Appendix 1 Supplementary Methods

Whole exome sequencing

DNA extraction

Fresh frozen tissues from tumor samples were used for genomic DNA extraction with QIAamp DNA FFPE Tissue Kit (QIAGEN) following the manufacturer's instructions. Genomic DNA of cellular sediments of pleural effusions were prepared with DNeasy Blood & Tissue kit (QIAGEN). Normal tissue DNA was sequenced together with tumor DNA samples for the purpose of identifying germline mutations. The DNA quality was assessed by Nanodrop2000 (Thermo Fisher Scientific) and the quantity was measured by dsDNA HS Assay Kit (Life Technologies) on Qubit 2.0.

Library preparation and sequencing

Extracted tumor genomic DNA was fragmented into 300–350 bp using Covaris M220 instrument (Covaris). Sequencing libraries were prepared with KAPA Hyper Prep kit (KAPA Biosystems) with optimized protocols. In brief, sheared tissue DNA were experienced with end-repairing, A-tailing, adapter ligation and size selection using Agencourt AMPure XP beads (Beckman Coulter). Libraries were then subjected to PCR amplification and purification before targeted enrichment.

DNA libraries from different samples were marked with unique indices during library preparation and up to 2 μg of different libraries were pooled together for targeted enrichment. Human cot-1 DNA (Life Technologies) and xGen Universal blocking oligos (Integrated DNA Technologies) were added to block nonspecific binding of library DNA to targeted probes. Customized xGen lockdown probes panel (Integrated DNA Technologies) were used to targeted enrich for AgilentV6+UTR predefined exon and UTR genes. The hybridization reaction was performed by using NimbleGen SeqCap EZ Hybridization and Wash Kit (Roche). Dynabeads M-270 (Life Technologies) was used to capture probe-bind fragments, followed by library amplification with Illumina p5 (5' AAT GAT ACG GCG ACC ACC GA 3') and p7 primers (5' CAA GCA GAA GAC GGC ATA CGA GAT 3') in KAPA HiFi HotStart ReadyMix (KAPA Biosystems), and purification by Agencourt AMPure XP beads. Library quantification was analyzed by KAPA Library Quantification kit (KAPA Biosystems). The size distribution of libraries was measured by Agilent Technologies 2100 Bioanalyzer (Agilent Technologies). The enriched libraries were sequenced on Hiseq 4000 NGS platforms (Illumina) to coverage depths of at 200x after removing PCR duplicates for FFPE.

Sequence Data Processing and Identification of Clinically-Actionable Mutations

Trimmomatic was used for FASTQ file quality control (QC). Leading/trailing low quality (below 15) or N bases were removed. Reads from each sample were mapped to the reference sequence hg19 (Human Genome version 19) using Burrows-Wheeler Aligner (BWA-mem, v0.7.12) with parameters (-t 8 -M). Local realignment around indels and base quality score recalibration were applied with the Genome Analysis Toolkit (GATK 3.4.0). GATK3.4.0 was applied to detect germline mutations from blood control samples. VarScan2 was employed for detection of somatic mutations (somatic P value =0.1, minimum quality score =15 and otherwise default parameters). Somatic variants presenting at less than 1% mutant allelic frequency in the paired control sample, but with at least 1% allelic frequency and at least 3 reads supporting variant alleles in tumor samples, were retained. We also filtered mutations population frequency ≥ 0.01 reported in the 1000 Genomes database, but still kept mutations if they were also present in COSMIC database (v76). Annotation was performed using ANNOVAR using the hg19 reference genome and 2014 versions of standard databases and functional prediction programs.

NGS panel of 425 genes

DNA extraction

5–8 of 10 μm tissue sections from tumor samples were used for genomic DNA extraction with QIAamp DNA FFPE Tissue Kit (QIAGEN) following the manufacturer's instructions. Genomic DNA of cellular sediments of pleural effusions were prepared with

https://dx.doi.org/10.21037/tlcr-21-751

DNeasy Blood & Tissue kit (QIAGEN). Normal tissue DNA was sequenced together with tumor DNA samples for the purpose of identifying germline mutations. The DNA quality was assessed by Nanodrop2000 (Thermo Fisher Scientific) and the quantity was measured by dsDNA HS Assay Kit (Life Technologies) on Qubit 2.0.

Library preparation and sequencing

Extracted tumor genomic DNA was fragmented into 300~350bp using Covaris M220 instrument (Covaris). Sequencing libraries were prepared with KAPA Hyper Prep kit (KAPA Biosystems) with optimized protocols. In brief, sheared tissue DNA were experienced with end-repairing, A-tailing, adapter ligation and size selection using Agencourt AMPure XP beads (Beckman Coulter). Libraries were then subjected to PCR amplification and purification before targeted enrichment.

DNA libraries from different samples were marked with unique indices during library preparation and up to 2 μg of different libraries were pooled together for targeted enrichment. Human cot-1 DNA (Life Technologies) and xGen Universal blocking oligos (Integrated DNA Technologies) were added to block nonspecific binding of library DNA to targeted probes. Customized xGen lockdown probes panel (Integrated DNA Technologies) were used to targeted enrich for 425 predefined genes. The hybridization reaction was performed by using NimbleGen SeqCap EZ Hybridization and Wash Kit (Roche). Dynabeads M-270 (Life Technologies) was used to capture probe-bind fragments, followed by library amplification with Illumina p5 (5' AAT GAT ACG GCG ACC ACC GA 3') and p7 primers (5' CAA GCA GAA GAC GGC ATA CGA GAT 3') in KAPA HiFi HotStart ReadyMix (KAPA Biosystems), and purification by Agencourt AMPure XP beads. Library quantification was analyzed by KAPA Library Quantification kit (KAPA Biosystems). The size distribution of libraries was measured by Agilent Technologies 2100 Bioanalyzer (Agilent Technologies). The enriched libraries were sequenced on Hiseq 4000 NGS platforms (Illumina) to coverage depths of at 500x after removing PCR duplicates for FFPE.

Sequence Data Processing and Identification of Clinically-Actionable Mutations

Trimmomatic was used for FASTQ file quality control (QC). Leading/trailing low quality (quality reading below 15) or N bases were removed. Reads from each sample were mapped to the reference sequence hg19 (Human Genome version 19) using Burrows-Wheeler Aligner (BWA-mem, v0.7.12) with parameters (-t 8 -M). Local realignment around indels and base quality score recalibration were applied with the Genome Analysis Toolkit (GATK 3.4.0). GATK3.4.0 was applied to detect germline mutations from blood control samples. VarScan2 was employed for detection of somatic mutations (somatic P value =0.1, minimum quality score =15 and otherwise default parameters). Somatic variant calls presenting at less than 1% mutant allelic frequency in the paired blood control sample, but with at least 1% allelic frequency and at least 3 reads supporting variant alleles in tumor samples, were retained. We also filtered mutations reported in dbSNP (v137) and the 1000 Genomes database, but still kept mutations if they were also present in COSMIC database (v76). Annotation was performed using ANNOVAR using the hg19 reference genome and 2014 versions of standard databases and functional prediction programs.

Genomic fusions were identified by FACTERA with default parameters. In short, we set minimum number of breakpoint-spanning reads to 5, minimum number of discordant reads to 2 and minimum similarity required for alignment of read to fusion template to 95%. Copy number variations (CNVs) were detected using ADTEx (http://adtex.sourceforge.net) with default parameters. The main advantage of ADTEx is that it can derive absolute copy numbers without any a priori knowledge of levels of normal DNA contamination or ploidy of the tumor samples. The algorithm takes not only depth of coverage (DOC) ratios but also allele frequency of germline heterozygous SNP (BAF) as inputs. The DOC ratios are smoothed by discrete wavelet transformation techniques prior to applying HMM to estimate polyploidy, normal contamination ratio and absolute CNVs. Germline CNVs from each patient were identified using the blood sample and normal human HapMap DNA sample NA18535 (Coriell Institute) for each captured region (exonic region). Somatic CNVs were identified using paired normal/tumor samples for each exon.

Calculation of TMS

Tumor mutation score (TMS) =1.0509288×mutation status of MUC4+0.26744718×mutation status of KRTAP10-6+ 0.1652883×mutation status of MEOX2+ 0.07419074×mutation status of NPIPB5+ 0.07250511×mutation status of FAM173B-

0.07430707×mutation status of ACP2+0.24224709×mutation status of PRDM7- 0.03014324×mutation status of SCN5A+ 0.23401033×mutation status of TCF20+0.24224673×mutation status of ZFHX4- 0.10444693×mutation status of NSRP1- 0.13372265. Gene mutation=1, and wild type=0.

Definition of TMB in WES and panel.

TMB was defined as the number of missense mutations per megabase of coding regions of the genome sequenced in WES. Panel TMB was counted by summing all base substitutions and indels in the coding region of targeted genes, including synonymous alterations to reduce sampling noise and excluding known driver mutations as they are over-represented in the panel.

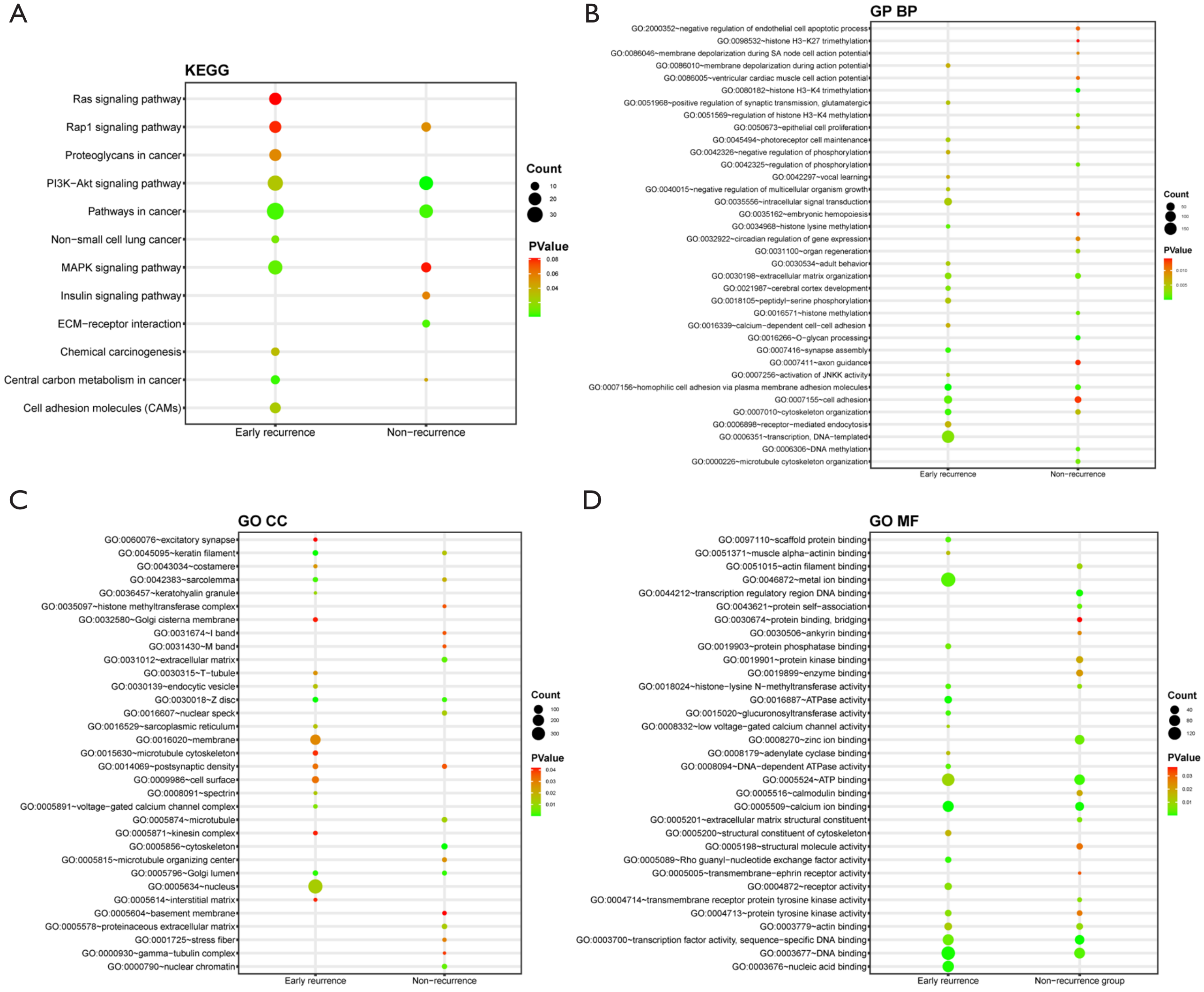

Figure S1 The mutant genes were analyzed by KEGG and GO enrichment for early recurrence and non-recurrence groups, respectively. KEGGE enrichment analysis found that mutant genes were significantly and only enriched in Ras signaling pathway in early recurrence group. KEGG, Kyoto Encyclopedia of Genes and Genomes; GO, Gene Ontology.

https://dx.doi.org/10.21037/tlcr-21-751

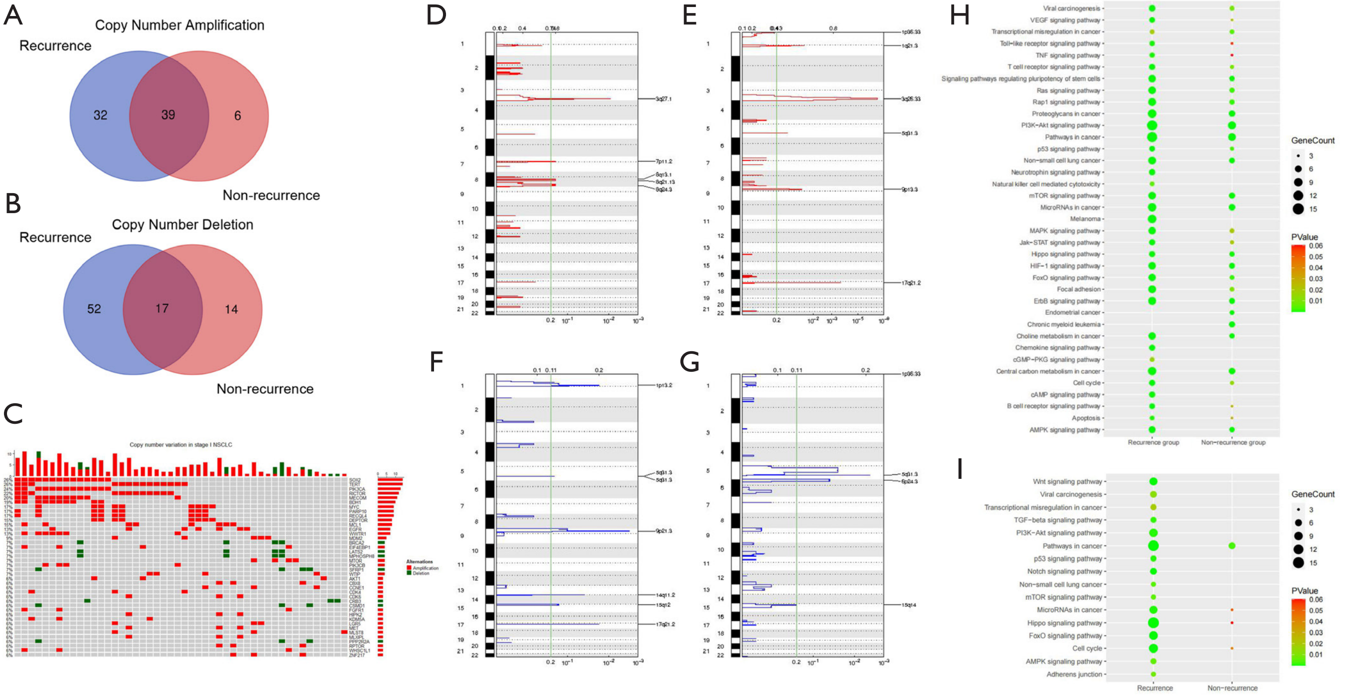

Figure S2 CNV in stage I NSCLC local cohort. (A) Thirty-nine genes with copy number amplification were identified in the early recurrence and non-recurrence groups, thirty-two genes were identified only in the early recurrence group, and six genes were only in the non-recurrence group. (B) Seventeen genes with copy number amplification were identified in the early recurrence and non-recurrence groups, 52 genes were identified only in the early recurrence group, and 14 genes were only in non-recurrence. (C) Top 40 CNVs in local cohort. Chromosome with copy number variations in local cohort: amplifications in chromosome 8 and deletions in chromosome 9, 14, and 17 were only identified in the early recurrence group. KEGG pathway enrichment analysis for copy number amplification (H) and deletion (I). CVN: Copy number variation; NSCLC, non-small cell lung cancer; KEGG, Kyoto Encyclopedia of Genes and Genomes.

https://dx.doi.org/10.21037/tlcr-21-751

Table S1 Clinical characteristics and early recurrence in TCGA cohort

Characteristic	Early recurrence (n=59)	Non-recurrence (n=38)	P value
Age (mean, years)	67	68	0.363
Gender (%)			0.605
Male	31 (52.5)	22 (57.9)	
Female	28 (47.5)	16 (42.1)	
Histologic type (%)			0.001
Adenocarcinoma	44 (74.6)	16 (42.1)	
Squamous carcinoma	15 (25.4)	22 (57.9)	
Smoking history (%)			0.130
Never smoker	9 (15.3)	2 (5.3)	
Former or current smoker	50 (84.7)	36 (94.7)	
Tumor sites (%)			0.114
Left	20 (33.9)	19 (50.0)	
Right	39 (66.1)	19 (50.0)	
Stage (%)			0.516
IA	24 (40.7)	18 (47.4)	
IB	35 (59.3)	20 (52.6)	

Table S2 Matched TCGA cohort by PSM

Clinical characteristic	Local cohort (n=74)	Matched TCGA cohort (n=74)	P value
Age (mean)	59.92 (8.64)	61.16 (9.77)	0.414
Gender = Male (%)	46 (62.2)	46 (62.2)	1
Histologic type = LUSC (%)	15 (20.3)	14 (18.9)	1
Tumor sites = Right (%)	40 (54.1)	45 (60.8)	0.506
Stage = IB (%)	49 (66.2)	50 (67.6)	1
Smoking history = Yes (%)	27 (36.5)	27 (36.5)	1

TCGA, the Cancer Genome Atlas; PSM, propensity score matching; LUSC, lung squamous carcinoma.

https://dx.doi.org/10.21037/tlcr-21-751