Cancer Screening and Prevention **2024** vol. 3(1) | 24–35
DOI: 10.14218/CSP.2023.00034

Original Article

Gene Cluster Expression Index and Potential Indications for Targeted Therapy and Immunotherapy for Lung Cancers

Aibing Rao*

Shenzhen Luwei (Biomanifold) Biotechnology Limited, Shenzhen, Guangdong, China

Received: July 24, 2023 | **Revised:** November 20, 2023 | **Accepted:** February 06, 2024 | **Published online:** March 25, 2024

Abstract

Background and objectives: About 30% of lung cancer patients are accessible to targeted therapy or immunotherapy based on the current criteria. In this study, a novel gene cluster expression analysis was introduced with a goal to potentially expand the treatments to more patients based on the proposed criteria.

Methods: Selected gene expression omnibus data sets were downloaded, normalized, and analyzed. A univariate recurrence prediction model was built based on the receiver operating characteristic, for which an optimal cutoff was determined to set abnormality status, called the gene cluster expression index (GCEI). Recurrence and survival risks were calculated and compared between two subgroups indexed by the GCEI. Moreover, a combinatory GCEI was also introduced and its performance was analyzed for combined multiple cluster statuses.

Results: The recurrence risks of the patient subgroups with abnormally expressed clusters with *GCEI* = 1 were much higher than for the corresponding normal subgroup with *GCEI* = 0. The higher risks ranged from 120–300% that of the corresponding lower-risk group.

Conclusions: The GCEI can be used to classify lung cancers with dramatically different recurrence risks and may be used to guide targeted therapy or immunotherapy for patients who are in a high-risk group but do not qualify for such treatment according to conventional companion tests.

Introduction

Gene expression analysis and transcriptome profiling have been extensively explored in lung cancer[1–5]; however, there has not been much research on gene expression profiling for targeted therapy and immunotherapy. The current standard approach to targeted therapy is via companion DNA tests,[6] while the immunotherapy option involves routine tests, such as pathological immunoassay for the protein expression of PD1 or PD-L1 or by DNA-based Next Generation Sequencing (NGS) assessment of the tumor mutation burden, mismatch repair, and microsatellite instability. Transcriptome profiling has emerged as a promising biomarker for cancer treatment and has shown encouraging clinical results.[7] In non-small cell lung cancer (NSCLC), a study showed that gene expression profiling might have better prognostic prediction power than considering the mutation status.[8] In this *in-silico* study, we present a framework with a novel analysis procedure to introduce the gene cluster expression index (GCEI) and demonstrate its power to stratify lung cancer patients with dramatically different prognostic risks.

Keywords: Transcriptome profiling; Gene cluster expression index; RNA expression analysis; Multivariate modeling; Lung cancer; Targeted therapy; Immunotherapy.

Abbreviations: ALK, anaplastic lymphoma kinase; AUC, area under the curve; cGCEI, combined gene cluster expression index; FPR, false positive rate; GCEI, gene cluster expression index; NSCLC, non-small cell lung cancer; PPV, positive prediction value; ROC, receiver operating characteristic; TPR, true positive rate..

***Correspondence to:** Aibing Rao, Shenzhen Luwei (Biomanifold) Biotechnology Limited, 10th Floor, Clou Building B, Baoshen Road, Nanshan District, Shenzhen, Guangdong 518057, China. ORCID: https://orcid.org/0009-0000-9966-2853. Tel: +86-13691947926, E-mail: aibing.rao@enlightendx.com

How to cite this article: Rao A. Gene Cluster Expression Index and Potential Indications for Targeted Therapy and Immunotherapy for Lung Cancers. *Cancer Screen Prev* 2024;3(1):24–35. doi: 10.14218/CSP.2023.00034.

Materials and methods

Preparation and preprocessing of the data sets

Training data set

Two lung cancer microarray data sets: GSE30219, originating from Rousseaux *et al*.,[9] and GSE31210, originating from Okayama *et al*.,[10] were downloaded from the Lung Cancer Explorer (LCE) web portal with standardized clinical data according to Cai *et al*.[11] There were 482 patients with non-empty recurrence labels,

among whom 168 cases (35%) were labeled as recurred within two years since diagnosis. The two data sets were further normalized by aligning the median of all the samples to 0 and then by aligning the median of all the genes to 0 independently. There were about 17,000 common genes in the selected data sets here and those used in the training data listed below. All samples with missing recurrence or missing expression value of a common gene were omitted. A combined data set was obtained by slicing and aligning the common genes and common clinical variables from the two normalized sets and then by stacking them together. There were 310 patients with Stage I cancers, 111 with Stage II, 53 with Stage III or IV, and 8 with unknown stages. The average patient age was 61 years old, with the youngest being 15 years old and the eldest 84 years old. There were 330 males and 152 females.

Testing data set

Other data sets, namely GSE37745, GSE41271, GSE50081, and GSE74777 with recurrence annotations, were also downloaded from the same LCE web portal as above and also normalized with the medians of the samples and genes aligned to 0, respectively. In addition, the expression of each gene in each data set was further normalized according to the distribution of the training set for the sake of applying thresholds from the training set directly to the testing set. The goal was to align the first and the third quartiles between the testing and the training data sets by linear mapping. Here, for a given gene, we let Q_1, Q_2 be the first and the third quartiles of its expression vector in the training data, and T_1, T_2 be the first and the third quartiles of its expression vector in the testing data. The normalized value $N(x)$ was then obtained from the original value x via the formula: $N(x) = (x - T_1)/(T_2 - T\,1) \times (Q_2 - Q1) + Q_1$. Finally, all 4 normalized data sets were stacked together. Note that T_1 and T_2 were calculated only using a subset with the same proportion of recurred samples as that of the training set. This normalization step was solely for directly applying the modelling recurrence threshold derived with the training set to the testing set.

Pre-selected gene clusters

To start the analysis, we chose 11 genes: *ALK, BRAF, EGFR, MET, NTRK, RAS, RET, ROS1, TP53, PDCD1*, and *CTLA4*. These were chosen because the first eight genes have been intensively studied and demonstrated to be drivers for lung cancers and their mutation status was used to guide targeted therapy and the last two were for immunotherapy. However, it should be noted that the procedure we report is very general and can be applied to any other genes and clusters. For a given gene, the literature and online information were used to select the cluster members. For example, the *ALK* cluster consisted of fusion partners cataloged in Ou *et. al*.[12] and some other genes from the String database and the gene card description, with 107 genes finally pre-selected for the *ALK* cluster. Table 1 lists the cluster members. Each cluster was analyzed independently using the same method. Note that the cluster members can be changed in the future when more insights about an important gene seed are added.

Gene cluster expression index (GCEI)

The goal was to assign samples with a binary index for a given gene cluster, called the *gene cluster expression index* (GCEI). This process comprises two steps: (1) Determination of the expression index of each member gene, such that a GCEI of 1 represents a higher recurrence risk stratified by the expression, or 0 otherwise; (2) Determination of the percentage of genes with a GCEI of 1 for each patient and labeling the patient as abnormal with a GCEI of 1 if there are too many abnormal members in the cluster. Both steps involve univariate prediction modeling via receiver operating characteristic (ROC) curve analysis.

Univariate modeling with the ROC curve and setting the cutoff value

The ROC curve is a basic technique in medical diagnostic test evaluation.[13] It is used for univariate modeling. Given a training set with a binary index vector, say recurrence, and a predicting vector, say expression vector of a gene, we can sort the training samples by the predictor values in increasing order, then by assuming a cutoff value that goes from the minimal to the maximal value with a fixed step size, each sample can be labeled as a binary prediction based on the cutoff. The prediction and the index (truth) give rise to a confusion matrix, such that a false positive rate (FPR) and true positive rate (TPR) are computed. The ROC curve is then plotted on a unit box with the FPR as the x-axis and TPR as the y-axis, as shown in Figure 1. The perfect prediction is at the top-left corner as $(FPR, TPR) = (0, 1)$; and therefore, along the curve from left to right, we can find the point closest to the corner (0, 1). This point is the optimal decision point leveraging both the specificity (1-FPR) and sensitivity (TPR) and the corresponding cutoff is thus set. The area of the bottom region is called the area under the curve (AUC), with values ranging from 0.5 to 1 (note: for a predictor with an AUC in between 0 and 0.5, a reversal with the 0-predictor flips the AUC to be above 0.5).

Determination of single gene expression abnormality concerning recurrence

For a given cluster, as listed in Table 1, for each member gene, we used its expression to predict recurrence and draw an ROC to obtain its optimal cutoff, which we then used to determine a sample expression status: normal or abnormal. Here, given a member gene g, we let T_g be the chosen cutoff, then the training samples were divided into two populations: one greater than or equal to T_g, the other less than. Now for each population, a recurrence percentage was computed, denoted as P_{above}, P_{below}, respectively.

We let $P_\delta = |P_{above} - P_{below}|$, which represents the prediction power of gene g by using its expression to stratify patients. In addition, if $P_{above} > P_{below}$, then g is considered over-expressed and showing a higher recurrence risk, or else it is under-expressed. Next, we set a significance level $T_{diff} = 5\%$ (note that this value was only for demonstration purposes, it can be set as another value based on a particular application), and the gene g was considered *significant* if $P_\delta \geq T_{diff}$. With respect to g, samples were 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 considered *abnormal*. In this way, all the member genes were labeled as 0 (normal) or 1 (abnormal, either up or down).

Cluster member voting and the GCEI

Next, we calculated the percentage of *abnormal* gene members for each sample to form a new feature vector. We plotted the ROC using the abnormal percentage to predict recurrence and denoted the chosen cutoff as T_p. We labeled the sample as 1 if the abnormal percentage is greater than or equal to T_p, or as 0 otherwise. This characteristic index is called the *gene cluster expression index* (GCEI). A GCEI of 1 represents an *abnormal* expression for the cluster, while a GCEI of 0 represents a *normal* expression. GCEI thus represents the abnormality of a gene cluster within which the percentage of abnormal member genes with

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, CEBPZ, 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, PRKAR1A, 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, TUBB, 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, ARHGEF2, 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, PRKAR1A, 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, TP53I11, TP53I13, 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*

GCEI = 1 is beyond T_p.

Combined GCEI (cGCEI)

A combined GCEI was defined first by concatenating a single cluster GCEI into binary string and second by counting the number of 1's in the string. This thus represents a summary of the expression abnormality of selected gene clusters. Here, for the targeted therapy genes, we fixed the ordered list of genes (*ALK, BRAF, EGFR, MET, NTRK, RAS, RET, ROS1, TP53*), and concatenated the corresponding GCEI of each cluster to obtain a binary string of 9 bits; for example, 000000000 represents that all 9 gene clusters were normally expressed, 100000000 represents that only the first *ALK* cluster was abnormally expressed and the rest 8 were normal, 111111111 represents that all 9 clusters were abnormally expressed, and so on. The 9-bit GCEI classified lung cancers into $2^9 = 512$ subtypes. For the immunotherapy gene couple (*CTLA4, PDCD1*), GCEI was a two-digit string with four combinations: 00, 01, 10, 11, representing that none, *CTLA4* only, *PDCD1* only, or both *CTLA4* and *PDCD1* clusters were abnormally expressed, respectively.

In practice, for the 9-bit GCEI string, since it would be difficult to accumulate enough patient cases for most of the 512 subtypes, we collapsed the 512 subtypes into only 10 super-subtypes by counting the number of digits that were 1 in the string, whereby patients were grouped into 10 subtypes with aggregated GCEIs of 0, 1, 2, 3, …, 9, respectively, denoted as *cGCEI*, with each cGCEI value representing how many gene clusters were abnormal among the nine clusters. To simplify it further, after analyzing the recurrence risk profiles of the 10 subtypes, we found that they could be further divided into two groups, denoted by the binary variable DGCntGT5, where the group of DGCntGT5 = 1 included all subtypes with cGCEI from 6 to 9, namely, all samples with at least 6 abnormal clusters; and DGCntGT5 = 0, which included all subtypes with cGCEI from 0 to 5, i.e. all samples with at most 5 abnormal clusters.

Recurrence and survival concerning GCEI status

Recurrence and survival were assessed with respect to the subgroups stratified by a single GCEI, a combinatory cGCEI, or by *DGCntGT5*. Given a binary index, this classified the samples into two subgroups with an index of 1 or 0, respectively. Recurrence/Survival risk was defined as the percentage of recurred/dead pa-

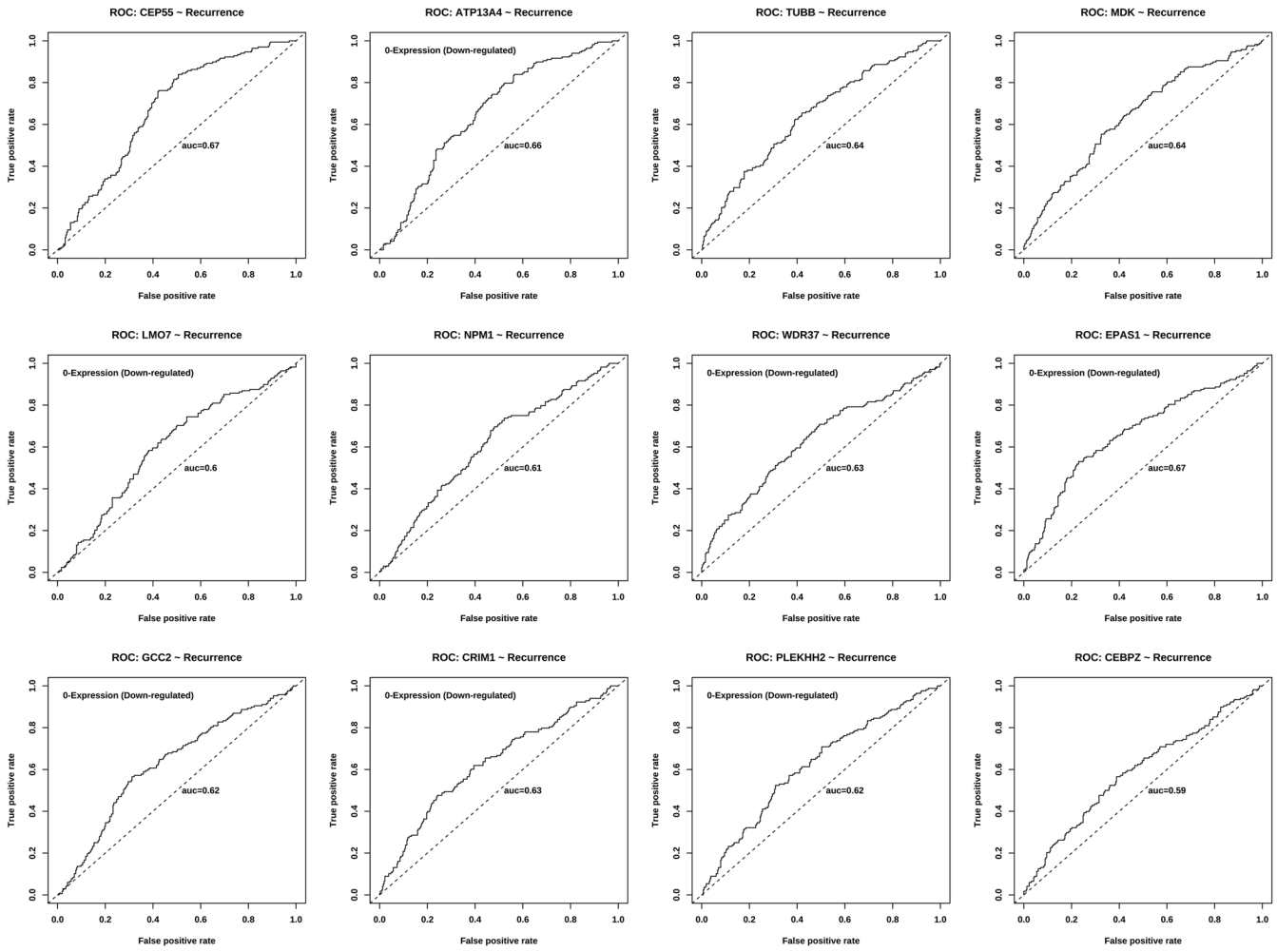

Fig. 1. Univariate ROCs of the top 12 genes in the ALK cluster in the decreasing order of P_δ. Inverted expression value (0 − *Expression*) was used to plot the ROC for the downregulated genes, similarly hereinafter. ALK, anaplastic lymphoma kinase; AUC, area under the curve; ROC, receiver operating characteristic.

tients within each subgroup.

Data analysis and software

The data analysis and plots were mostly performed using by R scripts in RStudio 2022.07.1 with R version 4.0.5 on the Mac platform with OS version darwin17.0. The ROC analysis was based on *prediction* and *performance* in the R package *ROCR*, where *performance* is a perfect function to obtain almost all the evaluation results of a prediction model, such as FPR, TPR, and AUC. Quartiles were calculated with the R function *quantile*.

Results

Univariate models of the ALK cluster members

Among the 107 pre-selected members in the *ALK* cluster, 72 *abnormal* genes had $P_\delta \geq 5\%$, accounting for 67% of the members, and the other 35 *normal* genes had $P_\delta < 5\%$. The corresponding AUCs, FPRs, TPRs, threshold T_g, and population risks of the *abnormal* and the *normal* genes are listed in Tables 2 and 3, respectively. As shown in Table 2, 33 genes were over-expressed for a 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*; while 39 were under-expressed for a 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*. Note that for the under-expressed genes, an inverted expression vector, namely 0-expression, should be used as a predictor to plot the ROC correctly.

For demonstration purposes, the ROC curves of the top 12 genes in decreasing order of P_δ are shown in Figure 1. The highest one in the first row of Table 2 is *CEP55*. Here, for the chosen cutoff $T_g = -0.0076$, a patient with a *CEP55* expression $\geq (-0.0076)$ has a recurrence risk of $P_{above} = 49.03\%$, while a patient with a *CEP55* expression $< (-0.0076)$ has a recurrence risk of $P_{below} = 18.39\%$, and hence the difference between the two is $P_\delta = 30.64\%$. *CEP55* is considered over-expressed 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*

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_δ (%)	Status
CEP55	0.68	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	0.64	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	0.6	0.39	0.58	0.0509	22.99	42.37	19.38	down
NPM1	0.61	0.46	0.68	−0.0053	43.63	24.66	18.97	up
WDR37	0.63	0.43	0.63	0.0118	24.54	43.23	18.69	down
EPAS1	0.67	0.31	0.58	0.0882	23.12	41.42	18.3	down
GCC2	0.62	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	0.62	0.37	0.57	0.079	24.73	41	16.27	down
CEBPZ	0.59	0.39	0.57	0.0299	43.58	27.65	15.93	up
TFG	0.58	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	0.59	0.41	0.58	0.005	27.04	42.17	15.13	down
LYPD1	0.6	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	0.59	0.5	0.66	−0.013	41.2	26.98	14.22	up
SMPD1	0.57	0.39	0.54	0.0281	26.87	40.57	13.7	down
LPIN1	0.56	0.36	0.5	0.053	42.86	29.37	13.49	up
MYT1L	0.57	0.47	0.62	−0.0011	41.27	27.83	13.44	up
YAP1	0.59	0.37	0.55	0.0335	27.52	40.91	13.39	down
MPRIP	0.59	0.46	0.63	0.0038	27.65	40.75	13.1	down
WNK3	0.54	0.35	0.48	0.0757	42.55	29.93	12.62	up
TANC1	0.62	0.32	0.57	0.0631	26.74	39.35	12.61	down
SEC31A	0.6	0.42	0.57	−0.0025	29.17	41.74	12.57	down
PRKAR1A	0.59	0.44	0.6	0.0074	27.98	40.53	12.55	down
TNIP2	0.56	0.46	0.6	7.00E-04	40.98	28.57	12.41	up
C12ORF75	0.58	0.48	0.62	−0.027	40.62	28.32	12.3	up
TPM4	0.56	0.5	0.63	−0.0131	40.46	28.18	12.28	up
TTC27	0.57	0.41	0.54	0.022	41.55	29.28	12.27	up
CYBRD1	0.59	0.48	0.61	0	28.51	40.55	12.04	down
SPTBN1	0.57	0.42	0.55	0.0219	28.14	39.58	11.44	down
ALKAL2	0.6	0.35	0.52	0.1104	28.04	39.25	11.21	down
SOS1	0.55	0.46	0.58	0.0069	40.42	29.34	11.08	up
ADAM17	0.57	0.46	0.58	−0.004	40.08	29.58	10.5	up
TSPYL6	0.56	0.48	0.6	−0.0022	39.84	29.44	10.4	up
KLC1	0.52	0.32	0.42	0.0302	41.52	31.19	10.33	up
PPFIBP1	0.55	0.46	0.57	−0.0145	39.92	29.92	10	up
SPECC1	0.57	0.47	0.58	−0.0119	39.75	29.83	9.92	up

(continued)

Table 2. *(continued)*

GENE	AUC	FPR	TPR	T_g	P_{above} (%)	P_{below} (%)	P_δ (%)	Status
WDPCP	0.56	0.44	0.55	0.0163	29.49	39.25	9.76	down
SLMAP	0.58	0.39	0.53	0.0339	29.02	38.75	9.73	down
CLIP4	0.58	0.34	0.5	0.0655	28.96	38.46	9.5	down
SLC16A7	0.58	0.44	0.54	0.0041	30.24	39.74	9.5	down
SWAP70	0.56	0.48	0.59	0.0024	29.73	39.23	9.5	down
LIMD1	0.57	0.49	0.58	0.0054	29.82	39.02	9.2	down
FRS2	0.52	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	0.52	0.34	0.43	0.0616	40.22	31.68	8.54	up
FBN1	0.53	0.46	0.55	−0.0041	39.15	30.77	8.38	up
SOCS5	0.56	0.36	0.47	0.0388	29.73	38.05	8.32	down
PLEKHA7	0.56	0.53	0.64	−0.0759	31.77	39.89	8.12	down
EIF2AK3	0.53	0.45	0.54	0	31.08	38.96	7.88	down
THADA	0.53	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	0.54	0.42	0.54	0.0205	30.88	38.11	7.23	down
KIF5B	0.53	0.44	0.55	0.0113	31.05	38.02	6.97	down
PHACTR1	0.57	0.41	0.52	0.0754	30.69	37.54	6.85	down
CLIP1	0.51	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	0.54	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	0.54	0.54	0.67	−0.0186	32.26	38.42	6.16	down
CLTC	0.51	0.39	0.45	0.0321	38.38	32.39	5.99	up
FBXO36	0.52	0.51	0.58	−0.0086	37.6	31.7	5.9	up
PTN	0.54	0.44	0.56	0.0341	31.63	37.45	5.82	down
PICALM	0.55	0.39	0.46	0.0318	31.35	37.04	5.69	down
FUT8	0.53	0.47	0.53	0.0417	37.55	32.24	5.31	up
VKORC1L1	0.53	0.47	0.54	0.0027	32.23	37.5	5.27	down
HIP1	0.54	0.44	0.54	0.0348	31.86	37.05	5.19	down
ITGAV	0.51	0.51	0.57	9.00E-04	37.25	32.16	5.09	up

ALK, anaplastic lymphoma kinase; AUC, area under the curve; FPR, false positive rate; TPR, true positive rate.

was found to be a fusion partner of *ALK* and a high *CEP55* expression was reported to be associated with a poor prognosis.[14,15] The second gene we consider is *ATP13A4*, which was under-expressed with P_{above} = 24.19%, P_{below} = 46.15%, for a difference of P_δ = 21.96%. *ATP13A4*, called ATPase 13A4, may enable ATPase-coupled cation transmembrane transporter activity and may be involved in cellular calcium ion homeostasis. In one lung cancer case study,[16] a 53-year-old metastatic Stage IV patient harboring *ATP13A4-ALK* and two other *ALK*-fusions *COX7A2L-ALK* and *LINC01210-ALK* underwent first-line crizotinib therapy, which showed 12 months of Progress Free Survival/Partial Remission (PFS/PR), then a new *SLCO2A1-ALK* fusion led to resistance. Afterward, second-line ceritinib therapy was applied and resulted in a further 8 months of PFS, and the NGS results demonstrated the loss of *ATP13A4-ALK* and *SLCO2A1-ALK*. Interestingly, the *ALK* expression itself was normal and only showed a difference of P_δ = 2.02%.

Note that the results for the remaining 10 gene clusters in this study are presented in the Supplementary File 1.

Cluster member voting models

Next, for the training sample, we calculated the percentage of *abnormal* members for each cluster. Again, we plotted the ROC but

Table 3. AUCs and recurrence risks of 35 *normal ALK* genes with P_δ < 5%

GENE	AUC	FPR	TPR	T_g	P_{above} (%)	P_{below} (%)	P_δ (%)	Status
TOGARAM2	0.56	0.51	0.62	−0.0095	32.37	37.34	4.97	normal
BCL11A	0.52	0.4	0.51	0.0391	32.37	37.34	4.97	normal
ATAD2B	0.51	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	0.55	0.38	0.49	0.0701	31.98	36.45	4.47	normal
AKAP8L	0.51	0.49	0.56	−0.0046	32.8	37.07	4.27	normal
CUX1	0.54	0.4	0.49	0.0396	32.28	36.52	4.24	normal
NCOA1	0.52	0.42	0.51	0.0344	32.26	36.49	4.23	normal
PLEKHM2	0.5	0.48	0.52	−0.0047	36.97	32.79	4.18	normal
SORCS1	0.51	0.54	0.59	−0.0075	33.08	36.99	3.91	normal
SMPDL3B	0.53	0.5	0.6	−0.0554	33.33	37.17	3.84	normal
CMTR1	0.51	0.47	0.51	0.0067	32.89	36.58	3.69	normal
MAPK1	0.52	0.49	0.52	−0.005	36.67	33.06	3.61	normal
TCF12	0.54	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	0.5	0.46	0.5	0.0021	36.68	33.2	3.48	normal
SMPD2	0.51	0.38	0.39	0.0535	36.63	33.87	2.76	normal
MAPK3	0.54	0.42	0.48	0.0142	33.33	36	2.67	normal
DCTN1	0.5	0.43	0.46	0.0209	36.32	33.7	2.62	normal
DCHS1	0.53	0.41	0.49	0.0531	36.47	33.97	2.5	normal
SMPD3	0.52	0.46	0.48	0.0141	36.16	33.72	2.44	normal
SRBD1	0.52	0.49	0.53	0.0003	33.75	35.95	2.2	normal
TPR	0.51	0.46	0.52	0.0063	33.76	35.89	2.13	normal
ALK	0.52	0.54	0.61	−0.0176	35.71	33.66	2.05	normal
TACR1	0.52	0.55	0.6	−0.0096	33.96	35.98	2.02	normal
VIT	0.53	0.42	0.52	0.015	33.67	35.66	1.99	normal
DYSF	0.51	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	0.51	0.41	0.45	0.0211	33.93	35.66	1.73	normal
CDK15	0.52	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	0.51	0.49	0.48	0.0012	34.32	35.37	1.05	normal
RANBP2	0.54	0.31	0.44	0.051	34.25	35.22	0.97	normal
C9ORF3	0.51	0.43	0.48	0.0195	35.35	34.46	0.89	normal

ALK, anaplastic lymphoma kinase; AUC, area under the curve; FPR, false positive rate; TPR, true positive rate.

with an abnormal percentage as a new recurrence predictor. The ROC curves are presented in Figure 2. For each ROC curve, the horizontal and vertical dashed lines mark the point on the curve that is the closest to the top-left corner (0, 1), and the corresponding FPR and TPR are shown near each dashed line. The AUC is also shown. Taking *ALK* as an example, the closest point to the top-left corner is (0.32, 0.73), indicating that the specificity (1-FPR) was 68% and the sensitivity (TPR) was 73%, and the AUC was 0.763. The corresponding cutoff was set as the voting threshold for the *ALK* cluster. Table 4 lists the corresponding AUCs, FPRs, TPRs, threshold T_p, P_{above}, and P_{below}. In summary, across the 11 studied clusters, the recurrence risk of the abnormal group

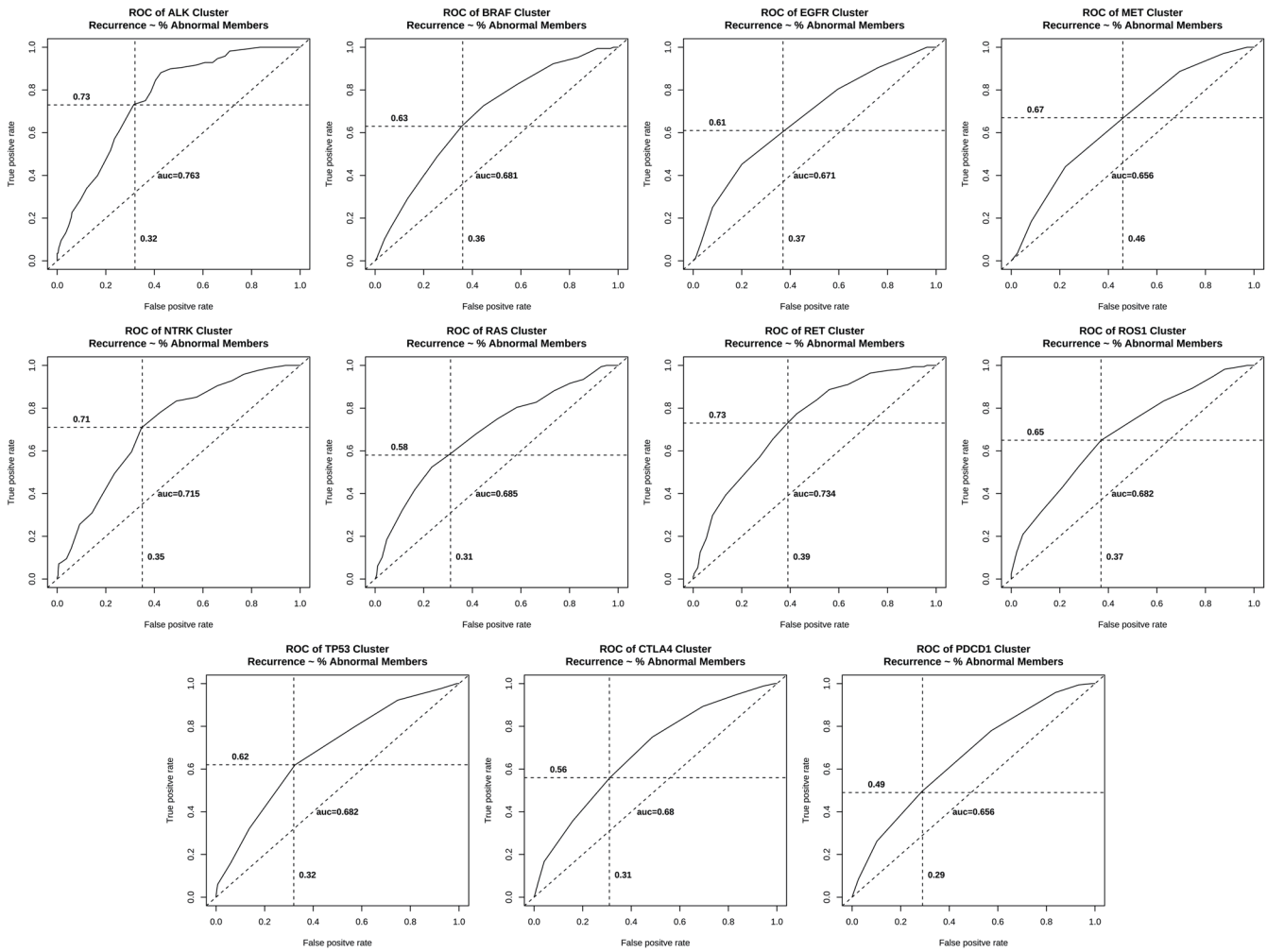

Fig. 2. Univariate ROCs of 11 clusters. The percentage of the *abnormal* members in each cluster was used as a recurrence predictor. For each ROC curve, the horizontal and vertical dashed lines mark the point on the curve that is the closest to the top-left corner (0, 1), and the corresponding FPR and TPR are shown near each dashed line. The AUC is also shown. Taking *ALK* as an example, the closest point to the top-left corner is (0.32, 0.73), indicating that the specificity (1-FPR) is 68% and sensitivity (TPR) is 73%, and the AUC is 0.763. The corresponding cutoff is set as the voting threshold for the *ALK* cluster. ALK, anaplastic lymphoma kinase; AUC, area under the curve; FPR, false positive rate; ROC, receiver operating characteristic; TPR, true positive rate.

Table 4. AUC, TPR, FPR, threshold T_p and recurrence risks for 11 clusters

SEED	AUC	T_p (%)	P_{above} (%)	P_{below} (%)	P_δ (%)	FPR	TPR	ACC	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

P_{above} is the recurrence risk of the patients with the corresponding abnormal cluster members ≥ T_p%, and P_{below} is the opposite group with < T_p%. ACC, accuracy; AUC, area under the curve; FPR, false positive rate; PPV, positive prediction value; TPR, true positive rate.

Table 5. Recurrence percentages of lung cancers in different stage groups flagged by the GCEI

Subpopulation	All (Stages I–IV)		Stage I		Stages 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

Group risk of *GCEI* = 1 is typically 120% to 300% that of the corresponding group of *GCEI* = 0. GCEI, gene cluster expression index.

(of all pathological stages) ranged from 174% (*PDCD1*) to 320% (*ALK*) of the corresponding normal group.

Recurrence and survival analysis

In the above, lung cancers were labeled as *normal* (*GCEI* = 0) or *abnormal* (*GCEI* = 1) using a given cluster GCEI 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 combinatory GCEI, the recurrence risk, defined as the percentage of recurred patients, was calculated based on the GCEI status of patients at different pathological stages, such as Stage I, Stages II–V, and all stages. Table 5 lists the recurrence risks for the subpopulations labeled by the atomic GCEI indicators and *DGCntGT*5 indicator. It can be seen that the *ALK* cluster gave the largest risk ratio for the lung cancer group with *GCEI* = 1 over *GCEI* = 0 for the 3 stage groups, with 320%, 332%, 188% for all stages, Stage I, and Stages II–IV, respectively. As for the minimal ratio, *PDCD1* gave 174% for all stages, MET 169% for Stage I, and *EGFR* 109% for Stages II–IV. On average, the risk ratios of the group with *GCEI* = 1 over *GCEI* = 0 were 222%, 247%, 134% for all stages, Stage I, Stages II–IV, respectively, indicating that on average the recurrence risk of patients with an abnormally expressed cluster was more than double that of the normal counterpart for all stages or Stage I, while even for the late Stages II–IV, the risk was still increased by 34%. This demonstrates the power of recurrence risk stratification with the GCEI.

In addition, the recurrence risks of the 10 subgroups of *cGCEI* = 0, 1, 2, …, 9, as defined by counting the number of 1's in the binary string of the ordered list (*ALK, BRAF, EGFR, MET, NTRK, RAS, RET, ROS1, TP53*), are listed in Table 6. It can be seen that the risk increased along with the cGCEI values, indicating that the

Table 6. Number of none-recurred and recurred cases and the recurrence risks of cGCEI derived from 9-digit string signatures (only evaluated for all stages)

cGCEI	Exemplary signatures	None-recurred	Recurred	Total	Recurrence (%)
0	000000000	53	4	57	7.02
1	100000000,000000001	61	11	72	15.28
2	110000000,000000011	39	10	49	20.41
3	111000000,000000111	33	7	40	17.5
4	111100000,000001111	30	12	42	28.57
5	111110000,000011111	21	11	32	34.38
6	111111000,000111111	24	31	55	56.36
7	111111100,001111111	24	26	50	52
8	111111110,011111111	20	32	52	61.54
9	111111111	9	24	33	72.73

cGCEI, combinatory gene cluster expression index.

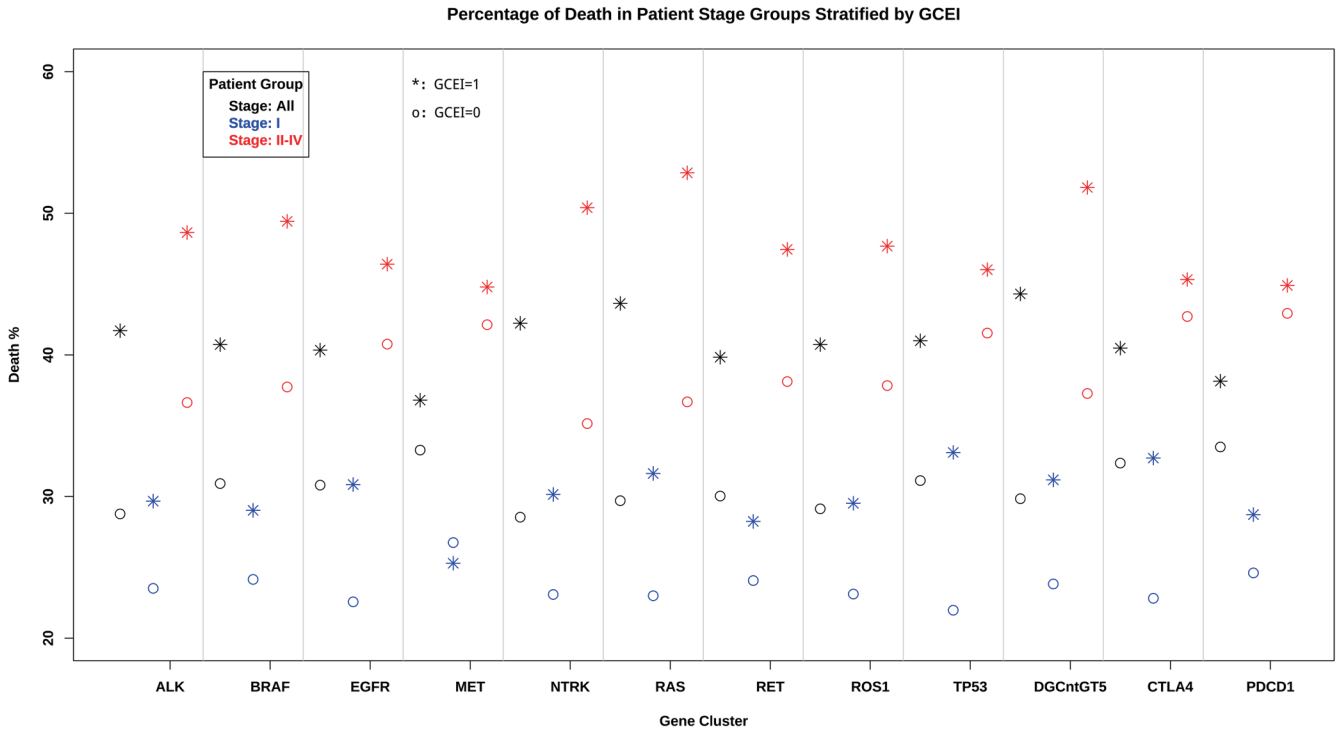

Fig. 3. Percentage of deaths of the lung cancer subgroups (GCEI = 0, 1) within different stages or all stages. For each gene cluster expression index or combinatory DCGCntGT5, 3 vertical pairs are plotted with different colors (All stages: black, Stage I: blue, Stages II–IV: red). Each pair consists of GCEI = 0 (circle) and GCEI = 1 (*). The vertical gap from the circle to * shows the increased percentage of GCEI = 1 compared to GCEI = 0. In summary, for Stage I, the median increase is 6.74% with a maximum of 11.14% (*TP53*, blue); for Stages II–IV, the median increase is 9.60% with a maximum of 16.18% (*RAS*, red); for all stages, the median increase is 9.85% with a maximum of 14.46% (DCGCntGT5, black). GCEI, gene cluster expression index.

more abnormal clusters there were, the higher the risk. For *cGCEI* = 0, where all the clusters were normally expressed, the recurrence risk was merely 7.02%; whereas, when there was one and only one abnormal cluster (*cGCEI* = 1), the risk was more than doubled to 15.28%; and it then increased to 20.41% for *cGCEI* = 2. However, a hiccup then occurred in the trend, whereby the risk went down to 17.50% for *cGCEI* = 3, which might be due to the data size. The risk then again kept increasing along with the *cGCEI*, albeit also with a hiccup. After *cGCEI* 6, the risk was beyond 56.36% until it hits an astonishing 72.73% for the group of patients with *cGCEI* = 9, where all 9 clusters showed abnormal expressions. This also shows the rationale of why we defined a new combined GCEI based on DGCntGT5 to collapse the 10 subtypes into only two.

Moreover, population survival analysis was applied to the subgroups of GCEI = 0 or 1 for Stage I, Stages II–IV, and all stages. Figure 3 shows the percentage of death of each subgroup within the different stages or all stages. In summary, for Stage I, the median increase was 6.74% with a maximum of 11.14% (*TP53*, Stage I); for Stages II–IV, the median increase was 9.60% with a maximum of 16.18% (*RAS*, Stages II–IV); for all stages, the median increase was 9.85% with a maximum of 14.46% (DCGCntGT5, all stages). However, an exception was noted for *MET* (Stage I, blue), where the percentages of death for GCEI = 0, 1 were 26.75% and 25.28%, respectively, and both subgroups had similar death risks. In conclusion, the survival population results indicated a modest survival risk difference based on GCEI status as defined by recurrence. The same procedure could be applied here by targeting the OS. Since the OS and RS are correlated but not the same, optimizing one of them may only guarantee a suboptimal risk profile for the other.

Validation

There were 703 patients in the validation set combined from the GSE37745, GSE41271, GSE50081, and GSE74777 data sets, within which there were 272 recurrences (39%) (vs. 35% in the training set), the average patient age was 66 years old (vs. 61 in the training set), and there were 278 females (40%) (vs. 31% in the training set), and 397 Stage I patients (49%) (vs. 64% in the training set). Table 7 shows the recurrence risks of GCEI = 0 vs. GCEI = 1, where the GCEI was determined based on the thresholds in the training phase. The average risk increase was 11.12%, and the maximum was 35.5% (*RAS*). This is a modest validation result compared with the training risk profiles (Table 5). Note that *CTLA4* showed a risk reversal while *ALK* and *NTRK* showed barely different risks between the two groups. These modest results might mainly be due to several reasons; first, the data were from different microarray chips: both gene expression omnibus sets in the training set came from *Affymetrix Human Genome U133 Plus 2.0 Array*, while in the validation set, although GSE37745 and GSE50081 were from the same chip, GSE41271 came out of RnaSeq of the *Illumina HumanWG-6 v3.0 expression beadchip* and GSE74777 was from the *Affymetrix Human Transcriptome (HT) Array 2.0*; second, the different patient profiles as stated in the above.

Comparison with conventional methods

Up to now, we have demonstrated that classification using the

Table 7. Recurrence risks of the normal group (GCEI = 0) vs. abnormal group (GCEI = 1) computed from the validation set using the same thresholds from the training phase

Cluster	% Recurrence (GCEI = 0)	% Recurrence (GCEI = 1)	% Increase of GCEI = 1 from 0
ALK	37.73	38.81	2.86
BRAF	36.36	41.13	13.12
CTLA4	39.51	35.92	−9.08
EGFR	36.36	40	10.01
MET	36.33	39.83	9.63
NTRK	38.01	38.38	0.97
PDCD1	37.06	40.27	8.66
RAS	34.37	46.57	35.5
RET	35.38	42.74	20.8
ROS1	36.39	40.49	11.27
TP53	35.42	42.07	18.77
DGCntGT5	36.99	41.05	10.98

cGCEI, combinatory gene cluster expression index.

GCEI could stratify lung cancers into groups with dramatically different recurrence risk profiles. Next, we compared the GCEI with other conventional characteristics, namely stage, node (N), and T of TNM, and prognosis of recurrence and survival via correlation analysis. The average correlation coefficients of GCEIs across these five clinical variables were: DGCntGT5 (0.39), *ALK* (0.36), *NTRK* (0.34), *BRAF* (0.32), *RAS* (0.32), *RET* (0.31), *EGFR* (0.27), *ROS1* (0.27), *MET* (0.23), *TP53* (0.22), *PDCD1* (0.12), *CTLA4* (0.09).

On the other hand, the average correlation coefficients of the clinical variables across 12 GCEIs were: recurrence (0.25), survival (0.23), stage (0.30), N (0.28), T (0.30), indicating that the GCEIs were modestly correlated with clinically important pathological variables and prognosis. The advantages of the GCEI include that the method involves molecular profiling and has the potential to guide targeted therapy and immunotherapy for lung cancers.

Discussion

The goal of this research was not to predict recurrence risk but to provide a novel approach to classify lung cancers based on gene cluster expression profiles. The original intention was to complement the current personalized approach with DNA-based classifications. Recurrence risk was used as a convenient guiding prognostic objective here to derive GCEIs, but this method can be applied to other objectives too, such as prognosis of survival, treatment response, and to clinically important pathological variables, such as stage, metastatic node count, or distance metastasis. As far as personalized medicine in lung cancer is concerned, although DNA-based tests have been successfully 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 the tested patients did not have their results before first-line treatment and fewer than 20% of tested patients had their results for all 4 driver mutations (*ALK, EGFR, ROS1, BRAF*), and *PD-L1* before first-line treatment. Moreover, although the turnaround time improved from the year 2017 to 2019, not all patients who tested positive for driver mutations received targeted therapy in the first-line setting.[17] Therefore it shows there is an unmet need for a large proportion of lung cancer patients who are not qualified for personalized medicines following the current paradigm. We can imagine that an RNA expression network (a cluster) centered around an important gene is disturbed not just by a particular DNA mutation, which might be just one thread in the whole picture, but by a lot of other factors. The abnormality of an RNA expression network is then gauged by the percentage of abnormally expressed nodes (cluster members). It is only after the percentage of abnormal nodes goes beyond a threshold is the collapse of the whole network triggered. GCEI was introduced here to label whether an RNA expression network looks normal or abnormal concerning the guiding objective, such as recurrence in the current study. When an expression network centered around an important gene for which there are available drug targets looks abnormal, the same drugs might come to the rescue and adjust the network to look more normal. Hence we propose that the patient group of abnormal status with *GCEI* = 1 who cannot access the same targeted therapy and immunotherapy might benefit from the same treatment. Evidence has already emerged in a study called the WINTHER trial (NCT01856296),[18] which 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). This study showed that transcriptome profiling is as useful as DNA tests for improving therapy recommendations and patient outcomes.

On the other front, novel RNA drugs have emerged and generated more and more enthusiasm in the pursuit of new lung cancer treatment.[19] Although the expression of a single targeted gene can be relatively easily evaluated, it will be important to know how the RNA expression of a gene network centered around the targeted gene is disturbed and how the disturbance is related to the clinical outcome. Hence it will be a routine requirement to measure whether a given RNA expression network is normal or abnormal clinically. The GCEI is a simple attempt to address this coming revolution.

Conclusions

Gene cluster expression index can be used to classify lung cancers with dramatically different recurrence risks and the recurrence risk (percentage) of the patient group with index 1 is typically 20% to 200% higher than the group with index 0. We expect that the higher risk group of index 1 may also be suitable for the corresponding targeted therapy or immunotherapy. Therefore, it may be used to guide targeted therapy or immunotherapy when the conventional companion tests give no recommendation. Nevertheless, this should be validated by clinical trials before it is applied in the clinical practice.

Acknowledgments

We thank UT Southwestern Medical Center for the curated data sets via https://lce.biohpc.swmed.edu/lungcancer/dataset.php.

Funding

All funding of the study is supported by the R&D department of Shenzhen Luwei (Biomanifold) Biotechnology Limited, Shenzhen, China.

Conflict of interest

Aibing Rao is a full-time employee of Shenzhen Luwei (Biomanifold) Biotechnology Limited, Shenzhen, China.

Ethical statement

The raw data sets were downloaded from the open web portal Lung Cancer Explorer (LCE). This was an *in-silico* research study and ethics approval was not applicable.

Data sharing statement

The post-processed data sets used in support of the findings of this study are available from the corresponding author at aibing.rao@enlightendx.com upon request.

References

[1] Tang H, Wang S, Xiao G, Schiller J, Papadimitrakopoulou V, Minna J, *et al*. Comprehensive evaluation of published gene expression prognostic signatures for biomarker-based lung cancer clinical studies. Ann Oncol 2017;28(4):733–740. doi:10.1093/annonc/mdw683, PMID:28200038.

[2] Woodard GA, Wang SX, Kratz JR, Zoon-Besselink CT, Chiang CY, Gubens MA, *et al*. Adjuvant Chemotherapy Guided by Molecular Profiling and Improved Outcomes in Early Stage, Non-Small-Cell Lung Cancer. Clin Lung Cancer 2018;19(1):58–64. doi:10.1016/j.cllc.2017.05.015, PMID:28645632.

[3] Bueno R, Richards WG, Harpole DH, Ballman KV, Tsao MS, Chen Z, *et al*. Multi-Institutional Prospective Validation of Prognostic mRNA Signatures in Early Stage Squamous Lung Cancer (Alliance). J Thorac Oncol 2020;15(11):1748–1757. doi:10.1016/j.jtho.2020.07.005, PMID:32717408.

[4] Luo Y, Deng X, Que J, Li Z, Xie W, Dai G, *et al*. Cell Trajectory-Related Genes of Lung Adenocarcinoma Predict Tumor Immune Microenvironment and Prognosis of Patients. Front Oncol 2022;12:911401. doi:10.3389/fonc.2022.911401, PMID:35924143.

[5] Yu J, Li G, Tian Y, Huo S. Establishment of a Lymph Node Metastasis-Associated Prognostic Signature for Lung Adenocarcinoma. Genet Res (Camb) 2023;2023:6585109. doi:10.1155/2023/6585109, PMID:36793937.

[6] Nagl L, Pall G, Wolf D, Pircher A, Horvath L. Molecular profiling in lung cancer. memo 2022;15:201–205. doi:10.1007/s12254-022-00824-7.

[7] Buzdin A, Sorokin M, Garazha A, Glusker A, Aleshin A, Poddubskaya E, *et al*. RNA sequencing for research and diagnostics in clinical oncology. Semin Cancer Biol 2020;60:311–323. doi:10.1016/j.semcancer.2019.07.010, PMID:31412295.

[8] Nagy Á, Pongor LS, Szabó A, Santarpia M, Győrffy B. KRAS driven expression signature has prognostic power superior to mutation status in non-small cell lung cancer. Int J Cancer 2017;140(4):930–937. doi:10.1002/ijc.30509, PMID:27859136.

[9] Rousseaux S, Debernardi A, Jacquiau B, Vitte AL, Vesin A, Nagy-Mignotte H, *et al*. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Sci Transl Med 2013;5(186):186ra66. doi:10.1126/scitranslmed.3005723, PMID:23698379.

[10] Okayama H, Kohno T, Ishii Y, Shimada Y, Shiraishi K, Iwakawa R, *et al*. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. Cancer Res 2012;72(1):100–111. doi:10.1158/0008-5472.CAN-11-1403, PMID:22080568.

[11] Cai L, Lin S, Girard L, Zhou Y, Yang L, Ci B, *et al*. LCE: an open web portal to explore gene expression and clinical associations in lung cancer. Oncogene 2019;38(14):2551–2564. doi:10.1038/s41388-018-0588-2, PMID:30532070.

[12] Ou SI, Zhu VW, Nagasaka M. Catalog of 5′ Fusion Partners in ALK-positive NSCLC Circa 2020. JTO Clin Res Rep 2020;1(1):100015. doi:10.1016/j.jtocrr.2020.100015, PMID:34589917.

[13] Hajian-Tilaki K. Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation. Caspian J Intern Med 2013;4(2):627–635. PMID:24009950.

[14] Couëtoux du Tertre M, Marques M, Tremblay L, Bouchard N, Diaconescu R, Blais N, *et al*. Analysis of the Genomic Landscape in ALK+ NSCLC Patients Identifies Novel Aberrations Associated with Clinical Outcomes. Mol Cancer Ther 2019;18(9):1628–1636. doi:10.1158/1535-7163.MCT-19-0105, PMID:31243098.

[15] Jiang C, Zhang Y, Li Y, Lu J, Huang Q, Xu R, *et al*. High CEP55 expression is associated with poor prognosis in non-small-cell lung cancer. Onco Targets Ther 2018;11:4979–4990. doi:10.2147/OTT.S165750, PMID:30154666.

[16] Cai C, Long Y, Li Y, Huang M. 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;13:8313–8316. doi:10.2147/OTT.S258067, PMID:32903930.

[17] Nadler E, Vasudevan A, Wang Y, Ogale S. 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 Treat Res Commun 2022;31:100522. doi:10.1016/j.ctarc.2022.100522, PMID:35189530.

[18] Rodon J, Soria JC, Berger R, Miller WH, Rubin E, Kugel A, *et al*. Genomic and transcriptomic profiling expands precision cancer medicine: the WINTHER trial. Nat Med 2019;25(5):751–758. doi:10.1038/s41591-019-0424-4, PMID:31011205.

[19] Khan P, Siddiqui JA, Lakshmanan I, Ganti AK, Salgia R, Jain M, *et al*. RNA-based therapies: A cog in the wheel of lung cancer defense. Mol Cancer 2021;20(1):54. doi:10.1186/s12943-021-01338-2, PMID:33740988.