Gelfond, J.A., Ibrahim, J.G., Gupta, M., Cheng, M.-H., and Cody, J.D. (2013) Differential expression analysis with global network adjustment. BMC Bioinformatics, 14 (258). ISSN 1471-2105

Copyright © 2013 The Authors.

http://eprints.gla.ac.uk/85826/

Deposited on: 17 Sep 2013

Enlighten – Research publications by members of the University of Glasgow
http://eprints.gla.ac.uk

BMC Bioinformatics

METHODOLOGY ARTICLE Open Access

Differential expression analysis with global network adjustment

Jonathan A Gelfond[1*], Joseph G Ibrahim[2], Mayetri Gupta[3], Ming-Hui Chen[4] and Jannine D Cody[5]

Abstract

Background: Large-scale chromosomal deletions or other non-specific perturbations of the transcriptome can alter the expression of hundreds or thousands of genes, and it is of biological interest to understand which genes are most profoundly affected. We present a method for predicting a gene's expression as a function of other genes thereby accounting for the effect of transcriptional regulation that confounds the identification of genes differentially expressed relative to a regulatory network. The challenge in constructing such models is that the number of possible regulator transcripts within a global network is on the order of thousands, and the number of biological samples is typically on the order of 10. Nevertheless, there are large gene expression databases that can be used to construct networks that could be helpful in modeling transcriptional regulation in smaller experiments.

Results: We demonstrate a type of penalized regression model that can be estimated from large gene expression databases, and then applied to smaller experiments. The ridge parameter is selected by minimizing the cross-validation error of the predictions in the independent out-sample. This tends to increase the model stability and leads to a much greater degree of parameter shrinkage, but the resulting biased estimation is mitigated by a second round of regression. Nevertheless, the proposed computationally efficient "over-shrinkage" method outperforms previously used LASSO-based techniques. In two independent datasets, we find that the median proportion of explained variability in expression is approximately 25%, and this results in a substantial increase in the signal-to-noise ratio allowing more powerful inferences on differential gene expression leading to biologically intuitive findings. We also show that a large proportion of gene dependencies are conditional on the biological state, which would be impossible with standard differential expression methods.

Conclusions: By adjusting for the effects of the global network on individual genes, both the sensitivity and reliability of differential expression measures are greatly improved.

Background

A goal of systems biology is to understand how a perturbation affects a network of interrelationships between genes. There are many models for gene networks, but few have shown accurate predictions across many datasets [1]. There is a growing collection of gene perturbation experiments in which a subset of transcipts' expression have been modulated, either through carcinogenesis, chemical treatments, or through siRNA [2]. These experiments comprise an information-rich dataset that allows us to construct global network models and test their predictive accuracy on the level of the full transcriptome.

Network models can be helpful for distinguishing between all differentially expressed genes and the genes that are immediately affected by a perturbation such as a gene deletion. There are clinically relevant reasons for this line of investigation. For instance, chromosomal abnormalities are the most common cause of mental retardation in the US, and the deletion of a chromosomal segment is a common subtype [3]. The deleted segments contain many genes, so that this is a type of sporadic, multiple gene knockout. The deletions are often *hemizygous*; that is, only one of the two (maternal or paternal) homologous chromosomes with the segment remains intact. Hence, the genes within the deleted segments are present with half of the normal copy number. Most genes exhibit some form of

*Correspondence: gelfondjal@uthscsa.edu
[1] Department of Epidemiology and Biostatistics, University of Texas Health Science Center San Antonio, San Antonio, Texas, USA
Full list of author information is available at the end of the article

© 2013 Gelfond et al.; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

dosage-sensitive decrease in expression (e.g., DNA deletion tends to decrease mRNA expression), but there is a possibility that the gene's regulatory network can compensate for the loss by selectively increasing the expression of the intact copy. The ability or inability of the regulatory networks to compensate with the intact copies on the homologous chromosome could explain which genes are responsible for common abnormal phenotypes. In this paper, we examined subjects with nearly identical somatic deleted segments of the q-arm of chromosome 18, and we performed comparative gene expression microarrays of these subjects with a parental control sample. Our goal is to: 1) identify the global gene expression differences, 2) to identify which genes are the most affected by the global expression differences relative to their regulatory networks, and 3) to identify genes that have different regulatory networks in case and control groups.

Various methods have been developed both for predicting gene expression and adjusting for the correlations between genes in differential expression analyses. Dahl et al [4] improved the detection of differential expression by grouping genes into co-expression clusters with a Dirichlet process mixture model. Leek at al [5] proposed a general framework for significance testing when a large-scale number of components have mutual dependency. This method may be applied to the differential expression problem and allows independent testing of hypotheses by conditioning on an orthogonal dependence kernel. These methods have shortcomings because they do not explicitly utilize the vast amount of data from prior experiments.

Some network methods do use extensive prior data such as Ruan et al [6] who used the cluster average of a set of co-expressed genes as a prediction for another gene within the cluster. This relatively simple method had similar or superior accuracy to models that used substantial auxiliary data including regulatory pathways and DNA-binding patterns of transcription factors such as the approach taken by Gustafsson et al [7] that used an "elastic-net" penalty [8] and information from multiple genomic modalities (microarray, ChIP-seq, etc.). Along the lines of Ruan et al [6], we present a simple and computationally efficient method for predicting gene expression. Our method extends the work of Congrove et al [9] who modeled a gene's expression as linearly dependent on other genes, but we have added features that improve computational efficiency and robustness to variations in transcriptional networks.

Methods

Suppose we have a series of microarrays represented by a matrix Y that is $G \times N$ where G is the number of the genes on the N arrays, and the expression measurements for each gene g will be the row vector $y_g, g \in [1 \ldots G]$. We define the perturbation vector X as an $N \times 1$ vector with elements $x_i = I(i^{th}\text{array is perturbed})$ where $I(\cdot)$ is the indicator function. Our goal is to estimate the effect of the perturbation conditional on the gene expression network $P(y_g|X, Y_{-g})$ where $Y_{-g} = \{y_{g'} : g' \neq g, g' \in [1 \ldots G]\}$. That is, we want to predict the effects of the perturbation on a gene given that it is within a regulatory network consisting of all other genes. Typically, the sample size (N) of the experimental data Y is small ≤ 10, but we assume that a database of experiments Y_D ($G \times N_D$) is also available with a substantially larger sample size $N_D \geq 100$. Our method will use this database Y_D as an independent training set to estimate the regulatory network within Y.

We make the simplifying assumption that the log-transformed gene expression y_g follows an approximately Gaussian distribution with mean $\beta_g Y$. A predictive model for each gene within Y and Y_D can be constructed by using Sparse Simultaneous Equations Model (SSEM) [9,10]. Under this model, the gene expression of a transcript y_g can be estimated by a weighted sum of the remaining transcripts. The key parameters are within β_g ($1 \times G$) where the $\beta_{gg'}$ elements are the influence that gene g' has on gene g and represent the overall transcription network with $\beta_{gg} = 0$ so that genes do not influence themselves. These β_g parameters can be estimated from a database of related array experiments. Thus, we have the model

$$y_g = \beta_g Y + \epsilon_g \tag{1}$$

where ϵ_g is an $1 \times N$ vector of Gaussian errors. This may be expressed in matrix form as

$$Y = BY + E \tag{2}$$

where B is $G \times G$ and composed of the $\beta_{gg'}$ parameters such that B has diagonal elements consisting of zeros, and E is the $G \times N$ matrix of Gaussian errors with rows ϵ_g. The intrinsic gene network can be described by B. Cosgrove et al [9] model the effect of the perturbation of this network by introducing ϕ_g

$$y_g = \beta_g Y + \phi_g + \epsilon_g, \tag{3}$$

and in matrix form

$$Y = BY + \Phi + E. \tag{4}$$

The Φ parameter is the $G \times N$ matrix composed of ϕ_g and is the *direct* effect of the perturbation X on gene expression that is not accounted for by the gene network. Their goal is to estimate these direct actions Φ upon gene expression to elucidate, for example, the direct drug targets. They estimated the B matrix by using *compendia* or a large ($N > 100$) database of experiments Y_D, and estimate Φ using the residuals

$$r = y^{pert} - By^{pert} = \phi^{pert} + \epsilon^{pert}. \tag{5}$$

Cosgrove et al [9] used ϕ^{pert} as the estimator of the direct action of the perturbation. Further, they did not take into account that the network B may not be the same in different cell types or biological conditions, which we consider in our approach below.

Although it was originally applied to large genomic databases, this model can be extended to apply to situations where the current experiment has insufficient data in order to accurately estimate the B parameter. Furthermore, estimating B with a large database Y_D and applying it to a new dataset will avoid the problem of using the data twice. For instance, if we have a large database of expression measures Y_D, we may estimate B_D such that $Y_D = B_D Y_D + E$, under the assumption that $B_D \approx B$. Given our estimate of B_D with rows β_{Dg}, we may then construct a linear predictor of each gene g in the new dataset Y by $Y_{\mathrm{N}g} = \beta_{Dg} Y$.

With a covariate matrix X that applies to the individual samples, we propose the model

$$y_g = \beta_{Dg} Y \nu_g + X' \delta_g + \epsilon = Y_{\mathrm{N}g} \nu_g + X' \delta_g + \epsilon. \tag{6}$$

This equation differs from Equation 3 as $\beta_{Dg} Y \nu_g = Y_{\mathrm{N}g} \nu_g$ is substituted for $\beta_g Y$ and $X' \delta_g$ is substituted for ϕ_g. The additional ν_g scalar parameter represents a very important increase in model flexibility compared to Cosgrove et al [9] where ν_g is constrained to be 1.0. In the new approach, ν_g is estimated for each gene so that if the historical network is inconsistent with the current network and is not a good predictor for gene expression in the current experiment, then its effect on inference approaches zero as ν_g does. The effect of the biological condition on gene g is described by the scalar δ_g, and this is distinct from the usual formulation of differential expression that tests the effect of X $(N \times 1)$ alone on y_g [11,12]. The δ_g parameter corresponds to an element of the Φ parameter, but unlike ϕ_g in Equation 3, δ_g is *jointly estimated* with ν_g and undergoes formal statistical testing in our approach rather than acting merely as a basis for ranking candidates. Hence, the hypothesis $H_0 : \delta_g = 0$ is the test of differential expression, conditional on the value of the historical gene network. The p-value can be derived through standard linear model theory, and the status of differential expression can be determined by setting a cutoff for p-values or false-discovery rates. The use of the database Y_D to construct $Y_{\mathrm{N}g}$ $(1 \times N)$ is key because it allows the network effect within Y to be estimated with only one additional unknown parameter ν_g, which is essential for datasets with relatively few biological replicates. It is noteworthy that the differential expression hypothesis using Equation 6 is an increase (or decrease) in a gene's expression conditional on the gene-network, which is a somewhat different hypothesis than standard differential expression. Additionally, we further extend the previous model of [9] by considering interactions between the network prediction and the covariate matrix by defining the $1 \times N$ element-wise multiplicative interaction term $(Y_{\mathrm{N}g} \times X)$ within

$$y_g = Y_{\mathrm{N}g} \nu_g + X' \delta_g + (Y_{\mathrm{N}g} \times X) \gamma_g + \epsilon. \tag{7}$$

The tests of the scalar $\gamma_g = 0$ correspond to the test of disruption of the gene network associated with the covariate X. By modeling the disruption of the network, we may test biologically interesting hypotheses and further improve the accuracy of predicting gene expression.

Even with large gene expression databases, regularization is required for accurate parameter estimation due to the condition $N \ll G$ because of the large number of parameters in the model for each gene. The prior model [9] used the Lasso [13] that has an L_1 penalty and produces a parsimonious model with many regression coefficients close to 0. We observed that the Lasso produced unstable regression coefficients with split sample training and validation sets, but prediction accuracy was nevertheless maintained. In a parsimonious model, some transcripts from a set of correlated transcript predictors would have non-zero coefficients whereas the other transcripts in the set would have coefficients close to zero. In this manner, the predictive weight would be unevenly shared by the correlated transcripts. However, Ruan et al [6] obtained good success in predicting gene expression by using a k-nearest-neighbor (KNN) method. The KNN method used the average of co-expressed genes as a predictor, which gives equal predictive weight to a set of correlated transcripts. We suggest a balance between unequal, *parsimonious* (Lasso) and equal, *robust* (KNN) weight among correlated transcript predictors. One such compromise is the elastic net that combines penalties on the L_1 and L_2 norms. The elastic net, however, carries a substantial burden in computational cost and risk of overfitting by optimizing over two parameters.

We offer a more efficient compromise and call our proposed approach Over-Shrinkage Ridge Regression (OSRR). The OSRR model is fit using standard ridge regression (RR) and ordinary least squares (OLS) implementations. The gene expression values from the database Y_D and the new data of interest Y are all centered at 0, which is numerically convenient. The gene expression variances are not scaled to all have equal variance because the scale of variability is biologically relevant. We are applying the network estimates from one array platform to another, which are on the log-scale, and we assume that fold-changes (i.e., a doubling of gene-expression) will be meaningful across platforms. After centering, the first step of the model fit is the estimation β_{Dg} with penalty λ based upon the database Y_D. In this approach, the regularization penalty takes the same form as standard ridge regression, and the first objective function to be minimized for each gene is

$$\|y_{Dg} - \beta_{Dg} Y_D\|^2 + \lambda \|\beta_{Dg}\|^2 \tag{8}$$

where $\|\cdot\|^2$ is the Euclidean norm. The β_{Dg} are the ridge regression coefficients with y_{Dg} as the outcome and $Y'_{D,-g}$ ($Y_{D,-g} = Y_D$without row$g$) as the covariate matrix. The identifiability of β_{Dg} when $\lambda > 0$ follows from the theory of ridge regression [14]. As $\lambda \to \infty$, the regression parameter estimates β_{Dg} will tend towards 0 as the penalty becomes the dominant term. Hence, the larger the λ, the more the $\beta_{Dg} Y_D$ is biased towards the null compared to the observed database values y_{Dg}. However, we introduce a second step and add another parameter ν_g to counteract the bias towards the null in predicting the data from a small experiment Y with rows y_g. The second step is the estimation of ν_g conditional on the β_{Dg} that minimizes $\|y_g - \beta_{Dg} Y \nu_g\|^2$ using ordinary least squares. If we let $Y_{Ng} = \beta_{Dg} Y$, then the estimation of ν_g can be seen as the univariate OLS coefficient of Y regressed onto Y_{Ng}. That is, the second objective function operates on the new data Y, and the estimate of row y_g is $\beta_{Dg} Y \nu_g$ where ν_g minimizes

$$\|y_g - Y_{Ng} \nu_g\|^2 \tag{9}$$

using ordinary least squares. The introduction of ν_g within a second stage of the regression has two effects. First, $\beta_g^{OSRR} = \beta_{Dg} \nu_g$ is no longer biased towards the null as $\lambda \to \infty$. The larger λ might make $\beta_{Dg} Y$ smaller than y_g by an order of magnitude, which is the motivation for the term “Over-Shrinkage". Second, larger values of λ will act to equalize the elements of β_g corresponding to correlated transcripts. This equalization property produces a similar effect as KNN's equal weighting, but OSRR still allows data-driven deviation from equality and allows genes that are not the nearest neighbors to influence prediction. Hence, we call this a global network adjustment. An important difference between OSRR and approaches such as ridge regression, LASSO, and the elastic net is that these methods penalize the magnitude of the parameter β norms with some function $P(\beta)$ whereas the OSRR approach considers a link function between the linear predictor $X\beta$ and the outcome Y. The link function $g(X\beta)$ is the second regression so that the model becomes $Y = g(X\beta)$, which may serve to minimize the effect of poor prediction.

The robustness of the OSRR method to variable gene regulatory networks and applicability to smaller sample sizes are its attractive features. If the gene expression network estimate B_D from Y_D results in a prediction Y_{Ng} that is uncorrelated with the smaller independent dataset Y, then the regression parameter ν_g converges to 0 under standard linear model theory. This property is quite distinct from the use of the LASSO model by [9] that does not consider larger training sets for estimating B or variable gene-gene dependencies due to the perturbation or biological state. The selection of λ is based upon the maximum correlation *across all genes* between the database prediction $\beta_{Dg} Y$ and the current experiment Y. Overfitting may result because the OSRR is applied to the same data it is tuned with (i.e., λ is selected based upon Y). OSRR largely avoids overfitting and gains simplicity by using a single tuning parameter λ for all transcripts. The method was implemented in the R statistical software [15], and we fitted the model in equation 6 with the lme4 R package [16]. A random intercept was used to account for the correlation within case-control pairs. The programs, example simulated data, and a demonstration of the penalty selection are available for download from the supplemental website as Additional file 1.

Simulation studies

We performed a simulation study to illustrate the properties of OSRR in the context of correlated covariates with similar effects on a continuous outcome. Here, we consider a single outcome y ($n \times 1$) with two independent sets of correlated predictors X_1 and X_2 both with dimension ($N \times G$). The $N = 201$ rows of X_1 and X_2 are drawn from a compound-symmetric, multivariate Gaussian distribution with dimension $G = 100$, mean 0 and variance 1.0 with correlation $\rho = 0.9$. Both X_1 and X_2 represent two correlated sets of genes that have an influence on the expression of the gene y. Within these sets, the influence of the genes is the same, which might reflect the activity of two pathways that are either associated with an increase or an inhibition of the expression of y. That is, $y = X_1 B_1 + X_2 B_2 + \epsilon$ where $B_1 = [\beta_1 \ldots \beta_1]'$ and $B_2 = [\beta_2 \ldots \beta_2]'$ are $G \times 1$ with identical elements $\beta_1 = 1.0$ and $\beta_2 = -1.0$, respectively. The errors ϵ have a Gaussian distribution with variance 1.0. The data were fit using 3 models: ordinary least squares (OLS), Elastic Net, and the OSRR with $\lambda = 10{,}000$. The elastic net penalty was selected using K=5 fold cross-validation. The results are shown in Figure 1, which is a network representation of the coefficient estimates. All models perform similarly in terms of prediction. The estimates of B_1 and B_2 are highly variable for OLS. The elastic net predicts as well as other models, but the estimated coefficients are highly variable with some coefficients clustered around 0. To offset the Lasso parameter estimates close to 0, the remaining coefficients are estimated to be larger in magnitude than the true values. The OSRR exhibits a strong smoothing property that equalizes the influence that each of the correlated predictors has, whereas the elastic net and the Lasso tend to reduce the number of influential predictors. This is an advantage in that it provides robustness,

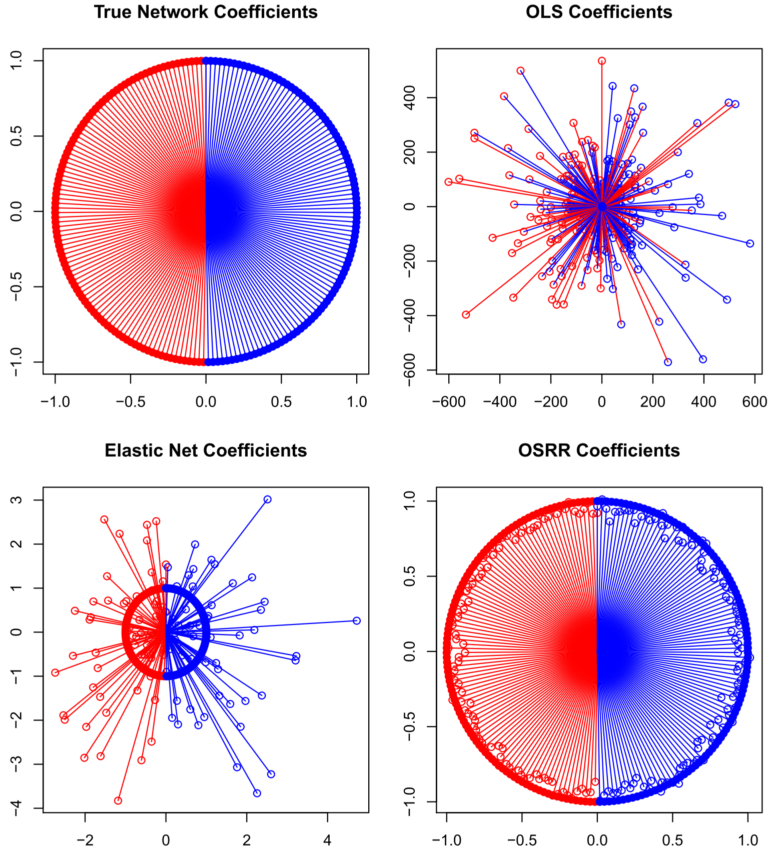

Figure 1 Network depiction of regression estimation. The regression coefficients are shown as rays from the origin. The absolute value of the coefficient is the distance from the origin and the sign is indicated by blue (+) or red (-). There are 200 rays indicating 200 estimated coefficients, and the true values of the coefficients (+1,-1) appear as rings. The Ordinary Least Squares (OLS) estimates are highly variable because of the small sample size and the correlation between the positive and negative predictor groups. The elastic net has less variable estimates than OLS, but the OSRR coefficients demonstrate the the strong smoothing property of that equalizes the influence of predictors.

but when there are sparse predictors or there is sufficient sample size to estimate the coefficients more precisely, then we anticipate that the OLS, elastic net or Lasso are advantageous.

As a proof of principle, we performed another simulation study to demonstrate the advantage of the OSRR methodology applied to differential expression compared to a naive approach that does not account for prior knowledge about correlations between transcripts. We simulated a database Y_D with $G = 200$ and $N = 100$. The transcripts were given a block diagonal correlation structure with 2 blocks of size 100 transcripts each with compound symmetry having variance 1 and correlation $\rho = 0.8$. Although this is a relatively simple multivariate Gaussian model compared to actual gene networks, it still represents a large dependency between genes. The OSRR model was fitted as described to estimate B_D with $\lambda = 1000$ selected out of the set $\{10^2, 10^3, 10^4, 10^5, 10^6\}$ based upon the out-sample prediction from training and validation sets. We then simulated a smaller ($N = 20$) dataset Y with the same correlation structure. Ten of the arrays were under the control condition ($x_i = 0$), and 10 were under a treatment condition ($x_i = 1$). Differential expression was simulated by randomly selecting genes with probability 0.2 and then adding or subtracting $\delta_g = 1.0$ to those genes. We compared the performance of the model in equation 6 to a model without correction for the network relationships $Y_{\mathrm{N}g}$. This naive model reduces to the t-test. The results are shown in Figure 2. The histograms for p-values corresponding to the null hypothesis ($\delta_g = 0$) are given for both models, and the OSRR model clearly has smaller p-values for the differentially expressed (DE) genes compared to the naive model. Also, the naive model p-values do not have a uniform distribution for the not differentially expressed genes as the OSRR does. The lack of uniform distribution of the null p-values is due to the dependence structure and could inflate the Type I error. This is consistent with higher sensitivity and specificity of OSRR relative to naive methods that do not utilize prior information on gene networks.

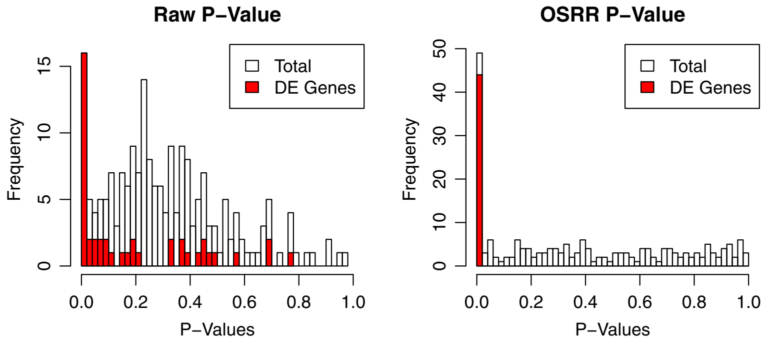

Figure 2 Proof-of-concept simulated differential expression experiment. Histograms of p-values are shown for both differentially expressed genes (DE) and not DE genes that follow the null hypothesis. The raw (naive) model has p-values for not DE genes that do not have the expected uniform distribution, though the p-values corresponding to the differentially expressed (DE) genes tend to be low. The OSRR model applied to the same dataset more clearly separates the DE from the not DE genes, and the distribution of the not DE genes is more uniform.

Empiricial studies

Chromosomal deletion study

In our chromosome deletion study, we investigated the effects of 18q chromosomal deletions (deleted segments were the $60,021,550 - 76,117,153$ bps of chromosome 18) on cell lines derived from blood samples called lymphoblastoid cell lines (LCLs). There were 10 case-control (c-c) pairs with the case subject having a deletion on chromosome 18 (18q-), and the control was a normal genotype, same-sex parent giving a total of $N = 20 = 2 \times 10$ subjects. We assayed each c-c pair with 4 Agilent 44K expression arrays. We used a dye-swap design such that for 2 arrays the case was labeled with Cy3 (control labeled with Cy5) and 2 with case labeled with Cy5 (control labeled with Cy3). Each case had associated clinical phenotypes including height, weight, age, and various laboratory measures of growth hormone responsiveness. The log-expression data for a given subject and gene are denoted as y_{grs} where the indices g, r, and s represent gene, replicate, and spot respectively. The Cy3 and Cy5 channels were treated separately, and we did not use the log-ratio. For each subject, we averaged the replicates and the spots corresponding to the same gene yielding the $N \times 1$ vector y_g. We did not model the dye factor because it was balanced and not predictive. There is a corresponding $N \times 1$ covariate matrix X where $x_i = I(i^{th}\text{subject is a case})$. We also examined the p-values of the δ_g parameter from the model compared to a standard analysis with the LIMMA software [12].

For Y_D, we used a much larger dataset from [17] who studied similar LCL cells that were treated with a variety of chemotherapies. There were a total of 374 Affymetrix chip assays in this dataset. We included all genes that were within the deleted region, but otherwise restricted our analysis to genes that had gene symbols in common to both the Affymetrix and the Agilent platforms and were expressed in the Affymetrix data. This leaves a total of 5,035 genes to estimate B_D. The expression values of the database genes were centered, but not scaled. A key issue is how to select the penalty parameter λ, and the selection process can be extremely computationally intensive as [9] used 200 CPU days for their cross-validation Lasso approach. To greatly simplify the selection of the penalty and minimize the tuning and hypothesis testing on the same dataset, we chose one penalty parameter for all genes and took a split sample approach so that the database was divided into two sets of 187 arrays. Further, we took into account three aspects of goodness-of-fit in assessing cross-validation. First, we considered prediction among the split samples and second, we considered the stability of the regression parameters B_D between the split samples. Third, we examined the prediction of B_D with respect to our independent experiment. The computation time for selecting λ took approximately 151,445 CPU seconds (8GB RAM, 2 Ghz) or about a 100 fold decrease in computing time relative to the Lasso approach for a similar sized dataset and computing environment [9]. The Based upon these three criteria, we selected $\lambda = 10^4$ out of the set $\{10^2, 10^3, 10^4, 10^5, 10^6\}$, and we applied the corresponding $\hat{B}_D$ to construct predictors $Y_{\mathrm{N}g} = \hat{\beta}_{Dg}Y$ for y_g from our expression experiment. Lastly, to further assess the robustness and Type I error, we performed a permutation analysis comparing OSRR and a conventional method. We permuted the case-control labels 100 times, and compared the histograms of the p-values from both methods.

Results

We observed that the network predictions $Y_{\mathrm{N}g}$ based upon $\hat{B}_D$ estimated from the database Y_D were strongly correlated with new data y_g. In Figure 3, the median correlation of $Y_{\mathrm{N}g}$ with y_g is 0.54 so that 25% $\approx \sqrt{0.54}$ of the variance of transcription regulation is explained. Note that the correlation of the observed expression with

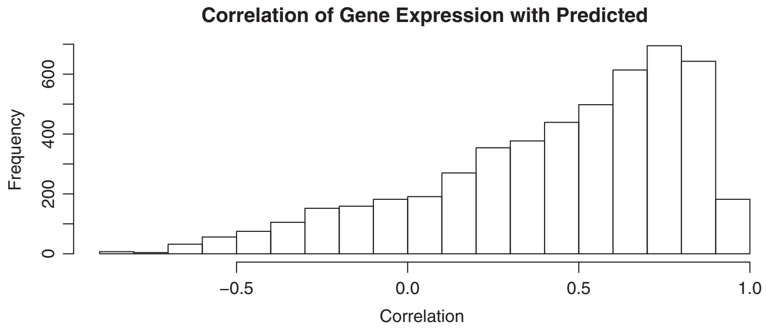

Figure 3 Histogram of correlations of the predictor Y_{Ng} based upon regulatory network from database Y_D with gene expression in an independent experiment Y. The positivity of the center of the distribution implies a dependency of gene expression that is consistent with the OSRR model prediction.

the network predictor is expected to be $>$ 0, and the mean correlation is significantly greater than 0 (p $<$ 0.0001). The degree of prediction is remarkable because Cosgrove et al [9] originally studied bacteria, whereas these data indicate that more complex human transcriptional patterns even on different microarray platforms are largely predictable. Nevertheless, for some genes, there is a negative correlation consistent with random noise. This underscores the importance of the estimation of ν_g in equation 6. That is, if the network model does not fit the smaller dataset, then ν_g is estimated to be close to 0 so that its effects are mitigated.

Figure 4, shows the distribution of p-values for differential expression of transcripts under both the standard LIMMA analysis and OSRR models. Under the null hypothesis p-values are uniformly distributed between 0 and 1. The p-values of the LIMMA model do not appear to be for from uniformly distributed, but the OSRR model not only has more significant (p $<$ 0.05) values, it also has a clear "spike" of small p-values towards 0. It is possible to model these types of distributions with a beta and uniform mixture (BUM) as suggested by Pounds et al [18] where the beta component models the spike of p-values that corresponds to differentially expressed transcripts and the uniform component corresponds to the transcripts that follow the null hypotheses. The BUM fit is shown in Figure 4, and the ratio of the the uniform (blue) component over the mixture density (green) is the empirical Bayes probability estimate of the false positive rate. The comparison of numbers of genes that at given p-values cutoffs is shown in Table 1. Clearly, the OSRR model identifies more than double (851 vs. 337) the number of significant transcripts, and those transcripts selected by the LIMMA approach are also selected by OSRR at a rate of 95% (319/337). The correlation between the $-\log$(p-values) from LIMMA and OSRR is 0.88, which indicates fairly good agreement in ranking the genes by probability of differential expression. The mean value for the false discovery rate (FDR) of the 21 genes in the deleted region in the chromosome according to LIMMA was 0.51 compared to the mean FDR of the q-value adjusted OSRR p-values of 0.17, indicating that the OSRR model is more likely to identify the deleted genes as differentially expressed.

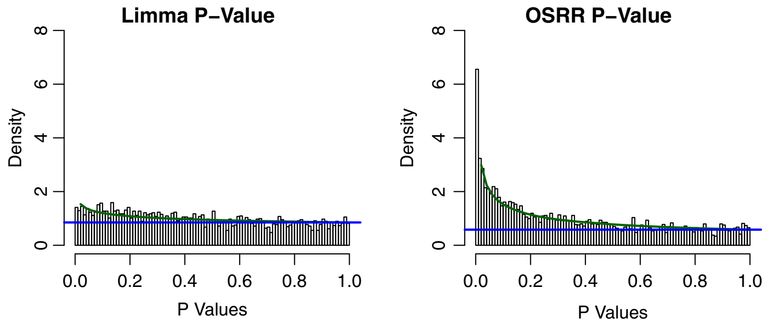

Figure 4 Histogram of p-values generated by conventional LIMMA and OSRR methods applied to the chromosomal deletion study. Note that LIMMA produces p-values that differ only slightly from a uniform distribution, but the OSRR method gives more significant p-values. The beta-uniform mixture model is shown in green and the uniform component is blue. The ratio of blue/green is the empirical Bayes probability of the false positive rate at a given p-value. This ratio approaches 0.5 for conventional analysis and it approaches 0.25 for the OSRR method.

Table 1 Number of significant transcripts in standard LIMMA and OSRR models at different p-value cutoffs

Up/Down	P-Value	LIMMA	OSRR	Both
Down	0.0500	94	209	92
	0.0100	25	79	25
	0.0010	9	19	9
	0.0001	5	8	5
Up	0.0500	241	642	227
	0.0100	46	251	42
	0.0010	6	52	5
	0.0001	2	15	2

Up/Down corresponds to higher/lower expression in the cases relative to controls.

Not only was the OSRR method shown to identify twice the number of differentially expressed genes, but the log fold-change estimates are more stable, and therefore less likely to be false positives based upon the split-sample analysis. We randomly split the data set 20 times into two sub-datasets of 5 families each, and compared the results of OSRR and the t-test from the two sub-datasets of the partition in estimating the log-fold change between cases and controls. Note that LIMMA is equivalent to the t-test in estimating log-fold change. Table 2 lists the correlations between the in-sample and out-sample comparisons of the differential expression estimates of the conventional model δ_g^{Std} and the OSRR model δ_g. The OSRR estimate was much more reproducible than the conventional estimate in independent samples with correlations of 0.321 (OSRR) vs. 0.201 (Standard) (95% CI for difference $[\,0.054, 0.183]$, $p < 0.001$). Furthermore, the out-sample results clearly show that OSRR is more consistent with both OSRR and conventional estimates in independent split datasets.

We compared the performance of the elastic net predictions with OSRR using a subset of 500 genes because of the computational time of the elastic net in this context. We selected a gene-specific penalty as did Cosgrove et al [9] who used the Lasso. We compared the average correlation of the predictions with the expression of the 500 genes in our smaller dataset, and we found that the mean correlation was 0.01 95% CI $[-0.01, 0.03]$ (paired t-test: $p = 0.43$) higher for the OSRR model. This is consistent with OSRR allowing equal prediction at a fraction of the computational cost of the elastic net.

Table 2 Reliability assessed by a repeated split-sample analysis

In/Out Sample	Type	Mean correlation	Std Dev
In-Sample	δ_g vs δ_g^{Std}	0.917	0.034
Out-Sample	δ_g^{Std} vs δ_g^{Std}	0.203	0.115
Out-Sample	δ_g vs δ_g	0.321	0.073
Out-Sample	δ_g vs δ_g^{Std}	0.313	0.089

The parameters δ_g and δ_g^{Std} denote the differential expression estimates for OSRR and standard methods.

The permutation study results are shown in Figure 5. The resulting estimates of the null distributions are different for the conventional and the OSRR model. The LIMMA model has a conservative Type I error rates indicated by nonuniform p-value distributions with a dip near 0, and the OSRR model has a slightly inflated Type I error with more smaller p-values than uniform. The OSRR model p-values are quite different from the non-permuted data in Figure 4. In contrast, the conventional analysis null distribution is more similar to the non-permuted distribution of p-values, which suggests a low signal-to-noise ratio. This is consistent with the OSRR model having a reasonably controlled Type I error rate, but much greater power to detect differences between the cases and controls.

We also applied the interaction model in Equation 7. A substantial proportion (10%, 484/5035) of genes had interactions ($H_0 : \gamma_g = 0$ with $p < 0.01$) between the dependence on the gene network and case-control status. This implies that the regulatory relationships are different for particular genes between cases and controls. The gene with the most significant interaction effect was the GARS gene, which has been linked to severe neurological conditions [19]. Interestingly, 93% of the 100 most significant interactions were negative ($\hat{\gamma}_g < 0$), which implies that the expression predictions are substantially less positively correlated with the observed expression in the abnormal cases than controls. See Additional file 2. This is consistent with overall disruption of the regulatory networks in the cases. Such observations would be impossible using standard differential expression models that only examine mean differences.

For the purpose of comparison, we applied the method of Cosgrove et al [9] to this dataset without the use of the larger training dataset using the elastic net estimators. Their method was not intended to be applied to datasets of this size (2 conditions (case/control) and 10 paired samples), but we wanted to investigate the necessity of using the database in a real example. Despite 5-fold cross-validation for the selection of the tuning parameters, there was overfitting of the gene network so that the estimates BY from the fit had a median correlation with Y of > 0.95 and the analysis of the residuals suggested in Equation 5 yielded p-values that did not differ from a uniform distribution (Data not shown). No differentially expressed gene could be identified due to the confounding of the gene expression network BY and the case-control status. However, we emphasize that their method was not intended for small datasets.

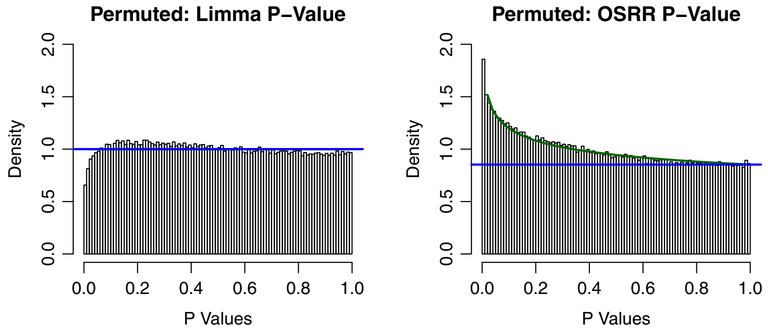

Figure 5 Histogram of the p-values of the from the permuted chromosome deletion study. The conventional analysis method has a slightly conservative, non-uniform null distribution. Whereas, the OSRR method has slightly anti-conservative, non-uniform null distribution.

Circadian rhythm study

As a proof of principle, we considered another dataset to see if the OSRR model had similar advantages in a mouse model. This dataset [20] consisted of a time-series of brain tissue collected in mice with two different genotypes (wild type and a mutation of the circadian rhythm "clock" gene). The brain tissue samples were measured on Affymetrix tissue arrays in duplicate every two hours for 44 and 24 hours for a total of 24 and 14 arrays, respectively, for wild type and mutant mice. We estimated the B matrix using a database Y_D from an experiment with the same Affymetrix platform with 122 samples of mixed tissue types [21]. We assessed the 3,000 genes with the highest variance in the circadian dataset. We fit a sinusiodal model with and without network adjustment term $Y_{\mathrm{N}g}\nu_g$ to identify genes associated with the circadian rhythm

$$y_g = Y_{\mathrm{N}g}\nu_g + a_g\cos(T + b_g) + d + \epsilon \tag{10}$$

where T $(1 \times N)$ is the time covariate in radians, d is an intercept, and the time dependencies are tested with scalars a_g and b_g ($H_0 : a_g = b_g = 0$) using a log-likelihood test assuming that ϵ is Gaussian noise. We did not use the information about the genotype of the mice within the models. Rather, we tested whether or not the OSRR model could account for the biological variation due to different genetic backgrounds and still identify time dependent genes. The results are shown in Figures 6 and 7. Figure 6 shows the unpermuted data analysis in both models with the OSRR model having a higher proportion of significant genes. Figure 7 shows the null distributions under permutations, and the null distributions of the two models are very similar. This indicates that the OSRR model has more power to detect time-dependent genes with a similar Type I error rate as the conventional model. We also compared the performance of the elastic net and the OSRR gene expression predictions by the average correlations of the network predictors $Y_{\mathrm{N}g}$ with the outsample gene expression y_g for a random subset of 200 genes. We found that the OSRR method had a 0.1 higher average correlation (95% CI [0.05, 0.15] with a paired t-test: $p = 0.0003$). That is, the OSRR had better prediction that the elastic net model for less computational costs; however, this advantage in prediction may be due to the smaller sample size of the training set ($N = 122$) relative to the chromosomal deletion training set ($N = 374$) for which the prediction of the two methods was approximately equivalent.

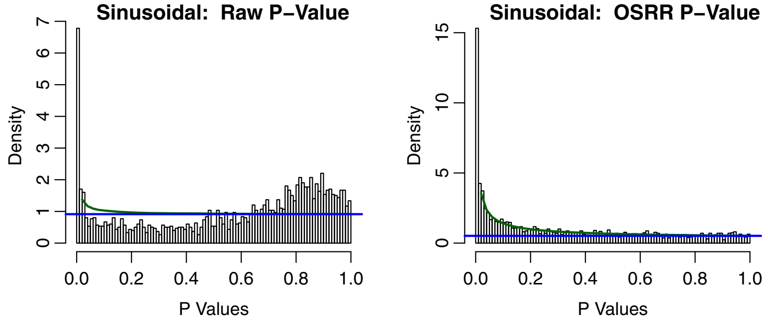

Figure 6 Histogram of p-values generated by conventional *raw* and OSRR method of the circadian rhythm study. The beta-uniform mixture model is shown in green and the uniform component is blue. The OSRR model identifies approximately double the number of significant genes.

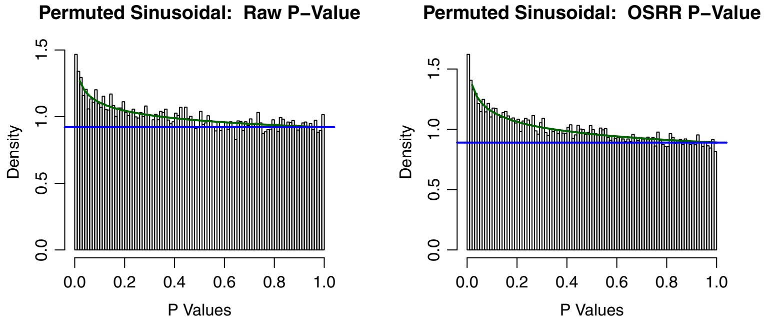

Figure 7 Histogram of the p-values of the from the permuted circadian rhythm study. Both the conventional analysis and the OSRR method have slightly non-uniform null distributions.

Discussion

We have utilized the estimates of gene-networks from large databases in the analysis of small, independent datasets assayed on a different microarray type to successfully predict about 25% of the variation in transcript expression. The prediction is shown to markedly increase the sensitivity and the reliability of detecting differentially expressed transcripts in two different datasets. These inferences are different from the standard differential expression analyses because they reflect an adjustment based upon a regulatory model for each gene, and by testing an interaction term, one may make a statistical assessment of whether the regulatory network has changed between biological states. This method is approximately 100 times computationally faster than the previously reported method [9]. The majority of the computational time is spent fitting predictive models for each gene independently, and this time is multiplied by the density of the grid of the tuning parameter(s). OSRR has the advantages that the tuning parameter grid can be sparse and that the algorithms for fitting linear models are relatively well-optimized. The computation can be accelerated by parallelizing the fit, but parallel analyses are nontrivial because of large memory requirements of the operations (i.e., solving linear equations with 5,000 variables). Attention should also be paid to improving the computational efficiency of more complex models such as the elastic net.

Unlike the previous model, the use of the network predictor as a covariate in the differential expression regression model also provides robustness against poor prediction of particular transcripts. Given OSRR's ease of implementation and its robustness, there is a broad set of potential applications to small sample size expression experiments that leverage the growing large-scale gene expression databases such as the Gene Expression Omnibus (GEO) [22]. We recommend that researchers select relevant datasets to construct network models by considering the species, type of tissue or cells, the microarray platform, and other sources of variation. We have seen that if the database Y_D used to estimate B_D has a different tissue type than the independent dataset Y, then the method tends to be less effective in accounting for variation.

Conclusion

We have developed a novel adaptation of ridge-regression called OSRR that robustly estimates models of transcriptional co-expression networks based upon large microarray experiments. There are many possibilities for future research. Because of the robustness across microarray platforms, the OSRR approach can applied to RNA sequencing (RNA-Seq) data as well. In preliminary studies, we have found that the predictions of OSRR derived from microarray data perform similarly when applied to RNA-seq data. As we suggested with Equation 7, we can consider tissue specific modulations of regulatory networks as an extension of the OSRR model. Different tissue types or experimental conditions may induce different correlations between genes, and this fact is utilized in the COXEN model in [23] for disease classification purposes. These authors found that genes have different correlations in different cell types, and that genes that have shared correlation structures between two subtypes can predict how those subtypes will respond to chemotherapy. Also, the steps in the two-step fit are suggestive of levels within a hierarchical model. We can use this framework to extend the prediction model by including extra terms. For example, some genes may be better predicted using another network model with prediction $Y_{\mathrm{N2}\,g}$. To account for this, we propose

$$y_g = Y_{\mathrm{N}g}\nu_{\mathrm{N}g} + Y_{\mathrm{N2}\,g}\nu_{\mathrm{N2}\,g} + X\delta_g + \epsilon. \tag{11}$$

If the other network model is a superior predictor, then this will be reflected by the $\nu_{\mathrm{N2}\,g}$ parameter. However, the more terms within the model, the more biological replicates are required for model stability. We may include information a known network of specific genes in a similar manner. We could further address another aspect of the SSEM problem. That is, the structure among the rows of B was not modeled. For example, if a gene had a relatively large influence on many other genes, then one could use this pattern for better estimation of B. This may potentially be achieved using a similar approach of Friedman et al [24] in the construction of the graphical Lasso.

Additional files

Additional file 1: File with link to R implementation programs.

Additional file 2: File with estimation results from chromosome 18 study for alterations in gene networks in cases relative to controls.

Competing interests

All of the authors deny competing interests, financial or otherwise, related to this research.

Authors' contributions

JAG helped conceive the methodology and performed the data analysis and simulation studies. JGI, MG, and MHC helped conceive and refine the methodology. JDC conceived and conducted the chromosomal deletion experiment and assisted in the interpretation of the experiment. All authors read and approved the final manuscript.

Acknowledgements

Research reported in this publication (JAG, JGI, MHC, MG) was supported by the National Institute of the General Medical Sciences of the the National Institutes of Health under Award Number 2R01GM070335-13A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Funding for the experiment, JDC, and JAG were supported by The Chromosome 18 Registry and Research Society.

Author details

[1]Department of Epidemiology and Biostatistics, University of Texas Health Science Center San Antonio, San Antonio, Texas, USA. [2]Department of Biostatistics, University of North Carolina School of Public Health, Chapel Hill, North Carolina, USA. [3]School of Mathematics and Statistics, University of Glasgow, Glasgow, Scotland, UK. [4]Department of Statistics, University of Connecticut, Storrs, Connecticut, USA. [5]Department of Pediatrics, University of Texas Health Science Center San Antonio, San Antonio, Texas, USA.

Received: 11 March 2013 Accepted: 25 July 2013
Published: 23 August 2013

References

1. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G: **Towards a rigorous assessment of systems biology models: the DREAM3 challenges.** *PLOS ONE* 2010, **5**(2):e9202.
2. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A: **NCBI GEO: archive for functional genomics data sets-10 years on.** *Nucleic Acids Res* 2011, **39**(Suppl 1):D1005—D1010.
3. Schoumans J, Ruivenkamp C, Holmberg E, Kyllerman M, Anderlid B, Nordenskjold M: **Detection of chromosomal imbalances in children with idiopathic mental retardation by array based comparative genomic hybridisation (array-CGH).** *J Med Genet* 2005, **42**(9):699–705.
4. Dahl DB, Newton MA: **Multiple hypothesis testing by clustering treatment effects.** *J Am Stat Assoc* 2007, **102**(478):517–526.
5. Leek JT, Storey JD: **A general framework for multiple testing dependence.** *Proc Natl Acad Sci* 2008, **105**(48):18718–18723.
6. Ruan J: **A top-performing algorithm for the DREAM3 gene expression prediction challenge.** *PLOS ONE* 2010, **5**(2):e8944.
7. Gustafsson M, Hornquist M: **Gene expression prediction by soft integration and the Elastic Net—Best performance of the DREAM3 gene expression challenge.** *PLoS ONE* 2010, **5**(2):e9134.
8. Zou H, Hastie T: **Regularization and variable selection via the elastic net.** *J R Stat Soc Ser B* 2005, **67**(2):301–320.
9. Cosgrove EJ, Zhou Y, Gardner TS, Kolaczyk ED: **Predicting gene targets of perturbations via network-based filtering of mRNA expression compendia.** *Bioinformatics* 2008, **24**(21):2482–2490.
10. Zellner A, Theil H: **Three-stage least squares: simultaneous estimation of simultaneous equations.** *Econometrica* 1962, **30:**54–78.
11. Newton M, Noueiry A, Sarkar D, Ahlquist P: **Detecting differential gene expression with a semiparametric hierarchical mixture method.** *Biostatistics* 2004, **5**(2):155–176.
12. Wettenhall J, Smyth G: **limmaGUI: a graphical user interface for linear modeling of microarray data.** *Bioinformatics* 2004, **20**(18):3705–3706.
13. Tibshirani R: **Regression shrinkage and selection via the Lasso.** *J R Stat Soc Ser B-Methodological* 1996, **58**(1):267–288.
14. Hoerl AE, Kennard RW: **Ridge regression: biased estimation for nonorthogonal problems.** *Technometrics* 1970, **12:**55–67.
15. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2012.
16. Bates D, Maechler M, Bolker B: *lme4: Linear Mixed-Effects Models Using S4 Classes*; 2011. [R package version 0.999375-39]. [http://CRAN.R-project.org/package=lme4]. Accessed 01/03/2013
17. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, De Jager PL, Shaw SY, Wolfish CS, Slavik JM, Cotsapas C, Rivas M, Dermitzakis ET, Cahir-McFarland E, Kieff E, Hafler D, Daly MJ, Altshuler D: **Genetic analysis of human traits in vitro: drug response and gene expression in Lymphoblastoid cell lines.** *PLOS Genet* 2008, **4**(11):e1000287.
18. Pounds S, Morris S: **Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of p-values.** *Bioinformatics* 2003, **19**(10):1236–1242.
19. Antonellis A, Ellsworth R, Sambuughin N, Puls I, Abel A, Lee-Lin S, Jordanova A, Kremensky I, Christodoulou K, Middleton L, Sivakumar K, Ionasescu V, Funalot B, Vance J, Goldfarb L, Fischbeck K, Green E: **Glycyl tRNA synthetase mutations in Charcot-Marie-Tooth disease type 2D and distal spinal muscular atrophy type V.** *Am J Human Genet* 2003, **72**(5):1293–1299.
20. Siepka SM, Yoo SH, Park J, Song W, Kumar V, Hu Y, Lee C, Takahashi JS: **Circadian mutant overtime reveals F-box protein FBXL3 regulation of CryptochromeandPeriodGene expression.** *Cell* 2007, **129:**1011–1023.
21. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, et al.: **A gene atlas of the mouse and human protein-encoding transcriptomes.** *Proc Natl Acad Sci U S A* 2004, **101**(16):6062–6067.
22. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles - database and tools update.** *Nucleic Acids Res* 2007, **35**(Sp. Iss. SI):D760–D765.
23. Smith SC, Baras AS, Lee JK, Theodorescu D: **The COXEN principle: translating signatures of In vitro chemosensitivity into tools for clinical outcome prediction and drug discovery in cancer.** *Cancer Res* 2010, **70**(5):1753–1758.
24. Friedman J, Hastie T, Tibshirani R: **Sparse inverse covariance estimation with the graphical lasso.** *Biostatistics* 2008, **9**(3):432–441.

doi:10.1186/1471-2105-14-258
Cite this article as: Gelfond *et al.*: Differential expression analysis with global network adjustment. *BMC Bioinformatics* 2013 **14**:258.