Biostatistics (2017) **18**, 1, *pp.* 1–14
doi:10.1093/biostatistics/kxw022
Advance Access publication on June 20, 2016

Prediction of cancer drug sensitivity using high-dimensional omic features

TING-HUEI CHEN

Department of Mathematics and Statistics, Laval University, 1045 Medicine Avenue, office 1056, Quebec, G1V 0A6, Canada

WEI SUN*

Department of Biostatistics, University of North Carolina at Chapel Hill, 135 Dauer Drive, 3101 McGavran-Greenberg Hall, Chapel Hill, NC 27599-7420 and Public Health Sciences Division, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Seattle, WA 98109-1024
wsun@fredhutch.org

Summary

A large number of cancer drugs have been developed to target particular genes/pathways that are crucial for cancer growth. Drugs that share a molecular target may also have some common predictive omic features, e.g., somatic mutations or gene expression. Therefore, it is desirable to analyze these drugs as a group to identify the associated omic features, which may provide biological insights into the underlying drug response. Furthermore, these omic features may be robust predictors for any drug sharing the same target. The high dimensionality and the strong correlations among the omic features are the main challenges of this task. Motivated by this problem, we develop a new method for high-dimensional bilevel feature selection using a group of response variables that may share a common set of predictors in addition to their individual predictors. Simulation results show that our method has a substantially higher sensitivity and specificity than existing methods. We apply our method to two large-scale drug sensitivity studies in cancer cell lines. Both within-study and between-study validation demonstrate the good efficacy of our method.

Keywords: Bilevel selection; Cancer cell lines; Drug sensitivity.

1. Introduction

Human cancer arises from an accumulation of somatic mutations during the lifetime of a patient. Interventions targeting mutated proteins or relevant pathways have proved to be effective treatment options. However, not all the patients with the targeted somatic lesions respond to the therapy. For example, several cancer drugs target over-expression of oncogene HER2. Among those breast cancer patients with HER2 over-expression, only 30% respond to such targeted therapy (de Palma and Hanahan, 2012). Such heterogeneous response may be due to underlying genomic heterogeneity, which can be manifested by different omic features, e.g., DNA alterations, gene expression, or epigenetic marks. Preclinical model systems

*To whom correspondence should be addressed.

such as cancer cell lines that reflect the genomic diversity of human cancers can be used to identify predictive omic features/biomarkers for drug sensitivity (Caponigro and Sellers, 2011). Recently, two groups (Garnett *and others*, 2012; Barretina *and others*, 2012) have studied drug sensitivity in a large number of cancer cell lines and measured several types of omic features including somatic mutations of cancer genes, genome-wide copy number aberrations, and gene expression. The sample size ranges from 200 to 500 per drug, while the number of omic features is greater than 10 000. The authors conducted drug-by-drug analysis to identify associated omic features, and they demonstrated that drugs with the same targets have some common predictive omic features in addition to their individual features. Therefore, a joint analysis of the drugs sharing a target may improve the sensitivity and specificity to identify their shared predictive omic features. To this end, we consider the feature selection method for multivariate responses to identify predictive omic features of drugs with the same target.

Two types of methods have been developed for feature selection for multivariate responses: group-wise selection and bilevel selection. Group-wise selection methods, such as group Lasso (Yuan and Lin, 2006) or group adaptive Lasso (Wang and Leng, 2008), assume that all the response variables within a group are associated with the same set of covariates (Huang *and others*, 2012). This assumption is not reasonable for cancer drug-sensitivity studies because drugs with the same target may have different predictive omic features. In contrast, bilevel selection methods encourage the selection of covariates associated with all the response variables, but they also allow some covariates to be associated with one or a few response variables (Breheny and Huang, 2009). The flexibilities in bilevel selection are desirable for cancer drug-sensitivity applications. A few methods have been developed for bilevel selection with one or more response variables, such as group bridge (gBridge, Huang *and others*, 2009), composite MCP (cMCP, Breheny and Huang, 2009), sparse-group lasso (SGL, Simon *and others*, 2013), group exponential Lasso (FEb, Breheny, 2015), and group variable selection via convex log-exp-sum penalty (Chen *and others*, 2014). Although these methods work satisfactorily in many real-data analyses, we find that their performance is limited in cancer drug sensitivity studies where the response variables are multivariate and the covariates are high-dimensional omic features with strong correlations. These issues motivate us to develop a new method to construct predictive models of cancer drug sensitivity using omic features.

In this article, we propose a new bilevel selection method called BipLog and apply it to analyze the aforementioned drug sensitivity data sets. We seek to answer a few important questions in our data analysis. First, by splitting the data from Garnett *and others* (2012) into training and testing sets, we assess the variation in the drug sensitivity that can be explained by our predictive model. Second, we use all the data from Garnett *and others* (2012) to select omic features associated with each drug target, and we evaluate their prediction performance using independent data from Barretina *and others* (2012). There are substantial differences in these two studies in terms of the drugs studied and the method to estimate the drug sensitivity. Therefore, this between-study comparison helps to evaluate the robustness and generality of our method. Third, we use this between-study comparison to compare the results of BipLog with those of the "drug-by-drug" analysis using the elastic net (Zou and Hastie, 2005).

The remainder of this article is organized as follows. In Section 2, we introduce of BipLog and its implementation. We present the simulation studies and real-data analyses in Sections 3 and 4. Section 5 provides concluding remarks.

2. METHODS

2.1 *Objective function*

Suppose in a particular drug group that shares a target, we observe measurements of drug sensitivity of q drugs (response variables), denoted by $\mathbf{y}_k = (y_{1k}, \ldots, y_{nk})^T$ $(1 \leq k \leq q)$, and p omic features (covariates),

denoted by $\mathbf{x}_j = (x_{1j}, \ldots, x_{nj})^T$ $(1 \le j \le p)$ for n samples. We assume that q is much smaller than the sample size n, but p is often larger or much larger than n. After standardizing $\mathbf{y}_k$ and $\mathbf{x}_j$ to have mean 0 and $\|\mathbf{y}_k\|_2^2 = \|\mathbf{x}_j\|_2^2 = n$, we assume a linear system: $E(\mathbf{y}_k) = \mathbf{X}\boldsymbol{\beta}_k = \sum_{j=1}^{p} \mathbf{x}_j \beta_{jk}$, where $\mathbf{X}_{n\times p} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ and $\boldsymbol{\beta}_k = (\beta_{1k}, \ldots, \beta_{pk})^T$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)$ and denote each row of $\boldsymbol{\beta}$ by $\mathbf{b}_j = (\beta_{j1}, \ldots, \beta_{jq})$. Let $|\mathbf{b}_j| = \sum_{k=1}^{q} |\beta_{jk}|$. The objective function that we aim to minimize is a penalized least squares:

$$Q(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{k=1}^{q} \|\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k\|_2^2 + \sum_{j=1}^{p} \sum_{k=1}^{q} p_{\boldsymbol{\theta}_1}\left(|\beta_{jk}|\right) + \sum_{j=1}^{p} p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|), \tag{2.1}$$

where $p_{\boldsymbol{\theta}_1}\left(|\beta_{jk}|\right) = \lambda_1 \log(|\beta_{jk}| + \tau_1)$, $p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|) = \lambda_2 \log(|\mathbf{b}_j| + \tau_2)$, $\boldsymbol{\theta}_1 = (\lambda_1, \tau_1)$, and $\boldsymbol{\theta}_2 = (\lambda_2, \tau_2)$. In its general form, $p_{\varpi}(\beta) = \lambda \log(|\beta| + \tau)$ is the Log penalty for a parameter β with tuning parameters $\varpi = (\lambda, \tau)$.

In our previous works (Sun *and others*, 2010; Chen *and others*, 2015), we have shown that the Log penalty has promising performances in genomic studies. The Log penalty was originally proposed by Friedman (2008), and it was discussed by Mazumder *and others* (2011) in the form of $\frac{\lambda}{\log(r+1)} \log(r|\beta|+1)$, with $r > 0$ and $\lambda > 0$. By applying L'H™ôpital's rule to this form of Log penalty, it can be shown that it bridges the L_1 penalty (as $r \to 0+$) and the L_0 penalty (as $r \to \infty$). The Log penalty in (2.1) is a reparameterization of this form. The Log penalty is a nonconvex penalty or more precisely a folded concave penalty (Fan and Lv, 2010) in the sense that it is concave for $\beta \in [0, \infty)$, with continuous derivative $p'_{\varpi}(\beta) \ge 0$, and $p'_{\varpi}(0+) > 0$. Similar to other types of folded concave penalties, the Log penalty mitigates the estimation bias produced by convex penalty (e.g., Lasso penalty) and it can achieve variable selection consistency without requiring restricted irrepresentable condition (Zhao and Yu, 2006). We achieved bilevel selection by applying Log penalties to each coefficient ($\sum_{j=1}^{p} \sum_{k=1}^{q} p_{\boldsymbol{\theta}_1}\left(|\beta_{jk}|\right)$) and each group of coefficients ($\sum_{j=1}^{p} p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|)$), respectively. In the following section, we will give more explanation and justification of this particular form of bilevel penalty.

2.2 *Computation*

We estimate the $\boldsymbol{\beta}$ that minimizes $Q(\boldsymbol{\beta})$ in (2.1) using a combination of local linear approximation (LLA) (Zou and Li, 2008) and a coordinate descent algorithm. Specifically, given initial values of $\boldsymbol{\beta}$ or the estimates from the tth iteration, denoted $\{\hat{\beta}_{jk}^{(t)}\}$, we apply LLA to the Log penalty functions $p_{\boldsymbol{\theta}_1}\left(|\beta_{jk}|\right)$ and $p_{\boldsymbol{\theta}_2}\left(|\mathbf{b}_j|\right)$ to update them at the $(t+1)$th iteration:

$$p_{\boldsymbol{\theta}_1}\left(|\beta_{jk}|\right) \approx p_{\boldsymbol{\theta}_1}\left(|\hat{\beta}_{jk}^{(t)}|\right) + \left.\frac{\partial p_{\boldsymbol{\theta}_1}\left(|\beta_{jk}|\right)}{\partial |\beta_{jk}|}\right|_{|\beta_{jk}|=|\hat{\beta}_{jk}^{(t)}|} \left(|\beta_{jk}| - |\hat{\beta}_{jk}^{(t)}|\right) = \frac{\lambda_1 |\beta_{jk}|}{|\hat{\beta}_{jk}^{(t)}| + \tau_1} + C_1,$$

$$p_{\boldsymbol{\theta}_2}(|\mathbf{b}_j|) \approx p_{\boldsymbol{\theta}_2}\left(|\hat{\mathbf{b}}_j^{(t)}|\right) + \sum_{k=1}^{q} \left.\frac{\partial p_{\boldsymbol{\theta}_2}\left(|\mathbf{b}_j|\right)}{\partial |\beta_{jk}|}\right|_{|\beta_{jk}|=|\hat{\beta}_{jk}^{(t)}|} \left(|\beta_{jk}| - |\hat{\beta}_{jk}^{(t)}|\right) = \sum_{k=1}^{q} \frac{\lambda_2 |\beta_{jk}|}{|\hat{\mathbf{b}}_j^{(t)}| + \tau_2} + C_2,$$

where C_1 and C_2 are constants with respect to β_{jk}. Then the objective function at the $(t+1)$th iteration, denoted $\tilde{Q}^{(t+1)}(\boldsymbol{\beta})$, can be written $\tilde{Q}^{(t+1)}(\boldsymbol{\beta}) = \frac{1}{2n} \sum_{k=1}^{q} \|\mathbf{y}_k - \mathbf{X}\boldsymbol{\beta}_k\|_2^2 + \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{\lambda_1 |\beta_{jk}|}{|\hat{\beta}_{jk}^{(t)}| + \tau_1} + \sum_{j=1}^{p} \sum_{k=1}^{q} \frac{\lambda_2 |\beta_{jk}|}{|\hat{\mathbf{b}}_j^{(t)}| + \tau_2}$. We use a coordinate descent approach to find each regression coefficient β_{jk}

sequentially. Let $\bar{\beta}_{jk}^{(t)} = (1/n)\sum_{i=1}^{n} x_{ij}\left(y_{ik} - \sum_{l\neq j} x_{il}\hat{\beta}_{lk}^{(t)}\right)$, to solve for β_{jk}, we minimize

$$\tilde{Q}(\beta_{jk}) = \frac{1}{2}\left(\beta_{jk} - \bar{\beta}_{jk}^{(t)}\right)^2 + \left\{\frac{\lambda_1}{|\hat{\beta}_{jk}^{(t)}| + \tau_1} + \frac{\lambda_2}{\sum_{k=1}^{q}|\hat{\beta}_{jk}^{(t)}| + \tau_2}\right\}|\beta_{jk}|. \tag{2.2}$$

This "LLA + coordinate descent" algorithm alternates through different iterations indexed by t, and within each iteration, it estimates all the regression coefficients sequentially. Finally, this algorithm is considered to have converged if the maximum difference in the coefficient estimates between consecutive iterations is less than a predefined threshold, say 10^{-4}.

The penalty term in (2.2) can be written as an adaptive Lasso (Zou, 2006) form of $\lambda_1\hat{w}_{jk}|\beta_{jk}|$ with $\hat{w}_{jk} = \left[\left(|\hat{\beta}_{jk}^{(t)}| + \tau_1\right)^{-1} + \frac{\lambda_2}{\lambda_1}\left(\sum_{k=1}^{q}|\hat{\beta}_{jk}^{(t)}| + \tau_2\right)^{-1}\right]$. In contrast to the adaptive Lasso, which adapts a weight function $1/|\hat{\beta}_{jk}^{(t)}|$, our weight function is a weighted sum of the contributions of the individual coefficient estimates $(|\hat{\beta}_{jk}^{(t)}| + \tau_1)^{-1}$ and the group-level estimates $(\sum_{k=1}^{q}|\hat{\beta}_{jk}^{(t)}| + \tau_2)^{-1}$, with weights 1 and λ_2/λ_1, respectively. Note that the inclusion of tuning parameters τ_1 and τ_2 prevents an infinite penalty for any regression coefficient with a previous estimate of 0. This is necessary for the iterative estimation procedure to proceed with one or more regression coefficients penalized to 0.

Next we show that an alternative group penalty using an L_2 norm (i.e., $\lambda_2 \log\left(\|\mathbf{b}_j\|_2 + \tau_2\right)$) is not appropriate. With the L_2 norm in group penalty, the intermediate objective function in (2.2) becomes $\tilde{Q}^{(t+1)}(\beta_{jk}) = \frac{1}{2}\left(\beta_{jk} - \bar{\beta}_{jk}^{(t)}\right)^2 + \frac{\lambda_1|\beta_{jk}|}{|\hat{\beta}_{jk}^{(t)}|+\tau_1} + \frac{\lambda_2|\hat{\beta}_{jk}^{(t)}||\beta_{jk}|}{\|\hat{\mathbf{b}}_j^{(t)}\|_2^2+\tau_2\|\hat{\mathbf{b}}_j^{(t)}\|_2}$. It is easy to show that this group-level penalty does not deliver the desirable property when $\|\hat{\mathbf{b}}_j^{(t)}\|_2 > 0$. Specifically, a nonzero coefficient β_{jk} may be penalized to zero in the tth iteration, and the group-level penalty may "rescue" it by borrowing information across $\beta_{jk'}$'s for $k' \neq k$. However, using this L_2 norm group penalty, the penalty for β_{jk} equals to 0 as long as $\hat{\beta}_{jk}^{(t)} = 0$, and thus it cannot borrow information across $\beta_{jk'}$'s.

Finally, we will discuss the role of the individual-level penalty $p_{\theta_1}\left(|\beta_{jk}|\right)$ in our method. If we remove this penalty term, the intermediate objective function of coordinate descent algorithm shown in (2.2) becomes $\tilde{Q}(\beta_{jk}) = \frac{1}{2}\left(\beta_{jk} - \bar{\beta}_{jk}^{(t)}\right)^2 + \left\{\frac{\lambda_2}{\sum_{k=1}^{q}|\hat{\beta}_{jk}^{(t)}|+\tau_2}\right\}|\beta_{jk}|$. A bilevel variable selection can still be achieved because the group information contributes to the weight term and the term $|\beta_{jk}|$ in the penalty function delivers individual penalty. In fact, Huang *and others* (2012) has pointed out that applying a folded concave penalty to the L1 norm of a group of coefficients achieves bilevel selection. However, in our method, by including $p_{\theta_1}\left(|\beta_{jk}|\right)$ in the overall objective function, the individual coefficient estimate can also contribute to the weight and thus provide more information for penalized estimation. The following simulation study shows that the performance of BipLog becomes worse without the individual-level penalty $p_{\theta_1}\left(|\beta_{jk}|\right)$.

2.3 *A Bayesian interpretation of BipLog*

The following Bayesian interpretation provides additional insight into our method and the role of the tuning parameters. Recall that $\mathbf{b}_j = (\beta_{j1}, \ldots, \beta_{jq})^T$ are the regression coefficients for the jth covariate across the q response variables. Our BipLog penalty can be derived from a Bayesian setup using the following priors: $p(\mathbf{b}_j|\omega_{j1}, \ldots, \omega_{jq}, \omega_j) = \left\{\prod_{k=1}^{q}\frac{1}{2}(\omega_{jk}^{-1} + \omega_j^{-1})\exp\left(-\frac{|\beta_{jk}|}{\omega_{jk}}\right)\right\}\exp\left(-\frac{\sum_{k=1}^{q}|\beta_{jk}|}{\omega_j}\right)$; $p(\omega_{jk}|\delta_1, \tau_1) =$ inv-Gamma$(\omega_{jk}; \delta_1, \tau_1)$; $p(\omega_j|\delta_2, \tau_2) =$ inv-Gamma$(\omega_j; \delta_2, \tau_2)$, where $\delta_1 > 0$, $\delta_2 > 0$, $\tau_1 > 0$, and $\tau_2 > 0$ are four hyperparameters. Given the above specification, after integrating out ω_{jk} and ω_j, we obtain the

density of $\mathbf{b}_j$: $f(\mathbf{b}_j|\delta_1,\tau_1,\delta_2,\tau_2) \propto \frac{\tau_2^{\delta_2}\delta_2}{2(\sum_{k=1}^q|\beta_{jk}|+\tau_2)^{1+\delta_2}}\prod_{k=1}^q\frac{\tau_1^{\delta_1}\delta_1}{2(|\beta_{jk}|+\tau_1)^{1+\delta_1}}$, and $-\log\{f(\mathbf{b}_j|\delta_1,\tau_1,\delta_2,\tau_2)\}$ gives exactly the same form of the BipLog penalty as in (2.1) if we set $n\lambda_1 = 1+\delta_1$ and $n\lambda_2 = 1+\delta_2$. This also gives more insight into the scale of the tuning parameters of λ_1 and λ_2. Empirically, the grids of possible values of λ_1 and λ_2 could be set at the scale of n^{-1} since both δ_1 and δ_2 are constant $O(1)$.

2.4 *Tuning parameter selection*

We choose the best set of tuning parameters by a grid search over an initial pool of tuning values. On the basis of the Bayesian interpretation of BipLog, we set $\lambda_1 = (1+\delta_1)/n$ and $\lambda_2 = (1+\delta_2)/n$. The initial values for δ_1 and δ_2 range from 0 to 5.0 with a 0.5 increment. The initial values for τ_1 and τ_2 are from 10^{-3} to two times of $\max_{j,k}\{|\hat\beta_{jk}^{\mathrm{mls}}|\}$ and two times of $\max_j\{\sum_{k=1}^q|\hat\beta_{jk}^{\mathrm{mls}}|\}$, respectively, where $\hat\beta_{jk}^{\mathrm{mls}}$ denotes marginal regression coefficient estimate for the kth response and the jth covariate. The rationale of this choice is similar for both τ_1 and τ_2. Use τ_1 as an example. If τ_1 is too small compared to the smallest nonzero value of $|\beta_{jk}|$, p_{θ_1} is mainly depend on $|\beta_{jk}|$. If τ_1 is too large, the effect of $|\beta_{jk}|$ on the penalty becomes negligible. We use the marginal regression coefficients as data-driven quantities to generate the approximate range of τ_1 and τ_2. Our simulation studies show that this solution works well in practice. We select a combination of tuning parameters using the extended BIC (Chen and Chen, 2008). The extended BIC for a model m is $\mathtt{BIC}_\varrho(m) = -2\log l_n\{\hat{\boldsymbol\theta}(m)\} + \kappa_m\log n + 2\varrho\log\varsigma(S_{\kappa_m})$, where $l_n\{\hat{\boldsymbol\theta}(m)\}$ is the log likelihood, $\hat{\boldsymbol\theta}(m)$ are the estimates of all the parameters, κ_m is the degree of freedom for model m, and $\varsigma(S_{\kappa_m})$ is the number of models with degree of freedom equal to κ_m. Specifically, $l_n\{\hat{\boldsymbol\theta}(m)\}$ is calculated using the penalized coefficient estimates assuming a multivariate Gaussian distribution. We set κ_m to be the number of nonzero coefficients and $\varsigma(S_{\kappa_m}) = \binom{pq}{\kappa_m}$, i.e., the number of choices of κ_m coefficients from a total of pq regression coefficients. In addition, following Chen and Chen (2008), we set $\varrho \approx 1 - 1/[2\log(pq)/\log n]$.

3. Simulation studies

3.1 *Simulation setup*

It is difficult to simulate high-dimensional genomic data with a realistic correlation structure except in a few special cases, for example, genetic data from experimental cross. We create the simulation setting as in Sun *and others* (2010). Using the function `sim.map` in R/qtl (Broman *and others*, 2003), we first simulated a genetic marker map of 2000 single-nucleotide polyneorphines (SNPs) from 20 chromosomes of length 90 cM, with 100 SNPs per chromosome. Then we used the function `sim.cross` in R/qtl to simulate the genotype data of an F2 cross with sample size $n = 200$ based on the simulated marker map. As expected, the simulated genotypes show strong correlations for nearby SNPs (average R^2 is 0.96 for SNPs within 1 cM) and negligible correlation for SNPs from different chromosomes. The following real data analysis will consider the data sets including p to be more than 13000 and about 500 samples. To save the computation time, we randomly selected $p = 600$ SNPs from the 2000 SNPs for the following simulation of quantitative traits.

We simulated a total of $q = 30$ quantitative traits from the multivariate linear model $\mathbf{Y}_{n\times q} = \mathbf{X}_{n\times p}\boldsymbol\beta_{p\times q} + \mathbf{E}_{n\times q}$, where $\mathbf{Y} = (\mathbf{y}_1,\ldots,\mathbf{y}_q)$. The residuals $\mathbf{E}$ were simulated from a multivariate Gaussian distribution with mean 0 and compound symmetry covariance structure with diagonal variance $(0.25 + 0.5)$ and off-diagonal covariance 0.5. Traits 1–10 share a pair of causal SNPs, and each has its own causal SNP. Traits 11–30 do not have individual causal SNPs. Traits 11–20 share two pairs of causal SNPs, and traits 21–30 share one pair of causal SNPs. The pairs of causal SNPs shared across traits may be located in different chromosomes (unlinked) or at the same chromosome with the effect sizes being $(\eta,\ \eta)$ (SNPs linked in coupling) or $(\eta, -\eta)$ (SNPs linked in repulsion). We consider two effect sizes $\eta = 0.3$ or 0.6. Given the

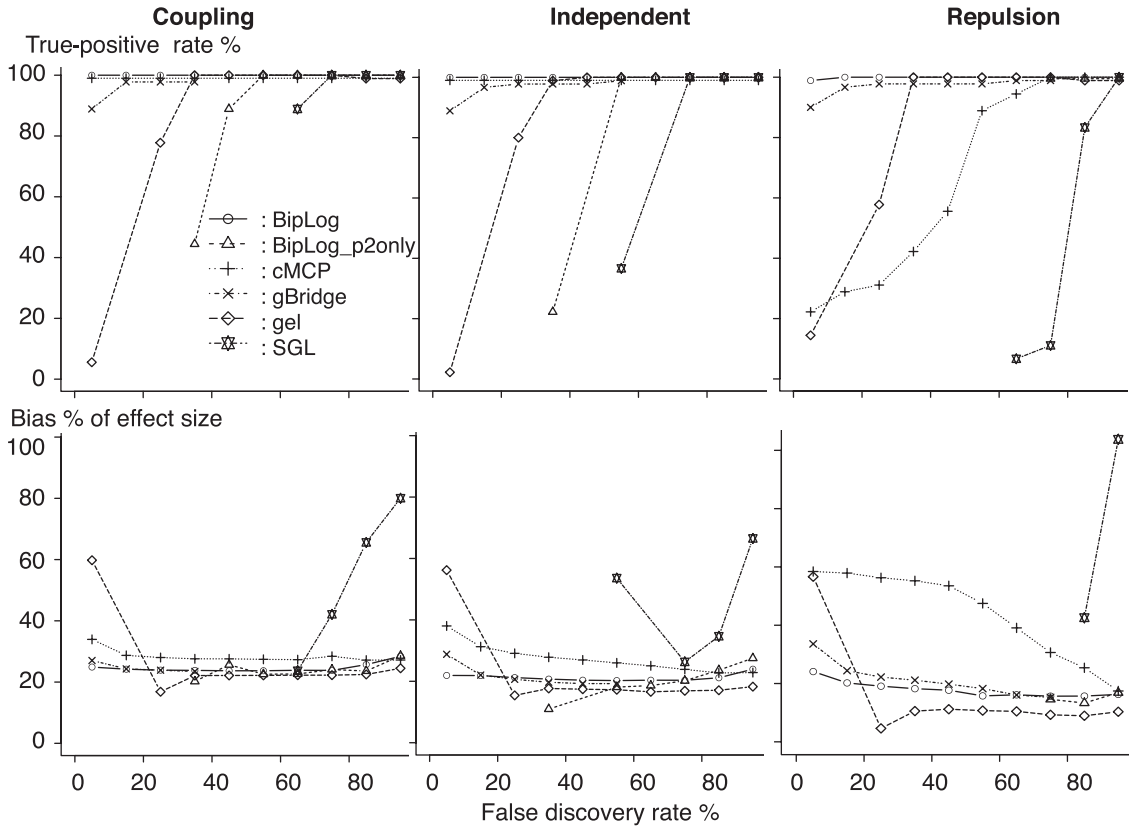

Fig. 1. Comparisons of six bilevel selection methods (SGL, gBridge, cMCP, gel), BipLog-p2only and BipLog) via simulation studies. For each of the 3 simulation scenarios with effect size $\eta = 0.6$, 30 traits are considered. The total number of true trait-SNP associations is 90, which includes 10 associations due to SNPs affecting only one trait (individual SNPs) and 80 associations due to SNP pairs shared across traits (shared SNPs).

three relationships between the causal SNP pairs and the two effect sizes, there are six simulation scenarios in total.

We compared BipLog with five other approaches: SGL (Simon *and others*, 2013), gel (Breheny, 2015), gBridge, cMCP (Huang *and others*, 2012), and BipLog with only group penalty p_{θ_2}. We used the implementation of gel, gbridge and cMCP in R/grpreg, with the default choice of 1000 tuning values for the tuning parameter λ, which were uniformly distributed on a log scale. In addition, we used the implementation of SGL in R/SGL with the default choice of 20 values of the tuning parameter. Because the R package R/SGL can only take univariate instead of multivariate response variable, we collapse $\mathbf{Y}_{n\times q}$ into a vector of length nq: $\tilde{\mathbf{y}} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_q^T)^T$ and generate a new covariate data matrix of dimension $nq \times pq$ by creating a block diagonal matrix $\tilde{\mathbf{X}} = \text{diag}\{\mathbf{X}, \ldots, \mathbf{X}\}$ with q identical matrix $\mathbf{X}_{n\times p}$. This approach allows us to analyze multivariate data by a univariate regression of $\tilde{\mathbf{y}}$ vs. $\tilde{\mathbf{X}}$, but it increases the dimensionality of the covariate matrix and hence the computation time. Using the default number of 20 values of tuning parameter generated by the R package R/SGL takes about 5 h to finish one simulation, and thus we did not increase the number of tuning values. Similar transformation of data is needed to apply group variable selection via convex log-exp-sum penalty (Chen *and others*, 2014), which takes much longer computational time, and thus we did not evaluate the results of this approach.

We compare the performance of these methods across the range of tuning parameters by ROC-like curves. Specifically, instead of comparing true-positive rate (i.e., sensitivity) vs. false-positive rate (i.e., 1 - specificity) in regular ROC curves, we compare true-positive rate vs. false discovery rate (FDR). In high-dimensional settings, FDR is a more appropriate measure of accuracy than specificity. Let s be the number of causal SNPs, and let D be the number of discoveries, i.e., the number of nonzero regression coefficient estimates. $D = \texttt{TD} + \texttt{FD}$, where TD and FD are the number of true and false discoveries. Then FDR = $\texttt{FD}/D$ and true-positive rate = $\texttt{TD}/s$. In addition, we also considered the average estimation bias of nonzero effect sizes as the average of $|\hat{\beta}_{jk} - \beta_{jk}|/|\beta_{jk}| \times 100\%$ for any $\beta_{jk} \neq 0$, where β_{jk}'s are the true effect sizes. Small bias is crucial for the success of tuning parameter selection using either BIC or cross-validation because both criterion rely on model fit, which in turn relies on unbiased estimates of effect sizes. We considered a 10% interval for the FDR from 0% to 100% and for all the models within the same FDR interval, we select the highest true-positive rate and the lowest estimation bias.

Figure 1 shows the median of those measurements across 100 simulations for each of the methods to be compared: BipLog, BipLog-p2only (BipLog with only group penalty p_{θ_2}), gBridge, cMCP, gel, and SGL when $\eta = 0.6$. In the "coupling" and "independent" setting, BipLog and cMCP have best performance in terms of FDR and sensitivity, and the estimates from BipLog have smaller bias than those from cMCP. In the "repulsion" setting, BipLog has much better performance than cMCP in terms of variable selection or bias. gBridge also has good performance (though slightly worse than BipLog) in all three settings. BipLog with only group penalty (BipLog_p2only) has poor performance in all the three settings. The results for $\eta = 0.3$ reach similar conclusions except that cMCP shows worse performance in the independent setting and gBridge shows better performance in the repulsion setting (see Figure 1 of supplementary material available at *Biostatistics* online).

4. Omic signatures of cancer drug sensitivity

To identify the omic features associated with the cancer drugs' sensitivity, both Garnett *and others* (2012) and Barretina *and others* (2012) generated omic and pharmacological data for a panel of human cancer cell lines, which represent the characteristics of various types of cancers. Garnett *and others* (2012) measured the mutation statuses of 64 commonly mutated cancer genes (exon sequencing), genome-wide copy number alterations (Affymetrix SNP array 6.0), and gene expression (Affymetrix HT-U133A microarray), while Barretina *and others* (2012) measured the mutation statuses of 1600 genes (targeted sequencing), genome-wide copy number alterations (Affymetrix SNP array 6.0), and gene expression (Affymetrix U133 plus 2.0 array). In the study of Garnett *and others* (2012), 130 drugs were screened for drug sensitivity analysis in a range of 275–507 cell lines from a panel of 639 human tumor cell lines. In the study of Barretina *and others* (2012), 24 drugs were screened for drug sensitivity analysis for 500 cell lines on average from a panel of 947 human tumor cell lines. In both studies, the drug sensitivity was assessed by IC_{50}, which is half-maximal inhibitory concentration. All the data used in this section were downloaded from http://www.broadinstitute.org/ccle/data/browseData and http://www.cancerrxgene.org/downloads/.

4.1 *Evaluation of prediction model in the study of Garnett and others (2012) by cross-validation*

Of the 130 drugs analyzed by Garnett *and others* (2012), 41 have nonmissing IC_{50} values in fewer than 331 cell lines, while the other 89 drugs have nonmissing IC_{50} values in more than 461 cell lines. These 89 drugs were grouped by their targets, and two drugs were excluded from our analysis because they do not group with any other drugs. We will first study these 87 drugs since a larger sample size is necessary for the following analyses using training/testing sets. Of the 87 drugs, 57, 69, and 56 are grouped by targeted family, targeted process, and targeted molecule, respectively. The three grouping strategies have a semihierarchical order: targeted family > targeted process > targeted molecule. One drug is

often grouped in multiple ways. There are four targeted families: chemotherapy, cytoplasmic/nonreceptor tyrosine kinase (CTK), receptor tyrosine kinase (RTK), and serine/threonine protein kinase (S/T Kinase), which include 12, 7, 10, and 30 drugs, respectively. There are 18 targeted processes (typically with less than 10 drugs per process) and 24 targeted molecules groups (typically with less than 10 drugs per molecule).

We randomly selected 65 cell lines from those with nonmissing IC_{50} values for all 87 drugs as testing set and used the remaining cell lines as the training set. The training set was used for feature selection, and the tuning parameter was selected by the extended BIC. If a drug belonged to more than one group, we took the union of the omic features associated with that drug across the groups. To obtain the regression coefficients estimates for the union of the omic features associated with each drug, we re-estimated the regression coefficients for each drug separately using the training data (denoted $\hat{\boldsymbol{\beta}}_{\text{train}}$) by linear regression. Thus a predictive model for each drug was obtained. Next, we used the testing data to estimate the percentage of the variance explained by the predictive model. Let SS_z be the sum of squares of z, and let $\mathbf{y}_{\text{test}}$ and $\mathbf{X}_{\text{test}}$ be the standardized $\log(IC_{50})$ and omic features in the testing set. Then Prediction R-square $\equiv 1 - \frac{SS\boldsymbol{\epsilon}_{\text{test}}}{SS\mathbf{y}_{\text{test}}}$, where $\boldsymbol{\epsilon}_{\text{test}} = \mathbf{y}_{\text{test}} - \mathbf{X}_{\text{test}}\hat{\boldsymbol{\beta}}_{train}$. The possible range for prediction R-square is $(-\infty, 1]$. To evaluate the significance of prediction R-square for a drug with k associated features, we randomly chose k features from the candidate 13 847 omic features to estimate prediction R-square and repeated it for 1000 times to generate a null distribution. Then we calculated a p value as the percentage of the null prediction R-squares $\geq$ the observed one. It took 8 s on average to generate the p value for each drug.

BipLog identified that 70 of the 87 drugs were associated with at least one omic feature (Figures 2(a) and (b)). Forty-nine drugs had prediction R-squares greater than 0, and 17 had values greater than 20%. Forty-one of the drugs had significant prediction R-squares at the 0.05 significance level (Figure 2c), which corresponds to an estimate of FDR = $87 \times 0.05/41 \approx 0.1$. Overall, these results suggest that the identified omic features could provide useful predictions of drug sensitivity.

4.2 *Construction of prediction model using all the data from Garnett and others (2012)*

Next, we combined the training and testing sets and selected the omic features using all the available data for the 87 drugs. A few examples were shown in Figure 3, and the complete results can be found in the Supplementary materials available at *Biostatistics* online. The abnormal gene BCR-ABL is formed by the fusion of genes BCR and ABL, and it is often observed in chronic myeloid leukemia (CML). BCR-ABL encodes a tyrosine kinase that is not regulated by cellular signals and thus causes unregulated cell proliferation, which may lead to cancer. Three drugs that target BCR-ABL protein products are included in this study (Figure 3A). The sensitivity of two of these drugs is negatively correlated with the occurrence of the BCR-ABL mutation, which is expected. The negative correlation indicates that the presence of the BCR-ABL mutation is related to a reduction in $\log(IC_{50})$, hence an increase in the drug sensitivity. There are two interesting new findings in this example. (i) The sensitivity of AP-24534 also increases as the expression of AZU1 increases, which is consistent with the tumor suppressor role of AZU1. (ii) The sensitivity of Bosutinib is associated with the expression of two cancer-related genes, EGFR and CAV2, instead of the BCR-ABL mutation. EGFR is a signaling protein that plays an important role in many types of cancer, and CAV2 is potentially a tumor suppressor (Lee *and others*, 2011).

Figure 3B shows that when the gene encoding PHLDA1 has a higher expression, all four drugs that target MEK1/MEK2 have higher sensitivity. Several previous studies have suggested that PHLDA1 may be functionally important in cancer, and some studies have shown that it functions in the MEK1/MEK2 pathway (Oberst *and others*, 2008). This finding suggests that the expression of PHLDA1 could be an informative biomarker for the efficacy of cancer drugs targeting MEK1/MEK2.

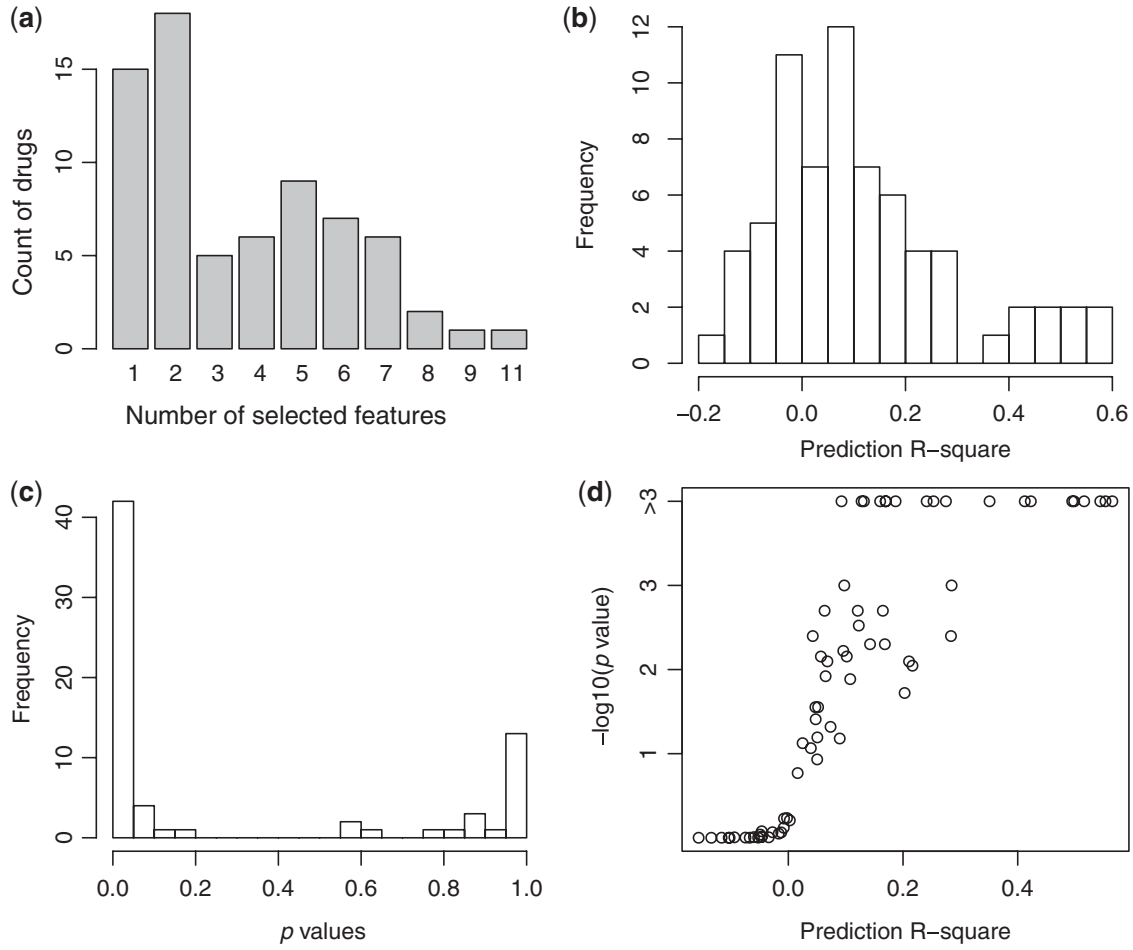

Fig. 2. Summary of the genomic feature selection results of the within-study analysis. (a) Distribution of the number of omic features selected per drug. (b) Distribution of the prediction R-squares for each drug. (c) Distribution of the p values of the prediction R-squares. (d) Scatter plot of the prediction R-squares and their corresponding p values. Since the null distribution was simulated as 1000 null samples, the smallest p value is 0.001.

The ERBB2 gene encodes a protein product that promotes the growth of cancer cells. Our analysis identifies several genes related to the two drugs targeting ERBB2 (Figure 3C): BIBW2992 and Lapatinib. BIBW2992 has been approved for use against nonsmall cell lung carcinoma, and its efficacy for breast cancer treatment is being evaluated. Lapatinib has been approved for treatment in advanced ERBB2-receptor-positive breast cancer patients. As expected, we identified the ERBB2 mutation or the copy number variation as omic features associated with these drugs. The novel findings are the association with the gene expressions of C1ORF116, CYR61, and STAM2. C1ORF116 interacts with SMD2, which is closely related to tumorigenesis. Several studies have shown that CYR61 is involved with breast cancer tumorigenesis and progression. In addition, STAM2 may be involved in "signaling by EGFR in cancer" (Croft *and others*, 2011). Therefore, the combined information from the ERBB2 mutation (or copy number alterations) and gene expression of C1ORF116, CYR61, and STAM2 may provide a more accurate prediction of drug efficacy than the ERBB2 mutation/copy number alteration alone.

A. BCR_ABL

	AP-24534	Nilotinib	Bosutinib
EGFR	0.00	0.00	–0.19
AZU1	–0.20	0.00	0.00
CAV2	0.00	0.00	–0.19
BCR_ABL_MUT	–0.39	–0.67	0.00

B. MEK1/MEK2

	RDEA119	CI-1040	PD-0325901	AZD6244
PHLDA1	–0.49	–0.39	–0.43	–0.36

C. ERBB2

	Lapatinib	BIBW2992
C1ORF116	–0.31	–0.36
CYR61	–0.35	0.00
ERBB2_CN	0.00	–0.26
ERBB2_MUT	–0.31	0.00
STAM2	0.00	–0.16

D. Mitosis

	Vinorelbine	EpothiloneB	Vinblastine	Docetaxel	BX-795	SL0101-1	BI-D1870	ZM-447439	RO-3306
ABCB1	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
YAP1	0.00	0.00	–0.25	–0.24	0.00	0.00	–0.10	0.00	0.00
AXL	0.00	–0.13	0.00	–0.17	–0.36	0.00	0.00	–0.19	–0.15
blood	0.00	0.00	0.00	0.28	0.18	0.00	0.33	0.21	0.15

Fig. 3. Omic features associated with four groups of drugs that share the molecular targets BCR_ABL (A), MEK1/MEK2 (B), ERBB2 (C), or the process target Mitosis (D). For each group, the regression coefficient matrix is shown for those omic features with at least one nonzero coefficient, where a row corresponds to a genomic feature and a column corresponds to a drug. The feature X_MUT is a binary indicator showing whether gene X has mutation; ERBB2_CN is the copy number of the gene ERBB2; blood is a binary indicator showing whether the cell line is derived from a blood tumor. The remaining features are gene expressions.

Figure 3D presents the estimated coefficient matrix for nine drugs that target the Mitosis process. The features shared by several drugs include the expression of genes YAP1 and AXL and the blood-tissue indicator. YAP1 encodes "YES-associated protein 1," which has been shown to be related to different types of cancer. AXL encodes a receptor tyrosine kinase, which is also involved with tumorigenesis. Previous studies have shown that the protein products of YAP1 and AXL may function together (Cui *and others*, 2011).

4.3 *Validation of the prediction model on the data from Barretina and others (2012)*

We treated the data of Garnett *and others* (2012) and Barretina *and others* (2012) as the training and testing study data, respectively. Of the 24 drugs analyzed by Barretina *and others* (2012), 12 were analyzed in the training study. Five of the 12 drugs had missing values in more than 325 cell lines in the training study, so they were not included in the 87 drugs in the above analysis. To address this issue, we conducted another group-wise analysis using the training data for groups involving any of these 12 drugs. Then we chose the features associated with each drug as the union of the features selected in this new analysis and those from the above analysis, whenever possible. For the 12 drugs that were analyzed using the testing data but not the training data, we fitted the prediction models using the features selected for their drug targets. For example, for the drug Topotecan, which targets the molecule TOP1, we used the features from the training study associated with the drug group that targeted TOP1 as the features associated with Topotecan.

To determine whether at least one of the selected features is associated with drug sensitivity in the testing data, we used an F test to compare the intercept-only model and the model with all the identified omic features (Figure 4). The p values are smaller than 0.05 in most cases. The drugs PLX4720 and Lapatinib are particularly significant, with p values of 10^{-39} and 10^{-17}, respectively.

To compare the omic features identified by our method and the elastic net in the study of Garnett *and others* (2012), we calculated the prediction R-square of $\log(IC_{50})$ in the testing study for the 12 drugs

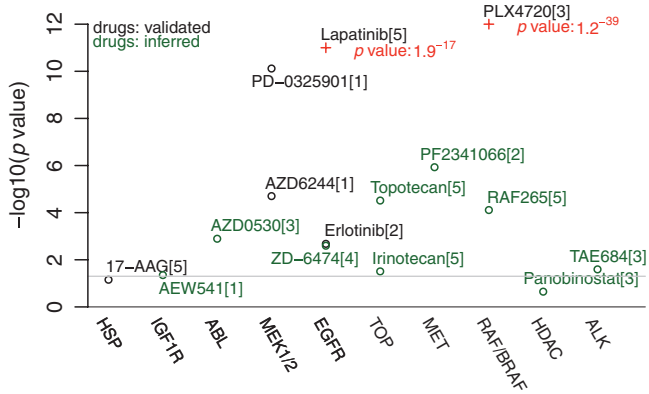

Fig. 4. Evaluation of the predictive model in the study of Barretina *and others* (2012); the models themselves were constructed using the data of Garnett *and others* (2012). The "validated drugs" are the drugs that were analyzed in both studies. The inferred drugs are the drugs that were analyzed only in the study of Barretina *and others* (2012). The x-axis shows the drug targets, and the y-axis shows the $-\log 10(p$ values) from the F test that compares the model with all the identified omic features to the intercept-only model using the data of Barretina *and others* (2012). The numbers in brackets are the number of features in the corresponding prediction model.

Table 1. *Prediction R-squares in the study of Barretina* and others *(2012).*

Drug	17-AAG	AZD6244	PD-0325901	PLX4720	Erlotinib	Lapatinib
			Groupwise analysis by BipLog			
Prediction R^2 [Num of **X**]	15% [5]	2.5% [1]	7.3% [1]	27% [3]	$<$ 0.0% [2]	21% [5]
p value	$<$ 0.001	$<$ 0.001	$<$ 0.001	$<$ 0.001	1.00	$<$ 0.001
			Drug-by-drug analysis by elastic net			
Prediction R^2 [Num of **X**]	15% [16]	10% [7]	29% [17]	29% [5]	1.5% [7]	14% [16]
p value	$<$ 0.001	0.275	$<$ 0.001	0.525	0.949	0.430

analyzed in both the training and testing studies. Because of the disparate ranges and scales of $\log(IC_{50})$ in the two studies, as shown in Figure 5 of supplementary material available at *Biostatistics* online, we could not directly use the regression coefficients estimated from training study. Instead, for each drug, we randomly split the cell lines in the testing data into two groups of equal size as set 1 and set 2. We used set 1 to estimate the linear regression coefficients of the features identified by the training data. Then we used set 2 to estimate the prediction R-square. We repeated this procedure 100 times to obtain median prediction R-squares as our final estimates. Then we applied Monte Carlo simulations to evaluate the significance of the prediction R-squares, similar to our approach in Section 4.1.

Of the 12 drugs, six had a prediction R-square greater than 0 using a set of features selected by our method or the elastic net analysis in the training study. The results for these six drugs are presented in Table 1. In general, BipLog tended to choose more parsimonious models than those chosen by the elastic net, and the estimates of the prediction R-square were statistically significant in all but one case. Some of the models selected by the elastic net had greater prediction R-squares than those from BipLog. However, because more variables were included in the model, they were not significantly larger than what was expected from the null distributions.

5. CONCLUSION AND DISCUSSION

We have presented a new method, BipLog, for the bilevel selection of omic features related to cancer drug sensitivity. BipLog can select the covariates shared by a group of response variables as well as the covariates that are associated with one or a few of the response variables. The application of BipLog to real-data analysis reveals many interesting results. This is partly due to the strong effect size in the data. In contrast to genome-wide association studies where a genetic variant may explain only a few percentage of the variation in the trait of interest, the omic features measured in tumor tissues have a strong influence on the cancer progression and its response to drug treatment.

We have implemented our method in an R package with C code for computationally intensive part of the computation. Although using four tuning parameters does incur higher computational load, our method is computationally feasible for the problems we aim to solve. For example, in the simulation setting with $n = 200$, $p = 562$, $q = 30$, and with a total 6000 combinations of the four tuning parameters, it takes about 30 min to finish the computation. Considering an example in our real application, it takes about 1 h to finish computation for a data set with $n = 462$, $p = 13,062$, and $q = 8$. We have also evaluated how the values of the four tuning parameters influence the performances of BipLog using the same data set used in our simulation studies with $\eta = 0.6$ (See Figures 2–4 of supplementary material available at *Biostatistics* online). Our results suggest the performance of our method does not change much with respect to the choice of δ_1 and δ_2 (note that $\lambda_1 = (1 + \delta_1)/n$ and $\lambda_2 = (1 + \delta_2)/n$), and thus it is possible to set them as fixed values such as 3.5 or 4.0.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGEMENTS

Conflict of Interest: None declared.

FUNDING

National Institutes of Health (R03 CA167684-01 and R01 GM105785).

REFERENCES

BARRETINA, J., CAPONIGRO, G., STRANSKY, N., VENKATESAN, K., MARGOLIN, A. A., KIM, S., WILSON, C. J., LEHÁR, J., KRYUKOV, G. V., SONKIN, D. AND *others*. (2012). The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607.

BREHENY, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics* **71**, 731–740.

BREHENY, P. AND HUANG, J. (2009). Penalized methods for bi-level variable selection. *Statistics and its Interface* **2**, 369–380.

BROMAN, K. W, WU, H., SEN, Ś. AND CHURCHILL, G. A. (2003). R/qtl: Qtl mapping in experimental crosses. *Bioinformatics* **19**, 889–890.

CAPONIGRO, G. AND SELLERS, W. R. (2011). Advances in the preclinical testing of cancer therapeutic hypotheses. *Nature Reviews Drug Discovery* **10**, 179–187.

CHEN, J. AND CHEN, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika* **95**, 759–771.

CHEN, T.H., SUN, W. AND FINE, J.P. (2015). Designing penalty functions in high dimensional problems: the role of tuning parameters. *Technical Report*. UNC Chapel Hill.

CHEN, Y., DU, P. AND WANG, Y. (2014). Variable selection in linear models. *Computational Statistics* **6**, 1–9.

CROFT, D., O'KELLY, G., WU, G., HAW, R., GILLESPIE, M., MATTHEWS, L., CAUDY, M., GARAPATI, P., GOPINATH, G., JASSAL, B. AND *others*. (2011). Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* **39**(suppl 1), D691–D697.

CUI, ZL, HAN, FF, PENG, XH, CHEN, X, LUAN, CY, HAN, RC, XU, WG, GUO, XJ AND *others*. (2011). YES-associated protein 1 promotes adenocarcinoma growth and metastasis through activation of the receptor tyrosine kinase axl. *International Journal of Immunopathology and Pharmacology* **25**, 989–1001.

de PALMA, M. AND HANAHAN, D. (2012). The biology of personalized cancer medicine: facing individual complexities underlying hallmark capabilities. *Molecular Oncology* **6**, 111–127.

FAN, J. AND LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101–148.

FRIEDMAN, J.H. (2012). Fast sparse regression and classification. *International Journal of Forecasting* **28**, 722–738.

GARNETT, M.J., EDELMAN, E.J., HEIDORN, S.J., GREENMAN, C.D., DASTUR, A., LAU, K.W., GRENINGER, P., THOMPSON, I.R., LUO, X., SOARES, J. AND *others*. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature* **483**, 570–575.

HUANG, J., BREHENY, P. AND MA, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science* **27**, 481–499.

HUANG, J., MA, S., XIE, H. AND ZHANG, C.H. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339–355.

LEE, S., KWON, H., JEONG, K., PAK, Y. AND *others*. (2011). Regulation of cancer cell proliferation by caveolin-2 down-regulation and re-expression. *International Journal of Oncology* **38**, 1395.

MAZUMDER, RAHUL, FRIEDMAN, JEROME H AND HASTIE, TREVOR. (2011). Sparsenet: coordinate descent with nonconvex penalties. *Journal of the American Statistical Association* **106**, 1125–1138.

OBERST, MICHAEL D, BEBERMAN, STACEY J, ZHAO, LIU, YIN, JUAN J, WARD, YVONA AND KELLY, KATHLEEN. (2008). Tdag51 is an erk signaling target that opposes erk-mediated hme16c mammary epithelial cell transformation. *BMC Cancer* **8**, 189.

SIMON, N., FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22**, 231–245.

SUN, W., IBRAHIM, J.G. AND ZOU, F. (2010). Genomewide multiple-loci mapping in experimental crosses by iterative adaptive penalized regression. *Genetics* **185**, 349–359.

WANG, H. AND LENG, C. (2008). A note on adaptive group lasso. *Computational Statistics & Data Analysis* **52**, 5277–5286.

YUAN, M. AND LIN, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**, 49–67.

ZHAO, P. AND YU, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research* **7**, 2541–2563.

ZOU, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

ZOU, H. AND HASTIE, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**, 301–320.

ZOU, H. AND LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Annals of Statistics* **36**, 1509–1533.

[Received February 08, 2015; revised January 07, 2016; accepted for publication April 12, 2016]