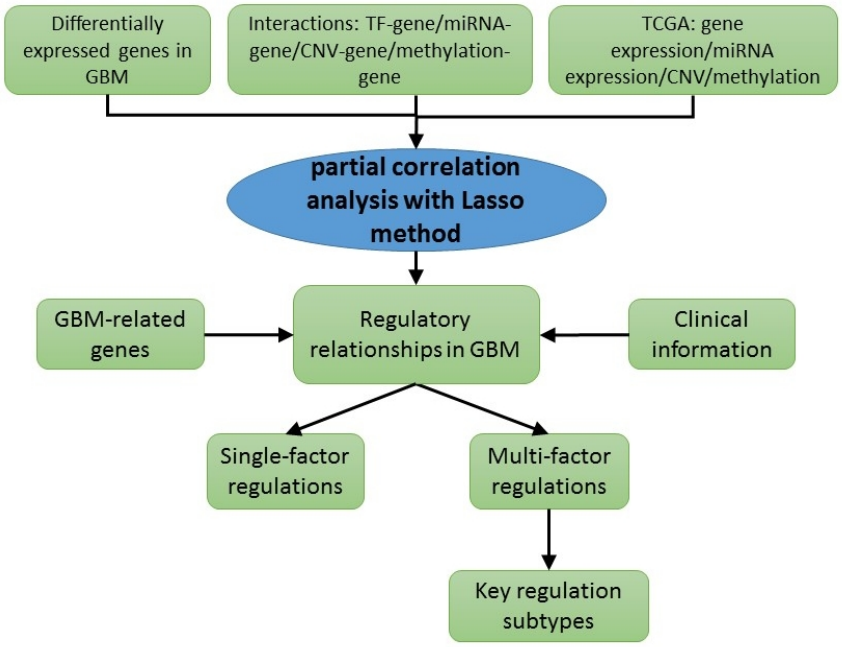
Electronic Supplementary Material (ESI) for Molecular BioSystems.

Discovering key regulatory mechanisms from single-factor and multi-factor regulations in glioblastoma utilizing multi-dimensional data

Chen Peng, Yi Shen, Mengqu Ge, Minghui Wang, Ao Li

1. Supplementary material and method

Overview of the method



The flow chart of the method in this study is shown above. Firstly we extracted differentially expressed genes in GBM based on gene expressions in normal and tumor. Secondly we extracted the predicted/possible regulators of each differentially expressed gene according to the prior information. Then the TCGA data containing gene expression, miRNA expression, CNV and methylation about these regulators and the differentially expressed target genes were imported to partial correlation analysis with Lasso method as imput. The results of regulatory relationships were further classified into two types: single-factor and multi-factor and further analysis of multi-factor regulations indicate three key regulation subtypes. In addition, GBM-related genes with evidence of involvement in GBM were used for investigating the

genes related to GBM in the results of regulatory relationships and the clinical information was also used in subsequent survival analysis.

Predicted regulatory interactions

Four types of regulatory relationships were investigated in this study: TF regulation (TF-gene), miRNA regulation (miRNA-gene), CNV regulation (CNV-gene) and methylation regulation (methylation-gene). Candidate TF-gene and miRNA-gene interactions were retrieved from [1], which have totally 130,338 TF-gene interactions including 214 human TFs, 16,534 target genes and 118,408 miRNA-gene interactions with 10,255 target genes and 276 human miRNAs. The regulations of CNV and methylation were also assigned to each gene if its CNV and methylation values were avaliable. The integration of above predicted interactions is regarded as prior information in this study. Since these prior information usually contains false positives, we further used partial correlation analysis with Lasso method to solve this problem.

Gene expression, miRNA expression, CNV, methylation and clinical data for GBM

We used multi-dimensional data of GBM in this study, which include four 'omics': miRNA expression, gene expression, CNV and methylation. To obtain these data we resorted to TCGA [2], which produces unprecedented multi-dimensional data of cancers. First we downloaded the normalized data (‘level 3’) of GBM patients for CNV, methylation, miRNA expression and gene expression. Then we extracted totally 277 GBM patients in these data, which contain the information of all these four 'omics' (i.e., each has one set of CNV, one set of methylation, one set of miRNA

expression and one set of gene expression). For differential expression analysis, we additionally downloaded the gene expression data of organ-specific normal samples. Besides, we also obtained all available public survival data for each patient. All these data were provided by the GBM catalog in TCGA Data Portal (https://tcga-data.nci.nih.gov/tcga). Tumor samples were taken from the participants in MD Anderson while normal samples from donor tissue of individules that did not have cancer. As for gene expression, miRNA expression and methylation, we directly extract the values of corresponding samples for each gene. Since the CNV profiles only provide mean values of chromosome segments, we further downloaded 'refGene.txt' from UCSC (http://genome.ucsc.edu/) that contains the chromosomal location of 44,914 genes and assigned the average intensity value of CNV to each gene. Finally, the four 'omics' data of GBM patients as well as the gene expression data of organ-specific normal samples were transformed into five normalized matrixes in which a row represent a gene and a column represent a sample. The data propressing process was carried out with matlab.

GBM-related genes with evidence of involvement in GBM

As for GBM-related genes with evidence of involvement in GBM, we resorted to six sources: the Online Mendelian Inheritance in Man (OMIM) [3], two genome-wide association studies [4-5], TCGA [2], the Catalogue Of Somatic Mutations In Cancer (COSMIC, version 67) [6], the Genetic Association Database (GAD) [7] and one integrative genomic analysis of GBM [8]. The collection were performed following the procedure described in [9]. Firstly, we downloaded gene information from

COSMIC database (version 66, released on 25th July 2013, http://cancer.sanger.ac.uk/cosmic). For each gene, we defined the number of samples that were labeled with 'GBM' as *D* and the number of samples with positive mutation records as *d*. Then 556 genes with *d* no less than 2 and a percentage (*d/D*) no less than 3% were obtained. Secondly, from supplementary Table 6 of [2] (data.nci.nih.gov/docs/somatic_mutations/tcga_mutations.htm), we retrieved 215 genes with validated non-silent somatic mutations. In addition to above resources, 3 GBM-related genes obtained by searching OMIM with key word "glioblastoma" (http://www.ncbi.nlm.nih.gov/omim) [3] and 5 GBM-related genes that have positive association reports from GAD (updated on 1th Sepetember 2014, http://geneticassociationdb.nih.gov) [7] were collected. Besides, we retrieved 42 genes from supplementary Table 7 of [8], which have accumulated mutations at significant frequencies, and 6 risk genes for GBM from 2 glioma genome-wide association studies [4] and [5]. After integrating these data and removing the duplicates, we finally got 760 experimentally validated GBM-related genes.

Reference

1. Yu, H. *et al*. Combinatorial network of transcriptional regulation and microRNA regulation in human cancer. *BMC. SYST. BIOL.* **6**, 61 (2012).
2. McLendon, R. *et al.* Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061-1068 (2008).
3. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-D517 (2005).
4. Shete, S. *et al*. Genome-wide association study identifies five susceptibility loci for glioma. *Nat. Genet.* **41**, 899-904 (2009).
5. Wrensch, M. *et al*. Variants in the CDKN2B and RTEL1 regions are associated

with high-grade glioma susceptibility. *Nat. Genet.* **41**, 905-908 (2009).
6. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945-D950 (2011).
7. Becker, K. G., Barnes, K. C., Bright, T. J. & Wang, S. A. The genetic association database. *Nat. Genet.* **36**, 431-432 (2004).
8. Parsons DW, Jones S, Zhang X, Lin JC, Leary RJ, *et al.* An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**: 1807-1812 (2008).
9. Sun, J., Gong, X., Purow, B. & Zhao, Z. Uncovering microRNA and transcription factor mediated regulatory networks in glioblastoma. *PLoS Comput. Biol.* 8, e1002488 (2012).

2. Supplementary Tables

Table S1 Detailed information of 164 experimentally validated GBM-related genes

Gene name	*p-value* of differential expression	*q-value* after BN correction
A2M	3.2E-08	2.4E-04
ABCC3	4.1E-26	4.1E-22
ABCC4	5.1E-13	4.6E-09
ABCC5	1.6E-08	1.3E-04
ACAN	9.6E-07	6.3E-03
ACTR3B	2.0E-08	1.5E-04
ADAM12	3.7E-07	2.5E-03
AIFM1	3.4E-07	2.4E-03
ANK2	4.9E-12	4.3E-08
ANXA1	1.9E-12	1.7E-08
ARHGEF6	5.0E-12	4.4E-08
ARNT2	2.3E-12	2.0E-08
ATM	8.5E-26	8.5E-22
ATP13A2	3.8E-08	2.8E-04
ATP6V1E1	2.0E-52	2.1E-48
ATP8B1	1.8E-07	1.2E-03
AURKB	5.0E-12	4.4E-08
BAI3	1.8E-11	1.6E-07
BAZ1A	2.4E-27	2.4E-23
BCL11A	6.1E-10	5.0E-06
BCL2L13	5.0E-13	4.6E-09
BCL6	9.0E-07	6.0E-03
CACNA1I	1.1E-07	7.8E-04
CAST	1.0E-19	9.9E-16
CCR5	1.6E-13	1.5E-09
CD46	1.6E-09	1.3E-05
CDK4	7.4E-11	6.4E-07
CENPF	1.6E-33	1.7E-29
CFI	9.0E-13	8.1E-09
CHD4	1.5E-08	1.1E-04
CHD5	6.7E-09	5.2E-05
CHEK1	6.7E-23	6.6E-19
CHEK2	1.3E-13	1.2E-09
CHL1	3.2E-07	2.2E-03
CNTN6	9.0E-08	6.5E-04
COG5	1.1E-07	8.0E-04
COL15A1	1.3E-10	1.1E-06

COL1A1	2.4E-14	2.2E-10
COL1A2	1.4E-06	9.3E-03
COL4A2	8.1E-14	7.4E-10
COL6A2	3.0E-22	3.0E-18
CPS1	8.5E-07	5.6E-03
CTBP2	6.5E-07	4.4E-03
CYP46A1	1.6E-29	1.6E-25
DARS2	1.7E-09	1.4E-05
DDX1	9.7E-07	6.4E-03
DHTKD1	2.1E-07	1.5E-03
DNAH17	1.4E-06	9.0E-03
DNAH9	2.3E-08	1.7E-04
DUSP22	1.6E-07	1.1E-03
DYNC1H1	2.3E-07	1.6E-03
EGFR	3.0E-21	2.9E-17
EIF4G2	3.2E-10	2.7E-06
ELAVL2	6.8E-07	4.6E-03
ENPP2	1.4E-08	1.1E-04
ERBB2	3.4E-09	2.7E-05
FBXW7	1.8E-11	1.6E-07
FEZ1	1.0E-10	8.6E-07
FGFR2	1.5E-07	1.1E-03
FLI1	2.0E-08	1.5E-04
FN1	1.8E-10	1.5E-06
FUBP1	5.6E-09	4.4E-05
GALNT3	4.3E-08	3.2E-04
GNAI1	3.9E-07	2.7E-03
GRIA2	1.3E-31	1.3E-27
GRM3	3.4E-21	3.4E-17
GSTM5	1.0E-06	6.7E-03
GYPC	9.2E-08	6.6E-04
H3F3A	2.1E-10	1.7E-06
HERC1	6.6E-09	5.2E-05
HIF1A	9.0E-21	8.9E-17
IBTK	1.5E-07	1.1E-03
ID3	6.6E-09	5.2E-05
IDH1	8.8E-11	7.5E-07
IQGAP1	2.2E-20	2.2E-16
ITGAV	3.9E-18	3.8E-14
JAG1	6.4E-19	6.3E-15
KCNJ12	3.5E-18	3.4E-14
KLF6	3.7E-17	3.6E-13

KNTC1	4.0E-14	3.7E-10
KPNA2	3.2E-13	2.9E-09
KRAS	8.0E-11	6.8E-07
KRIT1	6.4E-15	6.0E-11
KTN1	3.3E-09	2.7E-05
LGALS3BP	6.4E-13	5.8E-09
LRRN2	1.2E-07	8.8E-04
LTF	1.1E-19	1.1E-15
LYN	2.1E-08	1.6E-04
MAN2B1	1.2E-32	1.2E-28
MAP3K6	1.4E-09	1.1E-05
MAPK7	8.0E-14	7.4E-10
MAPK9	1.4E-07	9.8E-04
MEOX2	1.5E-16	1.4E-12
MLH1	2.1E-17	2.0E-13
MSH6	1.1E-09	9.2E-06
MSR1	2.1E-08	1.6E-04
MYO1B	2.1E-13	1.9E-09
MYO5C	1.6E-09	1.3E-05
NBN	7.4E-18	7.2E-14
NELL2	8.8E-49	9.0E-45
NOTCH1	6.4E-11	5.5E-07
NOTCH2	5.0E-13	4.6E-09
NPTXR	7.1E-08	5.2E-04
NRP1	1.2E-22	1.2E-18
NTRK3	4.6E-09	3.7E-05
NUP205	5.7E-16	5.4E-12
NUP88	2.2E-16	2.1E-12
PDGFRB	6.4E-10	5.3E-06
PDK2	1.3E-06	8.7E-03
PDPK1	3.2E-11	2.8E-07
PIK3C3	6.9E-13	6.2E-09
PIK3R1	2.1E-10	1.8E-06
PIM1	7.5E-09	5.9E-05
PMS2	3.6E-07	2.5E-03
PRKD2	3.7E-26	3.7E-22
PRKDC	2.4E-11	2.1E-07
QKI	3.4E-25	3.4E-21
QSOX1	6.7E-09	5.3E-05
RB1	7.2E-14	6.6E-10
RHOT1	2.1E-08	1.6E-04
RIMBP2	2.1E-11	1.8E-07

RIMS2	1.1E-13	1.0E-09
RINT1	7.7E-09	6.0E-05
ROS1	5.9E-07	4.0E-03
RPN1	7.1E-23	7.0E-19
RSF1	7.2E-09	5.6E-05
RTN1	4.1E-81	4.2E-77
RUNX1T1	3.5E-08	2.6E-04
RYR2	1.5E-09	1.2E-05
SHMT2	2.4E-21	2.3E-17
SIRPA	1.3E-09	1.1E-05
SLC25A13	2.9E-43	3.0E-39
SLCO2B1	1.1E-08	8.5E-05
SMAD4	5.9E-25	5.8E-21
SMURF2	6.5E-07	4.4E-03
SNX13	1.1E-09	8.6E-06
SOX11	1.8E-28	1.8E-24
SPTBN1	8.1E-08	5.9E-04
ST7	4.7E-08	3.5E-04
STAT1	1.9E-13	1.8E-09
STAT3	1.2E-09	9.4E-06
STEAP3	2.7E-44	2.7E-40
SV2B	2.3E-28	2.4E-24
SYNE1	4.6E-08	3.4E-04
SYP	8.7E-09	6.8E-05
TAF1B	3.6E-10	3.0E-06
TBK1	1.1E-06	7.3E-03
TCF12	5.1E-14	4.7E-10
TGFBR2	5.2E-09	4.1E-05
TIMP2	3.9E-14	3.6E-10
TNFRSF11B	5.0E-07	3.4E-03
TNFSF4	6.4E-10	5.3E-06
TOP1	5.5E-08	4.0E-04
TP53	4.0E-33	4.1E-29
TRIM24	1.2E-09	1.0E-05
TRIM3	2.1E-14	2.0E-10
TRMT5	1.4E-07	9.8E-04
TRRAP	2.2E-07	1.5E-03
TSC1	2.8E-07	1.9E-03
UNG	2.9E-08	2.2E-04
XPO1	6.7E-07	4.5E-03
ZEB1	2.2E-07	1.6E-03

Table S2 Detailed results of pathway enrichment analysis of single-factor and multi-factor regulations

Single-factor regulations	No. genes	*p-value*	Whether Cancer-related
Pathways in cancer	55	4.3E-04	YES
MAPK signaling pathway	46	8.0E-04	YES
Regulation of actin cytoskeleton	42	9.4E-05	NO
Neuroactive ligand-receptor interaction	37	4.1E-02	NO
Calcium signaling pathway	36	1.3E-04	NO
Focal adhesion	33	1.0E-02	NO
Lysosome	27	1.5E-04	NO
Neurotrophin signaling pathway	21	3.0E-02	NO
Arrhythmogenic right ventricular cardiomyopathy (ARVC)	20	2.3E-04	NO
Oocyte meiosis	20	1.9E-02	NO
Hypertrophic cardiomyopathy (HCM)	18	6.0E-03	NO
Prostate cancer	18	1.0E-02	YES
Fc gamma R-mediated phagocytosis	18	1.9E-02	NO
Pancreatic cancer	17	2.0E-03	YES
Colorectal cancer	17	1.3E-02	YES
Dilated cardiomyopathy	17	2.9E-02	NO
Melanogenesis	16	4.9E-02	NO
Acute myeloid leukemia	15	2.0E-03	YES
Long-term potentiation	15	1.0E-02	NO
Cardiac muscle contraction	15	3.1E-02	NO
ECM-receptor interaction	15	4.4E-02	NO
Multi-factor regulations			
Pathways in cancer	59	2.8E-02	YES
MAPK signaling pathway	56	1.0E-03	YES
Focal adhesion	41	1.0E-02	NO
Regulation of actin cytoskeleton	39	4.8E-02	NO
Wnt signaling pathway	36	1.2E-03	YES
Endocytosis	34	4.7E-02	NO
Cell cycle	30	3.0E-03	YES
Axon guidance	30	5.0E-03	NO
Oocyte meiosis	26	7.9E-03	NO
Vascular smooth muscle contraction	26	1.0E-02	NO
Lysosome	25	3.0E-02	NO
Epithelial cell signaling in Helicobacter pylori infection	24	2.3E-05	NO
T cell receptor signaling pathway	24	2.2E-02	YES
GnRH signaling pathway	22	2.6E-02	YES

DNA replication	21	4.8E-09	YES
Vibrio cholerae infection	21	3.3E-05	NO
Apoptosis	21	1.4E-02	YES
Melanogenesis	21	4.5E-02	NO
Colorectal cancer	20	2.0E-02	YES
ErbB signaling pathway	20	2.8E-02	YES
ECM-receptor interaction	19	3.8E-02	NO

Table S3 Detailed information of genes involved in pathway enrichment analysis of the key regulation subtypes

Regulation subtype	Gene name
CNV/TF	
MAPK signaling pathway	MAPKAPK5, GNA12, DUSP10, MAPKAPK3, CACNB2, CACNB3, NFKB1, FGF12, MAPKAPK2, DAXX, AKT1, MAP3K7, MAP3K6, TNFRSF1A, PAK2, MAPT, PPP3CB, PPP3CA, PRKACB, RAPGEF2, EGFR, TP53, FLNC, FLNA, CDC25B, NRAS, RPS6KA2, RRAS2, MAPK9, PLA2G6, MAPK8, MAPK8IP1, DUSP8
Wnt signaling pathway	CSNK1A1, PPP2R1A, CTBP2, ROCK1, BTRC, CAMK2G, TP53, SMAD4, TCF7L2, MAP3K7, CTNNBIP1, SFRP4, PPP3CB, MAPK9, MAPK8, WIF1, CAMK2B, PPP3CA, PRKACB, RUVBL1, NFATC3, APC, LRP5
Cell cycle	CDC7, YWHAZ, TP53, SMAD4, PRKDC, TTK, ANAPC10, CDC20, CHEK2, MCM3, MCM5, ATM, CDC25B, HDAC2, MCM7, PLK1, PCNA, BUB3
DNA replication	SSBP1, LIG1, POLA2, MCM3, RNASEH2A, MCM5, RPA1, POLD3, RFC4, MCM7, RFC1, RFC2, POLE3, POLD1, POLD2, PCNA
Spliceosome	RBM22, NCBP2, PRPF31, NHP2L1, MAGOH, LSM7, TRA2A, PRPF4, BUD31, DDX23, DHX15, LSM5, HNRNPC, SNRPF, PRPF40A, THOC1
Axon guidance	GNAI3, ROCK1, LIMK2, ABLIM3, DPYSL2, CDK5, ITGB1, SLIT3, NRAS, PAK2, SEMA4F, NCK1, SRGAP3, PPP3CB, PPP3CA, NFATC3
Apoptosis	CFLAR, TP53, NFKB1, CAPN2, ATM, CAPN1, AKT1, PRKAR2B, TNFRSF1A, PRKAR2A, CASP7, PRKAR1B, PPP3CB, PPP3CA, PRKACB
T cell receptor signaling pathway	VAV3, MALT1, NFKB1, AKT1, MAP3K7, PRKCQ, NRAS, PAK2, PLCG1, NCK1, PPP3CB, MAPK9, PPP3CA, NFATC3
Vibrio cholerae infection	KDELR2, ATP6V1H, PDIA4, ATP6V1B2, ATP6V1D, ATP6V0C, SEC61B, PLCG1, ATP6V1E1, PRKACB, SEC61G, SEC61A1, SEC61A2
Mismatch repair	POLD3, RPA1, MSH6, RFC4, RFC1, SSBP1, RFC2, POLD1, LIG1, POLD2, PCNA, MLH1
Epithelial cell signaling in Helicobacter pylori infection	EGFR, ATP6V0C, PLCG1, ATP6V1E1, ADAM17, ATP6V1H, MAPK9, NFKB1, MAPK8, ATP6V1B2, CSK, ATP6V1D
MiRNA/TF	
Pathways in cancer	DVL2, COL4A2, BMP2, FGFR3, COL4A1, FGF9, FZD1, CDK2, FZD7, TCF7L1, MAPK1, ACVR1B, HSP90B1, HDAC2, ITGA6, VEGFA, LAMC1, MYC, TRAF4
Focal adhesion	TLN1, COL4A2, VAV3, COL4A1, PPP1CB, COL5A1, PAK6, MAPK1, PAK7, ITGA6, PAK3, ITGB8, VEGFA, SHC1, PAK1, LAMC1, THBS1
Regulation of	GIT1, VAV3, FGFR3, FGF9, WASF1, RDX, ARHGEF12, PPP1CB,

actin cytoskeleton	PAK6, MAPK1, PAK7, ITGA6, PAK3, TIAM1, ITGB8, ARPC5L, PAK1
MAPK signaling pathway	FGFR3, FGF9, MAP2K3, MAPKAPK2, RPS6KA5, MAP4K3, ACVR1B, MAPK1, PPP3CB, PRKACA, PAK1, RAPGEF2, DUSP8, MYC, DUSP6
Axon guidance	ABLIM1, NRP1, GNAI1, ARHGEF12, SLIT2, PAK6, PAK7, MAPK1, RND1, PAK3, SRGAP3, PPP3CB, PAK1
Wnt signaling pathway	DVL2, PPP2R1A, CAMK2G, BTRC, FZD1, TCF7L1, FZD7, PPP2CB, PPP3CB, PRKACA, TBL1X, PLCB1, MYC
Oocyte meiosis	MAPK1, PPP2R1A, CAMK2G, BTRC, PPP2CB, PPP3CB, FBXO5, PRKACA, ESPL1, PPP1CB, SMC3, CDK2
Chemokine signaling pathway	MAPK1, VAV3, LYN, GNAI1, TIAM1, GNB1, PRKACA, SHC1, CX3CL1, PAK1, PLCB1, CCNL2
Melanogenesis	DVL2, MAPK1, GNAI1, CAMK2G, CREB3L2, FZD1, PRKACA, PLCB1, TCF7L1, FZD7
Cell cycle	HDAC2, DBF4, PCNA, ESPL1, MCM2, MYC, SMC3, CDK2, WEE1, MCM6
Insulin signaling pathway	MAPK1, PRKAR2A, SOCS2, TSC1, PHKA1, PRKACA, SHC1, RPS6KB1, TRIP10, PPP1CB
TF/methylation	
Vascular smooth muscle contraction	MYL6, ADCY1, ADCY2, ADCY7, ARHGEF12, KCNMB1, ACTG2, MYL6B, PLA2G12A, PLA2G2A, PRKACB, CALCRL, PLA2G5, CALM1
Oocyte meiosis	ADCY1, ADCY2, CCNB2, ADCY7, RPS6KA2, PPP3CB, FBXO5, ESPL1, ANAPC10, PRKACB, SMC3, CALM1
Antigen processing and presentation	HSPA2, PSME1, PSME2, TAP1, IFI30, HLA-C, CD4, HLA-E, CANX, CD74, B2M
Dilated cardiomyopathy	LAMA2, ADCY1, ADCY2, ADCY7, ITGA6, ITGAV, LMNA, CACNB2, CACNG2, PRKACB, TPM4
Gap junction	ADCY1, ADCY2, ADCY7, GNAI1, GRB2, TUBB2A, TUBB6, PDGFD, PRKACB, TUBA1C
DNA replication	PRIM1, RFC2, LIG1, POLD2, MCM2, RNASEH2A
Type I diabetes mellitus	CD86, GAD2, CPE, PTPRN2, HLA-C, HLA-E

3. Supplementary Figures

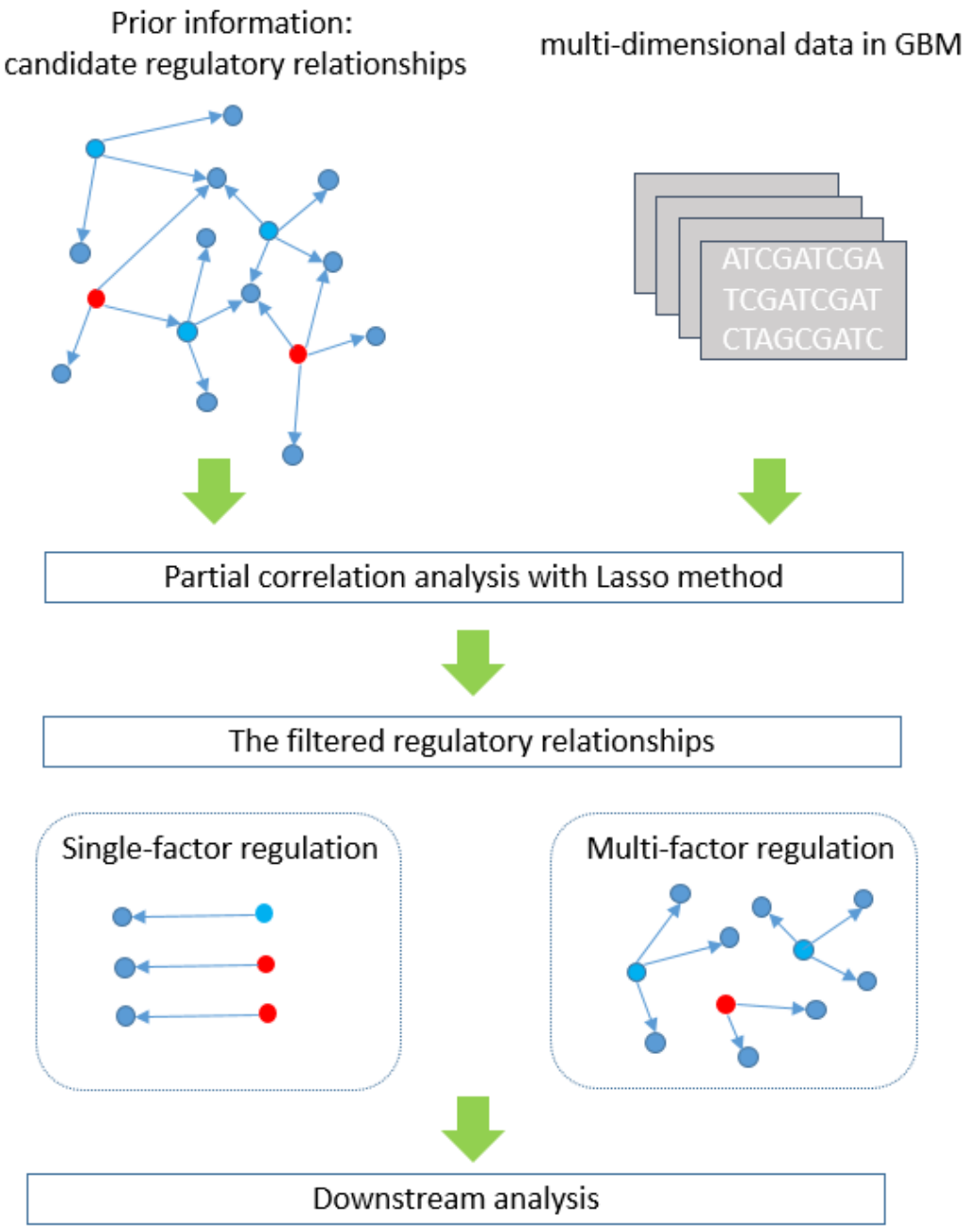


Figure S1 The flow chart of the filter process of prior regulatory relationships with partial correlation analysis with Lasso method.

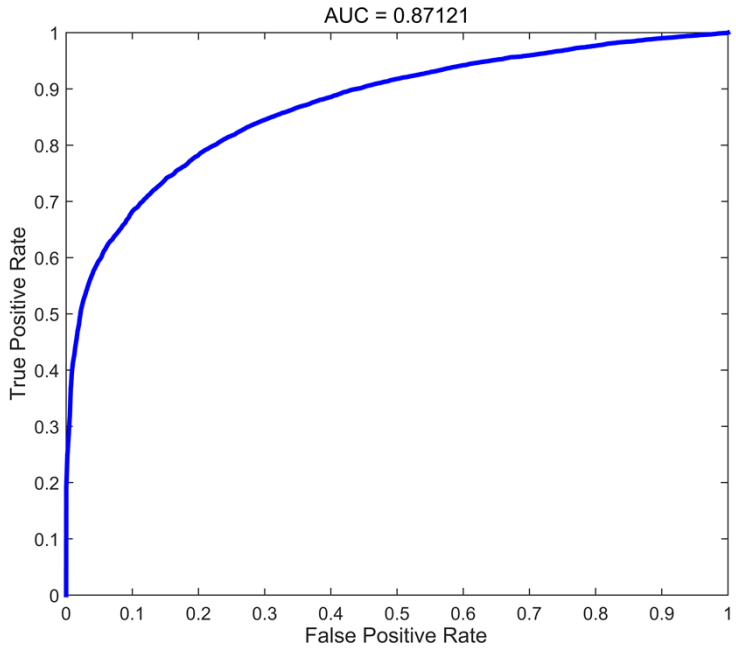


Figure S2 The ROC curve of the partial correlation analysis with Lasso method. X-axis and y-axis represent false positive rate and true positive rate, i.e. 1-Sp and Sn, respectively. The AUC value is labeled on the top of the figure.

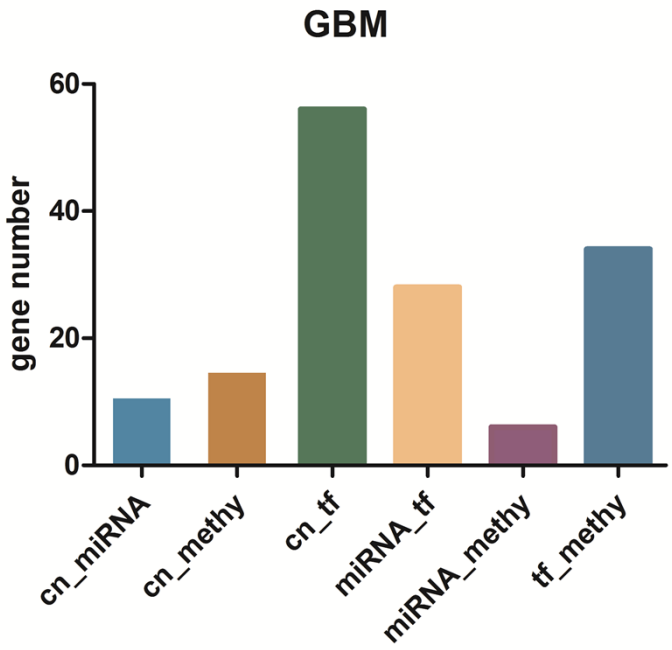


Figure S3 Distribution of regulation types of experimentally validated GBM-related genes appeared in multi-factor regulations

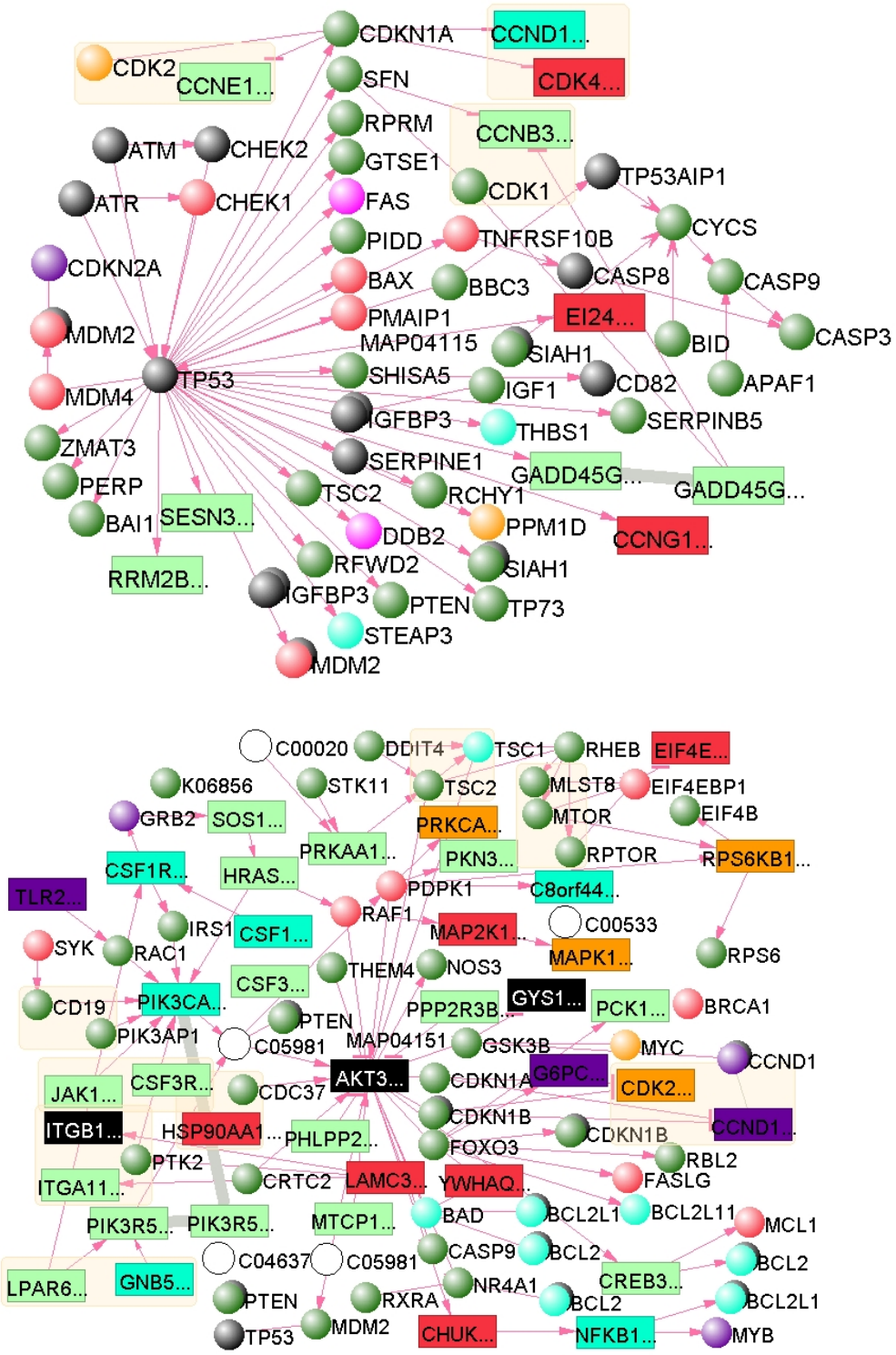


Figure S4 GBM-related P53 pathway (top) and PI3K-AKT pathway (bottom). The nodes are genes and rectangles are KEGG groups. Red, orange, black, purple, blue and pink represent single-factor, miRNA/TF, CNV/TF, TF/methy, CNV/methy and CNV/miRNA regulation, respectively. Here “methylation” and “copy number variation” is abbreviated to “methy” and “CNV” for convenience.