A comparison of two tests for the detection of QTL

----An example concerning chicken body weight

Author: **Wenjuan Hu**

Liwen Liang

Supervisor: Martin Sköld

Lars Rönnegård

D-level thesis in Statistics

Department of Economics and Society
Högskolan Dalarna, Sweden

Spring 2007

ABSTRACT

The problem of detection of QTL received much more attention in recent years. The aim of this paper is to compare two approaches of significant test for the detection of quantitative trait loci (QTL) affecting chicken body weight. Data used in the paper is from a practical intercross experiment. We employ regression mapping to calculate LOD(s) score which show the most likely QTL. Monte-Carlo permutation test is employed to measure the significance of the QTL effect as well as the simple likelihood ratio test. The comparison is given out at the end.

KEY WORDS: QTL, LOD score, Regression mapping, Permutation test, χ^2 Test

Contents

Chapter 1

Introduction

A Quantitative Trait Loci (QTL) is the genetic location[1] affecting the quantitative variation of biological trait in nature. Sometimes the trait variation is decided by more than one gene, called multiple loci (QTLs)[2]. The problem of identifying the QLT or QLTs is very important in biological research but also difficult. The variations in quantitative traits may come mainly from the variation of environment and get only quite small part of effect from the gene(s), in this case the genetic composition can not be inferred directly from the trait value and statistical methods are needed.

In this paper we consider the problem of detecting the single QTL that affects chicken body weight. We identify the most likely QTL along a chicken chromosome and compare two statistical tests for the detection of QTL. The data used in the paper is from an intercross experiment, which is developed and maintained at the Virginia Polytechnic Institute and State University in Blacksburg, Virginia, USA.

In this chapter, we describe the biological background of the data and the basic hypotheses. In Chapter 2, we give an explanation on different methods used to identify the QTL. In Chapter 3, we present our results and compare the two statistical tests. In Chapter 4, we get a conclusion of our analysis and take some further discussions.

1.1 Background

Along with the development of biochemical knowledge, the complete genetic code of several different species has recently been revealed. Increasing effort aimed at finding out how the differences between individuals in genetic code influence a certain quantitative trait have been made. This is in fact the problem of detection and

[1] Loci is some measure of location at which the gene has effect on the quantitative trait.
[2] Gregor Mendel (1866) and Nilsson-Ehle (1909) have proposed and demonstrated respectively that the quantitative variation of trait could be the result of the action of multiple genes.

characterization of the QTL(s).

Our study focus on comparing two statistical tests for the detection of QTL affecting chicken body weight, and the potential effects could be estimated by the data from a large experimental cross. Before the illustration of the practical data set, we first make a theoretical explanation of the experiment.

1.1.1 Experiment

The experiments for identifying quantitative trait loci (QTL) are usually divided into two approaches: backcross and intercross. A detailed account of representation of these two experiments has been taken in Karl William Broman's paper (1997).

Crossing animals from two breeds that differ genetically in the trait of interest can get two pure-breeding lines: the low (L) and the high (H) parental line. Each of the two lines is essentially homozygous at all locations. Because these lines are results of intensive inbreeding, so they received the same allele from both of their parents at each location. Then the first filial generation called F1 generation was given by crossing the two parental lines. These F1 individuals are heterozygous wherever the parental lines differ, since they receive a copy of each chromosome from each of the H and the L parental line. And they should be genetically identical (H*L). So the F1 generation has genotype of HL.

We can get the second filial generation by continuing crossing, and Broman (1997) used two figures to explain the two kinds of crossing experiment. Figure 1.1 shows the intercross experiment procedure clearly. In an intercross experiment, the F2 generation individuals are given by crossing the F1 individuals either itself or to each other, and each individual of F2 receives two chromosomes from F1 generation (HL*HL). So each of the F2 generation individuals will be a combination of the two parental HL chromosomes and the F2 will have genotypes LL, HL or HH at each location.

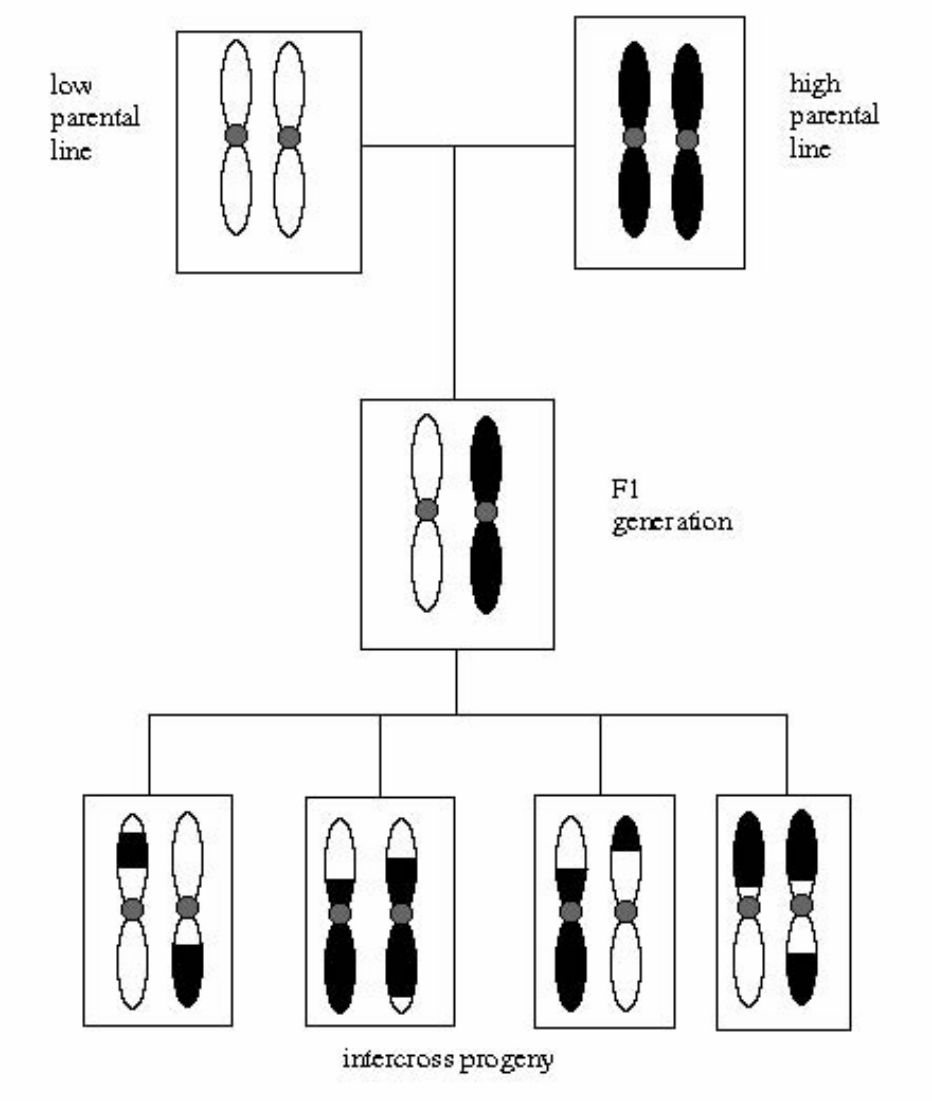

Figure 1.1: An intercross experiment with four progeny and only one pair of homologous chromosomes is shown.

Figure 1.2 shows the backcross experiment procedure. The F2 generation individuals are the offspring from mating of the F1 individuals (HL) and one of the parental lines (take the high line for example, H). Then the backcross individuals have the genotype either HL or HH since each of them receives one chromosome from the high parental line (H) and one from F1 generation (HL).

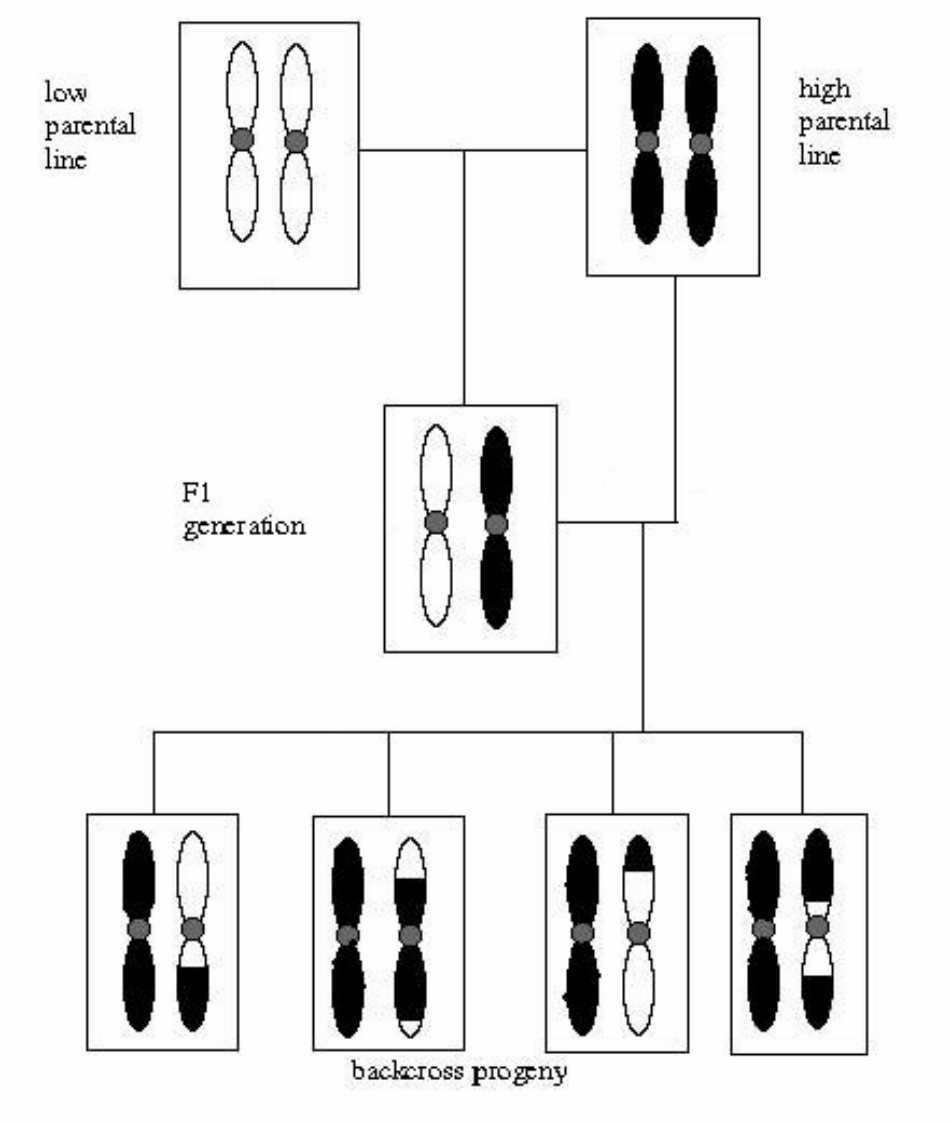

Figure 1.2: A backcross experiment with four progeny and only one pair of homologous chromosomes is shown.

The data set we used in our paper is generated from an intercross experiment, but we illuminate the theory of the QTL method by the genotypes of backcross experiment. Because the analytic substance of the approach is nearly the same for both experiments, and the backcross offspring have one of only two genotypes at each genetic location, so it is much simpler in calculation.

1.1.2 The practical intercross experiment

The data we used to present our analysis is from a real experiment. This experiment was initiated in 1957 by crossing seven partially inbred lines of White Plymouth Rock chickens at the Virginia Polytechnic Institute and State University in Blacksburg, Virginia, USA. The experiment researchers established two selection lines by divergent selection on a single trait of chicken body weight at 56 days of age. Then after more than 40 generations of selection in opposite directions, a high and a low weight line were obtained. These lines have a remarkable nine-fold difference in chicken body weight. Figure 1.3 shows the difference in each generation directly. The two lines were maintained as closed populations selected for either high or low body weight at 56 days of age. The high (H) and the low (L) chicken body weight lines are just the two pure-breeding lines resulted of intensive inbreeding that are needed for the cross experiment.

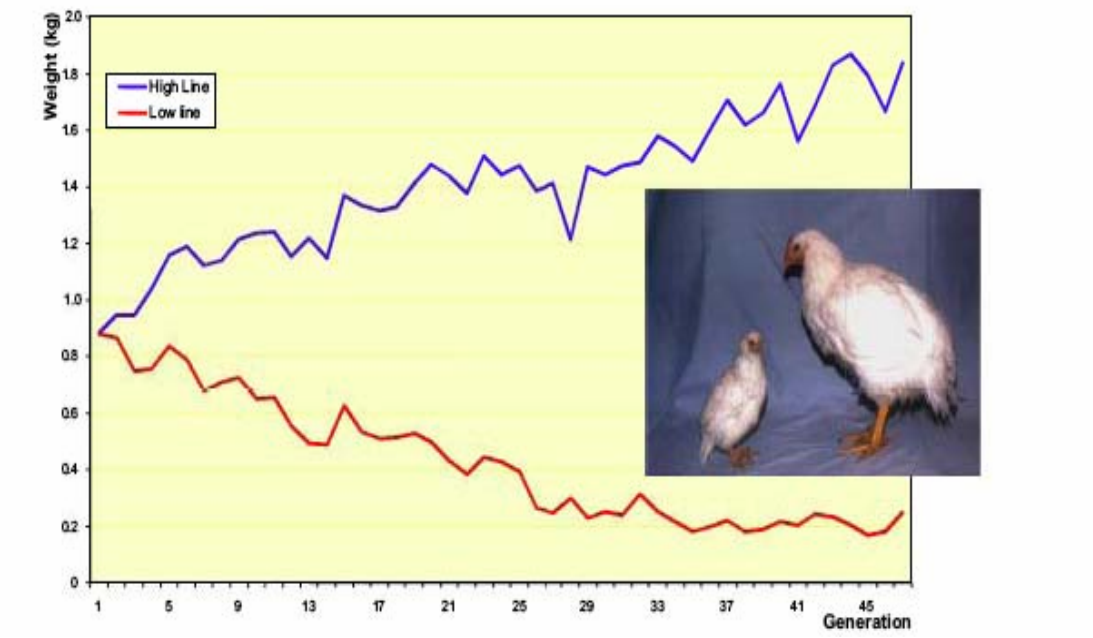

Figure 1.3: Chicken body weight of males at age of 56 days from generation 1 to 47 from the chicken lines selected for high and low weight. The chickens showed in the photo are from generation 37 which of 56 days old of high and low weight comparison.

An intercross was designed by the experiment researchers of the Institute and University using generation 41 of this long-term selection experiment. Then 10 high weight males (H) were mated to 22 low weight females (L) and 8 low weight males (L) were mated to 19 high weight females (H). Hence, there were 29 chickens from the high line (H) mated to 30 chickens from the low line (L), and 83 F1 generation offspring were born which contain 8 males and 75 females. Then from the intercross experiment, those 8 males and 75 females F1 generation offspring were intercrossed mated and 874 F2 generation chickens were born (Figure 1.4).

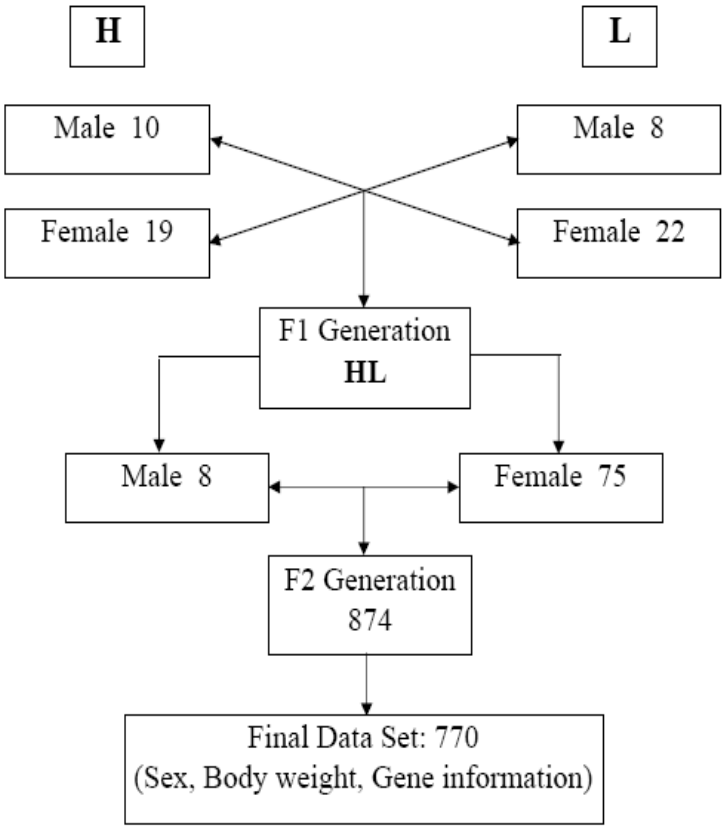

Figure 1.4: The intercross experimental procedure with the number and sex of the chickens of each generation and the final result.

770 chickens of the F2 generation were contained in the final data set since these individuals had recorded sex, body weight and known pedigree information. So the

final data information obtained from the intercross experiment and used in this paper are: chicken body weights (in grams) at 56 days of age for 770 individuals, sex (386 females which be noted as “0” and 384 males be noted as “1”), and the expected number of alleles (ranging from 0 to 2) inherited from the high line at 497 different locations along Chromosome 1[3].

1.1.3 Data explanation

In such an experiment, we assume that each of the offspring is scored for one trait and they can be typed at a number of genetic markers. In our example, the offspring are scored for body weight at 497 genetic locations.

We have 770 chickens with recorded weight and sex. Figure 1.5 shows the histogram of chicken body weight:

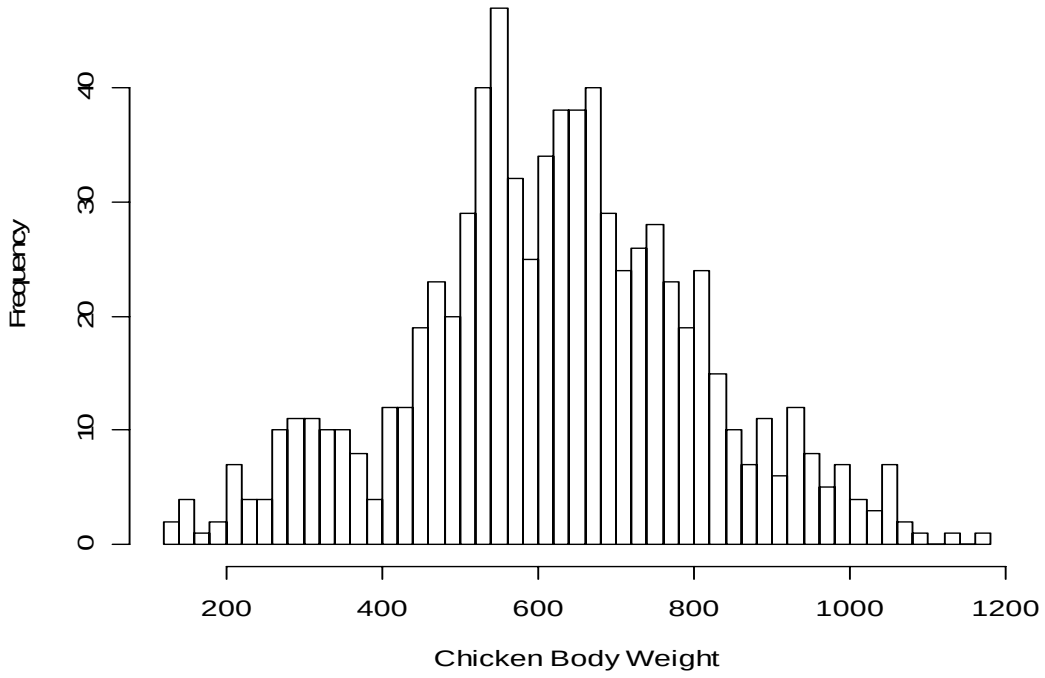

Figure 1.5: The histogram of chicken body weight

From the preceding introduction of the experiment, we know that for each

[3] Chickens have 31 chromosomes and Chromosome 1 is the longest of the 31ones.

offspring, there are three genotypes at each of these marker loci in an intercross experiment: LL, HL or HH. At each of 497 locations on chromosome 1, we have the estimated expected number of alleles (H). This is done by typing the chromosome at 74 so called marker locations and then interpolating linearly. The distance between these markers is given by the unit of centiMorgans (cM). The genotype “LL” is recorded of value 0 and “HH” of value 1, so the estimated expected number of alleles at each location is range from 0 to 2. Figure 1.6 below shows the estimated number of alleles inherited at each location on chromosome 1 of the third and the fourth chicken.

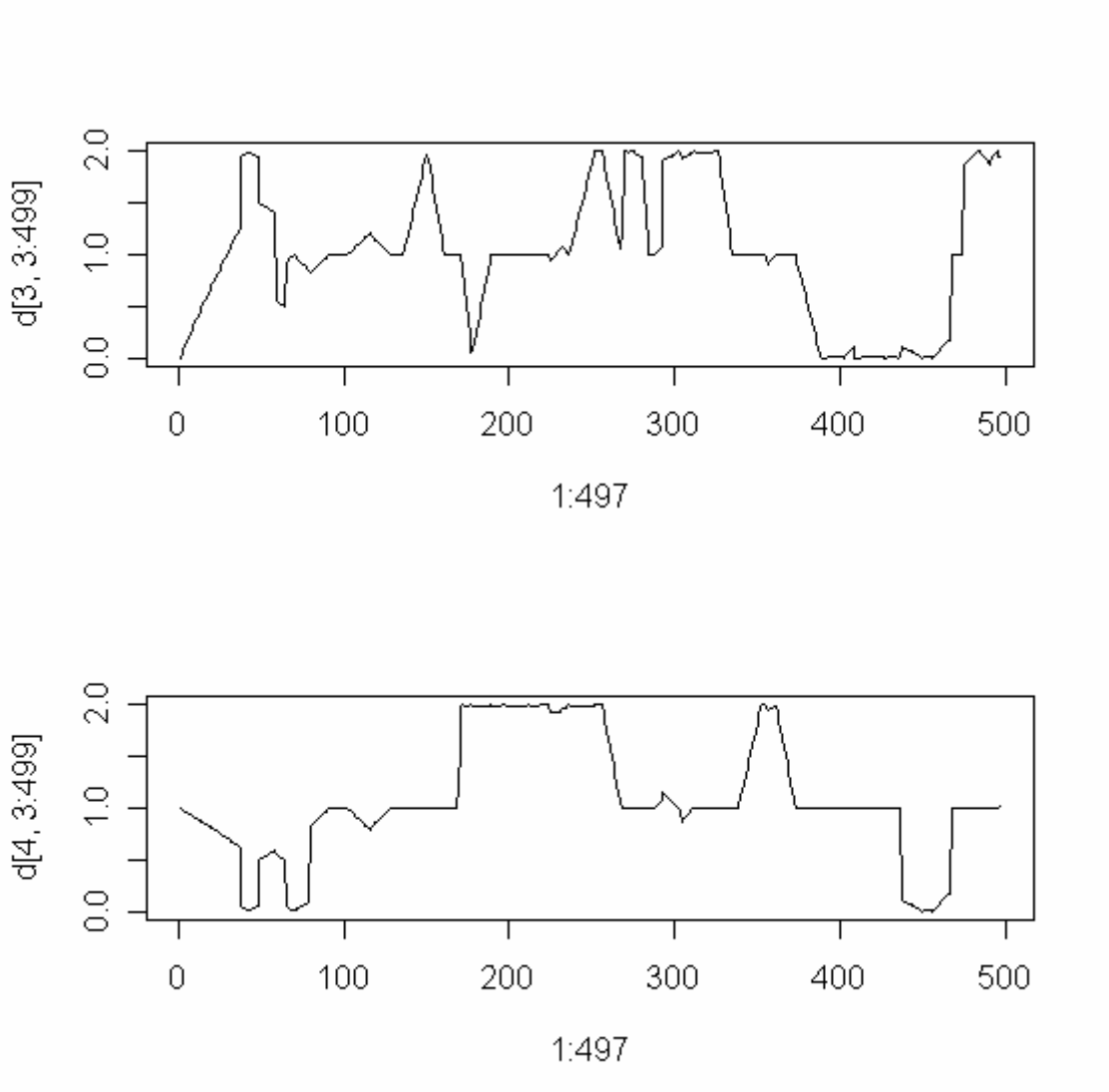

Figure 1.6: Estimated number of alleles for the example chromosomes

1.2 Goals

The main purpose of our paper is to compare the performance of chi-square and permutation tests for the detection of single QTL at a specified location (local) or chromosome (global) at a given level of significance. First we should make sure if there is a QTL at the chromosome, and then the specific location of QTL becomes the point we are interested in. According to these two levels of goals, we have two levels of hypothesis separately.

1.2.1 Global hypothesis

Assuming that Y_i has distribution of F_{x_i}, then the global null-hypothesis is:

$$H_0: \ F_{x_i} = F \quad \text{For all } i = 1, \ldots, 770 .$$

Which means "the distribution of Y_i does not depend on any of the locations". In other words, there is no gene on the chromosome affecting chicken body weight.

1.2.2 Local hypothesis

If the global null hypothesis has been rejected, the question concerned now is that given location j, if the gene at the $j'th$ location affects chicken body weight. This gives 497 null hypotheses at 497 locations.

H_0^j: The gene at the $j'th$ location does not affect chicken body weight

H_α^j: The gene at the $j'th$ location does affect chicken body weight

1.2.3 Multiple hypothesis testing

In our case, if we simply infer that the global hypothesis can be rejected if any one of the local hypotheses is rejected, then the multiple hypothesis testing problems occurs. Actually the global hypothesis is a multiple hypothesis of the 497 local

hypotheses. The global null hypothesis is true when all the 497 local null hypotheses are true. A multiple hypothesis testing problem may arise when we infer about the global hypothesis using results from the local ones. In this section, we will discuss how to solve this multiple comparison problem.

Because of the large number of local hypotheses, it is much easier to observe one local hypothesis being rejected even if all local hypothesizes were actually true. Thus it is much easier to reject the global hypothesis when the null hypothesis is true, which means an increasing of the type I error. This kind of type I error is called false positive. So we can not reject global hypothesis just simply because one local hypothesis is rejected.

To solve this problem, we employ a global statistic T^*, instead of the test statistic T used to test the local hypothesis. The larger T is, the less likely to reject the null hypothesis. So at the location with maximum T along the chromosome, it would be the hardest to reject the global null hypothesis. Then employing $T^* = \max(T)$ as the global test statistic should be an easy way to avoid the multiple hypothesis error. Using a global test statistics actually converts the multiple-testing problem to a single one. The problem is that it may be difficult to find the null distribution of the global test statistic, which we will discuss later.

Chapter 2

Method

2.1 Previous study

There are several of different methods for detecting the QTL in an experimental cross. Karl William Broman (1997) suggested distinguishing methods which model a single QTL at a time from those which attempt to model the effects of several QTLs at once. We focus on two popular single QTL methods here: Interval mapping and Regression mapping. A brief discussion about their advantages and disadvantages is also given at the end of this section.

The previous researchers often focus on the example of a backcross experiment because of its simplicity. At each location in the genome, the backcross offspring have one of only two possible genotypes, HL an HH. In practice, the intercross experiment is more complicated but also more commonly used. But the method developed for backcross experiment can also be employed while analyzing intercross experiment. In this chapter, since the aim is to introduce the basic idea of QTL method, we focus on backcross experiment as well.

2.1.1 Interval mapping

Lander and Botstein (1989) improved a method called “Interval mapping”, which is currently one of the most popular methods for identifying QTL.

They assume that there is a single QTL, and the backcross individuals have phenotypes which are normally distributed with mean μ_H or μ_L (according to whether their genotype is HH or HL) and common variance σ^2. Also the assumption of no interference and a genetic map specifying the locations of the markers are used in this method.

Consider a model with two markers separated by a recombination fraction r and putative QTL located between them, at a recombination fraction of r_L from the left marker, the distance between the two markers are d cM. According to Haldane map function (Haldane 1919), $r = \frac{1}{2}(1 - e^{-2d/100})$ and $r_L = \frac{1}{2}(1 - e^{-2d_L/100})$. Thus the recombination fraction between the QTL and the right marker is

$$r_R = \frac{1}{2}(1 - e^{-2((d-d_L)/100)}) = (r - r_L)/(1 - 2r_L)$$

Table 2.1 below shows the four possible sets of genotype at the two marker locations, and the conditional probability for each QTL genotype given the marker genotype respectively. For each of the four sets of marker genotypes, the conditional phenotype density function can be written in the form of a mixture of two normal distributions. Then the likelihood function for the four parameters (μ_H , μ_L , σ , r_L) can be obtained.

Table2.1: Conditional probability for the QTL genotype given the two marker genotypes

Marker genotype		**QTL genotype**	
Left	Right	HH	HL
HH	HH	$(1-r_L)(1-r_R)/(1-r)$	$r_L\ r_R/(1-r)$
HH	HL	$(1-r_L)\,r_R/r$	$r_L\ \ (1-r_R)/r$
HL	HH	$r_L\ \ (1-r_R)/r$	$(1-r_L)\,r_R/r$
HL	HL	$r_L\ r_R/(1-r)$	$(1-r_L)(1-r_R)/(1-r)$

Lander and Botstein proposed to use the maximum likelihood method to get the estimates of the four parameters by using the EM algorithm. The equation for the LOD score statistic is as follows:

$$LOD = \log_{10} \frac{\sup_{\Theta_0} L(\theta \mid x)}{\sup_{\Theta} L(\theta \mid x)}$$

H_0: No QTL, which means that the individuals' phenotypes follow a normal distribution.

H_α: It is a QTL at the current location

The LOD score needs to be calculated for each location and the likelihood under alternative hypothesis should also be calculated for each location. This method requires large amount of computation since the EM algorithm must be performed at each location.

The LOD score obtained from the above procedure is plotted against genetic locations and compared to a threshold. We infer the presence of a QTL at where the LOD curve exceeds the threshold, and the point that LOD curve reaches its maximum is used as the estimate of the QTL location. The region around the inferred QTL which included the estimate of the QTL location is used to be an interval estimate for QTL.

The genome-wide threshold, which indicates the significance of a peak in the LOD curve, is obtained by finding the 95th percentile of the maximum LOD score under the null hypothesis of no QTL.

The chief benefit of interval mapping is that it gives more precise estimates of the location and effect of a QTL than other methods according to Broman.

2.1.2 Regression mapping

Knapp et al. (1990), Haley and Knott(1992), and Martinez and Curnow(1992) independently developed a method called "regression mapping", which approximates interval mapping quite well but require much less computation. The main idea of this regression mapping method is followed.

Consider the same model described in previous section with two markers and a putative QTL between them. The conditional expected value of the phenotype for an individual with given marker genotype is

E (y|marker gen) = $\mu_L + (\mu_H - \mu_L)$ Pr [4](QTL genotype is HL|marker gen)

In regression mapping, the backcross individuals' phenotypes are regressed on Pr(), which is the conditional probabilities for given genotype of HL at the putative QTL. The likelihood function is calculated under the assumption of that given marker genotype, the individuals' phenotypes having a normal distribution, which is:

$$\text{y|marker gen} \sim N(\tilde{y}, \sigma^2),$$

where $\tilde{y}$ = E (y|marker gen).

Thus the LOD score can be calculated by:

$$LOD = \frac{n}{2}\log_{10}(\frac{RSS_0}{RSS}),$$

Where：

n = the number of individuals;

$RSS_0 = \sum_i (y_i - \bar{y}_i)^2$, means the residual sum of squares under null hypothesis of no QTL;

$RSS = \sum_i (y_i - \hat{y}_i)^2$, means the residual sum of squares from the regression above.

Following gives out a brief comparison between interval mapping and regression mapping. Like interval mapping, the LOD score is calculated at each location in the chromosome of regression mapping, and the similar inference procedure is also complemented. It is proved by Broman (1997) that the difference between the two methods is very small. But obviously this method is easier in calculation; we need to only perform a single linear regression at each location, instead of performing the EM algorithm at each location which saves a great computation time. Also because this method needs only simple linear regression, it is easy to include additional effects into the model, such as sex and so on. This may be practically attractive to certain research work.

[4] The Pr() is shown in Table 1

2.1 QTL method and model used in this paper

As we said in the previous chapter, the main aim of this paper is to compare two kinds of statistical test for the existence of a QTL at a specified location (local) or chromosome (global). So first, LOD curve and the estimate of QTL are needed as the basis of performing a significant test. In this section, we will introduce the methods and models that are actually used for the analysis

We choose regression mapping to calculate the LOD score for the two reasons that we mentioned in the previous section:

1) Regression mapping gives result as good as interval mapping but needs much less computation.

2) It is easy to include additional effects into the model, such as sex. In our case, sex is a real factor which may affects the chicken body weight, so it is reasonable to take sex into account in our model.

Using a regression mapping method, the most likely location of a QTL can be estimated by fitting a simple linear regression model for the 497 locations which are included in our data. Notice that the regression model is not the main purpose of this thesis and we just want to use it to construct a test statistic which has different statistical properties under the null and alternative hypothesis. The regression models that we use are as follows:

We let Y_i denote the phenotype of an individual i, which is the weight of the $i'th$ chicken in our data, $i = 1, \ldots, 770$, Y_i follows normal distribution. And X_i denote the expected number of alleles at certain locations of the $i'th$ chicken, which is ranged from 0 to 2. We imagine two models:

$$\text{Null model: } Y_i = \mu + sex_i + \varepsilon,$$

$$\text{Alternative model: } Y_i = \mu + sex_i + X_i\beta + \varepsilon,$$

where ε is independent and identically normal distributed error.

The regression mapping gives the LOD score as:

$$LOD = \frac{n}{2}\log_{10}(\frac{RSS_0}{RSS}),$$

where n is the number of F2 individuals and

$RSS_0 = \sum_i (y_i - \bar{y}_i)^2$ is the residual sum of squares under the null model,

$RSS = \sum_i (y_i - \hat{y}_i)^2$ is the residual sum of squares from the alternative model.

Now a question may come that concerns about the independent variable. In regression mapping, the individuals' phenotypes are regressed on Pr(), their conditional probabilities for having the genotype HL at the putative QTL given their marker genotypes. Here, the individuals' phenotypes are regressed on the numbers of H at the putative QTL. Then next step is how to come from Pr() in theory to number of H in practice.

As what we said in previous section, the QTL methods developed for backcross experiment also work for intercross experiment. But the question is that we do not have this simple probabilistic interpretation in the intercross experiment as in backcross experiment.

In backcross, with x (the number of H) equal to 0 or 1, we have

$$E(y|\text{marker gen}) = \beta_0 + \beta_1 * E(x|\text{marker gen}) = \beta_0 + \beta_1 * Pr(x=1|\text{marker gen})$$

When in intercross with x equals to 0, 1 or 2, E(y|marker gen) will not equal to a probability. We could write:

$$E(y|\text{marker gen}) = \beta_0 + \beta_1 * Pr(x=1|\text{marker gen}) + \beta_2 * 2 * Pr(x=2|\text{marker gen})$$

Under the assumption of additive effect of the genes, the contribution of two genes are the twice of one gene, which means $\beta_1 = \beta_2$. Then we can get the equation below:

$$\begin{aligned} E(y|\text{marker gen}) &= \beta_0 + \beta_1 * Pr(x=1|\text{marker gen}) + \beta_1 * 2 * Pr(x=2|\text{marker gen}) \\ &= \beta_0 + \beta_1 * (Pr(x=1|\text{marker gen}) + 2 * Pr(x=2|\text{marker gen})) \\ &= \beta_0 + \beta_1 * E(x) \end{aligned}$$

$$= \beta_0 + \beta_1 * X_i$$

So actually they are the same model.

2.2 Test for the significance

2.2.1 χ^2 Test

Since the LOD score itself is a likelihood ratio test statistic and approximately have a χ^2 distribution with 1 degree of freedom, we can just use the values of LOD score to get a series of P-value according to the location.

This method assumes that the regression model is correct and thus the LOD score indeed has a χ^2 distribution. This distribution assumption should be satisfied when doing this test.

If the P-value is smaller than a certain significant level α, H_0 could be rejected and the locations with these low P-values can be the possible locations affecting chicken body weight.

For a **global hypothesis**, which concerns whether a specific chromosome affect chicken weight, the equation of P-value is:

$$P_G = P\,(LOD \geq \max(LOD_j)) \qquad (1)$$

In order to calculate this P_G, we should know the distribution function of $\max(LOD_j)$. It may be possible in theorem to derive the distribution function using transformation. But it would be very complicated and unpractical to do this transformation, especially with 497, such a large number of locations. So it doesn't make sense to test the global hypothesis using χ^2 test

For a **local hypothesis**, which concerns whether there is a specific QTL affecting chicken weight, the equation of P-value is:

$$P_j = P_{\chi^2}(LOD \geq LOD_j) \tag{2}$$

If $p_j \leq \alpha$, reject $H_0{}^j$ at location j, which means the gene at the $j'th$ location does affect chicken body weight.

A further discussion about χ^2 Tests

χ^2 Tests can be directly given by LOD score and is supposed powerful if the distribution assumption is satisfied. The remaining of the section will take a further discussion on how to check the assumption and what will happen if the assumption is not satisfied.

The theory we need to use here is from "Bootstrap Methods and their Application" written by A.C. Davison and A.V. Hinkley.

If $X \in R$ is a random variable with a continuous distribution function F, then the random variable $Z = 1 - F(X)$ should have an uniform distribution on [0，1]. Hence, if F_{T_0} is the distribution function of some variable like $t(Y)$ under null hypothesis, then $p(Y) = 1 - F_{T_0}(t(Y))$ should have a uniform distribution under null hypothesis.

An easy way to check this is to plot the histogram of $p(Y)$. It should be uniformly distributed only if all null-hypothesizes are true. However, under the assumption that there is only one or a few QTLs, we can assume that almost all null hypotheses are true and hence the distribution is supposed to be close to uniform in our case.

If the histogram of $p(Y)$ is similar to the uniform distribution, the distribution assumption is supposed to be satisfied. But on the other hand, if the histogram is far different from uniform distribution, the distribution assumption is considered not well satisfied, which means that χ^2 Tests may be not suitable under this situation.

Moreover, QQ plot can be drawn against χ^2 distribution to check the LOD

score distribution.

2.2.2 Permutation test

As we just said, χ^2 Tests has a strong assumption about distribution. So if it is not satisfied, we need other method to get the significant level. Permutation test may be a good choice since no distribution assumption is needed.

The study of Churchill and Dogerge(1994)

A great effort has been put into finding an appropriate LOD threshold. Churchill and Doerge (1994) developed a method based on permutation to measure the significance of the QTL effect. In order to effectively sample from the distribution of the test statistics under a null hypothesis of no QTL, the quantitative trait data are permutated with respect to original data for a large number of times like 500 or 1000 usually. This method does not depend on the distribution assumption of the variety.

They also mentioned that the number of permuted times N is an important point of the test since it will determine the accuracy. And through experience the number of N=1000 is supposed to be adequate for estimating critical values at a significance level of α =0.05. We follow this premise is our research.

Monte-Carlo permutation test

A permutation test can be obtained based on Monte-Carlo methods. Monte-Carlo methods are usually used to approximate the P-value of the permutation test. The main procedure to perform this method can be mathematically described as follows:

1) Draw N samples $y^1,...,y^N$, from the distribution specified by H_0

The trait values are simulated N times among the n individuals to create a permutated data set. Sampling from H_0 by random permutation of the vector of traits is based on the assumption of exchangeability which will be discussed later.

2) Compute $t_i^s = t_i(y^s)$ for $i = 1, 2, \ldots, N$

A test statistic (such as LOD score) is computed at each analysis location for all the permutated data as well as the original data.

3) Compute $\hat{p}(y) = N^{-1}\Sigma_{i=1}^{N} 1\{t_i^s \geq t_i(y)\}$

where $1\{t^s \geq t(y)\}$ equals to 1 if $t_i^s \geq t_i(y)$, or equals to 0 if $t_i^s < t_i(y)$.

Hence, $\Sigma_{i=1}^{N} 1\{t_i^s \geq t(y)\}$ is the number of permutated test statistics which the value is larger than the original one.

Then the P-value can be computed by calculating the proportion compare with the number of permutated test statistics which are larger than the original one to the sample size N.

Now we consider a simple example to make it clearer of how to calculate the P-value following this step. Using "T" as the test statistic which is given by the original data. Draw 1000 samples of traits and calculate T^s (s is from 1 to 1000) for each permutated data. If there are 20 values of T^s larger than T among the 1000 T^s, then P-value equals to 20/1000, which is 0.02.

4) Reject H_0 if $\hat{p} \leq \alpha$

Given a significance level α, we can infer if the null hypothesis can be rejected.

In our case, we have 770 individuals with trait value of chicken body weight, and 497 genetic locations for each individual. Then simulate the trait value for 1000 times to create permutated data sets. The 384 male and 386 female individuals are permutated separately according to the exchangeable assumption.

For a **global hypothesis**, which concerns whether a specific chromosome affect chicken weight, this test can be performed with test statistic $LOD^* = \max(LOD_i^*)$. The P-value of the global hypothesis is defined by:

$$\hat{p}_G = \frac{1}{1000}\Sigma_{s=1}^{1000} 1\{LOD^{*s} \geq LOD\} \quad (3)$$

Where:

i) $LOD^{*s} = \max(LOD_i^{*s})$ is the maximum value of the LOD score in each time that after permutation.

ii) $LOD = \max(LOD_i)$, is the maximum value of original LOD score.

Given a significant level of α, H_0 can be rejected if $\hat{P}_G \le \alpha$, which means there is no gene on the chromosome affecting chicken body weight.

For a **local hypothesis**, which concerns whether a specific QTL affect chicken weight, the equation of P-value is:

$$\hat{p}_j = \frac{1}{1000}\Sigma_{s=1}^{1000} 1\{LOD_j{}^s \ge LOD_j\} \quad (4)$$

Where:

i) $LOD_j{}^s$ indicates the LOD score at the $j'th$ location after permutation.

ii) LOD_j is the original LOD score at the $j'th$ location.

Then for a given significant level α, $H_0{}^j$ can be rejected at location j if $\hat{p}_j \le \alpha$, which means the gene at that location dose affect the chicken body weight.

A further discussion about the permutation test

As we said before, permutation test does not need distribution assumption, but it should be proved that the data set is exchangeable. In this section, we will discuss about this problem.

First, we want to introduce the definition of "exchangeable":

$y = (y_1, ..., y_n)$ is exchangeable if its density function $f(y)$ is invariant under permutation of y.

For example, $y = (y_1, y_2) \sim f(y)$, y is exchangeable if:

$$f((y_1, y_2)) = f((y_2, y_1))$$

There is a theorem given below which can help us to check if a data is exchangeable:

If $y_1,...,y_n$ are independent and identically distributed (iid), then they are exchangeable.

In our case, under null hypothesis $H_0 : F_{x_i} = F$, the trait value y_i is identically distributed since the distribution of Y_i does not depend on any of the genes. And we can assume almost all null-hypothesizes are true under the assumption of there is only one or a few QTLs. Hence, y_i is supposed to be close to identically distributed. Additionally, it is reasonable to assume that y_i are independent.

To sum up, the trait value y_i is independent and identically distributed in our data, so the data set is supposed to be exchangeable and we indeed can use permutation test in our case.

Chapter 3

Results

3.1 Original LOD curve

For the method that we used, the LOD score should be calculated at each location and a LOD curve using original data by regression mapping is drawn against the 497 locations. Following Figure 3.1 shows that.

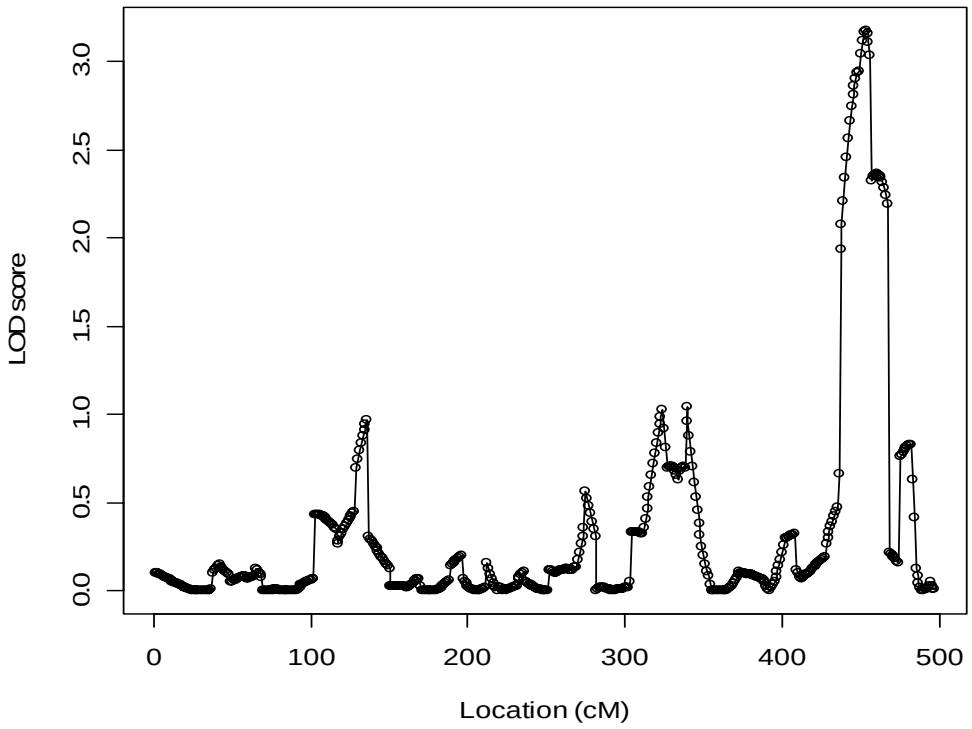

Figure 3.1: Original LOD curve

The location with the highest LOD score is supposed to be the most likely location affecting chicken body weight. From what been showed on Figure 3.1, the maximum LOD score appears at around location 450, which is much larger than other regions. Actually, it reaches the peak of 3.1782 at location 453. Now we are going to investigate if such a large value is likely to occur by chance only.

3.2 χ^2 Tests results

3.2.1 Global hypothesis

According to what we have mentioned before in Chapter 2, χ^2 test is not suitable for the global hypothesis, so we do not present the result here.

3.2.2 Local hypothesis

Following equation (2) in Section 2.2.1, we can get P_j. Figure 3.2 shows the plot of P_j against 497 locations.

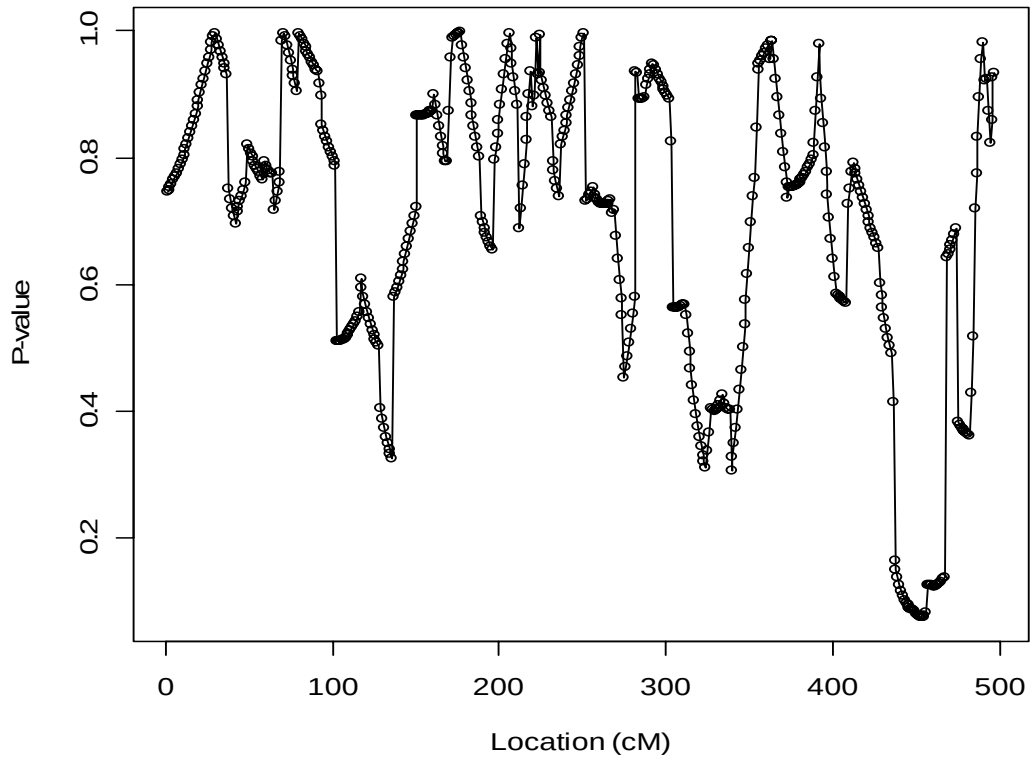

Figure 3.2: P-values for chi-square test under local hypothesis

Obviously, the P-value set reaches its lowest point around the location 450, where the LOD score also reaches its highest point. Specifically, P-value reaches its

bottom of 0.0746 at location 453, which can be seen as an estimator of the most likely QTL location.

If we choose α=0.1 as the significant level, the individual local H_0^j can be rejected at locations 445, 446,…, 456, which include location 453 (the estimator of QTL).

3.2.3 Check the distribution assumption for χ^2 Test

The P-value of χ^2 Test can be given directly by the LOD score and this test is supposed to be powerful if the distribution assumption can be satisfied. The remaining of the section will focus on the problem of checking the assumption satisfaction level.

A histogram of the P-value obtained from our χ^2 Test under local hypothesis is followed as Figure 3.3.

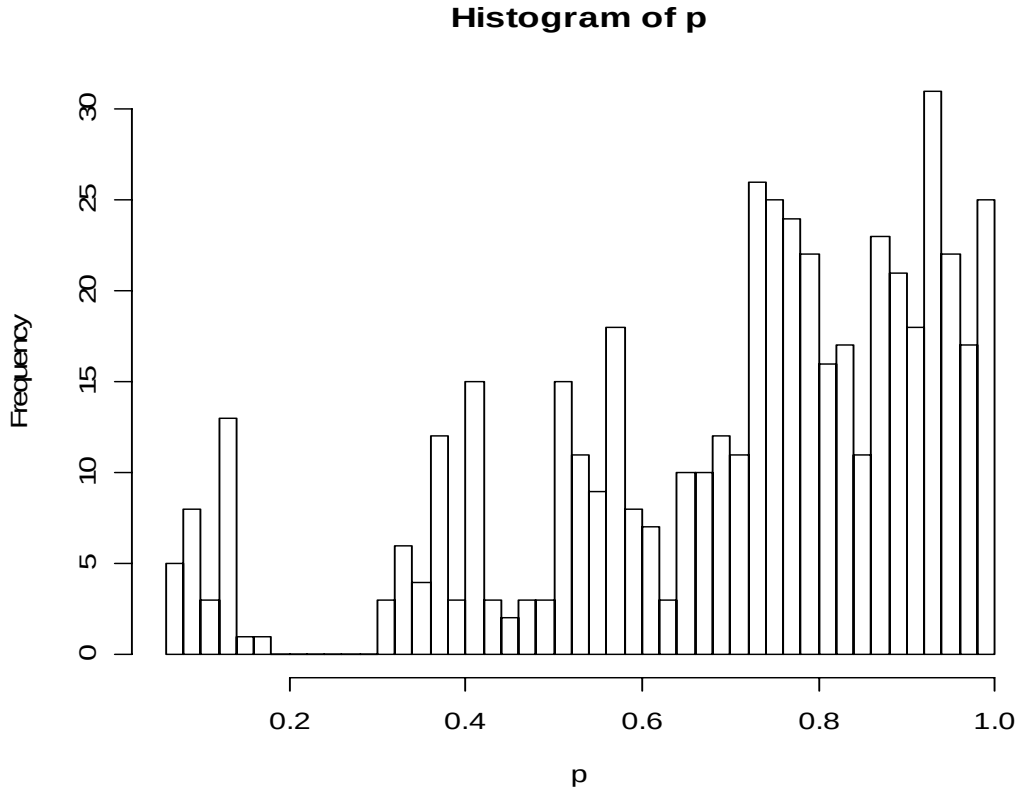

Figure 3.3: Histogram of P-value for chi-square test

The P-values will come from a uniform distribution if all the null-hypothesizes are true. And under the assumption that there are only one or a few QTLs, the P-values are closely come from a uniform distribution. Hence, the histogram should look like uniform.

But it can be seen in this Figure 3.3 that there is no P-value located in the region from 0.2 to 0.3, which means the distribution of P-value is far different from uniform distribution. Thus the distribution assumption is not well satisfied.

Also, we draw a QQ plot of LOD against χ^2 distribution to check the distribution assumption.

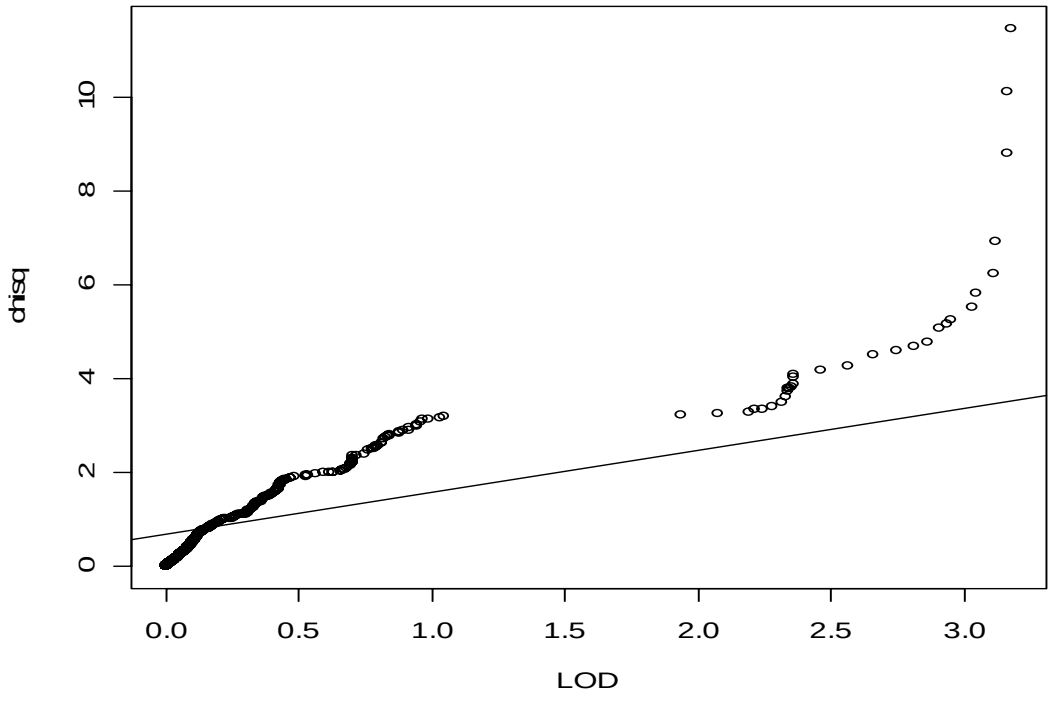

Figure 3.4: QQ plot of LOD score against Chis-quare

It can be seen clearly that the QQ plot is far away from the QQ line, which indicates that distribution assumption is not satisfied. So the χ^2 Test is not suitable here.

3.3 Permutation test results

3.3.1 LOD curves from permutation

We get 1000 LOD curves after permutation. We randomly choose 5 permutated LOD curve shown below from Figure 3.5 to Figure 3.9. These figures are generated from the 1st, 16th, 198th, 566th and 991st vector of the permutation LOD score separately.

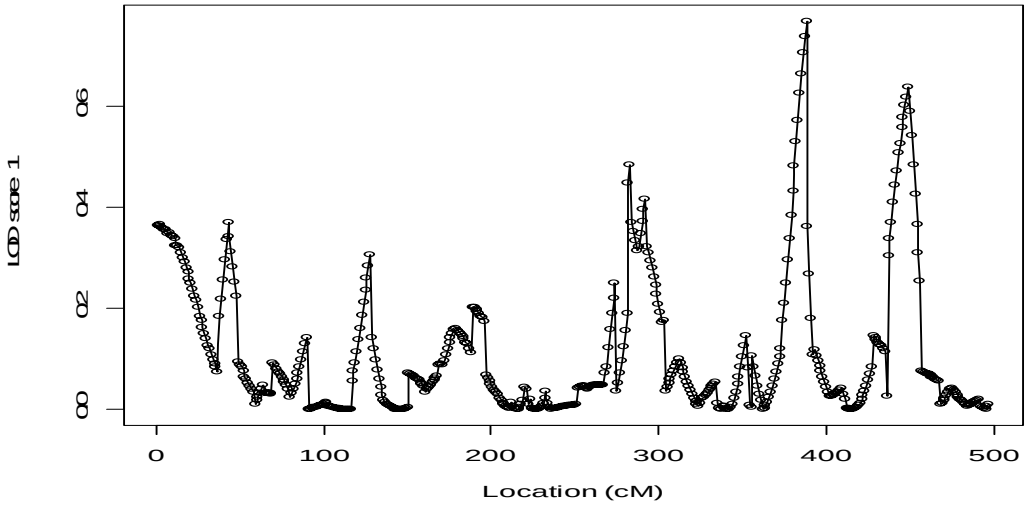

Figure 3.5: LOD curve after permutation of the 1st vector

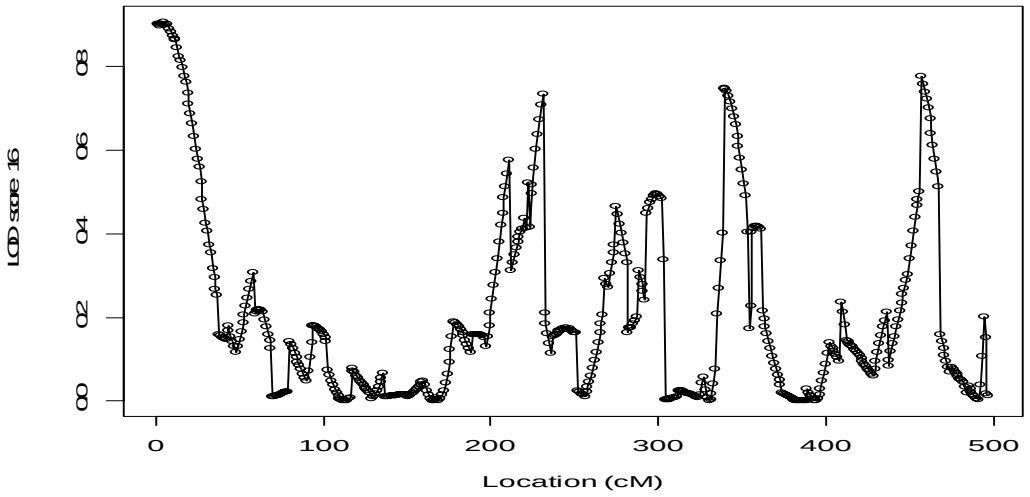

Figure 3.6: LOD curve after permutation of the 16th vector

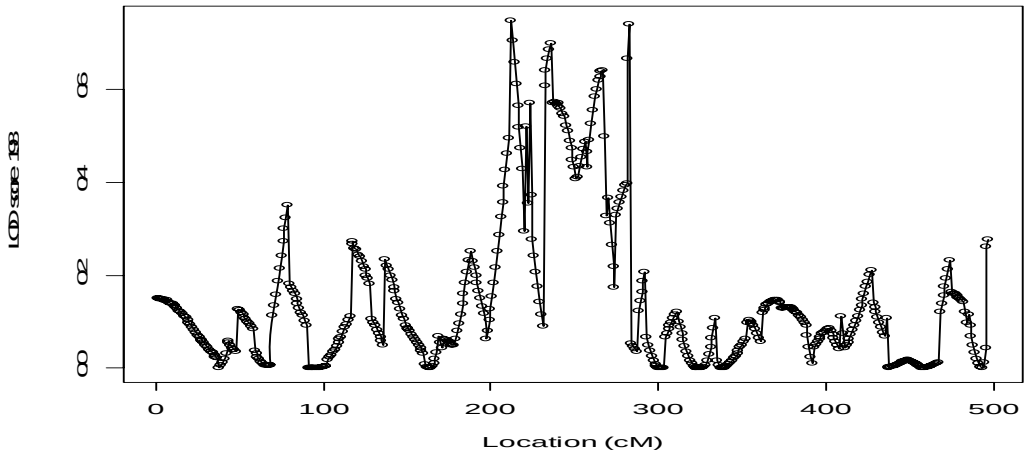

Figure 3.7: LOD curve after permutation of the 198th vector

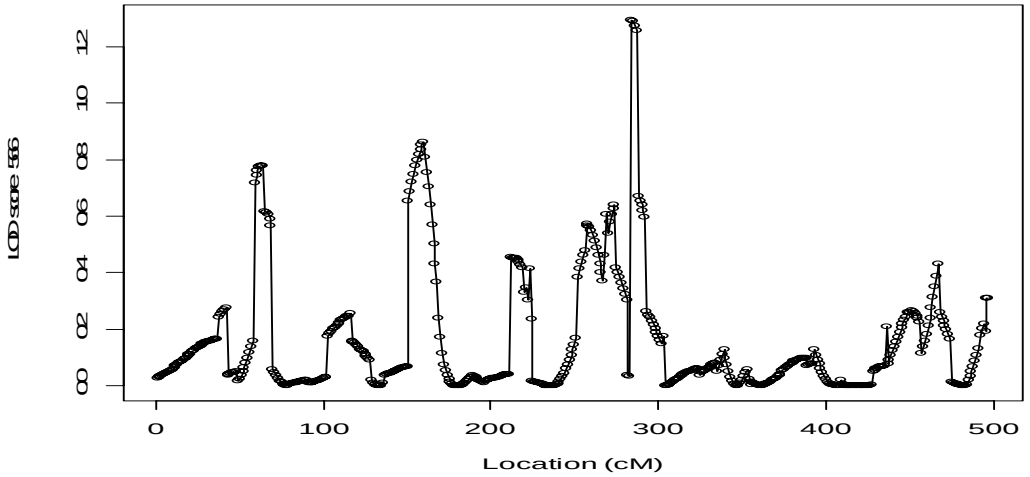

Figure 3.8: LOD curve after permutation of the 566th vector

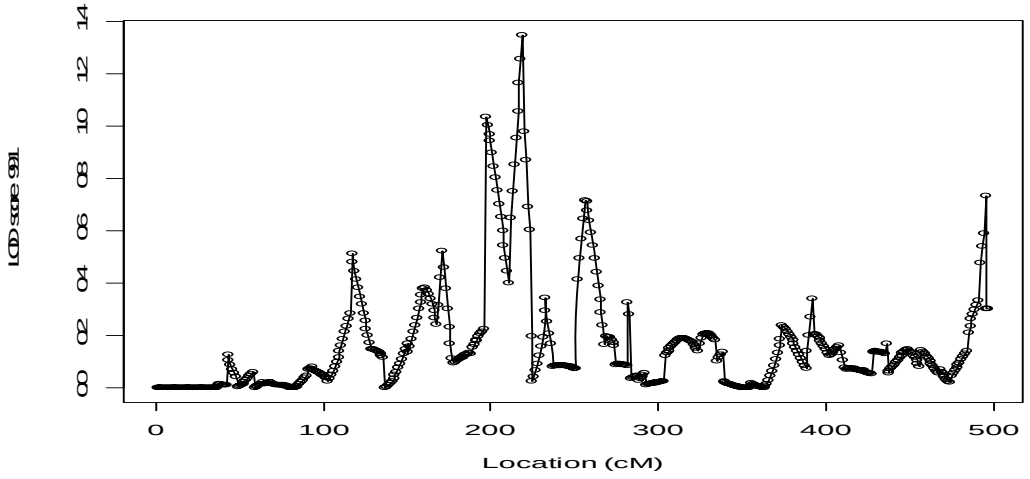

Figure 3.9: LOD curve after permutation of the 991st vector

As we can see in the figures above, the permutated LOD curves can be either similar or very different from the original one.

3.3.2 Global hypothesis

Follow equation (3) in Section 2.2.2, we get:

$$\hat{p}_G = 0$$

in our simulations.

Obviously, under the significant level α =0.05, the null hypothesis H_0 can be rejected, which means there is a gene somewhere on the chromosome affecting the trait value of chicken body weight. The maximum value of the original LOD score is 3.17822 and the maximum value of the permuted LOD score is 3.162535, so actually none of the LOD score after permutation is larger than the original one. So $\hat{p}_G$ equals to zero.

3.3.3 Local hypothesis

Follow the equation (4) in Section 2.2.2, we can get a series of $\hat{P}_j$ after permutation which called “PP”.

The graph of $\hat{P}_j$ for the permutation test is showed below in Figure 3.10:

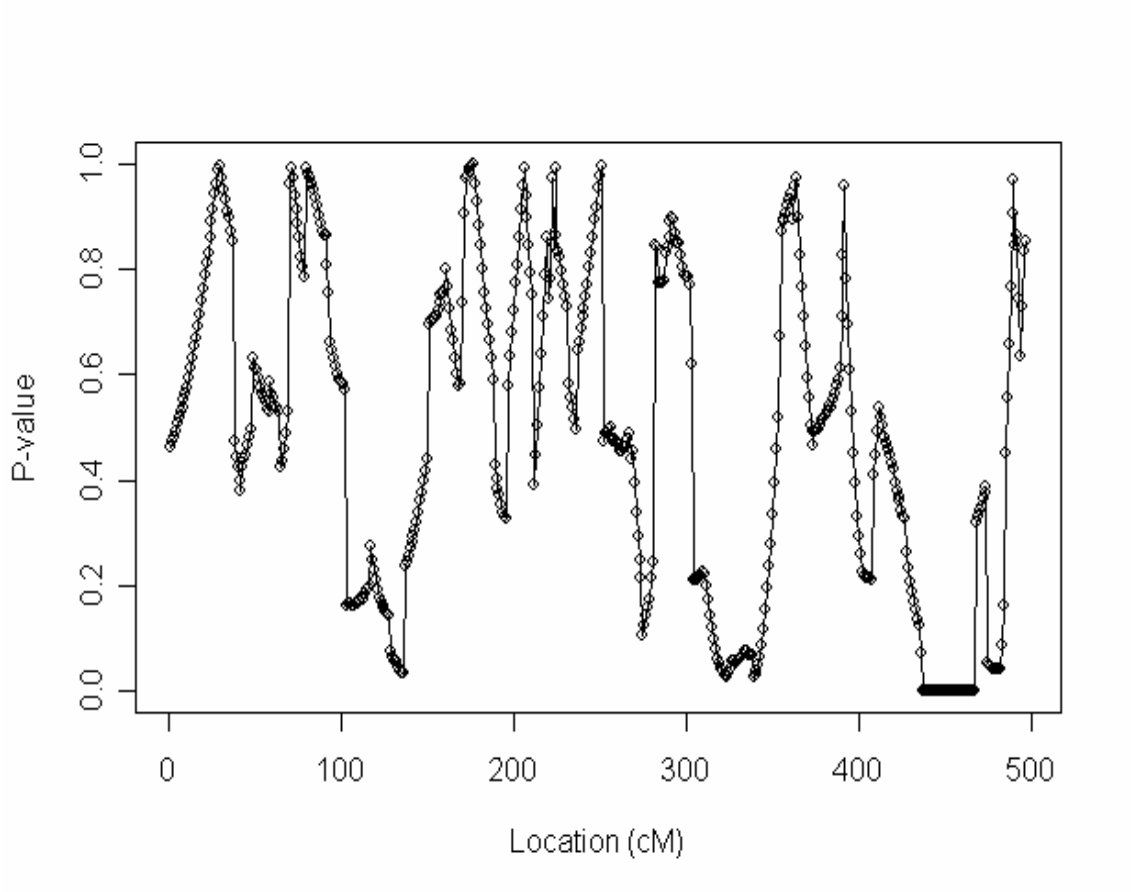

Figure 3.10: P-values for permutation test

We can see from the graph that the permutated P-value reaches its lowest point around location 450. Actually, P-values equal to 0 from location 439 to location 467. Choose α=0.05 as the significant level, so H_0^j can be rejected at locations 437 to 467, and locations 477 to 483. The estimate location of QTL, 453, is included in these regions.

Besides the most likely location region from 438 to 467 and from 477 to 483, Figure 3.10 also shows two regions with very small P-values. They are region from 100 to 150 and region from 300 to 350. The small P-values in these regions indicate that more than one QTL may exist in the chromosome. Thus, in our case, there may be a multiple quantitative trait loci (QTLs) problem.

We do not go further about this problem since it is not the point we focus on in this paper. But the problem of detecting multiple quantitative trait loci (QTLs) is important in much research work and it is worth to be further studied.

3.4 Comparison

Corresponding to the procedure of checking the assumption discussion of χ^2 Test in previous section, we also plot a histogram of permutated P-value here.

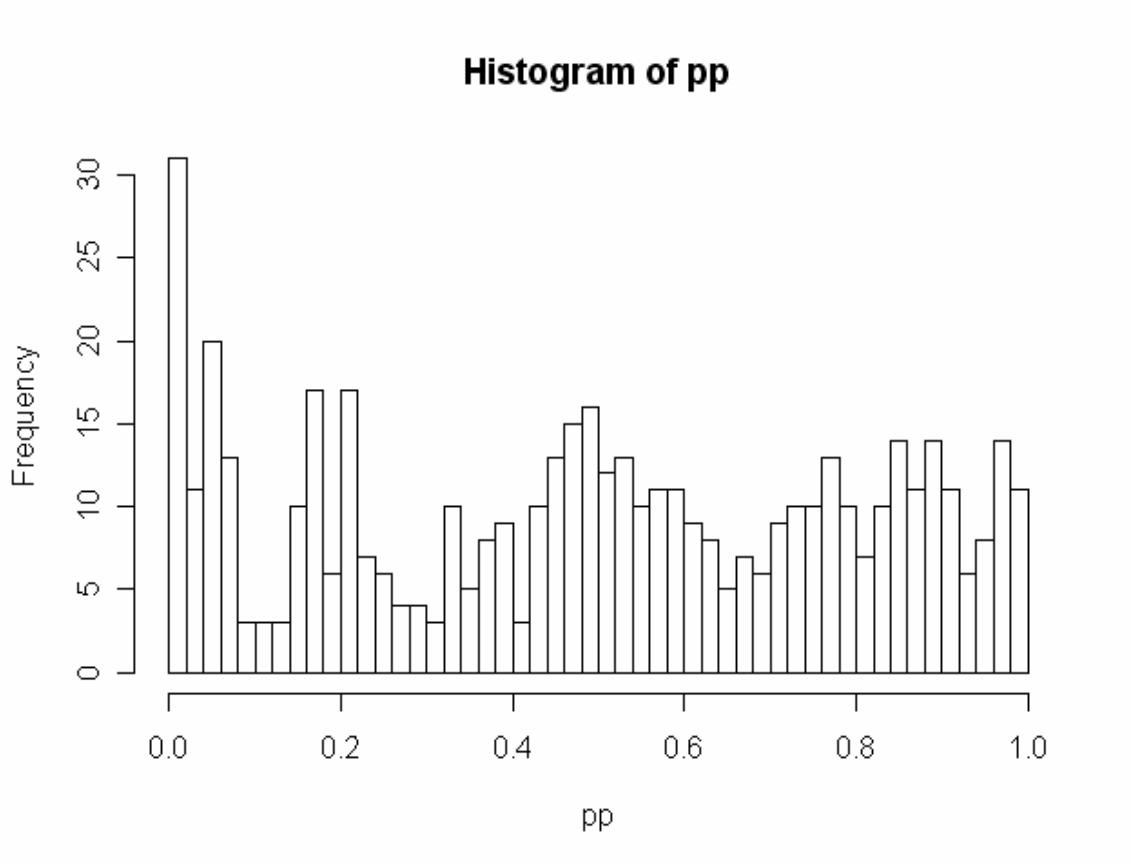

Figure 3.11: Histogram of P-values for permutation test

From the histogram of these P-values, we find that it is much closer to the uniform distribution than in χ^2 Test.

To sum up, we give a comparison of χ^2 Test and Permutation test:

Comparing to chi-square test, permutation test have two main advantages:

1) No distribution assumption is needed.

2) Can be used to test global hypothesis

Notice that chi-square could be powerful to test local hypothesis when distribution assumption is satisfied. Also, the data set should be exchangeable when using permutation test.

Chapter 4

Summary

Until now, we compare the performance of chi-square test and permutation test using a real data concerning the chicken body weight and get the conclusion after comparison. In our case, Monte-Carlo permutation test has two main advantages than chi-square test: no distribution assumption is needed, and can be used under global hypothesis.

Also, chi-square could be powerful to test local hypothesis when distribution assumption is satisfied, and the data set should be exchangeable when using permutation test.

Within the whole process of analysis, there are several problems that need further discussion.

1) Multiple QTLs

In section 3.3.2, we showed that there might be a multiple quantitative trait loci (QTLs) problem in our case. Detecting multiple QTLs is interesting in many biological studies and plenty of methods have been introduced. These methods include multiple regressions, interval mapping revisited, MQM mapping and so on. Doerge and Churchill (1996) also discussed the permutation test for multiple loci in their paper. Thus a further discussion can be made in this direction.

2) Other methods to solve multiple hypotheses problem

In section 1.2.3, we mentioned about the problem going along with multiple hypothesis testing, which we called false positive. To solve this problem, we employ $LOD^* = \max(LOD)$ as the global test statistic. Actually, previous studies have provided some alternative methods. For example, John D. Storey and Robert

Tibshirani (2003) introduced a concept named FDR (false discovery rate), and a measure of statistical significance called “q-value” is associated with each tested feature. And they suggested that FDR is a sensible measure of the balance between the numbers of true positive and false positive in many genome wide studies. However, in our study the local hypotheses are not independent, which makes computation of FDR difficult.

3) Develop other testing methods

We focus on chi-square test and permutation test in this paper. But there are still alternative test methods can be studied. This may be an interesting research direction.

Reference

1. Broman, Karl William (1997), *"Identification quantitative trait loci in experimental crosses"*
2. Churchill G.A., and, R.W.Doerge (1994) , *"Empirical threshold values for quantitative trait mapping"*, Genetics 138:963-971
3. Davison A.C., and D.V. Hinkley(1997),*"Bootstrap method and their application"*, Cambridge University Press
4. Doerge R.W., and G.A.Churchill (1996)，*"Permutation tests for multiple loci affecting a quantitative character"*, Genetics 142:285-294
5. Haley, C.S., and S.A.Knott (1992), *"A simple regression method for mapping quantitative trait loci in line crosses using flanking markers",*Heredity 69:315-324
6. John D.Storey, Robert Tibshirani (2003), *"Statistics significant for genome wide studies "*, Genetics statistics, 9440-9445
7. Knapp,S.J., Bridges, Jr., and D.Birkes (1990)*"Mapping quantitative trait loci using molecular marker linkage maps"*, Theoretical and Applied Genetics 79:583-592
8. Lander, E.S., and D.Botstein (1989) *"Mapping mendelian factors underlying quantitative trait using RFLP linkage maps"* , Genetics 121:185-199
9. Martinez, O., and R.N.Curnow (1992), *"Estimating the location and the sizes of the effects of quantitative trait loci using flanking markers"* , Theoretical and Applied Genetics 85:480-488

Appendix

Programmes in R:

1. Programme for Figure 3.1: Original LOD curve

```
rm(list=ls())
d=read.table("c:/ChickenData.txt")
for (i in 3:499){
g0=lm(d[,1]~d[,2])
g1=lm(d[,1]~d[,2]+d[,j])
RSS0=sum((fitted(g0)-d[,1])^2)
RSS=sum((fitted(g1)-d[,1])^2)
lod[i]=770/2*log10(RSS0/RSS)
}
plot(1:497,lodp,xlab="Location (cM)",ylab="LOD score")
lines(1:497,lodp)
```

2. Programme for Figure 3.2: P-values for chi-square test under local hypothesis

```
p=1-pchisq (lod,df=1)
plot(1:497,p,xlab="Location (cM)",ylab="P-value")
lines(1:497,p)
```

3. Programme for Figure 3.3: Histogram of P-value for chi-square test

```
hist(p,breaks=50)
```

4. Programme for Figure 3.4: QQ plot of LOD score against Chi- square

```
y=rchisq(1000,df=1)

qqplot(lod,y,xlab="LOD",ylab="chisq")

qqline(y)
```

5. Programme for Figure 3.5 to Figure 3.9: LOD curves from permutation

```
rm(list=ls())
d=read.table("c:/ChickenData_Order.txt")
d0=subset(d,V2==0,drop=TRUE)
d1=subset(d,V2==1,drop=TRUE)
lodp<-matrix(0,1000,497)
for (i in 1:1000){
```

```
attach(d0)
w0=sample(d0$V1)
attach(d1)
w1=sample(d1$V1)
cbind(w0)
cbind(w1)
W=c(w0,w1)
attach(d)
d$V1=W
for (j in 3:499){
g0=lm(d[,1]~d[,2])
g1=lm(d[,1]~d[,2]+d[,j])
RSS0=sum((fitted(g0)-d[,1])^2)
RSS=sum((fitted(g1)-d[,1])^2)
lodp[i,j-2]=770/2*log10(RSS0/RSS)
}
}
plot(1:497,lodp[1,],xlab="Location (cM)",ylab="LOD score 1")
lines(1:497,lodp[1,])
plot(1:497,lodp[16,],xlab="Location (cM)",ylab="LOD score 16")
lines(1:497,lodp[16,])
plot(1:497,lodp[198,],xlab="Location (cM)",ylab="LOD score 198")
lines(1:497,lodp[198,])
plot(1:497,lodp[566,],xlab="Location (cM)",ylab="LOD score 566")
lines(1:497,lodp[566,])
plot(1:497,lodp[991,],xlab="Location (cM)",ylab="LOD score 991")
lines(1:497,lodp[991,])
```

6. Programme for Figure 3.10: P-values for permutation test

```
pp=numeric()
for (s in 1:497){
n=0
for (t in 1:1000){
if (LOD[t,s]>=lod[s]){
n=n+1}
else {n=n}
}
pp[s]=n/1000}
```

7. Programme for Figure 3.11: Histogram of P-value for permutation test

```
hist(pp,breaks=50)
```