Chi-square processes for gene mapping in a population with family structure

Charles-Elie Rabier, Jean-Marc Azaïs, Jean Michel Elsen, Céline Delmas

Noname manuscript No.
(will be inserted by the editor)

Chi-square processes for gene mapping in a population with family structure

Charles-Elie Rabier · **Jean-Marc Azaïs** ·
Jean-Michel Elsen · **Céline Delmas**

Received: date / Accepted: date

Abstract Detecting a Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured), on a given chromosome is a major problem in Genetics. We study a population structured in families and we assume that the QTL location is the same for all the families. We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL on the interval $[0,T]$ representing a chromosome. We give the asymptotic distribution of the LRT process under the null hypothesis that there is no QTL in any families and under local alternative with a QTL at $t^\star \in [0,T]$ in at least one family. We show that the LRT is asymptotically the supremum of the sum of the square of independent interpolated Gaussian processes. The number of processes corresponds to the number of families. We propose several new methods to compute critical values for QTL detection. Since all these methods rely on asymptotic results, the validity of the asymptotic assumption is checked using simulated data. Finally we show how to optimize the QTL detecting process.

Keywords Chi-square process · Gaussian process · Likelihood Ratio Test · Mixture models · QTL detection · MCQMC

Mathematics Subject Classification (2000) 62M86 · 65C05 · 62P10

Charles-Elie Rabier
INRA, UR 875 MIAT, BP 52627, Castanet-Tolosan, 31326 Cedex, France
Tel.: +33-5-61285745
Fax: +33-5-61285335
E-mail: cerabier@insa-toulouse.fr, ce.rabier@gmail.com

Jean-Marc Azaïs
Institut de Mathématiques de Toulouse, CNRS UMR 5219, Université Paul Sabatier, 118 route de Narbonne, Toulouse, 31062, France

Jean-Michel Elsen
INRA, UMR 1388 GenPhySE, BP 52627, Castanet Tolosan, 31326 Cedex, France

Céline Delmas
INRA, UR 875 MIAT, BP 52627, Castanet-Tolosan, 31326 Cedex, France

1 Introduction

Detecting a Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured), on a given chromosome is a major problem in Genetics. For example, [Li et al. 2006] detected QTL responsible for reduction of grain shattering in cultivated rice, [Frary et al. 2000] highlighted the presence of a QTL responsible for tomato fruit size, [Silva et al. 2011a] and [Silva et al. 2011b] looked for QTL affecting lactose in Brazilian Gir dairy cattle. In this paper, we study a population structured in families and we assume that the QTL location is the same for all the families. Each family is a set of offsprings from one sire. The problem is that a QTL can be detected in one family if and only if the sire is heterozygous at the QTL. As a result, geneticists focus on a few families. The individual belongs to a family labeled by $i \in \{1,...,I\}$ and its random label $C \in \{1,...,I\}$ is distributed according to a multivariate distribution, i.e.

$$\left\{ \mathbb{P}(C=i) = \pi_i, i = 1,...,I; \sum_{i=1}^{I} \pi_i = 1 \right\}.$$

The chromosome will be represented by the segment $[0,T]$. The distance on $[0,T]$ is called the genetic distance, it is measured in Morgans. A so-called "genome information" at location t is denoted $X(t)$ which takes values in $\{-1,1\}$. The admitted model for the stochastic structure of $X(.)$ is due to [Haldane 1919] which states that :

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

where for any $a \in \mathbb{R}$, δ_a denotes the point mass at a and $N(.)$ is a standard Poisson process on the interval $[0,T]$. In a more practical point of view, this model assumes no crossover interference and the Poisson process represents the number of crossovers on $[0,T]$ which happen during meiosis.
Let us denote by Y, the so-called Quantitative Trait random variable. The stochastic model, associated to Y is defined by

$$Y = \mu_i \; + \; X(t^\star) \; q_i \; + \; \sigma\varepsilon \;\; , \;\; \text{if } C = i \; , \tag{1}$$

where μ_i and q_i are respectively the polygenic and QTL effects within family i, and ε is a standard normal random variable. $t^\star$ is the true location of the QTL. Recall that the location $t^\star$ of the QTL is the same for all the families.
In fact the "genome information" will be available only at certain fixed locations called "markers" $t_1 = 0 < t_2 < ... < t_K = T$ and the observation will be

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

Our dataset $(Y_j, \; X_j(t_1), \; ..., \; X_j(t_K), \; C)_{j=1,...,n}$ is supposed to be obtained by collecting n independent and identically distributed observations (i.i.d.) copies of the random vector $(Y, \; X(t_1), \; ..., \; X(t_K), \; C)$. A so-called Haldane's function denoted by r is considered. This function, from $[0,T]^2$ into $[0,1/2]$ is defined as follows:

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

with the convention $\bar{r}(t,t') = 1 - r(t,t')$. For each $(t,t') \in [0,T]^2$ the quantity $r(t,t')$ represents the probability of recombination of two loci located at t and t'.
It can be proved that, conditionally on $X(t_1),\ldots,X(t_K)$ and C, Y obeys to the following mixture model with known weights :

$$p(t^\star) f_{(\mu_i+q_i,\sigma)}(.) + \{1-p(t^\star)\} f_{(\mu_i-q_i,\sigma)}(.) \ \ , \ \text{if } C = i \ , \tag{2}$$

where $f_{(m,\sigma)}$ is the Gaussian density with parameters (m,σ) and where the function $p(t^\star)$ is the probability that $X(t^\star) = 1$ conditionally on the flanking markers. It can be expressed from the functions r and $\bar{r}$, see Sections 2 and 3.
$\Lambda_n(t)$ will denote the likelihood ratio test (LRT) statistic, at location t (see Section 2 for a precise definition) of the null hypothesis $\{q_i = 0, i = 1,...,I\}$ (i.e. no QTL in any family). The challenge is that the true location $t^\star$ is not known. As a result, at each location $t \in [0,T]$, the presence of a QTL is tested and considering the maximum of $\Lambda_n(.)$ gives the LRT of $\{q_i = 0, i = 1,...,I\}$ on the full chromosome. Note that $\arg\sup \Lambda_n(t)$ is a natural estimator for the QTL location.
Some theoretical results about the LRT process and using approximations, are present, in [Rebaï et al. 1995], [Rebaï et al. 1994], [Cierco 1998], [Azaïs and Cierco-Ayrolles 2002], [Azaïs and Wschebor 2009], [Chang et al. 2009]. In [Azaïs et al. 2014], the focus is on the exact model. However these papers deal with only one family ($I = 1$). In practice, geneticists look for the QTL not in one family but simultaneously in several families, each of them defined by a different sire. This design is called daughter design [Weller et al. 1990]. Since a QTL can be detected in one family if and only if the sire is heterozygous at the QTL, considering a few families increases the chances to study families whose sires are heterozygous at the QTL. As a result, in this paper, we address the problem of the asymptotic distribution of the LRT process when a few families are considered ($I \geqslant 1$). Our main result (Theorem 1 and 2) is that the distribution of the LRT statistic is asymptotically that of the maximum of the square of I independent and "non linear interpolated Gaussian processes". Then, using our theoretical results, we are able to propose methods, as a function of the genetic map, to compute thresholds (i.e. critical values) for QTL detection. Since all these methods rely on asymptotic results, the validity of the asymptotic assumption is checked using simulated data. Moreover, we show how to optimize the detecting process by comparing performances of a global test and a multiple testing procedure. Our methods are available in a Matlab package with graphical user interface : "imfamily.zip". It can be downloaded at http://charles-elie.rabier.pagesperso-orange.fr/doc/articles.html . These methods are alternatives to permutation methods (e.g. [Jung et al. 2007]), generally used in genetics, that enable to compute empirically the distribution of the maximum of the process ([Churchill and Doerge 1994]). Our methods present the advantage to be largely faster than permutation methods. We will also show on simulated data that the mixture model approach is more rewarding than the linearized likelihood approach [Haley 1992] which is very popular in our research field.
We refer to the book of [Van der Vaart 1998] for elements of asymptotic statistics used in proofs. We also refer to [Weller et al. 1990], [Siegmund and Yakir 2007], [Wu et al. 2007] for some genetic background and to [Ron et al. 2001], [Chen et al. 2006], [Weller et al. 2008] for the application field of our study. Typically, the study of

[Silva et al. 2011a], [Silva et al. 2011b] on QTL affecting lactose in Brazilian Gir dairy cattle is an example of application ($K = 27$, $I = 14$, $n = 657$, $\max \pi_i = 0.19$), based on a permutation threshold and on a linearized likelihood.

Our paper ends with an illustration inspired from human real data (Phase 3 release 2 of the HapMap Project). Although the daughter design is not realistic in humans, it is always interesting to use patterns from real data. For instance, as in animal data, human data present the problem of informativeness of genetic markers. These human data, which present a high density of markers (K=75,245), can be analyzed easily, with the help of our interpolation described in Theorems 1 and 2. Besides, we were able to recover, on simulated data, the QTL linked to human height on chromosome 7, highlighted by [Gudbjartsson 2008] in a Genome-Wide Association Study (GWAS).

2 Main results: two genetic markers

To begin, we suppose that there are only two markers ($K = 2$) located at 0 and T : $0 = t_1 < t_2 = T$. For $t \in [t_1, t_2]$ we define

$$p(t) = \mathbb{P}\left\{X(t) = 1 \middle| X(t_1), X(t_2)\right\}$$

and

$$x(t) = \mathbb{E}\left\{X(t) \middle| X(t_1), X(t_2)\right\} = 2p(t) - 1.$$

It is clear that $p(t^\star)$ is the probability appearing in (2). An application of the rule of total probabilities leads to

$$\begin{aligned} p(t) = {} & Q_t^{1,1}\ 1_{X(t_1)=1} 1_{X(t_2)=1} \ +\ Q_t^{1,-1}\ 1_{X(t_1)=1} 1_{X(t_2)=-1} \\ & + Q_t^{-1,1}\ 1_{X(t_1)=-1} 1_{X(t_2)=1} \ +\ Q_t^{-1,-1}\ 1_{X(t_1)=-1} 1_{X(t_2)=-1} \end{aligned} \tag{3}$$

where

$$\begin{aligned} Q_t^{1,1} &= \frac{\bar{r}(t_1,t)\ \bar{r}(t,t_2)}{\bar{r}(t_1,t_2)} \quad , \quad Q_t^{1,-1} = \frac{\bar{r}(t_1,t)\ r(t,t_2)}{r(t_1,t_2)} \\ Q_t^{-1,1} &= \frac{r(t_1,t)\ \bar{r}(t,t_2)}{r(t_1,t_2)} \quad , \quad Q_t^{-1,-1} = \frac{r(t_1,t)\ r(t,t_2)}{\bar{r}(t_1,t_2)}. \end{aligned}$$

We can remark that we have

$$Q_t^{-1,-1} = 1 - Q_t^{1,1} \quad \text{and} \quad Q_t^{-1,1} = 1 - Q_t^{1,-1}.$$

Let $\theta = (q_1,\ ...,\ q_I,\ \mu_1,\ ...,\ \mu_I,\ \sigma)$ be the parameter of the model at t fixed and $\theta_0 = (0,\ ...,\ 0,\ \mu_1,\ ...,\ \mu_I,\ \sigma)$ the true value of the parameter under H_0. The likelihood of the triplet $(Y,\ X(t_1),\ X(t_2),\ C)$ with respect to the measure $\lambda \otimes N \otimes N \otimes N$, λ being the Lebesgue measure, N the county measure on $\mathbb{N}$, is at a position t :

$$L_t(\theta) = \sum_{i=1}^{I} \left[p(t) f_{(\mu_i+q_i,\sigma)}(Y) + \{1-p(t)\} f_{(\mu_i-q_i,\sigma)}(Y)\right]\ 1_{C=i}\ \ g_i(t)$$

where

$$g_i(t) = \frac{\pi_i}{2}\left\{\bar{r}(t_1,t_2)\ 1_{X(t_1)=1}1_{X(t_2)=1}\ +\ r(t_1,t_2)\ 1_{X(t_1)=1}1_{X(t_2)=-1}\right\}$$
$$+\frac{\pi_i}{2}\left\{r(t_1,t_2)\ 1_{X(t_1)=-1}1_{X(t_2)=1}\ +\ \bar{r}(t_1,t_2)\ 1_{X(t_1)=-1}1_{X(t_2)=-1}\right\}\ .$$

Note that the notation $g_i(t)$ will be useful in the generalization to several markers (Section 3). In what follows, $l_t(\theta)$ will be the loglikelihood. We first compute the Fisher information at a point θ_0 that belongs to H_0 :

$$\left.\frac{\partial l_t}{\partial q_i}\right|_{\theta_0} = \frac{Y-\mu_i}{\sigma^2}\ x(t)\ 1_{C=i}\ , \tag{4}$$

$$\left.\frac{\partial l_t}{\partial \mu_i}\right|_{\theta_0} = \frac{Y-\mu_i}{\sigma^2}\ 1_{C=i}\quad , \qquad \left.\frac{\partial l_t}{\partial \sigma}\right|_{\theta_0} = -\frac{1}{\sigma}\ +\ \sum_{i=1}^{I}\frac{(Y-\mu_i)^2}{\sigma^3}\ .$$

After some calculations, we find

$$I_{\theta_0} = Diag\left[\frac{\pi_1}{\sigma^2}\ \ \mathbb{E}\left\{x^2(t)\right\},\ ...,\ \frac{\pi_I}{\sigma^2}\ \ \mathbb{E}\left\{x^2(t)\right\},\ \frac{\pi_1}{\sigma^2}\ ,\ ...\ ,\ \frac{\pi_I}{\sigma^2}\ ,\ \frac{2}{\sigma^2}\right]\ . \tag{5}$$

Before introducing our main theorem, let us define the LRT statistic and the alternative hypothesis. The LRT at t, for n independent observations, will be defined as

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

where $\widehat{\theta}$ is the maximum likelihood estimator (MLE), and $\widehat{\theta}_{|H_0}$ the MLE under H_0. On the other hand, in order to define the alternative hypothesis (so-called $H_{\lambda t^\star}$), the location $t^\star$ of the QTL has to be added in the definition. The alternative hypothesis will be the following :

$H_{\lambda t^\star}$: "there is a QTL at the position $t^\star$ in at least one family".

Besides, in order to deal with Le Cam (1986)'s theory, we will consider local alternatives.

Theorem 1 *Suppose that the parameters* $(q_1,\ ...,\ q_I,\ \mu_1,\ ...,\ \mu_I,\ \sigma)$ *vary in a compact and that* σ *is bounded away from zero. Let* H_0 *be the null hypothesis* $\{q_i = 0, i = 1,...,I\}$ *and define the following local alternative*

$H_{\lambda t^\star}$: *"there is at least one* $q_i = \lambda_i/\sqrt{n}$*, with* $\lambda_i \in \mathbb{R}^\star$*, at the position* $t^\star$*".*

With the previous defined notations,

$$\Lambda_n(.) \xrightarrow{F.d.} \sum_{i=1}^{I}\left\{Z^i(.)\right\}^2\quad , \quad \sup\Lambda_n(.) \xrightarrow{\mathscr{L}} \sup\sum_{i=1}^{I}\left\{Z^i(.)\right\}^2$$

as n tends to infinity, under H_0 *and* $H_{\lambda t^\star}$ *where :*

- $\overset{F.d.}{\rightarrow}$ *is the convergence of finite-dimensional distributions,* $\overset{\mathscr{L}}{\longrightarrow}$ *is the convergence in distribution*
- *the* $Z^i(.)$ *are independent Gaussian processes with unit variance*
- $Z^i(.)$ *is the the continuous and the non linear interpolated process such as :*

$$Z^i(t) = \frac{\alpha(t)Z^i(t_1) + \beta(t)Z^i(t_2)}{\sqrt{\mathbb{V}\{\alpha(t)Z^i(t_1) + \beta(t)Z^i(t_2)\}}} \tag{6}$$

where

$$\mathrm{Cov}\left\{\mathrm{Z^i(t_1), Z^i(t_2)}\right\} = \rho(\mathrm{t_1, t_2}) \quad , \quad \rho(\mathrm{t_1, t_2}) = \exp(-2|\mathrm{t_1 - t_2}|) \quad ,$$
$$\alpha(t) = Q_t^{1,1} - Q_t^{-1,1} \quad , \quad \beta(t) = Q_t^{1,1} - Q_t^{1,-1}$$

and with expectation :

- *under* H_0, $m(t) = 0$
- *under* $H_{\lambda t^\star}$

$$m^i_{t^\star}(t) = \frac{\alpha(t)\ m^i_{t^\star}(t_1) + \beta(t)\ m^i_{t^\star}(t_2)}{\sqrt{\mathbb{V}\{\alpha(t)Z^i(t_1) + \beta(t)Z^i(t_2)\}}}$$

where

$$m^i_{t^\star}(t_1) = \frac{\lambda_i\ \sqrt{\pi_i}\ \rho(t_1, t^\star)}{\sigma} \quad , \quad m^i_{t^\star}(t_2) = \frac{\lambda_i\ \sqrt{\pi_i}\ \rho(t^\star, t_2)}{\sigma}\ .$$

The proof of Theorem 1 is given in Appendix 1. Let us recall the definition of a Chi-square process from [Davies1987].

Definition 1 A Chi-square process $W(.)$ with d degrees of freedom is a process such as:

$$W(t) = V_1(t)^2 + \cdots + V_d(t)^2 \tag{7}$$

where the $V_i(t)$ are independent for each t and distributed as a standardized Normal under the null hypothesis.

As a consequence, the limiting process $\sum_{i=1}^{I}\left\{Z^i(.)\right\}^2$ of Theorem 1 is a Chi-Square process with I degrees of freedom where the $Z^i(.)$ are independent and identically distributed (a particular case of formula 7).

Note that Theorem 1 could easily be generalized to selective genotyping experiments, that allow to reduce genotyping costs, by genotyping only extreme individuals. In other words, under selective genotyping, the genome information at markers $X(t_1), ..., X(t_K)$ is available for one individual, if and only if $Y \geqslant S_+$ or $Y \leqslant S_-$, where S_+ and S_- denote two real thresholds (constant). Typically, an additional factor would appear in the mean function $m^i_{t^\star}(.)$ of the processes $Z^i(.)$ introduced in Theorem 1 (see [Rabier 2015] for more details).

On the other hand, Theorem 1 could also be adapted to the interference model, where contrary to Haldane, crossovers do not occur independently from each others. In that case, the functions $\alpha(.)$ and $\beta(.)$ would be linear functions (see [Rabier 2014]).

3 Several markers : the "Interval Mapping" of [Lander and Botstein 1989]

In that case suppose that there are K markers $0 = t_1 < t_2 < ... < t_K = T$. We consider values t, t' or $t^\star$ of the parameters that are distinct of the markers positions, and the result will be prolonged by continuity at the markers positions. For $t \in [t_1, t_K] \backslash \mathbb{T}_K$ where $\mathbb{T}_K = \{t_1, ..., t_K\}$, we define t^ℓ and t^r as :

$$t^\ell = \sup\{t_k \in \mathbb{T}_K : t_k < t\} \quad , \quad t^r = \inf\{t_k \in \mathbb{T}_K : t < t_k\}.$$

In other words, t belongs to the "Marker interval" (t^ℓ, t^r).

Theorem 2 *We have the same result as in Theorem 1, provided that we make some adjustments and that we redefine each process $Z^i(.)$ in the following way :*

- *in the definition of $\alpha(t)$ and $\beta(t)$, t_1 becomes t^ℓ and t_2 becomes t^r*
- *under the null hypothesis, the process $Z^i(.)$ considered at marker positions is the "squeleton" of an Ornstein-Uhlenbeck process: the stationary Gaussian process with covariance $\rho(t_k, t_{k'}) = \exp(-2|t_k - t_{k'}|)$*
- *at the other positions, $Z^i(.)$ is obtained from $Z^i(t^\ell)$ and $Z^i(t^r)$ by interpolation and normalization using the functions $\alpha(t)$ and $\beta(t)$*
- *at the marker positions, the expectation is such as $m^i_{t^\star}(t_k) = \lambda_i \sqrt{\pi_i}\, \rho(t_k, t^\star)/\sigma$*
- *at other positions, the expection is obtained from $m^i_{t^\star}(t^\ell)$ and $m^i_{t^\star}(t^r)$ by interpolation and normalization using the functions $\alpha(t)$ and $\beta(t)$.*

A proof is given in Appendix 2. Note that when the number of genetic markers is infinite, each process $Z^i(.)$ is an Ornstein-Uhlenbeck process. As a consequence, when the number of genetic markers is infinite, $\sum_{i=1}^I \left\{Z^i(.)\right\}^2$ is an Ornstein-Uhlenbeck Chi-Square process with I degrees of freedom (OUCS(I)) since the processes $Z^i(.)$ are independent.

In Figure 1, we consider a chromosome of length $T = 60$cM and 3 families (i.e. $I = 3$). We focus on two genetic maps :

- an infinite number of genetic markers
- only 4 markers located every 20cM.

One path of each asymptotic process is presented in this figure. We can notice that the path of the OUCS(3) is very jerky whereas the path of the process corresponding to the sparse map is smooth due to the interpolation between markers.

4 Different methods to obtain thresholds as a function of the map considered

4.1 Introducing the methods

We propose several new methods, as a function of the map considered, to compute thresholds for the supremum of the LRT process under H_0. In particular, two kinds of maps are studied :

- a sparse map : a few markers covering the chromosome
- a dense map : a high density of markers pretty close to each other.

We will assume that when the map is dense, tests are performed only on markers, whereas when the map is sparse, tests are also performed between markers (cf. exemple 11.3 p. 248 of [Wu et al. 2007]).

Under a sparse map, thresholds can be obtained according to the most appropriate method which depends on the number of families I :

- for $I = 1$, the problem is the same as computing the distribution of the maximum, i-absolute, value of a Gaussian vector. This can be done by a Discrete Monte-Carlo Quasi Monte-Carlo method (DMCQMC). In particular, the method for numerical computation of a multivariate normal probability ([Genz 1992]) can be considered. It uses a transformation that simplifies the problem and places it into a form that allows efficient calculation using MCQMC methods. A simple MC method using N points has errors that are typically $O(1/\sqrt{N})$ whereas a MCQMC method has errors approximatively $O(1/N)$, that's why the focus here is on MCQMC.
- for $I > 1$, a Discrete Monte-Carlo (DMC) method can be performed. According to Theorem 2, when we test only on markers, the asymptotic process is a Discrete Ornstein-Uhlenbeck Chi-Square process with I degrees of freedoms (DOUCS(I)). In this case, the processes $Z^i(.)$ are simply AR(1) processes. Then, in order to obtain values between markers, we can complete by interpolation using formula (6). As a result, the threshold is easily obtained by a DMC method based on a large number of sample paths (denoted $nspaths$) of the asymptotic process.

Under a dense map, we propose theoretical methods to obtain the thresholds. As mentioned previously, when the number of genetic markers is infinite, the LRT process is asymptotically an OUCS(I) process. In [Rabier and Genz 2014], we propose an approximative formula (named DF here) for the threshold of the supremum of the OUCS(I) process. It is based on [Delong 1981]'s work on Brownian motion. This formula is suitable when I and the threshold are large. Besides, statistical tables given by [Estrella 2003], for the threshold of the supremum of the OUCS(I), are also available. Note that, in order to obtain its exact tables, Estrella improved Delong's work on hypergeometrics functions. In the following, Estrella's method will be denoted ET.

Table 1 is a summary of the different methods.

4.2 Applications under the null hypothesis

In this section, the focus is on thresholds corresponding to the 95% quantile of the supremum of the LRT process under H_0. In order to illustrate the different methods, a sparse map and a dense map are considered. Since all the methods are based on asymptotic results (cf. Theorem 1 and 2), populations of different sizes have been simulated in order to check when the asymptotic regime is reached. In what follows, npop will denote the number of populations whereas n is the size of a population.

Sparse map

The sparse map consists of a chromosome of length $T = 60$cM with 4 genetic markers equally spaced every 20cM. The presence of a QTL is tested every 5cM.

In Table 2, thresholds are presented as a function of I. In Table 3, the focus is on the number of false positives (NFP) as a function of the number of individuals n (thresholds given in Table 2). Using the Binomial distribution, a 95% confidence interval is computed (in brackets in the tables) for the true percentage of the number of false positives.

According to Table 3, when there are on average 200 individuals per family (i.e. $n = 200\ I$), NFP is not significantly different from 5%. When $n = 50\ I$, we can consider that NFP is still fair (even if it is significantly different from 5%) whereas when $n = 30\ I$, NFP is not so nominal.

Dense map
The dense map consists of a chromosome of length $T = 50$cM with 501 genetic markers equally spaced every 0.1cM.

The thresholds and the NFP are respectively compared in Tables 4 and 5. This aspect suggests fast convergence to the asymptotic regime.

4.3 Remark

ET is not appropriate for the sparse map for two reasons :

- ET is based on Ornstein-Uhlenbeck (OU) process which is much more irregular than the process $Z^1(.)$ (OU can be viewed as a stationary version of the Brownian motion). When $I = 1$, this can be formalized by the use of Slepian type inequalities, specially lemma 2.1 in [Azaïs and Wschebor 2009] which comes from [Plackett 1984]. It can be proved that the covariances are smaller in the case of OU process than for the process $Z^1(.)$. It implies that the maximum of OU is stochastically greater than the maximum of $Z^1(.)$. Since $\mathbb{P}\left(\sup\left|Z^1(.)\right| > u\right) \approx 2\mathbb{P}\left(\sup Z^1(.) > u\right)$, this argument can be approximatively extended to the absolute value.
- for the sparse map, the focus is not on the continuous process but on the dicrete process : the maximum of a continuous process is always greater than the discrete one.

To sum up, ET will give too large thresholds.

5 Optimization of the QTL detecting process

5.1 Motivation

A few sires are heterozygous at the QTL and others are homozygous. As mentioned previously, a QTL can only be detected in a family defined by an heterozygous sire. Thus, two questions arise :

- is it always profitable to include all the families in the analysis ?

- do we have to analyze families all together or separately ?

We consider here, the sparse map of Section 4.2. As previously, tests are performed every 5 cM. We will consider tests at the 5% significance level.

5.2 About the QTL effects

When we deal with I families, since the total number of individuals is n, the expected number of individuals in family i is only $n\pi_i$. Hence, in order to study the evolution of the power of the Interval Mapping with the number of families, we will consider $\lambda_i = \frac{\lambda}{\sqrt{\pi_i}}$ (note that when $I = 1$, we have $\lambda_1 = \lambda$ since $\pi_1 = 1$). As a consequence, the mean function, $m^i_{t^\star}(t)$, of the asymptotic process $Z^i(.)$, is proportional to λ and does not depend on i (cf. Theorem 1 and 2).

5.3 How to optimize the QTL detecting process

Only asymptotic results are studied here (cf. Theorem 1 and 2). Figures 2, 3, 4 are related to question 1 whereas Figures 5, 6, 7 are related to question 2.
In Figures 2, 3, 4, the power is represented as a function of the QTL location, the number of families and the QTL effects. In Figures 5, 6, 7, we compare the power of the approach which consists in analyzing all families together (as previously), and the power of the approach which consists in analyzing families separately. Note that we used a discretization for the QTL location $t^\star$ (every 5cM).

Figures 2, 3, 4
As expected, when all the sires are heterozygous, the power increases with the number of families (cf. Figure 2 for $\lambda = 2$). Besides, for a given number of families, the power increases with the proportion of heterozygous sires (cf. Figure 3 for $I = 5$, $\lambda = 2$ and various number number nz of non zeros λ_i's).

According to Figure 4, it is almost as powerful to consider only one family whose sire is heterozygous (cf. curve $I = 1$ with $nz = 1$), as to consider 5 families with only two heterozygous sires (cf. curve $I = 5$ with $nz = 2$). As a result, it is much more powerful to consider one family of an heterozygous sire (cf. curve $I = 1$ with $nz = 1$) as to consider 5 families with only one heterozygous sire (cf. curve $I = 5$ with $nz = 1$). Hence, if the families could be sorted in advance, it would be more powerful to concentrate the analysis on the families with a segregating QTL (i.e. families of heterozygous sires). Furthermore, once the families targeted, it would be more powerful to remove the families with very small QTL effects (not illustrated here). Indeed, these families are a source of noise to our model.

Figures 5, 6, 7
In practice, the segregating families are not known before the statistical analysis and the true question is : do we have to analyze all the families together (so-called "global approach") or analyze families separately (so-called "Bonferroni approach") ? Indeed, since our results are asymptotic, the variance is not better estimated when the

global approach is considered.

Figures 5, 6, 7 are related to these two approaches. When the global approach is considered and when H_0 is rejected, it only comes out that there is a QTL in at least one family (i.e. at least one sire is heterozygous), but this family is not known. As a result, in order to answer the same question, we define, for the Bonferroni approach, the test statistic U and the critical region CR, which results from a Bonferroni correction :

$$U = \left(\sup \left\{ Z^1(.) \right\}^2, \ ..., \ \sup \left\{ Z^I(.) \right\}^2 \right) \quad ,$$

$$CR = \left\{ u = (u_1, ..., u_I) \in \mathbb{R}^I \ \text{ such as there is at least one } u_i \text{ verifying } \ u_i \geqslant c \right\} \quad ,$$

where c is the threshold such as $\mathbb{P}\left(\sup \left\{ Z_0^1(.) \right\}^2 \geqslant c \right) = \frac{0.05}{I}$.
$Z_0^1(.)$ is the analogue of the Gaussian process $Z^1(.)$ under the null hypothesis.
The Bonferroni correction allows to have $\mathbb{P}_{H_0}(U \in CR) \leqslant 0.05$. Obviously, the power of the Bonferroni approach is $\mathbb{P}_{H_{\lambda t^\star}}(U \in CR)$. Note that we could have considered other multiple testing procedures (e.g. [Benjamini and Hochberg 1995], [Didelez et al. 2006]).

In Figure 5, the focus is on the particular case for which there is only a QTL in family 1. The power of the two approaches is represented as a function of the QTL location $t^\star$ and the number of families. We can notice that the Bonferroni approach is more powerful than the global approach. In Figure 6, the focus is on the particular case for which there is a QTL in each family. In that case, the Bonferroni approach is outperformed by the global approach.

Figure 7 represents the mean power of the two approaches. Every alternative hypotheses have been considered (i.e. for a given I, we have considered $nz = 1, ..., I$). Equiprobability concerning all these hypotheses has been supposed. According to the figure, for a given number of families, there is a mean increase in terms of power of at least 15% when the global approach is considered.

5.4 Conclusion

It comes out from this study that in order to optimize the QTL detecting process, it is required: :

- to target, whenever possible, families with the largest QTL effects and then, to analyze all these families together.
- when it is not possible to target families, to analyze all the families together.

6 Comparison between different global tests

6.1 Mixture model vs linearized likelihood

Tables 6 and 7 compare on the sparse map, two approaches regarding the data analysis. Note that different number of families have been considered assuming that all

the sires were heterozygous ($\lambda = 2$, $\pi_i = 1/I$). The first approach is the one theoretically studied in this paper. It relies on the mixture model and MLE are computed every $5cM$, with the help of the EM algorithm. The thresholds used are the same as in Table 2. The second approach is the linearized likelihood method [Haley 1992]. It consists in approximating the mixture of formula (2) by only one distribution:

$$f_{(\mu_i+\{2p(t^\star)-1\}q_i,\sigma)}(.) \quad , \quad \text{if } C = i \ .$$

As a consequence, when this approximation is used, analytical formulas (as in a linear model) are available regarding the different estimators. Besides, thresholds are usually obtained from permutation tests, that enable to compute empirically the distribution of the maximum of the process [Churchill and Doerge 1994].

According to Tables 6 and 7, the method based on the mixture model is more powerful in all cases. The largest difference of power is observed for $n = 30I$ (approximately 6%). Note that for $n = 200I$, the approach relying on the mixture model is still slightly more interesting. We will show in Section 8 that a major drawback of the permutation method is that it requires a large amount of time to get computed, which is not the case of our proposed methods.

Last, let us focus on the method based on mixture model: we can notice in Tables 6 and 7 that the Theoretical Power is always located, whatever the value of n, in the 95% confidence interval for the true value of the power. It validates our asymptotic results.

6.2 Interpolated process vs discrete process

We propose to compare here the powers of two statistical tests. The first one is the LRT studied in details in this paper. Recall that it is based on the test statistic:

$$\sup_{t\in[0,T]} \sum_{i=1}^{I} \left\{Z^i(t)\right\}^2$$

where $Z^i(.)$ is the interpolated Gaussian process obtained in Theorems 1 and 2. The second one relies on the test statistic:

$$\max_{k\in\{1,\cdots,K\}} \sum_{i=1}^{I} \left\{Z^i(t_k)\right\}^2$$

where $t_1,\cdots,t_K$ are the markers positions. The aim of this comparison is to quantify the usefulness of the interpolated process in the QTL detection.

We consider a chromosome of size $T = 1$M and two different genetic maps:

- map 1 consists in 6 markers equally spaced every 20cM
- map 2 consists in 51 markers equally spaced every 2cM.

One QTL is present on the chromosome at $t^\star = 0.4$M (on a marker) or $t^\star = 0.5$M (between two markers) for map 1 and at $t^\star = 0.4$M (on a marker) or $t^\star = 0.51$M (between two markers) for map 2. For each test statistic, threshold and power are based on 100,000 paths of the corresponding process under the null hypothesis and

under the alternative of one QTL located at $t^\star$ in all families ($\lambda = 2$, $\pi_i = 1/I$). The power gain is defined as the difference between the power of the LRT and the power of the test relying only on marker locations. A confidence interval for the power gain is given based on 30 simulations of 100,000 paths. According to Table 8, when the QTL is located on a marker, there is no need to use the interpolation results to test for a QTL: the power gain is too low or even slightly negative in some cases. When the QTL is located between the markers, the power gain is all the more important as the map is more sparse; furthermore, when we increase the intensity of the Poisson process (modelling the number of recombinations), we decorrelate the markers and the power gain is more important using the interpolation results.

7 Behavior of the LRT when the main assumptions are violated

The proposed LRT is based on two main assumptions. First, we consider that the QTL location is the same for all the families. Secondly, we assume that there is only one QTL located on $[0,T]$. In this section, we investigate the behavior of the LRT when these assumptions are violated.

7.1 QTL locations are different across families

Let us consider the case where the true QTL location is different across families. In what follows, $t_i^\star$ will denote the QTL location for family i. In this context, the mean function of the process $Z^i(.)$ is still an interpolated function relying on the functions $\alpha(t)$ and $\beta(t)$, except that the quantities $m^i_{t^\star}(t_1)$ and $m^i_{t^\star}(t_2)$, from Theorem 1, are now replaced by the quantities $m^i_{t_i^\star}(t_1)$ and $m^i_{t_i^\star}(t_2)$ defined in the following way:

$$m^i_{t_i^\star}(t_1) = \frac{\lambda_i \sqrt{\pi_i}\, \rho(t_1, t_i^\star)}{\sigma} \quad , \quad m^i_{t_i^\star}(t_2) = \frac{\lambda_i \sqrt{\pi_i}\, \rho(t_i^\star, t_2)}{\sigma} \ .$$

The proof is the same as the proof of Theorem 1, provided that we replace $X(t^\star)$ by $X(t_i^\star)$ in formula (14). In presence of several markers, it can be seen easily that the formula becomes

$$m^i_{t_i^\star}(t^\ell) = \frac{\lambda_i \sqrt{\pi_i}\, \rho(t^\ell, t_i^\star)}{\sigma} \quad , \quad m^i_{t_i^\star}(t^r) = \frac{\lambda_i \sqrt{\pi_i}\, \rho(t_i^\star, t^r)}{\sigma} \ .$$

Tables 9 and 10 focus on the cases $I = 3$ and $I = 5$ respectively. The genetic map considered is the sparse map (introduced in Section 4.2). Table 9 investigates two different scenarios:

1. the QTLs present in the first two families, are located on the same genetic marker ($t_1^\star = t_2^\star = 0.2$)
2. the QTLs present in the first two families, are located at the middle of a marker interval ($t_1^\star = t_2^\star = 0.5$)

Note that for both scenarios, the location $t_3^\star$ of the QTL in the third family is allowed to vary along the genome. According to Table 9, for scenario 1, the Theoretical power reaches its maximum when $t_3^\star$ takes the same value as $t_1^\star$ and $t_2^\star$. In other words, the QTLs have to be located on the same genetic marker, for the three families. However, under scenario 2, the maximum power is reached when $t_3^\star$ is equal to 0.4, which is a different location from the QTL locations in the first two families. This surprising result is due to the fact that the signal is maximum when the QTL is located on a genetic marker, whereas there is a loss of power when the QTL is not located on a genetic marker. Figure 8 illustrates these two scenarios in a noiseless setting (i.e. ideal situation). As expected, for scenario 1, the signal is maximum when $t_1^\star = t_2^\star = t_3^\star = 0.2$. Under scenario 2, we can observe that it is maximum for $t_3^\star = 0.4$ and $t_3^\star = 0.6$. As a result, it is more rewarding to have a QTL located on one genetic marker in one family, than all QTLs located at the middle of a marker interval. Recall that the Theoretical power, computed in Table 9, was obtained by Monte Carlo and in presence of noise in the model. It was maximum for $t_3^\star = 0.4$ under scenario 2.

Last, we obtain the same kind of conclusions when we deal with 5 families (cf. Table 10).

7.2 Several QTLs are lying on the genome

Let us consider now that m QTLs lie on the genome, at locations $t^{\star(1)} < t^{\star(2)} < \ldots < t^{\star(m)}$. In order to make the reading easier, we will assume that the QTLs are located at the same location across families. Let $q_{s,i}$ denote the effect of the s-th QTL in family i and let $\lambda_{s,i}$ denote the constant such as $q_{s,i} = \lambda_{s,i}/\sqrt{n}$.

In this context, the mean function of the process $Z^i(.)$ is still an interpolated function relying on the functions $\alpha(t)$ and $\beta(t)$, except that $m^i_{t^\star}(t_1)$ and $m^i_{t^\star}(t_2)$ are now replaced by the quantities $m^i_{\tilde{t}^\star}(t_1)$ and $m^i_{\tilde{t}^\star}(t_2)$ defined in the following way:

$$m^i_{\tilde{t}^\star}(t_1) = \frac{\sum_{s=1}^m \lambda_{s,i}\ \sqrt{\pi_i}\ \rho(t_1, t^{\star(s)})}{\sigma} \quad , \quad m^i_{\tilde{t}^\star}(t_2) = \frac{\sum_{s=1}^m \lambda_{s,i}\ \sqrt{\pi_i}\ \rho(t^{\star(s)}, t_2)}{\sigma} \qquad (8)$$

The proof relies heavily on the proof of Theorem 1, provided that we replace formula (13) by the following formula:

$$S_n(t,i) = S_n^0(t,i)\ +\ \sum_{s=1}^{m}\sum_{j=1}^{n} \frac{\lambda_{s,i}}{n\ \sigma\ \sqrt{\pi_i}}\ 1_{C_j=i}\ X_j(t^{\star(s)})\ h_j(t)$$

where $S_n^0(.,i)$ is the process obtained under H_0.

In presence of several markers, it can be seen easily that formula (8) becomes

$$m^i_{\tilde{t}^\star}(t^\ell) = \frac{\sum_{s=1}^m \lambda_{s,i}\ \sqrt{\pi_i}\ \rho(t^\ell, t^{\star(s)})}{\sigma} \quad , \quad m^i_{\tilde{t}^\star_i}(t^r) = \frac{\sum_{s=1}^m \lambda_{s,i}\ \sqrt{\pi_i}\ \rho(t^{\star(s)}, t^r)}{\sigma}\ .$$

Table 11 focuses on the sparse map and investigates the power of the LRT when either 1, either 2 or 3 QTLs lie on the genome. We studied in particular the following configurations:

- $m = 1, t^{\star(1)} = 0.20$

- $m = 2$, $t^{\star(1)} = 0.20$, $t^{\star(2)} = 0.40$
- $m = 2$, $t^{\star(1)} = 0.20$, $t^{\star(2)} = 0.60$
- $m = 3$, $t^{\star(1)} = 0.20$, $t^{\star(2)} = 0.40$, $t^{\star(3)} = 0.60$.

Note also that the power is reported as a function of the QTL effect signs. For all cases, the absolute value of the constant linked to the QTL effect was equal to 2 (i.e. $\forall(s,i)$, $|\lambda_{s,i}\sqrt{\pi_i}| = 2$), in order to deal with small QTL effects. The Theoretical power was obtained by generating 100,000 paths of the asymptotic process, whereas 1,000 samples of size n were considered regarding the empirical power. We studied the cases $I = 1$, $I = 2$, $I = 3$, and a sample of size $n = 200I$, $n = 50I$, or $n = 30I$.

According to Table 11, there is a good agreement between the Empirical power and the Theoretical Power for $n = 200I$. Besides, the power increases with the number of families whatever the scenario studied. As expected, the power associated to the usual case $m = 1$ is fair in all cases, in particular for $I = 3,5$. However, when the number of QTLs increases ($m = 2$ or $m = 3$), the power higly depends on the QTL effect signs and the distance between QTLs. For instance, for $m = 2$, we observe a large decrease when the 2 QTLs of opposite signs become closer from each other. Last, we can notice that using the LRT under a 3 QTLs scenario can be more powerful than when only 1 QTL lie on the genome. To conclude, the use of the test statistic $\sup\Lambda_n(.)$ is appropriate for testing and localizing one QTL on $[0,T]$, but it is not so reliable when more than one QTL (i.e. $m > 1$) lie on $[0,T]$: it highly depends on the parameter values.

8 Illustration on human data

To conclude this study, we propose to give an illustration inspired from human real data. Although the daughter design is not realistic in humans, it is always interesting to use patterns from real data.

According to its dedicated website at http://hapmap.ncbi.nlm.nih.gov, “the International HapMap Project is a multi-country effort to identify and catalog genetic similarities and differences in human beings. The Project is a collaboration among scientists and funding agencies from Japan, the United Kingdom, Canada, China, Nigeria, and the United States. The goal of the International HapMap Project is to compare the genetic sequences of different individuals to identify chromosomal regions where genetic variants are shared.”

In this context, we downloaded data from Phase 3 release 2 of the HapMap Project. We focused on phased haplotype data of the following populations :

- Utah residents with Northern and Western European ancestry from the CEPH collection (CEU)
- Chinese in Metropolitan Denver, Colorado (CHD)
- Han Chinese in Beijing, China (CHB)
- Japanese in Tokyo, Japan (JPT)
- Mexican ancestry in Los Angeles, California (MEX)
- Toscans in Italy (TSI).

We investigated the presence of a QTL linked to human height on chromosome 7 in humans. In a previous Genome-Wide Association Study (GWAS), [Gudbjartsson 2008] highlighted the presence of a QTL located on a marker called "rs798544". As a consequence, our goal here is to check, with the help of simulated data, if we are able to recover this QTL using our global test.

To begin with, let us introduce the notion of marker informativity. At a given marker k, in order to know the genome information $X(t_k)$ of the offspring of one sire, the sire has to be heterozygous at this genetic marker, otherwise the marker is considered as uninformative. As a consequence, the "genome information" is available only at certain fixed locations called "informative markers", instead of "markers". These "informative markers" depend obviously on the family (i.e. on the sire) : one given marker can be informative in one given family, and uninformative in other families. In what follows, $0 \leqslant t_1^i < t_2^i < ... < t_{K^i}^i \leqslant T$ will denote the locations of informative markers in family i. Note that K^i refers to the number of informative markers in this family. As a result, an observation is now

$$\left(Y,\ X(t_1^C),\ ...,\ X(t_{K^C}^C),\ C\right).$$

Let $\mathbb{T}_{K^i}^i$ be the quantity such as $\mathbb{T}_{K^i}^i = \left\{t_1^i, ..., t_{K^i}^i\right\}$. In this context, it is straightforward to show that the asymptotic process $Z^i(.)$ verifies now $\forall t \in [t_1^i, t_{K^i}^i] \backslash \mathbb{T}_{K^i}^i$:

$$Z^i(t) = \frac{\alpha_i(t) Z^i(t^{\ell,i}) + \beta_i(t) Z^i(t^{r,i})}{\sqrt{\mathbb{V}\left\{\alpha_i(t) Z^i(t^{\ell,i}) + \beta_i(t) Z^i(t^{r,i})\right\}}}$$

where

$$t^{\ell,i} = \sup\left\{t_k^i \in \mathbb{T}_{K^i}^i : t_k^i < t\right\} \quad , \quad t^{r,i} = \inf\left\{t_k^i \in \mathbb{T}_{K^i}^i : t < t_k^i\right\}$$

and the $\alpha_i(t)$ and $\beta_i(t)$ are the analogue of $\alpha(t)$ and $\beta(t)$, relying now on the informative markers of family i. A proof is given in Section 8. The mean function of $Z^i(.)$ is still an interpolated function, based now on $\alpha_i(t)$ and $\beta_i(t)$. Last, the limiting process $\sum_{i=1}^I \left\{Z^i(.)\right\}^2$ is a Chi-Square process with I degrees of freedom where the $Z^i(.)$ are independent and not identically distributed. Note that the cases $t_1^i \neq 0$ and $t_{K^i}^i \neq T$ are discussed in the proof in Section 8 (see also below).

According to Hapmap data, chromosome 7 is of length $T = 1.86$M. A total of 75,245 markers are available and the locations of these markers are perfectly known. We considered a maximum of 14 families, and only 63,112 markers were found to be informative in at least one family. Table 12 gives the number of informative markers in each family. We can notice that the number of informative markers varies from 19,016 to 22,137: a large decrease in terms of marker density is observed, after filtering. In the same way as what has been done before, we focused on different number of families: $I = 6$, 10, 14. The set of chosen families was $\{1,4,6,8,10,12\}$ for $I = 6$, and $\{1,2,4,6,8,9,10,11,12,13\}$ for $I = 10$ (cf. Table 12), ensuring the presence of all the different kinds of populations (CEU, CHB, ...). The number of informative markers in at least one family was equal to 54,420 for $I = 6$, and equal to 60,819 for $I = 10$. In what follows, we will describe our simulation framework only for the case $I = 14$. Other cases can be deduced easily.

The genome of each sire was created by considering two haplotypes from the phased data of the chosen population. Note that once the sire's genome is built, we can extract easily the informative markers (i.e. the heterozygous markers). 50 offsprings were generated by family, and the LRT was computed over the grid defined by the 63,112 markers informative in at least one family. When the location did not match an informative marker in a given family, the value of $Z^i()$ was obtained by interpolation with the help of the functions $\alpha_i(.)$ and $\beta_i(.)$. Besides, when the first informative marker did not match the first extremity of the chromosome (i.e. $t_1^i \neq 0$), we considered (cf. proof in Section 8),

$$Z^i(t) = Z^i(t_1^i) \quad \forall t \in [0, t_1^i].$$

In the same way, when the last informative marker did not match the end of the chromosome (i.e. $t_{K^i}^i \neq 1.86$), we considered

$$Z^i(t) = Z^i(t_K^i) \quad \forall t \in [t_K^i, 1.86].$$

Table 13 compares thresholds obtained according to three methods available under a dense map: DF, DMC, and the permutation method. Note that ET is not reported since statistical tables given by [Estrella 2003] do not cover the case $T = 1.86$. The DMC method was computed with the help of I independent AR(1) processes: we generated paths of a DOUCS(I) with a constant discretization step equal to $1.86/63{,}111$. In other words, it assumes that the informative markers are the same across families, and equally spaced. Last, the permutation threshold was obtained by generating one population: it handles the fact that informative markers change across families. According to Table 13, as expected, DF gives the largest threshold. Recall that it relies on the continuous OUCS(I). We can also notice that, for $I = 10$ and $I = 6$ the DMC threshold is greater than the one obtained by permutation. It is not the case for $I = 14$.

Then, the true level associated to each method was computed by generating 1,000 populations. For instance, for $I = 10$, it was found respectively equal to 3.3%, 3.6%, and 4.1% for DF, DMC and the permutation method. In all cases, DF was the most conservative method, and the shuffling method, which handles the informativity correctly, seemed less conservative than DMC. However, the permutation method has a major drawback. It requires a large amount of time in order to compute the threshold (112h24 for $I = 14$, 78h10 for $I = 10$, 42h30 for $I = 6$). In opposite, DF can be obtained instantaneously, and DMC computation time remains reasonable (4h06 for $I = 14$, 2h40 for $I = 10$, 1h29 for $I = 6$).

Recall that our analysis relies only on one chromosome (number 7) in humans. In this context, our proposed methods seem to be the most appropriate for a whole genome study (23 chromosomes). Another advantage of our theoretical study is the following. In order to obtain the values of the process $Z^i(.)$ at each uninformative marker of family i, it is now possible to complete by interpolations with the functions $\alpha_i(.)$ and $\beta_i(.)$. Usually, geneticists perform either an EM algorithm to compute the MLE, or they choose a linearized likelihood method (cf. Section 6). The EM algorithm is time consuming, and performing a linearized likelihood method at each uninformative marker can also be challenging, specially when the number of such markers is large. In contrast, completing by interpolation is very fast.

To conclude, Table 14 investigates the power. In most cases, the permutation method is slightly more powerful than DF and DMC. Besides, as expected, considering families without QTL, decreases the power of the global test.

To conclude, in most cases, we were able, with our simulated data, to detect the QTL linked to human height on chromosome 7, highlighted by [Gudbjartsson 2008].

Acknowledgements This work has been supported by the the National Center for Scientific Research (CNRS), the Animal Genetic Department of the French National Institute for Agricultural Research, and SABRE. We thank Simon de Givry for help with human data.

References

[Azaïs and Cierco-Ayrolles 2002] Azaïs JM, Cierco-Ayrolles C (2002) An asymptotic test for quantitative gene detection. Ann. Inst. Henri Poincaré (B), 38:1087–1092

[Azaïs et al. 2014] Azaïs JM, Delmas C, Rabier CE (2014) Likelihood ratio test process for Quantitative Trait Locus detection. Statistics, 48:787–801

[Azaïs et al. 2009] Azaïs JM, Gassiat E, Mercadier C (2009) The likelihood ratio test for general mixture models with possibly structural parameter, ESAIM, 13:301–327

[Azaïs and Wschebor 2009] Azaïs JM, Wschebor M (2009) Level sets and extrema of random processes and fields. Wiley, New-York

[Benjamini and Hochberg 1995] Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological), 57(1):289–300

[Chang et al. 2009] Chang MN, Wu R, Wu SS, Casella G (2009) Score statistics for mapping quantitative trait loci. Statistical Application in Genetics and Molecular Biology, 8:16

[Chen et al. 2006] Chen HY, Zhang Q, Yin CC, Wang CK, Gong WJ, Mei G (2006) Detection of quantitative trait loci affecting milk production traits on bovine chromosome 6 in a Chinese holstein population by the daughter design, Journal of Dairy Science, 89:782–790

[Churchill and Doerge 1994] Churchill G, Doerge RW (1994) Empirical threshold values for quantitative trait mapping, Genetics, 138:963–971

[Cierco 1998] Cierco C (1998) Asymptotic distribution of the maximum likelihood ratio test for gene detection. Statistics, 31:261–285

[Davies1987] Davies RB (1987) Hypothesis testing when a nuisance parameter is present only under the alternative, Biometrika, 74:33–43

[Delong 1981] Delong D (1981) Crossing probabilities for a square root boundary by a Bessel process, Communication in Statistics Theory and Methods, 10:2197–2213

[Didelez et al. 2006] Didelez V, Pigeot I, Walter P (2006) Modifications of the Bonferroni-Holm procedure for a multi-way ANOVA, Statistical Papers, 47(2):181–209

[Estrella 2003] Estrella A (2003) Critical values and p values of bessel process distributions : computation and application to structural break tests, Econometric Theory, 19:1128–1143

[Frary et al. 2000] Frary A, Nesbitt TC, Frary A, Grandillo S, van der Knaap E et al (2000) fw2.2: A Quantitative Trait Locus Key to the Evolution of Tomato Fruit size, Science, 289:85–88

[Gassiat 2002] Gassiat E (2002) Likelihood ratio inequalities with applications to various mixtures , Ann. Inst. Henri Poincaré (B), 6:897–906

[Genz 1992] Genz A (1992) Numerical computation of multivariate normal probabilities, J. Comp. Graph. Stat., 1:141–149

[Gudbjartsson 2008] Gudbjartsson DF, Walters GB, Thorleifsson G, Stefansson H, Halldorsson BV, Zusmanovich P, ... Helgadottir A (2008) Many sequence variants affecting diversity of adult human height, Nature genetics, 40(5):609–615.

[Haldane 1919] Haldane JBS (1919) The combination of linkage values and the calculation of distance between the loci of linked factors. Journal of Genetics, 8:299–309

[Haley 1992] Haley CS, Knott S (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity, 69(4):315-324.

[Jung et al. 2007] Jung BC, Jhun M, Song SH (2007) A new random permutation test in ANOVA models. Statistical Papers, 48(1):47–62

[Lander and Botstein 1989] Lander ES and Botstein D (1989) Mapping mendelian factors underlying quantitative traits using RFLP linkage maps, Genetics, 138:235–240
[Li et al. 2006] Li C, Zhou A, Sang T (2006) Rice domestication by reducing shattering, Science, 311:1936–1939
[Plackett 1984] Plackett RI (1984) A reduction formula for normal multivariate integrals, Biometrika, 41:351–360
[Rabier and Genz 2014] Rabier CE, Genz A (2014) The supremum of Chi-Square processes, Methodology and Computing in Applied Probability, 16:715–729
[Rabier 2014] Rabier CE (2014) On Quantitative Trait Locus mapping with an interference phenomenon, TEST, 23(2):311–329
[Rabier 2015] Rabier CE (2015) On stochastic processes for Quantitative Trait Locus mapping under selective genotyping, Statistics, 49:19–34
[Rebaï et al. 1994] Rebaï A, Goffinet B, Mangin B (1994) Approximate thresholds of interval mapping tests for QTL detection. Genetics, 138:235–240
[Rebaï et al. 1995] Rebaï A, Goffinet B, Mangin B (1995) Comparing power of different methods for QTL detection. Biometrics, 51:87–99
[Ron et al. 2001] Ron M, Kliger D, Feldmesser E, Seroussi E, Ezra E, Weller JI (2001) Multiple quantitative trait locus analysis of bovine chromosome 6 in the Israeli holstein population by a daughter design, Genetics, 159:727–735
[Siegmund and Yakir 2007] Siegmund D, Yakir B (2007) The statistics of gene mapping. Springer, New York
[Silva et al. 2011a] Silva AA, Azevedo ALS, Gasparini K, Verneque RS, Peixoto MGCD, Panetto BR, Guimaraes SEF, Machado MA (2011a) Quantitative trait loci affecting lactose and total solids on chromosome 6 in Brazilian Gir dairy cattle, Genetics and Molecular Research, 10:3817–3827
[Silva et al. 2011b] Silva AA, Azevedo ALS, Verneque RS, Gasparini K, Peixoto MGCD, da Silva MVGB, Lopes PS, Guimaraes SEF, Machado MA (2011b) Quantitative trait loci affecting milk production traits on bovine chromosome 6 in zebuine Gyr breed, Journal of Dairy Science, 94:971–980
[Van der Vaart 1998] Van der Vaart AW (1998) Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics
[Weller et al. 2008] Weller JI, Golik M, Seroussi E, Ron M, Ezra E (2008) Detection of quantitative trait loci affecting twinning rate in Israeli holsteins by the daughter Design, Journal of Dairy Science, 91:2469–2474
[Weller et al. 1990] Weller JI, Kashi Y, Soller M (1990) Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. Journal of Dairy Science, 73:2525–2537
[Wu et al. 2007] Wu R, Ma CX, Casella G (2007) Statistical Genetics of Quantitative Traits. Springer

Appendix 1. Proof of Theorem 1

Preliminaries

Let t belong to the interval $[t_1,t_2]$ and let recall Lemma 3.1 of Azais, Delmas & Rabier (2012).

Lemma 1 *The conditional expectation $x(t)$ of $X(t)$ is linear in $X(t_1),X(t_2)$:*

$$x(t) = \alpha(t)X(t_1) + \beta(t)X(t_2)$$

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

Then, we have the following relationship

$$\mathbb{V}\left\{x^2(t)\right\} = \mathbb{E}\left\{x^2(t)\right\} = \alpha^2(t) \ + \ \beta^2(t) \ + \ 2 \ \alpha(t) \ \beta(t)\rho(t_1,t_2) \ .$$

Since the model is regular, we can apply Theorem 5.39 of Van der Vaart (98). As a result, according to formulae (4) and (5), we have

$$\Lambda_n(t) = \sum_{i=1}^{I} \left[\sum_{j=1}^{n} \frac{(Y_j - \mu_i)\, x_j(t)}{\sigma\, \sqrt{n\, \pi_i\, \mathbb{E}\{x^2(t)\}}}\, 1_{C_j=i} \right]^2 + o_{P_{\theta_0}}(1) \tag{9}$$

where $o_{P_{\theta_0}}(1)$ denotes a sequence of random vectors that converges to zero in probability under H_0.

Let $S_n(.,i)$ be the following process, for n observations:

$$S_n(t,i) = \sum_{j=1}^{n} \frac{(Y_j - \mu_i)\, x_j(t)}{\sigma\, \sqrt{n\, \pi_i\, \mathbb{E}\{x^2(t)\}}}\, 1_{C_j=i} \ . \tag{10}$$

According to Lemma 1,

$$S_n(t,i) = \{\ \alpha(t)\, S_n(t_1,i) \ +\ \beta(t)\, S_n(t_2,i)\ \} \ /\ \sqrt{\mathbb{E}\{x^2(t)\}}\ .$$

We will call $Z^i(.)$ the limiting process of $S_n(.,i)$.

Study under H_0

Without loss of generality, let us assume $n = 1$ and let us consider the process $S(.,i)$ defined in the following way:

$$S(t,i) = \frac{(Y - \mu_i)\, x(t)}{\sigma\, \sqrt{\pi_i\, \mathbb{E}\{x^2(t)\}}}\, 1_{C=i} = \frac{Y - \mu_i}{\sigma\, \sqrt{\pi_i}}\, 1_{C=i}\ \ h(t)\ .$$

where $h(t) = x(t)/\sqrt{\mathbb{E}\{x^2(t)\}}$.
$h(.)$ is a random process, independent of Y and C. It is easy to see that

$$\mathbb{E}\{S(t,i)\} = 0 \ \ ,\ \ \mathbb{V}\{S(t,i)\} = \mathbb{E}\{h(t)\}^2 = 1 \ \ .$$

Besides,

$$\mathrm{Cov}\{\mathrm{S}(\mathrm{t}_1,\mathrm{i}),\ \mathrm{S}(\mathrm{t}_2,\mathrm{i})\} = \mathbb{E}\{\mathrm{h}(\mathrm{t}_1)\mathrm{h}(\mathrm{t}_2)\} = \rho(\mathrm{t}_1,\mathrm{t}_2) \ \ .$$

So, we have

$$Z^i(t) = \left\{\ \alpha(t)\, Z^i(t_1) \ +\ \beta(t)\, Z^i(t_2)\ \right\} \ /\ \sqrt{\mathbb{E}\{x^2(t)\}}\ \ ,$$

$\mathbb{E}\left\{Z^i(t)\right\} = 0$, $\mathbb{V}\left\{Z^i(t)\right\} = 1$ and $\mathrm{Cov}\left\{\mathrm{Z}^{\mathrm{i}}(\mathrm{t}_1),\ \mathrm{Z}^{\mathrm{i}}(\mathrm{t}_2)\right\} = \rho(\mathrm{t}_1,\mathrm{t}_2)$.
A direct application of central limit theorem implies that $Z(t_1)$ and $Z(t_2)$ have a limit distribution which is a Gaussian distribution. According to formula (9), we have $\Lambda_n(t) = \sum_{i=1}^{I} S_n^2(t,i) + o_{P_{\theta_0}}(1)$. As a result, $\Lambda_n(.) \overset{F.d.}{\rightarrow} \sum_{i=1}^{I} \left\{Z^i(.)\right\}^2$.

Study under $H_{\lambda t^\star}$

In this part, we set

$$Y = \mu_i \ + \ \frac{\lambda_i}{\sqrt{n}}\, X(t^\star) + \ \sigma\varepsilon \ \ , \ \ \text{if } C = i \ , \tag{11}$$

where ε is a standard normal random variable. Recall that $t^\star$ denotes the QTL location.

According to formula (9), we have

$$\Lambda_n(t) = \sum_{i=1}^{I} S_n^2(t,i) \ + \ o_{P_{\theta_0}}(1) \ \ . \tag{12}$$

Recall that under $H_{\lambda t^\star}$, if there is a QTL within family i (i.e. $\lambda_i \neq 0$), the density of $Y \big| X(t_1), X(t_2), C$ verifies

$$p(t^\star) f_{(\mu_i+q_i,\sigma)}(Y) + \{1 - p(t^\star)\}\, f_{(\mu_i-q_i,\sigma)}(Y) \ \ , \ \ \text{if } C = i \ .$$

The model with t^* fixed is differentiable in quadratic mean, this implies that the alternative defines a contiguous sequence of alternatives. By Le Cam's first lemma, relation (12) remains true under the alternative. As a result, $\Lambda_n(.) \overset{F.d.}{\to} \sum_{i=1}^{I} \left\{Z^i(.)\right\}^2$.

Calculations of the mean function of $Z^i(.)$, so-called $m^i_{t^\star}(t)$, can be done using the process $S_n(.,i)$. According to formula (16) and (11), we have

$$\begin{aligned} S_n(t,i) &= \frac{1}{\sqrt{n\ \pi_i}} \sum_{j=1}^{n} \varepsilon_j \ 1_{C_j=i}\ h_j(t) \ + \ \sum_{j=1}^{n} \frac{\lambda_i}{n\ \sigma\ \sqrt{\pi_i}}\ 1_{C_j=i}\ X_j(t^\star)\ h_j(t) \\ &= S_n^0(t,i) \ + \ \sum_{j=1}^{n} \frac{\lambda_i}{n\ \sigma\ \sqrt{\pi_i}}\ 1_{C_j=i}\ X_j(t^\star)\ h_j(t) \end{aligned} \tag{13}$$

where $S_n^0(.,i)$ is the process obtained under H_0.

Recall that $h_j(.)$ is the equivalent of the process $h(.)$ for the individual j. According to the law of large number :

$$\frac{1}{n} \sum_{j=1}^{n} X_j(t^\star)\ h_j(t)\ 1_{C_j=i} \to \pi_i\ \mathbb{E}\left\{X(t^\star)h(t)\right\} \ . \tag{14}$$

Besides, we have $\mathbb{E}\left\{X(t^\star)h(t_1)\right\} = \rho(t_1,t^\star)$ and $\mathbb{E}\left\{X(t^\star)h(t_2)\right\} = \rho(t^\star,t_2)$.
As a result,

$$m^i_{t^\star}(t_1) = \ \lambda_i\ \sqrt{\pi_i}\ \rho(t_1,t^\star)\ /\ \sigma \ \text{ and } \ m^i_{t^\star}(t_2) = \lambda_i\ \sqrt{\pi_i}\ \rho(t^\star,t_2)\ /\ \sigma \ \ .$$

Due to the interpolation, we have

$$m^i_{t^\star}(t) = \left\{\alpha(t)\ m^i_{t^\star}(t_1) \ + \ \beta(t)\ m^i_{t^\star}(t_2)\right\} \ / \ \sqrt{\mathbb{E}\left\{x^2(t)\right\}} \ \ .$$

Study of the supremum of the LRT process

Since the model with t fixed is regular, we have the relationship (cf. section "Study under H_0")

$$\Lambda_n(t) = \sum_{i=1}^{I} S_n^2(t,i) + o_{P_{\theta_0}}(1)$$

under the null hypothesis. Our goal is now to prove that the rest above is uniform in t.

Let us consider now t as an extra parameter. Let t^*, θ^* be the true parameter that will be assumed to belong to H_0. Note that t^* makes no sense for θ belonging to H_0. It is easy to check that at H_0 the Fisher information relative to t is zero so that the model is not regular.

It can be proved that assumptions 1, 2 and 3 of [Azaïs et al. 2009] holds. So, we can apply Theorem 1 of [Azaïs et al. 2009] and we have

$$\sup_{(t,\theta)} l_t^n(\theta) - l_{t^*}^n(\theta^*) = \sup_{d\in\mathscr{D}} \left[\left\{ \frac{1}{\sqrt{n}} \sum_{j=1}^{n} d(X_j) \right\}^2 1_{d(X_j)\geqslant 0} \right] + o_P(1) \qquad (15)$$

where the observation X_j stands for $Y_j, X_j(t_1), X_j(t_2), C_j$ and where $\mathscr{D}$ is the set of scores defined in [Azaïs et al. 2009], see also [Gassiat 2002]. A similar result is true under H_0 with a set $\mathscr{D}_0$. Let us precise the sets of scores $\mathscr{D}$ and $\mathscr{D}_0$. This sets are defined at the sets of scores of one parameter families that converge to the true model p_{t^*,θ^*} and that are differentiable in quadratic mean.

It is easy to see that

$$\mathscr{D} = \left\{ \frac{\langle U, l_t'(\theta^*)\rangle}{\sqrt{\mathbb{V}(\langle U, l_t'(\theta^*)\rangle)}}, U \in \mathbb{R}^{2I+1}, t \in [t_1, t_2] \right\}$$

where l' is the gradient with respect to θ. In the same manner

$$\mathscr{D}_0 = \left\{ \frac{\langle U, l_t'(\theta^*)\rangle}{\sqrt{\mathbb{V}(\langle U, l_t'(\theta^*)\rangle)}}, U \in \mathbb{R}^{I+1} \right\},$$

where now the gradient is taken with respect to μ_1, ..., μ_I and σ only. Obviously, this gradient does not depend on t.

Using the transform $U \to -U$ in the expressions of the sets of score, we see that the indicator function can be removed in formula (15). Then, since the Fisher

information matrix is diagonal (see formula (5)) , it is easy to see that

$$\sup_{d\in\mathscr{D}}\left[\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}d(X_j)\right\}^2\right]-\sup_{d\in\mathscr{D}_0}\left[\left\{\frac{1}{\sqrt{n}}\sum_{j=1}^{n}d(X_j)\right\}^2\right]$$

$$=\sup_{t\in[t_1,t_2]}\left(\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\frac{\partial l_t}{\partial q_1}(X_j)\mid_{\theta_0}}{\sqrt{\mathbb{V}\left\{\frac{\partial l_t}{\partial q_1}(X_j)\mid_{\theta_0}\right\}}}\right]^2+\ldots+\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\frac{\partial l_t}{\partial q_I}(X_j)\mid_{\theta_0}}{\sqrt{\mathbb{V}\left\{\frac{\partial l_t}{\partial q_I}(X_j)\mid_{\theta_0}\right\}}}\right]^2\right)$$

$$=\sup_{t\in[t_1,t_2]}\left(\sum_{i=1}^{I}\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\frac{\frac{\partial l_t}{\partial q_i}(X_j)\mid_{\theta_0}}{\sqrt{\mathbb{V}\left\{\frac{\partial l_t}{\partial q_i}(X_j)\mid_{\theta_0}\right\}}}\right]^2\right).$$

This is exactly the desired result. Since the model with t^* fixed is differentiable in quadratic mean, the alternative defines a contiguous sequence of alternatives. By Le Cam's first lemma, relation (15) remains true under the alternative.

Appendix 2. Proof of Theorem 2

The proof of the theorem is the same as the proof of Theorem 1 as soon as we can confine our attention to the interval (t^ℓ,t^r) when considering a unique instant t and to the intervals $(t^\ell,t^r)(t'^\ell,t'^r)$ when considering two instants t and t'. For that we need to prove that

$$x(t)=\mathbb{E}\{X(t)|X(t_1),\ldots,X(t_K)\}=\mathbb{E}\left\{X(t)|X(t^\ell),X(t^r)\right\}$$

which is a direct consequence of the independance of the increments of Poisson process.

Proof of results introduced in Section 8

Recall that $\mathbb{T}^i_{K^i}=\left\{t^i_1,\ldots,t^i_{K^i}\right\}$. Let $t\in[t^i_1,t^i_{K^i}]\backslash\mathbb{T}^i_{K^i}$. Let define $x^i(t)$ the quantity such as $x^i(t)=\mathbb{E}\left\{X(t)\mid X(t^{\ell,i}),X(t^{r,i}),C=i\right\}$. Besides, $Q^{1,1}_{t,i}$, $Q^{1,-1}_{t,i}$, $Q^{-1,1}_{t,i}$ and $Q^{-1,-1}_{t,i}$ are the following quantities:

$$Q^{1,1}_{t,i}=\frac{\bar{r}(t^{\ell,i},t)\ \bar{r}(t,t^{r,i})}{\bar{r}(t^{\ell,i},t^{r,i})}\quad,\quad Q^{1,-1}_{t,i}=\frac{\bar{r}(t^{\ell,i},t)\ r(t,t^{r,i})}{r(t^{\ell,i},t^{r,i})}$$

$$Q^{-1,1}_{t,i}=\frac{r(t^{\ell,i},t)\ \bar{r}(t,t^{r,i})}{r(t^{\ell,i},t^{r,i})}\quad,\quad Q^{-1,-1}_{t,i}=\frac{r(t^{\ell,i},t)\ r(t,t^{r,i})}{\bar{r}(t^{\ell,i},t^{r,i})}.$$

Lemma 2 *We have the following relationship:*

$$x^i(t)=\alpha_i(t)X(t^{\ell,i})+\beta_i(t)X(t^{r,i})$$

with $\alpha_i(t) = Q_{t,i}^{1,1} - Q_{t,i}^{-1,1}$, $\beta_i(t) = Q_{t,i}^{1,1} - Q_{t,i}^{1,-1}$.

Let $S_n(.,i)$ be the following process, for n observations:

$$S_n(t,i) = \sum_{j=1}^{n} \frac{(Y_j - \mu_i)\ x_j^i(t)}{\sigma\ \sqrt{n\ \pi_i\ \mathbb{E}\left\{(x^i(t))^2\right\}}}\ 1_{C_j=i}\ . \tag{16}$$

According to Lemma 2

$$S_n(t,i) = \left\{\ \alpha_i(t)\ S_n(t^{\ell,i},i)\ +\ \beta_i(t)\ S_n(t^{r,i},i)\ \right\}\ /\ \sqrt{\mathbb{E}\left\{(x^i(t))^2\right\}}\ .$$

We will call $Z^i(.)$ the limiting process of $S_n(.,i)$.

Let us consider now the case where the first informative marker does not lie at the beginning of the chromosome ($0 < t_1^i$). Let $t \in [0, t_1^i[$, we have

$$S_n(t,i) = \sum_{j=1}^{n} \frac{(Y_j - \mu_i)\ \tilde{x}_j^i(t)}{\sigma\ \sqrt{n\ \pi_i\ \mathbb{E}\left\{(\tilde{x}^i(t))^2\right\}}}\ 1_{C_j=i}$$

where $\tilde{x}^i(t) = 2\ \mathbb{P}\left\{X(t) = 1 \mid X(t_1^i), C = i\right\} - 1$. Recall that in the classical situation, when t have two flanking markers : $x^i(t) = 2\ \mathbb{P}\left\{X(t) = 1 \mid X(t^{\ell,i}), X(t^{r,i}), C = i\right\} - 1$. In our case,

$$\begin{aligned}\tilde{x}^i(t) &= 2\left\{\bar{r}(t,t_1^i)1_{X(t_1^i)=1} + r(t,t_1^i)1_{X(t_1^i)=-1}\right\} - 1\\ &= 2\left\{\rho(t,t_1^i) + r(t,t_1^i)\right\} - 1 = \rho(t,t_1^i)\end{aligned}$$

Besides, we have

$$\sqrt{\mathbb{E}\left\{(\tilde{x}^i(t))^2\right\}} = \rho(t,t_1^i)\ .$$

As a result,

$$\forall t \in [0,t_1^i[\quad S_n(t,i) = S_n(t_1^i,i)\ .$$

By symmetry, when $t_{K^i}^i < T$, we have

$$\forall t \in]t_{K^i}^i, T]\quad S_n(t_{K^i}^i,i) = S_n(t,i).$$

To conclude, we just have to use same kind of arguments as in formula (9) in order to prove that the LRT process converges asymptotically to the process $\sum_{i=1}^{I}\left\{Z^i(.)\right\}^2$.

Table 1 Summary of all the methods studied as a function of the genetic map (DMC for Discrete Monte-Carlo, DMCQMC for Discrete Monte-Carlo Quasi Monte-Carlo, ET for Estrella Exact Table, DF for Delong Approximative Formula)

Map	Method
Dense (testing on markers)	ET (table available for $I \leqslant 20$) DF (formula available for I and threshold large)
Sparse (testing between markers)	DMCQMC (available only for $I = 1$) DMC for $I > 1$

Table 2 Thresholds obtained using the appropriate method as a function of the value of I considered (nspaths=1,000,000). The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is performed every 5cM.

Method	DMCQMC ($I = 1$)	DMC ($I = 3$)	DMC ($I = 5$)
Threshold	6.06	10.76	14.47

Table 3 Number of False Positives (NFP) as a function of the number of individuals n and the method considered. The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is performed every 5cM ($\sigma = 1$, $\mu_1 = -0.37$, $\mu_2 = 0.03$, $\mu_3 = 0.06$, $\mu_4 = -0.26$, $\mu_5 = 0.27$, npop=40,000).

n \ Method	DMCQMC ($I = 1$)	DMC ($I = 3$)	DMC ($I = 5$)
200 I	5.20% [4.98%;5.42%]	5.03% [4.82%;5.24%]	5.22% [5.00%;5.44%]
50 I	5.78% [5.55%;6.01%]	5.97% [5.74%;6.20%]	6.11% [5.88%;6.34%]
30 I	6.60% [6.36%;6.84%]	6.77% [6.52%;7.02%]	7.08% [6.83%;7.33%]

Table 4 Thresholds obtained using theoretical methods ET, DF as a function of the value of I considered. DMC for checking (nspaths=1,000,000). The map consists of 501 genetic markers equally spaced every 0.1cM (T=50cM). A test is performed on each marker.

	$I = 1$			$I = 3$			$I = 5$		
Method	ET	DF	DMC	ET	DF	DMC	ET	DF	DMC
Threshold	7.84	7.61	7.68	13.09	12.91	12.86	17.15	17.02	16.94

Table 5 Number of False Positives (NFP) as a function of the number of individuals n and the method used ($I = 5$). The map consists of 501 genetic markers equally spaced every 0.1cM (T=50cM). A test is performed on each marker ($\sigma = 1$, $\mu_1 = -0.37$, $\mu_2 = 0.03$, $\mu_3 = 0.06$, $\mu_4 = -0.26$, $\mu_5 = 0.27$, npop=40,000).

n \ Method	DF	DMC	ET
1000	4.78% [4.57%;4.99%]	5.13% [4.91%;5.35%]	4.41% [4.21%;4.61%]
500	4.96% [4.75%;5.17%]	5.15% [4.93%;5.37%]	4.64% [4.43%;4.85%]
150	5.67% [5.44%;5.90%]	5.91% [5.68%;6.14%]	5.34% [5.12%;5.56%]

Table 6 Theoretical Power and Empirical Power (EP) as a function of the method used, and the number of families. The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is performed every 5cM ($\lambda = 2$, $t^\star = 25$cM, nspaths=100,000 for the Theoretical Power and npop=10,000 for the Empirical Power, $\mu_1 = -0.37$, $\mu_2 = 0.03$, $\mu_3 = 0.06$, $\mu_4 = -0.26$, $\mu_5 = 0.27$, $\sigma = 1$, $nz = I$). Thresholds are given between parentheses, and confidence intervals for the true value of the power are given between brackets.

Method	Power	$I = 1$	$I = 3$	$I = 5$
Mixture model (this paper)	Theoretical	37.59% (6.06)	68.57% (10.76)	85.58% (14.47)
	EP for $n = 200$ I	38.08% [37.13%;39.03%]	68.80% [67.89%;69.71%]	85.00% [84.30%;85.70%]
	EP for $n = 50$ I	37.54% [36.59%;38.49%]	68.37% [67.46%;69.28%]	84.74% [84.04%;85.44%]
	EP for $n = 30$ I	37.83% [36.88%;38.78%]	68.57% [67.66%;69.48%]	85.15% [84.45%;85.85%]
Permutation + Linearized likelihood	EP for $n = 200$ I	36.91% (6.11) [35.96%;37.86%]	67.15% (10.83) [66.23%;68.07%]	84.37% (14.60) [83.66%;85.08%]
	EP for $n = 50$ I	35.43% (6.32) [34.49%;36.37%]	65.37% (11.09) [64.44%;66.30%]	82.96% (14.78) [82.22%;83.69%]
	EP for $n = 30$ I	31.37% (6.79) [30.46%;32.27%]	62.72% (11.53) [61.77%;63.67%]	79.84% (15.23) [79.05%;80.62%]

Table 7 Same legend as Table 6 except that $t^\star = 50$cM.

Method	Power	$I = 1$	$I = 3$	$I = 5$
Mixture model (this paper)	Theoretical	34.99% (6.06)	65.10% (10.76)	82.64% (14.47)
	EP for $n = 200$ I	34.89% [33.95%;35.82%]	64.26% [63.32%;65.20%]	81.60% [80.84%;82.35%]
	EP for $n = 50$ I	35.07% [34.13%;36.01%]	64.91% [63.97%;65.85%]	81.31% [80.55%;82.07%]
	EP for $n = 30$ I	35.71% [34.77%;36.64%]	64.18% [63.24%;65.11%]	80.63% [79.85%;81.40%]
Permutation + Linearized likelihood	EP for $n = 200$ I	34.05% (6.09) [33.12%;32.13%]	63.21% (10.83) [62.26%;64.15%]	81.53% (14.50) [80.77%;82.29%]
	EP for $n = 50$ I	31.22% (6.51) [30.31%;36.37%]	61.12% (11.24) [60.16%;62.07%]	78.56% (15.01) [77.75%;79.36%]
	EP for $n = 30$ I	28.80% (6.90) [27.81%;29.69%]	61.01% (11.50) [60.05%;61.97%]	75.80% (15.67) [74.96%;76.64%]

Table 8 Difference in terms of power between the LRT and the test relying only on marker locations ($T = 1, \forall i \;\; \lambda = \lambda_i\sqrt{\pi_i} = 2, \sigma = 1$, ζ refers to the intensity of the Poisson process)

Map	$t^\star$	ζ	$I = 1$	$I = 3$	$I = 5$
map 1	0.5	1	$[0.86\%, 0.93\%]$	$[1.69\%, 1.77\%]$	$[1.67\%, 1.73\%]$
		5	$[1.07\%, 1.13\%]$	$[2.52\%, 2.62\%]$	$[4.09\%, 4.19\%]$
	0.4	1	$[0.60\%, 0.69\%]$	$[0.53\%, 0.59\%]$	$[0.24\%, 0.29\%]$
		5	$[0.20\%, 0.31\%]$	$[-1.21\%, -1.12\%]$	$[-1.24\%, -1.18\%]$
map 2	0.51	1	$[0.011\%, 0.034\%]$	$[0.0030\%, 0.0079\%]$	$[-0.00027\%, 0.0010\%]$
		5	$[0.505\%, 0.577\%]$	$[0.365\%, 0.397\%]$	$[0.063\%, 0.072\%]$
	0.4	1	$[0.0032\%, 0.019\%]$	$[0.020\%, 0.042\%]$	$[0.011\%, 0.026\%]$
		5	$[0.144\%, 0.196\%]$	$[0.297\%, 0.355\%]$	$[0.205\%, 0.259\%]$

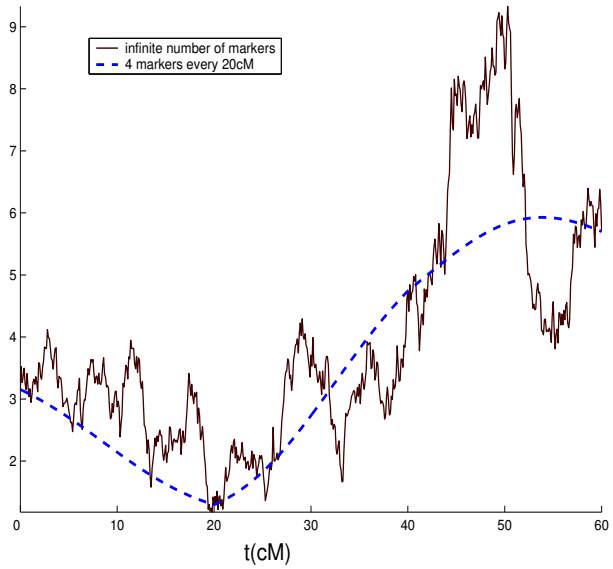


Fig. 1 Paths of the process $\sum_{i=1}^{3} \{Z^i(.)\}^2$ as a function of the genetic map (T=60cM).

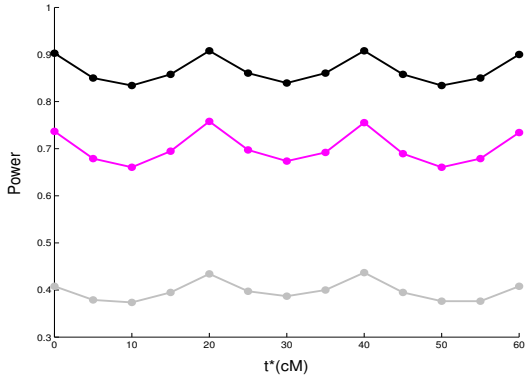


Fig. 2 Power as a function of $t^\star$ and I. From top to bottom, $I=5$, $I=3$, $I=1$ ($\forall$i $\lambda=\lambda_i\sqrt{\pi_i}=2$, $\sigma=1$, nspaths=100,000). The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is performed every 5cM.

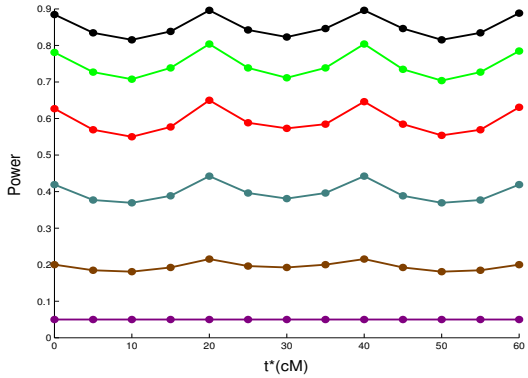


Fig. 3 Power as a function of $t^\star$ and the number nz of non zero. From top to bottom, $nz=5,\ 4,\ 3,\ 2,\ 1,\ 0$ ($I=5$, $\lambda=2$, $\sigma=1$, nspaths=100,000). The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is performed every 5cM.

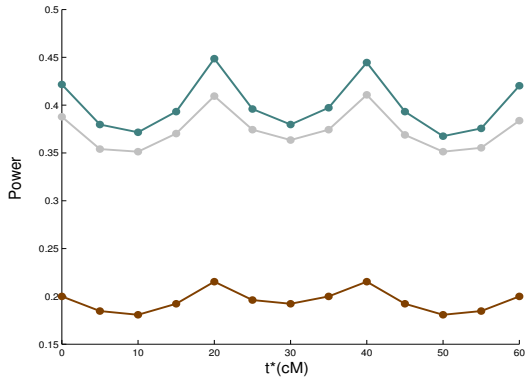


Fig. 4 Power as a function of $t^\star$, I and the number nz of non zero. From top to bottom : $I = 5$ with $nz = 2$, $I = 1$ with $nz = 1$, $I = 5$ with $nz = 1$ ($\lambda = 2$, $\sigma = 1$, nspaths=100,000). The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is performed every 5cM.

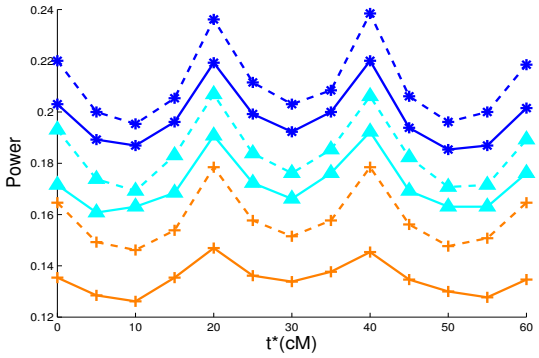


Fig. 5 Power of the global approach (in solid line) and power of the Bonferroni approach (in dashed line), as a function of $t^\star$ and in the particular case of $nz = 1$. Crosses refer to $I = 12$, rectangles to $I = 7$ and stars to $I = 5$ ($\lambda = 2$, $\sigma = 1$, nspaths=100,000). The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is performed every 5cM.

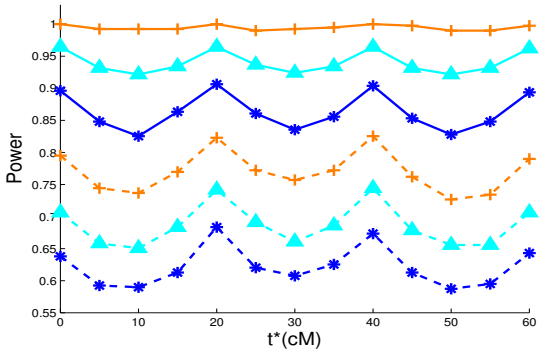


Fig. 6 Power of the global approach (in solid line) and power of the Bonferroni approach (in dashed line), as a function of $t^\star$ and in the particular case of $nz = I$. Crosses refer to $I = 12$, rectangles to $I = 7$ and stars to $I = 5$ ($\lambda = 2$, $\sigma = 1$, nspaths=100,000). The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is performed every 5cM.

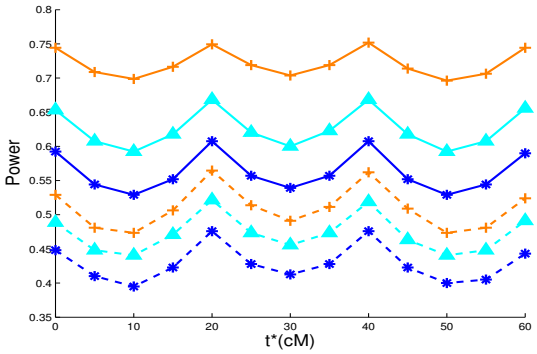


Fig. 7 Mean power of the global approach (in solid line) and mean power of the Bonferroni approach (in dashed line), as a function of $t^\star$. Crosses refer to $I = 12$, rectangles to $I = 7$ and stars to $I = 5$ ($\lambda = 2$, $\sigma = 1$, nspaths=100,000). The map consists of 4 genetic markers equally spaced every 20cM (T=60cM). A test is performed every 5cM.

$(t_1^\star,t_2^\star)$ \ $t_3^\star$	0	5	10	15	20	25	30	35	40	45	50	55	60
$(20,20)$	70.22%	70.11%	71.07%	72.66%	74.66%	72.79%	70.76%	70.74%	70.96%	68.58%	67.00%	66.79%	66.81%
$(50,50)$	60.07%	59.35%	59.92%	61.00%	63.44%	62.82%	63.41%	64.86%	67.67%	65.71%	64.88%	65.39%	66.82%

Table 9 Theoretical Power when the QTL locations (in cM) can be different between families ($I=3$, $T=0.60$, $K=4$, $\forall k=1,..,4 \quad t_k=0.20(k-1)$, $\forall i\ \lambda=\lambda_i\sqrt{\pi_i}=2$, $\sigma=1$, 100,000 paths for the Theoretical Power).

$(t_1^\star,t_2^\star,t_3^\star)$ \ $(t_4^\star,t_5^\star)$	$(0,0)$	$(5,5)$	$(10,10)$	$(15,15)$	$(20,20)$	$(25,25)$	$(30,30)$	$(35,35)$	$(40,40)$	$(45,45)$	$(50,50)$	$(55,55)$	$(60,60)$
$(20,20,20)$	86.46%	86.11%	87.01%	88.41%	90.56%	88.58%	87.21%	86.52%	87.10%	84.68%	83.04%	82.27%	82.99%
$(50,50,50)$	76.82%	75.73%	76.15%	78.05%	80.92%	80.21%	80.71%	82.41%	85.34%	83.29%	82.76%	82.92%	84.75%

Table 10 Theoretical Power when the QTL locations (in cM) can be different between families ($I=5$, $T=0.60$, $K=4$, $\forall k=1,..,4 \quad t_k=0.20(k-1)$, $\forall i\ \lambda=\lambda_i\sqrt{\pi_i}=2$, $\sigma=1$, 100,000 paths for the Theoretical Power)

I	n \ m	1 (+)	2^a (+ -)	2^b (+ -)	3 (+ - +)	3 (+ - -)
1	$+\infty$	41.28%	10.62%	20.60%	32.97%	57.71%
	200 I	40.60%	10.20%	21.20%	34.70%	58.20%
	50 I	41.50%	11.10%	19.00%	31.60%	56.30%
	30 I	39.90%	12.50%	21.50%	29.50%	54.20%
3	$+\infty$	74.77%	14.16%	35.67%	60.82%	92.26%
	200 I	74.90%	15.20%	34.20%	60.00%	90.30%
	50 I	73.70%	15.40%	34.30%	59.00%	91.10%
	30 I	72.80%	14.90%	34.10%	57.90%	88.80%
5	$+\infty$	90.58%	17.67%	49.56%	79.24%	99.00%
	200 I	91.20%	18.90%	48.50%	78.40%	98.90%
	50 I	88.60%	16.90%	46.90%	78.10%	98.10%
	30 I	91.30%	21.90%	44.90%	78.50%	97.00%

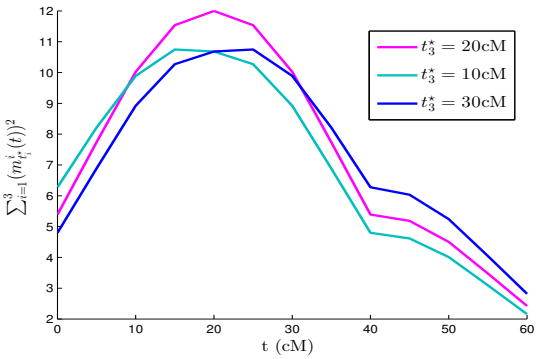
Table 11 Theoretical Power and Empirical Power as a function of the number m of QTLs, their effects, and the number I of families ($T = 0.60$, $K = 4$, $\forall k = 1,..,4 \quad t_k = 0.20(k-1)$). A test is performed every 5cM. (+) denotes a positive effect whereas $(-)$ is a negative effect. 1 refers to the situation ($m = 1$, $t^{\star(1)} = 0.20$), 2^a refers to ($m = 2$, $t^{\star(1)} = 0.20$, $t^{\star(2)} = 0.40$) , 2^b refers to ($m = 2$, $t^{\star(1)} = 0.20$, $t^{\star(2)} = 0.60$), 3 refers to ($m = 3$, $t^{\star(1)} = 0.20$, $t^{\star(2)} = 0.40$, $t^{\star(3)} = 0.60$). $\forall(s,i) \quad |\lambda_{s,i}\sqrt{\pi_i}| = 2$, $\sigma = 1$, 100,000 paths for the Theoretical Power, 1,000 samples of size n for the Empirical Power.

Family ID	Population	Nb Informative Markers
1	CEU	20,925
2	CEU	21,939
3	CEU	21,572
4	CHD	20,053
5	CHD	19,753
6	CHB	20,340
7	CHB	19,280
8	JPT	19,016
9	JPT	19,326
10	MEX	19,556
11	MEX	21,803
12	TSI	21,409
13	TSI	20,867
14	TSI	22,137

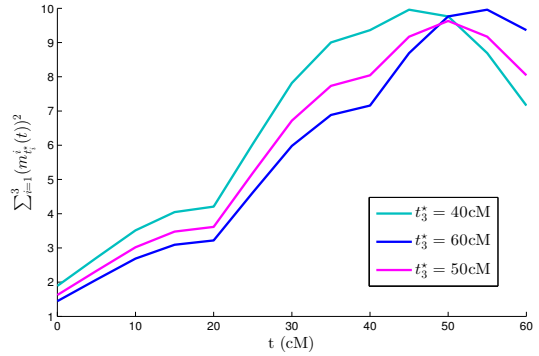
Table 12 Number of informative markers in each family (CEU=Utah residents with Northern and Western European ancestry from the CEPH collection, CHD=Chinese in Metropolitan Denver (Colorado), CHB=Han Chinese in Beijing (China), JPT=Japanese in Tokyo (Japan), MEX=Mexican ancestry in Los Angeles (California), TSI=Toscans in Italy).

Table 13 Thresholds, CPU time and Number of False Positives (NFP) according to DF, DMC, and Permutation methods (nspaths=100,000 for DMC, Permutation threshold based on 1,000 shufflings npop=1 and $n = 50I$, NFP based on npop=1,000).

	$I = 14$			$I = 10$			$I = 6$		
Method	DF	DMC	Permutation	DF	DMC	Permutation	DF	DMC	Permutation
Threshold	36.67	36.46	36.52	29.98	29.71	29.42	22.64	22.37	21.55
CPU time	-	4h06	112h24	-	2h40	78h10	-	1h29	42h30
NFP	3.3%	3.5%	3.5%	3.3%	3.6%	4.1%	3.2%	3.2%	4.1%



$t_1^\star = t_2^\star = 20$cM



$t_1^\star = t_2^\star = 50$cM

Fig. 8 Signal $\sum_{i=1}^{I} \left(m^i_{t_i^\star}(t)\right)^2$ as a function the QTL location in each family ($I = 3$, $T = 0.60$, $K = 4$, $\forall k = 1,..,4 \quad t_k = 0.20(k-1)$, $\forall i\ \lambda = \lambda_i\sqrt{\pi_i} = 2$, $\sigma = 1$).

Table 14 Empirical Power as a function of the method used, the number I of families and the number nz of non zero λ_i's (npop=1,000, $n = 50I$, $\lambda = 2$, when $\lambda_i \neq 0 \quad \lambda_i = \lambda\sqrt{I}$, $\sigma = 1$).

I	nz	DF	DMC	Permutation
	14	99.7%	99.7%	99.7%
14	10	92.8%	93.1%	93.1%
	6	60.6%	61.7%	61.5%
	10	96.7%	96.8%	97.3%
10	6	70.7%	71.8%	72.8%
	3	24.72%	25.62%	26.72%
	6	82%	83%	84.9%
6	3	34.9%	37%	40.2%
	2	17.1%	18%	21.2%