

Computational Identification of Guillain-Barre Syndrome Related Genes by an mRNA Gene Expression Profile and a Protein-Protein Interaction Network

Chunyang Wang

Tianjin Neurological Institute, Tianjin Medical University General Hospital

Shiwei Liao

Tianjin Huanhu Hospital, Tianjin Neurosurgical Institute

Xiaowei Hu

Tianjin Neurological Institute, Tianjin Medical University General Hospital

Jing Xu (✉ jingxu01@tmu.edu.cn)

Tianjin Neurological Institute, Tianjin Medical University General Hospital

Research Article

Keywords: GBS, genes, mRNA microarray analysis, protein-protein interaction, shortest path

Posted Date: October 28th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-461238/v2>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Version of Record: A version of this preprint was published at Frontiers in Molecular Neuroscience on March 17th, 2022. See the published version at <https://doi.org/10.3389/fnmol.2022.850209>.

Abstract

Background

In this study, we developed a computational method to identify Guillain–Barré syndrome (GBS) related genes based on (i) a gene expression profile, and (ii) the shortest path analysis in a protein-protein interaction (PPI) network.

Results

Totally 30 GBS related genes were screened out, in which 20 were retrieved from PPI analysis of up-regulated expressed genes and 23 were from down-regulated expressed genes (13 overlap genes). GO enrichment and KEGG enrichment analysis were performed respectively. Results showed that there were some overlap GO terms and KEGG pathway terms in both up-regulated and down-regulated analysis, which indicated these terms may play critical role during GBS process.

Discussion

These results could shed some light on the understanding of the Genetic and molecular pathogenesis of GBS disease, providing basis for future experimental biology studies and for the development of effective genetic strategies for GBS clinical therapies.

Background

Guillain–Barré syndrome (GBS) is a rapid-onset muscle weakness as a result of damage to the peripheral nervous system. The exact molecular mechanism and epigenetic feature of this disease are still unclear. Therefore, it is of great importance to identify GBS-related genes that could be used as biomarker for early diagnosis and effective genetic strategies for clinical therapies.

How to identify the disease genes has become one of the most important and challenging problems in biomedicine and genomics. Gene expression profiles can be used to select differentially expressed genes as disease genes. These methods are useful resources and have been widely used [1–4]. However, the false-positive problem and the errors in the high-throughput data have not been well solved yet [5]. Thus, it is not a good idea to use only the gene expression profiles to identify novel genes.

The “guilt by association” rule suggests that interacting proteins share the same or similar functions and hence may be involved in the same pathway. This assumption, first proposed by Nabieva *et al.* [6], can be used to identify disease related genes from protein-protein interaction (PPI) networks constructed from existing PPI data. Methods based on the PPI data have been widely used for gene function predictions.

In this study, we developed a computational method to identify GBS related genes by integrating a gene expression profile and a weighted functional PPI network, which can make up the defect of only using high-throughput data. We also have successfully applied the integrating method to gene function prediction and to identification of novel genes of various kinds of diseases, such as influenza A/H7N9 virus infection [7], colorectal cancer [5, 8], lung cancer [9], HBV virus infection related hepatocellular carcinoma [10], retinoblastoma [11], etc.

Implementation

The overall procedure of our method is illustrated in Figure 1. In the following subsections, details are presented:

Mrna Expression Profiles Of Gbs Patients And Healthy Controls

mRNA Microarray analyses were performed on the PBMCs of 4 GBS patients and 4 age- and gender-matched healthy controls.

GBS patients fulfilled the standard diagnostic criteria for GBS were recruited from Tianjin Medical University General Hospital [12]. Blood was sampled when patients were within the peak timing of manifesting GBS and before treatment with intravenous immune globulin (IVIG), plasma exchange or glucocorticoid. Informed consent was obtained at enrollment from all patients or legally acceptable surrogates. The present study was approved by the ethical review committees of Tianjin Medical University General Hospital. Human peripheral blood mononuclear cells (PBMCs) were isolated from all GBS patients and healthy controls. The labeled cRNAs were hybridized onto the human LncRNA Expression Microarray V3.0 (Arraystar, Rockville, MD), which was designed for the global profiling of human lncRNAs, mRNAs and protein-coding transcripts.

Totally an expression profile dataset of 8 samples, 21,620 probes was obtained. Then signal intensity was log₂ transformed and normalized and 14,707 genes were derived from source probes.

The Mrmr Method

We employed the mRMR (Maximum Relevance Minimum Redundancy) method [5, 13–15] to rank the importance of all the 14,707 genes examined, according to the Maximum Relevance Minimum Redundancy criterion. In this procedure, each gene was regarded as a feature. The Maximum Relevance criterion selects features most important in discriminating GBS samples and controls. The Minimum Redundancy criterion excludes features containing redundant information among the selected features. Briefly, to rank features using mRMR criteria, two values were calculated for each feature: value A for relevance and value B for redundancy. Then, the value A-B is used to measure the feature; the higher the value A-B is, the higher the feature ranks. For details of the mRMR method, please refer to [5, 13–15].

Two ordered list were generated by the mRMR method, one was called the MaxRel Table, and the other was called mRMR Table. In the MaxRel table, all the features were ranked only by the Maximum Relevance criterion. In the mRMR Table, they were ranked by the mRMR criteria, i.e., a feature with a smaller index in such a table indicated that it had a better trade-off between the maximum relevance and the minimum redundancy and thus could be more important. The two tables were provided in Supporting Information S1. In this study, we select the top 5%, i.e., 735 features, from the mRMR Table. The 735 genes were regarded as significant differentially expressed genes from the expression profiles and were analyzed in the next procedures.

In the 735 significant genes, there were 271 up-regulated genes and 464 down-regulated genes, producing 1,619 and 2,590 protein products respectively. The up-regulated genes and down-regulated genes were analyzed respectively in the next procedures. The number of the genes and proteins were summarized in Table 1.

Table 1
Number of genes or proteins in each step of our computational procedures.

	Significant differentially expressed genes by mRMR		Shortest path proteins	Shortest path proteins with betweenness > threshold	Final GBS related genes
	genes	proteins(ENSP)			
Up-regulated	271	1,619	858	20	20
Down-regulated	464	2,590	1,273	24	23

Ppi Data From String

STRING (Search Tool for the Retrieval of Interacting Genes) [16] (<http://string.embl.de/>), is an online database resource which compiles both experimental and predicted protein-protein interactions with a confidence score to quantify each interaction confidence. A weighted PPI network can be retrieved from STRING, in which proteins in the network are represented as nodes, while interactions between proteins are given as edges marked with confidence scores if they are in interaction with each other. Interacting proteins with high confidence scores in such a PPI network are more likely to share similar biological functions than non-interactive ones [16–18]. This is because the protein and its interactive neighbours may form a protein complex performing a particular function or may be involved in the same pathway.

In this study, we constructed a graph G with the PPI data from STRING (version 9.0). In such a graph, proteins were represented as nodes; however the weight of each interaction edge was assigned a d value instead of a confidence score (s). The d value was derived from the confidence score s according to the equation. Thus, the d value can be considered as representing protein distances to each other: the smaller the distance, the higher the interaction confidence score and the more similar the functions they have.

In this study, we analyzed in such a graph every two protein interactions from the significant differentially expressed proteins.

Shortest Path Tracing

The Dijkstra algorithm were often used to find the shortest path in the graph G between two given proteins. In this study, the Dijkstra algorithm was implemented with R package 'igraph'. A shortest path was traced from each of the 1,619 proteins to all the other 1,618 proteins in the graph, which was for the up-regulated genes. For down-regulated genes, the shortest path of each of the 2,590 proteins to all the other 2,589 proteins was traced in the graph.

Then we picked out all proteins existing on the shortest paths and ranked these proteins according to their betweenness. For up-regulated genes, 858 shortest path proteins were retrieved, while for down-regulated genes, 1,273 shortest path proteins were retrieved, as list in Table 1. The 858 and 1,273 shortest path proteins were ranked by betweenness respectively.

The betweenness threshold should be set in order to select significant ones from a ranked list. By a computational method, we can set the threshold differently to yield different number of gene results. The more the threshold is, the less the genes are. Generally speaking, it is practical to select the top most 20 to 30 significant genes for further analysis or for experimental validation. Furthermore, the threshold values should be different for up-regulated genes and for down-regulated ones, because the number of path tracing proteins and the number of shortest path proteins were different.

In this study, for shortest path proteins from the analysis of up-regulated genes, top 20 proteins (20 genes) with betweenness $>1,400$ were selected, while for down-regulated genes the top 24 proteins (23 genes) with betweenness $>4,000$ were selected. These 20+23 genes were regarded as the final significant GBS-related genes in this study and they were list in Table 2 and Table 3, respectively.

Table 2
The 20 GBS-related genes identified by PPI network
from the analysis of up-regulated expressed genes
from the expression profiles

ENSP	Gene	Betweenness
ENSP00000269305	TP53	13,036
ENSP00000344818	UBC	6,065
ENSP00000344456	CTNNB1	4,569
ENSP00000275493	EGFR	3,294
ENSP00000263253	EP300	2,807
ENSP00000326366	PSEN1	2,456
ENSP00000417281	MDM2	2,445
ENSP00000270202	AKT1	2,350
ENSP00000221494	SF3A2	2,184
ENSP00000264657	STAT3	2,182
ENSP00000339007	GRB2	2,150
ENSP00000324806	GSK3B	2,094
ENSP00000284981	APP	2,046
ENSP00000357879	PSMD4	1,754
ENSP00000350941	SRC	1,655
ENSP00000356425	UHL5	1,614
ENSP00000361626	YBX1	1,574
ENSP00000338018	HIF1A	1,444
ENSP00000262613	SLC9A3R1	1,438
ENSP00000252486	APOE	1,410

Table 3
The 23 GBS-related genes identified by PPI network
from the analysis of down-regulated expressed
genes from the expression profiles

ENSP	Gene	Betweenness
ENSP00000269305	TP53	36,055
ENSP00000344818	UBC	15,309
ENSP00000275493	EGFR	12,072
ENSP00000344456	CTNNB1	12,052
ENSP00000270202	AKT1	10,629
ENSP00000339007	GRB2	10,496
ENSP00000221494	SF3A2	9,484
ENSP00000206249	ESR1	8,165
ENSP00000263253	EP300	7,353
ENSP00000264657	STAT3	6,262
ENSP00000350941	SRC	6,111
ENSP00000417281	MDM2	6,002
ENSP00000362649	HDAC1	6,001
ENSP00000348461	RAC1	5,995
ENSP00000329357	SP1	5,560
ENSP00000361626	YBX1	5,343
ENSP00000264033	CBL	5,062
ENSP00000337825	LCK	4,852
ENSP00000314458	CDC42	4,798
ENSP00000304903	CD2BP2	4,549
ENSP00000358490	CD2	4,549
ENSP00000324806	GSK3B	4,281
ENSP00000046794	LCP2	4,043

Results And Discussion

GBS-related genes

In this study, results of two sets of GBS-related genes were identified. As shown in Table 2 and Table 3, 20 GBS-related genes were identified from the analysis of up-regulated significant expressed genes from the gene expression profile data, while 23 GBS-related genes were identified from the data of down-regulated ones.

From Table 2 and Table 3, it can be seen that there were 13 overlap genes, which were TP53, UBC, CTNNB1, EGFR, EP300, MDM2, AKT1, SF3A2, STAT3, GRB2, GSK3B, SRC, YBX1. These genes were identified both from up-regulated analysis and down-regulated analysis, indicating they could play more important roles in GBS.

TP53 is a critical tumor suppressor gene that regulates cell cycle progression, apoptosis, cellular senescence and many other properties critical for control of normal cellular growth and death [19]. MDM2 has p53-independent transcription factor-like effects in nuclear factor-kappa beta (NFκB) activation. Therefore, MDM2 promotes tissue inflammation and MDM2 inhibition has potent anti-inflammatory effects in tissue injury[20]. Increased levels of STAT3 proteins were observed in CD4+ T cells from GBS patients [21]. Previous data reveal that Grb2 facilitates the association of FasL with adaptin beta, and promotes sorting of FasL to the cell surface in Schwann cells. As FasL is a potent regulator of cell death, dynamic regulation of its cell surface localization is critical for controlling local tissue remodeling and inflammation [22].

Ppi Relationship Between The Gbs-related Genes

We mapped all the GBS-related genes to the PPI network constructed from the STRING database. The PPI relationships between the GBS-related genes were shown in Figure 2. The coding genes of the proteins were denoted as nodes. The 20 GBS-related genes identified from up-regulated analysis were represented as red circles, while the 23 genes from down-regulated ones were represented as blue circles. Note that there were 13 overlap genes, which were represented as green circles.

Go Enrichment Analysis

The functional annotation tool DAVID [23] was implemented for GO enrichment analysis on the GBS-related genes. The results were provided in Supporting Information S2. We also plot the GO enrichment results in Figure 3 from the data in Supporting Information S2. The overlap GO terms in both up-regulated and down-regulated analysis were listed as follows:

GO:0010604~positive regulation of macromolecule metabolic process

GO:0007242~intracellular signaling cascade

GO:0007166~cell surface receptor linked signal transduction

GO:0043232~intracellular non-membrane-bounded organelle

GO:0043228~non-membrane-bounded organelle

GO:0005886~plasma membrane

Kegg Pathway Enrichment Analysis

The functional annotation tool DAVID [23] was implemented for KEGG enrichment analysis on the GBS-related genes. The enrichment p -value was corrected to control family-wide false discovery rate under a certain rate (e.g. ≤ 0.05) with the Benjamin multiple testing correction method. The results were provided in Supporting Information S3. We also plot the KEGG pathway enrichment results in Figure 4 from the data in Supporting Information S3. The overlap KEGG pathway terms in both up-regulated and down-regulated analysis were listed as follows:

hsa05215: Prostate cancer

hsa05200: Pathways in cancer

hsa05213: Endometrial cancer

hsa05210: Colorectal cancer

hsa05214: Glioma

hsa04012: ErbB signaling pathway

hsa04510: Focal adhesion

hsa04722: Neurotrophin signaling pathway

hsa04310: Wnt signaling pathway

Conclusions

The problem of how to identify disease genes has become one of the most important and challenging problems in biomedicine and genomics. In this study, we developed a computational method to identify GBS related genes based on the shortest paths in the PPI network. Finally, 30 genes were screened out which could be the most significant, which to some extent imply their direct or indirect relationship with the development or formation of GBS, providing clues for further researches and experimental validations. These results may provide a new reference for researches into GBS pathogenesis and for new strategies of clinical therapies.

Declarations

Ethics and Consent to Participate

The present study was approved by the ethical review committees of Tianjin Medical University General Hospital.

Consent to Publish

Not applicable

Availability of Data and Materials

None

Competing interests

The authors declare that they have no competing interests

Authors' contributions

Authors individual contributions to this article are: Data curation and formal analysis, CYW and SWL; Project administration, Funding acquisition and Supervision, JX. Collect data: XWH. All authors read and approved the final manuscript.

Funding

The funding for the study was provided by the National Natural Science Foundation of China (grant nos. 81601041), Medical Foundation of Jieping Wu (grant. no. 320.6750.19089-56), and Youth incubation fund of General Hospital of Tianjin Medical University (grant. no.zyyfy2019007).

Author's Information

¹ Department of Neurology, Tianjin Medical University General Hospital, Tianjin, 300052, China

² Department of Neurorehabilitation and Neurology, Tianjin Huanhu Hospital, Tianjin Key Laboratory of Cerebral Vascular and Neurodegenerative Diseases, Tianjin Neurosurgical Institute, Tianjin 300350, China.

References

1. Huang, T., et al., *Classification and analysis of regulatory pathways using graph property, biochemical and physicochemical property, and functional property*. PLoS One, 2011. **6**(9): p. e25297.

2. Cai, Y.D., et al., *A unified 35-gene signature for both subtype classification and survival prediction in diffuse large B-cell lymphomas*. PLoS One, 2010. **5**(9): p. e12726.
3. Huang, T., et al., *Analysis and prediction of the metabolic stability of proteins based on their sequential features, subcellular locations and interaction networks*. PLoS One, 2010. **5**(6): p. e10972.
4. Liu, Q., et al., *TF-centered downstream gene set enrichment analysis: Inference of causal regulators by integrating TF-DNA interactions and protein post-translational modifications information*. BMC Bioinformatics, 2010. **11 Suppl 11**: p. S5.
5. Li, B.Q., et al., *Identification of colorectal cancer related genes with mRMR and shortest path in protein-protein interaction network*. PLoS One, 2012. **7**(4): p. e33393.
6. Nabieva, E., et al., *Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps*. Bioinformatics, 2005. **21 Suppl 1**: p. i302-10.
7. Zhang, N., et al., *Identification of Influenza A/H7N9 virus infection-related human genes based on shortest paths in a virus-human protein interaction network*. Biomed Res Int, 2014. **2014**: p. 239462.
8. Li, B.Q., et al., *An ensemble prognostic model for colorectal cancer*. PLoS One, 2013. **8**(5): p. e63494.
9. Li, B.Q., et al., *Identification of lung-cancer-related genes with the shortest path approach in a protein-protein interaction network*. Biomed Res Int, 2013. **2013**: p. 267375.
10. Jiang, M., et al., *Identification of hepatocellular carcinoma related genes with k-th shortest paths in a protein-protein interaction network*. Mol Biosyst, 2013. **9**(11): p. 2720–8.
11. Li, B.Q., et al., *Identification of retinoblastoma related genes with shortest path in a protein-protein interaction network*. Biochimie, 2012. **94**(9): p. 1910–7.
12. Asbury, A.K., *Diagnostic considerations in Guillain-Barre syndrome*. Ann Neurol, 1981. **9 Suppl**: p. 1–5.
13. Zhang, N., et al., *Computational prediction and analysis of protein gamma-carboxylation sites based on a random forest method*. Mol Biosyst, 2012. **8**(11): p. 2946–55.
14. Peng, H., F. Long, and C. Ding, *Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy*. IEEE Trans Pattern Anal Mach Intell, 2005. **27**(8): p. 1226–38.
15. Li, B.Q., et al., *Prediction of protein domain with mRMR feature selection and analysis*. PLoS One, 2012. **7**(6): p. e39308.
16. Szklarczyk, D., et al., *The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored*. Nucleic Acids Res, 2011. **39**(Database issue): p. D561-8.
17. Kourmpetis, Y.A., et al., *Bayesian Markov Random Field analysis for protein function prediction based on network data*. PLoS One, 2010. **5**(2): p. e9293.
18. Ng, K.L., J.S. Ciou, and C.H. Huang, *Prediction of protein functions based on function-function correlation relations*. Comput Biol Med, 2010. **40**(3): p. 300–5.
19. McCubrey, J.A., et al., *Roles of TP53 in determining therapeutic sensitivity, growth, cellular senescence, invasion and metastasis*. Adv Biol Regul, 2016.

20. Ebrahim, M., et al., *MDM2 beyond cancer: podoptosis, development, inflammation, and tissue regeneration*. *Histol Histopathol*, 2015. **30**(11): p. 1271–82.
21. Liu, H., et al., *IL-35 Is Involved in the Pathogenesis of Guillain-Barre Syndrome Through Its Influence on the Function of CD4+ T Cells*. *Immunol Invest*, 2015. **44**(6): p. 566–77.
22. Thornhill, P.B., et al., *The adaptor protein Grb2 regulates cell surface Fas ligand in Schwann cells*. *Biochem Biophys Res Commun*, 2008. **376**(2): p. 341–6.
23. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*. *Nat Protoc*, 2009. **4**(1): p. 44–57.

Figures

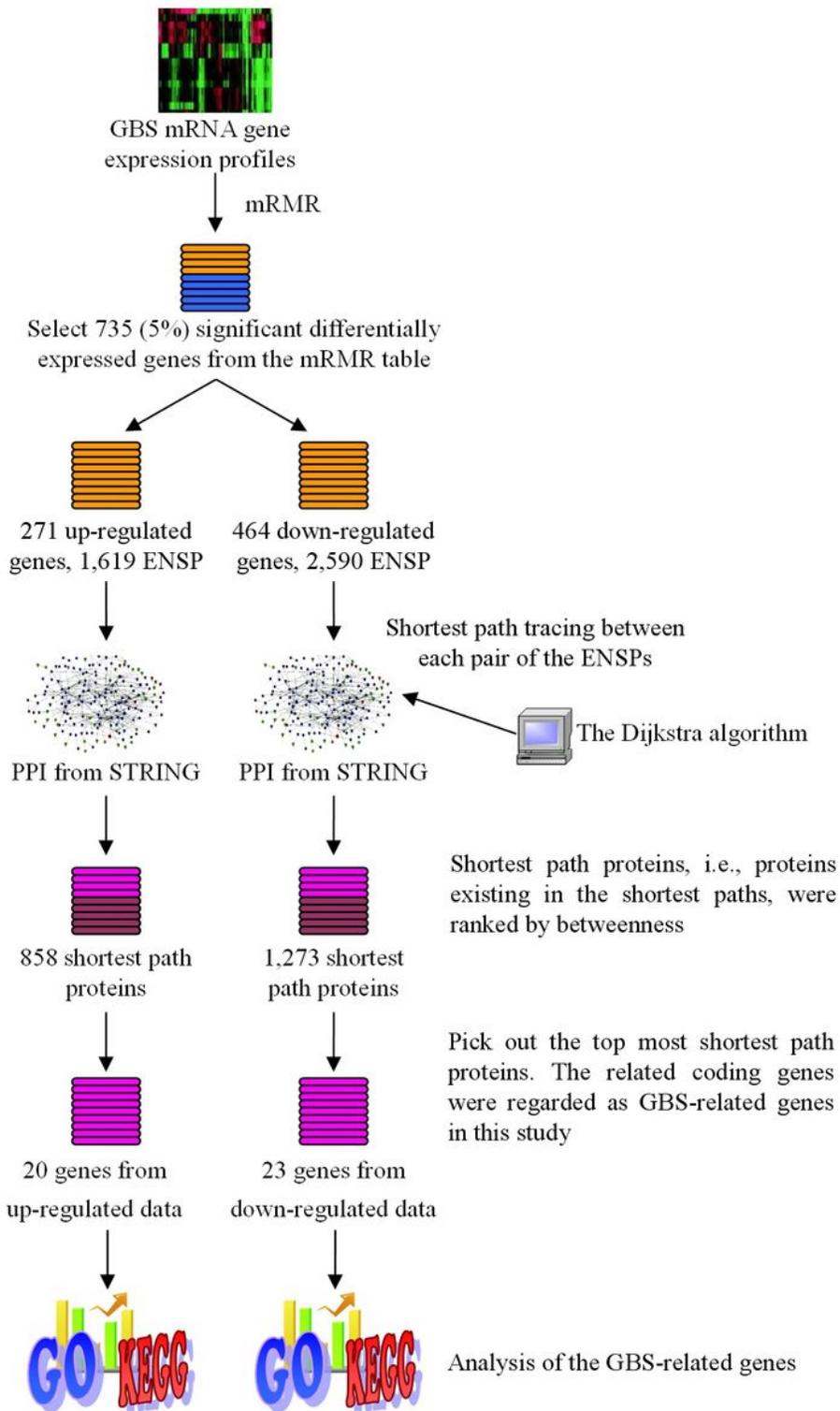


Figure 1

The flowchart of the method developed in this study to identify the influenza A/H7N9 infection-related human genes. Target Human Proteins interacted with the influenza A/H7N9 virus were obtained based on sharing GO terms. Shortest path proteins were calculated from the shortest paths between every pair of the Target Human Proteins, by searching by the Dijkstra algorithm in the network constructed from

STRING. Finally 20 shortest path proteins were screened out with betweenness >10,000, the related genes of which were considered as infection-related human genes.

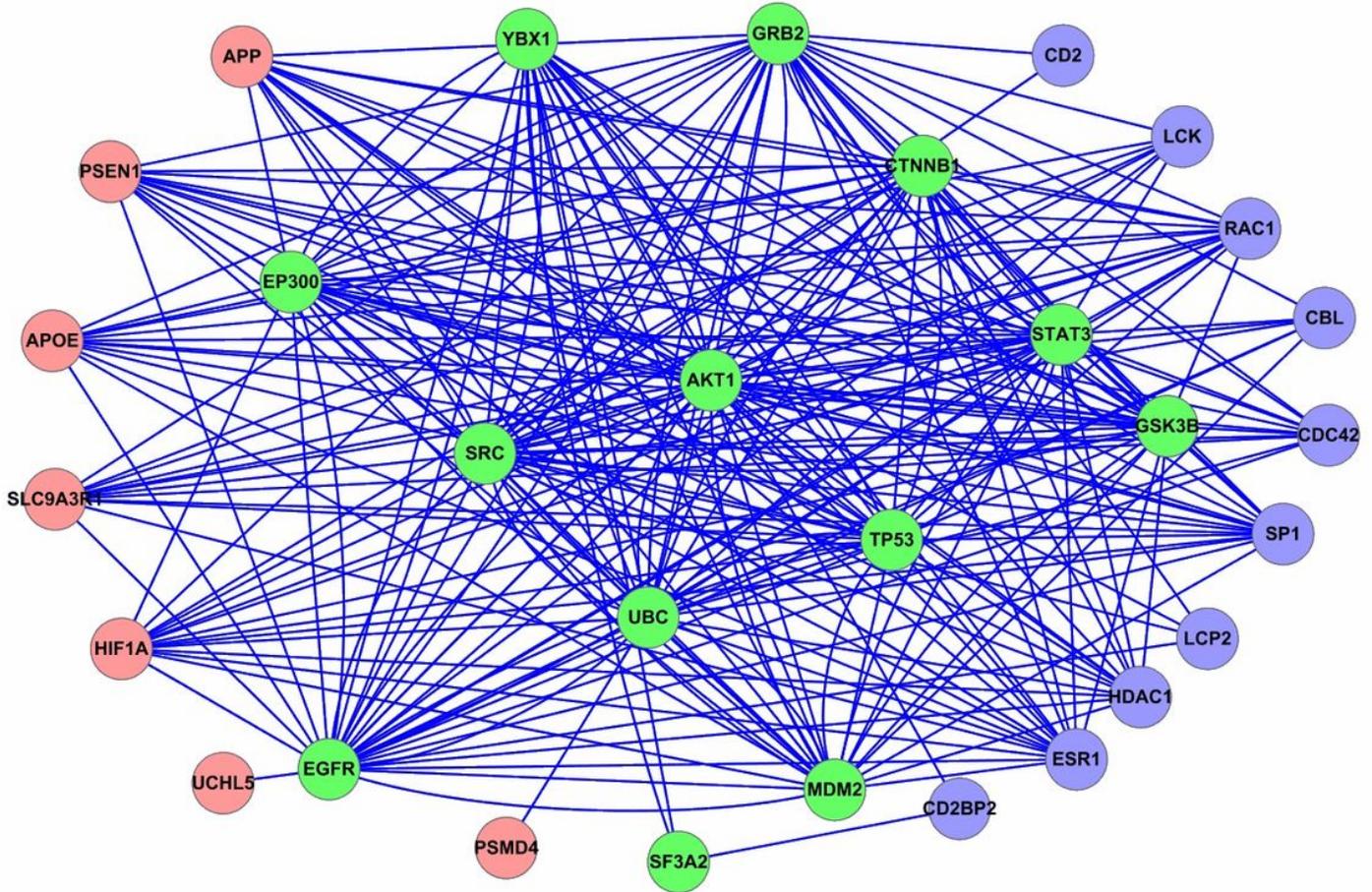
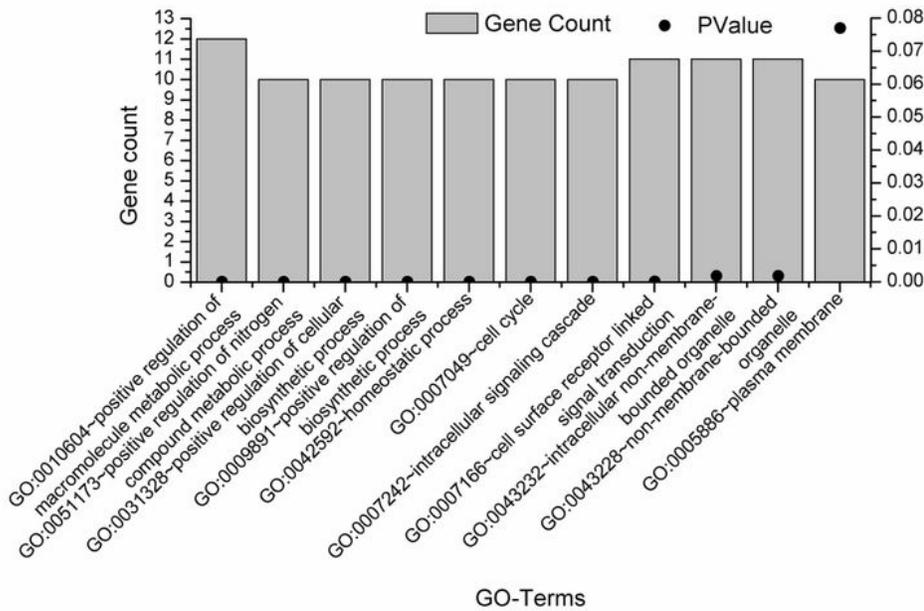
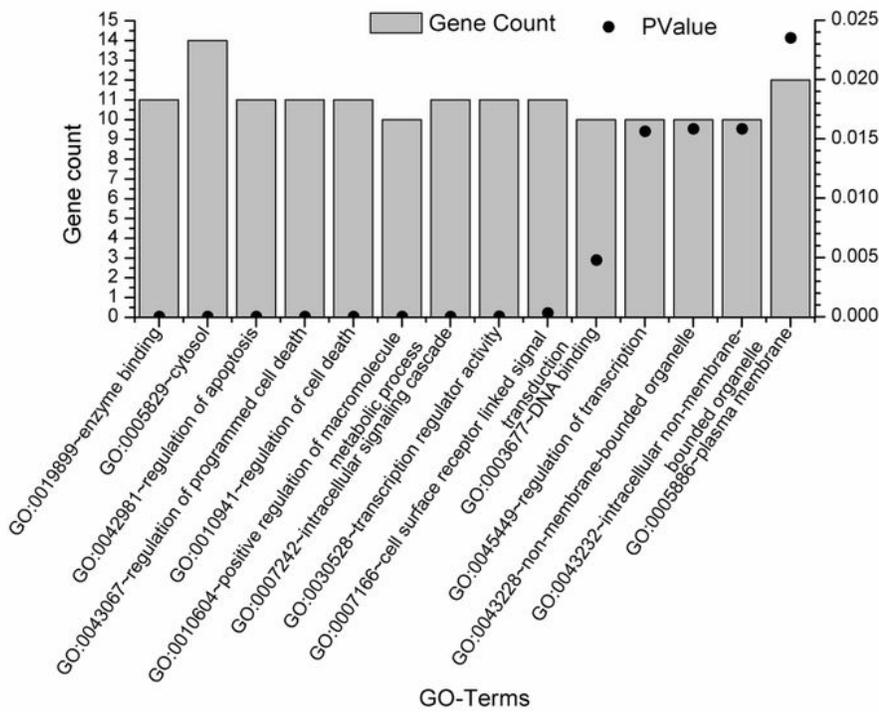


Figure 2

PPI relationship between all the GBS-related genes identified in this study. Red circle represents the GBS-related genes identified from up-regulated analysis. Blue circle represents the GBS-related genes identified from down-regulated analysis. Green circle represents the overlap GBS-related genes both from up-regulated and down-regulated analysis.



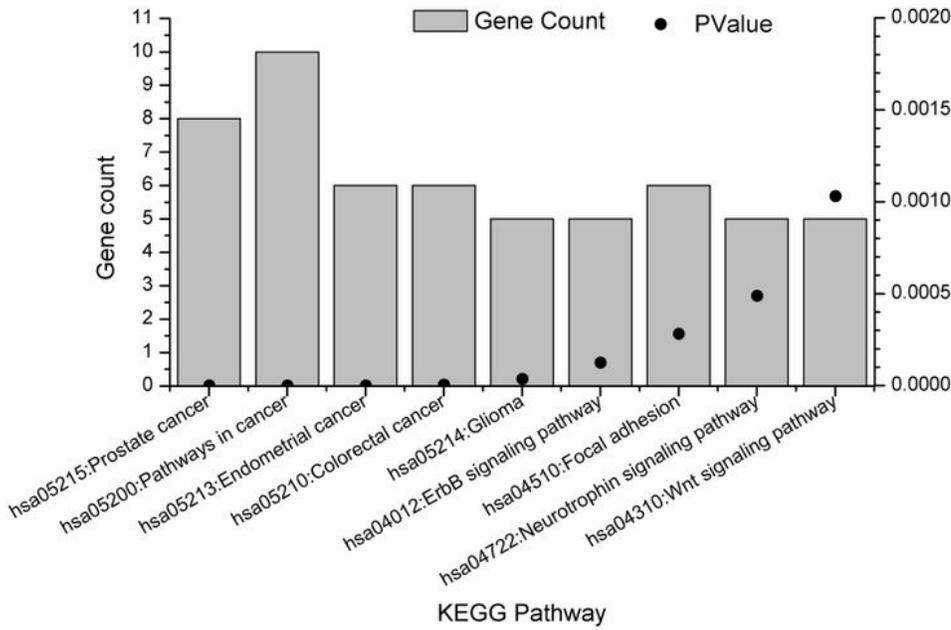
(A) Up-regulated analysis



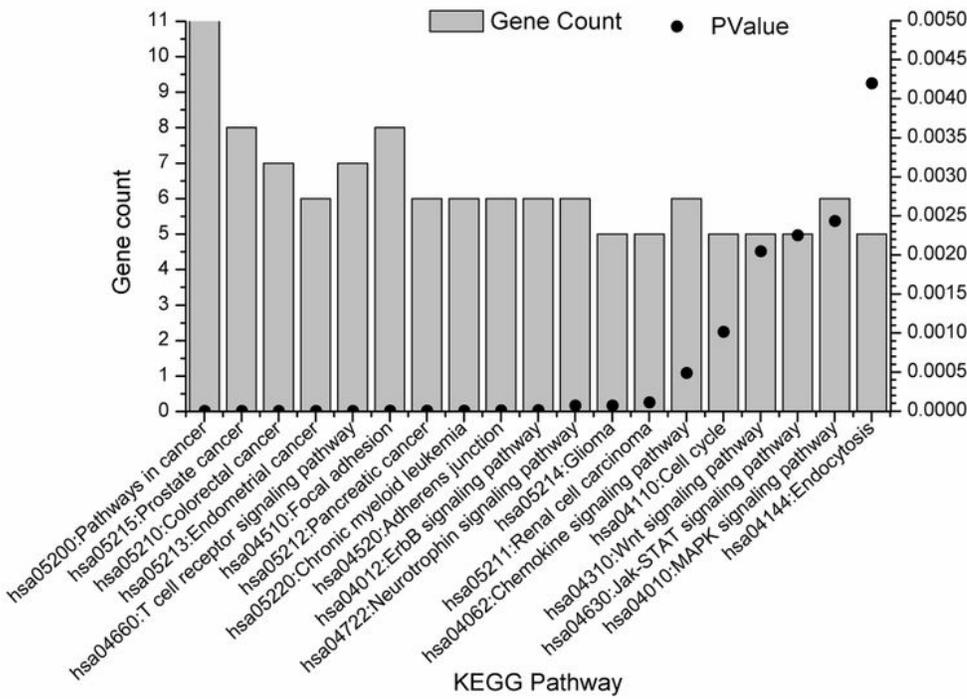
(B) Down-regulated analysis

Figure 3

GO term enrichment analysis of the GBS-related genes. GO term enrichment analysis were performed on the 20 genes from up-regulated analysis and 23 genes from down-regulated analysis, with results plotted in (A) and (B), respectively. Only pathways with gene count ≥ 10 were shown.



(A) Up-regulated analysis



(B) Down-regulated analysis

Figure 4

KEGG term enrichment analysis of the GBS-related genes. KEGG term enrichment analysis were performed on the 20 genes from up-regulated analysis and 23 genes from down-regulated analysis, with results plotted in (A) and (B), respectively. Only pathways with gene count ≥ 5 were shown.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupportingInformationS1.xlsx](#)
- [SupportingInformationS2.xlsx](#)
- [SupportingInformationS3.xlsx](#)