Evaluation of a KRAS gene expression signature in lung cancer

A. Lovše, J. Otoničar, R. Luštrik, L. Ausec, J. Kokošar, Genialis Inc., Boston MA, United States

Genialis

EACR23-0944

INTRODUCTION

Lung adenocarcinoma is a complex disease driven by multiple oncogenic drivers including KRAS mutations which occur in approximately 30% of tumors[1]. Previous clinical trials focusing on these patients have reported partial response rates which could not be informed by existing immunohistochemical and mutational biomarkers[2,3]. Despite sharing the same oncogenic mutations, there was significant heterogeneity in gene expression, suggesting diverse signaling patterns in responders[4].

Gene expression signatures have been reported in the literature that stratify KRAS-mutated lung cancers into phenotypic subgroups that are, in part, defined by a particular combination of genomic variants[4]. A subsequent study questioned the applicability of genetic surrogates to identify these tumor subtypes and instead derived concordant subtypes purely based on gene expression[5]. Both studies used unsupervised clustering of KRAS-mutant lung adenocarcinomas in The Cancer Genome Atlas (TCGA) data set and developed gene expression-based classifiers. The initial gene signature was evaluated in genotypically similar cohorts in treatment-naive and platinum-refractory KRAS-mutant lung adenocarcinomas[4].

In this study, the gene expression-based classifier is applied to a new data set to evaluate the model's robustness and transferability across demographically distinct clinical cohorts.

Methods

Classifier implementation

- Gene expression-based logistic regression classifier using 18-gene signature
- Trained on 69 RNA-seq samples from TCGA-LUAD that were used in the original study[4]

Cohort details

The cancer genome atlas lung adenocarcinoma cohort

- RNA-seq and somatic mutations data of 69 lung adenocarcinoma samples generated by the TCGA Research Network (TCGA-LUAD)
- Frozen samples taken from a cohort of roughly gender balanced (30M/39F), white (52/69), predominantly early stage (34 I, 16 II, 15 III, 3 IV, 1 NA), treatment naive patients

Independent lung adenocarcinoma cohort

- RNA-seq of 87 LUAD samples from Korean patients who underwent lobectomy (GEO accession: GSE40419)[6]
- Predominantly male (53M/34F), mostly early stage (55 I, 13 II, 13 III, 4 IV, 2 NA) patients, 18 with lymph node metastases
- This study focused on a subset of 15 KRAS-mutant samples, predominantly male (12M/3F) patients, mostly early stage (9 I, 1 II, 4 III, 0 IV, 1 NA), 4 with lymph node metastases

Bioinformatics processing

Expression values (TPM) and genetic variants for the TCGA cohort were obtained from Genomic Data Commons. For the independent cohort, gene expression was quantified with BBDuk-STAR-featureCounts-rnanorm pipeline[7], and variants were called from RNA-Seq data following GATK best practices[8]. Non-synonymous variants with clinical implications according to the ClinVar database[9] were kept for further analyses. Cluster comparisons were done in R (4.2.2) using DESeq2[10], ComplexHeatmap[11], ClusterProfiler[12] packages.

Prior Art

The original study by Skoulidis et al. (2015) defined three subgroups of KRAS mutated lung cancer tumors:

KC subgroup:

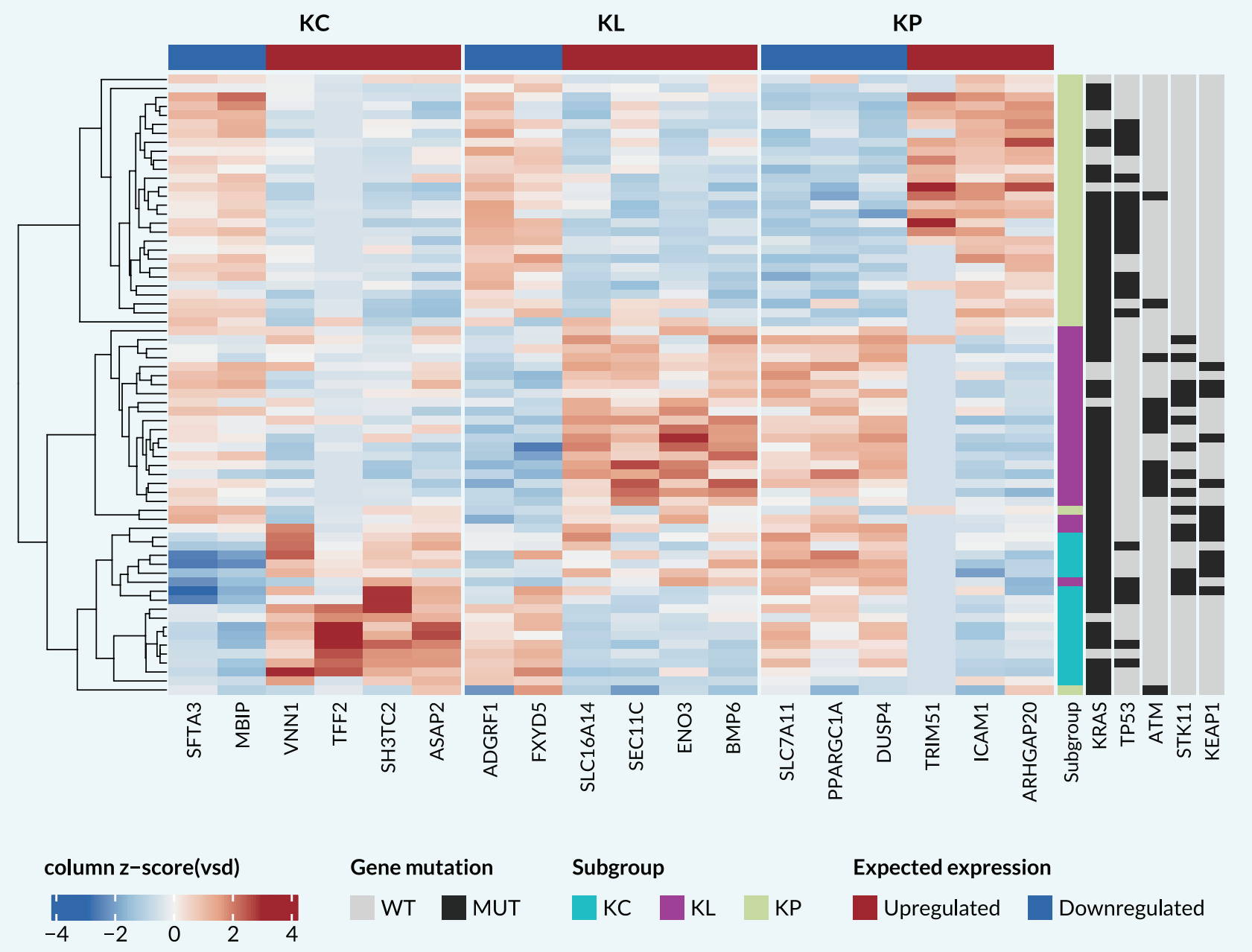
- Inactivation of CDKN2A/B coupled with low TTF1 expression.
- Enrichment of gene expression signatures reflecting both upper and lower GI neoplastic processes.
- Higher average expression of the embryonically restricted chromatin regulator HMGA2.
- Higher occurrence of invasive mucinous carcinomas.

KL subgroup:

- Co-mutations in STK11/LKB1.
- Functional inactivation of the LKB1-AMPK pathway.
- Activation of an NRF2-driven antioxidant and cytoprotective transcriptional program.

KP subgroup:

- Co-mutations in TP53.
- Higher overall mutational load.
- Elevated expression of immune checkpoint mediator/effector molecules, including PD-L1, PD-1, and CTLA-4.
- Improved relapse-free survival.



▲ Figure 1.

Reproduced gene expression and co-mutational landscape of 69 lung adenocarcinoma samples from TCGA used in the original study by Skoulidis et al. (2015).

The heatmap shows relative gene expression levels (standardized, log2 transformed and variance stabilized counts) of the 18 signature genes. Samples are in rows, while genes are in columns. Hierarchical clustering with optimal leaf reordering is done on rows (samples). Columns are grouped to include genes informative of a specific subset. The expected gene expression in each cluster, shown on the top of the columns, was determined based on the original article. Common co-mutations are reported on the right side of the heatmap. Wild-type (WT) is represented in gray, and nonsynonymous mutations (MUT) are represented in black. Other co-mutations mentioned in the original article are excluded from the plots, since they were not detected in the independent cohort. ***Note:*** *These results are our reproduction of Skoulidis et al. (2015) and serve as an independent validation of the findings in that study.*

Results and Discussions

Newly implemented model is similar to the original classifier

▼ Table 1.

Performance of the new implementation of the classifier compared to the original study. Standard machine learning performance metrics are given as weighted averages of metrics computed in a stratified 5-fold cross-validation on the training data set (TCGA, 69 samples). Baseline classifier used the empirical class distribution for predicted probabilities and predicted majority class (KP).

	Accuracy	Precision	Recall	AUC	Log loss
Logistic regression	0.928	0.928	0.928	0.995	0.124
Baseline	0.435	0.189	0.435	0.5	1.067

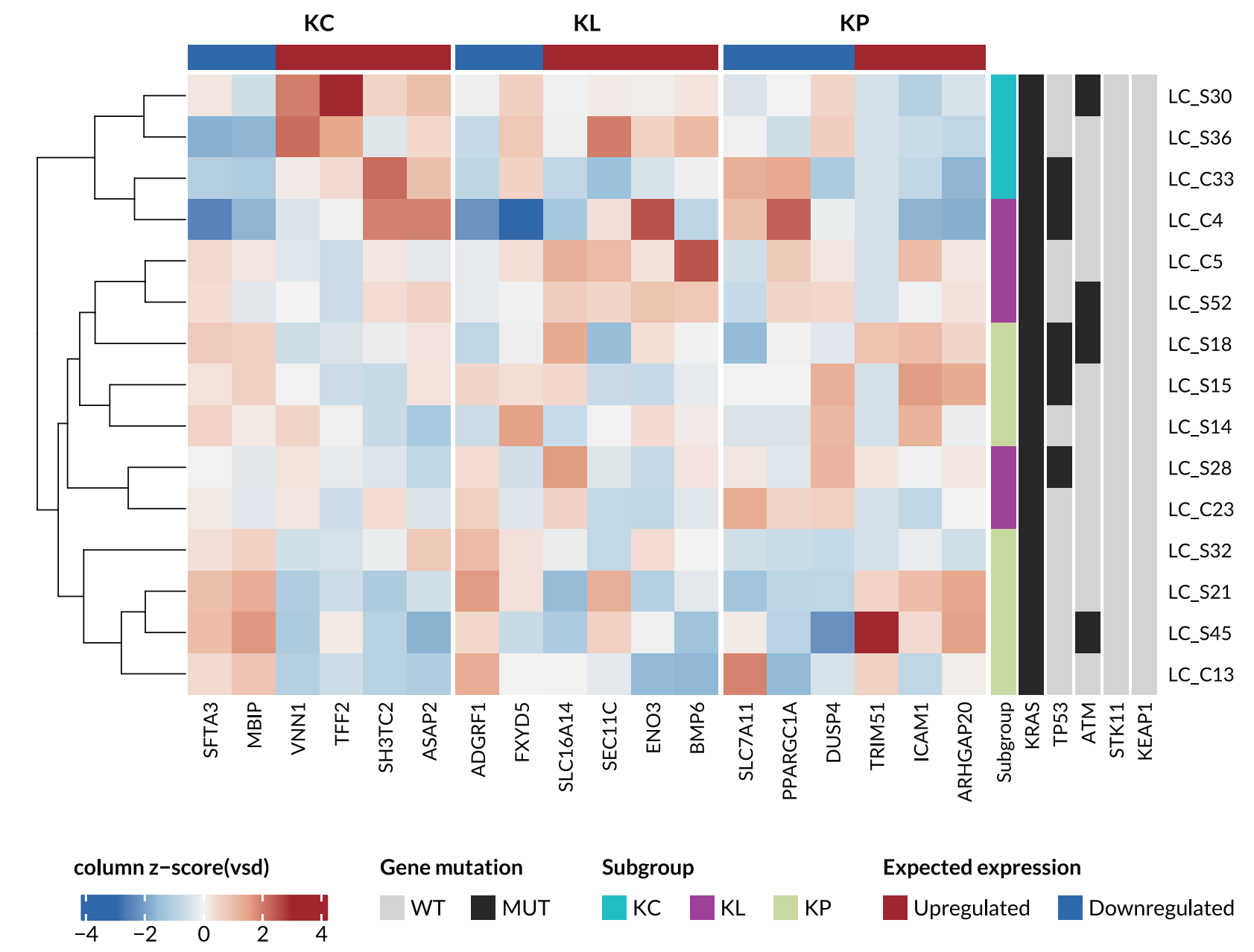
Accuracy: Number of correct predictions / Total number of predictions

Precision: Per-subgroup weighted average of Number of correct subgroup predictions / Total number subgroup of predictions

Recall: Per-subgroup weighted average of Number of correct subgroup predictions / Total number of samples in that subgroup

Area Under the Receiver Operator Curve (AUC): Per-subgroup weighted average of AUCROC scores calculated for one-versus-rest comparison

Log loss: Indicates the proximity to the true probability values



▲ Figure 2.

Gene expression and co-mutational landscape of 15 KRAS-mutant lung adenocarcinomas (LUAD) in the independent data set.

The heatmap shows relative gene expression levels (standardized, log2 transformed and variance stabilized counts) of the 18 signature genes. Samples are in rows, while genes are in columns. Hierarchical clustering with optimal leaf reordering is done on rows (samples). Columns are grouped to include genes informative of a specific subset. The expected gene expression in each cluster, shown on the top of the columns, was determined based on the original article. Common co-mutations are reported on the right side of the heatmap. Wild-type (WT) is represented in gray, and nonsynonymous mutations (MUT) are represented in black. Other co-mutations mentioned in the original article are excluded from the plots, since they were not detected in the independent cohort.

The model does not perform well on an independent cohort

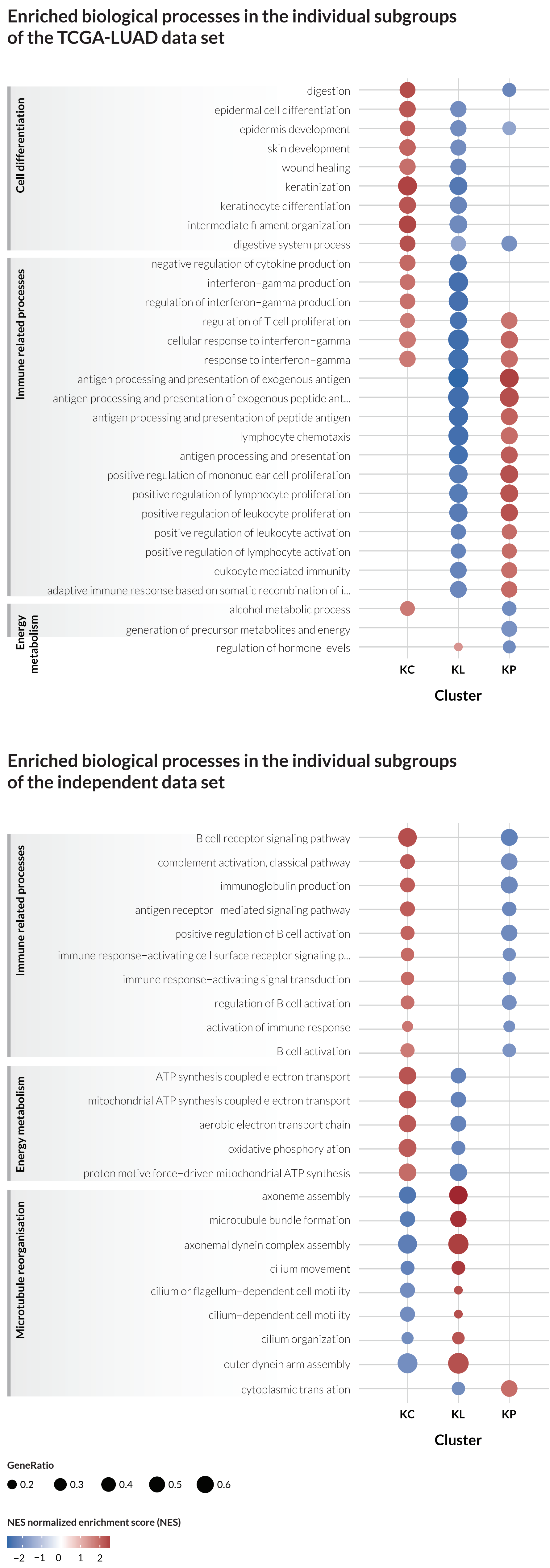
Gene expression patterns (heatmap, *Figure 2*) in an independent lung cancer cohort do not support the hypothesis of three subgroups as identified in the original study. Notably, the independent cohort did not exhibit distinct characteristics associated with KL subgroups. The majority of KL samples had lower predicted probabilities, highlighting the model's difficulties in distinguishing them from the other two subgroups. No KRAS mutated samples had co-occurring STK11 or KEAP1 mutations, typically found in the KL subgroup. Furthermore, the observed frequencies of these mutations in the whole Korean cohort are lower than expected based on TCGA-LUAD data, with only two STK11 mutations and no KEAP1 mutations in 87 patients. Comparably lower mutation rates of these two genes were also observed in other Asian cohorts[13–15], indicating that original observations may not hold for particular demographics. Additionally, co-mutations in TP53 were present in all three subgroups, which contradicted the findings based on the TCGA data.

Groups identified in an independent cohort have different characteristics compared to the original definitions

In the independent data set, differential gene expression analysis comparing each subgroup to the other two revealed several hundred differentially expressed genes (DEG) in the KC (231 up, 121 down; FDR < 0.05) and KP subgroups (468 up, 378 down; FDR < 0.05) and only 32 in the KL subgroup (20 up, 12 down; FDR < 0.05).

Gene set enrichment analysis highlighted negative enrichment of immune related pathways in the KP subgroup and upregulation of B-cell related pathways in the KC subgroup. Cell differentiation processes showed less prominent enrichment in the KC subset of the independent data set[not shown]. Instead, this subgroup was enriched in microtubule reorganization processes.

Overall, these results suggested that the predicted subgroups in the independent cohort could not be characterized by the same biological processes as in the original study (*Prior Art Box*).



▶ Figure 3.

Enriched biological processes in the individual subgroups of the TCGA-LUAD data set (above) and independent data set (below).

Dot plot shows the top ten significantly enriched biological processes in each subgroup identified by gene set enrichment analysis (GSEA). For GSEA, all expressed genes were ranked based on Wald's statistics from comparisons of one subgroup versus the other two. Dots are coloured by the normalized enrichment score (NES), with red representing positive enrichment scores and blue negative enrichment scores. Gene ratio is calculated as the number of core enriched genes divided by gene set size. Same groups are characterized by different biological processes in the two independent data sets.

CONCLUSION

The current study yielded an independent implementation of a published 18-gene expression signature for lung adenocarcinoma subtypes.

While the results of the original study could be faithfully recapitulated, the classification scheme did not transfer to a new cohort with different demographics and treatment history, indicating a need for a more robust or generalizable set of features and models.

Our reimplementation enables testing the signature in diverse clinical contexts and serves as a crucial foundation for developing a highly generalizable model.

References

1. Prior, I. A., Hood, F. E. & Hartley, J. L. The Frequency of Ras Mutations in Cancer. Cancer Res. **80**, 2969–2974 (2020).
2. Jänne, P. A. et al. Adagrasib in Non-Small-Cell Lung Cancer Harboring a KRASG12C Mutation. N. Engl. J. Med. **387**, 120–131 (2022).
3. Skoulidis, F. et al. Sotorasib for Lung Cancers with KRAS p.G12C Mutation. N. Engl. J. Med. **384**, 2371–2381 (2021).
4. Skoulidis, F. et al. Co-occurring genomic alterations define major subsets of KRAS-mutant lung adenocarcinoma with distinct biology, immune profiles, and therapeutic vulnerabilities. Cancer Discov. **5**, 860–877 (2015).
5. Daemen, A. et al. Transcriptional Subtypes Resolve Tumor Heterogeneity and Identify Vulnerabilities to MEK Inhibition in Lung Adenocarcinoma. Clin. Cancer Res. **27**, 1162–1173 (2021).
6. Seo, J.-S. et al. The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. **22**, 2109–2119 (2012).
7. In Detail: General RNA-Seq analysis pipeline (featureCounts). Genialis https://genialis.zendesk.com/hc/en-us/articles/360012780273-In-Detail-General-RNA-Seq-analysis-pipeline-featureCounts- (2023).
8. RNAseq short variant discovery (SNPs + Indels). GATK https://gatk.broadinstitute.org/hc/en-us/articles/360035531192-RNAseq-short-variant-discovery-SNPs-Indels- (2023).
9. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. **46**, D1062–D1067 (2018).
10. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. **15**, 1–21 (2014).
11. Gu, Z., Eils, R. & Schlesner, M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics **32**, 2847–2849 (2016).
12. Wu, T. et al. clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. The Innovation **2**, (2021).
13. Lengel, H. B. et al. Genomic mapping of metastatic organotropism in lung adenocarcinoma. Cancer Cell **41**, 970-985.e3 (2023).
14. Cerami, E. et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov. **2**, 401–404 (2012).
15. Gao, J. et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci. Signal. **6**, pl1 (2013).