

# Gene Cluster Expression Index (GCEI) and Potential Indications for Targeted Therapy and Immunotherapy

Aibing Rao

Shenzhen Luwei (Biomanifold) Biotechnology Limited

10th Floor, Clou Building B, Baoshen Road,

Nanshan District, Shenzhen, PR China, ZIP: 518057

*Email:* aibing.rao@enlightendx.com

**Abstract** Lung cancer recurrence risk was demonstrated to be related to driver gene and immunotherapy target gene cluster expression abnormality. Nine clusters seeded with driver genes ALK, BRAF, EGFR, MET, NTRK, RAS, RET, ROS1, TP53 and two immunotherapy target genes PDCD1 and CTLA4 were investigated respectively to predict lung cancer recurrence. The cluster of a seed was pre-selected to include fusion partner genes in the case of gene fusion, ligands, its pseudogenes, upstream and downstream co-expressors or inhibiting genes, effectors directly related to important pathways, etc. For each cluster, a gene cluster expression index (GCEI) was defined in two steps: Firstly, apply the univariate ROC of using each member's expression vector to predict recurrences to label a patient sample as either normal or abnormal; Secondly, apply the percentage of abnormal genes in the cluster to predict recurrences to derive an optimal threshold so that a cluster member voting strategy can be achieved and a sample is labeled as abnormal (with respect to the cluster expression profile) if the the percentage of abnormal genes for the sample is greater than or equal to the threshold and as normal vice versa. Combinatory GCEI was developed as a binary string concatenating the individual GCEI corresponding to the individual cluster in an ordered list of driver or other important gene seeds. It showed that the recurrence risk of the abnormal group is typically 50% to 200% higher than the normal counterpart. Finally it was proposed and discussed to expand targeted therapy and immunotherapy to the abnormal group defined by GCEI.

**Background** Molecular profiling such as DNA-based mutation panels and proteomics have been demonstrated great success in oncology for personalized medicine. Transcriptome profiling has emerged to be another promising opportunity as complement and expansions to the DNA-based approach and as new tools to further advance clinical oncology.

**Methods** Lung cancer gene expression GEO data sets were downloaded, normalized, combined and analyzed. A novel approach was presented to analyze expression abnormality of important gene clusters with seeds including drivers such as ALK, BRAF, EGFR, MET, NTRK, RAS, RET, ROS1, TP53 or immunotherapy target PDCD1 and CTLA4, etc. A cluster was pre-specified for each seed and included the fusion partners in the case of translocation, ligands, activators, inhibitors, effectors, co-stimulators in the important pathways, etc. Each cluster member was labeled as normal or abnormal (up or down) with the univariate ROC by using its expression to predict recurrences. Cluster level labeling of expression state (normal or abnormal) was via a dynamic voting strategy, of which the voting threshold was set as the optimal cutoff on the ROC

associated with the univariate model of using the percentage of the abnormal members to predict recurrences. Given an ordered list of important genes, a binary string of the same length was encoded by assigning 0 for *normal* and 1 for *abnormal* representing the cluster expression state of the corresponding position, called gene cluster expression index (GCEI) signature. Finally lung cancer recurrences were assessed and compared based on GCEI states and the combinations.

**Results** The recurrence risks of single gene normal group ( $GCEI = 0$ ) vs abnormal group ( $GCEI = 1$ ) were as follows, ALK: 17% vs. 55% for all stages, 13% vs. 42% for Stage I, 36% vs. 67% for Stage II-IV; BRAF: 23% vs. 49% for all stages, 15% vs. 36% for Stage I, 54% vs. 59% for Stage II-IV; EGFR: 25% vs. 47% for all stages, 17% vs. 33% for Stage I, 54% vs. 59% for Stage II-IV; MET: 25% vs. 44% for all stages, 17% vs. 29% for Stage I, 51% vs. 60% for Stage II-IV; NTRK: 19% vs. 52% for all stages, 13% vs. 40% for Stage I, 44% vs. 63% for Stage II-IV; RAS: 24% vs. 51% for all stages, 16% vs. 35% for Stage I, 47% vs. 65% for Stage II-IV; RET: 19% vs. 50% for all stages, 14% vs. 35% for Stage I, 40% vs. 65% for Stage II-IV; ROS1: 23% vs. 48% for all stages, 17% vs. 32% for Stage I, 45% vs. 64% for Stage II-IV; TP53: 23% vs. 50% for all stages, 15% vs. 38% for Stage I, 49% vs. 64% for Stage II-IV; and for the immunotherapy target gene: CTLA4: 26% vs. 49% for all stages, 14% vs. 38% for Stage I, 53% vs. 62% for Stage II-IV; PDCD1: 28% vs. 48% for all stages, 16% vs. 37% for Stage I, 54% vs. 61% for Stage II-IV. In addition, taking 9-driver gene GCEI and summarizing number of '1', the count of abnormal driver genes,  $N$ , and then comparing the population of  $N \leq 5$  vs.  $N > 5$ , the recurrence risks were: 19% vs. 59% for all stages, 13% vs. 49% for Stage I, 41% vs. 66% for Stage II-IV. Hence most of the cases the recurrence risk is 1.5 to 3 times higher for patient group with abnormally expressed gene clusters than normally expressed.

**Discussion** Precision medicine based on RNA expression analysis is discussed and it is conjectured to apply targeted therapy or immunotherapy to lung cancers based on the related gene expression status as determined by the cluster member voting strategy. This can serve as an extension and complement to the current DNA-based tests, especially for a majority of patients who have been tested negative based on the conventional tests and have possibly missed the potential treatment benefit.

**Abbreviations** GCEI: Gene Cluster Expression Index; GEO: Gene Expression Omnibus; ROC: Receiver Operating Curve; AUC: Area Under the Curve; FPR: False Positive Rate; TPR: True Positive Rate PPV: Positive Prediction Value.

**Keywords** Transcriptome Profiling; Gene Cluster Expression Index; RNA Expression Analysis; Multivariate Modeling; Lung Cancer; Targeted Therapy; Immunotherapy.

## 1 Introduction

Gene expression or transcriptome profiling has been extensively explored in the past 20 years in oncology and there are several multi-gene Rna tests have been put in practical clinical use for human cancers[44, 46, 58]. For lung cancer there have been a lot of gene expression signatures published for prognosis prediction[7, 9, 26, 29, 33, 63, 64, 67, 71] and a comprehensive evaluation was performed by Tang et al.[59] However there is little research regarding using gene expression profiling for targeted therapy or immunotherapy. The current standard approach for selection of targeted therapy is via matching particular gene mutations[39], and the selection of immunotherapy is via routine tests such as pathological immunoassay (IHC) for protein expression of PD1 or

PD-L1, via DNA-based NGS assessment of Tumor Mutation Burden (TMB), Mismatch Repair (MMR), and Micro-satellite Instability (MSI). Transcriptome profiling analysis has emerged as promising biomarkers to cancer treatment and showed encouraging clinical results[10, 57]. In NSCLC, study showed that gene expression profiling might have better prognostic prediction power than mutation status[38] in particular scenarios. Hence, RNA profiling analysis is promising and will be an important direction as a complement or even better choice than the current IHC and DNA-based approaches in precision medicine of cancers. In the following we present a framework with novel transcriptome analysis algorithms to assign expression abnormality status to important driver genes or immunotherapy target genes based on member smart voting within a clustered gene set seeded at considering gene for lung cancer. The patient populations in different expression state have been showed to have dramatically different clinical prognosis risks and hence requires different personalized treatment considerations. The general framework is applicable and implementable to other cancer types with strongly related genes with clusters.

## 2 Materials and methods

### 2.1 Gene Expression Data

Two microarray datasets downloaded from Gene Expression Omnibus (GEO) databases are GSE30219 [48] and GSE31210 [40]. There were 483 lung cancers with none-empty recurrence labels, of which 204 cases were labeled as recurred within two years since the diagnosis, accounting for 42%. Two data sets were respectively normalized using IQR (Inter Quantile Range) method, namely, the the quartiles ( $Q_1, Q_3$ ) of the original expression data were linearly mapped to the unit interval  $(0, 1)$ . The normalization procedure was applied first at the sample dimension and then at the gene dimension. Then taking the common genes, two normalized data subsets of the common genes plus clinical variables were stacked together to form a combined analysis data set. There are about 17000 common genes, the cases of different stages counted as 310 (I), 111 (II), 44 (III-IV), and 8 (unknown). Average age is 61, with the youngest 15 and the eldest 84. There are 331 male patients.

### 2.2 Pre-selected Gene Clusters

Nine driver gene ALK, BRAF, EGFR, MET, NTRK, RAS, RET, ROS1, TP53 and two immunotherapy target genes PDCD1, CTLA4 were used for gene cluster expression analysis. Take ALK as an example, in ALK-positive NSCLC population, almost 100 ALK fusion partners were cataloged [41], together with String and genecard.org description, 107 genes were selected. The gene clusters are listed in Table 1. Note that the clusters are not mutually exclusive and one member may appear in different clusters.

### 2.3 Gene Cluster Expression Index (GCEI)

The goal is to assign a sample a binary index for a given gene cluster, called *gene cluster expression index* (GCEI). It consists of two steps, first to determine the expression index of all cluster members, second to apply a smart voting procedure to define the GCEI for the sample.

#### 2.3.1 ROC of Univariate Model to Determine Single Gene Expression Abnormality with Respect to Recurrence

Given a gene seed, use each cluster member listed in Table 1 to predict recurrence and draw ROC to obtain an optimal cutoff, which is set at the ROC position closest to the top-left corner of

Table 1: Pre-selected Gene Clusters for Important Lung Cancer Genes.

SEED	GENE
ALK	ADAM17, AKAP8L, ALK, ALKAL2, ATAD2B, ATIC, ATP13A4, BCL11A, BIRC6, C12ORF75, C9ORF3, CAMKMT, CBL, CDK15, CEPBZ, CEP55, CLIP1, CLIP4, CLTC, CMTR1, CRIM1, CUX1, CYBRD1, DCHS1, DCTN1, DYSF, EIF2AK3, EML4, EML6, EPAS1, ERC1, FBN1, FBXO11, FBXO36, FRS2, FUT8, GCC2, HIP1, IRS1, ITGAV, KIF5B, KLC1, LCLAT1, LIMD1, LMO7, LPIN1, LYPD1, MAPK1, MAPK3, MDK, MPRIP, MSN, MTA3, MYT1L, NCOA1, NPM1, NYAP2, PHACTR1, PICALM, PLEKHA7, PLEKHH2, PLEKHM2, PPFIBP1, PPM1B, PRKARIA, PRKCB, PTN, RANBP2, RBM20, SEC31A, SHC1, SLC16A7, SLMAP, SMPD1, SMPD2, SMPD3, SMPDL3A, SMPDL3B, SOCS5, SORCS1, SOS1, SPECC1, SPTBN1, SQSTM1, SRBD1, SRD5A2, STRN, SWAP70, TACR1, TANC1, TCF12, TFG, THADA, TNIP2, TOGARAM2, TPM4, TPR, TRIM66, TSPYL6, TTC27, TUBE, VIT, VKORC1L1, WDPCP, WDR37, WNK3, YAP1
BRAF	BRAF, MAP2K1, MAP2K2, MAP2K3, MAP2K4, MAP2K5, MAP2K6, MAP2K7, MAP3K1, MAP3K10, MAP3K11, MAP3K12, MAP3K13, MAP3K14, MAP3K14.AS1, MAP3K19, MAP3K2, MAP3K20, MAP3K21, MAP3K3, MAP3K4, MAP3K5, MAP3K6, MAP3K7, MAP3K7CL, MAP3K8, MAP3K9, MAP4K1, MAP4K2, MAP4K3, MAP4K4, MAP4K5, RAF1
EGFR	AREG, BRAF, BTC, CTNNB1, EGF, EGFR, EREG, MUC1, NRG1, NRG2, NRG3, NRG4, NRGN, RGS16, SRC, TGFA
MET	GAB1, GRB2, HGF, MET, PIK3R1, PLCG1, SRC, STAT3
NTRK	AFAP1, AGBL1, AGBL2, AGBL3, AGBL5, ARHGEP2, BCAN, BCR, BTBD1, CD74, CHTOP, CTRC, DAB2IP, EML4, ETV6, GRIPAP1, HNRNPA2B1, IGFBP7, IRF2BP2, LMNA, LRRC71, LYN, MPRIP, MRPL24, MYO5A, NACC2, NFASC, NTRK1, NTRK2, NTRK3, PAN3, PDE4DIP, PLEKHA6, PPL, QKI, RABGAP1L, RBPMS, RFWD2, SCYL3, SLITRK1, SLITRK2, SLITRK3, SLITRK4, SLITRK5, SLITRK6, SQSTM1, STRN, TFG, TLE4, TP53, TPM3, TPM4, TPR, TRAF2, TRIM24, TRIM63, UBE2R2, VCL
RAS	FRAS1, GRASP, HRAS, HRASLS, HRASLS2, HRASLS5, KRAS, MRAS, NRAS, RASA1, RASA2, RASA3, RASAL1, RASAL2, RASAL3, RASD1, RASD2, RASEF, RASGEF1A, RASGEF1B, RASGEF1C, RASGRF1, RASGRF2, RASGRP1, RASGRP2, RASGRP3, RASGRP4, RASIP1, RASL10A, RASL10B, RASL11A, RASL11B, RASL12, RASSF1, RASSF10, RASSF2, RASSF3, RASSF4, RASSF5, RASSF6, RASSF7, RASSF8, RASSF9, RRAS, RRAS2
RET	ADD3, ALOX5, ANK3, ANKS1B, ARHGAP12, CCDC186, CCDC3, CCDC6, CCDC88C, CCNY, CCNYL1, CDC123, CLIP1, CTNNA3, CUX1, DOCK1, DUSP5, DYDC1, EML4, EML6, EPC1, EPHA5, ERC1, FRMD4A, GDNF, GFRA1, GFRA2, GFRA3, GFRA4, GPRC5B, IL2RA, KIAA1217, KIAA1468, KIF13A, KIF5B, LSM14A, MINDY3, MPRIP, MRPS30, MYO5C, NCOA4, NRP1, PARD3, PCM1, PICALM, PRKARIA, PRKCQ, PRKG1, PRPF18, PTER, PTK2, PTPRK, RASSF4, RBPMS, RET, RETN, RETNLB, RETREG1, RETREG2, RETREG3, RETSAT, RUFY2, SIRT1, SORBS1, TBC1D32, TRIM24, TRIM33, TSSK4, UBE2D1, WAC, ZNF43, ZNF438
ROS1	AKT1, CCDC6, CD74, CEP72, CLTC, EZR, GOPC, IRS1, KDELR2, KMT2C, LIMA1, LRIG3, MAPK1, MAPK3, MSN, MYO5C, PLCG2, PROS1, PTPN11, RBPMS, ROS1, SDC4, SLC34A2, SLC6A17, SLMAP, STAT3, TFG, TMEM106B, TPD52L1, TPM3, VAV3, WNK1, ZCCHC8
TP53	TP53, TP53BP1, TP53BP2, TP53H11, TP53H13, TP53I3, TP53INP1, TP53INP2, TP53RK, TP53TG1, TP53TG5
CTLA4	CD274, CD276, CD28, CD80, CD86, CTLA4, FOXP3, GRB2, LCK, NFAM1, NFAT5, NFATC1, NFATC2, NFATC2IP, NFATC3, NFATC4, PTPN11
PDCD1	CD247, CD274, CD3D, CD3E, CD4, CD80, FGL1, HLA.DQB1, HLA.DRB1, LAG3, PDCD1, PDCD1LG2, PRKCQ, PTPN11, ZAP70

the unit square, the cutoff is used to determine a sample expression status: normal or abnormal. Given a member gene  $g$ , let  $T_g$  be the cutoff, then the samples in the training data set are divided into two populations, one above  $T_g$ , the other below. Then for each population, the recurrence percentages are collected, denoted as  $P_{above}$ ,  $P_{below}$ , respectively. Let  $P_\delta = |P_{above} - P_{below}|$ , the absolute difference values between the populations, it represents the prediction power of a gene expression to recurrence.  $P_\delta$  represents the prediction power of  $g$  as a univariate predictor of the recurrence. Moreover, if  $P_{above} > P_{below}$ , then  $g$  is over-expressed for the population of higher recurrence risk, or else if  $P_{above} < P_{below}$ , then  $g$  is down-expressed. The risk difference between these two groups is called *significant* if  $P_\delta \geq T_{diff}$ , where  $T_{diff}$  is a pre-specified threshold and is set as 5% in the following. With respect to  $g$ , a sample is labeled as: (1). *normal* if  $P_\delta < T_{diff}$ ; (2). *up* if  $P_\delta \geq T_{diff}$  and  $P_{above} > P_{below}$ ; (3). *down* if  $P_\delta \geq T_{diff}$  and  $P_{above} < P_{below}$ . Both *up* and *down* are called *abnormal*.

### 2.3.2 Cluster Member Voting to Define GCEI

Now for the considering cluster, calculate the percentage of the *abnormal* gene members for each sample and use the percentage as a new univariate predictor of recurrence, following the same approach as in the above, ROC is plotted and an optimal percentage threshold  $T_p$  is obtained. Now for each sample, if the percentage of the abnormal members is greater than or equal to  $T_p$ , the sample is labeled with 1, or else 0. This characteristic index is called *Gene Cluster Expression Abnormality Index* (GCEI). GCEI value 1 represents *abnormal* expression with respect to the given gene cluster, while value 0 represents *normal*.

### 2.3.3 Lung Cancer Recurrence Risks of GCEI Status

Recurrence risks are assessed with respect to the status of a single GCEI or a combination of multiple GCEIs. For a single cluster GCEI, recurrence risk is calculated for  $GCEI = 0$  and  $GCEI = 1$  respectively. For 9 driver gene cluster combination, given the ordered list (ALK, BRAF, EGFR, MET, NTRK, RAS, RET, ROS1, TP53), concatenate the corresponding GCEI of each cluster to obtain a binary string of 9 bits, for example, 000000000 represents all nine gene clusters are normally expressed, 100000000 represents only ALK cluster is abnormally expressed, 111111111 represents all 9 clusters are abnormally expressed, and so on. 9-bit GCEI classifies lung cancers into  $2^9 = 512$  subtypes. Moreover, since in practice it might be difficult to accumulate enough patient cases for some of the 512 subtypes, we may collapse 512 subtypes into only 10 super-subtypes as follows, by counting number of digit 1 in the 9-bit string, patients are grouped into 10 subtypes with aggregated GCEI of 0, 1, 2, 3,  $\dots$ , 9 respectively, and each GCEI value tells how many gene clusters are abnormal among the nine clusters. For immunotherapy target couple (CTLA4, PDCD1), GCEI is a two-digit string with four combinations: 00, 01, 10, 11, representing none, CTLA4 only, PDCD1 only, or both of the two clusters are abnormally expressed respectively.

## 2.4 Data Analysis Software

Data analysis and plots were scripted in house using RStudio 2022.07.1 with R version 4.0.5 on Mac platform with OS version darwin17.0.

## 3 Results

### 3.1 Univariate Models

Univariate models are constructed first by using the expression of each cluster member as a recurrence predictor and second by aggregating the expression status of cluster members and using the percentage of abnormal members as a new recurrence predictor, called cluster member voting model. At last, recurrence risks are assessed with respect to various patient populations using these models with combination.

#### 3.1.1 ALK Cluster

There are 107 pre-selected members in ALK cluster, most of which are fusion partners [41]. The univariate models showed that 72 *abnormal* genes have  $P_\delta \geq 5\%$ , accounting for 67%, and the rest 35 *normal* genes have  $P_\delta < 5\%$ . The corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks of the *abnormal* and the *normal* genes are listed in Table 2 and Table 3 respectively. As shown in Table 2, 33 genes are over-expressed for higher recurrence risk: CEP55, TUBB, MDK, NPM1, CEBPZ, TFG, ATIC, LYPD1, LCLAT1, LPIN1, MYT1L, WNK3, TNIP2, C12ORF75, TPM4, TTC27, SOS1, ADAM17, TSPYL6, KLC1, PPFIBP1, SPECC1, FRS2, SHC1, FBN1, THADA, SQSTM1, CLIP1, CBL, CLTC, FBXO36, FUT8 and ITGAV; 39 are down-expressed for higher recurrence risk: ATP13A4, LMO7, WDR37, EPAS1, GCC2, CRIM1, PLEKHH2, TRIM66, FBXO11, SMPD1, YAP1, MPRIP, TANC1, SEC31A, PRKAR1A, CYBRD1, SPTBN1, ALKAL2, WDPCP, SLMAP, CLIP4, SLC16A7, SWAP70, LIMD1, BIRC6, SOCS5, PLEKHA7, EIF2AK3, PPM1B, KIF5B, PHACTR1, CAMKMT, RBM20, SRD5A2, NYAP2, PTN, PICALM, VKORC1L1 and HIP1.

For illustrating purpose, the ROCs of top 12 genes in decreasing order of  $P_\delta$  are shown in Figure 1. The highest one in the first row of Table 2 is CEP55. It shows that the normalized

Table 2: AUCs and recurrence risks of 72 *abnormal* ALK genes with  $P_{\delta} \geq 5\%$ .

GENE	AUC	FPR	TPR	$T_g$	$P_{above}(\%)$	$P_{below}(\%)$	$P_{\delta}(\%)$	Status
CEP55	0.675	0.42	0.76	-0.0076	49.03	18.39	30.64	up
ATP13A4	0.66	0.43	0.7	-0.0197	24.19	46.15	21.96	down
TUBB	0.644	0.39	0.62	0.0253	46.02	25	21.02	up
MDK	0.64	0.42	0.64	0.0403	44.77	25.1	19.67	up
LMO7	0.601	0.39	0.58	0.0509	22.99	42.37	19.38	down
NPM1	0.608	0.46	0.68	-0.0053	43.63	24.66	18.97	up
WDR37	0.628	0.43	0.63	0.0118	24.54	43.23	18.69	down
EPAS1	0.665	0.31	0.58	0.0882	23.12	41.42	18.3	down
GCC2	0.624	0.32	0.57	0.0458	24.87	41.52	16.65	down
CRIM1	0.63	0.39	0.62	0.0303	25.25	41.79	16.54	down
PLEKHH2	0.618	0.37	0.57	0.079	24.73	41	16.27	down
CEBPZ	0.591	0.39	0.57	0.0299	43.58	27.65	15.93	up
TFG	0.576	0.35	0.51	0.0718	43.88	28.67	15.21	up
ATIC	0.61	0.41	0.58	0.0254	42.92	27.73	15.19	up
TRIM66	0.588	0.41	0.58	0.005	27.04	42.17	15.13	down
LYPD1	0.598	0.45	0.61	0.011	42.15	27.5	14.65	up
FBXO11	0.62	0.34	0.56	0.041	25.57	40.2	14.63	down
LCLAT1	0.587	0.5	0.66	-0.013	41.2	26.98	14.22	up
SMPD1	0.572	0.39	0.54	0.0281	26.87	40.57	13.7	down
LPIN1	0.565	0.36	0.5	0.053	42.86	29.37	13.49	up
MYT1L	0.569	0.47	0.62	-0.0011	41.27	27.83	13.44	up
YAP1	0.589	0.37	0.55	0.0335	27.52	40.91	13.39	down
MPRIP	0.592	0.46	0.63	0.0038	27.65	40.75	13.1	down
WNK3	0.539	0.35	0.48	0.0757	42.55	29.93	12.62	up
TANC1	0.624	0.32	0.57	0.0631	26.74	39.35	12.61	down
SEC31A	0.601	0.42	0.57	-0.0025	29.17	41.74	12.57	down
PRKARIA	0.592	0.44	0.6	0.0074	27.98	40.53	12.55	down
TNIP2	0.561	0.46	0.6	7e-04	40.98	28.57	12.41	up
C12ORF75	0.581	0.48	0.62	-0.027	40.62	28.32	12.3	up
TPM4	0.558	0.5	0.63	-0.0131	40.46	28.18	12.28	up
TTCT27	0.569	0.41	0.54	0.022	41.55	29.28	12.27	up
CYBRD1	0.588	0.48	0.61	0	28.51	40.55	12.04	down
SPTBN1	0.571	0.42	0.55	0.0219	28.14	39.58	11.44	down
ALKAL2	0.598	0.35	0.52	0.1104	28.04	39.25	11.21	down
SOS1	0.551	0.46	0.58	0.0069	40.42	29.34	11.08	up
ADAM17	0.569	0.46	0.58	-0.004	40.08	29.58	10.5	up
TSPYL6	0.562	0.48	0.6	-0.0022	39.84	29.44	10.4	up
KLC1	0.523	0.32	0.42	0.0302	41.52	31.19	10.33	up
PFFIBP1	0.549	0.46	0.57	-0.0145	39.92	29.92	10	up
SPECC1	0.566	0.47	0.58	-0.0119	39.75	29.83	9.92	up
WDPCP	0.559	0.44	0.55	0.0163	29.49	39.25	9.76	down
SLMAP	0.584	0.39	0.53	0.0339	29.02	38.75	9.73	down
CLIP4	0.575	0.34	0.5	0.0655	28.96	38.46	9.5	down
SLC16A7	0.575	0.44	0.54	0.0041	30.24	39.74	9.5	down
SWAP70	0.563	0.48	0.59	0.0024	29.73	39.23	9.5	down
LIMD1	0.574	0.49	0.58	0.0054	29.82	39.02	9.2	down
FRS2	0.516	0.44	0.54	0.0062	39.47	30.71	8.76	up
BIRC6	0.55	0.38	0.52	0.0159	30.26	38.98	8.72	down
SHC1	0.521	0.34	0.43	0.0616	40.22	31.68	8.54	up
FBN1	0.527	0.46	0.55	-0.0041	39.15	30.77	8.38	up
SOCS5	0.556	0.36	0.47	0.0388	29.73	38.05	8.32	down
PLEKHA7	0.556	0.53	0.64	-0.0759	31.77	39.89	8.12	down
EIF2AK3	0.534	0.45	0.54	0	31.08	38.96	7.88	down
THADA	0.529	0.44	0.52	0.0089	38.94	31.25	7.69	up
SQSTM1	0.51	0.45	0.53	0.0163	38.7	31.35	7.35	up
PPM1B	0.544	0.42	0.54	0.0205	30.88	38.11	7.23	down
KIF5B	0.527	0.44	0.55	0.0113	31.05	38.02	6.97	down
PHACTR1	0.574	0.41	0.52	0.0754	30.69	37.54	6.85	down
CLIP1	0.508	0.4	0.46	0.0297	38.89	32.04	6.85	up
CAMKMT	0.54	0.43	0.53	0.0096	31.25	37.98	6.73	down
RBM20	0.52	0.47	0.56	0.0093	31.31	37.69	6.38	down
CBL	0.536	0.39	0.45	0.019	38.58	32.28	6.3	up
SRD5A2	0.56	0.52	0.64	-0.0121	32.09	38.32	6.23	down
NYAP2	0.542	0.54	0.67	-0.0186	32.26	38.42	6.16	down
CLTC	0.512	0.39	0.45	0.0321	38.38	32.39	5.99	up
FBXO36	0.515	0.51	0.58	-0.0086	37.6	31.7	5.9	up
PTN	0.538	0.44	0.56	0.0341	31.63	37.45	5.82	down
PICALM	0.549	0.39	0.46	0.0318	31.35	37.04	5.69	down
FUT8	0.533	0.47	0.53	0.0417	37.55	32.24	5.31	up
VKORC1L1	0.529	0.47	0.54	0.0027	32.23	37.5	5.27	down
HIP1	0.544	0.44	0.54	0.0348	31.86	37.05	5.19	down
ITGAV	0.506	0.51	0.57	9e-04	37.25	32.16	5.09	up



Table 3: AUCs and recurrence risks of 35 *normal* ALK genes with  $P_\delta < 5\%$ .

GENE	AUC	FPR	TPR	$T_g$	$P_{above}(\%)$	$P_{below}(\%)$	$P_\delta(\%)$	Status
TOGARAM2	0.564	0.51	0.62	-0.0095	32.37	37.34	4.97	normal
BCL11A	0.519	0.4	0.51	0.0391	32.37	37.34	4.97	normal
ATAD2B	0.505	0.36	0.41	0.0621	37.91	33	4.91	normal
MSN	0.55	0.39	0.51	0.0511	31.74	36.51	4.77	normal
PRKCB	0.546	0.38	0.49	0.0701	31.98	36.45	4.47	normal
AKAP8L	0.506	0.49	0.56	-0.0046	32.8	37.07	4.27	normal
CUX1	0.542	0.4	0.49	0.0396	32.28	36.52	4.24	normal
NCOA1	0.516	0.42	0.51	0.0344	32.26	36.49	4.23	normal
PLEKHM2	0.503	0.48	0.52	-0.0047	36.97	32.79	4.18	normal
SORCS1	0.513	0.54	0.59	-0.0075	33.08	36.99	3.91	normal
SMPDL3B	0.527	0.5	0.6	-0.0554	33.33	37.17	3.84	normal
CMTR1	0.512	0.47	0.51	0.0067	32.89	36.58	3.69	normal
MAPK1	0.519	0.49	0.52	-0.005	36.67	33.06	3.61	normal
TCF12	0.536	0.45	0.49	0.0036	36.77	33.2	3.57	normal
SMPDL3A	0.52	0.51	0.54	-0.0275	36.43	32.86	3.57	normal
MTA3	0.503	0.46	0.5	0.0021	36.68	33.2	3.48	normal
SMPD2	0.505	0.38	0.39	0.0535	36.63	33.87	2.76	normal
MAPK3	0.536	0.42	0.48	0.0142	33.33	36	2.67	normal
DCTN1	0.502	0.43	0.46	0.0209	36.32	33.7	2.62	normal
DCHS1	0.531	0.41	0.49	0.0531	36.47	33.97	2.5	normal
SMPD3	0.517	0.46	0.48	0.0141	36.16	33.72	2.44	normal
SRBD1	0.523	0.49	0.53	-3e-04	33.75	35.95	2.2	normal
TPR	0.512	0.46	0.52	0.0063	33.76	35.89	2.13	normal
ALK	0.518	0.54	0.61	-0.0176	35.71	33.66	2.05	normal
TACR1	0.515	0.55	0.6	-0.0096	33.96	35.98	2.02	normal
VIT	0.532	0.42	0.52	0.015	33.67	35.66	1.99	normal
DYSF	0.506	0.51	0.53	-0.0257	35.74	33.91	1.83	normal
IRS1	0.52	0.48	0.52	0.0129	33.91	35.71	1.8	normal
EML4	0.505	0.41	0.45	0.0211	33.93	35.66	1.73	normal
CDK15	0.517	0.45	0.51	0.0047	33.94	35.61	1.67	normal
ERC1	0.51	0.41	0.43	0.0162	35.82	34.16	1.66	normal
EML6	0.51	0.54	0.6	-0.0678	34.43	35.59	1.16	normal
STRN	0.511	0.49	0.48	0.0012	34.32	35.37	1.05	normal
RANBP2	0.538	0.31	0.44	0.051	34.25	35.22	0.97	normal
C9ORF3	0.512	0.43	0.48	0.0195	35.35	34.46	0.89	normal

expression cutoff  $T_g = -0.0076$ . Patience with CEP55 expression  $\geq (-0.0076)$  has a recurrence risk of  $P_{above} = 49.03\%$  while patience with CEP55 expression  $< (-0.0076)$  has a risk of  $P_{below} = 18.39\%$ , hence the difference  $P_\delta = 30.64\%$ . CEP55 is over-expressed (with respect to recurrence) because  $P_{above} > P_{below}$ . CEP55, called Centrosomal Protein 55, is related to DNA damage and cytoskeletal signaling and plays a role in mitotic exit and cytokinesis. CEP55 was found to be a fusion partner of ALK [13] and high CEP55 expression is associated with poor prognosis [25]. The second gene is ATP13A4, which is down-expressed with  $P_{above} = 46.15\%$ ,  $P_{below} = 24.19\%$  and a difference  $P_\delta = 21.96\%$ . ATP13A4, called ATPase 13A4, may enable ATPase-coupled cation transmembrane transporter activity and may be involved in cellular calcium ion homeostasis.

In a lung cancer case study [11], a 53-year-old metastatic Stage IV patient was harbored with ATP13A4-ALK and two other ALK-fusions COX7A2L-ALK and LINC01210-ALK, first-line crizotinib therapy showed 12 months of PFS/PR, then a new SLCO2A1-ALK fusion led to resistance, afterwards a second line ceritinib resulted in further 8 months of PFS and NGS results demonstrated the loss of ATP13A4-ALK and SLCO2A1-ALK.

Interestingly, ALK expression itself is normal and only gives a difference of  $P_\delta = 2.02\%$  with this training data set.

### 3.1.2 BRAF Cluster

BRAF cluster contains 33 members. The ROCs of only 12 genes with top  $P_\delta$  are presented in Figure 2. AUCs, FPRs, TPRs, threshold  $T_g$ , population risks for all members are listed in Table 4. BRAF phosphorylates MAP2K1 and thereby activates the MAP kinase signal pathway and here most of the cluster members are related to MAP. There are 19 members with  $P_\delta \geq 5\%$ , accounting for 58%, within which only 4 genes MAP2K2, MAP4K4, MAP3K7 and MAP2K1 are over-expressed. MAP2K1 (MEK1) and MAP2K2 (MEK2) activates BRAF via controlling KSR1[28]. On the other hand, RAF1 is down-expressed with modest  $P_\delta = 6.02\%$ . BRAF/RAF1

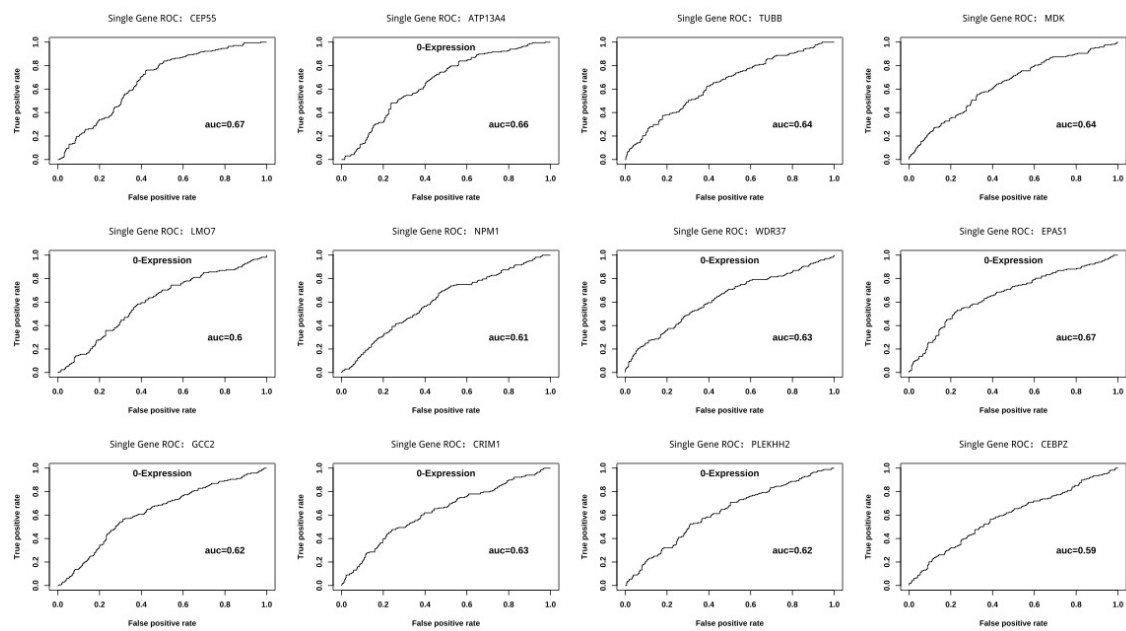


Figure 1: Univariate ROCs of the top 12 genes ALK cluster in the decreasing order of  $P_{\delta}$ . Inverted expression value ( $0 - Expression$ ) was used to plot ROC for down regulated genes, similarly hereinafter.



heterodimers are downstream receptors of RAS and are crucial activator of MAPK[62]. Moreover, MAP2K3/4/5 and MAP3K1/2/3/5/6/7CL/8/9/11/13/14-AS1 are down-expressed with  $P_\delta$  ranging from 5.2% to 13.78%, so is RAF1. Other remaining ones such as MAP2K6/7, MAP3K/4/10/12/14/19/20 and MAP4K1/2/3/5 are normal with  $P_\delta < 5\%$ . BRAF itself is deemed to be normal with  $P_\delta = 4.21\%$ .

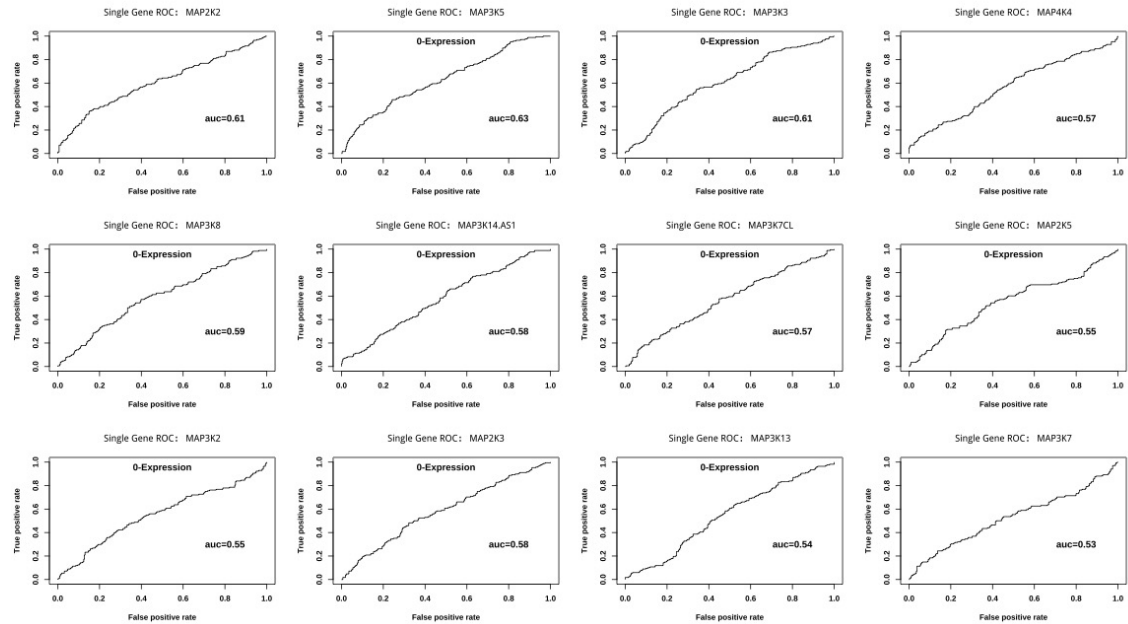


Figure 2: Univariate ROCs of top 12 genes in BRAF cluster in the decreasing order of  $P_\delta$ .

### 3.1.3 EGFR Cluster

EGFR cluster contains 16 members. The ROCs are presented in Figure 3 and the corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks are listed in Table 5. There are 12 members with  $P_\delta \geq 5\%$ , accounting for 75%, 7 of which including NRG4, EREG, SRC, RGS16, TGFA, CTNNB1 and NRG3 are over-expressed. NRG4, EREG, TGFA and NRG3 are known ligands of EGFR, while RGS16 is phosphorylated by EGFR to have GTPase activation, in addition, EGFR increasingly interacts with SRC and CTNNB1 by phosphorylating MUC1. On the other hand, EGFR itself and MUC1 are down-expressed for lung cancer recurrence, so are other two ligands BTC and AREG. Lastly, The remaining four genes BRAF, NRG2, NRG1 and EGF are normal with  $P_\delta < 5\%$ .

### 3.1.4 MET Cluster

MET cluster contains 8 members. The ROCs are presented in Figure 4 and the corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks are listed in Table 6. There are 7 members with  $P_\delta \geq 5\%$ , accounting for 87.5%. Most MET effectors are PI3-kinase subunits. SRC, GRB2 and PLCG1 are over-expressed while PIK3R1, HGF, GAB1 and MET itself are down-expressed. However, STAT3 is normal with  $P_\delta = 0.09\%$ .

Table 4: AUCs and recurrence risks of 33 BRAF genes ordered by  $P_\delta$ .

GENE	AUC	FPR	TPR	$T_g$	$P_{above}(\%)$	$P_{below}(\%)$	$P_\delta(\%)$	Status
MAP2K2	0.607	0.37	0.54	0.0282	44.39	27.8	16.59	up
MAP3K5	0.629	0.36	0.54	0.05	26.56	40.34	13.78	down
MAP3K3	0.612	0.34	0.55	0.0439	26.74	39.35	12.61	down
MAP4K4	0.568	0.51	0.64	-0.0223	40.23	28.24	11.99	up
MAP3K8	0.588	0.4	0.57	0.0754	27.98	39.45	11.47	down
MAP3K14-AS1	0.576	0.51	0.64	-0.0104	30.94	39.63	8.69	down
MAP3K7CL	0.568	0.45	0.58	0.0071	30.22	38.91	8.69	down
MAP2K5	0.552	0.39	0.54	0.0244	29.85	38.43	8.58	down
MAP3K2	0.553	0.42	0.55	0.0094	30.43	38.89	8.46	down
MAP2K3	0.581	0.37	0.52	0.0306	29.73	38.05	8.32	down
MAP3K13	0.541	0.48	0.58	-0.0201	31.5	39.23	7.73	down
MAP3K7	0.525	0.45	0.53	0.0179	38.7	31.35	7.35	up
MAP3K1	0.526	0.5	0.6	-0.0183	31.7	38.71	7.01	down
MAP3K6	0.543	0.46	0.57	0.0087	31.19	37.88	6.69	down
MAP2K1	0.514	0.38	0.45	0.0407	38.66	32.29	6.37	up
RAF1	0.536	0.39	0.46	0.0363	30.82	36.84	6.02	down
MAP2K4	0.523	0.44	0.53	0.0085	31.78	37.8	6.02	down
MAP3K11	0.539	0.5	0.6	-0.0042	32.07	37.55	5.48	down
MAP3K9	0.542	0.5	0.6	-0.0449	32.75	37.95	5.2	down
MAP4K2	0.516	0.51	0.57	-0.0159	37.11	32.3	4.81	normal
MAP3K21	0.525	0.47	0.53	0.0201	37.29	32.52	4.77	normal
MAP4K3	0.511	0.49	0.55	-0.0021	32.5	37.19	4.69	normal
MAP3K20	0.501	0.42	0.45	0.0607	32	36.48	4.48	normal
MAP4K5	0.54	0.43	0.48	0.0222	32.37	36.73	4.36	normal
MAP2K6	0.506	0.36	0.4	0.0443	37.57	33.22	4.35	normal
BRAF	0.508	0.4	0.45	0.0458	37.31	33.1	4.21	normal
MAP3K4	0.512	0.43	0.47	0.0202	37.09	33.09	4	normal
MAP2K7	0.524	0.32	0.43	0.0242	32.83	36.27	3.44	normal
MAP4K1	0.514	0.5	0.55	-0.0175	33.73	36.12	2.39	normal
MAP3K19	0.521	0.47	0.51	0.0101	33.64	35.88	2.24	normal
MAP3K10	0.516	0.41	0.47	0.0101	33.94	35.63	1.69	normal
MAP3K14	0.519	0.37	0.48	0.059	35.75	34.32	1.43	normal
MAP3K12	0.517	0.45	0.5	0.0124	35.5	34.4	1.1	normal

Table 5: AUCs and recurrence risks of EGFR genes ordered by  $P_\delta$ .

GENE	AUC	FPR	TPR	$T_g$	$P_{above}(\%)$	$P_{below}(\%)$	$P_\delta(\%)$	Status
NRG4	0.559	0.43	0.58	0.0034	41.74	28.57	13.17	up
EREG	0.549	0.59	0.73	-0.0948	39.61	26.44	13.17	up
SRC	0.54	0.37	0.51	0.0188	42.29	29.54	12.75	up
MUC1	0.567	0.43	0.56	0.0151	29.63	40.17	10.54	down
EGFR	0.57	0.3	0.46	0.0658	29.67	38.83	9.16	down
RGS16	0.549	0.43	0.52	0.0228	39.64	30.77	8.87	up
TGFA	0.513	0.39	0.49	0.0691	39.81	31.16	8.65	up
BTC	0.563	0.42	0.54	0.0483	30.19	38.52	8.33	down
AREG	0.549	0.45	0.56	0.0266	30.87	38.49	7.62	down
NRGN	0.562	0.37	0.45	0.0989	30.81	37.1	6.29	down
CTNNB1	0.52	0.5	0.57	-0.0256	37.55	31.88	5.67	up
NRG3	0.523	0.57	0.62	-0.0287	37.1	31.66	5.44	up
BRAF	0.508	0.4	0.45	0.0458	37.31	33.1	4.21	normal
NRG2	0.532	0.44	0.5	0.0083	32.5	36.52	4.02	normal
NRG1	0.528	0.4	0.45	0.0314	33.16	35.99	2.83	normal
EGF	0.51	0.41	0.4	0.2169	34.87	34.84	0.03	normal

Table 6: AUCs and recurrence risks of MET genes ordered by  $P_\delta$ .

GENE	AUC	FPR	TPR	$T_g$	$P_{above}(\%)$	$P_{below}(\%)$	$P_\delta(\%)$	Status
PIK3R1	0.621	0.4	0.63	0.0369	25.76	41.2	15.44	down
HGF	0.579	0.45	0.62	0.0079	27.23	40.89	13.66	down
SRC	0.54	0.37	0.51	0.0188	42.29	29.54	12.75	up
GAB1	0.612	0.36	0.57	0.0203	27.94	39.93	11.99	down
GRB2	0.559	0.47	0.6	-0.0076	40.49	28.94	11.55	up
PLCG1	0.524	0.44	0.54	0.0036	39.3	30.83	8.47	up
MET	0.544	0.4	0.48	0.0288	32.23	37.5	5.27	down
STAT3	0.53	0.37	0.45	0.0507	34.91	34.82	0.09	normal

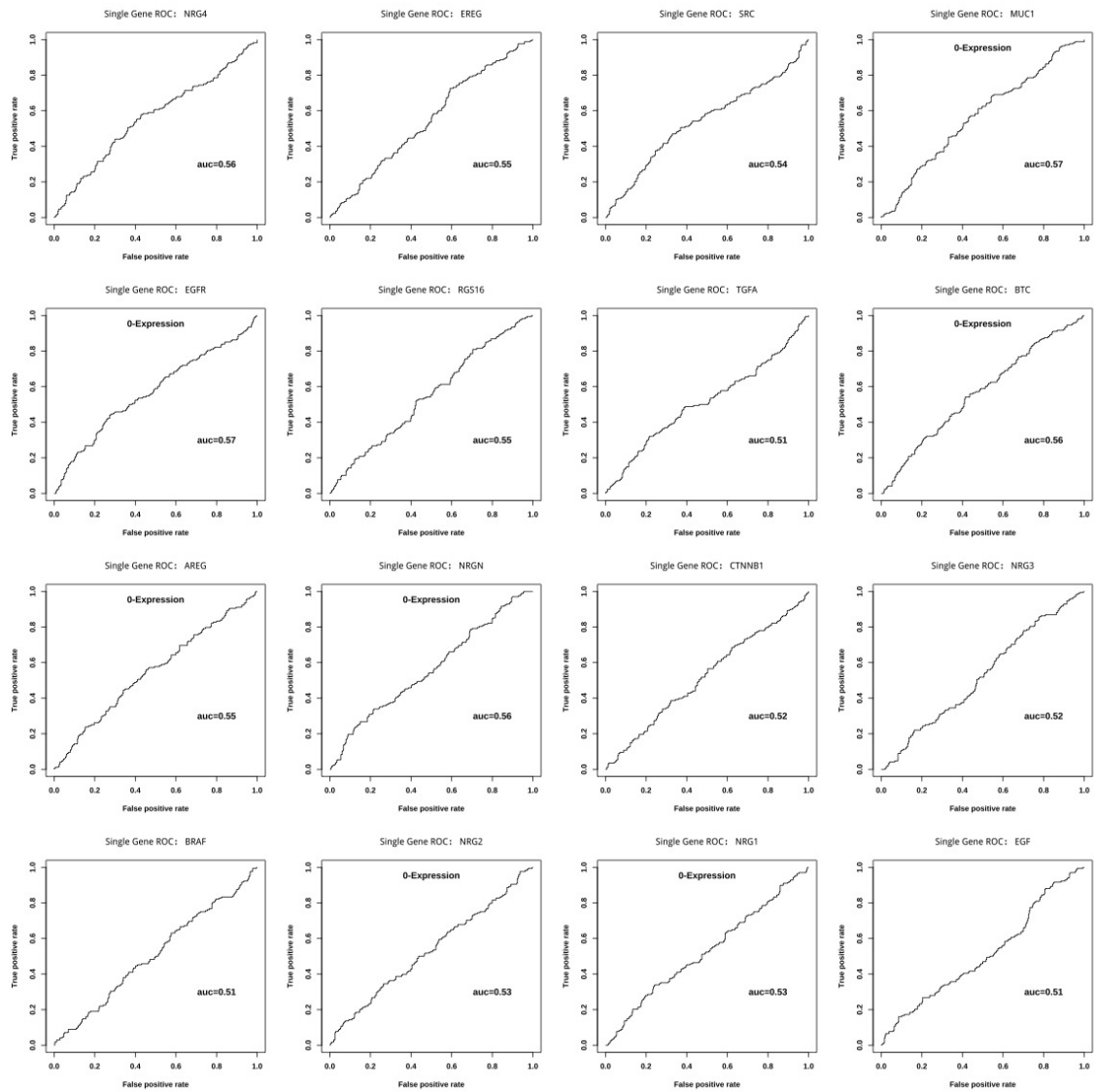


Figure 3: Univariate ROCs of 16 genes in EGFR cluster.

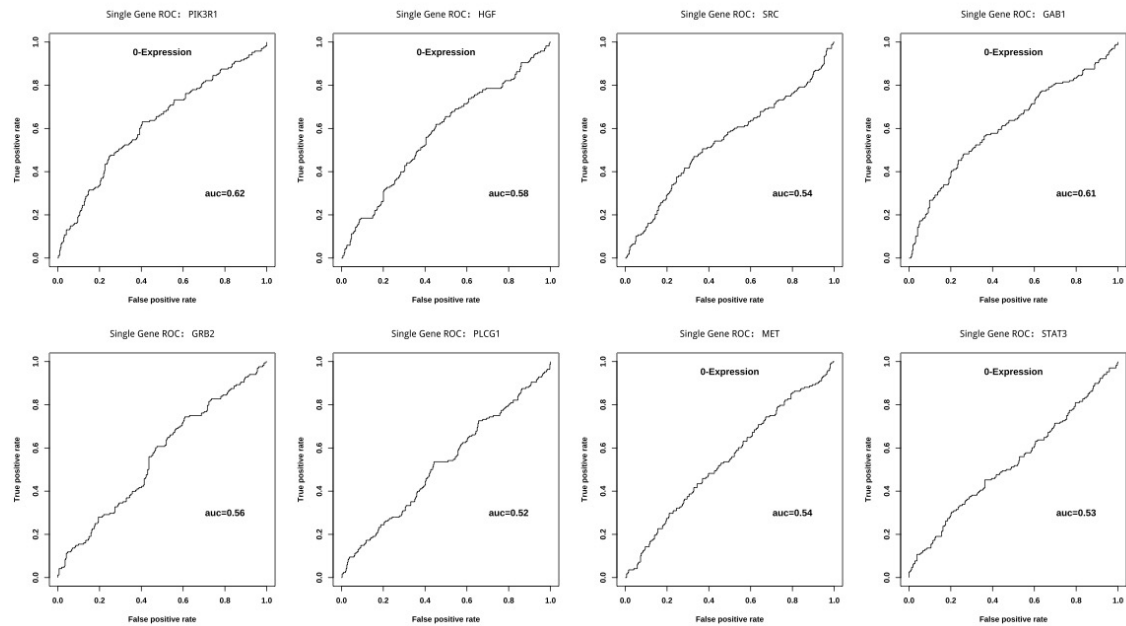


Figure 4: Univariate ROCs of 8 genes in MET cluster.

### 3.1.5 NTRK Cluster

NTRK cluster contains 58 members, most of the which are NTRK fusion partners listed in Cocco E et al [12] and are re-organized in Table 7. The ROCs are presented in Figure 5 and the corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks are listed in Table 8. There are 37 members with  $P_\delta \geq 5\%$ , accounting for 64%, within which 24 are over-expressed: ETV6, TPM3, SLITRK1, TFG, SLITRK4, CHTOP, SLITRK5, TPM4, TP53, TRAF2, AGLB3, LYN, RFWD2, NTRK1, AFAP1, AGLB5, UBE2R2, SQSTM1, SLITRK2, MRPL24, NTRK2, GRI-PAP1, SLITRK6 and TRIM24 and 13 are down-expressed: MPRIP, TLE4, RBPMS, NFASC, NTRK3, ARHGEF2, CD74, RABGAP1L, NACC2, TRIM63, IGFBP7, DAB2IP, AGLB1. The remaining 24 genes: AGLB2, PPL, BCR, SCYL3, LMNA, MYO5A, CTCR, PLEKHA6, BCAN, PDE4DIP, HNRNPA2B1, VCL, TPR, PAN3, QKI, SLITRK3, EML4, BTBD1, STRN, LRRC71 and IRF2BP2 don't show much differences with  $P_\delta < 5\%$ .

Table 7: NTRK fusion partners as listed in Cocco E et al [12]

Domain	NTRK1	NTRK2	NTRK3
Coiled-coil	MPRIP, TFG, SQSTM1, TPM3, TRIM63, ARHGEF2, TPR, LMNA, PPL	TRIM24, PAN3, SQSTM1	TPM4, TFG, MYO5A
Zinc Finger	IRF2BP2	TRAF2	
WD domain	RFWD2	STRN	EML4
Alternative	CD74, NFASC, BCAN, TP53, CTCR	QKI, ETV6, NACC2, BCR, TLE4	ETV6, BTBD1
Unknown	RABGAP1L, GRIPAP1, PLEKHA6, CHTOP, LRRC71, PDE4DIP, AFAP1, SSBP2, MIR548F1, IGFBP7, MRPL24, SCYL3	DAB2IP, VCL, AGLB4, AFAP1	LYN, RBPMS, UBE2R2, HNRNPA2B1

As shown in Table 7, in the first row ETV6 is over-expressed with  $P_{above} = 45.45\%$ ,  $P_{below} = 27.46\%$  and a difference of  $P_\delta = 17.99\%$ . ETV6 is an ETS family transcription factor repres-

ing transcription. ETV6-NTRK3 fusion has been found in different types of cancers and its expression activates the MAPK and PI3K pathways [12]. The second is actin-binding TPM3, tropomyosin 3, with  $P_\delta = 16.93\%$ . TPM3-NTRK1 fusion has been reported broadly in many different tumor types, but it is very rare in lung cancer. TPM3-NTRK1 fusion was confirmed in a Chinese lung cancer study [74]. Choi et al. [14] reported a NSCLC case of acquired TPM3-NTRK1 fusion resistant to larotrectinib with EML4-ALK fusion progressed on lorlatinib. For NTRK family itself, NTRK1/3 having  $P_\delta$  around 9% and NTRK2 having  $P_\delta = 5.78\%$ , however NTRK1/2 are over-expressed while NTRK3 is down-expressed. Additionally, all 6 SLIT and NTRK like family members [6] were selected into the NTRK cluster and they are over-expressed. SLITTRK1/4/5 have  $P_\delta$  greater than 12%, SLITTRK2/6 have modest  $P_\delta$  around 7% while SLITTRK3 has neglective  $P_\delta = 1.78\%$ . SLITRK5 mediates BDNF-dependent NTRK2 (TrkB) trafficking and signaling [56]. SLITRK3 activates NTRK3 in squamous cell lung cancer [8]. On the other hand, MPRIP stands at the top with  $P_\delta = 13.1\%$ . MPRIP-NTRK1 and CD74-NTRK1 fusions were identified by Vaishnavi A et al. [61] while CD74 has modest  $P_\delta = 7.34\%$ . Both fusions lead to TRKA kinase activity. MPRIP targets myosin phosphatase to the actin cytoskeleton and enables cadherin binding. MPRIP can be also fusion partner of other drive gene in lung cancer. A lung cancer case was reported to be sensitive to ALK inhibitor with MPRIP-ALK fusion[19]. Another late stage case was shown to be sensitive to crizotinib with MPRIP-ROS1 fusion [54]. The second on the down-regulation side is transcription corepressor TLE4, which inhibits the transcriptional activation mediated by PAX5, by CTNNB1 and by TCF family members in Wnt signaling.

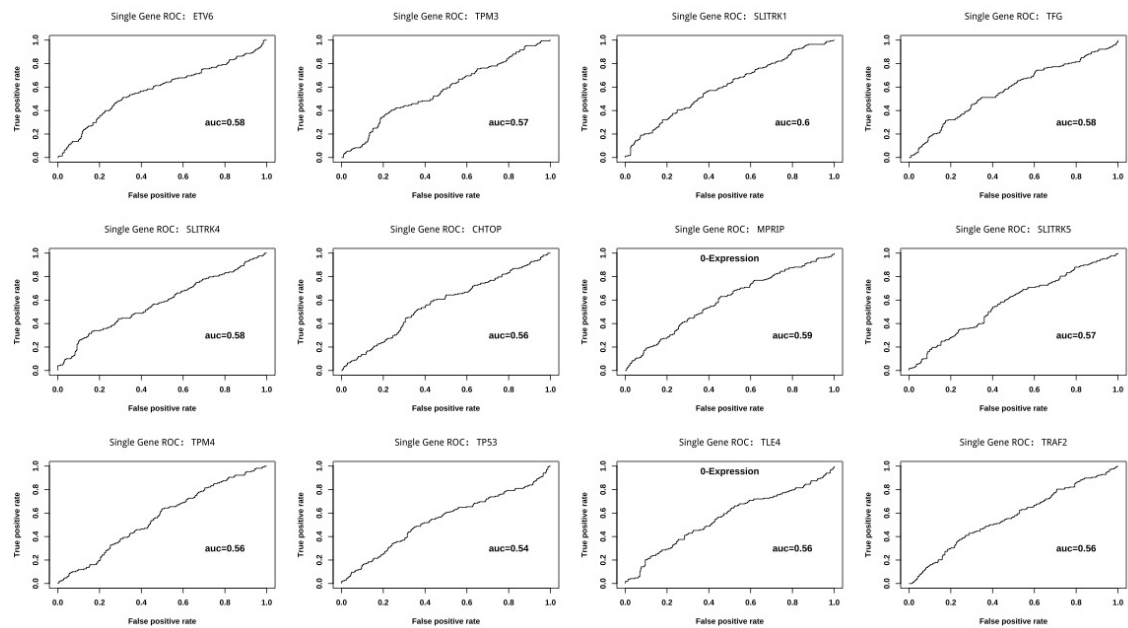


Figure 5: Univariate ROCs of top 12 genes in NTRK cluster.

Table 8: AUCs and recurrence risks of NTRK genes ordered by  $P_\delta$ .

GENE	AUC	FPR	TPR	$T_g$	$P_{above}$ (%)	$P_{below}$ (%)	$P_\delta$ (%)	Status
ETV6	0.579	0.35	0.54	0.048	45.45	27.46	17.99	up
TPM3	0.573	0.26	0.42	0.0359	46.41	29.48	16.93	up
SLITRK1	0.604	0.39	0.57	0.0076	43.58	27.65	15.93	up
TFG	0.576	0.35	0.51	0.0718	43.88	28.67	15.21	up
SLITRK4	0.577	0.29	0.44	0.04	44.85	29.65	15.2	up
CHTOP	0.565	0.43	0.59	0.0186	42.24	28	14.24	up
MPRIP	0.592	0.46	0.63	0.0038	27.65	40.75	13.1	down
SLITRK5	0.574	0.44	0.58	-0.001	41.45	28.63	12.82	up
TPM4	0.558	0.5	0.63	-0.0131	40.46	28.18	12.28	up
TP53	0.541	0.39	0.51	0.0377	41.55	29.82	11.73	up
TLE4	0.56	0.44	0.57	0.0176	28.3	40	11.7	down
TRAF2	0.565	0.37	0.49	0.0486	41.62	30.18	11.44	up
RBPMS	0.596	0.54	0.73	-0.057	30.07	41.26	11.19	down
NFASC	0.564	0.44	0.55	0.0215	28.5	39.64	11.14	down
AGBL3	0.564	0.4	0.51	0.0088	40.76	30.26	10.5	up
LYN	0.519	0.34	0.45	0.1237	41.21	31	10.21	up
RFWD2	0.555	0.39	0.49	0.0239	40.49	30.69	9.8	up
NTRK1	0.557	0.42	0.53	0.0078	40	30.53	9.47	up
NTRK3	0.57	0.44	0.53	0.0159	29.38	38.54	9.16	down
AFAP1	0.546	0.38	0.48	0.0339	40.2	31.1	9.1	up
AGBL5	0.558	0.5	0.6	-0.0037	39.66	30.09	8.97	up
UBE2R2	0.564	0.37	0.46	0.036	40.21	31.25	8.96	up
ARHGEP2	0.583	0.42	0.58	0.0254	29.8	38.38	8.58	down
SQSTM1	0.51	0.45	0.53	0.0163	38.7	31.35	7.35	up
CD74	0.547	0.45	0.55	0.0118	31	38.34	7.34	down
RABGAP1L	0.535	0.4	0.49	0.054	30.53	37.67	7.14	down
SLITRK2	0.541	0.45	0.54	0	38.6	31.5	7.1	up
NACC2	0.546	0.5	0.6	-0.0017	31.17	38.25	7.08	down
MRPL24	0.528	0.51	0.58	-4e-04	38.13	31.11	7.02	up
TRIM63	0.562	0.5	0.58	-0.0018	31.33	38.15	6.82	down
IGFBP7	0.572	0.34	0.48	0.0626	30.59	37.18	6.59	down
NTRK2	0.521	0.54	0.61	-0.0191	37.36	31.58	5.78	up
DAB2IP	0.531	0.51	0.58	-0.0012	31.86	37.5	5.64	down
GRIPAP1	0.507	0.43	0.49	0.0147	37.96	32.33	5.63	up
SLITRK6	0.529	0.43	0.49	0.0705	37.9	32.32	5.58	up
TRIM24	0.534	0.43	0.48	0.0624	37.85	32.46	5.39	up
AGBL1	0.519	0.5	0.57	0.001	32	37.35	5.35	down
AGBL2	0.539	0.5	0.59	-0.0119	32.37	37.34	4.97	normal
PPL	0.595	0.28	0.49	0.1933	31.47	36.28	4.81	normal
BCR	0.535	0.47	0.58	-0.0139	32.84	37.38	4.54	normal
SCYL3	0.548	0.46	0.54	-0.0153	33.1	37.31	4.21	normal
LMNA	0.501	0.44	0.48	0.026	37.16	32.95	4.21	normal
MYO5A	0.507	0.5	0.55	-0.0089	36.8	32.76	4.04	normal
CTRC	0.535	0.53	0.61	-0.0113	33.2	36.77	3.57	normal
PLEKHA6	0.522	0.45	0.5	-0.002	33.2	36.77	3.57	normal
BCAN	0.517	0.46	0.49	0.0054	36.56	33.33	3.23	normal
PDE4DIP	0.533	0.47	0.52	-0.0039	33.59	36.28	2.69	normal
HNRNPA2B1	0.502	0.41	0.45	0.0346	36.46	33.89	2.57	normal
VCL	0.525	0.35	0.41	0.0667	33.33	35.65	2.32	normal
TPR	0.512	0.46	0.52	0.0063	33.76	35.89	2.13	normal
PAN3	0.526	0.42	0.49	0.0316	33.67	35.66	1.99	normal
QKI	0.524	0.43	0.54	0.0215	33.82	35.61	1.79	normal
SLITRK3	0.509	0.47	0.49	0.0015	35.78	34	1.78	normal
EML4	0.505	0.41	0.45	0.0211	33.93	35.66	1.73	normal
BTBD1	0.519	0.37	0.43	0.042	35.54	34.49	1.05	normal
STRN	0.511	0.49	0.48	0.0012	34.32	35.37	1.05	normal
LRRC71	0.524	0.4	0.47	0.0297	34.72	34.95	0.23	normal
IRF2BP2	0.523	0.4	0.45	0.0233	34.74	34.94	0.2	normal



### 3.1.6 RAS Cluster

RAS cluster contains 35 members. The ROCs are presented in Figure 6 and the corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks are listed in Table 9. There are 25 members with  $P_\delta \geq 5\%$  accounting for 71%, with 13 over-expressed and 12 down-expressed, and the remaining 12 are normal with  $P_\delta < 5\%$ . They functionally belong to the following categories:

- *Ras/Rab GTPases*  
over-expressed HRAS, NRAS, KRAS, RRAS2, RASD2; down-expressed RASD1, RRAS; normal MRAS.
- *RAS like family or HRAS like suppressors*  
over-expressed RASL11A/11B, HRAS like suppressor HRASLS; down-expressed RASL12; and normal RASL10A/10B, HRASLS2/5.
- *Ras-association domain family (Rassf)*  
over-expressed RASSF6; down-expressed RASSF2/3/7/10; and normal RASSF1/4/5/8/9.
- *RasGAP (GTPase activating protein)*  
over-expressed RASAL1/2; down-expressed RASA1; normal RASAL3, RASA2/3. RASAL1 belongs to GAP1 and suppresses RAS function. RASAL2 encodes a characteristic domain of GAP and inhibits Ras-cyclic AMP pathway. RASAL3 encodes a protein with pleckstrin homology (PH), C2, and RasGAP domains and is important for liver natural killer T (NKT) cell expansion and functions by suppressing RAS activity and the down-stream ERK signaling pathway.
- *RasGEF (guanine nucleotide exchange factor)*  
over-expressed RASGEF1B/1C and normal RASGEF1A. RASGEF1A is specific for RAP2A, KRAS, HRAS, and NRAS in vivo. RASGEF1B is only specific for RAP2A.  
Moreover, it also includes guanyl-releasing factors (GRF), RASGRF2, down-expressed and normal RASGRF1; and guanyl-releasing proteins (GRP), down-expressed RASGRP1/2/3, and normal RASGRP4.

The top tier of  $P_\delta$  in between 17% and 25% contains 4 genes, in which NRAS, HRAS and RASAL1 are up and RASL12 is down. KRAS is also up but with modest  $P_\delta = 5.29\%$ . This is in consistent with the finding that lung cancer patients with lower RAS expression and treated with bevacizumab plus chemotherapy had a longer PFS and OS than with high RAS expression[5]. Interestingly, NRAS and HRAS suppress KRAS-driven lung cancer growth[60].

### 3.1.7 RET Cluster

RET cluster contains 72 members, most of which are fusion partners [42]. The ROCs are presented in Figure 7 and the corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks are listed in Table 10. There are 47 members with  $P_\delta \geq 5\%$ , accounting for 65%, within which 18 over-expressed and 29 down-expressed. The rest 25 members are normal. The over-expressed ones include MRPS30, CDC123, LSM14A, IL2RA, GPRC5B, KIAA1217, UBE2D1, PRPF18, PARD3, RETNLB, CLIP1, GFRA3, RET, KIAA1468, TRIM33, GDNF, TRIM24, RETREG1; The down ones include ANK3, GFRA1, EPC1, CCDC186, MPRIP, NCOA4, RETN, SORBS1, MINDY3, PRKAR1A, DOCK1, RBPMS, KIF13A, SIRT1, ARHGAP12, MYO5C, ZNF438, WAC, RETSAT, KIF5B, CCDC88C, TSSK4, CCDC3, PCM1, TBC1D32, PRKCQ,

Table 9: AUCs and recurrence risks of RAS genes ordered by  $P_{\delta}$ .

GENE	AUC	FPR	TPR	$T_q$	$P_{above}$ (%)	$P_{below}$ (%)	$P_{\delta}$ (%)	Status
NRAS	0.625	0.38	0.64	0.0328	47.14	23.92	23.22	up
RASL12	0.645	0.38	0.62	0.0353	23.22	43.91	20.69	down
RASAL1	0.614	0.4	0.61	0.0241	44.74	25.98	18.76	up
HRAS	0.608	0.47	0.66	-0.0101	42.69	25.68	17.01	up
RASGRP2	0.615	0.29	0.51	0.051	26.14	38.91	12.77	down
RASSF6	0.581	0.37	0.51	0.1273	42.29	29.54	12.75	up
RASD2	0.559	0.35	0.48	0.038	42.55	29.93	12.62	up
RASGEF1C	0.559	0.44	0.57	0.003	40.6	29.44	11.16	up
RASSF10	0.571	0.48	0.63	-0.0877	30.04	41.15	11.11	down
HRASLS	0.543	0.43	0.54	0.1177	40.62	29.84	10.78	up
RASSF2	0.572	0.34	0.49	0.1436	28.05	38.36	10.31	down
RASL11A	0.547	0.46	0.58	-0.0054	39.67	30	9.67	up
RASAL2	0.55	0.43	0.52	0.0032	39.55	30.92	8.63	up
RASGEF1B	0.544	0.47	0.57	-0.0018	39.09	30.54	8.55	up
RASSF7	0.567	0.32	0.46	0.054	29.9	38.19	8.29	down
RRAS	0.573	0.43	0.55	0.0342	30.58	38.04	7.46	down
RASL11B	0.553	0.4	0.48	0.0847	39.02	31.77	7.25	up
RASD1	0.567	0.33	0.48	0.192	30.56	37.42	6.86	down
RASA1	0.555	0.37	0.49	0.027	31.16	37.83	6.67	down
RRAS2	0.506	0.41	0.48	0.055	38.57	31.99	6.58	up
RASGRP1	0.532	0.21	0.34	0.2924	30.15	36.71	6.56	down
RASGRF2	0.511	0.49	0.57	-4e-04	31.78	37.8	6.02	down
RASGRP3	0.547	0.36	0.47	0.0947	31.21	36.89	5.68	down
RASSF3	0.523	0.45	0.55	0.0189	31.71	37.18	5.47	down
KRAS	0.513	0.44	0.49	0.0191	37.73	32.44	5.29	up
RASGEF1A	0.527	0.41	0.46	0.1617	37.68	32.73	4.95	normal
RASL10B	0.506	0.35	0.4	0.0348	37.64	33.22	4.42	normal
RASSF4	0.507	0.46	0.51	0.0294	37.12	32.81	4.31	normal
RASSF1	0.533	0.53	0.63	-0.0083	32.77	36.84	4.07	normal
RASA2	0.517	0.42	0.48	0.0236	32.73	36.64	3.91	normal
RASGRF1	0.541	0.47	0.55	0.0079	32.71	36.57	3.86	normal
RASSF8	0.557	0.43	0.52	0.0287	32.64	36.33	3.69	normal
RASAL3	0.547	0.54	0.61	-0.027	33.33	36.65	3.32	normal
MRAS	0.526	0.45	0.48	0.0214	36.71	33.45	3.26	normal
RASSF9	0.536	0.46	0.56	0.031	33.18	36.16	2.98	normal
FRAS1	0.52	0.35	0.45	0.1116	36.59	33.96	2.63	normal
RASIP1	0.531	0.53	0.58	-0.004	33.62	36.03	2.41	normal
RASEF	0.507	0.41	0.44	0.0531	33.5	35.79	2.29	normal
HRASLS2	0.526	0.43	0.48	0.0123	33.92	35.69	1.77	normal
HRASLS5	0.534	0.38	0.49	0.0128	34.16	35.36	1.2	normal
RASSF5	0.568	0.32	0.48	0.0734	34.1	35.28	1.18	normal
RASGRP4	0.521	0.55	0.61	-0.0132	34.62	35.14	0.52	normal
GRASP	0.503	0.47	0.46	-4e-04	34.67	35.02	0.35	normal
RASL10A	0.516	0.54	0.62	-0.0099	34.7	35.05	0.35	normal
RASA3	0.537	0.43	0.53	0.0138	34.69	34.97	0.28	normal

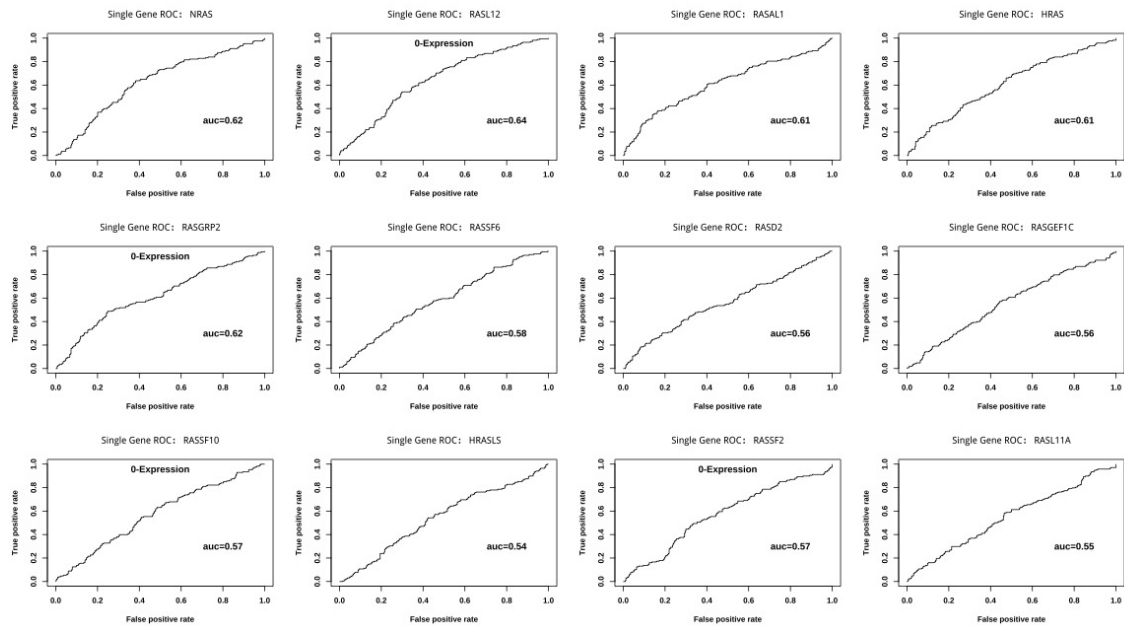


Figure 6: Univariate ROCs of top 12 genes in RAS cluster.

NRP1, PRKG1, PICALM; The normal ones are: CTNNA3, GFRA2, PTPRK, PTK2, RASSF4, DYDC1, CUX1, RUFY2, EPHA5, ADD3, ANKS1B, CCNY, DUSP5, FRMD4A, PTER, ZNF43, GFRA4, RETREG3, EML4, ERC1, CCNYL1, EML6, RETREG2, CCDC6, ALOX5. Among the top 21 genes with  $P_{\delta} \geq 8\%$ , there are only 5 over-expressed ones and the rest 16 are down ones.

### 3.1.8 ROS1 Cluster

ROS1 cluster contains 33 members, most of which are fusion partners [43]. The ROCs are presented in Figure 8 and the corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks are listed in Table 11. There are 21 members with  $P_{\delta} \geq 5\%$ , accounting for 64%, within which PTPN11, TPM3, TFG, KDELR2, CEP72, TPD52L1, VAV3, CLTC, WNK1 are over-expressed, and SLC34A2, SDC4, RBPMS, LRIG3, SLMAP, KMT2C, PLCG2, MYO5C, PROS1, CD74, EZR, ROS1 are down-expressed, and the rest 12 genes MSN, MAPK1, TMEM106B, SLC6A17, MAPK3, LIMA1, ZCCHC8, IRS1, GOPC, CCDC6, AKT1, STAT3 are normal with  $P_{\delta} < 5\%$ . At the top, PTPN11, named as protein tyrosine phosphatase non-receptor type 11, more commonly alias as SHP2, has the highest  $P_{\delta} = 20.71\%$  and is over-expressed for higher risk recurrence. ROS1 mediates the phosphorylation of PTPN11 to activate the downstream pathway. The second top over-expressed is actin-binding TPM3 with  $P_{\delta} = 16.93\%$ , which also appears in NTRK cluster. A case reported that EML4-ALK and TPM3-ROS1 fusion coexistence in an advanced NSCLC Chinese man[76]. The third over-expressed is TFG with  $P_{\delta} = 15.21\%$ , called trafficking from ER to Golgi regulator, also called TRK-fused gene protein, is required for secretory cargo traffic from the ER to the Golgi apparatus, TFG-ROS1 fusion was reported in lung cancers[1] and other cancers[4]. On the down-expression side, SLC34A2 is at the top with  $P_{\delta} = 11.76\%$ . SLC34A2-ROS1 fusion was reported in lung cancer tissues[16]. SDC4, RBPMS and LRIG3 are

Table 10: AUCs and recurrence risks of RET genes ordered by  $P_{\delta}$ .

GENE	AUC	FPR	TPR	$T_g$	$P_{above}(\%)$	$P_{below}(\%)$	$P_{\delta}(\%)$	Status
ANK3	0.613	0.46	0.64	0	26.09	42.86	16.77	down
MRPS30	0.613	0.42	0.6	0.0277	43.53	26.8	16.73	up
GFRA1	0.61	0.47	0.65	-0.0044	27.47	41.77	14.3	down
EPC1	0.581	0.43	0.6	0.0088	27.19	41.13	13.94	down
CCDC186	0.606	0.4	0.57	0.0312	27.57	40.67	13.1	down
MPRIP	0.592	0.46	0.63	0.0038	27.65	40.75	13.1	down
NCOA4	0.615	0.49	0.68	-0.0107	28.79	41.78	12.99	down
RETN	0.577	0.46	0.6	0.0126	27.7	40.52	12.82	down
SORBS1	0.577	0.45	0.55	0.0444	26.82	39.6	12.78	down
MINDY3	0.591	0.41	0.58	0.0257	27.32	39.93	12.61	down
PRKARIA	0.592	0.44	0.6	0.0074	27.98	40.53	12.55	down
CDC123	0.565	0.42	0.55	0.0189	41.44	29.23	12.21	up
LSM14A	0.55	0.35	0.48	0.0324	42.19	30	12.19	up
DOCK1	0.56	0.44	0.58	-5e-04	29.03	41.03	12	down
IL2RA	0.573	0.46	0.58	0.0204	40.66	29.05	11.61	up
RBPMS	0.596	0.54	0.73	-0.057	30.07	41.26	11.19	down
KIF13A	0.59	0.37	0.54	0.0433	28.73	38.54	9.81	down
SIRT1	0.542	0.45	0.6	0.0094	30	38.93	8.93	down
ARHGAP12	0.566	0.43	0.54	0.0048	30.61	39.24	8.63	down
GPRC5B	0.561	0.48	0.58	-0.0146	38.96	30.47	8.49	up
MYO5C	0.573	0.36	0.5	0.1109	29.83	37.87	8.04	down
KIAA1217	0.521	0.44	0.53	0.0304	39.04	31.1	7.94	up
ZNF438	0.551	0.49	0.58	0.0014	30.7	38.58	7.88	down
UBE2D1	0.533	0.39	0.46	0.0565	39.39	31.69	7.7	up
PRPF18	0.53	0.4	0.48	0.0215	39.22	31.65	7.57	up
PARD3	0.529	0.36	0.43	0.0724	39.34	32.11	7.23	up
WAC	0.561	0.39	0.53	0.0249	30.48	37.63	7.15	down
RETSAT	0.552	0.39	0.51	0.0544	30.43	37.58	7.15	down
KIF5B	0.527	0.44	0.55	0.0113	31.05	38.02	6.97	down
RETNLB	0.514	0.6	0.67	-0.0171	37.46	30.6	6.86	up
CLIP1	0.508	0.4	0.46	0.0297	38.89	32.04	6.85	up
CCDC88C	0.552	0.35	0.49	0.0518	30.93	37.5	6.57	down
TSSK4	0.553	0.31	0.46	0.0283	30.52	36.89	6.37	down
GFRA3	0.52	0.54	0.61	-0.0485	37.64	31.28	6.36	up
CCDC3	0.541	0.52	0.6	-0.0149	31.65	37.96	6.31	down
PCM1	0.537	0.42	0.54	0.0419	31.02	37.29	6.27	down
TBC1D32	0.516	0.43	0.51	0.0089	31.82	37.92	6.1	down
PRKCC	0.546	0.42	0.52	0.0642	31.28	37.28	6	down
NRP1	0.556	0.45	0.54	0.0318	31.55	37.32	5.77	down
PRKG1	0.563	0.35	0.49	0.0436	31.35	37.04	5.69	down
PICALM	0.549	0.39	0.46	0.0318	31.35	37.04	5.69	down
RET	0.534	0.51	0.57	-0.0032	37.5	31.86	5.64	up
KIAA1468	0.511	0.47	0.54	0.0065	37.66	32.1	5.56	up
TRIM33	0.538	0.35	0.41	0.0455	38.33	32.78	5.55	up
GDNF	0.518	0.44	0.5	0.0092	37.84	32.31	5.53	up
TRIM24	0.534	0.43	0.48	0.0624	37.85	32.46	5.39	up
RETREG1	0.531	0.56	0.61	-0.156	37.05	31.86	5.19	up
CTNNA3	0.524	0.43	0.48	0.0052	37.56	32.71	4.85	normal
GFRA2	0.528	0.43	0.5	0.0148	31.94	36.77	4.83	normal
PTPRK	0.559	0.33	0.44	0.0769	31.82	36.6	4.78	normal
PTK2	0.538	0.38	0.48	0.0328	32.23	36.9	4.67	normal
RASSF4	0.507	0.46	0.51	0.0294	37.12	32.81	4.31	normal
DYDC1	0.507	0.59	0.63	-0.0217	36.59	32.31	4.28	normal
CUX1	0.542	0.4	0.49	0.0396	32.28	36.52	4.24	normal
RUFY2	0.504	0.44	0.49	0.0063	37.1	32.95	4.15	normal
EPHA5	0.502	0.41	0.45	0.0111	37.17	33.33	3.84	normal
ADD3	0.535	0.4	0.48	0.0666	32.97	36	3.03	normal
ANKS1B	0.503	0.48	0.52	-0.0024	36.29	33.47	2.82	normal
CCNY	0.506	0.41	0.48	0.0317	36.52	33.88	2.64	normal
DUSP5	0.52	0.49	0.52	0	33.61	36.1	2.49	normal
FRMD4A	0.502	0.57	0.62	-0.0188	35.97	33.62	2.35	normal
PTER	0.513	0.38	0.43	0.0714	36.22	33.92	2.3	normal
ZNF43	0.539	0.36	0.47	0.0868	33.51	35.69	2.18	normal
GFRA4	0.506	0.52	0.55	-0.007	34.03	36.08	2.05	normal
RETREG3	0.511	0.5	0.52	-0.0133	34.06	35.92	1.86	normal
EML4	0.505	0.41	0.45	0.0211	33.93	35.66	1.73	normal
ERC1	0.51	0.41	0.43	0.0162	35.82	34.16	1.66	normal
CCNYL1	0.507	0.52	0.51	-0.0089	35.63	34.04	1.59	normal
EML6	0.51	0.54	0.6	-0.0678	34.43	35.59	1.16	normal
RETREG2	0.505	0.36	0.43	0.0337	35.41	34.43	0.98	normal
CCDC6	0.502	0.43	0.45	0.0145	35.27	34.44	0.83	normal
ALOX5	0.514	0.61	0.68	-0.0708	35	34.68	0.32	normal

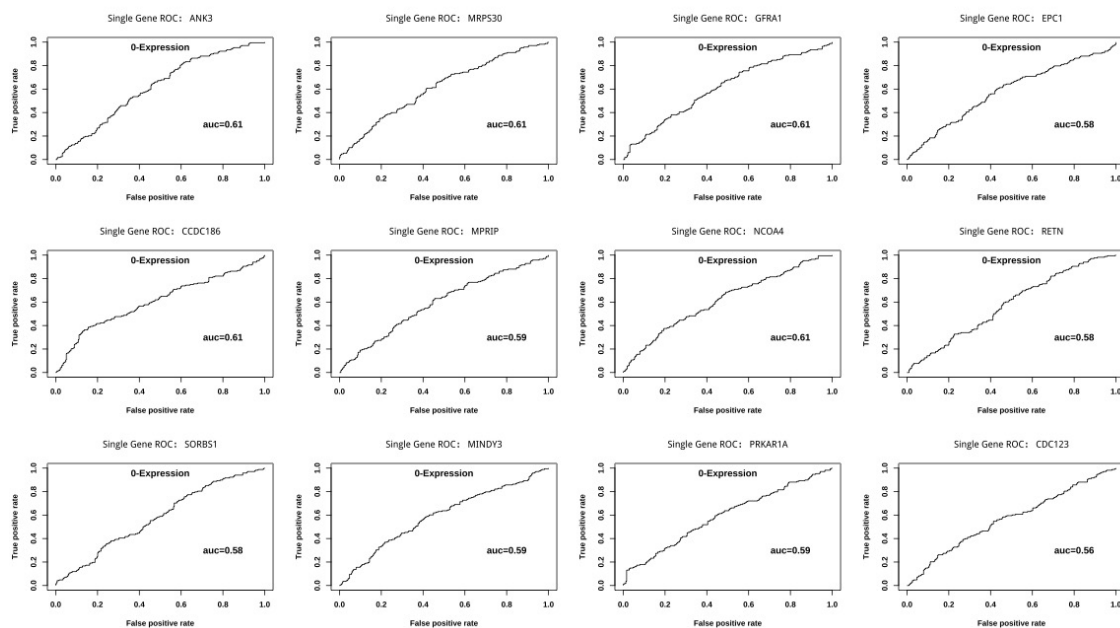


Figure 7: Univariate ROCs of top 12 genes in RET cluster.

the next 3 down-expressed genes and with similar  $P_{\delta}$ . SDC4 is a cell surface proteoglycan that bears heparan sulfate. SDC4-ROS1 fusion is rare in lung cancer, a case was reported that SDC4-ROS1 fusion positive was treated with crizotinib followed by three cycles of chemotherapy, after disease progression it was revealed the original SDC4-ROS1 fusion along with a KRAS point mutation (p.G12D)[78].

### 3.1.9 TP53 Cluster

TP53 cluster contains 11 members. The ROCs are presented in Figure 9 and the corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks are listed in Table 12. There are 9 members with  $P_{\delta} \geq 5\%$  within which TP53, TP53BP1, TP53BP2, TP53I13, TP53I3, TP53INP2, TP53RK are over-expressed while TP53INP1, TP53TG5 are down. The rest two normal ones are TP53I11 and TP53TG1. At the top is down-expressed TP53INP1 with the highest  $P_{\delta} = 18.42\%$  while the second is the over-expressed TP53BP2 with  $P_{\delta} = 16\%$ . Unlike other clusters of which the seeds have modest  $P_{\delta}$ , TP53 itself is over-expressed and stands at the third with  $P_{\delta} = 11.73\%$ . TP53INP1, named as tumor protein p53-inducible nuclear protein 1, is a tumor suppressor, over-expressed during stress responses including inflammation and regulating metabolic homeostasis[50]. Moreover it plays important role in DNA damage response[52]. On the contrary, TP53INP2 is over-expressed with notable  $P_{\delta} = 9.83\%$  and it plays dual roles and switches between transcription and autophagy by sensing the nutrient status [70]. TP53BP2, P53-binding protein 2, also called apoptosis stimulating protein 2 of P53 (ASPP2), is involved with multiple pathways in tumorigenesis[20]. Similarly, TP53BP1 is also over-expressed but with modest  $P_{\delta} = 7.94\%$  and plays critical roles in DNA damage response in cancer[36].

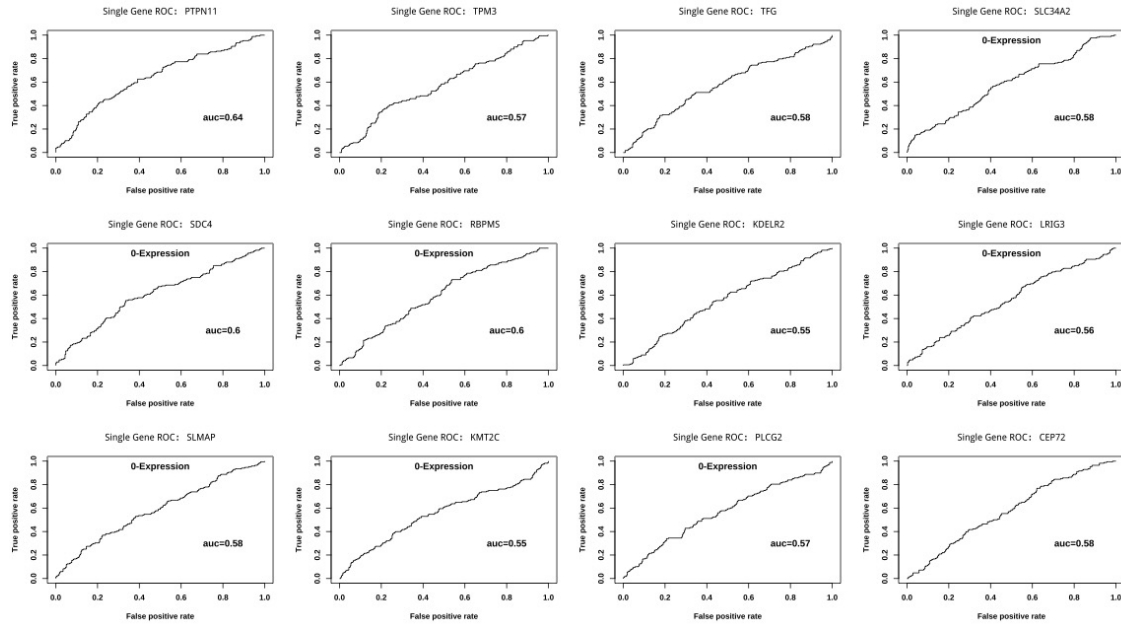


Figure 8: Univariate ROCs of top 12 genes in ROS1 cluster.

Table 11: AUCs and recurrence risks of ROS1 genes ordered by  $P_{\delta}$ .

GENE	AUC	FPR	TPR	$T_q$	$P_{above}(\%)$	$P_{below}(\%)$	$P_{\delta}(\%)$	Status
PTPN11	0.637	0.39	0.62	0.003	45.81	25.1	20.71	up
TPM3	0.573	0.26	0.42	0.0359	46.41	29.48	16.93	up
TFG	0.576	0.35	0.51	0.0718	43.88	28.67	15.21	up
SLC34A2	0.584	0.41	0.57	0.0376	28.63	40.39	11.76	down
SDC4	0.602	0.34	0.55	0.0815	27.89	39.38	11.49	down
RBPMS	0.596	0.54	0.73	-0.057	30.07	41.26	11.19	down
KDELR2	0.552	0.43	0.54	0.036	40.44	29.96	10.48	up
LRIG3	0.56	0.55	0.66	-0.1199	31.33	41.57	10.24	down
SLMAP	0.584	0.39	0.53	0.0339	29.02	38.75	9.73	down
KMT2C	0.554	0.39	0.53	0.0126	29.82	39.02	9.2	down
PLCG2	0.568	0.39	0.51	0.05	29.07	38.06	8.99	down
CEP72	0.576	0.45	0.55	0.0179	39.41	30.49	8.92	up
MYO5C	0.573	0.36	0.5	0.1109	29.83	37.87	8.04	down
PROS1	0.579	0.52	0.66	-0.0401	31.15	38.66	7.51	down
CD74	0.547	0.45	0.55	0.0118	31	38.34	7.34	down
TPD52L1	0.541	0.41	0.48	0.0895	38.57	31.99	6.58	up
EZR	0.539	0.49	0.6	-0.0131	31.71	38.14	6.43	down
ROS1	0.578	0.44	0.52	0.0206	31.8	37.86	6.06	down
VAV3	0.534	0.45	0.51	0.0014	38.05	32.03	6.02	up
CLTC	0.512	0.39	0.45	0.0321	38.38	32.39	5.99	up
WNK1	0.511	0.49	0.56	-0.0052	37.75	31.76	5.99	up
MSN	0.55	0.39	0.51	0.0511	31.74	36.51	4.77	normal
MAPK1	0.519	0.49	0.52	-0.005	36.67	33.06	3.61	normal
TMEM106B	0.541	0.41	0.47	0.011	33.2	36.68	3.48	normal
SLC6A17	0.524	0.54	0.58	-0.0044	36.36	33.03	3.33	normal
MAPK3	0.536	0.42	0.48	0.0142	33.33	36	2.67	normal
LIMA1	0.539	0.34	0.44	0.0851	36.31	34.08	2.23	normal
ZCCHC8	0.506	0.53	0.6	-0.0185	33.84	36.07	2.23	normal
IRS1	0.52	0.48	0.52	0.0129	33.91	35.71	1.8	normal
GOPC	0.502	0.5	0.55	-0.0096	34.08	35.81	1.73	normal
CCDC6	0.502	0.43	0.45	0.0145	35.27	34.44	0.83	normal
AKT1	0.513	0.4	0.46	0.0371	35.15	34.64	0.51	normal
STAT3	0.53	0.37	0.45	0.0507	34.91	34.82	0.09	normal



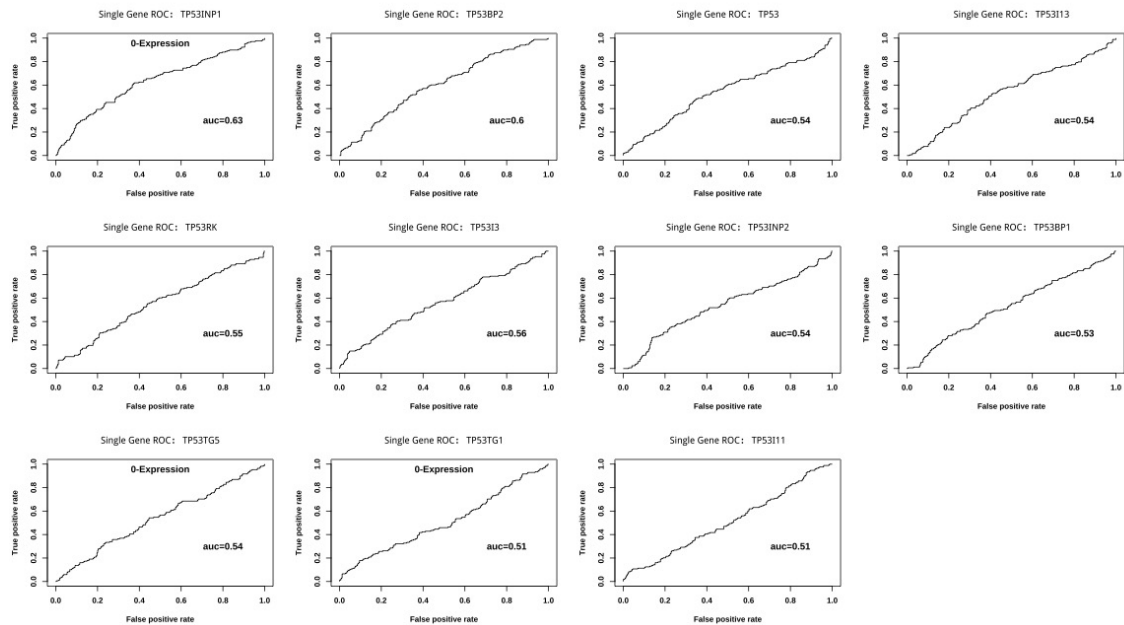


Figure 9: Univariate ROCs of 11 genes in TP53 cluster.

Table 12: AUCs and recurrence risks of TP53 genes ordered by  $P_{\delta}$ .

GENE	AUC	FPR	TPR	$T_g$	$P_{above}(\%)$	$P_{below}(\%)$	$P_{\delta}(\%)$	Status
TP53INP1	0.629	0.38	0.61	0.0189	25.11	43.53	18.42	down
TP53BP2	0.602	0.39	0.56	0.0261	43.72	27.72	16	up
TP53	0.541	0.39	0.51	0.0377	41.55	29.82	11.73	up
TP53I13	0.536	0.41	0.53	0.0241	40.64	30.04	10.6	up
TP53RK	0.55	0.44	0.55	0.0167	40	30.16	9.84	up
TP53I3	0.562	0.41	0.52	0.0458	40.28	30.45	9.83	up
TP53INP2	0.537	0.41	0.52	0.0178	40.28	30.45	9.83	up
TP53BP1	0.531	0.39	0.47	0.02	39.5	31.56	7.94	up
TP53TG5	0.535	0.45	0.54	0.0017	31.4	38.33	6.93	down
TP53TG1	0.508	0.39	0.42	0.0443	36.11	33.83	2.28	normal
TP53I11	0.507	0.45	0.45	0.0307	34.88	34.83	0.05	normal

### 3.1.10 PDCD1(PD1) Cluster

As shown in Table 1, only 15 genes were pre-selected for PDCD1 cluster. The ROCs are presented in Figure 10 and the corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks are listed in Table 13. There are 8 members with  $P_\delta \geq 5\%$ , accounting for 53%. PTPN11, which is also a member in ROS1 cluster, also appears at the top, PDCD1 suppresses T-cell activation through the recruitment of PTPN11[34]. The second LAG3 and the third PDCD1LG2 have  $P_\delta$  around 11%. Lymphocyte activation gene 3, LAG3, is a T cell activation inhibitory coreceptors similar to PDCD1 and CTLA4 and emerged as the third important immunotherapy target[35]. LAG3 and PDCD1 synergistically regulate T cell function[66], they collaborate to limit CD8+ T cell signaling and weaken anti-tumor immunity and dual blockade of them is a promising immunotherapy strategy[22]. Moreover, a over-expressed ligand of LAG3, Fibrinogen-like protein 1, FGL1, is also a cluster member and has  $P_\delta = 5.87\%$ . PDCD1LG2 (i.e. PD-L2) is one of two PDCD1 ligands and has been emerged as another immunotherapy target similar to PD-L1[65]. Next tier consists of 3 down-expressed genes with medium prediction power, HLA-DRB1, ZAP70 and PRKCQ, with  $P_\delta$  range in between 6% and 8.42%. HLA-DRB1 is a HLA Class II Antigen. ZAP70, called Zeta Chain Of T Cell Receptor Associated Protein Kinase, regulates motility, adhesion and cytokine expression of mature T cell. PDCD1 modulation of T cell involves inhibition of TCR-mediated phosphorylation of ZAP70 and association with CD3Z, and downstream inhibition of PKCQ which is required for T cell IL-2 production[53]. Lastly, CD80 with  $P_\delta = 5.85$  is also over-expressed. CD80 is a ligand of CTLA4, just like PD-L1 as a ligand of PD1, CD80 and PD-L1 interaction suggests significant crosstalk between PD1 pathways and CTLA4 pathways[51, 75]. However, PD1 itself, CD274/PD-L1, and other important PD1 related gene HLA-DQB1, CD3D/E, CD247(CD3Z), and CD4 have  $P_\delta < 5$ , with expression levels not strongly related to lung cancer recurrence by this training data set.

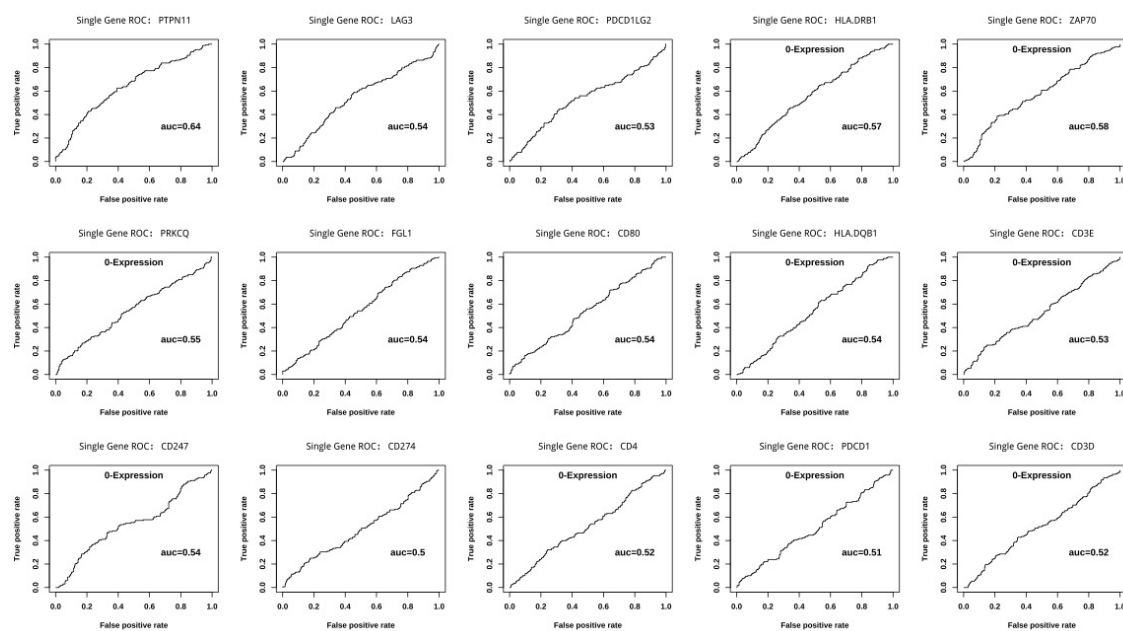


Figure 10: Univariate ROCs of 15 genes in PDCD1 cluster.

Table 13: AUCs and recurrence risks of PDCD1 genes ordered by the prediction power

GENE	AUC	FPR	TPR	$T_g$	$P_{above}(\%)$	$P_{below}(\%)$	$P_\delta(\%)$	Status
PTPN11	0.637	0.39	0.62	0.003	45.81	25.1	20.71	up
LAG3	0.542	0.45	0.58	0.0026	40.76	29.1	11.66	up
PDCD1LG2	0.534	0.42	0.54	0.0038	40.81	29.73	11.08	up
HLA-DRB1	0.566	0.48	0.59	-0.0149	30.8	39.22	8.42	down
ZAP70	0.578	0.38	0.51	0.0569	30.81	37.68	6.87	down
PRKCQ	0.546	0.42	0.52	0.0642	31.28	37.28	6	down
FGL1	0.543	0.43	0.49	0.0245	38.07	32.2	5.87	up
CD80	0.537	0.47	0.54	0.0012	37.82	31.97	5.85	up
HLA-DQB1	0.536	0.53	0.62	-0.067	33.45	36.76	3.31	normal
CD3E	0.53	0.31	0.39	0.0709	32.9	35.78	2.88	normal
CD247	0.541	0.42	0.54	0.051	33.97	35.53	1.56	normal
CD274	0.504	0.46	0.49	0.0626	34.39	35.15	0.76	normal
CD4	0.52	0.34	0.4	0.1319	34.59	34.96	0.37	normal
PDCD1	0.505	0.37	0.4	0.0202	34.67	34.98	0.31	normal
CD3D	0.52	0.42	0.48	0.0562	34.98	34.75	0.23	normal

### 3.1.11 CTLA4 Cluster

CTLA4 cluster contains 17 members. The ROCs are presented in Figure 11 and the corresponding AUCs, FPRs, TPRs, threshold  $T_g$ , population risks are listed in Table 14. There are 10 members with  $P_\delta \geq 5\%$ , accounting for 59%. Similar to PD1, PTPN11 is still the one with maximal  $P_\delta$ . CD276 and CD86 are the next highest two over-expressed genes with  $P_\delta$  very close to PTPN11. CTLA4 itself, CD80 and GRB2 are over-expressed with modest  $P_\delta$ . CD276, also known as B7-H3, CD80 and CD86 belong to the same B7 family as PD-L1. CTLA4 is a homologue of CD28 and they are coreceptors, GRB2, called growth factor receptor-bound protein 2, is an important adaptor participating CD28 and CTLA-4 signaling mechanisms[49]. CD80/86 binds to CD28 while CTLA4 reduce their interaction time. Synergistically with CTLA4, CD276 inhibits T cell activation by inhibiting IL-2 secretion and evidence suggested that IL20RA is a receptor of CD276[31].

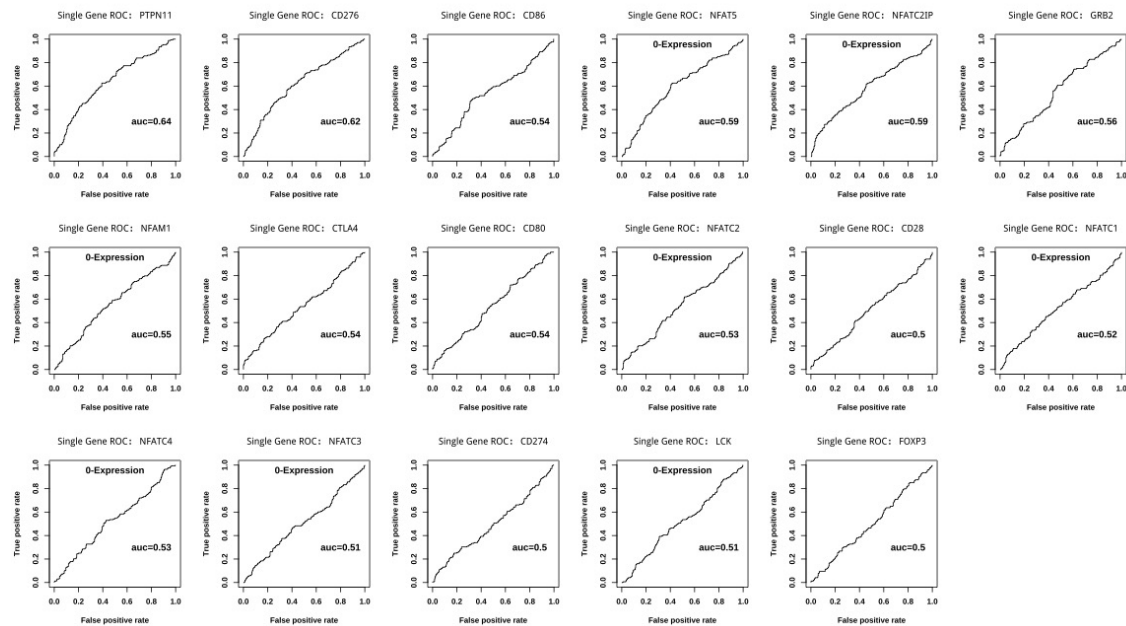


Figure 11: Univariate ROCs of 17 genes in CTLA4 cluster.

Table 14: AUCs and recurrence risks of CTLA4 genes ordered by  $P_\delta$ .

GENE	AUC	FPR	TPR	$T_p$	$P_{above}(\%)$	$P_{below}(\%)$	$P_\delta(\%)$	Status
PTPN11	0.637	0.39	0.62	0.003	45.81	25.1	20.71	up
CD276	0.62	0.39	0.6	0.0269	44.84	26.25	18.59	up
CD86	0.541	0.33	0.49	0.0829	44.15	28.91	15.24	up
NFAT5	0.594	0.41	0.62	0.0219	27.59	40.14	12.55	down
NFATC2IP	0.587	0.46	0.62	-0.0034	29.1	40.76	11.66	down
GRB2	0.559	0.47	0.6	-0.0076	40.49	28.94	11.55	up
NFAM1	0.555	0.4	0.52	0.0249	30.21	37.93	7.72	down
CTLA4	0.54	0.41	0.49	0.049	38.68	31.85	6.83	up
CD80	0.537	0.47	0.54	0.0012	37.82	31.97	5.85	up
NFATC2	0.533	0.52	0.62	-0.0134	32.16	37.89	5.73	down
CD28	0.505	0.45	0.49	0.0201	36.89	33.07	3.82	normal
NFATC1	0.522	0.38	0.45	0.0334	32.77	36.07	3.3	normal
NFATC4	0.526	0.43	0.53	0.0131	33.33	36.03	2.7	normal
NFATC3	0.507	0.41	0.47	0.0303	35.57	34.38	1.19	normal
CD274	0.504	0.46	0.49	0.0626	34.39	35.15	0.76	normal
LCK	0.512	0.39	0.46	0.0851	35.24	34.56	0.68	normal
FOXP3	0.501	0.44	0.43	0.0132	34.93	34.8	0.13	normal

### 3.1.12 Cluster Member Voting Models

Now that a sample is assigned a percentage of *abnormal* members for each given cluster, another ROC is plotted using the percentage as a recurrence predictor. The ROCs are presented in Figure 12. Table 15 lists the corresponding AUCs, FPRs, TPRs, threshold  $T_p$ ,  $P_{above}$  representing the recurrence risk of the patient group with abnormal cluster members  $\geq T_p\%$ , and  $P_{below}$  representing that of the opposite group with  $< T_p\%$ . In summary, for each cluster, the recurrence risk of the abnormal group (of all pathological stages) ranges from 74% (PDCD1) to 220% (ALK) higher, comparing to the opposite normal group, which is calculated via  $\frac{P_\delta}{P_{below}} \times 100\%$ . Next the recurrence risks are investigated in more details.

Table 15: AUC, TPR, FPR, Threshold  $T_p$  and Recurrence Risks for 11 Clusters.  $P_{above}$  is the recurrence risk of the patients with the corresponding abnormal cluster members  $\geq T_p\%$ , and  $P_{below}$  is the opposite group with  $< T_p\%$ .

SEED	AUC	$T_p(\%)$	$P_{above}(\%)$	$P_{below}(\%)$	$P_\delta(\%)$	FPR	TPR	ACCURACY	PPV
ALK	0.763	55.56	55.41	17.31	38.1	0.32	0.73	0.7	0.55
BRAF	0.681	57.89	48.62	23.48	25.14	0.36	0.63	0.64	0.49
EGFR	0.671	58.33	46.58	25.1	21.48	0.37	0.61	0.62	0.47
MET	0.656	57.14	43.75	24.78	18.97	0.46	0.67	0.59	0.44
NTRK	0.715	51.35	52.19	19.29	32.9	0.35	0.71	0.67	0.52
RAS	0.685	60	50.52	24.31	26.21	0.31	0.58	0.66	0.51
RET	0.734	55.32	50.21	19.25	30.96	0.39	0.73	0.65	0.5
ROS1	0.682	52.38	48.44	22.96	25.48	0.37	0.65	0.64	0.48
TP53	0.682	55.56	50.49	23.19	27.3	0.32	0.62	0.66	0.5
CTLA4	0.68	60	48.96	25.52	23.44	0.31	0.56	0.64	0.49
PDCD1	0.656	62.5	47.98	27.51	20.47	0.29	0.49	0.64	0.48

### 3.1.13 Clusters Defined Using Combinatory GCEI (cGCEI)

Given an ordered list of gene clusters represented by ALK, BRAF, EGFR, MET, NTRK, RAS, RET, ROS1 and TP53 in the fixed order, a patient is labeled as a 9-digit binary string  $i_1i_2 \dots i_9$ , each digit  $i_k (k = 1, 2, \dots, 9)$  stands for the corresponding gene cluster expression status where 0 is for *normal* while 1 for *abnormal*. This is called driver gene cluster expression signature, for example, 000000000 represents that all 9 clusters are *normal*, 100000000 represents that only cluster ALK is *abnormal* while 111111111 represents that all 9 clusters are *abnormal*, etc. The 9-driver gene cluster expression signature classifies lung cancers into 512 ( $= 2^9$ ) expression types. Similarly, two immunotherapy target genes: PDCD1, CTLA4 give rise to a two-bit signature string. Moreover, by counting the number of 1 in the 9-digit signature string, which is the number of abnormally expressed clusters in 9 driver gene clusters, called a combinatory *GCEI* and denoted as

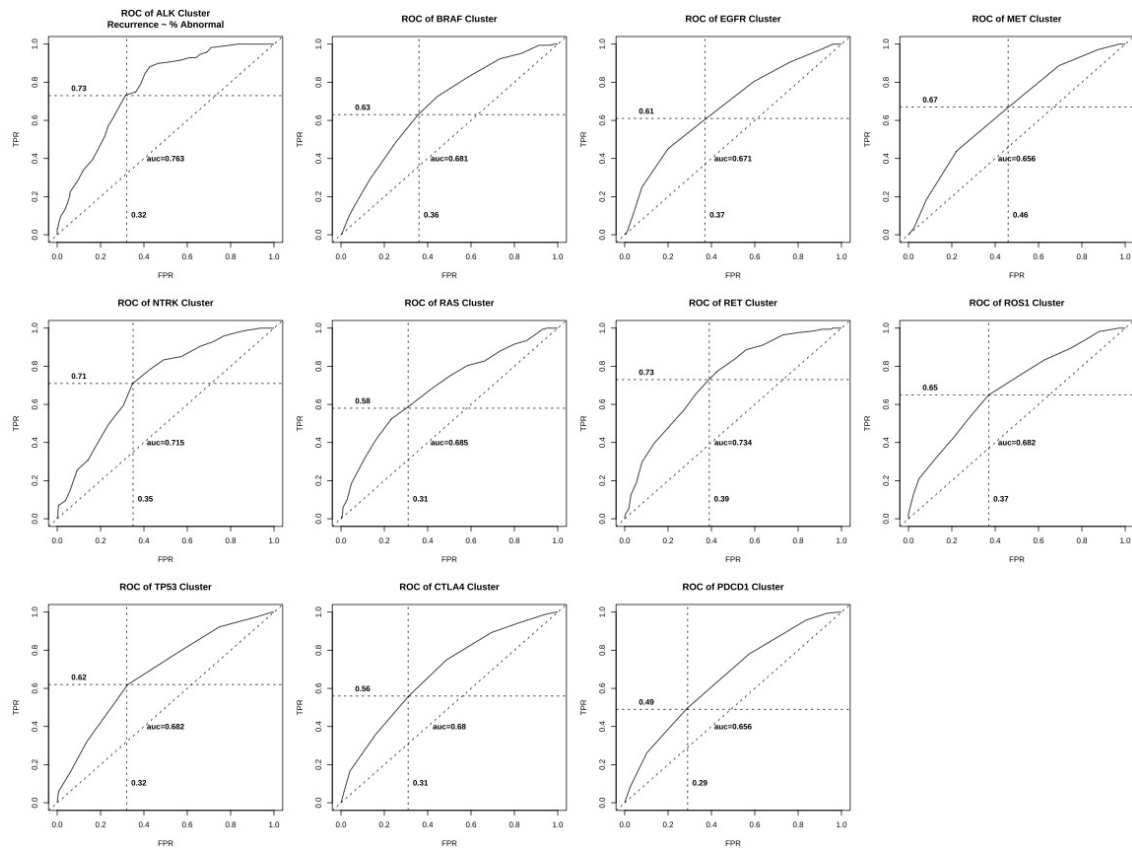


Figure 12: Univariate ROCs of 11 clusters. The percentage of *abnormal* members in each cluster was used as a predictor to recurrence.

*cGCEI*. Patients were then grouped into 10 groups with  $cGCEI = 0, 1, 2, 3, \dots, 9$  respectively. For example,  $cGCEI = 0$  is the patients with signature 000000000,  $cGCEI = 1$  is the patients with any signature with only one 1 and eight 0, such as 100000000, 010000000,  $\dots$ , 000000001), and  $cGCEI = 9$  is the patients with signature 111111111, etc. Similarly a two-bit binary string by combining GCEI of PDCD1 and CTLA4 is defined and has 3 status: 0, 1, or 2, representing none of, or one of, or both of PDCD1 and CTLA4 clusters are *abnormal*. Furthermore, another combinatory GCEI is defined by thresholding *cGCEI* values, a meaningful threshold value of 5 is used to collapse 10 groups with  $cGCEI = 0, 1, 2, 3, \dots, 9$  to only two groups, denoted as *DGCntGT5*, of which the value 1 stands for count of abnormal driver gene cluster is  $> 5$  and 0 for  $\leq 5$ . Therefore,  $DGCntGT5 = 1$  means that there are at least 6 abnormal clusters in 9 driver gene clusters, and  $DGCntGT5 = 0$  means that there are at most 5 abnormal driver gene clusters. All these labeling schemes for lung cancers have dramatic indication for recurrence risks, of which the *abnormal* group is in general 130% – 300% of the *normal* counterpart as shown below.

### 3.2 Recurrence Risks

In the above lung cancers were labeled as *normal* ( $GCEI = 0$ ) or *abnormal* ( $GCEI = 1$ ) with respect to a given cluster or a combination of atomic GCEIs. Next the recurrence risks were assessed for the subpopulations defined by individual GCEI status and combinations of GCEIs. For a given atomic or a combinatory GCEI, the recurrence risk, defined as the percentage of the recurred patients, was calculated with respect to the GCEI status for patients of different pathological stages, namely of stage I, of stage II-V, and of all stages respectively. Table 16 lists the recurrence risks for subpopulations labeled by the atomic GCEI indicators and *DGCntGT5* indicator. It shows that ALK cluster gives the largest risk ratio of lung cancer group with  $GCEI = 1$  over  $GCEI = 0$  for 3 stage groups, with 320%, 332%, 188% for all stages, Stage I, Stage II-IV respectively. As an example of the ratio calculation, take the values in Table 16 corresponding to ALK for all stages, 320% was derived by  $\frac{55.41}{17.31} \times 100\% = 320\%$ , other ratios were calculated similarly. As for the minimal ratio, PDCD1 gives 174% for all stages, MET gives 169% for Stage I, and EGFR gives 109% for Stage II-IV. On average, the risk ratio of group with  $GCEI = 1$  over  $GCEI = 0$  is 222%, 247%, 134% for all stages, Stage I, Stage II-IV respectively. This demonstrates the power of recurrence risk stratification with gene cluster expression voting strategy.

Furthermore, the recurrence risks were also calculated based on binary string signatures of the atomic GCEIs. As described in the above, *cGCEIs* corresponding to the ordered 9-gene list (ALK, BRAF, EGFR, MET, NTRK, RAS, RET, ROS1, TP53) separate lung cancers into 10 groups by counting number of *abnormal* clusters, or the number of 1 in the signature string. Table 17 listed the recurrence percentages for 10 groups defined by 9-gene signatures. It shows that the recurrence risk increases along with *cGCEI* values, namely the number of *abnormal* clusters. For  $cGCEI = 0$ , the recurrence risk is merely 7.02%, when there is one and only one abnormal cluster ( $cGCEI = 0$ ), the risk more than doubled to 15.28%, and then increases to 20.41% for  $cGCEI = 3$ . Interestingly, it then comes a hiccup where the risk goes down to 17.50% for  $cGCEI = 4$ , but this might be due to the data size. After  $cGCEI \geq 6$ , the risk goes beyond 56.36% to an astonishing 72.73% for the group of patients with  $cGCEI = 9$  where all 9 driver clusters show abnormal expressions based on member voting and it has only one signature 111111111. This explains the rationale that we defined a new GCEI based on *DGCntGT5* in the proceeding sub-section.



Table 16: Recurrence Percentages of Lung Cancers in Different Stage Groups Flagged by GCEI. Group risk of  $GCEI = 1$  is typically 130% to 300% of the corresponding group of  $GCEI = 0$ .

Subpopulation	All (Stage I-IV)		Stage I		Stage II-IV	
	GCEI=0(%)	GCEI=1(%)	GCEI=0(%)	GCEI=1(%)	GCEI=0(%)	GCEI=1(%)
ALK	17.31	55.41	12.56	41.75	35.85	67.23
BRAF	23.48	48.62	15.38	36.27	53.57	59.48
EGFR	25.1	46.58	16.67	33.02	54.24	59.29
MET	24.78	43.75	17.14	28.89	50.98	60.33
NTRK	19.29	52.19	12.87	39.81	44.23	63.33
RAS	24.31	50.52	16.36	35.42	47.3	65.31
RET	19.25	50.21	14.14	35.29	39.58	64.52
ROS1	22.96	48.44	16.92	32.11	44.64	63.79
TP53	23.19	50.49	14.56	37.5	48.57	63.73
CTLA4	25.52	48.96	13.79	38.32	52.87	62.35
PDCD1	27.51	47.98	16.13	36.56	54.35	61.25
DGCntGT5	18.84	59.47	13.3	49.35	40.68	66.37
Average	22.63	50.22	14.98	37.02	47.24	63.08

Table 17: Recurrence Risks of cGCEI based on 9-digit Signatures (Only Evaluated for all Stages).

cGCEI	Ex Signature	NoneRecurred	Recurred	Total	Recurrence(%)
0	000000000	53	4	57	7.02
1	000000001	61	11	72	15.28
2	000001100	39	10	49	20.41
3	000100101	33	7	40	17.5
4	101100010	30	12	42	28.57
5	001011011	21	11	32	34.38
6	111110001	24	31	55	56.36
7	111110101	24	26	50	52
8	111110111	20	32	52	61.54
9	111111111	9	24	33	72.73

## 4 Discussion

Although DNA-based genetic tests have been routinely used for targeted therapy and immunotherapy, the proportion of patients whose tumors can be targeted therapeutically is limited and is usually less than 30%. A retrospective study of 2257 metastatic NSCLC patients showed that more than half of tested patients did not have results prior to first-line treatment and fewer than 20% of tested patients had results for all 4 driver mutations (ALK, EGFR, ROS1, BRAF) and PD-L1 prior to first-line treatment. Moreover, although the turnaround time improved from year 2017 to 2019, not all patients who tested positive for driver mutations received targeted therapy in the first-line setting[37]. Hence the percent of patients who received targeted therapy was less than 30%. We propose that for a given driver gene cluster, the targeted therapy with respect to the gene may be beneficial to the patient group of  $GCEI = 1$ . In addition, immunotherapy may be beneficial to the patient group of  $GCEI = 1$  with respect to PDCD1 or CTLA4 clusters. The WINTHER trial (NCT01856296) [47] was the first clinical trial to navigate lung, colon, head and neck, and other cancer patients with previous treatments to therapy on the basis of fresh biopsy-derived DNA sequencing or RNA expression (tumor versus normal). It shows that transcriptome profiling is as useful as DNA tests for improving therapy recommendations and patient outcome, and hence transcriptome analysis can expand personalized treatment.

## References

- [1] Andre F et al. AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. *Cancer Discov.* 2017 Aug;7(8):818-831. doi:10.1158/2159-8290.CD-17-0151. Epub 2017 Jun 1. PMID: 28572459; PMCID: PMC5611790.

- [2] Pugh TJ et al. AACR Project GENIE Consortium, Genomics and Analysis Working Group. AACR Project GENIE: 100,000 Cases and Beyond. *Cancer Discov.* 2022 Sep 2;12(9):2044-2057. doi:10.1158/2159-8290.CD-21-1547. PMID: 35819403; PMCID: PMC9437568.
- [3] Scharpf RB et al. Genomic Landscapes and Hallmarks of Mutant RAS in Human Cancers. *Cancer Res.* 1 November 2022; 82 (21): 4058–4078. doi:10.1158/0008-5472.CAN-22-1731
- [4] Ambati SR et al. Entrectinib in two pediatric patients with inflammatory myofibroblastic tumors harboring ROS1 or ALK gene fusions. *JCO Precis Oncol.* 2018;2:PO.18.00095. doi:10.1200/PO.18.00095. Epub 2018 Sep 13. PMID: 31763577; PMCID: PMC6874363.
- [5] An S et al. Lower Ras expression as an independent predictor of patient outcomes in lung cancer treated with bevacizumab plus chemotherapy. *Cancer Gene Ther.* 21, 110–114 (2014). doi:10.1038/cgt.2014.5
- [6] Aruga J et al. Human SLITRK family genes: genomic organization and expression profiling in normal brain and brain tumor tissue. *Gene.* 2003 Oct 2;315:87-94. doi:10.1016/s0378-1119(03)00715-7. PMID: 14557068.
- [7] Biswas D et al. A clonal expression biomarker associates with lung cancer mortality. *Nat Med.* 2019 Oct;25(10):1540-1548. doi: 10.1038/s41591-019-0595-z. Epub 2019 Oct 7. Erratum in: *Nat Med.* 2020 Jul;26(7):1148. PMID: 31591602; PMCID: PMC6984959.
- [8] Bollig-Fischer A et al. Role of novel cancer gene SLITRK3 to activate NTRK3 in squamous cell lung cancer. *Mol Biomed.* 2, 26 (2021). doi:10.1186/s43556-021-00051-2.
- [9] Bueno R et al. Multi-Institutional Prospective Validation of Prognostic mRNA Signatures in Early Stage Squamous Lung Cancer (Alliance). *J Thorac Oncol.* 2020 Nov;15(11):1748-1757. doi: 10.1016/j.jtho.2020.07.005. Epub 2020 Jul 24. PMID: 32717408; PMCID: PMC8278226.
- [10] Buzdin A et al. RNA sequencing for research and diagnostics in clinical oncology. *Semin Cancer Biol.* 2020 Feb;60:311-323. doi: 10.1016/j.semcancer.2019.07.010. Epub 2019 Aug 11. PMID: 31412295.
- [11] Cai C et al. Coexisting of COX7A2L-ALK, LINC01210-ALK, ATP13A4-ALK and acquired SLCO2A1-ALK in a lung adenocarcinoma with rearrangements loss during the treatment of Crizotinib and Ceritinib: a case report. *Onco Targets Ther.* 2020 Aug 20;13:8313-8316. doi:10.2147/OTT.S258067. PMID: 32903930; PMCID: PMC7445493.
- [12] Cocco E et al. NTRK fusion-positive cancers and TRK inhibitor therapy. *Nat Rev Clin Oncol.* 2018 Dec;15(12):731-747. doi:10.1038/s41571-018-0113-0. PMID: 30333516; PMCID: PMC6419506.
- [13] Couëtoux du Tertre M et al. Analysis of the genomic landscape in ALK+ NSCLC patients identifies novel aberrations associated with clinical outcomes. *Mol Cancer Ther.* 2019 Sep;18(9):1628-1636. doi:10.1158/1535-7163.MCT-19-0105. Epub 2019 Jun 26. PMID: 31243098.
- [14] Choi H et al. Horyun Choi, Jinah Kim, Yeun Ho Lee, Leeseul Kim, Na Hyun Kim, Young Kwang Chae. Acquired TPM3-NTRK1 fusion resistant to larotrectinib in a non-small cell lung cancer with EML4-ALK fusion progressed on lorlatinib [abstract]. In: *Proceedings of the American Association for Cancer Research Annual Meeting 2022.* 2022 Apr 8-13. Philadelphia (PA): AACR; *Cancer Res* 2022;82(12.Suppl): Abstract nr 4120.

- [15] Correale P et al. Distinctive germline expression of class I human leukocyte antigen (HLA) alleles and DRB1 heterozygosity predict the outcome of patients with non-small cell lung cancer receiving PD-1/PD-L1 immune checkpoint blockade. *J Immunother Cancer*. 2020 Jun;8(1):e000733. doi:10.1136/jitc-2020-000733. PMID: 32554614; PMCID: PMC7304840.
- [16] Davies KD et al. Identifying and targeting ROS1 gene fusions in non-small cell lung cancer. *Clin Cancer Res*. 2012 Sep 1;18(17):4570-9. doi: 10.1158/1078-0432.CCR-12-0550. Epub 2012 Aug 23. PMID: 22919003; PMCID: PMC3703205.
- [17] Deng H et al. HLA-DRB1: A new potential prognostic factor and therapeutic target of cutaneous melanoma and an indicator of tumor microenvironment remodeling. *PLoS One*. 2022 Sep 21;17(9):e0274897. doi:10.1371/journal.pone.0274897. PMID: 36129956; PMCID: PMC9491554.
- [18] Farago AF et al. Clinicopathologic Features of Non-Small-Cell Lung Cancer Harboring an NTRK Gene Fusion. *JCO Precision Oncology*. 2018; 2:2-12. doi:10.1200/PO.18.00037.
- [19] Fang WF et al. Abstract 3460: MPRIP-ALK, a novel ALK rearrangement that responds to ALK inhibitor in non-small-cell lung cancer. *Cancer Res 1 July 2019*; 79 (13\_Supplement): 3460. doi:10.1158/1538-7445.AM2019-3460
- [20] Hou YF et al. TP53BP2: Roles in suppressing tumorigenesis and therapeutic opportunities, *Genes and Diseases*, 2022, ISSN 2352-3042, doi:10.1016/j.gendis.2022.08.014.
- [21] Hsiao SJ et al. Detection of Tumor NTRK Gene Fusions to Identify Patients Who May Benefit from Tyrosine Kinase (TRK) Inhibitor Therapy. *The Journal of Molecular Diagnostics*. 2019, 21(4):553-571, doi:10.1016/j.jmoldx.2019.03.008.
- [22] Huang RY et al. LAG3 and PD1 co-inhibitory molecules collaborate to limit CD8+ T cell signaling and dampen antitumor immunity in a murine ovarian cancer model. *Oncotarget*. 2015 Sep 29;6(29):27359-77. doi:10.18632/oncotarget.4751. PMID: 26318293; PMCID: PMC4694995.
- [23] Inoue Y et al. Clinical significance of PD-L1 and PD-L2 copy number gains in non-small-cell lung cancer. *Oncotarget*. 2016 May 31;7(22):32113-28. doi:10.18632/oncotarget.8528. PMID: 27050074; PMCID: PMC5078001.
- [24] Iwona K et al. Identification of PD-1 ligands: PD-L1 and PD-L2 on macrophages in lung cancer milieu by flow cytometry. *Translational Lung Cancer Research* 2021;10(4). <https://tlcr.amegroups.com/article/view/50051>.
- [25] Jiang C et al. High CEP55 expression is associated with poor prognosis in non-small-cell lung cancer. *Onco Targets Ther*. 2018 Aug 17;11:4979-4990. doi:10.2147/OTT.S165750. PMID: 30154666; PMCID: PMC6103653.
- [26] Kratz JR et al. A practical molecular assay to predict survival in resected non-squamous, non-small-cell lung cancer: development and international validation studies. *Lancet*. 2012 Mar 3;379(9818):823-32. doi: 10.1016/S0140-6736(11)61941-7. Epub 2012 Jan 27. PMID: 22285053; PMCID: PMC3294002.
- [27] Kelley MJ et al. Genetic analysis of the beta-tubulin Gene, TUBB, in non-small-cell lung cancer. *JNCI: Journal of the National Cancer Institute*, Volume 93, Issue 24, 19 December 2001, Pages 1886–1888, <https://doi.org/10.1093/jnci/93.24.1886>

- [28] Lavoie H et al. MEK drives BRAF activation through allosteric control of KSR proteins. *newblock Nature*. 2018 Feb 22;554(7693):549-553. doi:10.1038/nature25478. Epub 2018 Feb 12. PMID: 29433126; PMCID: PMC6433120.
- [29] Li B et al. Development and Validation of an Individualized Immune Prognostic Signature in Early-Stage Nonsquamous Non-Small Cell Lung Cancer. *JAMA Oncol*. 2017 Nov 1;3(11):1529-1537. doi: 10.1001/jamaoncol.2017.1609. PMID: 28687838; PMCID: PMC5710196.
- [30] Li B et al. Genomic characterization and outcome evaluation of kinome fusions in lung cancer revealed novel druggable fusions. *newblock npj Precis. Onc.* 5, 81 (2021). doi:10.1038/s41698-021-00221-z
- [31] Liu S et al. The role of CD276 in cancers. *Front. Oncol.* 11:654684. doi:10.3389/fonc.2021.654684
- [32] Liu F et al. NTRK Fusion in Non-Small Cell Lung Cancer: Diagnosis, Therapy, and TRK Inhibitor Resistance. *Front Oncol.* 2022 Mar 17;12:864666. doi:10.3389/fonc.2022.864666. PMID: 35372074; PMCID: PMC8968138.
- [33] Luo Y et al. Cell Trajectory-Related Genes of Lung Adenocarcinoma Predict Tumor Immune Microenvironment and Prognosis of Patients. *Front Oncol.* 2022 Jul 18;12:911401. doi: 10.3389/fonc.2022.911401. PMID: 35924143; PMCID: PMC9339705.
- [34] Marasco M et al. Molecular mechanism of SHP2 activation by PD-1 stimulation *Science Advances*. 2020; 6(5). doi:10.1126/sciadv.aay4458
- [35] Maruhashi T et al. LAG-3: from molecular functions to clinical applications. *J Immunother Cancer*. 2020 Sep;8(2):e001014. doi:10.1136/jitc-2020-001014. PMID: 32929051; PMCID: PMC7488795.
- [36] Mirza-Aghazadeh-Attari M et al. 53BP1: A key player of DNA damage response with critical functions in cancer *Review DNA Repair (Amst)*. 2019 Jan;73:110-119. doi:10.1016/j.dnarep.2018.11.008.
- [37] Nadler E et al. Real-world patterns of biomarker testing and targeted therapy in de novo metastatic non-small cell lung cancer patients in the US oncology network. *Cancer Treatment and Research Communications*. 2022;Vol. 31. doi:10.1016/j.ctarc.2022.100522.
- [38] Nagy TA et al. KRAS driven expression signature has prognostic power superior to mutation status in non-small cell lung cancer. *Int J Cancer*. 2017 Feb 15;140(4):930-937. doi: 10.1002/ijc.30509. Epub 2016 Nov 23. PMID: 27859136; PMCID: PMC5299512.
- [39] Nagl L et al. Molecular profiling in lung cancer. *memo* 15, 201–205 (2022). <https://doi.org/10.1007/s12254-022-00824-7>
- [40] Okayama H et al. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Res*. 2012 Jan 1;72(1):100-11. doi:10.1158/0008-5472. PMID: 22080568.
- [41] Ou S et al. Catalog of 5' Fusion Partners in ALK-positive NSCLC Circa 2020. *JTO Clin Res Rep*. 2020 Feb 19;1(1):100015. doi:10.1016/j.jtocr.2020.100015. PMID: 34589917; PMCID: PMC8474466.

- [42] Ou S et al. Catalog of 5' fusion partners in RET+ NSCLC Circa 2020. *JTO Clin Res Rep.* 2020 Mar 30;1(2):100037. doi:10.1016/j.jtocrr.2020.100037. PMID: 34589933; PMCID: PMC8474217.
- [43] Ou S et al. Catalog of 5' Fusion Partners in ROS1-Positive NSCLC Circa 2020. *JTO Clin Res Rep.* 2020 Apr 28;1(3):100048. doi:10.1016/j.jtocrr.2020.100048. PMID: 34589944; PMCID: PMC8474457.
- [44] Peng X et al. Molecular Characterization and Clinical Relevance of Metabolic Expression Subtypes in Human Cancers. *Cell Rep.* 2018 Apr 3;23(1):255-269.e4. doi: 10.1016/j.celrep.2018.03.077. PMID: 29617665; PMCID: PMC5916795.
- [45] Qian W et al. Fibrinogen-like protein 1 (FGL1): the next immune checkpoint target. *J Hematol Oncol.* 2021 Sep 15;14(1):147. doi:10.1186/s13045-021-01161-8. PMID: 34526102; PMCID: PMC8444356.
- [46] Qian Y et al. Prognostic cancer gene expression signatures: current status and challenges. *Cells.* 2021 Mar 15;10(3):648. doi: 10.3390/cells10030648. PMID: 33804045; PMCID: PMC8000474.
- [47] Rodon J et al. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. *Nat Med.* 2019 May;25(5):751-758. doi: 10.1038/s41591-019-0424-4. Epub 2019 Apr 22. PMID: 31011205; PMCID: PMC6599610.
- [48] Rousseaux S et al. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. *Sci Transl Med.* 2013 May 22;5(186):186ra66. doi:10.1126/scitranslmed.3005723. PMID: 23698379.
- [49] Rudd CE et al. CD28 and CTLA-4 coreceptor expression and signal transduction. *Immunol Rev.* 2009 May;229(1):12-26. doi:10.1111/j.1600-065X.2009.00770.x. PMID: 19426212; PMCID: PMC4186963.
- [50] Saadi H et al. The stress protein TP53INP1 plays a tumor suppressive role by regulating metabolic homeostasis. *Biochimie.* Volume 118, 2015, Pages 44-50, doi:10.1016/j.biochi.2015.07.024.
- [51] Sansom DM et al. Dimers Aren't Forever: CD80 Breaks up with PD-L1, *Immunity.* Volume 51, Issue 6, 2019, Pages 972-974, doi:10.1016/j.immuni.2019.11.011.
- [52] Shahbazi J et al. Tumor protein 53-induced nuclear protein 1 enhances p53 function and represses tumorigenesis. *Front. Genet.* 13 May 2013, Sec. Cancer Genetics, Volume 4, 2013, doi:10.3389/fgene.2013.00080.
- [53] Sheppard KA et al. PD-1 inhibits T-cell receptor induced phosphorylation of the ZAP70/CD3Z signalosome and downstream signaling to PKC $\theta$ , *FEBS Letters*, 574, doi:10.1016/j.febslet.2004.07.083
- [54] Shu Y et al. Identification of a novel MPRIP-ROS1 fusion and clinical efficacy of Crizotinib in an advanced lung adenocarcinoma patient: a case report. *Onco Targets Ther.* 2020 Oct 13;13:10387-10391. doi:10.2147/OTT.S270961. PMID: 33116618; PMCID: PMC7568639.
- [55] Solomon JP et al. NTRK fusion detection across multiple assays and 33,997 cases: diagnostic implications and pitfalls. *Mod Pathol.* 2020 Jan;33(1):38-46. doi:10.1038/s41379-019-0324-7. Epub 2019 Aug 2. PMID: 31375766; PMCID: PMC7437403.

- [56] Song M et al. Slitrk5 mediates BDNF-dependent TrkB receptor trafficking and signaling. *Developmental Cell*. Volume 33, Issue 6, 2015, Pages 690-702, doi:10.1016/j.devcel.2015.04.009.
- [57] Sorokin M et al. Personalized targeted therapy prescription in colorectal cancer using algorithmic analysis of RNA sequencing data. *BMC Cancer*. 2022 Oct 31;22(1):1113. doi: 10.1186/s12885-022-10177-3. PMID: 36316649; PMCID: PMC9623986.
- [58] Supplitt S et al. Current Achievements and Applications of Transcriptomics in Personalized Cancer Medicine. *Int J Mol Sci*. 2021 Jan 31;22(3):1422. doi: 10.3390/ijms22031422. PMID: 33572595; PMCID: PMC7866970.
- [59] Tang H et al. Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies. *Ann Oncol*. 2017 Apr 1;28(4):733-740. doi: 10.1093/annonc/mdw683. PMID: 28200038; PMCID: PMC5834090.
- [60] Tang R et al. Multiplexed screens identify RAS paralogues HRAS and NRAS as suppressors of KRAS-driven lung cancer growth. *Nat Cell Biol*. 25, 159–169 (2023). doi:10.1038/s41556-022-01049-w
- [61] Vaishnavi A et al. Oncogenic and drug-sensitive NTRK1 rearrangements in lung cancer. *Nat Med*. 2013 Nov;19(11):1469-1472. doi:10.1038/nm.3352. Epub 2013 Oct 27. PMID: 24162815; PMCID: PMC3823836.
- [62] Varga A et al. RAF1/BRAF dimerization integrates the signal from RAS to ERK and ROKA. *Sci Signal*. 2017 Mar 7;10(469):eaa18482. doi:10.1126/scisignal.aai8482. PMID: 28270557.
- [63] Wang M et al. The prognostic value of long non coding RNAs in non small cell lung cancer: A meta-analysis. *Oncotarget*. 2016 Dec 6;7(49):81292-81304. doi: 10.18632/oncotarget.13223. PMID: 27833074; PMCID: PMC5348393.
- [64] Wan YW et al. Signaling pathway-based identification of extensive prognostic gene signatures for lung adenocarcinoma. *Lung Cancer*. 2012 Apr;76(1):98-105. doi: 10.1016/j.lungcan.2011.09.016. Epub 2011 Nov 1. PMID: 22047960; PMCID: PMC3274612.
- [65] Wang Y et al. Evolving landscape of PD-L2: bring new light to checkpoint immunotherapy. *Br J Cancer (2022)*. <https://doi.org/10.1038/s41416-022-02084-y>
- [66] Woo SR et al. Immune inhibitory molecules LAG-3 and PD-1 synergistically regulate T-cell function to promote tumoral immune escape. *Cancer Res*. 2012 Feb 15;72(4):917-27. doi:10.1158/0008-5472.CAN-11-1620. PMID: 22186141; PMCID: PMC3288154.
- [67] Woodard GA et al. Adjuvant Chemotherapy Guided by Molecular Profiling and Improved Outcomes in Early Stage, Non-Small-Cell Lung Cancer. *Clin Lung Cancer*. 2018 Jan;19(1):58-64. doi: 10.1016/j.clc.2017.05.015. Epub 2017 May 31. PMID: 28645632.
- [68] Wu K et al. Circular RNA F-circSR derived from SLC34A2-ROS1 fusion gene promotes cell migration in non-small cell lung cancer. *Mol Cancer*. 2019 May 22;18(1):98. doi: 10.1186/s12943-019-1028-9. PMID: 31118036; PMCID: PMC6530145.
- [69] Xia H et al. Evidence of NTRK1 Fusion as Resistance Mechanism to EGFR TKI in EGFR+ NSCLC: Results From a Large-Scale Survey of NTRK1 Fusions in Chinese Patients With Lung Cancer. *Clin Lung Cancer*. 2020 May;21(3):247-254. doi:10.1016/j.clc.2019.09.004. Epub 2019 Sep 26. PMID: 31761448.



- [70] Xu YF et al. The bifunctional role of TP53INP2 in transcription and autophagy. *Autophagy*, 2020, 16:7, 1341-1343, DOI:10.1080/15548627.2020.1713646
- [71] Yu J et al. Establishment of a Lymph Node Metastasis-Associated Prognostic Signature for Lung Adenocarcinoma. *Genet Res (Camb)*. 2023 Jan 31;2023:6585109. doi: 10.1155/2023/6585109. PMID: 36793937; PMCID: PMC9904923.
- [72] Yu X et al. The miR-195 Axis Regulates Chemoresistance through TUBB and Lung Cancer Progression through BIRC5. *Mol Ther Oncolytics*. 2019 Aug 6;14:288-298. doi:10.1016/j.omto.2019.07.004. PMID: 31508486; PMCID: PMC6727248.
- [73] Zhang K et al. Clinical Characteristics and Molecular Patterns of RET-Rearranged Lung Cancer in Chinese Patients. *Oncol Res*. 2019 May 7;27(5):575-582. doi:10.3727/096504018X15344979253618. Epub 2018 Aug 21. PMID: 30131091; PMCID: PMC7848427.
- [74] Zhao R et al. Identification of NTRK gene fusions in lung adenocarcinomas in the Chinese population. *J Pathol Clin Res*. 2021 Jul;7(4):375-384. doi:10.1002/cjp2.208. Epub 2021 Mar 26. PMID: 33768710; PMCID: PMC8185368.
- [75] Zhao YL et al. PD-L1:CD80 Cis-Heterodimer Triggers the Co-stimulatory Receptor CD28 While Repressing the Inhibitory PD-1 and CTLA-4 Pathways *Immunity*. Volume 51, Issue 6, 2019, Pages 1059-1073, doi:10.1016/j.immuni.2019.11.011.
- [76] Zhu YC et al. Dual drive coexistence of EML4-ALK and TPM3-ROS1 fusion in advanced lung adenocarcinoma: ALK and ROS1 fusion in ADC. *Thoracic Cancer*. 2017; 9. 10.1111/1759-7714.12578.
- [77] Zhu YC et al. CEP72-ROS1: A novel ROS1 oncogenic fusion variant in lung adenocarcinoma identified by next-generation sequencing. *Thorac Cancer*. 2018 May;9(5):652-655. doi: 10.1111/1759-7714.12617. Epub 2018 Mar 8. PMID: 29517860; PMCID: PMC5928353.
- [78] Zhu YC et al. Concurrent ROS1 gene rearrangement and KRAS mutation in lung adenocarcinoma: A case report and literature review. *Thorac Cancer*. 2018 Jan;9(1):159-163. doi:10.1111/1759-7714.12518. Epub 2017 Oct 3. PMID: 28971587; PMCID: PMC5754306.
- [79] Xu Q et al. Identification of Master Regulators Driving Disease Progression, Relapse, and Drug Resistance in Lung Adenocarcinoma. *Front Bioinform*. 2022 Jan 28;2:813960. doi: 10.3389/fbinf.2022.813960. PMID: 36304306; PMCID: PMC9580914.
- [80] Rogers K et al. Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*. 2014 Jul 31;511(7511):543-50. doi: 10.1038/nature13385. Epub 2014 Jul 9. Erratum in: *Nature*. 2014 Oct 9;514(7521):262. [corrected to Rodgers, K]. Erratum in: *Nature*. 2018 Jul;559(7715):E12. PMID: 25079552; PMCID: PMC4231481.