Published in final edited form as:
Cell. 2012 September 14; 150(6): 1107–1120. doi:10.1016/j.cell.2012.08.029.

Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing

Marcin Imielinski[1,2,3,4,*], **Alice H. Berger**[1,4,*], **Peter S. Hammerman**[1,4,*], **Bryan Hernandez**[1,4,*], **Trevor J. Pugh**[1,4,*], **Eran Hodis**[1], **Jeonghee Cho**[5], **James Suh**[6], **Marzia Capelletti**[4], **Andrey Sivachenko**[1], **Carrie Sougnez**[1], **Daniel Auclair**[1], **Michael Lawrence**[1], **Petar Stojanov**[1,4], **Kristian Cibulskis**[1], **Kyusam Choi**[5], **Luc de Waal**[1,4], **Tanaz Sharifnia**[1,4], **Angela Brooks**[1,4], **Heidi Greulich**[1,4], **Shantanu Banerji**[1,4], **Thomas Zander**[7,8], **Danila Seidel**[7], **Frauke Leenders**[7], **Sascha Ansén**[7], **Corinna Ludwig**[7], **Walburga Engel-Riedel**[7], **Erich Stoelben**[7], **Jürgen Wolf**[7], **Chandra Goparju**[9], **Kristin Thompson**[1], **Wendy Winckler**[1], **David Kwiatkowski**[4], **Bruce E. Johnson**[4], **Pasi A. Jänne**[4], **Vincent A. Miller**[10], **William Pao**[11], **William D. Travis**[12], **Harvey Pass**[9], **Stacey Gabriel**[1], **Eric Lander**[1,13,14], **Roman K. Thomas**[7,8,15,16,17], **Levi A. Garraway**[1,4], **Gad Getz**[1], and **Matthew Meyerson**[1,3,4]

[1]Cancer Program, The Broad Institute of Harvard and M.I.T., 7 Cambridge Center, Cambridge, MA, 02142, USA

[2]Department of Pathology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA

[3]Department of Pathology, Harvard Medical School, 44 Binney Street, Boston, MA 02115, USA

[4]Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA, 02115, USA

[5] Samsung Research Institute, Samsung Medical Center, Seoul, Republic of Korea

[6]Department of Pathology, Langone Medical Center, New York University, New York, NY, 10016, USA

[7]Department of Internal Medicine and Center for Integrated Oncology Köln-Bonn, University of Cologne, 50924, Cologne, Germany

[8]Max Planck Institute for Neurological Research, 50924, Cologne, Germany

[9]Department of Cardiothoracic Surgery, Langone Medical Center, New York University, New York, NY, 10016. USA

[10]Thoracic Oncology Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA

[11]Division of Hematology/Oncology, Vanderbilt-Ingram Cancer Center, Nashville, TN, 37232, USA

[12]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, 10065, USA

[13]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA, 02142, USA

Address correspondence to: Matthew Meyerson (matthew_meyerson@dfci.harvard.edu).
*These authors contributed equally

[14]Department of Systems Biology, Harvard Medical School, 44 Binney Street, Boston, MA 02115

[15]Department of Translational Genomics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany

[16]Laboratory of Translational Cancer Genomics, Köln – Bonn, University of Cologne, 50924 Cologne, Germany

[17]Department of Pathology, University of Cologne, Kerpener Str. 62, 50937 Cologne, Germany

SUMMARY

Lung adenocarcinoma, the most common subtype of non-small cell lung cancer, is responsible for over 500,000 deaths per year worldwide. Here, we report exome and genome sequences of 183 lung adenocarcinoma tumor/normal DNA pairs. These analyses revealed a mean exonic somatic mutation rate of 12.0 events/megabase and identified the majority of genes previously reported as significantly mutated in lung adenocarcinoma. In addition, we identified statistically recurrent somatic mutations in the splicing factor gene *U2AF1* and truncating mutations affecting *RBM10* and *ARID1A*. Analysis of nucleotide context-specific mutation signatures grouped the sample set into distinct clusters that correlated with smoking history and alterations of reported lung adenocarcinoma genes. Whole genome sequence analysis revealed frequent structural re-arrangements, including in-frame exonic alterations within *EGFR* and *SIK2* kinases. The candidate genes identified in this study are attractive targets for biological characterization and therapeutic targeting of lung adenocarcinoma.

INTRODUCTION

Lung cancer is a leading cause of death worldwide, resulting in over 1.3 million deaths per year, of which over 40% are lung adenocarcinomas (World Health Organization, 2004; Travis, 2002). Most often tumors are discovered as locally advanced or metastatic disease, and despite improvements in molecular diagnosis and targeted therapies, the average five-year survival rate for lung adenocarcinoma is approximately 15% (Minna, 2008).

Molecular genotyping is now routinely used to guide clinical care of lung adenocarcinoma patients, largely due to clinical trials that demonstrated superior efficacy of targeted kinase inhibitors as compared to standard chemotherapy for patients with *EGFR* mutations or *ALK* fusions (Kwak et al., 2010; Pao and Chmielecki, 2010). In addition to *EGFR* and *ALK* alterations found in approximately 15% of U.S. cases, lung adenocarcinomas frequently harbor activating mutations in *KRAS, BRAF, ERBB2*, and *PIK3CA* or translocations in *RET* and *ROS1* (Pao and Hutchinson, 2012), all of which are being pursued as targets in ongoing clinical trials (http://clinicaltrials.gov/). Lung adenocarcinomas also often harbor loss-of-function mutations and deletions in tumor suppressor genes *TP53, STK11, RB1, NF1, CDKN2A, SMARCA4*, and *KEAP1* (Ding et al., 2008; Kan et al., 2010; Sanchez-Cespedes et al., 2002). Unfortunately, such alterations are difficult to exploit therapeutically. Therefore, knowledge of additional genes altered in lung adenocarcinoma is needed to further guide diagnosis and treatment.

Previous efforts in lung adenocarcinoma genome characterization include array-based profiling of copy number changes (Tanaka et al., 2007; Weir et al., 2007), targeted sequencing of candidate protein-coding genes (Ding et al., 2008; Kan et al., 2010), and whole genome sequencing of a single tumor/normal pair (Ju et al., 2012; Lee et al., 2010). These studies identified somatic focal amplifications of *NKX2-1*, substitutions and copy number alterations in known oncogenes and tumor suppressor genes, and recurrent in-frame fusions of *KIF5B* and *RET*. These studies have also nominated several putative cancer genes

with somatic mutations (*EPHA* family, *NTRK* family, *TLR4, LPHN3, GRM1, GLI*) but the functional consequence of many alterations is unknown. A recent study describing whole exome sequencing of 16 lung adenocarcinomas (Liu et al., 2012) enumerated several mutated genes but did not identify genes undergoing positive selection for mutation in the studied tumors.

In this study, we used next-generation sequencing to sequence the exomes and/or genomes of DNA from 183 lung adenocarcinomas and matched normal adjacent tissue pairs. In addition to verifying genes with frequent somatic alteration in previous studies of lung adenocarcinoma, we identified novel mutated genes with statistical evidence of selection and likely contributing to pathogenesis. Together, these data represent a significant advance towards a comprehensive annotation of somatic alterations in lung adenocarcinoma.

RESULTS

Patient cohort description

We sequenced DNA from 183 lung adenocarcinomas and matched normal tissues using paired-end massively parallel sequencing technology (Bentley et al., 2008). The cohort included 27 never smokers, 17 light smokers (defined by less than ten pack-years of tobacco use), 118 heavy smokers (greater than ten pack-years), and 21 patients of unknown smoking status (Table 1). The cohort included 90 stage I, 36 stage II, 22 stage III, and 10 stage IV lung adenocarcinoma cases as well as 25 patients with unknown stage. All tumors were chemotherapy-naïve, primary resection specimens except for one case with whole genome sequence data (LU-A08-43) that was a post-chemotherapy metastatic tumor from a never-smoker. Sample acquisition details are provided in Extended Experimental Procedures. Additional clinical descriptors of the cohort are provided in Table 1. Comprehensive clinical and histopathological annotations, sequence characteristics, and major variants for each patient in the study are provided in Table S1.

Mutation detection and validation

We examined 183 lung adenocarcinoma tumor/normal pairs with a combination of whole exome (WES) or whole genome sequencing (WGS): 159 WES, 23 WES and WGS, and 1 WGS only. Exomes were sequenced to a median fold coverage of 92 (range: 51-201) on 36.6 MB of target sequence (Fisher et al., 2011). Genomes were sequenced to a median coverage of 69 (range: 25-103) in the tumor and 36 (range: 28-55) in the normal, with the higher tumor coverage to adjust for stromal contamination. Complementary SNP array analysis of 183 pairs was used to detect genome-wide somatic copy number alterations. See Extended Experimental Procedures for more details.

We identified somatic substitutions and small insertions and deletions (indels) through statistical comparison of paired tumor/normal sequence data using algorithms calibrated for stromally-contaminated cancer tissues (Banerji et al., 2012; Stransky et al., 2011) (www.broadinstitute.org/cancer/cga, Extended Experimental Procedures). Exonic regions of the 183 cases contained 77,736 somatic variants corresponding to a median of 8.1 mutations/MB and a mean of 11.9 mutations/MB (range 0.04 to 117.4). These were comprised of 43,813 missense, 14,801 silent, 3,504 nonsense, 1,460 splice site, 2,310 deletions, 839 insertions, and 11,009 other mutations (predominantly 5’ and 3’ UTRs). Of the 3,149 indels, 182 were in-frame, 1,785 were predicted to cause a frame shift, 68 occurred at a splice site, and 1,114 were otherwise classified.

Mutation calls were validated by cross-comparison of coding mutations detected by WES and WGS from 24 cases with both data types. We validated 84% of 380 indel and 97% of 9,354 substitution variant calls identified by WGS at sufficiently powered sites in the

corresponding WES tumor sample. In the converse analysis, we validated 86% of 338 indel and 98% of 8,912 substitution WES variant calls at sufficiently powered sites in the corresponding WGS tumor sample (Figure S1, Table S2A, Extended Experimental Procedures). To validate mutations from cases with only WES data, we randomly selected 69 candidate mutations for ultra-deep (>1,000 fold) targeted resequencing. Somatic status was confirmed for 30 of 33 (91%) indel events and 33 of 36 (92%) substitution events (Table S2B). These validation rates generally meet or exceed those reported in similar sequencing studies (Banerji et al., 2012; Berger et al., 2012; Gerlinger et al., 2012; Nikolaev et al., 2011; Stransky et al., 2011; TCGA_Network, 2011; Totoki et al., 2011; Zang et al., 2012).

Somatic genetic signatures of mutagen exposure in lung adenocarcinoma

Consistent with previous studies (Ding et al., 2008; Kan et al., 2010; Zang et al., 2012), we observed significantly higher exonic mutation rates in tumors from smokers (median: 9.8/MB; mean: 12.9/MB; range: 0.04-117.4/MB) compared to never smokers (median: 1.7/MB; mean = 2.9/MB; range: 0.07-22.1/MB, $P = 3.0 \times 10^{-9}$, Wilcoxon rank sum). Lung adenocarcinoma mutation rates in our cohort exceeded those reported for other epithelial tumor types, except melanoma and squamous cell lung cancer (Hodis et al., 2012; Nikolaev et al., 2011; TCGA_Network, 2012; Wei et al., 2011).

To characterize the mutation spectrum of lung adenocarcinoma, we analyzed somatic substitutions and covered bases within their trinucleotide sequence context (Figure 1A). The most frequent mutation signatures were C→T transitions in the setting of CpG dinucleotides (CpG→T) and C→A transversions. The least frequent mutation type was A→C. Unbiased hierarchical clustering of context-specific mutation rates across 182 WES cases yielded 5 mutation spectrum clusters. These clusters represented grades of increasing mutational complexity: Cluster 1 was enriched for CpG→T mutations and marked by an overall low mutation rate. Cluster 2 was characterized by CpG→T transitions and CpG→A transversions. Cluster 3 showed additional C→A transversions outside of the CpG context. Cluster 4 showed additional C→T transitions outside of the CpG context and TpC transversions that mutated to either a T or a G. Cluster 5 comprised hypermutated tumors, containing a broad mutational spectrum that included rare mutation signatures, such as A→T transversions. Mutation spectrum clusters in tumors correlated with clinical features of patients. Cluster 1 was significantly enriched in never and light smokers ($P=1.4 \times 10^{-9}$, Fisher's exact test) while Cluster 4 was significantly enriched in patients with advanced (IIIB or IV) stage (P=0.0063, Fisher's Exact Test).

Differentiation of smokers and never smokers was evident from comparison of mutation counts from the most frequent mutational signatures, CpG→T and C→A (Figure 1B). These results were consistent with previous reports of signatures of DNA damage by tobacco (Hainaut and Pfeifer, 2001). Applying thresholds to a log-adjusted ratio of CpG→T and C→A mutations (see Experimental Methods), we imputed smoking status for 21 patients who lacked reported smoking history and accurately recapitulated reported smoking status for over 75% of the remaining cases (Figure 1B). Exonic and intronic mutation rates, context-specific mutation counts, imputed smoking status, and mutation spectrum cluster assignments for each patient are provided in Table S1.

Calibration of a statistical approach to the analysis of high mutation rate tumors

The high mutation rates in lung adenocarcinoma and other tumors (Hodis et al., 2012; TCGA_Network, 2012) present a challenge for unbiased discovery of mutated genes undergoing positive somatic selection. Over 13,000 of 18,616 genes with adequate sequence coverage had non-synonymous somatic mutations in at least one tumor, and over 3,000 were

mutated in at least 5 patients. These genes included those with very large genomic footprints (e.g. *TTN*), genes with low basal expression in lung adenocarcinomas (e.g. *CSMD3*), and genes accumulating high numbers of silent substitutions (e.g. *LRP1B*).

Application of a standard binomial background mutation model assuming a constant mutation rate in each patient and nucleotide context stratum (Berger et al., 2011) yielded profound test statistic inflation (Figure S2A) and identified over 1,300 significantly mutated genes. Genes with significant P values in this analysis had low basal expression in lung adenocarcinoma cell lines (Barretina et al., 2012) (Figure S2B), harbored high fractions of synonymous mutations, and were enriched in gene classes previously unassociated with cancer (e.g. olfactory receptors, solute transporters). Recalibration of this model by limiting to genes with evidence of expression improved but did not completely correct this statistical inflation (Figure S2C). These results suggested a high degree of variation in neutral somatic mutation rates among genes, including expression-dependent variation. This observation is consistent with reports of regional mutation rates correlated with density of H3K9 chromatin marks across cancers (Schuster-Bockler and Lehner, 2012), and with gene expression in multiple myeloma (Chapman et al., 2011).

To more adequately model variation of neutral somatic mutation rates among genes, we applied the InVEx algorithm (Hodis et al., 2012) to exploit the abundant non-coding mutations detected by both WES and WGS. InVEx permutes coding, untranslated, and intronic mutations within covered territories of each gene, patient, and nucleotide context to generate within-gene null distributions of “functional impact” across a sample set (see Experimental Procedures).

Our primary InVEx analysis employed a PolyPhen-2 (PPH2) based metric (Adzhubei et al., 2010) to assess the functional impact of observed and permuted mutations. Applying this analysis to 12,907 mutated genes with at least one PPH2-scored event yielded a well-distributed test statistic with minimal inflation (Figure S2D) and without gene expression bias in lung adenocarcinoma cell lines (Figure S2B). To increase specificity and power, we restricted our analysis to 7,260 genes demonstrating expression (median Robust Multi-Array Average (RMA) value ≥ 5) in a panel of 40 lung adenocarcinoma cell lines (Barretina et al., 2012), which resulted in a similarly well-calibrated test statistic (Figure S2E).

Next, we tested for enrichment of loss-of-function (LOF) mutations by considering only truncating mutations as functional and all remaining mutation types as neutral. We applied this method to 2,266 genes with evidence of expression in lung adenocarcinoma cell lines and at least one truncating mutational event. Finally, we applied both PPH2 and LOF InVEx analyses to a focused set of Cancer Gene Census (CGC) genes expressed in lung adenocarcinoma and mutated or amplified in one or more tumor types.

Statistical driver analysis yields previously reported and novel lung adenocarcinoma genes

The primary PPH2 InVEx analysis yielded 13 genes with statistical evidence of positive selection ($q<0.25$) (Table S3A). These included lung adenocarcinoma genes with non-synonymous mutation frequencies consistent with previous reports: *TP53* (50%), *KRAS* (27%), *EGFR* (17%), *STK11* (15%), *KEAP1* (12%), *NF1* (11%), *BRAF* (8%), and *SMAD4* (3%). This analysis also uncovered 5 novel candidates, including *CHEK2*, a gene driven by an apparent recurrent mapping artifact in three tumors and removed from all subsequent analyses (see Extended Experimental Procedures). The remaining candidates are mutated at frequencies lower than most previously reported genes, demonstrating the increased power of our large sample set. The LOF InVEx yielded 6 significantly mutated genes ($q<0.25$), including *BRD3*, an additional gene not contained in the PPH2 analysis (Table S3B). The

CGC-only PPH2 and LOF analyses yielded 15 and 10 genes respectively, including CTNNB1, FGFR3, ATM, CBL, PIK3CA, PTEN, FBXW7, ARID1A and *SETD2* (Table S3C-D). In total, the union of these four analyses nominated 25 genes as significantly mutated in our cohort (Figure 2A). Somatic coding mutations in significantly mutated genes and known lung adenocarcinoma genes are provided in Table S1. The entire list of somatic coding mutations for all covered genes is provided in Table S4.

To compare our results with previous reports, we reviewed the CGC and lung adenocarcinoma literature to identify genes with previous evidence for functional somatic mutation in lung adenenocarcinoma (see Extended Experimental Procedures for criteria and references). Of the 19 genes with reported functional mutations, 13 were significantly mutated genes nominated by our analysis (KRAS, TP53, EGFR, STK11, SMARCA4, NF1, RB1, BRAF, KEAP1, SMAD4, CTNNB1, PIK3CA and *ATM*). The alterations driving the statistical enrichment of these genes included previously reported and novel mutations (Figure S3A-C). The remaining 6 reported lung adenocarcinoma genes (CDKN2A, *ERBB2, AKT1, NRAS, HRAS, APC*) were not significant in our mutation analysis (Table S3E), although we did identify canonical driver mutations in these genes (e.g. *AKT1* p.E17K, *NRAS* p.Q61L) and although *CDKN2A* is significantly deleted (see Figure 2) and re-arranged (see below). This may reflect a power limitation of our cohort or analytic methods we applied, particularly when identifying infrequently mutated genes such as *AKT1, NRAS*, and *HRAS*. Also missing among our significantly mutated genes were 22 genes nominated by two previous large-scale targeted lung adenocarcinoma sequencing studies of similar or smaller size (Ding et al., 2008; Kan et al., 2010) (see Extended Experimental Procedures for the complete list). Most of these genes (20 of 22) did not pass our gene expression filter and thus were not included in our global analysis. Targeted analysis of these genes identified four with nominal evidence for positive selection via PPH2 InVEx (*EPHA3, LPHN3, GRM1, TLR4*), the most significant of these being *EPHA3* (P = 0.0027, PPH2 InVEx).

Correlations among alterations in significantly mutated genes and clinicopathologic and genomic features

We correlated mutation status of the 25 significantly mutated genes with clinical features (smoking, age, stage), genomic variables (mutation rate, mutation spectrum cluster, imputed smoking status), and presence of driver alterations in 25 genes frequently or functionally altered in lung adenocarcinoma. These alterations included genes with reported high frequency of somatic mutation (e.g. *KRAS*) or focal amplification (e.g. *NKX2-1*) or deletion (e.g. *TP53*). High frequency somatic copy number alterations used for this analysis were curated from published surveys of lung adenocarcinoma (Tanaka et al., 2007; Weir et al., 2007). See “Experimental Methods: Hallmarks analysis” for the strict definition of driver alterations. In our cohort, we observed gains of *TERT* (42% of cases, 15% focal), *MYC* (31%), *EGFR* (22%), and *NKX2-1* (18%, 10% focal). Frequent losses were seen in *TP53* (18%) and *CDKN2A* (24%, 10% homozygous) as well as in other significantly mutated genes including *SMAD4, KEAP1*, and *SMARCA4*.

EGFR mutation was significantly anti-correlated with *KRAS* mutation ($P = 3.3 \times 10^{-4}$) and somatic mutation rate ($P = 5.9 \times 10^{-4}$) *EGFR* mutations significantly correlated with never/light smoker status ($P = 2.0 \times 10^{-6}$), imputed never/light smoker status (1.5×10^{-4}), and membership in spectrum cluster 1 (P = 0.0015). *KRAS, STK11, SMARCA4*, and *KEAP1* mutations were significantly anti-correlated with both spectrum cluster 1 and imputed never/light smoking status (P<0.005). These findings are consistent with reported associations (Koivunen et al., 2008; Pao et al., 2004; Pao et al., 2005; Slebos et al., 1991). In addition, *NF1* mutations were significantly depleted in spectrum cluster 1 ($P = 4 \times 10^{-3}$) and co-occurred with *U2AF1* mutations (P = 0.0011). *KRAS* driver alterations (including both mutations and copy number alterations) significantly associated with spectrum cluster 3 (P =

0.00071). *STK11* driver alterations were significantly enriched in spectrum cluster 2 (P=0.0026). Correlation results are graphically summarized in Figure S3D.

Finally, we screened the 25 significantly mutated genes for association with progression-free survival (PFS) across 135 patients with PFS data. *U2AF1* and (P = 0.00011, log-rank test) and *TP53* mutations (P=0.0014 , log-rank test) were associated with significantly reduced survival (Figure S3E), The latter finding was consistent with previous reports (Kosaka et al., 2009; Mitsudomi et al., 1993). No other significant associations with PFS were seen.

Nomination of candidate lung adenocarcinoma genes

One of the most significantly mutated genes in this lung adenocarcinoma cohort was *U2AF1* ($P = 2.0 \times 10^{-6}$, PPH2 InVEx), which had non-synonymous mutations in 3% of cases (Figure 3A). Identical c.101C>T, p.S34F mutations were seen in 4 of 5 *U2AF1* mutant cases (Figure 3A), the exact mutation reported in myelodysplastic syndrome (MDS) (Graubert et al., 2012; Yoshida et al., 2011). To our knowledge, this study is the first report of *U2AF1* mutations in an epithelial tumor. One of four p.S34F mutations occurred with an activating event in *KRAS* (p.Q61H), suggesting that *U2AF1* mutations may confer tumorigenic capability independent of known proliferation-sustaining driver genes. As mentioned above, 4 patients with *U2AF1* mutations and survival data had significantly reduced PFS (Figure S3E). Non-synonymous mutations in genes encoding other members of the spliceosome complex (including *SF3B1, U2AF2*, and *PRPF40B*) were found in 14 additional cases (Yoshida et al., 2011).

RBM10 was frequently mutated (12/183 cases; 7%) and subject to recurrent nonsense, frameshift, or splice site mutations, present in 7 of 12 mutated cases (4% of overall cohort) (Figure 3B). This resulted in significant enrichment in the global PPH2 InVEx analysis (P = 0.00042) (Table S3). Like *U2AF1, RBM10* is an RNA-binding protein highly expressed in lung adenocarcinoma cell lines (data not shown) and its mutations co-occurred with those in known lung adenocarcinoma oncogenes (*KRAS, EGFR, PIK3CA*). *ARID1A*, encoding a key protein in the SWI/SNF chromatin-remodeling complex, was mutated in 8% of cases (Figure 3C) and showed significant accumulation of nonsense substitutions and frameshift indels (P=0.027, CGC LOF InVEx).

Whole genome rearrangement analysis reveals novel and recurrent structural variants

We used paired-end and split-read mapping of whole genome data (Banerji et al., 2012; Bass et al., 2011; Medvedev et al., 2009) to detect and map the breakpoints of 2,349 somatic rearrangements across 24 WGS cases. The majority of these were intra-chromosomal rearrangements (1,818 events), but included 531 inter-chromosomal events. Among these were 1,443 (61.4%) genic rearrangements (i.e. in which one breakpoint was contained within the promoter, untranslated region, intron, or exon of a gene) and 906 (38.6%) purely intergenic events. Lung adenocarcinomas harbored a wide range of total rearrangements (median: 98, range: 18-246), genic rearrangements (median: 50, range: 12-173) (Figure 4A), and overall genome complexity (Figure S4). The variability of rearrangement counts between cases did not correlate with clinical variables (Figure S4, Table S1) or mutation spectrum. Rearrangement coordinates and interpretations are provided as Table S5.

The reading-frame of affected genes was preserved by 3% of detected rearrangements (71 of 2,349). These included 34 protein fusions, 13 duplications, and 24 deletions. We found 44 rearrangements that fused untranslated regions (UTR) of two genes without affecting the protein-coding sequence of either gene. All 25 genic fusions we tested were confirmed by PCR and Illumina sequencing (see Extended Experimental Procedures) (Table S5).

The gene with the highest rate of rearrangements for its size was *CDKN2A* (4.3 rearrangements/sequenced MB). Two cases had out-of-frame, antisense fusions (with *MTAP* and *C9orf53*) and a third harbored an in-frame deletion (Figure 4B). As shown in lung squamous cell carcinomas, rearrangements represent an additional mechanism of *CDKN2A* inactivation, in addition to reported mutation, homozygous deletion, and methylation (TCGA_Network, 2012). Additional lung adenocarcinoma tumor suppressors affected by predicted null or truncating rearrangements included *STK11* (2.5 kb deletion removing the translational start site) and *APC* (mid-exon rearrangement) (Figure 4B).

We next focused on potentially activating in-frame re-arrangements of kinase genes. This analysis uncovered a two-exon deletion in *EGFR*, previously identified in glioblastoma multiforme but novel in lung adenocarcinoma, ablating a portion of the C-terminus of EGFR encoded by exons 25 and 26 (Figure 4B, Figure 5A, and Figure S5), including residues associated with interaction with PIK3C2B (Wheeler and Domin, 2001) and CBL (Grovdal et al., 2004). Similar C-terminal deletion variants (EGFR vIVb) have been previously identified in glioblastoma (Ekstrand et al., 1992), and shown to be oncogenic in cellular and animal models (Cho et al., 2011; Pines et al., 2010). This tumor contained a second somatic alteration in *EGFR*, a p.G719S mutation, suggesting possible synergy of activating *EGFR* mutations or presence of independent, subclonal activating mutations.

To assess oncogenicity of this novel EGFR variant, we ectopically expressed an *EGFR* transgene lacking exons 25 and 26 in NIH-3T3 cells. As has been previously observed for oncogenic *EGFR* mutations, cells stably expressing this transgene demonstrated colony formation in soft agar (Figure 5B) and increased EGFR and AKT phosphorylation in the absence of EGF (Figure 5C). In contrast, cells expressing wild-type EGFR formed colonies only in the presence of EGF (Figure 5B). Over-expression of the *EGFR* transgene in Ba/F3 cells led to interleukin-3 independent proliferation that was blocked by treatment with an EGFR tyrosine kinase inhibitor, erlotinib (Figure 5D) at concentrations previously shown to be sufficient for inhibition of activated variants of EGFR (Yuza et al., 2007).

Kinases with in-frame rearrangements in tumors without mutations in lung adenocarcinoma oncogenes included *SIK2* and *ROCK1* (Figure 4B). An in-frame kinase domain duplication in *SIK2* (salt inducible kinase 2) was identified and validated by qPCR. The duplication occurred 15 amino acids upstream of Thr-175, where a related kinase, SIK1, is activated by STK11 (Hashimoto et al., 2008). A 19-exon duplication was uncovered in *ROCK1*, a serine/threonine kinase that acts as an effector of Rho signaling (Pearce et al., 2010).

Notably, we did not identify any in-frame rearrangements involving kinase fusion targets in lung adenocarcinoma *ALK, RET1*, and *ROS1*. Given their reported 2-7% frequency in lung adenocarcinoma (Bergethon et al., 2012; Takeuchi et al., 2012), our study of 24 tumor/normal pairs may not be large enough to detect these rearrangements. Interestingly, an out-of-frame *ROS1-CD74* translocation was identified in a single patient, without evidence for the previously characterized reciprocal activating event. In-frame fusions and indels are annotated for each WGS case in Table S1.

DISCUSSION

Charting the next-generation hallmarks of lung adenocarcinoma

The “hallmarks of cancer”, as defined by (Hanahan and Weinberg, 2000, 2011), comprise a set of cellular traits thought to be necessary for tumorigenesis. They also represent a powerful framework to evaluate our understanding of genetic alterations driving lung adenocarcinoma. With this aim, we mapped each of 25 experimentally validated lung adenocarcinoma genes to one or more cancer hallmarks from (Hanahan and Weinberg,

2000, 2011) (Table S6, and Experimental Procedures). These 25 genes include the 19 previously reported genes discussed above in addition to 6 genes subject to frequent copy number alteration in lung adenocarcinoma (*NKX2-1, TERT, PTEN, MDM2, CCND1*, and *MYC)*. Next, we integrated this gene-hallmark mapping with our somatic mutation and copy number data to estimate the prevalence of cancer hallmark alterations in lung adenocarcinoma (Figure 6, Table S1).

For many cases in our cohort, we could attribute only a minority of the ten cancer hallmarks to a distinct genetic lesion (Figure S6). Only 6% of tumors had alterations assigned to all six classic hallmarks and none had alterations impacting all ten emerging and classic hallmarks. In contrast, 15% of our cohort did not have a single hallmark alteration and 38% had three or fewer. This finding is likely explained in part by alteration of cancer genes by mechanisms not assayed in our study and also suggests that many lung adenocarcinoma genes have not been identified. This may be especially relevant for the hallmarks of "Avoiding Immune Destruction" and "Tumor Promoting Inflammation", to which none of the recurrently mutated genes identified in our study or previous studies could be linked. One of the most important and therapeutically targetable cancer hallmarks is "Sustaining Proliferative Signaling" (Figure 6, Figures S6). Less than half (47%) of our cohort harbored a mutation in a known driver gene for this hallmark, and only slightly more (55%) when including high-level amplification in one or more proliferative signaling genes (e.g. *EGFR, ERBB2, MYC*).

Our mapping of somatic alterations to cancer hallmarks illuminates specific gaps in the understanding of the somatic genetic underpinnings of lung adenocarcinoma. Around half of the sequenced cohort lacked a mutation supporting sustained proliferative signaling and a majority lacked a genetic alteration explaining the phenotypes of invasion and metastasis or angiogenesis. This phenotypic gap may be explained by novel capabilities not yet attributed to alterations in known lung adenocarcinoma genes, or through novel alterations in genes previously unassociated with this disease that will emerge through additional unbiased analyses.

While annotating the 25 known lung adenocarcinoma genes, we noted that *SMARCA4*, an epigenetic regulator and tumor suppressor, could not be clearly mapped to any cancer hallmark. Given the frequent somatic mutations in epigenetic and splicing regulators found by recent cancer genome scans (Elsasser et al., 2011) and our study (*U2AF1, ARID1A, RBM10, SETD2*, and *BRD3*), we speculated that these alterations may represent a novel hallmark of "Epigenetic and RNA deregulation". Together, these genes implicate the proposed "11th hallmark" in a considerable proportion of cases (10% including only SMARCA4, 22% including nominated genes).

Efficiency and power in somatic genetic studies of lung adenocarcinoma

This study represents the largest sequencing analysis of lung adenocarcinoma to date. Our analysis reveals the genomic complexity of lung adenocarcinoma at the base-pair and structural levels, exceeding that observed in genome characterization studies of most other tumor types. We have applied a recently published statistical method (Hodis et al., 2012) for identifying somatically mutated genes displaying evidence of positive selection in cancer. This permutation approach exploits the abundant supply of intronic and flanking mutation events detected in both WES and WGS to adequately model the gene-specific variation in neutral mutation rates (Hodis et al., 2012). We believe that such a calibrated approach is required to identify signals of positive somatic selection in large unbiased cancer genome scans. This concern is particularly relevant to tumor types harboring high rates of somatic mutation, such as lung adenocarcinoma or melanoma.

This study has led to discovery of significant mutation of 25 genes in lung adenocarcinoma. Notably, our study did not identify a mutated oncogene in every tumor sample. Furthermore, we were unable to statistically nominate several important, but rarely mutated, lung adenocarcinoma genes (*AKT1, ERBB2, NRAS, HRAS*, each with ≤3 events in our cohort). Therefore, future studies of larger cohorts by The Cancer Genome Atlas and other consortia that combine analysis of data from RNA-seq, methylation profiling, and other „omic platforms, will likely yield an even more complete annotation of genes significant to lung adenocarcinoma.

Conclusion

This study represents a significant advance towards complete characterization of the genomic alterations of lung adenocarcinoma. These results are a testament to the power of unbiased, large-scale next generation sequencing technology to expand our understanding of tumor biology. The novel mutated genes identified in this study warrant further investigation to determine their biologic, prognostic and/or therapeutic significance in lung adenocarcinoma, potentially leading to clinical translation and improved outcomes for patients with this deadly disease.

EXPERIMENTAL PROCEDURES

Details of sample preparation and analysis are described in the Extended Experimental Procedures.

Patient and sample characteristics

We obtained DNA from tumor and matched normal-adjacent tissue from 6 source sites. DNA was obtained from frozen tissue primary lung cancer resection specimens for all samples, with the exception of one patient (LU-A08-14), for whom a liver metastasis was obtained at autopsy. The 183 lung adenocarcinoma diagnoses were either certified by a clinical surgical pathology report provided by the external tissue bank, collaborator, or verified through in-house review by an anatomical pathologist at the Broad Institute of MIT and Harvard. A second round of pathology review was conducted by an expert committee led by W. Travis. Informed consent (Institutional Review Board) was obtained for each sample using protocols approved by the Broad Institute of Harvard and MIT and each originating tissue source site.

Massively parallel sequencing

Exome capture was performed using Agilent SureSelect Human All Exon 50 Mb according the manufacturers instructions. All whole exome (WES) and whole genome (WGS) sequencing was performed on the Illumina HiSeq platform. Basic alignment and sequence QC was done on the Picard and Firehose pipelines at the Broad Institute. Mapped genomes were processed by the Broad Firehose pipeline to perform additional QC, variant calling, and mutational significance analysis.

External data

Gene expression data for 40 lung adenocarcinoma cell lines was obtained from the Cancer Cell Line Encyclopedia (CCLE) (http://www.broadinstitute.org/ccle/home) as robust microarray average (RMA) normalized tab-delimited text data (Barretina et al., 2012).

Statistical Analyses

We evaluated statistical evidence for somatic selection within the longest transcript of each gene using InVEx (Hodis et al., 2012) with PolyPhen-2 based (Adzhubei et al., 2010) and

loss-of-function (LOF) based scoring schemes. The method was implemented in Python (http://www.python.org) and is available for download (http://www.broadinstitute.org/software/invex/). Gene ranking according to a stratified binomial model was performed using the MutSig method from (Berger et al., 2011), implemented in MATLAB. Correlations between genotype status, mutation/rearrangement spectrum data, and clinical variables were performed by Fisher's exact test for dichotomous variables and Wilcoxon rank sum test for dichotomous variables versus numeric data (e.g. mutation status vs total mutation rate). All remaining statistical computing, including cluster analysis and visualization, was performed using standard packages in R (http://www.r-project.org).

Hallmarks analysis

We manually assigned 25 genes implicated by previous studies to be frequently or functionally altered in lung adenocarcinoma to one or more cancer “hallmarks” as defined by (Hanahan and Weinberg, 2000, 2011) (See Extended Experimental Procedures). We determined whether alterations in gene *i* could be implicated as a “driver” of one or more cancer hallmarks in case *j* by applying the following criteria: We inferred the activation status of genes annotated by the Sanger Gene Census as “dominant” cancer genes (e.g. *KRAS*) in each patient by evaluating every nonsynonymous variant in the gene for its presence within a COSMIC hotspot {Forbes, 2011 #245}. Mutations that were present in the COSMIC database (http://www.sanger.ac.uk/genetics/CGP/cosmic/) at least 10 times were considered oncogenic mutations. We considered a dominant gene activated if it harbored such a variant or a high-level, focal amplification. We considered recessive cancer genes (e.g. *TP53*) to be inactivated if the gene had (1) a truncating mutation, (2) compound missense mutations, (3) a hemizygous missense mutation, or (4) homozygous copy number loss. We mapped each patient *j* to hallmark *k* if the sample contained at least one activating or inactivating event in a dominant or recessive cancer gene, respectively, that mapped to hallmark *k*.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank all members of the Biological Samples Platform, DNA Sequencing Platform, and Genetic Analysis Platforms of the Broad Institute, without whose work this sequencing project could not have occurred. M.I. is supported by NCI training grant T32 CA9216-31. T.J.P. is supported by a Canadian Institutes of Health Research Fellowship. A.H.B. is supported by a postdoctoral fellowship from the American Cancer Society. P.S.H. is supported by a Young Investigator Award from the National Lung Cancer Partnership and a Career Development Award from the Dana-Farber/Harvard Cancer Center Lung Cancer SPORE P50 CA090578. E.H. is supported by NIGMS training grant T32GM07753. R.K.T is supported by the German Ministry of Science and Education (BMBF) as a member of the NGFNplus program (grant 01GS08100), by the Max Planck Society (M.I.F.A.NEUR8061), by the Deutsche Forschungsgemeinschaft (DFG) through SFB832 (TP6) and grant TH1386/3-1, by the EU-Framework Programme CURELUNG (HEALTH-F2-2010-258677), Stand Up To Cancer-American Association for Cancer Research Innovative Research Grant (SU2C-AACR-IR60109), by the Behrens-Weise Foundation and by an anonymous foundation. This work was supported by the National Human Genome Research Institute (E.S.L.) and by Uniting Against Lung Cancer, the Lung Cancer Research Foundation, and the American Lung Association (M.M.).

REFERENCES

Cancer Statistics. W.H. Organization; 2004.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods. 2010; 7:248–249. [PubMed: 20354512]

Banerji S, Cibulskis K, Rangel-Escareno C, Brown KK, Carter SL, Frederick AM, Lawrence MS, Sivachenko AY, Sougnez C, Zou L, et al. Sequence analysis of mutations and translocations across breast cancer subtypes. Nature. 2012; 486:405–409. [PubMed: 22722202]

Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehar J, Kryukov GV, Sonkin D, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012; 483:603–607. [PubMed: 22460905]

Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, Sougnez C, Voet D, Saksena G, Sivachenko A, et al. Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet. 2011; 43:964–968. [PubMed: 21892161]

Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–59. [PubMed: 18987734]

Berger MF, Hodis E, Heffernan TP, Deribe YL, Lawrence MS, Protopopov A, Ivanova E, Watson IR, Nickerson E, Ghosh P, et al. Melanoma genome sequencing reveals frequent PREX2 mutations. Nature. 2012; 485:502–506. [PubMed: 22622578]

Berger MF, Lawrence MS, Demichelis F, Drier Y, Cibulskis K, Sivachenko AY, Sboner A, Esgueva R, Pflueger D, Sougnez C, et al. The genomic complexity of primary human prostate cancer. Nature. 2011; 470:214–220. [PubMed: 21307934]

Bergethon K, Shaw AT, Ignatius Ou SH, Katayama R, Lovly CM, McDonald NT, Massion PP, Siwak-Tapp C, Gonzalez A, Fang R, et al. ROS1 Rearrangements Define a Unique Molecular Class of Lung Cancers. J Clin Oncol. 2012; 30:863–870. [PubMed: 22215748]

Chapman MA, Lawrence MS, Keats JJ, Cibulskis K, Sougnez C, Schinzel AC, Harview CL, Brunet JP, Ahmann GJ, Adli M, et al. Initial genome sequencing and analysis of multiple myeloma. Nature. 2011; 471:467–472. [PubMed: 21430775]

Cho J, Pastorino S, Zeng Q, Xu X, Johnson W, Vandenberg S, Verhaak R, Cherniack AD, Watanabe H, Dutt A, et al. Glioblastoma-derived epidermal growth factor receptor carboxyl-terminal deletion mutants are transforming and are sensitive to EGFR-directed therapies. Cancer Res. 2011; 71:7587–7596. [PubMed: 22001862]

Ding L, Getz G, Wheeler DA, Mardis ER, McLellan MD, Cibulskis K, Sougnez C, Greulich H, Muzny DM, Morgan MB, et al. Somatic mutations affect key pathways in lung adenocarcinoma. Nature. 2008; 455:1069–1075. [PubMed: 18948947]

Ekstrand AJ, Sugawa N, James CD, Collins VP. Amplified and rearranged epidermal growth factor receptor genes in human glioblastomas reveal deletions of sequences encoding portions of the N- and/or C-terminal tails. Proc Natl Acad Sci U S A. 1992; 89:4309–4313. [PubMed: 1584765]

Elsasser SJ, Allis CD, Lewis PW. Cancer. New epigenetic drivers of cancers. Science. 2011; 331:1145–1146. [PubMed: 21385704]

Fisher S, Barry A, Abreu J, Minie B, Nolan J, Delorey TM, Young G, Fennell TJ, Allen A, Ambrogio L, et al. A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries. Genome Biol. 2011; 12:R1. [PubMed: 21205303]

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. A census of human cancer genes. Nat Rev Cancer. 2004; 4:177–183. [PubMed: 14993899]

Gerlinger M, Rowan AJ, Horswell S, Larkin J, Endesfelder D, Gronroos E, Martinez P, Matthews N, Stewart A, Tarpey P, et al. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N Engl J Med. 2012; 366:883–892. [PubMed: 22397650]

Graubert TA, Shen D, Ding L, Okeyo-Owuor T, Lunn CL, Shao J, Krysiak K, Harris CC, Koboldt DC, Larson DE, et al. Recurrent mutations in the U2AF1 splicing factor in myelodysplastic syndromes. Nat Genet. 2012; 44:53–57. [PubMed: 22158538]

Grovdal LM, Stang E, Sorkin A, Madshus IH. Direct interaction of Cbl with pTyr 1045 of the EGF receptor (EGFR) is required to sort the EGFR to lysosomes for degradation. Exp Cell Res. 2004; 300:388–395. [PubMed: 15475003]

Hainaut P, Pfeifer GP. Patterns of p53 G-->T transversions in lung cancers reflect the primary mutagenic signature of DNA-damage by tobacco smoke. Carcinogenesis. 2001; 22:367–374. [PubMed: 11238174]

Hanahan D, Weinberg RA. The hallmarks of cancer. Cell. 2000; 100:57–70. [PubMed: 10647931]

Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011; 144:646–674. [PubMed: 21376230]

Hashimoto YK, Satoh T, Okamoto M, Takemori H. Importance of autophosphorylation at Ser186 in the A-loop of salt inducible kinase 1 for its sustained kinase activity. J Cell Biochem. 2008; 104:1724–1739. [PubMed: 18348280]

Hodis E, Watson IR, Kryukov GV, Arold ST, Imielinski M, Theurillat JP, Nickerson E, Auclair D, Li L, Place C, et al. A landscape of driver mutations in melanoma. Cell. 2012; 150:251–263. [PubMed: 22817889]

Ju YS, Lee WC, Shin JY, Lee S, Bleazard T, Won JK, Kim YT, Kim JI, Kang JH, Seo JS. A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. Genome Res. 2012

Kan Z, Jaiswal BS, Stinson J, Janakiraman V, Bhatt D, Stern HM, Yue P, Haverty PM, Bourgon R, Zheng J, et al. Diverse somatic mutation patterns and pathway alterations in human cancers. Nature. 2010; 466:869–873. [PubMed: 20668451]

Koivunen JP, Kim J, Lee J, Rogers AM, Park JO, Zhao X, Naoki K, Okamoto I, Nakagawa K, Yeap BY, et al. Mutations in the LKB1 tumour suppressor are frequently detected in tumours from Caucasian but not Asian lung cancer patients. British journal of cancer. 2008; 99:245–252. [PubMed: 18594528]

Kosaka T, Yatabe Y, Onozato R, Kuwano H, Mitsudomi T. Prognostic implication of EGFR, KRAS, and TP53 gene mutations in a large cohort of Japanese patients with surgically treated lung adenocarcinoma. J Thorac Oncol. 2009; 4:22–29. [PubMed: 19096302]

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. Circos: an information aesthetic for comparative genomics. Genome Res. 2009; 19:1639–1645. [PubMed: 19541911]

Kwak EL, Bang YJ, Camidge DR, Shaw AT, Solomon B, Maki RG, Ou SH, Dezube BJ, Janne PA, Costa DB, et al. Anaplastic lymphoma kinase inhibition in non-small- cell lung cancer. N Engl J Med. 2010; 363:1693–1703. [PubMed: 20979469]

Lee W, Jiang Z, Liu J, Haverty PM, Guan Y, Stinson J, Yue P, Zhang Y, Pant KP, Bhatt D, et al. The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature. 2010; 465:473–477. [PubMed: 20505728]

Liu P, Morrison C, Wang L, Xiong D, Vedell P, Cui P, Hua X, Ding F, Lu Y, James M, et al. Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. Carcinogenesis. 2012

Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009; 6:S13–20. [PubMed: 19844226]

Minna, JD.; Schiller, JH. Lung Cancer.; Harrison's Principles of Internal Medicine. 17th ed2008. p. 551-562.

Mitsudomi T, Oyama T, Kusano T, Osaki T, Nakanishi R, Shirakusa T. Mutations of the p53 gene as a predictor of poor prognosis in patients with non-small-cell lung cancer. J Natl Cancer Inst. 1993; 85:2018–2023. [PubMed: 8246288]

Nikolaev SI, Rimoldi D, Iseli C, Valsesia A, Robyr D, Gehrig C, Harshman K, Guipponi M, Bukach O, Zoete V, et al. Exome sequencing identifies recurrent somatic MAP2K1 and MAP2K2 mutations in melanoma. Nat Genet. 2011; 44:133–139. [PubMed: 22197931]

Pao W, Chmielecki J. Rational, biologically based treatment of EGFR-mutant non-small-cell lung cancer. Nat Rev Cancer. 2010; 10:760–774. [PubMed: 20966921]

Pao W, Hutchinson KE. Chipping away at the lung cancer genome. Nat Med. 2012; 18:349–351. [PubMed: 22395697]

Pao W, Miller V, Zakowski M, Doherty J, Politi K, Sarkaria I, Singh B, Heelan R, Rusch V, Fulton L, et al. EGF receptor gene mutations are common in lung cancers from “never smokers” and are associated with sensitivity of tumors to gefitinib and erlotinib. Proc Natl Acad Sci U S A. 2004; 101:13306–13311. [PubMed: 15329413]

Pao W, Wang TY, Riely GJ, Miller VA, Pan Q, Ladanyi M, Zakowski MF, Heelan RT, Kris MG, Varmus HE. KRAS mutations and primary resistance of lung adenocarcinomas to gefitinib or erlotinib. PLoS Med. 2005; 2:e17. [PubMed: 15696205]

Pearce LR, Komander D, Alessi DR. The nuts and bolts of AGC protein kinases. Nat Rev Mol Cell Biol. 2010; 11:9–22. [PubMed: 20027184]

Pines G, Huang PH, Zwang Y, White FM, Yarden Y. EGFRvIV: a previously uncharacterized oncogenic mutant reveals a kinase autoinhibitory mechanism. Oncogene. 2010; 29:5850–5860. [PubMed: 20676128]

Sanchez-Cespedes M, Parrella P, Esteller M, Nomoto S, Trink B, Engles JM, Westra WH, Herman JG, Sidransky D. Inactivation of LKB1/STK11 is a common event in adenocarcinomas of the lung. Cancer Res. 2002; 62:3659–3662. [PubMed: 12097271]

Schuster-Bockler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. Nature. 2012

Slebos RJ, Hruban RH, Dalesio O, Mooi WJ, Offerhaus GJ, Rodenhuis S. Relationship between K-ras oncogene activation and smoking in adenocarcinoma of the human lung. J Natl Cancer Inst. 1991; 83:1024–1027. [PubMed: 2072410]

Stransky N, Egloff AM, Tward AD, Kostic AD, Cibulskis K, Sivachenko A, Kryukov GV, Lawrence MS, Sougnez C, McKenna A, et al. The mutational landscape of head and neck squamous cell carcinoma. Science. 2011; 333:1157–1160. [PubMed: 21798893]

Takeuchi K, Soda M, Togashi Y, Suzuki R, Sakata S, Hatano S, Asaka R, Hamanaka W, Ninomiya H, Uehara H, et al. RET, ROS1 and ALK fusions in lung cancer. Nat Med. 2012; 18:378–381. [PubMed: 22327623]

Tanaka H, Yanagisawa K, Shinjo K, Taguchi A, Maeno K, Tomida S, Shimada Y, Osada H, Kosaka T, Matsubara H, et al. Lineage-specific dependency of lung adenocarcinomas on the lung development regulator TTF-1. Cancer Res. 2007; 67:6007–6011. [PubMed: 17616654]

TCGA_Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011; 474:609–615. [PubMed: 21720365]

TCGA_Network. Comprehensive Genomic Characterization of Squamous Cell Lung Cancers. Nature. 2012 Submitted.

Totoki Y, Tatsuno K, Yamamoto S, Arai Y, Hosoda F, Ishikawa S, Tsutsumi S, Sonoda K, Totsuka H, Shirakihara T, et al. High-resolution characterization of a hepatocellular carcinoma genome. Nat Genet. 2011; 43:464–469. [PubMed: 21499249]

Travis WD. Pathology of lung cancer. Clin Chest Med. 2002; 23:65–81. viii. [PubMed: 11901921]

Wei X, Walia V, Lin JC, Teer JK, Prickett TD, Gartner J, Davis S, Stemke-Hale K, Davies MA, Gershenwald JE, et al. Exome sequencing identifies GRIN2A as frequently mutated in melanoma. Nat Genet. 2011; 43:442–446. [PubMed: 21499247]

Weir BA, Woo MS, Getz G, Perner S, Ding L, Beroukhim R, Lin WM, Province MA, Kraja A, Johnson LA, et al. Characterizing the cancer genome in lung adenocarcinoma. Nature. 2007; 450:893–898. [PubMed: 17982442]

Wheeler M, Domin J. Recruitment of the class II phosphoinositide 3-kinase C2beta to the epidermal growth factor receptor: role of Grb2. Mol Cell Biol. 2001; 21:6660–6667. [PubMed: 11533253]

Yoshida K, Sanada M, Shiraishi Y, Nowak D, Nagata Y, Yamamoto R, Sato Y, Sato-Otsubo A, Kon A, Nagasaki M, et al. Frequent pathway mutations of splicing machinery in myelodysplasia. Nature. 2011; 478:64–69. [PubMed: 21909114]

Yuza Y, Glatt KA, Jiang J, Greulich H, Minami Y, Woo MS, Shimamura T, Shapiro G, Lee JC, Ji H, et al. Allele-dependent variation in the relative cellular potency of distinct EGFR inhibitors. Cancer Biol Ther. 2007; 6:661–667. [PubMed: 17495523]

Zang ZJ, Cutcutache I, Poon SL, Zhang SL, McPherson JR, Tao J, Rajasegaran V, Heng HL, Deng N, Gan A, et al. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. Nat Genet. 2012

HIGHLIGHTS

- Exome and genome characterization of somatic alterations in 183 lung adenocarcinomas
- *U2AF1, RBM10* and *ARID1A* are among novel recurrently mutated genes
- Structural variants include activating in-frame fusion of *EGFR*
- Epigenetic & RNA deregulation proposed as a potential lung adenocarcinoma hallmark

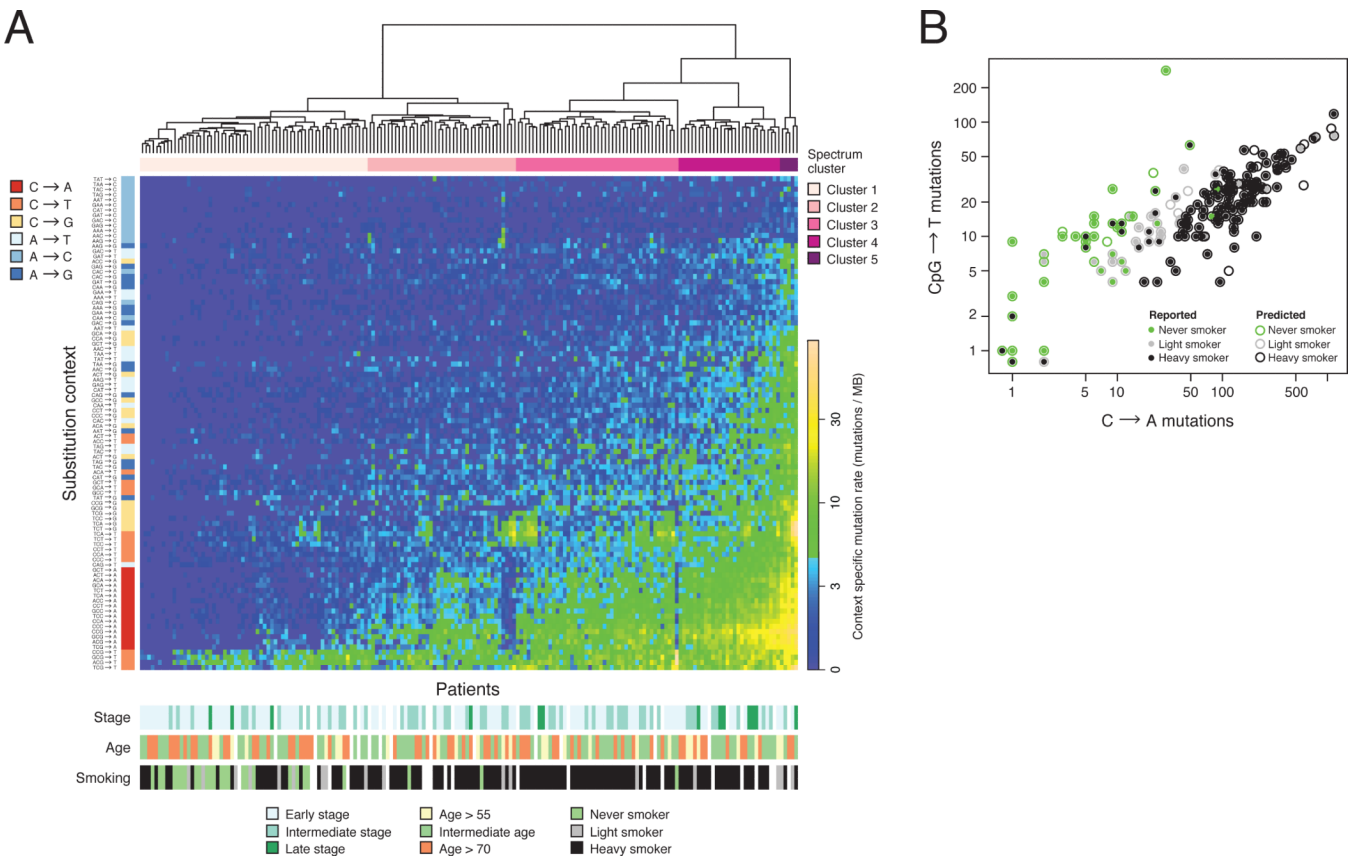

Figure 1. Mutation spectrum analysis of 183 lung adenocarcinomas

(A) Hierarchical clustering of 183 lung adenocarcinomas according to their nucleotide context-specific exonic mutation rates. Each column represents a case, and each row represents one of 96 strand-collapsed trinucleotide context mutation signatures. Top bar: patient-cluster membership. Left bar: simplified single-nucleotide context mutational signature. Bottom bars: reported tumor stage, age, and smoking status for each patient. Right gradient: mutation rate scale. **(B)** Stratification of reported versus imputed smoking status by the log transform of the adjusted ratio of C->A tranversion rates and CpG->T transition rates. The color of each inner solid point represents the reported smoking status for that particular patient. The color of each outer circle indicates that patient's imputed smoking status as predicted by the classifier. Additional analytic details are provided in the Extended Experimental Procedures.

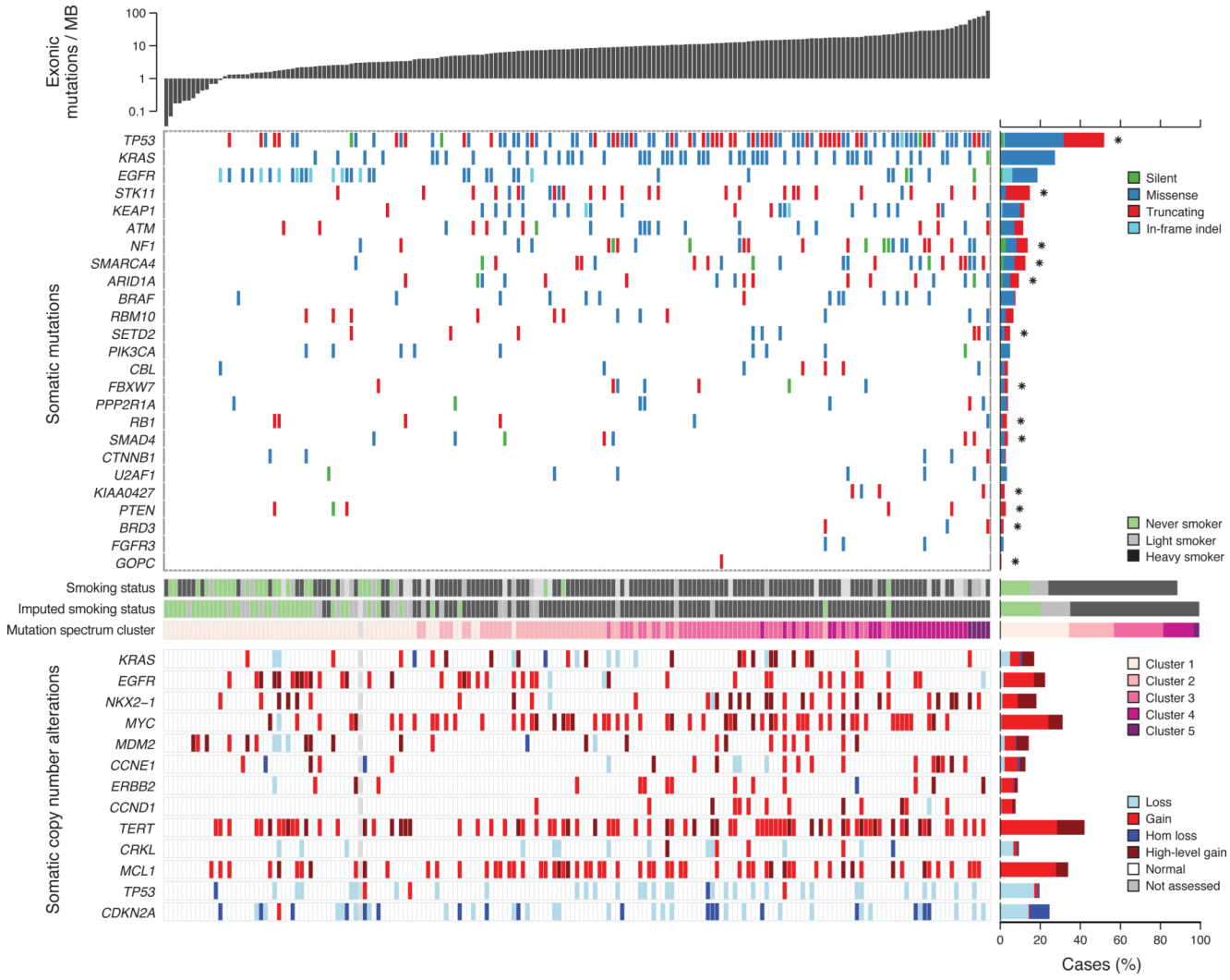

Figure 2. Somatic mutations and copy number changes in 183 lung adenocarcinomas

Top panel, summary of exonic somatic mutation of 25 significantly mutated genes (see text and Table S3 for details). Tumors are arranged from left to right by the number of non-silent mutations per sample, shown in the top track. Significantly mutated genes are listed vertically in decreasing order of non-silent mutation prevalence in the sequenced cohort. Colored rectangles: mutation category observed in a given gene and tumor. Bar chart (right): prevalence of each mutation category in each gene. Asterisks indicate genes significantly enriched in truncating (nonsense, frameshift) mutations. Middle bars: smoking status and mutation spectrum cluster for each patient. White boxes indicate unknown status. Bottom panel: summary of somatic copy number alterations derived from SNP array data. Colored rectangles indicate the copy number change seen for a given gene and tumor.

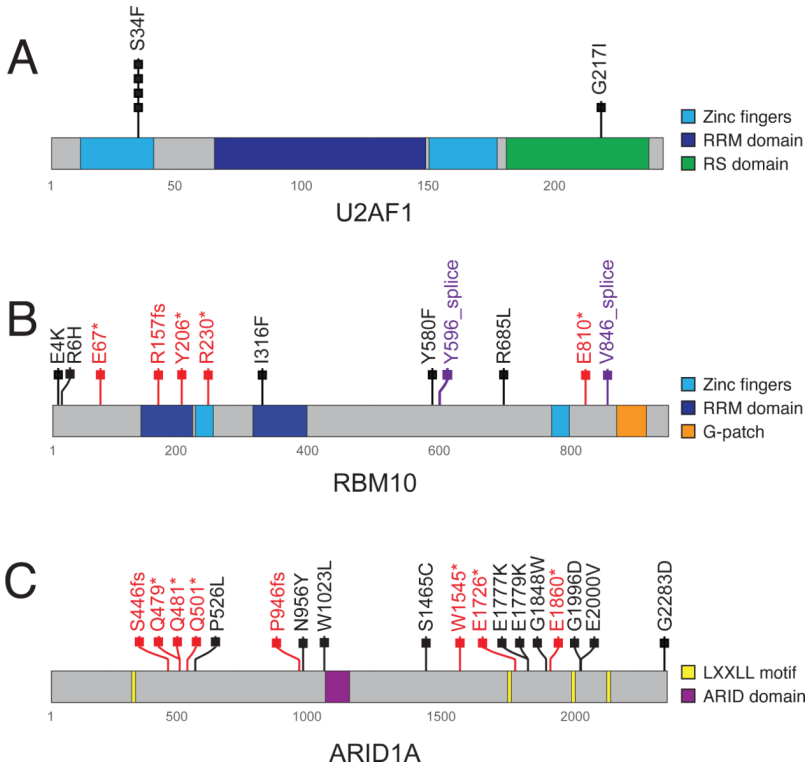

Figure 3. Somatic mutations of lung adenocarcinoma candidate genes *U2AF1*, *RBM10*, and *ARID1A*

(A) Schematic representation of identified somatic mutations in *U2AF1* shown in the context of the known domain structure of the protein. Numbers refer to amino acid residues. Each rectangle corresponds to an independent, mutated tumor sample. Silent mutations are not shown. Missense mutations are shown in black. **(B)** Schematic of somatic *RBM10* mutations. Splice site mutations are shown in purple; truncating mutations are shown in red. Other notations as in (A). **(C)** Schematic of somatic *ARID1A* mutations. Notations as in (A) and (B).

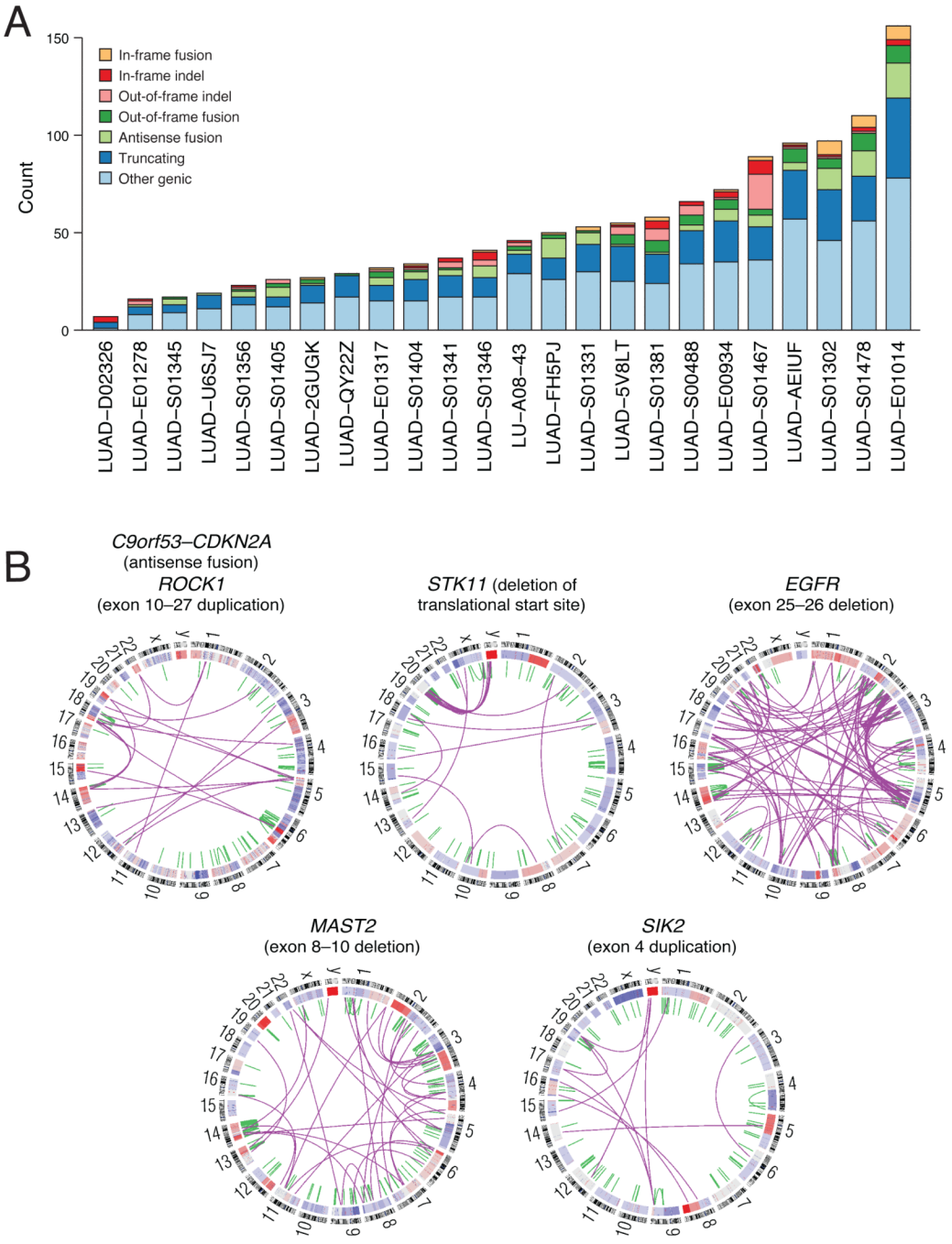

Figure 4. Whole genome sequencing of lung adenocarcinoma

(A) Summary of genic rearrangement types across 25 lung adenocarcinoma whole genomes. Stacked-bar plot depicting the types of somatic rearrangement found in annotated genes by analysis of whole genome sequence data from 25 tumor/normal pairs. The “Other Genic” category refers to rearrangements linking an intergenic region to the 3’ portion of a genic footprint. **(B)** Representative Circos (Krzywinski et al., 2009) plots of whole genome sequence data with rearrangements targeting known lung adenocarcinoma genes *CDKN2A, STK11* and *EGFR* and novel genes *MAST2, SIK2*, and *ROCK1*. Chromosomes are arranged circularly end-to-end with each chromosome's cytobands marked in the outer ring. The inner

ring displays copy number data inferred from whole genome sequencing with intrachromosomal events in green and interchromosomal translocations in purple.

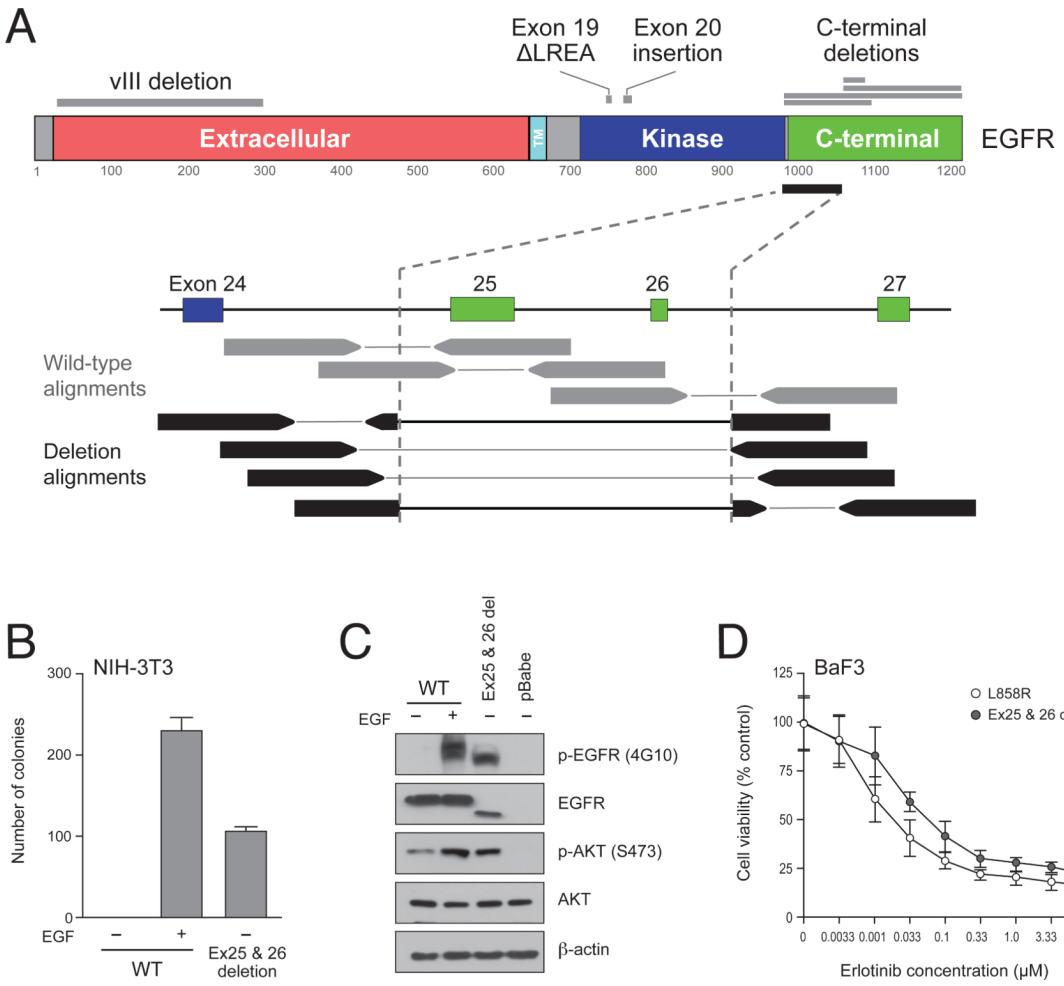

Figure 5. Identification of a novel lung adenocarcinoma in-frame deletion in *EGFR*

(A) Schematic representation of reported *EGFR* alterations (above protein model) for comparison with a C-terminal deletion event found in this study by whole genome sequencing (below protein model). A schematic depiction of sequencing data shows the expected wild-type reads (gray) in contrast with the observed reads (black) spanning or split by the deletion breakpoint. Supporting paired-end and split read mapping data are shown are Figure S5. **(B)** Soft agar colony forming assay of NIH-3T3 cells expressing exon 25 and 26-deleted EGFR (Ex25&26del) or wild-type EGFR in the presence or absence of ligand stimulation. The bar graph shows the number of colonies formed by indicated cells with or without EGF in soft agar (n=3, mean +SD). **(C)** Ex25&26del EGFR is constitutively active in the absence of EGF. The same NIH-3T3 cells used for the assay in (B) were subjected to immunoblotting with anti- phospho-tyrosine (4G10), anti-EGFR and anti-phospho-Akt (S473) antibodies. Blots were probed with anti-Akt and anti-B-actin antibodies (loading control). **(D)** Cell growth induced by the oncogenic EGFR deletion mutant is suppressed by erlotinib treatment. Ba/F3 cells transformed by either L858R or Ex25&26del mutants were treated with increasing concentrations of erlotinib as indicated for 72 hrs and were assayed for cell viability.

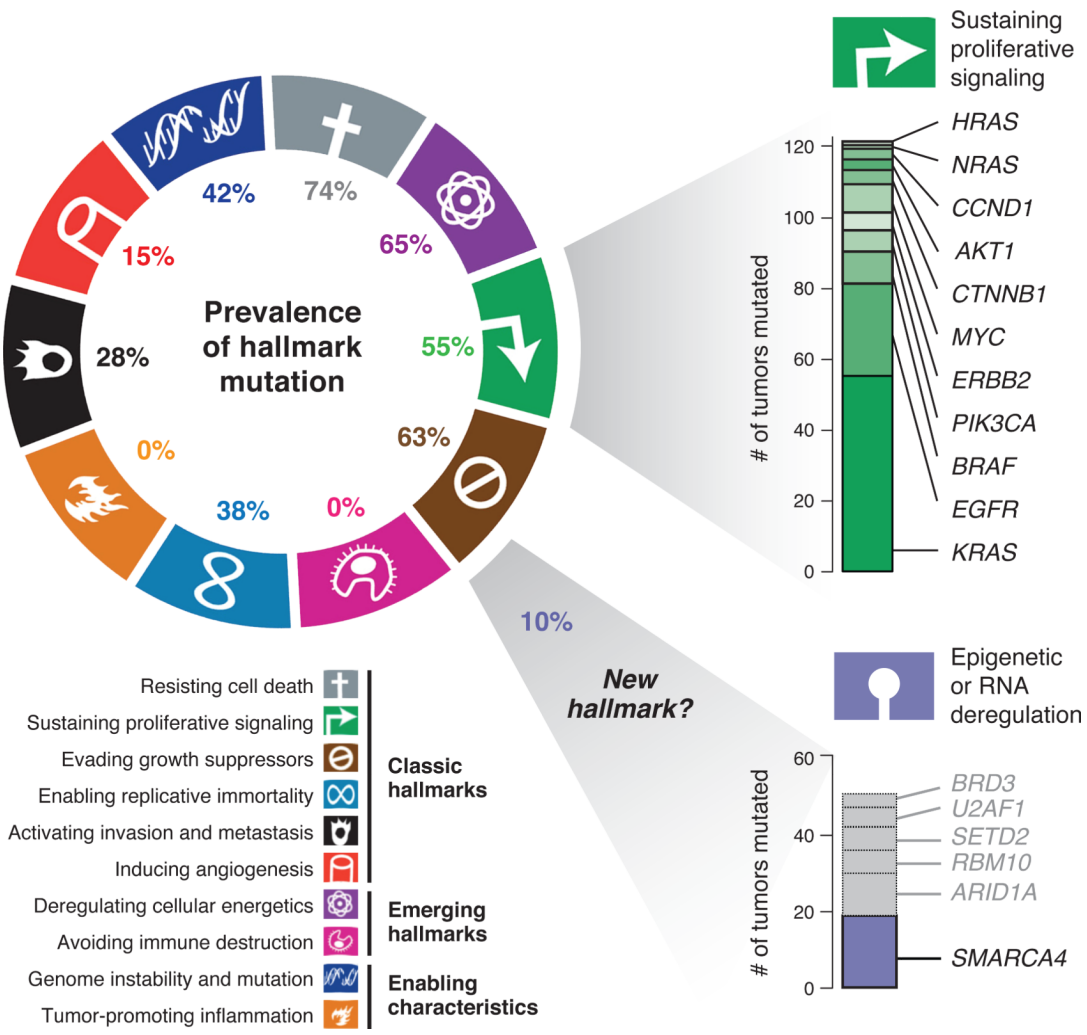

Figure 6. Next-generation hallmarks of lung adenocarcinoma

Left, the prevalence of mutation or SCNA of Sanger Cancer Gene Census (Futreal et al., 2004) genes mapping to cancer hallmarks defined by (Hanahan and Weinberg, 2011). Suspected passenger mutations were filtered out of the analysis, as described in Experimental Procedures. Top right, genes comprising the mutated genes in the hallmark of "sustaining proliferative signaling" are shown. Bottom right, a proposed "11th hallmark" of Epigenetic and RNA deregulation is shown, depicted as above. Genes shown in gray are candidate lung adenocarcinoma genes identified in this study that may additionally contribute to the hallmark.

Table 1
Whole genome and whole exome sequencing statistics

(A) Clinical features and (B) selected sequencing statistics for 183 whole exome (WES) and whole genome sequenced (WGS) cases. “Tumor Target Territory” and “Tumor Target Territory” refers to the exonic territory targeted by the exome capture baitset reported by (Fisher et al., 2011) and used in this study. (B) does not include data on 23 whole exome sequences that were obtained for 23 of 24 WGS cases.

(A)

Age at surgery (median; range)	66 (36-87)
Gender	
Male	95
Female	88
Smoking Status (AJCC 7th Edition)	
Never smoker	27
Smoker > 10 years	118
Smoker ≤ 10 years	17
NA	
Pack years (median; range)	30 (0-128)
Survival	
Follow-up available	135
Follow-up unavailable	48
PFS in months (median; range)	9 (0-63)
Tumor stage	
I	90
II	36
III	22
IV	10
NA	25

(B)

Statistic	Whole Exome Capture	Whole Genome
Tumor normal pairs sequenced	159	24
Total tumor Gb sequenced	1031.6	4946.0
Median fold tumor target coverage (range)	91 (51-201)	69 (25-103)
Median normal fold target coverage (range)	92 (62-141)	36 (28-55)
Median somatic mutation rate per MB in target territory (range)	6.8 (0.3-94.7)	13.3 (4.5-55.3)
Median # of coding mutations per patient (range)	216 (1-3512)	323 (63-2279)
Median # of nonsynonymous mutations per patient (range)	167 (1-2721)	248 (53-1770)
Median # of transcribed non-coding mutations per patient (range)	187 (13-2559)	18,314 (4,632-100,707)
Total # of number of structural rearrangements	N/A	2349
Total # of number of frame-preserving genic rearrangements	N/A	71
Total # of number of frame-abolishing genic rearrangements		235

(B)

Statistic	Whole Exome Capture	Whole Genome
Median # of genes powered at 20% exonic territory (range)	15647 (15046 - 16019)	16905 (10136 – 16952)
Median # of genes powered at 50% exonic territory (range)	6788 (6078 – 7402)	8771 (2634 – 8863)