Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data

Junhua Zhang*, Shihua Zhang*, Yong Wang, Xiang-Sun Zhang

From The 6th International Conference on Computational Systems Biology (ISB2012)
Xi'an, China. 18-20 August 2012

Abstract

Motivation: Understanding the molecular mechanisms underlying cancer is an important step for the effective diagnosis and treatment of cancer patients. With the huge volume of data from the large-scale cancer genomics projects, an open challenge is to distinguish driver mutations, pathways, and gene sets (or core modules) that contribute to cancer formation and progression from random passengers which accumulate in somatic cells but do not contribute to tumorigenesis. Due to mutational heterogeneity, current analyses are often restricted to known pathways and functional modules for enrichment of somatic mutations. Therefore, discovery of new pathways and functional modules is a pressing need.

Results: In this study, we propose a novel method to identify Mutated Core Modules in Cancer (iMCMC) without any prior information other than cancer genomic data from patients with tumors. This is a network-based approach in which three kinds of data are integrated: somatic mutations, copy number variations (CNVs), and gene expressions. Firstly, the first two datasets are merged to obtain a mutation matrix, based on which a weighted mutation network is constructed where the vertex weight corresponds to gene coverage and the edge weight corresponds to the mutual exclusivity between gene pairs. Similarly, a weighted expression network is generated from the expression matrix where the vertex and edge weights correspond to the influence of a gene mutation on other genes and the Pearson correlation of gene mutation-correlated expressions, respectively. Then an integrative network is obtained by further combining these two networks, and the most coherent subnetworks are identified by using an optimization model. Finally, we obtained the core modules for tumors by filtering with significance and exclusivity tests. We applied iMCMC to the Cancer Genome Atlas (TCGA) glioblastoma multiforme (GBM) and ovarian carcinoma data, and identified several mutated core modules, some of which are involved in known pathways. Most of the implicated genes are oncogenes or tumor suppressors previously reported to be related to carcinogenesis. As a comparison, we also performed iMCMC on two of the three kinds of data, i.e., the datasets combining somatic mutations with CNVs and secondly the datasets combining somatic mutations with gene expressions. The results indicate that gene expressions or CNVs indeed provide extra useful information to the original data for the identification of core modules in cancer.

Conclusions: This study demonstrates the utility of our iMCMC by integrating multiple data sources to identify mutated core modules in cancer. In addition to presenting a generally applicable methodology, our findings provide several candidate pathways or core modules recurrently perturbed in GBM or ovarian carcinoma for further studies.

* Correspondence: zjh@amt.ac.cn; zsh@amss.ac.cn
National Center for Mathematics and Interdisciplinary Sciences, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

© 2013 Zhang et al.; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Background

Cancer is a complex disease and multiple factors including genomic, epigenomic, and gene expression aberrations are involved in its formation and development [1]. Understanding the pathogenesis of cancer at the molecular level is a great challenge and will shed lights on the effective diagnosis and treatment of cancer patients. Rapid advances in high-throughput sequencing technologies create opportunities to address this task. Large-scale cancer genomics projects, such as the Cancer Genome Atlas (TCGA) [2], International Cancer Genome Consortium (ICGC) [3] and the Catalogue Of Somatic Mutations In Cancer (COSMIC) [4], have produced a large volume of data in recent years, providing a basis for systems level understanding of cancer formation and progression [5].

In general, cancer genomes possess a large number of mutations including somatic mutations and copy number variations (CNVs). Among them, some mutations contributing to cancer progression from the normal to the malignant state are called driver mutations, and those that accumulate in cells but do not contribute to cancer development are called passengers [6,7]. Therefore, distinguishing the functional driver mutations, driver pathways or core modules from random passengers will be a crucial step in understanding the molecular mechanisms of carcinogenesis, which can further aid in effective diagnosis, treatment and prognosis of cancer patients.

Initially, efforts were devoted to detect individual driver genes that cause tumors. A standard approach for this is to identify recurrent mutations in a large cohort of cancer patients. But the extensive mutational heterogeneity of cancer genomes [2,8,9] makes this kind of method sometimes ineffective because patients even from the same tumor type can have different driver mutations.

Further studies revealed that the acquisition of tumorigenic properties, such as cell proliferation, angiogenesis, or metastasis are mainly due to disruption of some cellular signaling and regulatory pathways [10,11]. Driver mutations either directly target such biological pathways or tend to cluster within closely knitted network modules which are closely linked to specific biological pathways [12,13]. Thus, identification of mutated driver pathways or core modules is of primary importance for understanding cancer initiation and progression. Moreover, a great deal of investigation indicates that genes in the driver pathway or core module usually cover a large number of samples and exhibit mutual exclusivity, these two criteria are commonly used in the pathway or module based methods. For example, Ding et al. [8] and Jones et al. [9] analyzed known pathways for enrichment of somatic mutations, Boca et al. [14] and Efroni et al. [15] detected known pathways which are significantly mutated across the patients, and Cerami et al. [16] and Ciriello et al. [17] identified oncogenic network modules by using somatic mutation and the human reference network. Although the priori knowledge (such as protein-protein interactions (PPI) and signal transduction pathways) can provide some useful information for the detection of driver mutations, the incompleteness of the human PPI network and the existence of many unknown pathways may limit the wide application of such methods in some extent. Recently, methods and algorithms were developed for de novo discovery of mutated driver pathways and functional modules in tumors based solely on cancer genomic data [18-20].

On the other hand, somatic mutations and CNVs in cancer genomes frequently perturb the expression level of affected genes and thus disrupt pathways controlling normal growth. Genes in the same pathway usually have similar gene expression profiles and thus can coordinate achieve a particular function [21]. Several studies have demonstrated the necessity of integrating gene expression information to identify candidate driver genes and driver pathways [19,22,23].

In this study we present an integrative method, called iMCMC (identify Mutated Core Modules in Cancer) that integrates gene sequence and expression information to identify mutated core modules in cancer. A typical characteristic of iMCMC is that it uses only cancer genomic data without any prior knowledge such as PPI networks and known pathways. First, somatic mutations and CNVs are used to generate a mutation network, similarly an expression network is obtained from the gene expression profiles. Then, an integrative molecular network is constructed by combining these two networks (i.e., integrating the three different kinds of data). Finally, an optimization model is used to identify coherent subnetworks (modules), which are further assessed by statistical tests. These are key contributions of our approach. The main consideration is that cooperative dysregulation of gene sequence and expression may contribute to cancer formation and progression. Furthermore, cellular networks contain functional modules, and tumors usually target specific modules critical to their growth. More importantly, our weighted integrative network is constructed to take into consideration possible features of genes in the driver pathways or core modules: large coverage, mutual exclusivity, strong influence of a gene’s mutation on other genes, and high correlation of gene mutation-correlated expressions. All these factors are reflected in the vertex weight or edge weight of the integrative network. Applying iMCMC to the TCGA glioblastoma multiforme (GBM) and ovarian carcinoma data, we identified five and two mutated core modules, respectively. In the GBM data, the involved pathways include parts of the RB signaling and RTK signaling pathways (CDKN2B, CDK4, EGFR, NF1), and in the ovarian carcinoma data a recurrent mutated module related to
cell cycle and DNA repair (CCNE1, MYC, RAD52) is revealed. Importantly, most of the implicated genes are oncogenes or tumor suppressors previously reported to be related to cancer pathogenesis (others include TP53, PTEN, RB1, MDM2 for GBM, and KRAS for ovarian carcinoma). Furthermore, to investigate the possible role of gene expressions or CNVs for the identification of mutated core modules, we also performed iMCMC on the datasets consisting of somatic mutations combined with CNVs or gene expressions. The results indicate that each indeed provides extra useful information to the original data for module detection. To conclude, as a generally applicable methodology, iMCMC can identify not only some known pathways but also provide candidate pathways or core modules recurrently perturbed in cancer for further studies.

**Results and discussion**

**Overview of our method**

Three kinds of data including somatic mutations, CNVs, and gene expressions were used in this study. All data were downloaded from the TCGA website (https://tcga-data.nci.nih.gov/tcga/). The proposed iMCMC method for identification of mutated core modules contains six steps. A schematic overview of iMCMC is displayed in Figure 1. For additional details please refer to the Materials and methods section.

**Step 1:** A mutation matrix is obtained by combining somatic mutations and CNVs, and an expression matrix is generated from gene expression profiles.

**Step 2:** A mutation network and an expression network are constructed based on the mutation and expression matrices, respectively.

**Step 3:** These two networks are integrated into an integrative network.

**Step 4:** Coherent subnetworks (modules) are identified using an optimization model.

**Step 5:** A random test is performed to assess significance of the selected subnetworks, for which a p-value $p_1$ is obtained.

**Step 6:** Finally, a Markov chain Monte Carlo permutation strategy is adopted to test mutual exclusivity of the subnetworks, and a p-value $p_2$ is calculated.

In the end, core mutated modules can be obtained if the selected subnetworks pass the last two statistical assessments.

**Application to glioblastoma multiforme (GBM)**

Among the glioblastoma dataset obtained from TCGA, DNA copy number variations are present in 169 samples, gene expression profiles in 202 and nucleotide sequence aberrations in 135 samples. Using the construction procedure of the integrative network 93 genes were left in the integrative network $\mathcal{M}$ (notice that some of these are metagenes - genes that are mutated in the same samples). These genes are present in 90 samples common to all three kinds of data. Five core modules are obtained by performing iMCMC on $\mathcal{M}$, where $\lambda = 1$ is used (see Materials and methods).

The first module consists of CDKN2A and CYP27B1 and covers 60 GBM samples. Initially, five genes including CDK4, CDKN2A, CDKN2B, CYP27B1, and MTAP were detected. This group of genes has a significant p-value of $p_1 < 0.001$ and the exclusivity p-value of $p_2 = 1$. After sequentially removing some co-occurring genes, we obtained CDKN2A and CYP27B1 with $p_2 < 0.001$. Previously, CDKN2B and CYP27B1 were identified as the most frequently sampled pair for GBM [18]. Both CDKN2A and CDKN2B are tumor suppressors located on 9p21.3-22.3 which is a common homozygous deletion region on the human chromosome. These two genes mutate almost simultaneously in all samples, so they have a very low exclusivity value (0.06), and it is not contradictory for us to identify CDKN2A instead of CDKN2B in the module for further analysis. CDKN2A encodes protein p16, which is a tumor suppressor protein with an important role in cell cycle regulation [24]. Mutations in CDKN2A are associated with increased risk in a wide range of cancers. Especially, recent studies showed that CDKN2A in high-grade glioma tissues was significantly down-regulated than in low-grade glioma tissues [25], which indicates that CDKN2A may be involved in malignant glioma carcinogenesis. CYP27B1 plays an important role in normal bone growth, calcium metabolism, and tissue differentiation. Gene amplification and mRNA splice variants of CYP27B1 in human glioblastoma were also previously reported [26].

The second module is obtained by removing CDKN2A and CYP27B1 from the integrative network $\mathcal{M}$ and performing iMCMC on the remaining genes (Figure 2). CDKN2B and a metagene including CDK4 and TSPAN31 were identified with a coverage rate of 63/90. This module is significant and the genes CDKN2B and CDK4/TSPAN31 are mutually exclusive with $p_1 < 0.001$ and $p_2 < 0.001$. Several studies have found that variants of CDKN2B are associated with high-grade glioma susceptibility [27]. Feng et al. [28] made an integrated analysis of multiple kinds of data at 9p21.3 in glioblastoma and showed that the complete loss of 9p21.3 and low CDKN2B expression were associated with worse prognosis for both tumor progression/recurrence-free survival. The functional importance of CDK4 in astrocytic tumourigenesis, particularly during the later stages of tumor progression has been reported [29]. This gene has also been a putative prognostic marker and related to the survival of GBM patients [30,31]. As an oncogene, CDK4 is suppressed by CDKN2B in the RB signaling pathway (Figure 2B). The gene TSPAN31 is thought to be involved
in growth-related cellular processes, because the encoded protein mediates signal transduction events thus plays a role in the regulation of cell development, activation and growth. TSPAN31 is associated with tumorigenesis although there is no report about its relationship with GBM. However, TSPAN31 was also found highly amplified in a number of GBM patients elsewhere. Here TSPAN31 and CDK4, as a metagene, mutate in the same samples, and both are relatively correlated to CDKN2B (Figure 2A).

A more detailed explanation of Figure 2A will help demonstrate the advantage of our framework, which can further enable understanding of how the three kinds of data are integrated and utilized for identification of the module. Both mutation (including somatic mutations and CNVs) and expression information exists not only in the edge but also in the vertex. In the vertex of Figure 2A, these correspond to the red and blue parts, respectively, which represent the proportion of coverage and the influence of mutation status on other genes while along the...
edge they point to mutual exclusivity and gene expression correlation respectively. This and other figures show that exclusivity in the detected module is always large due to the application of mutual exclusivity test in iMCMC. Although the expression correlation is not very high, the influence of some genes on other genes calculated from their expression is sometimes heavily utilized in the vertex. Therefore, we presume that gene expression indeed plays an important role in the identification of mutated core modules which is reflected either in the edge or in the vertex of the integrative network.

Performing iMCMC after removal of the foregoing two modules from $M$ results in the third module with three genes ($TP53$, $PTEN$ and $MTAP$) and the fourth module including three other genes ($EGFR$, $NF1$ and $MDM2$) and a metagene ($CHAT/SLC18A3$). Both modules are highly significant ($p_1 < 0.001$ and $p_2 < 0.001$) and cover 70 and 46 GBM samples, respectively. Finally, our method identifies a module including $RB1$ and a metagene $DKK1/PRKG1/CSTF2T$ significant at $p_1 < 0.001$ and $p_2 = 0.02$ levels. Besides these no other significant modules are detectable.

In the third module, both $TP53$ and $PTEN$ are important tumor suppressors [32]. When $PTEN$ is mutated or deleted its enzymatic activity will be inactivated which may lead to increased cell proliferation and reduced cell death. Several studies indicate that concomitant inactivation of $TP53$ and $PTEN$ promoted the development of glioblastoma. This co-operative nature was also validated in adult brain of mature mice [33]. The gene $MTAP$ encodes an enzyme that plays a major role in polyamine metabolism and is important for the adenine and methionine salvage pathway [34]. A number of studies indicate that $MTAP$ deficiency is a common occurrence in various cancers including glioblastomas, non-small cell lung cancer, melanoma, pancreatic and endometrial cancer [35,36]. Here $MTAP$ not only has stronger exclusivity but also higher gene expression correlation with $TP53$ than $PTEN$ in the identified module.

The genes $EGFR$ and $NF1$ in the fourth module are involved in the RTK signaling pathway (Figure 3B), which is one of the core pathways altered in the development of GBM [2]. $NF1$ is a human glioblastoma suppressor gene while $EGFR$ is frequently activated in primary glioblastomas. Both have been used as biomarkers for the identification of the glioblastoma subtypes [37]. Amplification of $MDM2$ or increased expression occurs in many tumors [38]. Although $TP53$ and $MDM2$ often form a negative feedback loop by $MDM2$ inhibiting $TP53$ activity which results in transcriptional up-regulation of $MDM2$ expression, functions of $MDM2$ independent of $TP53$ have also been identified. For example, Biernat et al. demonstrated the molecular mechanism of $MDM2$’s escape from $TP53$-regulated growth control [39]. The gene $CHAT$ encodes an enzyme which catalyzes the biosynthesis of the neurotransmitter acetylcholine. $SLC18A3$ is located within the first intron of $CHAT$ and aids in the transport of acetylcholine, synthesized by $CHAT$, into secretory vesicles for release into the extracellular space. $CHAT$ is presently the most specific indicator available to monitor the functional state of cholinergic neurons in the central and peripheral nervous systems [40]. Central cholinergic neurons are involved in several neurodegenerative diseases such as Alzheimer’s disease and amyotrophic lateral sclerosis. Abnormalities of $CHAT$ in the brain have also been demonstrated in schizophrenia and sudden infant death syndrome. In the fourth module, $MDM2$ and $EGFR$ as well as $MDM2$ and $NF1$ are highly exclusive with high correlation in expression (Figure 3). Moreover, high exclusivity or correlation is also observed between $NF1$ and $EGFR$, $NF1$ and $CHAT/SLC18A3$ as well as $EGFR$ and $SLC18A3$, respectively.

Application to ovarian cancer

The ovarian carcinoma dataset from TCGA describes DNA copy number variations in 559 high-grade serous ovarian adenocarcinomas, gene expression profiles in 489 tumors and DNA sequence aberrations in coding genes.
of 320 tumors. After preprocessing, we obtained 371 genes in the integrative network $M$ in 311 samples. Here, some genes merged as metagenes.

We notice that $TP53$ is the most commonly mutated gene and is present in more than 80% of the high-grade serous ovarian carcinomas while all other genes are mutated in less than 27% of samples. In addition, analysis of $TTN$ mutations indicates that these are likely to be artifacts [41]. Considering the prevalence of $TP53$ mutation and the possible inaccuracy of $TTN$ mutations, we removed these two genes from $M$ and performed iMCMC on the remaining genes for which two mutated core modules are identified.

The first module consisting of three genes, $CCNE1$, $MYC$ and $RAD52$, is statistically significant with $p_1 < 0.001$ and $p_2 < 0.001$. This module is approximately exclusively mutated in 150 samples. $CCNE1$ and $MYC$ are two important genes engaged in cell cycle progression (Figure 4B). The gene $CCNE1$ is essential for the control of the cell cycle at the G1/S transition. In many tumors overexpression of this gene results in chromosome instability that may contribute to tumorigenesis [42]. Nakayama et al. demonstrated that amplification of $CCNE1$ is related to poor survival suggesting that $CCNE1$ can be a potential therapeutic target in the treatment of ovarian cancer [43]. $MYC$ is a strong proto-oncogene that codes a transcription factor and is often found to be constitutively (persistently) expressed in many types of cancers [42]. This leads to the unregulated expression of many genes (presumably through DNA over-replication), some of which are involved in cell proliferation and result in cancer formation [44]. The gene $RAD52$ is involved in double-stranded break repair and plays a central role in genetic recombination and DNA repair. Experiments by Schilddraut et al. provide evidence for an association between several genes in the DNA repair and response pathways and risk of invasive serous ovarian cancer [45]. In addition to genes with strong support associations, the study is also supportive of associations between three SNPs in $RAD52$ and invasive serous ovarian cancer. More importantly, $RAD52$ in the current module is not only highly exclusive but also correlates with $MYC$ and $CCNE1$.

The second module consists of $KRAS$ and $PPP2R2A$ with $p_1 < 0.001$ and $p_2 = 0.05$, covering 77 samples. As an Oncogene, $KRAS$ is an important signal transducer involved in the regulation of various cellular responses during cell proliferation, differentiation, and survival. Mutations in $KRAS$ frequently occur in cancer cells such as specific ovarian cancer subtypes [46] and indicate poor prognosis and increased resistance to some cancer therapies [47]. The protein encoded by $PPP2R2A$ is also implicated in the negative regulation of cell growth and division, and is associated with a variety of regulatory subunits. Although $PPP2R2A$ has not been directly implicated in tumorigenesis, several findings suggest that deregulation of $CHEK2$ and/or $PPP2R2A$ has pathogenic
effects in at least a subset of germ cell tumors in childhood teratoma [48].

**Analysis using only somatic mutations and CNVs**
To further investigate if gene expression provides useful information for the identification of mutated core modules in cancer, we analyzed data only from somatic mutations and CNVs. In this case only one significant module is detected for each dataset. In GBM, the module contains three genes \(CDKN2B\), \(PTEN\) and \(TP53\). Compared to the module containing \(PTEN\), \(TP53\) and \(MTAP\) identified using all three kinds of data, the current module has lower exclusivity between several pairs of genes (Figure 5). In the ovarian carcinoma data the module consists of \(CCNE1\) and \(MYC\). Interestingly, \(RAD52\) is not detected, although it has a high correlation and very high exclusivity both with \(CCNE1\) and \(MYC\) (Figure 4). All these indicate that gene expression is helpful for the identification of biologically mutated core modules in cancer.

**Integration of somatic mutations and gene expressions data**
Recently, core modules were detected in the GBM data without using CNV information [49]. In this case three modules were identified which are significantly mutually exclusive (Table 1). A slightly different strategy was adopted in [49] for data integration: more weight is given to somatic mutations than gene expressions (i.e., \(k = 2\) was used in the integrated model); a smaller threshold is selected to detect a bigger subnetwork in the optimization algorithm; and the statistical test for mutual exclusivity is also slightly different.

Several oncogenes or tumor suppressors such as \(PTEN\), \(TP53\), \(EGFR\) and \(NF1\) were also detected when these two kinds of data are used for GBM. These four genes are involved in the RTK/RAS/PI(3)K signaling pathway, which is one of the core pathways altered in the development of glioblastoma and was deduced by the TCGA Research Network [2]. It should be noted that because of the lack of CNV data, several genes including \(CDKN2A\), \(CDKN2B\), \(CDK4\) and \(MDM2\) were not identified.

**Conclusions**
In this paper, iMCMC is presented to integrate somatic mutations, CNVs and gene expressions to detect mutated core modules in cancer. Unlike previous approaches exploring pathways or modules, iMCMC does not use any prior information such as human PPI networks and known pathways. We apply iMCMC to the GBM and ovarian carcinoma datasets and identified five and two mutated modules respectively. Many of the detected genes have been reported to be implicated in carcinogenesis and some modules are involved in known pathways. For example, \(CDKN2B\) and \(CDK4\) as well as \(EGFR\) and \(NF1\) are involved in the RB and RTK signaling pathways, respectively and the \(CCNE1\), \(MYC\) and \(RAD52\) module in ovarian carcinoma is involved in cell cycle.

For further improvement of the integrative network \(\mathcal{M}\) and the optimization algorithm in iMCMC, two
parameters, i.e., $k$ and $\lambda$, should be further explored. A typical feature of our method is to employ two parameters to balance not only different sources of data but also the vertices and edges of the weighted network. This provides flexibility for using the method because one can choose different parameters to emphasize on specific factors. It should be noted that different choices for parameters may result in slightly different results. This needs further consideration in practice based on actual data.

For comparison, we tested the proposed method on datasets integrating only two kinds of data, i.e., somatic mutations and CNVs or expressions. The results indicate that gene expressions or CNVs indeed provide extra useful information to the original data for the identification of mutated core modules in cancer.

In conclusion, our findings provide several candidate core modules recurrently perturbed in GBM or ovarian carcinoma for further studies. Our integrative method, iMCMC, will be a helpful complementary tool in the identification of cancer pathways and as a general methodology with practical significance, it has a potential to be employed in cancer research.

**Materials and methods**

**Data sets**
The GBM and ovarian carcinoma data were downloaded from TCGA website (https://tcga-data.nci.nih.gov/tcga/) in December, 2011. We used three kinds of data: somatic mutations, DNA copy number variations (CNVs), and gene expressions. We considered only the data from level 3. The GBM dataset contains CNVs in 1269 genes spanning 169 glioblastoma samples, gene expression profiles in 11861 genes in 202 samples and nucleotide sequence aberrations in 343 genes in 135 samples. For the ovarian carcinoma dataset, these three kinds of data are in 966 genes in 559 samples, 11864 genes in 489 samples and 8431 genes in 320 samples, respectively. All these data are primary materials required to construct the integrative network for further analysis. First, a matrix $A_0$ is generated by identifying common samples with somatic mutations and CNVs and merging their genes over the common samples. $A_0$ is binary: if any mutation occurs in a given gene in a particular sample or if the given gene is in a statistically significant variation region of the particular sample, which is determined by GISTIC [50], then the mutation is assigned the number 1; if these criteria are not met then 0 is assigned. A mutation matrix $A$ is then obtained by reducing the size of $A_0$ by combining genes that are mutated in the same samples into larger ‘metagenes’. An expression matrix $B$ is obtained by using the method described previously [37]. $B$ is a real matrix with each of its entries representing relative expression of a given gene in a particular sample. For all these matrices, rows and columns correspond to samples and genes, respectively.

| Module | Gene | $p$-value for exclusivity | Related to GBM | Reference |
|--------|------|--------------------------|----------------|----------|
| 1      | EGFR, NF1, PTEN, PIK3R1, TP53 | 0.01 | all the genes | [49] and the references wherein |
| 2      | COL6A2, DST, ERBB2, PIK3CA, RB1 | <0.001 | all except DST | same as above |
| 3      | PRA M, SYNE1 | <0.001 | both genes | same as above |

**Figure 5** A comparison of the GBM data with or without gene expressions. (A) When all three kinds of data are used, the identified module has high exclusivity and relative correlation between TP53 and MTAP. (B) When gene expressions are not used the detected module is not as exclusive as before.
The main idea of this study is to integrate three kinds of data resources described above via a network framework and identify mutated core modules in cancer by an optimization model and the following statistical tests. The preliminary version of this method was recently proposed in [49] for GBM somatic mutation and gene expression integrative analysis. For the completeness of this paper we describe the approach with some improvements in the following.

Construction of an integrative network $\mathcal{M}$

With the above data, we constructed an integrative network based on which an optimization model can be built to detect oncogenic modules and pathways. The construction procedure contains three steps.

**The network based on gene expression**

In this step a network based on gene expression called Expression Network (denoted by $\mathcal{EN}$) is constructed. $\mathcal{EN}$ is weighted both for its edges and vertices, where each vertex denotes a gene, and each edge is the correlation between expressions of two vertices (genes). Weight of each vertex reflects the extent of the influence of a gene mutation on the expression of other genes.

We notice that genes in $A$ and $B$ may be different and so the common genes are identified first. Let $(G_1, S_1)$ and $(G_2, S_2)$ be the sets of genes and samples contained in the two matrices, respectively. $G_0$ and $S$ are set as $G_0 = G_1 \cap G_2$ 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 as $n_{i}^{(1)}$ and $n_{i}^{(2)}$, respectively. Moreover, based on the elements in $A$ and $B$ we set $e_{i}^{(1)} = \{a_{ki} : a_{ki} = 1, k \in S\}$ and $e_{i}^{(2)} = \{a_{ki} : a_{ki} = 0, k \in S\}$, and so a mutation-correlated expression vector $e_{i} = (e_{i}^{(1)}, e_{i}^{(2)})$ can be obtained. 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. A prerequisite for this procedure is a minimum number of 2 for samples in the two groups. Therefore, the vertex set of the expression network $\mathcal{EN}$ is $G$ 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 $\mathcal{EN}$ can be defined as:

$$ f_{i} = 1 - 1/d \sum_{r=1}^{d} p_{ir} $$

where $d$ is the number of genes in $G_2$, and $p_{ir}$ is the $p$-value of differential expression of gene $r$ relative to $i$’s mutation status. This means that smaller the $p$-values stronger the influence of a gene mutation on others. That is, it is more likely to be a driver that should be given greater weights.

For any two genes $i$ and $j$ in $G$, the edge weight $w_{ij}$ is defined as the absolute value of Pearson correlation between $e_{i}$ and $e_{j}$ among the samples in $S$. Note that corresponding to metagene, weights of the vertex and edge in the expression network are obtained from averages of the values of related genes.

**The network based on somatic mutations and CNVs**

Based on the mutation matrix $A$ generated from somatic mutations and CNVs, a Mutation Network (MN) can be constructed. To hold the same vertex set as in the expression network $\mathcal{EN}$, the same gene set $G$ is used to construct MN. For any gene $i \in G$, $m_{i}$ denotes the number of mutations in $i$ across the samples in the mutation matrix $A$, i.e., $m_{i} = \sum_{k} a_{ki}$. 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 $u_{ij}$ is defined as the number of samples in which exactly one of the pair is mutated divided by the number of samples in which at least one of the pair is mutated in $A$. The vertex weight is a measure of mutation coverage and the edge weight is a measure of mutual exclusivity.

**The integrative network**

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

We observed that in $\mathcal{EN}$ or $\mathcal{MN}$ the vertex and edge weights have different measurement levels. To balance these two terms, we defined $f = \max f_{i}$ and $u = \max u_{ij}$ in $\mathcal{EN}$ and similarly, $h = \max h_{i}$ and $v = \max v_{ij}$ in $\mathcal{MN}$. We 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 $\mathcal{EN}$, respectively (similarly, $H = \{h_{i}\}$ and $V = \{v_{ij}\}$ in $\mathcal{MN}$). Then $U$ and $\xi F$ (similarly, $V$ and $\eta H$) have balanced values.

While integrating the two networks more importance can be given to $\mathcal{MN}$ than $\mathcal{EN}$ when gene expression values are considered to contain noises. Thus a parameter $k$ is introduced to reflect the relative importance of $\mathcal{MN}$ relative to $\mathcal{EN}$. Set $\delta = (u/v) = k$, then $\delta = k/(u/v)$. In this paper $k = 1$ is used.

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

$$ w_{ij} = \delta \cdot u_{ij} + v_{ij}, \quad c_{i} = \delta \xi \cdot f_{i} + \eta \cdot h_{i}, $$

where $n$ is the number of genes in $G$. From the above discussion it is clear that $\xi$ and $\eta$ can be directly
determined by the EN and MN networks, which is also similar for \( \delta \) once \( k \) is preassigned.

An optimization model for detecting coherent subnetworks

For the integrative network \( \mathcal{M}_t \) our goal is to extract some modules (subnetworks) with high weights in both edges and vertices. We used the previously reported optimization model [51] for this purpose. With \( w_{ij} \) and \( c_i \) defined as in Eq. (1), the model is as follows:

\[
\begin{align*}
\max & \quad \sum_i \sum_j w_{ij} x_{ij} + \lambda c_i x_i, \\
\text{s.t.} & \quad x_{ij}^\delta + x_{jk}^\delta + \cdots + x_{kn}^\delta = 1, \\
& \quad x_i \geq 0, i = 1, \ldots, n,
\end{align*}
\]

where the \( n \)-dimensional non-negative vector \( x = (x_1, x_2, \ldots, x_n) \), determined by solving the optimization model, represents the degree of each vertex that belongs to a 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 vertices and the subnetwork. In the model, a positive parameter \( \lambda \) is introduced to balance these two terms.

On the other hand, a trivial solution will be obtained when model (2) is unconstrained where all vertices from the original network can be included into the subnetwork, so a regularization constraint should be introduced to limit the number of vertices selected. This is the role of \( \beta \) which can adjust the strength of regularization applied to the variable \( x = (x_1, x_2, \ldots, x_n) \). \( \beta = 2 \) is an attractive option in many cases since the optimization of a quadratic function over a sphere is polynomially solvable in contrast to general non-convex programming [52] but tends to select all vertices in the network to the final subnetwork. The L1-type constraint when \( \beta = 1 \), leads to a sparse solution, i.e., many of the entries in the final optimal solution \( x \) will be zeros [53]. In general, we use \( \beta = 1 \) in model (2) to extract small-sized subnetworks from a larger network.

The optimization model (2) can be easily solved by quickly finding a local maximum from a predetermined initial solution using the following iterative algorithm [51]:

\[
x_i^{t+1} = \left( x_i^t + \frac{2(WX)}{2X^T WX + \lambda \sum_i c_i x_i^t} \right)^{1/2},
\]

where \( W = (w_{ij}) \) is the \( n \times n \) edge weight matrix, and \( X = (x_1^t, x_2^t, \ldots, x_n^t)^T \) 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 the cutoff, 0.1 is used in this study) 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 another locally optimal solution and its corresponding subnetwork based on the new network.

Significance test of the subnetwork (module)

We performed a random test to assess the significance of the results. For a selected subnetwork \( SN \) with \( b \) vertices, we obtained a quantity \( C \) which is the sum of all vertex weights and edge weights involved in \( SN \). Then we randomly selected \( b \) vertices from the original network and obtained a similar quantity \( CR \). This procedure is repeated 1,000 times and the number \( r \) of CRs which is larger than \( C \) can be calculated. The significant \( p \)-value of \( SN \) (denoted as \( p_1 \)) can be obtained from the quantity of \( r \) divided by 1,000.

Mutual exclusivity test of the subnetwork (module)

After a subnetwork passes the significance test, the following step is performed to evaluate whether it exhibits a pattern of mutually exclusive genomic alterations. For this we used the ‘switching permutation’ method proposed by Cirriello et al. [17], which adopts a Markov chain Monte Carlo permutation strategy based on random network generation models.

Furthermore, although a subnetwork \( SN \) with \( b \) (\( b > 2 \)) vertices is not significantly mutually exclusive, we cannot exclude the possibility that one of its subsets is. In this case we can 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 choosing the paired vertices with the smallest exclusivity and removing one vertex with the smaller entry value \( x \) in the solution of (3). This process is repeated until either of the two conditions is reached: \( SN' \) is significantly mutually exclusive or \( b = 2 \). In this study, \( p_2 \) denotes the exclusivity \( p \)-value for concise description.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

JZ and SZ conceived this project. JZ carried out the experiment and data analysis. JZ, SZ and YW carried out the biological analyses and wrote the manuscript. KXS supervised this project. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful to Prof. Edwin Wang and Prof. Raul Rabadan for their helpful discussion. This work was supported by the National Natural Science Foundation of China, No. 11001256, 11131009, and 61171007, the ‘Special Presidential Prize’ - Scientific Research Foundation of the CAS and the Foundation for Members of Youth Innovation Promotion Association, CAS. A preliminary version of this paper was published in the proceedings of IEEE ISB2012.
Declarations

The publication of this article is from the National Natural Science Foundation of China, No. 11001256, 11131009, and 61171007, the ‘Special Presidential Proje’ Scientific Research Foundation of the CAS and the Foundation for Members of Youth Innovation Promotion Association, CAS. This article has been published as part of BMC Systems Biology Volume 7 Supplement 2, 2013: Selected articles from The 6th International Conference of Computational Biology. The full contents of the supplement are available online at http://www.biomedcentral.com/bmcsystbiol/supplements/7/S2.

Published: 14 October 2013

References

1. Zhang S, Liu CC, Li W, Shen H, Laid P, Zhou XJ. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. Nucleic Acids Res 2012, 40:9379-9391.
2. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 2008, 455:1061-1068.
3. International cancer genome consortium: International network of cancer genome projects. Nature 2010, 464:993-998.
4. Forbes S, Clements J, Dawson E, Barnford S, Webb T, Dogan A, Flanagan A, Teague J, Wooster R, Futreal PA, Stratton MR. COSMIC. 2005. Br J Cancer 2006, 94:318-322.
5. Meyerson M, Gabriel S, Getz G. Advances in understanding cancer genomes through second-generation sequencing. Nat Rev Genet 2010, 11:685-696.
6. Greenman C, Stephens P, Smith R, Dalgliesh GL, Hunter C, Bignell G, Davies H, Teague J, Butler A, Stevens C, Edkins S, O’Meara S, Vastrick I, Schmidt EE, Avis T, Barthorpe S, Bhamra G, Buck G, Choudhury B, Clements J, Cole J, Dicks E, Forbes S, Gray K, Halliday K, Harrison R, Hills K, Hinton J, Jenkinson A, Jones D, et al. Patterns of somatic mutation in human cancer genomes. Nature 2007, 446:153-158.
7. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009, 458:719-724.
8. Ding L, Getz G, Wheeler DA, Mardis ER, McKellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, Fulton L, Fulton RS, Zhang Q, Wendl MC, Lawrence MS, Larson DE, Chen K, Dooling DJ, Sabo A, Hawes AC, Shen H, Jhangiani SN, Lewis LR, Hall O, Zhu Y, Mathew T, Ren Y, Yao J, Scherer SE, Clerc K, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature 2008, 455:1069-1075.
9. Jones S, Zhang X, Parsons DW, Lin JC, Leary RJ, Angenendt P, Mankoo P, Carter H, Kamihara Y, Jimeno A, Hong SM, Fu L, Lin MT, Calhoun ES, Kamiyama M, Walter K, Nikolaou T, Nikolsky Y, Hargan J, Smith DR, Hidalgo M, Leach SD, Klein AP, Jaffee EM, Goggins M, Maitra A, Iacobuzio-Donahue C, Eshleman JR, Kern SE, Hruban RH, et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. Science 2008, 321:1801-1806.
10. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004, 10:789-799.
11. Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell 2011, 144:646-664.
12. Jonsson PF, Bates PA. Global topological features of cancer proteins in the human interactome. Bioinformatics 2006, 22:2291-2297.
13. Qiu YQ, Zhang S, Zhang XS, Chen L. Detecting disease associated modules and prioritizing active genes based on high throughput data. BMC Bioinformatics 2010, 11:26.
14. Baca SM, Kinzler KW, Velculescu VE, Vogelstein B, Parmigiani G. Patient-oriented gene set analysis for cancer mutation data. Genome Biol 2010, 11:R112.
15. Efroni S, Ben-Hamo R, Edmonson M, Greenblum S, Schaefer CF, Buetow KH: Detecting cancer gene networks characterized by recurrent genomic alterations in a population. PLoS ONE 2011, 6:e14437.
16. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS ONE 2010, 5:e9819.
17. Casoello G, Cerami E, Sander C, Schultz N. Mutual exclusivity analysis identifies oncogenic network modules. Genome Res 2012, 22:308-308.
18. Vandin F, Uppal E, Raphael BJ. De novo discovery of mutated driver pathways in cancer. Genome Res 2012, 22:375-385.
19. Zhao J, Zhang S, Wu L, Zhang XS. Efficient methods for identifying mutated driver pathways in cancer. Bioinformatics 2012, 28:2940-2947.
20. Miller CA, Settle SH, Sulman EP, Aksape KD, Milosavljevic A. Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. BMC Med Genomics 2011, 4:34.
21. Segal E, Wang H, Koller D. Discovering molecular pathways from protein interaction and gene expression data. Bioinformatics 2003, 19(Suppl 1):i264-272.
22. Akavia UD, Levin O, Kim J, Sanchez-Garcia F, Kotlar D, Causton HC, Pochanard P, Mozes E, Getaway LA, Peter D. An integrated approach to uncover drivers of cancer. Cell 2010, 143:1005-1017.
23. Masica DL, Karchin R. Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer Res 2011, 71:4550-4561.
24. Roco JW, Sridransky D: p16 (MTS1/CDKN2/INK4a) in cancer progression. Exp Cell Res 2001, 264:62-55.
25. Liu W, Lv G, Li Y, Li L, Wang B. Downregulation of CDKN2A and suppression of cyclin D1 gene expressions in malignant gliomas. J Exp Clin Cancer Res 2011, 30:76.
26. Gracia E, Fischer U, Elkahloun A, Trent JM, Meese E, Meltzer PS. Isolation of genes amplified in human cancers by microdissection mediated cDNA capture. Hum Mol Genet 1996, 5:595-600.
27. Cerami E, Demir E, Schultz N, Taylor BS, Sander C. Automated network analysis identifies core pathways in glioblastoma. PLoS ONE 2010, 5:e9819.
28. Feng J, Kim ST, Liu W, Kim JW, Zhang Z, Zhu Y, Berens M, Sun J, Xu J. An integrated analysis of germline and somatic genetic and epigenetic alterations at 9p21.3 in glioblastoma. Cancer 2012, 118:232-240.
29. Nishi K, Tomaso Edi, Ng HK, Pang JC, Rossell MF, Hjelm NM. Expression of p19INK4d, CDK4, CDK6 in glioblastoma multiforme. Br J Neurosurg 2000, 14:28-32.
30. Jhoo YH, Hess KR, Liu L, Linskey ME, Yung WA. Modeling diagnosis for patients with malignant astrocytic gliomas: Quantifying the expression of multiple malignant genetic and clinical variables. Neuro Oncol 2005, 7:485-494.
31. Kim H, Huang W, Jiang X, Penncooke B, Park PJ, Johnson MD. Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. Proc Natl Acad Sci 2010, 107:2183-2188.
32. Zheng H, Ying H, Yan H, Kimmelman AC, Hiller DJ, Chen J, Perry SR, Tonon G, Chu GC, Ding Z, Stommel JM, Dunn KL, Wiedemeyer R, You MI, Brennan C, Alan Wang Y, Ligon KL, Wong WH, Chen LS, DePinho RA: p53 and Pten control neural and glioma stem/progenitor cell renewal and differentiation. Nature 2008, 455:1129-1133.
33. Chow LN, Endersby R, Cui X, Rankin S, Qu C, Zhang J, Broniscer A, Ellison DW, Baker SJ. Cooperation within and among Pten, p53, and p63 pathways induces high-grade astrocytoma in adult brain. Cancer Cell 2011, 19:305-316.
34. Schmid M, Men S, Rosenbach MD, Carrera CJ, Friedman H, Carson DA. A methylthioadenosine phosphorylase (MTAP) fusion transcript identifies a new gene on chromosome 9p21 that is frequently deleted in cancer. Oncogene 2000, 19:5747-5754.
35. Nusinovitch M, Kerr AR, Kang A, Meehan P, Chen A, Fragoso H, Widenreuther G, McDonald B, Mahdi M, et al. A methylthioadenosine phosphorylase (MTAP) fusion transcript identifies a new gene on chromosome 9p21 that is frequently deleted in cancer. Oncogene 2000, 19:5747-5754.
36. Huang HY, Li SH, Yu SC, Chou FF, Tseng CC, Hu TH, Uen YH, Tsan YF, Wang YH, Fang FM, Huang WW, Wei YC, Wu JM, Li CF. Homozygous deletion of MTAP gene as a poor prognosticator in gastrointestinal stromal tumors. Clin Cancer Res 2009, 15:6963-6972.
37. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Willerson MD, Ryan R, Miller C, Ding L, Golub T, Mesirov JP, Alexe G, Lawrence M, O’Kelly M, Tamayo P, Wehr BA, Gabrie S, Winckler W, Gupta S, Jakkula L, Fuller F, Nordenberg B, Christensen J, David James C, Sarkaria JN, Brennan C, Kahn A, Spellman PT, Wilson RK, Speed TP, Gray JW, Meyerson M, et al. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in P53/PTEN, IDH1, EGFR, and NF1. Cancer Cell 2010, 17:98-110.
38. Iwayama T, Lozano G. MDM2, an introduction. Mol Cancer Res 2003, 1:993-1000.
39. Biernat W, Kleihues P, Yonekawa Y, Ohgaki H. Amplification and overexpression of MDM2 in primary (de novo) glioblastomas. J Neuropath Exp Neurol 1997, 56:180-185.
40. Oda Y. Choline acetyltransferase: the structure, distribution and pathologic changes in the central nervous system. Pathol Int 1999, 49:921-937.
41. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature 2011, 474:609-615.
42. Engler DA, Gupta S, Growdon WB, Drapkin RI, Nitta M, Sergent PA, Alfred SF, Gross J, Deavers MT, Kuo WL, Karlan BY, Rueda BR, Orsulic S, Gershenson DM, Brer M, Gray JW, Mahapatra G. Genome wide DNA copy number analysis of serous type ovarian carcinomas identifies genetic markers predictive of clinical outcome. PloS ONE 2012, 7:e30996.
43. Nakayama N, Nakayama K, Shamima Y, Ishikawa M, Katagiri A, Iida K, Miyazaki K. Gene amplification CCNE1 is related to poor survival and potential therapeutic target in ovarian cancer. Cancer 2010, 116:2621-2634.
44. Saucier L, Whitfield J, Martins CP, Finch AJ, Murphy DJ, Sodir NM, Kaneznis AN, Brown Swigart L, Nas N, Ean G. Modelling Myc inhibition as a cancer therapy. Nature 2008, 455:679-683.
45. Schildkraut JM, Iversen ES, Wilson MA, Clyde MA, Moorman PG, Palmeni RT, Whittaker R, Bentley RC, Marks JR, Berchuck A. Association between DNA damage response and repair genes and risk of invasive serous ovarian cancer. PloS ONE 2010, 5:e10065.
46. Riccardelia C, Oehler MK. Diverse molecular pathways in ovarian cancer and their clinical significance. Mutat Res 2009, 62:270-275.
47. Jančík S, Drábek J, Radzioch D, Hajdúch M. Clinical relevance of KRAS in human cancers. J Biomed Biotechnol 2010, 2010:159601.
48. Jin Y, Mertens F, Kullendorff CM, Panagopoulos I. Fusion of the tumor-suppressor gene CHEK2 and the gene for the regulatory subunit B of protein phosphatase 2 PPP2R2A in childhood teratoma. Neoplasia 2006, 8:413-418.
49. Zhang J, Zhang S, Wang Y, Zhao J, Zhang XS. Identifying mutated core modules in glioblastoma by integrative network analysis. Proceedings of IEEE 6th International Conference on Systems Biology 2012, 304-309.
50. Beroukhim R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S, Du J, Kau T, Thomas RK, Shah K, Soto H, Perna S, Prinsen J, Debiase RM, Demichelis F, Hatton C, Rubin MA, Garraway LA, Nelson SF, Liu L, Mischel PS, Meyerson M, Golub TG, Lander ES, Mellinghoff IK, et al. Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. Proc Natl Acad Sci 2007, 104:20007-20012.
51. Wang Y, Xia Y. Condition specific subnetwork identification using an optimization model. Proceedings of the 2nd International Symposium on Optimization and Systems Biology 2008, 333-340.
52. Ye Y. A new complexity result on minimization of a quadratic function with a sphere constraint. In Recent advances in global optimization. Volume 1. NJ: Princeton University Press;Floudas CA, Pardalos PM 1992:19-31.
53. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Series B 1996, 58:267-288.

Cite this article as: Zhang et al. Identification of mutated core cancer modules by integrating somatic mutation, copy number variation, and gene expression data. BMC Systems Biology 2013, 7(Suppl 2):S4.