Supplementary Material

A multiple genomic data fused SF2 prediction model, signature identification and gene regulatory network inference for personalized radiotherapy

The details of MGPLS-UVE algorithm

VIP index in MGPLS algorithm

To be more convincing, we explored the *VIP* to calculate the importance of each gene to the response variable, which is the basis for selecting the signature genes.[1]

$$VIP=\sqrt{p\times(\boldsymbol{q}/\mathrm{sum}(\boldsymbol{s}))} \quad \text{(S1)}$$

where p is the number of genes in the training data set, and

$$\boldsymbol{s}=\mathrm{diag}(\boldsymbol{T}'\times\boldsymbol{T}\times\boldsymbol{Q}\times\boldsymbol{Q}') \quad \text{(S2)}$$

$$\boldsymbol{q}=\boldsymbol{s}'\times\boldsymbol{w} \quad \text{(S3)}$$

where the parameters $\boldsymbol{T}$, $\boldsymbol{Q}$,$\boldsymbol{w}$ are calculated through MGPLS, $\boldsymbol{w}$ is the unitized form of $\boldsymbol{W}$.

Cross-validation process and uninformative variable elimination (UVE)

In MGPLS regression, it is essential to determine the right complexity of the model, i.e., the number of latent variables (LVs). The bigger the number of LVs, the much easier the model is to overfit. The number of LVs can be optimized by using cross-validation.[2] In this paper, when $PRESS_h/RSS_{h-1} > 0.95^2$, we believe that the new component will not improve the accuracy of the model, so h at this time is the optimal number of LVs.

In order to improve the modeling accuracy, 10 times of 6-fold cross-validation method was used to the model training. To do this:

1) we randomly divided 60 samples into 6 groups, 5 of which were selected as the training sets and the remaining 1 group was the verification set.
2) Then the *VIP* value of each gene was calculated.
3) Repeated steps 1)-2) for 6 times until every group was used as the verification set for one and only one time.
4) We averaged the *VIP* values obtained through these 6 times of cross-validation and recorded them as $VIP_{i,j}$. $VIP_{i,j}$ is the *VIP* value for gene *i* in the *j*th round of cross-validation.
5) Repeated steps 1)-4) for 10 times and got the average value of $VIP_{i,j}$ (j=1...10) for each gene.
6) After sorting *VIP* values of all genes in descending order, UVE was performed (whose details are available below).
7) Repeated steps 1)-6) until the regression accuracy cannot be improved any more. The remaining genes were considered as signature genes.

In this paper, the number of genes removed in UVE process were different:

1) MGPLS rough selection procedure. According to the *VIP* value, we first removed 1 variable each time and repeated 22 times. As a result, 7600 variables were left. Then we removed 100 variables each time and repeated 71 times. Finally, 500 variables were left.

2) MGPLS fine selection procedure. According to the *VIP* value, we removed 1 variable each time and recorded the RMSE value of the model until all 500 variables were removed. Then the gene set with the lowest RMSE was considered as signature set.

Brief introduction of LASSO algorithm

Given predictors $\boldsymbol{x}_i$ and response values y_i for i=1, 2⋯, n, the optimization goal is to find regression coefficients $\boldsymbol{\beta}$ to minimize

$$\sum_{i=1}^{n}(y_i-\boldsymbol{\beta}'\boldsymbol{x}_i)^2+\lambda\|\boldsymbol{\beta}\|_1 \tag{S4}$$

where λ is the regularization parameter. Our GRN is a direct network that encodes the regulatory relationships among 113 signature genes. It is assumed that a gene can be directly regulated by other genes and a single CNV at most. In our case, response values denote the GE data of each gene; the matrix of GE data and CNV data of all genes are simultaneously used as predictors. After obtaining $\boldsymbol{\beta}$ using coordinate descent algorithm, ordinary least square algorithm was employed to re-estimate non-zero coefficient element of $\boldsymbol{\beta}$ to get final regression coefficients. *Cytoscape* toolkit was then used to visualize the obtained GRN.[3]

The details of sparse GRN inference

Let $\boldsymbol{E}\in\boldsymbol{R}_{113\times60}$ denote the matrix of GE data and $\boldsymbol{C}\in\boldsymbol{R}_{113\times60}$ denote the matrix of CNV data. $\boldsymbol{E}=[\boldsymbol{e}_1,\boldsymbol{e}_2,\dots\boldsymbol{e}_{113}]$ and $\boldsymbol{C}=[\boldsymbol{c}_1,\boldsymbol{c}_2,\dots\boldsymbol{c}_{113}]$ where $\boldsymbol{e}_i$, $\boldsymbol{c}_i$ are the ith row vector of matrix $\boldsymbol{E}$, $\boldsymbol{C}$ respectively. The GRN is defined as follows:

$$\boldsymbol{e}_i=\boldsymbol{b}_i\boldsymbol{E}+\boldsymbol{f}_i\boldsymbol{C}+\boldsymbol{\mu}_i+\boldsymbol{\varepsilon}_i \tag{S5}$$

where $\boldsymbol{b}_i,\boldsymbol{f}_i$ denotes ith row vector of adjacency matrix $\boldsymbol{B}\in\boldsymbol{R}_{113\times113}$, $\boldsymbol{F}\in\boldsymbol{R}_{113\times113}$ respectively. The element b_{ij} represents the activation (positive) or deactivation (negative) weight of edge from jth gene to ith gene; $\boldsymbol{\mu}_i$ is a model bias that can be removed by mean centering; and $\boldsymbol{\varepsilon}_i$ is a residual. Our goal is to estimate row vectors $\boldsymbol{b}_i,\boldsymbol{f}_i$ that minimize $\boldsymbol{\varepsilon}_i$.

(S5) can be rewritten in a least square minimization problem as:

$$\min_{\boldsymbol{b}_i,\boldsymbol{f}_i}\|\boldsymbol{e}_i-\boldsymbol{b}_i\boldsymbol{E}-\boldsymbol{f}_i\boldsymbol{C}\|_2^2 \tag{S6}$$

where $||\cdot||_2$ denotes 2 norm.

In order to obtain sparse model and avoid the overfitting, we add L1 regularization term to (S6) to make it a LASSO regression form as follow:

$$\min_{\boldsymbol{b}_i,\boldsymbol{f}_i}\|\boldsymbol{e}_i-\boldsymbol{b}_i\boldsymbol{E}-\boldsymbol{f}_i\boldsymbol{C}\|_2^2+\lambda_1\|\boldsymbol{b}_i\|_1+\lambda_2\|\boldsymbol{f}_i\|_1 \tag{S7}$$

where λs are penalty coefficients.

There are two hypothesizes in the model:

(1) There is no self-regulation, i.e., the diagonal elements of the $\boldsymbol{B}$ matrix are all zero.

(2) A gene can be directly regulated by only CNV that belong to the gene but no other genes, i.e., only diagonal elements of $\boldsymbol{F}$ matrix can be non-zero.

Based on these two hypothesizes, (S6) can be rewritten as follow:

$$L(\boldsymbol{\beta}_i)=\min_{\boldsymbol{\beta}_i}\|\boldsymbol{e}_i-\boldsymbol{\beta}_i\boldsymbol{Y}\|_2^2+\lambda\|\boldsymbol{\beta}_i\|_1 \tag{S8}$$

where $\boldsymbol{\beta}_i = [b_{i1}, b_{i2}, \dots\dots, b_{ii-1}, b_{ii+1}, \dots\dots, b_{i113}, f_{ii}]$

$\boldsymbol{Y} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots\dots, \boldsymbol{e}_{i-1}, \boldsymbol{e}_{i+1}, \dots\dots, \boldsymbol{e}_{113}, \boldsymbol{c}_i]$.

Given λ, the optimal $\boldsymbol{\beta}_i$ can be find thro ugh using coordinate descent algorithm to (S8).

$$\frac{\partial L}{\partial \beta_{ij}} = -\boldsymbol{y}_i \left(\boldsymbol{e}_i^{\mathrm{T}} - \boldsymbol{Y}_{(-j)}^{\mathrm{T}} \boldsymbol{\beta}_{i(-j)}^{\mathrm{T}} - \boldsymbol{y}_j^{\mathrm{T}} \beta_{ij} \right) + \lambda \partial_{\beta_{ij}} \|\boldsymbol{\beta}_i\|_1 \tag{S9}$$

where $\boldsymbol{Y}_{(-j)}$ denotes matrix $\boldsymbol{Y}$ whose *j*th row is removed, and $\boldsymbol{y}_j$ denotes the *j*th row vector of $\boldsymbol{Y}$. Then (S9) can be rewritten as:

$$\frac{\partial L}{\partial \beta_{ij}} = -c_{ij} + a_{ij}\beta_{ij} + \lambda \partial_{\beta_{ij}} \|\boldsymbol{\beta}_i\|_1 \tag{S10}$$

where $C_{ij} = \boldsymbol{y}_j(\boldsymbol{e}_i^{\mathrm{T}} - \boldsymbol{Y}_{(-j)}^{\mathrm{T}} \boldsymbol{\beta}_{(-j)}^{\mathrm{T}})$, $a_{ij} = \boldsymbol{y}_j \boldsymbol{y}_j^{\mathrm{T}}$. Then $\boldsymbol{\beta}_i$ can be calculated as follow:

$$\beta_{ij} = \begin{cases} \dfrac{(c_{ij} + \lambda)}{a_{ij}} & c_{ij} < -\lambda \\ 0 & |c_{ij}| \le \lambda \\ \dfrac{(c_{ij} - \lambda)}{a_{ij}} & c_{ij} > \lambda \end{cases} \tag{S11}$$

The overall procedure to construct sparse GRN is described in Table S1.

Table S1. Steps to construct sparse GRN

Sparse GRN algorithm
procedure SGRN ($\boldsymbol{e}_i, \boldsymbol{Y}, \lambda, \varepsilon$)
initialize $\boldsymbol{\beta}_i$
while error > ε do
$\boldsymbol{\beta}_i^{old} = \boldsymbol{\beta}_i$
for *j*=1:m
Update $\boldsymbol{\beta}_{ij}$ via (S11)
end for
error=$\|\|\boldsymbol{\beta}_i^{old} - \boldsymbol{\beta}_i\|\|_2$
end while
return $\boldsymbol{\beta}_i$
end procedure

After obtaining adjacency matrices $\boldsymbol{B}$ and $\boldsymbol{F}$, the genes with absolute values greater than 0.1 in $\boldsymbol{B}$ and $\boldsymbol{F}$ were selected, and prepared for the next least squares regression. Because the coefficients obtained by the above method are only used to select genes, they are not the final regression coefficients. For a gene g_i, other genes whose absolute values of regression coefficients were greater than 0.1 were selected as regulatory genes of g_i.

Table S2. 113 signature genes selected by MGPLS-UVE algorithm

Serial number	Gene name	Entrez gene id	Chromosome	Cytoband	Regression coefficient	Data type
	constant				0.5472	
1	YY1AP1	55249	1	1q22	-0.0617	CNV
2	INPP5A	3632	10	10q26.3	0.0616	CNV
3	DAP3	7818	1	1q22	-0.0593	CNV
4	GON4L	54856	1	1q22	-0.0567	CNV
5	JTB	10899	1	1q21	-0.0467	CNV
6	TM9SF4	9777	20	20q11.21	-0.0171	GE
7	NECAP2	55707	1	1p36.13	-0.0167	GE
8	RBBP9	10741	20	20p11.2	-0.0151	GE
9	YBX3	8531	12	12p13.1	-0.0155	GE
10	NISCH	11188	3	3p21.1	-0.0151	GE
11	SNRNP35	11066	12	12q24.31	0.0147	GE
12	CCDC130	81576	19	19p13.2	-0.0145	GE
13	HGH1	51236	8	8q24.3	-0.0144	GE
14	RBBP4	5928	1	1p35.1	-0.0142	GE
15	TGS1	96764	8	8q11	-0.0138	GE
16	WRAP73	49856	1	1p36.3	-0.0135	GE
17	VPS4B	9525	18	18q21.33	0.0131	GE
18	DCTN6	10671	8	8p12-p11	0.0129	GE
19	PSMG4	389362	6	6p25.2	0.0127	GE
20	TYK2	7297	19	19p13.2	-0.0127	GE
21	BAHD1	22893	15	15q15.1	0.0127	GE
22	NOP2	4839	12	12p13	-0.0122	GE
23	MIPEP	650794	13	13q12.11	0.0121	GE
24	KIAA1468	57614	18	18q21.33	0.0119	GE
25	PRCC	5546	1	1q21.1	-0.0118	GE
26	CCDC174	51244	3	3p25.1	-0.0118	GE
27	HAUS1	115106	18	18q21.1	0.0118	GE
28	PRR14	78994	16	16p11.2	-0.0118	GE
29	ATP5A1	498	18	18q21	0.0114	GE
30	SLC7A1	56301	19	19q13.1	0.0111	GE
31	SMARCC1	6599	3	3p21.31	-0.0110	GE
32	RPS6KB2	6199	11	11q13.2	-0.0107	GE
33	RNH1	6050	11	11p15.5	0.0102	GE
34	TMEM185B	79134	2	2q14.2	0.0099	GE
35	TTC8	123016	14	14q31.3	0.0098	GE
36	LSM6	11157	4	4q31.22	-0.0095	GE
37	RAD23B	5887	9	9q31.2	0.0094	GE
38	ANP32B	10541	9	9q22.32	0.0093	GE
39	RAD54L	8438	1	1p32	-0.0089	GE
40	PYCR2	29920	1	1q42.12	-0.0089	GE

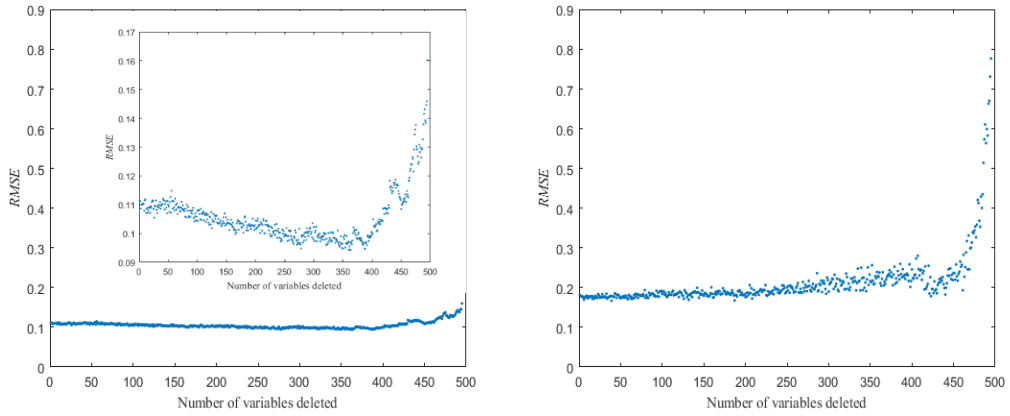
Table S2. 113 signature genes selected by MGPLS-UVE algorithm (continued)

41	TRIM28	10155	19	19q13.4	0.0086	GE
42	NDUFV1	4723	11	11q13	-0.0086	GE
43	TRMT44	152992	4	4p16.1	-0.0084	GE
44	SRRM1	10250	1	1p36.11	-0.0082	GE
45	PRPF19	27339	11	11q12.2	-0.0081	GE
46	NOP56	10528	20	20p13	-0.0081	GE
47	WDR4	10785	21	21q22.3	-0.0080	GE
48	ANAPC4	29945	4	4p15.2	-0.0079	GE
49	TUBGCP6	85378	22	22q13.31-q13.33	-0.0077	GE
50	HNRNPDL	9987	4	4q21.22	0.0077	GE
51	CTCF	10664	16	16q21-q22.3	0.0075	GE
52	C17orf62	79415	17	17q25.3	-0.0075	GE
53	TVP23B	51030	17	17p11.2	0.0074	GE
54	ZBED4	9889	22	22q13.33	-0.0070	GE
55	DDB1	1642	11	11q12-q13	-0.0070	GE
56	EWSR1	2130	22	22q12.2	0.0069	GE
57	LRCH1	23143	13	13q14.11	0.0068	GE
58	EXOSC7	23016	3	3p21.31	-0.0068	GE
59	FARSA	2193	19	19p13.2	-0.0068	GE
60	DENND4B	9909	1	1q21	-0.0066	GE
61	MED28	80306	4	4p16	-0.0066	GE
62	NAF1	92345	4	4q32.2	-0.0065	GE
63	CNOT10	25904	3	3p22.3	-0.0061	GE
64	GCDH	2639	19	19p13.2	-0.0060	GE
65	TRIO	11078	22	22q13.1	0.0056	GE
66	PPAT	5471	4	4q12	-0.0054	GE
67	RPL34	6164	4	4q25	-0.0053	GE
68	WBP4	11193	13	13q14.11	0.0053	GE
69	CTTNBP2NL	55917	1	1p13.2	0.0053	GE
70	BCLAF1	9774	6	6q22-q23	0.0053	GE
71	INTS5	80789	11	11q12.3	-0.0052	GE
72	ZNF764	92595	16	16p11.2	-0.0047	GE
73	UBIAD1	29914	1	1p36.22	-0.0047	GE
74	RAN	5901	12	12q24.3	0.0046	GE
75	RPUSD2	27079	15	15q13.3	0.0045	GE
76	CLN5	1203	13	13q21.1-q32	0.0041	GE
77	GAR1	54433	4	4q25	-0.0041	GE
78	ZBTB39	9880	12	12q13.3	-0.0039	GE
79	PMS2P1	5379	7	7q22.1	-0.0038	GE
80	SMARCAD1	56916	4	4q22-q23	-0.0037	GE
81	BLM	641	15	15q26.1	-0.0033	GE
82	USP7	7874	16	16p13.3	0.0032	GE
83	RNF138	51444	18	18q12.1	0.0029	GE

Table S2. 113 signature genes selected by MGPLS-UVE algorithm (continued)

84	RPL9	6133	4	4p13	-0.0028	GE
85	RIOK1	83732	6	6p24.3	0.0027	GE
86	MYB	4602	6	6q22-q23	-0.0026	GE
87	SNX7	51375	1	1p21.3	0.0025	GE
88	METTL14	57721	4	4q26	0.0023	GE
89	PAICS	10606	4	4q12	-0.0023	GE
90	NAT10	55226	11	11p13	-0.0022	GE
91	TAF11	6882	6	6p21.31	-0.0022	GE
92	POLD1	5424	19	19q13.3	0.0022	GE
93	SHPRH	257218	6	6q24.3	-0.0021	GE
94	NFATC3	4775	16	16q22.2	0.0018	GE
95	PMS2P3	5387	7	7q11.23	-0.0016	GE
96	PDCD2	5134	6	6q27	0.0016	GE
97	TBP	6908	6	6q27	-0.0014	GE
98	NOP14	8602	4	4p16.3	0.0014	GE
99	YWHAZ	7534	8	8q23.1	0.0012	GE
100	SNRPD1	6632	18	18q11.2	0.0011	GE
101	PTK2	5747	8	8q24.3	0.0010	GE
102	CLNS1A	1207	11	11q13.5-q14	-0.0009	GE
103	CENPC	1060	4	4q13.2	-0.0009	GE
104	ABHD18	80167	4	4q28.2	-0.0009	GE
105	MCM3	4172	6	6p12	0.0007	GE
106	MRPL16	54948	11	11q12.1	-0.0007	GE
107	MCM7	4176	7	7q21.3-q22.1	-0.0005	GE
108	WDR74	54663	11	11q12.3	0.0005	GE
109	COMMD6	170622	13	13q22	0.0005	GE
110	ABCE1	6059	4	4q31	-0.0003	GE
111	ENOPH1	58478	4	4q21.22	-0.0002	GE
112	KDELR2	11014	7	7p22.1	0.0001	GE
113	MRPL1	29088	8	8q11.2-q13	-0.00001	GE

*Gene sorted by the absolute value of the regression coefficient. 24 “Hub” genes are highlighted.



(a) (b)

Figure S1. The RMSEs obtained by different number of genes. (a) RMSE trend using CNV and GE data; (b) RMSE trend using CNV, GE and ME data. When we delete unimportant genes, the prediction accuracy will increase and the RMSE value will decrease; conversely, the RMSE value will increase. Therefore, as the number of genes decreases, the value of RMSE will first decrease and then increase.

GO and KEGG Pathway Enrichment Analysis Result

GO and KEGG pathway analysis were performed for 113 signature genes (Figure S2). For Biological process, signature genes were mostly enriched in cellular nitrogen compound metabolic process, heterocycle metabolic process and organic cyclic compound metabolic process (Figure S2a). For molecular function, signature genes were mostly involved in binding, protein binding, heterocycle metabolic binding and organic cyclic compound binding (Figure S2b). For cellular component, signature genes were mostly associated with intracellular, intracellular part and intracellular organelle (Figure S2c). GO secondary classification map can be seen in Figure S3. In addition, 113 signature genes were enriched in 111 KEGG pathways, of which ribosome biogenesis in eukaryotes, DNA replication, homologous recombination and so on were highly significant. The results showed that more signature genes contribute to tumorigenesis and progression mostly through involvement in translation, DNA replication and repair, signal transduction, and cell growth and death (Figure S2d).

Figure S2. GO and KEGG analysis of 113 signature genes. (a) Top 20 of GO enrichment in Biological Process; (b) Top 20 of GO in Molecular Function; (c) Top 20 of GO enrichment in Cellular Component; (d) Top 20 of KEGG enrichment and KEGG pathway number chart. All terms are sorted in ascending p-values.

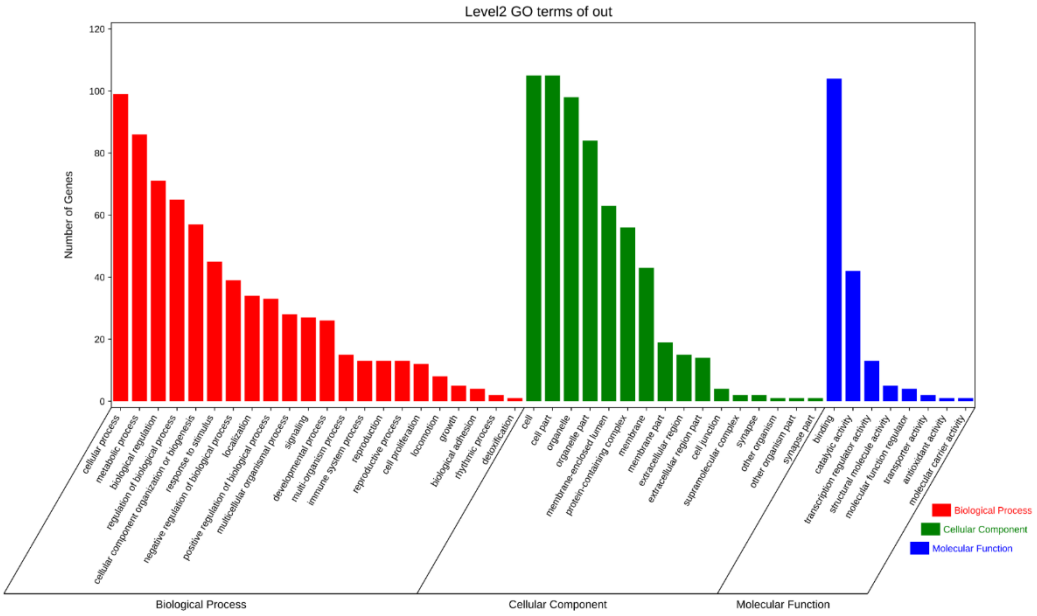


Figure S3. GO secondary classification map, which shows the number and enrichment condition of 113 genes in each GO term. Since a gene often corresponds to multiple GO terms, the same gene will appear under different classification items. In other words, it will be counted multiple times. If the number of genes of all the columns is added up, therefore, the value will be more than 113.

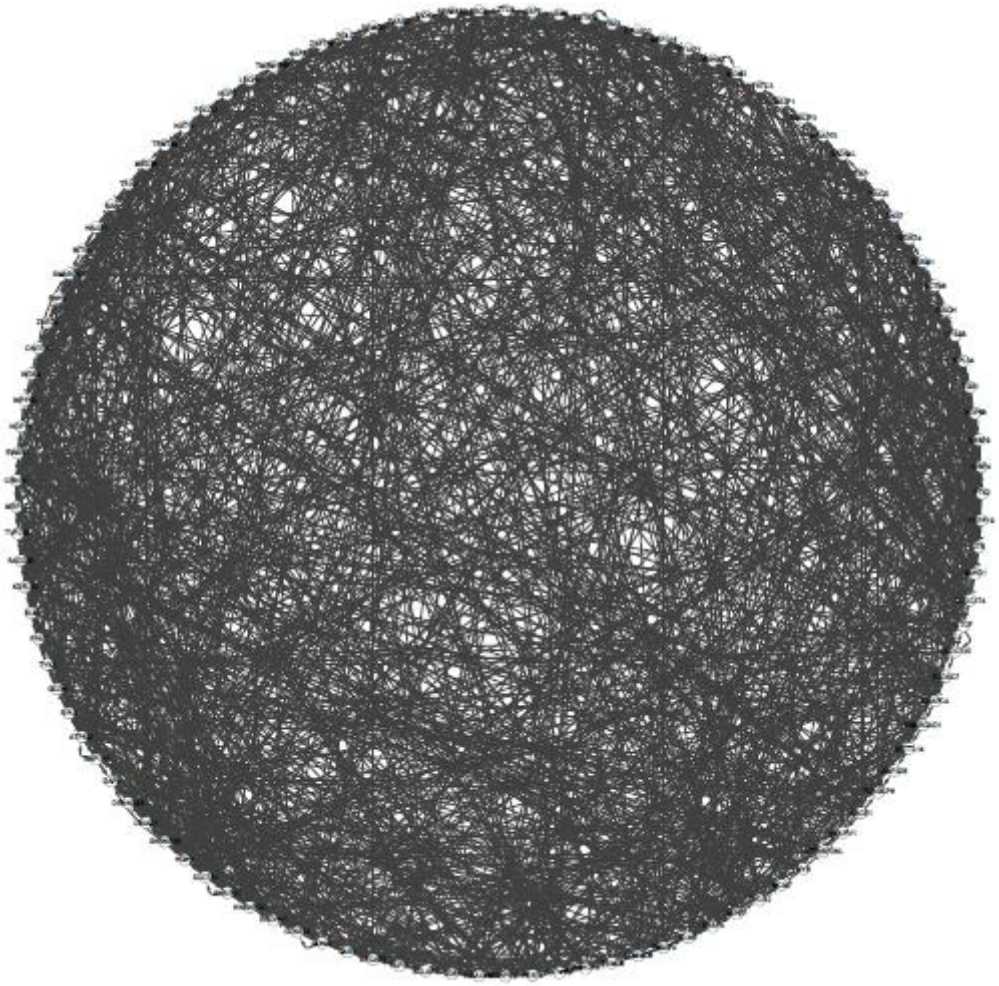


Figure S4. Gene regulatory network of all 113 genes (No further beautification).

Table S3. Pearson correlation coefficient of 12 genes

Gene	coefficient	*P* value	GE average value	GE standard deviation	CNV average value	CNV standard deviation
ATP5A1	0.76	1.97e-12	-0.01	0.88	-0.16	0.27
BLM	0.58	1.21e-6	-0.01	0.92	0.07	0.20
CLNS1A	0.75	4.26e-12	-0.01	0.87	0.08	0.30
EWSR1	0.53	1.12e-5	-0.004	0.86	-0.02	0.21
MCM3	0.56	2.68e-6	-0.006	0.92	0.05	0.20
MIPEP	0.65	1.89e-8	-0.01	0.91	-0.09	0.24
MYB	0.54	8.28e-6	-0.005	0.93	-0.03	0.27
PDCD2	0.58	1.03e-6	-0.02	0.89	-0.06	0.18
RPL9	0.59	8.06e-7	-0.02	0.89	-0.08	0.14
RPL34	0.73	4.77e-11	-0.01	0.86	-0.12	0.16
SNRPD1	0.60	5.28e-7	-0.01	0.87	-0.06	0.20
SRRM1	0.53	1.37e-5	-0.002	0.85	0.03	0.19

* Sample size is 60.

References

1. Mehmood T, Warringer J, Snipen L. Improving stability and understandability of genotype-phenotype mapping in Saccharomyces using regularized variable selectioninL-PLSregressio. *BMC Bioinformatics.* 2012;13(1):1-13.
2. Shao J. Linear model selection by cross-validation. *J Am Stat Assoc.* 1993;88(422):486-494.
3. Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* 2003;13(11):2498-2504.