SOME STATISTICAL PROBLEMS INVOLVED BY THE DETECTION OF QUANTITATIVE TRAIT LOCI IN ANIMAL GENETICS

by

M. San Cristobal-Gaudy*,** and C. Chevalet**

July 1993

Abstract

QTL (quantitative trait loci) are a finite number of loci that determine the expression of a quantitative trait (such as height or weight). QTL detection is a fashionable topic, which is one of the first applications of genetic maps. These maps are elaborated nowadays. Experimental designs are different in human, plant, or animal genetics. Statistical methods have to make use of inbreeding information and of available molecular markers of the map. Those methods should not require too much calculation, should not be too simplistic, should be powerful and have an acceptable robustness. For instance, a question that arises is to decide whether the means in a gaussian mixture are the same or not.

Keywords: genetics, genetic maps, statistical mixtures

* Institute of Statistics, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.
** Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique, Castanet Tolosan, France.

1. Introduction

The recent developments of biotechnologies have opened a large amount of interesting features in several domains, such as medicine, plant or animal breeding. In the latter cases, the aim is to improve a breed for, *e.g.* disease resistance or meat production.

The knowledge in molecular biology and cellular genetics has served to construct new probabilistic and statistical models, or to improve existing models. Moreover, the results obtained using statistical tools permit reducing the work of biologists and geneticists. For instance, genetic maps are the basis for locating major genes or genes with smaller effects such as those of "quantitative trait loci" (QTL), with statistical methods; then, the corresponding physical genes can be looked for in a restricted region of a particular chromosome, not anymore in the entire genome.

Statistical methods dealing with gene location using genic maps are highly related to experimental designs. Data in medicine, plant and animal breeding differ in nature, so that methods developed in one area are to be used with care in another area. Nevertheless, the main problem remains that of testing whether the distribution of the data is a mixture of two gaussians or not. This question is particularly difficult to answer when the means of the two components are close, which is the usual case in QTL detection.

This paper is partitioned into three parts. Some basic genetic concepts are explained in section 2. Related probabilistic and statistical models are then presented in section 3. Finally, an overview of QTL detection is given in section 4, including a short bibliography and future work perspectives.

2. Some genetic background

All living individuals are built up with cells. Each cell is a complex unit of several specific biological mecanisms, but a kernel is always present. This kernel includes *chromosomes* that contain information : the double helix macromolecule, *DNA* (DesoxyriboNucleic Acid), in which the *genetic code* is encoded. The structure of the DNA is very complex, but is summarized, for our purpose, as follows: the DNA can be represented by two complementary and oriented sequences of four bases, the thymin (T), the adenosin (A), the guanin (G), and the cytosin (C). These bases are paired: a A (resp. C) is always related with a T (resp. G). Hence, a particular sequence of DNA base pairs is given by the sequence of bases on only one side of the molecule (for instance: ATTCAGGCGT, since the corresponding DNA piece is

...AATTCAGGCGT...

...TTAAGTCCGCA...

A sequence responsible for the production of some substance (a protein) is named a *gene*. At the same location (or *locus*) on the DNA molecule, another individual can carry another gene: a modified protein is produced, or none. This matter of fact is called *polymorphism*, at the considered locus. Not all the loci are polymorphic.

In the genome of usual species, there exists loci, that generally do not code for proteins, which sequence is TGTG...TG, n times. If n varies from individual to individual, this polymorphism is interesting, since systematic biotechnological techniques for finding the length of this sequence are available. The particularly interesting situations are those where 2 to 10 variants are encountered in the working species. These non codant loci (called microsatellites) serve as *markers* along the genome. Other types of molecular markers exists: minisatellites (the repeated motif is longer than two base pairs), RFLP (restriction fragment length polymorphism), etc.. These molecular genetic markers are more numerous and polymorphic than phenotypic markers (colour of skin, form of horns) and than blood groups.

Most of our cells are such that the chromosomes are paired; one talks about 2N-cells. One chromosome is inherited from the sire (or father) and the "sister-chromosome" is inherited from the dam (or mother). During meiosis, where gametes (N-cells) are produced, *crossing-overs* arise. A crossing-over is in fact an exchange of genetic material, and is represented in figure 1. The DNA of one chromosome crosses the DNA of the sister chromosome, and a cut occurs at the junction point. Thousands of different gametes are produced in this way. When a male gamete (spermatozoid) and a female gamete (ovule) fuse and give an egg, a new individual is created, who will be different from his sisters and brothers.

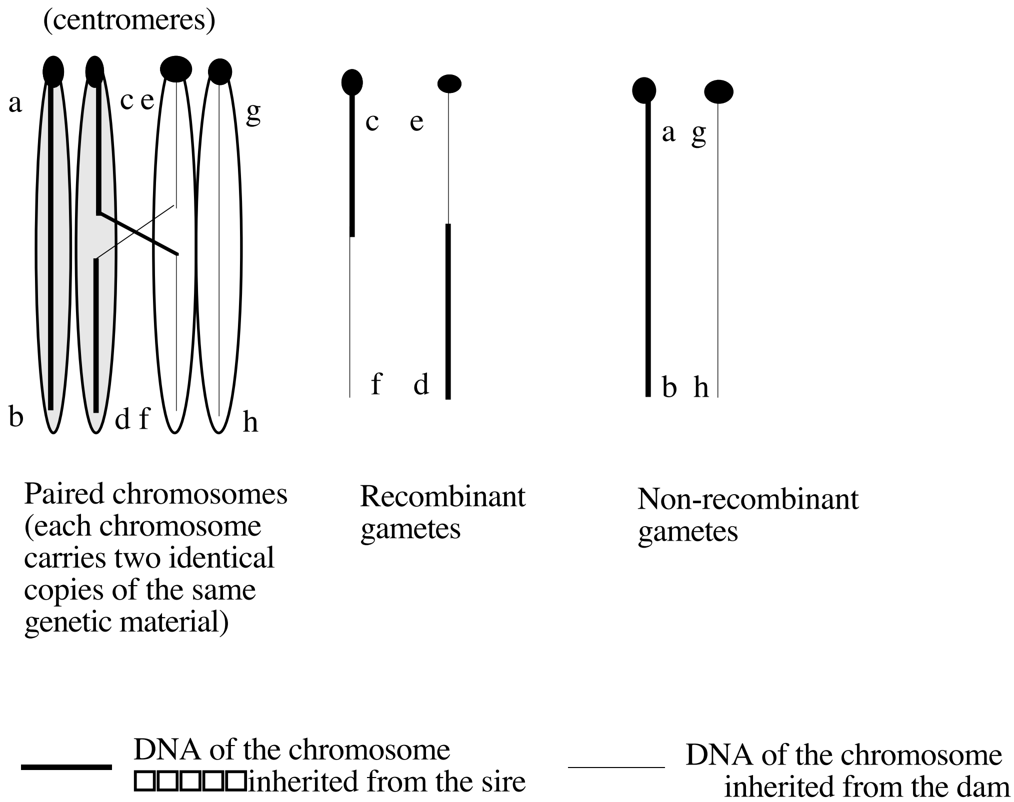


Figure 1. Scheme of a crossing-over arising in a small region of paired chromosomes.

3. Related probabilistic and statistical tools

3.1. Background

Since the causes of the phenomenon are largely unknown, the stochastic nature of the occurrence of crossing-overs is modeled in the following simple way.The random variable X , representing the number of occurrences of crossing-overs between two loci of the genome, is distributed according to a Poisson distribution with parameter d:

$$P(X=k) = d^k e^{-d} / k! \qquad (1)$$

where d is the distance between the two considered loci. The unit of this distance is the Morgan (M). Using this distance, one wants to construct a linear map along a line representing the DNA of the whole chromosome, and where points are loci. This map is called the *genetical map*.

It is in fact impossible to measure d directly. New concepts are useful. Two loci are called *recombinant* when an odd number of crossing-overs occurred during meiosis between them. This means that the region of the chromosome near the first locus is inherited from the father, while the region of the chromosome near the second locus is inherited from the mother - or the contrary. The recombination rate r is the probability that a recombination occurs between the two loci, *i.e.* is equal to the probability of an odd number of crossing-overs:

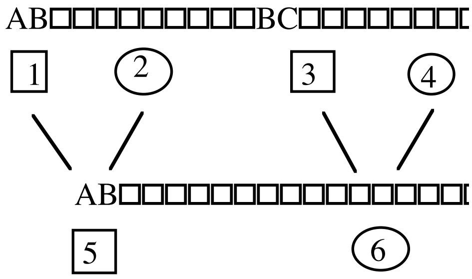
$$r = \sum_{k=0}^{+\infty} \mathrm{P}(X = 2k+1) = \frac{1}{2}(1 - \mathrm{e}^{-2d}) \qquad (2)$$

An estimation of r provides an estimation of the desired distance between the two considered loci. The relation (2) between r and d is known as the Haldane mapping function. According to this relation, we will in the sequel make no real distinction between r and d .

r is a real number lying between 0 and 1/2. It is equal to 0 when the two loci are transmitted always together to the progeny: the loci are genetically totally linked. On the contrary, when the probability that the two loci are transmitted together is 1/2 (if there are located on two separate chromosomes for example), then they are genetically independent. The parameter r quantifies the *genetic linkage between two loci.* They are linked if r is less than 1/2.

Let us consider a locus L and a population of individuals. Suppose that two genes, say A and B, are present at this locus in this population. An individual has a *genotype* AA, BB, or AB. In the two former cases, the individual is said to be *homozygous*. In the latter, he is *heterozygous*. Only *phenotypes* can be observed: [AA], [AB] and [BB]. For instance, the phenotype [AA] corresponds to white animals, [BB] to black animals, and [AB] to black and white animals. If so, the underlying genetic model is said to be *codominant*: the heterozygotes are different from the two homozygous types. The genetic model is totally *dominant* if, say, [AB]=[BB], *i.e.* the heterozygotes are black like the BB-homozygotes.

The biological definition of a gene is a functional one: for example a protein is produced with some biochemical properties. Two genes are considered to be alike if the corresponding properties are the same; on the contrary two homologous genes (*i.e.* at the same locus) are said to be *alleles* if the proteins differ in some trait. Population geneticists have developed another classification of genes, independent of their biological effects: the *identity by descent* (ibd) relationship. Two genes are ibd if they are copies of a common ancestor gene. Let us consider the following pedigree



where A, B and C denote genes at a locus L. If no common ancestor exists for the animals 1, 2, 3, 4, one can say that, *e.g.* the A gene of 5 is identical by descent (ibd) with the A gene of 1, and the B gene of 5 is ibd with the B gene of 2. The B gene of 6 is ibd with the B gene of 3 with probability 1/2, and with the B gene of 4 with probability 1/2. Different genes (*i.e.* that are not ibd) are called *alleles*. The B of 5 and the B of 1 are alleles in the ibd sense. This classification allows to make statements about the genetic structures of populations without any reference to observations of genotypes at some loci. It is the main genetic tool used in models by quantitative geneticists.

3.2. Maps

It was seen previously that a parallel was tempted between the molecular genetics level, and the population genetics level (dealing with probability, ibd relations between genes). In the former case, genes have a functional definition and are referred to as sequences of base pairs, loci are mapped on chromosomes, and alleles at these loci are listed. Distances between loci are base pairs. Tools for this *physical mapping* include cytogenetics (in situ hybridization of labelled gene sequences allows their rough location to be visualized), the ordering of a set of overlapping cloned sequences (as it has been recently achieved for complete arms of human chromosomes Y and 21) and ultimately the complete sequence of chromosomes (a goal already reached in the case of yeast chromosome III).

On the population genetics level, distances are expressed in Morgans, estimated with statistical tools (via the recombination rate). This procedure is the *genetical mapping*. The two maps are related, since the ordering of loci is the same, although relative distances are different. The union of the physical map and the genetical map is the *genic map*. Because the human genome has estimated length 3.10^9 base pairs (bp) on one hand and 30 Morgans ($1M=3.10^3cM$) on the other hand, 1cM represents more or less 10^6 bp.

In practice, it is useful to group genetically related loci, and to locate one of these loci on the physical genome. Then such related loci are on the same chromosome as the physically located locus.

Genic maps can have several uses (see for instance: Commission of the European Communities, 1992). The first one is the possibility to follow the transmission of a gene (G) responsible for big differences between individuals (*major gene*). If a marker (M) is tightly linked to the interesting gene, both are transmitted together to the progeny in most of the cases. If r is the recombination rate between the two loci where the marker gene M and the other gene G are located, then G and M are transmitted together with probability $1-r$, which is close to 1 if r is near 0. The gene responsible for the disease "cystic fibrosis" was located by using this procedure iteratively (Kitsis, Warren, Kaplan, 1988,).

In the case of quantitative traits like height or weight, the "infinitesimal model" was used by quantitative geneticists as a "black box" before the development of genic maps: an infinite number of loci are contributing to the trait with infinitesimal contributions. The observed trait is then assumed gaussian. But it is intuitively more appealing to suppose that a finite number of loci determine a genotype relative to the quantitative phenotype. These loci influencing the quantitative trait of interest are called *QTL* (quantitative trait loci). QTL detection is of major interest in agronomical research and the availability of genic maps should allow these QTL to be located, at least those with the greatest variation between alleles. Marker assisted selection may be more efficient than phenotypic selection (Lande and Thompson, 1990). One can also use the results of QTL detection to introgress one or several interesting genes into a line, without perturbing the rest of the genome. Statistical methods for detecting QTL will be discussed in Section 4.

Finally, maps of different species can be compared, since entire regions of the genome of one species are similar to other regions of another species. For example, the parallel between the human and the pig maps provides an animal model for malignant hyperthermia, because of the similarity of the chromosome around this gene and the pig Halothane gene, related to stress sensitivity (MacLennan, Phillips1992). It is now well-known that pigs which die during an anaesthesia with the halothane gas are also, in general, highly sensitive to stress and even can die if too stressed. The symptoms occurring with such pigs and those arising sometimes during anaesthesia for humans (malignant hyperthermia) are similar. The similarity of human and pig maps around these genes and a common mutation explaining both syndromes prove how powerful comparative mapping may be for elucidating mammalian genomes organization.

Statistical designs for mapping are usually similar as those for QTL detection. These designs are different if one is working in plant, human or animal genetics. The situation in plant breeding is the most favorable, one in the sense that totally inbred lines are available. This means that individuals of parental lines are homozygous at all marker and QTL loci, any allele being characteristic of one single parental line so that associations between alleles at different loci (*phase*) are known in the parental line. This most convenient situation is encountered in practice in tomato, wheat, etc... , but not in every plant species, as well as in some mouse lines.

In the human and domestic animal genetics, the phase is usually unknown: an A gene at locus L is related to an A' gene at locus L' for some individuals of the parental line, while this A at locus L is sometimes encountered with a B' gene at locus L', for the same parental line. Calculations required to built *e.g.* a likelihood, are hence more complicated since the number of possible situations (genotypes at a given locus, for the progeny) is huge.

Because of the small family size arising in human genetics, methods such as pedigree analysis were greatly developed. Unfortunately, they are not really appropriate in animal genetics because of the big family size. A method based on similarities between relatives can be extended from human (Haseman and Elston, 1972) to animal genetics (Götz and Ollivier, 1992), but lacks power.

From a pure statistical point of view, a great advantage of plant and animal designs over human designs is that they can be planned.

PiGMaP

Our main interest here is on animal breeding, and more specifically on the pig species. The elaboration of the pig map is a European project named PiGMaP (Pig Gene Mapping Project), included in the BRIDGE program devoted to genome study.

The design used to construct the map of the pig species, and afterwards to detect some QTL or major genes, is an F2 type (see figure 2).

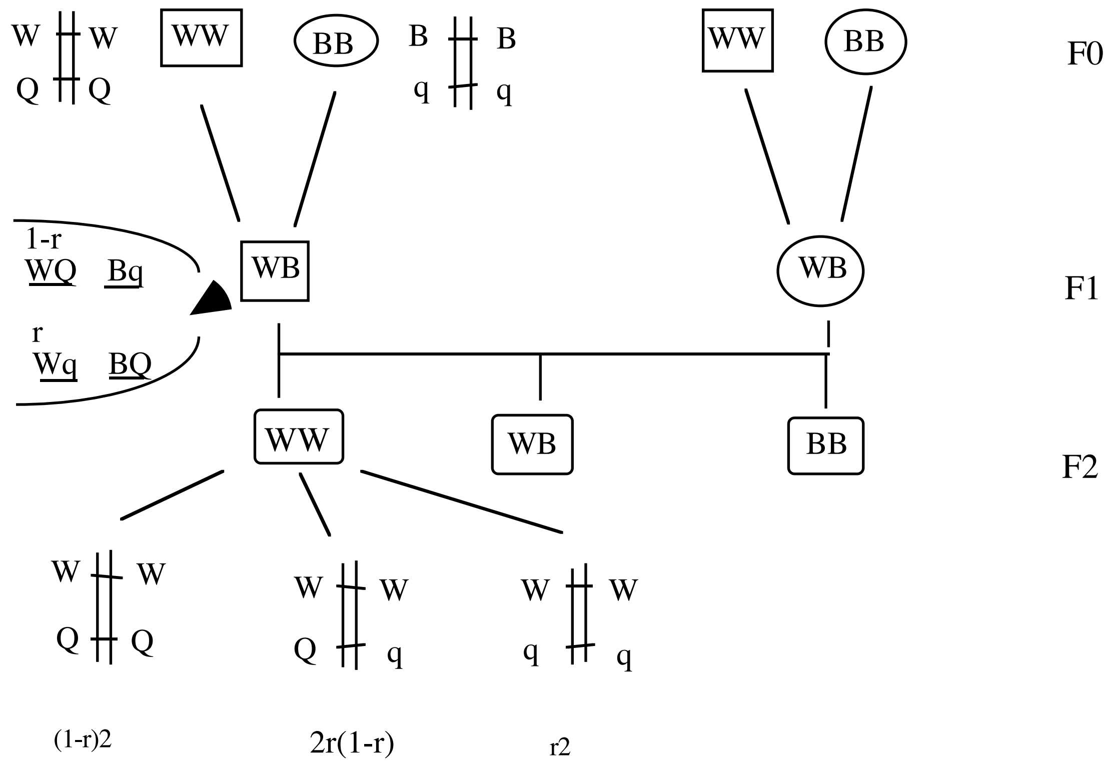


Figure 2. F2 breeding scheme used for mapping and QTL detection in pig.

The parental lines are the Large White (the most common pig in Europe) and the Meishan (a chinese pig). These species are phenotypically very different. Some zootechnical aspects are as following:

* the generation interval is relatively short, less than one year
* the prolificacy is in average 10 piglets per litter for european species, with 3 or 4 more piglets for chinese pigs
* the Large White has a rapid growth and a lean carcass compared to the Meishan; the chinese pig is more placid.

Concerning the map caracteristics, the pig has 18 pairs of chromosomes, plus the sexual ones (X and Y). The length of the genome is expected to be the same as the human: 3.10^9 bp and 30M. The aim is to find 150 markers (loci) equally spaced every 20 cM along the genome. These markers are mainly molecular markers, especially microsatellites.

In the absence of European project concerning QTL detection, the INRA (Institut National de la Recherche Agronomique) in France continues the breeding scheme and measures some traits which are candidates for QTL (Bidanel and Ollivier, 1992): psychological reactivity of piglets, resistance to neonatal diarrhoea, development of sexual function (females), growth, sexual odor of meat (males), and prolificacy components.

Six Large White males were mated with six Meishan females, giving six litters. Among these 72 piglets, six F1 males and 23 F1 females were chosen at random and crossed to give at least 500 F2 animals, on which the previous traits are being measured. 20 F2 animals provided the genetic material of the French contribution to PiGMaP.

In the F2 breeding scheme (figure 2), let W and B be the genes (alleles) at a marked locus. They can represent the colour of the skin of pigs: white and black, but also two lengths of a microsatellite. If the males of the parental lines are WW-homozygotes, and if the females are BB-homozygotes, then the F1 animals are all heterozygotes. But it can be seen that all types of animals are present in the F2 generation. The F2 scheme seems to be the optimal one for maximising the internal genetic variability and the between line variability.

Suppose further that a QTL is linked to the previous marker locus with recombination rate r. Let Q and q be the alleles at the QTL. For sake of simplicity, suppose that the parental lines have genotypes $\frac{\mathrm{WQ}}{\mathrm{WQ}}$ and $\frac{\mathrm{Bq}}{\mathrm{Bq}}$ respectively. The F1 individuals are $\frac{\mathrm{WQ}}{\mathrm{Bq}}$ and give two types of non recombinant gametes, WQ and Bq, with probability (1-r), and two types of recombinant gametes, Wq and BQ, with probability r. Then in the F2 population, several genotypes are present. For instance, the WW marked animals can have three different QTL genotypes. They can be non recombinant $\frac{\mathrm{WQ}}{\mathrm{WQ}}$ with probability $(1-r)^2$, recombinant $\frac{\mathrm{WQ}}{\mathrm{Wq}}$ with probability $2r(1-r)$, or double recombinant $\frac{\mathrm{Wq}}{\mathrm{Wq}}$ with probability r^2. If the putative QTL is segregating in the population, *i.e.* have an influence on the measured trait, then the distribution of the data is a mixture of three distributions. The density can be written as

$$\begin{aligned} \mathrm{f}(y) &= p_{QQ}\,\mathrm{h}_{QQ}(y) + p_{Qq}\,\mathrm{h}_{Qq}(y) + p_{qq}\,\mathrm{h}_{qq}(y) \\ &= \sum_{g\in G} \mathrm{P}(g|M)\,\mathrm{h}_g(y) \end{aligned} \tag{3}$$

where G is the finite set of all possible QTL genotypes g,

M is the marker genotype,

$\mathrm{h}_g(.)$ is the density of the data with QTL genotype g.

In most of the cases, it is assumed that the genotype g modifies the mean only:

$$\mathrm{h}_g(y) = \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(y-\mu_g)^2\} \tag{4}$$

Of course,

$P(g=QQ|\ M=WW)=(1-r)^2$
$P(g=Qq|\ M=WW)=2r(1-r)$
$P(g=qq|\ M=WW)=r^2$
etc...

If no QTL is segregating in the population, then all the genotypes g have the same influence on the measured trait, hence

$$f(y) = h_g(y) \qquad \forall\ g \in G \tag{5}$$

In the simplest case of gaussian distributions, the problem is to test whether the density of the data is a mixture of three gaussian densities, or is simply a gaussian density. If an additive genetic model with two alleles at the QTL is adopted ($\mu_{Qq} = (\mu_{QQ} + \mu_{qq})/2$), this last question is reduced to the decision of whether two or one gaussian distributions occur. This statistical problem is not simple, as it will be seen in the next section, wich also contains a short bibliography on that subject.

4. QTL detection

4.1. Available results

ANOVA methods

The simplest way to detect QTL is to use ANOVA methods. Their great advantage is the simplicity of the model. Suppose for instance that one wants to evaluate the effect of the allele j at the marker locus, in the progeny of sire i:

$$y_{ijk} = u_i + a_{ij} + e_{ijk}$$

where u_i is the effect of sire i
a_{ij} is the effect of allele j out of sire i
e_{ijk} is the residual effect.

LeRoy and Elsen (1991) showed that the test of equality of intra-sire marker effects can be interpretated in some cases by the absence of QTL, and in some other cases by the absence of linkage between the QTL and the marker locus.

Likelihood methods

Likelihood methods are more sophisticated. They take into account in the model the effects of the QTL genes on the trait, on one hand, and the recombination rate *r* between the putative QTL and one or two marker loci on the other hand. The power of the likelihood ratio test (LRT) increases when two flanking markers are considered.

The likelihood has the form

$$L = \prod_{i=1}^{n} \sum_{g \in G} \mathrm{P}(g|M_i)\ \mathrm{h}_g(y_i) \qquad (6)$$

where M_i is the marker genotype of individual i

G is the finite set of all possible QTL genotypes g

y_i is the observation of the random variable Y_i

$\mathrm{h}_g(.)$ is the probability density function of Y_i having QTL genotype g.

Often,

$$\mathrm{h}_g(y) = \phi\left(\frac{y - \mu_g}{\sigma_g}\right) \qquad (7)$$

where $\phi(.)$ is the standard normal density function.

If $|G|=2$, the null hypothesis is

$$\mathrm{H}_0\colon\ \mathrm{f}(y) = \phi\left(\frac{y - \mu}{\sigma}\right) \qquad (8)$$

while the alternative is

$$\mathrm{H}_1\colon\ \mathrm{f}(y) = p\ \phi\left(\frac{y - \mu_1}{\sigma_1}\right)\ +\ (1 - p)\ \phi\left(\frac{y - \mu_2}{\sigma_2}\right) \qquad (9)$$

with $p \in]0,1[$, $\mu_1 \neq \mu_2$ or $\sigma_1 \neq \sigma_2$.

As Titterington *et al* . (1985) pointed out, the LRT of H_1 *vs* H_0 can be viewed as:

* H'_0 *vs* H_1, where H'_0: $\mu_1=\mu_2$ and $\sigma_1=\sigma_2$. Hence, two degrees of freedom appear in the test. But under H'_0, the nuisance parameter p is not estimable; see Davies (1977) for some proposal in this case, unfortunately not directly interesting here.

* H''_0 *vs* H_1, where H''_0: p=0 (or p=1). Here, one degree of freedom appear. But the parameter p lies on the boundary of the parameter space.

In both cases, the standard asymptotic theory is not valid.

Pons (1993), following the works of Redner (1981), Ghosh and Sen (1985) and Self and Liang (1987), has considered the following problem:

$$\begin{aligned} &\mathrm{H}_1\colon\ \mathrm{f}(y,\theta) = (1-p)\ \mathrm{g}(y,\theta_1) \ +\ p\ \mathrm{g}(y,\theta_2) \\ &\qquad p \neq 0 \qquad p \neq 1 \qquad \theta_1 < \theta_2 \\ &\qquad \Omega = [0,1] \times \Re^d \times \Re^d \\ &\mathrm{H}_0\colon\ \mathrm{f}(y,\theta) = \mathrm{g}(y,\theta_1) \end{aligned} \tag{10}$$

The parameters are identifiables under H_1, not under H_0. She showed that the LRT statistic converges in law:

$$\lambda_n = 2[\sup_{\mathrm{H}_1} l_n(\theta) \ -\ \sup_{\mathrm{H}_0} l_n(\theta)] \underset{n\to+\infty}{\to} \quad \lambda = \sup_{\alpha\in\Re_+^d} T^2(\alpha)\ \mathrm{I}_{(T(\alpha)>0)} \tag{11}$$

where $l_n(\theta)$ is the log likelihood based on the sample (Y_1,... Y_n) and parameter θ = (p, θ_1, θ_2). T is a gaussian process with mean 0 and covariance function depending only on f. This result was extended to the situation of n families with m halfsibs each. In this case, $\lambda_{n,m}$ converges weakly to the distribution of

$$\sup_{\alpha\in\Theta} \{T^2(\alpha) + U^2(\alpha) + V^2(\alpha)\}\ \mathrm{I}_{(T(\alpha)>0,\ U(\alpha)>0,\ V(\alpha)>0)} \tag{12}$$

where T, U, V are independent gaussian processes depending only on the density f, and Θ is defined in Pons (1993).

In practice, simulations are needed to have a test significance level.

Lander and Botstein (1989) proposed also a LRT for detecting a QTL, but the presentation is slightly different. In each interval between two adjacent markers on the genetical map, the test statistic is evaluated at each point r giving a curve $\{\lambda_{l,n}(r)\}$ for the lth interval, with $r \in [0, r_{0,l}]$, where $r_{0,l}$ is the known recombination rate between the markers defining the lth interval. An approximate significance level is given as an asymptotic result when the map is becoming dense ($r_{0,l} \to 0, \forall l$). In this case, the process evaluated by the curve $\bigcup_l \{\lambda_{l,n}(.)\}$ converges to a Orenstein-Uhlenbeck diffusion process. This approach is available for inbred lines and was used with success in tomato genetics.

Remarks

In the preceding likelihood models, both p and (μ, σ) parameters were estimated. An estimate of p is interesting in the aim of mapping, where the exact location of the QTL is wanted. For the feature of marker assisted selection, the effect of each QTL allele is important, with a nereby marker locus. No precise estimate of the distance between the two loci is necessary.

If the proportion parameters of the mixture are fixed (Goffinet *et al.*, 1992), or if an a priori is defined (Aitkin and Rubin, 1985), when the inference is based on the mixture parameters (μ, σ), then the empirical information matrix is not positive definite with probability different from 1 as n tends towards infinity (Quinn *et al.*, 1987). Hence, even in this simple case, the standard asymptotic theory is not applicable.

4.2. Possible application of likelihood methods to PiGMaP

Let us consider the PiGMaP scheme, and assume that a map is given (and known). Because all the animals are typed, the transmission of the marker genes at each locus can be followed. Hence, for each marker locus, it is possible to say that the neighborhood of a marker locus of a particular F2 animal is inherited from a precise F0 animal, or at least with fixed probability. This is illustrated by figure 3.

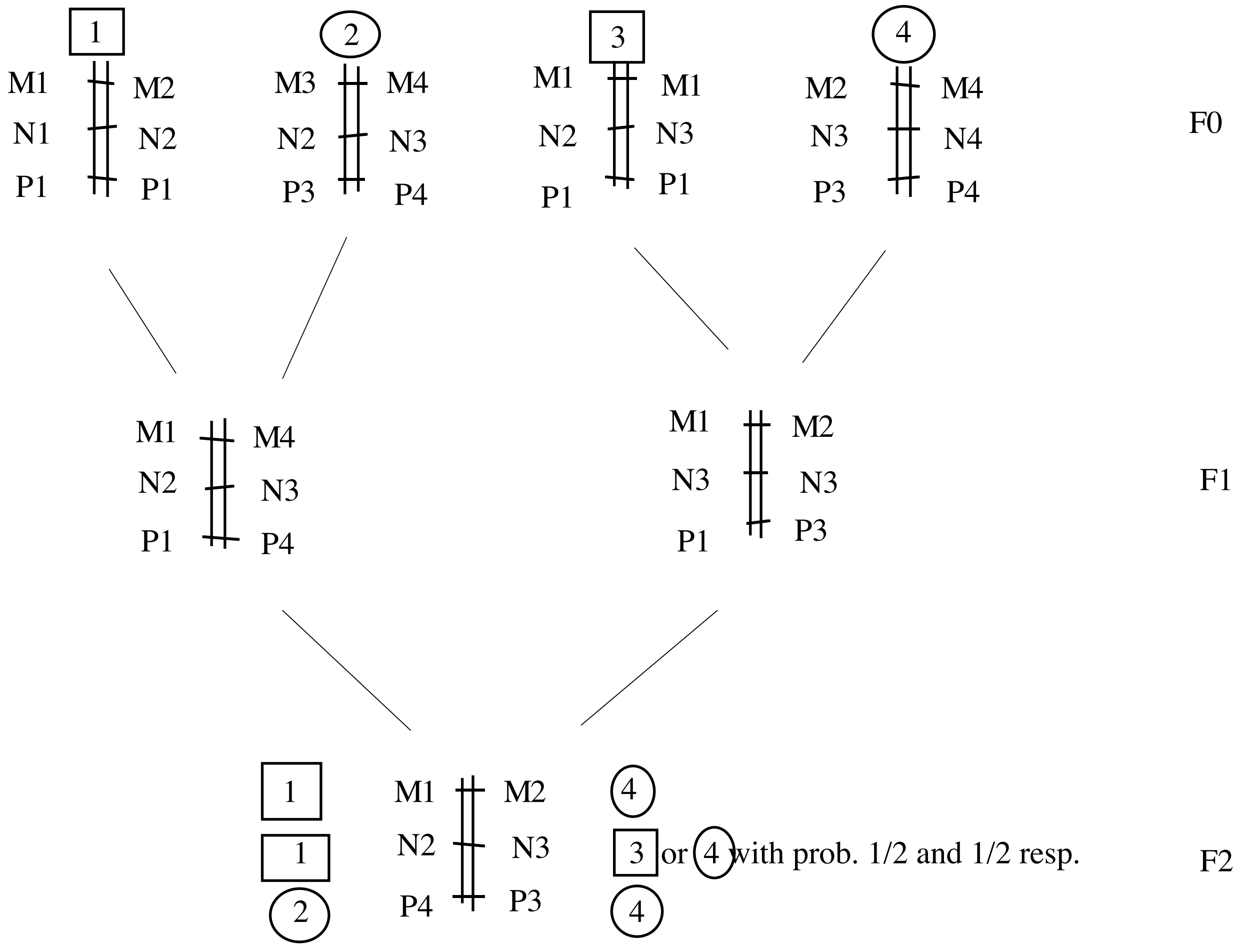


Figure 3. The transmission of marker genes can be followed through the F2 scheme. Here, three loci are considered.

Further, let us assume a naively simple genetical model, with two alleles at the QTL, say Q and q, Q for the European line (animals number 1, 3, ...), and q for the Chinese line (animals number 2, 4, ...). The QTL is supposed to affect only the mean of the trait; let μ_g be the trait mean relative to the QTL genotype g. We first have a look at the case where the QTL is in the neighborhood of a marker locus (r=0). The map is hence supposed to be dense enough. Notice that the marker locus and the QTL are indistinguishable here.

The statistical model is as follows. Let us consider an i.i.d. sample ($Y_{l,i}$, $A_{l,i}$, $B_{l,i}$), for i=1,...n, from the population r.v. (Y, A, B). The index l stands for the considered locus and will be omitted sometimes. The rv Y is continuous with conditional density function $f_{Y|A,B}$:

$$f_{Y|A=a,B=b}(y;\theta) = a\phi(y-\mu_{QQ}) + b\phi(y-\mu_{qq}) + (1-a-b)\phi(y-\mu_{Qq}) \qquad (13)$$

At each locus and for each individual of the scheme, a probability of beeing QQ-genotype (resp. qq-genotype) exists. The possible values of these probabilities are in finite number. These probabilities are the consequences of the information given by the transmission of markers through the F2-scheme. Let $\{s_{A1}, s_{A2}, \ldots, s_{Ak}\}$ be the finite set of values of A, with $0 = s_{A1} < s_{A2} < \ldots < s_{Ak} = 1$. It can be assumed that the distribution of $\underline{A}$ is multinomial $M(1; \{\alpha_{A1}, \alpha_{A2}, \ldots, \alpha_{Ak}\})$, where $\underline{A} = [\mathrm{I}\{A = s_{A1}\}, \ldots, \mathrm{I}\{A = s_{Ak}\}]$. The same assumption is made for B: $\underline{B} \sim M(1; \{\alpha_{B1}, \alpha_{B2}, \ldots, \alpha_{Bk'}\})$

$\theta = \theta_l = (\mu_{l,QQ}, \mu_{l,qq}, \mu_{l,Qq})$ in (13) is the parameter to be estimated. Remark that this parameterization is equivalent to $\mu_{QQ} = \mu - a$, $\mu_{qq} = \mu + a$, $\mu_{Qq} = \mu + d$, where μ is the common mean, a is the additive effect and d the dominance effect.

For each locus, a LRT is worked out to test whether there exists a QTL segregating and tighly linked to the marker locus l, or if there is no QTL:

$$\mathrm{H}_{l,0}\colon \mu_{l,QQ} = \mu_{l,qq} = \mu_{l,Qq} \tag{14}$$

The test statistics $\lambda_{l,n}$ converges in law to a χ_1^2.

The question is then how to detect the first QTL (the QTL with the greatest effect $|\mu_{l,QQ} - \mu_{l,qq}|$ on the trait), taking into account the fact that $\lambda_{l,n}$ and $\lambda_{l',n}$, $l \neq l'$, are correlated? Then how can one detect the second QTL, given the first one?

A second interesting issue would be to introduce nonparametric components in the model. This in order to avoid non robustness of the LRT with respect to departure from normality. For instance, a semiparametric model could be considered, with the standard normal $\phi(.)$ replaced by an unknown density.

Of course, it must be kept in mind that some extensions are needed to be closer to the genetical model that the geneticists believe to be acceptable:

* the random variables X_i in (13) could first be considered as independent, for sake of simplicity. But F2 animals are not independent because of the family structure: full sib and half sib relationships exist in the design (see explanations of the PiGMaP scheme). Quantitative geneticists believe in the importance of these family relationships. A known variance-covariance matrix of $X=\{X_i\}$ must hence be incorporated in the statistical model.

* the rv A_i and B_i are not independent either, in , say, 5-10% of the cases

* the number of QTL alleles is equal to twice the number of founders (F0 animals): one for each founder chromosome

* use of "interval mapping", *i.e.* method inspired by Lander and Botstein (1989) in constructing a curve for each interval of flanking markers, at each possible location of the QTL in this interval.

5. Further discussion

In the second and third sections, some genetic concepts were exposed, from the biological and from the mathematical point of view. Here, quantitative geneticists can improve their models concerned with qualitative or quantitative data by introducing determinant genes at given loci. Marker assisted selection is more efficient than usual selection. Statistical problems arising in QTL detection are very interesting for statisticians, since they are quite general (mixtures of distributions is an old problem) and difficult to solve quickly. A wide research area is still open. Geneticists can also profite the results of biometricians in limiting their search for a major gene in a very restricted part of the physical map. All this nice things are possible because of the parallel between the known biological mecanisms and the probabilistic modelling of this knowledge.

PiGMaP is a good example of common work: of quantitative geneticists for the breeding scheme, of qualified personnel to collect blood from pigs and extract DNA from the blood cells, of cellular geneticists to provide and use methods for finding molecular markers, of biometricians to order the loci and build the genetical map, of cytogeneticists to locate some genes on the physical map.

Multilocus problems

The still unresolved questions exposed at the end of Section 4 are consequences of "monolocus" problems: for each marker locus of the map, a statistical test is worked out to decide whether a QTL located in the neighborhood of the marker locus is segregating, or not. The method of Lander and Botstein (1989) considered successive intervals of flanking markers. A "window" of m markers could be envisioned, as a "multilocus" problem, with extreme cases the whole genome (segregation analysis) and the single markers, as already mentioned.

MultiQTL problems

It is even more realistic to extend the "monoQTL" case to the "multiQTL" situation, where at least two QTL are segregating in the considered population. In the case of two QTL, Haley and Knott (1992) showed by simulation that a three-dimensional surface is needed to avoid placing the QTL in the same position. Apart from the computational aspects that increase rapidly, it is of course even more difficult to decide whether two, one, or zero QTL are segregating, and quite impossible to have a (good) estimate of the number of segregating QTL.

Use of covariates

In model (13), only the QTL effect was taken into account, at a first step. A more satisfying model is this with μ_g (g=QQ, qq, Qq) replaced by

$$\mu_i = x_i'\beta + u_g = x_i'\beta + z_i'u \qquad (15)$$

where x_i and z_i are incidence vectors,

β is a vector of parameters related to "environmental" factors (sex, region of birth, season of birth, ...),

u={u_g}, $g \in$ G, is the vector of effects of QTL alleles. Its dimension is 3 in (13), but can be equal to twice the number of founders. In the latter case, this factor could be treated at random:

$$u \sim N(0, \sigma_u^2) \qquad (16)$$

The null hypothesis of no segregating QTL is then $\mathrm{H}_0: u_g = u_0 \quad \forall g$, for the fixed effect case, and $\mathrm{H}_0': \sigma_u^2 = 0$, for the random effect case.

The preceding ideas raise to the model choice problem. As usually, a mid term situation is to be chosen. Simple models or methods are easy to implement, the computational time is short, some properties are known, like asymptotic properties. These nice aspects are not shared in general by more sophisticated methods or models. But, as counterpart, a more satisfying genetic model is taken into account, and the power is increased.

Experimental designs

Proposals are for the moment related to designs. The experimental designs are quite different if one is interested in plant, human or animal genetics. Also for animal species designs are not the same, due to zootechnical constraints such as the generation interval. The first step is to develop *ad hoc* methods for each kind of design. It seems that the second step is far away, where a unifying theory would be given.

Optimisation of experimental designs was not discussed in this paper. Simulations have shown that the F2 scheme is a good one. In the particular situation of QTL detection in pig, some questions remain: is it preferable to increase the number of markers of the map or the number of animals? Is it preferable to have big families of little size, or few families of big size? In which proportion? These questions have been given intuitive responses until now.

To conclude, QTL detection is possible nowadays because of the great developments of biotechnologies. This domain lies in the intersection between probability, statistics, quantitative genetics, cellular genetics, molecular biology. Statistical problems are difficult, but general and highly interesting.

Acknowledgements

The first author would like to thank Jean-Michel ELSEN, Bruno GOFFINET, Brigitte MANGIN (INRA Toulouse), François RODOLPHE, Odile PONS, Didier BOICHARD, Pascale LE ROY, Jean-Pierre BIDANEL and Jean-Louis FOULLEY (INRA Jouy-en-Josas) for helpful discussions on this new topic. The scientists of the Laboratoire de Génétique Cellulaire are greatly acknowledged for their kind explanations of molecular genetics. The authors are greatful to Irène GIJBELS and André BOUCKAERT (UCL, Louvain-la-Neuve) for interesting suggestions and improvements of the paper.

References

Aitkin M. and Rubin D.B. (1985). Estimation and hypothesis testing in finite mixture models. Journal of the Royal Statistical Society Series B 47, 67-75.

Bidanel J.P. and Ollivier L. (1992). Projet PiGMaP. 2èmepartie. Détection de locus à effets quantitatifs. Expérience n°1 - Rouillé - Le Magneraud 1992.

Chevalet C., Gillois M. and Vu Tien Khang J. (1984). Conditional probabilities of identity of genes at a locus linked to a marker. Génétique Sélection Evolution 16, 431-444.

Commission of the European Communities; European Laboratory Without Walls (1992). Le projet de cartographie génétique du porc (Pig gene mapping project: PiGMaP). Bridge.

Davies R.B. (1977). Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64, 247-254.

Ghosh J.K. and Sen P.K. (1985). On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. In Proceedings of the Berkeley conference in honor of J Neyman and J Kiefer, Vol 11. Le Cam, Olshen eds.

Goffinet B., Loisel P. and Laurent B. (1992). Testing mixture models when the proportions are known. Biometrika 79,

Götz K.U. and Ollivier L. (1992). Theoretical aspects of applying sib-pair linkage tests to livestock species. Genet Sel Evol 24, 29-42.

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

Haseman J.K. and Elston R.C. (1972). The investigation of linkage between a quantitative trait and a marqueur locus. Behaviour Genetics 3, 3-19.

Kitsis, A., Warren P. and Kaplan J.C. (1988). Génétique inverse et mucoviscidose. Medicine Sciences 4, 151-156.

Lande R. and Thompson R. (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. Genetics 124, 743-756.

Lander E.S. and Botstein D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121, 185-199.

Le Roy P. and Elsen J.M. (1991). Détection et localisation de gènes à effet quantitatif. Détection de gènes majeurs chez les animaux : apport des marqueurs. Séminaire du département de Génétique et d'Amélioration des Plantes. Méribel, INRA, 79-84.

MacLennan D.H. and Phillips S.M. (1992). Malignant hyperthermia. Science 256, 189-793.

Mangin B., Goffinet B. and Elsen J.M. (1993). Testing in normal mixture models with some information on the parameters. Biometrical Journal.

Pons O. (1993). Homogeneity test in distribution mixtures and application to major genes detection. Statistics.

Quinn B.G., McLachlan G.J. and Hjort N.L. (1987). A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. Journal of the Royal Statistical Society Series B 49, 311-314.

Rodolphe F. and Lefort M. (1992). Molecular marqueurs and genetic determinism of quantitative characters: use of a model in which the effects are linked to the markers. Technical Report, Laboratoire de Biometrie, Inra, Jouy-en-Josas, France.

Titterington D.M., Smith A.F.M. and Makov U.E. (1985). Statistical analysis of finite mixture distribution. John Wiley and Sons. New-York.