Identifying Mutated Core Modules in Glioblastoma by Integrative Network Analysis

Junhua Zhang, Shihua Zhang, Yong Wang, Junfei Zhao and Xiang-Sun Zhang

National Center for Mathematics and Interdisciplinary Sciences,
Academy of Mathematics and Systems Science,
Chinese Academy of Sciences, Beijing 100190, China

Correspondece to: Junhua Zhang (zjh@amt.ac.cn) or Shihua Zhang (zsh@amss.ac.cn)

***Abstract*—Glioblastoma multiforme (GBM) is the most common and aggressive type of brain tumor in humans. Distinguishing "driver" mutations from passively selected "passengers" is a central challenge in computational cancer biology. Because of mutational heterogeneity, analyses that extend beyond single genes are often restricted to examine known pathways and functional modules for enrichment of somatic mutations. In this paper we present a network-based method to identify mutated core modules for tumors without any prior information other than the data of somatic mutations and gene expressions from tumor patients. Firstly, two networks with weighted vertices and weighted edges are constructed by using the mutations and expressions, respectively. Then these two networks are combined to get an integrative network, for which an optimization model is used to identify the most coherent subnetworks. With the significance and exclusivity tests we get the core modules for tumors. By applying our method to The Cancer Genome Atlas (TCGA) GBM data, we obtained three core modules, which contain not only oncogenes and tumor suppressors that have been previously implicated in GBM pathogenesis (e.g., *EGFR, TP53, PTEN, NF1* and *RB1*), but also some genes which have not or rarely been reported earlier in the context of glioblastoma multiforme (e.g., *DST, PRAME* and *SYNE1*). Thus, in addition to present generally applicable methodology, our findings provide several GBM candidate genes for further studies.**

***Index Terms*—Cancer; core module; somatic mutation; gene expression.**

I. Introduction

With the rapid advances in high-throughput genome sequencing, large-scale cancer genomics projects, such as the Cancer Genome Atlas (TCGA) and the International Genome Consortium (ICGC), are producing a large volume of data about genomic, epigenomic, and gene expression aberrations in tumor samples [1], [2], [3]. This unprecedented volume of information provides a basis for systems level understanding of cancer formation and progression. A key challenge is to distinguish the functional "driver" mutations, which contribute to tumorigenesis, from the random "passenger" mutations, which have accumulated in somatic cells but do not contribute to tumor development [4].

A standard approach to predict driver mutations is to identify recurrent mutations in a large cohort of cancer patients. For example, by comparing alteration rates in individual genes or regions of copy number alteration against an empirically derived background alteration rate [5]. But further studies revealed that cancer genomes exhibit extensive mutational heterogeneity with no two genomes containing exactly the same complement of somatic mutations [1], [6], [7]. That is, the driver mutations may be different for diverse patients – even those from the same tumor type.

On the other hand, driver mutations typically target genes in cellular signaling and regulatory pathways which lead to the acquisition of tumorigenic properties, such as cell proliferation, angiogenesis, or metastasis [8], [9]. Thus, examining mutations in the context of such biological pathways is an alternative approach. These studies include analyzing known pathways for enrichment of somatic mutations [1], [6], [7], identifying known pathways that are significantly mutated across many patients [10], [11], and *de novo* discovery of mutated driver pathways in cancer [12], [13].

In addition, algorithms that extend pathway analysis to genome-scale gene interaction networks have recently been introduced. For example, Cerami et al. [14] and Ciriello et al. [15] identified oncogenic network modules from a constructed network by using gene mutation information and the human reference network (including protein-protein interactions (PPI) and signal transduction pathways). The defect of such approaches exist in that human PPI network is incomplete, and many protein-protein interactions are unknown. Furthermore, while some pathways are well-characterized and cataloged in various databases [16], [17], knowledge of pathways remains incomplete, too. Recently, Miller et al. developed a method for detecting functional modules in tumors based solely on patterns of recurrent genomic aberration [18].

Some investigations indicate that cancer alterations tend to cluster within closely knitted network modules or communities, and that altered modules are closely linked to specific biological pathways. Furthermore, genes in the same pathway are usually activated together and thus have similar gene expression profiles. Some research already shows that genes with similar expression profiles are more likely to coordinately achieve a particular function [19], [20]. So it is very necessary to integrate gene expression information to identify oncogenic network modules and candidate pathways. Masica and Karchin proposed a method to examine the correlation among somatic

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

mutation and gene expression to identify important genes in human cancer [21]. In the procedure, each time a particular mutated gene is examined if there are genes differentially expressed with respect to its mutation status, then the mutated genes are selected out for further investigation which have 2 or more drastic mutation-correlated over- and under-expression genes. But whether the selected genes form oncogenic modules or pathways and whether their expressions are correlated are not discussed.

Glioblastoma multiforme is the most common brain tumor in adults, with median survival just over a year. Although some genes have been reported to be drivers for this cancer, the etiology and the molecular pathogenetic mechanisms are not entirely clear yet. In this paper we present a network-based method to integrate somatic mutations and gene expressions to identify mutated core modules for cancers without any other prior information such as PPI networks and known pathways. Our approach is based on the hypothesis that cooperative dysregulation of gene sequence and expression may contribute to cancer formation and progression. Core modules and candidate oncogenic processes are investigated with the consideration that cellular networks contain functional modules, and that tumors target specific modules critical to their growth. Key elements in our approach include combined analysis of somatic mutations and gene expressions, that is, analysis of an integrated molecular network constructed from mutations and expressions, respectively; identification of coherent subnetworks (modules) using an optimization model; and statistical assessment of identified core modules. We applied the method to the TCGA GBM data and obtained three core modules, which contain not only some well-known oncogenes and tumor suppressors that have been previously implicated in GBM pathogenesis (e.g., *EGFR, TP53, PTEN, NF1* and *RB1*), but also some others which have not or rarely been reported earlier in the context of glioblastoma multiforme (e.g., *DST, PRAME* and *SYNE1*). Thus, in addition to present generally applicable methodology, our findings provide several GBM candidate genes for further studies.

II. Methods

A. Data

We downloaded the GBM data from TCGA website (https://tcga-data.nci.nih.gov/tcga/). After processing the gene expression data using the method in [22], we obtained unified gene expression profiles in 202 samples. We collected the nucleotide sequence aberrations data in three different platforms and got the somatic mutations in 135 samples.

Finally we obtain the mutation matrix A and the gene expression matrix B. The rows and columns of these matrices correspond to samples and genes, respectively. At this stage, B is a standerization matrix and A is binary: 1 if any mutation occurs for a particular gene in a particular sample, otherwise the element is 0.

B. Construction of an integrative network

With the GBM data, we have constructed an integrative network based on which an optimization model can be used to detect oncogenic modules and pathways. Mainly the constructive procedure contains three steps.

1) The network based on gene expression: In this step we construct a network based on gene expression, called *Expression Network* (denoted by **EN**). **EN** is weighted both for its edges and vertices, in which each vertex denotes a gene, and the edge is weighted by the correlation between expressions of the two vertices (genes). As for the weight in each vertex, it reflects the influencing extents of the gene's mutation to other genes' expression.

We notice that the genes in A and B may be different, so the common genes are found out at first. Let (G_1, S_1) and (G_2, S_2) are the sets of genes and samples contained in these two matrices, respectively. Set $G_0 = G_1 \cap G_2$ and $S = S_1 \cap S_2$. For any gene $i \in G_0$, the samples in S are classified into two groups according to the binary mutation vector of i from the mutation matrix A, and the corresponding numbers of samples are denoted by $n_i^{(1)}$ and $n_i^{(2)}$, respectively. Then p-values for all genes in G_2 are calculated using the program *mattest* in MATLAB toolbox to evaluate the extents of differential expression of these genes related to i's mutation status. In this procedure the prerequisite that the minimum numbers of samples of these two groups are no less than 2 are required. So the vertex set of the expression network **EN** is in fact G where

$$G = \{i \in G_0 : n_i^{(1)} \geq 2, n_i^{(2)} \geq 2\}.$$

For any gene $i \in G$, the vertex weight of **EN** can be defined as:

$$f_i = 1 - 1/d \sum_{r=1}^{d} p_r,$$

where d is the number of genes in G_2, and p_r is the p-value of differential expression of gene r relative to i's mutation status. The meaning is that the smaller the p-values the stronger the influence of the gene's mutation to others. That is, it is more likely to be drivers, so it should have greater weights.

For any two genes i and j in G, the edge weight u_{ij} is defined as the absolute value of Pearson correlation between expressions of i and j across the samples in S.

2) The network based on somatic mutation: To hold the same vertex set with the expression network **EN**, in this subsection the gene set G is also used to construct the network based on somatic mutations, called *Mutation Network* (**MN**). For any gene $i \in G$, let m_i denote the number of mutations for i across the samples in the mutation matrix A, i.e., $m_i = \sum_r a_{ri}$. The vertex weight is defined as

$$h_i = m_i/m,$$

where m is the number of all samples in A. For any pair of genes i and j in G, the edge weight v_{ij} is defined as the number of samples where exactly one of the pair is mutated divided by the number of samples where at least one of the pair

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

is mutated in A. It is clear that the vertex weight measures the mutation coverage and the edge weight measures the mutual exclusivity.

3) The integrative network: An integrative network $\mathcal{M}$ can be obtained by synthesizing the expression network **EN** and the mutation network **MN**.

We notice that in **EN** or **MN** the vertex weights and edge weights have different measurement of levels. To balance these two terms, define $f = \max f_i$ and $u = \max u_{ij}$ in **EN** (similarly, $h = \max h_i$ and $v = \max v_{ij}$ in **MN**). Set $\xi = u/f$, and $\eta = v/h$. Let $F = \{f_i\}$ and $U = \{u_{ij}\}$ denote the sets of vertex weights and edge weights in **EN**, respectively (similarly, $H = \{h_i\}$ and $V = \{v_{ij}\}$ in **MN**). Then U and ξF (similarly, V and ηH) have the balanced values.

On the other hand, considering that gene expression values often contain noises, it is proper to give more importance for **MN** than **EN** while integrating these two networks. Set $\delta \cdot (u/v) = k$, where k is a quantity reflecting the relative importance of **MN** respective to **EN**. Then $\delta = k/(u/v)$. In this paper $k = 2$ is used.

The integrative network $\mathcal{M}$ with edge weights w_{ij} and vertex weights c_i can be defined as follows:

$$w_{ij} = \delta \cdot u_{ij} + v_{ij}, \quad c_i = \delta\xi \cdot f_i + \eta \cdot h_i, \quad i, j = 1, \cdots, n, \tag{1}$$

where n is the number of genes in G. From the above discussion we know that ξ and η can be directly determined by the networks **EN** and **MN**. Also is the case for δ once k is preassigned.

C. An optimization model for detecting coherent subnetworks

For the integrative network $\mathcal{M}$, our goal is to extract some modules (subnetworks) with high weights for both edges and vertices, the optimization model in [23] can be used for this purpose. With w_{ij} and c_i defined in Eq. (1), the model is as follows:

$$\begin{aligned} \max \quad & \textstyle\sum_i \sum_j w_{ij} x_i x_j + \lambda c_i x_i, \\ s.t. \quad & x_1^\beta + x_2^\beta + \cdots + x_n^\beta = 1, \\ & x_i \geq 0, i = 1, \cdots, n, \end{aligned} \tag{2}$$

where the n-dimensional non-negative vector $x = (x_1, x_2, \cdots, x_n)$, determined by solving the optimization model, represents the degree of each node belonging to some specific subnetwork. The first term in the objective function measures the interconnectivity within the subnetwork, while the second term measures the degree of association between the nodes and the subnetwork. In the model a positive parameter λ is introduced to balance these two terms.

On the other hand, a trivial solution will be obtained when model (2) is unconstrained where all nodes from the original network can be included into the subnetwork, so a regularization constraint should be introduced to limit the number of nodes selected. This is the role of β which can adjust the strength of regularization applied to the variable $x = (x_1, x_2, \cdots, x_n)$. When $\beta = 2$, it is very attractive in many cases since the optimization of a quadratic function over a sphere is polynomially solvable in contrast to general nonconvex programming [24] but tends to select all the nodes in the network to the final subnetwork. When $\beta = 1$, this L_1-type constraint will lead to a sparse solution, i.e., many of the entries in the final optimal solution x will be zeros [25]. Usually we use $\beta = 1$ in model (2) in order to extract small-sized subnetworks from a large network.

The optimization model (2) can be easily solved by quickly finding a local minimum from a predetermined initial solution through the following iterative algorithm [23]:

$$x_i^{t+1} = \left(x_i^t \frac{2(WX)_i + \lambda c_i}{2X^T WX + \lambda \sum_i c_i x_i^t} \right)^{\frac{1}{\beta}}, \tag{3}$$

where $W = (w_{ij})$ is the $n \times n$ edge weight matrix, and $X = (x_1^t, x_2^t, \cdots, x_n^t)^T$ is the n-dimensional solution vector at time t. Algorithm (3) is convergent and the non-zero entries in solution x (determined in practice as entries that are greater than a cutoff, 0.03 is used in this paper) define a certain subnetwork (module). After one locally optimal solution is obtained, these corresponding vertices are eliminated from the network, and the whole procedure is then iterated, i.e., we solve for another locally optimal solution and its corresponding subnetwork based on the new network.

D. Significance test of the subnetwork (module)

We perform a random test to assess the significance of the results. For a selected subnetwork **SN** with b vertices, we get a quantity C by summing up all the vertex weights and edge weights involved in **SN**. Then we randomly select b vertices from the original network and also get a similar quantity CR. This procedure is repeated for 1000 times and the number r of CRs which is larger than C can be calculated. The significance p-value of **SN** (denoted by p_1) can be obtained through the quantity of r divided by 1000.

E. Mutual exclusivity test of the subnetwork (module)

After a subnetwork has passed the significance test, the following step is to evaluate whether it exhibits a pattern of mutually exclusive genomic alterations. Here the "switching permutation" method proposed by Ciriello et al. [15] was used for this purpose, in which a Markov chain Monte Carlo permutation strategy is adopted based on random network generation models.

Furthermore, it is imaginable that even though a subnetwork **SN** with b ($b > 2$) vertices is not significantly mutually exclusive, we cannot exclude that one of its subsets actually is. In this case two strategies can be adopted. One is to reduce the scale of the subnetwork sequentially, that is, a subset **SN**$'$ of size $b-1$, contained in **SN**, is selected which is more likely to be significant among all the subsets of **SN** with $b-1$ vertices. This can be realized by removing from **SN** the less informative vertex (gene), i.e., the one with the smallest number of unique alterations [15]. The process is repeated recursively until either of the two conditions is reached: **SN**$'$ is significantly mutually exclusive or $b = 2$. On the other hand, because high weights

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

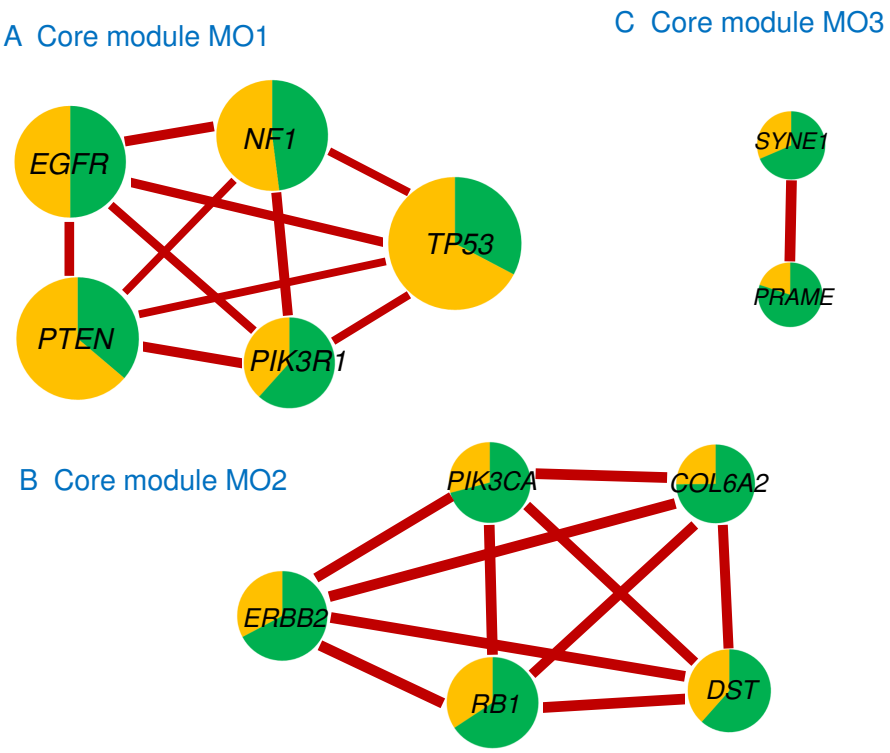

Fig. 1. The schematic figure of the mutated core modules in GBM obtained by our integrative network analysis. The area of each vertex (gene) and the width of each edge are proportional to their weights in the integrative network $\mathcal{M}$. The orange and green in the vertex correspond to the gene's coverage and the influence of its mutation status to other genes, respectively, whereas the dark red edge corresponds to the mutual exclusivity and expression correlation between the genes. Here all the weights are extracted from the 118-gene network $\mathcal{M}$ wherein the module $MO1$ in A is directly identified. The modules $MO2$ in B and $MO3$ in C can be obtained by removing the detected module(s) from $\mathcal{M}$ and recalculating the weights using Eq. (1). For clarity, we do not display the recalculated weights here.

are simultaneously required for the subnetwork, an alternative strategy is to hold the largest entry vertex and examine which else gene is mutually exclusive with it. Hereafter in this paper, we denote the exclusivity p-value by p_2 for concise description.

III. Results

In this section the results will be presented for the GBM data from TCGA (see the above Data subsection). Through the construction procedure of the integrative network we have 118 genes left in $\mathcal{M}$. Three core modules are obtained by performing algorithm (3) on $\mathcal{M}$, where $\lambda = 1$ is used. A schematic figure of the modules are displayed in Fig. 1.

A. Core module $MO1$

The first module $MO1$ consists of five genes: *EGFR, NF1, PTEN, PIK3R1* and *TP53*. $MO1$ is significant with $p_1 = 0.00$ and $p_2 = 0.01$.

The first four genes in $MO1$ (i.e., *EGFR, NF1, PTEN* and *PIK3R1*) are in the *RTK/RAS/PI(3)K* signalling pathway, one of the three core pathways altered in the development of glioblastoma deduced by the TCGA Research Network [1]. *NF1* is a human glioblastoma suppressor gene, and *EGFR* is frequently activated in primary glioblastomas [1]. Both of them have been used as biomarkers for identifying the glioblastoma subtypes [26]. *PTEN* functions primarily by regulating *RTK/PI3K/AKT* signaling through its lipid phosphatase activity. As a tumor suppressor gene, *PTEN*'s mutations and deletions inactivate its enzymatic activity which may lead to increased cell proliferation and reduced cell death. Frequent genetic inactivation of *PTEN* occurs in glioblastoma, endometrial cancer, and prostate cancer. The gene *PIK3R1*, except occurring in the core glioblastoma pathway in [1], it has also been reported to be involved in human cancers before, including glioblastoma [27]. *TP53* is an important tumor suppressor and it is the most commonly mutated gene for the samples of the TCGA GBM data ($\sim 28.9\%$). It is known that mutations in *TP53* and *PTEN* are both obligate events in the pathogenesis of human glioblastoma. However, there are also studies indicate that *PTEN* loss may disrupt cellular homeostasis enough to be detected as a cellular stress inducing a low-level astrogliosis response, but it is insufficient to drive proliferation, consistent with the inability to initiate gliomagenesis in the absence of other mutations. In [28] the authors examined the cooperativity between these two tumor suppressors *TP53* and *PTEN* in mature mice and they concluded that combined inactivation of *PTEN* and *TP53* induced high-grade astrocytomas. So maybe it is reasonable to

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

think that *TP53* loss is required for the genesis of glioblastoma if *PTEN* deletion appears [29]. On the other hand, Zhu et al. investigated the cooperativity between *TP53* and *NF1* and they found that early inactivation of *TP53* tumor suppressor gene cooperating with *NF1* loss induces malignant astrocytoma [30]. All these indicate that although *TP53* is in another core pathway in [1], it is of certain reasonability to include *TP53* in the module $MO1$.

B. Core module $MO2$

The second module $MO2$ is obtained by removing $MO1$ from $\mathcal{M}$ and performing (3) on the remaining genes. Now five genes including *COL6A2, DST, ERBB2, PIK3CA* and *RB1* were detected with $p_1 = 0.00$ and $p_2 = 0.00$, which indicates that $MO2$ is very significant for both statistical tests.

The gene *RB1* is a principal member of the *RB* signalling core pathway for glioblastoma in [1]. In fact, it is a key regulator of entry into cell division that acts as a tumor suppressor, and one of its functions is to prevent excessive cell growth by inhibiting cell cycle progression until a cell is ready to divide. *RB1* is dysfunctional in several major cancers [31]. *COL6A2* encodes one of the three alpha chains of type VI collagen. Its product contains several domains which have been shown to bind extracellular matrix proteins, an interaction that explains the importance of this collagen in organizing matrix components. Recently a multi-cancer stage-associated gene expression signature has been identified, consisting of a set of genes that are coordinately overexpressed only in samples of cancer that have exceeded a particular stage specific to each cancer type. The signature contains numerous epithelial-mesenchymal transition (EMT) markers, such as the EMT-inducing transcription factor *Slug* (*SNAI2*), as well as *FAP, COL6A2*, etc [32]. And in [32] the authors demonstrated that the signature was strongly associated with the phenotype "Days to Tumor Recurrence" in glioblastoma. Moreover, *Slug* has also recently been found to be associated with invasiveness in glioma [33], and *COL6A2* is one of the several genes with high correlation expression with *Slug* [32]. In [34] *COL6A2* was selected as one of the endothelial marker genes and has been confirmed to be conserved in primary and metastatic brain tumors. *COL6A2* has also been reported to be related to glioblastoma in other papers such as [35], [36].

The genes *ERBB2* and *PIK3CA* are also contained in the *RTK/RAS/PI(3)K* signalling pathway [1]. *ERBB2* mutation has also previously been reported in glioblastoma tumor in [37]. Like *PIK3R1*, gene *PIK3CA* also belongs to *PI(3)Ks*, which are a family of lipid kinases capable of phosphorylating the 3'OH of the inositol ring of phosphoinositides. They are responsible for coordinating a diverse range of cell functions including proliferation, cell survival, degranulation, vesicular trafficking and cell migration. Moreover, frequent activating missense mutations of *PIK3CA* have been previously reported in multiple tumor types, including glioblastoma [38], [39]. The gene *DST* encodes a member of the plakin protein family of adhesion junction plaque proteins. Multiple alternatively spliced transcript variants encoding distinct isoforms have been found for this gene. It has been known that some isoforms are expressed in neural and muscle tissue, and some isoforms are expressed in epithelial tissue. Consistent with the expression, mice defective for this gene shows skin blistering and neurodegeneration [40]. Because there has not been any report about the relationship between *DST* mutation and glioblastoma, maybe this is a new gene for the pathogenesis of human glioblastoma.

C. Core module $MO3$

After removing $MO1$ and $MO2$ from $\mathcal{M}$ and performing (3) on the remaining network the third module $MO3$ is obtained which contains two genes: *PRAME* and *SYNE1*. Actually, at first seven genes including $MO3$ and five others were detected with $p_1 = 0.00$ and $p_2 = 0.47$. Then sequential removing less informative genes was performed but none of the subsets passed the exclusivity test. Finally, we held the largest entry gene *SYNE1* and respectively examined each gene pair. The gene *PRAME* was identified which was very significantly exclusive with *SYNE1* with a p-value $p_2 = 0.00$.

Gene *SYNE1* mutation is known to influence cerebellar ataxia, and is associated with lung, ovarian, and colorectal cancers. But it has not been highlighted in previous studies using TCGA GBM data and there has not been any correlation between GBM and *SYNE1* mutation in the literature until recently. In [21] the authors found one large network in which the genes' expression changes are associated with the mutated *SYNE1* gene, wherein several known oncogenes are included. The results suggest that *SYNE1* mutation is important in TCGA GBM tumor samples. *SYNE1* and its associated genes may be new targets for future treatments. In [41] *SYNE1* was highlighted to be associated with the GBM patients' lifetime, so it is an important biomarker of glioblastoma survival. *PRAME* was previously reported to be associated with melanoma and acute leukemias. And recently it has been reported to be involved in the pathogenesis of glioblastoma also [42], [43].

IV. Conclusion

In this paper, a network-based method is presented which integrates somatic mutations and gene expressions to find out mutated core modules in cancer. Different from some previous approaches exploring pathways or modules, our method does not use any prior information such as human PPI networks and known pathways.

In the construction of the integrative network $\mathcal{M}$ and the performing process of the optimization algorithm (3) there are two parameters, i.e., k and λ, need to be further explained. This is a typical feature of our method, which employ two parameters to balance not only different sources of data but also the vertices and edges of the weighted network constructed from the data. On one hand, this provides flexibility for using the method because one can choose different parameters depending on which factor he/she puts more emphases. On the other hand, different choices of parameters may result in

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

slightly different results. This needs further consideration in practice according to the actual data.

Three core modules are identified in this paper when the method with $k = 2$ and $\lambda = 1$ was applied to the TCGA GBM data (referring to Fig. 1). Among the modules, there are not only some well-known oncogenes and tumor suppressors that have been previously implicated in GBM pathogenesis, but also some others which have not or rarely been reported earlier in the context of glioblastoma multiforme. These results indicate that the presented method of integrative network analysis can be expected to provide useful information for the study of pathogenesis in cancer.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China, No. 11131009 and 11001256, the 'Special Presidential Prize' – Scientific Research Foundation of the CAS and the Foundation for Members of Youth Innovation Promotion Association, CAS.

REFERENCES

[1] The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061-1068 (2008).
[2] International cancer genome consortium. International network of cancer genome projects. *Nature* 464, 993-998 (2010).
[3] M. Meyerson, S. Gabriel and G. Getz, Advances in understanding cancer genomes through second-generation sequencing. *Nat Rev Genet* 11, 685-696 (2010).
[4] C. Greenman et al., Patterns of somatic mutation in human cancer genomes. *Nature* 446, 153-158 (2007).
[5] R. Beroukhim et al., Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc Natl Acad Sci* 104, 20007-20012 (2007).
[6] L. Ding et al., Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* 455, 1069-1075 (2008).
[7] S. Jones et al., Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* 321, 1801-1806 (2008).
[8] B. Vogelstein and K. W. Kinzler, Cancer genes and the pathways they control. *Nat Med* 10, 789-799 (2004).
[9] D. Hanahan and R. A. Weinberg, Hallmarks of cancer: The next generation. *Cell* 144, 646-674 (2011).
[10] S. M. Boca et al., Patient-oriented gene set analysis for cancer mutation data. *Genome Biol* 11, R112 (2010).
[11] S. Efroni et al., Detecting cancer gene networks characterized by recurrent genomic alterations in a population. *PLoS ONE* 6, e14437 (2011).
[12] F. Vandin, E. Upfal and B. J. Raphael, De novo discovery of mutated driver pathways in cancer. *Genome Res* 22, 375-385 (2012).
[13] J. Zhao, S. Zhang, L. Y. Wu and X. S. Zhang. Efficient methods for identifying mutated driver pathways in cancer. In submission (2012).
[14] E. Cerami et al., Automated network analysis identifies core pathways in glioblastoma. *PLoS One* 5, e8918 (2010).
[15] G. Ciriello et al., Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res* 22, 398-406 (2012).
[16] M. Kanehisa and S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 28, 27-30 (2000).
[17] G. Joshi-Tope et al., Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 33(suppl 1): D428-432 (2005).
[18] C. A. Miller et al., Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med Genomics* 4, 34 (2011).
[19] T. Ideker, O. Ozier, B. Schwikowski and A. F. Siegel, Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics* 18(Suppl 1), S233-240 (2002).
[20] E. Segal, H. Wang and D. Koller, Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics* 19(Suppl 1), i264-272 (2003).
[21] D. L. Masica and R. Karchin, Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. *Cancer Res* 71, 4550-4561 (2011).
[22] G. W. Roel et al., Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98-110 (2010).
[23] Y. Wang and Y. Xia, Condition specific subnetwork identification using an optimization model. *Proceedings of the second international symposium on Optimization and Systems Biology (OSB'08)*, 333-340 (2008).
[24] Y. Ye, A new complexity result on minimization of a quadratic function with a sphere constraint. in *Recent advances in global optimization*, New York: Princeton University Press Princeton, 19-31 (1992).
[25] R. Tibshirani, Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58, 267-288 (1996).
[26] R. G. Verhaak et al., Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* 17, 98-110 (2010).
[27] M. Mizoguchi et al., Genetic alterations of phosphoinositide 3-kinase subunit genes in human glioblastomas. *Brain Pathol* 14, 372-377 (2004).
[28] L. M. Chow et al., Cooperativity within and among Pten, p53, and Rb pathways induces high-grade astrocytoma in adult brain. *Cancer Cell* 19, 305-316 (2011).
[29] H. Zheng et al., p53 and Pten control neural and glioma stem/progenitor cell renewal and differentiation. *Nature* 455, 1129-1133 (2008).
[30] Y. Zhu et al., Early inactivation of p53 tumor suppressor gene cooperating with NF1 loss induces malignant astrocytoma. *Cancer Cell* 8, 119-130 (2005).
[31] M. Nakamura et al., Promoter hypermethylation of the RB1 gene in glioblastomas. *Lab Invest* 81, 77-82 (2001).
[32] W. Y. Cheng et al., A Multi-Cancer Mesenchymal Transition Gene Expression Signature Is Associated with Prolonged Time to Recurrence in Glioblastoma. *PLoS ONE* 7, e34705 (2012).
[33] H. W. Yang et al., SNAI2/Slug promotes growth and invasion in human gliomas. *BMC Cancer* 10: 301 (2010).
[34] Y. Liu et al., Vascular gene expression patterns are conserved in primary and metastatic brain tumors. *J Neurooncol* 99, 13-24 (2000).
[35] C. L. Tso et al., Distinct Transcription Profiles of Primary and Secondary Glioblastoma Subgroups. *Cancer Res* 66, 159-167 (2006).
[36] X. L. Xu and A. M. Kapoun, Heterogeneous activation of the TGF pathway in glioblastomas identified by gene expression-based classification using TGF-responsive genes. *Journal of Translational Medicine* 7: 12 (2009).
[37] P. Stephens et al., Lung cancer: intragenic ERBB2 kinase mutations in tumours. *Nature* 431, 525-526 (2004).
[38] Y. Samuels et al., High frequency of mutations of the PIK3CA gene in human cancers. *Science* 304, 554 (2004).
[39] G. L. Gallia et al., PIK3CA gene mutations in pediatric and adult glioblastoma multiforme. *Mol Cancer Res* 4, 709-714 (2006).
[40] Entrez Gene: DST dystonin.
[41] N. VL Serão et al., Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival. *BMC Medical Genomics* 4: 49 (2011).
[42] H. Dong et al., Integrated analysis of mutations, miRNA and mRNA expression in glioblastoma. *BMC Systems Biology* 4: 163 (2010).
[43] F. M. Hofman et al., Immunotherapy of Malignant Gliomas Using Autologous and Allogeneic Tissue Cells. *Anti-Cancer Agents in Medicinal Chemistry* 10, 462-470 (2010).

978-1-4673-4398-5/12/$31.00 ©2012 IEEE