SIBS 中国科学院上海生命科学研究院
Shanghai Institutes for Biological Sciences, CAS

砺志求真　笃学明德

Deciphering the mechanisms of complex diseases using machine learning and network analysis approaches

The 15th KJC Bioinformatics Symposium, Seoul, Korea

Tao Huang, Ph.D.　　**June 21-22, 2017**

Contents

01 Pan-Cancer CNV

02 lncRNA in EBVaGC

03 SCZ-ATDH Occurrence

04 BRCA Metastasis

PART

Pan-Cancer CNV

ONE

TCGA CNV from cBioPortal

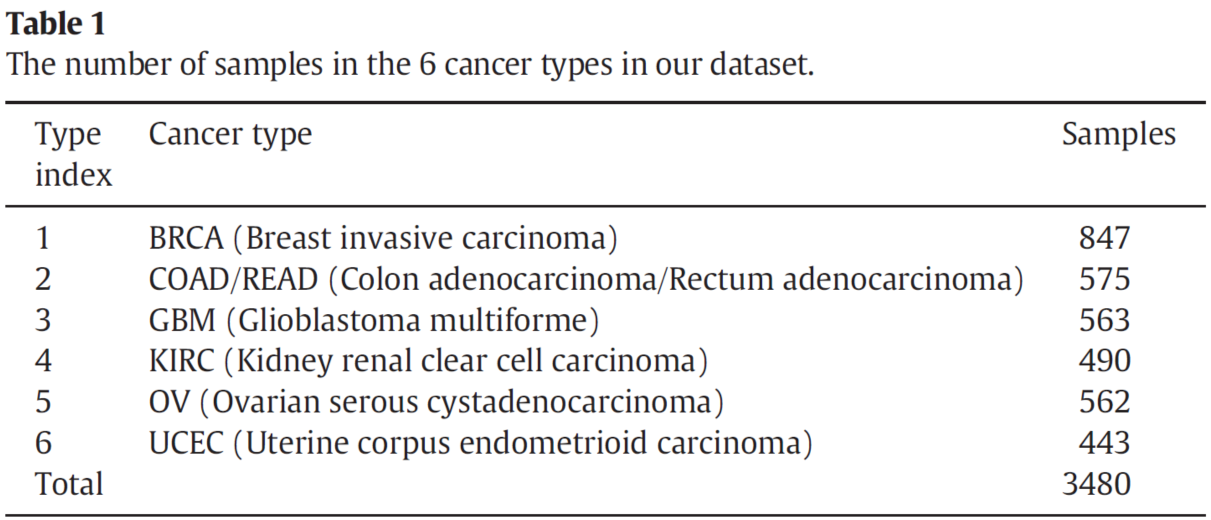


Table 1
The number of samples in the 6 cancer types in our dataset.

Type index	Cancer type	Samples
1	BRCA (Breast invasive carcinoma)	847
2	COAD/READ (Colon adenocarcinoma/Rectum adenocarcinoma)	575
3	GBM (Glioblastoma multiforme)	563
4	KIRC (Kidney renal clear cell carcinoma)	490
5	OV (Ovarian serous cystadenocarcinoma)	562
6	UCEC (Uterine corpus endometrioid carcinoma)	443
Total		3480

Cancer Types with Sample Size > 400

CNV values for 24,174 genes

Value	Meaning
-2	a deep loss (possibly a homozygous deletion)
-1	a shallow loss (possibly heterozygous deletion)
0	diploid
1	a low-level gain
2	a high-level amplification

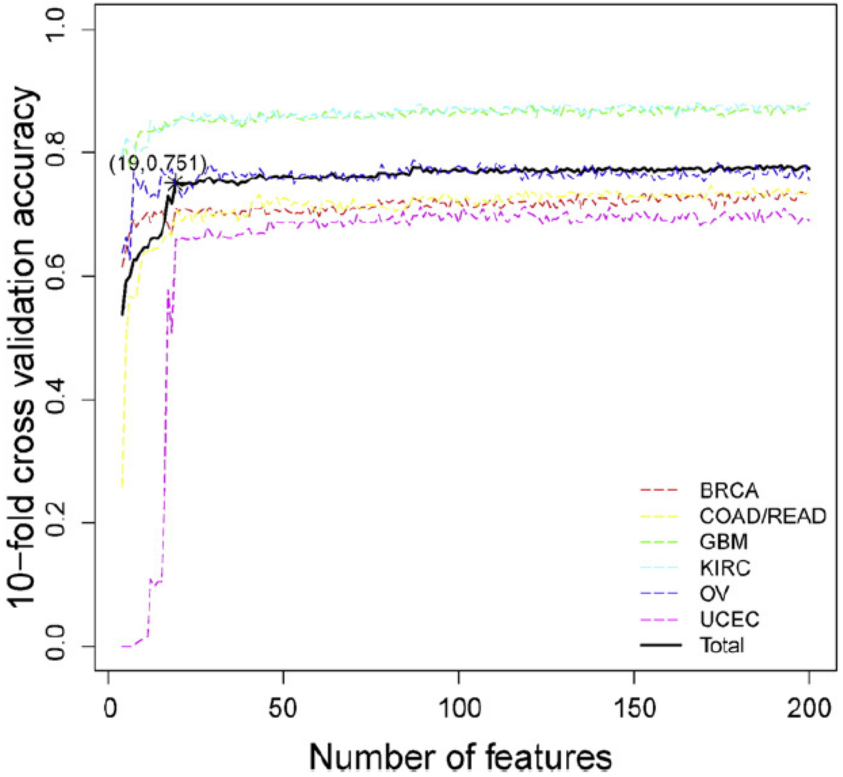
Most discriminative CNVs for different cancers

mRMR

minimal-Redundancy-Maximal-Relevance

$$\max_{f_j \in \Omega_t} \left[I\left(f_j, c\right) - \frac{1}{mm} \sum_{f_i \in \Omega_s} I(f_j, f_i) \right]$$

Incremental feature selection



IFS

Top 19 CNVs

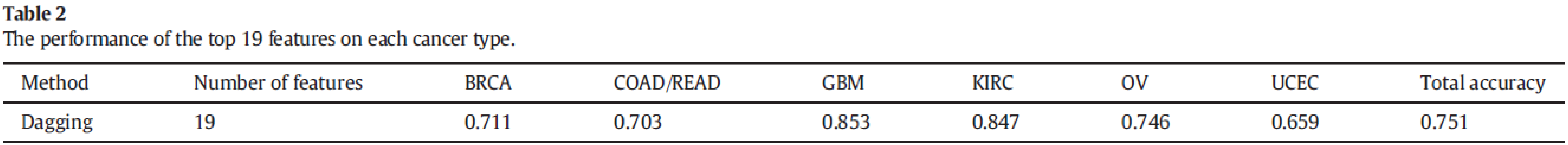


Table 2
The performance of the top 19 features on each cancer type.

Method	Number of features	BRCA	COAD/READ	GBM	KIRC	OV	UCEC	Total accuracy
Dagging	19	0.711	0.703	0.853	0.847	0.746	0.659	0.751

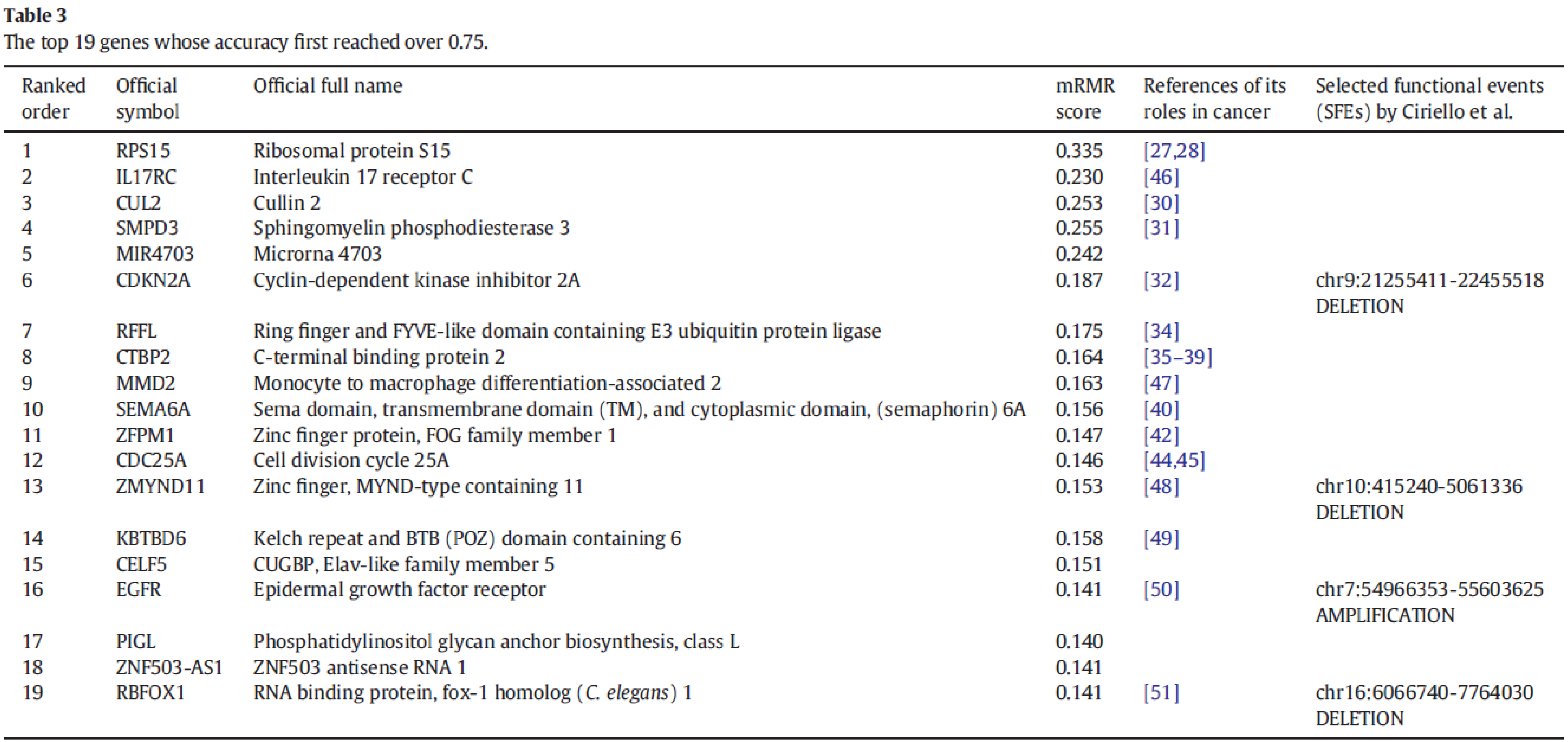


Table 3
The top 19 genes whose accuracy first reached over 0.75.

Ranked order	Official symbol	Official full name	mRMR score	References of its roles in cancer	Selected functional events (SFEs) by Ciriello et al.
1	RPS15	Ribosomal protein S15	0.335	[27,28]	
2	IL17RC	Interleukin 17 receptor C	0.230	[46]	
3	CUL2	Cullin 2	0.253	[30]	
4	SMPD3	Sphingomyelin phosphodiesterase 3	0.255	[31]	
5	MIR4703	Microrna 4703	0.242		
6	CDKN2A	Cyclin-dependent kinase inhibitor 2A	0.187	[32]	chr9:21255411-22455518 DELETION
7	RFFL	Ring finger and FYVE-like domain containing E3 ubiquitin protein ligase	0.175	[34]	
8	CTBP2	C-terminal binding protein 2	0.164	[35–39]	
9	MMD2	Monocyte to macrophage differentiation-associated 2	0.163	[47]	
10	SEMA6A	Sema domain, transmembrane domain (TM), and cytoplasmic domain, (semaphorin) 6A	0.156	[40]	
11	ZFPM1	Zinc finger protein, FOG family member 1	0.147	[42]	
12	CDC25A	Cell division cycle 25A	0.146	[44,45]	
13	ZMYND11	Zinc finger, MYND-type containing 11	0.153	[48]	chr10:415240-5061336 DELETION
14	KBTBD6	Kelch repeat and BTB (POZ) domain containing 6	0.158	[49]	
15	CELF5	CUGBP, Elav-like family member 5	0.151		
16	EGFR	Epidermal growth factor receptor	0.141	[50]	chr7:54966353-55603625 AMPLIFICATION
17	PIGL	Phosphatidylinositol glycan anchor biosynthesis, class L	0.140		
18	ZNF503-AS1	ZNF503 antisense RNA 1	0.141		
19	RBFOX1	RNA binding protein, fox-1 homolog (*C. elegans*) 1	0.141	[51]	chr16:6066740-7764030 DELETION

The CNV, deletion, amplification occurrence frequencies

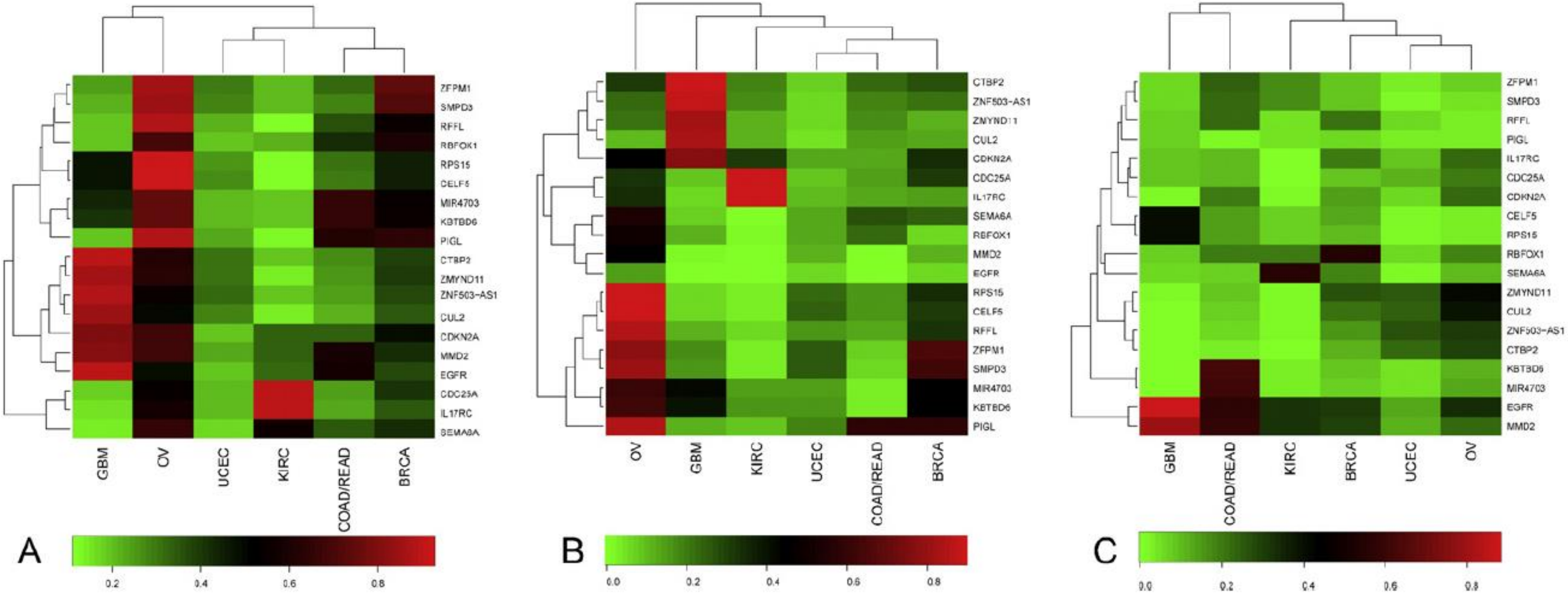


Fig. 2. The CNV, deletion, amplification occurrence frequencies of the selected 19 in six cancer types. (A) The CNV (deletion and amplification) occurrence frequencies of the selected 19 in six cancer types; (B) The deletion occurrence frequencies of the selected 19 in six cancer types; (C) The amplification occurrence frequencies of the selected 19 in six cancer types.

PART

lncRNA in EBVaGC

TWO

EBVaGC: Epstein–Barr virus (EBV)-associated gastric carcinoma

Gastric cancer

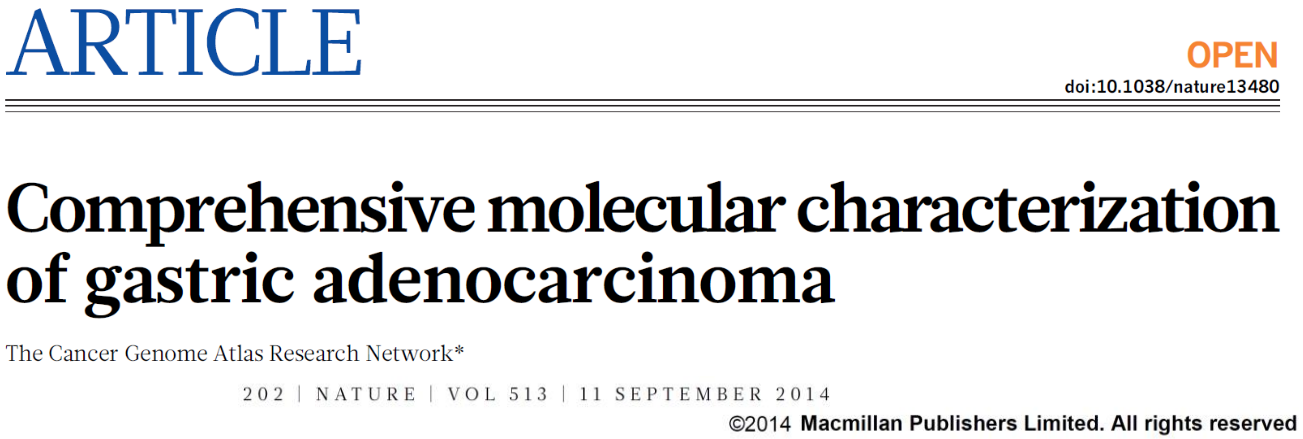
4th most common cancer worldwide
2nd on cancer death

EBVaGC

Epstein–Barr virus (EBV)-associated gastric carcinoma
1/10 of all gastric carcinomas

Identification

It is identified by the expression of EBV-encoded small ribonucleic acid 1 (EBER1) in cancer cell nuclei, using in situ hybridization.



ARTICLE

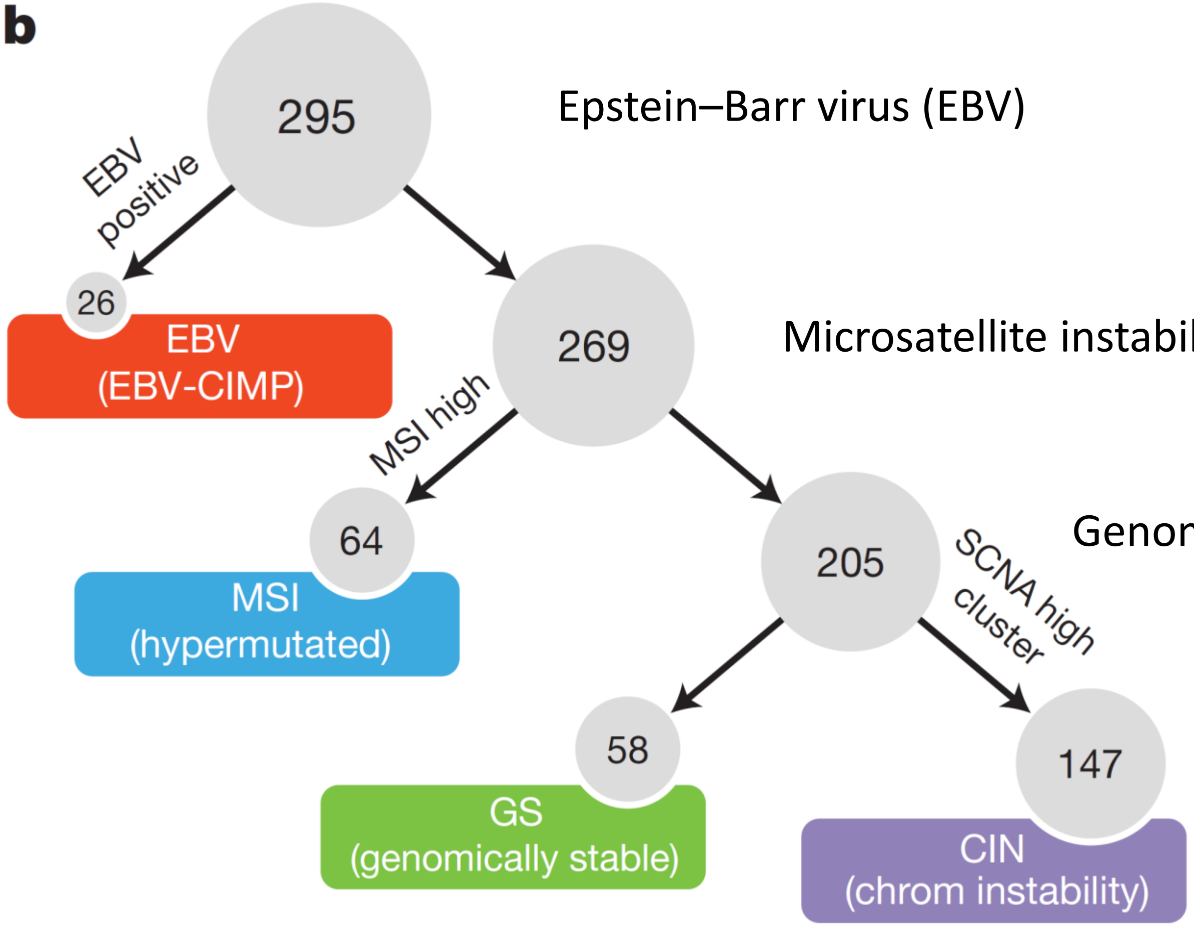
OPEN

doi:10.1038/nature13480

Comprehensive molecular characterization of gastric adenocarcinoma

The Cancer Genome Atlas Research Network*

202 | NATURE | VOL 513 | 11 SEPTEMBER 2014



Epstein–Barr virus (EBV)

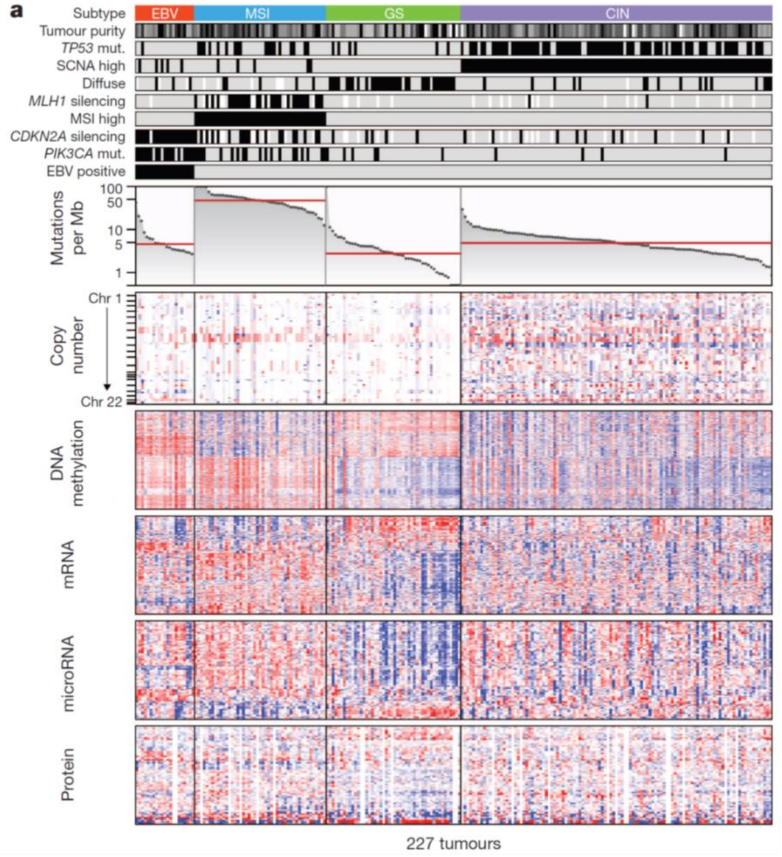
Microsatellite instability (MSI)

Genomically stable (GS)

Chromosomal instability (CIN)

lncRNA?

TCGA



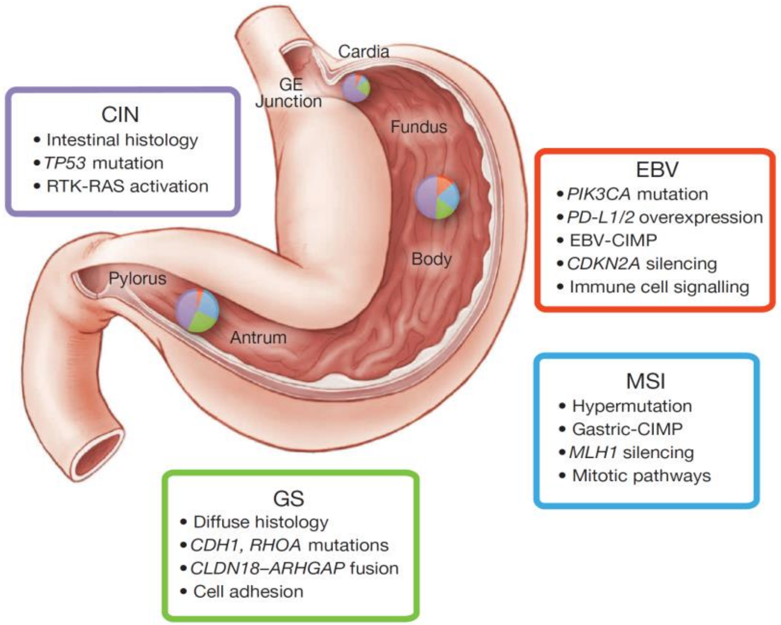
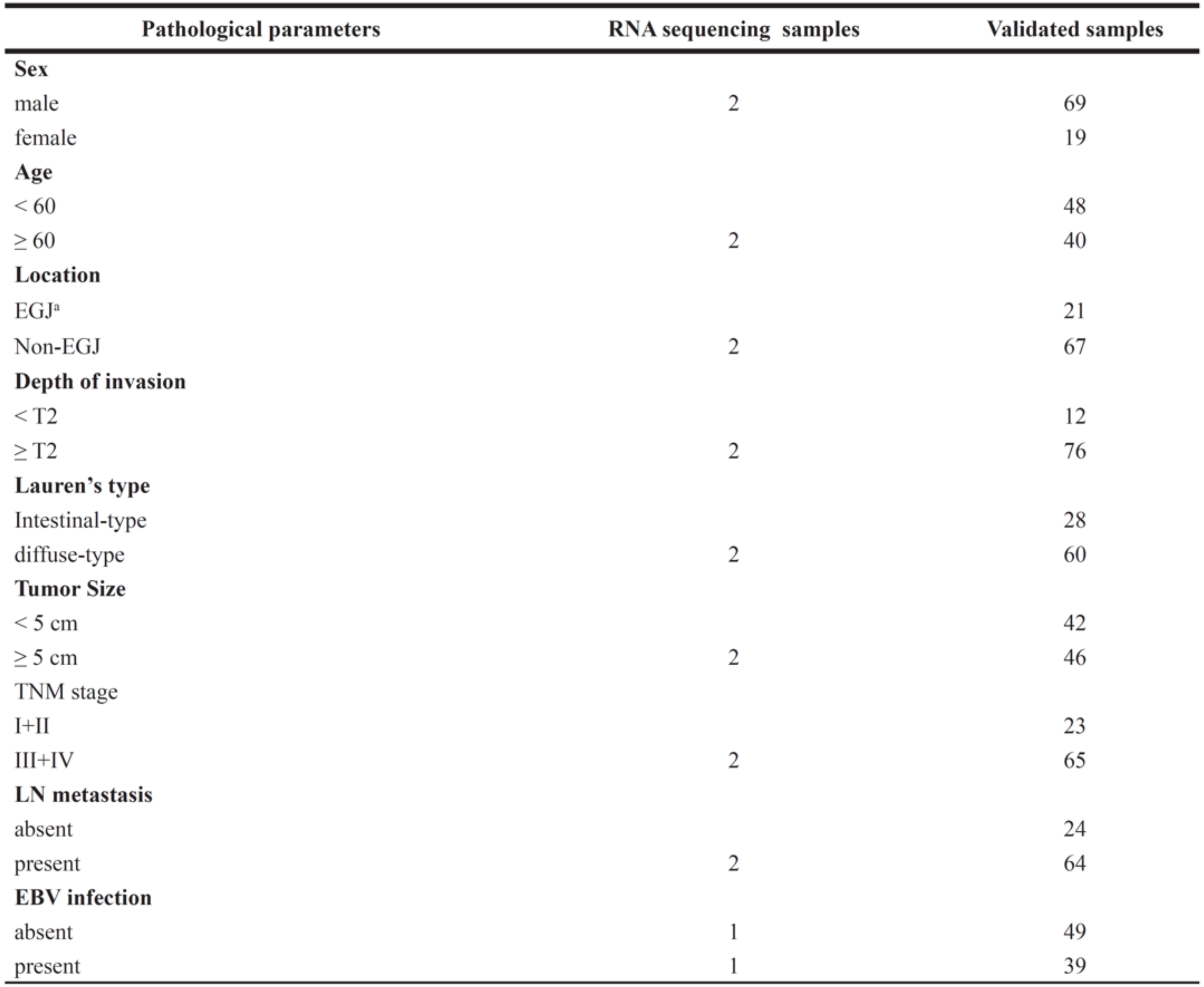


Figure 6 | Key features of gastric cancer subtypes. This schematic lists some of the salient features associated with each of the four molecular subtypes of gastric cancer. Distribution of molecular subtypes in tumours obtained from distinct regions of the stomach is represented by inset charts.

Samples



Pathological parameters	RNA sequencing samples	Validated samples
Sex		
male	2	69
female		19
Age		
< 60		48
≥ 60	2	40
Location		
EGJ[a]		21
Non-EGJ	2	67
Depth of invasion		
< T2		12
≥ T2	2	76
Lauren's type		
Intestinal-type		28
diffuse-type	2	60
Tumor Size		
< 5 cm		42
≥ 5 cm	2	46
TNM stage		
I+II		23
III+IV	2	65
LN metastasis		
absent		24
present	2	64
EBV infection		
absent	1	49
present	1	39

EBV-specific lncRNAs

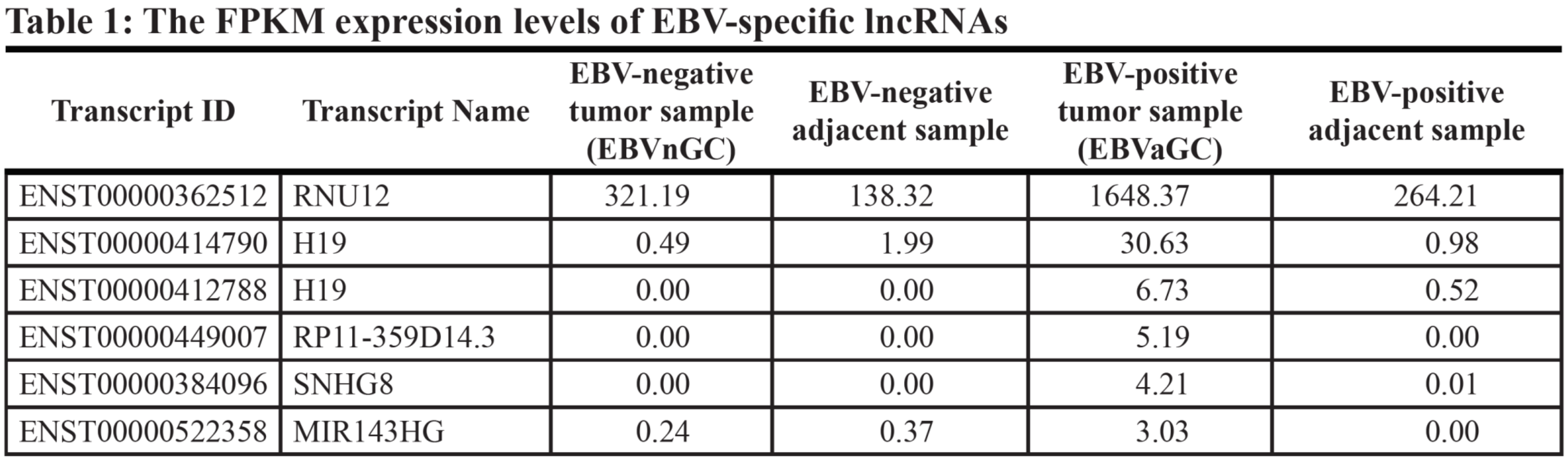


Table 1: The FPKM expression levels of EBV-specific lncRNAs

Transcript ID	Transcript Name	EBV-negative tumor sample (EBVnGC)	EBV-negative adjacent sample	EBV-positive tumor sample (EBVaGC)	EBV-positive adjacent sample
ENST00000362512	RNU12	321.19	138.32	1648.37	264.21
ENST00000414790	H19	0.49	1.99	30.63	0.98
ENST00000412788	H19	0.00	0.00	6.73	0.52
ENST00000449007	RP11-359D14.3	0.00	0.00	5.19	0.00
ENST00000384096	SNHG8	0.00	0.00	4.21	0.01
ENST00000522358	MIR143HG	0.24	0.37	3.03	0.00

Validation of SNHG8

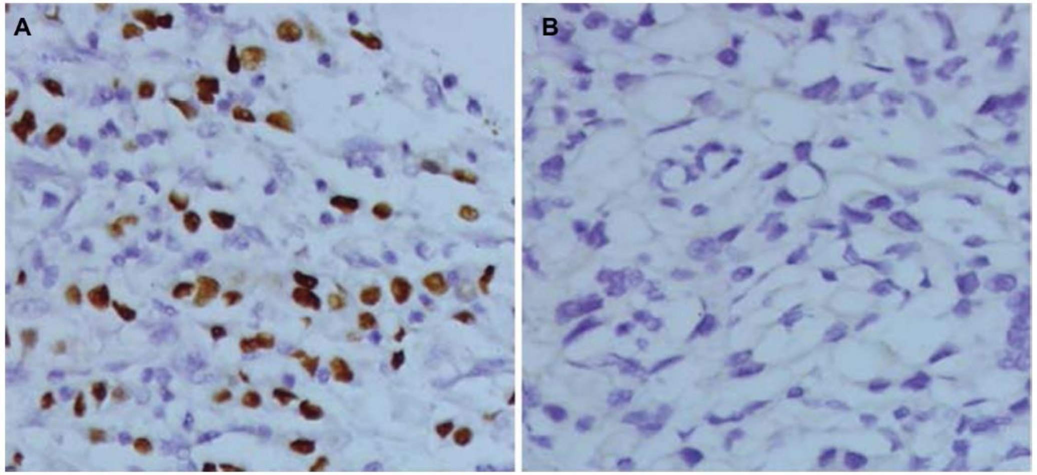


Figure 1: *In situ* hybridization of EBER1 in gastric cancer tissue. (**A**) EBVaGC, EBER(+) tissue. (**B**) EBVnGC, EBER(−) tissue. Magnification, ×400. EBER, EBV-encoded small RNA; EBVaGC, EBV-associated gastric carcinoma; EBVnGC, non-EBV-infected gastric cancer.

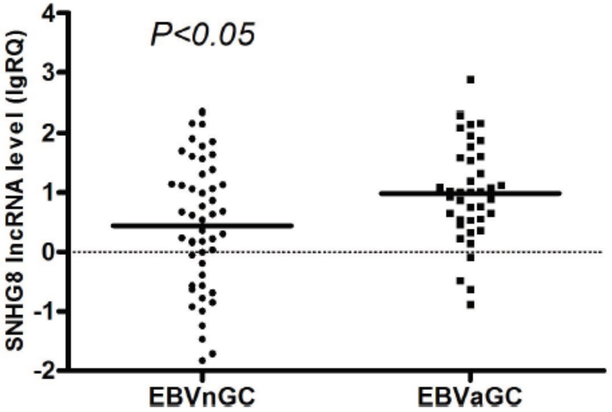


Figure 2: Distribution of SNHG8 lncRNA levels in EBVnGC and EBVaGC. Bold lines represent the mean value for each patient cohort; RQ = $2^{-\Delta\Delta Ct}$

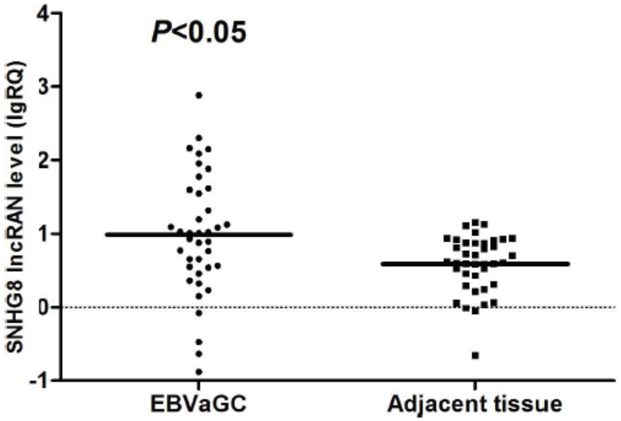


Figure 3: Distribution of SNHG8 lncRNA levels in EBVaGC and adjacent tissue. Bold lines represent the mean value for each patient cohort; RQ = $2^{-\Delta\Delta Ct}$

The Role of SNHG8 in EBVaGC?

Targets of SNHG8

Data

- RNA sequencing dataset of the Illumina Body Map
- 14,886 lncRNAs from the LNCipedia
- 21,721 protein-coding genes from UCSC hg19

Method

- Expression levels of lncRNAs and 21,721 protein-coding genes were calculated using TopHat and Cufflinks
- Co-expression pair of with absolute Pearson correlation coefficient > 0.5

Results

- The coexpressed mRNAs were considered to constitute the microenvironment around the lncRNA and were used to annotate the functions of the lncRNA.
- **SNHG8 : 577 targets**

Targets of EBV

Data

- EBV Genomics (https://ebv.wistar.upenn.edu)
- Human gene expression levels and EBV expression levels
- 201 samples

Method

- Calculate Pearson correlation coefficient between the human and EBV genes
- Human genes with an absolute Pearson correlation coefficient > 0.5 were considered as the target genes of an EBV gene

Results

- **Target genes of EBV Proteins**

How SNHG8 interacts with EBV

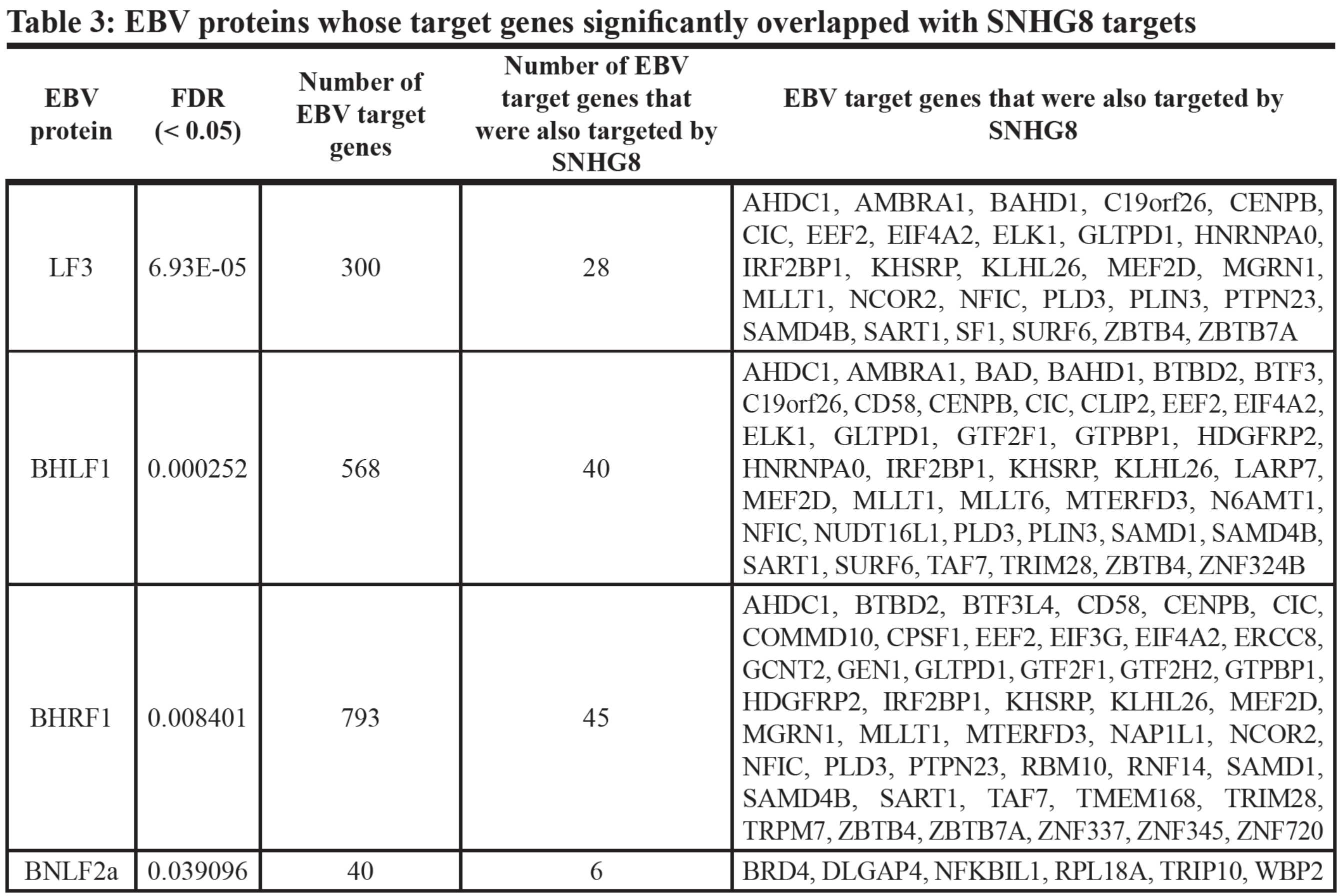
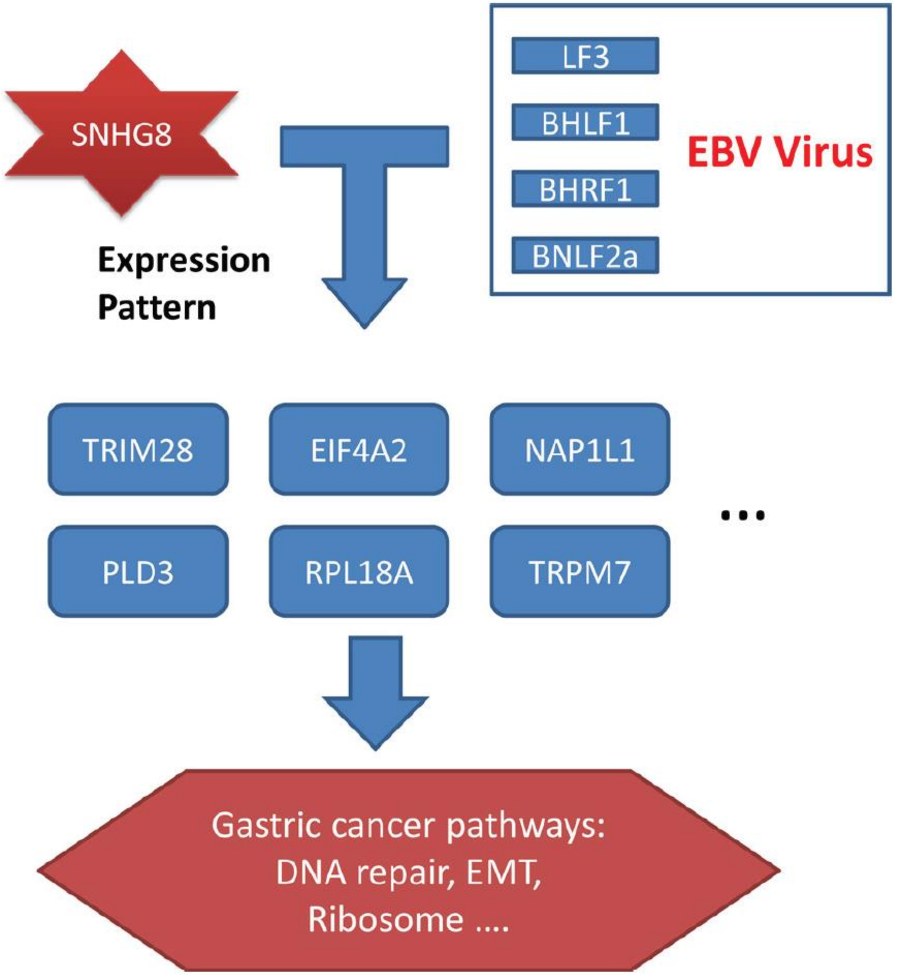


Table 3: EBV proteins whose target genes significantly overlapped with SNHG8 targets

EBV protein	FDR (< 0.05)	Number of EBV target genes	Number of EBV target genes that were also targeted by SNHG8	EBV target genes that were also targeted by SNHG8
LF3	6.93E-05	300	28	AHDC1, AMBRA1, BAHD1, C19orf26, CENPB, CIC, EEF2, EIF4A2, ELK1, GLTPD1, HNRNPA0, IRF2BP1, KHSRP, KLHL26, MEF2D, MGRN1, MLLT1, NCOR2, NFIC, PLD3, PLIN3, PTPN23, SAMD4B, SART1, SF1, SURF6, ZBTB4, ZBTB7A
BHLF1	0.000252	568	40	AHDC1, AMBRA1, BAD, BAHD1, BTBD2, BTF3, C19orf26, CD58, CENPB, CIC, CLIP2, EEF2, EIF4A2, ELK1, GLTPD1, GTF2F1, GTPBP1, HDGFRP2, HNRNPA0, IRF2BP1, KHSRP, KLHL26, LARP7, MEF2D, MLLT1, MLLT6, MTERFD3, N6AMT1, NFIC, NUDT16L1, PLD3, PLIN3, SAMD1, SAMD4B, SART1, SURF6, TAF7, TRIM28, ZBTB4, ZNF324B
BHRF1	0.008401	793	45	AHDC1, BTBD2, BTF3L4, CD58, CENPB, CIC, COMMD10, CPSF1, EEF2, EIF3G, EIF4A2, ERCC8, GCNT2, GEN1, GLTPD1, GTF2F1, GTF2H2, GTPBP1, HDGFRP2, IRF2BP1, KHSRP, KLHL26, MEF2D, MGRN1, MLLT1, MTERFD3, NAP1L1, NCOR2, NFIC, PLD3, PTPN23, RBM10, RNF14, SAMD1, SAMD4B, SART1, TAF7, TMEM168, TRIM28, TRPM7, ZBTB4, ZBTB7A, ZNF337, ZNF345, ZNF720
BNLF2a	0.039096	40	6	BRD4, DLGAP4, NFKBIL1, RPL18A, TRIP10, WBP2

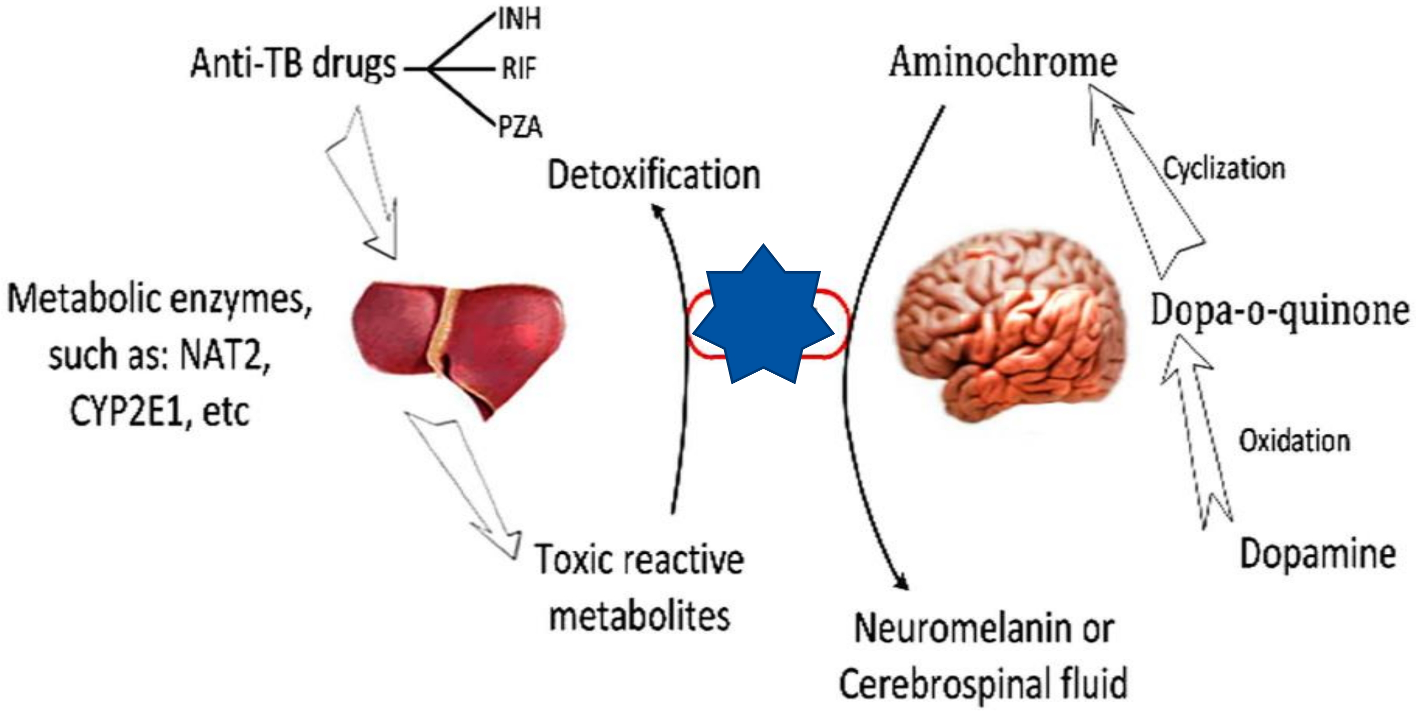


PART THREE

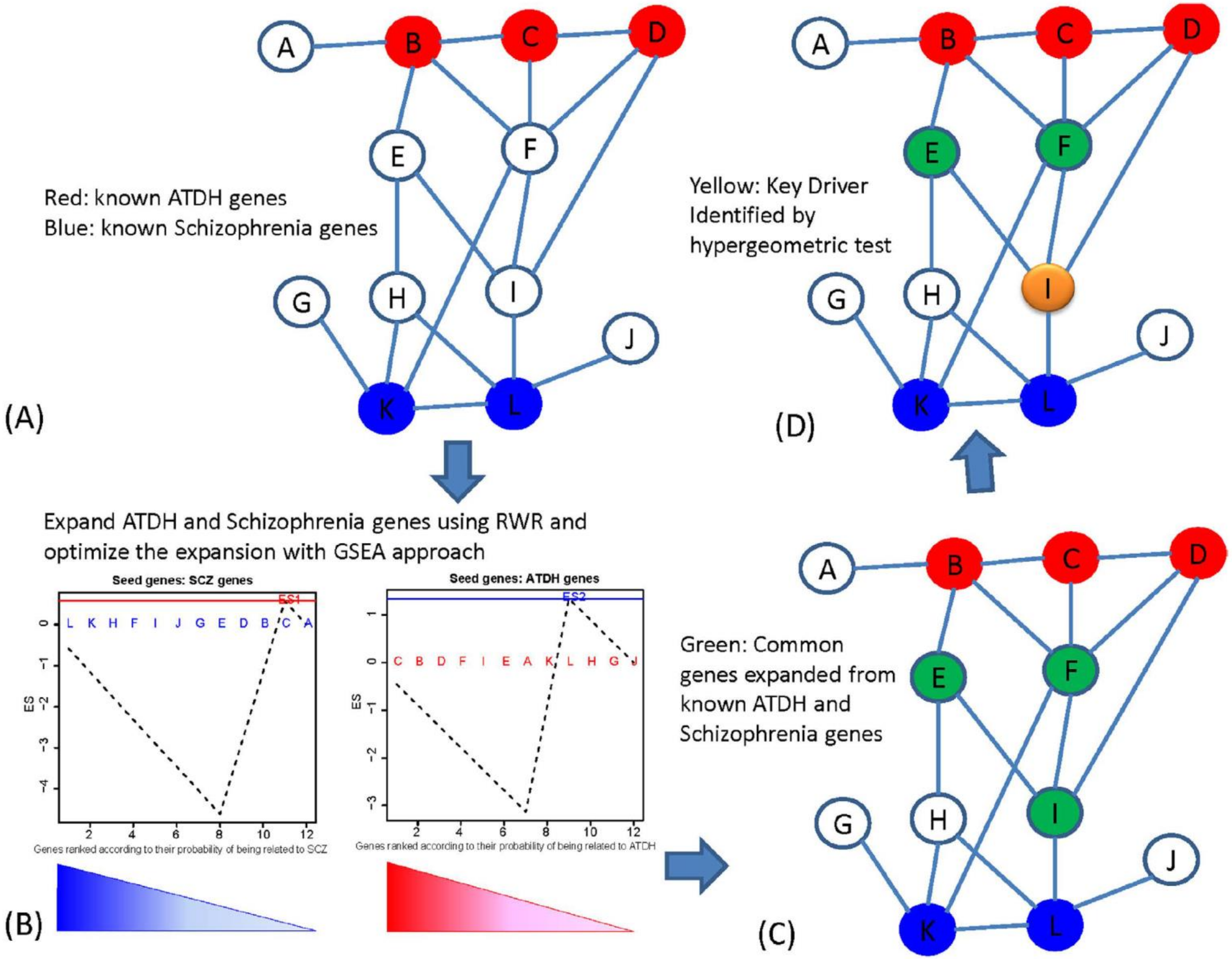
SCZ-ATDH Occurrence

Clinical Observation of SCZ and ATDH

The treatments for tuberculosis can induce anti-tuberculosis drug-induced hepatotoxicity (ATDH) and Schizophrenia (SCZ)-like disorders.


Anti-TB drugs
INH
RIF
PZA
Metabolic enzymes, such as: NAT2, CYP2E1, etc
Detoxification
Toxic reactive metabolites
Aminochrome
Cyclization
Dopa-o-quinone
Oxidation
Dopamine
Neuromelanin or Cerebrospinal fluid

Key Driver of SCZ and ATDH



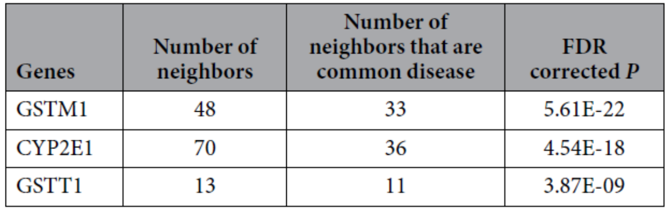
Red: known ATDH genes
Blue: known Schizophrenia genes

Expand ATDH and Schizophrenia genes using RWR and optimize the expansion with GSEA approach

Green: Common genes expanded from known ATDH and Schizophrenia genes

Yellow: Key Driver Identified by hypergeometric test

Validate shared key drivers with GWAS



Genes	Number of neighbors	Number of neighbors that are common disease	FDR corrected *P*
GSTM1	48	33	5.61E-22
CYP2E1	70	36	4.54E-18
GSTT1	13	11	3.87E-09

Table 1. The shared key causal genes for ATDH and SCZ.

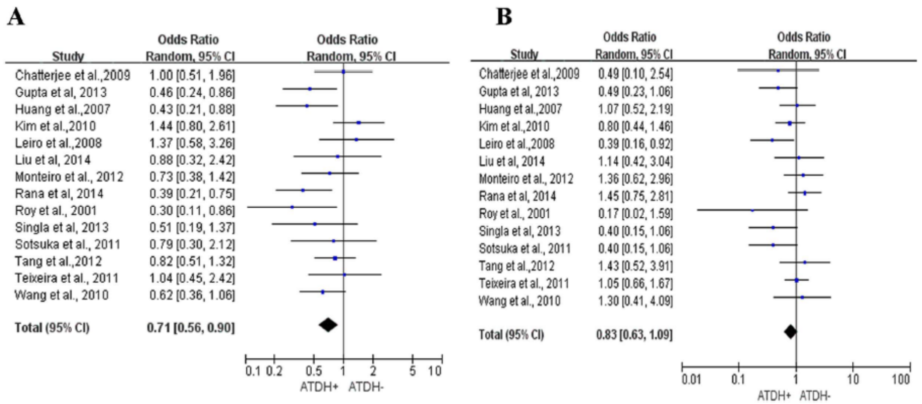


Figure 1. Forest plots from meta-analysis of *GSTM1/GSTT1* polymorphisms and ATDH. (A) Summary of the ORs and corresponding 95% CIs for the *GSTM1* present genotype; (B) summary of the ORs and 95% CIs for the *GSTT1* present genotype.

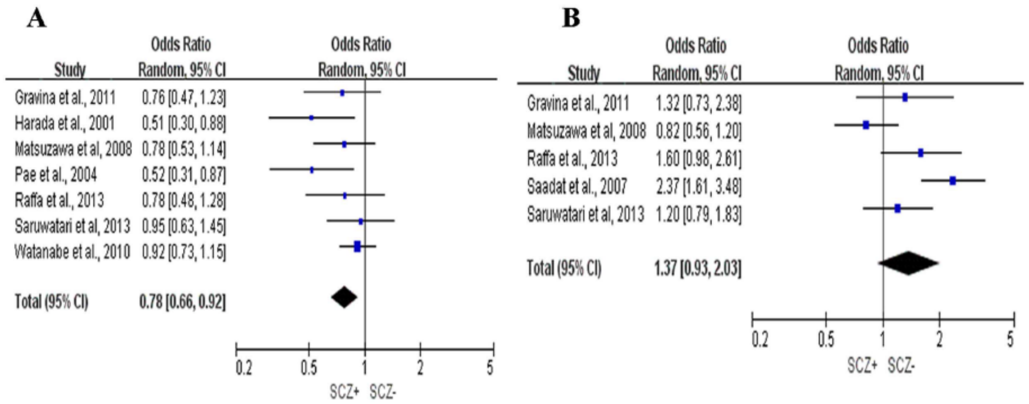
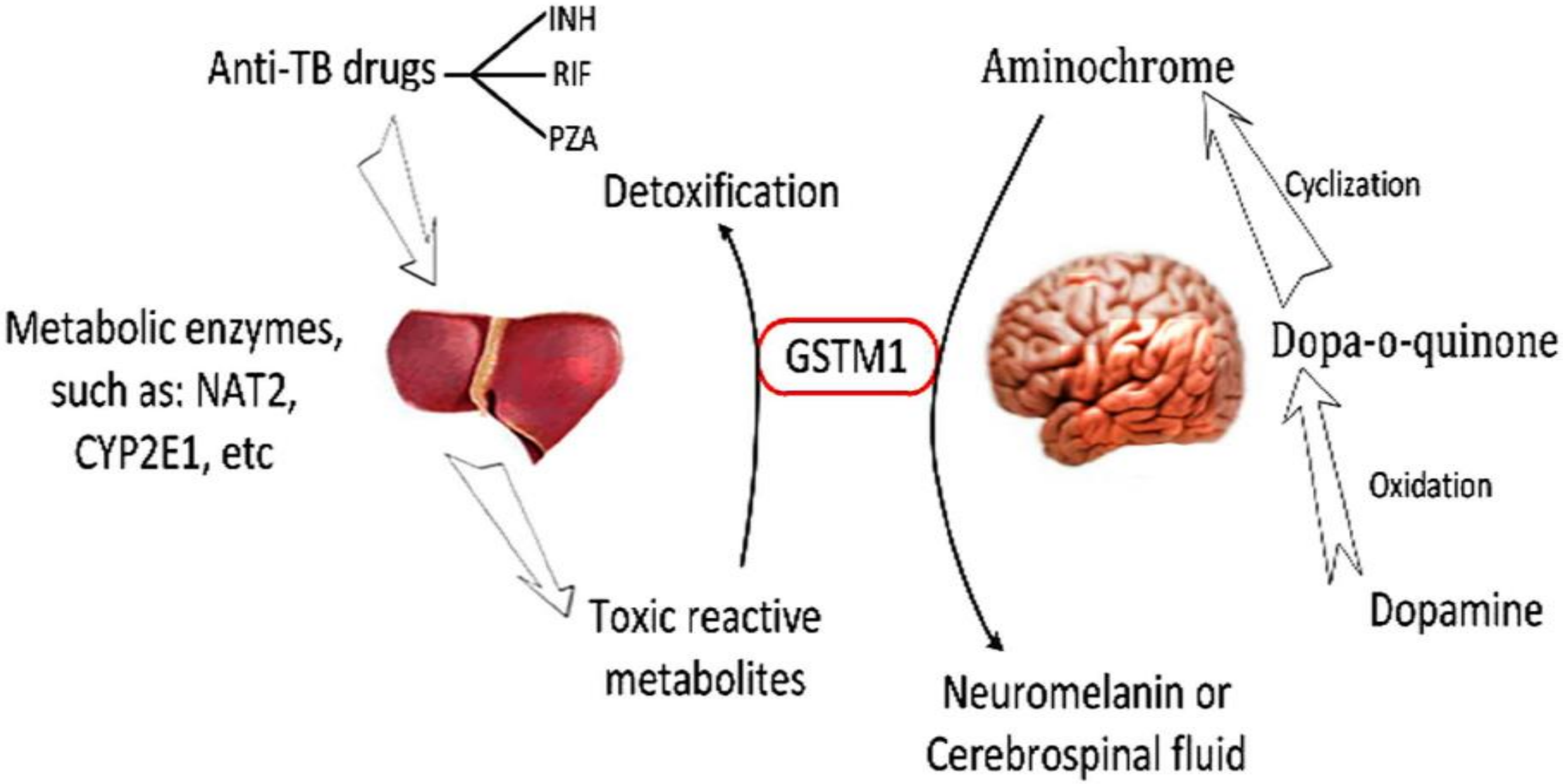


Figure 2. Forest plots from meta-analysis of *GSTM1/GSTT1* polymorphisms and SCZ. (A) Summary of the ORs and corresponding 95% CIs for the *GSTM1* present genotype; (B) summary of ORs and 95% CIs for the *GSTT1* present genotype.

The *GSTM1 present genotype was confirmed to be significantly* associated with both ATDH and SCZ.

GSTM1 in SCZ and ATDH

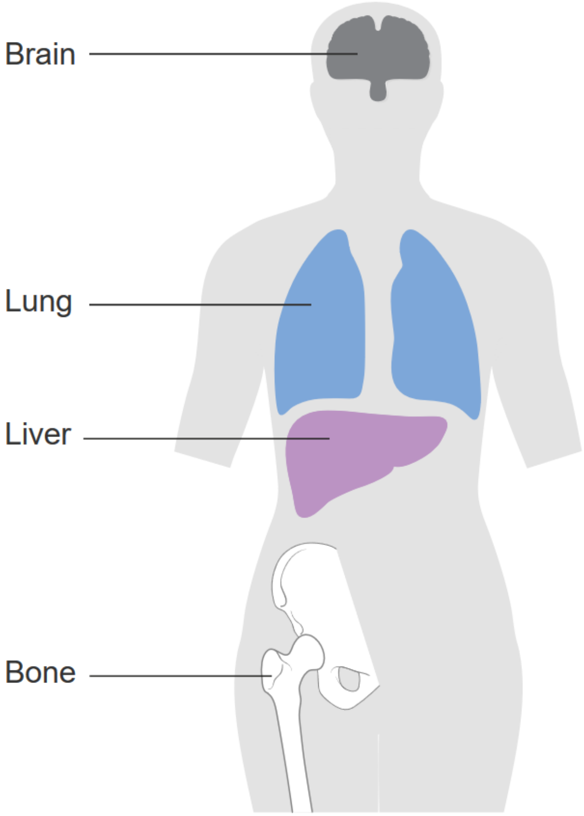

Anti-TB drugs
INH
RIF
PZA
Detoxification
Aminochrome
Cyclization
Metabolic enzymes, such as: NAT2, CYP2E1, etc
GSTM1
Dopa-o-quinone
Oxidation
Dopamine
Toxic reactive metabolites
Neuromelanin or Cerebrospinal fluid

PART

BRCA Metastasis

FOUR

Breast Cancer Metastasis To Bone

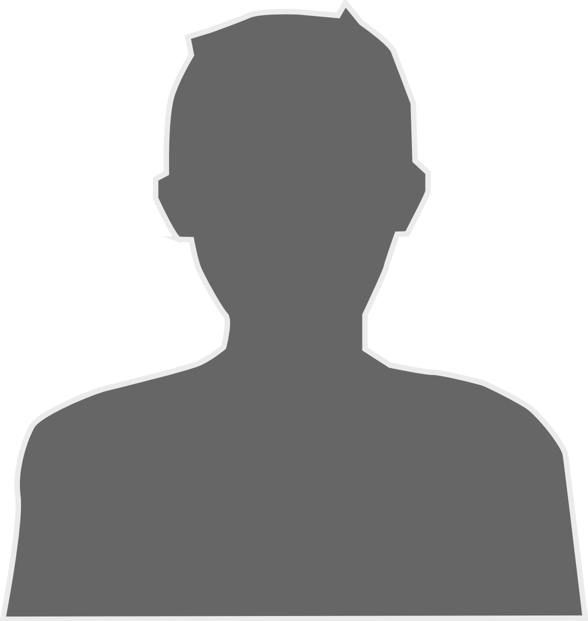


Common sites of metastasis for breast cancer

More than 70% of breast cancer deaths can be attributed to bone metastasis from breast cancer.

It is 60% greater than the next common metastasis pattern.

Breast/Bone Cancer Genes



Manually Curated

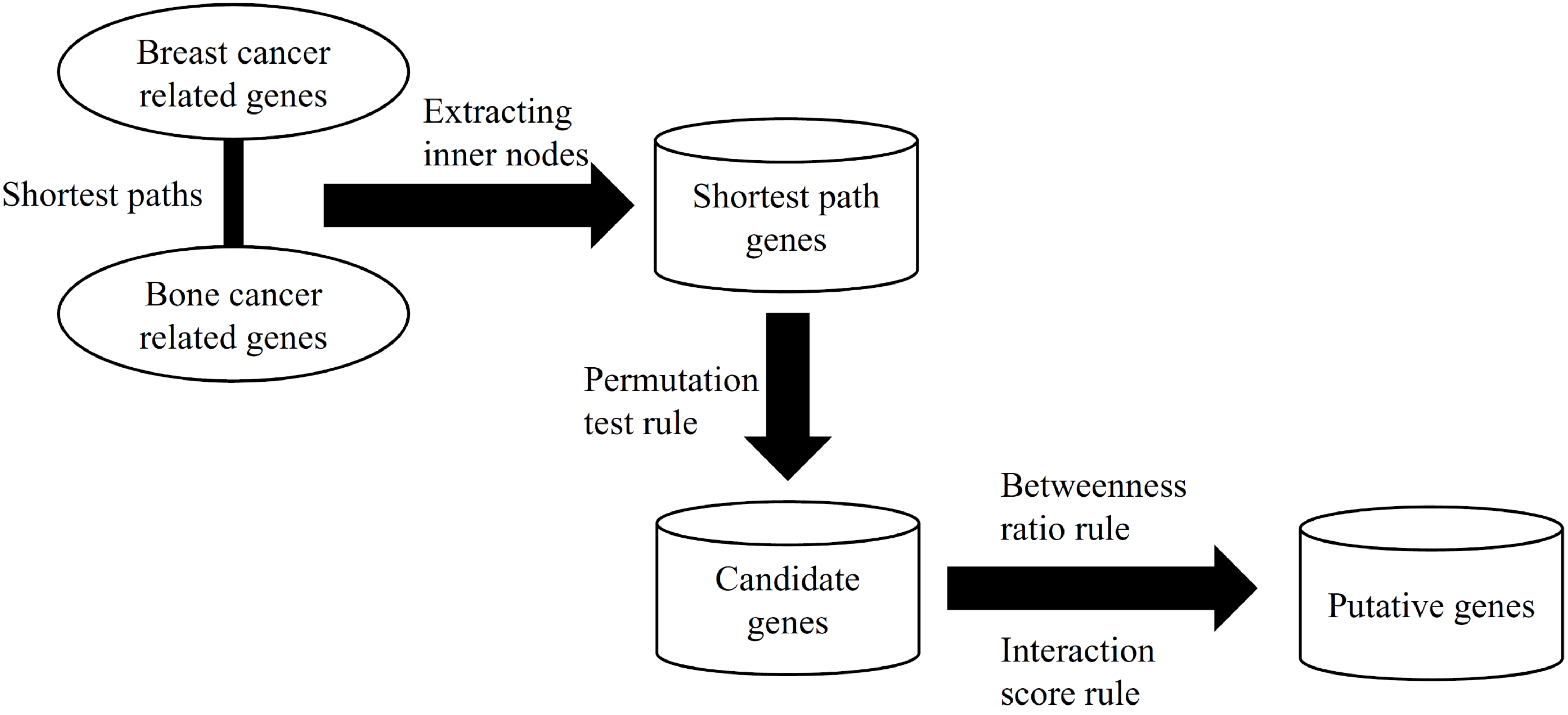
369 Breast Cancer Genes

UniProtKB: Protein Knowledgebase
TSGene: Tumor Suppressor Gene Database

603 Bone Cancer Genes

UniProtKB: Protein Knowledgebase
NCG: Network of Cancer Genes
CTD: Comparative Toxicogenomics Database

Flowchart to illustrate the SP method procedure



Rules

Shortest Path

Nodes on the shortest Path

Permutation Test Rule

$$\mathrm{FDR}(g) = \frac{\theta}{1000}$$

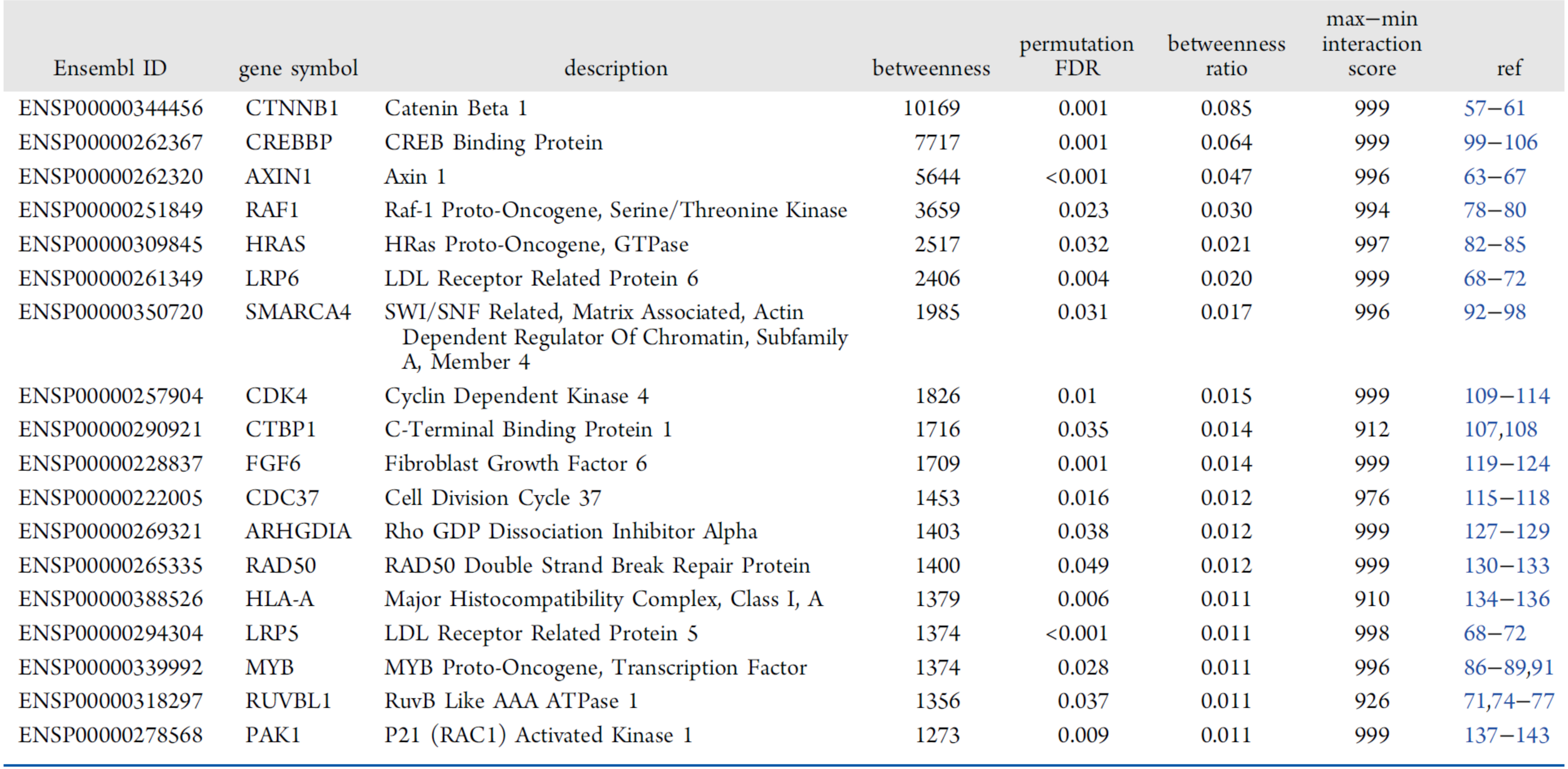
Betweenness Ratio Rule

$$R(g) = \frac{\text{betweenness of } g}{|S_{\text{bone}}| \cdot |S_{\text{breast}}|}$$

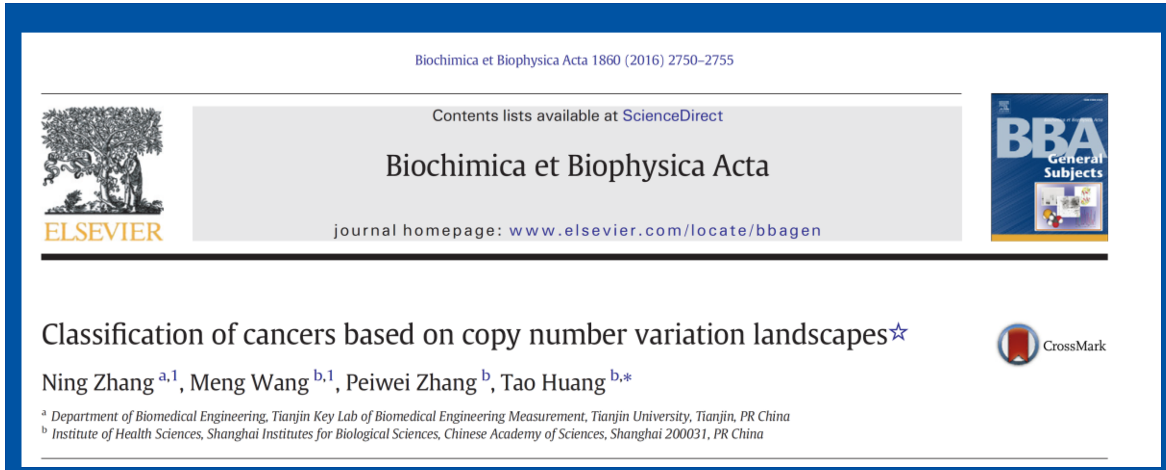
Interaction Score Rule

$$\max{-}\min(g) = \min\{\max\{S(g, x) : x \in S_{\text{bone}}\}, \max\{S(g, x) : x \in S_{\text{breast}}\}\}$$

18 Putative Genes and Their Measurements



Ensembl ID	gene symbol	description	betweenness	permutation FDR	betweenness ratio	max−min interaction score	ref
ENSP00000344456	CTNNB1	Catenin Beta 1	10169	0.001	0.085	999	57−61
ENSP00000262367	CREBBP	CREB Binding Protein	7717	0.001	0.064	999	99−106
ENSP00000262320	AXIN1	Axin 1	5644	<0.001	0.047	996	63−67
ENSP00000251849	RAF1	Raf-1 Proto-Oncogene, Serine/Threonine Kinase	3659	0.023	0.030	994	78−80
ENSP00000309845	HRAS	HRas Proto-Oncogene, GTPase	2517	0.032	0.021	997	82−85
ENSP00000261349	LRP6	LDL Receptor Related Protein 6	2406	0.004	0.020	999	68−72
ENSP00000350720	SMARCA4	SWI/SNF Related, Matrix Associated, Actin Dependent Regulator Of Chromatin, Subfamily A, Member 4	1985	0.031	0.017	996	92−98
ENSP00000257904	CDK4	Cyclin Dependent Kinase 4	1826	0.01	0.015	999	109−114
ENSP00000290921	CTBP1	C-Terminal Binding Protein 1	1716	0.035	0.014	912	107,108
ENSP00000228837	FGF6	Fibroblast Growth Factor 6	1709	0.001	0.014	999	119−124
ENSP00000222005	CDC37	Cell Division Cycle 37	1453	0.016	0.012	976	115−118
ENSP00000269321	ARHGDIA	Rho GDP Dissociation Inhibitor Alpha	1403	0.038	0.012	999	127−129
ENSP00000265335	RAD50	RAD50 Double Strand Break Repair Protein	1400	0.049	0.012	999	130−133
ENSP00000388526	HLA-A	Major Histocompatibility Complex, Class I, A	1379	0.006	0.011	910	134−136
ENSP00000294304	LRP5	LDL Receptor Related Protein 5	1374	<0.001	0.011	998	68−72
ENSP00000339992	MYB	MYB Proto-Oncogene, Transcription Factor	1374	0.028	0.011	996	86−89,91
ENSP00000318297	RUVBL1	RuvB Like AAA ATPase 1	1356	0.037	0.011	926	71,74−77
ENSP00000278568	PAK1	P21 (RAC1) Activated Kinase 1	1273	0.009	0.011	999	137−143



Biochimica et Biophysica Acta 1860 (2016) 2750–2755

Contents lists available at ScienceDirect

Biochimica et Biophysica Acta

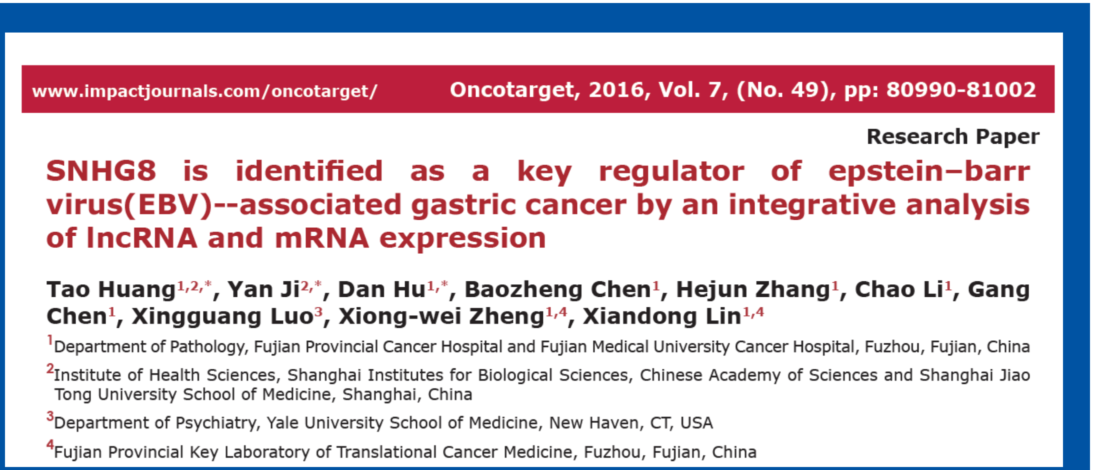
journal homepage: www.elsevier.com/locate/bbagen

ELSEVIER
BBA General Subjects
CrossMark

Classification of cancers based on copy number variation landscapes☆

Ning Zhang [a,1], Meng Wang [b,1], Peiwei Zhang [b], Tao Huang [b,*]

[a] Department of Biomedical Engineering, Tianjin Key Lab of Biomedical Engineering Measurement, Tianjin University, Tianjin, PR China
[b] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, PR China



www.impactjournals.com/oncotarget/ Oncotarget, 2016, Vol. 7, (No. 49), pp: 80990-81002

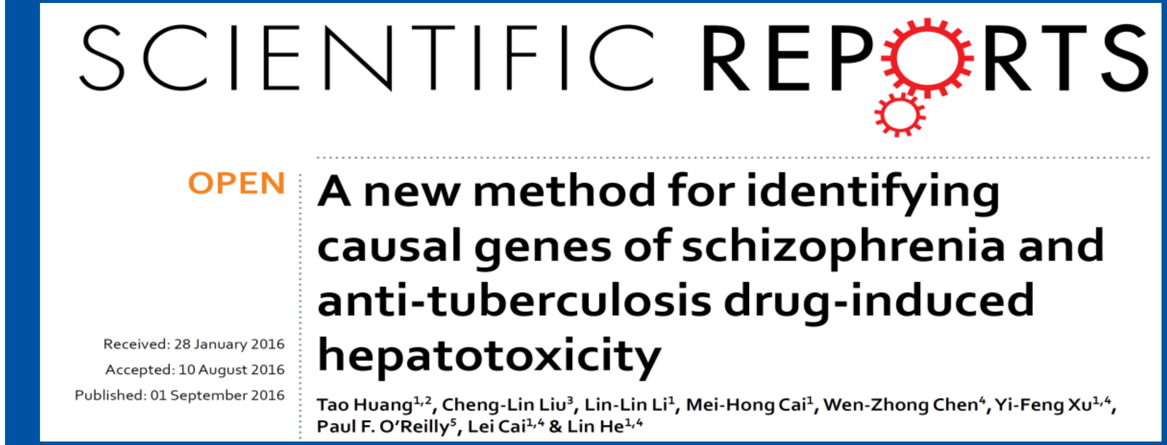
Research Paper

SNHG8 is identified as a key regulator of epstein–barr virus(EBV)--associated gastric cancer by an integrative analysis of lncRNA and mRNA expression

Tao Huang[1,2,*], Yan Ji[2,*], Dan Hu[1,*], Baozheng Chen[1], Hejun Zhang[1], Chao Li[1], Gang Chen[1], Xingguang Luo[3], Xiong-wei Zheng[1,4], Xiandong Lin[1,4]

[1] Department of Pathology, Fujian Provincial Cancer Hospital and Fujian Medical University Cancer Hospital, Fuzhou, Fujian, China
[2] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Shanghai Jiao Tong University School of Medicine, Shanghai, China
[3] Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA
[4] Fujian Provincial Key Laboratory of Translational Cancer Medicine, Fuzhou, Fujian, China

THANKS



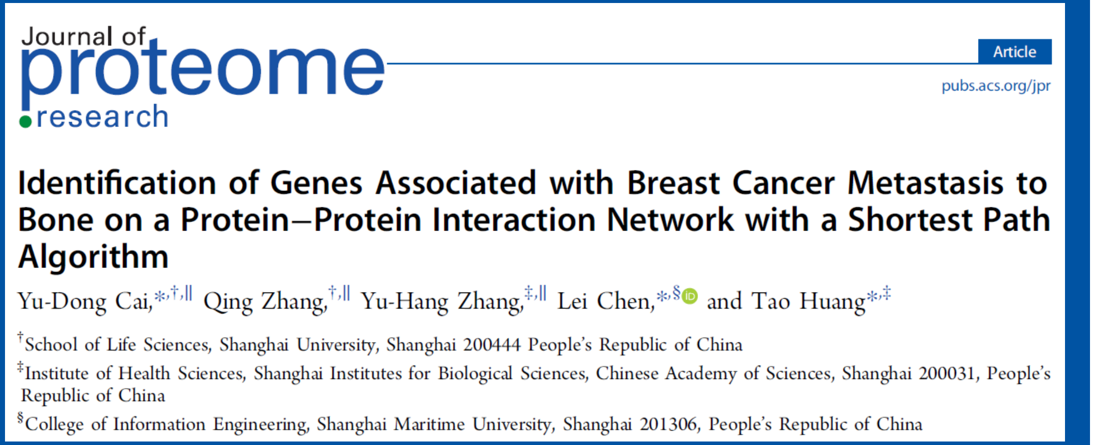
SCIENTIFIC REPORTS

OPEN

A new method for identifying causal genes of schizophrenia and anti-tuberculosis drug-induced hepatotoxicity

Received: 28 January 2016
Accepted: 10 August 2016
Published: 01 September 2016

Tao Huang[1,2], Cheng-Lin Liu[3], Lin-Lin Li[1], Mei-Hong Cai[1], Wen-Zhong Chen[4], Yi-Feng Xu[1,4], Paul F. O'Reilly[5], Lei Cai[1,4] & Lin He[1,4]



Journal of proteome research

Article

pubs.acs.org/jpr

Identification of Genes Associated with Breast Cancer Metastasis to Bone on a Protein–Protein Interaction Network with a Shortest Path Algorithm

Yu-Dong Cai,[*,†,∥] Qing Zhang,[†,∥] Yu-Hang Zhang,[‡,∥] Lei Chen,[*,§] and Tao Huang[*,‡]

[†] School of Life Sciences, Shanghai University, Shanghai 200444 People's Republic of China
[‡] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, People's Republic of China
[§] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, People's Republic of China