PathExt: a general framework for path-based mining of omics-integrated biological networks

Narmada Sambaturu[1], Vaidehi Pusadkar[2], Sridhar Hannenhalli*[3], and Nagasuma Chandra*[1,2]

[1]IISc Mathematics Initiative, Indian Institute of Science, Bangalore - 560012, Karnataka, India
[2]Department of Biochemistry, Indian Institute of Science, Bangalore - 560012, Karnataka, India
[3]Cancer Data Science Laboratory, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA

Abstract

Motivation: Large scale transcriptomic data are routinely used to prioritize genes underlying specific phenotypes. Current approaches largely focus on differentially expressed genes (DEGs), despite the recognition that phenotypes emerge via a network of interactions between genes and proteins, many of which may not be differentially expressed. Furthermore, many practical applications lack sufficient samples or an appropriate control to robustly identify statistically significant DEGs.
Results: We provide a computational tool - PathExt, which, in contrast to differential genes, identifies differentially active paths when a control is available, and most active paths otherwise, in an omics-integrated biological network. The sub-network comprising such paths, referred to as the TopNet, captures the most relevant genes and processes underlying the specific biological context. The TopNet forms a well-connected graph, reflecting the tight orchestration in biological systems. Two key advantages of PathExt are (i) it can extract characteristic genes and pathways even when only a single sample is available, and (ii) it can be used to study a system even in the absence of an appropriate control. We demonstrate the utility of PathExt via two diverse sets of case studies, to characterize (a) Mycobacterium tuberculosis (M.tb) response upon exposure to 18 antibacterial drugs where only one transcriptomic sample is available for each exposure; and (b) tissue-relevant genes and processes using transcriptomic data from GTEx (Genotype-Tissue Expression) for 39 human tissues. Overall, PathExt is a general tool for prioritizing context-relevant genes in any omics-integrated biological network for any condition(s) of interest, even with a single sample or in the absence of appropriate controls.
Availability: The source code for PathExt is available at https://github.com/NarmadaSambaturu/PathExt.
Contact: nchandra@iisc.ac.in, sridhar.hannenhalli@nih.gov

1 Introduction

Whole-genome transcriptomic data are routinely harnessed to probe genes and processes underlying specific biological contexts, including diseases (Blumenberg (2019); Jiang *et al.* (2015)). Extracting biological insights from such high-dimensional data remains an important challenge (Esteve-Codina (2018)). A standard approach to interpreting such data is to first identify differentially expressed genes (DEGs) and then to identify enriched functions among such genes (Esteve-Codina (2018)). However, biological phenotypes emerge from complex interactions among numerous biomolecules, resulting in a highly heterogeneous transcriptional landscape, thus adversely affecting the power to detect critical genes and pathways based on DEGs alone. Moreover, such high-coverage data encodes a vast amount of information beyond DEGs, warranting exploration using multiple complementary approaches. Genome-wide molecular interaction networks constructed from experimentally identified physical, regulatory, signaling, and metabolic interactions have shown great

promise as a framework for integrating and interpreting such data (Sambarey *et al.* (2017a,b)). The identification of sub-networks in such biological networks, which encode the processes perturbed by a stimulus, or active processes in general, can lead to mechanistic insights, as well as help prioritize genes for intervention (Mitra *et al.* (2013)). Several methods have been proposed to integrate transcriptomic data with biological networks, that identify 'active modules' or connected sub-networks which show changes across conditions (Mitra *et al.* (2013)). Despite the availability of interaction data, these methods largely rely on network scoring schemes which prioritize DEGs (Mitra *et al.* (2013)). However, in many practical scenarios including clinical settings, lack of appropriate controls or sufficiently large number of samples preclude robust identification of statistically significant DEGs (Stretch *et al.* (2013)).

In this work, to complement the conventional differential expression-based analyses, we provide PathExt, a path-based approach to mining omics-integrated biological networks. PathExt uses a network weighting scheme that prioritizes edges/interactions rather than nodes/genes, and identifies differentially active paths when comparing conditions, or highly active paths when studying a single condition. The sub-network comprised of these differential paths, referred to as the TopNet, captures the genes and pathways characterizing the biological condition under study. Deviating from traditional approaches to active sub-network identification, PathExt does not use the selection of a connected module as a constraint. Rather, the method results in a well-connected sub-network, reflective of the interconnectedness of biological processes responding to any stimulus.

PathExt can be used to address the following biologically important questions: (i) What are the most significantly differential paths between conditions, and what are the most critical genes underlying the differentially active paths (note that the critical genes themselves may not be differentially active)?; (ii) What is the response to a given perturbation?; and (iii) What are the most active paths and processes in a condition for which there is no appropriate control?

We demonstrate the wide applicability of PathExt by applying it to two diverse sets of case studies. (a) Exposure of the pathogen Mycobacterium tuberculosis to 18 antibacterial drugs, where only one sample is collected for each such exposure. We find that the TopNet for each sample reveals the pathways known to be affected by the corresponding drug. (b) Transcriptomic data for 39 human tissues. Application of PathExt reveals tissue-relevant genes and processes despite the absence of a clear control. In all applications, we find that the TopNet forms a well-connected graph (not expected by chance). Overall, PathExt is a general framework for the integration and analysis of knowledge-based biological networks and omics data, to reveal context-relevant genes and processes. This can be done even with a single sample, or in the absence of appropriate controls. We provide the open source PathExt tool at https://github.com/NarmadaSambaturu/PathExt.

2 Methods

2.1 PathExt

We provide an overview of PathExt in Figure 1. The inputs to PathExt are (a) a directed gene network and (b) gene-centric omics data for the conditions of interest. The omics data can represent a variety of quantities pertaining to the node, such as gene expression level, differential expression, protein, metabolite level, etc., in one or more conditions. The output of PathExt is a sub-network, that we refer to as the TopNet, consisting of the most significant differential or active paths, and is interpreted based on the application context.

PathExt can be used to interrogate any combination of knowledge-based networks and omics data. For clarity, we describe the steps for a protein-protein interaction network (PPIN) and gene expression data. The pipeline consists of the following steps (Figure 1): (1) Integrate inputs, (2) Compute top k shortest paths, (3) Estimate statistical significance of the top k shortest paths, and (4) Construct TopNet by retaining the edges in the significant shortest paths.

STEP 1, Integrate inputs: We integrate the inputs by computing (sample-specific or condition-specific) node and edge weights in the knowledge-based network using the omics data. In the specific scenario when comparing conditions (e.g. pre- and post-treatment), we encode the 'response' of the system to the change in conditions by assigning the node weight as either the fold change in gene expression ($N_i = FC$), or fold change in combination with simple gene expression ($N_i = SI \times FC$). Here N_i is the weight of node i, and SI is the normalized signal intensity, or expression level, of a particular gene. Such a response can be in terms of up-regulated/activated pathways (Activated Response TopNet), obtained by computing

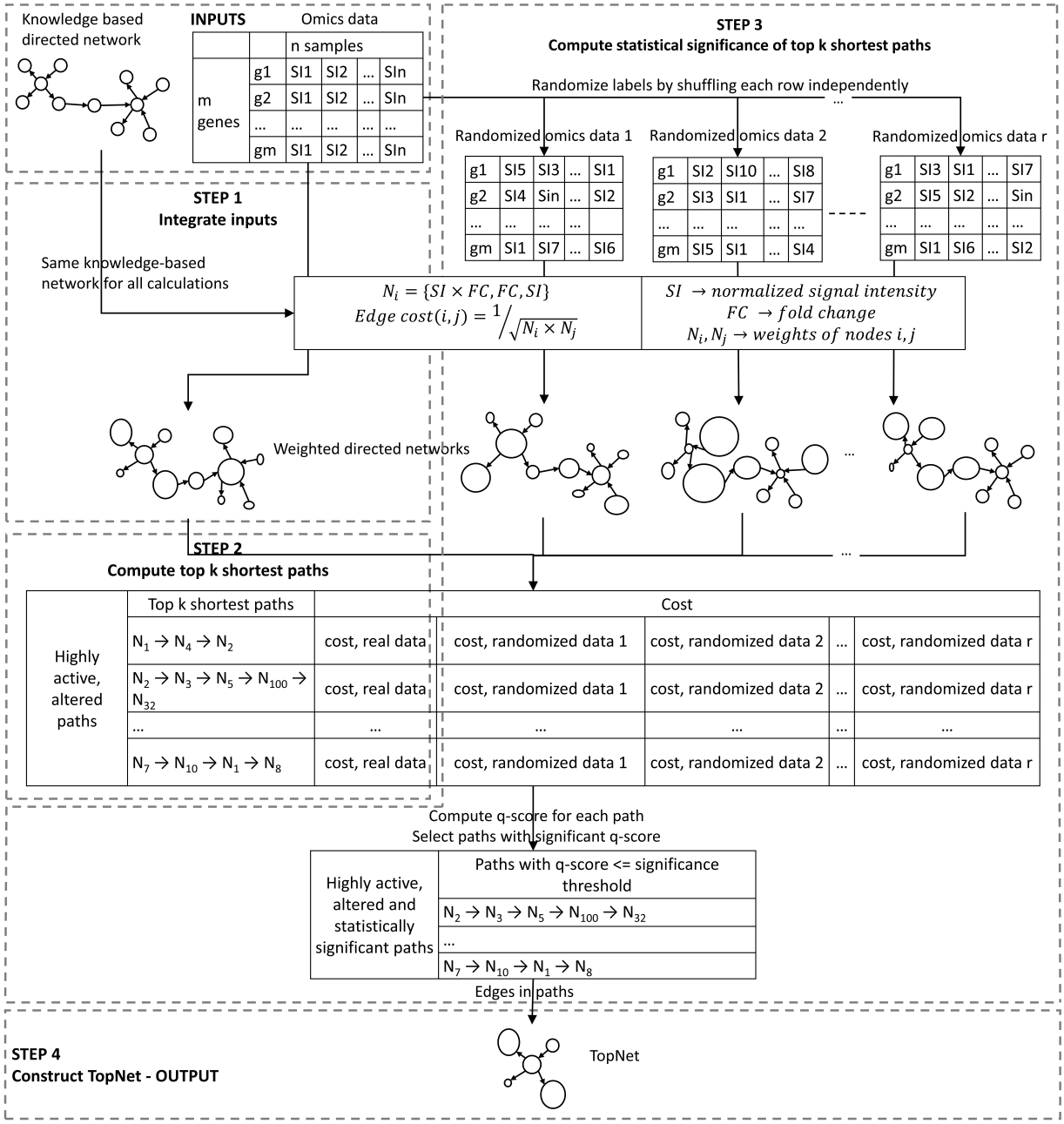

Figure 1: PathExt overview. PathExt uses a knowledge-based directed network and omics data as inputs, and outputs a sub-network consisting of context-relevant genes and processes, referred to as the TopNet.

$FC = SI_{perturbed}/SI_{control}$, or down-regulated/repressed pathways (Repressed Response TopNet), obtained using $FC = SI_{control}/SI_{perturbed}$. The Response TopNet is a union of these two TopNets, and provides a holistic view of the active, altered genes and processes. Exclusively applying the expression value as the node weight ($N_i = SI$) is useful either when no control is available, or when the emphasis is on identifying highly active processes in each state. This TopNet is referred to as the Highest Activity TopNet (HA TopNet). Even in this case, comparisons between states can be carried out after the TopNet is generated for each state.

To assign edge weight, we interpret an edge to represent a 'reaction' between the two nodes, and following the principles of mass action kinetics, an edge between highly abundant nodes is given Edge weight$_{(i,j)}$ = $N_i \times N_j$, where N_i and N_j are the weights of the incident nodes i, j. This choice of edge weight prioritizes highly active interactions in a given context.

STEP 2, Compute top k shortest paths: To achieve a biological outcome, typically a sequence of active reactions is involved, represented by a series of high weight edges in our network. In order to enumerate such high weight paths, we first transform the edge weights into edge costs as Edge cost$_{(i,j)}$ = $1/\sqrt{\text{Edge weight}_{(i,j)}}$, and use Dijkstra's algorithm (Dijkstra *et al.* (1959)) to identify all-pair-shortest-paths. We then normalize the path cost for each node pair by the number of edges along the shortest path to get Normalized path cost = $\frac{\sum_{\text{edges in path}} \text{Edge cost}}{\text{Number of edges in path}}$, and retain the top k shortest paths, where k is a user-defined, application-specific threshold.

STEP 3, Statistical significance of shortest path costs: We assess the statistical significance of the normalized cost of each selected path empirically as follows. Given an $m \times n$ matrix of gene expression data for m genes in n samples/conditions, we randomly shuffle data in each row (gene) independently. The edges are re-weighted with the randomized gene expression data, and the cost of each path from step 2 is computed. After r such randomizations, for each path selected in step 2, r randomization-based costs are computed, based on which a z-score and p-value is estimated for each path. The p-value is finally transformed into a q-value (Benjamini and Hochberg (1995)) to account for multiple hypotheses testing. All paths with significant q-value are retained.

STEP 4, Construct TopNet: The edges in the significant paths from step 3 form a sub-network, which we refer to as the TopNet. The TopNet provides a snapshot of the active and/or significantly altered processes in the system, and can be studied to gain mechanistic insights. To further prioritize critical genes and paths in the TopNet, we apply network centrality measure —Ripple Centrality (Sambaturu *et al.* (2016)).

In cases where a single condition is being examined, or the number of conditions is too small to generate a sufficiently large number of randomized gene expression matrices, step 3 can be skipped, and top k shortest paths can be taken to represent highly active, altered paths, albeit without the statistical filter. In such cases, Step 4 can be directly applied to these paths to generate a TopNet.

2.2 Ripple centrality

Ripple centrality (Sambaturu *et al.* (2016)) prioritizes nodes which can reach a large fraction of the network along highly active and perturbed paths. It is measured as Ripple centrality$(u) = C(u) \times R_{out}(u)$, where $R_{out}(u) = |\text{nodes reachable from } u|$ denotes the outward reachability of node u, and $C(u) = (n-1)/\sum_{v=1}^{n-1} \sigma(u,v)$ gives the closeness centrality of node u. Here $\sigma(u,v)$ denotes the cost of the shortest path from node u to all $n-1$ other nodes in the graph.

2.3 M.tb drug exposure

2.3.1 Data

Transcriptomic data for M.tb H37Rv exposed for 16 hours to 2xMIC of 18 drugs was obtained from GSE71200 (Ma *et al.* (2015)). The list of 18 drugs along with their mechanism of action and TopNet details can be found in Supplementary Table S1. A knowledge-based network composed of experimentally validated protein-protein interactions as well as regulatory interactions in M.tb was obtained from (Mishra *et al.* (2017)), consisting of 3,686 genes and 34,223 edges.

2.3.2 Gold standards

INH is known to affect the mycolic acid synthesis and processing pathways in M.tb (Wishart *et al.* (2017)). To create a gold standard for INH treatment, we searched for the term 'mycolic acid' in Mycobrowser (Kapopoulou et al. (2011)), a database of manually curated annotations for pathogenic mycobacteria, including M.tb. This resulted in a list of 17 M.tb genes, to which we added katG and fas, the known targets of INH (Wishart *et al.* (2017)). Similarly, gold standards were created for 5 other drugs by searching for terms related to their known mechanisms of action - 'RNA polymerase' for Rif, 'mycolic acid' for ethionamide, 'protein synthesis' for capreomycin, and '30s ribosomal protein' as well as '16s rrna' for kanamycin and streptomycin (Wishart *et al.* (2017)) (Supplementary Table S3).

2.3.3 TopNet creation

For all 18 drugs in GSE71200 (Ma *et al.* (2015)), Activated Response TopNets were constructed using $N_i = SI_{drug} \times (SI_{drug}/SI_{control})$, while $N_i = SI_{control} \times (SI_{control}/SI_{drug})$ was used to construct the Repressed Response TopNets. Only shortest paths with 2 or more edges were considered, and 1,000 randomizations of the gene expression matrix were carried out for computing statistical significance of shortest paths. The percentile and q-value thresholds were chosen such that the resulting TopNets were of similar size for all cases (Supplementary Table S1). Activated and Repressed TopNets are provided in Supplementary Files S1 and S2, respectively.

2.3.4 Functional enrichment

Functional enrichment was carried out using ClueGO v2.3.4 (Bindea *et al.* (2009)), a plugin in the network visualization tool Cytoscape 3.2 (Shannon *et al.* (2003)). Enrichment was against GO Biological Processes, GO Cellular Components and GO Molecular Functions, with a q-value cutoff of 0.01. Enriched pathways for all 18 drug exposure cases are provided in Supplementary File S3.

2.3.5 Significance of TopNet connectedness

Significance of TopNet connectedness was tested by comparing against comparable sub-networks induced by (a) the top DEGs, (b) 1,000 sets of randomly sampled genes, and (c) 1,000 sets of randomly sampled edges. Here the number of DEGs and sampled genes (or edges) corresponds to the number of nodes (or edges) in the TopNet.

2.4 Human tissues

2.4.1 Data

Normalized gene expression data was collected from GTEx (Carithers and Moore (2015)) (dbGaP accession number phs000424.v7.p2) for 39 human tissues, corresponding to 23 organs and 2 cell lines. The signal intensities of each tissue were summarized using the LMFit function in R (Limma package; Ritchie *et al.* (2015)). The antilog of the fitted value was used for further analysis as PathExt requires non-negative values. Human protein-protein interaction network (hPPIN) comprising regulatory, signaling and metabolic pathways was obtained from (Sambarey *et al.* (2017a)). This network has 17,062 proteins (nodes) and 208,759 interactions (edges).

2.4.2 TopNet creation

Since no control was available, we constructed two types of TopNets - HA TopNets using $N_i = SI$, and z-score TopNets using $N_i = |z - score|_i$. Here z-score for a gene i in a given tissue was computed with respect to all tissues, and statistical significance of shortest paths was computed by randomizing the $|z - score|$ matrix 1,000 times. The size of the TopNet can vary across tissues and across percentile and false discovery rate thresholds. For the z-score TopNets, we explored percentile thresholds in the range [0.001, 1.0] and q-value thresholds from the set {0.001, 0.005, 0.01, 0.05} in each tissue to adjust the TopNet size to ≈300 nodes. Then for the HA TopNet of each tissue, we explored the same set of percentile thresholds so as to

have a comparable size between HA and z-score TopNets; the percentiles across tissues were either 0.001 or 0.002 in all cases. Thresholds for all tissues are available in Supplementary Table S9. HA TopNets and z-score TopNets for all tissues are provided in Supplementary Files S4 and S5, respectively.

2.4.3 Gold standards

The human protein atlas (HPA; Uhlén *et al.* (2015)), a compiled list of Disease genes (Feiglin *et al.* (2017)), and genes from the Disease Ontology browser of the Mouse Genome Informatics (MGI) database (Bult *et al.* (2018)) were used to validate the results. HPA provides lists of genes whose mRNA expression is elevated in a particular tissue. The elevated expression can correspond to one of three categories: (i) $\geq$ 5-fold mRNA levels in a particular tissue as compared to all other tissues, (ii) $\geq$ 5-fold higher mRNA levels in a group of 2-7 tissues, and (iii) $\geq$ 5-fold higher mRNA levels in a particular tissue as compared to average levels in all tissues. The union of genes from the above three categories form the gold standard. HPA data was downloaded on the 26th of December, 2018. Disease genes were compiled by Feiglin *et al.* (2017) by cross-referencing data from two databases - Online Mendelian Inheritance in Man (OMIM, Hamosh *et al.* (2005)), and the Human Phenotype Ontology (HPO, Köhler *et al.* (2013)). OMIM is a compendium of associations between genetic variations and predominantly Mendelian disorders, while HPO provides a standardized vocabulary for working with such phenotypic abnormalities. The Disease Ontology browser of the MGI lists genes whose mutation is associated with phenotypes characteristic of human disease (Bult *et al.* (2018)). A list of housekeeping genes obtained from Eisenberg and Levanon (2013), comprising of 3,804 genes with constant expression level across a panel of tissues, is used as a negative control to test whether tissue TopNets are enriched in ubiquitously active genes.

2.4.4 Functional enrichment and ranking of pathways

Enrichment was carried out using the enrichGO function of the R package clusterProfiler v3.6.0 (Yu *et al.* (2012)), using Biological Processes as the ontology, and with a Benjamini Hochberg cutoff of 0.01. For each tissue, the background for enrichment was set to be the list of genes for which both expression and interaction data were available. Pathway enrichment results for HA TopNets, z-score TopNets, their corresponding baselines, gold standards, as well as housekeeping genes, are provided for all tissues in Supplementary File S6. Pathways enriched in the TopNets were ranked based on the cost of the first TopNet shortest path involving a gene from that pathway. Ties were broken based on the fold enrichment of TopNet genes in a pathway relative to expectation.

3 Results

3.1 PathExt reveals pathways related to drugs' mechanism of action in treated M.tb

In a previous study, the *Mycobacterium tuberculosis* (M.tb) strain H37Rv was exposed to different concentrations of various anti-tuberculosis drugs, and the transcriptional response was measured (GEO accession number GSE71200; Ma *et al.* (2015)). We obtained the transcriptomic data for 2xMIC (twice the minimum inhibitory concentration) dose of 18 drugs, for bacteria surviving 16 hours of drug exposure, suggesting a degree of drug resistance. Only one replicate per MIC per drug and a single untreated control sample were measured, making robust estimation of differential expression impractical. For 6 drugs where the mechanism of action is well studied (Wishart *et al.* (2017)), we obtained gold standard sets of genes experimentally verified to be perturbed upon drug exposure (Methods section 2.3.2). In all 6 cases, the Response TopNets generated by PathExt are concordant with the gold standards, and reveal genes and pathways relevant to the action of each drug (Table 1). In contrast, the genes with 1.5-fold differential expression have consistently poor overlap with gold standards (Table 1). We discuss the Isoniazid and Rifampicin exposures in detail below.

Drug	Gold standard	Activated Response TopNet			Repressed Response TopNet			Response TopNet			1.5 FC DEGs			
		Nodes	Gold standard	p-value	Nodes	Gold standard	p-value	Nodes	Gold standard	p-value	DEGs	Base network	Gold standard	p-value
Capreomycin	29	184	2	0.45	195	5	0.02	372	7	0.03	1796	1676	16	0.26
Ethionamide	15	201	5	0.001	202	0	1	394	5	0.02	610	562	6	0.02
Isoniazid	17	195	6	0.0002	213	5	0.002	401	11	2.2e-7	1078	987	8	0.07
Kanamycin	30	191	5	0.02	199	2	0.5	379	7	0.03	790	731	4	0.23
Rifampicin	22	196	7	0.0001	197	4	0.03	380	11	4.3e-6	1579	1462	13	0.07
Streptomycin	30	193	7	0.0009	181	4	0.06	338	10	0.0003	308	286	1	0.29

Table 1: Response TopNets for M.tb exposed to 6 drugs whose mechanism of action is well known, are concordant with gold standards and reveal genes relevant to the action of each drug.

3.1.1 PathExt links INH exposure to mycolic acid synthesis and processing

The anti-bacterial drug Isoniazid (INH) inhibits the synthesis of mycolic acids, which are long fatty acids found in the cell walls of mycobacteria (Wishart *et al.* (2017)). The Activated Response TopNet (selecting for up-regulated paths), Repressed (down-regulated paths), and merged Response TopNets (Methods) identified by PathExt were all significantly enriched in gold standard genes related to mycolic acid synthesis and processing (Table 1, Supplementary Table S1). In stark contrast, the DEGs with $\geq$ 1.5-fold differential expression had poor overlap with the gold standard (Table 1, Supplementary Table S1). The central genes (Methods section 2.2) in the Activated Response TopNet consist mainly of genes involved in mycolic acid biosynthesis, whereas the Repressed Response TopNet has unsaturated acyl-CoA hydratases responsible for oxidizing fatty acids, and genes involved in lipid degradation as the central nodes (Supplementary Tables S4 and S5). These results unambiguously point to the up-regulation of fatty acid synthesis and down-regulation of its degradation as a resistance response to INH exposure.

A previous study (Takayama *et al.* (2005)) consolidated experimental and computational evidence to list the 7 main processes in the mycolic acid synthesis and processing pathway, namely, the FAS-I (fatty acid synthetase-I) system, transition from the FAS-I system to the FAS-II system, the FAS-II system, cyclopropane synthases and methyltransferases, oxidation-reduction, Claisen-type condensation, and mycolic acid processing. Of the 42 genes described in their work, interaction and expression data were available for 39, of which 16 were present in the INH exposed Response TopNet (3.63 fold enrichment; Fisher's p-value = 1.68e-6), while the 1,078 DEGs comprise only 14 of these genes (Fisher's p-value = 0.27). Notably, the TopNet sub-network induced by the 16 genes from the mycolic acid synthesis and processing pathway (Figure 2) and their immediate neighbors, represent all 7 component processes. Interestingly, NADH dehydrogenases (highlighted in violet in Figure 2) are also picked up in this sub-network. It has been hypothesized that M.tb may gain resistance to INH by regulating NADH dehydrogenase and the intracellular NADH/NAD+ ratio (Miesel *et al.* (1998)). This is consistent with the fact that the bacteria under study are the ones which survived exposure to 2xMIC of INH and thus likely to have triggered their resistance processes.

Finally, as an additional control, we directly compared the Response TopNet genes with same number of top DEGs in terms of their functional enrichment (Figure 2, Methods section 2.3.4). The genes in the Response TopNet are enriched in the functional terms relevant to INH exposure, such as *cell periphery*, which is the part of the cell most affected by INH (Wishart *et al.* (2017)), and stress response terms such as *oxidoreductase activity* and *oxidation-reduction process*. We also find the term *regulation of metabolic*

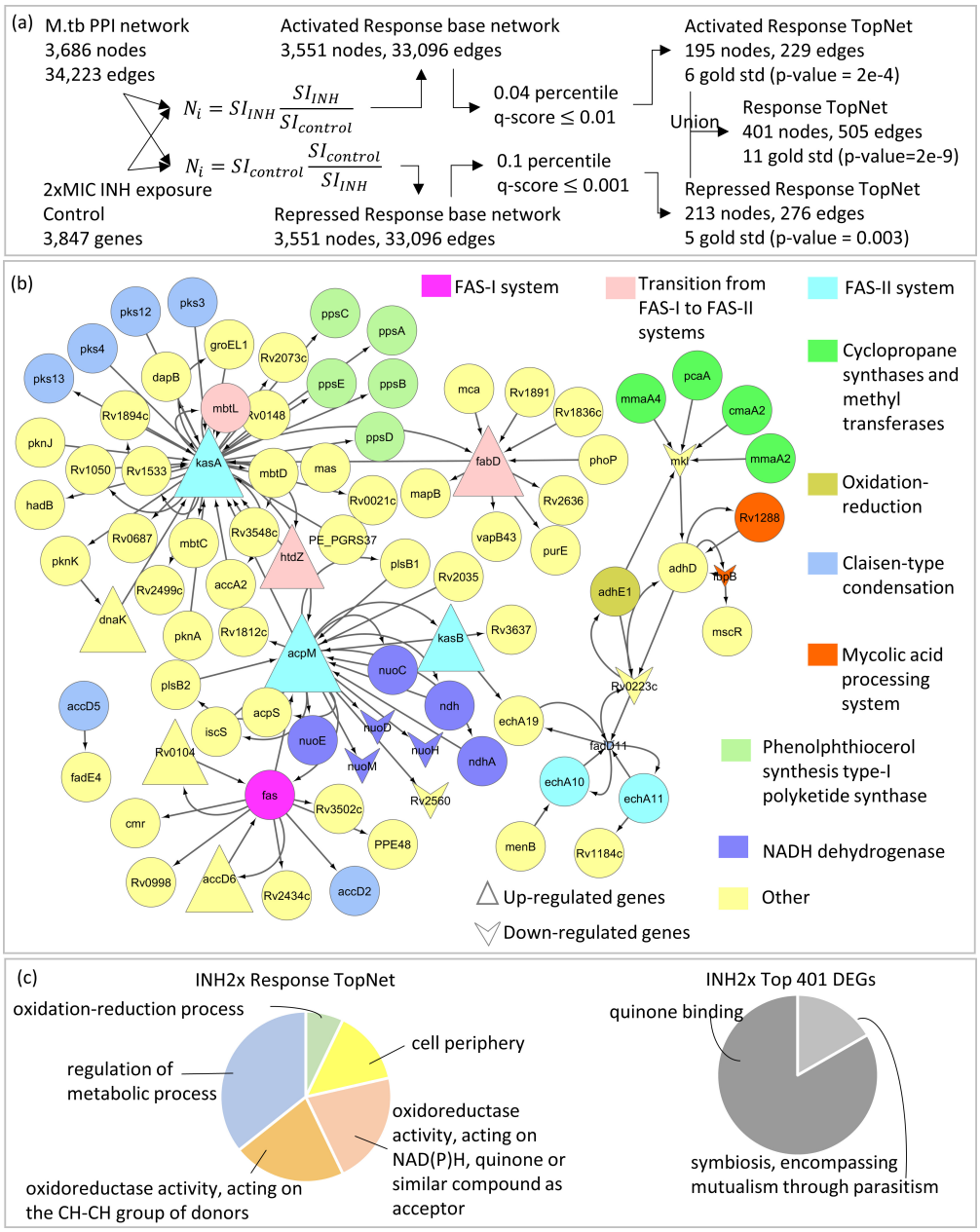

Figure 2: Response to 2xMIC INH. (a) Gene expression data for a single sample of M.tb exposed to 2xMIC of INH for 16 hours is integrated with a knowledge-based protein-protein interaction network for M.tb using two weighting schemes. $N_i = SI_INH \times (SI_{INH}/SI_{control})$ prioritizes genes up-regulated after exposure to INH, and results in an Activated Response TopNet. $N_i = SI_{control} \times (SI_{control}/SI_{INH})$ prioritizes genes and processes down-regulated after exposure to INH, and results in a Repressed Response TopNet. Thresholds for the top k shortest paths and statistical significance are chosen such that the two TopNets are of comparable sizes. The union of the two TopNets gives a Response TopNet. All three TopNets are enriched in gold standard genes. (b) Sub-network of the Response TopNet formed by extracting genes from the mycolic acid synthesis and processing pathway (Takayama *et al.* (2005)), the known target pathway of INH, and their immediate interactors. Every component process of this pathway is represented in the Response TopNet by at least 1 gene. (c) GO enrichment of Response TopNet gives pathways relevant to INH exposure, such as cell-periphery and oxidation-reduction process. Enrichment of an equal number of top DEGs does not provide drug-specific insights.

processes, which is an expected energy conservation response. In contrast, the top 401 DEGs are enriched for the terms *quinone binding* and *symbiosis encompassing mutualism through parasitism*, which are not informative of the condition under study. Together, these results show that the Response TopNet for M.tb exposure to 2xMIC of INH is indeed characteristic of its action and reveals genes and processes that would be missed by a conventional approach relying on differential gene expression alone.

3.1.2 Rif exposure TopNet reveals the perturbation of nucleotide synthesis pathway

Rifampicin (Rif) inhibits DNA-dependent RNA polymerase activity, thus suppressing transcriptional initiation (Wishart *et al.* (2017)). Once again, the Activated, Repressed and union Response TopNets are enriched in gold standard genes, whereas the DEGs are not (Table 1, Supplementary Table S1). The gene rpoB (Rv0667) is central in the Activated Response TopNet, effectively recapitulating previous reports which suggest that Rif resistance can be caused by transcriptional up-regulation of rpoB (Zhu *et al.* (2018)). The error prone DNA repair synthesis protein DnaE2 (Rv3370c), and the genetic recombination and nucleotide excision repair protein RecA (Rv2737c) are also central in this network. Exposure to antibiotics such as Rif has been shown to result in a recA-dependent SOS response, and a corresponding increase in dnaE2 levels (McGrath *et al.* (2013)). Also, the up-regulation of dnaE2 has been identified as a critical factor in the emergence of drug resistance both *in-vitro* and *in-vivo* (Boshoff *et al.* (2003)). Other central genes (full list in Supplementary Table S4) include the 16S ribosomal RNA methyltransferase Rv2372c, and the replicative DNA helicase dnaB (Rv0058). These genes reflect perturbations in the nucleotide synthesis pathway, the very pathway known to be affected upon exposure to Rif. Central genes in the Repressed Response TopNet include, among others, dnaK (Rv0350) and Rv0232, a transcriptional regulator of the tetR/acrR-family. Disruption of Rv0232 has been shown to provide a growth advantage to H37Rv *in-vitro* (DeJesus *et al.* (2017)). We found that Rv0232 was 4.5-fold down-regulated and centrally involved in repressed paths, suggesting this as a possible resistance mechanism.

Interestingly, dnaK is central in the Repressed Response TopNet for Rif, whereas it is central in the Activated Response TopNet for INH exposure. It has been shown that dnaK is repressed by Rif (Eltringham *et al.* (1999)), whereas cells with higher levels of dnaK are more likely to persist upon exposure to INH (Jain *et al.* (2016)). This result underscores the biological and mechanistic relevance, as well as the condition-specificity of the TopNets generated by PathExt.

Although the exact pathway for DNA-dependent RNA polymerase activity is not known, examining the central genes from the Rif Activated and Repressed Response TopNets along with their immediate interactors provides valuable insights. These genes form two connected components, connected by two linker genes, fadE18 (Rv1933c) and fadD11 (Rv1550) (Figure 3). This sub-network highlights three major processes, namely, (i) transcription and nucleotide synthesis, (ii) error-prone synthesis and repair, and (iii) lipid metabolism. Figure 3 also shows the GO-term based enrichment of the genes in the Response TopNet, and for an equal number of top DEGs. The genes in the Response TopNet are enriched for terms relevant to exposure to Rifampicin, such as *translation*, which is the process targeted by Rif, *plasma membrane* and *acyl-CoA dehydrogenase activity*, which are related to lipid metabolism. On the other hand, the 380 top DEGs are enriched for the terms *cell periphery* and *plasma membrane*, which are not specifically informative of cellular response to the drug.

As demonstrated by the INH and Rif case studies, each Response TopNet reveals drug-specific mechanisms. Drug-specificity of the TopNets is further emphasized by the fact that there is no node or edge common to all 18 Response TopNets, despite the same knowledge-based network being used as input in all cases.

The Response TopNet is a connected graph with $> 50\%$ nodes in the largest component in each of the 18 drug exposures. This connectedness, reflective of biological pathways, is shown to be non-random (Methods section 2.3.5), and not captured by the sub-networks induced by top DEGs (Table 2, Supplementary Table S2). This suggests that our Response Network captures crosstalk between the dysregulated paths, which simple differential gene expression analysis may not.

Taken together, these results show that PathExt captures drug-specific responsive genes and processes, even when only a single sample was available per condition.

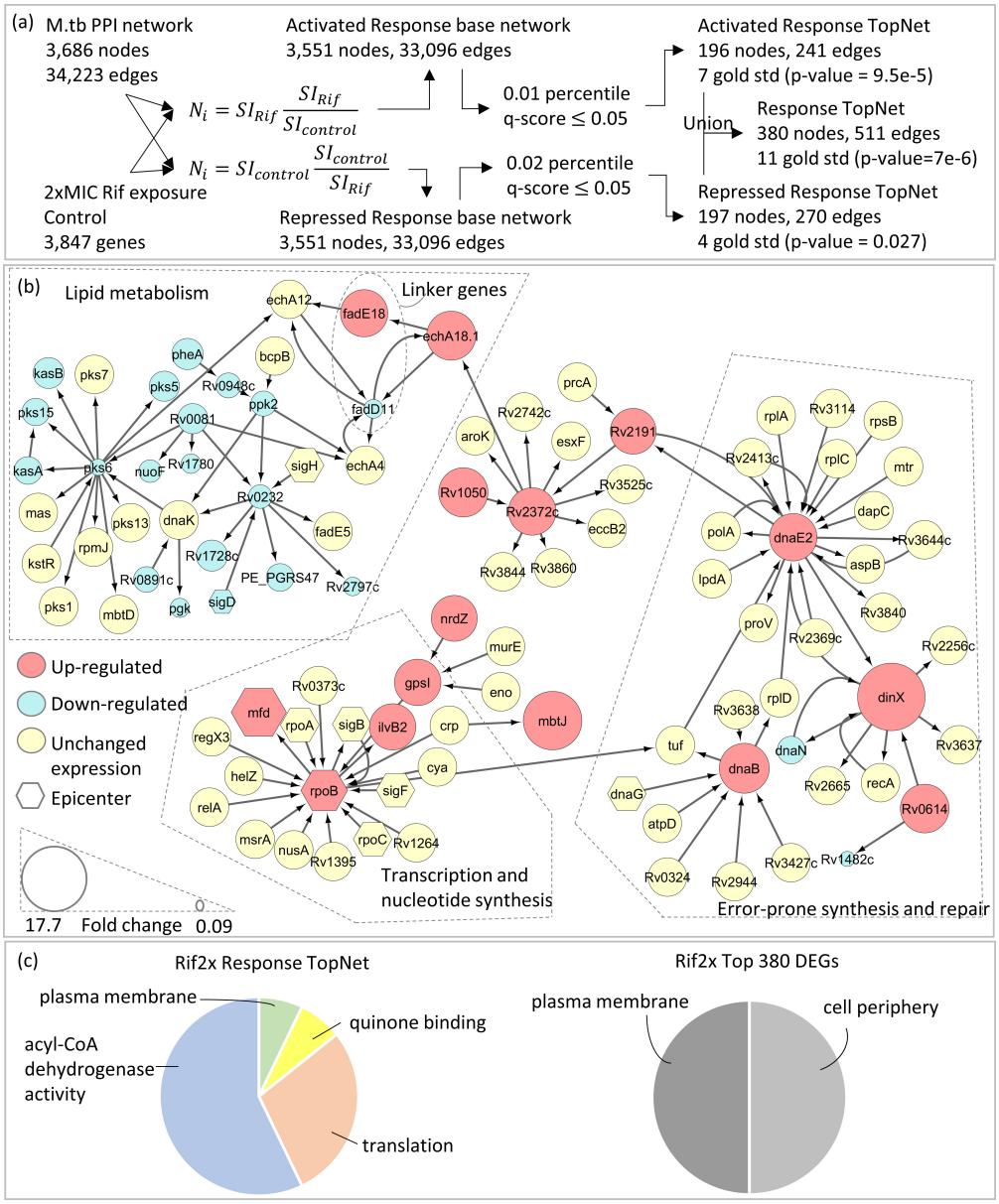

Figure 3: Response to 2xMIC Rif. (a) Gene expression data for a single sample of M.tb exposed to 2xMIC of Rif for 16 hours is integrated with a knowledge-based protein-protein interaction network for M.tb using two weighting schemes. $N_i = SI_{Rif} \times (SI_{Rif}/SI_{control})$ prioritizes genes up-regulated after exposure to Rif, and results in an Activated Response TopNet. $N_i = SI_{control} \times (SI_{control}/SI_{Rif})$ prioritizes genes and processes down-regulated after exposure to Rif, and results in a Repressed Response TopNet. Thresholds for the top k shortest paths and statistical significance are chosen such that the two TopNets are of comparable sizes. The union of the two TopNets gives a Response TopNet. All three TopNets are enriched in gold standard genes. (b) Central genes from Activated and Repressed Response TopNets as well as their immediate interactors, extracted from the union Response TopNet for Rif. The inclusion of two linker genes, fadE18 and fadD11, links the resulting components into a single connected component. The size of the nodes reflects the extent of dysregulation of the genes. Up-regulated genes are colored red, while down-regulated genes are blue. This module contains genes related to transcription and nucleotide synthesis, the known target pathway of Rif (Wishart *et al.* (2017)). Other pathways represented here are lipid metabolism and error-prone synthesis and repair, both known mechanisms of resistance to Rif (Howard *et al.* (2018); Boshoff *et al.* (2003)). (c) GO enrichment of Response TopNet gives pathways relevant to Rif exposure, such as translation. Enrichment of an equal number of top DEGs does not provide drug-specific insights.

Drug	Accession number	Response TopNet				Top k DEGs induced CC	Node sampling mean CC	Edge sampling mean CC
		Nodes	Edges	CC	Largest CC%			
Capreomycin	GSM1829654	372	518	4	97.6	159	181.8	205.9
Cycloserine	GSM1829656	378	486	2	98.4	172	183.4	206.4
Ethionamide	GSM1829659	394	516	3	90.1	151	187.2	207.3
LY 83583	GSM1829661	357	473	1	100	155	179.7	205.1
Aminothiadiazole	GSM1829664	352	434	2	96.9	131	178.4	202
Tunicamycin	GSM1829667	380	484	2	97.1	145	184.4	206
Acivicin	GSM1829683	359	432	3	96.4	136	180.3	201.2
Fenamisal	GSM1829685	377	466	1	100	157	183.6	204.4
Iodotubercidin	GSM1829687	385	450	6	92.7	161	185.3	202
Disulfiram	GSM1829690	383	685	7	55.1	128	184.4	207.2
Methoxy-methylellipticinium	GSM1829709	378	490	4	97.4	175	182.1	197.8
Chloroxine	GSM1829712	357	432	2	98.6	156	179.4	201.2
Isoniazid	GSM1829740	401	505	1	100	211	188.1	207
Kanamycin	GSM1829743	379	483	3	96.6	149	185.3	205.7
Moxifloxacin	GSM1829746	383	513	5	86.2	155	185	207.1
Pretomanid (PA-824)	GSM1829749	368	469	5	93.2	187	179.3	197.1
Rifampicin	GSM1829752	380	511	6	92.9	196	184.7	207.4
Streptomycin	GSM1829755	338	463	3	93.2	172	175.9	204.2

Table 2: Response TopNets of all 18 drugs are connected graphs with > 50% nodes in a single connected component (CC). In contrast, the graphs induced by the same number of top DEGs, and by randomly sampling the same number of nodes or edges as the TopNet are highly disconnected. All randomizations are carried out 1,000 times.Standard deviations, z-scores and p-values are provided in Supplementary Table S2.

3.2 Human tissue TopNets reveal tissue-related genes and processes

In a second set of case studies, we applied PathExt to identify tissue-related pathways using gene expression data for 39 human tissues in GTEx (Carithers and Moore (2015)), corresponding to 23 organs and 2 cell lines. In this scenario, there is no control. Therefore, we constructed two types of TopNets independently in each tissue (Methods section 2.4.2). A Highest Activity TopNet (HA TopNet) where $N_i = SI$, and a z-score TopNet where $N_i = |z - score|_i$. Here N_i is the weight of node i, and SI is the normalized signal intensity (expression level). The z-score for gene i in a given tissue is computed relative to all tissues, thus using all tissues as a control for each tissue.

We assessed the tissue-specific TopNets against three gold-standards (Methods): (1) the Human Protein Atlas (HPA) (Uhlén *et al.* (2015)) where genes with ≥5-fold higher abundance in each tissue are labelled tissue-specific, (2) a set of curated tissue-relevant Disease genes (Feiglin *et al.* (2017)), and (3) a list of genes associated with tissue-specific human diseases from the MGI (Bult *et al.* (2018)). These comparisons are carried out for 37 out of 39 tissues, as corresponding gold standards could not be obtained for the 2 cell lines. We also use a list of housekeeping genes (Eisenberg and Levanon (2013)) as a negative control. To assess the utility of the z-score TopNets, we use the same number of genes with the highest $|z - score|$ as a baseline control. Likewise, for the HA TopNets, the baseline used is the set of genes with highest expression levels.

The MGI had ≥ 25 genes with both gene expression and interaction data for 5 tissues. Of these, the HA TopNets were significantly enriched in tissue-associated genes in 3 tissues, and z-score TopNets in 4 tissues

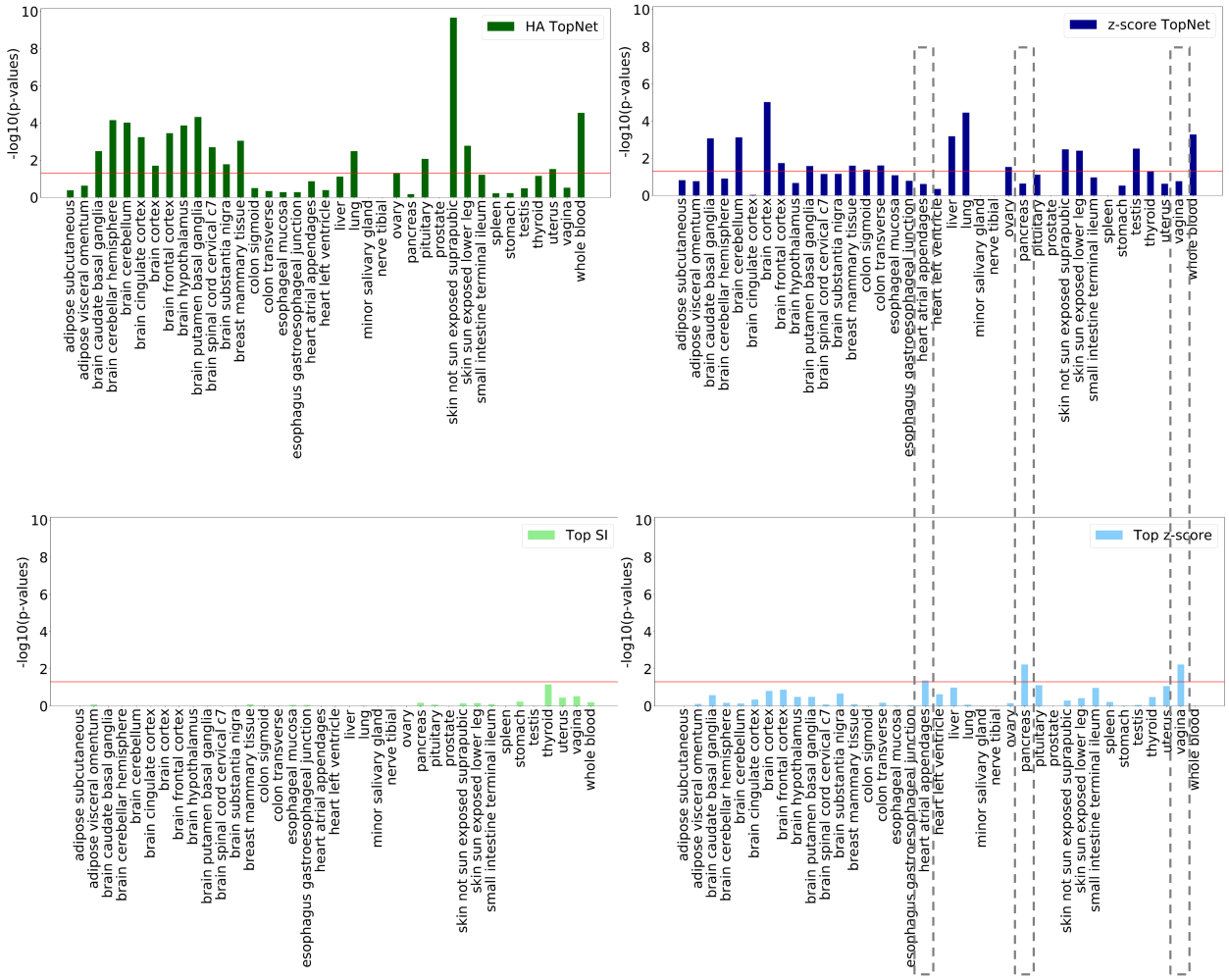

Figure 4: Overlap with Disease genes. p-values for overlap between the gold standard Disease genes, and nodes from (a) HA TopNets, (b) z-score TopNets, and their corresponding baseline controls, (c) Top SI (genes with highest expression) and (d) Top z-score. The p-values are plotted along the Y-axis as $-log_{10}(p-value)$, such that a taller bar corresponds to a more statistically significant overlap. The red horizontal line corresponds to p-value = 0.05. Tissues are plotted along the X-axis in lexicographic order. In every case except the three highlighted using dashed boxes, the overlap between TopNet nodes and the gold standard is better than the corresponding baseline. For pancreas and vagina, the genes with top z-score agree better with Disease genes than the genes in the z-score TopNet. These tissues are explored in greater detail in Supplementary Info S1.

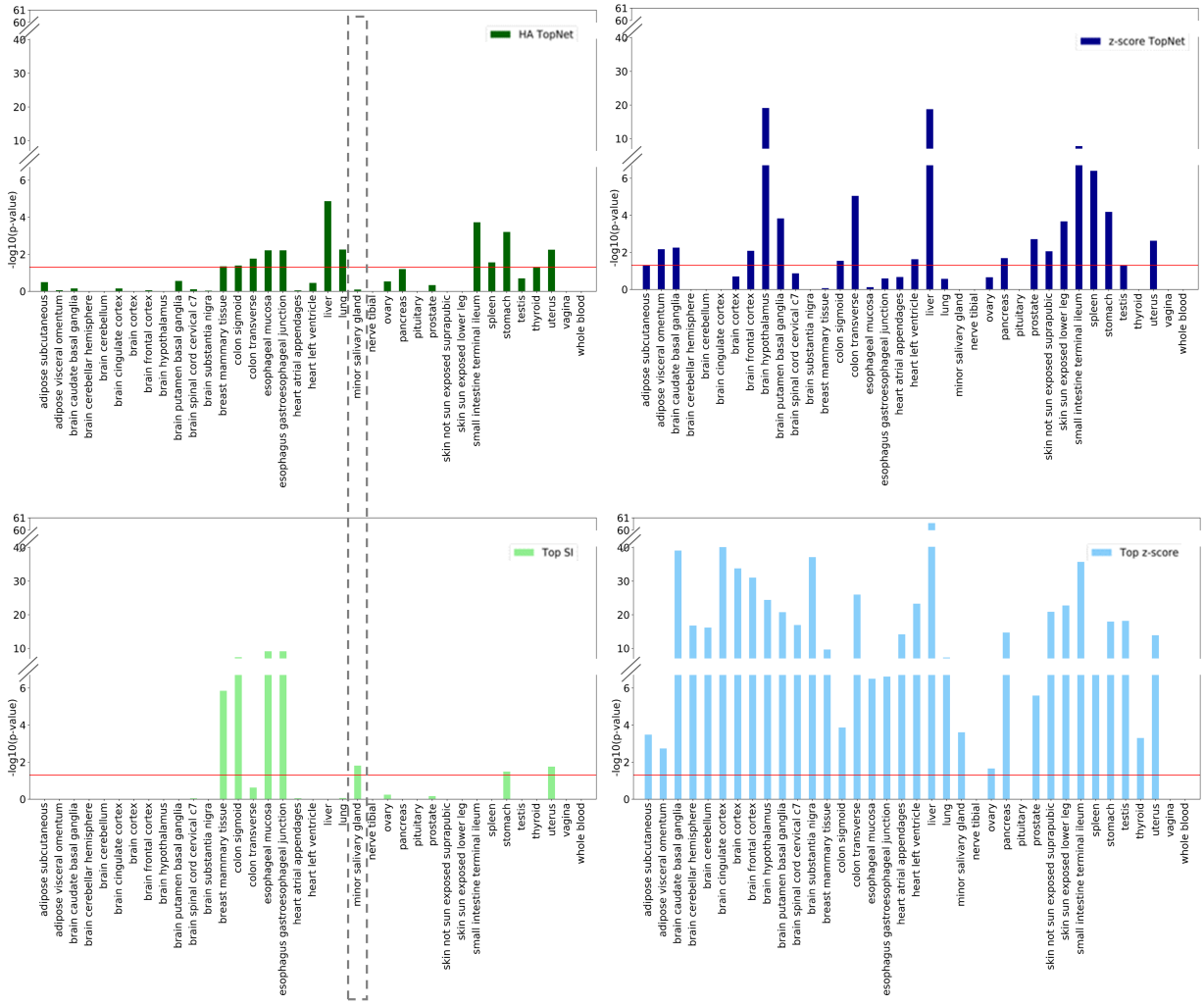

Figure 5: Overlap with HPA. p-values for overlap between the gold standard HPA and nodes from (a) HA TopNets, (b) z-score TopNets, and their corresponding baseline controls, (c) Top SI (genes with highest expression) and (d) Top z-score. The p-values are plotted along the Y-axis as $-log_{10}(p - value)$, such that a taller bar corresponds to a more statistically significant overlap. The red horizontal line corresponds to p-value = 0.05. Tissues are plotted along the X-axis in lexicographic order. Since genes with ≥5-fold differential expression are labelled tissue-specific in HPA, the genes with top z-score always have statistically significant overlap with this gold standard, as expected. Notwithstanding this comparison, the overlap between HA TopNet nodes and the gold standard is better than the corresponding baseline in all but one case (highlighted with dashed box); discussed in Supplementary Info S1.

(Supplementary Table S6). In every case, the TopNet picked up equal or more gold standard genes than the corresponding baseline.

Figures 4 and 5 show the Fisher's p-value of the overlap between the genes in the TopNets (and their corresponding baselines) and the two gold standards HPA and Disease genes. Since HPA is constructed based on differential abundance, as expected, genes with top z-score are highly concordant with the HPA-derived tissue-specific genes. In all other comparisons across tissues, genes identified by PathExt agree better with gold standards than the corresponding baselines. We found 4 exceptions out of 74 comparisons (37 tissues x 2 gold standards), marked with dashed boxes in Figures 4 and 5. Even in these cases, the pathways enriched in the TopNets are relevant to the functions of that tissue (Supplementary Info S1).

An ideal tissue-specific network should exclude housekeeping genes, which by their very definition are broadly active. We find that the TopNets identified by PathExt have this property, and are under-enriched in housekeeping genes in all but 1 case (Supplementary Table S6). This suggests that the paths prioritized by PathExt correspond to tissue-related functions rather than universally active processes.

3.2.1 PathExt-identified pathways enriched exclusively in a tissue correspond to known tissue-relevant functions

Figure 6 shows the top pathway exclusively enriched in the HA TopNet of each of 32 tissues (Methods section 2.4.4). The 7 excluded tissues had no exclusively enriched pathways. Figure 6 also highlights additional significant (0.01 ¡ q-value $\leq$ 0.05 and 0.05 ¡ q-value $\leq$ 0.1) pathway-tissue pairs. We found direct literature evidence supporting 20 out of 32 exclusive pathway-tissue associations, and indirect evidence for an additional 9 (Supplementary Table S7). A similar table along with literature evidence for the pathways enriched in the z-score TopNets is provided in Supplementary Table S8.

Some of the pathway-tissue pairs correspond to well-established functions of the tissue, such as *regulation of bile acid metabolic process* in liver (Chiang (2013)), *ethanol catabolic process* in lung (Bernstein (1982)), etc. PathExt reveals a few surprising associations as well. *Sensory perception of smell* is the top pathway exclusively enriched in the testis. At first glance this seems counter-intuitive. However, prevalence of olfactory receptors in the testis and sperm has been experimentally verified, and testicular olfactory receptor signaling has been implicated in sperm flagellar motility (Kang and Koo (2012)). As another example, *regulation of rhodopsin mediated signaling pathway* is enriched exclusively in the pancreas. Interestingly, rhodopsin regulates insulin receptor signaling in rod photoreceptor neurons (Rajala and Anderson (2010)), and loss of Arf4, a GTPase important for localizing rhodopsin to the eye and kidney, has been shown to result in damage of exocrine pancreas in mice (Pearring *et al.* (2017)). This surprising link between rhodopsin and the pancreas is not picked up by any of the gold standards or the controls.

Figure 6 also highlights the specificity of functions of the different regions of the brain. For instance, the *gamma-delta T cell activation pathway* is enriched in the brain cortex. Gamma-delta T cells have been implicated in Rasmussen encephalitis, a disease characterizing inflammation of the cerebral cortex (Owens *et al.* (2015); Varadkar *et al.* (2014)). The *adenylate cyclase-activating dopamine receptor signaling pathway* is enriched exclusively in the putamen basal ganglia region of the brain. The dorsal region of the basal ganglia comprises of the putamen, and the caudate nucleus (Lanciego *et al.* (2012)). Experiments involving homogenates of the caudate nucleus of the rat brain point at dopamine-sensitive adenylate cyclase as the receptor for dopamine in the mammalian brain (Kebabian *et al.* (1972)). This finding could indicate the presence of caudate nucleus cells in the putamen sample, or a shared function between these two adjacent regions of the brain. Several processes expected to be ubiquitous, such as *regulation of receptor activity* and *response to extracellular stimulus*, are enriched in all the tissues under consideration.

Overall, PathExt-identified tissue-specific TopNets recapitulate gold standard genes with known tissue-specific functions, and provide unique insights into tissue functions, not reflected in conventional differential expression-based analyses.

4 Discussion

We provide PathExt, a computational tool to identify sub-networks of an omics-integrated biological network, which capture the response to a perturbation, or the active processes in a particular condition. PathExt builds on our prior work which mined omics-integrated networks to (i) identify tuberculosis biomarkers

	adipose subcutaneous	adipose visceral omentum	brain caudate basal ganglia	brain cerebellar hemisphere	brain cerebellum	brain cortex	brain frontal cortex	brain hypothalamus	brain putamen basal ganglia	brain spinal cord cervical c-1	brain substantia nigra	breast mammary tissue	cells EBV transformed lymphocytes	cells transformed fibroblast	colon sigmoid	colon transverse	heart atrial appendages	heart left ventricle	liver	lung	minor salivary gland	pancreas	pituitary	prostate	skin not sun exposed suprapubic	skin sun exposed lower leg	small intestine terminal ileum	spleen	stomach	testis	thyroid	uterus
positive regulation of CREB transcription factor activity	**8.72E-03**	5.33E-01	4.67E-01	1.53E-01	1.02E-01	5.64E-01	4.13E-01	1.01E-01	4.03E-01	1.00E+00	5.42E-01	5.31E-01	4.06E-01	1.00E+00	1.00E+00	1.00E+00	4.49E-01	3.11E-01	3.72E-01	5.62E-01	5.36E-01	1.00E+00	1.18E-01	1.00E+00	4.60E-01	5.68E-01	1.00E+00	1.00E+00	1.00E+00	6.11E-01	1.55E-01	1.00E+00
central nervous system neuron development	9.53E-01	**8.38E-03**	1.00E+00	6.55E-01	8.49E-01	1.00E+00	1.00E+00	8.41E-01	1.00E+00	8.96E-01	1.00E+00	4.60E-02	7.43E-01	8.25E-01	6.51E-01	5.13E-01	3.05E-01	6.50E-01	4.49E-01	6.96E-01	1.33E-01	2.46E-01	4.12E-01	4.16E-01	5.82E-01	4.98E-01	1.26E-01	6.74E-01	2.78E-01	4.32E-01	8.55E-02	4.16E-01
negative regulation of monooxygenase activity	2.77E-02	4.60E-01	**9.69E-03**	1.00E+00	6.92E-02	1.35E-01	5.68E-02	3.47E-01	5.79E-02	4.98E-01	1.19E-01	1.08E-01	1.00E+00	1.00E+00	1.00E+00	1.31E-01	3.90E-01	4.06E-02	1.00E+00	1.00E+00	4.52E-01	1.00E+00	1.00E+00	4.57E-01	3.84E-01	5.19E-01	3.70E-01	1.00E+00	1.00E+00	1.76E-01	4.63E-01	1.00E+00
chondroitin sulfate metabolic process	1.56E-02	7.57E-01	4.24E-01	**1.19E-03**	1.03E-02	1.00E+00	1.00E+00	3.65E-01	1.00E+00	1.00E+00	5.42E-01	5.26E-01	1.24E-01	7.21E-01	7.99E-01	8.29E-01	1.70E-01	1.00E+00	1.00E+00	3.68E-02	8.26E-01	1.00E+00	1.85E-01	7.67E-01	1.00E+00	8.27E-01	1.00E+00	1.00E+00	6.80E-01	8.50E-01	7.92E-01	3.80E-01
positive regulation of lipoprotein lipase activity	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**7.94E-03**	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.37E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.51E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
gamma-delta T cell activation	1.00E+00	1.53E-01	4.38E-01	4.56E-01	9.02E-02	**4.39E-04**	7.44E-02	3.83E-01	3.74E-01	5.34E-01	5.25E-01	5.13E-01	1.38E-02	1.00E+00	1.00E+00	1.65E-01	4.29E-01	3.00E-01	8.65E-02	1.00E+00	3.36E-02	2.52E-01	4.48E-01	2.49E-02	1.72E-02	1.78E-01	1.00E+00	1.00E+00	4.08E-01	5.77E-01	5.04E-01	2.25E-01
cellular protein-containing complex assembly	7.78E-01	8.50E-01	2.87E-02	2.72E-01	1.16E-01	5.50E-02	**7.95E-03**	1.62E-02	7.53E-02	5.96E-01	5.08E-01	2.57E-01	4.76E-01	1.64E-01	2.08E-01	1.11E-01	2.53E-01	4.07E-02	7.77E-01	1.00E+00	4.52E-01	5.48E-02	9.56E-01	5.27E-01	6.74E-02	5.35E-01	8.90E-01	9.69E-01	5.32E-01	3.20E-01	8.57E-01	5.38E-01
regulation of protein localization to plasma membrane	4.80E-01	8.23E-01	4.73E-01	5.48E-01	7.63E-01	8.26E-01	7.05E-01	**9.72E-03**	8.23E-01	5.56E-01	6.28E-01	8.54E-01	4.92E-01	7.81E-01	8.55E-01	6.36E-01	4.88E-01	7.61E-01	4.70E-01	4.52E-01	6.38E-01	3.94E-01	5.30E-01	7.64E-01	2.54E-01	8.74E-01	7.55E-01	1.00E+00	4.93E-01	6.67E-01	6.03E-01	7.28E-01
adenylate cyclase-activating dopamine receptor signaling pathway	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**3.58E-04**	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	2.81E-02	1.00E+00	1.00E+00
monocarboxylic acid catabolic process	1.10E-01	8.95E-01	9.50E-01	6.12E-02	1.37E-01	9.76E-01	7.47E-01	9.48E-01	6.96E-01	**6.36E-03**	4.80E-01	9.27E-01	1.07E-01	6.45E-01	4.49E-01	1.00E+00	6.64E-01	1.92E-01	5.01E-02	1.56E-01	6.67E-01	9.09E-01	8.67E-01	2.09E-02	1.73E-01	5.11E-01	1.00E+00	5.97E-01	9.49E-01	4.71E-01	4.18E-01	9.18E-01
regulation of protein kinase A signaling	3.24E-01	5.92E-01	5.23E-01	1.00E+00	5.02E-01	1.00E+00	3.14E-02	1.00E+00	2.75E-02	1.00E+00	**2.98E-03**	2.59E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.89E-01	1.40E-01	1.00E+00	1.00E+00	7.08E-01	1.99E-01	1.00E+00	1.00E+00	1.00E+00	4.67E-01	1.00E+00	1.00E+00	1.54E-01	1.00E+00	3.37E-01
regulation of B cell apoptotic process	1.00E+00	1.00E+00	4.96E-01	1.00E+00	1.25E-01	1.00E+00	4.26E-01	2.13E-02	4.28E-01	4.98E-01	1.00E+00	**1.37E-03**	3.87E-01	1.00E+00	1.00E+00	2.09E-02	4.96E-01	3.63E-01	4.41E-01	1.00E+00	1.00E+00	3.47E-01	1.00E+00	5.45E-01	8.53E-02	1.00E+00	7.35E-02	3.71E-01	1.00E+00	5.23E-01	4.63E-01	6.53E-01
blood vessel development	5.23E-01	6.98E-01	1.00E+00	6.69E-01	4.86E-01	7.33E-01	4.70E-01	1.34E-01	5.60E-01	6.87E-01	5.42E-01	5.83E-01	**1.58E-08**	6.56E-01	7.32E-01	6.06E-01	5.14E-01	4.79E-01	4.08E-01	5.62E-01	1.79E-01	6.32E-01	1.44E-01	5.32E-02	3.35E-02	7.55E-01	5.94E-01	4.75E-01	5.94E-01	6.11E-01	7.00E-01	2.25E-01
negative regulation of T cell apoptotic process	3.01E-01	1.00E+00	1.00E+00	1.00E+00	4.86E-01	2.41E-01	4.41E-01	1.23E-01	1.00E+00	1.00E+00	5.96E-01	5.68E-02	2.79E-02	**6.88E-03**	1.29E-02	6.00E-02	1.00E+00	1.00E+00	1.29E-01	1.00E+00	6.19E-02	1.00E+00	5.15E-01	1.00E+00	1.63E-01	1.00E+00	1.39E-01	3.71E-01	2.55E-02	6.52E-01	5.72E-01	3.14E-01
cytokine-mediated signaling pathway	8.90E-01	8.57E-01	7.08E-01	2.75E-01	7.02E-01	6.04E-01	4.12E-02	3.60E-01	3.60E-01	8.31E-01	3.18E-01	8.02E-01	4.06E-01	6.21E-01	**6.85E-08**	5.49E-01	7.37E-01	8.11E-02	3.55E-01	8.09E-01	5.68E-01	9.00E-01	7.45E-01	7.76E-01	7.33E-01	5.58E-01	1.40E-01	2.88E-01	7.20E-01	4.18E-01	7.84E-01	3.39E-02
CD8-positive alpha-beta T cell activation	6.28E-01	1.53E-01	1.00E+00	1.00E+00	3.66E-01	1.92E-02	3.50E-01	9.02E-02	3.63E-01	5.34E-01	3.27E-02	4.21E-02	1.67E-02	1.00E+00	1.00E+00	**4.19E-03**	1.86E-02	3.00E-01	8.65E-02	1.00E+00	3.36E-02	6.16E-01	4.65E-01	1.28E-01	1.10E-01	1.00E+00	4.20E-01	1.00E+00	1.02E-01	5.93E-01	2.84E-02	6.07E-01
negative regulation of fibroblast proliferation	2.31E-02	4.17E-01	6.31E-01	1.45E-01	2.95E-01	4.25E-01	5.72E-01	2.68E-01	1.00E+00	4.31E-01	4.40E-01	4.16E-01	2.30E-02	6.35E-01	7.11E-01	4.37E-01	**5.14E-03**	1.71E-01	1.54E-02	4.60E-01	4.08E-01	5.81E-01	3.30E-01	6.90E-01	6.43E-01	4.47E-01	1.63E-02	1.00E+00	2.59E-01	4.91E-01	3.85E-01	5.58E-01
purine-containing compound metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.77E-01	3.27E-01	1.13E-01	1.00E+00	1.48E-01	**3.20E-07**	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
regulation of bile acid metabolic process	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	6.10E-02	1.00E+00	1.00E+00	1.00E+00	1.08E-01	1.00E+00	1.00E+00	1.00E+00	4.70E-01	1.00E+00	1.00E+00	**8.60E-03**	1.45E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.35E-02	2.92E-01	1.00E+00	1.00E+00	1.00E+00	1.57E-01
ethanol catabolic process	5.35E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	4.54E-01	1.35E-01	4.70E-01	1.00E+00	1.00E+00	1.00E+00	**2.25E-04**	1.14E-01	1.00E+00	1.00E+00	4.39E-01	1.00E+00	1.22E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
aorta morphogenesis	9.72E-02	7.08E-01	3.25E-01	4.02E-01	6.35E-01	4.42E-01	2.53E-01	6.04E-01	1.00E+00	1.00E+00	7.56E-01	7.40E-01	1.00E+00	3.58E-01	4.42E-01	7.27E-01	6.26E-01	1.00E+00	5.63E-01	4.60E-01	**2.85E-03**	5.44E-01	3.45E-01	6.98E-01	1.00E+00	1.00E+00	2.49E-01	1.00E+00	5.94E-01	7.80E-01	7.00E-01	2.43E-02
regulation of mononeon mediated signaling pathway	6.66E-01	5.76E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.56E-01	1.00E+00	1.00E+00	6.05E-01	1.00E+00	1.00E+00	4.25E-01	1.71E-01	2.25E-01	6.06E-01	4.82E-01	3.49E-01	1.00E+00	1.00E+00	1.51E-02	**3.91E-03**	2.56E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.33E-01	7.38E-01	4.53E-02	6.80E-01
uterine smooth muscle contraction	1.00E+00	1.00E+00	4.01E-01	1.00E+00	1.00E+00	5.01E-01	3.50E-01	1.00E+00	1.00E+00	1.00E+00	4.89E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.21E-01	1.00E+00	1.00E+00	**2.16E-06**	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.55E-01	1.00E+00	1.00E+00
negative regulation of NF-kappaB transcription factor activity	9.72E-01	1.00E+00	8.61E-01	8.86E-01	6.04E-01	5.35E-01	1.00E+00	1.58E-01	1.00E+00	8.99E-01	9.28E-01	9.29E-01	3.85E-01	6.22E-01	7.32E-01	7.84E-01	8.86E-01	1.00E+00	5.67E-01	9.26E-01	7.18E-02	3.60E-01	4.27E-01	**2.07E-04**	6.53E-01	7.78E-01	5.90E-01	7.12E-01	3.38E-01	4.55E-01	2.59E-01	8.65E-01
regulation of DNA recombination	7.58E-01	6.61E-01	8.41E-01	1.00E+00	3.53E-01	7.10E-01	1.00E+00	5.20E-01	1.00E+00	8.91E-01	2.96E-01	6.38E-02	8.80E-01	6.14E-01	7.24E-01	1.65E-01	6.42E-01	1.00E+00	3.08E-01	1.00E+00	1.00E+00	8.31E-01	4.27E-01	1.00E+00	**7.97E-03**	1.00E+00	5.35E-01	7.12E-01	3.12E-01	7.27E-01	4.18E-02	4.25E-01
estrogen biosynthetic process	1.00E+00	1.55E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.64E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.84E-01	**8.93E-05**	6.41E-02	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
spliceosomal tri-snRNP complex assembly	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	5.01E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.54E-02	1.00E+00	1.00E+00	1.00E+00	4.70E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	7.40E-02	1.22E-01	**6.93E-03**	1.00E+00	1.00E+00	1.00E+00	1.23E-02	1.57E-01
positive regulation of insulin secretion involved in cellular response to glucose stimulus	6.83E-01	2.37E-01	5.35E-01	5.64E-01	5.02E-01	5.47E-01	4.93E-01	5.07E-01	4.66E-01	2.78E-01	6.41E-01	2.59E-01	5.36E-01	1.00E+00	1.00E+00	1.08E-01	1.81E-01	1.00E+00	3.62E-02	1.06E-01	2.66E-01	5.05E-01	6.09E-02	2.66E-01	5.24E-01	1.00E+00	1.98E-01	**1.65E-03**	2.00E-01	7.05E-01	5.88E-01	7.11E-01
negative regulation of bone mineralization	1.69E-01	2.40E-02	1.16E-01	1.23E-01	1.27E-02	1.88E-01	3.99E-01	9.02E-02	1.00E+00	1.00E+00	5.42E-01	1.39E-01	1.38E-02	1.37E-02	2.12E-02	2.58E-02	1.14E-01	3.11E-01	1.30E-02	1.94E-01	3.36E-02	6.92E-02	1.00E+00	1.28E-01	1.41E-02	1.59E-01	3.87E-01	5.61E-02	**7.81E-05**	2.55E-01	1.24E-01	4.42E-02
detection of chemical stimulus involved in sensory perception of smell	1.00E+00	1.00E+00	1.00E+00	1.00E+00	3.87E-01	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	**5.65E-33**	1.00E+00	1.00E+00
regulation of membrane depolarization	4.14E-01	1.24E-01	1.99E-01	2.55E-01	4.17E-01	5.64E-01	3.79E-01	3.95E-01	6.61E-01	7.85E-01	8.34E-01	1.24E-01	1.09E-01	1.96E-01	1.52E-01	1.31E-01	4.50E-01	5.81E-01	3.55E-01	3.16E-01	1.17E-01	2.39E-01	9.35E-02	2.50E-01	4.05E-01	5.37E-01	3.70E-01	5.63E-01	1.49E-01	5.43E-01	**5.95E-03**	5.38E-01
regulation of glomerulus development	2.77E-02	4.95E-01	1.00E+00	1.00E+00	4.06E-01	1.00E+00	1.00E+00	6.10E-02	1.00E+00	5.09E-01	1.37E-01	4.53E-01	1.00E+00	1.00E+00	1.00E+00	1.48E-01	4.29E-01	1.00E+00	1.00E+00	5.50E-01	1.00E+00	2.52E-01	9.35E-02	1.00E+00	1.00E+00	1.00E+00	7.35E-02	3.01E-01	1.07E-02	3.55E-02	4.43E-01	**1.20E-03**
regulation of ERK1 and ERK2 cascade	6.85E-16	2.90E-09	8.91E-10	2.40E-04	4.90E-13	3.89E-08	9.77E-08	2.94E-11	3.45E-05	3.56E-05	3.59E-09	1.68E-09	1.32E-19	4.14E-12	1.72E-10	2.45E-26	2.11E-17	2.63E-04	9.47E-17	9.35E-03	1.39E-23	4.45E-05	8.11E-16	3.39E-15	1.00E-09	4.29E-08	9.64E-11	6.07E-05	1.00E-26	4.51E-07	5.40E-11	1.40E-12
response to extracellular stimulus	2.74E-14	1.42E-06	1.90E-07	2.48E-06	4.00E-09	8.66E-04	4.41E-09	4.24E-09	1.12E-06	2.06E-04	1.00E-09	1.68E-13	5.25E-05	1.64E-03	1.25E-04	4.55E-11	3.44E-08	7.89E-08	1.03E-11	7.30E-05	9.06E-07	8.06E-03	5.03E-11	3.99E-08	1.62E-13	2.78E-07	4.23E-13	2.64E-05	2.81E-08	8.22E-04	4.65E-08	8.79E-10
regulation of receptor activity	1.11E-19	1.04E-11	2.69E-20	3.05E-06	6.93E-21	7.83E-13	4.90E-16	8.49E-23	2.91E-14	3.02E-07	3.82E-17	3.35E-27	1.74E-30	1.76E-17	4.78E-23	2.44E-46	5.22E-34	8.64E-11	1.00E-25	4.71E-03	6.54E-28	2.02E-12	6.20E-29	2.26E-17	1.26E-21	1.11E-17	6.88E-33	4.90E-21	1.83E-43	1.62E-04	1.24E-24	1.87E-25

Fill legend	Text color legend	Border legend
q-value <= 0.01	Direct literature evidence	Pathways enriched with qvalue <= 0.01 only in 1 tissue
0.01 < q-value <= 0.05	Indirect literature evidence	
0.05 < q-value <= 0.1	No literature evidence found	

Figure 6: Tissue-exclusive pathways, HA TopNets. Top GO Biological Process exclusively enriched in each tissue, as well as the top 3 processes enriched in all tissues. The q-value of enrichment is provided for all cases. Green filling corresponds to cases with q-value ≤ 0.01, orange to $0.01 <$ q-value ≤ 0.05, and blue to $0.05 <$ q-value ≤ 0.1. Literature evidence support 29 out of 32 tissue-pathway pairs with q-value ≤ 0.01 (bold green text; Supplementary Table S7).

(Sambarey *et al.* (2017b)), (ii) discriminate between primary and metastatic melanoma (Metri *et al.* (2017)), and (iii) identify influential genes in the condition under study (Sambaturu *et al.* (2016)). Substantially extending our prior work, PathExt provides a general framework to address all the above questions, while employing rigorous statistical significance estimation to identify critical paths. Importantly, PathExt is designed to operate even when a single sample is available for each condition, and in the absence of an experimental control sample.

Current approaches to identify active sub-networks are largely built on the work by Ideker *et al.* (2002), called jActiveModules, which formulates a sub-network scoring scheme based on the statistical significance of differential gene expression, and then identifies high-scoring sub-networks using a simulated annealing approach. Cabusora *et al.* (2005) use the same scoring method but identify sub-networks by listing k-shortest paths in the interaction network among a set of 'seed' nodes. The best scoring sub-network is then identified by sampling seed nodes based on their differential expression. Although this method computes paths (unlike jActiveModules), all edges in the interaction network are given equal importance, and the scoring scheme as well as seed node prioritization still focuses on DEGs. Other methods along similar ideas have been proposed, that filter sub-networks based, for example, on network motifs (Milo *et al.* (2002)). In contrast, PathExt assigns weights to the interactions in the biological network as a function of the given omics data, thus transferring importance from individual genes to paths, and potentially capturing the way in which biological phenotypes emerge from interconnected processes. Interestingly, even though connectedness is not used as a criterion to identify sub-networks, the TopNet resulting from the identified paths forms a well-connected graph.

PathExt relies on two user defined parameters, the threshold k used to select the top k shortest paths, and the q-value for statistical significance of the paths selected to construct TopNet. These values have been set at very stringent values in this paper, allowing us to focus on the most active paths. Different thresholds can give different layers of information, with different levels of false discovery.

In summary, PathExt is a general framework for path-based mining of omics-integrated biological networks. While the paths identified by PathExt may not constitute a comprehensive or exhaustive listing of all the active, altered processes in the system, the resulting TopNet can be thought of as a starting point from which hypotheses can be generated. In this work, we have gathered, for each drug and each tissue, the top central genes, along with their fold change for drug exposure (Supplementary Tables S4, S5), and z-score for human tissues (Supplementary Tables S10, S11). Further examining the network or genomic neighborhood of these and other genes comprising the TopNet can provide additional insights, or strengthen the insights gained.

Funding

This work was supported by the Department of Biotechnology (DBT) - Indian Institute of Science (IISc) Partnership Program - Phase II [BT/PR27952/IN/22/ 212/2018] and the Mathematical Biology Initiative [DSTO/PAM/GR-1303] of the Government of India. S.H. was supported in part by National Science Foundation (NSF) award 1564785 and in part by the Intramural Research Program of the National Cancer Institute, Center for Cancer Research, National Institutes of Health (NIH).

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**(1), 289–300.

Bernstein, J. (1982). The role of the lung in the metabolism of ethanol. *Research communications in chemical pathology and pharmacology*, **38**(1), 43–56.

Bindea, G., Mlecnik, B., Hackl, H., Charoentong, P., Tosolini, M., Kirilovsky, A., Fridman, W.-H., Pagès, F., Trajanoski, Z., and Galon, J. (2009). Cluego: a cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics*, **25**(8), 1091–1093.

Blumenberg, M. (2019). Introductory chapter: Transcriptome analysis. In *Transcriptome Analysis*. IntechOpen.

Boshoff, H. I., Reed, M. B., Barry III, C. E., and Mizrahi, V. (2003). Dnae2 polymerase contributes to in vivo survival and the emergence of drug resistance in mycobacterium tuberculosis. *Cell*, **113**(2), 183–193.

Bult, C. J., Blake, J. A., Smith, C. L., Kadin, J. A., and Richardson, J. E. (2018). Mouse genome database (mgd) 2019. *Nucleic acids research*, **47**(D1), D801–D806.

Cabusora, L., Sutton, E., Fulmer, A., and Forst, C. V. (2005). Differential network expression during drug and stress response. *Bioinformatics*, **21**(12), 2898–2905.

Carithers, L. J. and Moore, H. M. (2015). The genotype-tissue expression (gtex) project.

Chiang, J. Y. (2013). Bile acid metabolism and signaling. *Comprehensive Physiology*, **3**(3), 1191–1212.

DeJesus, M. A., Gerrick, E. R., Xu, W., Park, S. W., Long, J. E., Boutte, C. C., Rubin, E. J., Schnappinger, D., Ehrt, S., Fortune, S. M., *et al.* (2017). Comprehensive essentiality analysis of the mycobacterium tuberculosis genome via saturating transposon mutagenesis. *MBio*, **8**(1), e02133–16.

Dijkstra, E. W. *et al.* (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, **1**(1), 269–271.

Eisenberg, E. and Levanon, E. Y. (2013). Human housekeeping genes, revisited. *TRENDS in Genetics*, **29**(10), 569–574.

Eltringham, I., Drobniewski, F., Mangan, J., Butcher, P., and Wilson, S. (1999). Evaluation of reverse transcription-pcr and a bacteriophage-based assay for rapid phenotypic detection of rifampin resistance in clinical isolates of mycobacterium tuberculosis. *Journal of clinical microbiology*, **37**(11), 3524–3527.

Esteve-Codina, A. (2018). Rna-seq data analysis, applications and challenges. *Data Analysis for Omic Sciences: Methods and Applications*, **82**, 71.

Feiglin, A., Allen, B. K., Kohane, I. S., and Kong, S. W. (2017). Comprehensive analysis of tissue-wide gene expression and phenotype data reveals tissues affected in rare genetic disorders. *Cell systems*, **5**(2), 140–148.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., and McKusick, V. A. (2005). Online mendelian inheritance in man (omim), a knowledgebase of human genes and genetic disorders. *Nucleic acids research*, **33**(suppl_1), D514–D517.

Howard, N. C., Marin, N. D., Ahmed, M., Rosa, B. A., Martin, J., Bambouskova, M., Sergushichev, A., Loginicheva, E., Kurepina, N., Rangel-Moreno, J., *et al.* (2018). Mycobacterium tuberculosis carrying a rifampicin drug resistance mutation reprograms macrophage metabolism through cell wall lipid changes. *Nature microbiology*, **3**(10), 1099.

Ideker, T., Ozier, O., Schwikowski, B., and Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, **18**(suppl_1), S233–S240.

Jain, P., Weinrick, B. C., Kalivoda, E. J., Yang, H., Munsamy, V., Vilcheze, C., Weisbrod, T. R., Larsen, M. H., O'Donnell, M. R., Pym, A., *et al.* (2016). Dual-reporter mycobacteriophages (ϕ2drms) reveal preexisting mycobacterium tuberculosis persistent cells in human sputum. *MBio*, **7**(5), e01023–16.

Jiang, Z., Zhou, X., Li, R., Michal, J. J., Zhang, S., Dodson, M. V., Zhang, Z., and Harland, R. M. (2015). Whole transcriptome analysis with sequencing: methods, challenges and potential solutions. *Cellular and molecular life sciences*, **72**(18), 3425–3439.

Kang, N. and Koo, J. (2012). Olfactory receptors in non-chemosensory tissues. *BMB reports*, **45**(11), 612.

Kebabian, J. W., Petzold, G. L., and Greengard, P. (1972). Dopamine-sensitive adenylate cyclase in caudate nucleus of rat brain, and its similarity to the "dopamine receptor". *Proceedings of the National Academy of Sciences*, **69**(8), 2145–2149.

Köhler, S., Doelken, S. C., Mungall, C. J., Bauer, S., Firth, H. V., Bailleul-Forestier, I., Black, G. C., Brown, D. L., Brudno, M., Campbell, J., *et al.* (2013). The human phenotype ontology project: linking molecular biology and disease through phenotype data. *Nucleic acids research*, **42**(D1), D966–D974.

Lanciego, J. L., Luquin, N., and Obeso, J. A. (2012). Functional neuroanatomy of the basal ganglia. *Cold Spring Harbor perspectives in medicine*, **2**(12), a009621.

Ma, S., Minch, K. J., Rustad, T. R., Hobbs, S., Zhou, S.-L., Sherman, D. R., and Price, N. D. (2015). Integrated modeling of gene regulatory and metabolic networks in mycobacterium tuberculosis. *PLoS computational biology*, **11**(11), e1004543.

McGrath, M., Gey van Pittius, N., Van Helden, P., Warren, R., and Warner, D. (2013). Mutation rate and the emergence of drug resistance in mycobacterium tuberculosis. *Journal of Antimicrobial Chemotherapy*, **69**(2), 292–302.

Metri, R., Mohan, A., Nsengimana, J., Pozniak, J., Molina-Paris, C., Newton-Bishop, J., Bishop, D., and Chandra, N. (2017). Identification of a gene signature for discriminating metastatic from primary melanoma using a molecular interaction network approach. *Scientific reports*, **7**(1), 17314.

Miesel, L., Rozwarski, D. A., Sacchettini, J. C., and Jacobs Jr, W. R. (1998). Mechanisms for isoniazid action and resistance. In *Genetics and Tuberculosis: Novartis Foundation Symposium 217*, pages 209–221. Wiley Online Library.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**(5594), 824–827.

Mishra, S., Shukla, P., Bhaskar, A., Anand, K., Baloni, P., Jha, R. K., Mohan, A., Rajmani, R. S., Nagaraja, V., Chandra, N., *et al.* (2017). Efficacy of β-lactam/β-lactamase inhibitor combination is linked to whib4-mediated changes in redox physiology of mycobacterium tuberculosis. *Elife*, **6**, e25624.

Mitra, K., Carvunis, A.-R., Ramesh, S. K., and Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature Reviews Genetics*, **14**(10), 719–732.

Owens, G. C., Erickson, K. L., Malone, C. C., Pan, C., Huynh, M. N., Chang, J. W., Chirwa, T., Vinters, H. V., Mathern, G. W., and Kruse, C. A. (2015). Evidence for the involvement of gamma delta t cells in the immune response in rasmussen encephalitis. *Journal of neuroinflammation*, **12**(1), 134.

Pearring, J. N., San Agustin, J. T., Lobanova, E. S., Gabriel, C. J., Lieu, E. C., Monis, W. J., Stuck, M. W., Strittmatter, L., Jaber, S. M., Arshavsky, V. Y., *et al.* (2017). Loss of arf4 causes severe degeneration of the exocrine pancreas but not cystic kidney disease or retinal degeneration. *PLoS genetics*, **13**(4), e1006740.

Rajala, R. V. and Anderson, R. E. (2010). Rhodopsin-regulated insulin receptor signaling pathway in rod photoreceptor neurons. *Molecular neurobiology*, **42**(1), 39–47.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, **43**(7), e47–e47.

Sambarey, A., Devaprasad, A., Baloni, P., Mishra, M., Mohan, A., Tyagi, P., Singh, A., Akshata, J., Sultana, R., Buggi, S., *et al.* (2017a). Meta-analysis of host response networks identifies a common core in tuberculosis. *NPJ Systems Biology and Applications*, **3**(1), 4.

Sambarey, A., Devaprasad, A., Mohan, A., Ahmed, A., Nayak, S., Swaminathan, S., D'Souza, G., Jesuraj, A., Dhar, C., Babu, S., *et al.* (2017b). Unbiased identification of blood-based biomarkers for pulmonary tuberculosis by modeling and mining molecular interaction networks. *EBioMedicine*, **15**, 112–126.

Sambaturu, N., Mishra, M., and Chandra, N. (2016). Epitracer-an algorithm for identifying epicenters in condition-specific biological networks. *BMC genomics*, **17**(4), 543.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research*, **13**(11), 2498–2504.

Stretch, C., Khan, S., Asgarian, N., Eisner, R., Vaisipour, S., Damaraju, S., Graham, K., Bathe, O. F., Steed, H., Greiner, R., *et al.* (2013). Effects of sample size on differential gene expression, rank order and prediction accuracy of a gene signature. *PloS one*, **8**(6), e65380.

Takayama, K., Wang, C., and Besra, G. S. (2005). Pathway to synthesis and processing of mycolic acids in mycobacterium tuberculosis. *Clinical microbiology reviews*, **18**(1), 81–101.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., *et al.* (2015). Tissue-based map of the human proteome. *Science*, **347**(6220), 1260419.

Varadkar, S., Bien, C. G., Kruse, C. A., Jensen, F. E., Bauer, J., Pardo, C. A., Vincent, A., Mathern, G. W., and Cross, J. H. (2014). Rasmussen's encephalitis: clinical features, pathobiology, and treatment advances. *The Lancet Neurology*, **13**(2), 195–205.

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., *et al.* (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, **46**(D1), D1074–D1082.

Yu, G., Wang, L.-G., Han, Y., and He, Q.-Y. (2012). clusterprofiler: an r package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, **16**(5), 284–287.

Zhu, J.-H., Wang, B.-W., Pan, M., Zeng, Y.-N., Rego, H., and Javid, B. (2018). Rifampicin can induce antibiotic tolerance in mycobacteria via paradoxical changes in rpob transcription. *Nature communications*, **9**(1), 4218.