Dissection and Fine-mapping of trans-eQTL hotspots

by

Jianan Tian

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2015

Date of final oral examination: 08/25/2015

The dissertation is approved by the following members of the Final Oral Committee:
Karl Broman, Professor, Biostatistics & Medical Informatics
Christina Kendziorski, Professor, Biostatistics & Medical Informatics
Brian Yandell, Professor, Statistics and Horticulture
Cécile Ané, Professor, Statistics and Botany
Alan Attie, Professor, Biochemistry

ACKNOWLEDGMENTS

First, I would like to express my deepest gratitude towards my advisor Karl Broman for his advise, guidance and help from the very first years of my PhD life. I could never say enough thanks to him. His passion and diligence for his work, sharp insights of both statistical and biological problems, elegant coding style and organization skills have inspired me in every aspect as an productive researcher. I felt fully energized every time after our Tuesday meeting, even when I didn't sleep at all the night before. I am truly indebted to him more than he knows, for his unconditional support, patience and encouragement throughout these five years.

I would also like to thank all professors in my committee: Christina Kendziorski, Brian Yandell, Cécile Ané and Alan Attie, for their wonderful questions and helpful suggestions. I enjoy the discussions we had in every committee meeting.

Lastly, to my husband Chenliang and our son Erik, for all the moments and emotions we share. This work is dedicated to them.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The measurement of genome-wide gene expression levels in experimental crosses can accelerate the identification of genes contributing to complex phenotypes and further enables the genetic dissection of gene expression. Efforts to map the genetic loci (called expression quantitative trait loci, eQTL) that contribute to gene expression variation reveal numerous local-eQTL, where a gene's mRNA abundance is associated with genotype near its genomic location, and also *trans*-eQTL, where a genetic locus affects the mRNA abundance of genes elsewhere in the genome. In some cases, an eQTL can affect expression of numerous genes throughout the genome. This thesis describes methods for the dissection and fine-mapping of such *trans*-eQTL hotspots.

This work was motivated by a large mouse intercross that was undertaken to identify genes and pathways contributing to obesity-induced type II diabetes. Over 500 intercross mice were generated, and gene expression was assayed, via microarrays, in six tissues.

We first consider a particular *trans*-eQTL hotspot, on mouse chromosome 6, that showed broad effect on gene expression, solely in pancreatic islets. We developed and applied a method for fine-mapping this hotspot. We used individuals that showed no recombination event in the region of the eQTL to develop a classifier for predicting eQTL genotype from the gene expression phenotypes. Applying this classifier to the recombinant individuals, we converted the expression phenotypes into a co-dominant Mendelian trait. With additional genotyping in selected recombinant mice, we were able to reduce the eQTL interval from 3.4 Mbp to a 298 kb region containing just three genes.

This fine-mapping approach relies on an assumption of a single eQTL in the hotspot region.

In order to test this hypothesis, of a single eQTL versus multiple linked eQTL, we developed both exploratory graphical methods and a formal likelihood-based test. We applied our methods to the motivating example, and further performed a simulation study to evaluate the power of the formal statistical test as a function of the distance between two linked eQTL.

We have implemented our methods in an R package, *qtlpvl*. We illustrate the use of the software in a small simulated data set.

1 INTRODUCTION

The measurement of genome-wide gene expression levels in experimental crosses can accelerate the identification of genes contributing to complex phenotypes and further enables the genetic dissection of gene expression. Efforts to map the genetic loci (called expression quantitative trait loci, eQTL) that contribute to gene expression variation reveal numerous local-eQTL, where a gene's mRNA abundance is associated with genotype near its genomic location, and also *trans*-eQTL, where a genetic locus affects the mRNA abundance of genes elsewhere in the genome. In some cases, an eQTL can affect expression of numerous genes throughout the genome. This thesis describes methods for the dissection and fine-mapping of such *trans*-eQTL hotspots.

The work was motivated by a large mouse intercross conducted by Alan Attie (Department of Biochemistry, University of Wisconsin–Madison) and colleagues, which was undertaken to identify genes and pathways contributing to obesity-induced type II diabetes. Over 500 intercross mice were generated, and gene expression was assayed, via microarrays, in six tissues.

eQTL studies

Experimental crosses

Experimental cross between inbred strains have long been used to study the relationship between quantitative traits and underlying genetic loci. The chromosomes of F_2 offspring from intercross experiments have segments from either of the two parental stains, as a result of crossing over during meiosis. Figure 1.1 illustrates an intercross between inbred strains A and B, on autosomes. The F_1 hybrids all have genotype AB, with one chromosome set from strain A and the other from strain B. By crossing F_1 siblings, the F_2 offspring will have genotypes AA, AB and BB with probabilities $\frac{1}{4}$, $\frac{1}{2}$, and $\frac{1}{4}$ respectively.

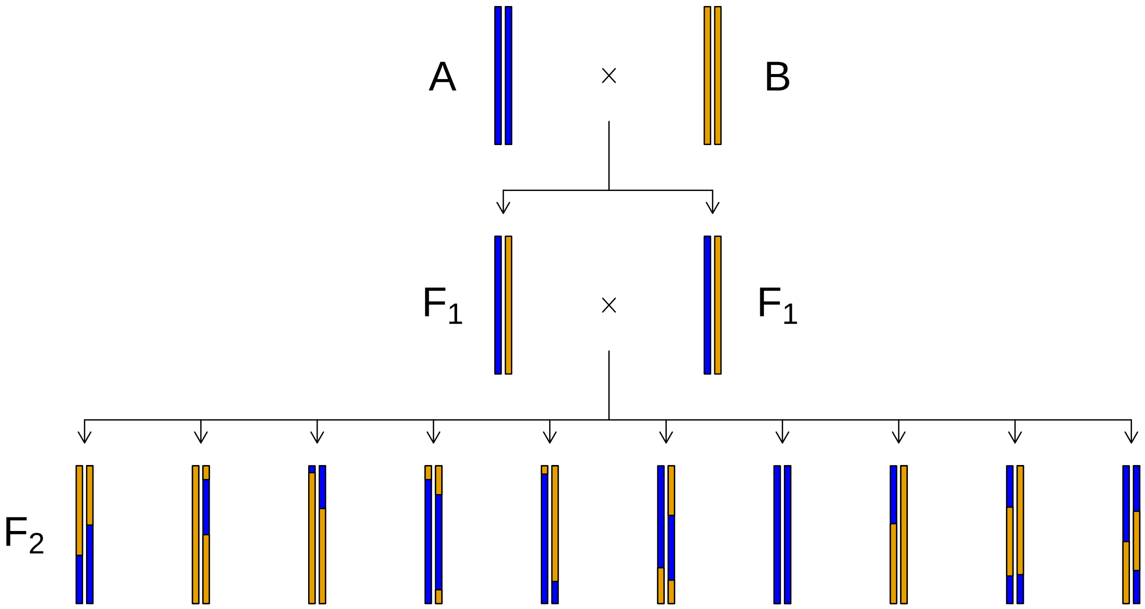


Figure 1.1: Illustration of an autosome in an intercross between two inbred strains, A and B.

LOD curves

For QTL analysis, we focus on Haley-Knott regression (Haley and Knott, 1992). A LOD score, measuring evidence for a QTL, is calculated for each position θ across the genome, as

$$\mathrm{LOD}(\theta) = \frac{n}{2} \log_{10} \frac{\mathrm{RSS}_0}{\mathrm{RSS}_\theta},$$

where RSS_θ is the residual sum of squares of the linear regression $y \sim \beta g_\theta + e$, where $g_\theta = 0, 1, 2$ for genotype AA, AB and BB, respectively, and RSS_0 is the residual sum of squares for the null model, with no QTL. With a pre-determined threshold T, we announce the existence of a quantitative trait locus (QTL) when $LOD(\theta) > T$. The threshold, T, is usually determined by either simulation or by a permutation test, to adjust for the multiple testing of all chromosome positions.

When measurement of gene expression (mRNA abundance) is taken as the quantitative trait, the underlying locus is called an expression quantitative trait locus (eQTL). As an illustration, Figure 1.2 shows the LOD curve for one gene expression trait mapping to chromosome 6, with a LOD score of 24.6 at 91.4 cM.

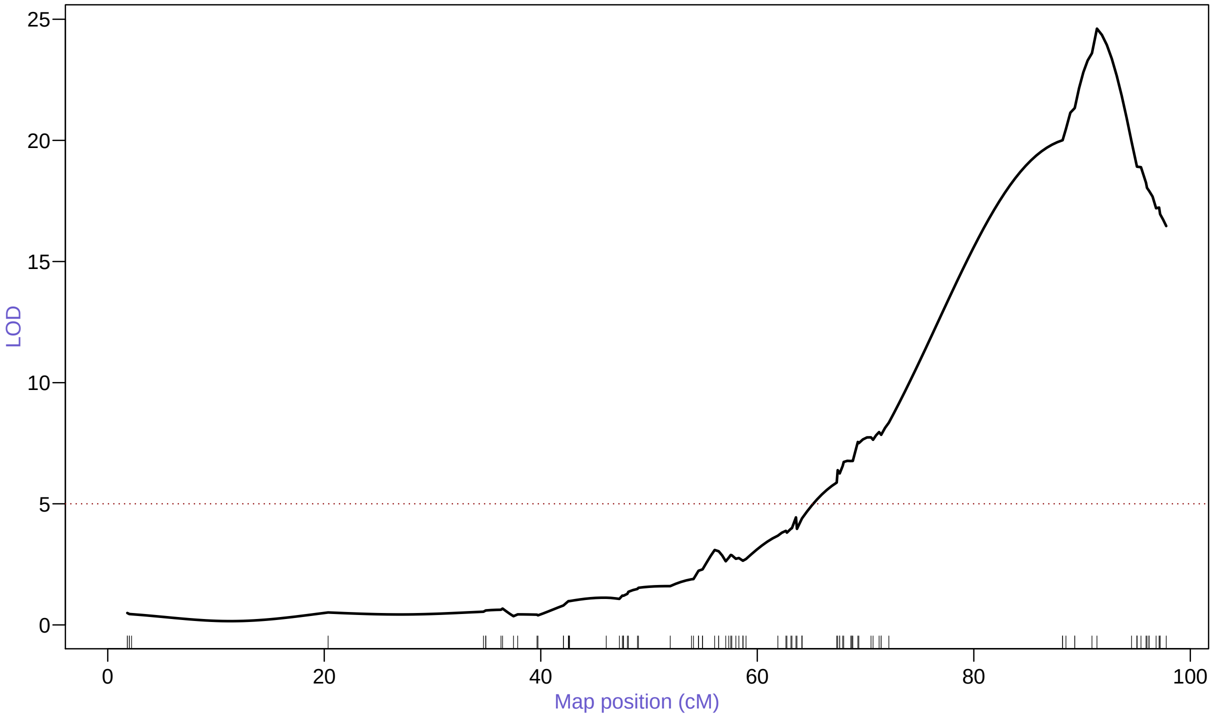


Figure 1.2: Illustration of LOD curve for one gene expression trait. The dashed red line marked threshold value at 5. Tick marks on the x-axis indicate the locations of the genetic markers. The eQTL is estimated to be at 91.4 cM, with LOD score 24.6.

Gene expressions can also be influenced by other variables, such as sex of the individual, environmental differences, and experimental condition in the process of measuring the expression. Some of these effects act additively, while others may interact with the QTL genotype. Controlling for these effects can help to reduce the residual variance and improve the power to identify the eQTL as well as the precision in estimating its position.

trans-eQTL hotspots

Figure 1.3 displays the results of applying QTL mapping to all gene expression traits from a single tissue, pancreatic islets. We controlled for batch as an additive covariate and sex as an interactive covariate. Each point in the figure corresponds to a peak in a LOD curve as in Figure 1.2. The y-axis is the genomic position of the gene, and the x-axis is the inferred QTL position. Along the diagonal, the expression traits sit near their eQTL, and these are called local-eQTL; otherwise they are called *trans*-eQTL. There are a number of vertical bands, where many expression traits map to the same location. These regions are called *trans*-eQTL hotspots.

We can count the number of traits in each region to define the hotspots more rigorously. We use a 10 cM sliding window and draw a curve of how many expression traits mapped in $(\theta - 5, \theta + 5)$ for every position θ on a chromosome. Figure 1.4 shows the curve for islet chromosome 2. The peak is near 75 cM with > 600 probes having an eQTL. There is also a minor peak near 35 cM with < 100 probes.

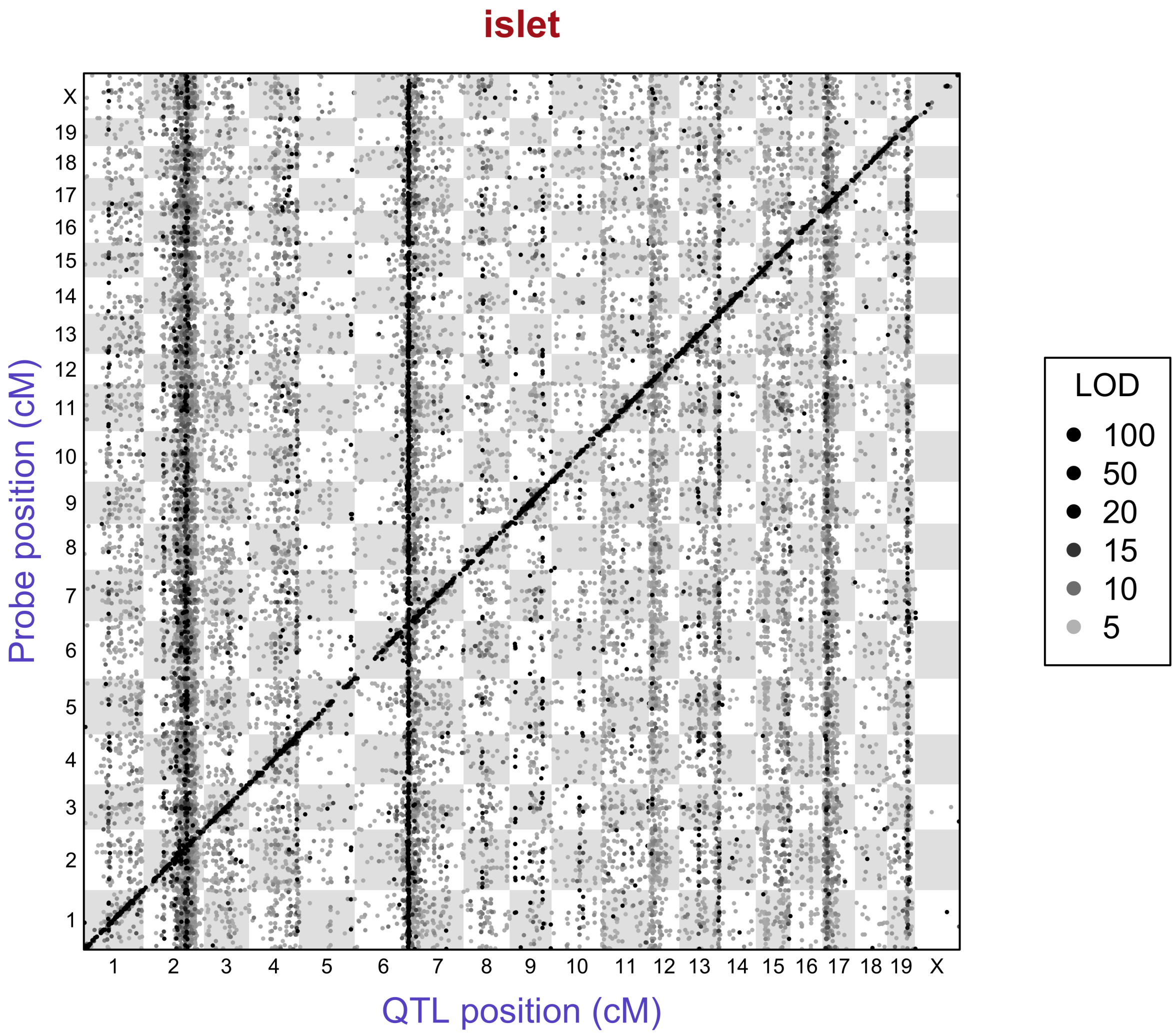


Figure 1.3: Inferred eQTL with LOD ⩾ 5 in pancreatic islets. Points correspond to peak LOD scores from single-QTL genome scans, for microarray probes with known genomic position. The y-axis is the position of the probe and the x-axis is the inferred QTL position. Points are shaded according to the corresponding LOD score, though we threshold at 100: all points with LOD ⩾ 100 are black.

Are *trans*-eQTL hotspots real? Statistical significance of hotspots

QTL hotspots are observed in many genetic studies (Breitling et al., 2008) and they are of particular interest because gene expressions mapping to the same location may indicate the existence of a genetic regulator. Several permutation approaches have been studied to

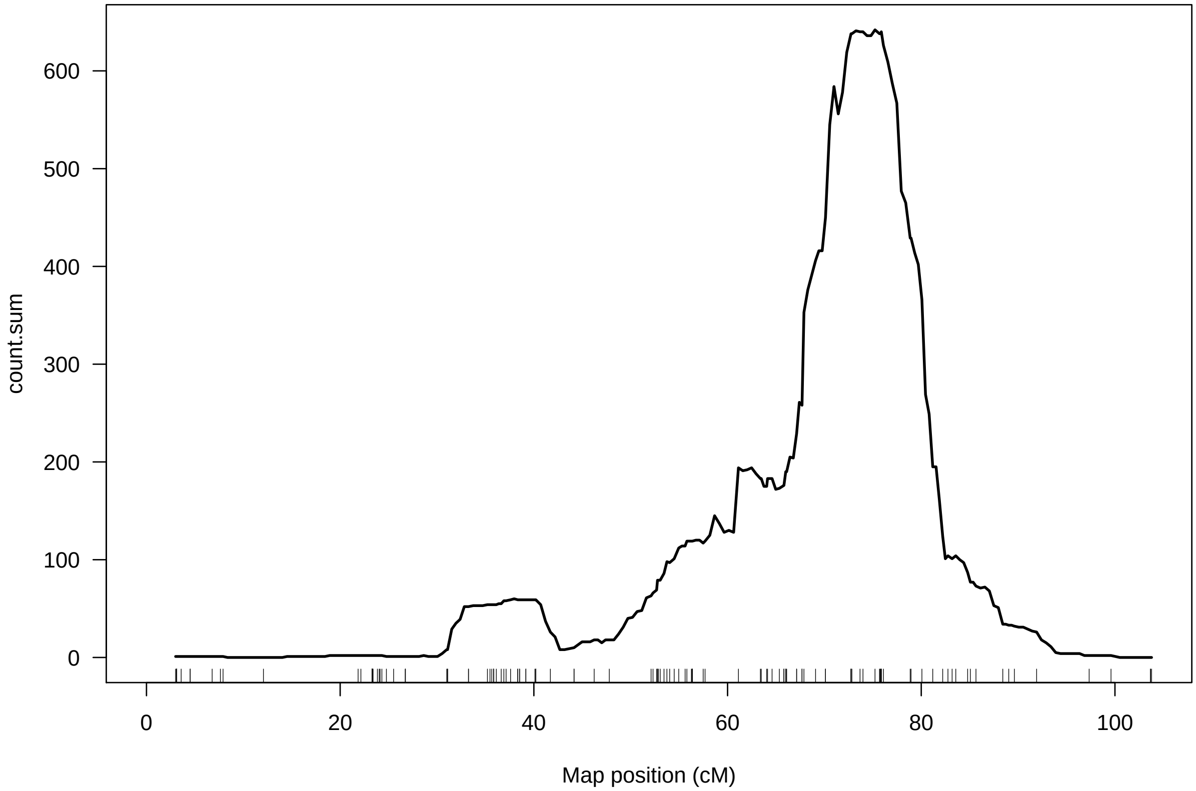


Figure 1.4: Number of expression traits with LOD $\geqslant 10$ in a 10 cM sliding window across chromosome 2, for pancreatic islets.

assess the statistical significance of such hotspots: Wu et al. (2008) proposed to permute on the observed QTL positions across the whole genome for each expression trait and use the distribution of maximal counts of expressions to assess statistical significance. This method permutes the expression traits independently and breaks the correlation structure among expressions. Breitling et al. (2008) criticized the method of Wu et al. (2008) as being too liberal in the presence of correlation among genes. They proposed a permutation method that shuffles the rows of the genotype data and thus breaks the linkage between genotype and expression but preserves the correlation structure. They used a predetermined LOD threshold and counted the number of traits with LOD score above this value. The choice of the LOD threshold has important influence on the results. To account for the magnitude of the LOD scores, Neto et al. (2012) proposed a quantile-based permutation approach that simultaneously accounts for the number and the LOD scores of expression traits within the hotspots. By considering a sliding scale of mapping thresholds, this method can assess the statistical significance of both small and large (in terms of LOD score) hotspots.

Are *trans*-eQTL hotspots artifacts? Batch effects of microarray data

Batch effects (i.e., artifacts arising from technical or environmental factors) are common in microarray experiments, and several methods has been proposed to remove or control them for eQTL analysis. However, these methods cannot distinguish *trans*-eQTL hotspots from batch effects, and there is some controversy about whether *trans*-eQTL hotspots are themselves artifacts and whether one should control for them, as one does for batch effects, in eQTL analysis.

Surrogate Variable Analysis (SVA: Leek and Storey, 2007) is a popular method in differential expression analysis. After controlling for the primary variable, the authors select several principal components (PCs) of the residual matrix and build surrogate variables from the subset of the expression data that are most associated with the PCs. These surrogate variables are used as covariates in the subsequent analysis. In eQTL analysis where there is no primary variable, surrogate variables could be built directly from the standardized expression data.

The ICE (*Intersample Correlations Emending*) QTL mapping method used a linear mixed effect model where the random effect follows a multivariate normal distribution with covariance matrix proportional to the intersample covariance matrix (Kang et al., 2008). The authors argued that regulatory hotspots are 'spurious' since: *i*) the locations of regulatory hotspots are inconsistent across disjoint sets of biologically replicated samples and *ii*) stronger *trans*-regulatory hotspots frequently appear with randomly permuted SNPs. In one of their examples, a true hotspot is also removed.

Stegle et al. (2010) proposed a Bayesian framework, PEER, that jointly models genotype, known factor and hidden factor to increase power of eQTL studies. The method could find up to three times more eQTLs than a standard method.

Listgarten et al. (2010) proposed a linear mixed effect model that controls both batch effects and population structure, if present. They suggest that their approaches provide better-calibrated P-values than ICE and SVA and maximum power of eQTL detection.

Gagnon-Bartsch and Speed (2012) proposed a method, RUV-2, to *Remove Unwanted Variation*, that uses a pre-defined set of negative control genes. They first run factor analysis on the negative control genes and then use the factors as covariates. The performance depends greatly on the choice of negative controls. The method of factor analysis is not quite important, but the choice of k, the number of factors used, is critical.

Overall, the principal component based methods such as SVA have the problem that the PCs might be highly correlated with the hotspot QTL, thus controlling for the PCs would remove the *trans*-eQTL hotspots; ICE buries the hotspot QTL in the intersample correlation; PEER avoided controlling for the hotspot QTL by assuming the hidden factor follows a normal distribution; Listgarten et al. (2010) takes genotype structure into the mixed effect model and thus preserves the hotspots. We believe that controlling for known batch effects is important and necessary, but controlling for hidden batch effects should be performed with caution, since real signals may be eliminated.

Organization

Many eQTL studies analyze each expression trait separately. In addition to increasing false discovery rate (FDR), the power to detect *trans*-eQTL hotspot can be low. Multiple trait analysis has been proposed for both composite interval mapping method (Jiang and Zeng, 1995) and least square regression (Knott and Haley, 2000), but when it comes to gene expression data, the dimension of the traits is so big that these can not be immediately applied.

Dimension reduction methods, such as principal component analysis (PCA), are also used in some studies (Lan et al., 2003; Alimi et al., 2013; Weller et al., 1996; Mangin et al., 1998). These methods analyze the first several PCs of the expressions and thus save a lot of computation, but the results can be hard to interpret.

In Chapter 2, we develop a method for defining and fine-mapping an eQTL interval. By converting the multivariate gene expression phenotype into a Mendelian trait and genotyping with additional markers, we were able to reduce islet chromosome 6 hotspot form 3.4 Mbp to 298 kb in the study of gene expressions in a mouse intercross experiment. In Chapter 3, we describe methods to distinguish between total pleiotropy versus close linkage for *trans*-eQTL hotspots: Is there a single QTL underlying a *trans*-eQTL hotspot, or are there multiple tightly linked QTLs? Both exploratory methods and formal likelihood-based methods for this problem are provided. In chapter 4, we introduce and illustrate the usage of a software package, R/qtlpvl, that implements these methods.

REFERENCES

Alimi, N., M. Bink, J. Dieleman, J. Magán, A. Wubs, A. Palloix, and F. van Eeuwijk. 2013. Multi-trait and multi-environment QTL analyses of yield and a set of physiological traits in pepper. *Theoretical and Applied Genetics* 126(10):2597–2625.

Breitling, R., Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. de Haan, A. I. Su, et al. 2008. Genetical genomics: spotlight on QTL hotspots. *PLoS Genetics* 4(10):e1000232.

Gagnon-Bartsch, J. A., and T. P. Speed. 2012. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 13(3):539–552.

Haley, C. S., and S. A. Knott. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69:315–324.

Jiang, C., and Z. B. Zeng. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140(3):1111–1127.

Kang, H. M., C. Ye, and E. Eskin. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180(4): 1909–1925.

Knott, S. A., and C. S. Haley. 2000. Multitrait least squares for quantitative trait loci detection. *Genetics* 156(2):899–911.

Lan, H., J. P. Stoehr, S. T. Nadler, K. L. Schueler, B. S. Yandell, and A. D. Attie. 2003. Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics* 164(4):1607–1614.

Leek, J. T., and J. D. Storey. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* 3(9):e161.

Listgarten, J., C. Kadie, E. E. Schadt, and D. Heckerman. 2010. Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences* 107(38):16465–16470.

Mangin, B., P. Thoquet, and N. Grimsley. 1998. Pleiotropic QTL analysis. *Biometrics* 54(1): 88–99.

Neto, E. C., M. P. Keller, A. F. Broman, A. D. Attie, R. C. Jansen, K. W. Broman, and B. S. Yandell. 2012. Quantile-based permutation thresholds for quantitative trait loci hotspots. *Genetics* 191(4):1355–1365.

Stegle, O., L. Parts, R. Durbin, and J. Winn. 2010. A bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Computational Biology* 6(5):e1000770.

Weller, J., G. Wiggans, P. VanRaden, and M. Ron. 1996. Application of a canonical transformation to detection of quantitative trait loci with the aid of genetic markers in a multi-trait experiment. *Theoretical and Applied Genetics* 92(8):998–1002.

Wu, C., D. L. Delano, N. Mitro, S. V. Su, J. Janes, P. McClurg, S. Batalov, G. L. Welch, J. Zhang, A. P. Orsth, et al. 2008. Gene set enrichment in eQTL data identifies novel annotations and pathway regulators. *PLoS Genetics* 4(5):e1000070.

2 FINE-MAPPING TRANS-EQTL HOTSPOTS

Identification of *Slco1a6* as a candidate gene that broadly affects gene expression in mouse pancreatic islets [1]

Jianan Tian*, Mark P. Keller†, Angie T. Oler†, Mary E. Rabagalia†,
Kathryn L. Schueler†, Donald S. Stapleton†, Aimee Teo Broman‡,
Wen Zhao**, Christina Kendziorski‡, Brian S. Yandell*,§,
Bruno Hagenbuch**, Karl W. Broman‡, Alan D. Attie†

Departments of *Statistics, †Biochemistry, ‡Biostatistics & Medical Informatics, and
§Horticulture, University of Wisconsin–Madison, Madison, Wisconsin 53706, and
**Department of Pharmacology, Toxicology and Therapeutics,
The University of Kansas Medical Center, Kansas City, Kansas 66160

[1] This paper is accepted pending minor revisions at Genetics (2015).

Abstract

We surveyed gene expression in six tissues in an F_2 intercross between mouse strains C57BL/6J (abbreviated B6) and BTBR *$T^+ tf$* /J (abbreviated BTBR) made genetically obese with the *$Leptin^{ob}$* mutation. We identified a number of expression quantitative trait loci (eQTL) affecting the expression of numerous genes distal to the locus, called *trans*-eQTL hotspots. Some of these *trans*-eQTL hotspots showed effects in multiple tissues, whereas some were specific to a single tissue. An unusually large number of transcripts (~ 8% of genes) mapped in *trans* to a hotspot on chromosome 6, specifically in pancreatic islets. By considering the first two principal components of the expression of genes mapping to this region, we were able to convert the multivariate phenotype into a simple Mendelian trait. Fine-mapping the locus by traditional methods reduced the QTL interval to a 298 kb region containing only three genes, including *Slco1a6*, one member of a large family of organic anion transporters. Direct genomic sequencing of all *Slco1a6* exons identified a non-synonymous coding SNP that converts a highly conserved proline residue at amino acid position 564 to serine. Molecular modeling suggests that Pro564 faces an aqueous pore within this 12-transmembrane domain-spanning protein. When transiently overexpressed in HEK293 cells, BTBR OATP1A6-mediated cellular uptake of the bile acid taurocholic acid (TCA) was enhanced compared to B6 OATP1A6. Our results suggest that genetic variation in *Slco1a6* leads to altered transport of TCA (and potentially other bile acids) by pancreatic islets, resulting in broad gene regulation.

Introduction

The measurement of genome-wide gene expression levels in segregating populations (such as the offspring of a cross between two inbred mouse strains), an endeavor termed genetical genomics (Jansen and Nap, 2001) or expression genetics (Broman, 2005), offers the promise of accelerating the identification of genes contributing to variation in complex phenotypes and further enables the genetic dissection of gene expression regulation (reviewed in Albert and Kruglyak, 2015). Early experiments in yeast (Brem et al., 2002; Yvert et al., 2003) mouse (Schadt et al., 2003), and human (Cheung et al., 2003; Morley et al., 2004) revealed some of the basic features of the genetic architecture of gene expression variation, including the prominent effects of local expression quantitative trait loci (eQTL), where a gene's mRNA abundance is strongly associated with genotype near its genomic location. There are also *trans*-acting effects, where single loci affect the mRNA abundance of large numbers of genes located throughout the genome.

To identify genes and pathways that contribute to obesity-induced type II diabetes, we constructed a large mouse F_2 intercross between diabetes-resistant (C57BL/6J, abbreviated B6) and diabetes-susceptible (BTBR $T^{+}tf$ /J, abbreviated BTBR) mouse strains. Greater than 500 F_2 offspring were generated. All were genetically obese through introgression of the leptin mutation ($Lep^{ob/ob}$) and all were sacrificed at 10 weeks of age, the age when essentially all BTBR ob/ob mice are diabetic. Thus, the screen focused on genetic differences between the B6 and BTBR mouse strains. In addition to measuring numerous diabetes-related clinical phenotypes (e.g., circulating levels of insulin and glucose), we conducted genome-wide gene expression profiling in six tissues of every F_2 mouse; adipose, gastrocnemius muscle, hypothalamus, pancreatic islets, kidney, and liver.

We identified numerous *trans*-eQTL hotspots. Some were common to more than one tissue, whereas others were strongly tissue-specific. We identified a particularly striking *trans*-eQTL hotspot on distal chromosome 6, at ~142 Mb. This hotspot was observed exclusively in pancreatic islets and affected ~2,400 transcripts encoded by genes located throughout the genome. Using principal component analysis of the transcripts mapping to the locus with the strongest LOD scores (LOD ~100), we were able to infer the eQTL genotype for all mice, thereby converting the multivariate gene expression phenotype to a co-dominant Mendelian trait. We applied the traditional method for fine-mapping a Mendelian trait. It involved identifying recombinant mice and genotyping additional markers to precisely define the location of the recombination events. This allowed us to refine the location of the *trans*-eQTL to a 298 kb interval containing just three genes.

Among the three genes included in the 298kb region are two members of the *SLCO* gene family, *Slco1a5* and *Slco1a6*, which encode the organic anion transporting polypeptide (OATP) 1A5 and OATP1A6 (Hagenbuch and Stieger, 2013). OATPs have been shown to transport numerous endogenous substrates, including bile acids such as taurocholic acid (Hagenbuch and Stieger, 2013). In addition to their role as detergents, bile acids have hormone-like properties due to their ability to activate nuclear hormone receptors (Parks et al., 1999). This makes genetic variation in a bile acid transporter a plausible mechanism for the differential regulation of numerous genes. We characterized functional differences between the two *Slco1a6* variants of B6 and BTBR mice and conclude that genetic variation in *Slco1a6* leads to altered transport of taurocholic acid (TCA), and potentially other bile acids, by pancreatic islets, resulting in broad gene regulation.

Materials and Methods

Mice and genotyping

C57BL/6J (abbreviated B6 or B) and BTBR $T^{+}tf$/J (abbreviated BTBR or R) mice were purchased from the Jackson Laboratory (Bar Harbor, ME) and bred at the Biochemistry Department at the University of Wisconsin–Madison. The $Lep^{ob/ob}$ mutation, which originated in the B6 strain (Ingalls et al., 1950), was introgressed into the BTBR strain. F_1 $Lep^{ob/ob}$ were made fertile by employing adipose tissue transplants from wild type mice to restore leptin.

All mice were genotyped with the 5K GeneChip (Affymetrix). A large set of DNA sample mix-ups were identified, by comparing observed genotypes to predictions based on large-effect local eQTL, and corrected (Broman et al., 2014). After data cleaning, there were 519 F_2 mice genotyped at 2,057 informative markers, including 20 on the X chromosome. The leptin gene resides on proximal chromosome 6, at ~29.0 Mbp, and the proximal 32 Mbp of this chromosome showed marked segregation distortion, with excess B6 homozygotes and reduced BTBR homoyzogotes. However, the *trans*-eQTL hotspot under study is at the opposite end of the chromosome from the leptin gene, and the region around the hotspot segregated normally.

Gene expression microarrays

Gene expression was assayed with custom two-color ink-jet microarrays manufactured by Agilent Technologies (Palo Alto, CA). RNA preparations were performed at Rosetta Inpharmatics (Merck & Co.). Six tissues from each F_2 mouse were used for expression profiling; adipose, gastrocnemius muscle (abbreviated gastroc), hypothalamus (abbreviated hypo), pancreatic islets (abbreviated islet), kidney, and liver. Tissue-specific mRNA pools for each tissue were used for the reference channel, and gene expression was quantified as the ratio of the mean $\log_{10}$ intensity (mlratio). For further details, see Keller et al. (2008). In the final data set, there were 519 mice with gene expression data on at least one tissue (487 for adipose, 490 for gastroc, 369 for hypo, 491 for islet, 474 for kidney, and 483 for liver). The microarray included 40,572 total probes; we focused on the 37,797 probes with known location on one of the autosomes or the X chromosome.

QTL analysis

For QTL analysis, we first transformed the gene expression measures for each microarray probe in each of the six tissues to normal quantiles, taking $\Phi^{-1}[(R_i - 0.5)/n]$, where Φ is the cumulative distribution function for the standard normal distribution and R_i is the rank in $\{1, \ldots, n\}$ for mouse i. We then performed single-QTL genome scans separately for each probe in each tissue, by Haley-Knott regression (Haley and Knott, 1992) with microarray batch included as an additive covariate and with sex included as an interactive covariate (i.e. allowing the effects of QTL to be different in the two sexes). Calculations were performed at the genetic markers and at a set of pseudomarkers inserted into marker intervals, selected so that adjacent positions were separated by $\leqslant 0.5$ cM. We calculated conditional genotype probabilities, given observed multipoint marker genotype data, using a hidden Markov model assuming a genotyping error rate of 0.2%, and with genetic distances converted to recombination fractions with the Carter-Falconer map function (Carter and Falconer, 1951).

For each probe in each tissue, we focused on the single largest LOD score peak on each chromosome, and LOD score peaks $\geqslant 5$ (corresponding to genome-wide significance at the 5% level, for a single probe in a single tissue, determined by computer simulations under the null hypothesis of no QTL).

Inference of eQTL genotype

To infer the eQTL genotype of each mouse at the islet chromosome 6 *trans*-eQTL hotspot, we selected the 181 microarray probes that mapped to the locus with LOD $\geqslant 100$, but were not associated with a gene located anywhere on chromosome 6, thereby ensuring they represented a *trans*-eQTL. We computed the first two principal components, using the mlratio measure of gene expression (i.e. prior to the normal quantile transformation that was used in the QTL analyses). The scatterplot of the two principal components revealed three distinct clusters of mice, corresponding to BB (homozygous B6), BR (heterozygous) or RR (homozygous BTBR). We used the mice without a recombination event in the region to define the genotype for these clusters, from which we could then infer the eQTL genotype for the mice with a recombination event in the region of the QTL.

Fine-mapping the eQTL

Fine-mapping of the chromosome 6 islet eQTL proceeded by the traditional method for a simple Mendelian trait: we identified the smallest interval in which individuals' genotypes at genetic markers were consistent with their inferred eQTL genotype. We selected individuals with recombination events flanking this interval and genotyped these recombinant animals at additional SNP markers in the interval, in order to more precisely localize their recombination events and so refine the QTL interval. We used SNPs predicted from high-throughput sequencing of genomic DNA from two BTBR mice (Eric Schadt, Mt. Sinai, personal communication) to guide our initial selection of the additional markers that were used to genotype recombinant individuals. Predicted SNPs proximal to the eQTL locus on chromosome 6 were confirmed with PCR amplification followed by Sanger Sequencing in BTBR and B6 genomic DNA. Confirmed SNPs were then used to further map the eQTL region using F_2 mice whose recombinations were within the eQTL interval. File S1 provides a complete list of the SNPs that were used to narrow our eQTL interval.

Characterization of the transport protein OATP1A6 encoded by *Slco1a6* gene

Radiolabeled [^{3}H]-taurocholic acid (10 Ci/mmol), [^{3}H]-cholic acid (30 Ci/mmol) and [^{3}H]- methotrexate (21.6 Ci/mmol) were purchased from American Radiolabeled Chemicals, Inc. Radiolabeled [^{3}H]-estrone-3-sulfate (45 Ci/mmol), [^{3}H]- (D-Pen^2, D-Pen^5)-enkephalin (44 Ci/mmol), [^{3}H]-estradiol-17-β-glucuronide (41.4 Ci/mmol) were obtained from PerkinElmer (Boston, MA). [^{3}H]-bromosulfophthalein (11.5 Ci/mmol) was obtained from International Isotope Clearing House (Leawood, KS, USA). All other chemicals used for the characterization were purchased from Sigma-Aldrich.

Cloning of mouse OATP1A6 and expression in HEK293 cells

The open reading frame (ORF) encoding mouse OATP1B6 was PCR amplified and cloned using cDNA prepared from RNA that was isolated from kidneys from B6 or BTBR mice using the following primers: forward primers containing an *NheI* restriction site were: 5′-AGAGGCTAGCACCATGGGAGAACCTGGGA-3′ for B6, and 5′-AGAGGCTAGCACCAT GGGAGAACCTGAGA-3′ for BTBR, and a reverse primer for both strains containing a *Not I* restriction site: 5′-AGAGGCGGCCGCCCTACAGCTTAGTTTTCAGTTCTCCA-3′. The PCR products were first cloned directionally into pcDNA5/FRT. In order to achieve the same level of expression for the B6 *vs.* BTBR ORFs, they were subcloned into a vector that contained the 3′untranslated region of mouse OATP1A6 (strain FVB/N, accession number BC071214; purchased from Thermo Fisher Scientific, formerly Open Biosystems). All clones were sequence verified on both strains. Human embryonic kidney (HEK) 293 cells (ATCC, Manassas, VA) were grown at 37 °C in a humidified 5% CO_2 atmosphere in DMEM High Glucose (ATCC) supplemented with 10% FBS (Hyclone), 100 U/ml penicillin, and 100 µg/ml streptomycin. HEK293 cells were plated at 250,000 cells per well in 24-well-plates pre-coated with 0.1 mg/ml poly-D-lysine. Twenty-four hours later cells were transfected with 0.5 µg plasmid DNA and 1.5 ul Fugene HD (Promega) per well. The uptake assays described below were performed 48 hours following transfection.

Site directed mutagenesis of mouse OATP1A6

The serine at position 564 in B6 was mutated to a proline and the proline at position 564 in BTBP to a serine using the Quick Change site directed mutagenesis kit (Agilent

Technologies, Santa Clara, CA) using the following primers: 5′-CAAAGCGCCAAAGTAA ATAGATGCAGGAATACCAGCAAATA-3′ and 5′-TATTTGCTGGTATTCCTGCATCTAT TTACTTTGGCGCTTTG-3′ for B6; 5′-CAAAGCGCCAAAGTAAATAGGTGCAGGAATA CCAGCAAATA-3′ and 5′-TATTTGCTGGTATTCCTGCACCTATTTACTTTGGCGCTTTG -3′ for BTBR. All mutants were verified by DNA sequencing.

Uptake activity assays for mouse OATP1A6

HEK293 cells transfected with either an empty vector (EV), the ORF encoding B6 OATP1A6 or the ORF encoding BTBR OATP1A6 were washed three times with 1 ml of pre-warmed (37 °C) uptake buffer containing 142 mM NaCl, 5 mM KCl, 1 mM KH_2PO_4, 1.2 mM $MgSO_4$, 1.5 mM $CaCl_2$, 5 mM glucose, and 12.5 mM HEPES, pH 7.4. Following the wash, 200 µl of uptake buffer (37 °C) containing the radiolabeled substrates were added to the well. At indicated times, uptake was terminated by four washes of 1 ml of ice-cold uptake buffer. Cells were then lysed with 300 µl 1% Triton X-100 in PBS at room temperature for 20 min. Two hundred µl of cell lysate were transferred to a 24-well scintillation plate (Perkin Billerica, MA) and 750 µl Optiphase Supermix scintillation cocktail (Perkin Elmer) was added to each well. Radioactivity was measured in a Microbeta Trilux liquid scintillation counter. The remaining cell lysates were transferred to 96-well plates to determine total protein concentration using the bicinchoninic acid protein assay (Pierce Biotechnology, Inc. Rockford, IL). All transport measurements were corrected by the total protein concentration and by the specific activities of the different substrates to yield pmoles per mg protein. All experiments were performed three independent times, with each study including triplicate technical measurements.

Software

All statistical analyses were conducted with R (R Development Core Team, 2015). QTL analyses were performed with the R package, R/qtl (Broman et al., 2003).

Data availability

The gene expression microarray and genotype data are available at the QTL Archive, now part of the Mouse Phenome Database:

http://phenome.jax.org/db/q?rtn=projects/projdet&reqprojid=532

Results

Identification of eQTL

We performed a single-QTL genome scan for all genes in six tissues of ~500 F_2 mice constructed from an intercross of diabetes-resistant (B6) and diabetes-susceptible (BTBR) mice. For all genes, we identified the single largest LOD score on each chromosome. The inferred expression quantitative trait locus (eQTL) for all genes with LOD $\geqslant 5$ are displayed in Figure 2.1, with the y-axis corresponding to the genomic position of the microarray probe and the x-axis corresponding to the estimated eQTL position. As expected, we see a large number of local-eQTL along the diagonal for each tissue-specific panel. These local-eQTL correspond to genes for which expression or mRNA abundance is strongly associated with genotype near their genomic position and are often referred to as cis-eQTL.

In addition to the local-eQTL, there are a number of prominent vertical bands: genomic loci that influence the expression of genes located throughout the genome. These distally-mapping eQTL are often referred to as *trans*-eQTL, and by virtue of their common genetic architecture (co-mapping), yield the *trans*-eQTL hotspots.

Overall, we detected many more *trans*-eQTL than *cis*-eQTL. The *trans*-eQTL hotspots can show either remarkable tissue specificity or be observed in multiple tissues. For example, a locus near the centromere of chromosome 17 (11.73 cM) shows effects in all tissues, with an average of 1498 microarray probes mapping to this region with LOD $\geqslant 5$ in the tissues surveyed. In contrast, a *trans*-eQTL hotspot located at the distal end of chromosome 6 was only observed in pancreatic islets. Furthermore, this islet-specific *trans*-eQTL hotspot was the strongest that we detected, with more eQTL mapping to this locus than any other locus identified in our study. We chose this *trans*-eQTL hotspot for further study.

Figure 2.2 displays, for each tissue, the estimated QTL location and LOD score for all probes mapping to chromosome 6 with LOD $\geqslant$ 5. The peak marker for the chromosome 6 *trans*-eQTL hotspot is rs8262456, located at 91.4 cM (141.52 Mbp in NCBI build 37). For pancreatic islets, in the 10 cM interval centered at this marker, there are 2889 probes with LOD peaks $\geqslant$ 5 (7.6% of the 37,797 probes considered), including 1700 probes with LOD $\geqslant$ 10 (4.5% of all probes) and 199 probes with LOD $\geqslant$ 100. These 2889 co-mapping probes are located throughout the genome, including all autosomes and the X chromosome, and include 756 with no annotated gene and 2133 probes with a gene annotation, corresponding to 2085 distinct genes (2038 genes with one probe, 46 genes with two probes, and one gene with three probes). This *trans*-eQTL hotspot is specific to islet cells; the other five tissues have 66–140 probes mapping to this 10 cM interval with LOD $\geqslant$ 5, and 36–66 probes with LOD $\geqslant$ 10. These results suggest that a gene located at ~141.5 Mbp on chromosome 6 regulates the expression of genes throughout the genome in an islet-specific fashion.

The *trans*-eQTL shows approximately additive allele effects in all microarray probes (Figure S1), and approximately equal numbers of probes having an effect in each direction: Among the 2889 probes mapping with LOD $\geqslant$ 5 to the 10 cM interval centered at the peak marker for the *trans*-eQTL, there are 1432 for which the BTBR homozygote has higher expression than the B6 homozygote, and 1457 for which the BTBR homozygote has lower expression.

Prediction of QTL genotype from phenotype

To identify gene candidates at the islet chromosome 6 locus, our first step was to narrow the effective interval. We calculated the first two principal components of the expression values for the 181 microarray probes that mapped with LOD $\geqslant$ 100 to the 10 cM interval centered at the peak marker for the chromosome 6 islet *trans*-eQTL hotspot but that did not reside on chromosome 6. (We considered probes with local-eQTL to be more likely affected by a separate polymorphism.) A scatterplot of these two principal components reveals three distinct clusters of mice (Figure 2.3). Points corresponding to mice without a recombination event in the 10 cM interval are colored according to their genotype in the interval. This indicates that the three clusters correspond to the three possible eQTL genotypes. There were 74 mice with recombination events in the interval (displayed in yellow in Figure 2.3). We inferred the eQTL genotype for these recombinant mice, based on the cluster in which they reside. In this manner, the multivariate gene expression phenotype was converted to a co-dominant Mendelian phenotype.

Fine-mapping the eQTL

By comparing the inferred eQTL genotypes at the islet chromosome 6 *trans*-eQTL hotspot to the observed marker genotypes in the region, we localized the QTL to the 3.4 Mbp interval between markers rs8262456, at 141.52 Mbp, and rs13479085, at 144.91 Mbp; see Figure 2.4A, in which the diamonds represent the inferred eQTL genotypes. There were 29 mice with recombination events in this interval.

Genotyping of 5 additional markers in the region (Figure 2.4B) refined the QTL position to a 904 kb interval, from 141.517–142.421 Mbp. There were eight mice with recombination events in this interval.

Genotypes at an additional 14 markers in these eight mice (Figure 2.4C) refined the QTL position to the 298 kb interval from 141.979–142.277 Mbp. The interval contains three genes, *Slco1a6*, *Slco1a5*, and *Iapp*. Recombination events in three mice define this interval: two at the proximal end and one at the distal end. Additional genotyping that refined the locations of these recombination events would not permit us to exclude any of the three genes. The expression levels of *Slco1a6* and *Iapp* are strongly associated with genotype in the region, with LOD scores of 161 and 10.6, respectively. *Slco1a5,* on the other hand, has a maximum LOD score on chromosome 6 of 2.4 (and at the other end of the chromosome) and shows evidence for a QTL on chromosome 7, with a LOD score of 9.2. That *Slco1a6* shows such a strong local-eQTL makes it our strongest candidate.

Characterization of *Slco1a6*

To characterize the genetic differences between B6 *vs.* BTBR *Slco1a6*, we sequenced all the exons from the *Slco1a6* gene in both mouse strains. In addition, we cloned cDNA prepared from total kidney RNA purified from each strain. Compared to B6 as the reference sequence, the BTBR sequence differed in 9 nucleotides, resulting in 6 amino acid changes (Table S1). Alignment of the amino acid sequence for BTBR OATP1A6 (the gene product for mouse *Slco1a6*) with 10 other OATP1A family members from human, rat and mouse (including the B6 reference sequence), revealed the degree of conservation for these 6 amino acid changes. Two SNPs were selective for either B6 or BTBR among the OATP1A sub-family:

1) at position 5, B6 OATP1A6 is a Glycine residue, whereas in all other proteins of this sub-family this residue is a Glutamate residue; and 2) at position 564, BTBR OATP1A6 is a Serine residue, whereas it is a Proline residue in the other family members. At amino acid positions 456 and 658, B6 OATP1A6 and the human OATP1A2 are the same, and different than all other members of the sub-family as well as BTBR OATP1A6. Finally, the remaining two differences between B6 and BTBR OATP1A6 were also different among the other family members.

Molecular modeling of the protein structure for mouse OATP1A6 suggests that the conversion of the highly conserved Pro^{564} forms a kink in the alpha-helix for one of the transmembrane domains, and this Pro residue faces into the central aqueous pore of the protein through which substrates are translocated (Figure 2.5A) (Meier-Abt et al., 2006). Conversion of this to Ser^{564} in BTBR OATP1A6 may alter the substrate selectivity and/or the transport activity of the protein. The Gly^{5} difference between B6 and BTBR OATP1A6 falls into the unstructured amino-terminal region of the protein.

To functionally characterize mouse OATP1A6, encoded by *Slco1a6*, and to evaluate potential differences between the B6 and BTBR proteins that may be related to the coding variants we detected in our sequencing, we performed cellular uptake studies of several known substrates for the OATP family of transporters. HEK293 cells were transiently transfected with the complete open reading frame (ORF) encoding either B6 or BTBR OATP1A6. With the exception of methotrexate (MTX), the uptake of all other substrates was significantly enhanced in cells overexpressing either variant of mouse OATP1A6 (Figure 2.5B). Furthermore, cells expressing BTBR OATP1A6 showed significantly elevated uptake of several substrates compared to cells expressing B6 OATP1A6. Of all substrates tested, uptake of the primary conjugated bile acid, taurocholate (TCA) showed the greatest signal above empty-vector control cells, for both B6 and BTBR OATP1A6. The primary unconjugated bile acid cholate (CA) was also transported, but less efficiently than TCA (Figure 2.5B) whereas taurochenodeoxycholic acid (TCDCA) was not a substrate for either B6 or BTBR OATP1A6 (data not shown). These results suggest that mouse OATP1A6 can mediate cellular uptake of certain bile acids and that one or more of the polymorphisms we identified in B6 vs. BTBR OATP1A6 yielded enhanced transport activity for the BTBR protein.

The studies illustrated in Figure 2.5B surveyed the substrate selectivity of mouse OATP1A6. In order to evaluate kinetic differences between B6 and BTBR OATP1A6, we measured the TCA uptake at increasing concentrations in HEK cells overexpressing one or the other mouse variant (Figure 2.5C). Uptake measurements were made at 1 min (initial linear range: 0 to 2 min at low concentration and 0 to 1 min at high concentration). BTBR-OATP1A6 has a much higher transport capacity (V_{max} = 673 ± 22 pmol/mg protein x min) than B6-OATP1A6 (V_{max}= 330 ± 18 pmol/mg protein x min). The calculated affinity for TCA is slightly lower for BTBR OATP1A6 ($K_m = 8 \pm 1\mu M$) than for B6 OATP1A6 ($Km = 3.5 \pm 0.8\mu M$). Because position 564 could be functionally important, we mutated this position in the two OATPs to the corresponding amino acid found in the other protein and tested them. As compared to B6-OATP1A6, B6-OATP1A6 with a serine at position 564 showed a 1.2 fold increased TCA uptake (Figure 2.5D). Similarly, as compared to BTBR-OATP1A6, BTBR-OATP1A6, with a proline at position 564, showed a 1.3 fold decreased TCA uptake (Figure 2.5D). These results support the hypothesis that position 564 is crucial for OATP1A6 function and/or expression.

Discussion

Analysis of gene expression variation in a large mouse intercross revealed a *trans*-eQTL hotspot on distal chromosome 6 (Figures 2.1 and 2.2), affecting the mRNA abundance for ~ 8% of the ~ 22,000 annotated genes, many with LOD scores ⩾ 100. This *trans*-eQTL hotspot affected more genes than any other hotspot detected in the six tissues profiled, and was solely observed in pancreatic islets, adding to its unique character. Principal component analysis of the *trans*-eQTL hotspot revealed three distinct clusters of mice (Figure 2.3) that corresponded to the three eQTL genotypes. This allowed us to infer the eQTL genotypes of all mice and, in doing so, convert the multivariate gene expression phenotype to a co-dominant Mendelian trait. With the initial marker genotypes, the eQTL could be localized to a 3.4 Mbp interval. Following two rounds of genotyping with additional markers in recombinant mice (Figure 2.4), we were able to reduce this to a 298 kb interval containing just three genes: *Slco1a5*, *IAPP*, and *Slco1a6*.

Only *Slco1a6* mRNA showed a clearly different expression pattern between B6 and BTBR mice, and it showed significantly higher expression than *Slco1a5* mRNA (Keller et al., 2008), and so we focused on *Slco1a6*. OATP1A6, the protein encoded by mouse *Slco1a6*, was reported to be a kidney-specific organic anion transporter whose mRNA was not detectable for the first three weeks after birth in BALB/c mice and reached adult levels more than six weeks after birth (Choudhuri et al., 2001), but no functional data has been published. Because most OATPs are able to transport bile acids, we first tested whether OATP1A6 was able to mediate uptake of bile acids. Indeed, we demonstrated that both conjugated (TCA) as well as unconjugated bile acids (CA) are substrates for OATP1A6 (Figure 2.5B). In addition, like many other OATPs (Roth et al., 2012), OATP1A6 is a multispecific transporter and can transport not just bile acids but also the hormone conjugates estradiol-17$beta$-glucuronide and estrone-3-sulfate, the dye bromosulfophthalein and the opioid peptide (D-Pen^2, D-Pen^5)-enkephalin (DPDPE). All the tested substrates were consistently better transported by BTBR OATP1A6 as compared to B6 OATP1A6 (Figure 2.5B,C). Among the 6 amino acid changes in OATP1A6, between B6 and BTBR, position 564 is the most likely to be involved in the functional difference between the two OATPs. The functional data with the mutants confirm this suggestion, at least in part, and demonstrate that indeed the substitution of the proline to serine at position 564 in B6 results in a transporter that shows increased TCA uptake (Figure 2.5D), and vice versa; that substitution of serine at position 564 in BTBR to proline decreases TCA transport. Although these single amino acid substitutions do not convert the B6 into the BTBR protein, the functional trend supports the suggested importance of this amino acid position. Additional experiments are required to determine whether these changes affect protein expression levels or turnover numbers.

Abe et al. (2010), using immunohistochemistry, demonstrated that the closely related OATP1A1 was expressed in β-cells and that OATP1A5 was expressed in islets and acinar cells of rat pancreas. In contrast, OATP1A4 was shown to be expressed in Î±-cells (Abe et al., 2010). However, the antibody used to detect rat OATP1A1 in the pancreas was raised against the C-terminal end of the protein, which is identical in OATP1A1 and OATP1A6, and therefore this antibody is not able to distinguish between these two OATPs. Given that obesity-dependent upregulation in BTBR mice was demonstrated for *Slco1a6* but not *Slco1a1* mRNA (Keller et al., 2008) it is unlikely, at least in mouse pancreas, that

OATP1A1 is expressed. The suggested OATP1A1-mediated pravastatin uptake into β-cells to increase insulin secretion (Abe et al., 2010) could also be explained by OATP1A6-mediated pravastatin transport (data not shown).

An OATP1A6 knockout mouse has so far not been reported. However, OATP1A/1B knockout mice where the whole *Slco1a/1b* locus was deleted are viable and have increased plasma levels of unconjugated bile acids as well as bilirubin and bilirubin monoglucuronide levels (van de Steeg et al., 2010) further demonstrating that bile acids are important substrates of OATPs.

While we have not yet tested the effect of bile acids on the transcripts that map to the chromosome 6 *trans*-eQTL hotspot, bile acids can affect gene regulation via farnesoid X receptor (FXR), a nuclear receptor and transcription factor that is activated by bile acids. Düfer et al. (2012) showed that certain bile acids can stimulate insulin secretion via FXR activation.

There has been some controversy about *trans*-eQTL hotspots potentially being artifacts (for example, see (Breitling et al., 2008; Kang et al., 2008)). If multiple transcripts show strongly correlated expression, perhaps induced by some confounding factor, such as a batch effect, and if one such transcript shows chance association to genotype at some location, then many transcripts will share this common, false-positive eQTL. This has led to the development of a number of methods to control for underlying confounding factors (Leek and Storey, 2007; Kang et al., 2008; Stegle et al., 2010; Listgarten et al., 2010; Gagnon-Bartsch and Speed, 2012; Fusi et al., 2012). However, the very strong associations to our chromosome 6 *trans*-eQTL hotspot (with 177 transcripts, from other chromosomes, mapping to the region with LOD $\geqslant$ 100), precludes the possibility that this is an artifact due to confounding: If these associations arose due to some underlying confounder, such as a batch effect, the strength of the association between the confounder and genotype in the region would have to be yet larger.

Our approach to convert the multivariate gene expression phenotype, for transcripts mapping to the chromosome 6 *trans*-eQTL hotspot, to a simple Mendelian trait could be applied more generally. While it was sufficient for us to consider the first two principal components to define the clusters of mice with a common eQTL genotype, one could focus on the non-recombinant mice (whose eQTL genotype is known) and apply linear discriminant analysis or other classification algorithms in order to infer the eQTL genotype in the recombinant mice. A key consideration that must be addressed is the possibility of multiple linked polymorphisms.

Efforts to improve the QTL mapping precision in experimental crosses have focused on increasing the density of recombination events, e.g. with advanced intercross lines (Darvasi and Soller, 1995) and heterogeneous stock (Mott et al., 2000). But the precision of QTL mapping may be hampered more by residual phenotype variation than by lack of recombination events. With a co-dominant Mendelian trait in an intercross, the distribution of the length of the interval defined by recombination events flanking the trait locus follows a gamma distribution with shape=2 and rate=$2n$, where n is the sample size. With 100 intercross mice, the average interval is just 1 cM. This points to the importance of integrating multiple phenotypes that map to a common locus in order to develop composite traits with markedly reduced residual variation.

Acknowledgments

The authors thank Amit Kulkarni for providing annotation information for the gene expression microarrays, and Eric Schadt for providing predicted SNPs based on his high-throughput sequencing of two BTBR mice. This work was supported in part by National Science Foundation grant DMS-12-65203 (to C.K.) and National Institutes of Health grants GM074244 (to K.W.B.), DK066369 (to A.D.A), GM077336 (to B.H.), and GM102756 (to C.K.)

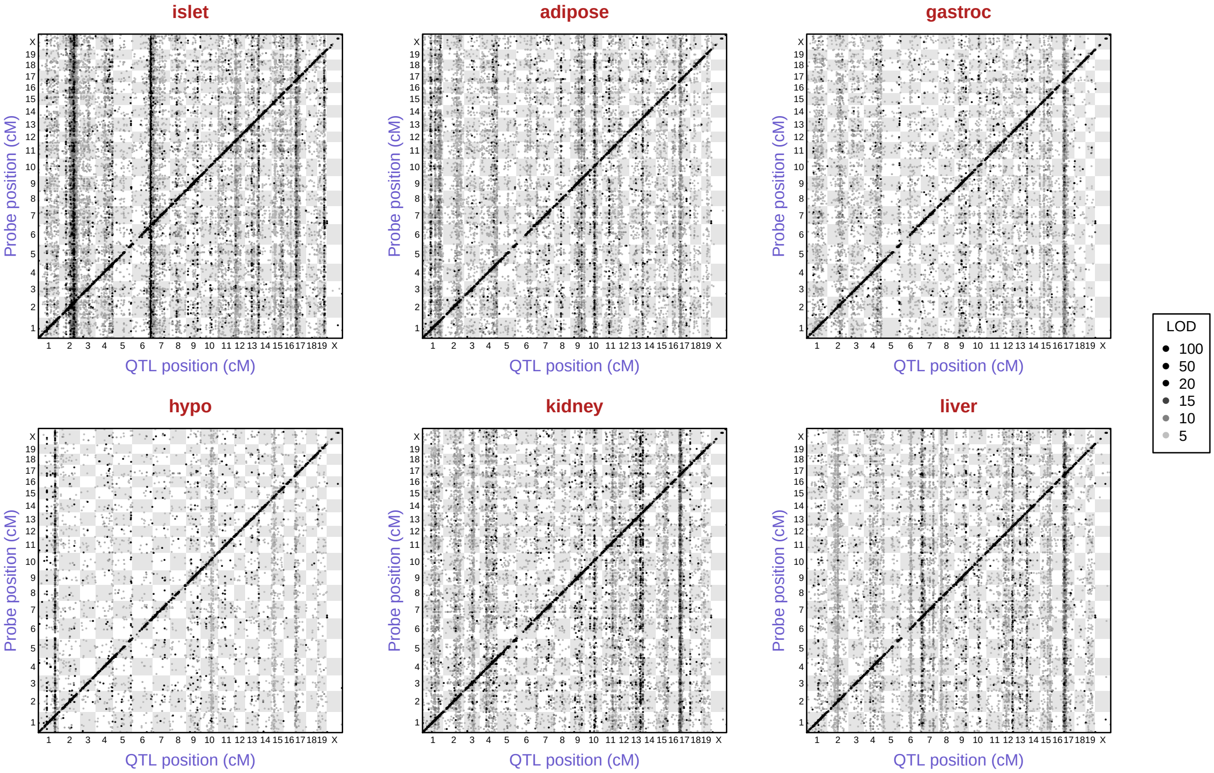


Figure 2.1: Inferred eQTL with LOD ⩾ 5, by tissue. Points correspond to peak LOD scores from single-QTL genome scans with each microarray probe with known genomic position. The y-axis is the position of the probe and the x-axis is the inferred QTL position. Points are shaded according to the corresponding LOD score, though we threshold at 100: all points with LOD ⩾ 100 are black.

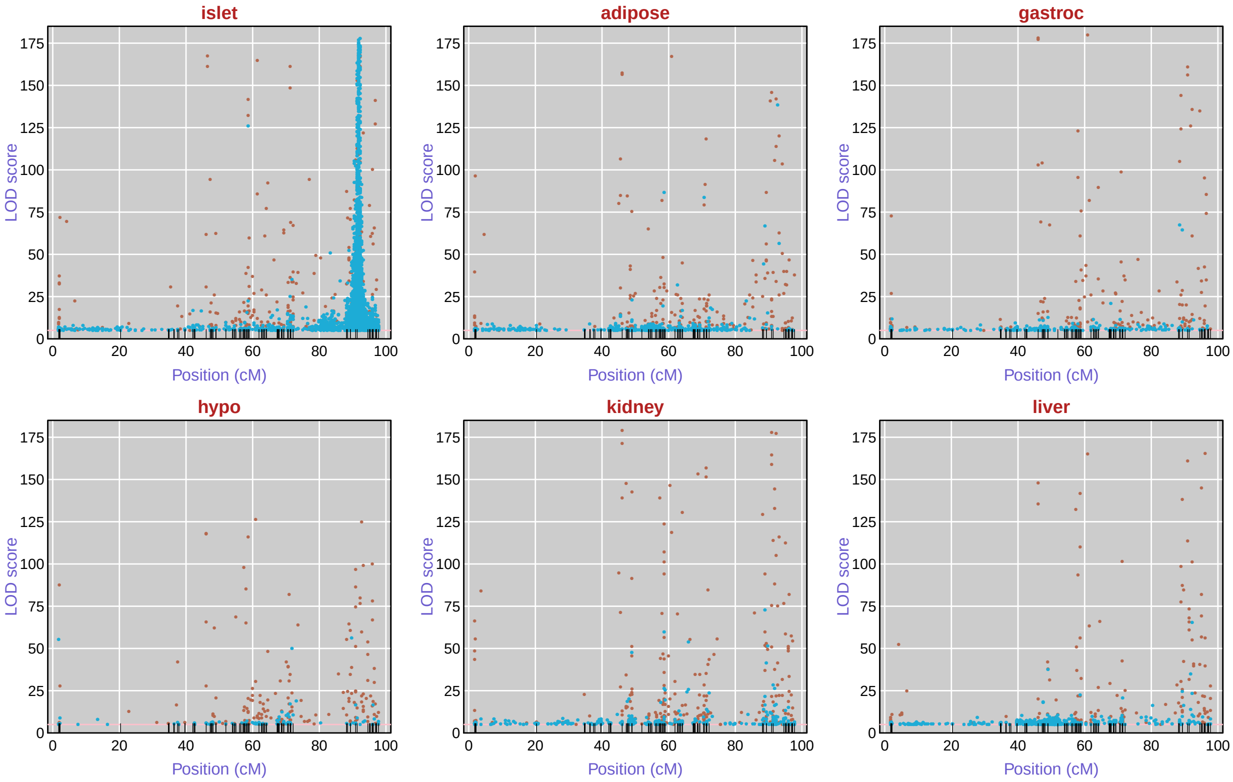


Figure 2.2: Inferred chromosome 6 eQTL with LOD ⩾ 5, by tissue. Each point corresponds to a microarray probe, and indicates the maximum LOD score on chromosome 6 for that probe, and the position of the peak LOD score. Brown points correspond to probes whose genomic location is on chromosome 6; blue points are for probes on other chromosomes.

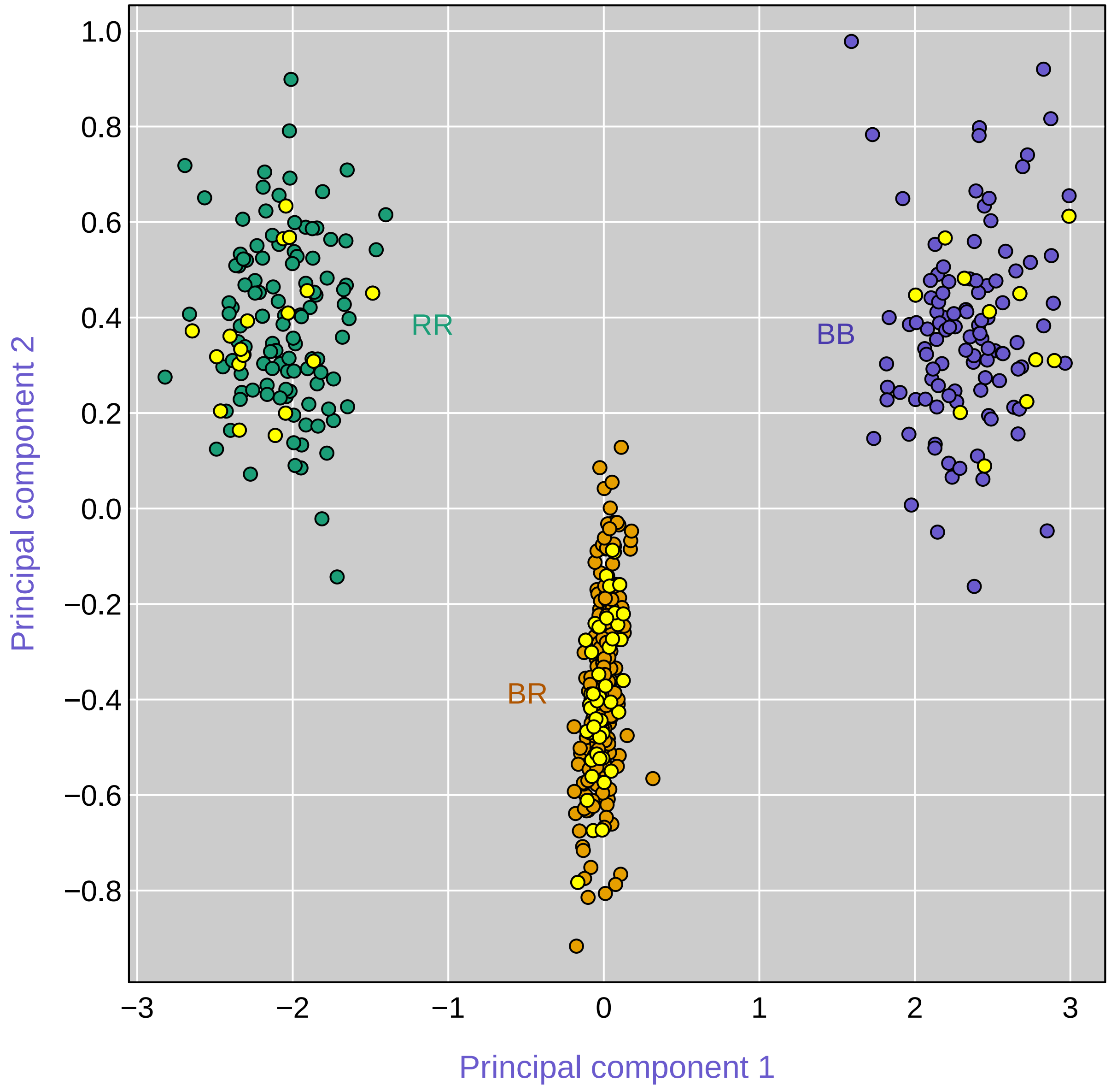


Figure 2.3: The first two principal components for the islet gene expression data for the 181 microarray probes that map to the chromosome 6 *trans*-eQTL hotspot with LOD ⩾ 100 but do not reside on chromosome 6. Each point is a mouse; points for mice without a recombination event in the 10 cM interval centered at the peak marker are colored by their genotype in the region. Yellow points correspond to the 74 mice with recombination events in the interval.

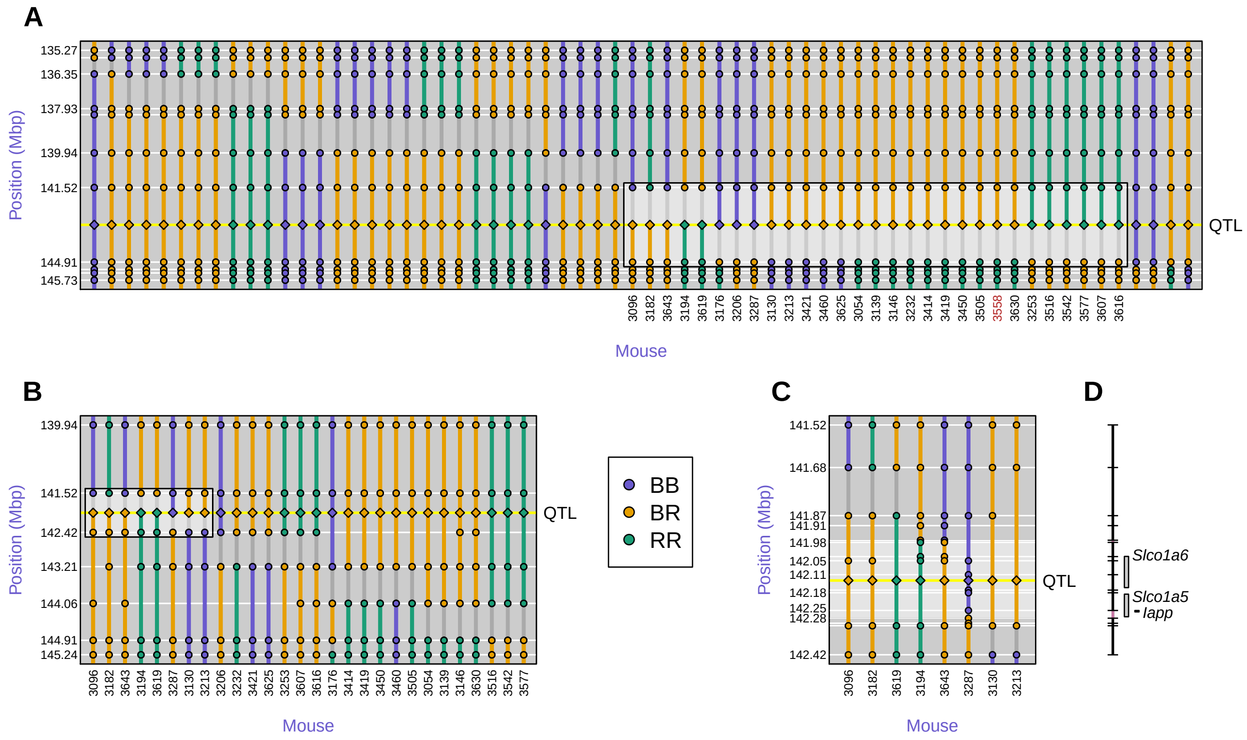


Figure 2.4: Fine-mapping of the islet chromosome 6 eQTL. **A:** Initial SNP genotypes of the 64 mice with recombination events in the 10 Mbp (7 cM) region around the QTL, along with their inferred QTL genotypes (shown at the center of the inferred interval). The highlighted box indicates 29 mice with recombination events in the QTL interval. **B:** Additional genotypes on five markers in the QTL interval, for 28 of the 29 mice with recombination event flanking the QTL. The highlighted box indicates eight mice with recombination event flanking the QTL. **C:** Additional genotypes in the reduced QTL interval, for eight mice with recombination events flanking the QTL. The QTL interval is further reduced to 300 kbp (141.98–142.28 Mbp), a region containing three genes (**D**).

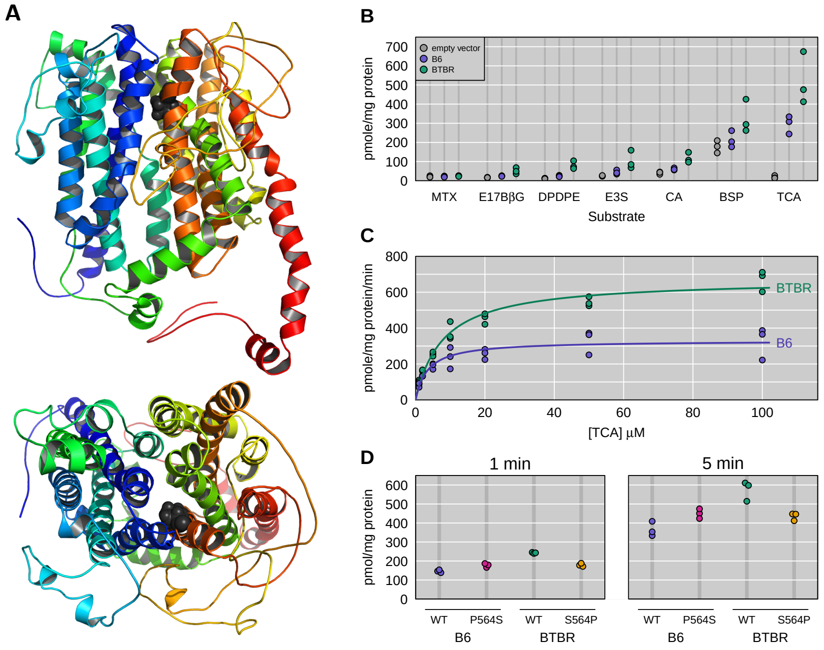


Figure 2.5: Functional characterization of mouse OATP1A6. **A:** Pymol views of a homology model of mouse OATP1A6 with the proline at position 564 highlighted **B:** Mouse OATP1A6-mediated uptake of some common OATP substrates was determined in HEK293 cells transiently transfected with empty vector (EV, gray points), C57BL/6-OATP1A6 (blue points) or BTBR-OATP1A6 (green points). Uptake of 10 ÂµM taurocholic acid (TCA), cholic acid (CA), estrone-3-sulfate (E3S), methotrexate (MTX), (D-Pen^2, D-Pen^5)-enkephalin (DPDPE), estradiol-17-β-glucuronide (E17βG), and bromosulfophthalein (BSP) was measured at 37 °C for 5 min. **C:** Uptake of TCA was measured at various concentrations from 1 to 100 μM at 37 °C for 1 min with empty vector and B6- or BTBR-OATP1A6-expressing HEK293 cells. Net uptake was calculated by subtracting the values of empty vector-transfected cells from either B6- or BTBR-OATP1A6 transfected cells. Resulting data were fitted to the Michaelis-Menten equation to obtain K_m and V_{max} values. **D:** Uptake of 10 μM TCA was measured at 1 and 5 min with normal B6 OATP1A6-Pro^{564} or mutant Ser^{564} and with normal BTBR OATP1A6-Ser^{564} or mutant Pro^{564}. For the uptake experiments, each point shown is derived from an individual experiment performed with triplicate determinations.

REFERENCES

Abe, M., T. Toyohara, A. Ishii, T. Suzuki, N. Noguchi, Y. Akiyama, H. O. Shiwaku, R. Nakagomi-Hagihara, G. Zheng, E. Shibata, et al. 2010. The hmg-coa reductase inhibitor pravastatin stimulates insulin secretion through organic anion transporter polypeptides. *Drug metabolism and pharmacokinetics* 25(3):274–282.

Albert, F. W., and L. Kruglyak. 2015. The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*.

Breitling, R., Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. de Haan, A. I. Su, et al. 2008. Genetical genomics: spotlight on QTL hotspots. *PLoS Genetics* 4(10):e1000232.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296(5568):752–755.

Broman, K. W. 2005. Mapping expression in randomized rodent genomes. *Nat. Genet.* 37(3):209–210.

Broman, K. W., M. P. Keller, A. T. Broman, C. Kendziorski, B. S. Yandell, S. Sen, and A. D. Attie. 2014. Identification and correction of sample mix-ups in expression genetic data: A case study. *arXiv* 1402.2633.

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

Carter, T., and D. Falconer. 1951. Stocks for detecting linkage in the mouse and the theory of their design. *J. Genet.* 50:307–323.

Cheung, V. G., L. K. Conlin, T. M. Weber, M. Arcaro, K.-Y. Jen, M. Morley, and R. S. Spielman. 2003. Natural variation in human gene expression assessed in lymphoblastoid cells. *Nat. Genet.* 33(3):422–425.

Choudhuri, S., K. Ogura, and C. D. Klaassen. 2001. Cloning, expression, and ontogeny of mouse organic anion-transporting polypeptide-5, a kidney-specific organic anion transporter. *Biochemical and biophysical research communications* 280(1):92–98.

Darvasi, A., and M. Soller. 1995. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* 141(3):1199–1207.

Düfer, M., K. Hörth, R. Wagner, B. Schittenhelm, S. Prowald, T. F. Wagner, J. Oberwinkler, R. Lukowski, F. J. Gonzalez, P. Krippeit-Drews, et al. 2012. Bile acids acutely stimulate insulin secretion of mouse β-cells via farnesoid x receptor activation and katp channel inhibition. *Diabetes* 61(6):1479–1489.

Fusi, N., O. Stegle, and N. D. Lawrence. 2012. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comp. Biol.* 8(1):e1002330.

Gagnon-Bartsch, J. A., and T. P. Speed. 2012. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 13(3):539–552.

Hagenbuch, B., and B. Stieger. 2013. The SLCO (former SLC21) superfamily of transporters. *Mol. Aspects Med.* 34(2):396–412.

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):315–324.

Ingalls, A. M., M. M. Dickie, G. Shell, et al. 1950. Obese, a new mutation in the house mouse. *Journal of Heredity* 41:317–318.

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

Kang, H. M., C. Ye, and E. Eskin. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180(4): 1909–1925.

Keller, M. P., Y. Choi, P. Wang, D. B. Davis, M. E. Rabaglia, A. T. Oler, D. S. Stapleton, C. Argmann, K. L. Schueler, S. Edwards, et al. 2008. A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.* 18(5):706–716.

Leek, J. T., and J. D. Storey. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3(9):1724–1735.

Listgarten, J., C. Kadie, E. E. Schadt, and D. Heckerman. 2010. Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci USA* 107(38): 16465–16470.

Meier-Abt, F., Y. Mokrab, and K. Mizuguchi. 2006. Organic anion transporting polypeptides of the OATP/SLCO superfamily: identification of new members in nonmammalian species, comparative modeling and a potential transport mode. *J. Membr. Biol.* 208(3): 213–227.

Morley, M., C. M. Molony, T. M. Weber, J. L. Devlin, K. G. Ewens, R. S. Spielman, and V. G. Cheung. 2004. Genetic analysis of genome-wide variation in human gene expression. *Nature* 430(7001):743–747.

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint. 2000. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci. USA* 97(23): 12649–12654.

Parks, D. J., S. G. Blanchard, R. K. Bledsoe, G. Chandra, T. G. Consler, S. A. Kliewer, J. B. Stimmel, T. M. Willson, A. M. Zavacki, D. D. Moore, et al. 1999. Bile acids: natural ligands for an orphan nuclear receptor. *Science* 284(5418):1365–1368.

R Development Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Roth, M., A. Obaidat, and B. Hagenbuch. 2012. OATPs, OATs and OCTs: the organic anion and cation transporters of the SLCO and SLC22A gene superfamilies. *Br. J. Pharmacol.* 165(5):1260–1287.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422(6929):297–302.

van de Steeg, E., E. Wagenaar, C. M. van der Kruijssen, J. E. Burggraaff, D. R. de Waart, R. P. O. Elferink, K. E. Kenworthy, and A. H. Schinkel. 2010. Organic anion transporting polypeptide 1a/1b–knockout mice provide insights into hepatic handling of bilirubin, bile acids, and drugs. *J. Clin. Invest.* 120(8):2942.

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

Yvert, G., R. B. Brem, J. Whittle, J. M. Akey, E. Foss, E. N. Smith, R. Mackelprang, and L. Kruglyak. 2003. Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nat. Genet.* 35(1):57–64.

The dissection of expression quantitative trait locus hotspots

SUPPLEMENT

Jianan Tian[*], Mark P. Keller[†], Aimee Teo Broman[‡], Christina Kendziorski[‡], Brian S. Yandell[*,§], Alan D. Attie[†] Karl W. Broman[‡,1]

Departments of [*]Statistics, [†]Biochemistry, [‡]Biostatistics and Medical Informatics, and [§]Horticulture,

University of Wisconsin–Madison, Madison, Wisconsin 53706

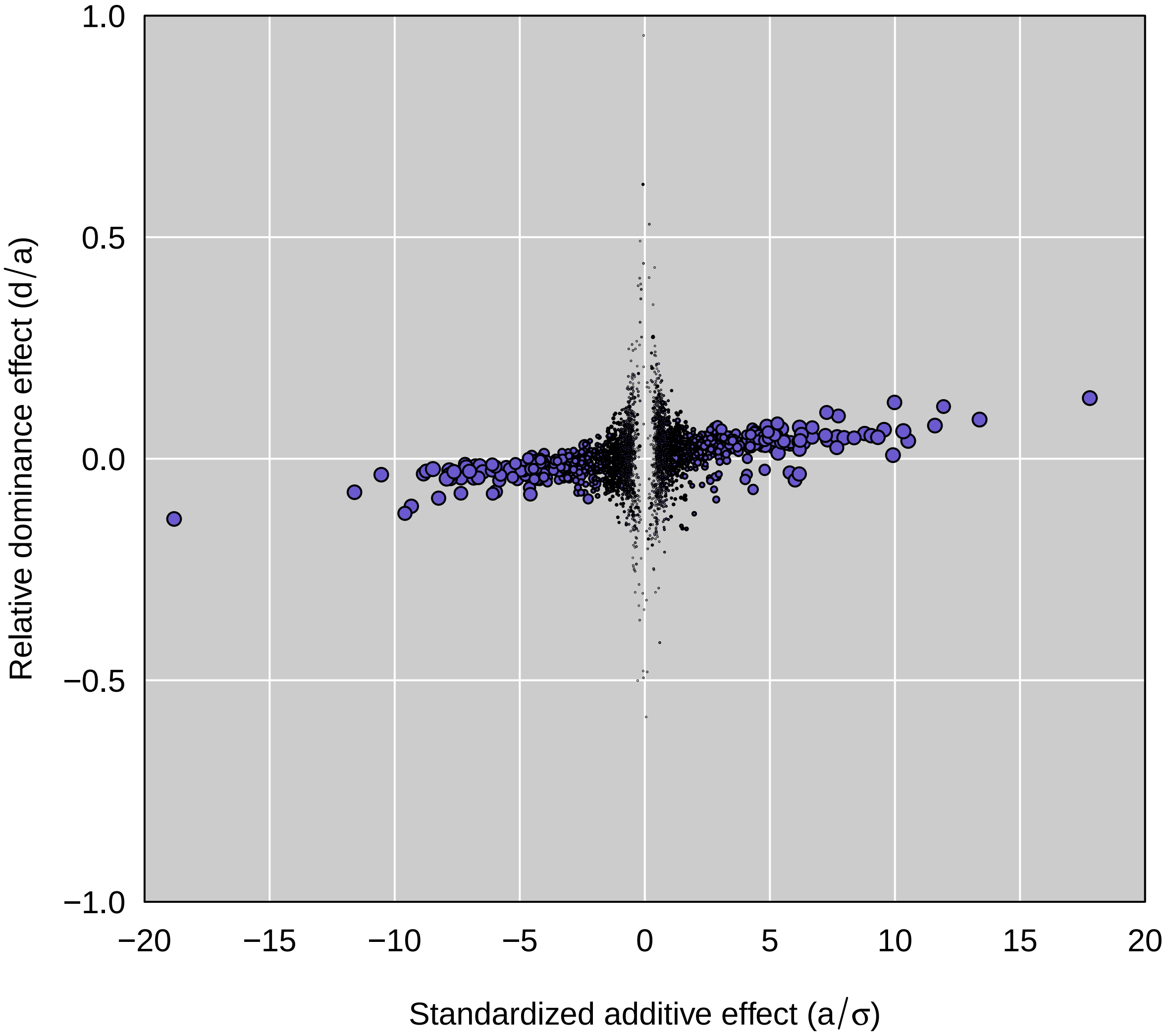


Figure 2.6: Estimated allelic effects of the distal chromosome 6 eQTL on gene expression in pancreatic islets, for the 2,889 microarray probes that map with LOD $\geqslant 5$ to the 10 cM interval centered at the peak marker. The additive effect, a, is defined as half the difference between the phenotype averages for the two homozygotes, with positive values corresponding to BTBR homozygotes having larger phenotype than the B6 homozygotes. The dominance effect, d, is the difference between the average phenotype for the heterozygotes and the midpoint between the two homozgotes. σ is the residual standard deviation. Plotted is the relative dominance effect (d/a) versus the standardized additive effect (a/σ), with points sized according to the LOD score.

Table 2.1: Sequences differences in the *Slco1a6* gene, between the B6 and BTBR mouse strains.

	B6		BTBR	
Position	**Nucleotide**	**Amino acid**	**Nucleotide**	**Amino acid**
14	G	Glycine (G)	A	Glutamic acid (E)
354	A	—	G	—
720	C	—	A	—
1063	G	Valine (V)	A	Isoleucine (I)
1074	C	—	T	—
1366	A	Asparagine (N)	G	Aspartic acid (D)
1690	C	Proline (P)	T	Serine (S)
1851	C	Phenylalanine (F)	A	Leucine (L)
1972	T	Serine (S)	C	Proline (P)

File S1 Single nucleotide polymorphisms (SNPs) between the B6 and BTBR mouse strains that were used to narrow the chromosome 6 eQTL interval.

	Gene symbol	Chr	bp (build37)	SNP ID	flanking sequence	B6 ref	BTBR	Forward primer	Reverse primer	SNP type	AA change
1	Gm5724	chr6	141684769	rs37215304	GGAGCACACTTATAAGGAGATAAAGCATATAAATTGGATTGCAGGAGAGGYTCTTCATTAATGGAAAGAAATCTGAAACAAAAGCATGTAGAGATCACCA	T	C	CCATATTGCTGTTCCAGGTACT	GAACTTTAGCCACTAGGGATCAT	non synonymous coding	N311S
2	Slco1a1	chr6	141874184	N/A	GTAGTTCTTTCTTTGGCAGTGTCTTTGGAAAGAAGAAAAAGGGGATGCTGRTCAGGATATTCACTCCTGCACAGACCAAAAAGCCAATCCACCAAGCACC	G	A	AGTGACAGATGCCCACTAAAC	GACCACAAGTTCAAGACACAAAC	non synonymous coding	T259I
3		chr6	141913339	rs36413998	GCATTATACTTGTTCGTTATCTGTAATAAAATTCACATTGTGGAACGAGCRGAAGGTTTATGTTTTTGAGTCAGTAAAAGTTTGCCCTCCCAAACAATAG	G	A	GGGAAACTAAGGTTGTTGGTAGA	AGAGGAGAGAAGAACAGTAGGG	intergenic	N/A
4		chr6	141970950	N/A	GAAGCCACTGGGAAAGAAAATCATTGAGTAGTGATGGGTAGCTTTTGCCAYTGTGCACTTAATTTCAGAGGGAAGAAAAGGACTGAGGAAAACCATTCTT	T	C	CAAGACAGTAGCAGGTGTTAGG	TCTGAGAGACCCGTGATACAA	intergenic	N/A

5	chr6	141970951	N/A	AAGCCACTGGGAAAGAAAATCATTGAGTAGTGATGGGTAGCTTTTGCCATKGTGCACTTAATTTCAGAGGGAAGAAAAGGACTGAGGAAAACCATTCTTA	T	G	CAAGACAGTAGCAGGTGTTAGG	TCTGAGAGACCCGTGATACAA	intergenic	N/A
6	chr6	141970986	N/A	GGTAGCTTTTGCCATTGTGCACTTAATTTCAGAGGGAAGAAAAGGACTGARGAAAACCATTCTTAATTCCAGTTGTCACCCGCCCTGCACCACCTTGGGT	G	A	CAAGACAGTAGCAGGTGTTAGG	TCTGAGAGACCCGTGATACAA	intergenic	N/A
7	chr6	141971021	N/A	GAAGAAAAGGACTGAGGAAAACCATTCTTAATTCCAGTTGTCACCCGCCCYGCACCACCTTGGGTGTGAACTCAGAGGACAGTCCCAAGTTCCCCAGCGA	T	C	CAAGACAGTAGCAGGTGTTAGG	TCTGAGAGACCCGTGATACAA	intergenic	N/A
8	chr6	141979254	rs52543135	GTGTTTGTGTATGTGTTTGGGTGTTTGAAGATGTACATATGTGCAGAACARAACCACATCGGGTATTGGTCCTCACCTTCCATTTGATTGAGGCAGGGTA	A	G	ACTTGGCTGTGACTCCATTAG	ATGCAAGAGGAGATTGGAAAGA	intergenic	N/A

9	Slco1a6	chr6	142034947	rs51801967	GAAACCTCATTACAGCTTAGTTTTCAGTTCTCCATCGTTCTCGACCTTAGRACTTTTGTGCATGTCTGTGCACTCACTTTCCTTCTCTGTGGGCTTCGTC	A	G	GTGCTACCCTGAAGTTCTGAAA	CAATAGGCGCCTCTATCTTGG	non synonymous coding	S658P
10	Slco1a6	chr6	142051511	rs13462749	AGGGGAAGTACTACCAAGCAAATAATGAAAGTACATACCAATGAGGAAGAYTGCCTTTGCTGTGGAGATTCCAAAATGATGTTCCAGGTATTTAGGCTTG	C	T	GACGACTTGTTTCCCTGTCAT	GACTTGATGGTCTCAACACTCC	non synonymous coding	V355I
11	Slco1a6	chr6	142106246	rs32147628	TGTTCTCTTTTTCTGCATTCTCAAGCACAACACAGGATGCAATTGTCAGARAAGAGTGAAGAGTGACCTATGGGGTTGTTTTTCTCCTTTGTACAATATT	G	A	TGCACTCTGAACACACAGAATAG	GGCAGGTAGTTGGAAGTGTTT	downstream	N/A
12		chr6	142166806	rs32153144	CTAAAACCTTTAATTATGACTCATCAGTAGTACTTGGTGTAATTAGATGASGTTATACATTTATTACTAACCAGGCCATAATTATGAGTTAGTGGAAATT	G	C	GGATCCTTTGGGAGAAGAGTAAG	GGAAAGGAAAGCAGCAACTATG	intergenic	N/A

13		chr6	142177340	rs32152292	AAAGGTAAATTCAATGTGTACATGCAACACATTGAAGAAGAAACTCATAGRGAAGAACCTGCTGTCCTCACATTGGCTGAGAAGTAAGGAGAGATTGGTT	G	A	GACTAACAGACTGGCTACACAA	CATTAAGAGGTGGTGCCTAGAA	intergenic	N/A
14	Slco1a5	chr6	142246802	N/A	CGCCATCACCCACACAAAGGCACTCAGTATGTGGTCTACTAATATTTACCYAAGTGTTAATGTGTTGATGGCCCACCTGCTCCATCTGCCAATGAAACCC	T	C	GCCCTTGGCTATGACTTCTC	CACAGGGTATCCTTTACCTTTCA	intronic	N/A
15		chr6	142277172	rs32160582	AAAAAAAAGGAAACAAATTGTATGGGTAGAGGAGGAGAGATGGTAGAGATRAATCTGGGAAGAGTTAGAGAAGACGGTGAATATGATCAAAACATGAAAC	G	A	TCCACAAGATGAAGCCCATAC	GGACGTGAAGTTGGGAAAGA	intergenic	N/A
16	Pyroxd1	chr6	142297143	N/A	GCTGGAAAGTCGCTTTCCCAACATTAAGGTCATCGAATCTGGAGTAAAGCRGTTGAAAAGTGAAGACCATGTAAGACTTCTTCCTGTTAAATGATATGTT	A	G	GACCGCCAGACCTTTAACTT	CATCCCAGCAGGTGACTTATC	non synonymous coding	Q86R

17	Pyroxd1	chr6	142306955	rs32163119	TCTTTTCTTTAAAGAAATGTGGCCTGTGTATGTGGAACTGACCAATGGGAMAATATACGGCTGTGATTTCCTTGTCAGCGCTACAGGAGTCACCCCAAAC	C	A	TCGCCGTTCCTGTAACATTC	GCCATCTTGGGTTACAGCTAATAy	non synonymous coding	T304K
18		chr6	142421188	rs32181060	CATCACCACTGGAGTCCCCTATGTCTCTGGCCCACCATGGCCAGTGTTGCRAGTAGTACAAACCTTTGTTGGTCACCTCCCAAGAAACTTCAAAAAG	A	G	CCTATTCAGCCACTCACTAC	CCACACAGCTGGAAGAAGAA	intergenic	N/A
19		chr6	143213027	rs36537203	TTTAAAATTTAAATTAAGTTTACTTTAATGGATATGCAAACACAGAATACYTGCACCTTCTGTCAAAGCTGAGGCTGTTGGGCTGACAAATGGTTGTTGGC	T	C	ACGCAGACACGACAAACA	CAGACAAAGACAGAGAGACAGAC	intergenic	N/A
20		chr6	143213530	rs36821985	GTCACAGCGTACAGAAAGTAACTTCTGGAGTTACCTAGTCCTTGTGATAAYTGTTGGGTCCAGGTACTGTTTTCAATGCTATTCATATAGCAATTCATTTA	C	T	CTAAGCCCAGGAAGACAGAAG	TGACAGGAGCTCAGAACATAAC	intergenic	N/A

21	chr6	144058889	rs30707584	CAGTATGGCAATAGTAATGAAGGAGATTAAGCTTAGTTTTTACAAAAAATWTTTTGAATATTTAGTATGACTTTCTGAACTTTAATGAGTATAATATTGGAGTAAAGTTAA	T	A	ATCACCCACGTCAGCAGTAT	AGGAGGAGGAGGAGGAGAAA	intergenic	N/A

3 DISSECTION OF TRANS-EQTL HOTSPOTS

The dissection of expression quantitative trait locus hotspots

Jianan Tian[*], Mark P. Keller[†], Aimee Teo Broman[‡],
Christina Kendziorski[‡], Brian S. Yandell[*,§], Alan D. Attie[†],
Karl W. Broman[‡]

Departments of [*]Statistics, [†]Biochemistry, [‡]Biostatistics and Medical Informatics, and
[§]Horticulture, University of Wisconsin–Madison, Madison, Wisconsin 53706

Abstract

Studies of the genetic loci that contribute to variation in gene expression frequently identify loci with broad effect on gene expression: expression quantitative trait locus (eQTL) hotspots. We describe a set of exploratory graphical methods as well as a formal likelihood-based test for assessing whether a given hotspot is due to one or multiple polymorphisms. We first look at the pattern of effects of the locus on the expression traits that map to the locus: the direction of the effects, as well as the degree of dominance. A second technique is to focus on the individuals that exhibit no recombination event in the region, apply dimensionality reduction (such as with linear discriminant analysis) and compare the phenotype distribution in the non-recombinants to that in the recombinant individuals: If the recombinant individuals display a different expression pattern than the non-recombinants, this indicates the presence of multiple causal polymorphisms. In the formal likelihood-based test, we compare a two-locus model, with each expression trait affected by one or the other locus, to a single-locus model. We apply our methods to a large mouse intercross with gene expression microarray data on six tissues.

Introduction

There is a long history of efforts to map the genetic loci (called quantitative trait loci, QTL) that contribute to variation in quantitative traits in experimental organisms, particularly to learn about the etiology of disease (Broman, 2001; Jansen, 2007). But it remains difficult to identify the genes underlying QTL (Nadeau and Frankel, 2000). There has recently been much interest in measuring gene expression in disease-relevant tissues in QTL experiments, as a way to speed the process from QTL to gene (Jansen and Nap, 2001; Albert and Kruglyak, 2015). The genetic control of gene expression is itself of great interest.

Expression quantitative trait loci (eQTL) analysis tries to find the genomic locations that influence variation in gene expression levels (mRNA abundances). eQTL near the genomic location of the influenced gene are called local eQTL, and eQTL far away from the influenced gene are called *trans*-eQTL. When a genomic region influences the expression of many genes, the region is called a *trans*-eQTL hotspot.

eQTL hotspots have been observed in many genetic studies (e.g., Brem et al., 2002; Yvert et al., 2003; Schadt et al., 2003; Chesler et al., 2005), and they are of particular interest because gene expressions mapping to the same location may indicate the existence of a genetic regulator.

Batch effects (artifacts arising from technical or environmental factors) are common in microarray experiments. This has led to the development of a number of methods to control for underlying confounding factors (Leek and Storey, 2007; Kang et al., 2008; Stegle et al., 2010; Listgarten et al., 2010; Gagnon-Bartsch and Speed, 2012; Fusi et al., 2012). However, these methods generally cannot distinguish *trans*-eQTL hotspots from batch effects. There is some controversy about whether *trans*-eQTL hotspots are themselves artifacts and whether one should control for them, as one does for batch effects, in eQTL analysis (e.g., Breitling et al., 2008; Kang et al., 2008). But in many cases, the associations between genotype and expression phenotypes are extremely strong (with LOD $>$ 100), which largely precludes the possibility of a batch effect artifact, as the strength of association between batch and genotype in the region would have to be even stronger.

In Tian et al. (2015), we considered a large mouse intercross between the strains C57BL/6J (abbreviated B6) and BTBR T$^+$ *tf*/J (abbreviated BTBR), with gene expression microarray data on six tissues (adipose, gastrocnemius muscle, hypothalamus, pancreatic islets, kidney, and liver), and mapped a *trans*-eQTL hotspot to a 298 kb region containing just three genes. This sort of fine-mapping approach is meaningful only if the hotspot is due to polymorphisms in a single gene.

This raises an important question about *trans*-eQTL hotspots: are they the result of polymorphisms in a single gene, or are there multiple underlying genes? In other words, is there complete pleiotropy, or are there multiple linked eQTL? Methods for testing pleiotropy versus tight linkage of multiple QTL (Jiang and Zeng, 1995; Knott and Haley, 2000) do not scale well to the case of the very large number of expression traits that map to a *trans*-eQTL hotspot. We developed a likelihood-based test that is a variation on the method of Knott and Haley (2000), as well as a number of exploratory data visualizations, to test whether multiple eQTL underlie a hotspot. We apply our approaches to the data considered in Tian et al. (2015).

Methods

We focus on the case of an intercross between two inbred strains, B and R (these labels were chosen to match the strains used in the application, below). We assume dense marker genotype data and genome-wide gene expression phenotype data (such as from microarrays or RNA-sequencing). We first perform a genome scan to identify quantitative trait loci (QTL), considering each expression trait individually. We use Haley-Knott regression (Haley and Knott, 1992) for this purpose, for the sake of speed. For each expression trait and each chromosome, we consider the location of the single-largest LOD score, provided that it exceeds a significance threshold that adjusts for the genome scan but not the search across expression traits.

We count the number of expression traits that show a *trans*-eQTL within a sliding window (e.g., of 10 cM) across the genome, and use peaks in these counts to define *trans*-eQTL hotspots. We then focus on one such hotspot, and on the set of expression traits that map to an interval centered at the peak count.

We then ask: Are the multiple expression traits that map to this *trans*-eQTL hotspot all affected by a common eQTL, or could there be multiple causal polymorphisms in the region? We have developed a set of exploratory data visualizations to address this question, as well as a formal likelihood-based test. We prefer to exclude expression traits whose genomic position is near (or even on the same chromosome as) the hotspot of interest, as these may be driven by separate, local-eQTL.

Exploratory data visualizations

We first consider the pattern of effects of the locus on the expression traits that map to the region: the direction of the effects, as well as the degree of dominance. We use a pair of data visualizations: first, a plot of the signed LOD score for each trait (with the sign taken from the estimated additive effect) versus the estimated eQTL location, for each expression trait analyzed separately. That is, for each expression trait, we find the largest LOD score on the chromosome, multiply it by ± 1, according to the sign of the estimated additive effect of the locus, and plot this signed LOD score versus the location at which that maximum LOD score was attained. If there are two nearby loci with effects in opposite directions, they may be revealed by this plot.

In addition, we plot the estimated dominance effect against the estimated additive effect, for each expression trait. Let B and R denote the two alleles in the cross, and let $\hat{\mu}_{BB}$, $\hat{\mu}_{BR}$ and $\hat{\mu}_{RR}$ denote the average expression levels for genotypes BB, BR, and RR, respectively. We estimate the additive effect as half the difference between the two homozygotes, $\hat{a} = (\hat{\mu}_{RR} - \hat{\mu}_{BB})/2$, and the dominance effect as the difference between the heterozygote and the midpoint between the two homozygotes, $\hat{d} = \hat{\mu}_{BR} - (\hat{\mu}_{BB} + \hat{\mu}_{RR})/2$. We then plot $\hat{d}$ vs $\hat{a}$ for all expression traits mapping to the hotspot. If there are two nearby loci with different inheritance patterns (e.g., one has additive allele effects and the other has an allele that is dominant), they may be revealed by this plot.

As a second technique, we consider the individuals that have no recombination event in the region. For these individuals, we know their eQTL genotype. We apply linear discriminant analysis (LDA, Hastie et al., 2009, Ch. 4) to the top 100 traits with the largest

LOD scores and make a scatterplot of the first and second linear discriminants; this should show three distinct clusters (or, for a fully dominant locus, two clusters). We calculate the linear discriminants for individuals that show a recombination event in the region, and add them as points to the plot. If the recombinant individuals fall within the clusters defined by the non-recombinants, this is consistent with there being a single causal locus. If, on the other hand, the recombinants look distinctly different from the non-recombinants, then multiple polymorphisms are indicated.

The basic idea underlying this visualization is that the non-recombinants can be used to derive an estimate of the conditional distribution of the multivariate expression phenotype, given the eQTL genotype. We use LDA as a dimension-reduction technique. The goal of the visualization is to compare the expression pattern in the recombinants and non-recombinants. If there is a single eQTL, the recombinants should look no different from the non-recombinants; if there is a difference, we can conclude that there are multiple eQTL.

Formal statistical test

To formally assess evidence of multiple linked loci versus complete pleiotropy at a *trans*-eQTL hotspot, we developed a likelihood-based test to compare the null hypothesis of a single eQTL affecting all expression traits to the hypothesis of two eQTL, with each expression trait affected by one or the other eQTL (but not both). The approach can handle only a limited number of expression traits, and so we focus on the 50 traits with largest LOD scores (when considered individually) in the interval centered at the hotspot.

We assume that the traits follow a multivariate normal distribution, conditional on eQTL genotype, and apply the multivariate QTL analysis method of Knott and Haley (2000). For a given QTL model, we have $Y = X\beta + E$, where Y is an $n \times p$ matrix of phenotypes, with n as the number of F_2 individuals and p as the number of traits, X is an $n \times q$ matrix of covariates (including additive covariates, interactive covariates, genotype probabilities for the position under investigation, and the interactive covariates times the genotype probabilities), and β is a $q \times p$ matrix of coefficients. We obtain $\hat{\beta} = (X'X)^{-1}X'Y$, calculate the matrix of estimated residuals $\hat{E} = Y - X\hat{\beta}$, and calculate the residual sum of squares matrix $\text{RSS} = \hat{E}'\hat{E}$. The LOD score is then $\frac{n}{2}\log_{10}\{|\text{RSS}_0|/|\text{RSS}|\}$, where $|\text{RSS}|$ denotes the determinant of the RSS matrix, and RSS_0 is the residual sum of squares matrix for the null model (with additive and interactive covariates, but no genotype probabilities).

We perform a QTL scan over the interval; at each putative QTL location, denoted λ, we calculate the LOD score, $\text{LOD}_1(\lambda)$, comparing this single-QTL model to the null model of no QTL. Let $M_1 = \max_\lambda \text{LOD}_1(\lambda)$.

We compare this to a two-QTL model, in which each QTL is affected by one or the other QTL but not both. In principle, one would need to consider, with p expression traits, 2^{p-1} possible assignments of the expression traits to the left and right QTL. This is a prohibitively large number, and so we make an approximation: we sort the expression traits according to their estimated QTL location, when considered individually, and we consider only the $p - 1$ cut-points of this list. We randomly order any expression traits that map to the same position. For each cut-point, we perform a two-dimensional scan over possible two-QTL models, and calculate $\text{LOD}_2^{(c)}(\lambda_1, \lambda_2)$, comparing the two-QTL model, with QTL at positions λ_1 and λ_2, and with the first c expression traits affected by the QTL at λ_1, and the last $p - c$ traits affected by the QTL at λ_2. Let $M_2^{(c)} = \max_{\lambda_1, \lambda_2} \text{LOD}_2^{(c)}(\lambda_1, \lambda_2)$, and let $M_2 = \max_c M_2^{(c)}$. The estimated cut-point is $\hat{c} = \arg\max_c M_2^{(c)}$, and the estimated QTL positions are $(\hat{\lambda}_1, \hat{\lambda}_2) = \arg\max_{\lambda_1, \lambda_2} \text{LOD}_2^{(\hat{c})}(\lambda_1, \lambda_2)$.

As evidence for the presence of two QTL, we consider the $\log_{10}$ likelihood ratio, $\text{LOD}_{2v1} = M_2 - M_1$.

The exhaustive two-dimensional scan of $\text{LOD}_2^{(c)}(\lambda_1, \lambda_2)$ is computationally intensive. We can accelerate this calculation by iteratively searching for the maximum on each of the two dimensions. There is no guarantee that this will converge to the overall maximum, especially when there are multiple modes in the two-QTL likelihood surface, but in practice we've found this algorithm works well. As a starting point of this iterative search, we can use either the estimated QTL location under the single-QTL model, or a randomly selected position.

Statistical significance: To assess the statistical significance of the result, we need an approximation of the distribution of the test statistic under the null hypothesis of a single QTL. We consider two approaches: a parametric bootstrap, and a stratified permutation test.

In the parametric bootstrap, we simulate new phenotype data using the estimated single-QTL model. In the stratified permutation test, we randomly permute the rows in the phenotype data relative to the rows in the genotype data, within each QTL genotype group. When there are unmeasured genotypes at the inferred QTL, we infer the QTL genotype, for each individual, to be that with maximum probability, conditional on the observed marker data. These conditional QTL genotype probabilities are calculated by a hidden Markov model (Broman and Sen, 2009, App. D).

For each procedure, we generate 1000 data sets, perform the full likelihood analysis (the scan for the single-QTL model, and the two-dimensional scan for the two-QTL model, for each possible partition of the traits) and calculate the test statistic. The P-value for the test is taken to be the proportion of simulated or permuted data sets with a test statistic that is greater or equal to the observed test statistic.

Visualizations: To visualize the results for the two-QTL model, we plot profile LOD curves for the left and right QTL, using the estimated cut-point, $\hat{c}$, for the expression traits into those mapping to the left QTL and those mapping to the right QTL. For the left QTL, we plot the slice $\text{LOD}_2^{(\hat{c})}(\lambda_1, \hat{\lambda}_2)$ against λ_1, for varying values of λ_1. Similarly, for the right QTL, we plot $\text{LOD}_2^{(\hat{c})}(\hat{\lambda}_1, \lambda_2)$ against λ_2, for varying values of λ_2. These sorts of profile LOD curves follow from an innovation of Zeng et al. (2000).

As a further diagnostic plot, we consider the statistic $\text{LOD}_{2v1}^{(c)} = M_2^{(c)} - M_1$, as a function of the cut-point, c. This is evidence for two vs one QTL, for a given cut-point, of the expression traits, into those that map to the left QTL and those that map to the right QTL. This displays the evidence for two vs one QTL as well as the evidence for a particular split of the expression traits.

Application

To illustrate our methods for the dissection of *trans*-eQTL hotspots, we consider a large mouse F_2 intercross (Tian et al., 2015) with gene expression microarray data on six tissues. The data are available at the QTL Archive, now part of the Mouse Phenome Database: http://phenome.jax.org/db/q?rtn=projects/projdet&reqprojid=532

Materials

The experiment was carried out in order to identify genes and pathways that contribute to obesity-induced type II diabetes. Greater than 500 offspring were generated from an F_2 intercross between diabetes-resistant (C57BL/6J, abbreviated B6) and diabetes-susceptible (BTBR T^+ *tf* /J, abbreviated BTBR) mouse strains. All mice were genetically obese through introgression of the leptin mutation ($Lep^{\mathrm{ob}/\mathrm{ob}}$) and were sacrificed at 10 weeks of age, the age when essentially all BTBR ob/ob mice are diabetic.

Mice were genotyped with the 5K GeneChip (Affymetrix). After data cleaning, there were 519 F_2 mice genotyped at 2,057 informative markers. Gene expression was assayed with custom two-color ink-jet microarrays manufactured by Agilent Technologies (Palo Alto, CA). Six tissues from each F_2 mouse were used for expression profiling: adipose, gastrocnemius muscle (abbreviated gastroc), hypothalamus (abbreviated hypo), pancreatic islets (abbreviated islet), kidney, and liver. Tissue-specific mRNA pools were used for the reference channel, and gene expression was quantified as the ratio of the mean $\log_{10}$ intensity (mlratio). For further details, see Keller et al. (2008). In the final data set, there were 519 mice with gene expression data on at least one tissue (487 for adipose, 490 for gastroc, 369 for hypo, 491 for islet, 474 for kidney, and 483 for liver). The microarray included 40,572 total probes; we focused on the 37,797 probes with known location on one of the autosomes or the X chromosome.

QTL analysis

For QTL analysis, we first transformed the gene expression measures for each microarray probe in each of the six tissues to normal quantiles, taking $\Phi^{-1}[(R_i - 0.5)/n]$, where Φ is the cumulative distribution function for the standard normal distribution and R_i is the rank in $\{1, \dots, n\}$ for mouse i. We then performed single-QTL genome scans, separately for each probe in each tissue, by Haley-Knott regression (Haley and Knott, 1992) with microarray batch as an additive covariate and with sex as an interactive covariate (i.e. allowing the effects of QTL to be different in the two sexes). Calculations were performed at the genetic markers and at a set of pseudomarkers inserted into marker intervals, selected so that adjacent positions were separated by $\leqslant 0.5$ cM. We calculated conditional genotype probabilities, given observed multipoint marker genotype data, using a hidden Markov model assuming a genotyping error rate of 0.2%, and with genetic distances converted to recombination fractions with the Carter-Falconer map function (Carter and Falconer, 1951). Calculations were performed with R/qtl (Broman et al., 2003), an add-on package to the general statistical software R (R Core Team, 2015).

For each probe in each tissue, we focused on the single largest LOD score peak on each chromosome, and on LOD score peaks $\geqslant 5$ (corresponding to genome-wide significance at the 5% level, for a single probe in a single tissue, determined by computer simulations under the null hypothesis of no QTL).

The inferred eQTL for all genes with LOD $\geqslant 5$ are displayed in Figure 3.1, with the y-axis corresponding to the genomic position of the microarray probe and the x-axis corresponding to the estimated eQTL position. As expected, we see a large number of local-eQTL along the diagonal for each tissue-specific panel. These local-eQTL correspond to genes for which expression or mRNA abundance is strongly associated with genotype near their genomic position.

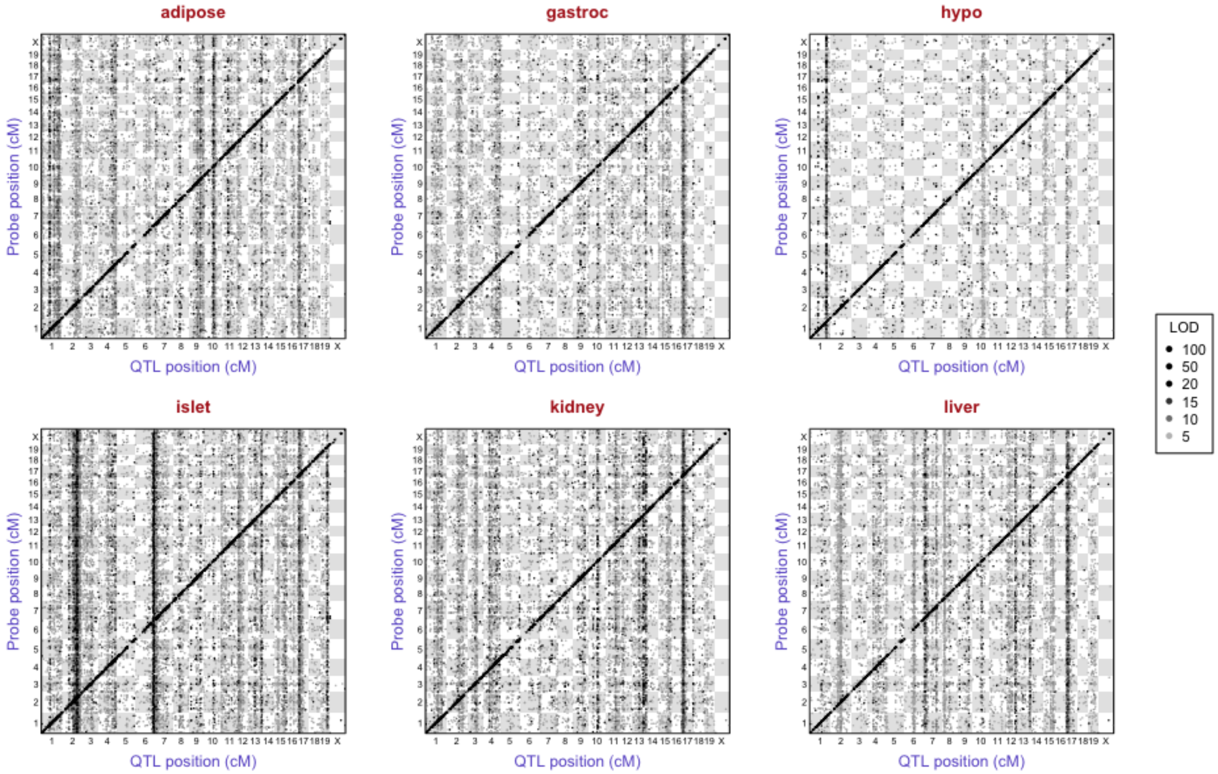


Figure 3.1: Inferred eQTL with LOD $\geqslant 5$, by tissue. Points correspond to peak LOD scores from single-QTL genome scans with each microarray probe with known genomic position. The y-axis is the position of the probe and the x-axis is the inferred QTL position. Points are shaded according to the corresponding LOD score, though we threshold at 100: all points with LOD $\geqslant 100$ are black. (A version of this figure appeared as Figure 1 in Tian et al., 2015)

In addition to the local-eQTL, there are a number of prominent vertical bands: genomic loci that influence the expression of genes located throughout the genome. These are the *trans*-eQTL hotspots. Overall, we detected many more *trans*-eQTL than local-eQTL. The *trans*-eQTL hotspots can show either remarkable tissue specificity or be observed in multiple tissues. For example, a locus near the centromere of chromosome 17, at 11.7 cM, shows effects in all tissues. In contrast, the *trans*-eQTL hotspot located at the distal end of chromosome 6 was only observed in pancreatic islets.

To define *trans*-eQTL hotspots of potential interest, we focused on a more conservative threshold for eQTL, of LOD $\geqslant 10$. We further excluded local-eQTL, defined here to be those for which the distance between the gene's genomic position and its inferred eQTL position was < 10 cM. We then counted the number of expression traits with a *trans*-eQTL in a sliding interval of length 10 cM (Figure S1).

For each *trans*-eQTL of interest, we widened the interval to be considered beyond that initial 10 cM window, to consider the interval in which the count of expression traits with eQTL was > 50, and then padded this further by adding 5 cM on either end.

We will focus on a set of six hotspots: adipose chromosome 1 at 39 cM, adipose chromosome 10 at 48 cM, islet chromosome 2 at 75 cM, islet chromosome 6 at 91 cM, kidney chromosome 13 at 68 cM, and liver chromosome 17 at 18 cM. Results for additional hotspots are displayed in File S1.

Visualization of QTL effects

We first consider the estimated effects of a locus on the expression traits that map to the region (Figure 3.2). In the left panels, we display the signed LOD score (with positive values indicating that the BTBR allele is associated with larger average expression and negative values indicating that the B6 allele is associated with larger average expression) versus the estimated eQTL location. In the right panels, we plot the estimated dominance effect versus the estimated additive effect for all transcripts mapping to the hotspot. The key value, in these visualizations, is for the case that two linked QTL show distinct inheritance patterns.

The islet chromosome 6 hotspot, at 92 cM, shows approximately equal numbers of expression traits for which the BTBR allele causes an increase or decrease in gene expression (Figure 3.2A), and the allele effects are approximately additive (Figure 3.2B), with estimated dominance effect near 0. These results are consistent with there being a single QTL. In Tian et al. (2015), this locus was resolved to a 298 kb interval containing just three genes, with good evidence for *Slco1a6* as the causal gene.

The kidney chromosome 13 hotspot, at 68 cM, shows clear evidence for two QTL. In Figure 3.2C, we see that for expression traits mapping to ~57 cM, the BTBR allele is associated with a decrease in expression, while for traits mapping to ~68 cM, the BTBR alleles is predominantly associated with an increase in expression, though with some traits having effects in the opposite direction. From Figure 3.2D, we can infer that, for the traits mapping to ~57 cM, the B6 allele is nearly dominant ($d \approx -a$, along the line with slope –1), while for the traits mapping to ~68 cM, the BTBR allele is dominant ($d \approx a$, along the line with slope +1).

The islet chromosome 2 hotspot, at 75 cM, shows expression traits with high LOD scores across a broad region (Figure 3.2E). For traits mapping to 70–75 cM, the B6 allele is associated with increased expression, while for traits mapping to 55–60 cM, the effect is in the opposite

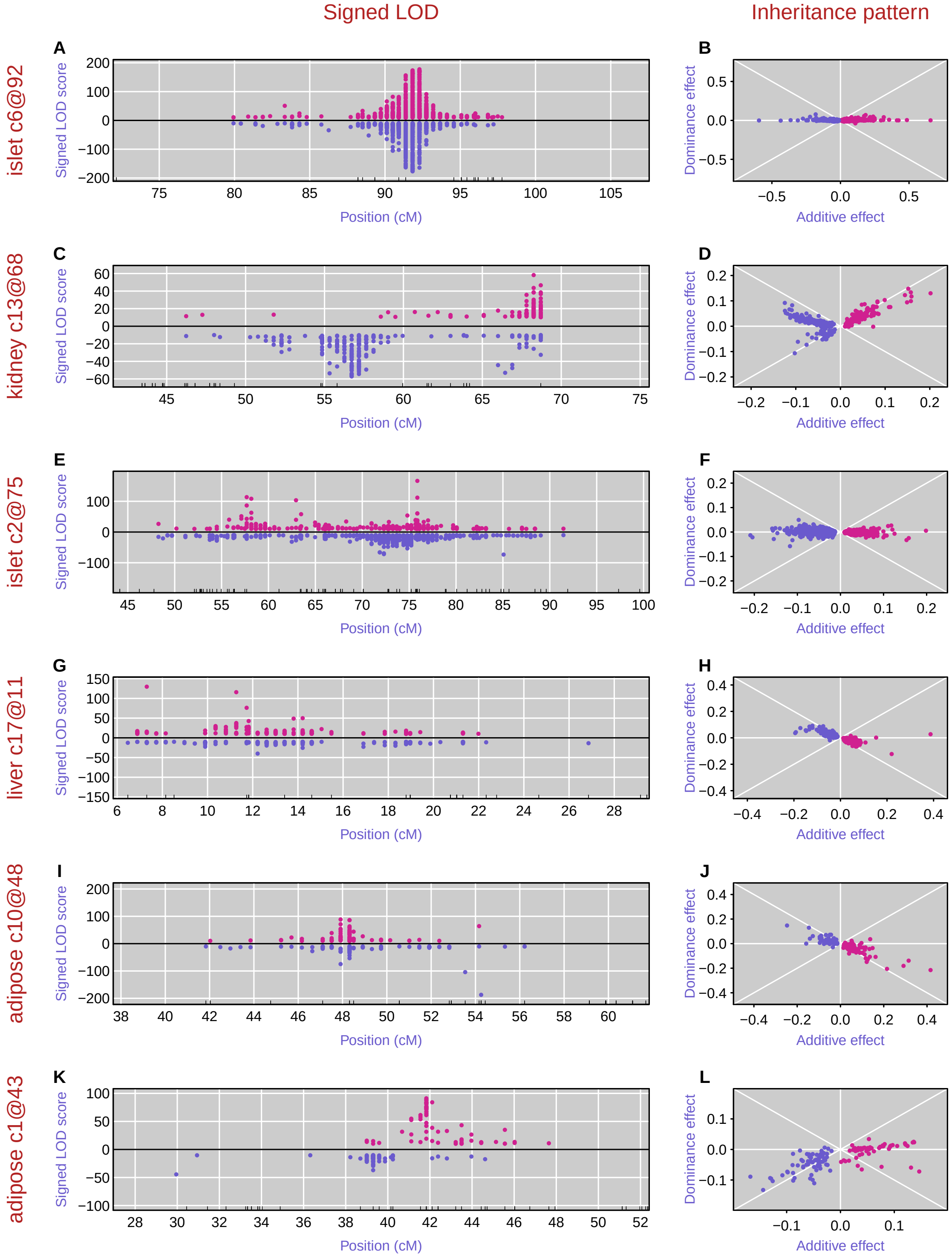


Figure 3.2: Visualizations of the QTL effects on the multiple expression traits that map with LOD $\geqslant 10$ to a *trans*-eQTL hotspot. Each row is a hotspot. The left panels are scatterplots of signed LOD scores (with positive values indicating that the BTBR allele is associated with larger average gene expression and negative values indicating that the B6 allele is associated with larger average gene expression) versus the estimated QTL location. Each point is a single expression trait. Tick marks at the bottom indicate the locations of the genetic markers. The right panels are scatterplots of the estimated dominance effects versus the estimated additive effects.

direction. The allele effects are nearly additive for all expression traits (Figure 3.2F).

The liver chromosome 17 hotspot, at 11 cM, has approximately equal numbers of traits with effects in each direction (Fig 3.2G), and the B6 alleles appears to be nearly dominant in most cases (Figure 3.2H). The adipose chromosome 10 hotspot, at 48 cM, is similar, with effects in both directions (Figure 3.2I) and with the B6 allele being nearly dominant (Figure 3.2J).

The adipose chromosome 1 hotspot, at 43 cM, again shows evidence for two QTL. For the expression traits mapping to 38–40 cM, the B6 allele is associated with increased expression (Figure 3.2K), but the BTBR allele appears dominant (Figure 3.2L). For traits mapping to 42-46 cM, however, the BTBR allele is associated with increased expression and the allele effects appear additive.

In summary, for two out of these six hotspots, these visualizations of the estimated QTL effects provide good evidence for two QTL. In one case (kidney chromosome 13), the two QTL are well-separated, but in the other case (adipose chromosome 1), the two loci are tightly linked.

Comparison of recombinants and non-recombinants

Our second graphical technique is to consider the individuals exhibiting no recombination event in the region of a *trans*-eQTL hotspot (for these individuals, we know their eQTL genotype), apply linear discriminant analysis (LDA) using the top 100 expression traits that map to the region, and make a scatterplot of the first two linear discriminants. Superposing points for the recombinant individuals, we can make a direct comparison of the recombinants and non-recombinants (Figure 3.3). If there is a single eQTL in the region, the recombinant individuals should reside within the clusters defined by the non-recombinant individuals. If the recombinant individuals appear different from the non-recombinants, this indicates the presence of a second QTL.

For the islet chromosome 6 hotspot, the non-recombinant mice form three distinct clusters, and the recombinant mice (in yellow) fit reasonably well within those clusters (Figure 3.3A). This is consistent with there being a single eQTL.

For the islet chromosome 2 (Figure 3.3C) and adipose chromosome 10 (Figure 3.3E) hotspots, the non-recombinant mice again form tight clusters, but the recombinant mice fall clearly outside those clusters. This is evidence for the presence of more than one eQTL.

In the other three cases, kidney chromosome 13 (Figure 3.3B), liver chromosome 17 (Figure 3.3D), and adipose chromosome 1 (Figure 3.3F), the clusters of non-recombinants are not so tight, and the recombinants are not obviously different from the recombinants. However, for the liver chromosome 17 hotspot (Figure 3.3D), one might make the case that the majority of recombinants are at the boundaries between the clusters, and so multiple eQTL may be indicated.

Returning to the adipose chromosome 10 hotspot (Figure 3.3E), note how the recombinant mice form tight clusters that are distinct from the non-recombinants. If there are two eQTL in the region, perhaps these clusters correspond to different two-locus recombinant genotypes? Through the fit of a two-QTL model (in the next section), we estimate the two QTL to be at 48 and 54 cM. If we color the points by the two-locus genotypes for these two positions, we see that the clusters of non-recombinants do share a common two-locus genotype (Figure 3.4A).

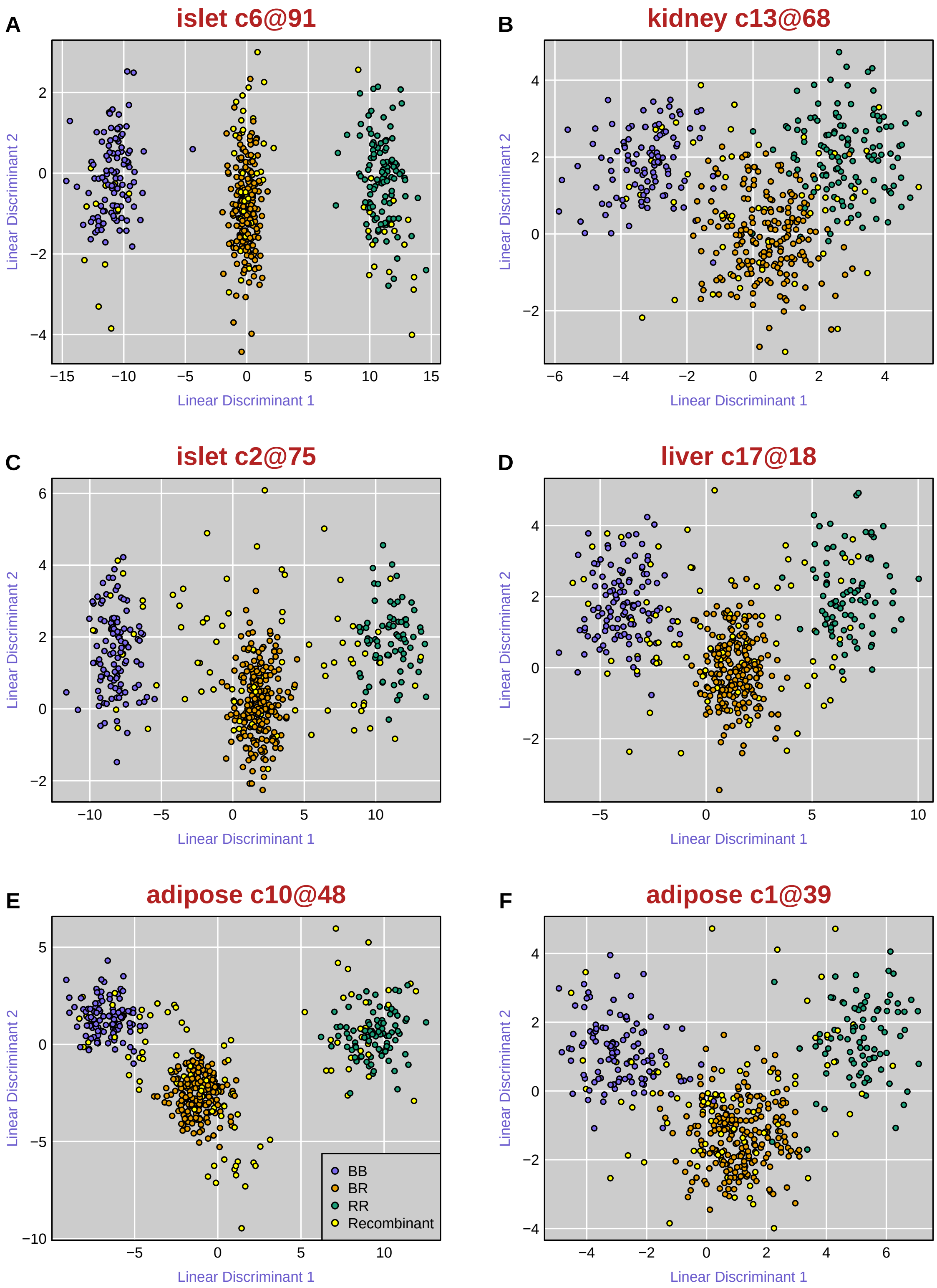


Figure 3.3: Scatterplots of the first two linear discriminants from application of linear discriminant analysis to the 100 expression traits that map to the region with the highest LOD score, with mice that show no recombination event in the region of a *trans*-eQTL hotspot. Blue, orange, and green points correspond to the non-recombinant mice with genotype BB, BR, and RR, respectively, at the eQTL. Yellow points correspond to recombinant mice.

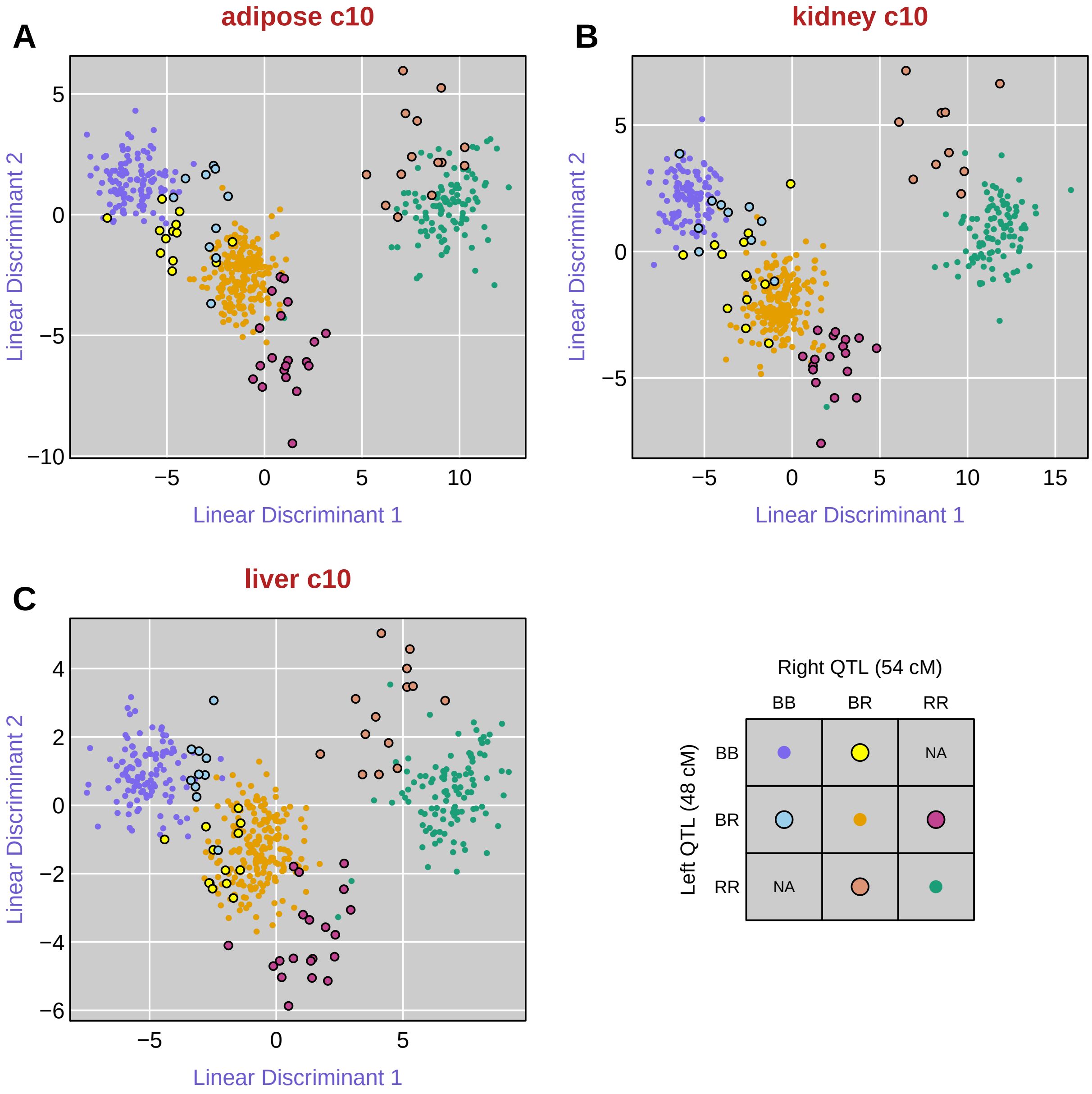


Figure 3.4: Scatterplots of the first two linear discriminants, as in Figure 3.3E, for the *trans*-eQTL hotspot on chromosome 10, here considering three tissues: adipose, kidney, and liver. Points correspond to mice, and they are colored according to their two-locus genotypes, for the inferred two QTL model, with one locus at ~48 cM and the other at ~54 cM.

This chromosome 10 hotspot also shows effect in kidney and liver, and so we applied this technique for this same region, with expression data for these tissues (Figures 3.4B and 3.4C). The three tissues give consistent results. Mice that are heterozygous at one QTL and homozygous BB at the other (yellow and light blue) sit between the non-recombinants that are homozygous BB (dark blue) and those that are heterozygous (orange). Mice that are homozygous RR at the left QTL and heterozygous at the right QTL (brown) sit above the non-recombinant RR mice (green), while mice that are heterozygous at the left QTL and homozygous RR at the right QTL (red) sit below the non-recombinant heterozygotes (orange).

There is one green point (non-recombinant RR) sitting among the red points (BR at the left QTL and RR at the right QTL). This mouse (with ID 3117) has a recombination event just to the left of the left QTL; if we moved that QTL slightly to the left, it would become a red point (BR at the left QTL and RR at the right QTL). In principle, a series of graphs of this form, with varying locations for the left and right QTL, could be used to define the QTL intervals in the context of this two-QTL model.

There is one additional green point among the red points in Figure 3.4C (liver). This mouse (with ID 3317) sits at the center of the cluster of green points in each of Figures 3.4A and 3.4B, and shows no recombination event in the region of these two QTL.

This example illustrates that consideration of the two-locus genotypes can help to strengthen evidence for two loci underlying a *trans*-eQTL hotspot. However, as we will describe below, in this particular case, the right eQTL appears to affect just three of the expression traits. Just one single trait, if affected by a separate locus, can have a great deal of leverage on these sorts of plots.

Formal tests for two QTL

To supplement these visualization techniques, we developed a formal statistical test for whether a *trans*-eQTL hotspot harbors one vs two eQTL. The results of this approach, for the six eQTL under consideration, are displayed in Figure 3.5.

Let's begin by considering the kidney chromosome 13 hotspot (Figures 3.5C and 3.5D). For these multivariate likelihood analyses, we focus on the top 50 expression traits mapping to the region, in terms of their LOD scores when considered individually. In the left panel (Figure 3.5C), the black curve is the LOD curve for the multivariate QTL analysis with a single-QTL model. The estimated QTL location is at 67.4 cM. The blue and pink curves are LOD profiles for the estimated two-QTL model, for which the estimated QTL locations are at 54.8 and 67.8 cM. The blue and pinks points indicate the LOD scores and estimated QTL locations for the individual expression traits, with blue points affected by the left QTL and pink points affected by the right QTL. The right panel (Figure 3.5D) shows the evidence for two versus one QTL as a function of the choice of cut-point for the list of expression traits, into those affected by the left and right QTL. The inferred cut-point has 40 traits affected by the left QTL and 10 traits affected by the right QTL, and has a LOD_{2v1} score of 45.8, indicating very strong evidence for two QTL, and for this particular cut-point.

The results for the islet chromosome 6 hotspot are displayed in Figures 3.5A and 3.5B. The inferred two-QTL model has QTL at 91.4 and 91.8 cM, with only the two expression traits affected by the left QTL. And $\text{LOD}_{2v1} = 2.3$, indicating weak evidence for two QTL.

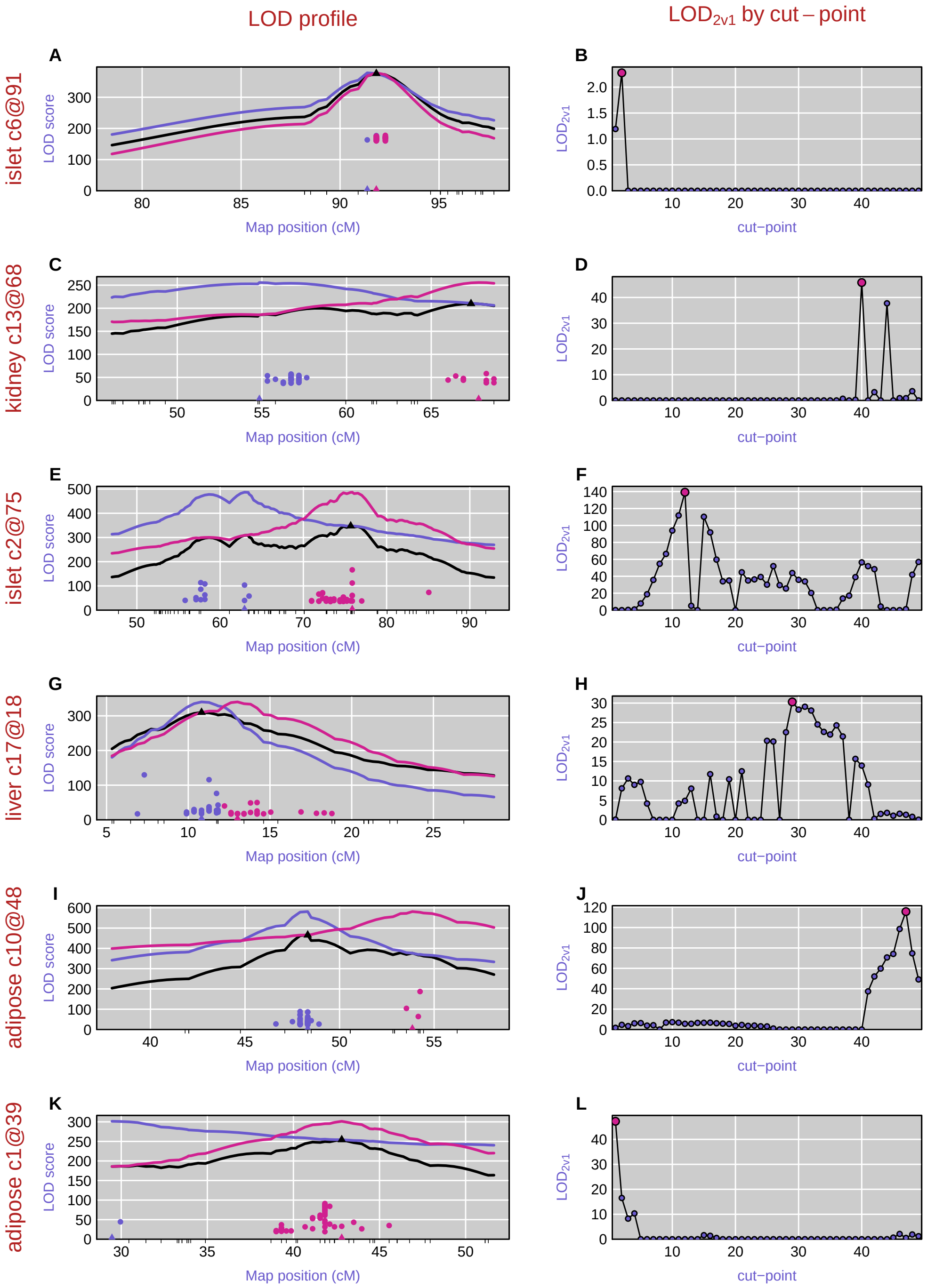


Figure 3.5: Results of a test of one versus two QTL at a *trans*-eQTL hotspot, considering the top 50 traits, in terms of LOD score, that map to the region. Each row is a hotspot. In the left panels, the black curve is the LOD curve for the single-QTL model, with estimated QTL location indicated by a black triangle. The blue and pink curves are profile LOD curves for the left and right QTL, respectively, for the estimated two-QTL model(with the estimated cut-point). Points indicate the LOD score and estimated QTL position for the 50 expression traits, analyzed separately. The points are colored according to whether they are estimated to be affected by the left QTL (blue) or right QTL (pink). The right panels show the $\text{LOD}_{2v1}^{(c)}$ score, indicating evidence for two versus one QTL, for each possible cut-point, c, of the list of expression traits into those that map to the left and those that map to the right QTL.

The results for the islet chromosome 2 hotspot (Figures 3.5E and 3.5F) indicate strong evidence for two QTL, with $\mathrm{LOD}_{2v1} = 139$. The estimated QTL are at 62.9 and 75.7 cM. The left QTL is inferred to affect 12 of the 50 expression traits.

The liver chromosome 17 hotspot (Figures 3.5G and 3.5H) has strong evidence for two QTL, with $\mathrm{LOD}_{2v1} = 30$, and the estimated QTL locations at 10.8 and 13.0 cM. The choice of cut-point of the expression traits is not so clear. We estimate 29 of the expression traits are affected by the left eQTL, but a model with 32 traits affected by the left eQTL gives a similar likelihood.

For the adipose chromosome 10 (Figures 3.5I and 3.5J) and the adipose chromosome 1 (Figures 3.5K and 3.5L) hotspots, the evidence for two QTL is strong, but only three eQTL are inferred to be affected by the right eQTL at the chromosome 10 hotspot, and only 1 trait is inferred to be affected by the left eQTL at the chromosome 1 hotspot. If we trim off these expression traits and apply the procedure again, we find that, for the adipose chromosome 10 locus (see Figure S2), there is little evidence for more than one eQTL affecting the remaining traits. Further, the analysis of two-locus genotypes in the LDA plot in Figure 3.4 is largely driven by these three expression traits that map to 54 cM. Similarly, if we trim off the first expression trait for the adipose chromosome 1 hotspot (Figure S3), there is limited evidence for multiple QTL affecting the remaining traits.

In summary, the formal statistical test provides strong evidence for two eQTL in five of these six cases, but in two of the cases, the majority of the traits are affected by a single eQTL.

Simulations

To further assess the performance of the proposed likelihood-based test for whether a *trans*-eQTL hotspot harbors more than one eQTL, we performed a set of computer simulation studies.

We generated 500 intercross offspring with 100 markers on a 100 cM chromosome, and then simulated $p = 10$ or 40 traits, with half of the traits affected by a QTL at 50 cM and the other traits affected by a QTL 0 – 20 cM away (at 50 – 70 cM). We also considered an unbalanced case with 5 traits affected by the left QTL and 35 traits affected by the right QTL. We assumed additive allele effects, with the additive effect of each QTL being $\alpha =$ 0.1, 0.2, 0.3, 0.4, or 0.5. Residual variation followed a normal distribution with mean 0 and standard deviation 1, with traits conditionally independent given the QTL genotypes.

We used 100 simulation replicates for each situation and calculated P-values by a parametric bootstrap with 1000 simulation replicates.

The estimated power to detect two linked QTL, as a function of the distance between the QTL, is shown in Figure 3.6. When the QTL effect is smaller than 0.3, the power to distinguish two QTL within distance of 10 cM is low for $p = 10$. For any fixed effect, the power to detect two QTL is higher for $p = 40$ than for $p = 10$. When the QTL effect is larger than 0.4, the power to distinguish two QTL separated by more than 5 cM is almost 100%.

The power to detect two QTL in the unbalanced case, with the left QTL affecting 5/40 traits, (Figure 3.6C) is considerably lower than for the balanced case (Figure 3.6B).

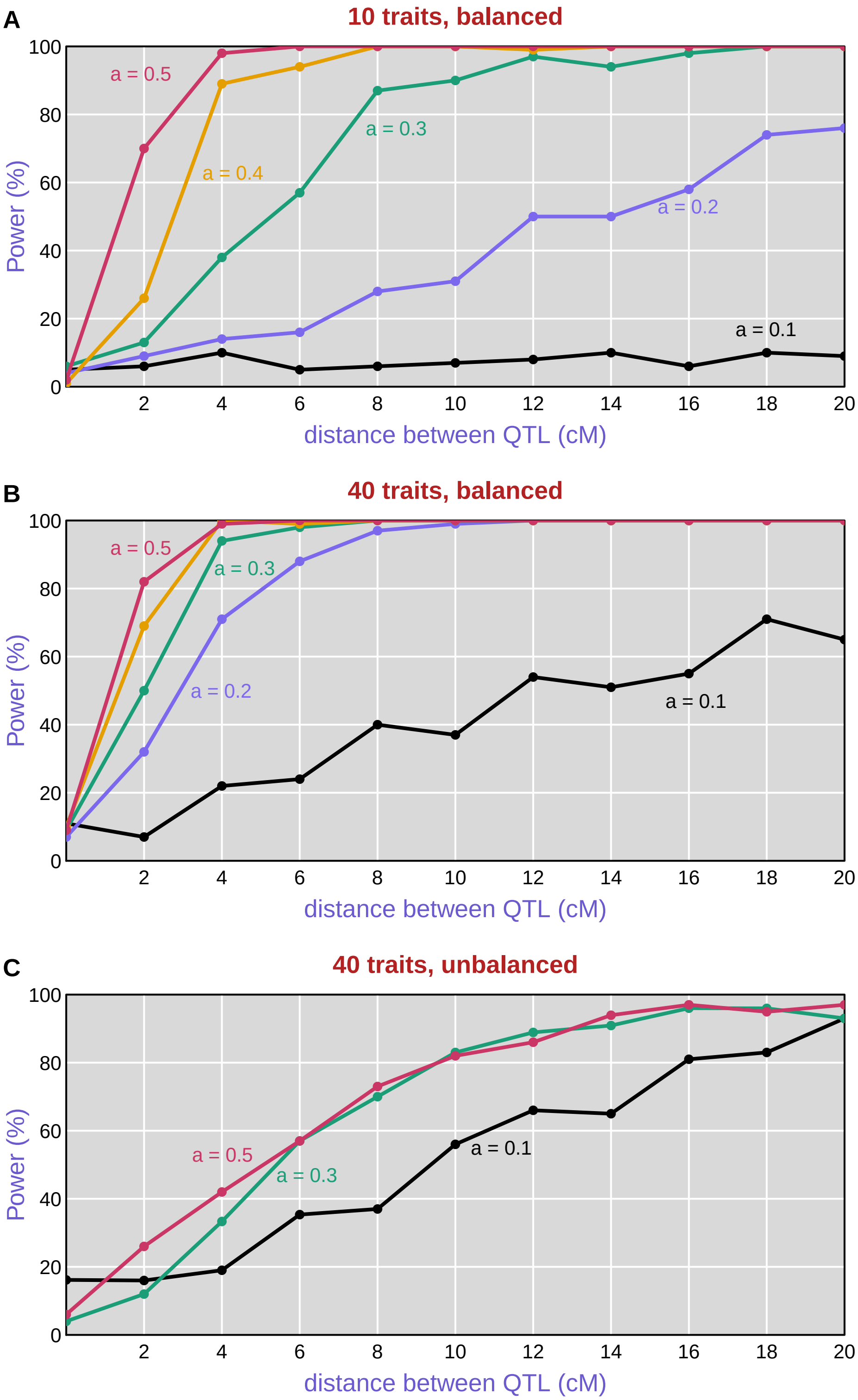


Figure 3.6: Power curve for various QTL effects as the distance between two QTL grows. p=10 and p=40

Discussion

In this paper, we have proposed exploratory methods and a formal inference method for dissecting *trans*-eQTL hotspots. We applied these approaches to data on a large mouse intercross with gene expression microarray data on six tissues, and we performed a simulation study to investigate the performance of the formal inference method. Both the exploratory methods and the formal inference method are helpful in dissecting *trans*-eQTL hotspots, and can give improved estimates of the eQTL positions.

The exploratory methods have the advantage of providing insight into the underlying evidence: that the multiple eQTL show distinct inheritance patterns, or that the recombinant and non-recombinant individuals show differences in expression. However, while the visualization methods can be strongly informative, they will not necessarily reveal the presence of two eQTL, as the inheritance pattern of the two linked eQTL may be the same, or the first two linear discriminants may not be revealing of the difference between the recombinants and non-recombinants.

In forming a multivariate test statistic, we chose to follow the method of Knott and Haley (2000), but other multivariate analysis of variance (MANOVA) statistics could also be used, including Pillai's trace, Lawley-Hotelling's test, and Roy's Lambda (Anderson, 2003). Similarly, in the exploratory data visualization based on linear discriminant analysis, other dimension-reduction techniques could be used. A supervised (i.e., classification) method, that makes use of the known eQTL genotypes of the non-recombinant individuals, is preferred.

The main issue in the formal statistical test is the choice of expression traits, as we can't handle a very large number of expression traits. Regularized methods (see Hastie et al., 2009, Sec 5.8), or a Bayesian approach, might have an advantage in this context. Such approaches could also be used to relax some of our modeling assumptions. For example, one might consider a model with two eQTL, where each expression trait can be affected by both.

We considered a single tissue at a time. The joint consideration of multiple tissues could provide additional power to dissect *trans*-eQTL hotspots that are in common across tissues.

We ignored the effects of eQTL elsewhere in the genome and considered just one region in isolation. In doing so, the effects of any other eQTL become part of the residual variation. Local-eQTL are a particularly important case, as they are quite common and often have large effect. Controlling for the effect of local-eQTL could give better precision in the dissection of a *trans*-eQTL hotspot.

We have implemented our methods in an R package (R Core Team, 2015), *qtlpvl*, available at https://github.com/jianan/qtlpvl.

Acknowledgments

The authors thank Amit Kulkarni for providing annotation information for the gene expression microarrays. This work was supported in part by National Science Foundation grant DMS-12-65203 (to C.K.) and National Institutes of Health grants GM074244 (to K.W.B.), DK066369 (to A.D.A), and GM102756 (to C.K.).

REFERENCES

Albert, F. W., and L. Kruglyak. 2015. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* 16:197–212.

Anderson, T. W. 2003. Testing the general linear hypothesis; multivariate analysis of variance. In *An introduction to multivariate statistical analysis*, 3rd ed., chap. 8, 291–380. Wiley.

Breitling, R., Y. Li, B. M. Tesson, J. Fu, C. Wu, T. Wiltshire, A. Gerrits, L. V. Bystrykh, G. de Haan, A. I. Su, et al. 2008. Genetical genomics: spotlight on QTL hotspots. *PLoS Genetics* 4(10):e1000232.

Brem, R. B., G. Yvert, R. Clinton, and L. Kruglyak. 2002. Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752–755.

Broman, K. W. 2001. Review of statistical methods for QTL mapping in experimental crosses. *Lab Animal* 30(7):44–52.

Broman, K. W., and S. Sen. 2009. *A guide to QTL mapping with R/qtl*. New York: Springer.

Broman, K. W., H. Wu, S. Sen, and G. A. Churchill. 2003. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* 19:889–890.

Carter, T. C., and D. S. Falconer. 1951. Stocks for detecting linkage in the mouse, and the theory of their design. *J. Genet.* 50:307–323.

Chesler, E. J., L. Lu, S. Shou, Y. Qu, J. Gu, J. Wang, H. C. Hsu, J. D. Mountz, N. E. Baldwin, M. A. Langston, et al. 2005. Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature genetics* 37: 233–242.

Fusi, N., O. Stegle, and N. D. Lawrence. 2012. Joint modelling of confounding factors and prominent genetic regulators provides increased accuracy in genetical genomics studies. *PLoS Comp. Biol.* 8:e1002330.

Gagnon-Bartsch, J. A., and T. P. Speed. 2012. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* 13:539–552.

Haley, C. S., and S. A. Knott. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity* 69:315–324.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. New York: Springer.

Jansen, R. C. 2007. Quantitative trait loci in inbred lines. In *Handbook of statistical genetics*, ed. DJ Balding, M Bishop, and C Cannings, vol. 1, 3rd ed., 589–622. Chichester: Wiley.

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

Jiang, C., and Z. B. Zeng. 1995. Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* 140(3):1111–1127.

Kang, H. M., C. Ye, and E. Eskin. 2008. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics* 180:1909–1925.

Keller, M. P., Y. Choi, P. Wang, D. B. Davis, M. E. Rabaglia, A. T. Oler, D. S. Stapleton, C. Argmann, K. L. Schueler, S. Edwards, H. A. Steinberg, E. C. Neto, R. Kleinhanz, S. Turner, M. K. Hellerstein, E. E. Schadt, B. S. Yandell, C. Kendziorski, and A. D. Attie. 2008. A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Res.* 18:706–716.

Knott, S. A., and C. S. Haley. 2000. Multitrait least squares for quantitative trait loci detection. *Genetics* 156(2):899–911.

Leek, J. T., and J. D. Storey. 2007. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* 3:1724–1735.

Listgarten, J., C. Kadie, E. E. Schadt, and D. Heckerman. 2010. Correction for hidden confounders in the genetic analysis of gene expression. *Proc. Natl. Acad. Sci USA* 107: 16465–16470.

Nadeau, J. H., and W. N. Frankel. 2000. The roads from phenotypic variation to gene discovery: mutagenesis versus QTLs. *Nat. Genet.* 25(4):381–384.

R Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schadt, E. E., S. A. Monks, T. A. Drake, A. J. Lusis, N. Che, V. Colinayo, T. G. Ruff, S. B. Milligan, J. R. Lamb, G. Cavet, et al. 2003. Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422:297–302.

Stegle, O., L. Parts, R. Durbin, and J. Winn. 2010. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comp. Biol.* 6:e1000770.

Tian, J., M. Keller, A. Oler, M. Rabagalia, K. Schueler, D. Stapleton, A. T. Broman, W. Zhao, C. Kendziorski, B. S. Yandell, B. Hagenbuch, K. W. Broman, and A. D. Attie. 2015. Identification of *Slco1a6* as a candidate gene that broadly affects gene expression in mouse pancreatic islets. *bioRxiv* doi:http://dx.doi.org/10.1101/020974.

Yvert, G., R. B. Brem, J. Whittle, J. M. Akey, E. Foss, E. N. Smith, R. Mackelprang, and L. Kruglyak. 2003. Trans-acting regulatory variation in saccharomyces cerevisiae and the role of transcription factors. *Nature genetics* 35:57–64.

Zeng, Z.-B., J. Liu, L. F. Stam, C.-H. Kao, J. M. Mercer, and C. C. Laurie. 2000. Genetic architecture of a morphological shape difference between two Drosophila species. *Genetics* 154:299–310.

Identification of *Slco1a6* as a candidate gene that broadly affects gene expression in mouse pancreatic islets

SUPPLEMENT

Jianan Tian*, Mark P. Keller†, Angie T. Oler†, Mary E. Rabaglia†, Kathryn L. Schueler†,

Donald S. Stapleton†, Aimee Teo Broman‡, Wen Zhao**, Christina Kendziorski†, Brian S. Yandell*,§,

Bruno Hagenbuch**, Karl W. Broman‡, Alan D. Attie†

*Departments of *Statistics, †Biochemistry, ‡Biostatistics & Medical Informatics, and §Horticulture,

University of Wisconsin–Madison, Madison, Wisconsin 53706, and **Department of Pharmacology, Toxicology

and Therapeutics, The University of Kansas Medical Center, Kansas City, Kansas 66160

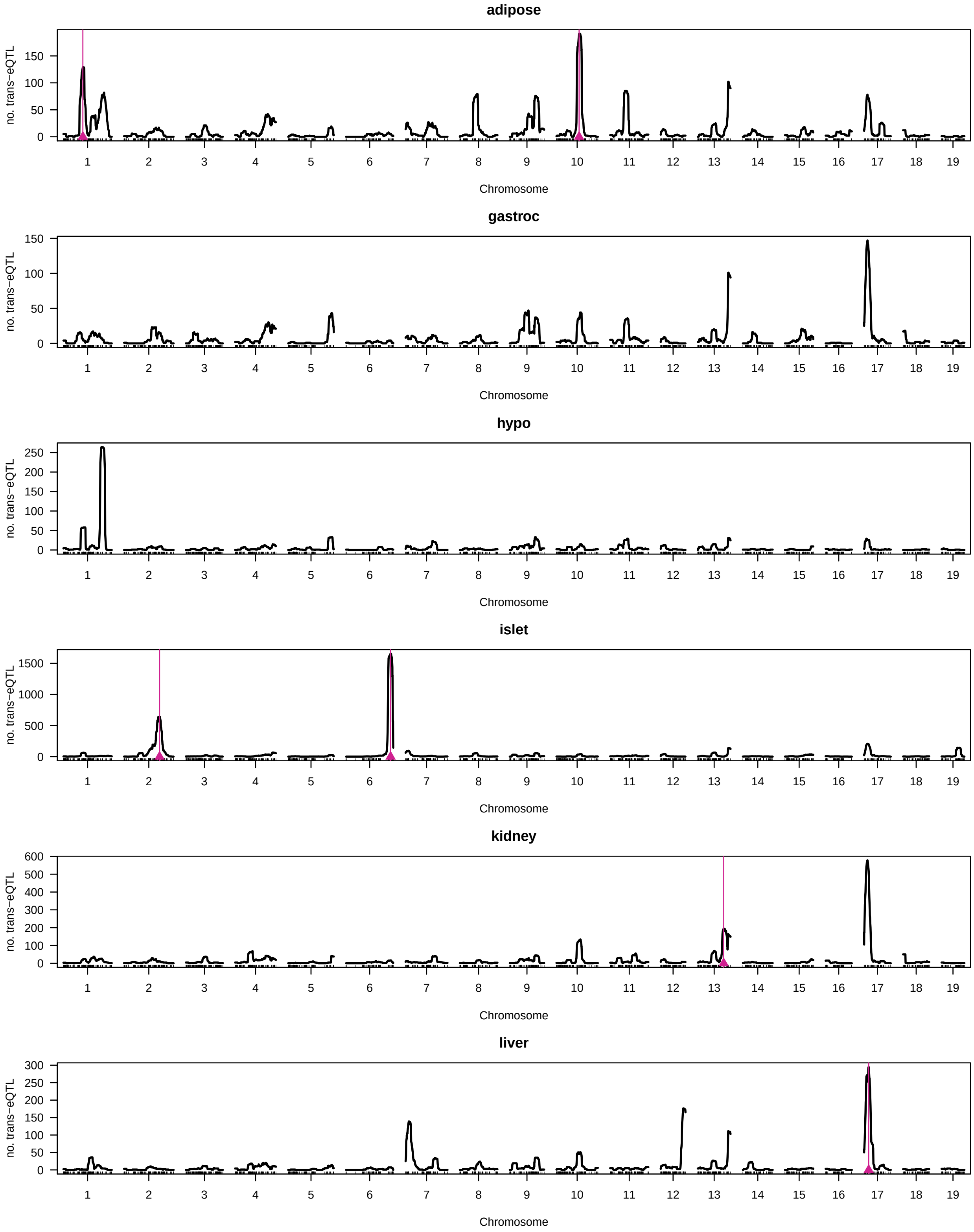


Figure 3.7: Number of expression traits with LOD $\geqslant 10$ in a 10 cM sliding window across the genome. Red triangles indicate the six *trans*-eQTL hotspots used as examples in Figures 2, 3, and 5.

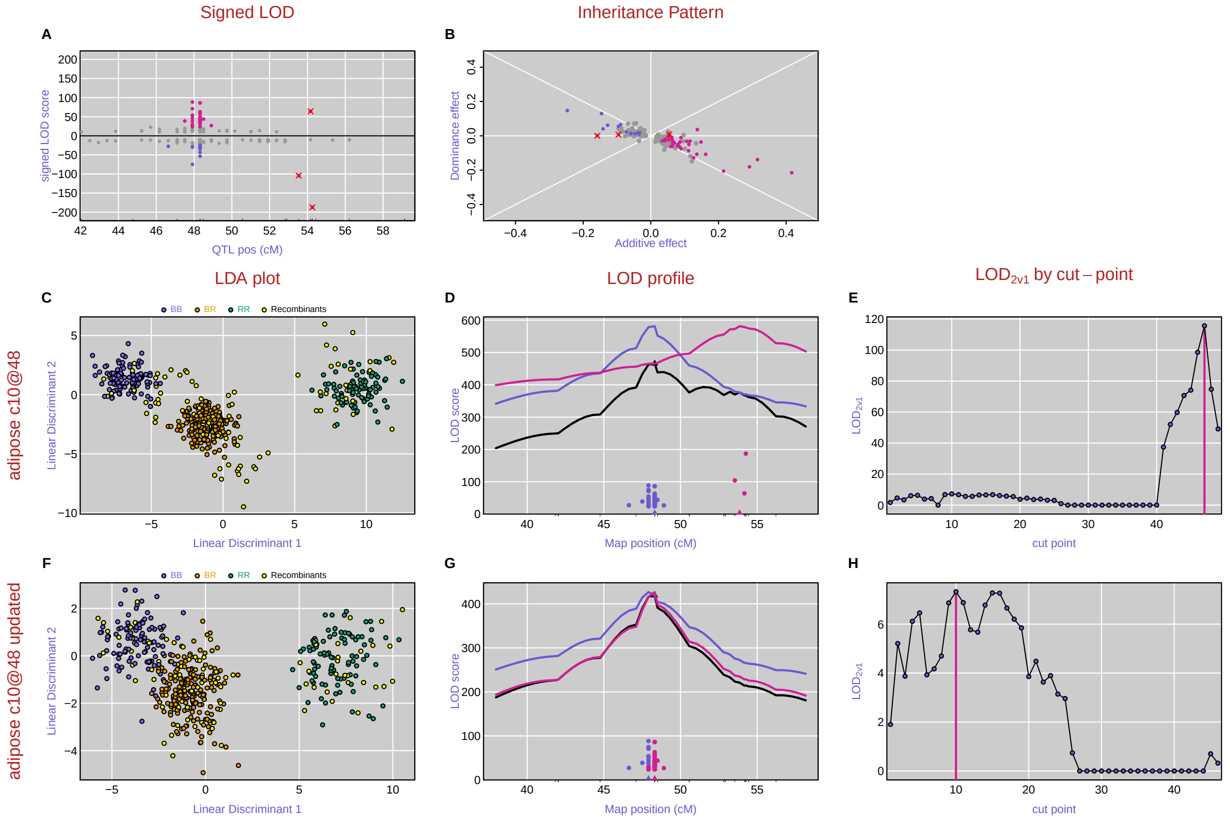


Figure 3.8: Effect of omitting three transcripts from the chromosome 10 *trans*-eQTL hotspot on the analysis results. **A**: Signed LOD scores, with transcripts having LOD ⩾ 10 highlighted. The three transcripts to be omitted are indicated with X's. **B**: Scatterplot of dominanance versus additive effects, with transcripts having LOD ⩾ 10 highlighted. The three transcripts to be omitted are again indicated with X's. **C-E**: LDA and likelihood results, as in Figures 3 and 5. **F-H**: LDA and likelihood results with the three transcripts omitted.

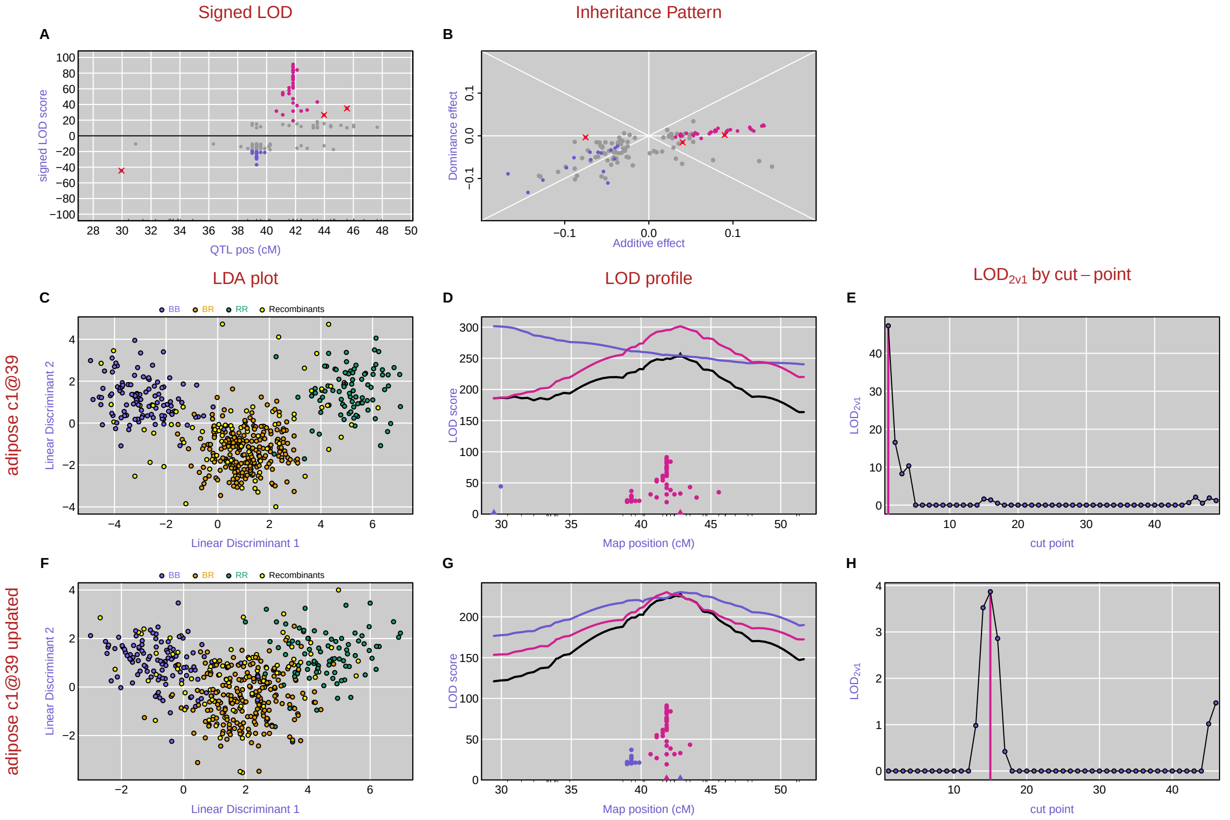
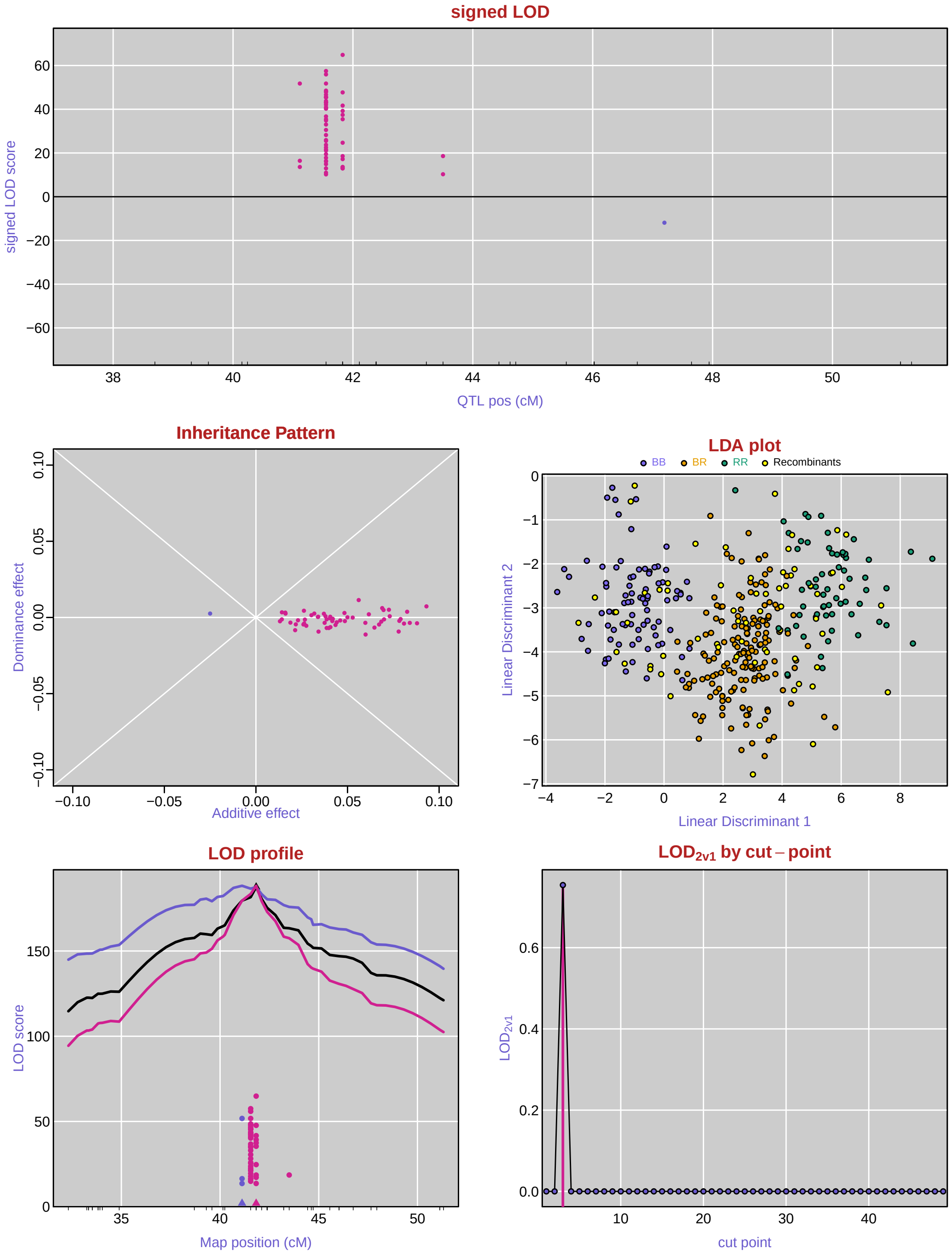
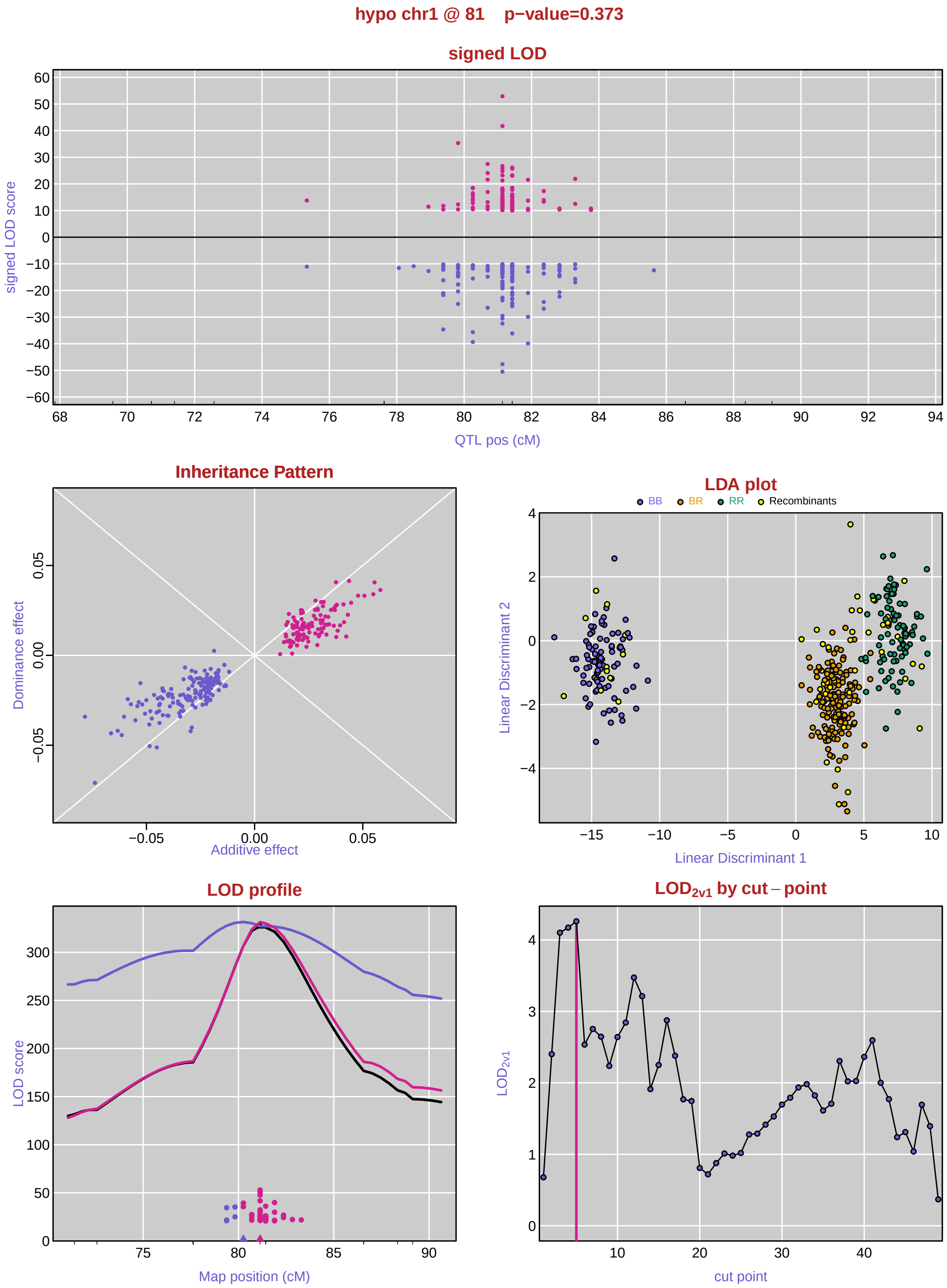


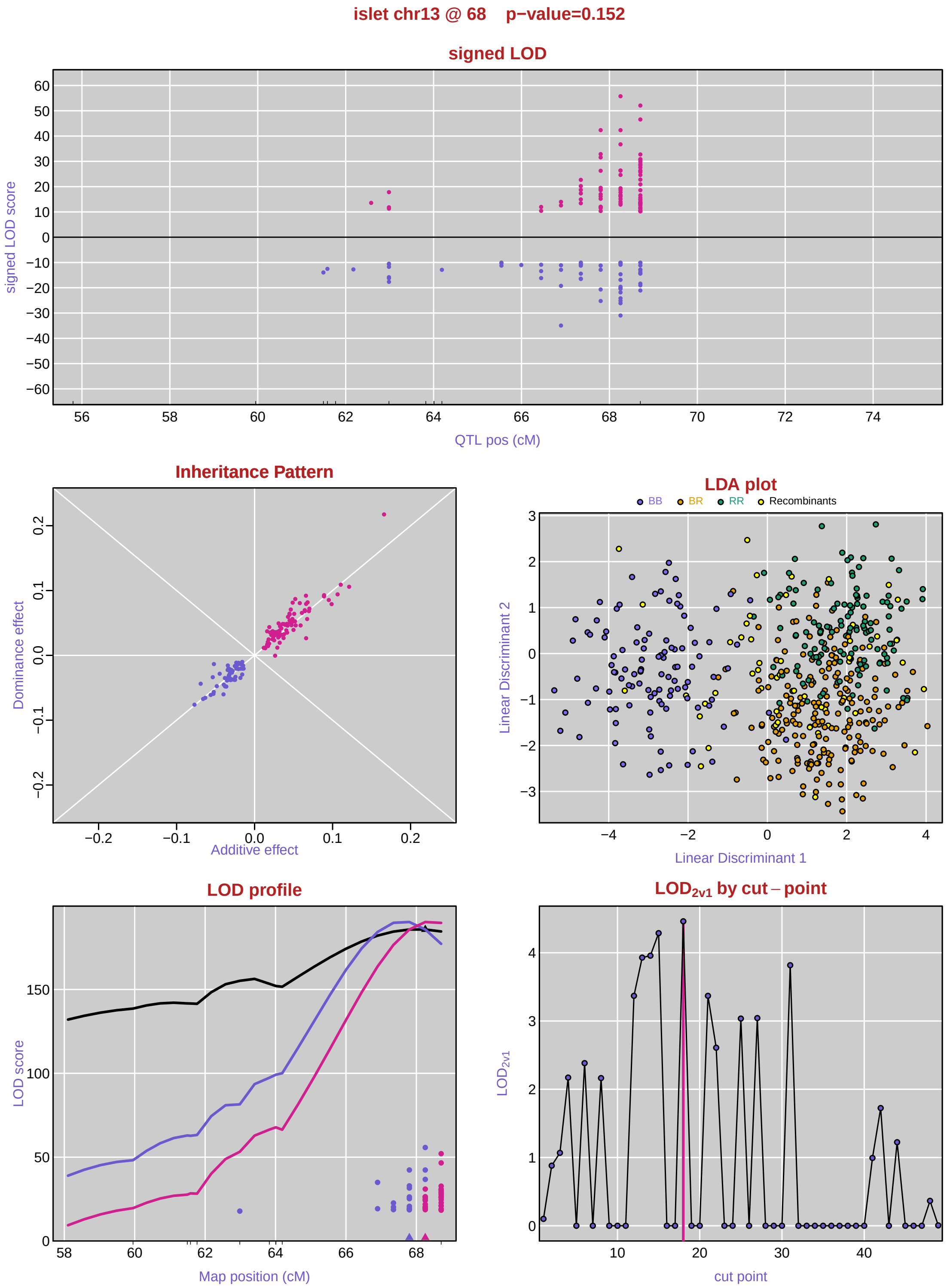
Figure 3.9: Effect of omitting one transcript from the chromosome 1 *trans*-eQTL hotspot on the analysis results. **A**: Signed LOD scores, with transcripts having LOD $\geqslant 10$ highlighted. The transcript to be omitted is indicated with an X. **B**: Scatterplot of dominanance versus additive effects, with transcripts having LOD $\geqslant 10$ highlighted. The transcript to be omitted is again indicated with an X. **C-E**: LDA and likelihood results, as in Figures 3 and 5. **F-H**: LDA and likelihood results with one transcript omitted.

FileS1.pdf A 35-page PDF with the results as in Figures 2, 3, and 4, for all 35 *trans*-eQTL hotspots identified.

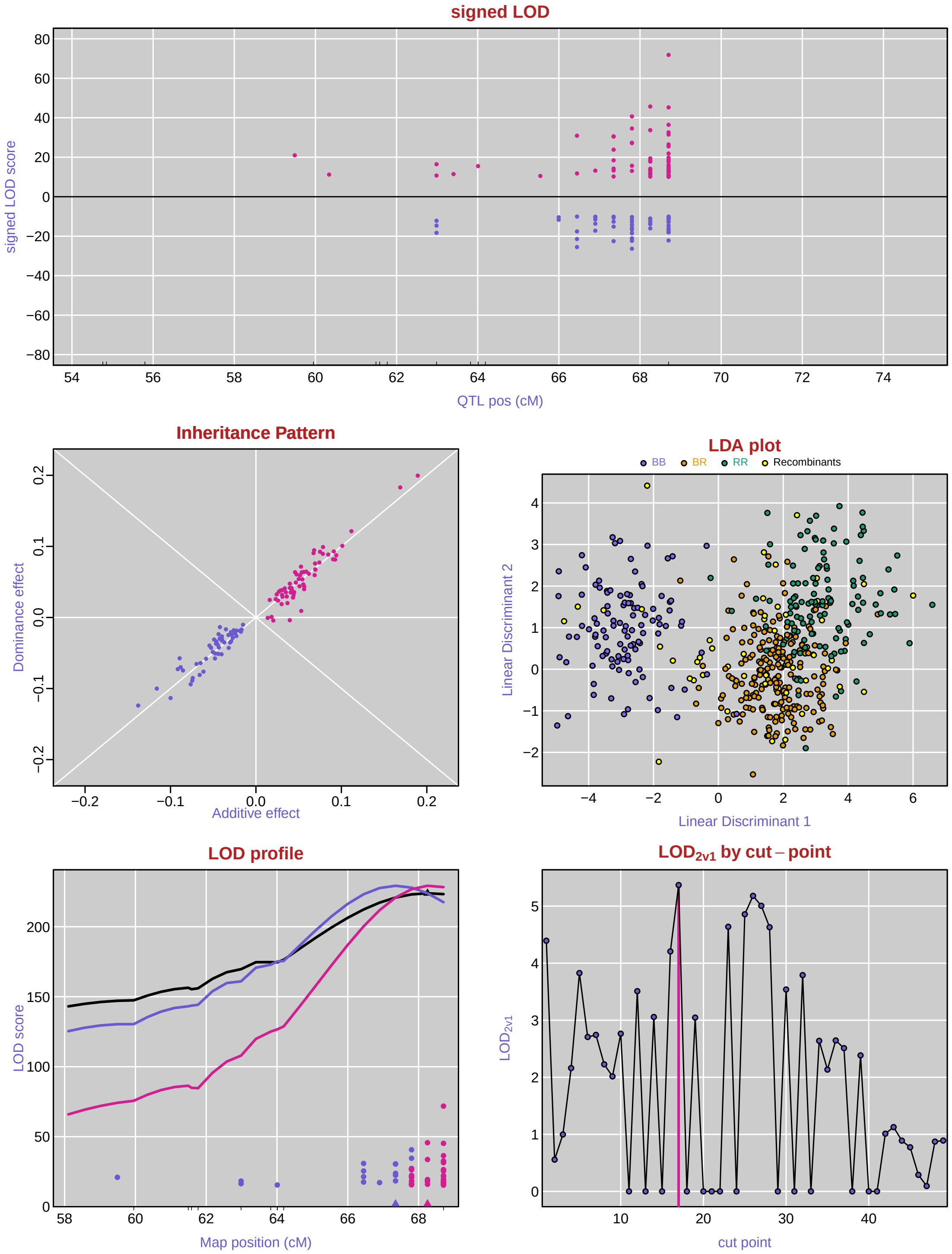
hypo chr1 @ 41 p−value=0.984

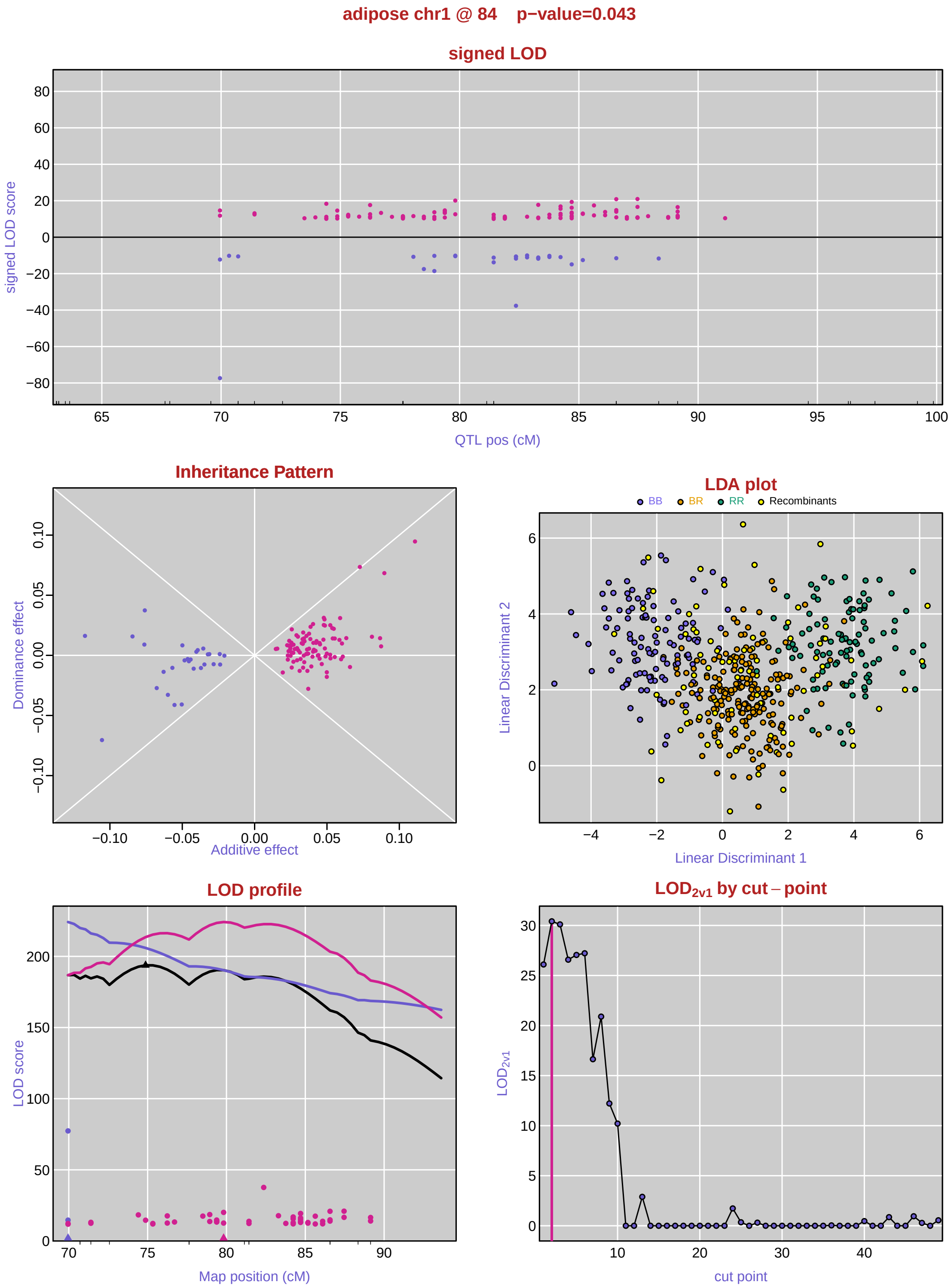


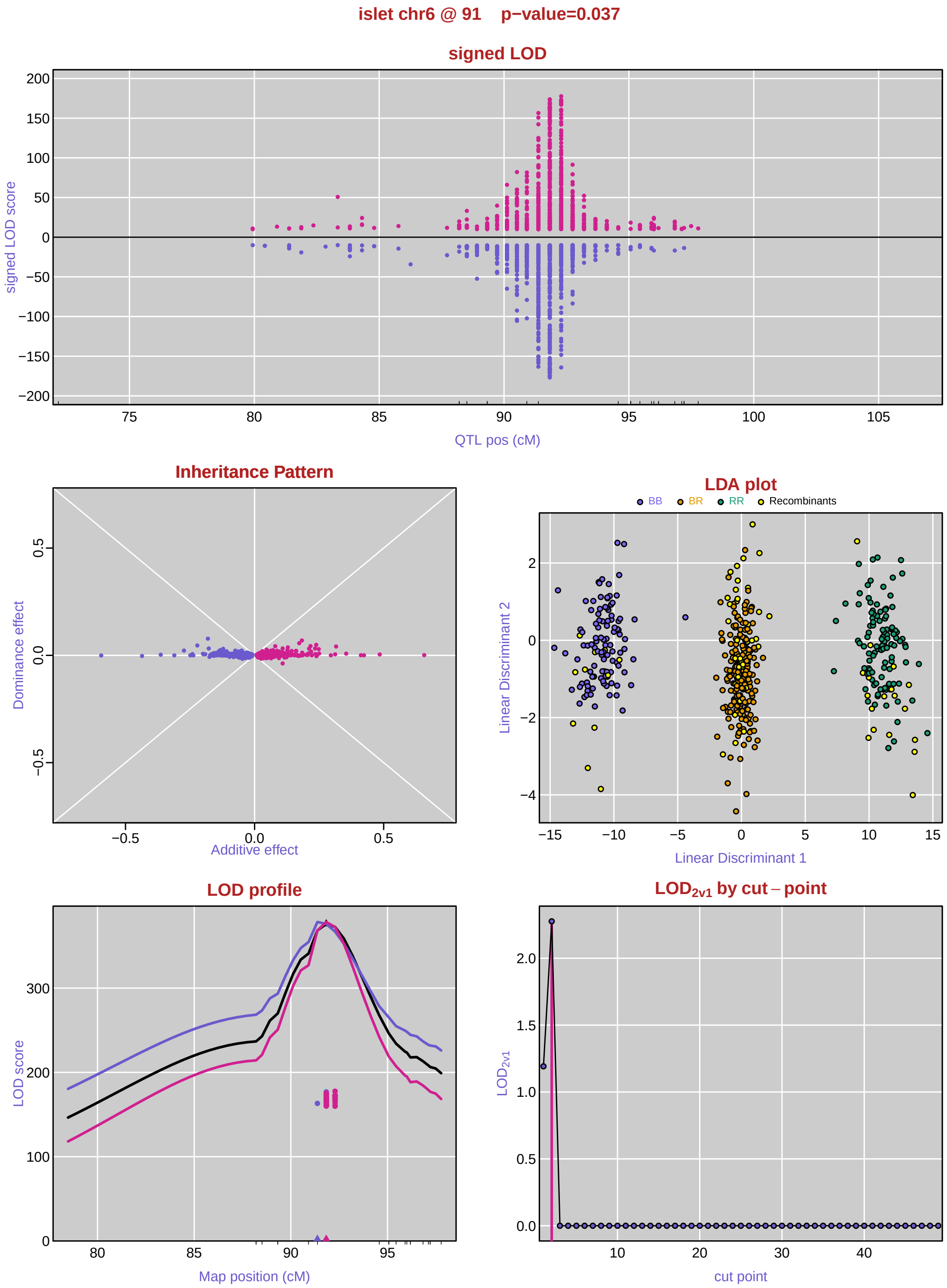

hypo chr1 @ 81 p−value=0.373
signed LOD
signed LOD score
60
50
40
30
20
10
0
−10
−20
−30
−40
−50
−60
68
70
72
74
76
78
80
82
84
86
88
90
92
94
QTL pos (cM)
Inheritance Pattern
Dominance effect
−0.05
0.00
0.05
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
4
2
0
−2
−4
−15
−10
−5
0
5
10
Linear Discriminant 1
LOD profile
LOD score
300
250
200
150
100
50
0
75
80
85
90
Map position (cM)
LOD_{2v1} by cut − point
LOD_{2v1}
4
3
2
1
0
10
20
30
40
cut point


islet chr13 @ 68 p−value=0.152
signed LOD
signed LOD score
60
50
40
30
20
10
0
−10
−20
−30
−40
−50
−60
56
58
60
62
64
66
68
70
72
74
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.2
0.1
0.0
−0.1
−0.2
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
3
2
1
0
−1
−2
−3
−4
4
Linear Discriminant 1
LOD profile
LOD score
150
100
50
0
58
60
62
64
66
68
Map position (cM)
LOD_{2v1} by cut − point
LOD_{2v1}
4
3
2
1
0
10
20
30
40
cut point

liver chr13 @ 68 p−value=0.119




adipose chr1 @ 84 p−value=0.043
signed LOD
signed LOD score
80
60
40
20
0
−20
−40
−60
−80
65
70
75
80
85
90
95
100
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.10
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
6
4
2
0
−4
−2
0
2
4
6
Linear Discriminant 1
LOD profile
LOD score
200
150
100
50
0
70
75
80
85
90
Map position (cM)
LOD2v1 by cut − point
LOD2v1
30
25
20
15
10
5
0
10
20
30
40
cut point



islet chr6 @ 91 p−value=0.037

signed LOD

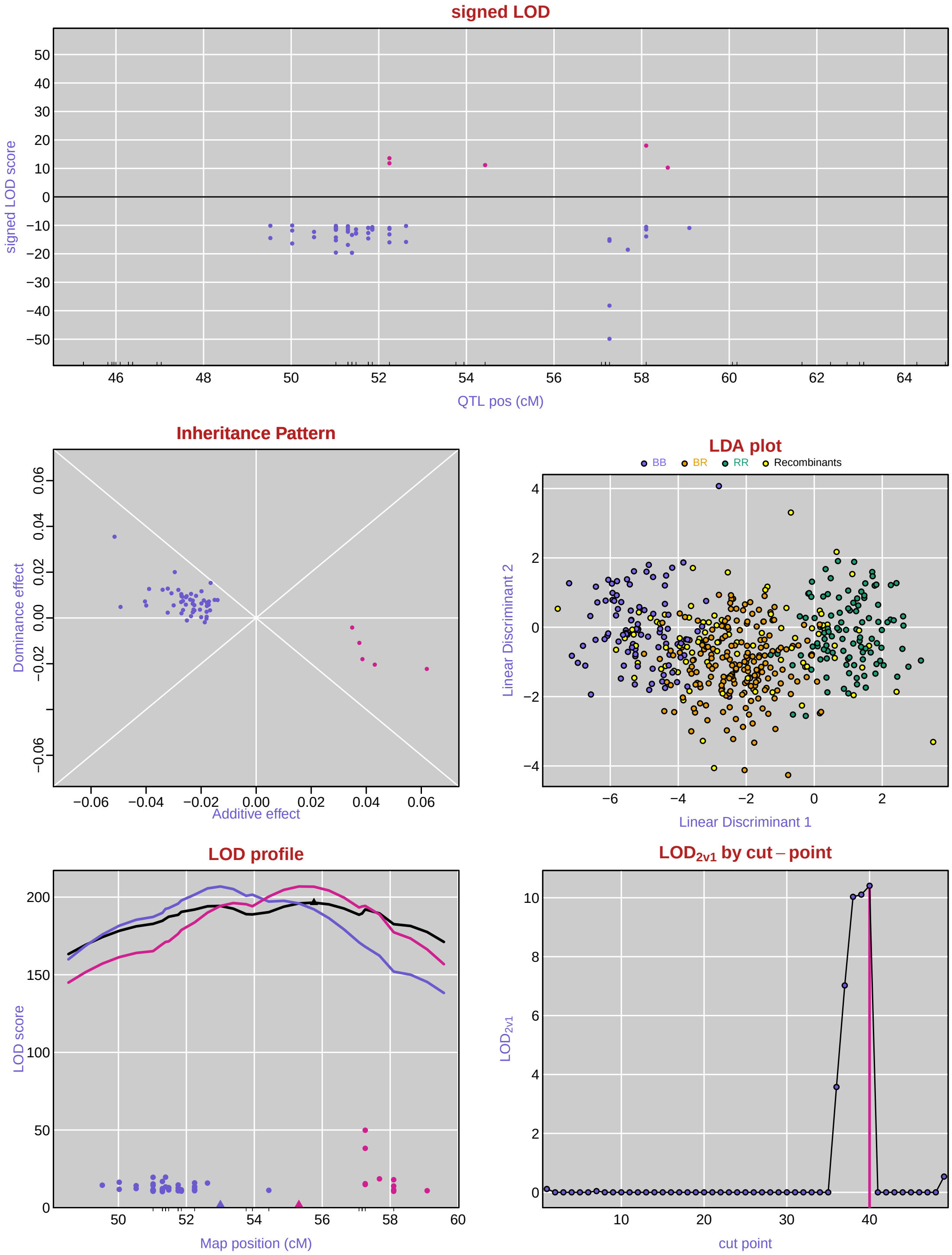
Inheritance Pattern

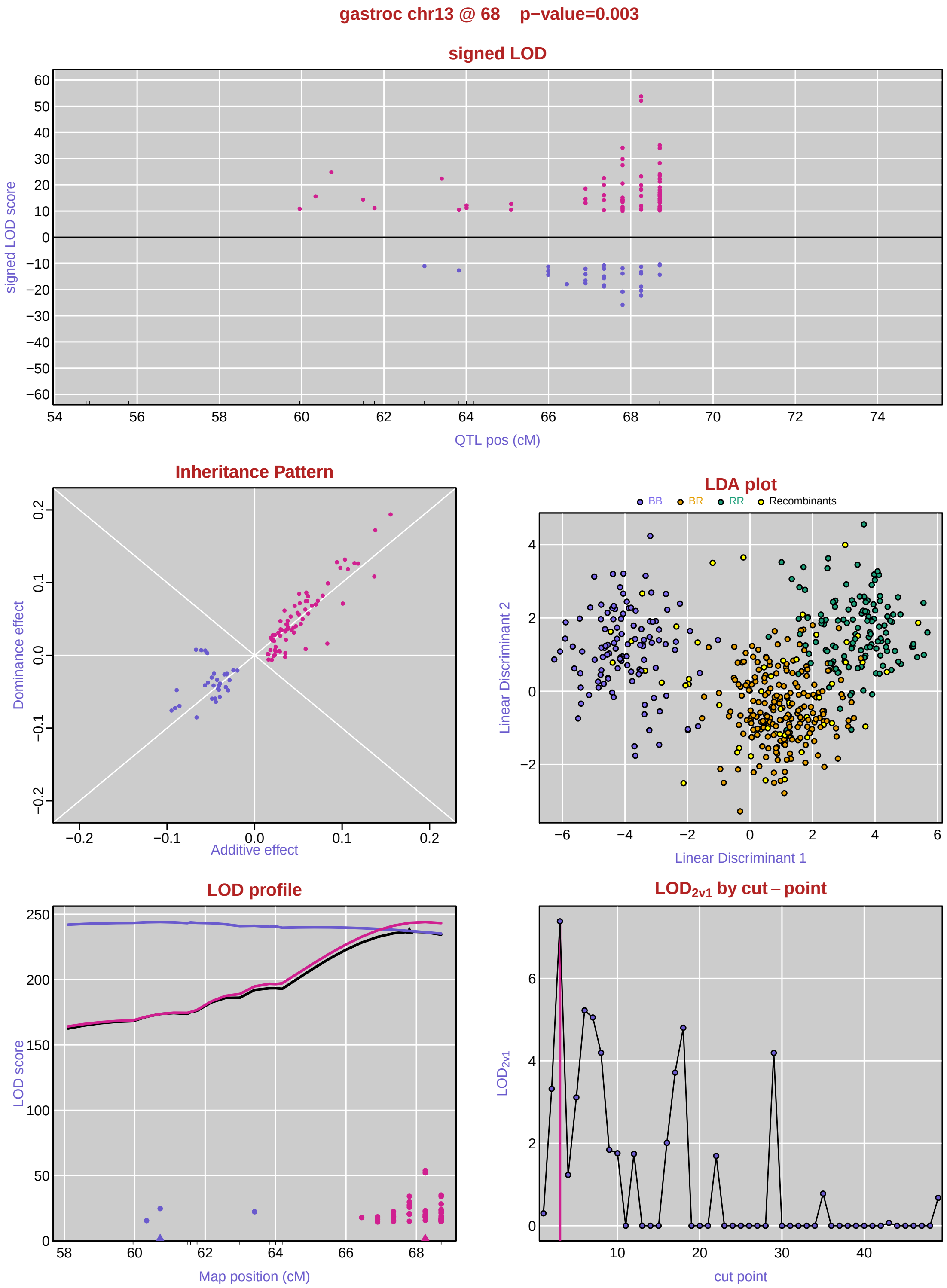
LDA plot

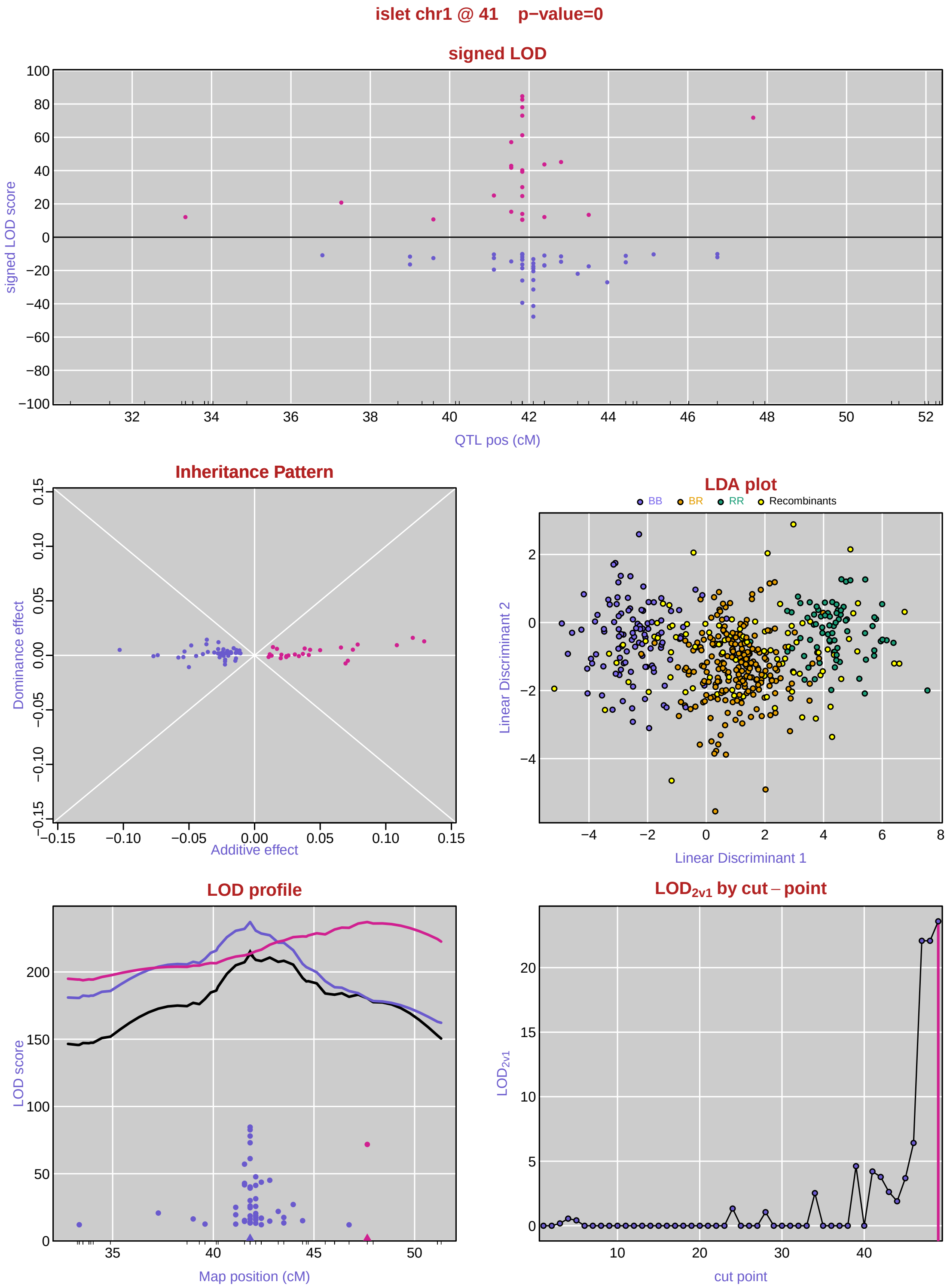
LOD profile

LOD_{2v1} by cut − point

kidney chr11 @ 54 p−value=0.021




gastroc chr13 @ 68 p−value=0.003
signed LOD
signed LOD score
60
50
40
30
20
10
0
−10
−20
−30
−40
−50
−60
54
56
58
60
62
64
66
68
70
72
74
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
4
2
0
−2
−6
−4
−2
0
2
4
6
Linear Discriminant 1
LOD profile
LOD score
250
200
150
100
50
0
58
60
62
64
66
68
Map position (cM)
LOD_{2v1} by cut − point
LOD_{2v1}
6
4
2
0
10
20
30
40
cut point



islet chr1 @ 41 p−value=0

signed LOD

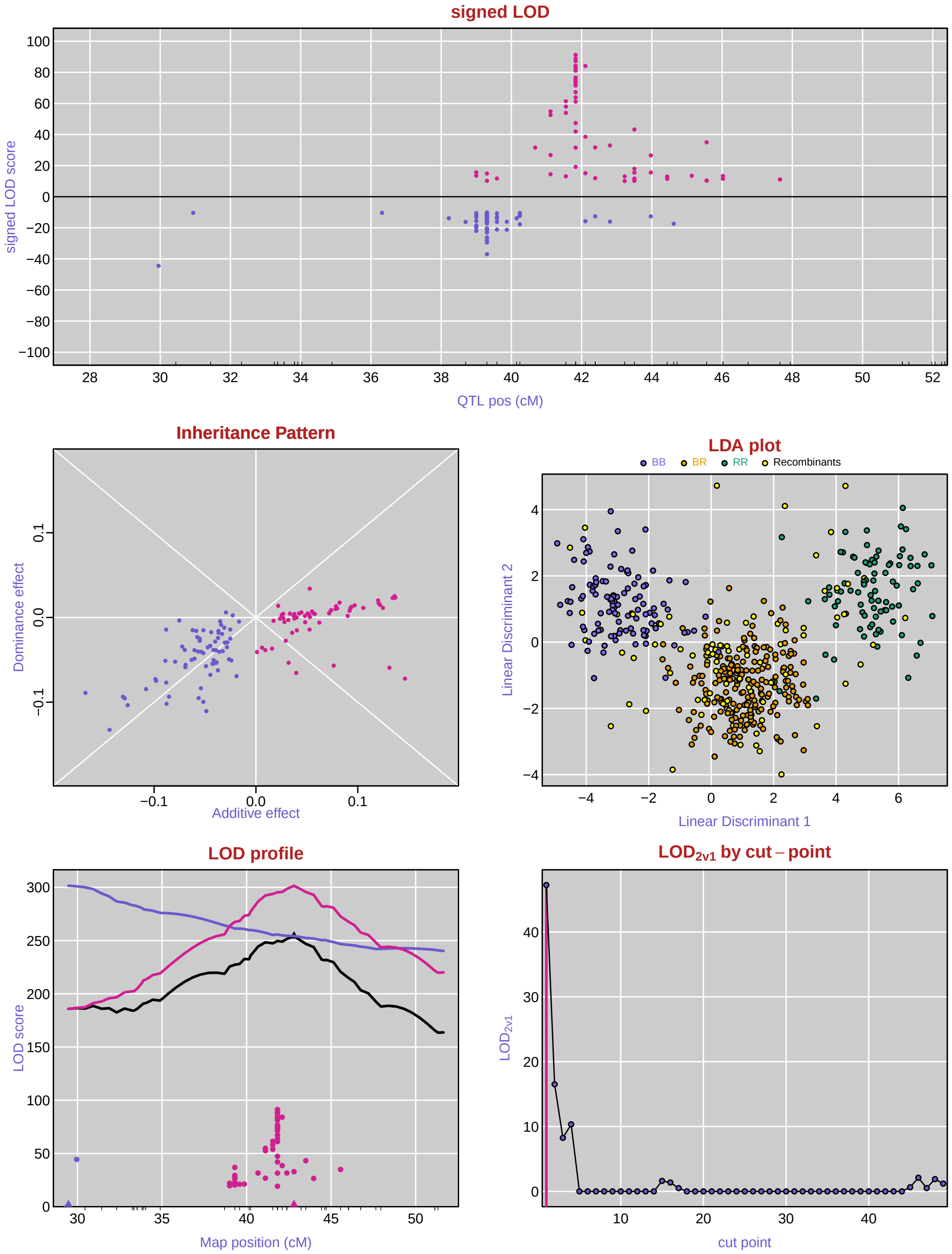
Inheritance Pattern

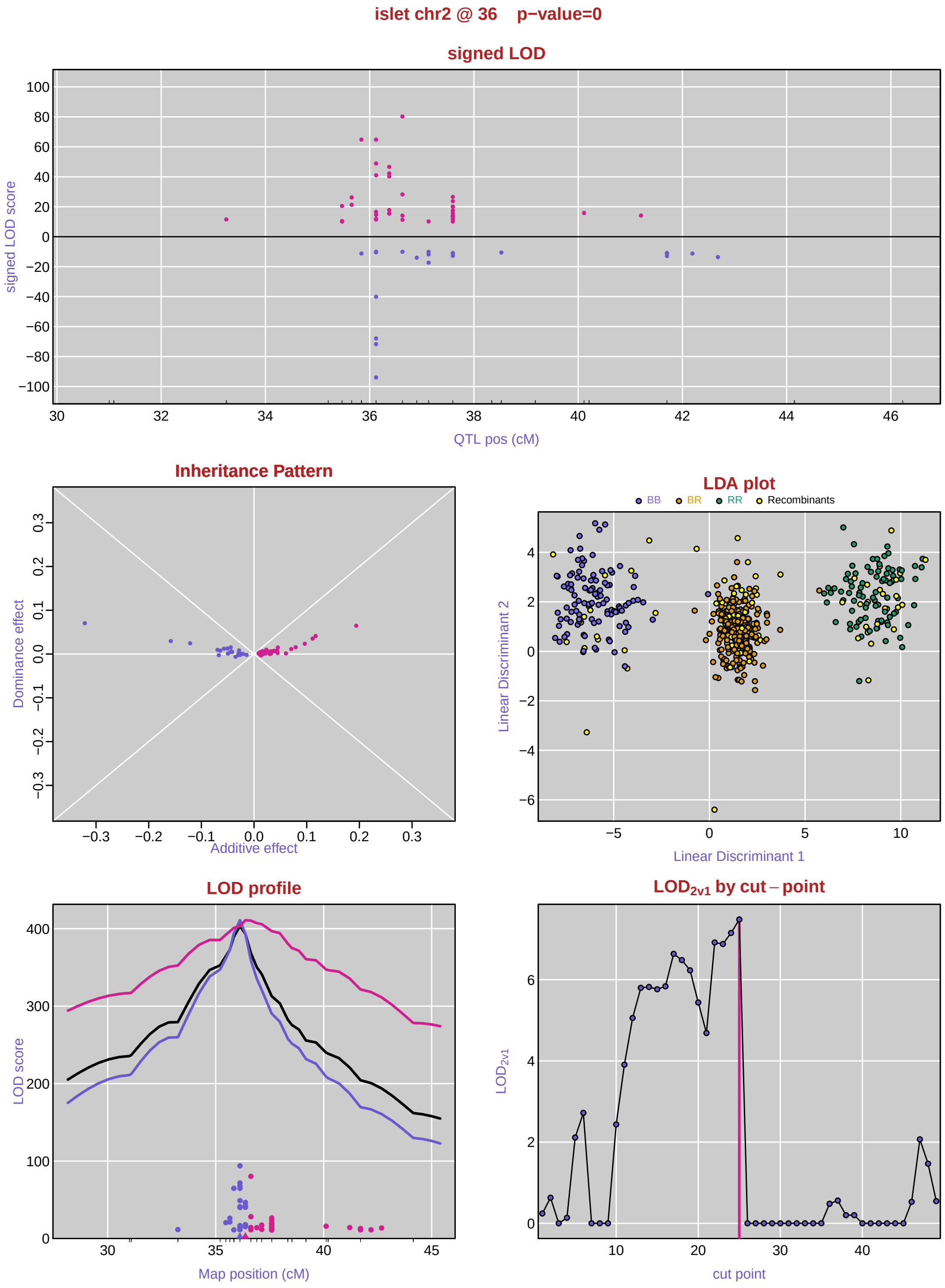
LDA plot

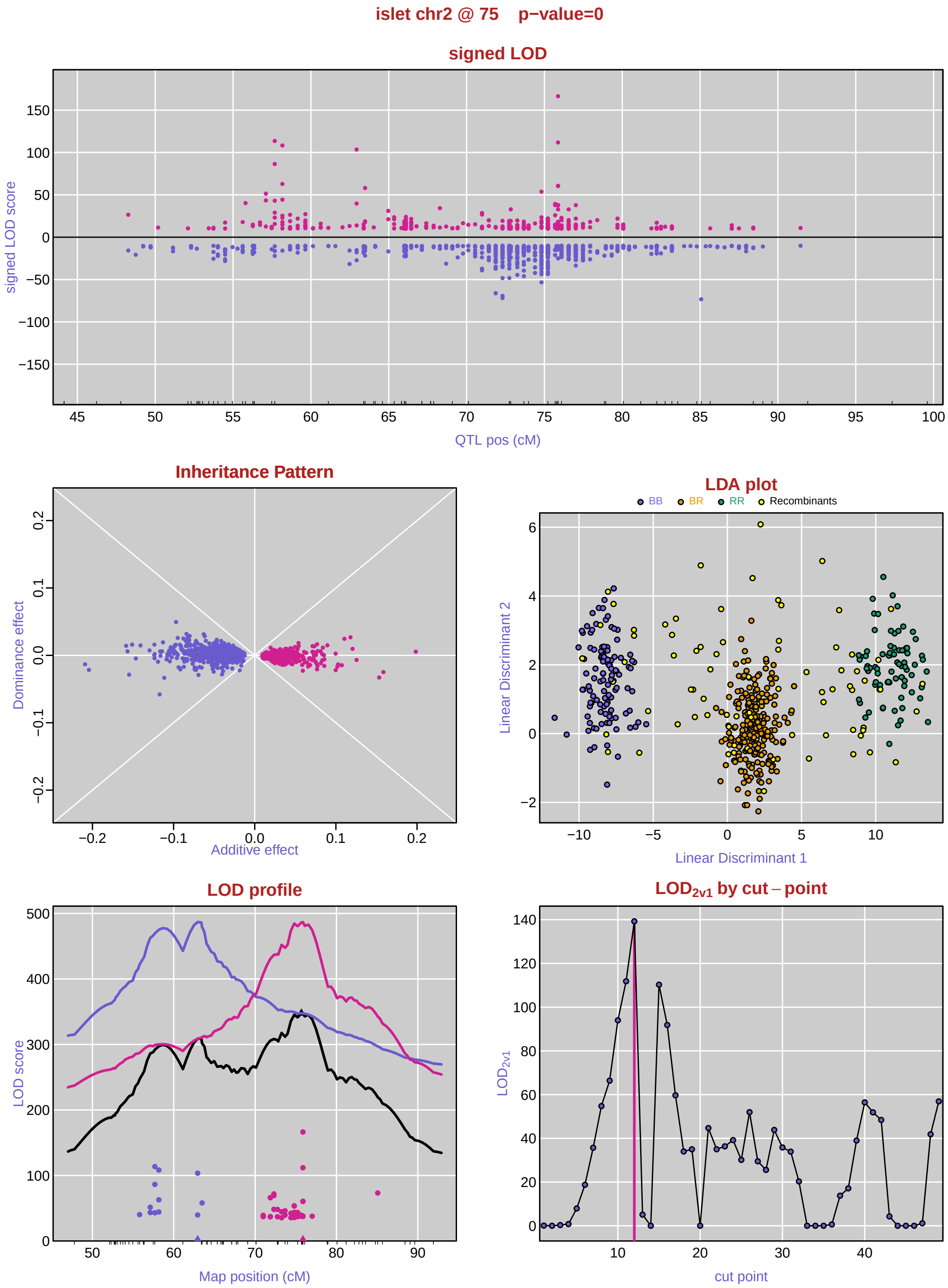
LOD profile

LOD$_{2v1}$ by cut − point

adipose chr1 @ 39   p−value=0




islet chr2 @ 36 p−value=0
signed LOD
signed LOD score
100
80
60
40
20
0
−20
−40
−60
−80
−100
30
32
34
36
38
40
42
44
46
QTL pos (cM)
Inheritance Pattern
Dominance effect
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
4
2
0
−2
−4
−6
−5
0
5
10
Linear Discriminant 1
LOD profile
LOD score
400
300
200
100
0
30
35
40
45
Map position (cM)
LOD$_{2v1}$ by cut − point
LOD$_{2v1}$
6
4
2
0
10
20
30
40
cut point



islet chr2 @ 75   p−value=0

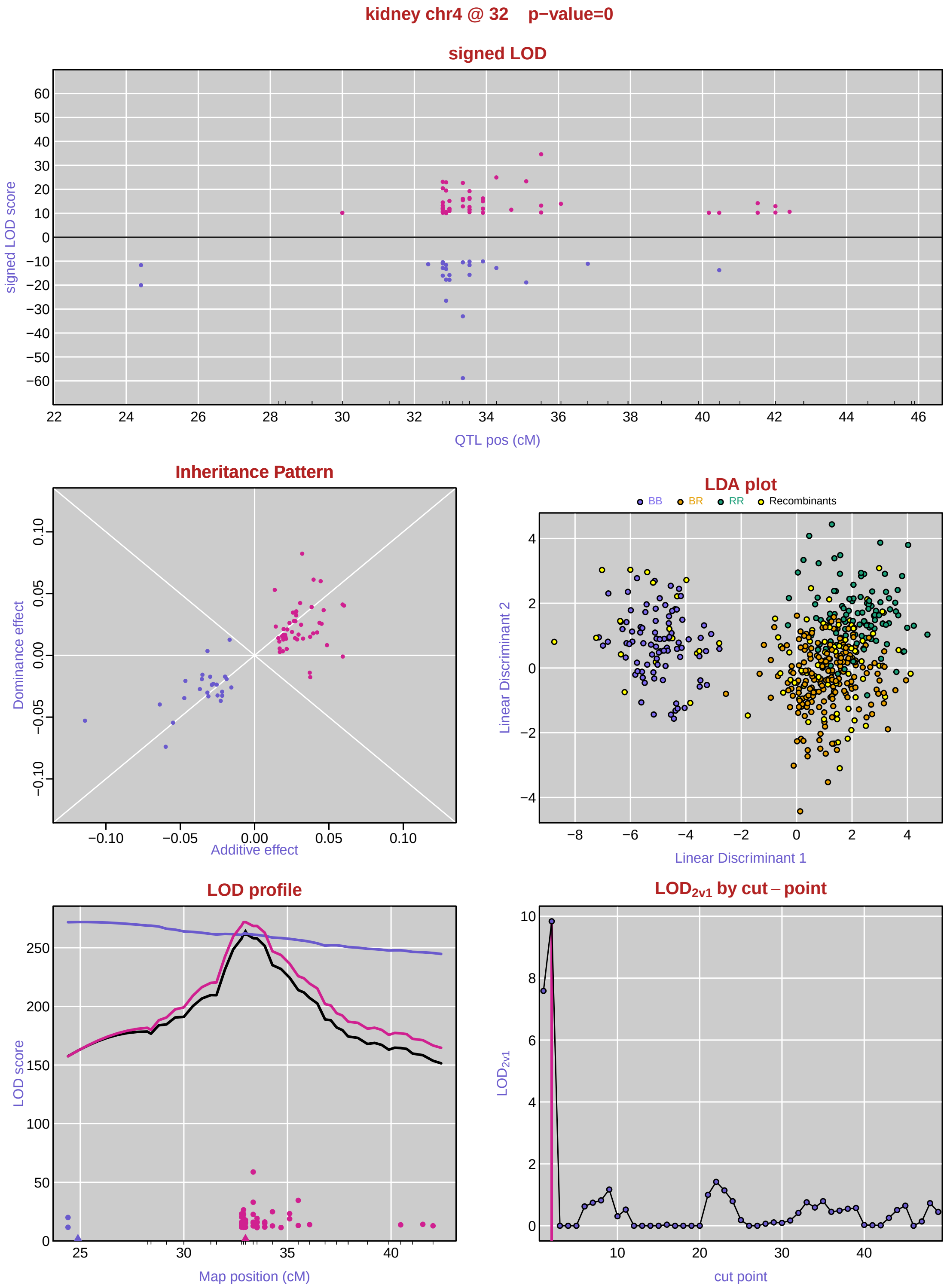
signed LOD

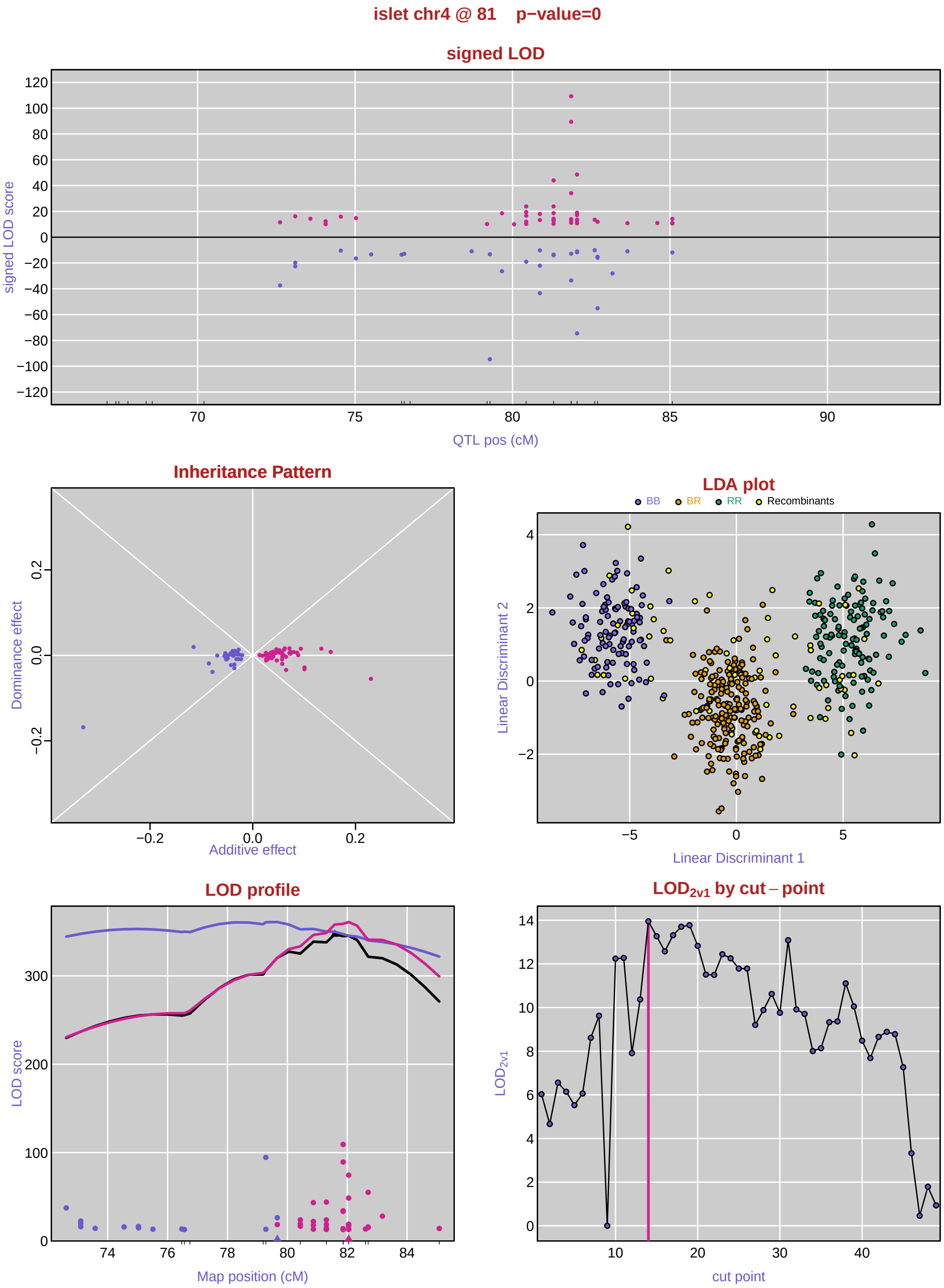
Inheritance Pattern

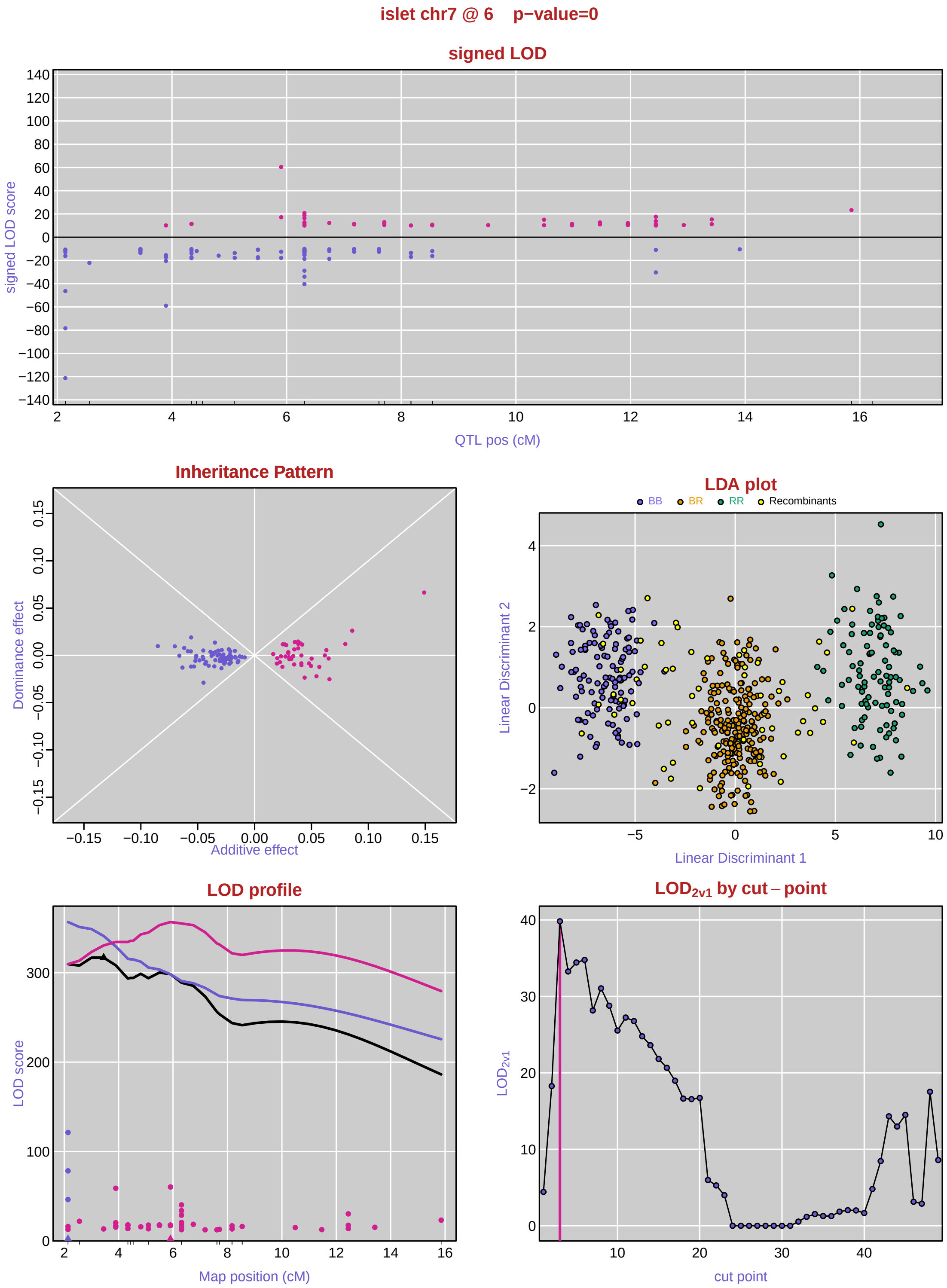
LDA plot

LOD profile

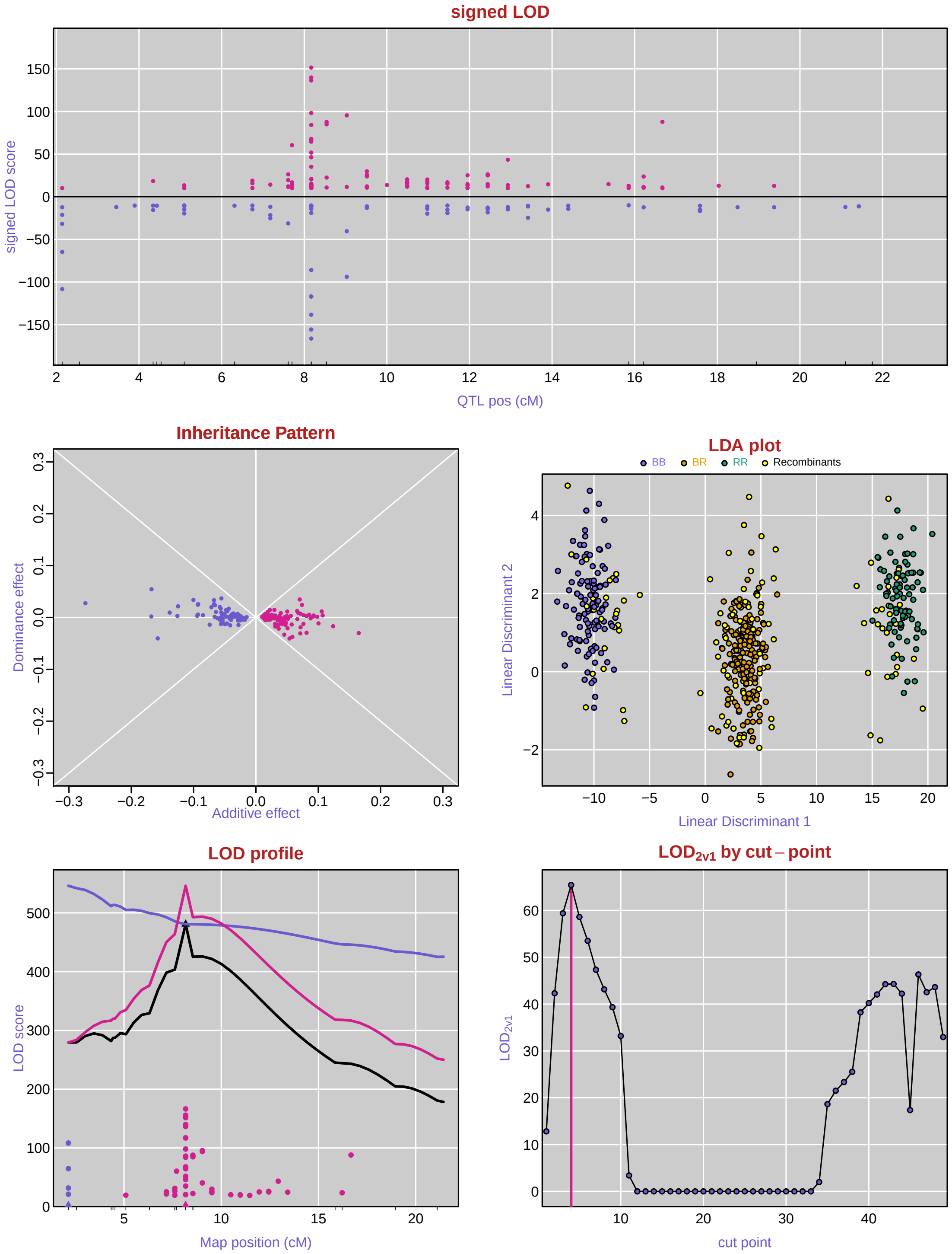
$\mathbf{LOD_{2v1}}$ by cut − point

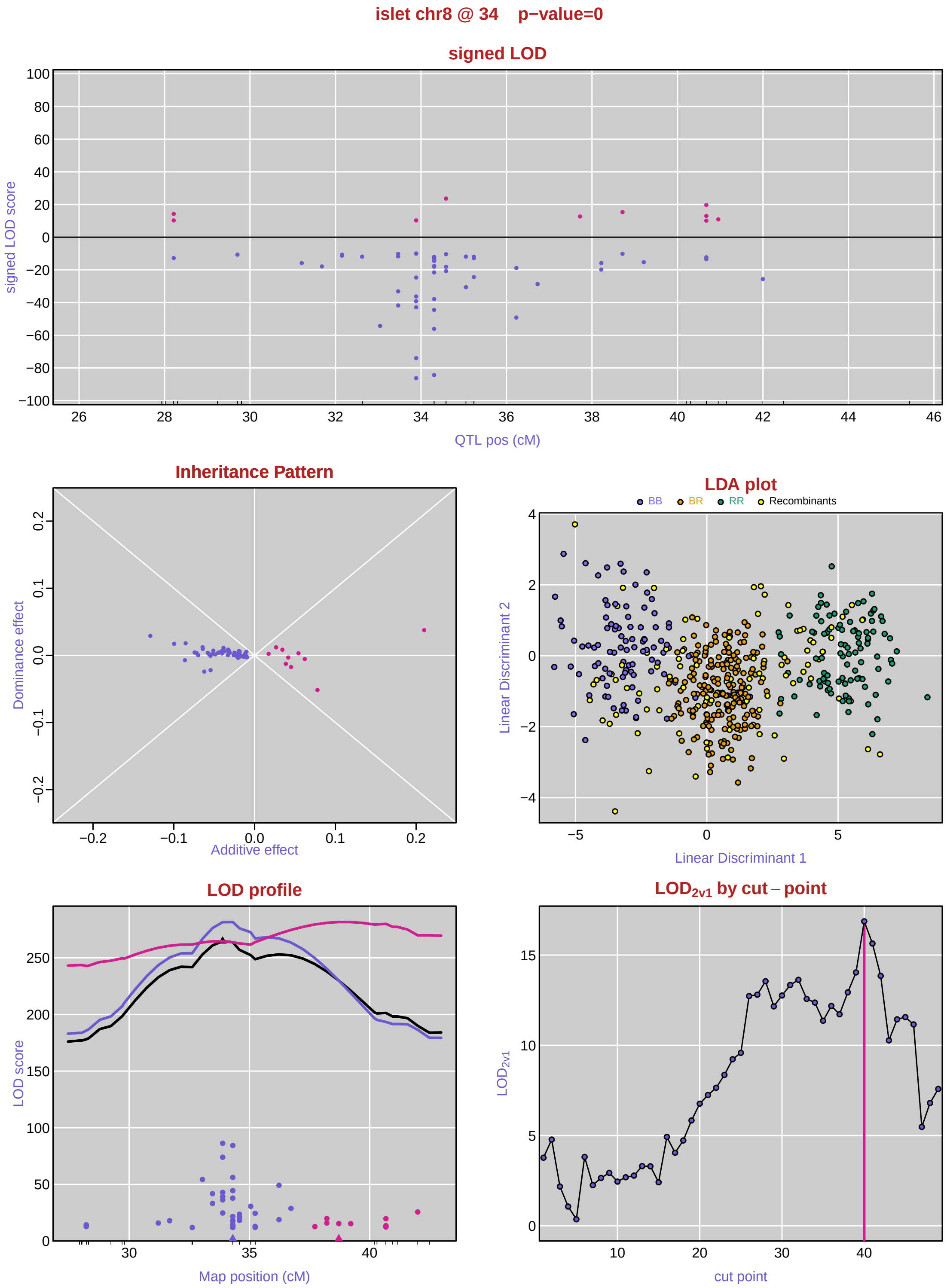

kidney chr4 @ 32 p−value=0
signed LOD
signed LOD score
60
50
40
30
20
10
0
−10
−20
−30
−40
−50
−60
22
24
26
28
30
32
34
36
38
40
42
44
46
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.10
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
4
2
0
−2
−4
−8
−6
−4
−2
0
2
4
Linear Discriminant 1
LOD profile
LOD score
250
200
150
100
50
0
25
30
35
40
Map position (cM)
LOD_{2v1} by cut − point
LOD_{2v1}
10
8
6
4
2
0
10
20
30
40
cut point

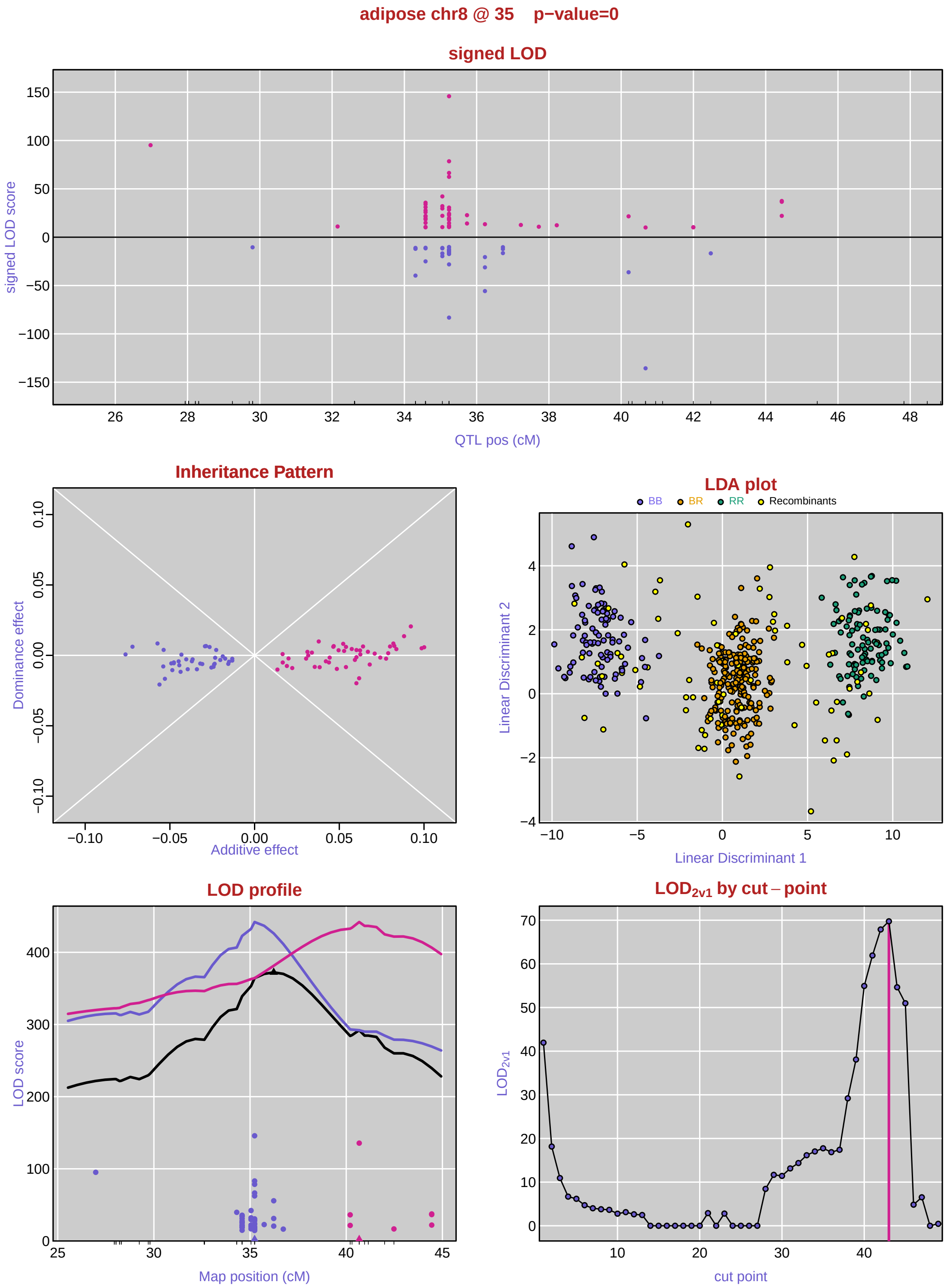

islet chr4 @ 81 p−value=0
signed LOD
signed LOD score
QTL pos (cM)
120
100
80
60
40
20
0
−20
−40
−60
−80
−100
−120
70
75
80
85
90
Inheritance Pattern
Dominance effect
Additive effect
−0.2
0.0
0.2
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
Linear Discriminant 1
4
2
0
−2
−5
5
LOD profile
LOD score
Map position (cM)
300
200
100
74
76
78
80
82
84
LOD2v1 by cut − point
LOD2v1
cut point
14
12
10
8
6
4
10
20
30
40


islet chr7 @ 6 p−value=0
signed LOD
signed LOD score
QTL pos (cM)
140
120
100
80
60
40
20
0
−20
−40
−60
−80
−100
−120
−140
2
4
6
8
10
12
14
16
Inheritance Pattern
Dominance effect
Additive effect
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
Linear Discriminant 1
4
2
0
−2
−5
0
5
10
LOD profile
LOD score
Map position (cM)
300
200
100
0
2
4
6
8
10
12
14
16
LOD2v1 by cut − point
LOD2v1
cut point
40
30
20
10
0
10
20
30
40

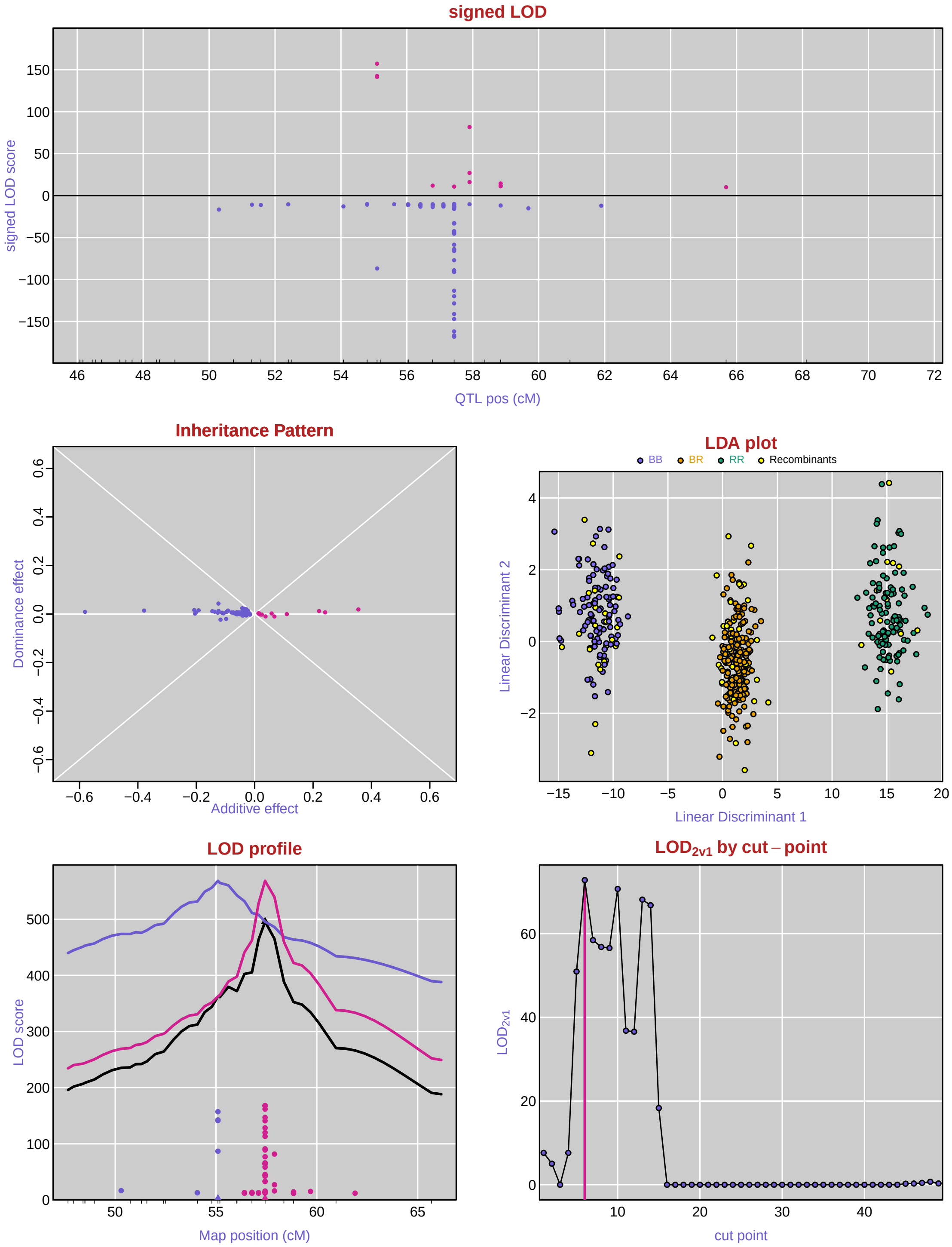
liver chr7 @ 8 p−value=0

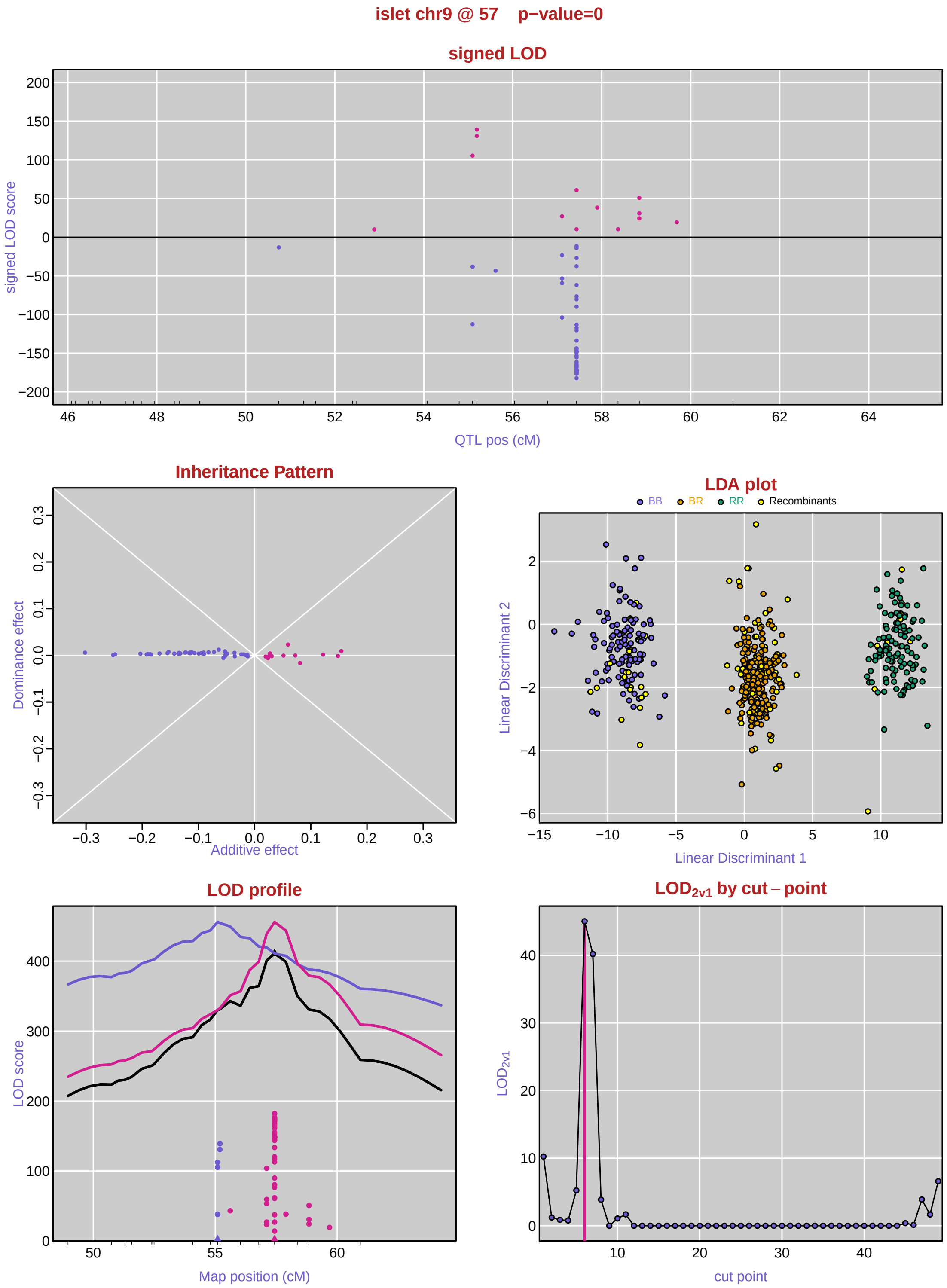


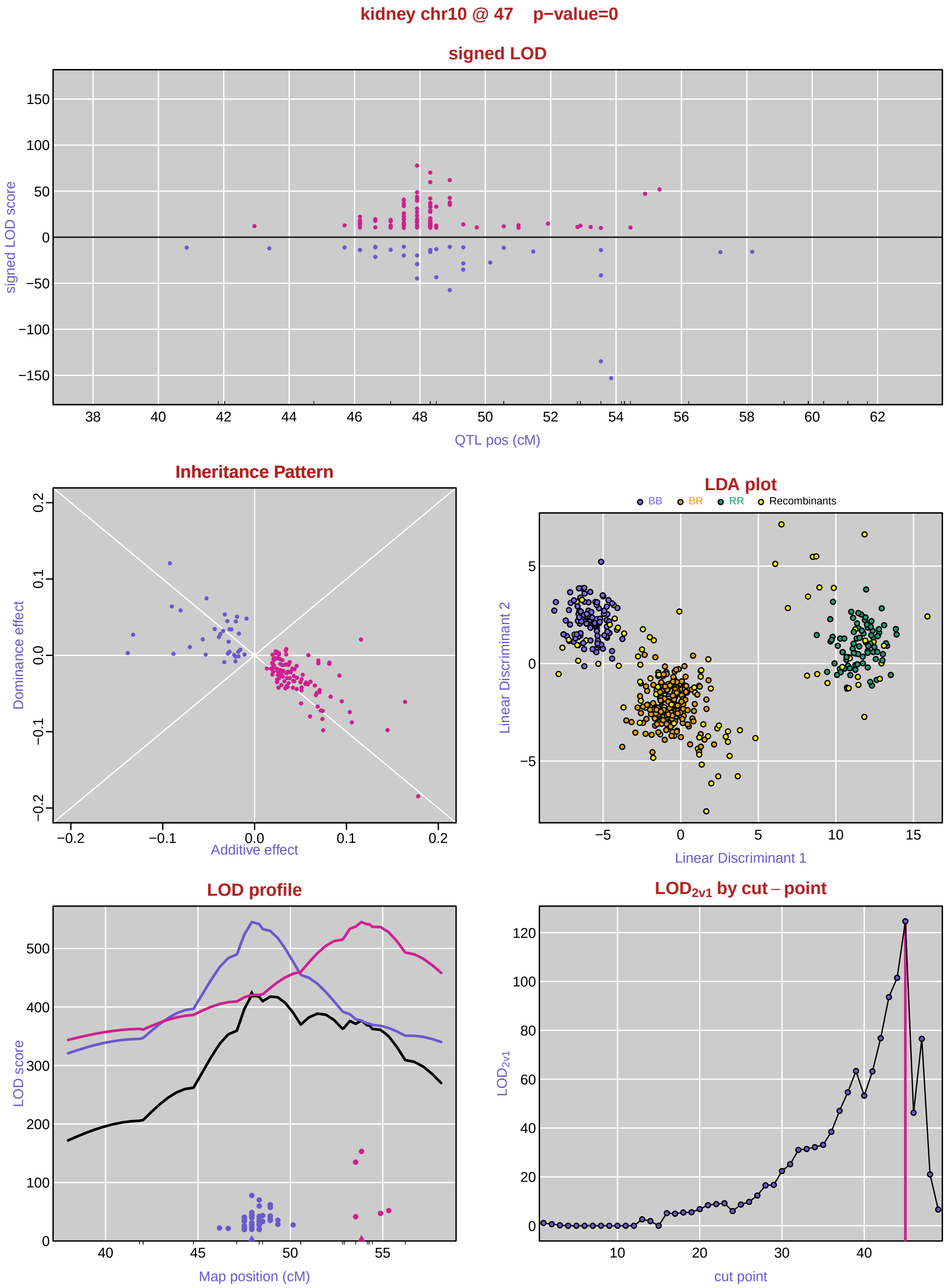

islet chr8 @ 34 p−value=0
signed LOD
signed LOD score
100
80
60
40
20
0
−20
−40
−60
−80
−100
26
28
30
32
34
36
38
40
42
44
46
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.2
0.1
0.0
−0.1
−0.2
Additive effect
−0.2
−0.1
0.0
0.1
0.2
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
4
2
0
−2
−4
Linear Discriminant 1
−5
0
5
LOD profile
LOD score
250
200
150
100
50
0
Map position (cM)
30
35
40
LOD$_{2v1}$ by cut − point
LOD$_{2v1}$
15
10
5
0
cut point
10
20
30
40


adipose chr8 @ 35   p−value=0
signed LOD
signed LOD score
150
100
50
0
−50
−100
−150
26
28
30
32
34
36
38
40
42
44
46
48
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.10
0.05
0.00
−0.05
−0.10
−0.10
−0.05
0.00
0.05
0.10
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
4
2
0
−2
−4
−10
−5
0
5
10
Linear Discriminant 1
LOD profile
LOD score
400
300
200
100
0
25
30
35
40
45
Map position (cM)
LOD_{2v1} by cut − point
LOD_{2v1}
70
60
50
40
30
20
10
0
10
20
30
40
cut point

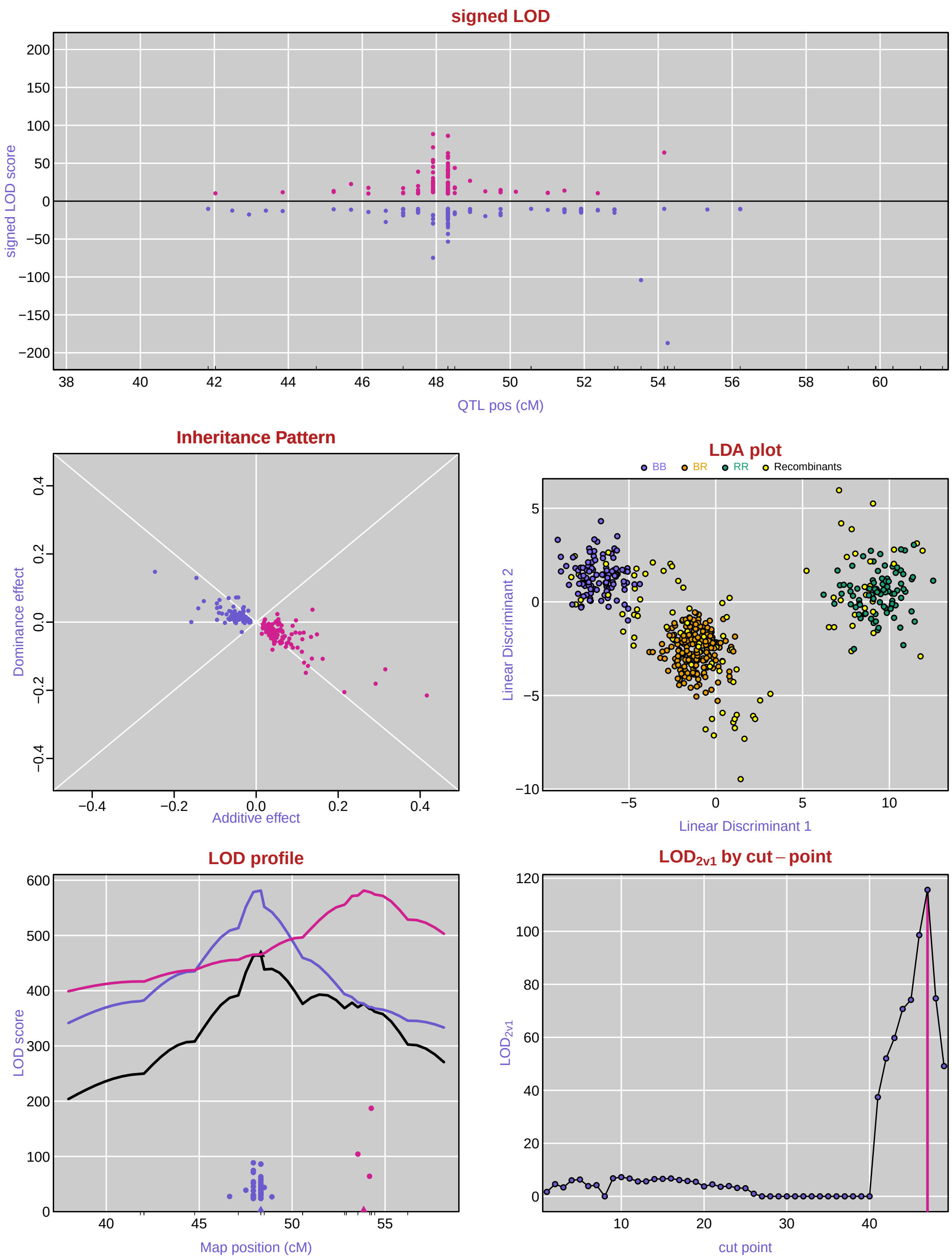
adipose chr9 @ 57 p−value=0

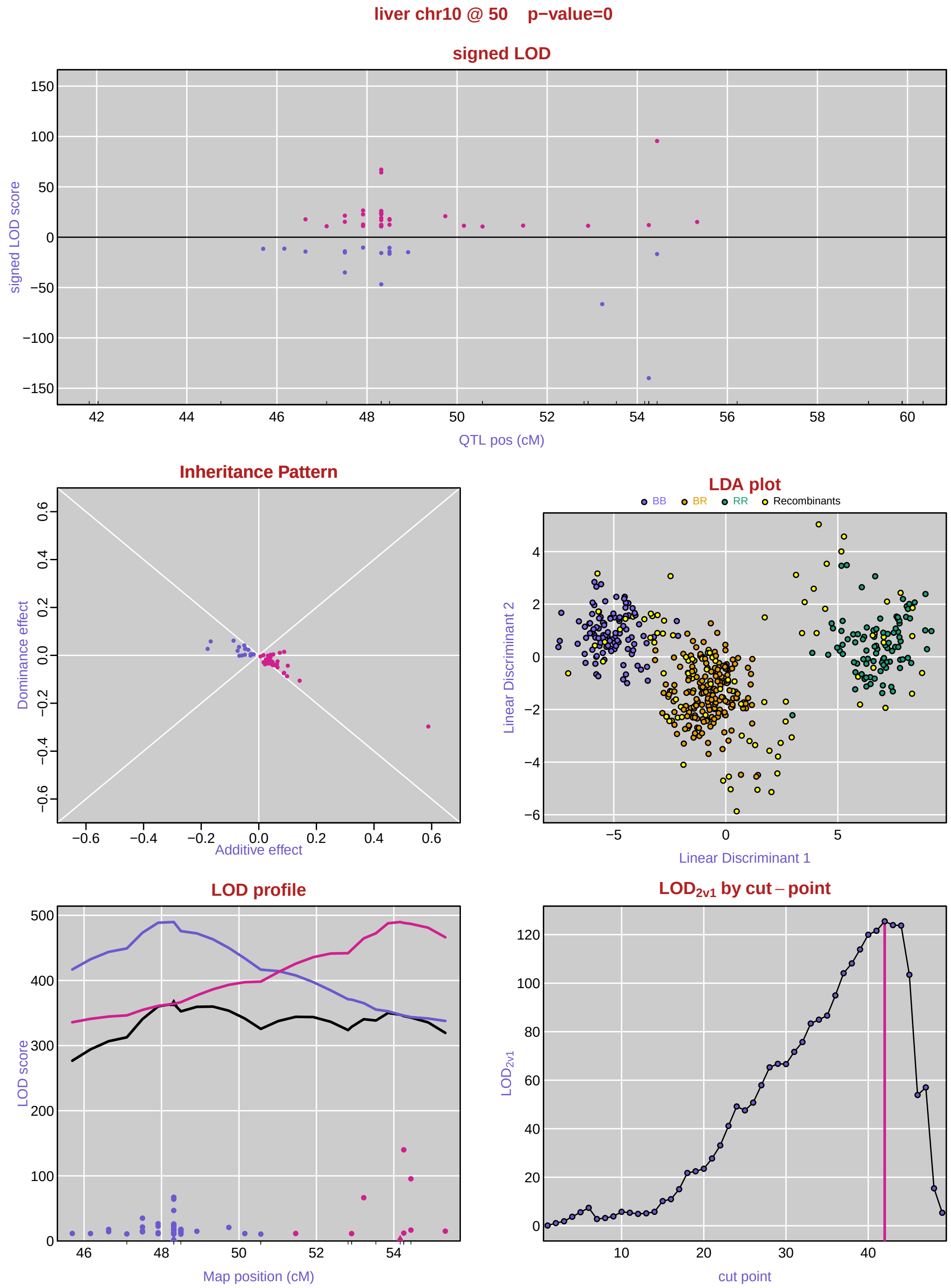



islet chr9 @ 57 p−value=0
signed LOD
signed LOD score
200
150
100
50
0
−50
−100
−150
−200
46
48
50
52
54
56
58
60
62
64
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.3
−0.2
−0.1
0.0
0.1
0.2
0.3
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
2
0
−2
−4
−6
−15
−10
−5
0
5
10
Linear Discriminant 1
LOD profile
LOD score
400
300
200
100
0
50
55
60
Map position (cM)
LOD_{2v1} by cut − point
LOD_{2v1}
40
30
20
10
0
10
20
30
40
cut point

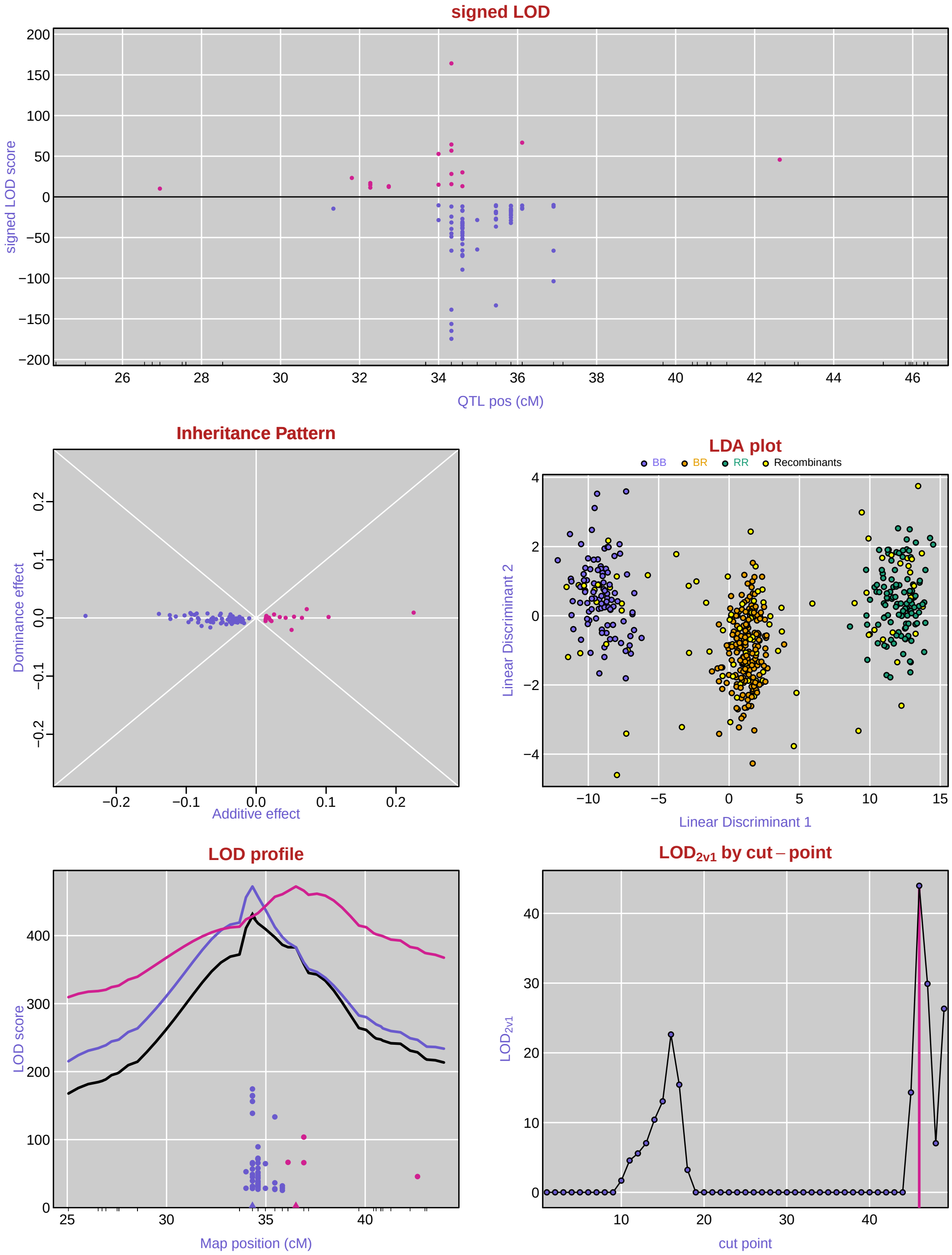

kidney chr10 @ 47 p−value=0
signed LOD
signed LOD score
150
100
50
0
−50
−100
−150
38
40
42
44
46
48
50
52
54
56
58
60
62
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
5
0
−5
−5
0
5
10
15
Linear Discriminant 1
LOD profile
LOD score
500
400
300
200
100
0
40
45
50
55
Map position (cM)
LOD_{2v1} by cut − point
LOD_{2v1}
120
100
80
60
40
20
0
10
20
30
40
cut point

adipose chr10 @ 48   p−value=0

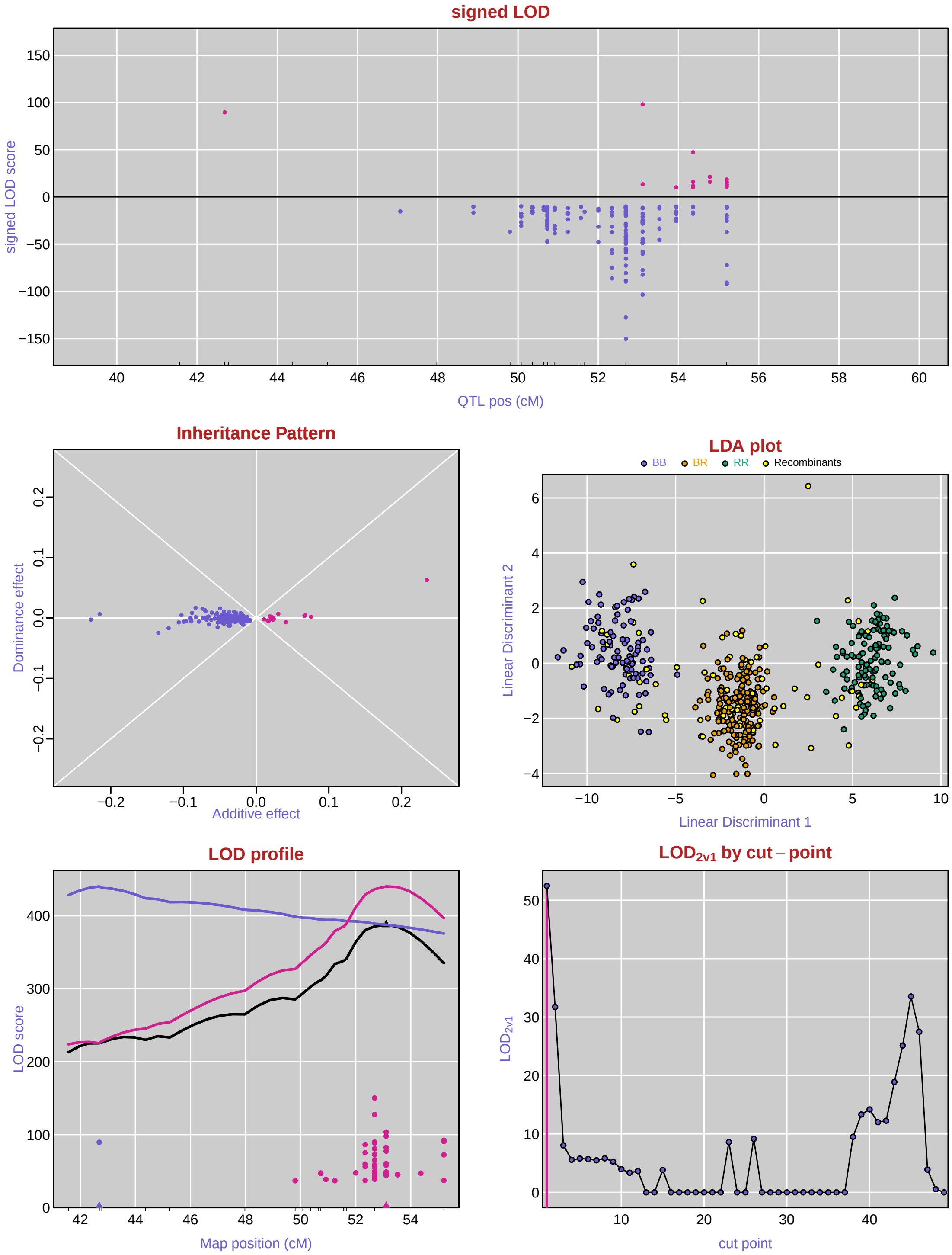


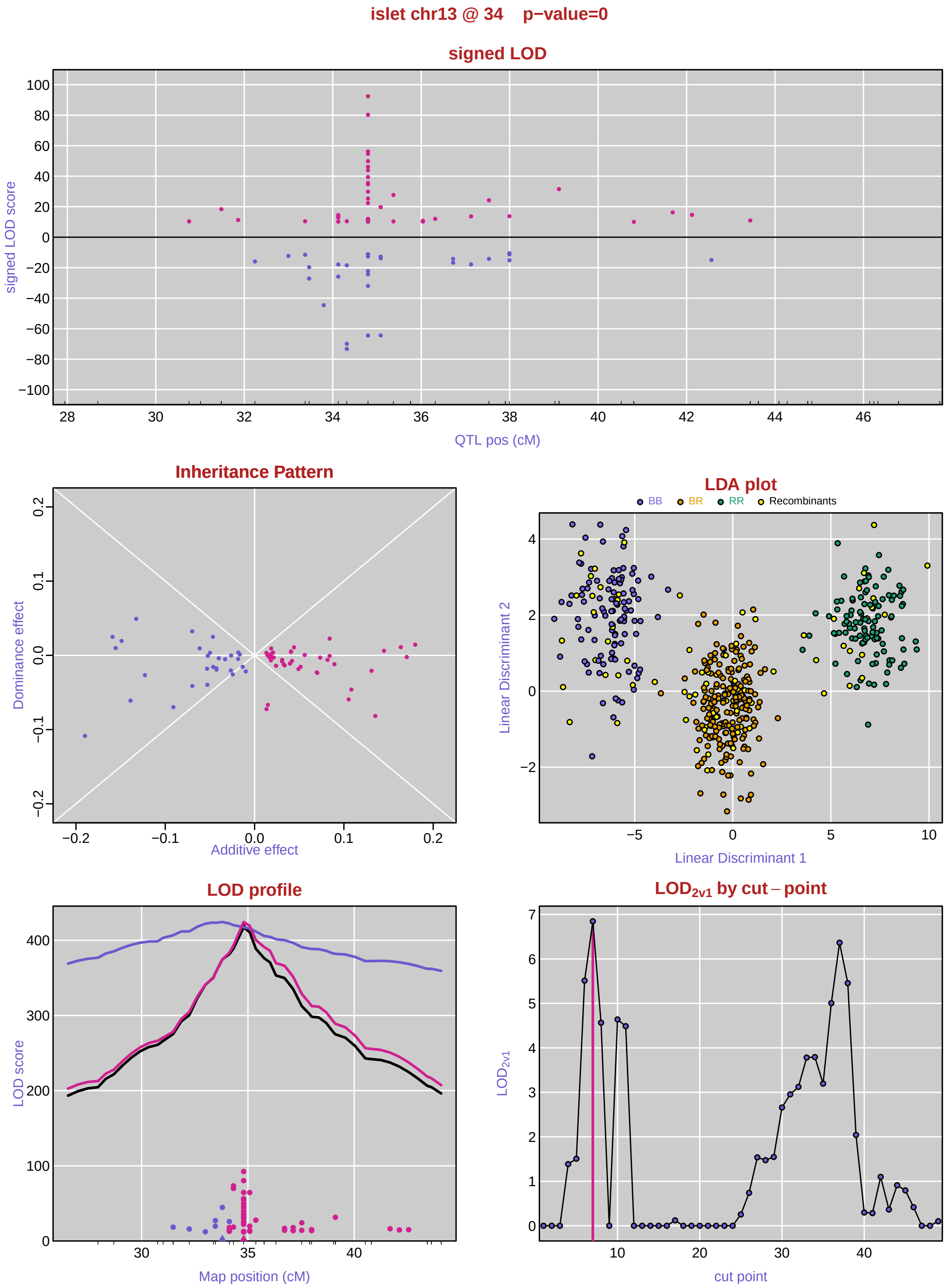

liver chr10 @ 50 p−value=0
signed LOD
signed LOD score
150
100
50
0
−50
−100
−150
42
44
46
48
50
52
54
56
58
60
QTL pos (cM)
Inheritance Pattern
Dominance effect
Additive effect
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
Linear Discriminant 1
4
2
0
−2
−4
−6
−5
0
5
LOD profile
LOD score
500
400
300
200
100
0
46
48
50
52
54
Map position (cM)
LOD_{2v1} by cut − point
LOD_{2v1}
120
100
80
60
40
20
0
10
20
30
40
cut point

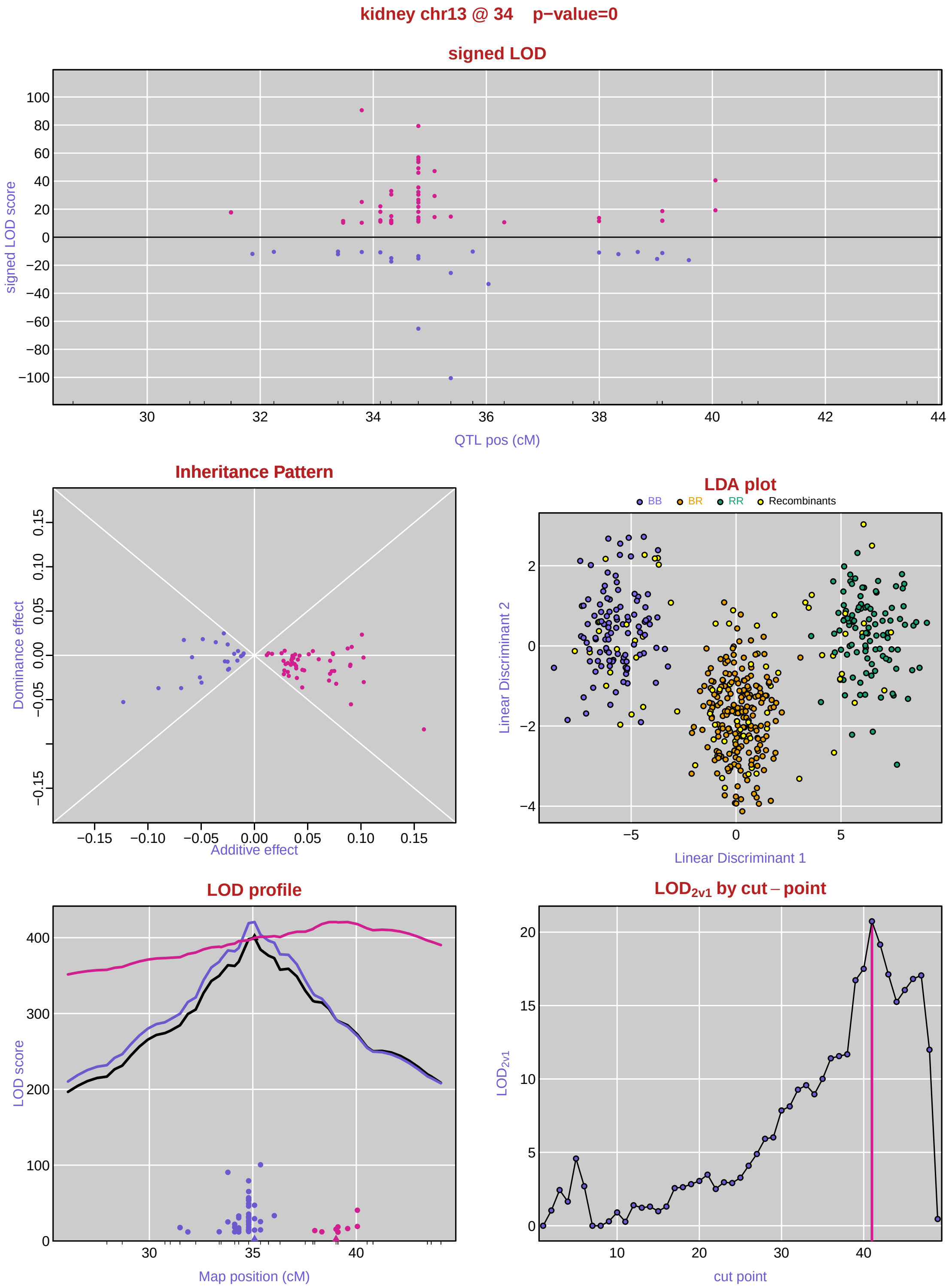
adipose chr11 @ 34 p−value=0



liver chr12 @ 52   p−value=0




islet chr13 @ 34 p−value=0
signed LOD
signed LOD score
100
80
60
40
20
0
−20
−40
−60
−80
−100
28
30
32
34
36
38
40
42
44
46
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.2
0.1
0.0
−0.1
−0.2
−0.2
−0.1
0.0
0.1
0.2
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
4
2
0
−2
−5
0
5
10
Linear Discriminant 1
LOD profile
LOD score
400
300
200
100
0
30
35
40
Map position (cM)
LOD_{2v1} by cut − point
LOD_{2v1}
7
6
5
4
3
2
1
0
10
20
30
40
cut point



kidney chr13 @ 34 p−value=0

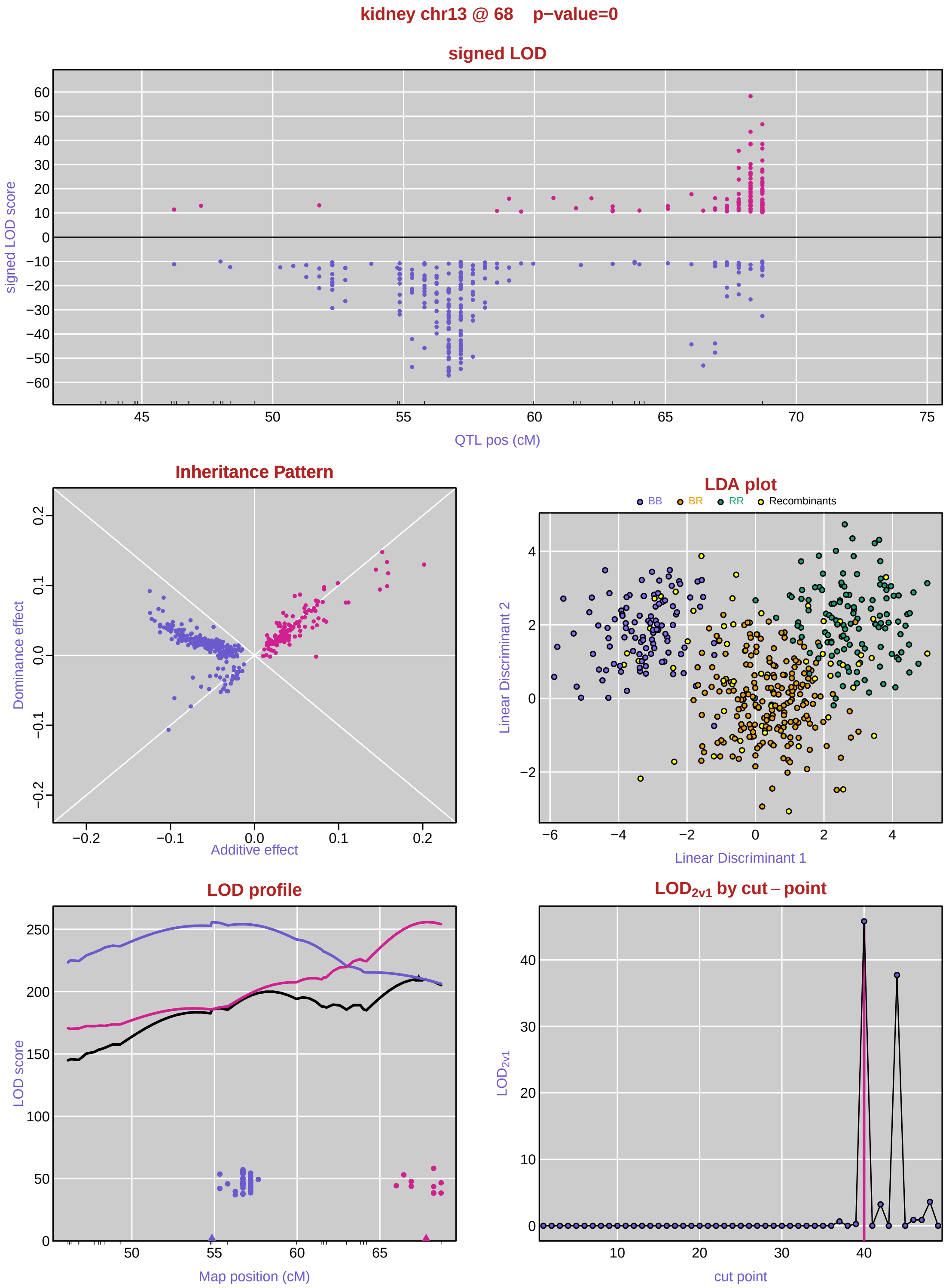
signed LOD

Inheritance Pattern

LDA plot

LOD profile

LOD_{2v1} by cut − point



kidney chr13 @ 68 p−value=0

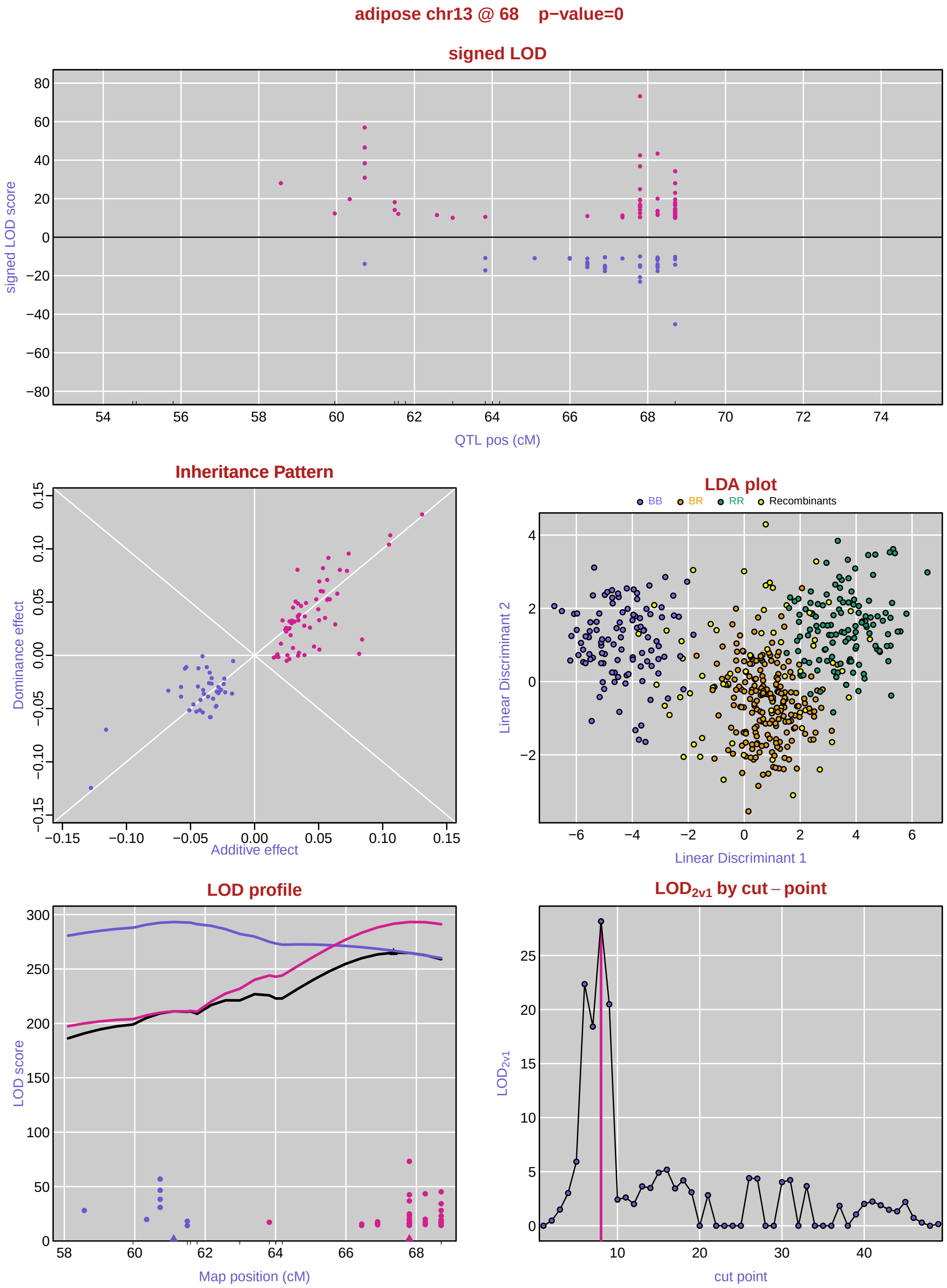
signed LOD

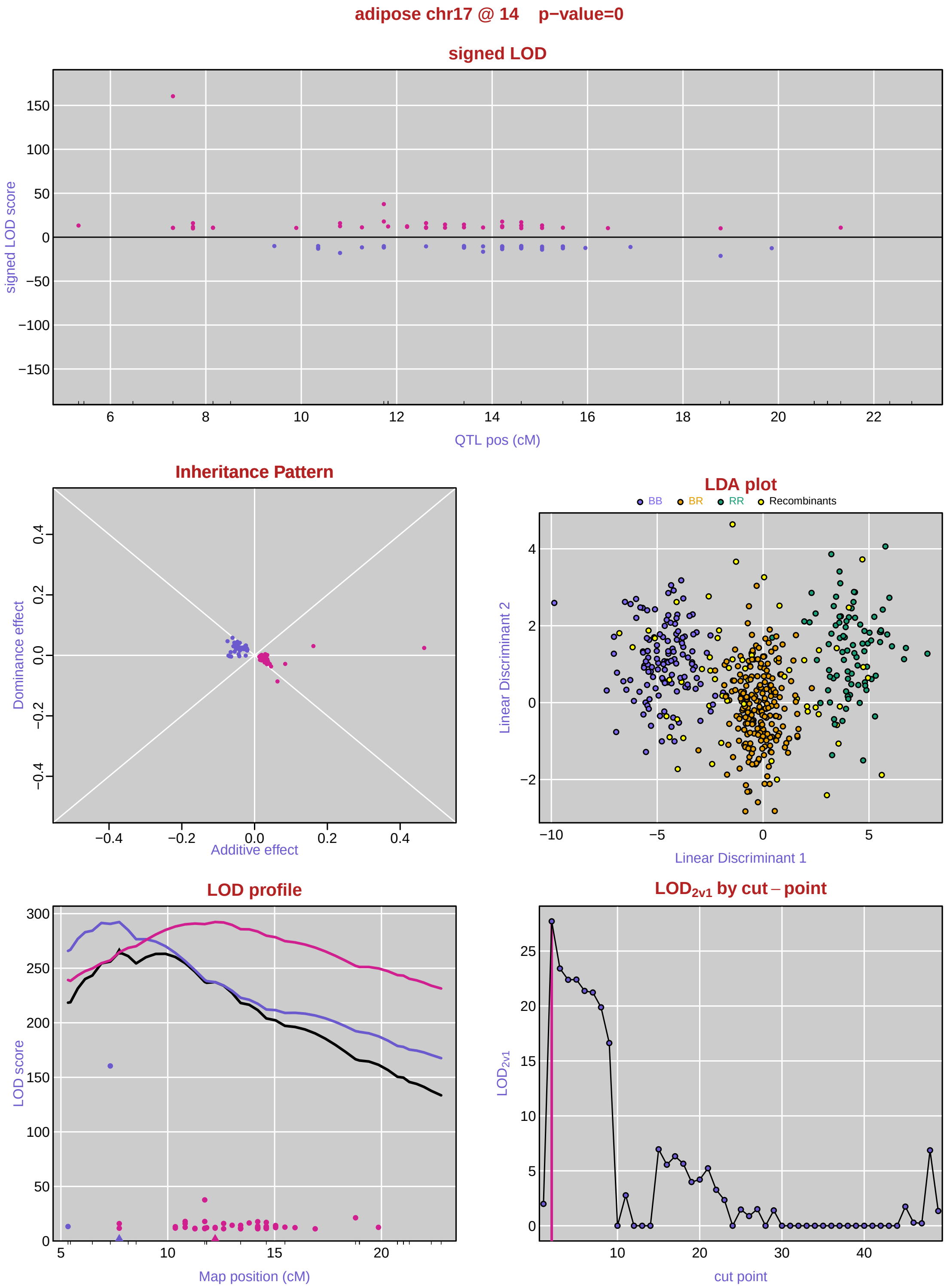
Inheritance Pattern

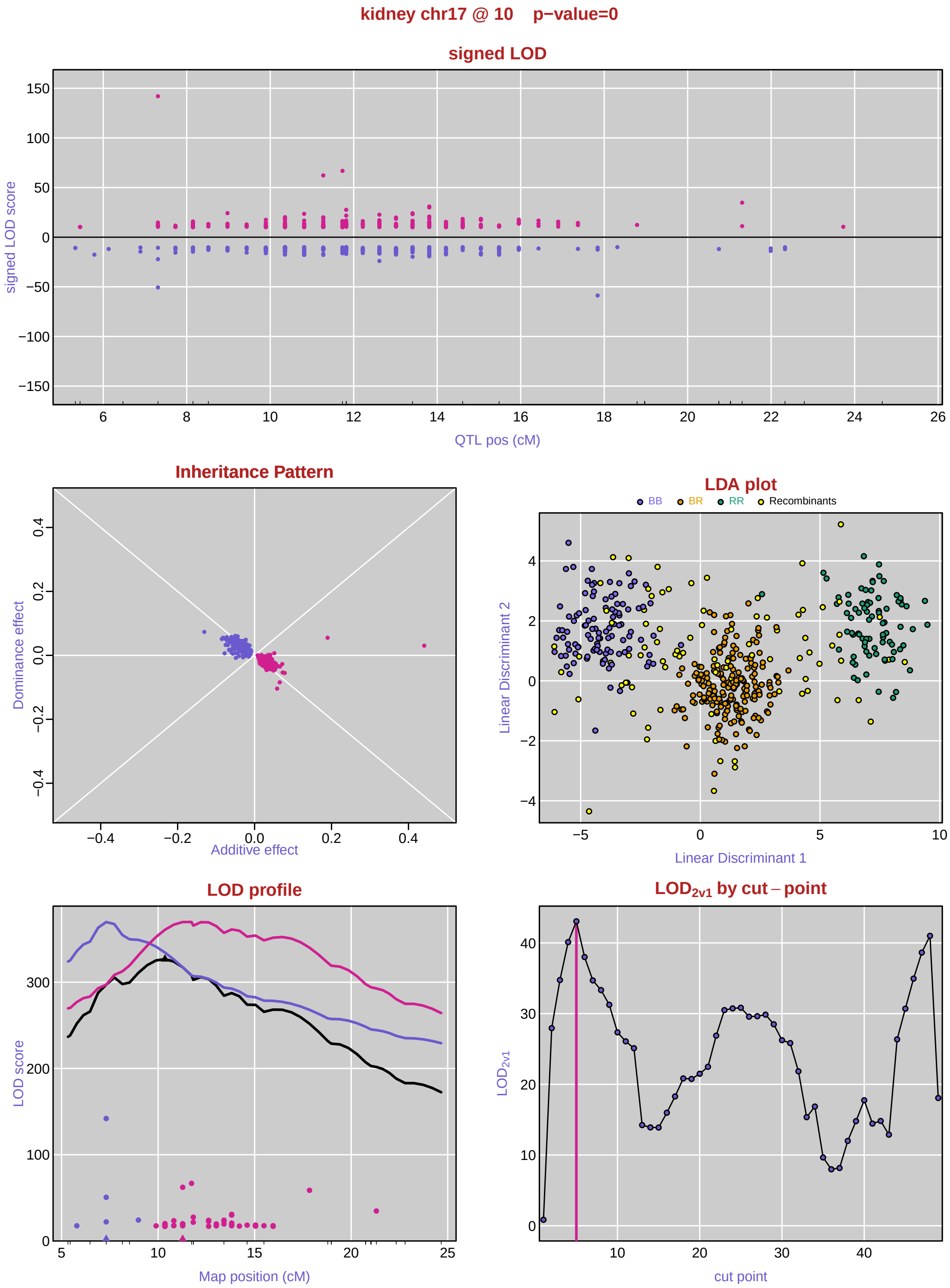
LDA plot

LOD profile

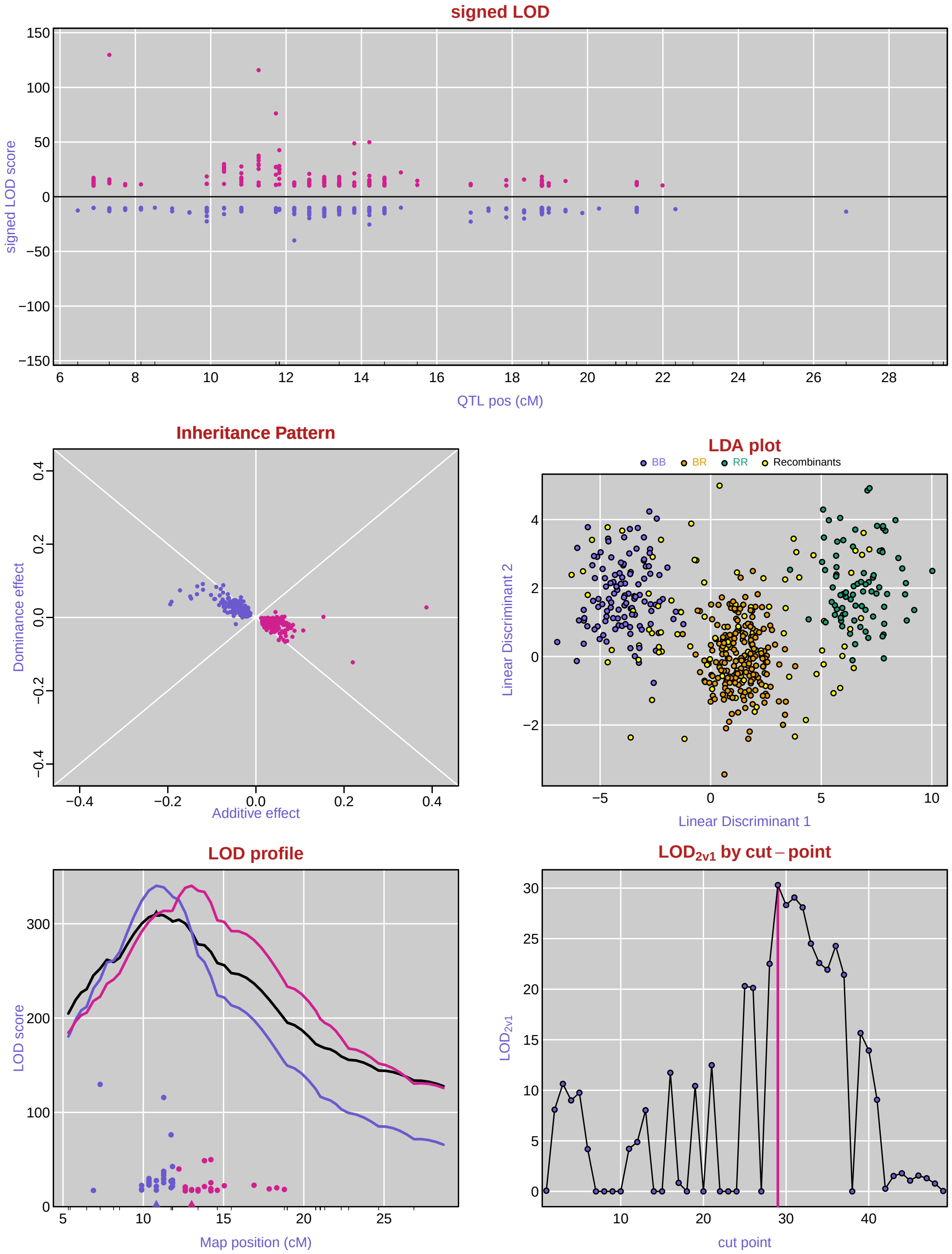
LOD$_{2v1}$ by cut − point


adipose chr13 @ 68 p−value=0
signed LOD
signed LOD score
80
60
40
20
0
−20
−40
−60
−80
54
56
58
60
62
64
66
68
70
72
74
QTL pos (cM)
Inheritance Pattern
Dominance effect
Additive effect
0.15
0.10
0.05
0.00
−0.05
−0.10
−0.15
−0.15
−0.10
−0.05
0.00
0.05
0.10
0.15
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
Linear Discriminant 1
4
2
0
−2
−6
−4
−2
0
2
4
6
LOD profile
LOD score
Map position (cM)
300
250
200
150
100
50
0
58
60
62
64
66
68
LOD$_{2v1}$ by cut − point
LOD$_{2v1}$
cut point
25
20
15
10
5
0
10
20
30
40


adipose chr17 @ 14 p−value=0
signed LOD
signed LOD score
150
100
50
0
−50
−100
−150
6
8
10
12
14
16
18
20
22
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
4
2
0
−2
−10
−5
0
5
Linear Discriminant 1
LOD profile
LOD score
300
250
200
150
100
50
0
5
10
15
20
Map position (cM)
LOD$_{2v1}$ by cut − point
LOD$_{2v1}$
25
20
15
10
5
0
10
20
30
40
cut point


kidney chr17 @ 10 p−value=0
signed LOD
signed LOD score
150
100
50
0
−50
−100
−150
6
8
10
12
14
16
18
20
22
24
26
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.4
0.2
0.0
−0.2
−0.4
Additive effect
−0.4
−0.2
0.0
0.2
0.4
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
4
2
0
−2
−4
Linear Discriminant 1
−5
0
5
10
LOD profile
LOD score
300
200
100
0
Map position (cM)
5
10
15
20
25
LOD2v1 by cut − point
LOD2v1
40
30
20
10
0
cut point
10
20
30
40

liver chr17 @ 18 p−value=0



signed LOD

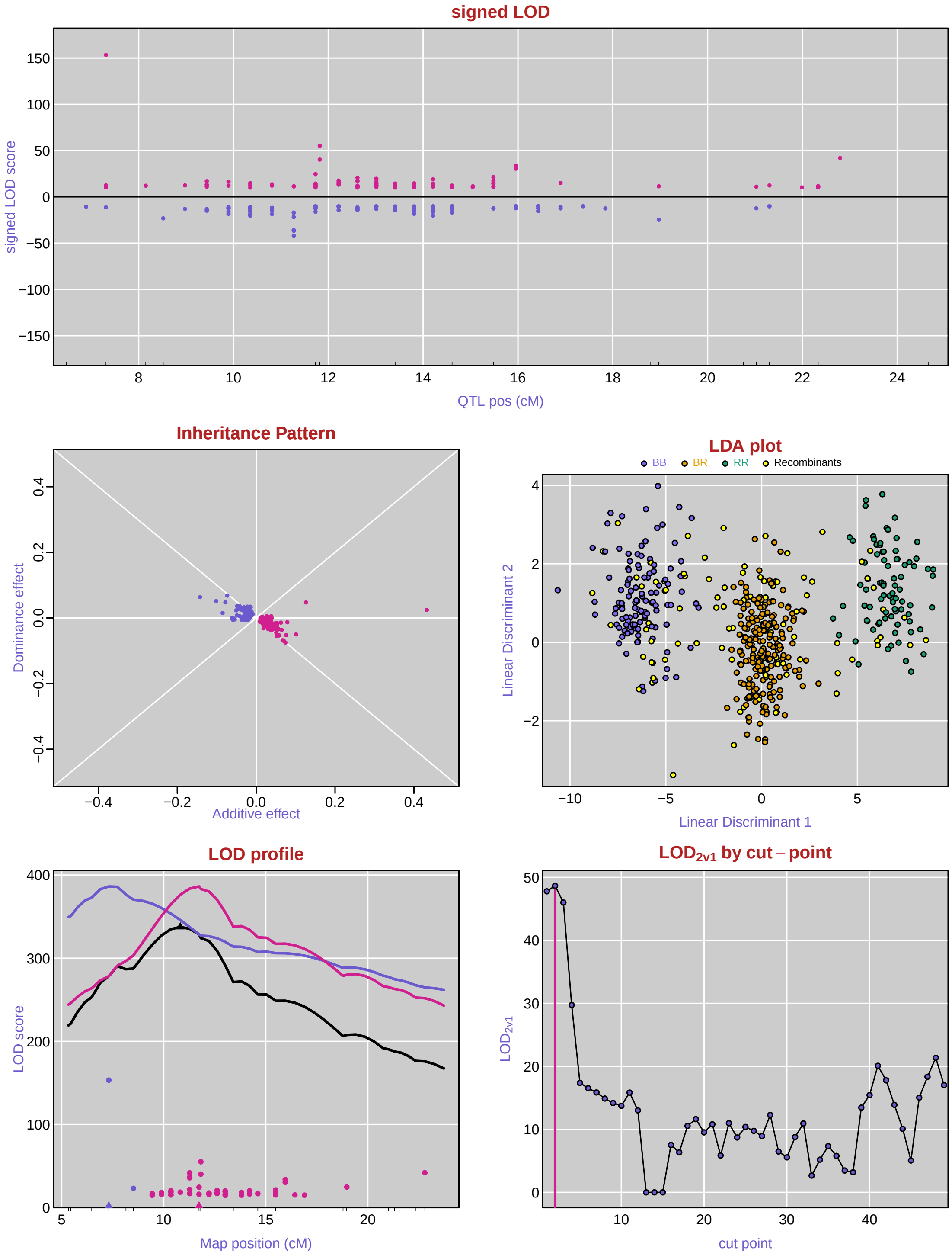
Inheritance Pattern

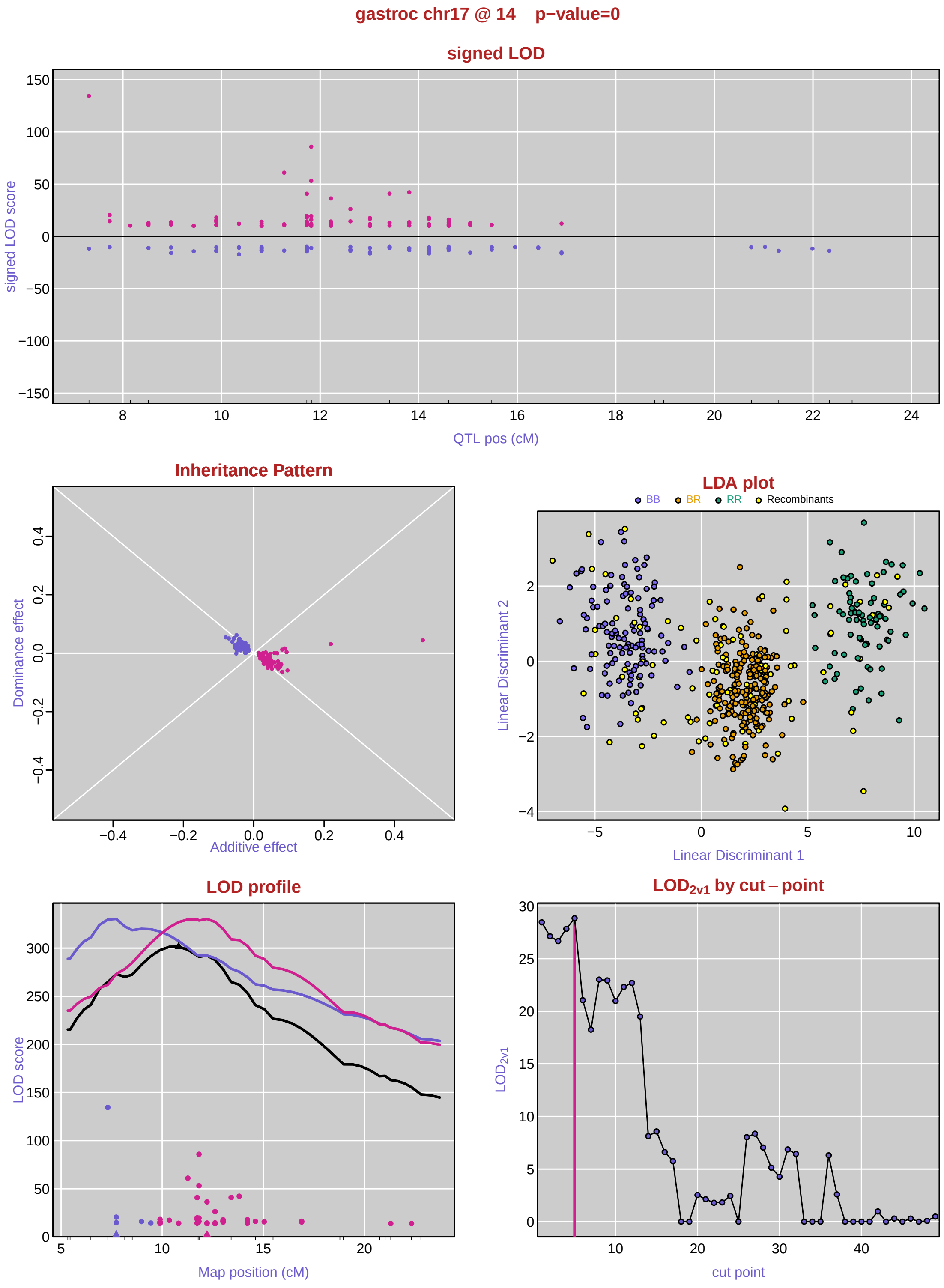
LDA plot

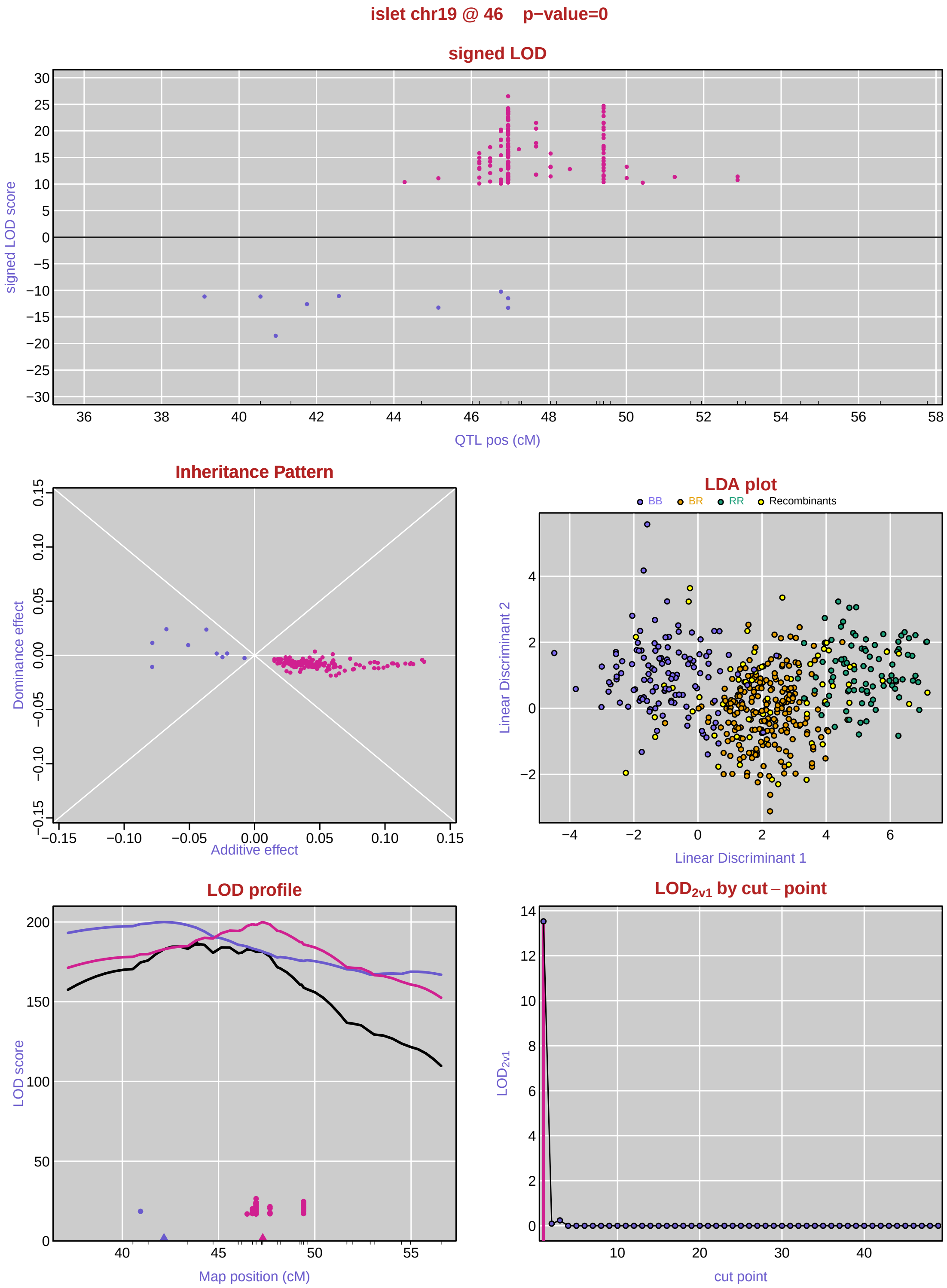
LOD profile

LOD_{2v1} by cut − point

islet chr17 @ 13 p−value=0




gastroc chr17 @ 14   p−value=0
signed LOD
signed LOD score
150
100
50
0
−50
−100
−150
8
10
12
14
16
18
20
22
24
QTL pos (cM)
Inheritance Pattern
Dominance effect
0.4
0.2
0.0
−0.2
−0.4
−0.4
−0.2
0.0
0.2
0.4
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 2
2
0
−2
−4
−5
0
5
10
Linear Discriminant 1
LOD profile
LOD score
300
250
200
150
100
50
0
5
10
15
20
Map position (cM)
LOD$_{2v1}$ by cut − point
LOD$_{2v1}$
30
25
20
15
10
5
0
10
20
30
40
cut point


islet chr19 @ 46 p−value=0
signed LOD
signed LOD score
QTL pos (cM)
Inheritance Pattern
Dominance effect
Additive effect
LDA plot
BB
BR
RR
Recombinants
Linear Discriminant 1
Linear Discriminant 2
LOD profile
LOD score
Map position (cM)
LOD2v1 by cut − point
LOD2v1
cut point

4 SOFTWARE: INTRODUCTION TO R/QTLPVL

The methods described in this thesis have been implemented as an R package, *qtlpvl*. This package focuses on QTL mapping with multiple traits and on testing of pleiotropy vs. close linkage when multiple traits map near each other, particularly for the of a trans-eQTL hotspot. We provide both exploratory plots and formal statistical tests aimed at dissecting trans-eQTL hotspots.

Data

To illustrate the package, we use the genotype data from the `listeria` data set (an F_2 population) provided with R/qtl along with a set of simulated phenotypes, included in qtlpvl as the data set `fake.phenos`. The phenotypes were simulated using two markers from chromosome 1 as QTL, with the first QTL having an additive allelic effect, and with one of the alleles at the second QTL being strictly dominant. There are 10 phenotypes. The first 5 are controlled by the first QTL, and the other 5 traits are controlled by the second QTL (and with a negative and larger effect). The 10 phenotypes were generated with these QTL effects plus independent, normally distributed residual variation. Treating these traits as gene expression measurements, we assigned genomic positions at random. The phenotype data is stored in matrix `fake.phenos` and their positions are stored in data frame `fake.probepos`.

We load the qtlpvl package, the simulated phenotype data, and the `listeria` data set as follows. (The R/qtl package is automatically loaded when qtlpvl is loaded.)

```
library(qtlpvl)
```

```
#> Loading required package: qtl
#> Loading required package: Rcpp
#> Loading required package: plyr
```

```
data(fake.phenos)
data(listeria)
```

We will first perform single-trait analysis. For each trait, we plot its LOD score versus its QTL position, when the LOD is bigger than a threshold (the default value is 3). To do so we first calculate the QTL genotype probabilities for `listeria` and then use the function `plotLOD`. We will only scan chromosome 1, and we set the variable `chr=1` for use throughout.

```
listeria <- calc.genoprob(listeria, step=1)
chr <- 1
plotLOD(Y=fake.phenos, cross=listeria, chr=chr)
```

Joint mapping: `scanone.mvn`

Having the cross object `listeria` and phenotype trait matrix `fake.phenos` loaded, we can perform joint, multi-trait QTL mapping with the function `scanone.mvn`. This assumes a multivariate normal model. `chr` is set to be “1”, which indicates the mapping is only on

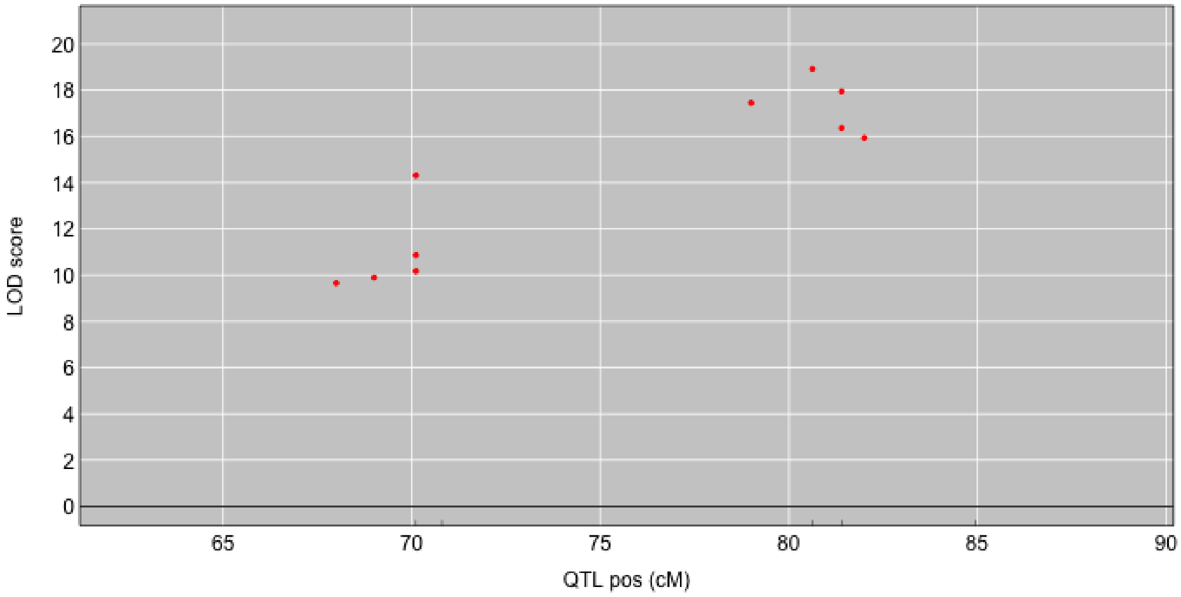


Figure 4.1: LOD score and QTL positions for each of the simulated traits

chromosome 1. Function `scanone.mvn` has two more parameters `addcovar` and `intcovar` that can be used to control for additive and interactive covariates. `intcovar` will also be used as additive covariates during mapping, and there is no need to manually add them into `addcovar`.

The result of joint mapping is a data frame with class `scanone` (that is, in the same form as returned by the R/qtl function `scanone`). We can then use `summary` and `plot` to look at the results.

```
out <- scanone.mvn(cross=listeria, Y=fake.phenos, chr=chr)
summary(out)
```

```
#>           chr pos  lod
#> c1.loc78   1  78 63.5
```

```r
plot(out)
```

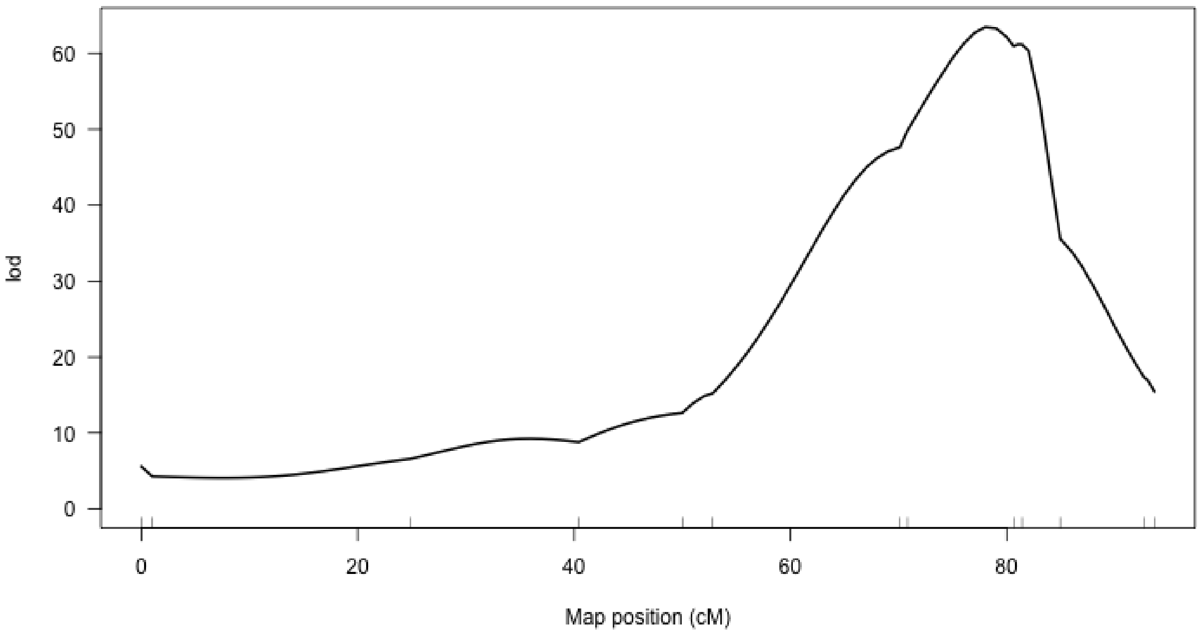


Figure 4.2: Joint mapping result under multivariate normal model for all the simulated traits

With this single-QTL analysis, the QTL is estimated to be at 78.0 cM, with LOD score 63.5. This is closer to the second QTL, as in our simulation, the second QTL had a larger effect than the first QTL.

Statistical Testing

test of 1 vs 2 QTL: `testpleio.1vs2`

We now turn to the question of pleiotropy vs. close linkage. Specifically, we test the hypotheses: * H_0: There is only one QTL affecting all the phenotypes; * H_1: There are two QTL, with each phenotype affected by one or the other QTL.

We again assume that the residual variation in the traits follows a multivariate normal distribution. We perform joint, multivariate QTL mapping to locate the single QTL, under the null hypothesis that there is one pleiotropic QTL affecting all the traits. We then run single-trait analysis on each trait and find the trait-specific QTL, sort the traits by their estimated QTL position and search for the best separation of the traits into two groups, where the first group (contains the first several traits) is controlled by the right QTL and the second group (contains the rest traits) is controlled by the left QTL. The LOD score for this two-QTL model is then subtracted by the LOD score for the single-QTL model, to give the final test statistic, LOD_{2v1}.

To get the null distribution of the statistics, we have two methods: * parametric bootstrap: generate data from the parameters estimated under the null hypothesis * stratified permutation: randomly permute the genotype data versus phenotype data within each QTL genotype group,

We repeat the entire procedure on data from either method and save the test statistics. The P-value is calculated from this empirical distribution.

The function `testpleio.1vs2` is used to do this test. Input parameters: * `cross`: An R/qtl cross object. * Y: matrix of multiple traits, with samples in the row, traits in the column. * chr: Character string referring to chromosome of interest by name. * region.l, region.r: left and right bounds for the interval of interest. * int.method: method to calculate the interval of interest if region.l and region.r is not specified (either bayes or 1.5lod) * search.method: method for searching the two-QTL model (either fast or complete; the default is fast). * RandomStart: When search.method is fast, indicates whether to use a random starting point for the search over two-QTL models; the default is TRUE. * RandomCut: Indicates whether to use random cutting or not when there are traits mapped to the same location. The default is FALSE (traits mapped to the same location will be bound together). * simu.method: method for determining the null distribution (parametric or permutation; the default is parametric). * n.simu: number of simulations for p-value. * tol: Tolerance value for the qr decomposition in fitting the linear models with lm. * addcovar, intcovar: Optional additive and interactive covariates to include when mapping.

This function will return a list with class testpleio.1vs2. We can use summary and plot to study the results. The result of summary includes the estimated QTL position and LOD score for the single-QTL and two-QTL models, as well as the test statistic LOD_{2v1} and its P-value.

As a quick illustration, we will perform just 10 bootstrap simulation replicates.

```
obj <- testpleio.1vs2(cross=listeria, Y=fake.phenos, chr=chr, n.simu=10,
                      region.l=60, region.r=90,
                      search.method="complete")
summary(obj)
```

```
#> Single QTL model: lod 63.46, pos 78.00 cM

#> Two QTLs model: lod 81.66, pos 70.11 cM, 81.40 cM

#>   Traits influenced by the left QTL: 1 2 3 4 5

#>   Traits influenced by the right QTL: 6 7 8 9 10

#> Difference of LOD score: 18.20

#> P-value is: 0 (from 10 simulations)
```

```
plot(obj)
```

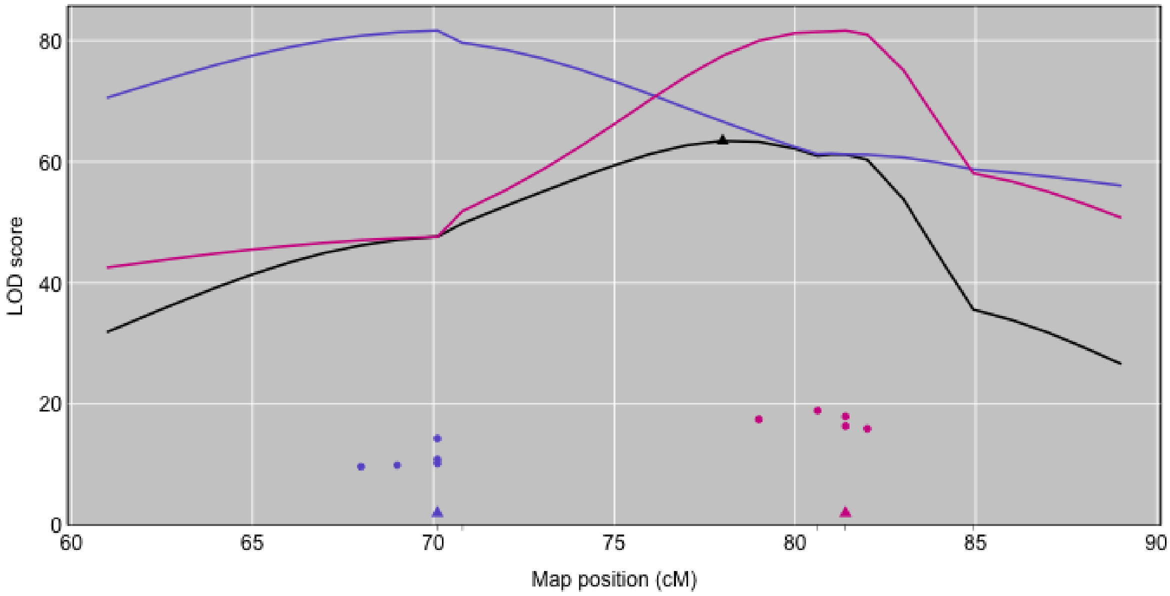


Figure 4.3: LOD profile plot for result of test1vs2 on simulated traits

The above figure shows the joint mapping result in black and the profile LOD curves for each of the two QTL under H_1:

- red line: $L_1(\lambda_1) = \max_{\lambda_2} \text{LOD2}(\lambda_1, \lambda_2)$

- blue line: $L_2(\lambda_2) = \max_{\lambda_1} \text{LOD2}(\lambda_1, \lambda_2)$

Triangles indicate the estimated positions of the QTL and solid points indicate QTL positions and LODs from single trait mapping.

We use `plottrace` to see how LOD_{2v1} changes when we move the cut-point of the left versus right group.

```
plottrace(obj)
```

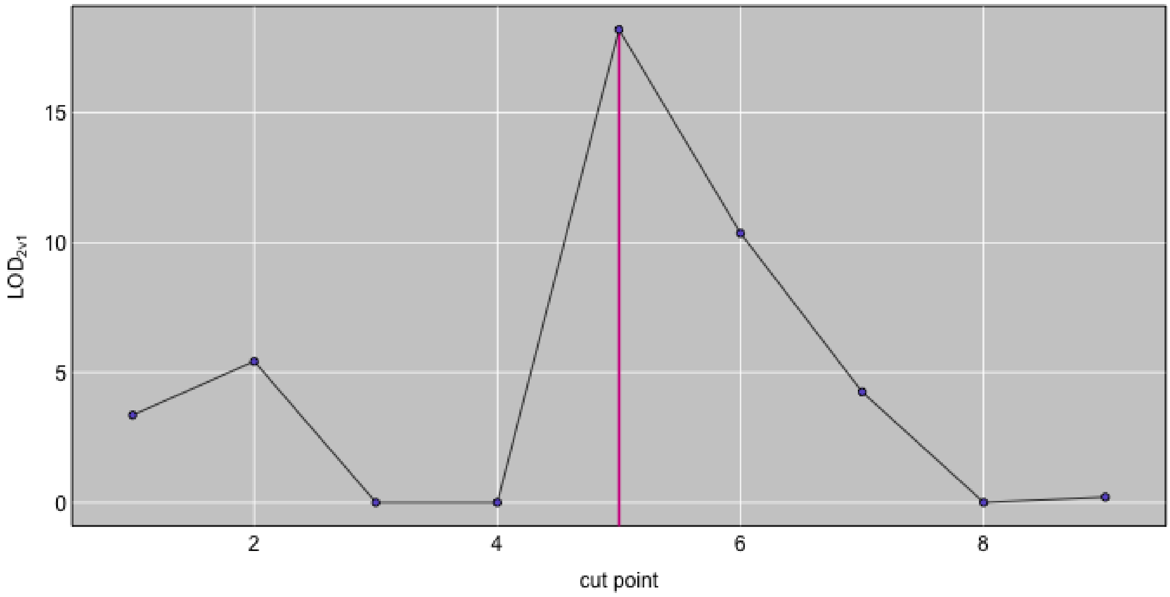


Figure 4.4: plot of LOD_{2v1} by cut-point

There are 7 possible QTL positions from the above plot with the results of single-trait mapping. Thus we have 6 possible ways of grouping these positions into the left versus right group. The 6 dots indicate the possible cut-points and the corresponding value for

LOD_{2v1}.Thus when the first 5 traits and the last 5 traits are grouped separately, the LOD_{2v1} is the biggest, it is our best two QTL model. This figure can be used as a diagnostic, to see how our inferred two-QTL model compares to other possible models.

test of 1 vs p QTL: `testpleio.1vsp`

We could also test the hypotheses: * H_0: There is only one QTL affecting all the phenotypes; * H_1: There are p QTL, each affecting one of the phenotypes (i.e., each phenotype is allowed to have its own QTL).

This function has fewer parameters than `testpleio.1vs2`. The arguments are `cross`, `Y`, `chr`, `addcovar`, `intcovar` and `n.simu`, and their usage is the same as before.

This function will return a list with class `testpleio.1vsp`. We can again use `summary` and `plot` to look at the results. The summary includes the estimated QTL positions and LOD score for the single-QTL and p-QTL models, as well as test statistics LOD_{pv1} and its P-value.

```
obj2 <- testpleio.1vsp(cross=listeria, Y=fake.phenos, chr=chr, n.simu=10)
summary(obj2)
```

```
#> Single QTL model: lod 63.46, pos 78.00 cM
#> Multiple QTLs model: lod 82.28
#>          Trait      POS       LOD
#> phenos1      1 70.11204 10.862842
#> phenos2      2 70.11204 10.174272
#> phenos3      3 69.00000  9.895711
#> phenos4      4 70.11204 14.309338
#> phenos5      5 68.00000  9.652426
#> phenos6      6 80.62324 18.919347
#> phenos7      7 79.00000 17.460231
#> phenos8      8 81.39623 17.944594
#> phenos9      9 82.00000 15.928201
#> phenos10    10 81.39623 16.367111
#>
#> Difference of LOD score: 18.82
#> P-value is: 0 (from 10 simulations)
```

```
plot(obj2)
```

Exploratory plots

plot of genetic pattern: `plotGenetpattern`

The function `plotGenetpattern` takes two kinds of input parameters: * a phenotype matrix `Y` and a genotype vector `genotype`, the later being genotypes at a common QTL. * a phenotype matrix `Y` and a cross object `cross` and chromosome number `chr`. This is used when there is uncertainty in the QTL locations for the traits. In this case, each trait is mapped separately and the genotypes at corresponding estimated QTL position are used.

We'll use the labels B and R for the alleles in the cross. (These correspond to the strains in the main application that motivated this work). The additive QTL effect is defined as

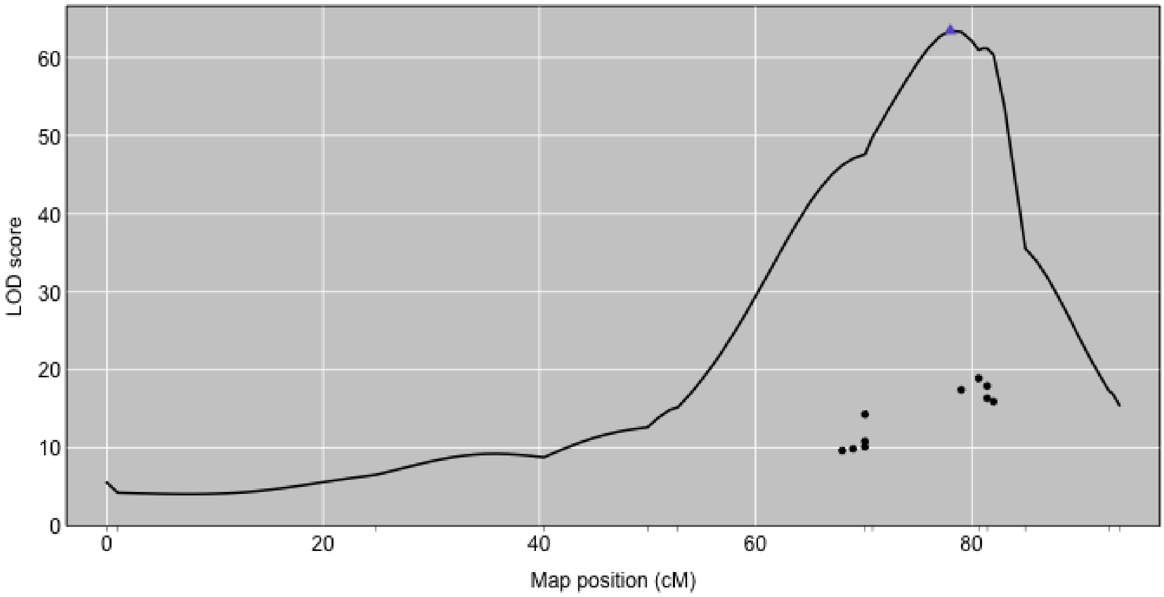


Figure 4.5: plot for result of testpleio.1vsp, showing joint mapping LOD curve and max LOD score for each trait separately

$a = (\mu_{RR} - \mu_{BB})/2$. The dominance effect is $d = \mu_{BR} - (\mu_{BB} + \mu_{RR})/2$. When plotted against each other, traits with pure additive effects are near the x-axis ($d = 0$) and traits with dominant effects are along the diagonals, $d = a$ (that is, $\mu_{BR} = \mu_{RR}$) and $d = -a$ (that is, $\mu_{BR} = \mu_{BB}$).

Here, we'll make plots by each method: with a common QTL and with separate QTL for each trait.

```
par(mfrow=c(1,2))

qtlpos <- max(out)$pos

m <- find.pseudomarker(listeria, chr, qtlpos, "prob", addchr=FALSE)

qtlgeno <- apply(listeria$geno[[chr]]$prob[,m,], 1, which.max)
```

```
plotGenetpattern(Y=fake.phenos, genotype=qtlgeno, main="by common QTL genotype")
plotGenetpattern(Y=fake.phenos, cross=listeria, chr=chr,
                 main="by individual QTL genotype")
```

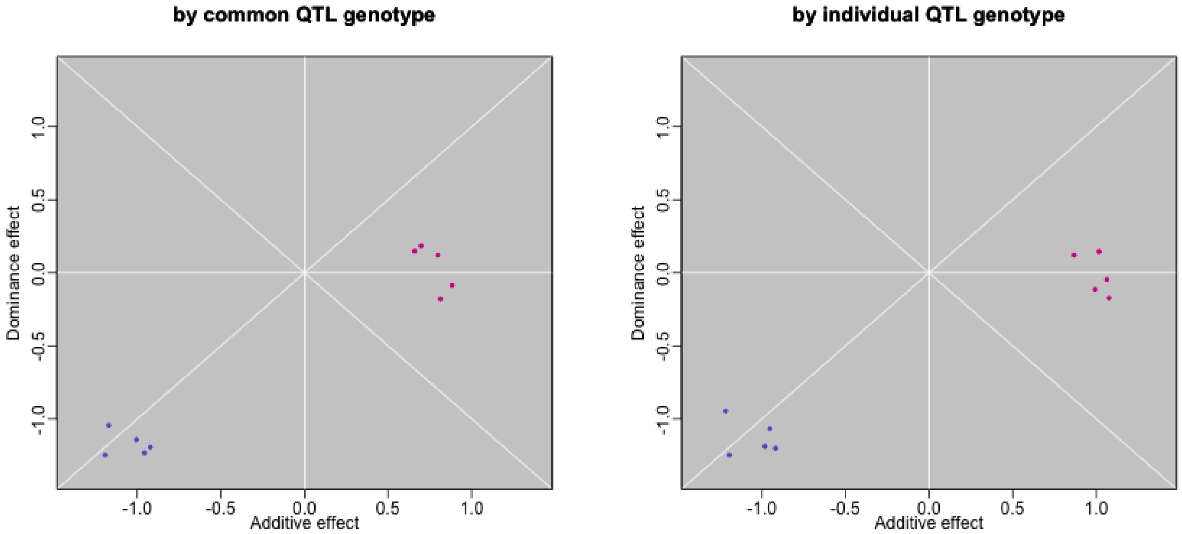


Figure 4.6: plot of inheritance pattern for all the simulated traits

In both plots, we see one set of traits for which the QTL is dominant, with the B allele associated with larger average trait value ($d \approx -a$ and $a < 0$), and another set of traits for which the QTL alleles are additive ($d \approx 0$), with the R allele associated with larger average trait value ($a > 0$). This is evidence for two QTL in the region, with different inheritance patterns.

plot signed LOD score: `plotLODsign`

The function `plotLODsign` gives a second exploratory plot to display the direction of the QTL effects. The input arguments are `cross`, `Y`, and `chr`, as well as `addcovar` and `intcovar`, all as above.

For each trait, the estimated additive QTL effect is used as the sign of the LOD score. We first run single-trait mapping to obtain LOD scores and estimated QTL positions. We use the R/qtl function `argmax.geno` to obtain imputed QTL genotypes, and estimated the additive QTL effect as $\mu_{RR} - \mu_{BB}$. Only traits with LOD score bigger than a threshold will be displayed, the default value of the threshold (`LOD.threshold`) is 3.

```
plotLODsign(Y=fake.phenos, cross=listeria, chr=chr)
```

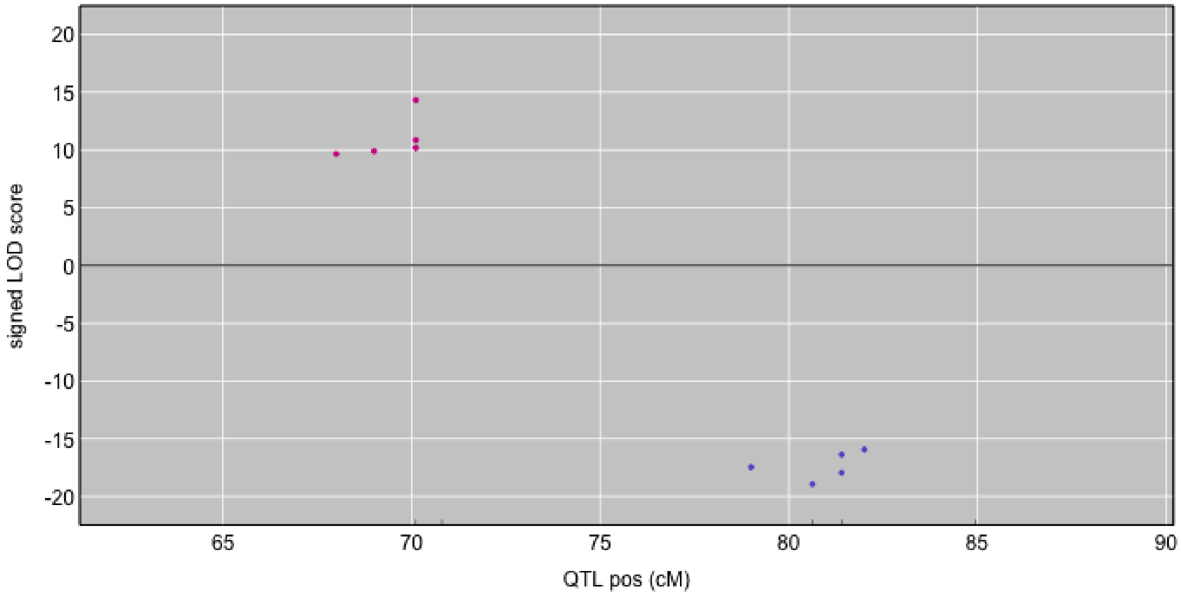


Figure 4.7: plot of signed LOD score for all the simulated traits

We see that the traits mapping to the left side of the region all show positive QTL effects, with the R allele associated with larger average trait, while the traits mapping to the right side all show negative QTL effects, with the B allele associated with larger average trait. This is again evidence for two QTL in the region.

Tick marks at the bottom of the plot indicate the positions of genotyped markers. We can pass these values to parameter `map`, but the default is pull this information from the input `cross` object on the given chromosome, `chr`.

If we had already performed QTL analysis on the traits, we can pass the signed LOD scores and their mapped positions to `LODsign` and `maxPOS`, respectively. By skipping the QTL analysis of the traits, this can speed up the procedure. The use of this feature will be shown at the last section.

Linear discriminant plot: `plottrans.LDA`

In our final exploratory plot, we identify individuals with no recombination event in the region of the QTL. For these individuals, we will know their QTL genotype. We apply linear discriminant analysis to the traits with the QTL genotype as the label and plot the first two linear discriminants. If the QTL effect is strong, this should show three distinct clusters. We then calculate the linear discriminants for the individuals that show a recombination event in the region, using the same coefficients, and add them as points in the plot. If the recombinant individuals fall within the clusters defined by the non-recombinants, this is consistent with there being a single QTL. If, on the other hand, the recombinants look distinctly different from the non-recombinants, this suggests more than one QTL.

In the following code, we first find which individuals have no recombinant event, and then call the function `plottrans.LDA` with phenotype matrix `Y`, QTL genotype `qtlgeno` and nonrecombinant IDs nonrecomb.

```
out <- out[out$chr==chr, ]

m <- which(out$pos >= qtlpos-5 & out$pos <= qtlpos+5)

g <- apply(listeria$geno[[chr]]$prob[,m,], 1:2, which.max)

nonrecomb <- which(sapply(apply(g, 1, unique), length) == 1)

names(nonrecomb) <- rownames(fake.phenos)[nonrecomb]

plottrans.LDA(Y=fake.phenos, qtlgeno, nonrecomb)
```

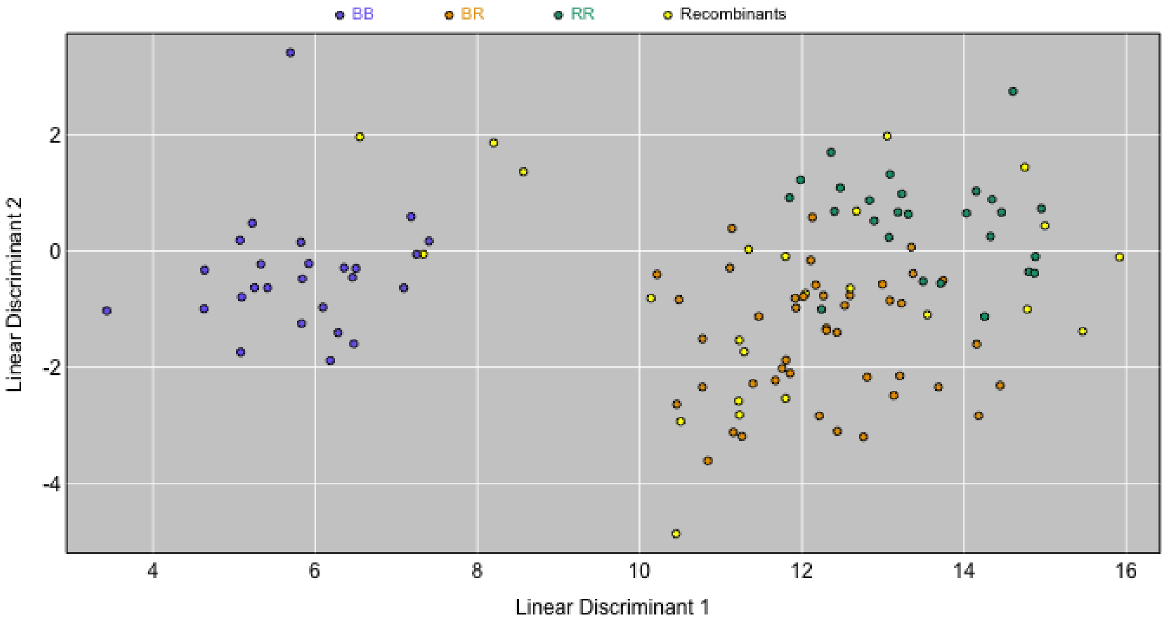


Figure 4.8: LDA plot for the simulated traits, non-recombinant mice are colored as yellow

In the above plot, each point is a individual colored by its QTL genotype if it is non-recombinant, or yellow if it was recombinant. This particular example is not terribly informative about whether there is one or two QTL.

Analysis of a trans-eQTL hotspot with gene positions

Lastly, we will demonstrate the entire procedure, as recommended for the analysis of expression data. We start by mapping each phenotype individually, then count the number of eQTL in windows of 10~cM. When there seems to be a trans-eQTL hotspot, we can build an object `transband` with the function `make.transbands`. This object contains basic information on the transband: chromosome and position of the transband, number of eQTL in the transband, and an estimate of the QTL position and LOD score under the multivariate normal assumption. It also contains all of the information needed for the exploratory plots: LOD scores and estimated QTL positions, estimated additive and dominance effects, QTL genotypes, and the IDs for the non-recombinant individuals. By extracting such information from this object, we can run all of the statistical tests and create the exploratory plots as a batch.

```
data(fake.probepos)
phenoname <- colnames(fake.phenos)
listeria$pheno <- data.frame(listeria$pheno, fake.phenos)
out <- scanone(listeria, pheno.col=phenoname, chr=1:19)
out1 <- convert_scan1(out, phenoname, chr=1:19)

marker.info <- get.marker.info(listeria, chr)
out.count <- count.trans(out1, fake.probepos, chr, marker.info)

par(mfrow=c(2,3))
```

```
plot(out.count, main="Count of trans-eQTL")

trans <- make.transbands(out1, fake.probepos, cross=listeria, chr=1:19,
                         mlratio = fake.phenos, lod.thr = 5,
                         trans.cM = 5, kernal.width = 1,
                         window.cM = 10, trans.count.thr = 0,
                         regn.cM = 5)
transband <- trans[[1]]

geno <- attr(transband, "geno")
nonrecomb <- attr(transband, "nonrecomb")
out <- attr(transband, "out")
map <- pull.map(listeria)

plot(obj, main="LOD profile")
plottrace(obj, main="Trace")
plotLODsign(maxPOS=out$pos, LODsign=sign(out$eff.a)*out$lod1, map=map[[chr]],
            main="Signed LOD")
plotGenetpattern(a=out$eff.a, d=out$eff.d, main="Inheritance Pattern")
plottrans.LDA(Y=fake.phenos, geno, nonrecomb, main="LDA plot")
```

The results are as before, with one added plot in the top left: a count of traits with a trans-eQTL in a sliding 10 cM window.

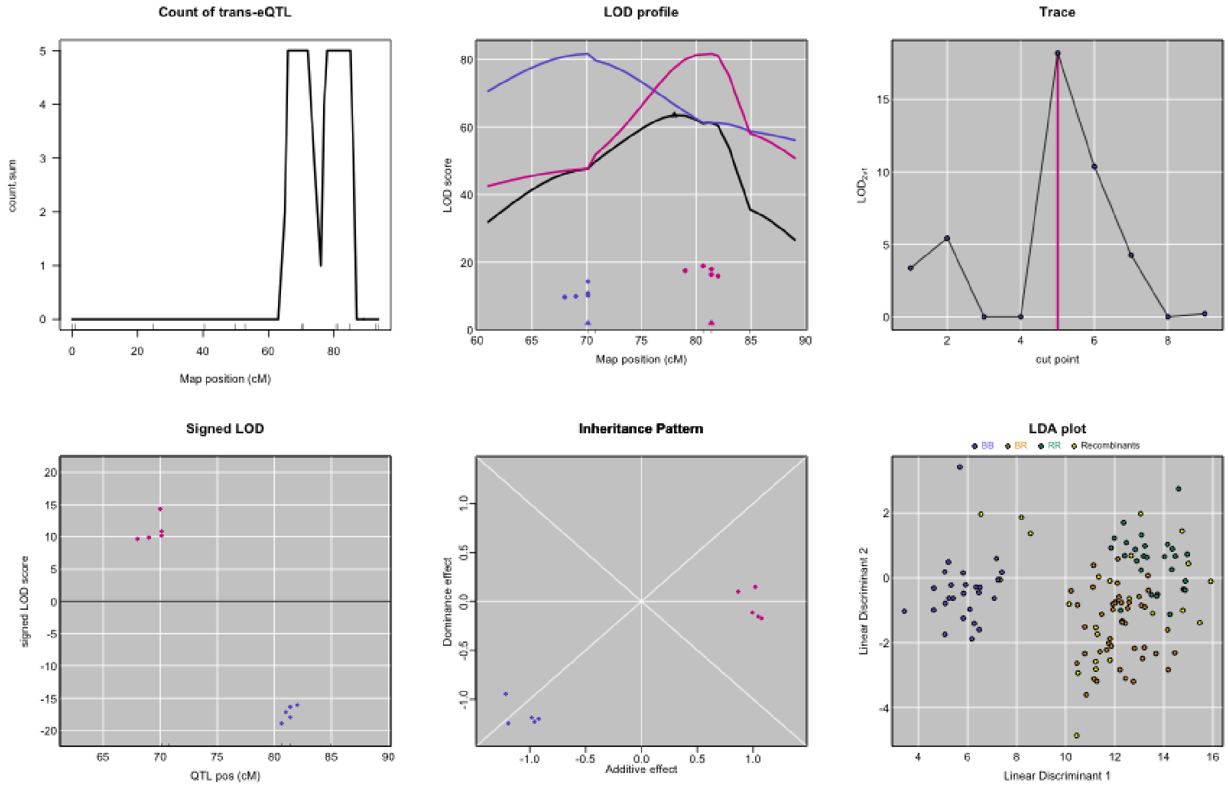


Figure 4.9: Summary plot for a trans-eQTL hotspot using simulated data

5 DISCUSSION

We proposed exploratory methods and formal statistical inference methods for dissecting trans-eQTL hotspots, and illustrated them with a large mouse intercross population with gene expression measured in six tissues. In the case of a hotspot with a single eQTL, we proposed a method to fine-map the hotspot by converting the multiple gene expression phenotypes into a co-dominant Mendelian trait. Application of this technique to a large-effect hotspot on mouse chromosome 6, with effects specific to pancreatic islets, reduced a 3.4 Mbp interval down to a 298 kb region with just three genes.

Our approach to convert the multivariate gene expression phenotype, for transcripts mapping to the chromosome 6 *trans*-eQTL hotspot, to a simple Mendelian trait could be applied more generally. While it was sufficient for us to consider the first two principal components to define the clusters of mice with a common eQTL genotype, one could focus on the non-recombinant mice, whose eQTL genotypes are known, and apply linear discriminant analysis (LDA) or other classification algorithms in order to infer the eQTL genotype in the recombinant mice. A key consideration that must be addressed is the possibility of multiple linked polymorphisms. If the assumption of a pleiotrophic single polymorphism is sound, we can evaluate the classifier by prediction errors in the training sample (non-recombinant mice) and use cross validation to improve the prediction, so that the inference of eQTL genotype in the recombinant mice could be more precise. In some classification methods, the quality of the prediction is measurable, and we could use the recombinant mice with better prediction scores to define the eQTL interval and exclude those with larger measurement errors, instead of using all recombinant mice available.

Efforts to improve the QTL mapping precision in experimental crosses have focused on increasing the density of recombination events, e.g. with advanced intercross lines (Darvasi and Soller, 1995) and heterogeneous stock (Mott et al., 2000). But the precision of QTL mapping may be hampered more by residual phenotype variation than by lack of recombination events. With a co-dominant Mendelian trait in an intercross, the length of the interval defined by recombination events flanking the trait locus is 1 cM on average with 100 intercross mice. This points to the importance of integrating multiple phenotypes that map to a common locus in order to develop composite traits with markedly reduced residual variation.

For dissecting *trans*-eQTL hotspots, the exploratory methods provided insight into the underlying evidence for more than one eQTL: that the multiple eQTL show distinct inheritance patterns, or that the recombinant and non-recombinant individuals show differences in expression. However, while the visualization methods can be strongly informative, they will not necessarily reveal the presence of the multiple eQTL in all cases, as the inheritance pattern of the two linked eQTL may be the same, or the first two linear discriminants may not be revealing of the difference between the recombinants and non-recombinants.

In forming a multivariate test statistic, we chose to follow the method of Knott and Haley (2000), but other multivariate analysis of variance (MANOVA) statistics could also be used, including Pillai's trace, Lawley-Hotelling's test, and Roy's Lambda (Anderson, 2003). Similarly, in the exploratory data visualization based on linear discriminant analysis, other dimension-reduction techniques such as PCA or tSNE (t-distributed Stochastic Neighbor Embedding Van der Maaten and Hinton, 2008), could be used. A supervised classification method, that makes use of the known eQTL genotypes of the non-recombinant individuals, is preferred.

The main issue in the formal statistical test is the choice of expression traits, as we can't handle a very large number of expression traits. Regularized methods (see Hastie et al., 2009, Sec 5.8), or a Bayesian approach, might have an advantage in this context. Such approaches could also be used to relax some of our modeling assumptions. For example, one might consider a model with two eQTL, where each expression trait can be affected by both.

We have seen examples with one group of the traits mapped to the right when considered jointly, while if mapping separately they are all on the left side. This happened when the single trait LOD curves had a bimodal shape and the two peaks were close together, such that the estimation of the QTL position was not stable. Hence our algorithm, which depends heavily on the QTL positions, failed to find the optimal partition of the traits.

Sometimes there were only a small number of traits mapping to the second eQTL versus the first one, with a large estimated distance between the two QTL. In cases like this one might want to repeat the test procedure with a smaller, refined group of traits. It can be tricky to choose which traits to include or exclude, and this can influence the results for the traits that map in-between. Ideally we want to search for a best partition of the traits without the constraint on the number of eQTL, but this would be computationally intensive.

We ignored the effects of eQTL elsewhere in the genome and considered just one region in isolation. In doing so, the effects of any other eQTL become part of the residual variation. Local-eQTL are a particularly important case, as they are quite common and often have large effect. Controlling for the effect of local-eQTL could give better precision in the dissection of a *trans*-eQTL hotspot.

We have implemented our methods in an R package (R Development Core Team, 2015), *qtlpvl*, available at https://github.com/jianan/qtlpvl. With a set of simulated gene expression data and position information for the expression traits, we showed how to define a *trans*-eQTL hotspot, to use exploratory tools to look at various patterns, and then how to test for one versus multiple eQTL for the hotspot. We can also jointly map multiple traits for a single QTL, or for two QTL under a multivariate normal assumptions, with the option to control for additive or interactive covariates. The covariates will be used for all the traits in our current package. It would be helpful to allow for trait-specific covariates, such as for the control of local-eQTL effects.

The methods and software we have developed can be used in a broad range of applications. We have focused on the case of microarray-based gene expression measurements, but the methods could be applied to any situation in which a large number of correlated traits are influenced by genotype in a common region.

REFERENCES

Anderson, T. W. 2003. Testing the general linear hypothesis; multivariate analysis of variance. In *An introduction to multivariate statistical analysis*, 3rd ed., chap. 8, 291–380. Wiley.

Darvasi, A., and M. Soller. 1995. Advanced intercross lines, an experimental population for fine genetic mapping. *Genetics* 141(3):1199–1207.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The elements of statistical learning: data mining, inference and prediction*. 2nd ed. New York: Springer.

Knott, S. A., and C. S. Haley. 2000. Multitrait least squares for quantitative trait loci detection. *Genetics* 156(2):899–911.

Van der Maaten, L., and G. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9(2579-2605):85.

Mott, R., C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint. 2000. A method for fine mapping quantitative trait loci in outbred animal stocks. *Proc. Natl. Acad. Sci. USA* 97(23): 12649–12654.

R Development Core Team. 2015. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.