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

Charles-Elie Rabier

23 Avril 2015

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 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

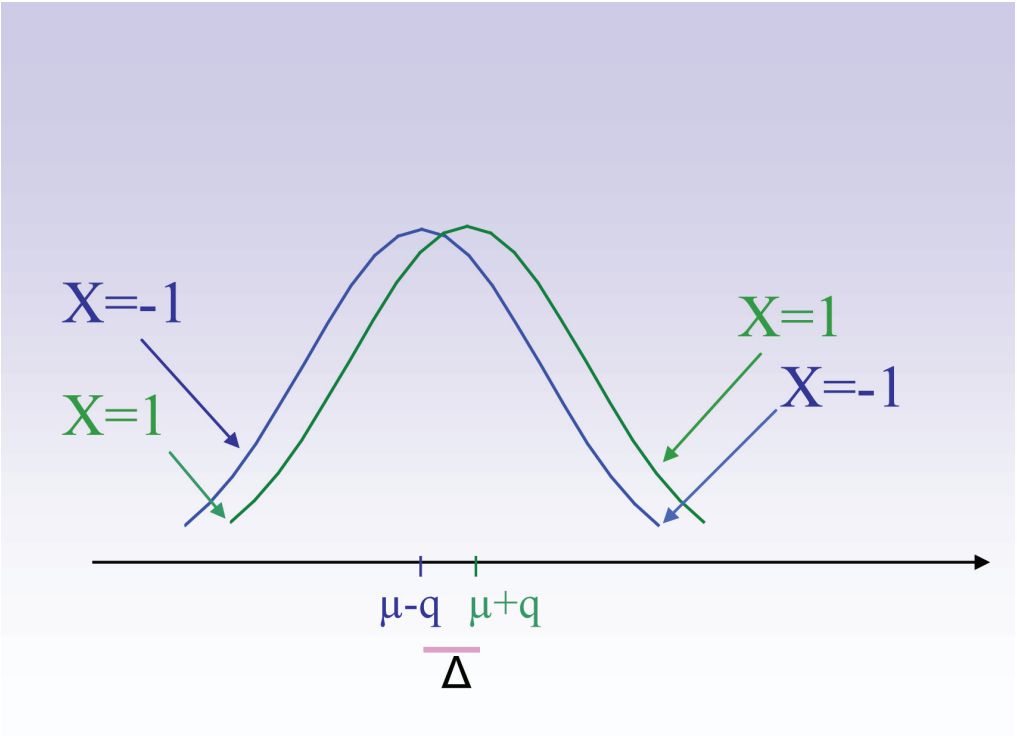
$$X = \begin{cases} -1 & \text{avec probabilité } 1-p \\ 1 & \text{avec probabilité } p \end{cases}$$

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

- Y : variable aléatoire correspondant au phénotype

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

Modèle en l'absence de censure



Distribution des phénotypes Y

Test statistique oracle (μ, q, σ)

- A l'aide de n observations (X_j, Y_j) i.i.d., 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 coûtait 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)
Darvasi et Soller (Theoretical and Applied Genetics, 1992)

Gène de l'obésité chez le porc (Fontanesi et al. 2012)
Géne pour la longueur de la coquille d'un coquillage Mérétrix (Lu et al. 2013)

Modèle correspondant au Selective Genotyping

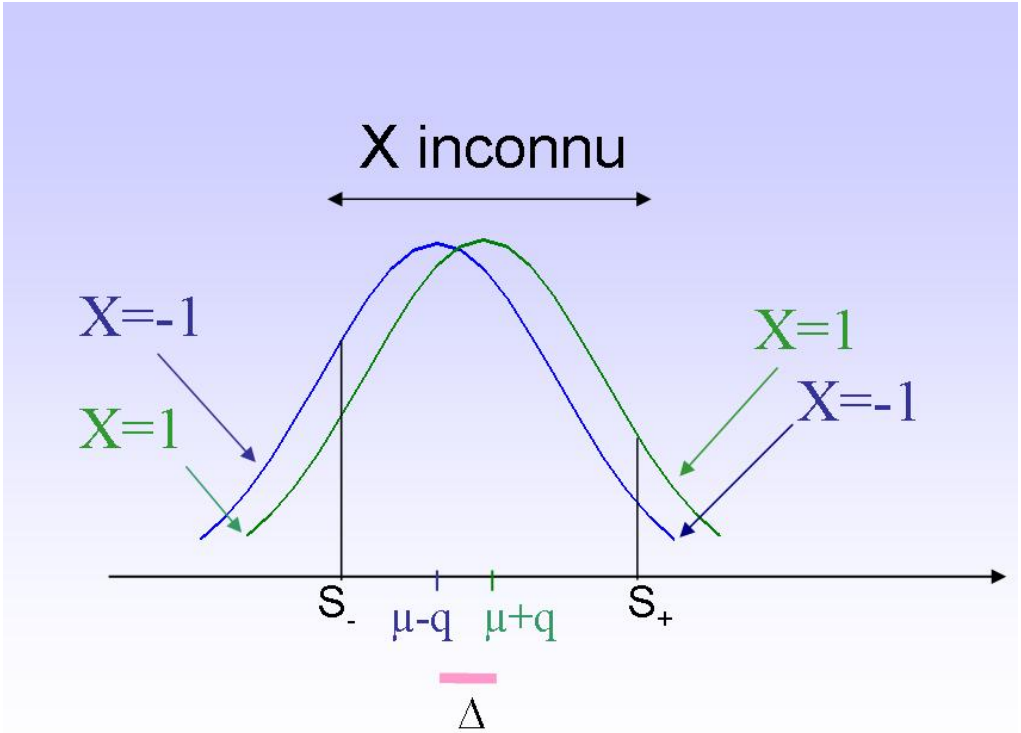
X disponible uniquement pour les individus présentant un phénotype extrême Y

$\Rightarrow$ 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



Distribution des phénotypes Y

Test de Wald (μ, q, σ)

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

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

- Statistique de Wald

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

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

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

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

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 $\gamma_+ = \gamma_- = \gamma/2$

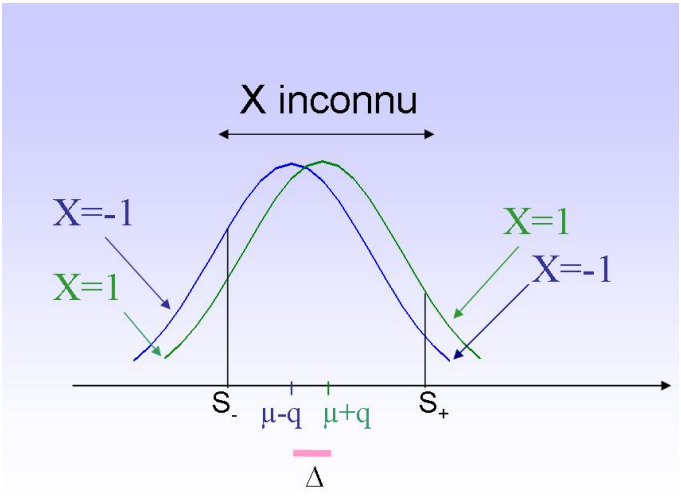
$$M = \gamma + 2\, z_{\gamma/2}\, \varphi(z_{\gamma/2})$$

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

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 (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

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

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

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

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
- $X(.)$: génome d'un individu
- On considère le modèle Poissonien de Haldane

Modélisation de Haldane

- Pas d'interférence lors des crossing overs
- $N(.)$: processus de Poisson standard sur $[0, T]$ représentant le nombre de crossings overs

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

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

- $r(t, t')$: probabilité de recombinaison entre deux loci

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

Modélisation de Haldane

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

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

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

$$t_1 = 0 < t_2 < ... < t_K = T$$

Configuration classique (oracle)

Une observation est le couple

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

et le challenge est dans le fait que $t^\star$ 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 $t^\star$ 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

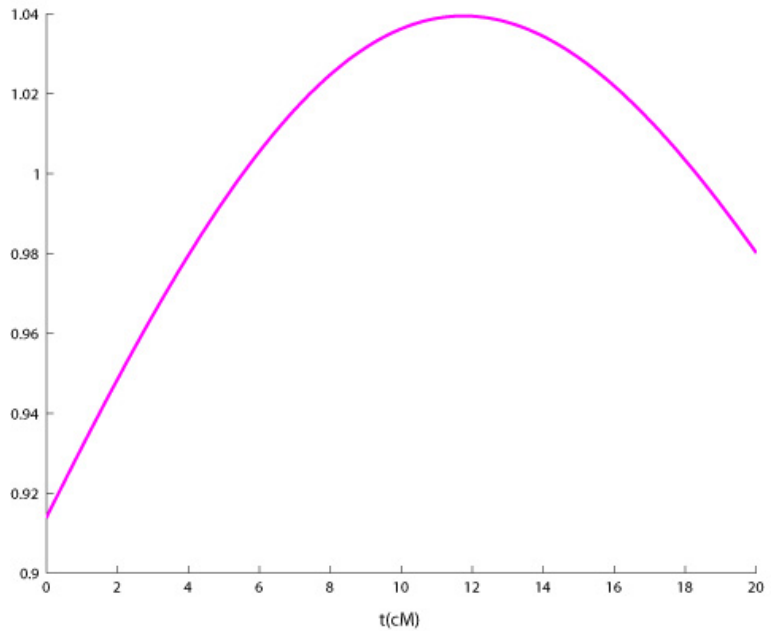
$\Rightarrow$ modèle de mélange de mélange

L'Interval Mapping de Lander et Botstein (1989)

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

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

$\Rightarrow$ LRT sur tout le chromosome : $\sup \Lambda_n(.)$



Une trajectoire du processus $\Lambda_n(.)$ ($T = 20$cM, $K = 2$)

Configuration selective genotyping

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

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

alors, dans notre cas, une observation sera le couple

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

Vraisemblance du couple $\left(Y, \overline{X}(t_1), \ldots, \overline{X}(t_K)\right)$

En $t = t^\star$

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

où

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

et

$$p(t^\star) = Q_{t^\star}^{1,1}\, 1_{\overline{X}(t^{\star\ell})=1} 1_{\overline{X}(t^{\star r})=1} \;+\; Q_{t^\star}^{1,-1}\, 1_{\overline{X}(t^{\star\ell})=1} 1_{\overline{X}(t^{\star r})=-1} + Q_{t^\star}^{-1,1}\, 1_{\overline{X}(t^{\star\ell})=-1} 1_{\overline{X}(t^{\star r})=1} \;+\; Q_{t^\star}^{-1,-1}\, 1_{\overline{X}(t^{\star\ell})=-1} 1_{\overline{X}(t^{\star r})=-1}$$

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

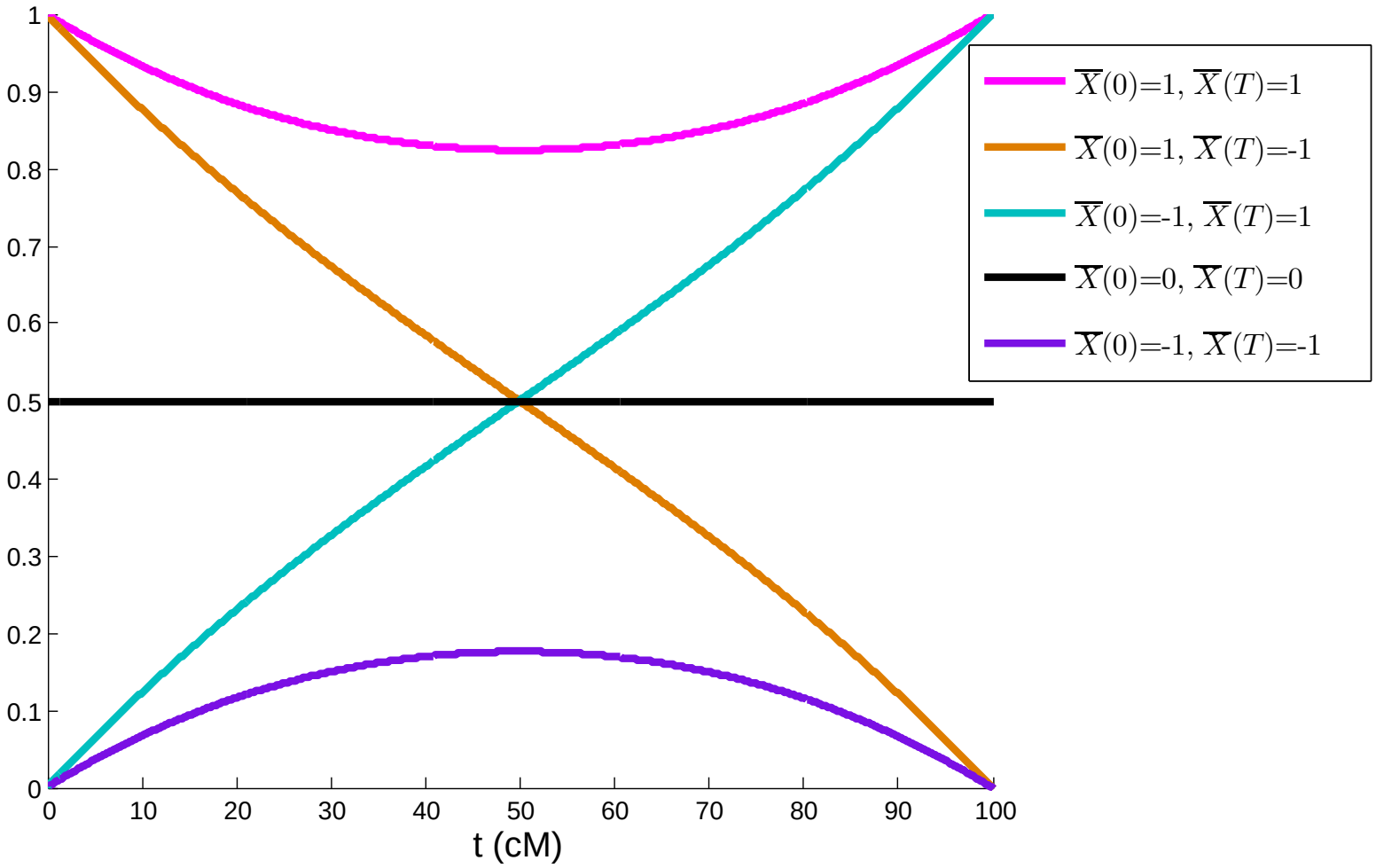
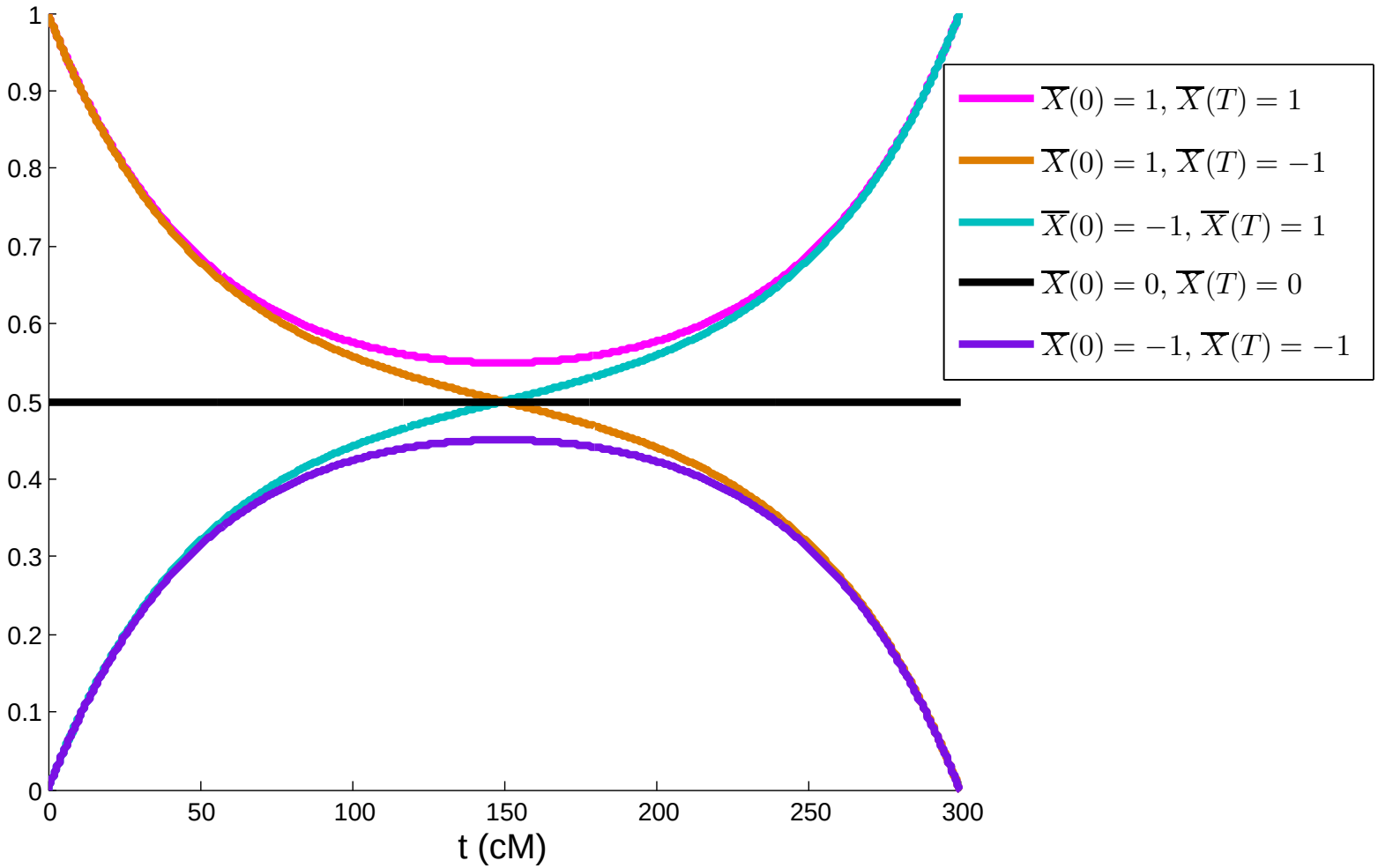


Illustration des différents poids ($K = 2$, $T = 3$M)



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

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

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

- $t^\star$ connu $\Rightarrow$ modèle régulier
- $t^\star$ 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 $[0, T]$”

$H_{at^\star}$: “le QTL est situé en $t^\star \in [0, T]$ 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 :

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

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

$$\frac{\partial l_t^n}{\partial q} |_{\theta_0} = \sum_{j=1}^{n} \frac{Y_j - \mu}{\sigma^2} \{2p_j(t) - 1\} 1_{Y_j \notin [s_-, s_+]}$$

$$= \frac{\alpha(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j \overline{X}_j(0) + \frac{\beta(t)}{\sigma} \sum_{j=1}^{n} \varepsilon_j \overline{X}_j(T)$$

Le processus limite est un processus interpolé ! ! !

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

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

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

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

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

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

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

$$\mathrm{Cov}_{H_0}\left\{S_n(0), I^n_{t^\star}(\theta) - I^n_{t^\star}(\theta_0)\right\} = \frac{a\sqrt{\mathcal{A}}\,\rho(0, t^\star)}{\sigma^2}\,.$$

Par le même genre de raisonnement,

$$\mathrm{Cov}_{H_0}\left\{S_n(T), I^n_{t^\star}(\theta) - I^n_{t^\star}(\theta_0)\right\} = \frac{a\sqrt{\mathcal{A}}\,\rho(t^\star, T)}{\sigma^2}\,.$$

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 δ, avec $0 < \delta < T$, et un entier n_0 tels que $\mathbb{P}\left(w_{S_n}(\delta) \geq \eta\right) \leq \varepsilon \quad \forall n \geq n_0$

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

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,

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

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

On montre que

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

Par conséquent,

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

Une interpolation non linéaire ($K = 2$, $t_1 = 0$, $t_2 = T$)

Théorème (R., Statistics 2013)

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

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

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

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

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

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

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

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

Efficacité κ du LRT sur tout le chromosome

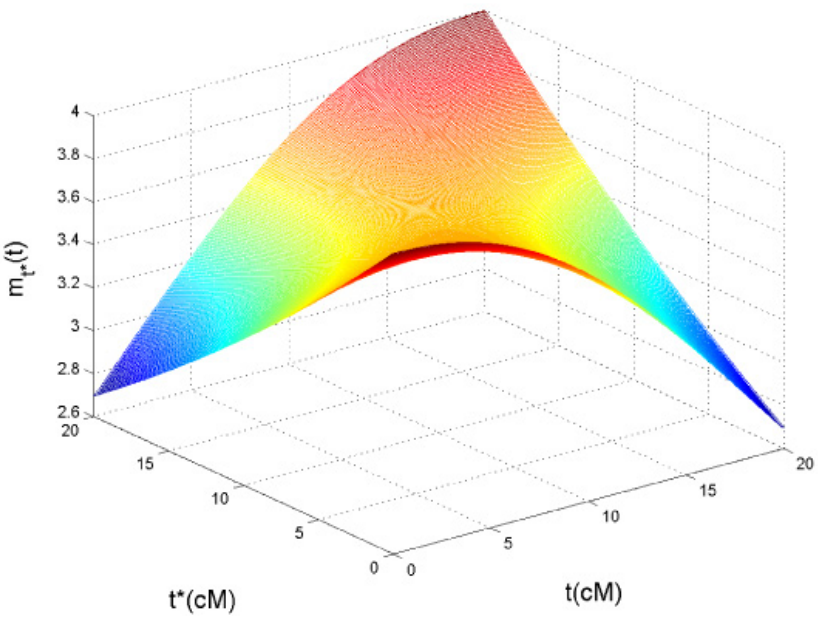
Oracle : pas de selective genotyping (i.e. tous les individus sont génotypés aux marqueurs)

Lemme

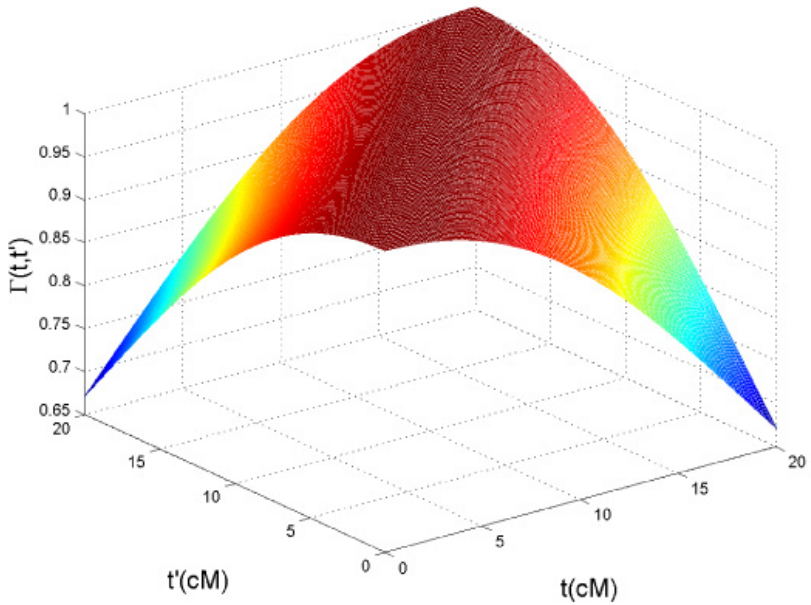
$$\kappa = \gamma \; + \; z_{\gamma_+}\varphi(z_{\gamma_+}) \; - \; z_{1-\gamma_-} \; \varphi(z_{1-\gamma_-}) = \mathcal{A}/\sigma^2$$

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

Fonction moyenne et fonction covariance ($a = 4$, $\sigma = 1$, $K = 2$, $T = 20$cM, $\gamma = 1$)

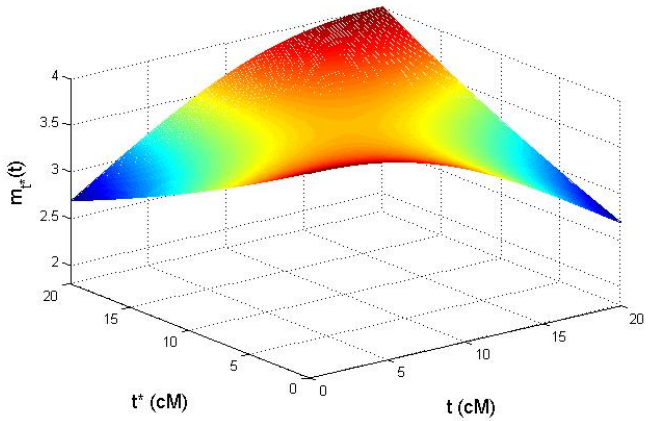


Fonction moyenne

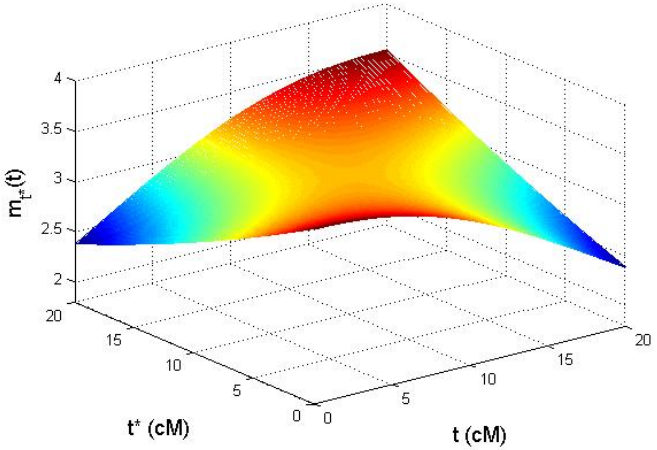


Fonction covariance

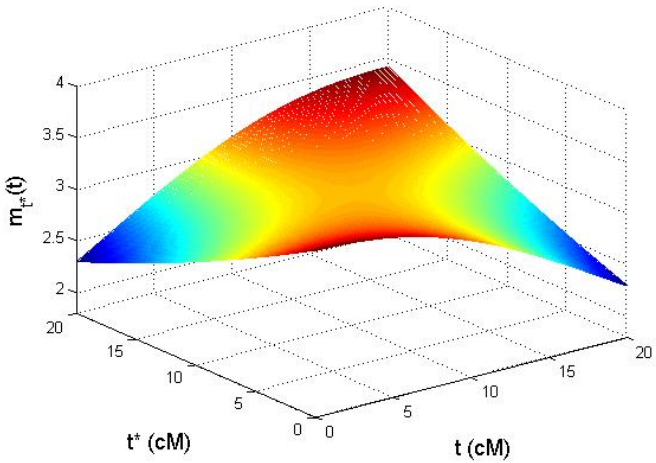
Fonction moyenne en selective genotyping ($K = 2$)



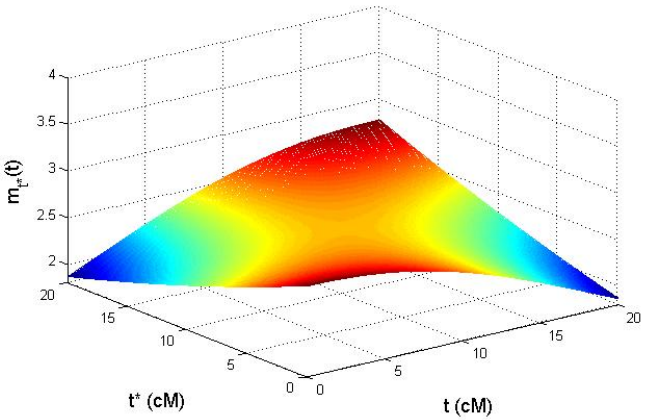
$\gamma = 1$



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



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



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

K marqueurs en $t_1 = 0 < t_2 < \ldots < t_K = T$

Le cas $t^\star \notin [t_k, t_{k+1}]$

$$I^n_{t^\star}(\theta) \; - \; I^n_{t^\star}(\theta_0)$$

$$= \frac{a}{\sigma\sqrt{n}} \left\{ \alpha(t^\star) \sum_{j=1}^{n} \varepsilon_j \, \overline{X}_j(t^{\star\ell}) + \beta(t^\star) \sum_{j=1}^{n} \varepsilon_j \, \overline{X}_j(t^{\star r}) \right\}$$

$$- \frac{a^2}{2\sigma^4} \mathcal{A} \left\{ \alpha^2(t^\star) + \beta^2(t^\star) + 2\alpha(t^\star)\beta(t^\star)\rho(t^{\star\ell}, t^{\star r}) \right\} + o_P(1)$$

$$\mathrm{Cov}_{H_0} \left\{ S_n(t_k), \frac{a\,\alpha(t^\star)}{\sigma\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \, \overline{X}_j(t^{\star\ell}) \right\} = \frac{a\,\alpha(t^\star)\,\sqrt{\mathcal{A}}\,\rho(t_k, t^{\star\ell})}{\sigma^2}$$

$$\mathrm{Cov}_{H_0} \left\{ S_n(t_k), \frac{a\,\beta(t^\star)}{\sigma\sqrt{n}} \sum_{j=1}^{n} \varepsilon_j \, \overline{X}_j(t^{\star r}) \right\} = \frac{a\,\beta(t^\star)\,\sqrt{\mathcal{A}}\,\rho(t_k, t^{\star r})}{\sigma^2}$$

K marqueurs en $t_1 = 0 < t_2 < \ldots < t_K = T$

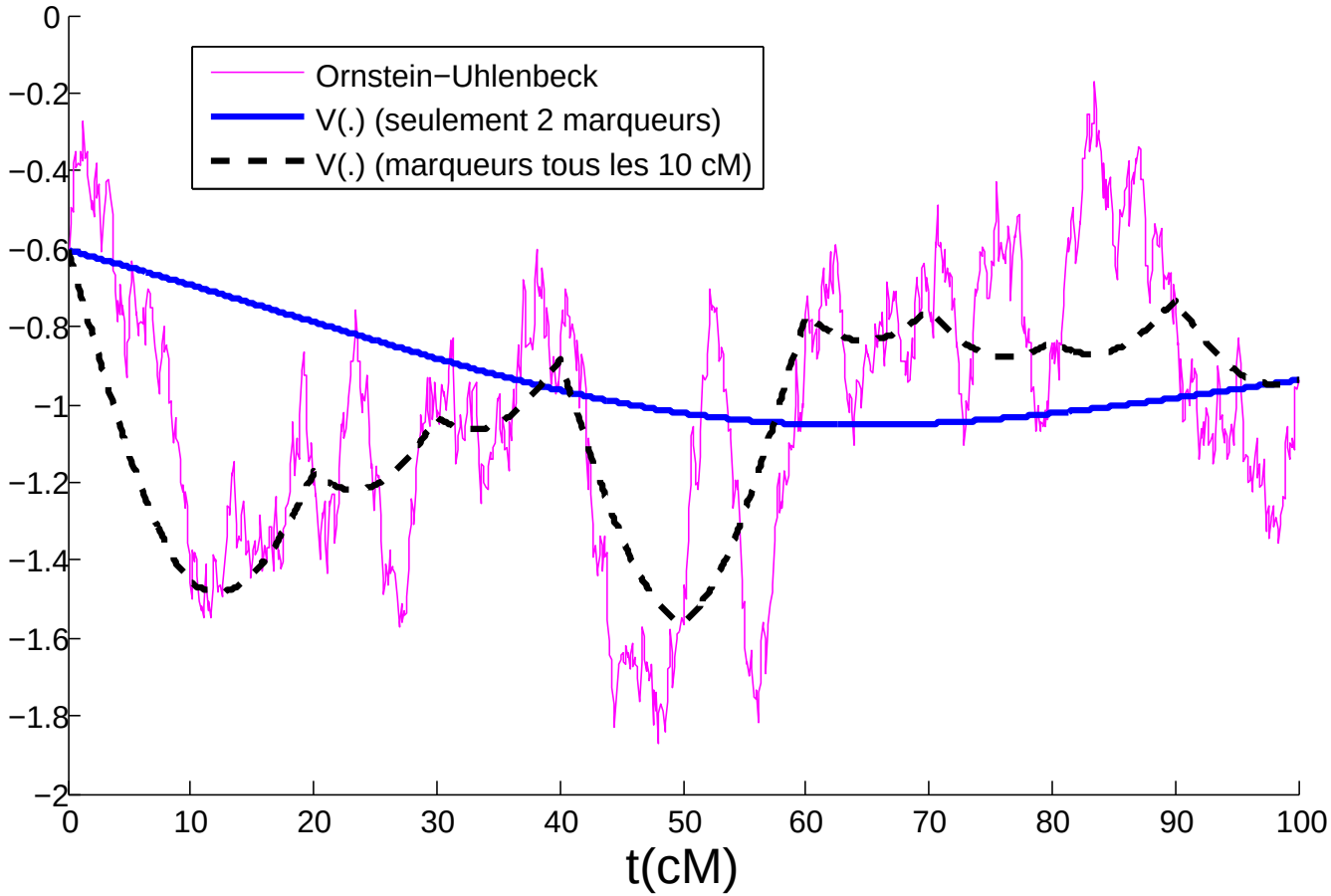
Le cas $t^\star \notin [t_k, t_{k+1}]$

$$\mathrm{Cov}_{H_0}\left\{S_n(t_k), I^n_{t^\star}(\theta) - I^n_{t^\star}(\theta_0)\right\} = \frac{a\,\alpha(t^\star)\,\sqrt{\mathcal{A}}\,\rho(t_k, t^{\star\ell})}{\sigma^2} + \frac{a\,\beta(t^\star)\,\sqrt{\mathcal{A}}\,\rho(t_k, t^{\star r})}{\sigma^2}$$

Comme $\alpha(t^\star)\rho(t_k, t^{\star\ell}) + \beta(t^\star)\rho(t_k, t^{\star r}) = \rho(t_k, t^\star)$,

$$S_n(t_k) \xrightarrow{\mathcal{L}} N\left(\frac{a\,\sqrt{\mathcal{A}}\,\rho(t_k, t^\star)}{\sigma^2},\ 1\right).$$

L'Interval Mapping lisse les trajectoires



Une interpolation non linéaire

Lemme

Soit $T_n(.)$ le processus tel que

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

$$T_n(.) \Rightarrow V(.) \quad et \quad T_n^2(.) \Rightarrow V^2(.) \ .$$

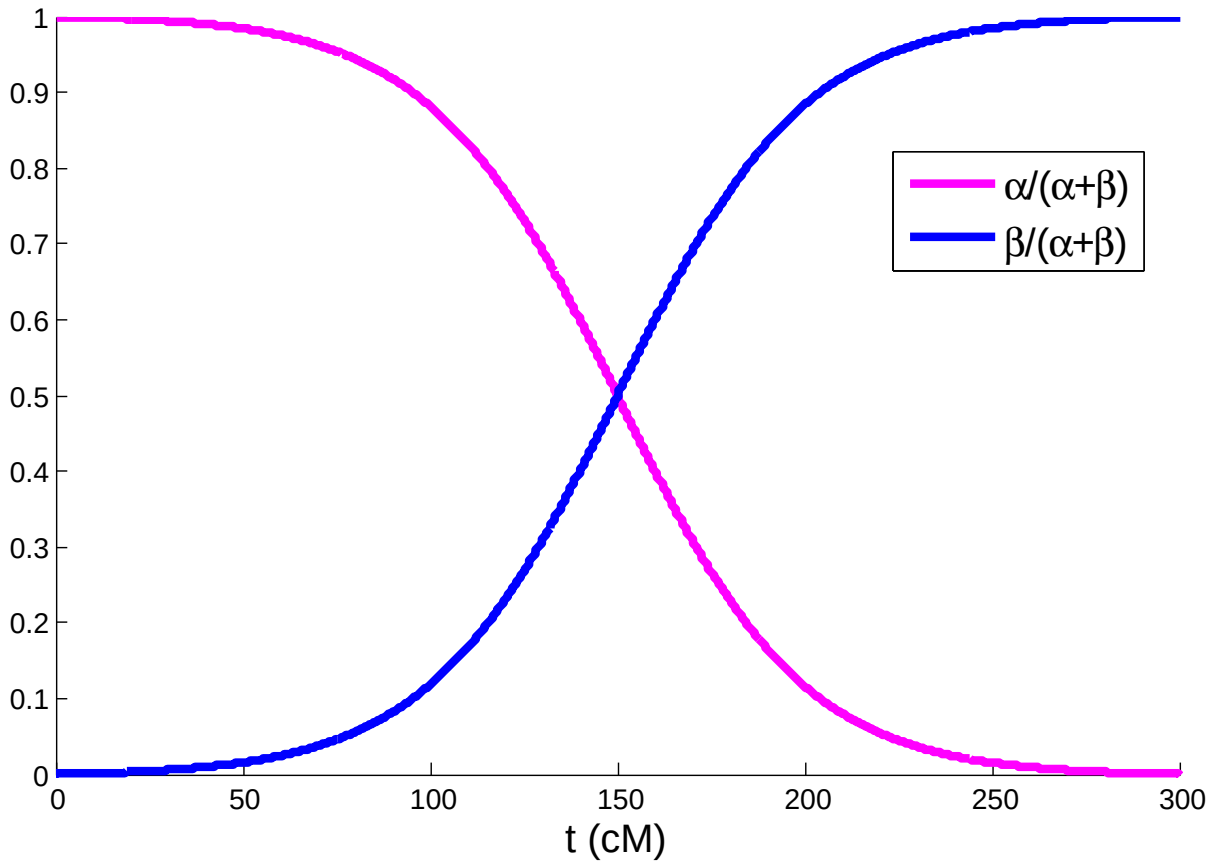
Un lemme utile pour les processus interpolés

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*

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

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

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

Inutile d'effectuer des tests partout sur le chromosome !

Application au calcul de valeurs critiques

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

$\Rightarrow$ fonction QSIMVNEF de Genz (1992)

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

$T = 1$M, 101 marqueurs équidistants,
$\gamma = 1$, 10000 échantillons

Un exemple avec un maximum de 657 tests sur le génome

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

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

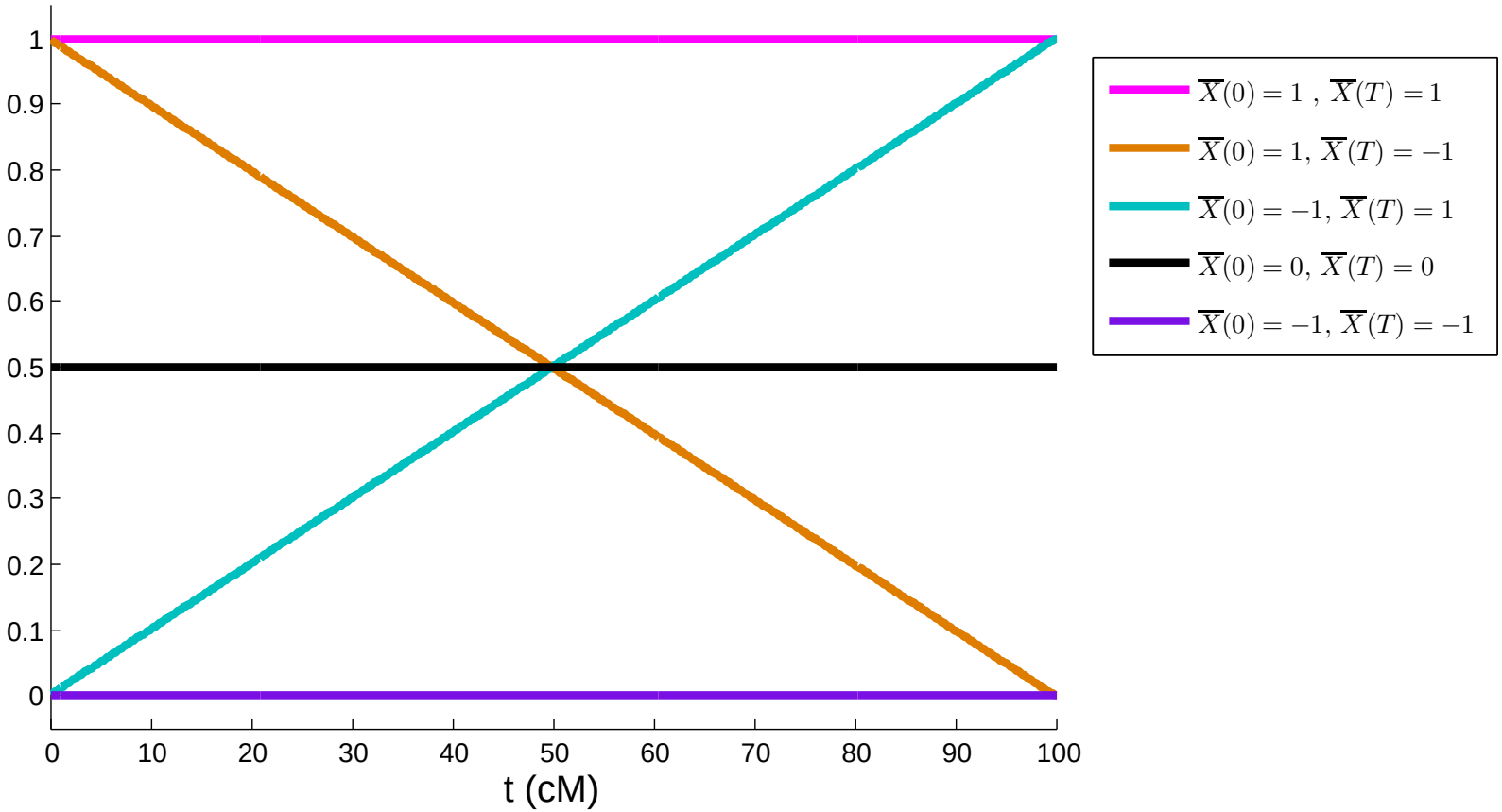
En $t = t^\star$

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

où

$$\widetilde{p}(t^\star) = 1_{\overline{X}(t^{\star\ell})=1} 1_{\overline{X}(t^{\star r})=1} + \frac{t^{\star r}-t^\star}{t^{\star r}-t^{\star\ell}} 1_{\overline{X}(t^{\star\ell})=1} 1_{\overline{X}(t^{\star r})=-1} + \frac{t^\star - t^{\star\ell}}{t^{\star r}-t^{\star\ell}} 1_{\overline{X}(t^{\star\ell})=-1} 1_{\overline{X}(t^{\star r})=1}$$

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



Une interpolation linéaire

Théorème (R., JSPI 2014)

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

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

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

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

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

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

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

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

Interpolation linéaire/Interpolation non linéaire

Lemme

Sous l'hypothèse nulle H_0,

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

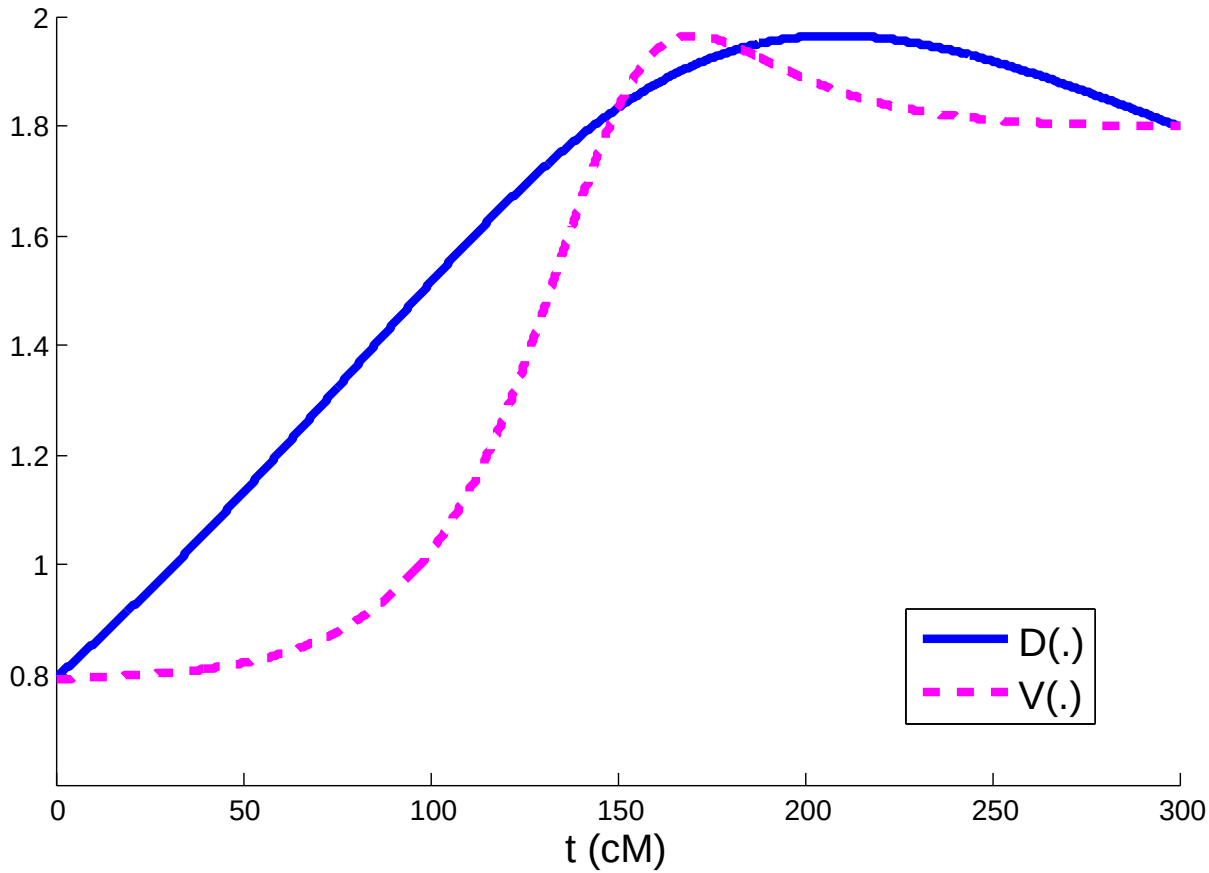
où $V(.)$ est le processus d'interpolation non linéaire, obtenu sous le modèle de Haldane.

Le seuil est donc le même
sous Haldane et sous le modèle d'interférence !

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

Robustesse asymptotique du LRT (R., EJS 2014)
car on se promène sur le génome ....

Une trajectoire des processus $D(.)$ et $V(.)$ sous H_0 ($K = 2$)



A propos de l'argmax des processus ($K = 2$) lorsqu'il est obtenu entre les marqueurs

- $D(0) = V(0)\,, \quad D(T) = V(T)$
- $\tilde{\xi} = \arg\max D^2(.), \quad \xi = \arg\max V^2(.)$

Lemme

Sous H_0 et $H_{at^\star}$,

- *Si $D(T)/D(0) \in]\rho(0,T), 1/\rho(0,T)[$, alors*

$$\tilde{\xi} = \frac{T\{\rho(0,T)D(0) - D(T)\}}{\{\rho(0,T) - 1\}\{D(0) + D(T)\}}$$

$$\frac{T\beta(\xi)}{\alpha(\xi) + \beta(\xi)} = \tilde{\xi}$$

Génotypage des non extrêmes (expérience inverse du selective genotyping)

$\mathcal{B}$ est la quantité suivante

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

Théorème

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

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

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

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

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

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

Oracle : tous les individus sont génotypés aux marqueurs

Théorème

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

κ *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

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

Théorème

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

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

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

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

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

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

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

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

On établit la relation :

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

avec $G(.)$ OUCS et $\vec{W}(t) = \begin{pmatrix} W_1(t) \\ \vdots \\ W_l(t) \end{pmatrix}$ 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

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

- d'une borne inf, obtenue par MCQMC (en collaboration avec Alan Genz, MCAP 2013)

Conclusions

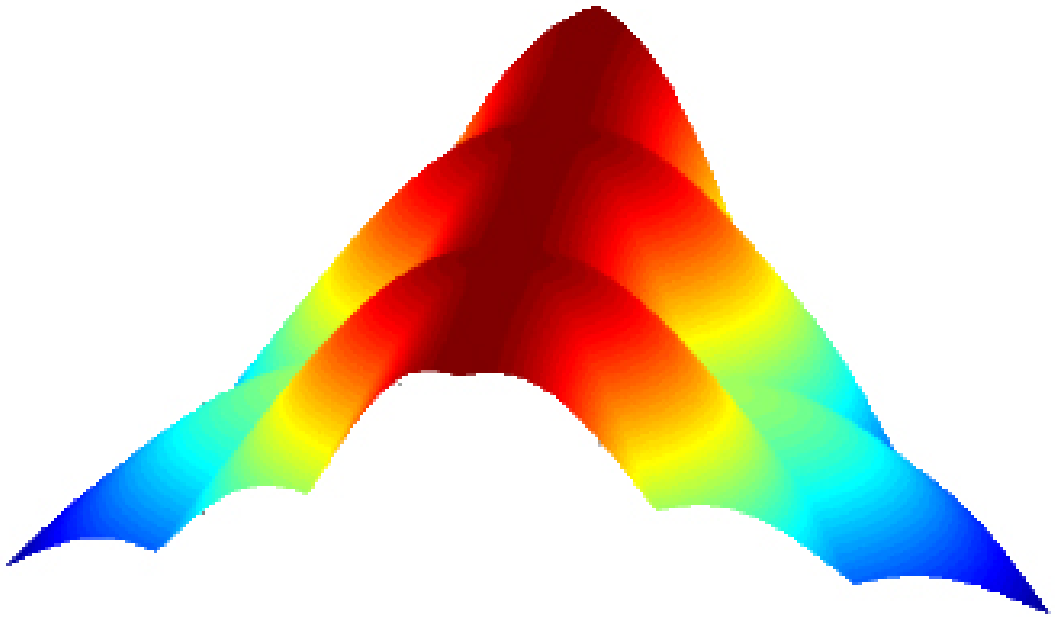
Selective Genotyping sur un marqueur :

- 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 l familles :
  $\Rightarrow$ Processus de Chi deux d'Ornstein-Uhlenbeck à l degrés de liberté

Merci de votre attention



Puissance théorique/Puissance empirique sous l'alternative $H_{at^\star}$

γ_+ \ $t^\star$	12cM	36cM	52cM	75cM
γ	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%)

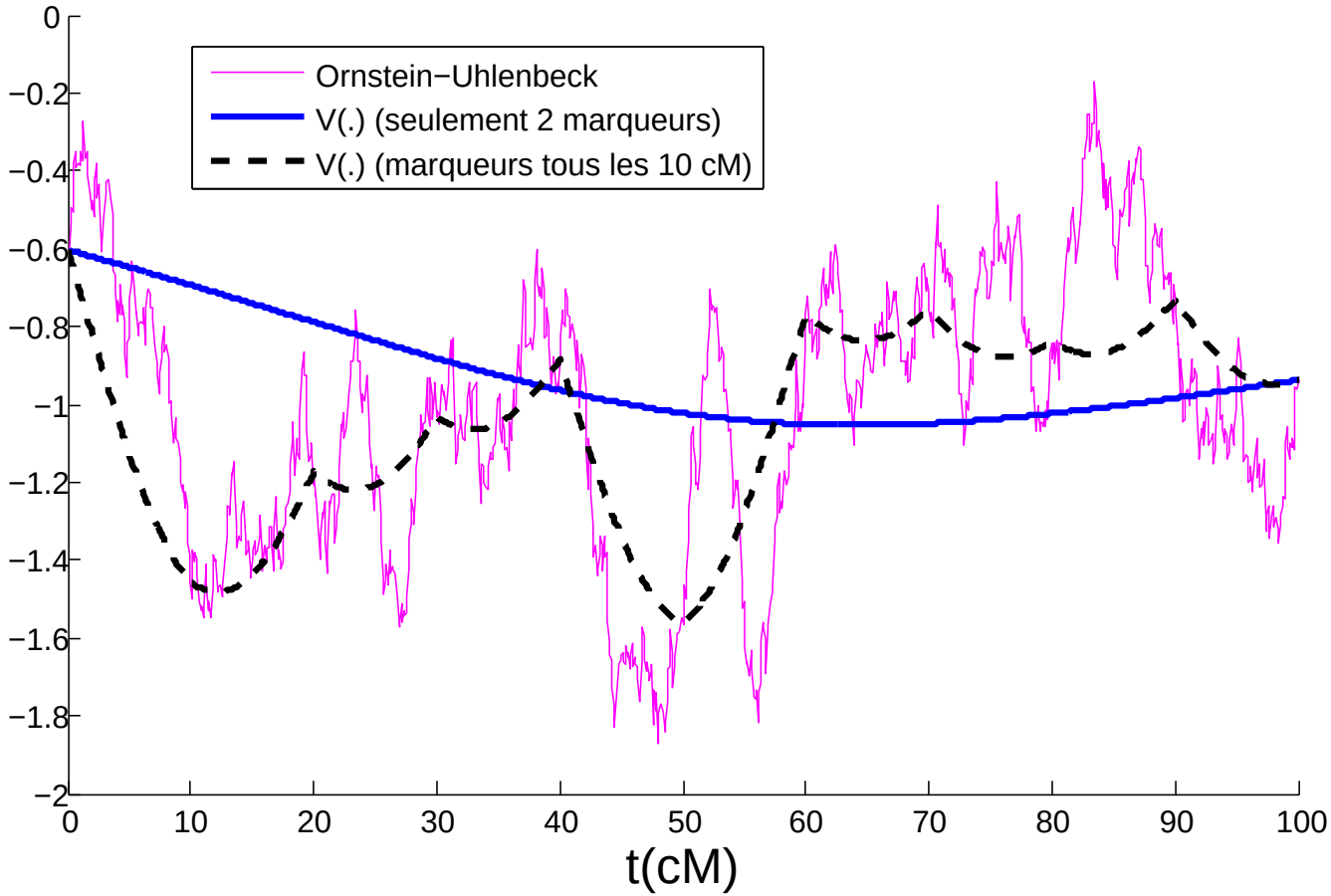
$T = 1$M, $K = 11$, marqueurs équidistants tous les 10cM, $\gamma = 0.3$, $a = 4$, 10000 échantillons de taille $n = 1000$, 100000 trajectoires pour la puissance théorique

Puissance théorique (expérience inverse du selective genotyping)

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

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

L'Interval Mapping lisse les trajectoires ($\gamma = 1$)



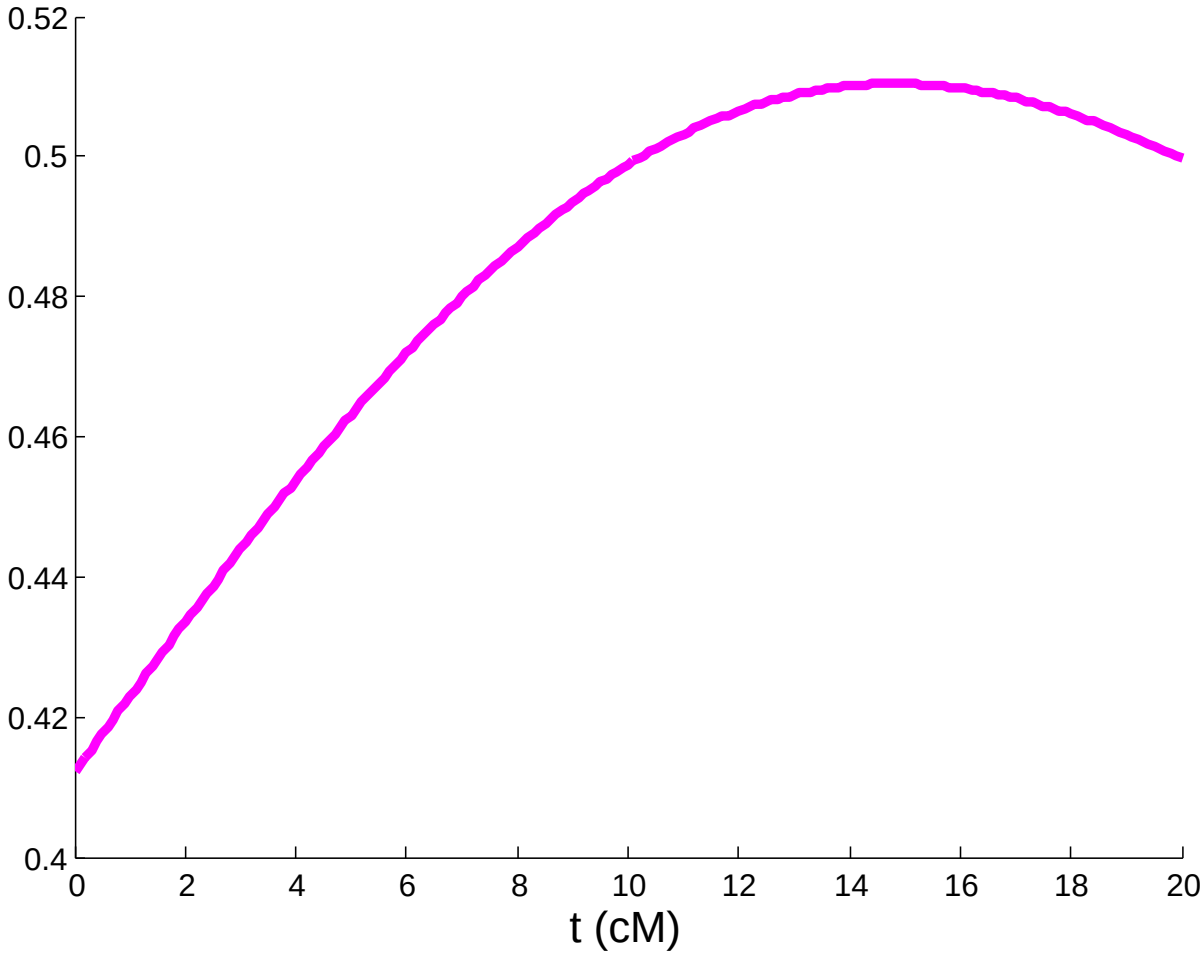
Conclusions sur le Selective Genotyping

- 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 côut génotypage/phénotypage)
- Le test de comparaison de moyenne est optimal

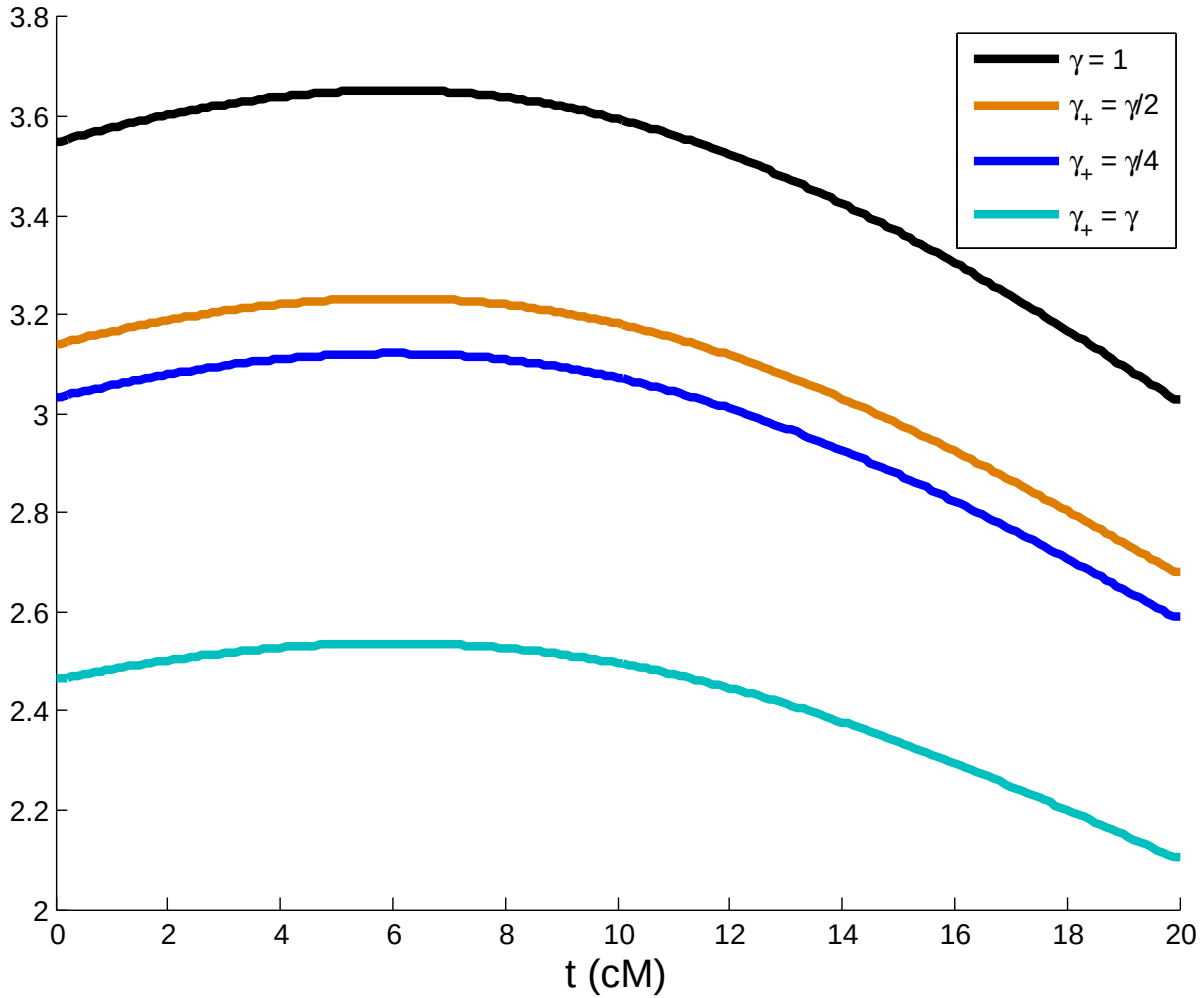
	$n = 50$		$n = 100$	
Nombre de QTLs	W_1	T_2	W_1	T_2
1	0.0020	0.0005	0.0041	0.0005
1000	2.7871	0.1267	5.1131	0.1384

Comparaison en temps de calcul
($q = 0.3$, $p = 1/2$, $\gamma = 0.3$, $\gamma_+/\gamma = 1/2$)

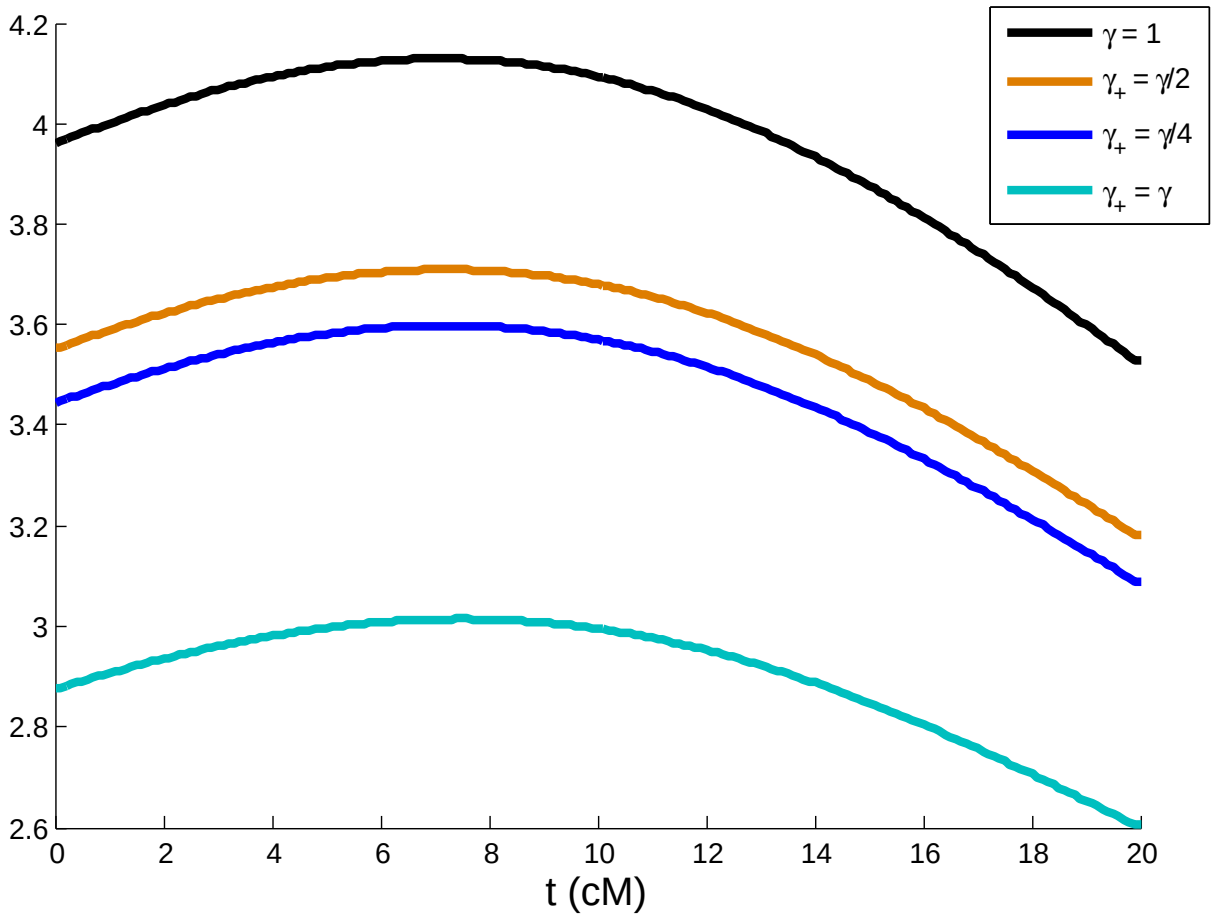
Une trajectoire du processus $V(.)$ sous H_0



Fonction moyenne du processus $V(.)$ sous $H_{at^\star}$ ($t^\star = 6$cM, $\gamma = 0.3$)



Une trajectoire du processus $V(.)$ sous $H_{at^\star}$ ($t^\star = 6$cM, $\gamma = 0.3$)



Pourcentage de faux positifs sous l'hypothèse nulle H_0

γ_+ \ n	1000	200	100	50
γ	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%

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

Puissance théorique en fonction du modèle

γ_+	$t^\star$	interference	Haldane
γ	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%

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

Application au calcul de valeurs critiques

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

$\Rightarrow$ 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%

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

Efficacité du test de Wald (μ, q, σ)

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

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

De la même manière, on note :

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

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 $H_{at^\star}$ ($K = 2$, $t_1 = 0$, $t_2 = T$)

On a

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

De la même façon,

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

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

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

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

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

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

Une seule recombinaison autorisée entre le QTL et les deux marqueurs

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

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

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

- Proportionnalité

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

- Modèle "d'analyse de variance"

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

Selective genotyping + interférence

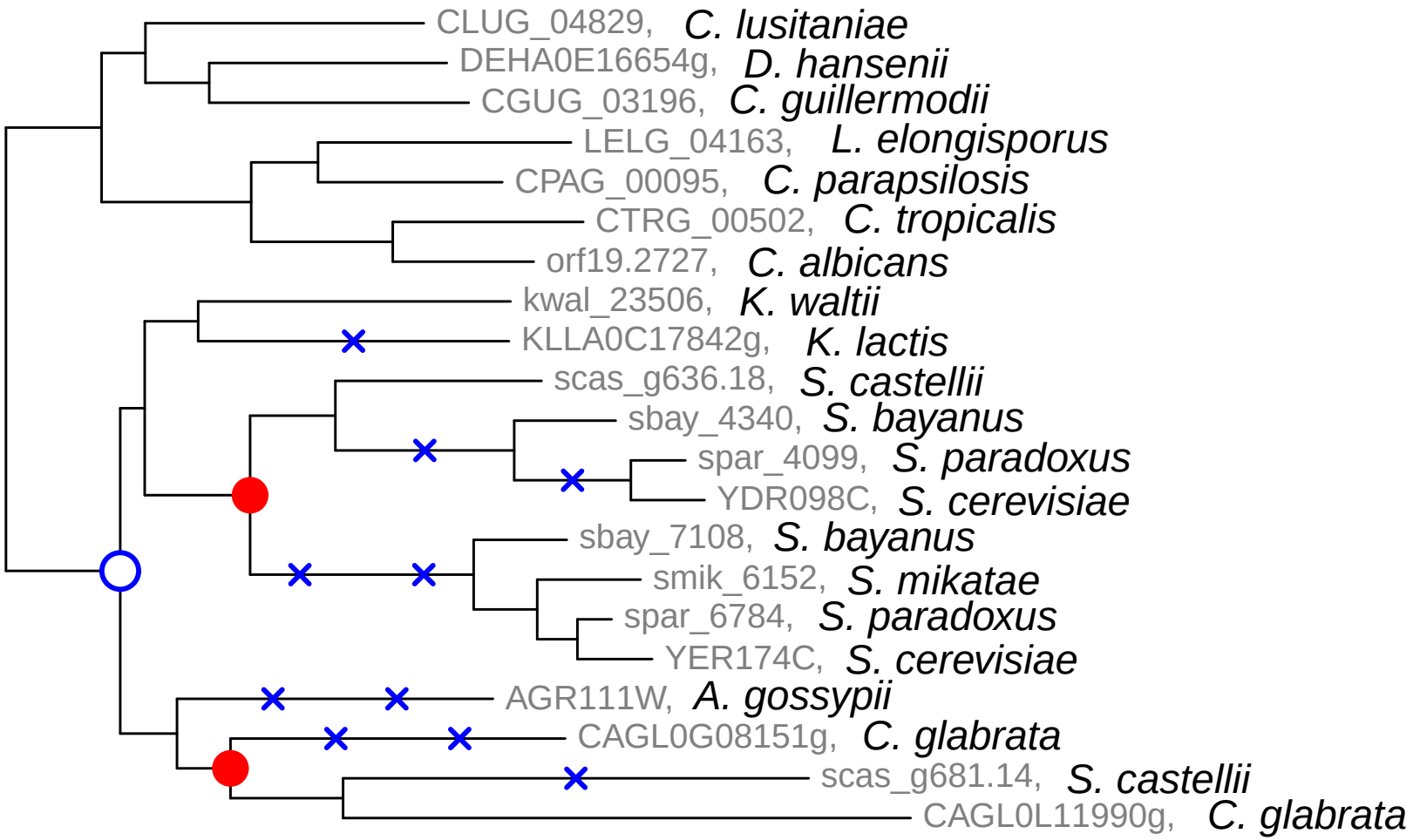
Configuration classique :

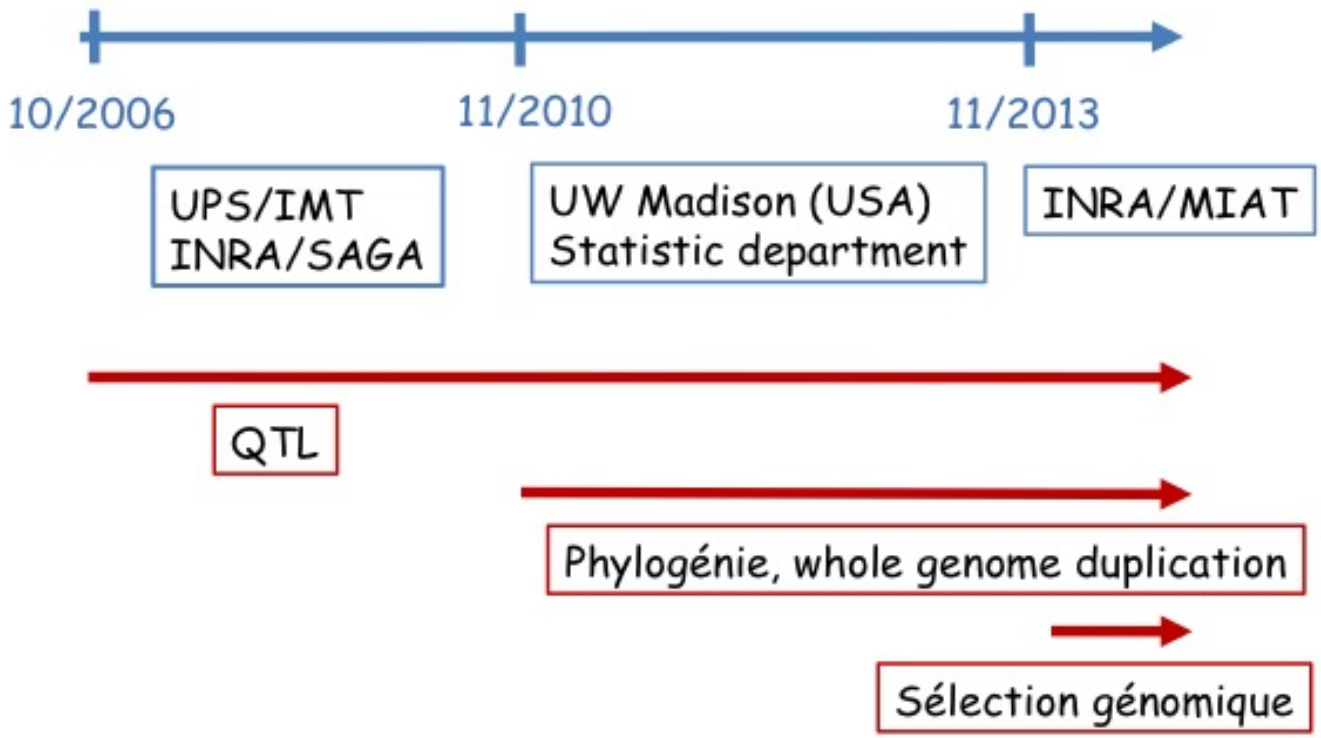
- On observe $\left(\tilde{Y},\ X(0),\ X(T)\right)$

Configuration selective genotyping :

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





Thème de recherche: Méthodes statistiques et probabilistes pour le génome