Bioinformatics Advance Access published August 27, 2013

# Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE)

Evan O. Paull[1], Daniel E. Carlin[1], Mario Niepel[2], Peter K. Sorger[2], David Haussler[1,3], and Joshua M. Stuart[1,*]

[1]Department of Biomolecular Engineering, University of California, Santa Cruz, CA 95064.

[2]HMS LINCS Center, Department of Systems Biology, Harvard Medical School, Boston, MA.

[3]Howard Hughes Medical Institute, Santa Cruz, CA.

Associate Editor: Prof. Martin Bishop

**ABSTRACT**

**Motivation:** Identifying the cellular wiring that connects genomic perturbations to transcriptional changes in cancer is essential to gain a mechanistic understanding of disease initiation, progression, and ultimately to predict drug response. We have developed a method called Tied Diffusion Through Interacting Events (TieDIE) that uses a network diffusion approach to connect genomic perturbations to gene expression changes characteristic of cancer subtypes. The method computes a sub-network of protein-protein interactions, predicted transcription factor-to-target connections, and curated interactions from literature that connects genomic and transcriptomic perturbations. **Results:** Application of TieDIE to The Cancer Genome Atlas (TCGA) and a breast cancer cell line dataset identified key signaling pathways, with examples impinging on MYC activity. Interlinking genes are predicted to correspond to essential components of cancer signaling and may provide a mechanistic explanation of tumor character and suggest subtype-specific drug targets. **Availability:** Software is available from the Stuart lab's wiki: https://sysbiowiki.soe.ucsc.edu/tiedie. **Contact:** jstuart@ucsc.edu.

## 1 INTRODUCTION

To optimize cancer treatment, whole-genome sequencing and expression data for an individual patient must be synthesized into a coherent explanation of disease-causing changes. Gene networks encapsulate our understanding of how genes and their products interact in the cell to mutually influence activity through protein-protein, protein-DNA, and coupled metabolic reactions. However, different tumors usually harbor unique combinations of mutations, or other genomic or epigenomic changes. A key question is how best to infer the structures of gene networks important for normal and diseased phenotypes using high-throughput data and *a priori* biological knowledge. Viewing cancer from a gene network perspective is expected to enhance our understanding of disease initiation, progression and therapy.

Given genes with functions disrupted in a particular type of cancer, newly implicated genes can be identified by searching for those with known regulatory connections to the input set. However, this task is complicated by the presence of many mutations whose functional significance in cancer is unclear, leading to many false positive discoveries. For example, using data on copy number alterations, gene mutations, and methylation status, it may be difficult to distinguish the genomic changes that exert a physiologically meaningful influence on tumor biology from numerous passenger events that result from loss of genome integrity. One can identify sub-networks that interconnect mutated genes, enriching the set of events for those proteins participating in common pathways. The assumption underlying this approach is that such mutations are more likely to be functionally relevant. Approaches such as MEMo (Ciriello, et al., 2012) and Dendrix (Vandin, et al., 2012) have been applied successfully in this manner. Several sub-network enrichment methods have been developed to identify regions of a network that contain an unexpectedly high number of relevant genes (see *Supplemental Methods* for an overview). Importantly, most methods suffer from the influence of curation bias in the network. The presence of "hub" genes that have many connections simply due to being studied to a greater extent in the literature are selected at high frequency even given random input genes. One promising class of approaches that helps mitigate this problem is the class based on heat-diffusion such as the HotNet algorithm (Vandin, et al., 2012). Intuitively, a diffusion strategy makes the *a priori* relevance of a hub comparable to a sparsely-connected gene because hubs might receive more total heat than a non-hub but the hub can also lose the same proportion out of its many connections.

The Tied Diffusion of Interacting Events (TieDIE) method described here extends on the heat diffusion strategies by leveraging different types of genomic inputs to find relevant genes on a background network with high specificity, in an attempt to reduce the false positive rate of previous approaches. Figure 1 shows a simple schematic of TieDIE applied to two distinct sets of genes; a *source set* (e.g. mutated genes) and a *target set* (e.g. transcription factors). Using two diffusion processes, the method discovers newly implicated genes as linking nodes residing on paths connecting sources to targets where the diffusion processes overlap. A logically consistent solution can then be extracted from the resulting network of sources, targets, and linkers (Supplemental Figure S1).

*To whom correspondence should be addressed.

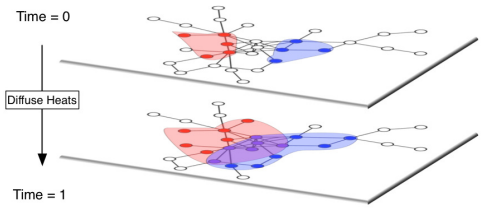


**Figure 1. Schematic of TieDIE.** Relevant genes from two distinct sets are shown as nodes colored by dies diffusing on a network from a source set (e.g. mutated genes; red nodes) and target set (e.g. transcription factors; blue nodes). "Linker" genes (purple nodes) residing between the source and target sets are revealed through a diffusion process evolved over time; two time slices shown as stacked layers of the same network.

We demonstrate that using two diffusion processes improves our ability to recover pathway models. The overlay of genomic perturbations in a single tumor sample with TieDIE solutions reveals patient-specific networks that may provide insights into therapy.

## 2 METHODS

TieDIE code is written in Python and is available at https://sysbiowiki.soe.ucsc.edu/tiedie.

### 2.1 Tied Diffusion

The TieDIE approach searches for relevant interconnecting genes on a background network using a diffusion strategy. The method is given as input an interaction network graph $G$ containing $N$ vertices $V = \{v_1..v_N\}$ that represent genes, proteins, or other biological pathway features such as gene products, protein complexes and cellular abstract processes. The nodes in $G$ are interlinked by $I$ edges $E = \{e_1..e_I\}$ representing both directed interactions as well as undirected relations such as protein-protein interactions. The interactions can be derived from curated sources such as NCI's Pathway Interaction Database (PID) (Schaefer, et al., 2009), from functional genomics predictions such as undirected high-throughput protein-protein assays or directed transcription-factor to target interactions such as from genome-wide chromatin-immunoprecipitation experiments, or from a mixture of both sources such as Reactome's Functional Interaction Network (Joshi-Tope, et al., 2005). The diffusion approaches used here make use of the adjacency matrix $\boldsymbol{A}$ of $G$, where $\boldsymbol{A}_{ij}$=1 if node $i$ activates node $j$, $\boldsymbol{A}_{ij}$=-1 if node $i$ represses or inactivates node $j$, and 0 otherwise.

In addition to the graph, the method is given a set of scores for each node in $G$. Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_N]$ be a vector of scores assigned to the nodes in the graph. Typically only a limited set of genes (e.g. in the range 10 to 50) will have known involvement in the disease process being studied. For example, this *involved* set may consist of genes for which a minimum number of mutations or copy number changes or DNA methylation silencing events have been detected in a given cohort of patient samples. Nodes corresponding to these involved genes are assigned scores between -1 and +1 to reflect a positive or negative association of activity with the disease state under study. Nodes associated with genes not known to be involved in the disease process are assigned a value of 0 to reflect an a priori belief in no involvement. The values in $\boldsymbol{x}$ can represent different types of measurements on the genes. For instance, the scores might reflect a how often a gene is mutated in one subtype of patients compared to another (e.g. from a –log of the P-value computed from a Fisher's exact test to detect differential mutation frequency). Alternatively the scores can reflect a gene's differential expression in tumor versus normal. Statistical techniques such as Significance Analysis of Microarrays (SAM) (Tusher, et al., 2001) may be used to compute the significantly differentially expressed genes, and the SAM score (d-statistic) for each gene normalized to this range. The genomic events may be pre-filtered with algorithms such as MutSig (Chin, et al., 2011) and MEMo (Ciriello, et al., 2012) that use sample statistics to find events that are likely to be "driving" the cancer phenotype. Note here that all mutations are assumed to lower a gene's activity (even though a minority, oncogenic mutations in particular, increase gene activity) and to take precedence over copy number alterations (e.g. amplifications) and expression events (e.g. over-expression), which will inflate the false-negative rate of the algorithms tested here as some true paths will not be counted. Relaxing this assumption is an important topic discussed in the *Supplemental Text*. The input vector of positive scores is scaled to match an intuition that the scores reflect a stationary probability distribution of occupancy on the nodes in the network obtained from a random walk process; i.e. $\sum_{i-1}^{N} x_i = 1$.

A new vector of scores, $\boldsymbol{x'}$, can be obtained for all of the genes in $G$ by diffusing the scores of the involved set onto the rest of the graph. The diffusion process places high scores on nodes that are near the input set of genes, which may implicate new genes with roles in the disease state. We consider methods here that update the state of all of the nodes in $G$ by some function computed on the original vector of scores from the known involved set; i.e. $\boldsymbol{x'} = r(\boldsymbol{x},\boldsymbol{A})$ for some *relevance function* that is a function of the input set of scores and of the full adjacency matrix. In this work, we tested three different approaches for use as the relevance functions; HotNet, which uses an undirected heat diffusion process; Google's PageRank, which incorporates direction of the links in a random walk; and Signaling Pathway Impact Analysis (SPIA), which incorporates both directionality and the excitatory and inhibitory nature of the interactions. These relevance functions are described in more detail in the *Supplemental Methods*.

In contrast to previous approaches that implicate sub-networks by running the relevance function with a single input set of genes, our algorithm extends the strategy by using multiple diffusion processes and then identifying overlapping regions in $G$, to find genes in a network that are proximal to multiple input sets. We

develop the approach here for the special case of two input sets but the method generalizes to any arbitrary number of input sets.

Suppose we are given a *source* set ***S*** and *target* set **T** where ***S*** acts upstream of ***T***. In the cancer setting ***S*** may correspond to genes involved in genomic alterations – mutations, deletions, and amplifications – while the target set may correspond to genes involved in transcriptional and post-transcriptional activation or de-activation. However, any features in the pathway diagram in addition to proteins can be included in the input sets including protein complexes, small molecules or metabolites, and cellular processes. In the same spirit as the spanning tree approach (Huang and Fraenkel, 2009), we are interested in identifying parsimonious networks that connect ***S*** to ***T***. We now let ***x*** represent the vector of scores for the source set and ***y*** the scores for the target set.

We can visualize the process for two input sets by imagining the intermixing of different color dyes, diffusing from two sources through a lattice representing *G*. The intensity and hue of the dye reveals whether a particular node is close to either the source, target, or to both sets. To disambiguate these cases and identify points that reside only between the sets, we determine a score for all nodes based on the relevance scores already computed:

$$z = f\left(r(x,A), r\left(y,A^T\right)\right), \quad (1)$$

where $\boldsymbol{A}^T$ is the transpose of the full adjacency matrix, the function *f*() is chosen to assign high relevance scores to nodes where both $r(\boldsymbol{x},\boldsymbol{A})$ and $r(\boldsymbol{y},\boldsymbol{A}^T)$ are high, and lower scores when either of the two are low. Note that the transpose of the adjacency matrix is used to force the diffusion to proceed upward from the targets by supplying a graph containing reversed edges. When applied to directional diffusion approaches like PageRank and SPIA this has the effect of running the algorithm backward. Of course the transpose makes no difference for undirected approaches like HotNet's heat-diffusion since $r_{HotNet}(\boldsymbol{y},\boldsymbol{A})=r_{HotNet}(\boldsymbol{y},\boldsymbol{A}^T)$. We refer to $z_i$ as the *linking score* for node *i*. This form is attractive because it decomposes over the separate relevance calculations, which can be "plugged in" to the tied-diffusion calculation. In this study we chose *f()* equal to the *min()* operator to extract genes that have intersecting evidence from both input data sets. A set of linking genes is obtained by thresholding the linking scores using a chosen value $\alpha$ selected to guarantee a desired level of specificity as a fixed multiple of the input set size described in *Supplemental Methods*.

The network solutions that interconnect sources and targets can be large, containing hundreds of genes. In addition, plug-in diffusion approaches like HotNet and PageRank ignore the logical consistency of the set of identified genes identified. While it may be advantageous to maintain such contradictory influences (e.g. to highlight places of possible model discrepancy), it is generally difficult and cumbersome to extract meaningful information from large networks. We therefore introduce a filter to specifically select for consistent regions of the network, which focuses attention on better-defined and more interpretable subsets of the solution space. To this end, we add an edge between each pair of nodes belonging to the source, target, or linker gene sets. This set of edges and nodes defines an initial graph *G'* that is reduced by finding the subset of edges that connect ***S*** to ***T*** through paths that are logically consistent with both the input data and the network model (see *Supplemental Methods*).

## 3 RESULTS

### 3.1 Motivating synthetic example

Consider a toy network in which a pre-defined "core" sub-network is embedded. Alterations to the state of expression of genes in the core network are assumed to contribute, or "drive," a disease process, while alterations outside of the core are assumed to have little effect. To create a small, but realistic scale-free network (Barabasi and Albert, 1999) for comparing single-source versus tied-diffusion, the scale_free_graph simulator available within the NetworkX library (Version 1.7) was used (see Supplemental Methods). A core sub-network of twelve genes containing four "source" genes and two "target" genes was then embedded in the NetworkX network, for a total of 125 nodes and 195 edges. In addition, six linking genes were simulated to connect the sources to targets. The four sources and two targets are assumed known and their scores are provided. However, the scores for the genes in the remainder of the embedded sub-network are not provided, simulating a case in which their involvement in the disease process is unknown *a priori*. To identify linker genes, scores for the core-network nodes were also simulated. For sources the scores reflect the degree of alteration such as mutation frequency across a cohort or its predicted impact from sequence-based analyses. For targets, the scores represent the degree of differential expression observed in the targets of a transcription factor. We also simulated six false-positive sources representing genes deleted or amplified by neutral "hitchhiking" events or genes with non-disease-specific tendencies to pick up passenger mutations .

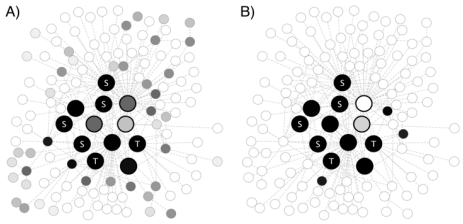


**Figure 2**. **Frequency of discovered core and off-core genes in single-source and tied-diffusion in a simulated network. A.** Single-source diffusion over the synthetic network. Darker colors indicate genes in a larger fraction of network solutions in repeated simulated trials at a fixed recall of 4 out of 6 signaling genes. **B.** The corresponding tied-diffusion frequencies at identical recall and test conditions.

We assessed the precision of both single-source and tied-diffusion when applied to finding a majority (four out of six) of the linking

genes in the core sub-network (Figure 2). In the case of single-source diffusion, the algorithm was given the same mixture of true and false positive sources including the two true-positive targets. As expected, the single-source approach produced more high-scoring nodes outside the core-network after diffusion compared to the tied-diffusion, which reflects a lower precision (Figure 2A). Note that the several dozen small, dark, "hot" nodes distal to the core pathway do not lie between the source and targets, and thus represent false positives. In comparison, the tied-diffusion approach produces far fewer off-core dark nodes; the only three such false-positives all reside between a source and target (Figure 2B). The example demonstrates how the single-source method can diffuse heat into peripheral regions of the network in situations whereas tied-diffusion focuses near the core sub-network. This occurs because we expect false positives to occur less frequently between two sets than proximal to any one set.

We also tested the simple k-nearest-neighbors (KNN) approach for predicting new genes based on the fraction of sources and targets in each gene's neighborhood. Both the single and double-diffusion approaches achieved significantly higher precision on average compared to the KNN approach (see Supplemental Figure S2). Thus, even though the starting number of known genes is the same for all methods, the biased diffusion toward the target leads to a higher quantifiable precision in distinguishing between the true- and false-positive core genes in this setting.

### 3.2 Tied-diffusion predicts breast cancer-related genes with high precision

We next asked whether the tied diffusion approach maintains this improvement when applied to actual patient datasets. The main assumption behind TieDIE is that more accurate pathways are obtained by directing diffusion processes to connect genes involved in somatic mutations to observed transcriptional effects. We therefore tested whether this approach could achieve higher accuracy than the comparable single-source approaches like HotNet that only consider the genomic perturbations without regard to the state of the transcriptome. We chose breast cancer as a test system because many mutations have been identified and diverse types of data are available. Lists of implicated genes were collected from both WikiPathways-WP1984 (Kelder, et al., 2012), a list of frequently mutated breast cancer genes from the Catalogue of Somatic Mutations in Cancer (COSMIC) version 57 (Forbes, et al., 2011), and collections from the breast cancer analysis working group of the Cancer Genome Atlas (TCGA) project (TCGA Network, 2012). At the time of acquisition, the breast cancer dataset included patient tumor samples for 533 patients and matched normal samples for a smaller subset, each with genomic sequencing and microarray expression data. To incorporate a diverse set of genomic and epigenomic alterations specific to a subtype, genes within regions of predicted copy number gain or loss based on the GISTIC algorithm (Beroukhim, et al., 2007) were identified as having at least five samples with either high-copy amplifications or homozygous deletions (TCGA Network, 2012). We also included as sources the frequently mutated genes published by the TCGA Network (i.e. genes mutated in more than ten tumors). Altogether, 110 genes were collected with significant numbers of events involving 41 amplified genes in 1258 samples, 14 deleted genes in 147 samples, and 54 mutated genes in 1115 samples. One gene, BRCA1, was methylated in 15 samples and had gene expression inversely correlated in these tumors and so was also included. The background network (SuperPathway; see Supplemental Methods) contained a collection of curated transcriptional, protein-level, and complex interactions for 4,737 genes, proteins and abstract concepts, with 101,526 interactions (Heiser, et al., 2011). To enrich for nodes with measured data, we removed complexes from the published SuperPathway after transferring their interactions to the incident constituent members (see Supplemental Data 1). All transcription factors used in this study were taken as proteins linked by a transcriptional regulatory interaction to at least one other gene in the SuperPathway.

Compared to single-source diffusion, the tied diffusion approach demonstrated higher precision over an appreciable range of recall (by varying the $\alpha$ parameter; Supplement Methods S.1.5) when used to predict breast cancer-implicated genes from COSMIC and WikiPathways gene sets (Supplemental Figure S3). Because WikiPathways includes genes related to breast cancer without any documented mutations, these results imply that the method may increase our ability to identify cancer essential genes as new drug targets (see Discussion). In realistic settings with 20% recall, the tied-diffusion approach approximately doubles the precision for finding the genes. At recall levels higher than 30-40% the performance of TieDIE relative to single-source diffusion decreases but TieDIE is always equal or better than single diffusion methods. Including more sets may help increase the precision further (e.g. proteins with perturbed kinase activity). However, the linear decline in precision past these levels of recall indicates that any diffusion process searching sufficiently far from the input sets may be no more effective than randomly drawing candidate genes.

### 3.3 TieDIE reveals subtype-specific networks

We used our approach to elucidate breast cancer-related networks that distinguish the major breast cancer subtypes—the so-called Basal and Luminal A subtypes. We chose to compare these two because they have a clear transcriptional signature compared to other subtypes that have a more intermediate or heterogeneous signature such as the Luminal B and HER2-amplified. Basal breast cancer is known to be a more proliferative form of cancer compared to the luminal type and the subtypes are thought to result from mutagenic transformation of different originating cell types. Since these subtypes respond differently to both general cytotoxic and targeted therapies, it is important to identify pathway mechanisms that differentiate the two subtypes to discover new treatments. The subtypes are easily identifiable from transcriptome

signature analysis. However, different genomic alterations within each subtype can lead to the same transcriptional profile. For example, PIK3CA mutations or PTEN deletions are both correlated with the luminal-A expression subtype. On the other hand, it is possible that seemingly synonymous genomic alterations can lead to different transcriptional subtypes. For example, while TP53 mutations are enriched in basal cancers, some luminal cancers also harbor TP53 mutations. Thus, network diffusion approaches may reveal how genomic alterations correlate with transcriptional signatures in a subtype-specific manner.

To find the significant pathway differences between these cancer subtypes, we performed a differential analysis between 99 Basal and 235 Luminal-A samples from the TCGA dataset. We used a set of 110 genomic perturbations published by the Cancer Genome Atlas Network (TCGA Network, 2012) and applied a chi-square proportions test to find those that occur with significantly different frequency in one subtype as compared to the other. This uncovered 12 genes with mutations significantly associated with either the basal or luminal subtypes at the p=0.05 level, used as the source gene set ***S***. We used this set for network search, while also weighting each source gene in proportion to the absolute log-ratio of its perturbation frequencies in basal vs. luminal-A tumors. A small constant was added to the frequencies to avoid log transforms of zero. The absolute log-ratio was normalized to values between 0 and 1, by dividing by the maximum absolute log-ratio.

The target set ***T*** was defined similarly. Transcription factors (TFs) were determined by inspecting the differential expression of each TF's predicted target genes. TFs with activity more associated with basal tumors should have a set of targets with high and/or low expression compared to chance expectation. We used a simplified version of the Califano lab's MARINa algorithm (Lim, et al., 2009) to find transcription factors with targets differentially expressed in basals compared to luminals. Predicted target genes were collected from the SuperPathway. SAM (Tusher, et al., 2001) was run to derive "delta" scores representing the degree of differential expression in basal tumors compared to luminal-A tumors for each target gene. Gene set enrichement analysis (GSEA) (Subramanian, et al., 2005) analysis was then used to identify TFs having targets with a non-random distribution of SAM delta scores. Rather than apply a strict multiple-hypothesis correction at the level of significance, we instead retained TFs with scores at a relaxed cutoff of p = 0.05 and then associated a relevance score to each retained TF by dividing its absolute GSEA score by the maximum absolute score. Similarly, source genes were retained that had a Fisher's exact of 0.05 and were assigned relevance scores equal to the normalized absolute log-transformed P-value. These data were used to test various relevance measures for use in single-source and tie-diffusion described next.

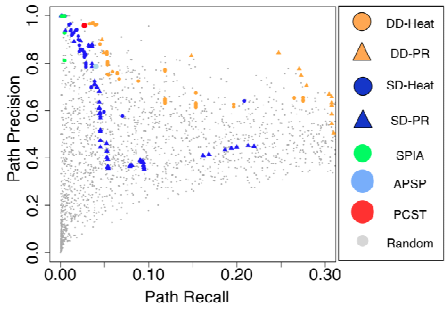


**Figure 3. Precision of single-source (blue points) and tied-diffusion (orange points) with different relevance scores for identifying pathways in a breast cancer.** Any paths containing even a single randomly injected "decoy" link were considered false-positives. Recall measures the number of logically consistent paths (x-axis; see Methods) out of the total possible; precision measures the number of such consistent paths in the total number returned. Relevance scores tested are heat diffusion (circles), Personalized PageRank (triangles) and SPIA (green circles). For comparison, included are all-pairs shortest paths (APSP; blue circle) and Prize-Collecting Steiner Trees (PCST; red dot). Randomly generated networks of various sizes were obtained to estimate the background distribution (gray dots). Different levels of precision and recall were obtained by varying algorithm parameters (e.g. the $\alpha$ parameter for single and tied diffusion; Supplemental Methods S.1.5).

### *3.3.1* Evaluation of relevance scores and competing methods

We evaluated the precision of different methods for their ability to find true network paths in the presence of randomly generated false-positive "decoy" links, which were added to the SuperPathway to increase its number of interactions by 50%. Any paths in the final solutions that contained even a single decoy link were considered to be false positive. By computing the number of such paths that exist in the entire SuperPathway network we calculated the total number of true-positive paths and used this to compute the precision and recall of the solutions (Figure 3).

We compared tied-diffusion to its single-source counterpart and to competing approaches such as Prize-Collecting Steiner Trees (PCST) as implemented by Dietrich et al. (2008) (Dittrich, et al., 2008), and Dijkstra's all-pairs-shortest-paths algorithm as a baseline approach. Three plug-in relevance score functions were included: heat-diffusion, Google's Personalized PageRank, and SPIA (see *Methods*). Diffusion strategies performed comparably to the three competing methods at the specific levels of recall obtained with each approach (green, blue, and red points in Figure 3). In addition, we found that tied-diffusion had higher precision over varying recall compared to the single-source equivalents. Also, the heat-diffusion relevance function performs comparably to Personalized PageRank over moderate levels of recall, after which Personalized PageRank outperforms the heat-diffusion kernel. Thus, given the significant computational benefits of using a pre-computed kernel (see Methods), we opted to use the heat diffusion approach as the principal algorithm for the remainder of our analysis as it allowed for a greater amount of experimentation.

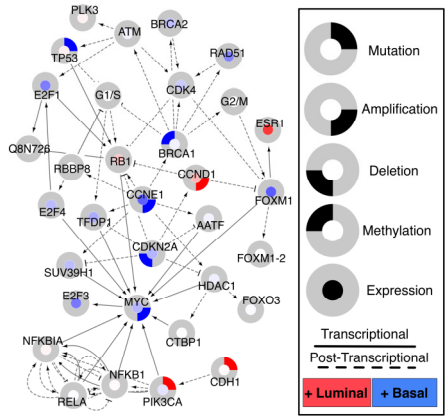


**Figure 4.** Tied-diffusion result for LumA vs. Basal breast cancer subtypes. The inner coloring of the rings represents the differential expression in luminal A as compared to basal samples. The outer ring represents differential frequency of genomic perturbations in luminal samples as compared to basal samples: differential mutation (upper right sector), amplification (lower right), deletion (lower left), and DNA methylated CpG islands near the promoter (upper left).

We next ran TieDIE with heat-diffusion as the relevance score to solve for a basal-luminal network as shown in Figure 4. A network connecting basal-enriched genomic events such as TP53 to basal-associated activated transcription factors such as MYC, was found to be significant ($p < 0.002$) using a permutation-based simulation (*Supplemental Methods*; Supplemental Figure S4). As expected, these paths involve several DNA damage checkpoint genes such as ATM, retinoblastoma (RB1), Cyclin E1 (CCNE1), and p16-ARF (CDKN2A). On the other hand, the luminal pathways include expected genes such as the often-targeted estrogen receptor protein (ESR1), the frequently mutated kinase PIK3CA and E-cadherin (CDH1), a protein that interacts with PIK3CA in the cell membrane. The two subtypes also differ in their copy number profiles with basal tumors characterized by amplifications in MYC and Cyclin E1 (CCNE1) and deletions in p16-Arf (CDKN2A). In contrast, luminal-A tumors tend to have amplifications in Cyclin D1 (CCND1), which interacts in an opposing fashion on the retinoblastoma gene (RB1) compared to Basal samples. This suggests that these Luminal A tumors may have either "flipped" or lost the functional interaction between CCND1 and RB1, or that the increase in transcriptional activity of RB1 can be explained by other upstream nodes such as TP53 or E2F1. Linking genes in the map may represent breast cancer "essential" genes whose functions are required for altered signaling logic in tumors. In support of this idea, inhibitors to PLK3 and HDAC1 were found to sensitize breast cancer cell lines (Heiser, et al., 2011) and targeting chromatin remodelers such as the HDACs is a current focus of clinical trial work. Thus, TieDIE's breast cancer network represents a data-driven graphical summary of testable hypotheses that can explain the simultaneous protein activation, transcriptional activity, and edge interactions found between many of the key genes involved in breast cancer.

### *3.3.2* Confirmation of subtype-specific network in cell lines

Studies in immortalized cell lines provide a convenient method for exploring the biology of various cancer subtypes. One strategy is to identify perturbations such as drug exposure, siRNA knock-down, or sets of extracellular ligands that block or induce cell death in cancer cell lines. An assumption in such studies is that results in cell culture can be transferred to real tumors *in vivo*. At a minimum this requires that the molecular networks in cell lines and primary tumors have significant similarity. We therefore tested the ability of TieDIE to infer basal-luminal networks generated from a completely independent data set collected in breast cancer cell lines.

Data for a panel of 36 breast cancer cell lines (17 basal, 19 luminal) was obtained from the Gray lab at OHSU (Heiser, et al., 2011). We used microarray gene expression data from these lines to get the target input scores and repeated the TieDIE analysis, using the same set of TCGA sample-derived genomic perturbations as the source set. The resulting network was found to have a high degree of overlap with the Basal vs. Luminal-A network derived from TCGA data. Removing either source or target genes used in either the cell line or tumor input sets from consideration, the linking 416 edges and 77 nodes in the cell line derived network overlapped with 213 edges and 52 nodes in the TCGA-derived network. We found this result to be surprising, given the fact that a much smaller overlap of 14 nodes (22% of the TCGA and 25% of the cell line target sets, respectively) was found in the "downstream" input sets of each. Also the significance of the overlap was much higher according to the hypergeometric test for the tied-diffusion result, with a p-value of $4 \times 10^{-73}$ (Supplemental Figure S5), compared with $2 \times 10^{-3}$ for the downstream sets only. We therefore conclude that each of the input sets is close in pathway space despite the low fractional overlap, allowing TieDIE to find a similar set of linker genes that represent the connecting topology between each pair of input sets. Interestingly, metabolism and biosynthetic processes were found only in the tumor-derived networks, reflecting the exceptional characteristics of cell line growth and their media.

### *3.3.3* Application to sample-specific networks

We next applied diffusion approaches to characterize the specific pathways of individual samples. For each sample we identified which of the transcription factors in the "downstream" set—identified by GSEA on the cohort-wide expression data—had at least one differentially expressed target for that sample. We then connected these tumor-specific active transcription factors to the genomic perturbations in that sample present in a *scaffold network*, a background network derived from a TieDIE solution from the cohort similar to the one shown in Figure 4 except using a smaller $\alpha$ parameter to obtain a larger starting network of 106 nodes and 423 edges.

To test if TieDIE provides an accurate scaffold, we performed a search for sample specific networks over TieDIE networks of multiple sizes as well as the entire SuperPathway. Using the procedure of adding random decoy links as described earlier, we measured the ratio of 'true' and 'false' paths in each sample-specific network solution, and plotted them for each choice of network 'scaffold' (Supplemental Figure S6). The precision of the sample-specific networks was plotted, and the mean precision was significantly higher when using the TieDIE summary networks compared to using the entire SuperPathway. This precision declines gradually as the TieDIE summary networks increase in size, which is expected, given that they capture a larger fraction of the possible edges in the starting network and in the limit are identical to the SuperPathway. This shows that the improved precision achieved by the TieDIE summary networks is transferred to sample-specific networks, thereby using data from the entire cancer cohort to inform the network predictions for individual samples.

We evaluated the ability of TieDIE to connect differentially expressed genes to at least one of the genomic perturbations represented in the sample-specific network. Differentially expressed genes were collected from those downstream of the input set of transcription factors in the cohort's TieDIE network. For the majority of samples these specific networks explained a significant fraction of differential expression (20-80%), although for a subset of samples none of the expression could be explained (Supplemental Figure S7). This may be due to missing links in the starting network, which greatly affects the TieDIE solutions, because the method cannot infer missing links. In addition, genomic perturbations that only occur in a small minority of samples may not be represented in the TieDIE network, as the method, by design, will filter out pathway elements that cannot be supported by a sufficient amount of sample data. This tradeoff increases the overall quality of the networks, at the expense of missing potentially novel, but rare, molecular mechanisms that may drive the cancer phenotype in a small minority of samples.

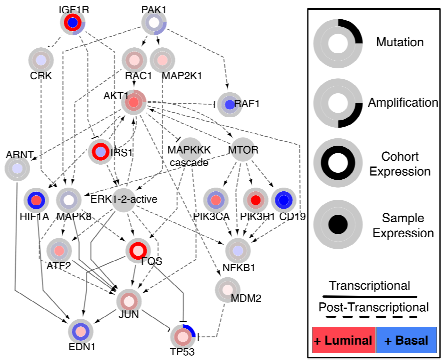


**Figure 5. Luminal A sample TCGA-BH-A0BR specific network reveals Basal-like molecular behavior**. The network connects genomic perturbations in the sample, red or blue rings around nodes, to transcriptional changes in the same sample, inner node coloring. Red and blue colors indicate higher and lower cohort mutation rates in luminal A samples as compared with Basal, outer ring; overall cohort differential expression of luminal A compared with Basal samples, second ring; individual sample expression, inner circle. Transcriptional interactions, solid lines; post-transcriptional interactions, dashed. Activating interactions, arrow at the target node; inactivating, flat bars.

### *3.3.4* Mapping the network of an abnormal luminal-A tumor

A major goal in cancer systems biology is to infer a specific network for each patient's tumor and, as data become available, each subclone identified within the tumor. Accurate network models could be used to explore a large space of potential targets to kill the tumor *in silico*. We therefore applied TieDIE to identify a pathway solution for every tumor sample in the TCGA breast cohort. To illustrate the results, we focus here on the non-canonical tumor sample TCGA-BH-A0BR, which had an intermediate pathway state between the classic basal and luminal-A subtypes as evidenced in its CircleMap plot (Figure 5). Tumor heterogeneity may contribute to such "mixed" samples or may reflect a tumor evolution distinct from the classic basal and luminal pattern. The sample-specific analysis (*Supplemental Methods*; S.2.5) of this tumor reveals it has a hybrid set of genomic perturbations with both luminal-A-like events such as an AKT1 mutation and several basal-like events such as a TP53 mutation, and amplifications in IGF1R and PAK1.

Even though the patient sample has a luminal-associated AKT1 mutation the surrounding network is more consistent with wild-type AKT1 activity. Namely, HIF1A is active, reflecting a basal program of hypoxic response and angiogenesis, further evidenced by increased EDN1 expression. In addition, IRS1 and PIK3CA expression are basal-like, and these network properties are maintained by a basal-like PAK1 amplification that, in this patient, may promote the activity of RAC1 and MAP2K1.

Interestingly, insulin like growth factor receptor (IGF1R) is known to be involved in the control of breast cancer cell growth. Blocking or reducing the activity of this receptor has been found to reduce growth in at least one luminal breast cancer cell line (Guvakova and Surmacz, 1997) and increase sensitivity in trastuzumab resistant cells are associated with IGFR1 as well as PAK1 (Rayala, et al., 2006). The TieDIE network suggests that IGF1R and PAK1 amplifications may be driving the growth signaling pathways of this tumor, in the absence of HER2 amplification. Experimental validation of such hypotheses are difficult to perform within patient samples, but the concordance of TCGA sample-expression derived and cell line-expression derived networks (see above) suggests that experiments in cell lines could be used to address such patient-specific hypotheses in the future.

## 4 DISCUSSION

We describe a new tied diffusion (TieDIE) method able to integrate transcriptional and genomic perturbation data with biological pathway models yielding sub-networks that connect the

two distinct data sources. We demonstrate the ability of TieDIE to generate better precision than single-source diffusion methods in recovering both genes and paths across a wide range of recall, with both a simulated toy example and patient data. Current limitations of the method are described in the *Supplemental Methods* section.

Because many transcription factors are not conventionally regarded as being druggable, approaches such as TieDIE that pinpoint influences upstream of these factors but still in neighborhoods proximal to key driving mutations may provide key starting points for identifying new drug targets. A unique set of genes may be essential for a specific tumor to thrive in its transformed environment – the so-called synthetic-lethal partners to those genes mutated in the tumor. Essential genes would be less likely to appear among a list of "cancer drivers" based on frequently mutated genes because negative selection would eliminate alterations to essential genes. Thus, methods like TieDIE can aid in finding these important genes to potentially inhibit their activity, which can be tested with specific inhibitors. In support of these observations, we found that the TieDIE solutions enrich for genes that are more sensitive to siRNA targeting in cell lines (*Supplemental Results;* Figure S8). Thus, TieDIE might be used to provide a scaffold for simulation of the effects of gene knockouts or drug treatments, potentially improving both computational costs and pathway specificity, when compared with a simulation performed on the entire starting network.

The method can be used to characterize individual tumor samples, which may display unique changes in pathway logic that do not conform to the canonical clinical subtypes. In such cases tumor heterogeneity and tumor evolution may produce a mixture of subtypes. The challenge is to identify which pathways are significantly impacted in complex primary tumors and metastases, and how we can leverage such information to specifically target therapy for a particular patient.

Using data from the TCGA breast cancer cohort, we identified a core-signaling pathway that recapitulates several known aspects of the biology distinguishing the major basal and luminal-A subtypes that was highly concordant with models derived from breast cancer cell lines. In addition, the non-overlapping regions of these pathways lie in areas where we expect to find differences such as tumor metabolism. One exciting possibility would be to compare a patient-specific network to a previously recorded set of networks derived from a bank of stable cell lines grown under different conditions. Exploring therapeutic options for that patient would then involve searching the library of cell lines and growth conditions for those that maximally match the patient's network.

TieDIE holds promise for uncovering pathways relating genomic perturbations to downstream transcriptional changes. Higher precision over single diffusion was obtained both in simulations and in human tumor sample datasets where a large set of genes involved in oncogenesis are known. Even though we chose breast cancer as a test case, the approach can be applied to a wide variety of datasets for both basic molecular biology applications and human disease applications outside of cancer.

## ACKNOWLEDGEMENTS

*Funding*: NIH LINCS Consortium Grant U54HG006097 (http://www.lincsproject.org/); National Cancer Institute TCGA Grant [5U24CA143858].

## REFERENCES

Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks, *Science*, **286**, 509-512.
Beroukhim, R., *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma, *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 20007-20012.
Chin, L., *et al.* (2011) Making sense of cancer genomic data, *Genes & development*, **25**, 534-555.
Ciriello, G., *et al.* (2012) Mutual exclusivity analysis identifies oncogenic network modules, *Genome research*, **22**, 398-406.
Dittrich, M.T., *et al.* (2008) Identifying functional modules in protein-protein interaction networks: an integrated exact approach, *Bioinformatics*, **24**, i223-231.
Forbes, S.A., *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer, *Nucleic acids research*, **39**, D945-950.
Guvakova, M.A. and Surmacz, E. (1997) Overexpressed IGF-I receptors reduce estrogen growth requirements, enhance survival, and promote E-cadherin-mediated cell-cell adhesion in human breast cancer cells, *Experimental cell research*, **231**, 149-162.
Heiser, L.M., *et al.* (2011) Subtype and pathway specific responses to anticancer compounds in breast cancer, *Proceedings of the National Academy of Sciences of the United States of America*.
Huang, S.S. and Fraenkel, E. (2009) Integrating proteomic, transcriptional, and interactome data reveals hidden components of signaling and regulatory networks, *Science signaling*, **2**, ra40.
Joshi-Tope, G., *et al.* (2005) Reactome: a knowledgebase of biological pathways, *Nucleic acids research*, **33**, D428-432.
Kelder, T., *et al.* (2012) WikiPathways: building research communities on biological pathways, *Nucleic acids research*, **40**, D1301-1307.
Lim, W.K., Lyashenko, E. and Califano, A. (2009) Master regulators used as breast cancer metastasis classifier, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 504-515.
Rayala, S.K., Molli, P.R. and Kumar, R. (2006) Nuclear p21-activated kinase 1 in breast cancer packs off tamoxifen sensitivity, *Cancer research*, **66**, 5985-5988.
Schaefer, C.F., *et al.* (2009) PID: the Pathway Interaction Database, *Nucleic acids research*, **37**, D674-679.
Subramanian, A., *et al.* (2005) 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.
TCGA Network (2012) Comprehensive molecular portraits of human breast tumours, *Nature*, **490**, 61-70.
Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116-5121.
Vandin, F., *et al.* (2012) Discovery of mutated subnetworks associated with clinical data in cancer, *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 55-66.
Vandin, F., Upfal, E. and Raphael, B.J. (2012) De novo discovery of mutated driver pathways in cancer, *Genome research*, **22**, 375-385.