Identification of expression QTL (eQTL) of genes expressed in mouse islets

Yue Bai
UFID: 3979-8481
April 25, 2020

1. INTRODUCTION

1.1. BACKGROUND AND GOAL OF ANALYSIS

Genetical genomics (Jansen and Nap, 2001) is a popular method that utilizes genome-wide gene expression data and marker genotypes to detect and identify loci that affect variation in gene expression (Ponsuksili et al., 2010). These detected genomic loci are called expression QTL (eQTL).

To detect genomic loci that affect obesity-induced type II diabetes, an F_2 intercross between the diabetes-resistant $C57BL/6J$ (abbreviated B6 or B) and diabetes-susceptible $BTBRT^{+}tf/J$ (abbreviated BTBR or R) mouse strains were constructed at the University of Wisconsin–Madison (Tian et al., 2015).

In this study, I applied this F_2 intercross dataset to investigate the gene expression in mouse islets. To detect and identify the QTL and locations of the QTL, I performed the single-QTL genome scan on each probe. By integrating high-dimensional gene expression data into some composite traits, I inferred the eQTL genotype for all individuals based on the clustering of those composite traits. I also estimated the effects of multiple e-QTL on the gene expressions by applying the multiple mapping method.

1.2. DESCRIPTION OF DATA

The data are available at the Mouse Phenome Database: `https://phenome.jax.org/projects/Attie1`. There are mainly three parts data.

MICE GENOTYPE DATA There were 519 F_2 mice genotyped at 2,057 informative markers, including 20 on the X chromosome in the clean dataset.

GENE EXPRESSION MICROARRAYS DATA The Gene expression microarrays dataset included 40,572 total probes and contained three different forms of data:

- Gene expression profiling of 500 individuals (mice) for each tissue
- Tissue-specific mRNA pools for the tissue, which were used for the reference channel;
- mlratio data, which was attenuated as the ratio of the mean $\log_{10}$ intensity: $\log_{10}$(individual/pool).

MICROARRAY ANNOTATION DATA The microarray annotation data includes 37,827 probes information with genomic locations.

In this study, I only used the mlratio data and focused on the 37,796 probes with known cM position on some chromosome in the microarray annotation dataset. I used the key identifier for the probes to match the probe in annotation data with the gene expression data in microarray data.

2. METHODS AND MODELS

2.1. SINGLE INTERVAL MAPPING BY HALEY-KNOTT REGRESSION

For each probe in the islet tissue, I did single-QTL analysis by Haley-Knott regression (Haley and Knott, 1992). To deal with the missing genotype data problem, the hidden Markov model was used to compute the conditional probability distributions of the true genotypes based on the multi-point marker genotype data. The Carter-Falconer map function was used for mapping genetic distances into recombination fractions and the genotyping error rate was .2%. I also tried the standard interval mapping using the EM algorithm. As shown in Figure 2.1, the results from EM algorithm and Haley–Knott regression are almost the same. Therefore, for the rest of the paper, I only performed the interval mapping by Haley-Knott regression for the single-QTL analysis .

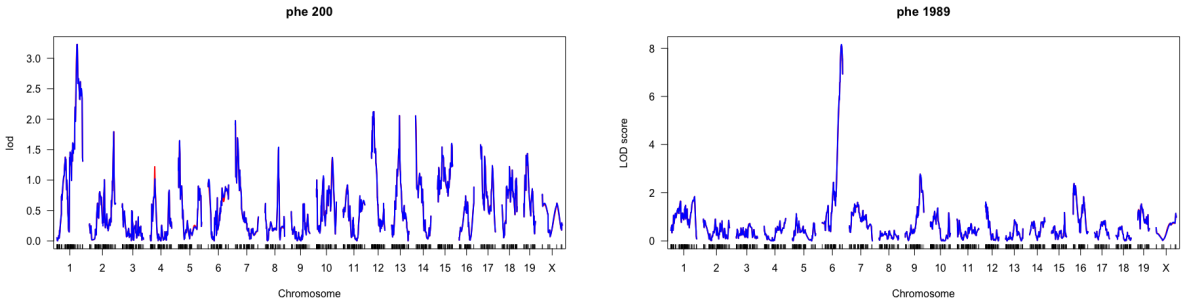

Figure 2.1: LOD curves from EM interval mapping (red) and Haley–Knott regression (blue).

2.2. QTL × COVARIATE INTERACTIONS

Suppose some covariates such as height or weight is related with the gene expression data, including those covariates into the QTL analyses may increase variation explained by the model and so increase power to detect QTL (Broman and Sen, 2009). Before performing the single-QTL genome scan, I needed to detect if there is any interaction effect between the genotype in some locus and the sex variable.

I selected some markers and probes, then constructed some plots of the sex-specific estimated effects of the inferred QTL. The results, in Figure 2.2, show that the marker "rs8262456" on chromosome 6 has some effect in both males and females. However, for different sex, the patterns of the effect are not the same. For probe "502269" (left panel), the BB (homozygous B6) individuals exhibit significantly smaller average gene expression than those RR (homozygous BTBR) or RB heterozygotes individuals in the male group. For probe "1002907112" (right panel), the RR (homozygous BTBR) individuals exhibit significantly larger average gene expression than those BB (homozygous B6) or RB heterozygotes individuals in the male group.

Therefore, in this study, I considered adding an interactive covariate "Sex" into the QTL mapping model. This means that the effect of the QTL can change with the covariate "Sex". The model is shown below:

$$y_i = \mu + \beta_x xi + \beta_g g_i + \gamma x_i g_i + \epsilon_i,$$

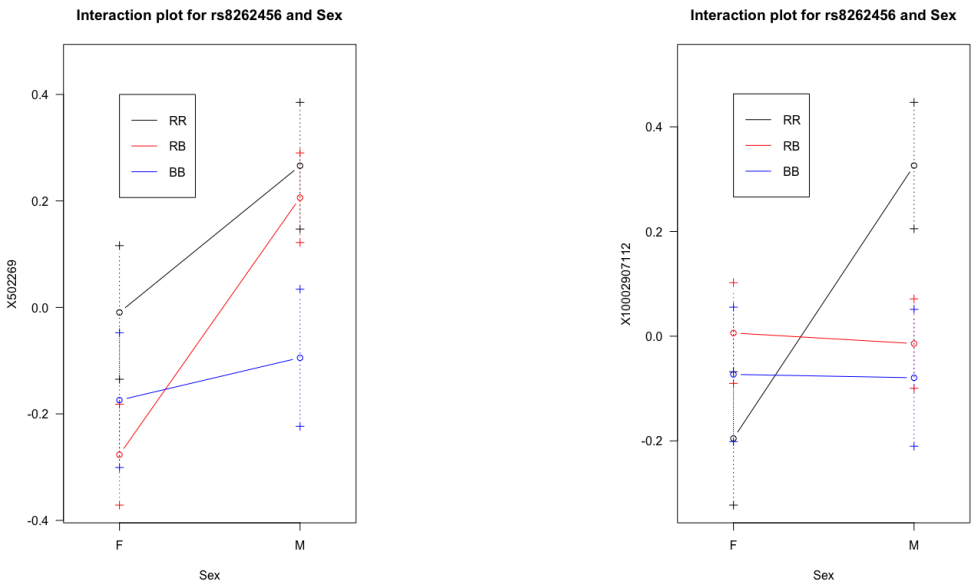

Figure 2.2: Estimated gene expression averages $\pm SE$ at selected probe "502269" (left) and probe "1002907112" (right) as a function of genotype at selected marker "rs8262456" and the "Sex" covariate.

where y_i is the gene expression measure of individual i, g_i is the genotype of individual i and x_i is the sex of individual i.

2.3. Principal component analysis

Many gene expression data in the probe were correlated or dependent on each other. I performed the principal component analysis (PCA) (Jolliffe, 1986) of the gene expression data that mapping to some locus with the large LOD scores, I applied the principle components with a large loading of gene expression information to convert the high-dimensional gene expression data to the low-dimensional data, i.e., some composite traits.

2.4. Multiple-QTL mapping

In the single-QTL analysis, for each probe and each genomic position, the hypothesis testing method is used to detect if there is a QTL. This single-QTL genome scan method may be incorrect if the gene expression phenotype is affected by multiple QTL (Broman and Sen, 2009). When dealing with some complex traits, one would expect that the trait is affected by multiple QTL. Therefore, I performed multiple QTL mapping in this study.

2.5. R package

All statistical analyses were performed by R. Package "R/qtl" (Broman et al., 2003) was used for QTL analysis, package "dplyr" (Wickham et al., 2018) was used for data manipulation, package "best-Normalize" (Peterson and Cavanaugh, 2019) was used for normalization transformations, package "ggplot2" (Wickham, 2016), package "ggcorrplot" (Kassambara, 2019) and package "factoextra" (Kassambara and Mundt, 2019) were used for further data visualization.

3. RESULTS

3.1. DATA DIAGNOSTICS AND SUMMARY

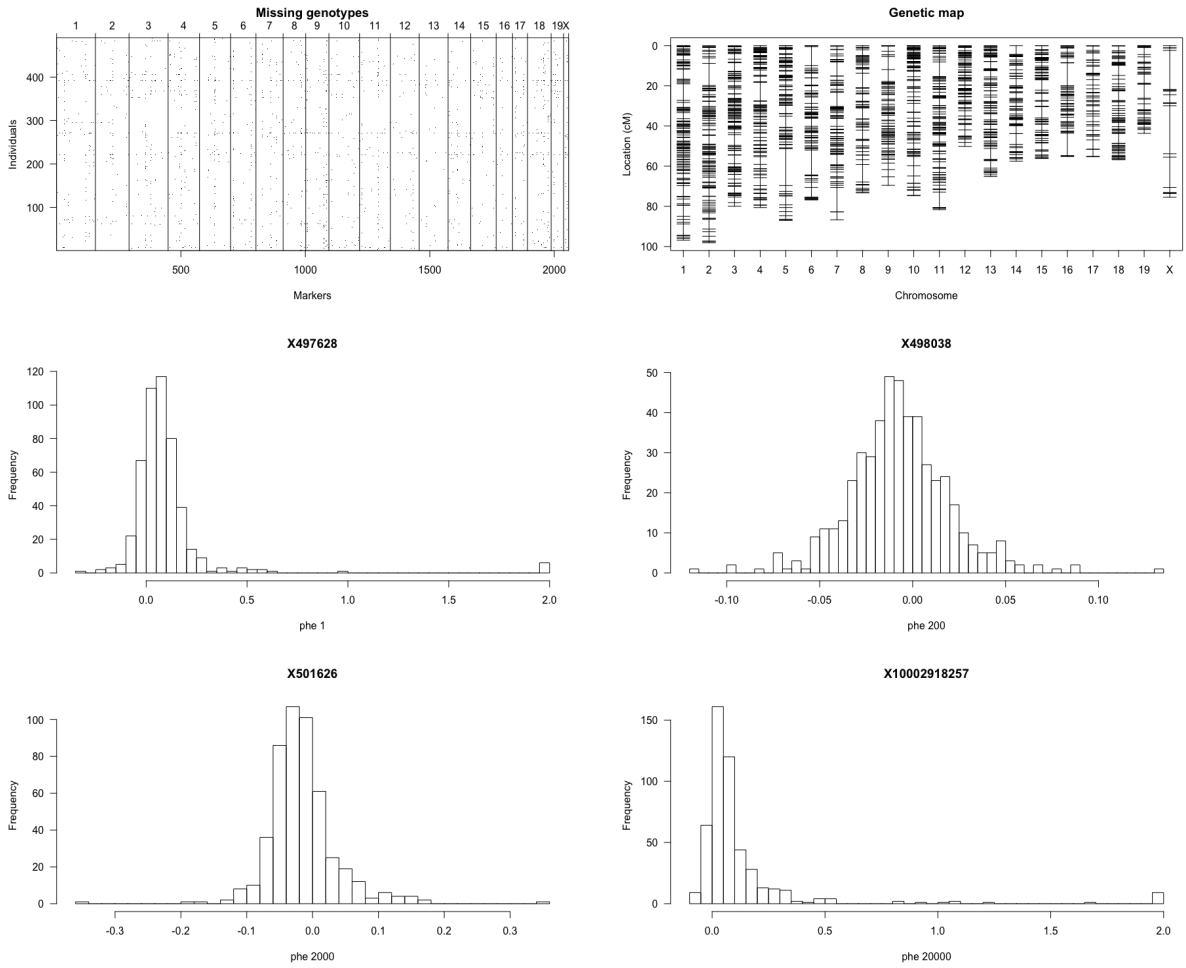

Figure 3.1: The summary of the "Attie1" dataset. Upper left: the plot of missing genotype data, black points represents missing genotype data; Upper right: the genetic map of all the markers on 20 chromosomes ; Lower: histograms of the four gene expression phenotypes.

Summary plots of the data are shown in Figure 3.1. The first plot of missing genotype data shows that there is no large scale of missing genotype data. The second plot shows that some large gaps of more than 20 cM existed in a few chromosomes, such as chromosome 5. Overall, the first two plots show that this dataset has a large number of markers with relatively small number of missing genotype data.

Also, there are some non-normal, highly skewed gene expression measures as we can see from the histograms in Figure 3.1. Most of the methods for QTL analysis requires the assumption that, the phenotype or the gene expression data follows a normal distribution based on the QTL genotype. The marginal distribution of the phenotype or gene expression data should be close to a normal distribution under the condition that there is no QTL (Broman and Sen, 2009). One strategy for dealing with the non-normal phenotypes is to transform the original data to quantiles of the standard normal distribution by using the ranked phenotype data.

To deal with the non-normal gene expression data, for each probe, I converted the gene expression data to quantile of the standard normal distribution. The transformation is:

$$z_i = \Phi^{-1}[(R_i - 0.5)/n],$$

where Φ^{-1} is the inverse of the standard normal cumulative distribution function, R_i refers to each observation's rank, and n refers to the number of individuals. All the QTL analyses are based on the transformed data.

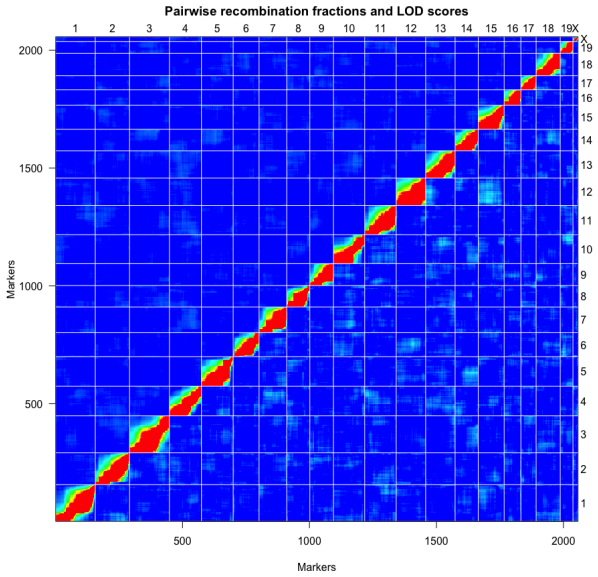

Figure 3.2: Plot of LOD scores for pairwise linkage (lower right triangle) and estimated recombination fractions (upper left triangle) for all the markers in the "Attie1" dataset. Red represents a small recombination fraction or a large LOD, while blue represents a large recombination fraction or a small LOD.

I also plotted the estimated recombination fractions and pairwise LOD for the 20 chromosomes. In Figure 3.2, the LOD scores for tests of pairwise linkage shows that little evidence for linkage between pairs of markers on different chromosomes.

3.2. Single-QTL analysis and identification of eQTL

Since there are a large number of probes and markers in this dataset, what I want to do first is to identify and analyse those QTL with large LOD score. Here, I performed a single-QTL analysis with an interactive covariate sex included for all gene expression measures in pancreatic islets tissue of 490 F_2 mice.

I only dealt with the single marker with maximum LOD on each chromosome for each probe in the islet tissue. After performing single-QTL genome scan for all the gene expression measures in 37,796 probes, I got 755,920 candidate eQTL which have the largest LOD score on the chromosome. By setting the threshold value of LOD score to be 10, there are 5,867 inferred eQTL with LOD $\geq$ 10. 2,484 eQTL at LOD score $\geq$ 25, 1,157 eQTL at LOD score $\geq$ 50 and 70 eQTL at LOD score $\geq$ 100. The putative eQTL for each chromosome and probe in the islet tissue with $LOD \geq 10$ are presented in Figure 3.3. The y-axis represents the genomic position (cM) of the probe in the islet tissue and the x-axis represents the genomic position (cM) of the marker on each chromosome.

In the Figure 3.3, there are mainly two significant patterns. First, there are extensive inferred eQTL points in the diagonal of the plot. It means that there are many local eQTL locate near the gene-of-origin, those genetic variants affect the expression of nearby genes. In total, there are 1,504 eQTL with LOD $\geq$ 10 that affect the expression of genes on the same chromosome. We referred to these eQTL as cis-eQTL. Second, there are also a lot of eQTL that strongly affect the expression of genes that have large distance from it or even on different chromosomes. We referred to these eQTL as trans-eQTL. There are multiple vertical lines in the plot. For example, there is a trans-eQTL hot-spot located at the chromosome 2 which influences the expression of a large number of genes located in multiple

chromosomes. Besides, the darkest vertical line is located at the end of chromosome 6. It shows that there is a extremely strong trans-eQTL hot-spot which affect the expression of genes located all around the genome.

The peak marker on the chromosome 6 is "rs8262456", located at 73.806 cM. One gene expression measures at probe "10002912806 is plotted against the genotype at the peak markers "rs8262456" in Figure 3.4. The RR (homozygous BTBR) individuals exhibit a larger average gene expression measure at probe "10002912806" than the RB heterozygotes, while the BB (homozygous B6) individuals exhibit a smaller average gene expression measure than the RB heterozygotes. This shows that this marker is linked to a eQTL. There are 1,382 probes with LOD ≥ 10, including 648 probes with LOD ≥ 25 and 23 probes with LOD≥ 100, in the 10 cM interval with this peak marker in the middle.

Besides the strongest trans-eQTL hot-spot on chromosome 6, there are four more inferred trans-eQTL that have LOD score ≥ 100. The peak marker on the chromosome 2 is "rs13476830", located at 73.804 cM. The peak marker on the chromosome 7 is "rs6396580", located at 61.7314 cM. The peak marker on the chromosome 9 is "rs13480398", located at 57.3410 cM. The peak marker on the chromosome 11 is "rs13481142", located at 52.2086 cM. The locations of these eQTL plus the trans-eQTL in chromosome 6 are displayed on the genetic map in Figure 3.5.

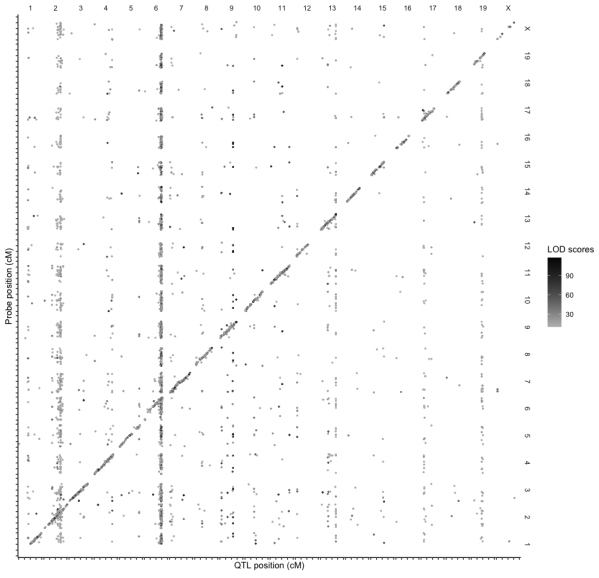

Figure 3.3: Inferred eQTL with LOD ≥ 10. The y-axis represents the genomic position (cM) of the probe in the islet tissue and the x-axis represents the genomic position (cM) of the marker on each chromosome.

3.3. INTEGRATION OF GENE EXPRESSION AND TRANS-EQTL GENOTYPE INFERENCE

There are 23 probes that mapped to some locus on a different chromosome with LOD ≥ 100. These eQTL are putative trans-eQTL.

I used principal components analysis (PCA) to transform the data into lower dimensions. The correlation matrix plot of gene expression data in 23 probes and the scree plot are shown in Figure 3.6. I obtained 23 principal components, which are called PC1-23. PC1 explains 69% of the total variance, which means that more than two-thirds of the information in the 23 probes dataset can be encapsulated by the first principal component. PC2 explains 6% of the total variance. In total, PC1 and PC2 can explain 75% of the variance.

From the correlation matrix plot in Figure 3.6, there are 19 probes' gene expression data that are highly correlated with each other, while the rest 4 probes' gene expression data are uncorrelated. For

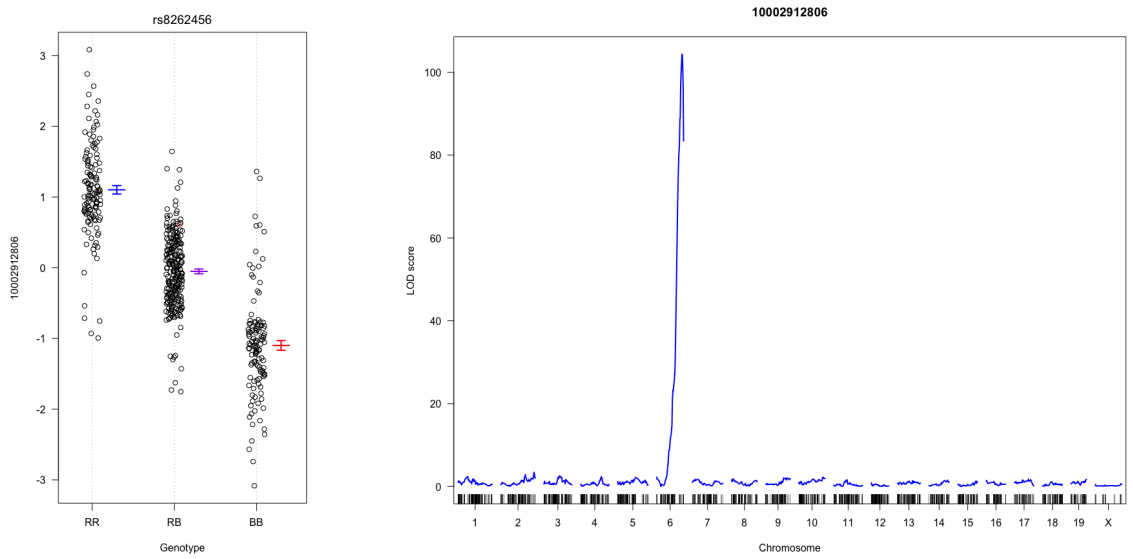

Figure 3.4: Dot plots of the gene expression measure at probe "10002912806" against the genotype at selected marker "rs8262456". For each genotype group, the colored segment represents the confidence interval for the average gene expression.

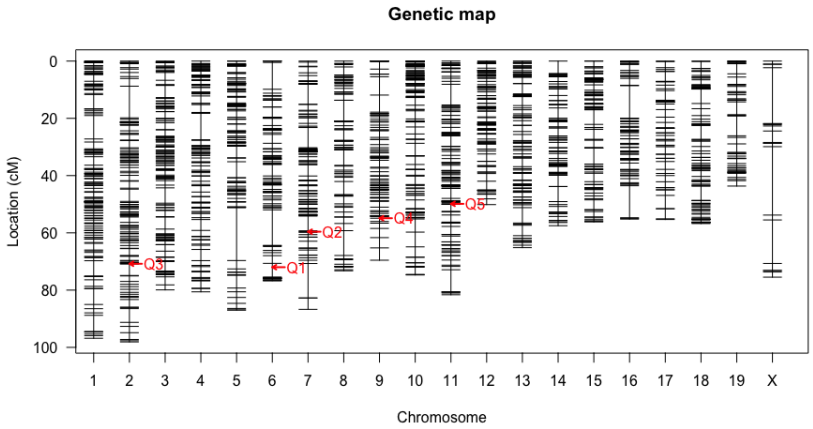

Figure 3.5: Locations of the eQTL with LOD ≥ 100 on the genetic map.

the 19 correlated probes, 11 of them are positively pairwise correlated while the rest of 8 probes are negatively correlated with the first 11 probes.

To distinguish the genotype of the trans-eQTL hot-spot on chromosome 6 for each individual, I chose these correlated 19 probes. All these 19 probes are mapped to some locus on chromosome 6 with LOD ≥ 100, while none of these probes is located on chromosome 6. I performed the principal components analysis of the gene expression data in these 19 probes. PC1 explains 83% of the total variance. PC2 explains 7% of the variance. In total, PC1 and PC2 can explain more than 90 % of the information.

The first two principal components of the gene expression data for each individual were displayed in Figure 3.7. The x-axis corresponds to the first principal component, and the y-axis corresponds to the second principal component. Points corresponding to mice are colored according to their genotype in the trans-eQTL hot-spot on chromosome 6. We can see from the scatterplot that there are three distinct clusters of mice.

Mice individuals with RR genotype on the chromosome 6 trans-eQTL hot-spot have larger amounts

of the first principal component than the other individuals with BB or BR genotype. Mice individuals with BB genotype on the chromosome 6 trans-eQTL hot-spot have smaller amounts of the first principal component than the other individuals with RR or BR genotype. In this case, by using the first principal component, the high-dimensional gene expression phenotype data was converted to a single composite trait for distinguishing and inferring the genotype of trans-eQTL hot-spot on chromosome 6.

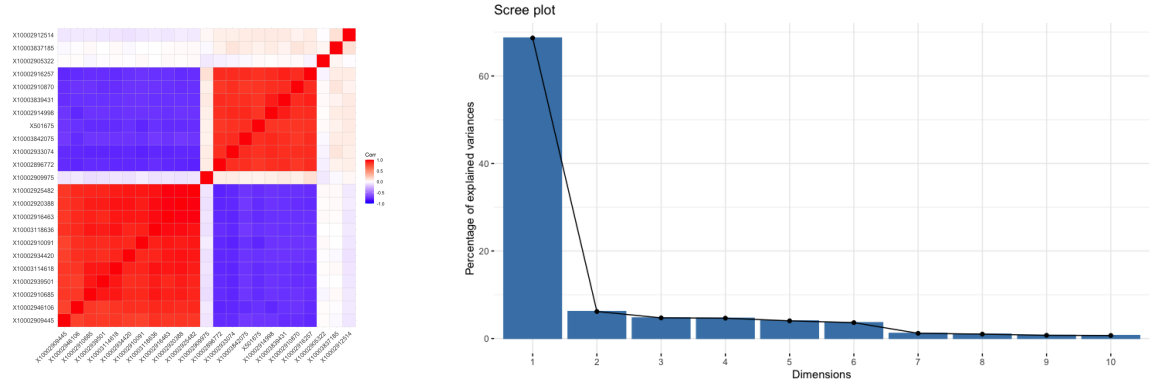

Figure 3.6: A correlation matrix plot of 23 selected probes which mapped to some locus with LOD ≥ 100 (left). Red represents large positive correlation, blue represents large negative correlation, and white represents small correlation (close to 0). A scree plot of the eigenvalues of principal components (right). The percentage of data variances explained by each principal component was shown in the scree plot.

Figure 3.7: The scatterplot of the first two principal components of the gene expression for each individual in the 19 probes. All these 19 probes are mapped to some locus on the chromosome 6 with LOD ≥ 100, while none of these probes is not located on chromosome 6. Each point represents a mouse individual.

3.4. MULTIPLE-QTL ANALYSIS

Based on the result in single-QTL genome scan in Figure 3.3, I found that many gene expression measure in the probe were affected by multiple loci in different chromosomes. In this part, I estimated the effects of multiple e-QTL on the gene expressions on some selected probe by applying the multiple mapping method.

Here, the probe "504047" on chromosome 5 was selected, and the multiple mapping was performed. I first performed a drop-one-QTL analysis to investigate the effect of each term in the model. The result in Figure3.8 indicates that the estimated percentage of the gene expression information explained by the full model is 42.5%. Under the null model that there is no effect, the LOD score is 59. In the drop-one-QTL analysis, one locus is removed from the full model each time, and the difference in the error between the reduced model and the full model is compared. The results show strong evidence for the effects of all of these loci but little evidence for the interaction.

The additive effect of an intercross is obtained by the half the difference between the mean of the gene expression data for the two homozygotes. The dominance effect is obtained by the difference between the mean of the gene expression data for the heterozygotes and middle value of the mean of the gene expression data for the two homozygotes (Broman and Sen, 2009). The estimated addictive effects for the chromosome 2, 7 and 13 loci being positive (0.748284, 0.416458, 0.174598, respectively) indicates that the RR homozygous have higher value of gene expression in probe "504047" than BB homozygous. The estimated dominance effects for the chromosome 2, 7 loci being negative (-0.007337, -0.038862 , respectively) indicates that the BR heterozygotes have lower value of gene expression in probe "504047" than RR homozygous. Because the amounts of dominance effects of the chromosome 2, 7 loci are much less than the addictive effects, it means that the BR heterozygotes have higher value of gene expression in probe "504047" than BB homozygous. Note that the estimated dominance effects for the chromosome 13 locus is positive (0.177801) and the amount of it is quite close to the estimated addictive effects, this means that the BR heterozygotes at the chromosome 13 locus have similar value of gene expression as RR homozygous. The interaction effect for the loci, however, is not significant.

Several other probes were selected to perform the multiple interval mapping. In most of the cases, I found that the gene expression in the selected probe is affected by multiple loci in different chromosomes but no interaction. In a few cases, however, the drop-one QTL analysis shows strong evidence for the loci as well as for the interaction. See the result of probe "10003843955" in Figure 3.9.

```
Full model result
----------------------------------
Model formula: y ~ Q1 + Q2 + Q3 + Q1:Q2 + Q1:Q3 + Q2:Q3 + Q1:Q2:Q3

       df       SS        MS      LOD     %var Pvalue(Chi2) Pvalue(F)
Model  26 207.9056 7.9963690 58.95911 42.54199            0         0
Error 463 280.8012 0.6064821
Total 489 488.7068


Drop one QTL at a time ANOVA table:
----------------------------------
                    df Type III SS    LOD    %var F value Pvalue(Chi2) Pvalue(F)
2@72.6              18     134.458 41.630 27.5130 12.3167        0.000   < 2e-16 ***
7@23.9              18      53.237 18.472 10.8935  4.8767        0.000  4.23e-10 ***
13@64.8             18      23.048  8.393  4.7161  2.1112        0.003   0.00501 **
2@72.6:7@23.9       12       5.456  2.048  1.1165  0.7497        0.666   0.70241
2@72.6:13@64.8      12       7.959  2.974  1.6286  1.0936        0.321   0.36326
7@23.9:13@64.8      12       7.419  2.775  1.5180  1.0194        0.385   0.42949
2@72.6:7@23.9:13@64.8  8       3.947  1.485  0.8077  0.8136        0.554   0.59085
---
        Estimated effects:
        -----------------
                                    est        SE      t
        Intercept                -0.032985  0.036799 -0.896
        2@72.6a                   0.748284  0.055365 13.516
        2@72.6d                  -0.007337  0.073600 -0.100
        7@23.9a                   0.416458  0.050965  8.171
        7@23.9d                  -0.038862  0.073733 -0.527
        13@64.8a                  0.174598  0.050351  3.468
        13@64.8d                  0.177801  0.073600  2.416
```

Figure 3.8: ANOVA table and estimated effects by multiple QTL mapping for the gene expression data at probe "504047". The estimated effect of interaction terms were ignored due to the insignificance.

```
Full model result
----------------------------------
Model formula: y ~ Q1 + Q2 + Q1:Q2

       df       SS         MS      LOD     %var Pvalue(Chi2) Pvalue(F)
Model   8 130.3687 16.2960851 33.01538 26.67647            0         0
Error 481 358.3341  0.7449774
Total 489 488.7028


Drop one QTL at a time ANOVA table:
----------------------------------
             df Type III SS    LOD   %var F value Pvalue(Chi2) Pvalue(F)
2@57.2        6      117.24 30.118 23.990  26.229        0.000   < 2e-16 ***
6@1.8         6       18.99  5.494  3.886   4.248        0.000   0.00035 ***
2@57.2:6@1.8  4        8.41  2.468  1.721   2.822        0.023   0.02459 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3.9: ANOVA table for the gene expression data at probe "10003843955".

4. Conclusion and Discussion

In this study, I used the same dataset as Tian et al. (2015) used in their research. By conducting the data diagnostics and summary, I detected the non-normality of multiple gene expression and transformed the gene expression data for further analysis.

Tian et al. (2015) performed the single-QTL analysis with batch data included as an additive covariate and with sex included as an interactive covariate. Because of the incompleteness of the batch data (several mice don't have the assignment record to the batch), I only considered including sex as an interactive covariate in this study. Despite the fact that the QTL × sex interaction model that I used leads fewer cases of some probe linked to some locus with large LOD score compared with the model that includes batch covariate, it does not influence the detection of trans-eQTL hot-spot. Both studies detected the same trans-eQTL hot-spot on distal chromosome 6. Besides the strongest trans-eQTL hot-spot on distal chromosome 6, I also detected four more inferred trans-eQTL in chromosome 2, 7, 9 and 11 that have LOD score ≥ 100.

Similar to Tian et al. (2015) did, I also performed principal component analysis to convert high-dimensional gene expression data into some composite traits. By using the first principal components, the genotype of trans-eQTL hot-spot on chromosome 6 can be easily and accurately distinguished and inferred. One could consider the first principal components as a composite trait of gene expression in islet tissue and use it for further study.

Besides the identification of eQTL by single-QTL analysis, I also considered evaluating the effect of those inferred trans-eQTL in other chromosomes. I performed the drop-one-QTL analysis to look at the effect of each term in the model. The result shows that there are loci in different chromosome affect the gene expression in a single probe. In some cases, there is also a significant interaction effect among different loci. It would be helpful to further explore the interaction effect among loci in different chromosome.

References

Broman, K. W. and S. Sen (2009). *A Guide to QTL Mapping with R/qtl*, Volume 46. Springer.

Broman, K. W., H. Wu, Ś. Sen, and G. A. Churchill (2003). R/qtl: Qtl mapping in experimental crosses. *Bioinformatics 19*(7), 889–890.

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*(4), 315.

Jansen, R. C. and J.-P. Nap (2001). Genetical genomics: the added value from segregation. *TRENDS in Genetics 17*(7), 388–391.

Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis*, pp. 129–155. Springer.

Kassambara, A. (2019). *ggcorrplot: Visualization of a Correlation Matrix using 'ggplot2'*. R package version 0.1.3.

Kassambara, A. and F. Mundt (2019). *factoextra: Extract and Visualize the Results of Multivariate Data Analyses*. R package version 1.0.6.999.

Peterson, R. A. and J. E. Cavanaugh (2019). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, 1–16.

Ponsuksili, S., E. Murani, M. Schwerin, K. Schellander, and K. Wimmers (2010). Identification of expression qtl (eqtl) of genes expressed in porcine m. longissimus dorsi and associated with meat quality traits. *BMC genomics 11*(1), 572.

Tian, J., M. P. Keller, A. T. Oler, M. E. Rabaglia, K. L. Schueler, D. S. Stapleton, A. T. Broman, W. Zhao, C. Kendziorski, B. S. Yandell, et al. (2015). Identification of the bile acid transporter slco1a6 as a candidate gene that broadly affects gene expression in mouse pancreatic islets. *Genetics 201*(3), 1253–1262.

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis*. Springer.

Wickham, H., R. François, L. Henry, and K. Müller (2018). *dplyr: A Grammar of Data Manipulation*. R package version 0.7.6.

A. R Code

```
rm(list=ls())
##########Loading the packages
library("qtl")
library("dplyr")
library("ggplot2")
library("ggcorrplot")
library("factoextra")
library("bestNormalize")

##########Read the data from .csv files
geno <- read.csv(file = 'Clean/genotypes_clean.csv')
sex <- read.csv(file = 'Clean/gender.csv')
islet <- read.csv(file = 'Clean/islet_mlratio_clean.csv')
annot <- read.csv(file = 'Clean/microarray_annot.csv')
sum(is.na(annot$pos.cM))
#[1] 30
a_gene_id.new <-annot$a_gene_id[!(is.na(annot$pos.cM))]
a_gene_id.new <- paste0("X",a_gene_id.new)

##########Clean the data
MouseNum.islet <- islet$MouseNum
MouseNum.geno <- sex$MouseNum
##Find the intersection of genotype and phenotype based on MouseNum
MouseNum.intersect <- intersect(MouseNum.islet,MouseNum.geno )
length(MouseNum.intersect)
geno.idx <- MouseNum.geno[! MouseNum.geno %in% MouseNum.intersect ]
islet.idx <-MouseNum.islet[! MouseNum.islet %in% MouseNum.intersect ]
geno.new <- geno[ !(geno$MouseNum %in% geno.idx ), ]
dim(geno.new)
geno.new.sorted <- geno.new[order(geno.new$MouseNum),]
sex.new <- sex[!(sex$MouseNum %in% geno.idx ),]
islet.new <- islet[ !(islet$MouseNum %in% islet.idx ), ]


##put the sex data into phenotype data
islet.sex.new <- merge(islet.new, sex.new,  by.x="MouseNum", by.y="MouseNum",
                       sort = FALSE)
islet.sex.new.sorted <- merge(islet.new, sex.new,
                              by.x="MouseNum", by.y="MouseNum")

##select probes with known cM position on some chromosome
islet.colnames <- colnames(islet.sex.new.sorted)[-c(1,40574,40575)]
length(islet.colnames[islet.colnames %in% a_gene_id.new ])
#[1] 37796
islet.colnames.new <- islet.colnames[islet.colnames %in% a_gene_id.new ]
islet.sex.new.sorted.sub <-subset(islet.sex.new.sorted,
                                  select = c("MouseNum",islet.colnames.new,
                                             "Sex" ,"pgm"))

##save the cleaned data into .csv file
write.csv(geno.new.sorted,'islet/geno_new_sorted.csv', row.names=FALSE)
write.csv(islet.sex.new.sorted.sub,'islet/pheno_new_sorted_sub.csv',
          row.names=FALSE)

##########single-QTL analysis
##read data for a QTL experiment
```

```
dat <- read.cross("csvs", dir="islet", genotypes=c("RR","BR","BB"),
                  alleles=c("R","B"), genfile="geno_new_sorted.csv",
                  phefile="pheno_new_sorted_sub.csv")

##summary of cross data
summary(dat)
plotMissing(dat, chr= c(1:19,"X"))
plotMap(dat, chr= c(1:19,"X"))

pheno_col <- 2
plotPheno(dat, pheno.col=pheno_col, xlab =paste("phe", pheno_col-1))

##normalization transformation
orderNorm_obj <- apply(dat$pheno[,-c(1,37798,37799)],2, orderNorm)
for( i in 2:37797){
    dat$pheno[,i] <-  orderNorm_obj[[i-1]]$x.t
}

##check the distribution of phenotype after transformation
pheno_col <- 2
plotPheno(dat, pheno.col=pheno_col, xlab =paste("phe", pheno_col-1))

##estimated recombination fractions between markers
dat <- est.rf(dat)
plotRF(dat,chr= c(1:19,"X"), col.scheme=c("redblue"))

##########Single QTL mapping no covariate
N <- 37796
lod.res <- vector("list", N);
pos.res <- vector("list",N);
flag <- rep(0,N);
pheno.idx <- which(flag < 1)
pheno.idx <- pheno.idx +1
for(pheno_col in pheno.idx ){
    print(pheno_col-1)
    out.hk <- scanone(dat, chr = c(1:19,"X") ,pheno.col=pheno_col, method="hk")
    mylist <- split(out.hk, out.hk$chr)
    lod.group.max <- c()
    pos.group.max <- c()
    for(i in c(1: length(unique(out.hk$chr))) ){
        lod.max.temp <- max(mylist[[i]]$lod)
        lod.group.max <- c(lod.group.max, lod.max.temp)
        pos.group.max <- c(pos.group.max, mylist[[i]][which.max(mylist[[i]]$lod),2]
    }
    lod.res[[pheno_col-1]] <- lod.group.max
    pos.res[[pheno_col-1]] <- pos.group.max
    flag[pheno_col-1] <- 1
    #plot(out.hk,col = c("red", " blue"), main = paste("phe",pheno_col-1 ) ,ylab="LO
}

##########Single QTL mapping with sex covariate
N <- 37796
lod.cov.res <- vector("list", N);
pos.cov.res <- vector("list",N);
flag <- rep(0,N);
pheno.idx <- which(flag < 1)
pheno.idx <- pheno.idx +1
sex <- as.numeric(pull.pheno(dat,"Sex") == "Male")
```

```
for(pheno_col in pheno.idx ){
    print(pheno_col-1)
    out.covar <- scanone(dat, chr = c(1:19,"X") ,pheno.col=pheno_col,
                         method="hk", addcovar=sex, intcovar=sex)
    mylist <- split(out.covar, out.covar$chr)
    lod.group.max <- c()
    pos.group.max <- c()
    for(i in c(1: length(unique(out.covar$chr))) ){
        lod.max.temp <- max(mylist[[i]]$lod)
        lod.group.max <- c(lod.group.max, lod.max.temp)
        pos.group.max <- c(pos.group.max, mylist[[i]][which.max(mylist[[i]]$lod),2]
    }
    lod.cov.res[[pheno_col-1]] <- lod.group.max
    pos.cov.res[[pheno_col-1]] <- pos.group.max
    flag[pheno_col-1] <- 1
}

##find the peak marker
data.full <- data.frame(vars.temp = c(),
                        pos.temp = c(),
                        probe.pos.temp =c(),
                        lod.temp = c(),
                        colvars.temp = c())
for(c in levels(annot.sub.sorted$chr)){
    annot.temp <-subset(annot.sub.sorted,annot.sub.sorted$chr==c )
    print(max(annot.temp$pos.cM))
    print(min(annot.temp$pos.cM))
    #pheno.idx <- as.numeric(rownames(annot.temp))
    pheno.names <- colnames(dat$pheno)[-c(1,37798,37799)]
    pheno.idx <- match(annot.temp$a_gene_id, pheno.names)
    pheno.idx <- pheno.idx +1
    lod.temp <- c()
    pos.temp <- c()
    for(pheno_col in pheno.idx ){
        lod.temp <- c(lod.temp, lod.cov.res[[pheno_col-1]])
        pos.temp <- c(pos.temp, pos.cov.res[[pheno_col-1]]   )
    }
    chr.temp <- c(1:19,"X")
    chr.temp <- rep(chr.temp,times=length(pheno.idx))
    vars.temp <- factor(chr.temp, levels = c(1:19,"X"))
    data.temp = data.frame(vars.temp = vars.temp,
                           pos.temp = pos.temp,
                           probe.pos.temp =rep(annot.temp$pos.cM,each=20),
                           lod.temp = lod.temp,
                           colvars.temp = rep(c,length(lod.temp)))
    data.full <- rbind(data.full,data.temp)
}
data.full$colvars.temp_f <- factor(data.full$colvars.temp, levels = c("X",19:1))

threshold.val <- 10
normalized.lod.temp  <-(lod.temp-min(lod.temp) )/(max(lod.temp)-min(lod.temp))
p1 <- ggplot(subset(data.full,data.full$lod.temp >= threshold.val),
             aes(pos.temp, probe.pos.temp, color=lod.temp)) +
    geom_point(size = 0.5, alpha =normalized.lod.temp[data.full$lod.temp >= thresho
)
p1 <- p1 + theme_classic() +theme( axis.text.x=element_blank(),
    axis.text.y=element_blank(), strip.background = element_blank(),
    panel.margin.y = unit(0, "lines"), panel.margin.x = unit(0,"lines")) +
```

```
    labs(x = "QTL position (cM)", y = "Probe position (cM)", colour = "LOD scores")
p1 <- p1 + scale_color_gradient(low ="grey" , high = "black") +
        coord_fixed(ratio = 1)
p1 <- p1 + facet_grid( cols = vars(vars.temp) , rows = vars(colvars.temp_f))
p1

## Summary of inferred LOD
# subset data that LOD score >= 10, 25, 50, 100
data.sub.10 <- subset(data.full, lod.temp>= 10)
dim(data.sub.10)
#[1] 5867    6
data.sub.25 <- subset(data.full, lod.temp>= 25)
dim(data.sub.25)
#[1] 2484    6
data.sub.50 <- subset(data.full, lod.temp>= 50)
dim(data.sub.50)
#[1] 1157    6
data.sub.100 <- subset(data.full, lod.temp>= 100)
dim(data.sub.100)
#[1] 70  6

## identify the strongest e-QTL in chr 6
#Find marker closest to a specified position
mar6 <- find.marker(dat, 6, 73.806)
phe6 <- subset(annot.sub.sorted, (chr == 6)& (pos.cM <91.862464) &
                    (pos.cM > 91.862462))
phe.col <- (1:length(colnames(dat$pheno)) )[colnames(dat$pheno) ==  phe6$a_gene_id]
plotPXG(dat, mar6,pheno.col=phe.col, ylab=sub("X","",phe6$a_gene_id))
out.covar <- scanone(dat, chr = c(1:19,"X") ,pheno.col=phe.col, method="hk",
                    addcovar=sex, intcovar=sex)
plot(out.covar, col= "blue",main = sub("X","",phe6$a_gene_id) ,ylab="LOD score")

########## Principal component analysis
##data manipulation
#find the probes that mapped to some locus with large LOD
#by removing duplicated rows based on colvars.temp_f and probe.pos.temp
data.sub.10.dinstinct <- data.sub.10 %>% distinct(probe.pos.temp, colvars.temp_f, .k
dim(data.sub.10.dinstinct)
#[1] 5674    6
data.sub.25.dinstinct <- data.sub.25 %>% distinct(probe.pos.temp, colvars.temp_f, .k
dim(data.sub.25.dinstinct)
#[1] 2469    6
data.sub.50.dinstinct <- data.sub.50 %>% distinct(probe.pos.temp, colvars.temp_f, .k
dim(data.sub.50.dinstinct)
#[1] 1147    6
data.sub.100.dinstinct <- data.sub.100 %>% distinct(probe.pos.temp, colvars.temp_f,
dim(data.sub.100.dinstinct)
#[1] 70  6

dim(subset(data.sub.10.dinstinct, colvars.temp != vars.temp))
#[1] 2699    6
dim(subset(data.sub.25.dinstinct, colvars.temp != vars.temp))
#[1] 879   6
dim(subset(data.sub.50.dinstinct, colvars.temp != vars.temp))
#[1] 404   6
dim(subset(data.sub.100.dinstinct, colvars.temp != vars.temp))
#[1] 23  6
data.sub.100.dinstinct.trans <-subset(data.sub.100.dinstinct,
```

```
                    colvars.temp != vars.temp)

##find pheno gene id
data.sub.dinstinct.trans <- data.sub.100.dinstinct.trans %>% distinct(vars.temp,
                .keep_all = TRUE)
pheno.idx.trans <- c()
for( i in 1: dim(data.sub.dinstinct.trans)[1]){
    pheno.idx.trans <- c(pheno.idx.trans, subset(annot.sub.sorted,
                        pos.cM == data.sub.dinstinct.trans$probe.pos.temp[i] &
                    chr == data.sub.dinstinct.trans$colvars.temp[i])$a_gene_id )
}

pheno.trans <- subset(dat$pheno, select =pheno.idx.trans )
dim(pheno.trans)
sub.pca <-  prcomp(pheno.trans, center = TRUE,scale. = TRUE)
summary(sub.pca)
plot(sub.pca)
fviz_eig(sub.pca)
fviz_pca_var(sub.pca,
             col.var = "contrib", # Color by contributions to the PC
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE     # Avoid text overlapping
)

ggcorrplot(cor(pheno.trans),hc.order =TRUE)


########## evaluate the effect of multiple loci
sub2 <- subset(data.sub.25,  colvars.temp!=vars.temp & vars.temp ==2 )
sub9 <- subset(data.sub.25,  colvars.temp!=vars.temp & vars.temp ==9 )
for( i in 1: dim(sub2)[1]){
    chr.temp <- sub2$colvars.temp[i]
    #print(chr.temp)
    if(length(sub9$probe.pos.temp[which(sub9$colvars.temp == chr.temp)])>0){
        if(any(abs(sub2$probe.pos.temp[i]-
                sub9$probe.pos.temp[which(sub9$colvars.temp==chr.temp)])<0.5)){
            print(sub2$probe.pos.temp[i])
        }
    }
}

phe.29 <- subset(annot.sub.sorted,(chr == 5)& (pos.cM <=  53.14375)&
                                (pos.cM >= 53.1436))
phe.col <- (1:length(colnames(dat$pheno)) )[colnames(dat$pheno)== phe.29$a_gene_id]
effectplot(dat, pheno.col=phe.col, mname1=mar6, mname2="Sex",
                        mark2=dat$pheno$Sex, geno2=c("F","M"),add.legend = FALSE)
legend(list(x = 1,y = 0.463), legend=c("RR", "RB", "BB"),
            col=c("black","red", "blue"), lwd= c(1,1,1))
plotPXG(dat, mar6,pheno.col=phe.col, ylab=sub("X","",phe.29$a_gene_id))

out.covar <- scanone(dat, chr = c(1:19) ,pheno.col=phe.col, method="hk",
                        addcovar=sex, intcovar=sex)
plot(out.covar, col= "blue",main = sub("X","",phe.29$a_gene_id) ,ylab="LOD␣score")
summary(out.covar)

chr <- c(2,7 ,13)
pos <- c(72.00, 23.90,  64.81)
qtl <- makeqtl(dat, chr, pos)
```

```
my.formula <- y~Q1*Q2*Q3
out.fitqtl <- fitqtl(dat,pheno.col = phe.col, qtl=qtl, formula=my.formula)
summary(out.fitqtl)
out.fitqtl2 <- fitqtl(dat,pheno.col = phe.col,  qtl=qtl,formula=my.formula,
                    dropone=FALSE, get.ests=TRUE)
summary(out.fitqtl2)
```