Bioinformatics, 32(10), 2016, 1509–1517
doi: 10.1093/bioinformatics/btw007
Advance Access Publication Date: 21 January 2016
Original Paper

OXFORD

Systems biology

Multi-task consensus clustering of genome-wide transcriptomes from related biological conditions

Zhen Niu[1,2], Deborah Chasman[2], Amie J. Eisfeld[3], Yoshihiro Kawaoka[3,4] and Sushmita Roy[1,2,5,*]

[1]Department of Computer Sciences, University of Wisconsin–Madison, Madison, WI 53706, USA, [2]Wisconsin Institute for Discovery, University of Wisconsin–Madison, Madison, WI 53715, USA, [3]Influenza Research Institute, Department of Pathobiological Sciences, School of Veterinary Medicine, University of Wisconsin–Madison, Madison, WI, 53711, USA, [4]Division of Virology, Department of Microbiology and Immunology, Institute of Medical Science, University of Tokyo, Tokyo, Japan and [5]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI 53792, USA

*To whom correspondence should be addressed.
Associate Editor: Igor Jurisica

Received on 6 July 2015; revised on 29 December 2015; accepted on 4 January 2016

Abstract

Motivation: Identifying the shared and pathogen-specific components of host transcriptional regulatory programs is important for understanding the principles of regulation of immune response. Recent efforts in systems biology studies of infectious diseases have resulted in a large collection of datasets measuring host transcriptional response to various pathogens. Computational methods to identify and compare gene expression modules across different infections offer a powerful way to identify strain-specific and shared components of the regulatory program. An important challenge is to identify statistically robust gene expression modules as well as to reliably detect genes that change their module memberships between infections.

Results: We present MULCCH (MULti-task spectral Consensus Clustering for Hierarchically related tasks), a consensus extension of a multi-task clustering algorithm to infer high-confidence strain-specific host response modules under infections from multiple virus strains. On simulated data, MULCCH more accurately identifies genes exhibiting pathogen-specific patterns compared to non-consensus and nonmulti-task clustering approaches. Application of MULCCH to mammalian transcriptional response to a panel of influenza viruses showed that our method identifies clusters with greater coherence compared to non-consensus methods. Further, MULCCH derived clusters are enriched for several immune system-related processes and regulators. In summary, MULCCH provides a reliable module-based approach to identify molecular pathways and gene sets characterizing commonality and specificity of host response to viruses of different pathogenicities.

Availability and implementation: The source code is available at https://bitbucket.org/roygroup/mulcch

Contact: sroy@biostat.wisc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Transcriptional regulatory networks play an important role in driving genome-wide mRNA levels in dynamic biological processes, such as response to stress or pathogenic infections. Identification of unique and shared components of transcriptional regulatory networks can be useful for designing more sensitive therapeutics or to exploit shared vulnerabilities of the immune system. Recently, compendia of genome-wide mRNA levels measuring mammalian host response have become available for multiple related conditions that represent responses to similar types of infections (e.g. different influenza viruses) or from similar diseases (e.g. different respiratory diseases; Aevermann *et al.*, 2014). Such datasets can be computationally interrogated to identify common and pathogen-specific components of transcriptional regulatory networks. Transcriptional regulatory networks exhibit a modular organization, defined by sets of genes that are co-expressed under different environmental conditions (Bonneau *et al.*, 2006; Ihmels *et al.*, 2002; Marzorati *et al.*, 2008; Polanski *et al.*, 2014; Segal *et al.*, 2003). Thus a module-based analysis for each condition, followed by a comparison of module patterns across conditions, can provide a systematic identification of the major patterns of conserved and condition-specific transcriptional response between different conditions. Differentially responding genes can be identified as those genes that change their module assignments between conditions. To study condition-specific modules and identify genes that switch module assignments, we need a correspondence between modules across multiple conditions, and, be able to identify statistically robust changes in module assignments between conditions. Consensus clustering is an established procedure for improving the robustness of clustering algorithms (Monti *et al.*, 2003), which are typically used for module identification. However, consensus clustering has thus far not been applied to module identification across multiple conditions. A second challenge is that the individual conditions themselves might be related, which if exploited can improve the confidence in detecting shared and context-specific patterns (Kirk *et al.*, 2012). A third related challenge is that identification of differentially responding genes using module-based analysis have typically focused on two conditions (Amar *et al.*, 2013), however, as the number of conditions increase, it is important to find sets of genes that might be induced in a subset of conditions but repressed in another subset.

To reliably identify condition-specific and common patterns of expression across multiple conditions, such as infection to multiple viruses, we propose MULti-task spectral Consensus Clustering for Hierarchically related tasks (MULCCH), a new consensus clustering extension for existing multi-task clustering algorithms. Multi-task clustering provides a principled machine-learning framework to cluster multiple datasets, one from each condition, simultaneously (Caruana, 1997). Each dataset from a condition is viewed as a separate task, and is clustered separately while information is shared between the tasks. Multi-task extensions have been developed for specific clustering algorithms, including *k*-means (Cai *et al.*, 2013) and mixture models (Bickel and Scheffer, 2004; Roy *et al.*, 2013). Multi-task clustering frameworks can provide a correspondence between the clusters across the datasets, making it straightforward to identify gene sets that respond in a conserved manner across the conditions. Here we provide a consensus clustering extension to an existing multi-task clustering algorithm Arboretum, (Roy *et al.*, 2013), which additionally exploits hierarchical relationships among the clustering tasks (Section 2). Arboretum is a Gaussian mixture model-based multi-task clustering approach that infers modules of co-expressed genes from multiple gene expression datasets, while incorporating the hierarchical relationships across different conditions. The Arboretum algorithm was originally applied to expression datasets from different species in a phylogeny, but is broadly applicable to multi-task clustering problems that are related by a hierarchy. The consensus clustering component of our approach relies on spectral clustering (Von Luxburg, 2007), which is applicable for clustering entities related by a graph. We represent an ensemble of multiple clusterings of the same dataset as a weighted graph of genes, with edge weights representing the confidence of gene pairs that are in the same cluster.

We compared our method to other existing clustering algorithms using simulated and real expression data. We evaluated the algorithms based on well-established measures of cluster quality, as well as new criteria that are important for identifying condition-specific (e.g. virus-specific and strain-specific) genes. Based on simulations, we found that MULCCH and consensus versions of tested clustering algorithms dominated most other clustering algorithms in terms of cluster quality and ability to accurately detect genes that change module assignments. We applied our approach to measurements of host transcriptional response to influenza virus infections in human Calu-3 cell line and mouse lung samples (Aevermann *et al.*, 2014). Our results identified modules with common and pathogenicity-specific patterns that are associated with various immune response processes. Taken together, our approach provides a systematic way to detect common and condition-specific patterns of transcriptional response across multiple conditions, while exploiting known relationships among the conditions.

2 Methods

2.1 Multi-task consensus clustering approach for hierarchically related contexts

To identify common and condition-specific patterns of transcriptional response from different conditions (e.g. host response to diverse viral infections), we developed MULCCH, a multi-task spectral consensus clustering approach. MULCCH accomplishes two goals: (i) identify high-confidence gene modules in each condition and (ii) reliably detect genes that transition between modules from different conditions.

Our approach extends an existing multi-task Gaussian mixture model-based algorithm, Arboretum (Roy *et al.*, 2013). The inputs to Arboretum are (i) a set of gene expression datasets, one for each condition, (ii) a binary tree that describes the relationship between the conditions represented by the datasets and (iii) a desired number of clusters, k. The parameters of Arboretum include transition matrices for each branch in the tree to model the tendency for genes to maintain or switch their assignments from the parent, and the mixture model parameters. These parameters are learned using expectation maximization (EM). The clustering produced by Arboretum gives k clusters for each condition, mapped one-to-one between conditions, which enables easy identification of both co-expression patterns and genes that switch their module membership between conditions. Genes that switch modules between different sets of conditions can indicate pathways associated with strain and pathogen-specific responses.

Figure 1 illustrates our entire approach with a simple example of three datasets, one each for a pathogenic virus (Fig. 1(1)). The relative similarities between the viruses are represented by the tree structure. Our consensus clustering procedure starts by running Arboretum many times separately with different initializations. In an individual Arboretum model, each leaf node represents one

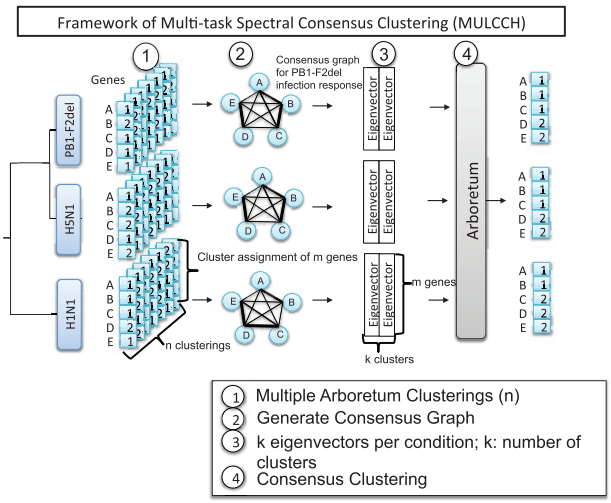

Fig. 1. Overview of MULCCH: MULti-task spectral Consensus Clustering for Hierarchically related tasks. Shown is an example of clusters learned in three conditions representing infections from three viruses of different pathogenicities: H1N1, H5N1, PB1-F2Del. This figure shows the framework of our spectral consensus clustering approach for five genes, A-E. For each condition, Arboretum clusters genes into k modules, $k = 2$ in this example. The main steps of MULCCH are (1) Run Arboretum multiple times on the same datasets to generate multiple clustering solutions; (2) Generate a consensus graph for each condition by evaluating the co-clustering frequency of pairs of genes, thickness of the edge denoting different gene co-clustering frequencies; (3) Extract the top k eigenvectors from each consensus graph; (4) Apply Arboretum again to obtain final consensus clustering

condition, and a Gaussian mixture with k components models the expression data of each module at the leaf node. In Figure 1(1), Arboretum is shown to produce $k = 2$ clusters for each of the conditions. Next, consensus clustering is approached as a spectral clustering problem, which we describe in more detail in the next section. We obtain one consensus graph over genes for each infection condition (Fig. 1(2)), where the edges in each graph are weighted by the frequency at which genes are co-clustered in that condition. We then compute eigenvectors from the Laplacian of the graph for each condition, and use them to generate new data vectors for that condition Figure 1(3). See Supplementary Materials for details. Finally, we run Arboretum on the new data vectors to obtain a consensus multi-task clustering Figure 1(4) consisting of k clusters per virus infection, with a one-to-one mapping between the clusters. Note that gene C exhibits an example of module transition between the pathogenicity groups: it is clustered into Module 1 for H5N1 and PB1-F2del virus infections, and Module 2 for the H1N1 virus infection.

2.2 Spectral consensus clustering

Consensus clustering obtains a consensus result from multiple runs of a clustering algorithm. Our consensus clustering approach relies on spectral clustering (Von Luxburg, 2007), an approach well-suited for clustering entities connected by a graph. We first define a weighted graph of genes to summarize the co-clustering relationships of genes for each condition and apply spectral clustering to this graph.

Let $x_1, \ldots, x_n$ denote the set of genes to be clustered. Suppose we run a clustering algorithm m times (typically, $m = 50$). For every pair of genes clustered, we compute a co-clustering index defined as the fraction of times genes i and j are in the same cluster, divided by m. We represent the co-clustering relationship between all pairs of genes using a weighted graph $G = (V, E)$. Each vertex v_i in this graph is a gene, and each edge is weighted by the co-clustering index between the two genes.

Given a graph, G, spectral clustering makes use of the eigenvalues of the Laplacian matrix of the graph. The Laplacian is a special operator on the graph which is inherently tied to the topology of the graph (Qin and Rohe, 2013; Von Luxburg, 2007). Following (Mohar and Alavi, 1991), we use the symmetric normalized graph Laplacian defined as:

$$L_{\text{sym}} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} A D^{-1/2}$$

where D is an $n \times n$ diagonal degree matrix, each entry $D(i, i)$ is the degree of a vertex, and A is the adjacency matrix, with each entry $A(i, j)$ specifying the status of the edge from i to j. To obtain k clusters, we compute the first k eigenvectors $u_1, \ldots, u_k$ of L_{sym}. Let $U \in R^{n \times k}$ be the matrix containing the vector $u_1, \ldots, u_k$ as columns. For $i = 1, \ldots, n$ let $y_i \in R^k$ be the ith row of U, which is our new data point. Spectral clustering entails applying the K-means clustering algorithm on the new data points $y_1, \ldots, y_n$. However, instead of applying K-means independently to each graph, we apply Arboretum to the sets of eigen vectors (Supplementary Materials).

We considered two versions of the graph to represent the co-clustering relationships: a binary graph and a weighted graph. We create a binary graph by applying a threshold, τ, on the co-clustering index. In the binary graph, we keep edges with weights above that threshold, and set all the weights to be 1. In the weighted graph version, we keep edges weighted above the threshold, but exclude edges with weights below the threshold τ. We experimentally examined different values of $\tau = \{0, 0.1, 0.3, 0.5, 0.7\}$. $\tau = 0$ means that we keep all weighted edges to produce a fully connected weighted graph. Based on multiple cluster evaluation criteria, we determined that a fully connected weighted graph provides the best results (Supplementary Fig. S1).

2.3 Dataset description

We applied our method to expression data measuring host response to six influenza viruses in two host systems: human Calu-3 cell line (GEO Accessions GSE28166, GSE37571, GSE40844, GSE40844, GSE43203, GSE43204) and lungs of 20-week old C57BL/6 mice (GSE33263, GSE37569, GSE37572, GSE43301, GSE44441, GSE44445). The mouse dataset comprised a time series of three or four samples per virus strain and the human cell-line dataset comprised six to nine samples per strain. Values were normalized to paired time points from a mock treatment. Each virus infection represented a different condition.

For the human cell line data, the viruses included three wild-type strains and three laboratory-generated mutants. The wild-type viruses included two low-pathogenicity strains of H1N1, A/California/04/2009 (H1N1) and A/Netherlands/602/2009 (NL), and one high-pathogenicity strain of H5N1, A/Vietnam/1203/2004 (H5N1). The other three viruses were laboratory-generated mutants of H5N1: PB2-627E, NS1trunc124 and PB1-F2del. PB2-627E has a mutation in the PB2 protein that reduces the virus's ability to replicate in mammals. NS1trunc124 has a shortened version of the NS1 protein that hampers the virus's ability to suppress the host antiviral response. PB1-F2del is missing PB1 entirely, which interferes with various viral functions. The mouse data included two wild-type strains, H1N1 and H5N1 and four H5N1 mutant strains. In addition to the mutants mentioned above, there was a mouse-only low-pathogenicity strain, HAavir, which exhibits altered tissue tropism compared to wild-type H5N1.

We modeled the relationship among the viruses according to their pathogenicity. Although the Arboretum algorithm was developed for expression data from multiple species in a phylogeny, it can be applied to any hierarchically related set of datasets. Our motivation for using the pathogenicity tree was to enable a systematic comparison of host response to viruses of different pathogenicities. Because several of the viruses were mutant strains of the same wild-type virus, a phylogenetic tree capturing the sequence divergence of different viruses may not reflect the pathogenicity of the virus, which is measured in the host system. Pathogenicity is measured by minimum lethal dose 50 (MLD_{50}), the dose that is lethal to 50% of infected mice (Tchitchek *et al.*, 2013). Higher doses indicate lower pathogenicity. The viruses represent different groups of pathogenicity: low (H1N1 and NL), medium (PB2-627E and NS1trunc124) and high (H5N1 and PB1-F2del). The pathogenicity tree used in our analysis connected the high and medium pathogenicity viruses first, and then connected the root of these two groups with the low pathogenicity group.

We applied an initial pre-processing on the original expression data to focus on genes that have orthologs between human and mouse, and that were differentially expressed between two pathogenicity groups in either species. We assessed differential expression with a *T*-test, $P < 0.01$ for human and $P < 0.05$ for mouse. We chose these different thresholds to obtain similar number of genes for each host system, $P < 0.05$ being too generous for the human cell line data and $P < 0.01$ being too restrictive for the more heterogeneous mouse data. These thresholds created a dataset of 7192 human genes and 7240 mouse genes.

2.4 Simulation experiments

We used simulations to evaluate our multi-task consensus clustering approach for two properties: (i) the ability of consensus clustering to recover better clusters and (ii) the ability to reliably detect genes that switch between clusters from different strains. We used Arboretum's generative model to simulate expression datasets along with simulated true cluster assignments. Using those cluster assignments, we then identified genes that transitioned between clusters for pairs of conditions. The parameters of Arboretum's generative model were learned from the real human cell line expression data with $k = 5$ clusters (covering 7192 genes). Therefore, the simulated data had the same structure as the real expression data, with six viruses each contributing a short time course. In order to ensure that the simulated data had good cluster separation, we first generated ten simulated datasets and sorted them based on average silhouette index, a measure of cluster separation. We used the two simulated datasets with the best cluster separation for our experiments.

2.5 Evaluation of clusters and transitions between clusters

We used multiple cluster evaluation criteria to assess the quality of clusters from different consensus, non-consensus, multi-task and nonmulti-task settings. These include (i) silhouette index, (ii) module stability and (iii) expression coherence. Silhouette index is computed for each data point i and measures the sharpness of boundaries of the clustering as $\frac{a_i - b_i}{\max(a_i, b_i)}$, where a_i is the average dissimilarity of i with all other data points in i's cluster and b_i is the average dissimilarity of i to the next best cluster of which i is not a member. We used two distance measures to measure the dissimilarity of genes: Euclidean distance and Pearson correlation. The larger the silhouette index, the better the clustering. Cluster stability is defined as the proportion of gene pairs that are in the same module under different initializations. Expression coherence is defined as the average proportion of genes whose expression profile has >0.8 correlation with the module's mean.

We introduced a fourth evaluation to measure the ability of each method to detect genes that change their module between different conditions. An important property of our multi-task clustering approach is that it outputs a one-to-one mapping between clusters from multiple clustering tasks. This is beneficial because the cluster ID then corresponds to a phenotype (pattern) of expression that can be compared across multiple conditions. In particular, the transitions of a gene between different clusters across conditions is indicative of differential expression. In order to test the ability of our method to reliably detect such transitions between different clusters, we used the simulated data described above. The simulated datasets provide the correct cluster assignment and are used to evaluate a clustering method's ability to detect true transitions in modules between any pair of conditions. We used precision, recall and *F*-score to quantify the predicted switched genes/transitions between each pair of conditions. Let S_{ab} denote the set of genes that transition their cluster assignment between conditions, a and b, and let S'_{ab} represent the set of genes predicted to transition between clusters inferred using the data from conditions a and b. Precision, P_{ab}, for each pair of conditions, a and b is $\frac{|S_{ab} \cap S'_{ab}|}{|S'_{ab}|}$, recall R_{ab} is $\frac{|S_{ab} \cap S'_{ab}|}{|S_{ab}|}$, and *F*-score is $\frac{2 * P_{ab} * R_{ab}}{P_{ab} + R_{ab}}$. For all three measures, higher values mean better results.

While Arboretum provides a natural mapping of cluster IDs across different conditions, the standard Gaussian mixture model does not. We mapped the clusters IDs between different conditions by computing a mean and variance value for each cluster and then applying the Hungarian matching algorithm with Euclidean distance. Because each strain had different number of experimental time points, we collapsed the per time point mean and variance to the average across time points.

2.6 Number of clusters

For all clusterings, we set k, the desired number of clusters, to five. We evaluated our choice by empirically trying to learn k from the range {5, 6, 7}. We found that five clusters performed the best based on the silhouette index and the ability to match cluster patterns between virus strains (Supplementary Table S1). This choice was also motivated by our particular datasets, in which the expression values of each gene tended to be constant or change monotonically over time within each viral infection. Clustering can then be seen as an extension of differential expression assessment. Instead of dividing genes into only two clusters (differentially expressed and not), a five-way clustering divides genes according to both sign of expression change and magnitude.

2.7 Clustering algorithms assessed

We compared several configurations of clustering algorithms that differed in whether they performed multi-task or consensus clustering, or how they performed consensus clustering. To assess the advantage of multi-task clustering over single task clustering in the context of immune response to infection from multiple influenza virus strains, we compared the original Arboretum to independent Gaussian mixture models (GMMs). The Arboretum algorithm was already shown to be more advantageous than traditional clustering in a multi-species context (Roy *et al.*, 2013); here our experiments tested this property in the application of multi-task learning to analyze multiple host response datasets. To assess methods of extracting consensus clusters from repeated clustering runs, we compared

spectral clustering to hierarchical clustering, which has been previously used for consensus clustering (Nguyen and Caruana, 2007). In both cases, we used the consensus edge weights as pairwise similarities between data points. In all cases, GMM clustering was used as the base line clustering algorithm. The complete set of tested configurations are:

- GMM (separate clustering for each of the conditions; neither multi-task nor consensus)
- Arboretum (multi-task clustering, non-consensus)
- MULCCH (multi-task clustering with spectral consensus clustering)
- Consensus GMM with spectral clustering (single task clustering for each condition independently, with spectral consensus clustering)
- Consensus GMM with hierarchical clustering (single task clustering for each condition independently, with hierarchical consensus clustering)

2.8 Analysis of host-response modules

Enrichment of modules for curated gene sets was tested using a hypergeometric test followed by Benjamini Hochberg correction for multiple hypothesis correction. A corrected P-value <0.05 threshold was used to call a module enriched in a process. The same criteria was applied to gene sets exhibiting the same pattern of module reassignment between pathogenicity groups. To assess significance of module similarity and reassignment of genes we defined a score S_{ij}^{XY}, for module i of strain X and module j of another strain Y, $S_{ij}^{XY} = \frac{-(\log(p_1)+\log(p_2))}{2}$. Here p_1 and p_2 are from a hypergeometric test of overlap significance. Pairs of modules with a score $S_{ij}^{XY} > 20$ are discussed in the text which corresponds to *P*-value <1E-8. These score values are displayed for all pairs of viruses in Figure 5B and Supplementary Figure S5B.

3 Results

3.1 Multi-task consensus clustering can more reliably detect transitions between modules

We first compared the ability of the clustering algorithms to reliably detect transitions between pairs of conditions using simulated data. We compared Arboretum, MULCCH, GMM, and Consensus GMM by computing precision, recall and *F*-score to evaluate the predicted transitions between pair of strains. For six different strains, there are $\binom{6}{2}$ possible combinations. F-scores are shown in Figure 2 as 15 box plot pairs and precision and recall shown in Supplementary Figure S2. Each subplot shows the distribution of F-score given by different algorithms for the switches between a pair of conditions estimated from 50 runs of each algorithm. We observed that for most of these 15 comparisons, the MULCCH *F*-scores are much higher than the average line of Arboretum, which supports that MULCCH performs better than Arboretum in terms of the ability to predict switched genes. On the other hand, the *F*-scores of Consensus GMM are located about the same height as the average of GMM for most of 15 runs; that is, we did not see much improvement of Consensus GMM compared with GMM. A similar result is also found for another simulated dataset (Supplementary Fig. S3), in which MULCCH dominated Arboretum while Consensus GMM performed about as well as or better than GMM. Importantly, MULCCH performed better than non-consensus clustering (either Arboretum or GMM), suggesting that the multi-task framework together with the consensus clustering framework enables us to infer high-confidence cluster transitions.

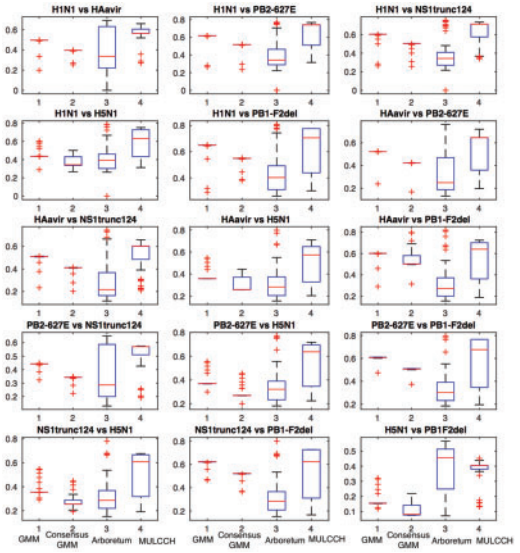

Fig. 2. Comparison of consensus and non-consensus clustering methods for reliably detecting cluster transitions. Each subplot shows a comparison of a pair of viruses: H1N1, H5N1, HAavir, PB2-627E, NS1trunc124, and PB1-F2del. Box plots represent the distribution of F-score of 50 runs each for four methods compared: two non-consensus methods (Arboretum and GMM) and two consensus methods (MULCCH and consensus GMM). The higher the F-score the better the ability to detect cluster transitions between different virus strains

3.2 Spectral consensus clustering produces better clusters than hierarchical consensus clustering

We next compared spectral consensus clustering (using either Consensus GMM or MULCCH) against Hierarchical Consensus Clustering (Section 2). We performed these comparisons on the real expression data measuring genome-wide transcriptome levels in Calu-3 cells under infection with influenza viruses of different pathogenicities (Section 2.3). Specifically, we applied Arboretum, MULCCH, Consensus GMM, GMM and Consensus Hierarchical clustering algorithms on the six influenza expression time courses and used silhouette index to evaluate the clustering quality. We used two measures of dissimilarity between pair of genes: Pearson correlation distance (Fig. 3) and Euclidean distance (Supplementary Fig. S4). We found that both consensus and non-consensus GMM and Arboretum yielded similarly good results, except for on one virus, PB2-627E. For that virus, Consensus GMM produced better clustering (see second row, first column). On the other hand, consensus hierarchical clustering performed the worst compared to all others. We found similar results using Euclidean distance (Supplementary Fig. S4). These results suggest that the specific strategy used to extract consensus clusters from multiple clustering results can greatly influence the results. Importantly, our graph-based spectral clustering approach performed significantly better than the hierarchical clustering approach.

3.3 Consensus clustering identifies high quality and stable clusters compared to non-consensus approaches

We next compared the clustering results using two additional metrics: (i) expression coherence (Fig. 4A, C) and (ii) module stability (Fig. 4B, D). Consensus methods performed well in all measures.

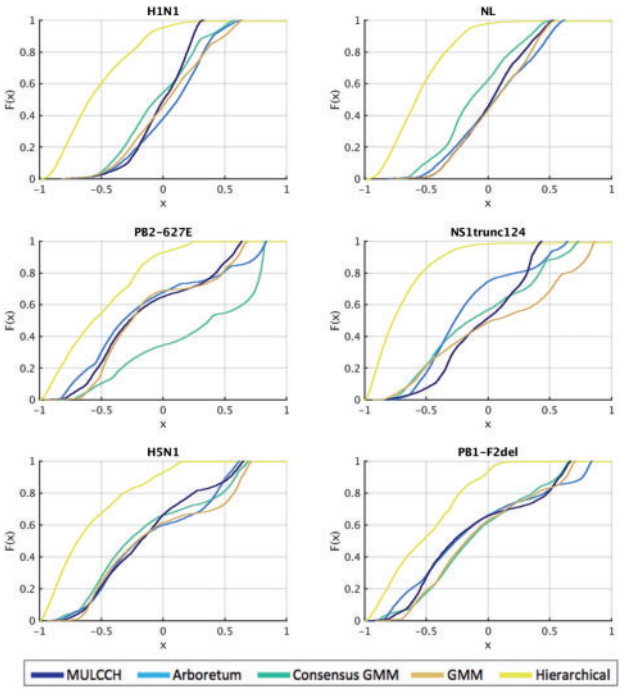

Fig. 3. Comparison of methods using silhouette index. Shown is a comparison of MULCCH's performance (dark blue) to that of Consensus GMM (green), Arboretum (light blue), GMM (brown) and Consensus Hierarchical Clustering (yellow). The X axis represent the silhouette value and Y axis represent the proportion of genes that get equal or higher silhouette value. The more right shifted the curve is, the better clustering it is

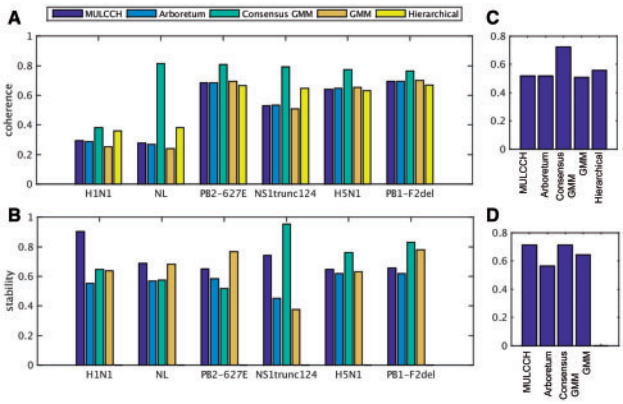

Fig. 4. Comparison of MULCCH to other clustering algorithms. MULCCH (dark blue), Consensus GMM (green), Arboretum (light blue), GMM (brown) and Hierarchical Consensus Clustering (yellow) were compared on the basis of expression coherence (A,C) and module stability (B,D). **(A)** shows expression coherence measured as the average proportion of genes whose expression profiles had >0.8 correlation with the module's mean using correlation distance; **(B)** shows stability, which measured the proportion gene pairs that are in the same module under different random initializations. **(C)** Averaged coherence over all six strains; **(D)** Averaged stability over all six strains. Hierarchical clustering will always have a stability of 1 and is therefore not shown in B and D

First, modules inferred by consensus methods were more stable or at least as stable as non-consensus methods for four of the six strains (H1N1, NL, PB2-627E, PB1-F2del). MULCCH displays much better stability than simple Arboretum, as shown by the large difference in the proportion of consistently co-clustered gene pairs (dark blue

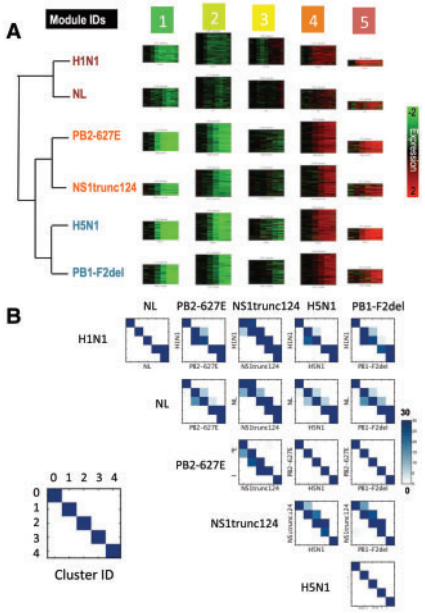

Fig. 5. Modules identified by MULCCH from human Calu-3 cell line expression data. **(A)** shows the tree structure among six viruses, and five modules for each virus. Each row represents a virus and each column represents a module. Modules with similar expression profiles received the same cluster ID across six virus strains. **(B)** shows the similarity of modules between any two conditions. Off-diagonal elements show similarity between modules of different IDs, and diagonal elements show similarity between modules of same ID

bars compared to light blue bars in Fig. 4). Second, in terms of expression coherence, Consensus GMM achieved the best result for all six strains among these four methods. MULCCH produced equally coherent modules as non-consensus Arboretum. In summary, compared to a non-multi task approach (consensus GMM), MULCCH recovers more accurate module transitions, and compared to a nonconsensus approach (Arboretum), MULCCH has higher stability and silhouette index, maintaining the same levels of expression coherence. Thus, combining multi-task and consensus clustering makes MULCCH an overall more powerful approach than traditional clustering as well as their consensus versions.

3.4 Shared immune response modules across viral infections

We applied our method to human and mouse expression data measuring transcriptional response under infection to influenza virus strains of different pathogenicities (Section 2.3). Figure 5A shows the modules identified from the human Calu-3 cell-line data. In all six viruses, we observed five major patterns of expression: strong repression (modules 0 and 1), mild repression (module ID 2), moderate induction (module ID 3) and strong induction (module ID 4). We observed similar patterns of expression in mouse lung as well, although mouse modules exhibited greater similarity in the expression patterns (Supplementary Fig. S5A).

Although the general pattern of expression is similar for modules of the same ID, the genes within each module might not be same. To this end we assessed the significance of module overlap between pairs of viruses (Section 2, Fig. 5B and Supplementary Fig. S5B). When modules of the same ID exhibit high overlap, matrices in Figure 5B have strong diagonals. All pairs of strains exhibit strong

diagonal elements, suggesting that there are common genes between the same module IDs in different conditions. We also observe off-diagonal elements in some human modules. This suggests that the specific genes assigned to the same module ID across two conditions are generally the same, however, there are several genes that switch their module assignments (e.g. genes that switch from module 4 to 5 between H1N1 and NL, Fig. 5B).

To biologically interpret these modules we tested them for statistical enrichment of curated gene set from Gene Ontology, KEGG, REACTOME and Biocarta pathways (Croft *et al.*, 2011; Gene Ontology Consortium, 2015; Kanehisa *et al.*, 2014; Liberzon *et al.*, 2011), transcription factor binding motifs from MSigDB (Liberzon *et al.*, 2011) and gene sets identified by various screening studies (Brass *et al.*, 2009; de Chassey *et al.*, 2013; Hao *et al.*, 2008; Karlas *et al.*, 2010; König *et al.*, 2010; Shapira *et al.*, 2009; Sui *et al.*, 2009; Tafforeau *et al.*, 2011; Watanabe *et al.*, 2014; Zhang *et al.*, 2009). Module 5, which exhibited conserved induction in all six virus strains, was enriched for immune response related processes and GPCR signaling (Supplementary Table S2). Module 3, exhibiting mild repression, was associated with several REACTOME pathways in all but H1N1. Such pathways have been observed to be down regulated in cellular response to pathogens (e.g. proteasome regulation (Widjaja *et al.*, 2010), destabilization of mRNA (Cathcart and Semler, 2014); Supplementary Table S2). Similarly, Module 5 of mouse, which exhibited strong induced expression, is associated with innate immune response processes such as NFKB signaling and cytokine signaling (Supplementary Table S3). Furthermore, in both human and mouse the most induced module was enriched for the binding sites of IRF7 and ISRE, which are important regulators of innate immune response genes.

3.5 Gene sets exhibiting pathogenicity specific patterns

We next examined the modules and genes for differential pattern of expression. We did this in two ways: comparing between pairs of strains, as well as between groups of strains. For pairs of strains, we examined the off-diagonal elements in Figure 5B and observed several off-diagonal elements that indicate that a significant number of genes switch their module assignments between viruses. For example, between the low pathogenicity virus H1N1 and the medium pathogenicity virus PB2-627E, we found transitions between Module 4 (mild induction) and Module 5 (high induction). Most of these transitions represented changes in the level of expression but not in direction. That is, we generally observed changes involving the mild repression and strong repression modules or mild and strong induction modules.

One virus participated in several notable transitions: NS1trunc124, a medium pathogenicity virus. We observed a significant fraction of genes that switched between Module 3 (mild repression) in both low pathogenicity viruses (H1N1 and NL) to Module 4 of NS1trunc124 (moderate induction). Similarly, we find that Module 4 of PB2-627E has significant overlap with Module 3 (moderate repression) of NS1trunc124. Overall, we found that NS1trunc124 had the greatest number of off-diagonal elements. This is consistent with NS1trunc124's global expression profile being the least correlated with any of the other viruses (Supplementary Fig. S6) as well as the knowledge that the NS1trunc124 mutation impeded the virus's ability to modulate host antiviral response (Tchitchek *et al.*, 2013). Importantly, although there is little correlation of NS1trunc124's transcriptional host response with the transcriptional response of other viruses, we still identify conserved components of this virus's response with other virus responses.

To systematically identify pathways and genes that are differentially expressed between more than two viruses, we compared module assignments for sets of viruses of different pathogenicity levels using a rule-based pattern extraction approach. Specifically, we represented each gene's module assignment across all six viruses as a string of six numbers, each number denoting a module assignment. We searched these strings for patterns where genes were in one module in one subtree (defined by the virus hierarchy in Fig. 5A), but a different module in another subtree. For example, the pattern representing genes in Module 3 (mild repression) in the low pathogenicity viruses and Module 4 (mild induction) in the high and medium pathogenicity viruses is 334444. Rule-based pattern extraction to individual genes identified 871 genes that exhibited a pathogenicity-specific pattern, representing ≈12% of the total number of genes that were analyzed. We focused on patterns that represented transitions between clusters with opposite signs of expression. The pattern 334444 encoding repression (Module 3) in the low pathogenicity viruses and mild induction (Module 4) in the high and medium pathogenicity viruses had the largest number of genes (162 genes). Genes in this pattern were associated with chromatin modification and RNA splicing machinery. Another interesting signature was 224444, which exhibited a transition between reduced expression in low pathogenicity viruses to induced expression in high and medium pathogenicity viruses (Supplementary Dataset S1). The genes having this pattern were enriched for the REACTOME cell cycle pathway. A third signature pattern of note was 115555, which indicates repression in the low pathogenicity viruses and induction in all other viruses (Fig. 6). While as a whole they were not enriched for any specific pathway, several of these genes were associated with receptor activity, cell migration and adhesion and immune response processes (represented by the genes HEY1, EPHB3). These gene sets exhibit examples of pathogen-specific expression and implicate splicing, cell cycle and chromatin modification related processes to be differentially regulated between the different pathogenicity groups.

We also applied our rule-based pattern extraction approach to entire modules by identifying pathways or GO processes associated with a particular pathogenicity type. We find numerous annotation categories that were associated with different pathogenicity groups (Table 1, Supplementary Tables S2 and S3), that are indicative of host immune response pathways (Type 1 interferon signaling), host factors that are part of influenza viral replication, as well as immune response transcription factors. Interestingly, we found that the influenza transcription and replication gene sets were enriched only for the high pathogenicity viruses in the mildly induced expression modules. Among the transcription factors whose motifs were enriched in different modules were important immune response regulators

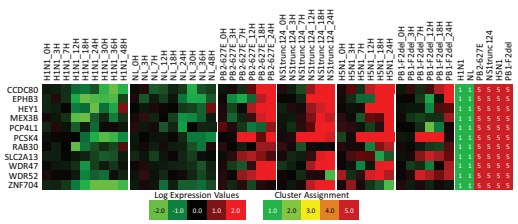

Fig. 6. Example of a pattern (signature string 115555) of genes exhibiting changes in module assignment between low pathogenicity viruses and all others. Shown are a set of genes that switched their module assignment from a repressed module (1) to an induced module (5) between the low and high/medium pathogenicity viruses

Table 1. Selected annotation categories that are differentially associated with host response to viruses of different pathogenicities; each pathogenicity group consisting of two viruses

		Pathogenicity		
Source	Annotation	Low	Medium	High
GO process	Transport	2		
	Actin mediated movement		5	
	Type 1 interferon	5		
	G1/S checkpoint	3		
	Respiration			5
REACTOME	Influenza life cycle			4
	Viral transcription and replication			4
Motif	LFA1			5
	E12		5	5
	FOXO4		5	5
	GATA		5	5
	OCT1/POU2F1		5	5

The numbers in each row specify the module in which the process was enriched. Only processes that were enriched in both viruses of a pathogenicity group were considered.

(LFA1, E12, OCT1) as well regulators that are specifically associated with the myeloid cell lineage (GATA, FOXO4).

4 Discussion

With our ever increasing ability to profile transcriptional responses from multiple diverse biological contexts—different diseases, cell types, or infections—the need to identify common and context-specific regulatory network components is becoming increasingly important. This is especially a challenge in systems biology studies of infectious disease research where such shared and unique network components can guide common treatment regimens, as well as identify host response aspects that are triggered as a function of strain virulence. To address this challenge, we proposed MULCCH, a multi-task consensus clustering approach to identify sets of host genes that are co-expressed in a similar manner under infection with multiple viruses, as well as to identify genes that exhibit differential patterns of expression between subsets of virus infections. Application of our approach to human and mouse transcriptomic data enabled us to find genes and processes that responded in a virus- or pathogenicity-specific manner.

Module-based analysis of transcriptional responses is a powerful approach to identify the major patterns of expression during a dynamic process and several approaches have been developed to identify these modules from one condition (Marzorati *et al.*, 2008). However, when considering multiple labeled conditions, such as the different viruses in our problem, methods that can use existing condition labels while finding gene groups within each condition are needed (Bickel and Scheffer, 2004; Roy *et al.*, 2013). The benefit of multi-task clustering is that it is more likely to find commonalities in cases where an independent clustering might not. A global correlation of profiles of human Calu-3 cell response to different influenza infections showed that one medium pathogenicity virus, NS1trunc124, was not correlated with the other similarly pathogenic viruses (Supplementary Fig. S6). However, using our approach, we were still able to find modules in this virus that significantly overlapped with modules exhibiting the same pattern from other viruses. This suggests that there can be subsets of genes that are correlated despite there being little correlation between genome-wide profiles.

The ability to reliably detect module transitions for individual genes is extremely important. Because clustering approaches in general are known to be prone to reaching local optima, we used a consensus clustering approach to extract high confidence cluster assignments and cluster transitions. The choice of the consensus clustering algorithm was important and influenced the quality of results. In particular, the spectral method produced better clusters than the hierarchical method as measured by silhouette index. While consensus GMM was able to learn stable clusters, it was not able to identify module transitions reliably, likely because it does not solve the matching problem across strains, which we had to do as a post clustering step. Incorporating the pathogenicity relationships using the hierarchical relatedness approach of the Arboretum framework was useful to identify genes exhibiting patterns of pathogen specificity. Instead of considering all possible types of transitions, we applied our rule-based pattern extraction procedure on virus groups defined by pathogenicity subtrees to identify genes that exhibited the same module assignment between the same group of pathogenicity, but differed in module assignment between pathogenicity groups. Most genes that exhibit such patterns are similar among the medium and high pathogenicities.

In this work, we relied on the accuracy of the pathogenicity tree to relate the viruses. One direction of future research is to explore simultaneous learning of modules as well as the tree topologies to relate the different viruses. An alternate route would be to use multi-task feature learning approaches (Argyriou *et al.*, 2008), that do not assume any structure on the tasks and aim to learn most informative features across all tasks simultaneously. Another promising extension to the model would be to integrate different types of -omic measurements (e.g. proteomics, transcriptomics, metabolomics), with appropriate data specific generative models. In conclusion, our consensus-based multi-clustering approach is a novel and robust approach to identify common and context-specific sets of co-expressed genes that can be applied to large numbers of related viruses. As systems biology studies expand to additional viruses, approaches such as ours will be increasingly useful for guiding targeted therapeutics.

Acknowledgements

This project is funded in part with funds from the National Institute of Allergy and Infectious Diseases, National Institutes of Health, grant U19AI106772 to YK, and in part by University of Wisconsin–Madison startup funds and a Sloan Foundation fellowship to SR.

Conflict of Interest: none declared.

References

Aevermann,B.D. *et al.* (2014) A comprehensive collection of systems biology data characterizing the host response to viral infection. *Scientific Data*, **1**, 140033.

Amar,D. *et al.* (2013) Dissection of regulatory networks that are altered in disease via differential co-expression. *PLOS Comput. Biol.*, **9**, e1002955

Argyriou,A. *et al.* (2008) Convex multi-task feature learning. *Mach. Learn.*, **73**, 243–272.

Bickel,S. and Scheffer,T. (2004) Multi-View Clustering. *ICDM*, **4**, 19–26.

Bonneau,R. *et al.* (2006) The Inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets *de novo*. *Genome Biol.*, **7**, R36.

Brass,A.L. *et al.* (2009) The IFITM proteins mediate cellular resistance to influenza A H1N1 virus, West Nile virus, and dengue virus. *Cell*, **139**, 1243–1254.

Cai,X. *et al.* (2013). Multi-view k-means clustering on big data. In: Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), pp. 2598–2604. AAAI Press.

Caruana,R. (1997) Multitask learning. *Mach. Learn.*, **28**, 41–75.

Cathcart,A.L. and Semler,B.L. (2014) Differential restriction patterns of mRNA decay factor AUF1 during picornavirus infections. *J. Gen. Virol.*, **95**, 1488–1492.

Croft,D. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.

de Chassey,B. *et al.* (2013) The interactomes of influenza virus NS1 and NS2 proteins identify new host factors and provide insights for ADAR1 playing a supportive role in virus replication. *PLoS Pathog.*, **9**, e1003440.

Gene Ontology Consortium (2015) Gene Ontology Consortium: going forward. *Nucleic Acids Res.*, **43**, D1049–D1056. (Database issue),

Hao,L. *et al.* (2008) Drosophila RNAi screen identifies host genes important for influenza virus replication. *Nature*, **454**, 890–893.

Ihmels,J. *et al.* (2002) Revealing modular organization in the yeast transcriptional network. *Nat. Genet.*, **31**, 370–377.

Kanehisa,M. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205. (Database issue),

Karlas,A. *et al.* (2010) Genome-wide RNAi screen identifies human host factors crucial for influenza virus replication. *Nature*, **463**, 818–822.

Kirk,P. *et al.* (2012) Bayesian correlated clustering to integrate multiple datasets. *Bioinformatics*, **28**, 3290–3297.

König,R. *et al.* (2010) Human host factors required for influenza virus replication. *Nature*, **463**, 813–817.

Liberzon,A. *et al.* (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

Marzorati,M. *et al.* (2008) How to get more out of molecular fingerprints: practical tools for microbial ecology. *Environ. Microbiol.*, **10**, 1571–1581.

Mohar,B. and Alavi,Y. (1991) The Laplacian spectrum of graphs. *Graph Theory Combin. Appl.*, **2**, 871–898.

Monti,S. *et al.* (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. *Mach. Learn.*, **52**, 91–118.

Nguyen,N. and Caruana,R. (2007). Consensus Clusterings. In: Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007), pp. 607–612.

Polanski,K. *et al.* (2014) Wigwams: identifying gene modules co-regulated across multiple biological conditions. *Bioinformatics*, **30**, 962–970.

Qin,T. and Rohe,K. (2013). Regularized spectral clustering under the degree-corrected stochastic blockmodel. *Adv. Neural Inf. Process. Syst.* **26**, 3120–3128.

Roy,S. *et al.* (2013) Arboretum: reconstruction and analysis of the evolutionary history of condition-specific transcriptional modules. *Genome Res.*, **23**, 1039–1050.

Segal,E. *et al.* (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat. Genet.*, **34**, 166–176.

Shapira,S.D. *et al.* (2009) A physical and regulatory map of host-influenza interactions reveals pathways in H1N1 infection. *Cell*, **139**, 1255–1267.

Sui,B. *et al.* (2009) The use of Random Homozygous Gene Perturbation to identify novel host-oriented targets for influenza. *Virology*, **387**, 473–481.

Tafforeau,L. *et al.* (2011) Generation and comprehensive analysis of an influenza virus polymerase cellular interaction network. *J. Virol.*, **85**, 13010–13018.

Tchitchek,N. *et al.* (2013) Specific mutations in H5N1 mainly impact the magnitude and velocity of the host response in mice. *BMC Syst. Biol.*, **7**, 69.

Von Luxburg,U. (2007) A tutorial on spectral clustering. *Stat. Comput.*, **17**, 395–416.

Watanabe,T. *et al.* (2014) Influenza virus-host interactome screen as a platform for antiviral drug development. *Cell Host Microbe*, **16**, 795–805.

Widjaja,I. *et al.* (2010) Inhibition of the ubiquitin-proteasome system affects influenza A virus infection at a postfusion step. *J. Virol.*, **84**, 9625–9631.

Zhang,L. *et al.* (2009) Systems-based candidate genes for human response to influenza infection. *Infect. Genet. Evol.*, **9**, 1148–1157.