Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1016/j.cell.2012.08.029">http://dx.doi.org/10.1016/j.cell.2012.08.029</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Elsevier</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Fri Dec 21 22:26:10 EST 2018</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/91638">http://hdl.handle.net/1721.1/91638</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Article is made available in accordance with the publisher's policy and may be subject to US copyright law. Please refer to the publisher's site for terms of use.</td>
</tr>
</tbody>
</table>

Detailed Terms
Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing

Marcin Imielinski,1,2,3,5,18 Alice H. Berger,1,5,18 Peter S. Hammerman,1,5,18 Bryan Hernandez,1,18 Trevor J. Pugh,1,5,18 Eran Hodis,1 Jeonghee Cho,6 James Suh,7 Marzia Capelletti,5 Andrey Sivachenko,1 Carrie Sougnez,1 Daniel Auclair,1 Michael S. Lawrence,1 Petar Stojanov,1,5 Kristian Cibulskis,1 Kyusum Choi,6 Luc de Waal,1,5 Tanaz Shariffnia,1,6 Angela Brooks,1,6 Heidi Greulich,1,6 Shantanu Banerji,1,5 Thomas Zander,1,11 Danila Seidel,9 Frauke Leenders,9 Sascha Ansén,9 Corinna Ludwig,9 Walburga Engel-Riedel,9 Erich Stoeiben, Jürgen Wolf,9 Chandra Goparaju,9 Kristin Thompson,1 Wendy Winckler,7 David Kwiatkowski,5 Bruce E. Johnson,5 Pasi A. Jänne,5 Vincent A. Miller,12 William Pao,14 William D. Travis,13 Harvey I. Pass,8 Stacey B. Gabriel,1 Eric S. Lander,1,4,15 Roman K. Thomas,6,10,11,16,17 Levi A. Garraway,1,5 Gad Getz,1 and Matthew Meyerson1,3,5,*

1Broad Institute of Harvard and MIT, 7 Cambridge Center, Cambridge, MA 02142, USA
2Department of Pathology, Massachusetts General Hospital, 55 Fruit Street, Boston, MA 02114, USA
3Department of Pathology
4Department of Systems Biology
Harvard Medical School, 25 Shattuck Street, Boston, MA 02115, USA
5Department of Medical Oncology, Dana Farber Cancer Institute, 450 Brookline Avenue, Boston, MA 02115, USA
6Samsung Research Institute, Samsung Medical Center, Seoul 135-967, Republic of Korea
7Department of Pathology
8Department of Cardiothoracic Surgery
Langone Medical Center, New York University, New York, NY 10016, USA
9Department of Internal Medicine and Center for Integrated Oncology
Köln-Bonn, University of Cologne, 50924 Cologne, Germany
10Laboratory of Translational Cancer Genomics
Max Planck Institute for Neurological Research, 50924 Cologne, Germany
11Thoracic Oncology Service, Department of Medicine
Köln-Bonn, University of Cologne, 50924 Cologne, Germany
12Department of Pathology
Memorial Sloan Kettering Cancer Center, New York, NY 10065, USA
13Division of Hematology/Oncology, Vanderbilt-Ingram Cancer Center, Nashville, TN 37232, USA
14Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
15Department of Translational Genomics, University of Cologne, Weyertal 115b, 50931 Cologne, Germany
16Department of Pathology, University of Cologne, Kerpener Strasse 62, 50937 Cologne, Germany
17These authors contributed equally to this work
*Correspondence: matthew_meyerson@dfci.harvard.edu
http://dx.doi.org/10.1016/j.cell.2012.08.029

SUMMARY

Lung adenocarcinoma, the most common subtype of non-small cell lung cancer, is responsible for more than 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 rearrangements, 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 more than 1.3 million deaths per year, of which more than 40% are lung adenocarcinomas (World Health Organization, 2012; Travis, 2002). Most often, tumors are discovered as locally advanced or metastatic disease, and despite improvements in molecular diagnosis and targeted therapies, the average 5 year...
survival rate for lung adenocarcinoma is ~15% (Minna and Schiller, 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 ~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, \( TLR4 \), \( LPHN3 \), \( GRM1 \), and \( GLI \)), 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 that likely contribute to pathogenesis. Together, these data represent a significant advance toward a comprehensive annotation of somatic alterations in lung adenocarcinoma.

RESULTS

Patient Cohort Description

We sequenced DNA from 183 lung adenocarcinomas and matched normal tissues by 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 (more 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-naive, primary resection specimens except for one case with whole-genome sequence data (LU-A08-43) that was a postchemotherapy 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 (available online).

Table 1. Summary of Clinical Features

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age at Surgery (Median; Range)</td>
<td>66 (36–87)</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>male</td>
<td>95</td>
</tr>
<tr>
<td>female</td>
<td>88</td>
</tr>
<tr>
<td>Smoking Status (AJCC 7th Edition)</td>
<td></td>
</tr>
<tr>
<td>never-smoker</td>
<td>27</td>
</tr>
<tr>
<td>smoker &gt;10 years</td>
<td>118</td>
</tr>
<tr>
<td>smoker ≤10 years</td>
<td>17</td>
</tr>
<tr>
<td>n/a</td>
<td>21</td>
</tr>
<tr>
<td>pack years (median; range)</td>
<td>30 (0–128)</td>
</tr>
<tr>
<td>Survival</td>
<td></td>
</tr>
<tr>
<td>follow-up available</td>
<td>135</td>
</tr>
<tr>
<td>follow-up unavailable</td>
<td>48</td>
</tr>
<tr>
<td>PFS in months (median; range)</td>
<td>9 (0–63)</td>
</tr>
<tr>
<td>Tumor Stage</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>90</td>
</tr>
<tr>
<td>II</td>
<td>36</td>
</tr>
<tr>
<td>III</td>
<td>22</td>
</tr>
<tr>
<td>IV</td>
<td>10</td>
</tr>
<tr>
<td>n/a</td>
<td>25</td>
</tr>
</tbody>
</table>

Distribution of selected clinical variables from 183 lung adenocarcinoma cases.

Mutation Detection and Validation

We examined 183 lung adenocarcinoma tumor/normal pairs with a combination of whole-exome sequencing (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 Table 2 and 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 by 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
We validated 86% of 338 indels and 98% of 8,912 substitutions identified by WGS at sufficiently powered sites in the corresponding WES tumor sample (Figure S1, Table S2A, and Extended Experimental Procedures). To validate mutations from cases with only WES data, we randomly selected 69 candidate mutations for ultradeep (>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 met or exceed those reported in similar sequencing studies (Banerji et al., 2012; Berger et al., 2012; Gerlinger et al., 2012; Nikolaev et al., 2012; Stransky et al., 2011; Cancer Genome Atlas Research 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 test), 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., 2012; Cancer Genome Atlas Research 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 five mutation spectrum clusters. These clusters represented grades of increasing mutational complexity: cluster 1 was enriched for CpG→T mutations and was 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; and 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 \( (p = 1.9 \times 10^{-9}, \) Fisher’s exact test), whereas 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 Procedures), we imputed smoking status for 21 patients who lacked reported smoking history and accurately recapitulated reported smoking status for more than 75% of the remaining cases (Figure 1B). Exonic and intronic mutation rates, context-specific mutation counts, imputed

### Table 2. Whole-Genome and Whole-Exome Sequencing Statistics

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Whole-Exome Capture</th>
<th>Whole Genome</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tumor/normal pairs sequenced</td>
<td>159</td>
<td>24</td>
</tr>
<tr>
<td>Total tumor Gb sequenced</td>
<td>1,031.6</td>
<td>4,946.0</td>
</tr>
<tr>
<td>Median fold tumor target coverage (range)</td>
<td>91 (51–201)</td>
<td>69 (25–103)</td>
</tr>
<tr>
<td>Median normal fold target coverage (range)</td>
<td>92 (62–141)</td>
<td>36 (28–55)</td>
</tr>
<tr>
<td>Median somatic mutation rate per Mb in target territory (range)</td>
<td>6.8 (0.3–94.7)</td>
<td>13.3 (4.5–55.3)</td>
</tr>
<tr>
<td>Median number of coding mutations per patient (range)</td>
<td>216 (1–3,512)</td>
<td>323 (63–2,279)</td>
</tr>
<tr>
<td>Median number of nonsynonymous mutations per patient (range)</td>
<td>167 (1–2,721)</td>
<td>248 (53–1,770)</td>
</tr>
<tr>
<td>Median number of transcribed noncoding mutations per patient (range)</td>
<td>187 (13–2,559)</td>
<td>18,314 (4,632–100,707)</td>
</tr>
<tr>
<td>Total number of structural rearrangements</td>
<td>n/a</td>
<td>2,349</td>
</tr>
<tr>
<td>Total number of frame-preserving genic rearrangements</td>
<td>n/a</td>
<td>71</td>
</tr>
<tr>
<td>Total number of frame-abolishing genic rearrangements</td>
<td>235</td>
<td></td>
</tr>
<tr>
<td>Median number of genes powered at 20% exonic territory (range)</td>
<td>15,647 (15,046–16,019)</td>
<td>16,905 (10,136–16,952)</td>
</tr>
<tr>
<td>Median number of genes powered at 50% exonic territory (range)</td>
<td>6,788 (6,078–7,402)</td>
<td>8,771 (2,634–8,863)</td>
</tr>
</tbody>
</table>

Selected sequencing statistics for 183 WES and WGS cases. “Tumor Target Territory” refers to the exonic territory targeted by the exome capture bait set reported by (Fisher et al., 2011) and used in this study. The “Whole-Exome Capture” column does not include data on 23 cases analyzed by both WES and WGS.

### Notes

- A median of 8.1 mutations/Mb and a mean of 11.9 mutations/Mb (range: 0.04–117.4). These comprised 43,813 missense, 14,801 silent, 3,504 nonsense, 1,460 splice-site, 2,310 deletions, 839 insertions, and 11,009 other mutations (predominantly residing in 5’ and 3’ untranslated regions [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 indels and 97% of 9,354 substitutions identified by WGS at sufficiently powered sites in the corresponding WES tumor sample. In the converse analysis, we validated 86% of 338 indels and 98% of 8,912 substitutions from WES at sufficiently powered sites in the corresponding WGS tumor sample (Figure S1, Table S2A, and Extended Experimental Procedures). To validate mutations from cases with only WES data, we randomly selected 69 candidate mutations for ultradeep (>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., 2012; Stransky et al., 2011; Cancer Genome Atlas Research Network, 2011; Totoki et al., 2011; Zang et al., 2012).

#### References

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; Cancer Genome Atlas Research Network, 2012) present a challenge for unbiased discovery of mutated genes undergoing positive somatic selection. More than 13,000 of 18,616 genes with adequate sequence coverage had nonsynonymous somatic mutations in at least one tumor, and more than 3,000 were mutated in at least five 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 more than 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 and 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-Böckler 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 noncoding mutations detected by both WES and WGS. InVEx permutes coding, untranslated, and intronic mutations within covered territories of each

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 $C\rightarrow A$ tranversion rates and $CpG\rightarrow T$ 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. See also Figure S1 and Tables S2 and S3.
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 Multiarray Average [RMA] value ≥ 5) in a panel of 40 lung adenocarcinoma cell lines (Barrettina 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 nonsynonymous mutation frequencies that were 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 five 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 were mutated at frequencies lower than most previously reported genes, demonstrating the increased power of our large sample set. The LOF InVEx yielded six significantly mutated genes (q < 0.25), including BRD3, which is 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 (Tables S3C and S3D). 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 adenocarcinoma (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 (Figures S3A–S3C). The remaining six reported lung adenocarcinoma genes (CDKN2A, ERBB2, AKT1, NRAS, HRAS, and 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 rearranged (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, LPN3, GRM1, and 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, and stage), genomic variables (mutation rate, mutation spectrum cluster, and 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 Hallmarks Analysis in the Experimental Procedures 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 mutations were significantly anticorrelated with KRAS mutations (p = 3.3 × 10⁻⁹), and somatic mutation rate (p = 5.9 × 10⁻⁴) EGFR mutations significantly correlated with never-/light smoker status (p = 2.0 × 10⁻⁵), 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 anticorrelated with both spectrum cluster 1 and imputed never-/light smoking status (p < 0.005). These findings are consistent with reported associations (Kolvunen et al., 2008; Pao et al., 2004, 2005; Siebos et al., 1991). In addition, NF1 mutations were significantly depleted in spectrum cluster 1 (p = 4 × 10⁻⁵) 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. \textit{U2AF1} (p = 0.00011, log rank test) and \textit{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 \textit{U2AF1} (p = 2.0 \times 10^{-6}, PPH2 InVEx), which had nonsynonymous mutations in 3% of cases (Figure 3A). Identical c.101C > T, p.S34F mutations were seen in four of five \textit{U2AF1} mutant cases (Figure 3A); this is 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 \textit{U2AF1} mutations in an epithelial tumor. One of four p.S34F mutations occurred with an activating event in \textit{KRAS} (p.Q61H), suggesting that \textit{U2AF1} mutations may confer tumorigenic capability independent of known proliferation-sustaining driver genes. As mentioned above, four patients with \textit{U2AF1} mutations and survival data had significantly reduced PFS (Figure S3E). Nonsynonymous mutations in genes encoding other members of the spliceosome complex (including \textit{SF3B1}, \textit{U2AF2}, and \textit{PRPF40B}) were found in 14 additional cases (Yoshida et al., 2011).

\textit{RBM10} was frequently mutated (12/183 cases; 7%) and subject to recurrent nonsense, frameshift, or splice-site mutations,
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). See also Figure S3.

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 intrachromosomal rearrangements (1,818 events) but included 531 interchromosomal events. Among these were 1,443 (61.4%) genic rearrangements (i.e., in which one breakpoint was contained within the promoter, UTR, 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 and Table S1) or mutation spectrum. Rearrangement coordinates and interpretations are provided as 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 (Cancer Genome Atlas Research 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 (midexon rearrangement) (Figure 4B).

We next focused on potentially activating in-frame rearrangements of kinase genes. This analysis uncovered a two-exon deletion in EGFR, which was previously identified in glioblastoma multiforme but is novel in lung adenocarcinoma, ablating a portion of the C terminus of EGFR encoded by exons 25 and 26 (Figures 4B, 5A, and S5), including residues associated with interaction with PIK3C2B (Wheeler and Domin, 2001) and CBL (Grøvdal et al., 2004). Similar C-terminal deletion variants (EGFR vIVb) have been previously identified in glioblastoma (Ekstrand et al., 1992) and have been 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...
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 (Krzewinski 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 WGS with intrachromosomal events in green and interchromosomal translocations in purple.

See also Figure S4 and Table S5.
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). Overexpression 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 quantitative PCR (qPCR). The duplication occurred 15 amino acids downstream of the kinase domain, and it was not detected in normal tissues or in lymphocytes from patients with lung adenocarcinoma. The duplication was confirmed by Sanger sequencing and was observed in 12% of lung adenocarcinoma tumors (Figure 4C).

See also Figure S5.
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, which is 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 six 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 and 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 (Figures 6 and S6). Less than half (47%) of our cohort harbored a mutation in a known driver gene for this hallmark, and only slightly more (55%) did so when including high-level amplification in one or more proliferative signaling genes (e.g., EGFR, ERBB2, and 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 (Elsässer 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 eleventh 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 genespecific 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, and 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 sequencing (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 toward 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 six 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 or collaborator or was 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.D.T. Informed consent (Institutional Review Board) was obtained for each sample by using protocols approved by the Broad Institute of Harvard and MIT and each originating tissue source site.

Massively Parallel Sequencing

Exome capture was performed by using Agilent SureSelect Human All Exon 50 Mb according the manufacturer’s instructions. All WES and WGS was performed on the Illumina HiSeq platform. Basic alignment and sequence quality control were done by using the Picard and Firehose pipelines at the Broad Institute. Mapped genomes were processed by the Broad Firehose pipeline to perform additional quality control, variant calling, and mutational significance analysis.

External Data

Gene expression data for 40 lung adenocarcinoma cell lines were obtained from the Cancer Cell Line Encyclopedia (CCLE) (http://www.broadinstitute.org/ccle)
org/ccle/home) as 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 by using InVex (Hodis et al., 2012) with PolyPhen-2-based (Adzhubei et al., 2010) and 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 by using the MutSig method from Berger et al. (2011) and was 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 by Wilcoxon rank sum test for dichotomous variables versus numeric data (e.g., mutation status versus total mutation rate). All remaining statistical computing, including cluster analysis and visualization, was performed by 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 $j$ 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 hot spot (Forbes et al., 2011). Mutations that were present in the COSMIC database (http://www.sanger.ac.uk/genetics/CGP/cosmic/) at least ten times were considered oncogenic mutations. We considered a dominant gene activated if it harbored such a variant or with 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 $k$.

ACCESSION NUMBERS
The dbGAP accession number for the data reported in this paper is phs000488.v1.p1.

SUPPLEMENTAL INFORMATION
Supplemental Information includes Extended Experimental Procedures, six figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.08.029.

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 the EU-Framework Programme CURELUNG (HEALTH-F2-2010-258677), by the Max Planck Society (M.I.F.A.NEUR8061), by the Deutsche Forschungsgemeinschaft (DFG) through SFB883 (T6) and grant TI1386-3/1, by the EU-Framework Programme CURELUNG (HEALTH-F2-2010-258677), by the Cancer Genome Atlas Research Network (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 the National Cancer Institute, Uniting Against Lung Cancer, the Lung Cancer Research Foundation, and the American Lung Association (M.M.). M.M., E.L., and L.A.G. are founders and equity holders of Foundation Medicine, a for-profit company that provides next-generation sequencing diagnostic services. V.A.M. is an employee of Foundation Medicine.

Received: May 8, 2012
Revised: July 27, 2012
Accepted: August 27, 2012
Published: September 13, 2012

REFERENCES


