Bioinformatics, 32(14), 2016, 2167–2175
doi: 10.1093/bioinformatics/btw151
Advance Access Publication Date: 19 March 2016
Original Paper

OXFORD

Data and text mining

Characterizing gene sets using discriminative random walks with restart on heterogeneous biological networks

Charles Blatti[1] and Saurabh Sinha[1,2,*]

[1]Department of Computer Science, University of Illinois, Urbana, IL 61801, USA and [2]Institute of Genomic Biology, University of Illinois, Urbana, IL 61801, USA

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on November 12, 2015; revised on February 17, 2016; accepted on March 14, 2016

Abstract

Motivation: Analysis of co-expressed gene sets typically involves testing for enrichment of different annotations or 'properties' such as biological processes, pathways, transcription factor binding sites, etc., one property at a time. This common approach ignores any known relationships among the properties or the genes themselves. It is believed that known biological relationships among genes and their many properties may be exploited to more accurately reveal commonalities of a gene set. Previous work has sought to achieve this by building biological networks that combine multiple types of gene–gene or gene–property relationships, and performing network analysis to identify other genes and properties most relevant to a given gene set. Most existing network-based approaches for recognizing genes or annotations relevant to a given gene set collapse information about different properties to simplify (homogenize) the networks.

Results: We present a network-based method for ranking genes or properties related to a given gene set. Such related genes or properties are identified from among the nodes of a large, heterogeneous network of biological information. Our method involves a random walk with restarts, performed on an initial network with multiple node and edge types that preserve more of the original, specific property information than current methods that operate on homogeneous networks. In this first stage of our algorithm, we find the properties that are the most relevant to the given gene set and extract a subnetwork of the original network, comprising only these relevant properties. We then re-rank genes by their similarity to the given gene set, based on a second random walk with restarts, performed on the above subnetwork. We demonstrate the effectiveness of this algorithm for ranking genes related to *Drosophila* embryonic development and aggressive responses in the brains of social animals.

Availability and Implementation: DRaWR was implemented as an R package available at veda.cs.illinois.edu/DRaWR.

Contact: blatti@illinois.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

A common task in bioinformatics is to characterize co-expressed gene sets using enrichment methods, such as Hypergeometric tests or gene set enrichment analysis (GSEA) (Subramanian *et al.*, 2005), associating the gene set with other previously annotated sets. These pre-existing gene sets may be defined from many diverse types of biological knowledge, such as shared protein domains, evolutionary origins, biological processes, etc. Public databases of curated

annotations that enable this paradigm of gene set characterization are highly diverse and rapidly increasing. This work addresses the challenge of incorporating heterogeneous data from multiple public resources into the task of characterizing the shared properties of a given gene set and identifying additional genes that are important and related.

One broad approach employed to perform gene set analysis with these different public resources is to represent the data as a biological network. Rather than using each data source one at a time to analyze a co-expressed gene set, sources may be integrated within a network and simultaneously leveraged to identify related genes. This idea was tested in the 'MouseFunc' challenge (Pena-Castillo *et al.*, 2008), where nine algorithms for integrating heterogeneous genomic evidence on mouse genes were evaluated for their ability to discover genes related to a set of co-functional genes. Network-based algorithms have also been applied to other important bioinformatics tasks such as understanding causes of diseases and effects of therapies (Chen *et al.*, 2012; Greene *et al.*, 2015; Hou and Ma, 2014; Jacquemin and Jiang, 2013; Vaske *et al.*, 2010). Network-based analysis of gene sets in particular has been designed to extend and annotate gene functions and modules (Reimand *et al.*, 2008; Wang *et al.*, 2015), quantify gene set enrichment for functional molecular networks (Cornish and Markowetz, 2014; Tarca *et al.*, 2009), identify subnetworks affected in or shared across diseases (Leiserson *et al.*, 2015; Shen *et al.*, 2012), or cluster and find signatures of cancer subtypes (Hofree *et al.*, 2013; Liu *et al.*, 2014).

Most gene set analyses performed on a biological network encompassing heterogeneous data types lose a significant portion of the data during network construction. Frequently, the rich and diverse public datasets are converted to homogeneous gene–gene networks containing only nodes representing genes of a single species and unweighted edges of a single type (Cornish and Markowetz, 2014; Hofree *et al.*, 2013; Hou and Ma, 2014). In these homogenous networks, an edge only represents a relationship between a pair of genes, but details about the different types of evidence for that relationship are lost. Algorithms that rely on these networks assume that all relationships in the network are as reliable as any other. Other algorithms rely on improved, weighted networks of gene–gene interactions that estimate the confidence of each edge by integrating the strengths and types (i.e. source database or experimental assay) of the evidences for each relationship (Cornish and Markowetz, 2014). There are also some studies that utilize biological networks containing more than one edge or node type (Chen *et al.*, 2012; Li and Patra, 2010). However, the networks in these studies usually have a structure specific to their system of interest, most often containing nodes of two different types and three types of edges capturing similarity within each type of node sets and the known relationships between them. Although they construct heterogeneous networks, these studies strictly rely on the structure of the problem and do not attempt to incorporate data from all possible sources.

GeneMANIA (Warde-Farley *et al.*, 2010) is a popular, network-based gene ranking algorithm that performed well in the MouseFunc evaluations. Its approach specifically integrates data from many different sources without sacrificing the edge source information. Data from each source informs the creation of its own 'affinity' network of gene–gene interactions. Different affinity networks are up- or down- weighted based on their relevance to the original functional gene set before being combined into a single composite and homogeneous network (Mostafavi and Morris, 2012). While the GeneMANIA approach works well and considers the types of sources that are most important to the ranking task, it still discards the specific details about the gene–gene relationships when constructing each affinity network. For example, the edges within a Pfam protein domain affinity network indicate that a pair of genes share a protein domain sequence, but does not preserve which domain(s) it may have been.

We developed the DRaWR ('Discriminative Random Walk with Restarts) method to rank genes for their relatedness to a given gene set, using biological networks that maintain detailed information from public data sources. Our algorithm is explicitly designed to work on heterogeneous networks with multiple node types that are able to represent a complete collection of public, genomic knowledge. We believe that DRaWR is the first method of its genre with this ability. We utilized the algorithm to perform the gene ranking task and simultaneously return the most relevant network features. Like many other network-based algorithms that rely on 'guilt-by-association' approaches (Hofree *et al.*, 2013; Hou and Ma, 2014; Ivan and Grolmusz, 2011), our algorithm implements a modified random walk with restart (RWR). However, unlike other methods, we employed two rounds of RWR: a first round of RWR on the large, noisy network of all public data, which reports the network nodes related to the given ('query') gene set, and a second stage RWR on a smaller network that includes only the query-relevant nodes from the original network. We evaluated our method's ability to recover left out genes from the expression domain gene sets of *Drosophila* embryonic development. We showed that our gene ranking method improves when multiple data sources are combined and when data from additional species are added to the original network. We also found that the novel 'two-round' RWR approach performs better than the more common single-round RWR. We finally applied the DRaWR algorithm to a multi-species study of intruder response in social animals (Rittschof *et al.*, 2014) to identify subtle and shared genetic 'toolkits' that underlie aggressive behavior.

2 Methods

2.1 Building a heterogeneous network

Our first task was to construct a heterogeneous network of biological knowledge, which represents prior information from multiple public resources. Our network was composed of 'gene' nodes representing the corresponding gene and proteins from each of eight different species (Supplementary Table S1) and 'feature' nodes that represent experimentally or computationally derived characteristics or properties of genes or proteins. The first type of edge we added to the network was an undirected 'homology' edge. These edges connect a pair of gene nodes with significant protein sequence similarity (BLAST *e*-value score < 0.01). Additionally, we assigned weights to the homology-based edges that are calculated from the *z*-transform of their *e*-value significance (maximum value is set to a *z*-score of 8).

The other edge types we created connect feature nodes to gene nodes with undirected edges and weights proportional to the reliability of the feature annotation. To incorporate protein structure data into our network, we included ~3700 feature nodes (Supplementary Table S2) representing different protein domains from Pfam (Finn *et al.*, 2014). We then connected each such feature node (called 'prot_domain' nodes) to all of the gene nodes whose protein contained that domain, as identified by HMMER (Finn *et al.*, 2011) scans (*e*-value < 0.01). The weight of the new edge was the *z*-transform of the HMMER *e*-value score of that domain in that gene. Homology and protein domain information was included for every species included in the network.

Additionally, in our application to *Drosophila* gene sets, we introduced hundreds of 'motif' feature nodes that represent distinct binding specificities of fruitfly transcription factors (TFs). A motif node was connected to all genes whose 5 kb upstream regulatory region contain the motif, i.e. if the regulatory region includes one of the top 0.5% of the highest scoring 500 bp windows genome-wide for that motif, as scored by the Stubb program (Sinha *et al.*, 2006). The weights on these edges were the z-transform of that window's empirical *P*-value (see Supp Methods SM1, Supplementary Table S3). Also for the *D.melanogaster* study, we incorporated 75 'ChIP' feature nodes, representing TF occupancy obtained from separate ChIP-seq experimental datasets corresponding to the early fruit fly embryo (SM2, Supplementary Table S4). Each 'ChIP' feature node represented an experimental assay and was connected to a gene node if the TF binds to the gene's 5 kb upstream regulatory region in the developmental stage assayed. Edge weights were assigned in the same way as 'motif'-gene edges mentioned above.

For the network used to study aggression across species, we defined 1827 'Gene Ontology' feature nodes, each one representing a term from Gene Ontology (Ashburner *et al.*, 2000). GO annotations for three species (human, mouse and fly) were downloaded from Ensembl (Cunningham *et al.*, 2014) and only terms with at least 20 annotated genes in each of the three species became feature nodes and were connect to their annotated genes in the three species (Supplementary Table S5). GO-gene edges had weight 2 if the GO annotation was curated and weight 1 if it was inferred computationally. Also for the aggression study, we added 12 mouse-specific 'brain atlas' feature nodes derived from gene expression information produced as part of the Allen Brain Atlas (Lein *et al.*, 2007) (Supplementary Table S6). Each 'brain atlas' node corresponded to a specific region of the mouse brain and was connected with an edge of weight 1 to the 100 genes that are most specifically expressed in that region.

For each application of our algorithm, we created a weighted, undirected network, choosing some or all of the above-mentioned components, as appropriate (described in Section 3). Our initial network was constructed with gene nodes G and sets of feature nodes for each different type, $F_1, F_2, \ldots F_k$, (e.g. 'motif', 'brain atlas', etc.). We represented the edges of this network with an adjacency matrix with the form

$$M = \begin{bmatrix} M_{GG} & M_{GF_1} & \cdots & M_{GF_k} \\ M_{F_1G} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ M_{F_kG} & \cdots & \cdots & M_{F_kF_k} \end{bmatrix} \quad (1)$$

where all of the homology edges were contained in the submatrix M_{GG}, while M_{F_iG} and M_{GF_i} were the submatrices that represent (weights of) edges between all feature nodes of type i and gene nodes in G. There were no edges between feature nodes, meaning $M_{F_iF_j} = 0$ for all i, j.

2.2 Functional annotation from two-stage random walk

Given a heterogeneous biological network M, a gene set Q referred to as the 'query' set, and the universe U of all genes to rank ($U \subseteq G$), we employed a two-stage algorithm based on a modified random walk with restart (RWR) approach (Tong *et al.*, 2006) to rank the gene nodes of U. The algorithm additionally ranks the feature nodes in the network M by their relevance to the query set Q. The intuition of how an RWR algorithm works is often understood with a 'walker' that traverses the nodes of a network. With probability (1-c), where c is the restart parameter, the walker follows an outgoing edge to a neighboring node and with probability c, the walker resets the walk by transporting directly to one of the genes in the 'restart set', defined as the query set Q in our algorithm. In properly formed networks in the long run, the probability distribution of the walker over all nodes will converge to a stationary distribution. This distribution produces a ranking on all nodes that incorporates the connectedness of the node in the network as well as the proximity of the node to the query set. In the first stage of our DRaWR algorithm, we applied RWR to find the highest-ranking feature nodes related to the query set Q to extract a relevant subnetwork (those feature nodes, all gene nodes G and edges involving them) of the initial network. The results of the second stage RWR on the subnetwork provide us the final rankings of gene nodes in U. Both stages are described in detail below and summarized in Figure 1.

2.2.1 Algorithm design

Before applying our DRaWR algorithm, we first must normalize the edge weights in the initial heterogeneous, biological network. We normalized the weights of all edges of the same type (e.g. all homology edges, or all edges connecting genes to feature nodes of a particular type) to create the normalized adjacency matrix N. In terms of our notation in Equation 1, all the entries of each non-zero submatrix M_{XY} are normalized to sum to 1:

$$(N_{XY})_{i,j} = {}^{(M_{XY})_{i,j}}\Big/_{\sum_{i,j}(M_{XY})_{i,j}} \quad (2)$$

We did this to equalize the global probability of the walker following a specific edge type. For example, even though edges connecting genes to motif nodes might account for 10 times the total weight as edges involving prot_domain nodes, this heuristic adjusted the edge weights so the walker takes motif edges as often as prot_domain edges overall.

Next we normalized each of the columns the matrix N to form a transition probability matrix, A.

$$A_{i,j} = {}^{N_{i,j}}\Big/_{\sum_i N_{i,j}} \quad (3)$$

The value $A_{i,j}$ is the probability that the walker following an outgoing edge will transition from node j to node i.

We define $\mathbf{v^t}$ to be the probability distribution of the walker over all nodes in the network after t steps of the RWR algorithm. We initialized this probability distribution, $\boldsymbol{v}^0$, to be the uniform distribution over all nodes by default. A single step of the random walk follows the equation:

$$\mathbf{v}^{t+1} = (1 - \mathrm{c})\mathbf{A}\mathbf{v}^{\mathbf{t}} + \mathrm{c}\boldsymbol{\alpha} \quad (4)$$

where c is the restart probability and $\boldsymbol{\alpha}$ reflects the probability of jumping to a gene in the restart set. When the restart set is defined as the set of query genes Q, then

$$\boldsymbol{\alpha}_i^{\mathrm{Q}} = \begin{cases} {}^1\Big/_{|\mathrm{Q}|} & \text{for gene nodes in Q} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

As the random walk is irreducible and aperiodic, the iterative update of this procedure is guaranteed to converge to the stationary distribution of the random walk regardless of the initial probability distribution $\boldsymbol{v}^0$. We ran iterations of the RWR with the query set defining the restart set ($\boldsymbol{\alpha} = \boldsymbol{\alpha}^{\mathrm{Q}}$) until the vector $\boldsymbol{v}^t$ converged

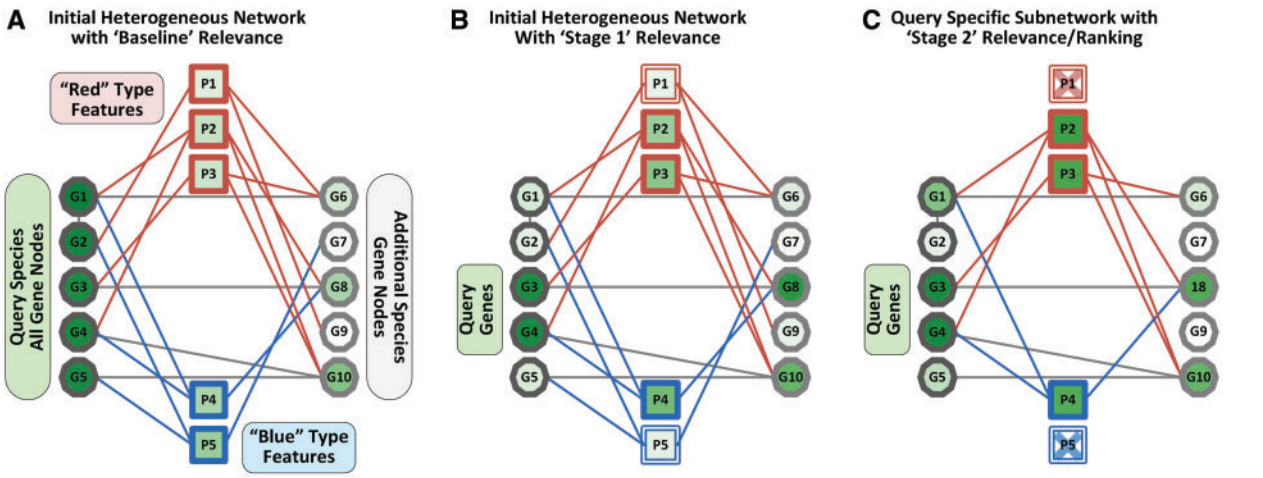

Fig. 1. Illustration of DRaWR Method. Given a set of genes called the query set, $Q=\{G3, G4\}$, DRaWR will rank all remaining genes from the query species, $U=\{G1, G2, \ldots, G5\}$, based on their relevance from a random walk with restart (RWR) in a heterogeneous network of biological knowledge. In this example, the heterogeneous network contains gene nodes from two species (circles) and feature nodes from two public resources (squares) which are connected by gene–gene sequence homology and by feature-gene annotations (edges). First, DRaWR will find the subset of feature nodes that are specific to Q, (*P2, P3, P4* in this example), by comparing the relevance (shading) of the nodes between (**a**) a 'baseline' RWR on the entire network with U as the restart set and (**b**) a 'stage 1' RWR on the entire network with Q as the restart set. A subnetwork is created using only the feature nodes that are specific to Q and (**c**) a 'stage 2' relevance is calculated from a RWR on this subnetwork with Q as the restart set. DRaWR finally ranks the genes in U for their similarity to the initial query set, Q, based on the 'stage 2 'relevance scores

($|\boldsymbol{v}^{t+1} - \boldsymbol{v}^{t}| < 0.05$). We notate this converged probability distribution as $\tilde{v}_Q$ (see Fig. 1B). The ranking of all nodes of M by the probabilities of $\tilde{v}_Q$ is referred to as the 'stage 1 query ranking'.

We wanted the ranking from the first stage to discriminate feature nodes that are related to the query set Q from those feature nodes that have high ranking in $\tilde{v}_Q$ simply due to their high connectivity in the network. To do this, we must also produce a ranking of nodes that does not depend on the query set. Therefore, in the first stage of DRaWR, we repeated the RWR procedure using the universe set U of all genes as the restart set (in place of set Q above). We thus arrived at a second converged relevance vector $\tilde{v}_U$ (see Fig. 1A) and refer to the ranking it induces on all nodes as the 'stage 1 baseline ranking'. Note, $\tilde{v}_U$ captures the overall relevance/importance of each node in the network without regard to the query set, whereas $\tilde{v}_Q$ incorporates overall network structure as well as proximity to the query set. Therefore, to find the feature nodes most specifically relevant to the query genes, we examined the difference between these vectors, $\tilde{v}_Q - \tilde{v}_U$.

For the second stage of our two-stage RWR, we selected the $50 \times k$ (k is the number of feature types) most query-specific feature nodes, defined as having the greatest values in $\tilde{v}_Q - \tilde{v}_U$, and created a subnetwork M' from the initial matrix M by removing all other feature nodes and their adjacent edges. Thus,

$$M' = \begin{bmatrix} M_{GG} & M_{GF_1'} & \cdots & M_{GF_k'} \\ M_{F_1'G} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ M_{F_k'G} & \cdots & \cdots & M_{F_k'F_k'} \end{bmatrix} \quad (6)$$

where F_i' represented only the selected feature nodes of feature type i. Using the same normalization procedure as above, we renormalized M' by type and converted it to the transition probability matrix A'. We repeated the random walk using A' and $\boldsymbol{\alpha}^{Q}$ (restart set defined from the query set Q) until we converged to the new relevance vector $\tilde{v}_Q'$ (see Fig. 1C). The ranking of all nodes induced by this new relevance vector was called the 'stage 2 query ranking'.

2.2.2 Evaluation of two stage RWR algorithm

We employed a cross validation scheme to evaluate the results of our ranking method. For each given query gene set, we held out 10% of the genes for testing, Q_{Te}, and the remaining 90% of the gene set are supplied to the algorithm as the query set Q_{Tr}. With a query set Q_{Tr}, we produced the 'stage 1 query rankings', identified the relevant features nodes, extracted the query-specific subnetwork, and repeated the RWR to produce the stage 2 query ranking. From the calculated rankings and the held out test sets Q_{Te}, we produced receiver operating characteristic (ROC) curves and quantified the performance of our algorithm with the area under these curves (AUROC).

3 Results

3.1 Applications to *Drosophila* developmental genes

We first applied the DRaWR algorithm to sets of genes defined based on *in situ* hybridization images of gene expression in *Drosophila* embryos from BDGP (Tomancak *et al.*, 2002). For this analysis, we focused on 92 spatio-temporal expression patterns (or 'domains') that contained between 100 and 1200 genes with the specific expression pattern. We applied the DRaWR algorithm to genes of each expression domain separately and evaluated gene rankings with the AUROC on the held out test set. In this application, we tested the feasibility of our algorithm to find additional genes related to each query set (using the AUROC measures described above). This application is important in instances where experimental annotation of genes has a non-trivial cost (as with constructing and imaging *in situ* hybridizations). Predicting other genes that share the expression pattern of the query set can provide investigators a manageable number of additional genes to assay.

We began by creating a *Drosophila*-specific heterogeneous network that contained gene nodes connected by 'homology' edges as

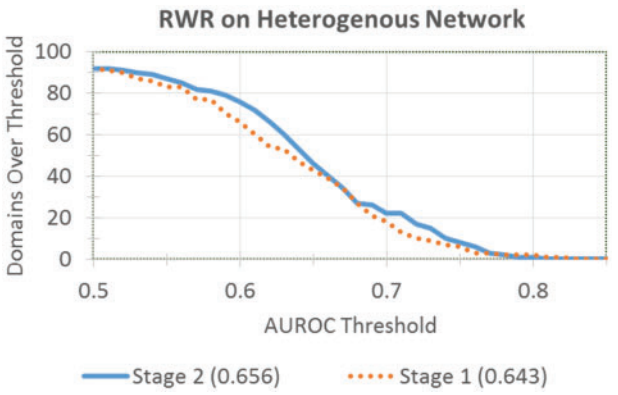

Fig. 2. Comparison of Stage 1 and 2 Rankings on *Drosophila* Heterogeneous Network. We compared the rankings produced at the end of the first stage random walk to the second stage random walk on query specific networks. We calculated the average stage 1 and stage 2 AUROCs for each of the 92 expression domains and then plot the number of domains (*y*-axis) that were above each possible AUROC threshold (*x*-axis)

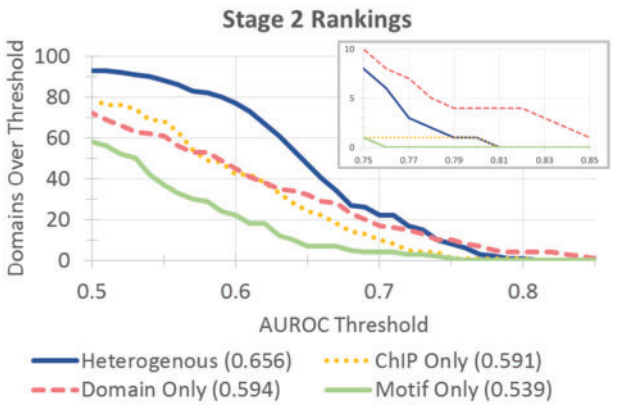

Fig. 3. Comparison of RWR on Different Drosophila Networks. We compared the stage 2 rankings produced by our algorithm when the initial network was defined by single ('Domain', 'ChIP', 'Motif') or 'Heterogeneous' feature types. We calculated the stage 2 AUROCs for each of the 92 expression domains and then plot the number of domains (*y*-axis) that were above each possible AUROC threshold (*x*-axis). The inset shows more detail for the chart region of high AUROC

well as three types of feature nodes with their corresponding edges, 'prot_domain', 'motif' and 'ChIP', as described in Section 2.1. Nodes and edges derived from Gene Ontology were excluded since they include information on the cell types corresponding to the 92 expression domains. Overall, this network had 17 482 nodes and 580 270 edges (Supplementary Table S7). For each of the 92 expression domain gene sets, we ranked the 13 609 gene nodes in this network and reported on where the held out genes fall in this ranking. We also noted the most relevant feature nodes for each gene set.

3.1.1 Results on Drosophila networks

Our first observation was that the rankings produced by two-stage RWR are better than those from the query-specific RWR in the first stage (Fig. 2 and Supplementary Table S8). For instance, the AUROC of the two-stage procedure is >0.6 for 76 of the 92 gene sets, while that of the first stage alone is >0.6 for only 66 gene sets. The improvement in the second stage RWR presumably resulted from removing features unrelated to the query gene set and performing the random walk on a more 'relevant' network. Since we do not know *a priori* which features may be important to any given set, this two-stage approach allows us to begin with all known data encoded in the network, reduce to a relevant subnetwork and produce better rankings. This is an important improvement over a majority of RWR algorithms that only produce rankings from the original networks that contain a large number of edges potentially irrelevant to the query gene set.

We next tested if and found that rankings are better due to our use of a heterogeneous network that combines data from multiple sources. Instead of the heterogeneous network (with four different edge types) that was used in the tests reported above, we produced four separate networks each with edges of a single type. We ran our two-stage algorithm on the 92 expression domain gene sets on each network and found that the heterogeneous network provides the highest AUROC on average (0.656). In general, the heterogeneous network outperformed the homogeneous 'prot_domain' and homogenous 'ChIP' networks, which were much better than the homogenous 'motif' network (Fig. 3 and Supplementary Table S9). For instance, the heterogeneous network leads to AUROC > 0.65 for 47 expression domains, significantly more than the 32 that the homogenous 'prot_domain' network achieves (32). The 'ChIP' only network was expected to outperform the 'motif' only network because the ChIP data was from the corresponding developmental stage. Interesting, the 'prot_domain' network is able to achieve very high AUROC values (>0.8) for four expression domains, while the heterogeneous network leads to this high level of accuracy only for one expression domain.

We performed several tests to evaluate different components of the DRaWR method. First, we observed that when we remove the normalization procedure that equalizes the global probability that a walker follows a particular edge type, the average AUROC results of our two-stage method on the heterogeneous network are somewhat worse (0.646) (Supplementary Table S10). We also examined the main parameter of the RWR method, the restart parameter, *c*. We ran the two-stage procedure on the heterogeneous network with six different values of the restart probability between 0 and 1. We found the best performance with the relatively high restart probability of 0.7 (Supplementary Fig. S1). The restart probability controls the influence of the network structure and the proximity of the query set on the final relevance vector. A high restart probability may be needed in the first stage to select relevant feature nodes that are more proximal to the query set than those functioning as hubs in the network. Finally, we tested whether incorporating additional gene–gene edges into the heterogeneous network (including protein and genetic interactions from BioGRID (Chatr-Aryamontri *et al.*, 2015), DIP (Salwinski *et al.*, 2004) and IntAct (Orchard *et al.*, 2014)) affected the outcome (Supplementary Methods SM3, Supplementary Table S11). As before, we found that adding additional edge types to the network overall improved the performance (average AUROC of 0.69, Supplementary Fig. S2).

3.1.2 Two stage RWR on multi-species networks

Our algorithm is designed to work with large, heterogeneous networks built from many public databases of biological knowledge. With improving high throughput sequencing techniques, the number of publicly available genomes is rapidly growing. We next sought to test whether including additional genomes in our biological network would improve ranking performance on the developmental gene sets. To this end, we constructed a '5 Insect' network with gene nodes representing genes from the fruit fly *D.melanogaster*, the mosquito *A.gambiae*, the honeybee *A.mellifera*, the jewel wasp

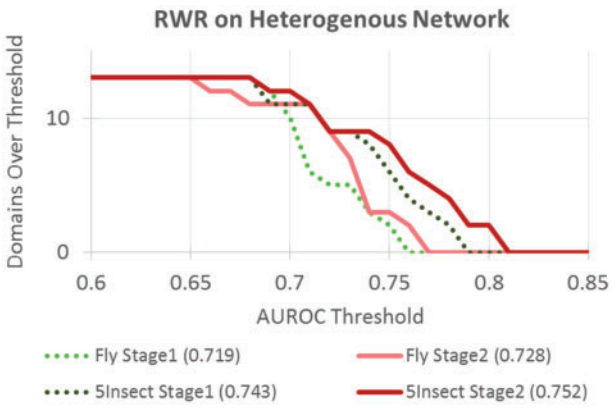

Fig 4 Comparison between Single and Multi-Species Networks. We compared the stage 1 and stage 2 rankings when the initial network was defined as the heterogeneous network either from a single species ('Fly') or from multiple species ('5Insect'). We calculated the AUROCs from each stage's rankings for each of the 12 selected Drosophila expression domain genes sets and then plot the number of domains (*y*-axis) that were above each possible AUROC threshold (*x*-axis)

N.vitripennis and the beetle *T.castaneum*. As described in Section 2.1, the gene nodes within and between the five species were connected with weighted 'homology' edges when they share high protein sequence similarity according to BLAST. Additionally, all 'prot_domain' and 'motif' feature nodes were connected to gene nodes in all five species in the manner described in Section 2.1. Since the ChIP experiments were only available for *Drosophila*, the 'ChIP' feature nodes only connect to fruit fly gene nodes. The new network had five times the number of species, but thirteen times the number of edges (Supplementary Table S12). This was mostly due to the homology edges, which account for 78% of the edges in the '5 Insect' network.

Although there were 58 147 gene nodes, spanning five species, in this new network, our task was still to rank the 13 604 gene nodes in *Drosophila* for their relatedness to a specific developmental gene set; genes from the other species were included in the network only to improve accuracy. For this reason, we calculated the stage one 'baseline' probabilities by defining the restart set as only the fruit fly genes and therefore identifying the relevance of the features nodes with respect to the network and the *Drosophila* genes. This careful construction of the baseline ranking prevents features like the 'ChIP' nodes that are *Drosophila* specific from always being selected as relevant features for the second stage simply because they are only connected to genes from the same species as the query genes. Apart from this modification, the two-stage RWR ranking algorithm and its evaluations were run on the '5 Insect' network in the same manner as the *Drosophila* network discussed above. Because of the increased size of the data, number of iterations required to converge and computational demands to perform the algorithm on the '5 Insect' network, we focused on only 12 of the 92 expression domains (Supplementary Table S13).

The average AUROC value for the stage 2 query rankings using the '5 Insect' heterogeneous network was higher (0.752) than the corresponding value on the *Drosophila* only heterogeneous network (0.728) (Fig. 4 and Supplementary Table S14).

As before, the stage 2 rankings in the '5 Insect' heterogeneous network were also better than the stage 1 rankings. The improvement upon incorporating additional species was in addition to the improvement we observed with heterogeneous over homogenous networks. The '5 Insect' network contains many additional nodes and edges that do not directly relate to the fruit fly genes being ranked. However, the advantage of the network approach is that many indirect connections contribute meaningfully to the rankings. We speculate that the '5 Insect' network provides more accurate ranking of fruit fly genes because more meaningful 'motif' or 'prot_domain' features are conserved across orthologous genes in multiple species and form dense subnetworks within the '5 Insect' heterogeneous network.

3.1.3 Query-specific feature nodes reveal shared properties of co-expressed gene sets

To create the query-specific subnetwork for the second stage RWR, our method identifies the set of feature nodes that are the most relevant to the query gene set. If there are k feature types, it selects $50k$ feature nodes to be included in the subnetwork. Of the 150 feature nodes ($k = 3$) selected from our heterogeneous *Drosophila* network as being relevant to a specific query gene set, on average, 6 were 'motif' nodes, 107 were 'prot_domain' nodes and 38 were 'ChIP' nodes. This reflected a strong enrichment for ChIP feature nodes, which only account for 2% of all feature nodes. This enrichment is not surprising given that the ChIP features were derived from experiments performed in the same developmental stages as query gene sets. This was a crude confirmation that our feature selection procedure is selecting query-relevant features. Some ChIP feature nodes were selected for many (>65) of the 92 different query gene sets. These nodes corresponded to the DNA-binding of pioneer factors TRL and VFL or important developmental regulators, such as TWI, HB and EVE. At the level of protein domains, the zinc-finger, homeobox and helix–loop–helix DNA-binding domains appeared as selected features for more than 50 of the 92 expression domains. These were the most common DNA-binding protein domains, and their appearance on the list of most relevant features is consistent with the understanding that transcription factors are a key component of gene expression control during development.

3.2 Comparison to GeneMANIA

We attempted to compare the performance of our two-stage random walk-based ranking procedure to the popular tool GeneMANIA. This tool implements label propagation on a 'gene–gene affinity network' to rank genes on their similarity to a given set. The only feature-gene data type from our analysis above that has already been preprocessed into a GeneMANIA affinity network is the 'prot_domain' feature type, representing Pfam domain annotations. In the GeneMANIA affinity network, two genes are joined if they share Pfam domains, but the exact number and types of the domains shared are lost in the representation. In our network, we explicitly connected gene nodes that share a protein domain to the feature node representing that protein domain, preserving the specific details of gene–gene relationships. Using 10% of each expression domains as test sets and the AUROC evaluation metric, we compared the GeneMANIA algorithm with its Pfam protein domain affinity network (SM4) to our two-stage RWR method applied to the homogeneous 'prot_domain' network in *Drosophila*. We found that our two-stage algorithm outperforms GeneMANIA at high values of AUROC threshold (Supplementary Fig. S3 and Supplementary Table S15). For example, at an AUROC threshold of 0.7, our RWR procedure produced rankings that yield an AUROC $>= 0.7$ for 17 expression domains, while GeneMANIA rankings reach this level of accuracy for only 8 expression domains. We also performed a comparison between DRaWR and GeneMANIA on a heterogeneous

network defined from Pfam domain edges as well as genetic and protein interaction edges (SM5 and Supplementary Table S16). We found that our random walk based approach on type-normalized heterogeneous networks produces similar average AUROC (0.7051) to the GeneMANIA label propagation method on the weighted combination of the corresponding gene–gene affinity networks (Supplementary Fig. S4). In both analyses, our algorithm was also able to report the most relevant protein domains, a capability that GeneMANIA lacks.

3.3 Application to multi-species behavioral aggression sets

Finally, we applied our DRaWR algorithm to experimentally derived gene sets that are challenging to analyze with common existing tools. In a recent study (Rittschof *et al.*, 2014), the authors attempted to understand if there are conserved neuromolecular mechanisms that underlie the common behavior of aggressive response to territorial intrusion in social animals. This study examined the transcriptomic state of brains in three greatly diverged social animals, the mouse *M.musculus*, the stickleback fish *G.aculeatus* and the honeybee *A.mellifera*. The analysis in the original study separately examined data from each species to find Gene Ontology terms and *cis*-regulatory elements significantly associated with differentially expressed (DE) genes in each species, and then honed in on associations that are shared across species. Our method, on the other hand, offers the potential for studying DE gene sets from the three species in an integrated framework that may enable more subtle signals of shared genetic 'toolkits' to reveal themselves.

3.3.1 Construction of network and definition of query sets

To construct the network for analysis of this dataset, we incorporated heterogeneous information from all three of the species in the study (mouse, stickleback fish and honeybee) as well as two additional, highly annotated species *D.melanogaster* and *H.sapiens*. We constructed a weighted network with nodes and edges described in detail in Section 2.1. We connected the gene nodes within and between species with 'homology' edges defined from all-pairs BLAST results. We connected 3671 'prot_domain' feature nodes to gene nodes in all five species based on the corresponding HMMER scans results. We also included 'Gene Ontology' feature nodes for 1827 GO terms, connecting them to nodes representing human, mouse and fruit fly genes, as per available gene annotations. We did not include any edges between 'Gene Ontology' feature nodes and genes nodes of stickleback fish or honey bee because most of their GO annotations included in Ensembl (Cunningham *et al.*, 2014) are derived from orthology rather than direct annotation. Finally, we added 'brain atlas' nodes and edges that connected these feature nodes to mouse gene nodes that are specifically expressed in one of twelve brain regions defined in the atlas. This new five species network (Supplementary Table S17) has 76 060 genes and over 13 million edges, with homology edges accounting for 95% of all edges.

We obtained one gene set of differentially expressed (DE) genes from each species from the aggression study (Rittschof *et al.*, 2014). These included 153 bee genes, 499 fish genes and 883 mouse genes deemed to be differentially expressed in the brains of the social animals when exposed to an intruder. Each of these three species-specific gene sets was to serve as a query gene set for DRaWR. Since we were interested in ranking genes and features for their relatedness to all three DE gene sets simultaneously, we additionally created a single gene set by combining all 1535 DE genes. For each of the four DE gene sets, we created an appropriate gene universe set (genes that need to be ranked by our procedure), comprising genes from only the corresponding subset of species.

Table 1. Ten query-specific features

Rank	Feature node name	Feature node type
1	Striatum	Brain Atlas
2	Retrohippocampal	Brain Atlas
3	Hippocampus	Brain Atlas
4	Pallidum	Brain Atlas
5	MRJP	Prot Domain
6	PMP22_Claudin	Prot Domain
7	JHBP	Prot Domain
8	Globin	Prot Domain
9	Olfactory	Brain Atlas
10	Claudin_2	Prot Domain

The top ten feature nodes selected with our algorithm on the '3 species' query set and multi-species, heterogeneous aggression network. Each node is listed along with its feature type.

3.3.2 Aggression related features

Application of the DRaWR pipeline to the multi-species query set comprising aggression-related DE genes from mouse, fish and bee revealed feature nodes that are most related to the query set (greatest value of $\tilde{\mathbf{v}}_Q - \tilde{\mathbf{v}}_U$) (Supplementary Table S18); we report the top ten features in Table 1.

The feature node corresponding to the 'Striatum' brain region was ranked first. This is consistent with the striatum being the part of the brain responsible for coordinating movement with motivation, an important component of an aggressive behavior response to an intruder. It has been demonstrated that damage to the striatum can result is aberrant social behavior (Glenn and Yang, 2012; Johansson and Hansen, 2001), and that the ventral striatum is active in maternal defense (Hansen *et al.*, 1991) and punishment behavior against a rival (Buades-Rotger *et al.*, 2015; Cikara *et al.*, 2011). The next most relevant feature nodes include the retrohippocampus, the hippocampus and the pallidum, which are known to be involved in emotions and movement or motivation and behavior. We also found the protein domain feature nodes for major royal jelly protein (MRJP) and juvenile hormone binding protein (JHBP) domains in our top ten list. Genes containing the MRJP domain have been previously implicated in behavior because of their expression in the mushroom bodies of honeybee brains (Drapeau *et al.*, 2006; Kucharski *et al.*, 1998). JHBP domain genes have also been correlated with hygienic behaviors in honeybees in response to infestations of parasitic mites (Parker *et al.*, 2012). There were several 'Gene Ontology' features identified by our method as relevant to our multi-species DE query set that were ranked in the top forty feature nodes. These included terms involving the plasma membrane, protein binding and ribosome. The fifth most related Gene Ontology feature node was for the term 'Hormone activity', which was also discovered in the original study (Rittschof *et al.*, 2014).

3.3.3 Observations about gene rankings in aggression study

In addition to identifying the features most relevant to the DE genes, we also evaluated the gene ranking produced after the second stage of DRaWR, conducting our tests on 10% of the genes held out from the query set. As Supplementary Table S19 (column '3 species') shows, we found the heterogeneous multi-species network (AUROC 0.690) to yield better rankings than any homogeneous, multispecies network containing a single feature type. Our method successfully

enabled us to integrate experimental results from different species with knowledge from many different sources in a single framework.

We also examined the aggression-related DE gene set of each species separately to check if these gene sets have varying levels of coherence that may make it more or less difficult to identify related genes. For each species, we tested ranking accuracy on held out DE genes, using either the 5-species network or a single-species network appropriate for that species. In general, we found that the species-specific DE gene sets that were the most difficult to correctly rank their related genes using only the appropriate single species network showed the greatest improvement when using the multi-species networks. In particular, we poorly ranked the mouse DE genes in the mouse single species heterogeneous network. However, when incorporating information from additional species, we see a great improvement (AUROC in heterogeneous, multi-species network 0.788).

4 Discussion

We have developed the DRaWR method to rank genes for their relatedness to a given gene set in the context of extensive, heterogeneous information represented as a network. We showed that the rankings improve when more sources of information are incorporated into the network and when data from additional species are appended. Our algorithm applies a two-stage RWR to rank related genes and, as a byproduct, produces a list of features that are specifically related to the gene set. We have shown its application in characterizing embryonic expression domains in *Drosophila* and transcriptomic responses to social intruders in a cross-species study.

With genome sequencing projects like the 10 000 Vertebrate Genomes (Genomes 10k) and 5000 Insect Genomes (i5k) underway and high throughput technologies becoming less expensive and more efficient, a biological network containing all public data would need to scale to thousands of species, covering tens of millions of genes and potentially billions of functional interactions. However, in runs with the 80 000 node, 13 million edge multi-species heterogeneous network above, representing the data required at least 4 GB of RAM and processing it took several hours using our R implementation. With these requirements, it becomes difficult to optimize the restart parameter or the number of selected features in the second stage subnetwork for each query set. Since all of our results suggest that we are able to produce the best rankings when given the largest, most diverse initial network, scalability of the algorithm is an important issue and one of the driving reasons for selecting a random walk with restart approach.

One common approach to address computational scalability is the paradigm of data and computation distribution offered by MapReduce (Dean and Ghemawat, 2008). The reliability and efficiency of this framework has led to its widespread adoption, and public instances (e.g. the Amazon Elastic Compute Cloud) provide a platform for users to store large networks and deploy analysis tools on them. A message passing version of the RWR algorithm maps easily to a MapReduce framework. It has been implemented in the graph mining software PEGASUS (Kang *et al.*, 2009) has been shown to scale to graphs with billions of nodes and edges. More recent software, B_LIN (Tong *et al.*, 2006), Pregel (Rozowsky *et al.*, 2009), GraphLab (Low *et al.*, 2012) and GraphX (Xin *et al.*, 2013), are explicitly designed to improve performance in scalable graph processing by carefully distributing data and minimizing communication costs.

There are several limitations to the random walk based approach. First, we are only able to represent positive information. Edges are only able to convey how closely related two nodes are and nodes are only allowed to be annotated as belonging to the given gene set. However, proper use of negative information may perhaps create a more nuanced network and produce better outcomes. For example, we may want to add edges that represent mutual exclusivity or strong anti-correlation between two nodes in the network. We may also have negative examples of a property of interest that we would like to incorporate to make rankings more accurate. Many of these properties may be addressed by remapping our random walk on a connectivity network algorithm into an application of belief propagation on probabilistic graphical models (Kang *et al.*, 2011; Koller and Friedman, 2009). Additionally, although we normalize our edges by type, the RWR does specifically treat different types of edges in a distinguishable way. Some studies have attempted to control how information is passed through different edge types by defining specific meta-paths (Yu *et al.*, 2014) that dictate a sequence of node types that must be followed to inform a relationship between two nodes. Our simple, two-stage RWR algorithm for gene ranking provides a solution to and highlights the challenges of performing analysis of experimental data on massive, heterogeneous networks of biological knowledge.

Funding

This work was supported by the National Institutes of Health Big Data to Knowledge (BD2K) initiative [1U54GM114838 to SS] and by the Cohen Graduate Fellowship awarded to CB.

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Buades-Rotger,M. *et al.* (2015) Winning is not enough: ventral striatum connectivity during physical aggression. *Brain Imaging Behav*, **10**, 105–114.

Chatr-Aryamontri,A. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.

Chen,X. *et al.* (2012) Drug-target interaction prediction by random walk on the heterogeneous network. *Mol. Biosyst.*, **8**, 1970–1978.

Cikara,M. *et al.* (2011) Us versus them: social identity shapes neural responses to intergroup competition and harm. *Psychol. Sci.*, **22**, 306–313.

Cornish,A.J. and Markowetz,F. (2014) SANTA: quantifying the functional content of molecular networks. *PLoS Comput. Biol.*, **10**, e1003808.

Cunningham,F. *et al.* (2014) Ensembl 2015. *Nucleic Acids Res*, **43**, D662–D669.

Dean,J. and Ghemawat,S. (2008) MapReduce: simplified data processing on large clusters. *Commun. ACM*, **51**, 107–113.

Drapeau,M.D. *et al.* (2006) Evolution of the Yellow/Major Royal Jelly Protein family and the emergence of social behavior in honey bees. *Genome Res.*, **16**, 1385–1394.

Finn,R.D. *et al.* (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.

Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.

Glenn,A.L. and Yang,Y. (2012) The potential role of the striatum in antisocial behavior and psychopathy. *Biol. Psychiatry*, **72**, 817–822.

Greene,C.S. *et al.* (2015) Understanding multicellular function and disease with human tissue-specific networks. *Nat. Genet.*, **47**, 569–576.

Hansen,S. *et al.* (1991) The effects of 6-OHDA-induced dopamine depletions in the ventral or dorsal striatum on maternal and sexual behavior in the female rat. *Pharmacol. Biochem. Behav.*, **39**, 71–77.

Hofree,M. *et al.* (2013) Network-based stratification of tumor mutations. *Nat. Methods*, **10**, 1108–1115.

Hou,J.P. and Ma,J. (2014) DawnRank: discovering personalized driver genes in cancer. *Genome Med.*, **6**, 56.

Ivan,G. and Grolmusz,V. (2011) When the Web meets the cell: using personalized PageRank for analyzing protein interaction networks. *Bioinformatics*, **27**, 405–407.

Jacquemin,T. and Jiang,R. (2013) Walking on a tissue-specific disease–protein-complex heterogeneous network for the discovery of disease-related protein complexes. *Biomed. Res. Int.*, **2013**, 732650.

Johansson,A.K. and Hansen,S. (2001) Increased novelty seeking and decreased harm avoidance in rats showing Type 2-like behaviour following basal forebrain neuronal loss. *Alcohol. Alcohol.*, **36**, 520–524.

Kang,U. *et al.* (2009) PEGASUS: a peta-scale graph mining system implementation and observations. In: *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*. IEEE Computer Society, pp. 229–238.

Kang,U. *et al.* (2011) Mining large graphs: Algorithms, inference, and discoveries. In: Proceedings of the 2011 IEEE 27th International Conference on Data Engineering. IEEE Computer Society, pp. 243–254.

Koller,D. and Friedman,N. (2009) *Probabilistic Graphical Models: Principles and Techniques*. Cambridge, MA: MIT Press.

Kucharski,R. *et al.* (1998) A royal jelly protein is expressed in a subset of Kenyon cells in the mushroom bodies of the honey bee brain. *Naturwissenschaften*, **85**, 343–346.

Lein,E.S. *et al.* (2007) Genome-wide atlas of gene expression in the adult mouse brain. *Nature*, **445**, 168–176.

Leiserson,M.D. *et al.* (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.*, **47**, 106–114.

Li,Y. and Patra,J.C. (2010) Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network. *Bioinformatics*, **26**, 1219–1224.

Liu,Y. *et al.* (2014) A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression. *BMC Bioinformatics*, **15**, 37.

Low,Y. *et al.* (2012) Distributed GraphLab: a framework for machine learning and data mining in the cloud. *Proc. VLDB Endow*, **5**, 716–727.

Mostafavi,S. and Morris,Q. (2012) Combining many interaction networks to predict gene function and analyze gene lists. *Proteomics*, **12**, 1687–1696.

Orchard,S. *et al.* (2014) The MIntAct project–IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.

Parker,R. *et al.* (2012) Correlation of proteome-wide changes with social immunity behaviors provides insight into resistance to the parasitic mite, Varroa destructor, in the honey bee (*Apis mellifera*). *Genome Biol.*, **13**, R81.

Pena-Castillo,L. *et al.* (2008) A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. *Genome Biol.*, **9**, S2.

Reimand,J. *et al.* (2008) GraphWeb: mining heterogeneous biological networks for gene modules with functional significance. *Nucleic Acids Res.*, **36**, W452–W459.

Rittschof,C.C. *et al.* (2014) Neuromolecular responses to social challenge: common mechanisms across mouse, stickleback fish, and honey bee. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 17929–17934.

Rozowsky,J. *et al.* (2009) PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls. *Nat. Biotechnol.*, **27**, 66–75.

Salwinski,L. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–D451.

Shen,R. *et al.* (2012) Mining functional subgraphs from cancer protein-protein interaction networks. *BMC Syst. Biol.*, **6**, S2.

Sinha,S. *et al.* (2006) Stubb: a program for discovery and analysis of cis-regulatory modules. *Nucleic Acids Res.*, **34**, W555–W559.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 15545–15550.

Tarca,A.L. *et al.* (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.

Tomancak,P. *et al.* (2002) Systematic determination of patterns of gene expression during Drosophila embryogenesis. *Genome Biol.*, **3**, RESEARCH0088.

Tong,H. *et al.* (2006) Fast random walk with restart and its applications. In: *Proceedings of the Sixth International Conference on Data Mining*. IEEE Computer Society, pp. 613–622.

Vaske,C.J. *et al.* (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–i245.

Wang,S. *et al.* (2015) Exploiting ontology graph for predicting sparsely annotated gene function. *Bioinformatics*, **31**, i357–i364.

Warde-Farley,D. *et al.* (2010) The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. *Nucleic Acids Res.*, **38**, W214–W220.

Xin,R.S. *et al.* (2013) GraphX: a resilient distributed graph system on Spark. In, *First International Workshop on Graph Data Management Experiences and Systems*. New York, New York: ACM, pp. 1–6.

Yu,X. *et al.* (2014) Personalized entity recommendation: a heterogeneous information network approach. In, *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*. New York, New York, USA: ACM, pp. 283–292.