The transcriptional landscape and mutational profile of lung adenocarcinoma

Author list

Jeong-Sun Seo[1,2,3,4,5,$,#], Young Seok Ju[4,$], Won-Chul Lee[1,3,$], Jong-Yeon Shin[1,5], June Koo Lee[1,6], Thomas Bleazard[1], Junho Lee[1], Yoo Jin Jung[7], Jung-Oh Kim[9], Jung-Young Shin[9], Saet-Byeol Yu[5], Jihye Kim[5], Eung-Ryoung Lee[4], Chang-Hyun Kang[8], In-Kyu Park[8], Hwanseok Rhee[4], Se-Hoon Lee[1,6,7], Jong-Il Kim[1,2,3,5], Jin-Hyoung Kang[10,#] and Young Tae Kim[1,7,8,#]

Affiliation

1. Genomic Medicine Institute (GMI), Medical Research Center, Seoul National University, Seoul 110-799, Korea

2. Department of Biochemistry, Seoul National University College of Medicine, Seoul 110-799, Korea

3. Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 110-799, Korea

4. Macrogen Inc., Seoul 153-781, Korea

5. Psoma Therapeutics Inc., Seoul 153-781, Korea

6. Department of Internal Medicine, Seoul National University Hospital, Seoul 110-799, Korea

7. Cancer Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea

8. Department of Thoracic and Cardiovascular Surgery, Seoul National University Hospital, Seoul 110-799, Korea

9. Division of Medical Oncology, Research Institute of Medical Science, The Catholic University of Korea, Seoul 137-040, Republic of Korea

10. Division of Medical Oncology, Seoul St. Mary's Hospital, The Catholic University of Korea, Seoul 137-040, Korea

$ These authors contributed equally to this work

 # Corresponding authors:

Jeong-Sun Seo, MD, PhD

Professor & Director
Genomic Medicine Institute and
Department of Biochemistry and Molecular Biology
Seoul National University College of Medicine
Phone: +82-2-740-8246
Fax: +82-2-741-5423
E-mail: jeongsun@snu.ac.kr

Jin-Hyoung Kang, MD, PhD

Professor
Division of Medical Oncology,
Seoul St. Mary's Hospital,
The Catholic University of Korea, Seoul 137-040, Korea
Phone: +82-2-2258-6043
Fax: +82-2-594-6043
E-mail: jinkang@catholic.ac.kr

Young Tae Kim, MD, PhD

Professor
Department of Thoracic and Cardiovascular Surgery,
Seoul National University College of Medicine, Seoul 110-799, Korea
Phone: +82-2-2072-3161
Fax: +82-2-765-7117
E-mail: ytkim@snu.ac.kr

Abstract

All cancers harbor molecular alterations in their genomes. The transcriptional consequences of these somatic mutations have not yet been comprehensively explored in lung cancer. Here we present the first large scale RNA sequencing study of lung adenocarcinoma, demonstrating its power to identify somatic point mutations as well as transcriptional variants such as gene fusions, alternative splicing events and expression outliers. Our results reveal the genetic basis of 200 lung adenocarcinomas in Koreans including deep characterization of 87 surgical specimens by transcriptome sequencing. We identified driver somatic mutations in cancer genes including *EGFR*, *KRAS*, *NRAS*, *BRAF*, *PIK3CA*, *MET* and *CTNNB1*. Candidates for novel driver mutations were also identified in genes newly implicated in lung adenocarcinoma such as *LMTK2*, *ARID1A*, *NOTCH2* and *SMARCA4*. We found 45 fusion genes, 8 of which were chimeric tyrosine kinases involving *ALK*, *RET*, *ROS1*, *FGFR2*, *AXL* and *PDGFRA*. Among 17 recurrent alternative splicing events, we identified exon 14 skipping in the proto-oncogene *MET* as highly likely to be a cancer driver. The number of somatic mutations and expression outliers varied markedly between individual cancers and was strongly correlated with smoking history of patients. We identified genomic blocks within which gene expression levels were consistently increased or decreased that could be explained by copy number alterations in samples. We also found an association between lymph node metastasis and somatic mutations in *TP53*. These findings broaden our understanding of lung adenocarcinoma and may also lead to new diagnostic and therapeutic approaches.

Introduction

Lung cancer is one of the most common cancers in humans, as well as the leading cause of cancer-related death worldwide (Jemal et al. 2011). Although diagnosis at an early stage is increasing with the introduction of low-dose computerized tomography screening, lung cancer is still a devastating disease which has a very poor prognosis (Aberle et al. 2011). Lung cancer can be classified based on histopathologic findings: adenocarcinoma is the most common type (Travis et al. 2005). Recently, deeper understanding of the major genetic alterations and signaling pathways involved has suggested a reclassification of lung adenocarcinoma based on underlying driver mutations. Cancer cells with these genetic alterations have survival and growth advantages over cells without such changes (Haber et al. 2007). Currently, approximately ten driver genes have been discovered in lung adenocarcinoma (Pao et al. 2011). Clinical trials using new chemotherapeutic agents targeting such alterations have demonstrated remarkable improvements in patient outcome, for example gefitinib (Maemondo et al. 2010; Mok et al. 2009) and crizotinib (Kwak et al. 2010) for lung adenocarcinoma harboring *EGFR* mutations and *EML4-ALK* (Soda et al. 2007) fusion, respectively. More recently, not only point mutations but also tyrosine kinase gene fusions, such as *KIF5B-RET*, were also identified as driver mutations (Ju et al. 2012). Nevertheless, we still do not know the molecular drivers of about 40% of lung adenocarcinomas (Pao et al. 2011). Interestingly, the frequencies of some driver mutations have been shown to be significantly different between ethnic groups (Shigematsu et al. 2006), and therefore comprehensive cancer genome studies in a range of human populations will help to find new molecular alterations which can be targeted in treatments of lung cancer.

In this study, we broadly surveyed genetic alterations in 200 fresh surgical specimens of lung adenocarcinoma in Koreans. 87 of these were analyzed by transcriptome sequencing

combined with whole-exome (n=76) and transcriptome sequencing (n=77) for matched adjacent normal tissue samples. Transcriptome sequencing is a powerful method for detecting driver mutations in cancer, since not only somatic point mutations but also aberrant RNA variants such as fusion genes and alternative splicing can be examined. Though advances in genomic technologies have enabled genome-wide analyses of cancers, to our knowledge this is the first large-scale study of lung adenocarcinoma using RNA sequencing.

Results

Cancer samples analyzed in this study

We collected 200 fresh surgical specimens of primary lung adenocarcinoma from patients who underwent major lung resection (Supplementary Figure 1). For each patient, we recorded diagnosis, gender, cancer stage and smoking status (Supplementary Table 1). Among the 200 cancer patients, the proportions of females and never-smokers were 54.5% (n=109) and 58.0% (n=116), respectively. Of these, 87 cancer tissues whose driver mutations were not detected by screening tests (Sanger sequencing for *EGFR* and *KRAS* point mutations and fluorescence *in situ* hybridization (FISH) for *EML4-ALK* fusion; Supplementary Table 1; Supplementary Figure 1), were analyzed by transcriptome sequencing combined with whole-exome (n=76) and transcriptome (n=77) sequencing of matched normal lung tissue samples (Supplementary Table 2; Supplementary Information). All these sequencing experiments were done as described previously (Ju et al. 2011). We generated 14,038,673,860 paired-end 101 bp-long reads from RNA sequencing of 164 samples (87 cancer and 77 corresponding normal tissues). On average, the RNA sequencing throughputs were 9.77 and 7.38 Gbp for cancer and normal tissues respectively. In the whole-exome sequencing of normal tissues, we obtained 32.96-fold read depth per tissue for regions targeted by the exome capture platform used in this study.

Somatic point mutations

Using our transcriptome data, we identified 4,607 somatic non-synonymous single nucleotide substitutions and 373 coding short-indel mutations (Supplementary Figure 2; Supplementary Table 3). Whole-exome sequencing of two randomly selected cancer

samples provided an estimate of 89.2% for the accuracy of our somatic mutation discovery (Supplementary Table 4). The median number of somatic mutations in each cancer sample was 25. Among a total of 87 samples, 45 carried driver mutations in well-known cancer genes in lung adenocarcinoma, such as *EGFR* (n=22; in-frame deletions of exon 19, L858R and G719A) and *KRAS* (n=18; G12C, G12V, G12D, G12S, G13C and G13D) (Figure 1 and Supplementary Table 3). In addition to *EGFR* and *KRAS*, other known mutations were also detected in *NRAS* (Sequist et al. 2011) (n=3; Q61H, Q61L, Q61K), *PIK3CA* (Ding et al. 2008) (n=2; H1047R and E555K) and *BRAF* (Ding et al. 2008) (n=1; V600E), which have all been reported as driver mutations of lung cancer. In addition, 2 samples carried known activating mutations in well-known oncogenes confirmed in other cancers (D32G of *CTNNB1* (Chan et al. 1999), M1124D of *MET* (Schmidt et al. 1999)), suggesting those mutations are also able to induce lung adenocarcinoma. Overall, 47 specimens harbored known point driver mutations in 7 cancer genes (*EGFR*, *KRAS*, *NRAS*, *PIK3CA*, *BRAF*, *CTNNB1* and *MET)*, which we here refer to as “canonical point driver mutations”. These mutations were mutually exclusive with one exception of 1 *EGFR* and *PIK3CA* double-mutant. In addition to these known driver mutations, we also identified a set of genes, which were frequently mutated or highly over-expressed in a subset of cancers. Of note, *TP53* was the most frequently mutated gene. *CDKN2A*, *RET*, *NOTCH2*, *SMARCA4*, *LMTK2*, *ARID1A* and *MTOR* were also frequently altered and are also worthy of note, since the functions of these genes are likely to be related to tumorigenesis or cancer maintenance (Figure 1). The pair-wise mutual exclusion and concurrence analysis for these mutated genes is shown in Supplementary Table 5.

Fusion gene analysis

One of the advantages of transcriptome sequencing over genome sequencing is that detection of transcriptional variants, such as fusion genes and alternative splicing, is feasible.

Using the gene fusion program (GFP) introduced previously (Ju et al. 2012) and typical gene expression patterns, we identified 45 in-frame fusion transcripts from the 87 cancer tissues (Figure 2; Table 1; Supplementary Figure 3; Supplementary Table 6). We attempted to validate all of the fusion genes using PCR amplification of cDNA and Sanger sequencing (Supplementary Figure 4; Supplementary Table 7). Among 43 fusion genes where PCR primer was available, 39 were successfully validated (Supplementary Table 7). Interestingly, the four invalidated fusion genes all included a surfactant gene (i.e. *SFTPB* or *SFTPA2*) or *H19*, wildtype gene expression of which were extremely high (> ~2,000 RPKM) in the corresponding specimens. This indicates that the fusion transcript may have been artificially synthesized during sequencing library construction.

Of the fusion genes we identified, 8 were chimeric tyrosine kinases which are highly likely to play an important role in cancer development. Cancer specimens carrying one of the tyrosine kinase fusions (n=10) did not harbor any of the canonical point driver mutations (*p*-value = 2.12×10^{-4}; Supplementary Table 8). Of these eight fusion genes, four have been reported previously (*EML4-ALK* (Soda et al. 2007), *KIF5B-RET* (Ju et al. 2012; Kohno et al. 2012; Lipson et al. 2012; Takeuchi et al. 2012), *CD74-ROS1* (Rikova et al. 2007; Takeuchi et al. 2012) and *SLC34A2-ROS1* (Rikova et al. 2007; Takeuchi et al. 2012)), and we refer to them here as “canonical transforming fusion genes”. The remaining four fusion genes were novel (*CCDC6-ROS1*, *FGFR2-CIT*, *AXL-MBIP* and *SCAF11-PDGFRA*) (Figure 3A). Of these four novel fusion genes, *CCDC6-ROS1*, *FGFR2-CIT* and *AXL-MBIP* carry protein tyrosine kinase domains and dimerization units (Alberti et al. 2003; Ju et al. 2012) (coiled-coil or leucine zipper domains), both of which are essential to activate chimeric tyrosine kinases. The *SCAF11-PDGFRA* fusion is an example of promoter swapping (Kas et al. 1997). Because the cancer specimens harboring these four novel fusion genes did not carry any known driver mutations (neither canonical point driver mutations (n=47) nor canonical transforming fusion genes (n=6); *p*-value = 0.021; Supplementary Table 8), they may play

important roles in cancer transformation. Other fusion genes identified in this study, such as *MAP4K3-PRKCE*, *BCAS3-MAP3K3*, *ERBB2IP-MAST4* and *APLP2-TNFSF11* may also have functional importance since the genes are serine-threonine kinases or involved in signaling pathways. The co-occurrence of 45 fusion genes with canonical point driver mutations is shown in Supplementary Table 6.

Alternative splicing

Alternative splicing is known to be related to the pathogenesis of colon cancer (Gardina et al. 2006). We assessed exon skipping events preferentially occurring in the cancer tissue using the transcriptome sequencing data. From a total of 87 tissues, we identified 17 recurrent exon skipping events, where a specific exon of a gene is not in full transcription in cancer (Supplementary Table 9). In particular, three cases of skipping of exon 14 in *MET* were interesting (Figure 3B), since *MET* protein tyrosine kinase is a well-known oncogene and this event was reported in lung adenocarcinoma previously (Kong-Beltran et al. 2006). Although the number of specimens harboring *MET* exon-skipping (n=3) is not sufficient for statistical testing, these three cancer genomes did not carry any of the canonical point driver mutations (n=47) or the canonical transforming fusion genes (n=6) (*p*-value = 0.057), suggesting that this exon-skipping event may have an independent transforming effect in the cancer. In addition, three cancer specimens expressed a short form of *FBLN2*, skipping exon 9. Skipping of exon 9 of *FBLN2* is also worthy of note (Supplementary Figure 5), since this gene was recently introduced as a tumor suppressor candidate in nasopharyngeal carcinoma (Law et al. 2012).

Lung adenocarcinoma in smokers

We compared the transcriptional landscape of lung cancers between ever-smokers and

never-smokers. There was a significant difference in the number of point mutations between the two groups (Figure 4A). On average, smokers had significantly more amino-acid-altering single nucleotide and short-indel mutations (65.0 and 20.6 mutations per cancer tissue of smokers (n=40) and never-smokers (n=33), respectively; *p*-value = 0.0011). Interestingly, the amount of smoking (pack-years) was positively correlated with the number of somatic point mutations in the cancer genome (*p*-value = 0.01; Supplementary Figure 6). We also identified differences in mutational spectrums. Cancer tissues from smokers showed similar mutational signatures to those identified previously (Pleasance et al. 2010) (C>A transversion most frequent; T>G transversion least frequent), whereas cancers from never-smokers did not. C>A transversion was more frequent in smokers (*p*-value = 3.1×10^{-6}), while T>G transversion was more common in never-smokers (*p*-value = 8.1×10^{-14}) (Figure 4B). In addition, from the gene expression profiles (Supplementary Information; Supplementary Table 10; Data access), we detected a total of 6,719 cancer outlier genes (COGs), which were extremely highly expressed in a small number of cancer tissues (Supplementary Table 11). The number of COGs per cancer tissue varied markedly (Figure 4C), ranging from 0 to 989. The lung adenocarcinomas of smokers carried significantly more COGs than those of never-smokers (*p*-value = 0.0078; Figure 4C and 4D; Supplementary Figure 7). These findings demonstrate that lung adenocarcinoma in smokers harbors more somatic mutations and greater perturbation of gene expression levels.

Co-localization of over- and under-expressed genes

Next, we assessed the gene expression pattern of each specific cancer specimen relative to the general transcriptional landscape of all 87 cancer tissues. After identifying genes which were relatively over-expressed and under-expressed in each cancer, we interestingly observed that these sets were spatially grouped together in the genome (Figure 5A; Supplementary Figure 8). We defined those regions containing such groups as jointly

regulated blocks (JRBs). The number of JRBs was highly variable among cancer tissues. In order to investigate the cause of these JRBs, we performed comparative genomic hybridization (CGH) array (Park et al. 2010) experiments on a subset of cancer samples (n=9). Interestingly, combined analyses between array results and JRBs showed that ~ 70% of JRBs can be explained by the copy number status of the cancer genome (Figure 5A and 5B). Recent reports have also shown that cancer genomes harbor large hypo-methylated (and hyper-methylated) blocks (Hansen et al. 2011; Wen et al. 2009), suggesting the combined effect of somatic copy number alterations and DNA methylation patterns are likely to induce the diversity of gene expression profiles in cancer tissues.

We merged the JRBs identified from 87 lung adenocarcinoma samples. This clearly showed that the blocks are not randomly distributed in the genome (Figure 5C). For example, gene expression is frequently increased on the short arm of chromosome 7, while expression is frequently decreased on the short arm of chromosome 3. These patterns correlate with frequent copy number alterations of cancer genomes identified in previous studies (Job et al. 2010; Weir et al. 2007).

Lymph node metastasis and *TP53* mutation

We investigated the correlation of somatic alterations with lymph node metastasis (information for lymph node metastasis is available in Supplementary Table 1). We divided the cancer samples into two groups: those with known or candidate driver mutations (canonical point driver mutations (n=47), canonical transforming fusion genes (n=6), novel tyrosine kinase fusion genes (n=4; CCDC6-*ROS1*, *FGFR2*-*CIT*, *AXL*-*MBIP*, *SCAF11*-*PDGFRA*) and *MET* exon 14 skipping (n=3)) and those without. We performed multivariate logistic regression for the presence or absence of lymph node metastasis including gender, age, cancer stage and smoking status as factors. Cancers with known or candidate driver mutations did not show higher rate of lymph node metastasis than those without (*p*-value =

0.15; multivariate logistic regression; Supplementary Table 12). However, cancer patients harboring both a known or candidate driver mutation and *TP53* mutations showed significantly higher rates of lymph node metastasis (*p*-value = 0.017; multivariate logistic regression; Supplementary Table 12). This implies that activated oncogenes and disrupted tumor suppressor genes such as *TP53* may together contribute to cancer metastasis. In addition to driver and *TP53* mutations, cancer stage showed significant association with lymph node metastasis (*p*-value = 0.00018; multivariate logistic regression; Supplementary Table 12).

Summary of driver mutations in lung adenocarcinoma

We summarize the mutational profiles of the 200 lung adenocarcinomas in Figure 6, including the results from transcriptome sequencing and from screening tests (99 with *EGFR* mutations, 6 with *KRAS* mutations and 7 with *EML4-ALK* fusions). The frequencies of *EGFR* and *KRAS* mutations in Korean patients were 60.5% (n=121) and 12.0% (n=24). Overall, ~ 75.5% (n=151/200) of lung adenocarcinomas are considered to be driven by point mutations in the 200 patients. In addition to point mutations, we found 17 tissues with fusion protein tyrosine kinase genes (*ALK*, *RET*, *ROS1*, *FGFR2*, *AXL* and *PDGFRA;* 10 from transcriptome sequencing and 7 from FISH study), which comprises 8.5% of all samples. Three samples (1.5%) carried activating exon skipping of *MET* tyrosine kinase, suggesting that around 10% of lung adenocarcinoma drivers are transcriptional variants that can be best investigated through transcriptome sequencing. Although we could not identify canonical driver mutations in 26 cancer tissues, we suggest some specific aberrations of note for each individual tissue (Supplementary Table 13).

Discussion

The landscape of lung cancer genomes has been widely investigated using genotyping microarray, sequencing of targeted cancer genes, CGH array, exome sequencing and whole-genome sequencing (Beroukhim et al. 2010; Ding et al. 2008; Lee et al. 2010). These studies provided the large repertoire of known genomic abnormalities in cancer genes (i.e. *EGFR* and *KRAS*), identified critical pathways (i.e. MAPK and PI3K pathway) for cancer transformation, and suggested putative druggable targets to develop more efficient treatments (Herbst et al. 2008). The transcriptional landscape of lung adenocarcinomas, however, has not yet been widely explored although fusion genes, alternative splicing events, and gene expression outliers may have critical roles in cancer transformation. In addition, there are several unique characteristics of lung cancer in East Asians (Bell et al. 2008) including a large proportion of female patients, the predominance of adenocarcinoma over other types and a high frequency of *EGFR* point mutations compared to Europeans. Investigation of such features may provide insights for the treatment or prevention of lung cancer in East Asians. Yet large-scale analysis of lung cancer genomes has not been performed in this ethnic group.

In this study, we have extensively analyzed the transcriptomes of 87 lung adenocarcinomas in Korean patients. Additional whole-exome and transcriptome sequencing for the adjacent paired-normal tissues was also performed to increase the specificity in identifying somatic mutations.

Transcriptome sequencing is a powerful tool to understand cancer because it captures a snapshot of diverse aspects of transformed cells. For instance, through whole-genome or whole-exome sequencing we can check for the presence of somatic mutations in cancer. However, transcriptome sequencing also provides a picture of dynamic consequences rather than just the mutations themselves. We can profile gene expression levels, gene fusions,

and alternative splicing events simultaneously, all of which contribute to the proliferation of cancer cells. Moreover, RNA-seq is a very sensitive tool to identify point mutations. For example, six specimens which were negative for *EGFR* point mutations in the conventional screening test were discovered to harbor *EGFR* point mutations by transcriptome sequencing in this study. We believe that RNA sequencing is likely to outperform genome sequencing in detection of cancer driver mutations, especially when tumor purity is relatively low. Genes with driver mutations in cancer cells are likely to be more highly expressed than in normal cells, therefore enhancing the signal-to-noise ratio in RNA-seq. For both approaches, it is important that systems are implemented which ensure the efficient collection and preservation of cancer tissues from clinic to bench.

Tumor heterogeneity is an important issue in cancer genome studies (Marusyk et al. 2010; Shah et al. 2012). Cancer tissue specimens with a low proportion of cancer cells require deeper read-depth when sequencing. We collected specimens from the center of tumors to attempt to obtain pure samples. The differences in numbers of somatic mutations per case as well as intrinsic variability in read allele frequency in RNA-seq (i.e. allele-specific expression and expression differences between cells) do not allow accurate estimation of tumor heterogeneity from cancer transcriptome sequencing data. Ignoring these issues, the distribution of read allele frequencies in cancer transcriptome sequencing for somatic point mutations suggests that the purity of the 87 specimens studied is approximately 80% on average (Supplementary Figure 9). In addition, to increase the sensitivity in detection of somatic mutations, we performed deep RNA-sequencing (approximately 10 Gb was sequenced for each cancer specimen) and applied relatively tolerant criteria for variant detection (i.e. read allele ratio should be ≥ 10% for SNV detection). However, somatic mutations from clones with lower frequencies may be unidentified in this study. One point worthy of note is that transcriptome sequencing can only detect somatic mutations from genes active in transcription. Somatic mutations in non-transcribed genes cannot be

detected. For example, a recent study reported that only around 36% of somatic mutations were expressed in breast cancers (Shah et al. 2012). However, this may be an advantage of transcriptome sequencing because somatic driver mutations must be transcribed to have a functional impact on the progression of cancer cells. Transcriptome sequencing enables us to focus on functionally active somatic mutations, thus providing functional insights into cancer genomes.

One of the aims of this project was to discover transforming fusion genes in lung adenocarcinomas. We estimated that studying 200 cancer tissues would provide 95.1% power to detect transforming fusion genes with a frequency of 1.5% in lung adenocarcinoma (Supplementary Information). By utilizing transcriptome sequencing technology, which is the most powerful method currently available to identify novel fusion genes (Maher et al. 2009), we found that *EML4-ALK*, *KIF5B-RET* and *ROS1* fusions are the three major transforming fusion genes of adenocarcinoma in Koreans. We also identified novel gene fusions, including three protein tyrosine kinase fusions. These novel fusion genes appear to be rare events in lung adenocarcinomas, because we identified only a single case of each fusion gene among 200 samples. However, they may be good druggable targets, and more functional assessments of them are required in further studies.

From this study, we obtained evidence that expression levels of genes within a specific genomic region can be over-expressed (or under-expressed) together in jointly regulated blocks, which are likely affected by differentially methylated regions (DMRs) or copy number alterations which have been reported to be important in carcinogenesis (Beroukhim et al. 2010; Hansen et al. 2011). Our CGH array analysis confirmed that genomic structural variations have a large-scale impact on the control of gene expression levels within such blocks. Future integrative studies, combining genomic structural variations, epigenomic changes and gene expression levels of cancers are necessary to understand the fine-scale mechanisms that control gene expression in cancer cells.

Finally, we observed that the expression landscapes of the cancer tissues were extremely

heterogeneous. As seen in the somatic point mutation and gene expression profile analyses, a subset of cancers harbored an extreme number of somatic point mutations and outlier genes. The pattern was unpredictable, but was not random and was associated with cigarette smoking. A recent study analyzing the impact of smoking on human normal lung tissues also supports our finding (Bosse et al. 2012).

In summary, we have comprehensively identified the genomic and transcriptional aberrations underlying lung adenocarcinoma in Koreans. The successful discovery of many aberrations in cancer genes, such as somatic mutations, gene fusions, alternative splicing events and cancer outliers, is most likely due to the strong power and comprehensive nature of whole-transcriptome sequencing. Our approach suggests a paradigm for large-scale deep transcriptome sequencing initiatives for a number of different cancer types. Our findings provide guidance for future translational studies correlating characteristics of cancer tissues and clinical features, such as drug response, recurrence and survival.

Methods

RNA and exome sequencing

All protocols of this study were approved by the Institutional Review Board of Seoul National University Hospital (Approval # C-1111-102-387) and Seoul St. Mary's Hospital (Approval # KC11TISI0678).

All the cancer and adjacent paired-normal tissue specimens used in this study were acquired from surgical specimens. Cancer and normal tissue specimens were grossly dissected and preserved immediately in liquid nitrogen after surgery. For RNA-seq, we extracted RNA from tissue using RNAiso Plus (Takara Bio Inc.), followed by purification using RNeasy MinElute (Qiagen Inc.). RNA was assessed for quality and was quantified using an RNA 6000 Nano LabChip on a 2100 Bioanalyzer (Agilent Inc.). The RNA-seq libraries were prepared as previously described (Ju et al. 2011).

For exome sequencing of matched normal tissues (and cancer specimens LC_C5 and LC_C21 for validation purposes), genomic DNA was extracted from normal lung. Genomic DNA (3μg) from each sample was sheared and used for the construction of a paired-end sequencing library as described in the protocol provided by Illumina. Enrichment of exonic sequences was then performed for each library using the SureSelect Human All Exon 50Mb Kit (Agilent Inc.) following the manufacturer's instructions.

Libraries for RNA and exome sequencing were sequenced with Illumina TruSeq SBS Kit v3 on a HiSeq 2000 sequencer (Illumina Inc.) to obtain 100-bp paired-end reads. The image analysis and base calling were performed using the Illumina pipeline (v1.8) with default settings.

Screening tests

Screening genetic tests were performed for identification of three well-known driver mutations in a subset of the 200 lung adenocarcinoma tissues as previously described: (1) Exon 18-21 of *EGFR* by PCR and Sanger sequencing (Lynch et al. 2004) (n=164); (2) Exon 2 of *KRAS* by PCR and Sanger sequencing (Eberhard et al. 2005) (n=37); (3) *EML4-ALK* fusion genes by FISH (Kwak et al. 2010) (n=163). The results of these studies are summarized in Supplementary Information and Supplementary Table 1.

Smoking history

Of the 87 individuals whose cancer specimens were RNA sequenced, smoking history before diagnosis of lung cancer was provided by 83 (47 smokers, 36 never-smokers and 4 unknowns). Information about the amount of smoking (pack-years) was available for 23 out of 47 smokers.

Sequence analyses

RNA and exome sequencing reads were aligned to the NCBI human reference genome assembly (build 37.1) using GSNAP (Wu et al. 2010) with allowance for 5% mismatches. In the same manner, the RNA sequencing reads were also aligned to a cDNA set consisting of 161,250 mRNA sequences obtained from public databases (36,742 RefSeq, 73,671 UCSC, and 161,214 Ensembl) to decrease the false positives and false negatives in variant detection from RNA sequencing data (Ju et al. 2011; Ju et al. 2012). The expression levels for 36,742 RefSeq genes were measured by uniquely-aligned RNA sequencing reads. For each gene, the number of reads aligned to it (raw read count) was normalized by RPKM (reads per kilobase per million mapped reads) (Mortazavi et al. 2008).

Somatic single nucleotide and short indel discovery

We first identified single nucleotide variations (SNVs) and short indel variants in cancer using the transcriptome sequencing data. To minimize false positive calls generated by misalignment, we used variant calls commonly identified from both the genome and the mRNA alignment. SNVs were defined according to the following three conditions: (i) the number of uniquely-mapped reads at the position should be ≥ 3; (ii) the average base quality for the position should be ≥ 20; (iii) the allele ratio at the position should be ≥ 10% for SNVs. Indels were called by the same procedure. Gene annotations for the variants were done using RefSeq genes. To identify somatic mutations, we removed SNVs and indels that were also identified in the normal tissue counterparts (76 whole-exome and 77 transcriptome sequencing). To remove potential germline variants which might not be filtered by this step, common germline SNPs which exist in normal human populations (variants identified from phase I of 1000 Genomes Project (Consortium 2010), variants with minor allele frequency > 1% from dbSNP 132 (Altshuler et al. 2010) and variants identified in normal Korean individuals (Ju et al. 2011)) were assumed to be unrelated to cancer transformation and were removed. These filtration steps might fail to remove some rare germline variants if the position was insufficiently covered in normal exome and transcriptome sequencing. However, we mostly focused on recurrent mutations in the later analysis, which are highly unlikely to be rare germline variants. Of note, among the 87 cancer specimens where RNA sequencing was performed, 11 did not have whole-exome sequencing of normal counterparts. Therefore, our lists of somatic mutations for the 11 cancer specimens are likely to include more rare germline variants than those of the other specimens. We did not consider the 11 specimens in the statistical analysis of the number of somatic mutations and smoking history of lung cancer patients.

Fusion gene detection

We detected in-frame fusion genes by utilizing the GFP software described in our previous report (Ju et al. 2012). Briefly, the method makes use of discordant read pairs, where two ends of a pair are mapped to different genes and exon-spanning reads across the exonic fusion breakpoint of chimeric transcripts. Additional filtration cascades (homology filter, fusion-spanning read filter and the fusion point filter) were applied to remove false positives (Ju et al. 2012). In addition to GFP analysis, we also assessed read depth along each exon of tyrosine kinases (e.g. *ALK*, *RET* and *ROS1*), since abrupt over-expression after fusion gene breakpoints is a hallmark of fusion for these genes (Ju et al. 2012; Lipson et al. 2012) (Supplementary Figure 3). We do not report intrachromosomal fusions between adjacent genes (<200 kb) because these are assumed to be due to read-through transcription and so do not originate in genomic rearrangement, e.g. translocation, inversion or large deletion (Ju et al. 2012). Finally, off-frame fusion genes were removed.

Alternative splicing

First, exon-skipping reads were extracted from the collection of sequencing reads for each cancer sample. The GSNAP alignment tool conveniently allows reads to be split into two segments and mapped to different exons when genomic positions of exons are provided to the program. We collected those spliced sequencing reads where non-adjacent exons of a gene were joined and defined them as exon-skipping sequencing reads. For candidate skipped exons supported by at least two exon-skipping reads, we obtained the expression levels for the candidate exon and its neighboring exons in terms of RPKM. Since the expression level for an exon is correlated with the expression level for the gene from which it derives, the expression of the exons was normalized by the expression level of the gene to which they belong. Next, the fold changes between the 5' neighboring and the 3' neighboring exons and the candidate exon were calculated and averaged. With an assumption of normal

distribution, a z-value for the fold change was obtained by considering all the fold changes calculated in the same manner in all the normal samples (n=77). If any of the normal averaged fold changes was less than 0.9, the candidate skipped exon was not further considered because dropped coverage was not specific to cancer.

Differentially Expressed Gene (DEG) selection

We selected DEGs by clustering genes by expression levels. First, genes with either RPKM < 1.0 or coefficient of variation (CV) < 0.7 were excluded to remove genes non-informative for clustering. This resulted in a total of 3,051 unique genes. Log_2 transformation and additional row-wise and column-wise normalization was applied: row-wise normalization was applied to each gene by subtracting the gene's median expression value from individual expression values. Similarly, column-wise normalization was applied to each sample by subtracting the sample's median expression value from individual expression values. This guarantees that row-wise and column-wise median expression values were both set to zero. Then, hierarchical clustering was done by Gene Cluster 3.0 with default parameters (de Hoon et al. 2004), correlation (uncentered) and complete linkage. A heatmap was drawn by R package function heatmap.2. (http://cran.r-project.org/web/packages/gplots/index.html). Finally, we referred to the hierarchical tree generated by the clustering process and selected three types of DEGs (cancer-UP, cancer-DOWN and mixed).

Cancer outlier gene analysis

Cancer outlier gene analysis was performed for 22,427 RefSeq genes across a total of 164 RNA-seq results with modified criteria suggested previously (MacDonald et al. 2006; Tomlins et al. 2005). The analysis pipeline is as follows. (i) All the expression values are subtracted by their median, which sets the gene's median to zero (location normalization). (ii) All the

expression values are then divided by their median absolute deviation (MAD) (scale normalization). (iii) Given a set of normalized expression values, (q75 + 3 x IQR) is defined as an outlier cutoff where q75 is the 75th percentile expression value and IQR (inter quartile range) is the absolute difference between 25th and 75th percentile expression values. An expression value is accepted as an outlier when its original RPKM is greater than or equal to 1.0 and its normalized expression value exceeds the estimated outlier cutoff. (iv) Genes showing an outlier pattern in any normal samples are excluded. (v) Finally, genes which exhibit an outlier pattern in at least one cancer sample are chosen as candidate cancer outlier genes.

Jointly Regulated Block (JRB) identification

To generate the gene expression signatures of each cancer sample, we calculated the ratio of gene expression levels in a cancer tissue to the average expression levels in 77 adjacent normal tissues. The gene set for analysis was formed from 36,742 transcripts in the RefSeq database, which yielded 22,427 genes after filtering out redundant entries.

To quantify relative expression among cancer samples, we compared the gene expression ratio of a gene (gene A) in the ith cancer sample with the ratio in all 87 cancer tissues and calculated a normalized z-score as follows.

$$\left(Z_{i,A} = \frac{\text{expression ratio}_{i,A} - \text{average expression ratio}_{A}}{SD_{A}}\right),$$

where SD is standard deviation of expression ratios.

During this process, genes with low expression levels (maximum expression < 3 RPKM) or genes with small variance in expressions (relative standard variation < 0.1) were removed, since signal to noise ratio for these genes is not sufficient. This left 16,419 genes for further investigation.

Given a set of z-scores, we calculated the moving-average of 10 z-scores by walking

through the chromosomes. An increased-expression JRB is defined as starting at a gene with a z-score > 1.5 and extending in both directions until a z-score < 0.5 is reached. Once its boundary is determined, the JRB must satisfy at least one of the three following criteria. (i) more than 40 genes within a block (ii) more than 20 genes in a block, and an average z-score > 1.2 (iii) an average z-score > 2.0. On the other hand, we applied slightly different conditions in discovering decreased-expression JRBs to increase the sensitivity. A decreased-expression JRB is defined as starting at a gene with a z-score < -1.0 and extending in both directions until a z-score > -0.5 is reached. Then the JRB must satisfy at least one of the three following criteria. (i) more than 40 genes within a block (ii) more than 20 genes in a block with an average z-score < -0.8. (iii) an average z-score < -1.0.

The comparison of JRBs and copy number alterations provided by Comparative Genomic Hybridization array was done by calculating the correlation (r^2) between the averaged z-scores of JRBs and the averaged log2ratio values of probes within the same JRBs. CGH array analyses were performed using 9 cancer samples (LC_C7, LC_C21, LC_C25, LC_C35, LC_S19, LC_S23, LC_S39, LC_S42 and LC_S51) and their normal counterparts using a customized Agilent 180K CGH array platform (Park et al. 2010). CGH array experiments were conducted according to the manufacturer's instructions (Agilent Inc.). Briefly, genomic DNA from cancer and adjacent-normal specimens were labeled by Cy-5 and Cy-3 dye, respectively, and hybridized. Log2 ratio was calculated by image analysis of CGH array using the CGH-105_Jan09 protocol (Agilent Inc.) for background subtraction and normalization.

Validation of fusion genes

Fusion genes were validated using PCR amplification of fusion gene breakpoints of chimeric cDNA and Sanger sequencing. The PCR reactions were 10 min at 95℃; 30 cycles of 30 sec

at 95℃, 30 sec at 62℃ and 30 sec at 72℃ and finally 10 min at 72℃. All the Sanger sequencing experiments were performed at Macrogen Inc. (http://www.macrogen.com). PCR and Sanger sequencing primers are shown in Supplementary Table 7.

Statistical analyses

The differences in number of somatic mutations and COGs between smokers and never-smokers (Figure 4A and 4C) were tested using Student's t-test. Chi-square tests were applied on the difference in mutation spectrums (Figure 4B). Logistic regression analysis was performed to assess the relationship between somatic mutations (known or candidate driver mutations and *TP53* mutations) and lymph node metastasis using gender, age, cancer stage and smoking status as covariates (Supplementary Table 12). Two-sided *p*-values were calculated for all these statistical tests.

Mutual exclusion and concurrence analysis

We carried out pair-wise mutual exclusion and concurrence analysis for genes that showed more than three mutations (including somatic point mutations, fusion genes and skipped exons). For a given pair of mutated genes A and B, we recorded the number of samples in possible four categories (A mutated only, B mutated only, A and B both mutated and neither). Then Fisher's exact test was performed to infer the mutual dependency between the two genes. Once two genes were determined to be mutually dependent on each other by the test, their mutual exclusion/concurrence was determined by calculating Pearson's correlation coefficient, r (mutual exclusion: $r < 0$ and concurrence: $r > 0$).

Data access

Gene expression values for 87 lung adenocarcinomas and 77 adjacent normal tissues can be viewed at http://gene.gmi.ac.kr and at the Gene Expression Omnibus (GEO) under accession number GSE40419. Transcriptome and exome sequencing data are uploaded to EBI-SRA under accession number ERP001058 (transcriptome sequencing; http://www.ebi.ac.uk/ena/data/view/ERP001058) and ERP001575 (exome sequencing), and these data are available via our FTP (ftp://ftp.gmi.ac.kr/asianGenome/data/Lung_cancer/) during review process.

Acknowledgments

We acknowledge the anonymous lung adenocarcinoma patients who enrolled in this study. We also thank Whijae Roh at GMI-SNU for his help on the statistical analyses and Sung-Soo Park at W Corporation for his advice on constructing the web-site for browsing gene expression values. This work has been supported in part by the Korean Ministry of Knowledge Economy (grant # 10037410 to J.-S.S.), by the National Research Foundation (NRF) of Korea funded by Korea Government (MEST) (Grant No. 2011-0016106 to Y.T.K and project of Global Ph.D. Fellowship to W.-C.L).

Figure Legends

Figure 1. The transcriptional landscape and mutational profile of 87 lung adenocarcinomas.

Each column characterizes the signature of cancer in one patient. Patients are classified into smokers and never-smokers. Each row represents a selected gene of interest, including known driver genes of lung adenocarcinoma, genes observed to be frequently mutated (≥3 cancers) and protein kinase genes involved in fusion and exon-skipping (ES) events. Cells are colored to indicate discovery of somatic mutations (red), gene over-expression (blue), both of these two (purple), and fusion and ES events (green) in cancer tissue. Patients with lymph node metastasis are indicated with “M” on the bottom row.

Figure 2. Graphical representation of 45 fusion genes identified from transcriptome sequencing of 87 lung adenocarcinomas.

Protein kinase-containing fusion genes are indicated with red lines joining the two genomic loci, while other fusions are indicated by blue lines. The protein kinase genes and their fusion partners are labeled in red and green respectively (outer layer).

Figure 3. Fusion genes and alternative splicing events revealed by RNA sequencing.

(A) Schematic figures of the domain structures of novel protein kinase fusion genes.

(B) Exon 14 skipping in *MET* proto-oncogene demonstrated by read depth across gene model. TM: trans-membrane domain.

Figure 4. Mutational and transcriptional variation in cancer between never-smokers and smokers.

(A) The number of somatic mutations (non-synonymous single nucleotide and short-indel mutations) in the cancer tissue of each patient. Patients are classified into never-smokers and smokers, and further sorted by mutation count. (Inset) Box plot of somatic mutation counts for never-smokers and smokers. The two groups are significantly different (p = 0.001079).

(B) The proportion of the six possible non-synonymous substitutions found within smokers and never-smokers. The two groups were significantly different with respect to transversions C>A and T>G (**, $p < 0.001$) and transversion T>A (*, $p < 0.01$).

(C) The number of cancer-outlier genes (COGs; extremely high-expressed genes in a subset of cancer specimens; see methods for details) in each cancer tissue. Patients are sorted as above. (Inset) Box plot showing that lung adenocarcinoma in smokers contains more cancer-outlier genes.

(D) Gene expression within cancer tissues against average expression in normal tissue. Scatter plots for patients LC_S33 (a never-smoker patient) and LC_S51 (a smoker patient) are shown, providing an example of the variation in gene expression perturbation. Selected genes of interest are labeled. Genes were categorized as “Cancer-up” where generally over-expressed and “Cancer-down” where generally under-expressed in lung cancer compared to paired-normal tissue by hierarchical clustering (see Supplementary Information).

Figure 5. Jointly regulated blocks (JRBs) identified from gene expression signatures.

(A) Large JRBs observed on chromosome 5 in one cancer sample (patient LC_S51) and its high correlation with CGH array results. (Top row) Relative expression levels of the genes on chromosome 5 (gray dots), their moving averages (red line) and detected JRBs (red horizontal bars). (Middle row) CGH array results for patient LC_S51. Log_2 ratio of probes (blue dots) and identified copy number alterations (blue horizontal bars). (Bottom row) Karyogram of chromosome 5.

(B) Correlation between JRBs and CGH array data for three cancer specimens. The x-axis represents the averaged Z-scores of JRB and y-axis indicates the averaged CGH array $\log_2$ ratios for the genomic area.

(C) The genomic location of JRBs and number of cancer tissues involved. Increased- and decreased-expression JRBs are shown in blue and red bars respectively.

Figure 6. A summary of the mutational profiles of 200 lung adenocarcinomas.

Pie chart shows the distribution of driver mutations identified in 200 lung adenocarcinoma patients in this study.

Tables

Table 1. The list of 15 selected gene fusions identified from 87 lung adenocarcinomas.

PTK; protein tyrosine kinase.

Donor gene	Acceptor gene	Chromosome	# samples	Distance (Mb)	Function
Known					
EML4	*ALK*	chr2;chr2	1	12.252	PTK
KIF5B	*RET*	chr10;chr10	4	11.227	PTK
CD74	*ROS1*	chr5;chr6	1	Interchromosomal	PTK
SLC34A2	*ROS1*	chr4;chr6	1	Interchromosomal	PTK
Novel					
CCDC6	*ROS1*	chr10;chr6	1	Interchromosomal	PTK
SCAF11	*PDGFRA*	chr12;chr4	1	Interchromosomal	PTK
FGFR2	*CIT*	chr10;chr12	1	Interchromosomal	PTK
AXL	*MBIP*	chr19;chr14	1	Interchromosomal	PTK
MAP4K3	*PRKCE*	chr2;chr2	1	6.215	Ser/Thr kinase
BCAS3	*MAP3K3*	chr17;chr17	1	2.23	Ser/Thr kinase
ERBB2IP	*MAST4*	chr5;chr5	1	0.515	Ser/Thr kinase
KRAS	*CDH13*	chr12;chr16	1	Interchromosomal	signaling
APLP2	*TNFSF11*	chr11;chr13	1	Interchromosomal	signaling
ZFYVE9	*CGA*	chr1;chr6	1	Interchromosomal	signaling
TPD52L1	*TRMT11*	chr6;chr6	1	0.723	tumor protein

References

Aberle, D.R., A.M. Adams, C.D. Berg, W.C. Black, J.D. Clapp, R.M. Fagerstrom, I.F. Gareen, C. Gatsonis, P.M. Marcus, and J.D. Sicks. 2011. Reduced lung-cancer mortality with low-dose computed tomographic screening. *N Engl J Med* **365:** 395-409.

Alberti, L., C. Carniti, C. Miranda, E. Roccato, and M.A. Pierotti. 2003. RET and NTRK1 proto-oncogenes in human diseases. *J Cell Physiol* **195:** 168-186.

Altshuler, D.M., R.A. Gibbs, L. Peltonen, E. Dermitzakis, S.F. Schaffner, F. Yu, P.E. Bonnen, P.I. de Bakker, P. Deloukas, S.B. Gabriel et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* **467:** 52-58.

Bell, D.W., B.W. Brannigan, K. Matsuo, D.M. Finkelstein, R. Sordella, J. Settleman, T. Mitsudomi, and D.A. Haber. 2008. Increased prevalence of EGFR-mutant lung cancer in women and in East Asian populations: analysis of estrogen-related polymorphisms. *Clin Cancer Res* **14:** 4079-4084.

Beroukhim, R., C.H. Mermel, D. Porter, G. Wei, S. Raychaudhuri, J. Donovan, J. Barretina, J.S. Boehm, J. Dobson, M. Urashima et al. 2010. The landscape of somatic copy-number alteration across human cancers. *Nature* **463:** 899-905.

Bosse, Y., D.S. Postma, D.D. Sin, M. Lamontagne, C. Couture, N. Gaudreault, P. Joubert, V. Wong, M. Elliott, M. van den Berge et al. 2012. Molecular Signature of Smoking in Human Lung Tissues. *Cancer Res* **72:** 3753-3763.

Chan, E.F., U. Gat, J.M. McNiff, and E. Fuchs. 1999. A common human skin tumour is caused by activating mutations in beta-catenin. *Nat Genet* **21:** 410-413.

Consortium, T.G.P. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061-1073.

de Hoon, M.J., S. Imoto, J. Nolan, and S. Miyano. 2004. Open source clustering software. *Bioinformatics* **20:** 1453-1454.

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

Eberhard, D.A., B.E. Johnson, L.C. Amler, A.D. Goddard, S.L. Heldens, R.S. Herbst, W.L. Ince, P.A. Janne, T. Januario, D.H. Johnson et al. 2005. Mutations in the epidermal growth factor receptor and in KRAS are predictive and prognostic indicators in patients with non-small-cell lung cancer treated with chemotherapy alone and in combination with erlotinib. *J Clin Oncol* **23:** 5900-5909.

Gardina, P.J., T.A. Clark, B. Shimada, M.K. Staples, Q. Yang, J. Veitch, A. Schweitzer, T. Awad, C. Sugnet, S. Dee et al. 2006. Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics* **7:** 325.

Haber, D.A. and J. Settleman. 2007. Cancer: drivers and passengers. *Nature* **446:** 145-146.

Hansen, K.D., W. Timp, H.C. Bravo, S. Sabunciyan, B. Langmead, O.G. McDonald, B. Wen, H. Wu, Y. Liu, D. Diep et al. 2011. Increased methylation variation in epigenetic domains across cancer types. *Nat Genet* **43:** 768-775.

Herbst, R.S., J.V. Heymach, and S.M. Lippman. 2008. Lung cancer. *N Engl J Med* **359:** 1367-1380.

Jemal, A., F. Bray, M.M. Center, J. Ferlay, E. Ward, and D. Forman. 2011. Global cancer statistics. *CA Cancer J Clin* **61:** 69-90.

Job, B., A. Bernheim, M. Beau-Faller, S. Camilleri-Broet, P. Girard, P. Hofman, J. Mazieres, S. Toujani, L. Lacroix, J. Laffaire et al. 2010. Genomic aberrations in lung adenocarcinoma in never smokers. *PLoS One* **5:** e15145.

Ju, Y.S., J.I. Kim, S. Kim, D. Hong, H. Park, J.Y. Shin, S. Lee, W.C. Lee, S.B. Yu, S.S. Park et al. 2011. Extensive genomic and transcriptional diversity identified through massively parallel DNA and RNA sequencing of eighteen Korean individuals. *Nat Genet* **43:** 745-752.

Ju, Y.S., W.C. Lee, J.Y. Shin, S. Lee, T. Bleazard, J.K. Won, Y.T. Kim, J.I. Kim, J.H. Kang, and J.S. Seo. 2012. A transforming KIF5B and RET gene fusion in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing. *Genome Res* **22:** 436-445.

Kas, K., M.L. Voz, E. Roijer, A.K. Astrom, E. Meyen, G. Stenman, and W.J. Van de Ven. 1997. Promoter swapping between the genes for a novel zinc finger protein and beta-catenin in pleiomorphic adenomas with t(3;8)(p21;q12) translocations. *Nat Genet* **15:** 170-174.

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

Kong-Beltran, M., S. Seshagiri, J. Zha, W. Zhu, K. Bhawe, N. Mendoza, T. Holcomb, K. Pujara, J. Stinson, L. Fu et al. 2006. Somatic mutations lead to an oncogenic deletion of met in lung cancer. *Cancer Res* **66:** 283-289.

Kwak, E.L., Y.J. Bang, D.R. Camidge, A.T. Shaw, B. Solomon, R.G. Maki, S.H. Ou, B.J. Dezube, P.A. Janne, D.B. Costa et al. 2010. Anaplastic lymphoma kinase inhibition in non-small-cell lung cancer. *N Engl J Med* **363:** 1693-1703.

Law, E.W., A.K. Cheung, V.I. Kashuba, T.V. Pavlova, E.R. Zabarovsky, H.L. Lung, Y. Cheng, D. Chua, D. Lai-Wan Kwong, S.W. Tsao et al. 2012. Anti-angiogenic and tumor-suppressive roles of candidate tumor-suppressor gene, Fibulin-2, in nasopharyngeal carcinoma. *Oncogene* **31:** 728-738.

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

Lipson, D., M. Capelletti, R. Yelensky, G. Otto, A. Parker, M. Jarosz, J.A. Curran, S. Balasubramanian, T. Bloom, K.W. Brennan 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., D.W. Bell, R. Sordella, S. Gurubhagavatula, R.A. Okimoto, B.W. Brannigan, P.L. Harris, S.M. Haserlat, J.G. Supko, F.G. Haluska 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.

MacDonald, J.W. and D. Ghosh. 2006. COPA--cancer outlier profile analysis. *Bioinformatics* **22:** 2950-2951.

Maemondo, M., A. Inoue, K. Kobayashi, S. Sugawara, S. Oizumi, H. Isobe, A. Gemma, M. Harada, H. Yoshizawa, I. Kinoshita et al. 2010. Gefitinib or chemotherapy for non-small-cell lung cancer with mutated EGFR. *N Engl J Med* **362:** 2380-2388.

Maher, C.A., C. Kumar-Sinha, X. Cao, S. Kalyana-Sundaram, B. Han, X. Jing, L. Sam, T. Barrette, N. Palanisamy, and A.M. Chinnaiyan. 2009. Transcriptome sequencing to detect gene fusions in cancer. *Nature* **458:** 97-101.

Marusyk, A. and K. Polyak. 2010. Tumor heterogeneity: causes and consequences. *Biochim Biophys Acta* **1805:** 105-117.

Mok, T.S., Y.L. Wu, S. Thongprasert, C.H. Yang, D.T. Chu, N. Saijo, P. Sunpaweravong, B. Han, B. Margono, Y. Ichinose et al. 2009. Gefitinib or carboplatin-paclitaxel in pulmonary adenocarcinoma. *N Engl J Med* **361:** 947-957.

Mortazavi, A., B.A. Williams, K. McCue, L. Schaeffer, and B. Wold. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* **5:** 621-628.

Pao, W. and N. Girard. 2011. New driver mutations in non-small-cell lung cancer. *Lancet Oncol* **12:** 175-180.

Park, H., J.I. Kim, Y.S. Ju, O. Gokcumen, R.E. Mills, S. Kim, S. Lee, D. Suh, D. Hong, H.P. Kang et al. 2010. Discovery of common Asian copy number variants using integrated high-resolution array CGH and massively parallel DNA sequencing. *Nat Genet* **42:** 400-405.

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

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

Schmidt, L., K. Junker, N. Nakaigawa, T. Kinjerski, G. Weirich, M. Miller, I. Lubensky, H.P. Neumann, H. Brauch, J. Decker et al. 1999. Novel mutations of the MET proto-oncogene in papillary renal carcinomas. *Oncogene* **18:** 2343-2350.

Sequist, L.V., R.S. Heist, A.T. Shaw, P. Fidias, R. Rosovsky, J.S. Temel, I.T. Lennes, S. Digumarthy, B.A. Waltman, E. Bast et al. 2011. Implementing multiplexed genotyping of non-small-cell lung cancers into routine clinical practice. *Ann Oncol* **22:** 2616-2624.

Shah, S.P., A. Roth, R. Goya, A. Oloumi, G. Ha, Y. Zhao, G. Turashvili, J. Ding, K. Tse, G. Haffari et al. 2012. The clonal and mutational evolution spectrum of primary triple-negative breast

cancers. *Nature* **486:** 395-399.

Shigematsu, H. and A.F. Gazdar. 2006. Somatic mutations of epidermal growth factor receptor signaling pathway in lung cancers. *Int J Cancer* **118:** 257-262.

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

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

Tomlins, S.A., D.R. Rhodes, S. Perner, S.M. Dhanasekaran, R. Mehra, X.W. Sun, S. Varambally, X. Cao, J. Tchinda, R. Kuefer et al. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. *Science* **310:** 644-648.

Travis, W.D., K. Garg, W.A. Franklin, Wistuba, II, B. Sabloff, M. Noguchi, R. Kakinuma, M. Zakowski, M. Ginsberg, R. Padera et al. 2005. Evolving concepts in the pathology and computed tomography imaging of lung adenocarcinoma and bronchioloalveolar carcinoma. *J Clin Oncol* **23:** 3279-3287.

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

Wen, B., H. Wu, Y. Shinkai, R.A. Irizarry, and A.P. Feinberg. 2009. Large histone H3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nat Genet* **41:** 246-250.

Wu, T.D. and S. Nacu. 2010. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26:** 873-881.

Figure 1

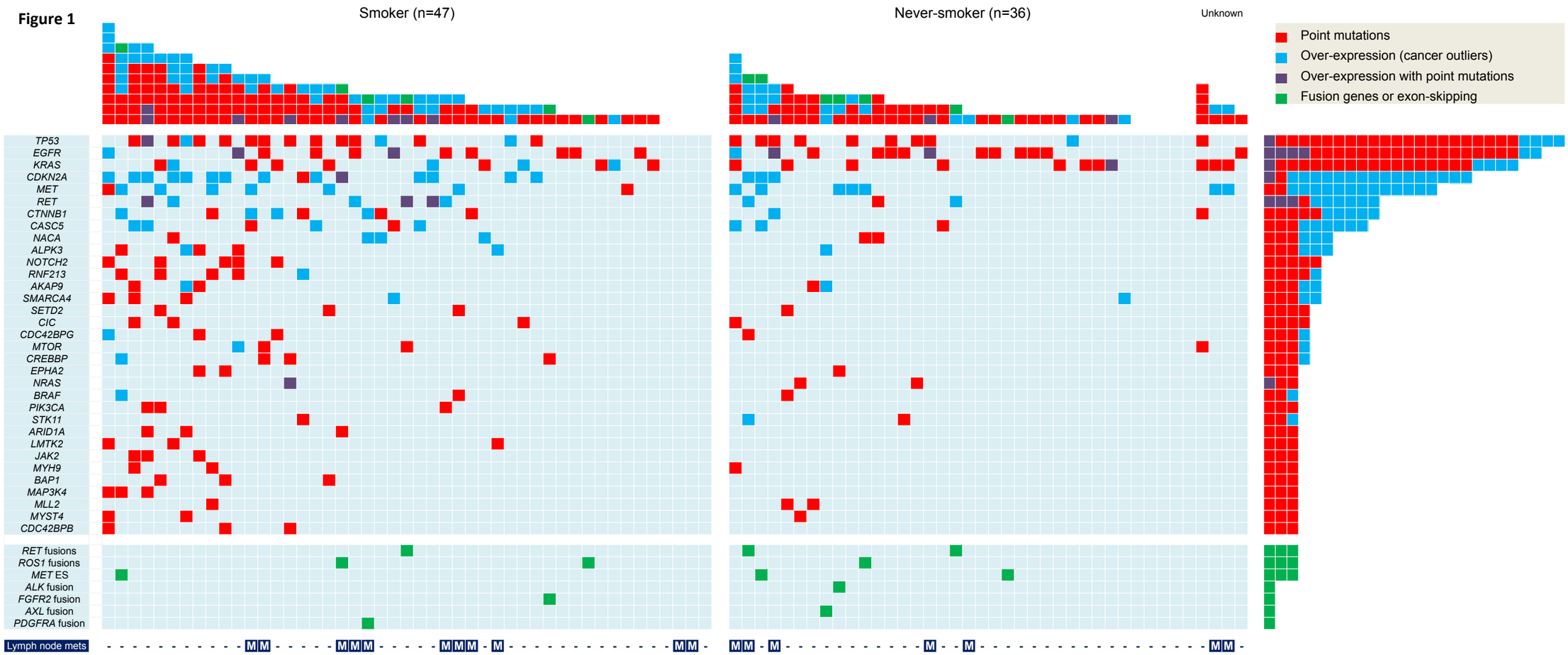
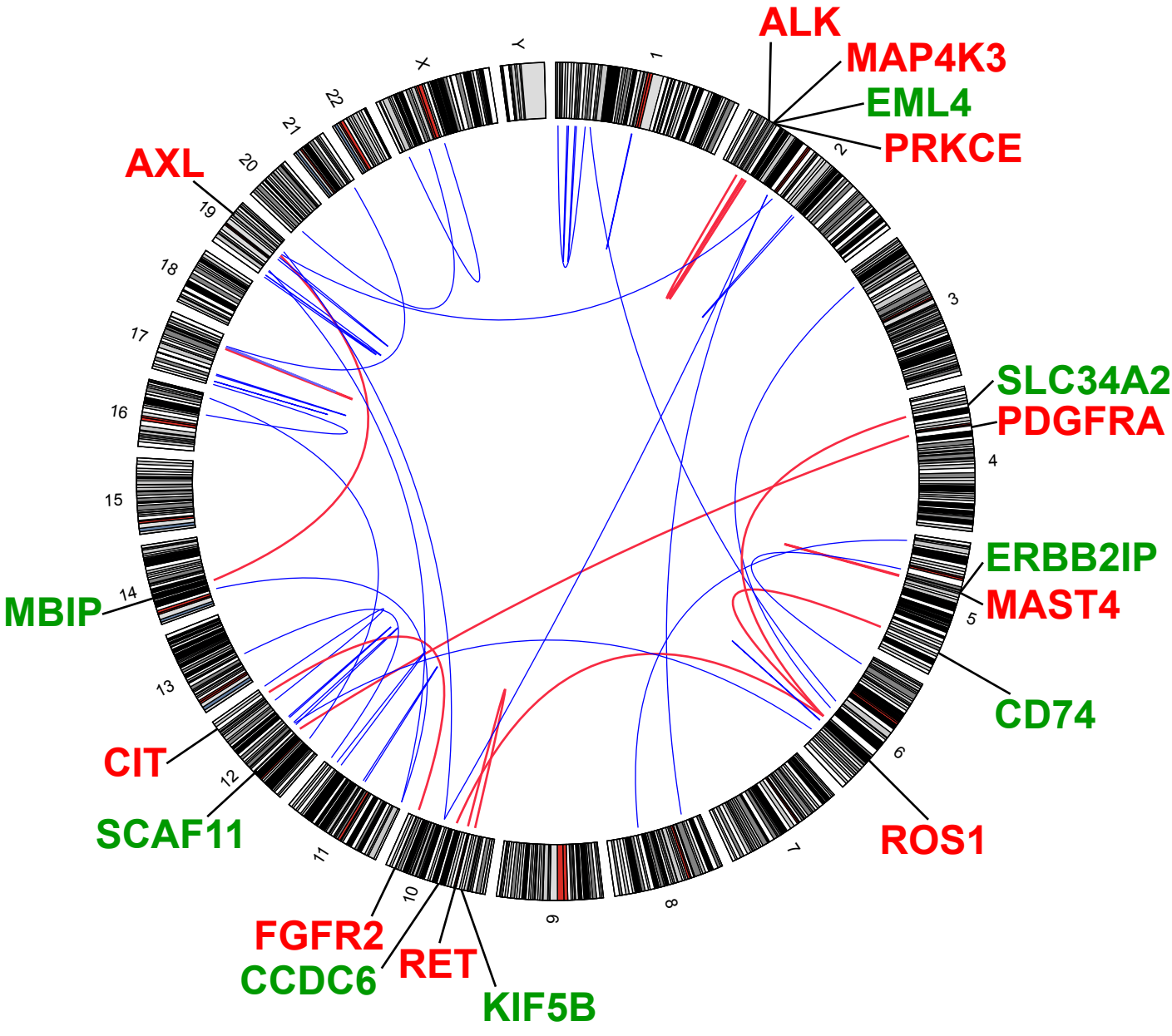
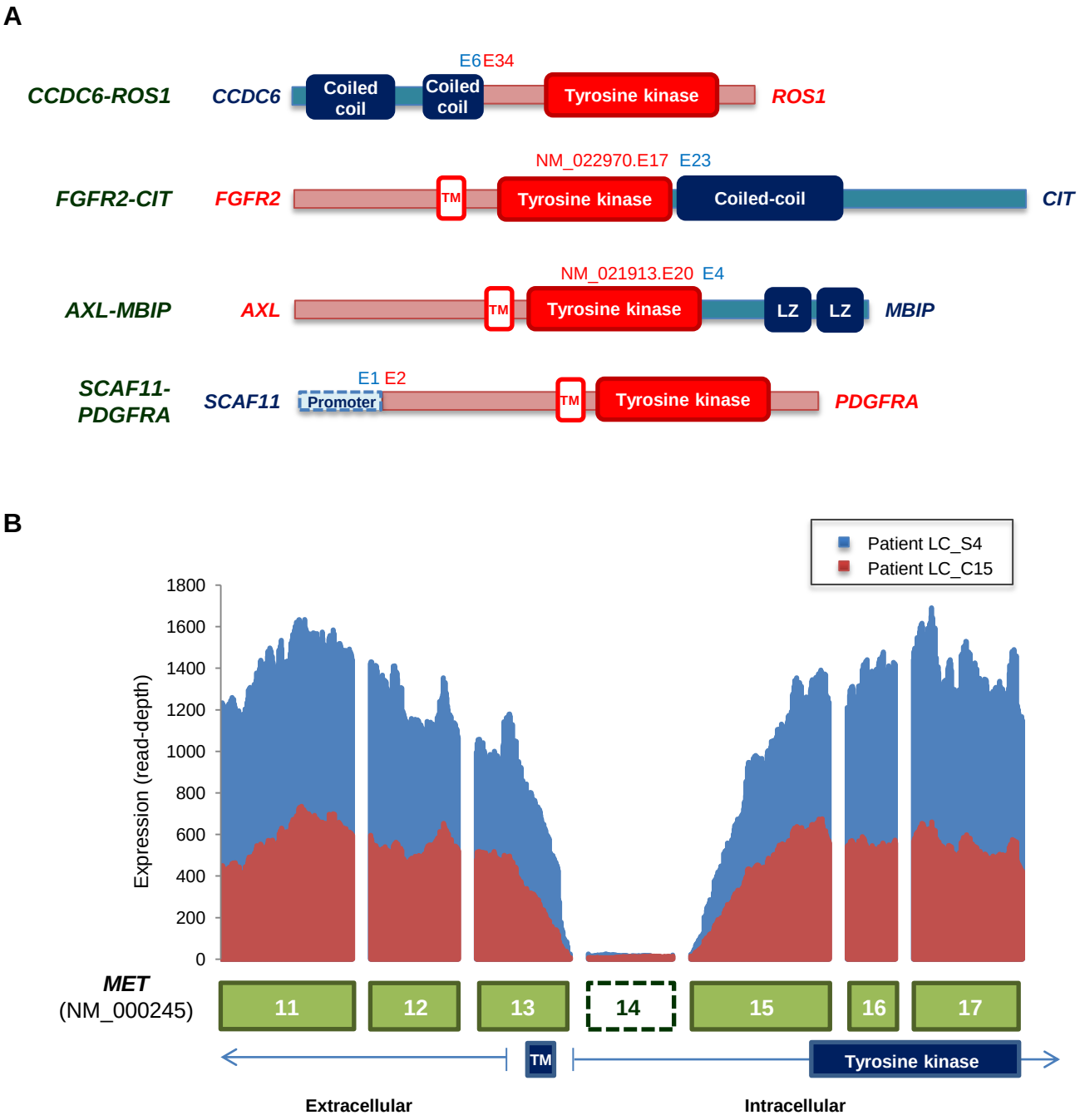
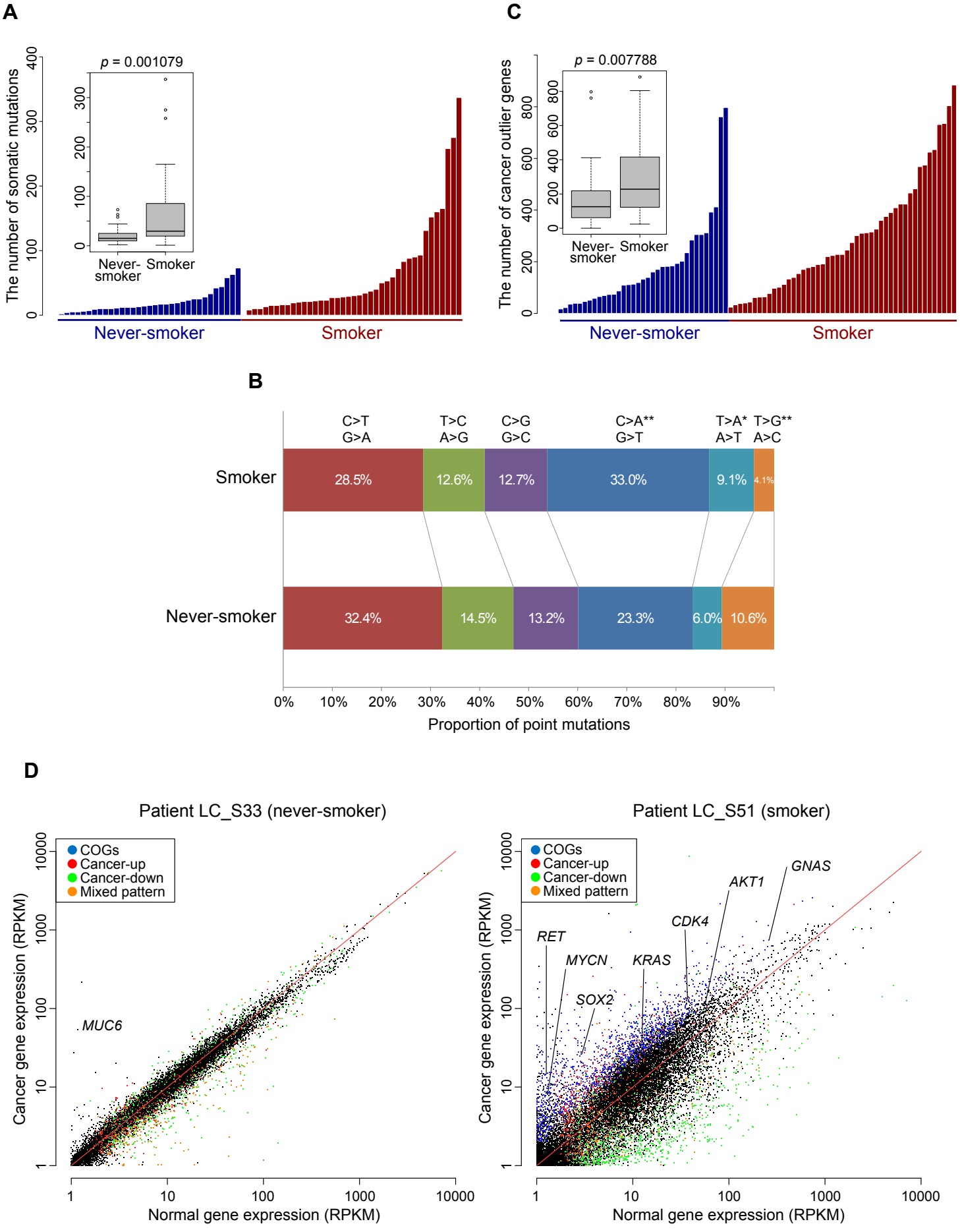

Smoker (n=47)
Never-smoker (n=36)
Unknown
Point mutations
Over-expression (cancer outliers)
Over-expression with point mutations
Fusion genes or exon-skipping
TP53
EGFR
KRAS
CDKN2A
MET
RET
CTNNB1
CASC5
NACA
ALPK3
NOTCH2
RNF213
AKAP9
SMARCA4
SETD2
CIC
CDC42BPG
MTOR
CREBBP
EPHA2
NRAS
BRAF
PIK3CA
STK11
ARID1A
LMTK2
JAK2
MYH9
BAP1
MAP3K4
MLL2
MYST4
CDC42BPB
RET fusions
ROS1 fusions
MET ES
ALK fusion
FGFR2 fusion
AXL fusion
PDGFRA fusion
Lymph node mets
M

Figure 2


ALK
MAP4K3
EML4
PRKCE
AXL
SLC34A2
PDGFRA
ERBB2IP
MAST4
MBIP
CD74
CIT
SCAF11
ROS1
FGFR2
CCDC6
RET
KIF5B
X
Y
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22


A
CCDC6-ROS1
CCDC6
E6 E34
Coiled coil
Coiled coil
Tyrosine kinase
ROS1
FGFR2-CIT
FGFR2
NM_022970.E17 E23
TM
Tyrosine kinase
Coiled-coil
CIT
AXL-MBIP
AXL
NM_021913.E20 E4
TM
Tyrosine kinase
LZ
LZ
MBIP
SCAF11-PDGFRA
SCAF11
E1 E2
Promoter
TM
Tyrosine kinase
PDGFRA
B
Patient LC_S4
Patient LC_C15
Expression (read-depth)
1800
1600
1400
1200
1000
800
600
400
200
0
MET
(NM_000245)
11
12
13
14
15
16
17
TM
Tyrosine kinase
Extracellular
Intracellular


A
p = 0.001079
The number of somatic mutations
Never-smoker
Smoker
C
p = 0.007788
The number of cancer outlier genes
Never-smoker
Smoker
B
C>T G>A
T>C A>G
C>G G>C
C>A** G>T
T>A* A>T
T>G** A>C
Smoker
28.5%
12.6%
12.7%
33.0%
9.1%
4.1%
Never-smoker
32.4%
14.5%
13.2%
23.3%
6.0%
10.6%
Proportion of point mutations
D
Patient LC_S33 (never-smoker)
Patient LC_S51 (smoker)
COGs
Cancer-up
Cancer-down
Mixed pattern
Cancer gene expression (RPKM)
Normal gene expression (RPKM)
MUC6
RET
MYCN
SOX2
KRAS
CDK4
AKT1
GNAS

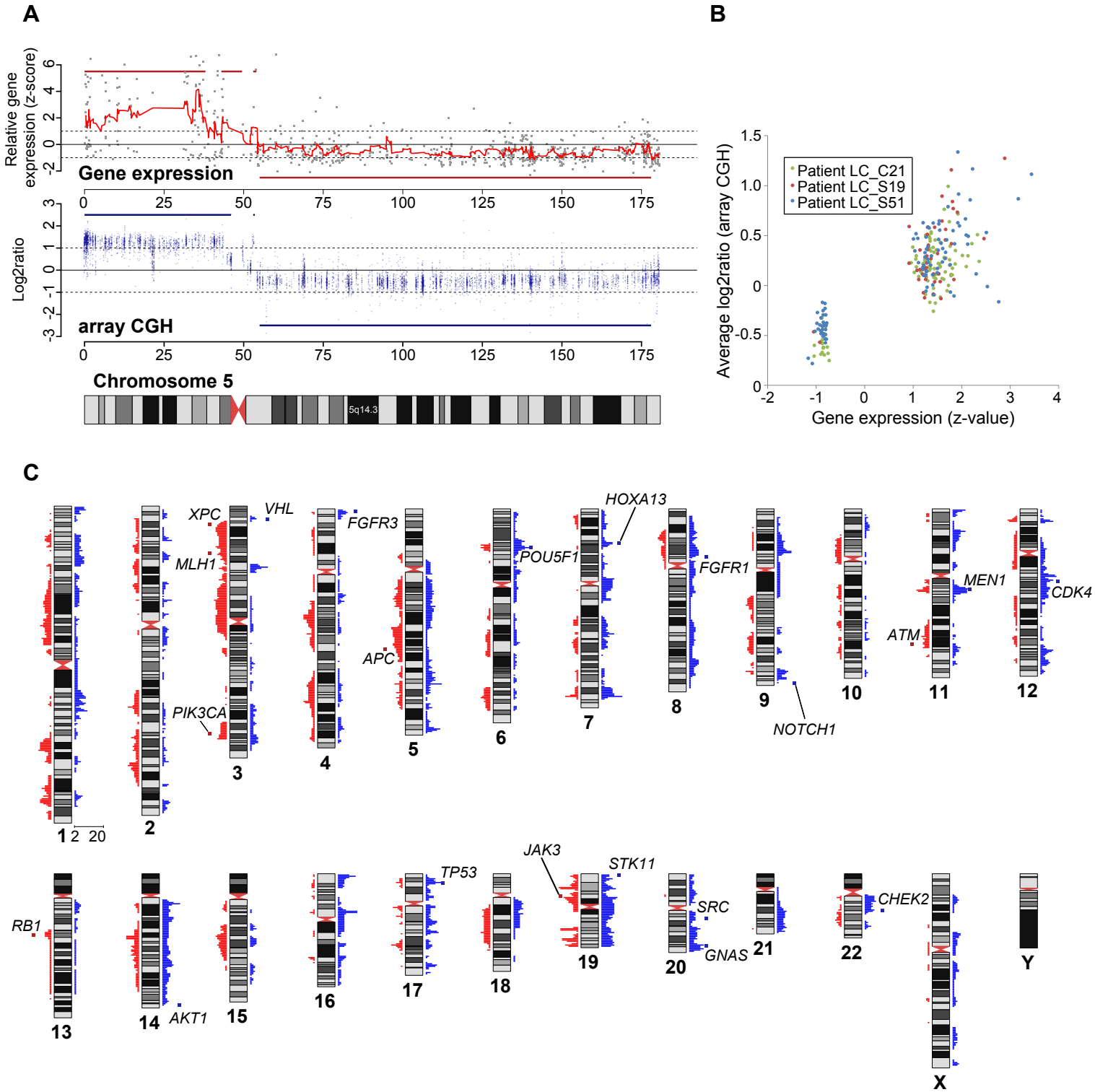
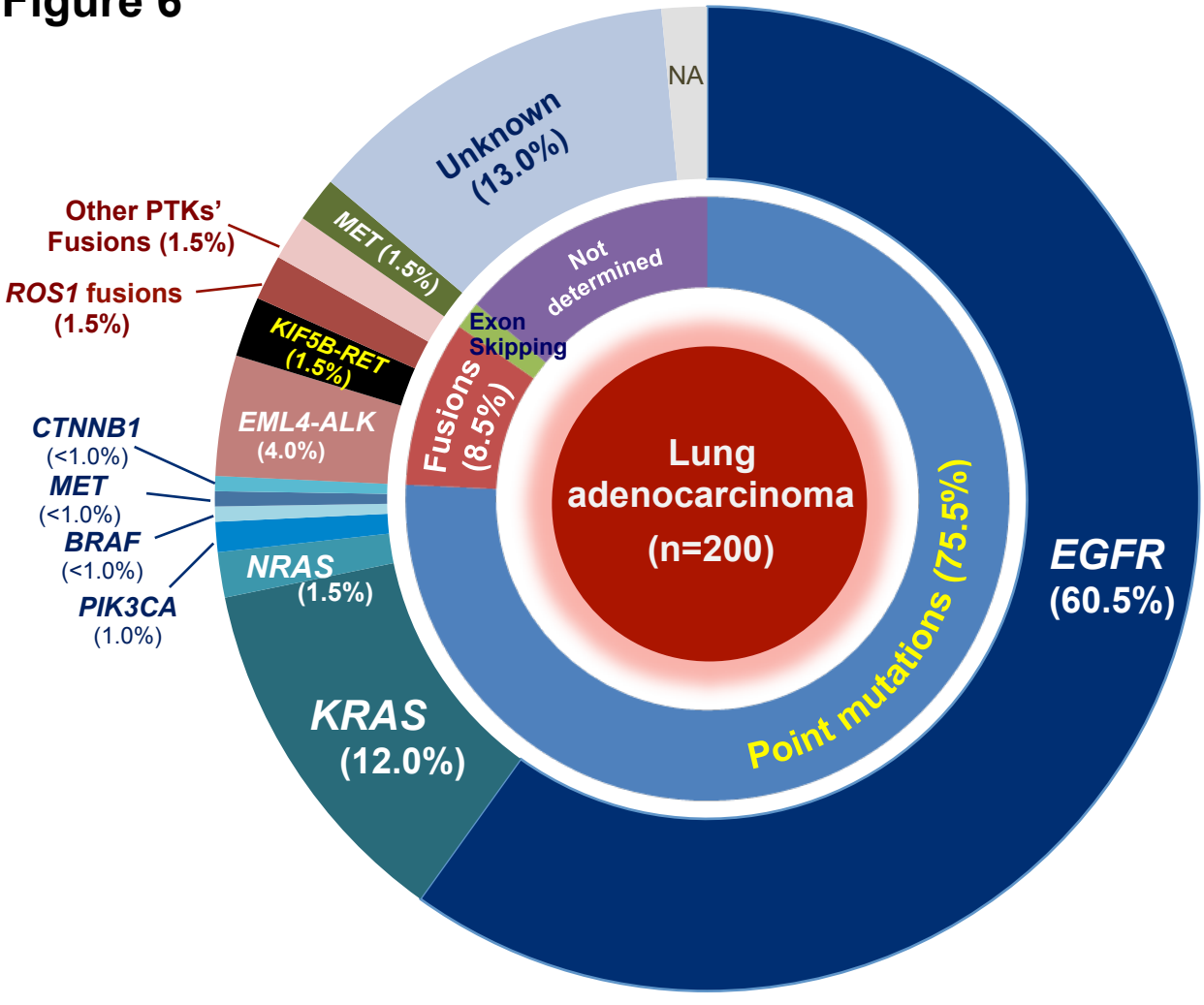

A
Relative gene expression (z-score)
Gene expression
Log2ratio
array CGH
Chromosome 5
5q14.3
B
Average log2ratio (array CGH)
Gene expression (z-value)
Patient LC_C21
Patient LC_S19
Patient LC_S51
C
XPC
VHL
MLH1
PIK3CA
FGFR3
APC
POU5F1
HOXA13
FGFR1
NOTCH1
MEN1
CDK4
ATM
RB1
AKT1
TP53
JAK3
STK11
SRC
GNAS
CHEK2

Figure 6


Lung adenocarcinoma (n=200)
Point mutations (75.5%)
Fusions (8.5%)
Exon Skipping
Not determined
EGFR (60.5%)
KRAS (12.0%)
NRAS (1.5%)
BRAF (<1.0%)
MET (<1.0%)
CTNNB1 (<1.0%)
PIK3CA (1.0%)
EML4-ALK (4.0%)
KIF5B-RET (1.5%)
ROS1 fusions (1.5%)
Other PTKs' Fusions (1.5%)
MET (1.5%)
Unknown (13.0%)
NA