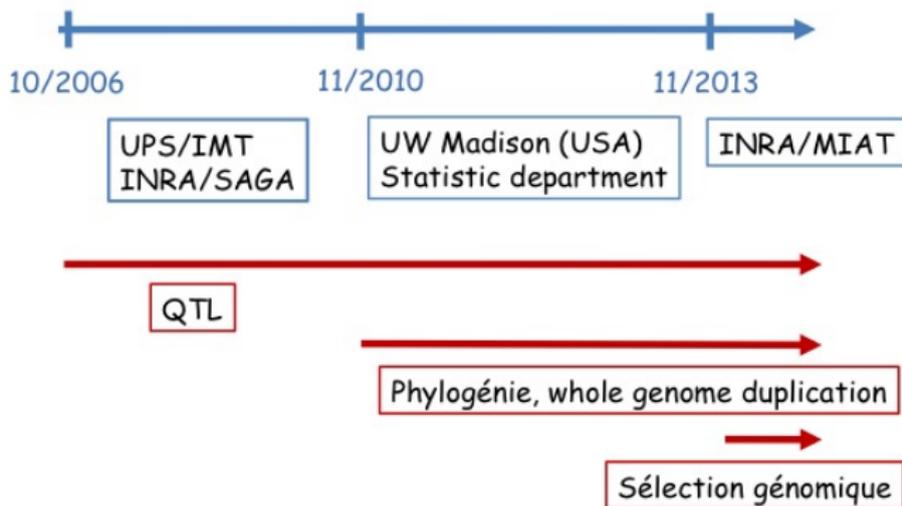


# Processus Gaussiens et de Chi-Deux pour la détection de gènes

Charles-Elie Rabier

24 Février 2015



Thème de recherche: Méthodes statistiques et probabilistes pour le génome

# 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

⇒ les méthodes statistiques vont nous permettre de détecter et localiser le QTL

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

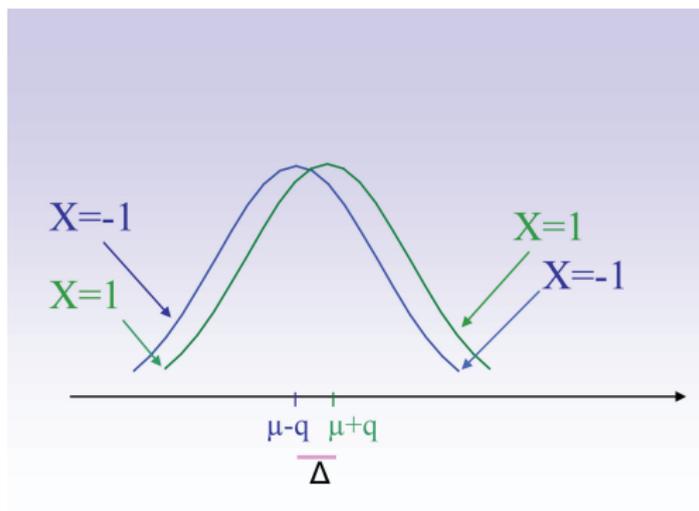
$$X = \begin{cases} -1 & \text{avec probabilité } 1 - p \\ 1 & \text{avec probabilité } p \end{cases}$$

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

- $Y$  : variable aléatoire correspondant au phénotype

$$Y = \mu + qX + \sigma\varepsilon \quad \text{où } \varepsilon \sim N(0, 1)$$

# Modèle en l'absence de censure



Distribution des phénotypes  $Y$

# Test statistique oracle $(\mu, q, \sigma)$

- A l'aide de  $n$  observations  $(X_j, Y_j)$  i.i.d., on souhaite tester :

$$H_0 : q = 0 \text{ vs } H_1 : q \neq 0$$

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

- Test statistique oracle :

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

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

# Selective Genotyping

## Génotyper coûtait cher

⇒ 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)*

*Darvasi et Soller (Theoretical and Applied Genetics, 1992)*

Gène de l'obésité chez le porc (Fontanesi et al. 2012)

Gène pour la longueur de la coquille d'un coquillage Mérétrix (Lu et al. 2013)

# Modèle correspondant au Selective Genotyping

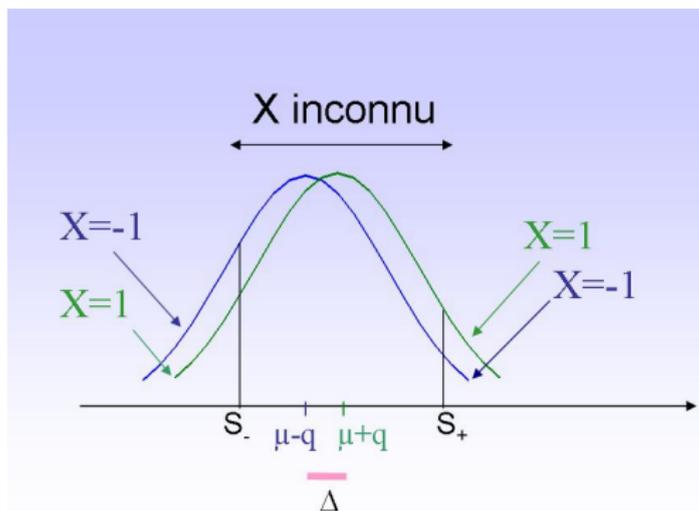
$X$  disponible uniquement pour les individus présentant un phénotype extrême  $Y$

⇒ On n'observe plus  $X$  mais  $\bar{X}$  :

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

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

# Modèle correspondant au Selective Genotyping



Distribution des phénotypes  $Y$

# Test de Wald ( $\mu, q, \sigma$ )

$$H_0 : q = 0 \text{ vs } H_1 : q \neq 0$$

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

- Statistique de Wald

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

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

$$\mathcal{A} = E_{H_0} \left[ (Y - \mu)^2 1_{\bar{X} \neq 0} \right] = \sigma^2 \left\{ \gamma + z_{\gamma+} \varphi(z_{\gamma+}) - z_{1-\gamma-} \varphi(z_{1-\gamma-}) \right\}$$

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

# Optimisation du génotypage

- On souhaite génotyper uniquement un pourcentage  $\gamma$  de la population

⇒ Comment choisir les  $\gamma_+$  et  $\gamma_-$  optimaux ?

$\forall p$ ,  $\kappa_1$  atteint son maximum  $M$  pour  $\gamma_+ = \gamma_- = \gamma/2$

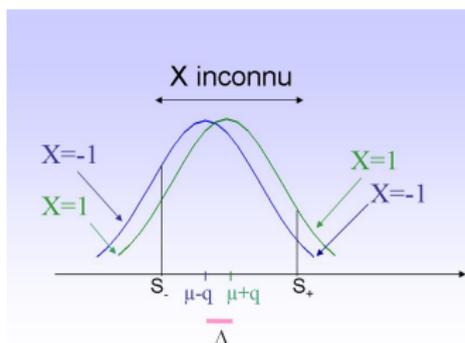
$$M = \gamma + 2 z_{\gamma/2} \varphi(z_{\gamma/2})$$

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

# 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



*Rabbee, Speca, Armstrong, Speed (Genet. Res. Camb., 2004)*

# Comparaison des 3 stratégies ( $\mu$ , $q$ , $\sigma$ )

## Lemme

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

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

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

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

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

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

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

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

# Conclusions sur le Selective Genotyping

- On doit génotyper le même pourcentage d'individus aux deux extrêmes
- Les phénotypes non extrêmes n'apportent pas d'information
- Le génotypage optimal est d'environ 30% (dépend du ratio coût génotypage/phénotypage)
- Le test de comparaison de moyenne est optimal

*R., JSPI 2013*

Deuxième partie :

Génome Scan

La position du QTL est inconnue

# Contexte

- On modélise un chromosome par un segment  $[0, T]$
- La distance sur  $[0, T]$  est appelée distance génétique
- $X(\cdot)$  : génome d'un individu
- On considère le modèle Poissonien de Haldane

# Modélisation de Haldane

- Pas d'interférence lors des crossing overs
- $N(\cdot)$  processus de Poisson standard sur  $[0, T]$

$$X(0) = \begin{cases} 1 & \text{avec probabilité } 1/2 \\ -1 & \text{avec probabilité } 1/2 \end{cases}$$

$$X(t) = X(0)(-1)^{N(t)}$$

- Les calculs sur le processus de Poisson conduisent à

$$\begin{aligned} r(t, t') &= \mathbb{P}(X(t)X(t') = -1) = \mathbb{P}(|N(t) - N(t')| \text{ impair}) \\ &= \frac{1}{2} (1 - e^{-2|t-t'|}) = \frac{1}{2} (1 - \rho(t, t')) \end{aligned}$$

# Modélisation de Haldane

- $t^*$  : position du QTL.
- $Y$  : variable aléatoire correspondant au phénotype

$$Y = \mu + q X(t^*) + \sigma \varepsilon \quad \text{où } \varepsilon \sim N(0, 1)$$

- Information génome  $X(\cdot)$  disponible seulement a des positions fixes, appelés marqueurs génétiques
- $K$  marqueurs génétiques sur  $[0, T]$  en

$$t_1 = 0 < t_2 < \dots < t_K = T$$

# Configuration classique (oracle)

Une observation est le couple

$$(Y, X(t_1), \dots, X(t_K)) .$$

et le challenge est dans le fait que  $t^*$  est inconnu !!!

# 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  $t^*$  du QTL inconnue

⇒ on scanne l'intervalle  $[0, T]$

⇒ tests du rapport de vraisemblance sur tout l'intervalle

## Construction du LRT

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

⇒ calcul des probabilités du génotype au QTL grâce aux recombinaisons et à la formule de Haldane

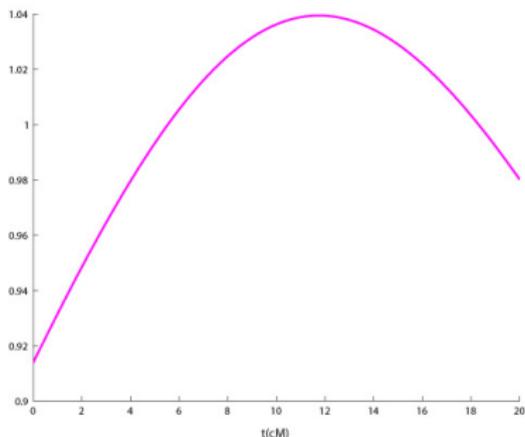
⇒ modèle de mélange de mélange

# L'Interval Mapping de Lander et Botstein (1989)

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

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

⇒ LRT sur tout le chromosome :  $\sup \Lambda_n(\cdot)$



Une trajectoire du processus  $\Lambda_n(\cdot)$  ( $T = 20\text{cM}$ ,  $K = 2$ )

# Configuration selective genotyping

$\bar{X}(t)$  est la variable aléatoire telle que

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

alors, dans notre cas, une observation sera le couple

$$\left( Y, \bar{X}(t_1), \dots, \bar{X}(t_K) \right).$$

# Vraisemblance du couple $(Y, \bar{X}(t_1), \dots, \bar{X}(t_K))$

En  $t = t^*$

$$L_{t^*}(\theta) = [p(t^*) f_{(\mu+q,\sigma)}(Y) 1_{Y \notin [S_-, S_+]} + \{1 - p(t^*)\} f_{(\mu-q,\sigma)}(Y) 1_{Y \notin [S_-, S_+]} + \frac{1}{2} f_{(\mu+q,\sigma)}(Y) 1_{Y \in [S_-, S_+]} + \frac{1}{2} f_{(\mu-q,\sigma)}(Y) 1_{Y \in [S_-, S_+]}] g(\cdot)$$

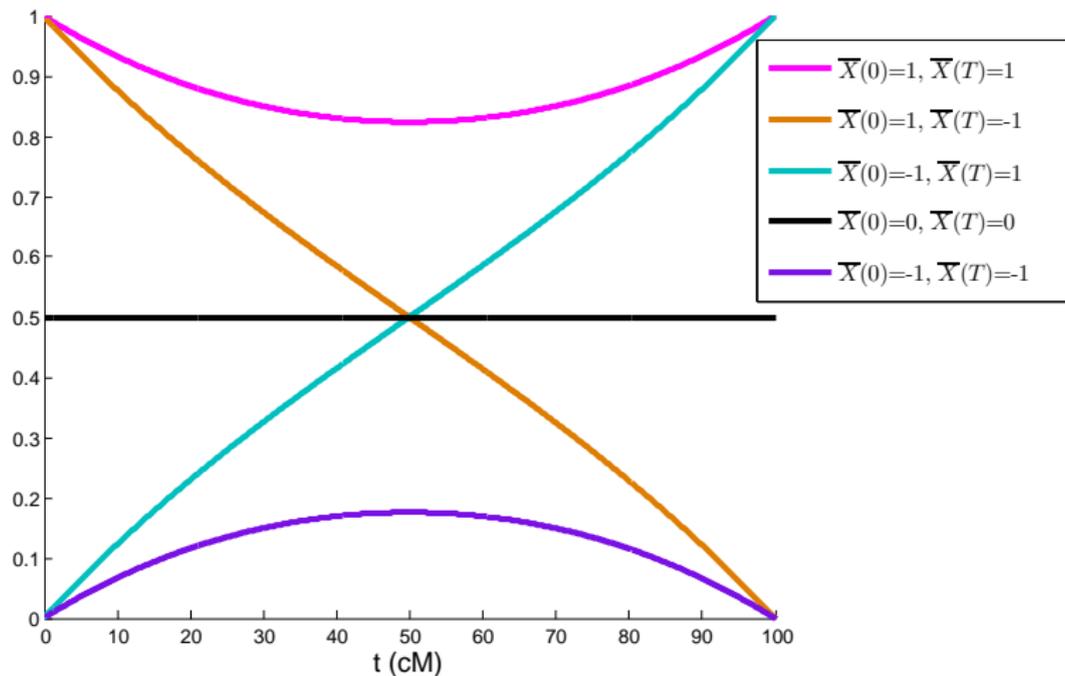
où

- $f_{(m,\sigma)}$  est la densité Gaussienne avec paramètres  $(m, \sigma)$
- $g(\cdot)$  est une fonction qui ne dépend pas des paramètres  $\mu$ ,  $q$  et  $\sigma$

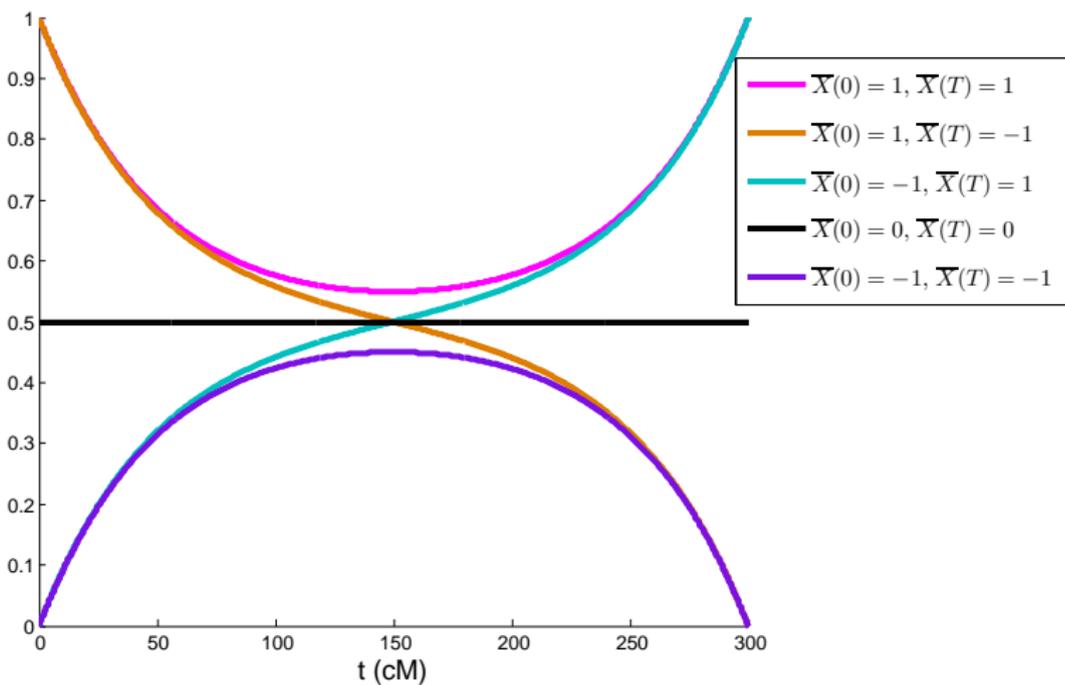
et

$$p(t^*) = Q_{t^*}^{1,1} 1_{\bar{X}(t^{*\ell})=1} 1_{\bar{X}(t^{*r})=1} + Q_{t^*}^{1,-1} 1_{\bar{X}(t^{*\ell})=1} 1_{\bar{X}(t^{*r})=-1} + Q_{t^*}^{-1,1} 1_{\bar{X}(t^{*\ell})=-1} 1_{\bar{X}(t^{*r})=1} + Q_{t^*}^{-1,-1} 1_{\bar{X}(t^{*\ell})=-1} 1_{\bar{X}(t^{*r})=-1}$$

# Illustration des différents poids ( $K = 2$ , $T = 1M$ )



# Illustration des différents poids ( $K = 2$ , $T = 3M$ )



# Statistiques du score et du LRT

- $\theta = (q, \mu, \sigma)$  paramètre du modèle à  $t$  fixe.
- $\theta_0 = (0, \mu, \sigma)$  représente  $H_0$

## Statistique du score en $t$

$$S_n(t) = \frac{\frac{\partial l_t^n}{\partial q} |_{\theta_0}}{\sqrt{\mathbb{V} \left( \frac{\partial l_t^n}{\partial q} |_{\theta_0} \right)}} ,$$

avec  $l_t^n(\theta)$  log vraisemblance en  $t$ , associée à  $n$  observations.

## Statistique du LRT en $t$

$$\Lambda_n(t) = 2 \left\{ l_t^n(\hat{\theta}) - l_t^n(\hat{\theta}_{|H_0}) \right\} ,$$

où  $\hat{\theta}$  est l'EMV, et  $\hat{\theta}_{|H_0}$  l'EMV sous  $H_0$ .

- $t^*$  connu  $\Rightarrow$  modèle **régulier**
- $t^*$  inconnu  $\Rightarrow$  modèle **irrégulier** (sous  $H_0$ , l'information de fisher par rapport à  $t$  est nulle)

# Quelques précisions sur les hypothèses testées

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

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

# Etude du processus de score sous $H_0$

$(K = 2, t_1 = 0, t_2 = T)$

## Lemme

On a la relation suivante :

$$\{2p(t) - 1\} 1_{Y_j \notin [s_-, s_+]} = \alpha(t) \bar{X}(0) + \beta(t) \bar{X}(T)$$

avec  $\alpha(t) = Q_t^{1,1} - Q_t^{-1,1}$  et  $\beta(t) = Q_t^{1,1} - Q_t^{1,-1}$ .

$$\begin{aligned} \frac{\partial l_t^n}{\partial q} \Big|_{\theta_0} &= \sum_{j=1}^n \frac{Y_j - \mu}{\sigma^2} \{2p_j(t) - 1\} 1_{Y_j \notin [s_-, s_+]} \\ &= \frac{\alpha(t)}{\sigma} \sum_{j=1}^n \varepsilon_j \bar{X}_j(0) + \frac{\beta(t)}{\sigma} \sum_{j=1}^n \varepsilon_j \bar{X}_j(T) \end{aligned}$$

Le processus limite est un processus interpolé !!!

Etude du processus de score sous  $H_{at^*}$  $(K = 2, t_1 = 0, t_2 = T)$ 

Comme notre modèle est différentiable en moyenne quadratique, on utilise Theorem 7.2 of Van der Vaart (98). Sous  $H_0$ , le ratio de log vraisemblance vérifie

$$l_{t^*}^n(\theta) - l_{t^*}^n(\theta_0) = \frac{a}{\sqrt{n}} \frac{\partial l_{t^*}^n}{\partial \mathbf{q}} \Big|_{\theta_0} - \frac{a^2}{2} \mathbb{E}_{H_0} \left\{ \left( \frac{\partial l_{t^*}^n}{\partial \mathbf{q}} \Big|_{\theta_0} \right)^2 \right\} + o_P(1)$$

où  $o_P(1)$  est une séquence qui converge en probabilité vers 0.

$$\begin{aligned} & l_{t^*}^n(\theta) - l_{t^*}^n(\theta_0) \\ &= \frac{a}{\sigma\sqrt{n}} \left\{ \alpha(t^*) \sum_{j=1}^n \varepsilon_j \bar{X}_j(0) + \beta(t^*) \sum_{j=1}^n \varepsilon_j \bar{X}_j(T) \right\} \\ & - \frac{a^2}{2\sigma^4} \mathcal{A} \left\{ \alpha^2(t^*) + \beta^2(t^*) + 2\alpha(t^*)\beta(t^*)\rho(0, T) \right\} + o_P(1) \end{aligned}$$

# Etude du processus de score sous $H_{at^*}$

$(K = 2, t_1 = 0, t_2 = T)$

Comme  $\alpha(t^*) + \beta(t^*)\rho(0, T) = \rho(0, t^*)$ ,

$$\text{Cov}_{H_0} \{S_n(0), l_{t^*}^n(\theta) - l_{t^*}^n(\theta_0)\} = \frac{a \sqrt{A} \rho(0, t^*)}{\sigma^2} .$$

Par le même genre de raisonnement,

$$\text{Cov}_{H_0} \{S_n(T), l_{t^*}^n(\theta) - l_{t^*}^n(\theta_0)\} = \frac{a \sqrt{A} \rho(t^*, T)}{\sigma^2} .$$

# Convergence faible du processus de score ( $K = 2, t_1 = 0, t_2 = T$ )

On a

$$S_n(t) = \frac{\alpha(t)S_n(0) + \beta(t)S_n(T)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(0, T)}},$$

D'après le théorème de l'image continue

$$S_n(t) \xrightarrow{\mathcal{L}} V(t) \quad \forall t \in [0, T].$$

Cela prouve la convergence finie dimensionnelle.

Tension + convergence finie dimensionnelle  
 $\Rightarrow$  convergence faible

# Convergence faible du processus de score

$(K = 2, t_1 = 0, t_2 = T)$

**Théorème 8.2 de Billingsley (1999)**, le processus de score est tendu ssi :

- 1  $S_n(0)$  est tendue
- 2 Pour tout  $\varepsilon > 0$  et  $\eta > 0$ , il existe  $\delta$ , avec  $0 < \delta < T$ , et un entier  $n_0$  tels que  $\mathbb{P}(w_{S_n}(\delta) \geq \eta) \leq \varepsilon \quad \forall n \geq n_0$

$$\text{où } w_{S_n}(\delta) = \sup_{|t'-t|<\delta} |S_n(t') - S_n(t)|$$

# Convergence faible du processus de score ( $K = 2, t_1 = 0, t_2 = T$ )

On pose

$$\tilde{\alpha}(t) = \alpha(t) / \sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(0, T)},$$

$$\tilde{\beta}(t) = \beta(t) / \sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(0, T)}.$$

On a,

$$\begin{aligned} w_{S_n}(\delta) &= \sup_{|t'-t|<\delta} |S_n(t') - S_n(t)| \\ &= \sup_{|t'-t|<\delta} \left| (\tilde{\alpha}(t') - \tilde{\alpha}(t)) S_n(0) + (\tilde{\beta}(t') - \tilde{\beta}(t)) S_n(T) \right| \\ &\leq \max(|S_n(0)|, |S_n(T)|) \left( w_{\tilde{\alpha}}(\delta) + w_{\tilde{\beta}}(\delta) \right) \end{aligned}$$

# Convergence faible du processus de score ( $K = 2, t_1 = 0, t_2 = T$ )

On montre que

$$\mathbb{P} \left( \max (|S_n(0)|, |S_n(T)|) \left( w_{\tilde{\alpha}}(\delta) + w_{\tilde{\beta}}(\delta) \right) \geq \eta \right) \leq \varepsilon.$$

Par conséquent,

$$\forall n \geq 1 \quad \mathbb{P} (w_{S_n}(\delta) \geq \eta) \leq \varepsilon.$$

# Une interpolation non linéaire

## Théorème (R., Statistics 2013)

$$S_n(\cdot) \Rightarrow V(\cdot) \quad , \quad \sup \Lambda_n(\cdot) \xrightarrow{\mathcal{L}} \sup V^2(\cdot) \quad \text{où}$$

- $V(\cdot)$  est le processus d'interpolation non linéaire tel que

$$\forall t \in [0, T] \quad V(t) = \frac{\alpha(t) V(0) + \beta(t) V(T)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(0, T)}}$$

avec  $\text{Cov}\{V(0), V(T)\} = \rho(0, T)$

- $V(\cdot)$  est un processus Gaussien de variance 1 et de fonction moyenne :

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

$$\text{sous } H_{at^*} : m_{t^*}(0) = \frac{a \sqrt{A}}{\sigma^2} \rho(0, t^*) \quad , \quad m_{t^*}(T) = \frac{a \sqrt{A}}{\sigma^2} \rho(t^*, T)$$

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

# Efficacité $\kappa$ du LRT sur tout le chromosome

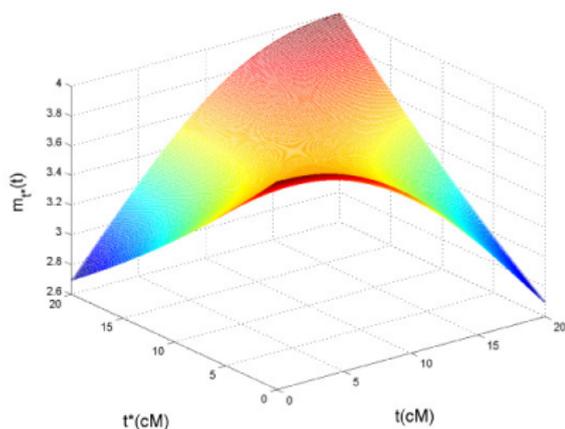
Oracle : pas de selective genotyping (i.e. tous les individus sont génotypés aux marqueurs)

## Lemme

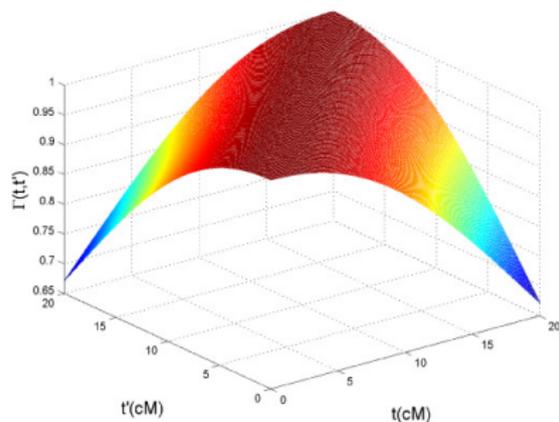
$$\kappa = \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) = \mathcal{A}/\sigma^2$$

$\kappa$  atteint son maximum pour  $\gamma_+ = \gamma_- = \gamma/2$

# Fonction moyenne et fonction covariance ( $a = 4$ , $\sigma = 1$ , $T = 20\text{cM}$ , $\gamma = 1$ , $\sigma = 1$ )

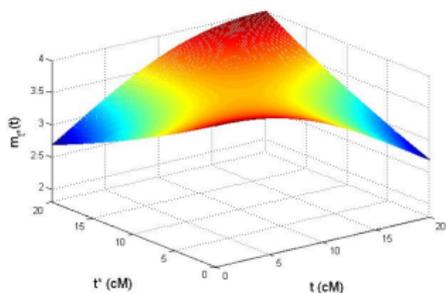


Fonction moyenne

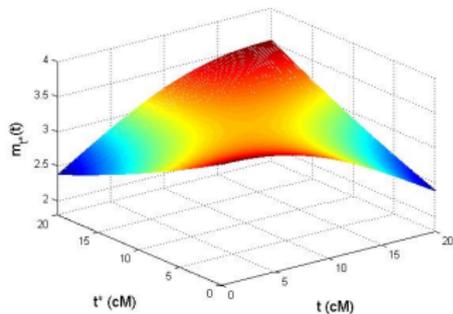


Fonction covariance

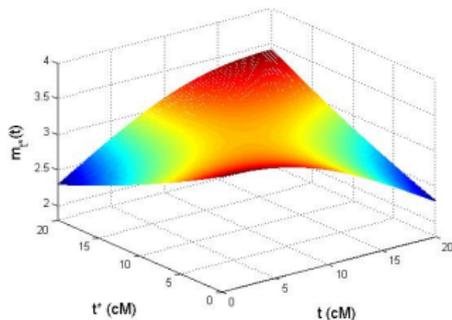
# Fonction moyenne en selective genotyping



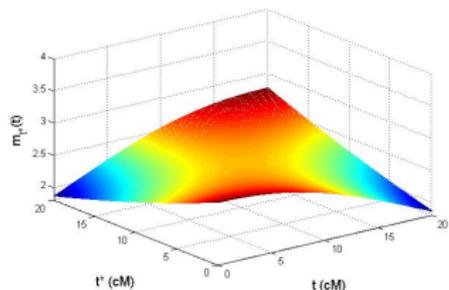
$$\gamma = 1$$



$$\gamma = 0.3, \gamma_+ = \gamma/2$$

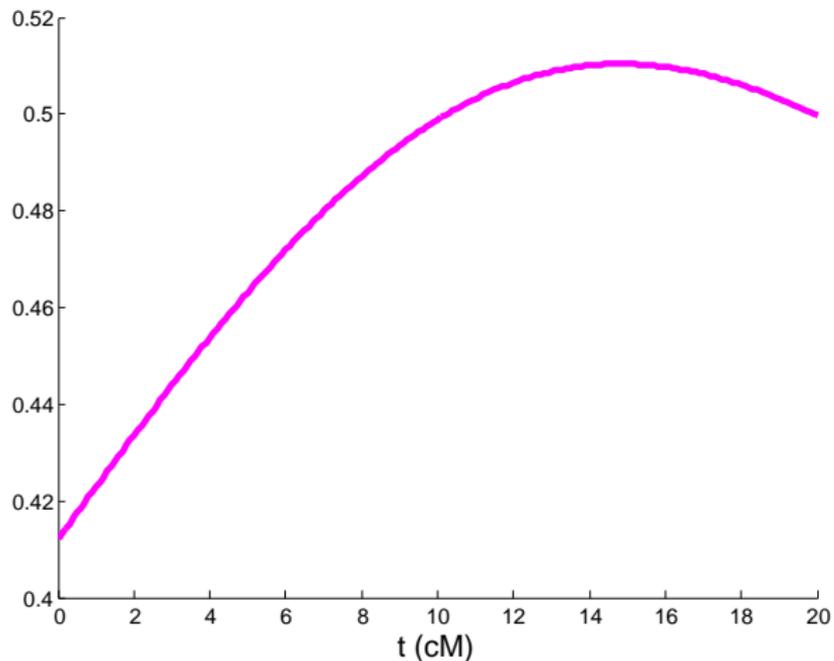


$$\gamma = 0.3, \gamma_+ = 3\gamma/4$$

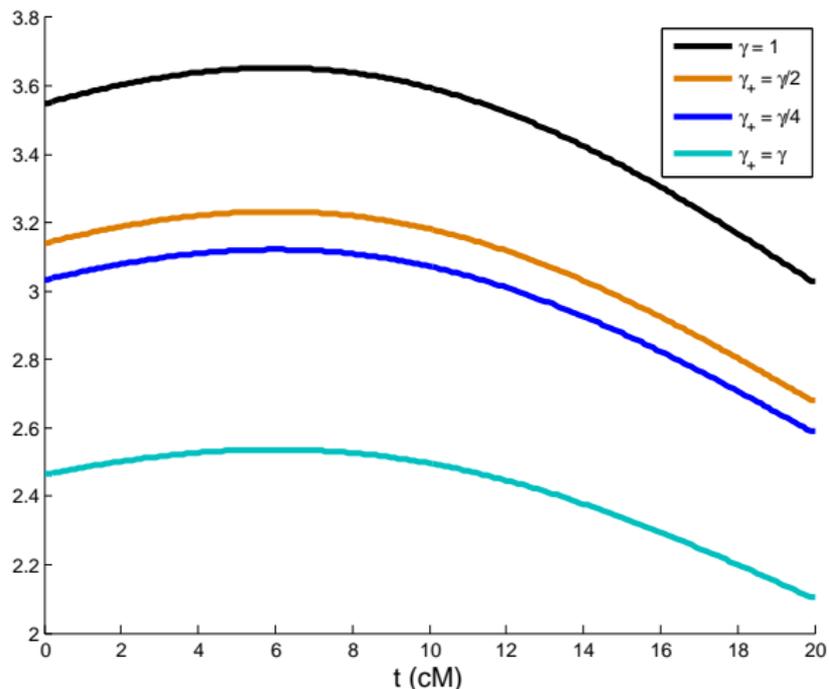


$$\gamma = 0.3, \gamma_+ = \gamma$$

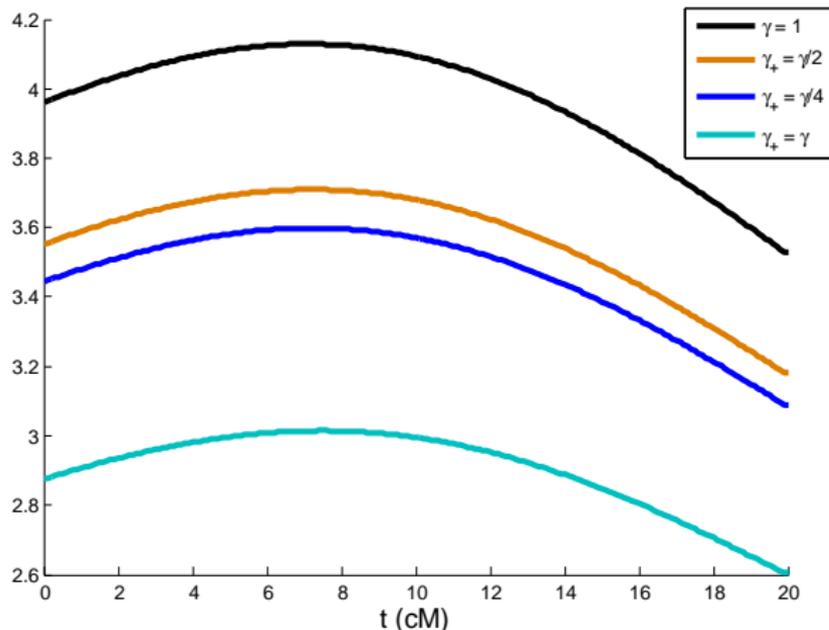
# Une trajectoire du processus $V(\cdot)$ sous $H_0$



# Fonction moyenne du processus $V(\cdot)$ sous $H_{at^*}$ ( $t^* = 6\text{cM}$ , $\gamma = 0.3$ )



# Une trajectoire du processus $V(\cdot)$ sous $H_{at^*}$ ( $t^* = 6\text{cM}$ , $\gamma = 0.3$ )



# Une interpolation non linéaire

## Lemme

Soit  $T_n(\cdot)$  le processus tel que

$$T_n(t) = \frac{\alpha(t)T_n(0) + \beta(t)T_n(T)}{\sqrt{\alpha^2(t) + \beta^2(t) + 2\alpha(t)\beta(t)\rho(0, T)}} , \text{ alors}$$

$$T_n(\cdot) \Rightarrow V(\cdot) \quad \text{et} \quad T_n^2(\cdot) \Rightarrow V^2(\cdot) .$$

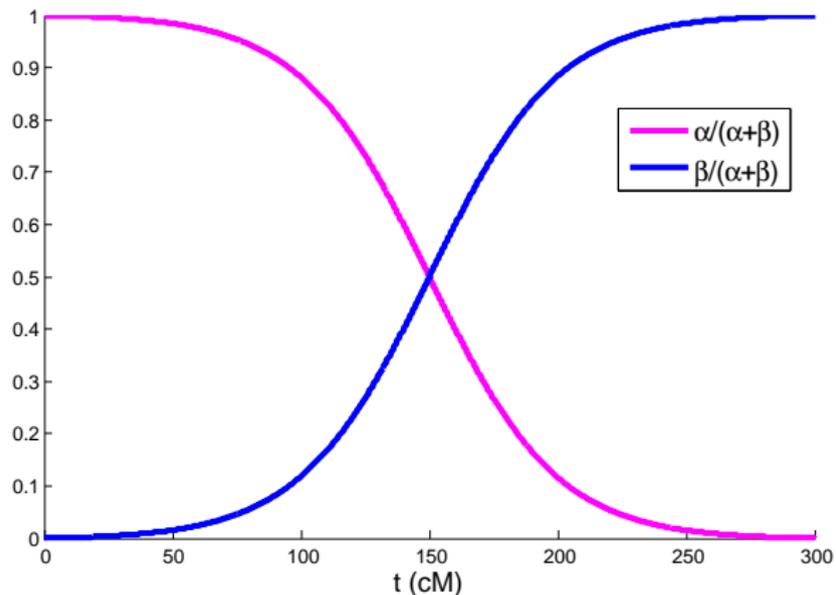
# Un lemme utile pour les processus interpolés

Lemme (Azaïs, Delmas, R., Statistics 2012)

Soient  $\psi_1(t)$  et  $\psi_2(t)$  deux fonctions telles que  $\frac{\psi_i(t)}{\psi_1(t)+\psi_2(t)}$  prend toutes les valeurs de  $[0, 1]$ ,  $i = 1, 2$ . Soient  $C_1$  et  $C_2$  deux nombres réels et  $0 < \tilde{\rho} < 1$  alors

$$\begin{aligned} & \max_{t \in [0, T]} \frac{\{\psi_1(t)C_1 + \psi_2(t)C_2\}^2}{\psi_1^2(t) + \psi_2^2(t) + 2\tilde{\rho}\psi_1(t)\psi_2(t)} \\ & = \max \left( C_1^2, C_2^2, \frac{C_1^2 + C_2^2 - 2\tilde{\rho}C_1C_2}{1 - \tilde{\rho}^2} \mathbb{1}_{\frac{C_2}{C_1} \in ]\tilde{\rho}, \frac{1}{\tilde{\rho}}[} \right). \end{aligned}$$

# Un lemme utile pour les processus interpolés



# Un lemme utile pour les processus interpolés

Le lemme s'applique en prenant

- $\tilde{\rho} = \rho(0, T)$
- $C_1 = T_n(0), C_2 = T_n(T)$
- $\psi_1(t) = \alpha(t), \psi_2(t) = \beta(t)$

On a donc

$$\sup_{[0, T]} T_n^2(t) = \max \left\{ T_n^2(0), T_n^2(T), h_n(0, T) \right\}$$

où

$$h_n(0, T) = \frac{T_n^2(0) + T_n^2(T) - 2\rho(0, T)T_n(0)T_n(T)}{1 - \rho^2(0, T)} \mathbf{1}_{\frac{T_n(T)}{T_n(0)} \in ]\rho(0, T), \frac{1}{\rho(0, T)}[}$$

Inutile d'effectuer des tests partout sur le chromosome !

# Application au calcul de valeurs critiques

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

⇒ fonction QSIMVNEF de Genz (1992)

	K	101
	$c$	9.74
Rebaï	$n = 200$	2.55%
	$n = 100$	2.52%
	$n = 50$	2.01%
	$c$	8.45
Feingold	$n = 200$	4.67%
	$n = 100$	4.72%
	$n = 50$	3.92%
	$c$	8.41
Nous	$n = 200$	4.76%
	$n = 100$	4.80%
	$n = 50$	3.97%

$T = 1M$ , marqueurs équidistants,  $\gamma = 1$

# Un exemple avec un maximum de 657 tests sur le génome

- $T = 10M, K = 329, \gamma = 1$
- $\forall k = 1, \dots, 301 \quad t_k = 0.01(k - 1)$
- $\forall k = 302, \dots, 329 \quad t_k = 3.25 + 0.25(k - 302)$

Feingold	$c$	12.55
	$n = 200$	2.85%
	$n = 100$	2.72%
	$n = 50$	2.02%
Nous	$c$	11.70
	$n = 200$	4.64%
	$n = 100$	4.20%
	$n = 50$	3.39%

# Modélisation du phénomène d'interférence (Rebaï et al. 95, 94)

Une seule recombinaison autorisée  
dans chaque intervalle de marqueurs

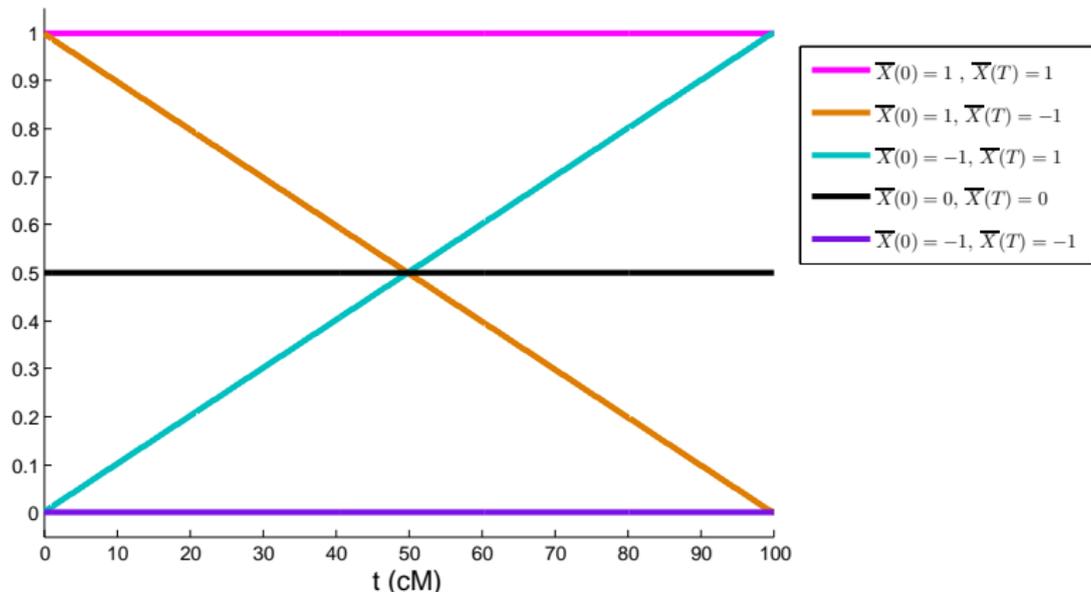
En  $t = t^*$

$$L_{t^*}(\theta) = \left[ \tilde{p}(t^*) f_{(\mu+q,\sigma)}(Y) 1_{Y \notin [S_-, S_+]} + \{1 - \tilde{p}(t^*)\} f_{(\mu-q,\sigma)}(Y) 1_{Y \notin [S_-, S_+]} + \frac{1}{2} f_{(\mu+q,\sigma)}(Y) 1_{Y \in [S_-, S_+]} + \frac{1}{2} f_{(\mu-q,\sigma)}(Y) 1_{Y \in [S_-, S_+]} \right] g(\cdot)$$

où

$$\begin{aligned} \tilde{p}(t^*) &= 1_{\bar{X}(t^{*\ell})=1} 1_{\bar{X}(t^{*r})=1} + \frac{t^{*r} - t^*}{t^{*r} - t^{*\ell}} 1_{\bar{X}(t^{*\ell})=1} 1_{\bar{X}(t^{*r})=-1} \\ &+ \frac{t^* - t^{*\ell}}{t^{*r} - t^{*\ell}} 1_{\bar{X}(t^{*\ell})=-1} 1_{\bar{X}(t^{*r})=1} \end{aligned}$$

# Illustration des différents poids ( $K = 2$ , $T = 1M$ )



# Une interpolation linéaire

## Théorème (R., JSPI 2014)

$$S_n(\cdot) \Rightarrow D(\cdot) \quad , \quad \sup \Lambda_n(\cdot) \xrightarrow{\mathcal{L}} \sup D^2(\cdot) \quad \text{où}$$

- $D(\cdot)$  est le processus d'interpolation linéaire tel que

$$\forall t \in [0, T] \quad D(t) = \frac{\tilde{\alpha}(t) D(0) + \tilde{\beta}(t) D(T)}{\sqrt{\tilde{\alpha}^2(t) + \tilde{\beta}^2(t) + 2\tilde{\alpha}(t)\tilde{\beta}(t)\rho(0, T)}}$$

$$\text{où } \tilde{\alpha}(t) = \frac{T-t}{T-0} \quad , \quad \tilde{\beta}(t) = \frac{t-0}{T-0} \quad \text{et} \quad \text{Cov}\{D(0), D(T)\} = \rho(0, T)$$

- $D(\cdot)$  est un processus Gaussien de variance 1 et de fonction moyenne

$$\text{sous } H_{at^*} : m_{t^*}(0) = \frac{a\sqrt{A}}{\sigma^2} \left\{ \tilde{\alpha}(t^*) + \tilde{\beta}(t^*)\rho(0, T) \right\}$$

$$m_{t^*}(T) = \frac{a\sqrt{A}}{\sigma^2} \left\{ \tilde{\alpha}(t^*)\rho(0, T) + \tilde{\beta}(t^*) \right\}$$

$$\forall t \in [0, T] \quad m_{t^*}(t) = \frac{\tilde{\alpha}(t) m_{t^*}(0) + \tilde{\beta}(t) m_{t^*}(T)}{\sqrt{\tilde{\alpha}^2(t) + \tilde{\beta}^2(t) + 2\tilde{\alpha}(t)\tilde{\beta}(t)\rho(0, T)}}$$

# Interpolation linéaire/Interpolation non linéaire

## Lemme

*Sous l'hypothèse nulle  $H_0$ ,*

$$\max_{t \in [0, T]} D^2(t) = \max_{t \in [0, T]} V^2(t) ,$$

*où  $V(\cdot)$  est le processus d'interpolation non linéaire, obtenu sous le modèle de Haldane.*

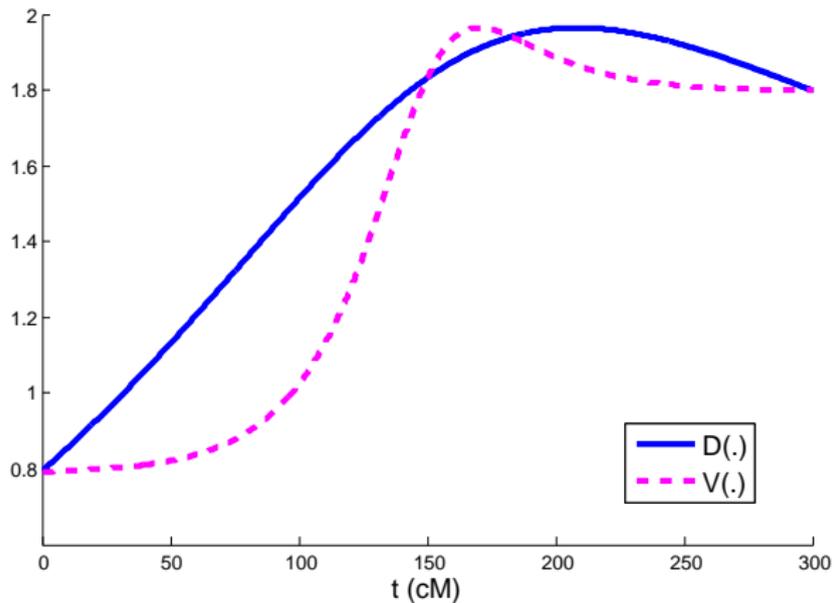
**Le seuil est donc le même**

**sous Haldane et sous le modèle d'interférence !**

Attention, le lemme n'est plus valable sous l'alternative  $H_{at^*}$

**Robustesse asymptotique du LRT (R., EJS 2014)**

# Une trajectoire des processus $D(\cdot)$ et $V(\cdot)$ sous $H_0$



# Génotypage des non extrêmes (expérience inverse du selective genotyping)

$\mathcal{B}$  est la quantité suivante

$$\mathcal{B} = \sigma^2 \{1 - \gamma - z_{\gamma+} \varphi(z_{\gamma+}) + z_{1-\gamma-} \varphi(z_{1-\gamma-})\}$$

## Théorème

$$S_n(\cdot) \Rightarrow U(\cdot) \quad , \quad \sup \Lambda_n(\cdot) \xrightarrow{\mathcal{L}} \sup U^2(\cdot) \quad \text{où}$$

- $U(\cdot)$  est le processus d'interpolation non linéaire
- $U(\cdot)$  est un processus Gaussien de variance 1 et de fonction moyenne :

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

$$\text{sous } H_{at^*} : m_{t^*}(0) = \frac{a \sqrt{\mathcal{B}}}{\sigma^2} \rho(0, t^*) \quad , \quad m_{t^*}(T) = \frac{a \sqrt{\mathcal{B}}}{\sigma^2} \rho(t^*, T)$$

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

# Efficacité du LRT sur tout le chromosome (expérience inverse du selective genotyping)

Oracle : tous les individus sont génotypés aux marqueurs

## Théorème

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

$\kappa$  atteint son maximum pour  $\gamma_+ = \gamma$  ou  $\gamma_- = \gamma$

L'information est bien présente dans les extrêmes !

# Si plusieurs QTLs sont présents sur le chromosome

$H_{at^*}$  : "il existe  $m$  QTLs situés en  $t_1^*$ , ...,  $t_m^*$  avec effets  $q_1 = a_1/\sqrt{n}$ , ...,  $q_m = a_m/\sqrt{n}$ "

## Théorème

$$S_n(\cdot) \Rightarrow Z(\cdot) \quad , \quad \sup \Lambda_n(\cdot) \xrightarrow{\mathcal{L}} \sup Z^2(\cdot) \quad \text{où}$$

- $Z(\cdot)$  est le processus d'interpolation non linéaire tel que

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

$$\text{avec } \text{Cov}\{Z(0), Z(T)\} = \rho(0, T)$$

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

$$\text{sous } H_{at^*} : m_{t^*}^{-}(0) = \sum_{s=1}^m \frac{a_s \sqrt{A}}{\sigma^2} \rho(0, t_s^*) \quad , \quad m_{t^*}^{-}(T) = \sum_{s=1}^m \frac{a_s \sqrt{A}}{\sigma^2} \rho(t_s^*, T)$$

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

# Conclusions

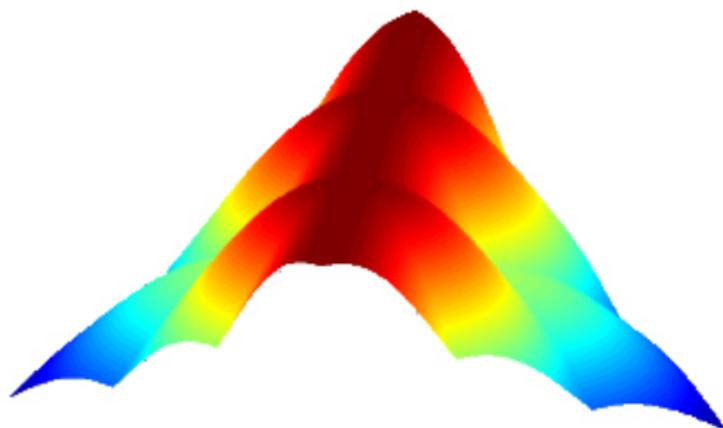
## Selective Genotyping sur un marqueur :

- On doit génotyper le même pourcentage d'individus aux deux extrêmes
- Les phénotypes non extrêmes n'apportent pas d'information
- La comparaison de moyenne est optimale

## Génome Scan + Selective Genotyping :

- Interpolation non linéaire (Haldane) / linéaire (interférence)
- Plus de puissance en interférence que sous Haldane
- Robustesse asymptotique du LRT
- Un seul test au maximum entre deux marqueurs
- Seuil identique avec/sans interférence et avec/sans selective genotyping
- Statistique de test simple équivalente au LRT
- Si Carte dense et  $I$  familles :  
⇒ Processus de Chi deux d'Ornstein-Uhlenbeck à  $I$  degrés de liberté

Merci de votre attention







# Puissance théorique/Puissance empirique sous l'alternative $H_{at^*}$

$\gamma$ \ $t^*$	12cM	36cM	52cM	75cM
$\gamma$	61.37% (60.74%)	62.03% (61.36%)	62.82% (62.23%)	61.68% (61.10%)
$\gamma/2$	83.83% (83.54%)	83.70% (83.15%)	84.61% (84.49%)	83.28% (83.79%)
$\gamma/4$	80.97% (80.95%)	80.86% (80.18%)	81.85% (81.10%)	80.67% (80.61%)
$\gamma/8$	76.15% (75.70%)	75.98% (75.36%)	76.75% (76.75%)	75.63% (75.14%)

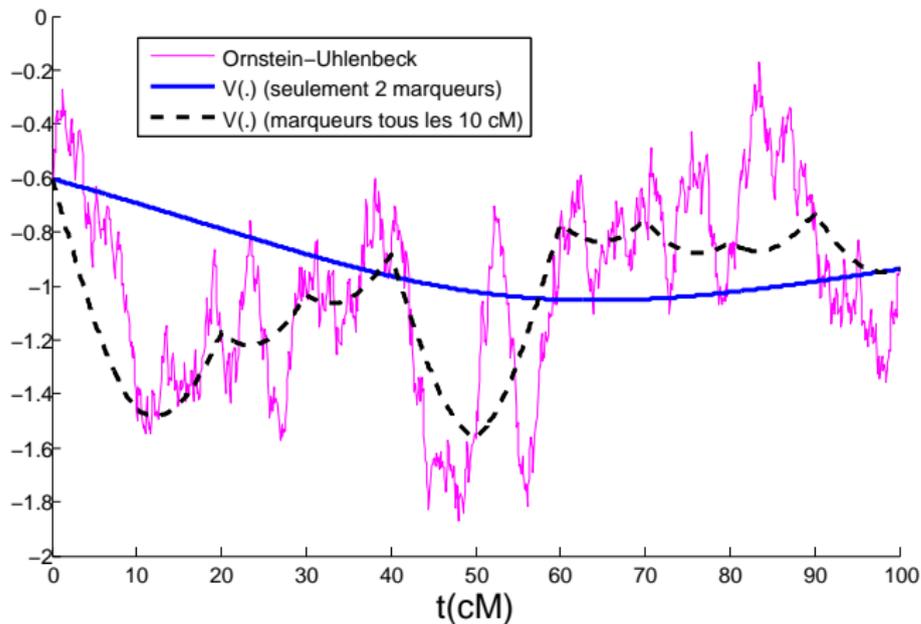
$T = 1M$ ,  $K = 11$ , marqueurs équidistants tous les 10cM,  
 $\gamma = 0.3$ ,  $a = 4$ , 10000 échantillons de taille  $n = 1000$ , 100000  
trajectoires pour la puissance théorique

# Puissance théorique (expérience inverse du selective genotyping)

$t^*$ \ $\gamma_+$	12cM	35cM	48cM	77cM
$\gamma/4$	83.99%	86.39%	84.91%	87.87%
$\gamma/2$	80.14%	82.75%	80.96%	84.00%
$\gamma$	95.36%	96.23%	95.48%	96.85%

$T = 1M$ ,  $K = 6$ , marqueurs équidistants tous les 20cM,  
 $\gamma = 0.2$ ,  $1 - \gamma = 0.8$ ,  $a = 6$ , 100000 trajectoires

# L'Interval Mapping lisse les trajectoires



# Conclusions sur le Selective Genotyping

- On doit génotyper le même pourcentage d'individus aux deux extrêmes
- Les phénotypes non extrêmes n'apportent pas d'information
- Le génotypage optimal est d'environ 30% (dépend du ratio coût génotypage/phénotypage)
- Le test de comparaison de moyenne est optimal

	$n = 50$		$n = 100$	
Nombre de QTLs	$W_1$	$T_2$	$W_1$	$T_2$
1	0.0020	0.0005	0.0041	0.0005
1000	2.7871	0.1267	5.1131	0.1384

Comparaison en temps de calcul

( $q = 0.3$ ,  $p = 1/2$ ,  $\gamma = 0.3$ ,  $\gamma_+/\gamma = 1/2$ )

# Pourcentage de faux positifs sous l'hypothèse nulle $H_0$

$\gamma_+$ \ $n$	1000	200	100	50
$\gamma$	5.05%	4.58%	4.20%	3.47%
$\gamma/2$	4.94%	4.73%	4.59%	4.21%
$\gamma/4$	4.82%	4.56%	4.65%	3.83%
$\gamma/8$	5.02%	4.87%	4.31%	3.40%

$T = 1M$ ,  $K = 11$ , marqueurs équidistants tous les 10cM,  
 $\gamma = 0.3$ ,  $a = 0$ , 10000 échantillons de taille  $n$

# Puissance théorique en fonction du modèle

$\gamma_+$	$t^*$	interference	Haldane
$\gamma$	0.80	45.08%	35.93%
	1.30	45.26%	36.07%
	2.05	54.66%	50.89%
	2.65	46.34%	38.53%
$\gamma/2$	0.80	72.12%	58.99%
	1.30	72.05%	60.22%
	2.05	82.13%	78.68%
	2.65	74.27%	64.56%
$\gamma/4$	0.80	68.06%	56.66%
	1.30	68.08%	56.82%
	2.05	78.79%	75.24%
	2.65	70.49%	60.52%

$\gamma = 0.4$ ,  $T = 3M$ ,  $K = 7$ , marqueurs équidistants tous les 50cM,  $a = 4$ , 100000 trajectoires

# Application au calcul de valeurs critiques

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

⇒ fonction QSIMVNEF de Genz (1992)

	K	101	51	26	6
Rebaï	$c$	9.74	9.09	8.43	6.92
	$n = 200$	2.55%	3.23%	3.82%	4.53%
	$n = 100$	2.52%	2.90%	3.59%	4.52%
	$n = 50$	2.01%	2.51%	3.53%	4.51%
Feingold	$c$	8.45	8.17	7.81	6.59
	$n = 200$	4.67%	4.91%	5.37%	5.25%
	$n = 100$	4.72%	4.67%	5.07%	5.38%
	$n = 50$	3.92%	4.35%	5.00%	5.43%
Nous	$c$	8.41	8.27	7.91	6.76
	$n = 200$	4.76%	4.71%	5.17%	4.76%
	$n = 100$	4.80%	4.40%	4.78%	4.95%
	$n = 50$	3.97%	4.15%	4.75%	4.88%

$T = 1M$ , marqueurs équidistants,  $\gamma = 1$

# Efficacité du test de Wald ( $\mu, q, \sigma$ )

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

A la fois sous  $H_0$  et sous  $H_a$ ,  $\gamma$  correspond

asymptotiquement au pourcentage d'individus génotypés

De la même manière, on note :

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

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

Efficacité du test de Wald :

$$\begin{aligned} \forall p, \kappa_1 &= \gamma + z_{\gamma_+} \varphi(z_{\gamma_+}) - z_{1-\gamma_-} \varphi(z_{1-\gamma_-}) \\ &= \mathcal{A}/\sigma^2 \end{aligned}$$

# Etude du processus de score sous $H_{at^*}$

$(K = 2, t_1 = 0, t_2 = T)$

On a

$$\begin{aligned} & \text{Cov}_{H_0} \left\{ S_n(0), \frac{a \alpha(t^*)}{\sigma \sqrt{n}} \sum_{j=1}^n \varepsilon_j \bar{X}_j(0) \right\} \\ &= \text{Cov}_{H_0} \left\{ \sum_{j=1}^n \frac{\sigma \varepsilon_j \bar{X}_j(0)}{\sqrt{n \mathcal{A}}}, \frac{a \alpha(t^*)}{\sigma \sqrt{n}} \sum_{j=1}^n \varepsilon_j \bar{X}_j(0) \right\} \\ &= \frac{a \alpha(t^*) \sqrt{\mathcal{A}}}{\sigma^2}. \end{aligned}$$

De la même façon,

$$\text{Cov}_{H_0} \left\{ S_n(0), \frac{a \beta(t^*)}{\sigma \sqrt{n}} \sum_{j=1}^n \varepsilon_j \bar{X}_j(T) \right\} = \frac{a \beta(t^*) \sqrt{\mathcal{A}} \rho(0, T)}{\sigma^2}.$$

# Modélisation du phénomène d'interférence (Rebaï et al. 95, 94)

## Cas $K = 2$ (i.e. $t_1 = 0, t_2 = T$ )

- Conservation du modèle de Haldane sur les marqueurs  
 $r(0, T) = \mathbb{P} \{X(0)X(T) = -1\} = (1 - \rho(0, T))/2$
- $U(t^*)$  : information génome au QTL
- $r_0(t^*)$  : probabilité de recombinaison entre le premier marqueur et le QTL

$$r_0(t^*) = \mathbb{P} \{X(0)U(t^*) = -1\}$$

- $r_T(t^*)$  : probabilité de recombinaison entre le deuxième marqueur et le QTL

$$r_T(t^*) = \mathbb{P} \{X(T)U(t^*) = -1\}$$

# Modélisation du phénomène d'interférence (Rebaï et al. 95, 94)

Une seule recombinaison autorisée entre le QTL  
et les deux marqueurs

$$\{X(0)X(T) = -1\} \Leftrightarrow \{X(0)U(t^*) = -1\} \cup \{X(T)U(t^*) = -1\}$$

Comme  $\{X(0)U(t^*) = -1\} \cap \{X(T)U(t^*) = -1\} = \emptyset$ , on a

$$r(0, T) = r_0(t^*) + r_T(t^*).$$

- Proportionnalité

$$r_0(t^*) = \frac{t^* - 0}{T - 0} r(0, T), \quad r_T(t^*) = \frac{T - t^*}{T - 0} r(0, T).$$

- Modèle "d'analyse de variance"

$$\tilde{Y} = \mu + U(t^*) q + \sigma \varepsilon \quad \text{où } \varepsilon \sim N(0, 1)$$

# Selective genotyping + interférence

## Configuration classique :

- On observe  $(\tilde{Y}, X(0), X(T))$

## Configuration selective genotyping :

- On observe  $(\tilde{Y}, \tilde{X}(0), \tilde{X}(T))$ , où
  - $\tilde{X}(0) = X(0) 1_{\tilde{Y} \notin [s_-, s_+]}$
  - $\tilde{X}(T) = X(T) 1_{\tilde{Y} \notin [s_-, s_+]}$

# Un arbre phylogénétique chez la levure

