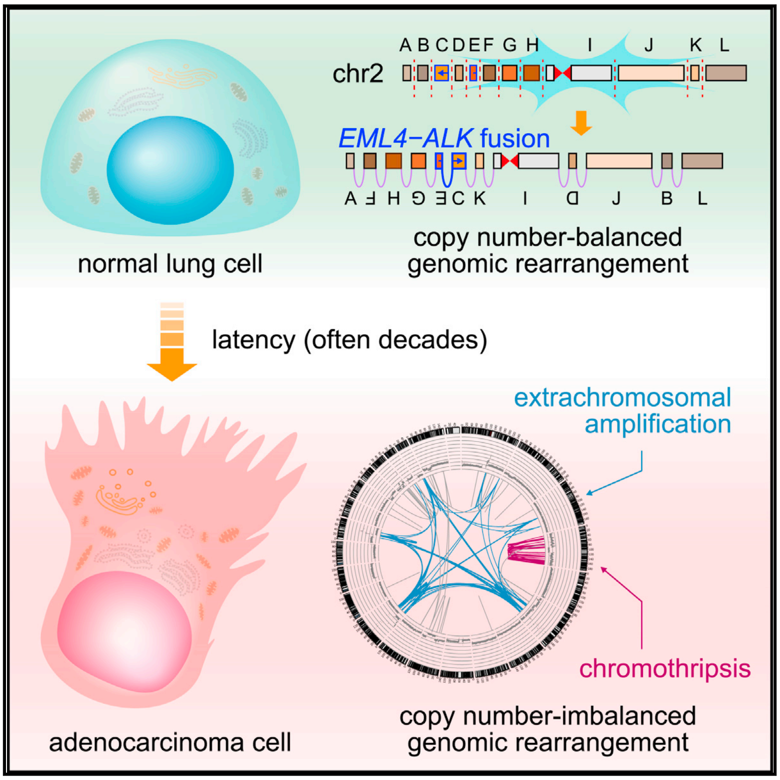
Cell

Article

Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma

Graphical Abstract



Highlights

- Driver fusion oncogenes in LADCs are generated from complex genomic rearrangements
- These rearrangements are frequently copy-number balanced, resembling germline events
- Fusions often arise in early decades of life, leaving long latency to diagnosis
- *SETD2* inactivation is cooperative with fusion oncogenes in *TP53*-wild-type LADCs

Authors

Jake June-Koo Lee, Seongyeol Park, Hansol Park, ..., Peter J. Park, Young Seok Ju, Young Tae Kim

Correspondence

ysju@kaist.ac.kr (Y.S.J.),
ytkim@snu.ac.kr (Y.T.K.)

In Brief

Driver fusion oncogenes in human lung adenocarcinoma of non-smokers are generated from complex genomic rearrangements and often arise in early decades of life, long before diagnosable disease.

Lee et al., 2019, Cell *177*, 1842–1857
June 13, 2019 © 2019 Elsevier Inc.
https://doi.org/10.1016/j.cell.2019.05.013

CellPress

Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma

Jake June-Koo Lee,[1,2,3,11] Seongyeol Park,[1,11] Hansol Park,[4] Sehui Kim,[5] Jongkeun Lee,[6] Junehawk Lee,[7] Jeonghwan Youk,[1] Kijong Yi,[1] Yohan An,[4] In Kyu Park,[8] Chang Hyun Kang,[8] Doo Hyun Chung,[5] Tae Min Kim,[9,10] Yoon Kyung Jeon,[5,9] Dongwan Hong,[6] Peter J. Park,[2,3] Young Seok Ju,[1,4,12,*] and Young Tae Kim[8,9,*]

[1]Graduate School of Medical Science and Engineering, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, United States
[3]Ludwig Center at Harvard, Harvard Medical School, Boston, MA 02115, United States
[4]Biomedical Science and Engineering Interdisciplinary Program, Korea Advanced Institute of Science and Technology, Daejeon 34141, Korea
[5]Department of Pathology, Seoul National University College of Medicine, Seoul 03080, Korea
[6]Clinical Genomics Analysis Branch, National Cancer Center, Goyang 10408, Korea
[7]Korea Institute of Science and Technology Information, Daejeon 34141, Korea
[8]Department of Thoracic and Cardiovascular Surgery, Seoul National University Hospital, Seoul 03080, Korea
[9]Seoul National University Cancer Research Institute, Seoul 03080, Korea
[10]Department of Internal Medicine, Seoul National University Hospital, Seoul 03080, Korea
[11]These authors contributed equally
[12]Lead Contact
*Correspondence: ysju@kaist.ac.kr (Y.S.J.), ytkim@snu.ac.kr (Y.T.K.)

https://doi.org/10.1016/j.cell.2019.05.013

SUMMARY

Mutational processes giving rise to lung adenocarcinomas (LADCs) in non-smokers remain elusive. We analyzed 138 LADC whole genomes, including 83 cases with minimal contribution of smoking-associated mutational signature. Genomic rearrangements were not correlated with smoking-associated mutations and frequently served as driver events of smoking-signature-low LADCs. Complex genomic rearrangements, including chromothripsis and chromoplexy, generated 74% of known fusion oncogenes, including *EML4-ALK*, *CD74-ROS1*, and *KIF5B-RET*. Unlike other collateral rearrangements, these fusion-oncogene-associated rearrangements were frequently copy-number-balanced, representing a genomic signature of early oncogenesis. Analysis of mutation timing revealed that fusions and point mutations of canonical oncogenes were often acquired in the early decades of life. During a long latency, cancer-related genes were disrupted or amplified by complex rearrangements. The genomic landscape was different between subgroups—*EGFR*-mutant LADCs had frequent whole-genome duplications with p53 mutations, whereas fusion-oncogene-driven LADCs had frequent *SETD2* mutations. Our study highlights LADC oncogenesis driven by endogenous mutational processes.

INTRODUCTION

Lung adenocarcinoma (LADC) is the most common type of lung cancer, which is the leading cause of cancer-related death worldwide (Herbst et al., 2018). A large proportion of LADCs are attributed to chronic tobacco smoking. By generating genome-wide base substitutions that often target cancer-related genes (e.g., *KRAS* and *TP53*) and inducing global epigenetic modifications in the airway epithelial cells, tobacco smoking directly causes LADCs (Alexandrov et al., 2013; Vaz et al., 2017). In contrast, about 25% of lung cancers develop in individuals who have not been smoking, and most of them are LADCs (Sun et al., 2007). LADCs of non-smokers have been associated with female sex and Asian ethnicity. Although studies have postulated various environmental factors—including second-hand smoke, radon, air pollution, household coal use, and occupational carcinogens (Sun et al., 2007)—the etiology and oncogenesis of LADCs in non-smokers are still enigmatic.

LADCs of non-smokers present frequent genetic alterations activating several oncogenes, e.g., point mutations in *EGFR* (20%–50% of LADCs) and gene fusions involving *ALK*, *ROS1*, and *RET* (>10% of all LADCs) (Herbst et al., 2018). Previous studies confirmed transforming and tumorigenic activities of these oncogene alterations in experimental models (Jackson et al., 2001; Ji et al., 2006; Soda et al., 2007), and the tumors were dependent on downstream signaling of these oncogenes for their growth and survival. Pharmacologic inhibition of corresponding kinases in patients with LADCs harboring these oncogene alterations has revolutionized treatment in advanced stages (Mok et al., 2009; Shaw et al., 2013a, 2014). All these findings point to the central role of these oncogene alterations in LADCs of non-smokers. However, the mutational processes generating these oncogenic alterations and the timing of their acquisition have not been fully characterized.

Whole-genome sequencing (WGS) provides a unique opportunity to explore the mutational processes operating in cancer genomes. Genome-wide patterns of base substitutions, termed mutational signatures, have provided insights into the mutagenic processes (Alexandrov et al., 2013), and observations of unique patterns of genomic structural variation have led to the discovery of catastrophic rearrangement processes, including chromothripsis (Stephens et al., 2011) and chromoplexy (Baca et al., 2013). Furthermore, integrative analyses of point mutations and copy-number alterations have deciphered the temporal sequence of somatic mutations including key oncogenic drivers in individual cancers (Mitchell et al., 2018; Gerstung et al., 2017). Although there have been landmark studies characterizing the genetic alteration landscape and evolutionary history of LADCs (Cancer Genome Atlas Research Network, 2014; Jamal-Hanjani et al., 2017), they were largely based on exome and transcriptome sequencing and put more emphasis on smoking-related LADCs.

In this study, we investigate the mutational processes operating in LADCs, with a focus on cases with minimal genomic evidence of smoking-associated mutagenesis, especially those driven by fusion oncogenes. Through a case-by-case characterization of 138 LADCs, our study deciphers the mutational processes and the timing of complex genomic rearrangements that lead to the development of LADCs.

RESULTS

Genome-wide Mutational Profile of Lung Adenocarcinomas

Our study cohort is comprised of 138 LADCs (Figure 1A), including 49 newly sequenced and 89 publicly available cases, largely from three previous studies (Imielinski et al., 2012; Cancer Genome Atlas Research Network, 2014; Wu et al., 2015) and others (Table S1). All tumors were surgically resected, histologically confirmed lung adenocarcinomas, or its related histology (pathology findings of newly sequenced cases are available in Table S1). The cohort includes 39 LADCs with known fusion oncogenes. Twenty-eight of these are newly sequenced samples, in which a fusion oncogene was detected in a screening of 350 surgically resected, treatment-naive LADCs (Figure S1A; STAR Methods).

We analyzed whole-genome sequences of 138 LADCs in a unified pipeline (median sequencing coverages of 42X and 32X for tumor and matched normal tissues, respectively; Table S1; STAR Methods). In 79 out of 138 LADCs, RNA-sequencing (RNA-seq) datasets were also combined (newly sequenced: n = 34, publicly available: n = 45). We found high-confidence somatic mutations including 4,136,995 base substitutions and 186,170 indels (STAR Methods).

In addition, we identified 21,420 high-confidence somatic genomic rearrangements with a wide variation in quantity between tumors (median = 117, range = 3−867). A substantial fraction of the rearrangements were linked with at least two other rearrangements, indicating that they were a part of a complex genomic rearrangement. These rearrangements were systematically clustered based on their spatial proximity (STAR Methods). We identified 893 complex rearrangement clusters accounting for 77% (16,585 out of 21,420) of the genomic rearrangements. Most tumors contained multiple clusters (median = 5, range = 0−32), each harboring 3 to 700 rearrangements (median = 5). All the variant information is summarized and publicly available online (http://genome.kaist.ac.kr).

Mutational Processes in Lung Adenocarcinomas

To characterize the mutational processes generating point mutations, we first analyzed mutational signatures for the 138 LADCs based on the COSMIC catalog (Table S2; Alexandrov et al., 2015). Among 131 tumors with available smoking history, the vast majority of LADCs of clinical never-smokers (n = 58) showed minimal contribution of mutational signature 4 (a C:G>A:T-dominant signature, related to exposure to smoking carcinogen), which was consistent with their clinical history (Figure S1B). However, the LADCs of ever-smokers (n = 73) exhibited variable contribution of this mutational signature to their mutational profiles. This variation was not fully explained by the amount of smoking exposure (in terms of pack-years), nor by other clinical variables (Figure S1B). Many smoker LADCs showed strong activity of signature 4. However, about one-third of smoker LADCs (24 of 73) displayed minor to no contribution of signature 4 despite their smoking history (median 25 pack-years; range = 1.5−80), indicating that smoking-associated mutagenesis minimally contributed to their mutational history. These tumors were largely the LADCs harboring activating mutations or fusions of canonical oncogenes (21 of 24), and their genomic characteristics were highly similar to LADCs of clinical never-smokers (Figure S1C). Therefore, we used mutational signature-based criteria rather than clinical records in classifying the LADCs for downstream analysis. This classification clustered our cohort into two groups, based on the count and the relative burden of signature 4 mutations (by k-means clustering; Figure 1B): 55 signature 4-high (S4-high) and 83 signature 4-low (S4-low) LADCs (Figures 1B–1D; STAR Methods).

S4-high LADCs were associated with older age of diagnosis (65 versus 58 years, $p = 0.0003$, Student's t test), had a higher proportion of male patients (72% versus 37%, $p < 0.0001$, Fisher's exact test), carried more mutations in *KRAS* (n = 18 versus 4), and showed a base substitution burden >10 times greater than S4-low LADCs (24.01/Mbp versus 1.89/Mbp, $p < 0.0001$). Number of indels ($r = 0.83$, $p < 0.0001$) was strongly correlated with the number of signature 4 mutations (Figure S1D), indicating that this could be a part of the smoking-related mutational signature.

S4-low LADCs (n = 83) included LADCs harboring activating mutations of *EGFR* (n = 28), *KRAS* (n = 4; G12D and G12A; both from non-C:G>A:T mutations), *ERBB2* (n = 2), and *MET* (n = 2; exon 14 splice site mutations), and all LADCs with known fusion oncogenes involving *ALK*, *ROS1*, *RET*, and others (n = 39) (Figure 1A). Their mutational spectra (Figure 1D) were predominantly explained by two endogenous mutational signatures: signatures 5 (median = 58%) and 1 (17%). These signatures are known to have clock-like properties and have been observed in most germline and normal somatic cells (Alexandrov et al., 2015).

In contrast to base substitutions and indels, the number of genomic rearrangements was not significantly different between S4-high and -low LADCs (171 versus 145, $p = 0.2665$; Figure 1E). Although the number of simple deletions was positively correlated with the number of signature 4 mutations ($r = 0.42$, $p < 0.0001$, Pearson's correlation), no other types of simple rearrangements were correlated as such (Figure S1D). Furthermore, the number of breakpoints in complex rearrangement clusters

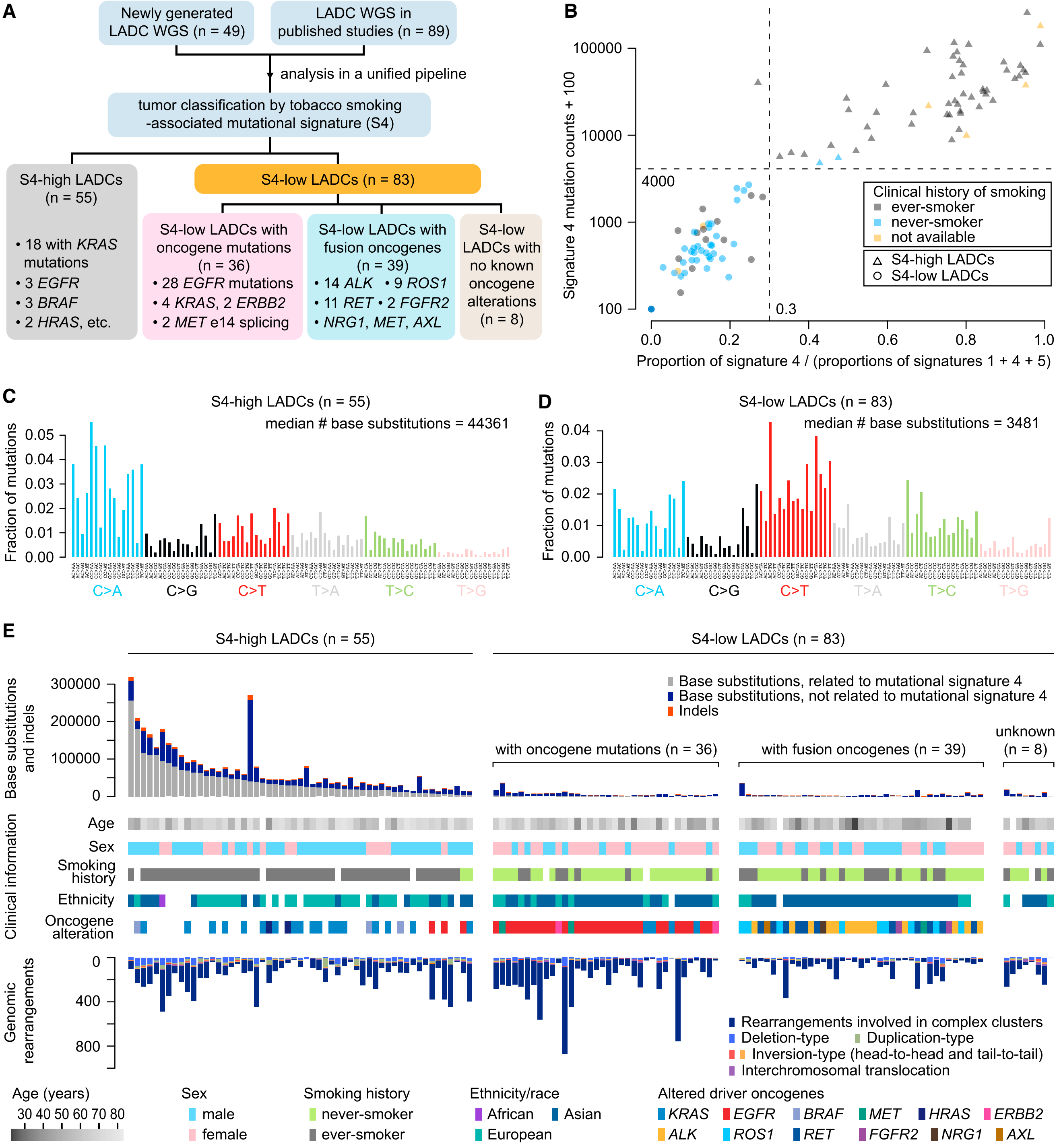


Figure 1. Mutational Processes of Lung Adenocarcinoma

(A) Composition of the study cohort and classification of LADCs into subgroups.

(B) K-means clustering separated the patient cohort (n = 138) into S4-high (n = 55) and S4-low (n = 83) LADCs (STAR Methods).

(C and D) Representative mutational spectra for the S4-high (C) and the S4-low (D) LADCs, calculated from the median values of fraction of mutations in 96-trinucleotide contexts.

(E) Clinical, demographic, and genomic characteristics of LADCs in this study. Complex genomic rearrangements were defined from clustering, and the remaining rearrangement events were classified by chromosomal positions and orientations of read pairs (STAR Methods).

See also Figure S1 and Tables S1 and S2.

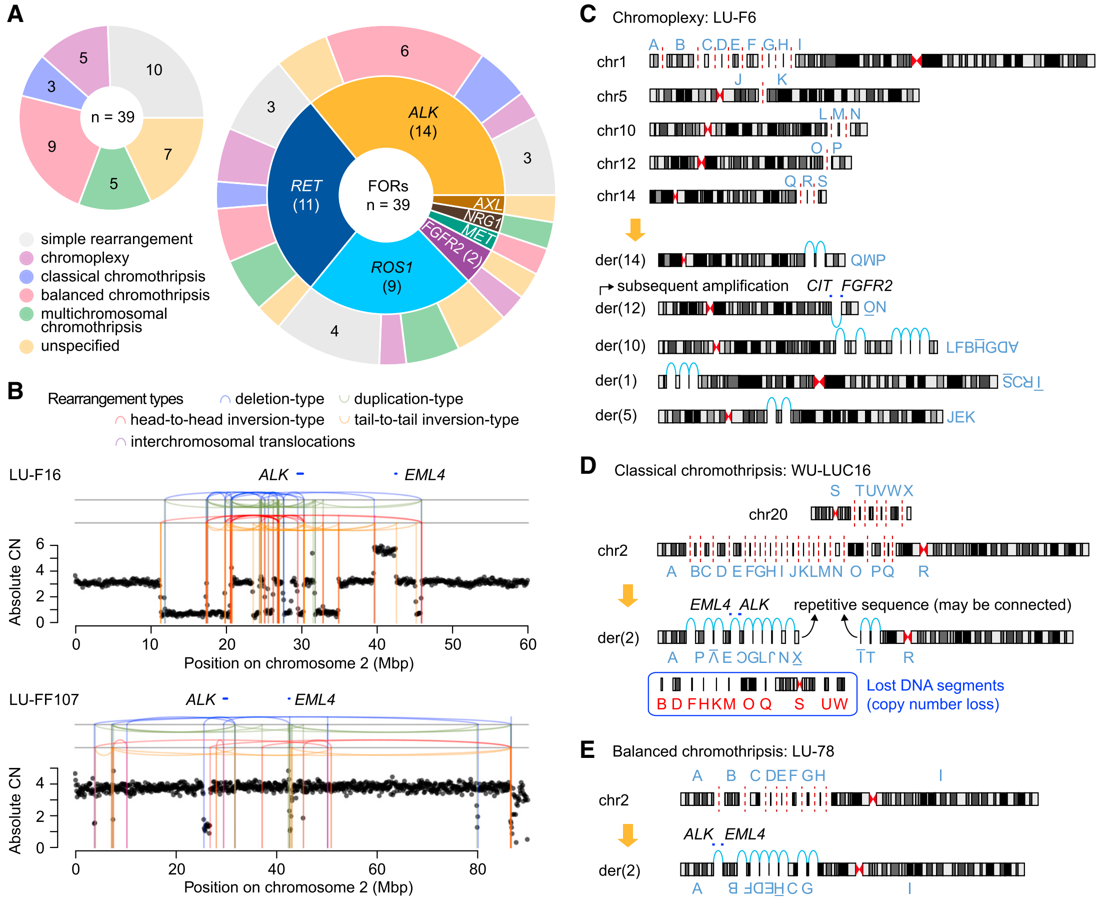


Figure 2. Complex Genomic Rearrangements Generating the Fusion Oncogenes

(A) Patterns of genomic rearrangements underlying the LADC fusion oncogenes.

(B) Examples of chromothripsis generating *EML4-ALK* fusion oncogenes. In LU-F16 (top), many DNA fragments are lost during the reassembly process after chromosomal shattering, and the final footprint shows a typical copy-number imbalance. In contrast, LU-FF107 (bottom) shows a copy-number-balanced pattern. Rearrangements are described as vertical lines and connecting arcs, whose colors indicate different types of rearrangements (the colors are consistently used throughout the manuscript). CN, copy number.

(C–E) An example of chromoplexy (C), classical chromothripsis (D), and balanced chromothripsis (E) generating the driver fusion oncogenes of LADCs. Capital letters indicate DNA fragments and their orientations correspond to their rearranged configuration in derivative chromosome. Red dashed lines indicate DNA double-strand breaks, and the light blue arcs indicate ligated breakpoints.

See also Figure S2.

was also uncorrelated with the burden of signature 4 mutations (r = 0.04, p = 0.6709). Instead, the number of rearrangements was different between the molecular subgroups (defined by major driver oncogenes) within the S4-low LADCs: a significantly larger burden of rearrangements in the LADCs driven by activating base substitutions or indels in canonical oncogenes (*EGFR*, *KRAS*, *ERBB2*, and *MET*; n = 36; hereafter we refer this group to as S4-low LADCs with oncogene mutations) compared to the LADCs driven by fusion oncogenes (n = 39) (211 versus 87; p = 0.0009; Figure S1E). This suggests that the rearrangement processes are not stochastic and could be related to the biological properties of driving oncogenes.

Although the ethnic compositions were different between the S4-high (largely patients of European ancestry) and the S4-low LADCs (mostly East Asian patients), we found no meaningful difference in the genomic landscapes of the two major ethnicities within the LADC subgroups, in considering point mutation burden, rearrangements, and copy-number variations (Figure S1F).

Complex Genomic Rearrangements Generating the Fusion Oncogenes

We next focused on the driver fusion oncogene-generating rearrangements (FORs) in 39 S4-low LADCs with fusion oncogenes (one in each). Robust in-frame transcription of these fusion oncogenes was confirmed by RNA-seq analysis in all available cases (n = 26; Figure S2A). In 10 LADCs (26%), the FORs were simple rearrangements (Figures 2A and S2B): large deletions (n = 2; *SFTPB-ALK* and *EZR-ROS1*), reciprocal inversions (n = 5; *EML4-ALK*, *KIF5B-RET*, and *CCDC6-RET*) and reciprocal translocations (n = 3; *CD74-ROS1* and *CCDC6-ROS1*). In contrast, in 29 LADCs (74%), FORs were complex (Figures 2A and 2B), involving a median of 20 rearrangement breakpoints (range = 4−281). We classified these 29 complex FORs into several groups by integrating various genomic features including copy number and breakpoint microhomology (Table S3). The FORs were also manually reconstructed to infer the structure of derivative chromosomes (Figure S2C; STAR Methods).

Among the 29 complex FORs, five had multiple breakpoints dispersed in more than two chromosomes, forming a closed chain pattern, as expected for chromoplexy (Figures 2C and S2C; Baca et al., 2013). Chromoplexy has been described as a source of fusion oncogenes in prostate cancers (Baca et al., 2013), NUT midline carcinomas (Lee et al., 2017b), and Ewing sarcomas (Anderson et al., 2018). Except for occasional imbalance due to deletion bridges and secondary amplification events (LU-F6), copy-number profiles in chromoplexy-related FORs were mostly balanced.

We also observed 3 LADCs whose fusion oncogenes were generated by classical chromothripsis (Figure 2D), a mutational process involving catastrophic chromosomal shattering followed by stochastic rejoining of the DNA segments (Stephens et al., 2011). These 3 FORs exhibited the typical features of chromothripsis: (1) a large number of clustered breakpoints (n = 144, 48, and 22); (2) even contributions of deletion-, tandem duplication-, head-to-head, and tail-to-tail inversion-type rearrangements; and (3) copy-number oscillation between two states (Figure 2B, top). Previously, chromothripsis was found to play a role in promoting oncogenesis by disrupting tumor suppressor genes or by amplifying oncogenes through double minutes (Zhang et al., 2015). Our analysis further supports its direct oncogenic role.

Interestingly, the largest proportion of complex FORs (31%; 9 out of 29 complex FORs) shared the features of both chromoplexy and chromothripsis. They were typically limited to one or two chromosomes with clustered breakpoints, similar to a classical chromothripsis event. However, their copy-number profile was predominantly balanced (Figure 2B, bottom), as seen in chromoplexy. This indicates that most of the DNA fragments were preserved and reassembled after a catastrophic chromosomal shattering (Figure 2E). This unique pattern of complex FORs, herein referred to as balanced chromothripsis, typically involved a smaller number of rearrangement breakpoints compared to classical chromothripsis (median [range]; 20 [14−59] versus 48 [22−144]), and the number was more similar to that of chromoplexy (20 [10−27]). Most breakpoints included microdeletions with short (1−3 bp) or no microhomologies, or had microduplications, implying that the DNA double-strand breaks were repaired by non-homologous end-joining (NHEJ; for microdeletions) or synthesis-dependent end-joining (for microduplications), where DNA synthesis by homologous recombination is followed by end-joining processes (Figure S2B). To the best of our knowledge, these copy-number-preserving chromothripsis patterns have been reported in germline samples (Kloosterman et al., 2011; Chiang et al., 2012; Redin et al., 2017) but not in cancer samples.

Complex FORs Involving Secondary Rearrangements

In five other cases, the FOR clusters exhibited copy-number oscillations that were typical for chromothripsis but involved multiple chromosomes (≥3). Through careful reconstruction of the complex FORs, we found that at least three of these events were explained by secondary complex rearrangements superimposed on the fusion oncogene-generating chromoplexy. For example, in LU-89, chromoplexy generated the *ATP1B1-NRG1* fusion oncogene in a derivative chromosome with two centromeres (dicentric chromosome) (Figure 3A). The two centromeres would then be pulled to opposite poles in the subsequent mitosis, forming an anaphase chromatin bridge. This would result in a rupture of the nuclear envelope and exposure of the DNA to cytoplasmic nucleases, triggering a secondary chromothripsis (Maciejowski et al., 2015). Two other FORs harboring *CD74-ROS1* in LU-FF58 and LU-SC126 were also explained by the same type of chromoplexy-induced chromothripsis (Figure S3A). In these cases (LU-89, LU-FF58, and LU-SC126), the primary chromoplexy chains involved at least 4, 2, and 9 genes, and their secondary chromothripsis events transected 8, 6, and 58 genes, respectively, including cancer-related genes such as *ARID1B* and *PARK2*. Theoretically, a large chromoplexy event could produce a dicentric chromosome involving multiple chromosomal segments (as suggested in LU-SC126 in Figure S3A), or even multiple dicentric derivative chromosomes and acentric segments. In this scenario, secondary rearrangement bursts could be triggered over extensive genomic regions and thus could have a large functional impact.

In one case (LU-F13), a highly amplified *AXL-MBIP* fusion oncogene was observed in a large (>15 Mbp) amplicon (Figure 3B). Overexpression of the fusion oncogene was confirmed in our reanalysis of RNA-seq from the same tumor (Seo et al., 2012). Six rearrangement breakpoints in chromosomes 14 and 19 were co-amplified with the amplicon, representing the ancestral structure of double minutes. Numerous other rearrangements in the amplicon were supported by smaller numbers of reads, indicating that those rearrangements were accumulated through or after the amplification cycles (Turner et al., 2017). This amplicon was inserted into the pericentromeric region of chromosome 21 (Figure 3B). Interestingly, very few base substitutions were co-amplified with the *AXL-MBIP* fusion, which implies that the fusion oncogene was acquired early in the mutational history of the cancer cell (described later in more detail).

In 6 LADCs, the complex FORs involved a small number of rearrangements (4−12 breakpoints), which made it difficult to infer their generative mechanisms (Mehine et al., 2013). One of them (TCGA-67-6215) showed a balanced, intrachromosomal complex rearrangement, while another (LU-51) exhibited segmental losses (Figure S3B).

Taken together, diverse types of complex rearrangements underlie the formation of driver fusion oncogenes in LADCs. Although their patterns are different, they commonly involve multiple, simultaneous DNA double-strand breaks in somatic cells. Until now, various cellular processes (chromosomal mis-segregation, telomere crisis, and programmed DNA breaks at transcriptional hubs, etc.) have been suggested to explain the chromosomal catastrophe (Zhang et al., 2015; Maciejowski et al., 2015; Haffner et al., 2010). Our findings suggest that these cellular conditions may serve as initiating events of malignant transformation in normal airway epithelial cells, especially in non-smokers.

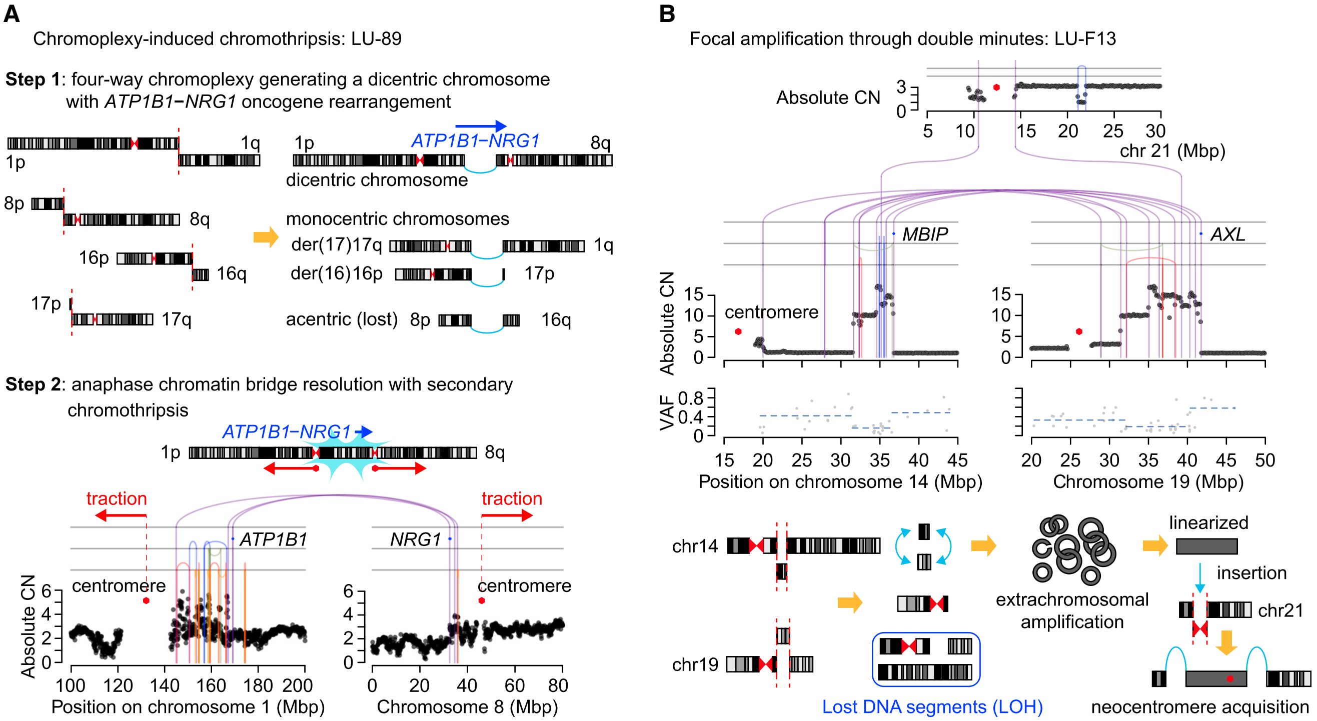

A
Chromoplexy-induced chromothripsis: LU-89
Step 1: four-way chromoplexy generating a dicentric chromosome with ATP1B1–NRG1 oncogene rearrangement
1p
1q
8p
8q
16p
16q
17p
17q
1p
ATP1B1–NRG1
8q
dicentric chromosome
monocentric chromosomes
der(17)17q
1q
der(16)16p
17p
acentric (lost) 8p
16q
Step 2: anaphase chromatin bridge resolution with secondary chromothripsis
ATP1B1–NRG1
1p
8q
traction
traction
ATP1B1
NRG1
centromere
centromere
Absolute CN
Position on chromosome 1 (Mbp)
Chromosome 8 (Mbp)
B
Focal amplification through double minutes: LU-F13
Absolute CN
chr 21 (Mbp)
MBIP
AXL
centromere
Absolute CN
VAF
Position on chromosome 14 (Mbp)
Chromosome 19 (Mbp)
chr14
chr19
extrachromosomal amplification
linearized
insertion
chr21
Lost DNA segments (LOH)
neocentromere acquisition

Figure 3. Complex FORs Involving Secondary Rearrangements

(A) Schematic description of 4-way chromoplexy generating the *ATP1B1-NRG1* fusion oncogene and its derivative chromosomes.

(B) A complex rearrangement generating amplified *AXL-MBIP* oncogene and the base substitutions residing in the amplicon (top). Dashed blue lines indicate the average variant allele fractions in chromosomal segments. Explanation model for the observed complex rearrangement is shown (bottom). VAF, variant allele fraction.

See also Figure S3.

Balanced DNA Copy Numbers of Fusion Oncogene Rearrangements

In addition to the 28 complex FORs, we observed 864 complex genomic rearrangements that did not generate the known fusion oncogenes (referred to as collateral rearrangements [CRs]) in our 138 LADCs. We questioned whether the FORs showed distinct genomic features compared to the CRs (Figure S4). Due to the difficulties of inferring the generative mechanism of complex rearrangements with a small number of breakpoints (<10), we limited this analysis to large complex clusters harboring 5 or more rearrangements (n = 482; Table S3).

Most noticeably, the copy-number-balanced rearrangements (balanced chromothripsis and chromoplexy) were nearly specific to the FORs. The complex rearrangement clusters with a high proportion of balanced breakpoints (≥ 0.4; STAR Methods) comprised only 6% (31 of 482) of all complex clusters, but they were strikingly enriched in FORs (13 out of 25; odds ratio = 25.89, $p < 0.0001$, Fisher's exact test) (Figure 4A). This odds ratio increases if we include the primary chromoplexy events in three cases of chromoplexy-induced chromothripsis (indicated in Figure 4A). A similar pattern was also observed among the simple rearrangements: we identified 51 balanced reciprocal inversions and translocations in our cohort (STAR Methods), and 8 (16%) of them were FORs, 156 times more common than expected by chance ($p < 0.0001$, by exact binomial test).

The complex FORs and the CRs were also distinguished by other genomic features, consistent with distinct generative mechanisms. For example, breakpoints with blunt-end ligation features (indicating NHEJ) were significantly more common in FORs than in CRs (odds ratio = 1.269; p = 0.0081), whereas microhomology or short-nucleotide insertion were less frequent in FORs than in CRs (odds ratio = 0.788; p = 0.0074) (Figure 4B). Moreover, the complex FORs were significantly enriched in DNase-I-hypersensitive and early replicating regions (Figure 4C).

These contrasting genomic features between the complex FORs and the CRs may be related to differences in selective pressure. Because the acquisition of fusion oncogenes is frequently the initiating event in oncogenesis (Mitelman et al., 2007), in principle, it is likely to occur in phenotypically normal cells. In these cells, a complex rearrangement with substantial DNA copy-number losses would induce cell cycle arrest and the cell would be subject to negative selection (Santaguida et al., 2017). This may explain the overrepresentation of balanced rearrangements including balanced chromothripsis and chromoplexy among the FORs. The structural similarity of complex FORs and germline complex rearrangements (Kloosterman et al., 2011; Chiang et al., 2012; Redin et al., 2017) further supports their common occurrence and possible selection in normal cells. Alternatively, the overrepresentation of balanced rearrangements could also be driven by a lower chance of generating functional fusion genes by copy-number-imbalanced

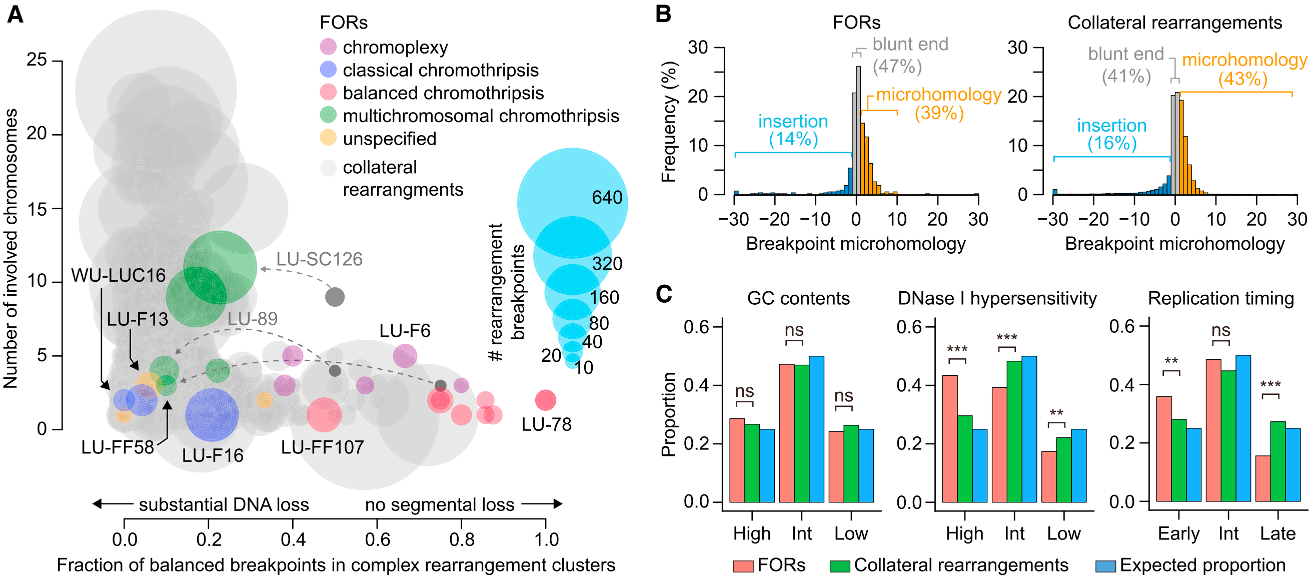


Figure 4. Genomic Features of FORs and CRs

(A) Classification of 482 complex genomic rearrangement clusters in 138 LADCs. The colored circles indicate the FORs, and the gray circles indicate the CRs. Size of the circle corresponds to the number of involved rearrangements, and the FORs with different genomic patterns are marked with different colors. Three dark-gray circles and related dashed arrows indicate the primary FORs and their secondary rearrangements into final footprints (colored circles). Identifiers of tumors that are described in the main text are annotated. FOR, fusion-oncogene-generating rearrangement.

(B) Breakpoint signatures of FORs (n = 25) and CRs (n = 457). The length of microhomology and inserted segments are plotted for complex rearrangements involving ≥5 rearrangements.

(C) Distribution of rearrangement breakpoints of the FORs (n = 25) and the CRs (n = 457) involving ≥5 rearrangement events and their relationship with three epigenomic markers from ENCODE dataset (STAR Methods). ns, not significant; Int, intermediate; ***p < 0.000001, **p < 0.001.

See also Figure S4 and Table S3.

complex rearrangements, which are more likely to disrupt the reading frame of genes.

Timing of Early Oncogenic Drivers in Mutational History of LADCs

To determine the timing of the fusion oncogenes and activating mutations in major oncogenes (*EGFR*, *KRAS*, *BRAF*, and *MET*), we conducted mutational timing analysis by integrating the variant allele fractions of point mutations and genomic rearrangements and the allele-specific copy numbers. Our approach, based on a previous method (Mitchell et al., 2018), estimates the timing of chromosomal copy-number gains or large focal amplifications (>10 Mbp) by measuring the burden of co-amplified somatic substitutions (mostly passengers) that were acquired before the copy-number gains (STAR Methods; Figure S5A). For the driver mutations acquired before the copy gain of the residing chromosomal segments (thus co-amplified as well), this analysis indicates the latest possible timing of the mutation or the fusion oncogene formation.

First, we assessed the temporal relationship between the oncogene alterations (mutations and fusions) and their copy-number amplifications in 74 informative tumors with at least one copy gain of the oncogene locus. In these tumors, the oncogene mutations and fusions were all clonal events (present in 100% of cancer cells; Table S4) and were mostly acquired before the amplifications (64 out of 74; 86%). For example, 24 out of 29 EGFR-activating mutations (those 5 were post-amplification events or in minor alleles), 12 out of 13 *KRAS*, 3 *BRAF*, and 2 *MET* splice site mutations were amplified to the maximal copy-number states of the amplified alleles, indicating that the mutations preceded the amplifications. We also confirmed that 21 out of 23 amplified fusion oncogenes predated the copy-number gains (Figure S5B). In one (LU-78) of the two exceptional cases, a very early whole-genome duplication (WGD) preceded or coincided with the fusion oncogene formation (Figure S5C).

Next, we studied the features of pre-amplification mutations in the genomic regions of the amplified oncogene alterations. As expected, the pre-amplification mutations in S4-high LADCs (n = 22) significantly outnumbered those in S4-low LADCs (n = 45), and their local mutational spectra showed a predominant contribution of mutational signature 4 (Figure S5D). In addition, 11 out of 13 amplified driver mutations of *KRAS* (10 of 10) and *BRAF* (1 of 3) in S4-high LADCs were from C:G>A:T base substitutions, consistent with smoking-associated mutations (Alexandrov et al., 2013). Altogether, we concluded that these oncogene mutations and their amplification occurred after the initiation of tobacco smoking in the patients' life history. However, further quantitative estimation of amplification timing was not possible due to the lack of clock-like features in the overwhelming smoking-associated mutations.

In contrast, the spectra of pre-amplification mutations in S4-low LADCs were well explained by two clock-like mutational signatures 1 and 5 (Figures S5D and S5E). Using the density of these pre-amplification mutations in 45 informative S4-low

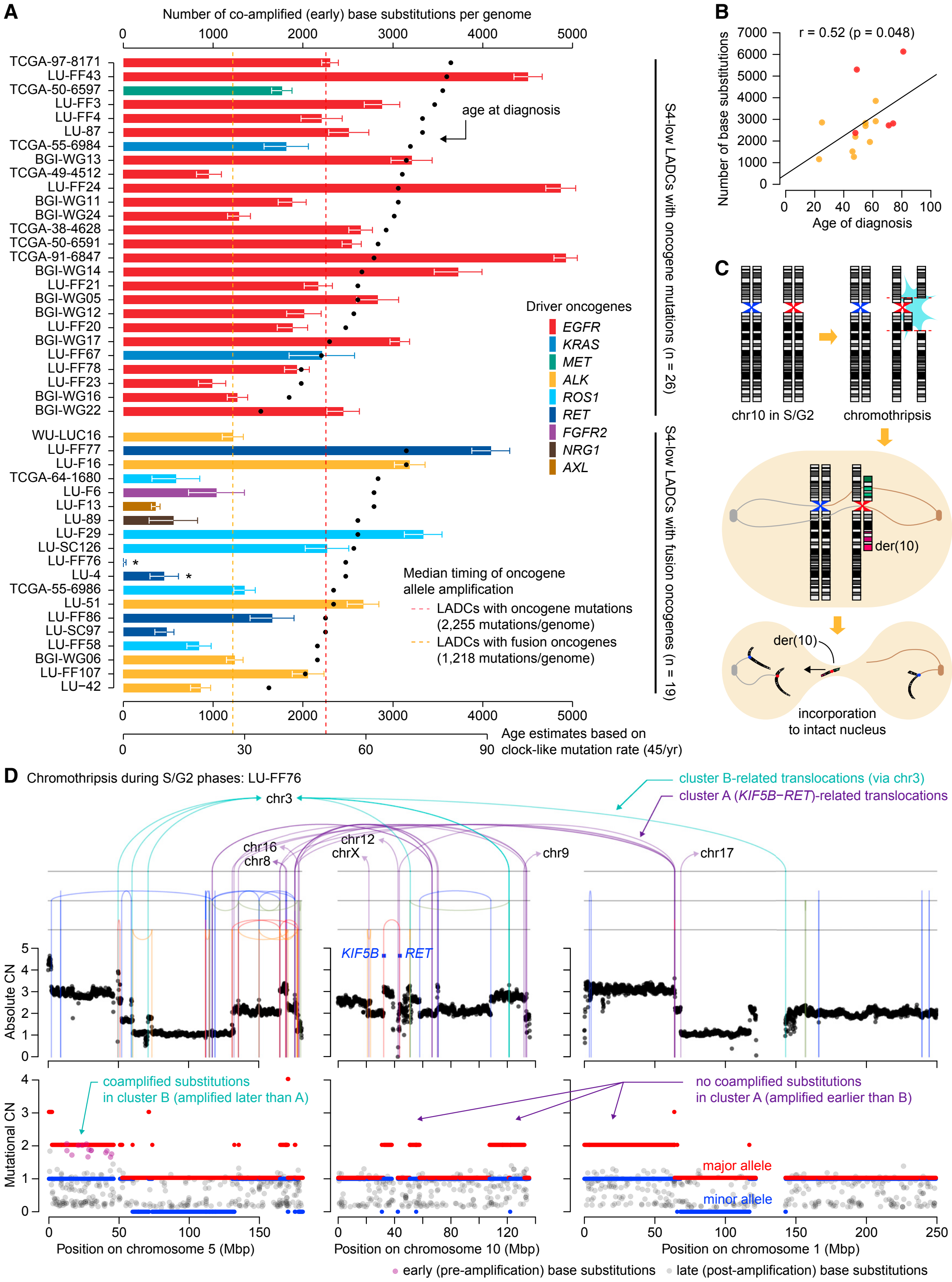


Figure 5. Amplification Timing of Early Oncogenic Drivers in Mutational History of LADCs

(A) Estimation of genome-wide mutation burden at the time points of amplification of fusions and point mutations of driver oncogenes. The 95% confidence interval was calculated from *Z* scores (STAR Methods). The age at diagnosis for each case is shown with black dots. Vertical dashed lines indicate median timing of copy-number amplifications in two subgroups. Two samples, where the timing of fusion oncogene acquisition can be pinpointed, are shown with asterisks. The number of mutations (in upper x axis) was translated into ages according to the mutation rate in (B) and plotted in parallel to the lower x axis.

(legend continued on next page)

LADCs, we identified that the amplification of fusion oncogenes occurred when the ancestral cell had a median of 1,218 substitutions, and the oncogene mutations started to amplify with a median of 2,255 substitutions (Figure 5A). From our dataset, we estimated that approximately 45 base substitutions had been accumulated per year in a set of S4-low, diploid LADCs (n = 15; r = 0.52, p = 0.048; Figure 5B) (discussed more in detail in STAR Methods). Because the mutational load of these tumors is linearly correlated with the ages of diagnosis, and their mutational spectra are explained by the clock-like signatures, the rate approximates the basal base substitution rate of the origin cells (Gerstung et al., 2017; Mitchell et al., 2018). Notably, this rate is broadly in agreement with the mutation rates observed in human normal tissues (Blokzijl et al., 2016; Yokoyama et al., 2019). If the rate is more or less constant throughout a patient's lifetime, the 1,218 mutations correspond to 27 years of age, which is almost three decades earlier than their clinical diagnosis (median = 55 years). Of note, because these time points represent the latest possible timing of fusion oncogenes, the actual timing of fusion oncogene acquisition would be even earlier.

In two cases (LU-FF76 and LU-4), we were able to determine the timing of fusion oncogene formation, because their segmental copy-number gain was coincident with the formation of the fusion oncogene (Figure 5A; shown with asterisks). In these tumors, the *KIF5B-RET* fusions were generated by chromothripsis involving multiple chromosomes after S phase (when the sister chromatids were already synthesized) and then were incorporated into an intact daughter cell harboring a full biparental set of chromosomes 10 in the next mitosis (Figure 5C). To precisely analyze the amplification timing with allele-specific copy numbers, we sequenced their genomes to a depth of 129X (LU-FF76) and 119X (LU-4). The co-amplified mutations in the derivative chromosome 10 indicated that the fusion oncogenes in these two cases were formed when the ancestral cell had accumulated ~30 and ~650 substitutions in their genomes (which correspond to only 1 and 14 years of age; Figures 5A, 5D, and S5F), which might be >40 years earlier than the clinical diagnosis in these two cases (55 years in both cases). This finding is reminiscent of acquisition of major oncogenic events in adolescent years or earlier in the origin cells of renal cell carcinomas (Mitchell et al., 2018) or in normal esophageal epithelium (Yokoyama et al., 2019).

It is well known that fusion oncogene-driven LADCs develop earlier than other types of LADCs, including those with *EGFR* mutations or smoking-related LADCs (Shaw et al., 2013b). This may reflect the different timing of the initial oncogenic driver events such as fusion oncogenes (often in early decades of life) or *KRAS* G12C mutations (after initiation of smoking), originating from different mutational processes. Furthermore, our data also indicates a long latency period from the acquisition of key oncogene mutations or fusions to the diagnosis of LADCs in S4-low LADCs. Cells harboring initial driver mutations or fusions could stay quiescent for a long time before their clinical presentation. This also implies that secondary driver events are essential for the development of LADCs.

Complex Rearrangements Establish Additional Driver Events

In contrast to the FORs, which were exclusively clonal and mostly occurred before the amplification events, the timing of complex CRs varied from early clonal events to subclonal events (Figure S6A). To elucidate the role of complex CRs in LADC oncogenesis, we further investigated their structure and their relationship with cancer-related genes.

The complex CRs varied widely in their numbers of involved rearrangement events and chromosomes and their copy-number alteration profiles (Table S3). Among them, classical chromothripsis was commonly observed: about half (64 out of 138) of the LADCs had such events (24 out of 55 in S4-high LADCs and 40 out of 68 in S4-low cases; STAR Methods). We found that those events could be further classified based on the distribution of breakpoints on chromosomes. For example, in 2 LADCs (LU-F31 and TCGA-97-8171) we observed classical chromothripsis events covering a whole chromosomal region with a large number of breakpoints (54 and 692) and a typical copy-number oscillation (Figure 6A). These cases would be well explained by catastrophic DNA damage in micronuclei after chromosomal mis-segregation (Zhang et al., 2015). In other 13 LADCs, we found a different pattern of chromothripsis, confined to two arms of different chromosomes, connected to each other by interchromosomal translocations (Figure 6B). These cases could be attributed to dicentric chromosome formation, similar to the chromoplexy-induced chromothripsis cases in FORs (Figure 3A), or to the telomere fusion-induced chromothripsis in previous studies (Rausch et al., 2012a; Maciejowski et al., 2015). Among the 64 LADCs harboring the chromothripsis CRs, the inactivation of tumor suppressor genes by chromothripsis was observed in 16 cases, suggesting their role as a driver. Notably, in at least 22 LADCs we observed multiple, independent classical chromothripsis events (Figure S6B), demonstrating that they could occur multiple times during the evolutionary history of the origin cell. In contrast, the coexistence of two independent balanced complex rearrangement clusters (portion of balanced breakpoints ≥ 0.4) was rare (only observed in BGI-WG06; Figure S6C).

In 61 LADCs, we observed highly amplified CRs involving many breakpoints (≥ 100) in multiple chromosomes (Figures 4A and 6C). These amplified CRs were especially common among the S4-low LADCs with EGFR-activating mutations or *MET* splice site mutations (observed in 23 cases out of 30). The congruent high copy numbers across segments in multiple

(B) Correlation between the number of somatic base substitutions and diagnosis ages in diploid LADCs without hypermutation (n = 15).

(C) Conceptual diagram of fusion oncogene formation in LU-FF76 and LU-4. Chromothripsis in S/G2 phase nucleus, followed by asymmetric segregation resulted in a copy-number gain of major allele (copy number = 2).

(D) Detailed structure of the derivative chromosome harboring *KIF5B-RET* in LU-FF76 (top) and the copy numbers of base substitutions (dots) and the allele-specific copy-number profile (solid lines; bottom). The pre-amplified substitutions are determined according to binomial probabilities (STAR Methods). CN, copy number.

See also Figure S5 and Table S4.

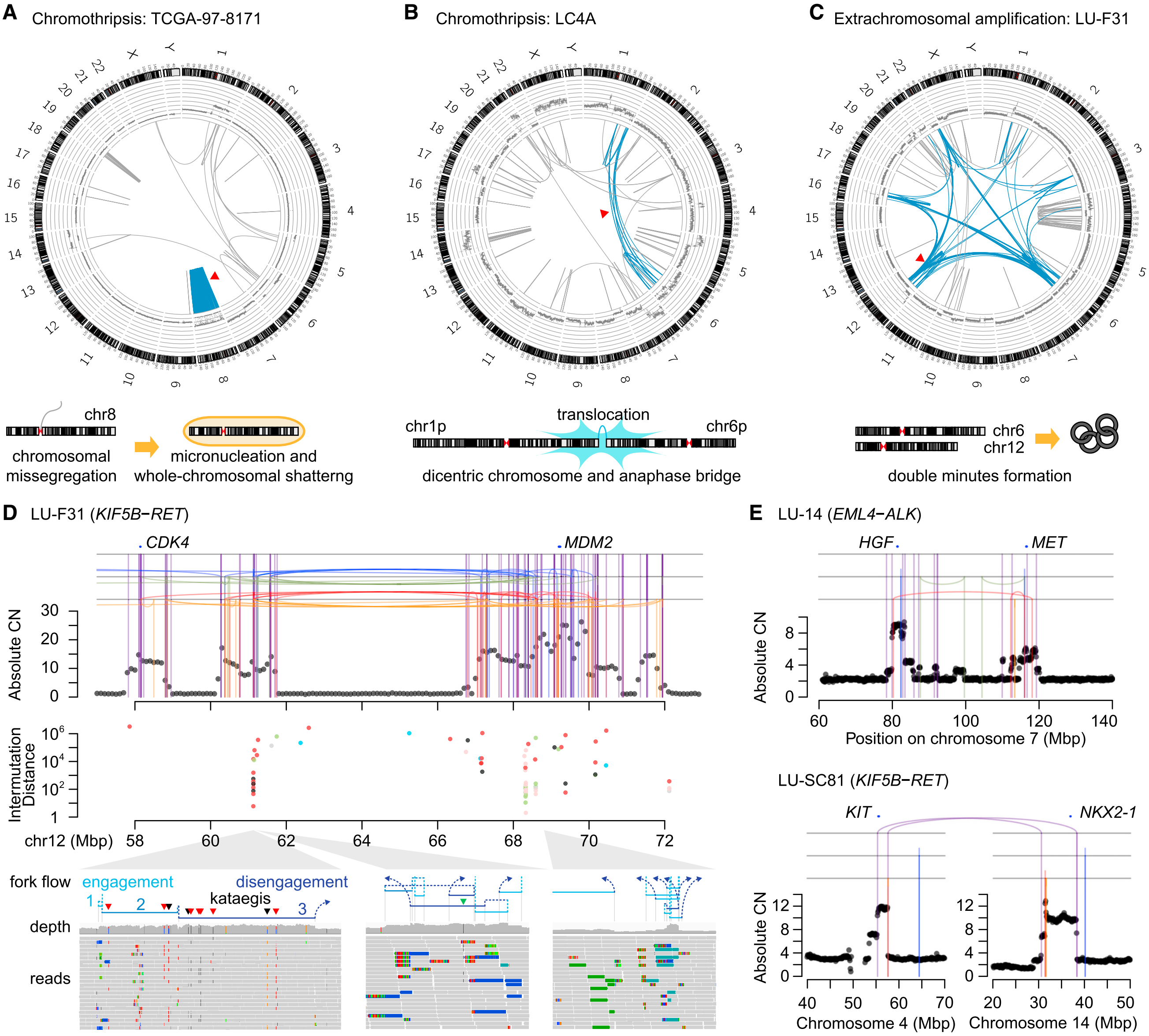


Figure 6. Complex CRs Providing Secondary Driver Events

(A–C) Circos plots showing the typical examples of whole-chromosomal chromothripsis (A), interchromosomal chromothripsis (B), and extrachromosomal amplification (C). Red triangles indicate the described complex rearrangements.

(D) *MDM2*/*CDK4* co-amplification in LU-F31 exhibiting frequent replication-mediated rearrangements. Its rearrangement structure (top), intermutation distances showing kataegis (middle), and representative regions with evidence of replication fork switching (bottom) are shown. Flows of replication forks are described with screenshots of Integrative Genomics Viewer (solid lines, strand synthesis; dashed lines, fork jumping). Base substitutions are indicated as dots (middle) and triangles (bottom) with the following colors: blue, C:G>A:T; black, C:G>G:C; red, C:G>T:A; gray, T:A>A:T; green, T:A>C:G; pink, T:A>G:C substitutions.

(E) Structures of rearrangements co-amplifying multiple oncogenes in LU-14 (top) and LU-SC81 (bottom).

See also Figure S6.

chromosomes indicate that extrachromosomal amplification (e.g., double minutes, homogeneously staining region, and neochromosomes) is likely the underlying structure (Turner et al., 2017; Garsed et al., 2014; Figure 6C). As observed in the amplicon of *AXL-MBIP* (Figure 3B), these amplified CRs also had several co-amplified breakpoints indicating their ancestral structure and numerous secondary breakpoints, which were acquired afterward (Garsed et al., 2014; Figure 6D). In many of these CRs, we observed local footprints with clustered rearrangements that have frequent microhomology at the breakpoints and short stretches of templated insertions, indicating a replication-mediated DNA double strand break repair (Helleday et al., 2014). Interestingly, these footprints frequently overlapped with kataegis, strand-coordinated clustered base substitutions with common C>T and C>G spectra, which could occur during replication-mediated DNA repair (Sakofsky et al., 2014). Kataegis substitutions were often co-amplified with the CRs (Figure S6D), indicating that the replication-mediated repair processes had been involved since the early stages of their structural evolution.

Known cancer-related genes were frequently amplified in these CRs; the most common examples were *TERT* (n = 21) and *MDM2* (n = 11), which were co-amplified with other cancer-related genes, including *CDK4*. Furthermore, the amplified CRs often contained oncogenic kinase genes (e.g., *MET* in LU-14, *KIT* in LU-SC81, *ERBB2* in BGI-WG24, etc.) that could mediate secondary resistance to targeted inhibitors of EGFR or ALK (Engelman et al., 2007; Katayama et al., 2012; Wilson et al., 2015; Figure 6E). If these treatment-naive tumors were treated with such inhibitors in an advanced clinical setting, these amplified oncogenes may confer rapid treatment failure to targeted inhibitors, which warrants clinical considerations. Overall, these cases show that complex genomic rearrangements also produce additional driver events after the initial oncogenic changes.

Distinct Oncogenesis Pathways Conferred by Different Oncogenic Drivers

To describe a full picture of secondary driver events in the development of LADCs, we next focused on common genomic alterations in LADC subgroups (Figure 7A). Here, we questioned whether the different mutational processes (S4-high versus -low) and different early driver oncogenes could lead to distinct genomic alteration landscapes.

In general, the mutations in cancer-related genes were largely clonal in both S4-high and -low LADCs, indicating that an effective clonal expansion started after accumulation of multiple driver mutations (Table S4). The number of mutations in cancer-related genes was larger in S4-high LADCs than in S4-low LADCs (19 versus 7, $p < 0.0001$, Student's t test), as expected from the crude mutational burden. However, the S4-high LADCs had mutations in canonical oncogenes (e.g., *KRAS*, *HRAS*, *BRAF*, and *EGFR*) only in about half of the cases (28 out of 55; Figure 7A), presumably due to the non-mutational, epigenetic activation of the MAPK pathway by smoking (Vaz et al., 2017). Among the S4-low LADCs (n = 83), 8 cases had no known alterations in major oncogenes. Of these, two cases commonly showed a recurrent fusion gene, *ERBB3-BCAR4* (Figure S7A). In one tumor (BGI-WG18), this fusion was formed by chromoplexy. In the other tumor with available RNA-seq data (TCGA-05-5429), an overexpression of corresponding fusion transcript was verified. The chimeric protein structure in both cases fully preserved the kinase domain of *ERBB3*, although its functional role has been controversial (Jaiswal et al., 2013). *BCAR4* is also known as an oncogene, of which fusion was previously reported in uterine cervical cancer (Cancer Genome Atlas Research Network, 2017). Interestingly, a recent WGS analysis of LADCs also reported *CD63-BCAR4* fusion from a female, never-smoking patient (Wang et al., 2018). The recurrent fusions of *BCAR4* with different partner genes indicate a key oncogenic role of *BCAR4*, especially among the patients with LADCs with no known alterations in major oncogenes. Overall, we estimate that >70% of LADC cases (102 out of 138) have genetic alterations that are clinically targetable or of which selective inhibitors are under development (Table S5) (Chakravarty et al., 2017).

The prevalence of WGD was different between the LADC subgroups, indicating their distinct pathways of oncogenesis (Figure 7A). WGD was common in S4-high LADCs (38 out of 55; 69%) and in S4-low LADCs with oncogene mutations (25 out of 36; 69%). *TP53* mutations and *MDM2* amplifications were enriched in these subgroups (61 out of 91; 67%, odds ratio = 5.248; $p < 0.0001$), upholding the role of p53 as a functional barrier for WGD (Aylon and Oren, 2011). In contrast, S4-low LADCs with fusion oncogenes had less frequent WGD events (12 out of 39; 31%). In several LADCs driven by fusion oncogenes (LU-57, LU-42, and LU-78), WGD occurred very early, when fewer than 1,000 genome-wide base substitutions were present (corresponding to ~22 years of age; Figure S7B). In contrast, some tumors had late WGD, or sequential amplifications of different chromosomes over time.

Frequently mutated genes were also different between the LADC subgroups (Figure 7A). S4-high LADCs had frequent clonal mutations in genes such as *STK11*, *KEAP1*, and *SMARCA4*, but these mutations were only rarely observed in S4-low cases. S4-low LADCs with oncogene mutations had frequent alterations in cell cycle regulator genes (27 out of 36; 75%) including bi-allelic inactivation of *CDKN2A* and *RB1* and amplifications of *CDK4*, *CDK6*, and *CCND1*.

LADCs with fusion oncogenes had a relatively silent mutational profile with rare *TP53* mutations (only 7 out of 39 cases had point mutations or bi-allelic inactivation; 18%). Instead, they had a frequent bi-allelic inactivation of *SETD2* (n = 7; 18%, 6 of them were clonal), indicating the gene's role as a tumor suppressor in this context. *SETD2*, a histone and microtubule methyltransferase, plays a critical role in recruiting the DNA damage repair machinery to active chromatin (Carvalho et al., 2014) and remodeling of the mitotic spindles (Park et al., 2016). We further analyzed 56 exomes and 58 targeted sequencing panels of LADCs (STAR Methods; Table S6) and confirmed that the *SETD2* mutations were significantly more frequent in fusion oncogene-driven LADCs (13 out of 82), compared to those with point mutations of oncogenes (2 out of 107; odds ratio = 9.783, p = 0.0006; Figure 7B). The enrichment of *SETD2* inactivation and depletion of *TP53* inactivation in fusion oncogene-driven LADCs were also consistently observed in our meta-analysis of publicly available datasets (n = 2,290; Table S6). In addition, bi-allelic inactivation of *PTEN* by focal deletion was observed in three LADCs with *RET* fusions (Figure S7C), which may also play an important role in this subgroup.

Lastly, we integrated the RNA-seq datasets (n = 79) to characterize the immune microenvironment of LADC subgroups (Table S7). S4-high LADCs showed a typical pro-inflammatory immune microenvironment, involving heavy CD8+ T cell infiltrations and polarization of macrophages toward M1 phenotype (Figure S7D). This is likely associated with their high mutational burden, which could result in greater load of neoantigens (304 versus 43, $p < 0.0001$) (Figure S7E; STAR Methods), compared to S4-low LADCs. The level of plasma cell was also higher in S4-high LADCs, suggesting a strong antitumor antibody response in this subgroup. We did not identify any significant difference of infiltrating immune cells between S4-low LADCs with oncogene mutations and fusion oncogenes.

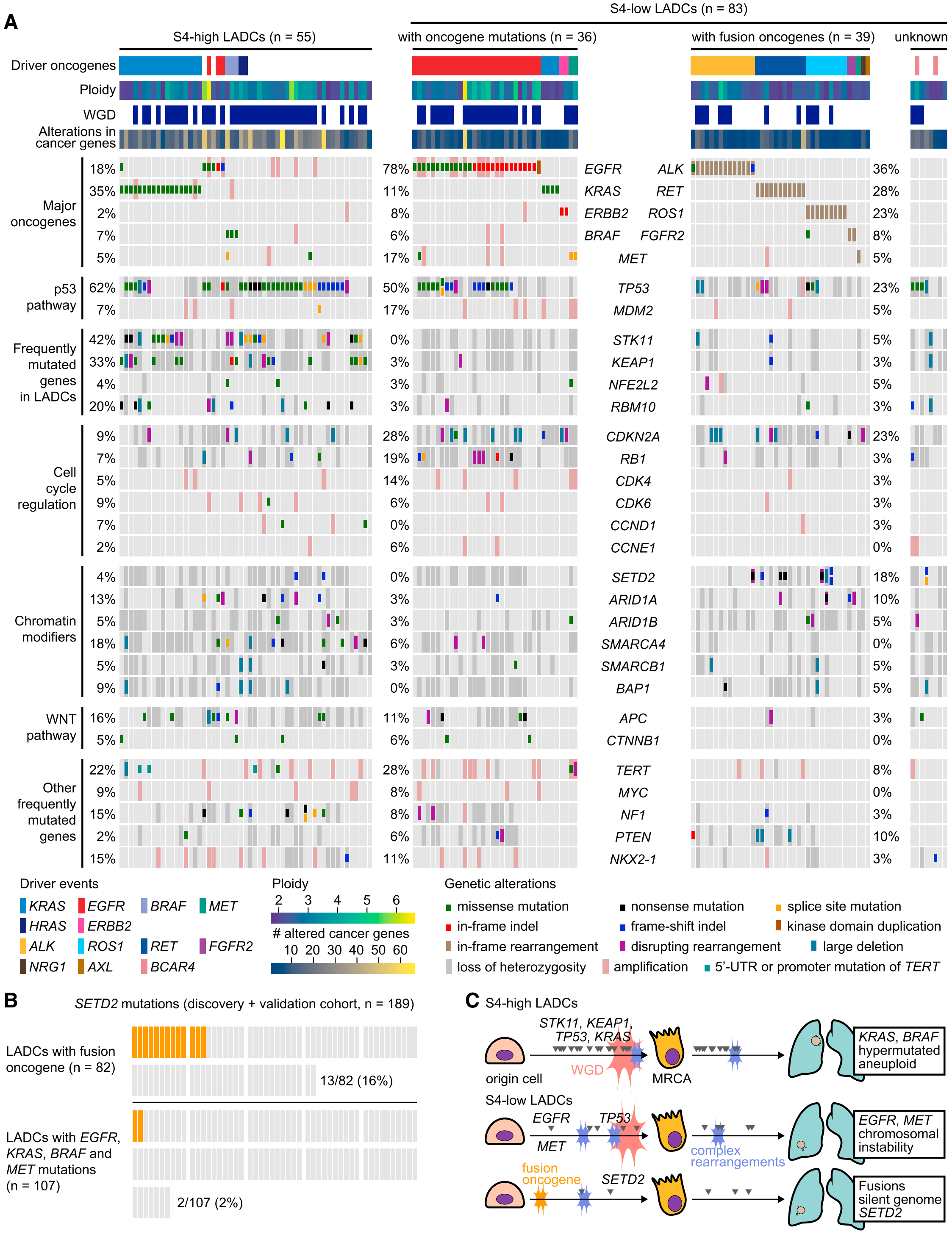


Figure 7. Distinct Mutational Landscape of LADC Subgroups

(A) Mutational landscape of LADCs in subgroups. Frequencies of mutations are described in percentages for each subgroup. Multiple mutations in one gene within a tumor were counted as one. Loss of heterozygosity was not considered in this counting.

(B) Validation of *SETD2* mutations in an expanded cohort.

(C) Conceptual diagrams of LADC oncogenesis.

See also Figure S7 and Tables S5, S6, and S7.

Altogether, our analysis provides a snapshot of distinct oncogenesis pathways of LADCs (Figure 7C). In non-smokers, oncogenesis is initiated by activation of canonical oncogenes and additional driver events are preferentially required to develop LADCs.

DISCUSSION

Our analysis shows that fusion oncogenes in LADCs are frequently generated from complex genomic rearrangement. These are catastrophic processes that have been explained by various errors of normal cellular physiology (Zhang et al., 2015; Maciejowski et al., 2015). The generative mechanism for the activating substitutions or indels of *EGFR* and *MET* is not yet clear. However, given the overwhelming prevalence of clock-like mutational signatures in the tumors with these mutations, they could also be associated with inevitable errors of normal cellular processes. Although we cannot exclude the possibility that these processes could be affected by environmental factors, our findings appreciate constitutive DNA damage and their illegitimate repair as the origin of LADCs from non-smokers (Zhu et al., 2016; Tomasetti et al., 2017). Basal rates of genomic rearrangements and their modifiers in normal airway epithelial cells will be helpful to more deeply understand LADC oncogenesis in non-smokers.

Our evolutionary analysis indicates that initiating oncogenic events are often acquired early in life, probably during adolescent ages. Two high-depth WGS cases showed that the fusion oncogenes could arise even before teenage years. These events likely take place in normal cells with competent DNA damage response. Notably, the LADCs driven by fusion oncogenes rarely had *TP53* inactivation (18%). These findings indicate that p53 inactivation is not a prerequisite for complex rearrangements including chromothripsis. Previous studies showed that chromothripsis was enriched in cancers harboring *TP53* mutations (Rausch et al., 2012a). These might be because the inactivation of p53 would make the cancer cells permissive to chromosomal instability, thus enabling cell survival after a chromosomal catastrophe like chromothripsis.

Mutations in *EGFR*, *MET*, and fusions involving *ALK*, *ROS1*, or *RET* define clinically relevant patient subsets (Herbst et al., 2018). Their different genomic alteration landscapes have clinical implications. Prevalent WGD and aneuploidy in *EGFR* mutant LADCs could partly explain their well-known adaptive resistance to treatment leading to frequent branched evolution (Lee et al., 2017a). Frequent *SETD2* mutations in fusion oncogene-driven LADCs may confer selective vulnerability (Pfister et al., 2015), which warrants further investigation.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - ○ Collection of Lung Adenocarcinoma Samples
- METHOD DETAILS
  - ○ Publicly Available Datasets
  - ○ Pathologic Examinations
  - ○ High-Throughput Sequencing
  - ○ Reads Alignment and Detection of Somatic Variants
  - ○ Mutational Signature Analysis
  - ○ Assessment of Smoking Based on Mutational Signatures
  - ○ Analysis of Genomic Rearrangements
  - ○ Classification of Complex Genomic Rearrangements
  - ○ Estimation of Purity, Ploidy, and Allele-Specific Copy Numbers
  - ○ Assessment for Copy Numbers and Clonal Status of Somatic Variants
  - ○ Determination of Pre- and Post-amplification Events
  - ○ Timing of Amplification
  - ○ Inferring Clock-like Mutation Rate of Origin Cells
  - ○ Validation of Enrichment of SETD2 Mutations in LADCs with Fusion Oncogenes
  - ○ Exploration of Immune Microenvironment in RNA-Seq Datasets
- DATA AND SOFTWARE AVAILABILITY

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.cell.2019.05.013.

ACKNOWLEDGMENTS

We thank Doga Gulhan, Max Sherman, Alon Galor, Isidro Cortes-Ciriano, Giorgio Melloni, Vinay Viswanadham, Cheng-Zhong Zhang, Emma V. Watson, Thomas Bleazard, Inigo Martincorena, David Pellman, and Peter Campbell for their comments and discussions. We also appreciate Jong-Yeon Shin, Seokjoo Chae, Santiago Gonzalez, Moritz Gerstung, and Peter Van Loo for their methodological advice. The results shown here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. We used dbGaP datasets (dbGaP: phs000488.v2.p1) with support from the National Human Genome Research Institute (NHGRI; 5U54HG003067). This study made use of data generated by the Cancer Research Center of BGI and Prof. Jianxing He's laboratory of The First Affiliated Hospital of Guangzhou Medical University. We used datasets that were funded by grants to Richard K. Wilson from Washington University in St. Louis and the NHGRI (U54 HG003079). Support was also provided by the Washington University Cancer Genome Initiative and Siteman Cancer Center. This research was supported by the Korea Health Technology R&D Project through the Korea Health Industry Development Institute, which was funded by the Ministry of Health & Welfare of Korea (HI16C2387 and HI17C1836 to Y.S.J.), the Suh Kyungbae Foundation (SUHF-18010082 to Y.S.J.), a National Research Foundation of Korea grant funded by the Korean government (MSIP; NRF-2014R1A2A2A05003665 to Y.T.K.), Research Program 2016 from the Seoul National University College of Medicine Research Foundation (800-20160251 to Y.T.K.), the Ludwig Center at Harvard (to J.J.-K.L. and P.J.P.), the Korea Institute of Science and Technology Information (K-17-L03-C022 to Junehawk Lee), and the Ministry of Trade, Industry, and Energy (Korea) under the Post-Genome Technology Development Program (10067758 to T.M.K.). Y.S.J. is a scientific fellow at the POSCO TJ Park Foundation.

AUTHOR CONTRIBUTIONS

Conceptualization, J.J.-K.L., Y.S.J., and Y.T.K.; Methodology, J.J.-K.L., S.P., and Y.S.J.; Software, S.P., H.P., Jongkeun Lee, D.H., and Y.S.J.; Validation,

H.P., S.K., J.Y., K.Y., Y.A., D.H.C., T.M.K., and Y.K.J.; Formal Analysis, J.J.-K.L., S.P., H.P., Junehawk Lee, J.Y., K.Y., Y.A., and Y.S.J.; Investigation, J.J.-K.L., S.P., S.K., I.K.P., C.H.K., T.M.K., Y.K.J., P.J.P., Y.S.J., and Y.T.K.; Resources, Junehawk Lee, I.K.P., C.H.K., T.M.K., Y.K.J., Y.S.J., and Y.T.K.; Data Curation, J.J.-K.L., S.P., H.P., J.Y., K.Y., and Y.A.; Writing – Original Draft, J.J.-K.L., S.P., and Y.S.J.; Writing – Review & Editing, all authors; Visualization, J.J.-K.L., S.P., H.P., S.K., Jongkeun Lee, Junehawk Lee, and D.H.; Supervision, Y.S.J. and Y.T.K.; Project Administration, J.J.-K.L. and S.P.; Funding Acquisition, Y.S.J. and Y.T.K.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 16, 2018
Revised: January 18, 2019
Accepted: May 3, 2019
Published: May 30, 2019

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Aparicio, S.A., Behjati, S., Biankin, A.V., Bignell, G.R., Bolli, N., Borg, A., Børresen-Dale, A.L., et al.; Australian Pancreatic Cancer Genome Initiative; ICGC Breast Cancer Consortium; ICGC MMML-Seq Consortium; ICGC PedBrain (2013). Signatures of mutational processes in human cancer. Nature *500*, 415–421.

Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407.

Anderson, N.D., de Borja, R., Young, M.D., Fuligni, F., Rosic, A., Roberts, N.D., Hajjar, S., Layeghifard, M., Novokmet, A., Kowalski, P.E., et al. (2018). Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. Science *361*, eaam8419.

Aylon, Y., and Oren, M. (2011). p53: guardian of ploidy. Mol. Oncol. *5*, 315–323.

Baca, S.C., Prandi, D., Lawrence, M.S., Mosquera, J.M., Romanel, A., Drier, Y., Park, K., Kitabayashi, N., MacDonald, T.Y., Ghandi, M., et al. (2013). Punctuated evolution of prostate cancer genomes. Cell *153*, 666–677.

Blokzijl, F., de Ligt, J., Jager, M., Sasselli, V., Roerink, S., Sasaki, N., Huch, M., Boymans, S., Kuijk, E., Prins, P., et al. (2016). Tissue-specific mutation accumulation in human adult stem cells during life. Nature *538*, 260–264.

Campbell, P.J., Getz, G., Stuart, J.M., Korbel, J.O., and Stein, L.D.; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Network (2017). Pan-cancer analysis of whole genomes. bioRxiv. https://doi.org/10.1101/162784.

Cancer Genome Atlas Research Network (2014). Comprehensive molecular profiling of lung adenocarcinoma. Nature *511*, 543–550.

Cancer Genome Atlas Research Network (2017). Integrated genomic and molecular characterization of cervical cancer. Nature *543*, 378–384.

Carvalho, S., Vítor, A.C., Sridhara, S.C., Martins, F.B., Raposo, A.C., Desterro, J.M., Ferreira, J., and de Almeida, S.F. (2014). SETD2 is required for DNA double-strand break repair and activation of the p53-mediated checkpoint. eLife *3*, e02482.

Chakravarty, D., Gao, J., Phillips, S.M., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yaeger, R., Soumerai, T., Nissan, M.H., et al. (2017). OncoKB: A precision oncology knowledge base. JCO Precis. Oncol. Published online May 16, 2017. https://doi.org/10.1200/PO.17.00011.

Chiang, C., Jacobsen, J.C., Ernst, C., Hanscom, C., Heilbut, A., Blumenthal, I., Mills, R.E., Kirby, A., Lindgren, A.M., Rudiger, S.R., et al. (2012). Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. Nat. Genet. *44*, 390–397, S1.

Cibulskis, K., Lawrence, M.S., Carter, S.L., Sivachenko, A., Jaffe, D., Sougnez, C., Gabriel, S., Meyerson, M., Lander, E.S., and Getz, G. (2013). Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nat. Biotechnol. *31*, 213–219.

Engelman, J.A., Zejnullahu, K., Mitsudomi, T., Song, Y., Hyland, C., Park, J.O., Lindeman, N., Gale, C.M., Zhao, X., Christensen, J., et al. (2007). MET amplification leads to gefitinib resistance in lung cancer by activating ERBB3 signaling. Science *316*, 1039–1043.

Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., and Eklund, A.C. (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann. Oncol. *26*, 64–70.

Garsed, D.W., Marshall, O.J., Corbin, V.D., Hsu, A., Di Stefano, L., Schröder, J., Li, J., Feng, Z.P., Kim, B.W., Kowarsky, M., et al. (2014). The architecture and evolution of cancer neochromosomes. Cancer Cell *26*, 653–667.

Gerstung, M., Jolly, C., Leshchiner, I., Dentro, S.C., Gonzalez, S., Mitchell, T.J., Rubanova, Y., Anur, P., Rosebrock, D., Yu, K., et al. (2017). The evolutionary history of 2,658 cancers. bioRxiv. https://doi.org/10.1101/161562.

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, N.D., Kanchi, K.L., Maher, C.A., Fulton, R., Fulton, L., Wallis, J., et al. (2012). Genomic landscape of non-small cell lung cancer in smokers and never-smokers. Cell *150*, 1121–1134.

Haffner, M.C., Aryee, M.J., Toubaji, A., Esopi, D.M., Albadine, R., Gurel, B., Isaacs, W.B., Bova, G.S., Liu, W., Xu, J., et al. (2010). Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. Nat. Genet. *42*, 668–675.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. Nat. Rev. Genet. *15*, 585–598.

Hendry, S., Salgado, R., Gevaert, T., Russell, P.A., John, T., Thapa, B., Christie, M., van de Vijver, K., Estrada, M.V., Gonzalez-Ericsson, P.I., et al. (2017). Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immunooncology Biomarkers Working Group: Part 1: assessing the host immune response, TILs in invasive breast carcinoma and ductal carcinoma in situ, metastatic tumor deposits and areas for further research. Adv. Anat. Pathol. *24*, 235–251.

Herbst, R.S., Morgensztern, D., and Boshoff, C. (2018). The biology and management of non-small cell lung cancer. Nature *553*, 446–454.

Imielinski, M., Berger, A.H., Hammerman, P.S., Hernandez, B., Pugh, T.J., Hodis, E., Cho, J., Suh, J., Capelletti, M., Sivachenko, A., et al. (2012). Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. Cell *150*, 1107–1120.

Jackson, E.L., Willis, N., Mercer, K., Bronson, R.T., Crowley, D., Montoya, R., Jacks, T., and Tuveson, D.A. (2001). Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. Genes Dev. *15*, 3243–3248.

Jaiswal, B.S., Kljavin, N.M., Stawiski, E.W., Chan, E., Parikh, C., Durinck, S., Chaudhuri, S., Pujara, K., Guillory, J., Edgar, K.A., et al. (2013). Oncogenic ERBB3 mutations in human cancers. Cancer Cell *23*, 603–617.

Jamal-Hanjani, M., Wilson, G.A., McGranahan, N., Birkbak, N.J., Watkins, T.B.K., Veeriah, S., Shafi, S., Johnson, D.H., Mitter, R., Rosenthal, R., et al.; TRACERx Consortium (2017). Tracking the evolution of non-small-cell lung cancer. N. Engl. J. Med. *376*, 2109–2121.

Ji, H., Li, D., Chen, L., Shimamura, T., Kobayashi, S., McNamara, K., Mahmood, U., Mitchell, A., Sun, Y., Al-Hashem, R., et al. (2006). The impact of human EGFR kinase domain mutations on lung tumorigenesis and in vivo sensitivity to EGFR-targeted therapies. Cancer Cell *9*, 485–495.

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

Katayama, R., Shaw, A.T., Khan, T.M., Mino-Kenudson, M., Solomon, B.J., Halmos, B., Jessop, N.A., Wain, J.C., Yeo, A.T., Benes, C., et al. (2012). Mechanisms of acquired crizotinib resistance in ALK-rearranged lung Cancers. Sci. Transl. Med. *4*, 120ra17.

Kloosterman, W.P., Guryev, V., van Roosmalen, M., Duran, K.J., de Bruijn, E., Bakker, S.C., Letteboer, T., van Nesselrooij, B., Hochstenbach, R., Poot, M., and Cuppen, E. (2011). Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. Hum. Mol. Genet. *20*, 1916–1924.

Koboldt, D.C., Zhang, Q., Larson, D.E., Shen, D., McLellan, M.D., Lin, L., Miller, C.A., Mardis, E.R., Ding, L., and Wilson, R.K. (2012). VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. *22*, 568–576.

Koh, J., Go, H., Keam, B., Kim, M.Y., Nam, S.J., Kim, T.M., Lee, S.H., Min, H.S., Kim, Y.T., Kim, D.W., et al. (2015). Clinicopathologic analysis of programmed cell death-1 and programmed cell death-ligand 1 and 2 expressions in pulmonary adenocarcinoma: comparison with histology and driver oncogenic alteration status. Mod. Pathol. *28*, 1154–1166.

Lee, J.K., Lee, J., Kim, S., Kim, S., Youk, J., Park, S., An, Y., Keam, B., Kim, D.W., Heo, D.S., et al. (2017a). Clonal history and genetic predictors of transformation into small-cell carcinomas from lung adenocarcinomas. J. Clin. Oncol. *35*, 3065–3074.

Lee, J.K., Louzada, S., An, Y., Kim, S.Y., Kim, S., Youk, J., Park, S., Koo, S.H., Keam, B., Jeon, Y.K., et al. (2017b). Complex chromosomal rearrangements by single catastrophic pathogenesis in NUT midline carcinoma. Ann. Oncol. *28*, 890–897.

Lee, J., Lee, A.J., Lee, J.K., Park, J., Kwon, Y., Park, S., Chun, H., Ju, Y.S., and Hong, D. (2018). Mutalisk: a web-based somatic MUTation AnaLyIS toolKit for genomic, transcriptional and epigenomic signatures. Nucleic Acids Res. *46* (W1), W102–W108.

Maciejowski, J., Li, Y., Bosco, N., Campbell, P.J., and de Lange, T. (2015). Chromothripsis and kataegis induced by telomere crisis. Cell *163*, 1641–1654.

Mehine, M., Kaasinen, E., Mäkinen, N., Katainen, R., Kämpjärvi, K., Pitkänen, E., Heinonen, H.R., Bützow, R., Kilpivaara, O., Kuosmanen, A., et al. (2013). Characterization of uterine leiomyomas by whole-genome sequencing. N. Engl. J. Med. *369*, 43–53.

Mitchell, T.J., Turajlic, S., Rowan, A., Nicol, D., Farmery, J.H.R., O'Brien, T., Martincorena, I., Tarpey, P., Angelopoulos, N., Yates, L.R., et al.; TRACERx Renal Consortium (2018). Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx Renal. Cell *173*, 611–623.

Mitelman, F., Johansson, B., and Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. Nat. Rev. Cancer *7*, 233–245.

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

Nik-Zainal, S., Van Loo, P., Wedge, D.C., Alexandrov, L.B., Greenman, C.D., Lau, K.W., Raine, K., Jones, D., Marshall, J., Ramakrishna, M., et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012). The life history of 21 breast cancers. Cell *149*, 994–1007.

Park, I.Y., Powell, R.T., Tripathi, D.N., Dere, R., Ho, T.H., Blasius, T.L., Chiang, Y.C., Davis, I.J., Fahey, C.C., Hacker, K.E., et al. (2016). Dual chromatin and cytoskeletal remodeling by SETD2. Cell *166*, 950–962.

Pfister, S.X., Markkanen, E., Jiang, Y., Sarkar, S., Woodcock, M., Orlando, G., Mavrommati, I., Pai, C.C., Zalmas, L.P., Drobnitzky, N., et al. (2015). Inhibiting WEE1 selectively kills histone H3K36me3-deficient cancers by dNTP starvation. Cancer Cell *28*, 557–568.

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. (2012a). Genome sequencing of pediatric medulloblastoma links catastrophic DNA rearrangements with TP53 mutations. Cell *148*, 59–71.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012b). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

Redin, C., Brand, H., Collins, R.L., Kammin, T., Mitchell, E., Hodge, J.C., Hanscom, C., Pillalamarri, V., Seabra, C.M., Abbott, M.A., et al. (2017). The genomic landscape of balanced cytogenetic abnormalities associated with human congenital anomalies. Nat. Genet. *49*, 36–45.

Saito, M., Shimada, Y., Shiraishi, K., Sakamoto, H., Tsuta, K., Totsuka, H., Chiku, S., Ichikawa, H., Kato, M., Watanabe, S., et al. (2015). Development of lung adenocarcinomas with exclusive dependence on oncogene fusions. Cancer Res. *75*, 2264–2271.

Sakofsky, C.J., Roberts, S.A., Malc, E., Mieczkowski, P.A., Resnick, M.A., Gordenin, D.A., and Malkova, A. (2014). Break-induced replication is a source of mutation clusters underlying kataegis. Cell Rep. *7*, 1640–1648.

Santaguida, S., Richardson, A., Iyer, D.R., M'Saad, O., Zasadil, L., Knouse, K.A., Wong, Y.L., Rhind, N., Desai, A., and Amon, A. (2017). Chromosome mis-segregation generates cell-cycle-arrested cells with complex karyotypes that are eliminated by the immune system. Dev. Cell *41*, 638–651.e5.

Saunders, C.T., Wong, W.S., Swamy, S., Becq, J., Murray, L.J., and Cheetham, R.K. (2012). Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics *28*, 1811–1817.

Seo, J.S., Ju, Y.S., Lee, W.C., Shin, J.Y., Lee, J.K., Bleazard, T., Lee, J., Jung, Y.J., Kim, J.O., Shin, J.Y., et al. (2012). The transcriptional landscape and mutational profile of lung adenocarcinoma. Genome Res. *22*, 2109–2119.

Shaw, A.T., Kim, D.W., Nakagawa, K., Seto, T., Crinó, L., Ahn, M.J., De Pas, T., Besse, B., Solomon, B.J., Blackhall, F., et al. (2013a). Crizotinib versus chemotherapy in advanced ALK-positive lung cancer. N. Engl. J. Med. *368*, 2385–2394.

Shaw, A.T., Hsu, P.P., Awad, M.M., and Engelman, J.A. (2013b). Tyrosine kinase gene rearrangements in epithelial malignancies. Nat. Rev. Cancer *13*, 772–787.

Shaw, A.T., Ou, S.H., Bang, Y.J., Camidge, D.R., Solomon, B.J., Salgia, R., Riely, G.J., Varella-Garcia, M., Shapiro, G.I., Costa, D.B., et al. (2014). Crizotinib in ROS1-rearranged non-small-cell lung cancer. N. Engl. J. Med. *371*, 1963–1971.

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.

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.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Sun, S., Schiller, J.H., and Gazdar, A.F. (2007). Lung cancer in never smokers–a different disease. Nat. Rev. Cancer *7*, 778–790.

Tomasetti, C., Li, L., and Vogelstein, B. (2017). Stem cell divisions, somatic mutations, cancer etiology, and cancer prevention. Science *355*, 1330–1334.

Turner, K.M., Deshpande, V., Beyter, D., Koga, T., Rusert, J., Lee, C., Li, B., Arden, K., Ren, B., Nathanson, D.A., et al. (2017). Extrachromosomal oncogene amplification drives tumour evolution and genetic heterogeneity. Nature *543*, 122–125.

Vaz, M., Hwang, S.Y., Kagiampakis, I., Phallen, J., Patil, A., O'Hagan, H.M., Murphy, L., Zahnow, C.A., Gabrielson, E., Velculescu, V.E., et al. (2017). Chronic cigarette smoke-induced epigenomic changes precede sensitization of bronchial epithelial cells to single-step transformation by KRAS mutations. Cancer Cell *32*, 360–376.

Wang, C., Yin, R., Dai, J., Gu, Y., Cui, S., Ma, H., Zhang, Z., Huang, J., Qin, N., Jiang, T., et al. (2018). Whole-genome sequencing reveals genomic signatures associated with the inflammatory microenvironments in Chinese NSCLC patients. Nat. Commun. *9*, 2054.

Wilson, F.H., Johannessen, C.M., Piccioni, F., Tamayo, P., Kim, J.W., Van Allen, E.M., Corsello, S.M., Capelletti, M., Calles, A., Butaney, M., et al. (2015). A functional landscape of resistance to ALK inhibition in lung cancer. Cancer Cell *27*, 397–408.

Wu, K., Zhang, X., Li, F., Xiao, D., Hou, Y., Zhu, S., Liu, D., Ye, X., Ye, M., Yang, J., et al. (2015). Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas. Nat. Commun. *6*, 10131.

Yokoyama, A., Kakiuchi, N., Yoshizato, T., Nannya, Y., Suzuki, H., Takeuchi, Y., Shiozawa, Y., Sato, Y., Aoki, K., Kim, S.K., et al. (2019). Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature *565*, 312–317.

Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat. Med. *23*, 703–713.

Zhang, C.Z., Leibowitz, M.L., and Pellman, D. (2013). Chromothripsis and beyond: rapid genome evolution from complex chromosomal rearrangements. Genes Dev. *27*, 2513–2530.

Zhang, C.Z., Spektor, A., Cornils, H., Francis, J.M., Jackson, E.K., Liu, S., Meyerson, M., and Pellman, D. (2015). Chromothripsis from DNA damage in micronuclei. Nature *522*, 179–184.

Zhu, L., Finkelstein, D., Gao, C., Shi, L., Wang, Y., López-Terrada, D., Wang, K., Utley, S., Pounds, S., Neale, G., et al. (2016). Multi-organ mapping of cancer risk. Cell *166*, 1132–1146.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD3 Rabbit Monoclonal Antibody (2GV6)	Ventana Medical Systems	Cat# 790-4341; RRID: AB_2335978
CD8 Rabbit Monoclonal Antibody (SP16)	Thermo Fisher	Cat# MA5-14548; RRID: AB_10984334
PD-1 Mouse Monoclonal Antibody (MRQ-22)	Cell Marque	Cat# NAT105
PD-L1 XP Rabbit Monoclonal Antibody (E1L3N)	Cell Signaling	Cat# 13684S; RRID: AB_2687655
PD-L2/B7-DC Mouse Monoclonal Antibody (176611)	R&D Systems	Cat# MAB1224; RRID: AB_2161995
Deposited Data		
WGS of tumor and normal samples	This paper	EGA: EGAS00001002801
RNA sequencing of tumor samples	This paper	EGA: EGAS00001002801
Human reference genome NCBI build 37, GRCh37	1000 Genomes Project	http://www.internationalgenome.org/category/grch37/
WGS data from TCGA LUAD	The Cancer Genome Atlas (TCGA)	dbGaP: phs000178.v9.p8
RNA sequencing files from TCGA LUAD	The Cancer Genome Atlas (TCGA)	https://portal.gdc.cancer.gov
WGS data from "Mapping the Hallmarks of Lung Adenocarcinoma with Massively Parallel Sequencing"	Imielinski et al., 2012	dbGaP: phs000488.v2.p1
WGS data from "Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas"	Wu et al., 2015	EGA: EGAS00001000982
WGS data from "Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers"	Govindan et al., 2012	Available upon request to the author
WGS data from "Fusion of KIF5B and RET transforming gene in lung adenocarcinoma revealed from whole-genome and transcriptome sequencing"	Ju et al., 2012	ENA: ERP001071
WGS data from "Clonal History and Genetic Predictors of Transformation Into Small-Cell Carcinomas From Lung Adenocarcinomas"	Lee et al., 2017a	EGA: EGAS00001001757
COSMIC Cancer Gene Census	Wellcome Sanger Institute	https://cancer.sanger.ac.uk/census
Software and Algorithms		
Burrows-Wheeler Aligner (BWA) MEM	See link	http://bio-bwa.sourceforge.net/
Samtools	See link	https://github.com/samtools/samtools
Sambamba	See link	http://lomereiter.github.io/sambamba/
Picard	See link	https://broadinstitute.github.io/picard/
GATK	See link	https://software.broadinstitute.org/gatk/
MuTect	Cibulskis et al., 2013	https://github.com/broadinstitute/mutect
Strelka	Saunders et al., 2012	https://sites.google.com/site/strelkasomaticvariantcaller/home
VarScan2	Koboldt et al., 2012	http://dkoboldt.github.io/varscan/
Delly	Rausch et al., 2012b	https://github.com/dellytools/delly
Sequenza	Favero et al., 2015	http://www.cbs.dtu.dk/biotools/sequenza/
Battenberg	Nik-Zainal et al., 2012	https://github.com/Wedge-Oxford/battenberg
Mutalisk	Lee et al., 2018	http://mutalisk.org
Annovar	See link	http://annovar.openbioinformatics.org/en/latest/
STAR	See link	https://github.com/alexdobin/STAR

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
RSEM	See link	https://github.com/deweylab/RSEM
OptiType	See link	https://github.com/FRED-2/OptiType
netMHCpan4.0	See link	http://www.cbs.dtu.dk/services/NetMHCpan/
CIBERSORT	See link	https://cibersort.stanford.edu
Filtering of structural variants from normal samples	This study	Data S1
Clustering of genomic rearrangements	This study	Data S1
Classification of complex rearrangement clusters	This study	Data S1
Amplification timing analysis	This study	Data S1
Nonamer generator script	This study	https://github.com/ju-lab/LungCancer_SV_timing_analyses
Other		
Integrative genomics viewer	See link	http://software.broadinstitute.org/software/igv/

CONTACT FOR REAGENT AND RESOURCE SHARING

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Young Seok Ju (ysju@kaist.ac.kr).

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Collection of Lung Adenocarcinoma Samples

To study genetic alteration landscape of LADCs, we carried out a cancer panel-based genotyping study for early-stage LADCs. Among 1,258 surgically resected, histologically-confirmed lung cancers that were registered to the thoracic malignancy tissue repository of Seoul National University Hospital, 947 tumors had adenocarcinoma histology. Among them, 350 LADCs were enrolled to this study with written informed consent (Figure S1A). We collected demographic and clinical information including age, sex, stage, and smoking history, and pathology findings including histologic features and molecular testing results per study protocol. Two different versions of targeted deep sequencing panels were used to detect mutations in key oncogene hotspots as well as fusion oncogenes, based on our previous study (Seo et al., 2012). The initial version (MG-LCDx panel) targeted 17 hotspot mutations and 7 known fusion oncogenes, and the later version (FIRST-cancer panel) probed 316 cancer-related gene mutations and 30 fusion oncogenes. Among the 350 LADCs, 51 tumors had known fusion oncogenes. These tumors did not have any coexisting point mutations in major oncogenes. Among the 299 tumors without known fusion oncogenes, 133 had *EGFR* mutations, and 56 had *KRAS* mutations. Among the tumors with sufficient amount of remaining fresh frozen tumor tissue and matched normal tissue specimens (n = 347), we randomly selected tumors for whole-genome sequencing. To compare the genomic alteration landscape between the clinically relevant subgroups of patients, we divided our samples into three groups based on the presence of known fusion oncogenes and smoking history (Figure S1), and selected tumors separately from these groups. A total of 53 LADCs were sequenced for their whole-genomes, and 49 were finally included in the analysis. The protocol of this study was reviewed and approved by the institutional review board of Seoul National University Hospital (IRB No: IRB 1109-108-379 and 1805-109-947).

METHOD DETAILS

Publicly Available Datasets

We included publicly available WGS datasets to our analysis, to reach a more complete picture of genomic alterations in LADCs. From the previous literature, we selected cases meeting our inclusion criteria: i) WGS datasets of surgically resected, and histologically-confirmed LADC cases, ii) treatment-naive tumors, iii) presence of whole-genome sequences of both tumor and paired normal tissues, and iv) datasets generated by paired-end Illumina high-throughput sequencing. We included a total of 89 WGS datasets of LADCs, largely from three previous studies (Cancer Genome Atlas Research Network, 2014; Imielinski et al., 2012; Wu et al., 2015) and others (Ju et al., 2012; Govindan et al., 2012; Lee et al., 2017a). We also analyzed published 43 exome sequencing datasets for validation of *SETD2* mutations (Wu et al., 2015). Raw sequencing data (FASTQ or BAM files by BWA-MEM alignment tool) were downloaded from public data repositories or from institutional repository. These genome sequences were processed using the same bioinformatics pipeline applied to the newly sequenced WGS datasets (shown below). RNA-seq FASTQ files were also downloaded for 37 tumors from TCGA study, and were processed together with the newly generated RNA-seq files.

Pathologic Examinations

Histology of 93 surgically resected LADCs were reviewed three thoracic pathologists (S.K., Y.K.J, and D.H.C.). Of these, 44 LADCs from outside of our LADC sequencing cohort were included for an immune marker profiling using immunohistochemical staining. Two pathologists (S.K., and Y.K.J.) initially examined all cases separately, and disagreed cases were discussed among three pathologists and finally reached consensus. Tumor histologic features were described according to the 2015 WHO classification system, focusing on predominant and associated histologic features in representative H&E sections. The degree of lymphocyte infiltration was assessed from the H&E slides according to the recent practical guideline (Hendry et al., 2017). Immunohistochemical staining of immune cell- or environment-related markers including CD3, CD8, PD-1, PD-L1, and PD-L2 was performed on 64 LADCs (including 20 of our whole-genome sequenced LADCs and 44 additional LADCs), as described in our previous study (Koh et al., 2015). Representative images of H&E tumor slides were taken under microscope, listed in Table S1, and uploaded to our study website with their related details (http://genome.kaist.ac.kr).

High-Throughput Sequencing

Fifty-three LADCs and their matched normal samples were sequenced for their whole genomes. We also sequenced whole genome of a LADC cell line (SNU-2612A). This cell line was established from malignant pleural effusion of patient AK55, from whom the *KIF5B−RET* fusion was first discovered (Ju et al., 2012). Genomic DNA materials were extracted from fresh frozen tumor tissues and their matched peripheral blood or normal lung tissue samples using DNeasy Blood and Tissue kits (QIAGEN, Venlo, Netherlands). DNA libraries for WGS were generated by a TruSeq PCR-Free Library Preparation Kit (Illumina, San Diego, CA) from 1 μg of genomic DNA materials. WGS was performed on a HiSeq X platform (Illumina) to generate a minimal read depth of 30X for both tumors and matched normal blood samples. For seven tumors with low purity (typically less than or around 0.2) in our initial analysis, we increased the sequencing depth to 70X with additional whole-genome sequencing. In addition, two tumors (LU-FF76 and LU-4) were finally sequenced to a genome-wide coverage of 130X, for a more detailed timing analysis. Among the 49 LADCs, 34 had remaining fresh frozen tissues, which were used for RNA-seq. An mRNA sequencing library was constructed according to the TruSeq protocol (Illumina, San Diego, CA) with 1 μg of mRNA materials, and then sequenced on a HiSeq 2500 machine with a target throughput of 10Gb. As an independent validation of high frequency of *SETD2* mutations among the LADCs driven by fusion oncogenes, we also carried out exome sequencing of 13 LADCs with their paired normal tissues and 58 targeted sequencing of LADCs harboring driver mutations (*EGFR, KRAS, BRAF*, and *MET*) or fusion oncogenes (*ALK, RET*, and *ROS1*), in the same manner that we described in our previous study (Lee et al., 2017a).

Reads Alignment and Detection of Somatic Variants

The raw FASTQ files were aligned to the human reference genome (GRCh37) using BWA-MEM algorithm. The duplicated reads were removed by Picard (available at http://broadinstitute.github.io/picard), and indel realignment and base quality score recalibration were performed by GATK (available at https://software.broadinstitute.org/gatk/). Depth of coverage of each sample, including publicly available WGS dataset is described in Table S1. To establish the highly sensitive somatic variant sets, we initially took the unions of variant calls from MuTect (Cibulskis et al., 2013) and Strelka (Saunders et al., 2012) for base substitutions and Strelka (Saunders et al., 2012) and Varscan2 (Koboldt et al., 2012) for indels. Next, we constructed a locus-specific background error matrix using all the normal tissue WGS used in this study, and this matrix was used to further filter out the sequencing artifacts. We used Delly (Rausch et al., 2012b) for initial somatic rearrangement calls and sequencing artifacts were filtered out using a locus-specific background dataset established from normal samples used in this study (Data S1). Rearrangements with poor mapping quality (median MAPQ < 40), with insufficient number of supporting reads, or those with many discordant reads in paired normal BAM files were considered to be false positives and removed. Short-sized deletions and duplications (< 1 Kbp) without soft-clipped reads, and unbalanced inversions (< 5 Kbp) without sufficient supporting reads (n < 5), which are mostly DNA library artifacts were also removed. We determined accurate breakpoint positions and microhomology sequences using "SA tag" of the clipped reads. Variant allele fractions of rearrangement breakpoints were calculated as the ratio between the number of concordant (supporting wild-type) and discordant (supporting rearrangement) read pairs from BAM files. Briefly, the number of read pairs spanning the rearrangement breakpoint in right orientation with normal insert size was defined as wild-type pairs. The number of i) discordant read pairs involving soft-clipped read by the rearrangement breakpoint, and ii) those without soft-clipped read but of which mapped position and orientation support the rearrangement event were counted as variant. After this process, we visually inspected all the rearrangements using Integrative Genomics Viewer to remove remaining false positive events and to rescue false negative events that were located nearby the breakpoints. More detailed step-by-step instructions are available in Data S1.

Mutational Signature Analysis

We analyzed mutational signatures by linear decomposition by using Mutalisk (Lee et al., 2018). To summarize, the relative contributions of mutational signatures were calculated by refitting 7 consensus mutational signatures previously identified and validated in lung cancers (COSMIC signatures 1, 2, 4, 5, 13, 17, and 18; available at https://cancer.sanger.ac.uk/cosmic/signatures). All possible combinations of the 7 mutational signatures were evaluated by non-negative least-squares function as described in our previous study (Lee et al., 2017a). Although mutational signatures 6 and 15 had been reported from lung adenocarcinoma cases with microsatellite instability (MSI) features, we excluded these two signatures in our decomposition analysis because we identified no tumor showing the typical mutational pattern of MSI (extreme number of indels and base substitutions). In three tumors, the mutational spectra required additional signatures for optimal decomposition. Addition of signature 28 for TCGA-67-6215, and a new signature (we used SBS37 of PCAWG signature catalog for this) for LUAD-TJW61 and LUAD-QY22Z showed significant improvements of decomposed results. The 96-trinucleotide based mutational spectra of 138 LADC cases and their mutational signature contributions are available in Table S2.

Assessment of Smoking Based on Mutational Signatures

Smoking information of our 138 LADCs were collected in various manners in each study (Ju et al., 2012; Govindan et al., 2012; Lee et al., 2017a; Wu et al., 2015; Cancer Genome Atlas Research Network, 2014). Therefore, we simplified the clinical history of smoking to classify the LADCs into two groups: ever-smoker and never-smoker. Since the mutational signature could provide reliable information about tobacco smoke exposure (Alexandrov et al., 2013), we assessed the extent and the proportion of base substitutions that were attributed to signature 4 in each case. Given that the smoking-induced mutagenesis predominantly contributes to early mutations along with clock-like mutational signatures (signatures 1 and 5) (Jamal-Hanjani et al., 2017), we used the proportion of signature 4 out of the sum of proportion of signatures 1, 4, and 5 as our readout. By k-means clustering ($k = 2$) based on the count and the proportion of signature 4 mutations, we found that our 138 LADCs were clearly separated into two groups (55 smoking-related and 83 smoking-unrelated LADCs). These groups well matched with their clinical history of smoking as well as the driver mutation status.

Analysis of Genomic Rearrangements

Given the prevalence of complex genomic rearrangements in human cancers (Campbell et al., 2017) and their distinct mechanisms from simple rearrangement processes (Zhang et al., 2013), we defined complex genomic rearrangements and analyzed them separately from the simple rearrangements. To define complex rearrangement events, we clustered the rearrangements based on spatial proximity with cut-offs at inter-breakpoint distance of < 5 Mbp. To infer their mechanisms, we systematically analyzed the genomic features of complex rearrangement clusters. First, we classified all breakpoints into copy-number balanced and unbalanced groups. Paired breakpoints located in tail-to-head orientation with distance $\leq$ 500 bp were defined as balanced breakpoints. Second, we assigned allele-specific absolute DNA copy numbers for each DNA segment. A DNA segment was defined as the region between the two breakpoints aligned in head-to-tail orientation. Third, we analyzed sequence microhomology and inserted nucleotides for each rearrangement breakpoint. Microhomologies of 2 bp or longer were considered as significant microhomology, and the blunt end ligation included 0 or 1 bp of microhomologies. Last, we analyzed genomic and epigenomic contexts of rearrangement breakpoints, including GC content, DNase I hypersensitivity sites, various histone marks, and replication timing. To do this, we used Mutalisk (Lee et al., 2018) with ENCODE A549 lung adenocarcinoma cell line as the reference epigenome in the most analyses. For DNA replication timing analysis, we used Repli-seq dataset from GM12878 (B lymphoblastoid cells) because the datasets from lung adenocarcinoma lineage were not available.

To analyze the impact of complex rearrangement clusters on cancer-related genes, we analyzed the copy number variations of involved genes. Rearrangement-related oncogene amplification was defined as the complex rearrangement cluster amplifying oncogenes in the cluster or nearby the breakpoints (distance < 5 Mbp). Here we defined the amplified genes as those whose absolute copy number was larger by at least 5 from the median copy number of the genes in the same chromosome. Rearrangement-related tumor suppressor gene (TSG) inactivation was defined as with a TSG transected by the rearrangement breakpoints. We regarded a TSG as homozygously deleted when the absolute copy number of the gene and its surrounding 100-Kbp genomic segment was 0.7 or less. The oncogenes and the TSGs were defined by Cancer Gene Census (COSMIC v84; available at https://cancer.sanger.ac.uk/census). We also analyzed kataegis (clustered base substitutions) in the vicinity of rearrangement breakpoints (breakpoint position ± 1 Kbp). Kataegis was defined as 3 or more clustered base substitutions within 1 Kbp.

Classification of Complex Genomic Rearrangements

Two co-first authors (J.J.-K.L. and S.P.) independently performed a manual curation of driver oncogene rearrangements in 39 LADCs with fusion oncogenes. Reconstruction was started from the breakpoints at the oncogenes (e.g., *ALK*, *ROS1*, and *RET*) or from their fusion partner genes (e.g., *EML4*, *CD74*, and *KIF5B*). Discordant read pairs and soft-clipped reads were used to reconstruct the whole structure of the complex rearrangements. Then, the results were compared with the output of Delly analysis to supplement the rearrangement structure. The final reconstructed maps of 39 oncogene rearrangements were cross-checked by two co-first authors and the consensus structures were finally generated with a corresponding author of this study (Y.S.J.). Based on the final structures of oncogene rearrangements and their copy number profile, we tried to classify them based on the known criteria of complex genomic rearrangements. We followed the definition of chromothripsis in the initial study (Stephens et al., 2011), referring to localized complex rearrangements exhibiting copy number oscillation between two or three states, and chromoplexy was defined as chains of reciprocal rearrangements involving multiple chromosomes, as commonly described in the initial study (Baca et al., 2013) and a review (Zhang et al., 2013). We also identified a group of rearrangements with chains of reciprocal rearrangements but not involving multiple (> 2) chromosomes, and called them as balanced chromothripsis. For 7 complex oncogene rearrangements, we could not classify their pattern because they were comprised of small number of rearrangements. Thus, we classified these events as unspecified complex rearrangements. Last, we found chromothripsis involving three or more chromosomes, which were similar to the examples in Figure 4 of the initial chromothripsis study (Stephens et al., 2011). We referred these to as multichromosomal chromothripsis. Three FORs (in LU-89, LU-FF58, and LU-SC126) from multichromosomal chromothripsis were commonly comprised of balanced and unbalanced breakpoints in different copy number states. These cases were explained by stepwise model of chromoplexy followed by secondary chromothripsis.

We also applied similar classification principle on collateral rearrangement (CR) clusters. Large CRs involving 10 or more rearrangement events were included in this analysis. Here we also considered copy number features of CRs, in conjunction with their number of involved chromosomes and the fraction of balanced breakpoints. We found that the variability of copy numbers in a CR cluster was helpful in distinguishing highly amplified rearrangement clusters (typically showing a large standard deviation of absolute copy numbers of breakpoints) from copy number loss-dominant clusters such as chromothripsis (typically oscillating between 2−3 copy number levels with small gaps). Therefore, focally amplified CRs were defined as CRs with highly variable absolute copy numbers at their breakpoints (standard deviation $\geq$ 1.7). Related computational scripts and step-by-step instructions are available in Data S1.

We plotted these events into the two-dimensional plane with fraction of balanced breakpoints on the x axis and number of chromosomes involved on the y axis and confirmed their grouped distribution on the plane (Figure 4A). Four groups were divided based on whether they involved two or more chromosomes, and whether the fraction of balanced breakpoints was less than 0.4, the cut-off value that offered the best separation of samples into two groups in histogram. Same criteria and cut-offs were applied to the analysis including the collateral rearrangements.

Estimation of Purity, Ploidy, and Allele-Specific Copy Numbers

We used Sequenza (Favero et al., 2015) and Battenberg (Nik-Zainal et al., 2012) algorithms to estimate tumor purity, ploidy, and segmented copy numbers. Purity and ploidy estimates from two different algorithms were largely consistent, but discordant solutions were made for several cases. In such cases we chose the solution with more reasonable copy number profile; for example, we avoided solutions in which extensive genomic regions were assigned as copy number = 0, or solutions assigning unreasonably large fractions of subclonal point mutations and rearrangements. In these cases, we re-ran the Sequenza algorithm with the refined purity and ploidy values as input parameters. We processed the output files of Sequenza to smoothen the copy number profile for further analysis. Absolute copy numbers were manually calculated for 100 Kbp-sized bins for whole genome using the read depth of tumor and paired normal samples, the purity and the ploidy estimates. To determine the WGD status, we used the ploidy estimates and the fraction of genomes showing loss of heterozygosity, as described in a recent large-scale analysis (Gerstung et al., 2017). For TCGA whole-genomes, we were guided by the purity, ploidy estimates and the allele-specific copy number profiles generated by cutting-edge bioinformatics algorithms by ICGC-PCAWG (Pan-Cancer Analysis of Whole-Genomes) consortium (Campbell et al., 2017). In the most of the cases, the ploidy and purity estimates by Sequenza were in good agreement with the results by PCAWG. However, when we saw discrepant values between the two solutions, we followed PCAWG estimates for our downstream analysis.

Assessment for Copy Numbers and Clonal Status of Somatic Variants

We estimated mutation copy number (n_{mut}) by previously described formula below (Nik-Zainal et al., 2012).

$$n_{mut} = f_s \frac{1}{\rho}\left[\rho n^{t}_{locus} + n^{n}_{locus}(1-\rho)\right]$$

In this formula, f_s indicates variant allele fraction, ρ indicates tumor cellularity. n^{t}_{locus} and n^{n}_{locus} are absolute copy numbers in tumor and normal cells, respectively, which were derived from following formula.

$$n_{locus} = 2 \times \frac{RD_{locus}}{RD_{\text{auto}}}$$

in which RD_{locus} indicates read depth of the locus of interest, and RD_{auto} indicates average haploid autosomal coverage that was obtained from paired normal WGS. Number of chromosomal copies of the mutation (n_{chr}) were inferred from n_{mut}, by selecting positive integer value that were closest to n_{mut}. Cancer cell fraction (CCF), which indicated fraction of cancer cells harboring a mutation, was calculated as below (Nik-Zainal et al., 2012).

$$CCF = \frac{n_{mut}}{n_{chr}}$$

Determination of Pre- and Post-amplification Events

To assess whether a specific somatic point mutation or a rearrangement was acquired before chromosomal copy gains, we compared mutation copy number (n_{mut}) with the allele-specific copy number of the chromosomal segment. A mutation was classified as pre-amplification event, when n_{mut} was larger than m, the positive integer value larger than the major copy number of tumor (n^t_{maj}) × 0.75. If n_{mut} was smaller than the above value but larger than 0.75, the mutation was assigned as post-amplification or minor allele mutations. Subclonal events (which are present in a fraction of cancer cells) were defined when the absolute value of n_{mut} was smaller than 0.75 copy. The probabilities of pre-amplification (B_{pre}), post-amplification or minor allele ($B_{post/min}$), and subclonal (B_{sub}) mutation were calculated using total read depth (DP) and variant read count (VC) of the mutation, n^t_{locus}, and ρ, as assessed by binomial probability as follows.

$$B_{pre} = \sum_{i=m}^{n^t_{maj}} \binom{DP}{VC} \frac{\rho i}{\rho n^t_{locus} + n^n_{locus}(1-\rho)}^{VC} \left(1 - \frac{\rho i}{\rho n^t_{locus} + n^n_{locus}(1-\rho)}\right)^{DP-VC}$$

$$B_{post/min} = \sum_{i=1}^{m-1} \binom{DP}{VC} \frac{\rho i}{\rho n^t_{locus} + n^n_{locus}(1-\rho)}^{VC} \left(1 - \frac{\rho i}{\rho n^t_{locus} + n^n_{locus}(1-\rho)}\right)^{DP-VC}$$

$$B_{sub} = \binom{DP}{VC} \frac{0.5\rho}{\rho n^t_{locus} + n^n_{locus}(1-\rho)}^{VC} \left(1 - \frac{0.5\rho}{\rho n^t_{locus} + n^n_{locus}(1-\rho)}\right)^{DP-VC}$$

The above values were converted into final probabilities of pre-amplification (P_{pre}), post-amplification or minor-allele ($P_{post/min}$), and subclonal mutations (P_{sub}) as below.

$$P_{pre} = \frac{B_{pre}}{(B_{pre} + B_{post/min} + B_{sub})}$$

$$P_{post/min} = \frac{B_{post/min}}{(B_{pre} + B_{post/min} + B_{sub})}$$

$$P_{sub} = \frac{B_{sub}}{(B_{pre} + B_{post/min} + B_{sub})}$$

To determine whether the FORs occurred before or after the copy number amplification, we followed the sequence as follows. Of the 39 FORs, 23 were overlapped with copy number amplifications. For these 23 FORs, we determined the temporal relationship between the rearrangement and the copy number amplification. For ten FORs, the amount of DNA copy number changes at the rearrangement junctions were 2 or higher, suggesting that the rearrangements preceded the amplifications (copy number changes in non-amplified rearrangement junctions should be 1 or 0). In other seven cases, the FORs were co-amplified with focal DNA amplicons involving multiple chromosomes, also indicating pre-amplification events. For the remaining six cases, the temporal relationship was determined using variant allele fractions of breakpoints, in the same manner for base substitutions as mentioned above. We determined the FORs as pre-amplification events when the median n_{mut} of the rearrangements cluster was more than n^t_{maj} × 0.65 (Figure S5B). Based on this criterion, four out of the six FORs were pre-amplification events. The remaining two cases (in LU-78 and LU-SC81) were confirmed as post-amplification events by phasing with heterozygous SNPs. In LU-78, we found that the FOR followed a very early WGD event (Figure S5C). None of these FORs were subclonal event.

Timing of Amplification

In order to estimate the timing of amplification events, we analyzed the expected number of pre-amplification substitutions (E_{pre}) by summation of the P_{pre} values of all substitutions in amplified segments larger than 10 Mbp. In cases of rearrangement clusters involving multiple arm-level amplifications, we used the minimum E_{pre} among them. The 95% confidence interval was obtained from z-scores using the sum of $P_{pre} \times (1 - P_{pre})$ as variance. If a substantial amount of APOBEC-mediated localized hypermutations was found in an amplicon (proportion of S2 + S13 > 20% among pre-amplification mutations), we corrected their fraction when we calculated E_{pre}. The density of pre-amplification mutations in the amplicon ($E_{pre}/(length\ of\ amplicon)$) was converted to a genome-wide density, and then translated into physical timescale with an assumption of constant mutation rate (45 substitutions/year; described in the next section).

We found that the pre-amplification mutations of S4-high LADCs were largely from smoking-induced mutagenesis, showing a profound contribution of S4 (Figure S5D). Due to the lack of clock-like property of S4, we excluded S4-high LADCs from the formal analysis of timing estimation. In contrast, the pre-amplification mutations of the most S4-low LADCs were largely explained by clock-like signatures 1 and 5 (Figure S5D), suitable for our timing analysis. Among the S4-low LADCs, we excluded hypermutated tumors with strong contribution of APOBEC-associated mutational signatures (total S2 and S13 mutation counts > 10,000), because we found that many pre-amplification mutations in these tumors were also APOBEC-associated.

Inferring Clock-like Mutation Rate of Origin Cells

Previous studies demonstrated a gradual and linear accumulation of mutations in somatic cells, including cancer cells. In a study of mutational signatures in multiple cancer types (Alexandrov et al., 2015), two signatures (COSMIC signatures 1 and 5) were linearly and positively correlated with age of diagnosis in the most tumor types, demonstrating their molecular clock-like properties. Likewise, another study using organoid models of normal human intestine and liver showed the number of mutations in adult stem cells were also linearly correlated with ages of donors, and the spectra of those mutations were largely explained by the clock-like signatures 1 and 5 (Blokzijl et al., 2016). These studies indicate that a majority of these clock-like somatic base substitutions in cancer cells are accumulated before the malignant transformation with more-or-less constant rate with aging (Alexandrov et al., 2015), and therefore the physical age of a specific somatic cell from the fertilized egg can be estimated by the number of somatic substitutions (Stratton et al., 2009). Based on this, a recent study estimated the timing of tumor-initiating unbalanced translocation in renal cell carcinoma with an assumption of constant mutation rate (Mitchell et al., 2018). A similar approach was also made in a large-scale analysis of multiple cancer types (Gerstung et al., 2017).

To implement this approach to our LADC dataset, we first aimed to infer the clock-like somatic mutation rate in the origin cells of LADCs. To this end, we used genome-wide base substitutions identified from cancer samples that were not affected by hypermutation processes such as APOBEC-associated mutagenesis and exogenous mutagens (e.g., smoking). This serves as a reasonable surrogate for the mutation rate of origin cells, as previously demonstrated (Mitchell et al., 2018). These LADCs were selected using multiple criteria as follows: smoking-unrelated, stable ploidy (< 2.5), adequate purity ($\geq$0.33), low fraction of APOBEC-associated mutations (COSMIC signatures 2 and 13 < 20%), and absence of dominant subclones. As a result, this analysis came up with 15 LADCs with stable genomes, and their numbers of somatic base substitutions and ages of diagnosis showed a linear correlation (r = 0.52, p = 0.0484), as expected.

$$\text{Number of substitution} = 44.44 \times \text{age} + 458.10$$

This rate broadly agreed with the mutation rates of normal tissue stem cells, which were reported from recent studies using organoid models (~40 mutations/year) (Blokzijl et al., 2016). Therefore, we assumed that somatic base substitutions had accumulated at an approximate rate of ~45/year in the average origin cells of our LADCs, and applied this rate to estimate the timings of copy number amplifications.

Validation of Enrichment of SETD2 Mutations in LADCs with Fusion Oncogenes

To directly validate the enrichment of *SETD2* mutations and the depletion of *TP53* mutations in LADCs with fusion oncogenes, we studied the presence of these mutations in an independent cohort of LADCs from Seoul National University Hospital (exome sequencing, n = 13; targeted sequencing, n = 58). We also re-analyzed 43 published exome sequencing datasets (Wu et al., 2015) to calculate the frequency of *SETD2* and *TP53* mutations.

We also searched the literature to find out the datasets or study results that could be used to analyze the enrichment or depletion of *SETD2* and *TP53* mutations between the three LADC subgroups defined by driver mutation status (fusion oncogene-driven, point mutations in canonical oncogene-driven, and driver unidentified groups). Three additional studies using exome or targeted sequencing (Cancer Genome Atlas Research Network, 2014; Saito et al., 2015; Zehir et al., 2017) along with our WGS-based analysis were used for this meta-analysis. Comparisons were made between three LADC groups. Odds ratios were calculated from the Fisher's exact test (two-sided p values were used).

Exploration of Immune Microenvironment in RNA-Seq Datasets

To investigate the immune microenvironment of our LADCs, we analyzed RNA-seq datasets from 79 LADCs in our merged cohort. Reads were aligned by STAR algorithm (https://github.com/alexdobin/STAR) with two-pass protocol. We used RSEM (https://github.com/deweylab/RSEM) in quantification of gene expression. Based on this result, we ran CIBERSORT (https://cibersort.stanford.edu) with transcript per million (TPM) values, and calculated the relative fraction of tumor-infiltrating immune cells. The fractions of different immune cell types were compared between the LADC subgroups.

We also predicted burden of tumor neoantigen by integrating the RNA-seq and the WGS datasets. We used OptiType algorithm (https://github.com/FRED-2/OptiType) with RNA-seq reads to obtain HLA class I type of each sample. Using our in-house script (available in https://github.com/ju-lab/LungCancer_SV_timing_analyses), we listed up all possible nonamers from the somatic mutational profile of 79 LADCs. For frameshift indels, we extracted nonamers based on the altered downstream protein sequence. Using the nonamer list and the HLA class I type information, we predicted tumor neoantigens using netMHCpan 4.0 (http://www.cbs.dtu.dk/services/NetMHCpan/).

DATA AND SOFTWARE AVAILABILITY

Newly generated sequencing datasets, including WGS (BAM files) and RNA-seq (FASTQ files), have been deposited in European Genome-Phenome Archive (EGA). The accession number for the sequence reported in this paper is EGA: EGAS00001002801. Computational scripts and related step-by-step instructions are available in Data S1 and in our public repository (https://github.com/ju-lab/LungCancer_SV_timing_analyses).

Variant calls including base substitutions, indels, and genomic rearrangements, mutational signatures, RNA-based gene expression profiles, pathology findings, and their key images are available to the public at our project website (http://genome.kaist.ac.kr).

Supplemental Figures

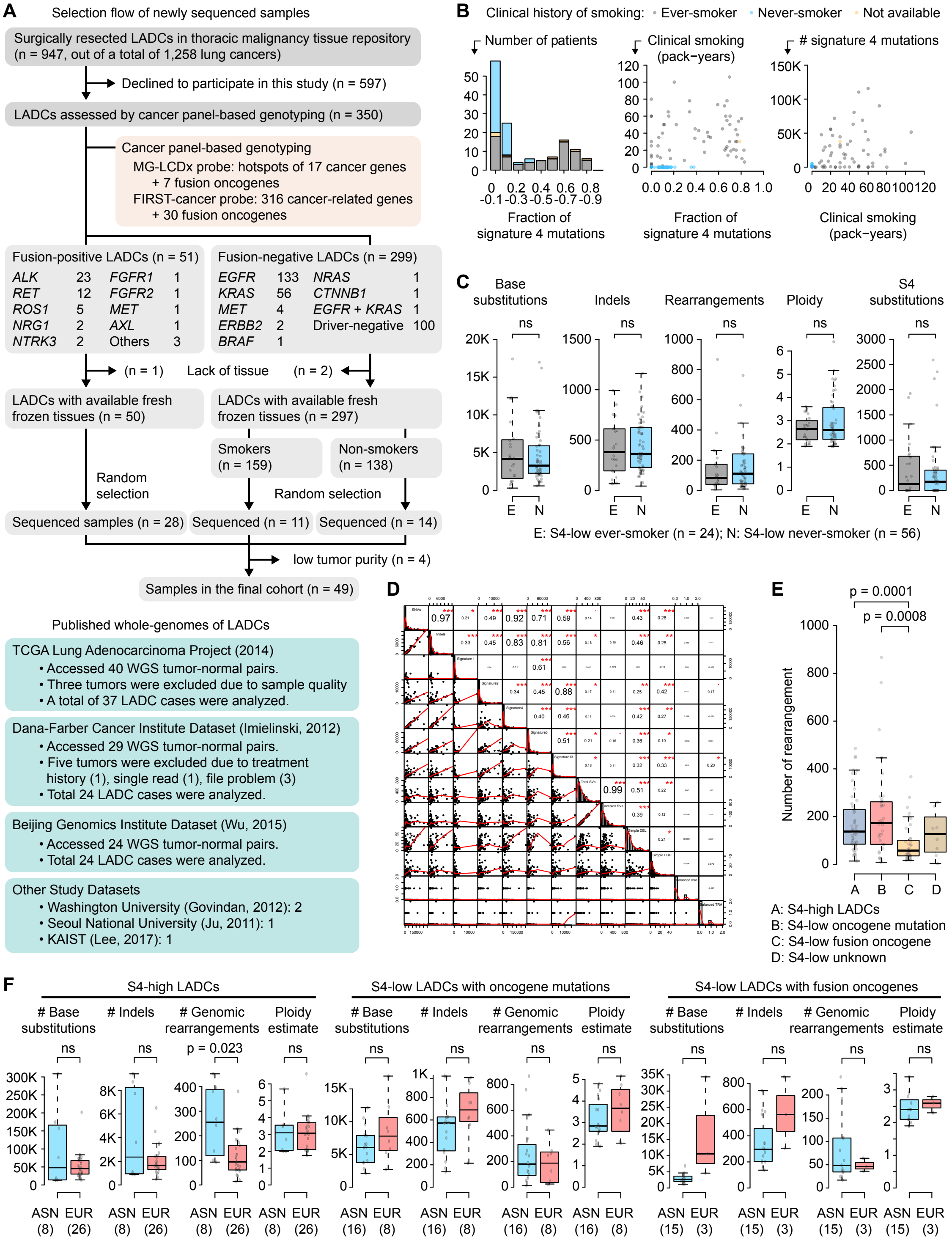


Figure S1. Study Design and Genomic Features of 138 LADCs, Related to Figure 1

(A) Collection process of newly sequenced LADCs (upper panel), and re-analyzed WGS datasets from previous publications (lower panel).

(B) Correlation between reported smoking history and exposure to mutational signature 4 (S4). Stacked histogram showing the distribution of the contribution of S4 to the mutational spectra and the clinical history of smoking (left panel). Scatterplot of the pack-year values from the patients' smoking history relative to the contribution of S4 mutations (middle panel). A weak but statistically significant correlation between the absolute count of S4 mutations and the pack-year values (right panel; r = 0.33, p = 0.0002).

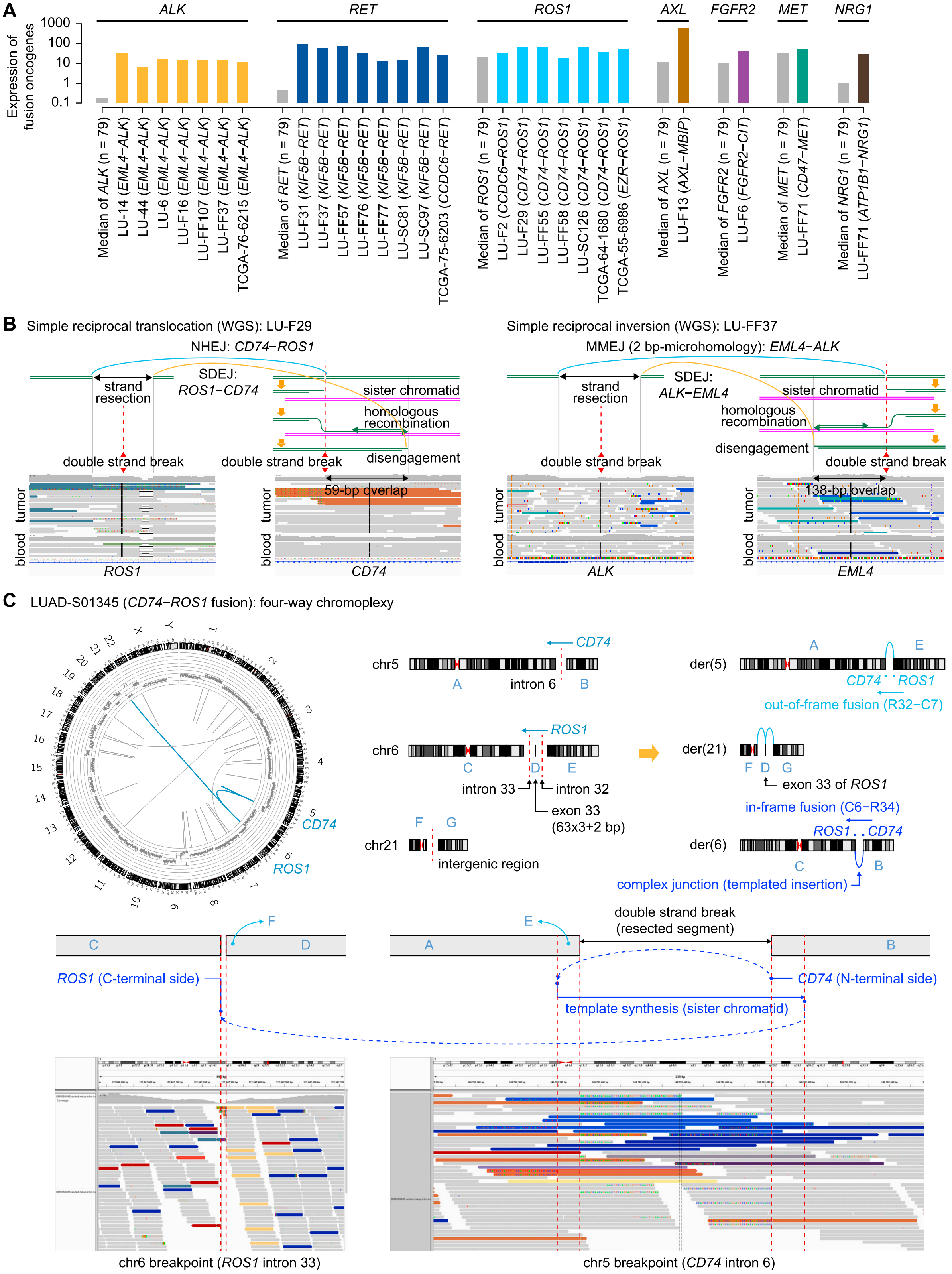
(legend continued on next page)

(C) Similar burden of genomic alterations between ever-smokers (n = 24) and never-smokers (n = 56) within the S4-low LADCs. Statistical comparisons are based on Student's t test (two-sided).

(D) Multiple correlation plot for mutational signatures and classes of genomic rearrangements. Diagonal cells show histograms for indicated variables. Off diagonal-cells in the lower left side show the correlation between two variables in scatter plots. Each dot indicates a data point from each tumor. Cells in the upper right side indicate Pearson's correlation coefficients (r), and the number of asterisks indicate statistical significance (***$p < 0.001$; **$p < 0.01$; *$p < 0.05$). In this plot, numbers of both simple deletions (r = 0.42) and duplications (r = 0.27) are positively correlated with S4 mutation counts. The correlation between simple deletions and S4 counts remains significant after exclusion of LADCs with no S4 mutations and no deletions (r = 0.43, $p < 0.0001$). However, the correlation between simple duplications and S4 counts was not significant after removing samples with no S4 mutations and no duplications (r = 0.18, p = 0.0901). SV, structural variation; DEL, deletion; DUP, duplication; INV, inversion; TRA, interchromosomal translocation.

(E) Burden of genomic rearrangements in subgroups of LADCs. Comparisons are made with a Student's t test (two-sided).

(F) Similar burden of genomic alterations among the patients of East Asian (ASN) and European (EUR) ancestry within the three LADC subgroups. The LADCs of purity less than 0.3 were excluded in this analysis. The p values (t test, two-sided) are shown at the top of boxplots. ns, not significant.



(legend on next page)

Figure S2. Genomic Features of Rearrangement Breakpoints Generating Fusion Oncogenes, Related to Figure 2

(A) Transcript abundance (transcripts per million; TPM) of oncogenes involved in driver fusions from RNA-seq analysis. S4-low LADCs driven by fusion oncogenes with available RNA-seq data (n = 26) are included in this analysis. Gray bar represents the median value of all available cases (n = 79). In-frame expression of the fusion oncogene transcript was confirmed in all cases.

(B) Detailed structure of simple genomic rearrangements generating fusion oncogenes. A reciprocal translocation in LU-F29 (left panel), and a reciprocal inversion in LU-FF37 (right panel). Both cases are shown because they have typical features of rearrangement breakpoints indicating synthesis-dependent end-joining (SDEJ). One side of the rearrangement (left) shows microdeletions at breakpoints, which can be attributed to the strand trimming process after the DNA double strand breaks. Another side (right) shows a duplicated sequence (59- and 138-bp long), which is present in both sides of the reciprocal rearrangements. This duplicated sequence is not explained by microhomology, and is consistent with the SDEJ model (Helleday et al., 2014).

(C) Circos plot and mechanistic illustration of a chromoplexy event generating a *CD74–ROS1* fusion in LUAD-S01345 (upper panel). Reconstructed complex junctional sequence connecting intron 6 of *CD74* (right) and intron 33 of *ROS1* (left), making an in-frame fusion (lower panel).

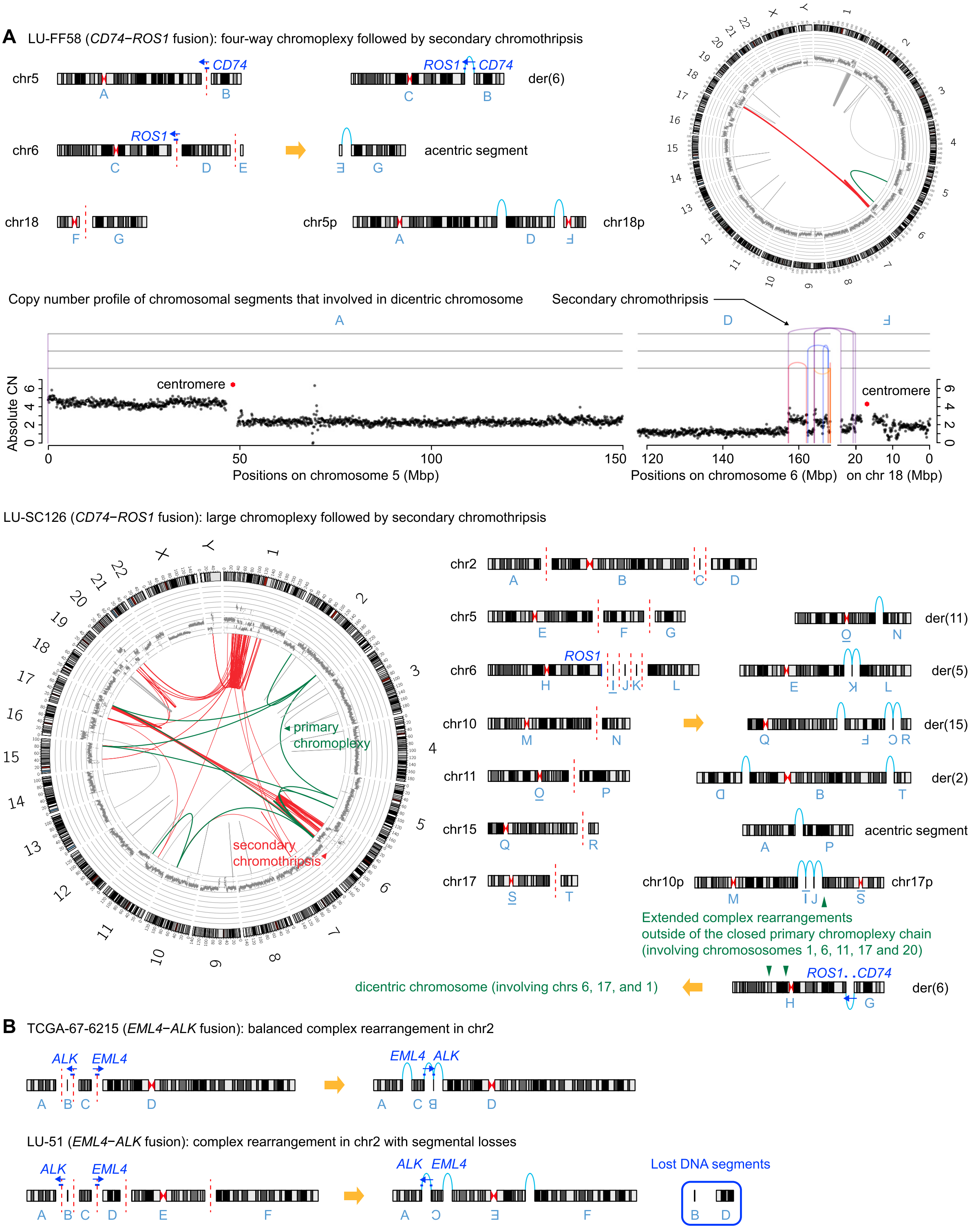

A
LU-FF58 (CD74–ROS1 fusion): four-way chromoplexy followed by secondary chromothripsis
chr5
CD74
A
B
chr6
ROS1
C
D
E
chr18
F
G
ROS1 CD74
der(6)
C
B
acentric segment
E
G
chr5p
A
D
F
chr18p
Copy number profile of chromosomal segments that involved in dicentric chromosome
Secondary chromothripsis
centromere
Absolute CN
0 2 4 6
0
50
100
150
Positions on chromosome 5 (Mbp)
120
140
160
Positions on chromosome 6 (Mbp)
20
10
0
on chr 18 (Mbp)
LU-SC126 (CD74–ROS1 fusion): large chromoplexy followed by secondary chromothripsis
primary chromoplexy
secondary chromothripsis
chr2
A
B
C
D
chr5
E
F
G
chr6
ROS1
H
I
J
K
L
chr10
M
N
chr11
O
P
chr15
Q
R
chr17
S
T
der(11)
O
N
der(5)
E
K
L
der(15)
Q
F
C R
der(2)
D
B
T
acentric segment
A
P
chr10p
M
I J
S
chr17p
Extended complex rearrangements
outside of the closed primary chromoplexy chain
(involving chromososomes 1, 6, 11, 17 and 20)
dicentric chromosome (involving chrs 6, 17, and 1)
ROS1 CD74
der(6)
H
G
B
TCGA-67-6215 (EML4–ALK fusion): balanced complex rearrangement in chr2
ALK
EML4
A
B
C
D
EML4
ALK
A
C
B
D
LU-51 (EML4–ALK fusion): complex rearrangement in chr2 with segmental losses
ALK
EML4
A
B
C
D
E
F
ALK
EML4
A
C
E
F
Lost DNA segments
B
D

Figure S3. Reconstruction of Initial Complex Genomic Rearrangement in Multichromosomal Chromothripsis, Related to Figure 3

(A) Complex FORs of two cases (LU-FF58 and LU-SC126) are likely explained by initial chromoplexy, followed by secondary chromothripsis. In LU-FF58 (upper panel), the allele-specific copy number profile indicates the co-occurrence of the FOR (green in circos plot) and an adjacent chromothripsis event (red). The initial event would be a chromoplexy involving three chromosomes (5, 6, and 18) generating a dicentric chromosome. The *CD74–ROS1* fusion oncogene is not affected by a secondary rearrangement burst because it is located in a monocentric chromosome 6. A long segment of the dicentric chromosome (from the centromere of chromosome 5 to ~160 Mbp region of chromosome 6) is likely lost during the chromoplexy-induced chromothripsis (two rearrangement junctions of the initial

(legend continued on next page)

chromoplexy event would also be lost from this event). This could explain the observed copy number pattern in chromosomes 6 and 18. In LU-SC126 (lower panel), the initial chromoplexy involved seven chromosomes (green). This chain is fully reconstructed and generates an acentric segment (which agrees with the copy number losses of the chromosomes 2 and 11), and a dicentric chromosome involving chromosomes 10 and 17. However, this putative dicentric chromosome could not explain the observed copy number patterns (the secondary chromothripsis is observed between chromosomes 6, 1, and 17, rather than 10 and 17). This suggests that the initial chromoplexy event may involve more than one closed chain of rearrangements, which could further exchange the genomic segments of the putative dicentric chromosome, with those of chromosomes 1 and 6. In the observed secondary chromothripsis event (red), we found several balanced chain-like rearrangements, which might be parts of an initial, large chromoplexy event.

(B) Complex FORs involving small numbers of rearrangements (<5). The FOR in TCGA-67-6215 is a balanced chromothripsis-like event involving three DNA double strand breaks (upper panel). A similar example was reported in uterine myomas (Mehine et al., 2013). In LU-51, two genomic segments are lost during the repair process (lower panel).

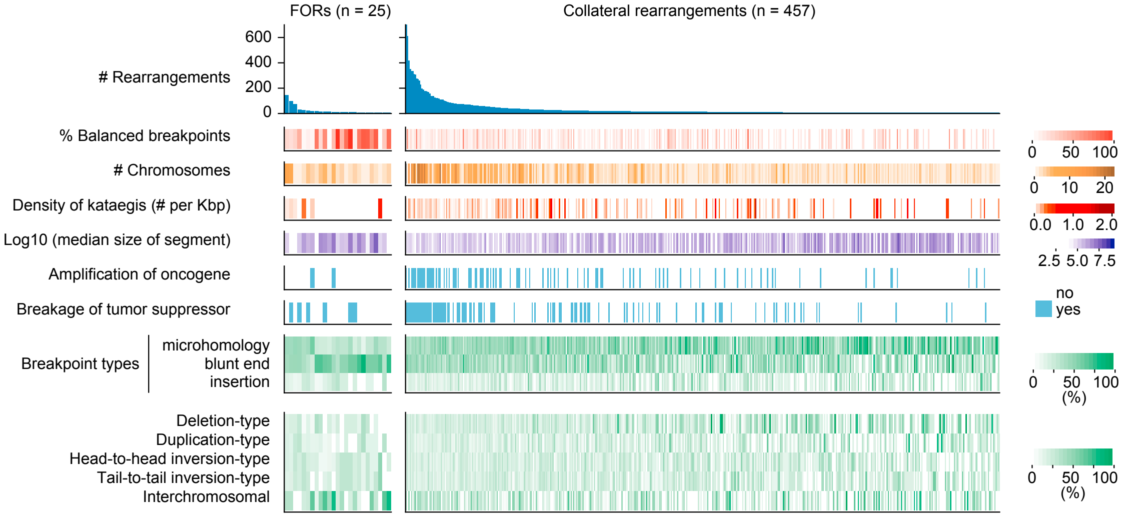


Figure S4. Genomic Features of Complex Rearrangements, Related to Figure 4

Genomic characteristics of complex rearrangement clusters involving five or more rearrangement events are shown in heatmaps (STAR Methods).

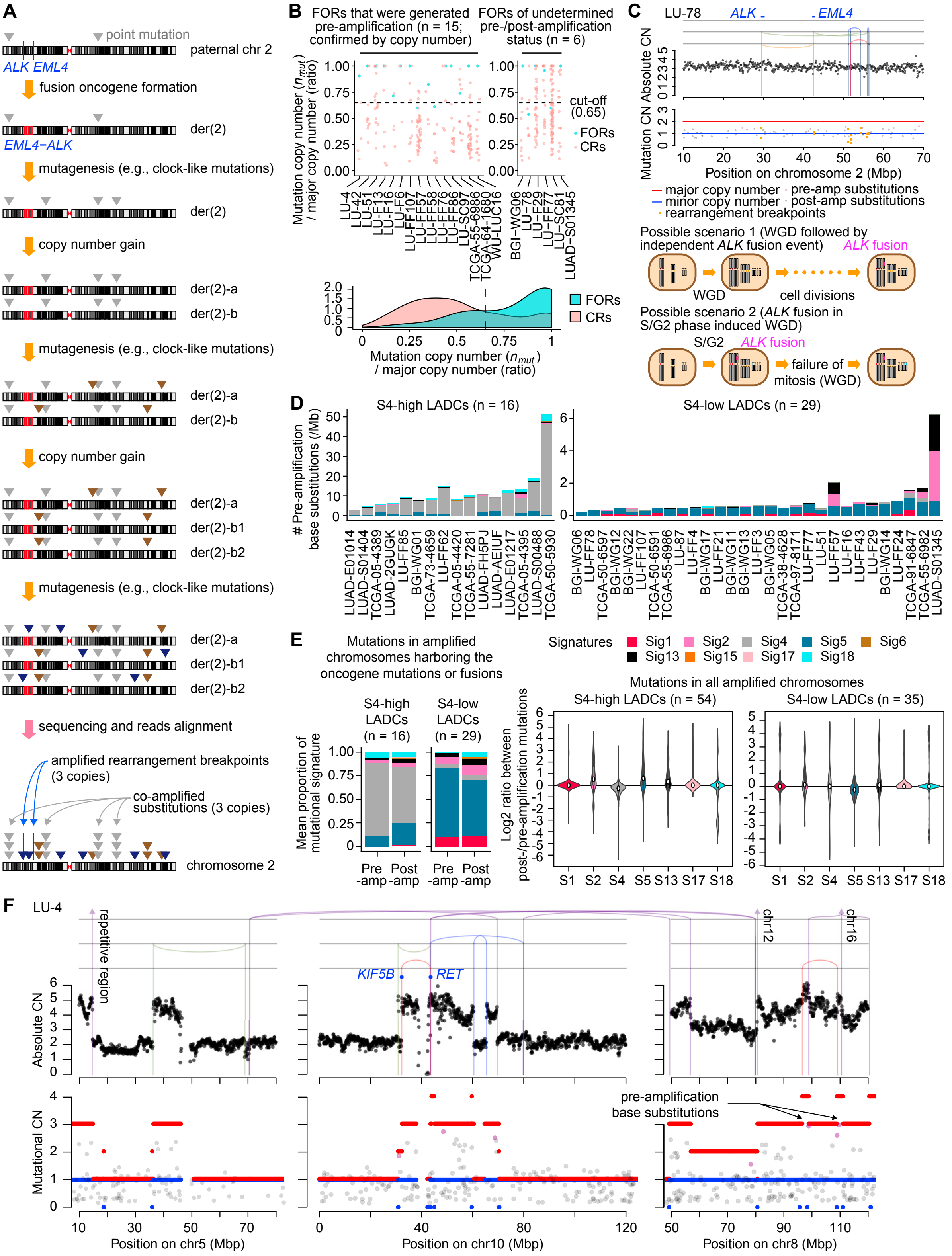


Figure S5. Timing Analysis in LADCs, Related to Figure 5

(A) Graphical summary of mutational timing analysis. We counted the amplified base substitutions to the maximal copy number of the chromosomal segments (i.e., the four positions of stacked gray triangles) to infer the time point of copy number amplification.

(legend continued on next page)

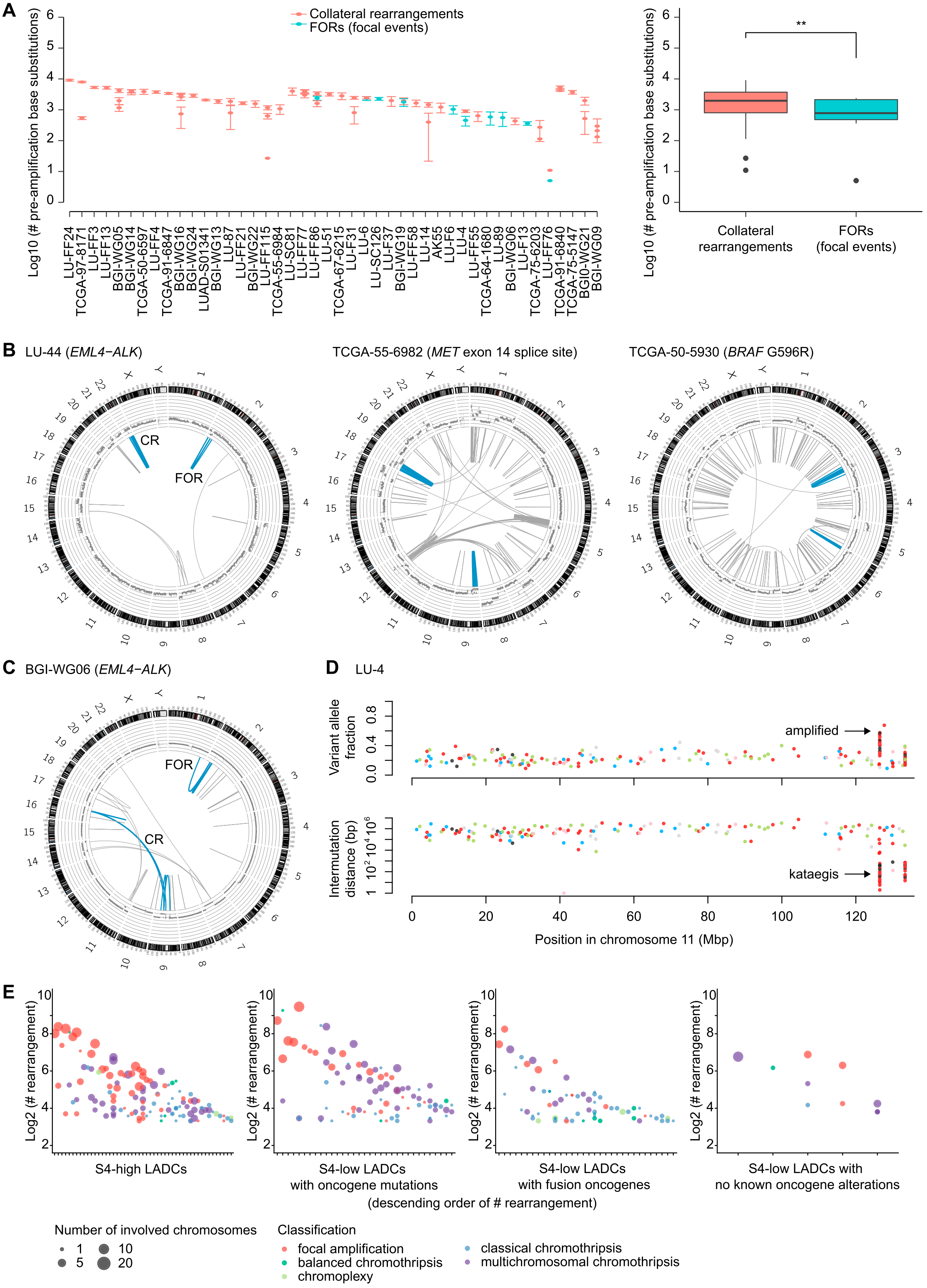
(B) Because of limited accuracy in counting reads at the rearrangement breakpoints, variant allele fractions of rearrangements, and their related parameters (e.g., mutation copy number) were unstable, even for FORs that were confirmed as pre-amplification events based on their copy number profile (upper panel, left graph). Therefore, when we determined the timing of balanced FORs (n = 6, upper panel, right graph), we used a ratio between mutation copy number and major copy number of 0.65. This cut-off accurately detected 13 of 15 pre-amplification FORs in our dataset (sensitivity = 87%; upper panel, left graph), and also separated well the FORs from the CRs in their density distributions (lower panel). In two LADCs (LU-57 and LU-FF55; not shown here), we confirmed that their FORs were in minor alleles.

(C) A post-amplification FOR generating an *EML4−ALK* fusion in LU-78. In this case, the FOR was preceded by a very early WGD. Structure of the FOR (upper panel), and the distribution of mutation copy number of base substitutions and rearrangement breakpoints (middle panel). Since we confirmed that the FOR was linked with the major allele of heterozygous SNP (rs6722025), we conclude that the FOR is present only in 1 copy out of the 2 major alleles. This could be explained by a serial model of WGD followed by *ALK* fusion (lower panel, scenario 1). However, we cannot exclude the scenario of fusion oncogene generation after S phase, and a subsequent failure of mitosis (lower panel, scenario 2).

(D) Contribution of mutational signatures to the pre-amplification mutations of key oncogene loci (oncogene mutations or fusions) in informative tumors from S4-high (n = 16) and -low (n = 29) groups (STAR Methods).

(E) Composition of mutational signatures in pre-amplification and post-amplification mutations across samples (STAR Methods). Comparisons of the local mutational spectra of the key oncogene loci (left panel), and global mutational spectra from all amplified chromosomes (right panel) are separately shown. For the local spectra analysis (left), we selected LADCs with at least 50 pre-amplification mutations (to maximize the number of analyzable samples), and decomposed the mutational spectra using the COSMIC catalog. In S4-high LADCs, the contribution of S4 was significantly greater in pre-amplification mutations, compared to post-amplification mutations (0.77 versus 0.60, p = 0.0052). The contribution of S5 showed the opposite trend. It was greater in post-amplification mutations (0.11 versus 0.23, p = 0.0045). In S4-low LADCs, S5 was higher in pre-amplification mutations compared to post-amplification mutations (0.73 versus 0.60, p = 0.0222). For the global spectra analysis (right), we selected all chromosomes harboring at least 100 pre- and 100 post-amplification mutations, and analyzed their mutational spectra from 786 chromosomes in S4-high LADCs (n = 54), and from 165 chromosomes in S4-low LADCs (n = 35), in the same manner using the COSMIC catalog. In S4-high LADCs, S4 was the only signature higher in the pre-amplification mutations (median ratio of pre-/post-amplification mutations = 0.833), whereas in S4-low LADCs, S5 was the only signature more active in pre-amplification mutations (median ratio = 0.8363), compared to post-amplification mutations.

(F) Complex genomic rearrangement generating a *KIF5B−RET* fusion in LU-4 (upper), the copy numbers of residing base substitutions (dots) and the allele-specific copy number profile (solid lines; lower). The pre-amplified substitutions are shown as purple dots and the post-amplification substitutions as gray dots. CN, copy number.



(legend on next page)

Figure S6. Timing and Structure of Collateral Complex Rearrangements, Related to Figure 6 and STAR Methods

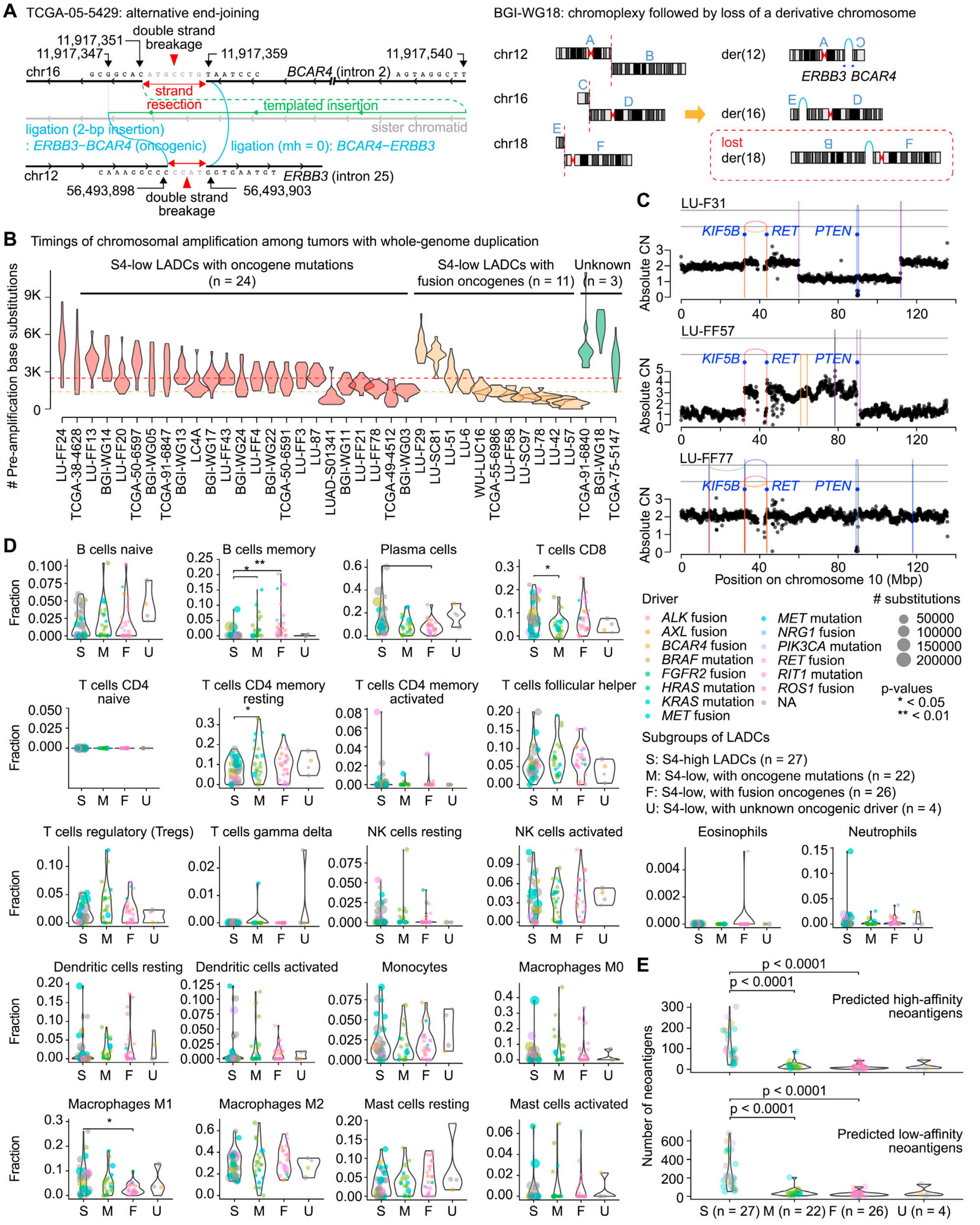
(A) Two plots compare the timing of CRs and FORs. In the left panel, the timing of individual complex rearrangement clusters is analyzed. In the right panel, the timing of CRs and FORs is shown in box plots.

(B) LADCs harboring multiple classical chromothripsis events.

(C) In BGI-WG06 tumor, we observed two independent events of balanced complex rearrangements.

(D) A kataegis event in the vicinity of the rearrangement breakpoint in extrachromosomal amplificon (in LU-4 tumor). Variant allele fraction shows the amplification of kataegis substitutions, indicating their early generation during the extrachromosomal amplification.

(E) Different features of complex genomic rearrangements (involving $\geq$ 10 rearrangements) by LADC subgroups. Size of the dots indicate the number of involved chromosomes, and the color of the dots represent their classes (STAR Methods). S4-high LADCs tend to have more focal amplification events (nearly one amplification event per tumor) compared to the other groups combined (0.98 versus 0.57, $p = 0.0225$). S4-low LADCs with oncogene mutations frequently have large focal amplicons (likely extrachromosomal; left upper side of the plot), whereas the S4-low LADCs with fusion oncogenes tend to have small-sized events (in terms of both number of rearrangements and chromosomes).



(legend on next page)

Figure S7. Recurrent Genetic Alterations in LADCs, Related to Figure 7

(A) Recurrent *BCAR4*−*ERBB3* fusions in two LADCs. In TCGA-05-5429, the FOR was generated from alternative end-joining, involving a templated insertion at the breakpoints from nearby sequences (left panel). In BGI-WG18, the FOR was generated from chromoplexy (right panel).

(B) Timing of chromosomal amplification in LADCs with WGD (n = 38). Some tumors show simultaneous amplifications of a majority of chromosomes (wide violin shapes), while other tumors indicate serial amplifications of chromosomes (narrow, but long violin shapes).

(C) Three LADCs harboring a *KIF5B*−*RET* fusion show common deletion of the *PTEN* tumor suppressor gene, indicating its relevance in this context.

(D) Composition of immune infiltrates by LADC subgroups (STAR Methods). The fractions of immune cell types are compared among the subgroups and the significant p values are indicated at the top of brackets (by Student's t test).; ***p < 0.001; **p < 0.01; *p < 0.05

(E) Number of predicted neoantigens from the mutational profile of LADCs by subgroups (STAR Methods). Data is shown for the predicted high-affinity and low-affinity neoantigens.