Peer-reviewed version available at Molecules 2018, 23, 183; doi:10.3390/molecules23020183

Article

The Integrative Method Based on Module-Network for Identifying Driver Genes in Cancer Subtypes

Xinguo Lu [1,*], Xing Li [1], Xin Qian [1], Qiumai Miao [1] and Shaoliang Peng [1,2,*]

[1] College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China; hnluxinguo@hnu.edu.cn(X.L); xingleo@hnu.edu.cn
[2] School of Computer Science, National University of Defense Technology, Changsha 410073, China
* Correspondence: hnluxinguo@hnu.edu.cn; pengshaoliang@nudt.edu.cn; Tel.: +1-397-585-0642

Abstract: With advances in next-generation sequencing(NGS) technologies, large number of multiple types of high-throughput genomics data are available. A great challenge in exploring cancer mechanism is to identify the driver genes from the mutation genes by analyzing and integrating multi-types genomics data. Breast cancer is known as a heterogeneous disease. The identification of subtype-specific driver genes is critical to guide the diagnosis, assessment of prognosis and treatment of breast cancer. We developed an integrated frame based on gene expression profilings and copy number variation(CNV) data to identify breast cancer subtype-specific driver genes. In this frame, we employed statistical machine-learning method to select gene subsets and utilized an module-network analysis method to identify potential candidate driver genes. The final subtype-specific driver genes were acquired by paired-wise comparison in subtypes. To validate specificity of the driver genes, the gene expression data of these genes were applied to classify the patient samples with 10-fold cross validation and the enrichment analysis were also conducted on the identified driver genes. The experimental results show that the proposed integrative method can identify the potential driver genes and the classifier with these genes acquired better performance than with genes identified by other methods.

Keywords: integrative analysis; module network; cancer subtypes; breast cancer; copy number variation; gene expression

1. Introduction

Breast cancer is one of the most common malignant tumors in women, the incidence rate is 7%-10% of all kinds malignant tumors which is usually associated with genetic[1]. Previous studies have classified breast cancer as five subtypes, including luminal A (LumA), luminal B (LumB), HER2-enriched (HER2), basal-like (Basal), and normal-like (Normal) types. Studies have shown that each cancer subtype has its own gene imprint and tumor markers, and genetic variation will increase the risk of cancer. Not all of the aberrations, however, have an important impact on tumor progression. To understand the mechanism of cancer, identifying genomic aberrations driving cancer process, which are termed drivers, has become the focus of research. Similarly, the researchers also considered that gene expression profiling plays an important role for understanding the pathogenesis of disease. For example, the over-expression of an oncogene or under-expression of a tumor suppressor gene also have a certain impact on cancer process. So that the driver genes may be the mutant genes of over/under-expressed genes. It is reasonable to think that an over/under-expressed driver gene has a footprint in a genome in the form of an aberration that can be used as a biomarker[2,3]. Meanwhile, with the availability of a large number of databases and software, it is beneficial to manage and analyze biological data. These useful biological techniques and tools enable researchers to explain how normal cellular activities are altered in different disease states, especially in cancers [4,5].

With the development of high-throughput sequencing technology, a large number of diseased-based histological data based on different technology have been provided in life science

research. These data are publicly available in databases, such as The Cancer Genome Atlas(TCGA) and International Cancer Genome Consortium(ICGC) which have generated several high throughput genomic data types for hundreds of sample on tens of cancer types. Recently, many methods have been developed which integrated genomic data of multi-types to reveal combinatorial patterns and discover new biological mechanisms[6]. Computational approach is an alternative method which can solve various biological problems, including analyzing complex biological network and identifying novel key genes[7]. For example, Zhang et al. developed an unbiased adaptive clustering approach to integrate and analyze the genomic data of multi-types of ovarian cancer, including genome-wide gene expression, DNA methylation, microRNA expression, and copy number alteration profiles. And they developed an algorithm to uncover molecular signatures that distinguish cancer subtypes[8]. It is worth noting that comprehensively analyzing multiple genomic data will improve the understanding of the role of biomarkers in breast cancer pathogenesis and procession[9]. In previous studies, many researchers have studied to discover novel candidate key genes by integrating gene expression and CNV data. Li et al. identify the breast cancer subtype-specific drivers by integrating and analyzing the copy number aberrations data and miRNA-mRNA dual expression profiling data[10]. However, many methods are based on linear methods, such as regression analysis and correlation analysis, which is not suitable for heterogeneous data that have extremely high within-group variations. Due to breast cancer data with is high heterogeneity, there are some defects in the linear integration method. Therefore, in this paper, we utilized the module network analysis[11] to integrate the datasets for identifying the driver genes. The main idea of module network is a form of Bayesian network that is similarly behaving variables classified into module and that can learn the same parents and parameters for each module, instead of each variable. A module is defined as a set of random variables(in this application, a set of genes) that share a statistical model. For example, a set of genes are co-expressed or co-regulated. So we applied the module-network analysis method to integrate the genomic data. Our work is to construct a novel computational framework to discovery the candidate driver genes of breast cancer subtypes by integrating the CNV data and gene expression data. Our methods were aimed at identifying a short of genes as candidate drivers of each breast cancer subtypes to advancing the drivers discovery process. Our results have shown that the proposed method was able to identify highly mutated gene as candidate driver genes of each subtype and the selected driver genes can classify cancer subtypes as well.

2. Results

Our goal is to identify the subtype-specific driver genes by integrating two genomic data types. The hypothesis is that driver genes are among over/under-expressed genes that also have abberations. From TCGA, we downloaded the clinical records and five breast cancer subtypes of high-throughout data of 825 patients together with their survival time from initial diagnoses to death, or to the last follow-up if they were still alive at the time of the TCGA study. All four tumor types of data(messenger RNA[mRNA] and miRNA expression, promoter DNA methylation, and CNA) were available for 485 of the 825 samples. We used the 485 samples that had both gene expression and copy number data for analysis. The gene expression data include expression level of 17268 genes. We obtained the CNV values of 20871 genes for each sample.

We applied the integrative method described in the methods section to identify the driver genes. The experiment includes three parts: 1) identifying the novel subtype-specific driver genes based on the proposed integrative method. 2) comparing the classification performance. The classifiers are constructed on these genes selected by the proposed method and other two methods separately, including information gain and Chi-squared. 3) analyzing biological functions of the obtained driver genes, such as Gene ontology(GO) functional enrichment, KEGG pathway enrichment analysis etc.

For gene expression and CNV data, the difference between each subtype and other three subtypes was analyzed. For gene expression data, we ranked the genes based on their q-values and selected genes with q-values <0.1. And for CNV data, we firstly selected genes with their q-values<0.1. We

calculated the frequency of amplification and deletion for each gene in each of the two subtypes samples and selected genes for which the difference of frequencies between two subtypes is more than 20%. Additional, we used a threshold of log2 copy number variation ratio of 0.3/-0.3 to call amplified/deleted genes. And the gene subsets are converged and the relevant subset of data is selected for each subtype. We selected the genes as candidate modulators with mutation frequency $fre > 0.6$. We compiled a list of 965 candidate modulators for the HER2 tumors, 336 candidate modulators for luminal-A tumors, 1213 candidate modulators for luminal-B tumors, 815 candidate modulators for basal-like tumors.

Through above data preprocessing, module-network analysis is used to select candidate driver genes. The whole learning algorithm is run 100 times. we filtered the set of candidate modulators and left only genes that appeared in at least one regulation program in at least 40% of the runs. These modulators are considered as candidate driver genes. Thus, the procedure resulted in 148 modules with 9 candidate drivers for HER2 samples, 550 modules with 28 candidate drivers for luminal-A samples, 258 modules with 12 candidate drivers for luminal-B samples, 201 modules with 14 candidate drivers for basal-like samples. Several modules shared the same modulators in each subtypes of samples. Interestingly, the outputs of module network analysis for HER2 and other subtypes shared three genes. Finally, the candidate driver genes in this subtype,but not those in other subtypes, are the subtype-specific driver genes, including 8 Her2 drivers, 11 Basal drivers, 28 LumA drivers and 10 LumB drivers.

2.1. Identified novel subtype-specific driver genes in LumA.

In this section, we applied the proposed frame to identify 28 candidate driver genes of LumA that are frequent mutatation and novel and significant. A heatmap of the 28 significantly differentially expressed genes of LumA is shown in Fig.2. As can be seen, expression profile can be clearly clustered into four subtypes by using the selected driver genes. In addition, we found that some samples were mixed between the LumA and the LumB in the process of clustering. That may be the fact that LumA and LumB belong to the luminal, and luminal subtype has the lowest overall mutation rate.

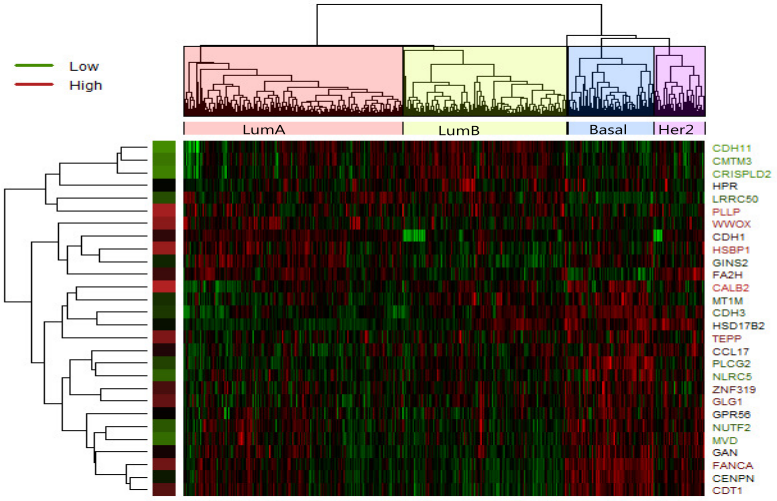
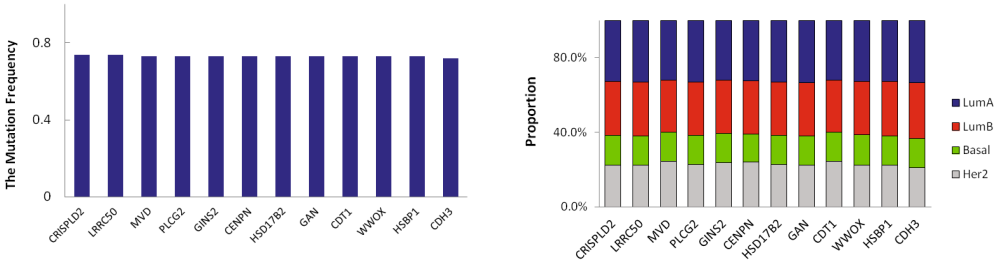


Figure 1. Heatmap of expression values of 28 most significant differently expressed gene in LumA-subtype. Clustering method on expression values was used to generate the heatmap. There are clear clusters of genes for the four tumor subtypes.

A novel significant driver of the subtype is a frequently mutated gene in specific subtype that has not been classified as a driver gene. Here we defined novel significant driver gene that satisfy the following requirements: (1) subtype-specific driver genes could be recurrently mutated in multiple

subtypes, but the mutation frequency of the driver gene is different from other three subtypes, and the difference is more than 0.2; (2)not previously classified as a driver by CGC database[12]. In LumA, we found 23 high-mutated novel significant driver genes. The top 12 mutation frequency of driver genes in LumA_subtype is shown in Fig.3a. From the Figure, we can see the mutation frequency of driver genes identified in LumA is more than 0.7. Moreover, the proportion of mutation samples in each breast cancer subtype is shown in Fig.3b. we can see that the mutation samples of LumA is higher than that of the other three subtypes.



(a) (b)

Figure 2. (**a**)The top 12 mutation frequency of driver genes in LumA. (**b**)The mutation proportion of the top 12 genes in each breast cancer subtype samples .

Of the 23 driver genes, CDH3, GLG1, CCL17 and PLCG2 are the most promising. These genes are frequently mutated and are involved in several cancer functions and pathways(see Additional file1). CDH3 belongs to the family of classic cadherins that are engaged in various cellular activities including motility, invasion, and signaling of tumor cells, in addition to cell adhesion[13]. CDH3 has the highest-mutated frequency in breast cancer subtypes, and it is mutated in 72% of LumA patient samples. GLG1 is a key ligand-receptor in the early response of cells[14]. CCL17 is a central regulator of Treg homeostasis, and CCL17 might be a target for vascular therapy[15]. CCL17 is mutated in 70% of LumA patient samples. A CCL17 gene is a candidate as one of the genetic factors in some allergic diseases[16]. PLCG2 encodes phospholipase C_{Y2}(PLC_{Y2}), an enzyme with a critical regulatory role in various immune and inflammatory pathways, and plays a key role in the regulation of immune response[17]. PLAID-associated deletions of PLCG2 cause diminished receptor-mediated activity at physiologic temperatures in B cells and natural killer cells with enhanced spontaneous signaling in mast cells and B cells at subphysiologic temperatures[18].

2.2. Identified novel subtype-specific driver genes in Basal/LumB/Her2.

We applied the same method to select driver genes in others subtypes. Similarly, we can identify the subtype-specific driver genes. In Basal, there are several novel significant driver genes we found. However, we found two genes that are strong novel drivers: FDPS and ATP1A2. FDPS is a key enzyme in the isoprenoid pathway responsible for cholesterol biosynthesis, post-translational protein modifications and synthesis of steroid hormones, whose expression is regulated by phorbol esters and polyunsaturated fatty acids[19]. FDPS is necessary for osteoclast survival and activity and is considered as a major molecular target of aminobisphosphonates[20]. The ATP1A2 gene encodes the α2 subunit of Na+-, K+- ATPase, a plasma membrane enzyme that counter transports Na+ and K+ across cell membranes[21]. ATP1A2 gene mutation result in degeneration of the amygdala and pyriform cortex[22].

In LumB, two potential novel driver we found is FASLG and RGS2. Alteration of FASLG pathway regulating cell death may lead to cancer development[23]. Studies have revealed that increased FASLG expression facilitate development and progression of tumors, including gastric cancer. These results suggest that variants of the FASLG gene is likely to be associated with the initiation and development

of gastric cancer[24]. The mutation frequency of FASLG is about 80% in subtype LumB. RGS2 is a member of a family of proteins that negatively modulate G-protein coupled receptor transmission. Variations in the RGS2 gene were found to be associated in humans with anxious and depressive phenotypes[25]. RGS2 is mutated in 78% Basal patient samples.

In Her2, there are fewer candidate driver genes than Luma, however, we found two genes that strong potential to become novel drivers: ZFPM2 and EGLN1. ZFPM2 protein is an important cofactor for GATA family of transcription factors. In adult tissues, the ZFPM2 protein of 1151 amino acids is expressed predominantly in brain, heart, and testis. ZFPM2 may act as repressor or activator, depending on specific promoter and cell type[26]. EGLN1 is a key oxygen sensor gene that negatively regulates the activity of hypoxia-in-ducible factor (HIF-1A). Owing to its important function as an oxygen sensor, EGLN1 is relevant to the human hypoxic response, both at high altitude in hypoxic conditions or in cellular hypoxia[27]. EGLN1 is a important gene functioning at the upstream of the HIF pathway and showed consistent selective signals across multiple studies[28].

2.3. Validation of classification between any two subtypes

To validate whether our subtype-specific driver genes obtained from integration analysis are also applicable to distinguish subtypes, we classified the samples between any two subtypes using SVM[29]. SVM is a linear maximum-margin model for classification[30]. Here, we applied the expression profiles of 216 LumA, 92 Basal, 122LumB and 55Her2, in which we selected half of the samples of two subtypes to train the classifier, and the rest of the samples were used to test the classification performance. In addition, to verify the effectiveness of the proposed method, we performed the classification performance comparison with the other two key gene selection methods, Information gain and Chi-squared. These two methods identified the key genes simply based on gene expression data but not considered the mutation data, while our method is based on the gene expression data and mutation data. And we also compared the classification performance with another integrative method, called Lemon-tree.

Also, to evaluate the performance of the model, we used accuracy measure. Accuracy can be computed as follows:

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{1}$$

Where TP, TN, FP and FN are true positive, true negative, false positive and false negative respectively.

In this study, we used the 10-fold cross-validation to evaluate the performance of the model. As shown in table1, while using the selected 28 driver genes of LumA and 11 driver genes of Basal together, the totally selected 39 genes for prediction makes the accuracy up to 99.4%. Moreover, while using the selected 11 driver genes of Basal and 8 driver genes of Her2 together, the prediction accuracy is 95.5%. This indicates that the subtype-specific driver genes we selected can divide samples into two classes well. However, the accuracy of two groups of LumA-LumB and LumB-Her2 is 82.4% and 81.4% respectively. Compared to the information gain, Chi-squared and Lemon-tree method, we can see the accuracy of four methods in LumA-Basal is up to 99.4% already. And, in LumA-Her2 and Basal-Her2, the prediction accuracy of module analysis ia higher than the other three methods. However, in LumA-LumB and LumB-Her2, the selected driver genes can not classify the subtypes very well. This may be that the selected driver genes have a similar regulation in gene expression.

Table 1. The accuracy of 10-fold cross validation between any two subtypes.

Subtypes	Module-network	Information Gain	Chi-squared	Lemon-tree
LumA-LumB	82.44%	84.04%	84.04%	76.60%
LumA-Basal	99.37%	99.37%	99.37%	99.37%
LumA-Her2	93.62%	90.78%	90.78%	90.78%
LumB-Basal	90.35%	95.61%	95.61%	80.70%
LumB-Her2	81.44%	89.69%	89.69%	74.23%
Basal-Her2	95.52%	91.04%	95.52%	91.04%

2.4. Functional analysis of driver genes based on each subtype

we did Gene Ontology(GO) biological process (BP) and Kyoto Encyclopedia of Genes and Genomes(KEGG) pathways enriched among the subtype driver genes with R package clusterProfiler (http://www.bioconductor.org/packages/release/bioc/html/clusterProfiler.html)[31]. Only the enriched GO terms with p-value less than 0.05 are shown and the enriched KEGG pathways with p-value less than 0.05 are shown.

For LumA, the important driver genes are mainly enriched in pathways in Cell adhesion molecules (CAMs), Terpenoid backbone biosynthesis, Thyroid cancer, African trypanosomiasis and so on after KEGG pathway enrichment. With respect to the biological process, nuclear DNA replication, steroid metabolic process, B cell receptor signaling pathway, organic hydroxy compound biosynthetic process and etc. are enriched after the GO functional enrichment.

In Basal, pathways in protein digestion and absorption, Lysosome, Natural killer cell mediated cytotoxicity, Apoptosis etc. are enriched in KEGG pathways. In terms of biological process, cholesterol biosynthetic process, secondary alcohol biosynthetic process, multicellular organism catabolic process, multicellular organismal macromolecule metabolic process, steroid biosynthetic process etc. are significantly enriched in GO functional enrichment.

In LumB, the pathways after KEGG enrichment are apoptosis, African trypanosomiasis, NOD-like receptor signaling pathway, Allograft rejection, Graft-versus-host disease etc.. In terms of biological process in GO functional enrichment, driver genes are enriched in response to positive regulation of peptidase activity, activation of innate immune response, positive regulation of innate immune response, cellular chloride ion homeostasis etc..

In Her2, the main pathways are the HIF-1 signaling pathway, MicroRNAs in cancer, Primary immunodeficiency, Bladder cancer and Pathways in cancer etc. after KEGG enrichment analysis. In terms of biological process, blood vessel morphogenesis, regulation of T cell proliferation, regulation of microtubule-based process, T cell proliferation and regulation of mononuclear cell proliferation and so on are enriched in GO functional enrichment.

3. Discussion

In this work, we introduced a module-based framework by integrating transcriptome and genomic data to identify significant driver genes in breast cancer subtypes. By virtue of the consideration the differential expression of genes, a subset of gene whose aberration/expression profile were significantly different between two subtypes may be selected. Also, we constructed a module network by integrating multi-genomic data to identify the specific driver genes.

We applied this method to the challenging problem of identifying driver genes of breast cancer subtypes. We selected breast cancer because breast cancer data are very heterogeneous, and the conventional method for analyzing heterogeneous data perform poorly. The final result demonstrate that the power of integrative analysis can generate a biological meaningful short list of genes as subtype-specific driver genes.

Furthermore, there are also some limitations of our method. Firstly, in computational and statistical analysis, the method is computationally expensive due to iteration in training the model and searching for the best model. In additional, because the method is based on the statistical

machine-learning, it will work better if there are more samples. If there are a few samples in each condition, this method will not perform well. In the word, we developed a novel integrated method based on statistical machine-learning analysis to discover the biomarkers of breast cancer subtypes, and we showed that the method can generate a short list of biologically meaningful genes that can promote the process of biomarkers discovery.

4. Materials and methods

4.1. Breast cancer patients materitals

We used processed and normalized breast cancer genomic data as given by TCGA. We downloaded gene expression data from the Agilent 244 K Custom Gene Expression platform, CNV data from the Affymatix Genome-Wide Human SNP 6.0 platform. In this data set, genomic data of 825 patients were obtained. We examined the clinical data from these patients to identify the subtype patients with gene expression data and CNV data. We selected 216 Luminal A, 122 Luminal B, 98 Basal-like and 58 HER2-enriched among the 825 patients, for which their gene expression and CNV data were available as well.

4.2. Differential expression analysis for data with intraclass heterogeneity

The within group variations are extremely high for the heterogeneous data of breast cancer. Conventional methods to select gene subsets perform poorly when applied to these data with high within class heterogeneity. In this manuscript, the approach of EMD is applied to acquire gene subset[32]. EMD is a measure of distance between two distributions that reflects the minimum cost of transforming one distribution into the other. Where test statistics generated by standard differential expression approaches reflect the likelihood that the difference of mean expression between two groups is non-zero or reflect the significance of the association between abundance of short reads of the two groups, the EMD test statistic reflects the overall difference between two normalized distributions.

Given two signatures P and Q which are represented as:$P = \{(p_1, w_{p1}), \ldots, (pm, w_{pm})\}$, where p_i is the center of the i_th histogram cell and w_{pi} is the weight of the cell; and $Q = \{(q_1, w_{q1}), \ldots, (qm, w_{qm})\}$, where p_j is the center of the j_th histogram cell and w_{pj} is the weight of the cell. Given P, Q, and d_{ij}(the Euclidean distance between p_i and q_j), f_{ij} (the flow between p_i and q_j).

First, the EMD scores are calculated as follows:

$$EMD(P,Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}d_{ij}}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}}. \tag{2}$$

Then, the FDRs are obtained for range of significance thresholds. Given $M = [m_1, \ldots, m_N]$, a vector of median of permuted EMDs, and $EMD = [emd_1, \ldots, emd_N]$, a vector of observed EMDs, the mathematical representation of FDR for gene$j$ and significance threshold i, t_i is as follows:

$$FDR_{ji} = \begin{cases} \frac{\sum_{k=1}^{N} I(m_k > t_i)}{\sum_{k=1}^{N} I(emd_k > t_i)} & if\ emd_j \geq t_i \\ 1 & otherwise \end{cases} \tag{3}$$

Where I is the indicator function, N is the number of genes in the dataset, and $t_i's$ are in descending order from T to zero with step $\triangle : \{T,\ T-\triangle,\ T-2\triangle,\ \ldots\ ,\ \triangle,\ 0\}$. Then, the q-value for gene j is calculated as:

$$q - value_j = min\left(FDR_j\right). \tag{4}$$

Peer-reviewed version available at Molecules 2018, 23, 183; doi:10.3390/molecules23020183

4.3. The selection of candidate modulators

The genes, within or neighboring to these regions of DNA that are recurrently amplified/deleted in tumors, are considered as candidate modulators. Thus we wanted to identified the genes that are mutated frequently and consider genes with high mutation frequency as candidate modulators across tumors. The mutation frequency is calculated as follows:

$$fre = \frac{\#Mutgenes}{N}. \tag{5}$$

Where $\#Mutgenes$ represents the number of mutation genes in across sample, N is the number of researched samples.

4.4. Initial modules construction based on Normal Gamma score

Innumerable functional studies suggest that driver mutation are expected to alter gene expression of their cognate proteins, their interacting partners, or genes that share the same biochemical pathway. This will lead to a correlated pattern of gene expression in a network of genes associated with a driver mutation. The initial modules construction step establishes an initial pairing between candidate modulators and gene expression modules by associated each target gene with the single modulator gene that fits it best.

First, for each candidate modulator gene, we use the gene expression values of the amplified/deleted samples to guide the choice of threshold, and consider the gene expression of the amplified/deleted samples to represent appropriate high/low expression levels. We use k-means clustering, using k=2 and the amplified/deleted samples as the two initial clusters to fit two normal distributions. The boundary between two clusters is the selected expression threshold level for this driver gene.

Then, the expression of each target gene is split into two sets: those in the tumor samples in which the driver's expression is below the threshold, and those in the tumor samples in which the driver's expression is above the threshold. The Normal Gamma scoring function is used to compute the quality of this split, thus measuring a target gene's fit with a candidate modulator. Given a vector $Leaf$ which represents the gene expression values contained in the leaf; α and λ are the parameters and N is the size of $Leaf$. The Normal Gamma score is described below:

$$\begin{aligned} Score = &- N * \ln\left(\sqrt{2\pi}\right) + \frac{\ln\left(\frac{\lambda}{\lambda+N}\right)}{2} + \ln\left(\Gamma(\alpha^+)\right) - \\ &ln\left(\Gamma(\alpha)\right) + \alpha * \ln\left(\beta\right) - \alpha^+ * \ln\left(\beta^+\right) \end{aligned} \tag{6}$$

where,

$$\beta = Max(1, \frac{\lambda * (\alpha - 2)}{\lambda + 1}) \tag{7a}$$

$$\beta^+ = \beta + \frac{Var(Leaf) * N}{2} + N * \lambda * \frac{\overline{Leaf^2}}{2 * (N + \lambda)} \tag{7b}$$

$$\alpha^+ = \alpha + \frac{N}{2} \tag{7c}$$

After the score is computed for all pair-wise combinations of candidate modulators and target genes, each gene is assigned to the single highest scoring candidate modulator.

4.5. Module-network learning

The module-network learning step uses the modules generated by the initial modules step as a starting point and uses an iterative approach to improve the score of the modules and their regulation programs. The part iteratively process contains two tasks: (i)learning a regulation for each module. (ii)re-assigning each gene to the module whose program best predicts its behavior. These two steps are

iterated until that fewer than 10% of the target genes have been re-assigned to a different module. The process learning is shown in Figure 3.

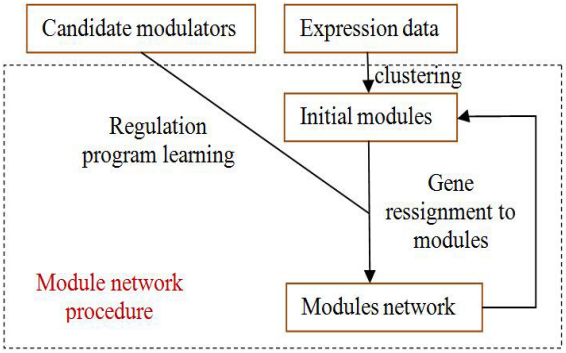


Figure 3. The process of module-network learning. It is an iterative procedure that determines both the partition of genes to modules and the regulation program for each module.

Thus, the Expectation Maximization(EM) algorithm[33] is applied to search for the model with the highest score. A important property of the EM algorithm is that each iteration is guaranteed to improve the likelihood of the model until convergence to a local maximum of the score is achieved.

4.6. The identification of candidate driver genes

We proposed an integrative frame based on module-network to identify candidate driver genes. The framework process is shown in Figure 4. First the EMD difference analysis and frequency analysis are used to select subset of data. Then the initial modules are constructed by k-means clustering and Normal Gamma scoring function. And the module-network learning is used to obtain the final modules and candidate modulators. the module-network learning step is run N times. We calculated the frequency of appearance of each candidate modulators as follows:

$$fre\_app = \frac{\#times\ of\ appearance}{N}. \tag{8}$$

where $\#times\ of\ appearance$ is the appearance times of each modulators in N runs. we filtered the set of candidate modulators and left only genes that appeared in at least one regulation program in at least 40% of the runs. The final run of the module-network learning algorithm is run using the filtered set of candidate modulators and these modulators are considered as candidate driver genes. The whole method is shown as algorithm 1.

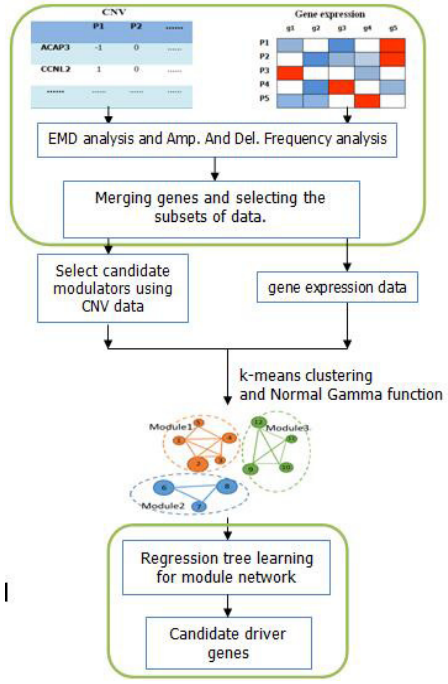


Figure 4. The frame diagram of the integrative method based on module-network. The first part indicates the pre-processing steps based on the differential expression analysis. The middle part is the construction of the initial modules. The bottom part represents the process of module network learning.

Algorithm 1 Integrative model based on module-network for cancer subtypes

Input: CNV data and gene expression data of two subtypes

Output: A short list of gene sets

The 1th step: Difference analysis EMDSort (P, Q, f_{ij}, d_{ij})

(a)compute the $EMD(P, Q)$ using f_{ij} and d_{ij} according to the Formula.2. f_{ij} is the flow and the d_{ij} is the Euclidean distance.

(b)compute the FDR_{ji} according to the emd-values.

(c)compute the $q - value$ according to the FDR_{ji}.

The 2th step: Initial modules construction

(a)fit two normal contributions by k-means clustering and select the threshold T for each modulator.

(b)split the expression of the target gene into two sets (A, B) according to the threshold T.

(c)Given a leaf vector $leaf$, the parameters α and λ, the size of $Leaf$ N.

(d)compute the $Score(target\_gene, modulator)$ of the split using the Formula.6.

(e)assign the target gene into the single highest scoring candidate modulator.

The 3th step: Module network learning

repeat

(a)search for a regulation program for each module.

(b)reassign each gene to the module whose program best predicts its behavior.

(b)compute the proportion of re-assigned genes pro.

until $(pro < 0.1)$

The 4th step: The identification of candidate driver genes.

Acknowledgments: The authors are grateful to Dana Pe'er Lab who provided the module networks learning analysis method with CONEXIC Toolkit and Guangchuang Yu's team who provided the functional enrichment

analysis method with clusterProfilter package. This work was supported by the National Natural Science Foundation grant of China and China Scholarship Council grant.

Conflicts of Interest: The authors declare that they do not have any conflicts of interest.

References

1. Mok TS, Wu YL, Thongprasert S, et al. Gefitinib or carboplatinpaclitaxel in pulmonary adenocarcinoma. *The Journal of Evidence-Based Medicine* **2009**, 361(10): 947-957.
2. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, Causton HC, Pochanard P, Mozes E, Garraway LA, Peer D. An integrated approach to uncover drivers of cancer. *Cell* **2010**, 143: 1005-17.
3. Lahti L, Schafer M, Klein H-U, Bicciato S, Dugas M. Cancer gene prioritization by integrative analysis of mRNA expression and DNA copy number data: a comparative review. *Brief Bioinform* **2013**, 14: 27-35.
4. Hanahan D, Weinberg R A. The hallmarks of cancer. *Cell* **2000**, 100(1): 57-70
5. Hanahan D, Weinberg R A. Hallmarks of cancer: the next generation. *Cell* **2011**, 144(5): 646-674.
6. Chen J, Zhang S. Integrative cancer genomics: models, algorithms and analysis[J]. *Frontiers of Computer Science* **2016**, 1-15.
7. Yue Z, Li HT, Yang Y, et al. Identification of breast cancer candidate genes using gene co-expression and protein-protein interaction information[J]. *Oncotarget* **2016**, 7(24): 36092.
8. Zhang W, Liu Y, Sun N, et al. Integrating genomic, epigenomic, and transcriptomic features reveals modular signatures underlying poor prognosis in ovarian cancer.[J]. *Cell Reports* **2013**, 4(3): 542-553.
9. Enerly E, Steinfeld I, Kleivi K, et al. miRNA-mRNA integrated analysis reveals roles for miRNAs in primary breast tumors.[J]. *Plos One* **2011**, 6(2): e16915.
10. Li D, Xia H, Li Z, et al. Identification of Novel Breast Cancer Subtype-Specific Biomarkers by Integrating Genomics Analysis of DNA Copy Number Aberrations and miRNA-mRNA Dual Expression Profiling[J]. *Biomed Research International* **2015**, 2015(2): 1-17.
11. Segal E, Al E. Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.[J]. *Nature Genetics* **2003**, 34(2): 166-76.
12. Futreal P A, Coin L, Marshall M, et al. A census of human cancer genes[J]. *Nature Reviews Cancer* **2004**, 4(3):177-83.
13. Taniuchi K, Nakagawa H, Hosokawa M, et al. Overexpressed P-cadherin/CDH3 promotes motility of pancreatic cancer cells by interacting with p120ctn and activating rho-family GTPases[J]. *Cancer research* **2005**, 65(8): 3092-3099.
14. Li C, Li R W, Elsasser T H, et al. Lipopolysaccharide-induced early response genes in bovine peripheral blood mononuclear cells implicate GLG1/E-selectin as a key ligand-receptor interaction[J]. *Functional & integrative genomics* **2009**, 9(3): 335-349.
15. Weber C, Meiler S, Doring Y,et al. CCL17-expressing dendritic cells drive atherosclerosis by restraining regulatory T cell homeostasis in mice.J Clin Invest[J]. *Journal of Clinical Investigation* **2011**, 121(7): 2898-2910.
16. Saeki H, Tamaki K. Thymus and activation regulated chemokine (TARC)/CCL17 and skin diseases.[J]. *Journal of Dermatological Science* **2006**, 43(2): 75.
17. Zhou Q, Lee G S, Brady J, et al. A hypermorphic missense mutation in PLCG2, encoding phospholipase Cγ2, causes a dominantly inherited autoinflammatory disease with immunodeficiency[J]. *The American Journal of Human Genetics* **2012**, 91(4): 713-720.
18. Aderibigbe O M, Priel D L, Lee C C R, et al. Distinct cutaneous manifestations and cold-induced leukocyte activation associated with PLCG2 mutations[J]. *JAMA dermatology* **2015**, 151(6): 627-634.
19. Romanelli M G, Lorenzi P, Sangalli A, et al. Characterization and functional analysis of cis-acting elements of the human farnesyl diphosphate synthetase (FDPS) gene 5' flanking region[J]. *Genomics* **2009**, 93(3): 227-234.
20. Olmos J M, Zarrabeitia M T, Hernandez J L, et al. Common allelic variants of the farnesyl diphosphate synthase gene influence the response of osteoporotic women to bisphosphonates[J]. *The pharmacogenomics journal* **2012**, 12(3): 227-232.
21. Fernandez D M, Hand C K, Sweeney B J, et al. A novel ATP1A2 gene mutation in an Irish familial hemiplegic migraine kindred.[J]. *Headache the Journal of Head & Face Pain* bf 2008, 48(1):101-108.
22. Harriott A M, Dueker N, Cheng Y C, et al. Polymorphisms in migraine-associated gene, atp1a2, and ischemic stroke risk in a biracial population: the genetics of early onset stroke study[J]. *SpringerPlus* **2013**, 2(1): 1-8.

23. Lei D, Sturgis E M, Wang L E, et al. FAS and FASLG genetic variants and risk for second primary malignancy in patients with squamous cell carcinoma of the head and neck[J]. *Cancer Epidemiology and Prevention Biomarkers* **2010**, 19(6): 1484-1491.
24. Wang M, Wu D, Tan M, et al. FAS and FAS ligand polymorphisms in the promoter regions and risk of gastric cancer in Southern China[J]. *Biochemical genetics* **2009**, 47(7-8): 559-568.
25. Lifschytz T, Broner E C, Zozulinsky P, et al. Relationship between Rgs2 gene expression level and anxiety and depression-like behaviour in a mutant mouse model: serotonergic involvement[J]. *International Journal of Neuropsychopharmacology* **2012**, 15(9): 1307-1318.
26. Greenbaum L, Smith R C, Lorberboym M, et al. Association of the ZFPM2 gene with antipsychotic-induced parkinsonism in schizophrenia patients[J]. *Psychopharmacology* **2012**, 220(3): 519-528.
27. Aggarwal S, Negi S, Jha P, et al. EGLN1 involvement in high-altitude adaptation revealed through genetic analysis of extreme constitution types defined in Ayurveda[J]. *Proceedings of the National Academy of Sciences* **2010**, 107(44): 18961-18966.
28. Xiang K, Peng Y, Yang Z, et al. Identification of a Tibetan-specific mutation in the hypoxic gene EGLN1 and its contribution to high-altitude adaptation[J]. *Molecular biology and evolution* **2013**, 30(8): 1889-1898.
29. Lu X, Peng X, Deng Y, et al. A Novel Feature Selection Method Based on Correlation-Based Feature Selection in Cancer Recognition[J]. *Journal of Computational & Theoretical Nanoscience* **2014**, 11(2): 427-433.
30. Firoozbakht F, Rezaeian I, Ngom A, et al. A new compact set of biomarkers for distinguishing among ten breast cancer subtypes[C]. *IEEE International Conference on Bioinformatics and Biomedicine. IEEE* **2015**, 1579-1585.
31. Yu G, Wang L G, Han Y, et al. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters[J]. *Omics A Journal of Integrative Biology* **2012**, 16(5):284.
32. Nabavi S, Schmolze D, Maitituoheti M, et al. EMDomics: a robust and powerful method for the identification of genes differentially expressed between heterogeneous classes[J]. *Bioinformatics* **2016**, 32(4): 533-541.
33. Dempster A P, Laird N M, Rubin D B. Maximum Likelihood from Incomplete Data via the EM Algorithm[J]. *Journal of the Royal Statistical Society* **1977**, 39(1): 1-38.