BMC Bioinformatics

METHODOLOGY   **Open Access**

DNLC: differential network local consistency analysis

Jianwei Lu[1,2†], Yao Lu[1†], Yusheng Ding[1†], Qingyang Xiao[3], Linqing Liu[1], Qingpo Cai[4], Yunchuan Kong[4], Yun Bai[5*] and Tianwei Yu[4*]

Abstract

Background: The biological network is highly dynamic. Functional relations between genes can be activated or deactivated depending on the biological conditions. On the genome-scale network, subnetworks that gain or lose local expression consistency may shed light on the regulatory mechanisms related to the changing biological conditions, such as disease status or tissue developmental stages.

Results: In this study, we develop a new method to select genes and modules on the existing biological network, in which local expression consistency changes significantly between clinical conditions. The method is called DNLC: Differential Network Local Consistency. In simulations, our algorithm detected artificially created local consistency changes effectively. We applied the method on two publicly available datasets, and the method detected novel genes and network modules that were biologically plausible.

Conclusions: The new method is effective in finding modules in which the gene expression consistency change between clinical conditions. It is a useful tool that complements traditional differential expression analyses to make discoveries from gene expression data. The R package is available at https://cran.r-project.org/web/packages/DNLC.

Keywords: Biological network, Gene expression, Local Moran's I

Background

The biological system operates by tightly controlling the abundance and activity of thousands of proteins. The regulations and interactions can be summarized as a scale-free network [1–3]. The known networks summarized from existing knowledge, e.g. protein-protein interaction and signal transduction networks, are static in nature. Yet in real biological systems, the activities of the edges on the network are dynamic [4]. In the context of gene expression, nodes on the biological network correspond to genes. The expression levels of genes that are close on the network can change between states of correlated, uncorrelated, or even reversely correlated, depending on the

biological condition [5]. Currently, a number of methods can analyze gene expression data in the context of an existing biological network. Most of the methods try to find "network markers", i.e. small subnetworks that change expression levels in response to clinical conditions [6–17]. Some other methods study the dynamic correlation patterns on the network, without considering the clinical outcome [18–20].

Given the biological network is dynamic, and physiological conditions influence the activity of the edges in the network, it is natural to consider the change in expression consistency, i.e. the co-expression patterns in subnetworks, in response to changing physiological states. So far, no method is available to find changes of expression consistency on the network. In this manuscript, our goal is to develop a computational method to detect genes around which the expression consistency changes significantly in response to physiological states. Finding such genes can reveal important mechanisms related to disease

* Correspondence: yunba@pcom.edu; tianwei.yu@emory.edu
†Jianwei Lu, Yao Lu and Yusheng Ding contributed equally to this work.
5Department of Pharmaceutical Sciences, School of Pharmacy, Philadelphia College of Osteopathic Medicine, Georgia Campus, Suwanee, GA, USA
4Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA
Full list of author information is available at the end of the article

Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 15):489

Page 2 of 10

development, by revealing biological functions that become more tightly regulated or de-regulated in association with disease status. Such a method should be able to complement existing differential expression methods to shed new light on the data.

For this purpose, we borrow the measure of Local Moran's I (LMI) from the field of spatial statistics, which quantifies spatial auto-correlation on a map [21]. We treat the network as a map, and calculate LMI for each node based on its expression value and the expression values of nearby nodes on the network. We then use the LMI values to quantify the local expression consistency around any given node. A high positive LMI value of a node in a specific sample implies that the node has a similar expression value to its neighbors in that sample, and their expression values are either very high or very low. In contrast, a large negative LMI value means the node is a spatial outlier, i.e. a node that has low consistency with its surrounding nodes on the network [22]. By combining LMI scores with the clinical data, and using regression models with local false discovery rate correction [23], our method finds nodes around which local expression consistency change significantly between different clinical conditions. It showed promising result in both simulations and real data analyses.

Methods

Calculating local Moran's I (LMI) score on the network

The overall workflow of the method is shown in Fig. 1. The data contains four pieces: $M_{p \times N}$ is the gene expression matrix with p genes in the rows and N samples in the columns; y is the clinical outcome vector of length N; $G = (V, E)$ is the network between the p genes, where the vertices V correspond to the genes, and the edges E represent functional relations between the genes; $C_{m \times N}$ is the matrix of other clinical variables, such as age, gender *etc*, with m variables in the rows and N samples in the columns. We assume there is a one-to-one match between the genes in the matrix and the nodes in the network. Any unmatched genes/nodes are eliminated from the analysis. To prepare for the analysis, the expression matrix is normalized using normal score transformation for every gene.

We calculate the LMI score for every gene in each sample. The goal of LMI is to quantify the extent to which nodes that are close to a given node have expression values similar to it. The formula of LMI for gene i in sample k is:

$$I_{i,k} = \frac{z_{i,k} - \overline{z_k}}{\sigma_k^2} \sum_{j \neq i} w_{ij} \left(z_{j,k} - \overline{z_k} \right),$$

where $z_{i,k}$ is the expression of gene i in sample k, $\overline{z_k}$ is the average gene expression in sample k, $z_{j,k}$ is the

expression of gene j for all the other genes on the network (where $j \neq i$); σ_k^2 is the variance of expression in sample k; w_{ij} is the weight assigned to gene j, which depends on its distance to gene i on the network.

There can be many strategies for the calculation of weights. The goal is to focus on the small region surrounding gene i on the network. One strategy is to assign the inverse of the distance d_{ij} between gene i and gene j as w_{ij}. Another strategy is to determine w_{ij} using a distance threshold: genes within a distance are given the same weight, while those farther away are given the weight of 0. In this study, we use a truncated Gaussian function to assign the weights,

$$w_{ij} = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-d_{ij}^2/2}, d_{ij} \leq 2 \\ 0, d_{ij} > 2 \end{cases},$$

Where d_{ij} is the length of the shortest path between nodes i and j. The weights are then normalized such that for gene i, $\sum_{j \neq i} w_{ij} = 1$.

The intuition of the approach is as follows: for a given node i, only nodes in its vicinity receive substantial weights. Then the calculation of $I_{i,k}$ essentially takes a weighted sum of the products of $(z_{i,k} - \overline{z_k})$ and all the nodes in the vicinity $(z_{j,k} - \overline{z_k})$, normalized by the variance of the expression levels in the sample. We can see that when $(z_{i,k} - \overline{z_k})$ and most of the $(z_{j,k} - \overline{z_k})$ are of the same sign, and have large absolute values, $I_{i,k}$ will have a large positive value. On the other hand, when $(z_{i,k} - \overline{z_k})$ and most of the $(z_{j,k} - \overline{z_k})$ are of opposite sign, and have large absolute values, then $I_{i,k}$ will be negative with a large absolute value. When there is no expression consistency between the nodes near node i, or if their values are close to zero, $I_{i,k}$ will be close to zero. Thus the LMI value $I_{i,k}$ is a good measure of the expression consistency of node i with its network vicinity.

Selecting differential consistency (DC) genes

After computing $I_{i,k}$ for every node i in every sample k, we have a matrix with the LMI values. The dimension of this LMI matrix is exactly the same as the original gene expression matrix, with p genes in the rows and N samples in the columns. We then find if a gene's LMI score changes significantly between different clinical conditions, while incorporating confounders such as age, race etc.

The procedure here is similar to traditional differential expression analysis where confounders are considered (Table 1). The relationship between the clinical outcome, the LMI score of a gene, and confounders can be described by a generalized linear model:
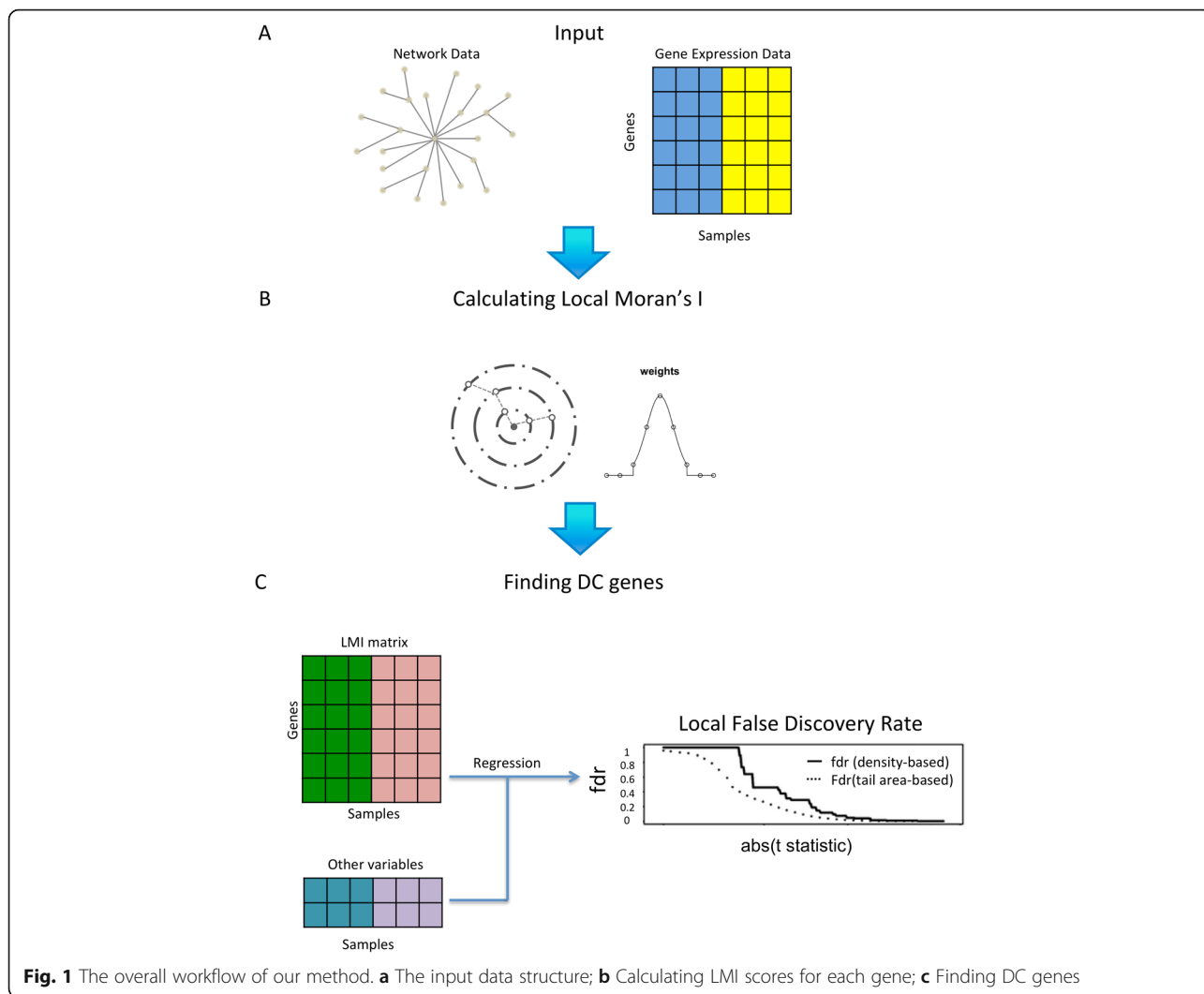
Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 15):489

Page 3 of 10

Fig. 1 The overall workflow of our method. **a** The input data structure; **b** Calculating LMI scores for each gene; **c** Finding DC genes

$$\mathrm{E}(y\,|\mathrm{LMI\_Score}, \mathrm{Confounders})$$
$$= g^{-1}\left(\alpha \times \mathrm{LMI\_Score} + \sum_{m}\beta_m \times \mathrm{Confounder}_m\right),$$

where $g^{-1}(\cdot)$ is an inverse link function, which can be chosen according to the specific type of the outcome variable. In this study we use the logistic regression for

Table 1 The pseudocode for conducting DC gene search on the network

Input: **G** (gene network), **M** (expression matrix), **y** (clinical outcome vector), *T* (local fdr threshold), **C** (confounder matrix)
Output: Collection of DC genes: S
Standardize each row of **M**
Local Moran's I Matrix **I** = Local _ Moran _ I(**G**, **M**)
For each node *i* in **G** do
t_i = t statistic from generalized linear model **y**~**I**$_i$ + **C**
End for
Fit $\{t_i\}_{i=1, ..., p}$ to mixture model using local fdr to find $\{lfdr_i\}_{i=1, ..., p}$
$S = \{i : lfdr_i \leq T\}$
Return S

binary outcome variable, and Cox proportional hazards model for survival outcome variable.

After the t-statistics for the parameter α for all genes are calculated, we follow the local false discovery rate (lfdr) procedure to adjust for multiple testing. For most genes, their local consistency on the network are unrelated to the clinical outcome, and their t-statistics will approximately follow a normal distribution. Genes around which local expression consistency change significantly between clinical conditions will have more extreme t-statistic values. Thus, we can consider the t-statistics of all the genes to follow a mixture model with two components:

$$f(t) = \pi_0 f_0(t) + (1-\pi_0)f_1(t),$$

where f is the mixture density for the observed *t*-statistics of all the genes, f_0 and f_1 are the densities of the *t*-statistics of the null (non-DC) and non-null (DC) genes respectively, and π_0 is the proportion of null genes [23].

Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 15):489

Page 4 of 10

We can estimate the probability that each gene belongs to the non-null category using mixture density estimation. In this study, we use the R package locfdr for the calculation [24]. By setting a threshold for the lfdr value, we can distinguish DC genes from the others.

Finding network communities of DC genes

After selecting the DC genes, we use a simple and efficient algorithm to group the DC genes and their directly connected genes into network communities for better data interpretation. We adopt the fast-greedy algorithm that directly optimizes modularity score to get the communities of a large graph [25]. After detecting several communities among the DC genes and their neighbors, biological function analysis is performed on each detected community. We use the GOstats method [26], which is based on the Gene Ontology biological processes, to perform the analysis.

Results

Simulation study

We conducted a systematic study using simulated data. In each simulation, data was generated using the following steps.

1. A scale-free network with m nodes was generated using the Barabasi-Albert model [27]. Based on this network structure, we calculated the Σ matrix, in which $\Sigma_{i,j} = c^{d_{i,j}}$, where c was a constant between 0 and 1, and $d_{i,j}$ was the shortest path between nodes i and j on the network.
2. An $m \times n$ gene expression matrix was generated using the multivariate normal distribution, using Σ as the variance-covariance matrix.
3. We then randomly selected five nodes from the network, the degree of which were within a certain range. Among the nodes within two hops of these five nodes, we changed the elements of the Σ matrix to $\Sigma_{i,j} = b^{d_{i,j}}$, where b was a constant between 0 and 1, and $d_{i,j}$ was the shortest path between nodes i and j.
4. Another $m \times n$ gene expression data matrix was generated using the multivariate normal density, using the modified Σ matrix as the variance-covariance matrix.
5. We joined the two matrixes horizontally to obtain the simulated expression matrix, which was of dimension $m \times 2n$, where m was the number of genes and $2n$ was the total number of samples.
6. The outcome variable corresponding to first n samples (original correlation samples) were set to 0, and the last n samples (changed correlation samples) were set to 1.

Four parameters were used to control the signal strength of the data: (a) The base correlation c, which controlled the background correlation strength. Four levels were used (base correlation = 0.2, 0.4, 0.6, 0.8). (b) The changed correlation b. We applied four levels (changed correlation = 0.2, 0.4, 0.6, 0.8) for simulation. (c) The degrees of the five selected nodes. Two ranges (degrees between 5 to 10, and between 15 to 20) were used. (d) Sample size in the simulated expression data (number of samples = 50, 100, 200, 300, 400, 500, 700, 1000).

Fifty datasets were simulated at each parameter setting. After a dataset was generated, we used three methods to analyze the data: (1) the DNLC algorithm; (2) the differential network analysis (DNA) method [28], and (3) simple differential expression (DE) analysis using t-test and local fdr correction. We evaluated the results by the PR-AUC (area under the precision-recall curve). Each node was assigned a status depending on the generation process: 1 (changed correlation with neighbors) or 0 (unchanged correlation with neighbors). The local fdr values calculated by each method were used as the predictor variable to generate the precision-recall curve. In each setting, the average area under the curve (AUC) was calculated to reflect the efficacy that the nodes with true local expression consistency changes were differentiated from other nodes.

As shown in Fig. 2, when the base correlation level c was equal to the changed correlation level b, at all settings the PR-AUC values were close to zero (Fig. 2). On the other hand, when the base correlation level c was different from the changed correlation level b, the PR-AUC values increased with the sample size, both in the cases of $b > c$ and in the cases of $b < c$. When the base correlation c was high (0.6 or 0.8), the power to detect the DC nodes was limited (Fig. 2, second and fourth columns). On the other hand, when the base correlation was at low or medium level (0.2 or 0.4), which was close to real data situations, the power was reasonably high when the difference between b and c was high (Fig. 2, first and third columns). In comparison, testing for differential expression didn't detect the differential consistency on the network in most cases. The differential network analysis (DNA) method exhibited some power to detect network differential local consistency, but its AUC values were lower than the DNLC method in most cases, except when the base correlation was high, and the network density was high (Fig. 2, lower-right panel). Although the DNA method seeks differential correlation, it doesn't use network information, which was likely the reason of the inferior performance. Overall, the simulation results validated that the DNLC method was effective in separating the nodes around which the expression consistency were changed.
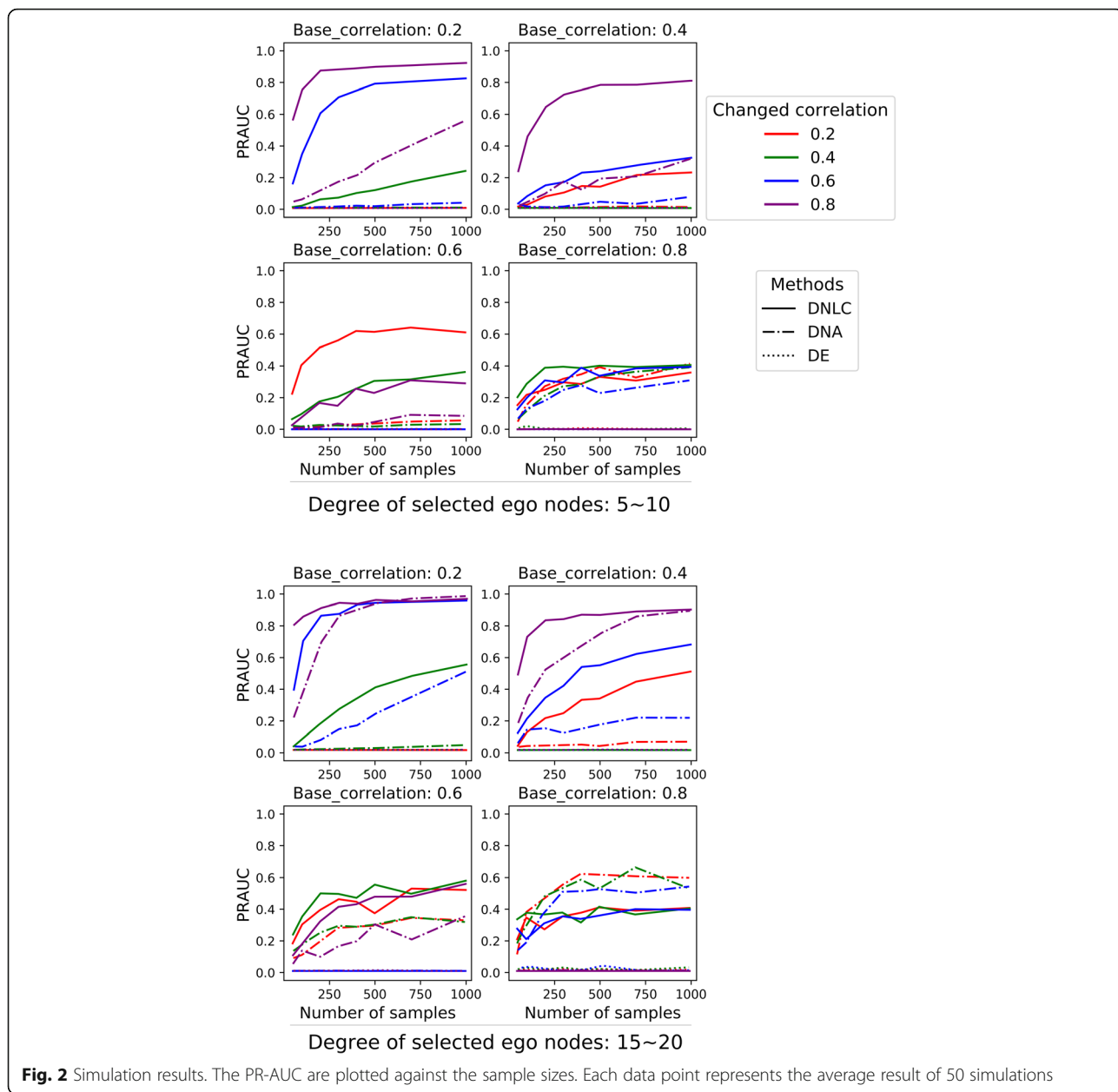
Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 15):489

Page 5 of 10

Fig. 2 Simulation results. The PR-AUC are plotted against the sample sizes. Each data point represents the average result of 50 simulations

Real data analysis: GSE10255

To test our method, we used the High-quality Interaction (HINT) database [29] for the human biological network. Two publicly available expression datasets were analyzed. The first dataset was the gene expression dataset of acute lymphoblastic leukemia (ALL) subjects in response to methotrexate (MTX) treatment (GSE10255) [30]. The dataset contained 12,704 rows (genes) and 161 columns (samples). The clinical outcome variable was the reduction of circulating leukemia cells after MTX treatment. At the lfdr threshold of 0.2, a total of 510 DC genes were selected. Furthermore, network modules were detected among the
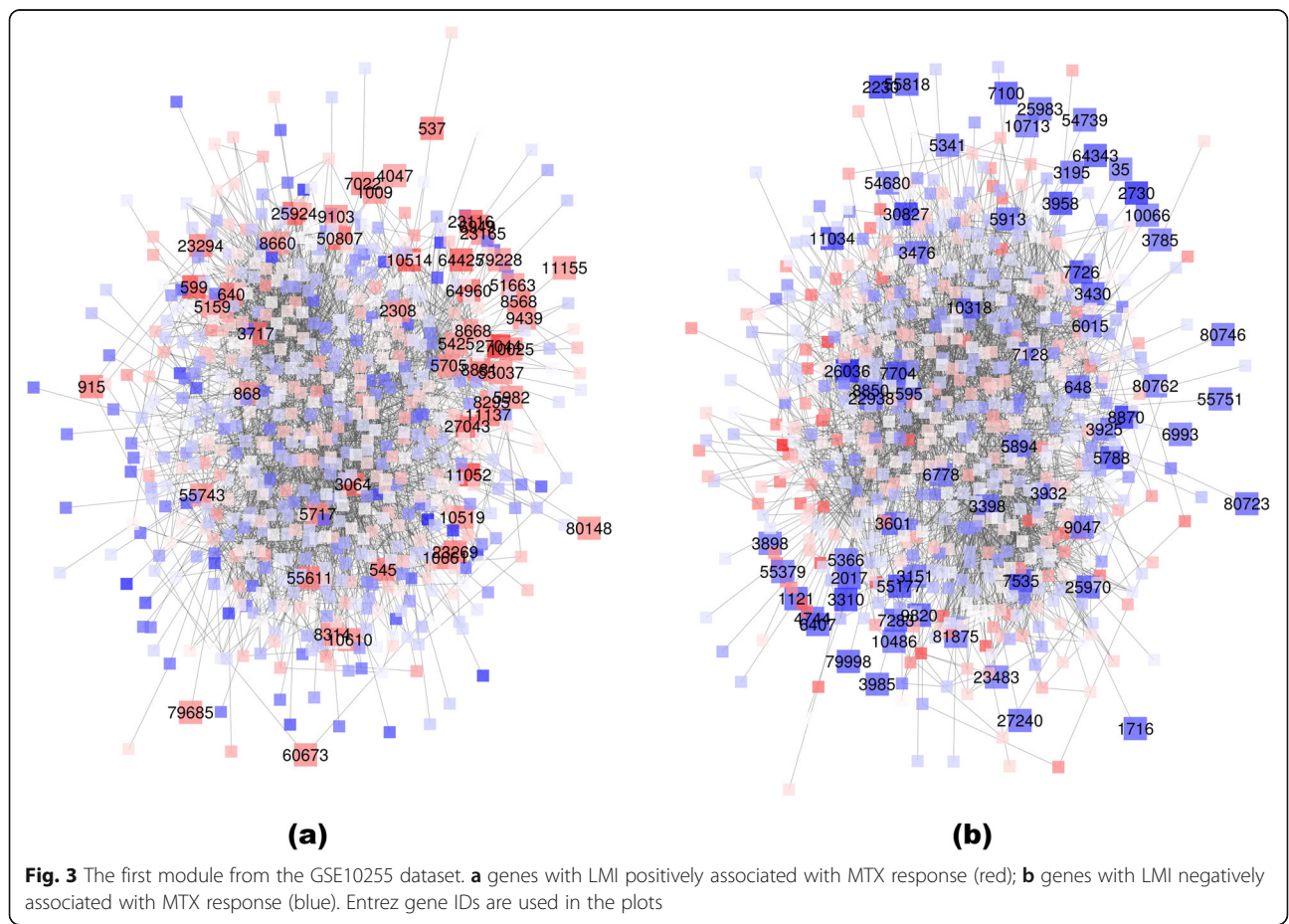
selected genes and their immediate neighbors on the network. In the following discussion, we focus on the largest module. The plots and functional annotations of all the modules are available at web1.sph.emory.edu/users/tyu8/DNLC/MTX.

We used the GOStats package to find gene ontology terms that were over-represented by the lists of genes [26]. For the largest network module (Fig. 3a), the biological processes overrepresented by the positive DC genes, i.e. genes with increased local consistency in patients with higher MTX response, could be categorized into five major groups: phosphorylation and

Lu et al. BMC Bioinformatics 2019, **20**(Suppl 15):489

Page 6 of 10

ubiquitination; peptide hormone secretion; catabolic process; DNA synthetic and repairing; apoptosis and response to hyperoxia. All these functions are closely related to MTX sensitivity in ALL. It has been well-documented that genes that regulate protein modification, apoptosis and DNA synthesis/repair influence caner development [31]. Both phosphorylation and ubiquitination of proteins have been reported to play important roles in MTX resistance in leukemia treatment. Phosphorylation of HSC70, an MTX-binding protein, regulates the transportation of MTX into the cells and contributes to MTX resistance in L1210 leukemia cells [32]. It has also been demonstrated that the MTX chemotherapeutic effect can be significantly reduced by antiepileptic drugs due to the downregulation of reduced folate carrier (Rfc1) activity, regulated by the ubiquitin-proteasome pathway [33]. Among the selected genes by our method, genes 868 (CBLB), 5705 (PSMC5) and 5717 (PSMD11) regulate protein modifications. Many research demonstrated the role of CBLB in leukemia [34, 35], while PSMC5 and PSMD11 were only reported to be involved in cancer development in very recent studies [36–38].

We next focus on the genes that regulate hormone secretion [39], catabolic process [40], and hyperoxia [41], whose roles in ALL treatment are not self-evident. The genes that regulate peptide/protein secretion interfere with the efficacy of chemotherapy to cancer cells by regulating hormone levels. It has been reported that the secretion of peptide hormones are changed in ALL [42]. In addition, some papers reported that insulin and insulin-like factor (IGF) secretions affect the resistance of chemotherapy in ALL patients [43]. Among the selected genes, genes 640 (BLK proto-oncogene), 3717 (JAK2), 8660 (IRS2) and 25,924 (MYRIP) are major genes involved in peptide secretions. JAK2 mutation is believed to have prognostic impact in acute myeloid leukemia [44–46]. The BLK proto-oncogene is involved in leukemia development [47]. IRS2 is an adaptor protein associated with the receptor of erythropoietin, insulin-like growth factor 1. Defective IRS2 expression plays a role in impaired hematopoietic cell differentiation [48].

The selected DC genes also included genes that regulate protein catabolic process. It has been reported that resistance to methotrexate (MTX) in leukemia is related

Fig. 3 The first module from the GSE10255 dataset. **a** genes with LMI positively associated with MTX response (red); **b** genes with LMI negatively associated with MTX response (blue). Entrez gene IDs are used in the plots

Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 15):489

Page 7 of 10

to hydrolase and thymidylate synthase activities, which are catabolic processes [49]. Among the selected genes, 2308 (FOXO1) and 5707 (PSMD1) are regulators of the catabolic process. It has been reported that FOXO1 can play a role in the development of acute myeloid leukemia (AML) [50]. Currently, there are no report about the relationship between PSMD1 and leukemia. However, PSMD1 has been reported to be overexpressed in other cancers [51].
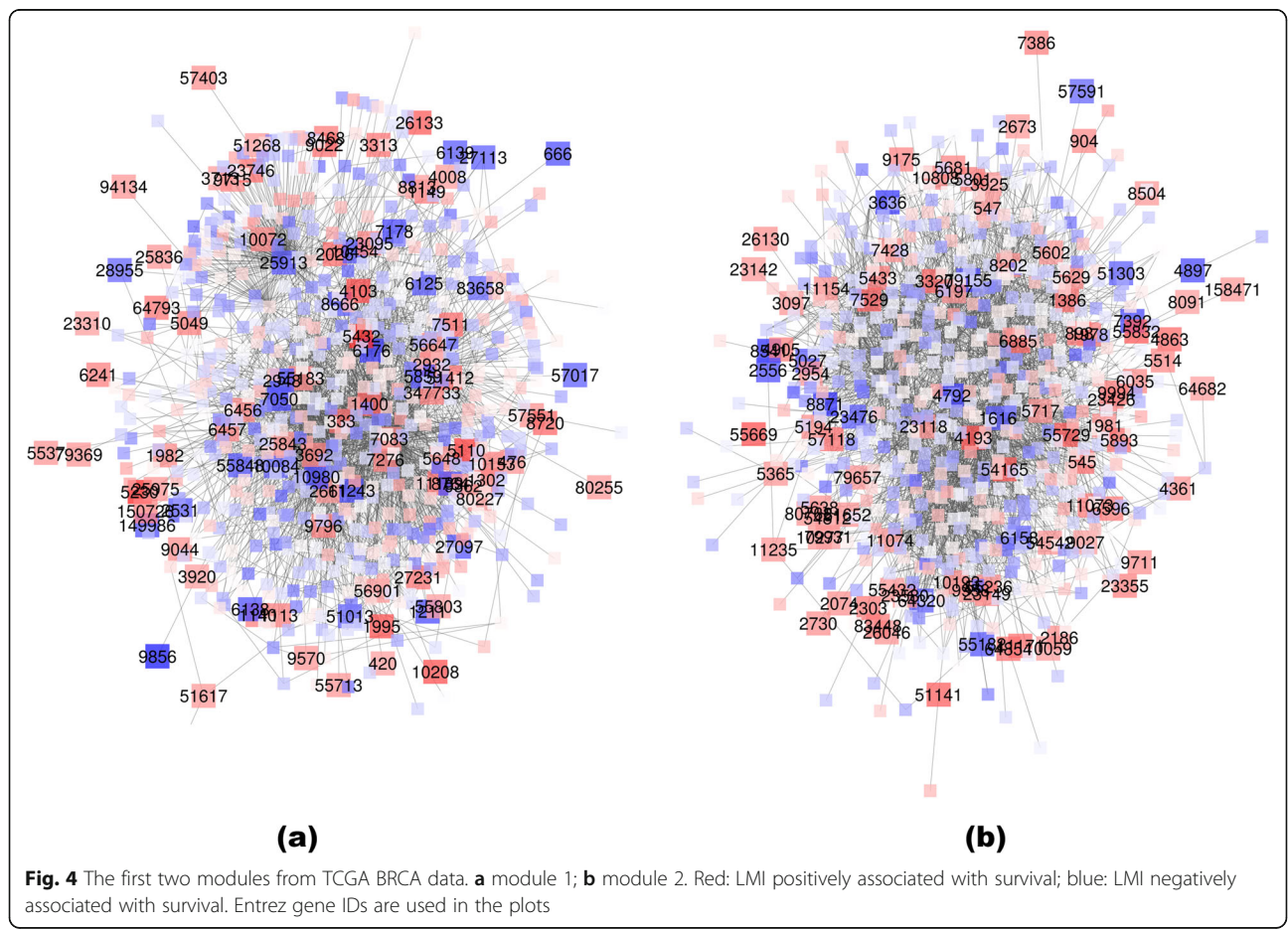
The negative DC genes, i.e. genes with decreased local consistency in patients with higher MTX response (Fig. 3b), were also clearly related to cancer development and progressions. The over-represented GO terms by the negative DC genes include immune cell development and activation [52, 53]; apoptosis [54]; mammary gland epithelium cell proliferation [55, 56]; cell-cell adhesion [57], and cell depolymerization [58]. A number of the selected DC genes are known to affect ALL progression. Also, some of them are known to regulate MTX resistance in leukemia treatment. For example, our method selected genes 595 (CCND1) and 3398 (ID2) that regulate mammary gland epithelial cell proliferation. It has been reported that CCND1 G870A polymorphism is associated with the risk of leukemia and toxicity of MTX

in ALL [59, 60]. ID2 is known to be associated with chemotherapy response and prognosis in acute myeloid leukemia [61].

Real data analysis: TCGA BRCA dataset

We applied the method to a second data set, the breast cancer (BRCA) gene expression dataset from The Cancer Genome Atlas (TCGA). We used the Cox proportional hazards model to link gene LMI values with patient survival outcome, while adjusting for baseline demographic variables including age, gender, and ethnicity. The plots and functional annotations of all the modules are at web1.sph.emory.edu/users/tyu8/DNLC/ BRCA. Again we focus on the largest modules for the discussion here.

In the first module (Fig. 4a), the negative DC genes, i.e. genes with decreased local consistency in patients with lower risk, appear to be more functionally coherent. The biological processes over-represented by the negative DC genes include protein/peptide metabolic process, biogenesis, or membrane targeting and transport, which are obviously related to breast cancer development. As examples, genes 6125 (RPL5) and 6138

Fig. 4 The first two modules from TCGA BRCA data. **a** module 1; **b** module 2. Red: LMI positively associated with survival; blue: LMI negatively associated with survival. Entrez gene IDs are used in the plots

Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 15):489

Page 8 of 10

(RPL15) were among the most significant genes in the list. RPL5 has been reported to be a tumor suppressor gene in breast cancer development [62]. While there is no research paper reporting the role of RPL15 in breast cancer, one study suggested the methylation of RPL15 may be involved in cancer development [63]. Genes 333 (APLP1), 476 (ATP1A1), 1113 (CHGA), and 2026 (ENO2) were on the positive gene list. ATP1A1 has been previously reported to be over-expressed in breast cancer [64]. The overexpression CHGA, a neuroendocrine carcinomas marker, characterizes 10% of infiltrative breast cancer [65]. ENO2 is used as a biomarker to help identify neuroendocrine differentiation in breast cancer [66].

In module 2 (Fig. 4b), the majority of the positive genes were involved in protein ubiquitination, which is a critical process in cancer development [67]. Functional groups of the negative genes include I-Kappa B kinase signaling. Nuclear factor kappa-beta (NF-kappaB) is a transcription factor that modulates the expression of many genes involved in cell proliferation, differentiation, apoptosis and metastasis. Nuclear factor-kappa B is used as a predictor of treatment response in breast cancer [68]. Expression of nuclear factor kappa B (NF-κB) is also used as a predictor of pathologic response to chemotherapy in patients with locally advanced breast cancer [69]. In the I-Kappa B kinase signaling pathway, our method found genes 4792 (NFKBIA), 23,476 (BRD4), and 79,155 (TNIP2) to be significantly associated with breast cancer survival. One study investigated common variants within the gene coding region for NF-kappaB and IkappaB, NFKB1 and NFKBIA, for involvement in sporadic breast cancer. However, the results did not support an involvement of the NFKBIA polymorphisms in sporadic breast cancer in the Caucasian population [70].

The local consistencies of genes 3636 (INPPL1) and 5027 (P2RX7) were also found to be negatively associated to breast cancer survival. They regulate phospholipid dephosphorylation and transport. INPPL1 is also known as SHIP2, which is involved in breast cancer development [71–73]. P2RX7 is also known as P2X7. Purinergic signaling has been implicated in the regulation of many cellular processes and is involved in tumor development and metastasis. Reports revealed that the activation of the P2X7 receptor promotes breast cancer cell invasion and migration, and the P2X7 receptor may be a useful therapeutic target for the treatment of breast cancer [74].
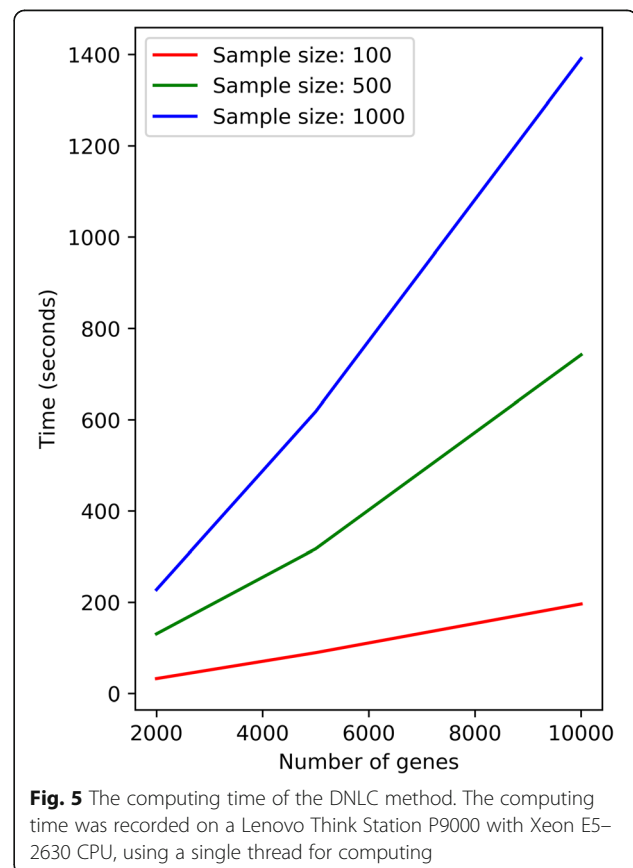
Discussions
In this manuscript, we presented a new method to detect differential consistency (DC) genes on the biological network, as well as network modules where DC genes are concentrated. By using the Local Moran's I (LMI) for measuring local expression consistency on the network, and using the regression framework, the method is versatile enough to be able to study continuous, categorical, and survival outcomes.

Given a large-scale network containing thousands of nodes, the number of possible sub-networks is astronomical. Thus we take the approach of focusing on a specific type of subnetwork: the ego-network, which is defined by the neighborhood of a given node [11]. This approach reduces the number of sub-networks to the number of nodes in the network. The algorithm focuses on the relations between the center node of each sub-network to its surrounding nodes, using the LMI to measure their expression consistency. The computing time of the method increases roughly linearly with the number of genes (nodes) and the sample size (Fig. 5). For example, with 10,000 genes and 500 samples, the method costs roug hly 12 min using single thread computation.

Conclusion
In simulations and real data analyses, we have shown the method is effective in finding genes around which

Fig. 5 The computing time of the DNLC method. The computing time was recorded on a Lenovo Think Station P9000 with Xeon E5–2630 CPU, using a single thread for computing

Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 15):489

Page 9 of 10

expression consistency changes in response to the clinical outcome. The method is a useful tool that complements traditional differential expression type of analyses to make discoveries from gene expression data.

Abbreviations
ALL: Acute lymphoblastic leukemia; BRCA: Breast invasive carcinoma cohort; DC: Differential Consistency; HINT: High-quality Interaction database; lfdr: local false discovery rate; LMI: Local Moran's I; PR-AUC: Area under the precision-recall curve; TCGA: The Cancer Genome Atlas

Authors' contributions
TY and YB conceived the study. QX and TY designed the algorithm. YL, YD, and LL programed the R package. JL, YL, YD, QC and YK conducted data processing. YB conducted results interpretation. YL, YD, YB and TY drafted the manuscript. All authors have read and approved the final manuscript.

Availability of data and materials
The R package is available at https://cran.r-project.org/web/packages/DNLC.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]School of Software Engineering, Tongji University, Shanghai, China. [2]Institute of Advanced Translational Medicine, Tongji University, Shanghai, China. [3]Department of Environmental Health, Emory University, Atlanta, GA, USA. [4]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA. [5]Department of Pharmaceutical Sciences, School of Pharmacy, Philadelphia College of Osteopathic Medicine, Georgia Campus, Suwanee, GA, USA.

References
1. Barabási A-L. Network medicine--from obesity to the "diseasome". N Engl J Med. 2007;357:404–7.
2. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nat Rev Genet. 2011;12:56–68.
3. Chan SY, Loscalzo J. The emerging paradigm of network medicine in the study of human disease. Circ Res. 2012;111:359–74.
4. Ideker T, Krogan NJ. Differential network biology. Mol Syst Biol. 2012;8:565.
5. Yu T. A new dynamic correlation algorithm reveals novel functional aspects in single cell and bulk RNA-seq data. PLoS Comput Biol. 2018;14(8):e1006391.
6. Chen L, Xuan J, Riggins RB, Wang Y, Clarke R. Identifying protein interaction subnetworks by a bagging Markov random field-based method. Nucleic Acids Res. 2013;41:e42.
7. Ciriello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. 2012;22:398–406.
8. Nie Y, Yu J. Mining breast cancer genes with a network based noise-tolerant approach. BMC Syst Biol. 2013;7:49.
9. Su J, Yoon B-J, Dougherty ER. Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. BMC Bioinf. 2010;11(Suppl 6):S8.
10. Taylor IW, Linding R, Warde-Farley D, Liu Y, Pesquita C, Faria D, Bull S, Pawson T, Morris Q, Wrana JL. Dynamic modularity in protein interaction networks predicts breast cancer outcome. Nat Biotechnol. 2009;27:199–204.
11. Yang R, Bai Y, Qin Z, Yu T. EgoNet: identification of human disease ego-network modules. BMC Genomics. 2014;15:314.
12. Zhao YZ, Kang J, Yu TW. A Bayesian nonparametric mixture model for selecting genes and gene subnetworks. Ann Appl Stat. 2014;8(2):999–1021.
13. Sanguinetti G, Noirel J, Wright PC. MMG: a probabilistic tool to identify submodules of metabolic pathways. Bioinformatics. 2008;24(8):1078–84.
14. Wei P, Pan W. Network-based genomic discovery: application and comparison of Markov random field models. J R Stat Soc: Ser C: Appl Stat. 2010;59(1):105–25.
15. Wei P, Pan W. Bayesian joint modeling of multiple gene networks and diverse genomic data to identify target genes of a transcription factor. Ann Appl Stat. 2012;6(1):334–55.
16. Lan Z, Zhao Y, Kang J, Yu T. Bayesian network feature finder (BANFF): an R package for gene network feature selection. Bioinformatics. 2016;32(23):3685–7.
17. Cai Q, Alvarez JA, Kang J, Yu T. Network marker selection for untargeted LC-MS metabolomics data. J Proteome Res. 2017;16(3):1261–9.
18. Yan Y, Qiu S, Jin Z, Gong S, Bai Y, Lu J, Yu T. Detecting subnetwork-level dynamic correlations. Bioinformatics. 2017;33(2):256–65.
19. Wang L, Liu S, Ding Y, Yuan SS, Ho YY, Tseng GC. Meta-analytic framework for liquid association. Bioinformatics. 2017;33:2140–7.
20. Ho YY, Parmigiani G, Louis TA, Cope LM. Modeling liquid association. Biometrics. 2011;67(1):133–41.
21. Anselin L. Local indicators of spatial association—LISA. Geogr Anal. 1995;27(2):93–115.
22. Lalor GC, Zhang C. Multivariate outlier detection and remediation in geochemical databases. Sci Total Environ. 2001;281(1):99–109.
23. Efron B, Tibshirani R. Empirical Bayes methods and false discovery rates for microarrays. Genet Epidemiol. 2002;23(1):70–86.
24. Efron B, Turnbull BB, Narasimhan B. locfdr: Computes local false discovery rates. R Package. 2011;1:1–7.
25. Clauset A, Newman ME, Moore C. Finding community structure in very large networks. Phys Rev E. 2004;70(6):066111.
26. Falcon S, Gentleman R. Using GOstats to test gene lists for GO term association. Bioinformatics. 2007;23(2):257–8.
27. Barabási A-L, Albert R. Emergence of scaling in random networks. Science. 1999;286(5439):509–12.
28. Gill R, Datta S, Datta S. A statistical framework for differential network analysis from microarray data. BMC Bioinf. 2010;11:95.
29. Das J, Yu H. HINT: high-quality protein interactomes and their applications in understanding human disease. BMC Syst Biol. 2012;6:92.
30. Sorich MJ, Pottier N, Pei D, Yang W, Kager L, Stocco G, Cheng C, Panetta JC, Pui CH, Relling MV, et al. In vivo response to methotrexate forecasts outcome of acute lymphoblastic leukemia and has a distinct gene expression profile. PLoS Med. 2008;5(4):e83.
31. Sun X, Saeki K, Masahito T. Apoptosis of B lymphocytic leukemia induced by anticancer drugs and their cell cycle specificity. Zhonghua Zhong Liu Za Zhi. 1998;20(1):25–7.
32. Liu T, Singh R, Rios Z, Bhushan A, Li M, Sheridan PP, Bearden SE, Lai JC, Agbenowu S, Cao S, et al. Tyrosine phosphorylation of HSC70 and its interaction with RFC mediates methotrexate resistance in murine L1210 leukemia cells. Cancer Lett. 2015;357(1):231–41.
33. Halwachs S, Schaefer I, Seibel P, Honscha W. Antiepileptic drugs reduce the efficacy of methotrexate chemotherapy through accelerated degradation of the reduced folate carrier by the ubiquitin-proteasome pathway. Chemotherapy. 2011;57(4):345–56.

Lu *et al. BMC Bioinformatics* 2019, **20**(Suppl 15):489

Page 10 of 10

34. Caligiuri MA, Briesewitz R, Yu J, Wang L, Wei M, Arnoczky KJ, Marburger TB, Wen J, Perrotti D, Bloomfield CD, et al. Novel c-CBL and CBL-b ubiquitin ligase mutations in human acute myeloid leukemia. Blood. 2007;110(3):1022–4.

35. Qu X, Li Y, Liu J, Xu L, Zhang Y, Hu X, Hou K, Liu Y. Cbl-b promotes chemotherapy-induced apoptosis in rat basophilic leukemia cells by suppressing PI3K/Akt activation and enhancing MEK/ERK activation. Mol Cell Biochem. 2010;340(1–2):107–14.

36. Jhaveri DT, Kim MS, Thompson ED, Huang L, Sharma R, Klein AP, Zheng L, Le DT, Laheru DA, Pandey A, et al. Using quantitative Seroproteomics to identify antibody biomarkers in pancreatic Cancer. Cancer Immunol Res. 2016;4(3):225–33.

37. Yim JH, Yun HS, Lee SJ, Baek JH, Lee CW, Song JY, Um HD, Park JK, Kim JS, Park IC, et al. Radiosensitizing effect of PSMC5, a 19S proteasome ATPase, in H460 lung cancer cells. Biochem Biophys Res Commun. 2016;469(1):94–100.

38. Qi T, Zhang W, Luan Y, Kong F, Xu D, Cheng G, Wang Y. Proteomic profiling identified multiple short-lived members of the central proteome as the direct targets of the addicted oncogenes in cancer cells. Mol Cell Proteomics. 2014;13(1):49–62.

39. Voorhess ML, Brecher ML, Glicksman AS, Jones B, Harris M, Krischer J, Boyett J, Forman E, Freeman AI. Hypothalamic-pituitary function of children with acute lymphocytic leukemia after three forms of central nervous system prophylaxis. A retrospective study. Cancer. 1986;57(7):1287–91.

40. Koskelo EK, Saarinen UM, Siimes MA. Low levels of serum transport proteins indicate catabolic protein status during induction therapy for acute lymphoblastic leukemia. Pediatr Hematol Oncol. 1991;8(1):53–9.

41. Wang J, Chen Q, Corne J, Zhu Z, Lee CG, Bhandari V, Homer RJ, Elias JA. Pulmonary expression of leukemia inhibitory factor induces B cell hyperplasia and confers protection in hyperoxia. J Biol Chem. 2003;278(33):31226–32.

42. Pallinger E, Horvath Z, Csoka M, Kovacs GT, Csaba G. Decreased hormone content of immune cells in children during acute lymphocytic leukemia (ALL) - effect of treatment. Acta Microbiol Immunol Hung. 2011;58(1):41–50.

43. Abe S, Funato T, Takahashi S, Yokoyama H, Yamamoto J, Tomiya Y, Yamada-Fujiwara M, Ishizawa K, Kameoka J, Kaku M, et al. Increased expression of insulin-like growth factor i is associated with Ara-C resistance in leukemia. Tohoku J Exp Med. 2006;209(3):217–28.

44. Ayatollahi H, Sadeghian MH, Keramati MR, Karimiani EG, Jafarian AH, Shirdel A, Rahimi H, Zangane-Far ME, Shajiei A, Sheikhi M. JAK2 V617F mutation in adult T cell leukemia-lymphoma. Indian J Hematol Blood Transfus. 2016; 32(4):437–41.

45. Vicente C, Vazquez I, Marcotegui N, Conchillo A, Carranza C, Rivell G, Bandres E, Cristobal I, Lahortiga I, Calasanz MJ, et al. JAK2-V617F activating mutation in acute myeloid leukemia: prognostic impact and association with other molecular markers. Leukemia. 2007;21(11):2386–90.

46. Montero-Ruiz O, Alcantara-Ortigoza MA, Betancourt M, Juarez-Velazquez R, Gonzalez-Marquez H, Perez-Vera P. Expression of RUNX1 isoforms and its target gene BLK in childhood acute lymphoblastic leukemia. Leuk Res. 2012; 36(9):1105–11.

47. Lynch SA, Brugge JS, Fromowitz F, Glantz L, Wang P, Caruso R, Viola MV. Increased expression of the src proto-oncogene in hairy cell leukemia and a subgroup of B-cell lymphomas. Leukemia. 1993;7(9):1416–22.

48. Machado-Neto JA, Favaro P, Lazarini M, da Silva Santos Duarte A, Archangelo LF, Lorand-Metze I, Costa FF, Saad ST, Traina F. Downregulation of IRS2 in myelodysplastic syndrome: a possible role in impaired hematopoietic cell differentiation. Leuk Res. 2012;36(7):931–5.

49. Banerjee D, Mayer-Kuckuk P, Capiaux G, Budak-Alpdogan T, Gorlick R, Bertino JR. Novel aspects of resistance to drugs targeted to dihydrofolate reductase and thymidylate synthase. Biochim Biophys Acta. 2002;1587(2–3):164–73.

50. Kode A, Mosialou I, Manavalan SJ, Rathinam CV, Friedman RA, Teruya-Feldstein J, Bhagat G, Berman E, Kousteni S. FoxO1-dependent induction of acute myeloid leukemia by osteoblasts in mice. Leukemia. 2016;30(1):1–13.

51. Deng S, Zhou H, Xiong R, Lu Y, Yan D, Xing T, Dong L, Tang E, Yang H. Over-expression of genes and proteins of ubiquitin specific peptidases (USPs) and proteasome subunits (PSs) in breast cancer tissue observed by the methods of RFDD-PCR and proteomics. Breast Cancer Res Treat. 2007; 104(1):21–30.

52. Catovsky D, Goldman JM, Okos A, Frisch B, Galton DA. T-lymphoblastic leukaemia: a distinct variant of acute leukaemia. Br Med J. 1974;2(5920):643–6.

53. Marculescu R, Vanura K, Montpellier B, Roulland S, Le T, Navarro JM, Jager U, McBlane F, Nadel B. Recombinase, chromosomal translocations and lymphoid neoplasia: targeting mistakes and repair failures. DNA Repair (Amst). 2006;5(9–10):1246–58.

54. Schuler D, Szende B, Borsi JD, Marton T, Bocsi J, Magyarossy E, Koos R, Csoka M. Apoptosis as a possible way of destruction of lymphoblasts after glucocorticoid treatment of children with acute lymphoblastic leukemia. Pediatr Hematol Oncol. 1994;11(6):641–9.

55. Moriuchi M, Moriuchi H. Induction of lactoferrin gene expression in myeloid or mammary gland cells by human T-cell leukemia virus type 1 (HTLV-1) tax: implications for milk-borne transmission of HTLV-1. J Virol. 2006;80(14):7118–26.

56. Quaglino A, Schere-Levy C, Romorini L, Meiss RP, Kordon EC. Mouse mammary tumors display Stat3 activation dependent on leukemia inhibitory factor signaling. Breast Cancer Res. 2007;9(5):R69.

57. Infante E, Heasman SJ, Ridley AJ. Statins inhibit T-acute lymphoblastic leukemia cell adhesion and migration through Rap1b. J Leukoc Biol. 2011; 89(4):577–86.

58. Chen CW, Lee YL, Liou JP, Liu YH, Liu CW, Chen TY, Huang HM. A novel tubulin polymerization inhibitor, MPT0B206, downregulates Bcr-Abl expression and induces apoptosis in imatinib-sensitive and imatinib-resistant CML cells. Apoptosis. 2016;21(9):1008–18.

59. Qin LY, Zhao LG, Chen X, Yang Z, Mo WN. The CCND1 G870A gene polymorphism and leukemia or non-Hodgkin lymphoma risk: a meta-analysis. Asian Pac J Cancer Prev. 2014;15(16):6923–8.

60. Xue Y, Rong L, Tong N, Wang M, Zhang Z, Fang Y. CCND1 G870A polymorphism is associated with toxicity of methotrexate in childhood acute lymphoblastic leukemia. Int J Clin Exp Pathol. 2015;8(9):11594–600.

61. Zhou JD, Ma JC, Zhang TJ, Li XX, Zhang W, Wu DH, Wen XM, Xu ZJ, Lin J, Qian J. High bone marrow ID2 expression predicts poor chemotherapy response and prognosis in acute myeloid leukemia. Oncotarget. 2017;8(54):91979–89.

62. Fancello L, Kampen KR, Hofman IJ, Verbeeck J, De Keersmaecker K. The ribosomal protein gene RPL5 is a haploinsufficient tumor suppressor in multiple cancer types. Oncotarget. 2017;8(9):14462–78.

63. Pavlova TV, Kashuba VI, Muravenko OV, Yenamandra SP, Ivanova TA, Zabarovskaia VI, Rakhmanaliev ER, Petrenko LA, Pronina IV, Loginov VI, et al. Technology of analysis of epigenetic and structural changes of epithelial tumors genome with NotI-microarrays by the example of human chromosome. Mol Biol (Mosk). 2009;43(2):339–47.

64. Bogdanov A, Moiseenko F, Dubina M. Abnormal expression of ATP1A1 and ATP1A2 in breast cancer. F1000Res. 2017;6:10.

65. Annaratone L, Medico E, Rangel N, Castellano I, Marchio C, Sapino A, Bussolati G. Search for neuro-endocrine markers (chromogranin a, synaptophysin and VGF) in breast cancers. An integrated approach using immunohistochemistry and gene expression profiling. Endocr Pathol. 2014;25(3):219–28.

66. Soh MA, Garrett SH, Somji S, Dunlevy JR, Zhou XD, Sens MA, Bathula CS, Allen C, Sens DA. Arsenic, cadmium and neuron specific enolase (ENO2, gamma-enolase) expression in breast cancer. Cancer Cell Int. 2011;11(1):41.

67. Huang X, Dixit VM. Drugging the undruggables: exploring the ubiquitin system for drug development. Cell Res. 2016;26(4):484–98.

68. Garg AK, Hortobagyi GN, Aggarwal BB, Sahin AA, Buchholz TA. Nuclear factor-kappa B as a predictor of treatment response in breast cancer. Curr Opin Oncol. 2003;15(6):405–11.

69. Prajoko YW, Aryandono T. Expression of nuclear factor kappa B (NF-kappaB) as a predictor of poor pathologic response to chemotherapy in patients with locally advanced breast cancer. Asian Pac J Cancer Prev. 2014;15(2):595–8.

70. Curran JE, Weinstein SR, Griffiths LR. Polymorphic variants of NFKB1 and its inhibitory protein NFKBIA, and their involvement in sporadic breast cancer. Cancer Lett. 2002;188(1–2):103–7.

71. Fu CH, Lin RJ, Yu J, Chang WW, Liao GS, Chang WY, Tseng LM, Tsai YF, Yu JC, Yu AL. A novel oncogenic role of inositol phosphatase SHIP2 in ER-negative breast cancer stem cells: involvement of JNK/vimentin activation. Stem Cells. 2014;32(8):2048–60.

72. Prasad NK. SHIP2 phosphoinositol phosphatase positively regulates EGFR-Akt pathway, CXCR4 expression, and cell migration in MDA-MB-231 breast cancer cells. Int J Oncol. 2009;34(1):97–105.

73. Prasad NK, Tandon M, Handa A, Moore GE, Babbs CF, Snyder PW, Bose S. High expression of obesity-linked phosphatase SHIP2 in invasive breast cancer correlates with reduced disease-free survival. Tumour Biol. 2008;29(5):330–41.

74. Xia J, Yu X, Tang L, Li G, He T. P2X7 receptor stimulates breast cancer cell invasion and migration via the AKT pathway. Oncol Rep. 2015;34(1):103–10.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.