Supplementary Materials

DGPathinter: a novel model for identifying driver genes via knowledge-driven matrix factorization with prior knowledge from interactome and pathways

Jianing Xi,[1,†] Minghui Wang[1,2,†] and Ao Li[1,2,*]

[a] School of Information Science and Technology, University of Science and Technology of China, Hefei AH230027, China.
[b] Centers for Biomedical Engineering, University of Science and Technology of China, Hefei AH230027, China.
*E-mail: aoli@ustc.edu.cn
† These authors contributed equally to this work.

Supplementary Information

List of Tables

List of Figures

1 Supplementary Information

1.1 delta score of HotNet2

The delta score of HotNet2 mentioned in our paper is the minimum edge weight parameter mentioned in subsection "1.4.2 Minimum edge weight" in the supplementary information of HotNet2 [3]. When different minimum edge weight parameters are chosen, different sets of genes that are strongly connected in the network will be selected as candidate driver genes. As shown in Supplementary Figure 26(d) in the supplementary information of HotNet2 [3], ROC curves are drawn according to the genes ranked by the scores of the minimum edge weight parameter delta. Therefore, by following HotNet2 [3], we use the delta scores (the minimum edge weight parameter delta) of genes to rank the genes.

1.2 Statistical validation on the results with STRING

When we further apply the statistical validation on the detection results of these competing methods when STRING is used, the validation results of the Friedman test demonstrate that the difference between the detection results of the investigated methods is statistically significant. For example, the p-values of the Friedman test on the AUCs of precision-recall curves (Supplementary Table S6) are 0.02 for NCG4.0 and 0.01 for CGC. For the average ranks of known NCG4.0 and CGC and genes in the detection results with STRING, the p-values are 0.01 and 0.01 respectively. When the Friedman test is applied on the proportions of known NCG4.0 and CGC genes in the top 100 genes, we yield p-values of 0.02 and 0.02 respectively. For the proportions of known cancer genes in the top 200 and 300 genes, the related p-values of the Friedman test are 0.02 (for NCG4.0, top 200), 0.02 (for CGC, top 200), 0.02 (for NCG4.0 top 300) and 0.02 (for CGC, top 300).

Table S1. The areas under the curve (AUCs) of precision-recall curves of the detection results of the competing methods on BRCA, GBM and THCA datasets, when evaluated by experimentally supported CGC genes [1] or statistically inferred NCG genes [2].

AUCs (CGC)	BRCA	GBM	THCA
DGPathinter	15.56%	26.19%	5.53%
HotNet2	0.39%	7.76%	0.43%
NBS	5.25%	15.81%	1.02%
MUFFINN-DNmax	0.29%	0.19%	0.19%
MUFFINN-DNsum	0.23%	0.46%	0.18%
AUCs (NCG4.0)	BRCA	GBM	THCA
DGPathinter	12.19%	12.54%	5.18%
HotNet2	2.98%	4.93%	0.43%
NBS	7.36%	5.08%	1.01%
MUFFINN-DNmax	2.84%	0.63%	0.20%
MUFFINN-DNsum	3.07%	0.69%	0.20%

Table S2. The top 100 genes prioritized by DGPathinter on BRCA dataset. Their ranks, gene symbols, mutation frequencies and whether the genes are curated as experimentally supported CGC genes [1] or statistically inferred NCG genes [2] are shown in the table, along with whether the genes are also detected by other competing methods.

Rank	GeneSymbol	MutFreq	Database	OtherMethods	Rank	GeneSymbol	MutFreq	Database	OtherMethods
1	PIK3CA	208	CGC & NCG4.0	NBS	51	SRRM2	11		
2	TP53	198	CGC & NCG4.0	NBS	52	PTPRB	11		
3	TTN	116		NBS	53	MYLK	10		
4	GATA3	60	CGC & NCG4.0	HotNet2;NBS	54	SSPO	22		
5	MAP3K1	67	CGC & NCG4.0	NBS	55	DYRK1A	8		
6	MUC16	62	NCG4.0		56	GCN1L1	9		
7	USH2A	34	NCG4.0		57	MED12	12		
8	RYR2	37	NCG4.0		58	CACNA1B	14		
9	SYNE1	24			59	TACC2	12	NCG4.0	
10	DST	16			60	FAM171A1	10		
11	MUC4	33			61	MYH7	11		
12	CDH1	34	CGC & NCG4.0		62	NF1	16	NCG4.0	
13	APOB	20		HotNet2	63	HECW1	13		
14	RYR3	24			64	KIAA1109	11		
15	BRCA2	25	CGC & NCG4.0		65	MGA	12		
16	BRCA1	21	NCG4.0		66	DNAH3	15		
17	LRP2	21			67	CROCC	15		
18	SPEN	24			68	PTEN	19	NCG4.0	
19	DNAH5	16			69	LAMA1	13	NCG4.0	
20	MDN1	16			70	SCN2A	10		
21	CSMD1	29	NCG4.0		71	CHD6	10		
22	FLG	30			72	MGAM	15		
23	OBSCN	33			73	TBL1XR1	12	NCG4.0	
24	SPTA1	19			74	TLN1	10	NCG4.0	NBS
25	COL12A1	13			75	ATM	22		
26	SYNE2	22			76	GRIN2B	12		NBS
27	MAP1A	15			77	ERBB3	9	NCG4.0	NBS
28	PLXNA4	12			78	NCOR1	19	CGC & NCG4.0	
29	AFF2	15	NCG4.0		79	LRP1	13		
30	XIRP2	17			80	WDR7	8		
31	AKT1	13	CGC & NCG4.0		81	ASPM	10		
32	COL1A1	8	NCG4.0		82	ERCC6	8		
33	PCNT	11			83	HUWE1	10		
34	DMD	19		HotNet2	84	TLR4	8	NCG4.0	
35	ACTN2	12		MUFFINN-DNmax	85	FBN1	13		
36	HSPG2	11			86	ITGAV	7		
37	WDFY3	13			87	NEB	19		
38	UTRN	11			88	ITPR2	7		
39	BRWD1	10			89	AKAP9	11	NCG4.0	
40	UBR4	18			90	ANK3	12		HotNet2
41	MUC12	12			91	MAP2	9		
42	DNAH7	11			92	SPI1	17		
43	RUNX1	19	NCG4.0		93	F5	12		HotNet2
44	MAP2K4	22	CGC & NCG4.0		94	TG	15	NCG4.0	
45	RELN	17			95	GOLGA4	8	NCG4.0	
46	RYR1	17			96	ABCA4	7		HotNet2
47	TBX3	14	CGC & NCG4.0		97	PREX2	13		
48	EPG5	9	NCG4.0		98	NFATC4	8		
49	MLLT4	14			99	SRCAP	12		
50	MACF1	12			100	PLCL2	8		

Table S3. The top 100 genes prioritized by DGPathinter on GBM dataset. Their ranks, gene symbols, mutation frequencies and whether the genes are curated as experimentally supported CGC genes [1] or statistically inferred NCG genes [2] are shown in the table, along with whether the genes are also detected by other competing methods.

Rank	GeneSymbol	MutFreq	Database	OtherMethods	Rank	GeneSymbol	MutFreq	Database	OtherMethods
1	TP53	31	CGC & NCG4.0	HotNet2;NBS	51	ZNF384	1		
2	PTEN	29	CGC & NCG4.0	HotNet2;NBS	52	BAI3	2		
3	ERBB2	7	NCG4.0	HotNet2;NBS	53	PIK3C2B	1		HotNet2
4	NF1	13	CGC & NCG4.0	NBS	54	AGAP2	2		
5	EGFR	15	CGC & NCG4.0	HotNet2;NBS	55	CHEK1	1		MUFFINN-DNsum
6	PIK3R1	9	CGC & NCG4.0	HotNet2;NBS	56	CAST	1		HotNet2;MUFFINN-DNsum
7	PIK3CA	6	CGC & NCG4.0	HotNet2;NBS	57	STAT1	1		NBS
8	RB1	9	NCG4.0	HotNet2;NBS	58	KRAS	1		HotNet2;MUFFINN-DNsum
9	DST	6			59	MDM4	1	CGC & NCG4.0	MUFFINN-DNmax;MUFFINN-DNsum
10	CDKN2A	3	NCG4.0	HotNet2	60	NBN	1		
11	FN1	3		HotNet2	61	PPP2R5D	1		HotNet2
12	TNK2	3		HotNet2	62	PAX5	1		HotNet2;NBS
13	CHEK2	4		HotNet2;NBS	63	PRKCB	1		
14	BCAR1	2		HotNet2	64	TAF1	2		HotNet2;NBS
15	COL1A1	2		HotNet2	65	PRKDC	2		
16	TNC	3		HotNet2	66	CCNB1	1		MUFFINN-DNsum
17	MSH6	4		HotNet2	67	PTPN11	2		
18	MLH1	2			68	PDK2	1		
19	FURIN	2		HotNet2	69	TRIM24	1		
20	EP400	2			70	TERT	2	CGC & NCG4.0	
21	PDGFRB	2		HotNet2;NBS	71	TRRAP	2		
22	ASPM	1		HotNet2	72	SHH	1		HotNet2;NBS
23	ING4	2		HotNet2	73	CTSH	1		MUFFINN-DNsum
24	MSH2	2		HotNet2;NBS	74	FLT4	1		NBS;MUFFINN-DNsum
25	ATM	1			75	NMBR	2		
26	IRS1	2		HotNet2;NBS	76	CYP3A4	1		MUFFINN-DNsum
27	MAPK9	2		HotNet2	77	RPS6KA3	1		NBS
28	BRCA2	3		HotNet2;NBS	78	GRIA2	1		MUFFINN-DNsum
29	EP300	3			79	WNT2	1		MUFFINN-DNmax;MUFFINN-DNsum
30	HSP90AA1	2		NBS	80	APBB1IP	1		MUFFINN-DNsum
31	CBL	1			81	PHLPP1	1		
32	PRKCZ	2			82	GLI1	1		HotNet2;MUFFINN-DNsum
33	NOTCH1	2		HotNet2	83	JAG1	1		HotNet2;MUFFINN-DNsum
34	KLF4	1		HotNet2	84	COL3A1	2	NCG4.0	HotNet2
35	ROS1	2	CGC & NCG4.0	HotNet2	85	RTN1	1		
36	MET	2		HotNet2	86	IL1RL1	1		NBS;MUFFINN-DNsum
37	KLF6	2	CGC & NCG4.0	HotNet2	87	LAMP1	2		
38	TCF12	2			88	POU2F1	1		
39	FBXW7	2		HotNet2;NBS	89	RPN1	1		
40	ITGB3	3		HotNet2	90	PRKD2	1		MUFFINN-DNmax;MUFFINN-DNsum
41	NOS3	1			91	SLC12A6	2		
42	TSC2	2			92	ST7	1		
43	LDHA	2			93	CENPF	2		
44	IQGAP1	2			94	NTRK3	1		
45	LTF	2			95	DTX3	1		NBS;MUFFINN-DNsum
46	GATA3	2			96	SIK2	2		
47	EPHA7	2			97	NDUFA10	1		
48	ROR2	2			98	MYLK2	1		
49	MSI1	2			99	KAT6B	2		
50	BCL11A	4			100	SYP	1		NBS

Table S4. The top 100 genes prioritized by DGPathinter on THCA dataset. Their ranks, gene symbols, mutation frequencies and whether the genes are curated as experimentally supported CGC genes [1] or statistically inferred NCG genes [2] are shown in the table, along with whether the genes are also detected by other competing methods.

Rank	GeneSymbol	MutFreq	Database	OtherMethods	Rank	GeneSymbol	MutFreq	Database	OtherMethods
1	BRAF	241	CGC & NCG4.0	HotNet2;NBS	51	PTPRD	5		
2	NRAS	34	CGC & NCG4.0	NBS	52	AXIN1	3		
3	TG	15		NBS	53	TSHR	3	CGC & NCG4.0	
4	HRAS	14		HotNet2;NBS	54	EZH1	2		
5	MUC16	21		NBS	55	IRF5	2		
6	CHEK2	5			56	DNMT3A	3		
7	EIF1AX	6			57	TJP1	2		
8	TTN	17		NBS	58	LAMA4	4		
9	CACNA1B	4		HotNet2	59	XPOT	3		
10	ATM	5		NBS	60	LCT	2		
11	TERT	3			61	BDP1	5		
12	OCRL	1			62	CPE	2		
13	ZXDA	1			63	DHX30	2		
14	MSI1	3			64	WDR33	3		
15	KRAS	4	CGC & NCG4.0	NBS	65	VWF	2		
16	COL20A1	4			66	ECSIT	2		
17	SLC12A4	5			67	PRKD1	3		
18	MYH13	3			68	IK	2		
19	CSMD1	4			69	THOC1	2		
20	ZFHX3	9		NBS	70	TMEM38B	2		
21	FBN1	3			71	NEB	2		
22	COL5A1	5		HotNet2	72	RIMS2	2		
23	OTUD4	8			73	ASPSCR1	2		
24	CFTR	2			74	STAT4	2		
25	LAMA5	2			75	MYH1	3		
26	RYR2	3			76	GTPBP4	4		
27	COL4A5	3			77	COL3A1	3		
28	COL4A1	3			78	PLXNB3	3		
29	PCNT	2			79	GLI2	3		HotNet2
30	PLEKHA5	2			80	DGCR8	2		
31	MECP2	2			81	GJA8	2		
32	MEPCE	2			82	NFAT5	4		
33	CTNND2	3		HotNet2	83	RALGAPA2	3		
34	PDE5A	2			84	PCSK1	4		
35	DNAJC11	2			85	NRXN1	5		
36	PCDHA6	3			86	ROBO2	2		MUFFINN-DNsum
37	FBN2	4			87	POMT2	1		
38	MUC7	3			88	ATP1A1	2		
39	ITGAL	4			89	DAAM1	3		
40	LRP1	6			90	ADAMTS2	4		HotNet2
41	GCN1L1	4			91	FASTK	2		HotNet2
42	PRKDC	5			92	MMP24	2		
43	TSG101	2		MUFFINN-DNsum	93	NPAS2	3		
44	S100A7	3			94	GEMIN5	2		
45	MYH6	3			95	CHD9	3		
46	HIST1H1C	1			96	ARVCF	2		
47	PDE4DIP	3			97	ALG13	4		
48	TBC1D4	2			98	TANC1	2		
49	POLR1B	2			99	NSD1	3		
50	FBXO11	3			100	NEK8	2		

Table S5. The average ranks of cancer specific known driver genes in the detection results of DGPathinter and the other network-based methods with STRING on BRCA, GBM and THCA dataset. The evaluation cancer specific known driver genes are from NCG4.0 [2] (left part of table) and CGC [1] (right part of table).

Known driver genes list	NCG4.0			CGC		
Method	BRCA	GBM	THCA	BRCA	GBM	THCA
DGPathinter(iRefIndex)	92.5	26.1	16.3	12.3	7.7	15.1
DGPathinter(STRING)	94.8	24.3	15.4	12.5	8.5	14.1
HotNet2(STRING)	504.1	62.1	850.9	258.3	42.9	794.2
NBS(STRING)	1290.2	418	1565	494	143.5	1460.3
MUFFINN-DNmax(STRING)	875.8	649.5	1683.4	1291.4	217	1726.5
MUFFINN-DNsum(STRING)	183.4	314.6	736.6	20.6	100.1	693.5
Random	6116.5	6116.5	6116.5	6116.5	6116.5	6116.5

Table S6. The areas under the curve (AUCs) of precision-recall curves of the detection results of DGPathinter and the other network-based methods with STRING on BRCA, GBM and THCA datasets, when evaluated by experimentally supported CGC genes [1] or statistically inferred NCG genes [2].

AUCs (CGC)	BRCA	GBM	THCA
DGPathinter(iRefIndex)	15.56%	26.19%	5.53%
DGPathinter(STRING)	14.36%	21.78%	5.49%
HotNet2(STRING)	0.81%	6.35%	0.25%
NBS(STRING)	0.21%	0.56%	0.19%
MUFFINN-DNmax(STRING)	0.43%	0.73%	0.19%
MUFFINN-DNsum(STRING)	1.96%	1.30%	0.28%
AUCs (NCG4.0)	BRCA	GBM	THCA
DGPathinter(iRefIndex)	12.19%	12.54%	5.18%
DGPathinter(STRING)	11.22%	12.92%	5.09%
HotNet2(STRING)	3.70%	5.56%	0.26%
NBS(STRING)	2.47%	0.94%	0.22%
MUFFINN-DNmax(STRING)	3.00%	0.86%	0.23%
MUFFINN-DNsum(STRING)	3.23%	1.22%	0.30%

Fig. S1. Bar plot of numbers of known cancer specific driver genes that are selected in the top 100 genes among the results of DGPathinter under different tuning parameters λ_C (top), λ_V (middle), λ_L (bottom). The dark blue bars represent the number of CGC genes [1], and the light blue bars represent the number of statistically inferred candidates genes in NCG4.0 (including both CGC genes and statistically inferred genes) [2]. The dark red texts at the top of the dark blue bars indicate the p-values of Fisher's exact test on the selected genes for cancer specific CGC gene, while the dark green texts at the top of the light blue bars represent the p-values for cancer-specific NCG4.0 genes.

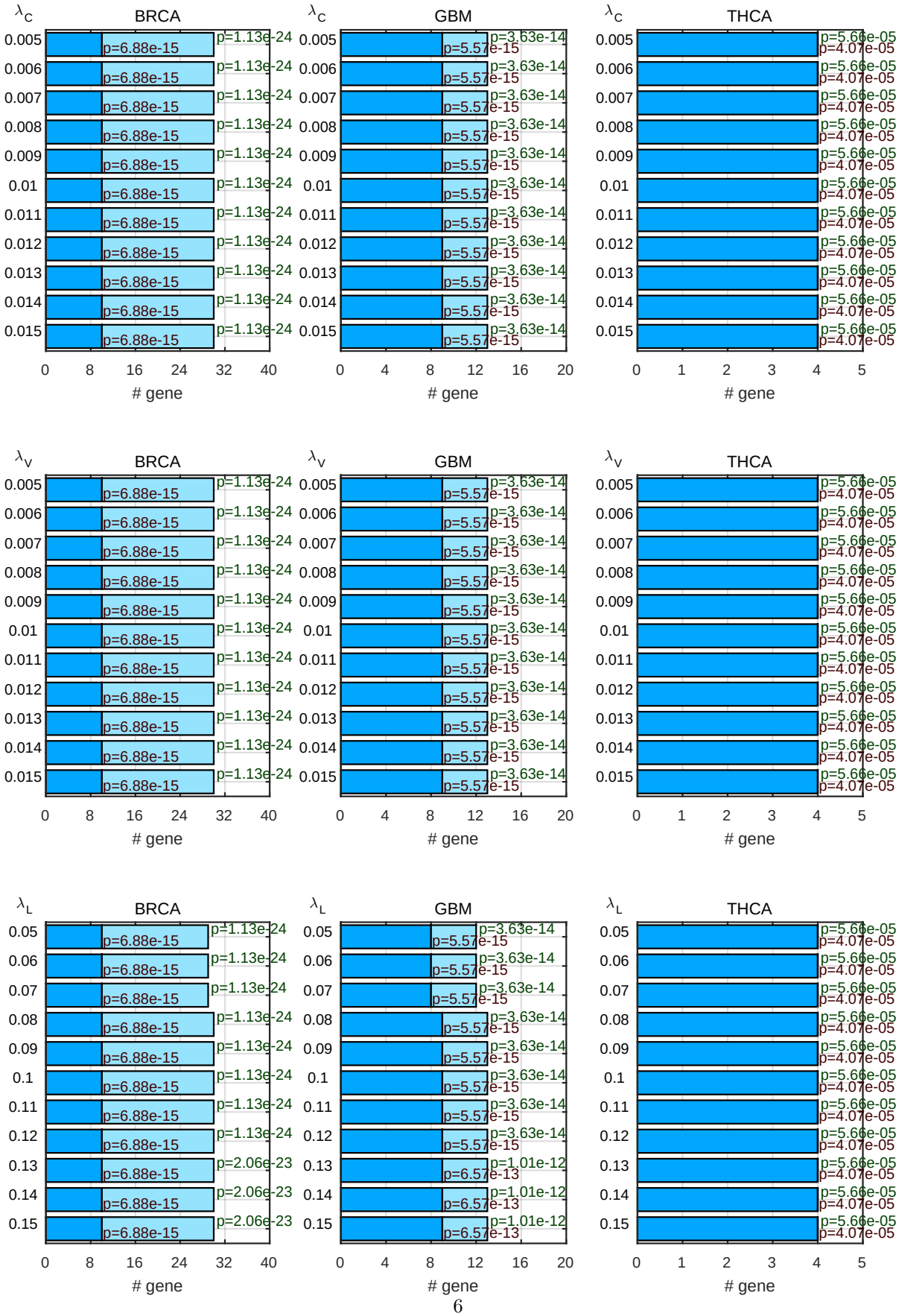
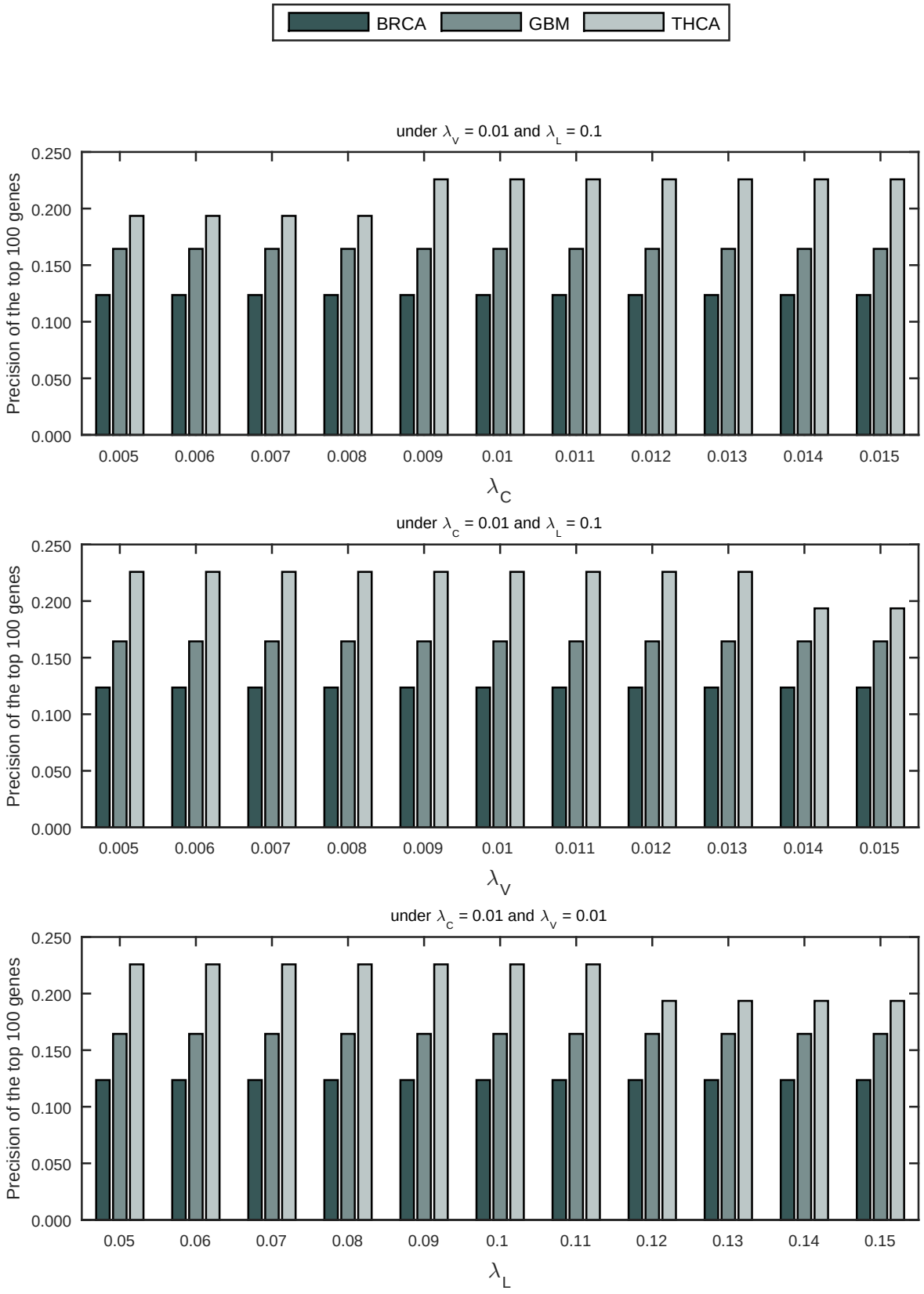
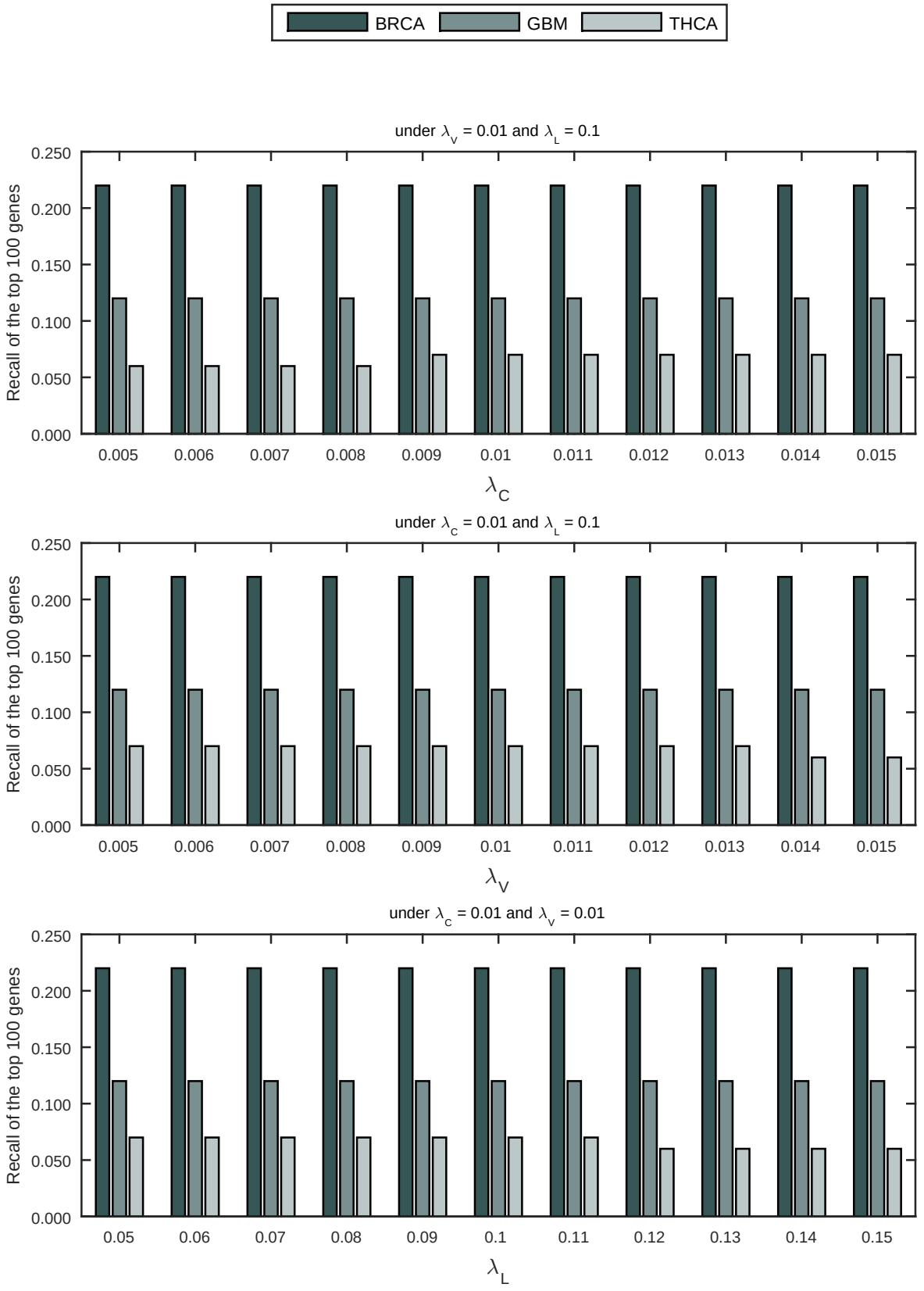


Fig. S2. Precision, recall and F-score (the harmonic mean of precision and recall) of the top 100 genes detected by DGPathinter for known NCG4.0 genes under different tuning parameters λ_C (top), λ_V (middle), λ_L (bottom). (A) Precisions; (B) Recalls; (C) F-scores.

(A)



(B)


BRCA
GBM
THCA
under λ_V = 0.01 and λ_L = 0.1
Recall of the top 100 genes
0.250
0.200
0.150
0.100
0.050
0.000
0.005
0.006
0.007
0.008
0.009
0.01
0.011
0.012
0.013
0.014
0.015
λ_C
under λ_C = 0.01 and λ_L = 0.1
Recall of the top 100 genes
λ_V
under λ_C = 0.01 and λ_V = 0.01
Recall of the top 100 genes
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
λ_L

(C)

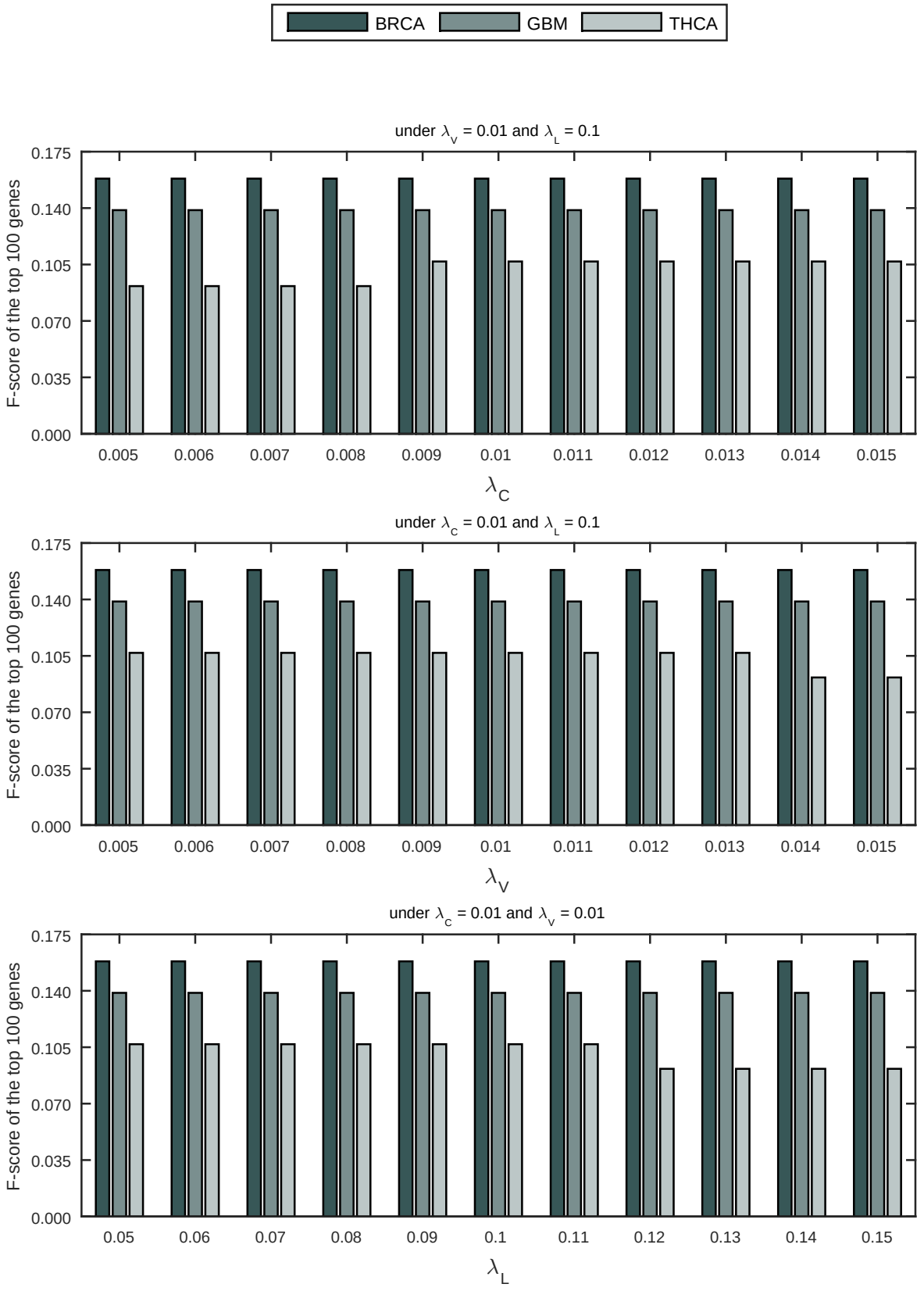
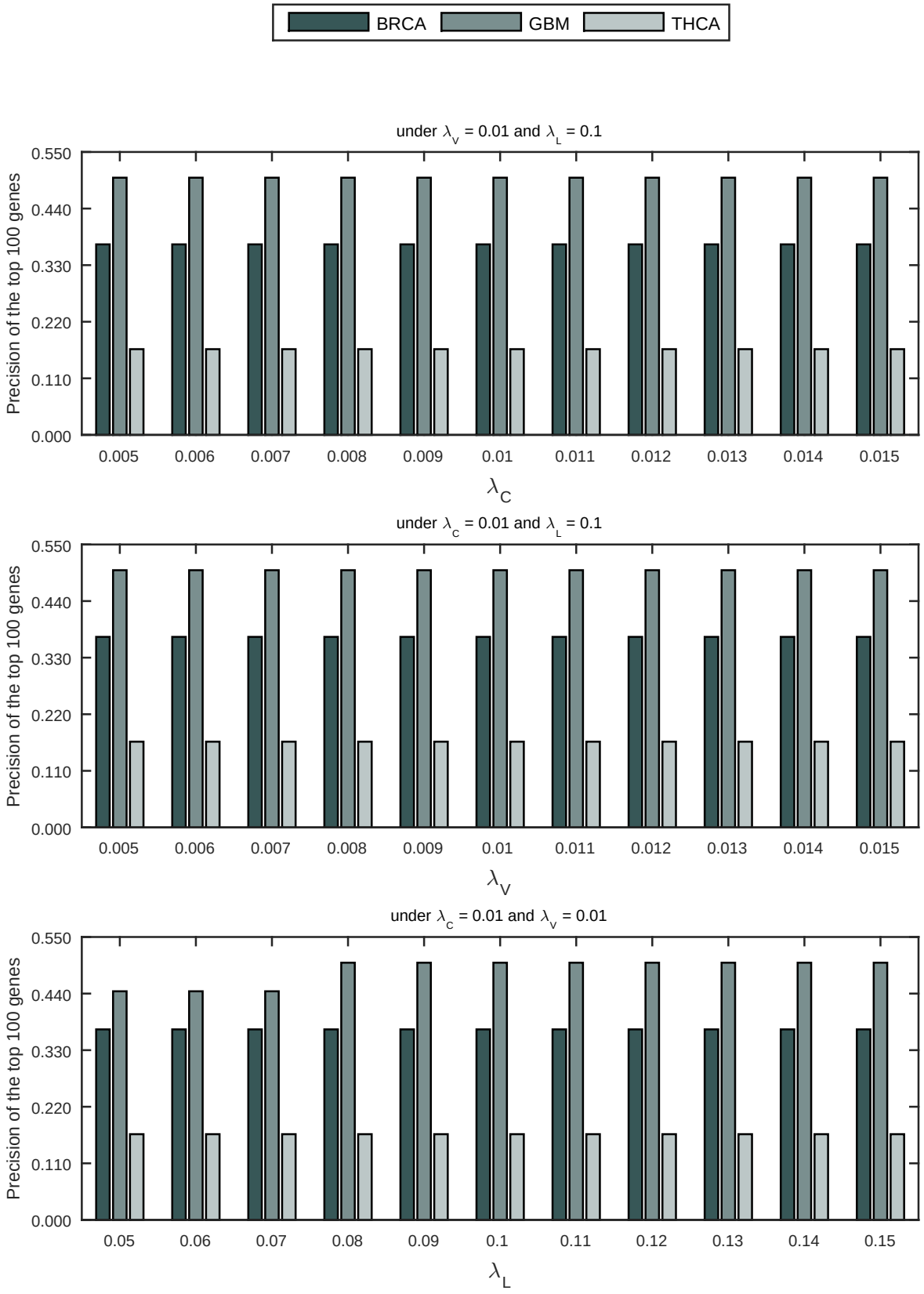
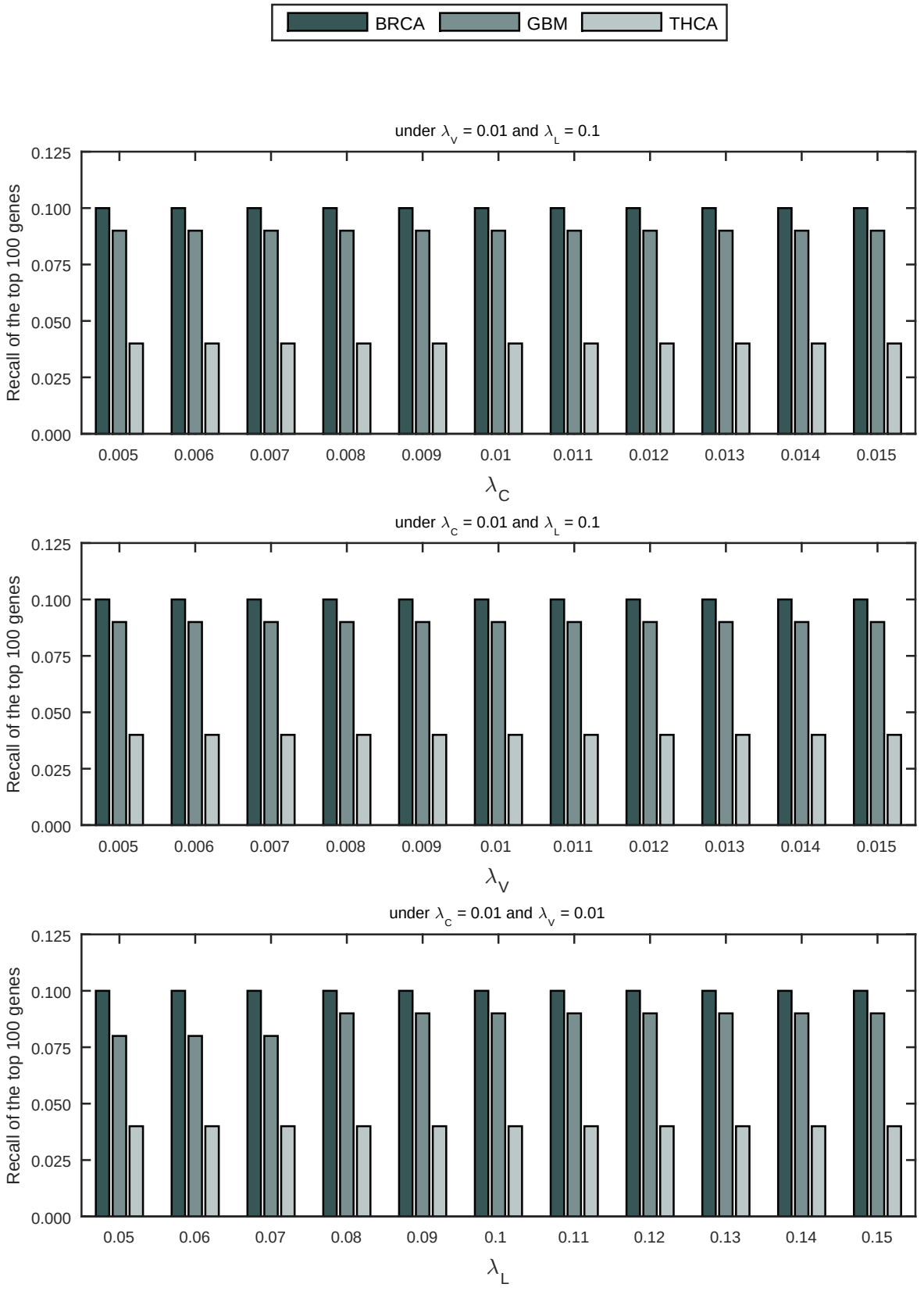

BRCA
GBM
THCA
under λ_V = 0.01 and λ_L = 0.1
F-score of the top 100 genes
0.175
0.140
0.105
0.070
0.035
0.000
0.005
0.006
0.007
0.008
0.009
0.01
0.011
0.012
0.013
0.014
0.015
λ_C
under λ_C = 0.01 and λ_L = 0.1
F-score of the top 100 genes
λ_V
under λ_C = 0.01 and λ_V = 0.01
F-score of the top 100 genes
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
λ_L

Fig. S3. Precision, recall and F-score (the harmonic mean of precision and recall) of the top 100 genes detected by DGPathinter for known CGC genes under different tuning parameters λ_C (top), λ_V (middle), λ_L (bottom). (A) Precisions; (B) Recalls; (C) F-scores.

(A)



(B)


BRCA
GBM
THCA
under λ_V = 0.01 and λ_L = 0.1
Recall of the top 100 genes
0.125
0.100
0.075
0.050
0.025
0.000
0.005
0.006
0.007
0.008
0.009
0.01
0.011
0.012
0.013
0.014
0.015
λ_C
under λ_C = 0.01 and λ_L = 0.1
Recall of the top 100 genes
λ_V
under λ_C = 0.01 and λ_V = 0.01
Recall of the top 100 genes
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
λ_L

(C)

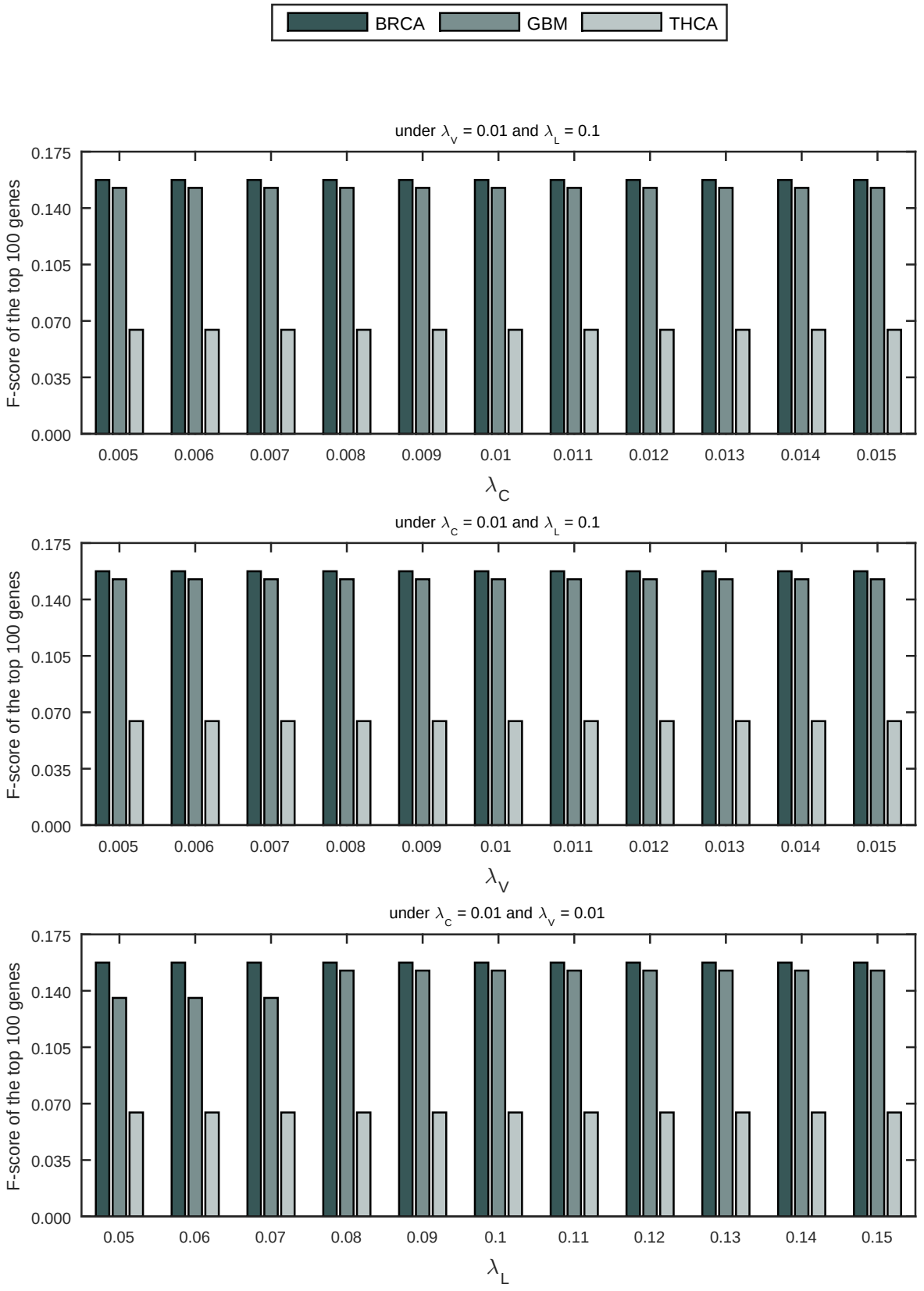

BRCA
GBM
THCA
under λ_V = 0.01 and λ_L = 0.1
F-score of the top 100 genes
0.175
0.140
0.105
0.070
0.035
0.000
0.005
0.006
0.007
0.008
0.009
0.01
0.011
0.012
0.013
0.014
0.015
λ_C
under λ_C = 0.01 and λ_L = 0.1
F-score of the top 100 genes
λ_V
under λ_C = 0.01 and λ_V = 0.01
F-score of the top 100 genes
0.05
0.06
0.07
0.08
0.09
0.1
0.11
0.12
0.13
0.14
0.15
λ_L

Fig. S4. Precision-recall curves of the prioritization results of the investigated methods for cancer specific known driver genes curated by CGC [1] on BRCA, GBM and HNSC datasets, where blue, dark green, light green, dark red and violet lines represent the curves of DGPathinter, HotNet2, NBS, MUFFINN-DNmax and MUFFINN-DNsum respectively. Different points on a same curve represent the precisions and recalls at different thresholds of the results.

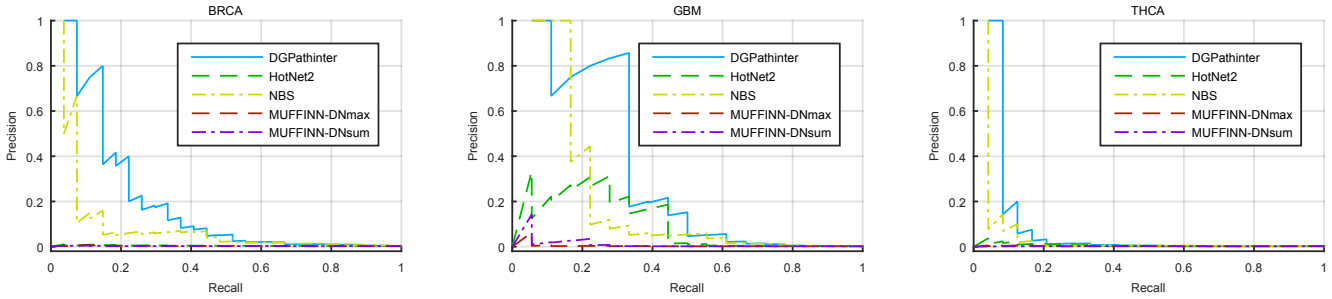


Fig. S5. Bar plot of numbers of known cancer specific driver genes that are selected in the top 100 genes among the results of DGPathinter with both network and pathway information (net+path), the results with only the pathway information (only path), the results with only the network information (only net) and the results with no prior information (no prior). The dark blue bars represent the number of CGC genes [1], and the light blue bars represent the number of statistically inferred candidates genes in NCG4.0 (including both CGC genes and statistically inferred genes) [2]. The dark red texts at the top of the dark blue bars indicate the p-values of Fisher's exact test on the selected genes for cancer specific CGC gene, while the dark green texts at the top of the light blue bars represent the p-values for cancer-specific NCG4.0 genes.

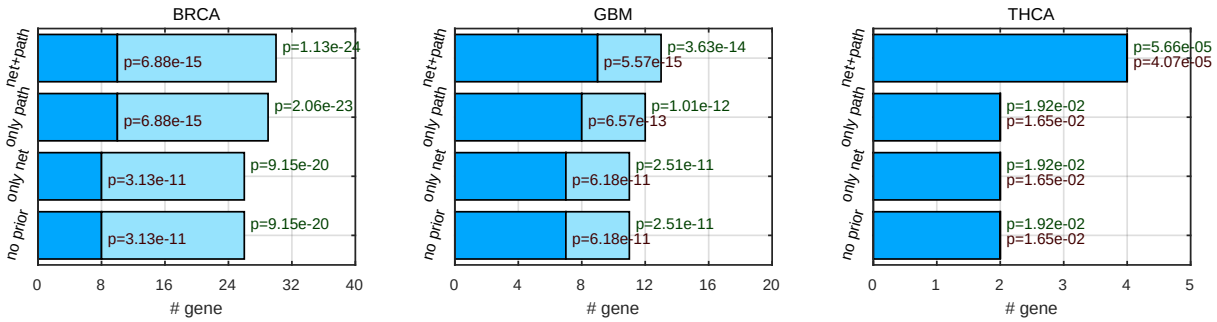


Fig. S6. Precision, recall and F-score (the harmonic mean of precision and recall) of the top 100 genes detected by DGPathinter with both network and pathway information (net+path), with only the pathway information (only path), with only the network information (only net) and with no prior information (no prior), for known CGC genes (left) and NCG4.0 genes (right) respectively.

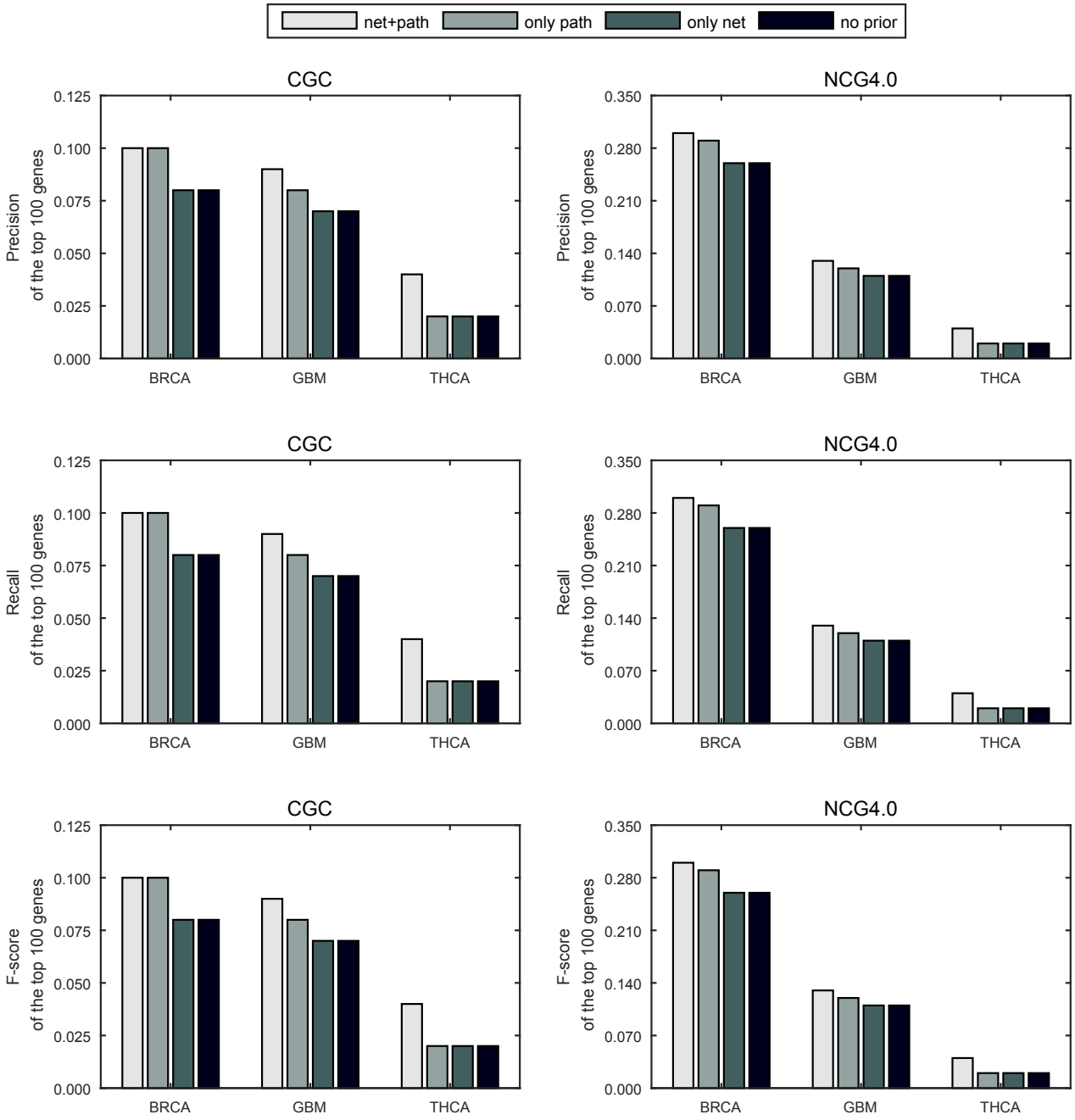


Fig. S7. Precision-recall curves of the results of DGPathinter(iRefIndex), DGPathinter(STRING), HotNet2(STRING), NBS(iRefIndex), MUFFINN-DNmax(iRefIndex) and MUFFINN-DNsum(iRefIndex) for cancer specific known driver genes curated by NCG4.0 [2] on BRCA, GBM and HNSC datasets, where blue, dark green, light green, dark red and violet lines represent the curves of DGPathinter, HotNet2, NBS, MUFFINN-DNmax and MUFFINN-DNsum respectively. Different points on a same curve represent the precisions and recalls at different thresholds of the results.

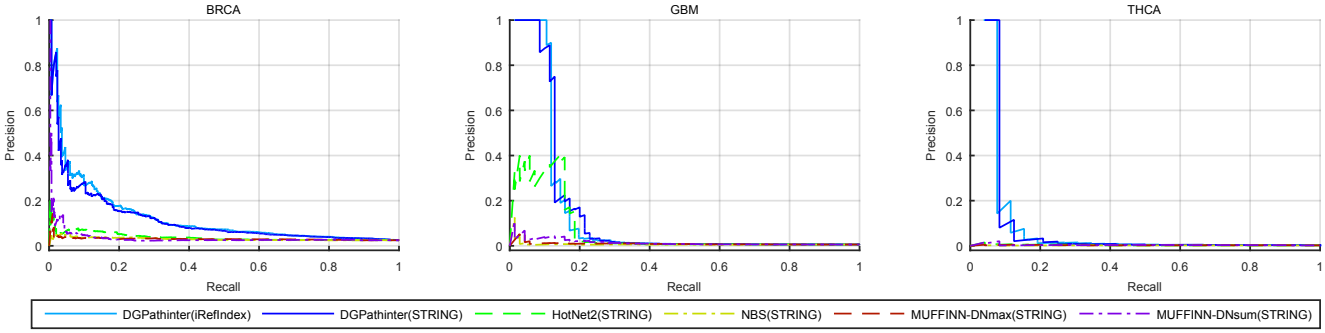


Fig. S8. Precision-recall curves of the results of DGPathinter(iRefIndex), DGPathinter(STRING), HotNet2(STRING), NBS(iRefIndex), MUFFINN-DNmax(iRefIndex) and MUFFINN-DNsum(iRefIndex) for cancer specific known driver genes curated by CGC [1] on BRCA, GBM and HNSC datasets, where blue, dark green, light green, dark red and violet lines represent the curves of DGPathinter, HotNet2, NBS, MUFFINN-DNmax and MUFFINN-DNsum respectively. Different points on a same curve represent the precisions and recalls at different thresholds of the results.

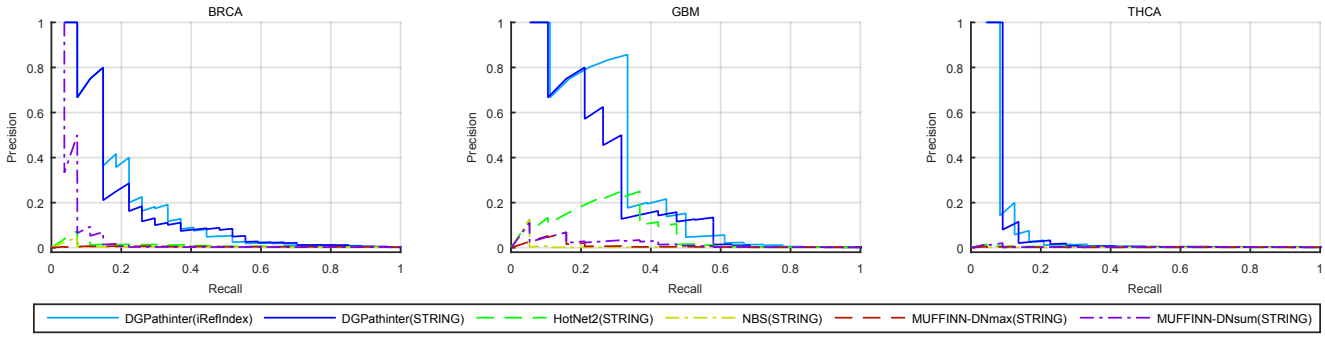
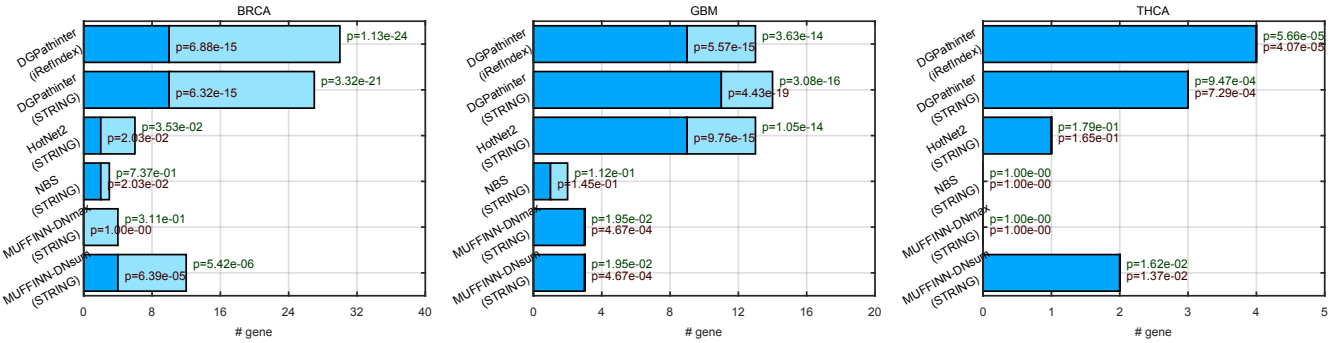


Fig. S9. Bar plot of numbers of known cancer specific driver genes that are selected in the top 100 genes among the results of DGPathinter(iRefIndex), DGPathinter(STRING), HotNet2(STRING), NBS(iRefIndex), MUFFINN-DNmax(iRefIndex) and MUFFINN-DNsum(iRefIndex). The dark blue bars represent the number of CGC genes [1], and the light blue bars represent the number of statistically inferred candidates genes in NCG4.0 (including both CGC genes and statistically inferred genes) [2]. The dark red texts at the top of the dark blue bars indicate the p-values of Fisher's exact test on the selected genes for cancer specific CGC gene, while the dark green texts at the top of the light blue bars represent the p-values for cancer-specific NCG4.0 genes.



References

[1] P Andrew Futreal, Lachlan Coin, Mhairi Marshall, Thomas Down, Timothy Hubbard, Richard Wooster, Nazneen Rahman, and Michael R Stratton. A census of human cancer genes. *Nature Reviews Cancer*, 4(3):177–183, 2004.

[2] Omer An, Vera Pendino, Matteo DAntonio, Emanuele Ratti, Marco Gentilini, and Francesca D Ciccarelli. NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database*, 2014:bau015, 2014.

[3] Mark D Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R Dobson, and Benjamin R Raphael. Pan-cancer identification of mutated pathways and protein complexes. *Cancer Research*, 74(19 Supplement):5324–5324, 2014.