Madhamshettiwar et al. BMC Bioinformatics 2013, 14(Suppl 16):S14
http://www.biomedcentral.com/1471-2105/14/S16/S14

BMC Bioinformatics

RESEARCH Open Access

RMaNI: Regulatory Module Network Inference framework

Piyush B Madhamshettiwar[1,2], Stefan R Maetschke[1,2], Melissa J Davis[1,2], Mark A Ragan[1,2*]

From Asia Pacific Bioinformatics Network (APBioNet) Twelfth International Conference on Bioinformatics (InCoB2013)
Taicang, China. 20-22 September 2013

Abstract

Background: Cell survival and development are orchestrated by complex interlocking programs of gene activation and repression. Understanding how this *gene regulatory network* (GRN) functions in normal states, and is altered in cancers subtypes, offers fundamental insight into oncogenesis and disease progression, and holds great promise for guiding clinical decisions. Inferring a GRN from empirical microarray gene expression data is a challenging task in cancer systems biology. In recent years, module-based approaches for GRN inference have been proposed to address this challenge. Despite the demonstrated success of module-based approaches in uncovering biologically meaningful regulatory interactions, their application remains limited a single condition, without supporting the comparison of multiple disease subtypes/conditions. Also, their use remains unnecessarily restricted to computational biologists, as accurate inference of modules and their regulators requires integration of diverse tools and heterogeneous data sources, which in turn requires scripting skills, data infrastructure and powerful computational facilities. New analytical frameworks are required to make module-based GRN inference approach more generally useful to the research community.

Results: We present the RMaNI (Regulatory Module Network Inference) framework, which supports cancer subtype-specific or condition specific GRN inference and differential network analysis. It combines both transcriptomic as well as genomic data sources, and integrates heterogeneous knowledge resources and a set of complementary bioinformatic methods for automated inference of modules, their condition specific regulators and facilitates downstream network analyses and data visualization. To demonstrate its utility, we applied RMaNI to a hepatocellular microarray data containing normal and three disease conditions. We demonstrate that how RMaNI can be employed to understand the genetic architecture underlying three disease conditions. RMaNI is freely available at http://inspect.braembl.org.au/bi/inspect/rmani

Conclusion: RMaNI makes available a workflow with comprehensive set of tools that would otherwise be challenging for non-expert users to install and apply. The framework presented in this paper is flexible and can be easily extended to analyse any dataset with multiple disease conditions.

Background

Complex cellular behaviour in cancer is orchestrated by the action of transcriptional regulatory networks [1,2]. Computational inference of transcriptional regulatory networks, referred to as Gene Regulatory Networks (GRN), from microarray gene expression data is one of the fundamental goals of systems biology and its translation to genomic medicine [3]. GRN inference and analysis, especially when integrated with experimental validation, has proven to be a powerful tool in understanding how regulatory networks are disrupted and rewired in normal and cancer conditions, and in identifying novel regulatory interactions as well as broader systemic disruptions in key oncogenic processes [4-6]. Many methods have been

* Correspondence: m.ragan@uq.edu.au
[1]Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Road, St Lucia, Brisbane, Queensland 4072, Australia
Full list of author information is available at the end of the article

developed to infer GRNs from microarray gene expression data. These approaches include unsupervised, semi-supervised and supervised methods based on computational mathematics, multivariate statistics and information science [7-11].

Although diverse computational and statistical approaches have been applied to this problem, the accuracy of edge-wise network inference methods remains poor [11-14]. Novel approaches are needed to address the genome-wide network inference problem. A promising direction is the inference of transcriptional modules instead of individual edges. Module inference is simpler than edge-wise network inference [15,16], and higher accuracies can be achieved [7,17].

Transcriptional module networks

Several studies have revealed that regulatory networks are modular in nature and organised hierarchically [18]. According to Oltvai and Barabasi's "complexity of life" pyramid, functional modules are less complex compared to individual transcriptional programs, which in turn are the building blocks for these modules [15]. Therefore, inferring modules instead of the individual interactions of complete networks drastically reduces the complexity of the inference problem, and shows great promise for network analysis in complex disease conditions including cancer [17,19-21]. A transcriptional-module network is composed of clusters of co-expressed genes collaboratively or alternatively regulated by one or several transcription factors (TFs) *via* convergent or divergent regulatory programs. A *convergent regulatory program* represents a particular set of target genes (TGs) regulated by different sets of TFs, whereas a *divergent regulatory program* represents a given set of TFs regulating distinct sets of TGs [7,22].

Several methods have been developed to infer modules from microarray data, including a range of clustering methods such as *k*-means, hierarchical clustering and self-organizing maps. However, all these approaches suffer from certain limitations; for instance, the number of clusters is not determined automatically but requires the number of clusters to be pre-specified [23-26]. WGCNA [27], based on the *weighted gene co-expression network analysis* approach [28], is the most widely used method and has been applied to a number of diseases [29-32]. It also uses a clustering approach to infer modules, but it optimizes the threshold to achieve a scale-free topology. Assuming scale-freeness, several model-based clustering approaches have been developed [33-35]. Model-based approaches allow a statistical analysis of the inferred modules and automatically estimate the number of modules [34]. For example, Genomica [20] uses expectation maximisation (EM) to identify modules [16,20].

Other methods [22,36-39] use additional experimental data such as protein-protein interactions, TF binding affinity data, *in vitro* DNA binding specificities, DNA motifs and ChIP-chip data. Such integrative approaches are attractive and promising approaches to infer modules, as they take into account different sources of biological information [40]. However, they do not natively integrate methods for module inference, identification of regulators, or comprehensive downstream analysis and visualization. Also, they support the analysis only of individual datasets arising from only one condition without differential analysis of other conditions or subtypes.

Integrating diverse data sources as well as multiple methods brings many challenges. These challenges can be diverse, range from methodological to practical in nature, and can arise due to the computational or statistical complexities of methods and the dimensionality of omic data [41,42]. For instance, combining heterogeneous data requires extensive file formatting at different stages of analysis, while integrating different methods involves the selection or optimization of diverse parameters and other user-control features. As a consequence of these challenges, it is difficult for biologists or clinicians (without strong informatic skills) to chain multiple methods together into comprehensive, flexible workflows to address substantial questions. For example, to identify the modules involved in any disease condition one must retrieve data from different repositories (*e.g.* motif data from Transfac [43] or Genomatix [44]), map the identifiers *e.g.* using Biomart [45], perform differential gene expression analysis *e.g* using LIMMA [46], infer the modules and identify regulators *e.g.* using Genomica [20], integrate the inferred modules and regulators for visualization *e.g.* using Cytoscape [47], and finally perform functional analysis of module genes *e.g.* using DAVID [48,49]. This work focuses on making available a workflow and computational resources for the inference of modules and their regulators, downstream analyses and visualization.

RMaNI - Regulatory Module Network Inference framework

Here, we present a novel integrative and automated analytical framework "RMaNI - Regulatory Module Network Inference" for disease condition or subtype-specific module network inference, analysis and data visualization. It uses the Learning Module Networks (LeMoNe) algorithm [50] and Regulatory Impact Factors (RIF) [51] to identify relevant regulatory TFs. The LeMoNe algorithm uses a Bayesian probabilistic model-based approach for clustering genes, and in selecting thresholds does not assume that networks necessarily have a scale-free topology [50].

RMaNI combines both transcriptomic as well as genomic data sources, and integrates heterogeneous knowledge resources and a set of complementary bioinformatic methods for microarray data processing, differential expression (DE) analysis, module detection and regulator identification, gene and module significance measure calculations,

functional enrichment analysis of module genes, and visualization of data and networks.

Case study - application to hepatocellular carcinoma

To demonstrate its utility, we applied RMaNI to a hepatocellular microarray dataset containing normal tissue and three disease conditions: pre-malignant (cirrhosis), cirrhosis with hepatocellular carcinoma (cirrhosisHCC), and hepatocellular tumor (HCC). We illustrate that the identification and analysis of transcriptional module network can give insight into the common and unique genetic architecture underlying hepatocellular carcinoma conditions.

Implementation

The RMaNI web interface has been created using Rwui [52], a Java-based application that uses the Apache Struts framework. The complete application is running on a Tomcat server on a high-performance computing cluster. The workflow integrates publicly available R[53], Bioconductor [54] and custom packages and functions for data import, processing, analysis, integration and visualization. All packages are currently running under R version 2.15.2, and can be easily updated as newer versions of R are released. RMaNI is freely available as a user-friendly web-application at http://inspect.braembl.org.au/bi/inspect/rmani, with a comprehensive manual available (Additional File 1).

In the next section, we describe the RMaNI workflow and provide a brief overview of the methods used in each step. Then, we present a case study showing how RMaNI can be employed to understand the genetic architecture underlying three hepatocellular carcinoma conditions.

RMaNI: structure and functionalities

Figure 1 illustrates the workflow in RMaNI. The workflow is divided into three main stages: 1) data preparation, 2) inference of modules and regulators, and 3) integration of module networks and analysis. In this section, we describe these stages and the individual steps involved therein.

Stage 1 - Data Preparation

At this stage, the pre-processed (background corrected and normalized) microarray gene expression data and sample annotations are imported from files uploaded by the user.

Step 1.1 - Dataset

sRMaNI can be applied to gene expression datasets arising from multiple conditions. Currently, we support datasets arising from 13 different types of Affymetrix chips: hgu133a, hgu133a2, hgu133b, hgu133plus2, hgu219, hgu95a, hgu95av2, hgu95b, hgu95c, hgu95d, hgu95e, hthgu133a and hthgu133b.

Step 1.2 - Feature selection for input to module inference workflow

Once a user has the microarray dataset, the question arises: which and how many features should one input to the network inference step? Because there is no standard feature-selection method or recommendation on the minimum number of features, the workflow compares different feature-selection methods for different gene sets, and identifies the optimal combination of these two parameters.

The user compares three feature-selection methods: differentially expressed genes between normal and all subtypes (DE_all), differentially expressed genes between normal and each subtype (DE_pair), and the most-variable genes across the dataset based on the coefficient of variation (Var). For differential expression analysis RMaNI uses the LIMMA package [46], and to select variable genes it uses a custom R function. To find the optimal number of genes, for each of the three feature selection methods, it selects eight subsets with 10 to 4000 genes (10, 50, 100, 200, 500, 1000, 2000 and 4000 genes) optimal for network inference step. To identify the optimal feature selection method and number of genes, it examines how well they group the samples into the known classes.

The user compares the different gene sets on seven different clustering methods (clues, kmeans, PAM, AGNES, Fanny, SOTA and MCLUST) [55-57]. The workflow uses the Rand Index (RI) [58] as a measure for evaluating the clustering performance. RI measures the similarity between two data clusterings (known against predicted). An RI equal to 1 indicates perfect clustering, while an RI of 0 indicates that the clustering is no better than chance. These methods are implemented in the R packages clValid [59], clues [55], cluster [57] and mclust [56]. A brief description of each clustering method is given below.

clues (clustering based on local shrinking) is a nonparametric clustering method using local shrinking [55]. It estimates the number of clusters and simultaneously finds a partition of a data set *via* three steps: shrinking, partition, and determination of the optimal number of partitions.

kmeans is a parametric, centroid-based clustering method. Given the number of clusters, it starts with an initial estimate for the cluster centroids, and each sample is assigned to the cluster with the nearest mean [60]. The cluster centroids are then updated, and the entire process is iterated until the cluster centroids become stable.

PAM (Partitioning Around Medoids) is a parametric method similar to *k*-means, but PAM is a medoid-based method. A *medoid* is a representative object of a cluster, such that its average dissimilarity to all objects in that cluster is minimal [61]. Given the number of clusters, PAM starts with an initial estimate for the cluster medoids, and calculates the dissimilarity matrix using the

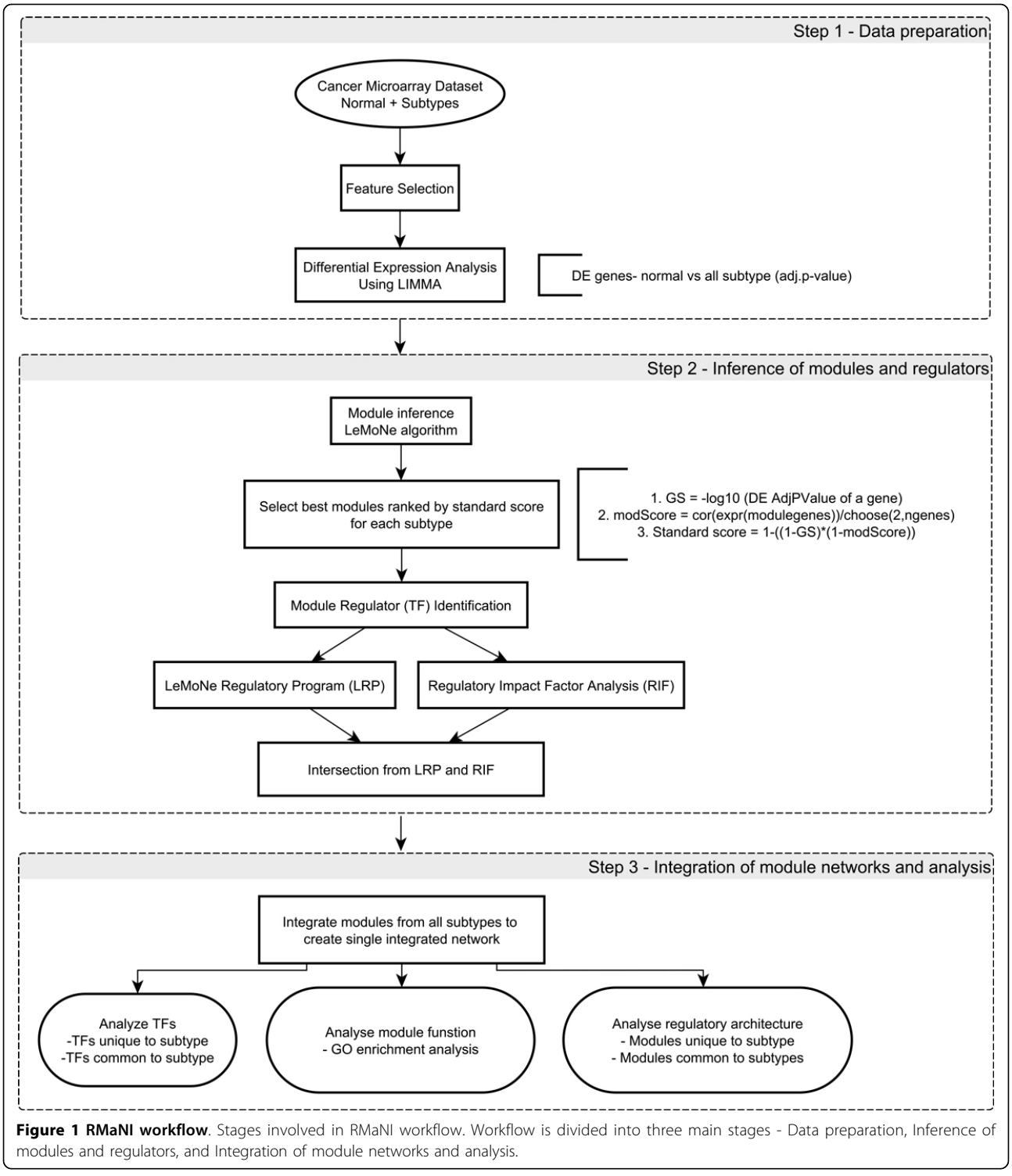

Figure 1 RMaNI workflow. Stages involved in RMaNI workflow. Workflow is divided into three main stages - Data preparation, Inference of modules and regulators, and Integration of module networks and analysis.

Euclidean or Manhattan distance [61]. Based on this matrix, each sample is assigned to the cluster with the nearest medoid.

AGNES (AGglomerative NESting) is a hierarchical clustering method which groups a dataset into a tree of clusters [61]. It is a bottom-up clustering method that starts with small clusters of single samples and then, at each step using a specified distance metric, merges the clusters into larger cluster. This is repeated iteratively until a single cluster is obtained, containing all samples.

Fanny is a fuzzy or soft clustering method [61]. With this method each sample has partial membership with each cluster rather than belonging exclusively to just a single cluster. Each sample describes the probability scores for its cluster membership. After optimizing the number of clusters, the method starts with assigning random cluster probabilities to each sample, and repeats this process until convergence.

SOTA (Self-Organising Tree Algorithm) is a divisive clustering method [62]. It generates an unsupervised neural network with a binary tree topology. Contrary to AGNES, SOTA is a top-down clustering method. It starts the clustering process with a binary tree consisting of a root node with two leaves, each representing one cluster. The self-organizing process then grows the tree by converting the leaf with the highest score into a node and attaching two new leaves to it. The score for each cluster is defined as the mean value of the distances between the cluster and the samples associated with it [63].

MCLUST (Model based clustering) is a nonparametric, model-based clustering method that uses finite normal mixture modelling and the expectation maximisation (EM) algorithm. Unlike other methods, it does not require the number of clusters as input, but instead infers the number of clusters from the data.

In summary, stage 1 provides an estimate on the feature-selection method and optimal number of genes which best explains the given data. The user can choose this feature selection method and this many genes for input to find clusters of co-expressed genes in the next step. For ease and flexibility of processing the user's own data, the feature selection step is not supported through the RMaNI web-interface.

Stage 2 - Clustering of genes to modules and identification of regulators

This is the main stage of the RMaNI workflow. It takes the gene set optimized in the feature selection step, and uses the corresponding gene expression data for module network inference. Below we provide the details of the individual steps.

Step 2.1 - Inference of transcriptional module networks

Given a gene expression dataset and a set of candidate regulators (TFs, microRNA or clinical variable of interest like stage or grade); inference of modules is composed of two steps: first clustering of co-expressed genes to identify modules, and second the inference of links between regulators and modules.

Step 2.1.1 - Clustering of genes

RMaNI uses the LeMoNe (Learning Module Networks) algorithm for inferring modules from microarray data, LeMoNe performs a two-way Bayesian clustering of genes and uses a Gibbs sampling procedure to iteratively update the cluster assignments of genes [34,50]. Each inferred module contains the genes for which the expression profiles best fit the same multivariate normal distribution [7]. LeMoNe has been successfully applied to different conditions including cancer [17,64-67]. LeMoNe outputs the ensemble of clustering solutions represented as a gene-to-cluster probability matrix reflecting the probability of the assignment of a gene to each module, referred to as fuzzy clustering (one gene can belong to multiple modules, each with certain probability). Using a graph spectral method and a probability cut-off, it then outputs tight clusters (in which one gene belongs to only one cluster) from fuzzy clusters [50].

Step 2.2 - Inferring the regulators

To identify and prioritize potential TFs regulating modules, candidate TFs are gathered by integrating lists of TFs from Vaquerizas [68], Ravasi [69], TCOF-DB [70] and Transfac [43]. To infer the potential regulator for each module, two methods are employed: LeMoNe's regulatory program (LRP) and the Regulatory Impact Factor analysis (RIF) algorithm. Below, we briefly describe these methods.

Step 2.2.1 - LeMoNe regulatory program

In LeMoNe's regulatory program, two types of regulators can be assigned, regulators with continuous or with discrete values. Continuous values include expression values measured, for example, for TFs, signal transducers, kinases and/or microRNAs. Discrete values can be clinical variables like tumor stage or grade. In this workflow the focus is on TFs. Transcriptional regulatory programs are inferred using a hierarchical decision-tree model. The regulator assigned to each module consists of the set of TFs for which the expression profiles best explain all or part of the conditions. TFs receive a regulatory score reflecting the statistical confidence with which a TF regulates genes in the cluster. The collection of the regulatory scores for each TF is then converted into a global score. Finally, the TFs are sorted by their scores to construct a ranked list of potential regulators.

Step 2.2.2 - Regulatory Impact Factor (RIF) analysis

RIF analysis was initially developed to identify TFs that contribute to the differential expression in a particular condition, although the TF itself is not differentially expressed [51]. RIF is based on the differential correlation between a TF and the genes differentially expressed (DE) under two conditions. To compute a regulatory confidence score, it integrates three sources of information into a single measure: (a) the change in correlation between the TF and the DE genes, referred to as differential wiring; (b) the amount of differential expression of DE genes; and (c) the abundance of DE genes under the two conditions.

It assigns a score (RIF1) to those TFs that are consistently most differentially co-expressed with the highly abundant and highly expressed DE genes, and another score (RIF2) to those TFs with the most altered ability to predict the abundance of DE genes [51].

Step 2.3 - Gene significance measures for module ranking

RMaNI uses two measures to rank the modules for each subtype, average gene significance (GS) and modScore. The average gene significance focuses on the differential expression of genes in two conditions in each module and the modScore represents the overall correlation between genes in each module. RMaNI combines these two scores into a single score referred as a standard score:

averageGS = average(-log10(DE pvalue of a gene))

modScore = sum(abs(correlation of genes))/choose (ngenes,2))

standardScore = 1 - ((1 - averageGS])*(1 - modScore))

Stage 3 - Integration of transcriptional module networks and topological analysis

At this stage, the workflow combines all subtype-specific modules and regulators to build a transcriptional module network. In the topological analysis of such a module network, RMaNI calculates the overlap of TFs, TGs and interactions across subtypes, and generates node and edge attributes to aid in visualization.

Step 3.1 - Functional enrichment analysis of the inferred modules

The genes in each of the modules are subjected to a functional GO enrichment analysis using BiNGO [71]. Significantly enriched GO terms are detected by a hypergeometric test with adjusted Benjamini-Hochberg False Discovery Rate (FDR) [72] correction at significance level 0.05 against the all other genes in the network as a background.

Step 3.2 - Cluster similarity measures

To visualize the similarities between different modules, RMaNI uses the Jaccard similarity index as an external measure and Biological Process (BP) and Molecular Function (MF) as biological measures. The Jaccard index is calculated as the number of unique genes common to two clusters divided by the total number of unique genes in two sets. The BP and MF similarity measures are calculated by the GOSemSim package in Bioconductor [73].

Step 3.3 - Visualization

Throughout the analysis a number of figures are generated for data visualization, including the representation of inferred modules, significance measures calculated for each module, and overlaps of TFs, TGs and interactions across all subtypes. To visualize the network, the workflow exports interactions to a Cytoscape [47] -compatible file. Node attributes such as subtype and module memberships, number of modules regulated by a TF, GO annotations, and edge attributes such as subtype membership of an interaction, and regulatory score for an edge, are also provided for further exploration.

Application of RMaNI to hepatocellular carcinoma

To demonstrate the utility of RMaNI, we applied this workflow to hepatocellular carcinoma dataset (GSE14323) [74], containing normal tissues and three disease conditions: pre-malignant (cirrhosis), cirrhosis with hepatocellular carcinoma (cirrhosisHCC), and hepatocellular tumor (HCC). We investigated the ability of RMaNI to infer condition-specific transcriptional module networks, find common and unique TFs and regulatory interactions to examine the genetic architecture and ultimately to understand the differences and similarities between conditions. Below, we present the results for the individual steps to demonstrate the workflow.

Dataset

We used a Robust Multiarray Averaging (RMA) normalised and standardised hapatocellular carcinoma microarray gene expression dataset, based on 115 samples (Table 1).

Inference of module networks in hepatocellular carcinoma conditions

We selected top 4000 differentially expressed genes (based on BH-adjusted p-value) between normal and three conditions to infer the modules as described in the workflow. For this step RMaNI uses the LeMoNe algorithm. Michoel et al. [65] evaluated performance of the LeMoNe against state-of-the-art method *genomica*, and Smet and Marchal [7] compared LeMoNe against other network inference methods. For each pair of the normal-to-condition datasets (Table 2), 10 clustering solutions were generated. For each run LeMoNe used the default setting of 50 burn-ins and 100 Gibbs sampling steps, where the minimum number of genes in a cluster was set to 4. The default probability score cut-off of 0.2 was used uniquely assign genes to clusters.

Table 3 summarizes the clustering results. For each condition, it shows the different number of clusters generated, with their number of genes, maximum and minimum

Table 1 Description of the hepatocellular carcinoma microarray dataset.

Dataset	No. of samples in each condition				Platform
	Normal	Cirrhosis	CirrhosisHCC	HCC	
GSE14323 115 samples	19	41	17	38	HG-U133A (12079 probes)

In the next step, we input dataset to the LeMoNe algorithm to infer modules.

Table 2 Summary of the datasets used in the study, five sets of normal and subtype pairs data were input to LeMoNe.

Datasets	No. of DE Genes	No. of Samples
Normal + cirrhosis	4000	60
Normal + cirrhosisHCC	4000	36
Normal + HCC	4000	57

module sizes. We also performed a GO enrichment analysis on each module using BiNGO to measure the functional coherence of genes in the modules. Table 3 also shows the total number of modules, in each subtype, with at least one significant GO category enriched (BH-adjusted p-value 0.05). For instance, in cirrhosis, RMaNI generated a set of 74 modules corresponding with a total of 3794 genes. The largest modules had 302 genes and the smallest 4. In the next step we identified the regulators of the modules.

Identification and ranking of regulators

To assign the potential regulators (TFs) to the inferred modules, two data-driven approaches were employed in RMaNI: LRP and RIF. This step resulted in the potential regulatory TFs ordered according to their LRP and RIF score. To find the most confident regulators for each cluster, RMaNI used the intersection of regulators identified by both methods and integrated both scores into one score (stdScore). Table 4 presents the TFs predicted that have a regulatory role in at least two conditions. For instance, it reveals that TF CBFB regulates at least one module in each of the cirrhosis, cirrhosisHCC and HCC conditions and has 557 interactions across the three conditions. Previous studies were limited to a set of prior candidate TFs only, e.g. differentially expressed TFs or TFs involved in a particular pathway but considering the fact that the detection of DE TFs from expression data is limited due to their low and sparse expression levels, RMaNI uses all the TFs of a species (human in this study) without the need of prior TF identification. However, its applicability in organisms without known TFs will largely be determined by the entirety of TF databases and annotations, which are expected to improve over time with advances in ChIP-chip and ChIP-seq studies. Other continuous regulatory factors such as microRNAs, signal transducers, kinases and discrete regulatory factors such as clinical parameter, e.g. stage, grade or treatments can also be used.

Identification of modules with the highest DE and correlation

To identify the modules with high DE as well as high correlation for each condition (referred as best modules), we ordered the standard score, generated from averagGS and modScore, and for each module the workflow detected the knee-point (the maximum inflection point of a graph) from standard score to select the best modules. Table 5 shows the total number of best modules selected for each condition and the number of TFs and target genes in selected modules. For instance, in cirrhosis, 7 modules corresponding to 200 genes were selected. The 200 genes include 191 TGs and 9 TFs.

Network analysis

We aggregated all the module networks inferred for each condition to construct an overall network. For this purpose, RMaNI generates the network around the regulators predicted with highest confidence according to stdScore. The generated hepatocellular carcinoma network includes 24 TFs and 557 TGs connected by 5897 edges. We found 144 nodes unique to cirrhosis, 342 nodes unique to cirrhosisHCC, and 71 nodes unique to HCC. 1296, 4104 and 497 edges were unique to cirrhosis, cirrhosisHCC and HCC conditions, respectively. Previous approaches do not identify unique or shared TFs between modules, and between subtypes or conditions. By contrast, in this analysis we performed the analysis of convergent and divergent regulatory programs via TF overlap analysis. Figure 2 illustrates the TFs overlap across three conditions. We found one TF (CBFB) associated with all the three conditions, two TFs (TCF4 and USF2) associated with two conditions (Table 4) and 21 TFs were unique to one condition.

Network visualization

We imported the inferred module network in Cytoscape for visualization and exploration. For demonstration of topological analysis of inferred network, we extracted a sub-network of 70 nodes (TFs and TGs). Figure 3 shows hepatocellular carcinoma sub-network which includes 6 TFs and 64 TGs connected by 110 edges. Nodes and edges are rendered as per different evidences. For instance, node

Table 3 Summary of gene clustering results.

Conditions	No. of Modules	No. of Genes	Max Module Size	Min Module Size
Normal + cirrhosis	74	3794	302	4
Normal + cirrhosisHCC	59	3813	342	4
Normal + HCC	78	3772	219	4

Table 4 TFs that are predicted to have a regulatory role in at least two conditions.

TFs	Conditions	TGs in cirrhosis	TGs in cirrhosisHCC	TGs in HCC	Total Edges
CBFB	cirrhosis, cirrhosisHCC, HCC	144	342	71	557
TCF4	cirrhosis, HCC	144	0	71	215
USF2	cirrhosis, cirrhosisHCC	144	342	0	486

Remaining TFs are unique to individual conditions.

Table 5 Summary of the modules with highest DE and correlation (best modules).

Conditions	Total Modules	No. of best Modules	No. of Genes	No. of TFs	No. of TGs
cirrhosis	74	7	200	9	191
cirrhosisHCC	59	6	183	11	172
HCC	78	6	255	30	225
Total	211	18	638	50	588
Unique	211	50	548	47	548

Best modules were selected, for each condition, from all the modules inferred.

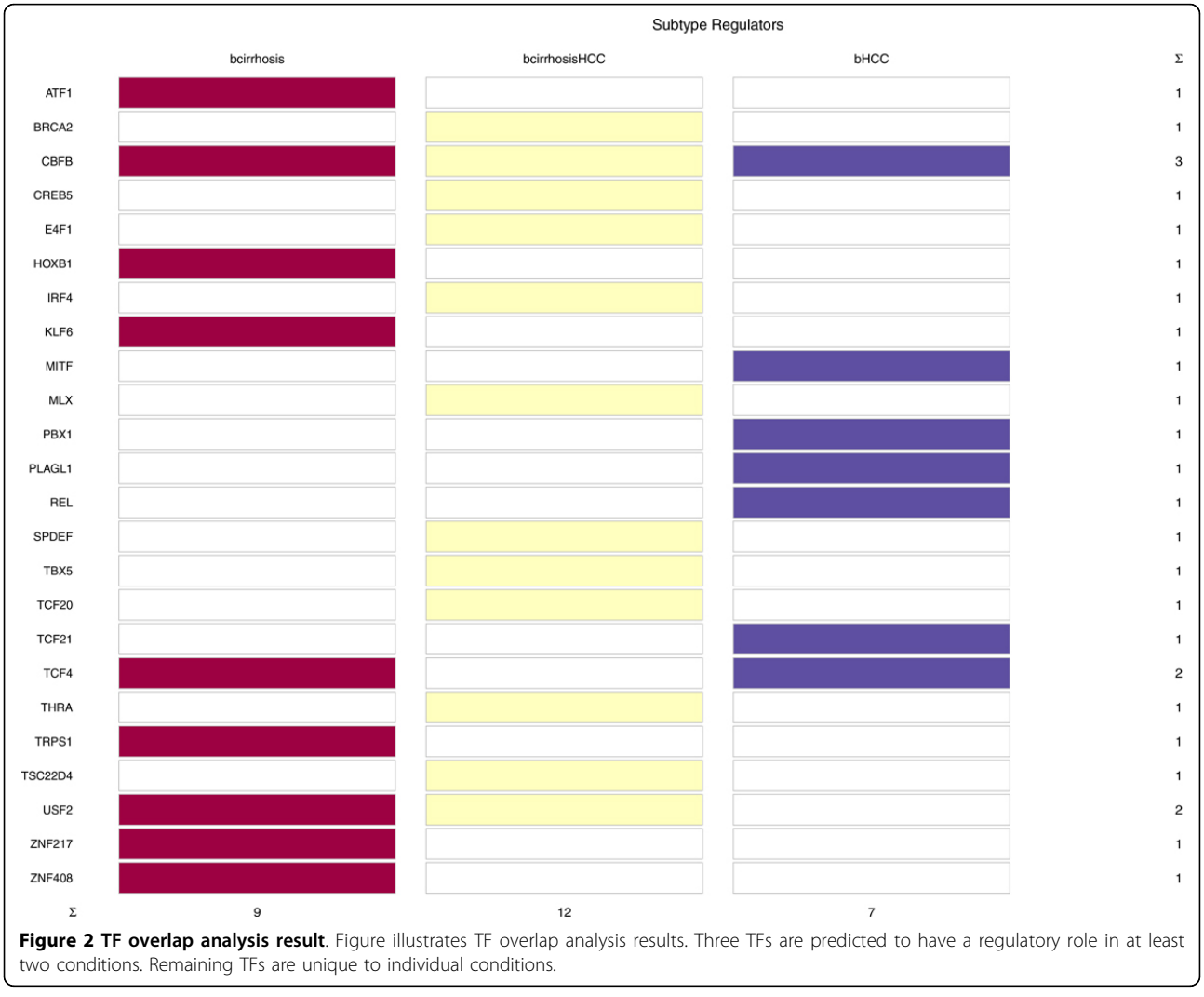

Figure 2 TF overlap analysis result. Figure illustrates TF overlap analysis results. Three TFs are predicted to have a regulatory role in at least two conditions. Remaining TFs are unique to individual conditions.

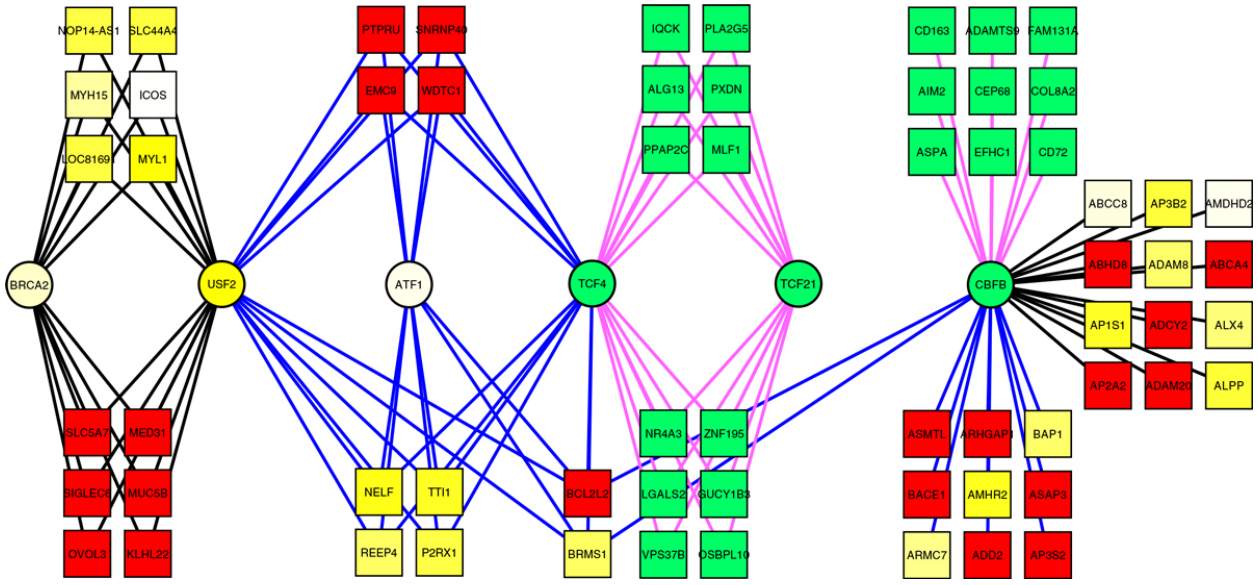

Figure 3 Hepatocellular carcinoma transcriptional sub-network. The hepatocellular carcinoma sub-network showing TGs (rectangles) and TFs (circles). Node colour gradient: red, up-regulation; green, down-regulation; yellow, no-change. Edge colors: blue, cirrhosis; black, cirrhosisHCC, cyan, HCC.

shape represents its type (TF or TG), and node colour gradient represents DE in cirrhosis against normal tissues. Edge colour represents condition membership. Figure shows distinct modules regulated by different TFs. It also illustrates the TF overlap analysis result (Table 4), for instance, CBFB, TCF4 and USF2 regulates the TGs in at least two conditions whereas TCF21, ATF1 and BRCA2 are predicted to have regulatory role only in one of the conditions.

Conclusions

We have presented the RMaNI workflow, developed for the end-user perspective of a biologist or clinician. It provides an easy-to-use interface to a comprehensive, integrated suite of tools for the inference of condition or subtype-specific transcriptional module networks and their analysis. We described the RMaNI workflow and applied it to hepatocellular carcinoma data. We demonstrated that identifying the transcriptional module network, and analysing and visualizing the inferred network, can give insight into the common as well as unique regulatory architecture underlying different disease conditions. We anticipate integrating additional tools and workflows in future to meet the distinct needs of researchers confronting the complexity of cancer.

Additional material

Additional File 1: RMaNI User Manual

List of abbreviations used

TF: Transcription Factor; TG: Target Gene; LeMoNe: Learning Module Networks; RIF: Regulatory Impact Factors; clues: clustering based on local shrinking; PAM: Partitioning Around Medoids; AGNES: AGglomerative NESting; SOTA: Self-Organising Tree Algorithm; MCLUST: Model based clustering; LRP: LeMoNe's regulatory program; averageGS: average Gene Significance; GO: Gene Ontology.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

PBM developed the RMaNI framework, wrote the code and the manuscript. SRM, MJD, MAR advised on design and features of RMaNI, provided overall scientific and technical guidance, and assisted with the manuscript. All authors read and approved the final manuscript.

Acknowledgements

We thank Mr Gavin Graham, Dr Gerald Hartig and Mr. A.Varlokov from Bioinformatics Resource Australia-EMBL for high-performance computing and web-development support; Dr Toni Reverter and Dr Sriganesh Srihari for helpful discussions; Dr Richard Newton for help with Rwui; and the R/Bioconductor research community, who have made their programs and source codes publicly available. Computational resources were provided by National Computational Infrastructure Specialised Facility in Bioinformatics. Access to Transfac was provided by QFAB Bioinformatics through Australian Research Council grant LE098933. PBM, SRM, MJD and MAR acknowledge support of Australian Research Council grants CE0348221, DP110103384 and LE0989334.

Declarations

Publication of this article was funded by The University of Queensland. This article has been published as part of *BMC Bioinformatics* Volume 14 Supplement 16, 2013: Twelfth International Conference on Bioinformatics (InCoB2013): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcbioinformatics/supplements/14/S16.

Authors' details

[1]Institute for Molecular Bioscience, The University of Queensland, 306 Carmody Road, St Lucia, Brisbane, Queensland 4072, Australia. [2]Australian

Research Council Centre of Excellence in Bioinformatics, The University of Queensland, 306 Carmody Road, St Lucia, Brisbane, Queensland 4072, Australia.

Published: 22 October 2013

References

1. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation.** *Cell* 2011, **144(5)**:646-674.
2. Gentles AJ, Gallahan D: **Systems biology: confronting the complexity of cancer.** *Cancer Res* 2011, **71(18)**:5961-5964.
3. Barabasi AL, Gulbahce N, Loscalzo J: **Network medicine: a network-based approach to human disease.** *Nat Rev Genet* 2011, **12(1)**:56-68.
4. Madhamshettiwar PB, Maetschke SR, Davis MJ, Reverter A, Ragan MA: **Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets.** *Genome Med* 2012, **4(5)**:41.
5. He F, Chen H, Probst-Kepper M, Geffers R, Eifes S, Del Sol A, Schughart K, Zeng AP, Balling R: **PLAU inferred from a correlation network is critical for suppressor function of regulatory T cells.** *Mol Syst Biol* 2012, **8**:624.
6. Choi JK, Yu U, Yoo OJ, Kim S: **Differential coexpression analysis using microarray data and its application to human cancer.** *Bioinformatics* 2005, **21(24)**:4348-4355.
7. De Smet R, Marchal K: **Advantages and limitations of current network inference methods.** *Nat Rev Micro* 2010.
8. Jérôme A, Annie R, Benoit M, Jean-Luc G: **Transcriptional Network Inference from Functional Similarity and Expression Data: A Global Supervised Approach.** *Statistical Applications in Genetics and Molecular Biology* 2012, **11(1)**.
9. Cerulo L, Elkan C, Ceccarelli M: **Learning gene regulatory networks from only positive and unlabeled data.** *BMC Bioinformatics* 2010, **11**:228.
10. de Jong H: **Modeling and simulation of genetic regulatory systems: a literature review.** *J Comput Biol* 2002, **9(1)**:67-103.
11. Maetschke SR, Madhamshettiwar PB, Davis MJ, Ragan MA: **Supervised, semi-supervised and unsupervised inference of gene regulatory networks.** *arXiv* 2013, arXiv:1301.1083.
12. Stolovitzky G, Prill RJ, Califano A: **Lessons from the DREAM2 Challenges.** *Ann N Y Acad Sci* 2009, **1158**:159-195.
13. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G: **Revealing strengths and weaknesses of methods for gene network inference.** *Proc Natl Acad Sci USA* 2010, **107(14)**:6286-6291.
14. Prill RJ, Marbach D, Saez-Rodriguez J, Sorger PK, Alexopoulos LG, Xue X, Clarke ND, Altan-Bonnet G, Stolovitzky G: **Towards a Rigorous Assessment of Systems Biology Models: The DREAM3 Challenges.** *PloS one* 2010, **5**:(2):e9202.
15. Oltvai ZN, Barabasi AL: **Systems biology. Life's complexity pyramid.** *Science* 2002, **298(5594)**:763-764.
16. Segal E, Friedman N, Koller D, Regev A: **A module map showing conditional activity of expression modules in cancer.** *Nat Genet* 2004, **36(10)**:1090-1098.
17. Michoel T, De Smet R, Joshi A, Van de Peer Y, Marchal K: **Comparative analysis of module-based versus direct methods for reverse-engineering transcriptional regulatory networks.** *BMC Syst Biol* 2009, **3**:49.
18. Ihmels J, Friedlander G, Bergmann S, Sarig O, Ziv Y, Barkai N: **Revealing modular organization in the yeast transcriptional network.** *Nat Genet* 2002, **31(4)**:370-377.
19. Bonneau R: **Learning biological networks: from modules to dynamics.** *Nat Chem Biol* 2008, **4(11)**:658-664.
20. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N: **Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.** *Nat Genet* 2003, **34(2)**:166-176.
21. Wong DJ, Chang HY: **Learning more from microarrays: insights from modules and networks.** *The Journal of investigative dermatology* 2005, **125(2)**:175-182.
22. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, *et al*: **Computational discovery of gene modules and regulatory networks.** *Nat Biotechnol* 2003, **21(11)**:1337-1342.
23. Jain AK, Murty MN, Flynn PJ: **Data clustering: a review.** *ACM Comput Surv* 1999, **31(3)**:264-323.
24. Dalton L, Ballarin V, Brun M: **Clustering algorithms: on learning, validation, performance, and applications to genomics.** *Current genomics* 2009, **10(6)**:430-445.
25. Thalamuthu A, Mukhopadhyay I, Zheng X, Tseng GC: **Evaluation and comparison of gene clustering methods in microarray analysis.** *Bioinformatics* 2006, **22(19)**:2405-2412.
26. Miller CA, Settle SH, Sulman EP, Aldape KD, Milosavljevic A: **Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors.** *BMC Med Genomics* 2011, **4**:34.
27. Langfelder P, Horvath S: **WGCNA: an R package for weighted correlation network analysis.** *BMC Bioinformatics* 2008, **9(1)**:559.
28. Zhang B, Horvath S: **A General Framework for Weighted Gene Co-Expression Network Analysis.** *Statistical Applications in Genetics and Molecular Biology* 2005, **4(1)**.
29. Winden KD, Karsten SL, Bragin A, Kudo LC, Gehman L, Ruidera J, chwind DH, Engel J Jr: **A systems level, functional genomics analysis of chronic epilepsy.** *PloS one* 2011, **6(6)**:e20763.
30. Rosen EY, Wexler EM, Versano R, Coppola G, Gao F, Winden KD, Oldham MC, Martens LH, Zhou P, Farese RV Jr, *et al*: **Functional genomic analyses identify pathways dysregulated by progranulin deficiency, implicating Wnt signaling.** *Neuron* 2011, **71(6)**:1030-1042.
31. Saris C, Horvath S, van Vught P, van Es M, Blauw H, Fuller T, Langfelder P, DeYoung J, Wokke J, Veldink J, *et al*: **Weighted gene co-expression network analysis of the peripheral blood from Amyotrophic Lateral Sclerosis patients.** *BMC Genomics* 2009, **10(1)**:405.
32. Horvath S, Zhang B, Carlson M, Lu KV, Zhu S, Felciano RM, Laurance MF, Zhao W, Qi S, Chen Z, *et al*: **Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target.** *Proc Natl Acad Sci USA* 2006, **103(46)**:17402-17407.
33. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data.** *Bioinformatics* 2001, **17(10)**:977-987.
34. Joshi A, Van de Peer Y, Michoel T: **Analysis of a Gibbs sampler method for model-based clustering of gene expression data.** *Bioinformatics* 2008, **24(2)**:176-183.
35. McNicholas PD, Murphy TB: **Model-based clustering of microarray expression data via latent Gaussian mixture models.** *Bioinformatics* 2010, **26(21)**:2705-2712.
36. Lemmens K, Dhollander T, De Bie T, Monsieurs P, Engelen K, Smets B, Winderickx J, De Moor B, Marchal K: **Inferring transcriptional modules from ChIP-chip, motif and microarray data.** *Genome Biol* 2006, **7(5)**:R37.
37. Reimand J, Tooming L, Peterson H, Adler P, Vilo J: **GraphWeb: mining heterogeneous biological networks for gene modules with functional significance.** *Nucleic Acids Res* 2008, **36(Web Server)**:W452-459.
38. Qi J, Michoel T, Butler G: **An integrative approach to infer regulation programs in a transcription regulatory module network.** *J Biomed Biotechnol* 2012, **2012**:245968.
39. McCord RP, Berger MF, Philippakis AA, Bulyk ML: **Inferring condition-specific transcription factor function from DNA binding and gene expression data.** *Mol Syst Biol* 2007, **3**:100.
40. Baitaluk M, Kozhenkov S, Ponomarenko J: **An integrative approach to inferring gene regulatory module networks.** *PLoS One* 2012, **7(12)**:e52836.
41. Vega VB, Woo XY, Hamidi H, Yeo HC, Yeo ZX, Bourque G, Clarke ND: **Inferring Direct Regulatory Targets of a Transcription Factor in the DREAM2 Challenge.** *Challenges of Systems Biology: Community Efforts to Harness Biological Complexity* 2009, **1158**:215-223.
42. Hurley D, Araki H, Tamada Y, Dunmore B, Sanders D, Humphreys S, Affara M, Imoto S, Yasuda K, Tomiyasu Y, *et al*: **Gene network inference and visualization tools for biologists: application to new human transcriptome datasets.** *Nucleic Acids Res* 2011.
43. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, *et al*: **TRANSFAC(R): transcriptional regulation, from patterns to profiles.** *Nucl Acids Res* 2003, **31(1)**:374-378.
44. Quandt K, Frech K, Karas H, Wingender E, Werner T: **MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data.** *Nucleic Acids Res* 1995, **23(23)**:4878-4884.
45. Durinck S, Moreau Y, Kasprzyk A, Davis S, De Moor B, Brazma A, Huber W: **BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.** *Bioinformatics* 2005, **21(16)**:3439-3440.
46. Gordon S: **Limma: linear models for microarray data.** In *Bioinformatics and Computational Biology Solutions using R and Bioconductor.* New York: Springer;Gentleman R, Carey V, Dudoit S, Irizarry R, Huber W 2005:397-420.

47. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res* 2003, **13(11)**:2498-2504.
48. Huang DW, Sherman BT, Lempicki RA: **Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.** *Nat Protocols* 2008, **4(1)**:44-57.
49. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37(1)**:1-13.
50. Joshi A, De Smet R, Marchal K, Van de Peer Y, Michoel T: **Module networks revisited: computational assessment and prioritization of model predictions.** *Bioinformatics* 2009, **25(4)**:490-496.
51. Reverter-Gomez A, Hudson NJ, Nagaraj SH, Perez-Enciso M, Dalrymple BP: **Regulatory Impact Factors: Unraveling the transcriptional regulation of complex traits from expression data.** *Bioinformatics* 2010, btq051.
52. Newton R, Wernisch L: **Rwui: A web application to create user friendly web interfaces for R scripts.** *R News* 2007, **7(2)**:32-35.
53. R Development Core Team: **R: A Language and Environment for Statistical Computing.** *Vienna, Austria: R Foundation for Statistical Computing* 2012.
54. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10)**: R80.
55. Fang C, Weiliang Q, Ruben HZ, Ross L, W X: **clues: An R Package for Nonparametric Clustering Based on Local Shrinking.** *Journal of Statistical Software* 2010, **33(4)**:1-16.
56. Fraley C, Raftery AE: **MCLUST Version 3: An R Package for Normal Mixture Modeling and Model-Based Clustering.** *Seattle, WA 98195-4322 USA: Department of Statistics, University of Washington* 2006.
57. Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K: **cluster: Cluster Analysis Basics and Extensions.** *R package version 1143* 2012.
58. Rand WM: **Objective Criteria for the Evaluation of Clustering Methods.** *Journal of the American Statistical Association* 1971, **66(336)**:846-850.
59. Datta S: **clValid: An R Package for Cluster Validation.** *Journal of Statistical Software* 2008, **25(4)**.
60. Hartigan JA, Wong MA: **Algorithm AS 136: A k-means clustering algorithm.** *Applied Statistics* 1979, **28(1)**:100-108.
61. Kaufman L, Rousseeuw P: **Finding Groups in Data: An Introduction to Cluster Analysis.** 1990, Wiley-Interscience.
62. Dopazo J, Carazo JM: **Phylogenetic Reconstruction Using an Unsupervised Growing Neural Network That Adopts the Topology of a Phylogenetic Tree.** *J Mol Evol* 1997, **44(2)**:226-233.
63. Yin L, Huang CH, Ni J: **Clustering of gene expression data: performance and similarity analysis.** *BMC Bioinformatics* 2006, , **7 Suppl 4:** S19.
64. Bonnet E, Tatari M, Joshi A, Michoel T, Marchal K, Berx G, Van de Peer Y: **Module network inference from a cancer gene expression data set identifies microRNA regulated modules.** *PloS one* 2010, **5(4)**:e10162.
65. Michoel T, Maere S, Bonnet E, Joshi A, Saeys Y, Van den Bulcke T, Van Leemput K, van Remortel P, Kuiper M, Marchal K, *et al*: **Validating module network learning algorithms using simulated data.** *BMC Bioinformatics* 2007, **8(Suppl 2)**:S5.
66. Bonnet E, Michoel T, Van de Peer Y: **Prediction of a gene regulatory network linked to prostate cancer from gene expression, microRNA and clinical data.** *Bioinformatics* 2010, **26(18)**:i638-i644.
67. Vermeirssen V, Joshi A, Michoel T, Bonnet E, Casneuf T, Van de Peer Y: **Transcription regulatory networks in Caenorhabditis elegans inferred through reverse-engineering of gene expression profiles constitute biological hypotheses for metazoan development.** *Mol Biosyst* 2009.
68. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM: **A census of human transcription factors: function, expression and evolution.** *Nat Rev Genet* 2009, **10(4)**:252-263.
69. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, Akalin A, Schmeier S, Kanamori-Katayama M, Bertin N, *et al*: **An atlas of combinatorial transcriptional regulation in mouse and man.** *Cell* 2010, **140(5)**:744-752.
70. Schaefer U, Schmeier S, Bajic VB: **TcoF-DB: dragon database for human transcription co-factors and transcription factor interacting proteins.** *Nucleic Acids Res* 2011, **39(Database)**:D106-110.
71. Maere S, Heymans K, Kuiper M: **BiNGO: a Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks.** *Bioinformatics* 2005, **21(16)**:3448-3449.
72. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society Series B (Methodological)* 1995, **57(1)**:289-300.
73. Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S: **GOSemSim: an R package for measuring semantic similarity among GO terms and gene products.** *Bioinformatics* 2010, **26(7)**:976-978.
74. Mas VR, Maluf DG, Archer KJ, Yanek K, Kong X, Kulik L, Freise CE, Olthoff KM, Ghobrial RM, McIver P, *et al*: **Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma.** *Mol Med* 2009, **15(3-4)**:85-94.

doi:10.1186/1471-2105-14-S16-S14
Cite this article as: Madhamshettiwar *et al.*: RMaNI: Regulatory Module Network Inference framework. *BMC Bioinformatics* 2013 **14**(Suppl 16):S14.