Published in final edited form as:

Genet Epidemiol. 2014 February ; 38(2): 144–151. doi:10.1002/gepi.21781.

Integrative Analysis of High-throughput Cancer Studies with Contrasted Penalization

Xingjie Shi^{1,2}, Jin Liu³, Jian Huang⁴, Yong Zhou², BenChang Shia⁵, and Shuangge Ma^{1,6,*}

¹Department of Biostatistics, School of Public Health, Yale University, West Haven, CT

²School of Statistics and Management, Shanghai University of Finance and Economics, West Haven, CT

³Division of Epidemiology and Biostatistics, UIC School of Public Health, West Haven, CT

⁴Departments of Statistics & Actuarial Science, and Biostatistics, University of Iowa, West Haven, CT

⁵Department of Statistics and Information Science, FuJen Catholic University, West Haven, CT

⁶VA Cooperative Studies Program Coordinating Center, West Haven, CT

Abstract

In cancer studies with high-throughput genetic and genomic measurements, integrative analysis provides a way to effectively pool and analyze heterogeneous raw data from multiple independent studies and outperforms “classic” meta-analysis and single-dataset analysis. When marker selection is of interest, the genetic basis of multiple datasets can be described using the homogeneity model or the heterogeneity model. In this study, we consider marker selection under the heterogeneity model, which includes the homogeneity model as a special case and can be more flexible. Penalization methods have been developed in the literature for marker selection. This study advances from the published ones by introducing the contrast penalties, which can accommodate the within- and across-dataset structures of covariates/regression coefficients and, by doing so, further improve marker selection performance. Specifically, we develop a penalization method that accommodates the across-dataset structures by smoothing over regression coefficients. An effective iterative algorithm, which calls an inner coordinate descent iteration, is developed. Simulation shows that the proposed method outperforms the benchmark with more accurate marker identification. The analysis of breast cancer and lung cancer prognosis studies with gene expression measurements shows that the proposed method identifies genes different from those using the benchmark and has better prediction performance.

Keywords

Integrative analysis; Contrasted penalization; Marker selection; High-throughput cancer studies

*Correspondence to: Shuangge Ma, 60 College ST, LEPH 206, New Haven, CT 06520 shuangge.ma@yale.edu.

Introduction

Profiling studies have been extensively conducted in cancer research, searching for genetic and genomic markers associated with clinical outcomes and phenotypes. Results generated from the analysis of a single dataset are often unsatisfactory [Guerra and Goldstein 2009; Huang et al. 2012]. Multiple factors may have contributed to the unsatisfactory performance, among which an important or perhaps the most important one is the small sample sizes of individual studies. Fortunately, for common cancer types and outcomes, there are often multiple independent studies with comparable designs. Two examples are provided in this article, and many more are available in the literature. Multi-dataset analysis can pool information across datasets and increase sample size. In multi-dataset analysis, classic meta-analysis methods first analyze each dataset separately and then pool summary statistics across datasets. In contrast, integrative analysis methods pool and analyze raw data from multiple studies and can outperform classic meta-analysis and single-dataset analysis methods [Huang et al. 2012; Liu et al. 2013b; Ma et al. 2011].

Consider the integrative analysis of multiple cancer datasets with high-dimensional genetic or genomic measurements. Here we use gene expression data, which are analyzed later in this article, as an example and note that the proposed method and discussions are also applicable to other types of high-dimensional measurements. The goal is to identify cancer outcome-associated genes (also referred to as “markers” hereafter). The genetic basis of multiple datasets, as measured by the sets of identified genes, can be described using the homogeneity model and the heterogeneity model [Liu et al. 2013b]. The homogeneity model postulates that multiple datasets share the same set of markers. That is, if a gene is identified as associated with outcome in one dataset, it is identified in all datasets. In comparison, under the heterogeneity model, a gene can be associated with outcome in some datasets but not others. The heterogeneity model includes the homogeneity model as a special case and can be more flexible. Multiple techniques can be used for marker selection in integrative analysis. We adopt penalization in this study following Liu et al. [2013b] and references therein, and conjecture that the contrasted approach can be extended to other marker selection techniques. Under the homogeneity model, one-level selection is needed and determines which genes are associated with outcomes. Under the heterogeneity model, two-level selection is needed. The first is to determine whether a specific gene is associated with outcome in any dataset at all, and the second is to determine, for an identified gene, in which dataset(s) it is associated with outcome.

The integrative analysis methods in Liu et al. [2013b] and others treat gene effects within the same dataset and/or across multiple datasets as exchangeable. Genes are interconnected. In single-dataset analysis, it has been shown that taking the interconnections of genes into consideration can improve marker identification and estimation [Huang et al. 2011; Liu et al. 2013c]. In integrative analysis, Liu et al. [2013a] describes the interconnections among genes using a network structure and incorporates it in marker selection. Note that here the network describes the relationships among genes within the same datasets, that is, the *within-dataset structure*.

In integrative analysis, the effects of a gene are measured in multiple datasets and represented by multiple regression coefficients. Even though the datasets are independent, it can still be reasonable to expect those regression coefficients to have some similarity. Such similarity is the basis of pooling multiple datasets. Consider for example when multiple datasets measure the same outcome and the same set of genes, as in Ma et al. [2011]. Here we expect similar sets of genes to be identified in multiple datasets. Furthermore, with proper normalization, the regression coefficients of the same gene are expected to be similar. When multiple datasets are on different outcomes or even different cancer types, as in Liu et al. [2013b] and others, it is of interest to identify genes shared by multiple datasets. For such genes, their regression coefficients not necessarily have similar magnitudes across datasets. But their signs can be “similar”. Under the above scenarios, we are interested in the relationship among the regression coefficients of the same gene in multiple datasets, that is, the *across-dataset structure*. Such a structure has not been investigated in the existing integrative analyses.

In this article, we conduct integrative analysis and penalized marker selection. We suggest that the within-dataset and across-dataset structures can be accommodated using contrasted penalization. Accommodating the within-dataset structure has been studied in Liu et al. [2013a] and will be briefly described here because of its connection with and distinction from the across-dataset structure. The main advancement is the introduction of a contrasted penalization method that accommodates the across-dataset structure. Although the proposed penalty can be used along with several existing penalties, to make the presentation concrete, we couple it with the group bridge [Huang et al. 2009]. As a data example, we analyze right censored data under the AFT (accelerated failure time) model. The proposed contrasted penalization is based on the notion of smoothing, which has been considered in penalization. However, this study is among the first to implement such a technique in integrative analysis. This study will be the first to show that accounting for the across-dataset structure of covariates/regression coefficients can effectively improve marker identification.

Integrative Analysis

Data and model settings

Assume that there are M independent datasets. We use the superscript “ (m) ” to denote the m th dataset. In dataset $m (= 1, \dots, M)$, there are $n^{(m)}$ iid observations. The total sample size is $n = \sum_{m=1}^M n^{(m)}$. Denote $Y^{(m)}$ as the outcome variable, which can be continuous, categorical, or survival. Let $X^{(m)}$ be the length- d vector of gene expressions. For simplicity of notation, in the downstream descriptions, it is assumed that the same set of genes is measured for all subjects in all datasets. In practical data analysis, missingness may occur, and different datasets may have mismatched gene sets. One possibility is to set the regression coefficients corresponding to missing genes as zero. Then the simple rescaling approach in Huang et al. [2012], which “scales up” the coefficients for measured genes, can be adopted. An alternative approach, especially when missingness is not serious, is to conduct imputation (using for example mean or median expressions across genes) and generate a complete set of gene expressions.

In dataset m , assume the model $Y^{(m)} \sim \varphi(X^{(m)'} \beta^{(m)})$, where φ is the known link function, and $\beta^{(m)}$ is the length- d vector of regression coefficients. Denote $L^{(m)}(\beta^{(m)})$ as the loss function. For example, it can be the negative log-likelihood function. In our numerical example, $L^{(m)}(\beta^{(m)})$ is built from an estimating equation. With M independent datasets, the overall loss function is $L(\beta) = \sum_{m=1}^M L^{(m)}(\beta^{(m)})$ where $\beta = (\beta^{(1)'}, \dots, \beta^{(M)'})'$.

Denote $\beta_j^{(m)}$ as the j th component of $\beta^{(m)}$. For gene $j (= 1, \dots, d)$,

$j (= 1, \dots, d)$, $\beta_j = (\beta_j^{(1)}, \dots, \beta_j^{(M)})'$ is the length- M vector of regression coefficients representing its effects across M datasets. Under the homogeneity model,

$I(\beta_j^{(1)}=0) = \dots = I(\beta_j^{(M)}=0)$ for $j = 1, \dots, d$, where $I(\cdot)$ is the indicator function. In

comparison, under the heterogeneity model, it is possible that $I(\beta_j^{(m)}=0) \neq I(\beta_l^{(m)}=0)$ for $j \neq l$.

Penalized marker selection

For marker selection, consider the penalized estimate $\hat{\beta} = \argmin \{L(\beta) + P(\beta)\}$, where $P(\beta)$ is the penalty function. A nonzero component of $\hat{\beta}$ indicates an association between the corresponding gene and outcome. $P(\beta)$ can be built on multiple penalties, such as bridge, SCAD, and MCP [Buhlmann and van de Geer 2011; Zhang 2010]. To make a concrete presentation, as a specific example, we consider $P(\beta)$ constructed using the bridge penalty [Huang et al. 2008].

Under the homogeneity model, selection needs to be conducted at only one level and can be achieved using the 2-norm gBridge (group bridge) where $P(\beta) = \lambda_1 \sum_{j=1}^d c_j \|\beta_j\|_2^\gamma$. Here $\lambda_1 > 0$ is the data-dependent tuning parameter. $c_j \propto M_j^{1-\gamma}$, where M_j is the number of datasets that measure gene j . When all datasets have exactly matched gene sets, $M_j \equiv M$. $\|\beta_j\|_2$ is the ℓ_2 -norm of β_j . $0 < \gamma < 1$ is the fixed bridge parameter. Under the heterogeneity model, selection needs to be conducted at two levels and can be achieved using the 1-norm gBridge where $P(\beta) = \lambda_1 = \sum_{j=1}^d c_j \|\beta_j\|_1^\gamma \cdot \|\beta_j\|_1 = \sum_{m=1}^M |\beta_j^{(m)}|$ is the ℓ_1 -norm of β_j .

The 2-norm (1-norm) gBridge is a composite of the outer bridge and inner ridge (Lasso) penalties. Consider for example the 1-norm gBridge. In our analysis, genes are the functional units. The overall penalty is the sum of d individual penalties, with one for each gene. For a specific gene, the first level of selection is to determine whether it is associated with any outcome at all, which is achieved using the bridge penalty. The second level is to determine in which datasets it is associated with outcome, which is achieved using the Lasso penalty. The composite of the two penalties can achieve the desired two-level selection. Using composite penalization in integrative analysis has been considered in Liu et al. [2013b]. Here the gBridge, which is based on the bridge penalty, is used as opposed to the group MCP in Liu et al. [2013b], which is based on the MCP. We find that when the

contrast penalty is present, gBridge has similar numerical performance as group MCP but lower computational cost.

Contrasted Penalization

The $P(\beta)$ penalty described in the last section treats gene effects as exchangeable. In this section, we introduce the contrast penalties to accommodate the structures of gene effects or equivalently their regression coefficients. The within-dataset structure has been studied in Liu et al. [2013a] and is very briefly described here because of its connection with and distinction from the across-dataset structure. The main advancement is the accommodation of the across-dataset structure.

Accommodating within-dataset structure

Network is constructed in Liu et al. [2013a] to describe the relationships among genes. In network analysis, a node corresponds to a gene, and an adjacency matrix can be computed to model the interconnectedness between any two nodes. Denote $A = (a_{jk}, 1 \leq j, k \leq d)$ as the $d \times d$ adjacency matrix. Consider the penalized estimate $\hat{\beta} = \argmin \{L(\beta) + P(\beta) + P_C(\beta)\}$,

where the contrast penalty $P_C(\beta) = \frac{1}{2} \lambda_2 \sum_{1 \leq j < k \leq d} a_{jk} \left(\frac{\|\beta_j\|_2}{\sqrt{M_j}} - \frac{\|\beta_k\|_2}{\sqrt{M_k}} \right)^2$, and $0 < \lambda_2 < \infty$ is data-dependent tuning parameter. The notion comes from the fact that $P_C(\beta)$ penalizes the

contrast between β_j and β_k . Denote $\theta = \left(\frac{\|\beta_1\|_2}{\sqrt{M_1}}, \dots, \frac{\|\beta_d\|_2}{\sqrt{M_d}} \right)'$. Define $G = \text{diag}(g_1, \dots, g_d)$

where $g_j = \sum_{k=1}^d a_{jk}$. In a network where a_{jk} is the weight of edge (j, k) , g_j is the degree of

vertex j . We have $P_C(\beta) = \frac{1}{2} \lambda_2 \theta' L \theta$ where $L = G - A$. For tightly connected nodes with large a_{jk} 's, the contrast penalty encourages their regression coefficients to be similar. We refer to Liu et al. [2013a] for more detailed development.

Accommodating across-dataset structure

As described in Introduction, there are scenarios under which the regression coefficients of the same gene may have certain similarity across datasets. However, there is no existing technique that promotes such similarity. To further fix idea, consider a small example with three datasets simulated under the distribution specified in the Simulation Study section under Simulation Study II. As a special case of the heterogeneity model, we adopt the homogeneity model under which all three datasets share the same ten important genes. And for a gene, its regression coefficients are identical across datasets. As can be seen from Table 4 (Appendix), the benchmark gBridge estimates may vary significantly across datasets. For example for gene 1 whose true regression coefficient is 0.4, the gBridge estimates are 0.186, 0.391, and 0.112 for the three datasets respectively. In comparison, the proposed contrast penalty can significantly improve estimation. Larger-scale simulation reported in the next section will show that under more realistic settings, such improvement can lead to more accurate marker identification.

With the motivation of promoting similarity of regression coefficients across datasets, we consider the penalized estimate $\hat{\beta} = \operatorname{argmin} \{L(\beta) + P(\beta) + P_C(\beta)\}$, where the contrast penalty

$$P_C(\beta) = \lambda_2 \sum_{j=1}^d \sum_{k \neq l} a_j^{(kl)} \left(\beta_j^{(k)} - \beta_j^{(l)} \right)^2. \quad (1)$$

Here $\lambda_2 \geq 0$ is a data-dependent tuning parameter. $a_j^{(kl)} = I(\operatorname{sgn}(\beta_j^{(k)}) \neq \operatorname{sgn}(\beta_j^{(l)}))$, where $\operatorname{sgn}(\cdot)$ is the sign function. Different from the penalty in the above subsection, $P_C(\beta)$

penalizes the contrast between $\beta_j^{(k)}$ and $\beta_j^{(l)}$, which are the regression coefficients of the same gene in different datasets.

When $\lambda_2 > 0$, the contrast penalty complements $P(\beta)$ and conducts finer tuning of estimation and selection. When $\operatorname{sgn}(\beta_j^{(k)}) \neq \operatorname{sgn}(\beta_j^{(l)})$, gene j demonstrates different effects in different datasets (for example, one positive and one negative effect, or one nonzero and one zero effect). Although the conflicting signs may seem counterintuitive, as described in Liu et al. [2013b], they have been observed in practical data analysis, and there are a few scenarios under which they can be informative (for example, the analysis of multiple “negatively correlated” diseases). However, as they are rare in practice, the contrast penalty is designed to have no effect with conflicting signs. When $\operatorname{sgn}(\beta_j^{(k)}) = \operatorname{sgn}(\beta_j^{(l)})$, gene j has qualitatively similar effects in datasets k and l . The contrast penalty shrinks the difference between $\beta_j^{(k)}$ and $\beta_j^{(l)}$ and encourages them to be similar. That is, it has a smoothing effect. Although the notion of smoothing is not new in penalization, most of the existing smoothing penalties are in the context of single-dataset analysis. The penalty described in the above subsection, although in the context of integrative analysis, still smoothes over regression coefficients in the same dataset.

With practical data, $\operatorname{sgn}(\beta_j^{(k)})$ needs to be estimated. There are several practically feasible proposals. The general strategy is to first conduct a simple estimation and then use the estimated signs. The first estimation approach is marginal analysis of each dataset separately [Huang et al. 2008]. This is in line with marginal screening. The second is penalized estimation (for example Lasso or bridge) with each dataset separately [Huang et al. 2008]. The third is penalized integrative analysis with all datasets [Ma et al. 2012]. In our numerical study, we find that the second and third approaches perform well. We refer to the aforementioned references for data and model conditions under which the three approaches can lead to asymptotically consistent sign estimation. Roughly speaking, it is assumed that $\log(d)/n \rightarrow 0$, and there are a small number of important genes with effects bounded away from zero. In addition, the expressions of important genes and unimportant genes are only weakly correlated. Sub-matrices of the design matrix are stable (for example, the Sparse Riesz Condition is satisfied [Zhang 2010]). Such assumptions can be sensible given that there are only a small number of known “cancer genes” and that genes with different biological functions tend to have weakly correlated expressions. We acknowledge that with

real data, it is hard to determine whether those assumptions are satisfied. In our simulation where real gene expressions are used (and it is not clear whether the assumptions are satisfied), the proposed method has satisfactory performance.

Computational algorithm—First consider continuous responses and linear regression models, where for subject i in dataset m , $Y_i^{(m)} = \beta_0^{(m)} + X_i^{(m)'} \beta^{(m)} + \epsilon_i^{(m)}$, $i = 1, \dots, n^{(m)}$. $\beta_0^{(m)}$ is the intercept, $\beta^{(m)}$ is the length- d vector of unknown regression coefficients, and $\epsilon_i^{(m)}$ is the random error. Assume that $Y_i^{(m)}$ and $X_i^{(m)}$ have been centered, and so the intercept $\beta_0^{(m)} = 0$.

Consider the least squares loss function $L^{(m)}(\beta^{(m)}) = \frac{1}{2n} \sum_{i=1}^{n^{(m)}} \left(Y_i^{(m)} - X_i^{(m)'} \beta^{(m)} \right)^2$.

Denote $Y^{(m)} = (Y_1^{(m)}, \dots, Y_{n^{(m)}}^{(m)})'$ and $X^{(m)} = (X_1^{(m)}, \dots, X_{n^{(m)}}^{(m)})'$. Further denote $\mathbf{Y} = (Y^{(1)'}, \dots, Y^{(M)'})'$ and $\mathbf{X} = \text{diag}(X^{(1)}, \dots, X^{(M)})$. Consider the overall loss function

$L(\beta) = \sum_{m=1}^M L^{(m)}(\beta^{(m)})$. Denote $\mathbf{X}_j$ as the submatrix of $\mathbf{X}$ that corresponds to β_j . Then

$$L(\beta) = \frac{1}{2n} \left\| \mathbf{Y} - \sum_{j=1}^d X_j \beta_j \right\|_2^2.$$

Optimization with the “least squares loss function + gBridge penalty” can be transformed into a series of weighted Lasso-type optimization. Proof follows from Proposition 1 of Huang et al. [2009] and is omitted here. Optimization with penalty function (1) is equivalent to minimizing

$$S(\beta, \theta) = \frac{1}{2n} \left\| \mathbf{Y} - \mathbf{X}\beta \right\|_2^2 + \sum_{j=1}^d \theta_j^{1-1/\gamma} c_j^{1/\gamma} \|\beta_j\|_1 + \tau_n \sum_{j=1}^d \theta_j + \lambda_2 \sum_{j=1}^d \sum_{k \neq l} a_j^{(kl)} \left(\beta_j^{(k)} - \beta_j^{(l)} \right)^2.$$

Here θ is a length- d vector with its j th component $\theta_j \geq 0$, $j = 1, \dots, d$. τ_n is a penalty parameter and can be computed from equation $\lambda_1 = \tau_n^{1-\gamma} \gamma^{-\gamma} (1-\gamma)^{\gamma-1}$. $c_j = M_j^{1-\gamma}$.

For fixed λ_1 , λ_2 , and γ , consider the following iterative algorithm. We use the superscript “[s]” to denote values calculated in iteration s .

1. Denote $\beta^{[0]}$ as the initial estimate. Possible choices of the initial estimate include the ridge or Lasso penalized estimate (from analyzing each dataset separately) and marginal regression estimate. In our numerical study, we use the Lasso estimate and find it satisfactory. Initialize $s = 0$.
2. $s = s + 1$.
3. Compute $\theta_j^{[s]} = c_j \left(\frac{1-\gamma}{\tau_n \gamma} \right)^\gamma \|\beta_j^{[s-1]}\|_1^\gamma$, $j = 1, \dots, d$.
4. Compute

$$\beta^{[s]} = \underset{\beta}{\operatorname{argmin}} \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \sum_{j=1}^d \left(\theta_j^{(s)}\right)^{1-1/\gamma} c_j^{1/\gamma} \|\beta_j\|_1 + \lambda_2 \sum_{j=1}^d \sum_{(k,l): k \neq l} a_j^{(kl)} \left(\beta_j^{(k)} - \beta_j^{(l)}\right)^2. \quad (2)$$

5. Iterate Step 2–4 until convergence.

In the above algorithm, the main computational task is Step 4. This step can be solved using a coordinate descent algorithm, which optimizes with respect to a single element of a group (where a group consists of the coefficients for a single gene across multiple datasets) at a time, and then iterates circularly through all elements and all groups until convergence.

Consider gene j . Let $\mathbf{r}_{-jk} = \mathbf{Y} - \sum_{j=1}^d X_j \beta_j + \mathbf{x}_j^{(k)} \beta_j^{(k)}$, where $\mathbf{x}_j^{(k)}$ is the submatrix of $\mathbf{X}$ that corresponds to $\beta_j^{(k)}$. At iteration $s-1$, define the index set

$A_j^{(k)} = \{l : l=1, \dots, M, l \neq k \text{ and } a_j^{(kl)} \neq 0\}$, and denote $m_j^{(k)}$ as its size. Let $\tilde{\beta}_j$ be the current estimate. Terms that involve $\beta_j^{(k)}$ in Step 4 can be written as

$$\begin{aligned} R\left(\beta_j^{(k)}\right) &= \frac{1}{2n} \|\mathbf{r}_{-jk}^{[s-1]} - \mathbf{X}_j^{(k)} \beta_j^{(k)}\|^2 + \zeta_j |\beta_j^{(k)}| + \frac{1}{2} \lambda_2 \sum_{l \in A_j^{(k)}} \left(\beta_j^{(k)} - \tilde{\beta}_j^{(l)}\right)^2 \\ &= \frac{1}{2} \left[\frac{1}{n} \mathbf{X}_j^{(k)'} \mathbf{X}_j^{(k)} + \lambda_2 m_j^{(k)} \right] \left(\beta_j^{(k)}\right)^2 - \left[\frac{1}{n} \left(\mathbf{r}_{-jk}^{[s-1]}\right)' \mathbf{X}_j^{(k)} + \lambda_2 \sum_{l \in A_j^{(k)}} \tilde{\beta}_j^{(l)} \right] \beta_j^{(k)} + \zeta_j |\beta_j^{(k)}| + c \\ &= \frac{1}{2} a_j^{(k)} \left(\beta_j^{(k)}\right)^2 - b_j^{(k)} \beta_j^{(k)} + \zeta_j |\beta_j^{(k)}| + c, \end{aligned}$$

where $a_j^{(k)} = \frac{1}{n} \mathbf{x}_j^{(k)'} \mathbf{x}_j^{(k)} + \lambda_2 m_j^{(k)}$, $b_j^{(k)} = \frac{1}{n} \left(\mathbf{r}_{-jk}^{[s-1]}\right)' \mathbf{x}_j^{(k)} + \lambda_2 \sum_{l \in A_j^{(k)}} \tilde{\beta}_j^{(l)}$,

$c = \frac{1}{2n} \|\mathbf{r}_{-jk}^{[s-1]}\|_2^2 + \frac{1}{2} \lambda_2 \sum_{l \in A_j^{(k)}} \left(\tilde{\beta}_j^{(l)}\right)^2$ and $\zeta_j = \left(\theta_j^{[s]}\right)^{1-1/\gamma} c_j^{1/\gamma}$. Since c is a constant with

respect to $\beta_j^{(k)}$, the minimizer of (3) is $\hat{\beta}_j^{(k)} = \frac{\operatorname{sgn}\left(b_j^{(k)}\right)}{a_j^{(k)}} \left(|b_j^{(k)}| - \zeta_j\right)_+$, where $u_+ = \max(0, u)$.

In numerical study, we use the ℓ_2 -norm of the difference between two consecutive estimates less than 0.001 as the convergence criterion (for both the coordinate descent and the overall iterative algorithm). Convergence of the coordinate descent algorithm can be derived following Tseng [2001]. The proposed overall algorithm always converges, since at each step, the nonnegative objective function decreases. As the gBridge type penalty is not convex, there is no guarantee that the algorithm converges to the global optimizers.

Consider other types of data and models. A generic choice of the loss function is the negative log-likelihood function. As shown in Appendix, estimating-equation based loss can also be adopted. Consider the following iterative algorithm: (i) Initialize $\beta^{(m)}$ ($m = 1, \dots, M$) as the Lasso or bridge penalized estimate from analyzing dataset m ; (ii) At the current estimate of β , make the Taylor expansion of $L(\beta)$. Keep the linear and quadratic terms; (iii)

Call the algorithm developed for linear regression; (iv) Repeat Step (ii) and (iii) until convergence.

The proposed method involves three tuning parameters γ , λ_1 , and λ_2 . With bridge-type penalties, the value of γ is often pre-fixed, as asymptotically, different values lead to similar results [Huang et al. 2008]. In our numerical study, we set $\gamma = 0.5$. λ_1 controls the overall sparsity of marker selection. λ_2 controls the smoothness among the coefficients for the same gene. In data analysis, λ_1 and λ_2 are chosen using V -fold cross validation ($V = 5$). For λ_1 , we search over the discrete grid of $2 \cdots, -2, -1, 0, 1, 2, \dots$. For λ_2 , our limited experience suggests that the estimates can be relatively less sensitive to its value. Thus to reduce computational cost, we search over the discrete grid of $10 \cdots, -2, -1, 0, 1, 2, \dots$. Research R code is available at <http://works.bepress.com/shuangge/46/>.

Simulation Study

The proposed contrast penalty can be applicable to multiple types of data and multiple models. In numerical study, as a specific example, we analyze right censored survival data under the AFT model. Details on the estimation procedure are provided in Appendix. Analysis of other types of data/models will be postponed to future studies.

Simulation study I

In the Data Analysis section below, we analyze three lung cancer datasets, which have 175, 79, and 82 subjects, respectively. Gene expression measurements on 22,283 probes are available. To closely mimic real data, we simulate three datasets with sample sizes equal to the lung cancer datasets. For each simulation replicate, the observed expressions of 1,000 randomly selected genes are used. We simulate multiple heterogeneity models. Under scenario (a), the three datasets share seven common markers. In addition, they also have two, three, and four dataset-specific markers. Under scenario (b), the three datasets share four common markers. In addition, they also have four, six, and eight dataset-specific markers. Under scenario (c), each dataset has ten unique markers. As a special case of the heterogeneity model, the homogeneity model is also considered, under which all three datasets have the same ten markers. Under all simulation scenarios, there are a total of 30 markers. For marker sets of different datasets, the simulated scenarios cover the whole spectrum from complete, partial, to no overlap. The absolute values of nonzero coefficients are simulated from uniform(0.4, 1.0). Their signs are determined by Binomial(1, 0.5). The random errors have standard normal distributions. We simulate the log event times from AFT models with intercept equal to 0. The log censoring times are independently generated as having uniform distributions. The censoring rate is about 35%.

Simulation suggests that the proposed method is computationally affordable. With fixed tunings, the analysis of one replicate takes 15 seconds on a regular PC. We evaluate marker identification accuracy using the number of true positives and number of false positives. In addition, we also evaluate prediction performance. For each simulated replicate (training data), three additional datasets (testing data) are generated under the same settings. Estimates are generated using the training data and then used to make prediction for subjects in the testing data. We follow Huang and Ma [2010]. For each subject, the prediction risk

score $X'\beta$ is calculated. For each dataset, we dichotomize the risk scores at the median and create two risk groups. The logrank statistic is computed to quantify whether the two groups have different survival functions. For each replicate, we compute the mean logrank statistics across the three datasets. The logrank statistic has a χ^2 distribution with degree of freedom 1. A value greater than 3.84 is significant at the 0.05 level.

Summary statistics based on 100 replicates are shown in Table 1. With the contrasted gBridge, we show the results under five λ_2 values (with cross-validated λ_1) to demonstrate the whole spectrum of the effects of contrast penalty. We also apply the gBridge as a benchmark. Table 1 suggests that the contrasted gBridge identifies a similar number of true positives as the gBridge. For example, under the heterogeneity model scenario (b), the numbers of true positives are 17.5 (gBridge), 17.2, 17.2, 16.9, 16.5 and 16.1 (contrasted gBridge with different λ_2). However, the contrast penalty may lead to a significant reduction in false positives. For example under the heterogeneity model scenario (a), the numbers of false positives are 23.9 (gBridge), 17.0, 14.2, 12.4, 9.8 and 9.0 (contrasted gBridge with different λ_2). The prediction logrank statistics are similar under different approaches. As the proposed method encourages similarity across datasets, its advantage over the benchmark gets more significant as the overlap of the three marker sets increases. It is also interesting to note that under heterogeneity model scenario (c) where the three datasets do not have any common marker, the proposed method has similar performance as the benchmark. Thus with practical data when the similarity across datasets is not clear, it is still “safe” to use the proposed method. As with other methods, performance of the proposed method improves as the “signals” get stronger (results omitted here and available from the authors).

Simulation study II

In this set of simulation, we follow many published studies and simulate gene expressions from parametric models. Specifically, three datasets are simulated, each with 100 subjects. For each subject, the expressions of 1,000 genes are simulated to have a multivariate normal distribution. The marginal means are equal to zero. The covariance matrix $\Sigma = (\sigma_{jk})_{d \times d}$ has entries $\sigma_{jj} = 1, j = 1, \dots, d$ and $\sigma_{jk} = \rho_{jk}, j \neq k$. Consider the following correlation structures. The first is the auto-regressive correlation, where $\rho_{jk} = \rho^{|j-k|}$ with $\rho = 0.2$ and 0.7 , corresponding to weak and strong correlations, respectively. The second is the banded correlation. Here two scenarios are considered. Under the first scenario, $\rho_{jk} = 0.33$ if $|j - k| = 1$ and 0 otherwise. Under the second scenario, $\rho_{jk} = 0.6$ if $|j - k| = 1$, 0.33 if $|j - k| = 2$, and 0 otherwise.

We also consider both the heterogeneity and homogeneity models. Under the heterogeneity model, the three datasets share five common markers. In addition, each dataset has five datasetspecific markers. Under the homogeneity model, the datasets share the same ten markers. Thus there are also a total of 30 markers. For the m th dataset, the logarithm of event time is simulated from the model $T^{(m)} = \beta_0^{(m)} + X^{(m)'} \beta^{(m)} + \epsilon^{(m)}$, where $\beta_0^{(m)} = 0.5$ and $\epsilon^{(m)} \sim N(0, 1)$. The regression coefficients of the markers are (0.4, 0.5, 0.6, 0.7, 0.8, -0.4, -0.5, -0.6, -0.7, -0.8), (0.4, -0.5, 0.6, -0.7, 0.8, -0.4, 0.5, -0.6, 0.7, -0.8) and (0.4, 0.5, 0.6, 0.7, 0.8, -0.4, -0.5, -0.6, -0.7, -0.8) for dataset 1–3, respectively. For markers shared by the three datasets, the regression coefficients may have the same or different signs in

different datasets. The censoring times are generated as uniformly distributed and independent of the event times. The censoring rate is about 30%.

Summary statistics based on 100 replicates are shown in Table 5 and 6 (Appendix) for the heterogeneity and homogeneity models, respectively. The format is the same as in Table 1. We again observe that the contrasted penalization identifies a similar number of true positives as the benchmark but a smaller number of false positives. The prediction logrank statistics are similar.

Remarks

Under certain simulation scenarios, for example heterogeneity model scenario (c) in Table 1, performance of the proposed method seems unsatisfactory. The power of integrative analysis lies in pooling common information across datasets. Under scenario (c), the three datasets do not have any common marker, which explains the inferior performance. As with other marker identification methods, performance of the proposed method also depends on the relative levels of “signals” and “noises”. We intentionally choose the settings under which the best possible performance (homogeneity model in Table 1) is less than perfect to reflect the reality that only a small fraction of the identified cancer markers have been confirmed. We have experimented with larger nonzero coefficient values and observed improvement performance (results omitted here). The improvement of the proposed method over the benchmark may not seem dramatic, which is also reasonable. The contrast penalty is meant to complement the gBridge penalty and accommodate finer data/model structures. In addition, the reduction in false positives can lead to more focused hypotheses testing and functional studies, which have important practical implications.

Data Analysis

Lung cancer study

Lung cancer is the leading causes of cancer death for both men and women in the United States. Gene profiling studies have been widely conducted on lung cancer, searching for markers associated with prognosis. We collect and analyze three lung cancer prognosis datasets with gene expression measurements. The UM (University of Michigan Cancer Center) dataset has a total of 175 patients, among whom 102 died during follow-up. The median follow-up is 53 months. The HLM (Moffitt Cancer Center) dataset has a total of 79 patients, among whom 60 died during follow-up. The median follow-up is 39 months. The CAN/DF (Dana-Farber Cancer Institute) dataset has a total of 82 patients, among whom 35 died during follow-up. The median follow-up is 51 months. We refer to Xie et al. [2011] and references therein for more detailed information. A total of 22,283 probe sets were profiled in all three datasets. Gene expression normalization is conducted for each dataset separately. In order to reduce computational cost, and as genes with higher variations are often of more interest, the probe sets are ranked using their variations, and the top 2,000 are screened out for analysis. For each gene expression and each dataset separately, the mean is normalized to zero, and the variance is normalized to one.

The contrasted gBridge identifies eight genes. Genes and their estimates are shown in Table 2. Gene PSPH is identified for the HLM and CAN/DF datasets but not for UM. All other

genes are identified for all three datasets. With the contrast penalty, some genes have estimated regression coefficients very close or even identical across datasets (for example for gene FOXM1, all the estimated coefficients are -0.0175), whereas others have different coefficients (for example for gene BMP2, the estimated regression coefficients are -0.0167 , -0.0167 , and -0.0045). The estimated regression coefficients are in general small. There are two main reasons. The first is that the log-transformed event times are “clustered”. As with simple linear regression, a rescaling of the event times will increase the magnitudes of estimates. In addition, penalization methods in general shrink estimates towards zero. As a comparison, gBridge identifies twelve genes (Table 7, Appendix). Among them, two genes are identified in all three datasets, five genes are identified in two datasets, and the rest five are identified in only one dataset. Compared with gBridge, the contrasted gBridge identifies gene sets that are much more coherent across datasets. Table 7 also shows that the gBridge estimates may vary significantly across datasets. For example for gene FOXM1, the estimated regression coefficients are -0.0238 , -0.0045 , and -0.0085 , respectively. Although there is considerable overlap, the two methods identify different sets of genes.

To provide a more comprehensive description of the identified genes, we evaluate their stability by computing the observed occurrence index [Huang and Ma 2010]. In particular, each dataset is randomly split into two sets with sizes 3:1. We apply the proposed method to the first set and identify markers. To avoid an extreme split, this process is repeated 100 times. For a gene, we compute the probability it is identified out of the 100 splits, and this probability is referred to as the observed occurrence index in Huang and Ma [2010]. Table 2 shows that the majority of identified genes have the observed occurrence indexes equal to one, suggesting satisfactory stability. The lowest occurrence index is 0.87, which is for gene FOS and the HLM and CAN/DF datasets.

Prediction performance is also evaluated. As we do not have independent testing data, we again resort to random splitting [Huang and Ma 2010]. Each dataset is split into a training set and a testing set with sizes 3:1. Estimates are generated using the training set. Then prediction is made for subjects in the testing set. Prediction performance is evaluated using the logrank statistic, which is computed in a similar manner as in simulation. To avoid an extreme split, this process is repeated 100 times, and the average logrank statistic is computed. This evaluation process is largely a byproduct of the occurrence index calculation and incurs little additional cost. The average logrank statistics are 9.112 (contrasted gBridge, p-value 0.0025) and 7.534 (gBridge, p-value 0.0061), respectively. The contrast penalty leads to improved prediction.

Breast cancer study

Breast cancer is the second leading cause of cancer deaths after lung cancer. It is one of the most extensively investigated cancers in the genomic era. Here data from three gene expression studies are collected and analyzed. The dataset reported in Huang et al. [2003] has 71 patients, among whom 35 died during follow-up. The median follow-up is 39 months. The dataset reported in Sotiriou et al. [2003] has 98 patients, and 45 died during follow-up. The median follow-up is 67.9 months. The dataset reported in van't Veer et al. [2002] has 78 patients, and 34 died during follow-up. The median follow-up is 64.2 months.

We analyze the 2555 genes that were profiled in all three datasets and refer to the original publications for more detailed information.

With the contrasted gBridge, seventeen genes are identified in at least one dataset (Table 3). Among them, twelve genes are identified in all three datasets, four genes are identified in two datasets, and one gene is identified in one dataset. Other observed patterns are similar to those with the lung cancer datasets. The gBridge identifies eighteen genes (Table 8, Appendix). Among them, eight are identified in three datasets, nine are identified in two datasets, and one gene is identified in one dataset. With the contrasted gBridge, most of the identified genes have satisfactory stability, except for gene Hs.33287 with the dataset in Sotiriou et al. [2003] which has the observed occurrence index equal to 0.30. Prediction performance of the two methods is evaluated using the approach described in the above subsection. The logrank statistics are 6.264 (contrasted gBridge, p-value 0.012) and 0.072 (gBridge, p-value 0.788), respectively. Prediction using the contrasted gBridge is significant and much better than that of gBridge.

Discussion

In cancer studies with high-dimensional genetic and genomic measurements, integrative analysis provides an effective way of pooling information across multiple independent datasets and can lead to improved marker identification. In this study, we have considered integrative analysis and penalized marker selection under both the heterogeneity and homogeneity models. In particular, the composite penalization based on the bridge penalty is adopted. Advancing from the published studies, we consider the within- and across-dataset structures of covariates. Such structures are concerned with the interconnections among covariates and their regression coefficients. Naturally, the contrasted penalization approach imposes penalties on the contrasts between coefficients to accommodate such interconnections. Advancing from the existing studies in particular Liu et al. [2013a], we have provided detailed development of a penalization method that accommodates the across-dataset structure. The proposed method is intuitively reasonable and can be realized using an effective iterative algorithm. Simulation shows that the contrasted penalization can outperform the benchmark penalization method by identifying a similar number of true positives but much fewer false positives. As the contrast penalty is concerned with secondary data structures, the observed improvement may not be dramatic. In practice, the significantly reduced number of false positives can be valuable. The analysis of lung cancer and breast cancer datasets shows that the proposed method can identify genes different from the benchmark. The proposed method identifies gene sets that are more coherent across datasets. In addition, the identified genes have satisfactory stability and better prediction performance.

There are multiple ways of specifying the covariate/coefficient structures. Even under the present specification, there can be multiple ways of constructing contrast penalties. For example, when accommodating the within-dataset structure, the contrast penalty penalizes $\|\beta_j\|_2 - \|\beta_k\|_2$. A natural alternative is to penalize $\|\beta_j - \beta_k\|_2$. When accommodating the across-dataset structure, the contrast penalty encourages $\beta_j^{(k)}$ and $\beta_j^{(l)}$ to have similar

magnitudes. This is sensible when multiple datasets measure the same outcome and have been properly normalized. When multiple datasets use different platforms and have not been properly normalized, it is more meaningful to encourage the similarity between $\text{sgn}(\beta_j^{(k)})$ and $\text{sgn}(\beta_j^{(l)})$. We have experimented with this possibility. However, as the sign function is not continuous, this alternative may bring tremendous computational challenges. Determination of a proper contrast penalty depends on the data settings, analysis goal, and computational feasibility. The investigation of other contrast penalties is nontrivial and will be postponed to future research. To demonstrate the proposed contrast penalty, we couple it with gBridge. It is also applicable with other composite penalties. In data analysis, we adopt the AFT models for the lung and breast cancer data. The applicability of contrast penalty is relatively “independent” of the loss function. Thus, alternative models, for example the Cox model, can be adopted. More comprehensive examination of the contrast penalty with other data/models will be pursued in the future. We compare the proposed “gBridge+contrast penalty” against gBridge to directly establish the merit of the contrast penalty. We recognize that there are multiple methods, for example the group MCP, that can analyze the simulated and real data. However, a comparison between “gBridge+contrast penalty” and for example group MCP may not be fair as they use different base methods for selection. Limitations of this study also include a lack of theoretical investigation and more detailed analysis of the identified genes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank the editor, associate editor, and two referees for careful review and insightful comments, which have led to a significant improvement of this article. This research was supported by NIH grants CA165923 and CA142774, National Social Science Foundation of China (13CTJ001), and National Bureau of Statistics Funds of China (2012LD001).

References

- [1]. Buhlmann, P.; van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer; 2011.
- [2]. Guerra, R.; Goldstein, DR. *Meta-Analysis and Combining Information in Genetics and Genomics*. 1st edition. Chapman and Hall/CRC; 2009.
- [3]. Huang E, Cheng S, Dressman H, Pittman J, Tsou M, Horng C, Bild A, Iversen E, Liao M, Chen C, West M, Nevins JR, Huang AT. Gene expression predictors of breast cancer outcomes. *The Lancet*. 2003; 361:1590–1596.
- [4]. Huang J, Horowitz JL, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Annals of Statistics*. 2008; 36:587–613.
- [5]. Huang J, Ma S. Variable selection in the accelerated failure time model via the bridge method. *Lifetime Data Analysis*. 2010; 16:176–195. [PubMed: 20013308]
- [6]. Huang J, Ma S, Li H, Zhang CH. The sparse Laplacian shrinkage estimator for high-dimensional regression. *Annals of Statistics*. 2011; 39:2021–2046. [PubMed: 22102764]
- [7]. Huang J, Ma S, Xie H. Regularized estimation in the accelerated failure time model with high-dimensional covariates. *Biometrics*. 2006; 62:813–820. [PubMed: 16984324]

- [8]. Huang J, Ma S, Xie H, Zhang CH. A group bridge approach for variable selection. *Biometrika*. 2009; 96:339–355. [PubMed: 20037673]
- [9]. Huang Y, Huang J, Shia BC, Ma S. Identification of cancer genomic markers via integrative sparse boosting. *Biostatistics*. 2012; 13:509–522. [PubMed: 22045909]
- [10]. Liu J, Huang J, Ma S. Incorporating network structure in integrative analysis of cancer prognosis data. *Genetic Epidemiology*. 2013a; 37(2):173–183. [PubMed: 23161517]
- [11]. Liu J, Huang J, Ma S. Integrative analysis of cancer diagnosis studies with composite penalization. *Scandinavian Journal of Statistics*. 2013b In press.
- [12]. Liu J, Huang J, Ma S, Wang K. Incorporating group correlations in genome-wide association studies using group Lasso. *Biostatistics*. 2013c; 14(2):205–219. [PubMed: 22988281]
- [13]. Ma S, Dai Y, Huang J, Xie Y. Identification of breast cancer prognosis markers via integrative analysis. *Computational Statistics and Data Analysis*. 2012; 56:2718–2728. [PubMed: 22773869]
- [14]. Ma S, Huang J, Wei F, Xie Y, Fang K. Integrative analysis of multiple cancer prognosis studies with gene expression measurements. *Statistics in Medicine*. 2011; 30:3361–3371. [PubMed: 22105693]
- [15]. Sotiriou C, Neo S, McShane L, Korn E, Long P, Jazaeri A, Martiat P, Fox S, Harris A, Liu E. Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences*. 2003; 100:10393–10398.
- [16]. Stute W. Distributional convergence under random censorship when covariates are present. *Scandinavian Journal of Statistics*. 1996; 23:461–471.
- [17]. Tseng P. Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*. 2001; 109:475–494.
- [18]. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*. 2003; 415:530–536. [PubMed: 11823860]
- [19]. Xie Y, Xiao G, Coombes K, Behrens C, Solis L, Raso G, Girard L, Erickson H, Roth J, Heymach J, Moran C, Danenberg K, Minna J, Wistuba I. Robust gene expression signature from formalin-fixed paraffin-embedded samples predicts prognosis of nonsmall-cell lung cancer patients. *Clinical Cancer Research*. 2011; 17:5705–5714. [PubMed: 21742808]
- [20]. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*. 2010; 38:894–942.

Table 1

Simulation study I. The first row: number of true positives (standard deviation). The second row: number of false positives (standard deviation). The third row: logrank statistic.

gBridge	Contrasted gBridge				
	$\lambda_2 = 0.01$	$\lambda_2 = 0.1$	$\lambda_2 = 1$	$\lambda_2 = 10$	$\lambda_2 = 100$
Heterogeneity model scenario (a)					
23.0(1.9)	23.1(1.9)	23.0(2.0)	22.9(1.8)	22.7(2.0)	22.7(2.0)
23.9(22.0)	17.0(19.9)	14.2(15.0)	12.4(13.8)	9.8(11.1)	9.0(10.8)
8.3	12.2	10.3	9.7	12.6	10.0
Heterogeneity model scenario (b)					
17.5(2.2)	17.2(2.1)	17.2(2.0)	16.9(2.2)	16.5(2.3)	16.1(2.5)
34.0(22.2)	26.6(19.2)	27.7(16.0)	27.3(15.2)	25.9(14.4)	22.9(14.5)
9.1	11.3	11.9	9.1	10.2	8.8
Heterogeneity model scenario (c)					
11.8(2.3)	11.7(2.4)	11.7(2.4)	11.4(2.4)	10.7(2.3)	10.4(2.3)
33.5(18.5)	31.0(17.0)	33.0(14.1)	38.0(16.3)	31.0(11.9)	31.0(11.9)
8.2	10.9	9.4	7.4	8.7	7.9
Homogeneity model					
28.3(2.6)	28.3(2.5)	28.3(2.5)	28.2(2.4)	27.9(2.7)	27.9(2.6)
17.2(25.5)	15.9(25.0)	12.6(22.4)	10.2(17.0)	8.4(13.4)	7.7(12.7)
9.6	11.4	10.4	9.6	11.9	9.6

Table 2

Analysis of the lung cancer datasets using contrasted gBridge: identified genes, estimates, and observed occurrence indexes.

Gene	Probe	UM	<i>P</i>	HLM	<i>P</i>	CAN/DF	<i>P</i>
FOXM1	202580_x_at	−0.0175	1.00	−0.0175	1.00	−0.0175	1.00
CX3CL1	203687_at	0.0079	1.00	0.0079	1.00	0.0079	1.00
PSPH	205048_s_at			−0.0112	1.00	−0.0111	1.00
BMP2	205289_at	−0.0167	1.00	−0.0167	1.00	−0.0045	1.00
SCGB1A1	205725_at	−0.0080	1.00	0.0097	1.00	−0.0080	1.00
MT1H	206461_x_at	−0.0074	1.00	0.0048	1.00	−0.0074	1.00
FOS	209189_at	−0.0167	1.00	−0.0002	0.87	−0.0002	0.87
PLA2G4A	210145_at	0.0147	1.00	−0.0076	1.00	0.0147	1.00

Table 3

Analysis of the breast cancer datasets using contrasted gBridge: identified genes, estimates, and observed occurrence indexes.

UniGene	Gene	Huang	<i>p</i>	Sotiriou	<i>p</i>	Vantveer	<i>p</i>
Hs.115907	DGKD	0.0111	1.00	0.0052	1.00	0.0076	1.00
Hs.117546	NNAT	0.0137	1.00	0.0121	1.00	-0.0152	1.00
Hs.155314	NUP93	-0.0016	0.60			-0.0028	0.64
Hs.166204	PHF1	0.0109	1.00	0.0036	1.00	0.0056	1.00
Hs.19904	CTH	-0.0170	1.00	0.0059	1.00	-0.0151	0.99
Hs.23103	BET1	-0.0010	0.99	0.0207	1.00	0.0026	1.00
Hs.283565	FOSL1	-0.0275	1.00	0.0332	1.00	-0.0277	1.00
Hs.3136	PRKAG1	0.0091	0.99	0.0072	0.99	0.0056	0.99
Hs.33287	NFIB	0.0163	1.00	-0.0016	0.30		
Hs.334534	GNS	0.0072	1.00	0.0097	1.00		
Hs.433300	FCER1G	0.0016	1.00				
Hs.433714	KRAS	0.0017	0.99	-0.0013	0.96	0.0009	0.98
Hs.4980	LDB2	-0.0135	1.00	-0.0107	1.00	-0.0093	1.00
Hs.82002	EDNRB	-0.0206	1.00	-0.0053	0.93	-0.0012	0.97
Hs.82508	THAP11	-0.0082	1.00	0.0110	1.00	-0.0035	1.00
Hs.9629	PRCC	0.0149	1.00	0.0161	1.00	-0.0062	1.00
Hs.2055	UBA1			-0.0179	1.00	-0.0129	1.00