METHODOLOGY ARTICLE

Open Access

CrossMark

Preservation affinity in consensus modules among stages of HIV-1 progression

Sk Md Mosaddek Hossain[1], Sumanta Ray[1*] and Anirban Mukhopadhyay[2]

Abstract

Background: Analysis of gene expression data provides valuable insights into disease mechanism. Investigating relationship among co-expression modules of different stages is a meaningful tool to understand the way in which a disease progresses. Identifying topological preservation of modular structure also contributes to that understanding.

Methods: HIV-1 disease provides a well-documented progression pattern through three stages of infection: acute, chronic and non-progressor. In this article, we have developed a novel framework to describe the relationship among the consensus (or shared) co-expression modules for each pair of HIV-1 infection stages. The consensus modules are identified to assess the preservation of network properties. We have investigated the preservation patterns of co-expression networks during HIV-1 disease progression through an eigengene-based approach.

Results: We discovered that the expression patterns of consensus modules have a strong preservation during the transitions of three infection stages. In particular, it is noticed that between acute and non-progressor stages the preservation is slightly more than the other pair of stages. Moreover, we have constructed eigengene networks for the identified consensus modules and observed the preservation structure among them. Some consensus modules are marked as preserved in two pairs of stages and are analyzed further to form a higher order meta-network consisting of a group of preserved modules. Additionally, we observed that module membership (MM) values of genes within a module are consistent with the preservation characteristics. The MM values of genes within a pair of preserved modules show strong correlation patterns across two infection stages.

Conclusions: We have performed an extensive analysis to discover preservation pattern of co-expression network constructed from microarray gene expression data of three different HIV-1 progression stages. The preservation pattern is investigated through identification of consensus modules in each pair of infection stages. It is observed that the preservation of the expression pattern of consensus modules remains more prominent during the transition of infection from acute stage to non-progressor stage. Additionally, we observed that the module membership values of genes are coherent with preserved modules across the HIV-1 progression stages.

Keywords: Gene co-expression network, Module eigengene, Hierarchical clustering, Consensus modules, Immune regulatory genes

Background

Acquired Immunodeficiency Syndrome (AIDS) is one of the cataclysmic diseases that have impaired the human species for decades. In spite of the enormous amount of efforts and resources employed to its study and even after thirty-three years of unveiling of the fact that Human Immunodeficiency Virus (HIV) as the cause of AIDS,

there is still no effective vaccine and no cure for this disease [1–3].

After initial infection, a person may not experience any symptom or may undergo a brief period of influenza-like illness, including fever, headache, rash or a sore throat. Typically, this is collocated with a prolonged period of time with no symptoms. As the infection develops, it interacts more with the immune system, intensifying the danger of common infections like tuberculosis, as well as other expedient infections, and tumors that seldom endanger people who have functioning immune systems (http://www.who.int/mediacentre/factsheets/fs360/

*Correspondence: sumantababai86@gmail.com
[1]Department of Computer Science and Engineering, Aliah University, Kolkata, 700156 West Bengal, India
Full list of author information is available at the end of the article

Hossain *et al. BMC Bioinformatics*   (2017) 18:181

Page 2 of 23

en/). These late, defenseless to grievous infections are categorized as AIDS. People often observed substantial weight loss at this stage (http://www.cdc.gov/hiv/basics/whatishiv.html).

There are three main stages of HIV infection: the acute stage (also known as primary HIV infection or acute retro-viral syndrome), chronic stage (this stage is sometimes called "asymptomatic HIV infection", "chronic infection" or "clinical latency stage"), and AIDS [4]. In the acute stage, the initial period following the contraction of HIV, it takes 2-3 weeks after infection until the copy number of HIV-1 virus increases, and the number of CD4+ T (T helper) cells remarkably reduces [5]. However, usually, patients infected with HIV-1 overcome from the acute stage without any treatments within 3-6 weeks and have a clinical latency period of 8 to 10 years (chronic stage) [6].

Although mostly there are few or no symptoms at first and CD4+ T cell count is almost recovered during the clinical latency stage, it has been discovered that immune damage occurs persistently [7]. A small proportion (about 5 to 8%) of HIV-infected patients maintain high levels of CD4+ T cells (T helper cells) without antiretroviral therapy and stay clinically stable for decades. They are called HIV controllers or long-term non-progressors (LTNP) [8]. Nevertheless, the most of the HIV-1 infected patients have a perceptible viral load and in lack of treatment will eventually advance to AIDS, a stage where the CD4+ T cell count falls below 200 cells / μL, and hence T cell-mediated immunity fails to defend the body from pathogens [9].

In recent years, researchers are practicing an extensive use of DNA microarray technology to analyze the expression levels of thousands of genes simultaneously to understand the rationales of cellular systems, molecular networks, disease mechanisms, etc. To reveal system-level properties of genes, construction and analysis of biological networks have been extensively used in [10–13]. Examples of such biological networks are gene regulatory networks, protein-protein interaction networks, metabolic networks, signaling networks, gene co-expression networks, etc. Amidst these biological networks, gene co-expression networks have numerous advantages [14], and empower us to endure a global overview of different diseases. In a co-expression network, genes are interconnected to each other on the basis of the resemblance of their expression profiles and such co-expressed genes tend to participate in the same pathway or form complexes [15, 16] that perform specific functions.

For analyzing the similarities and heterogeneity in network structures through co-expression modules, quite a significant number of computational methodologies have been put forward in [17–19]. To discover the preservation patterns in modules between the human brain and blood tissue, Cai et al., introduced a novel framework in [20]. Conservation and evolution of gene co-expression

networks across the human and chimpanzee brains have been studied by Oldham et al., in [17]. To reveal the association within the co-expression modules through eigengene networks, a revolutionary framework has been introduced by Langfelder et al. [21]. Ray et al. [22] proposed a novel framework to discover topological pattern changes in gene co-expression modules through eigengene networks among different stages of HIV-1 progression using a rank aggregation scheme. A novel framework has been proposed in [23] for discovering the preservation and expression pattern changes in co-expressed modules across three stages of HIV-1 disease progression through an eigengene-based analysis.

In this article, we have developed a novel framework to study the preservation and changes of modular structure in the gene co-expression networks across three stages of HIV-1 disease progression through eigengene-based approach. Initially, we have compiled three separate co-expression networks through Weighted Gene Co-Expression Network Analysis (WGCNA) framework [24] for three stages of HIV-1 disease progression. Next, consensus modules are identified by considering each pair of stages at a time. We have also searched for the immune regulatory genes which are preserved and not preserved among the HIV-1 infection stages across the consensus modules. Additionally, to investigate the topological characteristics of all the shared genes belonging to those consensus modules, we have computed their degree and betweenness centrality and identified the most significant Gene Ontology (GO) Terms and KEGG Pathways associated with them. The overlaps between each pair of consensus modules are investigated through an overlap score. The preservation patterns of the identified consensus modules are then discovered by using an eigengene-based measures. For consensus modules between a pair of stages, we have constructed eigengene network corresponding to each infection stage. We have also investigated preservation of the eigengene networks across two infection stages. The preserved eigengene networks form a higher order meta-network among the module eigengenes. Moreover, some of the meta-modules show a strong preservation during the transition of infection from one stage to another. We have also investigated the correlation between module membership (MM) values of genes with the preservation pattern of consensus modules.

Methods

In this section, we present our proposed model to detect and analyze the modules that are shared by two or more networks (also referred to as consensus modules) across acute, chronic and non-progressor stages of HIV-1 progression. To identify consensus modules, we have utilized the popular WGCNA [24] framework. Figure 1 outlines

Fig. 1 Overview of the whole framework for the present analysis
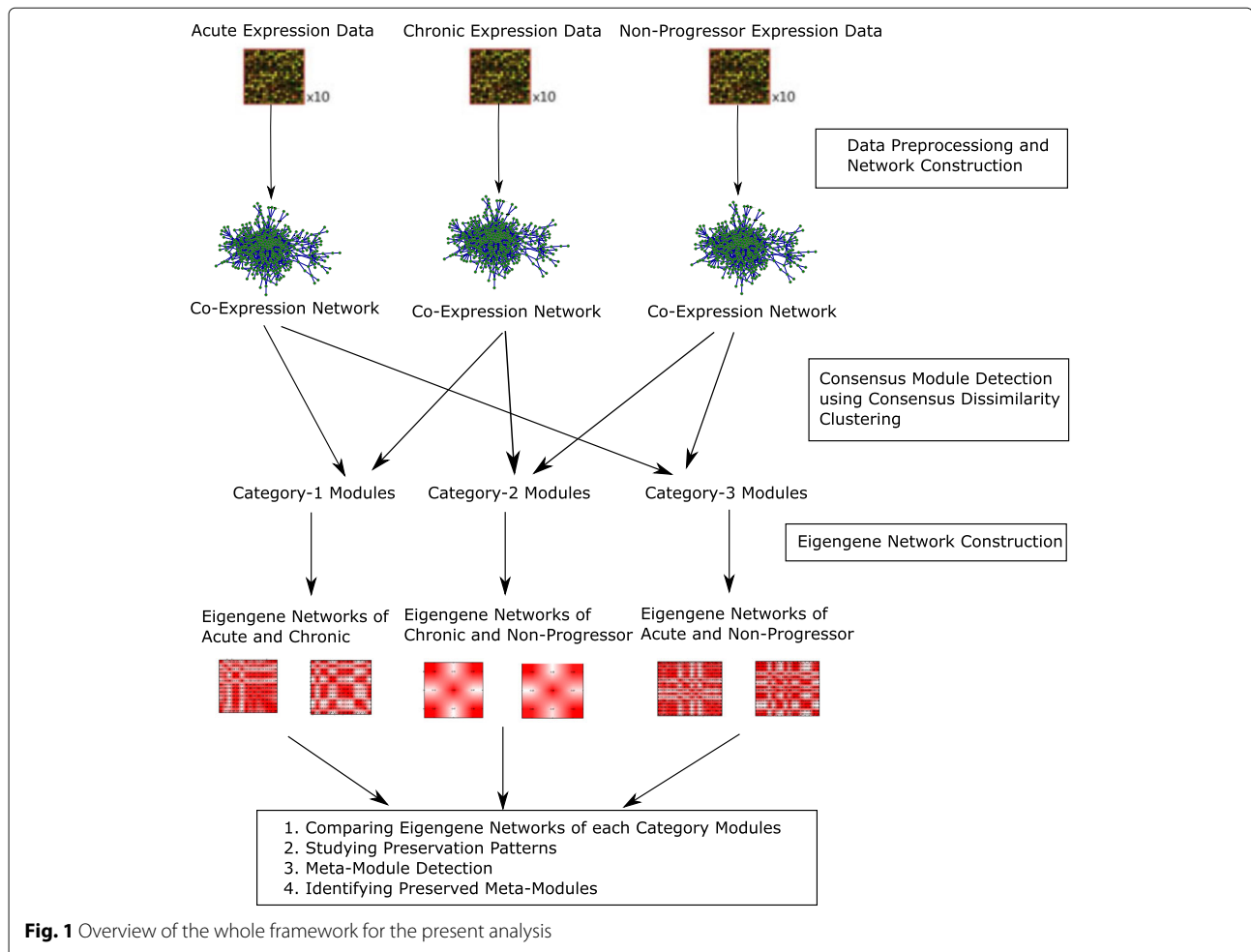
our approach for identifying preservation affinity in consensus modules among stages of HIV-1 progression.

Dataset used

In our present work, we have downloaded the HIV-1 microarray dataset from the Gene Expression Omnibus (GEO) database, submitted by Hyrcza MD, et al. with GEO Series accession no GSE6740 (http://www.ncbi.nlm.nih.gov/geo). It comprises of a stage-specific gene expressions of CD4+ T and CD8+ T cells from a cohort of untreated HIV-1 infected individuals and the dataset has been extracted from 10 gene chips (five gene chips from CD4+ T cells and five gene chips from CD8+ T cells, respectively) for each of the three HIV-1 stages viz. early HIV-1 infection (acute) samples, chronic infection HIV-1 samples, non-progressor HIV-1 infections samples with low or undetectable viral loads, and uninfected samples. All of the categories of datasets (acute, chronic, non-progressor, and uninfected) consist of 10 samples and 22283 genes.

Dataset preprocessing

In addition to the expression dataset acquired from the series matrix files, we have worked out on the CEL files available in the GEO database as mentioned above. At the outset, to winnow out the outliers and for reducing the data dimensionality for computational convenience, the Affy package in the Bioconductor toolbox of R statistical software has been employed here. We extracted the expressed genes from three expression datasets corresponding to acute, chronic, and non-progressor stages that are available in the CEL files through execution of the mas5calls() function [25] of Affy package. The mas5calls() function executes Wilcoxon signed rank-based algorithm for detection and comparison calls on microarray gene expression data. Detection calls are used to find whether the transcript of a gene is present or absent. For performing detection call the intensity differences of perfectly and mismatched probes are used. Comparison call uses the differences between target genes and perfectly matched probes intensities to define the studied genes

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 4 of 23

as increasing, marginally increasing, marginally decreasing, decreasing, or exhibit no expression change at all. One-sided Wilcoxon signed rank test is utilized to obtain "p-value" which is compared with two significance level α_1 and α_2. In this article, we have observed the detection call for $\alpha_1 = 0.05$. Thus, here a gene is said to be present in a sample if the associated "p-value" is less than 0.05. Now, a gene is called expressed in all samples, if it is present in all of the samples. This exposes 6521, 5939, and 6939 expressed genes for the acute, chronic and non-progressor stages, respectively. The list of Genes expressed in different stages of HIV-1 infection are available in Additional file 1 and the genes exclusively expressed in different stages of HIV-1 infection are listed in Additional file 2.

In the next step, we transformed the expression dataset corresponding to all the stages of HIV-1 progression in a multi-set format, by uniting dataset of two stages at a time. To prepare the multi-set datasets, and to restrict our analysis to the most connected genes (i.e. genes which have high correlations in their expression profiles) and to speed up calculations when it comes to module detection, at first, we have employed the Scale-free Topology Criterion proposed by Zhang and Horvath [24]. It is

observed that at soft threshold power (β) value of 24, the acute stage expression dataset with 6521 expressed genes satisfies scale-free topology criterion, as the scale-free topology model fitting index R^2, reaches a high threshold value (0.9), approximately (Fig. 2a(i) and (ii)). A linear relationship between $\log(p(k))$ and $\log(k)$, where $p(k)$ is the probability of the nodes having connectivity k, further confirms that the network is transformed into a scale free network at β value of 24, approximately (Fig. 2a(iii)). Applying same methodology, we have observed that the chronic stage and non-progressor stage expression datasets, approximately attained their scale-free topology criterion at $\beta = 40$ and $\beta = 30$, respectively. Next, we have calculated the connectivity of each expressed gene to all other genes for all the three expressed gene expression datasets by execution of softConnectivity() function of WGCNA package taking the (β) value as an argument. Thereafter, the 5600 most connected genes were extracted by computing the connectivity rank of all of the expressed genes from all the three expression datasets separately. We have also discarded the housekeeping genes which are expressed in all the samples and not associated with HIV-1 infection, from the most connected genes for our analysis.
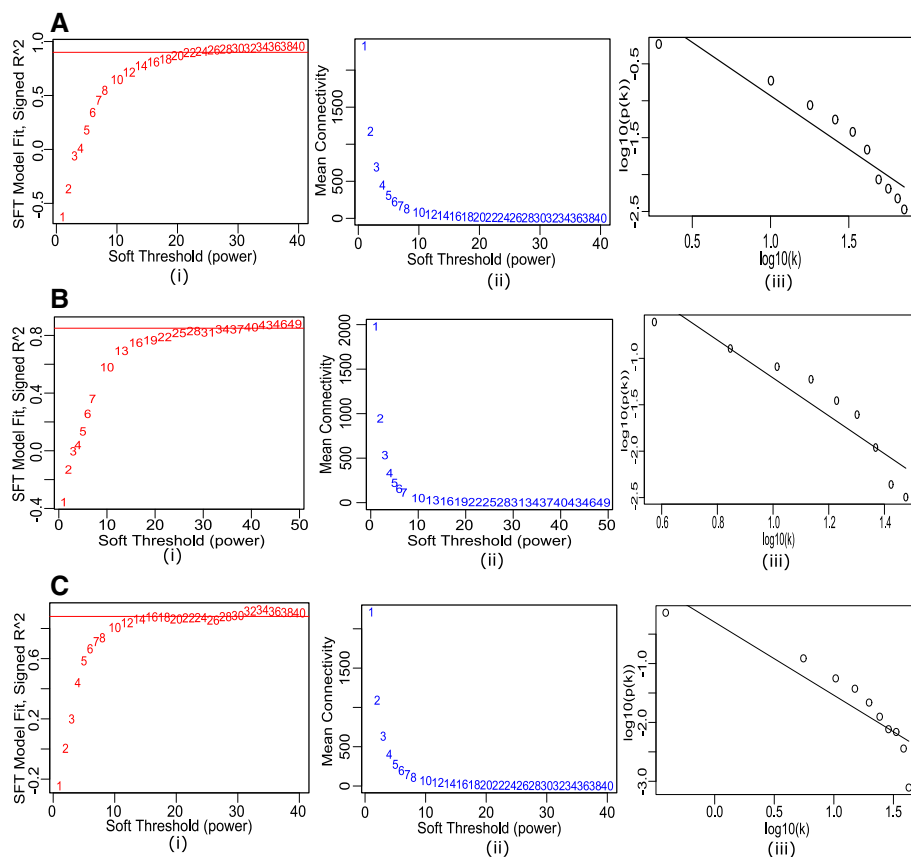

Fig. 2 Scale Free Topology (SFT) Transformation Plots for HIV-1 infection stages. **a** acute dataset **b** chronic dataset **c** non-progressor dataset

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 5 of 23

For this, first, we have selected genes that are expressed in all the samples. Next, we excluded those, which does not belong to HIV Dependency Factors (HDFs) sets (Brass et al. [26], Konig et al. [27], Zhou et al. [28]) and also do not interact with HIV-1 (from "HIV-1 Human Protein Interaction Database" (HHPID) dataset [29]) and also not included in any predicted interaction sets. For preparing predicted interaction set, we have taken union of all computationally predicted interactions from Tastan et al. [30], Dyer et al. [31], Doolittle et al. [32], Mukhopadhyay et al. [33, 34].

Adjacency matrix and connectivity of a network

A network can be interpreted with an adjacency matrix $Adj = [M_{ij}]$ which indicates how nodes are connected among themselves. A gene co-expression network can be represented through a symmetric adjacency matrix consisting of $n \times n$ elements where each node in the network is a gene [35].

In an unweighted network, an element M_{ij} of the adjacency matrix gets a value 1, if nodes i and j are connected (adjacent), or 0, if the nodes are not connected. In a weighted network, $0 \leq M_{ij} \leq 1$ corresponds to the connection strength between the nodes i and j.

$$0 \leq M_{ij} \leq 1, \tag{1}$$
$$M_{ij} = M_{ji},$$
$$M_{ii} = 1.$$

Here, we have constructed gene co-expression network for all the stages of HIV-1 dataset represented in a multi-set format by computing the Spearman correlation for every pair of genes of the gene expression profile matrices.

Transformation of the adjacency matrix

To emphasize the large adjacencies at the expense of low ones and to satisfy scale free topology criteria, we raised all the correlation values of the adjacency matrix to a fixed power β through power transformation law [24]

$$Power_{ij}(Adj, \beta) = M_{ij}^{\beta}. \tag{2}$$

The value of β for power transformation law is the same as soft threshold power (β) that we have already obtained in the "Dataset preprocessing" Section.

Topological Overlap Matrix (TOM) based similarity measure

A major objective of network analysis is to identify groups, or modules of densely interconnected genes which can be revealed by exploring similarity patterns in connection strengths, or high "topological overlap" of among genes. The Topological Overlap Matrix (TOM) based similarity measure [36–38], which indicates how two genes are similar in terms of the commonness of genes they are connected to, has been employed in our present analysis.

TOM is expressed as

$$TOM_{ij}(Adj) = \frac{\sum_{k \neq i,j} M_{ik} M_{kj} + M_{ij}}{min\left(\sum_{k \neq i} M_{ik}, \sum_{k \neq j} M_{jk}\right) + 1 - M_{ij}}. \tag{3}$$

TOM based dissimilarity measure

TOM based similarity matrix can be easily transformed into a dissimilarity matrix by applying the following equation:

$$\begin{aligned} D_{ij} &= Dissim_{ij}(TOM(Adj)) \\ &= 1 - TOM_{ij}(Adj). \end{aligned} \tag{4}$$

Quantile transformation

Topological Overlap Matrices (TOMs) of distinct datasets may possess different statistical features. For example, the TOM in the acute dataset may be systematically higher than the TOM in the chronic dataset. As consensus is expressed as the component-wise minimum of the two TOMs, a bias may result. Here, we illustrate a simple scaling that extenuates the effect of different statistical properties to some extent. We scale the chronic TOM such that the 95th percentile equals the 95th percentile of the acute TOM through Quantile transformation [21], which takes multiple TOMs of the same dimension as input and yields a single TOM whose component $Quant_{q,ij}$ is the qth Quantile of the corresponding components $TOM_{ij}^{(1)}$, $TOM_{ij}^{(2)}$ of the input matrices, computed as follows:
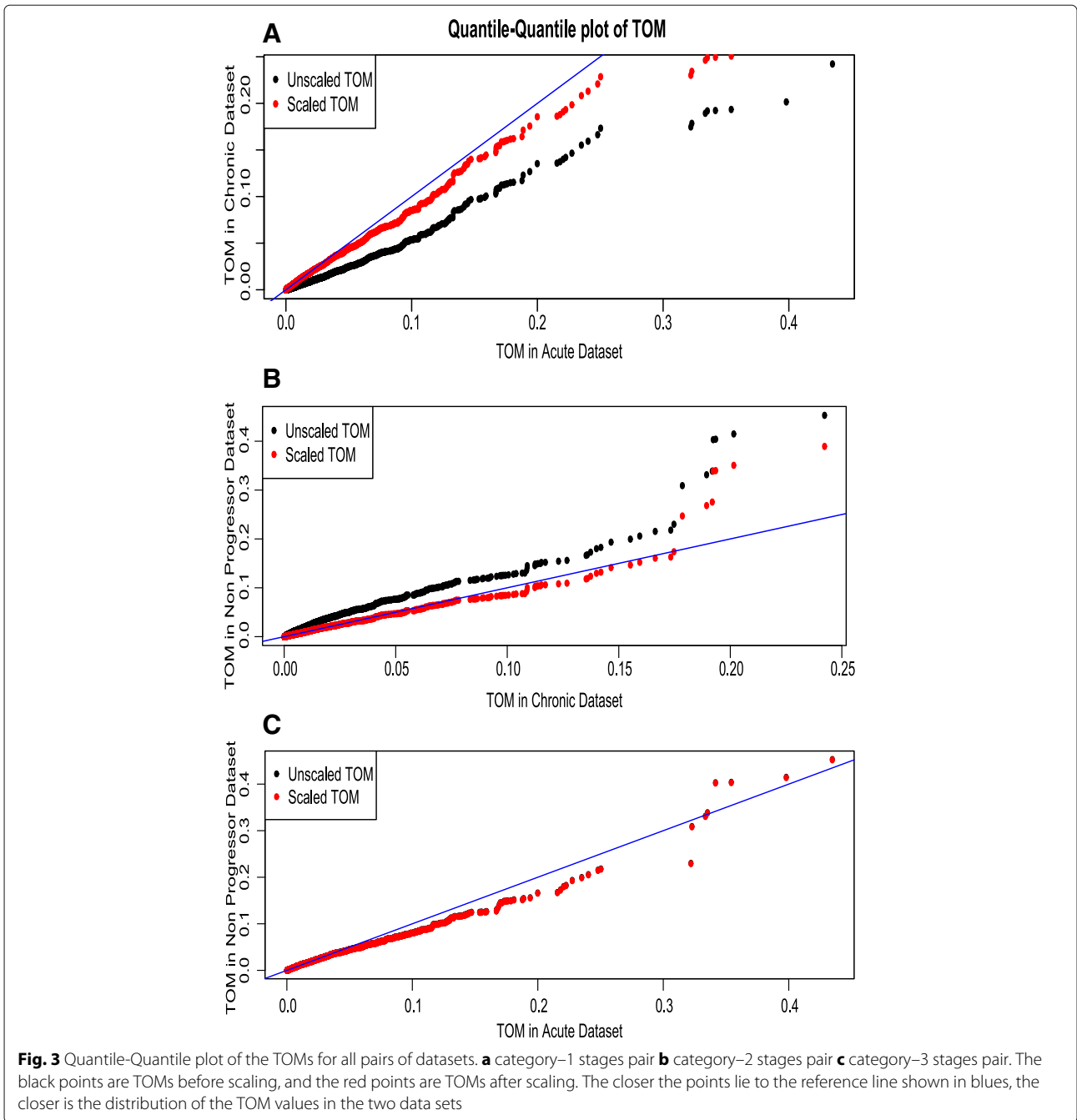
$$\begin{aligned} &Quant_{q,ij}\left(TOM^{(1)}, TOM^{(2)}\right) \\ &= quantile_q\left(TOM_{ij}^{(1)}, TOM_{ij}^{(2)}\right). \end{aligned} \tag{5}$$

Here, $TOM^{(s)}$, denotes the TOM of the dataset s.

To see what the scaling achieves, we form a quantile-quantile plot (Fig. 3) of all the pair of stages (e.g., acute-chronic) topological overlaps before and after scaling. From Fig. 3a, it is clearly visible that scaling changes the chronic TOM moderately, and brings it closer to the reference line shown in blue.

Consensus networks

A consensus network can be constructed from the co-expression networks expressed through adjacency matrices in such a way that, two nodes are connected with each

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 6 of 23

Fig. 3 Quantile-Quantile plot of the TOMs for all pairs of datasets. **a** category–1 stages pair **b** category–2 stages pair **c** category–3 stages pair. The black points are TOMs before scaling, and the red points are TOMs after scaling. The closer the points lie to the reference line shown in blues, the closer is the distribution of the TOM values in the two data sets

other if and only if, all of the input networks 'agree' on that connection. Thus, consensus network is defined as [21]:

$$Consensus_{ij} \left(TOM^{(1)}, TOM^{(2)}, \ldots \right)$$

$$= Min_{ij} \left(TOM^{(1)}, TOM^{(2)}, \ldots \right),$$

$$\text{where, } Min_{ij} \left(TOM^{(1)}, TOM^{(2)}, \ldots \right)$$

$$= min \left(TOM_{ij}^{(1)}, TOM_{ij}^{(2)}, \ldots \right). \quad (6)$$

Consensus modules

Modules in the consensus network are termed as consensus modules. In our present work, we have constructed consensus modules using pairwise gene dissimilarity measure defined analogously to Eq. (4):

$$Dissim \left(Consensus \left(TOM \left(Adj^{(1)} \right), TOM \left(Adj^{(2)} \right), \ldots \right) \right), \quad (7)$$

as input to the average linkage hierarchical clustering. The branches originating from the resulting cluster tree

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 7 of 23

(Dendrogram) are referred to as consensus modules. We have utilized a dynamic tree cut algorithm [39] for this purpose. Please note that here we have used the hierarchical clustering algorithm to group genes whose expression profiles are highly correlated across samples for a pair of stages. The alternative clustering procedures can also be employed to group the genes. In this article, we have followed the procedure described in [21] to perform such grouping.

Module summarization by Eigengene network

After constructing the consensus modules using hierarchical clustering technique as described above, we have summarized each consensus module expression profile by one representative gene: the module eigengene. Module eigengene is defined as the first right singular vector of a module expression matrix. Let, $C^{(k)} = (c_{ij}^{(k)})$ refers to the gene expression data corresponding to module k, where index $i = 1, 2, \ldots, p$ corresponds to the module genes and the index $j = 1, 2, \ldots, q$ corresponds to the microarray samples, and each row of $C^{(k)}$, has been standardized to mean 0 and variance 1. The singular value decomposition of $C^{(k)}[p \times q]$ is defined as:

$$C^{(k)} = UDV^T, \tag{8}$$

where, the columns of the orthogonal matrices $U = (u_1, u_2, \ldots, u_{(min(p,q))})$ and $V = (v_1, v_2, \ldots, v_{(min(p,q))})$ are the left- and right-singular vectors, respectively, and $D = (d_1, d_2, \ldots, d_{(min(p,q))})$ is a diagonal matrix containing singular values. Incorporating terminology from [17, 40–42], the first column of $V^{(k)}$ is referred to as the Module Eigengene:

$$ME^{(k)} = v_1^{(k)}. \tag{9}$$

Let, ME_I and ME_J denote the module eigengenes of the Ith and J^{th} modules, respectively, then the connection strength between eigengenes ME_I and ME_J is expressed as:

$$M_{Eigen,IJ} = \frac{1 + cor(ME_I, ME_J)}{2}. \tag{10}$$

Eigengenes of different modules of a gene co-expression network often exhibit correlations which we have used to constitute eigengene network [21]: Adj_{Eigen}, which is defined as follows:

$$Adj_{Eigen} = (M_{Eigen,IJ}). \tag{11}$$

Eigengenes of different consensus modules often exhibit correlations which we have used to constitute consensus eigengene networks:

$$Cons_{Eigen} = \left(Adj_{Eigen}^{(1)}, Adj_{Eigen}^{(2)}, \ldots \right). \tag{12}$$

Detecting meta-modules from Eigengene networks

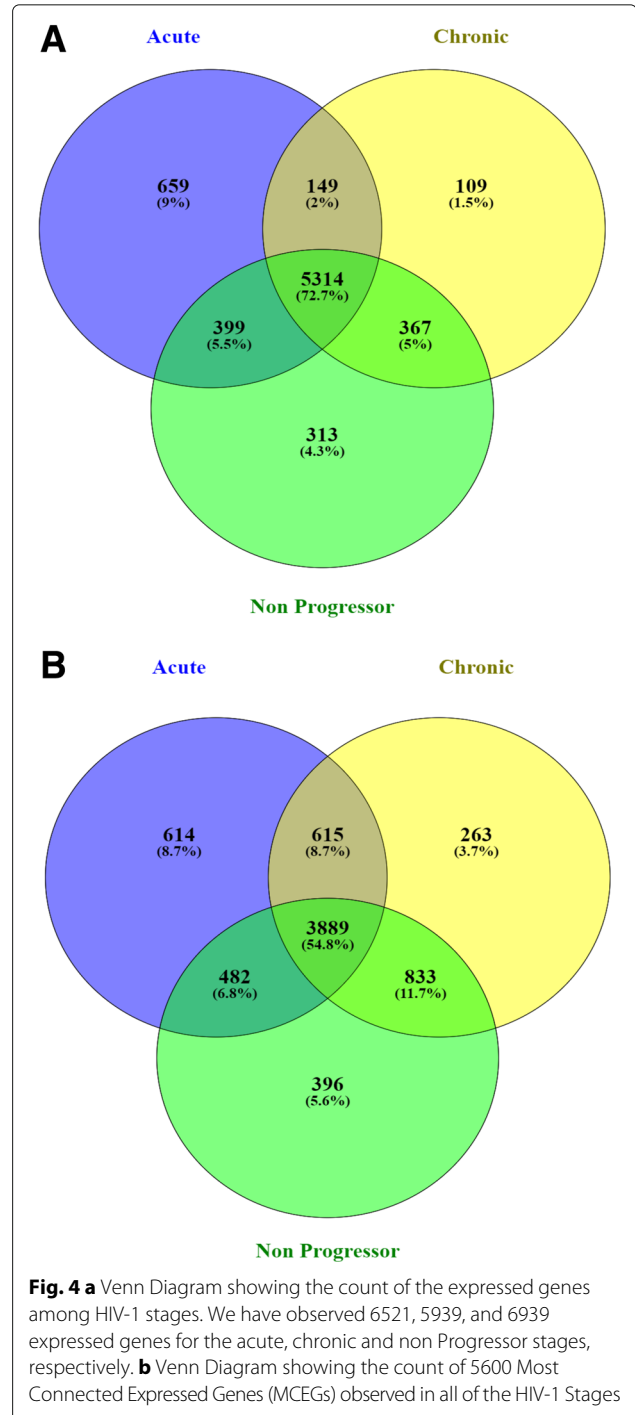After constructing the eigengene network, a module detection algorithm can be employed to detect modules

Fig. 4 a Venn Diagram showing the count of the expressed genes among HIV-1 stages. We have observed 6521, 5939, and 6939 expressed genes for the acute, chronic and non Progressor stages, respectively. **b** Venn Diagram showing the count of 5600 Most Connected Expressed Genes (MCEGs) observed in all of the HIV-1 Stages

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 8 of 23

in the eigengene networks that are referred to as meta-modules. The dissimilarity measure utilized here to detect such meta-modules is defined as [21]:

$$Dissim_{IJ}(Adj_{Eigen}) = \frac{1 - cor(ME_I, ME_J)}{2}, \qquad (13)$$

where $cor(ME_I, ME_J)$ refers to the correlation between the module eigengenes of Ith and J^{th} modules. We have used this the dissimilarity matrix as input to the average linkage hierarchical clustering, resulting in a cluster tree of modules (represented by eigengenes) and the branches of the cluster tree are referred to as meta-modules in our application.

Detecting consensus meta-modules

From the consensus eigengene network constructed through the method described above, a module detection algorithm can be employed again. The modules in the consensus eigengene networks hence detected are referred to as consensus meta-modules. The dissimilarity measure utilized here to detect such meta-modules is analogous to Eq. (13) and expressed as:

$$Dissim(Cons\left(Adj^{(1)}_{Eigen}, Adj^{(2)}_{Eigen}, \dots\right). \qquad (14)$$

The branches emanating from the cluster tree of modules resulting from average linkage hierarchical clustering using the above dissimilarity matrix as input correspond to consensus meta-modules in our application.

Identifying overlaps among the consensus modules

In the present article, we have also computed the overlaps among the identified consensus modules by taking two categories of modules at a time. To measure the overlap we have used Jaccard-based similarity metric defined as follows:

$$O_{i,j} = \frac{|M_i \cap M'_j|}{|M_i \cup M'_j|}, \qquad (15)$$

where $M_i \in$ category-i module, while $M'_j \in$ category-j module. For each pair of modules we have computed the overlap and constructed an overlap matrix as
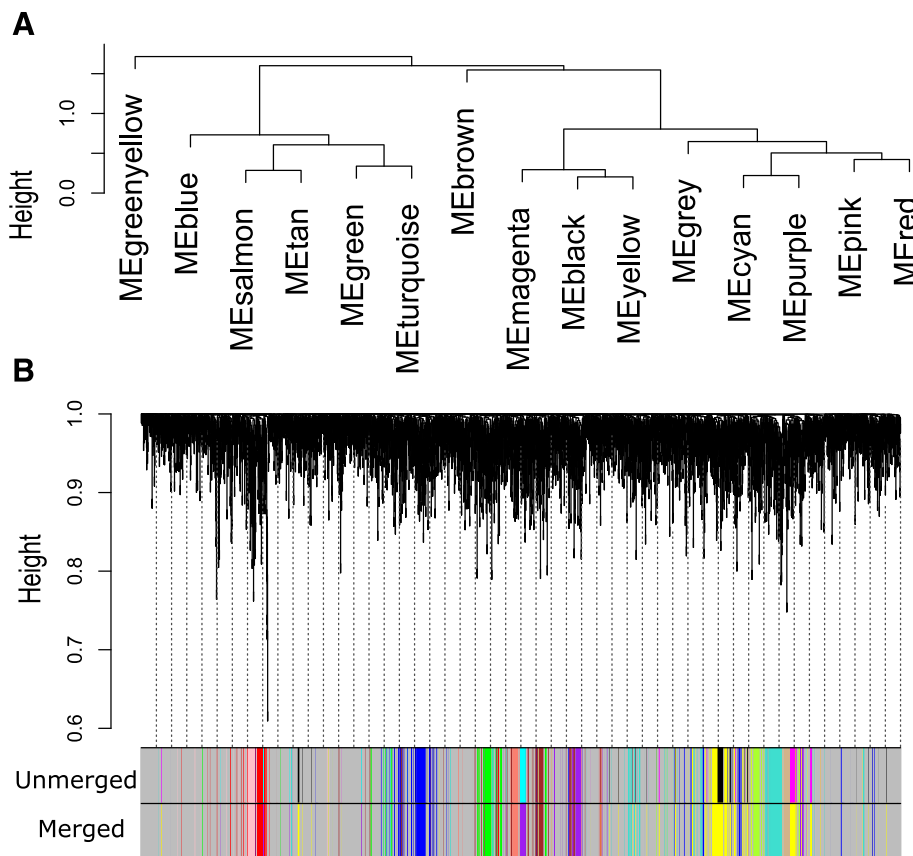

Fig. 5 a Hierarchical clustering tree for category-1 consensus modules. Each consensus module is described here as a module eigengene of it (e.g. module '*blue*' is represented as 'MEblue'). **b** The dendrogram for category-1 consensus modules. Here, 'unmerged' color codes signify the consensus modules whereas merged denote the consensus meta-modules. In this case, three consensus modules '*yellow*', '*black*' and '*magenta*' are grouped in '*yellow*' meta-module and '*cyan*' and '*purple*' are grouped into '*purple*' meta-module

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 9 of 23

$Overlap_{mat} = [O_{i,j}]_{m \times n}$. The overlap scores for category-*i* module M_i and category-*j* module M'_j are defined as

$$OvScore_{M_i} = max_{j=1}^{n} Overlap_{M_i,M'j}, and$$
$$OvScore_{M'_j} = max_{j=1}^{m} Overlap_{M_i,M'j}, \tag{16}$$

where *n* and *m* are the numbers of category-*i* and category-*j* modules, respectively. The OvScore metric of a module indicates the proportion of involvement of it in two other categories of modules.

Identifying preservation pattern in consensus modules among HIV-1 stages

To discover the changes in preservation patterns across each category of consensus modules, we have compared the two eigengene networks, each corresponds to an HIV-1 stage in a specific category. For example, we have compiled the eigengene networks corresponding to acute and chronic stages in category-1 module.

To compare eigengene networks (Eq. (11)), we have used the measures introduced in [21]. Let $Adj_{Eigen}^{(p)}$ and $Adj_{Eigen}^{(q)}$ denote the adjacency matrices of consensus eigengene networks of stages *p* and *q*. We construct a preservation

network between these two consensus eigengene networks as follows:

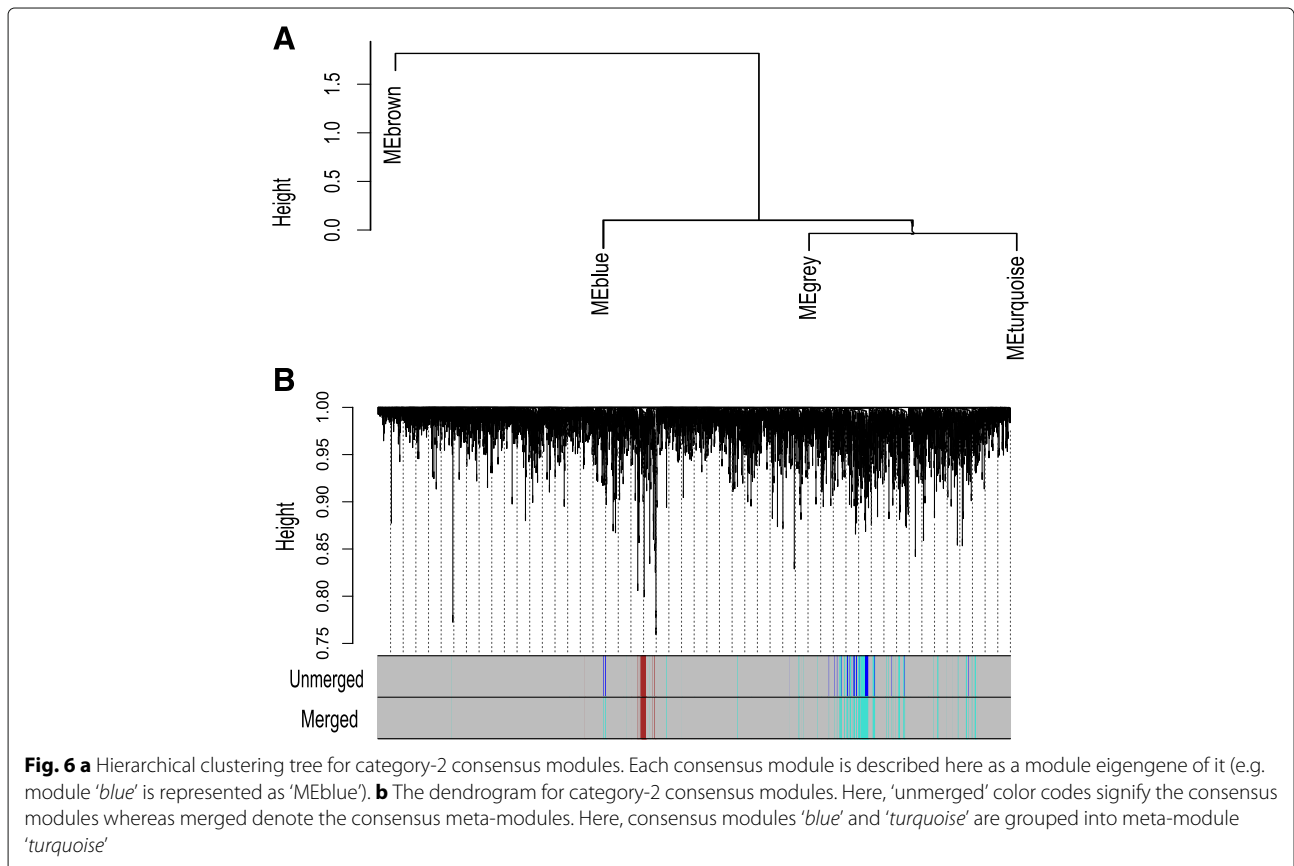$$Pres^{(p,q)} = Pres\left(Adj_{Eigen}^{(p)}, Adj_{Eigen}^{(q)}\right), \tag{17}$$

where the entries of the preservation network $Pres^{(p,q)}$ are defined as:

$$Pres_{I,J}^{(p,q)} = 1 - \frac{\left|cor\left(ME_I^{(p)}, ME_J^{(p)}\right) - cor\left(ME_I^{(q)}, ME_J^{(q)}\right)\right|}{2}. \tag{18}$$

Here, $ME_I^{(k)}$ signify the eigengene of the Ith consensus module in dataset *k*. Larger values of $Pres_{I,J}^{(p,q)}$ signify more preservation of correlation pattern among module eigengenes ME_I and ME_J across two networks.

Furthermore, to investigate the preservation between module eigengenes across two networks we have computed the Scaled Connectivity $C_I(Pres^{(p,q)})$ [21] of a module eigengene $ME_I^{(k)}$ which is given as:

$$C_I(Pres^{(p,q)}) = 1 - \frac{\sum_{J \neq I}\left|cor\left(ME_I^{(p)}, ME_J^{(p)}\right) - cor\left(ME_I^{(q)}, ME_J^{(q)}\right)\right|}{2(N-1)}. \tag{19}$$

Fig. 6 a Hierarchical clustering tree for category-2 consensus modules. Each consensus module is described here as a module eigengene of it (e.g. module '*blue*' is represented as 'MEblue'). **b** The dendrogram for category-2 consensus modules. Here, 'unmerged' color codes signify the consensus modules whereas merged denote the consensus meta-modules. Here, consensus modules '*blue*' and '*turquoise*' are grouped into meta-module '*turquoise*'

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 10 of 23

$C_I(Pres^{(p,q)})$ is close to 1 if the Ith module eigengene has a strong preservation pattern with the most of the other eigengenes. The density (D) [21] of the preservation network $Pres^{(p,q)}$ is given by:

$$D(Pres^{(p,q)}) = 1 - \frac{\sum_I \sum_{J \neq I} \left| cor\left(ME_I^{(p)}, ME_J^{(p)}\right) - cor\left(ME_I^{(q)}, ME_J^{(q)}\right) \right|}{2N(N-1)}.$$

(20)

Larger values of $D(Pres^{(p,q)})$ indicate a strong preservation of correlation patterns among the most of the eigengenes across the two networks.

Computing module membership of genes within consensus modules

Module membership (MM) of a gene is defined as the Pearson correlation value between the expression level of a gene on the microarray and the module eigengene.

The measure describes the extent of similarity between the expression level of the gene and the overall expression pattern of the module. Here, we compared the MM values of all the genes within a shared module between a pair of infection stages. In particular, we have computed the module membership values of all the genes within a shared module $M_i^{p_1-p_2}$, for stage p_1 as follows:

$$MM\_M_i^{p_1} = [v_1, v_2, \ldots v_n],$$
$$\text{where, } v_j = corr(ME_i, g_j).$$

(21)

Here, ME_i is the module eigengene of module M_i, g_j is the expression profile of jth gene in the module, and $corr(.)$ denotes the Pearson correlation operator. Similarly, we have computed the MM values of module $M_i^{p_1-p_2}$ for stage p_2. Therefore, for a shared module between a pair of stages, we obtained two sets of MM values, each of which corresponds to one stage. Next, we merged these two sets
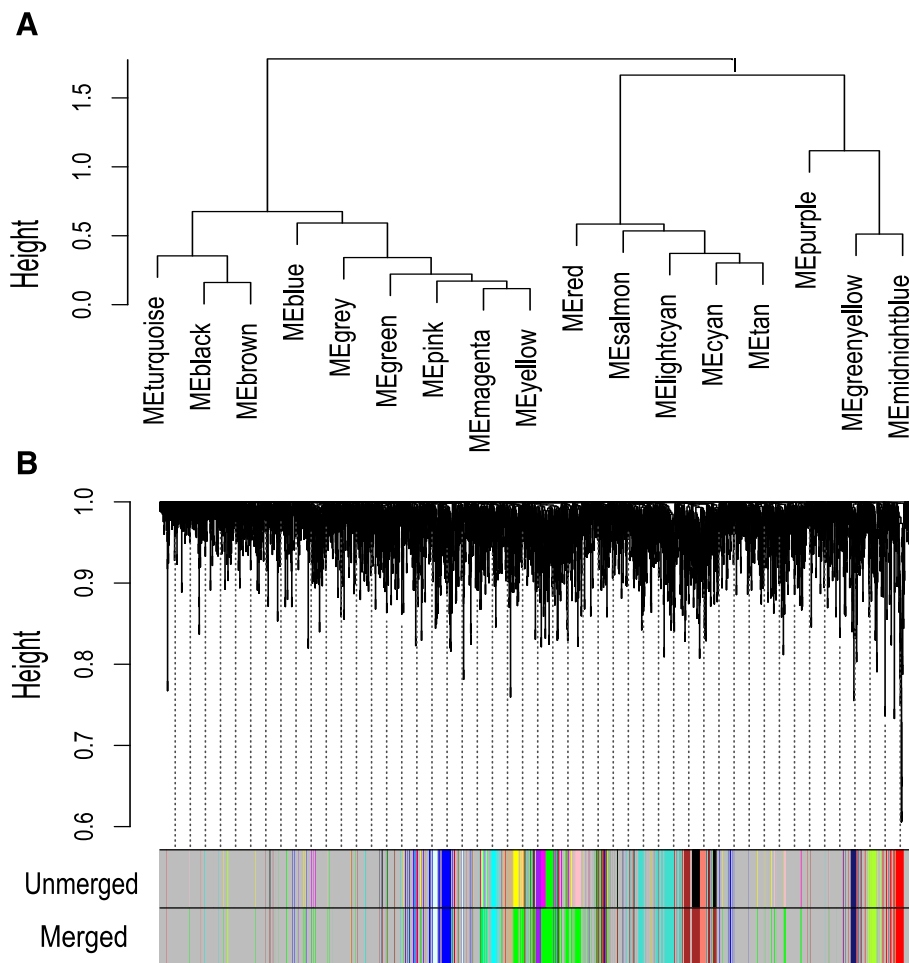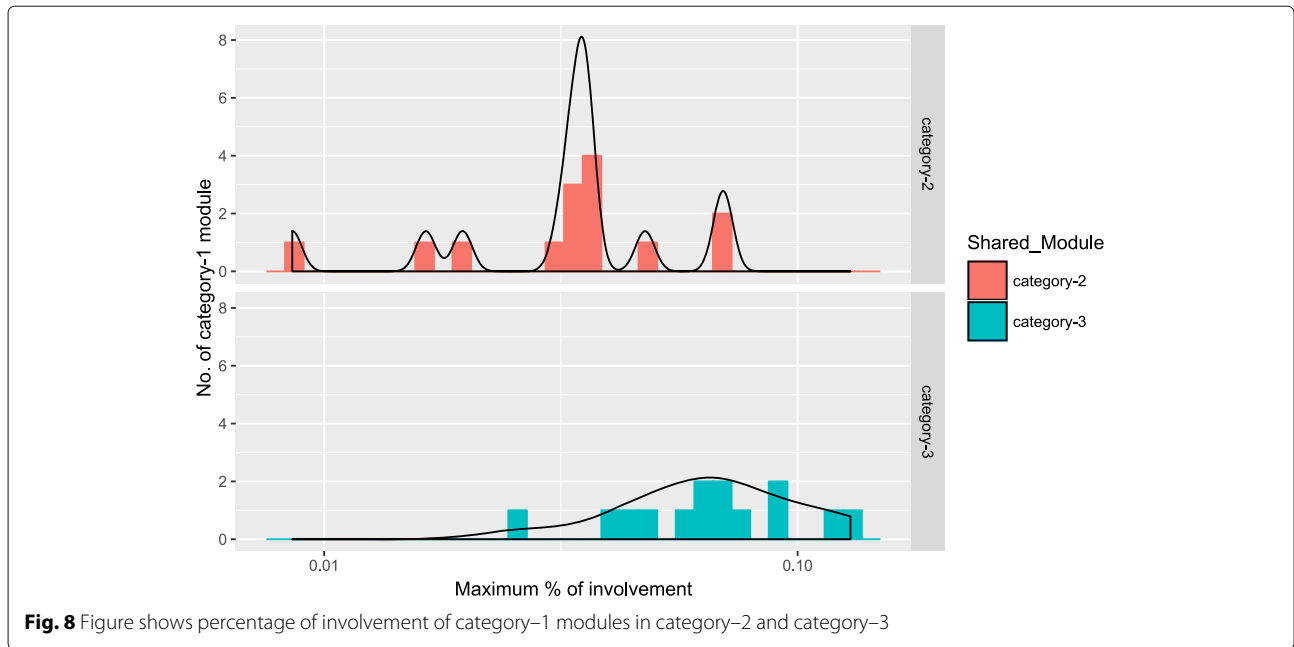

Fig. 7 a Hierarchical clustering tree for category-3 consensus modules. Each consensus module is described here as a module eigengene of it. (e.g. module '*blue*' is represented as 'MEblue'). **b** The dendrogram for category-3 consensus modules. Here, 'unmerged' color codes signify the consensus modules whereas merged denote the consensus meta-modules. Here, consensus modules '*black*' and '*brown*' are grouped into meta-module '*brown*', consensus modules '*green*', '*pink*', '*magenta*' and '*yellow*' are grouped to meta-module '*green*'

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 11 of 23

Fig. 8 Figure shows percentage of involvement of category–1 modules in category–2 and category–3
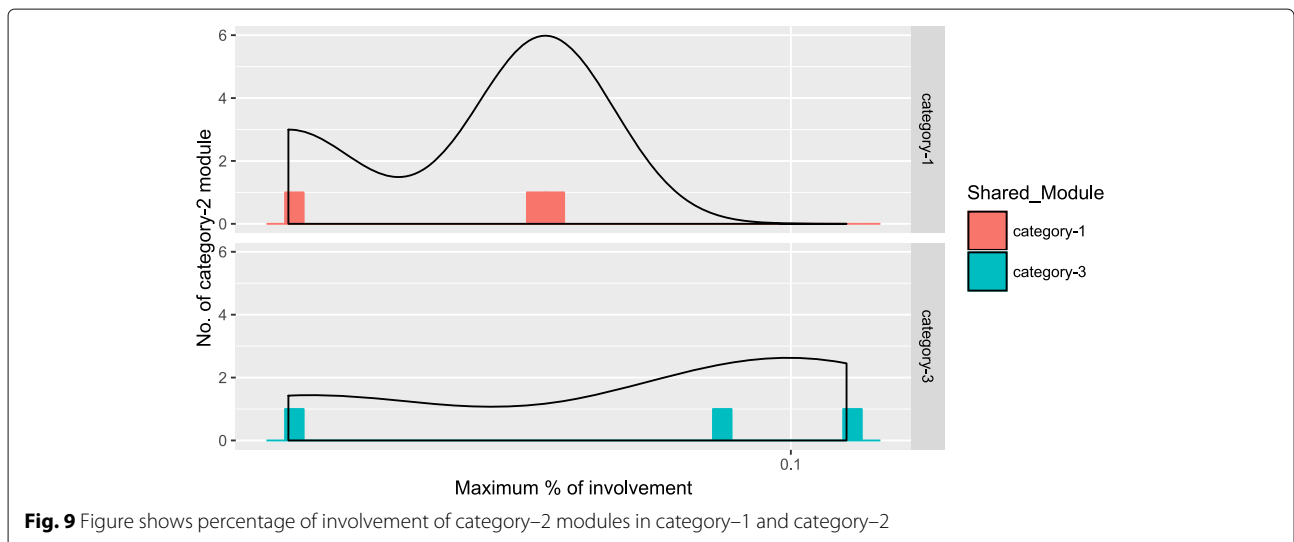
into one set of MM values and compared this set between the preserved shared modules.
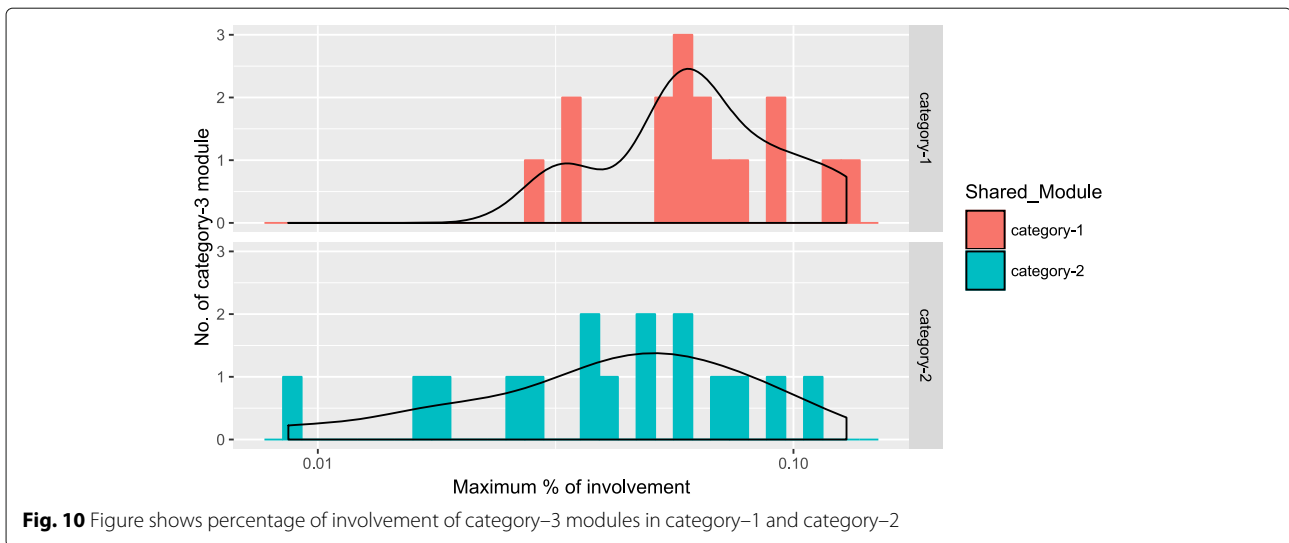
Results and discussion

Here we report the results of our eigengene based analysis of consensus modules identified at different stages of HIV-1 progression. For the rest of the paper, we will use the term category-1 modules for consensus modules of acute and chronic stages, category-2 modules for consensus modules of chronic and non-progressor stages and category-3 modules for consensus modules of acute and non-progressor stages.

Overlaps among the expressed genes

We have observed the overlaps among the selected expressed genes in acute, chronic and non-progressor stage. Figure 4 shows the overlaps among 6521 selected genes of acute stage, 5939 selected genes of chronic stage and 6393 selected genes of non-progressor stage. It can be noticed from Fig. 4a that all the stages share a good amount of common genes (72.7%) among themselves. A relatively small number of expressed genes (109/1.5%) of chronic stage have no overlaps with the expressed genes of other stages. For consensus module detection, we have transformed the expression profiles of these expressed

Fig. 9 Figure shows percentage of involvement of category–2 modules in category–1 and category–2

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 12 of 23

Fig. 10 Figure shows percentage of involvement of category–3 modules in category–1 and category–2
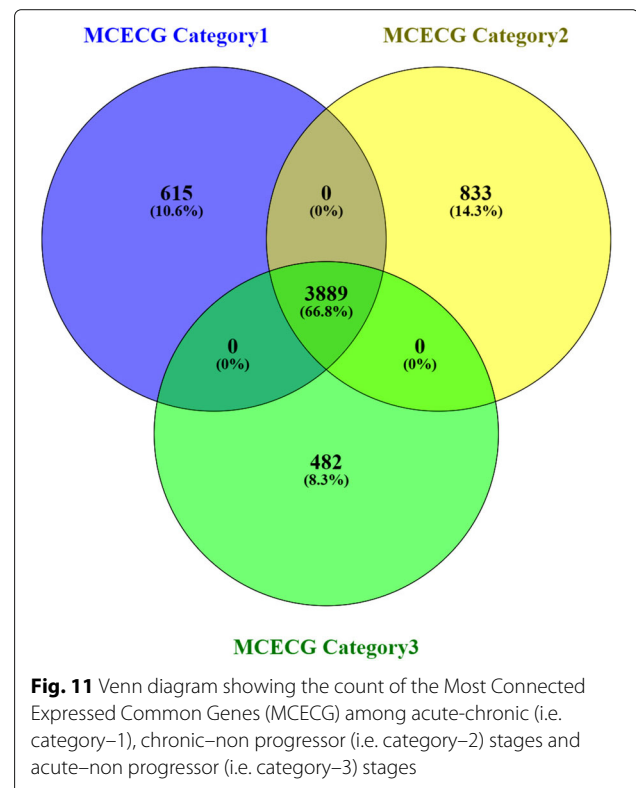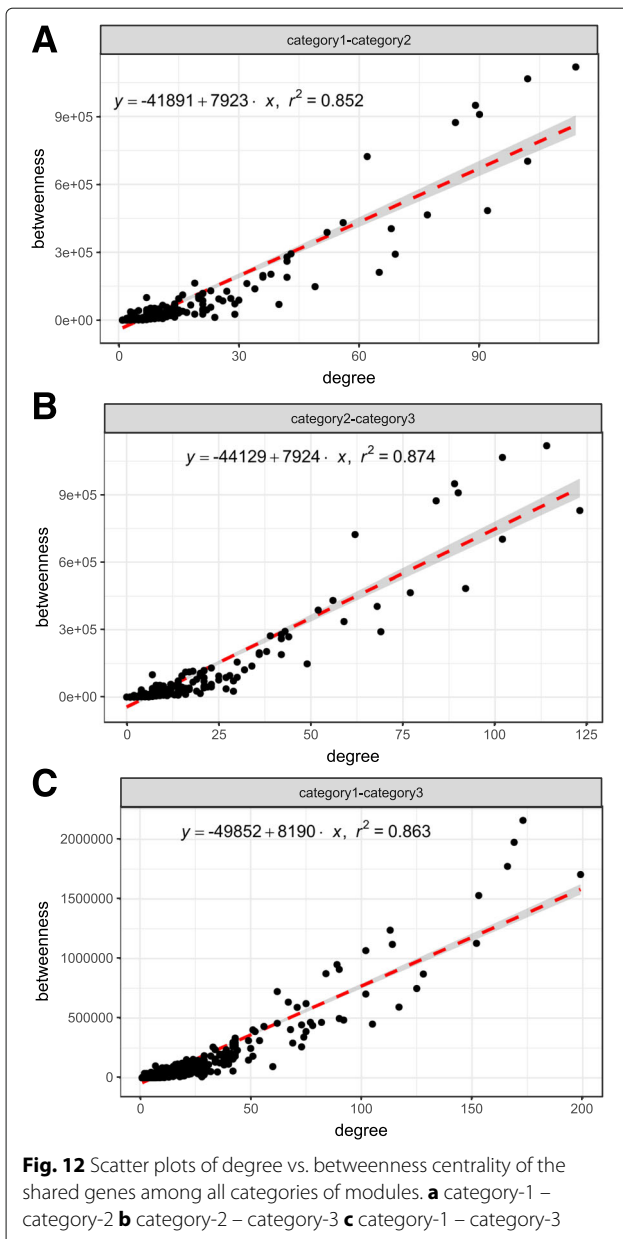
genes of all the stages into a multi-set format. We have chosen the 5600 most connected genes for all the stages of HIV-1 progression which we have discussed earlier in the "Methods" section. A closer look at the Fig. 4b reveals that the number of common genes (54.8%) among the stages has been decreased from our earlier observation. The possible reasons behind this, is that the connectivities among a significant number of expressed common genes with other genes are low compared to the connectivities among expressed non-common genes with other genes. As a supplementary information to the interested readers, we have included Additional files 3 and 4 which show the Venn diagrams of the expressed genes and the 5600 most connected expressed genes (MCEG), respectively, among the uninfected and three stages of HIV-1 infection.

Identification of consensus modules
We have utilized the consensus dissimilarity measure Eq. (7) in average linkage hierarchical clustering method to detect consensus modules. We take a pair of stages at a time and identify consensus modules from the expressed genes. The identified modules are given the same type of color code. The genes which are not assigned to any of the modules are labeled as gray color. We have obtained 14 consensus modules of category-1, as shown in Fig. 5. Similarly, we have obtained 3 category-2 and 16 category-3 modules (shown in Figs. 6 and 7). We have summarized each category modules by their corresponding module eigengenes through Eq. (10) and constructed an eigengene network among them using Eq. (11). For each category of consensus modules there are two sets of genes each corresponding to a specific HIV-1 infection stage. We have included the list of genes which are involved in the formation of all categories of consensus modules in Additional files 5, 6, 7, 8, 9 and 10.

It is worth noting that the number of category-2 modules for the chronic and non-progressor stages pair is relatively small and major genes didn't participate in modules formation (as they fall in gray module). This indicates that the commonness between the expression patterns of chronic and non-progressor is much lower than the other pair of stages.

Fig. 11 Venn diagram showing the count of the Most Connected Expressed Common Genes (MCECG) among acute-chronic (i.e. category–1), chronic–non progressor (i.e. category–2) stages and acute–non progressor (i.e. category–3) stages

Fig. 12 Scatter plots of degree vs. betweenness centrality of the shared genes among all categories of modules. **a** category-1 – category-2 **b** category-2 – category-3 **c** category-1 – category-3

Figures 8, 9 and 10 show the distribution of three categories of modules with their respective OvScore values. It is noticed from these figures that there is very little involvement among the three categories of modules. Most of the modules in each category have low OvScore. To investigate, whether these results have any correlation with the number of common genes that are involved in the consensus modules construction in each pair of the stages, we have performed the following analysis. We have drawn a Venn diagram in Fig. 11 to show the overlap among the common genes which are involved in the consensus modules construction for each pair of stages. It is observed from the figure that 66.8% genes are common among them. The complete list of all overlapped genes for Fig. 11 is provided in Additional file 11. The genes which are preserved between the stages across all the consensus modules are also listed in Additional file 12. After removing the housekeeping genes from the most connected expressed genes for each stage, we have found 57.05% genes are common among the genes involved in consensus module construction. Among those common genes we have also searched for the Immune Regulatory genes [43] and found some of them between the stages across all the consensus modules. We have collected and compiled a list of immune regulatory genes from Immunology Database and Analysis Portal (ImmPort), Immunogenetic Related Information Source (IRIS) and Immunome Database available in InnateDB [44]. The list of such Immune Regulatory genes preserved among the HIV-1 stages across all the consensus modules is available in Additional file 13 and the exclusive set of Immune Regulatory genes expressed in different stages of HIV−1 infection are listed in Additional file 14.

Furthermore, to explore the characteristics of the shared genes belonging to the consensus modules of each pair of stages, we have performed the following analysis. We have investigated the degree and betweenness centrality of the genes considering the whole human genome as an interaction network. Figure 12 (a), (b) and (c) show the scatter plots of degree vs. betweenness centrality of these shared genes. It can be observed from the figure that, there exists a strong correlation between degree and betweenness centrality of the genes in each category. For shared genes between category-1 and category-3, R^2 value (0.883) is slightly more than the shared genes of other pair of categories

Overlaps among the consensus modules

To detect the overlaps among the identified consensus modules, we have applied Eqs. 15 to 16 and obtained the overlap scores (OvScore) for all categories of consensus modules.

Table 1 Gene Ontology (GO) term and KEGG pathway of the shared genes of each category modules

Category pair	GO-Term (GO ID) (*p*-Value)	KEGG Pathway (*p*-Value)
Category-1–category-2	Translational initiation (GO:0006413) (*p*-value: 7.5E-19)	Ribosome (3.1E-7)
Category-1–category-3	Viral process (GO:0016032) (*p*-value: 8.0E-33)	HTLV-I infection (8.8E-12)
Category-2–category-3	Viral transcription(GO:0019083)(*p*-value: 2.4E-14)	Phagosome (4.7E-5)

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 14 of 23

(for category-2 and category-3: 0.874, for category-1 and category-2: 0.852). Some shared genes emerge as both hub (high degree) and bottleneck (high betweenness centrality). For example, genes: 'ACTB' , 'EEF1E1', 'CALM1', 'HSP90AA1', 'RAC1', 'STAT1', 'CSNK2A1' and 'STAT3' have degrees 102, 89, 114, 90, 92, 77 , 152 and 102 and betweenness centrality 1.066E+06, 9.49E+05,

9.9E+05, 1.118E+07, 4.85E+06, 4.65E+06, 1.12E+06, and 7.022E+06, respectively. To further explore the biological relevance of the shared genes, we have searched Gene Ontology (GO) terms and KEGG pathways that are associated with those genes. Table 1 summarizes the results for the shared genes of each category modules.
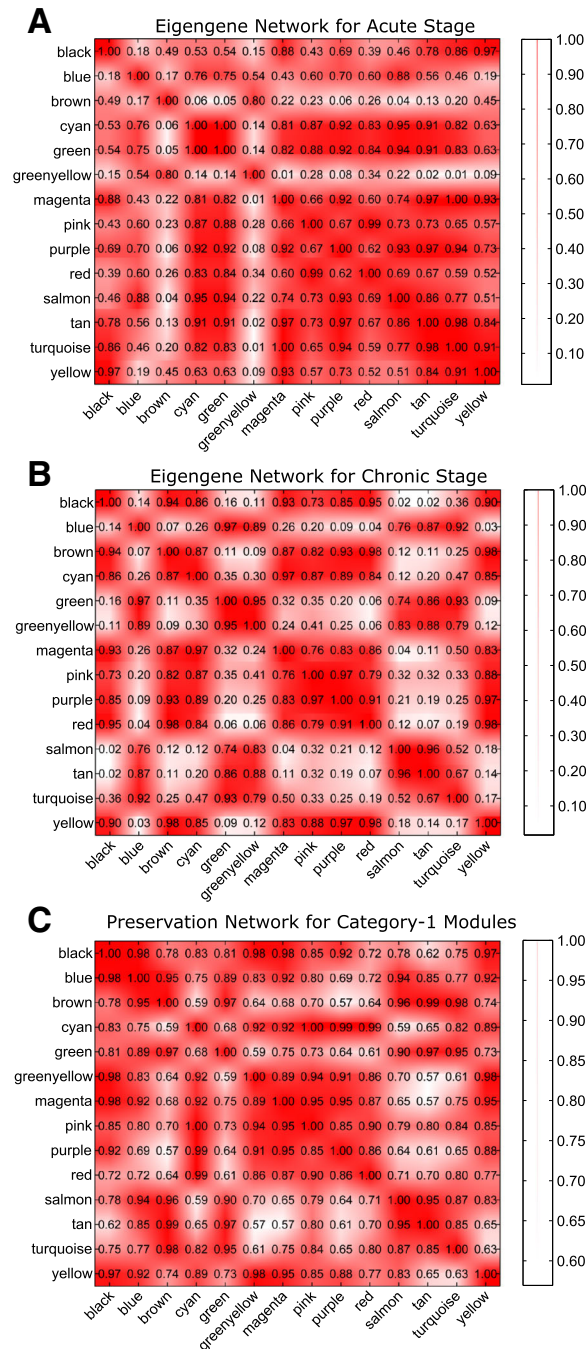
Fig. 13 Figure shows the heatmaps of eigengene networks for 14 category–1 modules. Panel **a** shows the eigengene network for acute stage and Panel **b** shows the same for chronic stage. Panel **c** shows the preservation network for the category-1 modules

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 15 of 23

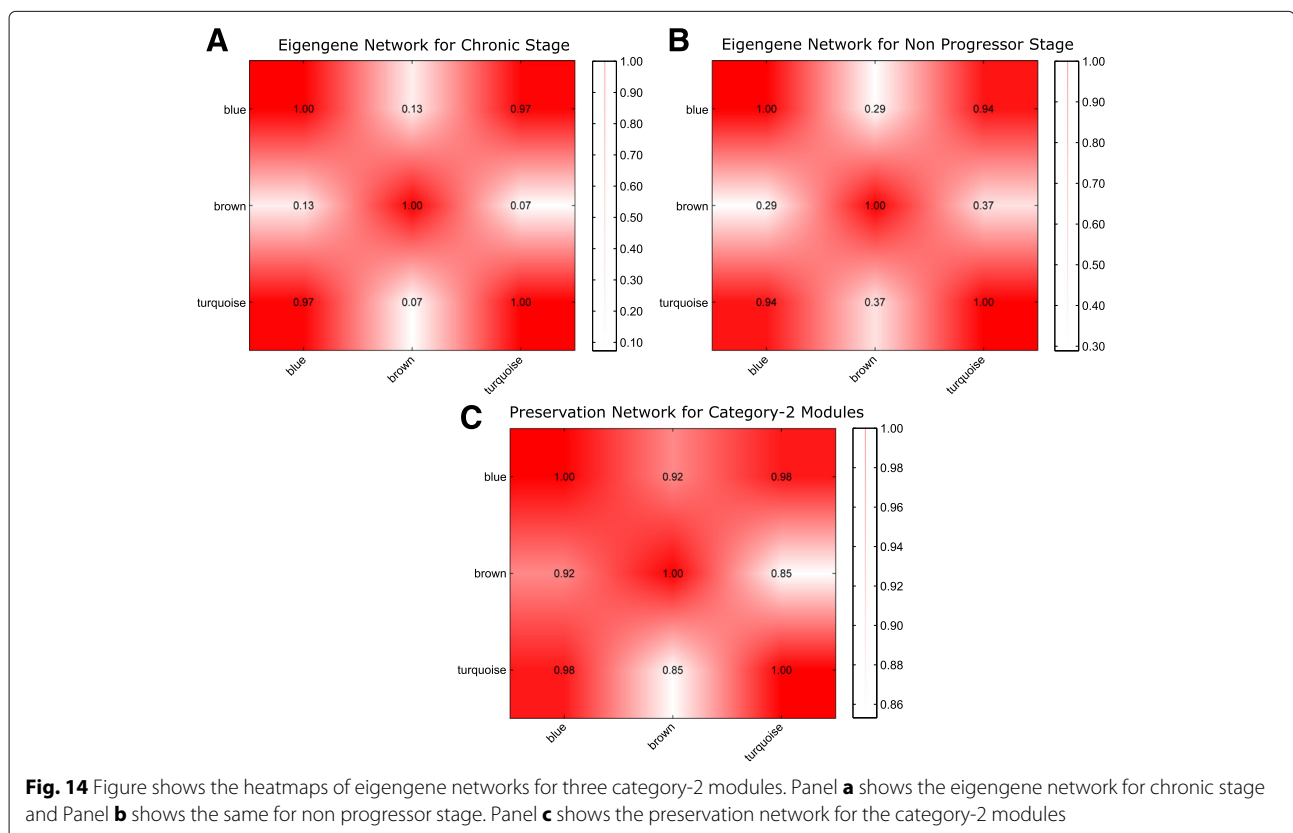## Preservation of consensus modules between each pair of stages

For each pair of stages, consensus modules are identified by using a consensus dissimilarity measure (Eq. (14)) which is utilized in the hierarchical clustering algorithm. The eigengene networks among the consensus modules represent how the characteristic expression patterns of modules are correlated with each other in a particular stage. We have constructed eigengene network corresponding to each infection stage for each category of consensus modules. For example, in category-1 module we have compiled the eigengene networks corresponding to acute and chronic stages. We have employed Eqs. 17 to 18 for comparing these two eigengene networks to know the changes in preservation patterns across each category of consensus modules.

Figure 13(a) and (b) show the heatmap of eigengene networks of category-1 modules corresponding to acute and chronic stages. Figure 13(c) shows the preservation network for the same. It can be noticed from this Figure that five consensus modules 'greenyellow', 'magenta', 'purple', 'pink', and 'red' retain their pairwise correlation pattern across acute and chronic stages. In other words, these modules preserve their expression patterns across acute and chronic stages. For category-2 and category-3

modules the heatmaps of eigengene networks and preservation networks are shown in Figs. 14 and 15, respectively. From Fig. 15(c) we noticed that there are several clusters of modules exist that preserve their expression pattern across acute and non-progressor stages. For example, purple, red, yellow and tan modules retain their pairwise correlation pattern same across acute and non-progressor stages. Another example includes black, blue and brown modules, or magenta, midnightblue and pink modules. For category-2 modules blue and brown modules have a same correlation pattern across chronic and non-progressor stages.

Additionally, for investigating the preservation between module eigengenes across two networks we have computed the Scaled Connectivity (Eq. (19)) of the module eigengenes for each category of modules and the density (Eq. (20)) of their preservation network.

Here, we report the results of preservation measures which are applied to the three categories of modules. Figure 16 shows the distribution of each category of modules with scale connectivity (C) values. As can be seen from the figure, category-1 and category-3 modules show similar types of distribution over the values of C. The density (D) value for category-1 module is 0.7956 whereas for category-3 module the value is 0.8060. The value D for category-2 modules is much higher (0.9195) than that

Fig. 14 Figure shows the heatmaps of eigengene networks for three category-2 modules. Panel **a** shows the eigengene network for chronic stage and Panel **b** shows the same for non progressor stage. Panel **c** shows the preservation network for the category-2 modules

Hossain *et al. BMC Bioinformatics* (2017) 18:181
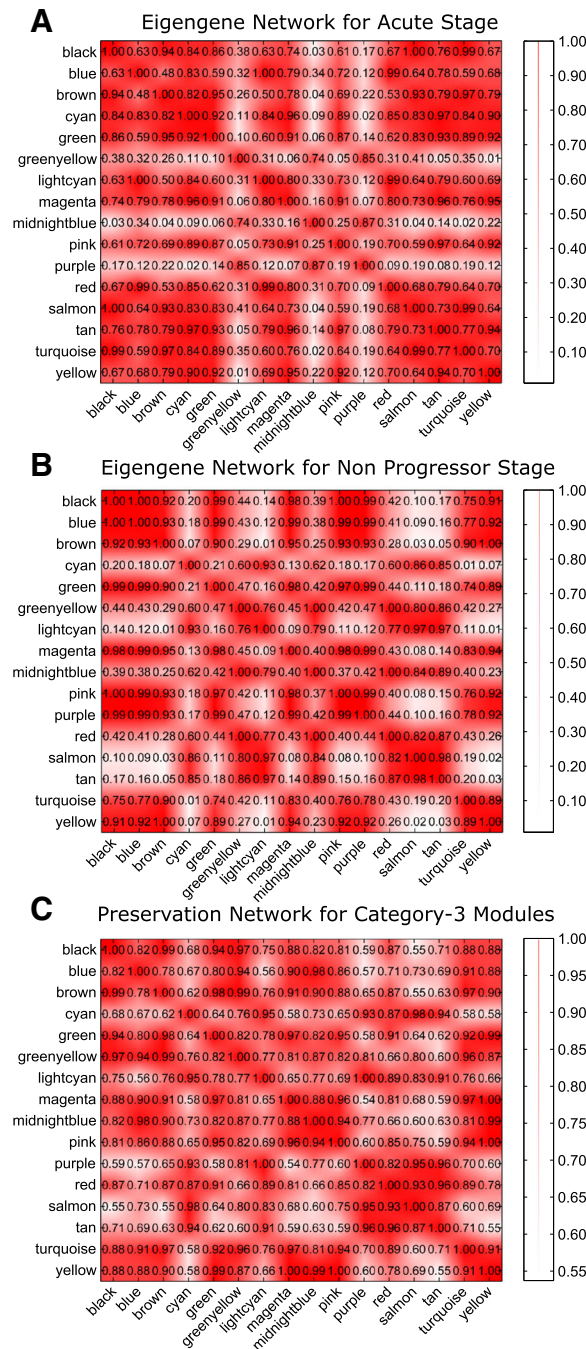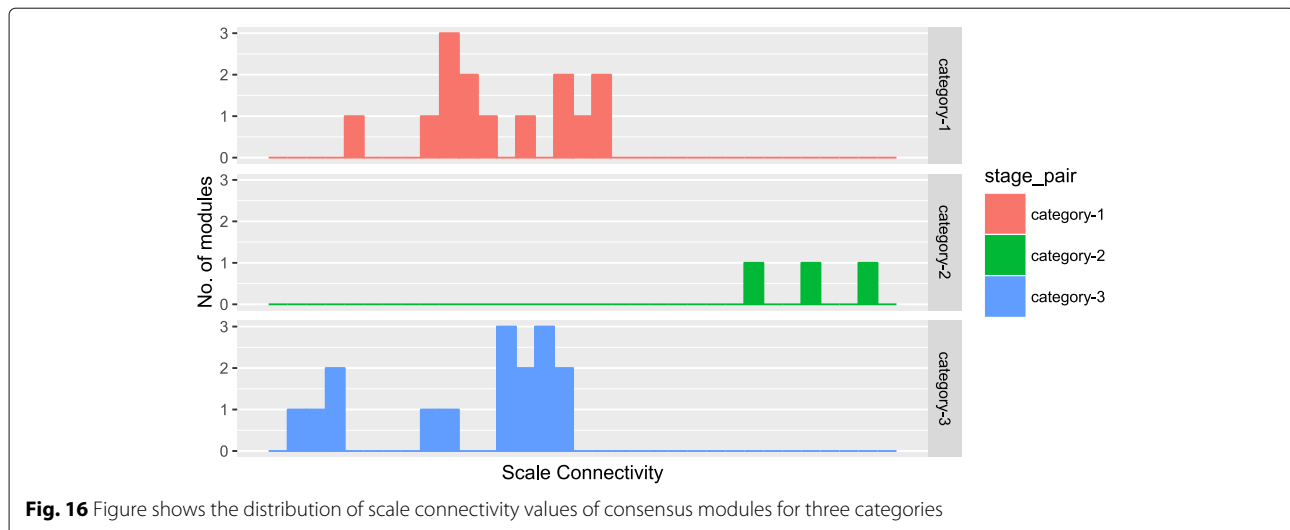
Page 16 of 23

Fig. 15 Figure shows the heatmaps of eigengene networks for 16 category-3 modules. Panel **a** shows the eigengene network for acute stage and Panel **b** shows the same for non progressor stage. Panel **c** shows the preservation network for the category-3 modules

for category-1 and category-3. The possible reason may be that the number of shared modules for chronic and non-progressor stages is only three.

To assess the significance of preservation among the shared modules, we have performed the following statistical test. For this, we have constructed three categories of shared modules randomly from the identified expressed genes of three infection stages. Thus, we obtained 14 random modules for category-1, 3 for category-2 and 16 for category-3. Random modules of each category are constructed by selecting genes randomly from the common expressed genes of a pair of stages. To investigate the preservation pattern of the constructed random modules, we computed the eigengene network and preservation

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 17 of 23

Fig. 16 Figure shows the distribution of scale connectivity values of consensus modules for three categories

matrix for each stage. From this, scale connectivity (C) values are computed for each category of random modules. We compared the C values of random modules with the original modules using Wilcoxon Ranksum test. The resulting *p*-values (7.4678e-06 for category-1 modules, 1.1296e-05 for categoty-3 and 3.2487e-05 for categoty-2) are very low, which signify the preservation of expression pattern between each pair of infection stages is statistically significant.

This suggests that there exists a strong preservation in the overall expression pattern of each category of modules.

Higher order organization of consensus modules
From Figs. 13, 14 and 15, it can be noticed that the shared modules not only preserve their correlation patterns across a pair of infection stages but also form groups or clusters corresponding to each infection stage. For example, in category-1 modules, magenta, green, purple and red modules have high correlation score among them in chronic stage. Similarly, salmon, tan and turquoise modules show high correlation among them in acute stage. This suggests to form a higher order structure of modules that reflects the relationship among them. To investigate the relationship among the modules in each stage, we have performed hierarchical clustering on each category of modules, by using the dissimilarity measure shown in Eq. (13). The identified meta-modules in each category, signify the association among the consensus modules. Figures 17 and 18 show the hierarchical clustering tree to detect meta-modules for category-1 and category-3 modules, respectively. For consensus modules of category-2, no meta-modules are found in both chronic and non-progressor stages. For category-1, we observed five meta-modules (yellow, turquoise, green, blue, and

brown) in acute stage and three meta-modules (turquoise, blue, brown) in chronic stage. In category-3, four meta-modules (turquoise, brown, blue, and yellow) are identified in acute stage, while three meta-modules (turquoise, brown, and blue) are found in non-progressor stage. Such groupings of consensus modules at each stage represent a strong correlation of expression patterns among the modules. From Fig. 17(a) and (b), one can observe the preservation of meta-modules across two stages. For example, in category-1, the first (yellow) and fifth (brown) meta-modules of acute stage are fully preserved in chronic stage. The second meta-module (turquoise) of acute stage is partially preserved in chronic stage. Similarly, from Fig. 18, it can be seen that the first (turquoise) and fourth (yellow) meta-modules of acute stage are highly preserved in non-progressor stage.

From Fig. 13(c), we noticed that a good amount of preservation exists among the consensus modules. It is also observed from Figs. 17 and 18, that the preservation exists in stage-specific meta-modules. So, it is tempting to investigate the preservation pattern among the consensus meta-modules. We detect consensus meta-modules by following the same methodology for consensus module detection. Identified eigengene networks for two stages are clustered using hierarchical clustering by using the dissimilarity measure mentioned in Eq. (14) to form consensus meta-modules. We have found 11 consensus meta-modules for category-1, 2 consensus meta-modules for category-2 and 12 consensus meta-modules for category-3. Figure 5 shows the hierarchical clustering tree for consensus meta-module detection. We noticed in the figure that modules black, yellow and magenta are merged to form a one consensus meta-module whereas cyan and purple form another consensus meta-module. Similarly, Figs. 6 and 7, show the consensus meta-module formation
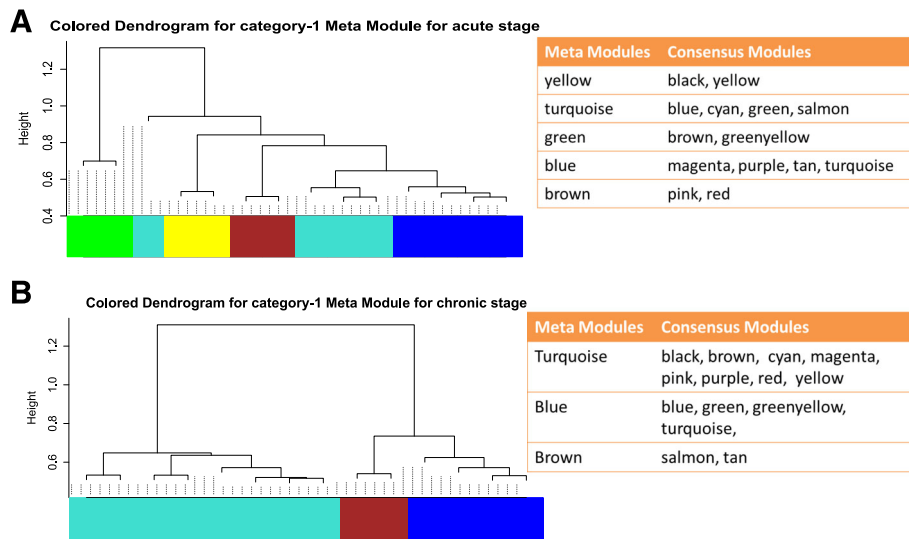
Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 18 of 23

Fig. 17 Figure shows higher order organization of category-1 modules in acute and chronic stages. Hierarchical clustering tree of consensus modules according to acute stage is shown in panel **a**. Panel **b** shows the same for chronic stage. In each panel the grouping of consensus modules into meta-modules are shown

for category-2 and category-3 modules, respectively. Such type of meta-modules represents a grouping of consensus modules between two stages. The difference between the consensus meta-module and simple meta-module is that the consensus meta-modules are constructed from consensus modules by considering a pair of stages. It represents shared meta-modules across two infection stages.

Consistency of module membership with the preservation pattern

In this article, we have also computed the module membership (MM) values of all the genes within a consensus module for the pair of infection stages in each category of modules using Eq. (21) and compared their MM values. Figure 19 shows the comparison of MM values for each pair of preserved category-1 modules. Each panel of Fig. 19 shows a density plot of MM values for preserved category-1 models. Here, we show distribution of MM values for six pairs of category-1 modules with high preservation score. It is evident from the figure that the MM values are consistent with the preservation score of the modules. For example, two preserved shared modules (score=0.99) module 4 (cyan) and module 9 (purple) show similar patterns in the
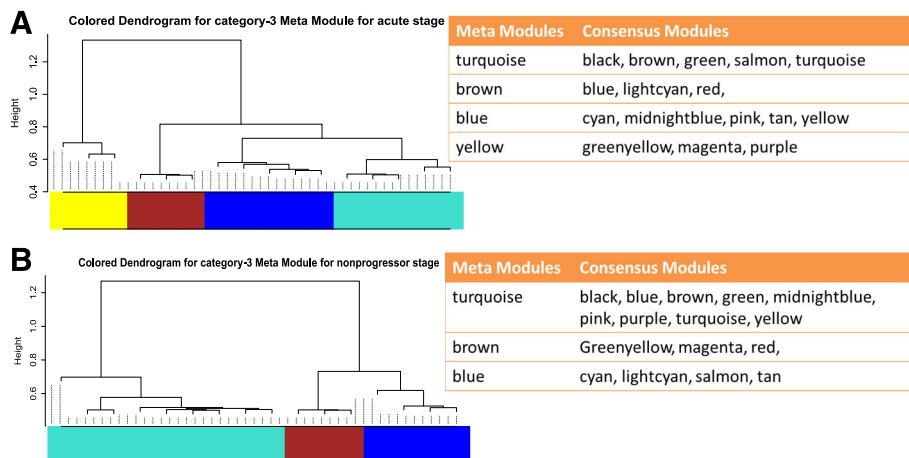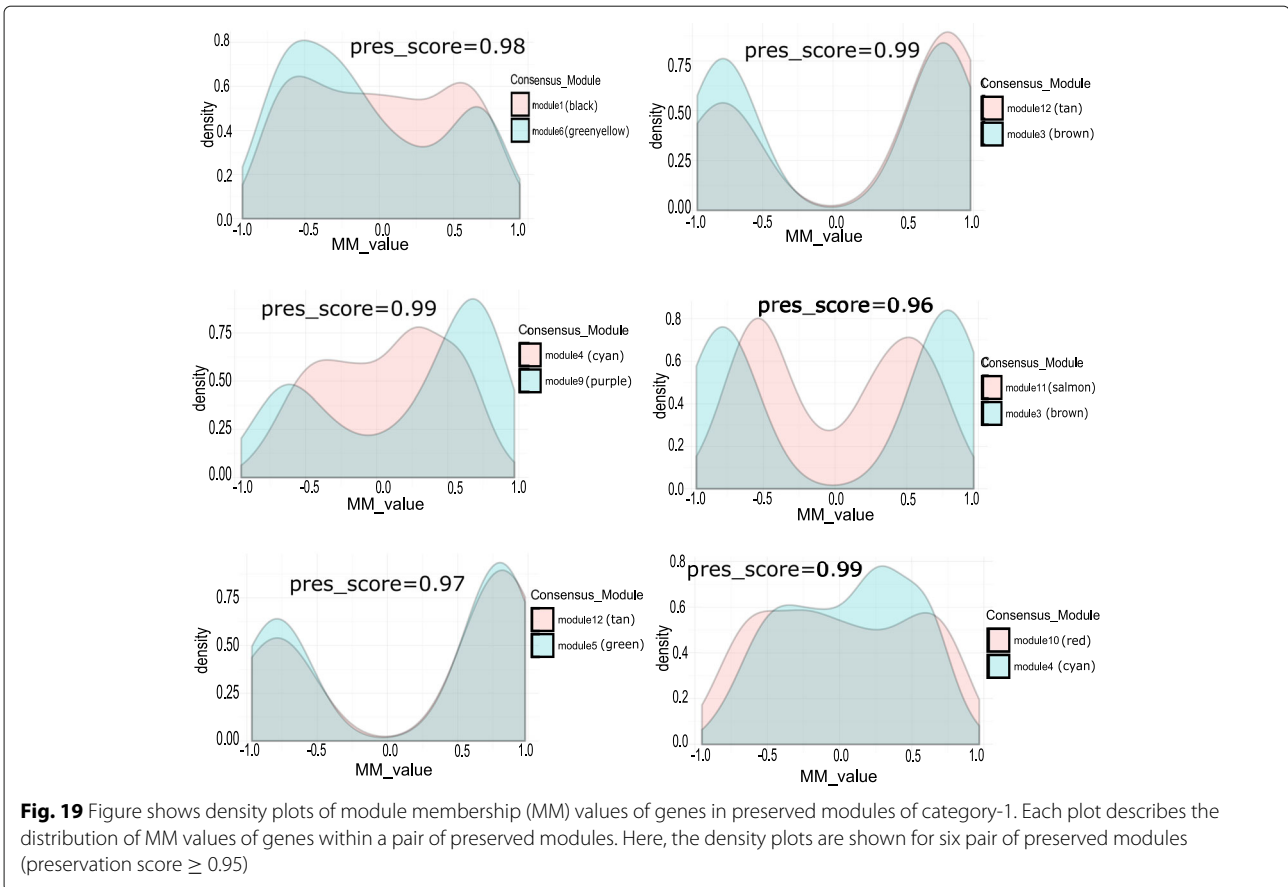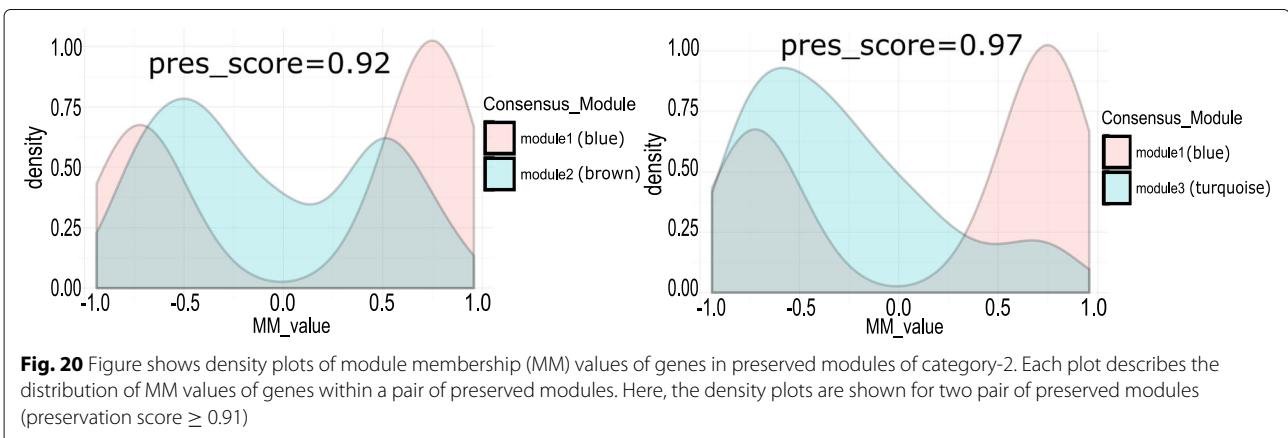

Fig. 18 Figure shows higher order organization of category-3 modules in acute and non-progressor stages. Hierarchical clustering tree of consensus modules according to acute stage is shown in Panel **a**. Panel **b** shows the same for non progressor stage. In each panel the grouping of consensus modules into meta-modules are shown

Fig. 19 Figure shows density plots of module membership (MM) values of genes in preserved modules of category-1. Each plot describes the distribution of MM values of genes within a pair of preserved modules. Here, the density plots are shown for six pair of preserved modules (preservation score ≥ 0.95)

distribution of their MM values. We can observe the same consistency in category-2 and category-3 modules. Figures 20 and 21 show the comparison of MM values between preserved category-2 and category-3 modules, respectively. This suggests that module membership is consistent with the preservation pattern of consensus modules.

Expression analysis of HIV infected individuals before and after ART

An effective suppression of viral replication (≤ 50 copies/mL) following an increase in CD4+ T-cell counts can be observed through Antiretroviral therapy (ART) in HIV infected individuals. In this experiment, we have performed a gene expression analysis with the microarray
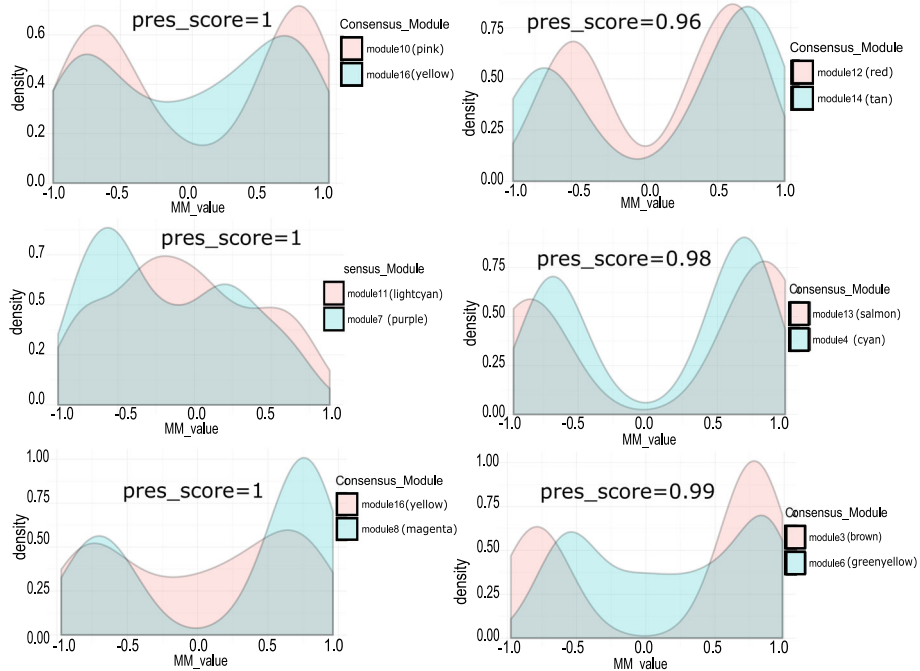

Fig. 20 Figure shows density plots of module membership (MM) values of genes in preserved modules of category-2. Each plot describes the distribution of MM values of genes within a pair of preserved modules. Here, the density plots are shown for two pair of preserved modules (preservation score ≥ 0.91)

Hossain *et al. BMC Bioinformatics*   (2017) 18:181

Page 20 of 23

Fig. 21 Figure shows density plots of module membership (MM) values of genes in preserved modules of category-3. Each plot describes the distribution of MM values of genes within a pair of preserved modules. Here, the density plots are shown for six pair of preserved modules (preservation score ≥ 0.95)

expression dataset (GSE44228 [45]) which provides gene expression values of 36 HIV infected individuals before and after antiretroviral therapy (ART). In this analysis, we have utilized 4,157 differentially expressed genes (DEGs) identified through multivariate permutation tests, provided in [45]. We have investigated expression values

the DEGs in treated and untreated samples, individually. Figure 22 shows box-plots of all the treated and untreated 36 samples. Moreover, to know which genes preserved their expression patterns in both treated and untreated samples, we have computed the Pearson correlations between the expression profiles of all the DEGs in
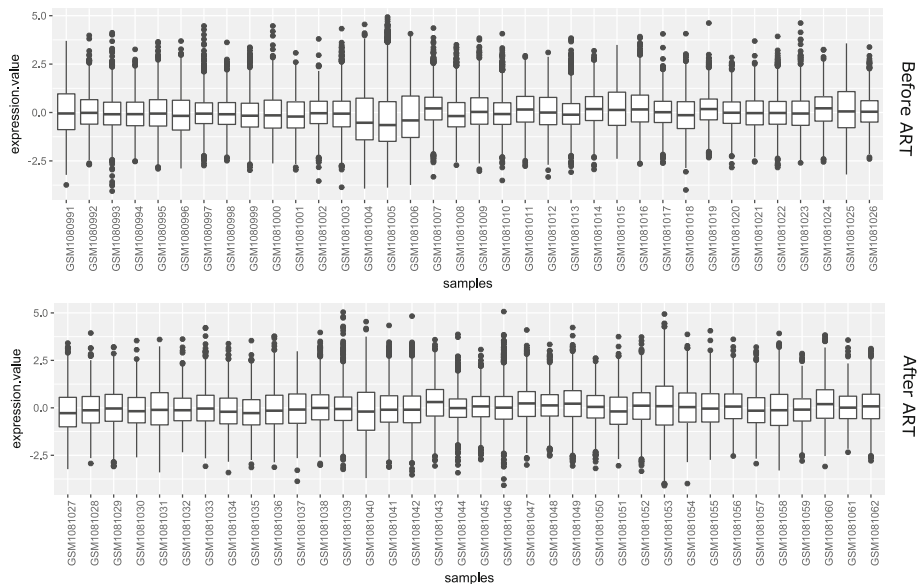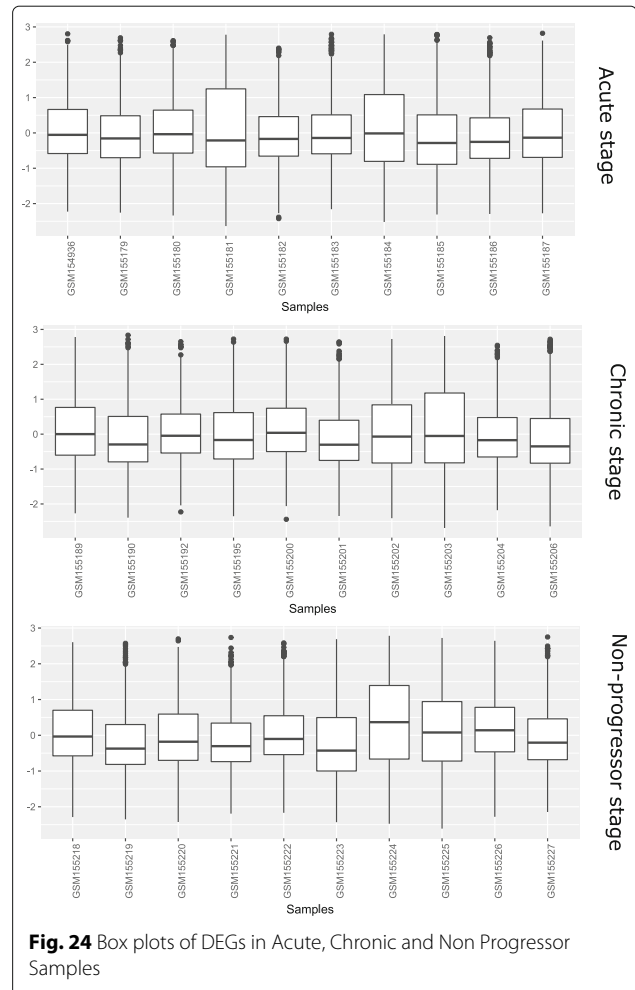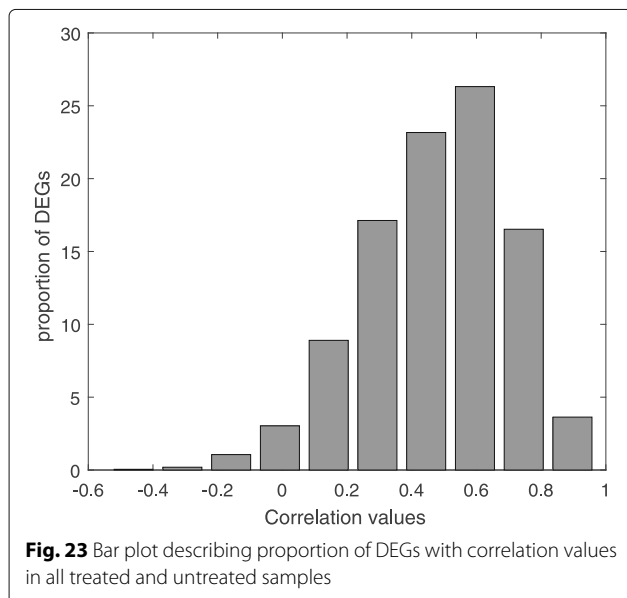

Fig. 22 Figure shows box plots of HIV infected 36 samples before and after Antiretroviral Therapy (ART)

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 21 of 23

both the samples. In Fig. 23, we have shown a bar plot which describes proportions of DEGs with correlation values. It can be observed from this figure that approximately 50% of the DEGs has correlations between 0.4 to 0.6. A small percentage ($\sim$ 5%) of the DEGs has correlations greater than 0.8 and a very few of the DEGs exhibit negative correlations.

We have also compared expression values of the DEGs in three HIV infections stages: acute, chronic and non-progressor. For this, we have collected expression profiles of the DEGs in acute, chronic and non-progressor samples. Figure 24 shows the box-plots of these samples in acute, chronic and non-progressor stages.

Conclusions

In the present article, we have carried out a comprehensive analysis to investigate the preservation pattern of coexpression network compiled from microarray gene expression data of HIV-1 progression stages. Here, three different categories of consensus modules are identified by considering each pair of infection stages at a time. For each category, we have compiled two eigengene networks of a consensus module, corresponding to each infection stage. We have found that eigengene networks are preserved in each pair of infection stages. The preservation pattern is more prominent in category-3 (consensus modules of acute and non-progressor pair) modules. However, there exists little involvement among the consensus modules between three categories. Moreover, the number of consensus modules of category-2 is only three, which indicates the preservation of network properties between chronic and non-progressor stage is not good. However, the preservation scores of blue, brown and turquoise modules in category-2 are high, which signifies the correlation

Fig. 24 Box plots of DEGs in Acute, Chronic and Non Progressor Samples

Fig. 23 Bar plot describing proportion of DEGs with correlation values in all treated and untreated samples

between eigengenes of each pair of modules remain the same in chronic and non-progressor stages. So, the preservation of eigengene in category-2 is high despite having low preservation of network properties between chronic and non-progressor stages.

Observing the preservation pattern of each category of modules in an individual infection stage, we have clustered the modules into groups of meta-modules. Each meta-module is identified in individual infection stage by performing hierarchical clustering which utilizes the dissimilarity measure defined in Eq. (13). The meta-modules are fully or partially preserved across a pair of infection stages. Some meta-modules in category-1 such as 'green' and 'blue' are not preserved between acute and chronic stages. List of the genes involved in those two meta modules are listed in Additional file 15. Similarly, for category-3, meta-module 'brown' is not preserved between acute and non-progressor stages. For category-2, no meta-modules are found and the possible reason behind this is the small number of identified consensus modules. Moreover, the preservation among the

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 22 of 23

consensus meta-modules are also discovered by identifying them using the Eq. (14) from consensus modules.

Apart from the eigengene networks, the preservation among the consensus modules is also observed while comparing the module membership (MM) values of genes within the modules. In other words, the MM values are found to be consistent with the preservation pattern of eigengene networks. In most of the cases, the distribution of MM values between two preserved modules in each category shows a strong correlation. This suggests that the way in which the genes within a pair of preserved module conforms to its characteristic expression pattern is similar.

Some issues still require to be explored further. It is worth mentioning that a clear investigation of the preserved modules through biological experiments can facilitate the understanding of key players or biomarkers that are essential for HIV infection. Apart from that, different machine learning approaches like support vector machines, rule-based systems, random forests, artificial neural networks, etc. may be useful tools for capturing the preservation structure among consensus modules of different stages of HIV-1 infection. Multilabel classifications would be an important tool to predict drug resistance in HIV-1 antiretroviral therapy [46, 47]. Beside, detecting preservation patterns module-wise, it is also interesting to identify the differentially co-expressed modules across a pair of stages in HIV-1 progression. We are now working in this direction.

Additional files

Additional file 1: List of genes expressed in different stages of HIV–1 infection. (XLSX 95.7 kb)

Additional file 2: List of genes exclusively expressed in different stages of HIV–1 infection. (XLSX 21.1 kb)

Additional file 3: Venn diagram showing the count of the expressed Genes among the uninfected and three stages of HIV–1 infection. (EPS 21.1 kb)

Additional file 4: Venn Diagram showing the count of the 5600 most connected expressed genes among the uninfected and three stages of HIV–1 infection. (EPS 210 kb)

Additional file 5: List of acute stage genes involved in category-1 (acute–chronic stage Pair) consensus module formation. (XLSX 23.0 kb)

Additional file 6: List of chronic stage genes involved in category-1 (acute–chronic stage Pair) consensus module formation. (XLSX 22.0 kb)

Additional file 7: List of chronic stage genes involved in category-2 (chronic–non-progressor stage pair) consensus module formation. (XLSX 12.0 kb)

Additional file 8: List of non-progressor stage genes involved in category 2 (chronic–non-progressor stage pair) consensus module formation. (XLSX 11.9 kb)

Additional file 9: List of acute stage genes involved in category-3 (acute–non-progressor stage pair) consensus module formation. (XLSX 21.5 kb)

Additional file 10: List of non-progressor stage genes involved in category-3 (acute–non-progressor stage pair) consensus module formation. (XLSX 21.5 kb)

Additional file 11: List of the most connected expressed common genes (MCECG) in each category of modules. (XLSX 76.2 kb)

Additional file 12: List of genes preserved in consensus modules of each category. (XLSX 11.5 kb)

Additional file 13: List of immune regulatory genes preserved in consensus modules in each category. (XLSX 9.80 kb)

Additional file 14: List of exclusive immune regulatory genes expressed in stages of HIV–1 infection. (XLSX 12.1 kb)

Additional file 15: List of the genes involved in the non–preserved category–1 meta-modules: "green" and "blue". (XLSX 22.7 kb)

Abbreviations
AIDS: Acquired immunodeficiency syndrome; ART: Antiretroviral therapy; Category–1 Modules: Consensus modules of acute and chronic stages; Category–2 Modules: Consensus modules of chronic and non-progressor stages; Category–3 Modules: Consensus modules of acute and non-progressor stages; DEG: Differentially expressed genes; GEO: Gene expression omnibus; GO: Gene ontology; HIV: Human immunodeficiency virus; HDFs: HIV dependency factors; HPID: HIV-1 human protein interaction database; MM: Module membership; ME: Module Eigengene; MCEG: Most connected expressed genes; MCECG: Most connected expressed common genes; OvScore: Overlap score; TOM: Topological Overlap Matrix; WGCNA: Weighted gene co-expression network analysis; SFT: Scale-Free Topology

Author details
[1]Department of Computer Science and Engineering, Aliah University, Kolkata, 700156 West Bengal, India. [2]Department of Computer Science and Engineering, University of Kalyani, Kalyani, 741235 West Bengal, India.

References
1. Sepkowitz KA. Aids-the first 20 years. N Engl J Med. 2001;344(23):1764–72.
2. Krämer A, Kretzschmar M, Krickeberg K. Modern Infectious Disease Epidemiology : Concepts , Methods,Mathematical Models, And Public

Hossain *et al. BMC Bioinformatics* (2017) 18:181

Page 23 of 23

Health. Statistics for Biology and Health. New York: Springer; 2010. doi:10.1007/978-0-387-93835-6.

3. Gallo RC, Montagnier L. The discovery of HIV as the cause of AIDS. N Engl J Med. 2003;349(24):2283–2285.

4. Pantaleo G, Menzo S, et al. Studies in subjects with long-term nonprogressive human immunodeficiency virus infection. N Engl J Med. 1995;332:209–16.

5. Chu C, Selwyn PA. Diagnosis and initial management of acute hiv infection. Am Fam Physician. 2010;81(10):1239–44.

6. Pantaleo G, Graziosi C, Fauci A. New concepts in the immunopathogenesis of human immunodeficiency virus infection. N Engl J Med. 1993;228(5):327–5.

7. Grossman Z, Meier-Schellersheim M, Paul W, Picker L. Pathogenesis of HIV infection: what the virus spares is as important as what it destroys. Nat Med. 2006;12(3):289–95.

8. Mothe B, Ibarrondo J, Llano A, Brander C. Virological, immune and host genetics markers in the control of hiv infection. Dis Markers. 2009;27(3):105–20.

9. Bennett JE, Dolin R, Blaser MJ, Vol. 2. Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases, 8th edn. Philadelphia: Elsevier Health Sciences; 2014.

10. Furlong L. Human diseases through the lens of network biology. Trends Genet. 2013;29(3):150–9.

11. Barabasi AL, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–13.

12. Cai JJ, Borenstein E, Petrov DA. Broker genes in human disease. Genome Biol Evol. 2010;2:815–25. doi:10.1093/gbe/evq064.

13. Bandyopadhyay S, Ray S, Mukhopadhyay A, Maulik U. A review of in silico approaches for analysis and prediction of HIV-1-human protein-protein interactions. Brief Bioinform. 2015;16(5):830–51. doi:10.1093/bib/bbu041.

14. Zhao W, Langfelder P, Fuller T, Dong J, Li A, Hovarth S. Weighted gene coexpression network analysis: state of the art. J Biopharm Stat. 2010;20(2):281–300.

15. Lee H, Hsu A, Sajdak J, Qin J, Pavlidis P. Coexpression analysis of human genes across many microarray data sets. Genome Res. 2004;14(6):1085–94.

16. Elo L, Jarvenpaa H, Oresic M, Lahesmaa R, Aittokallio T. Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process. Bioinformatics. 2007;23(16):2096–103.

17. Oldham M, Horvath S, Geschwind H. Conservation and evolution of gene coexpression networks in human and chimpanzee brains. Proc Natl Acad Sci U S A. 2006;103:17973–8.

18. Stuart J, Segal E, Koller D, Kim S. A gene co-expression network for global discovery of conserved genetic modules. Science. 2003;302(5643):249–55.

19. Carlson M, Zhang B, Fang Z, Mischel P, Horvath S, Nelson S. Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. BMC Genomics. 2006;7(40):. doi:10.1186/1471-2164-7-40.

20. Cai C, Langfelder P, Fuller T, Oldham M, Luo R, et al. Is human blood a good surrogate for brain tissue in transcriptional studies?. BMC Genomics. 2010;11(589):. doi:10.1186/1471-2164-11-589.

21. Langfelder P, Horvath S. Eigengene networks for studying the relationships between co-expression modules. BMC Systems Biol. 2007;1(54):. doi:10.1186/1752-0509-1-54.

22. Ray S, Bandyopadhyay S. Discovering condition specific topological pattern changes in coexpression network: an application to HIV-1 progression. IEEE/ACM Trans Comput Biol Bioinform. 2015;11(4):1086–1099.

23. Ray S, Hossain SMM, Khatun L. Discovering preservation pattern from co-expression modules in progression of HIV-1 disease: An eigengene based approach. In: 2016 IEEE International Conference on Advances in Computing, Communications and Informatics, ICACCI 2016. September 21-24. USA: IEEE; 2016. p. 814–20. doi:10.1109/ICACCI.2016.7732146. http://dx.doi.org/10.1109/ICACCI.2016.7732146.

24. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol. 2005;4:1128–1172. doi:10.2202/1544-6115.1128.

25. Liu WM, Mei R, Di X, Ryder TB, Hubbell E, Dee S, Webster TA, Harrington CA, Ho M-h, Baid J, Smeekens SP. Analysis of high density expression microarrays with signed-rank call algorithms. Bioinformatics. 2002;18(12):1593. doi:10.1093/bioinformatics/18.12.1593.

26. Brass A, Dykxhoorn D, Benita Y, Yan N, Engelman A, Xavier R, Lieberman J, Elledge S. Identification of host proteins required for hiv infection through a functional genomic screen. Science. 2008;319(5865):921–6. doi:10.1126/science.1152725.

27. König R, Zhou Y, et al. Global analysis of host-pathogen interactions that regulate early-stage HIV-1 replication. Cell. 2008;135(1):49–60. doi:10.1016/j.cell.2008.07.032.

28. Zhou H, Xu M, Huang Q, Gates A, et al. Genome-scale RNAi screen for host factors required for HIV replication. cell host microbe. Cell Host Microbe. 2008;4(5):495–504.

29. Fu W, Sanders-Beer B, Katz K, Maglott D, Pruitt K. Human immunodeficiency virus type-1, human protein interaction database at ncbi. Nucleic Acids Res (Database Issue). 2009;37:417–22.

30. Takada I, Kouzmenko A, Kato S. Wnt and ppargamma signaling in osteoblastogenesis and adipogenesis. Nat Rev Rheumatol. 2009;5(8):442–7.

31. Dyer M, Murali T, Sobral B. Supervised learning and prediction of physical interactions between human and hiv proteins. Infect Genet Evol. 2011;11:917–23.

32. Doolittle J, Gomez S. Structural similarity-based predictions of protein interactions between HIV-1 and homo sapiens. Virology. 2010;7(82):. doi:10.1186/1743-422X-7-82.

33. Mukhopadhyay A, Maulik U, Bandyopadhyay S. A novel biclustering approach to association rule mining for predicting HIV-1–human protein interactions. PLoS ONE. 2012;7:32289.

34. Mukhopadhyay A, Ray S, Maulik U. Incorporating the type and direction information in predicting novel regulatory interactions between HIV-1 and human proteins using a biclustering approach. BMC Bioinforma. 2014;15:26.

35. Dong J, Horvath S. Understanding network concepts in modules. BMC Systems Biol. 2007;1(24):. doi:10.1186/1752-0509-1-24.

36. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi A. Hierarchical organigation of modularity in metabolic networks. Science. 2001;297:1551–55.

37. Li A, Horvath S. Network neighborhood analysis with the multi-node topological overlap measure. Bioinformatics. 2007;23(2):222–231. doi:10.1093/bioinformatics/btl581.

38. Yip AM, Horvath S. Gene network interconnectedness and the generalized topological overlap measure. BMC Bioinforma. 2007;8(22):. doi:10.1186/1471-2105-8-22.

39. Langfelder P, Zhang B, Horvath S. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for R. Bioinformatics. 2008;24:719–20.

40. Alter O, Brown P, Botstein D. Singular value decomposition for genome-wide expression data processing and modelling. Proc Natl Acad Sci U S A. 2000;97(18):10101–6.

41. Fuller TF, Ghazalpour A, Aten JE, Drake TA, Lusis AJ, Horvath S. Weighted gene coexpression network analysis strategies applied to mouse weight. Mamm Genome. 2007;18:463–72.

42. Langfelder P, Mischel PS, Horvath S. When is hub gene selection better than standard meta-analysis?. Mamm Genome. 2013;17(8):1–16. doi:10.1371/journal.pone.0061505.

43. Paiardini M, Müller-Trutwin M. HIV-associated chronic immune activation. Immunol Rev. 2013;254(1):78–101.

44. Breuer K, Foroushani A, Laird M, Chen C, Sribnaia A, Lo R, Winsor G, Hancock R, Brinkman F, Lynn2 D. InnateDB: systems biology of innate immunity and beyond–recent updates and continuing curation. Antiviral Res. 2013;41:1228–33.

45. Massanella M, Singhania A, Beliakova-Bethell N, Pier R, Lada SM, White CH, Pérez-Santiago J, Blanco J, Richman DD, Little SJ, Woelk CH. Differential gene expression in HIV-infected individuals following ART. Antivir Res. 2013;100(2):420–8. doi:10.1016/j.antiviral.2013.07.017.

46. Heider D, Senge R, Cheng W, Hüllermeier E. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. Bioinformatics. 2013;29(16):1946–52. doi:10.1093/bioinformatics/btt331.

47. Riemenschneider M, Senge R, Neumann U, Hüllermeier E, Heider D. Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. BioData Mining. 2016;9:10. doi:10.1186/s13040-016-0089-1.