

# Expression quantitative trait loci mapping with multivariate sparse partial least squares regression

Hyonho Chun<sup>1</sup> and Südüz Keleş<sup>1,2</sup>

June 12, 2008

<sup>1</sup>Department of Statistics, University of Wisconsin at Madison, Madison, WI.

<sup>2</sup>Department of Biostatistics and Medical Informatics, University of Wisconsin at Madison, Madison, WI.

Running Head: eQTL mapping with SPLS regression.

Keywords: Gene Expression, Genetic Marker, eQTL, Sparse Partial Least Squares Regression, Clustering.

Corresponding author:

Sündüz Keleş

Department of Statistics,

1300 University Avenue,

Madison, WI 53705.

Email: keles@stat.wisc.edu.

URL: [www.stat.wisc.edu/~keles](http://www.stat.wisc.edu/~keles).

## Abstract

Expression quantitative trait loci (eQTL) mapping concerns elucidating which transcripts or groups of transcript are associated with which markers or sets of markers. This problem poses significant challenges due to high dimensionality of both the gene expression and genomic marker data. We propose a multivariate response regression approach with simultaneous variable selection and dimension reduction for the eQTL mapping problem. In our approach, transcripts with similar expression are clustered into groups, and their expression profiles are viewed as multivariate responses. Then, we employ our recently developed sparse partial least squares regression methodology to select markers associated with each cluster of genes. We demonstrate, with extensive simulations, when and why this multivariate response approach gains power. We illustrate that it compares competitively with other approaches for this problem and has a number of significant advantages including the ability to handle highly correlated genotype data and computational efficiency. We provide an application of this methodology to a case study from obesity and diabetes research of mice. This framework bypasses the issue of multiple transcript- or marker-specific analyses, thereby avoids potential elevation of Type-I error. Additionally, joint analysis of multiple transcripts by the means of multivariate response regression leads to increase in power for detecting weak linkages.

## INTRODUCTION

Expression quantitative trait loci (eQTL) mapping, an innovative combination of traditional quantitative trait mapping and microarray technology, is genetic mapping of genome-wide gene expression or transcriptome. It aims to identify genomic locations to which expression traits are linked. eQTL mapping provides an opportunity to investigate a large and unbiased set of traits as well as to study traits which are most immediately connected to DNA sequence variation. Therefore, it has the potential to furnish more detailed information about biological processes of gene networks than the traditional quantitative trait analysis. eQTL mapping studies have been applied in several model organisms and humans recently (BREM *et al.* 2002; SCHADT *et al.* 2003; MORLEY *et al.* 2004; CHESLER *et al.* 2005; STRANGER *et al.* 2005; WANG *et al.* 2006). These studies thus far have demonstrated several advantages of this line of research from identifying candidate genes (SCHADT *et al.* 2003) to elucidating regulatory networks (BREM *et al.* 2002; SCHADT *et al.* 2003; YVERT *et al.* 2003).

Typical eQTL studies involve a  $N \times G$  matrix of gene expression, where rows are different individuals (e.g., mice, in the order of tens) and columns are genes (in the order of thousands), and a  $N \times p$  matrix ( $X_p$ ) with genomic marker (in the order of hundreds or more) information. The eQTL analysis differs from the traditional single quantitative trait locus (QTL) analysis in the number of traits considered. We refer to KENDZIORSKI and WANG (2006) for a comprehensive review of general statistical issues concerning eQTL studies. Initial methods for eQTL mapping can be grouped into two (KENDZIORSKI *et al.* 2006): (1) transcript-specific analysis in which mapping of a single expression trait is considered at a time and the entire eQTL mapping analysis consists of thousands of transcript-specific analysis; (2) marker-specific analysis detecting differentially expressed transcripts based on their association with the discrimination pattern of an individual marker and the complete analysis requires scanning for all the markers under consideration. The fundamental challenge and impracticality with either of these approaches, which extend either the single trait mapping to multiple traits or differential expression identification for a single treatment (as implied by the marker genotype) to multiple treatments via multiple hypothesis testing, are due

to elevated number of false positives and loss of power due to disjoint (either at the transcript- or marker-level) analysis.

Methods such as clustering or principal components analysis on genome-wide gene expression (YVERT *et al.* 2003; LAN *et al.* 2003) are viable approaches for reducing the number of tests considered and possibly improving the power of linkage detection by combining multiple similarly behaving transcripts into a single trait. However, these methods suffer from the incapability of producing transcript-specific information, i.e., identified markers associate with a meta transcript trait that is composed of several distinct transcripts.

Recent efforts for eQTL analysis focus on combined analysis of all the transcript and marker data by innovatively collapsing the aforementioned approaches (1) and (2). Mixture Over Marker (MOM) model of KENDZIORSKI *et al.* (2006) is the first approach that allowed information sharing across transcripts by an Empirical Bayes method. Its operating principle depends on first identifying mapping transcripts and then characterizing one or more eQTL per mapping transcript by utilizing transcript-specific highest posterior density regions. Recently, GELFOND *et al.* (2007) improved on the MOM model by utilizing the genomic locations of the transcripts. In an attempt to identify mapping transcripts and related eQTL simultaneously, JIA and XU (2007) proposed a Bayesian shrinkage method called BAYES. This approach models individual transcript expression as a linear function of the marker data and treats eQTL mapping in a variable selection context. To achieve better power in linkage detection, it uses all the transcripts and markers simultaneously when fitting transcript-level regression models. These transcript-level regression models share the same set of prior distributions for the regression parameters. Although BAYES is flexible enough to map multiple markers per transcript, it is a highly parametric model which relies on prior specifications and requires intense computations. Furthermore, properties of BAYES when the markers are highly correlated are not studied. This is an important practical challenge because, typically, many markers are in close proximity of each other and are correlated due to linkage disequilibrium. Therefore, the highly correlated nature of marker data may hamper the performance of variable selection schemes that do not explicitly accommodate such a grouping structure.

In this paper, we propose a multivariate response regression framework for the eQTL mapping problem. This approach utilizes sparse partial least squares regression (SPLS) (CHUN and KELEŞ 2007), a novel statistical methodology for regression with multivariate response variables and with built-in dimension reduction and variable selection. Such a formulation is motivated by the apparent power advantages of multiple phenotype modeling observed in traditional multi-trait QTL mapping (JIANG and ZENG 1995; ALLISON *et al.* 1998). It aims to capitalize on the correlation among multiple transcripts while simultaneously dealing with all the markers. Recent eQTL mapping studies in yeast *Saccharomyces cerevisiae* provided extensive evidence that most of the eQTL have weak effects, and half of the transcripts require more than five loci (marker) under additive models (BREM and KRUGLYAK 2005). This study further elucidated the importance of joint analysis of multiple transcripts and markers to boost weak linkage signals. Our approach for utilizing multiple transcripts is aided by the successful applications of clustering in the study of gene expression analysis for identifying co-expressed groups of genes (EISEN *et al.* 1998). First, by utilizing the clustering principle, we partition genes into smaller groups based on their expression similarity across the experimental units. This helps to view the vector of expression of genes within a cluster as a multivariate response. Next, we form a cluster-level multivariate response regression and employ the principled simultaneous dimension reduction and variable selection method SPLS regression to identify markers affecting all or a subgroup of genes within the cluster. In the next two sections, we review the underlying principles of the SPLS regression methodology by focusing on the aspects important to our application and describe our proposed procedure for the eQTL mapping problem in detail. In the simulation studies section, we outline and present the results of our simulation studies benchmarking the operating characteristics of our approach and comparing it with other approaches. Our simulation studies are rooted by the recent observations on the eQTL architectures in yeast (ROCKMAN and KRUGLYAK 2006) and are more comprehensive than the eQTL mapping simulation studies present in the literature. In this section, we provide extensive evidence that the proposed framework has excellent power and very small Type-I error and significantly outperforms a comparable multiple univariate regression formulation of the eQTL mapping

problem. In the case study section, we illustrate our approach with a case study in obesity and diabetes research of mice (LAN *et al.* 2006) and then discuss potential extensions.

## eQTL MAPPING WITH SPLS REGRESSION

We start with a brief review of the sparse partial least squares regression since it forms the core of our methodology for eQTL mapping. Subsequently, we describe our approach in detail.

**Sparse partial least squares (SPLS) regression:** Partial least squares (PLS) regression has been used as an alternative to the ordinary least squares (OLS) regression in ill-conditioned linear regression models that arise in several disciplines such as chemistry, economics, medicine, psychology, and pharmaceutical science (DE JONG 1993). At the core of PLS regression is a dimension reduction technique that operates under the assumption of a basic latent decomposition of the response matrix ( $Y \in \mathcal{R}^{N \times q}$ ) and predictor matrix ( $X \in \mathcal{R}^{N \times p}$ ):

$$Y = TQ^T + F, \quad \text{and} \quad X = TP^T + E,$$

where  $T \in \mathcal{R}^{N \times K}$  is a matrix that produces  $K$  linear combinations (scores);  $P \in \mathcal{R}^{p \times K}$  and  $Q \in \mathcal{R}^{q \times K}$  are matrices of coefficients (loadings); and  $E \in \mathcal{R}^{N \times p}$  and  $F \in \mathcal{R}^{N \times q}$  are matrices of random errors.

In order to specify the latent component matrix  $T$  such that  $T = XW$ , PLS requires finding the columns of  $W = (w_1, w_2, \dots, w_K)$  from successive optimization problems. The criterion for  $k$ -th *estimated* direction vector  $\hat{w}_k$  is formulated as

$$\begin{aligned} \hat{w}_k &= \operatorname{argmax}_w w^T X^T Y Y^T X w \\ &\text{s.t. } w^T w = 1, \quad w^T S_{XX} \hat{w}_j = 0, \end{aligned} \tag{1}$$

for  $j = 1, \dots, k - 1$ , where  $S_{XX}$  is the sample covariance matrix of  $X$ . After estimating the latent components ( $T$ ), loadings ( $Q$ ) are estimated via OLS for the model  $Y = TQ^T + F$ .  $\beta^{PLS}$  is

estimated by  $\hat{\beta}^{PLS} = \hat{W}\hat{Q}^T$ , where  $\hat{W}$  and  $\hat{Q}$  are estimates of  $W$  and  $Q$ , since  $Y = XWQ^T + F = X\beta^{PLS} + F$ .

In [CHUN and KELEŞ \(2007\)](#), we investigated theoretical properties of the PLS regression and showed that although it had been traditionally promoted for regression problems with large number of explanatory variables, it suffers from the curse of dimensionality in the contemporary large  $p$ , small  $N$  setting. In particular, the components of the direction vectors get attenuated due to existence of large number of irrelevant variables. In order to address this issue, we developed Sparse Partial Least Squares which aims to promote sparsity by imposing  $L_1$  penalty to the direction vector objective function (1) of PLS. The final objective function that SPLS operates on is given by

$$\begin{aligned} \min_{\alpha, w} & -\kappa\alpha^T M\alpha + (1 - \kappa)(w - \alpha)^T M(w - \alpha) + \lambda_1|w|_1 + \lambda_2|w|_2 \\ \text{s.t.} & \alpha^T \alpha = 1, \end{aligned} \quad (2)$$

where  $M = X^T Y Y^T X$ . This formulation promotes exact zero property by imposing  $L_1$  penalty onto a surrogate of direction vector ( $w$ ) instead of the original direction vector ( $\alpha$ ), while keeping  $\alpha$  and  $w$  close to each other. The theoretical arguments for this formulation can be found in [CHUN and KELEŞ \(2007\)](#) where we also characterized the solutions to this minimization problem. The first  $L_1$  penalty encourages sparsity on  $w$ , and the second  $L_2$  penalty takes care of potential singularity in  $M$  when solving for  $w$ . The parameter  $\kappa$  is for reducing the concavity of the problem and avoiding locally optimal solutions. We show in [CHUN and KELEŞ \(2007\)](#) that a  $\kappa$  value of less than 0.5 performs good in practice and considering multiple  $\kappa$  values has the effect of initiating the algorithm with different starting values. After solving  $\alpha$  and  $w$ , we rescale the  $w$  solution to have norm one and use this scaled version as the estimated direction vector.

This direction vector objective function (2) is utilized in the course of the SPLS algorithm to select active, e.g., relevant, variables. We define  $\mathcal{A}$  to be an index set for active variables,  $K$  as the number of components, and  $X_{\mathcal{A}}$  as the matrix of covariates contained in  $\mathcal{A}$ . Then, the computational SPLS algorithm can be summarized as follows:

SPLS algorithm:

1. Set  $\hat{\beta}^{PLS} = 0$ ,  $\mathcal{A} = \{ \}$ ,  $k = 1$ , and  $Y_1 = Y$ .
2. While ( $k \leq K$ ),
  - 2.1. Find  $\hat{w}$  by solving the minimization problem in (2) with  $M = X^T Y_1 Y_1^T X$ .
  - 2.2. Update  $\mathcal{A}$  as  $\{i : \hat{w}_i \neq 0\} \cup \{i : \hat{\beta}_i^{PLS} \neq 0\}$ .
  - 2.3. Fit PLS with  $X_{\mathcal{A}}$  by using  $k$  number of latent components.
  - 2.4. Update  $\hat{\beta}^{PLS}$  by using the new PLS estimates of the direction vectors, and update  $Y_1$  and  $k$  through  $Y_1 \leftarrow Y - X \hat{\beta}^{PLS}$  and  $k \leftarrow k + 1$ .

As can be seen from the formulation in (2), SPLS has tuning parameters  $\lambda_1$ ,  $\lambda_2$  and  $K$ . We show in CHUN and KELEŞ (2007) that setting  $\lambda_2$  to  $\infty$  yields the solution to have the form of a soft thresholded estimator (ZOU and HASTIE 2005). As a result, step 2.1 of the SPLS algorithm takes the form of a simple soft thresholding driven only  $\lambda_1$ . To tune  $\lambda_1$ , we utilize the following adaptive form of soft thresholded estimator where we only need to tune  $\eta$  in place of direction vector specific  $\lambda_1$ s for all  $K$  direction vectors:

$$\tilde{w} = (|\hat{w}| - \eta \max_{1 \leq i \leq p} |\hat{w}_i|) I(|\hat{w}| \geq \eta \max_{1 \leq i \leq p} |\hat{w}_i|) \text{sign}(\hat{w}),$$

where  $0 \leq \eta \leq 1$ . This form of soft thresholding retains components those are greater than some fraction of the maximum component and avoids the necessity for tuning multiple thresholding parameters, one for each direction vector.

In summary, from a practical point of view, SPLS has two tuning parameters,  $\eta$  and  $K$ , and these are tuned by cross-validation (CV). Simulation studies with combinations of small and large numbers of predictors and sample sizes in CHUN and KELEŞ (2007) show that SPLS regression achieves both high predictive power and accuracy for finding the relevant variables and it outperforms other dimension reduction or variable selection methods such as sparse principle components

analysis (ZOU *et al.* 2006), supervised Principle Component (BAIR *et al.* 2006) and LASSO (TIB-SHIRANI 1996). Moreover, it has the capability to select a higher number of relevant variables than the available sample size since the number of variables that contribute to each direction vector is not limited by the sample size. This is a unique property of the SPLS regression and leads to simultaneous dimension reduction and variable selection.

As apparent from the formulation in (1), another unique property of SPLS is its ability to handle multivariate  $Y \in \mathcal{R}^{N \times q}$ ,  $q \geq 1$ , without additional computational complexity. This property motivates the use of SPLS within the context of eQTL mapping where the goal is to utilize transcript and marker information simultaneously.

**eQTL mapping with SPLS regression:** Our approach consists of two steps.

*Step 1. Clustering of the  $G \times N$  expression matrix.* We start out by clustering the  $G \times N$  gene expression data matrix. Current eQTL studies typically have a total of  $N$  experimental units from two or more distinct populations. There is an ample literature on clustering of gene expression data. Among the simplest methods are nonparametric clustering methods such as k-means, partitioning around medoids (KAUFMAN and ROUSSEEUW 1990), hierarchical clustering (EISEN *et al.* 1998), or parametric clustering methods, e.g., mixture of Gaussian distributions (FRALEY and RAFTERY 2002). We view the choice of the clustering method as a design dependent decision and will give an example of a clustering application within the context of our case study. The thrust of the clustering step is to provide a transition from transcript-level regression models to module/cluster-level regression models.

*Step 2. Cluster-specific SPLS regression with bootstrap confidence intervals.* After the clustering/grouping step, within each cluster  $k$ , we define a response vector  $Y_i^{(k)}$  of dimension  $1 \times G_k$  to denote the expression of all the  $G_k$  genes in this cluster measured on the  $i$ -th subject. Subsequently, a cluster-specific marker model is considered. For subject  $i$ , we have

$$Y_i^{(k)} = X_i \beta^{(k)} + E_i,$$

where  $E_i$  denotes the random error matrix and  $\beta^{(k)}$  is a  $p \times G_k$  matrix representing the contribution

of each marker  $m \in \{1, \dots, p\}$  to expression variation of each transcript  $g \in \{1, \dots, G_k\}$  of cluster  $k$ . Such a model is fit for every cluster using SPLS regression. In Figure 1, we provide a pictorial representation of the conceptual framework at the cluster-level when considering both univariate and multivariate responses. U-SPLS refers to SPLS regression with a single response which then corresponds to carrying out  $G_k$  number of SPLS regressions, one for each transcript. M-SPLS refers to SPLS regression with multivariate response and, as depicted in the figure, M-SPLS obtains linkage estimates (indicated by the regression coefficients) simultaneously.

There are two apparent gains to be expected from this clustering combined with SPLS regression approach. First, we expect this approach to be more powerful than both the individual transcript and the marker analyses as the genes with similar patterns will be considered simultaneously and inherent correlations among the transcripts will be taken into account. Thus, it will be able to detect weak linkages. Second, it is expected to avoid Type-I error inflation by avoiding multiple model fittings. One additional attractive property of SPLS regression is that when a single marker among a group of highly correlated markers is putatively linked to one or more transcripts, SPLS regression tends to select this whole set of correlated markers rather than randomly pick one. Many common variable selection methods such as LASSO (TIBSHIRANI 1996) randomly picks one of these markers. Selecting the whole set of correlated markers is more desirable as the data does not discriminate among these and one can always conduct further detailed studies for selected candidate markers.

The final stage of cluster-specific SPLS regression is constructing bootstrap confidence intervals for transcript selection. The outcome of multivariate SPLS regression is a set of selected markers that significantly associate with one or more transcripts in the cluster and their estimated regression coefficients. We provide an example of such an outcome for a dataset from our simulation study (case B.1 in simulation section in Figure 2. Panel (a) depicts true linkages simulated for a cluster of 100 genes over 145 markers. Panel (b) displays linkages estimated by the M-SPLS regression. As evident in this plot, M-SPLS is able to select the true set of markers but several false linkages, albeit with very weak effect sizes, are also revealed for the selected markers. This is not realistic because,

generally, a given marker or a set of markers is likely to associate with a subset of the genes within a cluster as cluster analysis is also prone to errors. To circumvent this, we construct bootstrap confidence intervals for transcript selection with the M-SPLS regression. Essentially, after the initial application of M-SPLS regression, subjects are randomly selected with replacement and multivariate response PLS regression is fitted using only the selected markers from the original fit. An empirical distribution of estimated regression coefficients is obtained for each marker/transcript combination after a large number of bootstrap iterations. Using these empirical distributions, a 95% confidence interval is constructed for each marker/transcript combination. The final summary of linkages contains marker/transcript combinations for which the confidence intervals exclude 0. Panel (c) of Figure 2 summarizes the linkages after the bootstrap confidence intervals are taken into account. Here, only the relevant transcripts have non-zero coefficients at the selected markers. For illustration purposes, we provide bootstrap confidence intervals for marker M136 across all the 100 transcripts in panel (d) of Figure 2.

## SIMULATION STUDIES

Our simulations are largely motivated by the empirical observations in eQTL analysis of yeast *S. cerevisiae* (Table 2 of BREM and KRUGLYAK (2005) and Figure 1 of ROCKMAN and KRUGLYAK (2006)). These two extensive studies in yeast reveal that the genetic architecture of most expression traits involves multiple QTL, and most of these QTL explain a minority of trait variation. In a yeast cross, only 3% of expression traits are consistent with single-locus inheritance, and most traits require more than two additive QTL.

We start our simulation section by first comparing our approach to BAYES (JIA and XU 2007) and MOM (KENDZIORSKI *et al.* 2006) by adopting the simulation experiments of JIA and XU (2007). Then, we devise simulation experiments that incorporate the eQTL architectures observed in the yeast *S. cerevisiae* and evaluate power gain and Type-I error attained by both the univariate and multivariate SPLS regressions.

**Simulation studies comparing cluster-specific SPLS regression based eQTL mapping to**

Table 1: Type-I error and power results based on the simulation set-up of [JIA and XU \(2007\)](#). Details of each scenario are pictorially depicted in [Figure 3](#).

Method	(Single Marker, Multiple Transcripts)		(Multiple Markers, Multiple Transcripts)	
	Type-I error	Power	Type-I error	Power
MOM	0	0.9800	0.0004	0.642
BAYES	0	0.9800	0	0.993
M-SPLS	0.007	0.9870	0.007	0.986
U-SPLS	0.126	0.9910	0.1430	0.928

### **BAYES and MOM:**

[JIA and XU \(2007\)](#) simulate expression and marker data in their simulations as follows. Ten markers ( $p = 10$ ) on 360 cM genome are generated by using the Haldane map function ([HALDANE 1919](#)) and four eQTL are located at markers 1, 3, 6 and 10. A total of  $G = 1000$  transcripts and  $N = 50$  samples are generated following the Bayesian regression model which forms the backbone of [JIA and XU \(2007\)](#)' BAYES method. Authors consider two scenarios that are depicted in [Figure 3](#). In the first scenario (single marker, multiple transcripts, panel A of [Figure 3](#)), each subgroup of transcripts is affected by only a single marker. Transcripts 1-50 are under the influence of marker 10, transcripts 601-604 of marker 3, transcripts 605-610 of marker 1, and transcripts 961-1000 of marker 6. The eQTL effects, which are essentially coefficients of the relevant markers in the BAYES's regression model, are generated from  $N(0, 3^2)$ , and error terms are generated from  $N(0, 0.1^2)$ . In the second scenario (multiple markers, multiple transcripts, panel B of [Figure 3](#)), multiple marker sets affect the expression of subgroups of transcripts in the following manner: transcripts 1-16 are controlled by markers 1 and 10, transcripts 17-20 by markers 1, 3, and 10, and transcripts 971-990 by markers 1 and 6. Data for the remaining transcripts as well as eQTL effects and error terms are generated in the same way as in the first scenario.

We perform 100 replicates of each simulation scenario and compare the operating characteristics of our method with the results from [JIA and XU \(2007\)](#). We note that [JIA and XU \(2007\)](#) use only 20 replicates which is presumably due to the computational complexity of the BAYES method but the results are overall comparable because our results for 20 versus 100 simulation replicates are very similar. We used 99% bootstrap confidence intervals based on 1000 bootstrap samples for

transcript selection whereas [JIA and XU \(2007\)](#) use some unspecified false discovery rate (FDR), which is much less than 1%, for linkage thresholding.

The simulation averages of power and Type-I error are reported in [Table 1](#). Here, U-SPLS refers to univariate SPLS where we fit an SPLS regression per transcript and M-SPLS refers to multivariate SPLS fitted using all the 1000 transcripts simultaneously. U-SPLS is expected to produce many false positives due to multiple fitting of the regression model. As seen in the table, indeed this approach has highly inflated Type-I error. It is possible to argue that the performance of U-SPLS can be improved by implementing a bootstrap confidence interval step similar to M-SPLS. However, this increases computation time considerably, i.e., if M-SPLS replicates 1000 bootstrap samples, U-SPLS would replicate 1000 for each  $G_k$  transcripts. We observe that M-SPLS has quite small Type-I error and performs comparably to BAYES in terms of power despite the fact that the underlying data generating model precisely follows the modeling assumptions of BAYES. Additionally, it is interesting to observe that M-SPLS can accommodate the case where multiple markers are associated with different subsets of the transcripts. In other words, M-SPLS does not require all the transcripts in the cluster to be linked to the same set of markers with similar effect sizes. This is a desired property since different groups of transcripts within a cluster could very well be associated with multiple marker sets.

**Simulation studies investigating the operating characteristics of cluster-specific SPLS regression based eQTL mapping:** In the previous subsection, we observed that M-SPLS regression overcomes the elevation of Type-I error by avoiding multiple model fits. In this section, we study how M-SPLS regression takes advantage of simultaneous modeling of multiple transcripts in more depth.

As stated earlier, our simulations for this section are motivated by the empirical observations in eQTL analysis of *S. cerevisiae* ([BREM and KRUGLYAK 2005](#); [ROCKMAN and KRUGLYAK 2006](#)). The simulations are divided into two as single eQTL architecture versus multiple eQTL architectures. [Figure 4](#) depicts these scenarios pictorially and we describe below each in more detail.

*Case A: Single eQTL architecture for a cluster of genes.* In this set-up, we assume that there is

only one eQTL architecture involving one or more loci and affecting a percentage of the genes in the cluster. In other words, the observed correlation mechanism among the genes is a result of a single eQTL architecture. This corresponds to considering three factors in the data generating scheme:  $r$ : # of eQTL (1, 3, 7);  $\rho$  : proportion of cluster genes affected by the eQTL architecture (10%, 30%, 60%, 90%);  $e$ : effect size (weak versus strong). By changing the proportion of cluster genes under the control of the eQTL mechanism, we aim to address the case where only a fraction of the similarity among the expression of genes within a cluster is attributable to the eQTL mechanism. This scenario also encompasses the case where the clustering procedure is noisy resulting in many irrelevant genes within a cluster.

The current literature on eQTL mapping utilizes only a very small number of markers and often does not impose a population structure when investigating the operating characteristics of the methods by simulations. We extend this by first utilizing only 10 markers as in the simulations of [JIA and XU \(2007\)](#) and then 145 markers, which is the size of the marker dataset in our case study. The first simulation uses 10 randomly selected markers (D2Mit2, D2Mit297, D2Mit327, D2Mit17, D2Mit26, D2Mit49, D2Mit148, D3Mit151, D3Mit22, D3Mit19) from chromosomes 2 and 3 of the mice study and the extended version uses all of the 145 markers. The rationale for using the real marker data is for accurately mimicking the ranges of the correlations that inherently exist among markers. The data generation starts with generating a norm 1 eQTL architecture direction vector with  $r$  nonzero components. The effect size is controlled by a constant multiplied to the eQTL architecture direction vector. We consider the constants  $e = 1$  and  $e = 2$  for weak and strong effects.  $\rho$  proportion of transcripts are controlled by this eQTL architecture and the random error terms are generated from  $N(0, 1)$ . We use 5-fold CV for marker selection and 95% bootstrap confidence intervals based on 1000 bootstrap samples for the transcript selection step of the M-SPLS regression. Each simulated dataset has  $G = 100$  transcripts with  $N = 60$  (number of mice in the case study) subjects and the number of simulation replicates is set to 100. More details on data generation of these simulations are provided in Table 3 of the Supplementary Materials Section.

The results in terms of power and Type-I error are presented in Figure 5. U-SPLS regression

exhibits inflated Type-I error as expected based on the earlier simulations following [JIA and XU \(2007\)](#)'s design. Additionally, the power of U-SPLS does not change as the proportion of transcripts associated with the eQTL mechanism increases. This is also expected since U-SPLS considers separate regression fits for each transcript and there is no information sharing. On the other hand, M-SPLS regression has very small Type-I error. Overall, M-SPLS also has significantly higher power than U-SPLS. We note that in our data generating scheme, the effect size is inherently decreasing as the number of markers  $r$  in the eQTL mechanism increases. This is because the effect size is proportional to the elements of the direction vector and the norm of the direction vector is by definition 1. As a result, the  $r = 7$  markers and weak effect size configuration has the highest noise level among the 24 configurations considered. Despite the decreasing overall signal at the transcript-level, the power of M-SPLS increases as the proportion of transcripts affected by the eQTL mechanism increases. This provides concrete evidence that M-SPLS successfully utilizes information across multiple transcripts; therefore, low signal linkages that might be missed by examining individual markers separately become detectable.

As an extension, we then consider a more realistic scenario with all the 145 markers. We assume that  $r = 3, 10$  markers control  $\rho \in \{0.1, 0.3, 0.6, 0.9\}$  proportion of the transcripts and consider both weak and strong effect sizes as in the previous case. [Figure 6](#) summarizes the results for both U-SPLS and M-SPLS and reassuringly our observations for the small number of markers simulation hold when the number of markers exceed the sample size. The power of M-SPLS again increases as the proportion of transcripts affected by the eQTL increases. Additionally, M-SPLS has more power than U-SPLS even when only 10% of the genes in the cluster are affected by the same set of markers (at both effect sizes when  $r = 10$ ).

*Case B: Multiple QTL architectures for a cluster of genes.* In this setting, subgroups of genes within a cluster have different eQTL architectures. This corresponds to having multiple hidden components. We study the case where two hidden components, therefore two different eQTL architectures, are present. These two hidden components are: (1) *eQTL mechanism 1*: linear combinations of markers 11, 12 and 13; (2) *eQTL mechanism 2*: linear combinations of markers 136

and 137. Mechanism 1 is set to have a weaker effect size than mechanism 2. We consider three sub cases for the multiple eQTL architectures simulation:

- *B.1*: Transcripts 1-50 are under the influence of mechanisms 1 and 2, and transcripts 51-90 are only affected by mechanism 1.
- *B.2*: Same as *B.1* but with a larger effect size.
- *B.3*: Similar to *B.1* except that transcripts 51 – 80 are associated with mechanism 1. All linkages with mechanism 2 have the same strong effect size. However, linkages with mechanism 1 exhibit weak effect sizes randomly generated from  $Unif(-0.3, 0.3)$  with the exception of the linkage with transcript 30. Transcript 30 has an effect size of 0.5 with mechanism 1.

More details on how the actual components of the direction vectors are set are provided in Table 4 of the Supplementary Materials Section.

The results of these multiple eQTL architecture simulations are provided in Figure 7. For cases *B.1* and *B.2*, M-SPLS has greater power than U-SPLS with a much smaller and negligible Type-I error. This observation is consistent with our earlier simulation experiments. In case *B.3*, linkages with the first eQTL mechanism cannot be detected by U-SPLS because the effect size is very small. This results in poor power for U-SPLS. On the other hand, although M-SPLS misses some of the linkages with this architecture, it has at least twice the power that U-SPLS has. This result is also reflected in the hot spot selection performance of the methods. Hot spot regions are traditionally defined as regions of a genome with multiple linkages. The first eQTL mechanism cannot be revealed as a hot spot from individual regression analyses by U-SPLS (bottom left panel of Figure 7). Here, we use the proportion of estimated linkages for hot spot detection. Although M-SPLS misses some of the mapping transcripts due to the eQTL mechanism 1 with weak effect size, it is able to identify markers involved in this mechanism as hot spots (bottom right panel of Figure 7).

## CASE STUDY: APPLICATION TO MICE DATA FROM A STUDY OF OBESITY AND DIABETES

We present an application our method to a mice data published by LAN *et al.* (2006). This dataset contains expression measurements of 45,265 transcripts from liver tissues of 60 mice. Mice were collected from a (B6 × BTBR)  $F_2$ -*ob/ob* cross where animals lacked a functional leptin protein hormone, known to be important for reproduction and regulation of body weight and metabolism (ZHANG *et al.* 1994), and segregated for obesity and diabetes related phenotypes. We utilized the pre-processed data which is publicly available at GEO (<http://www.ncbi.nlm.nih.gov/projects/geo/query/acc.cgi?acc=GSE3330>). The marker map for this data consists of 145 microsatellite markers from 19 mouse non-sex chromosomes. Following JIA and XU (2007), we performed an initial screening of the transcripts based on their variability across 60 mice and excluded transcripts with sample variances  $< 0.12$  from our analysis. This left a total of  $G = 1573$  transcripts.

Next, we clustered these remaining transcripts. As discussed earlier, clustering method in an application will be highly design dependent. For a time course experiment, methods that utilize dependencies among different time points (YUAN and KENDZIORSKI 2006) or methods specifically parameterizing cluster profiles (JÖRNSTEN and KELEŞ 2008) might be more desirable. For the mice data, we considered the following approach motivated by the successful use of the *topological overlap measure (TOM)* (RAVASZ *et al.* 2002) in clustering analysis (ZHANG and HORVATH 2005). First, we constructed undirected, unweighted gene networks based on the expression data using the Gaussian Graphical Model (GGM) approach of SCHFER and STRIMMER (2005). The constructed network is then used to compute TOM for each pair of transcripts. Dissimilarity measure 1-TOM between two transcripts represents a lack of closeness based on the number of shared neighbours in the expression network. Since 95 transcripts did not share any neighbours with other transcripts, they were set aside as singletons. Hierarchical clustering on the remaining transcripts using this dissimilarity measure resulted in about 47 clusters based on the average silhouette mea-

sure (KAUFMAN and ROUSSEEUW 1990). The within cluster Pearson correlations ranged from 0.027 to 0.948 with a mean of 0.226 across 47 clusters.

Here, we present the results for one of the clusters in more detail. This cluster contains three lipid metabolism transcripts, namely, *Scd1*, *Elovl6*, and *Fasn*, that were investigated by different analysis of the same dataset (LAN *et al.* 2006; JIA and XU 2007). There are a total of 83 transcripts in this cluster with a median within cluster correlation of 0.12. An application of our approach with M-SPLS yields 24 markers, presented in Table , that are associated with one or more transcripts. The total number of linkages identified for this cluster is 463 and there are 61 transcripts that do not map to any marker. An image plot of the estimated effects of this cluster across markers and transcripts is provided in Figure 8. The entire M-SPLS analysis, including both the tuning and the bootstrap steps, for this cluster of 83 transcripts took only 30 minutes on a 64 bit machine with 2.66Ghz CPU.

First, we note that many of the selected markers are in close proximity of each other on the mouse genome. Essentially, these 24 markers form clusters on the chromosomes that they appear. These physically close markers are highly correlated (bottom two panels of Figure 9 of the Supplementary Materials section). The fact that these highly correlated markers are identified closely relates to the correlated group selection property of the SPLS regression that we highlighted earlier. Since SPLS can select more than one variable at each step of the selection process (step 2.2 of the SPLS algorithm), it is able to capture all the relevant correlated variables rather than arbitrarily selecting one. Second, in each correlated marker group in Table , there is at least one marker previously declared as an obesity and diabetes related locus (literature search for the markers is performed via [www.informatics.jax.org](http://www.informatics.jax.org)). This result is encouraging since it provides a list of transcripts mapping to loci known to be related to obesity and diabetes.

Expression profiles of lipid metabolism transcripts *Scd1*, *Elovl6*, and *Fasn* are highly correlated (Figure 10 of the Supplementary Materials Section, minimum pairwise correlation is 0.756). Therefore, it is reasonable to expect similar linkages for these transcripts. Indeed, M-SPLS identifies that these transcripts map to similar markers, whereas BAYES yields different linkages for

these. This could be due to high correlation among markers. Unlike the markers generated by the Haldane map function in the simulation study section, markers from the mice study exhibit very high correlations. This multi-collinearity problem is not explicitly addressed in BAYES, and priors for regression coefficients are assumed to be independent. In fact, similar mixture priors were used by [SHA \*et al.\* \(2006\)](#) in the context of a different model and a decrease in the variable selection performance was observed for the correlated variable case. So, it is plausible that BAYES also suffers from a similar problem and tends to select only one of the variables among a set of correlated variables like other standard variable selection methods.

[LAN \*et al.\* \(2006\)](#) highlighted that transcripts which were highly correlated with *Scd1* mapped to the same genomic locations as *Scd1*, and found major QTL peaks for most of the 20 lipid metabolism traits at markers D2Mit263 and D5Mit240. These two markers are successfully identified by our approach. Among the five hot spots reported by MOM, two of them are also identified by M-SPLS with the group of transcripts we considered. These are D2Mit241, which is adjacent to the obesity-modifier locus D2Mit9 ([STOEHR \*et al.\* 2004](#)), and D8Mit249, which is close to the "fat" gene known to affect obesity and diabetes ([NAGGERT \*et al.\* 1995](#)). Instead of D5Mit1, that is known to affect triglyceride levels, M-SPLS identified D5Mit348 which is adjacent to D5Mit1. D15Mit63, emphasized in the findings of BAYES, is also identified by M-SPLS. Loci on chromosome 2 have been most popular candidates for obesity and diabetes ([STOEHR \*et al.\* 2000](#); [JEREZ-TIMAURE \*et al.\* 2005](#); [DIAMENT \*et al.\* 2004](#)), but hot spots from MOM and BAYES do not have noticeable indication of this. Also, BAYES does not find any hot spots on chromosome 2. However, M-SPLS yields strong effects for markers on chromosome 2. Furthermore, although marker D5Mit267, which is identified by M-SPLS but missed by MOM and BAYES, does not seem to be directly related to obesity and diabetes, it is associated with reproduction which is another known function of leptin protein hormone ([EWART-TOLAND \*et al.\* 1999](#)).

Table 2: Markers identified for a cluster of size 86 including three lipid metabolism transcripts: *Scd1*, *Elovl6*, and *Fasn*.

Marker	Map (cM)	# of Mapping Transcripts	Reference
D2Mit274	69.6	7	close to obesity modifier locus D2Mit9 (STOEHR <i>et al.</i> 2004), obesity/diabetes syndrome (STOEHR <i>et al.</i> 2000).
D2Mit17	73.9	19	obesity/diabetes syndrom (STOEHR <i>et al.</i> 2000).
D2Mit106	77.9	20	liver weight (JEREZ-TIMAURE <i>et al.</i> 2005), obesity/diabetes syndrome (STOEHR <i>et al.</i> 2000).
D2Mit194	85.4	19	obesity locus (DIAMENT <i>et al.</i> 2004), body weight and fat (JEREZ-TIMAURE <i>et al.</i> 2005), obesity/diabetes syndrome (STOEHR <i>et al.</i> 2000).
D2Mit263	98.7	20	lipid metabolism (LAN <i>et al.</i> 2006), obesity/diabetes syndrome (STOEHR <i>et al.</i> 2000).
D2Mit51	101.7	19	
D2Mit49	104.2	20	
D2Mit229	110.5	21	
D2Mit148	121.6	22	
D5Mit348	6.3	18	
D5Mit75	11.8	21	
D5Mit267	17.1	21	reproduction in leptin-deficient obese mice (EWART-TOLAND <i>et al.</i> 1999).
D5Mit259	43	21	
D5Mit9	46	22	
D5Mit240	49.1	22	lipid metabolism (LAN <i>et al.</i> 2006).
D5Mit136	54.9	22	
D8Mit249	58.1	17	fat gene (NAGGERT <i>et al.</i> 1995), triglyceride level (COLINAYO <i>et al.</i> 2003).
D8Mit211	72	22	
D8Mit113	77.6	18	
D9Mit8	58.6	20	fat-pad mass (MEHRABIAN <i>et al.</i> 1998).
D9Mit15	93.6	22	
D15Mit174	0	15	
D15Mit136	11.5	19	
D15Mit63	21	16	early life body weight (MILLER <i>et al.</i> 2002), diabetic modifier (TAKESHITA <i>et al.</i> 2006).

## DISCUSSION

The advent of microarray technology is providing an unprecedented opportunity for investigating complex genetics underlying inheritance of thousands of transcript levels in segregating populations. As discussed here, one of the statistical challenges is the eQTL mapping problem where the question of interest is to identify linkages between thousands of transcripts and markers. We have formulated the eQTL mapping problem as a variable selection problem in a multivariate response regression. We then utilized sparse partial least squares (CHUN and KELEŞ 2007) as a simultaneous variable selection and dimension reduction approach to identify linkages. This framework offers a computationally fast alternative to approaches that analyze multiple transcript and marker data simultaneously for gaining power and avoiding induced multiplicities for good error control. Our approach relies on an initial clustering of expression data that partitions transcripts into groups of similar expression. Then, expression profile across the genes in the cluster is treated as a multivariate response vector. We view the choice of the clustering method as an experiment/design dependent parameter and therefore do not advocate for a particular clustering method. As genome-wide eQTL studies expand measuring gene expression in multiple environments, different developmental stages, cell and tissue types, the rich literature on the clustering of expression data can readily be utilized.

We have demonstrated and validated the advantages of our method with simulation experiments motivated by the recent eQTL architecture observations in yeast *S. cerevisiae*. Our experiments included eQTL architectures with strong effects on a small fraction of transcripts as well as weak effects on a substantially larger fraction of genes. These studies showed that as the number of mapping transcripts increases, the power of M-SPLS increases whereas its univariate analog with transcript-level regressions cannot capitalize on this phenomena. We illustrated the utility of our approach with an example from mice obesity and diabetes research. This case study highlighted the ability of SPLS regression for selecting groups of correlated markers. BAYES, an alternative variable selection approach to eQTL problem, lacks this property and tends to select only one marker among the group of correlated markers. Our approach was able to consistently yield similar

linkages for highly correlated transcripts. Furthermore, we were able to identify a marker which was missed by the previous analysis of the same dataset but could potentially be important since it relates to another function of the leptin protein hormone (EWART-TOLAND *et al.* 1999).

In this paper, we allowed the markers to appear as main terms in the regression model. Identifying interactions among markers within the context of eQTL analysis is also a major research problem. Considering interactions is an exponentially complex problem. However, with an appropriate pre-screening of markers, SPLS regression has the potential to handle large number of interactions. In CHUN and KELEŞ (2007), this property is illustrated with as many as 5000 variables. Another important research question in eQTL mapping is allowing for linkages with locations between markers using interval mapping (CHEN and KENDZIORSKI 2007). Our current formulation only allows for mapping at exact marker locations. However, a first pass with our approach and then a more focused traditional interval mapping (SEN and CHURCHILL 2001) based on the selected markers might be a viable strategy.

## ACKNOWLEDGEMENTS

This research has been supported in part by a PhRMA Foundation Research Starer Grant in Informatics and the NIH grant HG003747 to S.K.

## LITERATURE CITED

- ALLISON, D. B., B. THIEL, P. S. JEAN, R. C. ELSTON, M. C. INFANTE, and N. J. SCHORK, 1998 Multiple phenotype modeling in gene-mapping studies of quantitative traits: power advantages. *American Journal of Human Genetics* **63**: 1190–1201.
- BAIR, E., T. HASTIE, D. PAUL, and R. TIBSHIRANI, 2006 Prediction by supervised principal components. *Journal of the American Statistical Association* **101**: 119–137.
- BREM, R. and L. KRUGLYAK, 2005 The landscape of genetic complexity across 5700 gene expression traits in yeast. *PNAS* **102**: 1572–1577.

- BREM, R. B., G. YVERT, R. CLINTON, and L. KRUGLYAK, 2002 Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**: 752–755.
- CHEN, M. and C. KENDZIORSKI, 2007 A statistical framework for expression quantitative trait loci (eQTL) mapping. *Genetics* **177**: 761–771.
- 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, D. W. THREADGILL, K. F. MANLY, and R. W. WILLIAMS, 2005 Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nature Genetics* **37**: 233–242.
- CHUN, H. and S. KELEŞ, 2007 Sparse partial least squares regression. Under revision. [http://www.stat.wisc.edu/~keles/Papers/SPLS\\_Nov07.pdf](http://www.stat.wisc.edu/~keles/Papers/SPLS_Nov07.pdf).
- COLINAYO, V. V., J. H. QIAO, X. P. WANG, K. L. KRASS, E. SCHADT, A. J. LUSIS, and T. A. DRAKE, 2003 Genetic loci for diet-induced atherosclerotic lesions and plasma lipids in mice. *Mammalian Genome* **14**: 464–471.
- DE JONG, S., 1993 SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18**: 251–263.
- DIAMENT, A., P. FARAHANI, S. CHIU, J. FISLER, and C. WARDEN, 2004 A novel mouse chromosome 2 congenic strain with obesity phenotypes. *Mammalian Genome* **15** (6): 452–459.
- EISEN, M. B., P. T. SPELLMAN, P. O. BROWN, and D. BOTSTEIN, 1998 Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**: 14863–14868.
- EWART-TOLAND, A., K. MOUNZIH, J. QIU, and F. F. CHEHAB, 1999 Effect of the genetic background on the reproduction of leptin-deficient obese mice. *Endocrinology* **140**: 732–738.
- FRALEY, C. and A. E. RAFTERY, 2002 Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association* **97**: 611–631.
- GELFOND, J. A. L., J. G. IBRAHIM, and F. ZOU, 2007 Proximity model for expression quantitative trait loci (eQTL) detection. *Biometrics* **63**: 1108–1116.
- HALDANE, J. B. S., 1919 The combination of linkage values and the calculation of distances

- between the loci of linked factors. *Journal of Genetics* **8**: 299–309.
- JEREZ-TIMAURE, N. C., E. J. EISEN, and D. POMP, 2005 Fine mapping of a QTL region with large effects on growth and fatness on mouse chromosome 2. *Physiological Genomics* **21**: 411–422.
- JIA, Z. and S. XU, 2007 Mapping quantitative trait loci for expression abundance. *Genetics* **176**: 611–623.
- JIANG, C. and Z. ZENG, 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. *Genetics* **140**: 1111–1127.
- JÖRNSTEN, R. J. and S. KELEŞ, 2008 Mixture models with multiple levels, with application to the analysis of multifactor gene expression data. *Biostatistics* doi:10.1093/biostatistics/kxm051.
- KAUFMAN, L. and P. ROUSSEEUW, 1990 *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons, New York.
- KENDZIORSKI, C. and P. WANG, 2006 A review of statistical methods for expression quantitative trait loci mapping. *Mammalian genome* **17**: 509–517.
- KENDZIORSKI, C. M., M. CHEN, M. YUAN, H. LAN, and A. D. ATTIE, 2006 Statistical methods for expression quantitative loci (eQTL) mapping. *Biometrics* **62**: 19–27.
- LAN, H., M. CHEN, J. B. FLOWERS, B. S. YANDELL, D. S. STAPLETON, C. M. MATA, E. T. N KEEN MUI, M. T. FLOWERS, K. L. SCHUELER, K. F. M. A ND ROBERT W. WILLIAMS, C. KENDZIORSKI, and A. D. ATTIE, 2006 Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genetics* **2**: e6.
- LAN, H., J. P. STOEHR, S. T. NADLER, K. L. SCHUELER, B. N S. YANDELL, and A. D. ATTIE, 2003 Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics* **164**: 1607–1614.
- MEHRABIAN, M., P. Z. WEN, J. FISLER, R. C. DAVIS, , and A. J. LUSIS, 1998 Genetic loci controlling body fat, lipoprotein metabolism, and insulin levels in a multifactorial mouse model. *The journal of clinical investigation* **101**: 2485–2496.
- MILLER, R. A., J. M. HARPER, A. GALECKI, and D. T. BURKE, 2002 Big mice die young: early

- life body weight predicts longevity in genetically heterogeneous mice. *Aging Cell* **1**: 22–29.
- MORLEY, M., C. M. MOLONY, T. M. WEBER, K. G. DEVLIN, J. L. SNEDEMAN, R. S. SPIELMAN, and V. G. CHEUNG, 2004 Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**: 743–747.
- NAGGERT, J. K., L. D. FRICKER, O. VARLAMOV, P. M. NISHINA, Y. ROUILLE, D. F. STEINER, R. J. CARROLL, B. J. PAIGEN, and E. H. LEITER, 1995 Hyperproinsulinaemia in obese fat/fat mice associated with a carboxypeptidase E mutation which reduces enzyme activity. *Nature Genetics* **10**: 135–142.
- RAVASZ, E., A. SOMERA, D. MONGRU, Z. OLTVAI, and A. BARABASI, 2002 Hierarchical organization of modularity in metabolic networks. *Science* **297**: 1551–1555.
- ROCKMAN, M. V. and L. KRUGLYAK, 2006 Genetics of global gene expression. *Nature* **7**: 862–872.
- SCHADT, E. E., S. A. MONKS, T. DRAKE, A. J. LUSIS, N. CHE, V. COLINAYO, T. G. RUFF, S. B. MILLIGAN, J. R. LAMB, G. CAVET, P. S. LINSLEY, M. MAO, R. B. STOUGHTON, and S. H. FRIEND, 2003 Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**: 297–302.
- SCHFER, J. and K. STRIMMER, 2005 Learning large-scale graphical gaussian models from genomic data. In *AIP Conference Proceedings 776. Science of Complex Networks: From Biology to the Internet and WWW (CNET 2004)*, edited by J. F. F. Mendes, S. N. Dorogovtsev, F. V. A. A. Povolotsky, and J. G. Oliveira, pp. 263–276, Aveiro, PT.
- SEN, S. and G. A. CHURCHILL, 2001 A statistical framework for quantitative trait mapping. *Genetics* **159**: 371–387.
- SHA, N., M. G. TADESSE, and M. VANNUCCI, 2006 Bayesian variable selection for the analysis of microarray data with censored outcomes. *Bioinformatics* **22**: 2262–2268.
- STOEHR, J., S. NADLER, K. SCHUELER, M. RABAGLIA, B. YANDELL, S. METZ, and A. ATTIE, 2000 Genetic obesity unmasks nonlinear interactions between murine type 2 diabetes susceptibility loci. *Diabetes* **49**: 1946–1954.

- STOEHR, J. P., J. E. BYERS, S. M. CLEE, H. LAN, I. V. BORONENKOV, K. L. SCHUELER, B. S. YANDELL, and A. D. ATTIE, 2004 Identification of major quantitative loci controlling body weight variation in ob/ob mice. *Diabetes* **53**: 245–249.
- STRANGER, B. E., M. S. FORREST, A. G. CLARK, M. J. MINICHELLO, S. DEUTSCH, R. LYLE, S. HUNT, B. KAHL, S. E. ANTONARAKIS, S. TAVARE, P. DELOUKAS, and E. T. DERMITZAKIS, 2005 Genome-wide associations of gene expression variation in humans. *PLoS Genetics* **1**: e78.
- TAKESHITA, S., M. MORITANI, K. KUNIKA, H. INOUE, and M. ITAKURA, 2006 Diabetic modifier QTLs identified in F2 intercrosses between akita and A/J mice. *Mammalian Genome* **17**: 927–940.
- TIBSHIRANI, R., 1996 Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **58**: 267–288.
- WANG, S., N. YEHYA, E. E. SCHADT, H. WANG, T. A. DRAKE, and A. J. LUSIS, 2006 Genetic and genomic analysis of a fat mass trait with complex inheritance reveals marked sex specificity. *PLoS Genetics* **2**: e15.
- YUAN, M. and C. KENDZIORSKI, 2006 Hidden markov models for microarray time course data in multiple biological conditions. *Journal of the American Statistical Association* **101**: 1323–1332.
- 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.
- ZHANG, B. and S. HORVATH, 2005 A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology* **4**: Article 17.
- ZHANG, Y., R. PROENCA, M. MAFFEI, M. BARONE, L. LEOPOLD, and J. M. FRIEDMAN, 1994 Positional cloning of the mouse obese gene and its human homologue. *Nature* **372**: 425–431.
- ZOU, H. and T. HASTIE, 2005 Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society: Series B (Statistical Methodology)* **67**: 301 – 320.
- ZOU, H., T. HASTIE, and R. TIBSHIRANI, 2006 Sparse principal component analysis. *Journal of*

Computational and Graphical Statistics **15**: 265–286.

**U-SPLS**

For  $g = 1, \dots, G_k$ :

$$\begin{array}{c}
 \boxed{Y_g} \\
 N \times 1
 \end{array}
 =
 \begin{array}{c}
 \boxed{X} \\
 N \times p
 \end{array}
 \times
 \begin{array}{c}
 \boxed{B_g} \\
 N \times 1
 \end{array}
 +
 \begin{array}{c}
 \boxed{E_g} \\
 N \times 1
 \end{array}$$

**M-SPLS**

$$\begin{array}{c}
 \boxed{Y} \\
 N \times G_k
 \end{array}
 =
 \begin{array}{c}
 \boxed{X} \\
 N \times p
 \end{array}
 \times
 \begin{array}{c}
 \boxed{B} \\
 p \times G_k
 \end{array}
 +
 \begin{array}{c}
 \boxed{E} \\
 N \times G_k
 \end{array}$$

Figure 1: *Cluster-specific univariate and multivariate SPLS regressions.* U-SPLS refers to SPLS regression with a univariate response. Application of U-SPLS at the transcript-level requires carrying out  $G_k$  number of SPLS regressions, one for each transcript. M-SPLS refers to SPLS regression with multivariate response where linkages are estimated simultaneously.

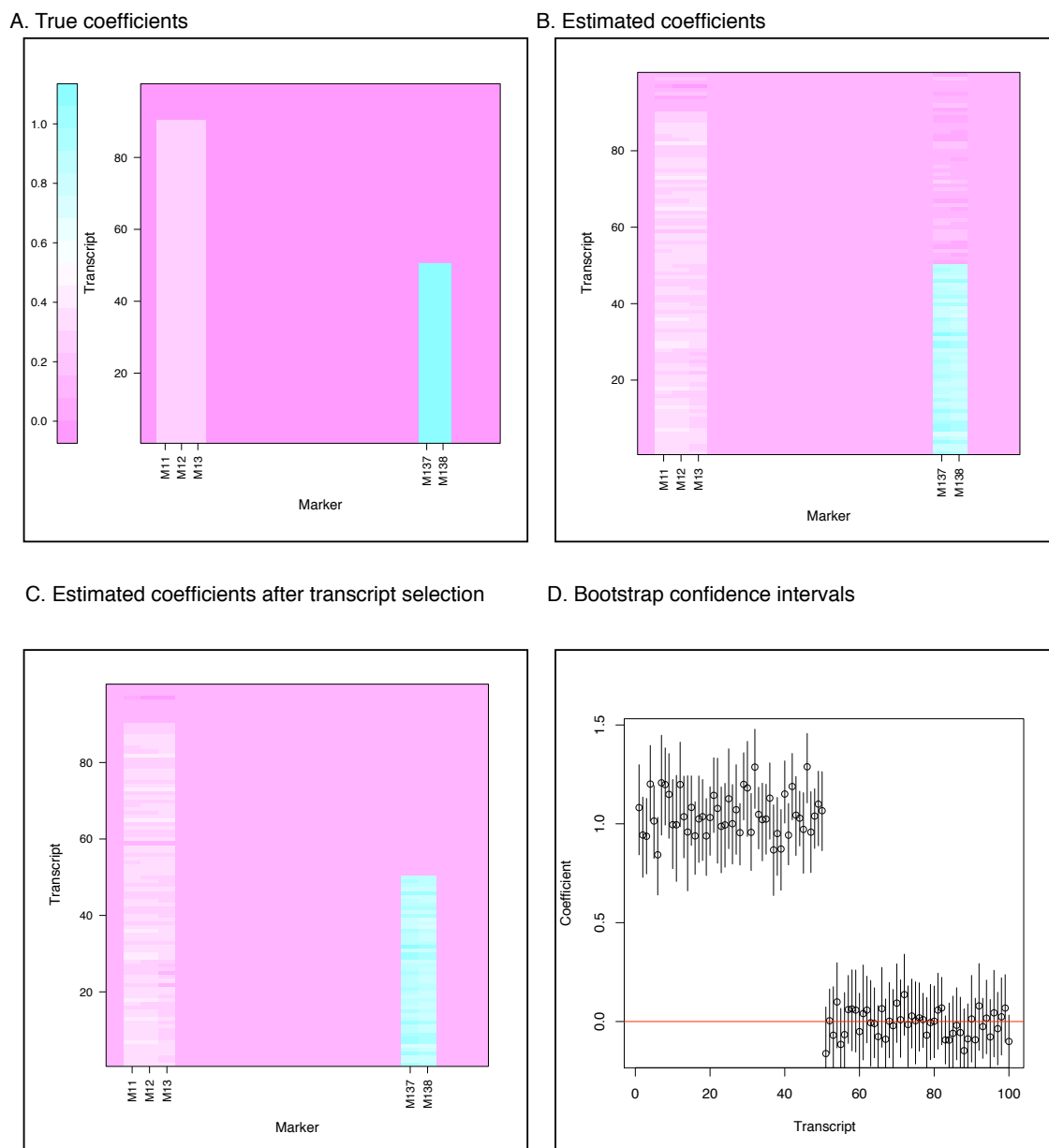
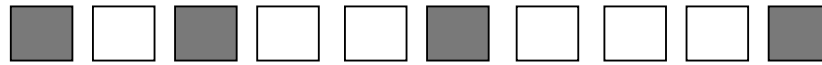


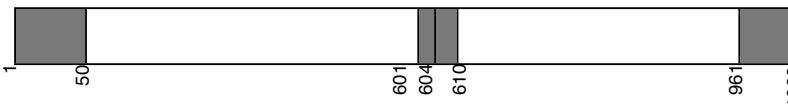
Figure 2: A. Set of true linkages. B. Linkages estimated by M-SPLS regression. C. Estimated linkages after considering the bootstrap confidence intervals. D. 95% CIs for marker M136 across all the transcripts in the cluster.

A: (Single marker, Multiple transcripts)

Markers:

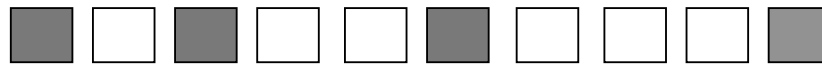


Transcripts:



B: (Multiple markers, Multiple transcripts)

Markers:



Transcripts:

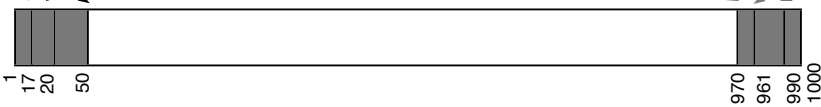


Figure 3: A. In the (Single marker, Multiple transcripts) scenario, only a single marker affects the expression of group of transcripts. Transcripts 1-50 are under the influence of marker 10, transcripts 601-604 of marker 3, transcripts 605-610 of marker 1, and transcripts 961-1000 of marker 6. B. In the (Multiple markers, Multiple transcripts) scenario, multiple marker sets affect the expression of subgroups of transcripts in the following manner: transcripts 1-16 are controlled by markers 1 and 10, transcripts 17-20 by markers 1, 3, and 10, and transcripts 971-990 by markers 1 and 6.

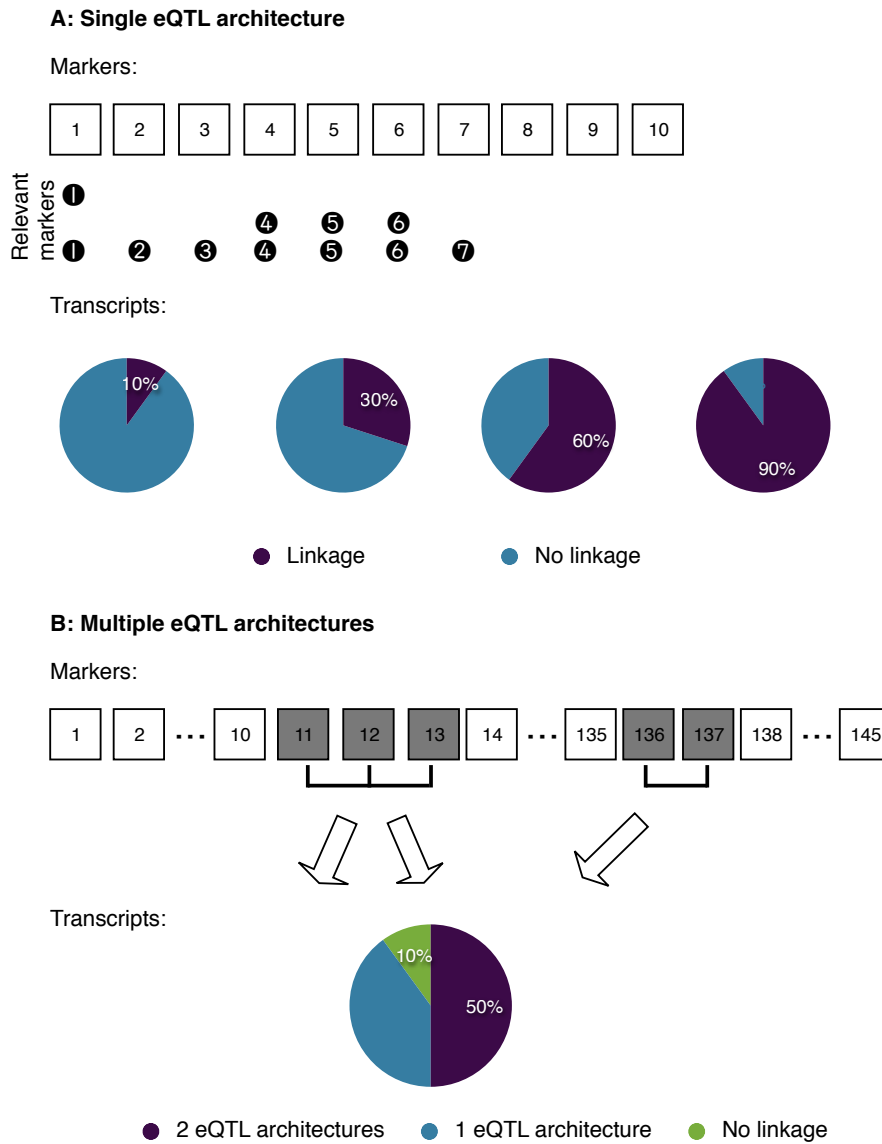


Figure 4: A. Single eQTL architecture. In a cluster,  $r \in \{1, 3, 7\}$  markers associate with a proportion  $\rho \in \{0.1, 0.3, 0.6, 0.9\}$  of the genes. For each case, both weak and strong effects are considered. B. Multiple eQTL architectures. In a cluster, one or two eQTL mechanisms (each with varying number of markers) associate with different proportions of genes.

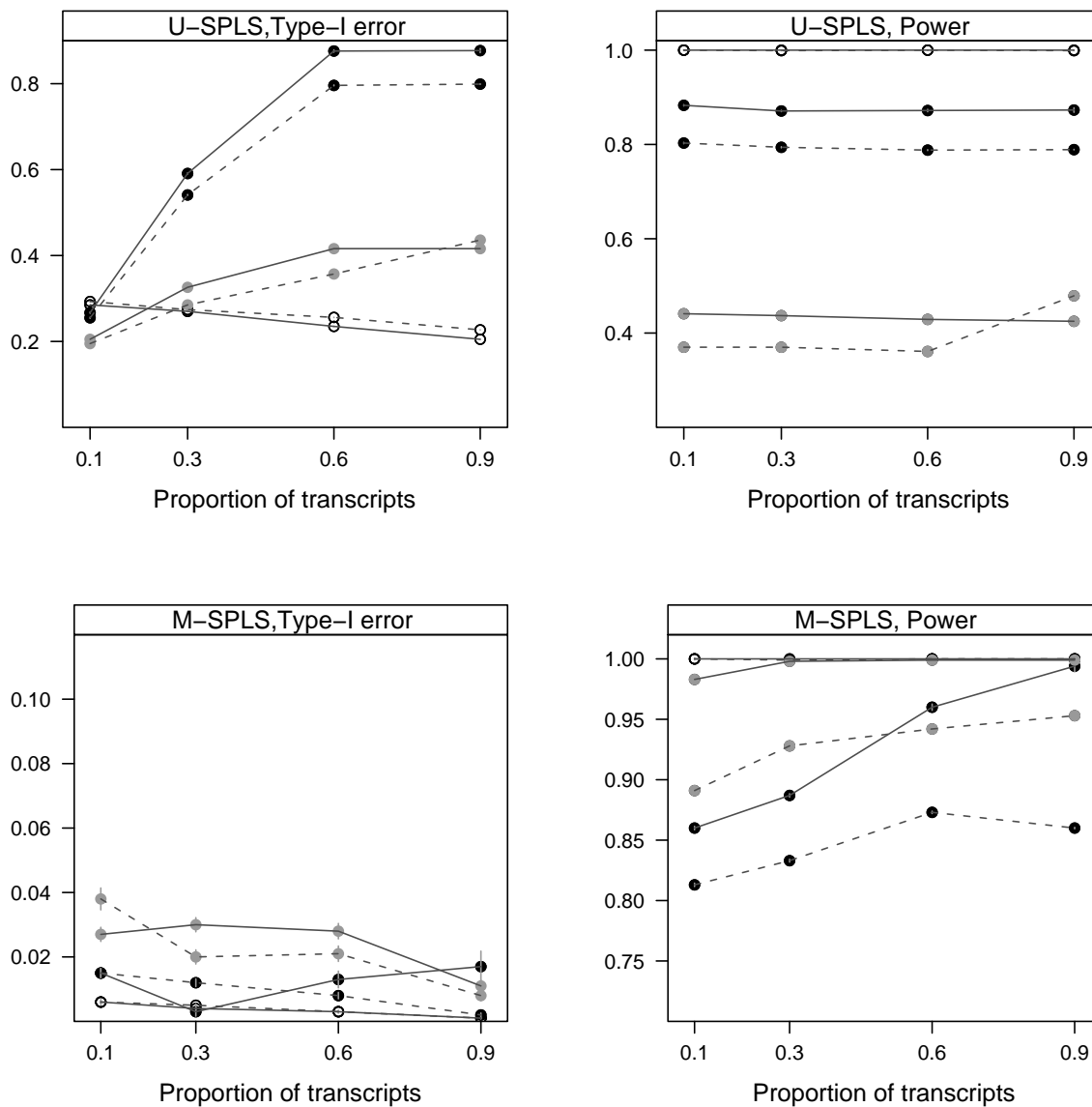


Figure 5: Results for Case A with  $p = 10$  markers. Symbols represent different number of markers associated with  $\rho \in \{0.1, 0.3, 0.6, 0.9\}$  proportion of genes in the cluster:  $\circ$ :  $r = 1$ ; gray filled  $\circ$ :  $r = 3$ ;  $\bullet$ :  $r = 7$ . Different line types indicate weak (dashed line) or strong (solid line) effects.

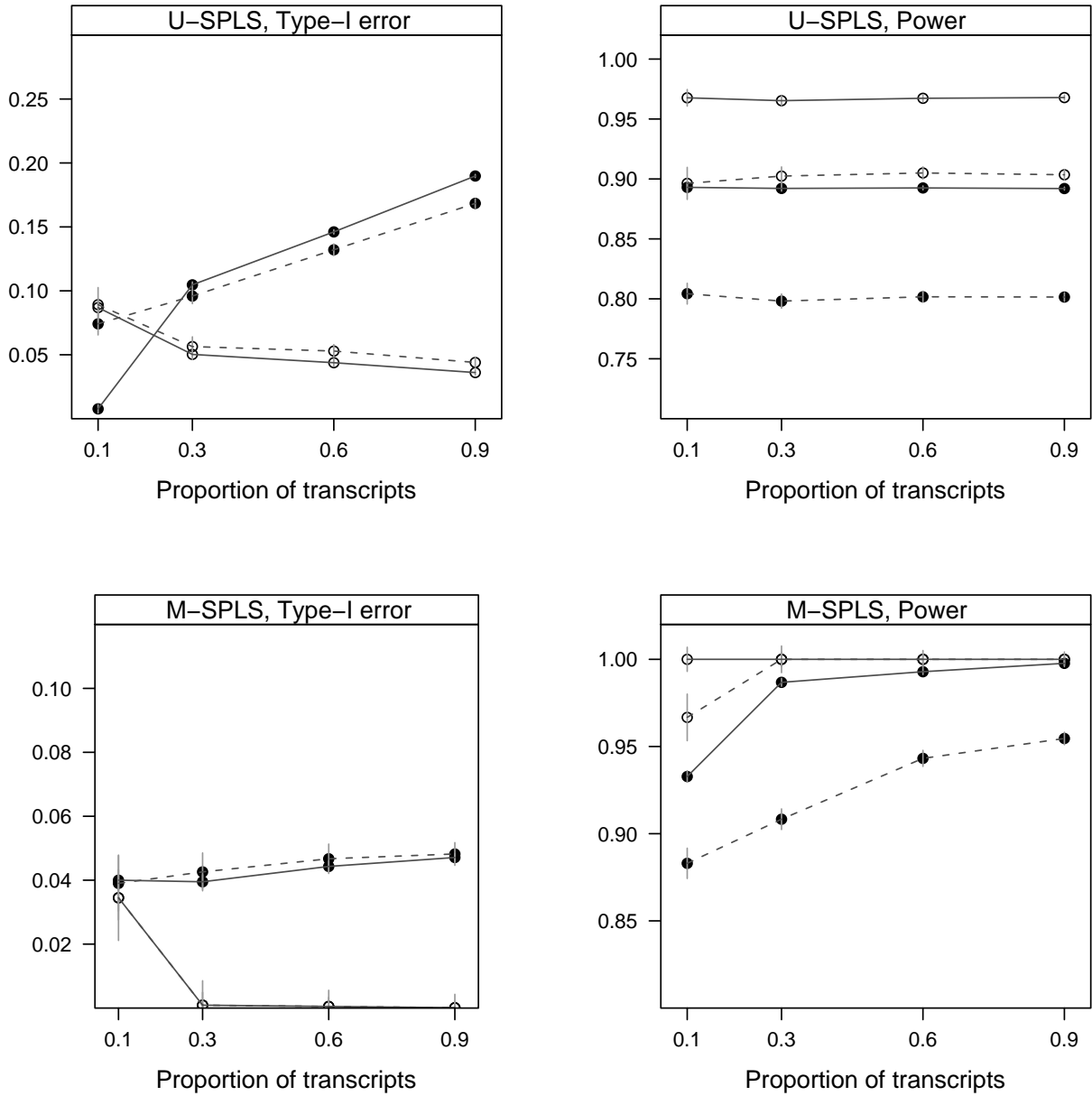


Figure 6: Results for Case A with  $p = 145$  markers. Symbols represent different number of markers associated with  $\rho \in \{0.1, 0.3, 0.6, 0.9\}$  proportion of genes in the cluster:  $\circ$ :  $r = 3$ ;  $\bullet$ :  $r = 10$ . Different line types indicate weak (dashed line) or strong (solid line) effects.

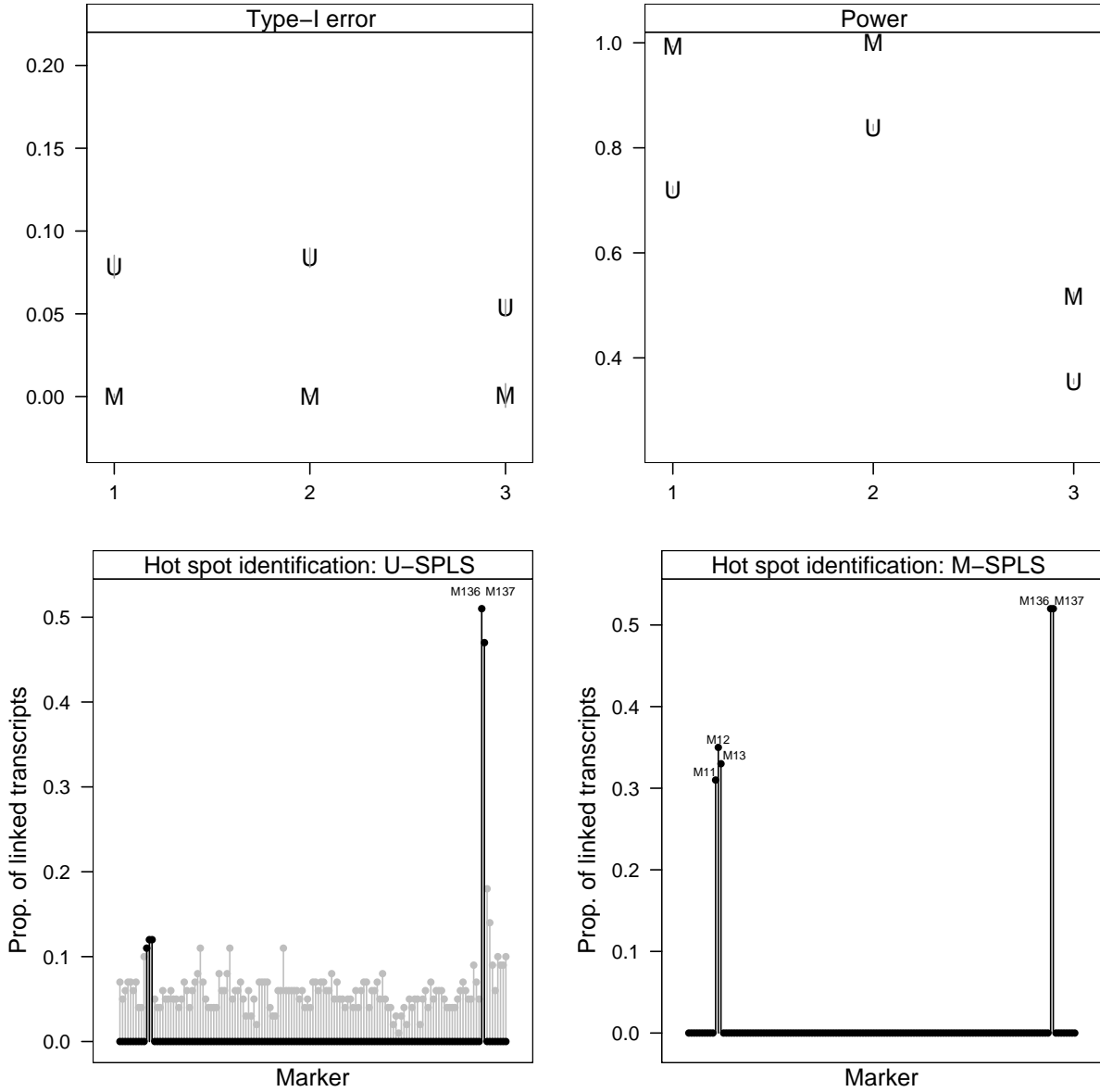


Figure 7: Results for Case B. Top panels represent Type-I error and power for U-SPLS (U) and M-SPLS(M) with vertical lines representing simulation standard errors. Bottom panels report the proportion of linked transcripts for each marker by U-SPLS (left-bottom panel) and M-SPLS (right-bottom panel) for simulation case B.3.

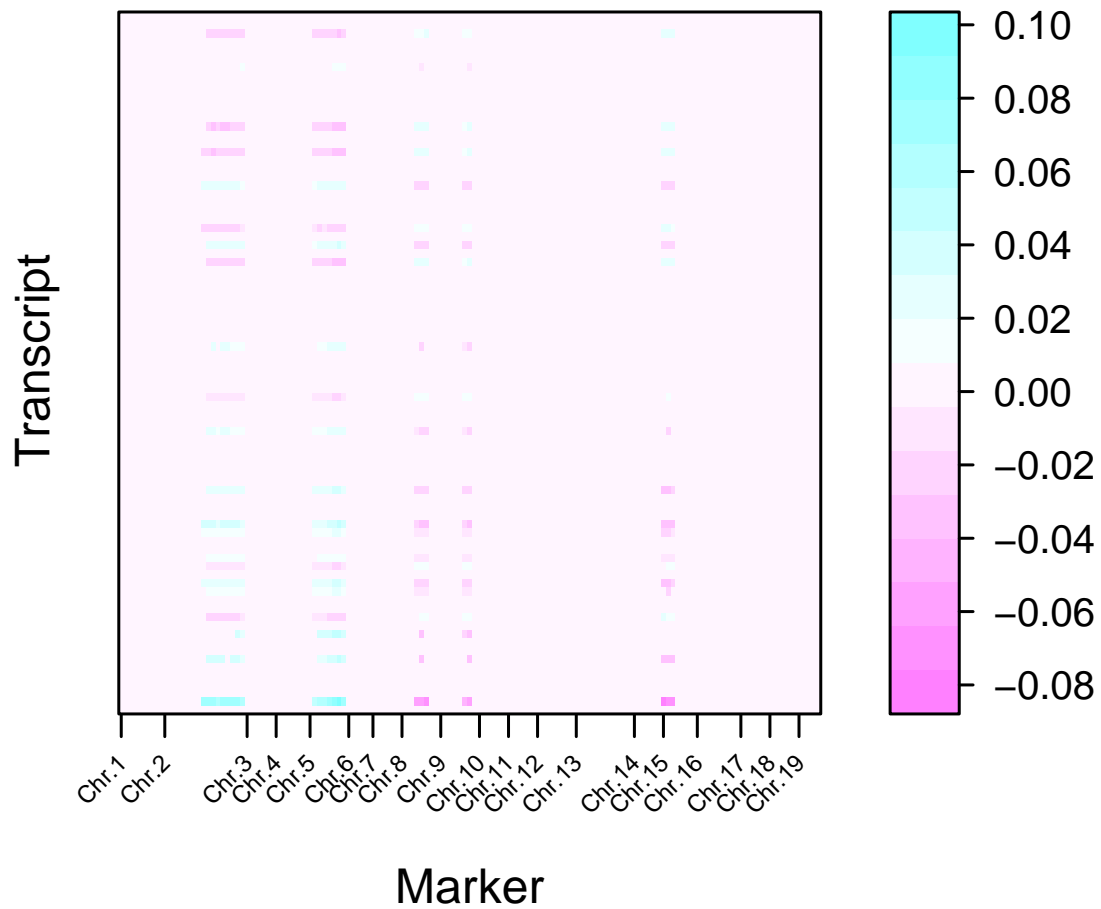


Figure 8: *M-SPLS* solution for a cluster of 86 genes including the three lipid metabolism transcripts.

## SUPPLEMENTARY MATERIALS

An R package will be made available on the last author's website upon publication.

Table 3: *Components of the direction vectors for simulation A.* The final marker-specific regression coefficients are obtained by multiplying the direction vectors with weak ( $e = 1$ ) or strong ( $e = 2$ ) effect sizes.

Case A				
	$r = 1$	$r = 3$	$r = 7$	$r = 10$
$p = 10$	$w_1 = 1.$ $w_j = 0,$ $j = 2, \dots, 10.$	$w_5 = 0.768.$ $w_6 = 0.512.$ $w_7 = 0.384.$ $w_j = 0,$ $j = 1, \dots, 4,$ $8, \dots, 10.$	$w_1 = 0.61.$ $w_2 = 0.49.$ $w_3 = 0.36.$ $w_4 = 0.24.$ $w_5 = 0.36.$ $w_6 = 0.24.$ $w_7 = 0.12.$ $w_j = 0,$ $j = 8, \dots, 10.$	
$p = 145$		$w_j = 0.577,$ $j = 11, \dots, 13.$ $w_j = 0,$ $j = 1, \dots, 10,$ $14, \dots, 145.$		$w_j = 0.316,$ $j = 11, \dots, 13,$ $40, \dots, 43,$ $74, 136, \dots, 137.$ $w_j = 0,$ everywhere else.

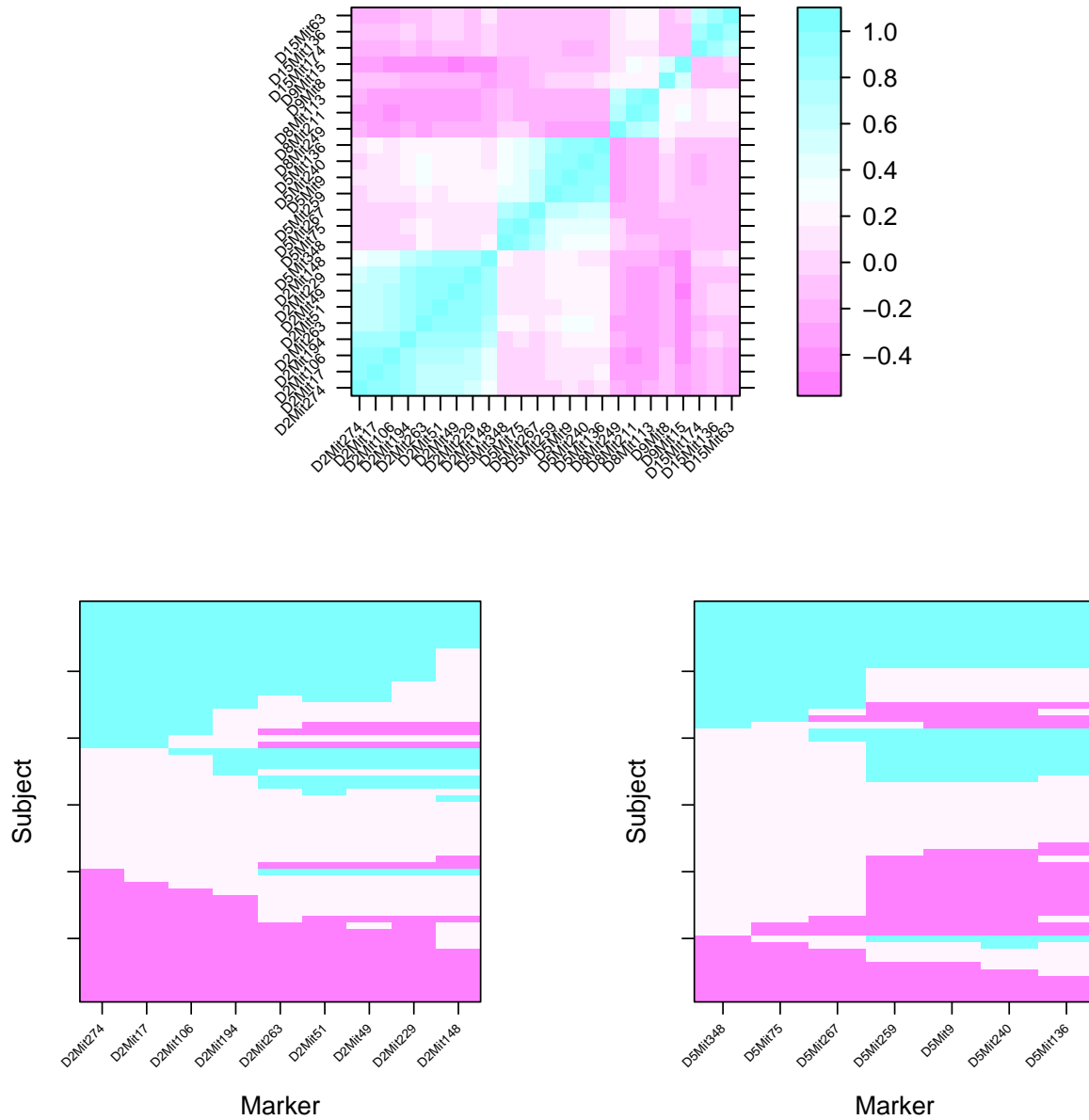


Figure 9: *Top: Correlation plot of all the 145 markers. Bottom: Genotype profiles of 60 mice across two clusters of markers: the left image plot represents a cluster of 9 markers on chromosome 2 whereas the right image plot represents a cluster of 7 markers on chromosome 5. Distinct colors depict the three genotypes from an F2 cross.*

Table 4: *Components of the direction vectors for simulation B.* Direction vectors for the first and second hidden components (i.e., eQTL mechanisms) are represented by  $w^1$  and  $w^2$  and the corresponding effect sizes are by  $e^1$  and  $e^2$ , respectively. \*: The effect size is set to 0.5 for transcript number 30.

Case B				
	$w^1$	$e^1$	$w^2$	$e^2$
B.1	$w_j^1 = 0.577,$ $j = 11, 12, 13.$ $w_j^1 = 0,$ $j = 1, \dots, 10, 14, \dots, 145.$	0.5	$w_j^2 = 0.707,$ $j = 136, 137.$ $w_j^2 = 0,$ $j = 1, \dots, 135, 138, \dots, 145.$	1.5
B.2	same as above	1	same as above	3
B.3	same as above	Unif( $-0.3, 0.3$ )*	same as above	1.5

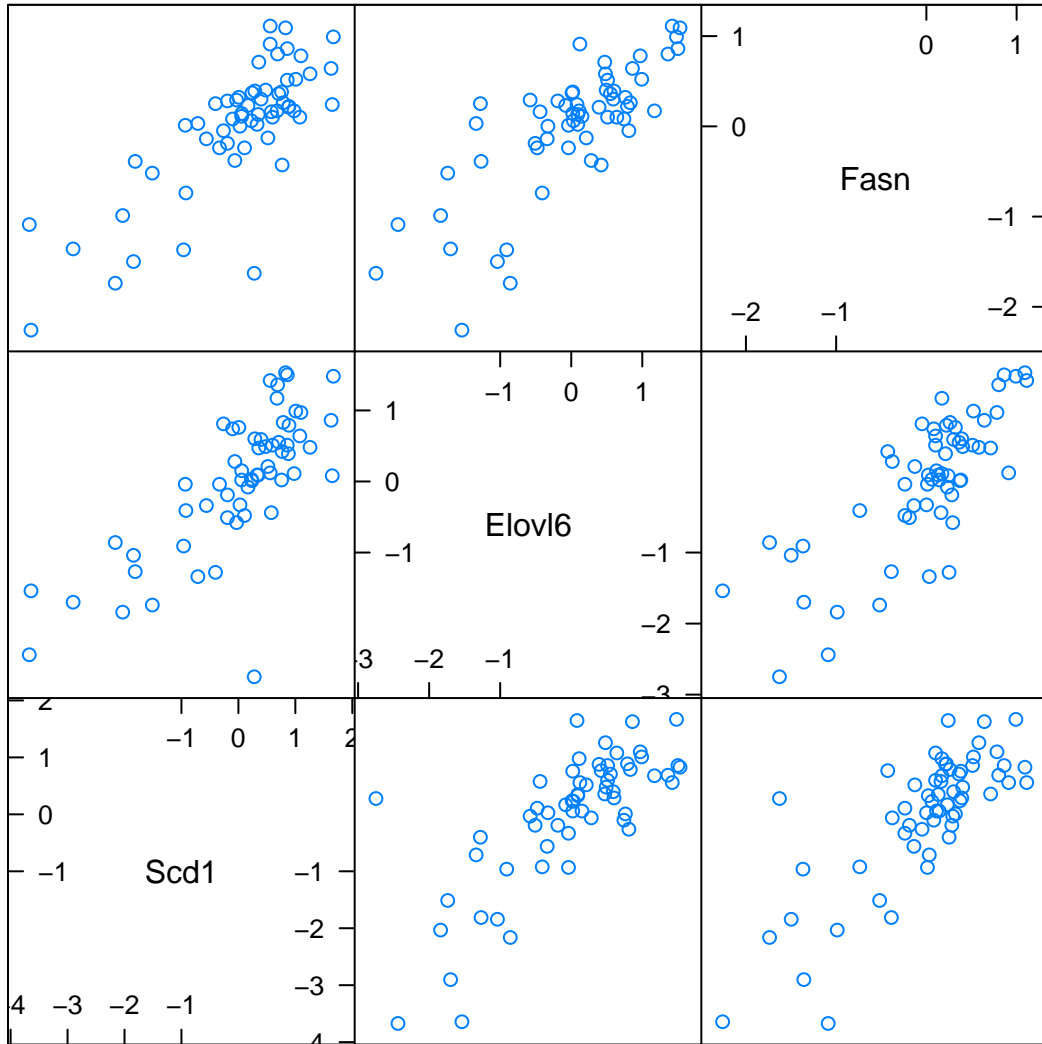


Figure 10: *Pair plots of the expression of three lipid metabolism transcripts across 60 mice.*