PHC 6937 Genetic Data Analysis
Final Project Report

Simulation Study on Confidence Interval of QTL Location

Huiyin Lu, Yang Wang, Yu Wang

Abstract

The confidence interval of Quantitative Trait Locus location is a very important quantity for genetic analysis. There were several constructive methods proposed including non-parametric bootstrap method and 1 drop LOD score method. In order to compare these two method, this study utilized the simulation to assess their performance in average proportion of coverage for true QTL location, average width of confidence interval by varying QTL location, marker density and effect size. We also investigated the impact of the missing marker genotype and marker genotyping error on the performance of two methods and the QTL estimation accuracy. We found there is no clear winner between these two method. Both of them can be biased depending on the QTL effect size and its relative position with markers. Hence, using these two method to derive the confidence interval is not proper in certain circumstance. Under 20% of marker genotype missing will not have significant impact on the performance of both method, except a little effect on the accuracy of QTL estimation. However, the marker genotyping error will have serious impact compare with mark genotype missing.

Key words: simulation, bootstrap, 1 drop LOD, confidence tnterval, marker genotype error, market genotype missing.

1.Introduction

Many genetics studies aim to locate the quantitative trait locus (QTL) which contribute to explain the genetics basis of variation in a quantitative trait (Miles,et al. 2008). And once

QTL position is determined, the confidence intervals for the position of quantitative trait loci are important to be investigated for further researches to identity the specific gene and the corresponding effects (Manichaikul et al. 2006). Nowadays, the LOD drop-off method is widely used to calculate the confidence intervals in QIL mapping. However, some literatures showed that the LOD supported confidence intervals have some drawbacks. For example, Mangin stated in their paper that when using LOD drop method, the test statistics does not actually chi-square distributed (Mangin et.al 1994). and as a result, when using LOD drop method, the coverage of the confidence intervals of QTL position is biased because it is affected by the effect of QTL (Manichaikul et al. 2006). Therefore, Visscher and his collaborators proposed to use the nonparametric bootstrapping as another method to find the confidence interval for the location of quantitative trait locic (Visscher et al. 1996). Nonparametric bootstrapping is a method to resample the data with replacement. The overall confidence interval for QTL position are obtain by combing the results from the CIs for QIL position which were estimated in each replicated dataset. However, recent researches found that the performance of the nonparametric bootstrap based confidence intervals of QTL position are not ideal in practice (Manichaikul et al. 2006). In reality, the coverage of nonparametric bootstrapping supported CI has prone to become biased and the degree of bias is affected by QTL position and its effects. Therefore, in our current project, our objective is to use simulation to implement LOD drop method and the nonparametric bootstrapping to find the confidence interval of the location of QTL to see how these two methods really perform in practice. After that, we are interested in the effects of the missing values in the missing genotypes and the genotyping errors.

PHC 6937 Genetic Data Analysis
Final Project Report

2. Methods

The dataset used in this project were simulated from a backcross population with sample size = 500.

Since the chromosome is expected to show symmetrical results around the mid point, in order to save the simulation running time, this project focused on the investigation of the QTL position between 0 to 50 cM (Walling et al. 1998). For each individual, single chromosome of 50 cM with the map density equals to 6 or 11 were simulated. In the current project, the regression model used to analyze the data is the method of Haley and Knott (1992). A constructive QTL at different positions on the chromosome was fitted by HK method to calculate the test statistic at each position (Visscher et al. 1996). And the QTL is assumed to locate at the position which has the largest test statistic. The effect sizes used in this project were varied to be 0.05, 0.5 and 1. Crossovers were generated under the assumption of Haldanes mapping function with our interference (Visscher et al. 1996). And the distribution of the environmental residuals follows the normal distribution. For each set of parameters, we simulated the BC populations with replication number = 1000.

In order to explore the effects of missing values in genotypes and the genotyping errors, we used the same method as stated above. The only difference between these procedures is to add the missing probability of the genotypes and genotyping errors when we simulated the data. We focused on to investigate the effects when map density = 11 and effect size = 0.05. The missing probabilities of genotypes and genotyping errors were set to be 0.05 and 0.20.

PHC 6937 Genetic Data Analysis
Final Project Report

3. Results

Our simulation study, consisting of 1000 replicates and 300 bootstraps for each confidence interval with 31 equal QTL positions on a chromosome of length 50 cM, allows us to figure out the confidence interval changing with a varying QTL position, effect and map density by bootstrap and 1 LOD drop methods. The results for the analyses of the bootstrap and 1 LOD drop method are shown in the following figures.

3.1 Coverage

QTL position: Results for the position of the QTL at altogether 31 equally spaced markers within 0~50 cM are presented in the following figure 1 and 2. When effect size is 0.05, maker space is 10, the coverage for both of the bootstrap and 1 LOD drop are higher than95% and perform more stable compared to the setting when maker space is 5. When effect size is 0.5, the coverage for the 1 LOD drop method are all higher than 95% under the two settings as maker spaces are 10 and 5. Whereas, the bootstrap methods performs relatively unstable, varying around 95%. A similar pattern was seen when effect size is 1.

In addition, the most striking feature of the three different effect sizes is that there are clear peaks for the QTL position to occur exactly at the marker location.

Effect size: Results for the three effect sizes setting are 0.05, 0.5 and 1, there corresponding plots are shown in figure 1 and 2. The left four plots in figure 1 display when effect size equals to 0.05, map density equals to 11, the coverage for both of the methods is greater than 95%. However, on the two ends with QTL position range from 0~5 and 45~50 cM, when map density equals to 6, the coverage is less than 95% for both of the two methods. The right four plots in figure 1 show that when effect size equals to 0.5, the coverage for both of the methods is close to 95% as the QTL is between the markers, while it's relatively higher than 95% as the QTL is exactly on the marker. The left four plots in figure 2 shows that when effect size equals to 1, the coverage with 1 LOD drop method is always higher than 95%. In contrast with bootstrap method, the coverage is fluctuating around 95%.

Map density: Results for different maker spaces are displayed in figure 1 and 2 as well. The marker spacing of 10 or 5 for the QTL mapping seems did not affect the coverage significantly. It's clear that the coverage perform similar results when marker spaces are 10 and 5 under the settings effect sizes equal to 0.5 and 1. The results may seem different when effect size is 0.05 with the two end of the chromosome.

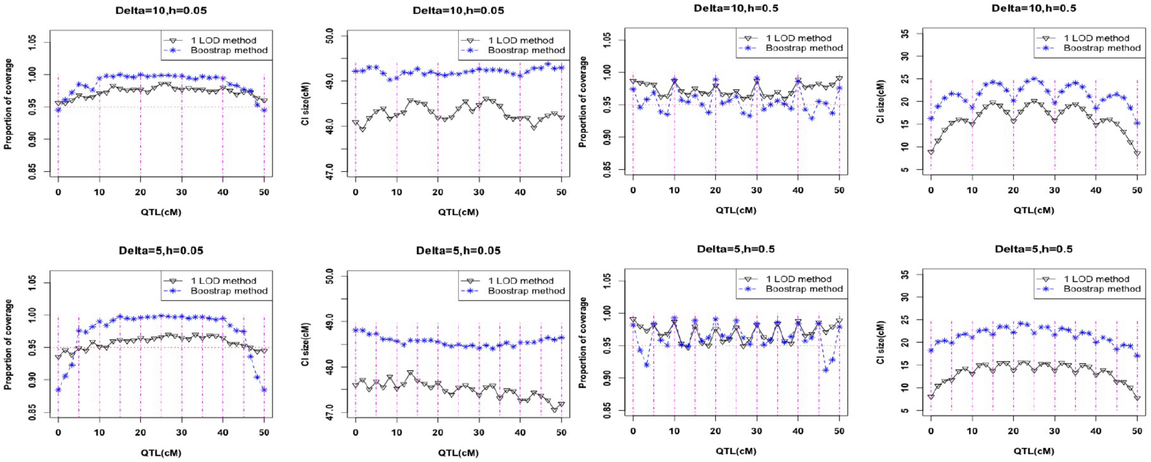

Figure 1. Coverage of 95% bootstrap confidence Intervals (blue), 1-LOD support intervals (black). With map density equals to 6 and 11, effect size equals to 0.05 and 0.5. The dashed vertical red lines denote marker positions on the chromosome.

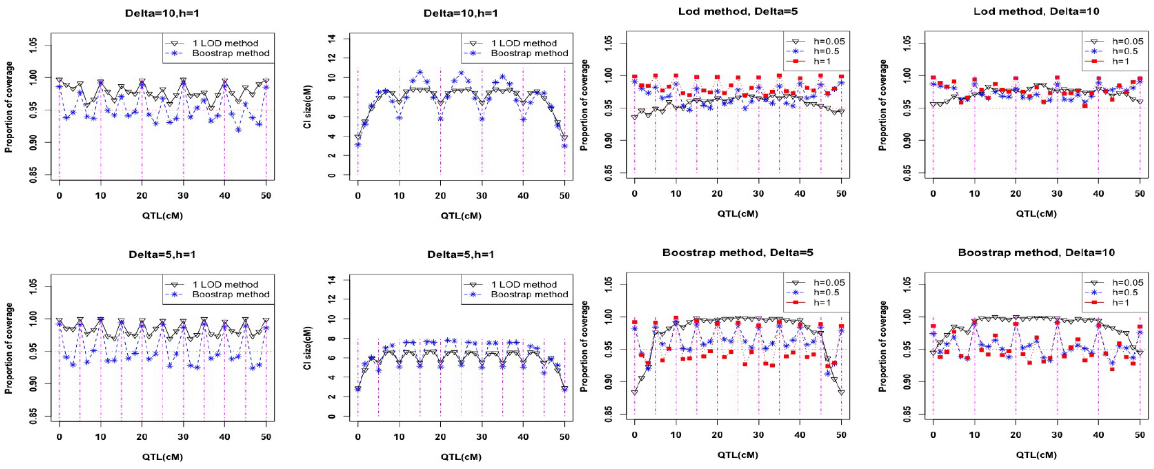

Figure 2. Coverage of 95% bootstrap confidence Intervals (blue), 1-LOD support intervals (black). With map density equals to 6 and 11, effect size equals to 1(The left four plots). Bootstrap and 1-LOD comparison with different effect sizes (The right four plots).The dashed vertical red lines denote marker positions on the chromosome.

3.2 Genotype missing

Results for the genotype missing with rate at 0.05 and 0.2 are presented in figure 3. The influence for both the confidence interval coverage and average CI width are very limited when marker genotypes are missing. Moreover, no obvious difference for the impact of missing marker genotypes was found by applying the bootstrap and 1 LOD drop methods.

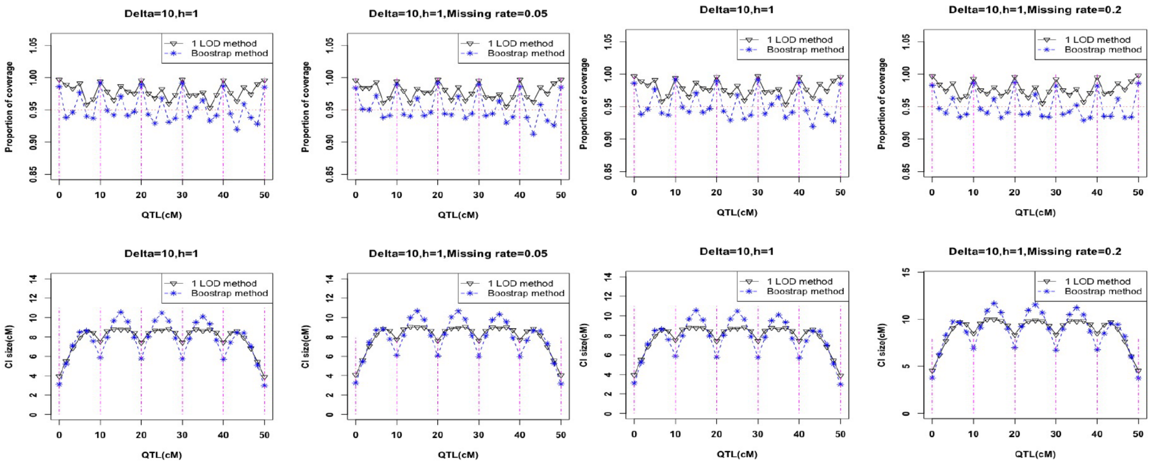

Figure 3. Coverage of 95% bootstrap confidence Intervals (blue), 1-LOD support intervals (black). With marker space 10, missing rate equals to 0.05 and 0.2. The dashed vertical red lines denote marker positions on the chromosome.

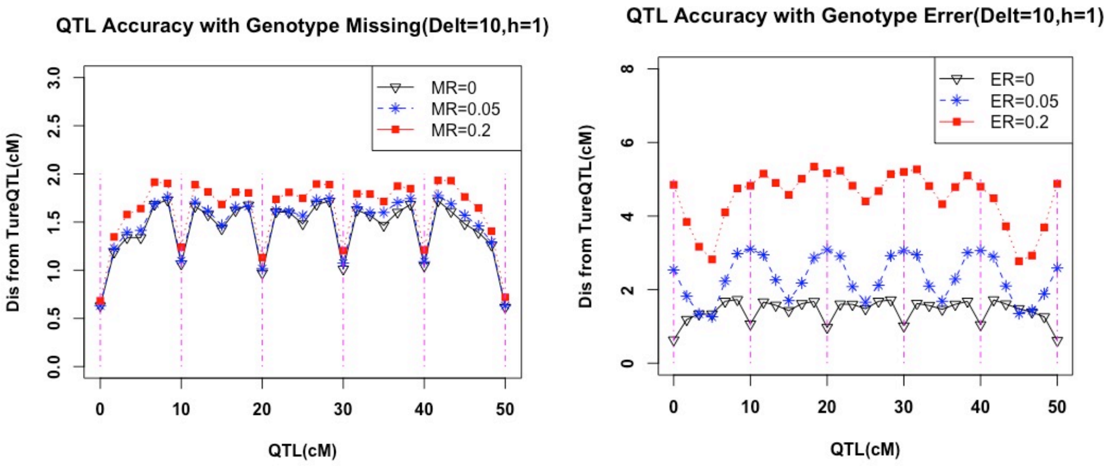

Figure4. Average distance from true TQL with marker genotype missing rate of 0(black),0.05 (blue) and 0.2 (red) (left). Average distance from true TQL with marker genotyping error rate of 0(black),0.05 (blue) and 0.2 (red) (right). With marker space 10, the dashed vertical red lines denote marker positions on the chromosome.

3.3 Genotype error

When marker genotyping error occurs, figure 5 shows the results with the error rate at 0.05 and 0.2. Both coverage curve and width curve of confidence intervals change a lot for both of the methods. For the 1 LOD drop method, the coverage fluctuates regularly that reaches the peak in the middle of two markers and declines to a nadir when it's exactly on the marker. For the bootstrap method, the coverage is generally stable around the 0.95 with error rate 0.05 and around 1 with error rate 0.2, but they both tend to perform poorly on the two ends of the chromosome. The influence of genotyping error to accuracy of QTL location estimation is showed as figure 4 left. The average distance from the true QTL position shift about 2cM when error rate increase from 0 to 0.05 or from 0.05 to 0.2.

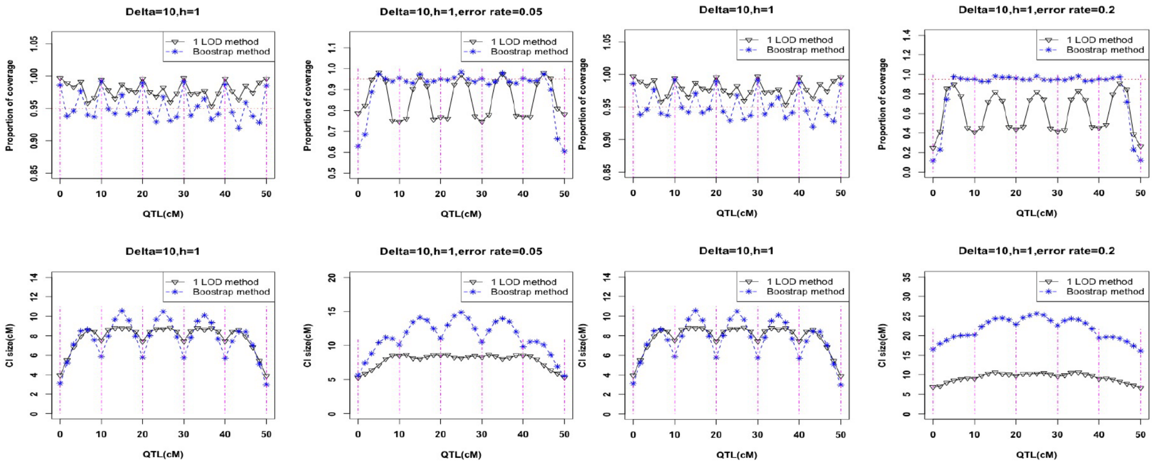

Figure5. Coverage of 95% bootstrap confidence Intervals (blue), 1-LOD support intervals (black). With marker space 10, genotype error rate equals to 0.05 and 0.2. The dashed vertical red lines denote marker positions on the chromosome.

4. Discussion and conclusion

As result indicated, there are many different parameters when comparing the performance of confidence intervals from the Bootstrap method and the 1 LOD drop method. Therefore, the discussion and conclusion regarding to different scenarios are necessary.

4.1 Comparison by effect size

I. When the effect size is relatively small (0.05).

Generally, the confidence interval from both methods are conservative (true coverage proportion bigger than 95%) with varied QTL location, especially the bootstrap method whose CI is even more conservative than the 1 LOD drop method. This probably because the average wide of CI from bootstrap is always bigger than the one for LOD drop method.

II. When the effect size is large (>= 0.5).

Bootstrap method works well when the QTL is between the marker, while tend to be conservative when the QTL is exactly on the marker. This may because its SE varies greatly regarding to the location of the QTL relative to the markers. While the 1 drop LOD method is tend to be conservative in this case.

III. When effect size varied from 0.05 to 1.

The 1 LOD drop method's proportion of the CI coverage is not affected significantly. The bootstrap method tends to perform better. The average width of the CI from bootstrap method is always higher than the 1 LOD drop method.

4.2 Comparison by marker density

When the marker density is varied from 5cM/marker to 10 cM/marker, the performance of both methods are not affected significantly. Except the true QTL is not on the marker before increasing the marker number while it is on the marker after increasing the marker number. This is result is consistent with other published literatures (Perer M Visscher et al.1996).

The average width of the CI also changes little with increasing marker spacing (consistent with Perer M Visscher et al.1996).

4.3 When there are missing marker genotypes

Its impact on both CI coverage and average CI width are very limited, with the missing rate varied from 0.05 to 0.2. However, the missing marker genotypes probably affect the accuracy of the QTL estimation, when the missing rate is higher than 0.2.

For different methods, the impact of missing marker genotypes did not show any obvious difference.

4.4 When marker genotyping error occurs

The performance of 1 LOD drop method is significantly affected, especially when the true QTL is on the markers or close to markers. This probably due to the wrong information was used, especially when the QTL is on the mark. The higher of the error rate, the stronger of the impact for 1 drop LOD method. However, the bootstrap method seems less sensitive to the genotyping errors; compared with the 1 LOD drop method. Its performance is quite stable even when error rate varied from 0.05 to 0.2. The non-parametric method is more robust in case of genotyping error.

The marker genotyping error also has much more serious impact on the accuracy of QTL estimation, compared with missing marker genotype.

Hence, both of 1 drop LOD method and bootstrap method are not always reliable and they are probably biased in certain cases. For 1 drop LOD method, it should be adjusted when using. If there is genotype error, it probably has serious effect for 1 drop LOD method.

There are other methods for deriving a QTL confidence interval proposed. A nominal 96.5 or 97% Bayes credible interval was shown to exhibit coverage near 95% for varied sample sizes, marker densities, and QTL effect size (Ani Manichaikul et al. 2006). A confidence interval based on a maximum likelihood ratio test was also proposed by B.Mangin (1995). A further comparison between all these methods simultaneously are needed in future.

LITERATURE CITED

Miles, C., and M. Wayne. "Quantitative trait locus (QTL) analysis." *Nature Education* 1.1 (2008): 208.

Manichaikul, A., J. Dupuis, S. Sen, and K. W. Broman. "Poor Performance of Bootstrap Confidence Intervals for the Location of a Quantitative Trait Locus." *Genetics* 174.1 (2006): 481-89. Web.

Mangin, B., B. Goffinet, and A. Rebai. "Constructing confidence intervals for QTL location." *Genetics*138.4 (1994): 1301-1308.

Visscher, Peter M., Robin Thompson, and Chris S. Haley. "Confidence intervals in QTL mapping by bootstrapping." *Genetics* 143.2 (1996): 1013-1020.

Walling, Grant A., Peter M. Visscher, and Chris S. Haley. "A comparison of bootstrap methods to construct confidence intervals in QTL mapping." *Genetical Research* 71.02 (1998): 171-180.

Peter M. Visscher, Robin Thompson and Chris S. Haley, 1996 Confidence Intervals in QTL Mapping by Bootstrapping. Genetics 143 1013-1020.

Ani Manichaikul, Jose ´e Dupuis, S ´aunak Sen and Karl W. Broman, 2006 Poor Performance of Bootstrap Confidence Intervals for the Location of a Quantitative Trait Locus. Genetics 174: 481–489.

Appendix A

PHC 6937 Genetic Data Analysis
Final Project Report

Major R code for simulation:

```
rm(list=ls())

library(qtl)

N <- 2 # The total number of confidence interval were generated.

n<- 300# The total number Bootstrap performed for each confidence interval.

Ef.size <-c(0.05,1)  #The effect size vector.

Chrom.length<-50

Posi <- c(0,1.7,3.3,5,6.7,8.3,10,11.7,13.3,15,16.7,18.3,20,21.7,23.3,

        25,26.7,28.3,30,31.7,33.3,35,36.7,38.3,40,41.7,43.3,45,46.7,48.3,

        50) # The QTL position vector.

Map.dens <- c(6,11)

map<- sim.map(Chrom.length, Map.dens[1], include.x=FALSE, eq.spacing=TRUE)

Effect<- Ef.size[2]

Indicator.1<-matrix(0,N,length(Posi))# coverage indicator for LOD drop method.

Indicator.2<-matrix(0,N,length(Posi))# coverage indicator for bootrap method.

CI.size1<-matrix(0,N,length(Posi))

CI.size2<-matrix(0,N,length(Posi))
```

```
Q.dist1<-matrix(0,N,length(Posi))
system.time(
   for(i in 1:length(Posi)){
      for(j in 1:N){
         set.seed(12+j)
         sim.data <- sim.cross(map, type="bc", n.ind=500, model =
rbind(c(1,Posi[i],Effect)))
         tt1<- calc.genoprob(sim.data, step=1)
         temp1<-scanone(tt1,chr=1, pheno.col=1, method="hk")
         CIs.1<-lodint(temp1,chr=1,drop=1)
         Indicator.1[j,i]<- as.numeric(Posi[i]>=CIs.1[1,2] & Posi[i]<=CIs.1[3,2])
         CI.size1[j,i]<-CIs.1[3,2]-CIs.1[1,2]
         Q.dist1[j,i]<-abs(CIs.1[2,2]-Posi[i])
         CIs.2<-scanoneboot(tt1,1,model="normal",method = 'hk',n.boot=n)
         lb<-quantile(CIs.2,0.025)
         ub<-quantile(CIs.2,0.975)
         Indicator.2[j,i]<- as.numeric(Posi[i]>=lb & Posi[i]<=ub)
         CI.size2[j,i]<- ub - lb
      }
   }
)
Coverage.1<- colMeans(Indicator.1)
Coverage.2<- colMeans(Indicator.2)
Size.1<-colMeans(CI.size1)
Size.2<-colMeans(CI.size2)
Distence<-colMeans(Q.dist1)
```