Network-based identification of disease genes in expression data: the GeneSurrounder method

Sahil Shah¹ and Rosemary Braun*¹,²

¹Engineering Sciences and Applied Mathematics
²Biostatistics, Feinberg School of Medicine and Northwestern Institute on Complex Systems
Northwestern University

June 1, 2017

Abstract

Motivation: The advent of high-throughput transcription profiling technologies has enabled identification of genes and pathways associated with disease, providing new avenues for precision medicine. A key challenge is to analyze this data in the context of the regulatory networks and pathways that control cellular processes, while still obtaining insights that can be used to design new diagnostic and therapeutic interventions. While classical differential expression analysis provides specific and hence targetable gene-level insights, it does not include any systems-level information. On the other hand, pathway analyses integrate systems-level information with expression data, but are often limited in their ability to indicate specific molecular targets.

Results: We introduce GeneSurrounder, an analysis method that takes into account the complex structure of interaction networks to identify specific genes that disrupt pathway activity in a disease-specific manner. GeneSurrounder integrates transcriptomic data and pathway network information in a novel two-step procedure to detect genes that (i) appear to influence the expression of other genes local to it in the network and (ii) are part of a subnetwork of differentially expressed genes. Combined, this evidence can be used to pinpoint specific genes that have a mechanistic role in the phenotype of interest. Applying GeneSurrounder to three distinct ovarian cancer studies using a global KEGG network, we show that our method is able to identify biologically relevant genes and genes missed by single-gene association tests, integrate pathway and expression data, and yield more consistent results across multiple studies of the same phenotype than competing methods.

Availability: The implementation in R is available at github.com/sahildshah1/gene-surrounder

Contact: rbraun@northwestern.edu

1 Introduction

Identifying genes whose expression is associated with a specific phenotype is a key step in understanding disease mechanisms and developing precisely targeted diagnostic and therapeutic interventions. Since the advent of high-throughput transcription profiling technologies (such as microarrays and RNA sequencing), gene expression data has become a powerful tool for the identification of disease-associated genes, providing a “snapshot” of transcriptional activity across the entire genome. Most commonly, differentially expressed genes are identified by testing each gene independently for significant differences in mean expression between the phenotypes [1]. However, while differential
expression analysis can identify specific (and hence targetable) disease-associated genes, it does not take into consideration the network of molecular interactions that govern cellular function, limiting the mechanistic insights that can be derived from the data. As a result, this analysis can miss crucial multi-gene interactions that underlie complex phenotypes, leading to poor agreement between different studies of the same conditions [2, 3, 4].

To address these challenges, network-based analysis techniques that can capture systems-level effects have been developed, enabled by databases that organize genes into experimentally derived networks of cellular interactions (such as KEGG [5]). Pathway analysis techniques combine the information encapsulated in the networks of functionally related genes or gene products (i.e. pathway maps) with gene expression data to identify the systems associated with a disease. By incorporating systems-level information, these techniques can capture multi-gene interactions, yielding mechanistically interpretable results that are more reliable than single–gene analyses [2, 3, 4].

Pathway analysis techniques can be broadly grouped into three categories: “functional scoring methods”, “network-based pathway methods”, and “active modules tools” [3, 4, 6]. *Functional scoring methods* identify functionally-related sets of genes (e.g., genes on a common pathway) that are enriched for association with the phenotype of interest. The key idea of functional scoring methods is to aggregate gene-level changes into a gene-set statistic to be able to identify pathways that may have small, but systematic changes. Gene Set Enrichment Analysis (GSEA) [7] is a popular example of this kind of analysis; a review of other such methods can be found in [3]. *Network-based pathway methods* also calculate pathway-level statistics to identify sets of differentially expressed genes, but differ from functional scoring methods in that they include information about the network of interactions between the genes (rather than treating pathways as simple gene sets). Examples of network–based pathway analysis methods include Signaling Pathway Impact Analysis (SPIA) [8], which aggregates perturbations along network edges, and Centrality Based Pathway Enrichment (CePa) [9, 10], which weights gene–level statistics by network centrality measures. Many other approaches have also been proposed; a comprehensive review of network–based pathway analysis methods may be found in [4]. *Active module tools* identify regions on a network (i.e. modules) that are associated with a phenotype (i.e. active). In contrast to functional scoring methods and network-based pathway methods that flag whole disease-associated pathways, active module tools attempt to find disease–associated subnetworks within pathways. jActiveModules [11] and HotNet [12] are two examples of active module tools that typify two different approaches to identifying regions on a network. The former scores genes and searches for high scoring modules; the latter diffuses “flow” (of, for example, mutation frequency) from genes and identifies modules that accumulate flow. [6] provides a recent review of active module detection methods.

While pathway analysis techniques integrate systems–level information with omic data to provide functional interpretations of the dataset, pathway–level results are generally not specific enough to suggest precise gene targets for follow-up studies. The “significant pathways” identified by such analyses may comprise tens or hundreds of genes, and the boundaries between pathways is often arbitrary. In addition, many techniques rely on user–settable parameters and arbitrary heuristics that depend on network size, limiting their interpretability and reliability [3, 13]. It is thus necessary to develop methods that analyze omic data in the context of the full network of interactions while still providing specific gene-level findings that can be targeted experimentally and therapeutically.

We present a novel analysis technique, GeneSurrounder, that takes into account the complex structure of interaction networks, in the form of a network model of cellular interactions, to identify specific disease-associated genes from expression data. The key idea of our method is to find the genes with neighbors on the network that are differentially expressed (with the magnitude of differential expression decreasing with distance from the putative disease gene) and have correlated expression with the putative disease gene. By finding the genes that affect and are affected by their neighbors in this way, our method identifies candidate genes that are “central” to the mechanisms underlying a given phenotype and does so without any reliance on user-set parameters or arbitrary pathway boundaries.

In this manuscript, we describe the GeneSurrounder algorithm and apply it to three gene expression data sets of high-vs-low grade ovarian cancer [14] to demonstrate the methodological and biological validity of our approach. We find that our method represents an integration of pathway and expression data to yield results that are not solely driven by either alone, detecting relevant
genes that would be missed by solely differential expression analysis. We also find that our results are more concordant than individual gene-level statistics and that our method identifies genes associated with ovarian cancer. Our novel analysis approach complements existing gene- and pathway-based analysis strategies to identify specific genes that control disease-associated pathways, providing a new strategy for identifying promising therapeutic targets.

2 Methods

Our goal is to identify candidate disease genes by analyzing gene expression (or other omic) data in the context of interaction networks to discover genes that drive the behavior of pathways associated with disease. We thus seek to identify genes with two defining characteristics: (i) they appear to influence other genes nearby in the network, as evidenced by strongly correlated expression with nearby genes; and (ii) their dysregulation is associated with disease, as evidenced by a pattern of differential expression centered about that gene.

To this end, the GeneSurrounder method consists of two tests that are run independently of each other (Figure 1) and then combined to determine if the putative disease gene is a “disruptive” candidate disease gene meeting both criteria.

The algorithm takes as input gene expression data and an independent network model of cellular interactions derived from a pathway database. In order to consider the full scope of a gene’s interactions and avoid artificially imposed pathway boundaries, we first construct a global interaction network comprising the graph union of all pathways. Using this global network and gene expression data, we compute evidence for each of the above criteria as follows.

2.1 Evidence of “Sphere of Influence”

The first step, dubbed “Sphere of Influence,” assesses if a candidate gene \(i\) meets the first criterion by testing if gene \(i\) is more strongly correlated with its network neighbors than with a random set of genes (Figure 2). We test this by computing the sum of the the absolute correlations between the expression of \(i\) and every gene within a network neighborhood of radius \(r\), and comparing this sum to a null distribution obtained by summing the absolute correlation between \(i\) and a set of \(M\) other genes selected at random (where \(M\) is the number of genes in the neighborhood of radius \(r\) about \(i\)).

The first step, therefore, of the Sphere of Influence procedure is to calculate the correlation Spearman rank correlation \(\rho_{ij}\) between gene \(i\) and every other gene \(j\) assayed and on the network.

From this set of correlations, we calculate the observed total absolute correlation between gene \(i\) and its neighbors within a neighborhood of radius \(r\),

\[
C_i(r) = \sum_{\{j: d_{ij} \leq r\}} |\rho_{ij}|, \tag{1}
\]

where \(d_{ij}\) indicates the geodesic distance of gene \(j\) from gene \(i\) on the network.

In order to compute the distribution of total correlation under the null, we randomly resample \(\rho_{ij}\) and recompute Equation \(1\). This procedure effectively redistributes the gene–gene correlations about the network, enabling a comparison of gene \(i\)'s influence in the true network neighborhood to its influence on a random selection of genes. This step tests the so-called “competitive null” described in \([3]\); that is, it tests whether gene \(i\) has a greater correlation with genes in its neighborhood, compared to a random set of genes.

The null distribution of the total absolute correlation for gene \(i\) as a function of the neighborhood radius is computed using \(10^3\) resamplings, and the observed total absolute correlation is compared to the resampled null distribution, yielding a \(p\)-value at each neighborhood radius for gene \(i\) \(p_{\text{Sphere}(r)}\) that quantifies whether \(i\) is more influential on its neighbors (or influenced by its neighbors) than expected by chance.
2.2 Evidence of “Decay of Differential Expression”

The previous step tested whether gene $i$ is strongly correlated with its network neighbors, independent of phenotype. We now turn our attention to whether the gene and its neighbors also exhibit an association with the phenotype of interest. In particular, if a gene $i$ is a source of dysregulation that drives the phenotype, we would expect that gene $i$ and its close neighbors will be differentially expressed, while genes farther away in the network will exhibit weaker differential expression. In other words, we expect a pattern of differential expression that is strongly localized about gene $i$. Hence, the second step, “Decay of Differential Expression,” tests whether the magnitude of differential expression of other genes $j$ decays as one moves farther from it in the network. Therefore, we expect a pattern of differential expression that is strongly localized about gene $i$ (Figure 3).

In order to do this, we must first compute a gene–level statistic $g_j$ that quantifies the magnitude of $j$’s association with the outcome of interest. (In the case of a two–class microarray experiment, $g_j$ could be the absolute value of the moderated $t$-statistic [1] from a test of differential expression; other statistics may be used as appropriate for other study designs. Note that in what follows, $- \log_{10} p$ values may be used in all cases without loss of generality.) We then quantify the “decay of differential expression” with the Kendall $\tau$ rank correlation coefficient between the differential expression and distance from gene $i$. We use the Kendall $\tau_B$ rank correlation coefficient because it accounts for ties.

The observed discordance is

$$D_i(r) = \tau_B \left( \left\{ g_j : d_{ij} \leq r \right\}, \left\{ d_{ij} : d_{ij} \leq r \right\} \right),$$

where $d_{ij}$ is the geodesic distance between gene $j$ and gene $i$.

To assess the statistical significance of $D_i(r)$, we randomly permute the phenotype labels and recompute the gene–level association statistics $g_j$ under the null hypothesis that the genes are not meaningfully associated with the phenotype. We then use the permuted $g_j^*$ to recompute $D_i^*$ according to Equation 2. A set of $10^3$ such re-computations forms a reference distribution against which we compare the observed $D_i$ to obtain a $p$ value $p_i^{\text{Decay}}(r)$ as the fraction of $D_i^* < D_i$ (which is expected to be negative).

It should be noted here that while $p_i^{\text{Sphere}}(r)$ (above) was obtained by randomly permuting genes, $p_i^{\text{Decay}}(r)$ is obtained by permuting the class labels. An important feature of the latter is that it preserves correlations between genes that were found in the $p_i^{\text{Sphere}}(r)$ calculation. In consequence, the null models, and hence the interpretations, of the two tests differ. $p_i^{\text{Sphere}}(r)$ quantifies whether the neighborhood surrounding gene $i$ is more strongly correlated with it than a random set of genes would be (independent of phenotype), testing the so–called “competitive null” [3]. In contrast, $p_i^{\text{Decay}}(r)$ assesses whether the neighborhood surrounding gene $i$ is more strongly associated with the phenotype of interest than those same genes would be with randomly–assigned phenotype labels (preserving the organization of genes in the network), thus testing the so–called “self-contained null” [3]. That is, it tests whether a specific set of genes in a neighborhood is more strongly associated with the phenotype of interest than the same set of genes would be for a random phenotype.

Because these two procedures permute orthogonal axes (genes vs. samples), they provide two independent tests with independent interpretations: $p_i^{\text{Sphere}}(r)$ tests whether gene $i$ influences its neighbors, and $p_i^{\text{Decay}}(r)$ tests whether that neighborhood is associated with disease. This enables us to combine the two independent pieces of evidence into a single assessment, as described below.

2.3 Combined Evidence

At this point in our algorithm, the Sphere of Influence and Decay of Differential Expression procedures have been run independently of each other, but neither component is sufficient by itself to determine if putative disease gene $i$ is in fact a “disruptive” candidate disease gene meeting both criteria. Therefore, the last step our method performs is to combine the $p$-values outputted by each component ($p_i^{\text{Sphere}}(r)$ and $p_i^{\text{Decay}}(r)$) using Fisher’s method [15],

$$X^2 = -2(\ln(p_i^{\text{Sphere}}(r)) + \ln(p_i^{\text{Decay}}(r))).$$


\(X^2\) follows a \(\chi^2\) distribution with 2 degrees of freedom, which can be used compute \(p_{\text{Combined}}(r)\), the combined evidence that gene \(i\) is a “disruptive” gene.

### 2.4 Neighborhood Size

As described, our method is only applied to the immediate neighborhood of a putative disease gene, but different genes may have different ‘reaches’ on the network and this extent is not known \textit{a priori}. Therefore, we have devised our analysis technique to be applied to the neighborhood of every radius (up to \(D\) the diameter of the network). Both the Sphere of Influence and Decay of Differential Expression procedures then are modified to output a \(p\)-value for every neighborhood (and the \(p\)-values of each neighborhood are then combined using Fisher’s method \cite{15}) by computing observed and null statistics for every neighborhood radius. Note that for both procedures, we compute the null statistics for every neighborhood from each re-sampling rather than re-sample \(N_{\text{Sphere}}\) and \(N_{\text{Decay}}\) times for each neighborhood. The \(p\)-value our method outputs for each gene (\(p_{\text{NIDG}}\)) is the smallest \(p_{\text{Combined}}\) across all distances. We do not adjust for multiple hypothesis because the \(p\)-values are not independent.

### 3 Results

#### 3.1 Application to Ovarian Cancer Data with Global KEGG Network Model

We applied our algorithm to three gene expression data sets of high-vs-low grade ovarian cancer from the publicly available and curated collection ‘curatedOvarianData’ \cite{14} to illustrate the components of the GeneSurrounder method and evaluate its performance. In order to test our algorithm, we evaluate its cross-study concordance, i.e., its consistency across different data sets that are measuring the same conditions, as previously described \cite{4}. The intuition underlying this approach is that methods that detect true biological signals should find them across different data sets measuring the same conditions. Here we use data from three independent studies of gene expression of 7680 genes (common in all three studies) in patients with high and low grade cancer (Table 1).

Our method combines gene expression data with an independent network model (i.e., not derived from the expression data) to detect the disruptive genes of the phenotype under consideration. We use the same global network model for each study, which we have constructed from KEGG pathways \cite{5}. The KEGG database organizes experimentally derived pathway information into individual networks of functionally related molecules. In the KEGG representation, the nodes (i.e. vertices) are genes or gene products and the links (i.e. edges) are cellular interactions. We merge the individual pathways to create a global network that does not have artificial boundaries and therefore do not assume that parts of the network are independent of one another. We merged the same 247 individual KEGG pathways we tested in \cite{4}. We input the largest connected component of the resulting network into our algorithm and it has \(N = 4867\) nodes, \(L = 42874\) links, and a diameter \(D = 34\). Of the \(N = 4867\) nodes, 2709 of them are also amongst the 7680 genes assayed in all three ovarian cancer studies.

We applied our method to each of the ovarian cancer studies with the global gene network to calculate the combined evidence \(p_{\text{Combined, min}}\) for each of the 2709 genes \(i\) that are assayed and on the network. With the results from each of the three ovarian cancer data sets, we evaluate not only the cross-study concordance of our analysis technique, but also its ability to identify biologically relevant genes, identify genes missed by single-gene association tests, and truly integrate pathway and expression data.

#### 3.2 Illustration of GeneSurrounder on MCM2

By way of illustration, we present the results for each component of our algorithm as applied to gene MCM2 using data from one study of high-vs-low grade ovarian cancer \cite{14} (GEO accession GSE14764). In Figure 4, each of the first three plots (from top to bottom) displays the \(-\log_{10}(p)\) from the Sphere of Influence, Decay of Differential Expression and Combined components of our
method. Since we compute these values as a function of network neighborhood size surrounding that gene, the \( p \)-values are plotted against the neighborhood radius (i.e. radius of geodesic distance from the putative “disruptive” disease disease gene MCM2.

Figure 4A (Sphere of Influence) illustrates the dilution of influence with distance and the effect that the size (i.e. number of assayed genes) of a neighborhood has on the decrease of influence. The putative disease gene in this example, MCM2, has significant influence in neighborhoods near to it, but this influence falls off and stays non-significant at far away distances. The steepest drop occurs between a radius of 5 and 6, where the number of assayed genes within the neighborhood (Figure 4D) increases sharply, contributing to the dilution of MCM2’s influence.

Figure 4B (Decay of Differential Expression) indicates a significant concentration of differential expression for neighborhoods with radii of 4–6. We observe that small neighborhoods immediately near a putative disease gene are not big enough to detect a decaying pattern of differential expression, such that the localized differential expression is only detectable at with a radius of at least 4. At the other end, big neighborhoods are too diverse to exhibit a consistent decay of differential expression; like the sphere of influence, the significance of the decay of differential expression flattens out at large distances.

Figure 4C illustrates the results of combining the results for each neighborhood. The \( p \)-value our method outputs for each gene is the most significant \( p_{\text{Combined}}(r) \) across all neighborhood radii; for MCM2 in this study, this occurs at a neighborhood radius of 4 with \( p_{\text{NIDG}} = 1.48 \times 10^{-5} \). Note that this value is returned for each gene regardless of how significant it is (i.e. without thresholding). Since our method returns the smallest \( p_{\text{Combined},j} \) for each gene (equivalently, the largest \( -\log_{10} p_{\text{Combined}}(r) \)) and the smallest \( p_{\text{Combined}}(r) \) of MCM2 is highly significant, MCM2 would be identified as a central candidate disease of high grade ovarian cancer. From a biological standpoint, this finding is sensible: MCM2 is a DNA replication factor, and therefore likely plays a role in the aggressive proliferation associated with high-grade ovarian carcinoma.

3.3 “Disruptive” genes found by GeneSurrounder are associated with ovarian cancer

To evaluate GeneSurrounder’s ability to identify biologically relevant genes, we compared our results in all three ovarian cancer studies (Table 4) to existing biological knowledge.

Our method found families of protein coding genes (CDC, MCM, ORC, FANC) that are in the union of the top 15 most significant genes in all three ovarian cancer studies. This finding is sensible because CDC genes encode for proteins involved in the cell division cycle, MCM and ORC genes encode for proteins involved in DNA replication, and FNC genes encode for proteins involved in DNA repair. These genes are therefore likely to play a role in high-grade ovarian cancer.

Four the genes found by our method (CDC7, ORC6L, DBF4, and CDC45) were consistently amongst the top–20 most significant genes in all three ovarian cancer studies. We found evidence in the literature that CDC7, ORC6L, and DBF4 are associated specifically with ovarian cancer. The inclusion of CDC45 amongst the top 20 most significant genes as found by our analysis technique in each of the three studies suggests the possibility that it is also associated with ovarian cancer. CDC7 encodes for a cell division cycle protein and has been found to both predict survival and be a powerful anticancer target in ovarian cancer [16]. ORC6L encodes for an origin recognition complex that is crucial for the initiation of DNA replication and has been found to highly expressed in ovarian cancer [17]. DBF4 encodes for a protein that activates the kinase activity of CDC7 and was found to be associated with ovarian cancer [18]. The finding of these genes from studies of high-vs-low grade ovarian cancers suggests the possibility that they are not only involved in ovarian cancer but, more specifically drive high grade ovarian cancer.

Note also that the genes that are not in top–15 for a given ovarian cancer study studies still have relatively high ranks. For each of the three ovarian cancer studies, we have applied our method to all 2709 genes that are assayed and on the network. Therefore RELB which has a rank of of 264 in GSE17260 and 675 in GSE9891 is still within the top 10% and top 25% respectively of genes tested. This finding is a demonstration of the consistency of our method.

This analysis therefore demonstrates that the central candidate disease genes found by GeneSurrounder are associated with ovarian cancer. A table of the full results is provided as a supplementary
3.4 GeneSurrounder identifies genes missed by single gene association tests

To investigate whether the “disruptive” genes detected by our method are those which would have been found using a single gene analysis, we compared the results from our method and a single gene analysis. For each of the ovarian cancer studies, we performed a moderated t-test [1], and calculated the corresponding evidence of differential expression $p_{DE}$ for each of the 2709 genes $i$ that are assayed and on the network. With the results from our analysis technique, we were able to compare $p_{combined}$ and $p_{DE}$ for each study and for each gene. The resulting pairs are shown as a scatter plot in Figure 5. For each of the three studies, we find from the top left of each panel of Figure 5 that there are genes with significant $p_{combined}$, $p_{min}$ and non-significant $p_{DE}$ which indicates that our analysis technique identifies genes that would have otherwise been missed.

3.5 GeneSurrounder represents a true integration of pathway and expression data

The method that we have developed combines gene expression data with an independent network model. To investigate whether our results are driven solely by the either the network or the expression data or represent a true integration of biological knowledge (the pathway networks) and experimental data, we consider the association between our results, the centrality, and the differential expression for each gene. If the results were driven solely by the network or by the expression data, the evidence a gene is a “disruptive” gene would correlate strongly with its centrality in the network (in the former case) or with its differential expression (in the later case). We therefore calculate the correlation between our results and two different measures of centrality and between our results and the differential expression for each of the studies. The results are given in Table 2. We find that for each of the studies, the correlations are small (on the order of 0.01), confirming that GeneSurrounder is not driven solely by network features or the expression data, but rather represents a true integration of biological knowledge (the pathway networks) with experimental data.

3.6 GeneSurrounder findings are more concordant than single–gene association statistics

To investigate the cross-study concordance of our analysis technique (i.e. its consistency across different data sets that are measuring the same conditions), we considered each pair of the three studies and calculated the correlation between our results. As a point of reference, we also calculated the correlation between the gene level statistics for each of the three studies. The results are given in Table 3. By analyzing the concordance of our method’s results across all three studies, we found that they are much more consistent than the classical approach of testing each gene for differences in expression between the cases and controls. This cross–study concordance suggests that our method reliably detects true biological effects (rather than noise) because they are consistent across studies measuring the same underlying condition [4]. The cross–study concordance also demonstrates the reproducibility of results obtained by our method. Importantly, we find that the list of “disruptive” genes detected by GeneSurrounder are more reproducible across studies than differentially expressed genes.

4 Discussion

In this manuscript, we have developed and presented a new analysis technique, GeneSurrounder, that integrates a network model with expression data to identify individual genes that can be targeted therapeutically. Our analysis technique identifies central candidate disease genes, the genes that impact disease associated mechanisms, by finding the genes with neighbors that are
differentially expressed and associated with the putative disease gene. The algorithm therefore consists of two tests which are run independently of each other and then their results are combined. The first test, Sphere of Influence, calculates the evidence that a putative disease gene is correlated with its neighbors, and the second test, Decay of Differential Expression, calculates the evidence that the neighbors of a putative disease gene are differentially expressed (with the magnitude of differential expression decreasing with distance). This approach complements existing methods and provides a new strategy for identifying promising therapeutic targets.

We applied our algorithm to three gene expression data sets of high-vs-low grade ovarian cancer [14] and combined each of them with the same global network model that we constructed from KEGG pathways. With the results from each of the three ovarian cancer data sets, we evaluated our analysis technique. By applying our method to three different data sets that are measuring the same conditions, we were able to show that it yields consistent (i.e. concordant) results across studies, suggesting its ability to detect biologically meaningful associations that are reproducible across studies. We also compare our results to existing biological knowledge and find that our method identifies biologically relevant genes. To show that our method identifies genes missed by single gene association tests and truly integrates pathway and expression data, we compare the results from our method to the results from a single gene analysis and the centrality of the genes in the network. Our positive results along these four dimensions of our analysis technique suggest that our method is a promising new strategy for identifying the genes that control disease.

Our algorithm identifies candidate disease genes from network and expression data by finding the genes with neighbors that are differentially expressed (with the magnitude decreasing with distance) and correlated with the putative disease gene. This approach differentiates our method from other analysis techniques that combine network data with expression data to identify disease genes. In an extension of SPIA [19], disease genes are found by calculating a ‘perturbation factor’ from upstream genes in individual pathways. In another method to identify disease genes [20], they are found using the Laplacian kernel to identify differentially expressed neighborhoods. In LEAN [21], disease genes are found by identifying local subnetworks of interest in a protein-protein interaction network. Unlike our method, LEAN considers only the immediate neighborhood (i.e. at a radius of one) and assess the enrichment of significant genes. These methods are complementary to our method and they can be used in conjunction with one another to identify genes that are promising therapeutic targets.

Finally, we note that although we developed and applied GeneSurrounder to expression data, our method can potentially be applied other other types of omic data. For instance, one might envision applying it to genomic sequence data to identify epistatic interactions, evidenced by gene neighborhoods that have a high level of correlations in their genetic variants. Additionally, while the approach as presented here does not explicitly consider edge directionality, it can be easily modified to do so by only considering downstream genes in the Sphere of Influence and Decay of Differential Expression computation. GeneSurrounder thus provides a computationally efficient and flexible means to detect genes that are central to a disease in the context of gene regulatory networks, providing insights into disease mechanisms and suggesting diagnostic and therapeutic targets.

Acknowledgements

We would like to thank Prof. William Kath, Dr. Marta Iwanaszko, Gary Wilk, and Phan Nguyen for helpful discussions.

Funding

This work was supported by the James S. McDonnell Foundation (grant #220020394) and the National Heart, Lung, and Blood Institute of the National Institutes of Health under award number R01HL128173. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.
Table 1: **Ovarian cancer datasets used in this study**: Comparisons were made between low- and high-grade serous ovarian carcinoma using public data. Sample sizes for each group in each dataset are given. The data are publicly accessible and available as part of the curatedOvarianData package [14].

| GEO Accession No. | N(low-grade) | N(high-grade) |
|-------------------|--------------|---------------|
| GSE14764          | 24           | 44            |
| GSE17260          | 67           | 43            |
| GSE9891           | 103          | 154           |

Table 2: **Correlation between GeneSurrounder results and network/gene statistics**: The three columns are the rank correlation between GeneSurrounder results ($p_{GS}$) and network/gene statistics (Degree, Betweenness, and $p_{DE}$) across all genes in each dataset. The Degree and Betweenness are two different network centrality measures. The Degree is the number of connections a node has and the Betweenness is the fraction of shortest paths that passes through the node. $p_{DE}$ is the p-value obtained from a standard differential expression $t$-test.

| Network/Gene Statistic | GSE14764 | GSE17260 | GSE9891 |
|------------------------|----------|----------|---------|
| Degree Cor.            | 0.044    | 0.070    | 0.038   |
| Betweenness Cor.       | 0.047    | 0.059    | 0.030   |
| $p_{DE}$ Cor.          | 0.060    | 0.103    | -0.051  |

Table 3: **Cross study concordance of GeneSurrounder results compared to differential expression analysis**: The first two columns $p_{GS}$ Cor. and $p_{DE}$ Cor. are the Spearman rank correlations between the results obtained from GeneSurrounder and differential expression analysis for each study pair. The third column is the ratio of the values in the first two columns. GeneSurrounder is more concordant than single gene association statistics.

| Ovarian Cancer Study Pair | $p_{GS}$ Cor. | $p_{DE}$ Cor. | Ratio |
|--------------------------|---------------|---------------|-------|
| GSE14764 - GSE17260      | 0.342         | 0.040         | 8.55  |
| GSE14764 - GSE9891       | 0.436         | 0.056         | 7.78  |
| GSE17260 - GSE9891       | 0.485         | 0.138         | 3.51  |
Table 4: “Disruptive” disease genes in high-grade ovarian cancer consistently found by GeneSurrounder: Listed are the union of the top 15 genes found in each study

| Gene     | GSE14764 | GSE17260 | GSE9891 |
|----------|----------|----------|---------|
|          | $-\log_{10} p_{GS}$ | Rank | $-\log_{10} p_{GS}$ | Rank | $-\log_{10} p_{GS}$ | Rank |
| CACNA1C  | 3.99     | 15       | 1.35    | 934    | 2.20    | 668   |
| CCL19    | 4.27     | 11       | 2.75    | 176    | 2.37    | 470   |
| CDC6     | 3.71     | 17       | 2.54    | 279    | 4.82    | 12    |
| CDC7     | 4.38     | 8        | 3.76    | 10     | 4.83    | 9     |
| CDC25A   | 2.58     | 188      | 2.99    | 79     | 4.83    | 1     |
| CDC45    | 4.27     | 12       | 3.74    | 12     | 4.82    | 15    |
| CNGB1    | 1.73     | 548      | 4.10    | 5      | 1.53    | 1132  |
| CTSG     | 2.55     | 198      | 4.04    | 7      | 1.68    | 1034  |
| DBF4     | 4.27     | 13       | 3.70    | 11     | 4.83    | 10    |
| FANCC    | 0.82     | 1599     | 4.04    | 8      | 3.39    | 81    |
| FANCE    | 2.27     | 320      | 4.54    | 2      | 2.58    | 283   |
| FANCI    | 1.03     | 1191     | 4.10    | 6      | 4.10    | 26    |
| FANCL    | 1.25     | 904      | 4.83    | 1      | 3.94    | 32    |
| MCM2     | 4.83     | 1        | 3.37    | 28     | 4.83    | 2     |
| MCM3     | 4.83     | 2        | 3.38    | 25     | 4.83    | 3     |
| MCM4     | 4.83     | 3        | 3.39    | 23     | 4.83    | 4     |
| MCM5     | 4.83     | 4        | 3.37    | 29     | 4.83    | 5     |
| MCM6     | 4.83     | 5        | 3.42    | 22     | 4.83    | 6     |
| ORC1     | 2.33     | 288      | 3.99    | 9      | 4.83    | 7     |
| ORC2L    | 4.54     | 6        | 4.27    | 4      | 2.15    | 738   |
| ORC4L    | 4.38     | 10       | 3.17    | 42     | 4.83    | 8     |
| ORC6L    | 4.38     | 9        | 3.69    | 15     | 4.83    | 11    |
| NFKB2    | 4.18     | 14       | 2.27    | 439    | 2.31    | 528   |
| RAD51C   | 0.90     | 1455     | 1.29    | 977    | 4.82    | 13    |
| RELB     | 4.54     | 7        | 2.58    | 264    | 2.20    | 675   |
| TRHR     | 2.48     | 233      | 4.38    | 3      | 2.44    | 405   |
| TTK      | 2.90     | 98       | 3.08    | 54     | 4.82    | 14    |
| USP1     | 1.80     | 514      | 3.71    | 13     | 3.74    | 43    |
| WNT5A    | 0.80     | 1638     | 3.69    | 14     | 1.71    | 1012  |
Figure 1: **Overview of GeneSurrounder algorithm.** The algorithm incorporates systems-level information, in the form of a network model of cellular interactions, with gene expression data to identify the genes that control disease-associated mechanisms. The algorithm then identifies “disruptive” genes by assessing the significance of the combined evidence that (1) a gene has a influence on others in the network and (2) that its influence is driving disease.
Figure 2: **Procedure for Sphere of Influence.** The Sphere of Influence computation tests if a putative driver gene is more correlated with its neighbors than a random sample of genes.

Figure 3: **Procedure for Decay of Differential Expression.** The Decay of Differential Expression computation tests if the discordance between differential expression and distance from the driver gene is greater with the phenotype labels we observe than with a random permutation of the sample labels.
Figure 4: **Illustration of Method.** Displayed are the results for the gene MCM2 when our algorithm was applied to Ovarian Cancer Study GSE14764. (a) shows $-\log_{10}(p_{\text{Sphere}})$ vs the Neighborhood Radius. (b) shows $-\log_{10}(p_{\text{Decay}})$ vs the Neighborhood Radius. (c) shows $-\log_{10}(p_{\text{Combined}})$ vs the Neighborhood Radius. (d) shows the Number of Assayed Genes vs the Neighborhood Radius. In the top three plots, the dashed and dotted lines correspond to a significance level of 0.05 and 0.01 respectively. In the bottom plot, the solid line corresponds to the total number of genes assayed and on the network.
Figure 5: **GeneSurrounder versus single–gene differential expression analysis.** Plotted are the $-\log_{10} p$ values obtained from GeneSurrounder ($-\log_{10}(p_{GS})$) versus a standard differential expression $t$-test ($-\log_{10}(p_{DE})$) for each gene in each study. The solid lines correspond to a significance level of 0.01 and the dashed line corresponds to a significance of 0.05. The top right quadrants contain significantly differentially expressed genes that are also identified as significant ”disruptive” genes by GeneSurrounder. In the upper left quadrants, GeneSurrounder identifies many significant genes that would not be detected as significant in a single–gene differential expression analysis. In the lower right quadrants, we find some genes that are significantly differentially expressed, yet *not* considered significant in the GeneSurrounder analysis, most likely due to their having low influence on the network.
References

[1] G Smyth. limma: Linear Models for Microarray Data. Bioinformatics and Computational Biology Solutions Using R and Bioconductor, (2005):397–420, 2005.

[2] T. Manoli, N. Gretz, H.J. Gröne, M. Kenzelmann, R. Eils, and B. Brors. Group testing for pathway analysis improves comparability of different microarray datasets. 22(20):2500, 2006.

[3] Purvesh Khatri, Marina Sirota, and Atul J Butte. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Computational Biology, 8(2):e1002375, 2012.

[4] Rosemary Braun and Sahil Shah. Network methods for pathway analysis of gene expression data. Computational and Structural Biotechnology Journal, In press (arXiv:1411.1993 [q-bio.QM]), 2014.

[5] Minoru Kanehisa, Michihiro Araki, Susumu Goto, Masahiro Hattori, Mika Hirakawa, Masumi Itoh, Toshiaki Katayama, Shuichi Kawashima, Shujiro Okuda, Toshiaki Tokimatsu, and Yoshihiro Yamanishi. KEGG for linking genomes to life and the environment. Nucleic Acids Research, 36(SUPPL. 1):480–484, 2008.

[6] Koyel Mitra, Anne-Ruxandra Carvunis, Sanath Kumar Ramesh, and Trey Ideker. Integrative approaches for finding modular structure in biological networks. Nature reviews. Genetics, 14(10):719–32, 2013.

[7] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael a Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America, 102(43):15545–50, October 2005.

[8] Adi Laurentiu Tarca, Sorin Draghici, Purvesh Khatri, Sonia S Hassan, Pooja Mittal, Jung-Sun Kim, Chong Jai Kim, Juan Pedro Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. Bioinformatics, 25(1):75–82, 1 January 2009.

[9] Zuguang Gu and Jin Wang. CePa: an R package for finding significant pathways weighted by multiple network centralities. Bioinformatics, 29(5):658–660, 1 March 2013.

[10] Zuguang Gu, Jialin Liu, Kunming Cao, Junfeng Zhang, and Jin Wang. Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. BMC Syst. Biol., 6:56, 6 June 2012.

[11] Trey Ideker, Owen Ozier, Benno Schwikowski, and AF Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. Bioinformatics, 18:233–240, 2002.

[12] Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. J. Comput. Biol., 18(3):507–522, March 2011.

[13] Biaobin Jiang and Michael Grishkov. Assessment of subnetwork detection methods for breast cancer. Cancer informatics, 13(Suppl 6):15, 2014.

[14] Benjamin Frederick Ganzfried, Markus Riester, Benjamin Haibe-Kains, Thomas Risch, Svitlana Tyekucheva, Ina Jazic, Xin Victoria Wang, Mahnaz Ahmadifar, Michael J. Birrer, Giovanni Parmigiani, Curtis Huttenhower, and Levi Waldron. curatedOvarianData: Clinically annotated data for the ovarian cancer transcriptome. Database, 2013:1–10, 2013.

[15] R A Fisher. Statistical methods for research workers. edinburgh: Oliver and Boyd, 1925. FisherStatistical Methods for Research Workers1925, 1932.

[16] Anjana A Kulkarni, Sarah R Kingsbury, Slavica Tuzdarova, Hye-Kyung Hong, Marco Loddo, Mohammed Rashid, Sara Rodriguez-Acebes, Andrew T Prevost, Jonathan A Ledermann, Kai Stoeber, and Gareth H Williams. Cdc7 kinase is a predictor of survival and a novel therapeutic target in epithelial ovarian carcinoma. Clin. Cancer Res., 15(7):2417–2425, 1 April 2009.

[17] Nathan J Bowen, L Deette Walker, Lilya V Matyunina, Sanjay Logani, Kimberly A Totten, Benedict B Benigno, and John F McDonald. Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells. BMC Med. Genomics, 2:71, 29 December 2009.
[18] Dorine Bonte, Charlotta Lindvall, Hongyu Liu, Karl Dykema, Kyle Purge, and Michael Weinreich. Cdc7-Dbf4 kinase overexpression in multiple cancers and tumor cell lines is correlated with p53 inactivation. *Neoplasia*, 10(9):920–931, September 2008.

[19] Adib Shafi, Michele Donato, and Sorin Draghici. A systems biology approach for the identification of significantly perturbed genes. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 423–432. ACM, 9 September 2015.

[20] Daniela Nitsch, Léon-Charles Tranchevent, Bernard Thienpont, Lieven Thorrez, Hilde Van Esch, Koenraad Devriendt, and Yves Moreau. Network analysis of differential expression for the identification of disease-causing genes. *PLoS One*, 4(5):e5526, 13 May 2009.

[21] Frederik Gwinner, Gwénola Boullday, Claire Vandiedonck, Minh Arnould, Cécile Cardoso, Iryna Nikolayeva, Oriol Guitart-Pla, Cécile V Denis, Olivier D Christophe, Johann Beghain, Elisabeth Tournier-Lasserve, and Benno Schwikowski. Network-based analysis of omics data: The LEAN method. *Bioinformatics*, 26 October 2016.