Bioinformatics
doi.10.1093/bioinformatics/xxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Systems Biology

A convex optimization approach for identification of human tissue-specific interactomes

Shahin Mohammadi [1,*] and Ananth Grama [1]

[1] Department of Computer Sciences, Purdue University, West Lafayette, IN 47907, USA.

* To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Analysis of organism-specific interactomes has yielded novel insights into cellular function and coordination, understanding of pathology, and identification of markers and drug targets. Genes, however, can exhibit varying levels of cell-type specificity in their expression, and their coordinated expression manifests in tissue-specific function and pathology. Tissue-specific/ selective interaction mechanisms have significant applications in drug discovery, as they are more likely to reveal drug targets. Furthermore, tissue-specific transcription factors (tsTFs) are significantly implicated in human disease, including cancers. Finally, disease genes and protein complexes have the tendency to be differentially expressed in tissues in which defects cause pathology. These observations motivate the construction of refined tissue-specific interactomes from organism-specific interactomes.

Results: We present a novel technique for constructing human tissue-specific interactomes. Using a variety of validation tests (ESEA, GO Enrichment, Disease-Gene Subnetwork Compactness), we show that our proposed approach significantly outperforms state of the art techniques. Finally, using case studies of Alzheimer's and Parkinson's diseases, we show that tissue-specific interactomes derived from our study can be used to construct pathways implicated in pathology, and demonstrate the use of these pathways in identifying novel targets.

Availability: `http://www.cs.purdue.edu/homes/mohammas/projects/ActPro`.

Contact: `mohammadi@purdue.edu`

1 Introduction

Proteins are basic workhorses of living cells, and their overall quantity is tightly regulated across different tissues and cell-types to manifest tissue-specific biology and pathobiology. These regulatory controls orchestrate cellular machinery at different levels of resolution, including, but not limited to gene regulation (Mele *et al.*, 2015; Göring, 2012), epigenetic modification (Chatterjee and Vinson, 2012; Mendizabal *et al.*, 2014), alternative splicing (Buljan *et al.*, 2012; Ellis *et al.*, 2012), and post-translational modifications (Vaidyanathan and Wells, 2014; Ikegami *et al.*, 2014). Transcriptional regulation is a key component of this hierarchical regulation, which has been widely used to study context-specific phenotypes. In the context of human tissues/ cell-types, genes can exhibit varying levels of specificity in their expression. They can be broadly classified as: (i) tissue-specific (unique to one cell-type); (ii) tissue-selective (shared among coherent groups of cell-types); and (iii) housekeeping (utilized in all cell-types). Tissue-specific/selective genes have significant applications in drug discovery, since they have been shown to be more likely drug targets (Dezso *et al.*, 2008). Tissue-specific transcription factors (tsTFs) are significantly implicated in human diseases (Raj *et al.*, 2014; Messina *et al.*, 2004), including cancers (Vaquerizas *et al.*, 2009). Finally, disease genes and protein complexes tend to be over-expressed in tissues in which defects cause pathology (Lage *et al.*, 2008).

The majority of human proteins do not work in isolation but take part in pathways, complexes, and other functional modules. Tissue-specific proteins are known to follow a similar trend. Perturbations that impact interacting interfaces of proteins are significantly enriched among tissue-specific, disease-causing variants (Wang *et al.*, 2012; Rolland *et al.*, 2014; Sahni *et al.*, 2015). This emphasizes the importance of constructing tissue-specific interactomes and their constitutive pathways for understanding mechanisms that differentiate cell-types and make them uniquely susceptible to tissue-specific disorders. Prior attempts

at reconstructing human tissue-specific interactomes rely on a set of *"expressed genes"* in each tissue, and use this set as the baseline of transcriptional activity. The node removal (NR) method (Bossi and Lehner, 2009) constructs tissue-specific interactomes by identifying the induced subgraph of the expressed genes. Magger *et al.* (Magger *et al.*, 2012) propose a method called *"Edge ReWeighting (ERW)"*, which extends the idea of Bossi *et al.* to weighted graphs. This method penalizes an edge once, if one of its end-points is not expressed, and twice, if both end-points are missing from the expressed gene-set.

While these methods have been used to study tissue-specific interactions, their underlying construction relies only on the immediate end-points of each interaction to infer tissue-specificity. Furthermore, they threshold expression values, often using ad-hoc choices of thresholds to classify genes as either expressed or not. Finally, it is hard to integrate expression datasets from multiple platforms, or from multiple labs into a single framework. These constraints are primarily dictated by limitations of high-throughput technologies for assaying gene expression. In these technologies, one can easily compare expression of the same gene across different samples to perform differential analysis; however, expression of different genes in the same sample are not directly comparable due to technical biases, differences in baseline expression, and GC content of genes. A recently proposed method called *Universal exPression Code (UPC)* (Piccolo *et al.*, 2013) addresses many of these issues by removing platform-specific biases, and converting raw expressions to a unified transcriptional activity score. These scores are normalized, and can be compared across different genes and platforms.

Leveraging the UPC method, we propose a novel approach that uses the topological context of an interaction to infer its specificity score. Our approach formulates the inference problem as a suitably regularized convex optimization problem. The objective function of the optimization problem has two terms – the first term corresponds to a *diffusion kernel* that propagates activity of genes through interactions (network links). The second term is a *regularizer* that penalizes differences between *transcriptional* and *functional* activity scores. We use these functional activity scores to compute tissue-specificity for each edge in the global interactome, which we show, through a number of validation tests, are significantly better than prior methods. Our method is widely applicable, and can be applied directly to single-channel, double-channel, and RNA-Seq expression datasets processed using UPC/SCAN. Furthermore, it can be easily adapted to cases where expression profiles are only available in preprocessed form.

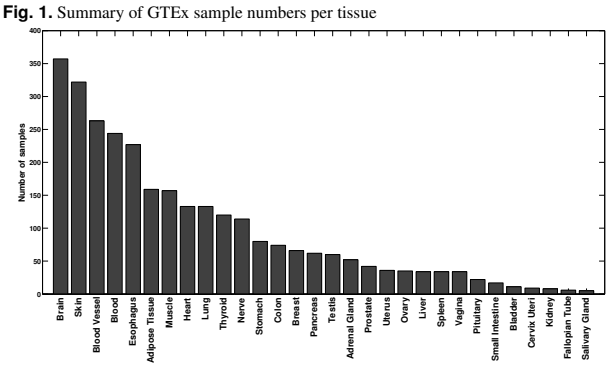
The rest of the paper is organized as follows: In Section 2.1 we provides details of the datasets used in our study. Next, we introduce our method, called *Activity Propagation (ActPro)*, along with a generalization of the ERW methods, and formalize different methods in a consistent notation. We evaluate the effectiveness of UPC transcriptional activity scores to predict tissue-specific genes in Section 3.1. Details of procedure for constructing tissue-specific networks and their parameter choices are discussed in Section 3.2. Sections 3.3-3.5 present validation studies for tissue-specific interactions using known pathway edges, co-annotation of proteins, and GWAS disease genes. Finally, in Section 3.6, we use the brain-specific interactome constructed using our method to identify novel disease-related pathways, and use them to identify candidate targets for neurodegenerative disorders.

2 Materials and Methods

2.1 Datasets

We downloaded the RNASeq dataset version 4.0 (*dbGaP accession phs000424.v4.p1*) from the The Genotype-Tissue Expression (GTEx) project (Ardlie *et al.*, 2015; Mele *et al.*, 2015). This dataset contains 2,916 samples from 30 different tissues/cell types, the summary of which is presented in Figure 1. We processed each sample using the UPC method (Piccolo *et al.*, 2013), a novel platform-independent normalization technique that corrects for platform-specific technical variations and estimates the probability of transcriptional activity for each gene in a given sample. The benefit of this method is that activation probability scores are highly consistent across different technologies, and more importantly, they are comparable across different genes in a given sample. For each gene, we recorded the transcript with the highest activation probability in the sample. Finally, we averaged replicate samples within each group to construct a unique transcription signature vector for each tissue/ cell type. The final dataset contains the expression value of 23,243 genes across 30 different tissues/ cell types.

Fig. 1. Summary of GTEx sample numbers per tissue



In addition, we extracted human protein-protein interactions from the iRefIndex database (Razick *et al.*, 2008), which consolidates protein interactions from different databases. Edges in this dataset are weighted using an MI (MINT-Inspired) score, which measures the confidence of each interaction based on three different evidence types, namely the interaction types (binary/complex) and experimental method used for detection, the total number of unique PubMed publications reporting the interaction, and the cumulative evidence of interlogous interactions from other species. Finally, we map transcription data to the human interactome by first converting all gene IDs to Entrez Gene IDs and only retaining genes that both have a corresponding node in the interactome and have been profiled by the GTEx project. This yields a global interactome with 147,444 edges, corresponding to protein-protein interactions, between 14,658 nodes, representing gene products.

2.2 Constructing human tissue-specific interactome

The global human interactome is a superset of all *possible* physical interactions that can take place in the cell. It does not provide any information as to which interactions actually occur in a given context. There are a variety of factors, including co-expression of genes corresponding to a pair of proteins, their co-localization, and post-translational modification, that mediate protein interactions at the right time and place. Quantifiable expression of both proteins involved in an interaction is one of the most important factors that determine the existence of an interaction. Different methods have been proposed in literature to utilize this source of information to construct human tissue-specific interactomes. Here, we briefly review existing methods, their drawbacks, and propose a new method, called *Activity Propagation (ActPro)*, which addresses noted shortcomings.

2.2.1 Prior Methods

Let us denote the adjacency matrix of the global interatome by $\mathbf{A}$, where element $\boldsymbol{a}_{ij}$ is the weight (confidence) of the edge connecting vertices v_i and v_j. Let $\boldsymbol{z}$ encode expression of genes in a tissue and $\underline{\boldsymbol{z}}$ be the binarized version of $\boldsymbol{z}$ for a fixed threshold. Finally, let $\boldsymbol{diag}$ operator applied to a given vector be the diagonal matrix with the vector on the main diagonal. Our aim is to compute a matrix $\hat{\mathbf{A}}$, which is the adjacency matrix of the tissue-specific interactome for a given expression profile. Using this notation, we can summarize prior methods for constructing tissue-specific interactomes as follows.

- **Node Removal (NR)** This method computes the induced subgraph of the "expressed" gene products (Bossi and Lehner, 2009).

$$\hat{\mathbf{A}} = \boldsymbol{diag}(\underline{\boldsymbol{z}}) * \mathbf{A} * \boldsymbol{diag}(\underline{\boldsymbol{z}}) \tag{1}$$

- **Edge Re-Weighting (ERW)** This method penalizes edges according to the expression state (active/ inactive) of its end points (Magger *et al.*, 2012). Given a penalty parameter $0 \leq rw \leq 1$, ERW penalizes each edge by rw once, if only one of its end-points is active, and twice, if both incident vertices are inactive. Formally:

$$\hat{\mathbf{A}} = \boldsymbol{diag}(rw * (\boldsymbol{e} - \underline{\boldsymbol{z}})) * \mathbf{A} * \boldsymbol{diag}(rw * (\boldsymbol{e} - \underline{\boldsymbol{z}})) \tag{2}$$

  where $\boldsymbol{e}$ is the vector of all ones.
- **Adaptive ERW** Adaptive ERW is a simple extension of ERW, in which we directly use transcriptional activity of genes to penalize each edge, instead of a fixed penalty. This is enabled by the UPC method, which normalizes expression values into compatible activation probabilities. In this formulation, we have:

$$\hat{\mathbf{A}} = \boldsymbol{diag}(\boldsymbol{z}) * \mathbf{A} * \boldsymbol{diag}(\boldsymbol{z}) \tag{3}$$

2.2.2 Proposed Method

The main assumption of *ERW* and *NR* methods is that *transcriptional activity* of a gene is a reliable proxy for its *functional activity*. While this holds in a majority of cases, there are situations in which these scores differ significantly. First, the basis for *transcriptional activity* estimation is that genes with higher expression levels have higher chance of being functionally active in a given context. While this is generally true, there are genes that only need a low expression level to perform their function; i.e., their functionally active concentration is much lower than the rest of genes. Second, there is noise associated with measurement of gene expression, and converting measured expression values to UPC scores can over/ under-estimate *transcriptional activity*. Finally, we note that there are genes whose *down-regulation* corresponds to their functional activity (as opposed to the other way around).

Based on these observations, we propose a novel framework, called *Activity Propagation (ActPro)* to identify the most functionally active subnetwork of a given interactome. Our method incorporates global network topology to propagate activity scores, while simultaneously minimizing the number of changes to the gene activity scores. To this end, we first define a smoothed functional activity score defined by the following optimization problem:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ (1-\alpha)\boldsymbol{x}\mathbf{L}\boldsymbol{x} + \alpha \parallel \boldsymbol{x} - \boldsymbol{z} \parallel_1 \right\}$$
$$\text{Subject to: } \begin{cases} \mathbf{1}^T\boldsymbol{x} = 1 \\ 0 \leq \boldsymbol{x} \end{cases} \tag{4}$$

In this problem, $\mathbf{L}$ is the *Laplacian* matrix, defined as $\mathbf{A} - \boldsymbol{\Delta}$, where element δ_{ii} of $\boldsymbol{\Delta}$ is the weighted degree of i^{th} vertex in the global interactome. The Laplacian operator $\mathbf{L}$ acts on a given function defined over vertices of a graph, such as $\boldsymbol{x}$, and computes the smoothness of $\boldsymbol{x}$ over adjacent vertices. More specifically, we can expand the first term in Equation 6 as $\sum_{i,j} w_{i,j} \parallel x_i - x_j \parallel_2^2$, which is the accumulated difference of values between adjacent nodes scaled by the weight of the edge connecting them. This term defines a *diffusion kernel* that propagates activity of genes through network links. The second term is a *regularizer*, which penalizes changes by enforcing sparsity over the vector of differences between *transcriptional* and *functional* activities. This minimizes deviation from original transcriptome. It should be noted here that use of norm-1 is critical, since norm-2 regularization diffuses the transcriptional activity scores and significantly reduces their discriminating power. This negative aspect of norm-2 minimization is confirmed by our experiments. Finally, constraint $\mathbf{1}^T\boldsymbol{x} = 1$, known as fixed budget. It ensures that vector $\boldsymbol{x}$ is normalized and bounded. Parameter α determines the relative importance of regularization versus loss. We can equivalently define a penalization parameter $\lambda = \frac{\alpha}{1-\alpha}$, which is the standard notation in optimization framework. This is a convex optimization problem, and we can solve it using efficient solvers to identify its global optimum.

Using the smoothed activity scores, we can re-weight the global human interactome as follows:

$$\hat{\mathbf{A}} = \boldsymbol{diag}(\boldsymbol{x}^*) * \mathbf{A} * \boldsymbol{diag}(\boldsymbol{x}^*) \tag{5}$$

We can also derive an alternative formulation for *ActPro* which, instead of using *transcriptional activity* scores computed by UPC, uses expression values computed through more common methods such as RMA or MAS5.0 (Lim *et al.*, 2007). We call this method *penalty propagation*, or *PenPro* for short. In this framework, computed expression values are not directly comparable and we need to threshold them to classify genes as either expressed or not. Using the same notation defined previously, we can define *functional activity* scores by solving the following problem:

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}}{\operatorname{argmin}} \left\{ (1-\alpha)\boldsymbol{x}\mathbf{L}\boldsymbol{x} + \alpha \parallel \boldsymbol{x} - \underline{\boldsymbol{z}} \parallel_1 \right\}$$
$$\text{Subject to: } \begin{cases} \mathbf{1}^T\boldsymbol{x} = 1 \\ 0 \leq \boldsymbol{x} \end{cases} \tag{6}$$

The only difference here is that, instead of *transcriptional activity* vector $\boldsymbol{z}$, we use the binarized expression vector $\underline{\boldsymbol{z}}$. We observe similar performance for *ActPro* and *PenPro*, with *ActPro* being marginally superior in all cases, and thus we will only present results for *ActPro*.

2.3 Implementation details

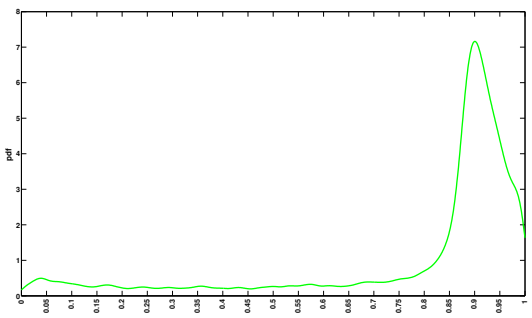
All software used in our experiments has been implemented in Matlab. To solve the convex problem in Equation 6, we used CVX, a package for specifying and solving convex programs (Grant and Boyd, 2008). We used Mosek together with CVX, which is a high-performance solver for large-scale linear and quadratic programs (MOSEK-ApS, 2015).

3 Results and discussion

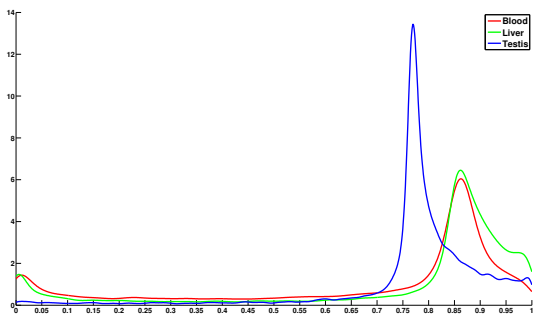
3.1 Evaluating tissue-specificity of genes

To validate the quality of UPC normalized expression values, we first analyze the distribution of gene expressions across all tissues. Figure 2(a) shows the distribution of transcriptional activities, averaged over all samples. The overall distribution exhibits a bimodal characteristic that has a clear separation point that distinguishes expressed genes from others. We set a global threshold of 0.75 for identifying genes that are expressed in each tissue. These genes are used in evaluating NR and ERW methods. It

should be noted that the distribution of UPC values vary across cell-types, as shown in Figure 2(b); however, the separation point is robust.



(a) Average expression across all samples



(b) Three tissues with low, medium, and high number of expressed genes

Fig. 2. Distribution of UPC normalized gene expression values

Expression value of genes across tissues can be classified as *specific, selective*, or *housekeeping*. Housekeeping (HK) genes are ubiquitously expressed across all tissues to perform core cellular functions. In order to investigate the typical transcriptional behavior of HK genes, we first downloaded the set of housekeeping genes from two separate studies(Eisenberg and Levanon, 2013; Uhlen *et al.*, 2015), measured using RNASeq and protein expression values, respectively, and defined a gold-standard set of HK genes as the intersection of these two datasets. Additionally, we downloaded genes in five well-known classes of HK genes from the Human Protein Atlas (Uhlen *et al.*, 2015). Figure 3(a) shows the expression values of each set averaged over different tissues. These expression values are exclusively above 0.8, which indicates that they are among the set of expressed genes that we identified in each tissue.

To further investigate the distribution of HK genes across tissues, we computed the total number of tissues that each gene is expressed in. Figure 3(b) shows the total number tissues that these genes are expressed in. We define HK genes in our dataset as the set of genes that are expressed in a majority of tissues ($\frac{2}{3}$ of all tissues). Finally, we define the set of preferentially expressed genes in each tissue as the set of expressed genes that are not ubiquitous (housekeeping). Figure 3(c) shows the total number of genes identified in each tissue as preferentially expressed (either specific or selective). Testis tissue exhibits the largest number of preferentially expressed genes (we refer to these as markers), with more that $1,400$ genes, while blood samples have the fewest markers with only ~ 250 marker genes.

Next, in order to assess whether the sets of preferentially expressed genes can predict tissue-specific functions, we performed GO enrichment analysis over different sets of tissue-specific markers using *GOsummaries*

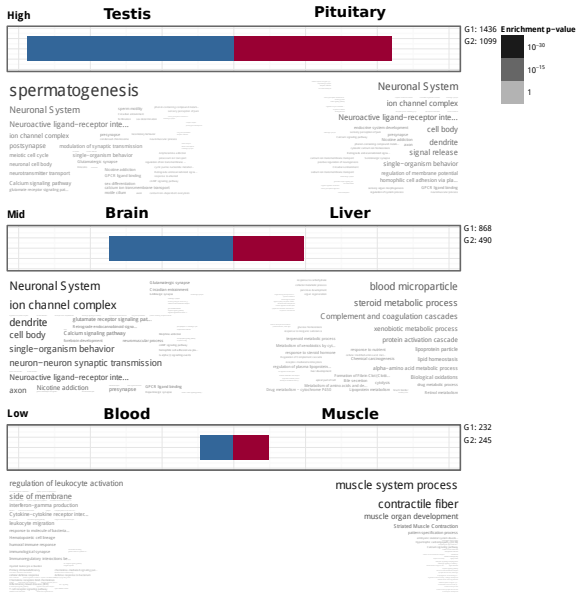


Fig. 4. Evaluation of tissue-specific markers using a threshold value of 0.75 to define genes as expressed. For tissues with high, medium, and low number of markers, we chose candidates for further enrichment analysis, which are presented in this figure.

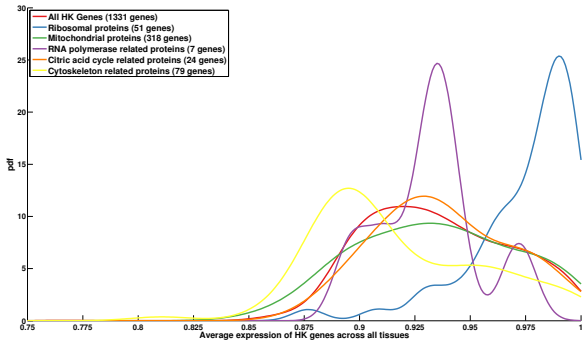
package in R/Bioconductor (Kolde, 2014). This package uses g:Profiler (Reimand *et al.*, 2011) as backend for enrichment analysis and provides a simple visualization of the results as a word cloud. The coverage of available annotations for different tissues is not uniform; that is, some tissues are better annotated for specific terms than the others. We chose six well-annotated tissues with high, mid, and low number of identified markers for further study. We limited terms to the ones with at least 20 and at most 500 genes to avoid overly generic/specific terms. Finally, we used a strong hierarchical filtering to remove duplicate GO terms and thresholded terms at p-value of 0.05. Figure 4 shows the enrichment word-cloud for each tissue. It can be seen that all terms identified here are highly tissue-specific and representative of main functions for each tissue, which supports the validity of computed transcriptional activity scores from UPC.

3.2 Constructing tissue-specific interactomes

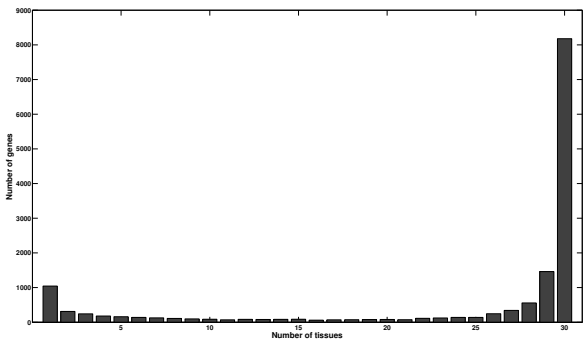
Node Removal (NR) and Edge ReWeighting (ERW) methods need a predefined set of expressed genes in each tissue to construct tissue-specific interactomes (or a given lower bound to threshold expression values). We use the set of all genes with transcriptional activity greater than or equal to 0.75 as the set of expressed genes for these methods. We chose this threshold based on the averaged distribution of gene expressions, as well as further manual curation of genes at different thresholds.

Node Removal (NR) method is known to disintegrate the network with stringent expression values(Magger *et al.*, 2012). To evaluate the performance of NR over different expression thresholds and assess its sensitivity to the choice of threshold, we computed the total number of components, as well as the size of largest connected components, while varying the value of expression threshold. Figure 5 shows stable behavior up to threshold value of 0.75, after which both the size of largest component and the total number of connected components exhibit a rapid shift and the network starts to disintegrate. This suggests that the expression value of 0.75 is also the optimal topological choice for NR method.

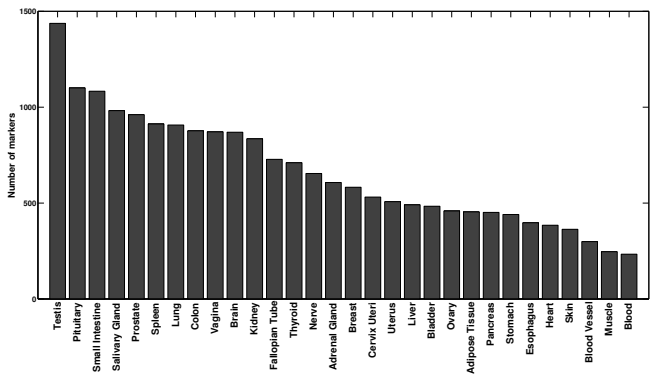
For the ActPro algorithm, we evaluated the results over three different values of α in set $\{0.15, 0.5, 0.85\}$ and reported the result for each case.



(a) Average expression of HK genes in different classes

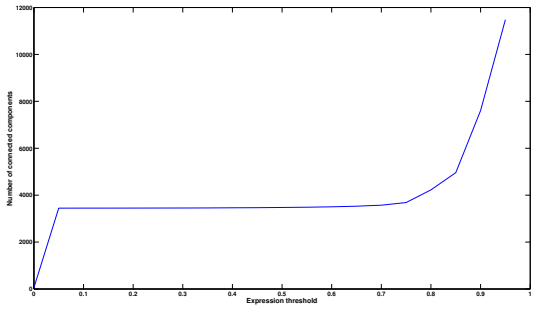


(b) Number of HK genes depending on expression threshold

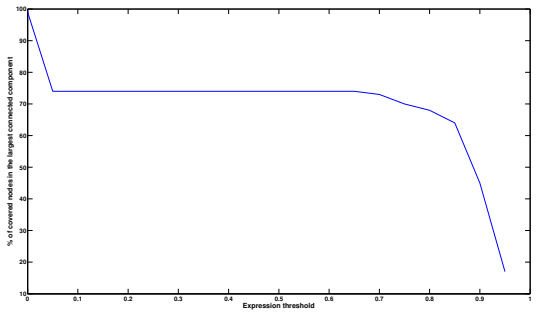


(c) Number of selected markers per tissue/cell type

Fig. 3. Summary statistics for tissue-specific and housekeeping genes in the GTEx dataset



(a) Number of connected components



(b) Coverage of the largest connected component (LCC)

Fig. 5. Effect of threshold values on topological characteristics of tissue-specific networks for the node removal (NR) method.

3.3 Tissue-specific interactome predicts context-sensitive interactions in known functional pathways

To evaluate the power of tissue-specific interactions in capturing context-sensitive physical interactions in known pathways, we first use Edge Set Enrichment Analysis (ESEA) to rank pathway edges according to their gain/loss of mutual information in each tissue context (Han *et al.*, 2015). ESEA aggregates pathways from seven different sources (KEGG; Reactome; Biocarta, NCI/Nature Pathway Interaction Database; SPIKE; HumanCyc; and Panther) and represents them as a graph with edges corresponding to biological relationships, resulting in over 2,300 pathways spanning 130,926 aggregated edges. It then uses an information-theoretic measure to quantify dependencies between genes based on gene expression data and ranks edges, accordingly. Formally, for each pathway edge, ESEA computes the differential correlation score (EdgeScore) as follows:

$$EdgeScore = \boldsymbol{MI}_{all}(i,j) - \boldsymbol{MI}_{control}(i,j) \qquad (7)$$

where $\boldsymbol{MI}_{all}$ is the mutual information of the gene expression profiles for genes i and j across all cell-types. Here, $\boldsymbol{MI}_{control}$ measures the mutual information only in the given tissue context. Each edge can be classified as either a gain of correlation (GoC), loss of correlation (LoC), or no change (NC) depending on the value of $EdgeScore$. We use GoC edges, that is, a pair of genes with positive gain of mutual information in the tissue context, as true positive edges in each tissue. Similarly, we use all positive edges in all tissues but the tissue of interest as true negatives.

To assess agreement between ESEA differential correlation scores over known pathway edges and computed tissue-specific interactions, we rank all pathway edges according to their edge weights in the human tissue-specific interactome and evaluate the enrichment of true positive pathway edges among top-ranked edges. We compute the receiver operating

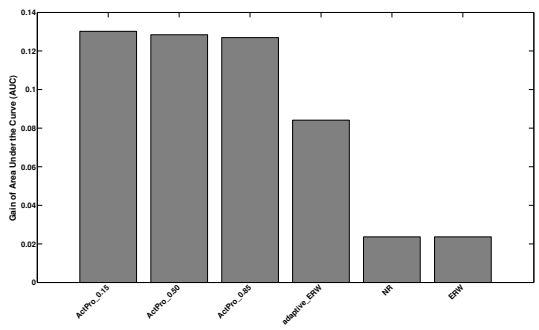


Fig. 6. Gain of Area Under the Curve (AUC) of known context-specific pathway edges among tissue-specific interactions

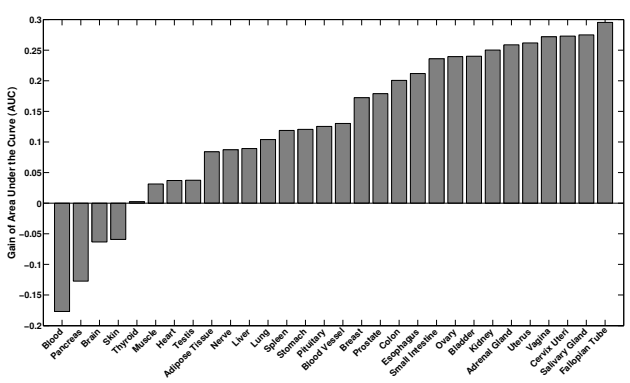


Fig. 7. Performance of ActPro with $\alpha = 0.15$ over different tissues

characteristic (ROC) curve for each tissue and average the area under the curve (AUC) gain, compared to random baseline, over all tissues. Figure 6 presents the relative performance of each method. All three configurations of the *ActPro* algorithm are ranked at the top of the list – demonstrating the superior performance of our proposed method.

To further investigate tissue-specific details for the top-ranked method, *ActPro* with $\alpha = 0.15$, we sorted AUC gain for each tissue, shown in Figure 7. This plot exhibits high level of heterogeneity, and surprisingly, four of the tissues had worse than random performance. This was consistent across all of the methods. To further understand this, we investigated the ranked list of edges and identified a high enrichment of edges with LoC among top-ranked edges. We performed enrichment analysis over these negative edges and identified significant tissue-specific functions among them, which suggests that the poor observed performance for these tissues is attributed to their misclassification as negative edges.

At the other end of the spectrum, *Fallopian Tube*, *Vagina*, and *Cervix Uteri* had consistently high AUC gain across different methods. Figure 8 shows the ROC curve for these tissues.

3.4 Tissue-specific interactions are enriched among proteins with shared tissue-specific annotations

We hypothesize that tissue-specific edges are enriched with proteins that participate in similar tissue-specific functions. To evaluate our hypothesis, we collected a set of manually curated tissue-specific Gene Ontology (GO) annotations from a recent study (Greene *et al.*, 2015). We mapped tissues to GTEx tissues and identified tissue-specific GO annotations for genes in each tissue-specific interactome. We excluded tissues with less that 100 edges with known annotations. This resulted in 10 tissues, *Adipose Tissue, Blood Vessel, Blood, Brain, Breast, Heart, Kidney, Lung*, and *Muscle*, for

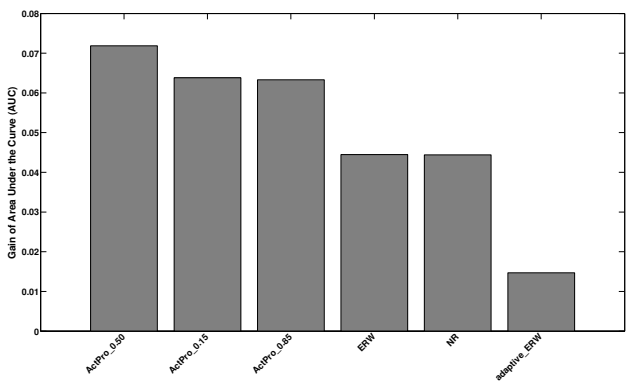


Fig. 9. Mean gain of Area under the curve (AUC) for predicting proteins co-annotated with tissue-specific functions

which we had enough annotations. We use the same strategy employed in previous section to identify the mean gain of AUC for each method, which is illustrated in Figure 9. It should be noted that the gain of AUC is much smaller here than the case with ESEA edges, which can be attributed to the sparsity of tissue-specific GO annotations. Unlike ESEA, *ActPro* with $\alpha = 0.5$ outperforms the case with $\alpha = 0.15$. Furthermore, adaptive ERW method performs the worst among all methods.

Among the ten tissues, *Adipose* and *Muscle* tissues performed marginally better than the others with AUC of 0.59 and 0.58, respectively. On the other hand, *Lung* tissue had the worst performance with lower than random AUC of 0.47.

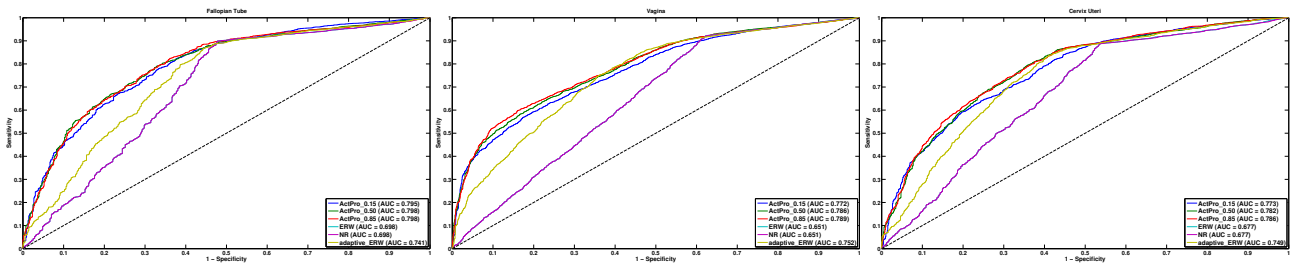
3.5 Tissue-specific interactions densely connect genes corresponding to tissue-specific disorders

Disease genes are densely connected to each other in the interactome, which provides the basis for a number of methods for network-based disease gene prioritization (Köhler *et al.*, 2008). Tissue-specific interactomes have been shown to have higher performance for predicting disease-related genes using random-walk method (Magger *et al.*, 2012). More recently, Cornish *et al.* (Cornish *et al.*, 2015) used the concept of "geneset compactness", and showed that the average distance among nodes corresponding to a given disorder is significantly smaller in tissue-specific networks compared to an ensemble of random graphs.

Here, we adopt this concept with a few modifications to measure how closely tissue-specific genes related to human disorders are positioned in networks constructed using different methods. First, we use a symmetric diffusion process instead of Random-Walk with Restart (RWR), which is a better measure of distance. Second, we use an alternate random model in which genes corresponding to tissue-specific disorders are strongly connected to each other, compared to random genesets of the same size. Second, we use an alternative random model in which we hypothesize that genes corresponding to tissue-specific disorders are strongly connected to each other, compared to random genesets of the same size.

To validate our hypothesis, we gather genes corresponding to tissue-specific disorders from a recent study (Himmelstein and Baranzini, 2015). These genes are extracted from the GWAS Catalog by mapping known associations to disease-specific loci. Among a total of 99 disorders, we focused on the gold standard set of 29 diseases with at least 10 high-quality primary targets. We successfully mapped 26 of these diseases to GTEx tissues, which are used for the rest of our study.

For a given tissue-specific interactome represented by its adjacency matrix, $\mathbf{A}_T$, we define a stochastic matrix $\mathbf{S} = \mathbf{\Delta}^{-\frac{1}{2}}\mathbf{A}_T\mathbf{\Delta}^{-\frac{1}{2}}$, where $\mathbf{\Delta}$ is the diagonal matrix with entries δ_{ii} being the degree of node i in



(a) Fallopian Tube (b) Vagina (c) Cervix Uteri

Fig. 8. Tissues with the highest gain of AUC for predicting tissue-specific pathway edges

the human tissue-specific interactome. Using this matrix, we can compute distances among gene pairs as:

$$\mathbf{D} = (1-\alpha)(\mathbf{I} - \alpha\mathbf{S})^{-1}, \tag{8}$$

Given a disease geneset Γ, we measure its compactness as the normalized average of distances for all pairs of nodes in the geneset:

$$\kappa_\Gamma = \frac{\sum_{i\neq j\in\Gamma} d_{ij}}{\binom{|\Gamma|}{2}} \tag{9}$$

Finally, we sampled without replacement, 100K vertex samples of size $|\Gamma|$ from the tissue-specific interactome and computed the compactness for each of the samples, individually. We defined an empirical *p*-value as the fraction of random instances with higher compactness (lower κ) compared to Γ. We removed disorders for which none of methods yield significant *p*-value given a threshold of 10^{-2}. These diseases either have errors in their disease-tissue/gene associations, had associations with multiple tissues, or their underlying subnetwork targeted disjoint functions within the network. The final dataset consists of 10 diseases with significantly compact interactions. To combine the *p*-values for different disorders, we use the Edgington method (Edgington, 1972). This method, unlike Fisher's or Stouffer's methods, is an additive method instead of multiplicative and is less susceptible to extremely low *p*-values. Formally, it gathers a statistic $\mathcal{S} = \sum_{i=1}^{k} p_i$ for a set of k given *p-values* and computes the meta *p*-value by assigning significance to $\mathcal{S}$ as:

$$\sum_{j=0}^{\lfloor\mathcal{S}\rfloor} -1^j \binom{k}{j} \frac{(\mathcal{S}-j)^k}{k!} \tag{10}$$

The list of all individual, and combined *p*-values is shown in Table 1. Node removal method was not significant for either of cases, as it disconnects the set of disease genes. Among other methods, *ActPro* with $\alpha = 0.15$ had the most significant combined *p*-value, closely followed by *ActPro* with $\alpha = 0.5$. *ERW* method was second to last, only better than NR method. This suggests that using actual activity scores instead of fixed penalty enhances construction of tissue-specific networks. Furthermore, using the global network context instead of only local context further improves the network quality.

3.6 Tissue-specific interactome identifies novel disease-related pathways – case study in neurodegenerative disorders

We now investigate whether tissue-specific interactomes can help in predicting novel pathways that are involved in the progression of neurodegenerative disorders. We perform a case study of *Alzheimer's* and *Parkinson's* diseases, both of which were among disorders with high compactness in brain tissue. We use Prize-Collecting Steiner Tree (PCST) algorithm to identify the underlying pathway among disease-genes identified by GWAS studies. Formally, PCST problem can be formulated as:

$$\underset{<v,e>\in T}{\operatorname{argmin}} \Big\{ \sum_e c_e - \lambda \sum_v b_v \Big\}, \tag{11}$$

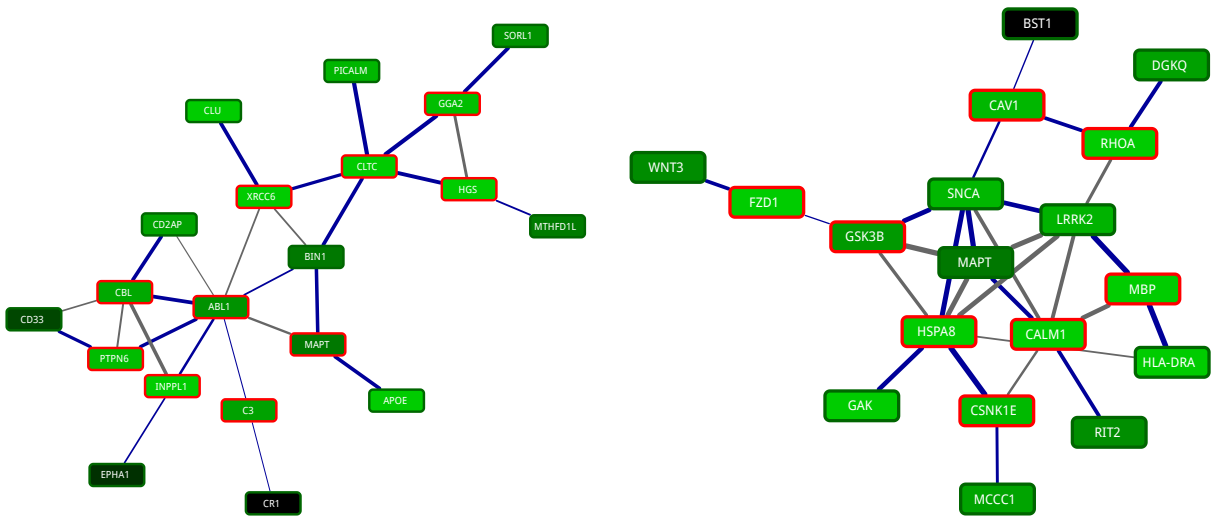
where T is an induced tree of the given graph, v and e are the set of vertices and edges in T, respectively, c_e is the cost of choosing edge e, and b_v is the reward/prize of collecting node v. Similar methods have been proposed previously to connect upstream signaling elements to downstream transcriptional effector genes (Tuncbag *et al.*, 2012, 2013).

To identify disease-related pathways, we first prune non-specific interactions in the network by removing vertices that have more than 500 interactions. We transform edge confidence values (conductances) to edge penalties (resistances) by inverting each edge weight. Node prizes are defined as the ratio of their incident edges that fall within disease-related genes to the total degree of a node. We assigned a node prize of 1,000 to disease genes to ensure that they are selected as terminal nodes. Finally, we use a recent message passing algorithm (Bailly-Bechet *et al.*, 2011) to identify PCST rooted at each disease-related gene and choose the best tree as the backbone of the disease-related pathway. Over each node, we use a maximum depth of 4 and $\lambda = 1$ as parameters to the message passing algorithm. Figure 10 shows final tissue-specific pathways for Alzheimer's and Parkinson's diseases.

Alzheimer's disease (AD) network contains two distinct subnetworks, one centered around *CLTC* and the other centered around *ABL1*. *PICALM, CLU, APOE*, and *SORL1* are all known genes involved in AD, which are also involved *negative regulation of amyloid precursor protein catabolic process*. All four of these genes converge on *CLCT* gene, but through different paths. *PICALM* gene is known to play a central role in clathrin-related endocytosis. This protein directly binds to CLTC and recruits clathrin and adaptor protein 2 (AP-2) to the plasma membrane (Carter, 2011). On the other hand, *CLU, APOE*, and *SORL1* are linked to the *CLTC* through novel linker genes *XRCC6, MAPT/BIN1* and *GG2A/HGS*, respectively. Gamma-adaptin gene, GGA2, binds to clathrins and regulates protein traffic between the Golgi network and the lysosome. This network is postulated to be an important player in AD (Carter, 2011). *HGS* gene is a risk factor age-related macular degeneration (AMD) and has been hypothesized to be a shared factor for AD (Logue *et al.*, 2014). Interestingly, *MAPT*, a novel marker identified in this study, is a risk factor for Parkinson's disease and very recently shown to also be linked to AD (Desikan *et al.*, 2015). A second component in AD network is centered around *ABL1* gene, which, together with *CBL, INPPL1, CD2AP*, and *MAPT* share the *SH3 domain binding* function. *INPPL1* gene, a metabolic syndrome risk factor, has been hypothesized to link AD and recently posed term *Type 3 diabetes* (Accardi *et al.*, 2012; de la Monte and Wands, 2008).

Table 1. Compactness of tissue specific disease genes in their tissue-specific interactome

	ActPro_0.15	ActPro_0.50	ActPro_0.85	ERW	NR	adaptive_ERW
type 2 diabetes mellitus	**2.10E-4**	2.90E-4	1.23E-3	6.60E-4	1.00E+0	1.38E-3
breast carcinoma	5.80E-4	**4.90E-4**	8.00E-4	3.92E-3	1.00E+0	8.50E-4
chronic lymphocytic leukemia	7.30E-4	5.10E-4	6.10E-4	8.10E-4	1.00E+0	**4.90E-4**
psoriasis	**7.40E-4**	1.44E-3	2.70E-3	5.25E-3	1.00E+0	3.30E-3
primary biliary cirrhosis	**1.47E-3**	3.25E-3	2.97E-3	2.86E-2	1.00E+0	3.45E-3
rheumatoid arthritis	**4.58E-3**	9.37E-3	1.56E-2	6.66E-2	1.00E+0	1.81E-2
Alzheimer's disease	6.69E-3	5.53E-3	5.83E-3	5.52E-3	1.00E+0	**5.28E-3**
metabolic syndrome X	**9.92E-3**	2.25E-2	5.26E-2	1.26E-1	1.00E+0	5.18E-2
Parkinson's disease	1.14E-2	9.75E-3	9.93E-3	1.42E-2	1.00E+0	**9.24E-3**
systemic lupus erythematosus	1.26E-2	7.36E-3	6.99E-3	**2.67E-3**	1.00E+0	6.86E-3
combined	**2.16E-20**	1.81E-19	2.56E-17	3.11E-13	1.00E00	2.97E-17



(a) Alzheimer's Disease (b) Parkinson's Disease

Fig. 10. Tissue-specific pathways in human neurodegenerative disorders. Nodes are colored according to their tissue-specific expressions, with novel identified genes marked in red, accordingly. The thickness of edges represent their confidence with tree edges marked as blue.

Finally, we note that *MAPT* gene is one of the central genes that links these two main components, the role of which warrants further investigation.

Parkinson's disease (PD) network, on the other hand, contains one densely connected core centered around *MAPT* gene. There are two main branches converging on *MAPT*. On the left, *WNT3, FZD1*, and *GSK3B* constitute upstream elements of the WNT signaling pathway, which is known to play an important role in PD neurodegeneration (Berwick and Harvey, 2012). *GSK3* gene product is postulated to directly interact with *MAPT* (τ and *LRKK2*, while implicitly regulating *SNCA* (α-Syn) in a β-cat dependent manner. However, we observed direct interaction between *GSK3B* and **SNCA**, and parallel pathways connecting it to *LRRK2* via *SNCA* and *MAPT*. Both *SNCA* and *MAPT* also take part in the right branch, together with *CAV1* and *RHOA*, which is enriched in *reactive oxygen species metabolic process*. Accumulation of ROS contributes to mitochondrial dysfunction and protein misfolding, both of which are linked to progression of PD. *RIT2* enzyme is identified independently and confirmed as PD susceptibility factor (Pankratz *et al.*, 2012). Pankratz *et al.* also suggested *CALM1* as the bridge linking *RIT2* with *MAPT* and *SNCA*, which confirms our findings. Cyclin G associated kinase (*GAK*) is a known risk factor for PD. We identified *HSPA8* as a key link between *GAK*, WNT signaling pathway, and *CSNK1E* with central PD genes, *MAPT, SNCA,* and *LRRK2*. *HSPA8* gene has been proposed as a biomarker for diagnosis of PD (Lauterbach, 2013). Finally, myelin basic protein (*MBP*) interacts closely with *CALM1* and *LRRK2*. This gene has been previously shown to be differentially expressed in PD and proposed as potential biomarker for PD (Kim *et al.*, 2006).

In summary, we show that the brain-specific interactome derived from our method helps in uncovering tissue-specific pathways that are involved in neurodegenerative diseases. Similar analysis of other human tissues can potentially contribute to identification of new therapeutic targets for other human disorders.

4 Conclusion

In this paper, we present a novel method for computing tissue-specific interactomes from organism-specific interactomes and expression profiles of genes in various tissues. Our method casts the problem as a convex optimization problem that diffuses functional activity of genes over the organism-specific interactome, while simultaneously minimizing perturbation of transcriptional activity. We show, using a number of validation studies, that the tissue-specific interactomes computed by our method, are superior to those computed using existing methods. Finally, we show, using a case study of brain-specific interactome for Alzheimer's and Parkinson's diseases, that our method is capable of constructing highly resolved disease-specific pathways, providing potential targets for novel drugs.

5 Funding

This work is supported by the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370, and by NSF Grant BIO 1124962.

References

Accardi, G. et al (2012). Can Alzheimer Disease Be a Form of Type 3 Diabetes? *Rejuvenation Research*, **15**(2), 217–221.

Ardlie, K.G. et al (2015). The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, **348**(6235), 648–660.

Bailly-Bechet, M. et al (2011). Finding undetected protein associations in cell signaling by belief propagation. *Proceedings of the National Academy of Sciences*, **108**(2), 882–887.

Berwick, D.C. and Harvey, K. (2012). The importance of Wnt signalling for neurodegeneration in Parkinson's disease. *Biochemical Society transactions*, **40**(5), 1123–8.

Bossi, A. and Lehner, B. (2009). Tissue specificity and the human protein interaction network. *Molecular systems biology*, **5**, 260.

Buljan, M. et al (2012). Tissue-Specific Splicing of Disordered Segments that Embed Binding Motifs Rewires Protein Interaction Networks. *Molecular Cell*, **46**(6), 871–883.

Carter, C. (2011). Alzheimer's Disease: APP, Gamma Secretase, APOE, CLU, CR1, PICALM, ABCA7, BIN1, CD2AP, CD33, EPHA1, and MS4A2, and Their Relationships with Herpes Simplex, C. Pneumoniae, Other Suspect Pathogens, and the Immune System. *International Journal of Alzheimer's Disease*, **2011**, 1–34.

Chatterjee, R. and Vinson, C. (2012). CpG methylation recruits sequence specific transcription factors essential for tissue specific gene expression. *Biochimica et biophysica acta*, **1819**(7), 763–70.

Cornish, A.J. et al (2015). Exploring the cellular basis of human disease through a large-scale mapping of deleterious genes to cell types. *Genome Medicine*, **7**(1), 95.

de la Monte, S.M. and Wands, J.R. (2008). Alzheimer's disease is type 3 diabetes-evidence reviewed. *J Diabetes Sci Technol*, **2**(6), 1101–1113.

Desikan, R.S. et al (2015). Genetic overlap between AlzheimerâĽ™s disease and ParkinsonâĽ™s disease at the MAPT locus. *Molecular Psychiatry*, **20**(12), 1588–1595.

Dezso, Z. et al (2008). A comprehensive functional analysis of tissue specificity of human gene expression. *BMC biology*, **6**, 49.

Edgington, E.S. (1972). An Additive Method for Combining Probability Values from Independent Experiments. *The Journal of Psychology*, **80**(2), 351–363.

Eisenberg, E. and Levanon, E.Y. (2013). Human housekeeping genes, revisited.

Ellis, J.D. et al (2012). Tissue-specific alternative splicing remodels protein-protein interaction networks. *Molecular cell*, **46**(6), 884–92.

Göring, H. (2012). Tissue specificity of genetic regulation of gene expression. *Nature Genetics*, **44**(10), 1077–1078.

Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In V. Blondel, S. Boyd, and H. Kimura, editors, *Recent Advances in Learning and Control*, Lecture Notes in Control and Information Sciences, pages 95–110. Springer-Verlag Limited. `http://stanford.edu/~boyd/graph_dcp.html`.

Greene, C.S. et al (2015). Understanding multicellular function and disease with human tissue-specific networks. *Nature genetics*, **32**(4), 453–465.

Han, J. et al (2015). ESEA: Discovering the Dysregulated Pathways based on Edge Set Enrichment Analysis. *Scientific reports*, **5**, 13044.

Himmelstein, D.S. and Baranzini, S.E. (2015). Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes. *PLOS Computational Biology*, **11**(7), e1004259.

Ikegami, K. et al (2014). Tissue-Specific Posttranslational Modification Allows Functional Targeting of Thyrotropin. *Cell Reports*, **9**(3), 801–809.

Kim, J.M. et al (2006). Identification of genes related to Parkinson's disease using expressed sequence tags. *DNA research : an international journal for rapid publication of reports on genes and genomes*, **13**(6), 275–86.

Köhler, S. et al (2008). Walking the interactome for prioritization of candidate disease genes. *American journal of human genetics*, **82**(4), 949–58.

Kolde, R. (2014). *GOsummaries: Word cloud summaries of GO enrichment analysis*. R package version 2.0.0.

Lage, K. et al (2008). A large-scale analysis of tissue-specific pathology and gene expression of human disease genes and complexes. *Proceedings of the National Academy of Sciences of the United States of America*, **105**(52), 20870–5.

Lauterbach, E.C. (2013). Psychotropics regulate Skp1a, Aldh1a1, and Hspa8 transcription âĽ" Potential to delay Parkinson's disease. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, **40**, 236–239.

Lim, W.K. et al (2007). Comparative analysis of microarray normalization procedures: effects on reverse engineering gene networks. *Bioinformatics*, **23**(13), i282–i288.

Logue, M.W. et al (2014). A search for age-related macular degeneration risk variants in Alzheimer disease genes and pathways. *Neurobiology of Aging*, **35**(6), 1510.e7–1510.e18.

Magger, O. et al (2012). Enhancing the prioritization of disease-causing genes through tissue specific protein interaction networks. *PLoS computational biology*, **8**(9), e1002690.

Mele, M. et al (2015). The human transcriptome across tissues and individuals. *Science*, **348**(6235), 660–665.

Mendizabal, I. et al (2014). Epigenetics and evolution. *Integrative and comparative biology*, **54**(1), 31–42.

Messina, D.N. et al (2004). An ORFeome-based analysis of human transcription factor genes and the construction of a microarray to interrogate their expression. *Genome research*, **14**(10B), 2041–7.

MOSEK-ApS (2015). *The MOSEK optimization toolbox for MATLAB manual. Version 7.1 (Revision 28).*

Pankratz, N. et al (2012). Meta-analysis of Parkinson's Disease: Identification of a novel locus, RIT2. *Annals of Neurology*, **71**(3), 370–384.

Piccolo, S.R. et al (2013). Multiplatform single-sample estimates of transcriptional activation. *Proceedings of the National Academy of Sciences of the United States of America*, **110**(44), 17778–83.

Raj, T. et al (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science (New York, N.Y.)*, **344**(6183), 519–23.

Razick, S., Magklaras, G. and Donaldson, I.M. (2008). iRefIndex: a consolidated protein interaction database with provenance. *BMC bioinformatics*, **9**(1), 405.

Reimand, J., Arak, T. and Vilo, J. (2011). g:Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, **39**(suppl), W307–W315.

Rolland, T. et al (2014). A Proteome-Scale Map of the Human Interactome Network. *Cell*, **159**(5), 1212–1226.

Sahni, N. et al (2015). Widespread Macromolecular Interaction Perturbations in Human Genetic Disorders. *Cell*, **161**(3), 647–660.

Tuncbag, N. et al (2012). SteinerNet: A web server for integrating 'omic' data to discover hidden components of response pathways. *Nucleic Acids Research*, **40**, 1–5.

Tuncbag, N. et al (2013). Simultaneous reconstruction of multiple signaling pathways via the prize-collecting steiner forest problem. *Journal of computational biology : a journal of computational molecular cell biology*, **20**(2), 124–36.

Uhlen, M. et al (2015). Tissue-based map of the human proteome. *Science*, **347**(6220), 1260419–1260419.

Vaidyanathan, K. and Wells, L. (2014). Multiple Tissue-specific Roles for the O -GlcNAc Post-translational Modification in the Induction of and Complications Arising from Type II Diabetes. *Journal of Biological Chemistry*, **289**(50), 34466–34471.

Vanunu, O. et al (2010). Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*, **6**(1), e1000641.

Vaquerizas, J.M. et al (2009). A census of human transcription factors: function, expression and evolution. *Nature reviews. Genetics*, **10**(4), 252–263.

Wang, X. et al (2012). Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nature Biotechnology*, **30**(2), 159–164.