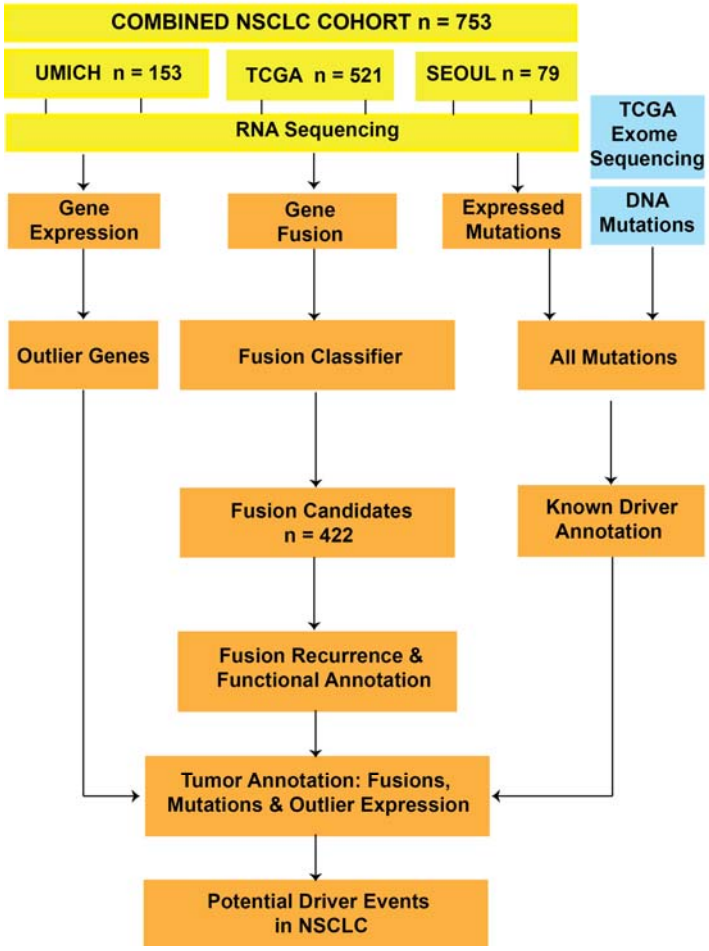
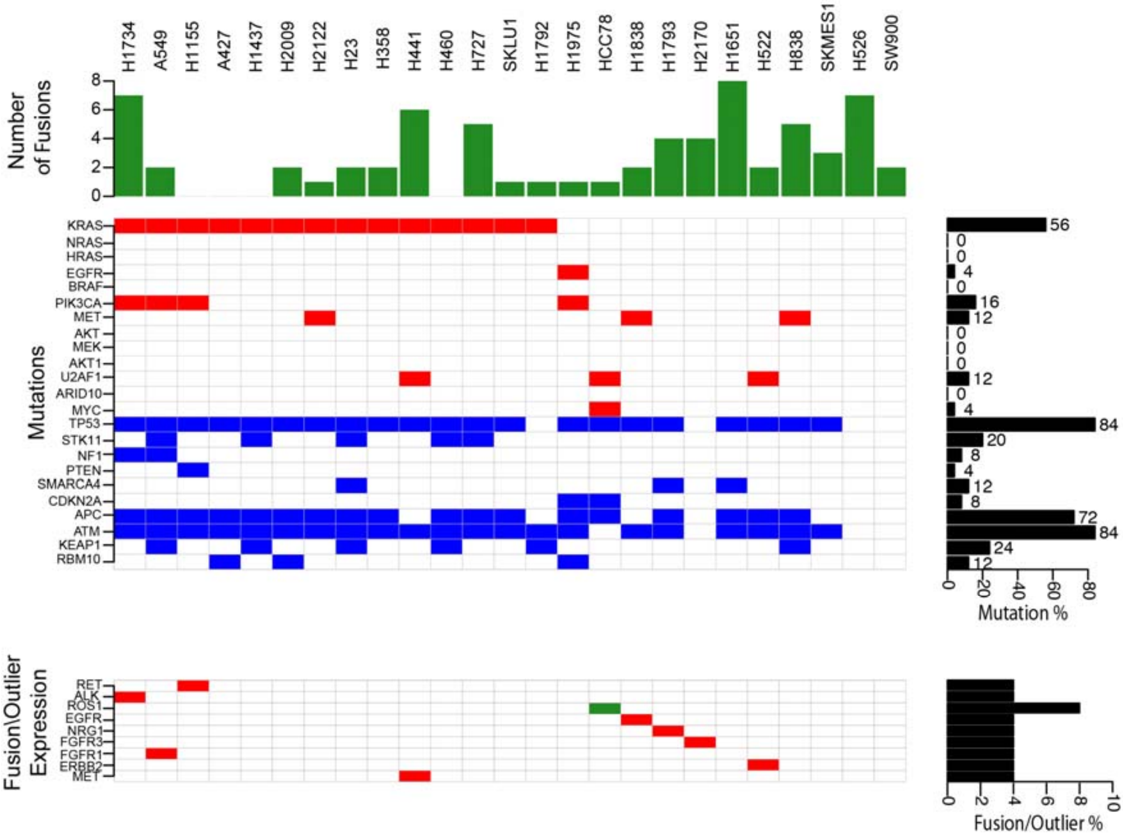
Supplementary Figure 1

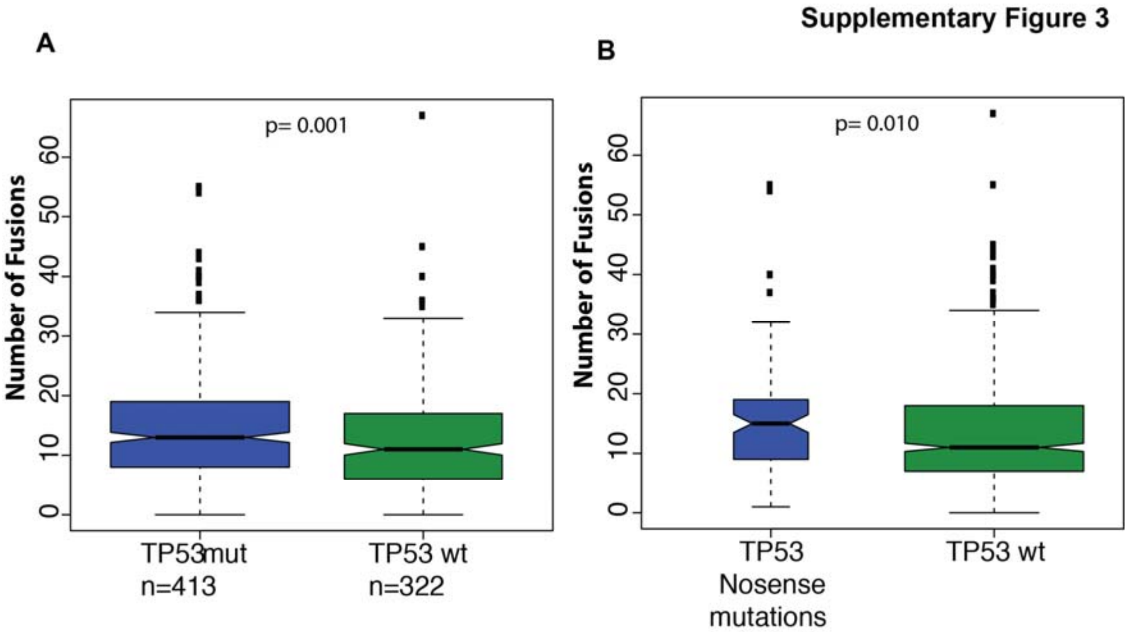


**Supplementary Figure 1.** Schematic diagram of the data generation and analysis workflow of lung cancer RNASeq data. A total of 753 lung cancer samples that include 728 clinical specimens and 24 cell lines, representing 451 LUAD and 251 LUSC, 11 LACC and 9 LCLC were interrogated for gene fusions and somatic mutations. The cohort was assembled combining 133 University of Michigan samples (UMICH), 79 Seoul National University samples (SEOUL), and 521 Cancer Genome Atlas samples (TCGA). The RNASeq data was mapped to human RefSeq Hg19 using TopHat2. Fusion calls were made with TopHat-Fusion (THF). In all cases fusions present in normal samples were considered false positives and filtered out. We developed and applied a fusion classifier that retained 422 gene fusions for further downstream analysis. The 422 fusions were classified into recurrent (>2 samples) and private fusions and further divided into inter chromosomal, intra chromosomal or fusions resulting from potential tandem duplication events. In addition samples were annotated for outlier expressions and somatic/COSMIC mutations in well-known lung oncogene drivers and tumor suppressors. Finally, both LUAD and LUSC cohorts were divided into either samples harboring oncogene driving mutations or samples without known driver genes.

## Supplementary Figure 2

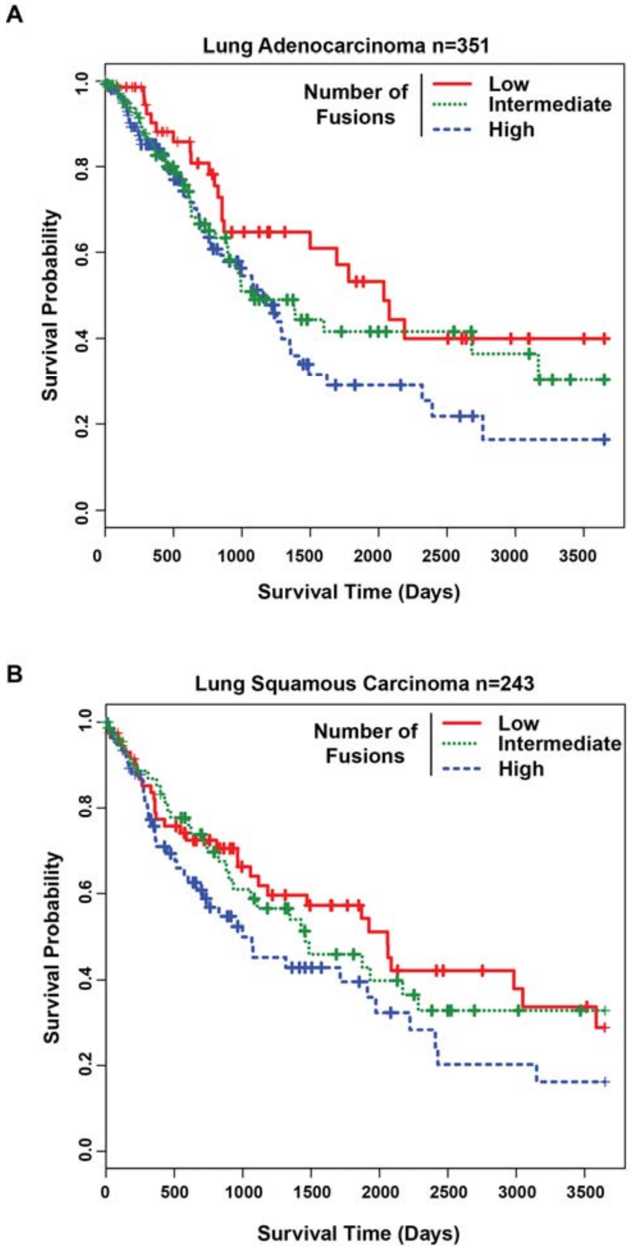


**Supplementary Figure 2:** Gene Fusions, Mutations and outliers in Lung Cancer Cell Lines. Histograms represent the number of prioritized fusions identified in each cell line sample. **Central Panel**: Heatmap denotes the presence or absence of activating mutations in known oncogenes (red) and deleterious mutations in tumor suppressors (blue). Samples are presented in columns and genes are in rows. **Right Middle Panel**: Bar plot summarizes the number of samples harboring activating or deleterious mutations for each gene. **Bottom Panel**: Heatmap displays samples harboring known gene fusions (green) involving receptor kinase genes. Samples in red indicate outlier expression pattern observed in the respective genes. The ordering of samples in center panels was dictated by mutation status in *KRAS*, *NRAS*, *HRAS*, *EGFR*, *BRAF*, *PIK3CA*, and *TP53* genes in that order. The bottom right histogram outlier percentage expression in the cell line panel. Data from the small cell lung cancer cell line H526 is shown alongside the NSCLC cell lines for comparison.



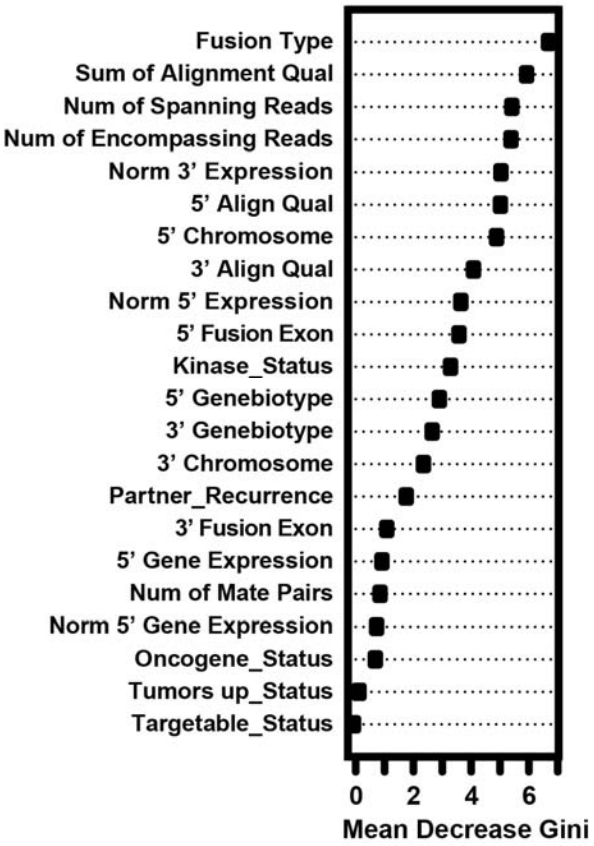
**Supplementary Figure 3:** Comparison between the Number of Fusions and *TP53* Mutation Status. **A**. Box plot representation of number of fusions in *TP53* wildtype vs all *TP53* mutated samples **B**. Box plot representation of number of fusions in *TP53* wildtype vs *TP53* nonsense mutations containing samples.

## Supplementary Figure 4



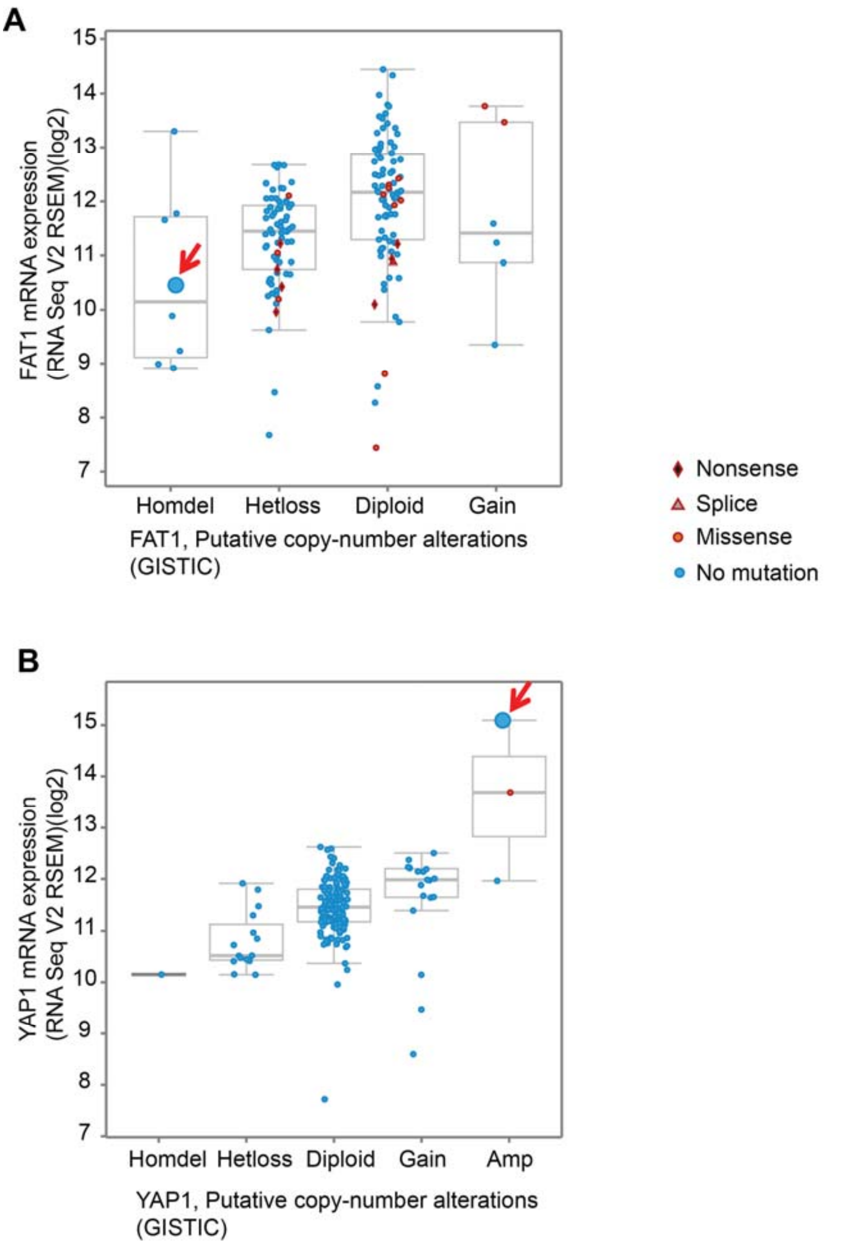
**Supplementary Figure 4:** Gene fusion frequency is a prognostic indicator in both LUAD and LUSC.**A**. Kaplan-Meyer survival curve for LUAD samples (n=351) with low (0-6) (n=55), intermediate (7-12) (n=185), or high (≥13) (n=111) number of fusions (Likelihood ratio test *P*=0.07562). Samples with high number of fusions have worst prognosis (Cox survival analysis *P*=0.0291). **B**. Kaplan-Meyer survival plot for LUSC samples (n=243) with low (0-11) (n=62), intermediate (12-18) (n=112) and high (≥19) (n=69) number of fusions (Likelihood ratio test *P*=0.1685). Samples with high number of fusions have worst prognosis (Cox survival analysis *P*= 0.0717).

## Supplementary Figure 5



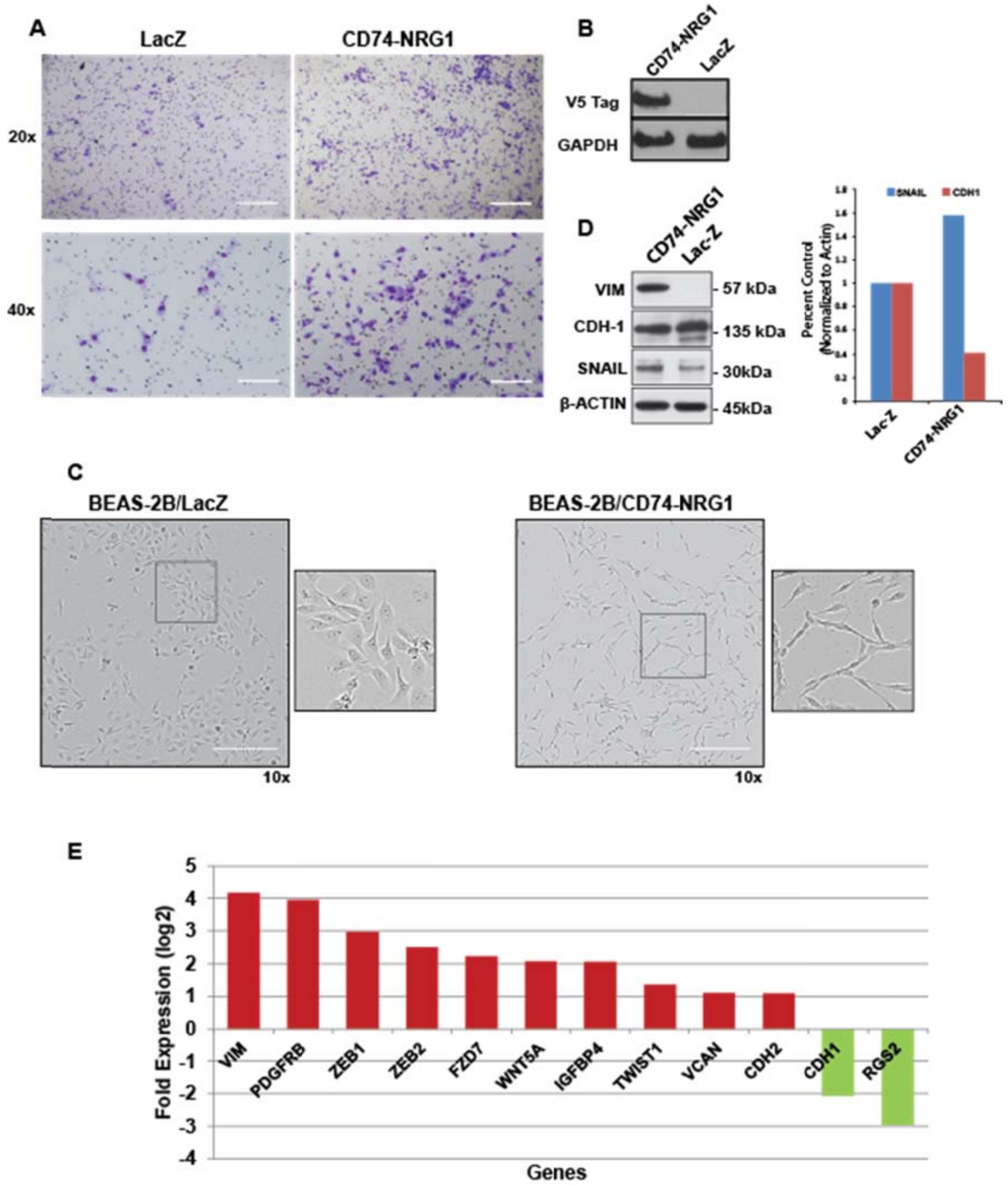
**Supplementary Figure 5:** Features Used by the Fusion Classifier. Importance of fusion classifier features in decreasing order (mean decrease GINI).

Supplementary Figure 6



**Supplementary Figure 6.** Integrative analysis of Hippo pathway gene aberrations in the fusion index cases. Box plot representation of mutation and copy number status along with mRNA expression values for the indicated Hippo pathway genes were generated using the analysis tool embedded in cbioportal (http://www.cbioportal.org) (A) *FAT1*(TCGA-43-3920) and (B) *YAP1* (TCGA-22-1016) aberration status in TCGA LUSC cohort (n=271) and in the corresponding index cases (sample-IDs in parenthesis, represented by large blue dots indicated with red arrow in the box plots). Red diamonds-Nonsense mutations; Red triangle-Splice mutation; Red dot: Missense mutation; Small blue dots- no mutation.

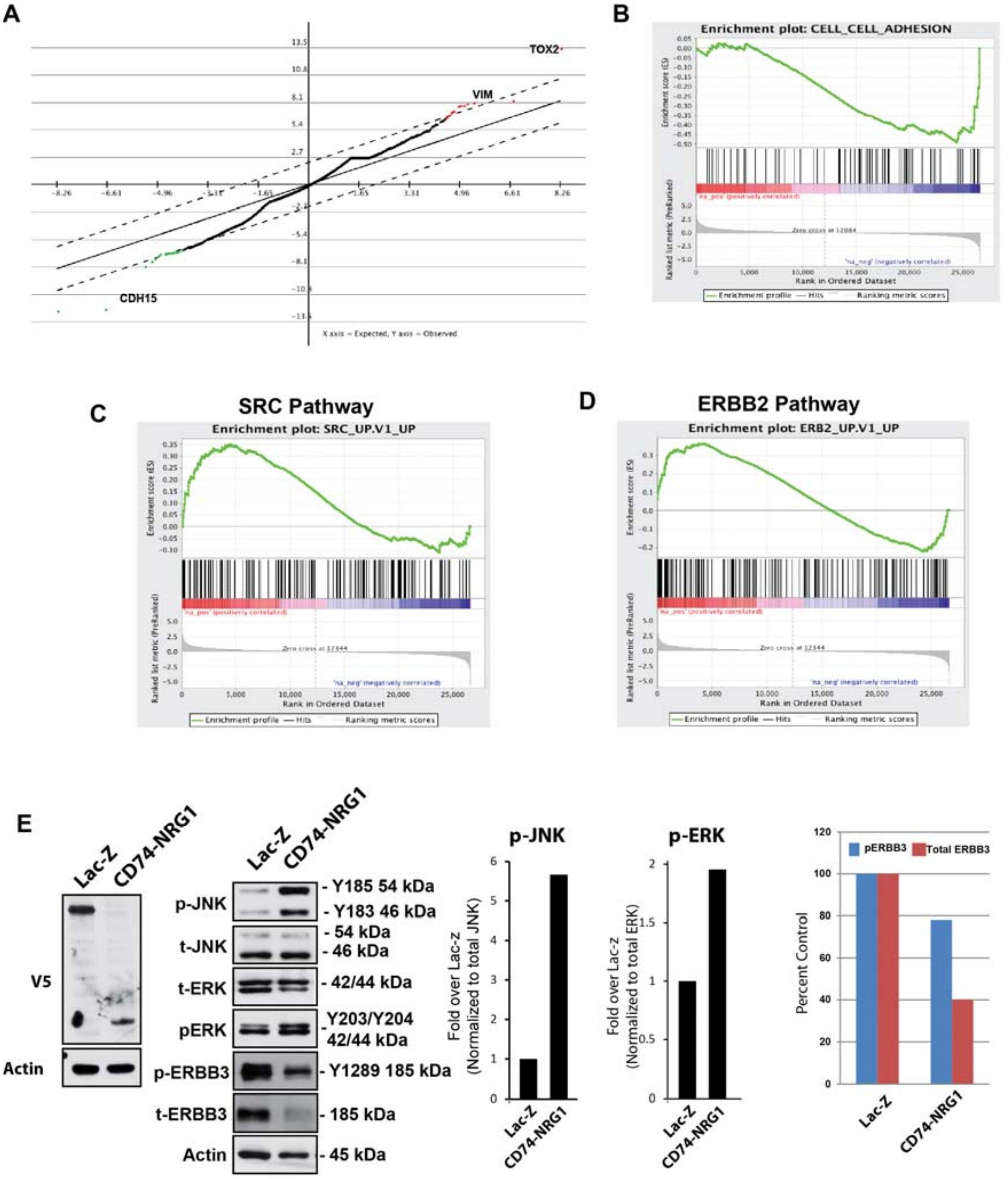
Supplementary Figure 7



**Supplementary Figure 7:** Functional Characterization of NRG1 fusion. **A**. Representative pictures of cells migrating to the basal side of the Boyden chamber membrane after Diff-Quick staining with an Olympus microscope at 20x magnification. White scale bar 100um **B**. Western blot analysis with V5epitope tag to monitor CD74-NRG1 fusion and control GAPDH protein expression in BEAS-2B stable cells. Blot images have been cropped for presentation; full size images are presented in Supplementary Figure 9. **C.** Representative pictures of BEAS-2B cells expressing the

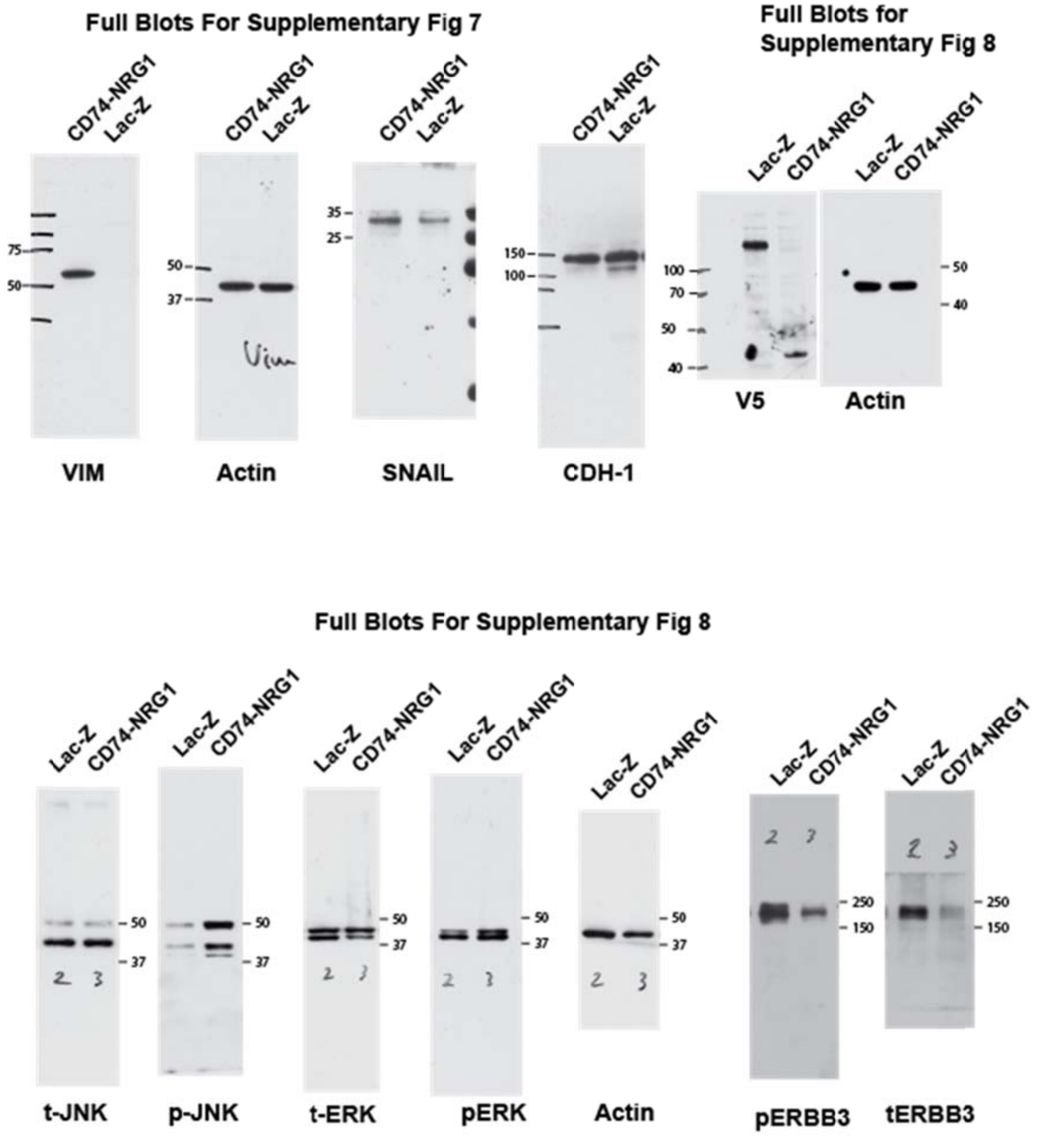
CD74-NRG1 fusion or Lac-Z. Cells expressing the CD74-NRG1 fusion appeared smaller and more fusiform as compared to Lac-Z, suggesting that they acquired a more mesenchymal phenotype. Boxed regions are enlarged sections on the right. White scale bar 200um **D**. Western blot analysis of E-cadherin (CDH-1) and Vimentin protein expression in transfected BEAS-2B cells. CD71-NRG1 transfected cells, showed a modest decrease in CDH-1 and a significant increase of Vimentin protein expression. Western Blot images have been cropped for presentation; full size images are presented in Supplementary Figure 9. E. Gene expression pattern for various epithelial mesenchymal transition markers from microarray data (log2 ratios). Western Blot images have been cropped for presentation; Full western blot images are presented in **Supplementary Fig 9.**

Supplementary Figure 8



**Supplementary Figure 8:** Gene Expression Analysis of *CD74-NRG1* Expressing Cells **A.** Differentially expressed genes identified by one class Significance Analysis of Microarrays (SAM). SAM-Plot represents the 35 genes with significant differential expression at 10% FDR (False Discovery Rate). **B, C and D**. Gene set enrichment analysis based on differentially-expressed genes among BEAS-2B cells transfected with the *CD74-NRG1* fusion or Lac-Z. Significant up-regulation of cell-cell adhesion, SRC and ERBB2 pathways was observed in CD74-NRG1 cells respectively. **E**. Western blot analysis to examine the levels of LacZ, CD74-NRG1, total and phosphorylated ERBB3, ERK and JNK1 proteins. Both lacZ and CD74-NRG1 fusion protein contains V5 epitope tag and their expression was evaluated by V5 antibody immunoblot. Barplots represent densitometric quantitation of phospho ERK, phospho JNK and ERBB3 protein levels. Western Blot images have been cropped for presentation; Full western blot images are presented in **Supplementary Fig 9.**

Supplementery Figure 9:



**Supplementary Figure 9:** Representative full Western blot images presented in Supplementary Figures 7 and 8.

**Supplementary Table 1:** Summary of Clinicopathological Characteristics for all Patients in the Combined Cohort.

| SAMPLES | | | | | | | |
|---|---|---|---|---|---|---|---|
| | **LUAD** | **LUSC** | **LUCL** | **Normal** | **LCLC** | **LACC** | **TOTAL** |
| **UMICH** | **67** | **36** | **24** | **6** | **9** | **11** | 153 |
| **SEOUL** | 79 | 0 | 0 | 0 | 0 | 0 | 79 |
| **TCGA** | 305 | 216 | 0 | 0 | 0 | 0 | 521 |
| **TOTAL** | 451 | 251 | 24 | 6 | 9 | 11 | **753** |

| GENDER | | |
|---|---|---|
| | **MALE** | **FEMALE** |
| **UMICH** | 64 | 58 |
| **SEOUL** | 48 | 31 |
| **TCGA** | 298 | 223 |
| **TOTAL** | 410 | 312 |

| FOLLOW UP TIME | | | | |
|---|---|---|---|---|
| | **MIN** | **MEDIAN** | **MAX** | **AVAILABLE** |
| **UMICH** | 0.26 | 4.64 | 17.37 | 111 |
| **SEOUL** | NA | NA | NA | 0 |
| **TCGA** | 0 | 0.92 | 18.66 | 436 |

| TUMOR STAGE | | | | |
|---|---|---|---|---|
| | **STAGE I** | **STAGE II** | **STAGE III** | **STAGE IV** |
| **UMICH** | 67 | 23 | 24 | 0 |
| **SEOUL** | NA | NA | NA | NA |
| **TCGA** | 250 | 112 | 101 | 19 |

| SMOKING STATUS | | | |
|---|---|---|---|
| | **NEVER SMOKER** | **LIGHT SMOKER** | **HEAVY SMOKER** |
| **UMICH** | 0 | 13 | 82 |
| **SEOUL** | NA | NA | NA |
| **TCGA** | 4 | 72 | 309 |

**Supplementary Table 2:** Non Synonymous Mutations in Lung Adenoid Cystic Carcinoma and Large Cell Lung Cancer Samples.

| S.No | Sample | Cancer Type | Cosmic Mutations | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Region | KRAS | NRAS | HRAS | BRAF | KIT | MET | TP53 | IDH1 | GNAS | ATM | NF1 |
| 1 | pt_lung_ACC_07-I-5699 | Lung Adenoid Cystic Carcinoma | Exonic | G12C | | | | | | | | | | |
| 2 | pt_lung_ACC_09-D-5737 | Lung Adenoid Cystic Carcinoma | Exonic | G12V | | | | | | | | | | |
| 3 | pt_lung_ACC_12-I-2857 | Lung Adenoid Cystic Carcinoma | Exonic | G12D | | | | | | | | | | |
| 4 | pt_lung_ACC_12-I-4664 | Lung Adenoid Cystic Carcinoma | Exonic | Q61H | | | | | | | | R186H, R187H | | |
| 5 | pt_lung_ACC_09-I-7040 | Lung Adenoid Cystic Carcinoma | Exonic | G13C | | | | M541L | | R141L | | | | |
| 6 | pt_lung_ACC_12-I-3344 | Pulmonary Met* | Exonic | | Q61R | | | M541L | T992I,T1010I | R141C | | | | |
| 7 | pt_lung_ACC_10-D-6174 | Lung Adenoid Cystic Carcinoma | Exonic | | | Q61L | | | | | | | | |
| 8 | pt_lung_ACC_09-I-5904 | Lung Adenoid Cystic Carcinoma | Exonic | | | | V600E | | | | | | | |
| 9 | pt_lung_ACC_11-T-230 | Lung Adenoid Cystic Carcinoma | Exonic | | | | | | | | V178I | | | |
| 10 | pt_lung_ACC_11-I-8842 | Lung Adenoid Cystic Carcinoma | No nominations | | | | | | | | | | | |
| 11 | pt_lung_ACC_12-I-318 | Lung Adenoid Cystic Carcinoma | No nominations | | | | | | | | | | | |
| | | | | | | | | | | | | | | |
| 12 | pt_lung_LC14 | Large Cell Lung Cancer | No nominations | | | | | | | | | | | |
| 13 | pt_lung_LC4 | Large Cell Lung Cancer | No nominations | | | | | | | | | | | |
| 14 | pt_lung_LC1 | Large Cell Lung Cancer | Exonic | | | | | | | C176S | | | | |
| 15 | pt_lung_L63 | Large Cell Lung Cancer | No nominations | | | | | | | | | | | |
| 16 | pt_lung_LC3 | Large Cell Lung Cancer | Exonic | R151G | | | | | | F113S | | P129L | K2530E, S85P | |
| 17 | pt_lung_LC13 | Large Cell Lung Cancer | No nominations | | | | | | | | | | | |
| 18 | pt_lung_LC2 | Large Cell Lung Cancer | Exonic | G12V | | | | | | | | | | |
| 19 | pt_lung_LC8 | Large Cell Lung Cancer | Exonic | R151G | | | | | | E166X | | | | Y1254H |
| 20 | pt_lung_LC9 | Large Cell Lung Cancer | Exonic | | | | | | | | | | | E149K |

* from Colon Adenocarcinoma

**Supplementary Table 3:** Comparison of the number of fusions among different tumor stages in LUAD and LUSC.

| LUAD | Stage I | Stage II | Stage III | Stage IV |
|---|---|---|---|---|
| Stage I | | 0.294 | 0.783 | 0.147 |
| Stage II | | | 0.219 | 0.049 |
| Stage III | | | | 0.201 |

| LUSC | Stage I | Stage II | Stage III | Stage IV |
|---|---|---|---|---|
| Stage I | | 0.014 | 0.726 | 0.093 |
| Stage II | | | 0.060 | 0.020 |
| Stage III | | | | 0.075 |

**Supplementary Table 4:** Univariate Cox regression for overall survival according to clinical variables (n = 621).

| | **Overall Survival** | | |
|---|---|---|---|
| | **HR** | **95% CI** | **p-value** |
| **Age, continuous** | 1.03 | 1.01 – 1.04 | < 0.001 |
| **Sex** | | | |
| Female | 1 | -- | |
| Male | 1.33 | 1.02 – 1.74 | 0.037 |
| **Stage, continuous** | 1.55 | 1.35 – 1.76 | < 0.001 |
| **Smoking status** | | | |
| Non-smoker | 1 | -- | |
| Smoker (<35 pack-year) | 1.31 | 0.52 – 3.30 | 0.565 |
| Smoker (≥35 pack-year) | 1.49 | 0.61 – 3.67 | 0.378 |
| **Histology** | | | |
| Adenocarcinoma | 1 | -- | |
| Squamous cell carcinoma | 0.99 | 0.76 – 1.29 | 0.989 |
| ***TP53* status** | | | |
| Wild-type | 1 | -- | |
| Mutant | 0.94 | 0.66 – 1.33 | 0.717 |
| ***KRAS* status** | | | |
| Wild-type | 1 | -- | |
| Mutant | 0.94 | 0.66 – 1.33 | 0.717 |
| ***EGFR* status** | | | |
| Wild-type | 1 | -- | |
| Mutant | 1.01 | 0.77 – 1.33 | 0.924 |

**Supplementary Table 5:** Multivariate Cox Regression for Overall Survival According to Number of Fusions in 621 NSCLC Patients Adjusted by Age, Gender and Stage.

| Covariates in the Model | | Hazard Ratio | 95% Confidence Intervals | p-value |
|---|---|---|---|---|
| Age, continuous | | 1.04 | 1.02 – 1.05 | <0.001 |
| Gender | Female | 1 | – | |
| | Male | 1.17 | 0.89 – 1.56 | 0.233 |
| Stage, continuous | | 1.63 | 1.42 – 1.87 | <0.001 |
| Number of Fusions | Low | 1 | | |
| | Intermediate | 1.13 | 0.79 – 1.61 | 0.484 |
| | High | 1.58 | 1.15 – 2.18 | 0.004 |

**Supplementary Table 6:** Multivariate Cox regression for overall survival according to number of fusions in 621 NSCLC patients adjusted by age, gender, stage and mutational status.

| Covariates in the Model | | Hazard Ratio | 95% Confidence Intervals | p-value |
|---|---|---|---|---|
| **Age, continuous** | | 1.04 | 1.02 – 1.06 | <0.001 |
| **Gender** | Female | 1 | – | |
| | Male | 1.22 | 0.96 – 1.85 | 0.082 |
| **Stage, continuous** | | 1.6 | 1.37 – 1.87 | <0.001 |
| **Smoking status** | Never-smoker | 1 | | |
| | Smoker (<35py) | 1.03 | 0.39 – 2.67 | 0.956 |
| | Smoker (≥35py) | 1.04 | 0.40 – 2.66 | 0.938 |
| **Histology** | LUAC | 1 | – | |
| | LUSC | 0.9 | 0.63 – 1.27 | 0.524 |
| **Number of fusions** | Low | 1 | | |
| | Intermediate | 1.34 | 0.91 – 2.08 | 0.125 |
| | High | 1.75 | 1.18 – 2.59 | 0.005 |
| ***KRAS* status** | WT | 1 | – | |
| | mutant | 1.17 | 0.76– 1.81 | 0.471 |
| ***EGFR* status** | WT | 1 | – | |
| | mutant | 1.32 | 0.87 – 2.01 | 0.183 |
| ***TP53* status** | WT | 1 | – | |
| | mutant | 1.2 | 0.86 – 1.69 | 0.385 |

**Supplementary Table 7:** NF1 Aberrations in NSCLC.

| S.No | Sample_ID | Cohort | Fusion partners | NF1_Aberration | NF1_Mutation | Other_Mutations | Fusion ORF |
|---|---|---|---|---|---|---|---|
| 1 | TCGA-43-6413 | TCGA | GOSR1-NF1 | Fusion | | | No |
| 2 | TCGA-69-7764 | TCGA | NLK-NF1 | Fusion | p.T2335P | | No |
| 3 | pt_lung_LS2 | UMICH | NF1-DRG2_AS | Fusion | | | No |
| 3 | pt_lung_LS2 | UMICH | NF1-MYO15A_AS | Fusion | | | No |
| 4 | TCGA-44-5644 | TCGA | NF1-PSMD11 | Fusion | p.H553Y | | Yes |
| | | | | | | | |

| S.No | Patient_UUID | Cohort | Gene | NF1_Aberration | NF1_mutation | Other_Mutations |
|---|---|---|---|---|---|---|
| 5 | 2c8877f9-ee7f-4216-97ad-d2939a13daa4 | TCGA | NF1 | Mutation | p.C904* | PIK3CA p.E545K; |
| 6 | 3c4ff061-d214-4d1c-8d2e-3034f207c252 | TCGA | NF1 | Mutation | p.E126* | cMET exon skipping 14 |
| 7 | 6bffe800-ec2b-4638-9333-97fe85dcd91c | TCGA | NF1 | Mutation | p.E1947*&p.W1831fs | |
| 8 | b4ff0f49-b787-48ec-91cc-ee26786ff1bf | TCGA | NF1 | Mutation | p.E2231* | |
| 9 | pt_lung_C115 | UMICH | NF1 | Mutation | p.E2306X&p.E2327X | |
| 10 | 66763a0c-6cda-4832-a0cc-e7b496d78eaa | TCGA | NF1 | Mutation | p.E2357* | |
| 11 | pt_lung_A70 | UMICH | NF1 | Mutation | p.E540X&p.S749X | |
| 12 | e16ca88f-488b-40f0-9169-e5a62482a2ff | TCGA | NF1 | Mutation | p.F1357fs | BRAF p.V600E; TP53 p.R196* |
| 13 | bcf2e591-9dae-440f-bd03-5f27c57db741 | TCGA | NF1 | Mutation | p.G2024* | BRAF p.G469V; |
| 14 | d721bfe0-90e3-415e-b9f3-1a270efa5fbb | TCGA | NF1 | Mutation | p.G751* | |
| 15 | e10568fe-0436-43f2-9f0f-48f9903868c4 | TCGA | NF1 | Mutation | p.L1267fs&p.L2538fs | |
| 16 | bf15f7ad-9d92-473b-91d1-f24aa373ab97 | TCGA | NF1 | Mutation | p.L2338H&p.S1938_splice | |
| 17 | e487c72f-2cb4-4a88-bd69-cd006d5b4c1a | TCGA | NF1 | Mutation | p.I2003_splice | |
| 18 | 81a0b2ff-a3d3-41bb-9ce6-765e6ae894af | TCGA | NF1 | Mutation | p.I396_splice | |
| 19 | f462cfef-f60a-4d3e-b92d-b8d8f50b6bb3 | TCGA | NF1 | Mutation | p.L2413fs | KRAS p.G12C; |
| 20 | 7f6455e8-fa3d-4452-acb2-8c9995073072 | TCGA | NF1 | Mutation | p.L494*&p.E1928fs | |
| 21 | c95957a7-1a1a-4c8d-bb61-7c99b500f224 | TCGA | NF1 | Mutation | p.P1463fs | TP53 p.R158fs; BRAF p.G469V |
| 22 | 591c068f-bbb1-4df2-9abb-d1a2e4a58372 | TCGA | NF1 | Mutation | p.P1951fs | |
| 23 | 4d687740-96ca-4d78-8c78-1a2024ce6b6c | TCGA | NF1 | Mutation | p.Q2492* | |
| 24 | 46592b7b-6968-42a6-83af-0917c9f4a9a5 | TCGA | NF1 | Mutation | p.R135fs | |
| 25 | ff9def3d-17e5-4ef6-b74e-933f11ed6f00 | TCGA | NF1 | Mutation | p.R1870_splice | PIK3CA p.N345K; KEAP1 p.P492fs; |
| 26 | bd3bf142-7c14-4538-8a76-3c6e140fa01a | TCGA | NF1 | Mutation | p.S2355_splice | |
| 27 | 42ca54fc-c1ae-41cd-bca1-7fe9810db460 | TCGA | NF1 | Mutation | p.T1273fs | BRAF p.L514P; TP53 p.V73fs |
| 28 | eeab558b-8d1e-4843-861d-dbfb06061758 | TCGA | NF1 | Mutation | p.T2565fs | EGFR p.L858R |
| 29 | 294cb595-0907-44c7-bbef-985a27c1e6e2 | TCGA | NF1 | Mutation | p.T317fs | |
| 30 | 8d0736fe-261c-445c-bfd2-a3ea3ceaf367 | TCGA | NF1 | Mutation | p.W2225* | |
| 31 | 90b02be3-5496-40a2-8c6e-460d2898aadb | TCGA | NF1 | Mutation | p.W336* | |
| 32 | 2e007464-f3f4-4eb2-bab8-91b8272c96d1 | TCGA | NF1 | Mutation | p.Y2285* | |
| 33 | 37c8d73a-45ae-40fc-ba9a-721b755c1160 | TCGA | NF1 | Mutation | p.Y2285fs | |

**Supplementary Table 8:** Lung Cancer Samples Harboring Fusions and/or Outlier Expression of NRG1.

| Patient_UUID | Cohort | Disease | Fusion Status | Oncogene Driven | Mutations | Expression Outlier Percentile | NRG1 Expression | Outlier NRG1 |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | FPKM | Score |
| pt_lung_A35 | UMICH | LUAD | YES | NO | TP53 p.P33R p.P72R; RBM10 p.A630P p.A696P; SMARCA4 p.R513W; APC p.V1822D p.V1804D; ATM p.N1983S | 99% | 29.08 | 0.93 |
| ju_lung_lc_s17 | SEOUL | LUAD | YES | NO | TP53 p.P33R | 99% | 33.08 | 0.93 |
| pt_lung_C028 | UMICH | LUAD | NO/TBD | NO | SMARCA4 p.E1056X; TP53 p.R248L p.R209L; APC p.V1822D p.V1804D; ATM p.N1983S | 99% | 83.92 | 0.93 |
| 0232d299-4cdf-4fd7-9a5e-8d13c208b40c | TCGA | LUAD | NO/TBD | NO | TP53 p.R156P; KEAP1 p.D236N; RBM10 p.S781L | 99% | 21.32 | 0.93 |
| 7b0622ab-63ea-483f-ae40-d3ea587bdbba | TCGA | LUAD | NO/TBD | NO | - | 99% | 25.86 | 0.93 |
| pt_lung_H1793 | UMICH | LUAD (LUCL) | NO/TBD | NO | SMARCA4 p.E514X; TP53 p.P33R p.R141H; APC p.V1822D p.E1991D; EGFR p.C311F; ATM p.N1983S | 99% | 281.86 | 10.13 |
| a3e1ac67-a1f2-44fb-8343-a7e8239fc24a | TCGA | LUSC | YES | NO | TP53 p.G244C; PIK3CA p.D1045V | 99% | 49.56 | 4.23 |
| ce8612ab-3149-4a6a-b424-29c0c21c9b8b | TCGA | LUSC | NO/TBD | NO | TP53 p.S314fs; CDKN2A p.P3fs; APC p.S966G; NF1 p.E1734V | 99% | 34.53 | 4.23 |
| 7e691df8-8ea6-472c-86bf-504c7ba6983d | TCGA | LUSC | NO/TBD | NO | APC p.S966G; CDKN2A p.P3fs; TP53 p.S314fs; NF1 p.E1734V | 99% | 49.33 | 4.23 |
| 791f1b21-695e-4db1-b41d-80590c09d257 | TCGA | LUSC | NO/TBD | NO | KEAP1 p.R320Q p.R470C; PIK3CA p.E453K | 99% | 31.24 | 4.23 |
| 14a4a93a-e24d-46f2-bee3-18bd792ef95a | TCGA | LUSC | NO/TBD | NO | TP53 p.E271* | 99% | 36.74 | 4.23 |
| 6394fe4a-6034-4c79-b28f-aa43e3753730 | TCGA | LUSC | NO/TBD | NO | - | 99% | 57.53 | 4.23 |
| 3351b902-9b7e-4b90-bf6b-bcf74be00bc1 | TCGA | LUSC | NO/TBD | NO | - | 99% | 32.85 | 4.23 |
| 48d-1296-44aa-b7b1-0795939 | TCGA | LUSC | Yes | NA | NA | 99% | 383.00 | ND |

**Supplementary Table 9:** Primer Sequences

| Primer | 5' Gene | 3' Gene | Primer ID | Primer Sequence (5'->3') | Length | Product Size |
|---|---|---|---|---|---|---|
| 1 | AHNAK | KAT5 | LF_77F | CTGCCAGACCCGCCCGGAAC | 20 | 152 |
| 2 | AHNAK | KAT5 | LF_77R | AGTCATGCGTGGTGCTGACGG | 21 | |
| 3 | AIM1L | ZNF683 | LF_57F | ACCTACAGCGGCACCCAGAGG | 21 | 163 |
| 4 | AIM1L | ZNF683 | LF_57R | GCCCCCTCGCCAGCTCTTTCT | 21 | |
| 5 | ANKRD11 | FANCA | LF_53F | GCAGCCCTCGGAGCACGAAT | 20 | 157 |
| 6 | ANKRD11 | FANCA | LF_53R | TGTGCGGCCACCAAAGACCA | 20 | |
| 7 | ATF6B | MUC5B | LF_84F | TGCTCCCAGCCATCAGCCACA | 21 | 110 |
| 8 | ATF6B | MUC5B | LF_84R | GCCGGCTCGGTCGGTCTTATTG | 22 | |
| 9 | C1ORF194 | UQCR10 | LF_118F | TACTTAGGAGAACTGCGGGC | 20 | 104 |
| 10 | C1ORF194 | UQCR10 | LF_118R | CGGAACAGCAGGGAGTACAA | 20 | |
| 11 | CD74 | NRG1 | LF_5F | CTGGATGCACCATTGGCTCCTGT | 23 | 110 |
| 12 | CD74 | NRG1 | LF_5R | GATGGCTTGTCCCAGTGGTGG | 21 | |
| 13 | CDK9 | AHCY | LF_1F | GAACAGCCAGCCCAACCGCTA | 21 | 145 |
| 14 | CDK9 | AHCY | LF_1R | ACGCATCAGGCCCGGCATCT | 20 | |
| 15 | CHST11 | TXNRD1 | LF_108F | CCAGGACAAAGCCATGAAGC | 20 | 159 |
| 16 | CHST11 | TXNRD1 | LF_108R | GCTGGGTTCTCTGGCAAAGT | 20 | |
| 17 | CHST11 | TXNRD1 | LF_21F | CATCCGCCGCAAGCCTCTCT | 20 | 148 |
| 18 | CHST11 | TXNRD1 | LF_21R | ACGGGAGCCTCTGACGACCA | 20 | |
| 19 | COX10 | PEMT | LF_100F | CATTGGCTCCGGGCCCTTTTG | 21 | 139 |
| 20 | COX10 | PEMT | LF_100R | CCGAAGGCCCTGCTCAGCTTG | 21 | |
| 21 | CPSF6 | TSPAN11 | LF_81F | CGCGGATGTCGGCGAAGAGTTC | 22 | 158 |
| 22 | CPSF6 | TSPAN11 | LF_81R | CAGATGCCCACAGCCAGGACG | 21 | |
| 23 | CPT1A | HRASLS2 | LF_88F | GCACGAGCCCAGACGCCTTT | 20 | 160 |
| 24 | CPT1A | HRASLS2 | LF_88R | GCCGGAGCCAGATGGACCAC | 20 | |
| 25 | CYP24A1 | C9ORF3 | LF_115F | GTCCGCAAATACGACATCCA | 20 | 193 |
| 26 | CYP24A1 | C9ORF3 | LF_115R | GATGTCCAGGGTCAGTTCGAG | 21 | |
| 27 | DAPK1 | GMDS | LF_75F | AGCGGAGCTGAAGTGCCCTG | 20 | 152 |
| 28 | DAPK1 | GMDS | LF_75R | CTCCGTCAACGTCCGCAGTGT | 21 | |
| 29 | EIF2AK2 | SULT6B1 | LF_26F | ACTGCCTAATTCAGGACCTCCACA | 24 | 262 |
| 30 | EIF2AK2 | SULT6B1 | LF_26R | CGCACGCATGGCTTGGAAGG | 20 | |
| 31 | ESRP1 | DOCK8 | LF_76F | CACAGCCTGGCACGGTGGTC | 20 | 192 |
| 32 | ESRP1 | DOCK8 | LF_76R | TGCGGGCGTGTCCGGTTTTC | 20 | |
| 33 | FAM60A | DPF3 | LF_19F | TTCCCGGCCAGCGGTAGCAA | 20 | 195 |
| 34 | FAM60A | DPF3 | LF_19R | TCCGGCAGTGCTCAATGGCT | 20 | |
| 35 | FGFR3 | TACC3 | LF_73F | AGCTACGGGGTGGGCTTCTTCC | 22 | 172 |
| 36 | FGFR3 | TACC3 | LF_73R | CGGACGTCCTGAGGGAGTCTCA | 22 | |
| 37 | GTF2E2 | GSR | LF_17F | ACACGGCATCAGCGAGGAGA | 20 | 158 |
| 38 | GTF2E2 | GSR | LF_17R | CCCTGCAGCATTTCATCACACCCA | 24 | |
| 39 | HLTF | HPS3 | LF_97F | ACGGCCATTGCAGTAATCCTTACCA | 25 | 143 |
| 40 | HLTF | HPS3 | LF_97R | TGGGGCACTTGCTTTGGCTCA | 21 | |
| 41 | IP6K1 | TRAIP | LF_3F | GGGAGCAACCTCGGCGCAA | 19 | 145 |
| 42 | IP6K1 | TRAIP | LF_3R | GGCCGGCGGAGCTTCAGATT | 20 | |
| 43 | ITSN1 | ENOX1 | LF_60F | GGCTCCTGCGTCCCTCCCAG | 20 | 219 |
| 44 | ITSN1 | ENOX1 | LF_60R | TGAACATGCGTGGCAGCCTCA | 21 | |
| 45 | JPH1 | NCOA2 | LF_27F | GGTGGACAGAGCAATTGAAGGCG | 23 | 166 |
| 46 | JPH1 | NCOA2 | LF_27R | TCCCATCCCACTCATCTTGAACACA | 25 | |
| 47 | MAPKAPK | ACAD10 | LF_83F | GCTCTGCGGCACTGTCACTT | 20 | 256 |
| 48 | MAPKAPK | ACAD10 | LF_83R | ACTGCGAAATCCCACGCCAGG | 21 | |
| 49 | MEAF6 | SCMH1 | LF_32F | AGGAAGCTGAGCGGCTCTTCAGT | 23 | 196 |
| 50 | MEAF6 | SCMH1 | LF_32R | GGCGATGGTGGCTCCTTGTGG | 21 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 51 | MRC2 | MAP3K3 | LF_15F | GCCTCGTCACCTGCTGCGCT | 20 | 149 |
| 52 | MRC2 | MAP3K3 | LF_15R | GACGTCGGTTCATCTGGAGGGC | 22 | |
| 53 | MYO5C | TNFAIP8L | LF_55F | GCATCCGTCATGAAGTTACCAGGC | 24 | 177 |
| 54 | MYO5C | TNFAIP8L | LF_55R | GGGCTTGAAGCGCAAGACTCTTTGA | 25 | |
| 55 | NFAT5 | VPS4A | LF_37F | AGCGCGGACCTAGACCTGGA | 20 | 100 |
| 56 | NFAT5 | VPS4A | LF_37R | ACTGCACGCACTTGGCTCGAA | 21 | |
| 57 | NFAT5 | VPS4A | LF_38F | CAGCGCGGACCTAGACCTGGA | 21 | 155 |
| 58 | NFAT5 | VPS4A | LF_38R | ACTGCACGCACTTGGCTCGAA | 21 | |
| 59 | NUSAP1 | EIF2AK4 | LF_56F | TGAGTCATCCAAACCTGGAA | 20 | 247 |
| 60 | NUSAP1 | EIF2AK4 | LF_56R | TCGCTGAGAAATGACTGCAC | 20 | |
| 61 | PCMT1 | LATS1 | LF_22F | AGCAACAATCAGTGCTCCACACA | 23 | 139 |
| 62 | PCMT1 | LATS1 | LF_22R | TGCTGCAGCCATCTGCTCTCG | 21 | |
| 63 | PPIG | UBR3 | LF_79F | GGAGGCGGTTAGCGGGCTTT | 20 | 151 |
| 64 | PPIG | UBR3 | LF_79R | ACGTTGGCAGAAGTCCCAGGC | 21 | |
| 65 | PTCH1 | AM120AO | LF_58F | TTTGGGGCCTTCGCGGTGGG | 20 | 112 |
| 66 | PTCH1 | AM120AO | LF_58R | GCCACAGCCTGTCGGTTGCAT | 21 | |
| 67 | PTPRD | LRMP | LF_7F | GCGGCTGCTTTAGTGAAGAAGTGAA | 25 | 160 |
| 68 | PTPRD | LRMP | LF_7R | ACGCGTTCAACACCATTCTCCA | 22 | |
| 69 | R3HDM2 | NFE2 | LF_106F | GACTCATGGAGGCTGAGCATT | 21 | 175 |
| 70 | R3HDM2 | NFE2 | LF_106R | TCTCCTGCCAAGTCAGTTCC | 20 | |
| 71 | R3HDM2 | NFE2 | LF_107F | CCCTTTTCTTCCCCTCTCC | 19 | 249 |
| 72 | R3HDM2 | NFE2 | LF_107R | GGAAAGCCCAGATGGCTCTA | 20 | |
| 73 | R3HDM2 | ARHGAP9 | LF_123F | CCGCCAAGGCCCGTGCGAG | 19 | 424 |
| 74 | R3HDM2 | ARHGAP9 | LF_123R | GCCATCTGCCCCAGTATAAG | 20 | |
| 75 | R3HDM2 | ARHGAP9 | LF_94F | CTTCCCAAGCCCCTTTCC | 18 | 233 |
| 76 | R3HDM2 | ARHGAP9 | LF_94R | ACCGGCTGGATAGCATTGTA | 20 | |
| 77 | RAF1 | TMEM40 | LF_87F | TGACCCAGTGGTGCGAGGGC | 20 | 152 |
| 78 | RAF1 | TMEM40 | LF_87R | TGAGGCTGGGAGGAGGATGCTG | 22 | |
| 79 | RARA | TCAP | LF_113F | TCCTGAATCGAGCTGAGAGG | 20 | 109 |
| 80 | RARA | TCAP | LF_113R | CAGCTCTGAGGTAGCCATGA | 20 | |
| 81 | RARA | TCAP | LF_114F | ATCGAGCTGAGAGGGCTTCC | 20 | 245 |
| 82 | RARA | TCAP | LF_114R | GCTGGTGGTAGGTCTCATGTC | 21 | |
| 83 | RARA | TCAP | LF_25F | GCTGAGAGGGCTTCCCCGGTT | 21 | 172 |
| 84 | RARA | TCAP | LF_25R | GCCCATCCGCATCATCAGCCA | 21 | |
| 85 | RBM12B | MMP16 | LF_62F | CCGGCCTTGTGTGTCCGACT | 20 | 254 |
| 86 | RBM12B | MMP16 | LF_62R | GAGCGGTGTGGGGGCACTGT | 20 | |
| 87 | RUNX1 | PTPRR | LF_13F | GCTGAGAAATGCTACCGCAGCCA | 23 | 144 |
| 88 | RUNX1 | PTPRR | LF_13R | ACCGGCTTCCCACTCTTCTTCTGA | 24 | |
| 89 | SLC12A7 | TERT | LF_2F | GCCTACGCCAGACAAGGTGCAG | 22 | 159 |
| 90 | SLC12A7 | TERT | LF_2R | TCTGCTTCCGACAGCTCCCGC | 21 | |
| 91 | SLC37A1 | TIAM1 | LF_122F | CTCGGCAACTGGTTTGGAA | 19 | 286 |
| 92 | SLC37A1 | TIAM1 | LF_122R | ACCATATGACCGTCAGGCTTC | 21 | |
| 93 | SLC37A1 | TIAM1 | LF_82F | GGGGCCTGTCCTTCGTCGTG | 20 | 125 |
| 94 | SLC37A1 | TIAM1 | LF_82R | GCGACCATCAACCGTCACCAGG | 22 | |
| 95 | SLC9A7 | VDR | LF_4F | GCTCACGCTCACCATCCTCACC | 22 | 160 |
| 96 | SLC9A7 | VDR | LF_4R | AGCAGGGGGCAGGTAAGTGGA | 21 | |
| 97 | SMARCB1 | BCL2L13 | LF_67F | TGGGCAGAAGCTGCGAGACG | 20 | 137 |
| 98 | SMARCB1 | BCL2L13 | LF_67R | CCTGGAACACACAGCGCCTGG | 21 | |
| 99 | SRGAP1 | MSRB3 | LF_20F | AGCAAAGACCATGCAACCTTGAGT | 24 | 123 |
| 100 | SRGAP1 | MSRB3 | LF_20R | ACGAGAAGCAAAGAACAGGGGCA | 23 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 101 | SRSF7 | SOS1 | LF_34F | ACCTGCCCGACGTCCCTTTG | 20 | 156 |
| 102 | SRSF7 | SOS1 | LF_34R | GGACCTCAGGGTTGCCTTCTTCTG | 24 | |
| 103 | TFG | GPR128 | LF_78F | TGCAACGAGTTTTCAGAGGA | 20 | 231 |
| 104 | TFG | GPR128 | LF_78R | GTAGGGGTGCTTGATGAGGA | 20 | |
| 105 | THADA | MTA3 | LF_43F | TGGGAGTCAGCCAGGAAGGTGT | 22 | 176 |
| 106 | THADA | MTA3 | LF_43R | GCTTCAGAGGCTGACCATTCCTCC | 24 | |
| 107 | TMEM131 | ARHGAP1 | LF_121F | TCACGAAATGCCCAGAAAACA | 21 | 206 |
| 108 | TMEM131 | ARHGAP1 | LF_121R | ACTTTGTGCAGATGAGAGCCA | 21 | |
| 109 | TMEM131 | ARHGAP1 | LF_80F | GGCAACACCAGTAGCTCAGAGGG | 23 | 95 |
| 110 | TMEM131 | ARHGAP1 | LF_80R | TGCAGTCTGAACAAGCTGCCAGG | 23 | |
| 111 | TNFRSF14 | IGHM | LF_102F | CTGACCCACAGACTCTGCAC | 20 | 252 |
| 112 | TNFRSF14 | IGHM | LF_102R | GGGAATTCTCACAGGAGACG | 20 | |
| 113 | TP53 | SAT2 | LF_24F | GGCTCCGGGGACACTTTGCG | 20 | 121 |
| 114 | TP53 | SAT2 | LF_24R | GCCAGCGAGGCGATCCTCTG | 20 | |
| 115 | TP53 | GLP2R | LF_31F | TTTGCGTTCGGGCTGGGAGC | 20 | 149 |
| 116 | TP53 | GLP2R | LF_31R | TGGCGGGCTAGCAAGAAGCG | 20 | |
| 117 | TSC1 | SMARCA4 | LF_30F | AAGCCAATGATGGAGCATGTGCG | 23 | 121 |
| 118 | TSC1 | SMARCA4 | LF_30R | ACCTTCACCGGGAGGTCGCT | 20 | |
| 119 | TTC1 | DOCK2 | LF_101F | AGGAGGAGCCAGGAGCGGAC | 20 | 177 |
| 120 | TTC1 | DOCK2 | LF_101R | AGTAGTGCTGGTCACCCATCTGGT | 24 | |
| 121 | UBA5 | MRAS | LF_99F | TTGCAGGAAGCAGCAGGAGGAA | 22 | 151 |
| 122 | UBA5 | MRAS | LF_99R | TTGTCACTGGGGACGGCGCT | 20 | |
| 123 | WASF2 | FGR | LF_11F | CGGGAGCACACTCTGTGCGGA | 21 | 108 |
| 124 | WASF2 | FGR | LF_11R | GCTGCGGCATGATCCTTGGGA | 21 | |
| 125 | ZNF544 | ZNF17 | LF_46F | CACCAGGCAGGTGACGCCTA | 20 | 189 |
| 126 | ZNF544 | ZNF17 | LF_46R | ACATTGCTTGGAAGGTGCCTCCTC | 24 | |
| 127 | ZNF664 | WSB2 | LF_47F | CGCCGGACGCCTCCATTGTT | 20 | 181 |
| 128 | ZNF664 | WSB2 | LF_47R | CCGGGCTTGAGTTCGGCCAG | 20 | |
| 129 | ZNF667 | OC1001282 | LF_69F | CGGAACCTGGTCTCGCTTGGT | 21 | 150 |
| 130 | ZNF667 | OC1001282 | LF_69R | GCTCTCCTGCGATCATTCGCCA | 22 | |
| 131 | ZNF704 | MYC | LF_14F | CCAGACGACGGCATCGACGAG | 21 | 151 |
| 132 | ZNF704 | MYC | LF_14R | ACGGCTGCACCGAGTCGTAG | 20 | |
| 133 | ZSWIM4 | RFX1 | LF_28F | CCTGAGCCCCCACTGCAAACC | 21 | 174 |
| 134 | ZSWIM4 | RFX1 | LF_28R | ACTGTCTCGCTGGCCCGCAT | 20 | |
| 135 | GAPDH | | | CTCTGCTCCTCCTGTTCGAC | 20 | 112 |
| 136 | GAPDH | | | ACGACCAAATCCGTTGACTC | 20 | |
| 137 | ACTB F | | | CTCTTCCAGCCTTCCTTCCT | 20 | 116 |
| 138 | ACTB R | | | AGCACTGTGTTGGCGTACAG | 20 | |

**Supplementary Data 1:** Clinicopathological Characteristics of the Combined Lung Cohort Used in this Study.

**Supplementary Data 2:** Fusions Recovered by the Classifier in the ju_lung Cohort.

**Supplementary Data 3:** Fusions Recovered by the Classifier in UMICH Cohort.

**Supplementary Data 4:** Lung Fusions Candidates after Classification.

**Supplementary Data 5:** Table with Recurrence of Known Fusions across the Full Cohort and in Samples with Unknown Drivers.

**Supplementary Data 6:** Fusions found in the HIPPO Pathway in NSCLC.

**Supplementary Data 7**: Differentially expressed genes in BEAS-2B cells expressing CD74-NRG1 fusion protein versus LacZ.

**Supplementary Data 8:** Fusions Used as True Positives for the Random Forest Classifier.

**Supplementary Data 9:** List of gene fusions identified by TOPHAT analysis in normal samples

# Supplementary Methods:

## Bioinformatics Methods

### Sequence Alignment

Sequence alignment was performed using the Tuxedo pipeline: Bowtie2 (Bowtie2/2.0.2) and Tophat2 (TopHat/2.0.6)[1, 2]. We supplied TopHat with the set of transcript models annotated in the *Homo sapiens* ensemble database version 69. The flag fr-firststrand was used for the strand specific RNASeq libraries while fr-unstranded was used for the unstranded libraries. All other parameters were used with default values.

### Fusion Calling

Fusion calling was performed with TopHat-fusion1 (THF) on the UMICH, TCGA and SEOUL cohorts. TopHat-fusion was run with the following arguments: bowtie1, fusion-search, keep-fasta-order, no-coverage-search, fusion-min-dist=0, fusion-anchor-length=13, fusion-ignore-chromosomes=chrM. TopHat post-processing was run with the arguments: skip-blast, num-fusion-reads=1, num-fusion-pairs=1, num-fusion-both=3.

### Fusion Annotation and Lung Cancer Fusion Database

A database of fusions in lung cancers was developed, and for each fusion structural and functional annotation was recorded. The structural information correspond to chromosome number of 3' and 5' partner genes, cohort, 3' and 5' chromosome location, 3' breaking exon, 5' breaking exons, median alignment quality of reads that support 3' gene, median alignment quality of reads that support 5' gene, number of spanning reads, spanning mate pairs and encompassing reads, 3' and 5' partner recurrence across the cohort and fusion type (Inter-chromosomal, Intra-chromosomal, Tandem-duplication).

The functional annotation corresponds to kinase status, oncogene status, tumor suppressor status and targetable status (TRUE/FALSE) of both 3' and 5' partner genes. Other functional annotations include the gene family of both fusion genes, as well as the gene biotype (protein-coding, ncRNA, rRNA, etc.). Moreover, the gene expression of each fusion gene was calculated in fragments per-kilobase per million (FPKM) using Cufflinks3 and stored in the database. In addition, an outlier score was calculated for the expression of both 5' and 3' partners in order to identify cases in which the 3' partner is highly expressed as consequence of the fusion event.

This database was created using pytables and hd5 format for fast access and storage and includes the following tables: patient table, patient clinical information table, fusions structural information table and expression table. In addition to these tables corresponding to fusion events, we create and additional table to store the mutation status for each patient, mutation table. The mutation table allows us to classify each patient as “driver positive” or “driver negative” according to mutation status of well-known cancer related genes (see below).

## Mutation Calling

UMICH cohort: Single nucleotide variants (SNVs) were called using Varscan2 (Varscan2/2.2.8)[3] on the ssRNAseq libraries of the UMICH cohort. Because, we did not have matched normal for each tumor sample, we consider only SNVs that were previously reported in the Catalogue of Somatic Mutations database (COSMIC version 56). Single nucleotide mutations in other positions were not considered for reporting or downstream analysis. SNVs present in dbSNP (v135) were filtered out, as well as SNVs with variant fraction smaller than 10%, or with less than six reads covering the position. Insertions and deletions were not called from the RNAseq data, because currently there are no available algorithms to efficiently assess these genetic aberrations on RNASeq libraries. SNVs for all tumor samples were aggregated and annotated using variant-tools[4]. TCGA cohort: All somatic mutations both SNVs and indels called on Exome sequencing data for the TCGA consortium were extracted from aggregated Mutation Annotation Format (MAF) files available at the Broad institute firehose Genome Data Analysis Center MAF dashboard on May 11 of 2013. SEOUL cohort: All SNV and insertion/deletion somatic mutations reported by Seo et al (2012) were used [5].

## Sample Annotation

We annotated the mutation status of well characterized oncogenes and tumor suppressors known to be involved in lung adenocarcinoma and squamous carcinoma. We considered activating mutations for *KRAS, NRAS, HRAS, EGFR, BRAF, PIK3CA, MET*, and missense or non-sense mutations for *TP53, STK11, NF1, PTEN, SMARCA4, CDKN2A,* and *APC* genes. Mutations reported in COSMIC were considered for *AKT, MEK, ATM, AKT1, KEAP1, U2AF1, RBM10, ARID10*, and *MYC* which have been recently implicated on these indications[6, 7]. Finally, we used the somatic mutation information to divide the combined cohort in two groups: samples with known drivers and samples of unknown drivers. The first group corresponds to samples with somatic mutations in *KRAS, NRAS, HRAS, EGFR, BRAF* and/or *PIK3CA*, while the second group to samples that do not harbor alterations in those well-known driver genes.

## Fusions Classifier Training

First, all fusions present in normal samples were considered false positives and filtered out **Supplementary Data 9**. For the classification step, we trained a random forest classifier with 10000 trees using the following features: chromosomes of 3 and 5’ genes, 3’ gene, 5’ gene, 3’ breaking exon, 5’ breaking exons, median alignment quality of reads that support 3’ gene, median alignment quality of reads that support 5’ gene, number of spanning reads, spanning mate pairs and encompassing reads, 3’ and 5’ partner recurrence, fusion type, gene biotype of both 3’ and 5’ genes, FPKM expression of both 3’ and 5’ genes, and FPKM expression of both 3’ and 5’ genes normalized across the combined cohort.

## Experimental Methods

### RNASeq Library Preparation

Transcriptome libraries were prepared following a modified protocol previously described for generating strand specific RNASeq libraries [8]. Briefly 2.5 µg of total RNA was subjected to polyA selection using oligodT beads (Invitrogen, Carlsbad, CA). Purified polyA RNA was fragmented and reverse transcribed using Superscipt-II (Invitrogen, Carlsbad CA). Second strand synthesis was performed with DNA Ploymerase I (New England Biolabs, Ipswich, MA) in the presence of dNTP mix containing dUTP instead of dTTP. The product was then subjected to end repair, A base addition and adaptor ligation steps. Libraries were next size selected in the range of 350 bps after resolving in a 3% Nusieve 3:1 (Lonza, Basel, Switzerland) agarose gel and DNA recovered using QIAEX II gel extraction reagent (Qiagen, Valencia, CA). Libraries were barcoded during the 14-cycle PCR amplification with Phusion DNA polymerase (New England Biolabs, Ipswich, MA) and purified using AMPure XP beads (Beckman Coulter, Brea, CA).

### Quantitative RT-PCR and PCR Fusion Validation

Complimentary DNA was synthesized from total RNA using Superscript III in presence of random primers (Invitrogen, Carlsbad, CA). Quantitative Real-time PCR (qPCR) was performed using SYBR Green Master mix on the StepOne Real-Time PCR System (Applied Biosystems). All oligonucleotide primers for the qPCR assays were obtained from Integrated DNA Technologies (Coralville, IA); *NRG1* forward 5'GATTCCTACCGAGACTCTCCTC3' and reverse 5'TGGAAGGCATGGACACCGTCAT3' and *GAPDH* forward 5'GTCTCCTCTGACTTCAACAGCG3' and reverse primer 5'ACCACCCTGTTGCTGTAGCCAA3'. Fold changes were calculated relative to *GAPDH* and normalized to the non-target control sample.

We validated a subset of nominated fusion genes by THF from UMICH cohort using real-time RT-PCR. Subsequently the products were resolved in a 2% agarose gel electrophoresis for size evaluation. Of the 29 attempted fusions, 28 were validated, representing a validation rate of 96.6% (**Supplementary Data 4**). The primer sequences for PCR fusion validation and PCR product sizes are listed in **Supplementary Table 9**.

### siRNA knockdown studies

Lung cancer cell line NCI-H1793 were plated in 6-well plates at a desired numbers and transfected with 2nmol of *NRG1* siRNAs (J-004608-11; and J-004608-12) or non-target control siRNA (Thermo Scientific). Transfection with Oligofectamine reagent (Invitrogen, Carlsbad, CA) was performed twice over a period of 48 hours. Knockdown efficiency was determined by qPCR. For cell proliferation assay 24 hours after transfection, cells were trypsinized and plated in triplicate at 8,000 cells per well in 24-well plates. The plates were incubated in the IncuCyte live-cell imaging system

(Essen Biosciences) at 37°C under 5% CO2 atmosphere. Cell proliferation rate was assessed by kinetic imaging confluence measurements at 3-hour time intervals.

### Protein Isolation and Western Blot Analysis

Cells were washed with ice cold PBS twice and harvested using cell lysis buffer (Cell Signaling). Protein concentrations were estimated (bicinchoninic protein assay, Pierce, Rockford, IL) and equal amounts were resolved under reducing conditions by 10% SDS-PAGE. The protein was transferred to PDVF membranes, blocked in 5% milk tris-buffered saline (TBS) containing 0.1% Tween 20 and incubated with respective antibodies against for overnight at 4°C. The membrane were washed three times with 0.1% Tween 20 - TBS and further incubated for 60 minutes with secondary HRP anti-rabbit IgG used at 1:2000 dilution. Antibodies against E-Cadherin, Vimentin, phospho-Erbb3, phospho-Erbb3, phosho-ERK and total-ERK were purchased from Cell Signaling Technology Inc. (Beverly, MA). Total Erbb3 and Erbb4 were purchased from Santa Cruz Biotechnology Inc. (Dallas, TX). The membrane-bound peroxidase activity was detected using ECL Prime Western Blotting Detection kits (Amersham, Arlington Heights, IL) and the chemiluminescence signal were captured by exposing to autoradiographic films.

## Supplementary References:

1. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14**, R36 (2013).

2. Trapnell C*, et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* **7**, 562-578 (2012).

3. Koboldt DC*, et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research* **22**, 568-576 (2012).

4. San Lucas FA, Wang G, Scheet P, Peng B. Integrated annotation and analysis of genetic variants from next-generation sequencing studies with variant tools. *Bioinformatics* **28**, 421-422 (2012).

5. Seo JS*, et al.* The transcriptional landscape and mutational profile of lung adenocarcinoma. *Genome Research* **22**, 2109-2119 (2012).

6. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525 (2012).

7. Weir BA*, et al.* Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**, 893-898 (2007).

8. Levin JZ*, et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods* **7**, 709-715 (2010).