Bioinformatics, 33, 2017, i208–i216
doi: 10.1093/bioinformatics/btx260
ISMB/ECCB 2017

Multiple network-constrained regressions expand insights into influenza vaccination responses

Stefan Avey[1,*], Subhasis Mohanty[2], Jean Wilson[2], Heidi Zapata[2], Samit R. Joshi[2], Barbara Siconolfi[2], Sui Tsang[3], Albert C. Shaw[2] and Steven H. Kleinstein[1,4,*]

[1]Interdepartmental Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT 06511, USA, [2]Section of Infectious Diseases, Department of Internal Medicine, [3]Department of Internal Medicine and [4]Departments of Pathology and Immunobiology, Yale School of Medicine, New Haven, CT 06520, USA

*To whom correspondence should be addressed.

Abstract

Motivation: Systems immunology leverages recent technological advancements that enable broad profiling of the immune system to better understand the response to infection and vaccination, as well as the dysregulation that occurs in disease. An increasingly common approach to gain insights from these large-scale profiling experiments involves the application of statistical learning methods to predict disease states or the immune response to perturbations. However, the goal of many systems studies is not to maximize accuracy, but rather to gain biological insights. The predictors identified using current approaches can be biologically uninterpretable or present only one of many equally predictive models, leading to a narrow understanding of the underlying biology.
Results: Here we show that incorporating prior biological knowledge within a logistic modeling framework by using network-level constraints on transcriptional profiling data significantly improves interpretability. Moreover, incorporating different types of biological knowledge produces models that highlight distinct aspects of the underlying biology, while maintaining predictive accuracy. We propose a new framework, Logistic Multiple Network-constrained Regression (LogMiNeR), and apply it to understand the mechanisms underlying differential responses to influenza vaccination. Although standard logistic regression approaches were predictive, they were minimally interpretable. Incorporating prior knowledge using LogMiNeR led to models that were equally predictive yet highly interpretable. In this context, B cell-specific genes and mTOR signaling were associated with an effective vaccination response in young adults. Overall, our results demonstrate a new paradigm for analyzing high-dimensional immune profiling data in which multiple networks encoding prior knowledge are incorporated to improve model interpretability.
Availability and implementation: The R source code described in this article is publicly available at https://bitbucket.org/kleinstein/logminer.
Contact: steven.kleinstein@yale.edu or stefan.avey@yale.edu
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Systems immunology leverages recent technological advancements in high-dimensional immune profiling to monitor the response to perturbations such as vaccination, as well as the dysregulation that occurs in disease. An increasingly common approach to gain insights from these large-scale profiling experiments involves the application of statistical learning methods, such as classification, to accurately predict immune state or clinical outcome (Larrañaga *et al.*, 2006). However, interpreting these models to gain insights into the underlying process remains a challenge.

Although a statistical model must be accurate to be meaningful, many systems biology studies focus on understanding the underlying biology and not on maximizing predictive accuracy. Even in cases where model accuracy is the main goal, such as diagnosis or prognosis, an accurate model that is interpretable in terms of underlying biology would likely be preferred over a similarly accurate model which lacked interpretability. Thus, it may be preferable for statistical learning models to be interpretable in terms of what is already known about a system rather than attempting to maximize accuracy using a 'black box' approach.

In order to interpret a model, the most predictive genes can be identified and tested for enrichment using collections of related gene lists known as gene set libraries. However, previous studies have shown that regression models built from transcriptional profiles can accurately predict immune state or cancer subtype using multiple completely distinct sets of genes (Gruvberger *et al.*, 2001; O'Hara *et al.*, 2013). Strikingly, O'Hara *et al.* (2013) found that removing all predictive genes from a model and refitting the model can be done for several iterations before a decrease in accuracy is observed. This result implies that single, parsimonious models may miss genes important to the underlying biology of the system of interest. Furthermore, the top predictive genes identified by a model can sometimes fail to be enriched in any gene set libraries, resulting in a lack of biological interpretability.

In an effort to improve model interpretability, many studies have proposed network-constrained regularization approaches that utilize prior knowledge (Chuang *et al.*, 2007; Li and Li, 2008; Rapaport *et al.*, 2007; Sun and Wang, 2012). These methods take advantage of large repositories of biological knowledge (e.g. pathways or protein–protein interactions) by encoding them in gene–gene networks and using these networks to enforce a constraint on the model. Including prior knowledge in the modeling process improves interpretability of the classifiers and in some cases can improve their accuracy over non-network methods. However, these studies were performed with a single source of prior knowledge. Thus, it remains unknown how prior knowledge network choice affects model performance and whether a single model is sufficient to capture the underlying biology.

Here we show that fitting multiple models, each incorporating a different source of prior biological knowledge, greatly improves model interpretability. We propose a new framework, Logistic Multiple Network-constrained Regression (LogMiNeR), which utilizes multiple models that each highlight distinct aspects of the underlying biology, while maintaining predictive accuracy. We first apply LogMiNeR to transcriptional profiling data to better understand differential influenza vaccination responses and subsequently show that LogMiNeR can be applied to classification of many immune as well as non-immune-mediated diseases. This new paradigm provides additional insights in systems biology studies which focus on finding predictive signatures that are interpretable in terms of prior biological knowledge.

2 Materials and Methods

2.1 Availability of transcriptional profiling data

The validation data (SDY80) is described in (Tsang *et al.*, 2014). The gene expression data from SDY63 and SDY404 are published and described in Thakar *et al.* (2015). The design of SDY400 is identical to that described in Thakar *et al.* (2015) except that the samples were collected during the 2012–13 vaccine season. Data are available from ImmPort (https://immport.niaid.nih.gov) and GEO (Discovery: GSE59635, GSE59654, GSE59743; Validation: GSE47353).

2.2 Defining vaccine response endpoint

Vaccination response was calculated from the fold change in antibody titer post-vaccination compared with pre-vaccination. Titers were measured at days 0 and 28 by hemagglutination inhibition assay in the discovery data and at days 0 and 70 by virus neutralization assay in the validation data. A titer of half the first dilution was assigned to samples in which the first dilution was negative and the largest dilution was reported if it was positive. High and low responders were defined as the top and bottom 30%, respectively, of the maximum adjusted fold change as defined by Tsang *et al.* (2014).

2.3 Data preprocessing

The discovery datasets were initially quantile normalized across arrays, and the processed validation data was used as provided. Following array normalization, each study went through several preprocessing steps independently in order to mitigate batch effects. First, probes were mapped to Entrez Gene IDs using the Bioconductor tool AnnotationDbi (Pages *et al.*, 2015), and probes were collapsed to unique genes by choosing the probe with maximum average expression. Next, genes located on the X and Y chromosomes were removed to avoid selection of sex-linked genes that may be confounded with vaccine response. The log fold change between day 7 and 0 was calculated for each gene, and the 1000 genes with the largest fold change magnitudes in the discovery datasets were selected as the initial feature set. The gene fold changes were standardized by subtracting the mean and dividing by the standard deviation. Finally, the pre-processed data from SDY63, SDY404 and SDY400 were combined to form the discovery data and SDY80 was used as the validation data.

The data for the additional case studies were downloaded from GEO using the Bioconductor R package GEOquery (Sean and Meltzer, 2007). The coefficient of variation was calculated for each gene and the 500 genes with largest variation were selected as the initial feature set.

GSE45291: The classifier was built to distinguish the 292 samples from subjects with SLE from the 20 control samples at baseline (time 0).

GSE37250: The classifier was built to distinguish the 195 samples from subjects with active tuberculosis from the 167 samples with latent tuberculosis.

GSE57338: The classifier was built to distinguish the 82 samples from subjects with idiopathic dilated cardiomyopathy from the 95 samples with ischemic heart disease.

2.4 Network-constrained logistic regression

Network-constrained logistic regression was performed as described by Sun and Wang (2012). To account for pairs of connected genes that were expected to have similar magnitude but potentially opposite effects on response, the adaptive regularization procedure described in Section 3.4 of Sun and Wang (2012) was performed.

Briefly, the signed Laplacian matrix was first calculated for each network by setting the *uv*th entry in the following way (Equation 1),

$$l^{uv} = \begin{cases} 1 & \text{if } u = v \text{ and } d_u \neq 0, \\ -s_u s_v/\sqrt{d_u d_v} & \text{if } u \text{ and } v \text{ are connected}, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

where d_u and d_v are the degrees of genes u and v and s_u and s_v are the signs of the coefficients for genes u and v estimated by correlation with the response variable during each round of cross validation. The model was then fit to minimize the objective function (Equation 2)

$$-\frac{1}{n}\sum_{i=1}^{n}[y_i \log(p(x_i)) + (1-y_i)\log(1-p(x_i))] + P(\theta) \qquad (2)$$

where response y_i is 0 for low responders or 1 for high responders and x_i is a vector of gene expression values for subject i. The penalty function took the form

$$P(\theta) = \lambda\alpha\sum_{j=1}^{p}|\theta_j| + \frac{1}{2}\lambda(1-\alpha)\sum_{u=1}^{p}\sum_{u\sim v}\left(\frac{s_u\theta_u}{\sqrt{d_u}} - \frac{s_v\theta_v}{\sqrt{d_v}}\right)^2 \qquad (3)$$

where $u \sim v$ indicates the set of node pairs which are connected to u in the network. The first term is the L1 penalty that results in model sparsity and the second term is the network constraint in Laplacian quadratic form that results in smoothness of coefficients over the network.

2.5 Fitting models

Lasso and Elastic Net Logistic Regression were performed using the glmnet R package v2.0.2 (Friedman *et al.*, 2010). Network-constrained logistic regression was performed using the pclogit R package v0.2 (Sun and Wang, 2012). 200 values of the tuning parameter lambda were chosen and 20 values of alpha were used in the interval [0,1]. Five-fold cross validation was performed to choose the optimal values of lambda and alpha. The cross-validation procedure was repeated 50 times for robustness to different splits of the discovery data. Cross-validation folds were balanced by study, response endpoint and gender for the flu vaccination case study and by response endpoint in the additional case studies. The parameters (lambda, alpha) were chosen by selecting the best model (Lasso: most sparse; Elastic Net and LogMiNeR: least sparse) with cross validated error within one standard error of the minimum.

2.6 Calculating Biological Interpretability

The mean-rank gene set enrichment test (Michaud *et al.*, 2008) was performed using the absolute value of the model coefficients to rank the genes. The test was performed using the geneSetTest function in the limma R package v3.24.15 (Ritchie *et al.*, 2015). Gene sets for KEGG, Reactome, and GO were downloaded from the Molecular Signatures Database v5.1 (Subramanian *et al.*, 2005). Gene sets for BTM (Li *et al.*, 2013) and CELLS (Abbas *et al.*, 2005) were obtained from the original publication. Only sets which had at least 15 genes overlapping with the gene expression feature set were used. False discovery rates were obtained using the method of Benjamini and Hochberg (1995).

2.7 Prior knowledge networks

Networks were defined for Reactome, GO, BTM and CELLS by connecting all pairs of genes within each gene set. The KEGG network incorporated pathway topology and was built using the KEGGgraph R package v1.26.0 (Zhang and Wiemann, 2009). ImmuneGlobal and ExpOnly_ImmuneGlobal networks were obtained from ImmuNet and edges were restricted to those with confidence of at least 0.1 (Gorenshteyn *et al.*, 2015). The STRING network incorporated all experimental evidence from the STRING database v10.0. The processed LINCS L1000 gene signatures generated as part of the Broad Institute Connectivity Map (Lamb *et al.*, 2006) were obtained from the Enrichr downloads page (Chen *et al.*, 2013), and a network was created by connecting all pairs of genes within each gene set.

2.8 Identifying potential drug interactions

The nearly 8000 sets included in the LINCS L1000 signatures were filtered to the drug response gene sets that were significantly enriched and reversed the signature of high vaccine response. These drug response gene sets contained genes which were significantly altered in the same direction upon treatment with the drug in the original dataset and when comparing low–high responders. To limit the number of false positives, we only considered gene sets that were significantly enriched (FDR < 0.001) in at least 50% of the cross validation runs.

2.9 Visualizing significantly enriched gene sets

The distance between each pair of gene sets which were significantly enriched in at least 10% of the 50 runs was calculated using the Jaccard distance. The minimum spanning tree was calculated for this network. This tree was then visualized in Cytoscape (Shannon *et al.*, 2003, 2013) and the MultiColoredNodes plugin (Warsow *et al.*, 2010) was used to color each node according to the prior knowledge network that led to its enrichment.

3 Results

3.1 Lasso and elastic net models are minimally interpretable

Systems vaccinology seeks to use high-throughput profiling of immune responses to vaccination in order to gain insights into the underlying biological mechanisms that lead to protection (Pulendran *et al.*, 2010). Such approaches may be especially useful for improving the response to influenza vaccination, which is an important public health tool, but fails to induce an antibody response in a significant fraction of individuals (Sasaki *et al.*, 2011). To better understand why some individuals successfully generate antibody responses, while others fail to do so, we recruited healthy young adults (21–30 years old) over three vaccination seasons and measured vaccine-specific antibody titers immediately prior to and 28 days post-vaccination with the seasonal trivalent inactivated influenza vaccine. Genome-wide transcriptional profiling of blood samples prior to and 7 days post-vaccination was carried out for a subset of these individuals with strong and weak antibody responses. To quantify the strength of the vaccine response, we used the adjusted Maximum Fold Change endpoint proposed by Tsang *et al.* (2014). This endpoint adjusts the maximum fold change in antibody titer for correlations with baseline titers and defines high and low vaccine responders as the top and bottom 30th percentile, respectively (see Section 2.2).

To predict the antibody response to influenza vaccination, we first applied standard Lasso logistic regression to the transcriptional profiling data (Tibshirani, 1996). The gene expression profiles were filtered to retain the 1000 genes which changed most 7 days post-vaccination (see Section 2.3). These profiles served as the discovery data to find a predictive signature to classify high and low vaccine responders. The models built from 50 runs of 5-fold cross validation on these discovery data were subsequently validated on an independent cohort from the NIH Center for Human Immunology (Tsang *et al.*, 2014). In total, we trained our model on 18 high and 17 low responders and validated our model on 11 high and 10 low responders (Supplementary Table S1). The Lasso models fit the discovery data with a median accuracy of

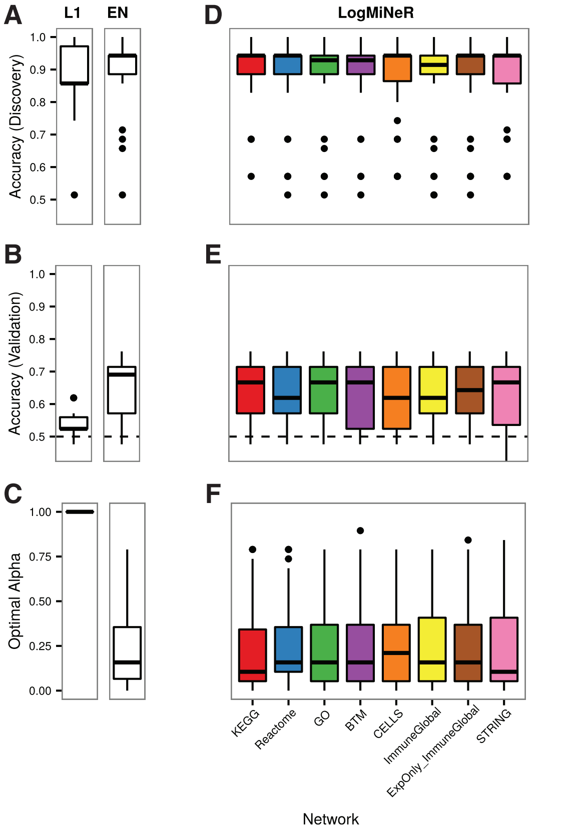

Fig. 1. The classification accuracy of models from 50 runs of 5-fold cross validation. Discovery **(A)** and Validation **(B)** accuracies for the Lasso (L1) and Elastic Net (EN) models. **(C)** The optimal value of the tuning parameter alpha [0,1] controlling the trade-off between the L1 and L2 constraints. Discovery **(D)** and Validation **(E)** accuracies for the network-constrained (LogMiNeR) models. **(F)** The optimal value of the tuning parameter alpha [0,1] controlling the trade-off between the L1 and network constraints

86% (Fig. 1A), but predicted response with poor accuracy in the validation data (Fig. 1B).

In order to understand what biological factors were able to discriminate high and low responders in the discovery data, we tested whether any gene sets were enriched among the predictive genes using a mean-rank gene set test (Michaud *et al.*, 2008) on the model coefficients (see Section 2.6). The gene set libraries used to test enrichment included KEGG (Kanehisa and Goto, 2000) and Reactome (Croft *et al.*, 2014; Milacic *et al.*, 2012) pathways, blood transcriptional coexpression modules (BTM) (Li *et al.*, 2013), gene ontology (GO) terms (Ashburner *et al.*, 2000), and blood cell subset signatures (CELLS) (Abbas *et al.*, 2005). On average, the models were significantly enriched (FDR < 0.05) for less than one gene set from four out of five gene set libraries tested (Fig. 2A). The Lasso models were enriched for an average of four GO terms including *catabolic process* and *cellular catabolic process*. These results show that Lasso logistic regression is predictive on the discovery data and interpretable in terms of GO terms, but does not validate well in an independent cohort.

The explicit goal of Lasso regression models is to reduce the number of features (genes) in the model, which may not be preferable when trying to interpret the underlying biology. In fact, the Lasso models chose only five genes on average, which may explain why these models do not validate and are minimally interpretable. In order to improve model accuracy and interpretability, we next fit Elastic Net logistic regression models (Zou and Hastie, 2005). The Elastic Net promotes a grouping effect that tends to include a larger number of relevant genes and was previously used to predict influenza vaccine response accurately from transcriptional profiles (Furman *et al.*, 2013). The Elastic Net models fit the discovery data with a median accuracy of 94% (Fig. 1A) and successfully predicted response in the validation data with a median accuracy of 69% (Fig. 1B). The hyperparameter controlling the trade-off between Lasso (alpha = 1) and ridge (alpha = 0) constraints favored the ridge constraint (Fig. 1C). Accordingly, the average number of genes used to predict vaccine response with Elastic Net (88) was much higher than for Lasso (5). However, when we tested whether any gene sets were enriched among these predictive genes, on average, the models were significantly enriched (FDR < 0.05) for less than one gene set from four out of five gene set libraries tested (Fig. 2A). Two GO terms, on average, were enriched using Elastic Net including *positive regulation of metabolic process* and *peptidase activity*.

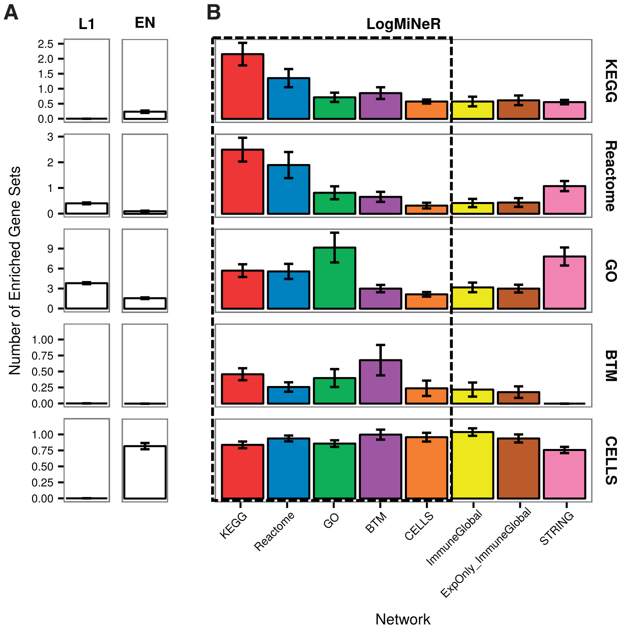

Fig. 2. The number of significantly enriched gene sets (FDR < 0.05) in multiple gene set libraries (rows) for L1 and EN models **(A)** or network-constrained (LogMiNeR) models **(B)**. The diagonal in the dashed box indicates models for which the same gene set library was used for prior knowledge and enrichment testing. Bars represent mean and whiskers represent one standard error

In order to test whether these models of vaccination response were more accurate than expected by chance, we permuted the class labels in the discovery set and retrained the Lasso and Elastic Net models. The validation accuracy after permuting class labels was 50%, which was significantly lower ($P = 0.03$, two-sided *t*-test) than Elastic Net but not significantly different from Lasso ($P = 0.66$, two-sided *t*-test), suggesting that the Elastic Net models were indeed fitting real differences in vaccination response and not noise (Fig. 1B). The difference in model accuracy was due to higher sensitivity using Elastic Net compared with Lasso (Supplementary Figs S3A, B, S4A and B). Overall, we found that Elastic Net models can predict vaccination response from transcriptional changes 7 days post-vaccination, but the genes underlying the predictions cannot be easily interpreted in terms of existing pathways or coexpression modules.

3.2 Network-constrained models improve interpretability

We next investigated whether biological interpretability could be improved by models that select different sets of genes. Prior work suggested that high-dimensional data often contain additional predictive genes beyond the ones chosen by standard machine learning methods. For example, iterative feature removal can lead to multiple mutually exclusive subsets of genes that predict equally well and implicate different pathways (O'Hara *et al.*, 2013).

We hypothesized that including prior knowledge directly into our modeling framework would improve interpretability by limiting the search space of the models according to known biological relationships. To test this, we coded prior knowledge in the form of gene–gene networks where edges represented associations between two genes. The type of association depended on the prior knowledge source (e.g. membership in the same pathway for KEGG). Multiple gene networks, representing a range of association types, served as prior knowledge to fit multiple models (Supplementary Table S2; see Section 2.7). All the gene set libraries used to test enrichment (KEGG, Reactome, BTM, GO, CELLS) were converted to networks and used as prior knowledge. Since existing databases are likely incomplete, we included additional prior knowledge networks derived from Bayesian integration of functional interactions from multiple evidence sources (ImmuneGlobal) or gene expression only (ExpOnly_ImmuneGlobal) (Gorenshteyn *et al.*, 2015) as well as protein–protein interactions with experimental evidence in the STRING database (Jensen *et al.*, 2009). The nine networks varied in edge density from 0.1% (KEGG) to 11.5% (GO) but all networks were filtered to contain the same set of genes used as input to the Lasso and Elastic Net models (Supplementary Table S2). A network constraint was added in addition to the Lasso constraint so that model coefficients would be smoothed over the network (Li and Li, 2008; Sun and Wang, 2012). The constraint is motivated by the assumption that genes closely connected in the prior knowledge network should contribute similarly to prediction. Connected genes are coerced to have model coefficients with similar magnitudes by penalizing the squared difference between coefficients. At the same time, we allow for the sign of the coefficients to vary. Thus we allow for anticorrelated genes, such as two gene products in a pathway where one negatively regulates the other, to contribute similarly to prediction but in opposite directions (see Section 2.4).

We applied this approach to the influenza vaccination data and found that the network-constrained models maintained similar predictive discovery accuracies as the Elastic Net models. Furthermore, the validation accuracies were predictive and not significantly different from Elastic Net ($P \geq 0.05$, two-sided *t*-test) (Fig. 1D and E). Each network-constrained model achieved a median discovery accuracy of at least 90% while maintaining a median validation accuracy of at least 62%. Interestingly, even when the prior knowledge networks were rendered biologically meaningless by gene label permutation, the network-constrained models retained similar validation accuracies, AUROC, sensitivities and specificities to the Elastic Net models (Supplementary Figs S1D, E, S2C, D, S3C, D, S4C and D). The subjects correctly classified as high or low responders in the validation dataset were consistent and did not depend on the source of prior knowledge. On average, the hyperparameter controlling the trade-off between model sparsity (alpha = 1) and network smoothness (alpha = 0) favored the network constraint (Fig. 1F). This indicates that the network constraint was enforced and consequently the magnitude of model coefficients was smooth over the network (see Section 2.5). Taken together, these results indicate that any of the network-constrained models are equally valid predictors of vaccination response, and are comparable with models built using Elastic Net.

In order to test whether network-constrained models improve interpretability compared with Lasso and Elastic Net, we again quantified interpretability as the average number of significantly enriched gene sets for each model. On average network-constrained models selected a comparable number of genes (45–150 depending on the network) to Elastic Net (88). Yet, in contrast to both the Lasso and Elastic Net models, the network-constrained models were highly enriched for known pathways and gene sets (Fig. 2B). For every gene set library, interpretability was better using a network-constrained model compared with using either standard Lasso or Elastic Net models. As expected, each prior knowledge network produced a large number of enriched gene sets based on that same network (e.g. using GO as prior knowledge leads to nine enriched GO sets). However, enrichment on the Reactome gene set library was greatest using the KEGG network as the prior knowledge source. Even though the accuracy of network-constrained models was similar when gene labels were permuted, the models built from permuted networks were minimally enriched for all gene set libraries tested (Supplementary Fig. S5B). In fact, the levels of enrichment using biologically meaningless networks closely mirrored the Elastic Net method. Thus, the network-constrained models are not only accurate, but also interpretable in terms of existing pathways, coexpression modules, GO terms, and blood cell subsets.

Although network-constrained models tend to choose more interpretable genes, individual genes that are not associated with any gene set libraries were still included in the predictive models. 240 of the 1000 genes used as the initial feature set were not present in any of the selected gene set libraries. All of these genes were used in at least one network-constrained model, and seven were consistently selected in at least 50% of cross-validation runs. One such gene, JCHAIN, is a positive predictor of high response in 72% of runs. JCHAIN encodes the joining chain for multermeric IgA and IgM antibodies and is highly expressed in antibody secreting cells (ASCs) (Nakaya *et al.*, 2011) suggesting an increase in circulating ASCs or antibody production 7 days post-vaccination in high responders. Thus, this method is able to identify novel predictive genes that are not annotated in existing gene set libraries.

3.3 Multiple models provide context-specific insights

We next asked whether the gene sets enriched by Lasso, Elastic Net, and network-constrained methods were context-specific or shared across multiple methods or prior knowledge networks. We defined 'consistently enriched' gene sets for each model as those that were significantly enriched (FDR < 0.05) in at least 10% of cross validation runs. Only a single gene set, GO *peptidase activity*, was consistently enriched regardless of the method used. (Supplementary Fig. S8). Of the 65 gene sets consistently enriched by any method 16 (25%) were enriched by two or more different methods. Although there were some shared insights, 49 (75%) gene sets were context-specific (i.e. consistently enriched by only one method).

Our results show that using different networks as prior knowledge leads to many predictive models that highlight context-specific as well as shared aspects of the underlying biology of vaccination response. Thus, we propose this approach, which we term 'LogMiNeR' (Logistic Multiple Network-constrained Regression), as a new framework for systems immunology studies. In contrast with previous methods, LogMiNeR fits multiple models, each using

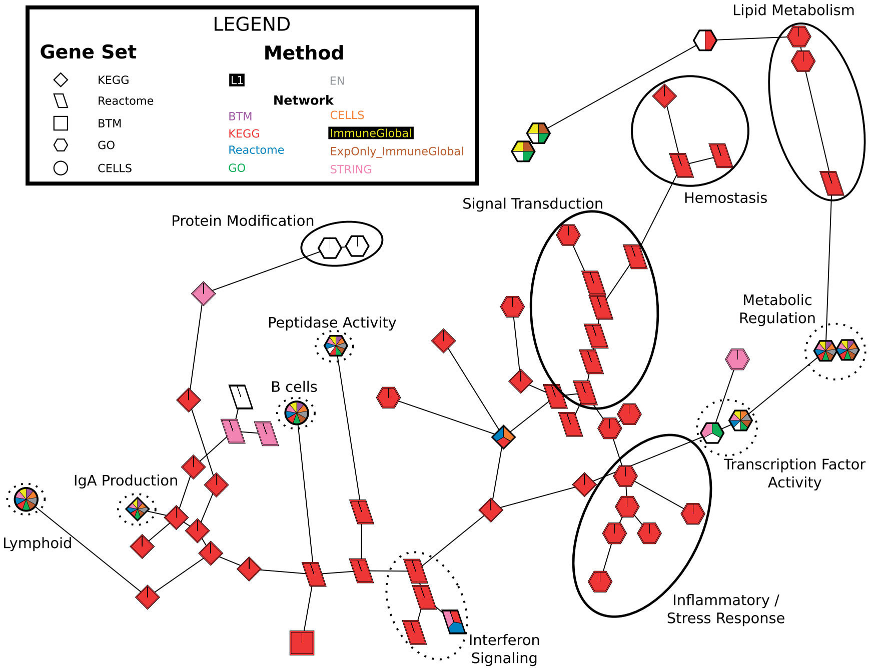

Fig. 3. Overview of the gene sets (nodes) arranged by Jaccard distance on a minimum spanning tree. Only gene sets significantly enriched (FDR < 0.05) by at least one model were included. We further limited the visualization to the sets consistently enriched in at least 10% of the runs. Groups of context-specific (solid ellipses) or shared (dashed ellipses) gene sets with similar annotations were manually identified

different sources of prior knowledge, to gain a broader understanding of the underlying biology.

LogMiNeR allows for the interpretation and visualization of multiple models simultaneously (Fig. 3). To visualize the results of LogMiNeR, gene sets were arranged by set similarity resulting in tightly connected groups of related gene sets. The gene sets identified by multiple models (25%) were annotated to processes such as interferon signaling, transcription factor activity, and metabolic regulation. In addition, lymphoid and specifically B cell signature genes were enriched by multiple models and were positive predictors of a successful vaccination response. The B cell enrichment was driven primarily by TNFRSF17, also known as B cell maturation antigen (BCMA). Other shared insights were only found by a subset of the methods. For example, the Reactome *interferon alpha/beta signaling* pathway positively predicted vaccine response and was only consistently enriched when the Reactome, KEGG or STRING networks were used as prior knowledge. (Supplementary Fig. S7). The many context-specific gene sets identified by only a single model (75%) were mainly found using the KEGG pathway network and included sets involved in inflammatory/stress response, signal transduction, lipid metabolism, and hemostasis. Overall, LogMiNeR allowed us to find both shared and context-specific insights into the influenza vaccination response which were dependent on the input prior knowledge network.

To further demonstrate the utility of LogMiNeR and the influence of the prior knowledge network, we applied this framework to identify drug interactions with the potential to negatively affect vaccine response (see Section 2.8). We hypothesized that if high vaccine responses were characterized by specific changes in gene expression, and if drug treatment caused the opposite changes in gene expression in model cell lines, then that drug has the potential to negatively impact a proper vaccine response. This is similar to the computational approach for drug repositioning suggested by Sirota *et al.* (2011) and provides a straightforward method to identify drugs that might interfere with a successful influenza vaccination response. To accomplish this analysis, we applied LogMiNeR using a drug response network built from the Library of Integrated Network-based Cellular Signatures (LINCS) L1000 dataset (Lamb *et al.*, 2006). The validation accuracy was predictive and was not significantly different than Elastic Net ($P \geq 0.05$, two-sided t-test). Out of the nearly 8000 drug response signatures tested, we identified seven consistently and significantly enriched (FDR < 0.001) drug response gene sets representing six distinct drugs (Supplementary Table S3). One of the six drugs, GDC-0980, is an mTOR inhibitor, suggesting that individuals taking mTOR inhibitors may have less effective responses to the seasonal influenza vaccination. The other drugs are kinase inhibitors specific to EGFR (gefitinib), MET/VEGFR2 (foretinib), MEK (AZD-8330, selumetinib) and Akt (A443654). Notably, these compounds target growth factor receptors (EGFR, VEGFR2, MET) or members of the Ras-ERK or PI3K-mTOR signaling pathways which lead to cell survival, proliferation, and motility (Mendoza *et al.*, 2011). Together, these results suggest potential drug interactions that may adversely affect vaccine response as well as highlight the importance of cell growth, proliferation and motility to generating a successful antibody response upon influenza vaccination.

3.4 LogMiNeR improves interpretability on additional datasets

In order to assess whether the results we obtained on the influenza vaccination data were generalizable to other datasets with larger sample sizes, we tested LogMiNeR on classification of publicly available gene expression data from an autoimmune disorder (SLE versus healthy), a bacterial infection (latent versus active TB), and a non-immune mediated disease (ischemic heart disease versus dilated cardiomyopathy) (see Section 2.1). In all three additional datasets tested, the accuracy of the network-constrained models was similar to the non-network methods (Supplementary Figs S11A, B, S12A, B, S13A and B).

The classification of SLE versus healthy subjects from peripheral blood was performed with nearly perfect accuracy (98–99%) regardless of which method was used. Although this highly accurate prediction suggests a strong signal in the data, the standard Lasso and Elastic Net methods were only enriched for an average of two GO terms—*negative regulation of biological process* and *cell–cell signaling* (Supplementary Fig. S11C). Using LogMiNeR with the STRING network, we found enrichment of the Reactome pathways *cytokine signaling in immune system* and, more specifically, *interferon alpha/beta signaling* (Supplementary Fig. S11D).

We next tested for enrichment of gene set libraries on models classifying latent versus active tuberculosis. The number of enriched signatures was similar for all methods, with the exception of GO sets which were more highly enriched with LogMiNeR (Supplementary Fig. S12C and D). Regardless of the method used, the Reactome pathways *immune system* and *cytokine signaling in immune system* were consistently and significantly enriched. In addition, LogMiNeR led to enrichment of GO terms beyond those shared by all methods including *response to external stimulus* and *defense response* which were positive predictors of active tuberculosis.

Finally, we applied LogMiNeR to classify patients with ischemic heart disease or dilated cardiomyopathy from transcriptional profiling data of heart tissue (GSE57338). This dataset is different from the others in that the samples are not from peripheral blood tissue and do not characterize an immune-mediated disease. Both Lasso and Reactome network-constrained models are enriched for multiple gene sets above background (Supplementary Fig. S13C and D). The enrichments of the Lasso models are more consistent across runs and include innate immune sets such as the KEGG *toll-like receptor signaling pathway*, Reactome *innate immune system* and BTM *regulation of antigen processing and presentation and immune response (M5.0)*. Thus, we suggest that Lasso and Elastic Net be used alongside of LogMiNeR as they can provide additional insights.

Our results demonstrate that LogMiNeR can be successfully applied to additional datasets and leads to models that are similarly accurate, but have the potential to expand the interpretability of standard logistic regression approaches.

4 Discussion

We propose LogMiNeR as a new framework for analyzing classification-based studies that leads to increased biological interpretability. A series of case studies using multiple transcriptional profiling datasets demonstrate that LogMiNeR leads to many accurate models that are biologically interpretable. In contrast, the application of standard classification methods to these same data produces models that are no more accurate, but are often severely limited in interpretability.

Our primary case study focused on understanding influenza vaccination responses. Specifically, we sought to use transcriptional profiling data from post-vaccination blood samples to predict whether individuals would have high or low vaccine responses as determined by antibody titers. Lasso models were not predictive above random on an independent validation dataset and few gene sets were consistently enriched by either standard classification method (Lasso or Elastic Net logistic regression). Elastic Net models were validated in an independent study and associated an increase in B cell signature genes, specifically TNFRSF17 (BCMA), 7 days post-vaccination with increased vaccination response. The expression of BCMA 7 days post-vaccination is a known positive predictor of antibody response to both yellow fever vaccine 17D (Querec *et al.*, 2009) and inactivated influenza vaccine (Li *et al.*, 2014; Nakaya *et al.*, 2011; Obermoser *et al.*, 2013). LogMiNeR models were similarly predictive to Elastic Net and expanded interpretability by identifying over 60 consistently and significantly enriched gene sets not found by standard classification methods. LogMiNeR allowed us to identify known components of the vaccination response that were not identified using Lasso or Elastic Net logistic regression. For example, we identified gene sets involved in interferon signaling to positively predict a successful influenza vaccination response. Interferon signaling has previously been reported 1–3 days post-immunization (Bucasas *et al.*, 2011; Li *et al.*, 2014; Tsang *et al.*, 2014), and it is possible that the network constraint helped enrich for this signal at later time points as well. We also identified 6 drugs from a library of nearly 8000 drug response profiles whose expression patterns suggested a potential negative effect on vaccination response in our cohort of young adults. We speculated that one such drug, the mTOR inhibitor GDC-0980, may negatively impact vaccination response. Although mTOR inhibitors were reported to improve immune responses to influenza vaccination in the elderly (Mannick *et al.*, 2014), these inhibitors reportedly decreased the response to pandemic influenza vaccination in a cohort of young adults receiving solid organ transplants (Cordero *et al.*, 2011), suggesting that the effect may be age-dependent. Furthermore, the drugs we identified were all inhibitors of growth factor receptors or members of the Ras-ERK or PI3K-mTOR signaling pathways (Mendoza *et al.*, 2011). This finding points to greater cell growth and proliferation in high vaccine responders. Whether these signatures are a primary cause leading to response or a by-product of B cell proliferation and migration, such as an increase in circulating plasmablasts (Tsang *et al.*, 2014), will require further study. Thus, LogMiNeR helped us identify known components of the vaccine response as well as generate new hypotheses about how drugs might affect response to seasonal influenza vaccination.

We also applied LogMiNeR to three larger datasets, each containing hundreds of samples, as case studies of disease classification. We found that all LogMiNeR models were similarly accurate, but had the potential to expand interpretability of standard logistic regression approaches. For example, when we classified SLE versus healthy samples from peripheral blood, LogMiNeR, but not standard approaches, resulted in models that were enriched for the interferon alpha/beta signaling pathway. This finding is consistent with many previous observations that individuals with SLE have a strong interferon signature in their peripheral blood (Rönnblom and Eloranta, 2013). In two other datasets the significantly enriched gene sets were similar whether or not network knowledge was incorporated. The LogMiNeR models classifying individuals with latent versus active TB were enriched beyond general immune system pathways to include more specific sets that positively predicted active TB including *response to external stimulus* and *response to stress*. On the other hand, Lasso models classifying ischemic heart disease versus dilated cardiomyopathy were more consistently enriched than LogMiNeR models. Notably, this dataset profiled heart tissue in a non-immune-mediated disease whereas many of our prior knowledge sources were, by design, blood tissue or immune-specific. Because Lasso and Elastic Net can provide additional insights, we suggest that non-network-constrained methods be used alongside LogMiNeR.

Interpretability was increased by each individual network-constrained model while maintaining model accuracy. Others have also reported this increased interpretability of network-constrained regression approaches (Chuang *et al.*, 2007; Li and Li, 2008; Rapaport *et al.*, 2007). The novelty of our approach is that we fit multiple models using different sources of prior knowledge. When calculating interpretability, we noticed that the largest number of enriched sets obtained with LogMiNeR tended to come from the network that corresponds to the same gene set library. On the flu vaccination data, the most enriched KEGG pathways are identified when a KEGG network is used as prior knowledge. In cases where this does not hold, the network is similar to the gene set library. For example, we found that the KEGG network led to the most enriched Reactome sets, likely because both are pathway databases and share many edges in the network. These results suggest that the prior knowledge networks given as input to LogMiNeR should be chosen based on the context used to interpret the models. On all four datasets tested, the accuracy of LogMiNeR was similar to Elastic Net models. These findings are consistent with a survey evaluating eight network-based methods applied to prognostic biomarker discovery which concluded that incorporating prior knowledge does not significantly improve classification accuracy. Surprisingly, LogMiNeR was similarly accurate to Elastic Net models even when gene labels in the network were permuted. This may be explained in part by the selection of a similar number of genes regardless of gene label permutation or by biases from isolated nodes in our prior knowledge networks.

LogMiNeR can be viewed as a hybrid approach between single-gene methods and gene set methods, which calculate set activity and are able to reduce the signal-to-noise ratio (Levine *et al.*, 2006). Each gene in LogMiNeR has its own weight, but these weights are coerced to be similar for genes connected in the prior knowledge network. We demonstrated that LogMiNeR not only improved interpretability on gene sets but also led to identification of predictive genes which were not annotated in existing gene set libraries. Thus, LogMiNeR models have the freedom to discover novel predictive genes while at the same time using prior knowledge to improve interpretability in terms of what is already known. Although we evaluated unweighted networks in this study, LogMiNeR could also be adapted to weighted networks. Furthermore, this method is not limited to gene expression data and can be applied whenever logistic regression is appropriate and prior knowledge networks exist.

In summary, we present a new framework, LogMiNeR, to perform classification while increasing interpretability. We used LogMiNeR to identify B cell-specific genes and mTOR signaling as predictive signatures of human influenza vaccination responses and show that using multiple prior knowledge networks can expand interpretability across many datasets. This framework could provide insights in systems biology studies which focus on finding predictive signatures that are interpretable in terms of prior biological knowledge.

Acknowledgement

We wish to thank Daniel Chawla for careful reading of this article.

Funding

Computational resources and support were provided by the Yale Center for Research Computing [National Institutes of Health grants RR19895 and RR029676-01]. This work was supported by the National Institutes of Health [grant number U19AI089992]. A.C.S. was supported by K24 AG042489. S.A. was supported by the National Science Foundation Graduate Research Fellowship Program [grant number DGE-1122492]. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Conflict of Interest: none declared.

References

Abbas,A.R. *et al.* (2005) Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data. *Genes Immun.*, **6**, 319–331.

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing when researchers tend to select pursuing multiple the (statistically) and support of conclusions. An unguarded use in a greatly results of single-inference inc. *J. R Stat. Soc.*, **57**, 289–300.

Bucasas,K.L. *et al.* (2011) Early patterns of gene expression correlate with the humoral immune response to influenza vaccination in humans. *J. Infect. Dis.*, **203**, 921–929.

Chen,E.Y. *et al.* (2013) Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, **14**, 128.

Chuang,H.-Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Cordero,E. *et al.* (2011) Therapy with m-TOR inhibitors decreases the response to the pandemic influenza A H1N1 vaccine in solid organ transplant recipients. *Am. J. Transplant.*, **11**, 2205–2213.

Croft,D. *et al.* (2014) The Reactome pathway knowledgebase. *Nucleic Acids Res.*, **42**, 472–477.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Furman,D. *et al.* (2013) Apoptosis and other immune biomarkers predict influenza vaccine responsiveness. *Mol. Syst. Biol.*, **9**, 659.

Gorenshteyn,D. *et al.* (2015) Interactive big data resource to elucidate human immune pathways and diseases. *Immunity*, **43**, 605–614.

Gruvberger,S. *et al.* (2001) Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns advances in brief estrogen receptor status in breast cancer is associated with remarkably distinct. *J. Cancer Res.*, **61**, 5979–5984.

Jensen,L.J. *et al.* (2009) STRING 8 - a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**(Suppl. 1), 412–416.

Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.

Lamb,J. *et al.* (2006) The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, **313**, 1929–1935.

Larrañaga,P. *et al.* (2006) Machine learning in bioinformatics. *Brief. Bioinformatics*, **7**, 86–112.

Levine,D.M. *et al.* (2006) Pathway and gene-set activation measurement from mRNA expression data: the tissue distribution of human pathways. *Genome Biol.*, **7**, R93.

Li,C. and Li,H. (2008) Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, **24**, 1175–1182.

Li,S. *et al.* (2013) Systems biological approaches to measure and understand vaccine immunity in humans. *Semin. Immunol.*, **25**, 209–218.

Li,S. *et al.* (2014) Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nat. Immunol.*, **15**, 195–204.

Mannick,J.B. *et al.* (2014) mTOR inhibition improves immune function in the elderly. *Sci. Transl. Med.*, **6**, 268ra179–268ra179.

Mendoza,M.C. *et al.* (2011) The Ras-ERK and PI3K-mTOR pathways: cross-talk and compensation. *Trends Biochem. Sci.*, **36**, 320–328.

Michaud,J. *et al.* (2008) Integrative analysis of RUNX1 downstream pathways and target genes. *BMC Genomics*, **9**, 363.

Milacic,M. *et al.* (2012) Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers*, **4**, 1180–1211.

Nakaya,H.I. *et al.* (2011) Systems biology of vaccination for seasonal influenza in humans. *Nat. Immunol.*, **12**, 786–795.

Obermoser,G. *et al.* (2013) Systems scale interactive exploration reveals quantitative and qualitative differences in response to influenza and pneumococcal vaccines. *Immunity*, **38**, 831–844.

O'Hara,S. *et al.* (2013) Iterative feature removal yields highly discriminative pathways. *BMC Genomics*, **14**, 832.

Pages,H. *et al.* (2015). AnnotationDbi: Annotation Database Interface. R package version 1.30.1.

Pulendran,B. *et al.* (2010) Systems vaccinology. *Immunity*, **33**, 516–529.

Querec,T.D. *et al.* (2009) Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans. *Nat. Immunol.*, **10**, 116–125.

Rapaport,F. *et al.* (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

Ritchie,M.E. *et al.* (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**,

Rönnblom,L. and Eloranta,M.-L. (2013) The interferon signature in autoimmune diseases. *Curr. Opin. Rheumatol.*, **25**, 248–253.

Sasaki,S. *et al.* (2011) Limited efficacy of inactivated influenza vaccine in elderly individuals is associated with decreased production of vaccine-specific antibodies. *J. Clin. Invest.*, **121**, 3109–3119.

Sean,D. and Meltzer,P.S. (2007) GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics*, **23**, 1846–1847.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Shannon,P.T. *et al.* (2013) RCytoscape: tools for exploratory network analysis. *BMC Bioinformatics*, **14**, 217.

Sirota,M. *et al.* (2011) Discovery and preclinical validation of drug indications using compendia of public gene expression data. *Sci. Transl. Med.*, **3**, 96ra77.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, **102**, 15545–15550.

Sun,H. and Wang,S. (2012) Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics*, **28**, 1368–1375.

Thakar,J. *et al.* (2015) Aging-dependent alterations in gene expression and a mitochondrial signature of responsiveness to human influenza vaccination. *Aging*, **7**, 38–52.

Tibshirani,R. (1996) Regression selection and shrinkage via the Lasso. *J. R Stat Soc B*, **58**, 267–288.

Tsang,J.S. *et al.* (2014) Global analyses of human immune variation reveal baseline predictors of postvaccination responses. *Cell*, **157**, 499–513.

Warsow,G. *et al.* (2010) ExprEssence–revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Syst. Biol.*, **4**, 164.

Zhang,J.D. and Wiemann,S. (2009) KEGGgraph: A graph approach to KEGG PATHWAY in R and bioconductor. *Bioinformatics*, **25**, 1470–1471.

Zou,H. and Hastie,T. (2005) Regularization and Variable Selection via the Elastic Net. *J. R Stat. Soc. B (Stat. Methodol.)*, **67**, 301–320.