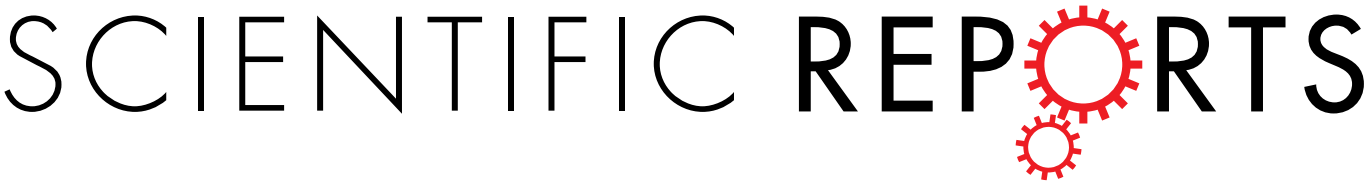

OPEN

Driver pattern identification over the gene co-expression of drug response in ovarian cancer by integrating high throughput genomics data

Received: 5 July 2017
Accepted: 9 November 2017
Published online: 23 November 2017

Xinguo Lu[1], Jibo Lu[1], Bo Liao[1], Xing Li[1], Xin Qian[1] & Keqin Li[1,2]

Multiple types of high throughput genomics data create a potential opportunity to identify driver patterns in ovarian cancer, which will acquire some novel and clinical biomarkers for appropriate diagnosis and treatment to cancer patients. To identify candidate driver genes and the corresponding driving patterns for resistant and sensitive tumors from the heterogeneous data, we combined gene co-expression modules with mutation modulators and proposed the method to identify driver patterns. Firstly, co-expression network analysis is applied to explore gene modules for gene expression profiles through weighted correlation network analysis (WGCNA). Secondly, mutation matrix is generated by integrating the CNV data and somatic mutation data, and a mutation network is constructed from the mutation matrix. Thirdly, candidate modulators are selected from significant genes by clustering vertexs of the mutation network. Finally, a regression tree model is utilized for module network learning, in which the obtained gene modules and candidate modulators are trained for the driving pattern identification and modulators regulatory exploration. Many identified candidate modulators are known to be involved in biological meaningful processes associated with ovarian cancer, such as CCL11, CCL16, CCL18, CCL23, CCL8, CCL5, APOB, BRCA1, SLC18A1, FGF22, GADD45B, GNA15, GNA11, and so on.

Ovarian cancer is known as a complex genomic disease. During tumorigenesis, many factors contribute to pathological gene expression changes, such as genomic affection and epigenomic affection. Hence, acquiring the etiopathogenesis and chemoresponse of cancer faces a great challenge. And the achievement can be utilized to deploy the high-performance diagnosis and treatment of cancer patients. Through many studies presented the prognosis of ovarian cancer, there is no completely validated clinical model for predicting ovarian cancer prognosis and drug response. Therefore, it remains an important research issue to identify prognostic and predictive driver patterns for improving ovarian cancer treatment.

Cancer genomes posses a large number of aberrations including somatic mutations and copy number variations (CNVs). Many mutations contribute to cancer progression from the normal to the malignant state. The researches have shown that some aberrations are vital for tumorigenesis and most cancers are caused by a small number of driver mutations developed over the course of about two decades[1,2]. The detection of these mutations with exceptionally high association between the copy number variations, somatic mutations and gene expression can ascertain disease candidate genes and potential cancer mechanisms. Several large scale cancer genomics projects, such as the Genomic Data Commons (GDC), The Cancer Genome Atlas (TCGA), and International Cancer Genome Consortium (ICGC), etc., have produced a large volume of data and provided us an opportunity to integrate different level of gene expression data to identify candidate driver genes and driver pathways and better understand the cancer at the molecular level.

[1]College of Computer Science and Electronic Engineering, Hunan University, Lushan Nan Rd., Changsha, 410082, China. [2]Department of Computer Science, State University of New York, New Paltz, NY, 12561, USA. Correspondence and requests for materials should be addressed to X.L. (email: hnluxinguo@hnu.edu.cn)

During exploring the large volume of genomics abnormalities generated from large-scale cancer projects, many computational and statistical methods have been proposed to search for this mutation driver patterns. Adib *et al*. predicted the potential specificity of several putative biomarkers by using gene expression microarray profile and examined their expression in a panel of epithelial tissues and tumors[3]. Youn *et al*. proposed a new method to identify cancer driver genes by accounting for the functional impact of mutations on proteins[4]. Tomasetti *et al*. combined conventional epidemiologic studies and genome-wide sequencing data to infer a number of driver mutations which is essential for cancer development[5]. Jang *et al*. identified differentially expressed proteins involved in stomach cancer carcinogenesis through analyzing comparative proteomes between characteristic alterations of human stomach adenocarcinoma tissue and paired surrounding normal tissue[6]. Xiong *et al*. proposed a general framework in which the feature (gene) selection was incorporated into pattern recognition to identify biomarkers[7]. Jung *et al*. identified novel biomarkers of cancer by combining bioinformatics analysis on gene expression data and validation experiments using patient samples and explored the potential connection between these markers and the established oncogenes[8]. Logsdon *et al*. identified a new category of candidate tumor drivers in cancer genome evolution which can be regarded as the selected expression regulators (SERs)–genes driving deregulated transcriptional programs in cancer evolution, and uncovered a previously unknown connection between cancer expression variation and driver events by using a novel sparse regression technique[9]. Naif *et al*. integrated exome sequencing data with functional RNAi screening data onto a human signaling network to predict breast cancer subtype-specific drug targets and survival outcome[10,11]. Some integrative network approaches on cancer omic data were developed to understand the function of genomic alterations in cancer[12,13]. A cancer genome embodies thousands of genomics abnormalities such as single nucleotide variants, large segment variations, structural aberrations and somatic mutations[14,15]. The corresponding malignant state reflects aberrant copy numbers (CN) and/or mutations and the expression patterns in which the mutation and copy number patterns are embeded[16]. Integration of copy number/mutation data and gene expression data was proposed to discover driver genes by quantifying the impact of these aberrations on the transcriptional changes[17,18]. The available implementations for the integrative analysis of genomics data includes regression methods, correlation methods and module network methods[19].

Driver patterns, including driver genes, driver mutations, driver pathways and core modules, which are considered as cancer biomarkers, are supposed to promote the cancer progression. Thus, identification of driver patterns is vital for providing insights into carcinogenic mechanism. The integration of gene expression, copy number and somatic mutations data to identifying genomics alterations which induce changes in the expression levels of the associated genes, becomes a common task in cancer mechanism and drug response studies. However, there is still a tough work to integrate information across the different heterogeneous omics data and distinguish driver patterns which can promote the cancer cell to propagate infinitely. Hence, we proposed a driver pattern identification method over the gene co-expression of drug response in ovarian by integrating high throughput genomics data. In the method, we integrated different level genomics data including gene expression, copy number and somatic mutation to identify drivers of resistance and sensitivity to anti-cancer drugs. First, weighted correlation network analysis is applied to acquire the co-expression gene modules. These modules are utilized as initial inputs for the following module networks learning. Then, mutation network is constructed from the mutation matrix generated by integrating the CNVs and somatic mutations. The candidate modulators are selected from the clusters of the vertexes of the mutation network. Finally, an optimization model is used to identify the driving pattern and explore the modulator regulatory. We used publicly available and clinically annotated gene expression, CNVs, and somatic mutation datasets from TCGA. The experimental results show that in the driver patterns many of the identified candidate modulators are known to be involved in biological meaningful processes associated with ovarian cancer, which can be regard as potential driver genes.

Results

We conducted the driver patterns identification and the driving regulatory analysis as following procedure. Firstly, co-expression network were constructed and the gene modules for gene expression profiles were explored. Then, CNVs and somatic mutations were integrated to build the mutation network. The candidate modulators were selected from clusters of the vertex of network. A regulation procedure was explored for candidate modulators and gene modules using network learning combined with regression tree model. At last, local polynomial regression fitting model was applied to conduct dosage-sensitive and dosage-resistant analysis.

Data preprocessing. To select a subset of genes whose aberration/expression profiles were significantly different between sensitive and resistant tumors, we applied differential analysis to the gene expression and CNV data between these two groups. For different gene expression analysis, we used the EMD approach: EMDomics, which is designed especially for heterogeneous data. We selected genes with a $q\text{-}value\text{-} < 0.1$. For CNV data, firstly, we mapped genes to the CNV regions to obtain CNV genes; then, we used the EMD approach for the differential analysis and ranking of the CNV genes. We also calculated the frequency of amplification and deletion for each gene in the two groups and selected genes for which the difference between their frequencies is more than 20%. We used a threshold of log2 copy number ratio of 0.3/−0.3 to call amplified/deleted genes. For somatic mutation data, we also calculated the frequency of mutations for all genes across the samples in each groups and selected the genes that were mutated in more than 2% of the tumors.

Gene modules with co-expression patterns. After the data preprocessing, 2690 genes are chosen for co-expression gene modules exploring. We constructed a weighted correlation network and identified co-expression modules, which was conducted by WGCNA, associated with cis-platinum sensitive and resistant. The gene co-expression modules and the corresponding hierarchical clustering dendrograms of these genes are shown in Fig. 1. The color bars correspond to the clusters of genes which can be seen as gene module. We

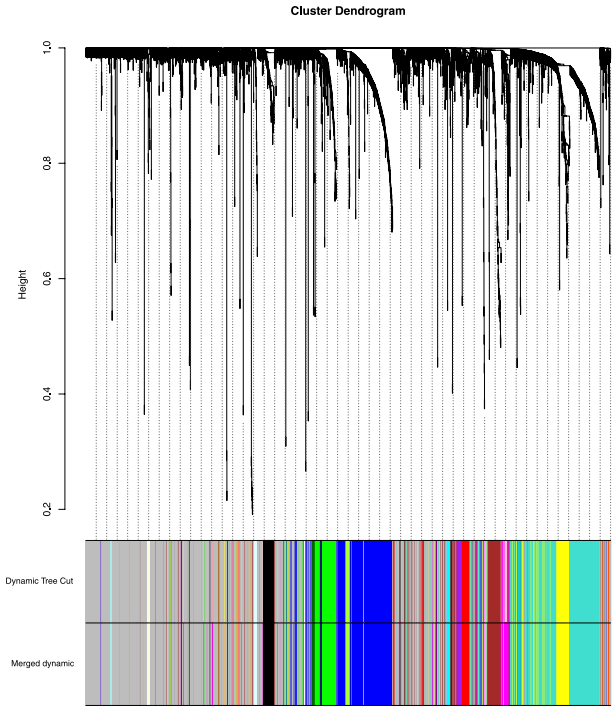

Figure 1. Gene co-expression modules for gene expression profiles. A network heatmap plot (interconnectivity plot) of a gene network together with the corresponding hierarchical clustering dendrograms, with dissimilarity based on topological overlap, together with assigned module colors.

color	turquoise	blue	brown	yellow	green	magenta	red	black	purple	greenyellow
count	357	288	145	130	115	78	62	61	29	23

Table 1. The top 10 co-expression modules.

identified thirty-nine modules are described in thirty-nine colors with turquoise, blue, brown, yellow, green and so on. The top 10 co-expression modules are shown in Table 1, The first row represents the color of the modules, and the second row represents the number of genes which contained in the corresponding module. The remaining genes are in grey. These grey genes are not clustered into any modules. The modules list is shown in Supplementary Table S1.

Candidate modulators selected after long gene filtering. There may be some genes which have mutation only because they are long. So, we filtered the frequently mutated genes by gene length. Finally we obtained 2334 genes including 1942 genes of copy number variation and 392 genes of somatic mutation.

For CNVs data, the mapping of CNVs can promote the search for causal links between genetic variation and disease susceptibility. Thus we mapped gene expression to CNVs regions to obtain CNVs genes, for a specific gene in a given sample, the value of amplified is assigned the number 1, the deleted is assigned the number -1, the normal is assigned the number 0; for somatic mutation data, we first obtained mutated genes and mapped the correlations between somatic mutations and their residing genes.

Then, a mutation matrix A is generated by integrating the CNVs and somatic mutations. The mutation matrix is binary: if any variation region in a given gene of the particular sample is in a statistically significant variation or any mutation arises in the specific gene in the given sample, the value of mutation is assigned the number 1; otherwise the value is assigned the number 0. The matrix rows and columns correspond to samples and genes, respectively. A mutation network (MN) was constructed according to the mutation matrix. For a mutated gene, m_i describes the number of mutations in gene i across the samples in the mutation matrix, i.e., $m_i = \sum_{i=1}^{s} a_{si}$, where m is the number of samples and a_{ij} is the value of mutation matrix of the gene i in the sample j. The network vertex weight is defined as $h_i = m_i/m$. Also the edge weight V_{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.

We selected a subgroup of significant genes by clustering the vertex of the mutation network. These genes can be regarded as candidate modulators (genes that regulate other genes in their module). We identified a list of 624 initial candidate modulators for the mutation genes shown in Supplementary Table S2, among which includes 40 data from somatic mutations. The gene expression heatmap of the top 50 candidate modulators is shown in Fig. 2.

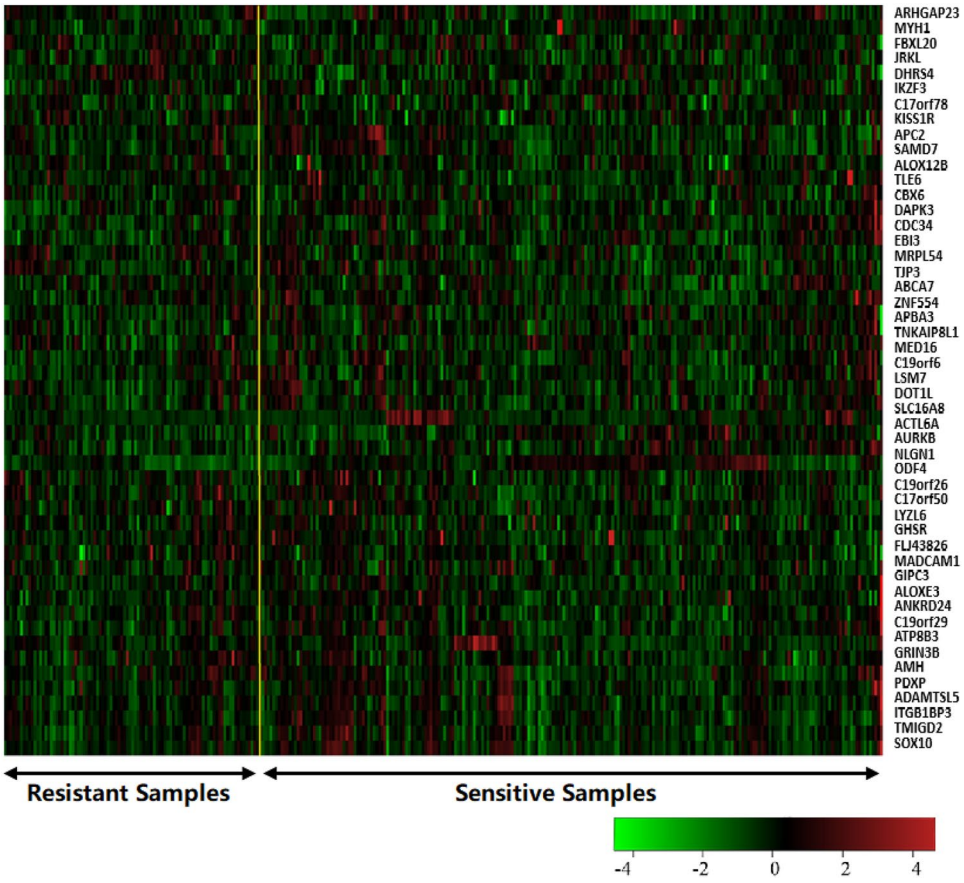

Figure 2. The gene expression heatmap of the top 50 candidate modulators.

Dosage-sensitive genes and dosage-resistant genes and driver patterns. We run the module networks learning using the gene modules and candidate modulators obtained from the above steps. It generally assumed that the modulators most likely regulate the expression of the genes in the corresponding gene modules. Because the gene modulators would have distortions in a amount of samples and they are likely to be the roots of the regression (drive) trees, the modulator gene are generally treat as a driver gene. Through the module network learning we obtained 456 modules and 93 modulators (see Supplementary Table S3). 5 out of these modulators are from somatic mutations. Some of modules and modulators identified by our method were also detected by previous studies[20,21]. Magically, several modules shared the same modulators, in which only one modulator is derived from somatic mutations.

We applied a local polynomial regression fitting model with the R package loess to the scatter diagram for 93 regulatory genes with 324 samples. *DS* threshold is set to (−0.25, 0.25) as the filtering criteria for dosage-sensitive genes(DSGs) and the others were selected as dosage-resistant genes(DRGs). Therefore, we obtained 44 DSGs and 49 DRGs (see Supplementary Fig. S4). The LRG1, HYDIN, POLRMT and TNFSF10 gene data scatter diagram for CNVs or somatic mutations vs. gene expression with local polynomial regression fitting model are shown in Fig. 3, in which HYDIN is derived from somatic mutations. Each red point represents an independent sample, the black line is the linear regression result, and the blue curve is the local polynomial regression fitting result. In Fig. 3A and B the slope of the black lines are close to 0, so the dosage sensitivity score are nearly equal to 0. In Fig. 3C and D the slope's absolute value of the black line are above 1, the blue curves are mostly overlapped with the black line, so the dosage sensitivity score are above 1. It can be concluded from the graph that the genes for Fig. 3A and B are dosage-resistant genes, whereas the genes for Fig. 3C and D are dosage-sensitive genes.

We identified three major driver patterns in which the top 3 gene modules with the largest number of regulatory genes (modulators) are explored. Fig. 4 shows these driver patterns with gene modulators and the corresponding gene modules, gene modulators shown in the left side with blue background ovals are the dosage-resistant genes. Gene modulators are in the right side are dosage-sensitive genes. In the first driver pattern with color pale blue, there are 15 genes in gene co-expression module 3 and 13 gene modulators. In the second driver pattern with color green, there are 7 genes in module 30 and 18 gene modulators. In third pattern with color red, 6 genes are in module 31 and 14 gene modulators are for regulatory. Remarkably, it shows that CCL11 regulate all the 3 modules; SLC18A1 and C9orf82 regulate the module 3 and module 30; PLXDC1, CCL16, CDC34 and HYDIN regulate the module 3 and module 31; ZFHX4 regulates the module 30 and module 31. Hence, these important regulatory genes regulate many gene modules, and these gene modulators are highly correlated with ovarian cancer[22–28].

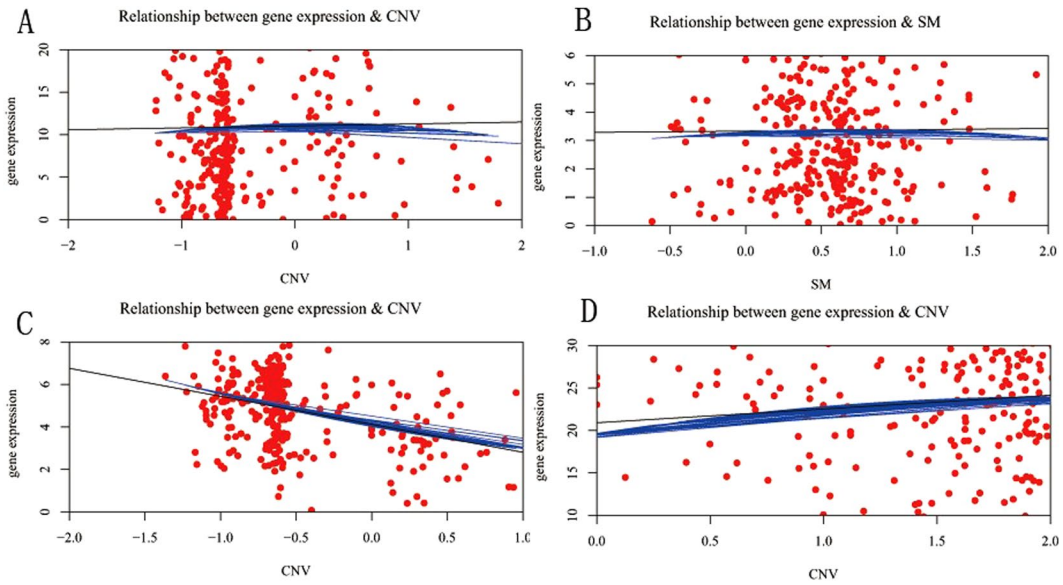

Figure 3. Scatter diagram with local polynomial regression fitting model. Each red point represents an independent sample. The black line is the linear regression, and the blue curve is the LOESS regression. (**A**) LRG1. (**B**) HYDIN. (**C**) POLRMT. (**D**) TNFSF10.

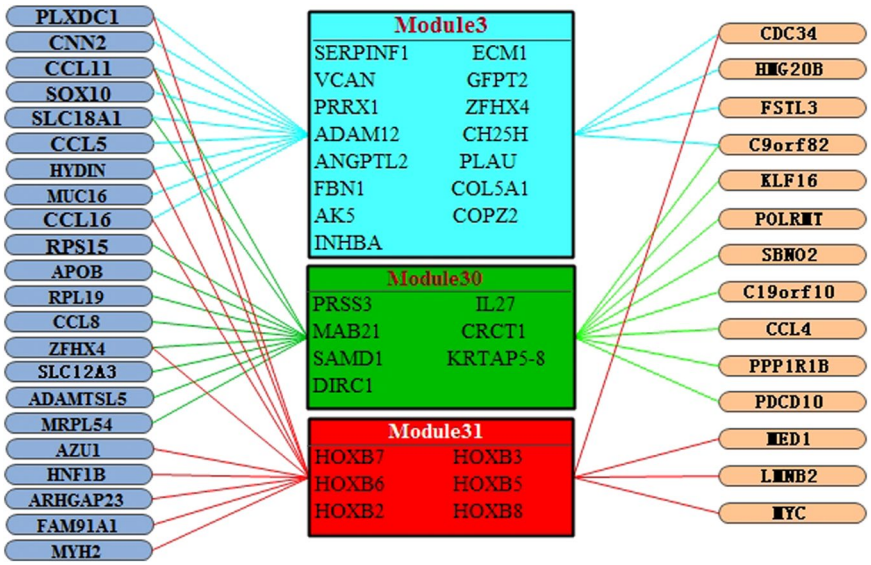

Figure 4. The 3 major driver patterns for dosage-sensitive and dosage-resistant. The graph depicts inferred regulators (left; ovals with blue background represents the DRGs, and ovals with orange background represents the DSGs) and their corresponding regulated gene modules (right).

In the above-mentioned analysis, we constructed gene co-expression modules using WGCNA. WGCNA is a well-established method for identifying significant genes from transcriptomic data. A gene significance(GS) measure is assigned to each gene. The higher the absolute value of GS_i, the more biologically significant is the i th gene. Gene significance of 0 indicates that the gene is not significant with regard to the biological question of interest. We applied WGCNA to select biologically and clinically significant genes with $GS_i\text{-}value > 0.7$. The result is compared with the genes identified by the proposed method. We found that there are 17 common genes including CCL5,CCL8,CCL4,CCL23,CCDC135 and so on. The comparison is shown in the Fig. 5. From the figure, we can see that 6 common genes are in the intersection between the key genes selected from WGCNA and the driver genes of dosage-resistant from the proposed method,and 11 common genes are in the intersection between the key genes selected from WGCNA and the driver genes of dosage-sensitive from the proposed method.

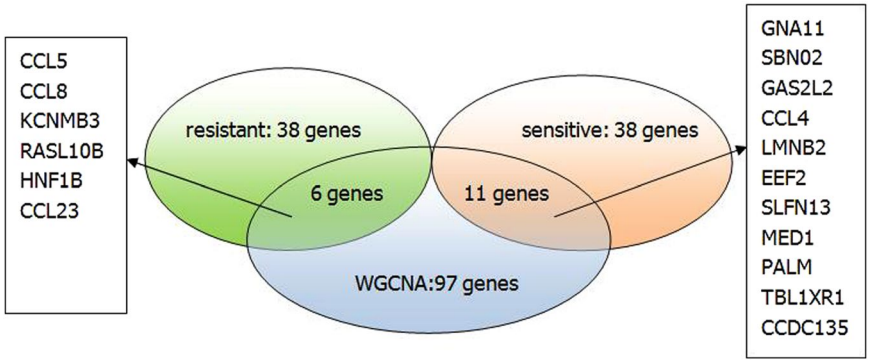

Figure 5. The comparison of significant genes between WGCNA and the proposed method.

GO-BP Term	pvalue	qvalue	gene
lymphocyte chemotaxis	3.43E-09	2.12E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
monocyte chemotaxis	4.76E-09	2.12E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
mononuclear cell migration	1.95E-08	5.80E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
chemokine-mediated signaling pathway	3.4E-08	5.87E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
lymphocyte migration	3.67E-08	5.87E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
neutrophil chemotaxis	3.95E-08	5.87E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
cellular response to interleukin-1	4.91E-08	6.25E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
neutrophil migration	6.05E-08	6.25E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
myeloid leukocyte migration	6.31E-08	6.25E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5/AZU1
granulocyte chemotaxis	1.09E-07	9.41E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
cellular response to tumor necrosis factor	1.16E-07	9.41E-06	APOB/CCL11/CCL16/CCL18/CCL23/CCL8CCL5/BRCA1
granulocyte migration	1.84E-07	1.36E-05	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5

Table 2. The top 12 enriched GO terms for DRGs.

GO-BP Term	pvalue	qvalue	gene
dopamine receptor signaling pathway	2.38E-06	0.001887	PALM/GNA15/GNA11/KLF16
skeletal muscle contraction	0.000127	0.032678	EEF2/TCAP/MYH8
multicellular organismal movement	0.000239	0.032678	EEF2/TCAP/MYH8
musculoskeletal movement	0.000239	0.032678	EEF2/TCAP/MYH8
trabecula morphogenesis	0.000255	0.032678	MED1/SBNO2/UBE4B
DNA-templated transcription, initiation	0.000289	0.032678	MED1/MED16/POLR2E
			MED24/POLRMT
phospholipase C-activating dopamine receptor signaling pathway	0.000289	0.032678	GNA15/GNA11
natural killer cell chemotaxis	0.000352	0.03489	CCL7/CCL4
negative regulation of keratinocyte proliferation	0.000498	0.038042	MED1/KDF1
androgen receptor signaling pathway	0.000544	0.038042	MED1/MED16/MED24
cardiac muscle tissue morphogenesis	0.000544	0.038042	MED1/TCAP/UBE4B
response to radiation	0.000691	0.038042	PPP1R1B/CCL7/GNA11/MYCECT2/UBE4B

Table 3. The top 12 enriched GO terms for DSGs.

Enrichment analysis for driver patterns. We applied the Gene Ontology (GO) category to expanded the results of DRGs and DSGs with the R package clusterProfiler, in which the significance threshold are set as p value < 0.05 and q-value < 0.05 separately. The top 12 enriched GO terms for DRGs and DSGs are presented in Tables 2 and 3, respectively. The DRGs are mainly enriched in receptor binding, signaling pathway, cell migration, cellular response, cell chemotaxis, activator activity, etc. These processes are highly relevant to the trait of tumor and contribute to the chief progression of tumor. And the DSGs are mainly enriched in receptor binding, signaling pathway, template transcription, cell proliferation, cell differentiation, tissue development, cell division, etc. Compared with DRGs, the GO annotated to DSGs are mostly management functions and basic functions of the cell and tissue. Figure 6 shows the top 12 biological process annotation analysis results on GO terms for DRGs and

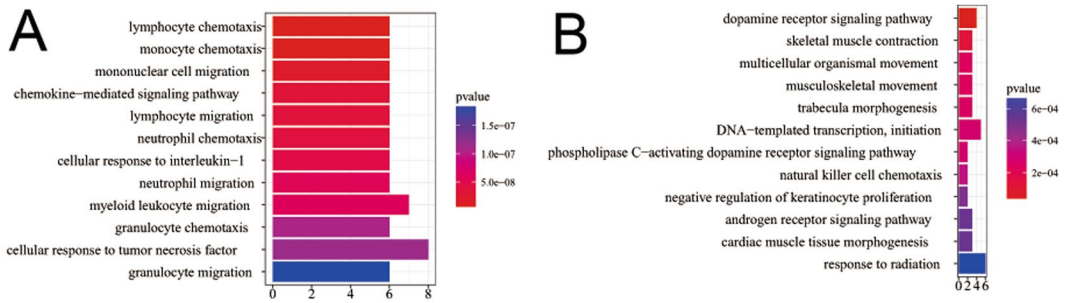

Figure 6. The analysis of GO biological process annotation analysis for DSGs and DRGs. (**A**) The GO biological process enrichment of DRGs. (**B**) The GO biological process enrichment of DSGs.

pathway	pvalue	gene
Chemokine signaling pathway	4.80E-06	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
Cytokine-cytokine receptor interaction	4.58E-05	CCL11/CCL16/CCL18/CCL23/CCL8/CCL5
Parkinson's disease	0.052537	NDUFB5/SLC18A1
Breast cancer	0.053866	BRCA1/FGF22
Other types of O-glycan biosynthesis	0.056251	MFNG
Ribosome	0.060691	RPS15/RPL19
Vitamin digestion and absorption	0.061213	APOB
Maturity onset diabetes of the young	0.066151	HNF1B
NOD-like receptor signaling pathway	0.070724	MFN1/CCL5
Phototransduction	0.071064	RCVRN
Asthma	0.078387	CCL11
Prion diseases	0.088067	CCL5

Table 4. The top 12 important pathways for DRGs.

DSGs, respectively. The length of the bar represents the number of selected genes annotated onto this GO term, and the colors of the bar represents the significance levels of enrichment. The GO enrichment analysis results for DRGs and DSGs are in Supplementary Table S5 and Supplementary Table S6, respectively. From GO biological process enrichment analysis, we achieved CCL11, CCL16, CCL18, CCL23, CCL8, CCL5, APOB, BRCA1 play extremely significant role in DRGs, and EEF2, GNA11, GNA15, MED1, MYH8 and TCAP act very important role in DSGs. CCL11, CCL16, CCL18, CCL23, CCL8 and CCL5 have effective chemical attraction on eosinophils catalysing their accumulation at allergic inflammation sites. CCL11 has been shown to negatively regulate neurogenesis with certain human tumors[22,29]. CCL16 is a chemokine prominently expressed in the liver, but also in ovarian and breast cancer[25,30]. CCL18 is associated with some human cancer types including ovarian cancer[31,32]. CCL23 plays a role in subclinical systemic inflammation and associated with atherogenesis[33]. CCL8 is strengthened in stromal fibroblasts at the tumor border and in tissues at which breast cancer cells incline to metastasize such as the lungs and the brain[34,35]. CCL5 is a chemokine that boosts cancer progression by arousing and adjusting the inflammatory diseases, which sequently reconstruct the cancer microenvironment. Moreover, CCL5 also boosts metastasis in ovarian and breast cancer cells[35,36]. APOB mutation is associated with some human cancer types such as steatosis, liver, hypocholesterolemia[37,38]. BRCA1 mutation confers high risks of ovarian and breast cancer, encodes a tumor suppressor[39]. EEF2 is a key components for protein synthesis, and ubiquitously expressed in normal cells[40]. GNA11 is associated with Congenital Hemangioma[41]. GNA15 is associated with inhibition of proliferation, activation of apoptosis and differential effects[42]. MED1 is associated with some biological process such as prostate cancer cell growth[43]. MYH8 is a key components for muscle development and regeneration[44]. TCAP is involved with energy regulation and metabolism, and is implicated in the regulation of stress-related behaviors[45].

Regulatory pathway analysis for driver patterns. To further elucidate the biological pathway of gene modulators in the driver pattern, we performed the Kyoto Encyclopedia of Genes and Genomes (KEGG)[46–48], pathway enrichment analysis of the results of DSGs and DRGs list with the R package clusterProfiler separately. The top 12 pathways for DRGs and DSGs are presented respectively in Tables 4 and 5. For DRGs, the most significant canonical pathways are mainly bound up with Chemokine signaling, Cytokine – cytokine receptor interaction, Parkinson's disease, Breast cancer, onset diabetes, Asthma, Prion diseases, etc. And the pathways for DSGs are major related to Thyroid hormone signaling, Apoptosis, Cytosolic DNA-sensing, Amoebiasis, Chagas disease, Cell cycle, WNT signaling, etc. Remarkablely for DRGs, Cytokine – cytokine receptor interaction is embedded in the Chemokine signaling pathway. For these two pathways, gene CCL11, CCL16, CCL18, CCL23, CCL8 and CCL5 have a key function (see Fig. 7). Chemokines play a role in the migration of many cells during development and are critical to nervous system development[49]. Cytokine interactions play an important role in

pathway	pvalue	gene
Thyroid hormone signaling pathway	0.000438	MED1/MED16/MED24/MYC
Apoptosis	0.009077	GADD45B/TNFSF10/LMNB2
Cytosolic DNA-sensing pathway	0.017271	POLR2E/CCL4
NF-kappa B signaling pathway	0.036048	GADD45B/CCL4
Amoebiasis	0.036744	GNA15/GNA11
Chagas disease (American trypanosomiasis)	0.041034	GNA15/GNA11
Cholinergic synapse	0.048583	PIK3R5/GNA11
Cytokine-cytokine receptor interaction	0.05237	CCL7/CCL4/TNFSF10
Cell cycle	0.058258	GADD45B/MYC
FoxO signaling pathway	0.065054	GADD45B/TNFSF10
Ubiquitin mediated proteolysis	0.069434	CDC34/UBE4B
Wnt signaling pathway	0.074816	MYC/TBL1XR1

Table 5. The top 12 important pathways for DSGs.

health and are vital to many cancer during immunological and inflammatory responses in disease. It can lead to antagonist, additive, or synergistic activities in keeping physiological functions such as body temperature, feeding and somnus, as well as in anorectic, ardent fever, and sleepiness neurological representations of acute and chronic disease[50]. Parkinson's disease is a pathway specifically associated to diseases of the central nervous system, NDUFB5 and SLC18A1 are enriched in this pathway (see Fig. 8). NDUFB5 and SLC18A1 are associated with multiple human diseases[23,51]. Breast cancer is the major cause of cancer death among the female worldwide, gene BRCA1 and FGF22 are enriched in this pathway (see Fig. 9). FGF22 is required for brain development and associated with hereditary and neoplastic disease[52]. For DRGs, the thyroid hormones are key regulators of metabolism, growth and other body systems, MED1, MED16, MED24 and MYC are play a role in this pathway (see Fig. 10). MYC is involved in cell cycle progression, apoptosis, and cellular transformation and is related to many tumors[53]. Apoptosis is an evolutionarily conserved signaling pathway, which plays a fundamental role in regulating cell number and eliminating damaged or redundant cells[54], GADD45B, TNFSF10, LMNB2 are enriched in this pathway (see Fig. 11). TNFSF10 is associated with some human cancer types including ovarian cancer[55]. LMNB2 acts as regulators of cell proliferation and differentiation, regards as a cancer risk biomarker in several cancer subtypes[56]. Cell cycle progression is completed through a reproducible sequence of events. It is a significant pathway associated with many diseases including ovarian cancer[57]. GADD45B and MYC are enriched in this pathway (see Fig. 12). GADD45B is implicated in some responses to cell injury including cell cycle checkpoints, apoptosis and DNA repair[58]. WNT signaling pathway is required for basic developmental processes and play an important role in human stem cells and cancers[59]. MYC and TBL1XR1 are enriched in this pathway (see Fig. 13). The KEGG pathway enrichment analysis result for DRGs and DSGs are shown in Supplementary Table S7 and Supplementary Table S8, respectively.

Discussion

The main purpose of this research was to distinguish the candidate driver genes and the corresponding driving mechanism for resistant and sensitive tumor from the heterogeneous data. We presented a machine-learning approach to integrate somatic mutations, CNVs and gene expression profiles to distinguish interactions and regulations for dosage-sensitive and dosage-resistant genes of ovarian cancer. We developed a general framework for integrating high dimensional heterogeneous omics data and can potentially lead to new insight into many related studies at the systems level. In the framework, weighted correlation network analysis (WGCNA) was applied to the co-expression network analysis; the mutation network was constructed by integrating the CNVs and somatic mutations and the initial candidate modulators was selected from the clustering vertexs of network; the regression tree model was utilized for module networks learning in which the obtained gene modules and candidate modulators were trained for the modulators regulatory mechanism; finally, a local polynomial regression fitting model was applied to identify dosage-sensitive and dosage-resistant driver patterns. From the Gene Ontology and pathway enrichment analysis, we obtained some biologically meaningful gene modulators, such as CCL11, CCL16, CCL18, CCL23, CCL8, CCL5, APOB, BRCA1, SLC18A1, FGF22, GADD45B, GNA15, GNA11, and so on, which can be conducive to appropriate diagnosis and treatment to cancer patients.

Methods

We integrated gene expression, CNV, and somatic mutation data to identify the driver patterns for resistant and sensitive tumors. The overview of the overall integrative analysis is shown in Fig. 14. As can be seen in Fig. 14, the procedure takes these datasets as inputs and then produces a short list of genes as candidate divers. The method contains the following steps: (1) extract a pool of candidate modulators as initial driver genes, which have significant CNVs/Somatic mutations in ovarian cancer samples and achieve the initial gene modules from the gene expression by weighted correlation network analysis (WGCNA); (2) construct the heterogeneous network by module networks learning and evaluate the regulatory mechanism between these different omics data by liner regression model; (3) identify driver patterns and modulators interactions in the constructed regulatory networks using gene ontology and pathway enrichment analysis.

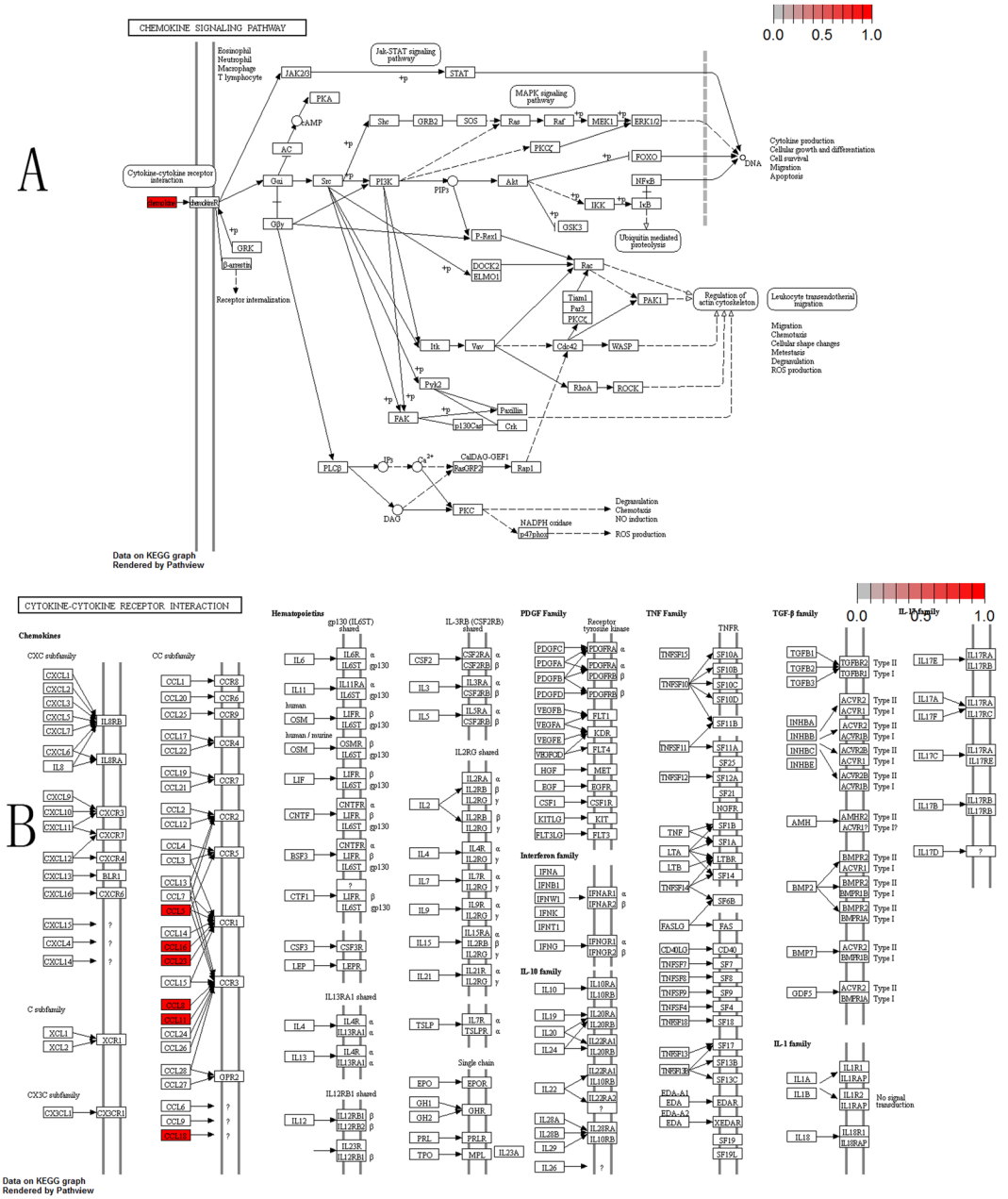

Figure 7. The six genes are enriched in the top 2 significant pathways(red, up-regulated). (**A**) Chemokine signaling pathway. (**B**) Cytokine-cytokine receptor interaction.

Data availability. We downloaded gene expression data from the Agilent 244 K Custom Gene Expression platform, CNV data from the Affymatrix Genome-Wide Human SNP Array 6.0 platform, and somatic mutation data from whole exome sequencing data obtained using Illumina Genome Analyzer DNA Sequencing. The clinical data of these patients are analyzed to determine cis-platinum chemotherapy response samples and these tumor samples are classified into resistant and sensitive groups, sensitive tumors have a platinum-free space of six months or more after the last primary treatment, no sign of residue or relapse, and the follow-up will at least six months later; and resistant tumors are recurred within six months after the last treatment[60]. According to the above definition, we identified 93 platinum-resistant and 231 platinum-sensitive primary tumors from the TCGA website, for which the gene expression profiles, CNV, and somatic mutation data were available as well. All the dataset used in this paper can be downloaded from Broad GDAC Firehose website (http://gdac.broadinstitute.org/runs/analyses_2014_10_17/data/OV/20141017).

Filtering mutation genes by gene length. The length of genes in human are very different and so the mutation probabilities of different genes are in vast difference. There may be some genes which have mutation only because they are long. In order to filter those long but non-qualified genes (genes which have mutations only because they are long yet they are not driver mutations) we adopted the filtering strategies of VarWalker[61]

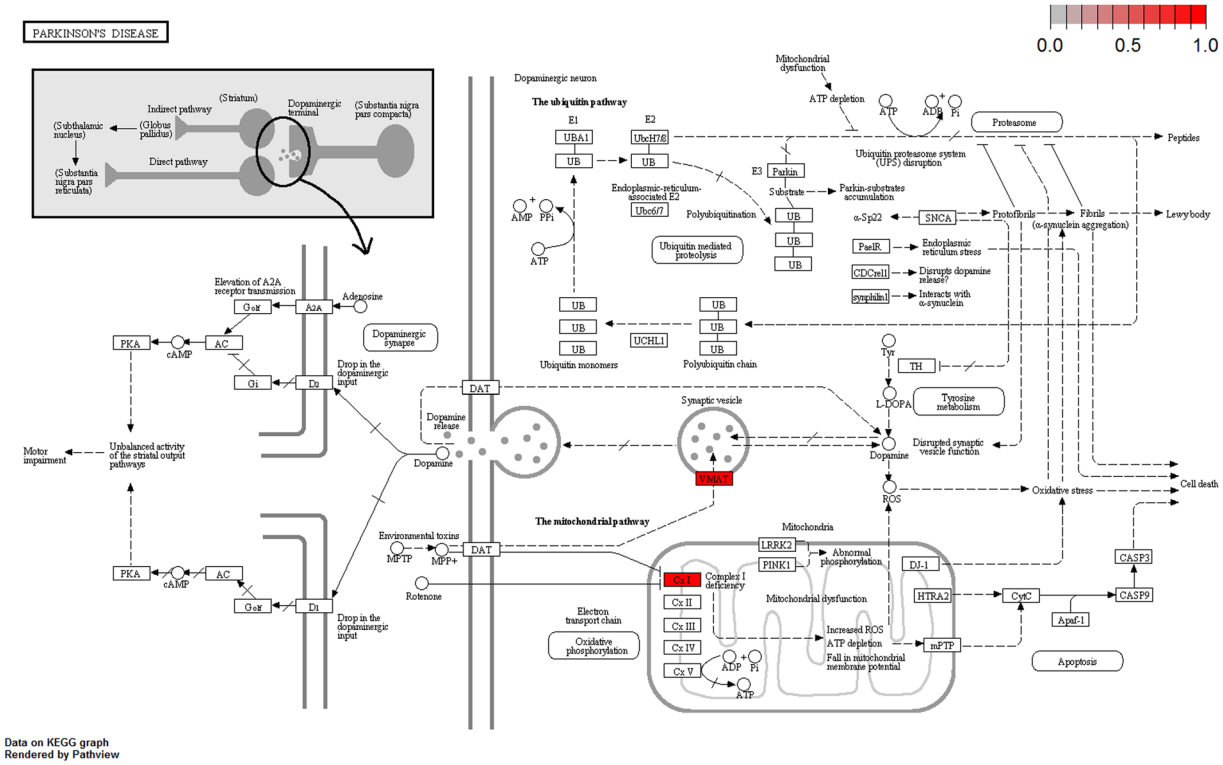

Figure 8. Parkinson disease pathway(red, up-regulated). NDUFB5 and SLC18A1 are enriched in this pathway. NDUFB5 is a gene alias of the complex 1 deficiency(Cx1) and SLC18A1 is a gene alias of the vesicular amine transporter(VMAT).

and computed a probability weight vector (PWV) by formulating a generalized additive model and estimated a relative mutation rate for each gene. The vector of X represents the gene length of cDNA. We applied the following model to estimate the mutated probability based on the cDNA length of genes,

$$\log it(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \sim f(X) \tag{1}$$

where $\pi = \frac{\#\ mutation\ \ genes}{\#\ CCDS\ \ genes}$ represents the proportion of mutant genes in the investigated samples. And the $f(\cdot)$ represents an unspecified smooth function. The function is then solved using a monotonic cubic spline with six knots. Through the above process of fitting, each gene was assigned a relative weight which would be used to select the important genes in the next resampling procedure.

Then a resampling test was applied to random gene sets for each samples. In the random selection process, the probability of each gene to be selected is based on the probability weight calculated in the above procedure. The weight resampling process was performed 1000 times in each sample. The mutation frequency was computed for each gene using following formula:

$$fre = \frac{\#(selecting\ \ the\ \ gene\ \ in\ \ resampling\ \ test)}{1000} \tag{2}$$

where $\#(selecting\ \ the\ \ gene\ \ in\ \ resampling\ \ test)$ represents the times and *fre* represents the frequency of the gene being selected across 1000 times in resampling test. Then we filter genes with *fre* >5% that indicates these genes may occurs at random. Those genes with *fre* >5% represents they are unlikely mutated at random. After resampling test, we obtained these significant genes as candidate modulators.

Gene co-expression modules analysis using WGCNA. Weighted correlation network analysis is a framework for co-expression analysis to explore gene modules of high correlation to external sample traits[62,63]. The gene expression correlation networks are based on quantitative measurements correlations. Assume that we are given an $n \times m$ expression matrix X where the row indices correspond to genes ($i = 1, \ldots, n$) and the column indices ($j = 1, \ldots, m$) correspond to samples:

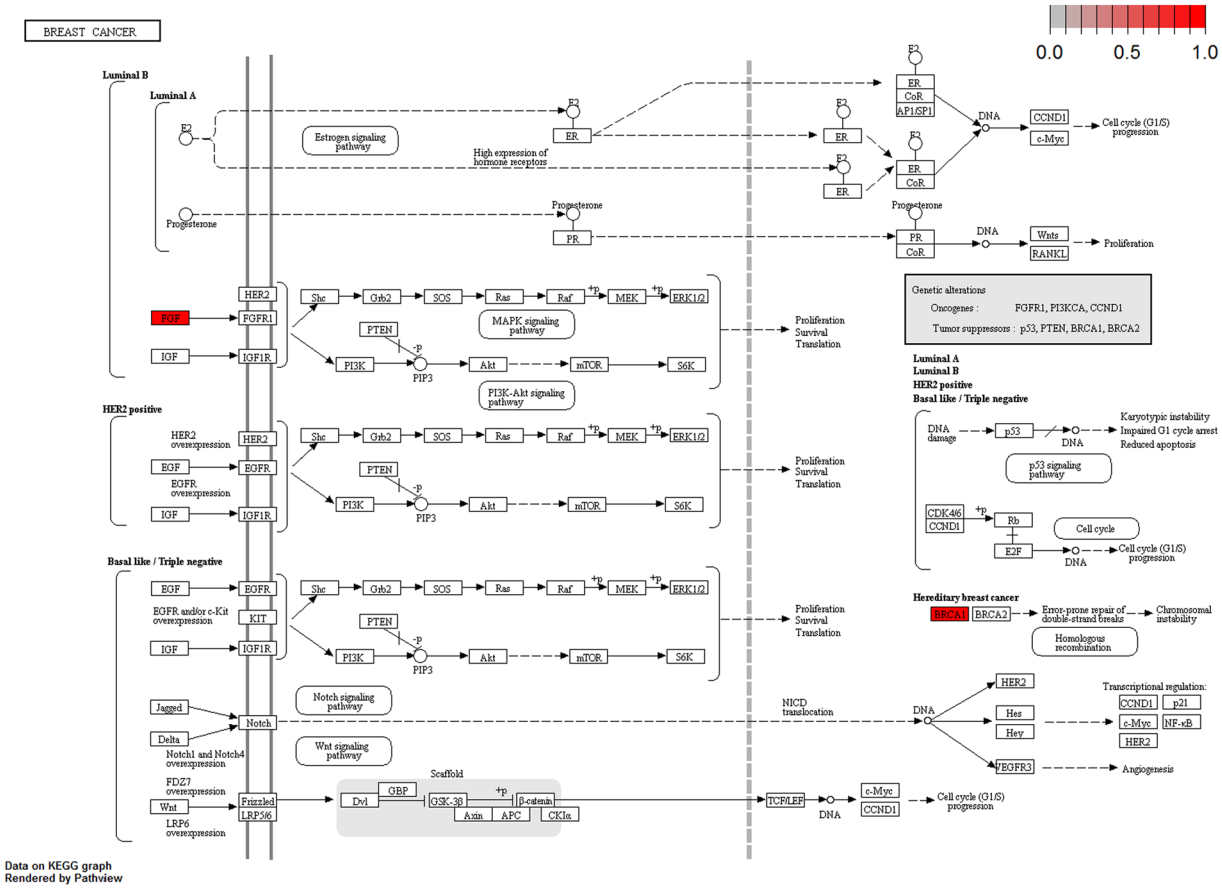

Figure 9. Breast cancer pathway(red, up-regulated). MED1, MED16, MED24 and MYC are enriched in this pathway.

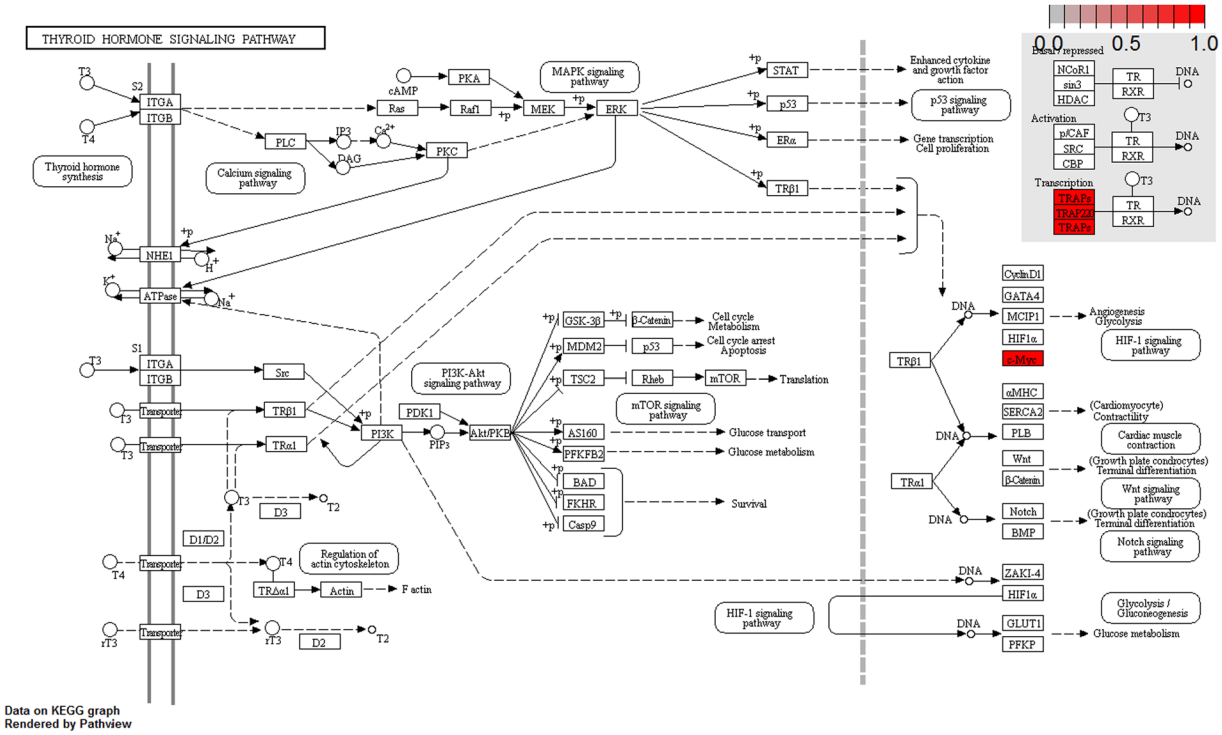

Figure 10. Thyroid hormones pathway (red, up-regulated). BRCA1 and FGF22 are enriched in this pathway.

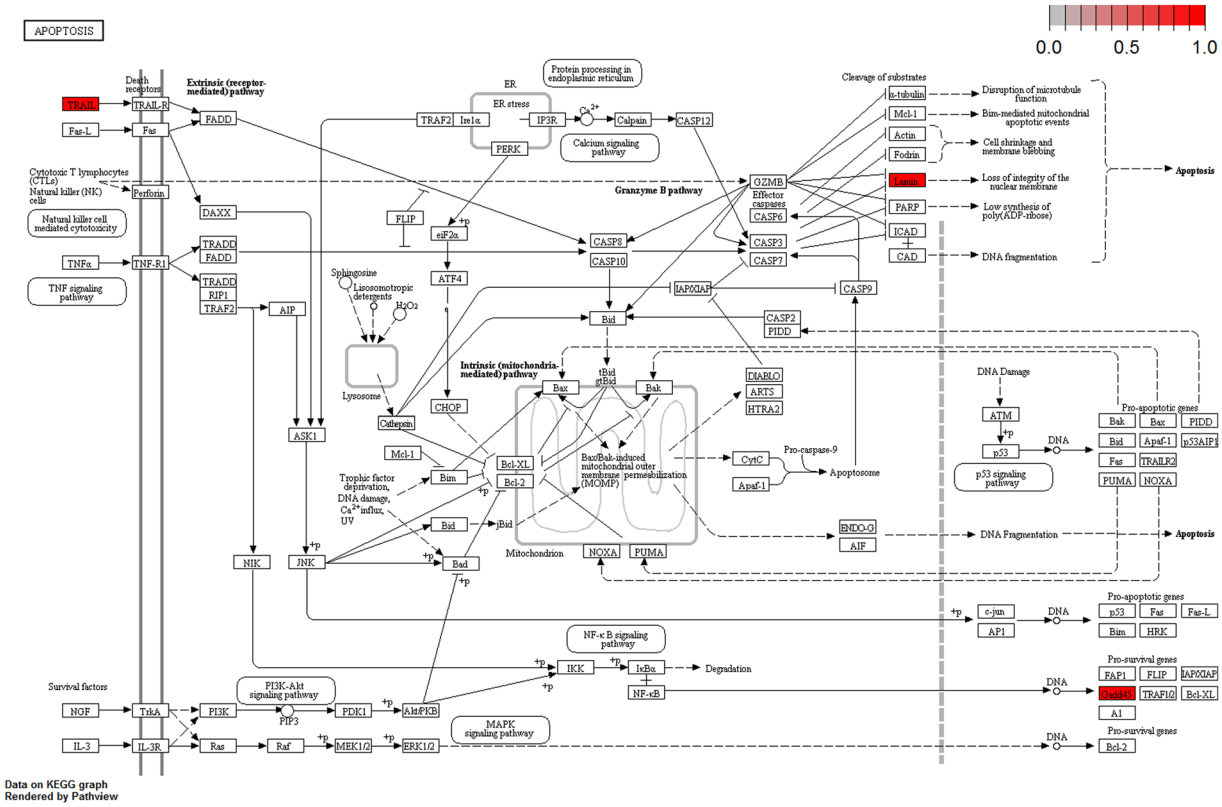

Figure 11. Apoptosis pathway(red, up-regulated). GADD45B, TNFSF10, LMNB2 are enriched in this pathway. TRAIL is a gene alias of TNFSF10, LMNB2 is a gene alias of the lamin.

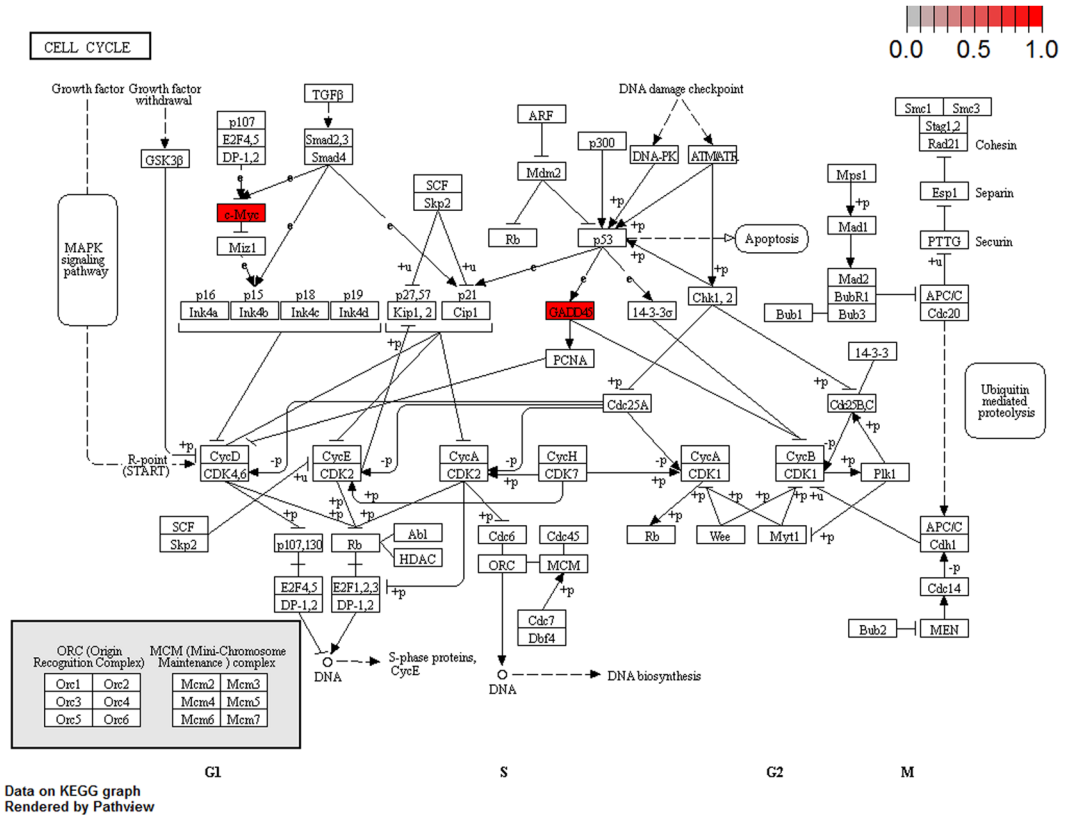

Figure 12. Cell cycle pathway(red, up-regulated). GADD45B and MYC are enriched in this pathway.

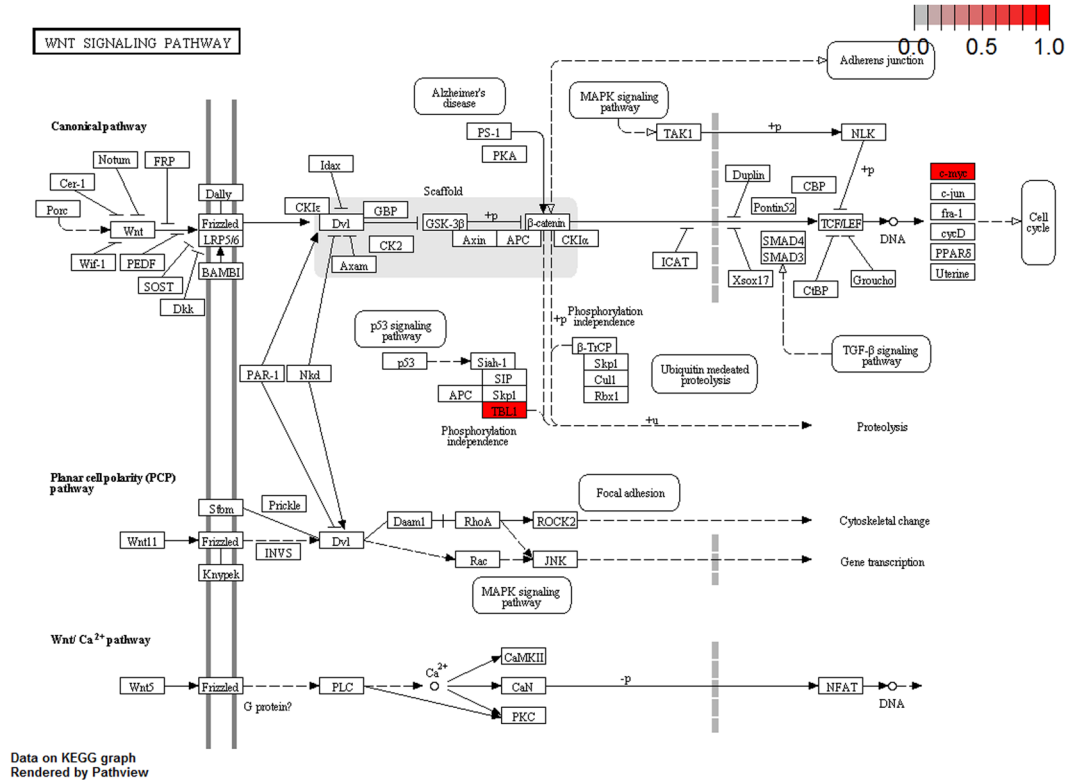

Figure 13. Wnt signaling pathway pathway(red, up-regulated). MYC and TBL1XR1 are enriched in this pathway.

$$X = \begin{pmatrix} X_{11} & X_{12} & \cdots & X_{1m} \\ X_{21} & X_{22} & \cdots & X_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nm} \end{pmatrix} \tag{3}$$

The co-expression similarity s_{ik} is applied to calculate the adjacency matrix:

$$s_{ij} = |cor(x_i, x_j)| \tag{4}$$

where s_{ij} is the absolute value of the correlation coefficient between gene i ($i = 1, \ldots, n$) and gene j ($j = 1, \ldots, n$) in the expression profiles matrix X. The value of s_{ij} is in [0, 1]. The adjacency function α_{ij} is defined with the co-expression similarity s_{ij} by soft threshold β as following,

$$\alpha_{ij} = power(s_{ij}, \beta) = s_{ij}^{\beta} \tag{5}$$

where $\beta \geq 1$. To measure the topological structure with adjacency function, topological overlap matrix (TOM) is applied to build the matrix $\Omega = [\omega_{ij}]$,

$$\begin{aligned} \omega_{ij} &= \frac{\alpha_{ij} + u_{ij}}{min(m_i, m_j) + 1 - \alpha_{ij}} \\ m_i &= \sum_{k, k \neq i} \alpha_{ik} \\ u_{ij} &= \sum_k \alpha_{ik} \alpha_{kj} \end{aligned} \tag{6}$$

in which the network structure is not only based on the two genes with direct adjacency correlation, but also on the indirect adjacency relationship.

The gene expression network is constructed based on matrix Ω. Gene modules are detected by unsupervised hierarchical clustering method. Give the q th gene co-expression module, the eigengene expression is $E^{(q)}$. The correlation coefficient for the gene x_i and the module $E^{(q)}$ can be defined as:

$$k_i^{(q)} = cor(x_i, E^{(q)}) \tag{7}$$

where $k_i^{(q)} \in [-1, 1]$.

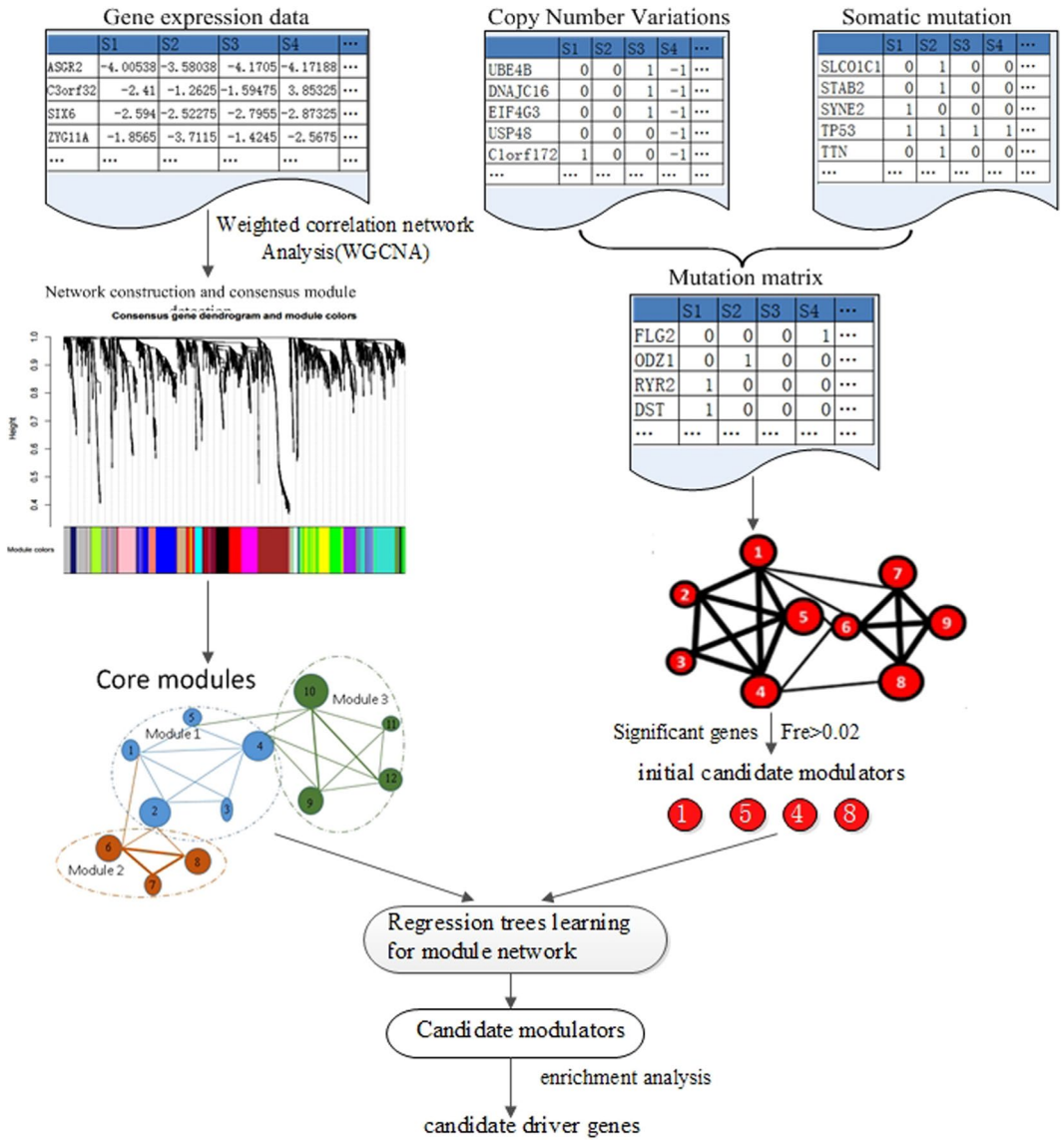

Figure 14. The scheme of driver patterns identification. (1) Acquire the initial gene modules for gene expression profiles via WGCNA network analysis (on the left) and extract a pool of candidate modulators as initial driver genes integrate CNVs and Somatic mutations in ovarian cancer samples (on the right); (2) Construct the heterogeneous network by module networks learning and evaluate the regulatory relationships by liner regression model; (3) Identify modulators interactions in the constructed regulatory networks and analyze the driver patterns through and gene ontology analysis and pathway enrichment analysis.

Modulators identification using module networks learning. From the initial gene co-expression modules, we focused on the candidate modulators' regulatory by regression tree model. Each candidate modulator is connected to a group of genes by maximizing the normal-gamma scoring function which describes the qualitative behavior of a candidate modulator that regulate the gene expression modules. A regression tree is formed with two basic blocks: decision nodes and leaf nodes. Each node corresponds to one of the selected modulators above. Our learning is an iterative process. In each iteration, a regulation procedure is found for each gene module and then each gene reassigned to the corresponding module due to prediction with the best regulatory function. We searched for the model with the highest score by using the Expectation Maximization (EM) algorithm. The scoring function defined as:

$$\log P(C, N) = \log P(C|N) + \log P(N) \quad (8)$$

where C is the candidate data and N the model structure of the network. The first term is the candidate data for a given model which meets normal gamma distribution. The second part is a penalty score on network complexity. The EM algorithm ensures the reliability of the model until convergence to a local maximum score. The essence of the algorithm consists of two steps: the first procedure is learning the best regulation program (regression tree)

for each module. The tree is constructed from the root to its leaves. The second procedure is to find the association rules for corresponding module.

Dosage-sensitive and dosage-resistant analysis. From the module networks learning, the candidate genes were generated. Then we applied a local polynomial regression fitting model to conduct dosage-sensitive and dosage-resistant analysis. In the model, a predict function was adopted to obtain n isometric points on the LOESS curve[64]. To achieve the interrelation between the copy number variation and the gene expression, a monotonicity function is defined as M:

$$M = \frac{2}{n(n-1)} \times \sum_{j=1}^{n}\sum_{i=1}^{j} S_{ij} \tag{9}$$

in which

$$S_{ij} = \begin{cases} 1, & s_j > s_i \\ 0, & s_j = s_i \\ -1, & s_j < s_i \end{cases} \tag{10}$$

and n is the number of samples. To quantize the association between gene expression and CNVs (somatic mutations), another linear fitting model was adopted and a straight line with slope of K was obtained. The slope of linear model can reflect the relationship of variables in LOSS model to some extent. Hence the dosage sensitivity (DS) score can be defined as:

$$DS = M \times |K| \tag{11}$$

Herein, for a gene, a larger DS indicates stronger dosage sensitivity.

References

1. Cheng, F., Zhao, J. & Zhao, Z. Advances in computational approaches for prioritizing driver mutations and significantly mutated genes in cancer genomes. *Briefings in Bioinformatics* **17**, 642–656 (2015).
2. Vogelstein, B. *et al*. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
3. Adib, T. R. *et al*. Predicting biomarkers for ovarian cancer using gene-expression microarrays. *British Journal of Cancer* **90**, 686–692 (2004).
4. Youn, A. & Simon, R. Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics* **27**, 175–181 (2010).
5. Tomasetti, C., Marchionni, L., Nowak, M. A., Parmigiani, G. & Vogelstein, B. Only three driver gene mutations are required for the development of lung and colorectal cancers. *Proceedings of the National Academy of Sciences* **112**, 118–123 (2014).
6. Jang, J. S. J., Cho, H. Y., Lee, Y. J., Ha, W. S. & Kim, H. W. The differential proteome profile of stomach cancer: identification of the biomarker candidates. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics* **14**, 491–499 (2004).
7. Xiong, M., Fang, X. & Zhao, J. Biomarker identification by feature wrappers. *Genome Research* **11**, 1878–1887 (2001).
8. Jung, Y. *et al*. Clinical validation of colorectal cancer biomarkers identified from bioinformatics analysis of public expression data. *Clinical Cancer Research* **17**, 700–709 (2011).
9. Logsdon, B. A. *et al*. Sparse expression bases in cancer reveal tumor drivers. *Nucleic Acids Research* **43**, 1332–1344 (2015).
10. Zaman, N. *et al*. Signaling network assessment of mutations and copy number variations predict breast cancer subtype-specific drug targets. *Cell Reports* **5**, 216–23 (2013).
11. Mcgee, S. R., Tibiche, C., Trifiro, M. & Wang, E. Network analysis reveals a signaling regulatory loop in pik3ca-mutated breast cancer predicting survival outcome. *Genomics, Proteomics & Bioinformatics* **15**, 121 (2017).
12. Wang, E. Understanding genomic alterations in cancer genomes using an integrative network approach. *Cancer Letters* **340**, 261 (2013).
13. Fu, C., Li, J. & Wang, E. Signaling network analysis of ubiquitin-mediated proteins suggests correlations between the 26s proteasome and tumor progression. *Molecular Biosystems* **5**, 1809 (2009).
14. Chan, K. C. A. *et al*. Cancer genome scanning in plasma: Detection of tumor-associated copy number aberrations, single-nucleotide variants, and tumoral heterogeneity by massively parallel sequencing. *Clinical Chemistry* **59**, 211–224 (2012).
15. Ding, J. *et al*. Systematic analysis of somatic mutations impacting gene expression in 12 tumour types. *Nature Communications* **6**, 8554 (2015).
16. Wei, P.-J., Zhang, D., Xia, J. & Zheng, C.-H. LNDriver: identifying driver genes by integrating mutation and expression data based on gene-gene interaction network. *BMC Bioinformatics* **17** (2016).
17. Zhang, J. & Zhang, S. The discovery of mutated driver pathways in cancer: Models and algorithms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* PP, 1–1 (2017).
18. Kumar, R. D., Swamidass, S. J. & Bose, R. Unsupervised detection of cancer driver mutations with parsimony-guided learning. *Nature Genetics* **48**, 1288–1294 (2016).
19. Huang, N., Shah, P. K. & Li, C. Lessons from a decade of integrating cancer copy number alterations with gene expression profiles. *Briefings in Bioinformatics* **13**, 305–316 (2011).
20. Gao, S. *et al*. Identification and construction of combinatory cancer hallmark-based gene signature sets to predict recurrence and chemotherapy benefit in stage ii colorectal cancer. *Jama Oncology* **2**, 1–9 (2015).
21. Wang, E. *et al*. Predictive genomics: A cancer hallmark network framework for predicting tumor clinical phenotypes using genome sequencing data. *Seminars in Cancer Biology* **30**, 4 (2014).
22. Nolen, B. M. & Lokshin, A. E. Targeting CCL11 in the treatment of ovarian cancer. *Expert Opinion on Therapeutic Targets* **14**, 157–167 (2010).
23. Willis, S. *et al*. Single gene prognostic biomarkers in ovarian cancer: A meta-analysis. *Plos One* **11**, e0149183 (2016).
24. Priebe, A. & Buckanovich, R. J. Ovarian tumor vasculature as a source of biomarkers for diagnosis and therapy. *Expert Review of Obstetrics & Gynecology* **3**, 65–72 (2008).
25. Manabe, S. *et al*. Expression and localization of CXCL16 and CXCR6 in ovarian endometriotic tissues. *Archives of Gynecology and Obstetrics* **284**, 1567–1572 (2011).
26. Kanska, J., Zakhour, M., Taylor-Harding, B., Karlan, B. & Wiedemeyer, W. Cyclin e as a potential therapeutic target in high grade serous ovarian cancer. *Gynecologic Oncology* **143**, 152–158 (2016).

27. Morrison, E. *et al*. Utilizing functional genomics screening to identify potentially novel drug targets in cancer cell spheroid cultures. *Journal of Visualized Experiments* (2016).
28. Chugh, S., Meza, J., Sheinin, Y. M., Ponnusamy, M. P. & Batra, S. K. Loss of n-acetylgalactosaminyltransferase 3 in poorly differentiated pancreatic cancer: augmented aggressiveness and aberrant ErbB family glycosylation. *British Journal of Cancer* **114**, 1376–1386 (2016).
29. Kuć, P. *et al*. Profiling of selected angiogenesis-related genes in serous ovarian cancer patients. *Advances in Medical Sciences* **62**, 116–120 (2017).
30. Gantsev, S. *et al*. The role of inflammatory chemokines in lymphoid neoorganogenesis in breast cancer. *Biomedicine & Pharmacotherapy* **67**, 363–366 (2013).
31. Urquidi, V. *et al*. CCL18 in a multiplex urine-based assay for the detection of bladder cancer. *Plos One* **7**, e37797 (2012).
32. Wang, Q. *et al*. CCL18 from tumor-cells promotes epithelial ovarian cancer metastasis via mTOR signaling pathway. *Molecular Carcinogenesis* **55**, 1688–1699 (2015).
33. Ignacio, R. M. C., Gibbs, C. R., Lee, E.-S. & Son, D.-S. Differential chemokine signature between human preadipocytes and adipocytes. *Immune Network* **16**, 189 (2016).
34. Farmaki, E., Chatzistamou, I., Kaza, V. & Kiaris, H. A CCL8 gradient drives breast cancer cell dissemination. *Oncogene* **35**, 6309–6318 (2016).
35. Zsiros, E., Dangaj, D., June, C. H., Kandalaft, L. E. & Coukos, G. Ovarian cancer chemokines may not be a significant barrier during whole tumor antigen dendritic-cell vaccine and adoptive t-cell immunotherapy. *OncoImmunology* **5**, e1062210 (2015).
36. Soria, G. & Ben-Baruch, A. The inflammatory chemokines CCL2 and CCL5 in breast cancer. *Cancer Letters* **267**, 271–285 (2008).
37. Cefalu, A. B. *et al*. A novel APOB mutation identified by exome sequencing cosegregates with steatosis, liver cancer, and hypocholesterolemia. *Arteriosclerosis, Thrombosis, and Vascular Biology* **33**, 2021–2025 (2013).
38. Williams, J. K. *et al*. Advanced nursing practice and research contributions to precision medicine. *Nursing Outlook* **64**, 117–123 (2016).
39. Rosenberg, S. M. *et al*. BRCA1 and BRCA2 mutation testing in young women with breast cancer. *JAMA Oncology* **2**, 730 (2016).
40. Yates, S. P., Jørgensen, R., Andersen, G. R. & Merrill, A. R. Stealth and mimicry by deadly bacterial toxins. *Trends in biochemical sciences* **31**, 123–133 (2006).
41. Ayturk, U. M. *et al*. Somatic activating mutations in GNAQ and GNA11 are associated with congenital hemangioma. *The American Journal of Human Genetics* **98**, 1271 (2016).
42. Zanini, S. *et al*. GNA15 expression in small intestinal neuroendocrine neoplasia: Functional and signalling pathway analyses. *Cellular Signalling* **27**, 899–907 (2015).
43. Liu, G. *et al*. MED1 mediates androgen receptor splice variant induced gene expression in the absence of ligand. *Oncotarget* **1**, 288–304 (2010).
44. Tonami, K. *et al*. Calpain-6 deficiency promotes skeletal muscle development and regeneration. *PLoS Genetics* **9**, e1003668 (2013).
45. D'Aquila, A. L. *et al*. Expression and actions of corticotropin-releasing factor/diuretic hormone-like peptide (CDLP) and teneurin c-terminal associated peptide (TCAP) in the vase tunicate, ciona intestinalis: Antagonism of the feeding response. *General and Comparative Endocrinology* (2016).
46. Ogata, H. *et al*. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **27**, 29–34 (2000).
47. Minoru, K., Yoko, S., Masayuki, K., Miho, F. & Mao, T. Kegg as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457 (2016).
48. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. Kegg: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Research* **45**, D353–D361 (2017).
49. Li, Q., Shirabe, K. & Kuwada, J. Y. Chemokine signaling regulates sensory cell migration in zebrafish. *Developmental Biology* **269**, 123–136 (2004).
50. Lippitz, B. E. Cytokine patterns in patients with cancer: a systematic review. *The Lancet Oncology* **14**, e218–e228 (2013).
51. Chen, M. *et al*. Impaired glucose metabolism in response to high fat diet in female mice conceived by *in vitro* fertilization (IVF) or ovarian stimulation alone. *PLoS ONE* **9**, e113155 (2014).
52. Heroult, M., Ellinghaus, P., Ince, S. & Ocker, M. Fibroblast growth factor receptor signaling in cancer biology and treatment. *Current Signal Transduction Therapy* **9**, 15–25 (2014).
53. Kim, T. *et al*. Role of MYC-regulated long noncoding RNAs in cell cycle regulation and tumorigenesis. *JNCI Journal of the National Cancer Institute* **107**, dju505–dju505 (2015).
54. Zeestraten *et al*. The prognostic value of the apoptosis pathway in colorectal cancer: A review of the literature on biomarkers identified by immunohistochemistry. *Biomarkers in Cancer* 13 (2013).
55. Charbonneau, B. *et al*. Risk of ovarian cancer and the NF- b pathway: Genetic association with IL1a and TNFSF10. *Cancer Research* **74**, 852–861 (2013).
56. Sakthivel, K. M. & Sehgal, P. A novel role of lamins from genetic disease to cancer biomarkers. *Oncology Reviews* **10** (2016).
57. Taylor-Harding, B. *et al*. Abstract 1749: Cell cycle requirements shape ovarian cancer progression. *Cancer Research* **73**, 1749–1749 (2013).
58. Salvador, J. M., Brown-Clay, J. D. & Fornace, A. J. Gadd45 in stress signaling, cell cycle control, and apoptosis. In *Advances in Experimental Medicine and Biology*, 1–19 (Springer Nature, 2013).
59. Krausova, M. & Korinek, V. Wnt signaling in adult intestinal stem cells and cancer. *Cellular Signalling* **26**, 570–579 (2014).
60. Prado, C. M. *et al*. The association between body composition and toxicities from the combination of doxil and trabectedin in patients with advanced relapsed ovarian cancer1. *Applied Physiology, Nutrition, and Metabolism* **39**, 693–698 (2014).
61. Jia, P. & Zhao, Z. Varwalker: personalized mutation network analysis of putative cancer genes from next-generation sequencing data. *Plos Computational Biology* **10**, e1003460 (2014).
62. Osterhoff, M. *et al*. Identification of gene-networks associated with specific lipid metabolites by weighted gene co-expression network analysis (WGCNA). *Experimental and Clinical Endocrinology & Diabetes* **122** (2014).
63. Liu, J., Jing, L. & Tu, X. Weighted gene co-expression network analysis identifies specific modules and hub genes related to coronary artery disease. *BMC Cardiovascular Disorders* **16** (2016).
64. Yan, Z. *et al*. The functional consequences and prognostic value of dosage sensitivity in ovarian cancer. *Mol. BioSyst.* **13**, 380–391 (2017).

Acknowledgements

The authors are grateful to Peter Langfelder's team who provided the weighted correlation network analysis method with WGCNA package, Prof Ripley B.D. who provided the Local Polynomial Regression Fitting analysis method with LOESS package, 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 clusterProfiler package. This work was supported by the National Natural Science Foundation grant of China and China Scholarship Council grant.

Author Contributions

X.G.L. and J.B.L. designed and supervised this study. X.G.L., J.B.L. and X.Q. performed the experiments, B.L. and K.Q.L. performed the data analysis, X.G.L., J.B.L. and X.L. wrote the main manuscript text, B.L. and K.Q.L. provided suggestions for the experimental design. All authors reviewed and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-017-16286-5.

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

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.