon \qquad \text{où} \quad \varepsilon \sim N(0, \sigma^2)$$

L'Interval Mapping de Lander et Botstein (1989)

On souhaite tester : $H_0 : q = 0$ vs $H_1 : q \neq 0$

L'Interval Mapping

- Position du QTL inconnue

$\Rightarrow$ on scanne l'intervalle $[0, T]$.

$\Rightarrow$ tests du rapport de vraisemblance sur tout l'intervalle

Construction du LRT

- Pour chaque position $t \in [0, T]$, génotype au QTL inconnu

$\Rightarrow$ calcul des probabilités du génotype au QTL grâce aux recombinaisons et à la formule de Haldane (1919)

$\Rightarrow$ modèle de mélange

L'Interval Mapping de Lander et Botstein (1989)

- Vraisemblance pour n observations j iid :

$$L_n(\theta, t) = \prod_{j=1}^{n} p_t^j f_{(\mu+q,\sigma)}(y_j) \; + \; (1 - p_t^j) f_{(\mu-q,\sigma)}(y_j)$$

où :

- $\theta = (q, \mu, \sigma)$
- $f_{(\mu,\sigma)}(.)$ densité Gaussienne de moyenne μ et de variance σ^2
- p_t^j probabilité que l'individu j soit de génotype Q_1Q_1 en t, sachant son génotype aux marqueurs A et B

L'Interval Mapping de Lander et Botstein (1989)

- $\Lambda_n(t)$ LRT à la position t
- les $\Lambda_n(t)$ définissent un processus $\Lambda_n(.)$

On recherche un seul QTL sur l'intervalle $[0, T]$

$\Rightarrow$ statistique naturelle : $\sup \Lambda_n(.)$

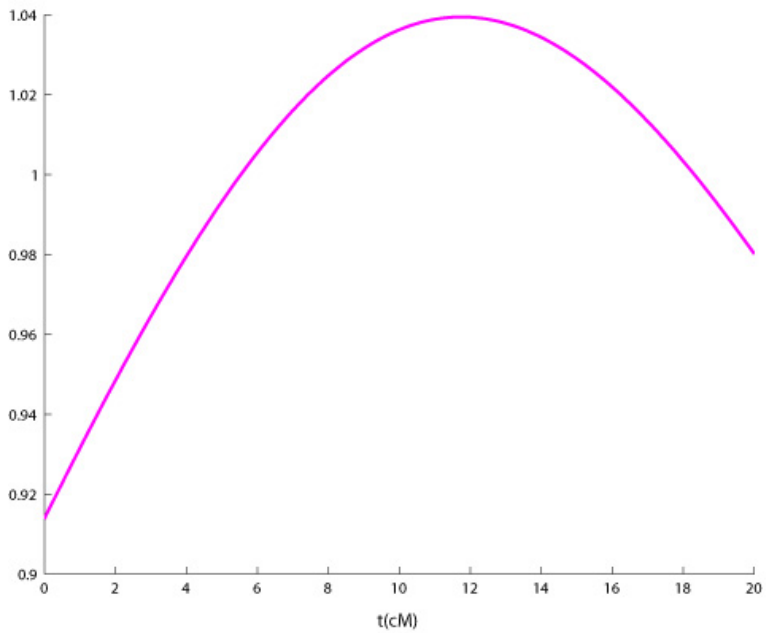


FIG.: Une trajectoire du processus $\Lambda_n(.)$ ($T = 20$cM)

Quelques précisions sur les hypothèses testées

H_0 : “il n’y a pas de QTL sur l’intervalle $[0, T]$”

$H_{at^\star}$: “le QTL est situé en $t^\star \in [0, T]$ avec un effet $q = a/\sqrt{n}$ ”

$\Rightarrow$ Théorie de Le Cam (1986)

Une interpolation non linéaire

Théorème

$$\Lambda_n(.) \overset{F.d.}{\rightarrow} \{Z(.)\}^2 \qquad \text{où}$$

- *$Z(.)$ est le processus d'interpolation non linéaire tel que*

$$\forall t \in [0,T] \quad Z(t) = \frac{\alpha(t)\, Z(0) \;+\; \beta(t)\, Z(T)}{\sqrt{\{\alpha(t)\}^2 + \{\beta(t)\}^2 + 2\alpha(t)\beta(t)e^{-2T}}}$$

$$\alpha(0) = 1, \beta(0) = 0, \alpha(T) = 0, \beta(T) = 1 \quad \text{et} \quad Cov\{Z(0), Z(T)\} = e^{-2T}$$

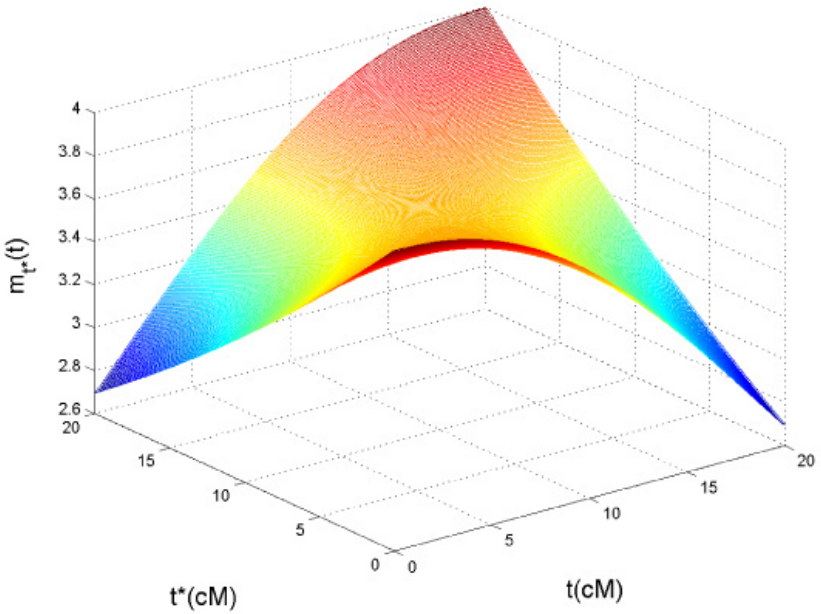
- *$Z(.)$ est un processus Gaussien de variance 1 et de fonction moyenne :*

$$\text{sous } H_0 : \; m(t) = 0 \;\; \forall t \in [0,T]$$

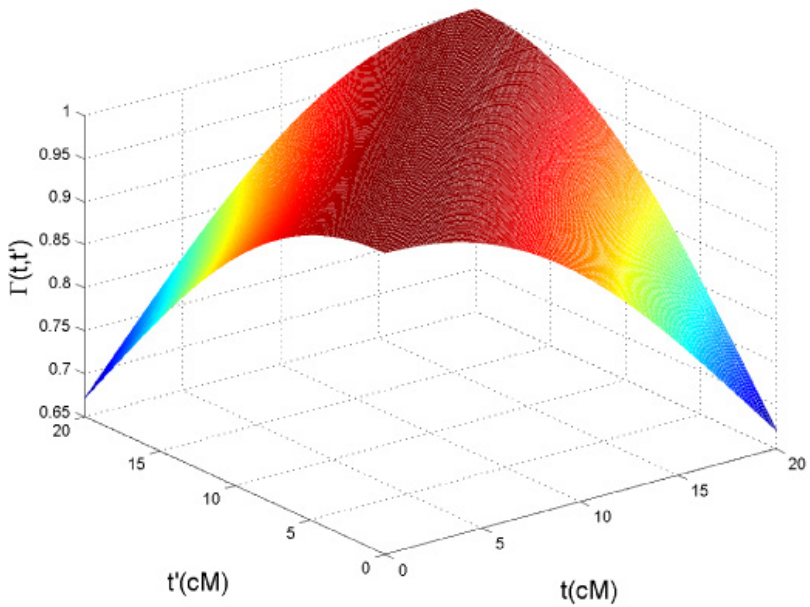
$$\text{sous } H_{at^\star} : \; m_{t^\star}(0) = \frac{a}{\sigma}\, g(0,t^\star) \;\; , \;\; m_{t^\star}(T) = \frac{a}{\sigma}\, g(T,t^\star)$$

$$\forall t \in [0,T] \quad m_{t^\star}(t) = \frac{\alpha(t)\, m_{t^\star}(0) \;+\; \beta(t)\, m_{t^\star}(T)}{\sqrt{\{\alpha(t)\}^2 + \{\beta(t)\}^2 + 2\alpha(t)\beta(t)e^{-2T}}}$$

Illustrations graphiques



Fonction moyenne



Fonction covariance

FIG.: Fonction moyenne et fonction covariance ($a = 4$, $\sigma = 1$, $T = 20$cM)

Illustrations graphiques

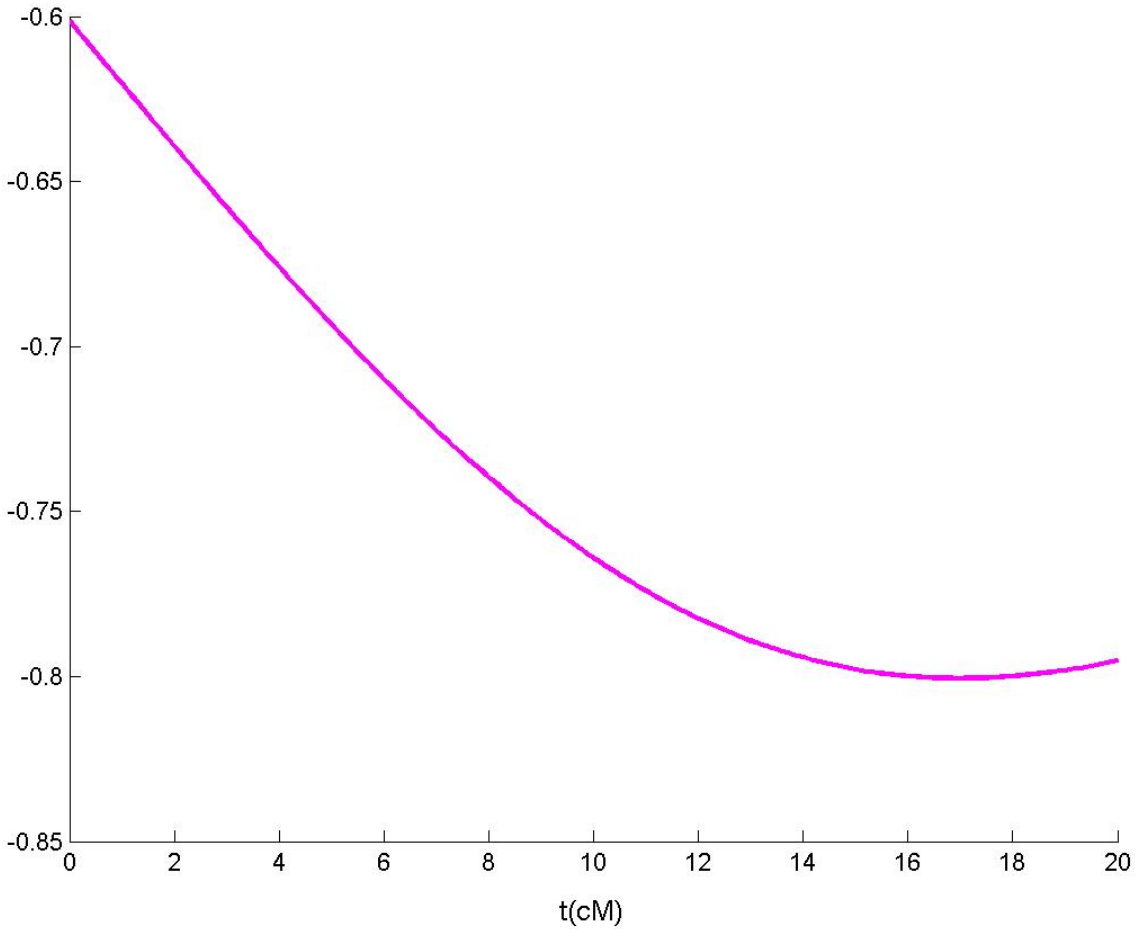


FIG.: Une trajectoire du processus $Z(.)$ sous H_0 ($T = 20$cM)

Illustrations graphiques

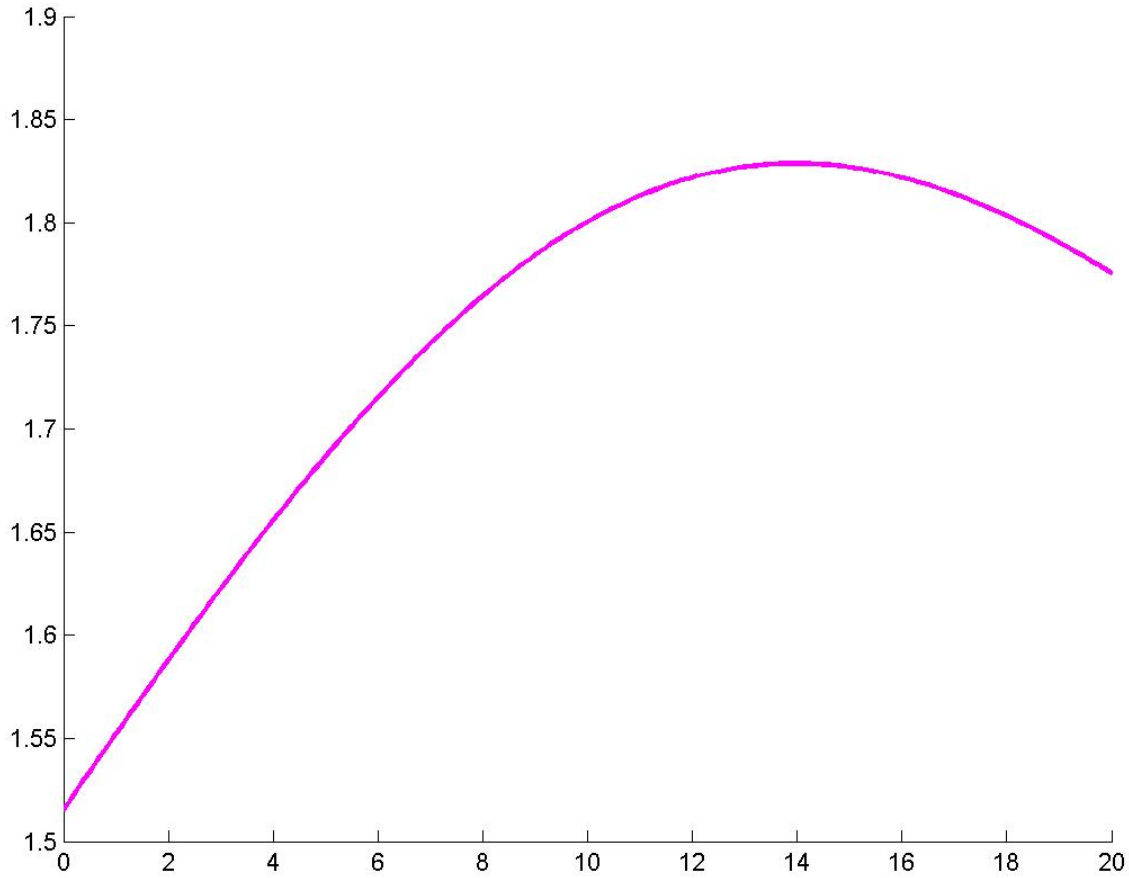


FIG.: Fonction moyenne ($a = 2$, $\sigma = 1$, $t^\star = 14$cM, $T = 20$cM)

Illustrations graphiques

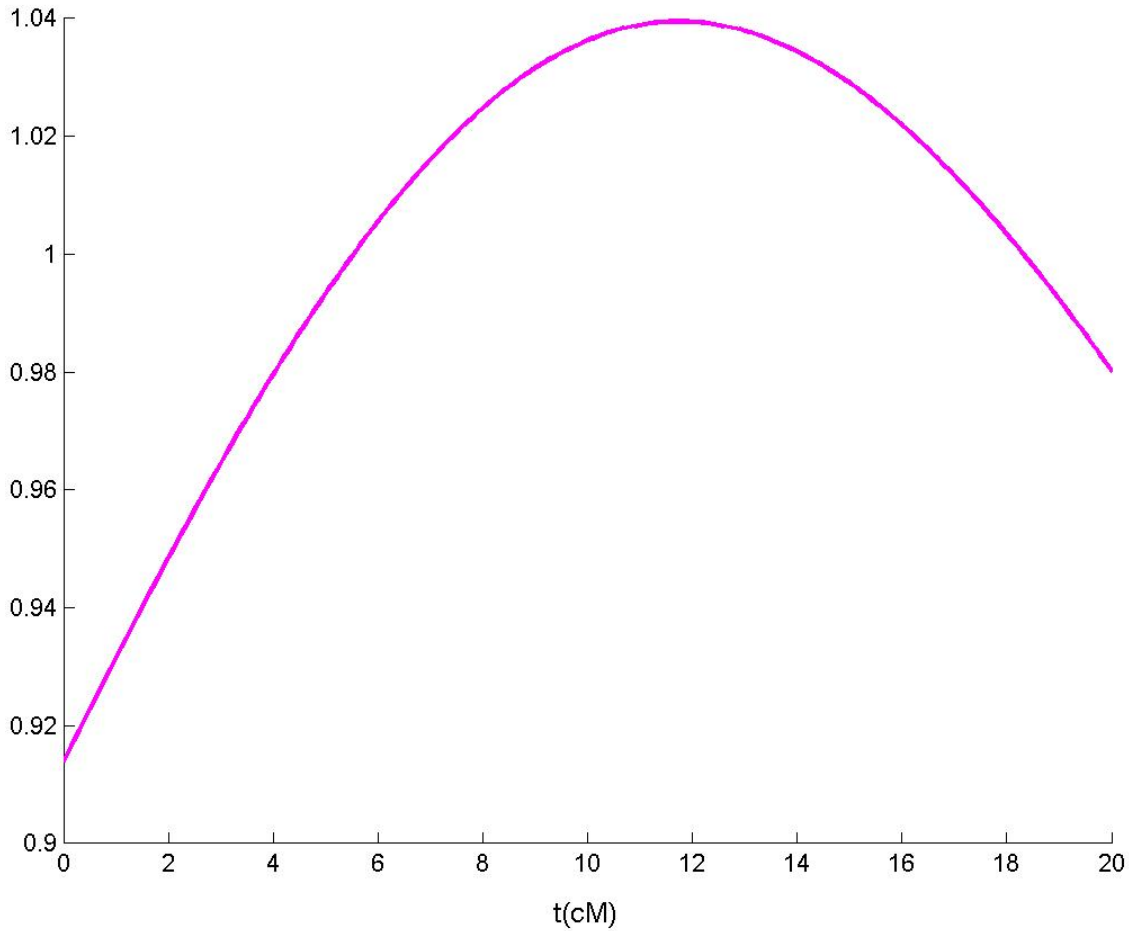


FIG.: Même trajectoire de $Z(.)$ que sous H_0 mais sous $H_{at^\star}$ ($a = 2$, $\sigma = 1$, $t^\star = 14$cM, $T = 20$cM)

Retour sur les poids du modèle de mélange

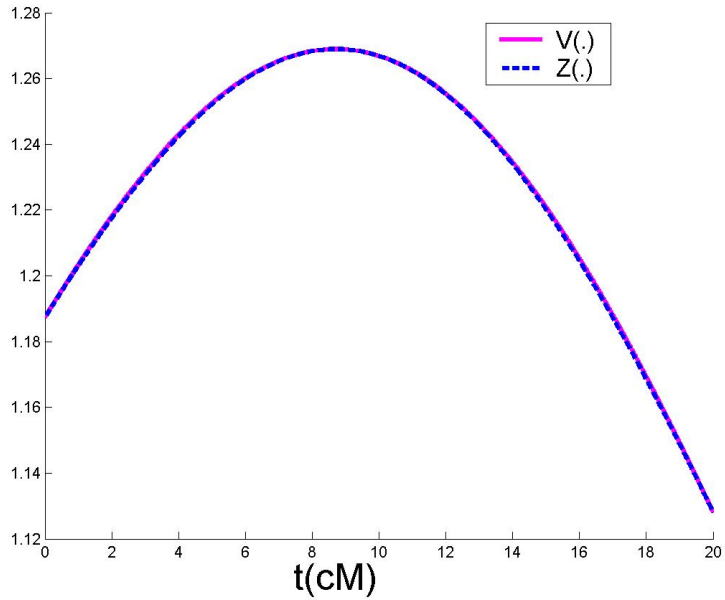
Ici, les poids du modèle de mélange considéré correspondent à la modélisation de Haldane

Modélisation de Haldane : le nombre de recombinaisons dans l'intervalle $[0, T]$ suit un processus de Poisson d'intensité 1

Dans Rebaï et al.(94 et 95), les auteurs n'autorisent qu'une seule recombinaison entre 2 marqueurs (Phénomène d'interférence)

$\Rightarrow$ nouveaux poids du modèle de mélange
$\Rightarrow$ le processus $\Lambda_n(.)$ tend vers le carré d'un processus d'interpolation linéaire $V(.)$

Interpolation linéaire/ Interpolation non linéaire



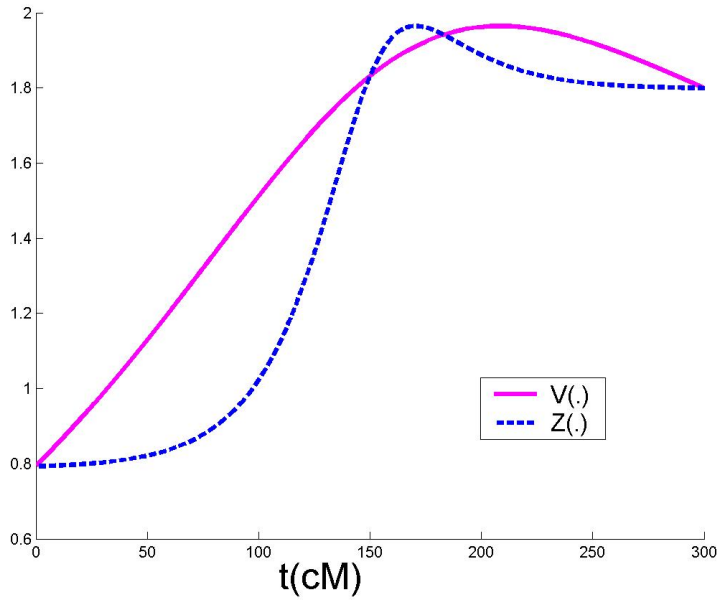


FIG.: Une trajectoire des processus $Z(.)$ et $V(.)$ sous H_0 ($T = 20$cM à gauche, $T = 300$cM à droite)

A propos des tests multiples

Lemme

On rappelle que $V(0) = Z(0)$ *et* $V(T) = Z(T)$.

Soient ξ *et* ξ' *tels que* $\xi = \dfrac{T\ \left\{e^{-2T}\ V(0) - V(T)\right\}}{\left\{e^{-2T} - 1\right\}\left\{V(0) + V(T)\right\}}$ *et* $\dfrac{T\,\beta(\xi')}{\alpha(\xi') + \beta(\xi')} = \xi$,

alors sous H_0 *et* $H_{at^\star}$

$$\left\{Z(\xi')\right\}^2 = \left\{V(\xi)\right\}^2 = \frac{\left\{V(0)\right\}^2 + \left\{V(T)\right\}^2 - 2\ e^{-2T}\ V(0)\ V(T)}{\left\{1 + e^{-2T}\right\}\left\{1 - e^{-2T}\right\}}$$

et

$$\sup_{t\in[0,T]} \left\{Z(t)\right\}^2 = \sup_{t\in[0,T]} \left\{V(t)\right\}^2$$

$$= \left\{V(\xi)\right\}^2\ 1_{\frac{V(T)}{V(0)} \in\ ]\ e^{-2T}\ ,\ e^{2T}\ [} + \max\left[\left\{V(0)\right\}^2\ ,\ \left\{V(T)\right\}^2\right] 1_{\frac{V(T)}{V(0)} \notin\ ]\ e^{-2T}\ ,\ e^{2T}\ [}$$

Inutile d'effectuer des tests partout sur le chromosome !

L'Interval Mapping lisse les trajectoires

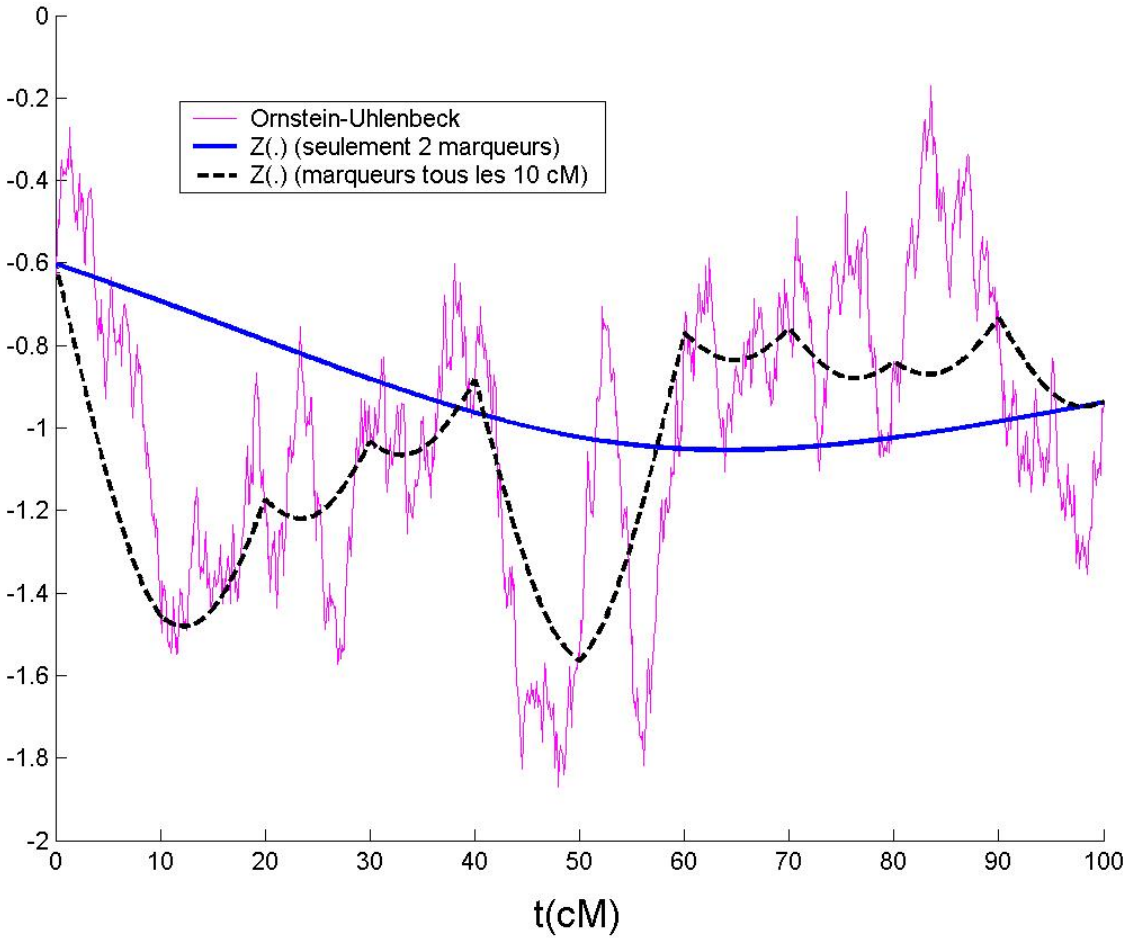


FIG.: 3 processus Gaussiens ($T = 100$cM)

Application au calcul de valeurs critiques

Calcul de la valeur critique c vérifiant $P_{H_0}(\sup\{Z(.)\}^2 > c) = 1 - \alpha$

$\Rightarrow$ fonction QSIMVNEF de Genz (1992)

Méthode	*la notre*	*Rebaï*	*Feingold*
Valeur critique	8.23	9.09	8.26

FIG.: Valeurs critiques en fonction de la méthode considérée (51 marqueurs positionnés tous les 2cM, $T = 1$M, $\alpha = 95\%$)

Méthode	*la notre*	*Feingold*
Valeur critique	5.40	5.78

FIG.: Valeurs critiques en fonction de la méthode considérée (2 marqueurs, $T = 1$M, $\alpha = 95\%$)

Approche multi-QTL

$H_{\vec{at}^\star}$: "il existe M QTL situés en $t_1^\star, ..., t_M^\star$ avec des effets $q_1 = \frac{a_1}{\sqrt{n}}, ..., q_M = \frac{a_M}{\sqrt{n}}$"

On supposera les effets QTL additifs

Approche multi-QTL

Théorème

$$\Lambda_n(.) \overset{F.d.}{\rightarrow} \{Z^\star(.)\}^2 \qquad \text{où}$$

- *$Z^\star(.)$ est le processus d'interpolation non linéaire tel que*

$$\forall t \in [0,T] \quad Z^\star(t) = \frac{\alpha(t)\, Z^\star(0) \;+\; \beta(t)\, Z^\star(T)}{\sqrt{\{\alpha(t)\}^2 + \{\beta(t)\}^2 + 2\alpha(t)\beta(t)e^{-2T}}}$$

$$\alpha(0) = 1, \beta(0) = 0, \alpha(T) = 0, \beta(T) = 1 \quad \text{et} \quad Cov\{Z^\star(0), Z^\star(T)\} = e^{-2T}$$

- *$Z^\star(.)$ est un processus Gaussien de variance 1 et de fonction moyenne :*

$$\text{sous } H_0 : \; m(t) = 0 \;\; \forall t \in [0,T]$$

$$\text{sous } H_{at^{\vec{\star}}} : \; m_{t^{\vec{\star}}}(0) = \sum_{s=1}^{M} \frac{a_s}{\sigma}\, g(0, t_s^\star) \;\; , \;\; m_{t^{\vec{\star}}}(T) = \sum_{s=1}^{M} \frac{a_s}{\sigma}\, g(T, t_s^\star)$$

$$\forall t \in [0,T] \quad m_{t^{\vec{\star}}}(t) = \frac{\alpha(t)\, m_{t^{\vec{\star}}}(0) \;+\; \beta(t)\, m_{t^{\vec{\star}}}(T)}{\sqrt{\{\alpha(t)\}^2 + \{\beta(t)\}^2 + 2\alpha(t)\beta(t)e^{-2T}}}$$

Population avec une structure de famille

On admettra qu'une population backcross = une famille de père

$\Rightarrow$ généralisation à une **population comprenant I familles de pères**

H_0 : "Il n'y a de QTL dans aucune des familles"

$H_{at^\star}$: "Un QTL est présent dans au moins une famille "

$$\Lambda_n(.) \overset{F.d.}{\rightarrow} \sum_{i=1}^{I} \left\{ Z^i(.) \right\}^2$$

Les $Z^i(.)$ sont des processus d'interpolation non linéaires indépendants

Chaque $Z^i(.)$ est :

- centré en l'absence de QTL dans la famille i
- décentré en présence de QTL dans la famille i

Processus de Chi deux d'Ornstein-Uhlenbeck

Hypothèse de carte dense

- Cas d'une seule famille :

$\Lambda_n(.)$ converge vers le carré d'un processus d'Ornstein-Uhlenbeck

Lander et Botstein (1989) et Cierco (1998)

- Cas de l familles :

$\Lambda_n(.)$ converge vers le carré d'un processus de Chi deux d'Ornstein-Uhlenbeck à l degrés de liberté

Processus de Chi deux d'Ornstein-Uhlenbeck

Définition (Processus d'Ornstein-Uhlenbeck)

Un processus O.U. est un processus Gaussien stationnaire, de moyenne nulle, de variance égale à un, et de covariance égale à $r(t) = exp(-2 \mid t \mid)$

Définition (Processus de Chi deux d'Ornstein-Uhlenbeck)

Soient $\tilde{Z}^1(.), ..., \tilde{Z}^I(.)$ *I processus O.U. indépendants.*

$S(t) = \sum_{i=1}^{I} \left\{\tilde{Z}^i(.)\right\}^2$ *est appelé Processus de Chi deux d'Ornstein-Uhlenbeck à I degrés de liberté*

Processus de Chi deux d'Ornstein-Uhlenbeck

On établit la relation :

$$\sup_{t\in[0,\ T]} S(t) = \sup_{t\in[1,\ e^{4T}]} \left(\frac{\left\| \vec{W}(t) \right\|}{\sqrt{t}} \right)^2$$

avec $\vec{W}(t) = \begin{pmatrix} W_1(t) \\ \vdots \\ W_l(t) \end{pmatrix}$ mouvement brownien en dimension l.

Ainsi, pour le calcul de valeurs critiques, on dispose :

- des tables de Delong (81) et de Estrella (2003)
- de la formule approximative de Delong (81) à condition que c et T soient grands

$$\mathbb{P}\left(\sup_{t\in[0,\ T]} S(t) < c \right) = \frac{(c/2)^{l/2}\, e^{-c/2}}{\Gamma(d/2)} \left[4T(1 - \frac{l}{c}) + \frac{2}{c} + O(\frac{1}{c^2}) \right]$$

- d'une borne inf, obtenue par MCQMC (en collaboration avec Alan Genz)

Conclusions et perspectives

Selective Genotyping :

- On doit génotyper le même pourcentage d'individus aux deux extrêmes
- Il n'y a pas d'information dans la bande
- Mêmes conclusions pour un selective genotyping avec deux caractères corrélés

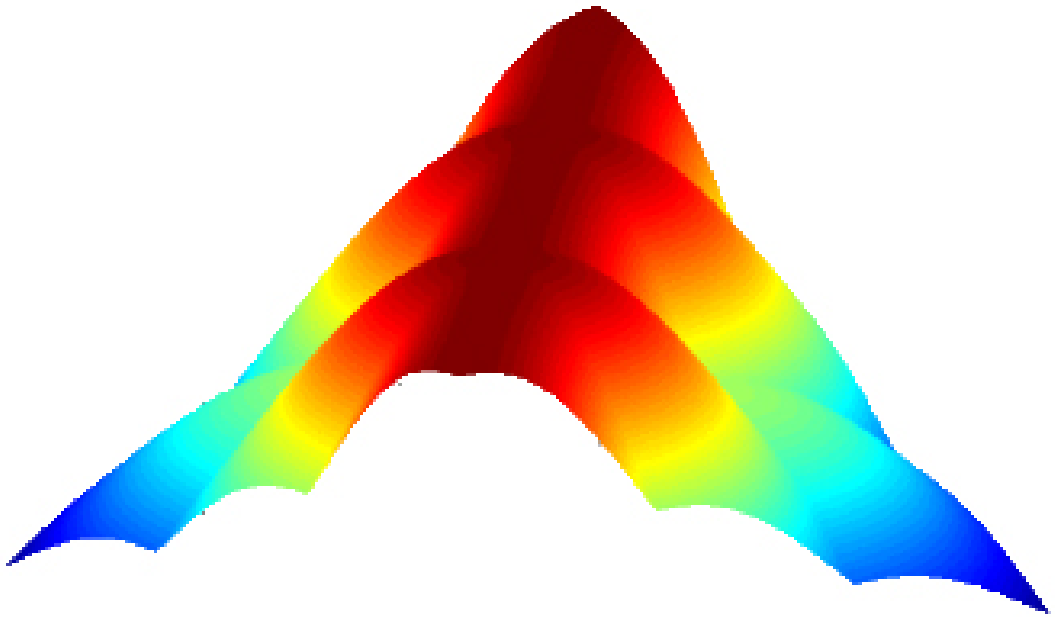
Génome Scan :

- L'Interval Mapping est une interpolation non linéaire
- On doit effectuer un seul test au maximum entre deux marqueurs
- Si Carte dense et I familles :
  $\Rightarrow$ Processus de Chi deux d'Ornstein-Uhlenbeck à I degrés de liberté

Perspectives :

- Sélection de modèle pour la recherche de plusieurs QTL
- Génome Scan couplé au Selective Genotyping
- Extension à l'analyse d'association

Merci de votre attention



Introduction d'un deuxième phénotype

2 phénotypes sont désormais disponibles : Y et Z

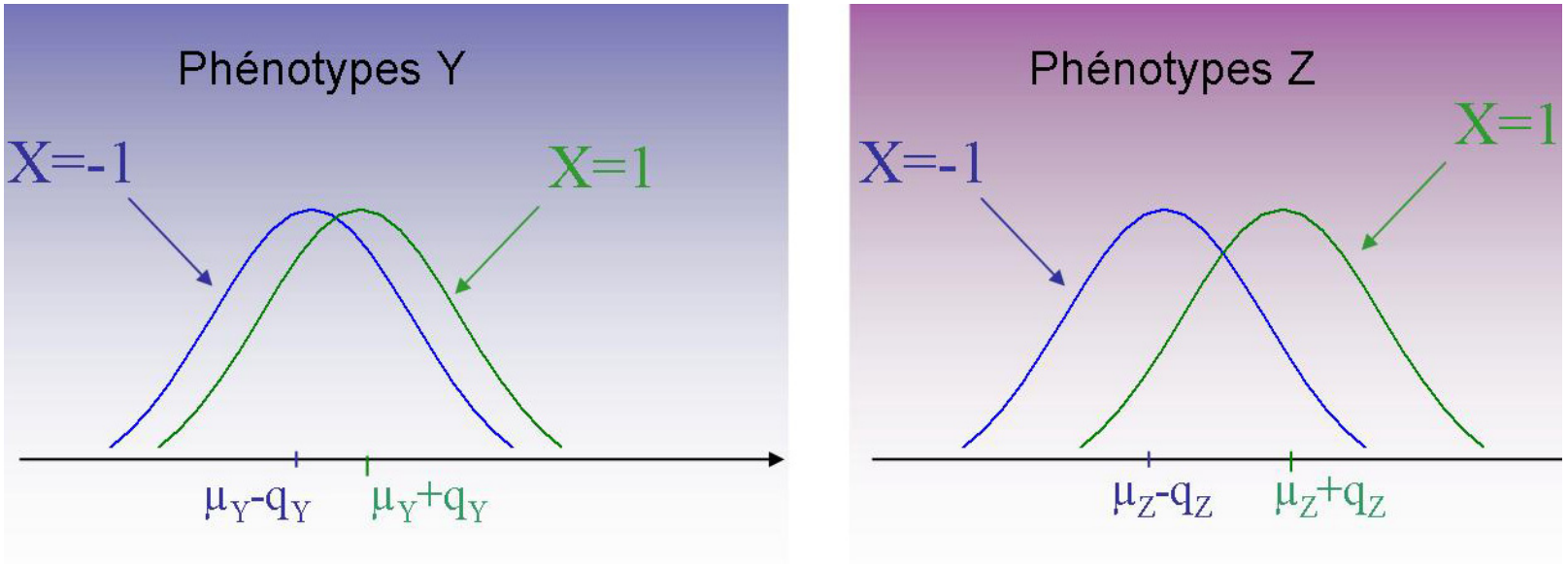


FIG.: Distribution des phénotypes Y et Z

Y/X et Z/X corrélées

Modèle en l'absence de censure

Les phénotypes Y et Z sont tels que :

$$\begin{pmatrix} Y \\ Z \end{pmatrix} = \begin{pmatrix} \mu_Y + q_Y X \\ \mu_Z + q_Z X \end{pmatrix} + \varepsilon$$

où

$$\varepsilon \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & r\,\sigma^2 \\ r\,\sigma^2 & \sigma^2 \end{pmatrix} \right)$$

On supposera :

- r et σ^2 connus
- $r \notin \{-1, 1\}$

Existe-t-il un QTL affectant le phénotype Z ?

Test statistique oracle (μ_Z, q_Z)

- A l'aide de n observations (X_j, Y_j, Z_j) iid, on souhaite tester :

$$H_{0Z} : q_Z = 0 \ vs \ H_{1Z} : q_Z \neq 0$$

On considère une alternative locale $H_{bZ} : q_Z = \frac{b}{\sqrt{n}}$

- Test statistique oracle :

$$T = \frac{\sum_{j=1}^{n} \frac{1}{p}\left(Z_j - \overline{Z}\right) 1_{X_j=1} - \frac{1}{1-p}\left(Z_j - \overline{Z}\right) 1_{X_j=-1}}{\sigma\sqrt{\frac{n}{p(1-p)}}}$$

$$T \overset{H_{0Z}}{\rightarrow} N(0,\ 1) \qquad T \overset{H_{bZ}}{\rightarrow} N\left(\frac{2\,b\,\sqrt{p(1-p)}}{\sigma},\ 1\right)$$

Selective Genotyping en présence de deux caractères corrélés

Génotyper coûte très cher

Le phénotype Z est difficile à mesurer pour des raisons biologiques

$\Rightarrow$ Selective Genotyping effectué sur Y

$\Rightarrow$ Z mesuré uniquement pour les individus présentant un phénotype Y extrême

Existe-t-il un QTL affectant le phénotype Z ?

Comparaison de deux stratégies (μ_Z, q_Z, μ_Y, q_Y)

2 stratégies pour l'analyse de données en Selective Genotyping :

1. Test de Wald basé sur l'ensemble des phénotypes
2. Test de Wald basé sur les phénotypes extrêmes

$$\tilde{\kappa}_1 = \tilde{\kappa}_2 = \left\{ \frac{1 - r^2}{\gamma} + \frac{r^2}{\kappa_1} \right\}^{-1}$$

où $\kappa_1 = \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-})$

$\forall p, \forall S_+, \forall S_-$, il n'existe pas d'information dans la bande !

$\forall p$, on doit génotyper le même pourcentage d'individus à "droite" qu'à "gauche" !

Illustration graphique

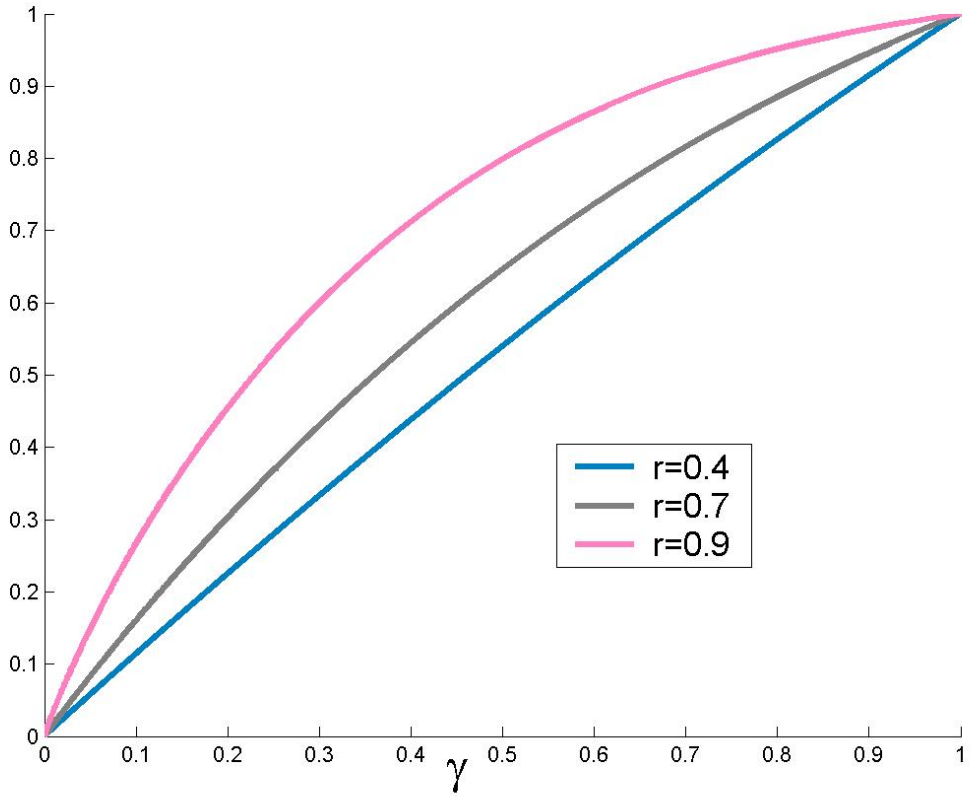


FIG.: Efficacité en fonction de γ et de r ($\gamma_+ = \gamma_- = \frac{\gamma}{2}$)

Doit-on analyser les familles simultanément ?

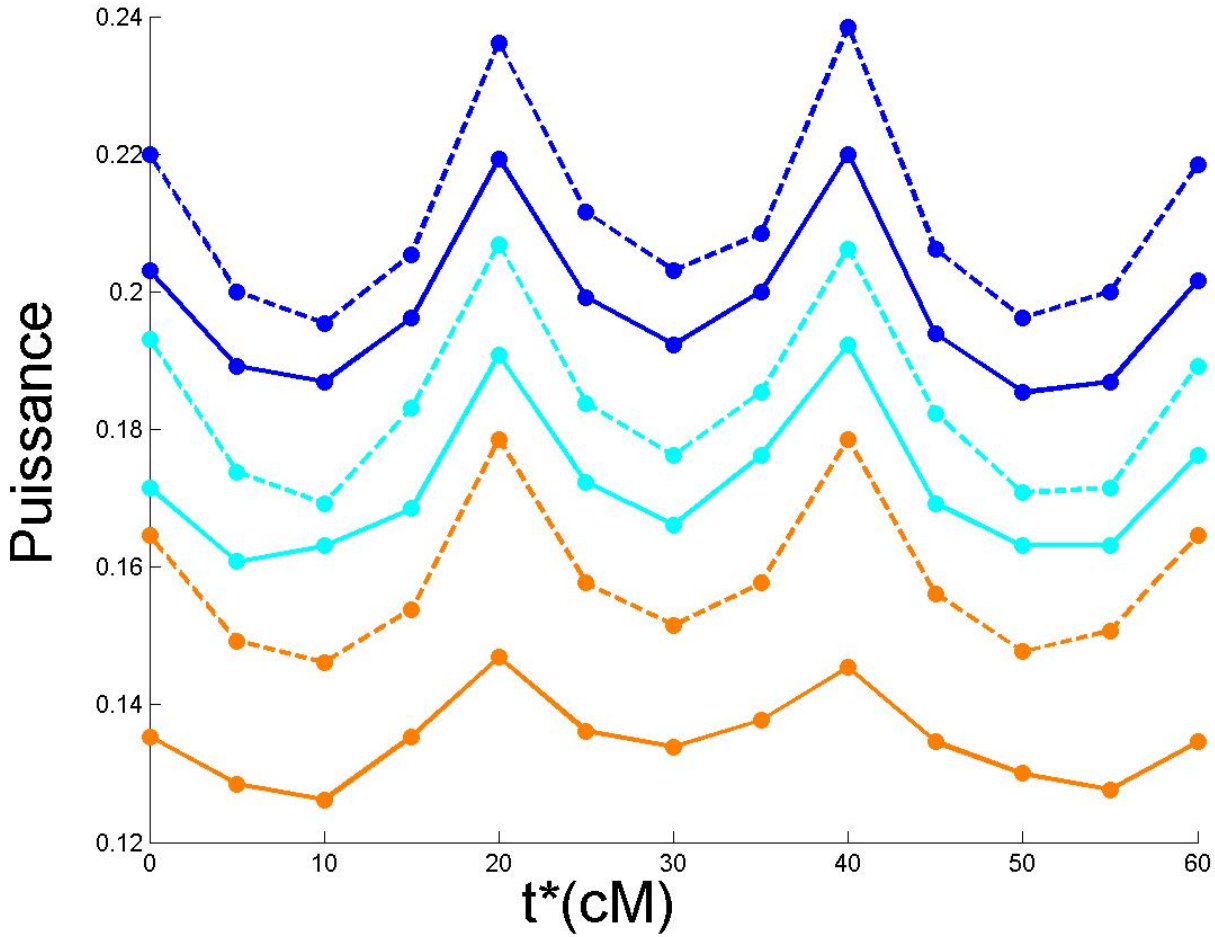


FIG.: Puissance approche globale (ligne continue) vs puissance approche Bonferroni (ligne pointillée). Une seule famille présente un QTL. Orange (I=12), Cyan (I=7), Bleu (I=5).