2
Statistical Theory in QTL Mapping

Benjamin Yakir, Anne Pisanté, and Ariel Darvasi

1. INTRODUCTION

Variability may be introduced in an observed phenotype by a range of elements. Inherited genetic factors, as well as environmental and behavioral conditions, may affect the phenotype. The blend of all these interactions gives rise to the unique being every living creature is. Experimental genetics has traditionally been, and still is, a very powerful tool for dissecting the genetic factors out of the blend that results in the observed phenotype complexity.

Unlike human genetics, a major advantage of experimental genetics is the ability to control the genetic background through inbred strain crosses, whereas nongenetic factors are kept relatively constant under controlled laboratory conditions. In reality, the ideal experiment is almost never feasible, and uncontrolled sources of variation and complex interactions may still obscure the underlying genetic effect.

Even under the best conditions, mapping quantitative trait loci (QTL) is a demanding endeavor. Any given QTL or genomic polymorphism contributes only a limited fraction of the phenotypical variation. This complex inheritance may involve partial penetrance, heterogeneity, the joint action of several genes, environmental effects, and more. The genetic dissection of complex traits is unavoidably based on a statistical approach. The aim in this chapter is to describe our view of the fundamental principles on which the statistical approach is based.

In a nutshell, the theory we will discuss involves the attempt to detect and locate weak signals in a noisy environment. This calls for the usage of large samples. Thereby, our probabilistical framework involves the distribution of statistics computed from large samples in the context of what is known as *local alternatives*. In statistical language, this theory is called *large sample theory*. Modern technology puts at our disposal the ability to genotype these

From: *Computational Genetics and Genomics*
Edited by: G. Peltz © Humana Press Inc., Totowa, NJ

samples over a practically unlimited collection of molecular genetic markers. The statistical investigation should make full use of this data. Stochastical processes are more appropriate as a model in this context than the separate investigation of individual markers. The statistical tools that were developed in the context of stochastical processes, in particular scanning statistics, are applicable also in the context of QTL mapping.

One should realize that QTL mapping is a multistage process that proceeds through several steps. The first step typically involves detection of chromosomes, or very large segments of chromosomes, which are likely to contain a QTL. In the next steps an attempt is made to narrow down the region containing the QTL. Finally, after a reduction to a small enough chromosomal segment, the gene associated with the variability in the investigated phenotype may be cloned, and its specific alleles may be identified. Several factors determine which tool is most appropriate at which stage. By the word tool we mainly mean here the selection of the cross and/or genetic resource. Phenotyping and genotyping methods may also be included in this context. A major factor, which determines to a large extent the advantages and disadvantages of a given tool for a given stage of the process, is the expected number of recombination events. This factor is directly determined by the breeding protocol. Statistically, recombination is reflected in the correlation among markers that reside on the same chromosome and between the markers and the QTL. In principle, an increase in the number of recombination events reduces the correlation. Reduction in the correlation is usually a blessing in the stages of fine-mapping but an obstacle in the first stage of detection. Another important factor is the strength of the statistic signal. The strength is usually summarized in the form of a noncentrality parameter; it is affected both by biological mechanisms by which the genetic variability is reflected as a phenotypic variability and by the breeding protocol. Other factors to be considered include, of course, the availability of the different resources and their respective costs.

Many attributes make the mouse an ideal mammalian model organism, especially for genetic investigations for which a wealth of resources have been established over the years. The relatively short generation time of the mouse, their easy breeding, and well-documented biological properties have led to the development of well-characterized, genetically designed specific strains *(13,14)*. These privileged circumstances have been exploited in both gene-to-phenotype and phenotype-to-gene studies. With the advent of molecular genetics, the use of DNA polymorphisms *(11)* has allowed for a refined identification of interstrains genetic divergence *(4)*. Correlation of the human and mouse genetic maps *(6,15)*, finally, makes the genetic analyses carried

out in mice applicable to human diseases by means of comparative mapping *(15)*. Mouse inbred strains are invaluable models for many complex diseases (for review *see* ref. *12*). The use of more specialized genetic resources, such as congenics, chromosome substitution strains, recombinant inbred (RI), along with various statistical packages *(12)* has already led to a primary dissection of a few complex, multigenic traits. A detailed description of the mouse strains and their use in genetics can be found in Lee Silver (*see* ref. *17* and http://www.princeton.edu/lsilver/book/MGcontents.html).

In the present chapter we examine the statistical aspects of QTL mapping, with special emphasis on the relevant parameters, their impact on the genetic design to be chosen, and reciprocally, their adjustment under the various genetic models.

2. DESIGN OF GENETIC EXPERIMENTS IN MICE

A genetic mapping program in mice is typically initiated by the selection of two pure inbred strains that exhibit a substantial difference in terms of the observed phenotype. An inbred strain lacks genetic variation. All mice within the strain carry two identical copies of each autosome and are thus genetically identical for all practical purposes. Conversely, genetic variation is present between strains, leading presumably to the between-strains average phenotypical difference. Crossing the two strains gives rise to offspring that are a genetic combination of the two parental strains. The process of recombination then blends the genomes further in subsequent crosses, generating mice with chromosomes that are a mosaic of segments from the two parental genomes. Correlating the parental origin of the genetic material at various loci with the measured level of the trait is the major statistical tool for identifying the genetic factors associated with the phenotype.

Several experimental designs have been developed in the context of QTL analysis. The most widely used designs are the backcross (BC), the intercross (F2) designs, and to some extent RI strains (*see* Fig. 1). The statistical theory we present here is given primarily in the context of those three designs.

3. THE STATISTICAL MODEL

Denote by Y the phenotype measurement for a random mouse. This measurement may vary both within and between lines of pure inbred mice. Some of this variability may be attributed to genetic and some to nongenetic factors. For a given locus showing polymorphism between two given lines of inbred strains, denote by A_1 the allele originating from one strain and by A_2 the allele originating from the other. Intercrossing the two inbred strains may give rise

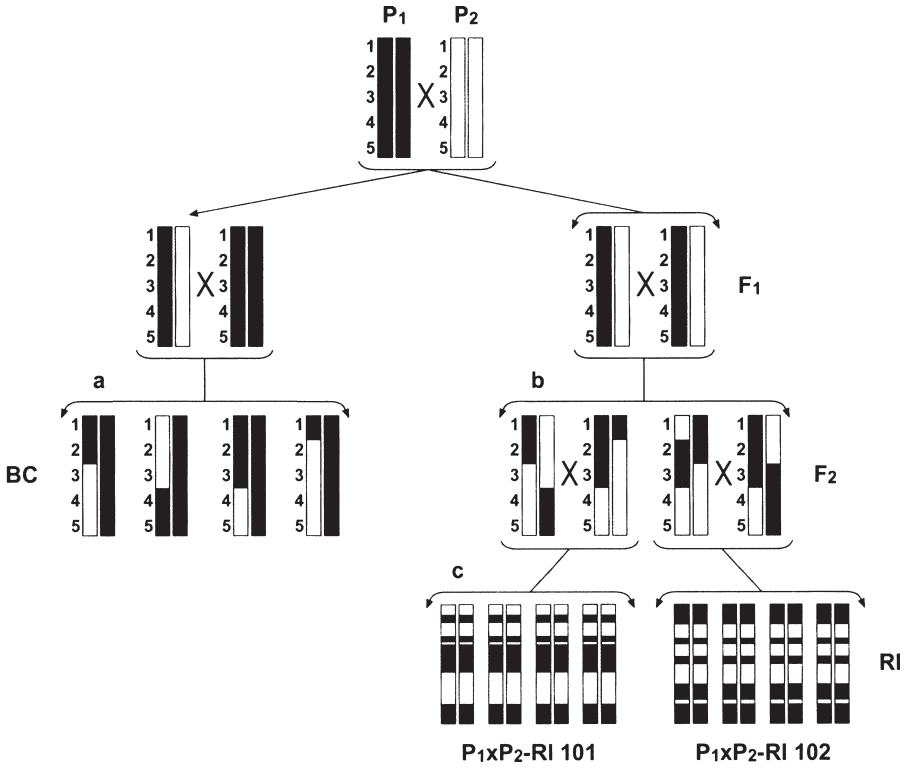

Fig. 1. The three cross designs: (a) backcross (BC), (b) intercross (F_2), and (c) recombinant inbred (RI). An outcross between two inbred lines (P_1 and P_2) produces the F_1 generation, with all the mice heterozygous over the whole genome; (a) the F_1 can be crossed back to one of the parental strain (P_1) to produce the BC; (b) an intercross within the F_1 will result in F_2 offspring; (c) strict inbreeding between F_2 pairs, for many generations, and following a single pair of chromosomes, will generate RI strains. All the individuals within a given RI strain carry the same homozygous, recombinant genotype.

to three genotypes. If X represents the copy number of allele A_2 in a genotype, then the variable X can take the values 0, 1, or 2. The model potentially assigns a different average level of Y for each genotype. At the same time, the variance of Y, or other characteristics of its distribution, is assumed to be independent of the genotype. The relation between the genotype and the phenotype is given in the regression formula:

$$Y = \mu + \alpha \cdot X + \delta \cdot I_{\{X=1\}} + e, \tag{1}$$

where e is a zero mean random deviate and I_A is the indicator function of an event A. (Specifically, $I_{\{X=1\}}$ is equal to 1 if the mouse is heterozygous and 0 if it is homozygous.) The coefficient α represents the *additive effect*, the coefficient δ represents the *dominant effect*, and the term μ is the intercept. μ is the expected level of Y for an (A_1A_1) homozygote. The expected level for an (A_2A_2) homozygote is 2α, and the expected level for a heterozygote is $\alpha + \delta$.

The deviate e incorporates all remaining factors that contribute to the variability. Such factors can include the genetic contribution from loci other than the one investigated, as well as environmental factors. We assume that this deviate is normally distributed and is uncorrelated with the genotype variable X.

The chosen cross design between the two inbred strains (BC, F2, RI, etc.) affects the distribution of the random variable X, as well as the distribution of the deviate e. For example, if the BC is formed by crossing back the F1 mice to the (A_1A_1) inbred strain, then X may take either the value 1 or the value 0, both with 0.5 probability. On the other hand, if the F1 mice are crossed one with another in order to form the intercross (F2), then X may take the values 0 or 2 with 0.25 probability and the value 1 with 0.5 probability. Finally, the RI mice are inbred strains, hence homozygous; X may take the value 0 or 2 with 0.5 probability each.

The phenotypical variance, the variance of Y, is a combination of variances that arises from several sources. The genetic variance is the part of the variance that is associated with the specific QTL. The source of the genetic variance is the variability of X. For the BC design, X may take only two values, the contribution of X to equation (1) simplifies to $(\alpha + \delta) \cdot X$. The variance of this component is $(\alpha + \delta)^2/4$, because the variance of X, a binomial (1, 1/2) random variable, is 1/4. For the F2 design, the genetic variance is the variance of $\alpha \cdot X + \delta \cdot I_{\{X=1\}}$. Because X and $I_{\{X=1\}}$ are statistically uncorrelated, this variance is the sum of the variances of its components. The variance of $\alpha \cdot X$ is $\alpha^2/2$, because X has a binomial (2, 1/2) distribution for the F2. The variance of $\delta \cdot I_{\{X=1\}}$ is $\delta^2/4$, because the variance of the indicator is 1/4. Overall, the genetic variance in the F2 is $\alpha^2/2 + \delta^2/4$. For the RI design, the relevant term is $\alpha \cdot X$ because RI lines are by definition homozygous. The variance here is α^2, because X equals 0 or 2 with 0.5 probability each.

Which of these three terms for the genetic variances is larger depends on the genetic model of the trait. For an additive model ($\delta = 0$) the genetic variance of an RI is twice as large as the genetic variance of the F2 and four times larger than that of the BC. However, for a dominant model ($\alpha = \delta$) the

genetic variance of the RI is equal to the genetic variance of the BC. The genetic variance of the intercross is 25% smaller.

The heritability coefficient (H^2) is a preliminary approach for assessing the efficiency of a design. This coefficient is the ratio between the variance of the genetic term—the term involving X—and the overall variability of Y. It may take values between 0 and 1. The closer the coefficient is to 1, the more informative the design is. In the opposite case, values of H^2 closer to 0 make the statistical inference more difficult.

The H^2 may give a rough idea regarding the statistical merits of a design. A much better insight is provided when considering two additional parameters—the parameter of noncentrality at the QTL and the between-markers correlation coefficient. We devote the rest of this section to the definition of these quantities and the computation of their values for the three designs. In the subsequent sections we illustrate the role of these terms in the assessment of the properties of various statistical inferential tools.

3.1. The Parameter of Noncentrality

Statistical inference is based on correlating the genetic information at given polymorphical loci (genetic markers) with the phenotypical expression Y. A given collection of mice can be subdivided according to their genotype at a given locus (the levels of the variable X). This leads to the formation of up to three subclasses. The statistical analysis proceeds by comparing the differences in the levels of phenotypical expression (Y) between the subclasses.

Consider the BC design, and let us assume initially that we know the genetic configuration at the functional polymorphism that affects the quantitative trait. The variable Xmay have here only two values. A natural summary statistic (Z) computes the difference between the average expression levels in each group.* For convenience, this statistic is standardized to have a standard deviation of 1. We define the parameter of noncentrality to be the expected value of this Z:

$$\mathbb{E}(Z) \approx \frac{\sqrt{n}(\alpha+\delta)}{2\sigma}, \tag{2}$$

where n is the number of BC mice that were genotyped and σ is the standard deviation of the deviate e from the regression model (1). The $(\alpha + \delta)$ arises as the expectation of the difference between the average phenotypes of the heterozygote and of the homozygote. The expectation in (2) is obtained by

*A slightly better statistic is the standardized sample regression coefficient. However, the statistic based on the difference of the averages is asymptotically equivalent and, in our view, is easier to interpret. Consequently, we will analyze this statistic.

dividing the expectation of the difference by its standard deviation, namely $\left(2\sigma/\sqrt{n}\right)$.* (As a matter of fact, one can justify using the statistic Z as an approximate score statistic for testing the null hypothesis that the locus is not associated with the trait, i.e., $\alpha + \delta = 0$. We will not follow this more formal route. The interested reader is referred to ref. *5*).

For the RI design also, X can take only two values. The standardized difference between the two types of homozygotes, which we call again Z, has the noncentrality parameter**:

$$\mathbb{E}(Z) \approx \frac{\sqrt{n}\alpha}{\sigma}. \tag{3}$$

In the F2 design, all three subBclasses can be realized. This leaves more flexibility in constructing statistics. For example, we shall use the statistic Z_α in order to make inference on the additive effect. Z_α is based on estimating the slope α in model (1). It essentially reflects the differences in phenotype average levels between the two homozygote types (standardized to have a standard deviation of 1). Likewise, we shall use the statistic $Z\delta$ (the difference in phenotypical average levels between the heterozygotes and homozygotes) in order to investigate deviations from the additive model. The expectations of these two types of statistics are***:

$$\mathbb{E}(Z_\alpha) \approx \frac{\sqrt{n}\alpha}{\sqrt{2}\sigma}, \quad \mathbb{E}(Z_\delta) \approx \frac{\sqrt{n}\delta}{2\sigma} \tag{4}$$

One can use a 2° of freedom χ^2 statistic in order to simultaneously test both the additive and dominance effects *(5)*. This statistic has the form: $Z_\alpha^2 + Z_\delta^2$.

Although this is not reflected in the above formulae, the σ value depends on the adopted breeding protocol and may vary between the BC, F2, and RI designs.

---

*About half the mice are heterozygotes. The variance of the phenotype across heterozygotes, as well as across homozygotes, is σ^2 (because X is given in each case). Therefore, the variance of the averages is $\sigma^2/(n/2)$. The variance of the difference is the sum of the variances, which leads to the expression of the standard deviation.

**About half the mice are homozygous for the alternative alleles. The expected difference between the two types of homozygotes is 2α. The variance of the averages for each homozygote type is $\sigma^2/(n/2)$. Hence, the expectation of Z is $(2\alpha)/\sqrt{2\sigma^2/(n/2)} = \sqrt{n}\alpha/\sigma$.

*** The expectation of the difference between the two homozygote types is 2α. The frequency of each homozygote type is about $n/4$. This leads to a variance of the average difference of $8\sigma^2/n$, which gives the expression of the expectation of Z is (2_α).

The contribution of an (A_1A_1) homozygote to the expectation is μ. The contribution of an (A_2A_2) homozygote to the expectation is $\mu + 2\alpha$. The relative frequency of (A_1A_1) among homozygotes is about 1/2. Consequently, the expectation of the average of the homozygotes is $\mu + \alpha$. The expectation among heterozygotes is $\mu + \alpha + \delta$. Therefore, the expectation of the difference is δ.

3.2. The Correlation Coefficient

The second parameter of interest is the intermarkers correlation coefficient. Given a pair of markers, consider the pair of computed Z statistics, one for each marker. The intermarkers correlation coefficient is the statistical correlation between these two statistics. This correlation is computed under the null assumption of both markers not being linked to a QTL ($\alpha = \delta = 0$). The value of this parameter is determined only by the recombination fraction between the two loci. It is independent of the additive and dominant coefficients of the trait (the parameters that determine the noncentrality parameter).

Consider a pair of markers on a random BC mouse, located at locus s and locus t on the same autosome denoted by $X(s)$ and by $X(t)$, the genotypes at both loci, respectively. Each may take either the value 0 or the value 1. Let θ be the recombination fraction between these two loci. The probability of the event $\{X(t) = 1\}$, given $\{X(s) = 1\}$, is $1–\theta$, because this event occurs if, and only if, the gamete inherited from the F1 parent is not recombinant. One can use the above conditional probability in order to show that the correlation between $X(t)$ and $X(s)$ is equal to $1–2\theta$. The associated test statistics $Z(t)$ and $Z(s)$ are (approximately) linear combinations, over a sample of mice, of the $X(t)$ and $X(s)$ variables. Consequently, for a large sample size,

$$\mathrm{corr}(Z(t), Z(s)) \approx \mathrm{corr}(X(t), X(s)) = 1–2\theta. \tag{5}$$

Consider next a random F2 mouse. Now, $X(s)$ and $X(t)$ may take three values each and two types of statistics are computed: Z and Z_{δ}. The probability transition matrix of going from $X(s)$ to $X(t)$ is given by:

$$\mathbb{P}(X(t) = i \mid X(s) = 2) = [(1-\theta)^2, 2\theta(1-\theta), \theta^2)],$$
$$\mathbb{P}(X(t) = i \mid X(s) = 1) = [\theta(1-\theta), \theta^2 + (1-\theta)^2, \theta(1-\theta)],$$
$$\mathbb{P}(X(t) = i \mid X(s) = 0) = [\theta^2, 2\theta(1-\theta), (1-\theta)^2)].$$

(As an example, observe that because in this case both gametes from the F1 parents should not be recombinant. Similar considerations provide the other entries in the above matrix.) Direct calculations give that here again the correlation between $X(t)$ and $X(s)$ is equal to $1–2\theta$. Consequently, because $Z_{\alpha}$is approximately a linear combination of X:

$$\mathrm{corr}(Z\alpha(t), Z\alpha(s)) \approx 1–2\theta. \tag{6}$$

In a similar way, the correlation coefficient between $I_{\{X(t) = 1\}}$ and $I_{\{X(s) = 1\}}$ is equal to $(1–2\theta)^2$. Z_{δ} is approximately a linear combination of $I_{\{X = 1\}}$. Hence,

$$\mathrm{corr}(Z_{\delta}(t), Z_{\delta}(\mathrm{s})) \approx (1–2\theta)^2. \tag{7}$$

The Z_{α} variables are not correlated with any of the Z_{δ} variables.

An RI mouse has two identical copies of each autosome. The parental origin at two loci is the same (X(t) = X(s)) if that chromosome is not recombinant and vice versa. Denote by θ_{RI} the recombination fraction for a random gamete in the RI sample. The classical result of Haldane and Waddington (8) can be used in order to attain the approximation:

$$\theta_{RI} \approx 4\theta/(1 + 6\theta) \tag{8}$$

(*See* ref. *9* for a general derivation of this and other results by a presentation of the problem in terms of finite population dynamics.) Considerations similar to those used for the BC give the following for the RI result:

$$\text{corr}(Z(t), Z(s)) \approx 1–2\theta_{RI} \approx 1–8\theta/(1 + 6\theta). \tag{9}$$

This completes the computation of the intermarkers correlations for the three designs.

4. LARGE SAMPLE THEORY AND GAUSSIAN PROCESSES

The selection of inferential statistics should not be taken lightly. The choice may substantially affect the efficiency of the statistical analysis. This selection is typically guided by the prior assumption of the way genetic and nongenetic factors interact with the measured phenotype. This prior assumption is reflected in the statistical model that formulates the interaction. The model presented in equation (1) is an example of such a statistical model. This model is consistent with a prior assumption that a single major locus is responsible for a substantial part of the phenotypical variation, with other genetic factors, if any, adding only a small contribution each to that variation. Moreover, this model disables some forms of nonadditive epistasis and some forms of gene/environmental interaction.

In the sequel we will consider separately the statistical properties of inferential statistics for each of the three experimental designs. These statistics are computed based on the phenotypical and genotypical data collected over a collection of markers. Model (1) of a single major gene and dense genotyping is consistent with the approach of computing an inferential statistic for each marker at a time. (If the markers are not so densely spaced, interval mapping may be preferred *(11,5)*.) The inferential statistic will be of the form Z in the BC and RI designs and of the form in the F2 design. Other models may propose the use of other types of test statistics. (i.e., in order to detect interacting genes, one may consider inferential statistics, computed from the phenotypical and genetic data for a pair of markers, for all such possible pairs; *see* ref. *10*.)

The first step in the investigation of the properties of statistical procedures involves the determination of the distribution of the inferential statistics.

Let us focus on a given autosome. Markers are genotyped at loci $t_1, t_2, \ldots, t_m$ (a total of m markers). For the BC and RI we denote the summary statistics by $Z(t_1), Z(t_2), \ldots, Z(t_m)$. For the F2 we denote them by $Z_\alpha(t_1), Z_\alpha(t_2), \ldots, Z_\alpha(t_m)$ and $Z_\alpha(t_1), Z_\delta(t_2), \ldots, Z_\delta(t_m)$. According to the large sample theory, the joint distribution of these statistics is approximately multinormal. Multinormal distributions are fully determined by the means and variances of the components and by the correlations between components. The components were standardized to have a variance of 1. The correlations were computed in (5) for the BC, in (6) and (7) for the F2, and in (9) for the RI. Thus, we are left only with the task of determining the means of the components.

The key concept in the determination of the means of the components (and a useful concept in general in statistics) is the concept of *sufficiency*. A statistic is called sufficient if it contains all the relevant information for making statistical inference (i.e., more formally, if conditioning on the statistic eliminates the dependency between parameters of the model and the distribution of the data). Assume a QTL is present at locus s. Let us figure that this locus is also genotyped and that an appropriate test statistic is computed (where the test statistic is $Z(s)$ in the BC and RI cases or $(Z_\alpha(s), Z_\delta(s))$ in the F2 case). Because, by assumption of model (1), the given QTL is the only genetic factor on the chromosome that contributes to the phenotypic variability, the particulars of these imaginary statistics, had we had them, would have been sufficient for determining the association with the trait. Therefore, the information at the other loci is no longer relevant, whether it is available or not.

This sufficiency assumption forces a given relation between the mean of a statistic computed at a marker ($Z(t)$) and the mean of the statistic ($Z(s)$) at the QTL, namely:

$$\mathbb{E}[Z(t)] = \mathrm{corr}(Z(t), Z(s)) \times \mathbb{E}[Z(s)]. \tag{10}$$

The right-hand side of (10) is the outcome of the noncentrality parameter, given in (2) for the BC, in (4) for the F2, and in (3) for the RI. The correlation coefficient between loci is computed in (5), (6), (7), and (9) for the three designs. In summary, the means of the components can be determined by identifying the parameter of noncentrality at the QTL and the correlation between the QTL and the various markers.

The Haldane model of crossovers is a popular model that leads to a simple relation between the genetic distance and the recombination fraction. Applying this function yields a correlation coefficient of the general form: $\exp\{-\beta|t - s|\}$ with β varying between the BC and the F2 designs (the correlation coefficient for the RI design does not have this form). Multinormal vectors with such correlation structure are denoted Orenstein–Uhlenbeck

processes. Yet, as we shall see in the following sections, the statistical properties of the inferential procedures based on the multinormal process do not depend on the exact form of the correlation function but on a rather weaker property.

5. DETECTING A QTL

Mapping a QTL is a multistage process. The first step, following the phenotypical and genotypical data collection, is the determination of the reflection in the collected data of the presence of a QTL. It should be noted that even when a genetic influence on the trait is undisputable, its effect may be too weak, and our data may not be sufficient, in order to distinguish it from random fluctuations. Therefore, the first question to be addressed is: Can we detect a strong enough signal for the presence of a QTL? If the answer for this question is affirmative then we can proceed in the process of mapping the QTL. If, however, the answer is negative, then we ought to revise our strategy. Such a revision may include an increase in the sample size within the framework of the current design, using a different cross design, and so on.

The field in statistics theory that deals with the issue of determining the presence of a signal in a noisy environment is called *hypothesis testing*. According to this theory, one should select a test statistic with a distribution that best reflects the presence of a signal, and base the conclusion on the computed value of that statistic. In the case of QTL mapping we identified such statistics—the statistics of the form Z in the BC and RI designs and the statistics $Z_\alpha^2 + Z_\delta^2$. in the F2 design. Large values of the statistics in the latter case or large absolute values of the statistics in the former cases are an indication of the presence of a QTL in the vicinity of the marker: a strong effect of the QTL will be reflected by a nonzero noncentrality parameter of the statistics, which will tend to deviate its value away from 0.

The simple theory of hypothesis testing, which is based on normal distribution, would have been applicable had we looked at a single marker, and a single marker only. However, in our case we examine a sequence of test statistics, one for each marker. An extreme value in *any* of the test statistics is an indication of the presence of a QTL. Thus, in reality, our test statistic is $\max_{ti} |Z_\alpha^2(t_i)|$ in the case of the BC and RI designs and $\max_{ti} [Z_\alpha^2(t_i) + Z_\delta^2.(t_i)]$ in the case of the F2 design, when the maximization is taken across all markers. It turns out that the distribution of these statistics is no longer normal, even though each component has a normal distribution. The determination of the threshold, which will assure a given significance level for the experiment,

is based on the distribution of the maximal test statistic in the absence of a QTL. This distribution, as we shall see in equations (11) and (12), depends on the form of the test statistics, the number of markers used for scanning, and on the correlation between the inferential statistics. This correlation is a function of the distance between markers and the design of the cross.

The probability of reaching the threshold is less than the product of the number of markers examined by the probability of reaching the threshold with a single marker. This last probability is easily computed using the normal distribution in the case of the BC and RI, or the χ^2 distribution on 2° of freedom in the case of the F2. This upper bound, also known as the *Bonferroni upper bound*, is actually a reasonable approximation of the true probability when the correlation between markers is not too high. However, when the correlation between markers is high, a better approximation of the probability takes the form:

$$\mathbb{P}(\max_{t_i} |Z(t_i)| \geq z) \approx [C + (\beta L z^2) \cdot \nu(2z^2 \beta \Delta)] \cdot \mathbb{P}(|Z| \geq z), \tag{11}$$

when the basic test statistic is a single normal variable Z, and the form:

$$\mathbb{P}(\max_{t_i} [Z_\alpha^2(t_i) + Z_\delta^2(t_i)] \geq u) \approx [C + (\beta L u) \cdot \nu(2u\beta\Delta)] \cdot \mathbb{P} \quad (Z_\alpha^2 + Z_\delta^2 \geq u), \tag{12}$$

when the basic test statistic is a χ^2 statistic. Here, the number of markers, used for the Bonferroni upper bound, is replaced by the term in the square brackets. The components that determine the value of these approximations are the number of chromosomes scanned (C); the sum of lengths between the first and the last marker in each chromosome, across all those chromosomes (L, measured in cM); the threshold (z in the first formula and u in the second); the probability of reaching the threshold with a single marker ($\mathbb{P}(|Z| \geq z)$ in the first formula and $\mathbb{P}(Z_\alpha^2 + Z_\delta^2 \geq u)$ in the second); and the components that reflect the correlation between markers. These components are the average distance between consecutive markers (Δ, measured in cM) and the rate with which the correlation between two markers approaches 1 as the markers get closer to each other (β). This last term is equal to 0.02 in the BC design, 0.08 in the RI design, and it turns out to be (0.02 + 0.04)/2 = 0.03 in the F2 design.

The function $\upsilon(\cdot)$ appearing in these formulae was originally developed in the context of random walks and renewal theory. It appears in other fields of statistics as well, including change-point detection and scanning statistics. It takes the form:

$$\nu(y) = \frac{2}{y} \exp\left\{-2\sum_{n=1}^{\infty} \Phi(-\sqrt{ny}/2)\right\}, \tag{13}$$

where $\Phi(\cdot)$ is the cumulative probability function of the normal distribution. It turns out that the function $\upsilon(\cdot)$ approaches the value of 1 as y approaches 0. The function can be approximated by $\exp\left\{-0.583\sqrt{y}\right\}$ for small values of y *(16)*. When markers become denser and denser, the distance between them, Δ, becomes smaller. This makes the arguments of the function $\upsilon(\cdot)$ in (11) and (12) approach 0. In the asymptotic case, the formula represents the probability of false detection with a continuum of markers. This formula is obtained by removing the function $\upsilon(\cdot)$ from the expressions in (11) and (12). At the other extreme, the function can be approximated by $2/y$ for large values of y. Substituting the function with this approximation reproduces the Bonferroni upper bound, because $\Delta = L / (m - C)$. Therefore, one can view the function $\upsilon(\cdot)$ as a correction term, which takes into account both the discreteness of the markers and the correlation between them.

Computing the power is an essential requirement for designing the experiment. The power is the probability of detecting the QTL, i.e., the probability of reaching the given threshold when a QTL is present. This probability depends on the expectations of the statistics computed at the markers. These expectations are tilted to have a nonzero value on the chromosome carrying the QTL. As we saw in (10), the expectations depend on the noncentrality parameter and on the correlations between the markers and the QTL. A simple lower bound for the power can be obtained by considering the probability of reaching the threshold in either of the two markers flanking the QTL. A refined approximation will take into account the possibility of reaching the threshold for markers that are further away from the QTL. We will not present these approximations here. The interested reader is referred to ref. *5*.

6. ESTIMATING MAP LOCATION

In the first stage of mapping a QTL the issue is to evaluate the reflection in the data of the presence of a QTL. If the answer to this evaluation is affirmative, then the continuation of the process of mapping involves narrowing down the candidate region likely to contain a QTL as much as possible. In its initial stage, this process involves the construction of a confidence interval (CI) for the QTL based on the data used for detection.

One procedure for constructing confidence intervals is by examining tests for the presence of a QTL at various loci. A QTL is assumed to exist somewhere along the chromosome. However, its exact location is unknown. According to this procedure, a locus s is included in the confidence interval if the hypothesis that s is the exact location of a QTL is *not* rejected. It follows that if the significance level for that test is 10%, then the confidence

level of the resulting CI is 90%. Likewise, if the significance level of the tests is 5%, then the confidence level is 95%.

One approach for constructing such location tests makes the simplifying assumption that the QTL is completely linked to one of the markers, or in other words, the correlation coefficient between the QTL and one of the markers is 1. Yet, the marker that is completely linked to the QTL remains unknown. The problem of constructing a CI reduces, through this assumption, to the problem of testing each of the markers for being completely linked to the QTL. In the CI, all the markers that were not rejected by the test are included. Naturally, this approach may produce better results when markers are densely spaced, in which case the simplification made does not introduce much error. It may be less satisfactory when the number of markers is limited. In the latter case one may try other approaches of constructing confidence intervals. We will not refer to such approaches. The reader may find an evaluation of several of these approaches in ref. *(5)*.

The decision to exclude a marker s from the CI (reject the hypothesis that s is the QTL) may be based on the relation:

$$\max_t Z^2(t) - Z^2(s) > x, \tag{14}$$

when a single degree of freedom statistic is used (BC, RI) or on the relation:

$$\max_t [Z_\alpha^2(t) + Z_\delta^2(t)] - [Z_\alpha^2(s) + Z_\delta^2(s)] > x, \tag{15}$$

when a 2° of freedom statistic is used (F2).

The selection of x to assure the desired confidence level may depend, however, on the unknown parameter of noncentrality, because the distribution of the statistics in (14) and (15) depends on that parameter. Still, a remedy to this problem may be provided by the notion of sufficiency. As was claimed before, the statistic $Z(s)$ in case (14) and the statistic $(Z_\alpha(s), Z_\delta(s))$ in case (15) are sufficient statistics for the parameters of model (1), including the noncentrality parameter. Consequently, the conditional probability of the events (14) or (15), given the value of the sufficient statistic, is independent of that unknown parameter. The threshold x can be selected based on this conditioned computation. The result is a confidence interval with the prescribed confidence level, regardless of what the true value of the noncentrality parameter is.

It should be noted that technically the problem of constructing a confidence interval for the QTL location is not like the problem of constructing a confidence interval for the population expectation. In the latter case, one typically takes an interval of about two standard deviations in each direction of the sample average in order to get a CI with a confidence level of 95%. This construction relies on the fact that the distribution of the sample average is

normal. The length of this interval decreases at a rate that is proportional $1/\sqrt{n}$, where n is the sample size (because the variance of the sample average is equal to the variance of a single observation, divided by the sample size). In QTL mapping, on the other hand, the estimate of the location of the QTL does not have a normal distribution, even when the sample size is large. Therefore, taking two standard deviations about its value will not result in a proper CI.

The difference between the normal case and QTL mapping is reflected also in the expression for the expected length of the CI. An approximation for this length for the case of a one degree of freedom statistic Z is provided in *(5)*:

$$\frac{x}{\beta\mu^2}+\frac{x^2}{2\beta\mu^4}+\frac{2(1-v(2\mu^2\beta\Delta))^{1/2}}{\beta\mu^2}+\frac{v^2(2\mu^2\beta\Delta)}{2\beta}. \tag{16}$$

Again, Δ is the average distance between markers, β is the rate of convergence to 1 of the covariance between markers as the distance between decreases, and $v(y)$ is the function presented in (13). The term μ is the noncentrality parameter. x is the threshold for the test. This threshold is essentially independent of the sample size. The noncentrality parameter, on the other hand, increases at a rate proportional to $\sqrt{n}$. It turns out, since the approximation is roughly proportional to $1/\mu^2$, that the expected length of the confidence interval decreases at a rate proportional to $1/n$ (compared to the $1/\sqrt{n}$, in the normal case).

7. FINE-MAPPING STRATEGIES

After detecting a QTL, a confidence interval for its location is computed. This confidence interval tends to be quite wide, perhaps 20 or 30 cM wide. Such wide intervals most likely contain dozens of genes that are good candidates to be the QTL. However, direct techniques of cloning, which may be used in order to verify that a given polymorphic sequence is the QTL, are lengthy and expensive. Therefore, it is critical to narrow down the search region, to below 1 cM, before the more direct measures can be applied. The process of narrowing down the interval containing the QTL is often called *fine-mapping*.

There is a major difference between fine-mapping of a Mendelian trait and fine-mapping of a QTL. In the former case there is a 1:1 relation between the presence or absence of the trait and the genetic composition at the functional composition. Thereby, one can barricade the functional polymorphism precisely by the identification of recombinant chromosomes and relating them to the phenotypical expression. In the latter case, on the other hand, there is

no such 1:1 relation, only statistical correlations between the genetic composition and the phenotypic expression. Consequently, one must revert to statistical procedures in order to carry out the task. These statistical procedures may be based on hypothesis testing, parallel in spirit to the task of QTL detection, or on the construction of a CI, similar to problem of estimating map location. The main concern in fine-mapping, however, is that the resulting region will be narrow enough.

Examining (16) we see that the two main parameters that determine its width are the parameter of noncentrality (μ) and the parameter that captures the rate of recombination in a close proximity to the QTL (β). The larger these parameters are, the shorter the confidence interval is expected to be. Fine-mapping is most efficiently conducted by selecting an experimental design that maximizes these parameters. An example of such design is to use an advanced intercross design, or F_i, as proposed in *(3)*. F_i stands for the *i*th generation of intercrossing. The rate of recombination (β) increases approximately linearly in *i*. This leads to a reduction in the width of the CI.

An alternative experimental design is the recombinant inbred segregation test (RIST). According to this design, RI strains are selectively crossed with their parental lines in such a way that ensures recombination in the investigated region. The simple chromosomes identification, which is used in Mendelian traits, is replaced by a statistical test to determine on which side of the recombination point the QTL is located. Choosing the appropriate RIST design, either the RIST-BC or the RIST-F2 design, will maximize the noncentrality parameter and improve the performance of the procedure. For a comprehensive review on fine-mapping strategies *see* ref. *(2)*.

8. DISCUSSION

In this chapter we have presented the statistical framework for QTL analysis in its various stages. Because any QTL will usually explain only a small fraction of the phenotypical variation, large samples cannot be avoided. We have emphasized on the two parameters that have the largest effect on this theory. The first is the noncentrality parameter, which reflects the proportion of variance explained by the QTL being studied, and the second is the extent of correlation between the functional polymorphism and the genetic marker tested and between pairs of markers. Different designs can be implemented for QTL analysis, and in this chapter we have described how the relevant parameters affect the use of each of the main experimental designs, namely, F2, BC, and RI strains.

QTL analysis consists of a number of steps as described throughout the chapter. The general theory presented here can serve as a basis for the analysis of any such stages. For example, although both the first and the second stages, QTL detection and map location, are affected by the same two parameters, their effect might be of opposite direction: the detection stage requires as little recombination as possible, whereas extensive recombination is preferred for localization.

The difficulty in QTL analysis lies in the large samples required for detection and the limited breakdown of the correlation in adjacent chromosomal regions. The sample sizes may reach unattainable numbers if the genetic architecture of the trait consists of many genes with small effect each. This has caused very few success stories in QTL analysis. Nevertheless, some have indeed succeeded in taking a QTL project all the way to the identification of the relevant genes. One such example is the *Mom1* gene affecting multiplicity and size of tumor induced by the Apc^{Min} mutation in mice *(1)*. In tomato, the *ORFX* gene was found to have an effect on fruit weight *(7)*. More recently, a complex genetic architecture influencing high-temperature growth could be resolved in yeast, using an elegant genetic approach *(18)*. With the advances of the postgenomic era other examples will undoubtedly follow. Multidisciplinary approaches, including comparative genetics, expression analysis, bioinformatics, proteomics, and so on, will undoubtedly help in this difficult endeavor.

REFERENCES

1. Cormier RT, Hong KH, Halberg RB, et al. Secretory phospholipase Pla2g2a confers resistance to intestinal tumorigenesis. Nat Genet 1997;17:88–91.
2. Darvasi A. Experimental strategies for the genetic dissection of complex traits in animal models. Nat Genet 1998;18:19–23.
3. Darvasi A, Soller M. Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 1995;141:943–951.
4. Dietrich W, Katz H, Lincoln SE, et al. A genetic map of the mouse suitable for typing intraspecific crosses. Genetics 1992;131:423–447.
5. Dupuis J, Siegmund D. Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 1999;151:373–386.
6. Eppig JT, Nadeau JH. Comparative maps: the mammalian jigsaw puzzle. Curr Opin Genet Dev 1995;5:709–716.
7. Frary A, Nesbitt TC, Grandillo S, et al. fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. Science 2000;289:85–88.
8. Haldane JBS, Waddington CH. Inbreeding and linkage. Genetics 1931; 16:357–374.
9. Kimura M. A probability method for treating inbreeding systems, especially with linked genes. Biometrics 1963;19:1–17.

10. Korol AB, Ronin YI, Kirzhner VM. Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics 1995;140:1137–1147.
11. Lander ES, Botstein D. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps [published erratum appears in Genetics 1994;136:705]. Genetics 1989;121:185–199.
12. Manly KF, Olson JM. Overview of QTL mapping software and introduction to map manager QT. Mamm Genome 1999;10:327–334.
13. Moore KJ, Nagle DL. Complex trait analysis in the mouse: the strengths, the limitations and the promise yet to come. Annu Rev Genet 2000;34:653–686.
14. Morse HC, III. The laboratory mouse: a historical perspective. In: Foster HL, Small JD, Fox JG, eds. The mouse in biomedical research. vol. 1. History, genetics, and wildmice. New York: Academic, 1981.
15. O'Brien SJ, Menotti-Raymond M, Murphy WJ, et al. The promise of comparative genomics in mammals. Science 1999;286:458–462, 479–481.
16. Siegmund D. Sequential analysis: tests and confidence intervals. New York: Springer, 1985.
17. Silver LM. Mouse genetics: concepts and applications. New York, Oxford: Oxford University Press, 1995. http://www.princeton.edu/lsilver/book/MGcontents.html.
18. Steinmetz LM, Sinha H, Richards DR, et al. Dissecting the architecture of a quantitative trait locus in yeast. Nature 2002;416:326–330.

Springer

http://www.springer.com/978-1-58829-187-5

Computational Genetics and Genomics
Tools for Understanding Disease
Peltz, G. (Ed.)
2005, X, 308 p., Hardcover
ISBN: 978-1-58829-187-5
A product of Humana Press