A systematic view of computational methods for identifying driver genes based on somatic mutation data

Supplementary Information

Yingxin Kan[1] , Limin Jiang[1] , Jijun Tang[1,2,3,] Yan Guo[4] and Fei Guo[1,*]

[1] School of Computer Science and Technology, College of Intelligence and Computing, Tianjin University, Tianjin, China
[2] School of Computational Science and Engineering, University of South Carolina, Columbia, U.S.
[3] Key Laboratory of Systems Bioengineering (Ministry of Education), Tianjin University, Tianjin, China
[4]Comprehensive cancer center, Department of Internal Medicine, University of New Mexico, Albuquerque, U.S.

Contents

Overview

We take total 10 kinds of cancer from The Cancer Genome Atlas (TCGA) (including Breast Invasive Carcinoma (BRCA), Colon Adenocarcinoma (COAD), Glioblastoma Multiforme (GBM), Head and Neck Squamous Cell Carcinoma (HNSC), Liver Hepatocellular Carcinoma (LIHC), Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Ovarian Serous Cystadenocarcinoma (OV), Pancreatic Adenocarcinoma (PAAD), Skin Cutaneous Melanoma (SKCM)) as datasets to evaluate the performance of three algorithms. They are MutSigCV[1], OncodriveCLUST[2] and OncodriveFML[3]. Then we analyze results in four various ways (CGC, OG/TSG, Q-value and QQ-plot). CGC is to show the proportion of genes from Cancer Gene Census (CGC) in candidate driver genes. OG/TSG classifies driver genes into oncogenes and tumor suppressor genes. Q-value shows significance of candidate drivers. QQ-plot reflects results on the whole.

The rest of related figures and tables are shown in this Supplementary Information. There are QQ-plots for OncodriveCLUST and OncodriveFML and heatmaps of OncodriveFML. In addition, we adjust parameters to conduct experiments based on TCGA-BRCA project and analysis of comparing results is shown in Supplementary Information Figure S3. We classify CGC and potential driver genes into ten kinds of cancer, which are included in Supplementary Information Table S1.

Figure S1

We make QQ-plots to confirm whether p-values obtained by the current statistical model meet the expected value and whether the statistical model is reasonable. We believe that the lower left points, as the passenger genes, should meet the expected value while points at the top right corner exceeding the expected values are regarded as driver genes. Here we show QQ-plots for two algorithms(OncodriveCLUST and OncodriveFML) in Figure S1, in which x-axis represents the expected $-\log_{10} p-value$ and y-aixs represents the observed -$\log_{10} p-value$.

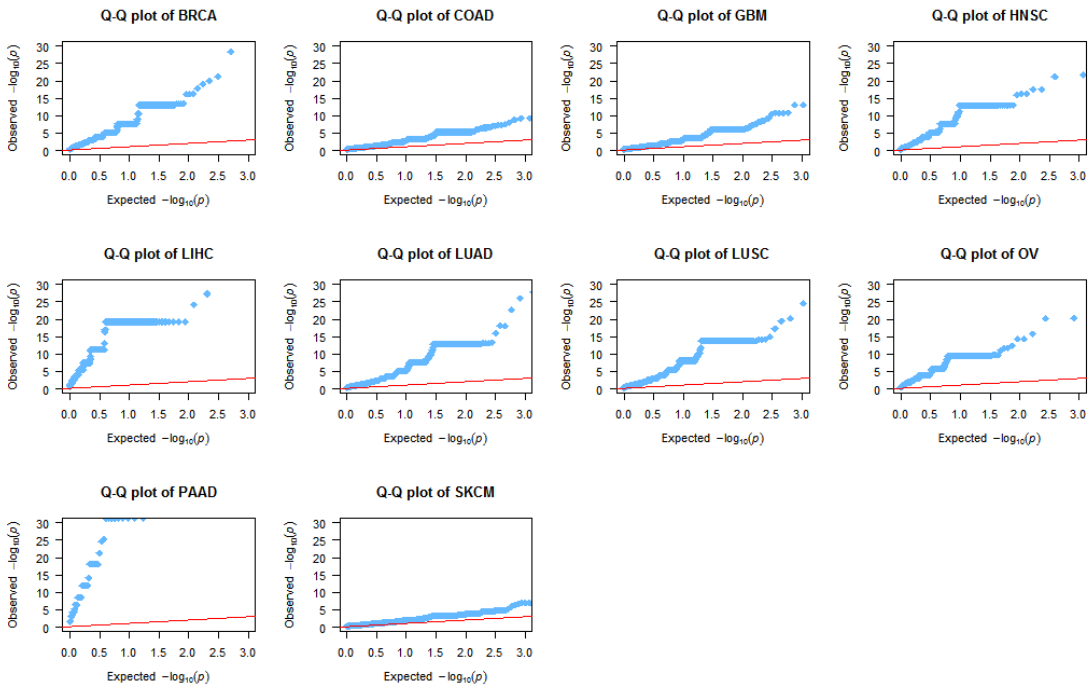

Figure S1A. QQ-plot of OncodriveCLUST. As we can see, the lower left points are far away expected p-values in most kinds of cancer, which indicates this model may predict many false positive genes.

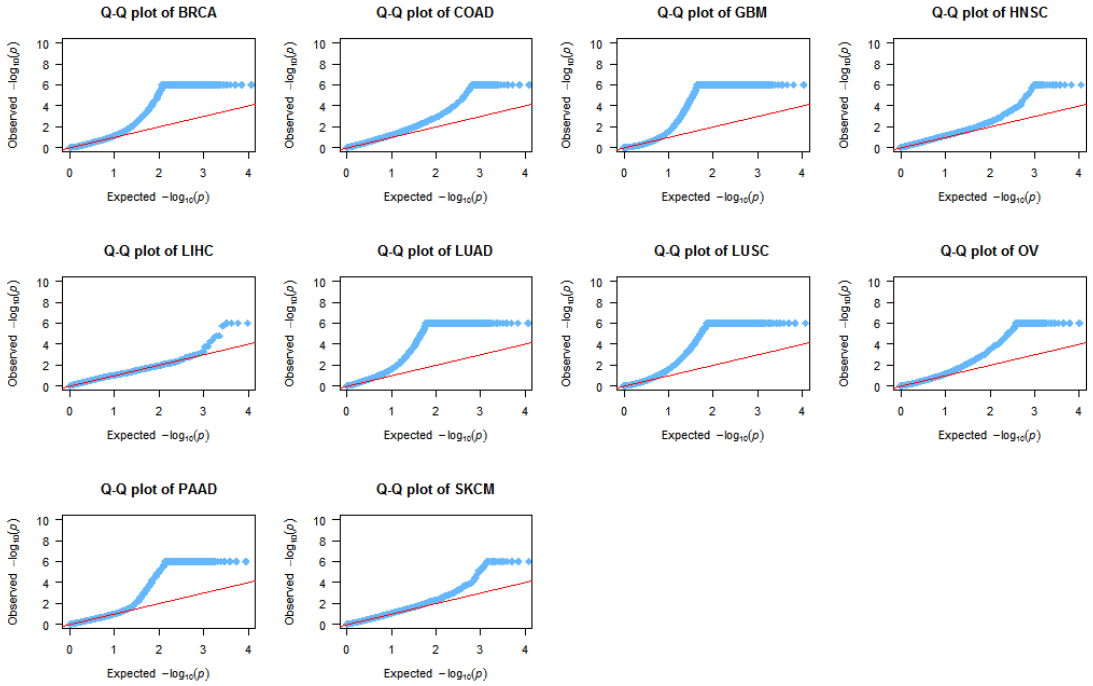

Figure S1B. QQ-plot of OncodriveFML. As we can see, the rationality of this model can be almost certain, while the observed values of points in the higher right corner are not high which means significance of candidate driver genes isn't evident.

Figure S2

We use -$\log_{10} q - value$ of each gene computed by every algorithm as the value of the cell in the heatmap and only candidate driver genes occurring in the more than five kinds of cancer are selected. Figure S2 shows the heatmap of OncodriveFML.

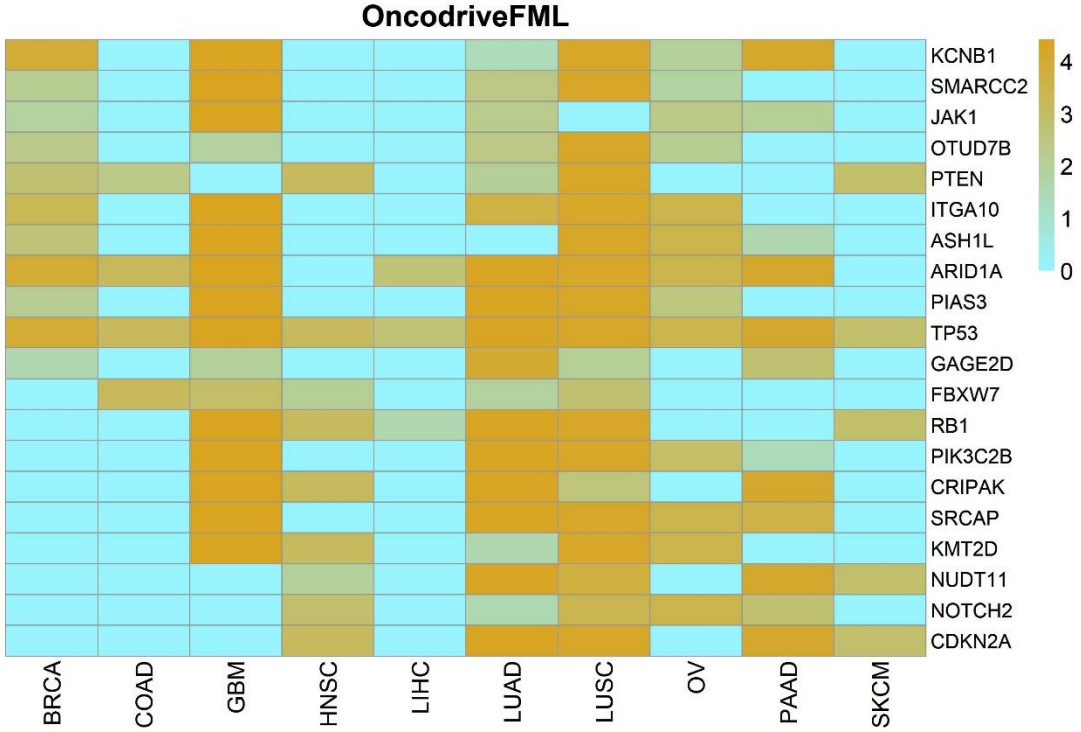

Figure S2. Heatmap of OncodriveFML. TP53 is also predicted as the driver genes in more than five types of cancer by OncodriveFML.

Figure S3

We adjust parameters of different methods to compare their performance based on TCGA-BRCA project. MutSigCV is an excellent tool considering heterogeneity of samples, genes and mutations. It is encapsulated so well that there's no adjustable param within command line interface. Therefore, we focus on other two methods and show the results of OncodriveCLUST and OncodriveFML. We apply venn plots to display overlap of drivers identified by each method on the basis of various parameters.

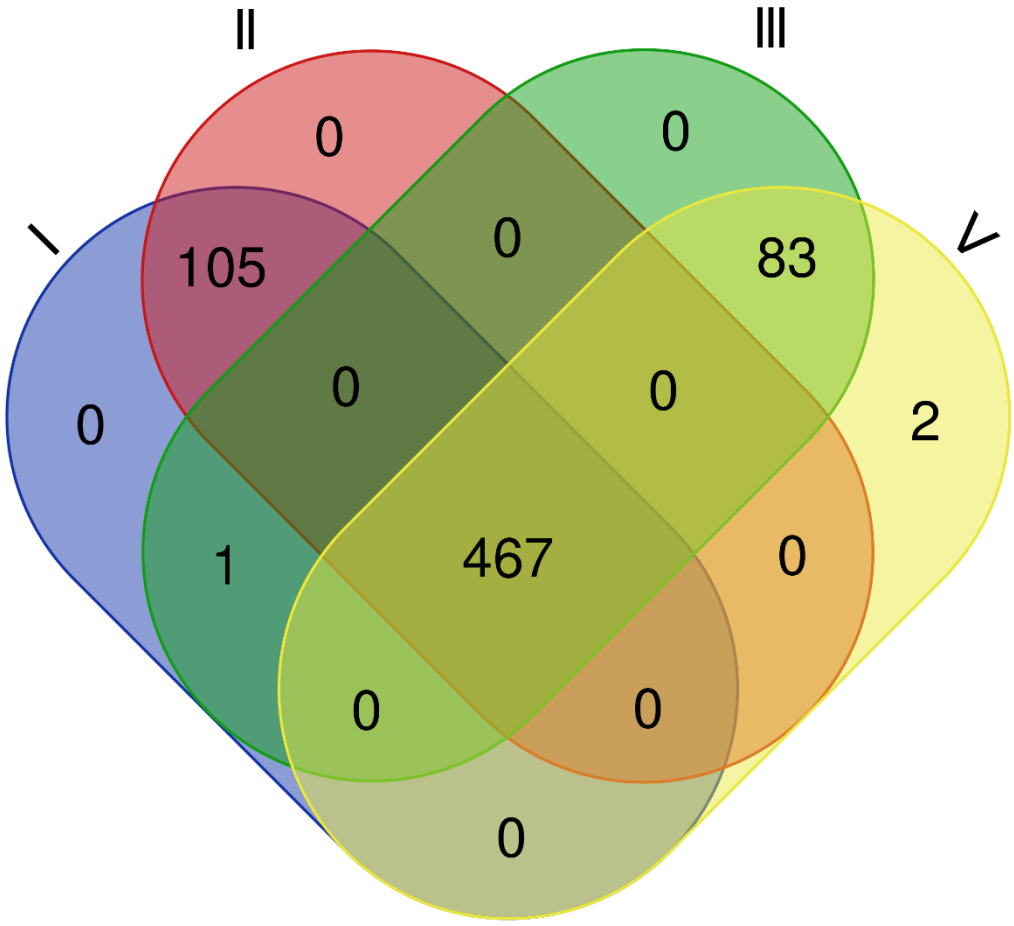

Figure S3A. Venn plot of OncodriveCLUST. We adjust two parameters: muts(minimum number of mutations of a gene to be included in the analysis) and dist(intra cluster maximum distances of amino acids between two positions that can be grouped). The values of muts and dist in each experiment are: I - muts=3, dist=5; II - muts=3, dist=10; III - muts=5, dist=5; IV - muts=5, dist=10. We also run OncodriveCLUST when muts is equal to 10 for two times but we only receive three

drivers. We believe that minimum number of mutations of a gene, i. e. muts, cannot be too large otherwise too many genes will be filtered.

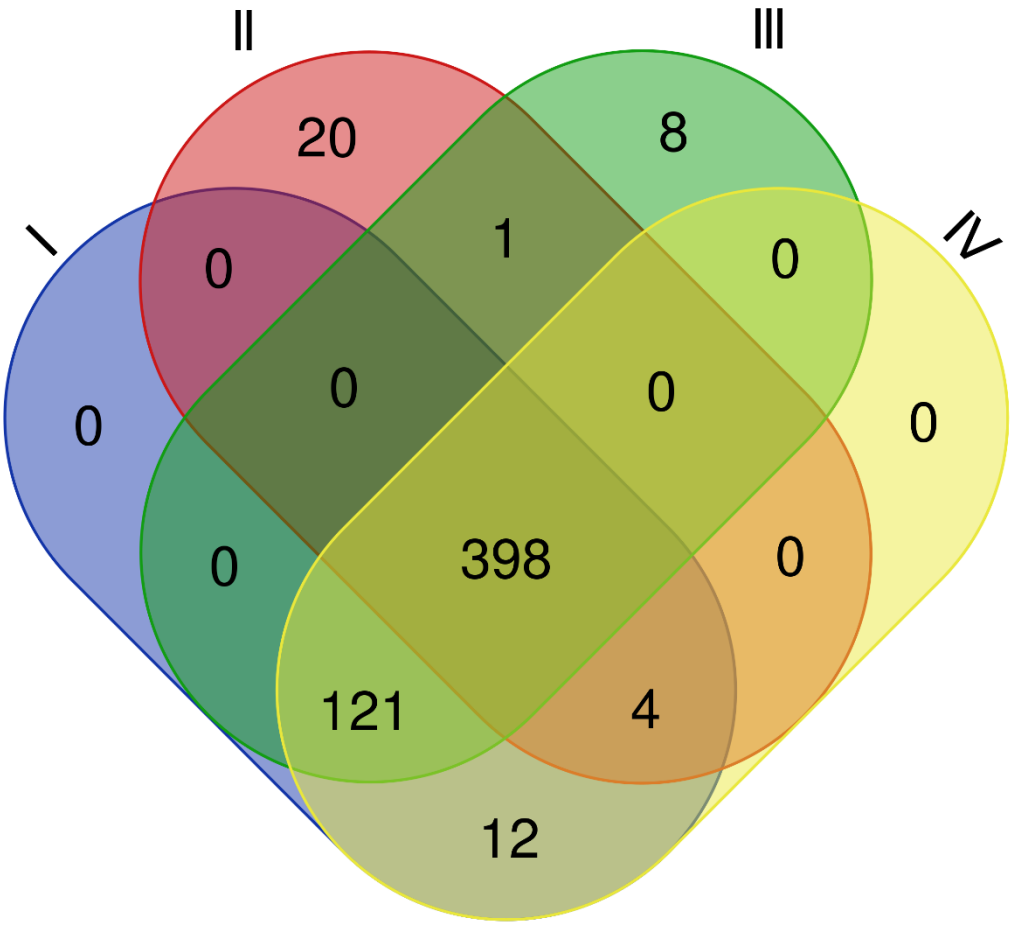

Figure S3B. Venn plot of OncodriveFML. We adjust three parameters: method (the type of operation that is applied to observed and simulated scores before comparing them), sampling (the minimum number of simulations to be performed) and sampling_max (the maximum number of simulations to be performed). The values of the three parameters in each experiment are: I - method: amean, sampling:100k, sampling:1000k; II - method: amean, sampling:10k, sampling:100k; III - method: gmean, sampling:100k, sampling:1000k; IV - method: gmean, sampling:10k, sampling:100k. We find that sampling and sampling_max have bigger effect on results according to the venn plot.

Supplementary Table S1

We classify CGC and potential driver genes (not occur in CGC database) identified by three methods that occur in more than 5 kinds of cancer into different tumor types, just shown in Table S1A and B. The two tables illustrate each gene in Table 3 of our paper belong to which cancer types and it will help researchers study specific cancers and genes they are interested in.

Table S1A. Summary of CGC Genes that Occur in More than 5 Kinds of Cancer Classified by Tumor Type

CGC Genes									
BRCA	**COAD**	**GBM**	**HNSC**	**LIHC**	**LUAD**	**LUSC**	**OV**	**PAAD**	**SKCM**
ARID1A	ARID1A	ARID1A	B2M	ARID1A	ARID1A	ARID1A	ARID1A	ARID1A	B2M
CLIP1	B2M	B2M	CDKN2A	CDKN2A	B2M	BRAF	BCL9L	BCL9L	BRAF
IDH1	BCL9L	BCL9L	FBXW7	GNAS	BCL9L	CDKN2A	CLIP1	CDKN2A	CDKN2A
JAK1	BRAF	BRAF	KMT2D	IDH1	BRAF	CLIP1	FBXW7	CLIP1	IDH1
KRAS	FBXW7	CDKN2A	NOTCH2	KEAP1	CDKN2A	FBXW7	GNAS	GNAS	KRAS
NF1	GNAS	CLIP1	PIK3CA	KRAS	FBXW7	IDH1	JAK1	JAK1	NF1
PIK3CA	KRAS	FBXW7	PTEN	PIK3CA	GNAS	KEAP1	KEAP1	KRAS	NRAS
PTEN	NRAS	IDH1	RB1	PTEN	IDH1	KMT2D	KMT2D	NOTCH2	PTEN
TP53	PIK3CA	JAK1	TP53	RB1	JAK1	KRAS	KRAS	TP53	RB1
	PTEN	KEAP1		TP53	KEAP1	NF1	NF1		TP53
	TP53	KMT2D			KMT2D	NOTCH2	NOTCH2		
		KRAS			KRAS	NRAS	NRAS		
		NF1			NF1	PIK3CA	PIK3CA		
		NRAS			NOTCH2	PTEN	PTEN		
		PIK3CA			NRAS	RB1	RB1		
		PTEN			PIK3CA	TP53	TP53		
		RB1			PTEN				
		TP53			RB1				
					TP53				

Table S1B. Summary of Potential Driver Genes that Occur in More than 5 Kinds of Cancer Classified by Tumor Type

Potential Driver Genes									
BRCA	**COAD**	**GBM**	**HNSC**	**LIHC**	**LUAD**	**LUSC**	**OV**	**PAAD**	**SKCM**
ASH1L	CD4	ASH1L	C6	CDC27	CD4	ASH1L	ASH1L	ASH1L	NBPF1
CD4	PRB2	C6	CRIPAK		COL11A1	C6	C6	C6	NUDT11
FAM104A		CD4	IL32		CRIPAK	CD4	CD4	CDC27	PRB2
FAM151A		CDC27	KRTAP4-1		FAM104A	CDC27	CDC27	COL11A1	
GAGE2D		COL11A1	NBPF1		FAM151A	COL11A1	COL11A1	CRIPAK	
GOT1L1		CRIPAK	NDUFAF2		GAGE2D	CRIPAK	FAM104A	GAGE2D	
HAX1		FAM104A	NUDT11		HAX1	FAM104A	FAM151A	GOT1L1	
ISG20L2		FAM151A	PODXL		IL32	FAM151A	GOT1L1	HAX1	
ITGA10		GAGE2D	PRB2		ITGA10	GAGE2D	HAX1	IL32	
KBTBD6		GOT1L1			KBTBD6	GOT1L1	IL32	ISG20L2	
KCNB1		HAX1			KCNB1	HAX1	ISG20L2	KBTBD6	
MT4		IL32			KRTAP4-1	IL32	ITGA10	KCNB1	
MUC7		ISG20L2			LNP1	ISG20L2	KCNB1	KRTAP4-1	
NBPF1		ITGA10			MT4	ITGA10	LNP1	LNP1	
NDUFAF2		KBTBD6			MUC7	KBTBD6	MT4	MUC7	
NUDT11		KCNB1			NBPF1	KCNB1	NBPF1	NBPF1	
OR10Z1		KRTAP4-1			NDUFAF2	KRTAP4-1	OR10Z1	NDUFAF2	
OTUD7B		LNP1			NUDT11	LNP1	OTUD7B	NUDT11	
PIAS3		MT4			OTUD7B	MT4	PIAS3	OR10Z1	
PODXL		MUC7			PIAS3	MUC7	PIK3C2B	PIK3C2B	
PRB2		NBPF1			PIK3C2B	NBPF1	PODXL	PODXL	
SMARCC2		NDUFAF2			PRB2	NUDT11	PRB2	PRB2	
UHRF1BP1		OR10Z1			SMARCC2	OR10Z1	SMARCC2	SRCAP	
		OTUD7B			SRCAP	OTUD7B	SRCAP	UHRF1BP1	
		PIAS3				PIAS3	UHRF1BP1		
		PIK3C2B				PIK3C2B			
		PODXL				PODXL			
		PRB2				PRB2			
		SMARCC2				SMARCC2			
		SRCAP				SRCAP			
		UHRF1BP1				UHRF1BP1			

References

1 Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancerassociated genes. Nature. 2013;499(7457):214–218.

2 Tamborero D, Gonzalezperez A, Lopezbigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29(18):2238–2244.

3 Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome biology. 2016;17(1):128.