Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer

Jie Peng [1], Ji Zhu [2], Anna Bergamaschi [3], Wonshik Han[4], Dong-Young Noh[4], Jonathan R. Pollack [5], Pei Wang[6,*]

[1]Department of Statistics, University of California, Davis, CA, USA;

[2]Department of Statistics, University of Michigan, Ann Arbor, MI, USA;

[3]Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Center, Oslo, Norway;

[4]Cancer Research Institute and Department of Surgery, Seoul National University College of Medicine, Seoul, South Korea;

[5]Department of Pathology, Stanford University, CA, USA;

[6]Division of Public Health Science, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.

December 11, 2008

*Correspondence author: pwang@fhcrc.org

Abstract

In this paper, we propose a new method `remMap` — REgularized Multivariate regression for identifying MAster Predictors — for fitting multivariate response regression models under the high-dimension-low-sample-size setting. `remMap` is motivated by investigating the regulatory relationships among different biological molecules based on multiple types of high dimensional genomic data. Particularly, we are interested in studying the influence of DNA copy number alterations on RNA transcript levels. For this purpose, we model the dependence of the RNA expression levels on DNA copy numbers through multivariate linear regressions and utilize proper regularizations to deal with the high dimensionality as well as to incorporate desired network structures. Criteria for selecting the tuning parameters are also discussed. The performance of the proposed method is illustrated through extensive simulation studies. Finally, `remMap` is applied to a breast cancer study, in which genome wide RNA transcript levels and DNA copy numbers were measured for 172 tumor samples. We identify a tran-hub region in cytoband 17q12-q21, whose amplification influences the RNA expression levels of more than 30 unlinked genes. These findings may lead to a better understanding of breast cancer pathology.

Key words: sparse regression, MAP(MAster Predictor) penalty, DNA copy number alteration, RNA transcript level, v-fold cross validation.

1 Introduction

In a few recent breast cancer cohort studies, microarray expression experiments and array CGH (comparative genomic hybridization) experiments have been conducted for more than 170 primary breast tumor specimens collected at multiple cancer centers (Sorlie et al. 2001; Sorlie et al. 2003; Zhao et al. 2004; Kapp et al. 2006;

Bergamaschi et al. 2006; Langerod et al. 2007; Bergamaschi et al. 2008). The resulted RNA transcript levels (from the microarray expression experiments) and DNA copy numbers (from the CGH experiments) of about $20K$ genes/clones across all the tumor samples were then used to identify useful molecular markers for potential clinical usage. While useful information has been revealed by analyzing expression arrays alone or CGH arrays alone, careful *integrative analysis* of DNA copy numbers and expression data are necessary as these two types of data provide complimentary information in gene characterization. Specifically, RNA data give information on genes that are over/under-expressed, but does not distinguish primary changes driving cancer from secondary changes resulting from cancer, such as proliferation rates and differentiation state. On the other hand, DNA data give information on gains and losses that are drivers of cancer. Therefore, integrating DNA and RNA data provides more complete information. Particularly, this helps to discern more subtle (yet biologically important) genetic regulatory relationships in cancer cells (Pollack et al. 2002).

It is widely agreed that variations in gene copy numbers play an important role in cancer development through altering the expression levels of cancer-related genes (Albertson et al. 2003). This is clear for *cis-regulations*, in which a gene's DNA copy number alteration influences its own RNA transcript level (Hyman et al. 2002; Pollack et al. 2002). However, DNA copy number alterations can also alter in trans the RNA transcript levels of genes from unlinked regions, for example by directly altering the copy number and expression of transcriptional regulators, or by indirectly altering the expression or activity of transcriptional regulators, or through genome rearrangements affecting cis-regulatory elements. The functional consequences of such *trans-regulations* are much harder to establish, as such inquiries involve assessment of a large number of potential regulatory relationships. Therefore, to refine our under-

standing of how these genome events exert their effects, we need new analytical tools that can reveal the subtle and complicated interactions among DNA copy numbers and RNA transcript levels. Knowledge resulting from such analysis will help shed light on cancer mechanisms.

The most straightforward way to model the dependence of RNA levels on DNA copy numbers is through a multivariate response linear regression model with the RNA levels being responses and the DNA copy numbers being predictors. While the multivariate linear regression is well studied in statistical literature, the current problem bears new challenges due to (i) high-dimensionality in terms of both predictors and responses; (ii) the interest in identifying *master regulators* in genetic regulatory networks; and (iii) the complicated relationships among response variables. Thus, the naive approach of regressing each response onto the predictors separately is unlikely to produce satisfactory results, as such methods often lead to high variability and over-fitting. This has been observed by many authors, for example, Breiman et al. (1997) show that taking into account of the relation among response variables helps to improve the overall prediction accuracy.

When the number of predictors is moderate or large, model selection is often needed for prediction accuracy and/or model interpretation. Standard model selection tools in multiple regression such as AIC and forward stepwise selection have been extended to multivariate linear regression models (Bedrick et al. 1994; Fujikoshi et al. 1997; Lutz and Bühlmann 2006). More recently, sparse regularizations have been utilized for model selection under high dimensional multivariate regression setting. For example, Turlach et al. (2005) propose to constrain the coefficient matrix of a multivariate regression model to lie within a suitable polyhedral region. Lutz and Bühlmann (2006) propose an L_2 multivariate boosting procedure. Brown et al. (1998, 1999, 2002) introduce a Bayesian framework to model the relation among the response

variables when performing variable selection for multivariate regression. Another way to reduce the dimensionality is through factor analysis. Related work includes Izenman (1975), Frank et al. (1993), Reinsel and Velu (1998), Yuan et al. (2007) and many others.

For the problem we are interested in here, the dimensions of both predictors and responses are large (compared to the sample size). Thus in addition to assume a sparse model, i.e., not all predictors affect a response, it is also reasonable to assume that a predictor may affect only some but not all responses. Moreover, in many real applications, there often exist a subset of predictors which are more important than other predictors in terms of model building and/or scientific interest. For example, it is widely believed that genetic regulatory relationships are intrinsically sparse (Jeong et al. 2001; Gardner et al. 2003). At the same time, there exist *master regulators* — network components that affect many other components, which play important roles in shaping the network functionality. Most methods mentioned above do not take into account the dimensionality of the responses, and thus a predictor/factor influences either all or none responses, e.g., Turlach et al. (2005), Yuan et al. (2007), and the L_2 row boosting by Lutz and Bühlmann (2006). On the other hand, other methods only impose a sparse model, but do not aim at selecting a subset of predictors, e.g., the L_2 boosting by Lutz and Bühlmann (2006). In this paper, we propose a novel method `remMap` — REgularized Multivariate regression for identifying MAster Predictors, which takes into account both aspects. `remMAP` uses an ℓ_1 norm penalty to control the overall sparsity of the coefficient matrix of the multivariate linear regression model. In addition, `remMap` imposes a "group" sparse penalty, which in essence is the same as the "group lasso" penalty proposed by Antoniadis and Fan (2001), Bakin (1999), Yuan and Lin (2006) and Zhao et al. (2006) (see more discussions in Section 2). This penalty puts a constraint on the ℓ_2 norm of regression coefficients for each

predictor, which controls the total number of predictors entering the model, and consequently facilitates the detection of *master predictors*. The performance of the proposed method is illustrated through extensive simulation studies. We also apply the `remMap` method on the breast cancer data set mentioned earlier. We identify a significant trans-hub region in cytoband 17q12-q21, whose amplification influences the RNA levels of more than 30 unlinked genes. These findings may shed some light on breast cancer pathology.

The rest of the paper is organized as follows. In Section 2, we describe the `remMap` model, its implementation and criteria for tuning. In Section 3, the performance of `remMap` is examined through extensive simulation studies. In Section 4, we apply the `remMap` method on a breast cancer data set. We conclude the paper with discussions in Section 5. Technical details are provided in the supplementary material.

2 Method

2.1 Model

Consider multivariate regression with Q response variables $y_1, \cdots, y_Q$ and P prediction variables $x_1, \cdots, x_P$:

$$y_q = \sum_{p=1}^{P} x_p \beta_{pq} + \epsilon_q, \quad q = 1, \cdots, Q, \tag{1}$$

where the error terms $\epsilon_1, \cdots, \epsilon_q$ have a joint distribution with mean 0 and covariance $\mathbf{\Sigma}_\epsilon$. The primary goal of this paper is to identify non-zero entries in the $P \times Q$ coefficient matrix $\mathbf{B} = (\beta_{pq})$ based on N i.i.d samples from the above model. Under normality assumptions, β_{pq} can be interpreted as the conditional correlation $\mathrm{Cor}(y_q, x_p | x_{-(p)})$, where $x_{-(p)} := \{x_{p'} : 1 \leq p' \neq p \leq P\}$. In the following, we use $Y_q = (y_q^1, \cdots, y_q^N)^T$

and $X_p = (x_p^1, \cdots, x_p^N)^T$ to denote the sample of the q^{th} response variable and that of the p^{th} prediction variable, respectively. We also use $\mathbf{Y} = (Y_1 : \cdots : Y_Q)$ to denote the $N \times Q$ response matrix, and use $\mathbf{X} = (X_1 : \cdots : X_P)$ to denote the $N \times P$ prediction matrix.

In this paper, we shall focus on the cases where both Q and P are larger than the sample size N. For example, in the breast cancer study discussed in Section 4, the sample size is 172, while the number of genes and the number of chromosomal regions are on the order of a couple of hundreds (after pre-screening). When $P > N$, the ordinary least square solution does not exist, and some sort of regularization becomes indispensable. The choice of suitable regularization depends heavily on the type of data structure we envision. In recent years, ℓ_1-norm based sparsity constraints such as *lasso* (Tibshirani 1996) have been widely used under such high-dimension-low-sample-size setting. This kind of regularization is particularly suitable for the study of genetic pathways, since genetic regulatory relationships are widely believed to be intrinsically sparse (Jeong et al. 2001; Gardner et al. 2003). In this paper, we impose an ℓ_1 norm penalty on the coefficient matrix $\mathbf{B}$ to control the overall sparsity of the multivariate regression model. In addition, we put constraints on the total number of predictors entering the model. This is achieved by treating the coefficients corresponding to the same predictor (one row of $\mathbf{B}$) as a group, and then penalizing their ℓ_2 norm. A predictor will not be selected into the model if the corresponding ℓ_2 norm is shrunken to 0. Thus this penalty facilitates the identification of *master predictors* — predictors which affect (relatively) many response variables. This idea is motivated by the fact that master regulators exist and are of great interest in the study of many real life networks including genetic regulatory networks. Specifically,

for model (1), we propose the following criterion

$$L(\mathbf{B};\lambda_1,\lambda_2)=\frac{1}{2}||\mathbf{Y}-\sum_{p=1}^{P}X_pB_p||_F^2+\lambda_1\sum_{p=1}^{P}||C_p\cdot B_p||_1+\lambda_2\sum_{p=1}^{P}||C_p\cdot B_p||_2, \qquad (2)$$

where $\mathbf{C}=(c_{pq})=(C_1^T:\cdots:C_P^T)^T$ is a pre-specified $P\times Q$ 0-1 matrix indicating on which coefficients penalization is imposed; B_p is the p^{th} row of $\mathbf{B}$; $||\cdot||_F$ denotes the Frobenius norm of matrices; $||\cdot||_1$ and $||\cdot||_2$ are the ℓ_1 and ℓ_2 norms for vectors, respectively; and "$\cdot$" stands for entry-wise multiplication. The indicator matrix $\mathbf{C}$ is pre-specified based on prior knowledge: if we know in advance that predictor x_p affects response y_q, then the corresponding regression coefficient β_{pq} will not be penalized and we set $c_{pq}=0$ (see Section 4 for an example). Finally, an estimate of the coefficient matrix $\mathbf{B}$ is $\widehat{\mathbf{B}}(\lambda_1,\lambda_2):=\arg\min_{\mathbf{B}}L(\mathbf{B};\lambda_1,\lambda_2)$.

In the above loss function, the ℓ_1 penalty induces the overall sparsity of the coefficient matrix $\mathbf{B}$. The ℓ_2 penalty on the row vectors $C_p\cdot B_p$ induces row sparsity of the product matrix $\mathbf{C}\cdot\mathbf{B}$. As a result, some rows are shrunken to be entirely zero (Theorem 1). Consequently, predictors which affect relatively more response variables are more likely to be selected into the model. We refer to the combined penalty in equation (2) as the `MAP` (MAster Predictor) penalty. We also refer to the proposed estimator $\widehat{\mathbf{B}}(\lambda_1,\lambda_2)$ as the `remMap` (REgularized Multivariate regression for identifying MAster Predictors) estimator. Note that, the ℓ_2 penalty is a special case (with $\alpha=2$) of the more general penalty form: $\sum_{p=1}^{P}||C_p\cdot B_p||_\alpha$, where $||v||_\alpha:=(\sum_{q=1}^{Q}|v_q|^\alpha)^{\frac{1}{\alpha}}$ for a vector $v\in\mathcal{R}^Q$ and $\alpha>0$. In Turlach et al. (2005), a penalty with $\alpha=\infty$ is used to select a common subset of predictor variables when modeling multivariate responses. In Yuan et al. (2007), a constraint with $\alpha=2$ is applied to the loading matrix in a multivariate linear factor regression model for dimension reduction. In the case of multiple regression setting, a similar penalty corresponding to $\alpha=2$ is

proposed by Bakin (1999) and by Yuan and Lin (2006) for the selection of grouped variables, which corresponds to the blockwise additive penalty in Antoniadis and Fan (2001) for wavelet shrinkage. Zhao et al. (2006) propose the penalty with a general α. However, none of these methods takes into account the high dimensionality of response variables and thus predictors/factors are simultaneously selected for all responses. On the other hand, by combining the ℓ_2 penalty and the ℓ_1 penalty together in the `MAP` penalty, the `remMAP` model not only selects a subset of predictors, but also limits the selected predictors to influence only some (but not all) response variables. Thus, it is more suitable for the cases when both the number of predictors and the number of responses are large.

In Section 3, we use extensive simulation studies to illustrate the effects of the `MAP` penalty. We compare the `remMAP` method with two alternatives: (i) the `joint` method which only utilizes the ℓ_1 penalty, that is $\lambda_2 = 0$ in (2); (ii) the `sep` method which performs Q separate lasso regressions. We find that, when there exist large hubs (master predictors), `remMAP` performs much better than `joint` in terms of identifying the true model; otherwise, the two methods perform similarly. This means that "simultaneous" variables selection enhanced by the ℓ_2 penalty pays off when there exist a small subset of "important" predictors and it costs little when such predictors are absent. In addition, both `remMAP` and `joint` methods impose sparsity of the coefficient matrix as a whole. This helps to incorporate information across different regressions and also amounts to a greater degree of regularization, which is usually desirable for the high-dimension-low-sample-size setting. On the other hand, the `sep` method controls sparsity for each individual regression separately and thus is subject to high variability and over-fitting. This is also noted by other authors including Turlach et al. (2005) and Lutz and Bühlmann (2006).

2.2 Model Fitting

In this section, we propose an iterative algorithm for solving the `remMAP` estimator $\widehat{\mathbf{B}}(\lambda_1, \lambda_2)$, which is a convex optimization problem when the two tuning parameters are not both zero. We first describe how to update one row of $\mathbf{B}$, when all other rows are fixed.

Theorem 1 *Given* $\{B_p\}_{p\neq p_0}$ *in (2), the solution for* $\min_{B_{p_0}} L(\mathbf{B}; \lambda_1, \lambda_2)$ *is given by* $\widehat{B}_{p_0} = (\widehat{\beta}_{p_0,1}, \cdots, \widehat{\beta}_{p_0,Q})$ *which satisfies: for* $1 \leq q \leq Q$

(i) If $c_{p_0,q} = 0$, $\widehat{\beta}_{p_0,q} = X_{p_0}^T \widetilde{Y}_q / \|X_{p_0}\|_2^2$ *(OLS), where* $\widetilde{Y}_q = Y_q - \sum_{p\neq p_0} X_p \beta_{pq}$;

(ii) If $c_{p_0,q} = 1$,

$$\widehat{\beta}_{p_0,q} = \begin{cases} 0, & \text{if } ||\widehat{\mathrm{B}}_{\mathrm{p_0}}^{\mathrm{lasso}}||_{2,\mathrm{C}} = 0; \\ \left(1 - \frac{\lambda_2}{||\widehat{\mathbf{B}}_{p_0}^{\mathrm{lasso}}||_{2,C} \cdot \|X_{p_0}\|_2^2}\right)_+ \widehat{\beta}_{p_0,q}^{\mathrm{lasso}}, & \text{otherwise,} \end{cases} \tag{3}$$

where $||\widehat{B}_{p_0}^{\mathrm{lasso}}||_{2,C} := \left\{\sum_{q=1}^{Q} c_{p_0,q} (\widehat{\beta}_{p_0,q}^{\mathrm{lasso}})^2\right\}^{1/2}$, *and*

$$\widehat{\beta}_{p_0,q}^{\mathrm{lasso}} = \begin{cases} X_{p_0}^T \widetilde{Y}_q / \|X_{p_0}\|_2^2, & \text{if } \mathrm{c_{p_0,q}} = 0; \\ \left(|X_{p_0}^T \widetilde{Y}_q| - \lambda_1\right)_+ \frac{\mathrm{sign}\left(\mathrm{X_{p_0}^T \widetilde{Y}_q}\right)}{\|X_{p_0}\|_2^2}, & \text{if } \mathrm{c_{p_0,q}} = 1. \end{cases} \tag{4}$$

The proof of Theorem 1 is given in the supplementary material (Appendix A).

Theorem 1 says that, when estimating the p_0^{th} row of the coefficient matrix $\mathbf{B}$ with all other rows fixed: if there is a pre-specified relationship between the p_0^{th} predictor and the q^{th} response (i.e., $c_{p_0,q} = 0$), the corresponding coefficient $\beta_{p_0,q}$ is estimated by the (univariate) ordinary least square solution (OLS) using current responses $\widetilde{\mathbf{Y}}$; otherwise, we first obtain the lasso solution $\widehat{\beta}_{p_0,q}^{\mathrm{lasso}}$ by the (univariate) soft shrinkage of the OLS solution (equation (4)), and then conduct a group shrinkage of the lasso

solution (equation (3)). From Theorem 1, it is easy to see that, when the design matrix $\mathbf{X}$ is orthonormal: $\mathbf{X}^T\mathbf{X} = I_p$ and $\lambda_1 = 0$, the `remMAP` method amounts to selecting variables according to the ℓ_2 norm of their corresponding OLS estimates.

Theorem 1 naturally leads to an algorithm which updates the rows of $\mathbf{B}$ iteratively until convergence. In particular, we adopt the `active-shooting` idea proposed by Peng et al. (2008) and Friedman et al. (2008), which is a modification of the `shooting` algorithm proposed by Fu (1998) and also Friedman et al. (2007) among others. The algorithm proceeds as follows:

1. Initial step: for $p = 1, ..., P$; $q = 1, ..., Q$,

$$\widehat{\beta}^0_{p,q} = \begin{cases} X_p^T Y_q / \|X_p\|_2^2, & \text{if} \quad c_{p,q} = 0; \\ \left(|X_p^T Y_q| - \lambda_1\right)_+ \frac{\mathrm{sign}\left(\mathrm{X_p^T Y_q}\right)}{\|X_p\|_2^2}, & \text{if} \quad c_{p,q} = 1. \end{cases} \tag{5}$$

2. Define the current *active-row set* $\Lambda = \{p : \text{ current } ||\widehat{B}_p||_{2,C} \neq 0\}$.

   (2.1) For each $p \in \Lambda$, update $\widehat{B}_p$ with all other rows of $\mathbf{B}$ fixed at their current values according to Theorem 1.

   (2.2) Repeat (2.1) until convergence is achieved on the current active-row set.

3. For $p = 1$ to P, update $\widehat{B}_p$ with all other rows of $\mathbf{B}$ fixed at their current values according to Theorem 1. If no $\widehat{B}_p$ changes during this process, return the current $\widehat{\mathbf{B}}$ as the final estimate. Otherwise, go back to step 2.

It is clear that the computational cost of the above algorithm is in the order of $O(NPQK)$, where K is the total number of iterations. The value of K depends on the overall sparsity of the final estimator $\widehat{\mathbf{B}}$, which is controlled by the tuning parameters.

2.3 Tuning

In this section, we discuss the selection of the tuning parameters (λ_1, λ_2). We briefly describe two different approaches: one based on a BIC criterion and another based on multi-fold cross validation.

In model (1), by assuming $\epsilon_q \sim \text{Normal}(0, \sigma^2_{q,\epsilon})$, the BIC criterion for the q^{th} regression can be defined as

$$\text{BIC}_q(\widehat{\beta}_{1q}, \cdots, \widehat{\beta}_{Pq}; \text{df}_q) = N \times \log(\text{RSS}_q) + \log N \times \text{df}_q, \tag{6}$$

where $\text{RSS}_q := \sum_{n=1}^{N}(y_q^n - \widehat{y}_q^n)^2$ with $\widehat{y}_q^n = \sum_{p=1}^{P} x_p^n \widehat{\beta}_{pq}$; and df_q is the degrees of freedom which is defined as (see supplementary material (Appendix B) for more details)

$$\text{df}_q = \text{df}_q(\widehat{\beta}_{1q}, \cdots, \widehat{\beta}_{Pq}) := \sum_{n=1}^{N} \text{Cov}(\widehat{y}_q^n, y_q^n)/\sigma^2_{q,\epsilon}. \tag{7}$$

For a given pair of (λ_1, λ_2), We then define the (overall) BIC criterion at (λ_1, λ_2):

$$\text{BIC}(\lambda_1, \lambda_2) = N \times \sum_{q=1}^{Q} \log(\text{RSS}_q(\lambda_1, \lambda_2)) + \log N \times \sum_{q=1}^{Q} \text{df}_q(\lambda_1, \lambda_2). \tag{8}$$

Efron et al. (2004) derive an explicit formula for the degrees of freedom of *lars* under orthogonal design. Similar strategy are also used by Yuan and Lin (2006) among others. In the following theorem, we follow the same idea and derive an unbiased estimator of df_q for `remMAP` when the columns of $\mathbf{X}$ are orthogonal to each other.

Theorem 2 *Suppose* $X_p^T X_{p'} = 0$ *for all* $1 \le p \ne p' \le P$. *Then for given* (λ_1, λ_2),

$$\begin{aligned} \widehat{df}_q(\lambda_1, \lambda_2) &:= \sum_{p=1}^{P} c_{pq} \times \mathbb{I}\left(||\widehat{B}_p^{\text{lasso}}||_{2,C} > \frac{\lambda_2}{||X_p||_2^2} \right) \times \mathbb{I}\left(|\widehat{\beta}_{pq}^{\text{ols}}| > \frac{\lambda_1}{||X_p||_2^2} \right) \\ &\times \left(1 - \frac{\lambda_2}{||X_p||_2^2} \frac{||\widehat{B}_p^{\text{lasso}}||_{2,C}^2 - (\widehat{\beta}_{pq}^{\text{lasso}})^2}{||\widehat{B}_p^{\text{lasso}}||_{2,C}^3} \right) + \sum_{p=1}^{P} (1 - c_{p,q}) \end{aligned} \tag{9}$$

is an unbiased estimator of the degrees of freedom $\text{df}_q(\lambda_1, \lambda_2)$ *(defined in equation (7)) of the* `remMAP` *estimator* $\widehat{\mathbf{B}} = \widehat{\mathbf{B}}(\lambda_1, \lambda_2) = (\widehat{\beta}_{pq}(\lambda_1, \lambda_2))$. *Here, under the orthogonal design,* $\widehat{\beta}_{pq}, \widehat{\beta}_{pq}^{\text{lasso}}$ *are given by Theorem 1 with* $\widetilde{Y}_q = Y_q$ $(q = 1, \cdots, Q)$, *and* $\widehat{\beta}_{pq}^{\text{ols}} := \frac{X_p^T Y_q}{||X_p||_2^2}$.

Theorem 2 is proved in the supplementary material (Appendix B). In Section 3, we show by extensive simulation studies that, as long as the correlations among the predictors $x_1, \cdots, x_P$ are not too complicated, (9) is a pretty good estimator of the degrees of freedom. However, when the correlations among the predictors are complicated, (9) tends to severely overestimate the actual degrees of freedom, and consequently the criterion (8) tends to select very small models.

As an alternative, v-fold cross validation is another commonly used tuning strategy. While it is computationally more demanding than BIC, v-fold cross validation requires much fewer assumptions and thus is more robust. To perform the v-fold cross validation, we first partition the whole data set into V non-overlapping subsets, each consisting of approximately $1/V$ fraction of total samples. Denote the i^{th} subset as $D^{(i)} = (\mathbf{Y}^{(i)}, \mathbf{X}^{(i)})$, and its complement as $D^{-(i)} = (\mathbf{Y}^{-(i)}, \mathbf{X}^{-(i)})$. For a given (λ_1, λ_2), we obtain the `remMAP` estimate: $\widehat{\mathbf{B}}^{(i)}(\lambda_1, \lambda_2) = (\widehat{\beta}_{pq}^{(i)})$ based on the i^{th} training set $D^{-(i)}$. We then obtain the *ordinary least square estimates* $\widehat{\mathbf{B}}_{\text{ols}}^{(i)}(\lambda_1, \lambda_2) = (\widehat{\beta}_{\text{ols},pq}^{(i)})$ as follows: for $1 \le q \le Q$, define $S_q = \{p : 1 \le p \le P, \widehat{\beta}_{pq}^{(i)} \ne 0\}$. Then set $\widehat{\beta}_{\text{ols},pq}^{(i)} = 0$ if $p \notin S_q$; otherwise, define $\{\widehat{\beta}_{\text{ols},pq}^{(i)} : p \in S_q\}$ as the ordinary least square estimates by regressing $Y_q^{-(i)}$ onto $\{X_p^{-(i)} : p \in S_q\}$. Finally, prediction error is calculated on the

test set $D^{(i)}$:

$$\texttt{remMAP.cv}_i(\lambda_1, \lambda_2) := ||\mathbf{Y}^{(i)} - \mathbf{X}^{(i)}\widehat{\mathbf{B}}^{(i)}_{\text{ols}}(\lambda_1, \lambda_2)||_2^2. \tag{10}$$

The v-fold cross validation score is then defined as

$$\texttt{remMAP.cv}(\lambda_1, \lambda_2) = \sum_{i=1}^{V} \texttt{remMAP.cv}_i(\lambda_1, \lambda_2). \tag{11}$$

The reason to use OLS estimates in calculating the prediction error is because the true model is assumed to be sparse. As noted by Efron et al. (2004), when there are many noise variables, using shrunken estimates in the cross validation criterion often results in over fitting. Similar results are observed in our simulation studies: if in (10) and (11), the shrunken estimates are used, the selected models are all very big which result in large numbers of false positive findings. In addition, we also try AIC and GCV for tuning and both criteria result in over fitting as well. These results are not reported in the next section due to space limitation.

3 Simulation

In this section, we investigate the performance of the `remMap` method and two alternatives coupled with two tuning strategies:

1. `remMap.cv`: `remMap` with (λ_1, λ_2) selected by 10-fold cross validation (11);
2. `remMap.bic`: `remMap` with (λ_1, λ_2) selected by BIC criterion (8) and degrees of freedom estimated by (9);
3. `joint.cv`: `remMap` with $\lambda_2 = 0$ and λ_1 selected by 10-fold cross validation (11);
4. `joint.bic`: `remMap` with $\lambda_2 = 0$ and λ_1 selected by BIC criterion (8) with degrees of freedom estimated by (9);

5. `sep.cv`: Q individual *lasso* regressions with the tuning parameter for each regression selected separately by 10-fold cross validation;
6. `sep.bic`: Q individual *lasso* regressions with the tuning parameter for each regression selected separately by a BIC criterion. Here, for each lasso regression, the degrees of freedom is estimated by the total number of selected predictors (Zou et al. 2007).

We simulate data as follows. Given (N, P, Q), we first generate the predictors $(x_1, \cdots, x_P)^T \sim \text{Normal}_\text{P}(0, \mathbf{\Sigma}_\text{X})$, where $\mathbf{\Sigma}_X$ is the predictor covariance matrix (for simulations 1 and 2, $\mathbf{\Sigma}_X(p, p') := \rho_x^{|p-p'|}$). Next, we simulate a $P \times Q$ 0-1 adjacency matrix $\mathbf{A}$, which specifies the topology of the network between predictors and responses, with $\mathbf{A}(p, q) = 1$ meaning that x_p influences y_q, or equivalently $\beta_{pq} \neq 0$. In all simulations, we set $P = Q$ and the diagonals of $\mathbf{A}$ equal to one, which is viewed as prior information (thus the diagonals of $\mathbf{C}$ are set to be zero). This aims to mimic `cis-regulations` of DNA copy number alternations on its own expression levels. We then simulate the $P \times Q$ regression coefficient matrix $\mathbf{B} = (\beta_{pq})$ by setting $\beta_{pq} = 0$, if $\mathbf{A}(p, q) = 0$; and $\beta_{pq} \sim \text{Uniform}([-5, -1] \cup [1, 5])$, if $\mathbf{A}(p, q) = 1$. After that, we generate the residuals $(\epsilon_1, \cdots, \epsilon_Q)^T \sim \text{Normal}_\text{Q}(0, \mathbf{\Sigma}_\epsilon)$, where $\mathbf{\Sigma}_\epsilon(q, q') = \sigma_\epsilon^2 \rho_\epsilon^{|q-q'|}$. The residual variance σ_ϵ^2 is chosen such that the average signal to noise ratio equals to a pre-specified level s. Finally, the responses $(y_1, \cdots, y_Q)^T$ are generated according to model (1). Each data set consists of N i.i.d samples of such generated predictors and responses. For all six methods, predictors and responses are standardized to have (sample) mean zero and standard deviation one before model fitting. Results reported for each simulation setting are averaged over 25 independent data sets.

For all simulation settings, $\mathbf{C} = (c_{pq})$ is taken to be $c_{pq} = 0$, if $p = q$; and $c_{pq} = 1$, otherwise. Our primary goal is to identify the `trans-edges` — the predictor-response pairs (x_p, y_q) with $\mathbf{A}(p, q) = 1$ and $\mathbf{C}(p, q) = 1$, i.e., the edges that are not pre-specified by the indicator matrix $\mathbf{C}$. Thus, in the following, we report the number

of false positive detections of `trans-edges` (FP) and the number of false negative detections of `trans-edges` (FN) for each method. We also examine these methods in terms of predictor selection. Specifically, a predictor is called a `cis-predictor` if it does not have any `trans-edges`; otherwise it is called a `trans-predictor`. Moreover, we say a false positive trans-predictor (FPP) occurs if a `cis-predictor` is incorrectly identified as a `trans-predictor`; we say a false negative trans-predictor (FNP) occurs if it is the other way around.

Simulation I

We first assess the performances of the six methods under various combinations of model parameters. Specifically, we consider: $P = Q = 400, 600, 800$; $s = 0.25, 0.5, 0.75$; $\rho_x = 0, 0.4, 0.8$; and $\rho_\epsilon = 0, 0.4, 0.8$. For all settings, the sample size N is fixed at 200. The networks (adjacency matrices $\mathbf{A}$) are generated with 5 master predictors (hubs), each influencing $20 \sim 40$ responses; and all other predictors are `cis-predictors`. We set the total number of `tran-edges` to be 132 for all networks. Results on `trans-edge` detection are summarized in Figures 1 and 2. From these figures, it is clear that `remMAP.cv` performs the best in terms of the total number of false detections (FP+FN), followed by `remMAP.bic`. The two `sep` methods result in too many false positives (especially `sep.cv`). This is expected since the Q tuning parameters are selected separately, and the relations among responses were not utilized at all. This leads to high variability and over-fitting. The two `joint` methods perform reasonably well, though they have considerably larger number of false negative detections compared to `remMAP`. This is because the `joint` methods incorporate less information about the relations among the responses caused by the master predictors. As to the impact of different model parameters, signal size s plays an important role for all six methods: the larger the signal size, the better these methods perform (Figure 1(a)).

On the other hand, the other three factors seem not to have such prominent effects, especially on the two `remMAP` methods. Dimensionality (P, Q) have some impacts on `sep`, but not much on `remMAP` or `joint` (Figure 1(b)). This is presumably because the network complexity does not increase with P, Q (all networks have 132 `trans-edges`). With increasing predictor correlation ρ_x, both `remMAP.bic` and `joint.bic` tend to select smaller models, and consequently result in less false positives and more false negatives (Figure 2(a)). This is due to the fact that, when the design matrix $\mathbf{X}$ is further away from orthogonality, (9) tends to overestimate the degrees of freedom. The residual correlation ρ_ϵ has little impact on `joint` and `sep`, and some (though rather small) impacts on `remMAP` (Figure 2(b)). This is expected since the former two methods depend less on relations among responses. Moreover, `remMAP` performs much better than `joint` and `sep` on predictor selection, especially in terms of the number of false positive `trans-predictors` (results not shown). This is due to the fact that the ℓ_2 norm penalty is more effective than the ℓ_1 norm penalty in screening out `trans-predictors`.

Simulation II

In this simulation, we study the performance of these methods on a network without big hubs. The data is generated similarly as before with $P = Q = 600$, $N = 200$, $s = 0.25$, $\rho_x = 0.4$, and $\rho_\epsilon = 0$. The network consists of 540 `cis-predictors`, and 60 `trans-predictors` with $1 \sim 4$ `trans-edges`. This leads to 151 `trans-edges` in total. As can be seen from Table 1, `remMAP` methods and `joint` methods now perform very similarly and both are considerably better than the `sep` methods. Indeed, under this setting, λ_2 is selected (either by `cv` or `bic`) to be small in the `remMAP` model, making it very close to the `joint` model.

Table 1: Simulation II. Network topology: uniform network with 151 `trans-edges` and 60 `trans-predictors`. $P = Q = 600, N = 200$; $s = 0.25$; $\rho_x = 0.4$; $\rho_\epsilon = 0$.

Method	FP	FN	TF	FPP	FNP
`remMAP.bic`	4.72(2.81)	45.88(4.5)	50.6(4.22)	1.36(1.63)	11(1.94)
`remMAP.cv`	18.32(11.45)	40.56(5.35)	58.88(9.01)	6.52(5.07)	9.2(2)
`joint.bic`	5.04(2.68)	52.92(3.6)	57.96(4.32)	4.72(2.64)	9.52(1.66)
`joint.cv`	16.96(10.26)	46.6(5.33)	63.56(7.93)	15.36(8.84)	7.64(2.12)
`sep.bic`	78.92(8.99)	37.44(3.99)	116.36(9.15)	67.2(8.38)	5.12(1.72)
`sep.cv`	240.48(29.93)	32.4(3.89)	272.88(30.18)	179.12(18.48)	2.96(1.51)

FP: *false positive*; FN: *false negative*; TF: *total false*; FPP: *false positive trans-predictor*; FNP: *false negative trans-predictor. Numbers in the parentheses are standard deviations*

Simulation III

In this simulation, we try to mimic the true predictor covariance and network topology in the real data. We observe that, for chromosomal regions on the same chromosome, the corresponding copy numbers are usually positively correlated, and the magnitude of the correlation decays slowly with genetic distance. On the other hand, if two regions are on different chromosomes, the correlation between their copy numbers could be either positive or negative and in general the magnitude is much smaller than that of the regions on the same chromosome. Thus in this simulation, we first partition the P predictors into 23 distinct blocks, with the size of the i^{th} block proportional to the number of CNAI (copy number alteration intervals) on the i^{th} chromosome of the real data (see Section 4 for the determination of CNAI). Denote the predictors within the i^{th} block as $x_{i1}, \cdots, x_{ig_i}$, where g_i is the size of the i^{th} block. We then define the *within-block* correlation as: $\text{Corr}(x_{ij}, x_{il}) = \rho_{\text{wb}}^{0.5|j-l|}$ for $1 \le j, l \le g_i$; and define the *between-block* correlation as $\text{Corr}(x_{ij,}, x_{kl}) \equiv \rho_{ik}$ for $1 \le j \le g_i$, $1 \le l \le g_k$ and $1 \le i \ne k \le 23$. Here, ρ_{ik} is determined in the following way: its sign is randomly generated from $\{-1, 1\}$; its magnitude is randomly generated from $\{\rho_{\text{bb}}, \rho_{\text{bb}}^2, \cdots, \rho_{\text{bb}}^{23}\}$. In this simulation, we set $\rho_{\text{wb}} = 0.9$, $\rho_{\text{bb}} = 0.25$ and use $P = Q = 600, N = 200$, $s = 0.5$, and $\rho_\epsilon = 0.4$. The heatmaps of the (sample)

correlations of the predictors of the simulated data and those of the real data are given in Figure S-2 of the supplementary material. The network is generated with five large hub predictors each having $14 \sim 26$ `trans-edges`; five small hub predictors each having $3 \sim 4$ `trans-edges`; 20 predictors having $1 \sim 2$ `trans-edges`; and all other predictors are `cis-predictors`. The results are summarized in Table 2. We observe that, `remMAP.bic` and `joint.bic` result in very small models, which is an indicator that (9) now severely overestimates the true degrees of freedom. This is due to the complicated correlation structure among the predictors. It can also be seen that, all three cross-validation based methods have large numbers of false positive findings, even though `remMAP.cv` method is still the best. Thus we propose a method called `cv.vote` to further control the false positive findings. The idea is to treat the training data from each cross-validation fold as a bootstrap sample. Then variables being consistently selected by many cross validation folds should be more likely to appear in the true model than the variables being selected only by few cross validation folds. Specifically, define $s_{pq}(\lambda_1, \lambda_2) = \sum_{i=1}^{V} I(\widehat{\beta}_{pq}^{(i)}(\lambda_1, \lambda_2) \neq 0)$. We then select edge (p, q) if $s_{pq}(\lambda_1, \lambda_2) > V_a$, where V_a is a pre-specified integer. In this simulation, we use $V_a = 5$ and thus `cv.vote` amounts to a "majority vote" procedure. From Table 2, `cv.vote` is very effective in decreasing the number of false positives, while only moderately increasing the number of false negatives for `remMap`. Interestingly, we note that for simulations where `remMap.cv` does not result in too many false positives, `remMap.cv.vote` gives very similar models as `remMap.cv`. For example, for a simulation similar as the one just mentioned, but having a simpler network topology, on average `remMap.cv` results in 2.04 false positive detections and 43.16 false negatives; by applying `remMap.cv.vote` with $V_a = 5$, we get (on average) 0.60 false positives, and 53.68 false negatives (detailed results omitted). These results indicate that `remMap.cv.vote` is an effective criterion in controlling false positive

rates while not sacrificing too much in terms of power.

Table 2: Simulation III. Network topology: five large hubs and five small hubs with 151 `trans-edges` and 30 `trans-predictors`. $P = Q = 600$, $N = 200$; $s = 0.5$; $\rho_{\text{wb}} = 0.9$, $\rho_{\text{bb}} = 0.25$; $\rho_\epsilon = 0.4$.

Method	FP	FN	TF	FPP	FNP
`remMap.bic`	0(0)	150.24(2.11)	150.24(2.11)	0(0)	29.88(0.33)
`remMap.cv`	93.48(31.1)	20.4(3.35)	113.88(30.33)	15.12(6.58)	3.88(1.76)
`remMap.cv.vote`	48.04(17.85)	27.52(3.91)	75.56(17.67)	9.16(4.13)	5.20(1.91)
`joint.bic`	7.68(2.38)	104.16(3.02)	111.84(3.62)	7(2.18)	10.72(1.31)
`joint.cv`	107.12(13.14)	39.04(3.56)	146.16(13.61)	66.92(8.88)	1.88(1.2)
`joint.cv.vote`	63.80(8.98)	47.44(3.90)	111.24(10.63)	41.68(6.29)	2.88(1.30)
`sep.bic`	104.96(10.63)	38.96(3.48)	143.92(11.76)	64.84(6.29)	1.88(1.17)
`sep.cv`	105.36(11.51)	37.28(4.31)	142.64(12.26)	70.76(7.52)	1.92(1.08)
`sep.cv.vote`	13.96(3.14)	96.08(3.59)	110.04(4.09)	0.44(0.51)	17.68(1.35)

FP: *false positive*; FN: *false negative*; TF: *total false*; FPP: *false positive trans-predictor*; FNP: *false negative trans-predictor. Numbers in the parentheses are standard deviations*

4 Real application

In this section, we apply the proposed `remMap` method to the breast cancer study mentioned earlier. Our goal is to search for genome regions whose copy number alterations have significant impacts on RNA expression levels, especially on those of the unlinked genes, i.e., genes not falling into the same genome region. The findings resulting from this analysis may help to cast light on the complicated interactions among DNA copy numbers and RNA expression levels.

4.1 Data preprocessing

The 172 tumor samples were analyzed using cDNA expression microarray and CGH array experiments as described in Sorlie et al. (2001), Sorlie et al. (2003), Zhao et al. (2004), Kapp et al. (2006), Langerod et al. (2007), Bergamaschi et al. (2006), and Bergamaschi et al. (2008). In below, we briefly describe the data preprocessing

steps. More details are provided in the supplementary material (Appendix C).

Each CGH array contains measurements (log_2 ratios) on about $17K$ mapped human genes. A positive (negative) measurement suggests a possible copy number gain (loss). After proper normalization, `cghFLasso` (Tibshirani and Wang 2008) is used to estimate the DNA copy numbers based on array outputs. Then, we derive *copy number alteration intervals* (CNAIs) — basic CNA units (genome regions) in which genes tend to be amplified or deleted at the same time within one sample — by employing the Fixed-Order Clustering (FOC) method (Wang 2004). In the end, for each CNAI in each sample, we calculate the mean value of the estimated copy numbers of the genes falling into this CNAI, which results in a 172 (samples) by 384 (CNAIs) numeric matrix.

Each expression array contains measurements for about $18K$ mapped human genes. After global normalization for each array, we also standardize each gene's measurements across 172 samples to median= 0 and MAD (median absolute deviation) = 1. Then we focus on a set of 654 breast cancer related genes, which is derived based on 7 published breast cancer gene lists (Sorlie et al. 2003; van de Vijver et al. 2002; Chang et al. 2004; Paik et al. 2004; Wang et al. 2005; Sotiriou et al. 2006; Saal et al. 2007). This results in a 172 (samples) by 654 (genes) numeric matrix.

As mentioned earlier, RNA transcription levels usually have complex correlation structure, which needs to be taken into account in modeling the influence of CNAIs on RNA levels. For this purpose, we apply the `space` (Sparse PArtial Correlation Estimation) method to search for associated RNA pairs through identifying non-zero partial correlations (Peng et al. 2008). The resulting (concentration) network (referred to as *Exp.Net.664* hereafter) has in total 664 edges — 664 pairs of genes whose RNA levels significantly correlated with each other after accounting for the expression levels of other genes.

Another important factor one needs to consider when studying breast cancer is the existence of distinct tumor subtypes. Population stratification due to these distinct subtypes could confound our detection of associations between CNAIs and gene expressions. Therefore, we introduce a set of subtype indicator variables, which later on is used as additional predictors in the `remMap` model. Specifically, we derive subtype labels based on expression patterns by following Sorlie et al. (2003), and divide the 172 patients into 5 distinct groups, which corresponds to the 5 subtypes suggested by Sorlie et al. (2003) — Luminal Subtype A, Luminal Subtype B, ERBB2-overexpressing Subtype, Basal Subtype and Normal Breast-like Subtype.

4.2 Interactions between CNAIs and RNA expressions

We then apply the `remMap` method to study the interactions between CNAIs and RNA transcript levels. First, for each of the 654 breast cancer genes, we regress its expression level on three sets of predictors: (i) expression levels of other genes that are connected to the target gene (the current response variable) in *Exp.Net.664*; (ii) the five subtype indicator variables derived in the previous section; and (iii) the copy numbers of all 384 CNAIs. We are interested in whether any unlinked CNAIs are selected into this regression model (i.e., the corresponding regression coefficients are non-zero). This suggests potential trans-regulations (`trans-edges`) between the selected CNAIs and the target gene expression. The coefficient of the linked CNAI of the target gene are not included in the `MAP` penalty (this corresponds to $c_{pq} = 0$, see Section 2 for details). This is because the DNA copy number changes of one gene often influence its own expression level, and we are also less interested in this kind of cis-regulatory relationships (`cis-edges`) here. No penalties are imposed on the expressions of connected genes either. In another word, we view the cis-regulations between CNAIs and their linked expression levels, as well as the inferred

RNA interaction network as "prior knowledge" in our study.

We select tuning parameters (λ_1, λ_2) in the `remMap` model through a 10-fold cross validation as described in Section 2.3. The optimal (λ_1, λ_2) corresponding to the smallest CV score from a grid search is $(355.1, 266.7)$. The resulting model contains 56 trans-regulations in total. In order to further control false positive findings, we apply the `remMap.cv.vote` procedure, and filter away 13 out of these 56 `trans-edges` which have not been consistently selected across different CV folds. The remaining 43 `trans-edges` correspond to three contiguous CNAIs on chromosome 17 and 31 distinct (unlinked) RNAs. Figure 3 illustrates the topology of the inferred regulatory relationships. The detailed annotations of the three CNAIs and 31 RNAs are provided in Table 3 and Table 4. Moreover, the Pearson-correlations between the DNA copy

Table 3: Genome locations of the three CNAIs having trans-regulations.

Index	Cytoband	Begin[1]	End [1]	# of clones[2]	# of Trans-Reg[3]
1	17q12-17q12	34811630	34811630	1	12
2	17q12-17q12	34944071	35154416	9	30
3	17q21.1-17q21.2	35493689	35699243	7	1

1. Nucleotide position (bp).
2. Number of genes/clones on the array falling into the CNAI.
3. Number of unlinked genes whose expression were regulated by the copy number of the CNAI.

numbers of CNAIs and the expression levels of the regulated genes/clones (including both `cis-regulation` and `trans-regulation`) across the 172 samples are reported in Table 4. As expected, all the cis-regulations have much higher correlations than the potential trans-regulations. In addition, none of the subtype indicator variables are selected into the final model, which implies that the detected associations between copy numbers of CNAIs and gene expressions are unlikely due to the stratification of the five tumor subtypes.

The three CNAIs being identified as trans-regulators sit closely on chromosome 17, spanning from $34811630bp$ to $35699243bp$ and falling into cytoband 17q12-q21.2.

This region (referred to as CNAI-17q12 hereafter) contains 24 known genes, including the famous breast cancer oncogene ERBB2, and the growth factor receptor-bound protein 7 (GRB7). The over expression of GRB7 plays pivotal roles in activating signal transduction and promoting tumor growth in breast cancer cells with chromosome 17q11-21 amplification (Bai and Louh 2008). In this study, CNAI-17q12 was highly amplified (normalized log_2 ratio$>$ 5) in 33 (19%) out of the 172 tumor samples. Among the 654 genes/clones considered in the above analysis, 8 clones (corresponding to six genes including ERBB2, GRB7, and MED24) fall into this region. The expressions of these 8 clones are all up-regulated by the amplification of CNAI-17q12 (see Table 4 for more details), which is consistent with results reported in the literature (Kao and Pollack 2006). More importantly, as suggested by the final `remMap` model, the amplification of CNAI-17q12 also influences the expression levels of 31 unlinked genes/clones. This suggests that CNAI-17q12 may harbor transcriptional factors whose activities closely relate to breast cancer. Indeed, there are 4 transcription factors (NEUROD2, IKZF3, THRA, NR1D1) and 2 transcriptional co-activators (MED1, MED24) in CNAI-17q12. It is possible that the amplification of CNAI-17q12 results in the over expression of one or more transcription factors/co-activators in this region, which then influence the expressions of the unlinked 31 genes/clones. Interestingly, some of the 31 genes/clones have been reported to have functions directly related to cancer and may serve as potential drug targets. For example, AGTR1 is a receptor whose genetic polymorphisms have been reported to associate with breast cancer risk and is possibly druggable (Koh et al. 2005). CDH3 encodes a cell-cell adhesion glycoprotein and is deemed as a candidate of tumor suppressor gene, as disturbance of intracellular adhesion is important for invasion and metastasis of tumor cells (Kremmidiotis et al. 1998). PEG3 is a mediator between p53 and Bax in DNA damage-induced neuronal death (Johnson et al. 2002) and may function as a tumor

suppressor gene (Dowdy et al. 2005). In a word, these 31 genes may play functional roles in the pathogenesis of breast cancer and may serve as additional targets for therapy. In the end, we want to point out that, besides RNA interactions and subtype stratification, there could be other unaccounted confounding factors. Therefore, caution must be applied when one tries to interpret these results.

5 Discussion

In this paper, we propose the `remMap` method for fitting multivariate regression models under the large P, Q setting. We focus on model selection, i.e., the identification of relevant predictors for each response variable. `remMap` is motivated by the rising needs to investigate the regulatory relationships between different biological molecules based on multiple types of high dimensional omics data. Such genetic regulatory networks are usually intrinsically sparse and harbor hub structures. Identifying the hub regulators (master regulators) is of particular interest, as they play crucial roles in shaping network functionality. To tackle these challenges, `remMap` utilizes a `MAP` penalty, which consists of an ℓ_1 norm part for controlling the overall sparsity of the network, and an ℓ_2 norm part for further imposing a row-sparsity of the coefficient matrix. This combined regularization takes into account both model interpretability and computational tractability. Specifically, the ℓ_2 norm penalty facilitates the detection of master predictors (regulators). As illustrated in Section 3, using the `MAP` penalty greatly improves the performance on both edge detection and master predictor identification.

We then apply the `remMap` method on a breast cancer data set. Our goal is to investigate the influences of DNA copy number alterations on RNA transcript levels based on 172 breast cancer tumor samples. The resulting model suggests the existence of a trans-hub region on cytoband 17q12-q21, whose amplification influences

RNA levels of 31 unlinked genes. Cytoband 17q12-q21 is a well known hot region for breast cancer, which harbors the oncogene ERBB2. The above results suggest that this region may also harbor important transcriptional factors. One way to verify this conjecture is through a sequence analysis to search for common motifs in the upstream regions of the 31 RNA transcripts, which remains as our future work.

Besides the above application, the `remMap` model can be applied to investigate the regulatory relationships between other types of biological molecules. For example, it is of great interest to understand the influence of single nucleotide polymorphism (SNP) on RNA transcript levels, as well as the influence of RNA transcript levels on protein expression levels. Such investigation will improve our understanding of related biological systems as well as disease pathology. We can also utilize the `remMAP` idea to other models. For example, when selecting a group of variables in a multiple regression model, we can impose both the ℓ_2 penalty (that is, the group lasso penalty), as well as an ℓ_1 penalty to encourage within group sparsity. Similarly, the `remMAP` idea can also be applied to vector autoregressive models and generalize linear models.

R package `remMap` will be available through CRAN shortly.

References

Albertson, D. G., C. Collins, F. McCormick, and J. W. Gray, (2003), "Chromosome aberrations in solid tumors," *Nature Genetics*, 34.

Antoniadis, A., and Fan, J., (2001), "Regularization of wavelet approximations," *Journal of the American Statistical Association*, 96, 939–967.

Bai T, Luoh SW., (2008) "GRB-7 facilitates HER-2/Neu-mediated signal transduction and tumor formation," *Carcinogenesis*, 29(3), 473-9.

Bakin, S., (1999), "Adaptive regression and model selection in data mining prob-

lems," *PhD Thesis* , Australian National University, Canberra.

Bedrick, E. and Tsai, C.,(1994), "Model selection for multivariate regression in small samples," *Biometrics*, 50, 226C231.

Bergamaschi, A., Kim, Y. H., Wang, P., Sorlie, T., Hernandez-Boussard, T., Lonning, P. E., Tibshirani, R., Borresen-Dale, A. L., and Pollack, J. R., (2006), "Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer," *Genes Chromosomes Cancer*, 45, 1033-1040.

Bergamaschi, A., Kim, Y.H., Kwei, K.A., Choi, Y.L., Bocanegra, M., Langerod, A., Han, W., Noh, D.Y., Huntsman, D.G., Jeffrey, S.S., Borresen-Dale, A. L., and Pollack, J.R., (2008), "CAMK1D amplification implicated in epithelial-mesenchymal transition in basal-like breast cancer," *Mol Oncol*, In Press.

Breiman, L. and Friedman, J. H., (1997), "Predicting multivariate responses in multiple linear regression (with discussion)," *J. R. Statist. Soc. B*, 59, 3-54.

Brown, P., Fearn, T. and Vannucci, M., (1999), "The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach," *Biometrika*, 86, 635C648.

Brown, P., Vannucci, M. and Fearn, T., (1998), "Multivariate Bayesian variable selection and prediction," *J. R. Statist. Soc. B*, 60, 627C641.

Brown, P., Vannucci, M. and Fearn, T.,(2002), "Bayes model averaging with selection of regressors," *J. R. Statist. Soc. B*, 64, 519C536.

Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, et al., (2004), "Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds," *PLoS Biol*, 2(2).

Dowdy SC, Gostout BS, Shridhar V, Wu X, Smith DI, Podratz KC, Jiang

SW., (2005) "Biallelic methylation and silencing of paternally expressed gene 3(PEG3) in gynecologic cancer cell lines," *Gynecol Oncol*, 99(1), 126-34.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *Annals of Statistics*, 32, 407–499.

Frank, I. and Friedman, J.,(1993), "A statistical view of some chemometrics regression tools (with discussion)," *Technometrics*, 35, 109C148.

Fu, W., (1998), "Penalized regressions: the bridge vs the lasso," *JCGS*, 7.

Friedman, J., Hastie, T. and Tibshirani, R.,(2008), "Regularized Paths for Generalized Linear Models via Coordinate Descent," *Techniqual Report*, Department of Statistics, Stanford University.

Friedman, J., Hastie, T. and Tibshirani, R.,(2007), "Pathwise coordinate optimization," *The Annals of Applied Statistics.*, 1(2), 302-332 .

Fujikoshi,Y. and Satoh, K., (1997), "Modified AIC and Cp in multivariate linear regression," *Biometrika*, 84, 707C716.

Gardner, T. S., D. DI Bernardo, D. Lorenz, and J. J. Collins, (2003) "Inferring genetic networks and identifying compound mode of action via expression profiling," *Science*, 301

Hyman, E., P. Kauraniemi, S. Hautaniemi, M. Wolf, S. Mousses, E. Rozenblum, M. Ringner, G. Sauter, O. Monni, A. Elkahloun, O.-P. Kallioniemi, and A. Kallioniemi, (2002), "Impact of dna amplification on gene expression patterns in breast cancer," *Cancer Res*, 62.

Izenman, A., (1975), "Reduced-rank regression for the multivariate linear model," *J. Multiv. Anal.*, 5, 248C264.

Jeong, H., S. P. Mason, A. L. Barabasi, and Z. N. Oltvai, (2001), "Lethality and centrality in protein networks," *Nature*, (411)

Johnson MD, Wu X, Aithmitti N, Morrison RS., (2002) "Peg3/Pw1 is a mediator between p53 and Bax in DNA damage-induced neuronal death," *J Biol Chem*, 277(25), 23000-7.

Kapp, A. V., Jeffrey, S. S., Langerod, A., Borresen-Dale, A. L., Han, W., Noh, D. Y., Bukholm, I. R., Nicolau, M., Brown, P. O. and Tibshirani, R., (2006), "Discovery and validation of breast cancer subtypes," *BMC Genomics*, 7, 231.

Kao J, Pollack JR., (2006), "RNA interference-based functional dissection of the 17q12 amplicon in breast cancer reveals contribution of coamplified genes," *Genes Chromosomes Cancer*, 45(8), 761-9.

Koh WP, Yuan JM, Van Den Berg D, Lee HP, Yu MC., (2005) "Polymorphisms in angiotensin II type 1 receptor and angiotensin I-converting enzyme genes and breast cancer risk among Chinese women in Singapore," *Carcinogenesis*, 26(2), 459-64.

Kremmidiotis G, Baker E, Crawford J, Eyre HJ, Nahmias J, Callen DF., (1998), "Localization of human cadherin genes to chromosome regions exhibiting cancer-related loss of heterozygosity," *Genomics*, 49(3), 467-71.

Langerod, A., Zhao, H., Borgan, O., Nesland, J. M., Bukholm, I. R., Ikdahl, T., Karesen, R., Borresen-Dale, A. L., Jeffrey, S. S., (2007), "TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer," *Breast Cancer Res*, 9, R30.

Lutz, R. and Bühlmann, P., (2006), "Boosting for high-multivariate responses in high-dimensional linear regression," *Statist. Sin.*, 16, 471C494.

Paik S, Shak S, Tang G, Kim C, Baker J, et al., (2004), "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *N Engl J Med*, 351(27), 2817-2826.

Peng, J., P. Wang, N. Zhou, J. Zhu, (2008), "Partial Correlation Estimation by Joint Sparse Regression Models," *JASA*, accepted.

Pollack, J., T. Srlie, C. Perou, C. Rees, S. Jeffrey, P. Lonning, R. Tibshirani, D. Botstein, A. Brresen-Dale, and Brown,P., (2002), "Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors," *Proc Natl Acad Sci*, 99(20).

Reinsel, G. and Velu, R., (1998), "Multivariate Reduced-rank Regression: Theory and Applications," *New York*, Springer.

Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She QB, et al., (2007), "Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity," *Proc Natl Acad Sci U S A*, 104(18), 7564-7569.

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese,J. C., Brown,P. O., Botstein, D., Lnning P. E., and Brresen-Dale, A.L., (2001), "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc Natl Acad Sci U S A*, 98, 10869-10874.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter,J., Perou,C. M., Lnning, P. E., Brown, P. O., Brresen-Dale, A.-L., and Botstein, D., (2003), "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proc Natl Acad Sci U S A*, 100, 8418-8423.

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al.,(2006), "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis," *J Natl Cancer Inst*, 98(4), 262-272.

Tibshirani, R., (1996) "Regression shrinkage and selection via the lasso," *J. R. Statist. Soc. B* , 58, 267C288.

Tibshirani, R. and Wang, P., (2008) "Spatial smoothing and hot spot detection for cgh data using the fused lasso," *Biostatistics* , 9(1), 18-29.

Turlach, B., Venables, W. and Wright, S.,(2005), "Simultaneous variable selection," *Technometrics*, 47, 349C363.

Wang, P., (2004) "Statistical methods for CGH array analysis," *Ph.D. Thesis, Stanford University*, 80-81.

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al.,(2005), "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, 365(9460), 671-679.

van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R, (2002) "A gene-expression signature as a predictor of survival in breast cancer," *N Engl J Med*, 347(25), 1999-2009.

Yuan, M., Ekici, A., Lu, Z., and Monterio, R., (2007) "Dimension reduction and coefficient estimation in multivariate linear regression," *J. R. Statist. Soc. B*, 69(3), 329C346.

Yuan, M. and Lin, Y., (2006) "Model Selection and Estimation in Regression with Grouped Variables," *Journal of the Royal Statistical Society, Series B,*, 68(1), 49-67.

Zhao, H., Langerod, A., Ji, Y., Nowels, K. W., Nesland, J. M., Tibshirani, R., Bukholm, I. K., Karesen, R., Botstein, D., Borresen-Dale, A. L., and Jeffrey, S. S., (2004), "Different gene expression patterns in invasive lobular and ductal

carcinomas of the breast," *Mol Biol Cell*, 15, 2523-2536.

Zhao, P., Rocha, G., and Yu, B., (2006), "Grouped and hierarchical model selection through composite absolute penalties," *Annals of Statistics*. Accepted.

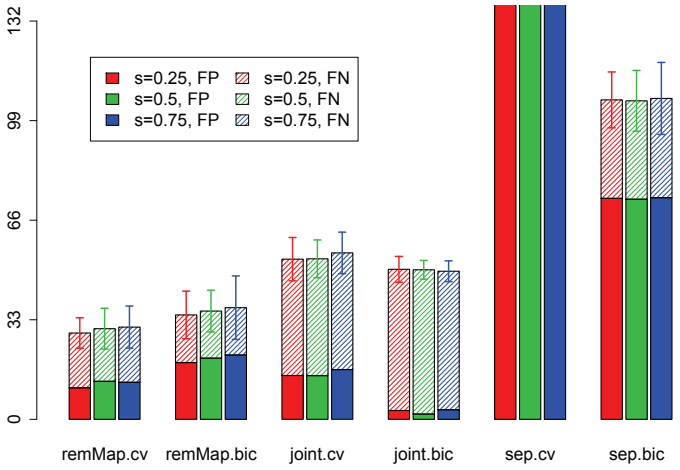
Zou, H., Trevor, H. and Tibshirani, R., (2007), "On degrees of freedom of the lasso," *Annals of Statistics*, 35(5), 2173-2192.

Table 4: RNAs[1] being influenced by the amplifications of the three CNAIs in Table 3.

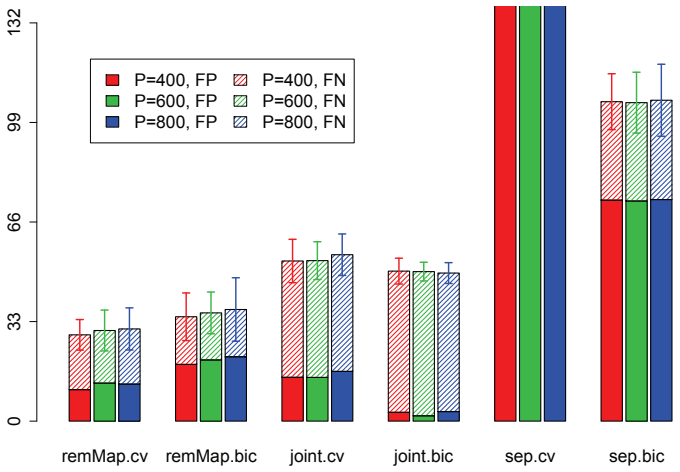
Clone ID	Gene symbol	Cytoband	Correlation
753692	ABLIM1	10q25	0.199
896962	ACADS	12q22-qter	-0.22
753400	ACTL6A	3q26.33	0.155
472185	ADAMTS1	21q21.2	0.214
210687	AGTR1	3q21-q25	-0.182
856519	ALDH3A2	17p11.2	-0.244
270535	BM466581	19	0.03
238907	CABC1	1q42.13	-0.174
773301	CDH3	16q22.1	0.118
505576	CORIN	4p13-p12	0.196
223350	CP	3q23-q25	0.184
810463	DHRS7B	17p12	-0.151
50582	FLJ25076	5p15.31	0.086
669443	HSF2	6q22.31	0.207
743220	JMJD4	1q42.13	-0.19
43977	KIAA0182	16q24.1	0.259
810891	LAMA5	20q13.2-q13.3	0.269
247230	MARVELD2	5q13.2	-0.214
812088	NLN	5q12.3	0.093
257197	NRBF2	10q21.2	0.275
782449	PCBP2	12q13.12-q13.13	-0.079
796398	PEG3	19q13.4	0.169
293950	PIP5K1A	1q22-q24	-0.242
128302	PTMS	12p13	-0.248
146123	PTPRK	6q22.2-q22.3	0.218
811066	RNF41	12q13.2	-0.247
773344	SLC16A2	Xq13.2	0.24
1031045	SLC4A3	2q36	0.179
141972	STT3A	11q23.3	0.182
454083	TMPO	12q22	0.175
825451	USO1	4q21.1	0.204
68400	BM455010	17	0.748
756253,365147	ERBB2	17q11.2-q12—17q21.1	0.589
510318,236059	GRB7	17q12	0.675
245198	MED24	17q21.1	0.367
825577	STARD3	17q11-q12	0.664
782756[2]	TBPL1	6q22.1-q22.3	0.658

1. The first part of the table lists trans-regulated genes. The second part of the table lists cis-regulated genes.

2. This cDNA sequence probe is annotated with `TBPL1`, but actually maps to one of the 17q21.2 genes.

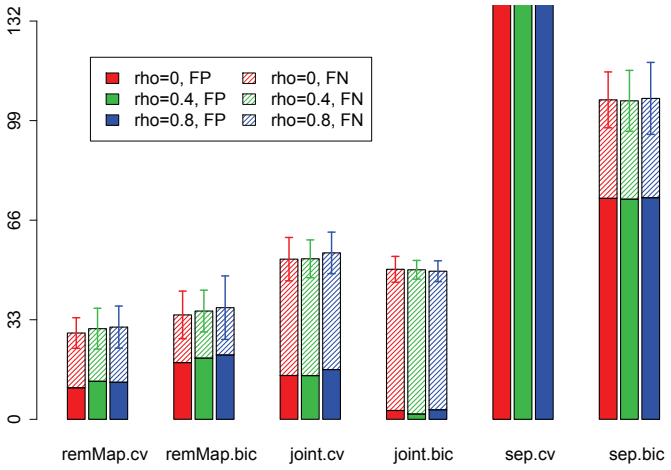


(a) Impact of signal size. $P = Q = 600$, $N = 200$; $\rho_x = 0$; $\rho_\varepsilon = 0$; the total number of `trans-edges` is 132.

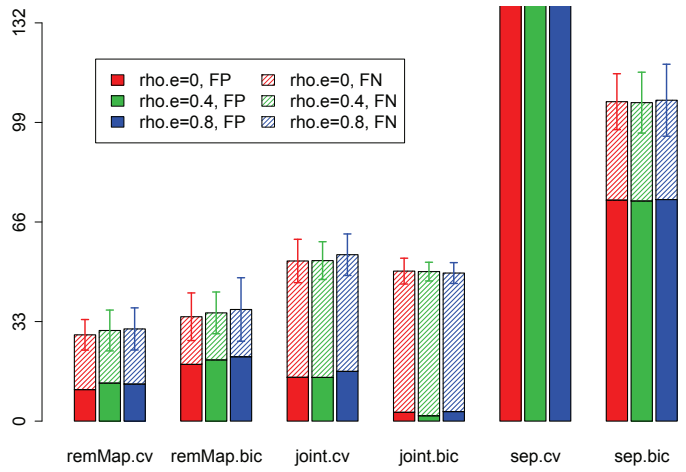


(b) Impact of predictor and response dimensionality. $Q = P$, $N = 200$; $s = 0.25$; $\rho_x = 0$; $\rho_\varepsilon = 0$; the total number of `trans-edges` is 132.

Figure 1: Impact of signal size and dimensionality. Heights of solid bars represent numbers of false positive detections of `trans-edges` (FP); heights of shaded bars represent numbers of false negative detections of `trans-edges` (FN). All bars are truncated at height=132.



(a) Impact of predictor correlation. $P = Q = 600$, $N = 200$; $s = 0.25$; $\rho_\varepsilon = 0$; the total number of `trans-edges` is 132.



(b) Impact of residual correlation. $P = Q = 600$, $N = 200$; $s = 0.25$; $\rho_x = 0$; the total number of `trans-edges` is 132.

Figure 2: Impact of correlations. Heights of solid bars represent numbers of false positive detections of `trans-edges` (FP); heights of shaded bars represent numbers of false negative detections of `trans-edges` (FN). All bars are truncated at height=132.

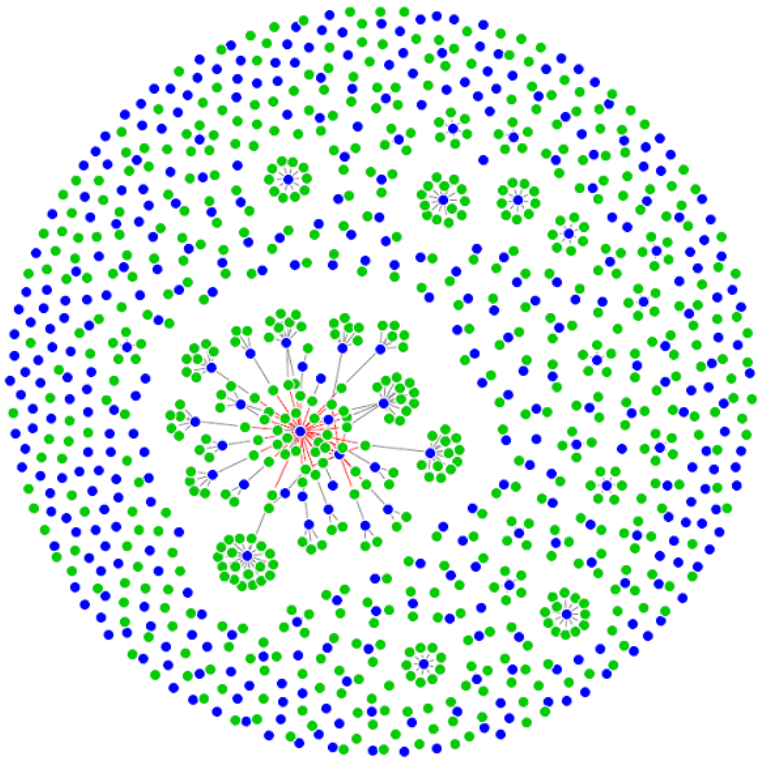


Figure 3: Network of the estimated regulatory relationships between the copy numbers of the 384 CNAIs and the expressions of the 654 breast cancer related genes. Each blue node stands for one CNAI, and each green node stands for one gene. Red edges represent inferred trans-regulations (43 in total). Grey edges represent cis-regulations.

Supplementary Material

Appendix A: Proof of Theorem 1

Define

$$L(\beta; Y, X) = \frac{1}{2}\sum_{j=1}^{q}(y_j - x\beta_j)^2 + \lambda_1 \sum_{j=1}^{q}|\beta_j| + \lambda_2\sqrt{\sum_{j=1}^{q}\beta_j^2}.$$

It is obvious that, in order to prove Theorem 1, we only need to show that, the solution of $\min_\beta L(\beta; Y, X)$, is given by (for $j = 1, \cdots, q$)

$$\widehat{\beta}_j = \begin{cases} 0, & \text{if } ||\widehat{\beta}^{\text{lasso}}||_2 = 0; \\ \widehat{\beta}_j^{\text{lasso}}\left(1 - \frac{\lambda_2}{||\widehat{\beta}^{\text{lasso}}||_2 x^2}\right)_+, & \text{otherwise,} \end{cases}$$

where

$$\widehat{\beta}_j^{\text{lasso}} = \left(1 - \frac{\lambda_1}{|xy_j|}\right)_+ \frac{xy_j}{x^2}. \tag{S-1}$$

In the following, for function L, view $\{\beta_{j'} : j' \neq j\}$ as fixed. With a slight abuse of notation, write $L = L(\beta_j)$. Then when $\beta_j \geq 0$, we have

$$\frac{dL}{d\beta_j} = -xy_j + (x^2 + \frac{\lambda_2}{||\beta||_2})\beta_j + \lambda_1.$$

Thus, $\frac{dL}{d\beta_j} > 0$ if and only if $\beta_j > \tilde{\beta_j}^+$, where

$$\tilde{\beta_j}^+ := \frac{xy_j}{x^2 + \frac{\lambda_2}{||\beta||_2}}(1 - \frac{\lambda_1}{xy_j}).$$

Denote the minima of $L(\beta_j)|_{\beta_j \geq 0}$ by $\beta^+_{j,\min}$. Then, when $\tilde{\beta_j}^+ > 0$, $\beta^+_{j,\min} = \tilde{\beta_j}^+$. On the other hand, when $\tilde{\beta_j}^+ \leq 0$, $\beta^+_{j,\min} = 0$. Note that $\tilde{\beta_j}^+ > 0$ if and only if $xy_j(1 - \frac{\lambda_1}{xy_j}) > 0$. Thus we have

$$\beta^+_{j,\min} = \begin{cases} \tilde{\beta}^+_j, & \text{if } xy_j(1 - \frac{\lambda_1}{xy_j}) > 0; \\ 0, & \text{if } xy_j(1 - \frac{\lambda_1}{xy_j}) \leq 0. \end{cases}$$

Similarly, denote the minima of $L(\beta_j)|_{\beta_j \leq 0}$ by $\beta^-_{j,\min}$, and define

$$\tilde{\beta_j}^- := \frac{xy_j}{x^2 + \frac{\lambda_2}{||\beta||_2}}(1 + \frac{\lambda_1}{xy_j}).$$

Then we have

$$\beta^-_{j,\min} = \begin{cases} \tilde{\beta}^-_j, & \text{if } xy_j(1 + \frac{\lambda_1}{xy_j}) < 0; \\ 0, & \text{if } xy_j(1 + \frac{\lambda_1}{xy_j}) \geq 0. \end{cases}$$

Denote the minima of $L(\beta_j)$ as $\widehat{\beta}_j$ (with a slight abuse of notation). From the above, it is obvious that, if $xy_j > 0$, then $\widehat{\beta}_j \geq 0$. Thus $\widehat{\beta}_j = \max(\tilde{\beta}^+_j, 0) = \frac{xy_j}{x^2 + \frac{\lambda_2}{||\beta||_2}}(1 - \frac{\lambda_1}{xy_j})_+ = \frac{xy_j}{x^2 + \frac{\lambda_2}{||\beta||_2}}(1 - \frac{\lambda_1}{|xy_j|})_+$. Similarly, if $xy_j \leq 0$, then $\widehat{\beta}_j \leq 0$, and it has the same expression as above. Denote the minima of $L(\beta)|_{||\beta||_2 > 0}$ (now viewed as a function of $(\beta_1, \cdots, \beta_q)$) as $\widehat{\beta}_{\min} = (\widehat{\beta}_{1,\min}, \cdots, \widehat{\beta}_{q,\min})$. We have shown above that, if such a minima exists, it satisfies (for $j = 1, \cdots, q$)

$$\widehat{\beta}_{j,\min} = \frac{xy_j}{x^2 + \frac{\lambda_2}{||\widehat{\beta}_{\min}||_2}} \left(1 - \frac{\lambda_1}{|xy_j|}\right)_+ = \widehat{\beta}^{\text{lasso}}_j \frac{x^2}{x^2 + \frac{\lambda_2}{||\widehat{\beta}_{\min}||_2}}, \tag{S-2}$$

where $\widehat{\beta}^{\text{lasso}}_j$ is defined by equation (S-1). Thus

$$||\widehat{\beta}_{\min}||_2 = ||\widehat{\beta}^{\text{lasso}}||_2 \frac{x^2}{x^2 + \frac{\lambda_2}{||\widehat{\beta}_{\min}||_2}}.$$

By solving the above equation, we obtain

$$||\widehat{\beta}_{\min}||_2 = ||\widehat{\beta}^{\text{lasso}}||_2 - \frac{\lambda_2}{x^2}.$$

By plugging the expression on the right hand side into (S-2), we achieve

$$\widehat{\beta}_{j,\min} = \widehat{\beta}_j^{\text{lasso}} \left(1 - \frac{\lambda_2}{||\widehat{\beta}^{\text{lasso}}||_2 x^2}\right).$$

Denote the minima of $L(\beta)$ by $\widehat{\beta} = (\widehat{\beta}_1, \cdots, \widehat{\beta}_q)$. From the above, we also know that if $||\widehat{\beta}^{\text{lasso}}||_2 - \frac{\lambda_2}{x^2} > 0$, $L(\beta)$ achieves its minimum on $||\beta||_2 > 0$, which is $\widehat{\beta} = \widehat{\beta}_{\min}$. Otherwise, $L(\beta)$ achieves its minimum at zero. Since $||\widehat{\beta}^{\text{lasso}}||_2 - \frac{\lambda_2}{x^2} > 0$ if and only if $1 - \frac{\lambda_2}{||\widehat{\beta}^{\text{lasso}}||_2 x^2} > 0$, we have proved the theorem.

Appendix B: Proof of Theorem 2

Before proving Theorem 2, we first explain definition (7) – the degrees of freedom. Consider the q^{th} regression in model (1). Suppose that $\{\widehat{y}_q^n\}_{n=1}^N$ are the fitted values by a certain fitting procedure based on the current observations $\{y_q^n : n = 1, \cdots, N; q = 1, \cdots, Q\}$. Let $\mu_q^n := \sum_{p=1}^P x_p^n \beta_{pq}$. Then for a fixed design matrix $X = (x_p^n)$, the expected re-scaled prediction error of $\{\widehat{y}_q^n\}_{n=1}^N$ in predicting a future set of new observations $\{\widetilde{y}_q^n\}_{n=1}^N$ from the q^{th} regression of model (1) is:

$$\text{PE}_q = \sum_{n=1}^N E((\widehat{y}_q^n - \widetilde{y}_q^n)^2)/\epsilon_{q,\epsilon}^2 = \sum_{n=1}^N E((\widehat{y}_q^n - \mu_q^n)^2)/\epsilon_{q,\epsilon}^2 + N.$$

Note that

$$(\widehat{y}_q^n - y_q^n)^2 = (\widehat{y}_q^n - \mu_q^n)^2 + (y_q^n - \mu_q^n)^2 - 2(\widehat{y}_q^n - \mu_q^n)(y_q^n - \mu_q^n).$$

Therefore,

$$\mathrm{PE}_q = \sum_{n=1}^{N} E((\widehat{y}_q^n - y_q^n)^2)/\epsilon_{q,\epsilon}^2 + 2\sum_{n=1}^{N} \mathrm{Cov}(\widehat{y}_q^n, y_q^n)/\epsilon_{q,\epsilon}^2.$$

Denote $RSS_q = \sum_{n=1}^{N}(\widehat{y}_q^n - y_q^n)^2$. Then an un-biased estimator of PE_q is

$$RSS_q/\epsilon_{q,\epsilon}^2 + 2\sum_{n=1}^{N} \mathrm{Cov}(\widehat{y}_q^n, y_q^n)/\epsilon_{q,\epsilon}^2.$$

Therefore, a natural definition of the degrees of freedom for the procedure resulting the fitted values $\{\widehat{y}_q^n\}_{n=1}^N$ is as given in equation (7). Note that, this is the definition used in Mallow's C_p criterion.

Proof of Theorem 2: By applying Stein's identity to the Normal distribution, we have: if $Z \sim N(\mu, \sigma^2)$, and a function g such that $E(|g'(Z)|) < \infty$, then

$$\mathrm{Cov}(g(Z), Z)/\sigma^2 = E(g'(Z)).$$

Therefore, under the normality assumption on the residuals $\{\epsilon_q\}_{q=1}^Q$ in model (1), definition (7) becomes

$$df_q = \sum_{n=1}^{N} E\left(\frac{\partial \widehat{y}_q^n}{\partial y_q^n}\right), \quad q = 1, \cdots, Q.$$

Thus an obvious unbiased estimator of df_q is $\widehat{df}_q = \sum_{n=1}^{N} \frac{\partial \widehat{y}_q^n}{\partial y_q^n}$. In the following, we derive $\widehat{df}_q$ for the proposed `remMap` estimator under the orthogonal design. Let $\widehat{\beta}_q = (\widehat{\beta}_{1q}, \cdots, \widehat{\beta}_{Pq})$ be a one by P row vector; let $\mathbf{X} = (x_p^n)$ be the N by P design matrix which is orthonormal; let $Y_q = (y_1^1, \cdots, y_q^N)^T$ and $\widehat{Y}_q = (\widehat{y}_q^1, \cdots, \widehat{y}_q^N)^T = \mathbf{X}\widehat{\beta}_q$ be N by one column vectors. Then

$$\widehat{df}_q = \mathrm{tr}\left(\frac{\partial \widehat{Y}_q}{\partial Y_q}\right) = \mathrm{tr}\left(\mathbf{X}\frac{\partial \widehat{\beta}_q}{\partial Y_q}\right) = \mathrm{tr}\left(\mathbf{X}\frac{\partial \widehat{\beta}_q}{\partial \widehat{\beta}_{q,\mathrm{ols}}}\frac{\partial \widehat{\beta}_{q,\mathrm{ols}}}{\partial Y_q}\right),$$

where $\widehat{\beta}_{q,\text{ols}} = (\widehat{\beta}^{\text{ols}}_{1q}, \cdots, \widehat{\beta}^{\text{ols}}_{Pq})^T$ and the last equality is due to the chain rule. Since under the orthogonal design, $\widehat{\beta}^{\text{ols}}_{pq} = X_p^T Y_q / ||X_p||_2^2$, where $X_p = (x_p^1, \cdots, x_p^N)^T$, thus $\frac{\partial \widehat{\beta}_{q,\text{ols}}}{\partial Y_q} = \mathbf{D}\mathbf{X}^T$, where $\mathbf{D}$ is a P by P diagonal matrix with the p^{th} diagonal entry being $1/||X_p||_2^2$. Therefore

$$\widehat{df}_q = \text{tr}\left(\mathbf{X}\frac{\partial \widehat{\beta}_q}{\partial \widehat{\beta}_{q,\text{ols}}}\mathbf{D}\mathbf{X}^T\right) = \text{tr}\left(\mathbf{D}\mathbf{X}^T\mathbf{X}\frac{\partial \widehat{\beta}_q}{\partial \widehat{\beta}_{q,\text{ols}}}\right) = \text{tr}\left(\frac{\partial \widehat{\beta}_q}{\partial \widehat{\beta}_{q,\text{ols}}}\right) = \sum_{p=1}^{P}\frac{\partial \widehat{\beta}_{pq}}{\partial \widehat{\beta}^{\text{ols}}_{pq}},$$

where the second to last equality is by $\mathbf{X}^T\mathbf{X} = \mathbf{D}^{-1}$ which is due to the orthogonality of $\mathbf{X}$. By the chain rule

$$\frac{\partial \widehat{\beta}_{pq}}{\partial \widehat{\beta}^{\text{ols}}_{pq}} = \frac{\partial \widehat{\beta}_{pq}}{\partial \widehat{\beta}^{\text{lasso}}_{pq}}\frac{\partial \widehat{\beta}^{\text{lasso}}_{pq}}{\partial \widehat{\beta}^{\text{ols}}_{pq}}.$$

By Theorem 1, under the orthogonal design,

$$\frac{\partial \widehat{\beta}_{pq}}{\partial \widehat{\beta}^{\text{lasso}}_{pq}} = \mathbb{I}\left(||\widehat{B}_p^{\text{lasso}}||_{2,C} > \frac{\lambda_2}{||X_p||_2^2}\right) \times \left[1 - \frac{\lambda_2}{||X_p||_2^2}\frac{||\widehat{B}_p^{\text{lasso}}||_{2,C}^2 - (\widehat{\beta}^{\text{lasso}}_{pq})^2}{||\widehat{B}_p^{\text{lasso}}||_{2,C}^3}\right],$$

and

$$\frac{\partial \widehat{\beta}^{\text{lasso}}_{pq}}{\partial \widehat{\beta}^{\text{ols}}_{pq}} = \begin{cases} 1, & \text{if } c_{p,q} = 0; \\ \mathbb{I}\left(|\widehat{\beta}^{\text{ols}}_{pq}| > \frac{\lambda_1}{||X_p||_2^2}\right), & \text{if } c_{p,q} = 1. \end{cases}$$

Note that when $c_{p,q} = 0$, $\widehat{\beta}_{pq} = \widehat{\beta}^{\text{ols}}_{pq}$, thus $\frac{\partial \widehat{\beta}_{pq}}{\partial \widehat{\beta}^{\text{ols}}_{pq}} = 1$. It is then easy to show that $\widehat{df}_q$ is as given in equation (9).

Appendix C: Data Preprocessing

C.1 Preprocessing for array CGH data

Each array output (log_2 ratios) is first standardized to have median= 0 and smoothed by `cghFLasso` (Tibshirani and Wang 2008) for defining gained/lost regions on the genome. The noise level of each array is then calculated based on the measurements

from the estimated normal regions (i.e., regions with estimated copy numbers equal to 2). After that, each smoothed array is normalized according to its own noise level.

We define *copy number alteration intervals (CNAIs)* by using the Fixed-Order Clustering (FOC) method (Wang 2004), which first builds a hierarchical clustering tree along the genome based on all arrays, and then cuts the tree at an appropriate height such that genes with similar copy numbers fall into the same CNAI. FOC is a generalization of the `CLAC` (CLuster Along Chromosome) method proposed by Wang et al. (2005). It differs in two ways from the standard agglomerative clustering. First, the order of the leaves in the tree is fixed, which represents the genome order of the genes/clones in the array. So, only adjacent clusters are joined together when the tree is generated by a bottom-up approach. Second, the similarity between two clusters no longer refers to the spatial distance but to the similarity of the array measurements (log_2 ratio) between the two clusters. By using FOC, the human genome is divided into 384 non-overlapping CNAIs based on all 172 CGH arrays. This is illustrated in Figure S-1. In addition, the heatmap of the (sample) correlations of the CNAIs is given in Figure S-2.

C.2 Selection of breast cancer related genes

We combine seven published breast cancer gene lists: the intrinsic gene set (Sorlie et al. 2003), the Amsterdam 70 gene (van de Vijver et al. 2002), the wound response (Chang et al. 2004), the 76-gene metastasis signature (Wang et al. 2005), the recurrence score (Paik et al. 2004), the Genomic Grade Index (GGI) (Sotiriou et al. 2006), and the PTEN signature (Saal et al. 2007). There are 967 genes in the current expression data set overlapping with the above combined breast cancer gene set. We further filter away genes with missing measurements in more than 20% of the samples, and 654 genes are left. Among these 654 selected genes, 449 are from the

intrinsic gene set (Sorlie et al. 2003), which are used to derive breast cancer subtype labels in Appendix C.4.

C.3 Interactions among RNA expressions

We apply the `space` (Sparse PArtial Correlation Estimation) method (Peng et al. 2008) to infer the interactions among RNA levels through identifying non-zero partial correlations. `space` assumes the overall sparsity of the partial correlation matrix and employs sparse regression techniques for model fitting. As indicated by many experiments that genetic-regulatory networks have a power-law type degree distribution with a power parameter in between 2 and 3 (Newman 2003), the tuning parameter in `space` is chosen such that the resulting network has an estimated power parameter around 2 (see Figure S-3(b) for the corresponding degree distribution). The resulting (concentration) network has 664 edges in total, whose topology is illustrated in Figure S-3(a). In this network, there are 7 nodes having at least 10 edges. These hub genes include PLK1, PTTG1, AURKA, ESR1, and GATA3. PLK1 has important functions in maintaining genome stability via its role in mitosis. Its over expression is associated with preinvasive in situ carcinomas of the breast (Rizki et al. 2007). PTTG1 is observed to be a proliferation marker in invasive ductal breast carcinomas (Talvinen et al. 2008). AURKA encodes a cell cycle-regulated kinase and is a potential metastasis promoting gene for breast cancer (Thomassen et al. 2008). ESR1 encodes an estrogen receptor, and is a well known key player in breast cancer. Moreover, it had been reported that GATA3 expression has a strong association with estrogen receptor in breast cancer (Voduc et al. 2008). Detailed annotation of these and other hub genes are listed in Table S-1. We refer this network as *Exp.Net.664*, which is used in our analysis to account for RNA interactions when investigating the regulations between CNAIs and RNA levels.

Table S-1: Annotations for hub genes (degrees greater than 10) in the inferred RNA interaction network *Exp.Net.664*.

CloneID	Gene Name	Symbol	ID	Cytoband
744047	Polo-like kinase 1 (Drosophila)	PLK1	5347	16p12.1
781089	Pituitary tumor-transforming 1	PTTG1	9232	5q35.1
129865	Aurora kinase A	AURKA	6790	20q13.2-q13.3
214068	GATA binding protein 3	GATA3	2625	10p15
950690	Cyclin A2	CCNA2	890	4q25-q31
120881	RAB31, member RAS oncogene family	RAB31	11031	18p11.3
725321	Estrogen receptor 1	ESR1	2099	6q25.1

C.4 Breast Cancer Subtypes

Population stratification due to distinct subtypes could confound our detection of associations between CNAIs and gene expressions. For example, if the copy number of CNAI A and expression level of gene B are both higher in one subtype than in the other subtypes, we could observe a strong correlation between CNAI A and gene expression B across the whole population, even when the correlation within each subtype is rather weak. To account for this potential confounding factor, we introduce a set of subtype indicator variables, which is used as additional predictors in the `remMap` model. Specifically, we derive subtype labels based on expression patterns by following the work of Sorlie et al. (2003). We first normalize the expression levels of each intrinsic gene (449 in total) across the 172 samples to have mean zero and MAD one. Then we use `kmeans` clustering to divide the patients into five distinct groups, which correspond to the five subtypes suggested by Sorlie et al. (2003) — Luminal Subtype A, Luminal Subtype B, ERBB2-overexpressing Subtype, Basal Subtype and Normal Breast-like Subtype. Figure S-4 illustrates the expression patterns of these five subtypes across the 172 samples. We then define five dummy variables to represent the subtype information for each tumor sample, which is used in the `remMap` model when investigating the interactions between CNAIs and RNA transcript levels.

References

Chang HY, Sneddon JB, Alizadeh AA, Sood R, West RB, et al., (2004), "Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds," *PLoS Biol*, 2(2).

Newman M, (2003), "The Structure and Function of Complex Networks," *Society for Industrial and Applied Mathematics*, 45(2), 167-256.

Paik S, Shak S, Tang G, Kim C, Baker J, et al., (2004), "A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer," *N Engl J Med*, 351(27), 2817-2826.

Peng, J., P. Wang, N. Zhou, J. Zhu, (2008), "Partial Correlation Estimation by Joint Sparse Regression Models," *JASA*, accepted.

Rizki A, Mott JD, Bissell MJ, (2007) "Polo-like kinase 1 is involved in invasion through extracellular matrix," *Cancer Res*, 67(23), 11106-10.

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter,J., Perou,C. M., Lnning, P. E., Brown, P. O., Brresen-Dale, A.-L., and Botstein, D., (2003), "Repeated observation of breast tumor subtypes in independent gene expression data sets," *Proc Natl Acad Sci U S A*, 100, 8418-8423.

Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al.,(2006), "Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis," *J Natl Cancer Inst*, 98(4), 262-272.

Saal LH, Johansson P, Holm K, Gruvberger-Saal SK, She QB, et al., (2007), "Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity," *Proc Natl Acad Sci U S A*, 104(18), 7564-7569.

Talvinen K, Tuikkala J, Nevalainen O, Rantanen A, Hirsimäki P, Sundström J, Kronqvist P, (2008), "Proliferation marker securin identifies favourable outcome in invasive ductal breast cancer," *Br J Cancer*, 99(2), 335-40.

Thomassen M, Tan Q, Kruse TA, (2008) "Gene expression meta-analysis identifies chromosomal regions and candidate genes involved in breast cancer metastasis," *Breast Cancer Res Treat*, Feb 22, Epub.

Tibshirani, R. and Wang, P., (2008) "Spatial smoothing and hot spot detection for cgh data using the fused lasso," *Biostatistics* , 9(1), 18-29.

van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R, (2002) "A gene-expression signature as a predictor of survival in breast cancer," *N Engl J Med*, 347(25), 1999-2009.

Voduc D, Cheang M, Nielsen T, (2008), "GATA-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value," *Cancer Epidemiol Biomarkers Prev*, 17(2), 365-73.

Wang, P., (2004) "Statistical methods for CGH array analysis," *Ph.D. Thesis, Stanford University*, 80-81.

Wang P, Kim Y, Pollack J, Narasimhan B, Tibshirani R, (2005) "A method for calling gains and losses in array CGH data," *Biostatistics*, 6(1), 45-58.

Wang Y, Klijn JG, Zhang Y, Sieuwerts AM, Look MP, et al.,(2005), "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, 365(9460), 671-679.

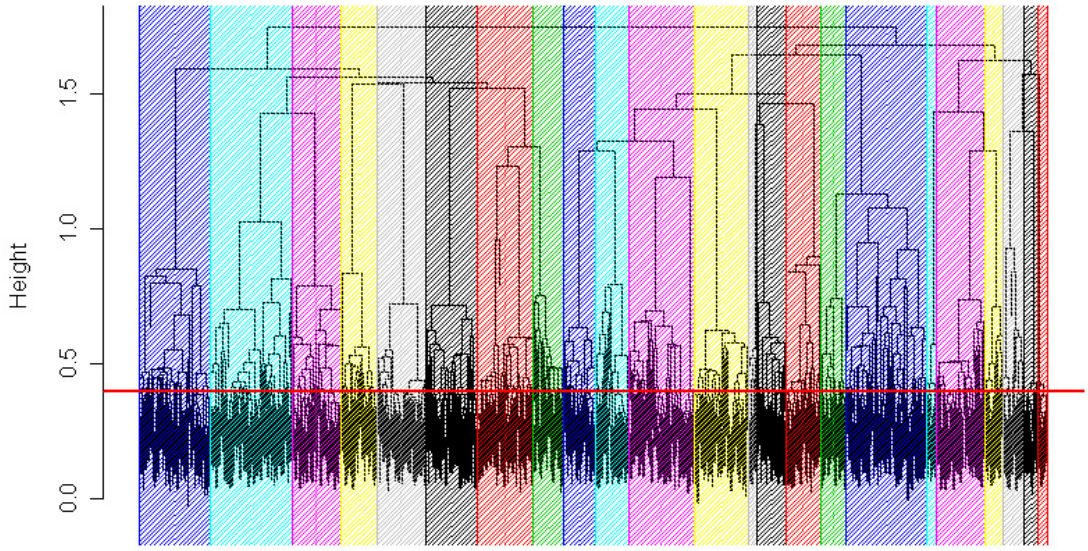


Figure S-1: Hierarchical tree constructed by FOC. Each leaf represents one gene/clone on the array. The order of the leaves represents the order of genes on the genome. The 23 Chromosomes are illustrated with different colors. Cutting the tree at 0.04 (horizonal red line) separates the genome into 384 intervals. This cutoff point is chosen such that no interval contains genes from different chromosomes.

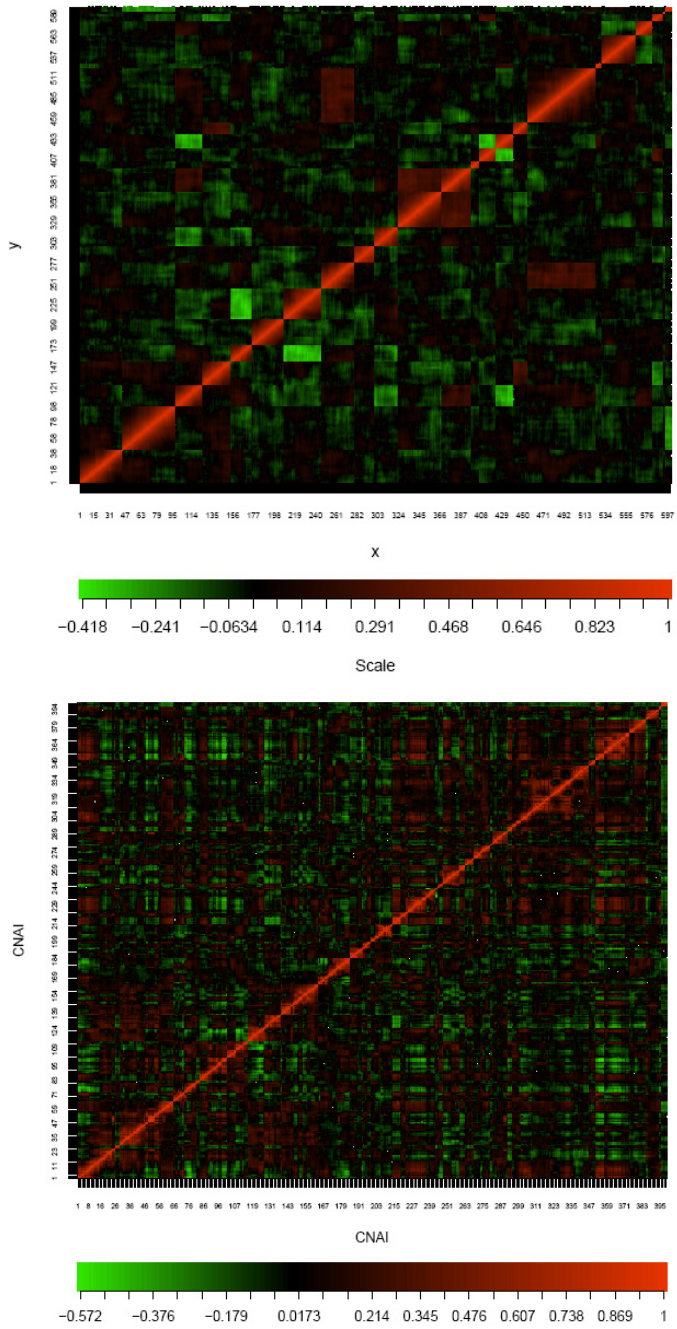
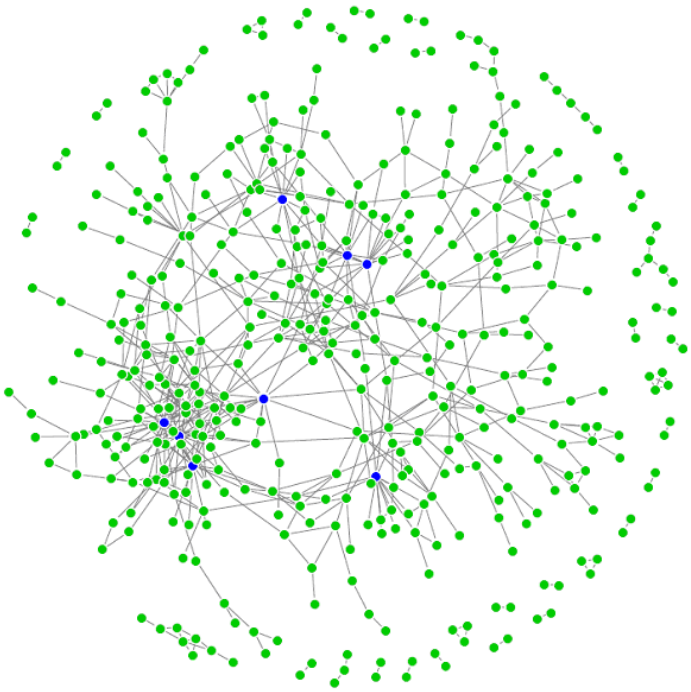
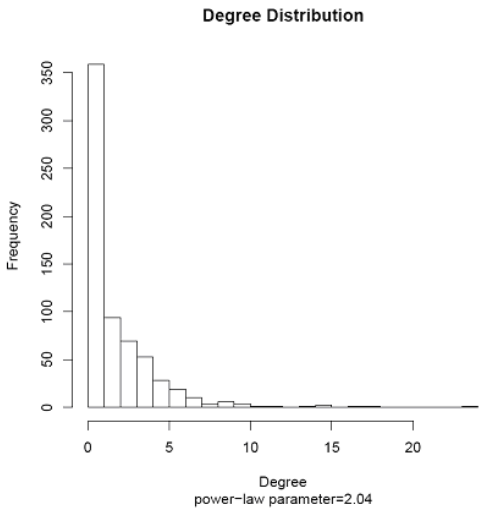


Figure S-2: Heatmaps of the sample correlations among predictors. Top panel: simulated data; Bottom panel: real data



(a) *Exp.Net.664*: Inferred network for the 654 breast cancer related genes (based on their expression levels) by `space`. Nodes with degrees greater than ten are drawn in blue.



(b) Degree distribution of network *Exp.Net.664*.

Figure S-3: RNA interaction network.

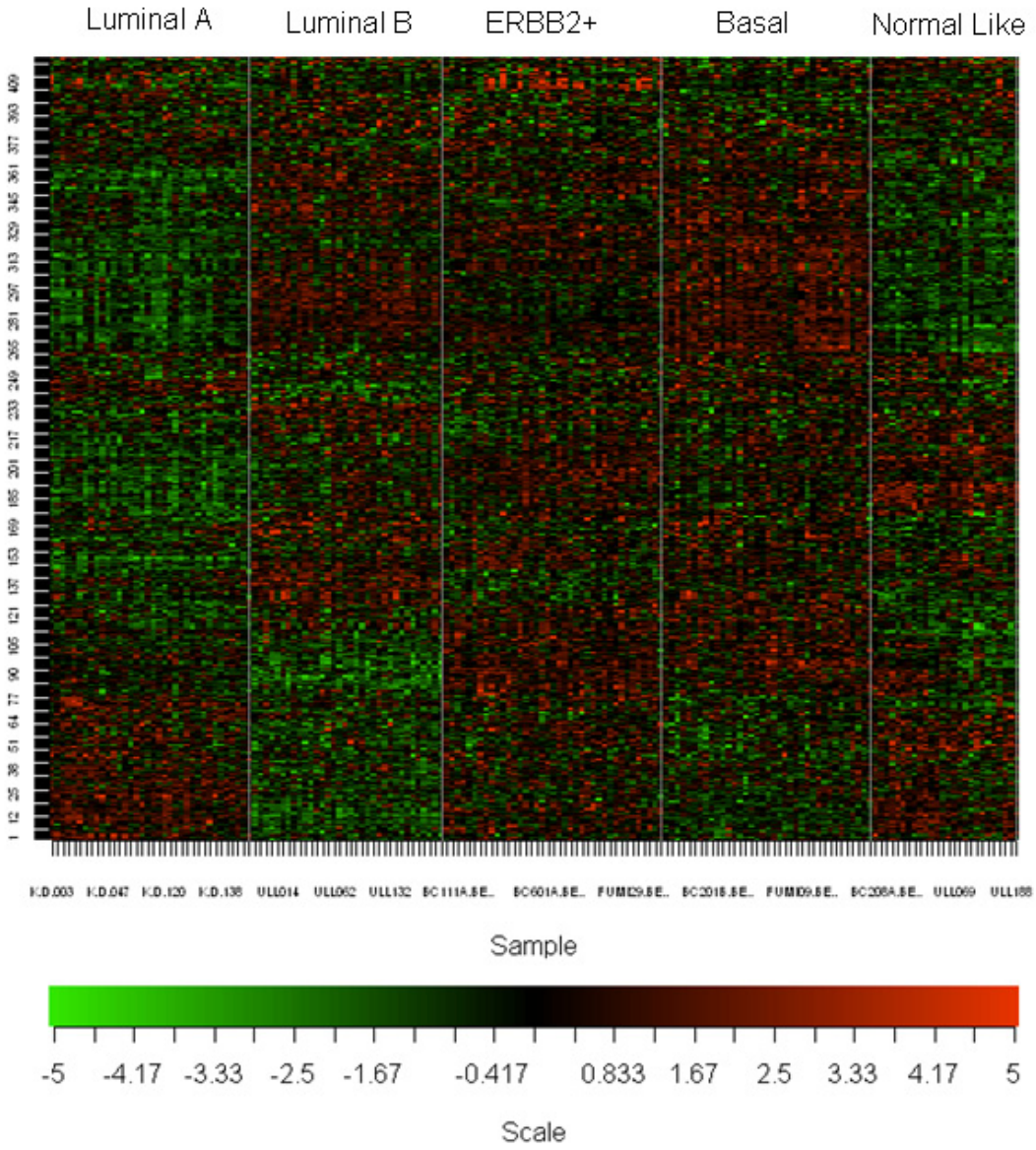


Figure S-4: From left to right, the 5 subtypes are: Luminal Subtype A, Luminal Subtype B, ERBB2-overexpressing subtype, Basal Subtype and Normal Breast-like Subtype.