ABSTRACT
Motivation: Considerable attention has been directed in recent years toward the development of methods for the contextual analysis of expression data using interaction networks. Of particular interest has been the identification of active subnetworks by detecting regions enriched with differential expression. In contrast, however, very little effort has been made toward the application of comparable methods to other types of high-throughput data.

Results: Here, we propose a new method based on co-clustering that is specifically designed for the exploratory analysis of large-scale, RNAi-based functional screens. We demonstrate our approach by applying it to a genome-scale dataset aimed at identifying host factors of the human pathogen, hepatitis C virus (HCV). In addition to recovering known cellular modules relevant to HCV infection, the results enabled us to identify new candidates and formulate biological hypotheses regarding possible roles and mechanisms for a number of them. For example, our analysis indicated that HCV, similar to other enveloped viruses, exploits elements within the endosomal pathway in order to acquire a membrane and facilitate assembly and release. This echoed a number of recent studies which showed that the ESCRT-III complex is essential to productive infection.

Supplementary Information: Supplementary data are available at Bioinformatics online.

Received on March 24, 2011; revised on July 18, 2011; accepted on August 7, 2011

1 INTRODUCTION
An important challenge in the analysis of high-throughput datasets, such as transcriptomic and proteomic data, is the integration of prior knowledge in their interpretation. One class of methods that has arisen to address this challenge is gene set analysis (GSA). (Dinu et al., 2009; Goeman and Buhlmann, 2007; Liu et al., 2007). Motivated by the increasing body of evidence that cellular responses are usually organized into pathways or active modules (Hartwell et al., 1999; Vidal, 2001), GSA methods try to assess the overall association between the phenotype of interest and known groupings of genes, by comparing the latter with the data. Although hugely successful, as evidenced by a large body of literature, GSA suffers from a number of drawbacks. Probably foremost of these, and somewhat ironic, is its reliance on predetermined gene sets. As a consequence of this dependency, GSA cannot be expected to discover novel pathways. In addition, with few exceptions (see Massa et al., 2010; Tacci et al., 2009), GSA algorithms treat pathways simply as sets; i.e. they completely ignore information on the individual interactions between entities (e.g. topology). Such information could potentially also be of use.

To complement the shortcomings of GSA, methods that try to more generally integrate interaction with expression data, outside the scope of canonical pathways, have emerged. We collectively refer to these as significant area search (SigArSearch) methods.

The optimization framework introduced by Ideker et al. (2002) for SigArSearch later inspired the development of several improvements. For example, a number of groups proposed the use of alternative heuristic search strategies, such as the greedy approaches of Sohler et al. (2004) and Nacu et al. (2007). Others proposed extensions to the scoring system, such as including scores defined on the edges (interactions) (Cabusora et al., 2005) instead of just the vertices (genes). Subsequently, Dittrich et al. (2008) pointed out that the existing methods (based on the optimization framework)
were heuristic in nature, and as such could not guarantee to identify the maximally scoring subgraph. In light of this, they presented an exact approach that was described as being able to deliver provably optimal solutions in reasonable time, which worked by reducing instances of the maximal subgraph problem to an integer-linear program. Other recent developments include extensions to the scoring system that incorporate measures of gene coexpression (whereas previous methods only used the individual data values of genes) (Guo et al., 2007; Ulitsky and Shamir, 2007), and an approach based on random walks that forgoes the optimization paradigm altogether (Koumuro et al., 2010). Although these methods are already quite popular for use with expression data, applications to other types of high-throughput datasets are still limited. In this study, we present a new method, inspired by the co-clustering approach of Hu et al. (2002), that is specifically intended for use with RNAI-based functional screens.

Functional studies in cultured cells were hampered in the past by the lack of a powerful method for perturbing gene activities (Echeverri and Perrimon, 2006). This changed with the discovery of RNA interference (RNAI) and the subsequent development of siRNA libraries aimed at targeting complete genomes for a number of organisms (Birmingham et al., 2009). Indeed, RNAI screens have proven to be effective at identifying genes associated with various biological processes, including cellular differentiation (Chia et al., 2010; Hu et al., 2009; Zhao and Ding, 2007), cancer (Bauer et al., 2010; Wurak et al., 2010; Zender et al., 2008), signaling (Barnes et al., 2004; DasGupta et al., 2005) and host-pathogen interactions (Brass et al., 2008; Liu et al., 2009; Tai et al., 2009; Zhou et al., 2008). Although there are now well-defined computational approaches for the various stages of the primary analysis of RNAI screens, including quality control (Zhang et al., 1999; Zhang, 2007b), normalization (Malouf et al., 2006; Wiles et al., 2008) and hit selection (Chung et al., 2008; König et al., 2007; Zhang, 2007a), the potential for methods that integrate interaction data has proven to be effective at identifying genes associated with other types of high-throughput datasets are still limited. In this study, we propose a new method for the contextual interpretation of RNAI screens that works by co-clustering together with interaction data.

2 METHODS
2.1 Data sources
We describe our approach by applying it to a genome-wide RNAI screen for host factors of the human pathogen, hepatitis C virus (HCV) (Tai et al., 2009). The screen provided ‘fold change’ data, which represents the ratio by which viral growth was inhibited or enhanced as a consequence of the knockdown of a particular gene, as well as P-values. We were able to map ~18 000 of the knockdowns to ENTREZ records. To serve as context for the analysis of the RNAI data, we assembled a network by collecting interactions defined in the STRING database (Jensen et al., 2009). Only those rated with at least a medium level of confidence (combined score>0.4) were included. As with before, all identifiers were mapped to ENTREZ records. Mapping information (Ensembl protein id to Entrez gene id) was retrieved from ENSEMBL BioMart. In situations where more than one STRING record could be mapped to a gene pair, the genes were considered to interact if at least one of the records fulfilled the minimum required combined score. This resulted in a network composed of 13 104 vertices (genes) and 330 523 edges (interactions).

2.2 Co-clustering
As mentioned earlier, the task of analyzing expression data in the context of an interaction network has often been framed as an optimization problem (Dittrich et al., 2008; Ideker et al., 2002; Nacu et al., 2007; Sohler et al., 2004). Roughly speaking, this means defining a scoring function on subnetworks, and then using a search algorithm to look for high scoring, connected regions. Although a similar approach is possible for RNAI data, such a formulation of the problem is probably not optimal. At the very least, it should not be the only one investigated. For one thing, simply searching for high scoring, connected subnetworks, where the score is based only on the individual data values of the genes included in a subgraph, would almost too often return regions that are, although connected, not functionally coherent. This is especially a problem when using interaction networks that have a small world property, which is the case for many biological networks (at least in an approximate sense). Although it is possible to ameliorate this problem in the case of expression data by including measures of coexpression in the scoring function (Guo et al., 2007; Ulitsky and Shamir, 2007), a similar concept unfortunately does not exist in RNAI datasets. As such, functional coherence needs to be enforced through some other means.

As an alternative (or complement) to viewing the joint interpretation task as an optimization problem, we propose to frame it as one of clustering. The general idea of our approach is to ‘co-cluster’ by simultaneously considering the two types of data, such that genes which are both near each other in the interaction network and at the same time showing strong links to the phenotype of interest (i.e. the RNAI data) tend to be clustered together.

We represented an interaction network as an undirected graph G=(V,E), where the set of vertices V correspond to genes, and the set of edges E to interactions. Given an edge eu, which connects two vertices u and v, a weight, which we view as a distance, was assigned to the edge in such a manner that it is smaller the stronger the data values of u and v are linked to the phenotype (i.e. the stronger the data values of u and v in the RNAI screen, the shorter the edge connecting them, if it exists). Specifically, if pu and pv are the P-values in the RNAI screen of the genes u and v, respectively, then we can view

\[ h_{uv} = (1 - pu)(1 - pv) \]  

(1)

as a measure of the ‘probability’ that both genes u and v are relevant to the phenotype of interest. Following Equation (1), we calculate the edge weight w(euv) as the probability of observing h(u,v) or higher. Formally,

\[ w(euv) = P(h_{uv} \leq (1 - X(1 - Y))) \]  

(2)

where X and Y are independent random variables with a standard uniform probability distribution. The right-hand side of Equation (2) can be expressed as

\[ \int_{0}^{1 - h_{uv}} (1 - X(1 - Y)) dX \]

which can then be simplified to

\[ 1 + h_{uv} \ln(h_{uv}) \]

(3)

Note that w(euv), which is visualized in Figure 1, can be viewed as a P-value for how high huv is. Roughly speaking, the more significant pu and pv are, the smaller w(euv) becomes.

Equation (3) was derived with the assumption that the P-values associated with the two vertices linked by an edge are independent. This, however, is not always true in actual biological data. Indeed, it is very reasonable to expect related genes, for example those that belong to the same protein complex, to induce similar phenotypes on the virus (if indeed they do induce a one). Nevertheless, this does not hurt the intended purpose of the edgeweight, which is simply to bring two genes closer to each other in the network if they both have strong data values (i.e. they both induce strong phenotypes).

With weights assigned to all the edges of the interaction network, we then construct the [V] × [V] distance matrix M, where mij is the length of the shortest path from vertex i to j. We use the well-known Floyd-Warshall algorithm to solve the all-pairs-shortest path problem. Using M as input, we then cluster the genes using a version of average linkage clustering.
Specifically, we calculated the score of a cluster as the numerator discovery rate (FDR) when that particular test is called significant, then the assigned to an edge as a function of the Edge weights. The surface corresponds to the weight (distance) Fig. 1. subnetwork score used by Ideker of signal in the different clusters. One possibility for this purpose is the this exploration, we use an additional scoring system to estimate the amount an impression as to which genes are both associated with a strong screening clustering procedure was conducted, the clusters along this path should give internal node to a cluster consisting of all the genes (leaves) that descend from it.

The dendrogram by itself could already be used for biological hypothesis generation. Starting from seed nodes, which could be genes of prior interest or simply genes with the strongest data values in the RNAi screen, one could gain insight into the mechanism of action of the seed gene by looking at the clusters that it belongs to. This can be done by exploring the path in the dendrogram which originates from the corresponding leaf, and then moving progressively upwards. Because of the manner in which the clustering procedure was conducted, the clusters along this path should give an impression as to which genes are both associated with a strong screening value and at the same time near the seed in the interaction network. To aid in this exploration, we use an additional scoring system to estimate the amount of signal in the different clusters. One possibility for this purpose is the subnetwork score used by Ideker et al. (2002) for expression data:

$$S_k = \frac{1}{A} \sum_{i=1}^{A} z_i$$

where A is the set of vertices in the subnetwork, k is the size of A and z_i is the z-score of vertex (gene) i. However, the problem with Equation (4) is that it is far too 'tolerant', such that search strategies using this scoring function often return huge subnetworks, even if there is actually no signal in the data (Dittrich et al., 2008; Rajagopalan and Agarwal, 2005). The reason for this is that even under the null hypothesis, where z_i are normally distributed, approximately half of them could still contribute to a positive score.

Because of the shortcomings of Equation (4), we defined an alternative scoring function, similar to that employed by Dittrich et al. (2008), which allows the control or estimation of the amount of signal in the clusters. Specifically, we calculated the score of a cluster C as

$$S_{q_0}(C) = \sum_{v \in C} \log \left( \frac{1 - q_v}{q_v} \right)$$

where q_v is the q-value of vertex v, and q_v(0,1] is a reference q-value parameter. Since the q-value of a test provides a measure of the false discovery rate (FDR) when that particular test is called significant, then the main numerator r_v = (1 - q_v)/q_v. Of Equation (5) can be roughly interpreted as a signal to noise ratio for vertex v. Inspired by a log-odds score, the final contribution of v to S_{q_0} was calculated by dividing r_v by the analogous ratio for the reference q-value, q_0, and then taking the logarithm. This results in a scoring system wherein only vertices associated with an FDR better than q_0 would be able to contribute positively. Consequently, even though clusters can contain negative scoring vertices, only those that include genes that actually have a chance of being significant can get a non-negative S_{q_0}.

We calculated q-values using the method of Storey (2002). In addition to exploring the dendogram by tracing paths from seed nodes (leaves), an alternative strategy for biological hypothesis generation is to directly extract 'active regions' from the network. We do this by scanning the whole tree, and then creating a super cluster consisting of all clusters that have a positive score given q_0 and larger than a minimum size. In this way, the parameter q_0 controls both the size of, and the amount of signal in, the resulting subnetwork. Note that if the RNAi data actually has no signal, then all genes would have a q-value of at least 1.0. This means that no cluster can have a positive score regardless of q_0, and accordingly an empty subnetwork would be returned (which is the desired behavior). Control of this parameter is further discussed in the Supplementary Material.

3 RESULTS AND DISCUSSION
HCV is a positive sense single-stranded RNA virus that in humans causes its namesake disease, hepatitis C (Senecal and Morelli, 2007). About 3% of the world’s population (270–300 million) is chronically infected with HCV. Of this number, approximately 21 000 knockdowns included in the study, we were able to map 17 821 to ENTREZ records. Of these, 13 104 participate in at least one interaction in the STRING-derived network. All subsequent analyses were limited to this subset.

One of the genes with the strongest RNAi data value is COPB1. It is a subunit of the COPI complex, which coats vesicles transporting proteins from the Golgi body back to the rough endoplasmic reticulum. The early parts of the clustering path originating from the COPB1 leaf going to the dendrogram root is visualized in Figure 2. In addition to COPB1, several other COPI subunits showed very strong RNAi data values. These were incrementally clustered together with the seed, culminating in cluster A (Fig. 2), which includes five of the seven COPI subunits and the genes ARF1 and ARCN1. Given that the latter two genes are also involved in vesicular trafficking, it is thus very likely that the strong phenotype induced by knockdown of COPB1 has indeed something to do with its role in the COPI complex. Note that although the subsequent additions to cluster A decreased the score (see clusters B and C), this does not mean that the added genes are likely to be noise, since a very stringent reference q-value was used (q_0 = 0.02). Indeed, all the genes in cluster C are associated with a q < 0.5. A comparison between using Equations (4) and (5) to score the clusters is made in Supplementary Figure S3.

In addition to manual exploration, the dendogram can also be used for biological hypothesis generation by directly extracting from the network regions with high activity and coherence. In the case of HCV, we do this by merging all positive-scoring clusters (for q_0 = 0.02) with at least four vertices. The resulting subnetwork
virus 1 (Crump et al., 2001), hepatitis B virus (Lambert et al., 2007) and herpes simplex virus 1 (Crump et al., 2007). It is not difficult to imagine that the same is also true for HCV. Indeed, a very recent study concluded this to be the case (Corless et al., 2010). In it, they showed that viral particle production is greatly reduced upon transfection of plasmids expressing various dominant negative forms of VPS4 and ESCRT-III components.

Edge weights, as calculated according to Equation (3), rely on the individual \(P\)-values of the adjacent vertices. However, it is not always possible to obtain such statistical measures of significance (i.e. \(P\)-values), particularly if the data are unusually distributed. Accordingly, we also developed a non-parametric version of our method that is able to deal with such cases. This is described and compared to the basic approach in the Supplementary Material.

To illustrate the difference between our method, which aims at greater functional coherency in the identified active subnetworks, and one that uses an optimization formulation, we also applied the algorithm of Dittrich et al. (2008) (as implemented in the BioNet R package) to the HCV data. The comparison is summarized in Figure 4. Note that for similarly sized subnetworks, the ones returned by our method (labelled as ‘COCLUST - ave-linkage’ in the figure) tend to have more connections between the vertices (i.e. the genes tend to have more interactions between them; see graph of Fig. 4A and C). However, this increase in functional coherency was achieved at the cost of an increase in the overall \(q\)-value level associated with the genes (Fig. 4B). Finer control of this trade-off is one of the directions that we intend to pursue in subsequent studies.

In addition to average-linkage, we also tried using a single-linkage criterion for the clustering aspect of our method when applied to the HCV data. This is because in contrast to average-linkage, where merge steps depend on the average distance between all pairs of genes from two clusters, single-linkage only considers the two genes that are closest to each other. That is, single-linkage, similar to BioNet (which only requires subnetworks to be connected), does not take group effects into account. In fact, for small network sizes (\(n < 100\)), single-linkage co-clustering even returned subnetworks with a better overall \(q\)-value level than the optimization approach (Fig. 4A). This latter behavior is due to the fact that, in contrast to BioNet, we do not require active regions to be fully connected; rather, we allow them to be composed of separate clusters or modules.

### 4 CONCLUSIONS AND FUTURE WORK

We presented a method for the contextual interpretation of RNAi-based functional screening data using interaction networks. Rather than posing the task as a maximal connected subgraph problem, as is often done with expression data, we framed it as one of clustering. In particular, we employed a co-clustering strategy where genes that are both near each other in the interaction network and at the same time show strong links to the phenotype of interest tend to be clustered together. The result of our method is a dendrogram that can be used for biological hypothesis generation in two ways: (i) by manually exploring the clustering from seed genes of a priori interest; or (ii) by directly extracting ‘active subnetworks’. We demonstrated our approach by applying it to a genome-wide screen for host factors of the human pathogen, hepatitis C virus. In addition to recovering known cellular modules relevant to HCV infection, our analysis...
allowed us to identify new potential candidates, and formulate biological hypotheses regarding possible roles and mechanisms for a number of them.

Our clustering approach enforces a level of functional coherency in the resulting subnetworks by considering ‘group effects’ at each merge step. As opposed to the more traditional optimization strategy usually employed for expression data, where subnetworks only need to be connected, the use of an average linkage score means that clusters would tend to comprise more of genes that are more closely connected to each other. In addition to promoting functional coherency, this also affords our method the possibility of highlighting groups of related genes that, although individually may only be moderately associated with the phenotype of interest, could collectively be more important than genes with more obvious data values. This feature is especially important for RNAi datasets since false negatives systematically arise from several properties of the experimental system. For example, even if a gene were genuinely linked to the phenotype of interest, knocking it down would not necessarily result in an observable effect if there is another gene that could also perform the respective function. Even if the alternative...
Fig. 4. Comparison with an optimization approach. We compared active regions found by our method (using both average- and single-linkage clustering) to those found by BauNet with respect to (A) average degree; (B) average q-value associated with the nodes; and (C) average clustering coefficient.

use of more sophisticated methods for extracting active regions from the dendrogram, the use of alternative clustering algorithms, and extending our method to be applicable to high-dimensional RNAi screens (i.e. screens that simultaneously monitor more than one phenotype). We will be applying the method to three new RNAi screens (i.e. screens that simultaneously monitor more than one phenotype). We will be applying the method to three new RNAi screens (i.e. screens that simultaneously monitor more than one phenotype).

Funding: German Research Foundation (Deutsche Forschungsgemeinschaft) through the International Research Training Group RECESS (IRTG 1563).

Conflict of Interest: none declared.

REFERENCES

Contextual analysis of RNAi-based screens