DISCOVERY OF FUNCTIONAL AND DISEASE PATHWAYS BY COMMUNITY DETECTION IN PROTEIN-PROTEIN INTERACTION NETWORKS

STEPHEN J. WILSON
Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: sw5@bcm.edu

ANGELA D. WILKINS
Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: aw11@bcm.edu

CHIH-HSU LIN
Graduate Program in Structural and Computational Biology and Molecular Biophysics, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: Chih-Hsu.Lin@bcm.edu

RHONALD C. LUA
Department of Molecular and Human Genetics, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: lua@bcm.edu

OLIVIER LICHTARGE
Departments of Molecular and Human Genetics, Structural and Computational Biology and Molecular Biophysics, Biochemistry and Molecular Biology, and Pharmacology, Baylor College of Medicine, One Baylor Plaza
Houston, Texas 77030, USA
Email: lichtarg@bcm.edu

Advances in cellular, molecular, and disease biology depend on the comprehensive characterization of gene interactions and pathways. Traditionally, these pathways are curated manually, limiting their efficient annotation and, potentially, reinforcing field-specific bias. Here, in order to test objective and automated identification of functionally cooperative genes, we compared a novel algorithm with three established methods to search for communities within gene interaction networks. Communities identified by the novel approach and by one of the established method overlapped significantly ($q < 0.1$) with control pathways. With respect to disease, these communities were biased to genes with pathogenic variants in ClinVar ($p << 0.01$), and often genes from the same community were co-expressed, including in breast cancers. The interesting subset of novel communities, defined by poor overlap to control pathways also contained co-expressed genes, consistent with a possible functional role. This work shows that community detection based on topological features of networks suggests new, biologically meaningful groupings of genes that, in turn, point to health and disease relevant hypotheses.

1. Introduction

How genes and proteins interact with each other is the basis of molecular biology and disease pathogenesis[1,2]. These functional interactions, which biologists place into pathways, have been characterized through hypothesis-driven experiments and then manually defined in the past[3,4]. This is necessarily knowledge intensive and painstaking, and it stands in sharp contrast to the massive amount of new gene interaction data from high-throughput experiments. Continued reliance on manual recognition of pathways may limit the overall capacity to characterize gene behavior, and potentially focus on already well-known sets of gene interactions. With at least 100,000 interactome hubs in humans, the number of potential interactions to annotate are in the billions[5]. Yet, the current estimate of interactions from the broadly used and expertly curated STRING database[6] that focus solely on proteins are in the millions. This large discrepancy suggests many unrecognized, or "dark," associations and pathways are simply missing.

In order to take a data-driven approach to annotate and detect novel biological pathways, clusters in biological networks were defined based on topological features to isolate functional and disease pathways[5,7]. One topological feature that has been extensively applied in social network analysis[8-10], but has not yet seen widespread use in biology, is community structure[11,12].

Communities are groups of nodes (i.e. proteins) that are more connected to each other than to anything else in a network[8,13]. Often these groups of nodes correspond to a common process, purpose, or function[5,9]. Therefore, it is reasonable to hypothesize that determining communities on biological networks may shed new light on groupings of genes with common biological function or features. Past efforts[13,14] were useful but did not comprehensively test various algorithms in functional and disease contexts. Given appropriate algorithms, community detection has the potential to automatically expand biological pathways, determine novel pathways, and perhaps even predict gene-disease associations.

This study sought to detect communities on a protein-protein interaction network and to evaluate their number and size against existing references. Several methods can evaluate performance in terms of the number and size of the overlap between communities and known control pathways. Moreover, beyond reference pathways, disease data can directly demonstrate the applicability of communities to formulate new and clinically relevant biomedical hypotheses.

2. Results

2.1. *Determining putative biological pathways*

In order to automatically determine putative biological pathways, several possible community detection methods exist. Clauset-Newman-Moore (CNM)[8] and Louvain[10] are well-established and extensively applied algorithms with more than 3000 citations each. BIGCLAM[15] is a more recent alternative that searches for densely overlapping, hierarchically nested communities in an orthogonal approach. Each of these approaches was tested on a STRING protein-protein interaction network[16], limited to high-quality direct biological associations. The communities that were obtained could then be compared to gold standard set of curated biological pathways, such as Reactome[17] and Canonical pathways from the GSEA tool[18], and, for disease pathways, DisGeNET[19].

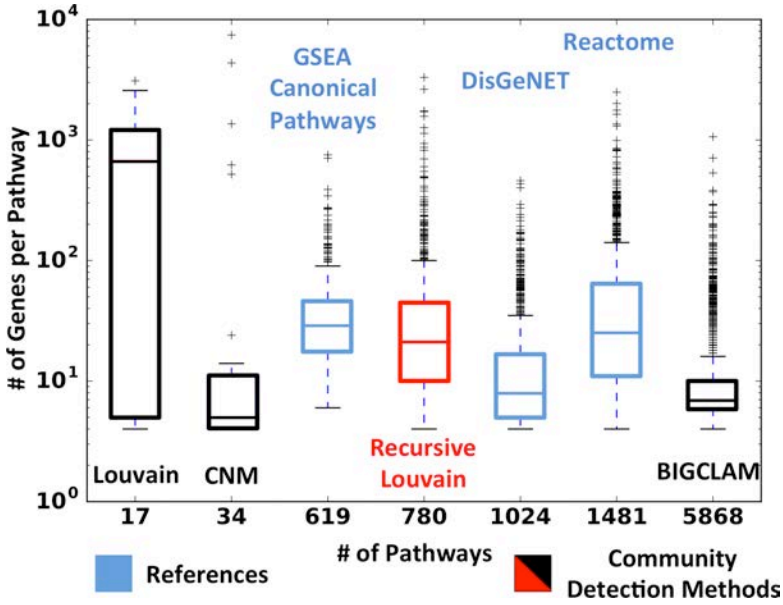

Figure 1: Community Algorithms Detect Variable Numbers and Sizes of Pathways. Recursive Louvain (Red) is a novel community detection method that detects a similar number of gene groups to the references with approximately the same number of genes per gene group given STRING 9.1 protein-protein interaction network.

A first assessment of performance was the granularity of the communities. That is, we compared the number of gene groups and the number of genes in each group in order to determine whether the communities resemble the references. CNM and Louvain community detection found an order of magnitude fewer groupings than the smallest reference set, and BIGCLAM detects five times more groups than the largest reference set (Figure 1). This is not surprising given that the methods were designed for social network analysis. Combined with different numbers of genes per group, these algorithms appear to poorly represent the reference pathways as defined by biologists.

To address this problem, we then introduced a novel community detection algorithm we called, Recursive Louvain (RL). RL applies the Louvain algorithm but iterates on the resulting communities so as to break them down stepwise into smaller and smaller groups until reaching a majority of right-sized communities, an idea that was in fact discussed in the original Louvain community detection paper[10]. In this way, RL generated communities that matched more closely the size of the control pathways (Figure 1, red).

2.2. *Assessing the biological relevance of communities*

Next, when comparing communities to the reference sets, careful consideration of what constitutes a pathway was necessary. First, we removed overly small reference pathways and communities (size ≤ 3 genes) to better focus on significant gene groupings. Additionally, pathways often share many genes, and in the extreme, they can share all but one gene. To avoid the over-counting of a pathway, or community, any with more than 90% of genes in common were combined. Finally, four different metrics were selected to gauge success. Jaccard similarity measures the similarity of a community to a reference by looking at the size of the intersection relative to the union of the genes; the modified Jaccard metric does not punish a community for being larger than the reference; the hypergeometric test measures the likelihood of getting an overlap between a reference and a community given all genes in a given community set; and the F_1 score measures the ability to recover an overlap (see methods for the mathematical details).

To test if communities represent biological information from functional and disease pathways, we compared each community to each reference pathway. This comparison was accomplished with the hypergeometric test, which allows a statistical probability and Benjamini-Hochberg False Discovery Rate (FDR) correction[20]. This correction is essential to account for

multiple testing. Encouragingly, many communities were significantly enriched (q-value ≤ 0.1) for a functional pathway (Reactome and Canonical Pathways), a disease pathway (DisGeNET), or a mixture of the two. Depending on the method, between 7-24% of communities were not enriched for any known pathway or disease and were regarded as novel. The exact breakdown of the community classification is shown in Figure 2A, and the majority of communities in BIGCLAM and RL are statistically overlapped with a function pathway and often with a disease pathway. Indeed, RL has the smallest fraction (7%) of novel communities, suggesting a higher positive predictive rate for the references. We noted that the number of genes in each community group generally increases from novel to mixed (Figure 2B). This could have a number of implications, including an observational annotation bias or a biological basis. These data show that community detection methods recover many commonly known functional and disease pathways but also discover new gene associations that possibly suggest novel pathways.

In order to assess the robustness of community detection we tested four metrics of

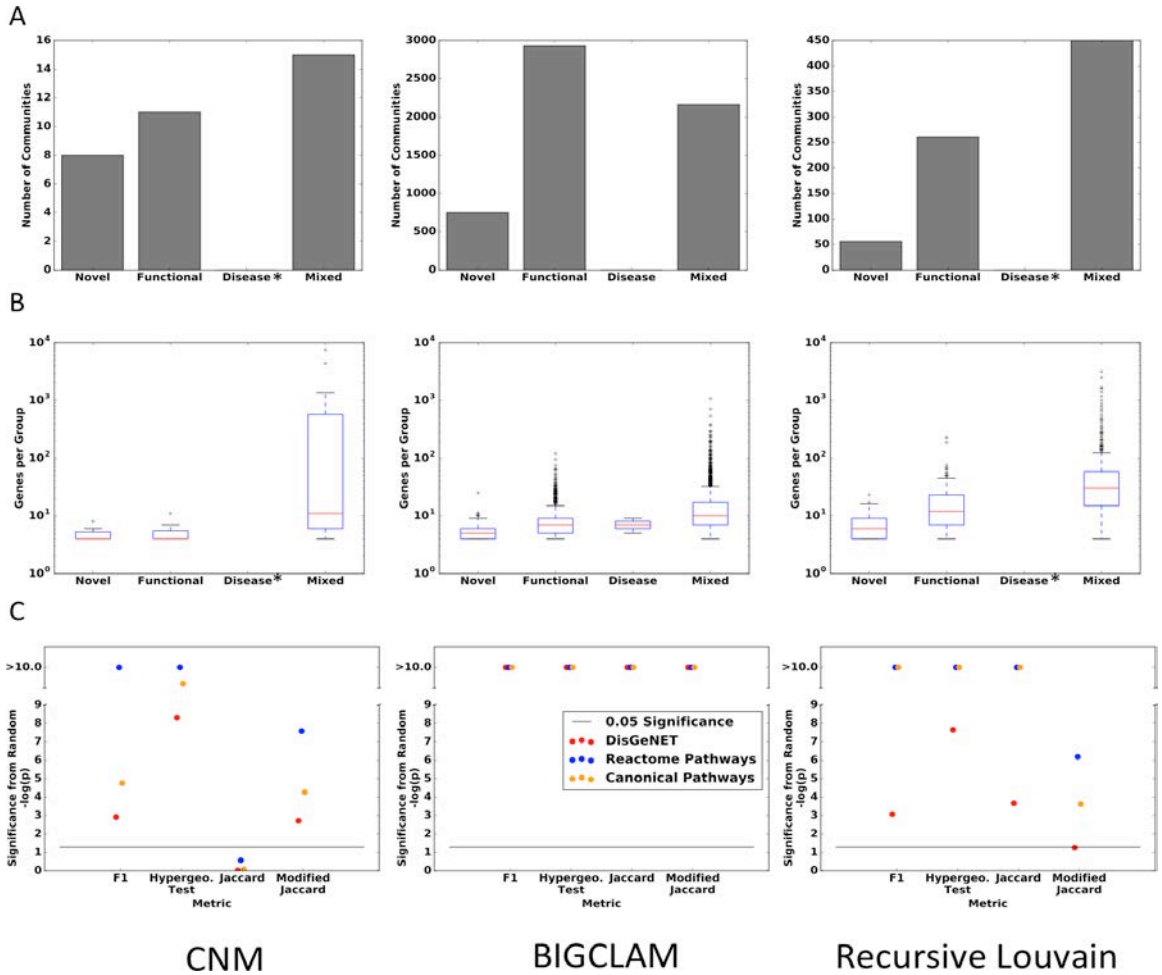

Figure 2: Communities Recapitulate Biological Knowledge. A) A hypergeometric test determines (q-value ≤ 0.1) whether a community is overlapped with a reference, and while many communities were overlapped with a disease or functional pathway, few or none (denoted by an *) were exclusively overlapped with only disease pathways. B) The number of genes in each group generally increases from a novel community to a community enriched for disease and functional pathways. C) All methods were non-randomly associated with every reference according to some metric, and many with p-values smaller than $10^{-10.}$

community overlap. Three random controls were generated to match the number and size of a set of communities and then scored against the references. Only the top score of a community or random against all pathways in a reference was kept. The distribution of random scores was then compared against the distribution of community scores using a Kolmogorov-Smirnov test. Due to poor performance and a lack of data (only 17 total communities), Louvain community detection (Supp. Figure 1-2) was assessed, but will not be shown because RL finds more total communities with overlap. As seen in Figure 2C, all three remaining methods were non-random by some metric; however, BIGCLAM and RL were significant on more metrics than CNM. In particular, BIGCLAM and RL appear highly significant in overlap with functional and disease pathways. RL has a higher percentage of communities that are enriched for both functional and disease pathways (Figure 2A), and this may suggest RL is better at recapitulating disease pathways. BIGCLAM has many more communities than the other methods (Figure 1A); this means we are more confident that BIGCLAM is performing different from random because we have more examples of overlap with the references. In contrast, Louvain community detection only found 17 communities, offering fewer opportunities to overlap with the references, and when we break those communities down further with RL, there is now more overlap with the references. These data show that BIGCLAM and RL recapitulate biological knowledge, while CNM appears to be less reliable.

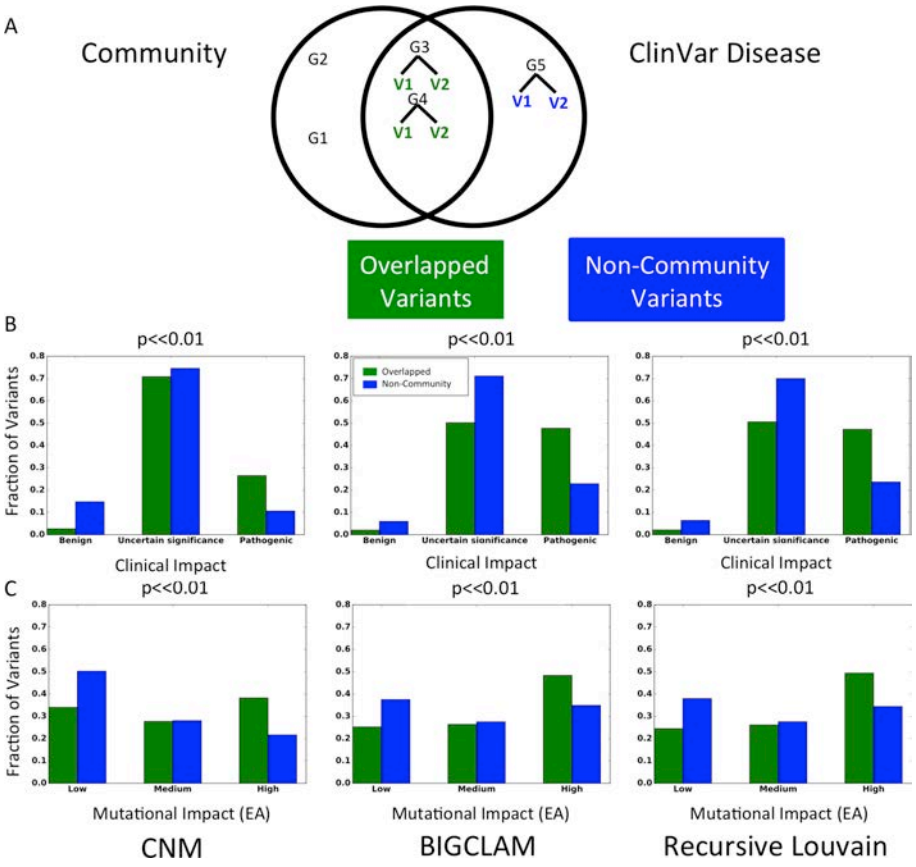

Figure 3: The Overlap between Communities and Diseases is Biased to Highly Pathogenic and Impactful Mutations. A) ClinVar groups variants into diseases. When a community and a disease from ClinVar both share genes, those genes possess a high B) clinical impact and C) mutation impact ($p<<0.01$) when compared to genes that are not found in communities. This implies that the communities are enriched towards variants that are pathogenic. Overlaps were only taken if the overlap was non-random ($q<0.05$).

2.3. *Clinical and disease relevance of communities*

A central question is whether these communities have real-world significance with respect to disease mechanisms. In order to address this question, communities were tested for overlap with diseases in the genetics database ClinVar. We hypothesized that communities represent units of biological function, and, if so, disrupting a gene that is part of a community would be more pathogenic than disrupting one that is outside of a community. Indeed, we find that mutations of disease genes that belong to communities have greater impact on the clinical phenotypes and on overall protein fitness

(Figure 3A). This was tested using the extensive annotations ClinVar[21] provides on the clinical impact of disease mutations. Specifically, for every community detection method, genes that fell inside communities showed are biased towards pathogenic variants (Chi-Square p-value << 0.01, Figure 3B). As an orthogonal control to test for bias in the impact of variations on disease genes, the Evolutionary Action (EA) provides an independent assessment of the deleterious impact of a mutation on protein fitness[22]. The same statistically significant trend emerges (Figure 3C, Chi-Square p << 0.01). These data show that mutations tend to have greater clinical and evolutionary deleterious impact if they affect genes that are part of communities.

To demonstrate a specific application of communities to disease pathways, we compared communities from BIGCLAM and RL, which outperformed CNM, against two diseases. These two diseases, Zellweger Syndrome (ZS) and Bardet-Biedl Syndrome (BBS) were both statistically associated with communities (p << 0.01). To associate diseases to communities, we used the disease-gene association information from two sources: (1) DisGeNet, a disease-gene association database integrating several public data resources and literature, and as shown in Figure 3, (2) ClinVar, a database providing the expert-asserted associations between genetic variants of genes and diseases. These disease-gene associations were used to calculate the statistical overlap between a disease and a community according to a hypergeometric distribution test of the overlap of genes, the unique genes of each, and all human genes. We hypothesized that when a community is statistically associated with a disease, any genes unique to the community are promising novel disease candidates. This hypothesis extends from a guilt-by-association assumption that has been successful in multiple systems[23,24]. As shown in Figure 4, when communities from multiple algorithms are compared to diseases, the overlaps possess high predictive power. For example, ZS is a peroxisomal biogenesis disorder characterized by severe hypotonia, epileptic seizures, and craniofacial abnormalities[25]. Because peroxisomal biogenesis depends highly on protein-protein interactions (PPIs), community detection on a PPI network reliably predicts and expands the disease definition. Indeed, using both community algorithms recovers all genes from DisGeNET and thirty additional genes are predicted. Of these thirty genes, fourteen are already annotated in the literature as being associated or causative of ZS. Two genes predicted to be associated with ZS are *ABCD1* and *ABCD2*, which are not known to be associated with ZS but transport very-long-chain fatty acids (*VLCFA*) across the peroxisome membrane

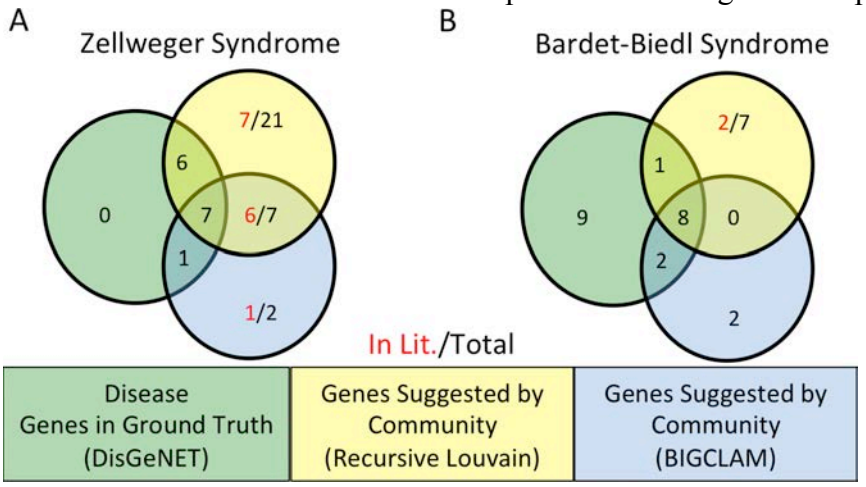

Figure 4: Communities Discover Novel Disease-Gene Associations. A) For Zellweger Syndrome, all known disease associated genes in DisGeNET are recovered between a community from Recursive Louvain and BIGCLAM. Fourteen out of thirty predictions possess some form of evidence in the literature. B) Bardet-Biedl Syndrome (BBS) possesses significant overlap with the ground truth but does not find all known genes. However, there is a high concordance of overlap between the methods and the ground truth, with 2/7 of the Recursive Louvain predictions with literature evidence.

and cause adrenoleukodystrophy, a related peroxisomal disorder[26].

Another example is BBS, a rare ciliopathy that affects multiple body systems, where over half of the known genes were recovered and nine genes were predicted. Of these nine genes, two genes are already known in the literature to be associated with BBS. BBS is characterized by obesity, polydactyly, hypogonadism, intellectual disability, and renal abnormalities[27]. The gene *FOPNL* is suggested by community analysis but possesses no literature evidence. Despite this, *FOPNL* is well known to be associated with the biogenesis of cilia and BBS causative mutations upset cilial function. Furthermore, *FOPNL* interacts with *PCM1*, a known BBS gene that is also suggested by community analysis[28]. For BBS, there is a lack of overlap between community predictions, which points to the fact that each method is dependent on different features and therefore provides unique insight. These data demonstrate that communities can be useful in predicting and expanding sets of genes related to diseases that depend on protein interactions.

We determined if novel communities that lacked overlap with functional and disease pathways are biologically relevant by analyzing the co-expression of community genes in breast invasive carcinoma (BRCA). BRCA was chosen as a test case because it has a large number of patients with whom to power a co-expression study, though other cancers will be investigated in the future. If the genes in a community are co-expressed together more than randomly selected genes within tumor tissue RNA sequencing data, then that community represents a biologically relevant disease module. To validate our co-expression analysis, we examined four Reactome pathways, which are related to breast cancer pathways (PI3K/AKT activation, Signaling to RAS, PI3K/AKT Signaling in Cancer, and Constitutive Signaling by AKT1 E17K in Cancer) and found they are significantly co-expressed/regulated in breast diseased tissues ($q<0.05$). For both BIGCLAM and RL, at least 30% of the communities were co-expressed more than random with a q-value < 0.1 (FDR corrected by Benjamini-Hochberg), and over 52 % of novel RL communities were co-expressed non-randomly (Figure 5). Moreover, CNM preformed weaker than RL and BIGCLAM in comparisons to references, but with co-expression, CNM showed no signal, suggesting that it may be a poor approach for biological analysis. Overall, the figure shows that all classes of communities, including novel communities, have co-expression in BRCA.

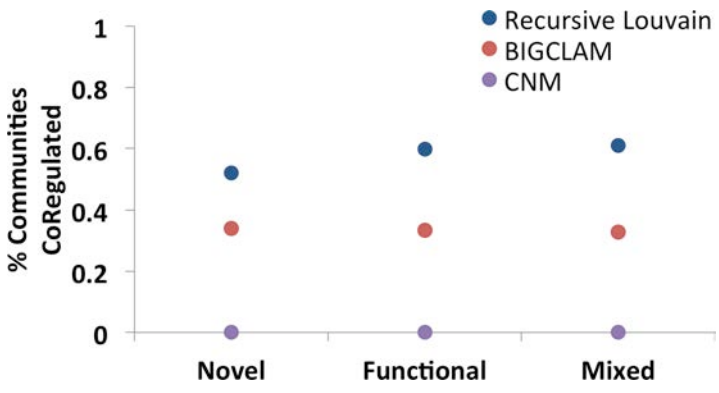

Figure 5: Communities are Significantly Perturbed in Cancer. In order to investigate whether novel communities had biological relevance, novel communities were investigated in the context of Breast Cancer (BRCA) co-expression data from TCGA. According to this analysis, the genes in novel communities are co-expressed to the same degree as communities with statistical overlap to functional and mixed pathways. This suggests that novel pathways represent a promising source of relevant biological knowledge.

As an application, one novel community (no. 657) detected by RL showed significant co-expression in BRCA (q = 0.00588) and has 14 gene members. Five members (*GPNTG*, *ECHS1*, *NACA*, *ABHD14B*, *NKX6-1*) were found to significantly coexist in the same subcellular location, extracellular vesicular exosome (GO:0070062; q = 0.01456; see Method). Furthermore, four members (*BTF3*, *GNPTG*, *CPEB2*, and *BICC1*) were found to be potentially co-regulated

by the same transcription factor, *DACH1*, in a triple-negative breast cancer cell line, MDA-MB-231, (q = 0.0077 in ChIP-Seq enrichment analysis; see Method). Although it has been shown that *DACH1* expression level can predict BRCA survival[29] and play roles in breast cancer metastasis[30], *DACH1* currently has no pathway annotation in KEGG and Reactome databases. Therefore, this community might be the pathway related to *DACH1*. These results showed that novel communities could be related not only by expression but also by subcellular location and transcription factors. These data show the potential of communities to expand our knowledge of biology and disease.

3. Discussion

Determining the relationships between genes is essential for molecular biology and medicine. These relationships often cluster together into functional and disease pathways, and the characterization of these pathways is necessary to improve disease classification, patient stratification and, ideally, personalized treatment[5]. Here, we investigated the automated discovery of pathways by comparing several community detection algorithms against known functional and disease pathways and leading us to a novel application of the well-known Louvain algorithm, which we call Recursive Louvain (RL).

First, the communities detected by both BIGCLAM and RL were associated non-randomly with all the control, reference pathways. This strongly supports the biological relevance of these communities. Second, these communities also show a bias towards genes that experience pathogenic and high-impact variants in ClinVar. And third, regardless of the enrichment to a particular reference, these communities are often statistically co-expressed in breast cancer, including those that are new, in the sense that they are not enriched for any known functional or disease pathway. Therefore, these novel communities of genes may point to currently unrecognized biological pathways. Finally, in at least several cases, communities appear to predict genes associated with diseases with high predictive power. In the case of Zellweger Syndrome, six out of seven of the highest confidence predictions were already found in the literature although they were missing from the reference. The data from these approaches therefore consistently show that communities are biologically relevant.

The breadth of information in the input network limits community analysis. With only direct protein-protein interaction information, protein associations via indirect biological mechanisms such as transcription regulation can be missed. Eventually, the addition of transcriptional, post-translational, and epigenetic associations should help better characterize biological processes and extend the ability of community detection to recognize a wider variety of pathways. This is important as we note that, so far, many diseases and pathways are not enriched for communities. Beyond the breadth of information, community detection is also limited by its quality. Low-confidence, spurious associations between proteins surely lead to incorrect memberships of proteins in pathways. Furthermore, the pathways found represent global averages of associations. The future addition of context-specific transcriptional networks, such as from ChIP-seq data in ENCODE[31], should help find context-specific communities relevant to individual tissues or disease states. Despite these limitations, this work reveals the potential of topological network analysis in the identification and expansion of biologically meaningful pathways and shows that diverse results can be achieved through careful algorithm choice.

4. Methods:

4.1. ***Collection of reference sets****:* Reactome was downloaded from http://www.reactome.org, and was filtered for all disease pathways by trimming the disease section of the hierarchy as well as filtering out any pathway with the following words: disorder, hiv, defect, cancer, mutant, host, disease, influenza, toxin, viral, carcinoma, deletions, deficiency, variant, or virus. Canonical Pathways from the GSEA tool were downloaded from http://software.broadinstitute.org/gsea/downloads.jsp. Both KEGG and Reactome pathways are included in the Canonical pathways. All Reactome pathways in this dataset were filtered out to eliminate redundancy and then KEGG pathways related to diseases were filtered out to eliminate overlap with disease pathways from DisGeNET. DisGeNET was downloaded from http://www.disgenet.org/web/DisGeNET/menu/downloads as the curated dataset.

4.2. ***Community detection:*** Louvain community detection was calculated with the python community detection, which can be downloaded at http://perso.crans.org/aynaud/communities/. This base module then was used to create RL. RL runs Louvain, then takes each community larger than ten genes and makes it a subgraph of the original network, then calls Louvain community detection again. It does this iteratively until all communities have been broke down to ten genes or less or a gene has been seen in more than three communities. CNM and BIGCLAM communities were detected using implementations in the SNAP software package[32]. All community detection algorithms were applied onto STRING 9.1 experimental network[16].

4.3. ***Comparison to reference sets:*** All groups of genes were filtered to exclude pathways that contained three or fewer genes. This eliminated pathways that could easily be randomly recapitulated and therefore skew results. Pathways often overlap with each other, with minor differences between them. To prevent over counting from this, pathways that are too similar were collapsed together. Given a set of reference gene groups $R_i \in R$ and a set of community gene groups $C_i \in C$, all gene groupings were collapsed if the Jaccard Similarity > 0.9, where:

$$Jaccard\ Similarity, J(C_i, R_i) = \frac{|C_i \cap R_i|}{|C_i \cup R_i|} \tag{1}$$

To collapse two gene groups, the union of the genes was taken. In addition to the Jaccard Similarity, we then adopted three mathematical measures to evaluate the community detection algorithms outputs against the references, including a Modified Jaccard Similarity, a Hypergeometric Distribution test, and a F_1 score.

$$Modified\ Jaccard\ Similarity,\ J_m(C_i, R_i) = \frac{|C_i \cap R_i|}{|R_i|} \tag{2}$$

$$Hypergeometric\ Test, P(X \geq |C_i \cap R_i|) = 1 - \sum_{j=0}^{|C_i \cap R_i|-1} \frac{\binom{|R_i|}{j}\binom{M-|R_i|}{|C_i|-j}}{\binom{M}{|C_i|}} \tag{3}$$

$$F_1\ Score = \frac{1}{2}(F_R + F_C) \tag{4}$$

$$F_R or\ F_C = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (5)$$

$$Precision = \frac{TP}{TP+FP}\ , Recall = \frac{TP}{TP+FN} \quad (6)$$

Where M=Number of genes in the original network and $F_R or\ F_C$ are the F_1 scores from the perspective of the reference or the community, respectively. TP represents the number of true positives; FP represents the number of false positives; FN represents the number of false negatives.

4.4 ***Comparison to ClinVar:*** In order to compare to ClinVar, we binned variants by clinical impact and Evolutionary Action, and the difference between each group of genes was assessed by a Chi-Square analysis. Only groups of genes from significant overlaps ($q < 0.1$ by hypergeometric analysis) between diseases and communities were assessed.

4.5. ***Generation and evaluation of random controls:*** Random controls were generated for each community set. For each community, a set of randomly generated genes were chosen from the protein interaction network such that the number of genes was identical to the number in the community. This was done three times in order to get a set of random communities that was then compared to the reference sets. The distribution of the random scores was compared against the distribution of the community scores using a Kolmogorov-Smirnov test. Each distribution was built with only the top score for a community or random against all pathways in a reference.

4.6. ***Co-expression analysis in tumor tissues using RNA-seq data:*** To determine if genes in a community have co-expression, RNA sequencing data version 2 of 1104 breast cancer tumor samples were downloaded from The Cancer Genome Atlas (TCGA) database (https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm) dated January, 2015. RNA-Seq by Expectation-Maximization (RSEM) normalized read counts (https://wiki.nci.nih.gov/display/TCGA/RNASeq) were used to represent mRNA expression level. The pairwise Spearman's rank correlation coefficients between the expression levels of pairs of genes in a community were computed. The distribution of absolute values of correlation coefficients was compared to the coefficient distribution of a random gene set, which is three times the size of a community, using a Kolmogorov-Smirnov test. All *p*-values were adjusted by Benjamini-Hochberg FDR correction[20]. A community was defined as co-expressed if the adjusted *p*-value is less than 0.1.

4.7. ***Gene Ontology and ChIP-Seq enrichment analysis:*** To understand the subcellular localization and potential upstream transcription factors of genes in a novel community, we analyzed the enrichment of Gene Ontology (GO) Cellular Component 2015 and ChIP Enrichment Analysis (ChEA) 2015 using Enrichr[33] (adjusted *p*-value < 0.1).

4.8. ***Computation:*** All calculations were done on an Ubuntu OS with 64 GB RAM and 4th Gen. Intel Core i7 3.7 GHz processor or equivalent machine.

4.9. ***Supplemental data*****:** Supplemental data can be seen at:
http://mammoth.bcm.tmc.edu/SupplementalPSB2016Data.pdf

5. Acknowledgements

The authors would like to acknowledge the kind support of Christie Buchovecky, Daniel Konecki, Teng-Kui Hsu, and Panos Katsonis for their discussions and general feedback of the work. Additionally, the authors would like to acknowledge funding by NLM training fellowship (Grant No. T15 LM007093) for SJW, as well as funding from DARPA (N66001-14-1-4027), National Science Foundation (NSF DBI-1356569, NSF DBI-0851393), and National Institutes of Health (NIH-GM079656, NIH-GM066099).

References

1. Pawson, T. & Linding, R. Network medicine. *FEBS Lett* **582**, 1266-1270, doi:10.1016/j.febslet.2008.02.011 (2008).
2. AlQuraishi, M., Koytiger, G., Jenney, A., MacBeath, G. & Sorger, P. K. A multiscale statistical mechanical framework integrates biophysical and genomic data to assemble cancer networks. *Nat Genet* **46**, 1363-1371, doi:10.1038/ng.3138 (2014).
3. Fabregat, A. *et al.* The Reactome pathway Knowledgebase. *Nucleic Acids Res* **44**, D481-487, doi:10.1093/nar/gkv1351 (2016).
4. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-462, doi:10.1093/nar/gkv1070 (2016).
5. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics* **12**, 56-68, doi:10.1038/nrg2918 (2011).
6. Szklarczyk, D. *et al.* STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**, D447-452, doi:10.1093/nar/gku1003 (2015).
7. Sharan, R., Ulitsky, I. & Shamir, R. Network-based prediction of protein function. *Mol Syst Biol* **3**, 88, doi:10.1038/msb4100129 (2007).
8. Clauset, A., Newman, M. E. & Moore, C. Finding community structure in very large networks. *Physical review E* **70**, 066111 (2004).
9. Yang, J. & Leskovec, J. Defining and evaluating network communities based on ground-truth. *Knowledge and Information Systems* **42**, 181-213 (2015).
10. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).
11. Ahn, Y. Y., Bagrow, J. P. & Lehmann, S. Link communities reveal multiscale complexity in networks. *Nature* **466**, 761-764, doi:10.1038/nature09182 (2010).
12. Palla, G., Derenyi, I., Farkas, I. & Vicsek, T. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* **435**, 814-818, doi:10.1038/nature03607 (2005).
13. Taylor, I. W. *et al.* Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nat Biotechnol* **27**, 199-204, doi:10.1038/nbt.1522 (2009).
14. Sah, P., Singh, L. O., Clauset, A. & Bansal, S. Exploring community structure in biological networks with random graphs. *BMC Bioinformatics* **15**, 220, doi:10.1186/1471-2105-15-220 (2014).
15. Yang, J. & Leskovec, J. in *Proceedings of the sixth ACM international conference on Web search and data mining.* 587-596 (ACM).

16. Franceschini, A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res* **41**, D808-815, doi:10.1093/nar/gks1094 (2013).
17. Croft, D. *et al.* The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-477, doi:10.1093/nar/gkt1102 (2014).
18. Subramanian, A. *et al.* 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**, 15545-15550, doi:10.1073/pnas.0506580102 (2005).
19. Pinero, J. *et al.* DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database : the journal of biological databases and curation* **2015**, bav028, doi:10.1093/database/bav028 (2015).
20. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300 (1995).
21. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862-868, doi:10.1093/nar/gkv1222 (2016).
22. Katsonis, P. & Lichtarge, O. A formal perturbation equation between genotype and phenotype determines the Evolutionary Action of protein-coding variations on fitness. *Genome research* **24**, 2050-2058, doi:10.1101/gr.176214.114 (2014).
23. Lee, I. *et al.* A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat Genet* **40**, 181-188, doi:10.1038/ng.2007.70 (2008).
24. McGary, K. L., Lee, I. & Marcotte, E. M. Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes. *Genome Biol* **8**, R258, doi:10.1186/gb-2007-8-12-r258 (2007).
25. Klouwer, F. C. *et al.* Zellweger spectrum disorders: clinical overview and management approach. *Orphanet J Rare Dis* **10**, 151, doi:10.1186/s13023-015-0368-9 (2015).
26. Burtman, E. & Regelmann, M. O. Endocrine Dysfunction in X-Linked Adrenoleukodystrophy. *Endocrinol Metab Clin North Am* **45**, 295-309, doi:10.1016/j.ecl.2016.01.003 (2016).
27. Khan, S. A. *et al.* Genetics of human Bardet-Biedl syndrome, an updates. *Clin Genet* **90**, 3-15, doi:10.1111/cge.12737 (2016).
28. Sedjai, F. *et al.* Control of ciliogenesis by FOR20, a novel centrosome and pericentriolar satellite protein. *J Cell Sci* **123**, 2391-2401, doi:10.1242/jcs.065045 (2010).
29. Wu, K. *et al.* DACH1 is a cell fate determination factor that inhibits cyclin D1 and breast tumor growth. *Mol Cell Biol* **26**, 7116-7129, doi:10.1128/MCB.00268-06 (2006).
30. Zhao, F. *et al.* DACH1 inhibits SNAI1-mediated epithelial-mesenchymal transition and represses breast carcinoma metastasis. *Oncogenesis* **4**, e143, doi:10.1038/oncsis.2015.3 (2015).
31. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100, doi:10.1038/nature11245 (2012).
32. Leskovec, J. & Sosič, R. SNAP: A General-Purpose Network Analysis and Graph-Mining Library. *ACM Transactions on Intelligent Systems and Technology (TIST)* **8**, 1 (2016).
33. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res* **44**, W90-97, doi:10.1093/nar/gkw377 (2016).