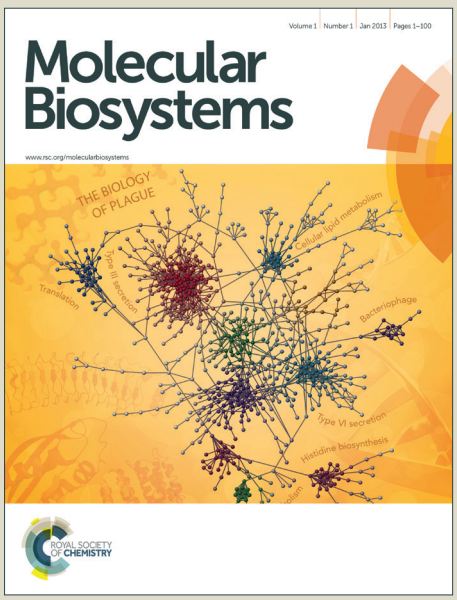
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems

Prioritization of candidate disease gene by enlarging the seed set and fusing the information of network topology and gene expression

Shao-Wu Zhang*, Dong-Dong Shao, Song-Yao Zhang, Yi-Bin Wang

College of Automation, Northwestern Polytechnical University, 710072, Xi'an, China

*Corresponding author: zhangsw@nwpu.edu.cn

Abstract: Identification of disease genes is very important not only for better understanding of gene function and cellular mechanisms driven human disease, but also for enhancing the level of human disease diagnosis and treatment. High-throughput techniques are recently applied frequently to detect dozens or even hundreds of candidate genes. However, the experimental approaches of validating these many candidates are usually time-consuming, tedious and expensive, and sometimes lack reproducibility. Therefore, numerous theoretical and computational methods (e.g. network-based approaches) have been developed to prioritize the candidate disease genes. Many network-based approaches implicitly utilize the observation that genes causing the same or similar diseases tend to correlate with each other in gene/protein relationship networks. Of these network approaches, the random walk with restart algorithm (RWR) is considered to be a state-of-the-art approach. To further improve the performance of RWR, we proposed a novel method named as ESFSC to identify the disease-related genes, by enlarging the seed set according to the centrality of disease genes in network and fusing the information of protein-protein interaction (PPI) network topological similarity and gene expression correlation. ESFSC algorithm restarts at all the nodes in seed set consisted of the known disease genes and their *k*-nearest neighbor nodes, then walks in the global network separately guided by the similarity transition matrix constructed with PPI network topological similarity properties and the correlational transition matrix constructed with the gene expression profiles. In the end, all the genes in the network are eventually ranked by weighted fusing above results of RWR guided by two kinds of transition matrices.

Comprehensive simulation results on the 10 diseases with 97 known disease genes collected from the Online Mendelian Inheritance in Man (OMIM) database show that ESFSC outperforms existing methods in prioritizing candidate disease genes. The top prediction results of Alzheimer's disease are in good consistent with the literature reports.

Keywords: candidate disease gene, *k*-nearest neighbor gene, protein-protein interaction network, fusion, random walk

I. Introduction

Discovery of disease-associated genes is an important step toward enhancing our understanding of the cellular mechanisms that drive human disease, with profound applications in modeling, diagnosis, prognosis, and therapeutic intervention [1]. Genetic approaches, such as linkage analysis (connecting loci with a tendency to be inherited together) and association studies (mapping correlation between alleles at different loci), have uncovered plenty of links between diseases and particular chromosomal regions potentially containing hundreds of candidate genes possibly associate with genetic disease [2]. Investigation of these candidates and other biological problems using experimental methods are usually time-consuming, tedious and expensive, and sometimes lacks reproducibility. Therefore, many studies from various research laboratories around the world have indicated that mathematical analysis, computational modeling, and introducing novel physical concept to solve important problems in biology and medicine, such as random walk models [3-4], protein-protein interaction network [5-9], protein structural class prediction [10,11], modeling 3D structures of targeted proteins for drug design [12-15], diffusion-controlled reaction simulation [16-19], cellular responding kinetics [20], bio-macromolecular internal collective motion simulation [21-23], identification of proteases and their types [24], membrane protein type prediction [25], protein cleavage site prediction [26,27], and signal peptide prediction [28], can timely provide

very useful information and insights for both basic research and hence are widely welcome by science community. The present study is related to the fundamental problems in system biology, network biology, and structural biology of proteins. The relationship between these systems will be of use for the global research. Recently, a number of computational approaches have also been proposed to prioritize candidate disease genes [29-54, 57-75]. According to the biological data and their representation that are primarily considered when scoring and ranking candidate disease genes, the prioritization approaches are categorized as: (i) Gene and protein characteristics [29-40]. These approaches are largely based on the similarity of characteristics of disease genes including sequence-based feature [32-34], expression patterns [35-37] and functional annotation data [38,39]. Although these approaches have better performance, they suffer from some inherent limitations, e.g. the incomplete and false-positive disease-causal genes data, ambiguous boundary between different disease, and highly heterogeneous of diseases [40]. (ii) Network information on molecular interactions [41-54, 68,69]. These approaches are largely based on the principle of 'guilt-by-association', in that, genes that are physically or functionally close to each other tend to be involved in the same biological pathways and have similarity effects on phenotypes [55,56]. (iii) Integrated biomedical knowledge [57-67]. These approaches make successful use of relatively simple integration procedures for a few different sources of functional information and annotations, which can achieve good performance and reduce the effect of noisy and incomplete datasets.

Many network-based approaches implicitly utilize the observation that genes causing the same or similar diseases tend to correlate with each other in gene/protein relationship networks. Existing network-based approaches can also be classified into two main families: (i) local approaches, which focus on the local network information such as direction interaction and shortest paths between disease genes and candidate genes [41-48, 72-75]; (ii) global approaches, which model the information flow in the cell to access the proximity and connectivity between known disease genes and candidate disease genes [40, 49, 50, 57, 68-70]. Several studies show that global

approaches, such as random walk and network propagation, are clearly superior to the local approaches[40, 49, 50, 57]. Random walk with restart (RWR) method simulates a random walk on the network to compute the proximity between two nodes by exploiting the global structure of the network. Although RWR and its improved methods have recently been applied to candidate disease gene prioritization and acquired better prioritizing results, most of the existing methods underutilize the centrality of disease genes in network, and the transition matrices used in these existing methods are normalized adjacency matrices or normalized intensity matrices of protein interaction, which cannot effectively represent the network status. In addition, most existing RWR methods just used single information source of protein-protein interaction (PPI) database. To further improve the performance of candidate disease gene prioritization, we proposed a novel method, called ESFSC, to identify the disease-related genes by enlarging the seed set based on the centrality of disease genes in network, and fusing the information of PPI network topological similarity and gene expression correlation. ESFSC algorithm restarts at all the nodes in seed set that consist of the known disease genes and their k-nearest neighbor nodes, then walks in the global network with RWR guided by one of the following two transition matrices: similarity transition matrix constructed with PPI network topological properties, correlational transition matrix constructed with the gene expression profiles. In the end, all the genes in the network are eventually ranked by weighted fusing the results of above two kinds of RWR. Through extensive simulations on the 10 diseases (*Adrenoleukodystrophy*, *Alzheimer Disease*, *Arrhythmogenic right ventricular dysplasia*, *Bladder Cancer*, *Breast cancer*, *Cornelia de Lange syndrome*, *Dilated cardiomyopathy*, *Ectodermal dysplasia*, *Hypercholesterolemia*, *Lung cancer*), our approach showed better performance than traditional RWR-based approaches. Moreover, it outperformed two representative network-based approaches, PRINCE [50] and ORIENT [70]. We also investigated the factors affecting the performance, and analyzed the top prediction results of Alzheimer's disease in detail which are in good consistent with the literature reports.

II. Materials and Methods

Protein-protein interaction data

The protein-protein interaction (PPI) network is modeled as undirected graph with nodes representing the proteins and edges representing the physical or binding interactions between proteins encoded by the genes. We downloaded five PPI datasets of *human*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae* from Biological General Repository for Interaction Datasets (BioGRID, Version 3.1.93 Release, thebiogrid.org/download.php) [76] to construct the PPI network. In order to enlarge the edges of human PPI network, the PPI interactions from four nonhuman species were mapped to homologous human genes identified by Inparanoid [77] with a threshold Inparalog score of 0.85. If both proteins in the interaction partner could be simultaneously mapped to human proteins, this interaction was used. After removing duplications and self-linked interactions, we obtained 78,525 interactions between 12,491 human genes.

Gene expression data and known disease-gene association data

The GSE34308, GSE4757, GSE29819, GSE31189, GDS3952, GSE12408, GSE29819, GSE16524, GDS3668 and GSE23066 datasets of *Adrenoleukodystrophy*, *Alzheimer Disease*, *Arrhythmogenic right ventricular dysplasia, Bladder Cancer*, *Breast cancer*, *Cornelia de Lange syndrome*, *Dilated cardiomyopathy*, *Ectodermal dysplasia*, *Hypercholesterolemia* and *Lung cancer* disease gene expression profiles were downloaded from NCBI Entrez Gene GEO (http://www.ncbi.nlm.nih.gov/geo/), which were derived from 10, 20, 12, 92, 82, 34, 13, 8, 19 and 10 disease and control samples respectively. The 97 known causative genes associated with these 10 diseases were collected from the Online Mendelian Inheritance in Man (OMIM) knowledgebase [78] (http://www.ncbi.nlm.nih.gov/omim). The OMIM ID and name of 97 known causative genes associated with 10 diseases are given in the Table S1 in the supplementary information.

ESFSC algorithm

RWR is a ranking algorithm which simulates a random walker, either starts on a seed node or on a set of seed nodes and moves to its immediate neighbors randomly at each step [69]. In the end, all the nodes in the graph are ranked by the probability of the random walker reaching this node. Given a connected weighted graph $G(V,E)$ with a set of nodes $V=\{v_1,v_2,\cdots,v_N\}$, and a set of links $E=\{(v_i,v_j)\,|\,v_i,v_j\in V\}$, RWR can be formally described as follows:

$$P_{t+1}=(1-\gamma)MP_t+\gamma P_0 \qquad (1)$$

where P_t is an $N\times1$ vector in which the ith element represents the probability of the walker being at node v_i at time step t, and P_0 is the $N\times1$ initial probability vector. M is an $N\times$N transition matrix of the graph, and γ is a fixed parameter which denotes the restarting probability at a given time step.

Although RWR and its variants have also been applied to candidate disease gene prioritization [40, 49, 50, 57, 69-71], most these methods start on a known causative disease gene [40] or on the set of known causative disease genes [49, 69, 70], and use PPI network or disease phenotype network to construct the transition matrix M [40, 57, 70]. By considering the centrality of disease genes in network and introducing the information of the gene expression profiles, we propose a novel method, namely ESFSC. ESFSC simulates a random walker, starts on a set of seed nodes that consist of the known disease genes and their k-nearest neighbor nodes, and moves to their immediate neighbors randomly at each step with RWR guided by one of the following two transition matrices: similarity transition matrix constructed with PPI network topological properties, correlational transition matrix constructed with the gene expression profiles. After the steady-state probability vectors of random walk-guiding with the similarity transition matrix and the correlational transition matrix respectively were obtained, all the genes in the network are eventually ranked by weighted fusing the results of above two kinds of RWR. We now describe the ESFSC algorithm in detail.

(i) All the known disease genes and their k-nearest neighbor genes in the

network were defined as the seed genes, which form a seed set *S*. Then, we constructed the initial probability vector P_0, in which the value of *i*th element was assigned as $1/|S|$ if $v_i \in S$ and 0 otherwise. |S| is the total number of seed gene in the set S.

(ii) According to the Leicht's similarity measure between two vertices [79], the topology similarity matrix *T* in the PPI network is defined as follows:

$$T = [I - \phi A]^{-1} \quad (2)$$

where *I* is the identity matrix, *A* is the adjacency matrix of the PPI network and ϕ is a free parameter whose value controls the balance between the neighbor term similarity and self-similarity of a pair proteins in PPI network. In general, selecting $\phi = 0.95/\lambda$, and λ is the largest eigenvalue of adjacency matrix *A*. With this topological similarity matrix, the transition matrix M^S was defined as:

$$M^{\mathrm{S}} = A \circ T = A \circ \begin{bmatrix} t_{11} & \cdots & t_{1N} \\ \vdots & \vdots & \vdots \\ t_{N1} & \cdots & t_{NN} \end{bmatrix} \quad (3)$$

where $A \circ T$ means the Hadamard product (or element-by-element product) of matrices *A* and *T*, and M^S is a column-normalized matrix.

(iii) According to the differential expression profiles of genes under the disease and health condition, we use the Pearson's correlation coefficient to measure the correlation between two genes and then construct the transition matrix M^C for a disease.

For a certain disease, let $Z_i^d = [z_1^d, z_2^d, \cdots, z_\tau^d, \cdots, z_L^d]$ and $Z_i^h = [z_1^h, z_2^h, \cdots, z_\tau^h, \cdots, z_L^h]$ denote the gene expression profiles under the disease and health condition, i and τ are the serial number of gene and sample respectively, then the differential gene expression profile can be defined as follows:

$$\Delta Z_i = [(z_1^d - z_1^h), (z_2^d - z_2^h), \cdots, (z_\tau^d - z_\tau^h), \cdots, (z_L^d - z_L^h)], \quad i = 1, 2, \cdots, N, \quad \tau = 1, 2, \cdots, L \quad (4)$$

The Pearson's correlation coefficient between any two differential gene expression

profiles ΔZ_i and ΔZ_j is defined as follows:

$$r_{ij} = \frac{Cov(\Delta Z_i, \Delta Z_j)}{\sigma_{\Delta Z_i}\sigma_{\Delta Z_j}}, \quad i, j = 1, 2, \cdots, N \tag{5}$$

where Cov is the covariance, $\sigma_{\Delta Z_i}$ is the standard deviation of ΔZ_i. With these correlation coefficients, the transition matrix M^C was defined as:

$$M^C = A \circ R = A \circ \begin{bmatrix} |r_{11}| & \cdots & |r_{1N}| \\ \vdots & \vdots & \vdots \\ |r_{N1}| & \cdots & |r_{NN}| \end{bmatrix} \tag{6}$$

where $A \circ R$ means the Hadamard product of matrices A and R, and M^C is a column-normalized matrix.

(iv) After the steady-state probability vectors P_∞^S and P_∞^C were obtained by repeating the iterations until $\|P_{t+1} - P_t\| < 10^{-10}$, all genes in the network are eventually ranked by the following weighted fusion results:

$$P_\infty = \alpha P_\infty^S + (1-\alpha) P_\infty^C \tag{7}$$

where α is a weighted parameter, P_∞^S and P_∞^C are the results of RWR-guiding with transition matrix M^S and P_∞^C respectively.

Figure 1 is a flow chart which shows the prioritization process of candidate disease gene of ESFSC algorithm.

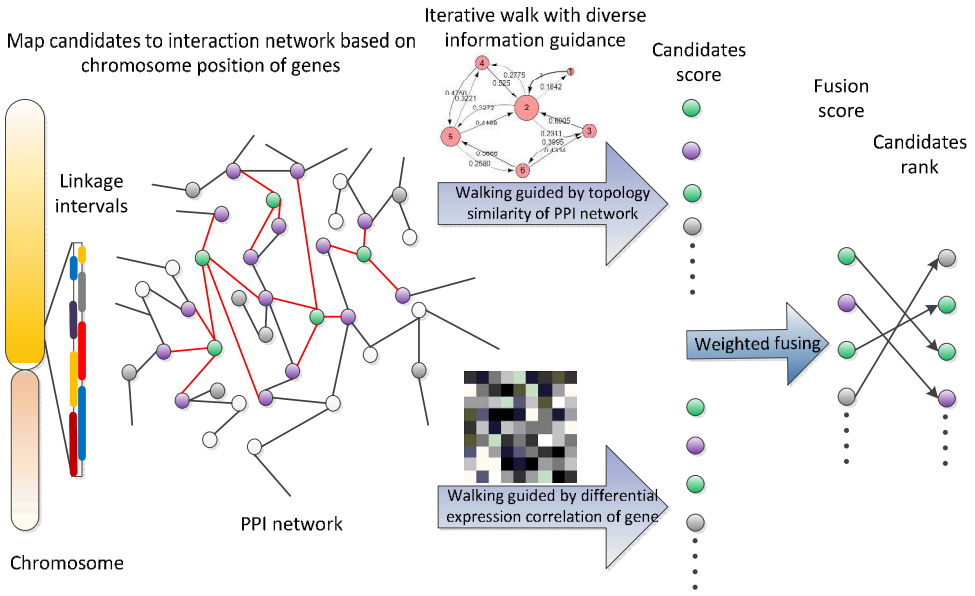


Figure 1 A flowchart to show the prioritization process of candidate disease gene of ESFSC algorithm. The candidate genes and the known disease genes relative to certain disease are mapped into the PPI network. The nodes in green, purple, gray and white represent the disease genes, the nearest neighbor of disease genes, candidate genes and other genes respectively. The bars represent the chromosome and linkage intervals respectively.

III. Results and discussion

Comparison with other network-based methods

To examine the performance of our ESFSC algorithm, we compared ESFSC with other three network-based methods, i.e. the RWR [69], PRINCE [50] and ORIENT [70] on the same PPI network, which achieved relatively better performance than that of linkage-based methods and graph partitioning-based methods [45]. The only difference between RWR and PRINCE is the construction of initial distribution of a disease, where the initial probability vector of RWR was constructed such that equal probability was assigned to each causing gene of a disease, and the prior vector in PRINCE was initialized by incorporating disease similarity information by using a logistic function. The main difference between RWR and ORIENT is the construction of transition matrix, where the transition matrix of RWR was a column-normalized adjacency matrix of PPI network, and the transition matrix of ORIENT was a

neighbor-favoring weighted adjacency matrix by considering the fact that the disease genes tend to be modularized in the network. In our ESFSC algorithm, we used two kinds of matrices. One is the similarity transition matrix M^S constructed with network topological properties and adjacency matrix in PPI network. Another is the correlation transition matix M^C constructed with differential gene expression profiles and Pearson's correlation coefficient. Based on the transition matrix M^S and M^C, we used RWR algorithm to obtain the rank results respectively. In the end, we employed the weighted fusion rule to form the final prioritization results.

Among the independent dataset test, sub-sampling (e.g., 5 or 10-fold cross-validation) test, and leave-one-out cross-validation (also called jackknife) test, which are often used for examining the accuracy of a statistical prediction method, the jackknife test was least arbitrary and can always yield a unique outcome [80] and has been widely and increasingly adopted by investigators to test the power of various prediction methods [81-89]. In this work, we used the leave-one-out cross validation to evaluate the performance of prioritization candidate gene for different network-based methods. An artificial linkage interval including one known disease gene s_i and 99 genes closest to s_i in terms of genomic distance were considered as candidate set X. In each validation stage, one known disease gene was removed from the know disease gene set $\Sigma=\{s_1,s_2,\ldots,s_i,\ldots,s_I\}$ and the rest know disease genes were used as training set to prioritize all the genes in the candidate set X. For a reliable performance comparison, we used the receiver operating characteristic (ROC) curve to show the relationship between *sensitivity* and 1-*specificity* at different threshold θ with the rank ratio (varying from 0 to 1 with scale 0.01), and computed the area under the curve (AUC) based on the rank of the removed gene s_i in set X. More specifically, given a threshold θ, the *sensitivity* (recall) is defined as the percentage of known disease genes that are ranked above threshold θ, whereas *specificity* is defined as the percentage of all non-known disease genes in set X that are ranked below threshold θ. Due to existing only one known disease gene (also called true positive

sample) in each cross-validation, AUC is a conservative measure. So, we also used other three measures R_v (average rank), R_p^1 (percentage of the disease genes ranked in top 1%) and R_p^5 (percentage of the disease genes ranked in top 1%) to evaluate the performance of the prioritization methods. R_v is defined as the average rank of one known disease gene among all the candidate genes. R_p^1 and R_p^5 are defined as the percentage of the known disease genes that are ranked as one of the genes in the top 1% and top 5% among all the candidates respectively. Clearly, larger AUC/ R_p^1 / R_p^5 and lower R_v values indicate a better prediction performance for a prioritization method.

With the 97 known causal genes related to 10 disease of *Adrenoleukodystrophy*, *Alzheimer Disease*, *Arrhythmogenic right ventricular dysplasia*, *Bladder Cancer*, *Breast cancer*, *Cornelia de Lange syndrome*, *Dilated cardiomyopathy*, *Ectodermal dysplasia*, *Hypercholesterolemia* and *Lung cancer*, the results of RWR [69], PRINCE [50], ORIENT [70] and ESFSC (k=1) are shown in figure 2 and table 1. As seen in the figure 2， the curve of ESFSC is above those of RWR, PRINCE and ORIENT, which suggest that our algorithm achieved both higher sensitivity and higher specificity. From table 1, we can see that the R_p^1, R_p^5 values of ESFSC are 35.05, 69.07 respectively, which are much higher than that of RWR (34.02, 62.86), ORIENT (32.99, 63.92) and PRINCE(28.87, 58.76) methods, and the average rank value of ESFSC is 5.74, which is 3.85, 3.18 and 3.58 lower than that of RWR, ORIENT and PRINCE methods. These results show that our ESFSC algorithm is superior to RWR, ORIENT and PRINCE methods, and have the best better performance for prioritizing candidate disease genes.

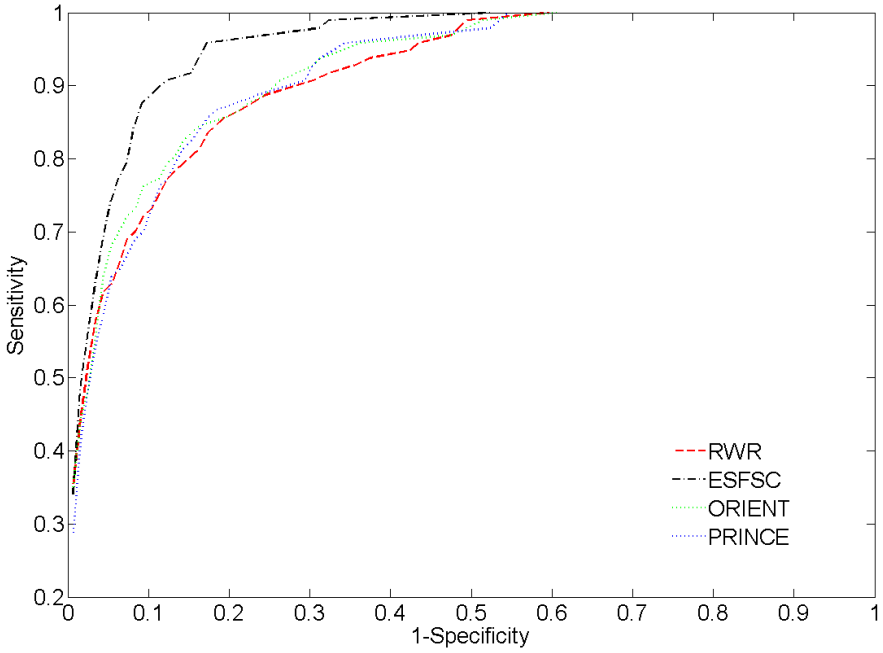


Figure 2 Comparison of the proposed ESFSC method with other existing state-of-the-art RWR, ORIENT and PRINCE methods.

Table 1 Performance comparison of RWR, ORIENT, PRINCE and ESFSC with the 97 known causal genes related to 10 diseases.

Method	AUC	R_p^1	R_p^5	R_v
RWR	0.91	34.02	61.86	9.59
ORIENT	0.92	32.99	63.92	8.92
PRINCE	0.92	28.87	58.76	9.32
ESFSC	0.96	35.05	69.07	5.74

Analysis of the globally top ranked genes

If we defined all known genes of one disease as training genes (i.e. source nodes), and the rest genes in the PPI network as testing genes (i.e. candidate genes), and used ESFSC algorithm to rank all the genes in the network. This stage was repeated for the 10 diseases. Comparing the rank results of these 10 diseases, we found that the some

genes ranked top overlap in most diseases. For example, the gene UBC, ELAVL1, SUMO2 and CUL3 appear in the top 10 genes of 10, 7, 6 and 5 diseases respectively. According to the annotation in the GeneCards web (http://www.genecards.org/), these genes involve in some important biological process related to some diseases.

UBC (ubiquitin C) represents an ubiquitin gene, which associates with protein degradation, DNA repair, cell cycle regulation, kinase modification, endocytosis, and regulation of other cell signaling pathways. Diseases associated with UBC include spinocerebellar ataxia type 3, and chromosome 5q deletion. The super-pathways of regulation of APC/C activators between G1/S and early anaphase and Fanconi anemia are related to UBC.

ELAVL1 (ELAV like RNA binding protein 1) links to a number of diseases, including paraneoplastic neurologic disorders, and hereditary breast cancer. The super-pathways of metabolism of RNA and destabilization of mRNA by AUF1 are related to ELAVL1. It is highly expressed in many cancers, and could be potentially useful in cancer diagnosis, prognosis, and therapy.

SUMO2 (small ubiquitin-like modifier 2) is a protein-coding gene. Diseases associated with SUMO2 include transient cerebral ischemia, and myocarditis. The super-pathways of proteolytic processing of SUMO and Wnt signaling pathway are related to SUMO2.

CUL3 (cullin 3) encodes a member of the cullin protein family. Diseases associated with CUL3 include pseudohypoaldosteronism type II, and glomuvenous malformation. The super-pathways of antigen processing, ubiquitination , proteasome degradation and immune system are related to CUL3.

To further explore the implications of the top ranked genes to disease, we conducted gene ontology (GO) and pathway enrichment analysis for the 1,223 genes which ranked top 200 in every disease. The Fisher's Exact test [90] was used to measure whether the top ranked gene group is more enriched with genes of a specific GO term or gene involved in a particular pathway than what would be expected by chance. Generally, the P-value smaller than 0.05 shows the low probability that the genes of same GO term or pathway appear in the group by chance, that is, this top

ranked group is significantly enriched in the annotation categories. We found that the gene groups significantly enriched in biological process (BP) are GO: 0010941 (covering 224 mapped genes), GO: 0043067 (covering 223 mapped genes) and GO: 0042981（covering 221 mapped genes）, whose terms are regulation of cell death, regulation of programmed cell death and regulation of apoptotic process respectively. These biological processes are highly associated with the process of disease. When mapping the 1,223 genes onto the KEGG pathway, we found that most of significantly enriched pathways are related with diseases. For example, map05200 (covering 148 mapped genes), map05220 (covering 53 mapped genes) and map04110 (covering 69 mapped genes) relate to cancer, chronic myeloid leukemia and cell cycle. The results of enrichment analysis about molecular function (MF) and cell component (CC) were listed in supplementary Table S2.

Effect of the parameters

There are three parameters γ, k and α in our ESFSC algorithm. The parameter γ is the restart probability, which adjust the preference between the importance of a protein with respect to the seed set and network topology. By selecting different γ values (varying from 0.05 to 0.95 with scale 0.05) to simulate, the AUC results are shown in figure 3 in which we found that the AUC value increased gradually in the range $0.05 \le \gamma < 0.5$, and decreased gradually in the range $0.5 < \gamma \le 0.95$. But the effect of this parameter is minor in the range of $\gamma \ge 0.3$. In this work, we fix $\gamma = 0.5$.

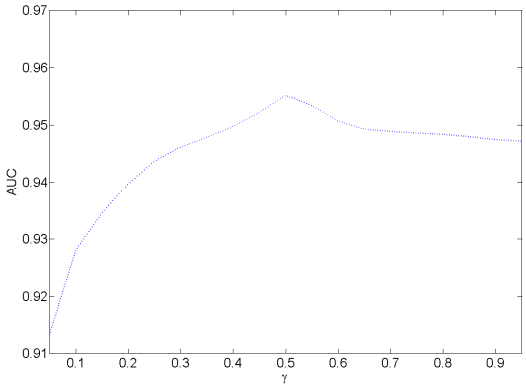


Figure 3 The relationship between the parameter γ and the AUC value for ESFSC algorithm.

The parameter k controls the size of seed set. Large k value lets the seed set including more neighbor genes of one known disease gene. To investigate the effect of this parameter, we set various k values varying from 0 to 10 with scale 1. The AUC results are shown in figure 4. From figure 4, we can see that AUC value increase gradually for $k < 1$, and decrease quickly in the range of $1 \le k \le 3$ and is stable for $k > 3$. This means that the nearest neighbor genes of known disease genes can effectively help to prioritize the candidate disease genes. If further increasing the k value, more noise will be introduced in the initial probability vector. In this work, we fix $k = 1$.

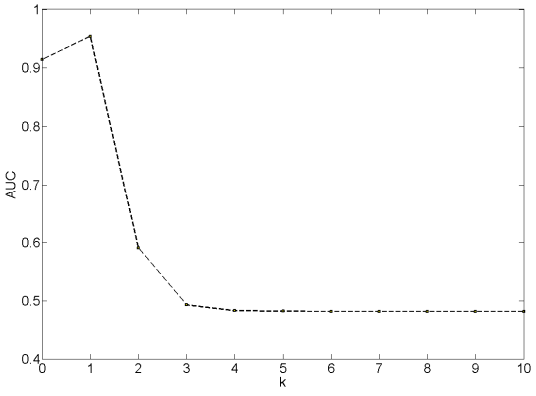


Figure 4 The relationship between the parameter k and the AUC value for ESFSC algorithm.

The parameter α controls the contribution of two kinds of random walk results, that is, random walk based on the similarity transition matrix (M^S) and random walk

based on the correlation transition matrix (M^C). Large α will introduce more contribution of random walk based on M^S . To investigate the effect of this parameter, we set various α values ranging from 0 to 1 with scale 0.01. The performance of ESFSC algorithm measured by AUC is shown in figure 5. From figure 5, we can see that the parameter α has bigger effect on the results. The AUC value increases gradually in the range of $0 \le \alpha \le 0.71$, and decreases quickly at $\alpha > 0.71$ .When $\alpha = 0.71$, ESFSC algorithm gets the best results. Therefore, we fix $\alpha = 0.71$ in this work.

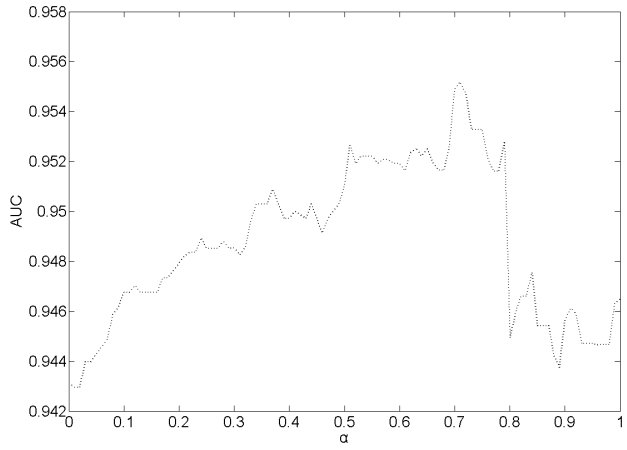


Figure 5 The relationship between the parameter α and the AUC value for ESFSC algorithm.

Effect of different transition matrices

To examine the effect of different transition matrices in the process of random walk, we constructed three transition matrices: adjacent matrix (A) based on the PPI network, similarity matrix (M^S) based on the PPI network topological similarity, and correlation matrix (M^C) based on the differential gene expression profiles. Using the known disease genes and their nearest neighbor genes as the seed set, the performance of ESFSC algorithm with different transition matrices (A, M^S, M^C) measured by four evaluation criteria are shown in table 2. We found that the performance of these three transition matrices (A, M^S, M^C) is almost same, especially the results of adjacent matrix is slight better than that of similarity matrix M^S and correlation matrix M^C . In other hand, above results indicate that our strategies of enlarging the seed set and

weighted fusion play an important role in enhancing the performance of prioritizing candidate disease genes.

Table 2 The performance of RWR with three transition matrices

Matrix	AUC	R_p^1	R_p^5	R_v
A	0.95	34.02	68.04	5.84
M^{S}	0.95	31.96	68.04	6.25
M^{C}	0.94	32.99	67.01	6.62
$M^{\mathrm{S}}+M^{\mathrm{C}}$	0.96	35.05	69.07	5.74

Case Example

Here, we provide a real example to demonstrate the power of our proposed ESFSC algorithm in identifying candidate disease genes. We focus on Alzheimer's disease (AD) since it is the most common form of dementia among older people. Dementia is a brain disorder that seriously affects a person's ability to carry out daily activities. The genes on the 12 disease-associated chromosomal regions from AD5 to AD16 were selected as research object. There are 14 known disease genes associated with AD on these 12 chromosomal regions, and the rest genes were used to prioritize with ESFSC algorithm. Top 5 ranked candidate genes had been selected for each chromosomal region, which are shown in Supplementary Table S3. The function and the relationship with AD of these top 5 ranked genes (CLU, EIF4EBP1, LPL, PTK2B and ERLIN2) on AD12 are explained as follows:

CLU (Clusterin) is significantly associated with human AD [91, 92]. Clusterin mRNA is distributed heterogeneously in the central nervous system with highest levels in ependymal cells, as well as in some neurons of the hypothalamus, brainstem, hebenula, and ventral horn of the spinal cord. It may be a suicide gene active in programmed cell death [93].

EIF4EBP1 (Eukaryotic translation initiation factor 4E-binding protein 1) gene influences both cell growth and proliferation, and controls the translation of protein

tau in homogenates of the medial temporal cortex of AD brain. Their aberrant changes may up-regulate the translation of tau mRNA, contributing to hyperphosphorylated tau accumulation in NFT-bearing neurons [94].

LPL (Lipoprotein lipase) helps transfer lipids from lipoprotein particles to cells. In the brain, LPL is present in Alzheimer's disease (AD) amyloid plaques. LPL mutations are associated with altered AD risk, and LPL is a potential role in the causation of AD [95].

PTK2B (Protein tyrosine kinase 2 beta) encodes a cytoplasmic protein tyrosine kinase which is involved in calcium-induced regulation of ion channels and activation of the map kinase signaling pathway. The encoded protein may represent an important signaling intermediate between neuropeptide-activated receptors or neurotransmitters that increase calcium flux and the downstream signals that regulate neuronal activity. PTK2B/PYK2 may also provide a mechanism for a variety of short and long-term calcium-dependent signaling events in the nervous system [96]. Aberrant PTK2B/PYK2 expression may play a role in cancer cell proliferation, migration and invasion, in tumor formation and metastasis. Elevated PTK2B/PYK2 expression is seen in gliomas, hepatocellular carcinoma, lung cancer and breast cancer.

ERLIN2 (ER lipid raft associated 2) encodes a member of the SPFH domain-containing family of lipid raft-associated proteins, which is associated with active γ-secretase in brain and affects amyloid β-peptide (Aβ) production. Aβ has a causative role in Alzheimer's disease [97].

Conclusions

In this work, we proposed a novel method named as ESFSC to identify the disease-related genes, by enlarging the seed set and fusing the information of the PPI network topological similarity and differential gene expression correlation. The novelty of our method lies in the three following aspects. Firstly, according to the centrality of disease genes, the known disease genes and their *k*-nearest neighbor genes in PPI network are used to construct the seed set for generating the initial

probability vector. Second, the PPI network topological similarity and differential gene expression correlation are introduced to construct the two kinds of transition matrices for respectively guiding the walker randomly walk in the global network. Third, the results of RWR guided by similarity matrix and correlation matrix were fused by weighted rule for ranking the candidate genes in the network. Leave-one-out cross-validation on the 10 diseases with 97 known disease genes show that our proposed ESFSC algorithm achieved higher precision (measured by AUC, R_p^1 and R_p^5) and lower average rank than the existing state-of-art network-based approaches. We also predicted the causing genes of Alzheimer's disease with ESFSC algorithm, and found that most of the top 5 ranked genes in our predicted results are in good accordance with current experimental reports.

In the future work, we will integrate some other genomic information such as functional annotations, pathway membership to further improve our method. We can also combine viral PPI network and human PPI network to predict novel candidate genes associated with disease.

Acknowledgment

This paper was supported by the National Natural Science Foundation of China (No. 61170134 and 60775012).

References

1 H. G. Brunner and M. A. van Driel, *Nature Reviews Genetics,* 2004, 5*,* 545-551.

2 D. Altshuler, M. J. Daly and E. S. Lander, *Science*, 2008, 322, 881-888.

3 G. P. Zhou, *Virulence*, 2013, 4, 669-670.

4 S. Yan and G. Wu, *Virulence*, 2013, 4, 716-725.

5 G. P. Zhou and R. B. Huang, *Current Topics in Medicinal Chemistry*, 2013, 13, 1152-1163.

6 G. P. Zhou, *Journal of Theoretical Biology*, 2011, 284, 142-148.

7 G. P. Zhou, *Protein & Peptide Letters,* 2011,18, 966–978.

8 G. P. Zhou, *Protein & Peptide Letters,* 2011,18, 964-965.

9 T. C. Bjorndahl, G. P. Zhou, X. H. Liu, R. P. Pineiro, V. Semenchenko, F. Saleem, S. Acharya, A. Bujold, C. A. Sobsey and D.S. Wishart, *Biochemistry,* 2011, 50, 1162–1173.

10 G. P. Zhou and N. Assa-Munt, *Proteins: Structure, Function, and Genetics,* 2001, 44, 57-59.

11 K. C. Chou and C. T. Zhang, *Critical Reviews in Biochemistry and Molecular Biology,* 1995, 30, 275-349.

12 G. P. Zhou and F. A. Troy, *Current Protein and Peptide Science,* 2005, 6, 399-411.

13 K. C. Chou, *Current Medicinal Chemistry,* 2004, 11, 2105-2134.

14 A. K. Sharma, G. P. Zhou, J. Kupferman, H. K. Surks, E. N. Christensen, J. J. Chou, M. E. Mendelsohn and A. C. Rigby, J*ournal of Biological Chemistry*, 2008, 283, 32860-9.

15 G. P. Zhou, H. K. Surks, J. R. Schnell, J. J. Chou, E. Michael, M. E. Mendelsohn and A. C. Rigby, *Blood,* 2004, 104, 963a.

16 G. Q. Zhou and W. Z. Zhong, *European Journal of Biochemistry,* 1982, 128, 383-387.

17 K. C. Chou and G. P. Zhou, *Journal of American Chemical Society,* 1982, 104, 1409-1413.

18 G. P. Zhou, T. T. Li and K. C. Chou, *Biophysical Chemistry*, 1981, 14, 277-281.

19 G. Z. Zhou, M. T. Wong and G. Q. Zhou, *Biophysical Chemistry,* 1983, 18, 125-132.

20 J. P. Qi, S. H. Shao, D. D. Li and G. P. Zhou, *Amino Acids,* 2007, 33, 75–83.

21 G. P. Zhou, *Physica Scripta,* 1989, 40, 698-701.

22 K. C. Chou, *Biophysical Chemistry,* 1988, 30, 3-48.

23 K. C. Chou, *Biopolymers*, 1987, 26, 285-295.

24 G. P. Zhou and Y. D. Cai, *PROTEINS: Structure, Function, and Bioinformatics,* 2006, 63, 681-684.

25 Y. D. Cai, G. P. Zhou and K. C. Chou, *Biophysical Journal,* 2003, 84, 3257-3263.

26 K. C. Chou, *Journal of Biological Chemistry,* 1993, 268, 16938-16948.

27 K. C. Chou, *Analytical Biochemistry,* 1996, 233, 1-14.

28 K. C. Chou, *Current Protein and Peptide Science*, 2002, 3, 615-622.

29 E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous and B. S. Pickard, *BMC bioinformatics,* 2005, 6, 55.

30 A. Schlicker, T. Lengauer and M. Albrecht, *Bioinformatics,* 2010, 26, i561-i567.

31 F. Ramírez, G. Lawyer and M. Albrecht, *Bioinformatics*, 2012, 28, 269-276.

32 F. S. Turner, D. R. Clutterbuck and C. A. Semple, *Genome biology,* 2003, 4, R75-R75.

33 E. A. Adie, R. R. Adams, K. L. Evans, D. J. Porteous and B. S. Pickard, *Bioinformatics*,2006, 22, 773-774.

34 N. López‐Bigas and C. A. Ouzounis, *Nucleic acids research*, 2004, 32, 3108-3114.

35 S. Aerts, D. Lambrechts, S. Maity, P. Van Loo, B. Coessens, F. De Smet, ... , and Y. Moreau, *Nature biotechnology*, 2006, 24, 537-544.

36 L. Franke, H. V. Bakel, L. Fokkens, E. D. De Jong, M. Egmont-Petersen and C. Wijmenga, *The American Journal of Human Genetics*, 2006, 78, 1011-1025.

37 M. A. Van Driel, K. Cuelenaere, P. P. C. W. Kemmeren, J. A. Leunissen, H. G. Brunner and G. Vriend, *Nucleic acids research*, 2005, 33, W758-W761.

38 J. Freudenberg and P. Propping, *Bioinformatics*, 2002, 18, S110-S115.

39 C. Perez-Iratxeta, P. Bork and M. A. Andrade, *Nature genetics*, 2002, 31, 316-319.

40 J. Zhu, Y. Qin, T. Liu, J. Wang and X. Zheng, *BMC bioinformatics*, 2013, 14, S5.

41 M. Krauthammer, C. A. Kaufmann, T. C. Gilliam and A. Rzhetsky, *Proceedings of the National Academy of Sciences of the United States of America*, 2004, 101, 15148-15153.

42 S. Karni, H. Soreq and R. Sharan, *Journal of Computational Biology*, 2009, 16, 181-189.

43 M. Oti, B. Snel, M. A. Huynen and H. G. Brunner, *Journal of medical genetics*, 2006, 43, 691-698.

44 J. Xu and Y. Li, *Bioinformatics*, 2006, 22, 2800-2805.

45 S. Navlakha and C. Kingsford, *Bioinformatics*, 2010, 26, 1057-1063.

46 P. F. Jonsson and P. A. Bates, *Bioinformatics*, 2006, 22, 2291-2297.

47 J. Lim, T. Hao, C. Shaw, A. J. Patel, G. Szabó, J. F. Rual, ... , and H. Y. Zoghbi, *Cell*, 2006, 125, 801-814.

48 I. Feldman, A. Rzhetsky and D. Vitkup, *Proceedings of the National Academy of Sciences,* 2008, 105 *,* 4323-4328.

49 S. Erten, G. Bebek, R. M. Ewing and M. Koyutürk, *Journal of Computational Biology,* 2011, 18, 1561-1574.

50 O. Vanunu, O. Magger, E. Ruppin, T. Shlomi and R. Sharan, *PLoS computational biology*, 2010, 6, e1000641.

51 S. Gao and X. Wang, *Journal of computer science and systems biology*, 2009, 2, 133.

52 R. Hoehndorf, P. N. Schofield and G. V. Gkoutos, *Nucleic acids research*, 2011, 39, e119.

53 P. Yang, X. Li, M. Wu, C. K. Kwoh and S. K. Ng, *PloS one*, 2011, 6, e21502.

54 T. Hwang, W. Zhang, M. Xie, J. Liu and R. Kuang, *Bioinformatics*, 2011, 27, 2692-2699.

55 S. Oliver, *Nature*, 2000, 403, 601-603.

56 D. Altshuler, M. Daly and L. Kruglyak, *Nature genetics*, 2000, 26, 135-138.

57 Y. Li and J. C. Patra, *Bioinformatics*, 2010, 26, 1219-1224.

58 X. Yao, H. Hao, Y. Li and S. Li, *BMC systems biology,* 2011, 5, 79.

59 Y. Chen, T. Jiang and R. Jiang, *Bioinformatics,* 2011, 27, i167-i176.

60 X.Guo, L. Gao, C. Wei, X. Yang, Y. Zhao and A. Dong, *PloS one,* 2011, 6, e24171.

61 Y. Li and J. C. Patra, *BMC bioinformatics*, 2010, 11, S20.

62 T. H. Pers, N.T. Hansen, K. Lage, P. Koefoed, P. Dworzynski, M. L. Miller, ... , and S. Brunak, *Genetic epidemiology*, 2011, 35, 318-332.

63 Y. Chen, W. Wang, Y. Zhou, R. Shields, S. K. Chanda, R. C. Elston and J. Li, *PloS one*, 2011, 6, e21137.

64 P. R. Costa, M. L. Acencio and N. Lemke, *BMC genomics*, 2010, 11, S9.

65 F. Mordelet and J. P. Vert, *BMC bioinformatics*, 2011, 12, 389.

66 I. Lee, U. M. Blom, P. I. Wang, J. E. Shim and E. M. Marcotte, *Genome research,* 2011, 21, 1109-1121.

67 S. Schuierer, L. C. Tranchevent, U. Dengler and Y. Moreau, *Bioinformatics*, 2010, 26, 1922-1923.

68 J. Chen, B. J. Aronow and A. G. Jegga, *BMC bioinformatics*, 2009, 10, 73.

69 S. Köhler, S. Bauer, D. Horn and P. N. Robinson, *The American Journal of Human Genetics,* 2008, 82, 949-958.

70 D. H. Le and Y. K. Kwon, *Computational biology and chemistry*, 2013, 44, 1-8

71 X. Wang, N. Gulbahce and H. Yu, *Briefings in functional genomics*, 2011, 10, 280-293.

72 B. Linghu, E. S. Snitkin, Z. Hu, Y. Xia and C. DeLisi, *Genome Biology*, 2009, 10, R91.

73 M. A. Care, J. R. Bradford, C. J. Needham, A. J. Bulpitt and D. R. Westhead, *Human mutation*, 2009, 30, 485-492.

74 P. Radivojac, K. Peng, W. T. Clark, B. J. Peters, A. Mohan, S. M. Boyle and S. D. Mooney, *Proteins: Structure, Function, and Bioinformatics*, 2008, 72, 1030-1037.

75 X. Wu, R. Jiang, M. Q. Zhang and S. Li, *Molecular Sys*tems Biology, 2008, 4, 189.

76 C. Stark, B. J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz and M. Tyers, *Nucleic acids research*, 2006, 34, D535-D539.

77 G. Östlund, T. Schmitt, K. Forslund, T. Köstler, D. N. Messina, S. Roopra, ... , and E. L. Sonnhammer, *Nucleic acids research*, 2010, 38, D196-D203.

78 A. Hamosh, A. F. Scott, J. S. Amberger, C. A. Bocchini and V. A. McKusick, *Nucleic acids research*, 2005, 33, D514-D517.

79 E. A. Leicht, P. Holme and M. E. J. Newman, *Physical Review E*, 2006, 73, 026120.

80 K. C. Chou, *Journal of Theoretical Biology,* 2011, 273, 236-247.

81 Z. C. Wu, X. Xiao and K. C. Chou, *Protein & Peptide Letters,* 2012, 19, 4-14.

82 H. Ding, L. Liu, F.B. Guo, J.Huang and H. Lin, *Protein & Peptide Letters,* 2011, 18, 58-63.

83 F. M. Li and Q. Z. Li, *Protein & Peptide Letters,* 2008, 15, 612-616.

84 L. F. Yuan, C. Ding, S.H. Guo, H. Ding, W. Chen and H. Lin, *Toxicology in Vitro,* 2013, 27, 852-856.

85 V. Tripathi and D. K. Gupta, *Journal of Biomolecular Structure and Dynamics, 2013,* doi: *10.1080/07391102.2013.827133.*

86 H. Lin, W. Chen and H. Ding, *PloS One,* 2013, 8, e75726.

87 X. Xiao, P. Wang, W. Z. Lin, J. H. Jia and K. C. Chou, *Analytical Biochemistry*, 2013, 436, 168-177.

88 H. Ding, S.H. Guo, E.Z. Deng, L.F. Yuan, F.B. Guo, J. Huang, N. Rao, W. Chen and H. Lin, *Chemometrics and Intelligent Laboratory System*, 2013, 124, 9-13.

89 J. Lin and Y. Wang, *Protein & Peptide Letters,* 2011, 18, 1219-1225.

90 R. A. Fisher, *Journal of the Royal Statistical Society*, 1922, 85, 87-94.

91 G. Liu, H. Wang, J. Liu, J. Li, H. Li, G. Ma, ... , and K. Li, *Neuromolecular medicine*, 2013, 1-9.

92 M. N. Braskie, N. Jahanshad, J. L. Stein, M. Barysheva, K. L. McMahon, G. I. de Zubicaray, ... , and P. M. Thompson, *The Journal of Neuroscience*, 2011, 31, 6764-6770.

93 P. Wong, J. Pineault, J. Lakins, D. Taillefer, J. Leger, C. Wang and M. Tenniswood, *Journal of Biological Chemistry*, 1993, 268, 5021-5031.

94 X. Li, I. Alafuzoff, H. Soininen, B. Winblad and J. J. Pei, *Febs Journal*, 2005, 272, 4211-4220.

95 L. Baum, L. Chen, E. Masliah, Y. S. Chan, H. K. Ng and C. P. Pang, *American journal of medical genetics*, 1999, 88, 136-139.

96 S. Lev, H. Moreno, R. Martinez, P. Canoll, E. Peles, J. M. Musacchio, ... , and J. Schlessinger, *Nature,* 1995, 376, 737-745.

97 Y. Teranishi, J. Y. Hur, G. J. Gu, T. Kihara, T. Ishikawa, T. Nishimura, ... , and L. O. Tjernberg, *Biochemical and Biophysical Research Communications*, 2012, 424, 476.