AN AUGMENTED BAYESIN MODEL TO PREDICT EQTL BY INTERGRATING DATA FROM MULTIPLE TISSUES

by

YONGHUA ZHUANG

M.D., Tongji University, 2001

Ph.D., Sichuan University, 2009

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science

Biostatistics Program

2016

This thesis for the Master of Science degree by

Yonghua Zhuang

has been approved for the

Biostatistics Program

by

Dr. Katerina Kechris, Chair and Advisor

Dr. Laura M. Saba,

Dr.Stephanie A. Santorico

Date 8/20/2016

Zhuang, Yonghua (M.S., Biostatistics)

An augmented Bayesin model to predict eQTL by intergrating data from multiple tissues

Thesis directed by Associate Professor Katerina Kechris and Assistant Professor Laura M. Saba.

ABSTRACT

To be added.

The form and content of this abstract are approved. I recommend its publication.

Approved: Katerina Kechris

ACKNOWLEDGEMENTS

I would like to extend my sincerest gratitude to my advisors, Dr. Katerina Kechris and Dr. Laura M. Saba, for their guidance, encouragement, and seemingly limitless patience. Thank you to my committee member, Dr.Stephanie A. Santorico, for sharing her expertise and sage advice. Thank you to Dr. Sam MaWhinney, Dr. Gary K. Grunwald and Dr. Deborah Glueck for their confidence.

And most importantly, thank you to my wife, Dan Wang, for your love, support, and generosity through all of the late nights and long weekends.

TABLE OF CONTENTS

LIST OF TABLES

TABLE

LIST OF FIGURES

FIGURE

CHAPTER I

INTRODUCTION

I.1 What is eQTL?

Genetic variation has recently been the focus of many researchers due to its relevance to differential disease susceptibility among individuals. Understanding the specific biological effect of genomic variants, commonly Single Nucleotide Polymorphisms (SNPs), in cells and tissues provide insight to the biology of the disease and complex phenotypes(Nica and Dermitzakis, 2013). Mediating the connection between genetic variants and disease susceptibility may be the effects of SNPs on the RNA expression levels of different genes. Genome-wide association studies (GWAS) have demonstrated that the majority of genetic variants are found in non-coding regions of the genome and may be involved in gene regulation(Manolio, 2010). The analysis of such variants in the context of gene expression measured in different tissues has established a big field in genetics investigating expression quantitative trait loci (eQTL).

An eQTL is a locus that explains a proportion of the variation in gene expression levels in either inbred populations, e.g., laboratory mice, or outbred populations, e.g., humans (Cookson *et al.*, 2009; Nica and Dermitzakis, 2013). An eQTL analysis can help reveal biological processes and discover the genetic factors associated with certain diseases. Determining if mRNA expression levels are altered by specific genetic variants provides evidence of a mechanical link between genetic variation and downstream biological events, of which the first step is often changes in gene expression. A standard eQTL study examines a direct association between markers of genetic variation (such as SNP) and gene mRNA expression levels typically measured in tens or hundreds of individuals. This association examination can be performed proximally or distally to the physical location of the gene of interest. The eQTLs that map to the approximate location of gene are referred to as cis-eQTLs while those that are far from the location of gene, often on different chromosomes, are referred to as trans-eQTLs (Rockman and Kruglyak, 2006). Figure 1 illustrates the concept of cis- and trans- eQTL and how they work. Although there is no uniform distance standard to define cis-eQTL, conventionally, variants within 1 Mb (megabase) on either side of a gene's

transcription start site (TSS) are considered cis while those variants affecting gene expression at a distance greater than 1 Mb from the TSS or on another chromosome were called trans-eQTL (Blauwendraat *et al.*, 2016; Webster *et al.*, 2009). Several studies suggest that most of the regulatory control takes place locally, in the vicinity of genes (Dixon *et al.*, 2007; Göring *et al.*, 2007; Schadt *et al.*, 2008). Numerous genes were detected to have cis eQTLs while detecting trans eQTLs has been less successful. Of note, some cis eQTLs are detected in many tissue types while the majority of trans-eQTLs are tissue-dependent (Gerrits *et al.*, 2009).

I.2 Current methods for eQTL analysis

The conventional eQTL analysis is to perform individual tests for each transcript-SNP pair using simple linear regression, which uses the number of minor alleles as covariate.

Figure 2 depicts the typical analysis strategy for single SNP-Gene association. The traditional approach entails simply selecting SNPs with the smallest association P values from standard maximum likelihood tests (Chen and Witte, 2007).

I.3 Challenges and limitations of current methods

This conventional method for eQTL study suffers several limitations. The eQTL analysis with linear regression assumes that every SNP is an equally likely causal and works independently on targeted gene, which might not be the case. The conventional eQTL linear regression is performed on each tissue separately and ignores the extensive information known about the SNPs on the other tissue(s), which results in low power and less accuracy due to a limited sample size in the tissue of interest.

To solve these problems, several approaches including Bayesian modeling have been developed.

I.4 Bayesian models

Bayesian prediction is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis when more information becomes available. Bayesian models have recently been introduced for eQTL studies (Scott-Boyer *et al.*, 2012; Veyrieras

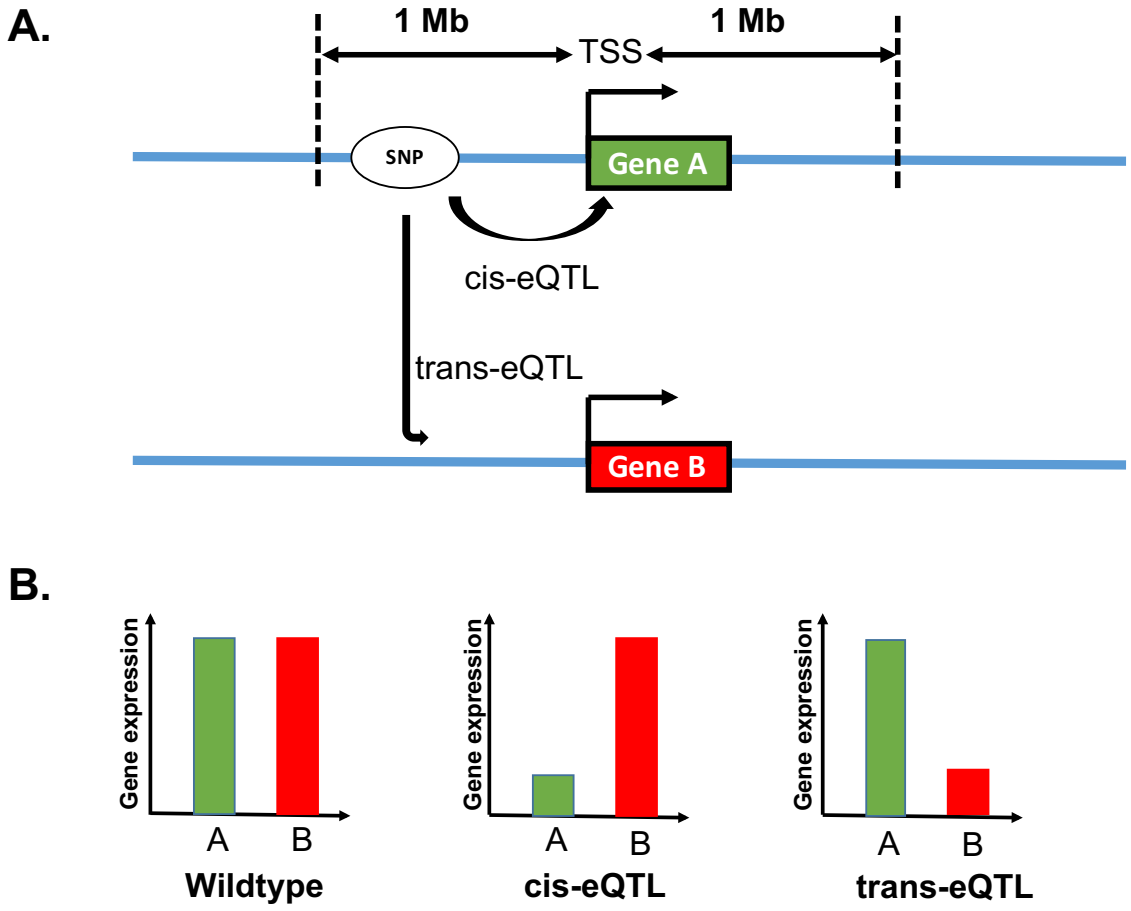

Figure I.1: Illustration of cis and trans expression quantitative trait loci (eQTLs). (**A**),SNP, white circle; genes A, green rectangle (same chromosome); genes B, red rectangle (different chromosome). Each blue line represents different chromosomes. (**B**), in wild-type, gene A (green bar) and gene B (red bar) are highly expressed in wild-type. In cis-eQTL, the gene A expression (green bar) is inhibited due to the SNP on the same chromosome while the transcription level of gene B is not changed. In trans-eQTL, the expression of gene B (red bar) is down-regulated by SNP on the other chromosome.

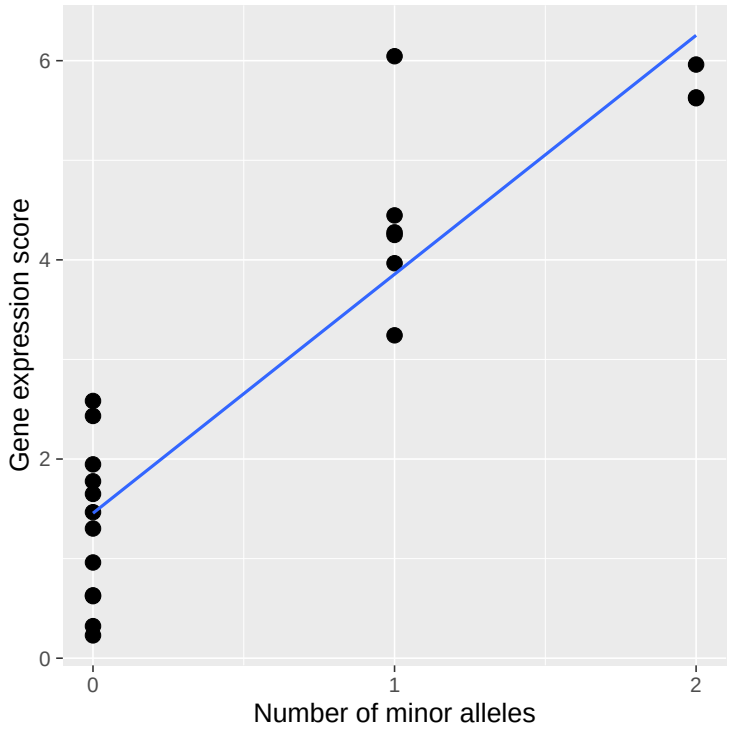

Figure I.2: eQTL analysis with a simple Linear Regression Model.

et al., 2008; Stegle *et al.*, 2010; Stephens and Balding, 2009; Chen and Witte, 2007). Bayesian methods provide a natural modeling framework for eQTL analysis, where information shared across markers and/or genes can increase the power to detect eQTLs (Chen and Witte, 2007; Imholte *et al.*, 2013). Bayesian models are usually based on some modification of a linear model relating expression to SNP genotype(s) (Veyrieras *et al.*, 2008; Stegle *et al.*, 2010; Chen and Witte, 2007). In most cases, uninformative priors are assigned or hyper-parameters for the priors are set to arbitrary values. To date, most eQTL analyses have studied the association of gene and SNP within a single tissue. In only a few studies, the informative priors of eQTL results in a tissue have been used to predict eQTL in other tissues citepLi:2016aa, Flutre:2013aa. A natural step in studying eQTL in a tissue of interest is to incorporate the known results in other tissues. Recently, Dr. Li and his colleagues developed an empirical bayes approach for multiple tissue eQTL analysis (MT-eQTL) (Li *et al.*, 2016). Although MT-eQTL accommodates variation in the number of samples, it was not designed to deal with the unequal number of gene transcripts among multiple tissues.

In terms of model performance evaluation, to our knowledge, current Bayesian models

have been evaluated on the power for detecting associated SNPs either on simulated data or based on the number of discoveries on the real data. Performance assessment on real data is often limited because of an overemphasis on the number of detected SNPs while ignoring potential false positive discovery. The performance of prediction models should be better assessed using other methods and metrics, such as allele specific eQTL (ASE).

I.5 Hypothesis and goals

At the molecular level, comparisons across tissues are often conducted to identify conserved expression changes. For eQTL, we hypothesize that mechanisms for transcriptional control through SNPs may be conserved across tissues and integrating mouse lung eQTL results to inform the prediction of mouse liver eQTL will improve power and accuracy, which can then be linked with genome wide association studies (GWAS) to study biological implications of genomic variants and their impact on disease development.

I.6 Novelty

In this study, we incorporate results of mouse lung eQTL (recombinant inbred mouse panel) to increase power and accuracy of liver eQTL prediction. We develop a novel Bayesian model for eQTL analysis, which takes prior eQTL information into account to better predict eQTL in another tissue.

Moreover, we first evaluate model performance with several methods based on alternative data rather than only utilizing simulated data. (Example of paper of ways to others have been evaluating eQTL methods.)

CHAPTER II

DATA

II.1 Study subjects: BXD inbred mice

Gene expression data and SNP genotypes in BXD inbred mice were downloaded from the Gene Network website (previously named as "WebQTL") (Chesler *et al.*, 2004; Wang *et al.*, 2003). The BXD family of recombinant inbred (RI) strains were derived by crossing C57BL/6J (B6) and DBA/2J (D2) inbred mouse strains and inbreeding progeny for 20 or more generations. The BXD RI strains has been successful used to study the genetics of several behavioral phenotypes including alcohol and drug addiction, stress, and locomotor activity (Tabakoff *et al.*, 2008; Phillips *et al.*, 1995). BXD inbred panel provides a remarkable resource because data for thousands of phenotypes and nearly 100 gene, protein, and metabolite expression data sets have been acquired over a nearly 40-year period. Another advantage of the BXD mice is that the sequencing of both parents have been completed (Source: http://www.genenetwork.org).

II.2 Gene expression data on liver

The liver gene expression data for BXD inbred mice from GEO series GSE16780 were downloaded from Gene Network website. These data were generated by Dr. Jake Lusis and colleagues at UCLA using GeneChip® Mouse Genome 430A Array and are currently listed as a BXD data set, although the study actually includes many other strains (Bennett *et al.*, 2010). The GeneChip® Mouse Genome 430A Array from Affymatrix is a single array representing approximately 14,000 well-characterized mouse genes that can be used to explore biology and disease processes.

RNA was isolated from liver samples from the 99 mouse strains including 30 BXD stains. Double stranded cDNAs were synthesized with 1 μg total RNA through reverse transcription with an oligodT primer using the cDNA Synthesis System. Biotin−labeled cRNA was generated from the cDNA and used to probe Affymetrix Mouse Genome HT−MG430A arrays. Array hybridization, washing and scanning were performed using the manufacturer?s protocol. The scanned image data was processed using the Affymetrix GCOS algorithm

utilizing quantile normalization or the Robust Multiarray method (RMA) to determine the specific hybridizing signal for each gene (Bennett *et al.*, 2010).

II.3 Gene expression data on lung

The lung gene expression data set for 57 strains of mice were generated using the M430 2.0 Affymetrix array and downloaded from Gene Network website. The Affymatrix Mouse Genome 430 2.0 Array offers complete coverage of the Mouse Expression Set 430 for analysis of over 39,000 transcripts on a single array. The data set includes 47 BXD strains and reciprocal F1 hybrids (B6D2F1 and D2B6F1). Data were created by Klaus Schughart, Lu Lu, and Rob Williams. Arrays were processed using RMA protocol by Yan Jiao and Weikuan Gu at the Memphis VA(Alberts *et al.*, 2011).

RNA was isolated from 47 strains of BXD mouse. Double?stranded cDNAs were synthesized with 8 ug total RNA using a standard Eberwine T7 polymerase method. The Affymetrix IVT labeling kit (Affy 900449) was used to generate labeled cRNA. 4-5 μg of each biotinylated cRNA preparation was fragmented and hybridized for 16 hours. After hybridization, GeneChips were washed, stained with SAPE, and read using an Affymetrix GeneChip fluidic station and scanner according to the manufacture protocol (Alberts *et al.*, 2011).

Expression of transcripts in the lung as well as most other Gene Network data sets is measured on a log2 scale. In other words, each unit corresponds approximately to a 2-fold difference in hybridization signal intensity. In order to simplify comparisons among different data sets, log2 RMA values of each array were adjusted to an average expression of 8 units and a standard deviation of 2 units (variance stabilized).

Of note, the gene expression from Gene Network in both liver and lung tissues includes 30 and 47 strains of BXD inbred mice, respectively.

II.4 Genotype data (SNP) on BXD

The smoothed BXD genotype data file were downloaded from Gene Network website (http://www.genenetwork.org/genotypes/BXD.geno) on November,30, 2016. The great majority of SNP genotypes were generated at Illumina. For a limited number of markers and

strains, the genotypes of BXDs seem to be heterozygous. It suggests that these strains were not fully inbred when were initially genotyped. The heterozygous SNPs were excluded from analysis due to its uncertainty.

II.5 allele-specific expression (ASE) in mouse liver

Dr. Lagarrigue and her colleagues have analyzed allele-specific expression (ASE) and parent-of-origin expression in adult mouse liver using next generation sequencing (RNA-Seq) of reciprocal crosses of heterozygous F1 mice from the parental strains C57BL/6J and DBA/2J (Lagarrigue *et al.*, 2013). In this study, they utilized a 10-Mb window on either side of the gene for the classification of local eQTL. An exon was considered to have ASE if P-value ≤ 0.05 and the B/D expression ratio is significantly greater than to 1.5 or less than 1/1.5 . The P value was calculated using a Fisher exact test with the Benjamini-Hochberg method adjustment to control false positive discoveries. Dr. Lagarrigue and her colleagues found, in average in three diet and sex contexts, 397 exons (284 genes) under ASE and shared by two replicates. They reported that a 60% overlap between genes exhibiting ASE and putative cis-eQTL identified in an intercross between the same strains. Among the 284 ASE genes that replicated among samples, 170 (60%) overlap with these 2382 local-eQTL genes published a previous study by Dr. Lagarrigue as well (Davis *et al.*, 2012).

We downloaded all signifiant ASEs from "http://www.genetics.org" website and used them as "standard" to evaluate the performance of newly developed bayesian methods. In other words, only these 287 ASE are considered to have true eQTL while the others do not have significant cis-eQTL.

CHAPTER III

METHODS

In this study, unless otherwise specified,all data manipulation and data analyses were performed using RStudio (version 0.98.1091)(RStudio Team, 2015), R (version 3.2.3)(R Core Team, 2015) using the following packages: "MatrixEQTL"(Shabalin, 2012), "ggplot2"(Wickham, 2009), "fBasics"(Team *et al.*, 2014), "xtable"(Dahl, 2016), "biomaRt"(Durinck *et al.*, 2005), and "flux" (Jurasinski *et al.*, 2014).

III.1 SNP Data Pre-processing

The original SNP data includes 3811 markers on 93 BXD stains mice. These SNPs are located on Chromosomes 1-19 and Chromosome X. The SNPs in BXD inbred mice were originally coded as "B", "D", "H" (heterozygous) and "U" (unknown). They were recoded them to "0", "1", "NA" and "NA", respectively. In other words, heterozygous genotypes and unknown genotypes were set to missing.

The SNP locations were updated to the Ensembl 84: Mus musculus genes (GRCm38.4) version. Among 3811 SNP markers, the chromosome locations were only available on 3025 SNPs in the GRCm38.4 annotation database.

III.2 RNA Expression Data Pre-processing

The gene expression data in mouse liver and lung obtained from Mouse Genome 430A Array and 430 Array were annotated with Ensembl 84: Mus musculus genes (GRCm38.4) to retrieve the transcript corresponding gene Emsembl ID and gene location.

III.3 Basic cis-eQTL analysis

We extended the basic Bayesian linear regression framework (Chen and Witte, 2007; **?**) and developed a model that does not assume uninformative or arbitrary priors. To get informative priors, we analyzed all lung eQTL on a panel of recombinant inbred mice (47 strains).

To get prior information from mouse lung tissue, a model for eQTL analysis is

$$y_{lgi} = \alpha_{lgk} + \beta_{lgk} x_{ki} + \varepsilon_{lgki}, \tag{III.1}$$

- y_{lgi} is the mean expression level of gene g in the strain i and the tissue l ;
- α_{lgk} is the tissue (l), gene (g), and SNP (k) specific intercept;
- β_{lgk} is the tissue (l), gene (g), and SNP(k) specific coefficient;
- x_{ki} is the genotype for SNP k and strain i coded as 0 and 1;
- ε_{lgki} is the error term for strain i, gene g, tissue, and SNP k;

where y_{lg} is the gene g expression value for lung tissue in inbred mouse , x_k is the genotype at SNP for mouse, α_{lgk} and β_{lgk} are the intercept and the regression coefficient for the effect of SNP for each gene-SNP, respectively, and ε_{lgk} is the error term assumed with Gaussian $N(0, \sigma^2_{lgk})$. Each SNP is modeled and regressed separately against each gene.

As with the mouse lung eQTL analysis, a similar basic model relating liver gene expression to genotype is

$$y_{vg} = \alpha_{vgk} + \beta_{vgk} x_k + \varepsilon_{vgk}, \tag{III.2}$$

where y_{vg} is the gene g expression value for liver tissue in mouse, x_k is the genotype at SNP for mouse, α_{vgk} and β_{vgk} are the intercept for the background gene expression level and the regression coefficient for the effect of SNP, respectively, for each gene-SNP pair of interest, and is the error term assumed with Gaussian $N(0, \sigma^2_{vgk})$. Each SNP is modeled and regressed separately against each gene. In inbred mouse, the environmental and genetic parameters were tightly controlled. Thus, no additional covariates were adjusted.

For simplicity, we only select the gene-SNP pair with minimum P value at each gene level for bayesian prediction. In other words, each gene have one and only a eQTL for further analysis. The SNP in selected eQTL for each gene might not be the same between liver and lung tissues.

Prior to developing Bayesian models, we examined whether the shared eQTLs between mouse lung and mouse liver are significant at different thresholds of P value using Chisq

test. A p value < 0.05 is considered significant.

The parameter of interest, regression coefficient for mouse liver, can be first estimated using the basic model (no prior) with Matrix eQTL package(Shabalin, 2012). In this study, we assumed that β is not directional since the direction of the effect in mouse lung is not relevant to mouse liver because we do not expect the exact same genetic variants in different tissues (will discuss more with mentors). Thus, we took the absolute values of β_{lgk} and β_{vgk} for further analyses. To further inform the estimation of $|\beta|_{vgk}$ for mouse liver genes using additional prior information, we assign a Normal prior distribution for $|\beta|_{vgk}$. We assume, $|\beta|_{vgk} \sim \mathcal{N}(z\gamma, \tau^2)$. In other words,

$$\left|\hat{\beta}\right| = z\gamma + U, \quad U \sim \mathcal{N}(0, \tau^2) \tag{III.3}$$

where $\hat{\beta}$ is a vector of the absolute first-stage coefficients (III.2), z is a vector of additional features (described below) for each gene-SNP pair, γ is an unknown vector of parameters corresponding to the additive contribution on the features to the prior mean $z\gamma$, τ^2 is the prior variance term for z.

The prior features we considered for include the significance level and effect of each mouse SNP and gene association (negative logarithm of p-value and absolute value of estimated β_{lgk}). The increase of the statistical significance level of a mouse eQTL lead to more influence of the prior.

The Gaussian conjugate prior assumption leads to a closed form solution to estimate β that simplifies computation. By completing the square, for one SNP the posterior distribution of β, given the data, is Gaussian with posterior mean,

$$\tilde{\beta} = (1 - \lambda) z\hat{\gamma} + \lambda \left|\hat{\beta}\right| \tag{III.4}$$

which is the weighted average of the maximum likelihood estimate (MLE) $\hat{\beta}$ using the basic model (no prior) and the prior mean $z\hat{\gamma}$. The matrix form for a multiple SNP model is given in one of Dr. Chen's papers(Chen and Witte, 2007).

The "shrinkage" term λ is a function of the two variances, σ^2_{vgk} from the basic model (III.2) and τ^2 from the prior in the second stage model. λ indicates how much the MLE is

shrunk towards the prior mean $z\hat{\gamma}$. λ increases to 1 when τ^2 is large (e.g., less informative prior of mouse lung eQTL) and σ^2_{vgk} is small, therefore giving less influence on prior, while λ decreases to 0 when τ^2 is small (more informative prior) and σ^2_{vgk} is large, thereby giving more influence to the prior. Least squares is used in the basic model to obtain estimates $\hat{\beta}$ and σ^2_{vgk}. For estimating $\hat{\gamma}$, τ^2, a two-stage procedure method can be employed(Chen and Witte, 2007; Heron *et al.*, 2011). We assume a common variance and independence across all SNPs and start modeling with an identity matrix and estimating with either the empirical bayes or semi-bayes approach. These estimates are substituted into the shrinkage term and the expression for the posterior mean ($\tilde{\beta} = (1-\lambda)z\hat{\gamma} + \lambda\left|\hat{\beta}\right|$).

III.3.1 Weighted Bayesian model

In standard Bayesian model, we found the majority of estimation ($z\hat{\gamma}$) in the second stage model are much less than their corresponding $\left|\hat{\beta}\right|$, the first estimated using the basic model without prior knowledge. Thus, we introduced a constant (c) weight to Bayesian model to rescale and obtain the final estimate $\tilde{\beta}$.

$$c = \frac{max(\left|\hat{\beta}\right|)}{max(z\hat{\gamma})} \tag{III.5}$$

The weighted Bayesian posterior mean is estimated by,

$$\tilde{\beta} = c(1-\lambda)z\hat{\gamma} + \lambda\left|\hat{\beta}\right| \tag{III.6}$$

III.3.2 Variance of posterior mean and posterior probability below 0

The conjugate prior for liver β was assumed to have normal distribution: $|\beta|_{vgk} \sim \mathcal{N}(z\gamma, \tau^2)$. The posterior mean was distributed as (Kulis, 2012):

$$P(|\beta|_{vgk} \sim \mathcal{N}(\tilde{\beta}, S) \tag{III.7}$$

$$S^{-1} = (\tau^2)^{-1} + (\sigma^2_{vgk})^{-1} \tag{III.8}$$

After calculating posterior mean and its standard deviation, we determined the probability of $\tilde{\beta}$ below 0 using "pnorm" function in R (version 3.2.3).

III.4 Model performance evaluation

Compared with simulation studies in pre-existing methods, we evaluated the developed model with several existing and novel strategies.

III.4.1 Models evaluation based on ASE

To evaluate our model, we compared the results with mouse benchmarks, the ones that are most consistent (or replicated) across mouse panels or cis-eQTL verified in allele specific expression studies. We used the signifiant ASEs identified in one of Dr. Lagarrigue studies as a standard to evaluate the performance of newly developed bayesian methods. Only these 287 ASE are considered to have true eQTL while the other mouse liver genes do not have significant cis-eQTL. We also ranked the lung P value ((III.1) and used them as thresholds to determine "positive" or "negative" cis-eQTL. For example, if a lung P value threshold is 0.001, the gene with "posterior probability below 0" < 0.001 would be considered to have signifiant cis-eQTL while the others do not have. According to this standard, we were able to determine the sensitivity and specificity of Bayesian models, which enables us to derive Receiver operating characteristic (ROC) cures and compare the power and accuracy between Bayesian models and other existing approaches.

III.4.2 Comparison with other methods

We compared the performance of newly developed methods with other existing methods, such as traditional method (linear regression in liver dataset without lung prior information), meta-analytic approach (Stouffer S, 1949; T., 1958), empirical Bayes approach for multiple tissue eQTL analysis (MT-eQTL) which was recently developed by Dr. Li and his colleagues. We also added the linear regression result on lung dataset only to assess liver ASE, which served as a control.

For meta analysis, Stouffer's method is also known as "inverse normal"(Stouffer S, 1949) while Lipták's method is Stouffer's method with weights; this method is commonly referred

to as the weighted Z-test (T., 1958). For MT-eQTL analysis, ...(will add more details)

III.4.3 Model evaluation by subsampling

At the end, to address the effect of sample size in newly developed Bayesian models, we subsampled the liver gene dataset but maintain the prior information from lung eQTL analysis. The original liver gene expression data includes 30 BXD strains and we randomly subset them to 10 strains, 15 stains, 20 strains, and 25 strains. Then we compared area under ROC curves between bayesian models and basic model without prior under different subsettings.

CHAPTER IV

RESULTS

IV.1 Overlap of lung and liver eQTL

First, we examined whether the mechanisms for transcriptional control through SNPs is conserved across tissues in mice and compared the actually shared cis-eQTL number and expected one between liver and lung. The expected number of shared cis-eQTL were calculated under the assumption that there is no correlation in terms of cis-eQTL between two tissues.

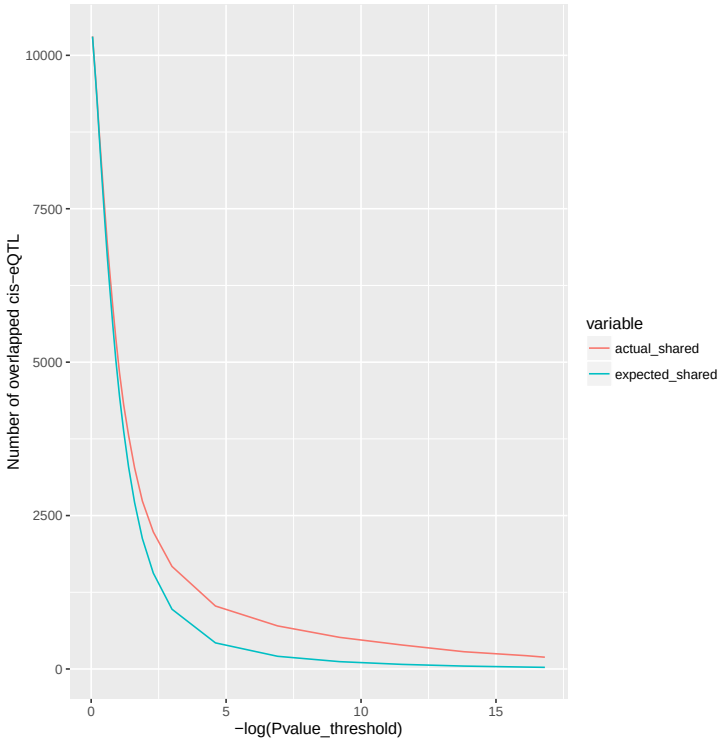

Figure IV.1: Overlap of lung and liver cis eQTL: actual vs expected.

Figure IV.1 shows the actually shared cis eQTL number and expected one between liver and lung at the different thresholds of P value. We found that the actually shared number of cis eQTL between liver and lung is significantly higher than the expected overlap when P value ≤ 0.85 (A.1).

Figure IV.2 clearly indicates that the fold change (fold change=$\frac{Actually\ \ shared\ \ cis\ \ eQTL}{Expected\ \ shared\ \ cis\ \ eQTL}$) is positively associated with negative log P value. The fold change is 1.06 when P value

= 0.4 while the fold change increases to 9.06 as P value goes low to 0.00000005 (A.1). All of above suggest that the mechanisms for gene expression control through SNPs is conserved across tissues and different tissues share cis-eQTL at a significant level. Thus, it would be take advantage of the known cis-eQTL information at one tissue (prior in Bayesian model) to help predict unknown cis-eQTL in another tissue.

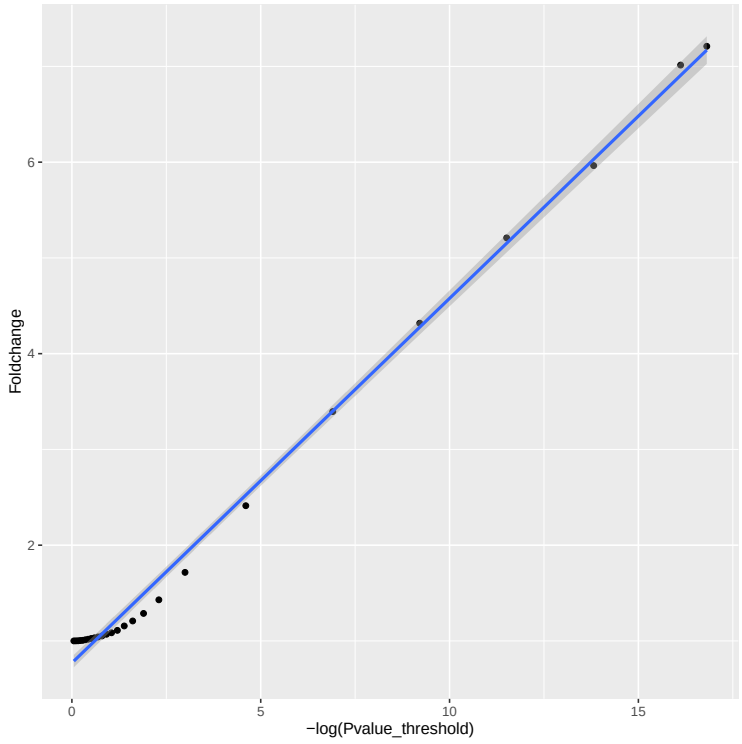

Figure IV.2: Fold change of shared eQTLs between liver and lung filtered by different P value thresholds.

IV.2 Unweighted Bayesian model

Next we integrated lung cis-eQTL results in mice (prior) and developed augmented Bayesian modeling to predict liver cis-eQTL. To get informative priors, we first analyzed all lung cis-eQTL on a panel of recombinant inbred mice using the standard linear regression approach. Figure IV.3 (left panel) depicts the distribution of absolute β value and P value in lung cis-eQTL analysis.

As with the liver cis-eQTL analysis, we also performed liver cis-eQTL analysis without prior using simple linear regression. Figure IV.4 (left panel) showes the histogram of absolute

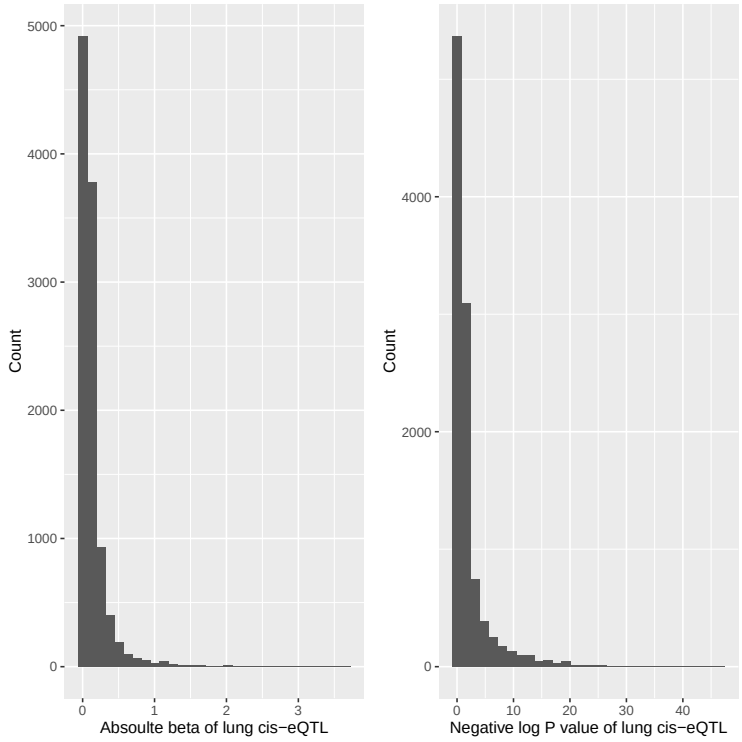

Figure IV.3: Histogram of absolute β values and P values for lung cis-eQTL derived from simple linear regression

β value and P value of liver cis-eQTL.

Then we tried to incorporate lung cis-eQTL known information including β and/or negative log P values into the Bayesian model to enhance liver cis-eQTL prediction. We observed that there is significant correlation between β and negative log P values($\rho = 0.84$). Thus, we only chose one of them as prior. Since β in lung cis-eQTL has similar scale to the β in liver lung cis-eQTl and the parameter of interest in this study is the regression coefficient for mouse liver, we finally selected β in lung cis-eQTL as prior in the following Bayesian model development.

Next we used standard Bayesian model (unweighted) to incorporate lung cis-eQTL information to update liver result. Table IV.1 summarized the statistics of mean, median, minimum and maximum of posterior estimation ($\tilde{\beta}$), original liver prediction ($\hat{\beta}$) and absolute value of coefficient in lung cis-eQTL ($|\beta|_{lgk}$). According to the table IV.1, we found that the maximum of $z\hat{\gamma}$ is 1.17, which is much lower than the maximum (5.18) of $\left|\hat{\beta}\right|$. To adjust for the distribution difference between $z\hat{\gamma}$ and $\left|\hat{\beta}\right|$, we introduced a weight ($c = \frac{max(|\hat{\beta}|)}{max(z\hat{\gamma})}$), to

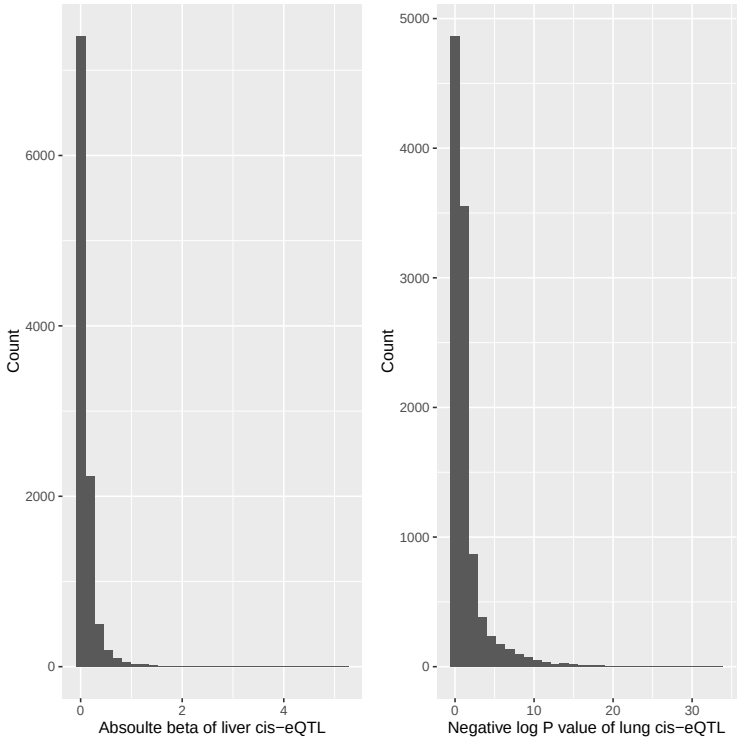

Figure IV.4: Histogram of absolute β values for cis-eQTL derived from simple linear regression

the Bayesian model.

Table IV.1: Summary of beta predictions with unweighted Bayesian model.

	$\tilde{\beta}$	$\left\|\hat{\beta}\right\|$	$z\hat{\gamma}$
Mean	0.10	0.11	0.11
Stdev	0.16	0.21	0.09
Median	0.05	0.05	0.08
Minimum	0.00	0.00	0.06
Maximum	3.92	5.18	1.17

IV.3 Weighted Bayesian model

Table IV.2 summarized the statistics of mean, median, minimum and maximum of posterior estimation ($\tilde{\beta}$), original liver prediction ($\left|\hat{\beta}\right|$) and $z\hat{\gamma}$. According to the table IV.2, we found that the mean ($\pm$sd) and maximum of ($\tilde{\beta}$) are 0.13($\pm$0.20) and 4.51, respectively.

Next the variance of posterior betas were calculated based on σ^2_{vgk}) and τ^2. To rank the liver cis-eQTL predicted by the weighted Bayesian model, the probability of posterior beta ($\tilde{\beta}$) less than 0 was determined based on the value of $\tilde{\beta}$ and its variance. Table IV.3 sum-

Table IV.2: Summary of beta predictions with weighted Bayesian model.

	$\tilde{\beta}$	$\hat{\beta}$	$z\hat{\gamma}$
Mean	0.13	0.11	0.11
Stdev	0.20	0.21	0.09
Median	0.07	0.05	0.08
Minimum	0.00	0.00	0.06
Maximum	4.51	5.18	1.17

marizes the standard deviation and probability of posterior beta below than 0 in weighted Bayesian model

Table IV.3: Summary of variance and probability of posterior beta below than 0 in weighted Bayesian model.

	$\tilde{\beta}$	$\sigma_{\tilde{\beta}}$	p (probability of $\tilde{\beta} < 0$)
Mean	0.13	0.05	0.09
Stdev	0.20	0.03	0.10
Median	0.07	0.04	0.05
Minimum	0.00	0.01	0.00
Maximum	4.51	0.19	0.44

IV.4 Model performance assessment

To assess the performance of the developed Bayesian model, we first evaluate it based on liver allele specific eQTL. Then we compared it with several existing methods in terms of sensibility and specificity according to the liver ASE. To test whether

IV.4.1 Evaluation based on allele specific eQTL (ASE)

To evaluate our model, we compared the results with a mouse benchmark, liver allele specific eQTL (ASE). We used the signifiant ASEs identified by Dr. Lagarrigue as a "gold standard" to evaluate the performance of newly developed bayesian method and other existing approaches. Of note, only these 287 ASE are considered to have true cis eQTL while the other mouse liver genes do not have significant ones. Figure IV.5 depicts negative log P values in both liver ASE group and non-ASE group in liver cis-eQTL and lung cis-eQTL analysis. In figure IV.5 , we can see that the mean of liver negative log P values is much bigger in ASE group than the ones in non-ASE group. This phenom also shows in lung cis-eQTL analysis. Figure IV.5 indicattes that ASE group has lower P value than non-ASE group in both liver and lung cis-eQTL analysis, which further suggests that the association

between SNP and genes are conserved among tissues, at least in liver and lung.

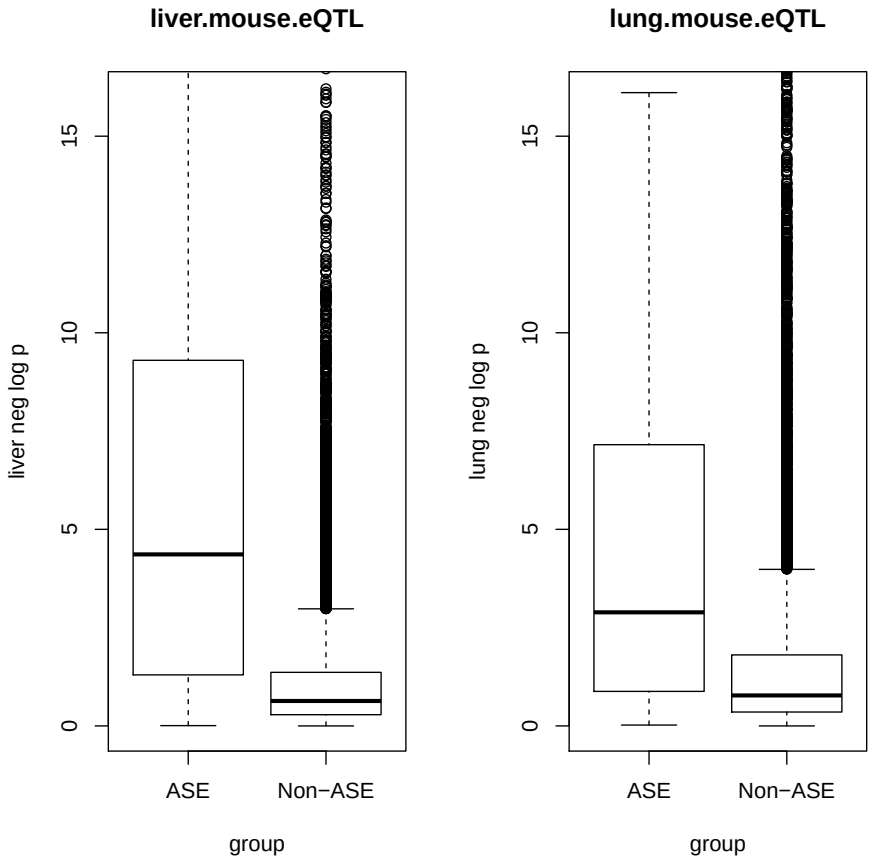

Figure IV.5: Negative log liver/lung P value distribution between ASE and Non-ASE groups.

IV.4.2 Comparison with other methods

Next we We also took the lung P values obtained from model III.1 as thresholds to determine "positive" or "negative" cis-eQTL. Then We were able to determine the sensitivity and specificity of models based on "ASE gold standard", which enables us to derive Receiver operating characteristic (ROC) cures and compare the power and accuracy between Bayesian models and other existing approaches. ROC curves in figure IV.6 show that in terms of sensitivity and specificity, the weighted Bayesian method we newly developed has better performance in predicting liver cis-eQTL when compared to the other four approaches including traditional liver eQTL analysis method with linear regression (marked as "Original"), multiple tissue Bayesian method, meta approach, and traditional lung eQTL analysis

method. As shown in table IV.4 , the area under ROC curves for original liver eQTL analysis method, newly developed Bayesian model, multiple tissue Bayesian method, meta-analysis method, and lung eQTL analysis strategy are 0.81, 0.84, 0.83, 0.81 and 0.71, respectively.

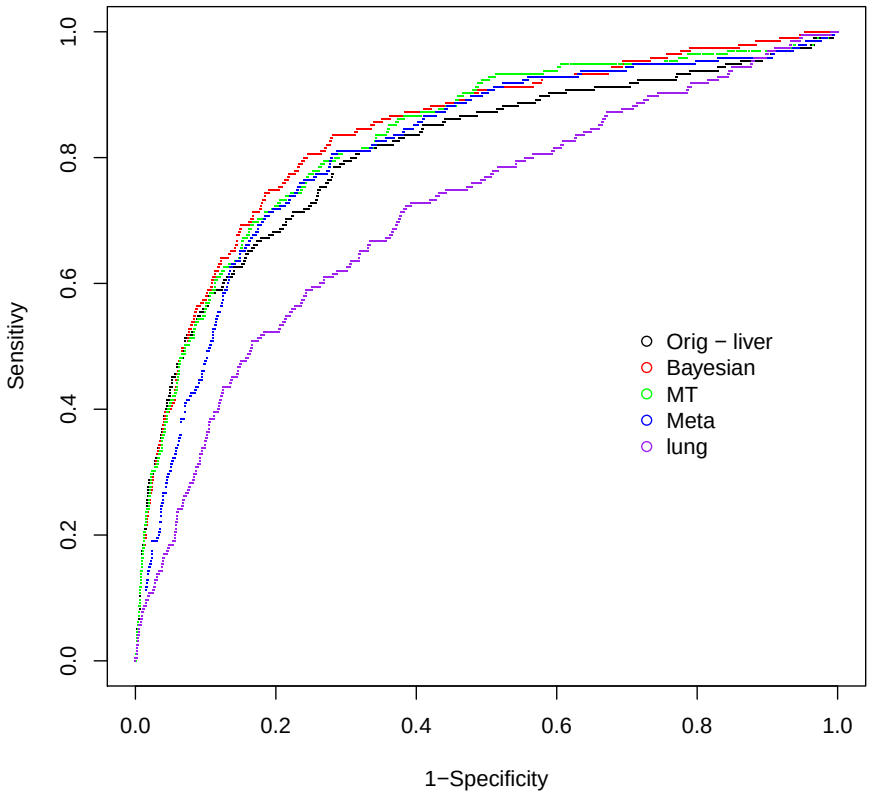

Figure IV.6: ROC curves among five predicting methods.

Table IV.4: AUC comparison among five predicting methods.

	AUC	Fold change
Original liver	0.81	1.00
Bayesian	0.84	1.04
MT	0.83	1.03
Meta	0.81	1.00
Original lung	0.71	0.88

Fold change $= \frac{AUC}{AUC_{original\ liver}}$.

IV.4.3 Evaluation based on subsampling

One of major goal in develop augmented Bayesian model is to improve the power and

accuracy for cis-eQTL prediction when sample size is small. To address the effect of sample size in newly developed Bayesian models, we subsampled the liver gene dataset but maintain the prior information from lung eQTL analysis. We compared area under ROC curves between weighted bayesian models we developed and other 4 approaches under different subsettings (33.33%, 50%, 66.67%, 83.33%).

Table IV.5 summarizes the AUCs of each prediction method under different subsettings. Figure IV.7 depicts the difference among ROC cures in weighted Bayesian, multiple tissue Bayesian and standard liver cis-eQTL(original) prediction methods. As shown in table IV.5, the AUC went down when the number of strains in liver gene expression got low. For example, if only including liver gene data from five strain BXD mice, the AUC of basic model for liver cis-eQTL analysis is 0.72 while it increased to 0.81 when we analyzed with full liver dataset (30 strains). According to table IV.5 and figure IV.7, the AUC in the Bayesian model we developed is always higher than the other 4 methods. In table IV.5, we also normalized AUC with the one from standard method (Original-liver) and calculated fold change for comparison. When the liver gene expression data got less, the fold change of weighted Bayesian, multiple tissue Bayesian and Meta increased, which suggests that the known cis-eQTL lung information is useful to improve liver cis-eQTL prediction.

Table IV.5: AUC comparison among subsetted dataset.

	Subsample (33.33%)		Subsample (50%)		Subsample (67.67%)		Subsample (83.33%)		Full sample	
	AUC	FC	AUC	FC	AUC	FC	AUC	FC	AUC	FC
Original liver	0.72	1.00	0.77	1.00	0.78	1.00	0.80	1.00	0.81	1.00
Bayesian	0.80	1.11	0.83	1.08	0.83	1.07	0.84	1.05	0.84	1.04
MT	0.78	1.09	0.81	1.06	0.78	1.00	0.83	1.03	0.83	1.03
Meta	0.75	1.05	0.79	1.03	0.80	1.03	0.81	1.01	0.81	1.00
Original lung	0.71	0.99	0.71	0.93	0.71	0.91	0.71	0.89	0.71	0.88

FC, fold change=$\frac{AUC}{AUC_{original\ liver}}$.

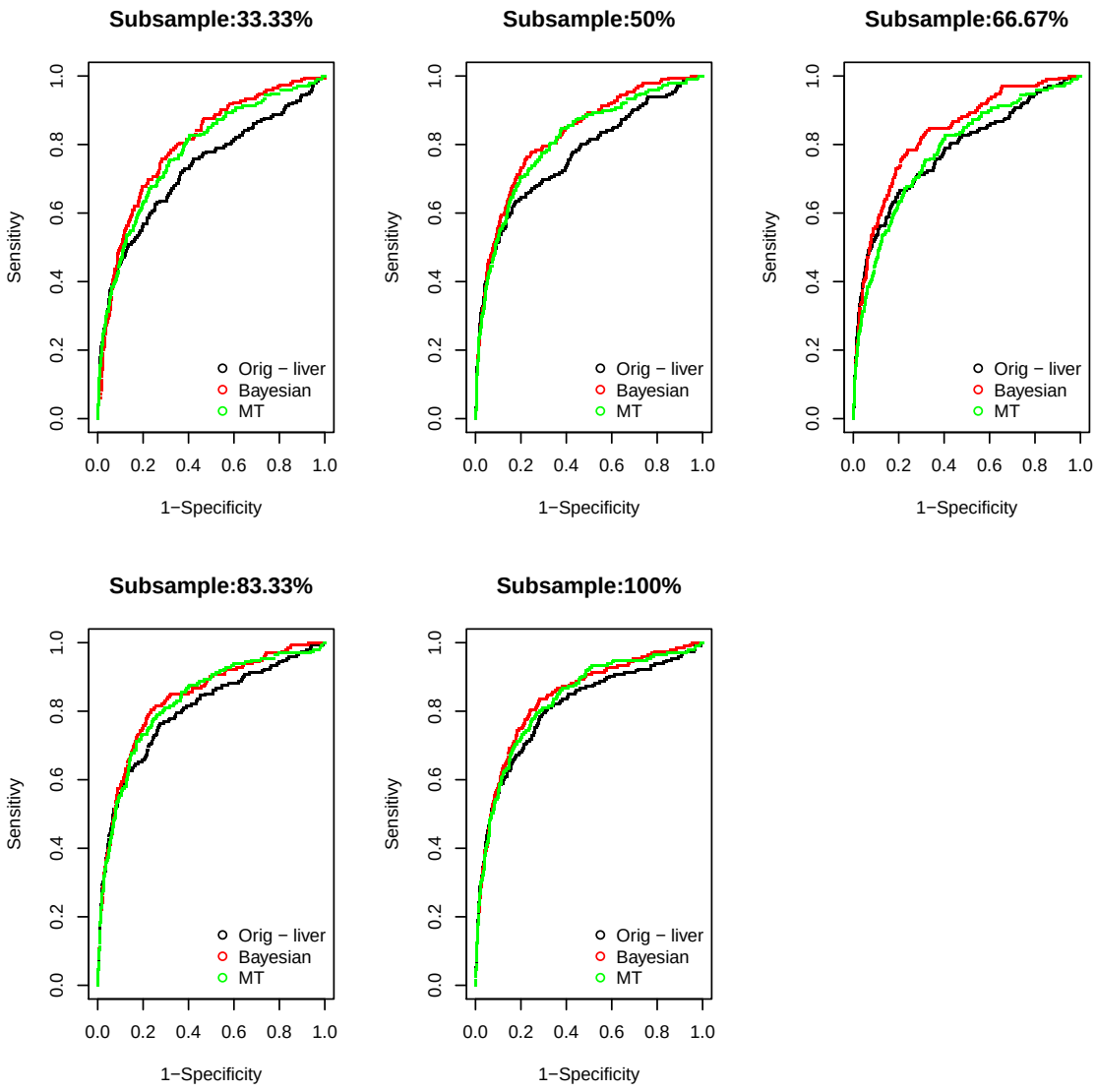

Figure IV.7: ROC curve with different subsampling settings.

CHAPTER V

DISCUSSION

V.1 Statistical discussion

V.2 Advantages and limitations

V.3 Future

REFERENCES

RStudio Team (2015). *RStudio: Integrated Development Environment for R*. RStudio, Inc., Boston, MA.

Alberts R, Lu L, Williams RW, Schughart K (2011). "Genome-wide analysis of the mouse lung transcriptome reveals novel molecular gene interaction networks and cell-specific expression signatures." *Respir Res*, **12**, 61. doi:10.1186/1465-9921-12-61.

Bennett BJ, Farber CR, Orozco L, Kang HM, Ghazalpour A, Siemers N, Neubauer M, Neuhaus I, Yordanova R, Guan B, Truong A, Yang Wp, He A, Kayne P, Gargalovic P, Kirchgessner T, Pan C, Castellani LW, Kostem E, Furlotte N, Drake TA, Eskin E, Lusis AJ (2010). "A high-resolution association mapping panel for the dissection of complex traits in mice." *Genome Res*, **20**(2), 281–90. doi:10.1101/gr.099234.109.

Blauwendraat C, Francescatto M, Gibbs JR, Jansen IE, Simón-Sánchez J, Hernandez DG, Dillman AA, Singleton AB, Cookson MR, Rizzu P, Heutink P (2016). "Comprehensive promoter level expression quantitative trait loci analysis of the human frontal lobe." *Genome Med*, **8**(1), 65. doi:10.1186/s13073-016-0320-1.

Chen GK, Witte JS (2007). "Enriching the analysis of genomewide association studies with hierarchical modeling." *Am J Hum Genet*, **81**(2), 397–404. doi:10.1086/519794.

Chesler EJ, Lu L, Wang J, Williams RW, Manly KF (2004). "WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior." *Nat Neurosci*, **7**(5), 485–6. doi:10.1038/nn0504-485.

Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009). "Mapping complex disease traits with global gene expression." *Nat Rev Genet*, **10**(3), 184–94. doi:10.1038/nrg2537.

Dahl DB (2016). *xtable: Export Tables to LaTeX or HTML*. R package version 1.8-2.

Davis RC, van Nas A, Castellani LW, Zhao Y, Zhou Z, Wen P, Yu S, Qi H, Rosales M, Schadt EE, Broman KW, Péterfy M, Lusis AJ (2012). "Systems genetics of susceptibility to obesity-induced diabetes in mice." *Physiol Genomics*, **44**(1), 1–13. doi:10.1152/physiolgenomics.00003.2011.

Dixon AL, Liang L, Moffatt MF, Chen W, Heath S, Wong KCC, Taylor J, Burnett E, Gut I, Farrall M, Lathrop GM, Abecasis GR, Cookson WOC (2007). "A genome-wide association study of global gene expression." *Nat Genet*, **39**(10), 1202–7. doi:10.1038/ng2109.

Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W (2005). "BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis." *Bioinformatics*, **21**(16), 3439–40. doi:10.1093/bioinformatics/bti525.

Gerrits A, Li Y, Tesson BM, Bystrykh LV, Weersing E, Ausema A, Dontje B, Wang X, Breitling R, Jansen RC, de Haan G (2009). "Expression quantitative trait loci are highly sensitive to cellular differentiation state." *PLoS Genet*, **5**(10), e1000692. doi:10.1371/journal.pgen.1000692.

Göring HHH, Curran JE, Johnson MP, Dyer TD, Charlesworth J, Cole SA, Jowett JBM, Abraham LJ, Rainwater DL, Comuzzie AG, Mahaney MC, Almasy L, MacCluer JW, Kissebah AH, Collier GR, Moses EK, Blangero J (2007). "Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes." *Nat Genet*, **39**(10), 1208–16. doi:10.1038/ng2119.

Heron EA, O'Dushlaine C, Segurado R, Gallagher L, Gill M (2011). "Exploration of empirical Bayes hierarchical modeling for the analysis of genome-wide association study data." *Biostatistics*, **12**(3), 445–61. doi:10.1093/biostatistics/kxq072.

Imholte GC, Scott-Boyer MP, Labbe A, Deschepper CF, Gottardo R (2013). "iBMQ: a R/Bioconductor package for integrated Bayesian modeling of eQTL data." *Bioinformatics*, **29**(21), 2797–8. doi:10.1093/bioinformatics/btt485.

Jurasinski G, Koebsch F, Guenther A, Beetz S (2014). *flux: Flux rate calculation from dynamic closed chamber measurements*. R package version 0.3-0.

Kulis B (2012). "Bayesain Linear Regression." *CSE 788.94: Topics in Machine Learning*.

Lagarrigue S, Martin L, Hormozdiari F, Roux PF, Pan C, van Nas A, Demeure O, Cantor R, Ghazalpour A, Eskin E, Lusis AJ (2013). "Analysis of allele-specific expression in mouse liver by RNA-Seq: a comparison with Cis-eQTL identified using genetic linkage." *Genetics*, **195**(3), 1157–66. doi:10.1534/genetics.113.153882.

Li G, Shabalin AA, Rusyn I (2016). "An Empirical Bayes Approach for Multiple Tissue eQTL Analysis." *arXiv:1311.2948 [stat.ME]*.

Manolio TA (2010). "Genomewide association studies and assessment of the risk of disease." *N Engl J Med*, **363**(2), 166–76. doi:10.1056/NEJMra0905980.

Nica AC, Dermitzakis ET (2013). "Expression quantitative trait loci: present and future." *Philos Trans R Soc Lond B Biol Sci*, **368**(1620), 20120362. doi:10.1098/rstb.2012.0362.

Phillips TJ, Huson M, Gwiazdon C, Burkhart-Kasch S, Shen EH (1995). "Effects of acute and repeated ethanol exposures on the locomotor activity of BXD recombinant inbred mice." *Alcohol Clin Exp Res*, **19**(2), 269–78.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rockman MV, Kruglyak L (2006). "Genetics of global gene expression." *Nat Rev Genet*, **7**(11), 862–72. doi:10.1038/nrg1964.

Schadt EE, Molony C, Chudin E, Hao K, Yang X, Lum PY, Kasarskis A, Zhang B, Wang S, Suver C, Zhu J, Millstein J, Sieberts S, Lamb J, GuhaThakurta D, Derry J, Storey JD, Avila-Campillo I, Kruger MJ, Johnson JM, Rohl CA, van Nas A, Mehrabian M, Drake TA, Lusis AJ, Smith RC, Guengerich FP, Strom SC, Schuetz E, Rushmore TH, Ulrich R (2008). "Mapping the genetic architecture of gene expression in human liver." *PLoS Biol*, **6**(5), e107. doi:10.1371/journal.pbio.0060107.

Scott-Boyer MP, Imholte GC, Tayeb A, Labbe A, Deschepper CF, Gottardo R (2012). "An integrated hierarchical Bayesian model for multivariate eQTL mapping." *Stat Appl Genet Mol Biol*, **11**(4). doi:10.1515/1544-6115.1760.

Shabalin AA (2012). "Matrix eQTL: ultra fast eQTL analysis via large matrix operations." *Bioinformatics*, **28**(10), 1353–8. doi:10.1093/bioinformatics/bts163.

Stegle O, Parts L, Durbin R, Winn J (2010). "A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies." *PLoS Comput Biol*, **6**(5), e1000770. doi:10.1371/journal.pcbi.1000770.

Stephens M, Balding DJ (2009). "Bayesian statistical methods for genetic association studies." *Nat Rev Genet*, **10**(10), 681–90. doi:10.1038/nrg2615.

Stouffer S DeVinney L SE (1949). "The American soldier: Adjustment during army life." *Princeton University Press*, **Vol. 1.**

T L (1958). "On the combination of independent tests." *Magyar Tud Akad Mat Kutato Int Közl*, (171-196).

Tabakoff B, Saba L, Kechris K, Hu W, Bhave SV, Finn DA, Grahame NJ, Hoffman PL (2008). "The genomic determinants of alcohol preference in mice." *Mamm Genome*, **19**(5), 352–65. doi:10.1007/s00335-008-9115-z.

Team RC, Wuertz D, Setz T, Chalabi Y (2014). *fBasics: Rmetrics - Markets and Basic Statistics*. R package version 3011.87.

Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, Pritchard JK (2008). "High-resolution mapping of expression-QTLs yields insight into human gene regulation." *PLoS Genet*, **4**(10), e1000214. doi:10.1371/journal.pgen.1000214.

Wang J, Williams RW, Manly KF (2003). "WebQTL: web-based complex trait analysis." *Neuroinformatics*, **1**(4), 299–308. doi:10.1385/NI:1:4:299.

Webster JA, Gibbs JR, Clarke J, Ray M, Zhang W, Holmans P, Rohrer K, Zhao A, Marlowe L, Kaleem M, McCorquodale 3rd DS, Cuello C, Leung D, Bryden L, Nath P, Zismann VL, Joshipura K, Huentelman MJ, Hu-Lince D, Coon KD, Craig DW, Pearson JV, NACC-Neuropathology Group, Heward CB, Reiman EM, Stephan D, Hardy J, Myers AJ (2009). "Genetic control of human brain transcript expression in Alzheimer disease." *Am J Hum Genet*, **84**(4), 445–58. doi:10.1016/j.ajhg.2009.03.011.

Wickham H (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-0-387-98140-6.

APPENDIX A

Supplemental results

Table A.1: Summary of overlap of lung and liver cis eQTL: actual vs expected.

	Pvalue_threshold	Pvalue_chisq.test	actual_shared	expected_shared	Foldchange
1	0.95	1.0000	10252	10252	1.00
2	0.9	0.0312	9881	9873	1.00
3	0.85	0.0009	9517	9500	1.00
4	0.8	0.0000	9109	9076	1.00
5	0.75	0.0000	8661	8600	1.01
6	0.7	0.0000	8207	8110	1.01
7	0.65	0.0000	7758	7623	1.02
8	0.6	0.0000	7251	7082	1.02
9	0.55	0.0000	6761	6550	1.03
10	0.5	0.0000	6255	6024	1.04
11	0.45	0.0000	5696	5432	1.05
12	0.4	0.0000	5175	4861	1.06
13	0.35	0.0000	4646	4271	1.09
14	0.3	0.0000	4102	3701	1.11
15	0.25	0.0000	3550	3106	1.14
16	0.2	0.0000	3035	2526	1.20
17	0.15	0.0000	2475	1919	1.29
18	0.1	0.0000	1927	1360	1.42
19	0.05	0.0000	1371	803	1.71
20	0.01	0.0000	723	282	2.56
21	0.001	0.0000	408	105	3.88
22	1e-04	0.0000	247	48	5.12
23	1e-05	0.0000	143	23	6.24
24	1e-06	0.0000	88	12	7.32
25	1e-07	0.0000	57	7	8.62
26	5e-08	0.0000	48	5	9.06

APPENDIX B

R codes

B.1 Step 1 - make eQTL

```
rm(list = ls())
gc()

#set directory
setwd("/Volumes/Transcend/Thesis_project/Subsetted_liver")
# subset dataset
sebsetn <- 30# full liver dataset has 30 strains

# subset liver gene expression dataset
mouse.liver.expression.eqtl <-read.table(file="2016-05-16 mouse.
    liver.expression.eqtl.txt",  header=T)
set.seed(50)
sub.mouse.liver.expression.eqtl <- mouse.liver.expression.eqtl[,
    c(1, sample(2:dim(mouse.liver.expression.eqtl)[2],sebsetn,
    replace=FALSE))]
write.table(sub.mouse.liver.expression.eqtl,file="sub.mouse.liver
    .expression.eqtl.txt", sep="\t", row.names=FALSE, quote=FALSE)

#subset liver snp expression data
BXD.geno.SNP.eqtl.for.liver <-read.table(file="2016-05-16 BXD.
    geno.SNP.eqtl.for.liver.txt",  header=T)
head(BXD.geno.SNP.eqtl.for.liver)
dim(BXD.geno.SNP.eqtl.for.liver)
set.seed(50)
sub.BXD.geno.SNP.eqtl.for.liver <- BXD.geno.SNP.eqtl.for.liver[,
    c(1, sample(2:dim(BXD.geno.SNP.eqtl.for.liver)[2],sebsetn,
```

```
    replace=FALSE))]
head(sub.BXD.geno.SNP.eqtl.for.liver)
dim(sub.BXD.geno.SNP.eqtl.for.liver)
write.table(sub.BXD.geno.SNP.eqtl.for.liver,file="sub.BXD.geno.
    SNP.eqtl.for.liver.txt", sep="\t", row.names=FALSE, quote=
    FALSE)

##########################
# liver eqtl analysis
base.dir = "/Volumes/Transcend/Thesis_project/Subsetted_liver"
# Linear model to use, modelANOVA, modelLINEAR, or modelLINEAR_
    CROSS
useModel = modelLINEAR; # modelANOVA, modelLINEAR, or modelLINEAR
    _CROSS
# Genotype file name
SNP_file_name = paste(base.dir, "/sub.BXD.geno.SNP.eqtl.for.liver
    .txt", sep="");
snps_location_file_name = paste(base.dir, "/2016-05-16 BXD.geno.
    loc.eqtl.for.liver.txt", sep="");
# Gene expression file name
expression_file_name = paste(base.dir, "/sub.mouse.liver.
    expression.eqtl.txt", sep="");
gene_location_file_name = paste(base.dir, "/2016-05-16 liver.gene
    .loc.txt", sep="");
# Covariates file name
# Set to character() for no covariates
covariates_file_name = character() ;
# Output file name
output_file_name_cis = tempfile();
```

```
output_file_name_tra = tempfile();

# Only associations significant at this level will be saved
pvOutputThreshold_cis = 1;
pvOutputThreshold_tra = 0.000000000000005;
# Error covariance matrix
# Set to numeric() for identity.
errorCovariance = numeric();
# errorCovariance = read.table("Sample_Data/errorCovariance.txt")
    ;
# Distance for local gene-SNP pairs
cisDist = 1e6; ##### 1 MB

## Load genotype data
snps = SlicedData$new();
snps$fileDelimiter = "\t";      # the TAB character
snps$fileOmitCharacters = "NA"; # denote missing values;
snps$fileSkipRows = 1;
snps$fileSkipColumns = 1;
snps$fileSliceSize = 2000;
snps$LoadFile(SNP_file_name);

## Load gene expression data
gene = SlicedData$new();
gene$fileDelimiter = "\t";
gene$fileOmitCharacters = "NA"; # denote missing values;
gene$fileSkipRows = 1;
gene$fileSkipColumns = 1;
gene$fileSliceSize = 2000;
```

```
gene$LoadFile(expression_file_name);

## Load covariates
cvrt = SlicedData$new();
cvrt$fileDelimiter = "\t";      # the TAB character
cvrt$fileOmitCharacters = "NA"; # denote missing values;
cvrt$fileSkipRows = 1;          # one row of column labels
cvrt$fileSkipColumns = 1;       # one column of row labels
if(length(covariates_file_name)>0) {
cvrt$LoadFile(covariates_file_name);
}

## Run the analysis
snpspos = read.table(snps_location_file_name, header = TRUE,
    stringsAsFactors = FALSE);
genepos = read.table(gene_location_file_name, header = TRUE,
    stringsAsFactors = FALSE);
head(genepos)
me = Matrix_eQTL_main(
  snps = snps,
  gene = gene,
  output_file_name = output_file_name_tra,
  pvOutputThreshold = pvOutputThreshold_tra,
  useModel = useModel,
  errorCovariance = numeric(),
  verbose = TRUE,
  output_file_name.cis = output_file_name_cis,
  pvOutputThreshold.cis = pvOutputThreshold_cis,
  snpspos = snpspos,
```

```
  genepos = genepos,
  cisDist = cisDist,
  pvalue.hist = TRUE,
  min.pv.by.genesnp = FALSE,
  noFDRsaveMemory = FALSE);

unlink(output_file_name_cis);
## Results:
cat('Analysis done in:', me$time.in.sec, ' seconds', '\n')
cat('Detected local eQTLs:','\n')
cis.eqtls<-me$cis$eqtls
head(cis.eqtls)
dim(cis.eqtls)
cis.eqtls$beta_se <-cis.eqtls$beta/cis.eqtls$statistic
write.table(cis.eqtls,file="sub.mouseliver.cis.1M.eqtls.txt", sep
    ="\t", row.names=FALSE, quote=FALSE)

#########################
# eqtl analysis for lung
## Settings
# Linear model to use, modelANOVA, modelLINEAR, or modelLINEAR_
    CROSS
useModel = modelLINEAR; # modelANOVA, modelLINEAR, or modelLINEAR
    _CROSS
# Genotype file name
SNP_file_name = paste(base.dir, "/2016-05-16 BXD.geno.SNP.eqtl.
    for.lung.txt", sep="");
snps_location_file_name = paste(base.dir, "/2016-05-16 BXD.geno.
    loc.eqtl.for.lung.txt", sep="");
```

```
# Gene expression file name
expression_file_name = paste(base.dir, "/2016-05-16 mouse.lung.
    expression.eqtl.txt", sep="");
gene_location_file_name = paste(base.dir, "/2016-05-16 lung.gene.
    loc.txt", sep="");
# Covariates file name
# Set to character() for no covariates
covariates_file_name = character() ;

# Output file name
output_file_name_cis = tempfile();
output_file_name_tra = tempfile();

# Only associations significant at this level will be saved
pvOutputThreshold_cis = 1;
pvOutputThreshold_tra = 0.000000000000005;

# Error covariance matrix
# Set to numeric() for identity.
errorCovariance = numeric();
# errorCovariance = read.table("Sample_Data/errorCovariance.txt")
    ;
# Distance for local gene-SNP pairs
cisDist = 1e6;

## Load genotype data
snps = SlicedData$new();
snps$fileDelimiter = "\t";      # the TAB character
snps$fileOmitCharacters = "NA"; # denote missing values;
```

```
146 snps$fileSkipRows = 1;
147 snps$fileSkipColumns = 1;
148 snps$fileSliceSize = 2000;
149 snps$LoadFile(SNP_file_name);
150
151 ## Load gene expression data
152 gene = SlicedData$new();
153 gene$fileDelimiter = "\t";
154 gene$fileOmitCharacters = "NA"; # denote missing values;
155 gene$fileSkipRows = 1;
156 gene$fileSkipColumns = 1;
157 gene$fileSliceSize = 2000;
158 gene$LoadFile(expression_file_name);
159
160 ## Load covariates
161 cvrt = SlicedData$new();
162 cvrt$fileDelimiter = "\t";      # the TAB character
163 cvrt$fileOmitCharacters = "NA"; # denote missing values;
164 cvrt$fileSkipRows = 1;          # one row of column labels
165 cvrt$fileSkipColumns = 1;       # one column of row labels
166 if(length(covariates_file_name)>0) {
167   cvrt$LoadFile(covariates_file_name);
168 }
169
170 ## Run the analysis
171
172 snpspos = read.table(snps_location_file_name, header = TRUE,
      stringsAsFactors = FALSE);
173 genepos = read.table(gene_location_file_name, header = TRUE,
```

```
     stringsAsFactors = FALSE);
head(genepos)

me = Matrix_eQTL_main(
  snps = snps,
  gene = gene,
  output_file_name = output_file_name_tra,
  pvOutputThreshold = pvOutputThreshold_tra,
  useModel = useModel,
  errorCovariance = numeric(),
  verbose = TRUE,
  output_file_name.cis = output_file_name_cis,
  pvOutputThreshold.cis = pvOutputThreshold_cis,
  snpspos = snpspos,
  genepos = genepos,
  cisDist = cisDist,
  pvalue.hist = TRUE,
  min.pv.by.genesnp = FALSE,
  noFDRsaveMemory = FALSE);

unlink(output_file_name_cis);
## Results:
cat('Analysis done in:', me$time.in.sec, ' seconds', '\n')
cat('Detected local eQTLs:','\n')
cis.eqtls<-me$cis$eqtls
head(cis.eqtls)
dim(cis.eqtls)
cis.eqtls$beta_se <-cis.eqtls$beta/cis.eqtls$statistic
write.table(cis.eqtls,file="mouselung.cis.1M.eqtls.txt", sep="\t"
```

```
  , row.names=FALSE, quote=FALSE)
```

B.2 Step 2 - Bayesian

```
####### Bayesian Method

# load mouse lung cis eqtl result
lung.mouse.eQTL<-read.table(file="mouselung.cis.1M.eqtls.txt",
    header=T)
# load mouse liver cis eqtl result
liver.mouse.eQTL<-read.table(file="sub.mouseliver.cis.1M.eqtls.
    txt",  header=T)

mouse4302ensembl_id<-read.table(file="2015-12-04 mouse4302ensembl
    _id.txt",  header=T)
mouse430aensembl_id<-read.table(file="2015-12-07 mouse430aensembl
    _id.txt",  header=T)
# Add ensemble id annoatation to the data
lung.mouse.eQTL<-merge(lung.mouse.eQTL, mouse4302ensembl_id, by.x
     = "gene", by.y="probe_id")
liver.mouse.eQTL<-merge(liver.mouse.eQTL, mouse430aensembl_id, by
    .x = "gene", by.y="probe_id")
head(lung.mouse.eQTL)
head(liver.mouse.eQTL)

library(data.table)
library(plyr)
# Select lung Gene-SNP pair with minimum P value
lung.mouse.eQTL.min <- data.table(lung.mouse.eQTL, key=c('ensembl
    _id', "pvalue"))
lung.mouse.eQTL.min<-lung.mouse.eQTL.min[J(unique(ensembl_id)),
    mult="first"]
```

```
lung.mouse.eQTL.min<-as.data.frame(lung.mouse.eQTL.min)

# Select liver Gene-SNP pair with minimum P value
liver.mouse.eQTL.min <- data.table(liver.mouse.eQTL, key=c('ensembl_id', "pvalue"))
liver.mouse.eQTL.min<-liver.mouse.eQTL.min[J(unique(ensembl_id)), mult="first"]
liver.mouse.eQTL.min<-as.data.frame(liver.mouse.eQTL.min)

lung.mouse.eQTL.min<-rename(lung.mouse.eQTL.min, c("pvalue"="lung_pvalue", "beta"="lung.beta", "beta_se"="lung.beta_se"))
liver.mouse.eQTL.min<-rename(liver.mouse.eQTL.min, c("pvalue"="liver_pvalue", "beta"="liver.beta", "beta_se"="liver.beta_se"))

head(lung.mouse.eQTL.min)
head(liver.mouse.eQTL.min)
tail(liver.mouse.eQTL.min)
dim(lung.mouse.eQTL.min)
dim(liver.mouse.eQTL.min)
# lung, liver eqtl with ensemble_id
merged.mouse.eQTL.min<-merge(lung.mouse.eQTL.min, liver.mouse.eQTL.min, by.x = "ensembl_id", by.y="ensembl_id")
head(merged.mouse.eQTL.min)
dim(merged.mouse.eQTL.min)
merged.mouse.eQTL.min<-data.frame(merged.mouse.eQTL.min)
merged.mouse.eQTL.min<-merged.mouse.eQTL.min[, c(1, 5, 7, 8, 12, 14, 15 )]
head(merged.mouse.eQTL.min)
```

```
write.table(merged.mouse.eQTL.min,file="mouse.liver.expression.
    min.txt", sep="\t", row.names=FALSE, quote=FALSE)

####### START HERE
merged.mouse.eQTL.min<-read.table(file="mouse.liver.expression.
    min.txt",  header=T)

###KK exploratory code
#plot(-log(merged.mouse.eQTL.min$lung_pvalue,10), -log(merged.
    mouse.eQTL.min$liver_pvalue,10))
#lungs = -log(merged.mouse.eQTL.min$lung_pvalue,10)
#livers = -log(merged.mouse.eQTL.min$liver_pvalue,10)
#mean(lungs[livers>10]>5)
#mean(livers[lungs>10]>5)

###KK added - didn't have abs beta variables

merged.mouse.eQTL.min$abs_liver.beta = abs(merged.mouse.eQTL.min$
    liver.beta)
merged.mouse.eQTL.min$abs_lung.beta = abs(merged.mouse.eQTL.min$
    lung.beta)
merged.mouse.eQTL.min$abs_liver.beta = abs(merged.mouse.eQTL.min$
    liver.beta)
merged.mouse.eQTL.min$abs_lung.beta = abs(merged.mouse.eQTL.min$
    lung.beta)
merged.mouse.eQTL.min$neg_log_lung_pvalue = -log10(merged.mouse.
    eQTL.min$lung_pvalue)
merged.mouse.eQTL.min$neg_log_liver_pvalue = -log10(merged.mouse.
    eQTL.min$liver_pvalue)
```

```
63
64 # Simple linear regression between abs_liver.beta and abs_lung.
     beta
65 # fit1<-summary(lm(abs_liver.beta ~ abs_lung.beta, data=merged.
     mouse.eQTL.min))
66 # fit1
67 # tau<-fit1$sigma**2
68 # check association between abs_liver.beta and abs.lung.beta
69
70 #Plots
71 #ggplot(merged.mouse.eQTL.min, aes(x=abs_lung.beta, y=abs_liver.
     beta)) +geom_point()+geom_smooth(method=lm)
72 #cor(merged.mouse.eQTL.min$abs_lung.beta, merged.mouse.eQTL.min$
     abs_liver.beta)
73 #ggplot(merged.mouse.eQTL.min, aes(x=lung.beta, y=liver.beta)) +
     geom_point()+geom_smooth(method=lm)
74 #cor(merged.mouse.eQTL.min$lung.beta, merged.mouse.eQTL.min$liver
     .beta)
75
76 merged.mouse.eQTL<-merged.mouse.eQTL.min
77 # retrieve ensembl_id
78 markers<-merged.mouse.eQTL[, 1]
79 # Yg=Ag + Bg*Xsnp+V
80 # retrieve betas.hat (liver.beta)
81 betas.hat<-merged.mouse.eQTL$abs_liver.beta
82 # retrieve liver.beta_se
83 se<-merged.mouse.eQTL$liver.beta_se
84
85 # create Z matrix with 2 columns: 1 for intercept,abs_lung.beta (
```

```
    merged.mouse.eQTL[,10])
Z<-as.matrix(merged.mouse.eQTL$abs_lung.beta)
Z<-as.matrix(merged.mouse.eQTL$neg_log_lung_pvalue) ##Use p-value
     as Z - didn't make a big difference
Z<-replace(Z,is.na(Z),0)
Z<-data.frame(1,Z)
Z<-as.matrix(Z)
rowLength<-length(markers)

#CHANGE, include both beta and pvalue
#Z1<-as.matrix(merged.mouse.eQTL$abs_lung.beta)
#Z2<-as.matrix(merged.mouse.eQTL$neg_log_lung_pvalue)
#Z1<-replace(Z1, is.na(Z1),0)
#Z2<-replace(Z2, is.na(Z2),0)
#Z<-data.frame(1,Z1,Z2)
#Z<-as.matrix(Z)
#rowLength<-length(markers)

# Regression: abs_liver.beta = intercept + beta*abs_lung.beta +
    error
lmsummary<-summary(lm(abs_liver.beta~-1+Z, data=merged.mouse.eQTL
    ))
lmsummary
model.prior = lm(abs_liver.beta~-1+Z, data=merged.mouse.eQTL)
# error ~ N(0, Tau)
tau<-lmsummary$sigma**2
tau
# output coeffieients (gamma matrix)
# gamma matrix
```

```
gamma<-as.matrix(lmsummary$coefficients[,1])
# transpose Z matrix
Z_transpose<-t(Z)
# create identity matrix
identity<-diag(nrow=rowLength)
# original betas.hat
betas.hat<-as.matrix(betas.hat)

#### WEIGHTS
useweights = 0 ##CHANGE TOGGLE
if(useweights ==1)
{
    val = 1
    weight = exp(-merged.mouse.eQTL.min$neg_log_lung_pvalue + val
       )
}

#create V matrix for liver_residual_variance
V <- matrix(0, rowLength, rowLength)
# V, liver residual variance
diag(V) <- merged.mouse.eQTL$liver.beta_se^2
# Creat Tau matrix
Tau<- diag(tau, rowLength, rowLength)
# follow Chen's paper and caculate s
s <-V + Tau
if(useweights ==1) {s <-V + diag(weight)*Tau}

# create inverse function for inversing diagnoal matrix
diag.inverse <- function(x){diag(1/diag(x), nrow(x), ncol(x))}
```

```
# create multiplication function for multiplicating two diagnoal
    matrix
diag.multi <- function(x,y){diag(diag(x)*diag(y), nrow(x), ncol(x
    ))}
# inverse s
S <-diag.inverse(s)
# follow chen's paper to caculate omega
omega<-diag.multi(S, V)
# retrieve omega value from the matrix
omega.diag<-diag(omega )
# summary the omega value
summary(omega.diag)

#regression beta
regbeta <- (Z %*% gamma)
head(regbeta)
summary(regbeta)
# caculate betas.tieda with the formula in Chen's paper
constant = max(merged.mouse.eQTL.min$abs_liver.beta)/max(regbeta)
     ###CHANGE
betas.tieda <- constant * omega %*% Z %*% gamma + (identity-omega
    ) %*% betas.hat

head(betas.tieda)
head(betas.hat)

markers1<-as.character(markers)
# combine ensemble_id, betas.hat and betas.tieda
outputVector<-c(markers1,betas.hat,betas.tieda)
```

```
write.table(matrix(outputVector,rowLength),file="hm_tau_hmresults
    .txt",col.names=FALSE,row.names=FALSE,quote=FALSE)
liver.mouse.eQTL.bayesian<-read.table(file="hm_tau_hmresults.txt"
    )
colnames(liver.mouse.eQTL.bayesian)<-c( "ensembl_id", "betas.hat"
    ,"betas.tieda")
head(liver.mouse.eQTL.bayesian)
# merge dataset with betas.hat and betas.tieda
liver.mouse.eQTL.bayesian<- merge(liver.mouse.eQTL.bayesian,
    merged.mouse.eQTL.min, by = "ensembl_id")
head(liver.mouse.eQTL.bayesian)

write.table(liver.mouse.eQTL.bayesian,file="liver.mouse.eQTL.
    bayesian.txt")
```

B.3 Step 3 - Posterior estimation

```
liver.mouse.eQTL.bayesian<-read.table(file="liver.mouse.eQTL.
    bayesian.txt")
head(liver.mouse.eQTL.bayesian)


# Caculate variance for beta.tieda by following Brian Kulis'
    lecture notes
# Invert Tau and V
Tau_invert<-diag.inverse(Tau)
V_invert<-diag.inverse(V)
PS_invert<-Tau_invert + V_invert

# S in Brian Kulis' lecture note:PS
PS <- diag.inverse(PS_invert)
# retrieve posterior variance
ps<-diag(PS)
range(ps)

# reshape posterior variance to long format
ps.long <- melt(ps)
head(ps.long)
# Caculate sd: square root on variance
ps.long$betas.tieda.se<-(ps.long$value)^0.5
# combine sd to the data.frame
liver.mouse.eQTL.bayesian<-cbind(liver.mouse.eQTL.bayesian,ps.
    long$betas.tieda.se)

# head(liver.mouse.eQTL.bayesian)
```

```
# rename betas.tieda.se
liver.mouse.eQTL.bayesian<-rename(liver.mouse.eQTL.bayesian, c("
    ps.long$betas.tieda.se"="betas.tieda.se", "liver.beta_se"="
    betas.hat.se"))

#liver.mouse.eQTL.bayesian<-subset(liver.mouse.eQTL.bayesian,
    select = c("ensembl_id", "betas.hat", "betas.hat.se", "betas.
    tieda",
#"betas.tieda.se","liver_pvalue", "abs_lung.beta", "neg_log_liver
    _pvalue", "neg_log_lung_pvalue"))

# caculate probability of betas.tieda below 0 based on betas.
    tieda and standard deviation
liver.mouse.eQTL.bayesian$p.below.0 <- pnorm(0,liver.mouse.eQTL.
    bayesian$betas.tieda, liver.mouse.eQTL.bayesian$betas.tieda.se
    )

head(liver.mouse.eQTL.bayesian)
dim(liver.mouse.eQTL.bayesian)
summary(liver.mouse.eQTL.bayesian$betas.tieda.se)

range(liver.mouse.eQTL.bayesian$p.below.0)
write.table(liver.mouse.eQTL.bayesian,file="liver.mouse.eQTL.
    bayesian with beta.txt")
```

B.4 Step 4 - Allele Specific Expression (ASE)

```
###START HERE
liver.mouse.eQTL.bayesian <- read.table(file="liver.mouse.eQTL.
    bayesian with beta.txt")
liver.mouse.eQTL.bayesian.tau <- liver.mouse.eQTL.bayesian

###ASE
##########
liver.ASE <- read.csv(file= "ASE.genetics.113.153882-6.csv")
dim(liver.ASE)
head(liver.ASE)
# 440 unique gene ID
length(unique(liver.ASE$geneID))

# verify ASE table
liver.ASE1 <- liver.ASE[which(liver.ASE$replicate == "M.CH. DxB
    and BxD"), ]
liver.ASE2 <- liver.ASE[which(liver.ASE$replicate == "M.HF DxB
    and BxD"), ]
liver.ASE3 <- liver.ASE[which(liver.ASE$replicate == "F.HF DxB
    and BxD"), ]
length(unique(liver.ASE1$geneID))
length(unique(liver.ASE2$geneID))
length(unique(liver.ASE3$geneID))
(length(unique(liver.ASE1$geneID))+length(unique(liver.ASE2$
    geneID))+length(unique(liver.ASE3$geneID)))/3
# As claimed in the paper: averaged 284 ASE for each replicate
sub.liver.ASE <-liver.ASE1
summary(sub.liver.ASE$pvalBH.DxB7)
```

```
sub.liver.ASE1 <- subset(sub.liver.ASE, pvalBH.DxB7 <
    0.00000000000001)
sub.liver.ASE2 <- subset(sub.liver.ASE, pvalBH.DxB7 >=
    0.00000000000001 & pvalBH.DxB7 < 0.0000058)
sub.liver.ASE3 <- subset(sub.liver.ASE, pvalBH.DxB7 >= 0.0000058
    & pvalBH.DxB7 < 0.0031000)
sub.liver.ASE4 <- subset(sub.liver.ASE, pvalBH.DxB7 >= 0.0031000
    & pvalBH.DxB7 >= 0.0031000 )
dim(sub.liver.ASE1)
dim(sub.liver.ASE2)
dim(sub.liver.ASE3)
dim(sub.liver.ASE4)

# sub.liver.ASE <- sub.liver.ASE[ sub.liver.ASE$geneID %in%
    names(table(sub.liver.ASE$geneID))[table(sub.liver.ASE$geneID)
     >1] , ]
# check the remain gene number after subsetting
dim(sub.liver.ASE)
liver.ASE.symbol <- unique(sub.liver.ASE$geneID)
liver.ASE.symbol1 <- unique(sub.liver.ASE1$geneID)
liver.ASE.symbol2 <- unique(sub.liver.ASE2$geneID)
liver.ASE.symbol3 <- unique(sub.liver.ASE3$geneID)
liver.ASE.symbol4 <- unique(sub.liver.ASE4$geneID)
length(liver.ASE.symbol)

# Annoate gene symbol with ensemble.ID
library(biomaRt)
mouse = useMart("ensembl", dataset = "mmusculus_gene_ensembl")
liver.ASE.ensembl <- getBM( attributes=c("ensembl_gene_id", "mgi_
```

```
    symbol") , filters=
47                                "mgi_symbol", values =liver.ASE.
                                    symbol, mart=mouse)
48 liver.ASE.ensembl1 <- getBM( attributes=c("ensembl_gene_id", "mgi
    _symbol") , filters=
49                                "mgi_symbol", values =liver.ASE.
                                    symbol1, mart=mouse)
50 liver.ASE.ensembl2 <- getBM( attributes=c("ensembl_gene_id", "mgi
    _symbol") , filters=
51                                "mgi_symbol", values =liver.ASE.
                                    symbol2, mart=mouse)
52 liver.ASE.ensembl3 <- getBM( attributes=c("ensembl_gene_id", "mgi
    _symbol") , filters=
53                                "mgi_symbol", values =liver.ASE.
                                    symbol3, mart=mouse)
54 liver.ASE.ensembl4 <- getBM( attributes=c("ensembl_gene_id", "mgi
    _symbol") , filters=
55                                "mgi_symbol", values =liver.ASE.
                                    symbol4, mart=mouse)
56 dim(liver.ASE.ensembl)
57 liver.ASE.ensembl <- unique(liver.ASE.ensembl)
58 # delete liver ASE ensemble ID which are not in the liver.mouse.
    eQTL.bayesian data frame
59 liver.ASE.ensembl <- liver.ASE.ensembl[liver.ASE.ensembl$ensembl_
    gene_id %in% liver.mouse.eQTL.bayesian.tau$ensembl_id, ]
60 dim(liver.ASE.ensembl)
61
62 liver.mouse.eQTL.bayesian.tau$eqtl[liver.mouse.eQTL.bayesian.tau$
    ensembl_id %in% liver.ASE.ensembl$ensembl_gene_id] <- 1
```

```
liver.mouse.eQTL.bayesian.tau$eqtl[!liver.mouse.eQTL.bayesian.tau
    $ensembl_id %in% liver.ASE.ensembl$ensembl_gene_id] <- 0

write.table(liver.mouse.eQTL.bayesian.tau, "liver.mouse.eQTL.
    bayesian.tau.txt")
```

B.5 Step 5 - ROC plot and AUC analysis

```
liver.mouse.eQTL.bayesian.tau <- read.table("liver.mouse.eQTL.
    bayesian.tau.txt")

Fcomb = function(ps) #chi-square (Fisher, 1932, Lancaster, 1961)
{
        k = length(ps)
        temp = -2*sum(log(ps))
        pchisq(temp, 2*k, lower.tail = F)
}

Ncomb = function(ps) #normal (Liptak, 1958, Stouffer 1949)
{
        k = length(ps)
        z = qnorm((1-ps))
        Ts = sum(z)/sqrt(k) # sum(1-Phi^-1(1-p))/sqrt(k)
        pnorm(Ts, lower.tail = F)  #Same as 1-Phi

}

#META
metapval = apply(cbind(liver.mouse.eQTL.bayesian.tau$lung_pvalue,
     liver.mouse.eQTL.bayesian.tau$liver_pvalue), 1, Ncomb)
liver.mouse.eQTL.bayesian.tau$metapval = metapval

#Multiple posterior prob by 2
##CHANGE?
#liver.mouse.eQTL.bayesian.tau$p.below.0 = 2*liver.mouse.eQTL.
    bayesian.tau$p.below.0
```

```
26
27 #MT Method
28 mtresults<-read.table(paste0("MTeQTLs_ASE_3c_",sebsetn,"s.txt"),
       header = TRUE)
29
30 minmtresults<-sapply(liver.mouse.eQTL.bayesian.tau$ensembl_id,
       function(x) max(mtresults[mtresults$ensembl_id == as.character
       (x),"marginalP.liver"]))
31 newmtresults = data.frame(liver.mouse.eQTL.bayesian.tau$ensembl_
       id, minmtresults, liver.mouse.eQTL.bayesian.tau$eqtl)
32 colnames(newmtresults) = c("ensembl_id", "marginalp", "eqtl")
33
34 newresults = liver.mouse.eQTL.bayesian.tau[,c("ensembl_id", "lung
       _pvalue", "liver_pvalue", "metapval", "p.below.0", "eqtl")]
35 pvals = sort(newresults[,"lung_pvalue"])
36 nvals = length(newresults[,"lung_pvalue"])
37 result.lung = result.liver = result.meta = result.pp = result.mt
       = matrix(0, nvals, 3)
38 totaltrue = sum(newresults[,"eqtl"])
39 totalfalse = sum(newresults[,"eqtl"]==0)
40 j = 1
41 for (i in pvals)
42 {
43    result.lung[j,1] = i
44    result.lung[j,2] = sum( newresults[newresults[,"lung_pvalue"]<
          i,"eqtl"])/totaltrue  #sens
45    result.lung[j,3] = sum( newresults[newresults[,"lung_pvalue"
          ]>=i,"eqtl"]==0)/totalfalse #spec
46
```

```
47    result.liver[j,1] = i
48    result.liver[j,2] = sum( newresults[newresults[,"liver_pvalue"
         ]<i,"eqtl"])/totaltrue  #sens
49    result.liver[j,3] = sum( newresults[newresults[,"liver_pvalue"
         ]>=i,"eqtl"]==0)/totalfalse #spec
50
51    result.mt[j,1] = i
52    result.mt[j,2] = sum( newmtresults[newmtresults[,"marginalp"]<
         i,"eqtl"])/totaltrue  #sens
53    result.mt[j,3] = sum( newmtresults[newmtresults[,"marginalp"
         ]>=i,"eqtl"]==0)/totalfalse #spec
54
55    result.meta[j,1] = i
56    result.meta[j,2] = sum( newresults[newresults[,"metapval"]<i,"
         eqtl"], na.rm  = TRUE)/sum(newresults[,"eqtl"]==1, na.rm =
         TRUE) #sens
57    result.meta[j,3] = sum( newresults[newresults[,"metapval"]>=i,
         "eqtl"]==0, na.rm  = TRUE)/sum(newresults[,"eqtl"]==0, na.
         rm = TRUE) #spec
58
59    result.pp[j,1] = i
60    result.pp[j,2] = sum( newresults[newresults[,"p.below.0"]<i,"
         eqtl"], na.rm  = TRUE)/sum(newresults[,"eqtl"]==1, na.rm =
         TRUE)  #sens
61    result.pp[j,3] = sum( newresults[newresults[,"p.below.0"]>=i,"
         eqtl"]==0, na.rm  = TRUE)/sum(newresults[,"eqtl"]==0, na.rm
          = TRUE) #spec
62
63    j = j+1
```

```
}


plot(1-result.liver[,3], result.liver[,2], xlim = c(0,1), ylim =
    c(0,1), xlab= "1-Specity", ylab = "Sensitivy", pch = ".")
points(1-result.pp[,3], result.pp[,2], col = "red", pch = ".")
points(1-result.lung[,3], result.lung[,2], col = "purple", pch =
    ".")
points(1-result.meta[,3], result.meta[,2], col = "blue", pch = ".
    ")
points(1-result.mt[,3], result.mt[,2], col = "green", pch = ".")
legend(.7, .55, legend = c("orig - liver", "bayesian", "lung", "
    meta", "mt"), col = c("black", "red", "purple", "blue", "green
    "), pch = 1)
title(paste0("Z:abs_lung_beta; subsample:", sebsetn))


 dev.copy(pdf,"comparison.pdf")
dev.off()

library(flux)
orig_auc <- auc(1-result.liver[,3], result.liver[,2])
bayesian_auc <- auc(1-result.pp[,3], result.pp[,2])
lung_auc <- auc(1-result.lung[,3], result.lung[,2])
meta_auc <- auc(1-result.meta[,3], result.meta[,2])
mt_auc <- auc(1-result.mt[,3], result.mt[,2])
auc <- rbind(orig_auc, bayesian_auc,lung_auc,meta_auc,mt_auc)
auc <- cbind(auc, auc[ , 1]/auc["orig_auc", 1])
colnames(auc) <- c("auc", "FC")
```

```
88 auc
```