Assembling cell context-specific gene sets: a case in cardiomyopathy

Mingming Liu [1], Vanessa King [2], Wei Keat Lim [2,*]

[1]Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA

[2]Siemens Corporation, Corporate Technology, Princeton, NJ, USA

Summary

An increasing amount of evidence suggests that canonical pathways and standard molecular signature databases are incomplete and inadequate to model the complex behavior of cell physiology and pathology. Yet, many Gene Set Analysis (GSA) studies still rely on these databases to identify disease biomarkers and molecular mechanisms within a specific cell context. While tremendous effort has been invested in developing GSA tools, there is limited number of studies focusing on *de novo* assembly of context-specific gene sets as opposed to simply applying GSA using the standard gene set database.

In this paper, we propose a pipeline to derive the entire collection of Cell context-Specific Gene Sets (CSGS) from a molecular interaction network, based on the hypothesis that molecular events linked to a specific phenotypic response should cluster within a subnet of interacting genes. Gene sets are assigned using both physical properties of the network and functional annotations of the neighboring nodes. The identified gene sets could provide a precise starting point such that the downstream GSA will cover all functional pathways in this particular cell context and, at the same time, avoid the noise and excessive multiple-hypothesis testing due to inclusion of irrelevant gene sets from the standard database. We applied the pipeline in the context of cardiomyopathy and demonstrated its superiority over MSigDB gene set collection in terms of: (i) reproducibility and robustness in GSA, (ii) effectiveness in uncovering molecular mechanisms associated with cardiomyopathy, and (iii) the performance in distinguishing diseased vs. normal states.

1 Introduction

The conflicting results repeatedly produced by many single gene-based approaches have been boosting the popularity of Gene Set Analysis (GSA) in omics study. Since the proposal of gene set enrichment analysis [1], various GSA methods have been developed using different statistical models. These methods generally fall into two categories, i.e. ‘self-constrained’ and ‘competitive’, based on the underlying null hypothesis and statistical methods [2]. The ‘self-constrained’ methods associate phenotype to a gene set through the members within the gene set only [3, 4], while the ‘competitive’ methods focus on contrasting a gene set with its counterpart in relation to the phenotype [1, 5]. The assumption in GSA is that unstable or weak changes in individual genes can be detected more robustly when studied as a group with sequential, expressional, or functional links. In the past decade, tremendous effort has been invested in developing GSA tools with the focus on obtaining biological interpretation from

[*] To whom correspondence should be addressed. Email: weikeat@gmail.com. Present address: Regeneron Pharmaceuticals, Inc., Tarrytown, NY, USA.

doi:10.2390/biecoll-jib-2013-234

differentially expressed genes [5-7]. Instead of running functional enrichment on the genes passing a certain cutoff, the GSA approaches establish a set-based association score and their significance level based on collective contributions from all members in the predefined gene sets. Similar strategies have been applied to genome wide association studies by integrating significant genotype-phenotype associations with gene set databases. The approaches, also known as Pathway-Wide Association Study (PWAS), examine whether members in the same functional pathway are jointly associated with a specific trait or disease [8-12]. While they offer various statistical methods to maximize analytical power without the necessity to increase sample size, starting GSA with a collection of high quality and appropriate gene sets should be equally important [13, 14].

Significant effort has been put into creating a catalog of gene sets through literature mining by aggregating genes with certain associations. One of the most popular collections is the Molecular Signatures Database (MSigDB) that has accumulated >6,000 gene sets in 5 major categories: (i) positional, (ii) curated, (iii) motif, (iv) computational, and (v) Gene Ontology (GO). However, most of these categories do not provide information of the cell context, except one subcategory named “gene expression signatures of genetic and chemical perturbations”, where the annotation is partially available. Among the gene sets annotated with their cell context, approximately half of them were extracted from experiments performed in either breast or liver tissues (see Figure 1), showing a strong bias in the data sources and the literature publications in general. Other gene set databases, including ConceptGen [15], WhichGenes [16], GeneSigDB [17], and GATHER [18], also assemble gene sets from various generic resources and offer additional tools for searching and customizing gene sets. While these databases may have facilitated various GSA studies, they are still subjected to certain limitations, such as the lack of cellular context information and the bias in literature studies toward certain diseases of high prevalence.

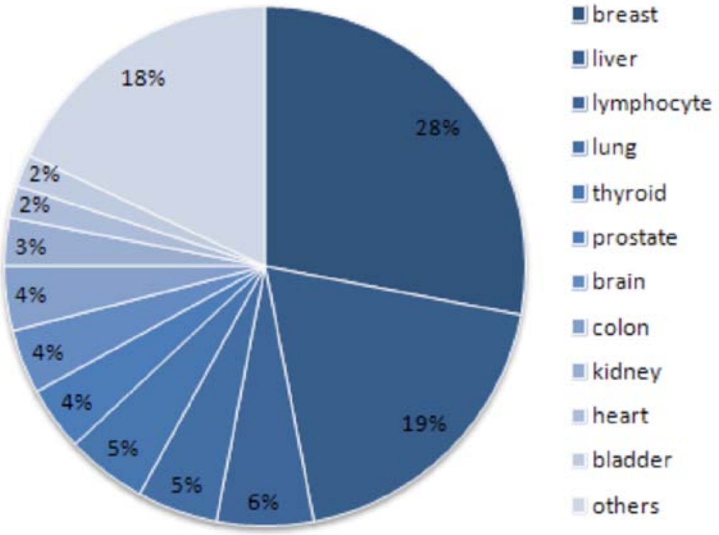


Figure 1: MSigDB cell context information.

Toward this end, we developed a procedure to create unbiased gene sets that are specific to a cellular context (see Figure 2). The approach makes use of gene expression profile (GEP), functional annotations, well-established reverse engineering algorithms, and molecular interactions databases. We integrated Protein-Protein Interaction (PPI) and Protein-DNA Interaction (PDI) network to systematically identify the gene sets, i.e. subnets that are strongly connected together both physically and functionally. The global network was first partitioned into the smallest units of physically connected subnets. Biological functions and pathway annotation were taken into consideration to re-evaluate if the neighboring subnets should be merged to form a combined gene set with certain functional ties. The most important distinction, compared to the standard gene set database, is that it takes into account the interactions among the member genes and it is orders of magnitude more complex than the canonical pathways that are mostly linear. Moreover, these gene sets are cell context specific and much higher in coverage than those that rely entirely on evidence collected from literature.

We demonstrate the gene sets assembly pipeline in the context of cardiomyopathy. While most cardiomyopathies are ischemic, i.e. related to coronary artery disease, the causes for many others remain unknown (idiopathic). Recent familial cardiomyopathy genetic studies have identified mutations in over 30 genes suitable for molecular genetics diagnostics, but the cause and effect relationships of the biomarkers are yet to be established [12]. We show that our newly generated cardiomyopathy gene sets are functionally relevant and the downstream analysis results using these gene sets are robust and stable, signifying the advantages of this integrated framework over GSA using the conventional gene set database for cardiomyopathy study.

2 Methods

2.1 Gene Expression Profile

A large cohort of cardiomyopathy GEPs was downloaded from Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/). The dataset (GSE5406) consists of 210 samples of human ischemic cardiomyopathy ($n=108$), idiopathic cardiomyopathy ($n=86$), and non-failing controls ($n=16$) [19]. Samples were extracted from the snap-frozen left ventricular myocardium at the time of cardiac transplantation from patients with advanced idiopathic or ischemic cardiomyopathy, or at the time of harvest from unused donor hearts that serve as non-failing controls. Samples were profiled using the Affymetrix HG-U133A array and data were normalized using the Robust Multi-array Average (RMA) method [20], probes matching to the same genes were filtered by selecting the most expressed one.

2.2 PPI/PDI Network Construction

To reconstruct a general human protein-protein interaction (PPI) network, we integrated data from four major PPI databases, i.e. BioGRID [21], IntAct [22], MINT [23] and REACTOME [24]. We only included interactions supported by at least one piece of direct experimental evidence demonstrating physical association between two human proteins. Interactions were further selected for context specificity by removing genes that do not express, and do not co-express for interacting pairs, using the cardiomyopathy gene expression profiles.

ARACNe (Algorithm for the Reconstruction of Accurate Cellular Networks) was used to assemble the cardiomyopathy-specific PDI network [25]. The algorithm employs information theory for the inference of TF-target interactions from large set of gene expression profiles, and further refined to determine directed interactions. The list of human TFs was first selected using "transcription factor activity" annotation in Gene Ontology, and then removed for nonspecific TFs (e.g. polymerases and TATA-box-binding proteins). To implement ARACNe, candidate interactions between a TF (x) and its potential target (y) were identified by computing pairwise mutual information, $I[x;y]$, and by applying a threshold based on the null-hypothesis of statistical independence ($p=0.05$, Bonferroni-corrected for the number of tested pairs). Indirect interactions were pruned using the data processing inequality (DPI), which states that if two genes interact only through a third gene, the mutual information between the two genes should be the least among the three pairwise measures. Thus, for each TF-target pair (x,y) we considered a path through all other TF (Z) and remove any interaction such that $I[x;y] < \min(I[x;z], I[y;z])$. These steps were repeated 100 times, each with a bootstrap dataset generated by randomly selecting samples with replacement from the original dataset. A consensus network was then constructed by retaining edges supported across a significant number of the bootstrap networks.

2.3 Network Partitioning

Isoperimetric partitioning algorithm [26, 27] was applied to identify physically strongly connected subnets in a network where nodes (V) are proteins and edges (E) are interactions. The algorithm formalizes the network partitioning problem into a linear system problem by introducing the isoperimetric constant *h*,

$$h = \inf_s \frac{|\partial S|}{Vol_s},$$

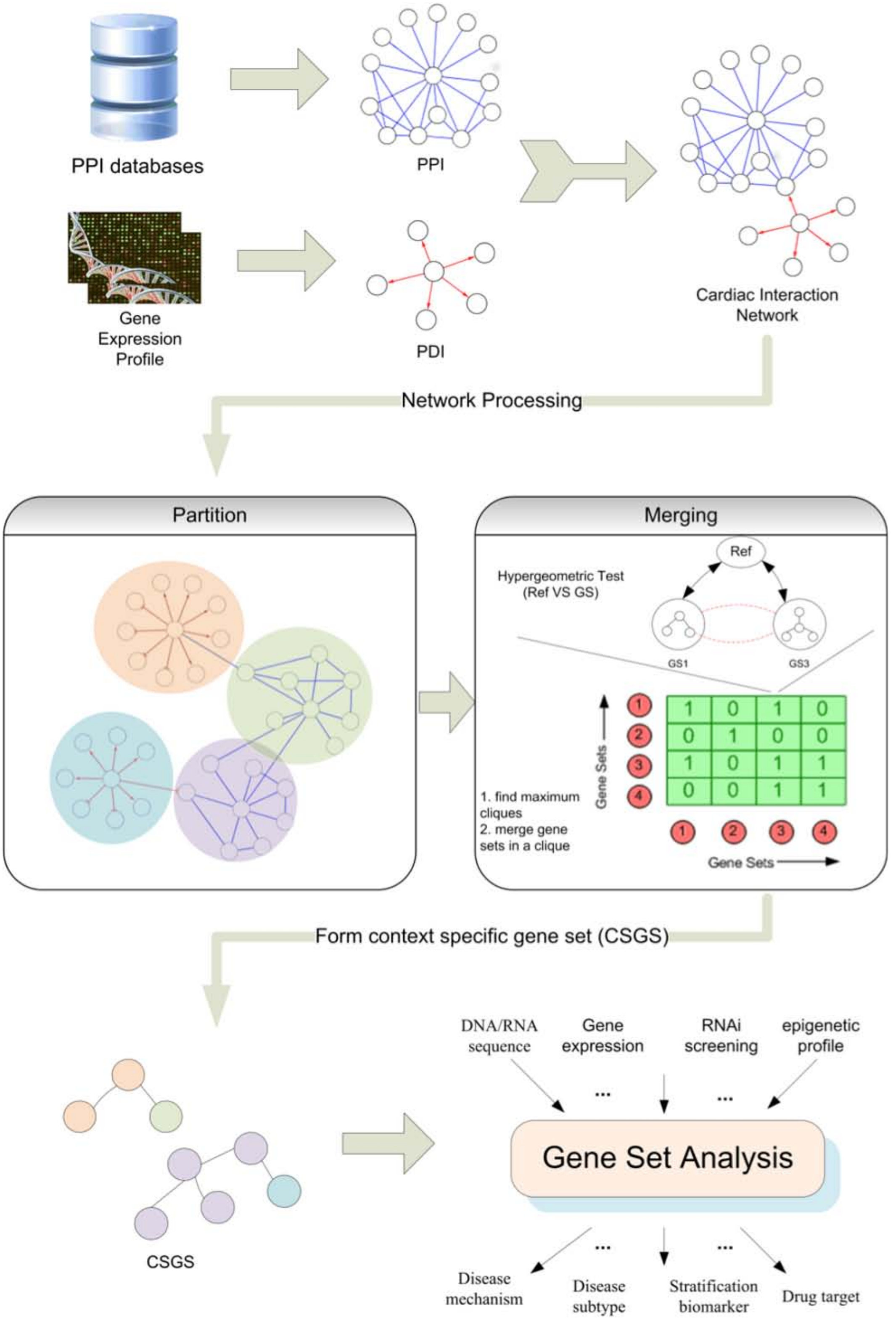
where S is the region in the manifold, Vol_S refers to the volume of region S, $|\partial S|$ is the area of the boundary of region S, and h is the infimum of the ratio over all possible S. In a graph, $|\partial S|$ and Vol_S can be defined with respect to the Laplacian matrix used in the isoperimetric algorithm. The Laplacian matrix, L, of a graph is defined as follows:

$$L = \begin{cases} d_i & \text{if } i = j, \\ -w_{ij} & \text{if } \exists \mathrm{e}_{\mathrm{ij}}, \\ 0 & \text{else} \end{cases},$$

where d_i is the weighted degree (sum of all edges) of node v_i, and w_{ij} is the weight of edge e_{ij}. The weights can be set according to the strength or evidence of the interactions, but we use unweighted graph by setting all $w_{ij} = 1$ in this study. Let $\mathrm{x_i}$ be an indicator vector that $x_i = 1$ if $v_i \in S$ ($S \subset V$), otherwise $x_i = 0$. As a result, $|\partial S| = \mathrm{x^T L x}$ and $Vol_s = x^T d$, thus, the isoperimetric constant *h* of a graph *G* can be rewritten in terms of the indicator vector as,

$$\min_x \frac{x^T L x}{x^T d}.$$

This optimization problem can be solved by a linear system $L_r x = d$ where L_r is a reduced Laplacian matrix by removing the corresponding row and column of the node with a maximum weighted degree. Through this step, the original singular matrix L is converted into a nonsingular matrix, which makes the computation more efficient and easier to be implemented in a parallelized environment [28]. By solving this linear system, each node will be assigned a real value. We chose a threshold β that minimize the isoperimetric ratio, and further generated an indicator vector y, where $y_i = 1$ if $x_i > \beta$, and $y_i = 0$ otherwise. The network can then be partitioned into two segments based on y. The algorithm was recursively implemented on each partition separately to generate smaller sub-partitions, and terminated when the isoperimetric ratio reaches a predefined threshold.



Gene Sets	1	2	3	4
1	1	0	1	0
2	0	1	0	0
3	1	0	1	1
4	0	0	1	1

Figure 2: The flowchart of our context-specific gene set assembly pipeline. The PPI network is integrated from 4 major databases and the PDI network is reconstructed using ARACNe, an information theory-based reverse engineering algorithm. The cardiac interactome consists of both PPI and PDI networks. Network partitioning is first implemented to identify initial subnets with strong connectivity (PPI) or sharing a common regulator (PDI). Then, pathway and functional annotations are taken into consideration to merge neighboring subnets with a functional tie. The final sets of subnets will form a collection of Cell context-Specific Gene Sets (CSGS) that could facilitate GSA in various applications, ranging from understanding disease pathogenesis to the discovery of patient stratification biomarkers.

2.4 Network Merging

The subnets obtained from the network partitioning stage were produced solely based on physical connectivity. However, protein functions should also be considered in gene sets identification. Thus, we considered merging the neighboring subnets if they share a functional annotation significantly. The initial subnets are considered to merge in terms of pathways and molecular functional information available from the Molecular Signature Database (MSigDB) [1]. All curated gene sets (c2), motif gene sets (c3), and GO terms sets (c5) were used (5329 sets in total). A hypergeometric test was used to determine the significance of overlap between a subnet and a MSigDB gene set [29]. Let s_i $(1 \le i \le l)$ be a gene set identified from network partition and $m_j\,(1 \le j \le r)$ an MSigDB gene set (i.e. the reference gene set, "ref" as showed in Figure 2), where l is the number of subnets identified and r is the number of reference gene sets used. The null hypothesis of the hypergeometric test assumes that there is no overlap between s_i and m_j. The 3-step procedure is described as follows: (i) compute the p value of the hypergeometric test between each s_i and m_j pair; (ii) create an $l \times l$ merging matrix $M : M[i,j]=1$, if gene sets s_i and s_j were both significantly overlapping ($p < 1\times10^{-6}$) with one of the MSigDB sets. The threshold was generated by permuting the MSigDB sets 1000 times while maintaining the original size for each MSigDB set, and the value satisfying 0.05 significant level was selected as the threshold; (iii) build a new graph from the merging matrix representing functional links between subnets, where nodes are subnets and edges are the connectivity obtained from step (ii). Maximum cliques, i.e. functionally related subnets, can be identified from the new graph, and merged to create a new subnet (see Figure 2).

2.5 GSEA for PDI-based Gene Sets

Gene sets were tested for their difference using Gene Set Enrichment Analysis (GSEA) [1]. An extended version of GSEA (termed GSEA2) [7] was utilized for the PDI-based gene sets. Since the regulon genes include both TF-induced and TF-repressed genes, treating these two subsets in the same way would dilute the significance of the enrichment analysis. For instance, if a TF gained additional activities from nonfailing/normal heart to cardiomyopathy, we would expect the TF-activated subset to be enriched in the upregulated genes, while the TF-repressed subset to be enriched in the downregulated genes. Thus, GSEA2 was formulated exactly to address this circumstance. The GSEA2 proceeds as follows: (a) Compute T-test between two sample groups, e.g. normal and diseased samples, for each of the N genes in microarray. Order the N genes by T-statistics from the most positive to the most negative values, denoted by Q; (b) identify overlaps independently for the positive target genes $G+$ in Q, and the negative target genes $G-$ in $\overline{Q}$, in which $\overline{Q}$ is the inversed ranking of Q with the inverted T-statistics; (c) Combine Q and $\overline{Q}$ and reorder the T-statistics by keeping the positions of overlap for both $G+$ and $G-$, denoted as Q_c; (d) Compute a running score by walking down the combined ranking Q_c. The score will increase by $|q_i|^p \big/ \sum_{1\le i\le|G|} |q_i|^p$ if the ith gene is a hit, or otherwise decrease by $1/(2N-|G|)$, where G is the combined set of $G+$ and $G-$. Finally, (e) an Enrichment Score (ES) is determined as the sum of the maximum and the minimum deviation from zero along the running score. We randomly permuted the phenotype labels and repeated steps (a) to (d) for 1000 times to compute the ES null distribution. Statistical significance of the ES can be computed by comparing the observed ES to the null distribution.

To compute PDI-based GSEP, GEP is first converted into z-score, for each of the N genes in microarray. In each sample, all genes were sorted by the z-score and then used as the reference list, Q, in steps (b) to (e) as above. The GSEP is approximated by the Normalized Enrichment Score (NES) computed for each regulon in all samples.

A MATLAB function implementing GSEA2 described above is available from Matlab Central (http://www.mathworks.com/matlabcentral/fileexchange/33599).

3 Results

3.1 Cardiomyopathy-specific gene sets assembly

Cardiomyopathy interaction network was reconstructed by combining PPI and PDI networks. The PPI network was integrated from 4 major PPI databases to define a network of 172,779 physical interactions that occur between 12,967 human proteins (see Method section). To bring in context specificity for the initial network, filters were applied by removing interactions in which individual genes are not expressed and gene pairs are not co-expressed across the large cohort of cardiomyopathy GEP. The filtered PPI network consists of 9,522 nodes and 117,199 edges. For PDI network, we chose to reconstruct the network *ab initio* due to the low coverage of the available PDI databases. The PDI network was built by applying ARACNe, an information theory-based reverse engineering algorithm that can infer genome-wide transcriptional interaction network from GEP (see Method section). The method has been widely utilized and proven effective in identifying transcriptional targets based on biochemical validations in various cell types [30-32]. The inferred cardiomyopathy PDI network contains 242,714 transcriptional interactions.

The PPI network was partitioned into smaller subnets by maximizing intra-subnet connectivity and minimizing inter-subnet connectivity using isoperimetric algorithm (see Method Section). A reference node (or ground node) was first selected, and all other nodes were then assigned a value, by solving a linear system with the isoperimetric algorithm, that reflects how tightly the nodes are connecting to the reference node. A threshold was determined at a value that provides partitions with the lowest isoperimetric ratio. The solutions can be interpreted as the expected number of steps taken by a random walker leaving node vi before reaching the reference node. A higher value indicates the existence of more paths for node vi to take in order to the reach the reference node and thus more likely to belong to the same cluster as the reference node, while a lower value indicates otherwise. In this paper, we partitioned the network recursively at each segment, and chose a stopping criterion that maximizes the number of subnets with node size greater than 5. A total of 694 subnets were produced.

The initial PPI subnets were produced solely based on physical connectivity and none of the proteins coexisted in any two subnets. However, many proteins serve multiple purposes in the cell and they should also appear in more than one functional gene set. Thus, we considered merging the neighboring subnets if they share a functional annotation significantly. The merging step results in 648 PPI subnets that will contribute to the CSGS. A mapping table between CSGS and MSigDB has been created as a cross-reference (Table S1). Figure 3 shows an example of merging two neighboring subnets to form a single gene set due to sharing a common functional pathway. The nodes were initially considered as two separate subnets by their connectivity. However, approximately half of the nodes in each subnet involve in MAPK signaling pathway, which makes it more functionally relevant to be considered as a single joined gene set. The merged PPI subnets constitute the first type of CSGS.

The second type of the CSGS comprises activities of transcription factor (TF) regulators. TF activities have been known to crucially involve in cell lineage determination, and recently shown to play a master integrator role in brain tumor pathogenesis [30]. Since the activity of a protein is not necessarily proportional to its mRNA concentration, due to post-transcriptional modifications, we used the regulon of a TF as an indicator of the TF's activity. Thus, the corresponding TF regulator divides the initial PDI network such that all genes sharing a common TF formed a gene set. Each gene set was divided into inducible targets and repressible targets by the TF. We identified 1,055 gene sets of size greater than 10.

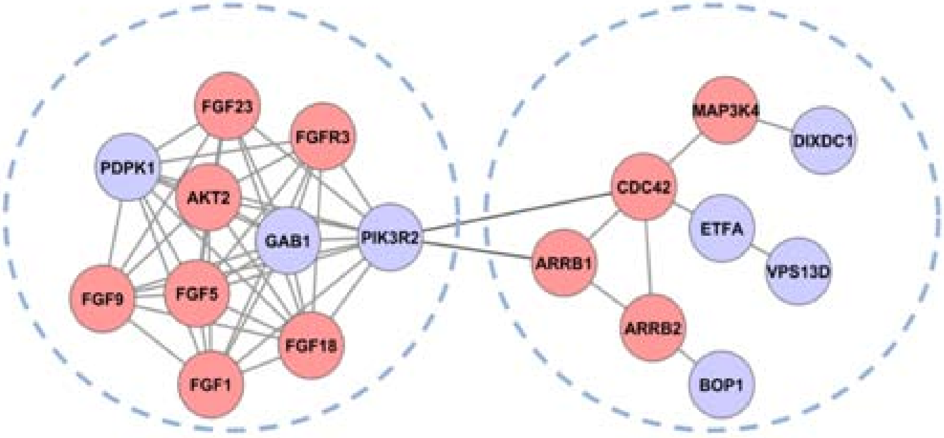


Figure 3: Example of a gene set formed by two subnets through functional merging. Dotted circles represent the initial subnets from network partitioning. Red nodes are proteins involved in MAPK signaling pathway.

3.2 Reproducibility and robustness in gene set analysis

The most appealing feature in GSA is the integration of additional knowledge into omics analysis to achieve more robust results. Here, we compare the robustness of differentially expressed gene sets using MSigDB and the CSGS collections. Gene sets in both collections were tested for their difference between the normal and idiopathic cardiomyopathy groups using GSEA or GSEA2. All gene sets in MSigDB and all PPI-based gene sets in CSGS were quantified using the standard GSEA statistics, while GSEA2 was utilized for the PDI-based gene sets. Finally all gene sets were sorted by statistical significance (p-value) computed using 1000 sample permutations.

We repeated the GSEA procedure 100 times, each time with a random subset of the nonfailing and cardiomyopathy samples. Figure 4 shows the occurring frequency (OF) of the top 50 most significant gene sets in CSGS and MSigDB. The OF is defined as the number of times a specific gene set appears in the top 50 most significant sets ranked by p-value. The most noteworthy result between the two OF distributions is the fatter tail of CSGS curve as compared to the distribution of MSigDB OF (11 CSGS vs. 3 MSigDB in the rightmost bin), indicating the higher robustness in gene sets selection using CSGS. To further explore the function of the highly conserved CSGS, GO term enrichment analysis was applied in the following section.

3.3 Molecular mechanisms underlying cardiomyopathy

A total of 1089 genes in the highly conserved CSGS were interrogated for their biological processes by GO term enrichment analysis, and visualized using Gorilla [33]. Table S2 shows the enriched GO terms along with descriptions and p-values. For comparison, Table S3 shows the MSigDB gene sets whose occurring frequency is greater than 70. Figure 5 illustrates two clusters of enriched GO terms as directed acyclic graphs, illustrating relationship of the terms and their degree of enrichment. The biological pathways and processes revealed by the functional enrichment analysis in CSGS recapitulate those previously known to associate with cardiomyopathies, whereas the MSigDB analysis shows enrichment in hypoxia-related gene sets and other irrelevant activities such as cancers.

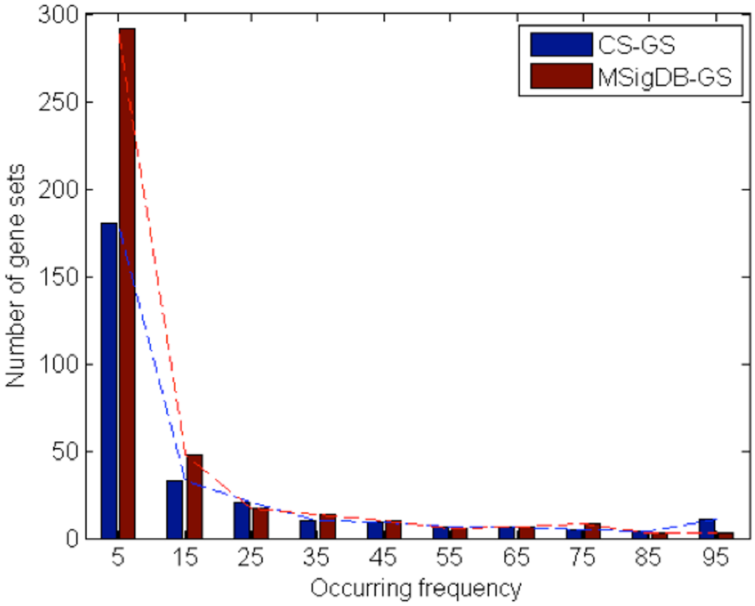


Figure 4: Occurring frequency distributions for the top 50 most significant gene sets. Dashed lines indicate the fitted distributions.

Figure 5(A) shows the hierarchical structure of cytokine-related GO terms for biological processes. Various studies have shown that cytokines are related to heart failure by modulating cardiovascular function through a variety of mechanisms [34]. For instance, increases in interleukin-1 (IL-1) and interleukin-6 (IL-6) cytokines are highly correlated to pulmonary hypertension, which is a sign of diastolic dysfunction leading to heart failure [35]. As shown in Figure 5(A), IL-1 and IL-6 mediated pathways are both significantly enriched in the highly conserved CSGS ($p=4.9\times10^{-4}$ and $p=7.8\times10^{-4}$ for IL-1 and IL-6 mediated pathway, respectively). Furthermore, cytokines are also playing an important role in the pathogenesis and pathophysiology of myocarditis and dilated cardiomyopathy [36].

Figure 5(B) shows another cluster of biological processes that are enriched in the highly conserved CSGS. This cluster is mainly associated with metabolic processes, which includes GO terms such as glucose metabolic process ($p=1\times10^{-9}$) and energy reserve metabolic process ($p=1\times10^{-7}$). Despite the poor understanding of dysregulated metabolic processes in cardiomyopathy, evidences are showing that decrease in protective glucose metabolism and increase in adverse free fatty acid metabolism will cause alteration in many aspects of cardiomyocyte energetics, which is one of the major reasons of heart failure [37]. As the myocardium fails, there are significant changes in the heart's ability to supply adequate energy for its needs and the increase in mortality has been observed among cardiomyopathy patients with lower cardiac energy reserve [38]. A better understanding and detailed characterization of mechanisms and pathways regulating cardiac metabolism will eventually lead to new therapies for heart failure [39].

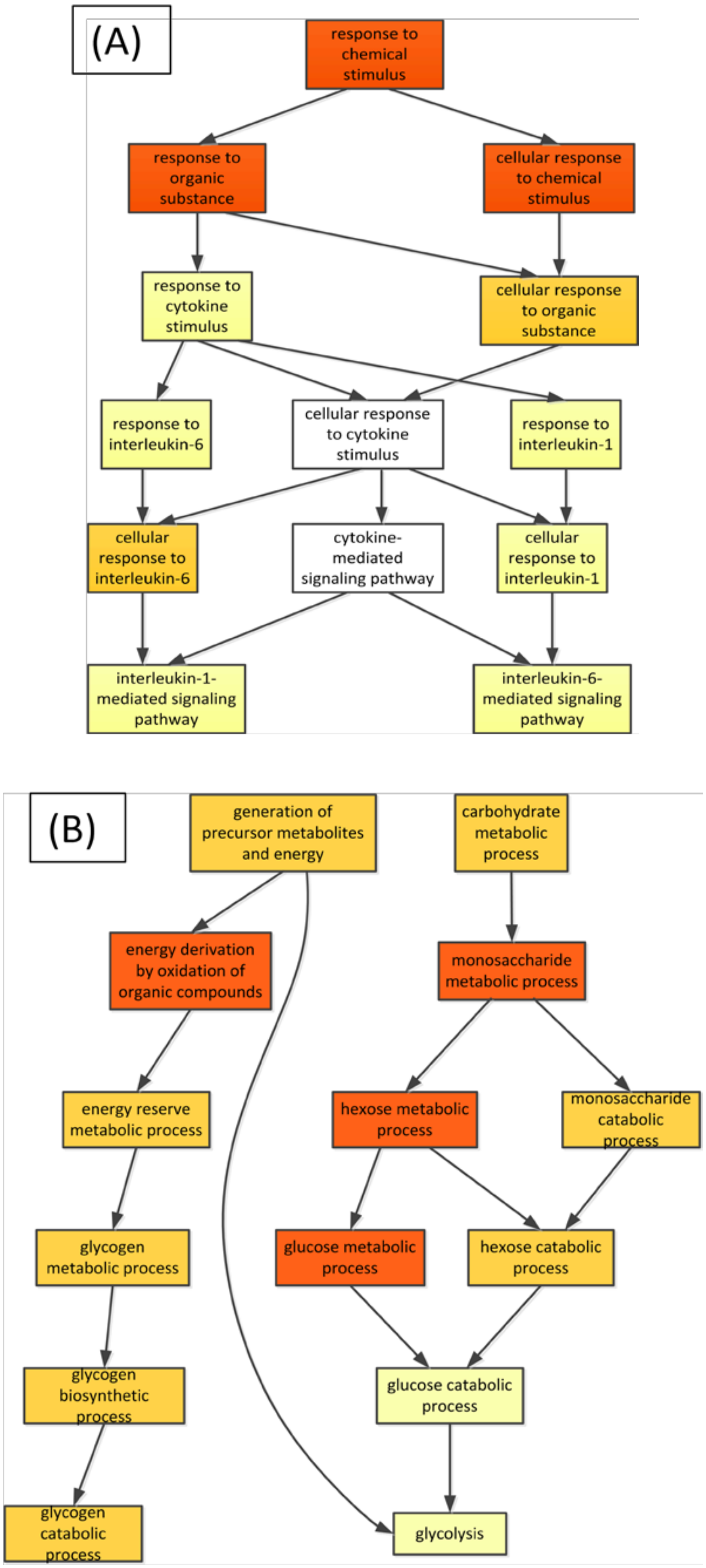

(A)
response to chemical stimulus
response to organic substance
cellular response to chemical stimulus
response to cytokine stimulus
cellular response to organic substance
response to interleukin-6
cellular response to cytokine stimulus
response to interleukin-1
cellular response to interleukin-6
cytokine-mediated signaling pathway
cellular response to interleukin-1
interleukin-1-mediated signaling pathway
interleukin-6-mediated signaling pathway
(B)
generation of precursor metabolites and energy
carbohydrate metabolic process
energy derivation by oxidation of organic compounds
monosaccharide metabolic process
energy reserve metabolic process
hexose metabolic process
monosaccharide catabolic process
glycogen metabolic process
glucose metabolic process
hexose catabolic process
glycogen biosynthetic process
glucose catabolic process
glycogen catabolic process
glycolysis

Figure 5: Hierarchical representation of the enriched GO terms: cytokine related process (left), and metabolic process (right). Boxes are color-coded by the enrichment *p*-value, where red is the most significant ($p<1\times10^{-9}$) and yellow is the least significant ($p>1\times10^{-3}$).

3.4 Gene set-based cardiomyopathy prediction

We assessed the MSigDB-based and CSGS-based classifier for their performance in distinguishing idiopathic cardiomyopathy from normal GEP samples. The GEP was first transformed into z-scores, gene by gene. In each sample, all genes were sorted by their expression z-score to produce a reference gene list in GSEA, and each gene set was interrogated against this list to compute a Normalized Enrichment Score (NES) as the surrogate for gene set expression. For PDI-based gene sets, we calculated NES using GSEA2, and converted the GEP matrix into gene set expression profile (GSEP) matrix.

Top n gene sets were selected by p-value of GSEP differential expression between nonfailing and idiopathic cardiomyopathy groups. The selected features were used to optimize parameters and build an RBF-kernel SVM classifier [40] in the training set: (a) linearly rescale each feature to the range [0, 1]; (b) Use 5-fold cross-validation and grid-search to find the best (C,γ) parameter in a subset of training data. In grid-search, parameter C increased from 2-5 to 215 and parameter γ increased from 2-15 to 23, in exponentially growing steps. Since the number of samples in these two groups is highly imbalanced, the weighted option was used to correct for the bias.

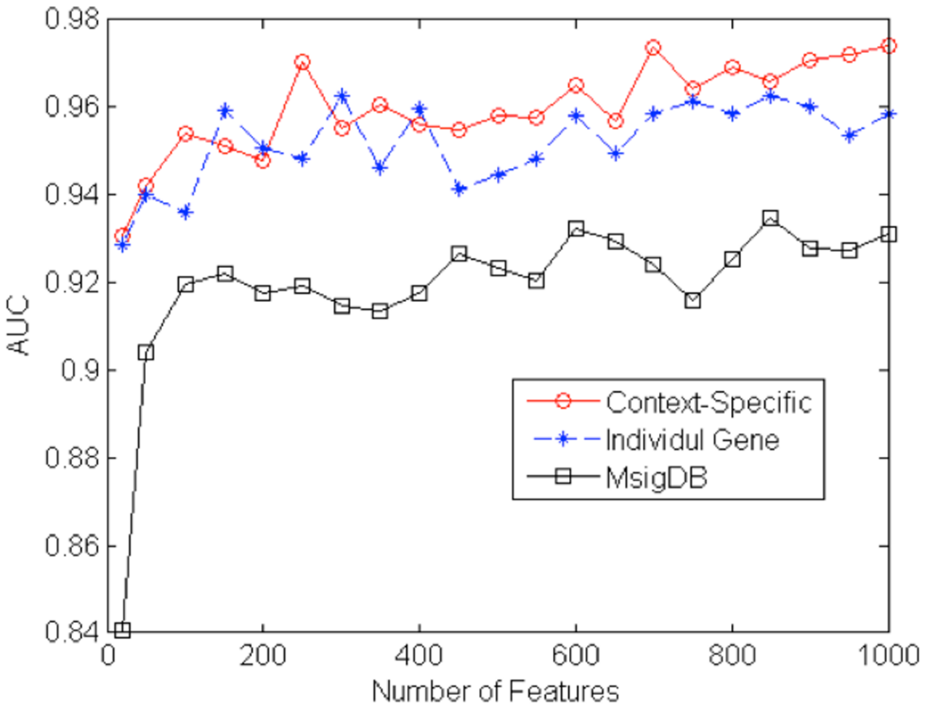


Figure 6: Comparison of the classification performance using (a) individual gene expression profile, (b) CSGS expression, and (c) MSigDB gene sets expression.

The classifications were performed 100 times with randomly selected training (2/3) and testing subsets (1/3). To evaluate the classification performance, we compared the results of the CSGS-based classifier with the MSigDB-based classifier, as well as with those using a convention gene-based classifier. All classifiers were trained/tested using the same scheme, with the only difference in the way a feature value is computed. A gene-based classifier uses the expression value of a gene as a feature, while a gene set-based (CSGS or MSigDB) classifier uses NES value of a gene set (GSEP). The area under receiver operating characteristic (AUC) was used as the comparison metric (Figure 6). The CSGS-based classifier produced a slightly higher AUC than that of the individual gene based classifier and this tendency becomes more apparent as the number of features increase. The MSigDB-based classifier had the worst performance among the 3 classifiers.

Besides accuracy, a robust classifier should hold a stable feature set across different studies. Figure 7 illustrates the variation of the top features in 100 runs using randomly selected data sets. Higher density at the bottom of the plots indicates higher agreement between two sets of features generated from two different training sets. The CSGS showed a more robust feature selection compared with MSigDB gene sets and individual genes. On average, the overlapping percentages of the top features (ranked by t-test) in 100 runs were 58%, 38%, and 41% respectively for CSGS, MSigDB and individual gene.

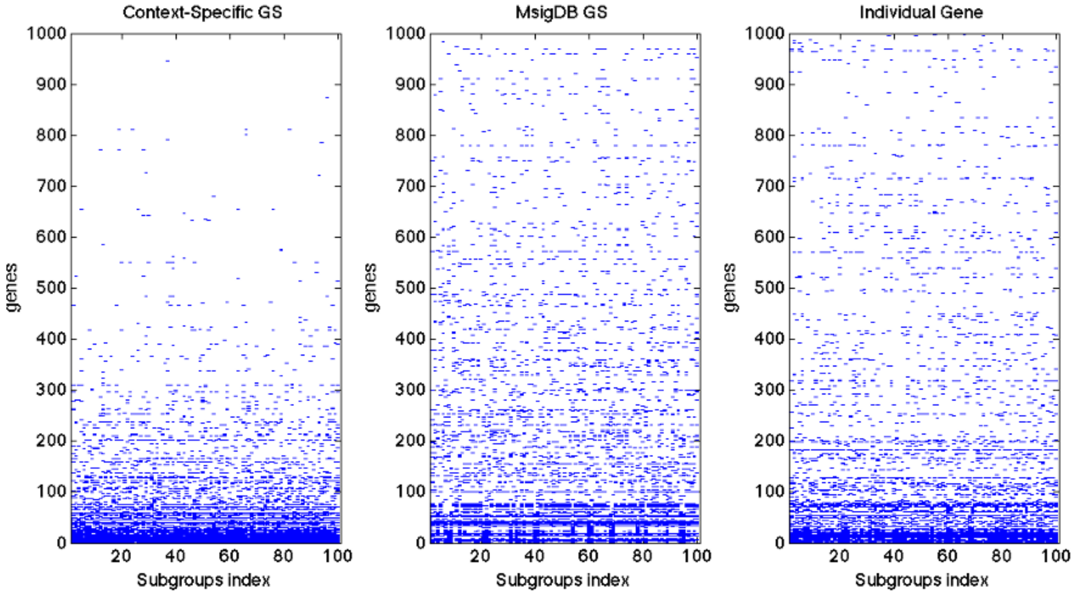


Figure 7: Comparison of the robustness in features selection for the classifier based on CSGS, MSigDB and individual gene. The x-axis represents different tests, and y-axis represents the rank of features in each test, as compared to the rank from the first test.

4 Discussion

GSA has established itself as an undisputed standard for omics data analysis, yet their implementation has always relied on the generic gene set collection until now. The standard database, e.g. MSigDB, is well suited in GSA aiming to identify common properties across different studies, such as identifying canonical pathways or molecular gene signature that are also dysregulated in other disease conditions. However, in order to uncover novel activities unique to a specific cell condition, GSA requires a new strategy to incorporate de novo assembly of a functional gene set collection that define physiological and pathological behavior within the cellular context of interest. These gene sets are substantially different from cell type to cell type due to the cell lineage determining gene expression and interaction that are driven by their unique regulatory elements. We demonstrated such a strategy here by generating a collection of cardiomyopathy-specific gene sets, but the pipeline is broadly applicable to produce gene sets for any other cellular context.

Our gene sets assembly utilizes the framework of network biology, envisioning cells as a complex web of macromolecular interactions [41]. Context specificity of the network was introduced through the gene expression and coexpression across a wide variety of cellular phenotypes. For instance, a PPI will be removed if (a) one of the interacting genes does not express, and (b) the gene pair does not coexpress in the cell context of interest. Although this is a rather naïve method that directly correlates protein activity with gene expression, the filtering criteria ensure fulfillment of the ‘minimum’ condition – no gene expression implies no protein activity. Such a basic criteria could remove >30% of the interactions from the generic PPI databases. The interaction network, including the transcriptional interactions inferred from GEP, constitutes an initial draft of cardiomyopathy interactome and the model should improve over time, as more valuable datasets are being produced to elucidate interaction dynamics [42]. Also, the current framework considers only interactions between proteins and mRNA, but when more advanced quantitative technology is available, impacts of the newly discovered molecular entities such as non-coding RNA (ncRNA) should be added into the equations of cell regulation, and form gene sets that could better represent the overall functional landscape of the biological system.

One of the challenges in the gene sets assembly is to define the boundary of subnets, due to the limited knowledge of true pathways. The issue is further complicated by the low coverage of the available context specific interaction information. Thus, our assembly pipeline includes a merging step after network partitioning in order to identify subnets that are connected both physically and functionally. The network partition step takes into account physical connectivity of the genes, while the merging step further considers pathway and functional annotations of the neighboring nodes. First, such strategy ensures that a *bona fide* subnet will not be disconnected due to missing links among the nodes. More importantly, the subnets boundary can overlap with neighbors from all directions, and proteins involved in multiple pathways will now coexist in all the corresponding gene sets.

A key application of omics profile analysis is the identification of small gene signatures to be used as the disease biomarkers. However, reproducing the signature genes in different studies has been extremely challenging, and methods relying on pathway/network-based gene sets have emerged as the broadly acceptable solutions [7, 43, 44]. In our comparison for cardiomyopathy classification, the results showed that AUC calculated based on CSGS-based features is slightly better than the ones based on individual gene. Advantages in terms of the final classification results are not so obvious here mainly due to the already high AUC. However, the robustness in feature selections clearly reveals the power of CSGS-based features in identifying valuable biomarker candidates and further elucidating molecular mechanisms underlying the disease.

Supplementary Information

Supplementary information is available at bioinformatics.cs.vt.edu/~mingming/csgs/.

Acknowledgements

The authors are grateful to Leo Grady for help with using the network partitioning package and reviewing the manuscript.

References

[1] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profile. *Proceedings of the National Academy of Sciences U.S.A.*, 102:15545-15550, 2005.

[2] D. Nam and S.-Y. Kim. Gene-set approach for expression pattern analysis. *Briefings in Bioinformatics*, 9:189-197, 2008.

[3] B. L. Fridley, G. D. Jenkins, and J. M. Biernacka. Self-contained gene-set analysis of expression data: an evaluation of existing and novel methods. *PLoS One,* 5:e12693, 2010.

[4] J. J. Goeman, S. A. v. d. Geer, F. d. Kort, and H. C. v. Houwelingen. A global test for groups of genes: testing association with clinical outcome. *Bioinformatics*, 20:93-99, 2004.

[5] S.-Y. Kim and D. J. Volsky. PAGE: Parametric Analysis of Gene Set Enrichment. *BMC Bioinformatics,* 6:144, 2005.

[6] H. Lee, W. Braynen, K. Keshav, and P. Pavlidis. ErmineJ: Tool for functional analysis of gene expression data sets. *BMC Bioinformatics,* 6:269, 2005.

[7] W. K. Lim, E. Lyashenko, and A. Califano. Master regulator used as breast cancer metastasis classifier. in *Pacific Symposium on Biocomputing*, 2009, pp. 504-515.

[8] Y. Kim, S. Wuchty, and T. M. Przytycka. Identifying Causal Genes and Dysregulated Pathways in Complex Diseases. *PLoS Computational Biology*, 7:e1001095, 2011.

[9] A. Califano, A. J. Butte, S. Friend, T. Ideker, and E. Schadt. Leveraging models of cell regulation and GWAS data in integrative network-based association studies. *Nature Genetics*, 44:841-847, 2012.

[10] D. Li, E. J. Duell, K. Yu, H. A. Risch, S. H. Olson, C. Kooperberg, B. M. Wolpin, L. Jiao, X. Dong, B. Wheeler, *et al.* Pathway analysis of genome-wide association study data highlights pancreatic development genes as susceptibility factors for pancreatic cancer. *Carcinogenesis*, 33:1384-1390, 2012.

[11] L. d. l. Fuentes, W. Yang, V. G. Dávila-Román, and C. C. Gu. Pathway-based genome-wide association analysis of coronary heart disease identifies biologically important gene sets. *European Journal of Human Genetics*, 20:1168–1173, 2012.

[12] B. Meder, J. Haas, A. Keller, C. Heid, S. Just, A. Borries, V. Boisguerin, M. Scharfenberger-Schmeer, P. Stähler, M. Beier, *et al.* Targeted Next-Generation Sequencing for the Molecular Genetic Diagnostics of Cardiomyopathies. *Circulation*, 4:110-122, 2011.

[13] B. Zhang, S. Kirov, and J. Snoddy. WebGestalt: an integrated system for exploring gene sets in various biological contexts. *Nucleic Acids Research*, 33:W741-W748, 2005.

[14] S. Jung, M. Verdicchio, J. Kiefer, D. V. Hoff, M. Berens, M. Bittner, and S. Kim. Learning contextual gene set interaction networks of cancer with condition specificity. *BMC Genomics,* 14:110, 2013.

[15] M. A. Sartor, V. Mahavisno, V. G. Keshamouni, J. Cavalcoli, Z. Wright, A. Karnovsky, R. Kuick, H. V. Jagadish, B. Mirel, T. Weymouth, *et al.* ConceptGen: a gene set enrichment and gene set relation mapping tool. *Bioinformatics,* 26:456-463, 2010.

[16] D. Glez-Peña, G. Gómez-López, D. G. Pisano, and F. Fdez-Riverola. WhichGenes: a web-based tool for gathering, building, storing and exporting gene sets with application in gene set enrichment analysis. *Nucleic Acids Research*, 37:W329-W334, 2009.

[17] A. C. Culhane, M. S. Schröder, R. Sultana, S. C. Picard, E. N. Martinelli, C. Kelly, B. Haibe-Kains, M. Kapushesky, A.-A. St Pierre, W. Flahive, *et al.* GeneSigDB: a

manually curated database and resource for analysis of gene expression signatures. *Nucleic Acids Research*, 40:D1060-D1066, 2012.

[18] J. T. Chang and J. R. Nevins. GATHER: a systems approach to interpreting genomic signatures. *Bioinformatics*, 22:2926-2933, 2006.

[19] S. Hannenhalli, M. E. Putt, J. M. Gilmore, J. Wang, M. S. Parmacek, J. A. Epstein, E. E. Morrisey, K. B. Margulies, and T. P. Cappola. Transcriptional genomics associates FOX transcription factors with human heart failure. *Circulation*, 114:1269-1276, 2006.

[20] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4:249-264, 2003.

[21] C. Stark, B.-J. Breitkreutz, A. Chatr-aryamontri, L. Boucher, R. Oughtred, M. S. Livstone, J. Nixon, K. Van Auken, X. Wang, X. Shi, *et al.* The BioGRID Interaction Database: 2011 update. *Nucleic Acids Research*, 39:D698-D704, 2011.

[22] S. Kerrien, B. Aranda, L. Breuza, A. Bridge, F. Broackes-Carter, C. Chen, M. Duesbury, M. Dumousseau, M. Feuermann, U. Hinz, *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, 40:D841-D846, 2012.

[23] A. Ceol, A. Chatr Aryamontri, L. Licata, D. Peluso, L. Briganti, L. Perfetto, L. Castagnoli, and G. Cesareni. MINT, the molecular interaction database: 2009 update. *Nucleic Acids Research*, 38:D532-D539, 2010.

[24] D. Croft, G. O'Kelly, G. Wu, R. Haw, M. Gillespie, L. Matthews, M. Caudy, P. Garapati, G. Gopinath, B. Jassal, *et al.* Reactome: a database of reactions, pathways and biological process. *Nucleic Acids Research*, 39:D691-D697, 2011.

[25] A. A. Margolin, K. Wang, W. K. Lim, M. Kustagi, I. Nemenman, and A. Califano. Reverse engineering cellular network. *Nature protocols*, 1:662-671, 2006.

[26] L. Grady and E. L. Schwartz. Isoperimetric graph partitioning for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28:469-475, 2006.

[27] L. Grady and E. L. Schwartz. Isoperimetric Partitioning: A new algorithm for graph partitioning. *SIAM Journal on Scientific Computing*, 27:1844-1866, 2006.

[28] L. Grady, T. Schiwietz, S. Aharon, and R. Westermann. Random walks for interactive organ segmentation in two and three dimensions: Implementatioin and validation. *Proc. of MICCAI*, 773-780, 2005.

[29] J. A. Rice. *Mathmatical Statistics and Data Analysis*: Duxbury Press, 2007.

[30] M. S. Carro, W. K. Lim, M. J. Alvarez, R. J. Bollo, X. Zhao, E. Y. Snyder, E. P. Sulman, S. L. Anne, F. Doetsch, H. Colman, *et al.* The transcriptional network for mesenchymal transformation of brain tumours. *Nature*, 463:318-325, 2010.

[31] T. Palomero, W. K. Lim, D. T. Odom, M. L. Sulis, P. J. Real, A. Margolin, K. C. Barnes, J. O'Neil, D. Neuberg, A. P. Weng, *et al.* NOTCH1 directly regulates *c-MYC* and activates a feed-forward-loop transcriptional network promoting leukemic cell growth. *Proceedings of the National Academy of Sciences U.S.A.*, 103:18261-18266, 2006.

[32] C. Lefebvre, P. Rajbhandari, M. J. Alvarez, P. Bandaru, W. K. Lim, M. Sato, K. Wang, P. Sumazin, M. Kustagi, B. C. Bisikirska, *et al.* A human B-cell interactome identifies MYB and FOXM1 as master regulators of proliferation in germinal centers. *Molecular Systems Biology*, 6:377, 2010.

[33] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: A Tool For Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists. *BMC Bioinformatics,* 10:48, 2009.

[34] A. Blum and H. Miller. Role of cytokines in heart failure. *American Heart Journal,* 135:181-186, 1998.

[35] A. Matsumori. Cytokines in myocarditis and dilated cardiomyopathy. *European Heart Journal Supplements,* 4:I42-I45, 2002.

[36] M. Humbert, G. Monti, F. Brenot, O. Sitbon, A. Portier, L. Grangeot-Keros, P. Duroux, P. Galanaud, G. Simonneau, and D. Emilie. Increased interleukin-1 and interleukin-6 serum concentrations in severe primary pulmonary hypertension. *American Journal of Respiratory and Critical Care Medicine,* 151:1628-1631, 1995.

[37] W. C. Stanley, F. A. Recchia, and G. D. Lopaschuk. Myocardial Substrate Metabolism in the Normal and Failing Heart. *Physiological Reviews,* 85:1093-1129, 2005.

[38] L. H. Opie, P. J. Commerford, B. J. Gersh, and M. A. Pfeffer. Controversies in ventricular remodelling. *Lancet,* 367:356-367, 2006.

[39] H. Ardehali, H. N. Sabbah, M. A. Burke, S. Sarma, P. P. Liu, J. G. Cleland, A. Maggioni, G. C. Fonarow, E. D. Abel, U. Campia, *et al.* Targeting myocardial substrate metabolism in heart failure: potential for new therapies. *European Journal of Heart Failure,* 14:120-129, 2012.

[40] C. C. Chang and C. J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology,* 2:27, 2011.

[41] M. Vidal, M. E. Cusick, and A. Barabási. Interactome Networks and Human Disease. *Cell,* 144:986-998, 2011.

[42] T. M. Przytycka, M. Singh, and D. K. Slonim. Toward the dynamic interactome: it's about time. *Briefings in Bioinformatics,* 11:15-29, 2010.

[43] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics,* 21:171-178, 2005.

[44] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology,* 3:140, 2007.