frontiers in Genetics

ORIGINAL RESEARCH
published: 09 May 2016
doi: 10.3389/fgene.2016.00080

Recursive Indirect-Paths Modularity (RIP-M) for Detecting Community Structure in RNA-Seq Co-expression Networks

Bahareh Rahmani[1], *Michael T. Zimmermann*[2,3], *Diane E. Grill*[2,3], *Richard B. Kennedy*[3], *Ann L. Oberg*[2,3], *Bill C. White*[1], *Gregory A. Poland*[3] *and Brett A. McKinney*[1,4]*

[1] Tandy School of Computer Science, University of Tulsa, Tulsa, OK, USA, [2] Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, USA, [3] Mayo Clinic Vaccine Research Group, Mayo Clinic, Rochester, MN, USA, [4] Department of Mathematics, University of Tulsa, Tulsa, OK USA

OPEN ACCESS

Edited by:
Galina Glazko,
University of Arkansas for Medical Sciences, USA

Reviewed by:
Peter Salzman,
University of Rochester Medical Center, USA
Yasir Rahmatallah,
University of Arkansas for Medical Sciences, USA

***Correspondence:**
Brett A. McKinney
brett-mckinney@utulsa.edu

Specialty section:
This article was submitted to Bioinformatics and Computational Biology,
a section of the journal Frontiers in Genetics

Received: 30 November 2015
Accepted: 20 April 2016
Published: 09 May 2016

Citation:
Rahmani B, Zimmermann MT, Grill DE, Kennedy RB, Oberg AL, White BC, Poland GA and McKinney BA (2016) Recursive Indirect-Paths Modularity (RIP-M) for Detecting Community Structure in RNA-Seq Co-expression Networks. Front. Genet. 7:80. doi: 10.3389/fgene.2016.00080

Clusters of genes in co-expression networks are commonly used as functional units for gene set enrichment detection and increasingly as features (attribute construction) for statistical inference and sample classification. One of the practical challenges of clustering for these purposes is to identify an optimal partition of the network where the individual clusters are neither too large, prohibiting interpretation, nor too small, precluding general inference. Newman Modularity is a spectral clustering algorithm that automatically finds the number of clusters, but for many biological networks the cluster sizes are suboptimal. In this work, we generalize Newman Modularity to incorporate information from indirect paths in RNA-Seq co-expression networks. We implement a merge-and-split algorithm that allows the user to constrain the range of cluster sizes: large enough to capture genes in relevant pathways, yet small enough to resolve distinct functions. We investigate the properties of our recursive indirect-pathways modularity (RIP-M) and compare it with other clustering methods using simulated co-expression networks and RNA-seq data from an influenza vaccine response study. RIP-M had higher cluster assignment accuracy than Newman Modularity for finding clusters in simulated co-expression networks for all scenarios, and RIP-M had comparable accuracy to Weighted Gene Correlation Network Analysis (WGCNA). RIP-M was more accurate than WGCNA for modest hard thresholds and comparable for high, while WGCNA was slightly more accurate for soft thresholds. In the vaccine study data, RIP-M and WGCNA enriched for a comparable number of immunologically relevant pathways.

Keywords: sequence analysis, RNA, gene expression profiling, newman modularity, weighted gene correlation network analysis, WGCNA, algorithms

INTRODUCTION

Modularity is a property of many complex systems where the components of the system are organized into functional subunits or modules. Modular organization can be observed in engineered systems like hardware components of a computer, packages in software, or parts of a vehicle. But modularity is also observed in evolved systems like DNA into chromosomes, spatial regions of the brain into specialized functions, or genes into transcriptional regulation networks (Shen-Orr et al., 2002). It is commonly held that cellular organization and biochemical function is modular in nature (Hartwell et al., 1999; Mitra et al., 2013). There are likely multiple selective pressures that lead to modularity in evolved systems, one of which may be the frequency of a changing environment (Parter et al., 2007). Intuitively, if an environment is relatively static, an evolving system has the luxury to build large modules, whereas in a rapidly changing environment, there is a greater advantage to building and reusing smaller, robust functional subunits (modules).

RNA-Seq is a next-generation sequencing technology for the genome-wide quantification of gene expression that is rapidly growing in prevalence and has some advantages over microarrays in characterizing transcriptomes (Ozsolak and Milos, 2011; Hitzemann et al., 2013). There are also many challenges to analyzing RNA-Seq data and few studies have investigated the network properties of RNA-seq based co-expression networks (Ballouz et al., 2015). One of the perennial statistical challenges of clustering data-driven networks (e.g., gene co-expression) is how optimally to determine the appropriate number and size of clusters. Hierarchical clustering relies on the relative distance between genes or samples, represented as trees. Clustering can be described as "cuts" of these trees, which can result in single clusters that encompass the majority of genes, or conversely, many small (potentially singleton) clusters. Accordingly, one of the motivations of the Weighted Gene Correlation Network Analysis (WGCNA) strategy is to avoid small clusters, enabled by parameters to specify a minimum module size and merge cut height (Langfelder et al., 2013). Model-based clustering uses likelihood and model-complexity to guide the user to an appropriate number of clusters and may be put into a hierarchical framework (Guo et al., 2010), but the size of the clusters may still have extremes.

Newman Modularity is a spectral clustering algorithm that performs multiple binary splits of the network and determines cluster memberships by optimizing a quadratic-form of the difference between observed and expected connections (under null model assumptions). Submodules are found automatically through recursive binary splits (Newman, 2006). Automated cluster number estimation is an attractive feature of modularity (data-driven); however, when applied to the RNA-Seq co-expression networks in the current study, modularity identifies a small number of large components that encompass the bulk of genes (e.g., **Figure 8**). To uncover the biologically relevant substructures, larger modules must be split and smaller modules merged in an optimal way.

The implicit goal of spectral clustering is to find a balance between groups in order to find more realistic communities and avoid trivial solutions (Bolla, 2011). For example, the potential for trivial solutions led to the proposal of normalized cut penalties to avoid minimal cut clusters that contain a single node separated from the rest of the nodes (Shi and Malik, 2000). In the case of gene co-expression networks, large modules may lead to the enrichment of functional categories that are too broad to be biologically informative for the given study. For example, when we use various settings of a hard threshold, Newman Modularity results in modules with over 1000 genes. Such large clusters should be further resolved to map onto more specific functional categories. At the same time, very small clusters may need to be broadened to identify statistically significant overlap with functional biological categories.

In the current study, we develop a generalized indirect-paths form of modularity and incorporate it into a merge-and-split algorithm we call "recursive indirect-paths modularity" (RIP-M). The RIP-M algorithm varies the order of finite power series of the adjacency matrix, which adaptively merges small clusters by incorporating indirect path evidence for module membership. Recent studies using powers of the adjacency matrix include subgraph centrality (Estrada and Rodriguez-Velazquez, 2005) and network deconvolution (Feizi et al., 2013). The subgraph centrality method used a sum of local moments of the adjacency matrix powers to derive a new centrality algorithm. Local moments use the diagonal of each matrix power. In our approach, we compute a modified modularity, as opposed to centrality, and we use the diagonal and off-diagonal of the power series up to a finite order. Also, our algorithm adapts the power series order to change the size and number of clusters. Network deconvolution reduces the noise in a network by assuming an infinite power series of the direct network, which allows one to invert the observed network to reverse the effect of transitive information flow from the theoretical indirect paths. Our approach adapts the order of the power series in subnetworks (modules) to reduce and magnify transitive effects within and between subnetworks in the observed network and thereby modify the size and number of clusters.

Including higher powers of the adjacency matrix increases the flow of information between a given pair of genes based on shared indirect connections, which, along with recursive splitting, allows RIP-M to develop modules with sizes near a user defined range. The goal of the RIP-M algorithm is to provide a user with the flexibility to specify a "Goldilocks" range that penalizes module sizes from being too large or too small for the biological questions being asked. We compare the accuracy of RIP-M with Newman modularity and WGCNA on simulated co-expression networks with homogeneous and heterogeneous cluster sizes. We compare the accuracy of methods for hard and soft thresholds, and we compare the methods based on functional enrichment of relevant pathways on real RNA-Seq data.

METHODS

In this section, we describe our new community detection algorithm, RIP-M, and our evaluation strategy based on simulated co-expression data and real data from an RNA-Seq experiment of an influenza vaccine study. Our aim is to detect communities or clusters of genes that are related by correlation across subjects in a gene expression study. To construct the network, we begin with a matrix of Pearson correlation coefficients $\left(-1 \leq \rho_{ij} \leq 1\right)$ between the expression levels of genes i and j. We then define a binary adjacency matrix $A_{ij} \in \{0, 1\}$ by the condition $\left|A_{ij}\right| \geq \tau$, where τ is a threshold.

```
input:
       Aadj          # Adjacency matrix
       minModuleSize # minimum allowed size of module
       maxModuleSize # maximum allowed size of module
       maxMergeOrder # maximum order for merging
output:
       list of modules, lists of node names, and number of iterations
algorithm:
while the stack is not empty and not max number of iterations
       consider a modules from the stack
       create a submatrix of Aadj from the popped module indices
       run modularity on the submatrix (Eqs 1-3)
       determine large and small modules based on minModuleSize and maxModuleSize
       return large modules onto the stack
       group small modules together into a new matrix for merging
  if there is more than one small module merge
       while mergerOrder < maxMergeOrder
              create sum of powers of the small modules matrix up to mergeOrder (Eq. 4)
               # indirect-path information
              run modularity to create a partition of the merged small modules matrix (Eq. 5)
              determine large, small and "Goldilocks" module sizes in the partition
       end while attempting merges and not found
       save Goldilocks modules and small unmergeable modules to module list
  end if
  increment iterations
end while stack not empty
return module list and number of iterations
```

FIGURE 1 | RIP-M algorithm pseudocode.

The value of τ is one factor that influences the size distribution of modules. We discuss the selection of τ and soft thresholds below.

Recursive Indirect-Paths Modularity (RIP-M)

Our RIP-M algorithm is a modification of Newman Modularity. RIP-M iteratively merges and splits modules to obtain a desired range of module sizes. If a module is larger than a user-defined threshold, the module is split based on iteration of modularity to this submodule. Merging is based on the application of Modularity to a power series of the adjacency matrix, which encodes indirect paths between nodes and has the effect of reassorting nodes into larger modules. We first review the relevant properties of Modularity.

Newman Modularity

Modules of an adjacency matrix, A_{ij}, are determined by recursive binary partitions of the nodes in the network. The base of the

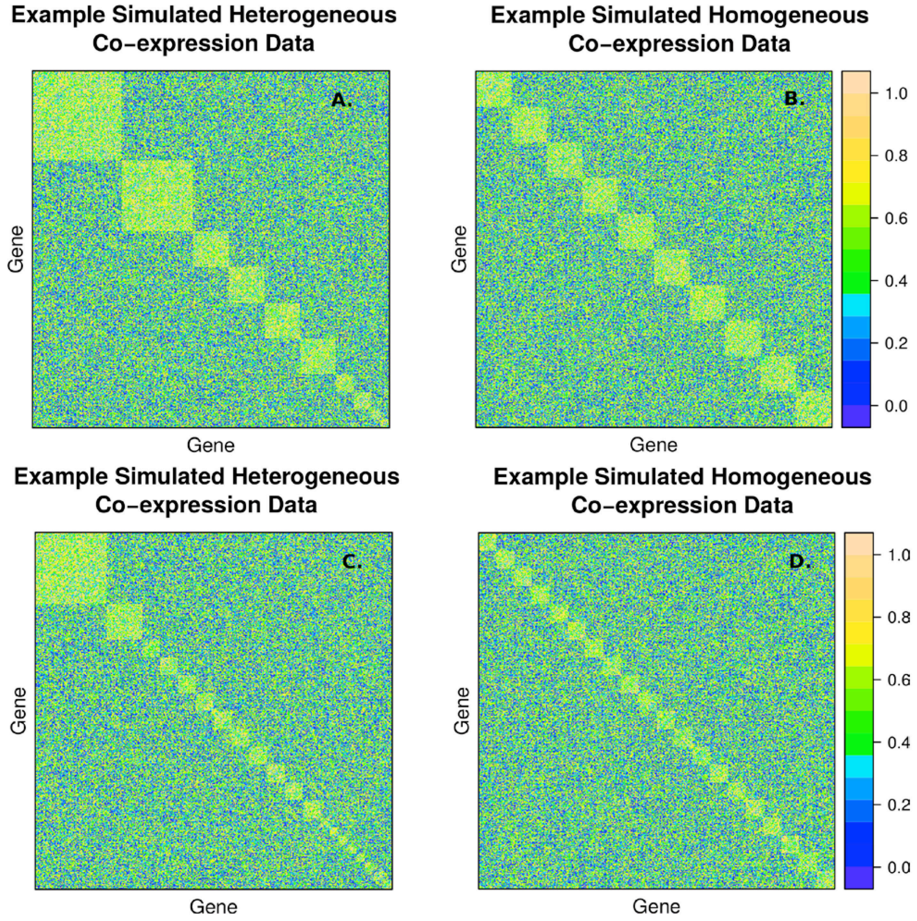

FIGURE 2 | Example of simulated correlation matrices with 10 clusters (A,B) and 20 clusters (C,D) of 400 nodes with heterogeneous (A,C) and homogeneous (B,D) cluster sizes. The block diagonal clusters are random uniform numbers between $r_{\text{signal,min}}$ and 1. Outside the blocks, the correlations are random uniform numbers between 0 and $r_{\text{noise,max}}$. The chosen signal and noise parameters are $r_{\text{signal,min}} = 0.2$ and $r_{\text{noise,max}} = 0.8$, respectively, to simulate a large amount of overlap between signal and background.

recursive algorithm is the binary partition of the network. The partition is encoded in a vector $s_i \in \{-1, 1\}$, where $s_i = 1$ if node i is in the first community and $s_i = -1$ if node i is in the second community. The vector s is obtained by maximizing the modularity Q:

$$Q = \frac{1}{4m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) s_i s_j, \tag{1}$$

where k_i, k_j are the degree of nodes i and j, and $m = \frac{1}{2} \sum_i k_i$. The quantity $k_i k_j / 2m$ is the expected number of edges between the nodes i, j for an Erdos-Renyi random network. We may rewrite Q in a quadratic matrix form as:

$$Q = \frac{1}{4m} s^T B\, s, \tag{2}$$

where the matrix $B_{ij} = (A_{ij} - \frac{k_i k_j}{2m})$. In order to maximize the modularity efficiently, one relaxes the requirement that the elements of s be dichotomous and one computes s as the dominant eigenvector of B. The values s are then dichotomized into 1 and -1 by taking the sign of the elements of the dominant eigenvector. This algorithm determines the initial binary partition of the network and is used recursively to form additional partitions. Within a subgroup of nodes g in a partition, the change in Q, $\triangle$Q, is computed by maximizing $B^{(g)}$:

$$\triangle Q = \frac{1}{4m} s^T B^{(g)} s, \;\; B^{(g)} = \sum_{i,j \in g} \left[B_{ij} - \delta_{ij} \sum_{k \in g} B_{ik} \right] \tag{3}$$

where $B^{(g)}$ is the subgroup modularity matrix properly accounting for the degree of nodes prior to binary splitting. Recursive partitioning stops when $\triangle Q$ is sufficiently small.

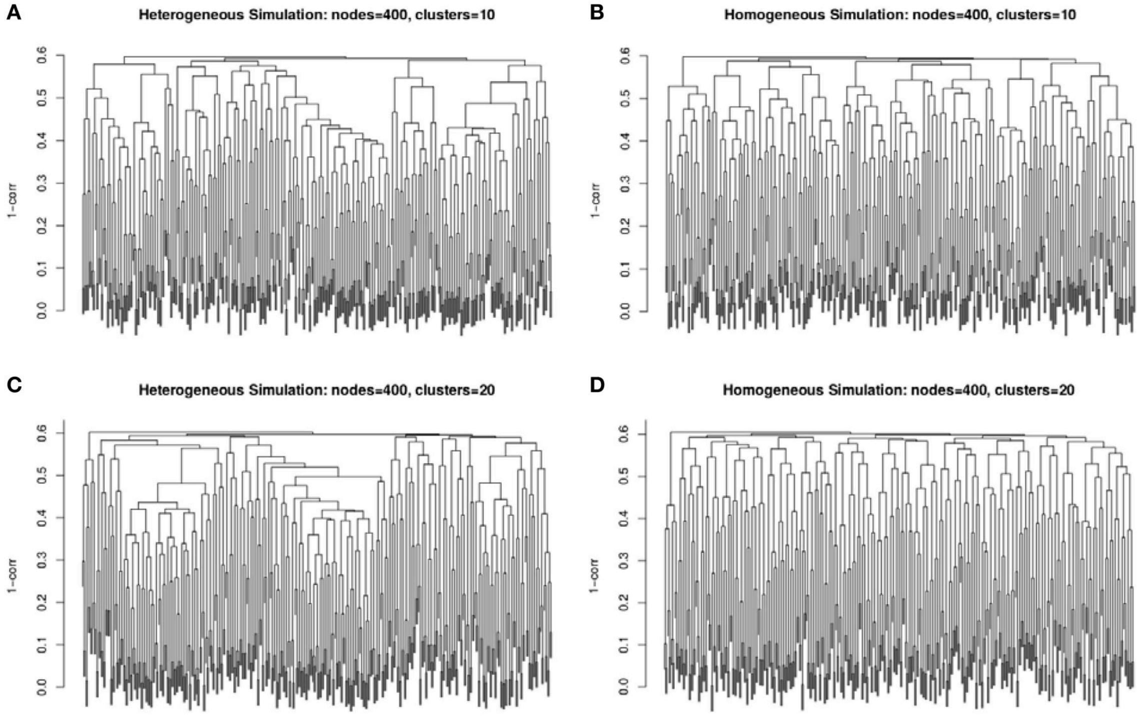

FIGURE 3 | Example dendrograms from the simulated correlation matrices with 10 clusters (A,B) and 20 clusters (C,D) of 400 nodes with heterogeneous (A,C) and homogeneous (B,D) cluster sizes. The simulated cluster blocks can be seen in **Figure 2**.

Generalized (Indirect-Paths) Modularity (IP-M)

We can generalize Equations (1–3) by writing Q as a function of a path matrix or power series of the adjacency matrix to order n:

$$P^{(n)} = A + A^2 + \cdots + A^n. \tag{4}$$

For a binary (non-weighted) adjacency matrix, an element of $P^{(n)}$ represents the number of paths with length smaller or equal to n between nodes i and j. This transforms the input network for modularity to a network weighted by indirect connections. For example, two nodes may not be directly connected, but multiple shared nodes may suggest that the two nodes should be in the same community. We find that this has a merging effect on communities. Then we can write the generalized-paths modularity as:

$$Q\left(P^{(n)}\right) = \frac{1}{4m^{(n)}} \sum\nolimits_{ij} \left(\left(P^{(n)}\right)_{ij} - \frac{k_i^{(n)} k_j^{(n)}}{2m^{(n)}} \right) s_i s_j. \tag{5}$$

We include the "n" superscripts to remind the reader that the degrees are computed for the path matrix P (Equation 4), not the original adjacency matrix A. For a given path order n, the algorithm for computing modularity is the same as Equations (1–3), but using P (Equation 4) instead of A. Next, we describe the RIP-M algorithm, which recursively splits large communities and varies the order n in $Q\left(P^{(n)}\right)$ (Equation 5) to automatically reassort clusters that are too small. The number of indirect paths (n) adapts within the recursion to attempt to achieve modules in the specified range.

RIP-M Algorithm

The recursion in RIP-M is implemented as a stack in the R language, described in pseudocode (**Figure 1**). The first input is the (weighted or un-weighted) adjacency matrix. In our simulation studies, we test hard and soft thresholding of the absolute value of the correlation matrix. A soft threshold is a transformation of the correlation matrix that raises values to a given power, element-wise. In applications, we show that the hard threshold, τ, and the soft-threshold power play important roles in determining the size of clusters. The target range of the module sizes is specified by the parameters minModuleSize and maxModuleSize. Modules below minModuleSize are placed together and fed into generalized modularity for merging with increasing orders of the indirect-paths power series $P^{(n)}$. The order increases and the merging iteration continues until maxMergeOrder is reached or the module-size target is reached. Modules above maxModuleSize are put in the stack to be split in subsequent iterations. Modules in the target range are added to the return module list.

Hierarchical WGCNA

Here, we review aspects of the WGCNA method that are relevant for comparison with modularity and the RIP-M method introduced in the current study. WGCNA is a hierarchical clustering method frequently used for clustering

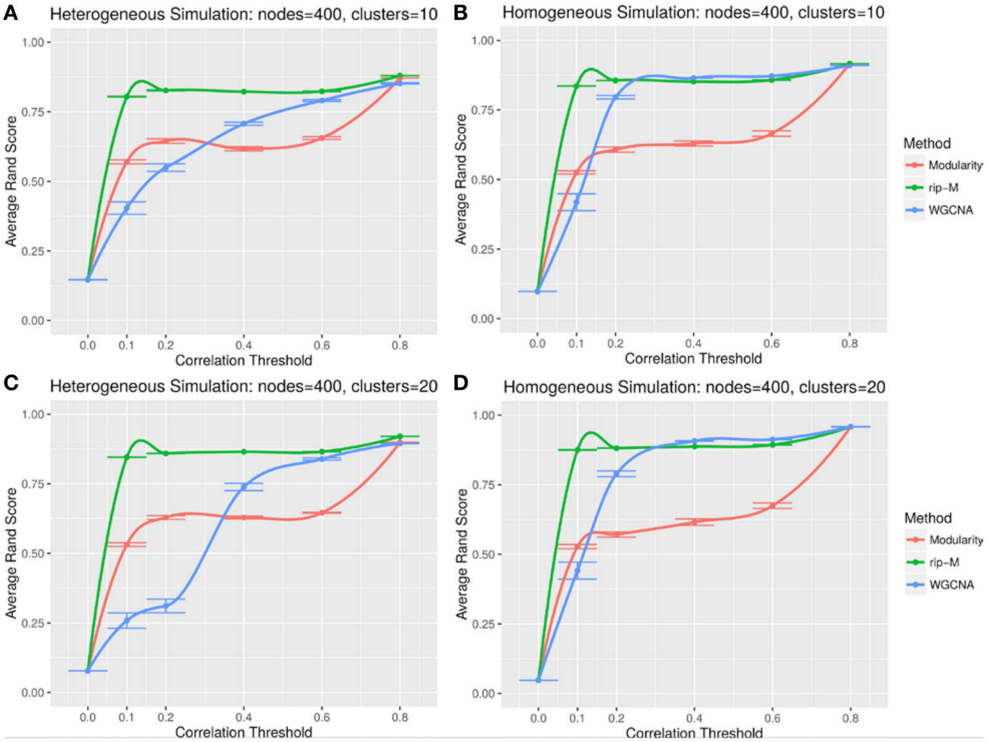

FIGURE 4 | Comparison of clustering accuracy using hard correlation threshold for random uniform simulated heterogeneous (A,C) and homogeneous (B,D) cluster sizes with 10 clusters (A,B) and 20 clusters (C,D). The signal and noise parameters are $r_{\text{signal,min}} = 0.2$ and $r_{\text{noise,max}} = 0.8$, respectively. 100 replicate simulations are created and clustering is carried out for each of the hard threshold cutoffs of the correlation matrix (horizontal axis). The accuracy of the clustering methods (average on the vertical axis) is based on the Rand index of the cluster identities compared with the true simulated cluster identities.

gene co-expression networks. It uses agglomerative hierarchical clustering with average linkage, but it uses a dynamic tree cutting to try to balance the size of clusters (Langfelder et al., 2013). This sometimes results in genes that are unassigned to any cluster. WGCNA also includes optional transformations of the correlation matrix, such as hard and soft thresholds and signed and unsigned networks. Parameters for these transformations may be guided by fitting a scale-free degree distribution. An important feature of WGCNA is the Topological Overlap Matrix (TOM) for assigning similarities between genes:

$$t_{ij} = \begin{cases} \dfrac{|N_1(i) \cap N_1(j)| + a_{ij}}{min\{|N_1(i)|, |N_1(j)|\} + 1 - a_{ij}} & if\ i \neq j \\ 1 & if\ i = j \end{cases},$$

where $|N_1(i)|$ is the number of neighbors of node i and a_{ij} is the corresponding element in the adjacency matrix. This first term in the numerator is the number of genes that genes i and j share as common connections. This term takes advantage of indirect evidence that two genes are similar. The denominator is a normalization factor. WGCNA also includes generalized TOM (GTOM) to perform hierarchical clustering with higher order indirect connections. The TOM matrix is transformed into a distance matrix for hierarchical clustering. We vary GTOM and thresholding parameters to identify the most uniform cluster sizes.

Methods for Cluster Comparison

Simulated Data

We simulate co-expression networks with known partitions or communities, and we compare clustering algorithms based on their accuracy in assigning nodes to their correct partitions. We create 100 replicate simulations for a given set of simulation parameters. We compare the accuracy of each method based on the Rand index (Rand, 1971). For all pairs of nodes, the Rand index determines whether the pair has the same cluster label predicted by the clustering algorithm and compares this concordance with the concordance in the true simulated clusters. We use two simulation strategies, both based on noisy block diagonal structures but with different noise distributions: random uniform and the beta distribution.

In the uniform random strategy, the background of the correlation matrix is composed of random uniform numbers between 0 and $r_{\text{noise,max}}$. This matrix is an example of a Wigner matrix (its entries are independent, uniformly bounded random variables, and we force the matrix to be symmetric; Bolla, 2011). On top of the background noise, we superimpose the clusters as block diagonal matrices composed of random uniform numbers between $r_{\text{signal,min}}$ and 1. Clusters become more difficult to resolve as the noise and signal limits overlap (i.e., as $r_{\text{block,min}}$ becomes larger and/or $r_{\text{block,max}}$ becomes smaller). In our simulations, we use highly overlapping

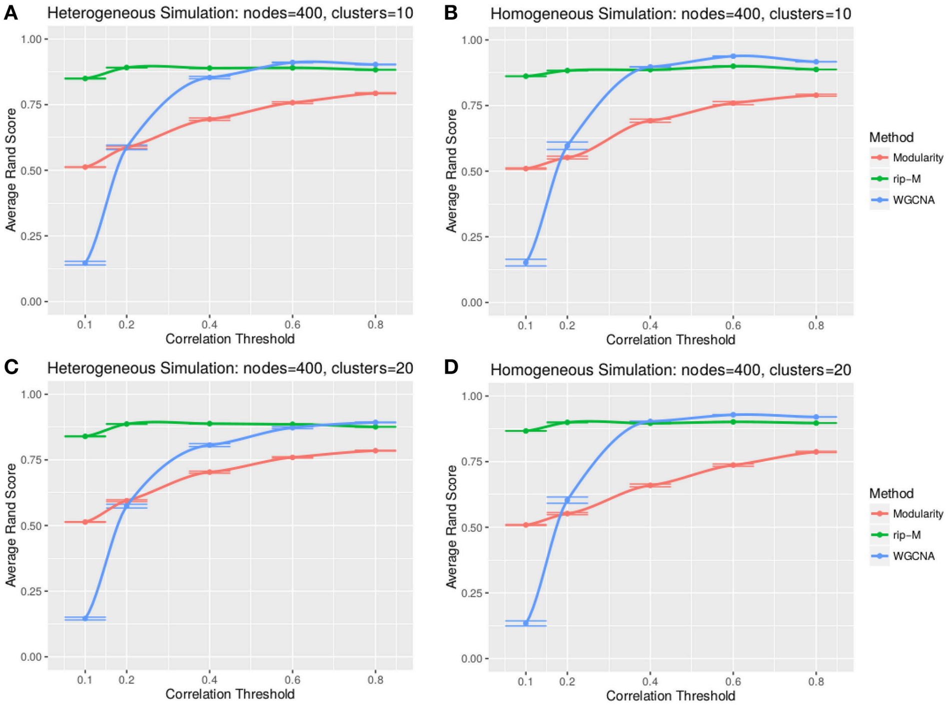

FIGURE 5 | Comparison of clustering accuracy using hard correlation threshold for beta-distributed noise simulated for heterogeneous (A,C) and homogeneous (B,D) cluster sizes with 10 clusters (A,B) and 20 clusters (C,D). The beta noise parameters are (3, 2) for the signal and (2, 3) for the background. One hundred replicate simulations are created and clustering is carried out for each of the hard threshold cutoffs of the correlation matrix (horizontal axis). The accuracy of the clustering methods (average on the vertical axis) is based on the Rand index of the cluster identities compared with the true simulated cluster identities.

parameters ($r_{\text{noise,max}} = 0.8$ and $r_{\text{signal,min}} = 0.2$) to make the community detection more challenging. We force the matrix to be symmetric, but it may not be positive definite.

The second simulation strategy uses a beta distribution with parameters (3, 2) for the signal and (2, 3) for the background noise. These parameters result in similarly challenging signal to noise as in the random uniform strategy, but with a less abrupt transition between clustered genes and non-clustered genes. In addition, we map the final correlation matrix to nearest positive definite matrix (Cheng and Higham, 1998). Since, we expect real co-expression data to display variation in the number and size of clusters, we simulate co-expression networks with heterogeneous (example in **Figures 2A,C**) and homogeneous (example in **Figures 2B,D**) cluster sizes. The challenge to cluster detection of the noise parameters can be seen in the flat hierarchy (small distances separating the simulated clusters) in **Figure 3**.

RNA-Seq Data

RNA-seq measures gene expression by sequencing, yielding the abundance of each transcript present in a sample. We applied each clustering method to RNA-Seq data from an influenza vaccine study of 105 individuals performed at the Mayo Clinic (Rochester, MN). All study participants provided written informed consent, and all study procedures were approved by Mayo Clinic's Institutional Review Board. Gene abundances were computed from transcriptomic sequencing using the MAP-RSeq bioinformatics pipeline tool (Kalari et al., 2014). The Illumin HiSeq 2000 (Illumina, San Diego, CA) sequencing reads were aligned to the human genome build 37.1 using TopHat (1.3.3) and Bowtie (0.12.7). Counting genes and mapping the reads to individual exons was carried by HTSeq (0.5.3p3) and BEDTools (2.7.1), respectively. Total number of counts per gene was obtained from the mRNA expression. Quality control was assessed pre- and post-normalization graphically with minus- vs.-average and box-and-whisker plots. The GC content and gene length adjustments were also evaluated graphically. Normalization of the gene counts was done with Conditional Quantile Normalization (CQN), which accounts for differences in library size and also adjusts for GC content and gene length (Hansen et al., 2012). These normalized values were used for subsequent analyses.

Gene expression levels were measured at Day 0 (before vaccination) and 3 days after vaccination. Clustering was performed on the difference of the log2 CQN normalized data between Day 3 and Day 0. Before clustering, we filtered to the top 4956 transcripts showing the largest variation across timepoints and largest read counts. Due to the perturbation of the gene network by a vaccine, we expect increased activity among immune system pathways. Thus, we used immunological pathways from the Reactome FI database to compare the enrichment from clustering methods of relevant biological pathways for the vaccine immune response experiment (Vastrik

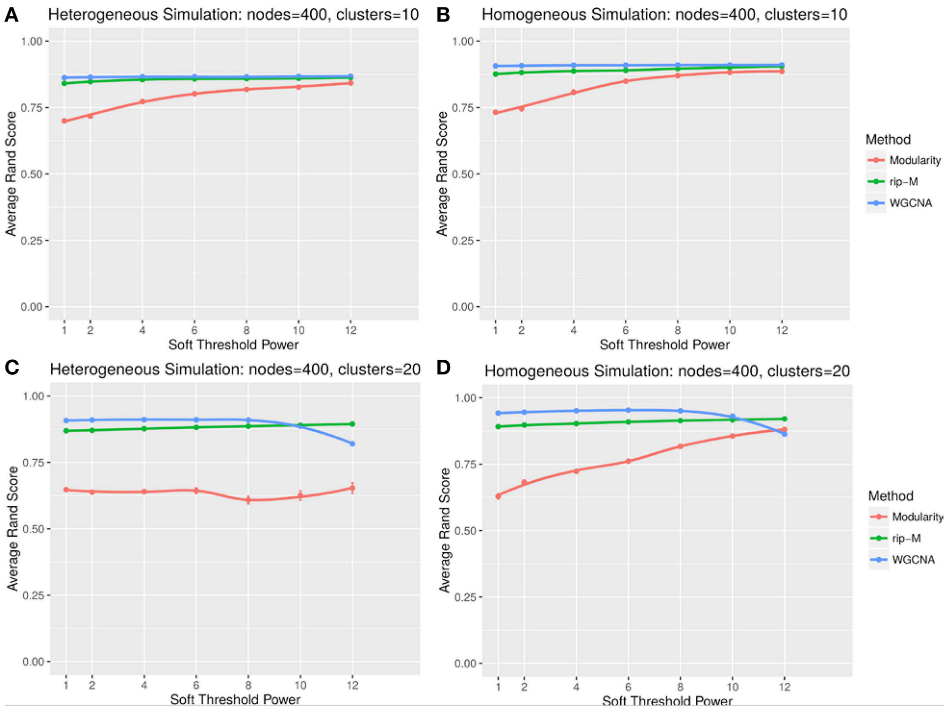

FIGURE 6 | Comparison of clustering accuracy using soft thresholds for random uniform simulated heterogeneous (A,C) and homogeneous (B,D) cluster sizes with 10 clusters (A,B) and 20 clusters (C,D). The signal and noise parameters are $r_{\text{signal,min}} = 0.2$ and $r_{\text{noise,max}} = 0.8$, respectively. One hundred replicate simulations are created and clustering is carried for varying soft threshold powers, which involves taking the element-wise powers of the correlation matrix. The accuracy of the clustering methods (on the vertical axis) is based on the Rand index of the cluster identities compared with the true simulated cluster identities.

et al., 2007). We calculated the fraction of overlap of each module with immunological gene sets and used the hypergeometric *p*-value to quantify significance of enrichment.

RESULTS

Simulation Results

We compared the average accuracy of Newman Modularity, RIP-M and WGCNA to find the correct number of clusters and the correct cluster identities of all 400 simulated genes based on the Rand index. We simulated 100 replicate datasets of co-expression data simulated with 10 and 20 heterogeneous-sized clusters (e.g., **Figures 2A,D**, respectively) and 10 and 20 homogeneous-sized clusters (e.g., **Figures 2B,C**, respectively). We specified a minimum cluster size of 10 and maximum of 50 for the RIP-M and WGCNA algorithm parameters. This range is more of a guideline than a strict rule because cluster sizes routinely go outside the range for both algorithms, and we find that the simulation results are not sensitive to this range. In addition to comparing homogeneous and heterogeneous cluster sizes, we compared methods based on hard and soft thresholding, which are common noise-reduction techniques for co-expression network clustering.

Hard Threshold Results

A hard threshold yields a binary adjacency matrix with elements equal to 1 when the correlation is above the threshold, and 0 otherwise. The threshold eliminates lower correlations that are more likely to be noise (due to chance). For hard thresholding, RIP-M has a consistently higher accuracy than WGCNA and Newman Modularity across all thresholds and simulation parameters (**Figures 4**, **5**). In the uniform random noise simulation (**Figure 4**), as the threshold increases, RIP-M reaches its maximum accuracy more quickly than the other methods. When the threshold is 0.8 in the uniform random uniform simulations (**Figure 4**), all of the methods are able to find the correct clusters with high accuracy because all noise connections are pruned at this threshold. This occurs at this particular threshold because we use $r_{\text{signal,min}} = 0.2$ and $r_{\text{noise,max}} = 0.8$ for the simulated signal and noise parameters, respectively. Many true within-cluster connections are also pruned at the 0.8 threshold, but this does not hinder the algorithms from detecting the community structure. In the simulations with beta-distributed noise (**Figure 5**), the hard threshold has a smoother effect on the cluster identification accuracy, but the accuracy trends are similar to the uniform random simulation strategies.

Soft Threshold Results

A soft threshold involves taking even powers of each element of the correlation matrix. Taking even powers gives negative correlations the same weight as positive correlations, but for simplicity we only simulate positive correlations. Higher powers force lower correlations closer to zero, reducing their role in the cluster analysis. For soft thresholds (**Figures 6**, **7**), WGCNA has slightly higher Rand index accuracy than RIP-M and both have higher accuracy than Newman Modularity. The accuracy of RIP-M and WGCNA is flat across soft threshold powers, whereas Newman Modularity tends to increase in accuracy with increasing threshold.

Influenza Vaccine RNA-Seq Results

To compare performance on real data, we applied RIP-M and WGCNA to an RNA-Seq co-expression network for an influenza vaccine experiment. Gene expression for subjects was measured before and 3 days after receiving influenza vaccine. We computed the log2 fold change between Days 3 and 0 for each subject's transcripts and then we removed transcripts showing low magnitude or variation across all subjects, which left 4956 transcripts. We calculated the Pearson correlation between all pairs of transcripts across all 105 subjects, and we applied the absolute value to the coefficients to treat activation and repression connections the same in the network.

To illustrate the types of extremes that can be observed in cluster sizes, we applied Newman Modularity to the RNA-Seq network for a range of hard threshold values (τ). For each value of τ, Modularity finds a giant component (**Figure 8**), which was the original motivation for the RIP-M modifications. We then swept the hard threshold parameter (τ) over a range of settings to identify clusters for RIP-M and WGCNA with limited small and large components. We specified a minimum cluster size of 100 and maximum of 300, with the goal of obtaining approximately 25 uniform clusters (average size of 200). This target was not quite achieved for RIP-M or WGCNA (**Figures 9**, **10**), but these hard thresholds yield the closest approximations. The WGCNA cluster sizes are more extreme than RIP-M, but of course the true size distribution is unknown. We provide rationale for choosing these clustering parameters in the Discussion.

In order to assess the biological relevance of the RIP-M and WGCNA influenza RNA-Seq clusters, we tested geneset enrichment of known immune system and other functional pathways that we predict may be related to vaccine response. We used the Reactome Functional Interaction (FI) database to identify enriched biological pathways in the genesets defined by each cluster (Vastrik et al., 2007). We calculated the fraction of overlap of each module with immunological gene sets and used the hypergeometric p-value to quantify significance of enrichment. We mapped each RIP-M module (**Figure 9**) and WGCNA cluster (**Figure 10**) onto the most enriched immunological gene set and sort them by cluster size. For example, "Influenza Life Cycle(R)" is enriched in one of the larger partitions of both methods, and "Cytokine-Cytokine Receptor" is enriched by both methods in more intermediate-sized clusters. To look beyond only the most immunological pathway in each cluster, we used a q-value cutoff of 0.1 for all RIP-M modules and WGCNA clusters. With this threshold, we identified 27

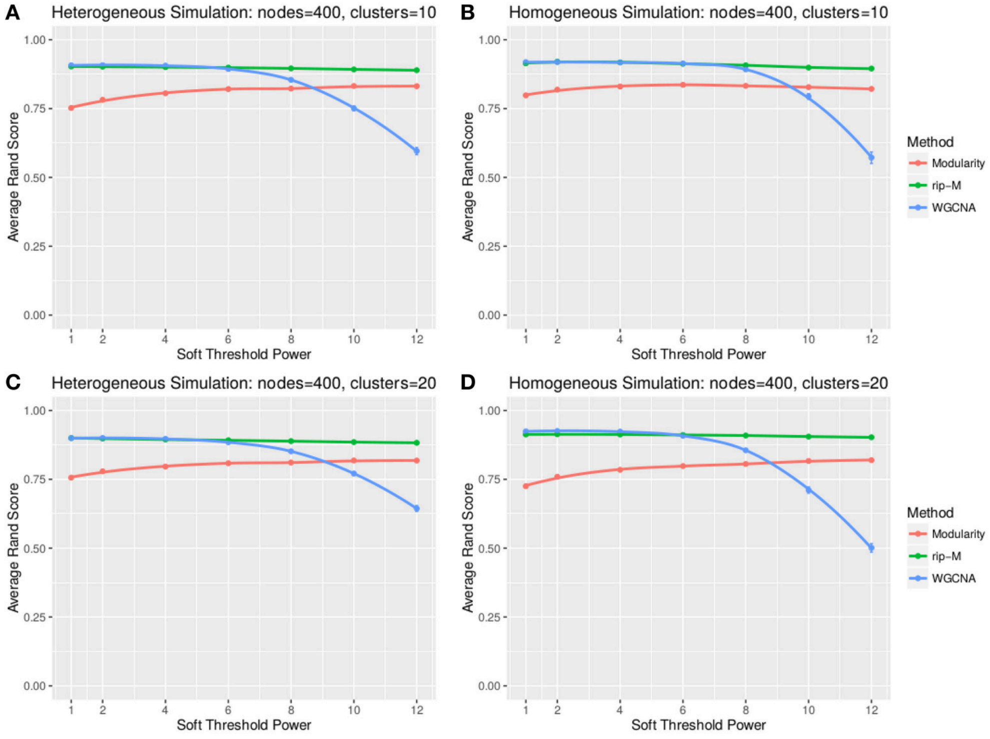

FIGURE 7 | Comparison of clustering accuracy using soft thresholds for beta-distributed noise simulated for heterogeneous (A,C) and homogeneous (B,D) cluster sizes with 10 clusters (A,B) and 20 clusters (C,D). The beta noise parameters are (3, 2) for the signal and (2, 3) for the background. One hundred replicate simulations are created and clustering is carried for varying soft threshold powers, which involves taking the element-wise powers of the correlation matrix. The accuracy of the clustering methods (on the vertical axis) is based on the Rand index of the cluster identities compared with the true simulated cluster identities.

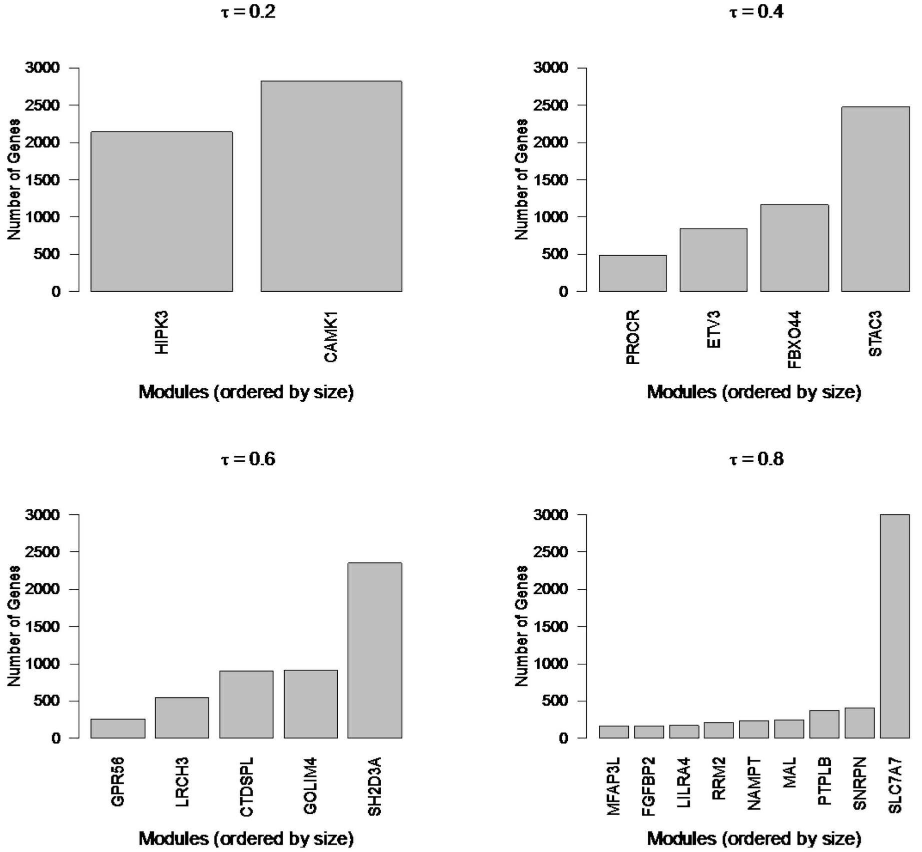

FIGURE 8 | Newman Modularity module sizes for flu vaccine RNA-Seq co-expression network for a range of hard thresholds (τ). The modules from each run (from 4965 transcripts) are labeled by their hub gene (gene with highest within module degree). Modularity finds large components for each value of τ.

immunological gene sets enriched by both algorithms. The RIP-M modules are enriched for an additional 17 immune gene sets that are distinct from WGCNA clusters, and WGCNA clusters are enriched for 10 immune gene sets that are distinct from RIP-M. This suggests that additional biological insights may be obtained through a combined clustering approach.

Hierarchical methods like WGCNA have a natural tree-based visualization of the relationships between genes. With RIP-M, it is more natural to visualize the relationships as a network graph (**Figure 11**). To balance the volume of information displayed on the graph with its interpretability, we collapsed the individual gene connections within a module onto one node and label the nodes based on their most enriched immune system pathway. Edges between collapsed nodes are weighted based on the number of edges between genes in each module. The aggregation of genes within modules/pathways and aggregation of the connectivity may increase sensitivity to detect regulation between pathways. For example, IGF-1 has been shown to promote cytotoxic activity in human natural killer cells (Ni et al., 2013), and our influenza vaccine pathway network (**Figure 11**) shows a strong connection between "IFG1 Pathway" and "Natural killer cell mediated cytotoxicity." Connections at the individual gene level within the two pathways may miss this relationship.

DISCUSSION

RIP-M is an extension of Newman Modularity for community detection that provides additional flexibility to identify substructure in modules that may be more relevant to the application domain. We compared and assessed this approach to find functional communities of genes in RNA-Seq gene co-expression networks; however, RIP-M is not limited to this application domain. We extended Newman Modularity to operate on indirect-path networks (IP-M) in order to increase information flow between nodes in the network and thereby merge modules with sizes that fall below a user specified size, below which biological pathway enrichment may be difficult to detect. Instead of using only the adjacency matrix, A, in the modularity B matrix, we use a power series of A (Equation 4), which represents higher order indirect connections between genes. Incorporating these indirect connections has the effect of merging smaller modules together into more reasonable sizes. The merging process, however, may leave large network components that contain unresolved substructure that have more specific biological function. Thus, we combined the IP-M merging capability with a splitting algorithm to decompose large modules and resolve finer functional pathways. We balanced the merging and splitting operations recursively in

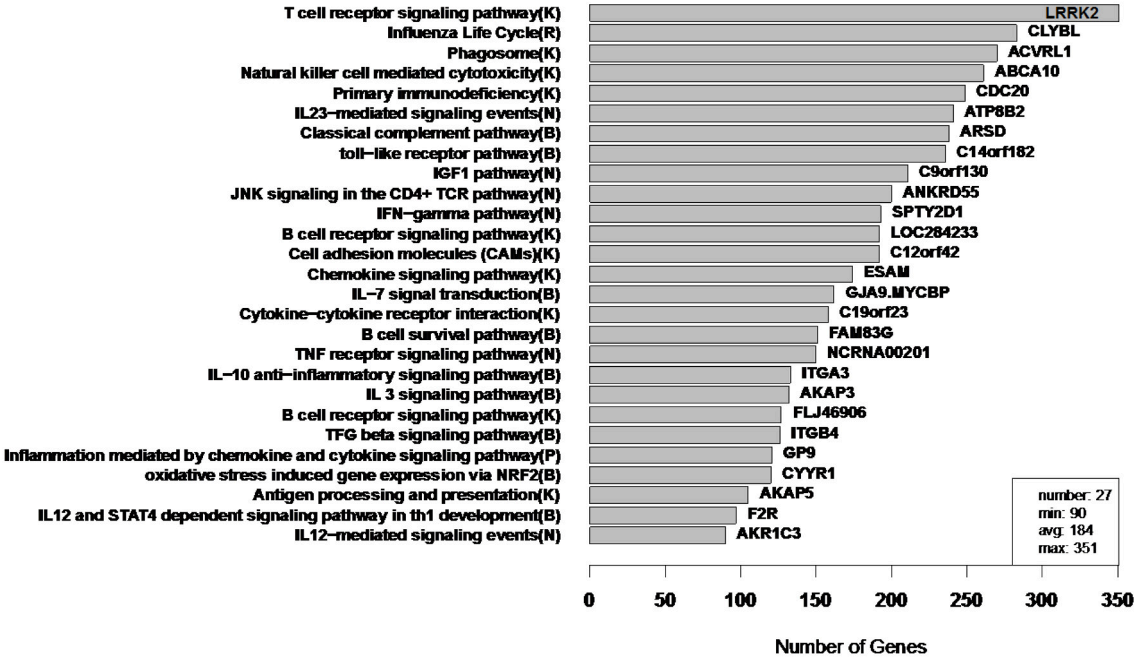

FIGURE 9 | RIP-M module sizes for flu vaccine RNA-Seq co-expression network. Each of the 27 modules (from 4965 transcripts) is labeled by its most enriched immune system pathway (on the left) and by its hub gene (gene with highest within-module degree) on the right. Hard threshold $\tau = 0.2$ of the absolute value of Pearson correlation was selected. The hard threshold choice was guided by the goal of having modules without extreme sizes, with a minimum size of 100 and maximum of 300. Parentheses in pathway names indicate source of annotation (Reactome, KEGG, NCI-PID, and BioCarta).

a modified version of modularity we call Recursive Indirect-paths Modularity. We compared this algorithm with standard modularity and hierarchical WGCNA on simulated correlation networks and real RNA-Seq co-expression from an influenza vaccine study.

It is common practice to use hard and soft thresholds to remove noise (reduce false-positive edges) from co-expression networks. We find that thresholds can also affect the number and size of clusters. A byproduct of this noise reduction is that some true direct connections may be lost (increase in false-negative edges). Using RIP-M, evidence for connections between nodes may be recovered from the adjacency matrix through evidence from indirect connections. Some pairs of genes that fall below the correlation threshold due to weak experimental or biological signal may show stronger indirect support for being linked based on evidence from the number of shared neighbors. We show that including higher-order connections in the generalized indirect-paths modularity has a merging effect on modules; some connections between nodes are strengthened and nodes from smaller clusters are rearranged into larger clusters.

The results on simulated and real co-expression data suggest that RIP-M has some advantages over Newman Modularity, on which RIP-M is based. For hard and soft thresholding of the simulated correlation matrices with homogeneous and heterogeneous cluster sizes, RIP-M and WGCNA have higher accuracy than Newman Modularity for finding the correct number of communities and correct partition identities of the nodes. In the simulated and real RNA-Seq co-expression data, Modularity has a tendency to create partitions with a small number of communities with some giant components, compared to WGCNA and RIP-M, which tend to resolve the co-expression networks into more intermediate sized clusters. WGCNA and RIP-M find comparable cluster numbers for simulated and real data. For the real data, WGCNA and RIP-M modules enrich for many of the same immunologically relevant pathways. Both methods enrich for some distinct immunological pathways, and there may be an advantage to taking the union of enriched pathways by both methods to carry forward for additional statistical analysis.

With the expectation that genes involving the immune system will be most stimulated by vaccination, we compared methods based on enrichment of communities of genes in immune functional pathways. RIP-M and WGCNA communities were enriched with the majority of the same immunological pathways, and each method found some unique pathways. Replication studies of these pathways will lead to a better understanding of the clustering mechanisms that lead to biologically useful communities and whether a hybrid approach may take advantage of unique mechanisms of each approach. We used enrichment of immunological pathways in the Reactome FI database; however, these pathways involve complex signaling and other interactions, and their genes are thus likely to have connections to multiple pathways, not simply connections within one pathway. This complexity adds to the noise of identifying partitions in this gene expression experiment, and in general one expects gene cluster membership to be fuzzy. Future work should involve testing for enrichment in other biological studies where the gold

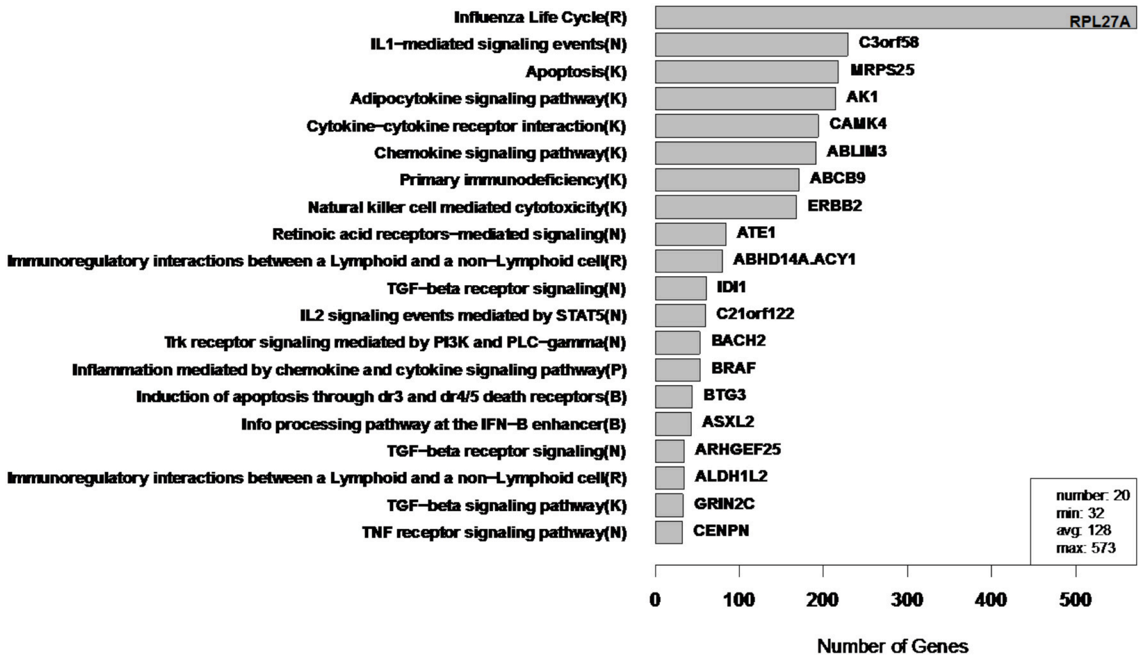

FIGURE 10 | WGCNA cluster sizes for flu vaccine RNA-Seq co-expression network. Each of the 20 clusters (from 4965 transcripts) is labeled by its most enriched immune system pathway (of the left) and by its hub gene (gene with highest within module degree) on the right. These clusters contain 2543 of the original 4965 because 2422 fell into the "unassigned" cluster. Algorithm parameters and hard threshold were swept with the goal of finding a module size distribution without extremely small or large modules. Parentheses in pathway names indicate source of annotation (Reactome, KEGG, NCI-PID, and BioCarta).

standard pathways may be more circumscribed or includes less heterogeneity, such as model organisms.

RIP-M and WGCNA allow the user to constrain cluster sizes. In this sense, RIP-M is similar to clustering algorithms that have the number of clusters as an input, like WGCNA, which may be suboptimal for data where cluster sizes fall outside the range. However, RIP-M and WGCNA are permitted to return cluster sizes outside of the user specified range (remaining true to actual relationships within the data). We simulated homogeneous and heterogeneous-sized clusters, where some cluster sizes fall outside the specified range, and RIP-M retains high accuracy. If no cluster size preference is provided (a wide range of cluster sizes is used by default), the RIP-M algorithm will still look for indirect evidence for connectivity and attempt to perform merging and splitting to find the best partition of the genes according to the relationships present within the supplied data. In the simulation studies, RIP-M is able to accurately find modules when the underlying data is composed of either heterogeneous or homogeneous cluster sizes.

Algorithm parameter selection is an important issue in clustering that is not unique to RIP-M, as WGCNA has a similar number of parameters that may be adjusted. In the real data, our choice of $\tau = 0.2$ was influenced by simulation results (**Figures 4**, **5**), but more by the fact that it resulted in clusters with the most uniform sizes. Increasing τ in the real data did not have a drastic effect on the number of clusters. We chose 25 as our target number of clusters because, of the roughly 5000 filtered genes, uniform clusters would contain approximately 200 genes each, which is a common number for geneset enrichment analysis. For example, the C7 immunologic signatures in MSigDB use geneset sizes of 200 (Godec et al., 2016). The C7 genesets correspond to top or bottom genes (FDR < 0.25 or maximum of 200 genes) for each comparison of conditions in immunologic gene expression experiments. In Chaussabel et al. the authors used $k = 30$ in k-means clustering of the coordinated gene co-expression across multiple immune phenotypes to create modules of genes suggestive of coordinated biological activity (Chaussabel et al., 2008). WGCNA recommends setting hard and soft thresholds based on scale free degree distribution, but, like our approach, this may lead to undesirable cluster assignments of genes. After deciding on a target of 200 genes, we then chose minimum and maximum cluster size parameters 100–300 in RIP-M because this range brackets the average module size target. Using different RIP-M parameters resulted in similar final cluster compositions. Our reported clusters with the choice of $\tau = 0.2$ had the most uniformly distributed sizes.

Another application of gene partitioning into modules is to use the collective information as a feature in a machine learning classifier of a phenotype or outcome. It has been shown for genetic association studies of multi-genic disorders that associations tend to be found in similar functional pathways (Franke et al., 2006). This suggests that constructing an attribute from the genes in a module or functional pathway may improve the ability to predict phenotypic class or variance while also reducing the multiple-hypothesis burden. Future study is needed to assess the potential advantage of combining clustering approaches with attribute construction approaches, such as single subject GSEA, GSVA (Hanzelmann et al.,

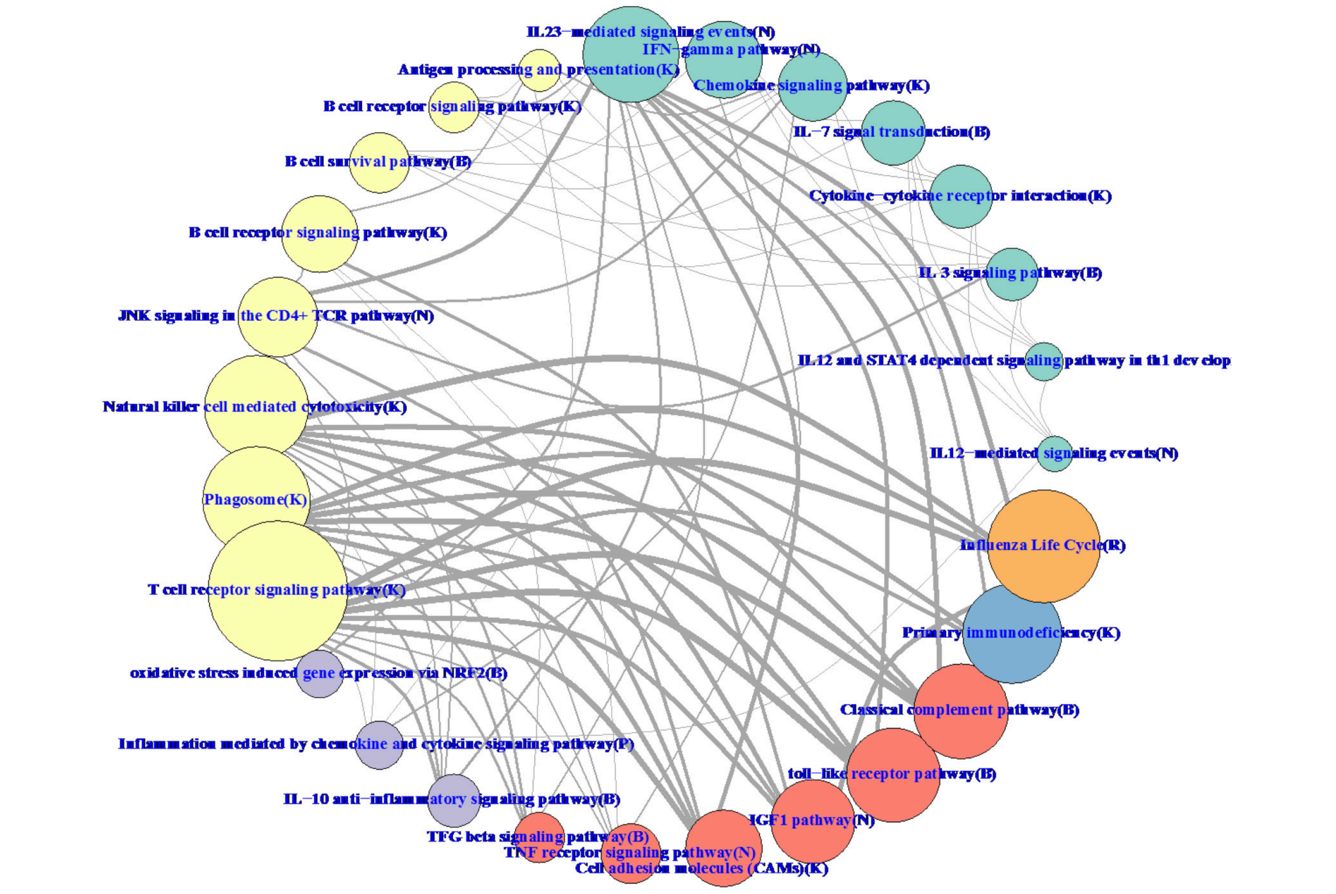
IL23−mediated signaling events(N)
IFN−gamma pathway(N)
Antigen processing and presentation(K)
Chemokine signaling pathway(K)
B cell receptor signaling pathway(K)
IL−7 signal transduction(B)
B cell survival pathway(B)
Cytokine−cytokine receptor interaction(K)
B cell receptor signaling pathway(K)
IL 3 signaling pathway(B)
JNK signaling in the CD4+ TCR pathway(N)
IL12 and STAT4 dependent signaling pathway in th1 development
Natural killer cell mediated cytotoxicity(K)
IL12−mediated signaling events(N)
Phagosome(K)
Influenza Life Cycle(R)
T cell receptor signaling pathway(K)
Primary immunodeficiency(K)
oxidative stress induced gene expression via NRF2(B)
Classical complement pathway(B)
Inflammation mediated by chemokine and cytokine signaling pathway(P)
toll−like receptor pathway(B)
IL−10 anti−inflammatory signaling pathway(B)
IGF1 pathway(N)
TFG beta signaling pathway(B)
TNF receptor signaling pathway(N)
Cell adhesion molecules (CAMs)(K)

FIGURE 11 | RIP-M pathway-collapsed network for RNA-Seq influenza vaccine data with 4965 transcripts showing highest variation. For each of the 27 RIP-M modules, genes and their within-module edges are collapsed into one node and labeled by its top immune system pathway. The radius of each node is proportional to the number of genes in the module. The weight of each edge is the total number of edges between genes of the two modules. The top five connections are shown for each module for clarity. The modules were obtained by hard threshold $\tau = 0.2$ of the absolute value of Pearson correlation. Colors were chosen based on pathway-type: yellow T-cell, B-cell, NK-Cell; green cytokine; purple inflammatory; blue immunodeficiency; orange influenza life cycle; red TLR and complement.

2013) or within-cluster hubs, for use as attributes in machine learning algorithms to improve their ability to predict disease susceptibility, immune response or other relevant outcomes. One may also combine RIP-M or other clustering methods with outcome information through differential co-expression analysis (Lareau et al., 2015).

We have extended the Newman Modularity method for community detection by including indirect-path information when computing modules and by creating the RIP-M merge-split algorithm that provides additional flexibility to identify network substructure. We tested RIP-M on synthetic and real RNA-Seq gene co-expression networks, identifying interpretable and biologically descriptive modules. For example, our results show a clear concentration of gene expression activity in several complementary areas of biological function important for both innate and adaptive immunity. The day 3 expression data delineates connections between several major pathways including JNK signaling in CD4 T cells and T cell receptor signaling, or B cell receptor signaling and B cell survival. It also illuminates the complex interplay between different immune processes such as IL-23 signaling (driving Th17 responses); IFNg, TNFa, and IL-12 (driving Th1 responses); anti-inflammatory signaling through TGF-b, IL-10, and the combination of IL-3 and IL-7 signaling known to drive hematopoietic production of lymphoid progenitor cells in both mice and humans (Grabstein et al., 1993; Puel et al., 1998; Crooks et al., 2000). RIP-M shows higher accuracy than Newman Modularity, and RIP-M's performance is comparable to WGCNA. Like WGCNA, RIP-M allows calibration of cluster sizes, but without removing genes. We believe RIP-M will be a valuable addition to the tools of computational biology for finding context-specific sets of genes for further pathway and integrative analysis.

AVAILABILITY

The raw data used in this study is available on ImmPort at https://immport.niaid.nih.gov/ under study number SDY67. Code for RIP-M and the normalized data used in this study are available at our website: http://insilico.utulsa.edu/index.php/ripm/.

AUTHOR CONTRIBUTIONS

BR, BW and BM conceived of the method and performed analyses. BR and BM wrote the initial draft of the paper. MZ, DG, GP, RK, and AO contributed to the design, interpretation, and revisions for content. All authors approved the final version and agree to be accountable to the accuracy and integrity of the work.

DISCLAIMER

Dr. GP is the chair of a Safety Evaluation Committee for novel investigational vaccine trials being conducted by Merck Research Laboratories. Dr. GP offers consultative advice on vaccine development to Merck & Co. Inc., CSL Biotherapies, Avianax, Dynavax, Novartis Vaccines and Therapeutics, Emergent Biosolutions, Adjuvance, Microdermis, Seqirus, NewLink, Protein Sciences, GSK Vaccines, and Sanofi Pasteur. Dr. GP holds two patents related to vaccinia and measles peptide research. These activities have been reviewed by the Mayo Clinic Conflict of Interest Review Board and are conducted in compliance with Mayo Clinic Conflict of Interest policies. This research has been reviewed by the Mayo Clinic Conflict of Interest Review Board and was conducted in compliance with Mayo Clinic Conflict of Interest policies.

ACKNOWLEDGMENTS

Research reported in this publication was supported by the National Institute of Allergy And Infectious Diseases of the National Institutes of Health under award number U01AI089859. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

REFERENCES

Ballouz, S., Verleyen, W., and Gillis, J. (2015). Guidance for RNA-seq co-expression network construction and analysis: safety in numbers. *Bioinformatics* 31, 2123–2130. doi: 10.1093/bioinformatics/btv118

Bolla, M. (2011). Penalized versions of the Newman-Girvan modularity and their relation to normalized cuts and k-means clustering. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 84(1 Pt 2):016108. doi: 10.1103/PhysRevE.84.016108

Chaussabel, D., Quinn, C., Shen, J., Patel, P., Glaser, C., Baldwin, N., et al. (2008). A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* 29, 150–164. doi: 10.1016/j.immuni.2008.05.012

Cheng, S. H., and Higham, N. (1998). A modified cholesky algorithm based on a symmetric indefinite factorization. *SIAM J. Matrix Analysis Appl.* 19, 1097–1110. doi: 10.1137/S0895479896302898

Crooks, G. M., Hao, Q. L., Petersen, D., Barsky, L. W., and Bockstoce, D. (2000). IL-3 increases production of B lymphoid progenitors from human CD34+CD38-cells. *J. Immunol.* 165, 2382–2389. doi: 10.4049/jimmunol.165.5.2382

Estrada, E., and Rodriguez-Velazquez, J. A. (2005). Subgraph centrality in complex networks. *Phys. Rev. E Stat. Nonlin. Soft. Matter Phys.* 71(5 Pt 2):056103. doi: 10.1103/PhysRevE.71.056103

Feizi, S., Marbach, D., Medard, M., and Kellis, M. (2013). Network deconvolution as a general method to distinguish direct dependencies in networks. *Nat. Biotechnol.* 31, 726–733. doi: 10.1038/nbt.2635

Franke, L., van Bakel, H., Fokkens, L., de Jong, E. D., Egmont-Petersen, M., and Wijmenga, C. (2006). Reconstruction of a functional human gene network, with an application for prioritizing positional candidate genes. *Am. J. Hum. Genet.* 78, 1011–1025. doi: 10.1086/504300

Godec, J., Tan, Y., Liberzon, A., Tamayo, P. S. B., Butte, A. J., Mesirov, J. P., et al. (2016). Compendium of immune signatures identifies conserved and species-specific biology in response to inflammation. *Immunity* 44, 194–206. doi: 10.1016/j.immuni.2015.12.006

Grabstein, K. H., Waldschmidt, T. J., Finkelman, F. D., Hess, B. W., Alpert, A. R., Boiani, N. E., et al. (1993). Inhibition of murine B and T lymphopoiesis *in vivo* by an anti-interleukin 7 monoclonal antibody. *J. Exp. Med.* 178, 257–264. doi: 10.1084/jem.178.1.257

Guo, J., Tian, D., and McKinney, B. A., Hartman J. L. IV. (2010). Recursive expectation-maximization clustering: a method for identifying buffering mechanisms composed of phenomic modules. *Chaos* 20:026103. doi: 10.1103/PhysRevE.84.016108

Hansen, K. D., Irizarry, R. A., and Wu, Z. (2012). Removing technical variability in RNA-seq data using conditional quantile normalization. *Biostatistics* 13, 204–216. doi: 10.1093/biostatistics/kxr054

Hanzelmann, S., Castelo, R., and Guinney, J. (2013). GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics* 14:7. doi: 10.1186/1471-2105-14-7

Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999). From molecular to modular cell biology. *Nature* 402(6761 Suppl.), C47–C52. doi: 10.1038/35011540

Hitzemann, R., Bottomly, D., Darakjian, P., Walter, N., Iancu, O., Searles, R., et al. (2013). Genes, behavior and next-generation RNA sequencing. *Genes Brain Behav.* 12, 1–12. doi: 10.1111/gbb.12007

Kalari, K. R., Nair, A. A., Bhavsar, J. D., O'Brien, D. R., Davila, J. I., Bockol, M. A., et al. (2014). MAP-RSeq: mayo analysis pipeline for RNA sequencing. *BMC Bioinformatics* 15:224. doi: 10.1186/1471-2105-15-224

Langfelder, P., Mischel, P. S., and Horvath, S. (2013). When is hub gene selection better than standard meta-analysis? *PLoS ONE* 8:e61505. doi: 10.1371/journal.pone.0061505

Lareau, C. A., White, B. C., Oberg, A. L., and McKinney, B. A. (2015). Differential co-expression network centrality and machine learning feature selection for identifying susceptibility hubs in networks with scale-free structure. *Bio Data Min.* 8:5. doi: 10.1186/s13040-015-0040-x

Mitra, K., Carvunis, A. R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nat. Rev. Genet.* 14, 719–732. doi: 10.1038/nrg3552

Newman, M. E. (2006). Modularity and community structure in networks. *Proc. Natl. Acad. Sci. U.S.A.* 103, 8577–8582. doi: 10.1073/pnas.0601602103

Ni, F., Sun, R., Fu, B., Wang, F., Guo, C., Tian, Z., et al. (2013). IGF-1 promotes the development and cytotoxic activity of human NK cells. *Nat. Commun.* 4:1479. doi: 10.1038/ncomms2484

Ozsolak, F., and Milos, P. M. (2011). RNA sequencing: advances, challenges and opportunities. *Nat. Rev. Genet.* 12, 87–98. doi: 10.1038/nrg2934

Parter, M., Kashtan, N., and Alon, U. (2007). Environmental variability and modularity of bacterial metabolic networks. *BMC Evol. Biol.* 7:169. doi: 10.1186/1471-2148-7-169

Puel, A., Ziegler, S. F., Buckley, R. H., and Leonard, W. J. (1998). Defective IL7R expression in T(−) B(+)NK(+) severe combined immunodeficiency. *Nat. Genet.* 20, 394–397. doi: 10.1038/3877

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *J. Am. Statist. Associat.* 66, 846–850. doi: 10.1080/01621459.1971.10482356

Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.* 31, 64–68. doi: 10.1038/ng881

Shi, J., and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 888–905. doi: 10.1109/34.868688

Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., Croft, D., de Bono, B., et al. (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.* 8:R39. doi: 10.1186/gb-2007-8-3-r39

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2016 Rahmani, Zimmermann, Grill, Kennedy, Oberg, White, Poland and McKinney. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.