KNOWLEDGE-ASSISTED APPROACH TO IDENTIFY PATHWAYS WITH DIFFERENTIAL DEPENDENCIES*

GIL SPEYER, JEFF KIEFER

Integrated Cancer Genomics Division, The Translational Genomics Research Institute
Phoenix, AZ 85004, U.S.A.
Email: gspeyer@tgen.org, jkiefer@tgen.org

HARSHIL DHRUV, MICHAEL BERENS

Cancer Cell Biology Division, The Translational Genomics Research Institute
Phoenix, AZ 85004, U.S.A.
Email: hdhruv@tgen.org, mberens@tgen.org

SEUNGCHAN KIM

Integrated Cancer Genomics Division, The Translational Genomics Research Institute
Phoenix, AZ 85004, U.S.A.
Email: skim@tgen.org

We have previously developed a statistical method to identify gene sets enriched with condition-specific genetic dependencies. The method constructs gene dependency networks from bootstrapped samples in one condition and computes the divergence between distributions of network likelihood scores from different conditions. It was shown to be capable of sensitive and specific identification of pathways with phenotype-specific dysregulation, i.e., rewiring of dependencies between genes in different conditions. We now present an extension of the method by incorporating prior knowledge into the inference of networks. The degree of prior knowledge incorporation has substantial effect on the sensitivity of the method, as the data is the source of condition specificity while prior knowledge incorporation can provide additional support for dependencies that are only partially supported by the data. Use of prior knowledge also significantly improved the interpretability of the results. Further analysis of topological characteristics of gene differential dependency networks provides a new approach to identify genes that could play important roles in biological signaling in a specific condition, hence, promising targets customized to a specific condition. Through analysis of TCGA glioblastoma multiforme data, we demonstrate the method can identify not only potentially promising targets but also underlying biology for new targets.

* This work was supported in part by the National Cancer Institute, National Institutes of Health [1U01CA168397].

1. Introduction

1.1. *Gene set analysis, DDN and EDDY*

Identification of biological features underlying disease phenotypes or conditions (e.g. differentially expressed or mutated genes) is critical in identifying therapeutic targets. As specific pathways are capable of complex rewiring between conditions, methods such as Gene Set Enrichment Analysis (GSEA) (1) and network-based analyses (2-4) have become increasingly attractive for extraction of such biological features from genomic data. One can use known genetic interactions as a ground truth network and overlay genomic data from different conditions to statistically evaluate regions with differential activities (5) or condition-specific sub-networks (6-8). Differential Dependency † Network (DDN) approaches are able to identify individual differential dependencies (9-13) or condition-specific sub-networks from genome-wide dependency networks such as a protein-protein interaction networks. Differential co-expression analysis methods (14), such as Gene Set Co-expression Analysis (GSCA), test gene sets for differential dependencies, but they are often overly sensitive to minor correlation changes and produce biased results with respect to the size of gene sets (15).

In our previous work, we have developed a novel, network-based computational method that overcomes the limitations of other network-based approaches (15). This novel computational approach – *EDDY: Evaluation of Differential DependencY* – combines GSEA's gene-set-assisted advantages with the robustness of assessment of differential network dependency. It interrogates gene sets (pathways) in a database to test if dependencies across genes are significantly rewired between conditions (see Fig. 1). It was shown to be capable of sensitive and specific identification of pathways with phenotype-specific dysregulation, i.e. *rewiring of dependencies between genes in different conditions*, with its robust network inference and low false discovery rate (15).

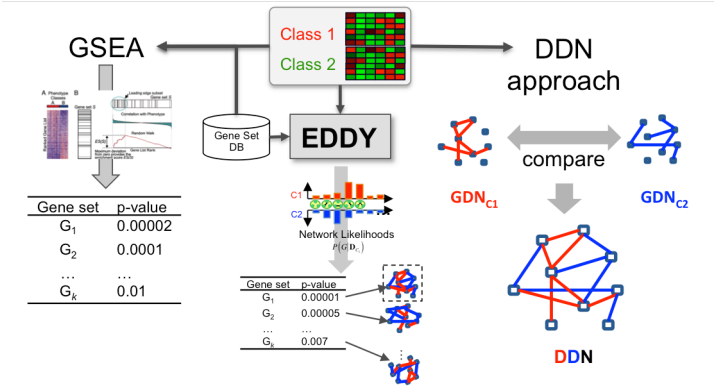

Figure 1. Advantages of EDDY compared to other tools

In this paper, we present a method to integrate known biological interactions to improve the performance of network inference and to enable better interpretation of inferred DDNs. The effect of the degree of prior knowledge integration on inferred DDNs is also analyzed. Finally, we describe the application of prior-knowledge assisted EDDY to glioblastoma (GB) gene expression downloaded from the Cancer Genome Atlas (TCGA).

† In this manuscript, we use 'dependency' to denote statistical dependencies derived from data such as co-expression, or conditional dependencies, and 'interaction' to denote *known* direct or indirect relationships between genes.

2. Methods

From two sets of samples representing different conditions, EDDY computes the discrepancy of gene dependency in a specific gene set by contrasting the two resulting probability distributions of candidate network structures (based on a likelihood of each network), constructed via a resampling approach, and evaluates its statistical significance to determine if the network structures are rewired between the conditions.

2.1. *EDDY: Evaluation of differential dependency*

Let a set of variables $\boldsymbol{G} = \{g_1, g_2, \ldots\}$ (each variable corresponds to a gene) denote the activity levels of the genes. For $\boldsymbol{G}$, there are N possible gene dependency network (GDN) structures $d_1, d_2, \ldots, d_N$ for the variables. Let a discrete random variable D take on $d_1, d_2, \ldots, d_N$ as its discrete values, then the posterior probability distribution $\Pr(D|\boldsymbol{S}_C)$ for a data $\boldsymbol{S}_C$ of a given condition C can represent the probability distribution of dependency network structures for $\boldsymbol{G}$ in the condition C. When two data sets, $\boldsymbol{S}_{C_1}$ and $\boldsymbol{S}_{C_2}$, are given for two different conditions C_1 and C_2, the divergence between the two corresponding probability distributions $\Pr(D|\boldsymbol{S}_{C_1})$ and $\Pr(D|\boldsymbol{S}_{C_2})$ is computed as a measure of difference between the conditions. The divergence between the conditions C_1 and C_2 is measured using the Jensen-Shannon (JS) divergence, an information-based metric to measure the similarity between two probability distributions (16) and the statistical significance of the divergence is computed using a permutation approach. This approach is a generalization of comparing the best networks from different conditions by considering many possible networks and their likelihoods instead of comparing the single best networks. The benefit of this generalization is a more reliable measure of discrepancy (15), especially when data is limited. Thus, there is a high chance of finding many local optima for the best network. By considering many probable dependency networks instead of one local optimal network, our approach can represent a more complete picture of dependencies at the cost of additional computation. EDDY then iterates through all gene sets in a database, for example, MSigDB (http://www.broadinstitute.org/gsea/msigdb/) to identify the dysregulated pathways.

2.2. *Inference of gene dependency network supported by known interactions*

To reduce computational complexity, EDDY uses a heuristic method that proposes probable dependency structures by independently evaluating each dependency between two variables. Specifically, χ^2-test is applied to test the independence between every pair of two variables g_i and g_j ($\in \boldsymbol{G}$), obtaining the resultant p-value p_{ij} ($=p_{ji}$). An edge e_{ij} between g_i and g_j is included when

$$\Pr(i; j|\boldsymbol{S}_C) = \left(1 - p_{ij}\right)^{\lambda} > \theta \qquad (1)$$

where $\lambda \geq 1$ and a user-specified parameter θ together control sensitivity of dependency discovery. We integrate known interactions retrieved from pathway databases to support dependency discovery. Formally, let $w_P \in [0, 1]$ denote a prior weight to control the level of prior knowledge to be incorporated into the inference of GDN and $E_P(i; j)$ be a binary-valued variable indicating the existence of known interaction between g_i and g_j. Known interactions can be retrieved from a pathway database such as Pathway Commons 2. Edge-specific threshold is given,

$$\theta_P(i; j) \leftarrow \theta \cdot [1 - w_P \cdot E_P(i; j)]. \qquad (2)$$

Prior weight (w_P) can be varied between 0 and 1, where $w_P = 0$ specifies no influence of the known gene interactions in GDN inference and all edges in inferred GDN requires full support from the data $\theta_P(i;j) = \theta$, and $w_P = 1$ makes inferred GDN include all the known interactions unconditionally, $\theta_P(i;j) = 0$. When $w_P = 0.5$, edges with half the support from the data will be included in the network. Edges are included in a network if they satisfy:

$$\Pr(i;j|\boldsymbol{S}_C) > \theta_P(i;j). \quad (3)$$

Since information on the condition-specificity of known interaction is generally not available, incorporating known interactions into GDN inference could potentially decrease the divergence between GDNs, hence, the sensitivity of the EDDY algorithm to detect pathways with condition-specificity. The specific effect of prior weight (w_P) on the sensitivity of EDDY will be discussed in the Results section.

Considerations: As opposed to data-derived edges, prior edges can have a direction, indicating, for example, the influence of one gene on another. While it is straightforward to incorporate the direction of an edge into EDDY, this may conflict with the acyclic requirement of Bayesian networks. For the computations in this work, directionality was determined not to create cycles. In addition, prior edge encompasses many types of interactions such as catalysis or phosphorylation. It also may describe various degrees of influence from explicitly controlling a state change to simply being a neighbor gene. For the work described here, we excluded these so-called "neighbor" interactions. In future work, we may examine a nuanced means of weighting other types of interactions.

2.3. *Estimating divergence between two conditions-specific probability distributions of GDNs*

The empirical estimate of the probability distribution, $\Pr(D|\boldsymbol{S}_C)$, is yielded from bootstrapping samples and the construction of GDNs as described above. Once the probability distribution of dependency network structures $\Pr(D|\boldsymbol{S}_{C_1})$ and $\Pr(D|\boldsymbol{S}_{C_2})$ are computed, the divergence between the conditions C_1 and C_2 is measured using the Jensen-Shannon (JS) divergence and the statistical significance is estimated using a permutation test. See (15) for more detail, and the overall workflow is shown in Fig. 2.

2.4. *Topological analysis of Differential Dependency Network (DDN)*

GDNs constructed for condition C_1 and C_2 are summarized into differential dependency networks (DDNs) where each edge is annotated as C1-specific, C2-specific, or common. While these condition-specific dependencies can be used to identify potential

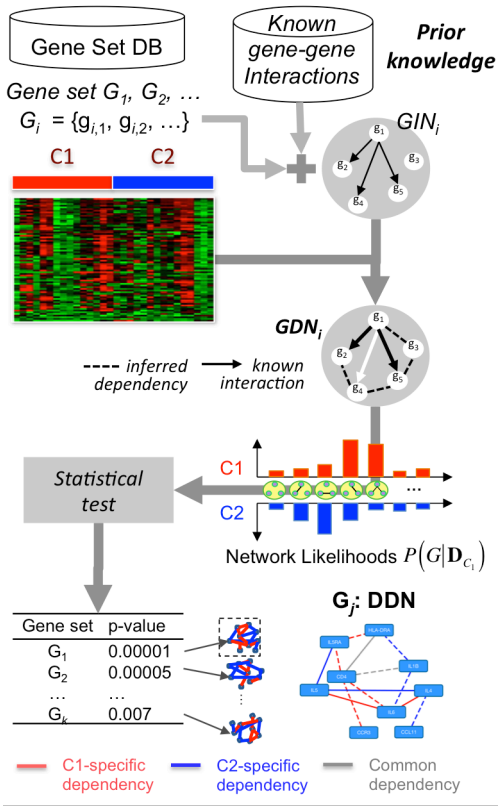

Figure 2. Workflow of knowledge-assisted EDDY

targets, the DDN often comprises hundreds of edges, rendering the prioritization of those dependencies non-trivial. We utilize the topological analysis of EDDY-derived DDNs to discern biologically important signaling nodes. These nodes could play important roles in biological signaling, hence, promising targets. For each node *i*, we will compute the normalized betweenness centrality metrics, $g(i|D_{C_1})$ and $g(i|D_{C_2})$ for GDNs, D_{C_1} and D_{C_2}, respectively (17). The regularized difference

$$\delta_{bw}(i|C_1, C_2) = \frac{g(i|D_{C_1}) - g(i|D_{C_2})}{g(i|D_{C_1}) + g(i|D_{C_2}) + \eta^2} \quad (4)$$

where η is a regularization parameter, is then used to assist in prioritization of genes.

2.5. *Comparison to Knowledge-fused Differential Dependency Network (KDDN)*

The KDDN (Knowledge-fused Differential Dependency Network) model (18; 19) extends the DDN method by incorporating prior knowledge into its regularized linear regression problem with sparse constraints, where the level of prior knowledge, w_P, is a parameter taking value in [0, 1] to adjust the degree of prior-knowledge integration into the determination of differential dependency. We compare the results of knowledge-assisted EDDY against KDDN's results. KDDN does not aggregate differential dependencies of genes in a gene set and assign a score to a gene set as EDDY does, but focuses on individual differential dependencies. Hence, we focus on those pathways enriched with differential dependencies, identified by EDDY, and compare corresponding differential dependency networks between two methods.

3. Results

3.1. *Data, Gene Sets and Analysis*

We used the gene expression data of 202 glioblastoma multiforme (GBM) samples assigned with GB subtype from TCGA to identify pathways enriched with differential dependency between mesenchymal (58 samples) and non-mesenchymal samples, and between proneural (57 samples) and non-proneural samples. The gene expression data were log-transformed, standardized, and quantized prior to EDDY analysis. The gene sets queried for the analysis were 472 gene sets in REACTOME category of MSigDB. We then mined known interactions from Pathway Commons 2 (http://www.pathwaycommons.org) and matched these to all pairings in the REACTOME gene sets for prior knowledge incorporation. To investigate the effect of the degree of prior knowledge in identifying condition-specific dependencies, the prior weights w_P = 0, 0.5, and 1 were used. $w_P = 0$ specifies no influence of the known gene interactions in GDN inference and all edges in inferred GDN requires full support from the data, and $w_P = 1$ makes inferred GDN include all the known interactions unconditionally. When $w_P = 0.5$, dependencies with known interactions are added with half the support from the data.

3.2. *Pathways identified by knowledge-assisted EDDY*

Across three different prior weights (w_P = 0, 0.5, and 1.0), EDDY identified 57 pathways with statistically significant divergence between mesenchymal (MES) and non-mesenchymal for at least one of the weights, and 75 pathways between proneural (PN) and non-proneural. Table 1 presents a subset (24 pathways) of 57 mesenchymal-specific pathways, and Table 2 a subset (38

pathways) of proneural-specific 75 pathways, based on their biological interest (bold-faced) or p-value (w_P = 0.5) < 0.05. For each pathway, we include the number of genes in the pathway, p-values, P_D (the proportion of newly discovered dependencies, E_D, compared to the total number of edges in GDN, E_D+E_P) and P_C (the proportion of condition-specific dependencies, E_C, compared to total edges, E_C+E_S), for different prior weights. As w_P increases, more known interactions are added to GDN without condition-specificity, and this has three possible effects. First, condition-specific edges with weak support from data can gain support from the prior weighting, thereby increasing P_C while reducing P_D. Second, condition-specific edges with prior support can lose specificity and hence, result in reduced P_C. Finally, the loss of condition-specific edges can reduce the diversity of networks in the score distribution, having the indirect effect of increasing the influence of the surviving condition-specific edges on the divergence calculation. Indeed, we observe a consistent decrease in the number of networks in the distribution as we increase prior weight. As a result of these competing effects, p-value does not correlate with prior weight, even when examined over the finer variation of 0.1 (data not shown). However, we did note that the number of pathways with statistically significant divergence tends to decrease with prior weight – 28, 20 and 16 pathways with statistically significant divergence between mesenchymal and non-mesenchymal, and 39, 36 and 28 pathways between proneural and non-proneural, as the prior weight increases from 0 to 0.5 to 1.0.

Table 1: A subset of the REACTOME pathways with significant differential dependency between GB mesenchymal and non-mesenchymal. P_D gives the proportion of newly discovered dependencies over the total number of edges in GDN and P_C the proportion of condition-specific dependencies over total number of edges. Systematic ID from MSigDB is used instead of full pathway for shorten description. Mapping from Systematic IDs for bold-faced pathways are provided in Table 3 and Table 4, and in Appendix at the end for the rest of pathways.

Systematic	#	p-value			$P_D = E_D/(E_D+E_P)$			$P_C = E_C/(E_C+E_S)$		
ID	genes	w_p=0	w_p=0.5	w_p=1	w_p=0	w_p=0.5	w_p=1	w_p=0	w_p=0.5	w_p=1
M760	27	0.0165	0.1314	0.2416	0.37			0.72		
M5113	29	0.1839	0.0173	0.4192		0.47			0.59	
M13748	34	0.1406	0.0299	0.0049		0.51	0.45		0.66	0.34
M9271	33	0.0122	0.0304	0.2399	0.77	0.66		0.75	0.68	
M506	23	0.0223	0.0478	0.1954	0.20	0.13		0.81	0.59	
M17157	19	0.0084	0.1605	0.6331	0.51			0.77		
M764	21	0.0019	0.1777	0.3609	0.73			0.83		
M571	38	0.6392	0.2754	0.0305			0.58			0.49
M9694	31	0.7833	0.0026	0.0705		0.04			0.35	
M1051	16	0.2921	0.0035			0.33			0.57	
M875	41	0.2310	0.0053	0.9018		0.58			0.76	
M612	23	0.3943	0.0104	0.8191		0.30			0.59	
M552	14	0.1828	0.0111	0.6727		0.19			0.58	
M3634	13	0.0091	0.0191		0.50	0.39		0.86	0.53	
M1062	21	0.1057	0.0222	0.1714		0.11			0.36	
M932	19	0.1187	0.0266	0.0606		0.64			0.79	
M16702	19	0.7982	0.0292	0.6791		0.39			0.61	
M1016	14	0.3862	0.0348	0.0561		0.47			0.66	
M1662	23	0.2844	0.0354	0.2397		0.33			0.64	
M6034	12	0.0568	0.0391	0.1070		0.92			0.64	
M17787	18	0.2575	0.0426	0.7349		0.69			0.33	
M7169	39	0.0082	0.0427	0.1184	0.85	0.81		0.80	0.76	

M901	35	0.0136	0.0427	0.0933	0.37	0.29		0.72	0.56	
M10122	13	0.3501	0.0433	0.6130		0.05			0.47	

Table 2: A subset of the REACTOME pathways with significant differential dependency between GB proneural and non-proneural.

Systematic	#	p-value			$P_D = E_D/(E_D+E_P)$			$P_C = E_C/(E_C+E_S)$		
ID	genes	$w_P=0$	$w_P=0.5$	$w_P=1$	$w_P=0$	$w_P=0.5$	$w_P=1$	$w_p=0$	$w_p=0.5$	$w_p=1$
M647	16	0.0020	0.0017	0.0014	0.89	0.83	0.78	0.93	0.94	0.72
M530	37	0.0648	0.0022	0.4847		0.25			0.68	
M1092	14	0.0154	0.0071	0.0072	0.87	0.79	0.78	0.90	0.84	0.71
M549	12	0.0335	0.0114	0.8563	0.25	0.16		0.82	0.65	
M1040	19	0.0141	0.0151	0.0463	0.59	0.52	0.51	0.51	0.43	0.23
M13408	21	0.1654	0.0202	0.0242		0.43	0.40		0.65	0.33
M714	38	0.0112	0.1503	0.5874	0.56			0.75		
M570	44	0.0440	0.2321	0.5892	0.56			0.78		
M947	25	0.0045	0.0000		0.11	0.07		0.87	0.68	
M9450	12	0.3631	0.0007			0.39			0.68	
M860	28	0.1070	0.0011	0.0704		0.20			0.68	
M12967	35	0.0534	0.0013	0.0395		0.09	0.07		0.58	0.06
M936	30	0.0050	0.0020	0.0684	0.67	0.48		0.86	0.73	
M15243	10	0.0559	0.0029			0.00			0.58	
M1075	31	0.0135	0.0040	0.1367	0.39	0.29		0.88	0.74	
M846	36	0.2413	0.0052	0.5402		0.22			0.69	
M1662	23	0.0026	0.0059	0.1335	0.48	0.36		0.86	0.73	
M801	11	0.0274	0.0061	0.8040	0.50	0.38		0.75	0.58	
M899	39	0.1676	0.0073	0.1689		0.48			0.76	
M769	10	0.1899	0.0103	0.7851		0.43			0.93	
M13115	27	0.0144	0.0122	0.2782	0.03	0.02		0.77	0.64	
M12627	11	0.0001	0.0139		0.00	0.00		0.86	0.72	
M564	10	0.1861	0.0152	0.7291		0.19			0.48	
M10272	11	0.0758	0.0168	0.0001		0.54	0.50		0.72	0.40
M11184	15	0.0242	0.0180	0.0070	0.88	0.86	0.85	0.75	0.69	0.64
M719	15	0.1317	0.0190	0.1944		0.06			0.71	
M794	13	0.0326	0.0215	0.3349	0.61	0.49		0.82	0.69	
M1014	11	0.3598	0.0232			0.03			0.63	
M907	11	0.0022	0.0273	0.7901	0.63	0.52		0.68	0.65	
M837	27	0.4998	0.0273	0.4145		0.39			0.74	
M918	13	0.0023	0.0285	0.7926	0.63	0.52		0.68	0.65	
M704	44	0.1173	0.0287	0.2284		0.21			0.66	
M1016	14	0.1716	0.0359	0.2208		0.35			0.76	
M3661	22	0.0774	0.0416	0.0697		0.35			0.73	
M15195	30	0.0953	0.0432	0.0659		0.42			0.70	
M661	30	0.2166	0.0448	0.4245		0.28			0.65	
M583	18	0.0162	0.0453	0.1178	0.59	0.43		0.81	0.65	
M1825	11	0.0229	0.0488	0.0961	0.50	0.37		0.93	0.89	

3.3. *Biological Significance of Selected Signaling Pathways Identified by EDDY*

3.3.1. *Condition-specificity of Integrin αIIb β3 signaling in mesenchymal GB*

EDDY analysis of mesenchymal vs non-mesenchymal GB show significantly different (p = 0.0165 at w_P = 0.5) dependency network for INTEGRIN_ALPHAIIB_BETA3_SIGNALING (**M760**; http://bit.ly/1Dlgidx). This pathway is representative of biological mechanisms of adhesion in platelets, but there are proteins that participate in other signaling process in a diverse array of tissues and diseases. The class dependent DDNs show interesting differences in the state of this pathway's genes in mesenchymal vs. non-mesenchymal GB. DDN and GDNs in Figure 3 show that mesenchymal GB loses dependency on the cell surface integrins ITGA2B (betweenness normalized difference, δ_{bw}=-0.83‡, rank, $R_{\delta_{bw}}$=2) and ITGB3 (δ_{bw}=-0.65, $R_{\delta_{bw}}$=7). Activation of ITGA2B/ITGB3-RAP1A-PTK2 signaling axis induces glioma cell proliferation (20). There is also a shift in the dependencies around SRC kinases between mesenchymal and non-mesenchymal GB samples with no SRC dependency evidence in mesenchymal samples but with new dependencies developed for Csk (δ_{bw}=0.12), also a member of Src-family kinase. In previous work, it is also demonstrated that Src family kinases plays very important role in migration and invasion cancer cells (21). Lastly, there is dependency shift in intracellular signaling effectors for integrins in the mesenchymal samples as evidenced by the δ_{bw} of PTPN1 (δ_{bw}=0.84, $R_{\delta_{bw}}$=1), APBB1IP (δ_{bw}=0.70, $R_{\delta_{bw}}$=6), SYK (δ_{bw}=0.43, $R_{\delta_{bw}}$=11), RAP1B (δ_{bw}=0.49, $R_{\delta_{bw}}$=9). These molecules have known roles in immunologic cell function, particularly cells of the monocytic origin (22-25). Mesenchymal GB samples have an appreciable amount of microglial (brain resident monocytic cells) cell infiltration that can be detected by RNA expression data (26), and it is interesting that EDDY appears to be detecting differential dependencies in molecules important for microglial function. In summary, this DDN demonstrates a differential wiring of ITGA2B/ITGB3 signaling network in mesenchymal vs non-mesenchymal GB. Functional validation of such differential wiring could help identifying novel nodes of vulnerability for treatment of subtype specific GB.

3.3.2. *Condition-specificity of PI3K events in ERBB2 signaling in proneural GB*

Another example of differential network dependency is illustrated in the analysis of proneural vs. non-proneural samples of GB. An example significant dependency network (p = 0.044 at w_P = 0)

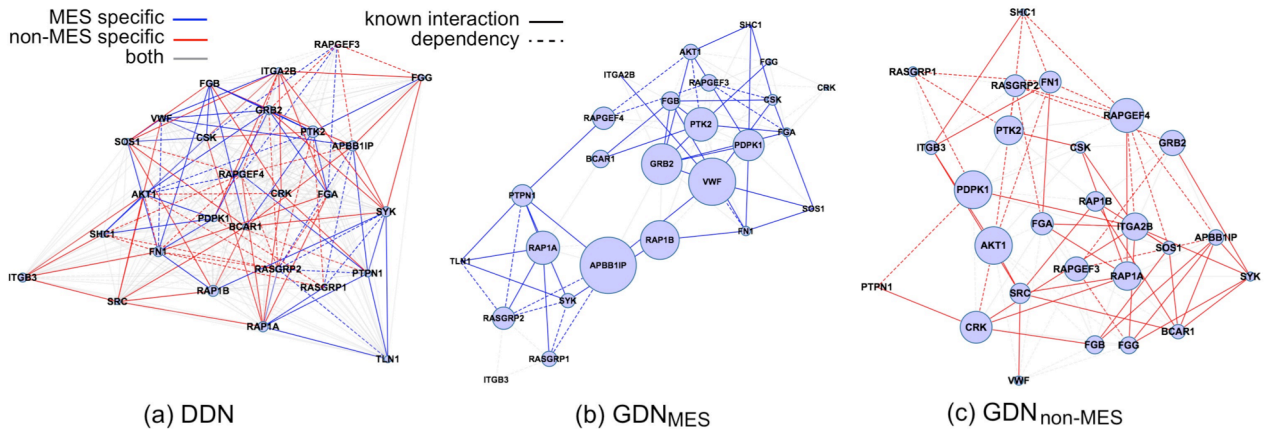

Figure 3: (a) DDN, (b) GDN$_{MES}$, and (c) GDN$_{non\text{-}MES}$ of Integrin αIIb β3 signaling (M760) pathway

‡ The full data for the betweenness centrality and their difference between GDNs are not shown due to the space constraint. However, the betweenness centrality is indicated by the size of nodes in the GDNs.

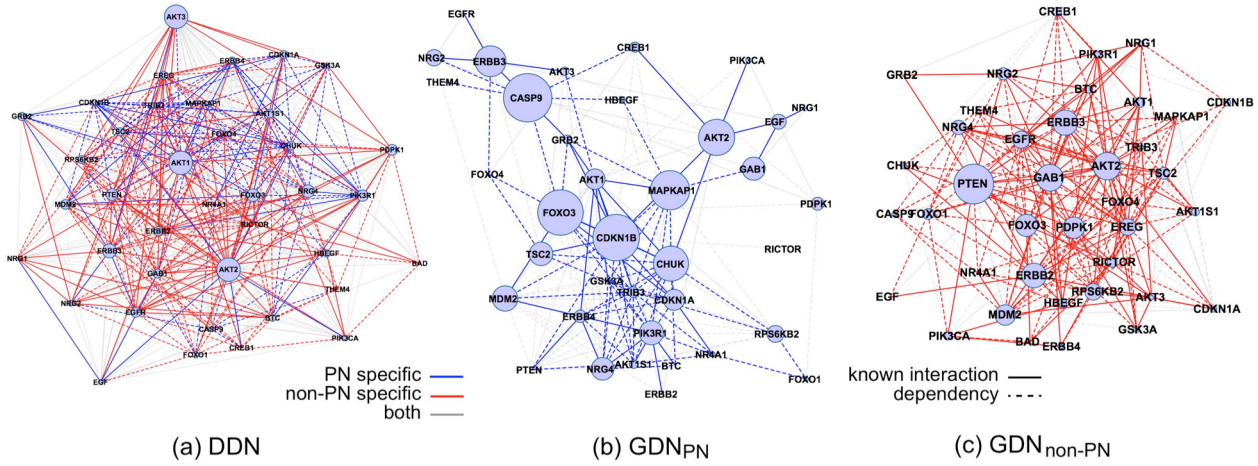

Figure 4: (a) DDN, (b) GDN$_{PN}$, and (c) GDN$_{non-PN}$ of PI3K events in ERBB2 signaling (M570) pathway

is PI3K_EVENTS_IN_ERBB2 Signaling (**M570**; http://bit.ly/1I87dUt). This pathway highlights the signaling events from ERBB2, add associated family members, signal down through PIK3CA to AKT and mTOR signaling (Figure 4). There is a shift in the dependency of the ERBB signaling receptors between the proneural and non-proneural with a lessened dependency in the proneural. This is consistent with the observation that the proneural subtype of GB seems to be more reliant on PDGFRA signaling than signaling through ERBB2 (δ_{bw}=0.77, $R_{\delta_{bw}}$=4) and EGFR (δ_{bw}=0.71, $R_{\delta_{bw}}$=7) (27). However, PIK3R1 (δ_{bw}=0.60, $R_{\delta_{bw}}$=10) does show differential dependency in proneural samples, which agrees with observation of enrichment of PIK3R1 mutations in proneural samples (27). This may suggest that PIK3R1 mutations drive PIK3CA based signaling rather than PIK3CA mutations or ERBB alterations in the proneural subtype. It may also argue that PI3K signaling may needs to be targeted differently in different subtypes of GB.

3.4. *Comparison to KDDN*

Since KDDN does not aggregate score and p-value for pathway as EDDY does, we first identify pathways enriched with differential dependency, and apply KDDN to the same data set using the same prior knowledge for comparison. We used KDDN Cytoscape plug-in with parameters λ_1 set to 0.2, λ_2 to 0.05, and δ to 0.1, the default settings. The results are summarized in Tables 3 and 4.

With the default settings, kDDN identifies fewer edges than EDDY. Nevertheless, the general trend is that EDDY and kDDN find more than twice as much agreement in condition-specific edges than disagreement (selecting edges for opposite conditions). Varying λ_1 and λ_2 can increase the number of kDDN edges to approach those found by EDDY, but we sought a consistent approach to setting these parameters for fair comparison, rather than fitting agreement *ad hoc*. A key difference between the two applications is that EDDY identifies both condition-specific and shared edges for both conditions. When we include these edges, the overlap improves somewhat, but in general, the alignment between kDDN and EDDY is not substantial. We attribute this disagreement to the enhanced sensitivity of the EDDY method in assessing significance over a distribution of network scores. This might raise a concern for potential false positive discoveries by EDDY. However, our previous analysis of EDDY with simulation data indicates the false positive rate for EDDY is low, which is also supported by low P_D (< 0.5) in Table 1 and Table 2 –

majority of edges identified by EDDY are known interactions. We leave more comprehensive comparisons between EDDY and kDDN or other similar methods to our future study.

Table 3: A comparison of DDNs found by EDDY and KDDN for GB mesenchymal. EDDY queries selected specific gene sets depending on prior weight, w_P. Statistics for the two networks are common dependencies E_S and condition-specific dependencies E_C for EDDY, and condition-specific dependencies E_K for KDDN. The last column represents concordance between KDDN and EDDY DDN, specifically $|[E_{C1} \cap E_{K1}] \cup [E_{C1} \cap E_{K1}]|$ where E_{Ci}, represents C_i-specific edges identified by EDDY and E_{Ki} represents C_i-specific edges identified by KDDN.

REACTOME Pathway (PN)	ID	w_P	$\|E_S\|$	$\|E_C\|$	$\|E_K\|$	concordance
INSULIN_RECEPTOR_RECYCLING	M506	0.0	25	108	28	8
INSULIN_SYNTHESIS_AND_PROCESSING	M764	0.0	15	75	22	7
INTEGRIN_ALPHAIIB_BETA3_SIGNALING	M760	0.0	41	104	34	9
PURINE_METABOLISM	M9271	0.0	62	190	63	21
PYRUVATE_METABOLISM	M17157	0.0	16	54	53	6
GLUCONEOGENESIS	M13748	0.5	96	183	41	12
GLYCOLYSIS	M5113	0.5	105	149	35	11
INSULIN_RECEPTOR_RECYCLING	M506	0.5	80	115	28	7
PURINE_METABOLISM	M9271	0.5	94	197	63	21
GLUCONEOGENESIS	M13748	1.0	205	106	41	7
NUCLEAR_SIGNALING_BY_ERBB4	M571	1.0	185	180	65	19

Table 4: A comparison of DDNs found by EDDY and KDDN for GB proneural

REACTOME Pathway (PN)	ID	w_P	$\|E_S\|$	$\|E_C\|$	$\|E_K\|$	concordance
ACTIVATED_POINT_MUTANTS_OF_FGFR2	M647	0.0	4	57	5	3
DOWNREGULATION_OF_ERBB2_ERBB3_SIGNALING	M549	0.0	5	23	8	3
FGFR1_LIGAND_BINDING_AND_ACTIVATION	M1092	0.0	5	47	5	3
G1_S_SPECIFIC_TRANSCRIPTION	M1040	0.0	33	35	8	3
PI3K_AKT_ACTIVATION	M714	0.0	61	186	58	19
PI3K_EVENTS_IN_ERBB2_SIGNALING	M570	0.0	78	271	83	31
ACTIVATED_POINT_MUTANTS_OF_FGFR2	M647	0.5	4	61	5	3
DOWNREGULATION_OF_ERBB2_ERBB3_SIGNALING	M549	0.5	15	28	8	2
ERK_MAPK_TARGETS	M13408	0.5	53	99	27	12
FGFR1_LIGAND_BINDING_AND_ACTIVATION	M1092	0.5	9	48	5	3
G1_S_SPECIFIC_TRANSCRIPTION	M1040	0.5	44	33	8	3
NEGATIVE_REGULATION_OF_FGFR_SIGNALING	M530	0.5	130	271	48	26
ACTIVATED_POINT_MUTANTS_OF_FGFR2	M647	1.0	19	50	5	3
ERK_MAPK_TARGETS	M13408	1.0	108	54	27	5
FGFR1_LIGAND_BINDING_AND_ACTIVATION	M1092	1.0	17	41	5	3
G1_S_SPECIFIC_TRANSCRIPTION	M1040	1.0	61	18	8	2

4. Discussion

Expression profiling and whole genome sequencing from hundreds of GB specimens by TCGA has revealed a broad spectrum of genetic alterations and discrete expression signatures and subtypes (27; 28). However, the issue of how to best target these molecular subtypes using pharmacological agents remains to be addressed. An obstacle in identifying subtype-specific drug vulnerabilities is how genetic alterations and gene expression affect wiring of key signaling networks that drives tumor phenotype (29). In this work we demonstrated that using knowledge-assisted EDDY, it is possible to identify subtype specific network wiring and gene dependencies, which may be used to identify subtype specific drug vulnerabilities.

Finally, we have recently started an implementation of the EDDY algorithm on a GPU, which has shown dramatic acceleration. Besides making computations faster and allowing for the running of larger datasets, we envision a prior weight optimization over the number of condition-specific edges. Additionally, experimental validation of highlighted differences is a main priority in the future. We have access to cohort of 64 patient derived GB xenografts that include all four GBM subtypes and are available to readily deploy to test novel hypothesis indicated through EDDY analysis.

5. Acknowledgments

This work was partly funded by National Cancer Institute, National Institutes of Health (NIH) [1U01CA168397 — Center for Target Discovery and Development (CTD2)].

References

1. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, et al. 2005. *Proc Natl Acad Sci U S A* 102:15545-50
2. Califano A. 2011. *Molecular systems biology* 7:463
3. de la Fuente A. 2010. *Trends in genetics : TIG* 26:326-33
4. Ideker T, Krogan NJ. 2012. *Molecular systems biology* 8:565
5. Guo Z, Li Y, Gong X, Yao C, Ma W, et al. 2007. *Bioinformatics* 23:2121-8
6. Hwang T, Park T. 2009. *BMC Bioinformatics* 10:128
7. Kim Y, Kim T-K, Kim Y, Yoo J, You S, et al. 2010. *Bioinformatics*
8. Ma H, Schadt EE, Kaplan LM, Zhao H. 2011. *Bioinformatics*
9. Lai Y, Wu B, Chen L, Zhao H. 2004. *Bioinformatics* 20:3146-55
10. Hu R, Qiu X, Glazko G, Klebanov L, Yakovlev A. 2009. *BMC Bioinformatics* 10:20
11. Mentzen W, Floris M, de la Fuente A. 2009. *BMC Genomics* 10:601
12. Zhang B, Li H, Riggins RB, Zhan M, Xuan J, et al. 2009. *Bioinformatics* 25:526-32
13. Zhang B, Tian Y, Jin L, Li H, Shih Ie M, et al. 2011. *Bioinformatics* 27:1036-8
14. Choi Y, Kendziorski C. 2009. *Bioinformatics* 25:2780-6
15. Jung S, Kim S. 2014. *Nucleic acids research* 42:e60
16. Lin J. 1991. *IEEE Transactions on Information Theory* 37:145-51
17. Freeman LC. 1977. *Sociometry* 40:35-41
18. Tian Y, Zhang B, Hoffman EP, Clarke R, Zhang Z, et al. 2014. *BMC systems biology* 8:87
19. Tian Y, Zhang B, Hoffman EP, Clarke R, Zhang Z, et al. 2015. *Bioinformatics* 31:287-9
20. Sayyah J, Bartakova A, Nogal N, Quilliam LA, Stupack DG, Brown JH. 2014. *The Journal of biological chemistry* 289:17689-98
21. Guarino M. 2010. *Journal of cellular physiology* 223:14-26
22. Traves PG, Pardo V, Pimentel-Santillana M, Gonzalez-Rodriguez A, Mojena M, et al. 2014. *Cell death & disease* 5:e1125
23. Jakus Z, Fodor S, Abram CL, Lowell CA, Mocsai A. 2007. *Trends in cell biology* 17:493-501
24. Li Y, Yan J, De P, Chang HC, Yamauchi A, et al. 2007. *Journal of immunology* 179:8322-31
25. Medrano-Fernandez I, Reyes R, Olazabal I, Rodriguez E, Sanchez-Madrid F, et al. 2013. *Cellular and molecular life sciences : CMLS* 70:2395-410
26. Engler JR, Robinson AE, Smirnov I, Hodgson JG, Berger MS, et al. 2012. *PloS one* 7:e43339
27. Verhaak RG, Hoadley KA, Purdom E, Wang V, Qi Y, et al. 2010. *Cancer cell* 17:98-110
28. Cancer Genome Atlas Research N. 2008. *Nature* 455:1061-8
29. Oh YT, Cho HJ, Kim J, Lee JH, Rho K, et al. 2014. *PloS one* 9:e103327

Appendix

Systematic ID	Pathway
M10122	RETROGRADE_NEUROTROPHIN_SIGNALLING
M1014	IL_6_SIGNALING
M1016	SYNTHESIS_OF_VERY_LONG_CHAIN_FATTY_ACYL_COAS
M1016	SYNTHESIS_OF_VERY_LONG_CHAIN_FATTY_ACYL_COAS
M10272	IONOTROPIC_ACTIVITY_OF_KAINATE_RECEPTORS
M1051	INTEGRATION_OF_PROVIRUS
M1062	ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_LOADING_OF_CLASS_I_MHC
M1075	INWARDLY_RECTIFYING_K_CHANNELS
M11184	ENDOGENOUS_STEROLS
M12627	DOPAMINE_NEUROTRANSMITTER_RELEASE_CYCLE
M12967	MRNA_3_END_PROCESSING
M13115	G_PROTEIN_ACTIVATION
M15195	MAPK_TARGETS_NUCLEAR_EVENTS_MEDIATED_BY_MAP_KINASES
M15243	GAP_JUNCTION_DEGRADATION
M1662	SIGNALING_BY_BMP
M1662	SIGNALING_BY_BMP
M16702	ACTIVATED_AMPK_STIMULATES_FATTY_ACID_OXIDATION_IN_MUSCLE
M17787	GLUCURONIDATION
M1825	REGULATION_OF_INSULIN_SECRETION_BY_ACETYLCHOLINE
M3634	CASPASE_MEDIATED_CLEAVAGE_OF_CYTOSKELETAL_PROTEINS
M3661	FGFR_LIGAND_BINDING_AND_ACTIVATION
M552	PROLACTIN_RECEPTOR_SIGNALING
M564	MEMBRANE_BINDING_AND_TARGETTING_OF_GAG_PROTEINS
M583	RIP_MEDIATED_NFKB_ACTIVATION_VIA_DAI
M6034	SEROTONIN_RECEPTORS
M612	CIRCADIAN_REPRESSION_OF_EXPRESSION_BY_REV_ERBA
M661	SIGNALING_BY_FGFR1_MUTANTS
M704	SIGNALING_BY_FGFR_MUTANTS
M7169	NCAM1_INTERACTIONS
M719	SHC1_EVENTS_IN_EGFR_SIGNALING
M769	ELEVATION_OF_CYTOSOLIC_CA2_LEVELS
M794	ACTIVATION_OF_CHAPERONES_BY_ATF6_ALPHA
M801	ACTIVATION_OF_CHAPERONE_GENES_BY_ATF6_ALPHA
M837	CREB_PHOSPHORYLATION_THROUGH_THE_ACTIVATION_OF_RAS
M846	FRS2_MEDIATED_CASCADE
M860	SHC_MEDIATED_CASCADE
M875	NETRIN1_SIGNALING
M899	IL1_SIGNALING
M901	GLOBAL_GENOMIC_NER_GG_NER
M907	CALNEXIN_CALRETICULIN_CYCLE
M918	N_GLYCAN_TRIMMING_IN_THE_ER_AND_CALNEXIN_CALRETICULIN_CYCLE
M932	SYNTHESIS_SECRETION_AND_INACTIVATION_OF_GLP1
M936	TRAF6_MEDIATED_IRF7_ACTIVATION
M9450	PLATELET_ADHESION_TO_EXPOSED_COLLAGEN
M947	INHIBITION_OF_VOLTAGE_GATED_CA2_CHANNELS_VIA_GBETA_GAMMA_SUBUNITS
M9694	ACTIVATION_OF_THE_PRE_REPLICATIVE_COMPLEX