Sparse expression bases in cancer reveal tumor drivers

Benjamin A. Logsdon1,2, Andrew J. Gentles3, Chris P. Miller4, C. Anthony Blau4, Pamela S. Becker4 and Su-In Lee1,5,*

1Department of Genome Sciences, University of Washington, Seattle, WA, 98195, USA, 2Sage Bionetworks, Seattle, WA, 98109, USA, 3Center for Cancer Systems Biology, Department of Radiology, Stanford University, CA, 94305, USA, 4Department of Medicine/Hematology, Center for Cancer Innovation, University of Washington, Seattle, WA, 98195, USA and 5Department of Computer Science & Engineering, University of Washington, Seattle, WA, 98195, USA

Received August 28, 2014; Revised November 10, 2014; Accepted November 25, 2014

ABSTRACT

We define a new category of candidate tumor drivers in cancer genome evolution: ‘selected expression regulators’ (SERs)—genes driving dysregulated transcriptional programs in cancer evolution. The SERs are identified from genome-wide tumor expression data with a novel method, namely SPARROW (SPARse selected expRession regulators identified With penalized regression). SPARROW uncovers a previously unknown connection between cancer expression variation and driver events, by using a novel sparse regression technique. Our results indicate that SPARROW is a powerful complementary approach to identify candidate genes containing driver events that are hard to detect from sequence data, due to a large number of passenger mutations and lack of comprehensive sequence information from a sufficiently large number of samples. SERs identified by SPARROW reveal known driver mutations in multiple human cancers, along with known cancer-associated processes and survival-associated genes, better than popular methods for inferring gene expression networks. We demonstrate that when applied to acute myeloid leukemia expression data, SPARROW identifies an apoptotic biomarker (PYCARD) for an investigational drug obatoclax. The PYCARD and obatoclax association is validated in 30 AML patient samples.

INTRODUCTION

Cancer genome evolution is characterized by the emergence of driver mutational events, successful mutations that are swept to high frequencies within the tumor cell population (1). Selection often favors driver events that alter expression levels of entire pathways, such as inhibition of apoptosis (2), increased cell proliferation (3), acquisition of stem-cell phenotypes (4), evolved resistance to therapy (5) and adaptation to local microenvironments (6), among others. Tumors that evolve under similar selective pressures in different patients can experience similar changes in transcriptomic output (7). These selective events induce across patient dependencies between recurrent driver events shared across patients and selected changes in gene expression (Figure 1a). One can classify driver events into a variety of molecular aberrations, including mutation drivers (m-drivers), copy number aberrant drivers (c-drivers), fusion drivers (f-drivers), epigenetic drivers (epi-drivers), among others (8). We define selected expression regulators (SERs) as genes with an underlying molecular aberration which affect transcriptomic activity, where the change in transcription has a marked effect on the fitness of the tumor cell population. However, shared expression profiles across patients do not indicate which gene’s expression is driving the cancer’s transcriptomic state because the expression of the genes in a pathway and their driver gene(s) may all be mutually correlated (Figure 1a). We propose a method to deconvolve this mutual correlation and prioritize the genes most likely to contain the underlying driver events. To do this, we take advantage of a statistical property shared by SERs as shown in Figure 1a. If the expression of a gene is driving the expression of other genes under selection, then the correlation among those genes should be explained away once one conditions on the expression of the true driver gene. These genes under selection become ‘conditionally independent’ with respect to one another, given the SER (Figure 1a).

In this paper, we present a novel computational method, called SPARROW (SPARse selected expRession regulators identified With penalized regression), to identify these SERs based purely on expression data from multiple tumor samples. Cancer genomes contain multiple pathways under

*To whom correspondence should be addressed. Tel: +1 206 685 1418; Fax: +1 206 543 2969; Email: suinlee@cs.washington.edu

© The Author(s) 2015. Published by Oxford University Press on behalf of Nucleic Acids Research.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com
Figure 1. (a) Driver mutational event in a subset of patients caused by selection, with a selected expression regulator (SER) mediating the effect of the driver mutant. The SER is correlated with genes under similar selection across patients. After conditioning on the SER, the correlation among the genes disappears. (b) There are many driver events across patients that target different pathways, depending on the selection pressure in different patients. The goal is to identify which SER best explains these patterns of expression.

selection across patients with many possible driver events (Figure 1b). To address this, we model the multifactorial nature of the transcriptomic response to selection. We assume that driver events that lead to similar outcomes in transcriptomic response for a subset of patients can be compactly represented in terms of a combined effect of a small subset of SERs (Figure 2). To identify these SERs, we use a sparse statistical model in which each gene’s expression level is modeled as a linear combination of a small set of SERs. Due to high dimensionality, we consider a set of SERs consisting of expression levels of ~3000 genes that can regulate other genes, including transcription factors, chromatin remodelers and signal transduction genes (Figure 2a). Even if the driver event is not in the same genomic location, if it has a similar effect on the expression of the SER(s) then the sparse basis will capture its transcriptomic effect. We fit a sparse basis for every gene with the SERs and determine how often an SER is chosen in the sparse basis for each gene (Figure 2b). If an SER is chosen for many genes, this increases our confidence that this SER contains (or is closely linked to) a driver event, such as a DNA mutation or epigenetic modification that is under selective pressure.

To fit these basis models, SPARROW uses a novel sparse regression technique. The most widely used sparse regression method is the least absolute shrinkage and selection operator (LASSO) (9). A number of authors applied LASSO to gene expression data to identify association among gene expression levels, genetic variation or complex traits (10–13). SPARROW uses a non-convex penalty with important advantages when compared to the popular convex LASSO penalty for analysis of cancer expression data. Specifically the LASSO penalty is known to choose false positives if there exist pathological correlations among variables (the irrepresentability condition (14)), whereas non-convex penalties can produce cleaner solutions, even in the presence of pathological correlations (15). We compare SPARROW with the LASSO and other methods that identify dependencies among expression variables, including two methods based on correlation, and popular network building approaches such as the algorithm for the reconstruction of accurate cellular networks (ARACNE (16)) and weighted graph correlation network analysis (WGCNA) algorithms (17).

We show that SPARROW identifies SERs with known tumor-specific cancer driver mutations in acute myeloid leukemia (AML) and glioblastoma multiforme (GBM), as compared to alternative methods to identify SERs. At-
ttempts to identify driver genes have been extensively focused on analyzing DNA sequence data (18–22). These approaches have several major limitations: (i) there are a large number of passenger mutations; (ii) driver events may be caused by epigenetic modifications, which are not captured by DNA sequencing; and (iii) due to a large number of passenger mutations, large sample sizes are required to have sufficient power to detect driver events. SPARROW provides a complementary criterion for revealing tumor regulators under selection, which takes downstream effects of drivers on genome-wide expression into account. We also show in AML that SERs are enriched for association with survival, apoptotic regulation and myeloid differentiation regulators. Finally, we determine that the gene \textit{PYCARD} is the most highly ranked apoptotic SER in AML, and show that \textit{PYCARD} expression is predictive of response to obatoclax, a pan-inhibitor of antiapoptotic \textit{BCL2} family, in tissue cultured from patients.

The resulting SERs for each gene in AML and GBM, the implementation of SPARROW and the data used in the study are freely available on our website http://sparrow-leelab.cs.washington.edu/.

MATERIALS AND METHODS

Microarray data processing

Raw cell intensity files (CEL) for Affymetrix U133 Plus 2.0 and U133A arrays (Affymetrix, Santa Clara, CA, USA) for gene expression data collected from patient bone marrow were retrieved from Gene Expression Omnibus (23) for accessions: GSE12417 (24) (AML), GSE14468 (25) (AML), GSE10358 (26) (AML) and GSE13159 (27) (AML, chronic lymphoblastic leukemia (CLL), childhood acute lymphoblastic leukemia (cALL), myelodysplastic syndrome (MDS)). Expression data were then processed using MAS5.0 normalization with the ‘Affy’ bioconductor package (28) and mapped to Entrez gene annotations (29) using custom chip definition files (CDF) (30). After the initial normalization and mapping to gene level expression levels, arrays were organized into groups based on the date the array was processed, which is encoded within the raw CEL data file for each sample. This temporal grouping was treated as a proxy for possible systematic biases in expression variation driven by batch effects. The ComBat batch effect correction algorithm (31) was then applied to each data set treating each group as a batch to correct for these possible batch effects. Dates where there was only one array processed were dropped from all data sets. MAS5.0 normalized intensities were natural log transformed prior to batch effect correction. For the GBM data from The Cancer Genome Atlas, Agilent g4502a gene expression data were accessed from Firehose for the 02–15 run (http://gdac.broadinstitute.org/runs/stddata_2014_02_15/data/GBM/20140215/). All normal samples were removed, and plate level batch effects were corrected with ComBat.

Learning SERs

Prior to combining the data across studies, all genes in each individual data set were scaled to have mean zero and variance one. After combining all patients with normal karyotype AML there were 682 patients with gene expression data for 17778 genes on the Affymetrix U133 Plus 2.0 platform (Affymetrix, Santa Clara, CA, USA). Similarly, there are 447, 235 and 206 patients for CLL, cALL and MDS, respectively. For the GBM data there are 502 patients and 17814 gene expression variables. We focused on a set of 3460 SERs, out of which 3052 are present in the Affymetrix U133 Plus 2.0 data. These SERs included transcription factors, chromatin modifiers and other regulatory factors similar to the list proposed by Gentles \textit{et al.} (32), 2993 of which are in the GBM data set. The full lists of SERs are provided in Supplementary Note 1.

SPARROW SERs

Cancer expression data contain complex correlation patterns due to the effect of selection on many different pathways in many different patients (1) (Figure 1b). In order to identify SERs, we developed a new method to filter out false positives in a sparse linear regression model. Specifically, we used the variational Bayesian spike regression (VBsr) model (33) as a baseline model and developed new methods to filter out false positives, by incorporating a data adaptive method for choosing the penalty parameter that does not require cross validation. The spike regression model is a Bayesian approximation of best subset selection. In this case, the best subset selection would be the choice of a subset of a fixed number of SERs, out of all possible SERs, that best explain the observed gene expression pattern of a given gene across patients. This is a difficult combinatorial problem, therefore we approach it with a variational Bayes approximation that treats the inclusion of any given SER in the regression model probabilistically. The variational Bayes approximation introduces additional probabilistic constraints that make inference extremely efficient (34).

The main novelty of our approach is the choice of the size of the subset of SERs to be used in the model. Previous work on the spike regression model suggested not only a posterior probability of inclusion in the model for each SER but also a statistic that measures the significance of a given SER in a given subset regression model (33). Yet the previous work also required a computationally expensive path search to determine an optimal best subset size to ensure that the null distribution of the test statistic for each feature was well calibrated (i.e. normally distributed). If the best subset size is too large, then there are too many features in the model given the sample size, and the test-statistics becomes inflated. Alternatively if the best subset size is too small, then there is the possibility of missing out on true associations. In this new version, we derive a fast analytical method to determine the best subset size to ensure that the test statistic is well calibrated without losing too many associations. This is a non-trivial problem, where many previous authors use \textit{ad hoc} methods such as cross validation to choose the analog of this model complexity parameter (35).

For each of the 17 788 genes we fit the following model

\[
y_{ik} = \sum_{j} x_{ij} \beta_{jk} + e_{ik},
\]

where \(y_{ik}\) is the expression level of the
kth gene for the ith patient for i = 1, . . . , n patients and k = 1, . . . , r genes, e_{ik} is a normally distributed error term (assuming independent and identically distributed errors), x_{ij} is the expression level of the jth SER for the ith patient and β_{jk} is the additive effect of the jth SER on the expression of the kth gene.

In addition, we implement the spike regression algorithm proposed by Logsdon et al. (33), where β_{jk} \sim I[β_{jk} = 0] (1 - p_β) + I[β_{jk} ≠ 0] p_β. The spike penalty is an improper prior, with a hyper-parameter, p_β, determining the prior probability that each β_{jk} is non-zero. We define an alternative parametrization of p_β with support −∞ < l_0 < ∞, l_0 = 2\log(p_β) − 2\log(1 - p_β) + log(2π).

This alternative parametrization motivates a new method for choosing p_β as a function of the number of SERs (m) and patients (n).

We propose to choose the penalty parameter for a given data set that is computationally efficient, l_0 = −(F^{-1}(1 - 0.05/m) - log(n) + 2\log(1 - p) - 2\log(p)).

This corresponds to p_β = \frac{\exp(\nu_0 + \sqrt{2\nu_0 \exp(\nu_0)})}{\exp(\nu_0) - \nu_0}. The speed up of the algorithm is significant, with a single run taking 10–100 times faster than the approach of Logsdon et al. (29). This makes the algorithm very computationally efficient, and scales approximately as O(nm). The most computationally expensive step is the choice of the number of restarts, but empirically there did not seem to be much gained by looking at more than 12 restarts. This choice of penalty parameter has F as the cumulative distribution function of a \chi^2 random variable, with n patients, m SERs and p = 0.95. In the variational Bayes spike regression model each β_{jk} coefficient has both a posterior probability of association and a Z-statistic. The posterior probability of a coefficient bounds the influence of the variable within the model, e.g. if the posterior probability is zero, then the SER will have no effect on the gene in question, and if the posterior probability is one, then the SER will have a full un-penalized effect. The Z-statistic is a measure of the significance of the β_{jk} parameter given the rest of the fitted model. The above choice of prior probability p_β corresponds to calibrating the posterior probability for any given β_{jk} to be 0.95 when its Z-statistic passes a Bonferroni correction for the total number of features—in this case a family-wise error rate of 0.05. This choice of penalty parameter allows us to still use the feature level Z-statistic proposed by Logsdon et al. (29), but without the computationally expensive path search over l_0, by having SERs whose association passes a Bonferroni correction contribute at least 95% of their maximal possible effect to the model. We found empirically that this choice of penalty parameter led to maximum power in the model with little to none inflation of the test statistic.

Accordingly, we define the sparse basis for each gene given the SERs in each individual regression model with \hat{z}_{jk}^2 > F^{-1}(0.05/m_p), with \hat{z}_{jk} being the Z-statistic for the effect of the jth SER on the kth target gene from the fitted spike model. For the SPARROW method, the driver sets are constructed after applying a Bonferroni correction across the driver genes of all models (0.05/(3052 × 17788) = 9.2 × 10^{-10}) to fix the family-wise error rate to 0.05 across every SER basis. The SPARROW and VBSR algorithms are freely available from CRAN (http://cran.r-project.org), and our project website (http://leelab-data.cs.washington.edu/sparrow.R).

**LASSO SERs**

As with SPARROW, for the LASSO we use a penalized regression methodology that identifies SERs for each gene. We solve the following lasso optimization (8) problem β^k = \arg\min_{β^k} \sum_i (y_{ik} - \sum_j x_{ij}β_{jk})^2 - λ_k \|β_{jk}\| with two different strategies for the choice of the penalty parameter λ_k for each gene’s SER basis. The first strategy chooses the λ_k that minimizes the 10-fold cross validation mean-squared error. The second strategy chooses the λ_k that produces a cross validation mean-squared error one standard deviation from the minimum mean-squared error, where the standard deviation is estimated across cross validation folds. All models are fit using the glmnet R-package (36). The lasso bases are defined as S_k = I[β_{jk} ≠ 0], where the SER scores computed based on how often each SER has a non-zero coefficient.

**Correlation and WGCNLA SERs**

Correlation SERs are inferred using a correlation network approach. In general, we use the number of edges an SER has in an inferred network as its SER score. A correlation network is constructed similarly to the hard thresholding WGCNA method of Langfelder and Horvat (16), where we test for the marginal dependence between each SER and each target gene using an additive linear model assuming independent and identically normally distributed errors e_{ik} \sim N(0, σ^2), y_{ik} = μ_k + x_{ij}β_{jk} + e_{ik}. The additive effect β_{jk} is estimated using the maximum likelihood estimator, μ̂_k, β̂_{jk}, σ̂^2_k = \arg\max_{μ_k, β_{jk}, σ^2_k} L(μ_k, β_{jk}, σ^2_k), and the null hypothesis H_0 : β_{jk} = 0 is tested using the standard Wald test to produce a t-statistic t_{jk} for each β_{jk}. We then construct the network using two different strategies. For the Correlation1 strategy, we construct a network by applying a Bonferroni correction for m × r tests to each pairwise test. The Correlation1 SERs for the kth gene are defined as S_k = |t_{jk}| > G^{-1}(1 - 0.05/m, n − 2), with G being the cumulative density function of a \chi^2 distributed random variable. For the second method of building the network (Correlation2), we choose a threshold for t_{jk} such that the sparsity of the entire correlation network is equivalent to the sparsity identified by the SPARROW bases across all genes. All estimates and test statistics are produced using the ‘lm’ function in R. We also construct a soft thresholded network similar to WGCNA with the scale free rule (16), where the absolute value of the correlation coefficients between SERs and each gene are raised to a power. The power is chosen to make the network as scale free as possible.

**ARACNE SERs**

We estimated a network using the ARACNE algorithm to generate ARACNE SER scores (16). The ARACNE
network first estimates a non-parametric joint probability density using a Gaussian kernel between each pair of genes in the data set, \( (\mathbf{z}) = 1/n \sum h^{-2} Q(h^{-1} | \mathbf{z} - \mathbf{z}^\prime |) \), where \( Q(\ldots) \) indicates a standard bivariate normal density function. The mutual information defined by this estimated density is \( I(x, y) = \frac{1}{n} \sum \log \left( \frac{(f_{x,y}(x,y))}{(f_{x}(x) f_{y}(y))} \right) \). The ARACNE algorithm then tests whether each pairwise mutual information is significant, and defines an edge between two genes when the mutual information is significant. Finally, for all triplets of genes with significant pairwise mutual information, the data processing inequality is used to prune away edges where a gene dissociates two other genes. The network that is produced after applying the data processing inequality to all triplets of genes is then the final ARACNE network. We used the adaptive partitioning option to speed up the run-time of the algorithm, and we set a significance cutoff based on a Bonferroni correction as with the SPARROW and correlation methods for \( 0.05/36 \). We treated the SERs as transcription factors in the ARACNE method, where edges between two non-transcription factor genes will not be removed by the data processing inequality unless the third gene is a transcription factor. This allows a fair comparison to the regression methods that focus only on the edges for SERs.

**SER definitions**

For the \( i \)th SER, the SER score \( H_i^k \) was computed as \( H_i^k = \sum_j I[\theta_{ji} \neq 0] \), where \( \theta_{ji} \) is a parameter indicating whether the \( i \)th SER is used in the bases for the \( j \)th gene, for \( k \) (SPARROW, LASSO\(^1\), LASSO\(^2\), Correlation\(^1\), Correlation\(^2\), ARACNE). For the soft-thresholding WGCNA method, the SER score was defined as \( \sum_j |\text{Cor}(y_i, x_j)|^{\beta} \), with \( y_i \) being the \( i \)th SER, \( x_j \) the \( j \)th gene and \( \beta \) the soft-thresholding parameter chosen to make the network as scale free as possible. SER scores define a ranking of SERs, where there is no absolute cutoff.

**Sparse SER simulation**

To test the performance of the SPARROW methods compared to the LASSO methods we simulated genes under selection using the expression data from 682 patients for the 3052 SERs in AML. Genes were simulated by independently selecting SERs with either a probability \( 3.9 \times 10^{-3} \) or \( 1.6 \times 10^{-2} \), such that the expected number of SERs \( (p) \) was either \( E[p] = 12 \) or \( E[p] = 50 \). The effects of the selected drivers were sampled from an independent and identically distributed (i.i.d.) standard normal distribution. The expected pattern of expression for the simulated gene was computed as the linear combination of sampled effects weighted by the sampled effects, with an i.i.d. error standard normal error term added. This process was replicated 100 times. Additionally, for the result shown in Figure 3 we also simulated a model with \( E[p] = 50 \), and the percentage variation explained by the regression model constrained to 80% to compare the performance of SPARROW and LASSO.

**Survival analyses**

We ran all survival analyses by fitting a Cox proportional hazards model (37). The log partial likelihood for a Cox proportional hazards function is defined as \( l(\beta) = \sum_{i, C_i = 1} \beta^T X_i - \sum_{i, C_i > 1} \exp(\beta^T X_i) \), where \( Y_i \) is the observed time for observation \( i \), \( C_i = 1 \) indicates that an event occurred at time \( Y_i \), \( C_i = 0 \) indicates censoring at time \( Y_i \), \( X_i \) is the vector of covariates for observation \( i \), and \( \beta \) as the log hazard ratios for the vector of covariates. The log partial likelihood is maximized using the ‘coxph’ function from the survival package in R (38). The reported test statistic for each covariate’s log hazard ratio \( \beta_j \) is generated using a Wald test based on the fitted model.

**Random effects model**

We tested the aggregate association between the variation in the top 100 SERs and the expression of each gene with the following probabilistic model \( y_{ik} = \sum_{j} x_{ij} y_j + \epsilon_{ik} \), where \( H_j^k = 1 \) indicates that SER \( j \) is in the top 100 SERs for all methods. The null hypothesis \( \gamma_{H_j^k} = 0, \ldots, \gamma_{H_0^k} = 0, \ldots, 0 \) is tested using a random effects score test as implemented in the Sequence Kernel Association Test R package (39) using uniform weights and a linear kernel matrix.

**Enrichment tests**

All enrichments for gene sets are computed with Fisher’s exact test. The enrichments are calculated relative to the en-
tire set of SERs. All receiver operating characteristic (ROC) curves and area under the curves (AUCs) are calculated using the ROC R-package (40). We used the Malacards and Genecards database (41) to generate disease-specific gene lists for AML, cALL, CLL and MDS (see Supplementary Note 1). The Malacards and Genecards database contains aggregated gene annotations from 59 sources (41). For the cancer driver gene list we used the genes annotated by the Cancer Gene Census (CGC) (42), which are identified as having evidence from multiple independent sources as being cancer drivers. For the tumor-specific drivers, we use the 23 genes identified as AML drivers (43) and 67 genes identified as GBM drivers (8). For SERs from AML (43) we also looked at the 260 genes with tier 1 mutations in at least two patients, 68 genes with tier 1 mutations in at least three patients and 37 genes with tier 1 mutations in at least four patients. Apoptosis and myeloid regulator gene sets are shown in the Supplementary Note 1. For the survival enrichments we calculated a one-sided Kolmogorov–Smirnov (KS) test for the top $m = (1, \ldots, 5) \times 100$ SERs for each method based on the distribution of marginal $P$-values from a survival analysis compared to the distribution of survival $P$-values of all SERs. To correct for the inflation in the KS test due to correlation between genes we generated an empirical null distribution of $P$-values for all 17 788 genes for 10 000 permutations of survival time and censoring. We then re-computed the KS test statistic using the null permutations to determine the corrected survival enrichment for the hubs from each network reconstruction method. Additionally, for the ROC curve for the survival enrichment, the set of genes nominally associated with survival based on a Cox proportional hazards model ($P$-value $<0.01$) was used among all 3052 candidate SERs.

**Leukemic stem cell score**

We computed the leukemic stem cell (LSC) score using the weights provided by Gentles et al. (44), for 28 of the 31 genes. These weights were applied to the training data after each gene was standardized to have mean zero and variance one.

**Expression and drug sensitivity profiling of 30 AML patient samples**

We profiled drug-induced cytotoxicity of an initial sample set of 30 viably cryopreserved primary AML specimens for which detailed clinical information on patients was available, including age, gender, cytogenetics, mutation status, antecedent hematologic disorder, initial blast count, initial platelet count, treatment regimen, response and survival. Screens of the drug sensitivity of primary AML samples was performed after thawing samples in the presence of DNAase followed by short-term culture (48 h) in media containing horse serum and fetal calf serum and low level hSCF (10 ng/ml), followed by density gradient centrifugation to prepare blast-enriched fractions with high (80–90%) viability and high blast fraction, $\geq80$–90. Blast-enriched cell fractions were plated in 384 well plates coated with matrix peptide at a density of 5000 cells/well. Once plated cells were treated with obatoclax at eight different concentrations ranging from the nanomolar to low micromolar range ($3.3 \times 10^{-9}$, $1.0 \times 10^{-8}$, $3.3 \times 10^{-8}$, $1.0 \times 10^{-7}$, $3.3 \times 10^{-7}$, $1.0 \times 10^{-6}$, $3.3 \times 10^{-6}$, $1.0 \times 10^{-5}$). Viability was assessed after 4 days in culture using CellTiter-Glo. These screens were performed, at the Quellios HTS Core (http://depts.washington.edu/uwhts/) at the UW’s Institute for Stem Cell and Regenerative Medicine (ISCRM).

RNA expression profiles of primary AML samples were performed in parallel at the UW Center for Ecogenetics and Environmental Health using the Affymetrix U133 Plus 2.0 platform.

**Microarray data processing for AML patients with drug sensitivity**

Gene expression data were collected on the Affymetrix U133 Plus 2.0 platform (Affymetrix, Santa Clara, CA, USA) for bone marrow samples from these 30 patients. As with the public AML data, expression data were first processed using a MAS5.0 normalization (28), and then probes were mapped to Entrez gene annotations (25) with the help of custom CDF (30). Potential batch effects were corrected using the ComBat algorithm (31) based on the three batches the arrays were run in. Signal intensities were natural log transformed prior to batch effect correction.

**Curve fitting to estimate drug sensitivity profiles**

Nonlinear curve fitting was performed using MATLAB’s ‘nlinfit’ function. After curve fitting, summary statistics were extracted based on the curve fits similar to the Cancer Cell Line Encyclopedia (CCLE) (45), including the activity area (AA), AUC, IC$_{50}$, EC$_{50}$ and $A_{\text{max}}$. Batch effects were corrected using ComBat (31) for each drug sensitivity summary based on the five batches that experiments were run in for the 30 patients. Curve fits for 30 patients are shown in Supplementary Figure S1.

**PYCARD and obatoclax association tests**

Associations between PYCARD expression and obatoclax sensitivity in patients were tested with either a standard linear regression model (with the ‘lm’ function in R) or a robust regression model (using the ‘rlm’ function from the MASS package in R (46)) for the five drug sensitivity summaries. Both expression data and drug sensitivity data were Winsorized at the 95% percentiles of the distributions. The robust regression method uses an M-estimator to decrease the influence of outliers on the test-statistic for the additive effect in a regression model (46). To control for a possible confounding factor, whether or not a patient achieved a complete response (CR) from the therapy they were given (not obatoclax) was included in the model.

**RESULTS**

**Methods overview**

To identify SERs, we fit our sparse learning algorithm (SPARROW) using $\approx 3000$ SERs in expression data from AML, GBM, cALL, CLL and MDS patients. Candidate SERs were chosen from transcription factors, chromatin
modifiers or regulators and signal transduction genes (32), since regulatory genes are attractive targets of selection for modulating expression of entire pathways (32,47). We prioritize SERs that are most frequently chosen in the sparse representations across all genes as more likely to drive tumor-specific evolution. To test the performance of our method, we compare highly ranked SERs of SPARROW with those of alternative statistical models that fall into four categories. These categories include models that leverage sparse conditional dependence relationships (SPARROW, LASSO1,2), conditional dependence (ARACNE), marginal dependence (Correlation1) and sparse marginal dependence (WGCNA, Correlation2). For the LASSO we implement two versions where the second version is sparser than the first. Details are presented in the Materials and Methods section.

Superior sparse estimation from SPARROW algorithm compared to LASSO

Similar to previous work (33,48,49), the SPARROW algorithm outperforms the LASSO algorithm with two different metrics for choosing the sparsity of the LASSO regression model. The two tuning parameter selection approaches for the LASSO include minimum 10-fold cross validation error and the 1 standard deviation rule (36). We see in Figure 3a that the SPARROW estimates for a simulated test case with 53 true SERs among 3052 candidates using the gene expression data from AML show improved performance compared to the LASSO estimates (Figure 3b and c), in terms of both the number of false positives (2 versus 158 and 62) and the bias in the regression coefficients. The LASSO method is known to produce biased estimates of the non-zero coefficients, especially those with large effects (14). Hence why the slope of the line is less than unity for the LASSO (where a slope of unity is the expectation in the case of an unbiased estimator).

AML mutation drivers

Recently, 23 genes with driver mutations were identified in AML within The Cancer Genome Atlas (43) (Supplementary Note 1). Driver mutations were defined as genes with a mutation frequency that was greater than the background mutation rate, as determined with the MuSiC suite of tools (50). Sixteen of these 23 genes were in our candidate SERs of 3052 genes in AML, and we see significant enrichment of these 16 AML drivers among the top 500 ranked SERs from the SPARROW method (P-value: 4.1 × 10−3, enrichment fold: 3.8) (Figure 4a), as well as the best classification performance (AUC = 0.82). The only other method with comparable performance is LASSO2 (P-value: 8.9 × 10−3, enrichment fold: 2.3) while the other approaches performed poorly at prioritizing genes with driver mutations. To test a larger gene set of candidate SERs in AML, we also tested for enrichment of genes mutated in at least two, three or four patients in the TCGA data set (Figure 4b–d). The enrichments and significance of the tests increase with the frequency of recurrent mutation. This supports our hypothesis that SERs are more likely to be genes that are under similar shared selection across patients, and therefore more likely to be recurrently mutated in their DNA.

Leukemic SERs

To further investigate the SERs in leukemia, we applied SPARROW to two other leukemias (cALL, CLL) and MDS (which frequently progresses to AML) using data from the Microarray Innovations in LEukemia (MILE) study (27). Leukemias and lymphomas are aggressive malignancies, characterized by the accumulation of immature blasts that fail to differentiate into the appropriate effector cells of erythropoiesis, lymphopoiesis and myelopoiesis. Many leukemia driver genes are also known regulators of normal hematopoiesis or myeloid regulators (43). To test the hypothesis that leukemia SERs were more likely to be myeloid regulators, we tested for enrichment of 17 myeloid regulators in the top SERs (Supplementary Note 1). There is sig-
significant enrichment across AML, cALL, CLL and MDS for these myeloid regulators, especially for the SPARROW methods. For example, the SPARROW enrichments for the top 500 SERs in AML, cALL, CLL and MDS had \( P \)-values of \( 1.4 \times 10^{-10}, 9.3 \times 10^{-7}, 3 \times 10^{-3}, 1.4 \times 10^{-10} \) and enrichment folds: 5, 3.9, 2.5 and 5, respectively (Supplementary Figure S2A–D). We tested for enrichment of gene sets specific to each leukemia based on Malacards and Genecards (41), and also saw statistically significant but modest enrichments, especially with SPARROW SERs (Supplementary Figure S3A–D), with \( P \)-values: 3.1 \times 10^{-10}, 2.3 \times 10^{-2}, 4 \times 10^{-5}, 6.2 \times 10^{-3} \) and enrichment folds: 1.7, 1.4, 1.5 and 1.4 for AML, cALL, CLL and MDS, respectively, among top 500 SPARROW SERs. Finally there is a nominal association for enrichment of genes identified in the CGC (42,51), a gold standard of general cancer driver genes, though these genes are not necessarily specific to the leukemias in question (Supplementary Figure S4A–D), \( P \)-values: 3.4 \times 10^{-4}, 3.4 \times 10^{-3}, 3.4 \times 10^{-3}, 6.3 \times 10^{-4} \) enrichment folds: 1.5, 1.4, 1.4 and 1.5 for AML, cALL, CLL and MDS, respectively, among top 500 SPARROW SERs.

**Comparison of methods for identifying SERs in AML**

We have shown that the SERs identified by SPARROW outperform standard correlation or convex penalized regression (LASSO) approaches in identifying genes with known or potential driver roles (Figure 4 and Supplementary Figures S2–S4). In addition we also show the AUC for all of these results in Supplementary Table S1. All methods have different orderings of SERs, as shown in pairwise plots of SER scores (Supplementary Figure S5). We predict that genes whose expression is driving the underlying tumor disease processes will be highly associated with the entire cancer transcriptome. To test this hypothesis, we took the top 100 SERs for each method in AML and tested whether their joint variation was significantly associated with variation for all genes (see the Materials and Methods section for further details). The distribution of \( P \)-values across all genes from this test is shown in Supplementary Figure S6. The SPARROW methods produce the most significant associations between top SERs and genome-wide gene expression (paired Mann–Whitney test \( P \)-value <2.2 \times 10^{-16} for all pairwise comparisons between SPARROW and other methods).

We predict that selection will favor fundamentally sparse bases. The summary statistics for the SER inference are shown in Supplementary Table S2. The SPARROW methods have both the sparsest median bases size (4) and median frequency that a given SER is used in a basis (24). To test whether SPARROW is under-powered or the data support sparse SER bases, we simulated 100 replicates of gene expression data using 12 or 50 randomly SERs in the AML data (Supplementary Table S3). We find that the SPARROW method is well powered to identify denser bases given the data (power 86% and 82%, respectively). This suggests that the number of SERs for any given gene is quite sparse and that the sparsity assumption leveraged by SPARROW is the key to identifying the SERs. The other sparsity methods (LASSO) have similar power to SPARROW but with higher false discovery rates (0.65–0.86) as compared to SPARROW (0.005–0.03). The genes downstream of each SER identified by SPARROW were tested for enrichment among canonical signaling pathways (52), and a diverse range of processes was identified (Supplementary Note 2). Enriched processes were not shared across SERs, suggesting that each SER is a unique indicator of sets of genes under selection in a subset of patients.

**AML SERs are associated with patient prognosis**

Next, we tested whether the expression levels of the identified SERs were more likely to be significantly associated with overall survival using Cox proportional hazards regression. For the GBM data we did not observe a significant enrichment for association with survival (results not shown), though the variation in survival for the TCGA GBM is much smaller (with patients doing uniformly poorly) than in AML, so there may be less signal to detect. In AML we modeled overall survival in several independent patient cohorts (24,26,53) using expression of each SERs as predictors with age, \( NPMI \) mutation status and \( FLT3 \) mutation status as covariates. We then computed a KS test statistic to test for an inflation in the top univariate \( P \)-values from the top ranked SERs with respect to the empirical distribution of \( P \)-values (see the Materials and Methods section for further details). This method of testing for enrichment was performed to test if there was still enrichment for association with survival after correcting for correlation among the top ranked SERs for each method. Figure 5a shows the permutation \( P \)-values from this KS enrichment.

<p>| Table 1. Association between expression of ( PYCARD ) and response to obatoclax therapy for the log_{10} IC50 estimated from obatoclax dosage response curves |
|-----------------|-----------------|-----------------|</p>
<table>
<thead>
<tr>
<th>( \log_{10} ) IC50</th>
<th>Model</th>
<th>Variable</th>
<th>Patients (n = 30)</th>
<th>( P )-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.4</td>
<td>Univariate OLS</td>
<td>( PYCARD )</td>
<td>-0.37 (-0.71, -0.032)</td>
<td>0.033</td>
</tr>
<tr>
<td>1.5</td>
<td>Bivariate OLS</td>
<td>( PYCARD )</td>
<td>-0.45 (-0.73, -0.17)</td>
<td>0.0025</td>
</tr>
<tr>
<td>1.5</td>
<td>Univariate robust</td>
<td>( PYCARD )</td>
<td>0.52 (0.25, 0.79)</td>
<td>0.00051</td>
</tr>
<tr>
<td>1.5</td>
<td>Bivariate robust</td>
<td>( PYCARD )</td>
<td>-0.38 (-0.74, -0.019)</td>
<td>0.04</td>
</tr>
<tr>
<td>1.4</td>
<td>Univariate OLS</td>
<td>( CR )</td>
<td>-0.43 (-0.7, -0.16)</td>
<td>0.0029</td>
</tr>
<tr>
<td>1.5</td>
<td>Bivariate OLS</td>
<td>( CR )</td>
<td>0.59 (0.33,0.85)</td>
<td>0.000077</td>
</tr>
</tbody>
</table>

We used two regression methods, ordinary least squares (OLS) and robust regression, to test for an association between \( PYCARD \) expression and obatoclax sensitivity. To control for potential, a potential confounder in terms of good versus poor responders, whether or not the patient experienced a complete response (CR) from the therapy they were given (in this case not obatoclax) was also included in the model.
test as a function of the SER ranking for the alternative methods. Only the SPARROW method passes the empirical P-value threshold of 0.05. In Figure 5b the ROC curves are shown for the prediction of survival associated genes (P-value <0.01) based on the SER scores, where the highly ranked SPARROW SERs are most likely to be associated with survival. This method shows that even when not correcting for correlation among genes, SPARROW still outperforms the other methods for survival enrichment. The AUC for these ROC curves along with the AUC excluding known m-drivers and candidate c-drivers in AML is shown in Supplementary Figure S7. We see that the SPARROW method has the best AUC, and none of the methods are significantly associated with poor outcomes. The selection of RAB37 and RAB27A as drivers together with the apparent importance of GUCY1A3 to LSC biology suggests further study. The fourth and fifth most highly scoring genes encode for zinc finger proteins. ZNF521 (also known as early hematopoietic zinc finger protein, EZHF) is highly expressed in early hematopoietic cell types and is significantly expressed in many AMLs (57). Its deregulation has been postulated to play a role in hematopoietic cancers. Less is known about ZNF185, but roles have been suggested for it in cellular differentiation and regulation of proliferation of chronic myelogenous leukemia cells.

Apoptotic SERs in AML

Dysregulation and inhibition of apoptosis is a critical feature of cancer, and in leukemias in particular. BCL-2 family genes with anti-apoptotic functions are commonly mutated or overexpressed in leukemias (58). Of 84 known apoptosis genes (59), 30 were SERs. We found that these 30 apoptosis regulators were more likely to be SERs in AML by SPARROW than those identified by other methods, especially in the top 500 SPARROW SERs (P-value: 5.7 × 10^{-3}, enrichment fold: 2) (Supplementary Figure S8). The gene-set sizes and enrichments for this result and for the other main gene-sets tested in AML are shown in Supplementary Table S5. The top three ranked apoptosis SERs based on SPARROW are PYCARD, CASP1 and DAPK1. Since PYCARD is relatively poorly characterized, particularly in the context of AML, we focused on it in further analyses.

The apoptotic SER PYCARD is a potential biomarker for obatoclax

PYCARD is known to play an important role in multiple cellular processes, including inflammation through the formation of the inflammasome (60), and mitochondrial driven apoptosis (61). We functionally analyzed the genes that selected PYCARD in their sparse SER basis. There were four distinct clusters after clustering these genes based...
on expression (Melst R package (62); Supplementary Figure S9). We tested the clusters for enrichment with Gene Ontology (63) and found that the two largest clusters were enriched for immune response (q-value: 2.5 × 10^{-15}) and mitochondria genes (q-value: 4.8 × 10^{-20}), respectively; see Supplementary Note S3.

The PYCARD protein is known to directly activate CASP9 and BAX. In addition, BCL-2 genes are known anti-apoptotic factors that inhibit normal BAX function (58) (Supplementary Figure S10). We hypothesize that if BCL-2 activity is reduced, PYCARD expression could be a marker for apoptosis induced by BAX. Because of its high ranking as an apoptotic SER in AML among the apoptotic regulators, along with PYCARD’s role in BAX regulation, we hypothesized that PYCARD expression is associated with response to therapy for the anti-apoptosis BCL-2 inhibitor obatoclax in patients with AML. To test this hypothesis, we assayed both genome-wide gene expression levels of 30 patients with AML on the Affymetrix U133 Plus 2.0 platform (Affymetrix, Santa Clara, CA, USA) along with in vitro drug sensitivity to obatoclax in leukemic tissue cultured from these 30 patients. The AA (area above the dosage response curve), AUC (area below the dosage response curve), A_{max} (maximum inhibition response observed), IC_{50} (concentration that gives 50% inhibition) and EC_{50} (concentration that gives 50% inhibition between maximum and minimum response) were estimated using a curve-fitting algorithm for blast counts at varying concentrations of obatoclax (see the Materials and Methods section).

We tested for association between response to obatoclax and PYCARD expression across 30 patients with both OLS (ordinary least squares) and robust regression models (see the Materials and Methods section). The drug sensitivity measure of IC_{50} had the strongest evidence for an association (Table 1 and Figure 6) with a negative effect −0.37 (95% CI: [−0.7, −0.032], P-value: 0.04) for the association with PYCARD for the robust regression model. The negative effect suggests higher expression of PYCARD is indicative of increased sensitivity to obatoclax. To control for possible confounding in terms of patients who were good responders to the therapy they were given (not the in vitro response) versus poor responders, we controlled for CR in the regression model and found that it increased the evidence of association between PYCARD and obatoclax with a conditional effect of −0.43 (95% CI: [−0.7, −0.16], P-value: 0.0029). The good responders (CR = 1) tended to have higher in vitro resistance to obatoclax with an effect 0.59 (95% CI: [0.33, 0.85], P-value: 0.000077) in the robust regression model (Table 1 and Figure 6). While A_{max} has little evidence of association (Supplementary Table S6 and Supplementary Figure S11), the other drug sensitivity measures EC_{50}, AA and AUC do have consistent evidence of association (Supplementary Tables S7–S9 and Supplementary Figures S12–S14). Additionally, to further test the robustness of PYCARD as a good biomarker and SER, we performed a joint test of association with survival and association with Obatoclax, and meta-analyzed the results (Supplementary Table S10) for the top 400 SPARROW SERs. We can see that PYCARD has by far the strongest association with both survival and response to Obatoclax therapy. Among the set of 3052 candidate SERs, only 17 had more significant associations with Obatoclax than PYCARD, but none of them showed as significant association with survival. Additionally, PYCARD had the highest SER score among those genes as well. Finally, among all 17 788 genes, 70 had more significant associations with Obatoclax than PYCARD, but none of these genes were as associated with survival and PYCARD.

DISCUSSION

As the main contribution of this paper we provide new criteria based on expression patterns to identify transcriptional regulators under selection and driving tumor progression. We define the genes under selective pressure that drive genome-wide expression changes as SERs, and propose a new statistical method, called SPARROW, to identify SERs based on gene expression data. SPARROW can identify genes with known driver mutations more accurately than similar methods based on the LASSO, and other popular approaches to learn gene association relationships from expression data. These results are driven by a new sparse regression technique we developed, which dramatically reduces false discovery rates compared to the LASSO.

The SPARROW method is designed to reveal vastly fewer false positives when considering sparse regression models across multiple genes as compared to methods like the LASSO, as shown in Figure 3. This reduction in the number of false positive edges will illuminate which genes are truly more conditionally dependent across all genes (i.e. the genes that have high SPARROW SER scores). We posit that dependencies among genes are generated by disease relevant processes, such as the effect of driver events on gene expression (Figures 1 and 2). This makes SPARROW SERs much more sensitive to genes that are driving some aspect of the
Selective pressure is an important criterion for understanding gene function, evolutionary processes and the basis of complex traits (64). There are many methods to estimate selection from DNA sequences in the field of evolutionary biology, population genetics and cancer biology (65). SPARROW is the first method to identify genes that are enriched for driver mutations purely from expression data. Existing approaches focus on the DNA sequence data, where SPARROW improves on these approaches in multiple ways. First, the majority of somatic mutations identified in tumor genomes are passenger mutations. Second, SPARROW can estimate the importance of a given gene on genome-wide expression levels, which may be more functionally relevant to the tumor biology (7). Finally, driver events in cancer may occur at different levels such as with epigenetic modifications, which are not detected by DNA sequencing.

Based on the SPARROW analysis, we propose that \textit{PYCARD} expression is a potential biomarker for \textit{in vitro} response to obatoclax. Obatoclax is a pan-inhibitor of anti-apoptotic members of the BCL2 family of proteins, which control the intrinsic or mitochondrial pathway of apoptosis (66). It is a BH3 mimic with activity against all of the antiapoptotic BCL2 family members (67), thus it is a ‘pan’ BCL2 antagonist. Obatoclax demonstrated activity against AML cell lines and primary cells, causing cell cycle blockade at S-G2 at lower concentrations, and inducing apoptosis at higher concentrations (68). In contrast to other BCL2 inhibitors, it has activity against \textit{MCL1}, and maintained activity in the absence of \textit{BAK}/\textit{BAX}, implying other mechanisms of action (68). Given we predict \textit{PYCARD} is an SER of overall apoptosis in AML, we expect that it would be applicable to a broadly acting anti-apoptotic drug like obatoclax. Furthermore, obatoclax synergized with other agents, including the \textit{BCL}-2 inhibitor ABT737, the \textit{FLRT3} inhibitor sorafenib and the chemotherapy drug cytobain to induce apoptosis (69). In combination with sorafenib, it reduced tumor formation in a xenograft mouse model with demonstrated downregulation of \textit{MCL1} (68). Given obatoclax’s potency in combination with orthogonal therapies, having a biomarker for obatoclax efficacy could potentially improve patient prognosis by informing patient selection for combination therapy involving obatoclax.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge James Anni, Timothy Martins and Sylvia Chien for their contributions to the gene expression and \textit{in vitro} drug sensitivity data generation. We would also like to thank two anonymous reviewers for their comments that significantly improved the quality of the manuscript. The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

FUNDING

National Institutes of Health [T32 HL 007312 to B.L.; U54 CA1149145 to A.G.]; National Science Foundation [DBI-1355899 to S.L.]; Life Sciences Discovery Fund [207088-01]; Philanthropic funding from Norman Metcalfe. Funding for open access charge: University of Washington. Conflict of interest statement. None declared.

REFERENCES


