Cell

Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers

Ramaswamy Govindan,[1,2,9] Li Ding,[1,3,4,9] Malachi Griffith,[3,4] Janakiraman Subramanian,[1,2] Nathan D. Dees,[3] Krishna L. Kanchi,[3] Christopher A. Maher,[1,2,3] Robert Fulton,[3,4] Lucinda Fulton,[3,4] John Wallis,[3,4] Ken Chen,[8] Jason Walker,[3] Sandra McDonald,[2,6] Ron Bose,[1,2] David Ornitz,[5] Donghai Xiong,[7] Ming You,[7] David J. Dooling,[3,4] Mark Watson,[2,6] Elaine R. Mardis,[2,3,4] and Richard K. Wilson[2,3,4,*]

[1]Department of Internal Medicine, Division of Oncology
[2]Siteman Cancer Center
[3]The Genome Institute
[4]Department of Genetics
[5]Department of Developmental Biology
[6]Department of Pathology and Immunology
Washington University School of Medicine, St. Louis, MO 63110, USA
[7]Department of Pharmacology and Toxicology, Medical College of Wisconsin, Milwaukee, WI 53226, USA
[8]Department of Bioinformatics and Computational Biology, Division of Quantitative Sciences, The University of Texas MD Anderson Cancer Center, Houston, TX 77030, USA
[9]These authors contributed equally to this work
*Correspondence: rwilson@wustl.edu
http://dx.doi.org/10.1016/j.cell.2012.08.024

SUMMARY

We report the results of whole-genome and transcriptome sequencing of tumor and adjacent normal tissue samples from 17 patients with non-small cell lung carcinoma (NSCLC). We identified 3,726 point mutations and more than 90 indels in the coding sequence, with an average mutation frequency more than 10-fold higher in smokers than in never-smokers. Novel alterations in genes involved in chromatin modification and DNA repair pathways were identified, along with *DACH1*, *CFTR*, *RELN*, *ABCB5*, and *HGF*. Deep digital sequencing revealed diverse clonality patterns in both never-smokers and smokers. All validated *EFGR* and *KRAS* mutations were present in the founder clones, suggesting possible roles in cancer initiation. Analysis revealed 14 fusions, including ROS1 and ALK, as well as novel metabolic enzymes. Cell-cycle and JAK-STAT pathways are significantly altered in lung cancer, along with perturbations in 54 genes that are potentially targetable with currently available drugs.

INTRODUCTION

Lung cancer is a leading cause of cancer-related death globally (Ferlay et al., 2010). Non-small cell lung cancer (NSCLC), the most common type, comprises three histological subtypes: adenocarcinoma, squamous cell carcinoma, and large cell carcinoma. Approximately 10%–40% of patients diagnosed with lung cancer report no history of tobacco smoking. The proportion of patients who are lifelong never-smokers is higher in parts of Asia (Subramanian and Govindan, 2007). Environmental and occupational exposures (Ng, 1994), as well as genetic susceptibility (Sellers et al., 1990; Yang et al., 1999), are thought to contribute to lung cancer risk in never-smokers.

Inhibitors of epidermal growth factor receptor (EGFR) tyrosine kinase (TK), gefitinib and erlotinib, have shown substantial activity in patients whose tumor cells harbor specific mutations in the EGFR TK domain (Lynch et al., 2004). The recent discovery of a fusion kinase involving the EMAP-like protein 4 (*EML4*) and anaplastic lymphoma kinase (*ALK*) genes in tumor specimens from some patients with NSCLC (mostly adenocarcinoma) and the dramatic response to crizotinib, as well as the identification of fusion kinases involving *RET* and *ROS1*, have reinvigorated efforts to identify novel genomic alterations that could be therapeutic targets (Kohno et al., 2012; Lipson et al., 2012; Soda et al., 2007; Takeuchi et al., 2012). EGFR TK domain mutations and fusion kinases involving *EML4-ALK* are present more often in the tumor specimens from lifelong never-smokers than from smokers (Soda et al., 2007; Subramanian and Govindan, 2007).

In our previous studies, SNP-array-based analysis of 371 lung adenocarcinomas has previously revealed 57 significant copy number alterations (Weir et al., 2007), including the most common amplification of *TITF1*, a lineage-specific transcription factor responsible for lung development (Kendall et al., 2007). In addition, sequencing of the coding exons of 623 candidate cancer genes in 188 lung adenocarcinomas identified 26 significantly mutated genes in lung adenocarcinoma, consisting of a set of oncogenes (*EGFR*, *KRAS*, ephrin receptor genes, *ERBB4*, *KDR*, *FGFR4*, and NTRK genes) and tumor

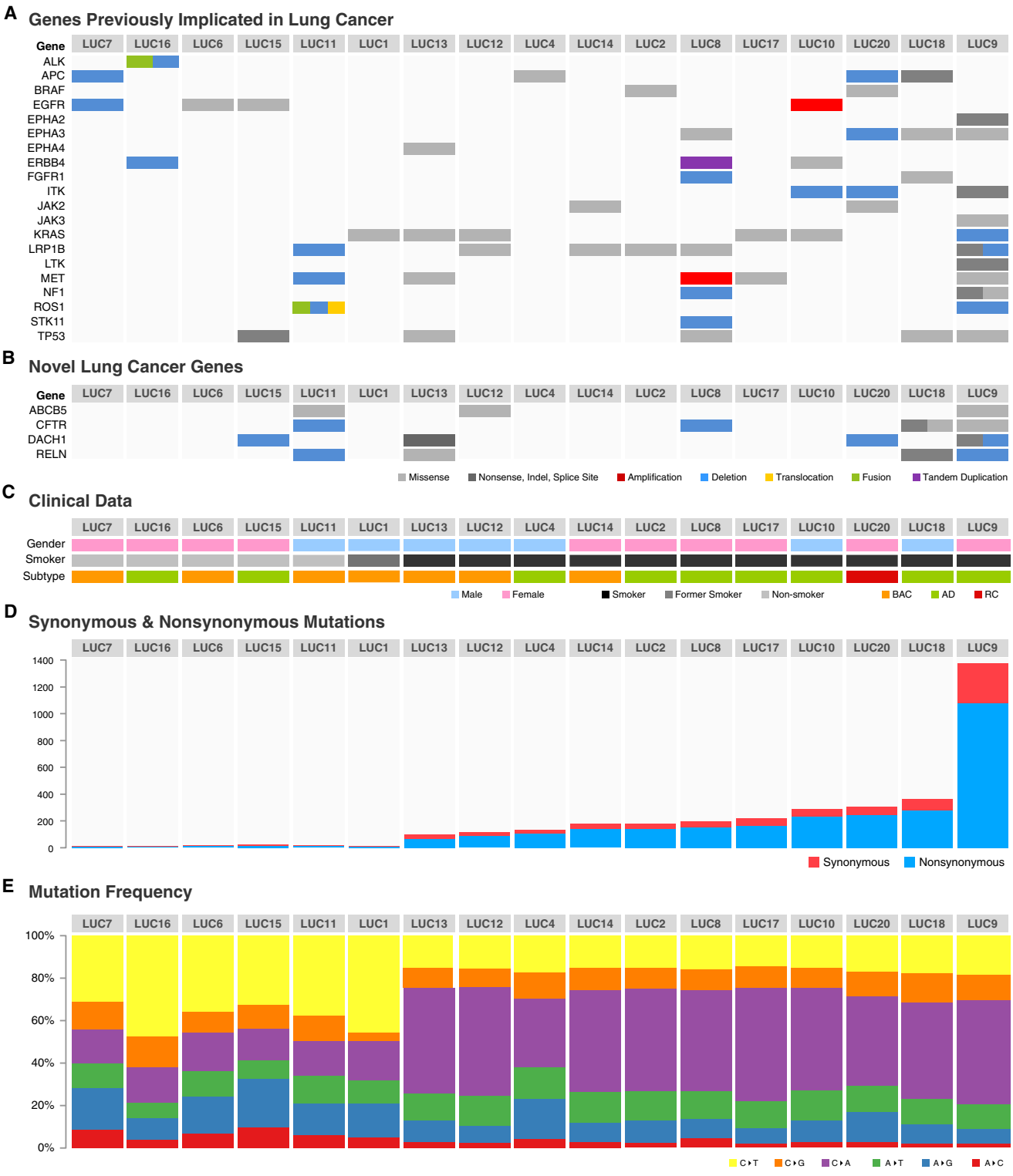

Figure 1. Mutation Landscape in Lung Cancer

(A and B) A heatmap of significant genetic events in 17 NSCLC samples is provided for both (A) genes previously implicated lung cancer and (B) novel genes found to be recurrently altered in the present study. Events, including point mutations, truncation mutations, copy number gains and losses, and larger structural variations are color coded according to the legend provided.

(C) Clinical characteristics of the 17 NSCLC patients.

suppressors (*TP53*, *STK11*, *NF1*, *RB1*, *ATM*, and *APC*) (Ding et al., 2008).

Clearly, discovering novel genetic alterations in lung cancer, from point mutations to large structural variants, requires a comprehensive genome-wide approach. We report a sequencing-based study of tumor specimens from 16 patients with adenocarcinoma and one patient with large cell carcinoma of the lung by using whole-genome and transcriptome sequencing. We identified several novel point mutations and novel fusions that are potentially targetable for therapy. Deep digital sequencing of somatic mutations, for the first time, revealed that lung cancers from both smokers and never-smokers are often heterogeneous and consist of subclonal populations. Our findings highlight the importance of comprehensive and integrated analysis of the genome and transcriptome of lung cancers for identifying novel pathways and therapeutic targets.

RESULTS

Study Design and Case Descriptions

Tumor and adjacent normal tissue samples for whole-genome sequencing were obtained from patients diagnosed with NSCLC who underwent definitive surgical resection prior to receiving chemotherapy or radiation at the Alvin J. Siteman Cancer Center at Washington University School of Medicine. All samples were subjected to pathology review to establish the histologic diagnosis and tumor cellularity. Only samples with tumor nuclei greater than or equal to 50% of total cellular nuclei in the section were utilized for this study. We identified 17 patients who met all of the above criteria, resulting in a cohort comprising 16 tumors with adenocarcinoma histology and one with large cell carcinoma histology. The median age of patients was 63 years (range 24–77). Five patients included in the study reported no history of tobacco smoking (referred to as ''never-smokers'' hereafter), and one patient had a very light history of tobacco smoking (ten packs per year), having quit smoking 38 years before developing lung cancer (referred to as ''former light smoker'' hereafter) (Figure 1). Clinical characteristics including tumor stage, treatment received, and outcome are provided in Table S1 (available online). Histological images are provided in Data S1. The study was approved by the Human Research Protection Office (HRPO) at the Washington University School of Medicine.

The initial data set for this study included whole-genome sequencing (WGS) paired-end sequencing data generated using 17 lung cancer (LUC) tumor-normal pairs, with haploid coverage ranging from 25.03- to 64.49-fold (Table S2). Point mutations, small (<30 bp) indels, copy number alterations, and structural variants (SVs) were discovered by using various computational approaches (Chen et al., 2009; Larson et al., 2012; Li et al., 2009; McKenna et al., 2010; Ye et al., 2009). Point mutations and indels identified by WGS were classified into four tiers as described previously (Mardis et al., 2009) (Extended Experimental Procedures and Tables S3 and S4). Custom sequence capture arrays were used to validate putative WGS mutations (Table S5). Variants of interest identified by WGS were extended by recurrence screening in an independent set of 94 primary lung adenocarcinomas (Table S6). RNA sequencing (RNA-seq) data were generated for all 17 tumors and a single, matched normal adjacent tissue to LUC9, with 11,578–14,507 genes detected as expressed in each tumor (Table S7).

Genomic Landscape of Lung Cancer in Relation to Tobacco Smoking

Substantial differences in the mutational burden, spectrum, and affected genes were found between smokers and never-smokers (Figure 1). Of the 12 samples from tobacco smokers (including the former light smoker), we observed one cancer genome (LUC9) with a significantly higher number of point mutations (tier 1: 1,363) when compared to the other tumor samples (Figure 1). This sample meets our criterion for ''hypermutation,'' which is defined as having a total number of tier 1 mutations at least 2 SD greater (1 SD = 329) than the rest of the samples. The total number of point mutations (tiers 1–3) was much higher in tobacco smokers (median 15,659, range 7,424–26,202, LUC9 not included) relative to never-smokers (median 888, range 842 –1,268). Similarly, the total number of point mutations involving coding regions (tier 1) also was much higher in smokers (median 209, range 104–1363) compared to never-smokers (median 18, range 10–22). The total number of point mutations in the former light smoker was 403 in tiers 1–3, with only 10 in tier 1 (Table S8). Consistent with previous reports (Ding et al., 2008; Lee et al., 2010), lung cancer due to tobacco smoking is associated with a significantly higher number of mutations per Mb (mutations per Mb: median 10.5, range 4.9–17.6, LUC9 not included) compared to never-smokers with lung cancer (mutations per Mb: median 0.6, range 0.6–0.9) and a single former light smoker with lung cancer (0.3 mutations per Mb) in our study. Figure 1 illustrates the different characteristics of mutations in patients according to their smoking status. In particular, C:G→A:T transversions were noted predominantly in tobacco smokers, whereas C:G→T:A transitions were the most frequent type of point mutations in never-smokers with lung cancer and the former light smoker, which is consistent with previously reported studies (Ding et al., 2008; Lee et al., 2010). The mutational spectrum of the single large cell carcinoma sample was not different from those of lung adenocarcinoma associated with tobacco smoking. Overall, the number of point mutations in the lung cancer genome appears to be closely related to the patient's tobacco smoking status, and the landscape of the former light smoker genome suggests a possible dose-response relationship between the amount and duration of tobacco smoke exposure and the extent of mutational burden. The hypermutated tumor (LUC9) was found to have point mutations involving several DNA repair genes, including *PRKDC*, *TP53*, *MSH3*, *POLK*, *MSH4*, *FANCM*,

(D) A stacked bar graph representing the total number of tier 1 mutations in each patient, color proportioned by the number of synonymous versus non-synonymous mutations.

(E) A stacked bar graph representing the frequency of each type of base substitution for all tier 1 point mutations in 17 NSCLC genomes.

See also Figure S1, Data S2 and S3, and Tables S1, S2, S3, S4, S5, S6, S8, S9, S10, S12, S13, S16, and S17.

FBXW7, *TOP2B*, *MLH1*, *RPA2*, *BUB1*, *FANCB*, and *TOP1* (Wood et al., 2001) (http://www.genesilico.pl/index.php/home.html). It is possible that these mutations in DNA repair genes resulted in an impaired ability to repair sustained DNA damage induced by chronic tobacco smoke.

Somatically Mutated Genes in Lung Cancer

Recurrent Mutations Previously Reported in Lung Cancer

Given the limited sample size of our study, to prioritize additional important mutations, we used an alternative analysis focusing on tier 1 mutations previously reported in lung cancer as reported in the Catalogue of Somatic Mutations in Cancer (COSMIC) (http://www.sanger.ac.uk/genetics/CGP/cosmic/). In addition to the well-known mutations involving *KRAS*, *EGFR*, and *TP53* genes, this approach revealed several other recurrent point mutations in kinase genes that may serve as potential therapeutic targets, including *BRAF* (D594N and V413L), *JAK2* (V615L and M532V), *JAK3* (A1090S), *EPHA3* (M320I, G187R, T393K, and R728L), *EPHA4* (E670D), *STK11* (D327fs), *LTK* (R669*), *MET* (Q99L and Y1003N), and *ITK* (Y588*) (Figure 1 and Table S3).

Novel Significantly Mutated Genes in Lung Cancer

We previously developed the significantly mutated gene (SMG) algorithm to detect, in an unbiased manner, biologically significant variants from cancer genome sequencing data (Dees et al., 2012) (Figure 1). The statistical significance of mutations in each gene is determined by comparing the mutation frequency of each gene with the background mutation rate across all samples. The algorithm identified nine genes that were highly significant (Table S9, false discovery rate $q \leq 0.05$ for two tests, see Extended Experimental Procedures). We did not find any correlations between gender and mutations (Table S10).

Of the nine SGMs, mutations involving *DACH1*, *RELN*, and *ABCB5* genes have not been previously reported in lung cancer. Low *DACH1* expression levels were associated with poor prognosis in patients with breast cancer (Wu et al., 2006). *DACH1* has been reported to have a tumor suppressor role in prostate cancer and in gliomas (Watanabe et al., 2011; Wu et al., 2009). In our study, two frameshift mutations (LUC9: K636fs and LUC13: A656fs) in the coiled-coil domain (CCD) of the *DACH1* gene were identified. Analysis of RNA-seq data for *DACH1* from the hypermutated sample pair revealed an FPKM (fragments per kilobase of exon per million fragments mapped; Extended Experimental Procedures) expression level of 2.257 in the normal sample, whereas the tumor sample had an expression level of 0.962. This result is corroborated in the WGS data, in which three samples (LUC9, LUC15, and LUC20) show a *DACH1* copy number loss in the tumor sample (Table S11). Lastly, our recurrent screening (n = 96) for mutations in *DACH1* identified two more nonsilent mutations, including one missense mutation (D584G) and one nonsense mutation (G430*).

Recurrent point mutations in the *RELN* gene were identified in three samples (LUC13: A1189D, LUC18: Y3301*; LUC9: H3224N, I1228N; and R301I). Mutations in the *RELN* gene have been identified in pediatric early T cell precursor acute lymphoblastic leukemia (Zhang et al., 2012). We also discovered three samples with nonsynonymous point mutations (LUC11: G347R, LUC12: M521L, and LUC9: P580S and A687S) in the *ABCB5* gene, which encodes a membrane transporter protein belonging to the ATP-binding cassette (ABC) protein family.

There were other genes that did not meet the threshold for significance on the SMG test but were included for testing for recurrence in our extension set (n = 96). Three candidate genes, *HGF*, *CFTR*, and *MICAL3*, were chosen for various reasons, including the possibility of being a therapeutic target for lung cancer (*HGF*) and their association with other lung diseases (*CFTR*, cystic fibrosis), or they were associated with never-smokers (*MICAL3*). The overall prevalence of these mutations in the combined set of 17 samples used for lung cancer whole-genome analysis and 96 samples used for validation was 4.4% (HGF), 4.4% (CFTR), and 0.9% (MICAL3) (Table S12). We identified five point mutations involving the *CFTR* gene in four samples; these included four missense (LUC18: M82V, LUC9: R170L, F354I, and A309S from panel screening) and one nonsense (LUC18: S478*) mutations. Two of the five point mutations involving *CFTR* (M82V and S478*) have been previously reported in patients with cystic fibrosis (Koukourakis et al., 2003). Recently, drugs that target specific *CFTR* point mutations (G551D and other nonsense mutations) have shown therapeutic benefit in patients with cystic fibrosis (Ramsey et al., 2011).

Chromatin-associated genes are found to be mutated in tumor samples from both never-smokers and smokers. We identified 73 nonsynonymous point mutations in 66 chromatin-associated genes, including mutations involving *SETD2*, *ARID1A*, and *ARID2* (Table S13). A nonsense mutation (Q1977*) in *SETD2* was identified in LUC11 from a never-smoker, and two missense mutations (E1735K in *ARID1A* and V465L in *ARID2*) were identified in two smokers (LUC14 and LUC18). Exome sequencing of hepatocellular carcinomas has reported recurrent mutations involving the *ARID2* gene (Li et al., 2011), and frequent mutations in *ARID1A* have been reported in ovarian clear cell carcinoma (Jones et al., 2010a) and endometriosis-associated ovarian cancer (Wiegand et al., 2010). Several point mutations in histone methyltransferase genes (*MLL3*, *MLL4*, *WHSC1L1*, and *ASH1L*) were identified as well.

Tumor Heterogeneity Analysis Using Deep Digital Sequencing Data

By performing targeted sequencing with high read coverage (mean depth of 381 reads) to validate variants detected by WGS, we were able to accurately estimate the variant allele frequencies (VAFs) for somatic mutations identified in each tumor sample. Based on the VAF distribution, we were able to estimate the number and size of the clonal populations in each tumor sample. Recent studies have shown the importance of clonal evolution in tumor progression and development of metastasis (Ding et al., 2012; Gerlinger et al., 2012). Using mutations from copy-number-neutral regions, we found that ten tumors had a multiclonal signature, and seven tumors were largely monoclonal (Table 1 and Figure 2). We did not find any correlation between smoking status and tumor clonality. Based on the VAFs of mutations, we were able to identify mutations that were present in the founding clone and/or the subclone(s). All *EFGR* and *KRAS* mutations validated in our cohort were present in the founder clones of the associated tumor samples (for example, the *EGFR* mutation in LUC15 at 19% VAF,

Table 1. Clonality and Purity Summary for 17 Cases

Case	Gender	Smoking Status	Dominant / Secondary Clone VAFs	Tumor Purity (Based on Dominant Clone VAFs)	Tumor Purity (Based on Chr. X VAFs)	Clonality Status
LUC1	male	light smoker	12.7%	25.4%	30.7%	monoclonal
LUC2	female	smoker	22.5% /12.9%	45.0%	n/a	biclonal
LUC4	male	smoker	14.9%	29.8%	29.8%	monoclonal
LUC6	female	never-smoker	24.7%	49.4%	n/a	monoclonal
LUC7	female	never-smoker	21.3% /10.8%	42.6%	n/a	biclonal
LUC8	female	smoker	22.7%	45.4%	n/a	monoclonal
LUC9	female	smoker	41.1% /20.4%	82.2%	n/a	biclonal
LUC10	male	smoker	43.1% /21.9%	86.2%	71.4%	biclonal
LUC11	male	never-smoker	28.8%	57.6%	41.0%	monoclonal
LUC12	male	smoker	10.4%	20.8%	24.3%	monoclonal
LUC13	male	smoker	41.9% /21.3%	83.8%	58.6%	biclonal
LUC14	female	smoker	29.5% /16.2%	59.0%	n/a	biclonal
LUC15	female	never-smoker	19.2% /10.8%	38.4%	n/a	biclonal
LUC16	female	never-smoker	47.2% /15.3%	94.4%	n/a	biclonal
LUC17	female	smoker	13.9% /11.2%	27.8%	n/a	biclonal
LUC18	male	smoker	18.8% /9.8%	37.6%	31.5%	biclonal
LUC20	female	smoker	39.9%	79.8%	n/a	monoclonal

Figure 2D, and the *KRAS* mutation in LUC10 at 48% VAF, Figure 2F). The clonal distributions of other mutations involving genes such as *HGF* were varied between samples. In the LUC9 tumor in particular, which exhibits two distinct mutation clusters at median VAFs 41.1% and 20.4%, an *HGF* mutation exists in both subclones (Figure 2E). We extended the subclonality analyses to copy number alterations (in particular, deletions) for LUC9 by using an algorithm that compares the observed read counts with the expected diploid read counts in the affected intervals. We found a biclonal pattern in the deletions that was similar to what we observed with SNV analysis described above (Table S14 and Figure S1). In LUC10, an *HGF* mutation exists in the secondary clone at 17% VAF (Figure 2F). It is likely that *EGFR* and *KRAS* mutations are initiating events for lung cancer, and other mutations such as *HGF* mutations are acquired later and perhaps are important for tumor maintenance and progression.

Of the ten tumor samples that had biclonal population, eight had one or more potentially targetable mutations in the subclone (Table S15). These eight samples also had at least one additional targetable mutation in the dominant clone. We believe that our analysis represents an underestimation of tumor heterogeneity due to the limitation of detecting low-frequency mutations in the tumor genome with an average of 30× haploid coverage. It is conceivable that future studies will need to focus on drug therapies affecting critical genes or key pathways not only in the dominant clone but also in the subclone. A treatment strategy focused mainly on the dominant clones could potentially fail, owing to emergence of subclones that are not originally targeted for therapy.

Structural Variants Identified by Whole-Genome and Transcriptome Sequencing

Among the validated 173 somatic rearrangements detected by WGS data were 59 interchromosomal translocations, 7 tandem duplications, 74 deletions, and 33 inversions (Table S16). The majority of the interchromosomal events were clustered in four samples: three from smokers and one from a never-smoker (Data S2). The never-smoker (LUC7) tumor genome is characterized by widespread chromosomal disruption that is consistent with chromothripsis (Stephens et al., 2011). We identified 15 validated interchromosomal translocation events between chromosome 5 and other chromosomes across the LUC7 tumor genome, with most events connecting the distal end of chromosome 5q with various locations on chromosomes 10, 12, 17, and 20. Copy number alterations often co-occur with translocation breakpoints, which is consistent with previously described chromothripsis events. We did not identify any *TP53* mutations in this tumor, though mutations involving the *TP53* gene have been reported to be associated with chromothripsis (Rausch et al., 2012).

We also analyzed the tumor genomes for novel fusion genes, an area of great interest therapeutically with the recent discovery of novel fusions involving kinase genes *ALK*, *ROS*, and *RET* in NSCLC (Takeuchi et al., 2012). With combined whole-genome and transcriptome sequencing, we were able to systematically identify and validate fusion genes. Three different algorithms, ChimeraScan (Iyer et al., 2011), defuse (McPherson et al., 2011), and BreakFusion (Chen et al., 2012), were used to identify fusion genes from the transcriptome sequencing data. High-confidence fusions were then orthogonally validated by analysis of the whole-genome DNA sequencing data. Based on this analysis, we identified 14 high-confidence fusions (Table S17 and Extended Experimental Procedures), including an in-frame novel fusion *KDELR2-ROS1* in LUC11 and an *EML4-ALK* fusion in LUC16. Even though ROS1 kinase fusions have been previously reported in patients diagnosed with NSCLC and cholangiocarcinoma (Bergethon et al., 2012; Gu et al., 2011; Rikova et al., 2007), we identified a novel 5′ partner (*KDELR2*)

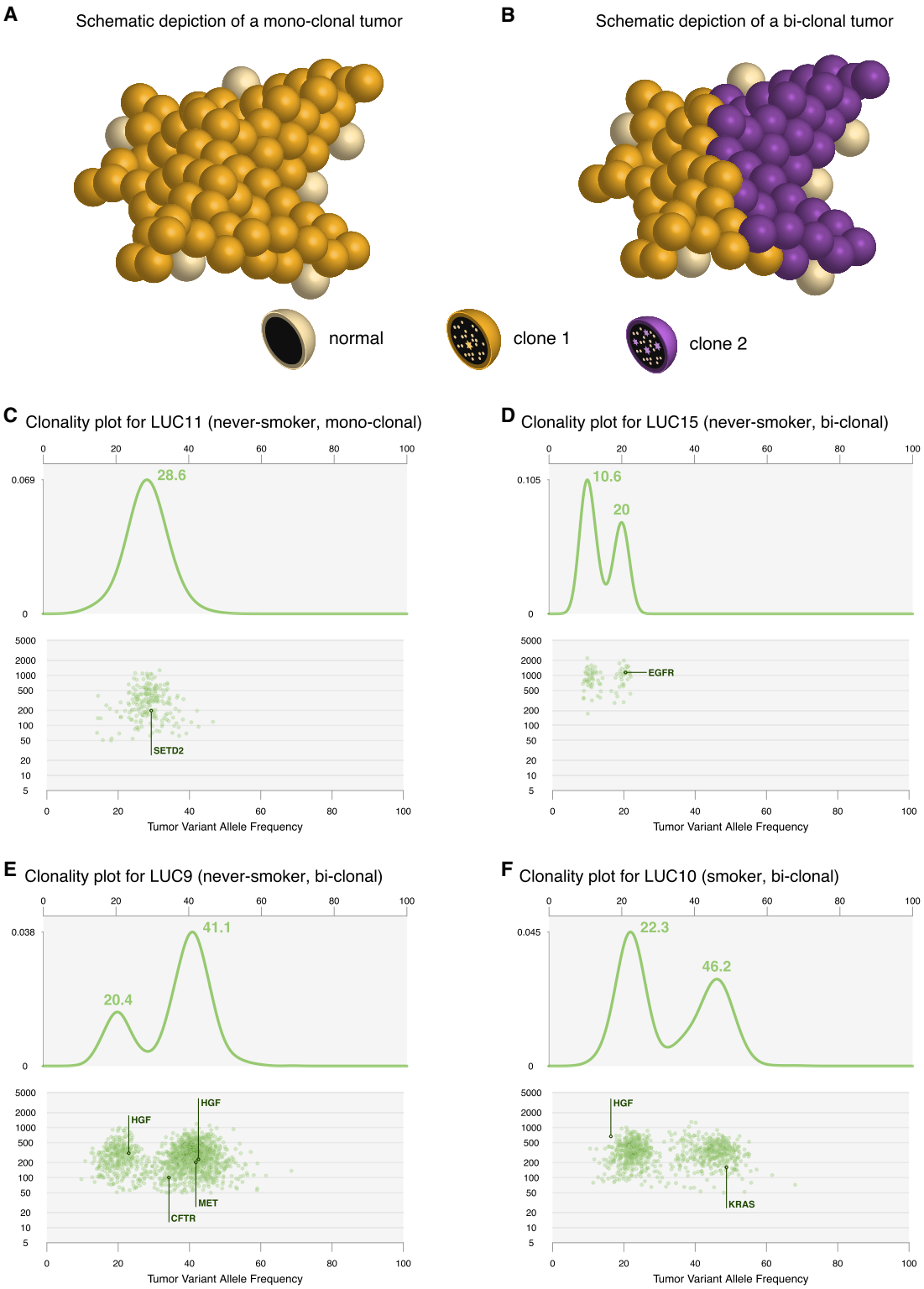

Figure 2. Tumor Clonality Analysis in Lung Cancer

(A) Schematic depiction of a monoclonal tumor sample with a higher tumor purity (i.e., few normal cells).

(B) Schematic depiction of a biclonal tumor sample consisting of a small number of contaminating normal cells, a primary or "founder" clone (pink tumor cells), and a secondary clone (purple tumor cells). The cells of the secondary clone contain the majority of mutations present in the founder clone but have acquired a distinct set of new mutations not shared with the founder.

in our never-smoker sample. A variety of genes have been reported to be 5′ partners in ROS1 fusions, and it is not known whether the 5′ partner plays a role in the oncogenic activity of the fusion kinase (Rikova et al., 2007; Takeuchi et al., 2012). Apart from fusion kinases, an in-frame fusion was detected between the *RASSF1A* (RAS association domain family protein 1) and *TTYH2* (Tweety, *Drosophila* homolog of 2) genes. Another novel fusion consisted of a transcription factor in the 3′ end: *FZR1-NFIC*. *NFIC* (nuclear factor I/C) is a dimeric DNA-binding protein and functions as a cellular transcription factor. *FZR1*, in association with the *APC* gene, is involved in the regulation of mitosis and meiosis.

Integrated Analyses of the Whole-Genome and Transcriptome Data

One of the major strengths of our study is the integration of whole-genome and transcriptome sequencing. Starting with 3,726 tier 1 variants (point mutations only) from all samples identified by WGS, we characterized the expression of each gene by digital (NGS-based) RNA-seq (Extended Experimental Procedures). The median read coverage from RNA-seq for all tier 1 variant positions was 24×, but in expressed genes, the median read coverage reached 129×. We observed significant concordance in variant identification between genome and transcriptome sequencing. Transcriptome sequencing confirmed the presence of 40% of the variants identified by WGS (at least one RNA-seq read) despite the observation that 34% of variants identified in WGS data were from a nonexpressed allele, and 3% of variants from highly expressed genes were not sufficiently covered at the variant positions. We utilized the RNA-seq data to further classify variants into four categories according to their expression patterns: expressed, mutant biased, wild-type biased, and silent gene (Figures 3A and 3B, Table S18, and Extended Experimental Procedures). The genomes of lung cancer from never-smokers had a higher proportion of expressed variants (49.4%) than tobacco smokers (29.1% or 27.0% if the hypermutated LUC9 is excluded). The number of expressed variants that are biased toward the mutant allele is a small proportion of all variants (9.6%) (Figure 3B). For these variants, the mutant allele had a significantly higher variant allele frequency (>20% higher) in the RNA compared to the DNA. Notably, a few genes (*KRAS*, *TP53*, *GTF3C1*, *PLEKHA6*, and *SGOL2*) showed mutant-biased overexpression relative to the wild-type allele in more than one sample. For example, *KRAS* mutations were detected in five of the 17 samples, and in all of these, the mutant allele was preferentially expressed (Table S19). We did not identify copy number amplification in the mutant-biased expression of the *KRAS* gene. *KRAS* and *TP53* were highly expressed (above the 75th percentile) in all 17 cases, and eight of nine *KRAS*/*TP53* mutations occurred in smokers. *KRAS* and *TP53* mutations were mutually exclusive in our 17 cases (Figure 3C), although previous studies showed that they could be present in the same samples (Ding et al., 2008). Although the VAFs observed in WGS and RNA-seq are generally correlated (Figure 3D), rare cases such as *KRAS* and *TP53* deviate from the expected VAF considerably. The mechanism underlying the observed difference in VAFs at the DNA and RNA level for these genes remains unknown.

Interestingly, we observed a lower mutation frequency in tier 1 than in tiers 2 and 3 for all 17 cases, and the average ratios of tier 1 versus tier 2 and 3 frequencies are 0.628 and 0.700 for never-smokers and smokers, respectively (Figures 4A and 4B) (the former light smoker, LUC1, was not included in the calculation and neither was the tier 3 mutation rate for the hypermutated sample, LUC9). This observation is statistically significant ($p = 1.526 \times 10^{-5}$), suggesting that selection pressure and transcription-coupled repair for coding mutations might be responsible for the reduced mutation rate in tier 1. Our result is consistent with the genome-wide analyses of mutation frequencies in a melanoma cell line (Pleasance et al., 2010a) and a lung cancer cell line from a smoker (Pleasance et al., 2010b). Further, we investigated the relationship between mutation frequency and gene expression level and found that highly expressed genes (FPKM >15) have less than four mutations per Mbp, whereas genes that are not expressed (FPKM = 0) have close to 14 mutations per Mbp. Thus, our analysis revealed a negative correlation between gene expression level and mutation frequency in lung cancers (correlation coefficient = −0.49, p = 0.1804), which is consistent with transcription-coupled repair mechanism (Figure 4C).

Somatically Altered Pathways

PathScan (Wendl et al., 2011) analysis was performed to identify significant clusters of point mutations involving genes in annotated KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways. The analysis identified 50 pathways with statistically significant ($p < 0.05$) enrichment of mutations (Table S20 and Extended Experimental Procedures). We subsequently incorporated information regarding indels, copy number variations, and messenger RNA (mRNA) expression level changes involving individual genes in the significant pathways. Based on this analysis, we identified several pathways that are affected in lung cancer, including JAK/STAT pathway (Figure 5A).

Genes involved in extracellular matrix (ECM) interaction, focal or cell adhesion, and cell-cycle pathways were significantly enriched in lung cancer. ECM interaction and cell adhesion genes play important roles in morphogenesis, maintenance of cellular and tissue structure, cell migration, and proliferation. Similarly, there was significant enrichment of genes involved in cell cycle, including the *PRKDC* gene that was recurrently mutated in three patient samples. In addition, there were

(C) Tumor clonality plot of a monoclonal tumor from a never-smoker (LUC11).
(D) Tumor clonality plot of a biclonal tumor from a never-smoker (LUC15) with an *EGFR* mutation in the founder clone.
(E) Tumor clonality plot from a tobacco smoker (LUC9) with two distinct clones. The founder clone has a mean tumor variant allele frequency of 41.1%, and the sub clone has a mean tumor variant allele frequency of 20.4%.
(F) Tumor clonality plot of a biclonal tumor from a tobacco smoker (LUC10) with a *KRAS* mutation in the founder clone.
See also Figure S1, Data S3, and Tables S11, S14, and S15.

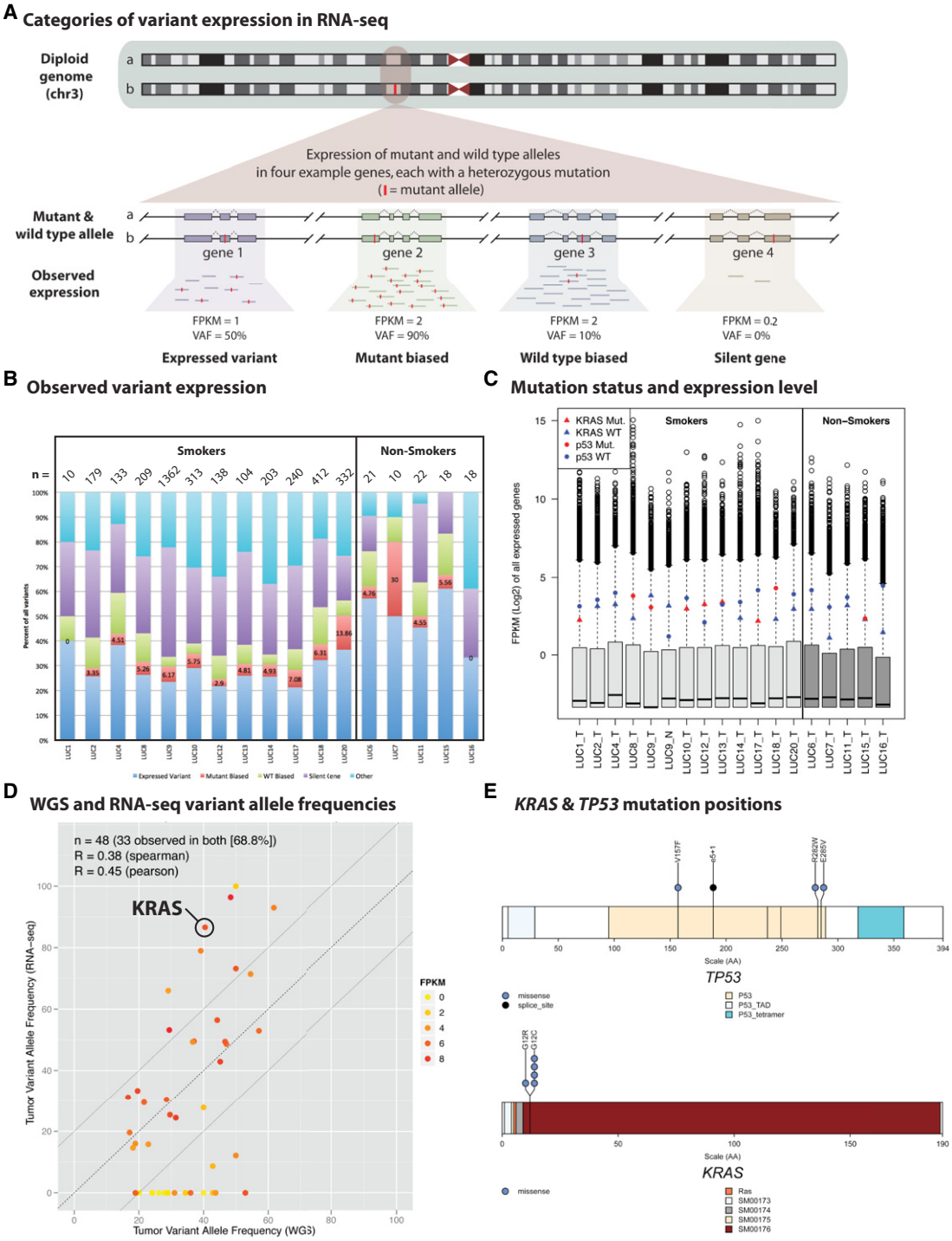

Figure 3. Mutant Biased Expression of *KRAS* and *TP53* Somatic Variants

(A) A line diagram depicting variant expression categories for heterozygous mutations from a diploid genome. A maternal (a) and paternal (b) allele of chromosome 3 is depicted with four example genes enlarged. Each gene contains a heterozygous mutation on the b allele, depicted as a red line. Each gene example illustrates a distinct variant expression pattern by displaying differing numbers of transcripts from each allele being generated from each locus. The FPKM is represented as differing numbers of transcripts generated from each locus, and the VAF is calculated as the proportion of these transcripts deriving from the mutant allele and containing the variant base.

(B) The proportion of variants corresponding to each of the four variant expression categories is summarized for all 17 lung cancers (“other” refers to cases in which the classification was ambiguous due to marginal sequencing coverage). The total number of variants (n) is provided for each patient, and the cases are grouped by smoker status.

mutations involving the cyclins *CCNA1* and *CCNB3* that are essential for activation of cyclin-dependent kinases (CDKs) and progression of cell cycle. We also identified significant enrichment of mutations in the JAK-STAT ($p = 0.04$) pathway in our sample set. Janus kinases are a family of tyrosine kinases that are involved in cytokine receptor signaling, and JAK2 in particular mediates signaling for class II cytokine receptors, cytokine receptors that utilize the γ_c receptor subunit, and receptors that utilize the gp130 subunit (Rodig et al., 1998). The gain-of-function V617F mutation in the pseudokinase domain of the *JAK2* gene leads to constitutive activation of the kinase domain and is associated with uncontrolled hematopoietic cell proliferation in various myeloproliferative neoplasms (Kralovics et al., 2005). However, it is not known whether mutations involving *JAK2* play a significant role in solid epithelial tumors, particularly in lung cancer. Recently, *JAK2* (V617F) mutations were reported in a small proportion (1%) of patients with lung cancer (Lipson et al., 2012). In our sample set, we identified two patients with missense mutations (M532V and V615L) in the protein kinase 1 domain of the *JAK2* gene. The detection of recurrent mutations in the protein kinase domain of the *JAK2* gene, as well as mutations in other genes involved in the JAK-STAT pathway (*JAK3* and *STAT1*), indicate that activation of this pathway may be oncogenic in a subset of patients with lung cancer (Figures 5A and 5B). These findings assume further importance with the development of drugs that effectively target activating mutations in the *JAK2* gene. In addition, we identified that several other pathways were significantly affected in lung cancer, including G-protein-coupled receptor, ion channels, chemokine signaling, calcium signaling pathways, immune modulation, and *ErbB* signaling.

Therapeutic Targets

The use of whole cancer genome sequencing and/or transcriptome sequencing to identify therapeutic targets has been recently reported (Jones et al., 2010b). Apart from previously characterized activating mutations in the tyrosine kinase domain of the *EGFR* gene, we identified potential therapeutic targets, including point mutations in the *HGF*, *MET*, *JAK2*, and *EPHA3* genes and fusions including *KDELR2-ROS1* and *EML4-ALK*. In an effort to comprehensively identify therapeutic targets in lung cancer, we matched gene alterations, including point mutations, copy number amplifications, and high gene expression levels with novel compounds that are currently being evaluated for the treatment of lung cancer (Somaiah and Simon, 2011) (Figure 6). As a result, we identified 54 genes with potentially druggable alterations in our 17 lung cancer patients with several novel therapeutic targets, including tyrosine kinases (*JAK*, *BRAF*, *PIK3CG*, *IGF1R*, *MET*, *RET*, and *FGFR1*), heat shock protein (*HSP90AA1*), and histone deacetylases (*HDAC1*, *HDAC2*, *HDAC6*, and *HDAC9*). A median of 11 (range 7–17) potentially druggable targets was found for each patient. Novel and recurrent druggable point mutations included *PRKCB2*, *MET*, *JAK2*, *HGF*, and *ERBB4*, in addition to previously well-known targets such as *KRAS*, *EGFR*, and *BRAF*. These findings clearly illustrate that there are several novel potential therapeutic targets in patients with lung cancer that require further exploration.

DISCUSSION

Lung cancer is a molecularly heterogeneous disease. The tumor genomic landscape is markedly distinct in never-smokers compared to smokers in several respects: 1) significantly higher mutation frequencies observed in smokers; 2) different mutation spectrum between smokers (C:G→A:T predominant) and never-smokers (C:G→T:A predominant); and 3) distinctive sets of mutations identified in never-smokers (*EGFR* mutations and *ROS1* and *ALK* fusions) and smokers (*KRAS*, *TP53*, *BRAF*, *JAK2*, and *JAK3* and mismatch repair gene mutations). Apart from point mutations, we identified a significant number of structural variations and fusion genes. Going forward, comprehensive genomic analyses of whole genomes and transcriptomes of a large number of lung cancer samples from lifelong never-smokers will be needed to better understand molecular genetics and to guide therapy in this unique subset of patients.

Aberrations in DNA repair pathway, chromatin modification genes, and novel fusions involving metabolic pathways identified in our study present novel therapeutic opportunities. It is possible that these previously poorly characterized molecular lesions in lung cancer may represent the proverbial Achilles' heel for targeted treatment. For example, certain DNA repair pathway lesions may confer unusual susceptibility of cancer cells to PARP inhibitors, much like those seen in BRCA-deficient cancer types. The role of epigenetic therapy in general—and histone deacetylase (HDAC) inhibitors in particular—should be studied in lung cancer, given the number of events in chromatin modifier genes we identified in this study.

Deep digital sequencing provides a large number of events that can be used to precisely estimate clonal size and mutational evolution over time during the natural course of disease progression and in response to selection pressure exerted by therapy. It is unlikely that current therapies would produce lasting remission or cure in advanced lung cancer unless dominant genetic alterations in the founder clone and emerging

(C) Box plots are used to display the expression of FPKM expression values for all detected genes in all 17 cases. The expression level of *KRAS* and *TP53* are displayed by colored triangles and circles, respectively, and patients with a mutation in these genes are indicated in red.
(D) The correlation between VAF calculated from WGS and RNA-seq read counts is depicted as a scatterplot for a single patient. The FPKM expression level of the gene harboring each variant is indicated by a yellow-to-red color scale where yellow indicates low gene expression and red indicates high gene expression. *KRAS* is highlighted as an example of a variant with a VAF that is higher in the RNA than in the WGS data for this patient.
(E) The amino acid position of each *KRAS* and *TP53* mutation is depicted relative to the open reading frame of the gene, along with the position of known protein domains.
See also Tables S7, S17, S18, and S19.

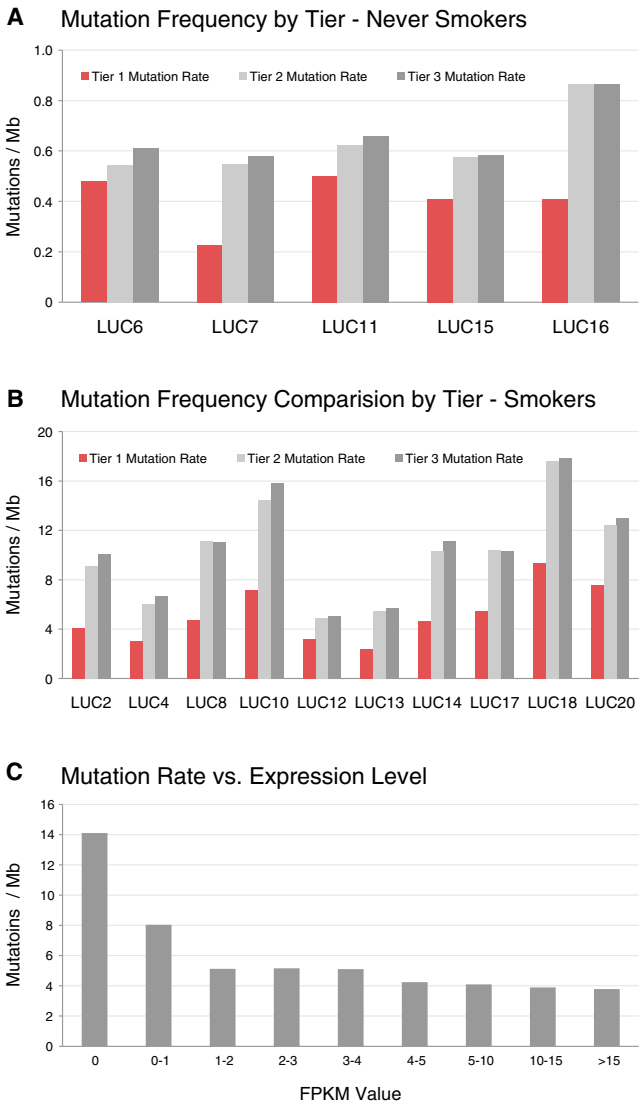

Figure 4. Analysis of Transcription-Coupled Repair across the Genome

(A and B) (A) The mutation rate is assessed independently in each of tiers 1–3 for never-smokers and (B) smokers. Smokers LUC1 and LUC9 are omitted from (B) due to their extremely low and extremely high mutation rates, respectively. Both (A) and (B) clearly show that the coding space in these tumor genomes incurs fewer mutations than other regions in the genomes.

(C) Genes were binned based on the FPKM values derived from RNA expression analysis of the tumor samples, and then the mutation rate (validated somatic mutations per adequately covered Mbp) was calculated for each expression level bin. The graph shows that the lowest mutation rates occur in the most highly expressed genes.

See also Data S2 and Tables S7, S8, and S9.

secondary clones are targeted specifically for therapy. A systematic approach to collect tissue samples—not only at the time of diagnosis but serially at the times of relapse to chronicle the dynamic clonal evolution that occurs over time and possibly at different metastatic sites—is absolutely critical to make major advances in therapy.

Only through a comprehensive assessment of WGS and transcriptomes in large numbers of carefully curated and well-annotated samples will we able to catalog potentially significant point mutations and structural variations that led to critical perturbations in the cellular homeostasis. Moreover, the cancer research community should radically overhaul the current approach to drug development and initiate a series of steps to study comprehensively genomic evolution over time in well-defined cohorts of patients enrolled in clinical trials. Comprehensive genomic characterization efforts to catalog somatically altered pathways will improve our understanding of the molecular genetics of lung cancer and will identify novel therapeutic targets. Functional studies in the laboratory and thoughtfully designed clinical studies will be needed to fully harness the data from genomic studies such as ours.

EXPERIMENTAL PROCEDURES

Point mutations and indels identified by WGS of 17 tumor-normal pairs were classified into four tiers as described previously (Mardis et al., 2009). Custom sequence capture arrays from Roche Nimblegen were used to validate all putative WGS mutations. Variants of interest identified by WGS were extended by recurrence screening in an independent set of 96 primary lung adenocarcinomas. RNA-seq data were generated for all 17 tumors and for a single, matched normal adjacent tissue with 11,578–14,507 genes detected as expressed in each tumor.

RNA-seq analysis involved alignment using TopHat (Trapnell et al., 2012) and assembly and expression estimation by Cufflinks (Trapnell et al., 2012) using a known set of reference transcripts from Ensembl v.58. The expression status of single-nucleotide variants was assessed by examination of TopHat alignment and Cufflinks transcript expression estimates, allowing the classification of each putative variant from WGS into one of five expression patterns (Extended Experimental Procedures).

Expressed gene fusions were identified by combining the results of ChimeraScan (Iyer et al., 2011), defuse (McPherson et al., 2011), and BreakFusion (Chen et al., 2012). Downstream analyses included the cross-validation of fusion events detected in RNA data with WGS SV predictions, tumor clonality estimates, and several analyses that are part of the MuSiC analysis suite (Dees et al., 2012). Tumor clonality estimation includes the identification of peaks in the kernel density estimates of deep-read count variant allele frequencies at somatic SNV sites in copy-number-neutral genomic regions from the tumor genomes. MuSiC analyses included the identification of significantly mutated genes under the statistical consideration of seven separate mutational mechanism categories, a proximity analysis used to identify recurrently mutated functional domains, and a comparison of the SNVs discovered in this data set with those in the COSMIC database. Global pathway analysis was performed by using PathScan (Wendl et al., 2011), followed by a more focused analysis of the KEGG cancer pathways (Kanehisa and Goto, 2000; Kanehisa et al., 2012) and the JAK-STAT pathway in particular.

To identify putative druggable targets, a candidate gene list was generated by identifying genes with nonsilent mutations,

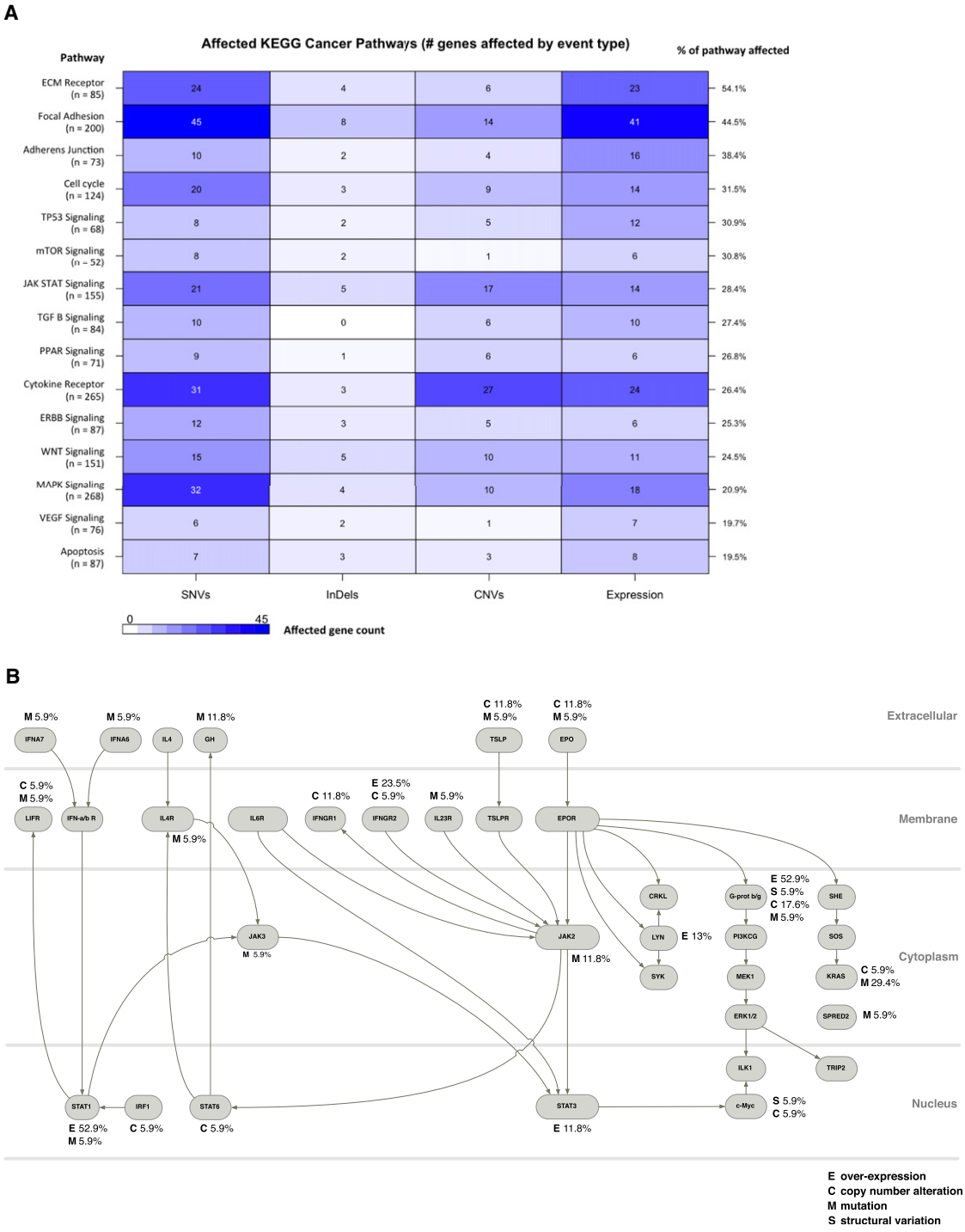

Figure 5. Alterations in JAK/STAT Pathway and Integration of Somatic Alterations and High RNA Expression in Significant KEGG Pathways

(A) Heatmap of significantly overrepresented gene pathways in lung cancer ($p < 0.05$). The number of gene members of each KEGG cancer pathway ("KEGG pathways in cancer" or "hsa05200") altered by one of four alteration types in at least one patient is summarized as a heatmap. The KEGG pathway name is listed on the y axis at the left, and the total number of genes comprising that pathway is provided (labeled as n). The number within each box represents the number of genes altered in at least one patient for each alteration type. The percentage of all gene members of the KEGG pathway altered in at least one patient by at least one alteration type is provided on the right side. The heatmap is sorted by this percentage.

(B) Molecular alterations in JAK-STAT pathway in patients with non-small cell lung cancer. Genes that were found to be altered in the 17 lung cancer samples are labeled with the type of molecular change (E, overexpression; C, copy number alteration; M, mutation; and S, structural variation) and the frequency.

See also Data S3 and Tables S3, S9, S11, S18, and S20.

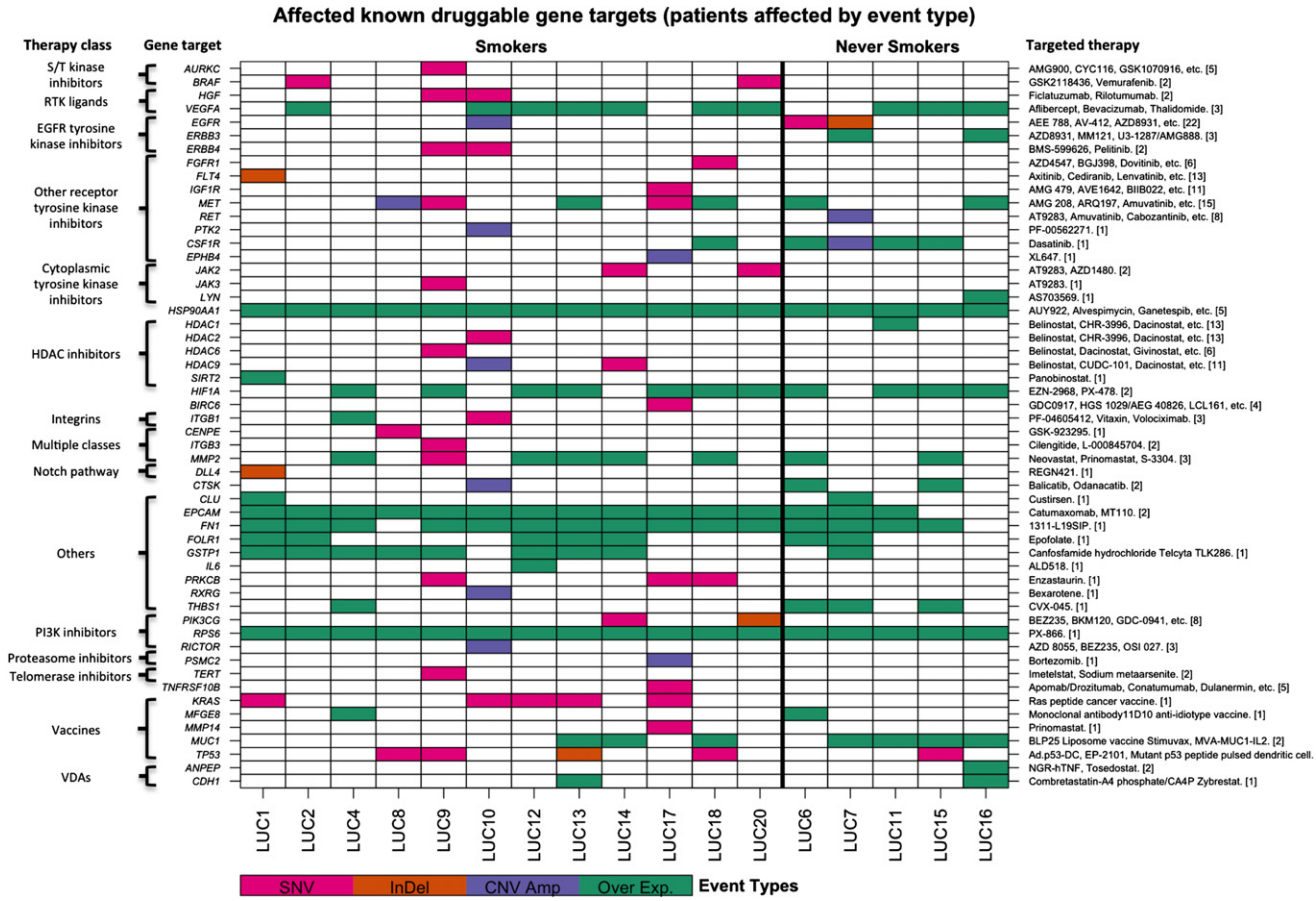

Figure 6. Potential Therapeutic Targets in Non-Small Cell Lung Cancer

Graphical representation of the various therapeutic targets in each patient sample. Patients are listed on the x axis. Target genes identified as altered in one or more patients and the drugs that targeted these genes are listed on the y axis (gene symbols on the left side and corresponding drug names on the right side). Where display of all drug names was not practical, the list was abbreviated. The numbers in parentheses indicate the total number of drugs currently available for each gene target. A box representing each gene-drug combination for each patient is colored according to the class of gene alteration: red for SNVs, orange for indels, purple for CNV amplifications, and green for RNA overexpression (Extended Experimental Procedures). Gene targets are grouped and labeled on the left side of the plot according to the therapeutic class of their targeted agents. See also Tables S3, S4, S9, S11, S15, S16, and S17.

copy number amplifications, and/or high RNA expression. The resulting gene list was intersected with a list of "known" drug-gene interactions currently used or under investigation in lung cancer. The same candidate gene list was also annotated against lists of genes that are thought to be potential targets for novel drug development according to a previously described approach (Hopkins and Groom, 2002; Russ and Lampel, 2005).

ACCESSION NUMBERS

The GEO accession numbers are pending for the RNA-seq data sets reported in this paper. Until they are available, readers can access the data from http://genome.wustl.edu/projects/detail/lung-cancer/.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, three data files, one figure, and 20 tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.08.024.

ACKNOWLEDGMENTS

We thank the following groups at The Genome Institute for their dedicated efforts in this work: the Production group for sequence data production and processing, the Technology Development group for formulation of methods and troubleshooting, the Analysis Pipeline group for developing the automated sequence analysis pipelines, the LIMS group for developing tools and software to manage samples and sequencing, and the Systems group for providing the IT infrastructure and HPC solutions required for sequencing and analysis. We thank Daniel C. Koboldt for his help on tumor clonality analysis, Mike Wendl for help with pathway analysis, Josh Peck for help with manuscript preparation, and Joshua McMichael for help with figure preparation. We also thank the Washington University Cancer Genome Initiative and Siteman Cancer Center for their support. This work was funded by grants to R.K.W. from Washington University in St. Louis and the National Human Genome Research Institute (NHGRI U54 HG003079).

Received: May 17, 2012
Revised: July 17, 2012
Accepted: August 23, 2012
Published: September 13, 2012

REFERENCES

Bergethon, K., Shaw, A.T., Ignatius Ou, S.H., Katayama, R., Lovly, C.M., McDonald, N.T., Massion, P.P., Siwak-Tapp, C., Gonzalez, A., Fang, R., et al. (2012). ROS1 rearrangements define a unique molecular class of lung cancers. J. Clin. Oncol. *30*, 863–870.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., et al. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nat. Methods *6*, 677–681.

Chen, K., Wallis, J.W., Kandoth, C., Kalicki-Veizer, J.M., Mungall, K.L., Mungall, A.J., Jones, S.J., Marra, M.A., Ley, T.J., Mardis, E.R., et al. (2012). BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. Bioinformatics *28*, 1923–1924.

Dees, N.D., Zhang, Q., Kandoth, C., Wendl, M.C., Schierding, W., Koboldt, D.C., Mooney, T.B., Callaway, M.B., Dooling, D., Mardis, E.R., et al. (2012). MuSiC: identifying mutational significance in cancer genomes. Genome Res. *22*, 1589–1598.

Ding, L., Getz, G., Wheeler, D.A., Mardis, E.R., McLellan, M.D., Cibulskis, K., Sougnez, C., Greulich, H., Muzny, D.M., Morgan, M.B., et al. (2008). Somatic mutations affect key pathways in lung adenocarcinoma. Nature *455*, 1069–1075.

Ding, L., Ley, T.J., Larson, D.E., Miller, C.A., Koboldt, D.C., Welch, J.S., Ritchey, J.K., Young, M.A., Lamprecht, T., McLellan, M.D., et al. (2012). Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature *481*, 506–510.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C., and Parkin, D.M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int. J. Cancer *127*, 2893–2917.

Gerlinger, M., Rowan, A.J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. N. Engl. J. Med. *366*, 883–892.

Gu, T.L., Deng, X., Huang, F., Tucker, M., Crosby, K., Rimkunas, V., Wang, Y., Deng, G., Zhu, L., Tan, Z., et al. (2011). Survey of tyrosine kinase signaling reveals ROS kinase fusions in human cholangiocarcinoma. PLoS ONE *6*, e15640.

Hopkins, A.L., and Groom, C.R. (2002). The druggable genome. Nat. Rev. Drug Discov. *1*, 727–730.

Iyer, M.K., Chinnaiyan, A.M., and Maher, C.A. (2011). ChimeraScan: a tool for identifying chimeric transcription in sequencing data. Bioinformatics *27*, 2903–2904.

Jones, S., Wang, T.L., Shih, IeM., Mao, T.L., Nakayama, K., Roden, R., Glas, R., Slamon, D., Diaz, L.A., Jr., Vogelstein, B., et al. (2010a). Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. Science *330*, 228–231.

Jones, S.J., Laskin, J., Li, Y.Y., Griffith, O.L., An, J., Bilenky, M., Butterfield, Y.S., Cezard, T., Chuah, E., Corbett, R., et al. (2010b). Evolution of an adenocarcinoma in response to selection by targeted kinase inhibitors. Genome Biol. *11*, R82.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic Acids Res. *40* (Database issue), D109–D114.

Kendall, J., Liu, Q., Bakleh, A., Krasnitz, A., Nguyen, K.C., Lakshmi, B., Gerald, W.L., Powers, S., and Mu, D. (2007). Oncogenic cooperation and coamplification of developmental transcription factor genes in lung cancer. Proc. Natl. Acad. Sci. USA *104*, 16663–16668.

Kohno, T., Ichikawa, H., Totoki, Y., Yasuda, K., Hiramoto, M., Nammo, T., Sakamoto, H., Tsuta, K., Furuta, K., Shimada, Y., et al. (2012). KIF5B-RET fusions in lung adenocarcinoma. Nat. Med. *18*, 375–377.

Koukourakis, M.I., Giatromanolaki, A., Brekken, R.A., Sivridis, E., Gatter, K.C., Harris, A.L., and Sage, E.H. (2003). Enhanced expression of SPARC/osteonectin in the tumor-associated stroma of non-small cell lung cancer is correlated with markers of hypoxia/acidity and with poor prognosis of patients. Cancer Res. *63*, 5376–5380.

Kralovics, R., Passamonti, F., Buser, A.S., Teo, S.S., Tiedt, R., Passweg, J.R., Tichelli, A., Cazzola, M., and Skoda, R.C. (2005). A gain-of-function mutation of JAK2 in myeloproliferative disorders. N. Engl. J. Med. *352*, 1779–1790.

Larson, D.E., Harris, C.C., Chen, K., Koboldt, D.C., Abbott, T.E., Dooling, D.J., Ley, T.J., Mardis, E.R., Wilson, R.K., and Ding, L. (2012). SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics *28*, 311–317.

Lee, W., Jiang, Z., Liu, J., Haverty, P.M., Guan, Y., Stinson, J., Yue, P., Zhang, Y., Pant, K.P., Bhatt, D., et al. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. Nature *465*, 473–477.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R.; 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Li, M., Zhao, H., Zhang, X., Wood, L.D., Anders, R.A., Choti, M.A., Pawlik, T.M., Daniel, H.D., Kannangai, R., Offerhaus, G.J., et al. (2011). Inactivating mutations of the chromatin remodeling gene ARID2 in hepatocellular carcinoma. Nat. Genet. *43*, 828–829.

Lipson, D., Capelletti, M., Yelensky, R., Otto, G., Parker, A., Jarosz, M., Curran, J.A., Balasubramanian, S., Bloom, T., Brennan, K.W., et al. (2012). Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. Nat. Med. *18*, 382–384.

Lynch, T.J., Bell, D.W., Sordella, R., Gurubhagavatula, S., Okimoto, R.A., Brannigan, B.W., Harris, P.L., Haserlat, S.M., Supko, J.G., Haluska, F.G., et al. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib. N. Engl. J. Med. *350*, 2129–2139.

Mardis, E.R., Ding, L., Dooling, D.J., Larson, D.E., McLellan, M.D., Chen, K., Koboldt, D.C., Fulton, R.S., Delehaunty, K.D., McGrath, S.D., et al. (2009). Recurring mutations found by sequencing an acute myeloid leukemia genome. N. Engl. J. Med. *361*, 1058–1066.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

McPherson, A., Hormozdiari, F., Zayed, A., Giuliany, R., Ha, G., Sun, M.G., Griffith, M., Heravi Moussavi, A., Senz, J., Melnyk, N., et al. (2011). deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput. Biol. *7*, e1001138.

Ng, T.P. (1994). Silica and lung cancer: a continuing controversy. Ann. Acad. Med. Singapore *23*, 752–755.

Pleasance, E.D., Cheetham, R.K., Stephens, P.J., McBride, D.J., Humphray, S.J., Greenman, C.D., Varela, I., Lin, M.L., Ordóñez, G.R., Bignell, G.R., et al. (2010a). A comprehensive catalogue of somatic mutations from a human cancer genome. Nature *463*, 191–196.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C., et al. (2010b). A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature *463*, 184–190.

Ramsey, B.W., Davies, J., McElvaney, N.G., Tullis, E., Bell, S.C., Dřevínek, P., Griese, M., McKone, E.F., Wainwright, C.E., Konstan, M.W., et al; VX08-770-102 Study Group. (2011). A CFTR potentiator in patients with cystic fibrosis and the G551D mutation. N. Engl. J. Med. *365*, 1663–1672.

Rausch, T., Jones, D.T., Zapatka, M., Stütz, A.M., Zichner, T., Weischenfeldt, J., Jäger, N., Remke, M., Shih, D., Northcott, P.A., et al. (2012). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell *148*, 59–71.

Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., Nardone, J., Lee, K., Reeves, C., Li, Y., et al. (2007). Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell *131*, 1190–1203.

Rodig, S.J., Meraz, M.A., White, J.M., Lampe, P.A., Riley, J.K., Arthur, C.D., King, K.L., Sheehan, K.C., Yin, L., Pennica, D., et al. (1998). Disruption of the Jak1 gene demonstrates obligatory and nonredundant roles of the Jaks in cytokine-induced biologic responses. Cell *93*, 373–383.

Russ, A.P., and Lampel, S. (2005). The druggable genome: an update. Drug Discov. Today *10*, 1607–1610.

Sellers, T.A., Bailey-Wilson, J.E., Elston, R.C., Wilson, A.F., Elston, G.Z., Ooi, W.L., and Rothschild, H. (1990). Evidence for mendelian inheritance in the pathogenesis of lung cancer. J. Natl. Cancer Inst. *82*, 1272–1279.

Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., et al. (2007). Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature *448*, 561–566.

Somaiah, N., and Simon, G.R. (2011). Molecular targeted agents and biologic therapies for lung cancer. J. Thorac. Oncol. *6* (*Suppl 4*), S1758–S1785.

Stephens, P.J., Greenman, C.D., Fu, B., Yang, F., Bignell, G.R., Mudie, L.J., Pleasance, E.D., Lau, K.W., Beare, D., Stebbings, L.A., et al. (2011). Massive genomic rearrangement acquired in a single catastrophic event during cancer development. Cell *144*, 27–40.

Subramanian, J., and Govindan, R. (2007). Lung cancer in never smokers: a review. J. Clin. Oncol. *25*, 561–570.

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

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. *7*, 562–578.

Watanabe, A., Ogiwara, H., Ehata, S., Mukasa, A., Ishikawa, S., Maeda, D., Ueki, K., Ino, Y., Todo, T., Yamada, Y., et al. (2011). Homozygously deleted gene DACH1 regulates tumor-initiating activity of glioma cells. Proc. Natl. Acad. Sci. USA *108*, 12384–12389.

Weir, B.A., Woo, M.S., Getz, G., Perner, S., Ding, L., Beroukhim, R., Lin, W.M., Province, M.A., Kraja, A., Johnson, L.A., et al. (2007). Characterizing the cancer genome in lung adenocarcinoma. Nature *450*, 893–898.

Wendl, M.C., Wallis, J.W., Lin, L., Kandoth, C., Mardis, E.R., Wilson, R.K., and Ding, L. (2011). PathScan: a tool for discerning mutational significance in groups of putative cancer genes. Bioinformatics *27*, 1595–1602.

Wiegand, K.C., Shah, S.P., Al-Agha, O.M., Zhao, Y., Tse, K., Zeng, T., Senz, J., McConechy, M.K., Anglesio, M.S., Kalloger, S.E., et al. (2010). ARID1A mutations in endometriosis-associated ovarian carcinomas. N. Engl. J. Med. *363*, 1532–1543.

Wood, R.D., Mitchell, M., Sgouros, J., and Lindahl, T. (2001). Human DNA repair genes. Science *291*, 1284–1289.

Wu, K., Li, A., Rao, M., Liu, M., Dailey, V., Yang, Y., Di Vizio, D., Wang, C., Lisanti, M.P., Sauter, G., et al. (2006). DACH1 is a cell fate determination factor that inhibits cyclin D1 and breast tumor growth. Mol. Cell. Biol. *26*, 7116–7129.

Wu, K., Katiyar, S., Witkiewicz, A., Li, A., McCue, P., Song, L.-N., Tian, L., Jin, M., and Pestell, R.G. (2009). The cell fate determination factor dachshund inhibits androgen receptor signaling and prostate cancer cellular growth. Cancer Res. *69*, 3347–3355.

Yang, P., Schwartz, A.G., McAllister, A.E., Swanson, G.M., and Aston, C.E. (1999). Lung cancer risk in families of nonsmoking probands: heterogeneity by age at diagnosis. Genet. Epidemiol. *17*, 253–273.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R., and Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics *25*, 2865–2871.

Zhang, J., Ding, L., Holmfeldt, L., Wu, G., Heatley, S.L., Payne-Turner, D., Easton, J., Chen, X., Wang, J., Rusch, M., et al. (2012). The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature *481*, 157–163.