Journal of Statistical Planning and Inference 153 (2014) 42–55

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

On empirical processes for quantitative trait locus mapping under the presence of a selective genotyping and an interference phenomenon

C.E. Rabier [a,b,*]

[a] University of Wisconsin-Madison, Statistic Department, Medical Science Center, 1300 University Avenue, Madison, WI 53706-1532, USA
[b] INRA, UR 875 Unité MIAT, F-31326, Castanet-Tolosan, France

ARTICLE INFO

Article history:
Received 21 December 2013
Received in revised form
20 January 2014
Accepted 29 May 2014
Available online 10 June 2014

Keywords:
QTL detection
Likelihood ratio test
Gaussian process
Selective genotyping
Interference phenomenon

ABSTRACT

We consider the likelihood ratio test (LRT) process related to the test of the absence of QTL (i.e. a gene with quantitative effect on a trait) on the interval $[0,T]$ representing a chromosome. The originality lies in the fact that we consider a selective genotyping (i.e. only the individuals with extreme phenotypes are genotyped) and an interference phenomenon (i.e. a recombination event inhibits the formation of another recombination event nearby). We show that, under the null hypothesis and contiguous alternatives, the LRT process is asymptotically the square of a "linear interpolated and normalized Gaussian process". We have an easy formula in order to compute the supremum of the square of this linear interpolated process. We prove that we have to genotype symmetrically and that the threshold is exactly the same as in the situation without selective genotyping and without interference.

1. Introduction

We study a backcross population: $A \times (A \times B)$, where A and B are purely homozygous lines and we address the problem of detecting a Quantitative Trait Locus, so-called QTL (a gene influencing a quantitative trait which is able to be measured) on a given chromosome. The trait is observed on n individuals (progenies) and we denote by Y_j, $j=1,\dots,n$, the observations, which we will assume to be Gaussian, independent and identically distributed (i.i.d.). The mechanism of genetics, or more precisely of meiosis, implies that among the two chromosomes of each individual, one is purely inherited from A while the other (the "recombined" one) consists of parts originated from A and parts originated from B, due to crossing-overs.

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 (see for instance Wu et al., 2007 or Siegmund and Yakir, 2007). K genetic markers are located at fixed locations $t_1=0<t_2<\cdots<t_K=T$. These markers will help us to find the QTL. $X(t_k)$ refers to the genetic information at marker k. For one individual, $X(t_k)$ takes the value $+1$ if, for example, the "recombined chromosome" is originated from A at location t_k and takes the value -1 if it is originated from B. We use the Haldane (1919) modeling for the genetic information at marker locations. It can be represented as follows: $X(0)$ is a random sign and $X(t_k)=X(0)(-1)^{N(t_k)}$, where $N(\cdot)$ is a standard Poisson process on $[0,T]$. Due to the independence of increments of the Poisson process, this model allows double recombinations between markers. For instance, if we consider three markers (i.e. $K=3$), we can have the scenario $X(t_1)=1$,

* Tel.: +1 608 265 9876; fax: +1 608 262 0032.
E-mail address: rabier@stat.wisc.edu

http://dx.doi.org/10.1016/j.jspi.2014.05.011
0378-3758/

$X(t_2)=-1$ and $X(t_3)=1$, which means that there has been a recombination between markers 1 and 2, and also a recombination between markers 2 and 3. Obviously, in the same way, we can have the scenario $X(t_1)=-1$, $X(t_2)=1$ and $X(t_3)=-1$.

A QTL is lying at an unknown position $t^\star$ between two genetic markers. $U(t^\star)$ is the genetic information at the QTL location. In the same way as for the genetic information at marker locations, $U(t^\star)$ takes value $+1$ if the "recombined chromosome" is originated from A at $t^\star$, and -1 if it is originated from B. In this study, inside the marker interval which contains the QTL, we will not consider the classical Haldane model (contrary to Chang et al., 2009; Azaïs et al., 2012), but we will focus on the model introduced by Rebaï et al. (1995) (see in particular their Section 2) in which double recombination between the QTL and its flanking markers is not allowed. As a consequence, under the model considered by Rebaï et al. (1995), if the QTL is lying for instance between the first two markers (i.e. $t^\star\in]t_1,t_2[$), we cannot have the scenario $X(t_1)=1$, $U(t^\star)=-1$ and $X(t_2)=1$. Indeed, this would have supposed that there had been a recombination between the first marker and the QTL, and also a recombination between the second marker and the QTL. In particular, the model considers that if we have a recombination between the QTL and one of its flanking marker, we could not have a recombination between the QTL and the other flanking marker. In other words, if $X(t_1)=1$ and $U(t^\star)=-1$, then we have automatically $X(t_2)=-1$. In the same way, if $X(t_2)=1$ and $U(t^\star)=-1$, then we have automatically $X(t_1)=-1$. Using a particular choice for the recombination probabilities between the QTL and the markers, it can be proved that the law of $U(t^\star)$ given its flanking markers (still assuming that they are located at t_1 and t_2) is the following (see Section 2 for details):

$$\mathbb{P}\{U(t^\star)=1|X(t_1),X(t_2)\}=\begin{cases}1 & \text{if } X(t_1)=1 \text{ and } X(t_2)=1\\ \dfrac{t_2-t^\star}{t_2-t_1} & \text{if } X(t_1)=1 \text{ and } X(t_2)=-1\\ \dfrac{t^\star-t_1}{t_2-t_1} & \text{if } X(t_1)=-1 \text{ and } X(t_2)=1\\ 0 & \text{if } X(t_1)=-1 \text{ and } X(t_2)=-1.\end{cases} \tag{1}$$

Note that when the distance between $t^\star$ and t_1 (resp. t_2) increases, it is more likely to have one recombination between the QTL and the first (resp. second) marker.

This way, in this study, inside the marker interval which contains the QTL, we model the interference phenomenon: a recombination event inhibits the formation of another recombination event nearby (see for instance McPeek and Speed, 1995). This phenomenon was noticed a long time ago by geneticists working on the Drosophila (Sturtevant, 1915; Muller, 1916). I refer to my recent study Rabier (2014b) where I largely describe the relevance of the interference model inside the marker interval, and the use of the classical Haldane model at marker locations.

We assume an "analysis of variance model" for the quantitative trait

$$Y=\mu+U(t^\star)q+\sigma\varepsilon \tag{2}$$

where ε is a Gaussian white noise.

Usually, in the problem of detecting a QTL on a chromosome, the genome information is available only at fixed locations $t_1=0<t_2<\cdots<t_K=T$, called genetic markers. So, usually an observation is

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

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

In this study, we consider the classical problem, but this time, in order to reduce the costs of genotyping, a selective genotyping has been performed: we consider two real thresholds S_- and S_+, with $S_-\le S_+$ and we genotype (i.e. we collect the genome information at markers) if and only if the phenotype Y is extreme, that is to say $Y\le S_-$ or $Y\ge S_+$. If we call $\tilde{X}(t)$ the random variable defined in the following way:

$$\tilde{X}(t)=X(t)1_{Y\notin[S_-,S_+]}$$

then, in our problem, one observation will be now

$$(Y,\tilde{X}(t_1),\ldots,\tilde{X}(t_K)).$$

We will observe n observations $(Y_j,\tilde{X}_j(t_1),\ldots,\tilde{X}_j(t_K))$ i.i.d.

It can be proved that $(Y,\tilde{X}(t_1),\ldots,\tilde{X}(t_K))$ obeys to a mixture model with known weights, times a function $g(\cdot)$ (fully given in Section 2) which does not depend of the parameters μ, q and σ

$$\Big[p(t^*)f_{(\mu+q,\sigma)}(Y)1_{Y\notin[S_-,S_+]}+\{1-p(t^*)\}f_{(\mu-q,\sigma)}(Y)1_{Y\notin[S_-,S_+]} +\frac{1}{2}f_{(\mu+q,\sigma)}(Y)1_{Y\in[S_-,S_+]}+\frac{1}{2}f_{(\mu-q,\sigma)}(Y)1_{Y\in[S_-,S_+]}\Big]g(\cdot) \tag{3}$$

where $f_{(m,\sigma)}$ is the Gaussian density with parameters (m,σ) and where the function $p(t^\star)$ is the conditional probability that $U(t^\star)=1$ conditionally on the flanking markers (cf. formula (1) if the flanking markers are located at t_1 and t_2).