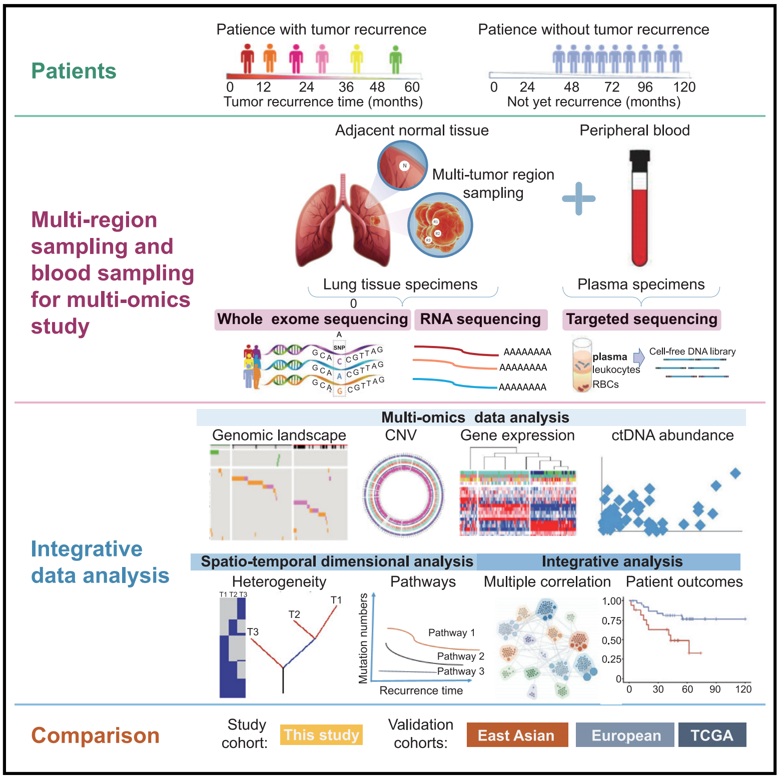
Cell Reports

Article

Spatiotemporal genomic analysis reveals distinct molecular features in recurrent stage I non-small cell lung cancers

Graphical abstract



Highlights

- Multi-sequencing with recurrent timing provides deep insights into early-stage NSCLC
- Asian NSCLCs exhibit diverse genomic features
- Integrative analysis can stratify stage I NSCLC prognosis, verified by ctDNA
- Dysfunction of DNA ICL-DSB events inducing GI contributes to early tumor recurrence

Authors

Kezhong Chen, Airong Yang,
David P. Carbone, ..., Shuangxiu Wu,
Mariam Jamal-Hanjani, Jun Wang

Correspondence

yangfan@pkuph.edu.cn (F.Y.),
bowusx@aliyun.com (S.W.),
m.jamal-hanjani@ucl.ac.uk (M.J.-H.),
wangjun@pkuph.edu.cn (J.W.)

In brief

Chen et al. describe genomic features that differ between Asian and Caucasian NSCLC and molecular features of early-relapse stage I NSCLC based on multiregional tumor sequencing and relapse-timing analysis. They develop an integrative clinical-genomic factor model stratifying stage I NSCLC prognosis and extend genomic insights into early-stage NSCLC.

Chen et al., 2022, Cell Reports 40, 111047
July 12, 2022 © 2022 The Authors.
https://doi.org/10.1016/j.celrep.2022.111047

CellPress

Cell Reports

CellPress
OPEN ACCESS

Article

Spatiotemporal genomic analysis reveals distinct molecular features in recurrent stage I non-small cell lung cancers

Kezhong Chen,[1,2] Airong Yang,[3] David P. Carbone,[4] Nnennaya Kanu,[2] Ke Liu,[3] Ruiru Wang,[3] Yuntao Nie,[1] Haifeng Shen,[1] Jian Bai,[3] Lin Wu,[3] Hui Li,[3] Yanbin Shi,[3] Tony Mok,[5] Jun Yu,[6] Fan Yang,[1,*] Shuangxiu Wu,[3,*] Mariam Jamal-Hanjani,[2,7,8,*] and Jun Wang[1,9,*]

[1]Thoracic Oncology Institute and Department of Thoracic Surgery, Peking University People's Hospital, Beijing 100044, China
[2]Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer Institute, London, UK
[3]Berry Oncology Corporation, Beijing 100102, China
[4]Ohio State University, Columbus, OH 43026, USA
[5]Department of Clinical Oncology, The Chinese University of Hong Kong, Hong Kong SAR, China
[6]University of Chinese Academy of Sciences, Beijing 100101, China
[7]Cancer Metastasis Laboratory, University College London Cancer Institute, London, UK
[8]Department of Medical Oncology, University College London Hospitals, London, UK
[9]Lead contact
*Correspondence: yangfan@pkuph.edu.cn (F.Y.), bowusx@aliyun.com (S.W.), m.jamal-hanjani@ucl.ac.uk (M.J.-H.), wangjun@pkuph.edu.cn (J.W.)

https://doi.org/10.1016/j.celrep.2022.111047

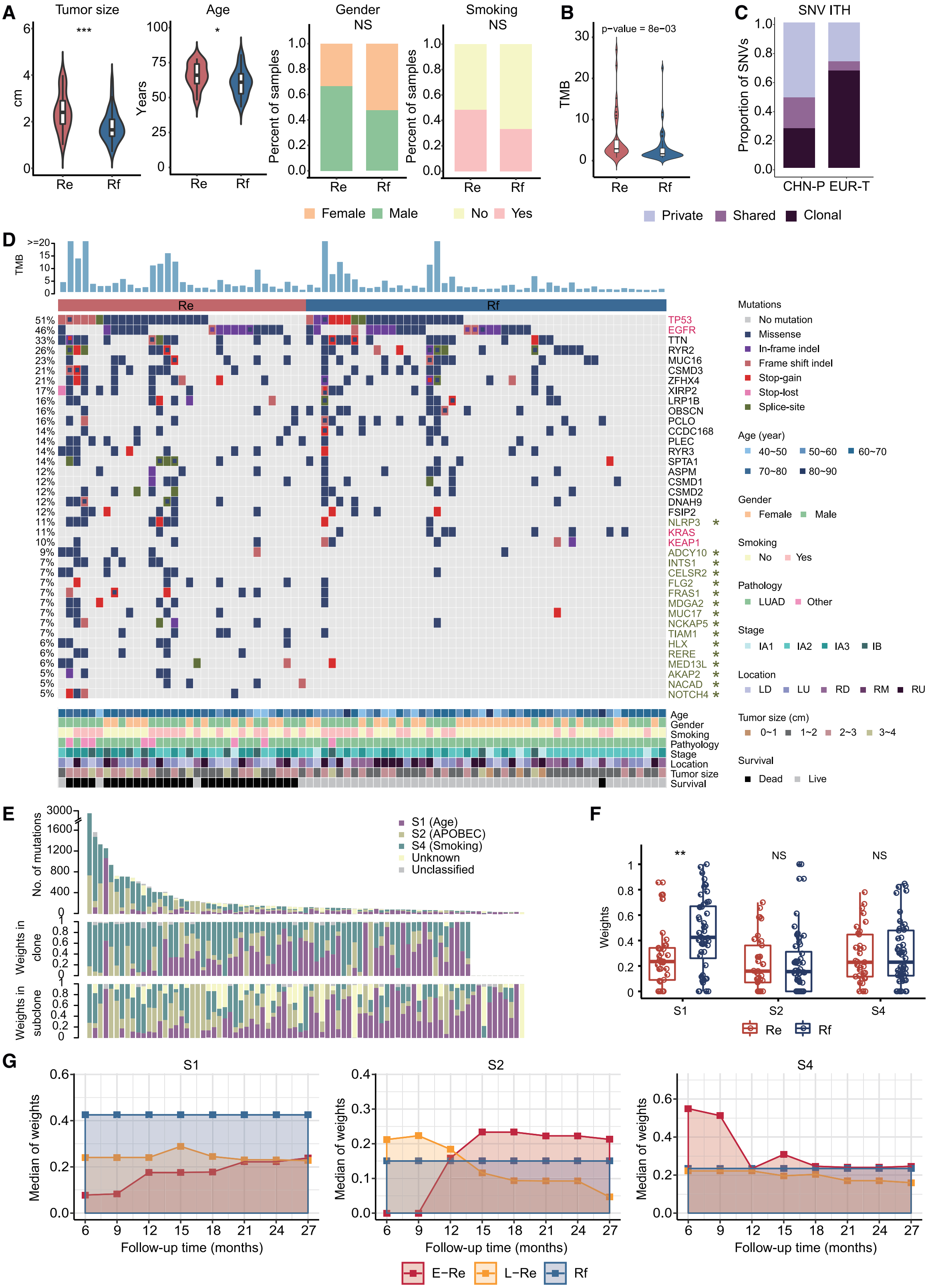
SUMMARY

Stage I non-small cell lung cancer (NSCLC) presents diverse outcomes. To identify molecular features leading to tumor recurrence in early-stage NSCLC, we perform multiregional whole-exome sequencing (WES), RNA sequencing, and plasma-targeted circulating tumor DNA (ctDNA) detection analysis between recurrent and recurrent-free stage I NSCLC patients (CHN-P cohort) who had undergone R0 resection with a median 5-year follow-up time. Integrated analysis indicates that the multidimensional clinical and genomic model can stratify the prognosis of stage I NSCLC in both CHN-P and EUR-T cohorts and correlates with positive pre-surgical deep next generation sequencing (NGS) ctDNA detection. Increased genomic instability related to DNA interstrand crosslinks and double-strand break repair processes is significantly associated with early tumor relapse. This study reveals important molecular insights into stage I NSCLC and may inform clinical postoperative treatment and follow-up strategies.

INTRODUCTION

Non-small cell lung cancer (NSCLC) is the leading cause of cancer death globally (Siegel et al., 2020). Surgery is the curative treatment for early-stage NSCLC, and the postoperative treatment strategy depends mainly on staging of disease according to tumor-node-metastasis (TNM) classification (Goldstraw et al., 2016). Multiple clinical guidelines recommend postoperative adjuvant chemotherapy for patients with resected stage II or III NSCLC, which is associated with a 5% improvement in overall survival (OS), whereas treatment after surgery is neither beneficial nor recommended for patients with stage I NSCLC as a group (Group et al., 2010).

However, approximately 20%–40% of patients with stage I lung cancer experience tumor recurrence (Re) after curative surgery, some of which even recur very shortly after surgery, and which patients will relapse and which will not is currently not predictable based on clinical characteristics alone (Vansteenkiste et al., 2014). Many attempts have been made to develop prognostic biomarkers for recurrence (Devarakonda et al., 2018; Li et al., 2017). Early studies focused on the prognostic value of clinical and pathological characteristics such as surgical procedure and histological subtype; however, these studies failed to explain why the prognosis of patients with similar clinical-pathological features are diversely different (Tsao et al., 2015). Subsequent studies focused on genome-based prognostic signatures, which was limited by the inconsistency and lack of validation. To date, none of these gene signatures have demonstrated clinical utility (Vargas and Harris, 2016). Only recently have next generation sequencing (NGS)-based genomic studies become feasible and focused on the molecular landscape of lung cancer. These studies have revealed that multiple genomic factors, including driver mutations, mutational signatures, pathway analysis, copy-number alterations, genomic heterogeneity, gene expression, and immune infiltration, contribute to the eventual prognosis of a patient (Table S1A). Most of these studies were conducted in Western countries with predominantly smokers and Caucasian subjects (Jamal-Hanjani et al., 2017; Rosenthal et al., 2019). In addition, none of these studies specifically focused on patients with stage I NSCLC. Postoperative



(legend on next page)

management of stage I disease continues to be challenging because there is no specific genomic guide for identification of patients with high risk of recurrence.

In this study, we performed a comprehensive analysis of resected stage I NSCLC by multiregional whole-exome sequencing (WES) with an average depth >500×, RNA sequencing, and plasma-targeted circulating tumor DNA (ctDNA) detection, and correlated the genomic findings with survival after median follow-up of more than 5 years. This analysis was also compared with published genomic datasets from an East Asian (EAS) cohort (EAS LUADs [lung adenocarcinomas]) (Chen et al., 2020), a European cohort (TRAcking Cancer Evolution through therapy [TRACERx]) (Jamal-Hanjani et al., 2017), and an American cohort (The Cancer Genome Atlas [TCGA]).

RESULTS

Patient cohort and sequencing quality statistics

We retrospectively enrolled 81 patients diagnosed with stage I NSCLC with no prior therapy as the Chinese CHN-P cohort. Patients were divided into two groups according to tumor relapse status: Re (n = 33) and recurrence-free (Rf, n = 48) groups, based on a median of 62 months of clinical follow-up until September 2020. The median recurrent time was 15 months (range: 2–62 months). The clinical demographics are summarized in Tables S1B and S1C. Except for dominant LUAD patients (p = 0.01), tumor sizes ($p = 6.0 \times 10^{-4}$) and patient ages (p = 0.01) of the Re group were higher than those of the Rf group, other clinical features did not differ between the two groups (Figure 1A; Table S1B).

Multiregional WES produced an average sequencing depth of 561× (range: 113–1,203×) using a total of 240 tumor tissue specimens and 81 matched adjacent normal tissues (Table S2A). RNA sequencing was also performed for 28 samples from 21 patients and generated an average of 33.1 million clean reads (14.2–45.7 million) per sample (Table S2B).

Significant somatic mutations in Re and Rf stage I NSCLCs

WES data revealed 31,531 somatic non-synonymous (non-silent) single-nucleotide variants (SNVs) or insertion/deletions (Indels), with a median of 115 (range: 30–1,044) mutations per patient for the Re group and 65 (range: 27–2,042) for the Rf group (p = 0.006; Figure S1A; Table S3A). The corresponding median tumor mutation burden (TMB) per patient was 2.89/Mb and 1.74/Mb, respectively (p = 0.008; Figure 1B), which was significantly different between Re and Rf groups. In addition, TMB was found to be significantly higher in patients who had smoked (n = 32/81, 39.5%) than in patients who were non-smokers in both groups ($p = 4.4 \times 10^{-4}$; Figure S1B; Table S3B). The multiregional WES also revealed that about 61.15% of mutations were pure subclonal ones (Tables S3C and S3D), and a high proportion of subclonal SNVs/Indels was observed in Re patients (p = 0.08; left graph of Figure S1C) but did not significantly affect patient disease-free survival (DFS) time (p = 0.052; right graph of Figure S1C). Compared with those patients with stage I NSCLCs in a European cohort of the TRACERx study (EUR-T) (n = 61) (Jamal-Hanjani et al., 2017), the CHN-P cohort had higher intratumor heterogeneity (ITH), characterized by lower clonal but higher shared and private subclonal mutation fractions in both smokers and non-smokers ($p < 0.014$; Figures 1C, S1D, and S1E) and a trend toward lower mutation diversity (p = 0.085; Figure S1F) based on SNVs/Indels. The ITH diagram of each patient is shown in Figures S1G and S1H.

MutsigCV and dNdScv ($q < 0.1$) were used to identify somatic driver mutation genes. *TP53* (61% versus 44%) and *EGFR* (52% versus 42%) were the top two driver genes in both Re and Rf groups. *KRAS* (3% versus 17%) and *KEAP1* (9% versus 10%) were additional driver mutation genes in the Rf group (Figure 1D). All four driver gene mutations had no significant impact on DFS time. *EGFR* mutations occurred frequently in women and patients who were non-smokers (Figures S1I and S1J), while *KRAS* and *KEAP1* mutations were enriched in men and patients who were smokers. *TP53* mutations occurred more frequently in smokers and tumors larger than the median size (Figure S1K).

Forty-one recurrently somatically mutated genes with frequencies >10% were detected (Table S3A), most of which had a higher frequency in the Re group than in the Rf group; however, they did not significantly impact Re or patient prognosis. Sixteen significantly differentially recurrent somatic mutations were identified in the CHN-P cohort based on Fisher's exact test ($p < 0.05$; Figure 1D; Table S3E), which were involved in functions related to signaling transduction, transcriptional regulation, tumor suppression, and cell adhesion or communication. All 16 mutations occurred at higher frequency in the Re group (18% to 12%) than that in the Rf group (4% to 0%). Patients harboring these somatic mutations had significantly shorter DFS times than individuals with wild-type genotypes, elucidating that those diverse somatic mutations promote or accompany lung carcinogenesis and progression.

A further comparison of somatic mutations in genes related to genomic integrity between the CHN-P cohort and the other EAS

Figure 1. Somatic mutation features

(A) Clinical feature comparisons between tumor recurrence (Re) and recurrence-free (Rf) groups. ***$p < 0.001$, **$p < 0.01$, *$p < 0.05$.

(B) Comparison of TMB between patients of Re and Rf groups ($p = 8.0 \times 10^{-3}$).

(C) Comparison of clonal, shared-, and private-subclonal SNV/Indel proportions in the patients between the CHN-P and the EUR-T cohorts.

(D) Somatic mutation profiles. Gene names are shown on the right side of the graph, and the corresponding mutational frequencies are shown on the left side. Potential driver genes are indicated in red color. Asterisks indicate the significantly differential mutation genes in the Re and Rf groups. Tumor mutation burden is shown on the upper panel. Clinical information is listed at the bottom area.

(E) Mutation gene numbers harboring mutational signatures S1, S2, and S4 statistically on total mutations, and their weights in clonal- and subclonal-level mutations in each patient are shown on the top, middle, and bottom graphs, respectively.

(F) Comparison of the weight of each mutational signature between Re and Rf groups.

(G) Dynamic changes of mutational signature weights in patients with different prognosis during follow-up. E-Re or L-Re, tumor recurred earlier or later than the specific time point on the x axis, respectively.

ITH, intratumor heterogeneity; NS, no significant difference. The p value less than 0.05 was defined as significant.

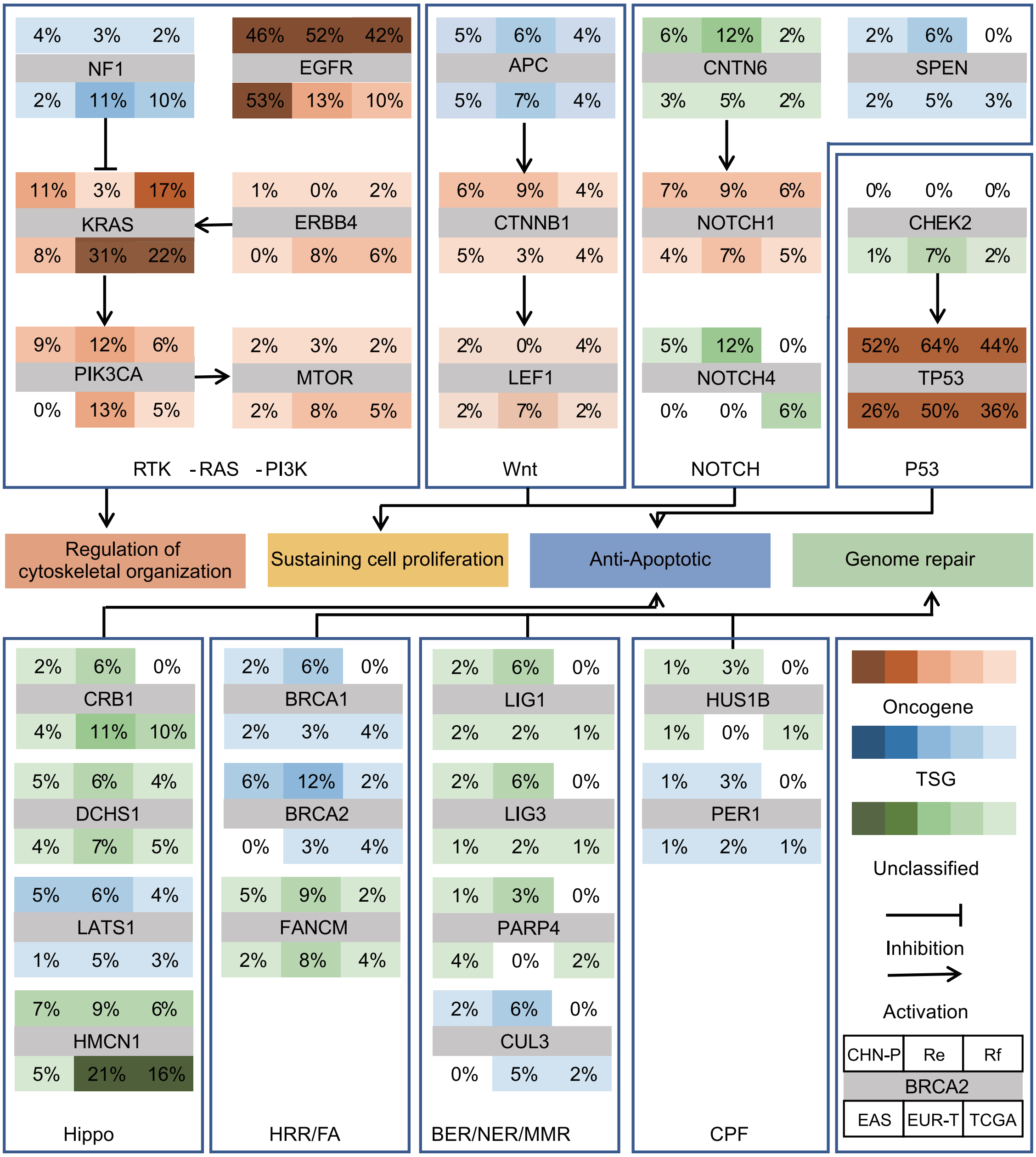
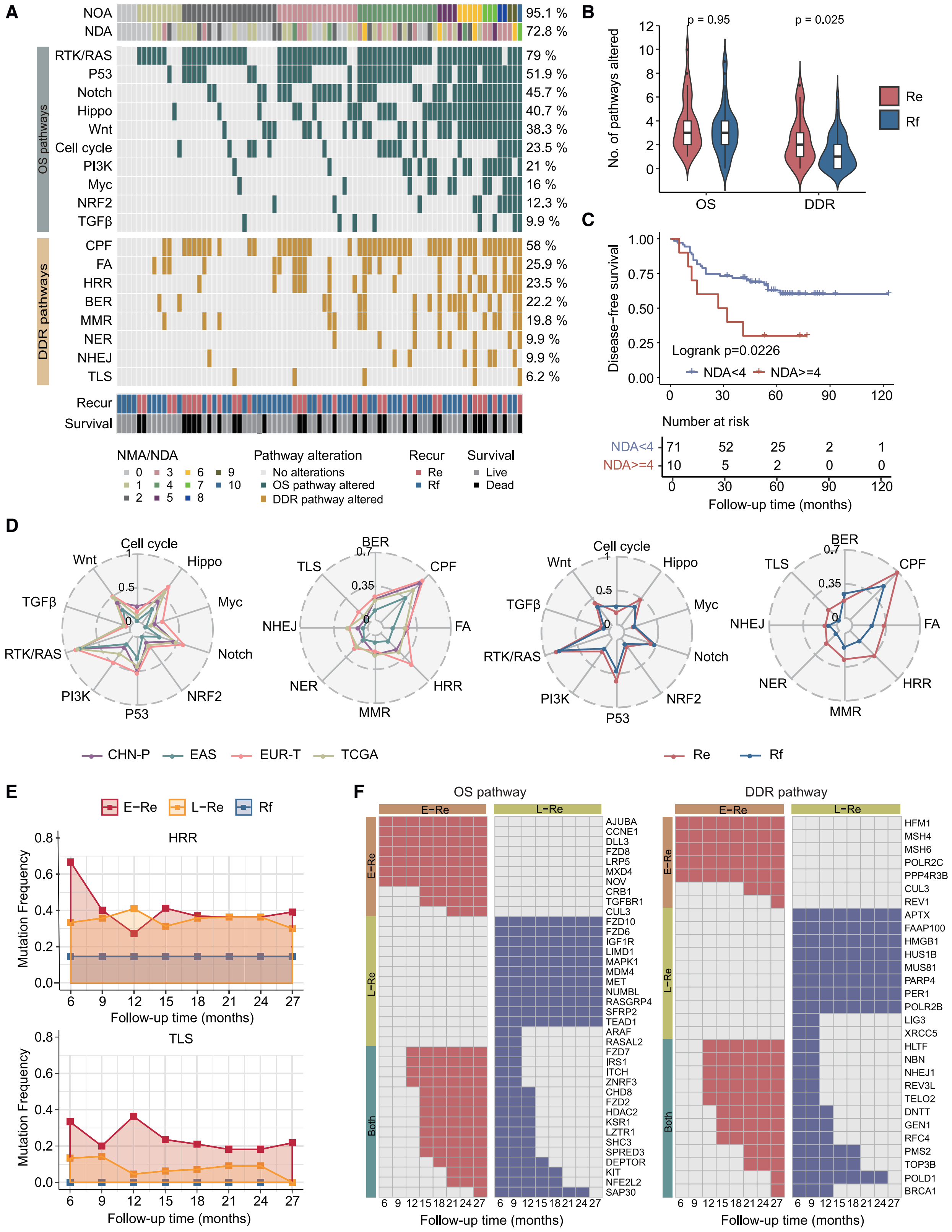


Figure 2. Somatic mutational comparison between the CHN-P cohort and other cohorts

Each gene box includes 6% values representing the mutational frequencies of genes in the CHN-P cohort on the whole (CHN-P), Re and Rf groups, EAS cohort, EUR-T cohort, and TCGA cohort, as shown in the graph. The color scale bar shows the mutation frequency from 0% to 100%. Genes are grouped by signaling pathways related to genome maintenance mechanisms. Interaction between genes is indicated by arrows.
TSG, tumor suppressor gene.

patients (n = 131) (Chen et al., 2020) and the Western stage I NSCLC cohorts (including the EUR-T [n = 61] and TCGA [n = 277] cohorts; Figure 2) identified that *NF1* (4%), *KRAS* (11%), *ERBB4* (1%), and *mTOR* (2%) of the receptor-tyrosine kinase (RTK)/RAS/phosphatidylinositol 3-kinase (PI3K) pathway, *CRB1* (2%) and *HMCN1* (7%) of the Hippo pathway, and *CHEK2* (0%) of the p53 pathway had lower mutation frequencies, while *EGFR* (46%) of the RTK/RAS/PI3K pathway had much higher mutation frequencies in our and EAS cohorts than in Western populations.

BRCA2 (12%), *BRCA1* (6%), and *FANCM* (9%) of the homologous recombination repair (HRR)/Fanconi anemia (FA) pathway were detected with higher mutation frequencies in the Re group compared with those in the Rf group (0%–2%), EAS cohort (0%–2%), and Western cohorts (3%–4% or 4%–8%). The same observations were also made in several genes of the base excision



(legend on next page)

repair (BER)/nucleotide excision repair (NER)/mismatch repair (MMR) and checkpoint factors (CPF) pathways, indicating possible different mutation fragile sites or genetic mechanisms causing lung carcinogenesis and relapse between Chinese and Western populations.

Somatic mutational signatures of stage I NSCLCs

Mutational signatures were *de novo* characterized according to the mutation spectrum, and three highly confident signatures with high similarity to the Catalogue of Somatic Mutations in Cancer (COSMIC) signatures were derived: age-related S1 (92%), activation-induced cytidine deamniase/apolipoprotein B mRNA editing enzyme catalytic polypeptide-related S2 (82%), and smoking-related S4 (93%) (Figures S2A and S2B). These three mutational signatures were also predominant in stage I NSCLCs of the EAS, EUR-T, and TCGA cohorts.

The number of patients with S1, S2, and S4 showed no significant difference between the Re and Rf groups (Figure S2C). However, the weight of age-related S1 predominated in clonal mutations (Figures 1E and S2D) and significantly accumulated in the Rf tumors compared with the Re tumors (p = 0.002; Figure 1F). The prevalence of S1 generally remained at low levels (median = 0.23) in patients with early recurrence (Line E-Re) in the dynamic analysis of mutational signature changes according to the relapse time of patients (Figure 1F). APOBEC-related S2 prevalence was higher in subclonal mutations (median = 0.17) than in clonal mutations (median = 0.08) (Figures 1E and S2D), suggesting that the timing of APOBEC mutagenesis was relatively late and induced a subclonal driver event. In contrast, smoking-related S4 dominated in clonal mutations (median weight = 0.3) (Figures 1E and S2D) and in some early Re (E-Re) patients (tumors recurred within 9 months of Line E-Re) (up to 0.51) (Figure 1G), implying its contribution to lung carcinogenesis and potential impact on particularly poor prognosis tumors. Further correlation analysis with clinical features also demonstrated that smoking-related S4 had a significant positive relationship with smoking and TMB features and was associated with the male gender (Figure S2E).

Somatic mutation-enriched pathways drive Re

Tumorigenesis, genomic evolution, and corresponding clinical phenotypes are driven by a group of frequently altered carcinogenesis pathways, including 10 canonical oncogenic signaling (COS) pathways (involving 335 genes), which cover 89% of 9,125 different solid tumor types as reported by TCGA PanCancer Atlas Project (Sanchez-Vega et al., 2018), and DNA damage repair (DDR) pathways (involving 233 genes), which have a considerable influence on genomic instability and drug resistance (Scarbrough et al., 2016; Wang et al., 2018). In the CHN-P cohort, alterations in COS pathways (95.1%) and DDR pathways (72.8%) were pervasive (Figure 3A; Table S4). For the 10 COS pathways, the most frequently altered pathway was the RTK/RAS/MAPK pathway (79.0%), followed by the p53 pathway (51.9%). For the eight DDR pathways, the most frequently altered DDR pathway was the CPF (58.0%) pathway, particularly focusing on *TP53* mutation (n = 41/47, 87.2%). Both numbers of COS pathway alterations (NOA) and DDR pathway alterations (NDA) positively correlated with TMB (p = 8.3×10^{-9} and 1.5×10^{-10}; Figures S3A and S3B); however, only NDA occurred significantly more in the Re patients (n = 28/33, 85.8%) in comparison with those in the Rf patients (n = 31/48, 64.6%) (p = 0.025; Figure 3B). When DDR pathway alterations occurred in four or more pathways (NDA $\geq$ 4) in each patient, they significantly affected DFS time (p = 0.023; Figure 3C), whereas NOA did not (p = 0.156; Figure S3C).

The comparison of COS and DDR pathway alterations in stage I patients between the CHN-P cohort and the other three cohorts demonstrated that similar pathway alteration patterns existed in the EAS and CHN-P cohorts, which were different from that of the EUR-T cohort (left two graphs of Figure 3D), while the pathway pattern of TCGA cohort was intermediate between the CHN-P cohort and the EUR-T cohort, possibly because of the inclusion of the 62 non-Caucasian patients. Pathway alteration numbers in the CHN-P cohort were more abundant than those in the EAS cohort, which was most likely caused by multiregional sampling in our study. Due to the lack of DFS information in the EAS and TCGA cohorts, we analyzed only the impact of pathway alterations on Re events in our and EUR-T cohorts. Alterations in translesion synthesis (TLS) and HRR pathways were significantly more frequent in the Re patients of the CHN-P cohort (p = 0.009 and 0.017, respectively) (right two graphs of Figure 3D) than in the Rf patients; however, no pathway alteration significantly differed between the two groups in the EUR-T cohort (Figure S3D). TLS is a fundamental mechanism to guarantee DNA replication across bulky barriers on DNA templates to prevent DNA double-strand breaks (DSBs) by specialized TLS DNA polymerases, including *Rev1*, *Rev3L*, and *HLTF* genes identified in the Re patients (n = 5), and all were subclonal (Table S4A). Loss-of-function mutations of these genes have been reported to increase frequencies of chromosome breaks and translocations, leading to genomic instability (Zafar and Eoff, 2017). The HRR pathway is also critical in repairing DNA DSBs and maintaining genome stability (Knijnenburg et al., 2018). The two most frequently mutated genes in the

Figure 3. Oncogenic signaling and DNA damage repair (DDR) pathway alteration analysis

(A) Statistical numbers of canonical oncogenic signaling (COS) pathway alterations (NOA) and DDR pathway alterations (NDA) of the CHN-P cohort. The pathway name is on the left and the corresponding alteration frequency is on the right of the graph.

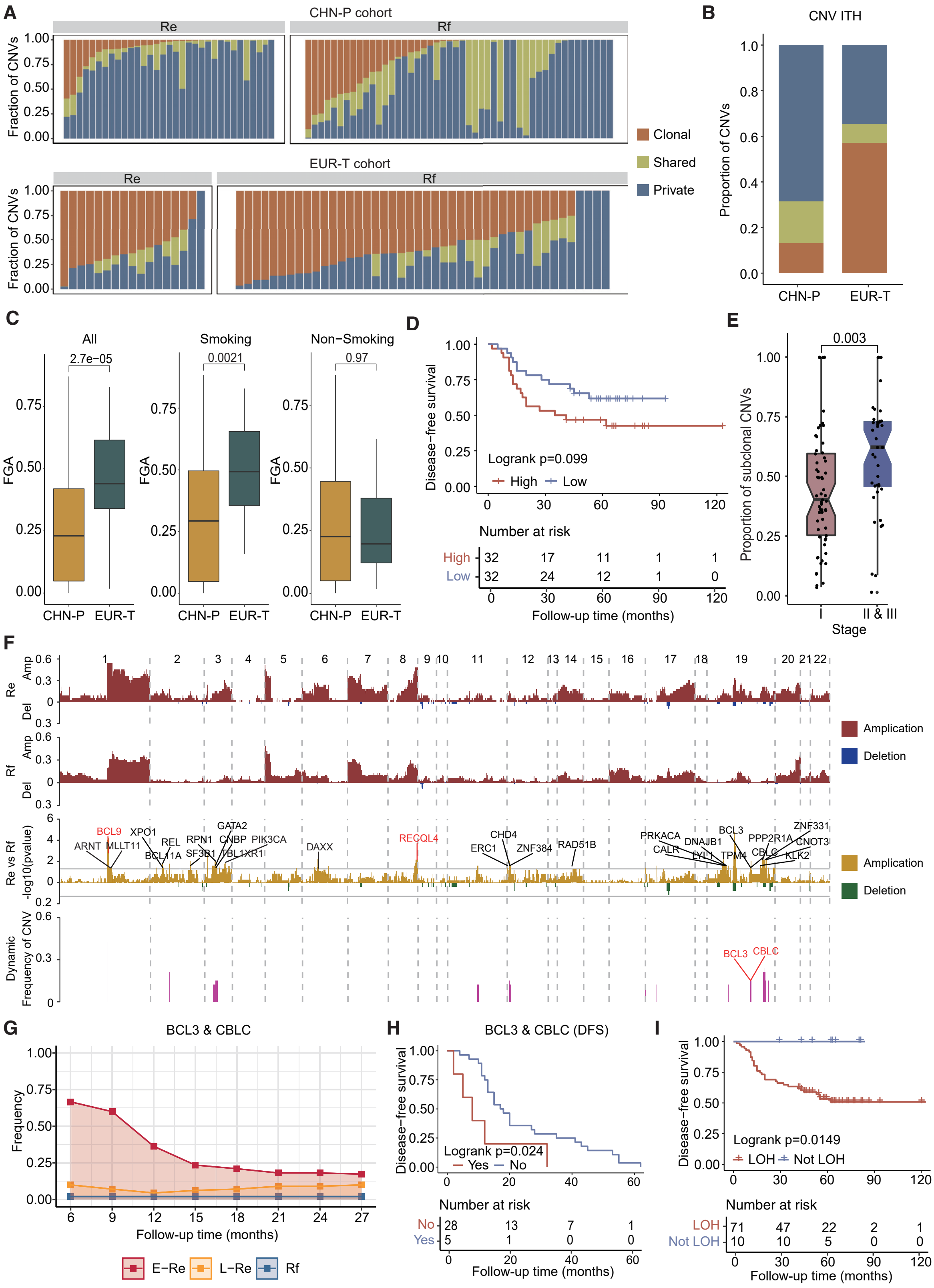
(B) Comparison of OS and DDR pathways between Re and Rf groups of the CHN-P cohort.

(C) Impact of NDA on patient disease-free survival (DFS) time. p = 0.0226 when NDA accumulates up to or more than four pathways.

(D) Comparison of mutation frequency of each pathway among CHN-P, EAS, EUR-T, and TCGA cohorts, as well as between the Re and Rf groups in the CHN-P cohort. Mutation frequency of HRR (p = 0.017) and TLS (p = 0.009) pathways significantly differed between the Re and Rf groups.

(E) Dynamic analysis on mutation frequencies of HRR and TLS pathways in the E-Re, L-Re, and Rf patients over a follow-up survey as shown on the x axis.

(F) Frequently mutating genes detected during dynamic analysis of canonical OS and DDR pathway alterations. Gene names are listed on the right of the graph. Both refer to the genes detected in both E-Re and L-Re types of tumor. The p value less than 0.05 was defined as significant.



(legend on next page)

HRR pathway were *BRCA2* (12%) and *BRCA1* (6%) in the Re patients (Figure 2; Table S4A), followed by 11 other mutated genes (including five helicase and three nuclease genes) at low mutation frequencies (<6%) (Table S4A).

Analysis of changes in the mutation frequency of each pathway among patients with different recurrence timing during follow-up revealed consistently higher mutation frequencies of TLS, HRR (Figure 3E), and FA, NER of DDR pathways, as well as Myc, PI3K, Hippo, p53, NRF2, and TGF-β of COS pathways in the E-Re patients than in either the late-Re (L-Re) or Rf patients (Figure S3E). The corresponding carcinogenesis-related gene mutations identified in the E-Re tumors are listed in Figure 3F (those within 15 months of E-Re and both groups; Table S4A), indicating their possible involvement in E-Re events. Among them, Myc, PI3K, Hippo, p53, NRF2, and TGF-β pathways regulate cell proliferation, growth, and apoptosis. TLS, HRR, FA, and NER pathways commonly act alone or synergize to repair DNA DSBs and complex lesion events, especially caused by DNA interstrand crosslinks (ICLs) during DNA replication at synthesis (S) phase of the cell cycle (Kass et al., 2016; Niraj et al., 2019). TLS pathway alterations had a significant impact on DFS time among all the patients ($p = 1.4 \times 10^{-7}$) (Figure S3F) and among the recurrent patients ($p = 0.03$) (Figure S3G), implying the potential roles for the TLS pathway and the corresponding gene mutations in E-Re.

Copy-number variations (CNVs) and focal driver genes related to cell proliferation in recurrent stage I NSCLCs

Clonal CNVs were defined as those detected in all the tumor regions of each patient, whereas subclonal CNVs were defined as those not detected in all tumor regions. Although the proportion of stage I patients harboring clonal CNVs in the EUR-T cohort (n = 54/61, 88.5%) and in the CHN-P cohort (n = 64/81, 79.0%) was similar (Figure 4A), a distinct genomic feature (Figure 4B), characterized by a significantly lower total fraction of genome altered (FGA) ($p = 2.7 \times 10^{-5}$; Figure 4C) and clonal CNV fraction ($p = 6.3 \times 10^{-14}$) but remarkably higher private-subclonal CNV fraction ($p = 1.2 \times 10^{-7}$), was detected in the CHN-P cohort compared with the EUR-T cohort (Figures S4A and S4B). Notably, the smokers (n = 32/81, 39.5%) in the CHN-P cohort exhibited consistent FGA (Figure 4C) and subclonal CNV proportions (Figure S4C) in comparison with those of the EUR-T cohort. On focal-CNV gene levels of the CHN-P cohort (Figure S4D), only 18% were clonal, 20% were shared-subclonal, and up to 62% were private-subclonal, revealing that the timing of CNV events was later in the CHN-P cohort compared with those in the EUR-T cohort.

Comparing the Re and Rf groups of the CHN-P cohort, FGA of CNV, clonal/subclonal CNV length, or percentage of clonal/subclonal-CNV length (Figures S4E–S4G) did not reach a significant level. Moreover, despite presenting a modest trend, the CNV ITH did not affect the DFS time for stage I NSCLCs in either the CHN-P ($p = 0.099$; Figure 4D) or EUR-T ($p = 0.51$; Figure S4H) cohort, suggesting that the impact of chromosomal instability on the clinical outcomes of patients in very early-stage NSCLC was in a kainogenesis stage. By comparing the subclonal CNV percentage between stage I and II–III patients of the EUR-T cohort, it was confirmed that stage II–III patients had significantly higher CNV heterogeneity ($p = 0.003$; Figure 4E).

At the chromosomal arm levels, a significant number of amplifications rather than deletions were detected in stage I NSCLCs in the CHN-P cohort (Figure 4F). Particularly for the Re patients, arm-level gains focused on 1q, 2p/q, 3q, 5q, 6p, 8q, 11q, 12p, 17p, and 19p/q, corresponding to 777 amplified genes annotated ($p < 0.05$; Figure 4F; Table S5). Among them, amplified 1q, 2p, 3q, 8q, and 12p have been previously reported in Chinese NSCLCs (Hu et al., 2019; Wu et al., 2015; Zhang et al., 2019). Twenty-nine CNV-related driver genes were identified by comparing with the COSMIC database. These events were enriched in the arm regions of 1q, 2p, 3q, 5q, 6p, 8q, 12p, and 19p/q, such as highly recurrent ($n \geq 10$ in the CHN-P cohort) and large-frequency difference between Re and Rf patients, a DNA replication-involved RecQ helicase gene *RECQL4* (8q24.3), and a transcriptional factor *BCL9* (1q21.2), whose functions are all related to cell proliferation and their amplification all had a significant influence on DFS ($p < 0.01$; Figure S4I).

Further analysis of CNV profiling according to the Re time revealed 265 CNV-amplified genes, including two frequently amplified driver genes closely correlated with E-Re events ($p < 0.05$) (bottom graph of Figure 4F; Table S5C). The two genes were transcription co-activator *BCL3* and cell signaling transduction gene *CBLC*, both as a proto-oncogene candidate on 19q13.32 and promoting tumor cell proliferation or migration (Turnham et al., 2020) (Figure 4G), which was significantly associated with DFS ($p = 0.024$; Figure 4H), suggesting that chromosomal amplifications on certain focal genomic loci may play an essential role in Re.

Genome doubling (GD) events and genome instability (GI) have been reported to be positively associated with lung carcinoma development from stages I to IV (López et al.,

Figure 4. Somatic copy-number variations and focal gene analysis

(A) Comparison of CNV-level intratumor heterogeneity between the CHN-P and EUR cohorts.

(B) Comparison of clonal, shared-, and private-subclonal CNV proportions in the patients of the CHN-P cohort and those in the EUR-T cohort.

(C) Significantly lower proportion of total fraction of genome altered (FGA) with CNVs in all patients ($p = 2.7 \times 10^{-5}$) and smoker patients ($p = 2.1 \times 10^{-3}$) of the CHN-P cohort than that in the EUR-T cohort.

(D) Impact of intratumor heterogeneity of CNV levels higher and lower than the median value on patient DFS time in the CHN-P cohort ($p = 0.099$).

(E) CNV-level intratumor heterogeneity in stage I NSCLCs compared with that in stage II–III NSCLCs of the EUR-T cohort.

(F) Significantly focal CNV amplifications across chromosomes 1–22 between Re and Rf groups and highly frequent CNV-driver genes significantly enriched in the Re group. Highly frequent CNV-related driver genes were significantly marked with red color between Re and Rf groups, as well as between patients with E-Re and L-Re patients.

(G) Dynamic analysis revealed that focal *BCL3* and *CBLC* amplifications (19q13.31–32) more frequently occurred in the E-Re patients than in L-Re and Rf patients.

(H) Significant impact of CNV-amplification driver genes *BCL3* and *CBLC* on patient DFS time.

(I) Significant influence of loss of heterozygosity (LOH) on patient DFS time. The p value less than 0.05 was defined as significant.

2020). Chromosomal instability events reportedly occurred as early as atypical adenomatous hyperplasia and adenocarcinoma *in situ* (Chen et al., 2019a; Hu et al., 2019). However, EAS LUADs have been reported to have a lower percentage of GD, ploidy, or percentage of genome altered (PGA) compared with those of the EUR cohort (Chen et al., 2020). In the CHN-P cohort, about 71.6% of patients (n = 58/81) were found to have GD events in at least one tumor region, and 27.6% (n = 16/58) were clonal events (Figure S4J). The numbers of patients with GD between the Re and Rf groups (p = 0.32) and the impact of GD on DFS (p = 0.2; Figure S4K) did not reach significant levels. Loss of heterozygosity (LOH) was found in all Re patients and 79% of Rf patients (n = 38/48), with a significant difference between the two groups (p = 0.005; Figure S4L), which influenced DFS time (p = 0.015; Figure 4I). However, fractions of the genome with LOH events between Re and Rf groups (p = 0.25; Figure S4M) and the influence on patient DFS time were not significant (p = 0.09; Figure S4N). Overall, for stage I NSCLCs, the impact of large-scale chromosome instability on patient prognosis did not reach significant levels; however, certain focal CNV regions/genomic loci and the corresponding driver genes, particularly for those related to cell proliferation regulation, were very important in E-Re versus L-Re.

Integrative analysis of multiple features related to Re

To evaluate the importance of each feature contributing to Re events, we included genomic features with a combination of clinical features that presented critical effects on Re ($p < 0.2$) and had little correlation with each other in previous analysis. These multiple features could be assigned into three groups: clinical features (including age, pathological type, tumor size, smoking, gender), molecular features (including TLS and HRR pathways, SNV clonality), and chromosomal instability (including focal CNV-related driver genes [$n \geq 10$] of *BCL9* and *RECQL4*, FGA, and LOH).

The integrative analysis of 16 features showed their correlation network structure in the CHN-P cohort (Figure 5A). Correlations among age, tumor size, SNV clonality, FGA, LOH, and ploidy were significant with each other. Figure 5B shows the weight of each feature in the multivariate model. Molecular features were found to be the strongest predictors, accounting for about 50.1% of weight, followed by clinical features (31.6%) and chromosomal instability (18.2%), similar to corresponding weights (31.7% clinical feature, 41.0% molecular, and 27.3% chromosomal instability) when using the top three features: tumor size, TLS pathway, and *BCL9* amplification sorted by $p < 0.0001$ (p = 0.26; Figure S5A). According to the median values of predicted hazard ratio from the multivariate Cox model, the patients could be partitioned into two survival groups: high risk and low risk, with quite differing prognoses based on the multiple genomic feature model (Figure 5C). This model was validated using the EUR-T cohort (the only multiregion cohort that had DFS data), as well as the LUAD patient subgroups of the CHN-P and EUR-T cohorts, and all presented good prognostic discrimination in stage I patients (Figures 5D and S5B). The area under curve (AUC) values of ROC curves validated that the model with a combination of genomic and clinical features together presented the best stratification performance (Figure S5C). Thus, the above integrative analysis across clinical and genomic features presented good prognostic discriminations for patients with Re of stage I NSCLC, underscoring the potential application of these genomic features to predict high-risk Re patients who may need to receive adjuvant therapy even if they are just stage I patients.

Plasma ctDNA profiling implied early metastasis and recurrence

ctDNA has proven to be a reliable biomarker in monitoring minimal residual disease (MRD) (Abbosh et al., 2017; Chaudhuri et al., 2017); however, few previous studies have specifically focused on stage I NSCLC, and correlations of ctDNA from blood taken before surgery and prognosis are getting more attention. We performed deep 457-gene-targeted NGS on plasma cell-free DNAs (cfDNAs) collected before radical tumor resection as described in the STAR Methods. Among the 55 plasma-qualified patients, 19 and 7 patients with ctDNAs were detected in blood plasma using both tumor-naive and tumor-informed methods, respectively (Figure S5D; Table S6). Moreover, mutation frequencies (variant allele frequency [VAF]) of tumor-naive ctDNA (Figure 5E) and tumor-informed ctDNA (Figure S5E) were higher in Re patients than those in Rf patients. Half of the mutations were clonal (Figure S5F), and most of these ctDNAs are associated with high cancer cell fraction (CCF > 0.5), no matter whether they are detected in one, two, or three tumor regions in both patient groups (Figure S5G). Positive ctDNA detection showed a significant association with shorter DFS time (p = 0.02 for tumor naive, Figure 5F; p = 0.0003 for tumor informed, Figure S5H) and positively correlated with recurrence risk predicted by the above-stated multiple-feature model ($p < 0.05$; Figures 5G and 5H). Taken together, this result validated the prediction effect of the model from another dimension and provided an alternative method to predict high-risk patients from blood, especially in case it is unable to do tumor tissue WES.

GI correlated with higher Re risk

The integrated prediction model suggested GI played essential roles in the Re of the CHN-P cohort (Figure 5B). We noted the patients who recurred within 15 months (median recurrent time) (n = 12/17, 70.6%) and harbored mutated genes related to DNA ICL-DSB and complex lesion repairs, i.e., core genes of TLS, HRR, and FA pathways (Table S4B), were significantly more than the ones who recurred after 15 months (n = 2/16, 12.5%; p = 0.038) and Rf ones (n = 7/48, $p = 8.89 \times 10^{-5}$; Figure 6A). Notably, the patients harboring DSB-related gene mutations were more likely to be male, smokers, and *TP53*-mutation patients and less likely to be *EGFR*-mutation patients (Table S4C). We further calculated GI scores of each patient (including the core gene mutations related to DNA ICL-DSB of the three pathways, TMB, FGA, LOH, SNV, and CNV clonalities) and found the GI scores had a significantly negative correlation with shorter DFS time ($R = -0.49$, $p = 3.8 \times 10^{-6}$; Figure 6B) and a positive correlation with recurrence risk predicted by the multiple-feature model ($R = 0.63$, $p < 2.2 \times 10^{-16}$; Figure 6C). Consistent correlation results were also observed in the Re group ($p < 0.02$), but not in the Rf group ($p > 0.05$) (Figures S6A and S6B). The difference of GI scores between E-Re and L-Re,

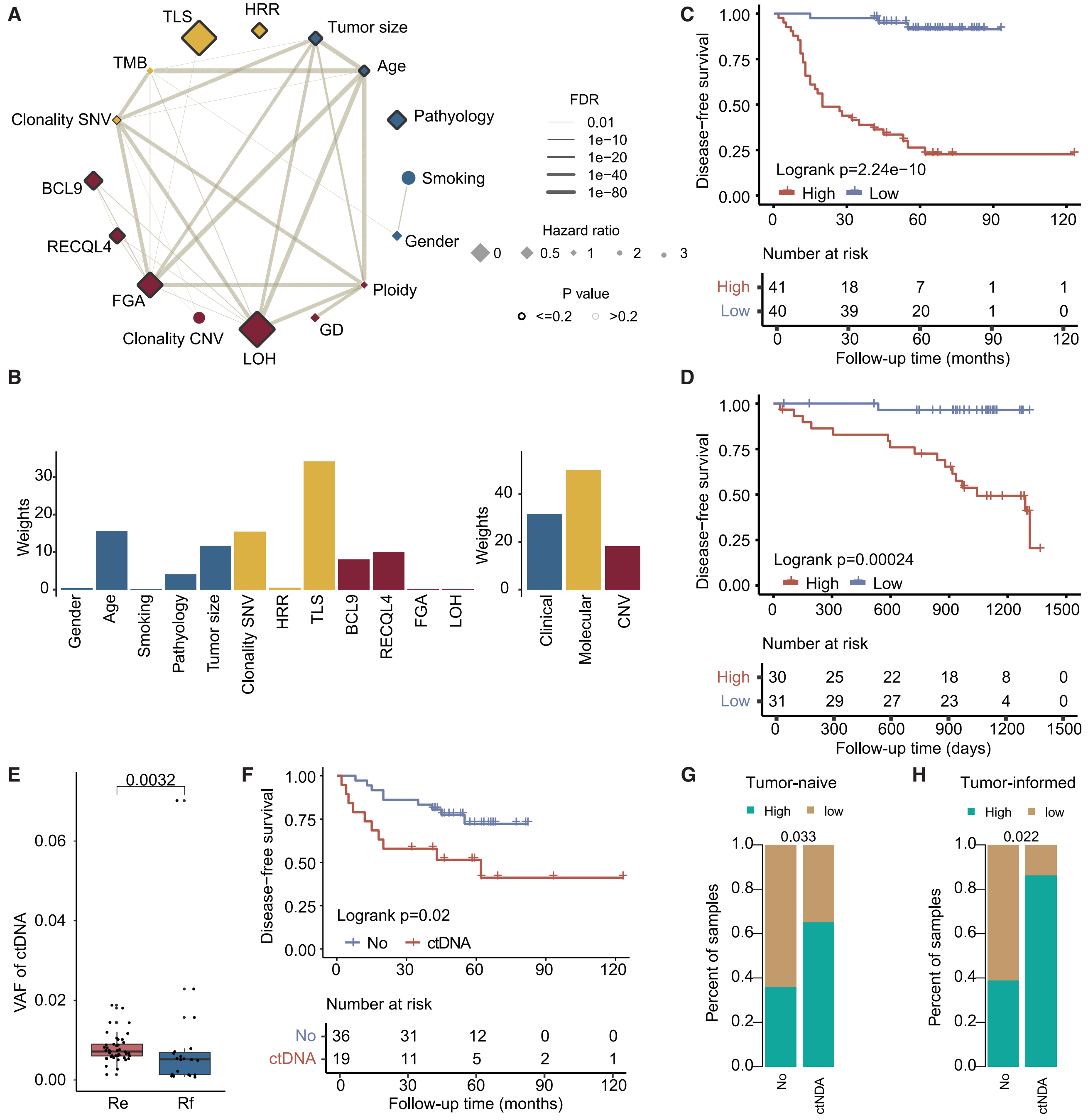


Figure 5. Integrative analysis of multiple features in a multidimensional system to predict Re events and correlation of circulating tumor DNA (ctDNA) in plasma and Re

(A) Correlation analysis showed prognosis importance for 16 features in the CHN-P cohort. Node sizes and borders denote the hazard ratios and statistical significance in the univariable Cox model, respectively. Lines in the circle denote the connection of significantly correlated features (false discovery rate [FDR] q < 0.01), and line thickness represents FDR q values.

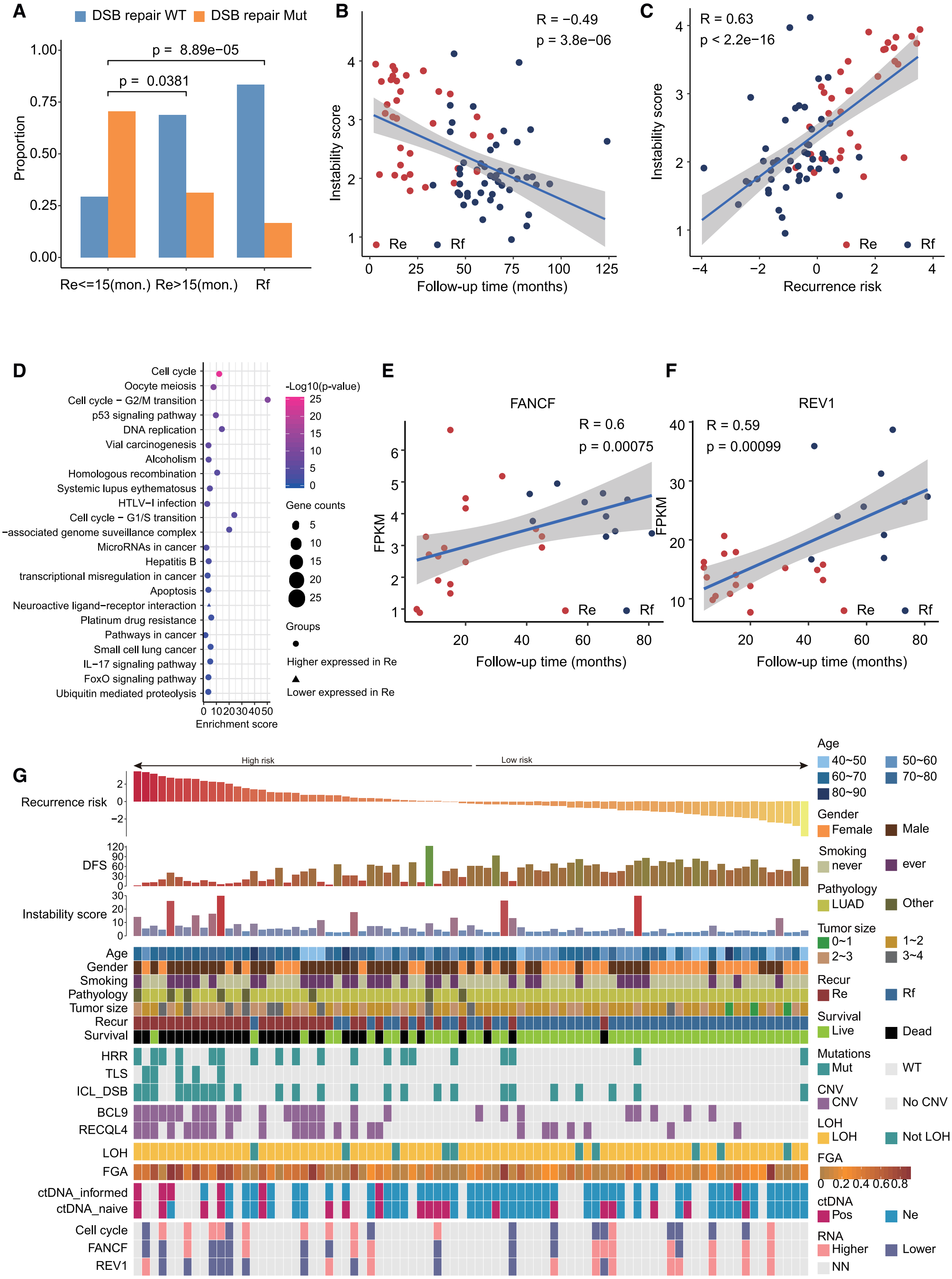
(B) Left graph: feature importance in the multivariate Cox model (expressed with percentages of the Wald statistic) for the patients in the CHN-P cohort. Right graph: important summary for each feature group.

(C and D) Kaplan-Meier survival curves stratify patients with high and low risks of Re according to the median values of predicted hazard ratio from the multivariate Cox model of (B). (C) CHN-P cohort. (D) EUR-T validation cohort.

(E) Comparison of plasma ctDNA abundance (expressed as VAF) between Re and Rf groups, detected using a tumor-naive method.

(F) Comparison of DFS between patients with and without ctDNAs detected in plasma samples using a tumor-naive method.

(G and H) Correlation analysis of ctDNA-positive events in plasma and Re risk predicted in the multivariate model. ctDNA was calculated using tumor-naive (G) and tumor-informed (H) methods, respectively. The p value less than 0.05 was defined as significant.



(legend on next page)

E-Re and Rf, and L-Re and Rf groups all reached significant levels ($p < 0.024$; Figure S6C) and also had a trend between the patients with positive and negative ctDNAs (Figure S6D). In the EUR-T cohort, which includes all stages, the molecular and genomic features of high-recurrence-risk stage I patients were close to those of stage II/III patients, higher than those of the low-risk stage I patients of the recurrence groups (Figure S6E), indicating these molecular alterations leading to poor prognosis may occur earlier than the progress of clinical tumor stage.

On gene expression levels, we found upregulation genes in Re patients (Re-up) were significantly enriched in cell-cycle pathways, p53 signaling, DNA replication, viral carcinogenesis/human T-lymphotropic virus type 1 infection, HR, and other cancer-related signaling transduction and regulation pathways, while downregulation genes in Re patients (Re-down) were mapped only to the neuroactive ligand-receptor interaction pathway ($p < 0.05$; Figure S6F; Table S7), implying that the cell-cycle pathway, which was closely related to DNA replication stress, was significantly upregulated in Re patients. In addition, transcriptions of *FANCF* (a core gene of FA core complex) and *Rev1* of the TLS pathway in Re patients were significantly lower than those in Rf patients ($p < 0.0008$; Figures 6E and 6F). Nine patients with high hazard risk predicted by the above multivariate model harbored upregulation of cell cycle and downregulation of FA/TLS core-gene transcriptions (n = 9/12), but none in the patients with low recurrence risk (n = 0/10) ($p = 0.0005$; Figure 6G). Therefore, we supposed that coincidences of upregulation of cell-cycle pathways and dysfunctions of DNA ICL-DSB repair function played important roles in GI and could indicate a poor prognosis.

DISCUSSION

The multiregional samples in the CHN-P cohort enabled us to be the first to compare genomic heterogeneity between EAS and Western lung cancer patients. Our work demonstrated a significant difference in ITHs between the CHN-P cohort and EUR-T cohort and identified distinct genomic features of high alteration frequencies of the DDR system observed in the recurrent tumors of the CHN-P cohort, particularly significantly involving the genes of the HRR and TLS pathways. Furthermore, the low total FGA and clonal CNVs but high subclonal CNVs (Figures 4A–4C) and SNVs/Indels (Figure S1D) detected in the CHN-P cohort revealed relatively lower overall GI in Asian patients than that in Western patients. Such a high percentage of subclonal alterations also reflected that late events during tumor evolution in Chinese lung cancer patients were different from those of the EUR-T cohort, which reflects genomic ancestry differences between Asian and Western patients or relatively simpler causes (smoking) of pathogenesis in Western patients than in Asian patients. The high subclonal SNVs also indicated the importance of multiregional sampling in Asian patients, which is lacking in previous Asian cohort genomic cancer studies.

Multiple studies have attempted to infer prognostic factors from either clinicopathological or genomic features. However, none of these studies specifically focused on stage I patients, although these tumors have a particular clinical need for risk stratification because there is no standard adjuvant therapy. Our study found that there are indeed differences between stage I and more advanced patients. For example, in the EUR-T cohort, stage I–III NSCLC showed elevated CNV heterogeneity paralleling an increased risk of Re or death, but no significant impact of CNV heterogeneity on prognosis was observed when patients with only stage I NSCLC were examined (Figure S4H). A similar result was also observed in our CHN-P cohort (Figure 4D). Therefore, although ITH has a certain influence on prognosis, it is not the predominant factor in such early-stage patients overall. The prognosis is affected by a series of factors. The interdependence and complex interactions of such variables may be the reason why none of the previous reports have been widely adopted in clinical practice because of poor reproducibility in independent external cohorts. Through network structure analysis across all clinical and molecular features that were found to have effects on Re in this study, we presented their correlations with each other and with Re events. Using a multivariate Cox model, including clinical features as predictors, we found that genomic features played a crucial role in predicting Re (Figure 5B), and comprehensive genomic and clinical information could stratify prognosis beyond the existing TNM stage for stage I patients, highlighting the utility of genomic sequencing for prognostic prediction. This Cox model stratification was also verified using the EUR-T cohort (Figure 5D). Interestingly, although stage I NSCLC is associated with low ctDNA release and rapid decay, and therefore was more challenging to detect in these patients (Abbosh et al., 2018; Chen et al., 2019b), we found that recurrence risk stratified by our model matched well with pre-surgical ctDNA detection. High-risk patients had a significantly higher ctDNA detection rate, which demonstrated the potential feasibility of combining our molecular model and ctDNA detection for more precise prognostication.

Our multiple-feature model and timing of relapse analysis revealed that molecular features related to ICL and DSB repair during DNA replication and GI mainly contributed to the high risk of E-Re (Figures 3E and 6G). In addition, significant gains in cell

Figure 6. Genome instability and RNA expression correlated with Re risk

(A) Comparison of numbers of patients harboring ICL-DSB recognition, resection, and translesion gene mutations between those with early recurrence (Re $\leq$ 15 months) and late recurrence (Re > 15 months) ($p = 0.038$), as well as Rf patients ($p = 8.89 \times 10^{-5}$). Fisher's exact test was used.

(B and C) Correlation analysis of genome instability scores and DFS time and Re risk. (B) DFS time. (C) Recurrence risk.

(D) Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment of the genes of two clusters based on the online tool Metascape ($p < 0.05$). Color scale bar of $\log_{10}|p$ value$|$ indicates significance enriched in the KEGG database.

(E and F) Correlation of gene expression levels of *FANCF* (an FA core complex gene) (E) and *Rev1* (of TLS pathway) (F) and DFS time.

(G) A diagram showing the correlation of predicted Re risk and genomic and clinical features. The predicted hazard risk of each patient is shown at the top of the graph.

DFS, disease-free survival; FGA, fragment of genome altered with CNVs; HRR, homologous recombination repair; ICL_DSB, interstrand crosslink and double-strand break; LOH, loss of heterozygosity; Recur, tumor recurrence; TLS, translesion repair. The p value less than 0.05 was defined as significant.

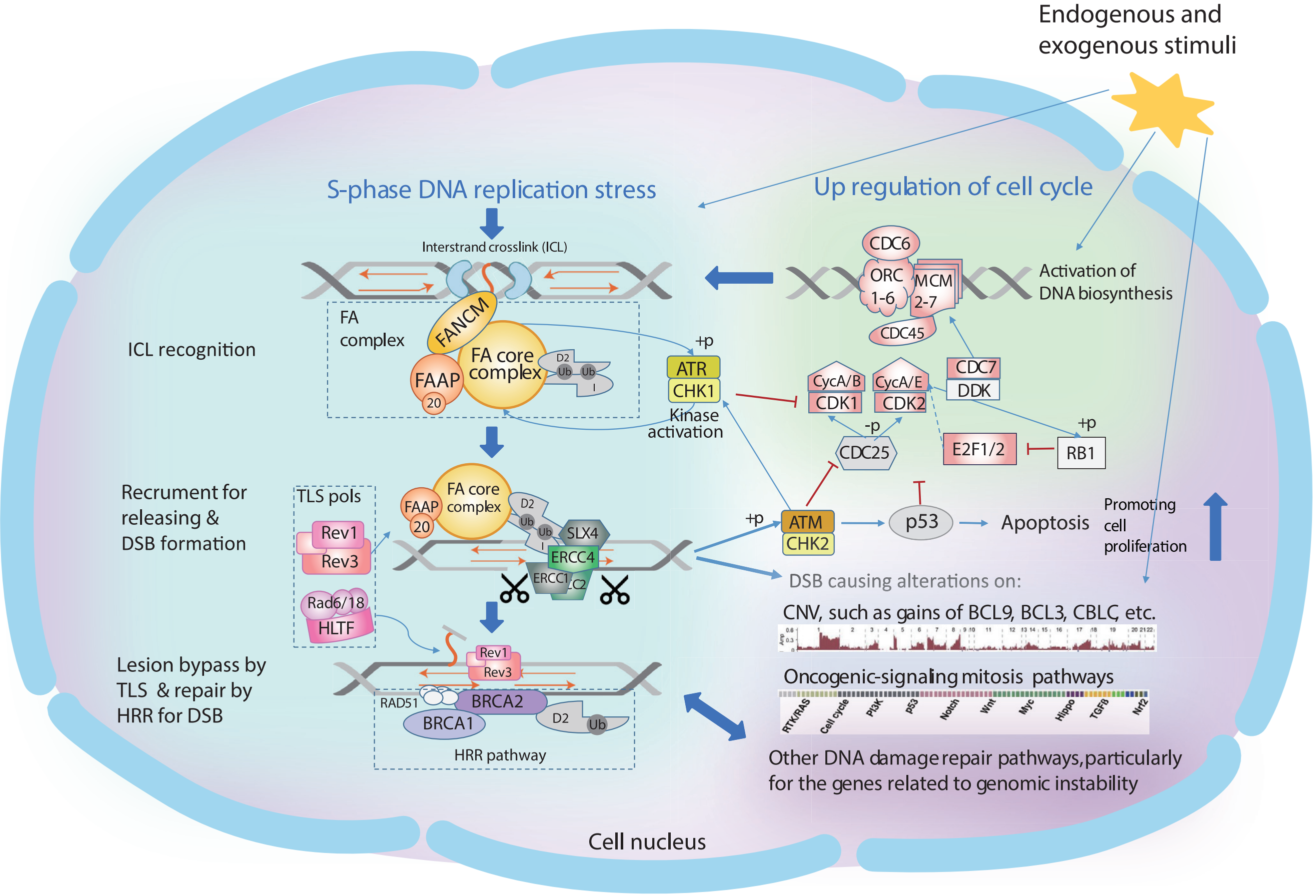

Endogenous and exogenous stimuli
S-phase DNA replication stress
Up regulation of cell cycle
Interstrand crosslink (ICL)
ICL recognition
FA complex
FANCM
FA core complex
FAAP 20
D2
Ub
Ub
I
+p
ATR
CHK1
Kinase activation
CDC6
ORC 1-6
MCM 2-7
CDC45
Activation of DNA biosynthesis
CDC7
DDK
CycA/B
CDK1
CycA/E
CDK2
-p
CDC25
E2F1/2
+p
RB1
Recrument for releasing & DSB formation
TLS pols
Rev1
Rev3
Rad6/18
HLTF
FA core complex
FAAP 20
D2
Ub
Ub
SLX4
ERCC4
ERCC1
+p
ATM
CHK2
p53
Apoptosis
Promoting cell proliferation
DSB causing alterations on:
CNV, such as gains of BCL9, BCL3, CBLC, etc.
Oncogenic-signaling mitosis pathways
RTK/RAS
Cell cycle
PI3K
p53
Notch
Wnt
Myc
Hippo
TGFB
Nrf2
Lesion bypass by TLS & repair by HRR for DSB
Rev1
Rev3
RAD51
BRCA2
BRCA1
D2
Ub
HRR pathway
Other DNA damage repair pathways, particularly for the genes related to genomic instability
Cell nucleus

Figure 7. Molecular processes illustrating E-Re events

Left: E-Re mechanism related to ICL-DSB repair. Key genes' coordination of the FA core complex for ICL recognition and release during DNA replication stress caused by variously endogenous or exogenous stimuli, particularly at the S-phase of the cell cycle, and then recruitment of TLS polymerase (pols) to bypass the ICL barriers to prevent DNA DSB. The DNA DSB was then repaired by the HRR pathway, as well as other DSB repair genes involved in the NER, non-homologous end joining (NHEJ), and MMR pathways that were in low-alteration frequencies in this study. The mutated genes of FA, TLS, and HRR pathways identified in the patients with early recurrence are indicated within the dotted-line boxes. Right: Re mechanism related to cell-cycle dysfunction. Unrepaired DNA DSB results in more SNV/Indels, CNVs, and chromatin remodeling accumulation, which may further cause pathogenic gene alterations and lead to uncontrollable cell cycle and promotion of cell proliferation, further causing DNA replication stress and accelerating genomic instability in cells, thus forming a vicious circle. Significant upregulation of the cell-cycle pathway was detected via RNA sequencing (RNA-seq) in recurrent tumors (shown in the green background).

proliferation-promoting focal CNV-driver genes (Figures 4F and 4G) and upregulation of cell-cycle genes (Figure 6D) were observed in the patients with Re, both E-Re and L-Re. Moreover, these two types of molecular event occurred with ICL-DSB recognition, resection, and translesion gene mutations in most E-Re cases (Figure 6G), which would aggravate GI and increase cancer risk and relapse susceptibility. Figure 7 illustrates the hypothetical molecular processes leading to GI and association with E-Re as detected in Chinese stage I NSCLCs of the CHN-P cohort. These processes were composed of upregulation of cell cycle (upper right part of Figure 7), gene alterations related to DNA ICL-DSB repair (left part), and activation of cell proliferation (lower right part). A recent study observed that DNA repair/replication gene mutations were significantly enriched in circulating tumor cells or relapse/metastatic tumors in comparison with the gene mutation frequencies in primary SCLC tumors (Su et al., 2019), implying there might be certain selection force in metastatic tumor cells with such mutation features compromising relapse, which needs in-depth investigation in the future studies. Tumor cells with features of DNA homologous recombination deficiency (HRD) and the core gene mutations of HRR, FA, and TLS pathways have been reported to be sensitive to platinum chemotherapy and poly-ADP-ribose polymerase inhibitors (Niraj et al., 2019; Soca-Chafre et al., 2019; Zafar and Eoff, 2017), and possibly also sensitive to immunotherapy because of high TMB in such tumor cells. Therefore, the core genes committed to ICL recognition and DSB repair, as well as those that strongly promote cell proliferation and cause DNA replication stress events identified in this study, may be valuable biomarkers for poor prognosis prediction and drug targets for Chinese patients with stage I NSCLC in clinical practice.

To summarize, we comprehensively investigated the genomic and molecular features of stage I lung NSCLCs and compared them with those from diverse international cohorts. Recurrent tumors have distinct somatic alteration events compared with Rf tumors and demonstrate a variety of alterations correlated with time to recurrence. These multiregional sequencing and multidimensional analyses provide important insights into predicting relapse after surgery for stage I NSCLC, which may help guide postoperative treatment strategies in this group of patients.

Limitations of the study

The main limitation is the small sample size of the cohort. The strict inclusion criteria of this study (such as long follow-up time) limited the sample size but reduced interference factors and insured the reliability of this study. In addition, we did not find any independent multiregional Chinese stage I NSCLC cohort with comprehensive genomic features and decent clinical information to validate our findings in this study. Another limitation is the poor RNA quality due to long time storage of the tissue samples, which resulted in only dozens of RNA samples meeting the quality requirements. These qualified samples helped us to reveal RNA expression differences between Re and non-recurrence patients but hindered a profound study on the tumor immune microenvironment.

STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - ○ Subtype classification
  - ○ DNA isolation
  - ○ Whole exome sequencing
  - ○ Mutation calling with WES data
  - ○ Driver gene identification and comparison
  - ○ Mutational signature derivation
  - ○ Phylogenetic tree construction
  - ○ ITH analysis on somatic mutations
  - ○ Comparison of mutational genes between the cohorts
  - ○ Pathway alteration analysis
  - ○ Copy number variation analysis
  - ○ ITH of copy number variation
  - ○ Focal CNV-related gene identification
  - ○ Genome doubling and loss of heterozygosity
  - ○ Dynamic analysis on genomic alterations
  - ○ Multilayer features for integrative analysis
  - ○ Feature importance in DFS analysis
  - ○ Plasma cfDNA sequencing
  - ○ Bioinformatics analysis of ctDNA mutation
  - ○ RNA isolation and sequencing
  - ○ RNA expression and functional analysis
  - ○ Public datasets
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.celrep.2022.111047.

ACKNOWLEDGMENTS

This study was supported by the Research Unit of Intelligence Diagnosis and Treatment in Early Non-small Cell Lung Cancer, Chinese Academy of Medical Sciences (2021RU002), National Natural Science Foundation of China (82072566 and 81602001), and Peking University People's Hospital Research and Development Funds (RS2019-01).

AUTHOR CONTRIBUTIONS

Conceptualization, K.C., S.W., and J.W.; formal analysis, K.C., A.Y., K.L., and S.W.; software, A.Y., K.L., and R.W.; molecular mechanism: S.W.; technique support, J.B., L.W., and Y.S.; resources, K.C., Y.N., H.S., and H.L.; writing, K.C., S.W., A.Y., D.P.C., N.K., T.M., J.Y., and M.J.-H.; supervision, F.Y. and J.W.

DECLARATION OF INTERESTS

M.J.-H. is a Cancer Research UK Fellow and has received funding from CRUK, National Institute for Health Research, Rosetrees Trust, UKI NETs, and NIHR University College London Hospitals Biomedical Research Center. M.J.-H. is a member of the Scientific Advisory Board and Steering Committee for Achilles Therapeutics. A.Y., K.L., R.W., J.B., L.W., H.L., Y.S., and S.W. are employees of Berry Oncology Corporation.

Received: September 8, 2021
Revised: April 26, 2022
Accepted: June 13, 2022
Published: July 12, 2022

REFERENCES

Abbosh, C., Birkbak, N.J., and Swanton, C. (2018). Early stage NSCLC - challenges to implementing ctDNA-based screening and MRD detection. Nat. Rev. Clin. Oncol. *15*, 577–586. https://doi.org/10.1038/s41571-018-0058-3.

Abbosh, C., Birkbak, N.J., Wilson, G.A., Jamal-Hanjani, M., Constantin, T., Salari, R., Le Quesne, J., Moore, D.A., Veeriah, S., Rosenthal, R., et al. (2017). Phylogenetic ctDNA analysis depicts early-stage lung cancer evolution. Nature *545*, 446–451. https://doi.org/10.1038/nature22364.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249. https://doi.org/10.1038/nmeth0410-248.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169. https://doi.org/10.1093/bioinformatics/btu638.

Chaudhuri, A.A., Chabon, J.J., Lovejoy, A.F., Newman, A.M., Stehr, H., Azad, T.D., Khodadoust, M.S., Esfahani, M.S., Liu, C.L., Zhou, L., et al. (2017). Early detection of molecular residual disease in localized lung cancer by circulating tumor DNA profiling. Cancer Discov. *7*, 1394–1403. https://doi.org/10.1158/2159-8290.cd-17-0716.

Chen, H., Carrot-Zhang, J., Zhao, Y., Hu, H., Freeman, S.S., Yu, S., Ha, G., Taylor, A.M., Berger, A.C., Westlake, L., et al. (2019a). Genomic and immune profiling of pre-invasive lung adenocarcinoma. Nat. Commun. *10*, 5472. https://doi.org/10.1038/s41467-019-13460-3.

Chen, J., Yang, H., Teo, A.S.M., Amer, L.B., Sherbaf, F.G., Tan, C.Q., Alvarez, J.J.S., Lu, B., Lim, J.Q., Takano, A., et al. (2020). Genomic landscape of lung adenocarcinoma in East Asians. Nat. Genet. *52*, 177–186. https://doi.org/10.1038/s41588-019-0569-6.

Chen, K., Zhao, H., Shi, Y., Yang, F., Wang, L.T., Kang, G., Nie, Y., and Wang, J. (2019b). Perioperative dynamic changes in circulating tumor DNA in patients with lung cancer (DYNAMIC). Clin. Cancer Res. *25*, 7058–7067. https://doi.org/10.1158/1078-0432.ccr-19-1213.

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

Chen, S., Zhou, Y., Chen, Y., Huang, T., Liao, W., Xu, Y., Li, Z., and Gu, J. (2019c). Gencore: an efficient tool to generate consensus reads for error suppressing and duplicate removing of NGS data. BMC Bioinf. *20*, 606. https://doi.org/10.1186/s12859-019-3280-9.

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. https://doi.org/10.1038/nbt.2514.

Devarakonda, S., Rotolo, F., Tsao, M.S., Lanc, I., Brambilla, E., Masood, A., Olaussen, K.A., Fulton, R., Sakashita, S., McLeer-Florin, A., et al. (2018). Tumor mutation burden as a biomarker in resected non-small-cell lung cancer. J. Clin. Oncol. *36*, 2995–3006. https://doi.org/10.1200/jco.2018.78.1963.

Gehring, J.S., Fischer, B., Lawrence, M., and Huber, W. (2015). SomaticSignatures: inferring mutational signatures from single-nucleotide variants. Bioinformatics *31*, 3673–3675. https://doi.org/10.1093/bioinformatics/btv408.

Goldstraw, P., Chansky, K., Crowley, J., Rami-Porta, R., Asamura, H., Eberhardt, W.E.E., Nicholson, A.G., Groome, P., Mitchell, A., and Bolejack, V. (2016). The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. J. Thorac. Oncol. *11*, 39–51. https://doi.org/10.1016/j.jtho.2015.09.009.

Group, N.M.-a.C., Arriagada, R., Auperin, A., Burdett, S., Higgins, J.P., Johnson, D.H., Le Chevalier, T., Le Pechoux, C., Parmar, M.K., et al. (2010). Adjuvant chemotherapy, with or without postoperative radiotherapy, in operable non-small-cell lung cancer: two meta-analyses of individual patient data. Lancet *375*, 1267–1277. https://doi.org/10.1016/S0140-6736(10)60059-1.

Hu, X., Fujimoto, J., Ying, L., Fukuoka, J., Ashizawa, K., Sun, W., Reuben, A., Chow, C.W., McGranahan, N., Chen, R., et al. (2019). Multi-region exome sequencing reveals genomic evolution from preneoplasia to lung adenocarcinoma. Nat. Commun. *10*, 2978. https://doi.org/10.1038/s41467-019-10877-8.

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. (2017). Tracking the evolution of non-small-cell lung cancer. N. Engl. J. Med. *376*, 2109–2121. https://doi.org/10.1056/NEJMoa1616288.

Jia, Q., Chiu, L., Wu, S., Bai, J., Peng, L., Zheng, L., Zang, R., Li, X., Yuan, B., Gao, Y., et al. (2020). Tracking neoantigens by personalized circulating tumor DNA sequencing during checkpoint blockade immunotherapy in non-small cell lung cancer. Adv. Sci. *7*, 1903410. https://doi.org/10.1002/advs.201903410.

Kass, E.M., Moynahan, M.E., and Jasin, M. (2016). When genome maintenance goes badly awry. Mol. Cell *62*, 777–787. https://doi.org/10.1016/j.molcel.2016.05.021.

Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat. Biotechnol. *37*, 907–915. https://doi.org/10.1038/s41587-019-0201-4.

Knijnenburg, T.A., Wang, L., Zimmermann, M.T., Chambwe, N., Gao, G.F., Cherniack, A.D., Fan, H., Shen, H., Way, G.P., Greene, C.S., et al. (2018). Genomic and molecular landscape of DNA damage repair deficiency across the cancer genome Atlas. Cell Rep. *23*, 239–254.e6. https://doi.org/10.1016/j.celrep.2018.03.076.

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. https://doi.org/10.1101/gr.129684.111.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218. https://doi.org/10.1038/nature12213.

Letouze, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.F., Tubacher, E., Bayard, Q., Bacq, D., Meyer, V., Semhoun, J., et al. (2017). Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nat. Commun. *8*, 1315. https://doi.org/10.1038/s41467-017-01358-x.

Li, B., Cui, Y., Diehn, M., and Li, R. (2017). Development and validation of an individualized immune prognostic signature in early-stage nonsquamous non-small cell lung cancer. JAMA Oncol. *3*, 1529. https://doi.org/10.1001/jamaoncol.2017.1609.

Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and 1000 Genome Project Data Processing Subgroup. (2009). The Sequence alignment/map (SAM) format and SAMtools, Bioinformatics *25*, 2078-2079. https://doi.org/10.1093/bioinformatics/btp352.

López, S., Lim, E.L., Horswell, S., Haase, K., Huebner, A., Dietzen, M., Mourikis, T.P., Watkins, T.B.K., Rowan, A., Dewhurst, S.M., et al. (2020). Interplay between whole-genome doubling and the accumulation of deleterious alterations in cancer evolution. Nat. Genet. *52*, 283–293. https://doi.org/10.1038/s41588-020-0584-7.

Lv, W., Wei, X., Guo, R., Liu, Q., Zheng, Y., Chang, J., Bai, T., Li, H., Zhang, J., Song, Z., et al. (2015). Noninvasive prenatal testing for Wilson disease by use of circulating single-molecule amplification and resequencing technology (cSMART). Clin. Chem. *61*, 172–181. https://doi.org/10.1373/clinchem.2014.229328.

Martincorena, I., Raine, K.M., Gerstung, M., Dawson, K.J., Haase, K., Van Loo, P., Davies, H., Stratton, M.R., and Campbell, P.J. (2017). Universal patterns of selection in cancer and somatic tissues. Cell *171*, 1029–1041.e21. https://doi.org/10.1016/j.cell.2017.09.042.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res. *31*, 3812–3814. https://doi.org/10.1093/nar/gkg509.

Niraj, J., Färkkilä, A., and D'Andrea, A.D. (2019). The fanconi anemia pathway in cancer. Annu. Rev. Cancer Biol. *3*, 457–478. https://doi.org/10.1146/annurev-cancerbio-030617-050422.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.C., and Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinf. *12*, 77. https://doi.org/10.1186/1471-2105-12-77.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616.

Rosenthal, R., Cadieux, E.L., Salgado, R., Bakir, M.A., Moore, D.A., Hiley, C.T., Lund, T., Tanić, M., Reading, J.L., Joshi, K., et al. (2019). Neoantigen-directed immune escape in lung cancer evolution. Nature *567*, 479–485. https://doi.org/10.1038/s41586-019-1032-7.

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B.S., and Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. *17*, 31. https://doi.org/10.1186/s13059-016-0893-4.

Ross, E.M., Haase, K., Van Loo, P., and Markowetz, F. (2020). Allele-specific multi-sample copy number segmentation in ASCAT. Bioinformatics *37*, 1909–1911. https://doi.org/10.1093/bioinformatics/btaa538.

Sanchez-Vega, F., Mina, M., Armenia, J., Chatila, W.K., Luna, A., La, K.C., Dimitriadoy, S., Liu, D.L., Kantheti, H.S., Saghafinia, S., et al. (2018). Oncogenic signaling pathways in the cancer genome Atlas. Cell *173*, 321–337. https://doi.org/10.1016/j.cell.2018.03.035.

Scarbrough, P.M., Weber, R.P., Iversen, E.S., Brhane, Y., Amos, C.I., Kraft, P., Hung, R.J., Sellers, T.A., Witte, J.S., Pharoah, P., et al. (2016). A cross-cancer genetic association analysis of the DNA repair and DNA damage signaling pathways for lung, ovary, prostate, breast, and colorectal cancer. Cancer Epidemiol. Biomarkers Prev. *25*, 193–200. https://doi.org/10.1158/1055-9965.epi-15-0649.

Siegel, R.L., Miller, K.D., and Jemal, A. (2020). Cancer statistics, 2020. CA Cancer J. Clin. *70*, 7–30. https://doi.org/10.3322/caac.21590.

Soca-Chafre, G., Montiel-Dávalos, A., Rosa-Velázquez, I.A.D.L., Caro-Sánchez, C.H.S., Peña-Nieves, A., and Arrieta, O. (2019). Multiple molecular targets associated with genomic instability in lung cancer. Int. J. Genom. *2019*, 1–8. https://doi.org/10.1155/2019/9584504.

Su, Z., Wang, Z., Ni, X., Duan, J., Gao, Y., Zhuo, M., Li, R., Zhao, J., Ma, Q., Bai, H., et al. (2019). Inferring the evolution and progression of small-cell lung cancer by single-cell sequencing of circulating tumor cells. Clin. Cancer Res. *25*, 5049–5060. https://doi.org/10.1158/1078-0432.ccr-18-3571.

Tarasov, A., Vilella, A.J., Cuppen, E., Nijman, I.J., and Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. Bioinformatics *31*, 2032–2034. https://doi.org/10.1093/bioinformatics/btv098.

Tsao, M.S., Marguet, S., Le Teuff, G., Lantuejoul, S., Shepherd, F.A., Seymour, L., Kratzke, R., Graziano, S.L., Popper, H.H., Rosell, R., et al. (2015). Subtype classification of lung adenocarcinoma predicts benefit from adjuvant chemotherapy in patients undergoing complete resection. J. Clin. Oncol. *33*, 3439–3446. https://doi.org/10.1200/jco.2014.58.8335.

Turnham, D.J., Yang, W.W., Davies, J., Varnava, A., Ridley, A.J., Conlan, R.S., and Clarkson, R.W.E. (2020). Bcl-3 promotes multi-modal tumour cell migration via NF-κB1 mediated regulation of Cdc42. Carcinogenesis *41*, 1432–1443. https://doi.org/10.1093/carcin/bgaa005.

Vansteenkiste, J., Crinò, L., Dooms, C., Douillard, J.Y., Faivre-Finn, C., Lim, E., Rocco, G., Senan, S., Van Schil, P., Veronesi, G., et al. (2014). 2nd ESMO Consensus Conference on Lung Cancer: early-stage non-small-cell lung cancer consensus on diagnosis, treatment and follow-up. Ann. Oncol. *25*, 1462–1474. https://doi.org/10.1093/annonc/mdu089.

Vargas, A.J., and Harris, C.C. (2016). Biomarker development in the precision medicine era: lung cancer as a case study. Nat. Rev. Cancer *16*, 525–537. https://doi.org/10.1038/nrc.2016.56.

Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. *38*, e164. https://doi.org/10.1093/nar/gkq603.

Wang, Z., Zhao, J., Wang, G., Zhang, F., Zhang, Z., Zhang, F., Zhang, Y., Dong, H., Zhao, X., Duan, J., et al. (2018). Comutations in DNA damage response pathways serve as potential biomarkers for immune checkpoint blockade. Cancer Res. *78*, 6486–6496. https://doi.org/10.1158/0008-5472.can-18-1814.

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. https://doi.org/10.1038/ncomms10131.

Zafar, M.K., and Eoff, R.L. (2017). Translesion DNA synthesis in cancer: molecular mechanisms and therapeutic opportunities. Chem. Res. Toxicol. *30*, 1942–1955. https://doi.org/10.1021/acs.chemrestox.7b00157.

Zhang, X.C., Wang, J., Shao, G.G., Wang, Q., Qu, X., Wang, B., Moy, C., Fan, Y., Albertyn, Z., Huang, X., et al. (2019). Comprehensive genomic and immunological characterization of Chinese non-small cell lung cancer patients. Nat. Commun. *10*, 1772. https://doi.org/10.1038/s41467-019-09762-1.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Biological samples		
Lung tumor tissues	This study	This study
Adjacent non-cancerous lung tissues	This study	This study
Peripheral blood samples just before surgery	This study	This study
Chemicals, peptides, and recombinant proteins		
ZR Genomic DNA Tissue MiniPrep kit	Zymo Research, USA	Cat# D3051
DNA Blood Midi/Mini kit	Qiagen, USA	Cat# 69504
MagMAX Cell-Free DNA Isolation Kit	Thermo Fisher Scientific, USA	Cat# A29319
5X WGS Fragmentation Mix	Qiagen, USA	Cat# Y9410L
An in-house process of DNA fragment end repair, A tailing, T-adaptors ligase, PCR amplification	Berry Oncology Corporation, China	This study
96 rxn xGen Exome Research Panel v1.0	Integrated DNA Technologies, USA	Cat#1056115
An in-house 457-gene panel	Berry Oncology Corporation, China	This study
RN28-EASYspin Plus RNA Mini Kit	Aidlab Biotechnologies Co., Ltd, China	Cat# RN2802
Ribo-Zero Magnetic (Human) kit	Epicentre Biotechnologies, USA	Cat# MRZH116
Deposited data		
Stage I NSCLC WES data	This study	HRA001278 in the China National Center for Bioinformation (https://ngdc.cncb.ac.cn/gsa/).
Stage I NSCLC RNA-seq data	This study	HRA001278 in the China National Center for Bioinformation (https://ngdc.cncb.ac.cn/gsa/).
Stage I NSCLC ctDNA data	This study	HRA001278 in the China National Center for Bioinformation (https://ngdc.cncb.ac.cn/gsa/).
East Asian cohort data	Chen et al., 2020	Nature Genetics. *52*, 177-186.
European cohort data	Jamal-Hanjani et al., 2017	The New England Journal of Medicine. *376*, 2109-2121.
TCGA cohort data	N/A	https://portal.gdc.cancer.gov
Software and algorithms		
FASTP (v0.14.1)	Chen et al., 2018	RRID: SCR_016962
Burrows-Wheeler Aligner (BWA, v0.7.15)	Li et al., 2009	RRID: SCR_010910
Sambamba (v0.6.8)	Tarasov et al., 2015	https://github.com/biod/sambamba
GATK (v4.0.11.0)	McKenna et al., 2010	RRID: SCR_001876
Mutect (v1.1.4)	Cibulskis et al., 2013	RRID: SCR_000559
ENCODE Data Analysis Consortium blacklisted regions	Letouze et al., 2017	http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz
ANNOVAR	Wang et al., 2010	RRID: SCR_012821
MutSigCV (v.1.4)	Lawrence et al., 2013	https://software.broadinstitute.org/cancer/cga/mutsig_download
dNdScv (v.0.1.0)	Martincorena et al., 2017	RRID: SCR_017093
SomaticSignatures R package (v 2.20.0)	Gehring et al., 2015	http://www.bioconductor.org/packages/release/bioc/html/SomaticSignatures.html
deconstructSigs (v1.9.0)	Rosenthal et al., 2016	https://github.com/raerose01/deconstructSigs
PHYLIP (v3.697)	N/A	RRID: SCR_006244
VarScan2 (v2.3.9)	Koboldt et al., 2012	RRID: SCR_006849
ASCAT (v2.5.2)	Ross et al., 2020	RRID: SCR_016868
FASTQC (v0.11.9)	N/A	RRID: SCR_014583

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
hisat2 (v2.1.0)	Kim et al., 2019	RRID: SCR_015530
HTseq (v0.11.0)	Anders et al., 2015	RRID: SCR_005514
edgeR (v.3.16.5)	Robinson et al., 2010	RRID: SCR_012802
Metascape online tools	Zhou et al., 2019	RRID: SCR_016620
Gencore	Chen et al., 2018	https://github.com/OpenGene/gencore
SAMtools	Li et al., 2009	RRID: SCR_002105
R/Bioconductor software packages	N/A	RRID: SCR_006442
Other		
Qubit® 4.0 Fluorometer	Life Technologies, USA	Cat# Q33238
Fragment Analyzer	Agilent Technologies, USA	Cat# G2938C
Illumina NovaSeq 6000	Illumina, USA	Cat# 20012850
HGVS variant description		http://varnomen.hgvs.org/
the 1000 Genomes Project (1000G)		http://browser.1000genomes.org
ExAC		http://exac.broadinstitute.org
dbSNP		https://www.ncbi.nlm.nih.gov/snp/
disease or phenotype databases OMIM		http://www.omim.org
COSMIC		https://cancer.sanger.ac.uk/cosmic/
ClinVar		http://www.ncbi.nlm.nih.gov/clinvar
PolyPhen-2		http://genetics.bwh.harvard.edu/pph2/
SIFT		http://www.blocks.fhcrc.org/sift/SIFT.html
KEGG		https://www.kegg.jp/kegg/pathway.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Jun Wang, wangjun@pkuph.edu.cn.

Materials availability

This study did not generate new unique reagents and biological materials.

Data and code availability

- Data: Raw data of WES, RNA-seq and targeted sequencing of ctDNA derived from human samples of the CHN-P cohort have been deposited at the China National Center for Bioinformation (https://ngdc.cncb.ac.cn/gsa/), and the accession number (HRA001278) is listed in the key resources table. Local law prohibits depositing raw WES and RNA-seq datasets derived from human samples outside of the country of origin. Prior to publication, the authors officially requested that the raw sequencing datasets reported in this paper be made publicly accessible. To request access, contact the Office of Human Genetic Resource Administration of The Ministry of Science and Technology for The Regulation of the People's Republic of China on the Administration of Human Genetic Resources.
- This paper does not report original code. The software used in this study is described in the above section and the key resources table in details.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

This retrospective study was carried out on a Chinese PKUPH cohort (named as CHN-P) of 81 Chinese non-small cell lung cancer (NSCLC) patients, including 73 lung adenocarcinoma (LUAD), 6 lung squamous cell carcinoma (LUSC), and 2 lung adenosquamous carcinoma (LUAS), who were diagnosed with pathological Stage I NSCLC in our department of the Peking University People's Hospital from August 2011 to April 2017 (Table S1). The patients' the age, sex, gender identities were also provided in Table S1. The patient inclusion criteria were (1) History of chest computed tomography (CT) scans, abdominal and adrenal gland ultrasonography or CT, brain magnetic resonance imaging and bone scans, or PET/CT before surgery; (2) Aged >18 years with no malignant tumor history within the past 5 years; (3) No neoadjuvant therapy was administered before surgery; and (4) Patients received R0 resection.

Lung tumor tissue and adjacent normal tissue specimens were collected by surgical removal for histopathologically pathological diagnosis and further study. To assess intra-tumor heterogeneity, the tumor tissue samples were separated into two to three regions (depending on the size of the tumor) immediately after pathological-diagnosis sampling and stored at -80°C. Paired peripheral blood samples were collected immediately before the surgery, separated into white blood cells and plasma, and then stored at -80°C in the laboratory for further experiments.

Clinical follow-up examinations were carried out for each patient after surgery, with a median time of 62.0 months (95% CI, 57.0–67.0). Among them, 33 patients experienced tumor recurrence (26 LUAD, 5 LUSC, and 2 LUAS), while 48 (47 LUAD and 1 LUSC) did not. The 7th American Joint Committee on Cancer edition of TNM staging was used in our study. We collected clinical variables for patients, including age, sex, smoking history, tumor histology, tumor location, and tumor size measured by CT and are summarized in Table S1. Chest CT scan and abdominal ultrasound/CT were performed on follow-up visits every 6 months after surgery for 5 years. Magnetic resonance imaging and bone scans were performed every 1 year for 5 years or any time with symptoms. The overall survival (OS) time was estimated from the date of surgical resection until death of any cause or the date of the last follow-up. Disease-free survival (DFS) time was defined as the time from the day of surgery until the first event (relapse or metastasis) or last follow-up. Comparison of clinical features between subgroups was based on Chi-square test, except for age and tumor size, which were based on the Wilcoxon rank-sum test.

The above-mentioned criteria reduced interference factors comparing to other prognostic studies and insured the reliability of this study. (1) All the enrolled patients underwent R0 resection and had no neoadjuvant therapy history. Many previous studies focusing on the prognosis of cancer patients after surgery did not completely exclude non-radical resection patients (R1, R2) so that the tumor relapse is related to surgery itself but not biological reasons. (2) All enrolled patients were not locally resected, therefore excluding the possibility of tumor recurrence that may be due to a nonstandard surgical technique (not enough groups of lymph node resected, the surgical margin is not distant enough from the tumor, etc.) or physical local tumor invasion (spread through air space, STAS). Distant metastasis for such stage I patients is more likely to be attributed to tumor biological features. (3) We clearly defined tumor recurrence but not the death of patients as tumor events for prognosis analysis, contrary to several studies that combined DFS and OS, which may complicate results because when patients experience tumor relapse, they may receive many terms of drug treatment that could have an important influence on the OS. (4) The follow-up time of this study was long enough (median: 5 years) to ensure the reliability of the tumor-related events, and the DFS was accurately recorded.

Whole-exome sequencing (WES) was performed on a total of 240 tumor tissue specimens and 81 matched adjacent normal tissues from 81 NSCLC patients. Twenty-two patients (13 recurrent and 9 recurrence-free patients) with 28 tumor samples passed RNA quantity and quality evaluation for RNA sequencing. Fifty-five plasma samples were successfully isolated from cell-free DNA (cfDNA) to perform deep-targeted sequencing with an in-house 457-gene panel of Berry Oncology Corporation (China).

Written informed consent for sample acquisition for research purposes in this study was obtained from all patients, and this study was approved by the Ethical Committee of Medical Research, Peking University People's Hospital in accordance with the Declaration of Helsinki Principles. This study did not generate new unique reagents.

METHOD DETAILS

Subtype classification

First, according to tumor recurrence or not, the patients were divided into two groups: tumor recurrence (Re) (n = 33) and recurrence-free (Rf) (n = 48) patients. For the Re patients, the median recurrence time was 15 months (range, 2–62 months). Patients with early tumor recurrence were defined according to their DFS time. Since the median DFS is 15 months in this cohort, we define recurrence less than 15 months after surgery as early relapse. Second, for the dynamic analysis, the Re patients were further defined as early (E-Re) and late (L-Re). E-Re referred to the tumor recurring earlier than the specific time point during the follow-up, while L-Re referred to the tumor recurring later than the specific time point during the follow-up survey.

DNA isolation

For genomic DNA (gDNA) isolation, ZR Genomic DNA Tissue Miniprep Kits (Zymo Research) for the tissue specimens and DNA Blood Midi/Mini kit (Qiagen) for the white blood cell samples were used according to the manufacturer's instructions. For the isolation of plasma cell-free DNA (cfDNA), MagMAX Cell-Free DNA Isolation Kit (Thermo Fisher Scientific) was used according to the manufacturer's protocol. The quality and quantification of purified DNA were assayed using gel electrophoresis and Qubit® 4.0 Fluorometer (Life Technologies), respectively. The DNA fragment size composition was assayed using a Fragment Analyzer (Agilent).

Whole exome sequencing

Purified 30 to 100 ng gDNA was first fragmented into DNA pieces of approximately 300 bp using an enzymatic method (5X WGS Fragmentation Mix, Qiagen). The DNA fragments then underwent an in-house process of end repair, A tailing, T-adaptors ligation on both ends, and PCR amplification to result in a pre-library. The final sequencing libraries were prepared using the 96 rxn xGen Exome Research Panel v1.0 (Integrated DNA Technologies) according to the manufacturer's protocol. 2 X 150-bp paired-end sequencing was performed with Illumina NovaSeq 6000 (Illumina). An average sequencing depth of 561 X (range: 194–1203 X) for tumor tissues and 513 X (range: 113-1062 X) for adjacent normal tissues (Table S2A) were obtained. The mean genomic coverage of $\geq$20 X was 98.5%.

Mutation calling with WES data

The raw sequencing reads were first subjected to quality control by trimming adaptor sequences and removing the reads with poly-N and low quality (less than Q20) preprocessed by FASTP (v0.14.1) (Chen et al., 2018). The clean reads in FASTQ format were aligned to the human UCSC reference genome (hg19/GRCh37) using Burrows-Wheeler Aligner (BWA, v0.7.15) with default parameters (Li and Durbin, 2009). Sambamba (v0.6.8) was used to process PCR duplicates for mapped BAM files (Tarasov et al., 2015). GATK (Genome Analysis Toolkit v4.0.11.0) (McKenna et al., 2010)for local realignment and base quality recalibration was employed to compute sequencing coverage and depth. Single nucleotide variations (SNVs) and small insertions and deletions (Indels: <50 bp) were identified using GATK MuTect2 (v1.1.4) (Cibulskis et al., 2013). Subsequently, we removed mutations, which were referred to the ENCODE Data Analysis Consortium blacklisted regions (Letouze et al., 2017). We filtered out the SNVs with <20 X depth or 4 X depth of the alternate alleles in tumor or SNVs with <10 X depth in normal or variant reads >1% of normal reads. For multiple-region samples, variants detected in more than one sample, but not all samples, were recalled because the absent variants in a part of samples might be due to low variant allele frequency (VAF), thus reducing false negative callings.

Variants were annotated using ANNOVAR software (Wang et al., 2010) based on multiple databases including HGVS variant description and population frequency databases (1000G, http://browser.1000genomes.org), ExAC (http://exac.broadinstitute.org), dbSNP (https://www.ncbi.nlm.nih.gov/snp/), disease or phenotype databases (OMIM (http://www.omim.org), COSMIC (https://cancer.sanger.ac.uk/cosmic/), ClinVar (http://www.ncbi.nlm.nih.gov/clinvar), and variant functional in silico predictive tools (PolyPhen-2, SIFT) (Adzhubei et al., 2010; Ng and Henikoff, 2003)to interpret the sequence variant at the nucleotide and amino acid levels. After annotation, we excluded the SNVs that were annotated as genomicSuperDups and VAF <0.2 or PopFreqMax >0.05 and kept the nonsynonymous SNVs with VAF >1% of cancer hotspots collected from the patient database or with VAF >3% of others for further analysis. All somatic mutations identified in the CHN-P cohort are summarized in Table S3A. For each mutation, the proportion of mutated reads (VAF) and the proportion of tumor cells harboring the mutation (cancer cell fraction, CCF) were calculated according to the methods described by Letouzé et al. (Letouze et al., 2017). For each sample, tumor mutation burden (TMB) was defined as the total number of nonsynonymous SNVs per megabase of coding area of a tumor genome based on WES (Table S3B), and SNVs/Indels' diversity was calculated and expressed by Shannon's Diversity Index based on clonal and subclonal mutation proportions. Comparison of SNV/Indel numbers, diversities, and TMB values between Re and Rf groups, as well as between smokers and nonsmokers, or between different cohorts was based on the Wilcoxon rank-sum test. Associations of specific mutated genes with clinical features, including gender, smoking, pathological type, tumor size, age, and stage, were based on Fisher's exact test.

Driver gene identification and comparison

To identify genes with significant frequency differences between groups, two-sided Fisher's exact tests were performed for all genes with a p-value cutoff of 0.05 to filter no significant values (Figure 1D and Table S3E). To identify driver mutations in the CHN-P cohort, MutSigCV (v.1.4) (Lawrence et al., 2013)and dNdScv (v.0.1.0) (Martincorena et al., 2017) were used with default parameters to infer significantly mutated driver genes (q < 0.1 in both callers) with the following results:

(1) MutSigCV (v1.4) was performed, and *EGFR* and *TP53* were the significantly mutated genes identified in both Re and Rf groups in the CHN-P cohort (q < 0.1).
(2) The dNdScv (v.0.1.0) R package was used to detect genes under positive selection in the CHN-P cohort. *EGFR* and *TP53* for both groups, as well as *KRAS* and *KEAP1* for the Rf group were identified (q < 0.1).

Mutational signature derivation

The gene mutational signatures of all specimens were *de novo* derived from WES data according to a non-negative matrix factorization (NMF) method using the SomaticSignatures R package (v2.20.0) (Gehring et al., 2015). Three stable and reproducible mutational signatures were deciphered (Figure S2A) and termed as signatures S1, S2, and S4. Cosine similarity was analyzed to compare these signatures to the catalog of COSMIC consensus signatures (Figure S2B). For each patient, signatures of clonal and subclonal somatic mutations were identified based on the signatures. Somatic mutations of the EUR-T cohort were also processed through the above analysis to *de novo* derive their mutational signatures.

To further determine the distribution of COSMIC signatures in each patient, deconstructSigs (v1.9.0) was used as previously described (Rosenthal et al., 2016), and the frequencies of these signatures in the CHN-P cohort are summarized. Patient numbers harboring S1, S2, and S4 were compared using Fisher's exact test (Figure S2C). Weights of S1, S2, and S4 in clonal and subclonal mutations between the two groups were compared using the Wilcoxon rank-sum test (Figure S2D). The associations between signatures and categorical variables of clinical features, including gender, smoking, and pathological types, were performed using the Wilcoxon rank-sum test, except for the stage variables that used the Cochran-Armitage Trend Test. Simple linear regression analysis was implemented using the R command lm to identify potential associations between signatures and continuous variables, including age, tumor size, TMB, and DFS time (Figure S2E).

Phylogenetic tree construction

All nonsilent mutations after filtering were considered for determining phylogenetic trees. We inferred phylogenetic trees of tumor blocks based on the mutation patterns in each of the patients using PHYLIP (v3.697; http://evolution.genetics.washington.edu/phylip.html) to perform the compatibility method Clique, generating unrooted trees. Branch lengths were inferred from the number of non-silent mutations acquired, and final trees were drawn manually and further optimized using Adobe Illustrator. The clonal, shared, and private branches of each tree represent mutations in all the tumor regions, in some but not all the tumor regions, and in only one tumor region, respectively.

ITH analysis on somatic mutations

To investigate possible mutagenic alteration processes during carcinogenesis, the mutation spectra of clonal mutations and branch mutations (i.e., intratumor heterogeneity, ITH) were compared based on their numbers and proportions between the Re and Rf groups using Student's *t*-test.

Comparison of mutational genes between the cohorts

To compare the mutational landscape of stage I NSCLC among different cohorts, we chose four cohorts, including the CHN-P cohort and East Asian LUADs (EAS) cohort, representing Chinese and East Asian race cohorts, as well as the European of TRACERx Project (EUR-T) cohort and TCGA (filtered our Asian patient samples) cohort, both representing Caucasian race cohorts (Tables S1C–S1E). For the CHN-P cohort, we divided patients into Re and Rf subgroups to screen the genes with higher mutation frequency in the Re group than that in the Rf group. After calculating the gene mutation rate of each cohort, genes enriched in at least one of the four cohorts or only in the Re group were selected for comparison. In the graph, the enriched genes are grouped by signaling pathways, and interactions between genes are indicated according to the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway maps. Four hallmarks related to genome maintenance mechanisms and carcinogenesis are listed to indicate the potential influence of mutations among signaling pathways, including regulation of cytoskeleton organization, sustaining cell proliferation, anti-apoptosis processes, and genome repair function (Figure 2).

Pathway alteration analysis

We evaluated the mapping of somatic mutations to 10 canonical oncogenic signaling (COS) pathways according to the templates in the signaling pathways manuscript from TCGA PanCancer Atlas project (Sanchez-Vega et al., 2018) and eight DNA damage repair (DDR) pathways according to previous reports (Wang et al., 2018). Ten COS pathways included (1) cell cycle, (2) Hippo, (3) Myc, (4) Notch, (5) oxidative stress response/Nrf2, (6) PI3K, (7) receptor-tyrosine kinase (RTK)/RAS/MAPK, (8) TGFβ, (9) p53, and (10) β-catenin/Wnt and involved a total of 335 genes (Table S4C). Eight DDR pathways included (1) MMR (mismatch repair), (2) BER (base excision repair), (3) NER (nucleotide excision repair), (4) HRR (homologous recombination repair), (5) NHEJ (non-homologous end-joining), (6) CPF (checkpoint factors), (7) FA (Fanconi anemia), and (8) TLS (translesion synthesis) and involved 233 genes (Table S4D).

A tumor was considered "altered" in the specific pathway when ≥ 1 gene is altered in the corresponding pathway. For each patient, the status of specific pathways was determined to be either altered or wild type. The number of COS pathway alterations (NOA) or the number of DDR pathway alterations (NDA) was calculated as the total number of altered pathways out of the 10 identified pathways or the eight DDR pathways for each patient, respectively. Comparison of NOA or NDA between the Re and Rf groups was performed using the Wilcoxon rank-sum test, and the association of NOA or NDA with TMB was calculated using the Cochran-Armitage test, both with the significance threshold $p \leq 0.05$. Comparison of mutation frequency of each pathway among the CHN-P cohort (All), East Asian cohort (EAS), European cohort (EUR-T), and TCGA cohort, as well as between the Re and Rf groups in the CHN-P cohort were performed using Fisher's exact test.

Copy number variation analysis

Copy number variation (CNV) calculation of multi-region tumors was performed according to the method of the TRACERx Project (Jamal-Hanjani et al., 2017). In brief, processed sample exome copy number data from paired tumor-normal was generated using VarScan2 (v2.3.9) copy number with default parameters except min-coverage = 8, min-segment-size = 50 (Koboldt et al., 2012). The VarScan2 copy number produced per-region LogR values, which were then adjusted by sequencing depth of paired tumor and normal samples. The B-allele frequency (BAF) of each SNP was calculated as the proportion of reads at a specific position that contained the reference base versus the variant for the SNP loci from the 1000 Genomes Project (http://browser.1000genomes.org). The logR and BAF values, which were GC corrected using function ascat.GCcorrect for each tumor region were processed with ASCAT (v2.5.2) (Ross et al., 2020) using default parameters except "gamma" set to 1, to provide segmented allele-specific copy number data plus cellularity and ploidy estimates for all samples. Per region copy number data called ASCAT (v2.5.2) is available in the raw segments sheet of Table S5, and floating-point copy number values were used for all copy number analyses. The maximal ploidy in multi-region tumors was assessed using ASCAT (v2.5.2).

ITH of copy number variation

To determine genome-wide copy number amplification/gain and deletion/loss, copy number data for each sample was divided by the sample mean ploidy and log2 transformed. Gain and loss were defined as log2(2.5/2) and log2(1.5/2), respectively.

For the multi-region sampling method to determine global CNV intratumor heterogeneity (ITH), all parts of the genome were considered independently and split into minimum consecutive segments of overlap within each tumor across all regions. Within each tumor, total CNV was defined as the genomic regions subjected to CNV in any region. The total fraction of genome altered (FGA) was calculated as the percentage of a tumor genome showing a copy number different from the whole genome, while CNV length was defined as the base number (Mb) of the genomic regions subjected to CNV in any region.

Any segment of CNV that overlapped across all regions was defined as clonal and all other segments of CNVs as subclonal. Subclonal CNVs detected only in one region were defined as private CNVs, while those detected in more than one region, but not in all the regions, were defined as shared CNVs. The proportion of clonal CNV or subclonal CNV was then defined as the percentage of the genome subjected to clonal or subclonal CNV divided by the percentage of total CNV, respectively. Numbers or proportions of total, clonal, and subclonal CNVs, as well as FGA, were compared between the Re and Rf groups using the Wilcoxon rank-sum test.

Focal CNV-related gene identification

Gene-level amplification was called the mean gene copy number > ploidy + 1 copy. Gene-level deletion was called the mean gene copy number <0.5 copy. Clonal CNV-related genes were defined as those occurring in all the tumor regions and others were defined as branch/shared and private CNV-related genes. To reveal the CNV differences between the Re and Rf groups, the frequency of focal CNV-related genes was counted and compared between the two groups using Fisher's exact test. The impact of significantly differential (Fisher's p-value<0.05) CNV-driver genes on tumor recurrence events were assayed using the Kaplan–Meier survival analysis method based on the LogRank test, and further analysis was focused on the genes with LogRank p-value <0.05. The driver CNV genes were identified according to the COSMIC cancer_gene_census database (website is stated in the key resources table), and those existing in >10 patients were subjected to further analysis.

Genome doubling and loss of heterozygosity

The genome doubling (GD) status for each sample was inferred using the genome-doubling algorithm described in https://github.com/hartwigmedical/hmftools/blob/master/purity-ploidy-estimator/README.md based on the copy number profile inferred by ASCAT. A patient was then considered as a GD sample with GD events in at least one tumor region, and the GD event was compared between the Re and Rf groups based on Fisher's exact test. Using ASCAT, segments were defined as loss of heterozygosity (LOH) if the minor allele copy number was <0.25 (López et al., 2020). We combined all the LOH regions of multi-region tumor samples from each patient to represent their LOH region, and then calculated the proportion of the genome containing LOH events over the whole genome. The sample numbers and factions of the genome containing LOH events in Re and Rf groups were compared using Fisher's exact test and Wilcoxon rank-sum test, respectively.

Dynamic analysis on genomic alterations

To reveal the correlation of signatures, pathway alterations, and focal CNVs related to tumor recurrence, we performed dynamic statistical calculations on signature weights, pathway mutation frequencies, and focal CNV frequencies in the CHN-P cohort during the follow-up survey. First, the patients were divided into three subgroups: E-Re, L-Re, and Rf, as described in the above section of Subtype Classification and shown in Figures 1G, 3E, 3F, 4G, and S3E. Subsequently, the median contributions of each signature, mutation frequency of each pathway, or a focal CNV gene frequency at a specific time was calculated. We critically focused on the factors that showed significant differences between the Re and Rf groups.

Multilayer features for integrative analysis

For the tumor recurrence events, a list of 16 features was generated, including features of key elements in describing clinical or genomic features and those found to have stratifying effects on tumor recurrence and patient outcome (Figures 5A and 5B). Basic clinical features including age, sex, smoking, pathological type, and tumor size were included. CNV clonality (i.e., subclonal CNV percentage), FGA, GD, LOH, and ploidy were included to represent different aspects of chromatin instability. The molecular features of these patients showed significant differences between the Re and Rf groups, including TMB, SNV clonality (i.e., subclonal SNV percentage) and pathways of TLS and HRR. two CNV driver genes with significant effects on tumor recurrence were included. DNA intercross link (ICL)-double strand break (DSB)-related somatic gene mutations, TMB, subclonal SNV and CNV percentages, FGA and LOH percentage, which represented different aspects of genome instability, were included to calculate genome instability score. For continuous variables, exact values after normalized to between 0 to 1 were calculated; while for categorical variables, 1 and 0 were assigned for the presence or absence, respectively. For each patient, the accumulation of the above values was calculated to be the genome instability score.

Feature importance in DFS analysis

Methods to evaluate feature importance for predicting patient DFS time were applied (Chen et al., 2020). In a univariate Cox model, the hazard ratio and p value of the feature were calculated for predicting patient DFS time. To evaluate the importance of multivariate models, Cox proportional hazard models were used. The Cox models were fitted using the coxph function in the survival R package with default parameters. The importance of each feature for the Cox model was determined by the proportion of the Wald statistic of each feature among the sum of all Wald statistics of the model. To better understand differences among patients with good and bad outcomes, patients were divided evenly into two survival groups based on the median of predicted hazard from the multivariate Cox model with features.

ROC curves were generated to evaluate the performance of the prediction algorithm using the pROC(Robin et al., 2011) library in the R package. Sensitivity and specificity were estimated at the score cut-off that maximizes the sum of sensitivity and specificity using the ROCR library in the R package.

Plasma cfDNA sequencing

For the targeted sequencing of cfDNA, the pre-libraries were prepared by end repair, A tailing, T-adaptors ligation on both DNA fragment ends, and PCR amplification according to the cSMART technique (Jia et al., 2020; Lv et al., 2015)were performed, and an in-house designed panel for 457-gene targets was applied to capture cfDNA fragments to generate sequencing libraries. Sequencing libraries were applied on NovaSeq 6000 platform (Illumina) in the 150PE mode. The average sequencing depth was not less than 15,000 folds for each sample.

Bioinformatics analysis of ctDNA mutation

FASTP was used to trim adapters and to remove low-quality sequences to obtain clean reads. The clean reads were aligned to the human Ensemble GRCh37/hg19 reference genome performed by BWA. PCR duplications were processed by gencore (Chen et al., 2019c), and consensus reads were generated. SAMtools was applied for the detection of SNVs/Indels. The nonsynonymous SNVs/Indels with VAF > 0.5% and reads numbers >5, or with VAF > 0.1% and reads numbers >3 in cancer hotspots were regarded as true mutations (i.e. tumor-naïve ctDNA). If the number of true mutations ≥ 2 in tumor sample, the sample was considered to be ctDNA-positive. To study the tumor tissue mutations in blood plasma, we extracted the read coverage of identical mutations from the bam files of plasma samples. If the number of altered reads > 2, the cfDNA was considered a true mutation from tumor samples. If the number of true mutations ≥ 2 in tumor sample (i.e. tumor-informed ctDNA), the sample was considered to be ctDNA-positive. The mutation sites, detected in tumor tissues but without read support in the blood plasma samples, were regarded not to enter the blood system. CCF values of ctDNA-identical mutations in the tumor samples were also checked to verify ctDNA mutations (Figure 6E and Table S6). The correlation between ctDNA detection and tumor recurrence risk predicted by the multiple-feature model was analyzed using Fisher's one-tailed exact test.

RNA isolation and sequencing

Total RNA was isolated using RN28-EASYspin Plus RNA Mini Kit (Aidlab Biotechnologies Co., Ltd, China) from tumor and adjacent normal tissue samples. Sequencing libraries were constructed using the Ribo-Zero Magnetic (Human) kit (Epicentre Biotechnologies, USA) to remove ribosomal RNA (rRNA) following the manufacturer's instructions. Purified libraries were quantified using a Qubit® 4.0 Fluorometer (Life Technologies, USA) and validated using an Agilent 2100 bioanalyzer (Agilent Technologies, USA) to confirm the RNA integrity and insertion size, as well as calculate the mole concentration before constructing sequencing libraries. The libraries (2 X 150 bp paired-end reads) were finally sequenced on the Illumina NovaSeq 6000 (Illumina) to generate an average of 33.1 M clean reads (14.2–45.7 M) per sample for the CHN-P cohort (Table S2B).

RNA expression and functional analysis

Raw sequencing reads were quality controlled with FASTQC (v0.11.9; http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) before mapping to the human reference genome (hg19/GRCh37) using hisat2 (v2.1.0) (Kim et al., 2019). The read counts were calculated by HTseq (v0.11.0) (Anders et al., 2015), and gene expression was normalized as fragments per kilobase of exon model per million reads mapped (FPKM) (Table S7). The differentially expressed genes (DEGs) between tumor recurrence and recurrence-free samples were identified by edgeR with fold change ≥ 2 and p value < 0.05 (Robinson et al., 2010) and visualized using R package pheatmap. The pathway of enriched DEGs was conducted using metascape online tools (https://metascape.org/gp/index.html#/main/step1) (Zhou et al., 2019) based on the KEGG databases. For a specific pathway, the mean of all the gene expression values in the pathway was used to determine the gene expression level of the pathway. The median value of a specific gene or pathway among the cohort was used to divide the cohort into two groups: high and low levels (as expressed in Figure 6G).

Public datasets

We searched all published articles on WES analyses of NSCLCs and downloaded three publicly published mutation files with large amounts of Stage I NSCLCs and enough sequencing depth from articles of East Asian-ancestry LUADs as an East Asian cohort (EAS, n = 131) (Chen et al., 2020), from TRACERx Project as a European cohort (EUR-T, n = 61) (Jamal-Hanjani et al., 2017), and from TCGA repository (n = 277; https://portal.gdc.cancer.gov) as a TCGA cohort (Tables S1C–S1E), and then processed through the bioinformatics analysis pipeline of this study. Corresponding data accession numbers are listed in the key resources table.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical analysis methods or quantification methods for each analysis are described in the main text and referenced in the related method sections above in detail. The Kaplan–Meier method was applied for survival curve analysis, based on the LogRank test. Statistical analysis and data visualization were conducted using R software. For all the tests, ***p < 0.001, **p < 0.01, and *p < 0.05. All final graphs were further manually optimized using Adobe Illustrator.