SOFTWARE TOOL ARTICLE

REVISED MSF: Modulated Sub-graph Finder [version 3; peer review: 3 approved]

Mariam R. Farman [1], Ivo L. Hofacker[1], Fabian Amman [1,2]

[1]Institute for Theoretical Chemistry,Theoretical Biochemistry Group,, University of Vienna, Vienna, 1090, Austria
[2]Department of Chromosome Biology, Max F. Perutz Laboratories,, University of Vienna, Vienna, 1030, Austria

v3 **First published:** 29 Aug 2018, **7**:1346 (https://doi.org/10.12688/f1000research.16005.1)
Second version: 22 Mar 2019, **7**:1346 (https://doi.org/10.12688/f1000research.16005.2)
Latest published: 14 Apr 2019, **7**:1346 (https://doi.org/10.12688/f1000research.16005.3)

Abstract

High throughput techniques such as RNA-seq or microarray analysis have proven to be invaluable for the characterizing of global transcriptional gene activity changes due to external stimuli or diseases. Differential gene expression analysis (DGEA) is the first step in the course of data interpretation, typically producing lists of dozens to thousands of differentially expressed genes. To further guide the interpretation of these lists, different pathway analysis approaches have been developed. These tools typically rely on the classification of genes into sets of genes, such as pathways, based on the interactions between the genes and their function in a common biological process. Regardless of technical differences, these methods do not properly account for cross talk between different pathways and most of the methods rely on binary separation into differentially expressed gene and unaffected genes based on an arbitrarily set *p*-value cut-off.

To overcome this limitation, we developed a novel approach to identify concertedly modulated sub-graphs in the global cell signaling network, based on the DGEA results of all genes tested. To this end, expression patterns of genes are integrated according to the topology of their interactions and allow potentially to read the flow of information and identify the effectors. The described software, named Modulated Sub-graph Finder (MSF) is freely available at https://github.com/Modulated-Subgraph-Finder/MSF.

Keywords

Differential gene expression analysis, pathway analysis, combining p-value, cell signalling network

Open Peer Review

Reviewer Status ✓ ✓ ✓

	Invited Reviewers		
	1	**2**	**3**
REVISED **version 3** published 14 Apr 2019	✓ report	✓ report	✓ report
REVISED **version 2** published 22 Mar 2019	? report		? report
version 1 published 29 Aug 2018	✗ report	✓ report	? report

1 **Guanming Wu**, Oregon Health and Science University (OHSU), Portland, USA

2 **Stefanie Widder**, Research Lab of Infection Biology, Waehringer Guertel 18-20, 1090 Vienna, Austria

3 **Haibo Liu**, Iowa State University, Ames, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Ivo L. Hofacker (ivo@tbi.univie.ac.at)

Author roles: Farman MR: Data Curation, Investigation, Methodology, Software, Validation, Visualization, Writing – Original Draft Preparation; **Hofacker IL**: Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Review & Editing; **Amman F**: Conceptualization, Project Administration, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was funded by the FWF ("Fonds zur Förderung der wissenschaftlichen Forschung") within the project Internationalen Kooperationsprojektes - Intl cooperation Project (Joint Project - Lead Agency Verfahren) with the project number (I 2353-B22). The grant was assigned to ILH. FA was funded by the Austrian Science Fund (FWF) project SFB F43.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Farman MR, Hofacker IL and Amman F. **MSF: Modulated Sub-graph Finder [version 3; peer review: 3 approved]** F1000Research 2019, **7**:1346 (https://doi.org/10.12688/f1000research.16005.3)

First published: 29 Aug 2018, **7**:1346 (https://doi.org/10.12688/f1000research.16005.1)

REVISED **Amendments from Version 2**

Version 3 of our manuscript addresses the points from the peer-reviewers, whom we thank for their constructive feedback. We have updated the manuscript and published a new version on MSF (GitHub and Zenodo).

1. Human genes names are in upper-case for all output files of MSF.
2. For source reliability the *t*-test is performed on log transformed p-values.

Furthermore, we corrected typos, improved the documentation, and rephrased some sentences. The project number in the funding section has been corrected from 'I 1988-B22' to 'I 2353-B22'. Figure 2 nodes label size have been increased.

See referee reports

Introduction

High throughput sequencing techniques have been widely used to yield differentially expressed genes (DEG)[1]. The changes in transcript abundance are measured, e.g. by next generation sequencing techniques and interpreted as an indicator of differential expression of genes. DEGs can be used to gain insights into the mechanisms underlying differences between conditions of samples, such as healthy versus infected. Differential gene expression analysis (DGEA) informs about the magnitude of expression changes, which are often expressed as log-fold change. The sign of log-fold change and the confidence level of observing an authentic change, often expressed as *p*-value. The information from DEGs is further interpreted to extract meaningful biological insights. For example, genes that could be involved in the response to a particular stimulus. To this end, pathway-based analysis has become an important tool to further interpret the results of a DGEA and to acquire understandings of the perturbations in a biological system. These pathway-based methods use predefined pathways or networks which are sets of genes with their interactions forming a functional unit. DEGs help to identify pathways or networks that may be altered during an infection, providing important information about diseases and its treatment process[2]. The expression measurements of the genes obtained from DGEA in combination with statistical methods and the predefined pathways are used to identify specifically modulated pathways and processes[3].

Well established resources for pathway annotations are KEGG (Kyoto Encyclopedia of Genes and Genomes)[4] and Reactome[5]. KEGG pathways is a branch of KEGG database that hosts a collection of manually drawn pathway maps representing the molecular interaction, reaction and relation networks of cellular functions. Similarly, Reactome is an open-source, manually curated, peer-reviewed database for signaling and metabolic molecules with their interactions formed into different biological pathways[5]. Both provide predefined pathways which are sets of genes and their interactions categorized into functional units.

Existing pathway-based analysis approaches use different research designs, which can be categorized into ORA (Over-representation analysis), FCS (Functional class scoring) and pathway topology based methods. All of which aim to find a subset of genes, e.g., significantly differentially expressed genes, genes associated with a certain pathway more often than expected given the total set of examined genes, e.g. the whole genome background. ORA is the first and the most basic method of pathway analysis[3]. It uses a DEG list with user defined cut-off for the log-fold change and *p*-value (most commonly using absolute log-fold change ≥ 2 and *p*-value ≤ 0.05). Subsequently, sets of genes associated with annotated pathways are tested for being over-represented in the set of DEGs. To this end, hyper-geometric distribution, chi-square tests, binomial probability or the Fisher's exact test are used, whereas, the information of the topology of genes in the pathways is neglected[6]. Furthermore, ORA assumes that the biological pathways are independent of each other and ignores the fact that they cross-talk and overlap[2,3].

Unlike ORA, FCS has no artificial cut-off to define a DEG list. FCS works in three steps. First it calculates the gene-level statistics including correlation of molecular measurements using differential expression of individual genes, ANOVA, t-test and Z-test. In the second step the statistics of individual genes in a pathway are transformed to an individual pathway-level statistic commonly using Kolmogorov-Smirnov statistic, mean or median. Finally the statistical significance of the pathway-level statistics is assessed. Although FCS overcomes some of the limitations from ORA, it still ignores the topology of genes in a pathway, cross-talk and overlap of the pathways[2,3]. Pathway topology based methods are similar to FCS except that they consider the topology of each gene during the gene-level statistics but still lacks to link the different functional pathways[2].

From another perspective, network based approaches do not categorize sets of genes into functional pathways, but they consider all interactions to be equal. Thereby, they avoid distinguishing arbitrarily between interactions within a pathway and interactions between pathways (i.e., cross-talk). With this they aim to identify subnetworks that show modulation between two conditions or upon a stimuli[7]. To find these active modules heuristics solutions like simulated annealing (SA), greedy methods, genetic algorithms (GA), network propagation and co-clustering methods are used[7]. jActiveModules has been the first of this kind using simulated annealing to find modulated sub-networks[8]. The benefit of omitting pathways is brought by reduced interpretability of the results due to the lack of functional labels on the networks.

On these grounds we propose a novel approach to make use of the rich gene and protein interaction annotation resources available and combining it to functional pathway annotations to gain additional insights from basic DGEA. To this, we start with the presupposition that expression of neighboring genes within a functional pathway are not independent from each other. Rather, they are often regulating each other's expression or are part of the same regulon[9]. We understand that the categorization of links between genes into labeled pathways is often an arbitrary one, given the extensive cross talk between different pathways. Although these categories have proven to be useful in many situations, they force a certain perspective onto the interpretation of novel data. Based on these two principles, we aim to find sub-graphs of connected genes within cell signaling network, which exhibit as

a whole significant differential expression changes. This approach differs in two main aspects from common pathway analysis tools. First, it does not aim to identify functional pathways enriched in differentially expressed genes, but detects sub-graphs or branches in a network graph (potentially spanning more than one functionally grouped pathway) which is coherently modulated. Second, it aims to improve the DGEA on the gene level, by collecting the information of neighboring genes, which as a whole might exhibit prominent enough signal to be called significantly modulated. All of this can be helpful to understand the cause and effect of a stimulus and might inform about potential points of intervention.

As input, information on functional links between genes provided by e.g. KEGG or Reactome and information on the differential expression status of single genes resulting from a DGEA, are required. As a result the analysis returns sub-graphs and their joint confidence scores, reflecting how the perturbation migrates through the network. Furthermore, the entry points of perturbation in the networks and overlap with conventional pathway categories are returned. To facilitate prioritization of the perturbation entry points, to each an impact score and a measure of its reliability are assigned. The impact score expresses the fraction of the sub-module being downstream of the entry point. The reliability is measured using a t-test on log transformed *p*-values of the immediate upstream and downstream genes. The output is prepared in a directed adjacency matrix file, convenient for visualization, e.g., with StringApp[10], available as a Cytoscape plug-in[11].

The proposed algorithm is named `Modulated Sub-graph Finder` (abbreviated `MSF`). `MSF` can help transform the information obtained from DGEA into comprehensible knowledge of signal transduction of genes, hence being a valuable complement to existing pathway based methods. `MSF` is freely accessible on GitHub under the terms of the Creative Commons Attribution 4.0 International License.

Methods

`MSF` was implemented as a Java program. It is developed as a novel heuristic approach to find concertedly modulated sub-graphs in networks of biological interactions. `MSF` does not use predefined gene sets grouped into functional units, but rather relies purely on the network of interacting genes. The input network consists of nodes corresponding to genes and edges representing interactions. Furthermore it utilizes comprehensive results from a differential gene expression analysis to discover the sub-graphs, or modules, which are as a whole modulated.

`MSF` uses the individual gene's *p*-values generated from the DGEA. The *p*-value expresses the probability that the null hypothesis of unmodified gene expression can't be rejected for a given statistical model. To find significantly modulated sub-graphs individual *p*-values of the vicinal genes in the global network are combined into a single combined *p*-value, using a statistical method for combining dependent *p*-values described by Hartung[12]. Hartung's method uses the inverse of standard normal distribution function. Using the inverse normal cumulative distribution function Φ^{-1}, individual gene *p*-values T_i are transformed to their corresponding normal score $t_i = \Phi^{-1}(1 - T_i)$ that is uniformly distributed on (0,1). Then using these normal scores, the correlation between genes is calculated $Cov(t_i, t_j) = \rho$. The normal scores and correlation are applied to the weighted inverse normal function to calculate the combined *p*-value $t(\rho)$ for all genes examined, namely the examined sub-graph

$$t(\rho) = \frac{\sum_{i=1}^{n} \lambda_i t_i}{\sqrt{(1-\rho)\sum_{i=1}^{n} \lambda_i^2 + \rho(\sum_{i=1}^{n} \lambda_i)^2}}$$

Lambda λ be the weights for each gene, currently all genes have equal weight, i.e. 1. The combined *p*-value $t(\rho)$ of a sub-graph will express the significance of all genes in the sub-graph being modulated together. The information from the different genes are used as, although not independent, replicated measurement of the behavior of the whole sub-graph. This potentially increases the power to detect also significant sub-graphs consisting of genes which are not significant on there own.

Overview of our method

To reduce the complexity to score all possible connected sub-graphs `MSF` applies a four step heuristic as described in the following. The proceeding identification of modulated sub-graphs from a network by `MSF` is presented as a flowchart diagram (Figure 1).

Initializing modulated sub-graphs. `MSF` constructs the first sub-graph starting with the genes associated with the lowest (most significant) *p*-value deduced from the DGEA. From this seed it tries to extend the sub-graph by adding directly neighboring genes, starting with the next most significant one. A single combined *p*-value is calculated for the extended sub-graph. If the combined *p*-value is smaller than the *p*-value of the original sub-graph, the extended sub-graph is accepted. This step is iteratively repeated until no further extension is accepted. In this case the process starts over with all remaining genes not yet in the significantly modulated sub-graph. This step identifies all simple sub-graphs that are modulated in the whole network.

Extending modulated sub-graphs. In the next step, we check if any of the initial modulated sub-graphs could further be extended beyond the immediate neighboring genes. Instead of testing single genes and their compatibility to be added, groups of genes are considered. If the combined *p*-value of the initial modulated sub-graph and the extension genes is smaller than the *p*-value of the initial sub-graph the extension is accepted. All possible extension paths up to *3* (default 2) genes at all nodes in the sub-graph are tested. Again, this step is iteratively repeated until no further genes are added to the significant differentially expressed sub-graphs. This step bridges small gaps of genes without a clear differential signal in the DGEA.

Merging modulated sub-graphs. After detection and extension of the modulated sub-graphs, each pair of so far identified sub-graphs is tested if its combination scores better than each on its own. The merging of the sub-graphs is done by depth first search traversal from one sub-graph to the other sub-graph. If the two

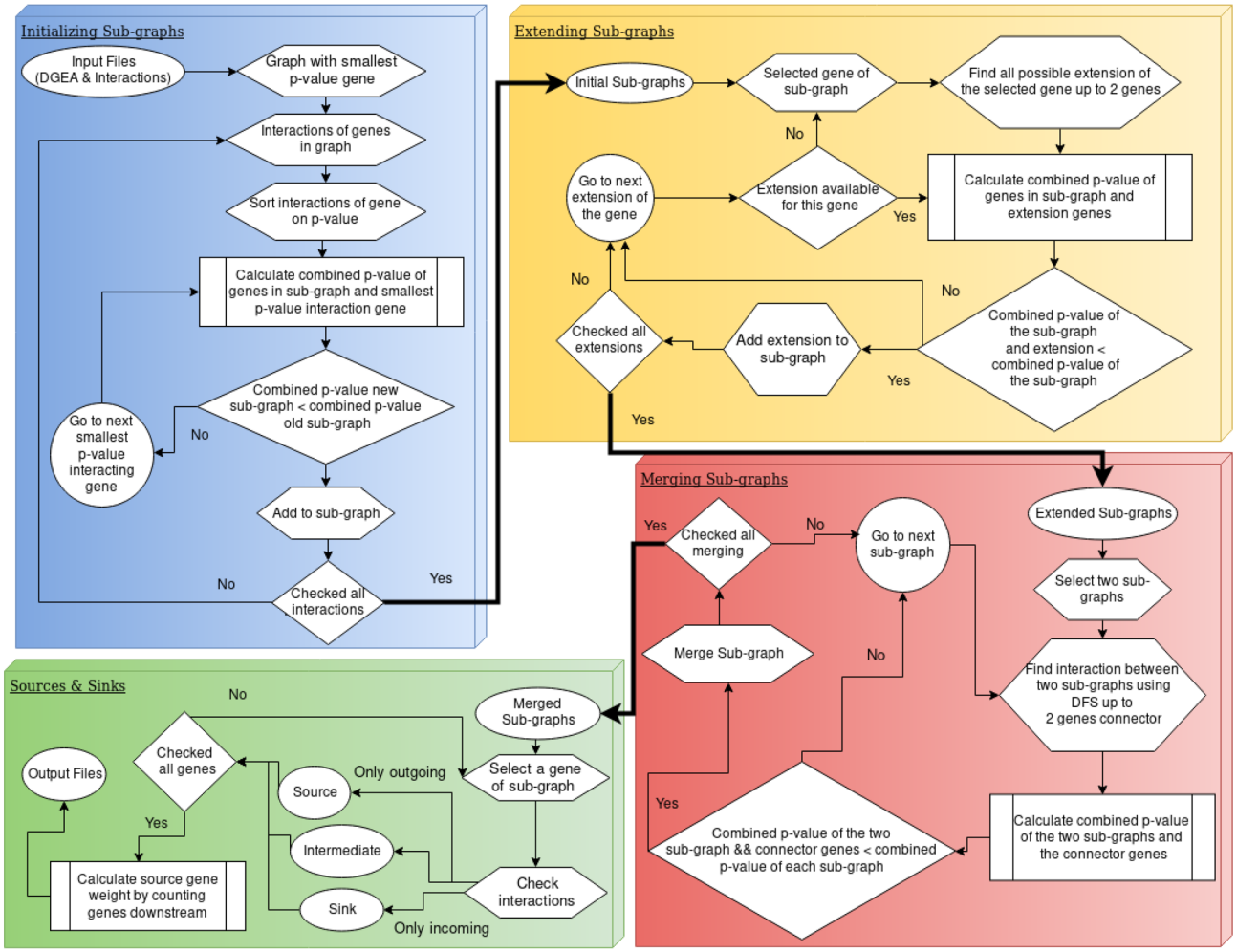

Figure 1. Graphical representation of the MSF heuristic approach to detect modulated sub-graphs in a global gene regulatory network.

sub-graphs merge with the connector of at most *3* genes (default 2 genes) and the combined *p*-value of the merged sub-graph including the bridging genes in between is less than the individual *p*-values of the two sub-graphs, the two sub-graphs are merged together to one bigger modulated sub-graph. This step is repeated iteratively until no sub-graphs can be merged anymore.

Finding sources & sinks. In a last post processing step MSF identifies the trigger points of the modulated sub-graphs. These trigger genes are the sources of the sub-graphs with only outgoing edges. These genes can be interpreted as the possible entry points of perturbation from where the stimulus causes downstream effects. Each individual source is given an impact score, expressing the relative number of downstream genes within the corresponding sub-graph directly connected by directed links. This score can be interpreted as an upper limit of how much of the sub-graphs' perturbation could have been introduced by the respective source, and thereby could be helpful to prioritize different identified sources for larger sub-graphs. In the same spirit as sources, the most downstream genes of the modulated sub-graphs are identified and defined as sinks. Due to loops not all sub-graphs are guaranteed to have sources or sinks. The reliability of each source is inspected using a t-test on log transformed *p*-values. The significant difference between the two groups, genes downstream the source and the genes upstream of the source is determined. This would help to assess if the source identified is reliable and indeed marks the border between two different regulation regimes.

MSF output. MSF generates a directed network file as an output, containing complete directed interactions of all modulated sub-graphs identified. This file could be imported into Cytoscape[11] for visualization. Additionally, a file containing details on all sources and sinks for all modulated sub-graphs is reported. Furthermore, for facilitated visualization in Cytoscape, a node attribute file is provided, containing the source weightage and the log-fold changes of all considered genes.

Operation. The only system requirements to run MSF are Java version 8 and JDK 1.8 or above. The few package dependencies are already been added to the release. The runtime of MSF is less than 10 minutes. To run MSF, the user must provide two text files, one containing the DGEA and the second one containing directed interactions in an adjacency matrix format file. Example files and a detailed tutorial to use MSF has been provided on GitHub https://github.com/Modulated-Subgraph-Finder/MSF.

Results

Case study

To demonstrate its usefulness, MSF is applied to RNA-seq data set of primary human monocyte derived macrophages (MDMs) infected with Ebola virus (GSE84188)[13]. Ebola Virus (EBOV) belongs to the Filoviridea family: filamentous, enveloped and single stranded RNA viruses. EBOV causes hemorrhagic fever in humans, inducing the host innate and adaptive immune response to be unable to control virus infection[14]. Currently, there are no approved antiviral drugs for the treatment of Ebola virus infection[15,16]. The initial targets of EBOV are the macrophages and dendritic immune cells[16,17]. EBOV inhibits the critical innate immune response of the host, which includes inhibiting the activation of alpha/beta interferon (IFN-α/β)[14,15,18]. It has been proposed that IFN-α/β should be tested against Ebola for its antiviral activity through clinical trials[15]. Ebola infection data was selected to test the approach because it has been well recognized for the last several decades, and vast literature is available for the pathogenesis of Ebola, hence facilitating the verification of the results of MSF with the vast literature present on Ebola infection. Especially, the detection of IFN-α/β as point of action for the virus could be considered as a basic indicator of the correctness and usefulness of the approach.

EBOV infection sequenced reads count data was downloaded from GEO (GSE84188), it describes the course of infection at three time-points 6, 24 and 48 hour post infection (hpi). Differential gene expression analysis was performed on the count data with edgeR package (version 3.4.2)[19] using an upper-quartile normalization. The DEG analysis results generated by edgeR were used as input for MSF. Cell signaling interactions were filtered from Reactome Functional interactions (FIs) Version 2016[20] for only direct interactions, which was used as a second input for MSF.

For the earliest time point at 6 hpi, three large modulated sub-graphs were identified with 42, 139, and 69 genes. The modulated sub-graphs consist predominantly of cytokines and chemokines (CXCL10, CCL8, CXCL9, CXCL11, CXCR4, CCR7, CCL4L1, CCL3L1, CCL4, CCL8, CCL20, CCL3, CCL19, IL6, IL27, IL23). IFNB1 and IFNA1 were both identified as two of the possible sources in the most significantly modulated sub-graph identified with 42 genes. IFNB1 has the highest impact score of 14.5. IFNA1 has an impact score of 8.7, in top 5 highest impact scores in the sub-graph it belongs to. Most of the sources identified by MSF were type I interferon induced genes (Supplement material). At 24 hpi seven modulated sub-graphs were identified with four main sub-graphs consisting of 61, 222, 130 and 242 genes, others being smaller than 6 genes. Again, IFNB1 and IFNA1 were identified as two sources out of the total sources with 3.9 and 1.6 impact score. IFNB1 was one of the top 5 impact score sources for the corresponding sub-graph. For the last time-point 48 hpi, six modulated sub-graphs were identified. Three of the sub-graphs were less than ten genes and main sub-graphs had 217, 224 and 276 genes. IFNB1 and IFNA1 were identified as sources in the most significantly modulated sub-graph with an impact score of 2.8 and 3.7, but not among the highest ranked sources (Supplement material).

As stated earlier IFN-α/β was reported to be one of the target genes of Ebola infection. We were able to successfully identify IFNA1 and IFNB1 as sources in all three Ebola infection time-points. Although IFNA1 and IFNB1 were already among the most significant genes in the DGEA during the later time points, MSF was also able to detect them as a source in the very early time-point when the genes were not significant based on the individual DGEA alone. Identifying the possible sources will reduce the search space for potential target genes and can help the biologist as the starting point of clinical testing for drugs and vaccines against an infection.

Table 1 compares the results of MSF, namely the number of detected sub-modules and their total genes numbers, to a simple analysis of mapping significantly modulated genes from the DGEA to the network and joining neighbors to modules. The numbers indicate that MSF detects less but larger and easier interpretable submodules, applying its statistical test. Furthermore, the

Table 1. Comparison of connected sub-graphs of modulated genes in the global network identified after the analysis with MSF and applying different *p*-value cut-offs from edgeR to genes in MSF identified modulated sub-graphs.

	Total number of genes in networks	Number of connected sub-graphs in network
6hpi		
edgeR + MSF	250	3
p-value ≤ 0.1	166	87
p-value ≤ 0.05	152	89
p-value ≤ 0.01	125	76
24hpi		
edgeR + MSF	656	7
p-value ≤ 0.1	457	183
p-value ≤ 0.05	418	198
p-value ≤ 0.01	332	216
48hpi		
edgeR + MSF	744	6
p-value ≤ 0.1	514	189
p-value ≤ 0.05	468	206
p-value ≤ 0.01	363	241

dependency of the results from the *p*-value cutoff choice is demonstrated for the DGEA, which is avoided for `MSF` altogether. It showcases how applying different cut-offs to the *p*-value of genes from edgeR to the sub-graphs identified by `MSF` breaks the larger sub-graphs to many smaller unconnected sub-graphs, many of which are single genes.

Modulated sub-graphs at 6 hpi

Three main modulated sub-graphs identified by `MSF` at 6 hpi are shown in Figure 2. The graphs represent the immediate output of the `MSF` analysis, visualized by StringApp[10] in Cytoscape[11]. Each node represents a gene part of a modulated sub-graph, whereby the associated colors code the functional annotation deduced from KEGG Pathways. The cross-talk between the pathways and also the multiple employment of many genes is evident. The flow of information between the sensors and effectors can be perceived given the directionality of each interaction, indicated by arrows. In more detail, sub-graph 1 (bottom) shows how the activation of Toll-like receptor, Cytokine, Chemokine activating Jak-STAT and MAPK genes, together with TNF leads into apoptosis. The next significant sub-graph (sub-graph 2: top right) reveals how information from the Extracellular matrix (ECM) receptor, which are reported to interact with Ebola glycoprotein (GP)[21], Chemokines, Cytokines, and Cytosolic DNA sensing are directly or indirectly controlling cell growth, differentiation, proliferation and apoptosis. It suggests that dysregulation of these pathways is responsible for modulation of apoptosis. Eventually, sub-graph 3 (top left) demonstrates how IFNA1 and IFNB1 modulates once more, via only a few intermediate steps, the apoptotic response of the cell. On the other side cAMP signaling genes activates platelet genes.

This display case might advertise with how little effort complex data can be processed and prepared for interpretation by the domain expert, to apprehend the dynamics of the underlying processes and suggest testable hypothesis and potential points of intervention.

Robustness

A potential concern is how noise in the gene expression measurements affects our analysis. To assess the robustness and stability of our method, we therefore added Poisson distributed noise to the read counts of the three time-points data set, used above. Then DGEA was carried out on the disturbed data with the same parameters as for the native data using edgeR, followed by analysis with `MSF`. This procedure was carried out 100 times. Every time the genes from the modulated sub-graphs identified from noisy data were compared to the genes of sub-graphs identified from the native data. The robustness of `MSF` analysis for the time-point 6, 24, and 48 hpi is shown in Figure 3. The procedure how data noise was modeled can be considered as rather stringent as `MSF` is sensitive to *p*-value changes across the

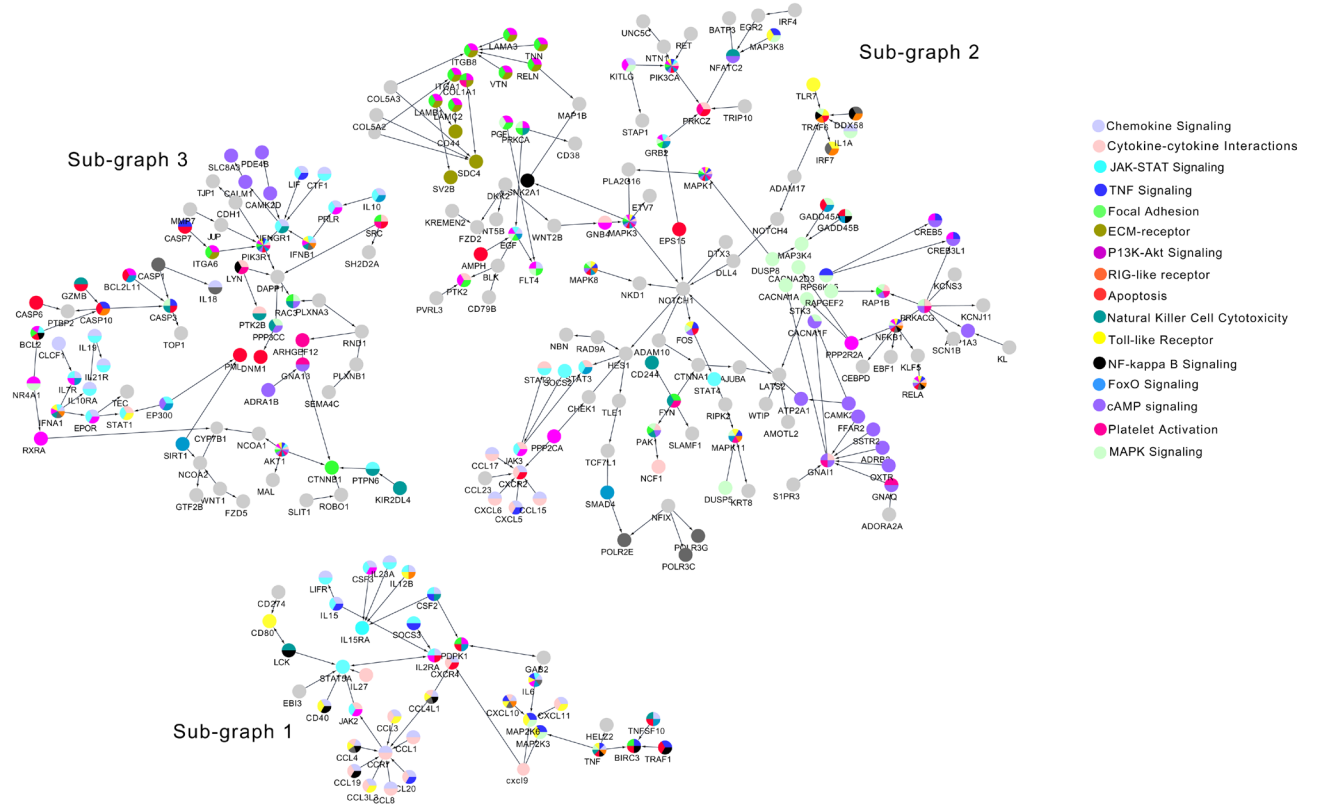

Figure 2. Visualization of the three modulated directed sub-graphs identified by `MSF` at 6 hpi. The node coloring is associated to KEGG pathways referring to the colors in the legend. The graph edges are from Reactome.

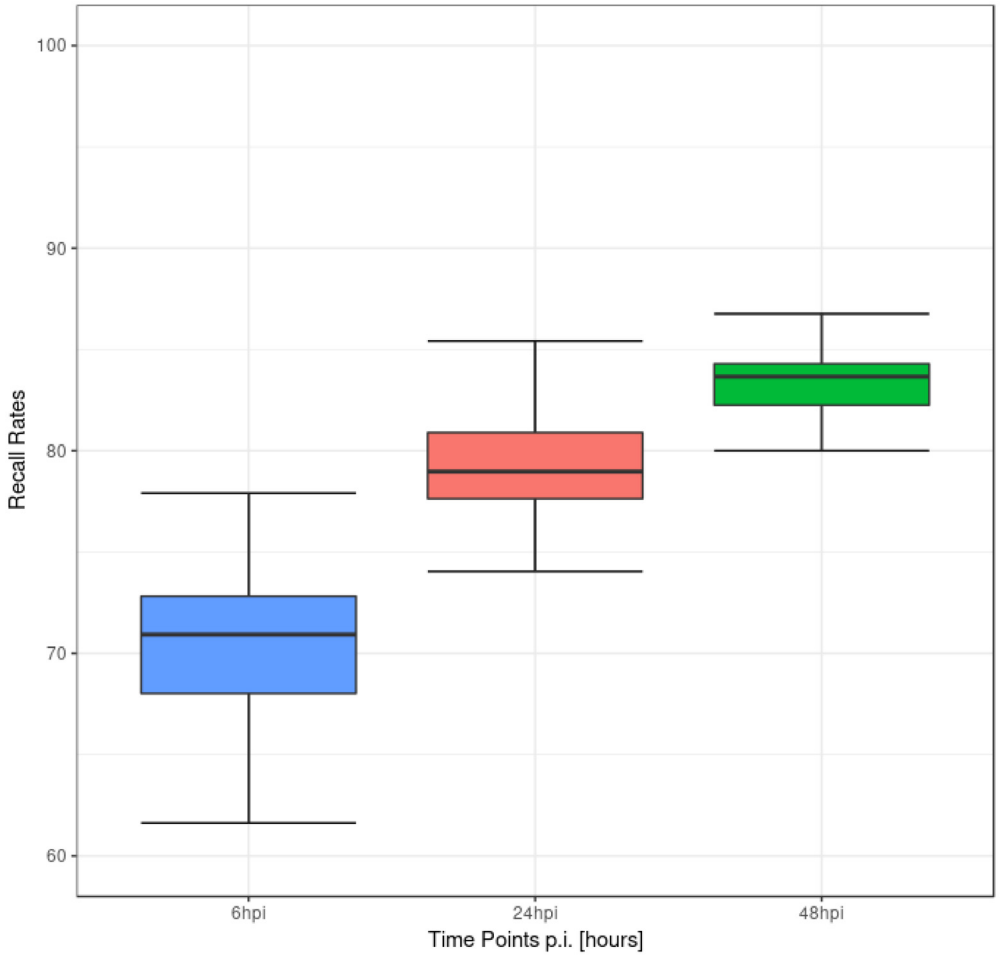

Figure 3. Recall rates for genes in MSF identified sub-graph for the three different time points of EBOV infection data for 100 simulations where Poisson distributed noise was added to the experimentally deduced reads per gene.

whole range of possible values. The observed median recall rates lay between 71 % (6 hpi) and 84 % (48 hpi).

Benchmark

The purpose of the comparisons to existing tools is to show the overall capabilities of MSF. MSF was compared to jActiveModules[8] since it uses similar approaches to find and score modules. For comparison to classical pathway enrichment analysis Reactome[5] and gene set enrichment analysis (GSEA)[22] was chosen since both are widely used and the latter does not rely on *p*-value cut-off.

jActiveModules. jActiveModules[8] is a plugin in Cytoscape that searches for molecular interaction network to find expression activated sub-networks. The method used to score the expression activated sub-networks is close to the method used in MSF. The difference is in how these sub-networks are identified. MSF starts building the sub-graphs from one gene, incorporating and combining the *p*-value of the next gene, with the check that the combined *p*-value of new sub-graph should be better than the original. On the other hand jActiveModules first transforms all the gene's *p*-values p_i to z-scores using $z_i = \Phi^{-1}(1-p_i)$, where Φ^{-1} is the inverse normal CDF and tries to find connected sets of genes with unexpectedly high levels of differential expression, in this case high z-scores. The overall score of the sub-network is calculated by combining the z-scores of the genes. Next using their extended simulated annealing method jActiveModules toggles multiple nodes to merge additional components.

The first time-point of Ebola infection data was analyzed using jActiveModules (Version 3.2.1) to compare the modulated sub-graphs identified by MSF and jActiveModules. The input files were same for both tools. From the modules identified by jActiveModules, the module with the highest pathScore was selected for comparison with MSF identified modulated sub-graphs. The module consists of a single graph with 314 genes. While MSF identified three directed modulated sub-graphs with 42, 69 and 139 genes. The overlap of the common genes identified between MSF and jActiveModules is shown in Figure 4. The sub-graphs identified by MSF are more fragmented than jActiveModules. Unfortunately there is no golden

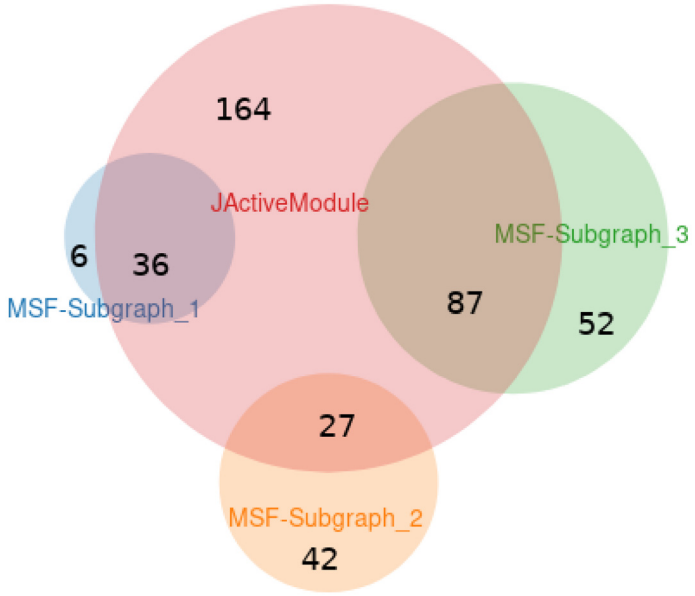

Figure 4. The Venn diagram shows the common genes identified as modulated from MSF identified sub-graphs and jActiveModules identified module.

standard example data that could help benchmark the method. MSF provides directionality, with the identification of possible perturbation sources of the sub-graphs. The predefined pathway labels could also be seen in MSF identified sub-graph with little effort using StringApp[10].

Reactome pathway analysis. Gene enrichment analysis was performed using Reactome analyze data tool[5] (version 67) on the different time-points of Ebola infection data. Reactome's over-representation analysis tool tests whether certain Reactome pathways are enriched for the lists of genes submitted to it. Genes from MSF identified sub-graphs for each time-point were analyzed for gene enrichment using this tool. For comparison the DEG results from edgeR for the three time-points were filtered using the cut-off of adjusted p-value < 0.05. These DEG lists were used for gene enrichment analysis. The compression of MSF identified sub-graphs gene lists and the DEG lists analysis is shown in Figure 5

All enriched pathways with a cut-off of p-value <0.05 for MSF and DEG lists for the three time-points were selected. The comparison shows most of the pathways known from literature to be dis-regulated by Ebola infection are enriched in both the enrichment analysis. EBOV glycoprotein (GP) interacts with the Toll-like recpetor signaling pathway, it triggers the activation of cytokines[13]. Toll-like receptor pathway is expected to be dis-regulated in the early stage of infection, this pathway was not identified as significantly dis-regulated when p-value cut off DEG lists were analyzed for enrichment. Nine Toll-like receptor cascades TLR10, TLR2, TLR3, TLR4, TLR5, TLR7/8, TLR9, TLR1:TLR2 and TLR6:TLR2 were identified as dis-regulated from gene enrichment analysis of MSF identified sub-graph genes, not a single one of these cascade was shown to be dis-regulated in pathway enrichment analysis from DEG cut-off lists. Since MSF considers the complete DEG results, even the weak signal at the earliest time-point was detected; for example Toll-like receptor signaling. While MSF is able to catch weak signals, it does not provide information about the functional relationships among genes like Reactome tool.

Gene set enrichment analysis. Gene set enrichment analysis (GSEA)[22] is a method to identify classes of genes or proteins that are overrepresented in a large set of genes or proteins. GSEA uses statistical approaches to identify significantly enriched or depleted groups of genes. The complete DEG list from DGEA of the first time-point 6 hpi was analyzed using bioconductor package GSEABase (version 1.44.0). GSEA was able to identify Toll-like receptor, chemokine signaling pathway, cytosolic DNA-sensing pathway, Jak-STAT signaling pathway, RIG-I-like receptor signaling pathway and apoptosis as the highest ranked pathways. Although GSEA identified the important pathways for Ebola infection, it did not show the topology of the genes in the different pathways identified and how they cross-talk. MSF and GSEA uses complete DEG list without any cut-off, that is why pathways important for Ebola infection showed up even with weak signal genes in it.

Discussion

Classic pathway analysis tools aim to detect in lists of significantly deregulated genes enriched associations with pathway genes categorized by their biological function and their interactions. Depending on the tool, the internal pathway topology is considered or neglected all together. Here presented tool, MSF, employs a different approach, by aiming to detect sub-graphs in whole gene regulatory networks which are significantly deregulated in a concerted manner. To this end, neighboring genes in the user provided network are tested for jointly common regulation. Exploiting that each gene's abundance, although not independent from its neighbors, is measured on its own, sensitivity can be increased by our applied p-value meta-analysis,

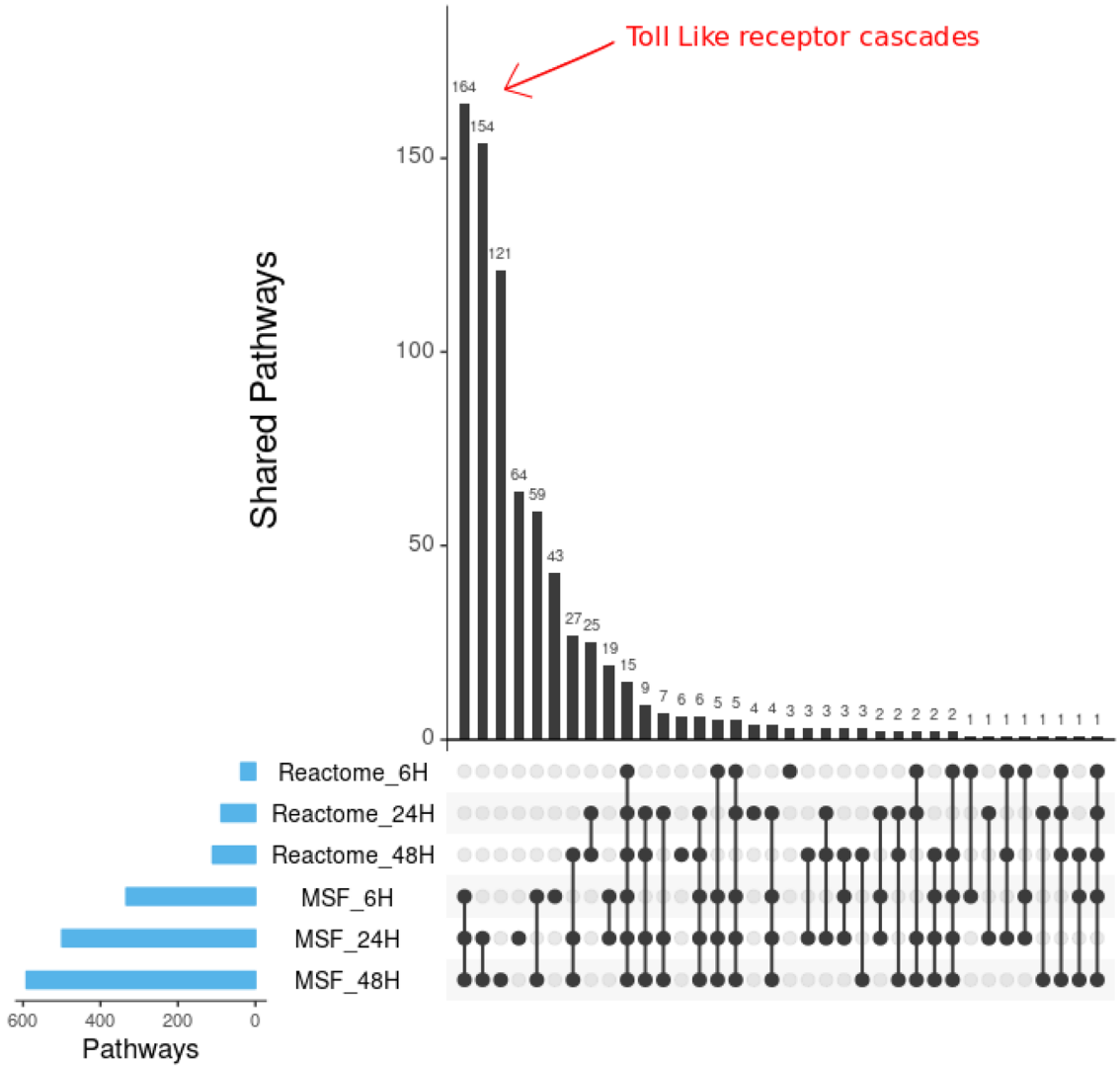

Figure 5. The Upset plot shows the number of shared pathways between MSF identified sub-graph gene list and DEG cut-off list for the three time-points. All 10 different Toll-like receptor cascades are in the set of 164 shared pathways only between MSF at different time-points.

namely Hartung's method. This potentially enables to call not just significant modulated genes based on the DGEA to be convincingly called to be part of a deregulated gene group. Furthermore, it allows to identify connected sub-graphs, representing the propagation of gene regulation perturbation in the input network. A better understanding of this propagation, especially the critical spots such as sensors, effectors, and hubs, facilitates the projection of potential intervention points, e.g., for drug development. Since MSF only uses interaction information in gene regulation network, but not the functional grouping of the genes into functional pathways, it is especially adapted to discover so called cross-talk between such pathways.

Conclusions

MSF is a fast and easy to use tool to find concertedly modulated sub-graphs in a given network. Its implementation in Java enables its use across many operating systems e.g. linux and windows. So far the raw output from edgeR[19] and DESeq2[23] are supported.

Data availability

The Ebola infection RNA-seq data set analyzed during the current study are available in the GEO repository (GSE84188)[13]. The cell signaling network file used is from Reactome Functional interactions (FIs) Version 2016[20]

Software availability

Source code is available from GitHub: https://github.com/Modulated-Subgraph-Finder/ MSF

Archived source code at time of publication: https://doi.org/10.5281/zenodo.2632973[24]

Grant information

This work was funded by the FWF ("Fonds zur Förderung der wissenschaftlichen Forschung") within the project Internationalen Kooperationsprojektes - Intl cooperation Project (Joint Project - Lead Agency Verfahren) with the project number (I 2353-B22). The grant was assigned to ILH. FA was funded by the Austrian Science Fund (FWF) project SFB F43.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Supplement material

Supplementary material is available form GitHub: https://github.com/Modulated-Subgraph-Finder/MSF

References

1. Malone JH, Oliver B: **Microarrays, deep sequencing and the true measure of the transcriptome.** *BMC Biol.* 2011; **9**(1): 34.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Khatri P, Sirota M, Butte AJ: **Ten years of pathway analysis: current approaches and outstanding challenges.** *PLoS Comput Biol.* 2012; **8**(2): e1002375.
PubMed Abstract | Publisher Full Text | Free Full Text
3. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E: **Pathway Analysis: State of the Art.** *Front Physiol.* 2015; **6**: 383.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res.* 2000; **28**(1): 27–30.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Fabregat A, Jupe S, Matthews L, *et al.*: **The Reactome Pathway Knowledgebase.** *Nucleic Acids Res.* 2018; **46**(D1): D649–D655.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Bayerlová M, Jung K, Kramer F, *et al.*: **Comparative study on gene set and pathway topology-based enrichment methods.** *BMC Bioinformatics.* 2015; **16**(1): 334.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Mitra K, Carvunis AR, Ramesh SK, *et al.*: **Integrative approaches for finding modular structure in biological networks.** *Nat Rev Genet.* 2013; **14**(10): 719–732.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Ideker T, Ozier O, Schwikowski B, *et al.*: **Discovering regulatory and signalling circuits in molecular interaction networks.** *Bioinformatics.* 2002; **18 Suppl 1**: S233–S240.
PubMed Abstract | Publisher Full Text
9. Michalak P: **Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes.** *Genomics.* 2008; **91**(3): 243–248.
PubMed Abstract | Publisher Full Text
10. Morris J, Jensen LJ, Doncheva NT: **stringApp 1.3.0**. [Online; accessed 19-Februar-2018]. 2018.
Reference Source
11. Shannon P, Markiel A, Ozier O, *et al.*: **Cytoscape: a software environment for integrated models of biomolecular interaction networks.** *Genome Res.* 2003; **13**(11): 2498–2504.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Hartung J: **A note on combining dependent tests of significance**. Technical report, Technical Report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998.
Reference Source
13. Olejnik J, Forero A, Deflubé LR, *et al.*: **Ebolaviruses Associated with Differential Pathogenicity Induce Distinct Host Responses in Human Macrophages.** *J Virol.* 2017; **91**(11): pii: e00179-17.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Prins KC, Cárdenas WB, Basler CF: **Ebola virus protein vp35 impairs the function of interferon regulatory factor-activating kinases IKKepsilon and TBK-1.** *J Virol.* 2009; **83**(7): 3069–3077.
PubMed Abstract | Publisher Full Text | Free Full Text
15. Konde MK, Baker DP, Traore FA, *et al.*: **Interferon β-1a for the treatment of Ebola virus disease: A historically controlled, single-arm proof-of-concept trial.** *PLoS One.* 2017; **12**(2): e0169255.
PubMed Abstract | Publisher Full Text | Free Full Text
16. Rhein BA, Powers LS, Rogers K, *et al.*: **Interferon-γ Inhibits Ebola Virus Infection.** *PLoS Pathog.* 2015; **11**(11): e1005263.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Falasca L, Agrati C, Petrosillo N, *et al.*: **Molecular mechanisms of Ebola virus pathogenesis: focus on cell death.** *Cell Death Differ.* 2015; **22**(8): 1250–9.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Cárdenas WB, Loo YM, Gale M Jr, *et al.*: **Ebola virus vp35 protein binds double-stranded RNA and inhibits alpha/beta interferon production induced by RIG-I signaling.** *J Virol.* 2006; **80**(11): 5168–5178.
PubMed Abstract | Publisher Full Text | Free Full Text
19. Robinson MD, McCarthy DJ, Smyth GK: **edger: a Bioconductor package for differential expression analysis of digital gene expression data.** *Bioinformatics.* 2010; **26**(1): 139–140.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Wu G, Feng X, Stein L: **Reactome FIs**. [Online; Version 2016]. 2016.
Reference Source
21. Veljkovic V, Glisic S, Muller CP, *et al.*: ***In silico* analysis suggests interaction between Ebola virus and the extracellular matrix.** *Front Microbiol.* 2015; **6**: 135.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Subramanian A, Tamayo P, Mootha VK, *et al.*: **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proc Natl Acad Sci U S A.* 2005; **102**(43): 15545–50.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**(12): 550.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Farman M: **Modulated-Subgraph-Finder/MSFv2.1 (Version v2.1).** *Zenodo.* 2019.
http://www.doi.org/10.5281/zenodo.2632973

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 3

Reviewer Report 29 April 2019

https://doi.org/10.5256/f1000research.20710.r47211

✔ **Guanming Wu**
Ontario Institute for Cancer Research, Toronto, ON, Canada

Almost all my comments have been addressed satisfactorily.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 15 April 2019

https://doi.org/10.5256/f1000research.20710.r46070

✔ **Stefanie Widder**
CeMM-Reseach Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

The authors have addressed all comments sufficiently, I therefore recommend the manuscript for indexing.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: networks, microbiome, cell types

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 15 April 2019

https://doi.org/10.5256/f1000research.20710.r47209

Haibo Liu
Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA

Almost all my questions/comments were addressed to a satisfactory degree. Now the manuscript is in good shape.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics and computational biology, transcriptomics and systems biology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 2

Reviewer Report 02 April 2019

https://doi.org/10.5256/f1000research.20442.r46068

Guanming Wu
Ontario Institute for Cancer Research, Toronto, ON, Canada

This revision has been improved a lot regarding the MSF software and the analysis and discussion of results. This reviewer appreciates the efforts the authors have made to improve the manuscript and the software tool. However, one of the previous major concerns about poor writing has not been addressed carefully. Many typos, grammar errors (especially related to correct use of singular and plural forms of nouns and followed verbs), and other types of errors still exist. There are many examples, here are just some:

1. In Abstract, “have proved tobe..”should be “have proved to be…”
2. In Abstract, “changesdue to” should be “changes due to”
3. In the first paragraph of Introduction, “during an infection providing…” should be “during an infection, providing…”
4. “Peerreviewed” should be “Peer-reviewed”

5. “…and combining it to functional pathway annotations” should be “…and combine it …”
6. “regulating each others expression” should be “regulating each other’s expression”
7. “to each a impact score and a measure of its reliability is assigned” should be “to each an impact score and a measure of its reliability are assigned”
8. “git hub” should be “GitHub”
9. In the equation for t(p), lambda in the first term in the denominator should have a subscript i.
10. “default 2 gene” should be “default 2 genes”
11. “the combined p-value of 3 the merged sub-graph”: should 3 be deleted?
12. “containing the source weightage and the log-fold chances of all considered genes.” Should be “containing the source weights and the log-fold changes of all considered genes.”
13. “…use MSF has been provided on git hub” should be “…use MSF have been provided on GitHub”
14. “the detection of IFN-α/β as point of action for the virus, could be”: comma should not be used
15. Table 1: “Number of connected sub-graph” should be “Number of connected sub-graphs”
16. “MSF was compared to jActiveModules8 since they use similar…” should be “MSF was compared to jActiveModules8 since it uses similar…”
17. “jActive-Modules” should be “jActiveModules”.
18. In the “Reactome pathway analysis” section: Please make sure “lists” are used in many places (not list).
19. “The here presented tool, MSF, employs a different approach…”: Delete “The” please

Other comments:

1. In the 'Abstract', authors pointed out “These methods … rely on binary separation into differentially expressed gene and unaffected genes based on an arbitrarily set p-value cut-off.”. However, as authors correctly described, the second type of classic pathway analysis approaches (e.g. GSEA) doesn’t do this. This should be changed.
2. T-test is usually applied for data having normal distribution or close to normal distribution. Using t-test for pvalues to calculate source genes’ significance is questionable.
3. In the “Case Study” section, “The modulated sub-graphs consist predominantly of Cytokines, chemokines (CXCL10, CCL8, CXCL9, CXCL11, CXCR4, CCR7, CCL4L1, CCL3L1, CCL4, CCL8, CCL20, CCL3, CCL19) and Interleukin genes (IL6, IL27, IL23).”: Interleukins are a type of cytokines. Therefore, this sentence should be modified.
4. Human genes should be in upper case. However, genes listed in the Supplement-Material use lower case, for example, https://github.com/Modulated-Subgraph-Finder/MSF/blob/master/Supplement-Material/6H/Sources . This should be changed.

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 15 Apr 2019

Mariam R Farman, University of Vienna, Vienna, Austria

We would like to thank the reviewer once again for the suggestions.

1. All the text modifications and rephrasing has been done.
2. The sentence in the abstract contradicting to the approach of GSEA has been changed.
3. T-test is now performed on the log-transformed p-values of the individual genes, since log-transformation brings the data close to normal distribution.

4. The sentence in the case study regarding interleukin genes has been changed.
5. MSF now outputs the human genes in upper case.

Competing Interests: No competing interests were disclosed.

Reviewer Report 01 April 2019

https://doi.org/10.5256/f1000research.20442.r46069

?

Haibo Liu
Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA

The 2nd version is improved much over the first version in readability and completeness. However, there are still quite a few language issues, unclear statements, and improper expression of concepts. See the commented manuscript in the PDF format for details.

In addition, the robustness test is not conclusive. Given that RNA-seq data themselves are noisy, I am not sure why it is necessary to add further noise to the RNA-seq data.

In the 2nd version, the authors compared their MSF tools to two existing tools for network or gene set enrichment analyses of an RNA-seq data from an experimental EBOV infection. But it is not clear how MSF outperform the other two tools.

The authors claimed their tool is fast, but no runtime is provided.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics and computational biology, transcriptomics and systems biology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 15 Apr 2019

Mariam R Farman, University of Vienna, Vienna, Austria

Thank you again for your helpful comments about our manuscript.

1. The language issues are thoroughly checked and unclear statements modified.
2. We agree that the RNA-seq data is already noisy, If you look at the robustness of MSF it varies from 71 % (6 hpi) and 84 % (48 hpi), the numbers are not outstanding but we

consider them to be reasonable. This shows that adding noise to the already noisy data the robustness is reasonable, if the data is less noisy the chances to get the truly dis-regulated genes increases.

3. We agree that the comparison with jActiveModule was not very conclusive since we did not have a golden standard example data to analyse by both the tools and then compare the results. Although MSF and jActiveModule have the same approach and do find the core genes, MSF additionally identifies the sources of the modulations in the networks which jActiveModule does not provide. The comparison to GSEA showed that although both the tools MSF and GSEA use no p-value cut-off lists and identify the true dis-regulated KEGG pathways, MSF out perform GSEA by providing the actual genes causing the dis-regulation from the lists and shows the cross-talk between the different KEGG pathways in the form of networks which GSEA does not provide.
4. The runtime for MSF has been provided now.

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 29 November 2018

https://doi.org/10.5256/f1000research.17481.r40208

?

Haibo Liu
Bioinformatics and Computational Biology Program, Iowa State University, Ames, IA

In this manuscript, the authors reported their newly developed tools, MSF, for interpreting gene lists from differential gene expression analysis. Their tool differs from existing pathway analysis tools: (1) it can identify concertedly modulated sub-graphs from user-provided gene networks, thus it can accounting for cross-talk between pathways; (2) it can potentially infer the flow of biological information in response to a perturbation from source to sink. Like the gene set enrichment analysis (GSEA), no arbitrary p-vale threshold is set to dichotomize whole gene lists before applying MSF analysis, instead all genes from DGEA are ordered by p-values from the smallest to the largest. An algorithm similar to the widely used network propagation algorithm is used for subgraph initialization and extension. The authors applied their tools to analyze an RNA-seq dataset from an Ebola virus infection experiment and showed their tool outperform the other tools. They concluded this tool, fast, robust and easy-to-use, is a good supplement to existing pathway-based analysis tools. However, the overall writing is very problematic and there are quite a few issues needing to be fixed. See the details listed as follows.

1. First of all, the code for the tool is not available via Github. I carefully checked multiple times the Github repository: https://github.com/Modulated-Subgraph-Finder/MSF, however I can't find the ModulatedSubPathFinder.jar file, which is the Java implementation of the proposed tool, MSF.
2. There are too many grammar issues and writing issues. Just mentioned a few, in the first paragraph of Introduction, "mechanism" in Line 6 should be plural, while "stimuli" in Line 14 should be "stimulus", "maybe" in Line 15 should be "may be". Careful proof-reading is strongly recommended.
3. The authors have a few misconceptions. For example, they treated "effectors" and "sinks" equally. In my opinion, effectors include sources, intermediate genes and sinks, i.e., all genes responding to perturbations. The authors think of the significance of statistical tests in the form of p-values as a confidence level of observing an authentic expression change. This might not be correct. Besides small p-values, the magnitude of fold changes is also important metric of authentic expression change. By the way, the fold change of gene expression is always non-negative. The expression of "sign of fold change" is not meaningful. Only log-transformed fold change is signed.
4. The flow of information/idea is not fluent in many places. For example, at the end of the first paragraph of Introduction, the authors mentioned the KEGG and Reactome Pathways. Then at the beginning of the second paragraph, they gave a detailed description of the two pathway databases, which might be unnecessary and disrupted the flow to set up the stage to introduce why their tool is necessary and useful. Some information about how their tool was implemented given in the last paragraph of Introduction should be moved to the Implementation section of Methods. Paragraph 3 under the section of “Case Study”, the DEGA results might better be described immediately after the second sentence of Paragraph 2. Similarly, some information in Conclusion should be move to the section of “Implementation”.
5. The title for the section of "Initial modulated sub-graphs” should be “Initializing modulated sub-graphs “. Under this section, “starting with the most significant one” should be “starting with the next most significant one”. “… not yet in a significantly modulated …” should be “…not yet in the significantly modulated …”.
6. Under the section of “Extending modulated sub-graphs”, it is not clear how the sub-graphs are extended by adding “MORE THAN ONE” gene at a time. If doing this way, there are infinite possibilities. The criterion to accept or reject added genes is not clear.
7. Under the section of “Merging modulated sub-graphs”, the authors mentioned that “After detection and extension of the modulated sub-graphs, they are tested if combined subgraphs SCORE better than on their own.” At this point, no merging has been done yet, how are the combined subgraphs tested? What is the SCORE used here? How can a depth-first search traverse from the FIRST sub-graph to the SECOND sub-graph before they are merged (Aren’t the subgraphs not necessarily connected?)?
8. Under the section of “Finding sources & sinks”, “circular loops” should be just “loop”.
9. Paragraph 2 Under “Case Study”, some details about edgeR-based DEGA are missing. How the directed cell signaling interactions ere filtered from the Reaction FIs? Based on what?
10. The directions of edges in right panel of Figure 2 should be showed, because the MSF can generate directed subgraphs.
11. In the last paragraph of the section of “Modulated sub-graphs at 6 hpi”, “This show cast example…” should be “This show case example…”.
12. The Robustness test seemed to show that the MSF is not robust enough to extra noise. What is the authors’ conclusion and explanation?
13. The authors compared the results from their tool to those from the Reactome pathway analysis tool and demonstrated better performance of their tool. Can they also compare the results from their tool to those from the GSEA tools, which don’t set arbitrary cutoff beforehand? The latter comparison might be more convincing.

14. In Discussion, what are “intermediate bottlenecks”? In Conclusion, the authors claimed their tool is “fast, robust and easy-to-use”. However, they did provide evidence to show their tool is fast. The robustness of their tool is not apparent.
15. Issues with figures: all legend titles are too long. In Figure 1, Texts in the flow chart symbols are not well-written. There are inconsistent case issues; “initial” should be “initializing”; symbol for condition test (“checked all interaction?”) should be diamond, not hexagon. Question marks might be added to condition test for readability. In the legend to Figure 1, what does “without exhaustively testing all connected subgraphs” mean? Does it the output from MSF analysis might not be comprehensive? Legends to Figure 3 and 4 are poorly written. Is the Toll-like receptor signaling one of the 149 shared pathways in Figure 4. It is not clearly described in the legend.
16. Limitations of their tool: In their implementation, the fold changes and direction of changes are not taken into account. If a resulting subgraph contains both down-regulated genes and up-regulated genes, how should the users interpret it? The authors didn’t test the sensitivity and specificity of their tool in this manuscript.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
No

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: bioinformatics and computational biology, transcriptomics and systems biology

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 18 Mar 2019

Mariam R Farman, University of Vienna, Vienna, Austria

Thank you very much for your suggestions.

1. So far the ModulatedSubPathFinder.jar was only available under the release tag on git hub (https://github.com/MariamFarman/Modulated-SubGraph-Finder/releases). Meanwhile we added the full source code to git hub.

2. The writing of the manuscript has been carefully checked and improved, especially the examples mentioned by the reviewers.
3. We agree with the reviewer that we used the word “effector” not very cautiously. Therefore, we changed it accordingly. The reviewer is again right to consider the magnitude of the fold change as an important metric of the system behaviour. We consciously ignore it since it is not straight forward to include it in our model, and we consider that the p-value at least partly reflect the magnitude of fold change since it expresses the probability that the observed fold change differs from 1.
4. The suggestions are taken into account and the text was modified accordingly. The details about KEGG and Reactome were considered necessary since later the information from both the databases would be used to showcase the results of MSF.
5. Text modified.
6. The use of extending sub-graph here is if the sub-graph could not be extending any more by one gene because the direct neighbouring genes have high p-values. To avoid producing fragmented sub-graphs, we try to, instead of single genes, append branches of up to three genes simultaneously. Thereby the accumulated signal of the whole branch can compensate for single unfavourable genes. Since we limit the procedure to branches of length three, the number of possibilities to be tested is limited. Again, a branch is only accepted if the overall score of the sub-module is better after extension. Details about criterion of rejection and acceptance of extension was added to the text.
7. The wording has been changed for better understanding of the paragraph describing merging of sub-graphs. The score to pass merging is that the combined p-value of the sub-graph after merging two sub-graphs (including connector genes) is smaller than the individual p-values of the two sub-graphs. The depth first search traverse is used to find connectors between the two sub-graphs to merge them.
8. Text modified.
9. Details about edgeR were added to the text. The interaction file downloaded from Reactome is filtered for only direct interactions . The tutorial on how the file was filtered is also provided on MSF git hub page.
10. The directions have been added to the figure and as a output for MSF to import in Cytoscape for easy visualization for the user.
11. Text modified.
12. While developing the algorithm we believed MSF would not only give insights into the network modulation but could also be used to increase robustness of DGEA of single genes by using additional information from their neighbours. Our analysis disproved this assumption, which we wanted to communicate with this robustness analysis. Given the comments by several reviewers we adapted the corresponding paragraph, omitting the comparison to the DGEA but showing only the overall robustness of our tool which is, after applying strong noise, still reasonable with a median recall rate of 71 to 84%.
13. MSF analysis comparison with GSEA analysis added. Although GSEA was able to find pathways known from literature to be dis-regulated during Ebola infection, it could not show the cross-talk between the different pathways like MSF does.
14. By intermediate bottlenecks we actually meant a gene that is actually connecting a number of sources with a number of sinks and thereby a potential point of intervention if one would like to uncouple the input stimuli from the downstream effects. As stated above the reviewer is right about the robustness, the text has been modified accordingly.
15. The figure legends have been rewritten for better understanding. Figure 1 chart symbols edited, symbols shapes modified as suggested. With the phrase “Without exhaustively testing all connected sub-graphs” we intended to say that not all possible connectors to

merge the graphs are tested but sub-graphs are connected with the first connector that passes the threshold. In figure 5, Toll-like receptor signaling is actually in the 164 shared pathways between different time-points of MSF.

16. We agree and are aware that magnitude and direction of the fold changes are important. On the magnitude we commented already further above. For the direction, we would like to mention that its interpretation is not straight forward without further information of the type of interaction between the genes, which is not always available. To illustrate, the up-regulation of an inhibitor and the down regulation of an activator can have the same effect on the network.

Competing Interests: No competing interests were disclosed.

Reviewer Report 12 November 2018

https://doi.org/10.5256/f1000research.17481.r40009

Stefanie Widder
CeMM-Reseach Center for Molecular Medicine of the Austrian Academy of Sciences, Vienna, Austria

The authors present a novel method for finding groups of genes that concomitantly change their expression profile upon signals and conditional changes. As opposed to standing tools in the field that rely on predefined functional classification, this approach is based on topology of the global interaction network and enables an unbiased identification of larger functional blocks highlighting pathway cross-talk. It furthermore includes a topological method for identifying source and sink pathways that provides a prediction of process causality. The latter is particularly useful for hypothesis generation and add-on experimental validation far beyond the field of biomedicine.

The paper is structured into the presentation of the algorithm, a biomedical use-case with validating background information and a slim benchmarking against differential expression analysis with different cut-offs.

While the method clearly fills an existing gap in high-throughput gene expression analysis and is very elegantly reasoned, I would like to raise a number of comments with regards to writing and benchmarking.

Introduction:

1. The main line of argumentation gets lost sometimes, in particular in paragraph 1. The narrative would furthermore benefit from actual examples instead of repeating a general statement of ‘changed conditions’. Sometimes, the same argument is repeated in differently phrased sentences.

2. Paragraph 2: Suppress ‘be subjective to some degree’.

3. Paragraph 4: Better highlight and delimit the novelty of the presented approach.

4. Overall, shorter sentences benefit the reader.

Methods:

The context of Hartung's method is described nicely, yet the actual way how the individual p-values are combined to result in a single measure is omitted. Please include as this information provides instructive benefit to the reader.

Results:

Case Study:

1. Paragraph 1: Readability would profit from focusing the background information, e.g. 'Ebola infection data was (->were) selected (...)' - rephrase into a half-sentence.

2. Wording: Until now->currently.

Robustness:

1. The recall-comparison (MSF to differential expression analysis) is weakening the proposed method, because no increase in recall for MSF can be achieved. It would be informative to see also precision stats in comparison. Generally, one would expect more robust statistics on larger subgraphs (MSF vs. DEG groups).

2. Also, I am not entirely convinced by adding noise to real and thus – already - noisy data. A small, artificial mock example with detailed known outcome (+/- noise) might be more suitable and supportive.

Figure 4:

This figure is very difficult to follow. I suggest to i) enhance the label sizes and texts, with particular emphasis on delimiting MSF and DEG, ii) improve the figure caption text including fundamental information of what is shown.

Github material:

Importantly - provide an example directory that contains complete examples cases (+*all* files) including the presented Ebola case + options and outcome to enable a swift recapitulation of the tool for the new user.

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Partly

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 18 Mar 2019

Mariam R Farman, University of Vienna, Vienna, Austria

Thank you for your helpful comments about our manuscript.

1. Paragraph 1 has been rephrased, “changed conditions” has been replaced with example of treated verse healthy example.
2. Paragraph 2 modified accordingly.
3. The novelty of MSF is provided in paragraph 5. Firstly, MSF does not use the predefined sets of genes to find the modulated sub-graphs but starts building the sub-graph by using the information from DGEA and interactions from whole cell signaling network. Second MSF considers the signal of the neighbouring genes to find significance of the modulated sub-graph.
4. We revisited the overall text and tried to emphasis more on clear readability.
5. More details about Hartung’s methods are provided in the appropriate section.
6. The case study is rephrased and wording modified.
7. The reviewer is correct about the recall-comparison weakening the proposed method. We have been very strict with testing the robustness of the method by adding extra noise to already noisy data. We expected the method to be more robust than DGEA but unfortunately that was not the case. Since for cut-off based DGEA robustness it does not matter if a few genes p-values go up or down as long as they are below the chosen cut-off. In contrast, for MSF the robustness analysis showed that it makes a difference for the sub-graphs identified. The reviewer is correct again to expect more robustness in the larger sub-graphs which is shown at later time-point 48 hpi that has larger sub-graphs than the other two time-points. If you look at the robustness of MSF alone it varies from 71 % (6 hpi) and 84 % (48 hpi) , we agree that the numbers are not outstanding but we consider them to be reasonable. Since our assumption that a post analysis with MSF can not only gain insights into the pathway modulation but also improve the DGEA for single genes, using their neighbours information did not hold, we removed that aspect to improve readability.

8. We agree that the data is already noisy even before adding noise. Again, we have been very strict to test the robustness of the method,. As mentioned above the larger the sub-graphs the more robust they are, we are not sure that having a small mock example would elaborate more.
9. Figure 5 (previously figure 4) set-up has been changed to one cut-off only i.e. 0.05. Figure label size and texts enhanced. Figure caption modified.
10. Unfortunately we can not provide the Ebola count files since they belong to Olejnik *et al.* A tutorial is provided to reproduce the results from the Ebola infection data, with details on how to obtain the raw data used from the third party source (GSE84188). MSF output files are also provided in the supplementary material.

Competing Interests: No competing interests were disclosed.

Reviewer Report 06 November 2018

https://doi.org/10.5256/f1000research.17481.r39844

Guanming Wu
Ontario Institute for Cancer Research, Toronto, ON, Canada

In this manuscript, Farman et al described an approach to search for network modules based on p-values collected from differential gene expression analysis to address pathway crosstalk, which cannot be addressed in conventional pathway enrichment analysis approaches. During the past decade, many network module-based approaches have been developed to understand the functions of genes collected from differential gene expression analysis and other omics approaches (for a review, see Mitra et al, 2013 [1]). Though the network approach described here has some innovative ideas (e.g. searching for sources and sinks in subgraphs), however, the authors introduce their approach in the context of pathway analysis, without mentioning these previously published similar approaches, let alone comparing their approach to others. Also, it is worthy of mentioning that the described approach in this manuscript is very similar to jActiveModule (Ideker et al, 2012[2]), the first of this kind, widely used for network-based data analysis.

The manuscript used the Ebola virus (EBOV) time course gene expression data set to show case MSF, trying to demonstrate the validity and usefulness of the approach. Indeed, the authors found that IFNA1 and IFNB1 are source genes in subgraphs across multiple time points, as reported by literature references. However, the authors have not discussed other genes in the found subgraphs, though the whole lists of them are provided in their GitHub site. IFNA1 and IFNB1 are among other source genes. The authors should develop a statistic approach to evaluate p-values and FDRs for subgraphs and individual source genes, therefore, providing users a way to choose the most significant genes for unknown phenotypes or biological processes. The current way to showcase the usefulness of the approach is not stringent and may not be useful if too many genes are collected in the subgraphs.

The authors compared MSF results with raw gene lists based on p-value cutoffs (Table 1). However, Table 1 is not a fair comparison. Only the largest subgraphs are listed for edgeR + MSF, while all subgraphs are listed for raw gene lists (e.g. for 6 hpi, 3 for edgeR + MSF, 39 for edgeR (p-value <= 0.1)). For a fair comparison, all subgraphs should be listed for edgeR + MSF too.

Section “Comparison to Reactome pathway analysis tool” and Figure 4 compare results produced by Reactome pathway enrichment analysis for different gene lists. The whole section is confusing. First, the section title is misleading: the comparison is for results generated from Reactome pathway analysis tool, not to that tool. Second, I cannot see too much value in this setting using different adjusted p-value cutoffs for gene lists, probably one (e.g 0.05) should be enough, to reduce the clutter in Figure 4. Third, the authors want to point out MSF can enrich Toll-like receptor signaling pathway, but not the raw gene lists. However, such a comparison has not clearly indicated in Figure 4. The set comparisons include too many gene lists. Fourth, the authors should point out what p-value or FDR cutoff value used to choose pathways from the Reactome analysis tool. It is not correct to choose all pathways listed in that tool for comparison. Finally, where “**Ten**” toll-like receptor cascade pathways come from?

Searching for sources/targets in individual MSF subgraphs based on the directions in the Rectome FI network and then drawing the schematic diagrams as illustrated in Figure 2 are interesting. It will be better to show directions in the Cytoscape network view (the right-side networks in Figure 2). The schematic diagrams in Figure 2 are interesting, but may dramatically simplify things occurring inside cells. The Reactome FI network provides functional relationships among genes or proteins, which are not necessary gene regulatory relationships. The authors should point this out in the manuscript.

Finally, the writing of this manuscript is under question. The authors should really read their manuscript much more carefully. There are far too many typos, wrong uses of punctuations, and grammar errors. For examples, “the Filoviridea family; filamentous” should be “the Filoviridea family: filamentous”; “pathogenesis of Ebola. Thereby facilitating” should be “pathogenesis of Ebola, thereby, facilitating”; “among the most significant gene in the DGEA” should be “among the most significant genes in the DGEA”, and many others.

References

1. Mitra K, Carvunis AR, Ramesh SK, Ideker T: Integrative approaches for finding modular structure in biological networks.*Nat Rev Genet*. 2013; **14** (10): 719-32 PubMed Abstract | Publisher Full Text
2. Ideker T, Ozier O, Schwikowski B, Siegel AF: Discovering regulatory and signalling circuits in molecular interaction networks.*Bioinformatics*. 2002; **18 Suppl 1**: S233-40 PubMed Abstract

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Partly

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Partly

Competing Interests: No competing interests were disclosed.

I have read this submission. I believe that I have an appropriate level of expertise to state that I do not consider it to be of an acceptable scientific standard, for reasons outlined above.

Author Response 18 Mar 2019

Mariam R Farman, University of Vienna, Vienna, Austria

First of all, we would like to say thank you for your valuable review and your critical comments.

1. Thank you very much for making us aware of the paper by Mitra et al and work by Ideker et al. Frankly, these papers skipped our attention. Meanwhile, the review and the tool JactiveModules has been cited in the paper. Using the same expression data and same interaction file we compared the results from MSF and JactiveModules. Although the methods to find the overall score for the modules is similar, there are differences in the sub-graphs identified. The differences seen in the modules identified by the two methods are because MSF starts building the sub-graphs from one gene, incorporating and combining the p-value of the next gene, with the check that the combined p-value of new sub-graph should be better than the original. On the other hand jActiveModules first transforms all the gene's p-values to z-scores and tries to find connected sets of genes with unexpectedly high levels of differential expression, in this case high z-scores. And then the overall score of the sub-network is calculated by combining the z-scores of the genes. Then using simulated annealing jActiveModules tries to find the highest scoring modules. Given the observed differences and our focus on the flow of perturbation, namely sinks and sources, we think that our contribution is not redundant to the previous work. To strengthen even further our perspective we also worked on the software itself, which now also scores the sources according to their reliability and the potential impact onto the modified sub-module. Since we believe that this improved the usability of the software critically, we would like to thank the reviewers particularly for point us to this.
2. Again, we thank the reviewer for the valuable input to evaluate the source genes. We acted on this suggestion by amending the software. We have now incorporated an impact score for each source gene, which expresses the percentage of genes in the sub-module which are downstream of the particular source. This should be helpful to prioritize different sources of one sub-module. MSF also performes a t-test for each source gene, testing if the p-values ofthe downstream genes are different from the upstream genes. This would help to see if the source identified indeed marks the border between two different regulation regimes.
3. The table (Table 1) has been modified. It shows number of MSF identified sub-graphs with the number of genes in it. When we apply different cut-offs to p-values of genes in the MSF identified sub-graph, it shows how they break from larger interpretable sub-graphs to smaller, less interpretable sub-graphs also consisting of single genes.
4. Section "Comparison to Reactome pathway analysis tool" has been modified. Figure 5 (previously figure 4) setup has been changed to one cut-off only, i.e. 0.05. Ten Toll-like receptor cascades were seen to be enriched from the genes in MSF identified sub-graphs

that did not appear in the cut-off gene list. Since MSF uses no cut-off, the sub-graphs identified had genes from Toll-like receptor cascades even when their signal was weak. Figure 5 caption modified for better understanding. A cut-off of 0.05 was used to choose pathways from Reactome pathway analysis for both MSF and cut-off gene list. The nine Toll-like receptor cascades have been mentioned in the manuscript, tenth being Toll-like receptor itself.

5. Directions have been added to the output of MSF that could be easily imported into Cytoscape. In addition, source impact score and log-fold sign for each individual gene can also be imported into Cytoscape. Functional relationship provided by Reactome FI has been mentioned. The schematic diagram, meant to give our take on the interpretation of the raw MSF output, was removed since we agree on the oversimplification criticism.
6. The writing of the manuscript has been carefully checked and improved.

Competing Interests: No competing interests were disclosed.