Gaussian processes and Phylogenetic trees in genomics

Charles-Elie Rabier

INRA, MIAT, Toulouse, France

06/24/2015

Research interest

Up to now related to mathematical methods for genomics

- statistical methods (empirical processes, statistical inference, bayesian statistics, high dimensional data analysis)
- probabilistic methods (random trees, combinatorics)

This talk : application to gene mapping and phylogenetics

Gene mapping

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

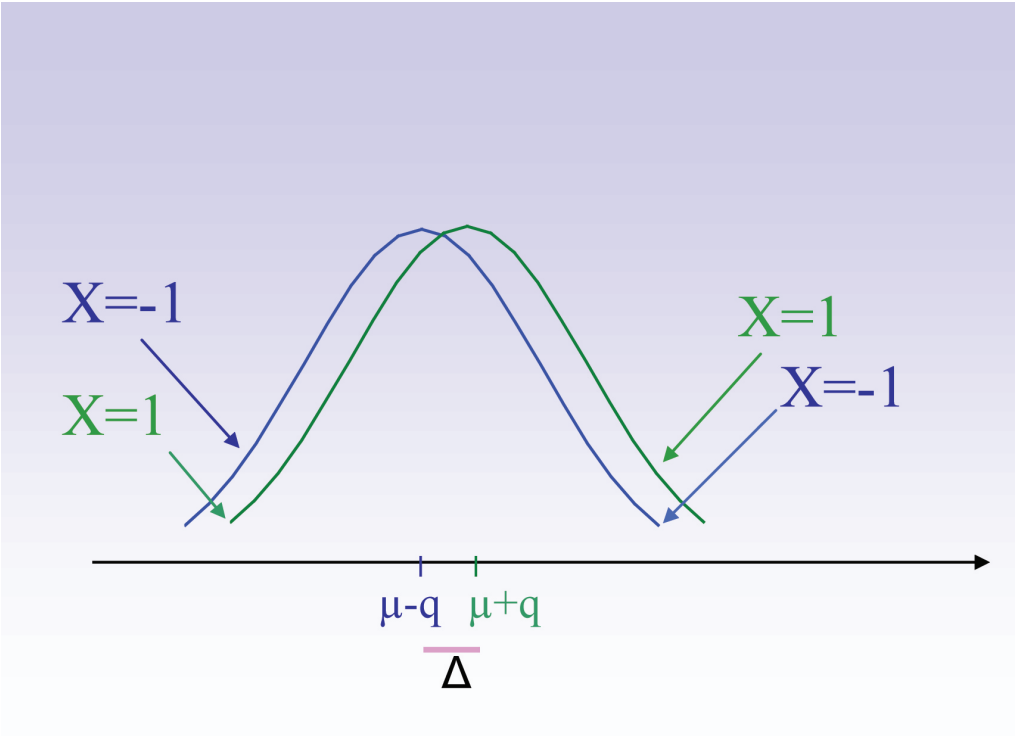
$$X = \begin{cases} -1 & \text{with probability } 1-p \\ 1 & \text{with probability } p \end{cases}$$

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 :

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

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

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

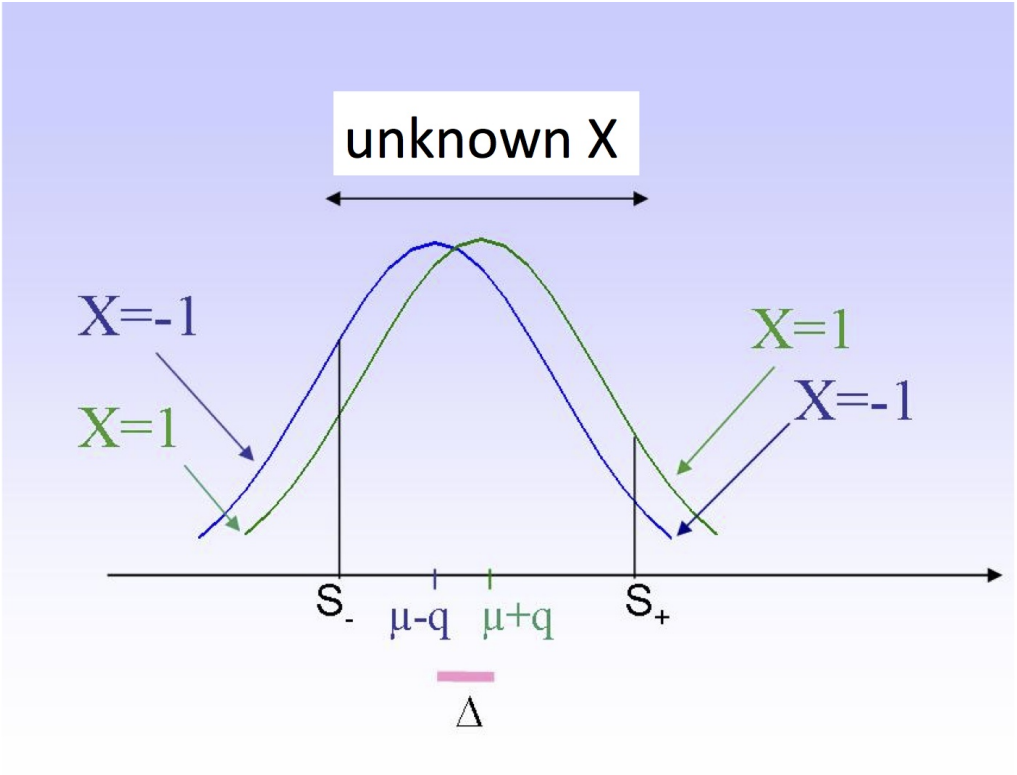
X is available only for individuals who present extreme phenotypes Y

$\Rightarrow$ 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_- ?

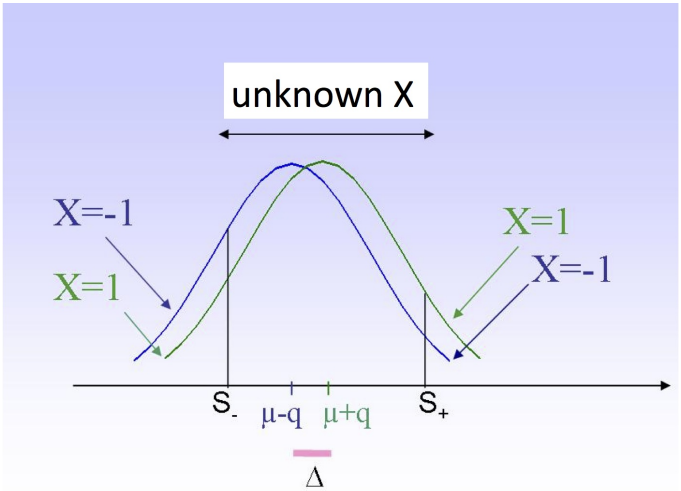
$\forall p$, κ_1 reaches its maximum M for $\gamma_+ = \gamma_- = \gamma/2$

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

$\forall p$, 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



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 distributions 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
- $X(.)$: random process representing the genome of one individual
- We consider Haldane modeling

Haldane Modeling (1919)

- no crossover interference
- $X(t)$: random variable corresponding to the genome information at t

$$X(0) \sim \frac{1}{2}(\delta_{+1} + \delta_{-1}), \quad X(t) = X(0)(-1)^{N(t)}$$

where $N(.)$ is a Poisson process with intensity 1 on $[0, T]$

- $r(t, t')$: probability of recombination between two loci

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

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

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

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

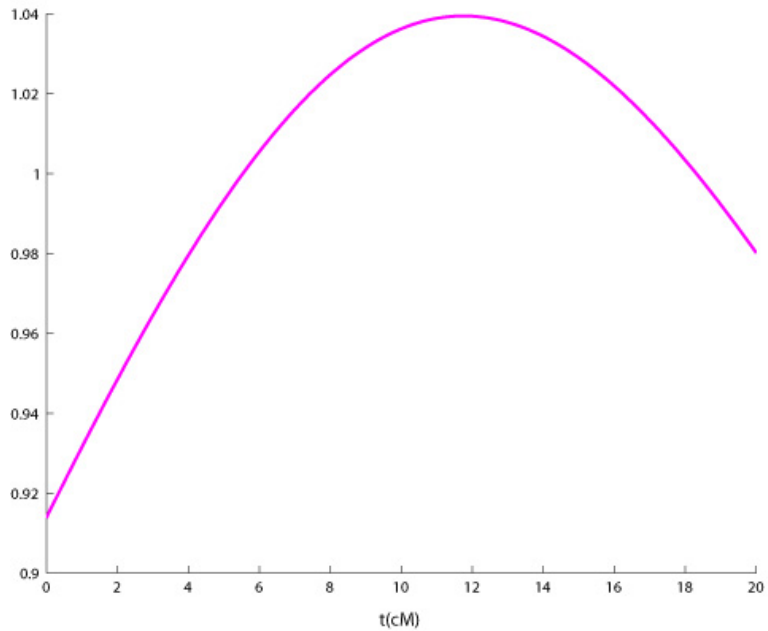
$\Rightarrow$ model of mixture of mixture

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, $K = 2$)

References on asymptotic study of $\sup \Lambda_n(.)$

- Cierco (Statistics, 1998)
- Azaïs et Cierco (Ann. Inst. Henri Poincaré (B), 2002)
- Chen and Chen (Statistica Sinica, 2005)
- Azaïs et Wschebor (Wiley, 2009)
- Chang, Wu, Ma, Casella (SAGMB, 2009)
- Azaïs, Delmas, R. (Statistics, 2012)
- Kim, Cui, Zhao (JSPI, 2013)

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

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

Illustration of the different weights ($K = 2$, $T = 1$M)

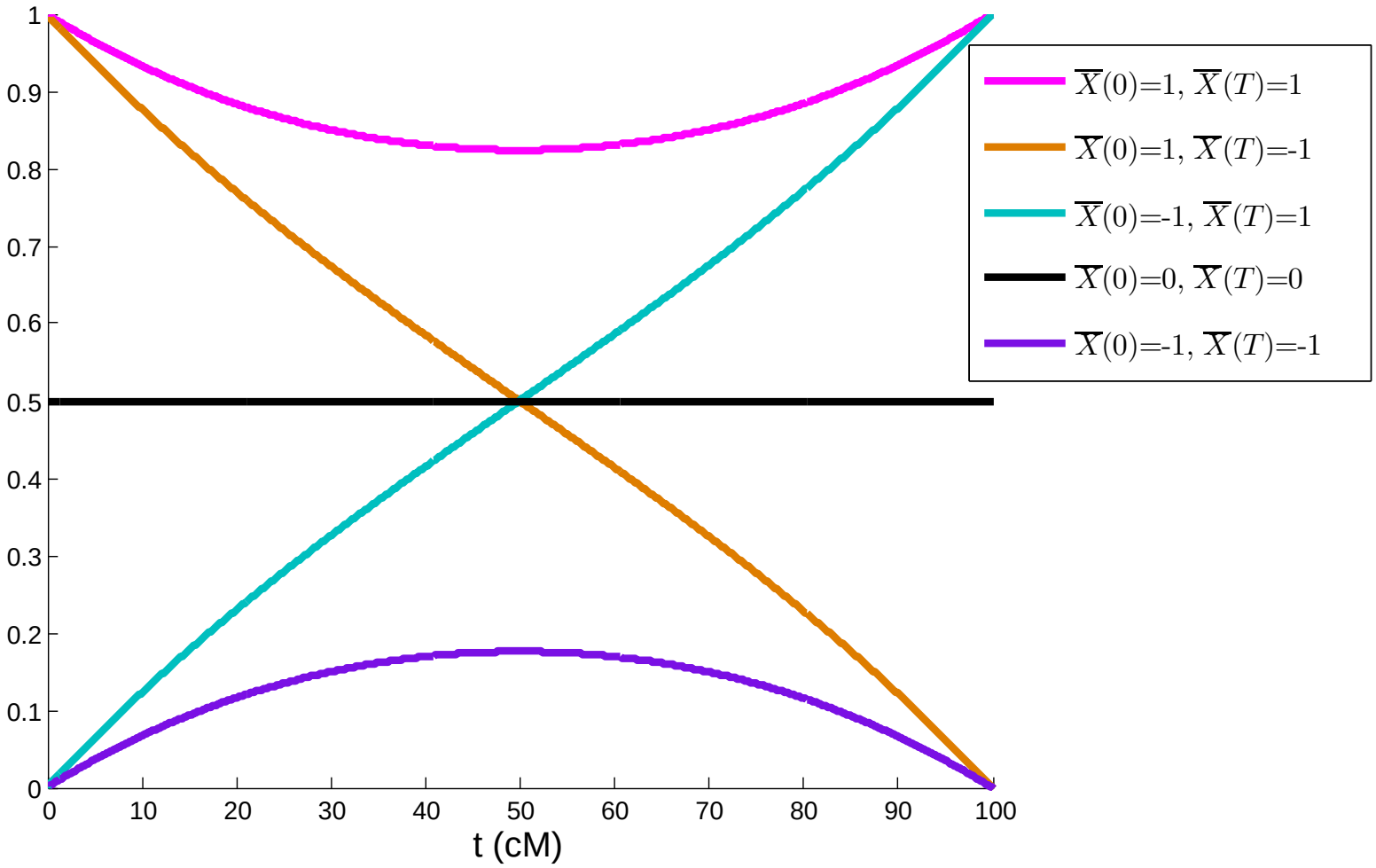
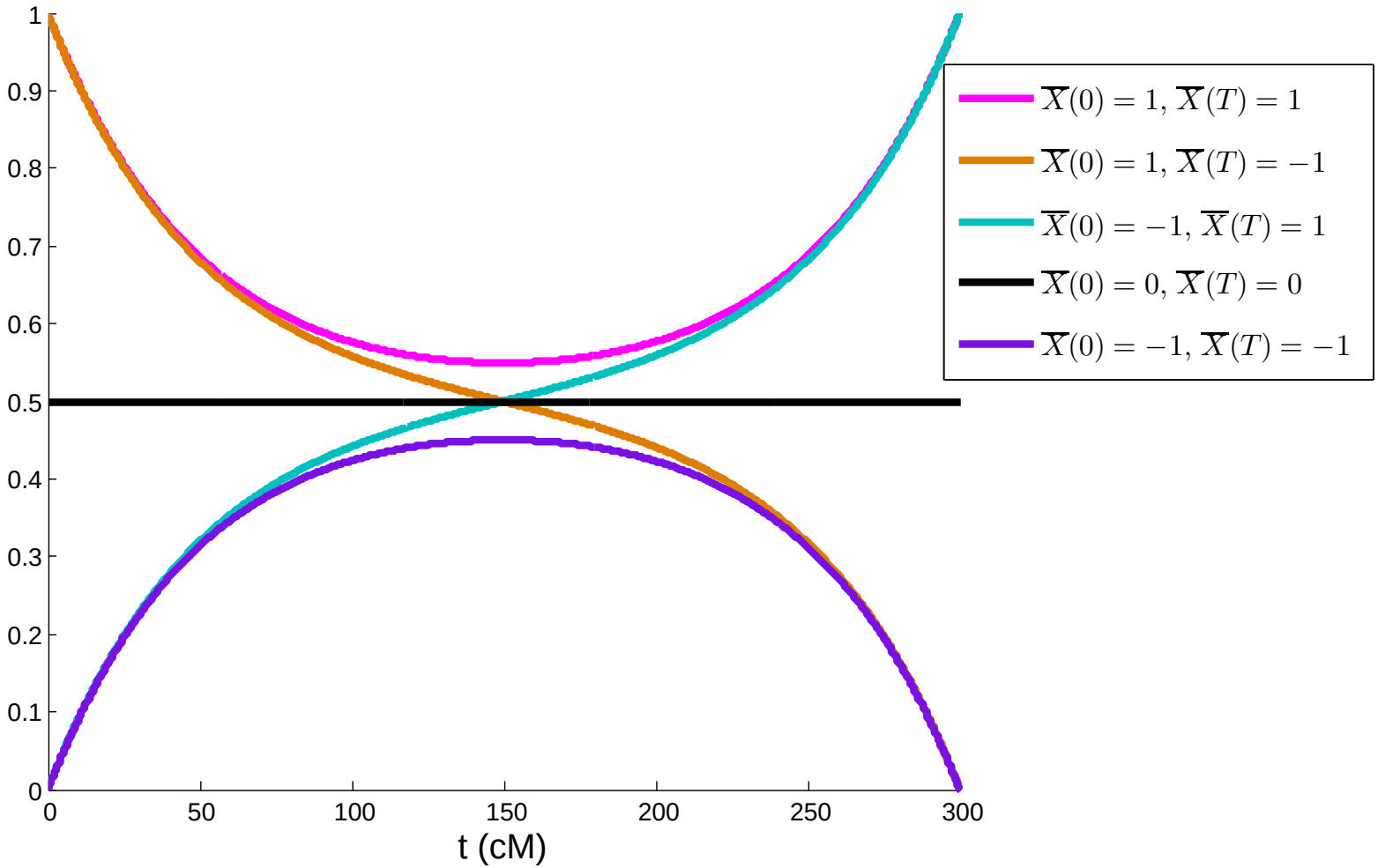


Illustration of the different weights ($K = 2$, $T = 3$M)



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 $q = a/\sqrt{n}$ "

Study of the score process under H_0 ($K = 2$, $t_1 = 0$, $t_2 = T$)

Lemma

We have the following relationship :

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

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

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

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

The limiting process is an interpolated process ! ! !

Study of the score process under $H_{at^\star}$ ($K = 2$, $t_1 = 0$, $t_2 = T$)

Since our model is differentiable in quadratic mean, we apply Theorem 7.2 of Van der Vaart (98). Under H_0, the log likelihood ratio verifies

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

where $o_P(1)$ is a sequence which converges in probability to 0.

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

Study of the score process under $H_{at^\star}$ ($K = 2$, $t_1 = 0$, $t_2 = T$)

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

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

In the same way,

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

Weak convergence of the score process ($K = 2$, $t_1 = 0$, $t_2 = T$)

We have

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

According to the continuous mapping theorem,

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

It proves the convergence of finite-dimensional

Tightness + convergence of finite-dimensional
$\Rightarrow$ weak convergence

Weak convergence of the score process ($K = 2$, $t_1 = 0$, $t_2 = T$)

Theorem 8.2 of Billingsley (1999), the score process is tight if and only if :

1. $S_n(0)$ is tight
2. For each positive $\varepsilon > 0$ and $\eta > 0$,
   there exists a δ, with $0 < \delta < T$ and an integer n_0
   such that $\mathbb{P}\left(w_{S_n}(\delta) \geq \eta\right) \leq \varepsilon \quad \forall n \geq n_0$

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

Weak convergence of the score process ($K = 2$, $t_1 = 0$, $t_2 = T$)

We set

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

We have,

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

Weak convergence of the score process ($K = 2$, $t_1 = 0$, $t_2 = T$)

We show that

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

As a result,

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

A non linear interpolation ($K = 2$, $t_1 = 0$, $t_2 = T$)

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*

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

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

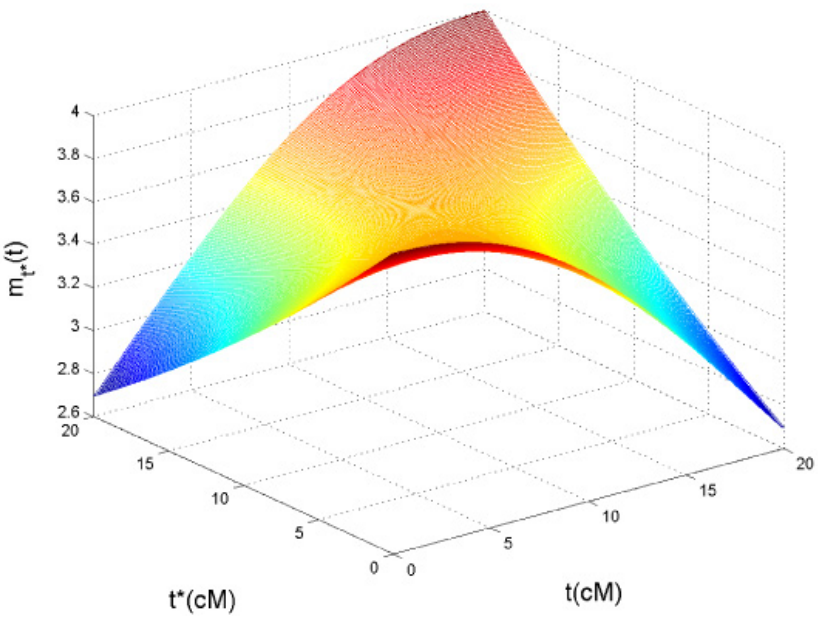
- $V(.)$ *is a Gaussian process with unit variance and with expectation :*

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

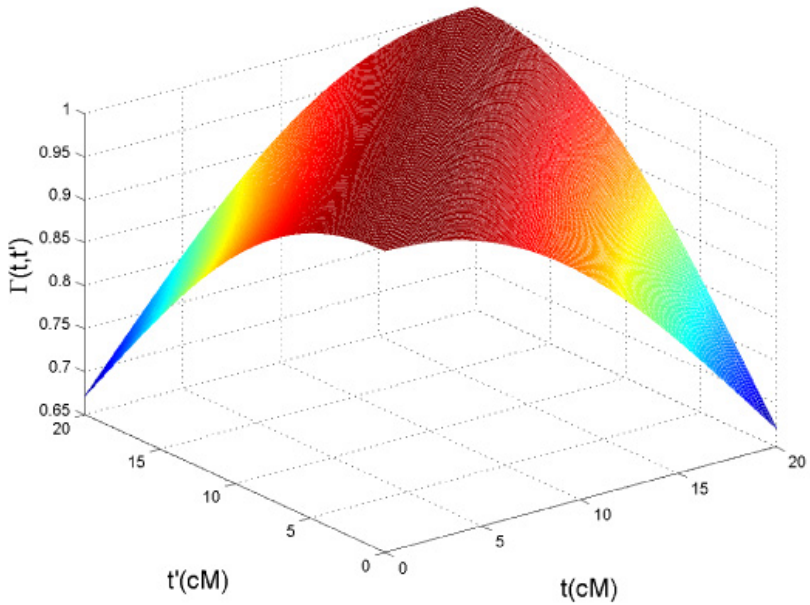
$$\textit{under } H_{at^\star} : \ m_{t^\star}(0) = \frac{a\sqrt{\mathcal{A}}}{\sigma^2}\ \rho(0, t^\star) \quad , \quad m_{t^\star}(T) = \frac{a\sqrt{\mathcal{A}}}{\sigma^2}\ \rho(t^\star, T)$$

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

Mean function and covariance function ($a = 4$, $\sigma = 1$, $K = 2$, $T = 20$cM, $\gamma = 1$)

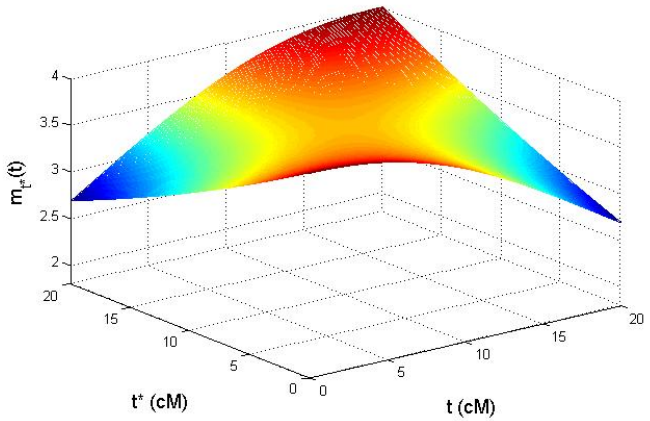


Mean function

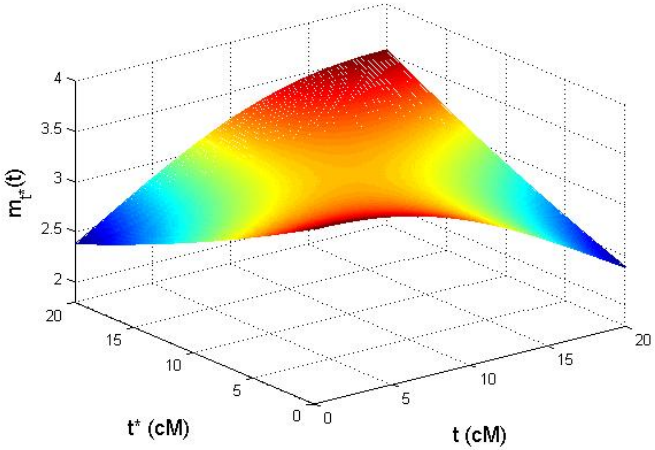


Covariance function

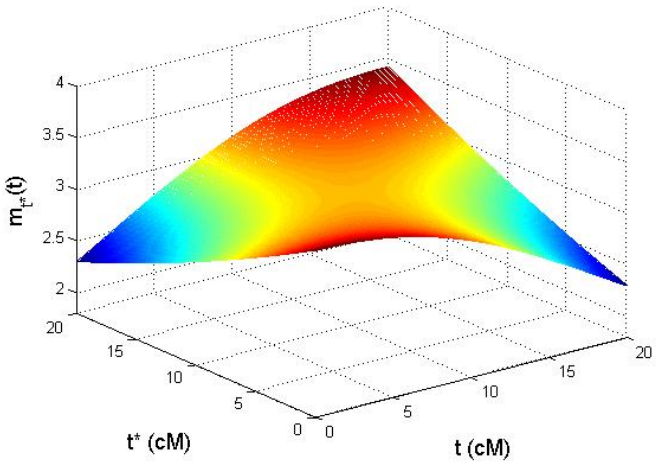
Mean function under selective genotyping ($K = 2$)



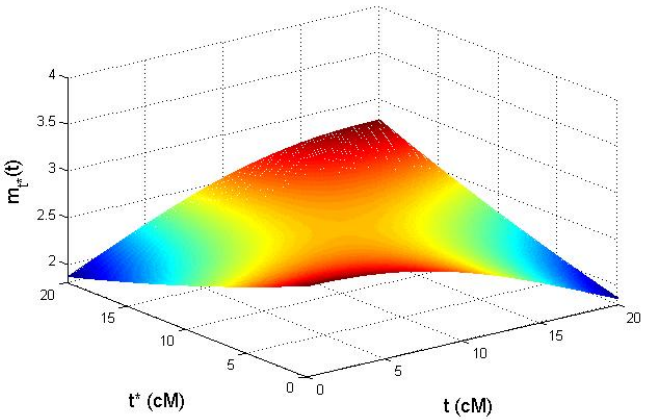
$\gamma = 1$



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



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



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

A non linear interpolation ($K = 2$, $t_1 = 0$, $t_2 = T$)

Lemma

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

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

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

A useful lemma for interpolated processes

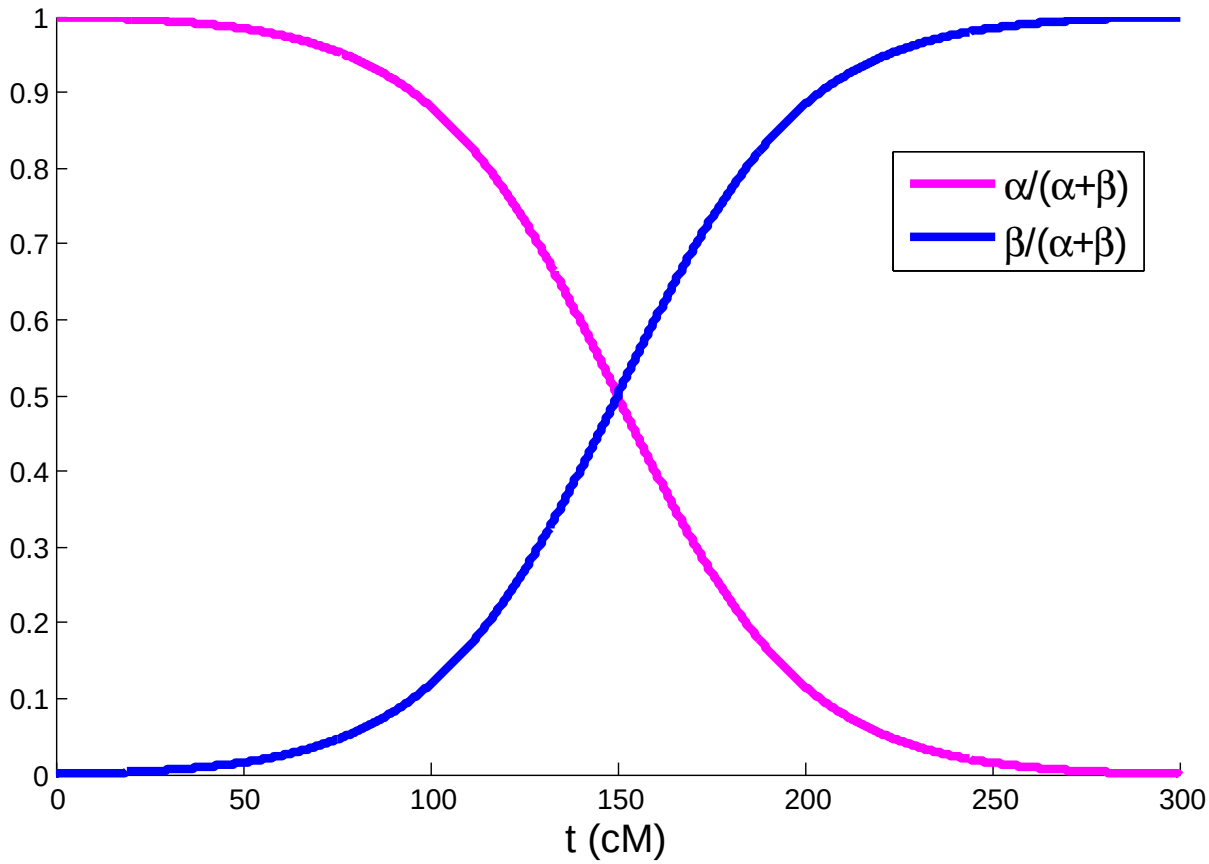
Lemma (Azaïs, Delmas, R., Statistics 2012)

Let $\psi_1(t)$ and $\psi_2(t)$ be two functions such that $\frac{\psi_i(t)}{\psi_1(t)+\psi_2(t)}$ takes every value in $[0,1]$, $i=1,2$. Let C_1 and C_2 be two real numbers and $0<\tilde{\rho}<1$, then

$$\max_{t\in[0,T]} \frac{\{\psi_1(t)C_1+\psi_2(t)C_2\}^2}{\psi_1^2(t)+\psi_2^2(t)+2\,\tilde{\rho}\,\psi_1(t)\,\psi_2(t)}$$

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

A useful lemma for interpolated processes



A useful lemma for interpolated processes ($K = 2$, $t_1 = 0$, $t_2 = T$)

We can apply the lemma by taking

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

Then, we have

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

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	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%
Our method	c	8.41
	$n = 200$	4.76%
	$n = 100$	4.80%
	$n = 50$	3.97%

$T = 1$M, markers equally spaced

$\gamma = 1$, 10000 samples

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

Interference phenomenon (Rebaï et al. 95, 94)

Only one recombination allowed

in each marker interval

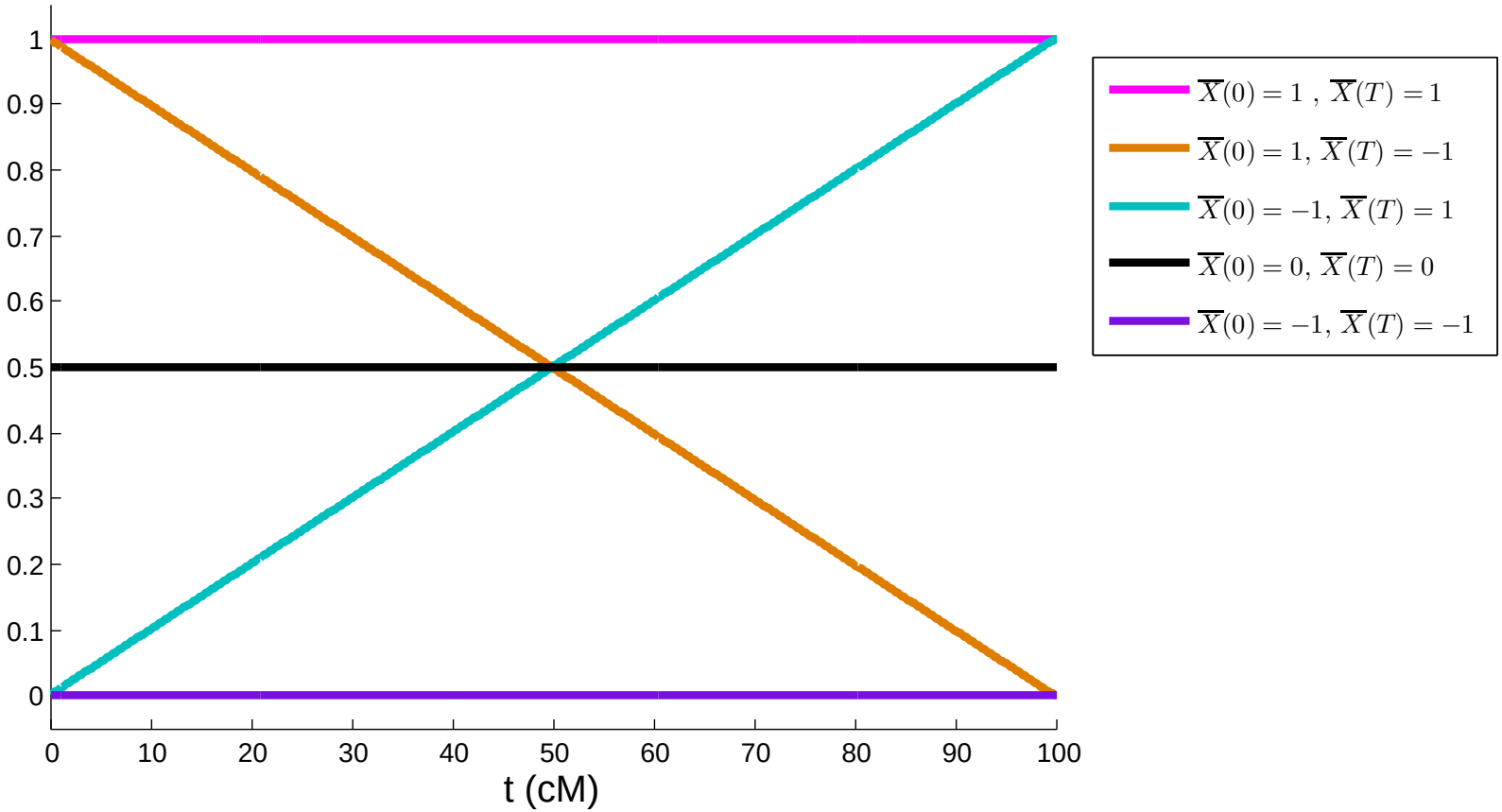
At $t = t^\star$,

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

where

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

Illustration of the different weights ($K = 2$, $T = 1$M)



A linear interpolation

Theorem (R., JSPI 2014)

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

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

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

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

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

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

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

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

Linear interpolation / Non linear interpolation

Lemma

Under the null hypothesis H_0,

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

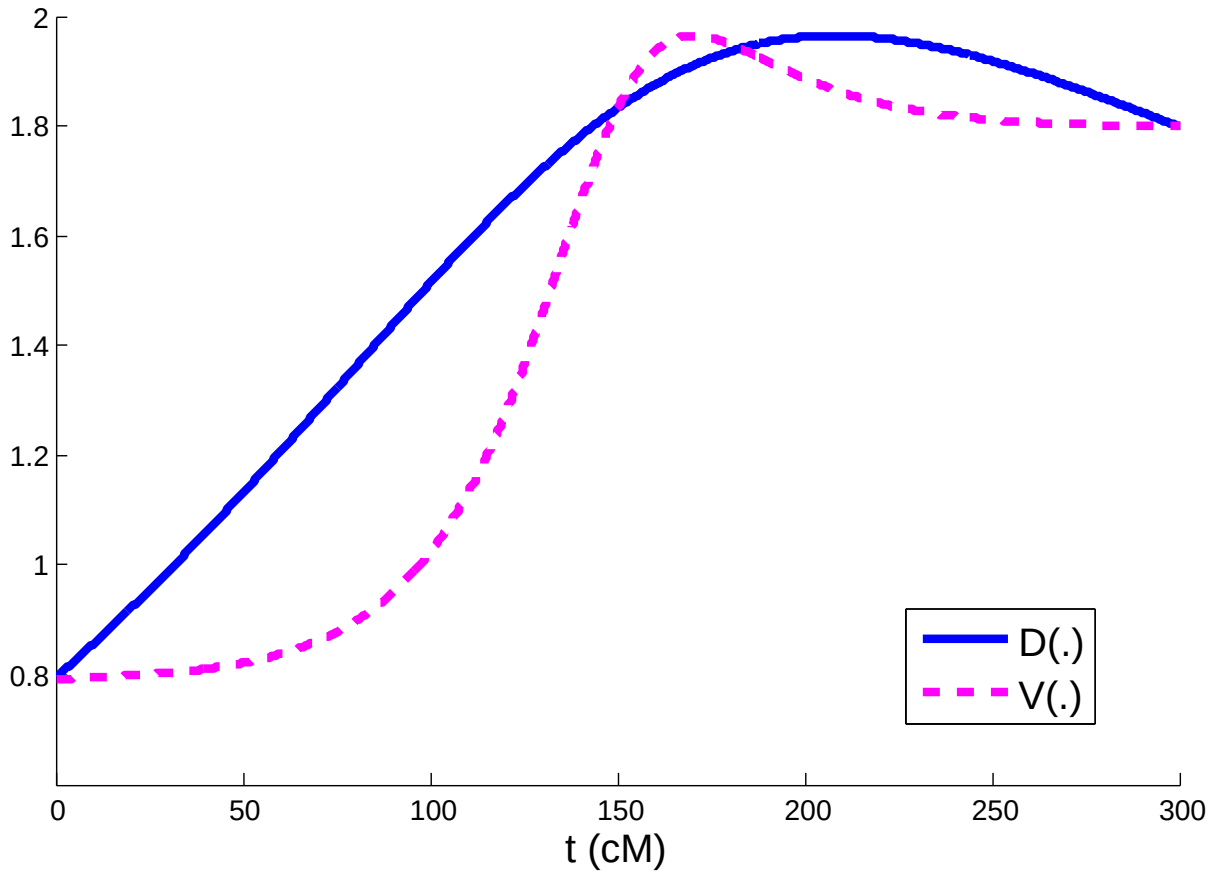
where $V(.)$ is the non linear interpolated process, obtained under Haldane.

The threshold is the same
under Haldane and under the interference model ! ! !

Be careful, the lemma is not true under the alternative $H_{at^\star}$

Asymptotic robustness of the LRT (R., EJS 2014)
since we scan the genome ....

One path of the processes $D(.)$ and $V(.)$ under H_0 ($K = 2$, $t_1 = 0$, $t_2 = 3M$)



About the argmax of the processes ($K = 2$) when it is obtained between markers

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

Lemma

Under H_0 and $H_{at^\star}$,

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

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

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

Ornstein-Uhlenbeck Chi-Square process (OUCS)

We have the relationship :

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

with $G(.)$ OUCS and $\vec{W}(t) = \begin{pmatrix} W_1(t) \\ \vdots \\ W_l(t) \end{pmatrix}$ brownian motion in dimension l.

Then, for computing critical values, are available :

- Delong (81) and Estrella (2003) tables
- Delong (81) approximative formula when c and T are large

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

- a lower bound, obtained by MCQMC (R. and A. Genz, MCAP 2013)

Third part :

Phylogenetics
Joint work with Cécile Ané and Tram Ta

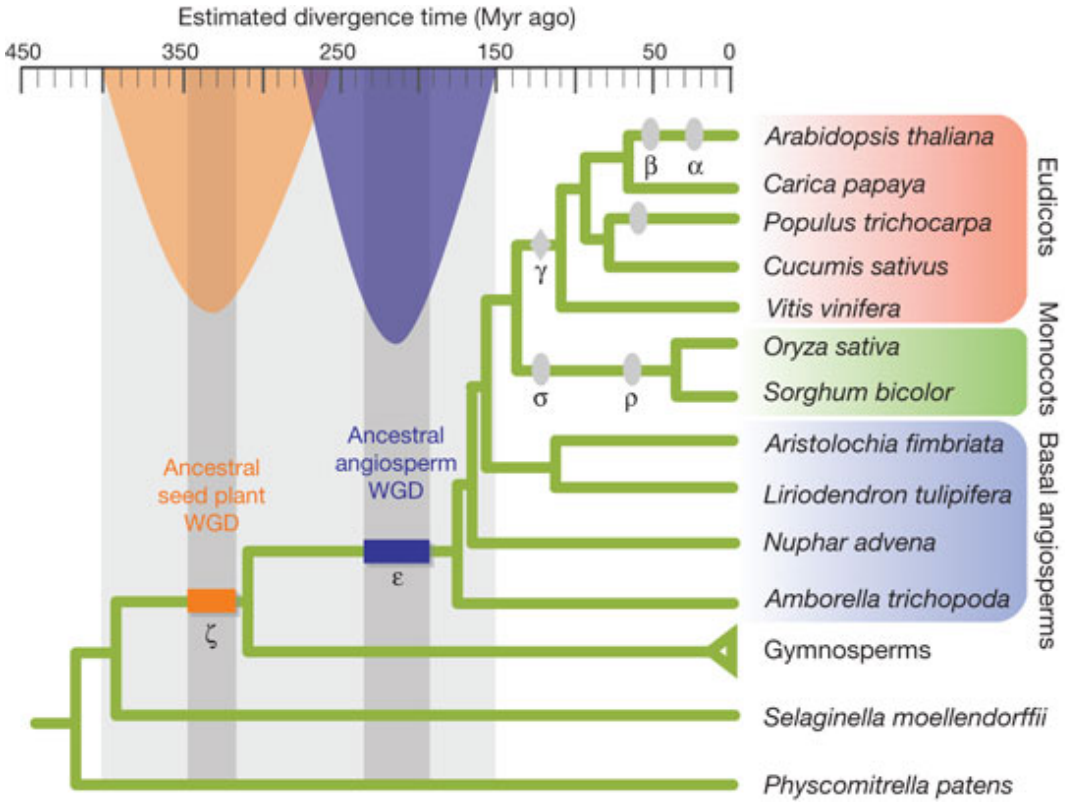
Whole Genome Duplication (WGD)

"Ancestral polyploidy in seed plants and angiosperms", Jiao et al. (Nature 2009)

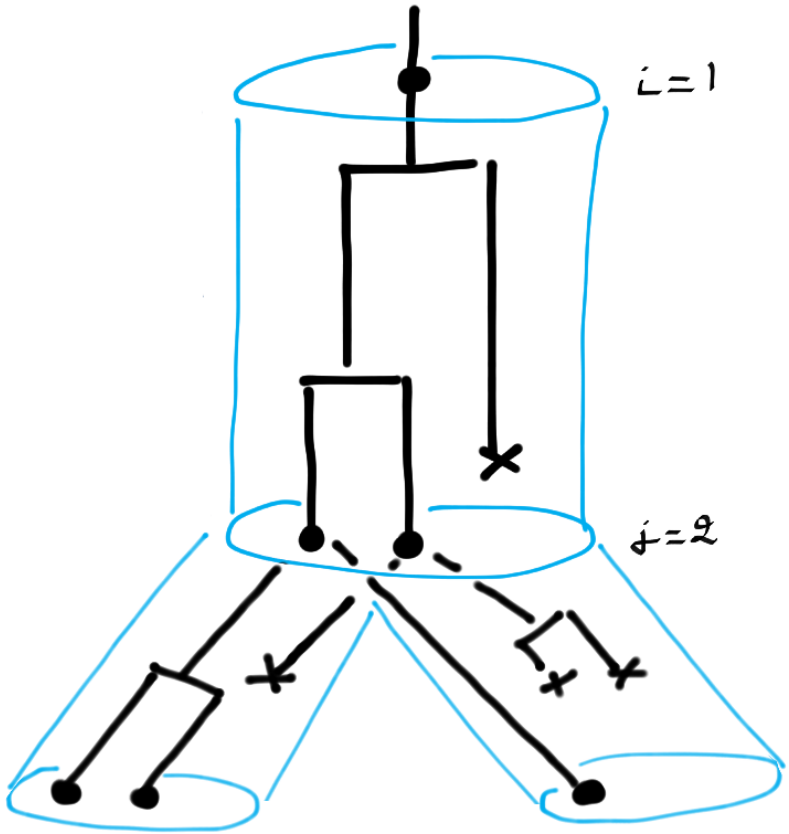
"Whole-genome duplication followed by gene loss and diploidization has long been recognized as an important evolutionary force in animals, fungi and other organisms, especially plants"

WGD in seed plants and angiosperms

Jiao et al. (Nature 2009)



Birth - death process for small scale events

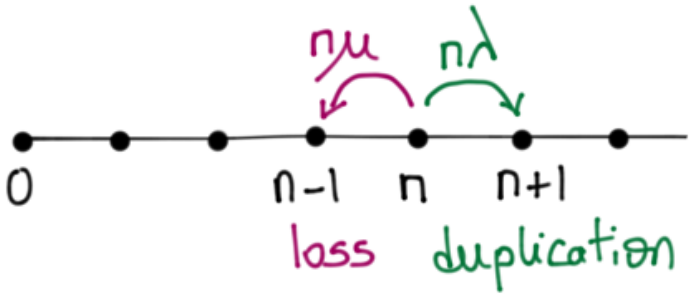


Data
(1 family)

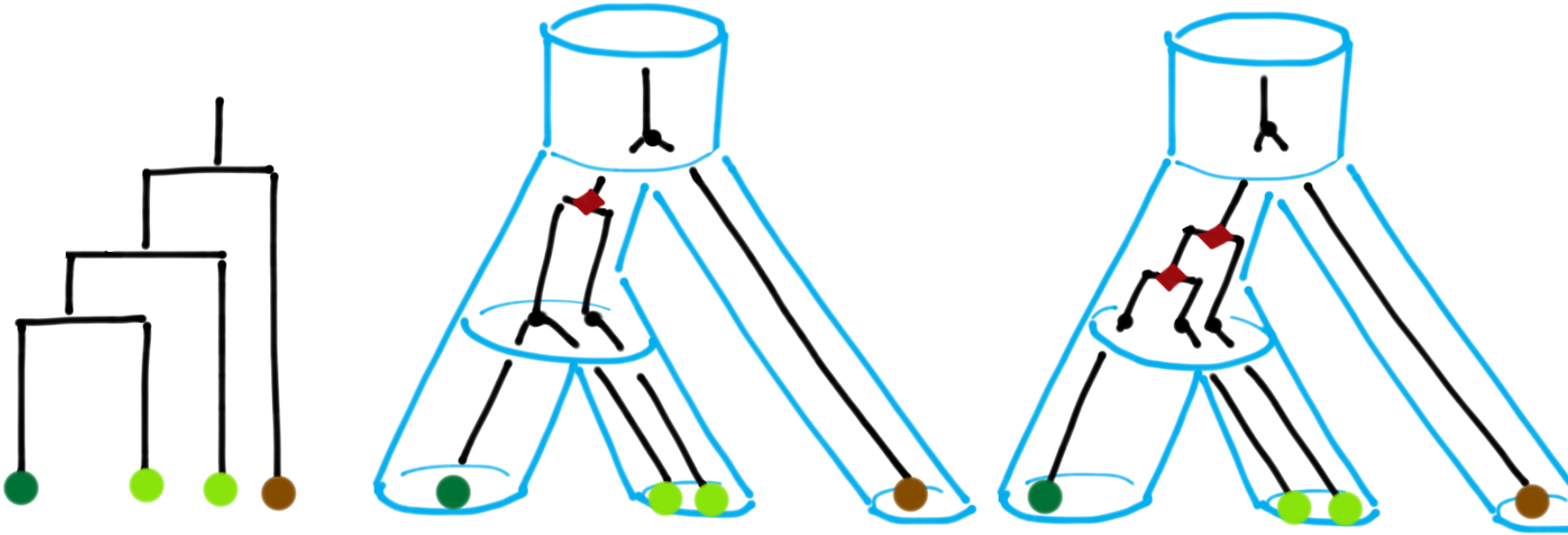
gene counts: X=2 X=1

sequences: a1=AACC b1=AATT
a2=AACT

Birth rate λ, death rate μ



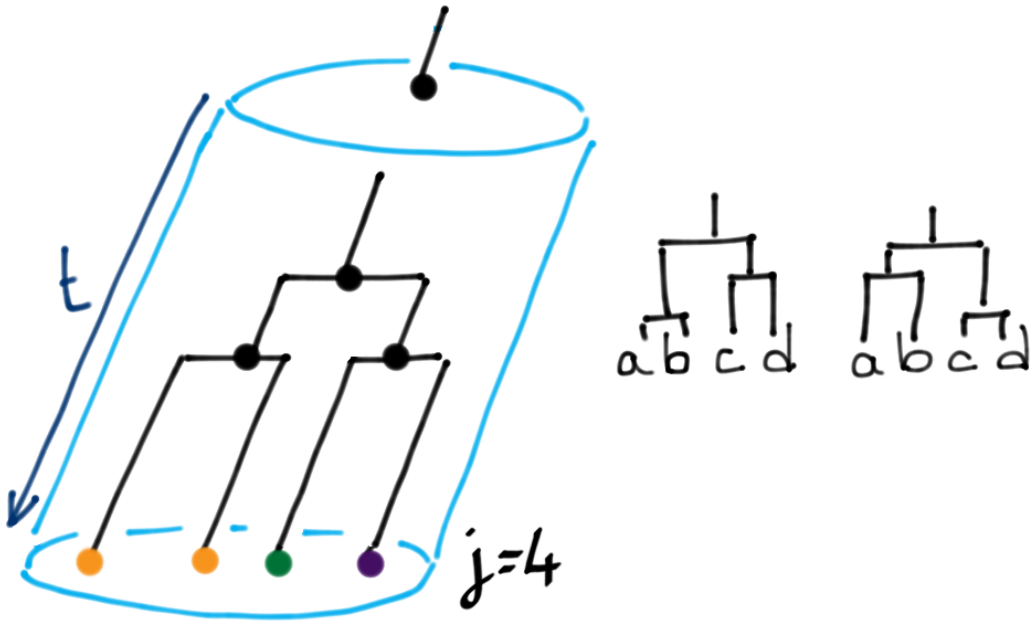
Likelihood of gene tree reconciliations, BD process



Problem 1 : each gene tree has many ”reconciliations” : to map gene tree inside species tree.

Likelihood of gene tree reconciliations, BD process

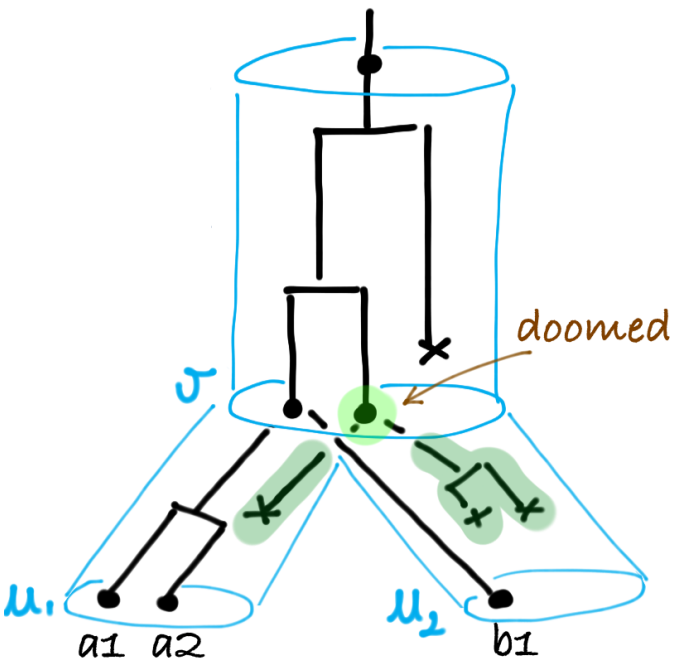
Problem 2 : labels

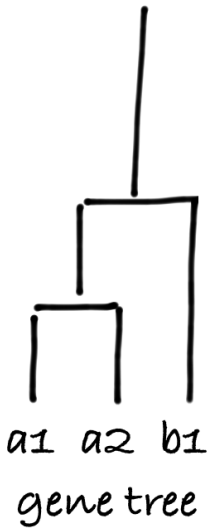


For a reconciled subtree within a ‘slice’, *j* tips, 3 colors

Arvestad et al. (2009), Rasmussen & Kellis (2011)

Likelihood of gene trees reconciliations, BD process



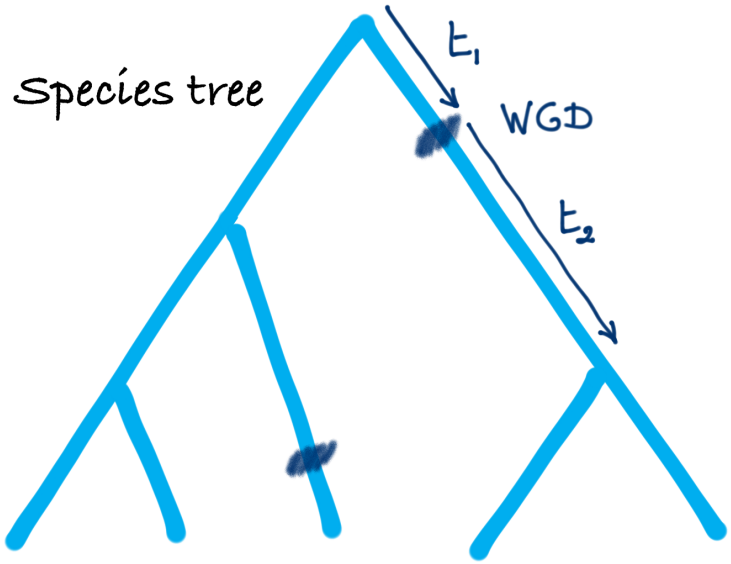


Problem 3 : gene trees lack doomed lineages

d_v : probability that a lineage starting at node v leaves no descendent (or : is doomed). Recursively :

$$d_v = \left(\sum_j p_{t_1}(1,j) d_{u_1}^j \right) \left(\sum_j p_{t_2}(1,j) d_{u_2}^j \right)$$

WGD model for large-scale events



At the WGD :

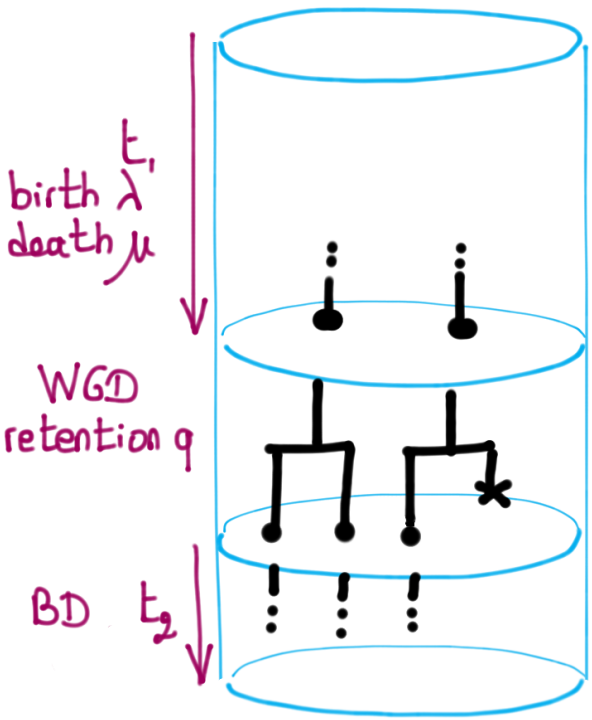
- each gene is duplicated
- second copy lost immediately with probability $1 - q$.

Each WGD has its own retention rate q, to explain :

- Large-scale events
- fragmentation : tendency to lose the extra copy, increased background loss rate shortly after WGD
- extension to whole genome triplications

R., Ta, Ané (2014)

Likelihood : birth-death + WGD model



Same recursive algorithm through the tree, but new transition probabilities along WGD edges :

$$p_{\mathrm{WGD}}(i,j) = \binom{i}{j-i} q^{j-i}(1-q)^{2i-j}$$

$$(i \leq j \leq 2i)$$

Two methods to detect WGDs

Using gene counts only :

- **fast** ($<$ 1s)
- exact likelihood
- optimize λ, μ and separate q's at each WGD
- but : **limited** information

R package `WGDgc`

Using full sequences :

- **rich** information and model, but
- **slow** (e.g. 1h/family) : integrate over tree, reconciliation, branch lengths (gene-specific and species-specific rates).
- approximate likelihood : search over gene trees, but most parsimonious reconciliation, new algorithm to find MP reconciliation with WGDs
- fixed $\hat{\lambda}, \hat{\mu}$

C++ program `spimapWGD`, based on `SPIMAP` (Rasmussen & Kellis 2011)

If you are interested in the gene tree ...

Some notations

- S : species tree
- D : data (ie. alignment data)
- T : gene tree topology
- ℓ : branch length
- R : reconciliation

Bayesian framework

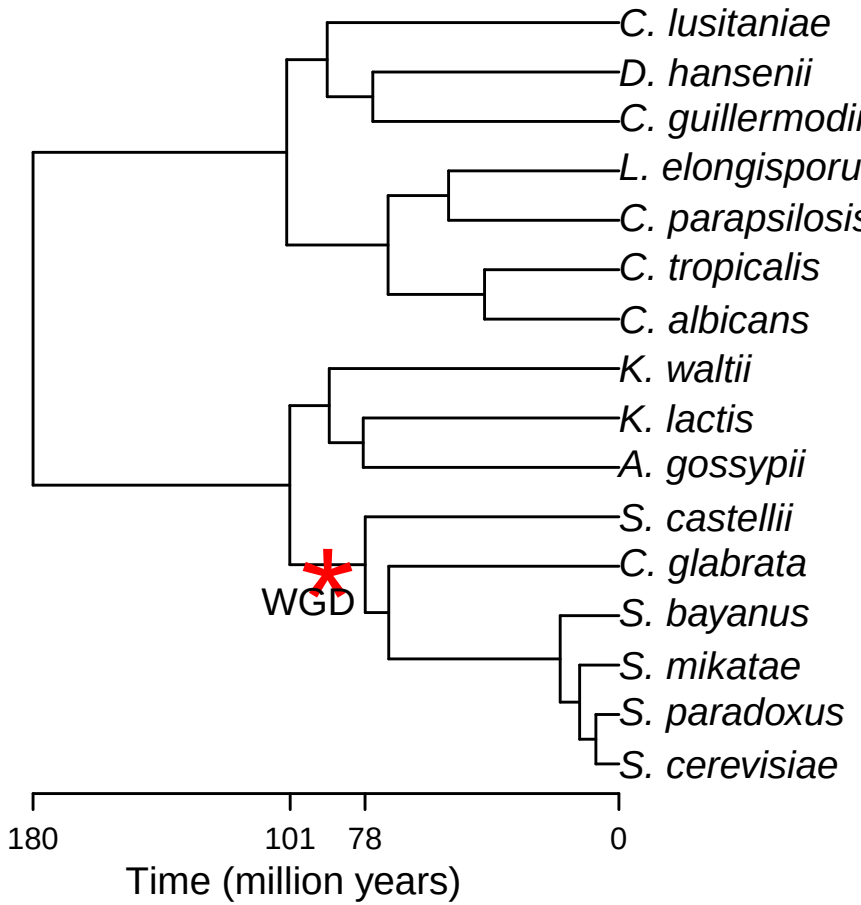
- $\mathbb{P}(T, R|S)$: topology prior
- $\mathbb{P}(\ell|T, R, S)$: branch length prior
- $\mathbb{P}(T, R, \ell|D, S)$: posterior

$\Rightarrow$ Markov Chain Monte Carlo (Hasting Metropolis) to estimate posterior distribution $\mathbb{P}(T, R, \ell|D, S)$

Yeast genome evolution

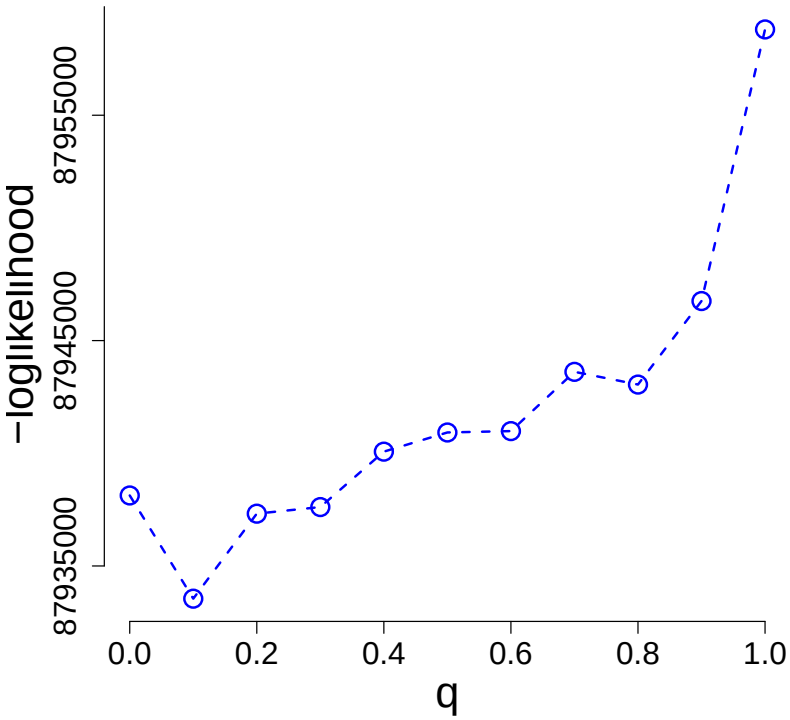
Kellis et al. (2004), from synteny on *Kluyveromyces waltii* and *S. cerevisiae* : "12% of the paralogous gene pairs were retained in each doubly conserved synteny block"

- 9209 gene families (Butler et al 2009)
- filter : 3932 families with ≥ 1 gene in both *Candida* and *Saccharomyces* subclades



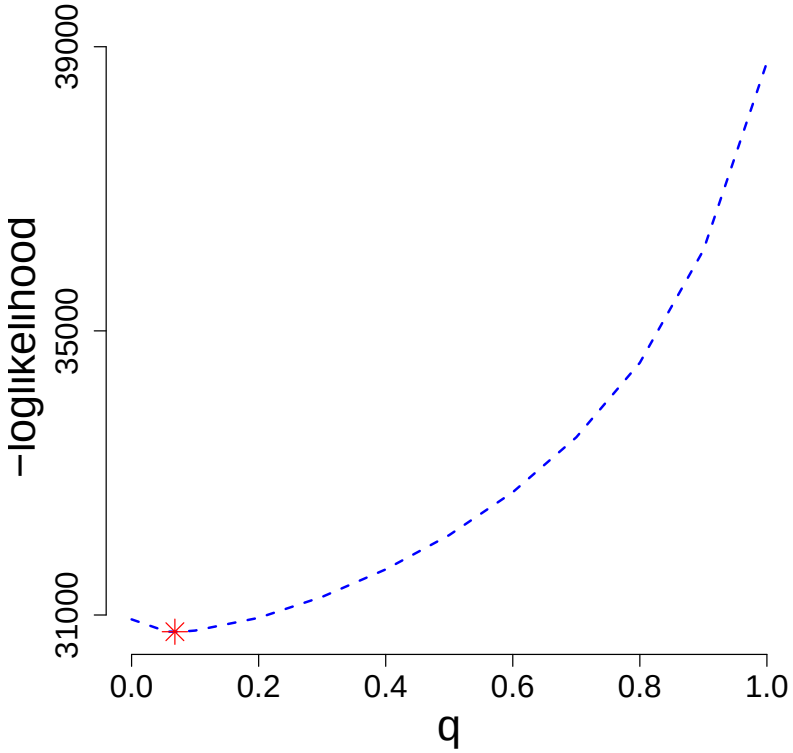
Testing the Yeast WGD

from sequences



LRT : 9159.5

from gene counts



LRT : 348.1

retention rate : $\hat{q} = 6.81\%$, in $[0.058, 0.079]$ with 95% confidence

Thanks to

Cécile Ané
Tram Ta
Jean-Marc Azaïs
Céline Delmas
Jean-Michel Elsen
Alan Genz

Matt Rasmussen
Bill Taylor

DEB-0949121