Bioinformatics Advance Access published June 19, 2013

Databases and ontologies

Human Interactome Resource and Gene Set Linkage Analysis for the Functional Interpretation of Biologically Meaningful Gene Sets

Xi Zhou [#], Pengcheng Chen [#], Qiang Wei, Xueling Shen and Xin Chen*

College of Life Sciences, Zhejiang University, Hangzhou, P.R. China, 310058

Associate Editor: Dr. John Hancock

ABSTRACT

Motivation: A molecular interaction network can be viewed as a network in which genes with related functions are connected. Therefore, at a systems level, connections between individual genes in a molecular interaction network can be used to infer the collective functional linkages between biologically meaningful gene sets.

Results: We present the Human Interactome Resource (HIR) and the Gene Set Linkage Analysis (GSLA) tool for the functional interpretation of biologically meaningful gene sets observed in experiments. GSLA determines whether an observed gene set has significant functional linkages to established biological processes. When an observed gene set is not enriched by known biological processes, traditional enrichment-based interpretation methods cannot produce functional insights, but GSLA can still evaluate whether those genes work in concert to regulate specific biological processes, thereby suggesting the functional implications of the observed gene set. The quality of HIR and the utility of GSLA are illustrated with multiple assessments.

Availability: http://www.cls.zju.edu.cn/hir/

Contact: xinchen@zju.edu.cn

1 INTRODUCTION

A wide range of genome-wide profiling technologies has emerged over the past decade, and these have become a foundation of current biological research (Farnham, 2009). The challenge is no longer obtaining the profiling data; instead, the challenge is the interpretation of molecular profiles, in order to elucidate biological mechanisms and overcome the "curse of spreadsheets" (Csermely., et al., 2012).

The traditional approach to interpreting a molecular profile is to focus on a handful of the most significantly changed genes. Because of the noise inherent in profiling technologies and the poor annotation of many genes, this approach has significant limitations (Subramanian, et al., 2005). To better interpret the biological significance of a molecular profile, annotation enrichment-based approaches have been developed. These approaches examine whether members of a gene set defined by an established biological concept (i.e., an annotation term, typically a biological process) tend to be enriched in the list of significantly changed genes (Li, et al., 2004; McLean, et al., 2010) or occur toward the top/bottom of the ranked list of all genes (Subramanian, et al., 2005). In this manner, these approaches determine whether an established biological concept (and the gene set defined by it) is correlated with the molecular profile. Because annotation enrichment-based analyses tend to produce results that are more reproducible and interpretable than individual gene-based analyses (Subramanian, et al., 2005), they have quickly become popular (Huang da, et al., 2009).

However, these approaches are not always useful, as that "whereof one cannot speak, thereof one must be silent" (Wittgenstein, 1922). Annotation enrichment-based approaches examine whether genes that constitute well-established biological concepts are enriched in a molecular profile. Therefore, these approaches essentially map the observed molecular profiles to established biological concepts (i.e., defining what these profiles are), and the researchers are left to interpret the functional implications of these molecular profiles using their knowledge regarding the biology of the mapped concepts. For a truly novel molecular profile that does not have established concepts to describe it, annotation enrichment-based approaches would fail to map to established concepts; consequently, researchers would be unable to appreciate the functional implications of observing such a profile by examining the "identity" of the profile.

However, it is important to understand the functional implications of novel molecular profiles, because novel molecular profiles are often the basis for the introduction of new biological concepts that expand our knowledge. To meet this analytical challenge, we developed a tool called Gene Set Linkage Analysis (GSLA), which detects significant functional linkages between "biologically meaningful gene sets", i.e., experimentally observed "significantly changed genes" and gene sets defined by established biological concepts. Functional linkages that relate an observed molecular profile to well-established biological concepts might assist in interpreting the functional implications of the observed molecular profile even in a case in which no established concept can be found to describe (i.e., be enriched in) the observed molecular profile.

To identify these collective functional linkages between gene sets, we developed the GSLA tool, which evaluates a largely homogenous functional interactome, the Human Interactome Resource (HIR). Here, "homogenous functional interactome" means that the functional interactions in the interactome have similar strengths and represent an unbiased sample of the true (complete) functional interactome. In a homogenous functional interactome, the density of inter-gene-set interactions will be able to indicate the overall strength of the functional linkage between two gene sets. The idea of exploring interaction density for functional inference has been discussed previously (Asur, et al., 2007; Bader and Hogue, 2003; Brun, et al., 2004). Based on the coverage and reliability of HIR, we designed a pair of tests that validate both the "strength" and the "biological significance" of the observed inter-

*To whom correspondence should be addressed.

# These authors made equal contributions.

gene-set interaction density to detect substantial functional linkage at the gene set level.

In the following sections, we provide a description of our method for obtaining a largely homogenous functional interactome HIR and the details of our GSLA approach. The quality of HIR and the utility of GSLA were evaluated in a series of systematic evaluations and case studies.

2 RESULTS

2.1 Method for constructing a homogenous functional interactome

GSLA relies on a simple strategy that considers the density of inter-gene-set interactions as the measure of the overall strength of the functional linkage between two gene sets. For this simple strategy to work, the underlining functional interactome needs to be homogenous, i.e., consist of functional interactions that are of similar strength and represent an unbiased sample of the true functional interactome. Obtaining a homogenous functional interactome is challenging because "functional interaction" is an umbrella term that describes diverse types of biological mechanisms that connect the functions of two genes. These diverse types of biological mechanisms are exhibited in diverse types of data. Functional interactions suggested by only one type of data are often unable to represent an unbiased sample of the complete functional interactome.

For example, a protein interaction network is a type of functional interactome in which functional interactions are suggested by only one type of data – physical protein interactions. The structure of protein interaction networks has been extensively studied (Bu, et al., 2003; Han, et al., 2004). In a typical protein interaction network, there are clusters of densely inter-connected genes, which are called modules (Han, et al., 2004). Network modules often encode cellular processes (or functions). There are bridge proteins that connect two modules and mediate the flow of information between the modules. Bridges are therefore effective regulators of modules (Yu, et al., 2007). For example, bridges were found to be effective drug targets (Hwang, et al., 2008; Nguyen and Jordan, 2010; Nguyen, et al., 2011), and unique bridges (bottlenecks) were found to be preferential targets of regulatory microRNAs (Wang, et al., 2011). The fact that bridges are effective module regulators indicates that bridges have strong functional interactions with the proteins in the modules that they regulate. In the true functional interactome, many functional interactions are expected between the bridges and proteins in the modules that the bridges regulate. However, in a protein interaction network, by definition, bridges interact with only a few proteins in the modules that the bridges regulate: proteins that interact extensively with module members are considered to be module members rather than bridges. In fact, many bridges that connect two modules have only two interactions, one interaction with each module. Therefore, the many functional interactions between the bridges and proteins in the modules that the bridges regulate are under-represented in a protein interaction network. This arrangement means that in a protein interaction network, the number of interactions between a bridge protein and proteins constituting a bridge-regulated module does not reflect the strength of the functional linkage between this bridge protein and the cellular process (or function) that is encoded by the module. Accordingly, the simple strategy of GSLA that uses inter-gene-set interaction density to measure the overall strength of the functional linkage between two gene sets will not work well if bridge proteins are involved in the gene sets.

One approach to constructing a homogenous functional interactome is to integrate multiple types of data that suggest functional interactions between genes. For example, the functional interactions between the bridge proteins and the proteins in the modules that the bridges regulate, which are under-represented in protein interaction networks, could be revealed in co-expression networks or co-evolution networks. However, integrating multiple types of data is not straightforward because we do not currently have an intuitive or established way to consistently and rigorously measure the "strength" of the functional interaction suggested by multiple types of data. Therefore, we took a rather rough approach to measuring the "strength" of a functional interaction. We assume that the strengths of the functional interactions between the genes that encode physically interacting proteins were approximately at the same level. Therefore, we used a statistical learning approach to identify pairs of genes that are most similar to the pairs of genes that encode physically interacting proteins, with respect to multiple types of data that suggest functional interaction. In this manner, we integrated multiple types of data to infer the functional interactions between genes, which were expected to represent the true functional interactome in a more unbiased manner. At the same time, these inferred functional interactions were of similar strengths, approximating the strengths of functional interactions between genes that encode physically interacting proteins.

Technically, this approach was similar to predicting physical protein interactions based on evidence of functional association, which has been previously explored for validating interactions using unreliable high-throughput experiments (Giot, et al., 2003) or directly predicting novel protein interactions (Basso, et al., 2005). However, in our methodological framework, this approach was used for a different purpose: implementing a consistent (although not rigorous) measure of the "strength of functional interaction" between two genes when functional interactions were suggested by multiple types of data.

The workflow of our interaction inference procedure is outlined in Figure 1 and is detailed in the Supplemental Text. To train an inference model, a gold-standard example dataset consisting of positive examples (reliably known interactions) and negative examples (random pairs of proteins) was assembled. For each example (a protein pair), 25 features (real numbered values) measuring five types of data that suggest functional association were computed. The quality of these features, i.e., their ability to distinguish positive and negative examples, was assessed. Then, a support vector machine prediction model was created to determine the contrast between the feature values in the positive and the negative examples. This contrast was used to predict whether an arbitrary pair of proteins was likely to interact based on its feature values. The inference model was evaluated using newly curated interactions in the major data repositories. Newly curated interactions refer to interactions that are curated after the time when all of the data that were used for interaction prediction were collected.

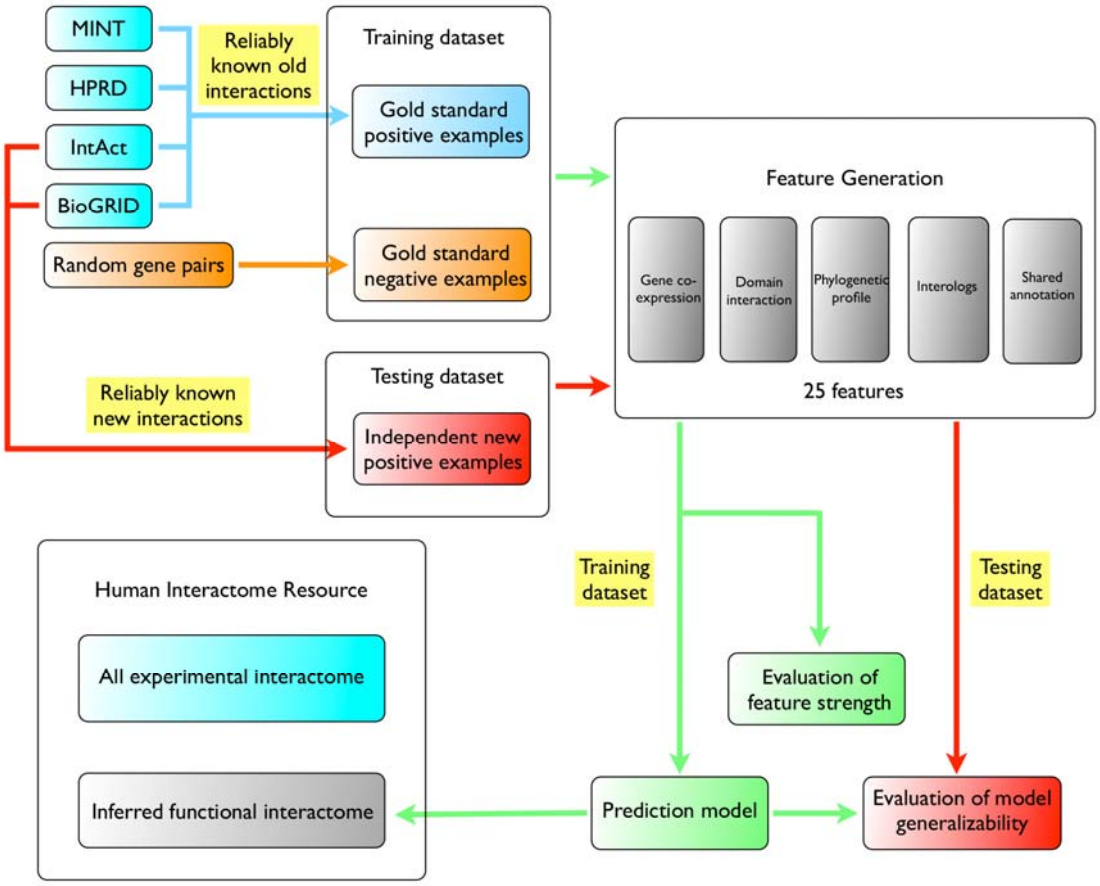

Fig.1. General workflow of interaction prediction.

The inferred interactions and the experimentally reported protein interactions were integrated to create the HIR interactome, which contains 155,974 non-redundant interactions. Evaluated as a physical protein interactome, HIR was expected to cover 22.1% of the entire human interactome with a per-interaction reliability of 41.0%. HIR can be considered a largely homogenous functional interactome. As shown in the Supplemental Text, the inferred interactions and the experimentally reported interactions were similar in the overall strength of functional interaction, and the integrated interactome showed an improved representation of functional linkages between the bridge proteins and the module proteins that the bridges regulate (which were under-represented in protein interactions). In addition, a case study was conducted to illustrate that HIR provided an up-to-date view of the functional associations between genes.

2.2 Gene set linkage analysis

Based on HIR, the gene set linkage analysis (GSLA) tool was created to identify functional linkages between user-defined gene sets (which can represent any type of biological significance) and established GO biological processes. The functional linkages between the GO biological processes are pre-computed and searchable on the HIR website. The topology of the GO biological process network is analyzed in the next section as part of the validation of the GSLA approach.

GSLA relies on a test of two hypotheses to detect substantial functional linkages between biologically meaningful gene sets. The first hypothesis (Q1) expects that the inter-gene-set interaction density between functionally linked gene sets is higher than the background interaction density between random genes. The second hypothesis (Q2) expects that the observed high density between functionally linked gene sets can be observed only in the biologically correct interactome, i.e., the density observed in HIR is higher than the densities observed in random interactomes consisting of the same genes, with each gene having the same number of neighbors. Q1 tests the strength of the functional linkage between two gene sets. Q2 verifies that the observed strong linkage is a result of the biologically correct network topology (i.e., our knowledge about the molecular mechanisms) rather than a result of the compositions of these two gene sets. Some genes known as hubs have considerably more neighbors than other genes. Gene sets that include many hubs are therefore more likely to connect to other gene sets. Q2 is used to remove the confounding factor of gene set composition and to ensure the "biological significance" of the detected functional linkages between gene sets. Q1 and Q2 are related but different tests. They complement each other to increase the sensitivity and specificity of GSLA.

To test the Q1 hypothesis, the density of the inter-gene-set interactions is first calculated as the observed number of inter-gene-set interactions divided by the total number of possible inter-gene-

set gene pairs. The number of inter-gene-set interactions was computed as follows. A common gene shared by two gene sets was treated as two distinct delegate genes, with each delegate gene belonging to only one gene set. Any gene that interacted with the shared gene was considered to interact with both delegate genes in both gene sets. The fact that two gene sets shared a gene was not by itself considered an inter-gene-set interaction.

According to the central limit theorem, the expected density of interactions in a sufficiently large number of pairs of genes, N, follows the normal distribution. As shown in Figure 2, the expected mean of this normal distribution is the density of the interactions among all genes (0.0008). The expected standard deviation approaches 0.0012 with an increase in N in our simulation, where N random gene pairs were sampled 100,000 times to calculate the mean and standard deviation of the interaction density. According to these data, a not-very-high inter-gene-set interaction density of 0.01 translates into a very small p- value ($p<10^{-10}$), when the two gene sets both have more than 20 members ($N \geq 400$). Functional linkages between two gene sets that are worthy of experimental investigation are anticipated to be much stronger than those linkages that marginally pass the significance threshold. Therefore, we used a density-based cutoff for Q1 (density $\geq$ 0.01) in later analyses. P values were no longer used because the densities have a more intuitive biological interpretation (the strength of the gene set interaction), and the p-values for the densities above this cutoff were always significant.

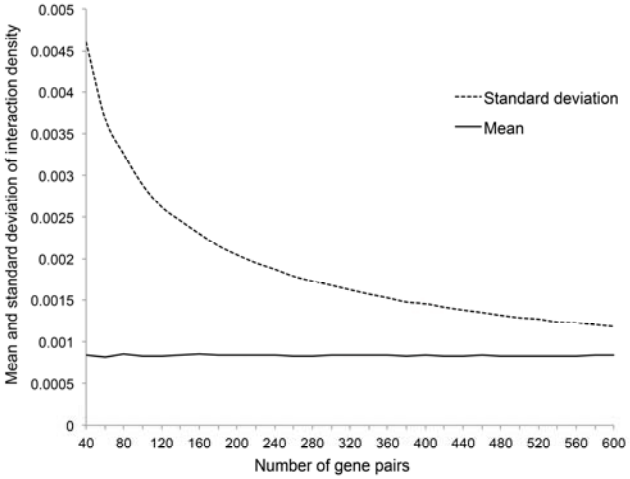

Fig. 2. The mean and standard deviation of the expected interaction density between random genes.

To test the Q2 hypothesis, the HIR interactome was randomized 100,000 times. Suppose that there were two interactions that involved four genes, G1-G2 and G3-G4. By replacing these two interactions with two new interactions, G1-G3 and G2-G4, the topology of HIR is changed, but the gene composition and the number of neighbors for each gene remains the same. This type of topology perturbation was performed 100,000 times to create one random interactome. With 100,000 random interactomes, the p-value for Q2 was computed as the fraction of random interactomes in which the densities of inter-gene-set interactions were higher than the density observed in HIR. A cutoff of $p \leq 1\times10^{-4}$ was used.

To interpret the biological significance of an experimentally observed molecular profile, we recommend using the top 20-200 "significant" genes to create a gene set that represents the phenotype. The GSLA algorithm outlined above treats all of the genes in the same gene set equally. Therefore, using too many genes will dilute the focus on the most prominent changes. On the other hand, because profiling data are intrinsically noisy and the interaction network has false positives and false negatives, using too few genes will not be able to suppress this noise and take advantage of the corroborative strength of a gene set. In our experience, using the top 20-200 significant genes to represent a molecular profile usually produced consistent results. In addition, GO biological processes that are functionally linked to gene sets of such sizes were typically specific in concept and therefore useful for further research. For example, as detailed in later sections, when interpreting the expression changes induced by statin treatment, using the top 20, 50, 100, and 200 genes to represent the expression profiles produced consistent results (Supplemental Table S1).

2.3 A functional linkage network connecting GO biological processes

The general, knowledge concerning the mechanisms of life is organized in the manner that complex cellular functions arise from composite coordination between biological processes, which consist of interacting molecules (Lu, et al., 2007). Extensive efforts have been made in the past to define these biological processes in terms of their constituent genes. However, the functional coordination between biological processes was poorly defined. For example, the GO biological processes are arguably the most popular definitions of biological processes (Ashburner, et al., 2000). GO provides semantic relationships between biological processes at different conceptual scales, but it does not provide the functional linkages between semantically different biological processes. Therefore, to evaluate whether GSLA could identify biologically meaningful functional linkages, we computed the functional linkages between all GO biological processes and analyzed them. The GO biological process linkage network (GO-BPLN) produced by GSLA was composed of 38,984 linkages that connect 1,850 biological processes (Supplemental Table S2).

In the GO-BPLN, a significant proportion of the functional linkages were supported by semantic similarity or by genes that were shared between biological processes. Wang et al. proposed an approach to measure the semantic similarity between two biological processes, which considered both the number and the location of their common ancestor processes in the GO semantic hierarchy (Wang, et al., 2007). For example, the two processes "translational initiation" (GO:0006413) and "translational termination" (GO:0006415) had a semantic similarity score of 0.61. Using this approach, we computed the average semantic similarity between processes that were connected in GO-BPLN. This average semantic similarity was significantly higher than that expected in random biological process networks that were composed of the same processes, with each process having the same number of neighbors (0.20 vs. 0.13, $p=1\times10^{-8}$). These data suggested consistency between the topology of GO-BPLN and the semantic hierarchy of GO. Notably, not all of the processes connected in GO-BPLN were semantically similar. In particular, there were 5,313 (13.6%) linkages that connected biological processes that shared no common ancestor other than the root process (GO:0008150, "biological process"). For example, "epidermal growth factor receptor (EGFR) transactivation by G-protein coupled receptor (GPCR) signaling pathway" (GO:0035625) was connected to "sensory perception" (GO:0007600). It was shown that EGFR can be trans-activated in a

calcium-dependent manner (Zwick, et al., 1997) and that sensory perception involves many GPCRs that regulate calcium levels (Dong, et al., 2001).

On the other hand, in GO-BPLN, 24,071 (61.7%) linkages connected biological processes that shared one or more genes. This percentage was significantly larger than the percentage of gene-sharing linkages expected in random biological process networks that were composed of the same processes, with each process having the same number of neighbors (61.7% or 24,071 interactions vs. 29.5% or 11,518 interactions, $p = 1 \times 10^{-3}$). These data suggested correlation between the topology of GO-BPLN and the compositions of these biological processes. It is important to note that two gene-sharing processes were not necessarily connected. For example, the "activin receptor signaling pathway" (GO:0032924) and the "cell-cell junction organization process" (GO:0045216) shared one gene, but this inter-gene-set interaction density did not qualify for a substantial functional linkage at the gene set level. Likewise, two biological processes that shared no genes were not necessarily disconnected. For example, "histone H3-K9 methylation" (GO:0051567) was connected to "negative regulation of transcription from RNA polymerase II promoter" (GO:0000122). This functional linkage was supported by the finding that histone H3 methylation is associated with transcriptional elongation by RNA polymerase II (Krogan, et al., 2003; Lehnertz, et al., 2003), and histone H3K9 methylation is required in siRNA-related centromeric gene silencing, in which centromeric RNA is transcribed by RNA polymerase II (Moazed, 2009). Notably, this functional linkage was identified based on the inferred functional interactions between the SETD8, SIRT1, and HDAC family genes (which have the annotation of negative regulation of transcription from the RNA polymerase II promoter) and the histone methylation enzyme SUV39H1. Therefore, successful identification of this functional linkage also suggested the usefulness of including inferred interactions in HIR.

The above results indicated that the topology of the GO-BPLN inferred by GSLA was consistent to our expectations, with respect to the semantics and gene compositions of the GO biological processes. In addition, we also conducted a case study to show that GSLA identified biological processes with functional linkages to the c-Jun N-terminal kinase (JNK) cascade process, which were mostly supported by previous research (Supplemental Text). In short, our evaluation suggested that GSLA reliably identified functional linkages between established biological processes.

2.4 A case study to illustrate the capability of GSLA to interpret the functional implications of observing a gene set in an experiment

In addition to identifying functional linkages between well-established biological processes, an important use of GSLA is to suggest the functional implications of observing a gene set in an experiment by connecting that gene set to established biological processes. This use of GSLA is illustrated in the following re-analysis of the expression profiles that describe statin-induced selective apoptosis.

Statins are a family of hydroxymethylglutaryl coenzyme A reductase (HMGCR) inhibitors commonly used to treat patients with hypercholesterolemia. Recently, statins were shown to induce apoptosis in a variety of tumor cells, including multiple myeloma (MM) cells. However, a reduction in cancer risk was observed only in some, but not all, MM patients undergoing early-phase clinical trials. It was therefore important to understand the selective mechanisms of statin. Clendening et al. reported that four MM cell lines, H929, KMS11, LP1, and SKMM1, showed dichotomized responses to lovastatin. H929 and KMS11 were sensitive, whereas LP1 and SKMM1 were not. The statin-induced expression changes that occurred in these cell lines were examined using traditional annotation enrichment analysis (AEA) (Clendening, et al., 2010). Their results showed that 22 biological processes were enriched for differentially expressed genes in the two insensitive cell lines. These processes included the cholesterol-related metabolic processes, which was consistent with the notion that lovastatin targets the cholesterol biosynthesis pathway, and this pathway was transactivated by its enzyme-level inhibition. In contrast, in the two sensitive cell lines, no biological processes were consistently enriched; i.e., the enzyme-level inhibition of the cholesterol biosynthesis pathway did not induce its expression.

Although AEA revealed the existence of different cellular responses to lovastatin, it identified no biological process related to apoptosis activation. However, the aim of measuring expression profiles before the onset of apoptosis was to identify mechanisms of statin action that are independent of the general apoptosis changes (Clendening, et al., 2010). For this purpose, AEA failed to provide useful insight.

The same expression profiles were re-analyzed with GSLA. The top 50 differentially expressed genes were used to define the "expression phenotype" of each cell line after lovastatin exposure (Supplemental Table S3). AEA analysis of these phenotype gene sets generated the same conclusions as those reported in the original publication (Supplemental Table S4), which suggested that the top 50 genes from each profile were adequate to capture the differences between the statin responses in sensitive and in insensitive cell lines.

GSLA analysis of these phenotype gene sets produced three notable results (Supplemental Table S4). First, consistent with AEA, GSLA indicated that cholesterol-related metabolic processes were regulated in the insensitive cell lines but not in the sensitive cell lines. However, unlike AEA, GSLA was able to indicate that apoptosis-related biological processes were regulated in the sensitive cell lines but not the insensitive cell lines, which predicted the observed apoptosis behavior of the sensitive cell lines. In addition, GSLA suggested cell line-specific biological processes that might lead to the activation of apoptosis in each sensitive cell line. In the KMS11 cell line, Toll-like receptor (TLR) signaling pathways were regulated, whereas in the H929 cell line, Rho/GTPase signaling pathways were regulated. These processes were known to be regulated by statin in other cell types and are known triggers of apoptosis (Supplemental Text).

To evaluate the usefulness of GSLA, we also compared two other analysis tools that interpret the biological significance of experimental observations. The Fixed Network Enrichment Analysis (FNEA) tool (Supplemental Table S5) is another network-based tool that interprets the biological significance of an experimentally observed gene set (Alexeyenko, et al., 2012). The Gene Set Enrichment Analysis tool (Supplemental Table S6) is a famous expression analysis tool that analyzes the entire expression profile instead of the highest altered genes. GSEA does not utilize network information. In the scenario of two phenotypes (treatment/control), GSEA can be considered to be a simple enrichment analysis, but GSEA uses a score that is similar to the score of the Kolmogorov-Smirnov test to measure the enrichment (Subramanian, et al., 2005), whereas AEA uses a hyper-geometric test to measure enrichment (Boyle, et al., 2004).

Statin sensitivity	Cell lines	Method used in the analysis of altered genes	Apoptosis related processes	Sterol metabolic related processes	Rho/GTPase related processes	Toll-like receptor related processes
Statin-sensitive cell lines	H929	GSLA	1,4	None	2,7	None
		FNEA	None	None	5,12	None
		AEA	None	None	None	None
		GSEA	None	None	None	None
	KMS11	GSLA	10,18	None	None	1,8,11,15,16,19
		FNEA	None	None	None	5,6,8,9,11,12,13,19
		AEA	None	None	None	None
		GSEA	None	14	None	None
Statin-insensitive cell lines	LP1	GSLA	None	1~11,13,18	16	None
		FNEA	None	1~5,7~9,11	None	None
		AEA	None	1~8,10,11,16	None	None
		GSEA	None	1	None	None
	SKMM1	GSLA	None	3~13	None	None
		FNEA	None	1~6,9,16~19	None	None
		AEA	None	1~8,11,13,15,16,17	None	None
		GSEA	None	1,10,15,17	None	None

Table 1. Ranks of four categories of biological processes in the top 20 biological processes that are identified with four analyses. FNEA, fixed network enrichment analysis; AEA, annotation enrichment analysis; GSEA, gene set enrichment analysis. The full GSLA, FNEA, AEA and GSEA results, together with significance scores, are provided in Supplemental Tables S4, S5, and S6.

Table 1 shows the ranks of four categories of biological processes, which were most likely relevant to this study, in the results produced by GSLA, AEA, FNEA and GSEA. The four categories included the apoptosis-related biological processes, which represented the observed phenotype of statin-sensitive cell lines; the cholesterol-related metabolic processes, which are the pharmacological targets of statin action and distinguished the statin-sensitive/insensitive cell lines in the classical AEA analysis; the Toll-like receptor (TLR) signaling pathways and the Rho/GTPase signaling pathways, which were implicated by GSLA as cell line-specific routes to active apoptosis (with indirect literature support).

AEA, FNEA and GSLA consistently identified the cholesterol metabolic processes as the distinguishing responses between the sensitive/insensitive cell lines, whereas with GSEA, one of the sensitive cell lines also showed differential expression in the cholesterol metabolic processes (similar to the insensitive cell lines). In addition, FNEA and GSLA consistently implicated that the TLR signaling pathways and the Rho/GTPase signaling pathways were regulated in the statin-sensitive cell lines. However, only GSLA was able to indicate that the expression changes of sensitive cell lines implied apoptosis activation.

In summary, this case study illustrated the difference between GSLA and traditional enrichment analyses. GSLA identifies the biological processes whose members have extensive functional interactions to the changed genes that change in an experiment. In other words, GSLA asks the question of what biological processes these changed genes might together regulate. In contrast, enrichment analysis asks the question of what biological processes constitute the changed genes. In other words, the GSLA question is about the "functional implication" of the changed genes, but the enrichment question is about the "identity" of the changed genes. Traditionally, with enrichment analyses, we first find the "identity" of the changed genes in terms of established biological processes; then, we use our knowledge about these processes to manually interpret the "functional implications" of observing these processes. In contrast, GSLA directly provides the "functional implications" without the need to explicitly define the "identity" of the changed genes, thereby circumventing the problem that some sets of changed genes observed in experiments might not have an "identity" in terms of established biological processes ("whereof one cannot speak, thereof one must be silent").

This case study was chosen to illustrate this difference. All of the expression profiles were measured before the first onset of apoptosis. Therefore, the "identity" of these profiles was not apoptosis. Enrichment analyses, by design, are not expected to annotate them to apoptosis. AEA analysis of the changed genes in sensitive cell lines produced biological processes that are relatively general in concept (Supplemental Table S4), which suggests that these changed genes did not have a good "identity" in terms of conceptually specific biological processes. Without such an identity, our knowledge about biological mechanisms cannot be used to interpret the functional implications of observing these changed genes. However, statin treatment induced apoptosis in sensitive cells. Therefore, apoptosis activation is a biologically important "functional implication" of statin-induced expression changes. GSLA is designed to detect such functional implications without the prerequisite to answer the question of what the changed genes are in terms of established biological processes.

In addition, in GSLA analysis, functional linkages between gene sets were supported by functional interactions between the individual genes in HIR. Approximately 41% of these gene level interactions represent physical protein interactions (Supplemental Text). These interactions can provide further assistance in the design of experiments to confirm the functional linkages at the gene set level.

2.5 Web interface to HIR and GSLA

The GSLA tool and its supporting HIR interactome have a unified website (http://www.cls.zju.edu.cn/hir/). A lightweight graphical network viewer was integrated within the website for visualizing multiple functional interactions between genes or gene sets. The functional linkages between GO biological processes are precomputed and can be searched with GO term accessions. To link an observation-derived molecular profile to established biological processes, the phenotype gene set must be uploaded as a list of HGNC gene IDs. An ID mapping tool is provided for converting other gene or protein IDs to HGNC IDs. For each identified functional linkage between gene sets, the gene level functional interac-

tions supporting this linkage are provided. All gene level functional interactions in the HIR interactome can be queried at the website or downloaded as compressed datasets. Functional interactions can be searched by specifying one or both of their component genes. Genes can be searched by their specific IDs or by sequence. For each functional interaction, the multiple types of data supporting this interaction are provided along with other details of this interaction. In addition, although predicting gene functions with the "guilt-by-association" strategy does not actually implement the idea of GSLA (see Supplemental Text), we nevertheless provide the predicted biological process annotations for each gene in HIR to facilitate related research.

3 DISCUSSION

GSLA summarizes functional linkages at the individual gene level to identify functional linkages at the gene set level. Therefore, the quality of the reference network has a critical impact on the usefulness of the approach. The reason why GSLA produced biologically interesting functional linkages in our assessments is most likely because GSLA used an interaction network that has balanced coverage and specificity.

For example, assuming that we aim to identify functional linkages between specific biological concepts and, in particular, that we aim to identify functional linkages between two small biological processes, each consisting of 20 genes, if 10 functional interactions between genes in these two processes were good evidence for a biologically interesting gene set interaction, then we would need a reference network that can cover at least 20% of the true interactome to observe at least 2 inter-gene-set functional interactions. In this case, a false positive rate of only 5×10^{-3} false interactions per pair of genes would result in 2 random false interactions between two 20-gene processes. If HIR had a false positive rate this high, such a signal-to-noise ratio would make it difficult to identify biologically meaningful functional linkages. As detailed in the Supplemental Text, the interactions in HIR were predicted using data prior to Jan 10, 2011. Using new interactions reported after this date, we estimated that HIR covered 22.1% of the physical human interactome with a per-interaction reliability of 41.0%. This per-interaction reliability corresponds to a false positive rate of 5.08×10^{-4} false protein interactions per pair of genes or 0.2 random false protein interactions between two 20-gene processes. This signal-to-noise ratio was likely to allow the identification of biologically meaningful functional linkages at the gene set level.

Before this work, several other human interactomes have been developed, e.g., the STRING database (Franceschini, et al., 2013) and the HumanNet (Franceschini, et al., 2013; Lee, et al., 2011). These interactomes were not rigorously assessed for their coverage and reliability using newly reported interactions. Typically, they included half a million interactions or more ($>2.5\times10^{-3}$ interactions per pair of genes). In contrast, the rate of yeast protein interactions is approximately 1.3×10^{-3} per pair of genes (Yu, et al., 2008), and we estimated in this work that the rate of human interactions is approximately 1.5×10^{-3} per pair of genes (Supplemental Text). Therefore, the false positive rates in these interactomes are likely high, which might not provide sufficient signal-to-noise ratios for GSLA to identify biologically meaningful gene set interactions. For example, if HumanNet or STRING were used in the above statin case study, GSLA would no longer be able to connect the expression changes in statin-sensitive cells to apoptosis-related biological processes.

To further improve our network-powered "functional implication" interpretation approach, we believe that the bottleneck is the creation of a high-quality network that can provide a sufficient signal-to-noise ratio for downstream approaches to accurately summarize gene-level functional interactions into gene set-level functional linkages. Our approach of building a protein interaction-like functional interactome shows promise: GSLA identified biologically meaningful functional linkages in a series of assessments. However, this approach is still limited, primarily because its theoretical basis involves many concepts that are still poorly defined, particularly, the "strength" of a functional interaction. We hope that this work inspires future research to define a rigorous and consistent measurement of the strength of a functional interaction as suggested by multiple types of data, which is likely the key to building accurate functional interactomes and useful network-powered gene set interpretation approaches.

FUNDING

This work is supported by the National Basic Research Program of China (973) Grant No. 2012CB944900, the National Science Foundation of China (NSFC) Grant No. 30970690 and No. 31071161, the Zhejiang Provincial Natural Science Foundation of China Grant No. LR13C020001, the Program for New Century Excellent Talents in University NCET-10-0722, and the Fundamental Research Funds for the Central Universities (Zhejiang University).

REFERENCES

Alexeyenko, A., *et al.* (2012) Network enrichment analysis: extension of gene-set enrichment analysis to gene networks, *BMC bioinformatics*, **13**, 226.

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

Asur, S., Ucar, D. and Parthasarathy, S. (2007) An ensemble framework for clustering protein-protein interaction networks, *Bioinformatics*, **23**, i29-40.

Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks, *BMC bioinformatics*, **4**, 2.

Basso, K., *et al.* (2005) Reverse engineering of regulatory networks in human B cells, *Nature genetics*, **37**, 382-390.

Boyle, E.I., *et al.* (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics*, **20**, 3710-3715.

Brun, C., Herrmann, C. and Guenoche, A. (2004) Clustering proteins from interaction networks for the prediction of cellular functions, *BMC bioinformatics*, **5**, 95.

Bu, D., *et al.* (2003) Topological structure analysis of the protein-protein interaction network in budding yeast, *Nucleic acids research*, **31**, 2443-2450.

Clendening, J.W., *et al.* (2010) Exploiting the mevalonate pathway to distinguish statin-sensitive multiple myeloma, *Blood*, **115**, 4787-4797.

Csermely., P., *et al.* (2012) Structure and dynamics of molecular networks: A novel paradigm of drug discovery. A comprehensive review. *Invited review to Pharmacology & Therapeutics*. arXiv.

Dong, X., *et al.* (2001) A diverse family of GPCRs expressed in specific subsets of nociceptive sensory neurons, *Cell*, **106**, 619-632.

Farnham, P.J. (2009) Insights from genomic profiling of transcription factors, *Nature reviews. Genetics*, **10**, 605-616.

Franceschini, A., *et al.* (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration, *Nucleic acids research*, **41**, D808-815.

Giot, L., *et al.* (2003) A protein interaction map of Drosophila melanogaster, *Science*, **302**, 1727-1736.

Han, J.D., *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network, *Nature*, **430**, 88-93.

Huang da, W., Sherman, B.T. and Lempicki, R.A. (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists, *Nucleic acids research*, **37**, 1-13.

Hwang, W.C., Zhang, A. and Ramanathan, M. (2008) Identification of information flow-modulating drug targets: a novel bridging paradigm for drug discovery, *Clinical pharmacology and therapeutics*, **84**, 563-572.

Krogan, N.J., *et al.* (2003) Methylation of histone H3 by Set2 in Saccharomyces cerevisiae is linked to transcriptional elongation by RNA polymerase II, *Molecular and cellular biology*, **23**, 4207-4218.
Lee, I., *et al.* (2011) Prioritizing candidate disease genes by network-based boosting of genome-wide association data, *Genome research*, **21**, 1109-1121.
Lehnertz, B., *et al.* (2003) Suv39h-mediated histone H3 lysine 9 methylation directs DNA methylation to major satellite repeats at pericentric heterochromatin, *Current biology : CB*, **13**, 1192-1200.
Li, K.C., *et al.* (2004) A system for enhancing genome-wide coexpression dynamics study, *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 15561-15566.
Lu, L.J., *et al.* (2007) Comparing classical pathways and modern networks: towards the development of an edge ontology, *Trends in biochemical sciences*, **32**, 320-331.
McLean, C.Y., *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions, *Nature biotechnology*, **28**, 495-501.
Moazed, D. (2009) Small RNAs in transcriptional gene silencing and genome defence, *Nature*, **457**, 413-420.
Nguyen, T.P. and Jordan, F. (2010) A quantitative approach to study indirect effects among disease proteins in the human protein interaction network, *BMC systems biology*, **4**, 103.
Nguyen, T.P., Liu, W.C. and Jordan, F. (2011) Inferring pleiotropy by network analysis: linked diseases in the human PPI network, *BMC systems biology*, **5**, 179.
Subramanian, A., *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545-15550.
Wang, C., *et al.* (2011) Topological properties of the drug targets regulated by microRNA in human protein-protein interaction network, *Journal of drug targeting*, **19**, 354-364.
Wang, J.Z., *et al.* (2007) A new method to measure the semantic similarity of GO terms, *Bioinformatics*, **23**, 1274-1281.
Wittgenstein, L. (1922) *Tractatus logico-philosophicus*. Wilhelm Ostwald's Annalen der Naturphilosophie, Germany.
Yu, H., *et al.* (2008) High-quality binary protein interaction map of the yeast interactome network, *Science*, **322**, 104-110.
Yu, H., *et al.* (2007) The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics, *PLoS computational biology*, **3**, e59.
Zwick, E., *et al.* (1997) Critical role of calcium- dependent epidermal growth factor receptor transactivation in PC12 cell membrane depolarization and bradykinin signaling, *The Journal of biological chemistry*, **272**, 24767-24770.