OrthoClust: An orthology-based network framework for clustering across multiple species

Koon-Kiu Yan[1,2,*], Daifeng Wang[1,2,*], Joel Rozowsky[1,2], Henry Zheng[2], Chao Cheng[4], Mark Gerstein[1,2,3]

1 Program in Computational Biology and Bioinformatics,

2 Department of Molecular Biophysics and Biochemistry, and

3 Department of Computer Science,

Yale University, Bass 432, 266 Whitney Avenue, New Haven, CT 06520

4 Department of Genetics, Dartmouth School of Medicine, Hanover NH 03755

Koon-Kiu Yan 1/23/14 1:23 PM
Deleted: Chao Cheng[4],

Koon-Kiu Yan 1/23/14 1:23 PM
Deleted: M

Abstract

Increasingly, high-dimensional genomics data is being collected in many organisms. Here, we develop OrthoClust for simultaneously clustering data across multiple species. OrthoClust is a general computational framework that integrates the co-association networks of individual species by utilizing the orthology relationships of genes between species. It outputs optimized modules that are fundamentally cross-species, which can either be conserved or species-specific. We demonstrated the application of OrthoClust using the RNA-Seq expression profiles of *C. elegans* and *D. melanogaster* from the modENCODE consortium. A potential application of cross-species modules is to infer putative analogous functions of uncharacterized elements like ncRNAs based on "guilt-by-association".

Koon-Kiu Yan 1/24/14 10:44 AM
Deleted: transcriptome

Koon-Kiu Yan 1/24/14 10:44 AM
Deleted: expression profiles

Koon-Kiu Yan 1/24/14 10:44 AM
Deleted: co-expression

Koon-Kiu Yan 1/24/14 10:48 AM
Deleted: ,

Koon-Kiu Yan 1/24/14 10:48 AM
Deleted: thus

Koon-Kiu Yan 1/24/14 10:49 AM
Deleted: ting

Koon-Kiu Yan 1/24/14 10:45 AM
Deleted: globally

Koon-Kiu Yan 1/24/14 10:48 AM
Deleted: modules

Koon-Kiu Yan 1/24/14 10:48 AM
Deleted: . The modules

Koon-Kiu Yan 1/24/14 10:55 AM
Deleted: and

Koon-Kiu Yan 1/24/14 10:52 AM
Deleted: applied

Koon-Kiu Yan 1/24/14 10:52 AM
Deleted: to

Koon-Kiu Yan 1/24/14 10:53 AM
Deleted: using RNA-Seq data

Koon-Kiu Yan 1/24/14 11:00 AM
Deleted: The resulting modules were used for inferring putative functions of ncRNAs based on "guilt-by-association"; worm and fly ncRNAs anchored to the same module may potentially have analogous functions.

Koon-Kiu Yan 1/24/14 11:00 AM
Formatted: Font color: Gray-80%

Keywords

clustering algorithm/ co-expression network/ orthology/ transcriptome profiling/ data integration

Background

Over the last decade, we have witnessed the burgeoning of comparative genomics. With the advancement of sequencing and other high-throughput techniques, 'omics'-scale data has been generated in many species [1][2]. Apart from genomic sequences, one can now compare two or more species in terms of their epigenome, regulome, transcriptome or interactome etc. As a result, computational frameworks that integrate such system-level data from different species are of particular interest. While different kinds of 'omics' scale data reflect different facets of a biological system, many of these high-dimensional data can be projected onto a network. For instance, the expression profiles of genes or the histone modification patterns in their upstream regions can be used to connect genes to form various association networks. Data from different species thus form species-specific networks that can in principle be integrated by incorporating evolutionary relationship.

For a set of genes, features associated with the topological properties of networks open additional windows to interpret their genomics features and annotation. Among which the concept of network modules is particularly important from a systems biology perspective. Through identifying modules, one can reduce the complexity of biological systems by collapsing the vast number of interconnections amongst its constituents into a smaller number of interactions between the modules [3][4]. While different 'omic' data arrive at different networks, genes clustered together to form modules are likely to have a common biological role; for instance, being regulated by a common transcription factor, being part of a protein complex, or being presented in a same pathway. One of the most widely studied 'omic'-scale data is genome-wide expression data. To analyze genome-wide expression profiles, network-based algorithms [5] together with approaches like hierarchical clustering [6], self-organized maps [7], spectral techniques [8] and superparamagnetic clustering [9] have been developed and extensively used since the dawn of the microarray era. While these methods have provided valuable biological insights, they were aimed at clustering within individual species only. A natural generalization that performs clustering across multiple species will therefore be instructive, especially because the recent advancements of transcriptome profiling techniques like RNA-Seq have generated

Koon-Kiu Yan 1/21/14 11:00 AM
Deleted: Indeed, genes with similar expression patterns across a variety of conditions are likely to have a common biological role; for instance, being regulated by a common transcription factor, being part of a protein complex, or being presented in a same pathway. By identifying expression modules, one can reduce the complexity of biological systems by collapsing the vast number of interconnections amongst its constituents into a smaller number of interactions between the modules [6][7].

Koon-Kiu Yan 1/21/14 11:16 AM
Deleted: clustering algorithms have been extensively used since the dawn of the microarray era [5]. Different types of algorithms have been applied or developed for clustering expression profiles, including

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [5]

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [8]

Unknown
Field Code Changed

Unknown
Field Code Changed

Koon-Kiu Yan 1/21/14 11:19 AM
Deleted: approaches

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [9]

Koon-Kiu Yan 1/21/14 11:16 AM
Deleted: ,

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [10]

Koon-Kiu Yan 1/21/14 11:17 AM
Deleted: and various network-based algorithms [11].

Koon-Kiu Yan 1/21/14 11:24 AM
Deleted: Recently,

Koon-Kiu Yan 1/21/14 11:24 AM
Deleted: such as

tremendous amounts of genome-wide expression data across many different species [10][11].

Here we present OrthoClust, a novel network based framework for clustering data across multiple species. OrthoClust integrates the networks of individual species using orthology relationships of genes between species. As connected genes within a species and orthologous pairs across species connect genes with the same function within and across species respectively, OrthoClust naturally extends the idea of functional modules into a cross-species dimension. The essence of OrthoClust is a cost function that detects modules across species. We present a solution of the optimization problem by using simulated annealing. As expression data is one of the most important class of 'omic-'data, we demonstrated OrthoClust using the genome-wide expression data of worm and fly generated by the modENCODE consortium, arriving at co-expression modules that range being highly conserved to species-specific. We then studied the results with traditional single-species clustering and network alignment. As more and more system-wide data is generated across different species, the concept of orthology-based meta clustering demonstrated by OrthoClust can serve as a general computational framework for integration of other 'omic'-scale data like protein-protein interactions.

Results

Cross-species Modules in a multiplex network

A co-association network is a network representation of certain genomics data. The data can be rather simple like protein binding profiles, in which two genes are connected if their corresponding proteins can physically interact. In many cases, it can be high dimensional such as genome-wide expression profiles. In this scenario, two genes are linked in a mathematically abstract way if their expression values across a variety of conditions are highly correlated. Despite the origin of the network, from a topological standpoint, a module is an interconnected region in which the density of edges is higher than the average density of the whole graph. Constitutes of a module are presumably genes working in a coordinated fashion, i.e. have a specific function. We combined the co-association networks from different species to form a network with two types of edges representing two types of functional similarities. Mathematically this structure is a multiplex

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [3]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [4]

Koon-Kiu Yan 1/21/14 10:24 AM
Deleted: Nevertheless, the expression profiles of different species can in principle be integrated by incorporating evolutionary relationship between species.

Koon-Kiu Yan 1/21/14 11:13 AM
Formatted: Font color: Auto

Unknown
Field Code Changed

Koon-Kiu Yan 1/21/14 11:25 AM
Deleted: co-expression pairs

Koon-Kiu Yan 1/21/14 11:32 AM
Deleted: a

Koon-Kiu Yan 1/21/14 11:39 AM
Deleted: spin system, in which machinery developed in statistical physics is used to solve the module detection problem - a global optimization problem with a rugged landscape.

Koon-Kiu Yan 1/21/14 11:28 AM
Deleted: As a demonstration, we applied

Koon-Kiu Yan 1/21/14 11:31 AM
Deleted: We further used the constructed modules to infer putative functions of worm and fly ncRNAs. A worm ncRNA and a fly ncRNA mapped to the same cluster are potential candidates for performing analogous functions in different species.

Koon-Kiu Yan 1/21/14 11:40 AM
Deleted: further

Koon-Kiu Yan 1/21/14 2:07 PM
Deleted: co-expression

Koon-Kiu Yan 1/21/14 2:09 PM
Deleted: expression profiles. Two

Koon-Kiu Yan 1/21/14 2:18 PM
Deleted: connected

Koon-Kiu Yan 1/21/14 12:02 PM
Deleted: co-expression

Koon-Kiu Yan 1/21/14 2:20 PM
Deleted: of

5

network [12]. Genes in a species are connected if they are co-associated, whereas genes from different species are connected if they are orthologs. Figure 1 shows a simple example of such a multiplex network. We extended the concept of modules used in co-association networks of individual species in a novel cross-species fashion. Here a module may comprise genes from multiple species, characterized by the two types of functional similarity in a cross-species manner. Within a module, from a molecular viewpoint, genes from the same species most likely share the same function as they are co-associated, co-expressed or physically binding together. Orthologs across different species, from an evolutionary standpoint, also might have similar biological function. Intuitively, a module should consist of nodes that form clique-like structures within a co-association network, as well as nodes that are linked by orthology relationships between layers of co-association networks. Nevertheless, as shown in Figure 1, it is entirely possible that a module in the multiplex network consists of genes from a single species. In fact, this is the case when a novel function emerges for a particular species and the genes corresponding to the specific function do not have corresponding orthologs.

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [18]

Koon-Kiu Yan 1/21/14 2:25 PM
Deleted: expressed

Koon-Kiu Yan 1/21/14 2:25 PM
Deleted: expression

Koon-Kiu Yan 1/21/14 2:27 PM
Deleted:

Koon-Kiu Yan 1/21/14 2:27 PM
Deleted: co-expression

Koon-Kiu Yan 1/21/14 2:27 PM
Deleted: expression

Overview of OrthoClust

Figure 2 shows the three major steps of OrthoClust: construction of the multiplex network, defining the cost function of the system and assigning nodes to modules by multiple runs of simulated annealing.

Construction of the multiplex network The inputs of OrthoClust are the co-association networks of two or more species, and the orthology relationships between genes of the species of interest. Of course, co-association networks are results of raw data, and there are various ways to arrive at the networks depending on the specific data and biological purposes. OrthoClust combines individual layers of co-association networks by connecting genes in different species via their orthology relationships. To account for the fact that many orthologous pairs are not one-to-one but many-to-many, orthologous links are weighted such that the weights are normalized by the number of orthologs of each node (see Materials and Methods).

Koon-Kiu Yan 1/21/14 2:35 PM
Deleted: genome-wide expression profiles

Koon-Kiu Yan 1/21/14 2:40 PM
Deleted: co-expression

Defining the energy function in the multiplex network OrthoClust defines a cost function in order to detect modules in a multiplex network. Specifically, every node can take a discrete label σ ranging from 1 to q. Nodes with the same label will be assigned to the same module. q is therefore a parameter chosen to be the maximum number of modules allowed in the system. If the network has N nodes, there will be N^q ways (configurations) to assign nodes to modules and each configuration is characterized by a cost function H defined as

$$H = -\left(\sum_{i,j \in S_1} \Lambda^1_{ij} \delta_{\sigma_i \sigma_j} + \sum_{i,j \in S_2} \Lambda^2_{ij} \delta_{\sigma_i \sigma_j} + \kappa \sum_{(i,j') \in O(S_1,S_2)} w_{ij'} \delta_{\sigma_i \sigma_{j'}} \right).$$

Here, S_1 and S_2 are the sets of genes for the two species respectively. $\Lambda_{ij} = A_{ij} - k_i k_j / 2m$, with $k_i = \sum_j A_{ij}$ and $m = \frac{1}{2} \sum_i k_i$. As A is a network adjacency matrix, the subtracted term is the expected number of links between nodes *i* and *j* in an ensemble of random graphs with the same degree distribution [13][14]. Its presence in *H* is to reduce the contribution of links between nodes with higher degree (i.e. hubs). The superscripts (1 or 2) correspond to the networks of two species. The value of the Kronecker delta $\delta_{\sigma_i \sigma_j}$ equals one if nodes *i* and *j* have the same label and zero otherwise. The cost function *H* is a generalization of existing network modularity functions [15][16]. In the standard modularity function, a network with high modularity means there is a high number of links between nodes in the same module, and low number of links between nodes in different modules. The novelty of OrthoClust is the last term regarding the orthologous links between nodes in different layers of the co-association networks. It sums over $O(S_1, S_2)$, i.e. all the orthologous pairs between S_1 and S_2. As mentioned above, each pair of orthologs is weighted by w_{ij} to take into account of the many-to-many orthology relationships (see Materials and methods). Configurations in which orthologs having the same label will lower the cost function. The relative contribution between co-association edges and orthologous edges is controlled by a coupling constant κ (for determination of the constant, see below). In the language of statistical physics, the entire framework can be interpreted as a spin system called a q-state Potts model [17], which is a generalization of the Ising model. The cost function

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [26]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [27]

Unknown
Field Code Changed

Koon-Kiu Yan 1/21/14 3:33 PM
Deleted: have lower energy

Koon-Kiu Yan 1/21/14 3:34 PM
Deleted: our

Koon-Kiu Yan 1/21/14 3:34 PM
Deleted: is essentially

Koon-Kiu Yan 1/21/14 3:34 PM
Deleted: a

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [28]

characterizes the energy of the spin (label) system and the optimal assignment of nodes to different modules is equivalent to the ground state of the Potts model.

Assigning nodes to modules by multiple runs of simulated annealing To optimize the cost function, OrthoClust employs a standard simulated annealing procedure similar to one used in [18]. Labels are randomly assigned initially, and updated via a heat bath algorithm. The temperature of the system is gradually lowered until the flipping rate of labels is lower than a certain threshold (see Figure S1, Materials and Methods). Although the labels have divided nodes of the network into modules, we do not directly use the resultant configuration due to the probabilistic nature of simulated annealing, but perform the annealing process T times. By summarizing the results using a co-appearance matrix (a matrix whose elements *(i,j)* represents how often the two nodes *i* and *j* co-appear in the same module), OrthoClust arrives at a set of modules by thresholding the co-appearance frequency and looking for nodes that co-appear often (see Materials and methods). The sizes of the modules follow a power law distribution; tiny modules are therefore neglected (see Materials and Methods). OrthoClust is in general not very sensitive to the value of *q*. This is because, even though the system starts with many different labels (a high value of *q*), the large range of states will coalesce into a few modules and only a few labels will remain to cover the appropriate number of modules as the system cools down. In other words, the exact value of *q* is not very important as long as *q* is chosen to be large.

often

Using OrthoClust for integrating expression profiles across species

A particular application for OrthoClust is to cluster expression profiles across species. Since OrthoClust is a network framework, raw expression profiles should be transformed into individual co-expression networks. Many algorithms have been proposed for this purpose based on calculating the N by N Pearson correlation matrix [19][20][21][22]. For our application, we found that a rank-based algorithm in which each gene is connected to the top *d* genes with the highest (absolute) Pearson correlation works best for resolving modules [19] (see Materials and Methods). It is well known that co-expression networks in many different species are modular,

Koon-Kiu Yan 1/21/14 3:36 PM
Deleted: The

Koon-Kiu Yan 1/21/14 3:36 PM
Deleted: finding

Koon-Kiu Yan 1/21/14 3:31 PM
Formatted: Font color: Gray-80%

Koon-Kiu Yan 1/21/14 3:37 PM
Deleted: determine the ground state,

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [23]

Koon-Kiu Yan 1/21/14 3:41 PM
Deleted: Spin values

Koon-Kiu Yan 1/21/14 3:41 PM
Deleted: spin states

Koon-Kiu Yan 1/21/14 3:41 PM
Deleted: spin values

Koon-Kiu Yan 1/21/14 3:42 PM
Deleted: spin

Koon-Kiu Yan 1/24/14 4:23 PM
Deleted: multiple

Koon-Kiu Yan 1/21/14 4:23 PM
Deleted:

meaning that a subset of genes (a module) have a specific function [5][23][24][25], therefore it is interesting to explore how these modules emerge in a cross species fashion. Like various co-association networks constructed by correlating high-dimensional data, a co-expression edge can have either a positive (+1) or a negative sign (-1) is assigned based on the sign of the correlation coefficient between two genes. Since anti-correlated genes do not work together, it is instructive to separate them into two different modules. This can be achieved by modifying the original cost function to separate the sets of positive and negative links in each species as specified by the superscripts (+ or -), i.e.,

$$H = -\left(\sum_{i,j\in S_1} \Lambda_{ij}^{1+}\delta_{\sigma_i\sigma_j} - \sum_{i,j\in S_1} \Lambda_{ij}^{1-}\delta_{\sigma_i\sigma_j} + \sum_{i,j\in S_2} \Lambda_{ij}^{2+}\delta_{\sigma_i\sigma_j} - \sum_{i,j\in S_2} \Lambda_{ij}^{2-}\delta_{\sigma_i\sigma_j} + \kappa \sum_{(i,j')\in O(S_1,S_2)} w_{ij'}\delta_{\sigma_i\sigma_{j'}} \right).$$

The minus sign in front of the negative links means the effects of the negative links are opposite to the positive links, meaning that in the favorable configurations, nodes in a same module are likely to be connected by positive links, nodes from different modules tend to be connected by negative links.

Simultaneous clustering of expression profiles in worm and fly via OrthoClust

As a demonstration, we applied OrthoClust to the transcriptomes of worm and fly generated by the modENCODE consortium [26]. In this analysis, the initial number of spin states q was chosen to be 250. We summarized the results of T_r (T=32, see below) annealing runs using a N-by-N co-appearance matrix, where N is the size of multiplex network (the total number of genes in worm (20377) plus fly (13623) in this case). As shown in Figure 3A, there are blocks of worm and fly genes along the diagonal. These blocks consist of genes that co-appear often in various runs of annealing, representing different worm and fly modules. Of particular interests are the blocks of worm and fly genes that co-appear with high frequency in the off-diagonal positions. For instance, as highlighted in Figure 3A, a block of worm genes and a block of fly genes form a conserved module. As expected, they share a significant fraction of GO terms ($P=3.3 \times 10^{-16}$, hypergeometric test). Figure 3B shows the common GO terms between a set of worm genes and a set of fly genes in the conserved module. Most of the common GO terms refer to fundamental

Koon-Kiu Yan 1/21/14 4:30 PM
Deleted: work in a coordinated fashion, i.e.

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [11]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [15]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [16]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [17]

Koon-Kiu Yan 1/21/14 4:26 PM
Deleted: The expression profiles of individual species are then transformed into individual co-expression networks. In general, given the expression profiles of N genes within a species, this can be done by calculating the N by N Pearson correlation matrix [19][20][21][22]. Here, OrthoClust employs a rank-based algorithm in which each gene is connected to the top *d* genes with the highest (absolute) Pearson correlation [19] (see Materials and Methods).

Koon-Kiu Yan 1/21/14 4:11 PM
Deleted: Under our framework, the co-expression networks are not weighted, but OrthoClust allows an edge to have e ... [1]

Koon-Kiu Yan 1/21/14 4:03 PM
Deleted: The reason for this distinc ... [2]

Koon-Kiu Yan 1/21/14 4:04 PM
Deleted: and therefore should belo ... [3]

Koon-Kiu Yan 1/21/14 4:32 PM
Deleted: .

Koon-Kiu Yan 1/21/14 4:06 PM
Deleted: In OrthoClust, *H* takes int ... [4]

Koon-Kiu Yan 1/21/14 4:49 PM
Deleted: .

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [29]

Koon-Kiu Yan 1/24/14 4:24 PM
Deleted: 32

biological functions like RNA processing and cell cycles processes. On the other hand, blocks that do not overlap in the off-diagonal positions correspond to specific worm and fly modules. For instance, GO terms related to chitin activities were found in certain fly specific modules. At a global level, we found that the size of modules follows a power-law distribution with an exponent of -1.89 (see Figure S2). The power-law distribution observed includes certain large modules. Practically, one could implement extra steps to break down these large modules recursively.

Separation of Modules in the GO space

As OrthoClust divides genes into modules based on how they are separated topologically in the multiplex network, it is instructive to examine systematically how these modules are separated in functional space as defined by GO terms. To do so, we constructed a metric to quantify the similarity between all worm and fly genes (both intra-species and inter-species) based on the overlap of GO terms. More specifically, we represented the relationship where gene *i* is annotated with GO term *j* by an adjacency matrix B, and further defined a matrix G such that $G_{ij} = B_{ij} \log \frac{n}{\sum_i B_{ij}}$. In matrix G, the contribution of a GO term *j* to a gene is weighted by its inverse document frequency, a quantity commonly used in text-mining [27]. High-level GO terms present in many genes will be weighted down [28]. The similarity score between two genes k_1 and k_2 was defined as the cosine of the two corresponding vectors (the k_1^{th} row and the k_2^{th} row in the G matrix). As shown in Figure 4, for the 150 modules obtained by clustering all worm and fly genes, the overlap between genes within a module is much higher than the overlap between genes across modules ($P=3\times10^{-83}$, Wilcoxon test). Nevertheless, since two orthologous genes tend to have very similar GO terms, we further investigated whether such a high level of overlap between genes within a module is merely the consequence of orthology. We therefore repeated the analysis by removing all orthologs inside the modules. We found that the overlap between genes within modules is still significantly higher than the overlap across modules ($P=1.5\times10^{-45}$, Wilcoxon test, see Figure S3). Thus, we conclude that in terms of GO annotation, OrthoClust has separated genes with different functions into different modules.

DISC.

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [30]

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [31]

Unknown
Field Code Changed

Benchmarking modules based on co-regulation patterns

Apart from GO analysis in Figure 4, we further tested whether genes inside a module are indeed more functionally relevant by examining the number of common regulators they possess. We identified the binding targets of a set of worm and fly transcription factors based on ChIP-Seq experiments generated by modENCODE consortium (see Materials and methods). These ChIP-Seq experiments and the RNA-Seq experiments for expression profiles were performed under similar developmental stages. We then counted for all pairwise combinations of modules, the number of common transcription factors for each pair of genes (see Figure S4). We found that pairs of genes within a module, in average, have more common transcription factors than pairs of genes in different modules (a 2.6 fold increase in worm and 1.6 fold increase in fly). This analysis is consistent with general observations that a transcription factor tends to regulate targets sharing similar biological functions.

Effects and the determination of the coupling constant κ

The cost function of OrthoClust takes into account two types of edges: co-association edges and orthology relationships. The coupling constant κ determines the relative contribution of the co-association links within a species and the orthologous links across species. A low value of κ means networks are likely to be clustered independently whereas a high value of κ means orthology links are more important and the label of a node tends to be aligned with its ortholog rather than its neighbors in the same network. In the clustering of gene expression profiles, we employed two independent methods to quantify the effects of tuning κ and determine its optimal value. First of all, we made use of a set of 1288 metagenes obtained from [23] as our gold standard. These metagenes were constructed based on orthologs whose expression relationships are conserved across multiple species including worm, fly and human. A metagene consists of genes from different genomes (worm and fly in this case) that presumably share the same function by considering their expression values across different conditions. Unlike our clustering approach, which is based on the optimization of a global cost function, metagenes were inferred by examining the likelihood that individual co-expression edges are transferred from

Koon-Kiu Yan 1/21/14 9:59 PM
Deleted: spin

Koon-Kiu Yan 1/21/14 9:59 PM
Deleted: value

Koon-Kiu Yan 1/21/14 10:00 PM
Deleted: W

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:29 PM
Deleted: [15]

Koon-Kiu Yan 1/21/14 10:16 PM
Deleted: energy

one species to another in a local manner. This complementarity makes the set of metagenes a good gold standard for validation. Following our clustering framework, the constituents of a metagene should appear in the same module. As shown in Figure 5A, for a low value of κ, clustering was performed independently and it was rare that the worm and fly components of a metagene fall into the same module. Nevertheless, both the worm and fly networks have high modularity, meaning the two networks were independently well separated into modules. On the other hand, for a high value of κ, most of the metagenes satisfied the criterion whereas the resultant modularity of individual networks became low. The value of κ can be tuned so as to balance this tradeoff.

Another method we used to examine the effects of κ is the similarity measure between genes based on their GO annotation as described in the previous section. The similarity scores between each pair of the 34,000 worm and fly genes define a weighted network W, where the nodes are the genes and the edges are weighted by the pair-wise scores. Since the weighted network is not a multiplex network but a single-layered network, its modularity can be quantified by a more traditional modularity function for weighted network defined as $\sum_{ij}\left(W_{ij}-\frac{k_i k_j}{2m}\right)\delta_{\sigma_i\sigma_j}$, where $k_i=\sum_j W_{ij}$, $m=\sum_i \frac{k_i}{2}$ and σ_i is the module label of node *i* [29]. A high modularity score means highly similar genes (in terms of GO annotation) are grouped in a module whereas distant genes are separated. In principle, this weighted network based on GO annotation serves as a benchmark for the multiplex network defined by OrthoClust. A favorable way of assigning nodes to modules by OrthoClust therefore should also be a favorable way to divide the weighted network into modules. As shown in Figure 5B, for each value of κ, we found the way to assign nodes to the modules by OrthoClust and then calculated the corresponding modularity score of the weighted network. When the value of κ is too high or too low, the modularity score of the weighted network is low. The κ that maximizes the modularity score of the weighted network should therefore be optimal κ for OrthoClust. Combing Figure 5A and Figure 5B, we picked κ=3 as our optimal value.

Weight associated with the orthology relationships

COULD WE USE CAPLET

Koon-Kiu Yan 1/21/14 10:47 PM
Deleted: We defined

Koon-Kiu Yan 1/21/14 10:47 PM
Deleted: in which the nodes are 34,000 genes in worm and fly, and the edges are the similarity score between each pair of gene.

Koon-Kiu Yan 1/21/14 10:58 PM
Deleted: good

Koon-Kiu Yan 1/21/14 10:56 PM
Deleted: should effectively

Koon-Kiu Yan 1/21/14 11:01 PM
Deleted: such that highly similar genes are grouped in a module, and distant genes are separated. This could effectively be quantified by the modularity function of the weighted network, which is defined as, where $k_i=\sum_j W_{ij}$, $m=\sum_i \frac{k_i}{2}$ and σ_i is the module label of node *i* [32].

Koon-Kiu Yan 1/21/14 11:02 PM
Deleted: we

Koon-Kiu Yan 1/21/14 11:02 PM
Deleted:

Koon-Kiu Yan 1/21/14 11:02 PM
Deleted: corresponding module assignment

Orthology relationships between species connect layers of co-association networks. While the coupling parameter κ defines the overall relative contribution between intra-species and inter-species connections, the weight of each orthology connection could in principle be adjusted. It is very common that in eukaryotes many orthologs are many-to-many instead of one-to-one, mathematically forming various bipartite cliques in the multiplex network. We tested OrthoClust by treating all the orthologous pairs equally in the cost function. We found that most of the cliques cannot be resolved, and their nodes got assigned to a single module (see Figure S5). It means the cost function is in favor of the huge cliques and bias against the conserved clusters that are linked by one-to-one orthologs. To account for this effect, OrthoClust therefore weights down the orthologous link of a node by the number of orthologs it possesses. As shown in Figure S5, the weighted approach works better in resolving the huge cliques.

Comparison with single-species clustering

The aim of OrthoClust is to perform clustering across multiple species in an integrated fashion. Naively, one could construct a cross-species module by perform clustering on individual species separately and concatenate the modules of different species by the orthologs they share. To examine this alternative approach, we performed single species clustering on the expression profiles of worm and fly separately using various standard methods (see Materials and methods). We then tested for each combination of worm and fly modules, whether or not there is an enrichment of orthologs between them based on a simple hypergeometric test (see Materials and methods). We found that even though there are certain combined worm-fly modules with significant enrichment of orthologous gene pairs, the enrichment is lower than the cross-species constructed by OrthoClust (Figure 6). This is of course not surprising because OrthoClust takes into account of the orthology relationships in the algorithm. Nevertheless, the analysis suggested that by using merely the co-expression data, it is in general less effective in finding the corresponding sets of genes in two species responsible for the same function. To show the result is not a consequence of the particular mathematical form of the cost function imposed by OrthoClust, we ran OrthoClust with κ=0. As there was no coupling between two species in the

cost function, the resultant sets of worm fly modules were essentially clustered independently. Again, we found that the combined worm-fly modules have lower enrichment of orthologous pairs compared to the case with optimal κ=3. Interestingly, this analysis also manifests how the coupling term in the cost function bring two sets of independent modules closer together in terms of the sharing of orthologs.

Comparison with network alignment

The concatenation of networks using orthology relationships resembles the problem of cross-species network alignment [30][31]. To compare OrthoClust with network alignment, we applied IsoRank [31] to align the worm and fly co-expression networks (see Materials and methods). Again, using the metagenes obtained from [23] as gold-standard, we found that 88% of metagenes were aligned by IsoRank (Figure S6), as compared to 81% by OrthoClust. Although IsoRank slightly outperformed OrthoClust in identifying the corresponding functional genes between two species, it does not immediately report how these pairs form clusters. Motivated by ref. [32], we looked for co-expression edges in two networks whose nodes are aligned by IsoRank. By connecting such edges in the network, we generated aligned subgraphs that could potentially be interpreted as modules conserved across two species. Among the gene-pairs that are predicted to be in the same module, we found that 43% are consistent with OrthoClust. The percentage is rather low probably because aligned subgraphs do not really possess the properties of clusters signified by the dense connections between genes within a species.

Robustness analysis

Simulated annealing was employed to optimize the cost function defined by OrthoClust. To reduce the effects of the stochastic nature of simulated annealing, we constructed the co-appearance matrix by repeating the annealing process R times. To determine R, we ran independent trials of R runs, resulting at different co-appearance matrices and thus different sets of clusters. We then compared the consistency between two sets of clusters by considering if two genes have been assigned to the same module by trial 1, whether or not they are assigned to a

common module in trial 2. This is essentially done by calculating a confusion matrix (see Materials and methods). By pairwise comparison of independent trials, we found that the overlap between trials increases as R increases (see Figure S7). More specifically, the overlap increases from 46% for R=8 to 65% for R=32, and 75% for R=64. Statistical significant results shown for the R=32 in the previous analysis showed the value offers a reasonable compromise between computational cost and robustness. We then further superposed different trials to construct a co-appearance matrix with 128 runs, and thus a set of "most accurate" clusters. We then calculated the consistency between the ultimate set with sets constructed with smaller values of R (see Figure S7). We found that the average consistency between clusters generated with R=32 and the ultimate set is 76%.

Mapping uncharacterized elements to modules

Apart from understanding the modular nature of biological processes, clustering expression profiles is very useful for inferring the putative functions of uncharacterized proteins [33] as well as ncRNAs [34][35]. The essence of this approach is "guilt by association": if an uncharacterized element is highly co-expressed with a core set of genes, one can infer the function of the gene based on the functions of genes within the core set. Nevertheless, most core sets were constructed by single-species clustering. The cross-species modules constructed by OrthoClust can potentially serve as anchor to relate uncharacterized but analogous elements from different species. To explore this avenue, we constructed modules using a set of core worm and fly genes (worm-fly orthologs) by OrthoClust (see Materials and Methods, Figure S8), arriving at a set of 21 core worm-fly modules with similar proportions of worm and fly genes (see Supplementary Dataset 1). We further investigated the functions of these modules based on their enriched GO terms (see Materials and Methods). For each module, by clustering the enriched GO terms, we assigned a list of representative keywords as their characteristic functions (see Figure 7). For instance, module 1 is signified by neurological system-process and module 2 by cellular-lipid-metabolism. As expected, many genes in these modules have orthologous partners within the module. In 18 out of the 21 modules, the fraction of genes with orthologous partners is higher

Koon-Kiu Yan 1/22/14 8:50 PM
Deleted: An important application for constructing modules of protein-coding genes is to infer putative functions of uncharacterized proteins as well as non-coding elements. To do so, we

Koon-Kiu Yan 1/22/14 4:46 PM
Deleted: 4

Koon-Kiu Yan 1/22/14 9:04 PM
Deleted: Table 1

Koon-Kiu Yan 1/22/14 10:02 PM
Deleted: 6

than 80%.

We then mapped worm and fly ncRNAs to the 21 modules based on their expression profiles (see Materials and methods). Though there is no gold standard available to evaluate systematically the performance of the mapping, we found examples suggesting that ncRNAs from different species could be linked together in terms of their potential functions. For instance sphinx, the fly lncRNA expressed in the brain, was shown to be involved in regulation of male courtship behavior [36]. In our analysis, this lncRNA was mapped to module 1 which is characterized by neurological system process and behavior. On the other hand linc-10 and linc-104, worm lincRNAs that are highly expressed in male-related stage [37], were mapped to the same module. In addition to the mapping, we also found that some modules are enriched with different classes of ncRNAs (see Figure S9). The list of worm and fly ncRNAs we tested and the modules they mapped to can be found in Supplementary Data 2.

Generalization to N species

OrthoClust is a general framework not only applicable to the clustering of expression profiles but in general other genomics data in the form of co-association network. In addition, the framework can be readily applied to more than two species by modifying the cost function. In general, for N species, the cost function will have N terms for the co-association networks, and N(N-1)/2 terms for the orthologs between all pairs of species. For instance, if N=3, the cost function can be written as

$$H = -\left(\sum_{k=1,2,3} \sum_{i,j \in S_k} \Lambda_{ij}^{k} \delta_{\sigma_i \sigma_j} + \sum_{(i,j') \in O(S_1,S_2)} \delta_{\sigma_i \sigma_{j'}} + \sum_{(i,j') \in O(S_1,S_3)} \delta_{\sigma_i \sigma_{j'}} + \sum_{(i,j') \in O(S_2,S_3)} \delta_{\sigma_i \sigma_{j'}} \right).$$

Here, S_1, S_2, S_3 stand for three different species. The inner summation is the modularity function for the network of a single species. The outer summation sums the three networks together. The extra terms represent the coupling (with coupling constant κ) between 3 pairwise combination, namely the orthologous pairs represented by $O(S_1,S_2)$, $O(S_2,S_3)$ and $O(S_1,S_3)$. The coupling constant κ can be determined by the same approach we explained in the example of two species. Of particular interest are the 1-1-1 triplets that consist of three genes from three species which

Koon-Kiu Yan 1/22/14 9:16 PM
Deleted: demonstrated the procedures of inference by focusing on ncRNAs. Unlike previous works in which ncRNAs were mapped to modules of individual species [13], we

Koon-Kiu Yan 1/22/14 9:16 PM
Deleted:

Koon-Kiu Yan 1/22/14 9:18 PM
Moved down [2]: By mapping ncRNAs from different species to such modules, we could naturally link ncRNAs from different species together in terms of their functions.

Koon-Kiu Yan 1/22/14 9:18 PM
Moved (insertion) [2]

Koon-Kiu Yan 1/22/14 9:17 PM
Deleted: .

Koon-Kiu Yan 1/22/14 9:21 PM
Deleted: .By mapping ncRNAs from different species to such modules, we could naturally link ncRNAs from different species together in terms of their functions. Table 1 summarizes the number of each class of ncRNAs (lncRNA, miRNA, tRNA, and snoRNA) in worm and fly that were mapped to the 21 conserved modules. More specifically, we mapped 852 lncRNAs (157 in worm, 695 in fly), 120 miRNAs (38 in worm, 82 in fly), 58 tRNAs (31 in worm, 27 in fly), and 187 snoRNAs (29 in worm, 158 in fly) to the modules. The list of worm and fly ncRNAs we used and the modules they belo... [5]

Koon-Kiu Yan 1/22/14 9:17 PM
Moved down [1]: The list of worm and fly ncRNAs we used and the module... [6]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:36 PM
Deleted: [35]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:36 PM
Deleted: [36]

Koon-Kiu Yan 1/22/14 9:58 PM
Deleted: We further investigated whether different classes of non-coding elem... [7]

Koon-Kiu Yan 1/22/14 9:17 PM
Moved (insertion) [1]

Koon-Kiu Yan 1/22/14 9:22 PM
Deleted: used

Koon-Kiu Yan 1/22/14 9:22 PM
Deleted: belong

are orthologous to one another in an one-to-one fashion, i.e. for triplet (g^{S1}, g^{S2}, g^{S3}), apart from g^{S1}, there exists no other gene in S_1 which is orthologous to g^{S2}, g^{S3} and vice versa. In this definition of *H,* a 1-1-1 triplet tends to have the same label. For simplicity, we do not specifically include a 3rd-order coupling term. Nevertheless, the 1-1-1 triplets are of particular importance among all the orthologous triplets because they correspond to particularly conserved biological function. Genes performing less conserved functions are more likely to undergo gene duplication and end up with many-to-many orthologs.

Discussion

In this study, we have presented OrthoClust, an orthology-based network framework that performs data clustering across multiple species. Due to the rapid increase of data from many species, a novel meta-clustering framework that integrates data from different species will be highly useful for comparative genomics. In OrthoClust, a module is defined based on evolutionary information as well as co-association information. A conserved module groups genes from multiple species corresponding to a common biological function; whereas a species-specific module consists of genes that are responsible for a specific function novel to a species. Though we have focused on expression data for illustration, OrthoClust can be readily applied to other high-dimensional data like histone modification patterns, or protein-protein interactions by appropriately modifying the cost function. For instance, in the context of interactome, OrthoClust can be used to detect modules in protein-protein interaction networks in a cross-species context and examine the evolution of protein complexes.

In single-species clustering, a conserved gene can be grouped to a species-specific module simply because of their strong tie. By incorporating orthology relationship between species, OrthoClust detects better the conserved modules. The concept of cross-species modules complements the principle of “guilt by association” because it may potentially lead to functionally analogous elements across species. This is of fundamental interest for elements like ncRNAs because only short regions of ncRNAs are constrained by structure or sequence-specific interactions [38]. Compared to protein-coding genes, this discrepancy in selection pressure

Koon-Kiu Yan 1/23/14 4:28 PM
Deleted: by incorporating orthology relationship between species,

Koon-Kiu Yan 1/23/14 4:08 PM
Deleted: a network framework

Koon-Kiu Yan 1/23/14 4:08 PM
Deleted: for detecting functional modules across multiple species. In summary, OrthoClust is a network framework that integrates the transcriptome of multiple species. Due to the rapid increase of data from many species, computational frameworks that integrate data from different species will be highly useful for comparative genomics.

Koon-Kiu Yan 1/23/14 4:12 PM
Deleted:

Koon-Kiu Yan 1/23/14 4:12 PM
Deleted: this orthology-based clustering framework

Koon-Kiu Yan 1/23/14 4:13 PM
Deleted: the function of

Koon-Kiu Yan 1/22/14 10:45 PM
Deleted: co-expression

makes it very hard to find orthologous ncRNAs by their sequence. RNA-Seq experiments have found that over 60% of the genome is transcribed, and many transcriptionally active regions (TARs) were identified [39]. Therefore, mapping onto OrthoClust modules based on expression profiles serves at least as an intermediate step to inferring the putative functions of the vast amount of TARs.

The essence of OrthoClust is a global optimization process. There are two aspects worth mentioning. First, as pointed out by several studies, finding network communities or modules using a global objective function has the issue of resolution limit [40][41]. The limit explains the existence of giant modules detected by OrthoClust. In principle, this could be complemented by a recursive scheme. Nevertheless, an extra coupling parameter $\boldsymbol{\lambda}$ can be added to the definition of Λ such that $\Lambda_{ij} = A_{ij} - \frac{\lambda k_i k_j}{2m}$. The parameter can control the size of resultant modules [18] but it is commonly chosen to be 1 by convention. In principle, the coupling parameter can be tuned to obtain a better resolution, or to obtain sizes that are more biologically relevant. Such a procedure is essentially a regularization process from the perspective of machine learning. Second, to minimize the cost function, we used simulated annealing as a conceptual demonstration. Though it is theoretically possible to obtain the optimal solution, it is computational expensive. There are faster approaches for module identification for single-layered network, for instance spectral techniques [15]. It is possible to generalize the concept for our particular multiplex network. A recent study combining various co-expression networks from the same species based on tensor computation point to a similar direction [42]. As simulated annealing scales not very well as the number of species increases, the analytical approaches described may present more efficient solutions.

To a certain extent, OrthoClust resembles the problem of network alignment. Nevertheless, the two problems are quite different. The essence of network alignment is to understand how individual nodes and edges in one network have their counterparts in another network, whereas OrthoClust focuses on whether genes working together in one species would preserve the collaboration in another species. Network alignment thus involves greater topological details, and to a certain extent it is a harder problem. As many of the networks

Koon-Kiu Yan 1/23/14 4:48 PM
Formatted: Font color: Black

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:30 PM
Deleted: [40]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:30 PM
Deleted: [41]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:30 PM
Deleted: [23]

Koon-Kiu Yan 1/23/14 4:53 PM
Deleted: find the

Koon-Kiu Yan 1/23/14 4:49 PM
Deleted: ground state of our energy

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:30 PM
Deleted: [26]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:30 PM
Deleted: [42]

Koon-Kiu Yan 1/23/14 4:51 PM
Deleted: While we have demonstrated that our framework can be generalized to three and even more species, it is in general not practical to implement it by simulated annealing.

Koon-Kiu Yan 1/23/14 4:51 PM
Deleted: T

Koon-Kiu Yan 1/23/14 5:20 PM
Formatted: Indent: First line: 0.5"

constructed are rather incomplete and there are possibly false positives, detecting modules is in general less sensitive to these errors compared to network alignment. It is worthwhile to point out that while the original motivation of some network alignment algorithms like IsoRank is to detect orthologous relationship between two species [31], what OrthoClust does could essentially be interpreted as refining the orthology relationships between two species via their corresponding co-association network. The rationale is actually important because common orthology detection approaches focus on the sequence level [43]. Because of the resemblance of network alignemtn and cross-species clustering, one could also modify OrthoClust by replacing the orthology relationships using aligned gene-pairs. As OrthoClust is a flexible computational framework, such modification would be technically straightforward but conceptually interesting to explore.

Materials and methods

Datasets of transcriptome and orthologous pairs

Transcriptome profiling data from worm and fly were generated by the modENCODE consortia using RNA-Seq. The expression values of worm and fly were measured across 33 and 30 developmental stages [26]. The total 11467 worm-fly orthologous pairs (including 1-to-1, 1-to-many, many-to-many relationships from 5769 unique worm orthologous genes, 5507 unique fly orthologous genes) between worm and fly were downloaded from the modENCODE website as they were compiled by the consortium [26]. At the genome-wide level, there are 20377 worm genes and 13623 fly genes. For each species, expression values in different developmental stages or cell lines were log-transformed, standardized and Pearson correlation coefficients were calculated for each pairs of genes.

The list of ncRNAs in worm and fly were obtained from wormbase [44] and flybase [30], including lncRNA (228 in worm, 852 in fly), miRNA (211 in worm, 215 in fly), snoRNA (141 in worm, 287 in fly) and tRNA (236 in worm, 238 in fly), all of which have a comparatively substantial number of RNAs (>100) in both worm and fly.

More details on OrthoClust

Koon-Kiu Yan 1/22/14 2:30 PM
Deleted: [29]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:30 PM
Deleted: [29]

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:30 PM
Deleted: [44]

The cost function

To take into account the fact that many orthologous pairs are not one-to-one but many-to-many, the contribution of a pair of orthologs to the generalized modularity function is not 1, but normalized by the number of orthologs. For example, if gene *i* from species 1 is orthologous to $d_i^{(o)}$ genes in species 2 including gene *j'* whereas gene *j'* in species 2 is orthologous to $d_{j'}^{(o)}$ genes in species 1 including gene *i*, the weight $w_{ij'}$ is defined as $(\frac{1}{d_i^{(o)}} + \frac{1}{d_{j'}^{(o)}})/2$. For simplicity, this modification is not displayed in the main equation.

Simulated annealing

Standard simulated annealing was employed. Spin values were randomly assigned initially, and updated via a heat bath algorithm. The initial temperature was chosen in a way such that the flipping rate (the probability that a node changes its spin state) was higher than 1-1/q. The temperature was gradually decreased with a cooling factor 0.9, until the flipping rate was less than 1%.

More details on applying OrthoClust to cluster expression profiles

Construction of individual co-expression networks

Many algorithms have been proposed to transform raw expression profiles into individual co-expression networks based on calculating the N by N Pearson correlation matrix [19][20][21][22]. There are two classes of algorithms: valued-based algorithms by thresholding the correlation coefficients, and rank-based algorithms in which each gene is connected to only a certain number of genes. We found that networks constructed by value-based algorithm are in general more difficult to resolve into smaller modules. Therefore, a rank-based algorithm in which each gene is connected to the top *d* genes with the highest (absolute) Pearson correlation was employed [19].

The value of *d* was chosen in order to keep the sparsity of networks. More specifically, d is the smallest value such that all genes from individual species independently form giant connected networks. In general, if *d* is very small, the resultant network by definition cannot form a giant connected graph. On the other hand, if d is very large, the network would not be sparse. In the

Koon-Kiu Yan 1/23/14 5:32 PM
Deleted: is

Koon-Kiu Yan 1/23/14 5:32 PM
Deleted: to

Koon-Kiu Yan 1/23/14 5:33 PM
Deleted: be

Koon-Kiu Yan 1/23/14 5:32 PM
Deleted: .

Koon-Kiu Yan 1/23/14 5:33 PM
Deleted: To keep the sparsity of networks, a rank-based algorithm in which each gene is connected to the top *d* genes with the highest (absolute) Pearson correlation coefficient was employed [19]. If *d* is very small, the resultant network by definition cannot form a giant connected graph.

worm fly analysis, d was chosen to be 5. Even though the number of nodes and edges in the two co-expression networks vary, the average number of links per node is quite similar (6.29 for worm and 6.56 for fly).

Koon-Kiu Yan 1/24/14 12:20 PM
Deleted: *d*

Koon-Kiu Yan 1/24/14 12:20 PM
Deleted: is

Decomposition of modules in worm and fly

In the genome-wide worm fly analysis, a stringent threshold (0.95) for co-appearance was employed for the co-appearance matrix shown in Figure 3A. Nodes that ended up with the same spin value for more than 95% of trials were assigned to the same module. Tiny clusters were neglected, arriving at a set of about 150 modules covering about 80% of nodes. Proper GO terms were found in the remaining modules.

Regulatory patterns of modules

ChIP-Seq data of 26 fly transcription factors and 79 worm transcription factors across various developmental stages (together 220 experiments in worm and 93 experiments in fly) were downloaded from the modENCODE consortium. For each ChIP-Seq experiment, binding targets of the transcription factors were identified by TIP [45] with a q-value cutoff of 0.01. The results of these experiments were superposed together to form the transcriptional regulatory networks for worm and fly (12600 edges for worm and 1200 edges for fly). The number of common transcription factors for a pair of genes was determined based on the resultant networks.

Comparison with single-species clustering

Standard clustering procedures, including k-means, hierarchical clustering and PAM, were applied for transcriptome profiling data from worm and fly. By neglecting resultant modules of size less than 5 genes, xxx worm modules and xxx fly modules remained. For each combination of these modules, the number of orthologous pairs between worm and fly genes was counted. The number of orthologous pairs was then compared to the expected number $\frac{n_w n_f}{N_w N_f} O_{wf}$, where n_w and n_f are the number of genes in the worm and fly modules, N_w and N_f are the total number of worm and fly genes, and O_{wf} is the number of orthologous pairs between worm and fly. Only

combinations with the enrichment of orthologous pairs over (P<0.05, hypergeometric test) were kept for the display in Figure 6. Modules of size less than 5 genes generated by OrthoClust were also neglected in the comparison.

Comparison with Network alignment

We applied IsoRank to align the worm and fly co-expression networks. The sequence identity between pairs of worm, fly proteins were downloaded from ref. [31]. We tuned the intrinsic parameters α but we did not find systematic trends. We then used α=0.5 and looked for co-expression edges in two networks whose nodes are aligned. Disconnected components formed by these aligned edges were used as potential seeds of conserved modules because they consist of sets of worm genes and fly genes that are perfectly aligned.

Robustness analysis

To compare two sets of clusters A and B, all possible N(N-1)/2 pairs of genes were divided into 4 categories: I. Assigned to the same module by both A and B; II. Assigned to the same module by A but not by B; III. Assigned to the same module by B but not by A. IV. Assigned to different modules by both A and B. Because the number of pairs in IV (true negative to a certain extent) is orders of magnitude higher than the others, the overlap between A and B was defined as I/(I+II+III). The number of pairs in I can be viewed as the true positive. The method is motivated by Ref. [46].

Koon-Kiu Yan 1/22/14 2:42 PM
Formatted: Font:Bold

Koon-Kiu Yan 1/22/14 2:42 PM
Formatted: Font:Bold, Font color: Gray-80%

Koon-Kiu Yan 1/24/14 4:11 PM
Formatted: Font:Bold

More details on inferring the functions of worm fly ncRNAs

Modules based on worm fly core set

OrthoClust was applied to the set of orthologs between worm and fly, consisting of 5059 worm genes and 4863 fly genes. The coupling constant was determined using the same scheme illustrated in the main text. A set of 21 modules with its number of genes greater than 15 and with enriched GO terms was arrived. As expected, the similarity between genes within modules was higher than the similarity between genes across modules ($P=1\times10^{-83}$, Wilcoxon test). To annotate

the functions of a module, the enriched GO terms among genes were obtained using the tool REVIGO [47]. The enriched GO terms were clustered into groups labeled by representative keywords given by REVIGO. The list of keywords was displayed by tag clouds in which the size of a keyword is proportional to the number of GO terms in the group (http://tagcrowd.com/).

Mapping ncRNAs to modules

Using RNA-Seq data generated by the modENCODE consortium, the expression profiles of ncRNAs were calculated under the same set of developmental stages as compared to the protein-coding genes. Annotation of ncRNAs was based on the latest version of wormbase [44] and flybase [48]. The ncRNAs were then mapped to the 21 modules based on the correlation between expression levels. More specifically, the correlation between the expression of the ncRNA and the expression of all protein-coding genes was calculated. A null distribution was constructed by randomly shuffling the expression values of the ncRNA 10 times, calculating the correlation coefficients between the randomized expression profile with expression profiles of all the protein coding genes, and pooling all values together. A set of protein-coding neighbors of the ncRNA was identified as the set of most correlated protein-coding genes with a false discovery rate of less than 5% with respect to the null distribution. The enrichment of the set of neighbors in every module was calculated by a hypergeometric test. The ncRNA would be mapped to a given module if $P<0.01$. An ncRNA could be mapped to multiple modules.

Enrichment of different classes of ncRNAs in modules

To obtain the enrichment of a particular class of ncRNA (miRNA, tRNA, snoRNA, lncRNA) with respect to the set of all ncRNA in a given module, a hypergeometric test was employed to calculate the significance of the fraction of mapped ncRNAs of this class to four classes in the module against the fraction of total mapped ncRNAs of this class to four classes across all modules.

Competing of Interest

The authors declare that they have no competing interests.

Unknown
Field Code Changed

Unknown
Field Code Changed

Koon-Kiu Yan 1/22/14 2:30 PM
Deleted: [44]

Unknown
Field Code Changed

Authors' contributions

KKY and MG conceived of the study. KKY and DW performed the analysis and drafted the manuscript. JR, HZ and CC contributed ideas and tools for the analysis and edited the manuscript.

Acknowledgements

We acknowledge Arif Harmanci, Michael Rutenberg Schoenberg and Gang Fang for valuable discussions. We acknowledge Wyatt Clark for critically reading the manuscripts. We acknowledge support from NIH and the AL Williams Professorship funds.

References

1. Berger B, Peng J, Singh M: **Computational solutions for omics data**. *Nat. Rev. Genet.* 2013, **14**:333–346.

2. Soon WW, Hariharan M, Snyder MP: **High-throughput sequencing for biology and medicine**. *Mol. Syst. Biol.* 2013, **9**.

3. Alon U: **Biological Networks: The Tinkerer as an Engineer**. *Science* 2003, **301**:1866–1867.

4. Hartwell LH, Hopfield JJ, Leibler S, Murray AW: **From molecular to modular cell biology**. *Nature* 1999, **402**:C47–C52.

5. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis**. *BMC Bioinformatics* 2008, **9**:559.

6. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc. Natl. Acad. Sci. U. S. A.* 1998, **95**:14863–14868.

7. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR: **Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation**. *Proc. Natl. Acad. Sci.* 1999, **96**:2907–2912.

8. Kluger Y, Basri R, Chang JT, Gerstein M: **Spectral biclustering of microarray data: coclustering genes and conditions**. *Genome Res.* 2003, **13**:703–716.

Koon-Kiu Yan 1/24/14 5:18 PM
Formatted: Bibliography, Widow/Orphan control, Adjust space between Latin and Asian text, Adjust space between Asian text and numbers

Unknown
Field Code Changed

Koon-Kiu Yan 1/24/14 5:18 PM
Formatted: Font:Cambria

9. Agrawal H, Domany E: **Potts ferromagnets on coexpressed gene networks: identifying maximally stable partitions**. *Phys. Rev. Lett.* 2003, **90**:158102.

10. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat. Methods* 2008, **5**:621–628.

11. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat. Rev. Genet.* 2009, **10**:57–63.

12. Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J-P: **Community Structure in Time-Dependent, Multiscale, and Multiplex Networks**. *Science* 2010, **328**:876–878.

13. Newman MEJ, Strogatz SH, Watts DJ: **Random graphs with arbitrary degree distributions and their applications**. *Phys. Rev. E* 2001, **64**:026118.

14. Maslov S, Sneppen K: **Specificity and Stability in Topology of Protein Networks**. *Science* 2002, **296**:910–913.

15. Newman MEJ: **Modularity and Community Structure in Networks**. *Proc. Natl. Acad. Sci.* 2006, **103**:8577–8582.

16. Traag VA, Bruggeman J: **Community detection in networks with positive and negative links**. *Phys. Rev. E* 2009, **80**:036115.

17. Wu FY: **The Potts model**. *Rev. Mod. Phys.* 1982, **54**:235–268.

18. Reichardt J, Bornholdt S: **Detecting Fuzzy Community Structures in Complex Networks with a Potts Model**. *Phys. Rev. Lett.* 2004, **93**:218701.

19. Ruan J, Dean A, Zhang W: **A general co-expression network-based approach to gene expression analysis: comparison and applications**. *BMC Syst. Biol.* 2010, **4**:8.

20. Zhou X, Kao M-CJ, Wong WH: **Transitive functional annotation by shortest-path analysis of gene expression data**. *Proc. Natl. Acad. Sci. U. S. A.* 2002, **99**:12783–12788.

21. Van Noort V, Snel B, Huynen MA: **The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model**. *EMBO Rep.* 2004, **5**:280–284.

22. Jordan IK, Mariño-Ramírez L, Wolf YI, Koonin EV: **Conservation and coevolution in the scale-free human gene coexpression network**. *Mol. Biol. Evol.* 2004, **21**:2058–2070.

23. Stuart JM: **A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules**. *Science* 2003, **302**:249–255.

24. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, Li M, Sousa AMM, Pletikos M, Meyer KA, Sedmak G, Guennel T, Shin Y, Johnson MB, Krsnik Ž, Mayer S, Fertuzinhos S, Umlauf S, Lisgo SN, Vortmeyer A, Weinberger DR, Mane S, Hyde TM, Huttner A, Reimers M, Kleinman JE, Šestan N: **Spatio-temporal transcriptome of the human brain**. *Nature* 2011, **478**:483–489.

25. Mao L, Van Hemert JL, Dash S, Dickerson JA: **Arabidopsis gene co-expression network and its functional modules**. *BMC Bioinformatics* 2009, **10**:346.

26. Gerstein M, Rozowsky J, Yan K-K, Wang D, et al: **Comparison of 3 metazoan transcriptomes**. .

27. Jones KS: **A statistical interpretation of term specificity and its application in retrieval**. *J. Doc.* 1972, **28**:11–21.

28. Yu H, Jansen R, Stolovitzky G, Gerstein M: **Total ancestry measure: quantifying the similarity in tree-like classification, with genomic applications**. *Bioinforma. Oxf. Engl.* 2007, **23**:2163–2173.

29. Newman MEJ: *Analysis of weighted networks*. 2004.

30. Berg J, Lassig M: **Cross-Species Analysis of Biological Networks by Bayesian Alignment**. *Proc. Natl. Acad. Sci.* 2006, **103**:10967–10972.

31. Singh R, Xu J, Berger B: **Global alignment of multiple protein interaction networks with application to functional orthology detection**. *Proc. Natl. Acad. Sci.* 2008, **105**:12763–12768.

32. Ficklin SP, Feltus FA: **Gene coexpression network alignment and conservation of gene modules between two grass species: maize and rice**. *Plant Physiol.* 2011, **156**:1244–1256.

33. Oliver S: **Proteomics: Guilt-by-association goes global**. *Nature* 2000, **403**:601–603.

34. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, Huarte M, Zuk O, Carey BW, Cassady JP, Cabili MN, Jaenisch R, Mikkelsen TS, Jacks T, Hacohen N, Bernstein BE, Kellis M, Regev A, Rinn JL, Lander ES: **Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals**. *Nature* 2009, **458**:223–227.

35. Liao Q, Liu C, Yuan X, Kang S, Miao R, Xiao H, Zhao G, Luo H, Bu D, Zhao H, Skogerbø G, Wu Z, Zhao Y: **Large-scale prediction of long non-coding RNA**

functions in a coding–non-coding gene co-expression network. *Nucleic Acids Res.* 2011.

36. Dai H, Chen Y, Chen S, Mao Q, Kennedy D, Landback P, Eyre-Walker A, Du W, Long M: **The evolution of courtship behaviors through the origination of a new gene in Drosophila**. *Proc. Natl. Acad. Sci.* 2008, **105**:7478–7483.

37. Nam J-W, Bartel DP: **Long noncoding RNAs in C. elegans**. *Genome Res.* 2012, **22**:2529–2540.

38. Pang KC, Frith MC, Mattick JS: **Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function**. *Trends Genet.* 2006, **22**:1–5.

39. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Röder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, Derrien T, Drenkow J, Dumais E, Dumais J, Duttagupta R, Falconnet E, Fastuca M, Fejes-Toth K, Ferreira P, Foissac S, Fullwood MJ, Gao H, Gonzalez D, Gordon A, Gunawardena H, Howald C, Jha S, Johnson R, Kapranov P, King B, Kingswood C, Luo OJ, Park E, Persaud K, Preall JB, Ribeca P, Risk B, Robyr D, Sammeth M, Schaffer L, See L-H, Shahab A, Skancke J, Suzuki AM, Takahashi H, Tilgner H, Trout D, Walters N, Wang H, Wrobel J, Yu Y, Ruan X, Hayashizaki Y, Harrow J, Gerstein M, Hubbard T, Reymond A, Antonarakis SE, Hannon G, Giddings MC, Ruan Y, Wold B, Carninci P, Guigó R, Gingeras TR: **Landscape of transcription in human cells**. *Nature* 2012, **489**:101–108.

40. Fortunato S, Barthélemy M: **Resolution limit in community detection**. *Proc. Natl. Acad. Sci.* 2007, **104**:36–41.

41. Kumpula JM, Saramäki J, Kaski K, Kertész J: **Limited resolution in complex network community detection with Potts model approach**. *Eur. Phys. J. B - Condens. Matter Complex Syst.* 2007, **56**:41–45.

42. Li W, Liu C-C, Zhang T, Li H, Waterman MS, Zhou XJ: **Integrative Analysis of Many Weighted Co-Expression Networks Using Tensor Computation**. *PLoS Comput Biol* 2011, **7**:e1001106.

43. Fang G, Bhardwaj N, Robilotto R, Gerstein MB: **Getting Started in Gene Orthology and Functional Analysis**. *PLoS Comput Biol* 2010, **6**:e1000703.

44. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, De La Cruz N, Davis P, Duesbury M, Fang R, Fernandes J, Han M, Kishore R, Lee R, Müller H-M, Nakamura C, Ozersky P, Petcherski A, Rangarajan A, Rogers A, Schindelman G, Schwarz EM, Tuli MA, Van Auken K, Wang D, Wang X, Williams G, Yook K, Durbin R, Stein LD, Spieth J, Sternberg PW: **WormBase: a comprehensive resource for nematode research**. *Nucleic Acids Res.* 2010, **38**:D463–467.

45. Cheng C, Min R, Gerstein M: **TIP: A probabilistic method for identifying transcription factor target genes from ChIP-seq binding profiles**. *Bioinformatics* 2011, **27**:3221–3227.

46. Brohée S, Helden J van: **Evaluation of clustering algorithms for protein-protein interaction networks**. *BMC Bioinformatics* 2006, **7**:488.

47. Supek F, Bošnjak M, Škunca N, Šmuc T: **REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms**. *PLoS ONE* 2011, **6**:e21800.

48. Marygold SJ, Leyland PC, Seal RL, Goodman JL, Thurmond J, Strelets VB, Wilson RJ, FlyBase consortium: **FlyBase: improvements to the bibliography**. *Nucleic Acids Res.* 2013, **41**:D751–757.

Figure Legends

Figure 1

An example to illustrate the idea of modules in a multiplex network. The co-association networks of species A and B are linked together to form a multiplex network via orthologous relationship between genes. There are three modules. The middle one is a conservative module with genes from both species, corresponding to fundamental biological functions across different species. The left and right ones are specific modules consisting of genes from species A and B respectively. They correspond to novel function emerged in two species.

Koon-Kiu Yan 1/24/14 7:07 PM
Deleted: expression

Figure 2

An outline of OrthoClust. The inputs of our pipeline are co-association networks from multiple species as well as orthology relationships. A cost function is defined based on the topology of the co-expression networks as well as orthology relationships. Each node can be in one of q possible states labeled by 1 to a. The cost function H is optimized by simulated annealing. In simulated annealing, labels are randomly assigned initially and are allowed to flip based on *H*. The temperature of the system is gradually lowered with a cooling factor α=0.9. The algorithm stops if the flipping rate is low enough. The labels of nodes at at the optimal configuration represent the assignment of nodes to modules. The algorithm is repeated multiple times. The resultant modules, represented by a set of Venn diagrams, could be specific or conserved.

Koon-Kiu Yan 1/24/14 7:10 PM
Deleted: expression profiles

Koon-Kiu Yan 1/24/14 7:10 PM
Deleted: Expression profiles of protein-coding genes are used to construct co-expression networks.

Koon-Kiu Yan 1/24/14 7:10 PM
Deleted: n energy

Koon-Kiu Yan 1/24/14 11:07 PM
Deleted: The system is mapped to a *q*-state spin system, and the ground state of *H* is obtained by

Koon-Kiu Yan 1/24/14 11:07 PM
Deleted: spin values

Koon-Kiu Yan 1/24/14 11:07 PM
Deleted: spin values

Koon-Kiu Yan 1/24/14 11:07 PM
Deleted: ground state

Figure 3

A. The co-appearance matrix of worm and fly genes. The worm and fly genes were sorted separately. Blocks along the diagonal are modules of worm and fly. Some blocks along the diagonal have strong co-appearance at the off-diagonal positions (see red circles as an example). These are conserved modules across worm and fly. In such modules, the corresponding worm and fly genes show strong overlap of GO terms ($P=3.3 \times 10^{-16}$, hypergeometric test). There are blocks along the diagonal that have no overlap at the off-diagonal positions (the blue pentagon and the green hexagon). They are the worm specific and fly specific modules. Such modules have rare overlap in terms of their GO terms ($P<0.05$, hypergeometric test).

B. Enriched GO terms of a conserved module in worm and fly. Each circle represents a GO term, and the color code stands for statistical significance. The terms project onto a semantic space in which the geometric distance between GO terms mirrors their sematic distance. GO terms with similar meanings are pack together. GO terms correspond to fundamental functions like RNA biology, cell cycle, etc.

Figure 4

The similarity of gene pairs within modules versus the similarity of gene pairs between modules. Genes within modules are significantly more similar than genes from different modules.

Figure 5

A. The effects of κ on clustering. As κ increases, the modularity scores of worm (green) and fly (blue) co-expression networks decrease. The fraction of metagenes whose components are found in the same modules decrease as κ increases.

B. The effects of κ on the modularity of GO reference network. The modularity peaks at $\kappa=3$, meaning that the modules defined by that particular value of coupling constant best separate the genes in terms of their GO annotations.

Figure 6

Comparison between single-species clusters and cross-species clusters generated by OrthoClust. Pairwise overlapping of clusters generated by k-means, hierarchical clustering or PAM has little enrichment of orthologous pairs compared to cross-species modules generated by OrthoClust.

Koon-Kiu Yan 1/22/14 2:14 PM
Formatted: Font:Bold

Figure 7

Koon-Kiu Yan 1/24/14 11:09 PM
Formatted: Font:Not Bold

Koon-Kiu Yan 1/24/14 11:14 PM
Deleted: 6

The set of conserved worm-fly modules and their annotated functions. The boxes represent modules. For each module, potential functions are summarized by keywords associated with enriched GO terms in a tag cloud. The font of a keyword is proportional to the frequency of occurrence of the corresponding GO terms in the module. The position of a module in the vertical direction represents the fraction of genes with orthologous partners in the module.

Additional Files

Additional file 1 -Supplementary Figures (.pdf)

Additional file 2 –Supplementary Dataset 1(.csv)

The list of worm and fly genes in the 21 core modules

Additional file 3 –Supplementary Dataset 2(.csv)

This file lists all worm and fly ncRNAs used in the analysis and modules they were mapped. The rows are ncRNAs. The columns are the modules. The values 1 and 0 mean whether a ncRNA is being mapped to a module or not.