Gaussian and Chi Square processes for Quantitative Trait Locus mapping under selective genotyping

Charles-Elie Rabier

INRA, MIAT, Toulouse, France

07/12/2014

What is a QTL ?

QTL = Quantitative Trait Locus

A QTL is a locus responsible
for the variation of a quantitative trait

How can we detect QTL ?

We need :

- a segregating population (obtained with the help of genetic crosses)
- genetic markers located on the genome
- phenotypes (i.e. trait)

$\Rightarrow$ statistical methods will help us to detect and find the QTL

First part :

Selective Genotyping

The QTL is located on a genetic marker

Oracle situation : all the genotypes are known

- X : random variable corresponding to the genotype at the QTL

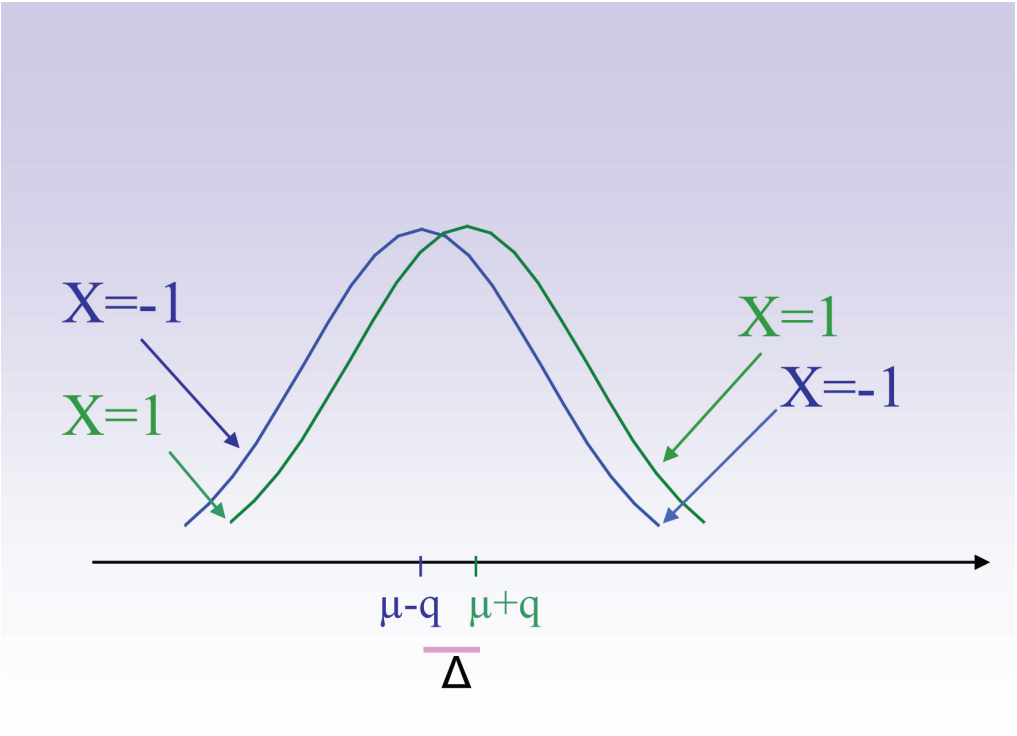
$$X = \begin{cases} -1 & \text{with probability } 1-p \\ 1 & \text{with probability } p \end{cases}$$

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

- Y : random variable corresponding to the phenotype

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

Oracle situation : all the genotypes are known



Probability distribution of the phenotypes Y

Oracle statistical test (μ, q, σ)

- Using a sample of n observations (X_j, Y_j) i.i.d., we test :

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

We consider a local alternative $H_a : q = \frac{a}{\sqrt{n}}$

- Oracle statistical test :

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

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

Selective Genotyping

Genotyping is expensive

$\Rightarrow$ Selective Genotyping : we genotype only individuals who present extreme phenotypes Y.

At a given power, a large increase of the number of individuals leads to a decrease of the number of individuals genotyped

Lebowitz et al. (Theoretical and Applied Genetics, 1987)
Darvasi and Soller (Theoretical and Applied Genetics, 1992)

Genes for fat deposition in Italian pigs (Fontanesi et al. 2012)
Genes for the Growth of the Clam Meretrix (Lu et al. 2013)

Model under selective genotyping

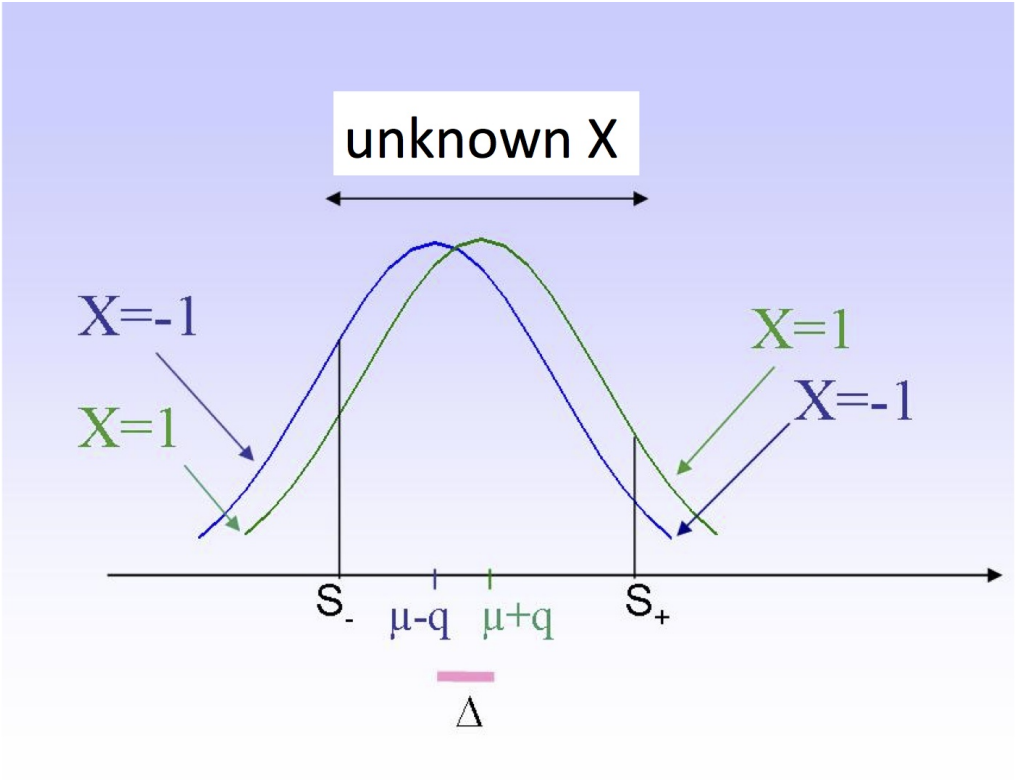
X is available only for individuals who present extreme phenotypes Y

$\Rightarrow$ We observe $\overline{X}$ instead of X :

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

where S_- et S_+ are two real thresholds such as $S_- \leq S_+$.

Model under selective genotyping



Probability distribution of the phenotypes Y

Wald test (μ, q, σ)

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

We consider a local alternative $H_a : q = \frac{a}{\sqrt{n}}$

- Wald statistic

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

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

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

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

How to optimize the selective genotyping

- We would like to genotype only a percentage γ of the population

$\Rightarrow$ How should we choose the optimal γ_+ and γ_- ?

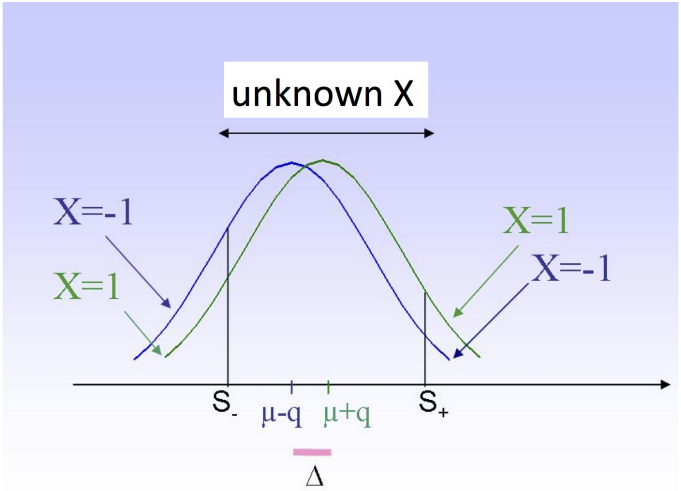
$\forall p$, κ_1 reaches its maximum M for $\gamma_+ = \gamma_- = \gamma/2$

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

$\forall p$, we should genotype symmetrically

3 strategies suitable for the data analysis under selective genotyping

1. Wald test based on all the phenotypes (even the phenotypes for which the genotypes are missing)
2. Comparison of means based on the extreme phenotypes
3. Wald test based only on the extreme phenotypes



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

Comparison of the 3 strategies (μ, q, σ)

Lemma

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

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

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

have the same asymptotic laws under H_0 and under H_a, that is to say :

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

where $\hat{q}_1$ and $\hat{q}_3$ denote the MLE of q for strategies one and three, and

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

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

Conclusions on Selective Genotyping

- We should genotype symmetrically
- The non extreme phenotypes don't bring any extra information for statistical inference
- We should genotype 30% of the individuals (it depends on the cost ratio genotyping/phenotyping)
- The comparison of means is optimal

R., JSPI 2014

	$n = 50$		$n = 100$	
QTL number	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

CPU time (in seconds)

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

Second part :

Genome Scan

The QTL location is unknown

Context

- The chromosome is represented by a segment $[0, T]$
- The distance on $[0, T]$ is called the genetic distance
- $X(.)$: genome of one individual
- We consider Haldane modeling

Haldane modeling

- No crossover interference
- $N(.)$: standard Poisson process on $[0, T]$

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

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

- Calculations on the Poisson distribution show that

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

Model

- $t^\star$: QTL location
- Y : random variable corresponding to the phenotype

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

- Genome information X(.) available only at fixed locations, called genetic markers
- K genetic markers on $[0, T]$ located at

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

Oracle situation

One observation is

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

and the challenge is that the QTL location $t^\star$ is unknown ! ! !

The Interval Mapping of Lander and Botstein (1989)

We want to test : $H_0 : q = 0$ vs $H_1 : q \neq 0$

The Interval Mapping

- the QTL location $t^\star$ is unknown

$\Rightarrow$ we scan the interval $[0, T]$

$\Rightarrow$ Likelihood Ratio Tests (LRT) on the whole interval

How to perform the LRT

- for each location $t \in [0, T]$ not on markers, "genome information" $X(t)$ unknown

$\Rightarrow$ probability of the genotypes at the QTL using the genome information on markers and Haldane formula

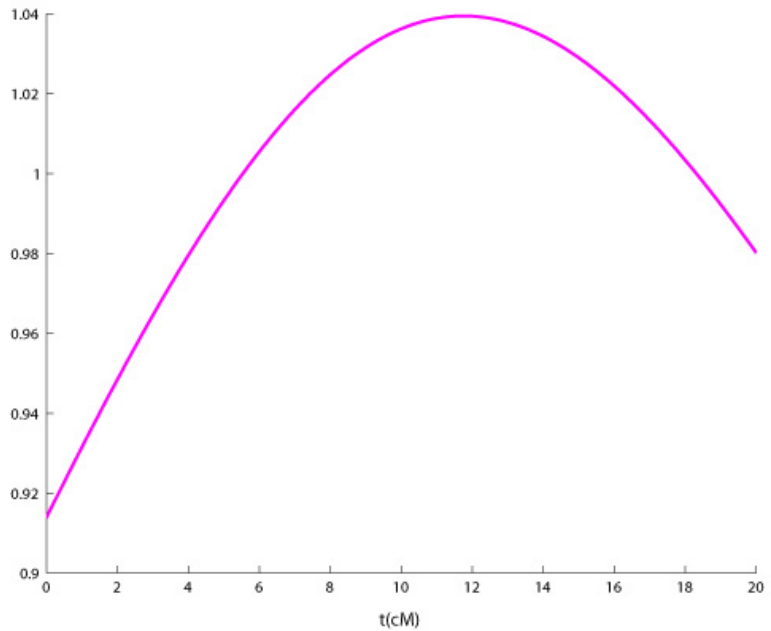
$\Rightarrow$ model of mixture of mixture

The Interval Mapping of Lander and Botstein (1989)

- $\Lambda_n(t)$: LRT at location t
- the $\Lambda_n(t)$ define a process $\Lambda_n(.)$

We look for only one QTL on the interval $[0, T]$

$\Rightarrow$ LRT statistic on the whole interval : $\sup \Lambda_n(.)$



One path of the process $\Lambda_n(.)$ ($T = 20$cM, $K = 2$)

Selective Genotyping framework

$\overline{X}(t)$ is the random variable such as

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

then, in our problem, one observation will be

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

Likelihood of $\left(Y, \overline{X}(t_1), \ldots, \overline{X}(t_K)\right)$

When $t = t^\star$

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

where

- $f_{(m,\sigma)}$ is the Gaussian density with parameters (m, σ)
- $g(.)$ is a function which does not depend on parameters μ, q and σ

and

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

Score statistic and LRT statistic

- $\theta = (q,\ \mu,\ \sigma)$ parameter of the model at t fixed
- $\theta_0 = (0,\ \mu,\ \sigma)$ stands for H_0

Score statistic at t

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

with $l_t^n(\theta)$ log likelihood at t, associated to n observations.

LRT statistic at t

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

with $\widehat{\theta}$ MLE, and $\widehat{\theta}_{|H_0}$ MLE under H_0.

- known $t^\star \Rightarrow$ regular model
- unknown $t^\star \Rightarrow$ irregular model (under H_0, the Fisher Information Matrix relative to t is equal to zero)

About the hypotheses tested

H_0 : "there is no QTL on $[0, T]$"

$H_{at^\star}$: "the QTL is located at $t^\star \in [0, T]$ with effect $q = a/\sqrt{n}$ "

A non linear interpolation

Theorem (R., Statistics 2013)

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

- $V(.)$ *is the non linear interpolated process such as*

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

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

- $V(.)$ *is a Gaussian process with unit variance and with expectation :*

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

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

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

Efficiency κ of the LRT on the whole chromosome

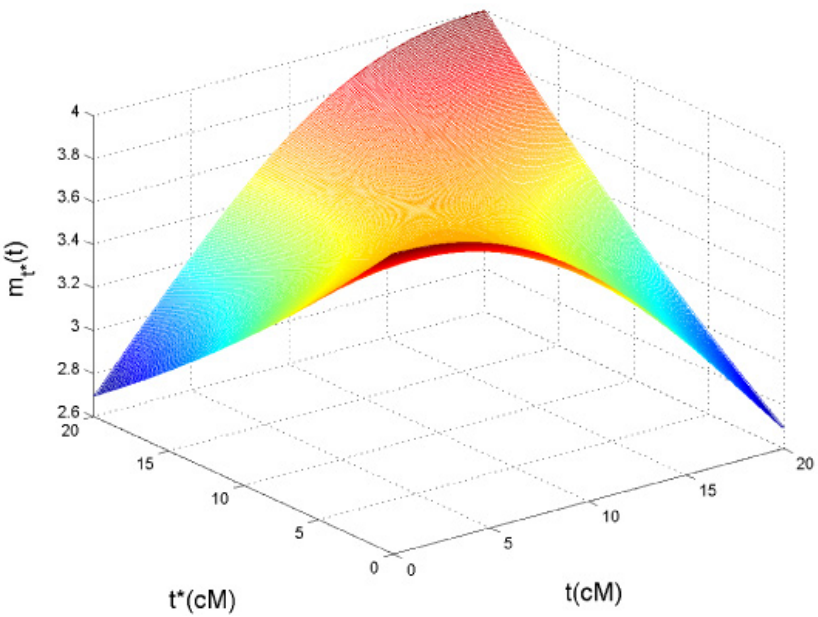
Oracle : no selective genotyping (i.e. the genome information on markers is available for all the individuals)

Lemma

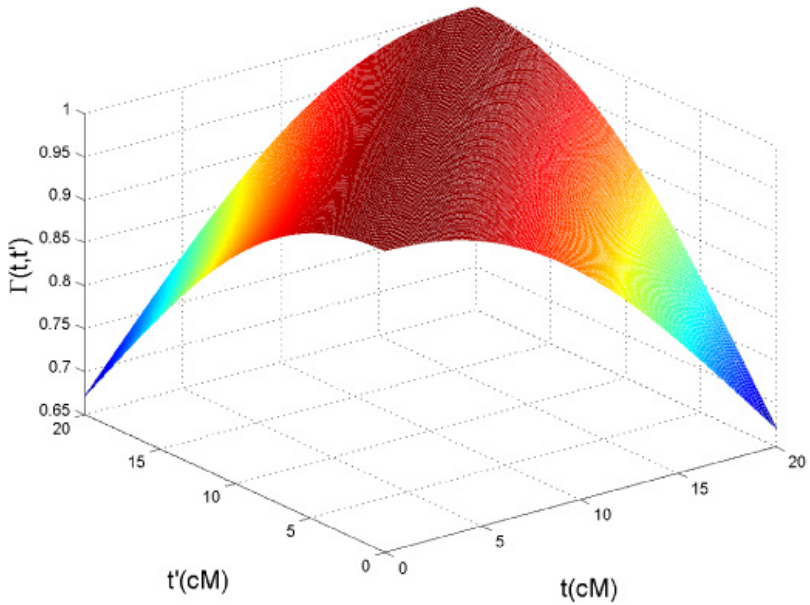
$$\kappa = \gamma + z_{\gamma_+}\varphi(z_{\gamma_+}) - z_{1-\gamma_-}\varphi(z_{1-\gamma_-}) = \mathcal{A}/\sigma^2$$

κ reaches its maximum for $\gamma_+ = \gamma_- = \gamma/2$

Mean function and covariance function ($a = 4$, $\sigma = 1$, $T = 20$cM, $\gamma = 1$, $\sigma = 1$)

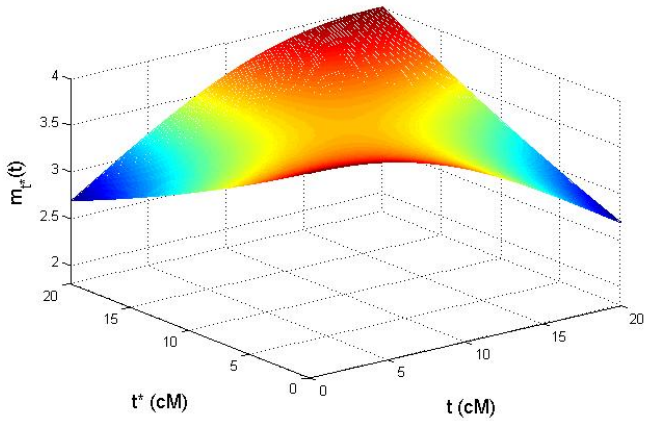


Mean function

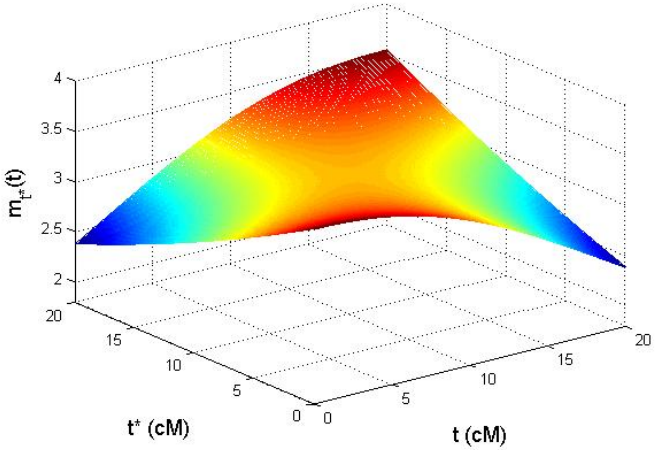


Covariance function

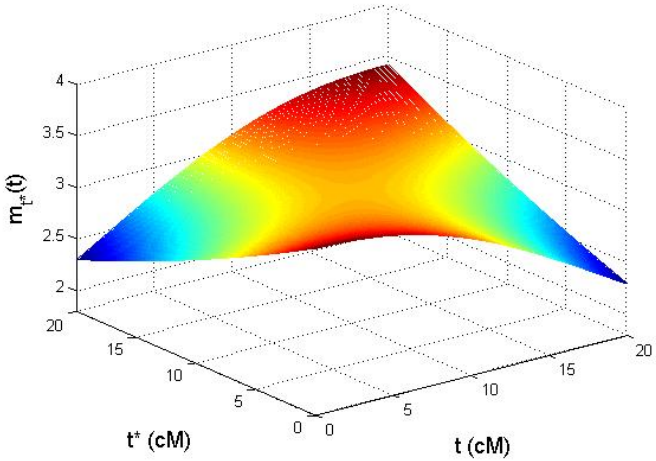
Mean function under selective genotyping



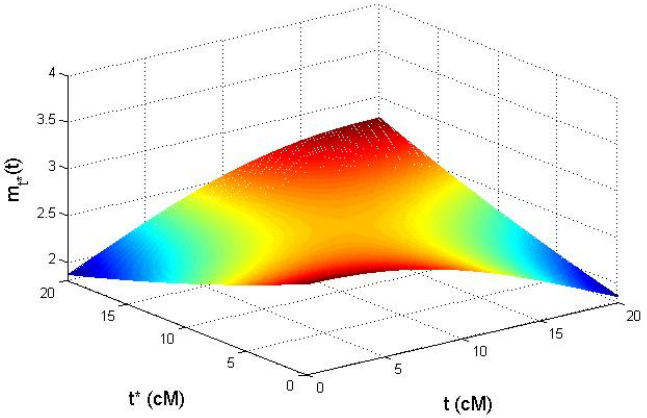
$\gamma = 1$



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

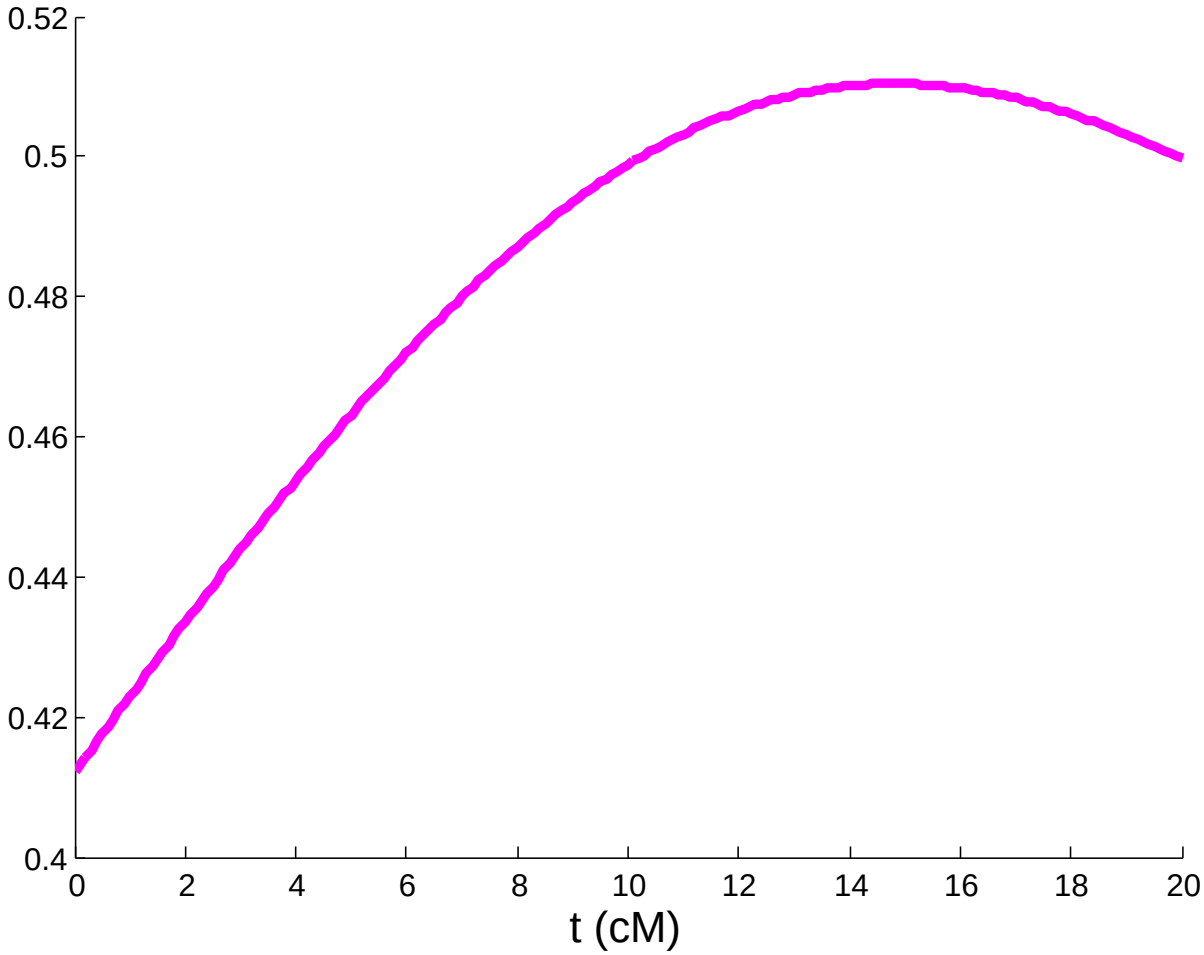


$\gamma = 0.3,\ \gamma_+ = 3\gamma/4$

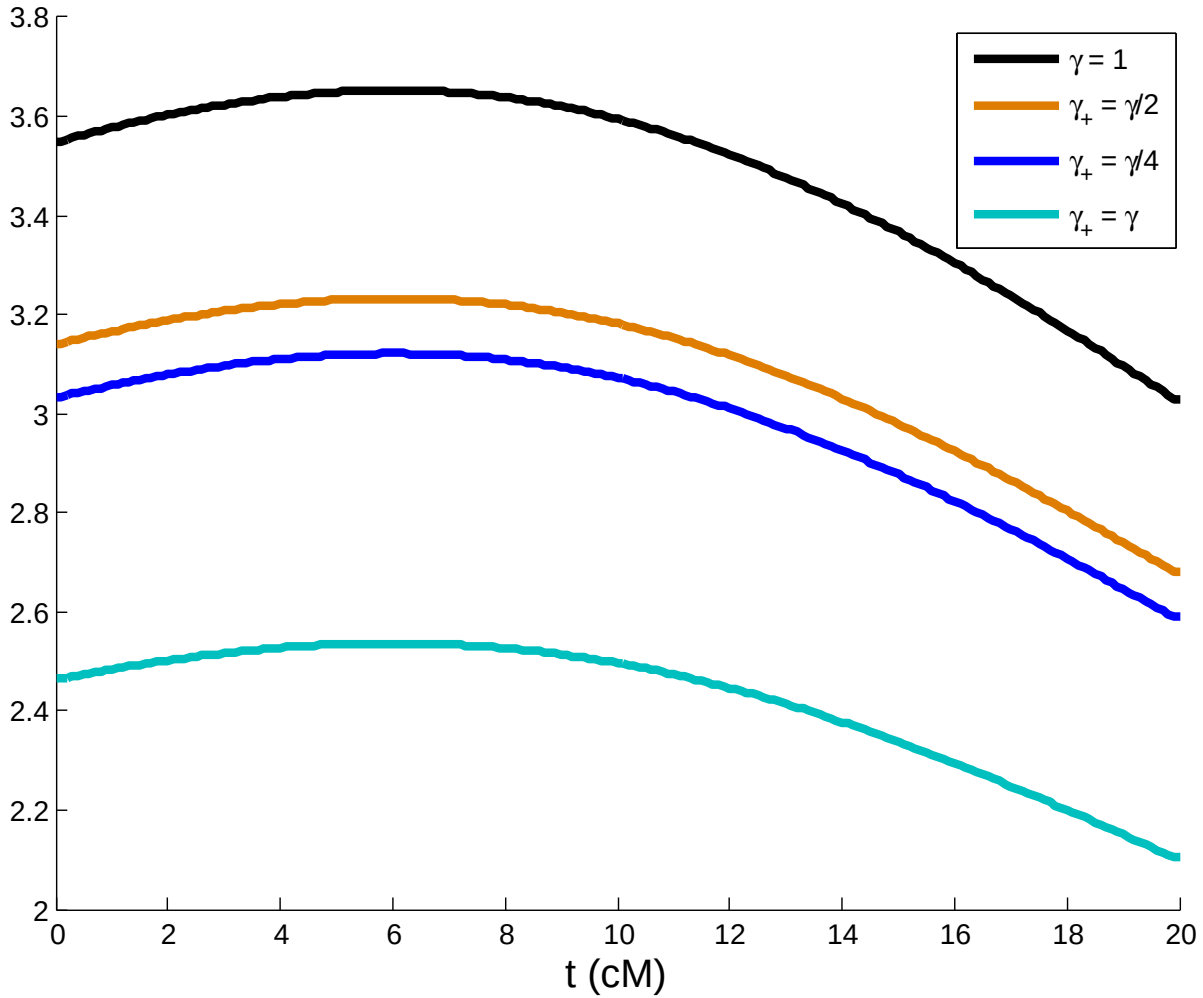


$\gamma = 0.3,\ \gamma_+ = \gamma$

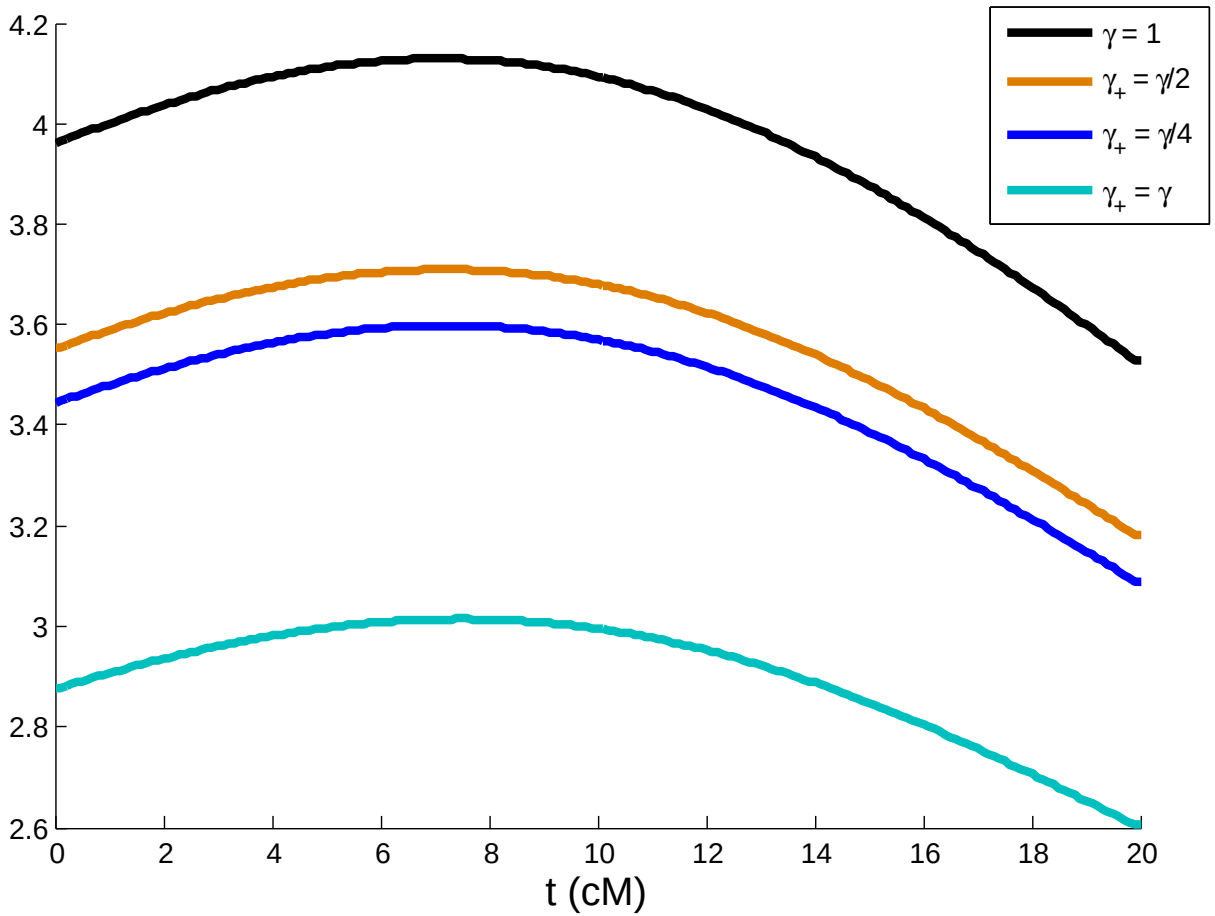
One path of the process $V(.)$ under H_0



Mean function of the process $V(.)$ under $H_{at^\star}$ ($t^\star = 6$cM, $\gamma = 0.3$)



One path of the process $V(.)$ under $H_{at^\star}$ ($t^\star = 6$cM, $\gamma = 0.3$)



A non linear interpolation

Lemma

Let $T_n(.)$ be the process such as

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

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

About the supremum ...

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

where

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

We should not perform tests everywhere on the chromosome ! ! !

Application to threshold calculations

Computation of the critical value c verifying $P_{H_0}(\sup V^2(.) > c) = 1 - \alpha$

$\Rightarrow$ QSIMVNEF function (Genz, 1992)

	K	101
Rebaï (94)	c	9.74
	$n = 200$	2.55%
	$n = 100$	2.52%
	$n = 50$	2.01%
Feingold (93)	c	8.45
	$n = 200$	4.67%
	$n = 100$	4.72%
	$n = 50$	3.92%
Azaïs (2012)	c	8.41
	$n = 200$	4.76%
	$n = 100$	4.80%
	$n = 50$	3.97%

$T = 1$M, markers equally spaced, $\gamma = 1$

An example with a maximum of 657 statistical tests on the genome

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

Feingold (93)	c	12.55
	$n = 200$	2.85%
	$n = 100$	2.72%
	$n = 50$	2.02%
Azaïs (2012)	c	11.70
	$n = 200$	4.64%
	$n = 100$	4.20%
	$n = 50$	3.39%

Other methods : Manichaïkul et al. (2007), Chang et al. (2009)

Conclusions

Selective Genotyping on one genetic marker :

- We should genotype symmetrically
- The non extreme phenotypes don't bring any extra information for statistical inference
- We should genotype 30% of the individuals (it depends on the cost ratio genotyping/phenotyping)
- The comparison of means is optimal

Genome Scan + Selective Genotyping :

- Non linear interpolation under Haldane
- The threshold is the same with/without selective genotyping
- Comparison of means on each marker
- Only one statistical test between markers
- Linear interpolation under interference
- Asymptotic robustness of the LRT

Thanks to

JM Azaïs, C Delmas, JM Elsen, C Ané