PLOS ONE

RESEARCH ARTICLE

Integrated micro/messenger RNA regulatory networks in essential thrombocytosis

Lu Zhao[1]*, Song Wu[1], Erya Huang[1], Dimitri Gnatenko[2], Wadie F. Bahou[2], Wei Zhu[1]

1 Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY, United States of America, **2** Department of Medicine, Stony Brook University, Stony Brook, NY, United States of America

* luzhao1986@gmail.com

Citation: Zhao L, Wu S, Huang E, Gnatenko D, Bahou WF, Zhu W (2018) Integrated micro/messenger RNA regulatory networks in essential thrombocytosis. PLoS ONE 13(2): e0191932. https://doi.org/10.1371/journal.pone.0191932

Editor: Geraldo A Passos, University of São Paulo, BRAZIL

Received: August 26, 2017

Accepted: January 15, 2018

Published: February 8, 2018

Data Availability Statement: The original data is publicly available at https://github.com/LaoZZZZZ/mirna_mrna_data.

Funding: Study Design is supported by HL091939 https://www.nih.gov/ National Institutes of Health WB. Data Collection is supported Subcontract G175 from https://www.nhlbi.nih.gov/research/resources/genetics-genomics/rsg NIH/NHLBI DNA Resequencing and Genotyping Service, WB. Modeling and data analysis is supported by HL091939 https://www.nih.gov/ National Institutes of Health WZ. Modeling and data analysis is also

Abstract

Essential thrombocytosis (ET) is a chronic myeloproliferative disorder with an unregulated surplus of platelets. Complications of ET include stroke, heart attack, and formation of blood clots. Although platelet-enhancing mutations have been identified in ET cohorts, genetic networks causally implicated in thrombotic risk remain unestablished. In this study, we aim to identify novel ET-related miRNA-mRNA regulatory networks through comparisons of transcriptomes between healthy controls and ET patients. Four network discovery algorithms have been employed, including (a) Pearson correlation network, (b) sparse supervised canonical correlation analysis (sSCCA), (c) sparse partial correlation network analysis (SPACE), and, (d) (sparse) Bayesian network analysis–all through a combined data-driven and knowledge-based analysis. The result predicts a close relationship between an 8-miRNA set (miR-9, miR-490-5p, miR-490-3p, miR-182, miR-34a, miR-196b, miR-34b*, miR-181a-2*) and a 9-mRNA set (CAV2, LAPTM4B, TIMP1, PKIG, WASF1, MMP1, ERVH-4, NME4, HSD17B12). The majority of the identified variables have been linked to hematologic functions by a number of studies. Furthermore, it is observed that the selected mRNAs are highly relevant to ET disease, and provide an initial framework for dissecting both platelet-enhancing and functional consequences of dysregulated platelet production.

1. Introduction

Platelets are anucleate blood cells generated from bone marrow megakaryocytes, and play an important role in haemostasis and thrombosis. Thrombocytosis is a disorder of platelet overproduction in the blood. It is classified as essential/primary thrombocytosis (ET) or reactive/secondary thrombocytosis (RT) by the causes. Essential thrombocytosis is a chronic myeloproliferative disorder with an unregulated surplus of platelets attributed to a malfunction in the body's feedback system. Complications of ET include stroke, heart attack, and formation of blood clots. Mutations involving *JAK2*, *CALR*, and *c-MPL* are identified in the majority of ET cohorts, although genetic risk substratification associated with thrombotic (or hemorrhagic) predisposition remains unknown [1].

Recent data have demonstrated that both megakaryocytes and platelets retain an abundant and diverse array of mRNAs and microRNAs (miRNAs) [2]. miRNAs are a class of non-coding 21- to 24-bp species that primarily regulate protein translation by post-transcriptional

supported by CA205172 https://www.nih.gov/ National Institutes of Health SW. Data collection is also supported by RR03211201A1 https://www.nih.gov/ National Institutes of Health DG.

Competing interests: The authors have declared that no competing interests exist.

targeting of 3'—UTRs [3], which subsequently regulates mRNA translation activity or stability [4]. Emerging evidence has implicated miRNAs in the control of megakaryocytopoiesis [5] and in progenitor fate during the megakaryocyte-erythroid transition. Distinct miRNA expression patterns have been described in differentiated hematopoietic cells [6] and in subsets of patients with myeloproliferative neoplasms [7, 8]. The miRNAs have effects on protein synthesis through regulating mRNA destabilization or translational repression [4]; indeed although quiescent platelets display minimal translational activity, maximally-activated platelets retain the capacity for protein synthesis, with implications for modulating arthritis-associated inflammation [9] or the production of platelet progeny *in vivo* [10].

Many computational methods have been developed to study interactions between miRNA and mRNA, which are largely based on two types of methods: one is computation-based method that uses the sequence complementarities of miRNA and its mRNA targets to build *in silicon* interaction databases, including MiRBase [11], TargetScan [12, 13] and so on; the second is experimental data-based method that examines expression profiles of miRNAs and mRNAs for negative correlations. For example, GenMiR++ [14, 15] and HOCTAR [16] predicts the interaction between miRNA and mRNA by integrating the expression profiling and sequence-based recognition software. Several other methods that are based solely on expression profile have also been published. Jayaswal et al. [17] developed a two-stage procedure that first clusters each expression data for miRNA and mRNA and then identify significant miRNA-mRNA relationship using t-test. Li et al. [18] proposed a method to find a set of differentially expressed miRNAs and mRNAs via Partial Least Squares Regression. It is very challenging to build causal relationship using observational data. Le et al. [19] designed an algorithm to uncover the causal regulatory relationship between miRNAs and mRNAs, using expression profiles of miRNAs and mRNAs without taking into consideration the previous target information. It is based on Intervention calculus when the Directed Acrylic Graph (DAG) is absent (IDA) [20]. While all the above methods focus on uncovering interaction between individual miRNA and mRNA, there is a growing body of literature showing that multiple miRNAs are coordinated by forming cohesive groups to collectively regulate one or more mRNAs [21]. The complex regulatory network formed between a group of miRNAs and a group of mRNAs acts as a vital force in catering similar functioning miRNAs and mRNAs together, and may provide better understandings on the underlying miRNA-mRNA regulatory modules (MMRMs) [22].

In this study, we explore the potential miRNA/mRNA regulatory networks associated to essential thrombocytosis based on a 43-member cohort (13 ET patients and 30 controls), through a combination of data-driven and knowledge-based analyses. Three classes of correlation network analyses methods, namely, the Pearson correlation network, the sparse canonical correlation network, and the sparse partial correlation network have been implemented, compared and integrated to obtain a more reliable and robust miRNA-mRNA pathway. This pathway was subsequently examined for its biological functionalities through an Ingenuity Pathway Analysis. Additionally, we have applied a sparse Bayesian Network analysis, the A* Lasso, to compare with the three Frequentist network analysis methods.

2. Methods

2.1 Patient recruitment, sample processing and data description

Subject recruitment (along with normal healthy controls) was completed by written consent through a study approved by the Stony Brook IRB (Institutional Review Board) Committee on Research Involving Human Subjects (approval period 1999 –present). Enrollment proceeded over a 3-year period and was restricted to adults (>21 years of age) meeting clinical and laboratory criteria for essential thrombocytosis as previously described (38). Patients were

randomly enrolled from the larger pool of patients referred for evaluation of thrombocytosis, and the primary ineligibility criteria were failure to provide consent; subject data are from the initial recruitment with no reentry to date. ET is rare in minors and no minors were included in this study. Subject gender distribution (9 females, 4 males) was designed to parallel the relative female preponderance of the disease; healthy controls identified from the ethnically diverse population of Long Island, NY were not matched with thrombocytosis cohorts, but were gender-equivalen (i.e. 15 females, 15 males). Methods for platelet isolation, sample processing, and sample quality control using highly-enriched peripheral blood platelets have been previously described [2, 23–25]. The miRNA data were obtained from sample hybridization to the Agilent G4470C human miRNA gene chip that incorporates 866 human and 89 viral miRNAs (miRBase database Version 12.0) and have been deposited into the public GEO database (GEO accession number GSE39046) [25]. The mRNA data were obtained from a custom 432-member oligonucleotide gene chip specifically designed to characterize human platelet-restricted gene expression data [24], and are publicly available (GEO accession number GSE12295).

In the remaining part of this section, we first introduce the Frequentist network analysis methods used in this study. Subsequently we present the integrated analysis combining the results from these different methods.

2.2 Sparse supervised canonical correlation analysis

Introduced by Hotelling in 1936 [26], (the first) canonical correlation between two variable sets looks for the weighted combination of all variables within each variable set such that the correlation of the two combinations is maximized. The weighted combinations are called canonical variables or components. Considering an $n * p$ matrix X and an $n * q$ matrix Y. Without loss of generality, we assume $p < q$. Canonical correlation analysis (CCA) [26] seeks coefficient vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, such that the correlation between the linear combinations $\omega = \boldsymbol{u}'X$ and $\xi = \boldsymbol{v}'Y$ is maximized, i.e.

$$\max_{\boldsymbol{u},\boldsymbol{v}} Corr(\omega, \xi) = \max_{\boldsymbol{u},\boldsymbol{v}} \frac{\boldsymbol{u}'\Sigma_{XY}\boldsymbol{v}}{\sqrt{\boldsymbol{u}'\Sigma_{XX}\boldsymbol{u}}\sqrt{\boldsymbol{v}'\Sigma_{YY}\boldsymbol{v}}}$$

where Σ_{XX}, Σ_{YY}, and Σ_{XY} are the variance for X, Y, and the covariance for X and Y, respectively. It is attained by the canonical variate pairs

$$\omega = \boldsymbol{u}'X = e'\Sigma_{XX}{}^{-\frac{1}{2}}X; \qquad \xi = \boldsymbol{v}'Y = f'\Sigma_{YY}{}^{-\frac{1}{2}}Y$$

with e and f from the singular value decomposition (SVD) of the matrix K given by $K = \Sigma_{XX}{}^{-\frac{1}{2}}\Sigma_{XY}\Sigma_{YY}{}^{-\frac{1}{2}} = eDf'$ [27].

In canonical correlation analysis, all variables are included in the linear combinations, yet for genetic data obtained via microarray studies or other high throughput methods, the number of variables usually surpasses tens of thousands, far exceeding the number of study subjects. Thus the fitted linear combinations may not be easily interpreted and the application of standard algorithms may fail. These problems can be solved by introducing sparse loadings in the canonical components, i.e. the sparse canonical correlation analysis (SCCA) proposed in 2007 [27]. The idea of SCCA is consistent with the belief that only a modest set of genes are truly associated with a given trait of interest.

Based on the foundation of SCCA, Witten and Tibshirani [28] further presented "sparse supervised canonical correlation analysis (sSCCA)", targeting on finding the sparse linear combinations of the two variable sets that are correlated with each other and also associated with the trait of interest. Still considering an $n * p$ matrix X and an $n * q$ matrix Y, and

assuming that the columns of X and Y have been standardized with mean 0 and standard deviation 1. Suppose in addition we have a categorical outcome vector $z \in \mathbb{R}^n$. The estimates of canonical vectors are defined as

$$\max_{u,v} u^T X^T Y v, \text{ subject to}$$

$$\|u\|^2 \leq 1,\ \|v\|^2 \leq 1,\ P_1(u) = \|u\|_1 \leq c_u,\ P_2(v) = \|v\|_1 \leq c_v, \tag{1}$$

$$u_j = 0\, \forall j \notin Q_u,\ v_j = 0\, \forall j \notin Q_v,$$

where P_1 and P_2 are convex penalty functions; c_u and c_v are assumed to be $1 \leq c_u \leq \sqrt{p}$ and $1 \leq c_v \leq \sqrt{q}$; Q_u and Q_v are the sets of variables with highest univariate association with the outcome z in X and Y, respectively; the threshold for variables to be included in Q_u and Q_v can either be fixed or defined as tuning parameters. The vectors u and v are obtained using an iterative algorithm with soft-thresholding. We have performed this sSCCA method on our genetic data set to investigate whether the expression of miRNA would have a significant effect on that of genes and vice versa.

2.3 Sparse partial correlation analysis

Given p continuous random variables $\{X_i, i = 1,2,\ldots, p\}$` from n samples, we can denote the set of measurements/data as

$$X = (X_1, X_2, \ldots \ \ldots, X_p)^T \in \mathbb{R}^{n \times p}$$

Here the rows of the matrix represent the samples and the columns the variables. Within each column (variable), the data are centered to the column mean. For any two random variables X_i and X_j, we denote the set of all other variables as $X_{-(i,j)}$, that is,

$$X_{-(i,j)} = X \setminus \{X_i, X_j\} = \{X_k, 1 \leq k \neq i, j \leq p\}$$

where X_i and $X_j \in \mathcal{R}^n$ are the ith and jth columns of X, and $X_{-(i,j)} \in \mathbb{R}^{n \times (p-2)}$ is the matrix obtained from X by deleting its ith and jth columns. Without loss of generality, we assume that $i < j$.

The Sparse partial Correlation Analysis (SPACE) is a modern method for estimating the partial correlation coefficient also relates to the least square regression problem [29]. This method starts with constructing p linear regression models

$$X_i = X_{-(i)}\beta^{(i)} + \boldsymbol{\varepsilon}_i = \textstyle\sum_{k \neq i} \beta_k^{(i)} X_k + \boldsymbol{\varepsilon}_i, i = 1, 2, \ldots, p \tag{2}$$

where $\boldsymbol{\varepsilon}_i$ are i.i.d. disturbance terms, the least square estimate of the regression coefficient vector is calculated as

$$\hat{\beta}^{(i)} = (\hat{\beta}_1^{(i)}, \hat{\beta}_2^{(i)}, \ldots, \hat{\beta}_{i-1}^{(i)}, \hat{\beta}_{i+1}^{(i)}, \ldots, \hat{\beta}_p^{(i)}) = \arg\min_{\beta \in \mathcal{R}^{p-1}} \|X_i - X_{-(i)}\beta\|^2$$

$$= (X_{-(i)}^T X_{-(i)})^{-1} X_{-(i)}^T X_i, \ \textit{for } i = 1, 2, \ldots, p$$

The sample partial correlation coefficient is then estimated as $\hat{\rho}_{ij} = sign(\hat{\beta}_j^{(i)}) \sqrt{\hat{\beta}_j^{(i)} \hat{\beta}_i^{(j)}}$.

2.4 Sparse Bayesian network analysis

The fundamental structure among a series of random variables is depicted by their joint probability distribution. Probabilistic graphical models are used to describe the conditional

independence or dependence structure implied by the joint distribution with a graph-induced decomposition of the joint density function. A Bayesian Network (BN), a branch of probabilistic graphical model, is a probabilistic graphical model defined over a DAG G with a set of $p = |V|$ nodes $V = \{v_1, \cdots, v_2\}$. In such a graph or network, a node is a random variable, and an edge between two nodes indicates certain stochastic association. The probability model associated with G in a Bayesian network factorizes as $p(X_1, \cdots, X_p) = \prod_{j=1}^{p} p(X_j|Pa(X_j))$, where $p(X_j|Pa(X_j))$ is the conditional probability distribution for X_j given its parents $Pa(X_j)$ with directed edges from each node in $Pa(X_j)$ to X_j in G. For Gaussian random variables, conditional independence of X and Y given Z is equivalent to a zero partial correlation: $\rho_{XY \cdot Z} = 0$. This provides certain insight into the relationship between the Bayesian network and the partial correlation network in that, the partial correlation, by controlling all other variables except the two targeting variables, should in general be more conservative than the Bayesian network.

A recently published paper [30] presented an algorithm entitled A* lasso, for learning a Sparse Bayesian Network structure for continuous variables in a high-dimensional space. Compared to the common two-stage inference methods, A* lasso is a single stage method that recovers the optimal sparse Bayesian network structure by solving a single optimization problem with A* search algorithm that uses lasso in its scoring system. The A*lasso method assumes continuous random variables and uses a linear regression model for the conditional probability distribution of each node $X_j = Pa(X_j)^* \beta_j + \epsilon$, where $\beta_j = \{\beta'_{jk} s for\ X_k \in Pa(X_j)\}$ is the vector of unknown parameters to be estimated from data and ϵ is the noise distributed as $N(0, 1)$. The BN's structure and parameters are obtained by minimizing the negative log likelihood of data with sparsity enforcing L_1 penalty as follows:

$$\min_{\beta_1, \cdots, \beta_p} \sum_{j=1}^{p} \|x_j - x'_{-j}\beta_j\|_2^2 + \lambda \sum_{j=1}^{p} \|\beta_j\|_1 \ s.t. G \in DAG, \tag{3}$$

where X_{-j} represents all columns of $\boldsymbol{X}$ excluding x_j, assuming all other variables are candidate parents of node v_j.

This lasso optimization problem can be solved efficiently with the shooting algorithm [31] if the acyclicity constraint is ignored, which is the most challenge part of the BN inference procedure. A heuristic scheme of A* lasso is proposed to prune search space when learning the Bayesian network structure by exploring a scoring algorithm based on lasso score generated by the shooting algorithm $f(Q_s) = g(Q_s) + h(Q_s)$ [31]. Here Q_s is the set of variables for which the ordering has been determined. And $g(Q_s)$ is the accumulated cost for reaching the Q_s state:

$$g(Q_s) = \sum_{v_j \in Q_s} LassoScore(v_j | \prod_{<v_j}^{Q_s} h(Q_s)) \tag{4}$$

Here $h(Q_s)$ is the estimated cost of reaching the goal stat from the current state

$$g(Q_s) = \sum_{v_j \in V \setminus Q_s} LassoScore(v_j | V \setminus v_j) \tag{5}$$

Furthermore, the Lasso Score is defined as

$$LassoScore(v_j | V \setminus v_j) = \min_{\beta_j} \|x_j - x'_{-j}\beta_j\|_2^2 + \lambda \sum_{j=1}^{p} \|\beta_j\|_1 \tag{6}$$

On top of the heuristic scheme, A* lasso further reduces the search space by limiting the size of intermediate search path via a size-limited priority queue that orders the promising intermediate search paths via the above scoring scheme. The combined strategy gives the A* lasso great advantage in efficiency over the common DP algorithms, which makes it scalable for high-dimension data, such as the miRNA and mRNA interaction problem in our study.

2.5 A novel joint network analysis pipeline

We proposed a novel pipeline for extracting miRNA and mRNA interaction network by combining the sSCCA and the SPACE methods. Our pipeline is designed for small/moderate sample size with large number of miRNAs and mRNAs. In order to extract meaningful insights from small/moderate datasets, the pipeline selects most relevant miRNAs and mRNAs that has the largest canonical correlation via sSCCA and then identifies links between these selected miRNAs and mRNAs through the SPACE method, where the latter would compute the pair-wise partial correlation coefficient conditioned on other features.

There are four steps in the pipeline (Fig 1). First, the differentially expressed (DE) miRNAs and mRNAs are selected via either limma [32] or SAM [33], which are commonly used methods for DE detection. Second, a subset of miRNAs and mRNAs are selected by performing the sSCCA method on the pooled DE miRNAs and mRNAs. In the third step, the pair-wise partial correlations are calculated by performing SPACE on the pooled DE miRNAs and mRNAs. Lastly, only the links that connects the selected miRNAs and mRNAs by sSCCA are kept and added to the sSCCA result.

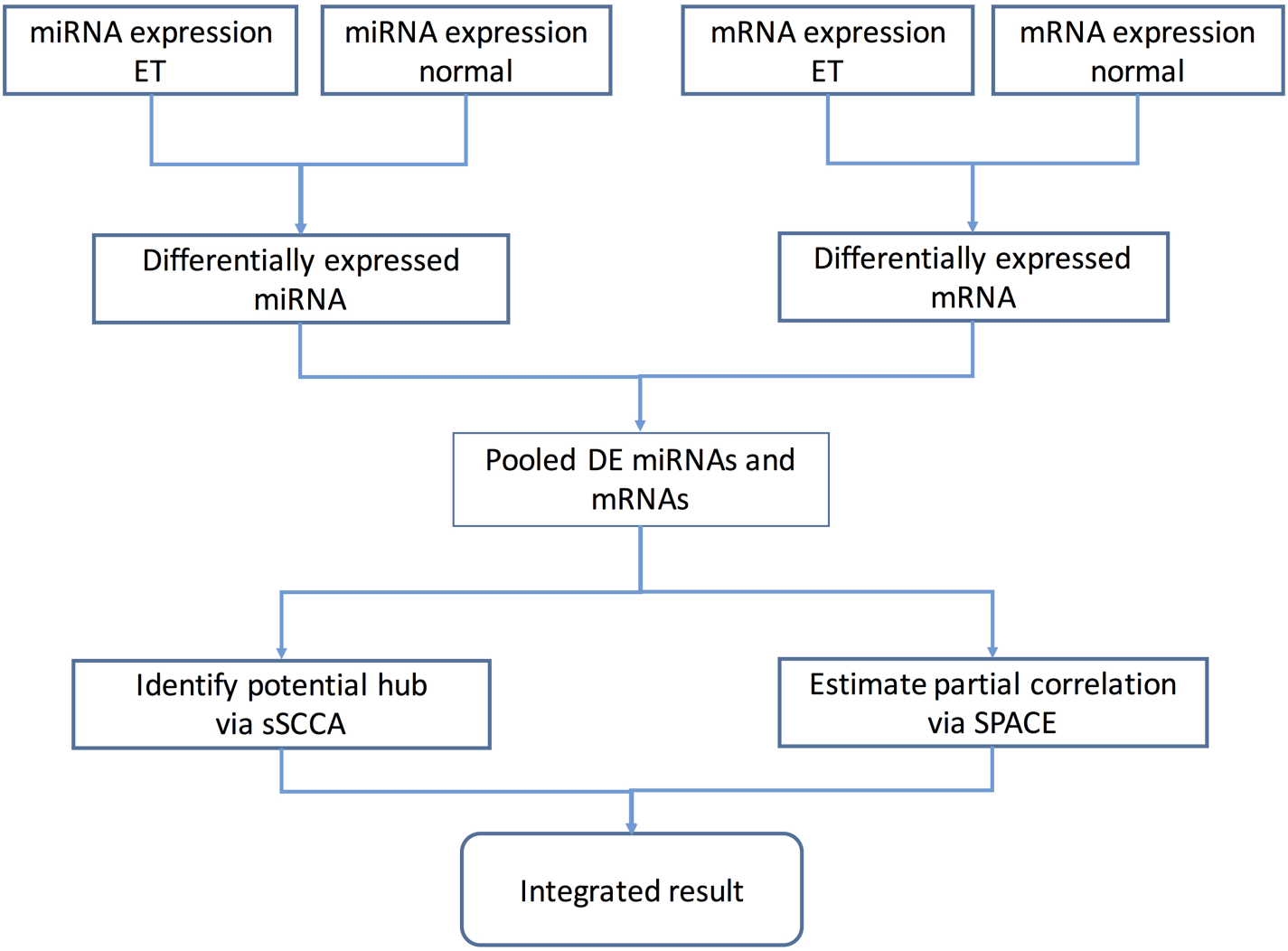

Fig 1. Pipeline of extracting the data-based miRNA and mRNA interaction networks through the joint sparse supervised canonical correlation network analysis (sSCCA) and sparse partial correlation network analysis (SPACE).

https://doi.org/10.1371/journal.pone.0191932.g001

3. Results

3.1 Data structure and processing

Our study integrated platelet mRNA/miRNA expression data from two distinct data sets: (1) mRNA expression data were obtained using a 432-member platelet-specific oligonucleotide custom array as previously described [24], and (2) miRNA expression data were obtained from sample hybridization to the Agilent G4470C human miRNA gene chip that incorporates 866 human and 89 viral miRNAs (miRBase database Version 12.0) [25]. Both mRNA and miRNA expression levels have been collected on 13 patients with essential thrombocytosis (ET) disease and 30 control subjects (S1 Table). Subject recruitment (along with normal healthy controls) was completed by written consent through a study approved by the Stony Brook IRB (Institutional Review Board) Committee on Research Involving Human Subjects (CORIHS), and was restricted to adults (>21 years of age) meeting clinical and laboratory criteria for essential thrombocytosis as previously described (38). Subject gender distribution (9 females, 4 males) paralleled the relative female preponderance of the disease; healthy controls were matched by gender (i.e. 15 females, 15 males).

Among 43 samples, there are 7 (3 ET, 4 NO) samples that have two technical replicates. The values of these samples are reset by the mean value of the sample replicates. The original miRNA data set was filtered in two steps: The first step is to filter out miRNAs with less than 30% non-absent cells in both groups. Next, miRNAs with more than 40% missing values in the sample sets were also dropped out. For the mRNA data, the proportion of missing expression data in the sample set for each mRNA was calculated and those with 50% or more absent data have been excluded. In addition, potential outliers were checked and filtered with a criterion of 3 standard deviations from the mean expression value. In both data sets, quantile normalization was applied to correct the between-array variation [34]. There are 93 out of 432 genes that have missing values in at least one sample. In general, it leads to selection biases if the missing values are simply discarded or the corresponding genes are removed; We decided to impute the missing values using the k-nearest neighbors algorithm [35] implemented in the *impute* R package, which takes into the consideration the correlation structure of the data.

After data filtering and processing, there are totally 327 platelet-specific mRNAs and 396 miRNAs left. To identify highly DE miRNAs and mRNAs, Linear models for microarray data (limma) [36] was applied to the expression data and design matrix. After fitting the linear model, the standard errors are moderated using a simple empirical Bayes model using eBayes function in limma package. Then top DE miRNAs and mRNAs are selected based on the adjusted p-value for the coefficient/contrast of interests. A total of 61 miRNAs and 19 mRNAs were selected at the significant level 0.01 adjusted by the Benjamini-Hochberg (BH) method.

3.2 Individual and combined network analysis results

With the 61 selected miRNAs as one variable set, the 19 mRNAs as the other, and the vector of subject disease status as a binary outcome vector, we applied four network analysis methods (Pearson correlation, sSCCA, SPACE and the Bayesian A* lasso) to the differentially expressed (DE) data sets (miRNA and mRNA).

On the Pearson correlation analysis, the pair-wise Pearson correlation coefficient is calculated using the "psych" R package, and 3164 non-zero coefficients are identified at the significant level 0.01 adjusted by the BH method. It covers all links from the results of SPACE and A* lasso, which indicates that the Pearson correlation may generate much more false positives than the other methods. Therefore, we have decided to focus on the results of the other three methods.

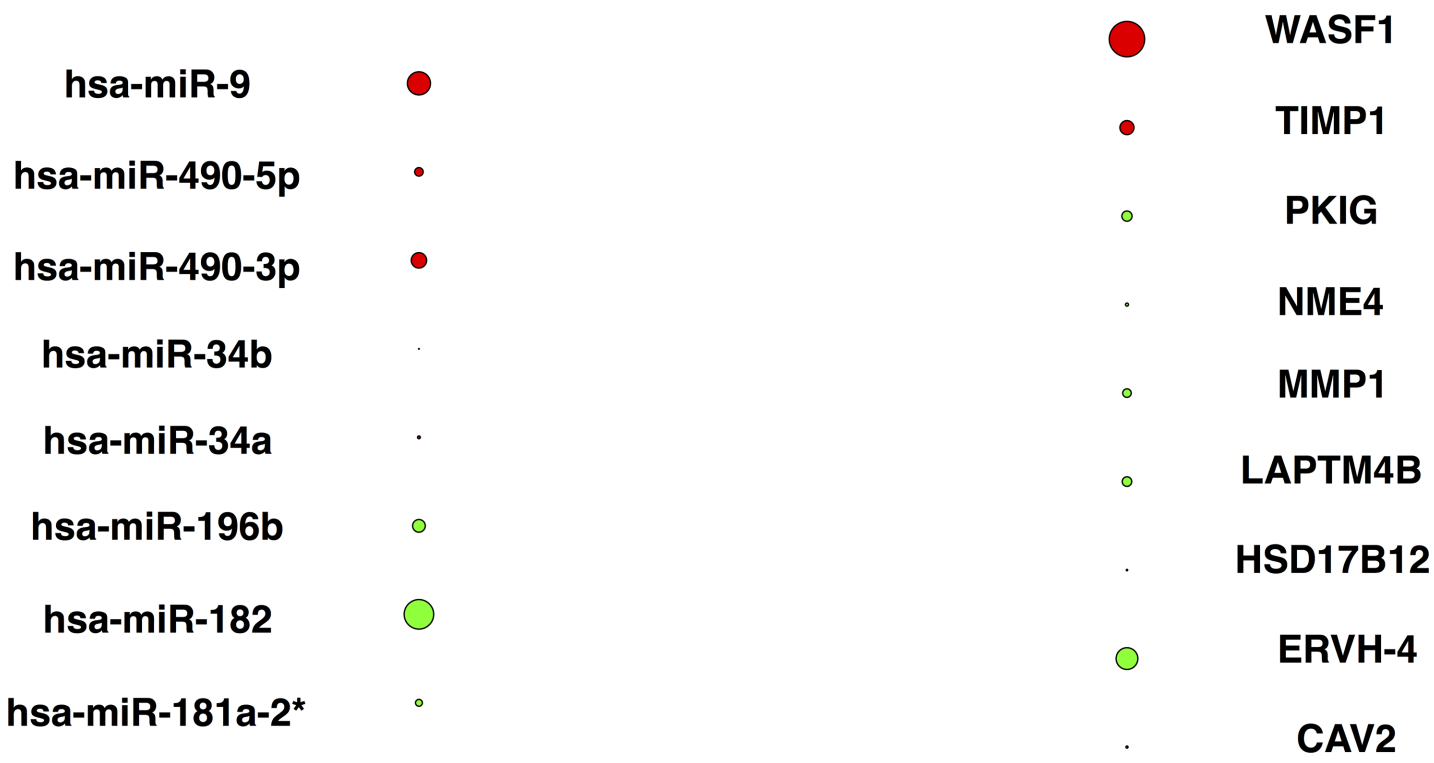

Fig 2. Bipartite plot of the sSCCA result. Red or green node represents positive or negative weight in vector *u* and *v*. The node size represents the absolute value of weight.

https://doi.org/10.1371/journal.pone.0191932.g002

On the sSCCA method, the miRNA and mRNA subsets were selected with the penalty of 0.3 (default value in R package SPACE) on vector *u* and 0.5 on vector *v*. As discussed previously, vector *u* restricts the number of selected miRNA, while vector *v* does the same to the mRNA. In the result, 8 miRNAs stand out with 9 corresponding mRNAs. Fig 2 visualizes the weights in the loadings of the first canonical correlation coefficient of selected miRNAs and mRNAs. The actual values are tabulated in supplementary S2 Table.

SPACE is a penalized method, which has one tuning parameter that controls the L_1 penalty on Lasso regression. The value is set as 0.5765849 as calculated by the following equation.

$$L_1 = \frac{\Phi(1 - \frac{\alpha}{2*p^2})}{\sqrt{n}} \tag{7}$$

Here n is the sample size (43), p is the number of features (80), and α is a constant (1).

Fig 3 illustrates the SPACE interaction network emphasizing the interactions between miRNAs and mRNAs. Those miRNAs and mRNAs that have direct links with each other are labeled. The network is connected and there is no isolated node. Within-group links accounts for most of the edges of the network, suggesting that interaction within group is more common than that between groups. There are only 14 (14/165) direct links between miRNAs and mRNAs.

On the result from the A* lasso algorithm, there are two critical parameters. One is the L_1 penalty on Lasso regression. We chose 0.2 (recommended value) as the L_1 value. The other parameter is the queue size that limits the search depth. In order to obtain a near optimal structure, 3,000 is chosen for this option. Since all mRNAs have direct links with miRNAs, the

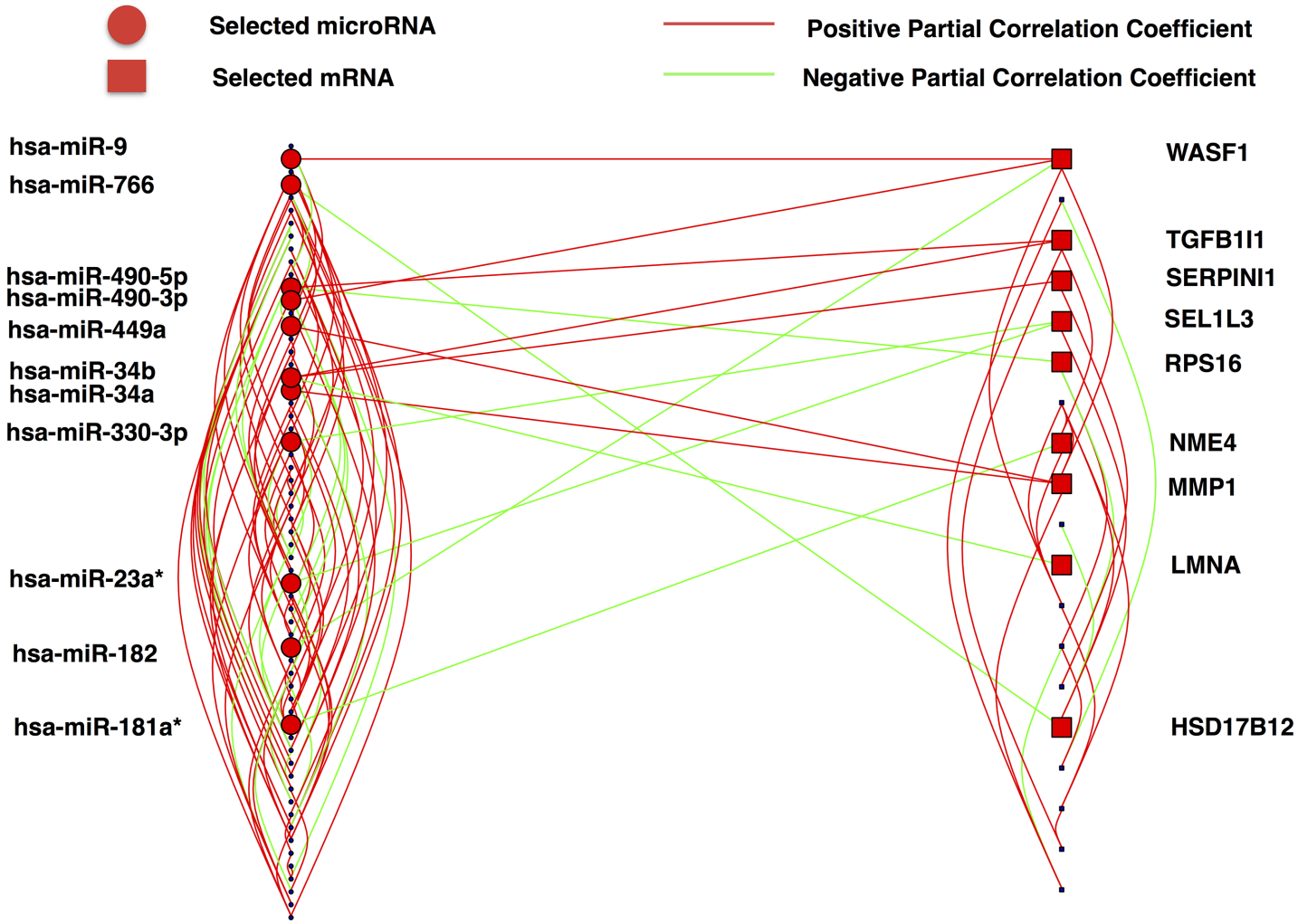

Fig 3. Bipartite plot of the SPACE result. Red circles represent miRNAs that have direct connection with the mRNAs, while the red squares denote the mRNAs that have direct link with the miRNAs. In addition, red and green lines represent positive or negative partial correlations between the pairs.

https://doi.org/10.1371/journal.pone.0191932.g003

names are not listed in the figure (S1 Fig), A* lasso shows the same pattern as the SPACE result, namely, within group interaction is more common than between group interaction.

A* lasso identified 306 links that covers 192 out of 250 links from SPACE result, which is consistent with our expectation that SPACE should be more conservative than A* lasso considering the methodological differences. Since it is very hard to interpret a network with too many links and nodes, we integrate the SPACE and A* lasso result with result from sSCCA by only keeping the selected miRNAs, mRNAs and the corresponding links from SPACE and A* lasso method (Fig 4) respectively. Fig 4 compares the integrated results using SPACE and A* Lasso method with sSCCA. The interaction within the selected mRNAs are strikingly consistent both on links and the value signs except A* Lasso has more links. Two miRNA and mRNA interactions are overlapped. One is the link between has-miR-182 and WASF1. The other is the link between has-mir-34a and MMP1 gene (miRNA).

To render the results more comprehensive, the expression value of those selected miRNAs and mRNAs are tabulated in S3 and S4 Tables. All selected miRNAs and mRNAs are differentially expressed with very small adjusted p-values (all less than 0.0001).

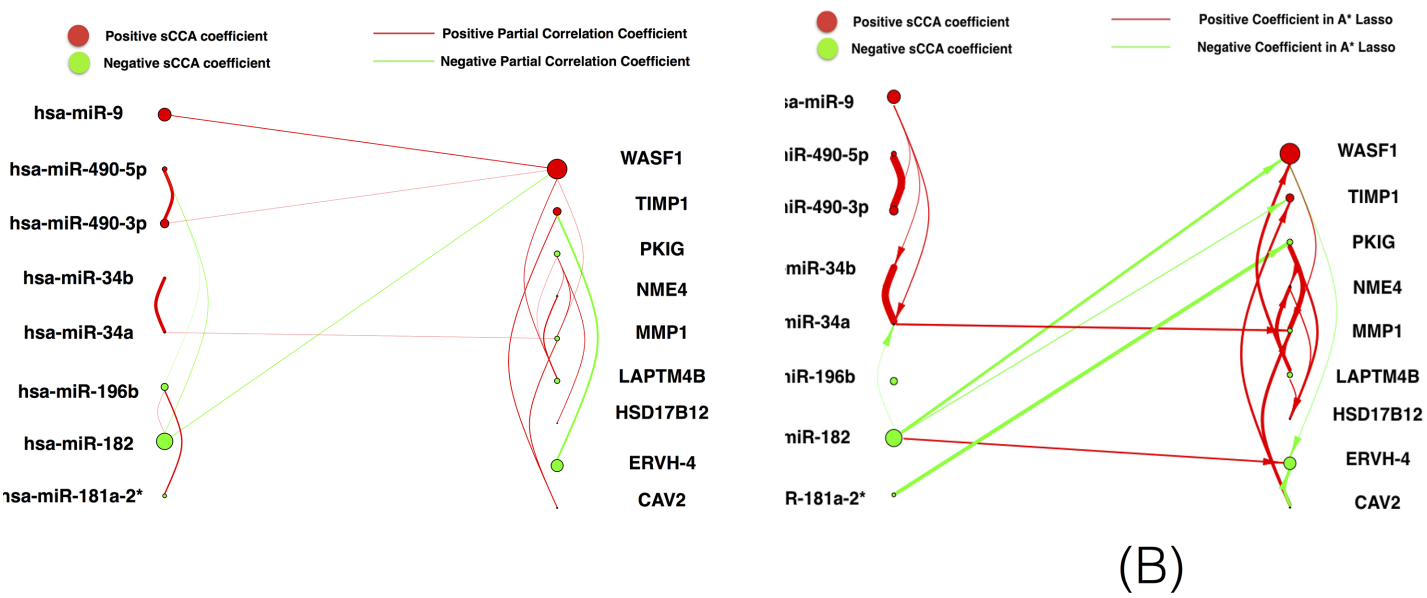

Fig 4. Integrated network analysis results. (A) is the integrated result between sSCCA and SPACE. (B) is the integrated result between sSCCA and A* Lasso method. The arrow is added back on figure (B). The red represents positive values (either weight or correlation coefficient) and green means negative values.

https://doi.org/10.1371/journal.pone.0191932.g004

4. Discussion

In this paper, we proposed a new integrative approach that extracts miRNA and mRNA interaction network by combining the sSCCA and the SPACE methods. Compared to the widely used methods (see S5 Table for detail comparison), such as HOCTAR and GenMiR++, our pipeline is designed for small/moderate sample size with large number of miRNAs and mRNAs and only focuses on the most relevant and sparse networks.

Our joint network analyses using miRNA and mRNA expression data have predicted a close relationship between 8 miRNAs (including miR-9, miR-490-5p, miR-490-3p, miR-182, miR-34a, miR-196b, miR-34b*, miR-181a-2*) and a 9-mRNA set (including CAV2, LAPTM4B, TIMP1, PKIG, WASF1, MMP1, ERVH-4, NME4, HSD17B12), collectively implicating distinct miRNA/mRNA subsets in an integrated network regulating the essential thrombocythemia (*vide infra*). The ET phenotype encompasses two distinct biological pathways, specifically (1) a regulatory network that controls excess platelet production either by effecting megakaryocyte proliferation or proplatelet formation, and (2) a presumably disconnected network that affects platelet functional activity leading to thrombotic or hemorrhagic risk known to accompany ET [37]. Despite these dichotomous functions, molecular defects causally implicated in platelet-associated bleeding or thrombosis remain largely unknown, sharply contrasting with genetic regulation of hematopoietic proliferation/differentiation signals known to accompany terminal megakaryocytopoiesis and platelet production. Application of our miRNA/mRNA network to platelet functional responses provides a logical framework for subsequent delineation of clinical thrombohemorrhagic outcomes in defined ET cohorts.

Notably, the network(s) identified by SPACE and A* lasso methods have significant overlap, serving to validate our conclusions by applying distinct approaches to yield comparable results. Two overlapped links (*miR-182-WASF1* and *miR-34a-MMP1*) are worthy targets for biological validation since all four mRNAs/miRNAs have been previously implicated in the ET phenotype. Indeed, *miR-34a* and *miR-182* identified by sparse SCCA have been previously described as demonstrating aberrant expression in polycythemia vera (PV) granulocytes [8];

furthermore, both *miR-34a* and *miR-182* are among the most significant differentially-expressed miRNA members among a cohort of thrombocytosis subjects [25]. The *miR 34* family members (*miR 34a* and miR *34b/c*) contain p53 binding sites, and *miR 34a* is widely studied as a tumor suppressor gene and as a potential therapeutic target in human cancer [38]. No prior evidence has demonstrated that *miR34a* regulates *MMP1* (matrix metalloproteinase 1) as demonstrated by our data [38]. Indeed, both *MMP1* and its inhibitor *TIMP1* (tissue inhibitor of metalloproteinases 1) are members of a well-characterized class of proteinases involved in tumor invasiveness and cancer metastases [39]. Furthermore, *TIMP1* has been predicted as a putative *miR-34a* target using the target prediction tools TargetScan [40], designed to identify regulatory targets using conserved complementary [12]. Members of the matrix metalloproteinase family have been implicated in the migration and invasion of leukemia cell (*MMP-2*) [41], and previously shown to mediate megakaryocyte transendothelial migration and proplatelet formation (*MMP-9*) [42]. *MMP1* has also been studied in the context of inflammation in several studies [43–46], thereby providing an additional link to the known function(s) of platelets in adaptive immunity [37].

In addition to *MMP1/TIMP1*, various other transcripts within the 9-member mRNA list have critical roles in platelet biology and function. Indeed, both *CAV2* (caveolin 2) and *WASF1* (WAS protein family, member 1) have fundamentally important functions in maintaining cytoskeletal function and viability of membrane/lipid rafts, key regulators of the platelet activation response. Moreover, the WAS protein family has been shown to be related to nucleosome and chromatin assembly, performing an important role in gene transcription that may regulate megakaryocytopoiesis and/or proplatelet formation [47]. A recent study in class prediction models of ET included a member from this family (*WASF3*) as one of the biomarkers segregating ET from reactive thrombocytosis and healthy controls [24], thereby extending the role of the WAS family of proteins in key regulatory functions of megakaryocytopoiesis and/or platelet activation. Finally, *HSD17B12* (hydroxysteroid (17-β) dehydrogenase 12) which catalyzes the penultimate step in testosterone synthesis, has been previously identified as a functionally-active dehydrogenase in ET platelets, serving as a putative link to gender-regulated differences in platelet function [23].

We also used the Ingenuity Pathway Analysis (IPA) software to further characterize the confirmed associations between 8-miRNAs and 9-mRNAs. IPA predicts that *miR-9* and *miR-196b* have interaction with *NME4* (NME/NM23 nucleoside diphosphate kinase 4) which links to several fundamentally important pathways regulating nucleotide synthesis expected to be active during enhanced megakaryoctyopoiesis (i.e. salvage pathways of pyrimidine ribonucleotides; pyrimidine ribonucleotides *de novo* biosynthesis; pyrimidine ribonucleotides interconversion; pyrimidine deoxyribonucleotides *de novo* biosynthesis 1). It also links *miR-34a/miR-34b** with *WASF1* and relates these two links to multiple pathways critical for platelet function (including actin cytoskeleton signaling; actin nucleation by ARP-WASP complex; epithelial adherens junction signaling; Rac signaling; regulation of actin-based motility by Rho; RhoA Signaling; RhoGDI Signaling; and signaling by Rho family GTPases). These pathways are relevant not only to megakaryocyte development and proplatelet formation, but also have fundamental relevance to platelet activation and signaling linked to cardio/cerebrovascular thrombotic diseases.

Supporting information

S1 Fig. Network generated by A* lasso.
(TIFF)

S1 Table. Data structure.
(DOCX)

S2 Table. Loadings of miRNA and mRNAs in the first canonical component of sSCCA result.
(DOCX)

S3 Table. Quantile normalized expression of selected miRNAs.
(DOCX)

S4 Table. Quantile normalized expression of selected mRNAs.
(DOCX)

S5 Table. Comparison with HOCTAR and GenMiR++.
(DOCX)

Author Contributions

Conceptualization: Song Wu, Wadie F. Bahou, Wei Zhu.

Data curation: Dimitri Gnatenko, Wadie F. Bahou.

Formal analysis: Lu Zhao.

Funding acquisition: Dimitri Gnatenko, Wadie F. Bahou, Wei Zhu.

Investigation: Wadie F. Bahou.

Methodology: Lu Zhao, Erya Huang, Wei Zhu.

Project administration: Song Wu, Wadie F. Bahou, Wei Zhu.

Resources: Song Wu.

Software: Lu Zhao.

Supervision: Wadie F. Bahou, Wei Zhu.

Validation: Lu Zhao, Erya Huang.

Writing – original draft: Lu Zhao, Erya Huang.

Writing – review & editing: Song Wu, Dimitri Gnatenko, Wadie F. Bahou, Wei Zhu.

References

1. Lundberg P, Karow A, Nienhold R, Looser R, Hao-Shen H, Nissen I, et al. Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. Blood. 2014; 123(14):2220–8. https://doi.org/10.1182/blood-2013-11-537167 PMID: 24478400.
2. Gnatenko DV, Dunn JJ, McCorkle SR, Weissmann D, Perrotta PL, Bahou WF. Transcript profiling of human platelets using microarray and serial analysis of gene expression. Blood. 2003; 101(6):2285–93. https://doi.org/10.1182/blood-2002-09-2797 PMID: 12433680.
3. Edelstein LC, Bray PF. MicroRNAs in platelet production and activation. Blood. 2011; 117(20):5289–96. https://doi.org/10.1182/blood-2011-01-292011 PMID: 21364189
4. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet. 2008; 9(2):102–14. https://doi.org/10.1038/nrg2290 PMID: 18197166.
5. Garzon R, Fabbri M, Cimmino A, Calin GA, Croce CM. MicroRNA expression and function in cancer. Trends Mol Med. 2006; 12(12):580–7. https://doi.org/10.1016/j.molmed.2006.10.006 PMID: 17071139.
6. Merkerova M, Belickova M, Bruchova H. Differential expression of microRNAs in hematopoietic cell lineages. Eur J Haematol. 2008; 81(4):304–10. https://doi.org/10.1111/j.1600-0609.2008.01111.x PMID: 18573170.

7. Girardot M, Pecquet C, Boukour S, Knoops L, Ferrant A, Vainchenker W, et al. miR-28 is a thrombopoietin receptor targeting microRNA detected in a fraction of myeloproliferative neoplasm patient platelets. Blood. 2010; 116(3):437–45. https://doi.org/10.1182/blood-2008-06-165985 PMID: 20445018.
8. Bruchova H, Merkerova M, Prchal JT. Aberrant expression of microRNA in polycythemia vera. Haematologica. 2008; 93(7):1009–16. https://doi.org/10.3324/haematol.12706 PMID: 18508790.
9. Boilard E, Nigrovic PA, Larabee K, Watts GF, Coblyn JS, Weinblatt ME, et al. Platelets amplify inflammation in arthritis via collagen-dependent microparticle production. Science. 2010; 327(5965):580–3. https://doi.org/10.1126/science.1181928 PMID: 20110505; PubMed Central PMCID: PMCPMC2927861.
10. Schwertz H, Koster S, Kahr WH, Michetti N, Kraemer BF, Weitz DA, et al. Anucleate platelets generate progeny. Blood. 2010; 115(18):3801–9. https://doi.org/10.1182/blood-2009-08-239558 PMID: 20086251; PubMed Central PMCID: PMCPMC2865870.
11. Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ. miRBase: tools for microRNA genomics. Nucleic Acids Res. 2008; 36(Database issue):D154–8. https://doi.org/10.1093/nar/gkm952 PMID: 17991681; PubMed Central PMCID: PMCPMC2238936.
12. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005; 120(1):15–20. https://doi.org/10.1016/j.cell.2004.12.035 PMID: 15652477.
13. Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Mol Cell. 2007; 27(1):91–105. https://doi.org/10.1016/j.molcel.2007.06.017 PMID: 17612493; PubMed Central PMCID: PMCPMC3800283.
14. Huang JC, Babak T, Corson TW, Chua G, Khan S, Gallie BL, et al. Using expression profiling data to identify human microRNA targets. Nat Methods. 2007; 4(12):1045–9. https://doi.org/10.1038/nmeth1130 PMID: 18026111.
15. Huang JC, Morris QD, Frey BJ. Bayesian inference of MicroRNA targets from sequence and expression data. J Comput Biol. 2007; 14(5):550–63. https://doi.org/10.1089/cmb.2007.R002 PMID: 17683260.
16. Gennarino VA, Sardiello M, Avellino R, Meola N, Maselli V, Anand S, et al. MicroRNA target prediction by expression analysis of host genes. Genome Res. 2009; 19(3):481–90. https://doi.org/10.1101/gr.084129.108 PMID: 19088304; PubMed Central PMCID: PMCPMC2661810.
17. Jayaswal V, Lutherborrow M, Ma DD, Yang YH. Identification of microRNA-mRNA modules using microarray data. BMC Genomics. 2011; 12:138. https://doi.org/10.1186/1471-2164-12-138 PMID: 21375780; PubMed Central PMCID: PMCPMC3065435.
18. Li X, Gill R, Cooper NG, Yoo JK, Datta S. Modeling microRNA-mRNA interactions using PLS regression in human colon cancer. BMC Med Genomics. 2011; 4:44. https://doi.org/10.1186/1755-8794-4-44 PMID: 21595958; PubMed Central PMCID: PMCPMC3123543.
19. Le TD, Liu L, Tsykin A, Goodall GJ, Liu B, Sun BY, et al. Inferring microRNA-mRNA causal regulatory relationships from expression data. Bioinformatics. 2013; 29(6):765–71. https://doi.org/10.1093/bioinformatics/btt048 PMID: 23365408.
20. Maathuis MH, Kalisch M, Bühlmann P. Estimating high-dimensional intervention effects from observational data. The Annals of Statistics. 2009; 37(6A):3133–64. https://doi.org/10.1214/09-aos685
21. Boross G, Orosz K, Farkas IJ. Human microRNAs co-silence in well-separated groups and have different predicted essentialities. Bioinformatics. 2009; 25(8):1063–9. https://doi.org/10.1093/bioinformatics/btp018 PMID: 19131366.
22. Masud Karim SM, Liu L, Le TD, Li J. Identification of miRNA-mRNA regulatory modules by exploring collective group relationships. BMC Genomics. 2016; 17 Suppl 1:7. https://doi.org/10.1186/s12864-015-2300-z PMID: 26817421; PubMed Central PMCID: PMCPMC4895272.
23. Gnatenko DV, Cupit LD, Huang EC, Dhundale A, Perrotta PL, Bahou WF. Platelets express steroidogenic 17beta-hydroxysteroid dehydrogenases. Distinct profiles predict the essential thrombocythemic phenotype. Thromb Haemost. 2005; 94(2):412–21. https://doi.org/10.1160/TH05-01-0037 PMID: 16113833.
24. Gnatenko DV, Zhu W, Xu X, Samuel ET, Monaghan M, Zarrabi MH, et al. Class prediction models of thrombocytosis using genetic biomarkers. Blood. 2010; 115(1):7–14. https://doi.org/10.1182/blood-2009-05-224477 PMID: 19773543; PubMed Central PMCID: PMCPMC2803693.
25. Xu X, Gnatenko DV, Ju J, Hitchcock IS, Martin DW, Zhu W, et al. Systematic analysis of microRNA fingerprints in thrombocythemic platelets using integrated platforms. Blood. 2012; 120(17):3575–85. https://doi.org/10.1182/blood-2012-02-411264 PMID: 22869791; PubMed Central PMCID: PMCPMC3482865.
26. Hotelling H. Relations between two sets of variates. Biometrika. 1936; 28(3–4):321–77.
27. Parkhomenko E, Tritchler D, Beyene J, editors. Genome-wide sparse canonical correlation of gene expression with genotypes. BMC proceedings; 2007: BioMed Central Ltd.

28. Witten DM, Tibshirani RJ. Extensions of sparse canonical correlation analysis with applications to genomic data. Statistical applications in genetics and molecular biology. 2009; 8(1):1–27.
29. Peng J, Wang P, Zhou N, Zhu J. Partial Correlation Estimation by Joint Sparse Regression Models. J Am Stat Assoc. 2009; 104(486):735–46. https://doi.org/10.1198/jasa.2009.0126 PMID: 19881892; PubMed Central PMCID: PMCPMC2770199.
30. Jing Xiang SK. A* Lasso for Learning a Sparse Bayesian Network Structure for Continuous Variables. Advances in neural information processing systems2013. p. 2418–26.
31. Fu WJ. Penalized Regressions: The Bridge versus the Lasso. Journal of Computational and Graphical Statistics. 1998; 7(3):397–416. https://doi.org/10.1080/10618600.1998.10474784
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712
33. Chu G, Li J, Narasimhan B, Tibshirani R, Tusher V. Significance Analysis of Microarrays Users Guide and Technical Document. 2001.
34. Pradervand S, Weber J, Thomas J, Bueno M, Wirapati P, Lefort K, et al. Impact of normalization on miRNA microarray expression profiling. RNA. 2009; 15(3):493–501. https://doi.org/10.1261/rna.1295509 PMID: 19176604; PubMed Central PMCID: PMCPMC2657010.
35. Olga Troyanskaya MC, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein and Russ B. Altman. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001; 17:6.
36. Smyth GK. limma: Linear Models for Microarray Data. 2005:397–420. https://doi.org/10.1007/0-387-29362-0_23
37. Bahou WF. Genetic dissection of platelet function in health and disease using systems biology. Hematol Oncol Clin North Am. 2013; 27(3):443–63. https://doi.org/10.1016/j.hoc.2013.03.002 PMID: 23714307; PubMed Central PMCID: PMCPMC3767180.
38. Li XJ, Ren ZJ, Tang JH. MicroRNA-34a: a potential therapeutic target in human cancer. Cell Death Dis. 2014; 5:e1327. https://doi.org/10.1038/cddis.2014.270 PMID: 25032850; PubMed Central PMCID: PMCPMC4123066.
39. Cathcart J, Pulkoski-Gross A, Cao J. Targeting Matrix Metalloproteinases in Cancer: Bringing New Life to Old Ideas. Genes Dis. 2015; 2(‘1):26–34. https://doi.org/10.1016/j.gendis.2014.12.002 PMID: 26097889; PubMed Central PMCID: PMCPMC4474140.
40. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015;4. https://doi.org/10.7554/eLife.05005 PMID: 26267216; PubMed Central PMCID: PMCPMC4532895.
41. He Y, Cao L., Yang M., Zhao M., Yu Y., and Xu W. [Role of WAVE1 in K562 leukemia cells invasion and its mechanism]. Zhonghua xue ye xue za zhi. 2009:5.
42. William J. Lane SD, Koichi Hattori, Beate Heissig, Margaret Choy, Sina Y. Rabbany, Jeanette Wood, Malcolm A. S. Moore, and Shahin Rafii. Stromal-derived factor 1–induced megakaryocyte migration and platelet production is dependent on matrix metalloproteinases. Blood. 2000; 96:8.
43. Andonovska B, Dimova C, Panov S. Matrix metalloproteinases (MMP-1, -8, -13) in chronic periapical lesions. Vojnosanit Pregl. 2008; 65(12):882–6. PMID: 19160981.
44. Brassart B, Fuchs P, Huet E, Alix AJ, Wallach J, Tamburro AM, et al. Conformational dependence of collagenase (matrix metalloproteinase-1) up-regulation by elastin peptides in cultured fibroblasts. J Biol Chem. 2001; 276(7):5222–7. https://doi.org/10.1074/jbc.M003642200 PMID: 11084020.
45. Herouy Y, Mellios P, Bandemir E, Dichmann S, Nockowski P, Schopf E, et al. Inflammation in stasis dermatitis upregulates MMP-1, MMP-2 and MMP-13 expression. J Dermatol Sci. 2001; 25(3):198–205. PMID: 11240267.
46. Zhang BB, Cai WM, Weng HL, Hu ZR, Lu J, Zheng M, et al. Diagnostic value of platelet derived growth factor-BB, transforming growth factor-beta1, matrix metalloproteinase-1, and tissue inhibitor of matrix metalloproteinase-1 in serum and peripheral blood mononuclear cells for hepatic fibrosis. World J Gastroenterol. 2003; 9(11):2490–6. https://doi.org/10.3748/wjg.v9.i11.2490 PMID: 14606082; PubMed Central PMCID: PMCPMC4656526.
47. Schulze H, Shivdasani RA. Molecular mechanisms of megakaryocyte differentiation. Semin Thromb Hemost. 2004; 30(4):389–98. https://doi.org/10.1055/s-2004-833474 PMID: 15354260.