Supplementary Materials for

Intra-tumor Heterogeneity in Localized Lung Adenocarcinomas Delineated by Multi-region Sequencing

Authors: Jianjun Zhang, Junya Fujimoto, Jianhua Zhang, David C. Wedge, Xingzhi Song, Jiexin Zhang, Sahil Seth, Chi-Wan Chow, Yu Cao, Curtis Gumbs, Kathryn A. Gold, Neda Kalhor, Latasha Little, Harshad Mahadeshwar, Cesar Moran, Alexei Protopopov, Huandong Sun, Jiabin Tang, Xifeng Wu, Yuanqing Ye, William N. William, Jack J. Lee, John V. Heymach, Waun Ki Hong, Stephen Swisher, Ignacio I. Wistuba, P. Andrew Futreal

Correspondence to: AFutreal@mdanderson.org (A.F.)

This PDF file includes:

Patients and Methods

Patients. We collected multi-region samples from 11 surgically resected lung adenocarcinomas including 8 stage I, 2 stage II (270 and 4990) and one stage III (283) tumors. Tumor size ranged from 2.0 cm to 4.6 cm. None of the patients had pre-operative chemotherapy or radiation therapy. Patient 283, who had N2 disease, received post-operative chemotherapy followed by radiation therapy. Patient 270, who had N1 disease, received post-operative chemotherapy. The other patient with N1 disease 4990 did not receive adjuvant chemotherapy due to comorbidities. Patient 324 received post-operative radiation because of positive surgical margins. The remaining 8 patients did not receive therapy in the adjuvant setting. Among these 11 patients, 3 were never smokers, 7 were former smokers and one patient was a current smoker. With a median follow-up of 21 months post-surgery, 3 patients (270, 330 and 4990) have had disease relapse. Please see **Table S1** for relevant clinical information. Collection and use of patient samples were approved by institutional review board (IRB) of University of Texas MD Anderson Cancer Center. Informed consent was obtained from all patients.

Sample collection and processing. Immediately after resection, eight regions from each tumor, as indicated in **Fig. S6A**, representing the spatial heterogeneity of the primary tumors, were collected by using 18-gauge needle core needle sampling. An H&E slide from each tumor region (**Fig. S6B)** was reviewed by experienced lung cancer pathologists to assess the percentage of tumor versus adjacent normal tissues, and the percentage of malignant cells versus tumor non-malignant stromal (inflammatory, vascular and fibroblasts) cells. In addition, tumor cell viability has been addressed by examining the presence of necrosis in the

tissues. Only tumor regions with at least 40% viable tumor cells were selected for DNA exaction and sequencing.

Multi-region WES. DNA was extracted from 48 qualified regional tumor samples (5 regions per tumor from 5 patients, 4 regions per tumor from 5 patients and 3 regions per tumor from one patient) as well as from matched peripheral blood leukocytes as germ line DNA control. Exome capture was performed on 500ng of genomic DNA per sample based on KAPA library prep (Kapa Biosystems) using the Agilent SureSelect Human All Exon V4 kit according to the manufacturer's instructions and paired-end multiplex sequencing of samples was performed on the Illumina HiSeq 2000 sequencing platform. The average sequencing depth was 277x per sample (ranging from 138x to 437x, standard deviation +/- 67).

Variant Calling. Paired-end reads in FastQ format generated by the Illumina pipeline were aligned to the reference human genome (UCSC Genome Browser, hg19) using Burrows-Wheeler Aligner (BWA) on default settings (*31*) except a seed length of 40, maximum edit distance of 3, and maximum edit distance in the seed of 2. Aligned reads were further processed following the GATK Best Practices of duplicate removal, indel realignment, and base recalibration. Somatic single-nucleotide substitutions were detected by using MuTect (*32*). In addition to MuTect's build-in filters, the following filtering criteria were applied: (i) total read count in tumor DNA ≥ 100; (*19*) total read count in germ line DNA ≥ 50; (iii) variant present on both strands; (iv) VAF in tumor DNA ≥ 5%; (v) VAF in germ line = 0; (vi) variants in positions listed in dbSNP129 were removed. Small indels were identified using Pindel (pindel024t) (*33*). Pindel outputs were furthered by applying (i) total tumor reads > 15; (*19*) total normal reads > 6; total number of

reads supporting a call > 4; (iii) VAF in tumor >5%; (iv) VAF in normal <1%. Different filtering criteria were applied to indels to increase sensitivity. Substitutions and indel were annotated using ANNOVAR based on UCSC known genes.

Variant validation by targeted capture deep sequencing. To validate variants identified by WES, we designed a pool of customized oligonucleotides using Agilent SureDesign for targeted capture enrichment and deep sequencing. Paired-end multiplex sequencing was applied to all samples on the Illumina HiSeq 2000 sequencing platform to an average sequencing depth of 863x (ranging from 456x to 1474x, sd +/- 231). The matching normal of case 317 failed capture, and the germ line exome data was used as reference for the targeted capture. Mutect and Pindel were applied to identify nucleotide substitutions and indels. Somatic variants were considered validated if they met the following criteria: (i) The same alteration at the same position was observed in the same tumor by both WES and deep sequencing; (*19*) VAF in tumor DNA was ≥1% of sequencing reads; (iii) VAF in germ line DNA was <1% of sequencing reads. We chose the threshold of 1% based on the coverage depth and error rate (*34, 35*).

Phylogenetic analysis. Mutation profiles were converted into binary format with 1 being mutated and 0 otherwise. For each patient, all validated mutations that were present in at least one tumor region were included. Ancestors were germ line DNA assuming with no mutations. Multistate discrete-characters Wagner parsimony method in PHYLIP (Phylogeny Inference Package) was used to generate phylogenic tree (*36-38*). Phylogenetic trees were redrawn in Adobe Illustrator with relative trunk and branch lengths proportional to the number of shared and distinct mutations on the corresponding trunk or branch.

Detection of copy number aberrations. Copy number data were derived from WES reads using an in house R package. Read counts in each exon region were determined using bedtools (*39*). The log2 ratios of tumor versus normal reads were then calculated for each tumor region after adjusting for the total mapped reads in that tumor region. Similar approach has been taken to estimate copy number status at base level (*40*). The log ratios were subjected to segmentation using the DNAcopy package of Bioconductor. Cancer genes known to be affected by amplification or homozygous deletion (*22*) were analyzed by comparing the coverage of segments containing candidate genes to the average coverage across the exome, after normalization using matched germ line exome sequencing data. A threshold of log2 ratio >1 or <-1 was used to screen for amplifications or deletions respectively. Manual inspection was applied to review all segments containing candidate genes in each tumor region to make amplification and deletion calls.

APOBEC mutation signature analysis. APOBEC mutation signatures were analysed as previously described (*41*). In brief, APOBEC signature enrichment E_{TCW} relating to the strength of mutagenesis at the TCW (where W is either A or T) motif was calculated as follows:

$$E_{TCT} = \frac{mutations_{TCW} \times context_{CorG}}{mutations_{CorG} \times context_{TCW}}$$

where $mutations_{TCW}$ is the number of mutated cytosines (and guanines) in a TCW (or WGA) motif, $mutations_{C\ (or\ G)}$ is the total number of mutated cytosines (or guanines), $context_{TCW}$ is the total number of TCW (or WGA) motifs within a 41-nucleotides region centered on the mutated cytosines (and guanines) and $context_{C\ (or\ G)}$ is the total number of cytosines (or guanines) within the 41-nucleotides region centered on the mutated cytosines (or guanines). Only TCW to TTW

or TGW, WGA to WAA or WCA, C to T or G, and G to A or C nucleotide subsitutions were included for this analysis. Over-representation of APOBEC mutation signature was determined using a two-sided Fisher's exact test comparing the ratio of the number of cytosine-to-thymine or cytosine-to-guanine substitutions and guanine-to-adenine or guanine-to-cytosine substitutions that occurred in and out of the APOBEC target motif (TCW or WGA) to an analogous ratio for all cytosines and guanines that reside inside and outside of the TCW or WGA motif within 41-nucleotide region centered on the mutation cytosine (and guanine). APOBEC mutation signature enrichment was determined for all mutations, trunk mutations and non-trunk mutations separately.

Subclonal analysis using ABSOLUTE. ABSOLUTE algorithm was applied to all validated mutations of each tumor region to estimate sample purity, ploidy and to infer cancer cell fractions of each mutation as described previously (*27*). Mutations were classified as clonal based on the posterior probability that the cancer cell fraction exceeded 0.95 and subclonal otherwise. We then combined all the sequencing reads from all regions of the same tumors and applied ABSOLUTE to the combined data to assess the distributions of clonal and subclonal mutations at the whole tumor level.

Subclonal analysis using Dirichlet process. For Dirichlet process (*13, 42*), tumor purity, ploidy and segmented log2 ratio values were used to estimate the copy number of the tumor DNA at the locus of each mutation, using equations derived in ASCAT (*43*). These copy numbers were then used as inputs to the Dirichlet process, allowing the conversion of allele frequencies to mutation copy numbers, as described previously (*42, 44*). A mutation

copy number of 1 indicates a mutation that is present in all cells of a tumor on one chromosome. Mutation copy numbers above 1 indicate mutations on chromosomes that have been gained after acquiring the point mutation, while mutation copy numbers below 1 indicate mutations that are present subclonally, i.e. only in a fraction of tumor cells.

Statistical analyses. ANOVA test was used to assess the association between mutation burden and age, gender, tumor size, lymph node status, or smoking status in each patient. A Fisher exact test was used to assess the significance of different mutation spectrum between trunk mutations and non-trunk mutations. A t-test was used to assess the association between percent trunk mutations and disease relapse in the phylogenetic analysis, and the association between disease relapse and percent subclonal mutations determined by the ABSOLUTE algorithms or Dirichlet processes. To determine the correlation of copy number changes between tumor regions, segment data were processed using the CNTools package of Bioconductor to generate a gene by tumor region copy number matrix. Correlations between tumor regions were then calculated to obtain the correlation coefficients.

Supplementary Figure Legends:

Fig. S1. Mutation burden of different lung adenocarcinomas. All validated mutations are included. The average numbers of mutations per region from each individual tumor are shown and error bars represent standard deviation of number of mutations in different regions of individual tumors.

Fig. S2. Copy number aberrations. (**A)** Copy number aberrations of each tumor region derived from WES. Y axis represents log2 ratios of tumor counts versus normal counts and x axis shows chromosomal numbers. **(B)** Correlation of copy number aberrations between 48 tumor regions. Heat map represents correlation coefficients R2 of log2 ratios of sequencing counts derived from WES in each tumor region to its matching germ line DNA.

Fig. S3. APOBEC mutation signatures. Percent mutations with APOBEC signatures (C>T/G substitution in TCW, where W is A or T) are shown for trunk (green), non-trunk (orange) and all (blue) mutations from each patient.

Fig. S4. Intra-regional heterogeneity. (A) VAF profiles of validated mutations from 48 tumor regions. VAF data were derived from WES counts. Individual mutations, arranged based on the number of regions in which the mutations were detected and their VAF, are shown on the *x* axis, and VAF are plotted on the *y* axis. (B) Intra-regional heterogeneity of 48 tumor regions by ABSOLUTE analysis. Copy number data and VAF data of all validated mutations were derived from WES. ABSOLUTE was applied to each tumor region.

Fig. S5. Distribution of clonal and subclonal mutations of 11 lung adenocarcinomas by ABSOLUTE analysis (A) and Dirichlet process (B)**.** Copy number and VAF data of all validated mutations derived from WES counts were combined from all regions of the same tumors. ABSOLUTE or Dirichlet process was applied to the combined data.

Fig. S6. Multi-region sampling. (A). Diagram of multiregional sampling. (B) Representative HE images from each tumor.