**LKB1 loss induces characteristic patterns of gene expression in human tumors associated with NRF2 activation and attenuation of PI3K-AKT.**

Jacob M. Kaufman1, Joseph M. Amann2, Kyungho Park1, Rajeswara Rao Arasada2

Haotian Li3, Yu Shyr4, David P. Carbone2

**Authors Affiliations:**

1: Vanderbilt University; Department of Cancer Biology;

2: Ohio State University; Department of Internal Medicine;

3: University of Kentucky College of Dentistry;

4: Vanderbilt University; Department of Biostatistics

**Keywords:**

Genomics, LKB1, NRF2, PI3K/AKT

**Grant support:**

SPECS grant: NCI U01CA114771.

**Corresponding Author:**

David P. Carbone, MD, PhD

Barbara J. Bonner Chair in Lung Cancer Research

Professor of Medicine

Director, James Thoracic Center

James Cancer Center

The Ohio State University Medical Center

488 Biomedical Research Tower

460 West 12th Avenue

Columbus OH 43210

email: david.carbone@osumc.edu

phone: (614) 685-4478

fax: (614) 293-4372

Supplementary Materials:

**Supplementary Figures S1-S20**

**Supplementary Tables S1-S7**

**Supplementary Methods**

**Supplementary methods:**

**Obtaining publicly available molecular and clinical data.**

For TCGA data, the following types of ‘Level 3’ processed and normalized molecular data were downloaded from the TCGA GDAC website:

 https://confluence.broadinstitute.org/display/GDAC/Home:

illumina hiseq RNAseq v2 gene expression (RSEM normalized),

illumina hiseq RNAseq v2 exon expression data (RPKM normalized),

illumina HumanMethylation450 BeadChip methylation data,

illuminahiseq mirnaseq miR gene expression data

RPPA proteomic data

TCGA analysis of lung adenocarcinoma samples had not been completed at the time of submission of this manuscript; data were obtained from the ‘stddata run’ performed on July 15, 2013. Copy number data was taken from the ‘analysis run’ performed on May 23, 2013 and gene level data was used. Other publicly available datasets were downloaded from GEO and ArrayExpress or from individual websites, as listed in Supplementary Table S1. Processed data uploaded to these sites by their original contributors were downloaded as ‘series matrix.txt’ files. In cases where data were presented as linear expression values, log2 transformed values were used. To collapse gene lists such that each gene was represented only once in our analyses, standardized scores from multiple probesets representing the same gene were averaged to give a single value.

**Determination of clinical variables and mutation status in samples.**

For resected tumors, stage, survival and smoking status were extracted from clinical information published for each dataset. Mutation status of LKB1, KRAS, and EGFR were also included for samples in the Wash U, MSKCC, MSKCC2, UNC, and USC datasets. For TCGA samples, unprotected ‘level 2’ somatic mutation calls were downloaded from the TCGA data portal website ‘https://tcga-data.nci.nih.gov/tcga/’. This gave mutation data for 519 unique tumor identifiers, of which 406 also had RNAseq gene expression data. LKB1 mutations were observed in 63 of these 406 tumors. Four additional TCGA samples (TCGA-50-7109, TCGA-55-6971, TCGA-44-6145, and TCGA-50-5939) had LKB1 mutations that were reported in COSMIC; these four samples were considered LKB1 mutant in our analyses. For KRAS, EGFR, and KEAP1, we included all non-silent mutations in our analysis, without attempting to distinguish ‘canonical’ mutations from those of unknown significance.

For cell lines, different sources have studied the LKB1 mutation status of many of the cell lines used in our work. We used data from exome capture sequencing performed by the GDSC (Garnett et al, Nature 2012; file: http://www.cancerrxgene.org/; gdsc\_manova\_input\_w2.xlsx) and CCLE (Barretina et al, Nature 2012; file: http://www.broadinstitute.org/ccle/home; files: CCLE\_Oncomap3\_2012-04-09.maf; CCLE\_hybrid\_capture1650\_hg19\_NoCommonSNPs\_NoNeutralVariants\_CDS\_2012.05.07.maf) studies, and also downloaded mutation files from the COSMIC database (Forbes et al, Nucl Acids Res 2011; http://cancer.sanger.ac.uk/cancergenome/projects/cosmic/) to determine mutations in LKB1.

Determination of LKB1 status in cell lines was not straightforward. The CCLE study identified 25 cell lines with LKB1 mutations while 30 were identified in the GDSC study. Of the 433 cell lines that were included in both studies, 28 cell lines were determined to have an LKB1 mutation in at least one study, but of these, only nine LKB1 mutations were seen in both studies, while the other LKB1 mutation status of the other 19 cell lines was discrepant between the CCLE and GDSC analyses, with an LKB1 mutation reported in one dataset and wild-type status reported in the other. Furthermore, in the CCLE study LKB1 nonsense mutations (p.Q37\*) in A549 and H460, which are well-known in the literature, were not found in the exome capture data, but were identified using the targeted Oncomap approach. Thus, we considered each of these studies to have high specificity for determination of LKB1 mutations, but low sensitivity. We supplemented these data by considering LKB1 mutations and loss of LKB1 by other mechanisms that were reported in an additional eight sources. A cell line was considered to have LKB1 loss if it was reported as such in any of these studies. Altogether, LKB1 loss was determined in 47 NSCLC cell lines, of which 39 had associated mRNA expression data to assess the LKB1-loss classifier. Because we showed that LKB1 mutations were missed by the GDSC and CCLE studies, we chose to define a cell line as LKB1 wild-type only if this was the result of sequencing performed by both studies. Cell lines that were determined to be LKB1 wild-type but were only included in one of the two studies were considered to have unknown LKB1 status for our analyses of sensitivity and specificity. The LKB1 status for all cell lines used in our work, as well as the LKB1-loss score determined by our gene expression analysis, is provided in ‘Supplementary Data 2’.

**Analysis of gene expression after LKB1 expression in H2122 and A549 cells.**

For our own LKB1 perturbation analyses, mRNA was isolated from A549 and H2122 after stable expression of pBABE, LKB1 or LKB1 K78I using a Qiagen mRNA isolation kit, with trizol extraction reagent. Three biological replicates were included for each condition, except for H2122 pBABE in which one replicate had poor RNA; thus, gene expression was analyzed for 17 total samples. RNA concentrations were measured, and the RNA integrity number and 28s:18s ratio were calculated for quality control purposes. Amplification of 130ng of total RNA was performed using Ambion WT Expression kit, and in vitro transcription was carried out overnight. cRNA was subsequently cleaned using Ambion-WT bead cleanup kit. 10.5ug of cRNA was used for second cycle cDNA synthesis and resulting cDNA was cleaned using Ambion-WT bead cleanup kit. 5.5ug of purified cDNA products were used in fragmentation and labeling reactions. Samples were hybridized overnight to a HT Human Gene 1.1 ST PM16 array plate utilizing a GeneTitan instrument. They were then scanned on the Affymetrix Gene Titan AGCC v. 3.2.3 and then analyzed on Affymetrix Expression Console v. 1.1 using a RMA normalization algorithm producing log base 2 results. These data were uploaded to the Gene Expression Omnibus data repository; GSE51266.

**Analysis of gene expression associated with LKB1 mutations.**

For clinical and cell line datasets in which LKB1 status was known, a student’s t-test was performed to determine statistical significance of differences in gene expression between LKB1 these comparisons. For cell line data, LKB1 mutation status was used as listed in Supplementary Data 2. For data from the Directors Challenge Lung consortium, to avoid potential issues with site-to-site differences or batch effects, we considered only the data from the Michigan cohort (n=178), as this was the largest of the four patient cohorts in the study. LKB1 mutation status was unknown, and associations with LKB1 expression were determined using p-values derived from linear regression analysis, by fitting the expression of each probeset to the expression of the LKB1 probeset 41657\_at. More than 70% of the mutations reported for LKB1 in the TCGA are nonsense, frameshift, or splice site, which strongly decrease LKB1 mRNA expression, as shown in Figure 4D. Thus, when mutation status is unknown, LKB1 mRNA levels can be used as a surrogate for LKB1 loss for the purpose of statistical comparisons across an entire aggregate dataset, although this would not be suitable to determine the LKB1 status of any single patient. Linear regression was performed using the R Bioconductor software platform, with the lm() function in the limma package. Finally, for gene expression data from LKB1-mutant cell lines stably transduced to over-express LKB1, average differences were determined for each probeset, comparing expression with wild-type LKB1 to expression with vector only or LKB1 K78I control.

**Comparison of LKB1-associated genes across multiple studies using overlap significance analysis.**

Associations with LKB1 were made for comparisons of sample groups listed in Supplementary Table 1. Genes represented by probesets on the Affymetrix U133A chips were used for further analysis, with each gene represented by a single probeset, resulting in 13211 genes. The top 200 genes over-expressed in LKB1-deficient samples were used as gene lists for each of these comparisons. Numerical overlap was determined between each pair of gene lists, and the statistical significance of this overlap was calculated using a hypergeometric distribution with the phyper() function in R bioconductor’s limma package. When comparisons were made between different platforms, only genes included in both platforms were considered for statistical calculations. P-values were log10 transformed and represented graphically using Cluster 3.0 and Java TreeView software, and is shown in Fig. 2A. Raw P-values are presented in this figure, without adjustment for multiple hypothesis testing. The actual number of overlapping genes and the statistical significance is reported in Supplementary Table S2.

 We also performed Gene Set Enrichment Analysis using the Broad Institute’s GSEA tool (http://www.broadinstitute.org/gsea/index.jsp). The TCGA analysis was taken as the comparator for the analysis because of its large sample size; genes were ranked according to their association with LKB1 mutations in this study, and enrichment plots were generated for comparison with gene lists representing the top 200 upregulated LKB1-associated genes in five clinical studies, four cell line studies, and four mouse comparisons. For the Ji mouse study we considered LKB1/KRAS mutant lung adenocarcinomas compared to either KRAS mutant alone or KRAS/p53 mutant tumors. Tumors in this model that result from KRAS mutations alone tend to be relatively benign and may be more comparable to human adenomas, while KRAS/p53 mutant tumors are aggressive and metastasize, so using these tumors as the comparison group may better reflect the phenotypes of adenocarcinomas.

**Development of LKB1-deficient gene signature.**

We used a training and testing approach to develop and test a gene signature capable of classifying LKB1-deficient tumors. We generated three gene lists using statistical comparisons from two training sets: the Wash U set with comparisons to documented LKB1 mutations and the Michigan samples from the Director’s Challenge Consortium with comparisons to LKB1 expression. The LKB1 classifier was taken as the intersection of these three lists.

ListA in Wash U:

All probesets ‘*x*’ such that raw p-value < 0.01 for student’s t-test comparing LKB1 mut (n=7) vs LKB1 WT (n=34), resulting in 601 selected probesets.

ListB in Mich:

All probesets ‘*x*’ such that raw p-value< 0.01 for linear regression model of 178 tumors, resulting in 3679 probesets:

*expr(204292\_x\_at) ~ a \* expr(x) + b*

ListC in Mich:

All probesets ‘*x*’ such that raw p-value< 0.01 for linear regression model of 178 tumors, resulting in 3467 probesets:

*expr(41657\_at) ~ a \* expr(x) + b*

Classifier = (ListA) ∩ (ListB) ∩ (ListC)

Lists B and C show a high degree of overlap, sharing more than half their genes, as they are derived from the same source and represent association with the two distinct LKB1 probesets. The intersection of the three lists results in a set of 167 probesets, a significantly larger intersection than expected by chance (p-value = 6.8e-38 by hypergeometric test). Standardized values were then averaged across probesets representing the same gene to give a final set of 129 standardized gene expression values.

**Derivation and validation of 16-gene LKB1 loss signature**

For the Michigan training cohort (n=178 tumors), expression data for each of these 129 genes was mean centered and normalized, and unsupervised hierarchical clustering was performed with Gene Cluster 3.0 utilizing uncentered Pearson’s correlations and the centroid linkage method. The same analysis was performed for the TCGA lung adenocarcinomas (n=446 tumors), although 19 genes of the 129 genes were not included in these data. Similar results were obtained for clustering of genes and tumors when using Spearman’s rank correlation as the similarity metric (data not shown). Resulting heat maps were visualized using Java TreeView application. Unsupervised clustering revealed correlation patterns within the genes of the LKB1-deficient signature that were reproducible across multiple resected LUAD datasets. Four distinct sets of co-regulated genes were identified from resulting dendrograms as gene clusters with internal centroid correlation values greater than 0.5, including three transcriptional nodes showing increased expression among LKB1-deficient tumors and one with decreased expression. The genes comprising these transcriptional nodes are given in Supplementary Data File 1. The same four nodes could be observed in classifiers independently derived from either the Michigan or the TCGA samples, demonstrating the reproducibility of these clustering patterns across multiple datasets. We compared the genes comprising each cluster for the Michigan cohort and the TCGA cohort, showing the amount of overlap with Venn diagrams in Supplementary Fig. S3 and calculating the statistical significance of this overlap using the Fisher’s exact test. We hypothesized that the expression of these transcriptional nodes was driven by different underlying phenotypes.

To determine which of these clusters had the strongest association with LKB1 loss, we first calculated a numeric score for each of these transcriptional nodes was first calculated by taking the average of the standardized expression values (subtracting mean expression and dividing by standard deviation) for the genes comprising the node. Then, again using LKB1 mRNA expression (probeset 41657\_at) as a surrogate marker for LKB1 loss in the Michigan training cohort, we determined the association of each cluster with LKB1 expression using a univariate linear regression. This revealed the 16-gene ‘FOX/CREB’ cluster to have the strongest association with LKB1 loss. Similarly, in the Wash U. training set, the FOX/CREB cluster was expressed by each of the seven LKB1 mutant tumors. These 16 genes were therefore used throughout the remainder of the work as the ‘LKB1-loss’ signature.

For calculation of the LKB1-loss score in cell line datasets we used gene expression from all cell lines, regardless of corresponding primary tumor type, to determine the mean and standard deviation for these 16 genes. These standardized expression scores were then averaged to give a single numeric LKB1-loss score. Gene expression data for cell lines came either from the Sanger collection (www.broadinstitute.org/cgi-bin/cancer/datasets.cgi; Sanger\_Cell\_Line\_Project\_Affymetrix\_QCed\_Data\_n798.gct) or from analysis by the CCLE (www.broadinstitute.org/ccle/home; CCLE\_Expression\_2012-09-29.res). In cases where cell lines were included in both of the datasets, the average of the two scores was used for analysis. Scores from the two sources were significantly correlated, with 91% concordance in predicted classification of LKB1 status between the two scores, showing that the expression of these genes is largely consistent across independent measurements by different studies.

Accuracy for predicting known LKB1 mutations in NSCLC was determined for the 67 LKB1 mutations in the TCGA resected lung adenocarcinomas, 26 LKB1 mutations in a pooled analysis of MSKCC2, UNC, and USC resected lung adenocarcinomas, and 39 LKB1 mutations in NSCLC cell lines. Sensitivity and specificity are reported in Fig. 4C, with P-values calculated using the Fisher’s exact test.

LKB1 mutation data was also available for the MSKCC1 dataset and these samples represented another potential test set. However, for unknown reasons univariate analysis comparing reported LKB1 mutant and wild-type tumors in this dataset yielded fewer significant gene associations than would be expected by chance. In this dataset only three genes passed a p-value cutoff of 0.001; in contrast, for the smaller Wash U cohort and the comparably sized MSKCC2 cohort, 47 genes and 72 genes passed this cutoff, respectively. Furthermore, the top ranked genes associated with LKB1 mutations in this dataset showed no significant overlap with the consistent pattern of gene expression observed in each of the other clinical and cell line datasets (Fig. 1A). Based on these findings we considered this dataset an outlier and excluded these data from our validation.

**Generating expanded gene lists for four transcriptional nodes.**

Statistical approaches used to study gene expression patterns can be limited in power when a list of genes is too short, but can be prone to yield spurious associations when gene lists are excessively long. For instance, the default settings on the Broad Institute’s GSEA program are to reject gene lists shorter than 25 genes or larger than 500. Thus, we used a general linear model with each of these four ‘cluster scores’ to identify the 200 genes with the strongest association to each of the clusters using the Michigan training cohort. The lm() function in the limma package of R bioconductor platform was used to determine the best fitting parameters to relate the expression of each probeset to the scores of the four LKB1-asociated gene clusters; interaction terms were not included in the model:

*expr(x) ~ a \* expr(LKB1\_loss) + b \* expr(Mito)+*

*c \* expr(NRF2)+ d \* expr(Down)*

Gene lists for each of the four nodes were then constructed taking the top 200 most significantly up-regulated genes for that node as determined by the corresponding p-value from this model (Supplementary Data 1).

**Statistical inference of biological significance for four transcriptional nodes.**

We used several approaches to identify candidate pathways that could potentially drive the expression of the four transcriptional nodes observed in our analysis. Gene set enrichment analysis was performed using the molecular signatures database (http://www.broadinstitute.org/gsea/msigdb/) to determine enrichment of transcription factor consensus sequences in the promoter regions of these gene lists. This tool was also used to compare our four gene lists to previously characterized perturbation and cancer-derived signatures.

The connectivity map was used to determine significant similarities between our eight gene lists and gene perturbations induced in the cell lines MCF7, HL60, and PC3 by six hours of treatment with 1309 different small molecules. We uploaded our gene lists onto the connectivity map online analysis tool to rank compounds that were significantly associated with the gene expression phenotypes we observed. For significant hits we then downloaded ranked perturbation lists from the Connectivity Map website to generate our own enrichment p-values.

Finally, we generated an association matrix using searches of GEO and ArrayExpress to obtain perturbations of interest to our study. Because the connectivity map did not employ a lung cancer derived cell line, we searched for all perturbations made to A549, a commonly studied lung adenocarcinoma cell line with a mutation in LKB1. We next performed targeted queries for perturbations related to the hypotheses suggested by our GSEA and connectivity map analyses; specifically we searched for perturbations involving pharmacologic or genetic modulations of the CREB and FOXO3 pathways, the NRF2 transcription factor, mitochondria, and protein translation. Also, for the connectivity map associations highlighted in Supplementary Tables S3-S5, we downloaded ranked perturbation lists from the Connectivity Map website to generate our own enrichment p-values for the corresponding perturbations. We eliminated redundant probesets to reduce our association matrix to a single probeset per gene, and then determined the top 200 over-expressed and under-expressed genes associated with each perturbation (roughly the top and bottom 2% of changes). Numeric overlap was then determined with each of the four cluster scores and statistical significance calculated using a hypergeometric distribution by the phyper() function in the Bioconductor limma package.

**Determining associations between LKB1 loss and other molecular features of lung adenocarcinoma.**

Lung adenocarcinomas with LKB1 mutations have previously been shown to exhibit an increased prevalence of KRAS mutations and are almost mutually exclusive of EGFR mutations. These tumors likely have unique biological features in addition to the ones that have already been described in the literature, and understanding these distinctive features may increase our understanding of their biology and potential treatment strategies. We took advantage of the extensive characterization of lung adenocarcinomas undertaken by the TCGA to identify other features associated with LKB1 loss. Statistical comparisons could be made on the basis of LKB1 sequence mutations. However, our results demonstrate that approximately half the instances of LKB1 loss occur without the presence somatic mutations detected in LKB1. We expect the statistical power of our observations to be increased by classifying tumors on the basis of the LKB1-loss gene signature, both because the number of LKB1-loss tumors is increased and because including these tumors among the LKB1 wild-type group would dilute the differences between the two groups. Therefore we chose to use the LKB1-loss classifier as the basis for determining LKB1 status for these comparisons.

For copy number data, we analyzed 79 regions chromosomal alterations determined to be statistically significant according to analysis performed by the TCGA Genome Data Analysis Center (Broad Institute TCGA Genome Data Analysis Center (2013): Analysis Overview for Lung Adenocarcinoma (Primary solid tumor cohort) - 23 May 2013. Broad Institute of MIT and Harvard. [doi:10.7908/C1D798FR](http://dx.doi.org/10.7908/C1D798FR)). The denotation of these regions as 'amplification' or 'deletion' peaks was determined by the TCGA analysis. Although the actual copy number of a given gene in a disrupted region is a discrete integer value, the measured copy number data are reported as a single continuous variable across tumors; thus, interpreting the significance of a particular value is not straightforward. More extreme values indicate a greater likelihood of copy number loss and are more associated with homozygous loss than heterozygous loss, but no single cutoff crisply discriminates two copies from one heterozygous copy number loss, and no value clearly distinguishes heterozygous from homozygous loss; likewise for amplifications, greater numbers are indicative of increased copy numbers. Because of the uncertainty inherent in specifying a cutoff to delineate loss or amplification of a given chromosomal region, we chose to perform analysis on these data as continuous variables, and reported p-values representing Student's t-test comparing LKB1-loss to LKB1-wild type tumors (Supplemental Table S7). Q-values were then reported to adjust for the multiple hypothesis testing of 79 distinct regions.

For miRNA expression and RPPA protein expression, Student's t-tests were also used to compare these two groups. Data were obtained as described previously; expression data were log2 transformed prior to analysis of miRNA expression, while RPPA data were used as given, with arbitrary units. For miRNA, we limited our investigation to species with an average log2 expression of 4.0, representing 173 out of 1046 total miRNA species reported. Raw p-values were reported, as well as Q-values representing correction for multiple hypothesis testing.

**Association Between LKB1 Loss and Prevalence of Other Mutations**

Co-occurrence or mutual exclusivity of mutations can reveal information about how particular pathways are activated or interact with each other within cancer cells. Mutual exclusivity – a statistically significant decrease in the co-occurrence of two somatic mutations, can occur because of functional redundancy, i.e. that both mutations could result in the same effects in the cell, or because of negative interactions between pathways downstream of the mutations, for instance if the activity of one mutation leads to attenuation of the other mutated gene due to feedback inhibitory effects. On the other hand, increased likelihood of two mutations occurring in the same tumor suggests that the effects are synergistic, perhaps by significantly amplifying oncogenic signaling through a single pathway, or by activating distinct pathways that have cooperative effects. There could be alternative explanations for co-occurrence of different mutations independent of the function of the genes. A single phenotype might increase the likelihood of the two genes undergoing mutations; for instance, this could result if a tumor had disrupted mismatch repair genes, or if a large chromosomal gain or loss led to increased chances of mutation in closely positioned genes.

To determine the association of LKB1 loss with other mutations we used 403 lung adenocarcinomas characterized by the TCGA that had both RNAseq gene expression data and determination of somatic mutations by exon capture DNAseq. Mutations were found in 16302 genes, with a total of 102425 somatic alterations observed. Because of the high levels of exposure to carcinogens in cigarette smoke, non-small cell lung cancer has one of the highest rates of DNA mutations in cancer, with a median of five to ten nucleotide changes per megabase of DNA (Lawrence et al., 2013). Thus, most of the observed alterations are ‘noise’ mutations that occur at random throughout the genome having no functional significance. There are sophisticated algorithms to rank the genes affected by mutation, taking into account effects such as gene length, the context of the specific nucleotides surrounding an observed mutation, and the pattern and rate of specific base changes seen in a given tumor, in order to determine which genes are most likely to have been mutated in a nonrandom way (Imielinski et al., 2012). Non-random mutations are more likely to have been selected during the evolution of a cancer because of their functional effects. For our analysis, we narrowed our search to known cancer genes listed in the COSMIC Cancer Gene Census, as well as NRF2 and KEAP1, which are known to be mutated in NSCLC; we further limited our search to genes with at least a 5% prevalence of mutation in our dataset, because less prevalent mutations are unlikely to reach statistical significance in this analysis. These filters yielded a list of 32 genes, which were then subjected to statistical, testing using the Fisher’s exact test to determine if their prevalence is significantly different between LKB1 wild-type and LKB1-deficient tumors. Any genes included in the analysis that are affected only by chance mutations would be unlikely to segregate significantly with LKB1 status. Thus, the inclusion of such genes is unlikely to affect the results of our analysis, except to increase the likelihood of a false positive result, necessitating correction of raw P-values by multiple hypothesis testing.

Using this approach we identified five genes with significantly different mutation rates in LKB1-deficient and wild-type lung adenocarcinomas. KRAS, KEAP1, and ATM each showed significantly higher mutation rates among LKB1-deficient tumors, while EGFR and p53 had a lower rate of mutation.

Several previously published datasets were available in which tumors were characterized by gene expression profiling and sequencing of targeted mutations, and these were used to verify the observations made in the TCGA cohort, when possible. We pooled previously published datasets by combining the Director’s Challenge tumors (n=449) with MSKCC1 (n=91), MSKCC2 (n=102), UNC (n=116), USC (n=58), Wash U (n=68), and Duke (GSE3141, n=58). Overlapping tumors from MSKCC1 and the Director’s Challenge tumors were excluded from the Director’s Challenge samples for a total of 851 unique tumors. LKB1 status was classified for all tumors based on the expression of the LKB1-loss score and prevalence of LKB1 loss was determined for EGFR mutant vs. EGFR wild-type, KRAS mutant vs. KRAS wild-type, and p53 mutant vs. p53 wild-type. Statistical comparison of LKB1-loss prevalence between these tumor groups was made using the Fisher’s exact test.

**Determination of clinical associations.**

Our LKB1 classification score was used to predict LKB1-loss status among lung adenocarcinomas from the Director’s Challenge Consortium (n=449, of which 441 had survival data) and the TCGA (n=446, of which 372 had survival data). Kaplan-Meier analysis of overall survival performed between these two groups, using the R statistical platform with the survival package. Statistical significance was determined using the log rank test.

Initial clinical stage at presentation and smoking status were also examined in the TCGA dataset and in the pooled cohort described above. Prevalence of LKB1 loss was determined for each of these clinical categories. Prevalence of LKB1 loss for stage II, III, and IV was compared to that of stage I using Fisher’s exact test, which revealed no significant difference according to stage. Similarly, LKB1 loss prevalence was compared between ever-smokers and never-smokers, which revealed a decreased prevalence of LKB1 loss in never smokers in both the TCGA and pooled cohorts.