Supplementary Document for the Manuscript entitled: A Big Data Pipeline: Identifying Dynamic Gene Regulatory Networks from Time Course GEO Data with Applications to Influenza Infection

Michelle Carey and Juan Camilo Ramírez and Shuang Wu and Hulin Wu

1 Article template produced by the pipeline for GSE52428 subject 1

Title of submission

Name1 Surname[1,2], Name2 Surname[2], Name3 Surname[2,3], Name4 Surname[2], Name5 Surname[2‡], Name6 Surname[2‡], Name7 Surname[1,2,3*], with the Lorem Ipsum Consortium

1 Affiliation Dept/Program/Center, Institution Name, City, State, Country
2 Affiliation Dept/Program/Center, Institution Name, City, State, Country
3 Affiliation Dept/Program/Center, Institution Name, City, State, Country

These authors contributed equally to this work.
‡These authors also contributed equally to this work.
Current Address: Dept/Program/Center, Institution Name, City, State, Country
* correspondingauthor@institute.edu

Abstract

Please add abstract here

Introduction

Please review the below GEO Citation(s) and and add an Introduction here. Woods CW, McClain MT, Chen M, Zaas AK et al. A host transcriptional signature for presymptomatic detection of infection in humans exposed to influenza H1N1 or H3N2. PLoS One 2013;8(1):e52198. PMID: 23326326

Methods

Pre-processing

Affymetrix Genechip® arrays are currently among the most widely used high-throughput technologies for the genome-wide measurement of expression profiles. To minimize mis- and cross-hybridization problems, this technology includes both perfect match (PM) and mismatch (MM) probe pairs as well as multiple probes per gene (Lipshutz et al., 1999). As a result, significant pre-processing is required before an absolute expression level for a specific gene may be accurately assessed. In general, pre-processing probe-level expression data consists of three steps: background adjustment (remove local artifacts and "noise"), normalization (remove array effects), and summarization at the probe set level (combine probe intensities across arrays to obtain a measure of the expression level of corresponding mRNA).

Detect the Dynamic Response Genes (DRGs)

(a) Obtain the estimated gene expression curves

We assume that the centered expression levels of the i^{th} gene, belonging to subject j, denoted here by $x_{i,j}$, is a smooth function over time t and that the centered gene expression measurement $\tilde{y}_{i,j}$ is a discrete observation from this smooth function, which has been distorted by noise, *i.e.*,

$$\tilde{y}_{i,j} = x_{i,j}(t_k) + \epsilon_{i,j},$$

for $i = 1, \dots, n$, $j = 1, \dots, N$ and $k = 1, \dots, K_{i,j}$, where n is the number of genes, N is the number of subjects (or experimental conditions), $K_{i,j}$ is the number of time points observed for the i^{th} gene, belonging to subject j. The noise $\epsilon_{i,j}$ is assumed to be an independently identically distributed (i.i.d.) Gaussian random variable with mean 0 and variance σ^2.

The $K_{i,j} \times 1$ vector of the estimated centered expression levels evaluated at the points $\mathbf{t}$, for the i^{th} gene, belonging to subject j, $\hat{\mathbf{x}}_{i,j}$ is obtained by spline smoothing [1, 2]. This approach approximates $\mathbf{x}_{i,j}$ by a linear combination of L independent basis functions, $\mathbf{x}_{i,j} \approx \sum_{l=1}^{L} \mathbf{b}_{i,j,l} c_{i,j,l} = \mathbf{B}_{i,j}\mathbf{c}_{i,j}$, where the $K_{i,j} \times L$ matrix $\mathbf{B}_{i,j}$ denotes the basis functions evaluated at time $\mathbf{t}$ and the vector $\mathbf{c}_{i,j}$ provides the corresponding coefficients.

The coefficients $\mathbf{c}_{i,j}$ can be estimated by minimizing

$$[\tilde{\mathbf{y}}_{i,j} - \mathbf{B}_{i,j}\mathbf{c}_{i,j}]'[\tilde{\mathbf{y}}_{i,j} - \mathbf{B}_{i,j}\mathbf{c}_{i,j}] + \lambda_j \mathbf{c}'_{i,j} \underbrace{\left[\int \frac{\mathrm{d}^2 B_{i,j}(t)}{\mathrm{d}t^2} \frac{\mathrm{d}^2 B_{i,j}(t)}{\mathrm{d}t^2} \mathrm{d}t\right]}_{\mathbf{R}_{i,j}} \mathbf{c}'_{i,j} \qquad (1)$$

where the first term defines the squared discrepancy between the observed centered gene expression measurements $\tilde{\mathbf{y}}_{i,j}$ and the estimated measurements $\hat{\mathbf{x}}_{i,j}$, and the second term containing the $L \times L$ matrix $\mathbf{R}_{i,j}$ which is the inner product of the second derivative of the basis functions penalizes the curvature of $\hat{\mathbf{x}}_{i,j}$ and hence requires it to be sufficiently smooth. The parameter λ_j controls the trade-off between the fit to the data and the smoothness requirement and hence ensures that $\hat{\mathbf{x}}_{i,j}$ has an appropriate amount of regularity. All the genes for each subject are assumed to have the same λ_j. The minimum of (1) for fixed λ_j is

$$\hat{\mathbf{c}}_{i,j} = (\mathbf{B}'_{i,j}\mathbf{B}_{i,j} + \lambda \mathbf{R}_{i,j})^{-1}\mathbf{B}'_{i,j}\tilde{\mathbf{y}}_{i,j}$$

and the estimated centered expression levels are $\hat{\mathbf{x}}_{i,j} = \mathbf{B}_{i,j}\hat{\mathbf{c}}_{i,j}$.

We expect that only a small fraction of genes respond to the external stimulus and the majority of the genes have no significant response with relatively flat expression levels over time. Therefore, estimating the parameter λ_j using the conventional method of minimizing the prediction error with generalized cross validation (`GCV`), see [3] for details, of all the genes together is not ideal as `GCV` will tend to select a λ_j that is large to minimize the prediction error of the majority of unresponsive genes. As we are interested in obtaining an appropriate amount of regularity for the responsive genes, we apply an approach similar to [4] and [5] and choose a subset of the genes that exhibit time course response patterns with relatively smooth trajectories that do not fluctuate widely. Then we rank these genes by their interquartile range and select 200 of the top ranking genes as our estimation subset. The regularity parameter λ_j is estimated by minimizing the `GCV` of the responsive genes in our estimation subset, this parameter is then used to smooth the time course data for all the genes.

(b) Perform a hypothesis test to identify the genes with expressions that change significantly over time

Dynamic response genes (DRGs) can be defined as genes with expressions that change significantly over time. In order to determine which genes can be considered DRGs, we use an F-statistic which compares the goodness-of-fit of the null hypothesis $H_0 : \hat{\mathbf{x}}_{i,j} = 0$ versus the alternative hypothesis $H_a : \hat{\mathbf{x}}_{i,j} \neq 0$. The F-statistic is given by,

$$F_{i,j} = \frac{\frac{\mathrm{RSS}^0_{i,j} - \mathrm{RSS}^1_{i,j}}{\mathrm{df}_{i,j} - 1}}{\frac{\mathrm{RSS}^1_{i,j}}{K_{i,j} - \mathrm{df}_{i,j}}},$$

where $\mathrm{df}_{i,j} = \mathbf{B}_{i,j}(\mathbf{B}'_{i,j}\mathbf{B}_{i,j} + \lambda \mathbf{R}_{i,j})^{-1}\mathbf{B}'_{i,j}$ is the degrees of freedom of the estimated curve $\hat{\mathbf{x}}_{i,j}$, $\mathrm{RSS}^0_{i,j} = \tilde{\mathbf{y}}'_{i,j}\tilde{\mathbf{y}}_{i,j}$ and $\mathrm{RSS}^1_{i,j} = [\tilde{\mathbf{y}}_{i,j} - \hat{\mathbf{x}}_{i,j}]'[\tilde{\mathbf{y}}_{i,j} - \hat{\mathbf{x}}_{i,j}]$ are the residual sum of squares under the null and the alternative models for the i-th gene, belonging to subject j. The genes with large F-ratios can be considered as exhibiting notable changes with respect to time.

Cluster the DRGs into temporal gene response modules (GRMs)

As many of the DRGs exhibit similar expression patterns over time, we wish to cluster them into co-expressed modules (groups of genes which have similar gene expression patterns over time). This step not only reduces the modeling dimension but also eases the identifiability problem. It is widely recognized that many co-expressed genes may follow similar temporal patterns, but at the same time, some genes may have very few or even no co-expressed genes, and thus may exhibit unique temporal response patterns. Consequently, the GRMs can vary greatly in size, with some being large and containing many genes and others being small or even containing a single gene. To obtain these clusters, we adopt the Iterative Hierarchical Clustering (IHC) method introduced in [6]. This approach requires a single parameter α that controls the trade-off of the between- and within-cluster correlations. In particular, the average within-cluster correlation will be approximately α, and the between- cluster correlation will be below α. The IHC algorithm is outlined below:

Initialization: Cluster the data for the standardized DRGs using the hierarchical agglomerative clustering approach. Let the distance metric be the Spearman rank correlation with a threshold of α, and the linkage method be the average of the genes in each cluster.

Merge: Treat each of the cluster centers as 'new genes' and use the same rule as in the initialization step to merge the centers into new clusters. The cluster centers provide the average time-course pattern of the cluster members.

Prune: Let c_i be the center of cluster i. If the correlation between the cluster center and gene j, which will be denoted by $\rho_{i,j}$, is less than α, then remove $gene_j$ from the cluster i. Let P be the number of genes removed from the existing S clusters. Assign all P genes into single-element clusters. Hence, there is now $(S + P)$ clusters in total.

Repeat Merge-Prune Steps until the index of clusters converges.

Repeat Merge Step until the between-cluster correlations are less than α.

Construct the high-dimensional gene regulatory network (GRN) that determines the interactions between the GRMs

High-dimensional gene regulatory networks map how the change in the expression of any single gene is regulated by its own expression level and other gene expression levels. There is an abundance of literature regarding the use of ordinary differential equation (ODE) modeling to construct a high-dimensional gene regulatory network (GRN) [7, 8, 9]. ODEs model gene regulations using rate equations. ODEs differ from the classical regression models as they capture not only the direct effects that are strong interactions between two gene response modules but they also include indirect effects high correlations that may exist between two gene response modules that are not directly connected but influence each other via a third gene response module they both directly interact with. In general, such indirect interactions may be induced not only by the third gene response module, but equally by the entire collective dynamics of a network.

Here we model the interactions between GRMs using the following ODE

$$\frac{\mathrm{d}m_{q,j}}{\mathrm{d}t} = \sum_{p=1}^{Q} \beta_{p,q,j} m_{p,j}, \qquad \text{for} \quad q = 1, \ldots, Q, \tag{2}$$

where $\frac{\mathrm{d}m_{q,j}}{\mathrm{d}t}$ represents the instantaneous rate of change in q^{th} gene response module for subject j, $\{\beta_{p,q,j}\}_{p,q=1}^{Q}$ quantifies the regulation effects of the p^{th} gene response module on the rate of change of the q^{th} gene response module for subject j. The standard approach for estimating the parameters of differential equations from noisy measurements is non-linear least squares (NLS) [10, 11]. However, this method requires initial estimates of the regulation effects and initial conditions for the expression levels of the gene response modules at time t_0.

(a) Initial estimates of the regulation effects

The two-stage smoothing-based estimation method [12, 13] decouples the system of differential equations in (2) and approximates it by a set of pseudo-regression models as in (3). The first step obtains estimates of the average trajectory of the p^{th} GRM for subject j, $\hat{\mathbf{m}}_{p,j}$ and its derivative $\frac{\mathrm{d}\hat{\mathbf{m}}_{p,j}}{\mathrm{d}t}$ for $p = 1, \ldots, Q$. The estimated trajectories $\hat{\mathbf{m}}_p$ are the average of the smoothed trajectories attained by the spline smoothing approach in 1, over a fine mesh of values of $\{t_r\}_{r=1}^{R}$ for all the genes contained in the p^{th} GRM. Similarly, $\frac{\mathrm{d}\hat{\mathbf{m}}_{p,j}}{\mathrm{d}t}$ is estimated by averaging the derivative of the smoothed trajectories obtained by the spline smoothing approach in 1, $\frac{\mathrm{d}\hat{\mathbf{x}}_{i,j}}{\mathrm{d}t} = \frac{\mathrm{d}\mathbf{B}_{i,j}}{\mathrm{d}t}\hat{\mathbf{c}}_{i,j}$, over the same fine mesh of values for all the genes contained in the p^{th} GRM. The set of Q pseudo-regression models is

$$\frac{\mathrm{d}\hat{\mathbf{m}}_{q,j}}{\mathrm{d}t} = \sum_{p=1}^{Q} \theta_{p,q,j} \hat{\mathbf{m}}_{p,j} + \epsilon_{p,j} \quad q = 1, \ldots, Q \text{ and } j = 1, \ldots, N, \tag{3}$$

where $\theta_{p,q,j}$ denotes the direct effects that is the strong relationships between the p^{th} GRM and the rate of change in the q^{th} GRM for the j^{th} subject.

It is widely accepted that gene regulatory networks are sparse, *i.e.*, only a few of the $\theta_{p,q,j}$ are non-zero. In order to determine which of the regulation effects are significant (*i.e.*, non-zero) we apply the least absolute shrinkage and selection operator (LASSO) [14] approach to the pseudo-regression model in (3).

The LASSO approach requires minimizing

$$\left[\frac{\mathrm{d}\hat{\mathbf{m}}_{q,j}}{\mathrm{d}t} - \sum_{p=1}^{Q} \theta_{p,q,j}\hat{\mathbf{m}}_{p,j}\right]^2 + \gamma \sum_{p=1}^{Q} \|\theta_{p,q,j}\| \quad q = 1, \ldots, Q.$$

with respect to $\theta_{p,q,j}$ When γ is zero, the result will be same as conventional regression; when the value of γ is large, the coefficients $\theta_{p,q,j}$ will approach zero. This means that the LASSO estimator is a smaller model, with fewer predictors. The L1 regularization parameter γ enforces the amount of sparsity in $\boldsymbol{\theta}$ and can be chosen by minimizing the GCV. As such, LASSO is a model selection and dimensionality reduction technique that determines the significant coefficients $\theta_{p,q,j}$ or initial estimates of the weighted network edges.

(b) Refined estimation of the regulation effects

The parameter estimates from the two-stage method in the above pseudo-regression model are not efficient in terms of estimation accuracy when the model selection is performed simultaneously, and there can be significant approximation error in $\hat{\mathbf{m}}_{p,j}$ and its derivatives. The estimation of significant coefficients or network edges can be improved or refined using nonlinear least squares (NLS), maximum likelihood or other more efficient estimation methods once the model selection from part (a) of Step 6 is completed. We now adopted the NLS approach which minimizes the squared discrepancy between the numerical approximation to the solution of the differential equation (2) and the observations. The non-zero estimates of $\{\hat{\theta}_{p,q,j}\}_{p,q=0}^{Q}$ from part (a) are used as initial estimates for the regulation effects $\boldsymbol{\beta} = \{\beta_{p,q,j}\}_{p,q=1}^{Q}$ in equation (2). Given the initial estimates and a set of p initial values, $\hat{\mathbf{m}}_{p,j}$, which are attained by the spline smoothing approach in part (a), an initial numerical approximation of the solution to differential equation (2) can be computed.

The variability in the GRN is assessed by calculating the confidence intervals of the parameters $\hat{\boldsymbol{\beta}}$ using the delta method see [15] for details.

Results

Study

Use the study summary and overall design on GEO:
Summary
Diagnosis of influenza A infection is currently based on clinical symptoms and pathogen detection. Use of host peripheral blood gene expression data to classify individuals with influenza A virus infection represents a novel approach to infection diagnosis We used microarrays to assay peripheral blood gene expression at baseline and every 8 hours for 7 days following intranasal influenza A

H1N1 or H3N2 inoculation in healthy volunteers. We determined groups of co-expressed genes that classified symptomatic influenza infection. We then tested this gene expression classifier in patients with naturally acquired respiratory illness.

Overall design
We experimentally inoculated healthy volunteers with intranasal influenza A H1N1 and H3N2. Symptoms were documented and peripheral blood samples drawn into PAXgene RNA tubes for RNA isolation. We further enrolled patients presenting to the Emergency Department with naturally acquired respiratory illness, and documented symptoms and collected PAXgene RNA samples for RNA isolation.

DRGs

The F-test statistic was used to identify the top 3000 DRGs. These genes might translate into clinically valuable bio markers. Investigate the Annotation Cluster(s) attained from DAVID given below:
Annotation Cluster 1 Enrichment Score: 9.75

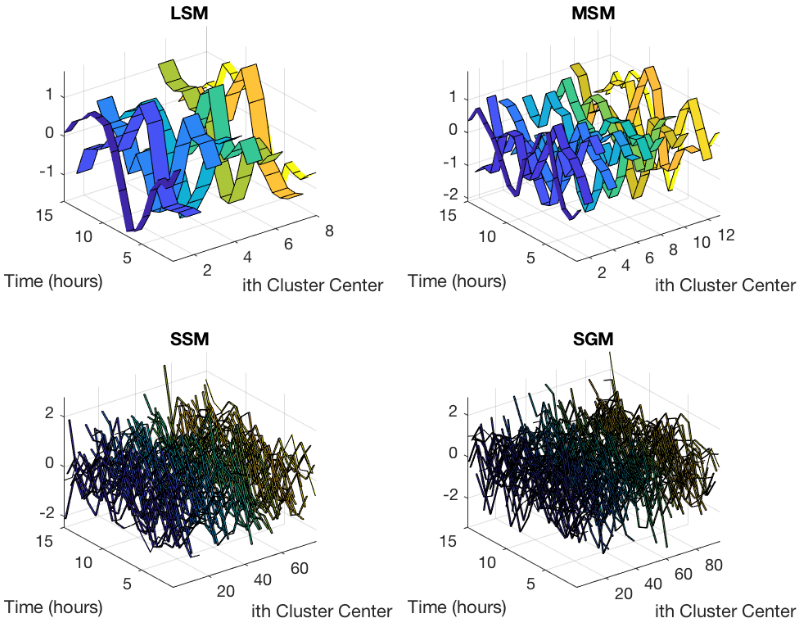
		Count	P Value	Benjamini
UP KEYWORDS	Antiviral defense	10	6.6E-16	5.0E-14
GOTERM BP DIRECT	defense response to virus	10	1.0E-13	2.4E-11
GOTERM BP DIRECT	type I interferon signaling pathway	8	6.2E-13	7.0E-11
UP KEYWORDS	Innate immunity	11	9.7E-12	2.4E-10
GOTERM BP DIRECT	defense response to virus	10	1.0E-13	2.4E-11
GOTERM BP DIRECT	negative regulation of viral genome replication	6	8.9E-10	6.8E-8
UP KEYWORDS	Cytoplasm	13	4.9E-4	7.3E-3
GOTERM CC DIRECT	cytosol	10	5.2E-3	1.1E-1

Annotation Cluster 2 Enrichment Score: 4.32

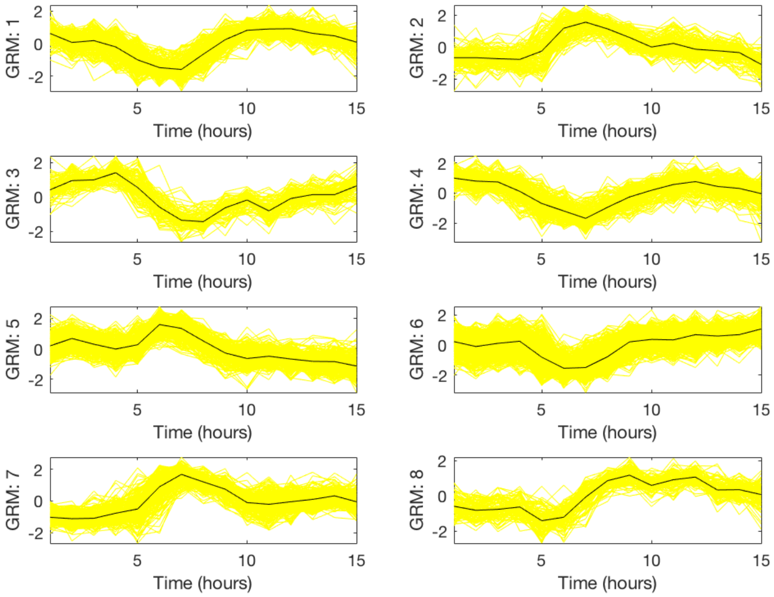
		Count	P Value	Benjamini
GOTERM BP DIRECT	response to virus	7	2.5E-9	1.4E-7
GOTERM MF DIRECT	double-stranded RNA binding	3	1.7E-3	9.9E-2
GOTERM BP DIRECT	type I interferon signaling pathway	8	6.2E-13	7.0E-11
UP KEYWORDS	RNA-binding	4	2.5E-2	1.8E-1

GRMs

The IHC method with a correlation threshold of $\alpha = 0.7$ was used to group the 3000 DRGs into GRMS. These modules can be classified into single-gene modules (SGM) with only one gene in each cluster, small-size modules (SSM) that contain between 2-10 genes in each cluster, medium-size modules (MSM) that consist of 11-99 genes in each of the clusters and large-size modules (LSM) which contain over 100 genes in each cluster.



The large size modules are often of interest as so many genes follow the same pattern.



The Annotation Cluster(s) attained from DAVID of the genes in each of the GRMS is given below:
GRM 1: Annotation Cluster 1 Enrichment Score: 14.85

		Count	P Value	Benjamini
UP KEYWORDS	Immunity	117	2.8E-21	3.7E-19
UP KEYWORDS	Innate immunity	71	8.4E-17	8.4E-15
GOTERM BP DIRECT	innate immune response	84	1.2E-8	9.9E-6

Discussion

Conclusion

Acknowledgements

This research was partially supported by the NIH grants HHSN272201000055C and R01 AI087135. The authors are grateful to the students, postdoctoral fellows and faculty members who are participating in the GEO-BigData project at the University of Texas Health Science Center at Houston.

References

[1] Green PJ, Silverman BW. Nonparametric regression and generalized linear models: a roughness penalty approach. CRC Press; 1993.

[2] Silverman B, Ramsay J. Functional Data Analysis. Springer; 2005.

[3] Golub GH, Heath M, Wahba G. Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics. 1979;21(2):215–223.

[4] Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. Journal of the American Statistical Association. 2005;100(470):577–590.

[5] Wu S, Wu H. More powerful significant testing for time course gene expression data using functional principal component analysis approaches. BMC bioinformatics. 2013;14(1):6.

[6] Carey M, Wu S, Gan G, Wu H. Correlation-based iterative clustering methods for time course data: the identification of temporal gene response modules for influenza infection in humans. Infectious Disease Modelling, in press, http://dxdoiorg/101016/jidm201607001. 2016;.

[7] Hecker M, Lambeck S, Toepfer S, Van Someren E, Guthke R. Gene regulatory network inference: data integration in dynamic models—a review. Biosystems. 2009;96(1):86–103.

[8] Lu T, Liang H, Li H, Wu H. High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. Journal of the American Statistical Association. 2011;106(496).

[9] Wu S, Liu ZP, Qiu X, Wu H. High-Dimensional Ordinary Differential Equation Models for Reconstructing Genome-Wide Dynamic Regulatory Networks. In: Topics in Applied Statistics. Springer; 2013. p. 173–190.

[10] Hemker P. Numerical methods for differential equations in system simulation and in parameter estimation. Analysis and Simulation of biochemical systems. 1972;p. 59–80.

[11] Bard Y, Bard Y. Nonlinear parameter estimation; 1974.

[12] Voit EO, Almeida J. Decoupling dynamical systems for pathway identification from metabolic profiles. Bioinformatics. 2004;20(11):1670–1681.

[13] Liang H, Wu H. Parameter estimation for differential equation models using a framework of measurement error in regression models. Journal of the American Statistical Association. 2008;103(484).

[14] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996;p. 267–288.

[15] Bates DM, Watts DG. Nonlinear regression: iterative estimation and linear approximations. Nonlinear regression analysis and its applications. 1988;p. 32–66.