EUROPEAN RESPIRATORY *journal*

FLAGSHIP SCIENTIFIC JOURNAL OF ERS

Early View

Original article

Genomic characterisation of pulmonary subsolid nodules: mutational landscape and radiological features

Yanmeng Li, Xiao Li, Hao Li, Yifan Zhao, Ziyang Liu, Kunkun Sun, Xiang Zhu, Qingyi Qi, Bei An, Danhua Shen, Ruoyan Li, Taorui Liu, Jiahui Mi, Lientu Wang, Fan Yang, Fan Bai, Jun Wang

Please cite this article as: Li Y, Li X, Li H, *et al*. Genomic characterisation of pulmonary subsolid nodules: mutational landscape and radiological features. *Eur Respir J* 2019; in press (https://doi.org/10.1183/13993003.01409-2019).

Genomic characterization of pulmonary subsolid nodules: mutational landscape and radiological features

Yanmeng Li[1#], Xiao Li[1#], Hao Li[1#], Yifan Zhao[1#], Ziyang Liu[1#], Kunkun Sun[2], Xiang Zhu[3], Qingyi Qi[4], Bei An[4], Danhua Shen[2], Ruoyan Li[1], Taorui Liu[1], Jiahui Mi[1], Lientu Wang[5], Fan Yang[1]*, Fan Bai[1]*, Jun Wang[1]*

1. Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences & Department of Thoracic Surgery, People's Hospital, Peking University, Beijing 100871, China
2. Department of Pathology, Peking University People's Hospital, Beijing 100044, China
3. Department of Pathology, Peking University Third Hospital, Beijing 100191, China
4. Department of Radiology, Peking University People's Hospital, Beijing 100044, China
5. Berry Oncology Co., Ltd, Fujian 350200, China

[#] These authors contributed equally to the study

*** Correspondence:**

Fan Bai, Biomedical Pioneering Innovation Center (BIOPIC), Integrated Research Building Room 330, Peking University, Yiheyuan Road No. 5, Haidian District, Beijing, 100871, China. Tel: +86(10)62756164; Fax: +86(10)62760616; Email: fbai@pku.edu.cn

Fan Yang, Department of Thoracic Surgery, Peking University People's Hospital, No. 11 Xizhimen South Street, Beijing, 100044, China. Tel: +86(10)88326657; Email: yangfan@pkuph.edu.cn

Jun Wang, Department of Thoracic Surgery, Peking University People's Hospital, No. 11 Xizhimen South Street, Beijing, 100044, China. Tel: +86(10)88325594; Email: wangjun@pkuph.edu.cn

Summary

Our study provides the first comprehensive description of the mutational landscape of SSNs and reveals the potential genomic grounds supporting the use of radiological parameters as important clinical references for diagnosis and treatment of SSNs.

ABSTRACT

Background: Lung adenocarcinomas (LUADs) that radiologically display as subsolid nodules (SSNs) exhibit more indolent biological behavior than solid LUADs. SSNs, commonly encompassing preinvasive and invasive but early-stage adenocarcinomas, can be categorized as pure ground-glass nodules (pGGNs) and part-solid nodules (PSNs). The genomic characteristics of SSNs remain poorly understood.

Methods: We subjected 154 SSN samples from 120 treatment-naive Chinese patients to whole exome sequencing. Clinical parameters and radiological features of these SSNs were collected. The genomic landscape of SSNs and differences from that of advanced stage LUADs were defined. We also investigated the intratumor heterogeneity and clonal relationship of multifocal SSNs and conducted radiogenomic analysis to link imaging and molecular characteristics of SSNs. Fisher's exact and Wilcoxon rank sum tests were used in the statistical analysis.

Results: The median somatic mutation rate across the SSN cohort was 1.12 mutations/Mb. Mutations in *EGFR* were the most prominent and significant variation, followed by those in *RBM10, TP53, STK11,* and *KRAS*. The differences between SSNs and advanced-stage LUADs at

a genomic level were unraveled. Branched evolution and remarkable genomic heterogeneity were demonstrated in SSNs. Although multi-centric origin was predominant, we also detected early metastatic events among multifocal SSNs. Using radiogenomic analysis, we found that higher ratios of solid components in SSNs were accompanied by significantly higher mutation frequencies in *EGFR*, *TP53*, *RBM10*, and *ARID1B*, suggesting that these genes play roles in the progression of LUADs.

Conclusions: Our study provides the first comprehensive description of the mutational landscape and radiogenomic mapping of SSNs.

Keywords: pulmonary subsolid nodules (SSNs), early stage lung adenocarcinoma, whole exome sequencing, radiological features, radiogenomic mapping, somatic mutation

Introduction

Early-stage lung adenocarcinomas (LUADs) can present as a spectrum of radiological appearances, ranging from pure ground-glass nodules (pGGNs) generally thought to be indolent, to more aggressive solid LUADs. Radiological subsolid nodules (SSNs) can be further categorized as pGGNs, which are defined as nodules manifesting as an area of hazy increased lung opacity with preservation of bronchial and vascular margins, and part-solid nodules (PSNs) comprised of both ground-glass (GG) and solid components [1]. A proportion of pGGNs eventually develop into PSNs. Pathologically, malignant SSNs encompass preinvasive lesions (atypical adenomatous hyperplasias (AAHs) and adenocarcinomas in situ (AISs)), minimally invasive adenocarcinomas (MIAs) and invasive adenocarcinomas (IACs).

To date, comprehensive studies describing the genomic landscape [2-5] and clonal architecture [6, 7] of LUAD mainly focused on relatively late stage tumors. Genomic features of the very early stage of LUAD [8], especially the subgroup that radiologically displays as SSNs [9], remain poorly understood. Further, the diverse clinical trajectories of pGGNs and PSNs pose significant management challenges, and there have been limited predictive and prognostic biomarkers to risk-stratify patients. Radiological features of SSNs such as the subclassification (pGGN or PSN) [10], the solid size [11] and the consolidation tumor ratio (CTR) [12], can serve as important clinical references for diagnosis and treatment of this disease. However, the potential genomic grounds supporting the use of radiological parameters as important clinical references for diagnosis and treatment of SSNs were not deciphered.

In this study, we performed whole-exome sequencing (WES) of 154 surgical SSN specimens from 120 patients. We defined the genomic characteristics of SSNs and delineated potential driver

somatic mutations underlying the development and progression to advanced-stage LUAD. We also revealed the intratumor heterogeneity between GG and solid regions of the same SSN (n=11) and the clonal relationship among multifocal SSNs from the same patient (n=26). Finally, we provided radiogenomic mapping of SSNs. Our findings shed light on the evolution of LUAD and provide valuable biological and clinical insights into SSNs.

Methods

SSN cohort

Pathologically confirmed malignant tumor samples from 154 SSNs along with paired controls from 120 treatment-naive Chinese patients were subjected to WES. The study design was summarized in **Figure 1A-B**, and detailed clinical features of the cohort were summarized in **Table 1** and **Supplementary Table S1**. Of note, 11 large PSN samples were divided into 2-5 regions for WES according to the gross appearance of the resected tumor and the radiological characteristics (**Figure 1C, Supplementary Table S2**). Pathological diagnoses were classified as AAH, AIS, MIA, or IAC according to the 2015 WHO classification system (**Figure 1D**).

Radiological evaluation

The maximum diameters of whole SSNs and solid components on lung windows were measured (**Figure 3B**). The consolidation tumor ratio (CTR) was calculated as the ratio of the maximum diameter of consolidation divided by the tumor size. Volumetric parameters, such as solid volume percentage, were semi-automatically obtained using the Lung VCAR workstation (GE Healthcare). The SSN mass was calculated using the following equation: mass = volume ×[(mean CT attenuation+1000)×0.01]. Image analysis details were included in the

Supplementary Methods.

WES library preparation and data analysis

Protocols for genomic DNA extraction and WES library preparation were detailed in **Supplementary Methods**. We used the Genome Analysis Toolkit (GATK, version 3.8.0) to pre-process the WES data, and MuTect (version 1.1.4) and Strelka (version 2.8.4) to call SNVs and INDELs, respectively. Further details of bioinformatic analyses were provided in the **Supplementary Methods**. We used Fisher's exact test and Wilcoxon rank sum test for statistical analysis.

RESULTS

The mutational landscape of SSNs

The median somatic mutation rate across the SSN cohort (n=154) was 1.12 mutations/Mb (range 0.03-6.87), which is significantly lower than that of Chinese advanced stage LUADs [5] (**Supplementary Figure S1**). The median sequencing depth per tumor was 217. The mutational landscape of SSNs was summarized in **Figure 2A**. Twelve significant mutated genes were identified by the MutSigCV algorithm. Notably, mutations in *EGFR* were the most prominent and significant variation (50%), followed by mutations in *RBM10* (16%), *TP53* (13%), *STK11* (8%), and *KRAS* (5%). Other frequent but not significant mutated genes included *BIRC6* (5%), *ARID1B* (4%), *CHD2* (4%), *FBXO11* (4%), and *GNAQ* (4%). Concurrent and mutually exclusive mutation patterns in SSNs were shown in **Figure 2D**. Similar to previous reports in LUADs [3, 5, 13], mutations in *KRAS* were mutually exclusive with those in *EGFR*. Both *RBM10* and *TP53* mutations commonly co-occurred with *EGFR* mutations. Interestingly *TP53* was found to be co-altered in only 19% (15/77) of *EGFR* mutant population, lower than the anticipated 50-60% in early-stage solid tumors or metastatic disease [5, 14], suggesting that these combinations may facilitate the transformation of early stage lung neoplasm to malignancy [15].

We further explored the etiological factors that underlie mutagenesis in Chinese SSNs. Mutational spectrum analysis revealed a strong enrichment of C > T transversions and C > A transitions which are associated with a history of smoking (**Figure 2B**). From 154 SSNs, we confidently identified three mutational signatures, which displayed high similarity to COSMIC signatures 4, 6, and 5 (**Figure 2C**) [16]. COSMIC signature 4 is associated with smoking and has been frequently reported in LUADs [4, 5]. Intriguingly, COSMIC signature 6, which is often

identified in colorectal cancers but less in LUADs [2], is characteristic of tumors with defective DNA mismatch repair. These results suggested that smoking and dysregulation of DNA repair contribute to SSN tumorigenesis.

Comparing the genomic features of SSNs to advanced-stage LUADs

To unravel the differences between SSNs and advanced-stage LUADs at a genomic level, we compared the mutation frequencies of driver genes in 10 oncogenic pathways [17] and other previously reported mutations related to LUAD [3, 5]. The mutation frequencies of these genes in advanced-stage LUADs were calculated by reanalyzing data from TCGA (The Cancer Genome Atlas; 106 samples of stage II-IV LUADs [3]) and published Chinese LUADs (Kui Wu et al.; 75 samples of stage II-IV LUADs [5]) (**Figure 2E, Supplementary Figure S2**).

We first compared the mutation frequencies of oncogenic pathways in SSNs and stage II-IV LUADs and found the differences between the three groups (**Supplementary Table S3**). Next, we examined the frequencies of driver mutations in SSNs and Chinese stage II-IV LUADs (**Figure 2E, Supplementary Figure S2**). The smoking status of these two cohorts were similar (73.3% non-smokers in Chinese SSNs and 61.3% in Chinese stage II-IV LUADs). Of note, *EGFR* (50.0% vs. 30.7%) and *RBM10* (16.2% vs. 1.3%) mutations exhibited significantly higher frequencies in SSNs. Of all the 83 *EGFR* mutations from 77 SSNs, 60% (50/83) were L858R, and 20% (17/83) were exon 19 deletions, while the remaining (16/83) were other rare mutations (T790M, L858M, L861Q, L833F and so on) (**Supplementary Figure S3**). This was consistent with a previous study which reported that *EGFR* mutations were found in 64% of Japanese SSNs [9]. Interestingly, *RBM10* exhibited a significantly higher mutation frequency in SSNs with specific mutational

types (nonsense, frameshift, and splice site) (**Supplementary Figure S3**) [4] when compared to Chinese advanced-stage LUADs. *RBM10*, which encodes an RNA-binding protein, was found to be frequently mutated (7%) in Caucasian LUADs [4] and highly mutated in preinvasive and early-stage LUADs [18]. We speculated that alterations in RNA splicing alterations are a hallmark of SSNs but do not confer a sustained evolutionary advantage during progression to advanced stage adenocarcinoma.

A number of genes displayed notably lower mutation frequencies in SSNs when compared with Chinese stage II-IV LUADs (**Figure 2E)**. These included tumor suppressor genes in key oncogenic pathways, including *TP53*, *APC*, *FAT1/2/3/4*, *CRB1/2*, *NOTCH1/2/3/4*, *SPEN*, *ARID1A*, and oncogenes including *KRAS*, *ERBB4*, *ROS1*, *PIK3CA/B*, *TEAD1/2*, *NFE2L2*, as well as genes associated with cytoskeleton remodeling including *PHPN2* and *TRIO*. Above all, we speculated that *EGFR* and *RBM10* mutations are critical for SSN tumorigenesis, but progression of SSNs to advanced-stage LUADs is driven by mutations in other important cancer-related genes.

Genomic profile of intra-patient multifocal SSNs

Multiple synchronous occurrence is a feature of SSNs. To explore the clonal relationship between multifocal SSNs and solid tumors from the same patient, we sequenced 66 samples from 26 patients (2-5 per patient). The interlesional genomic profile (the median sequencing depth of these regions is 186.9) was distinct in 23/26 patients and generally confirmed the multi-centric origin of SSNs (**Supplementary Figure S4**).

Of note, however, P119 with 4 pGGNs (T1, T3-5) and 1 PSN (T2) was found to have a metastatic event (between T4 and T5). P119-T4 and P119-T5, both of which were located in the

apical-posterior segment of left upper lobe (LUL) and histologically presented as MIAs, were demonstrated to share 10 non-synonymous and 8 synonymous mutations (**Figure 3A**). The potential driver *MLLT1* mutation was shared by these two pGGNs, while *EGFR(L858R)*, *TP53*, and *PASK* mutations were exclusive to P119-T4. Taken together, multi-centric tumors and intrapulmonary metastatic lesions with a common clonal origin co-existed in this patient, which can be visualized in the regional distribution of mutations and phylogenetic tree.

We identified 3 PSNs (T1-T3) and 1 solid nodule (T4) of P114 that originated from the same clonal origin (**Figure 3B**). P114-T1 showed predominant solid features by CT scan (CTR: 0.87, solid volume percentage: 68%) mainly acinar in pathology. P114-T2 (CTR: 0.39, solid volume percentage, 30%) and P114-T3 (CTR: 0.66, solid volume percentage: 41%) were GG-predominant lesions with AIS and mainly acinar in their pathology, respectively. P114-T1 was very large and therefore underwent multiregion sequencing (R1-R5). The phylogenetic tree of this patient revealed that P114-T1-R1~R5, P114-T2, P114-T3, and P114-T4 all shared a substantial amount of mutations. Interestingly, P114-T1 displayed remarkable regional heterogeneity and the most central part (R5) of P114-T1 was the nearest to the putative metastases (P114-T2, P114-T3, and P114-T4) in evolutionary distance, suggesting that metastatic lesions of SSNs can either stay as SSNs or progress to solid tumors.

Intra-tumor heterogeneity between the GG and solid components of SSNs

To investigate the evolution and intra-tumor heterogeneity of SSNs, we performed multiregion sequencing on 35 spatially separate regions sampled from 11 large SSNs. The regional mutation distributions and phylogenetic trees of P118, P117, and P116 were shown in **Figure 3C,**

3D, and 3E, respectively (results for the remaining 8 patients are included in **Supplementary Figure S5**). Branched evolution was evident in SSNs, with potential driver mutations presenting both on the trunks and branches in the phylogenetic trees (**Figure 3C, 3D, and 3E**). This demonstrated that genomic heterogeneity and branched evolution are salient features of LUADs, even in their early stages when they radiologically display as SSNs.

Next, we focused on the difference between the GG and solid components of PSNs in the hopes of uncovering genomic factors that shape the distinct radiological phenotypes within the same lesion. Interestingly, we found that solid components had comparable or even less somatic mutations numbers than the GG components, and no significant enrichment of specific mutations was identified in GG or solid components. The evolutionary distance between regions of the same radiological phenotype varied among patients, as demonstrated in the phylogenetic trees (**Figure 3B, 3C, 3D, and 3E**). For example, R4 and R5 sampled from the solid component of P114-T1 were remarkably heterogeneous. The same result was observed among R1, R2, and R3, which were sampled from the GG component of P114-T1 (**Figure 3B**). In contrast, R2 and R3 from the GG component of P117-T displayed high genomic similarity, whereas R1, R4, and R5 from the solid component grouped together on the phylogenetic tree (**Figure 3D**), consistent with their separation in the radiological phenotype. In this regard, mutational features are not deterministically correlated with radiological features of different regions in the same SSN.

Radiogenomic analysis links imaging and molecular characteristics of SSNs

Radiological features of SSNs can serve as important clinical references for diagnosis and treatment. Given the clinical significance of these radiological parameters of SSNs, we

investigated the differences in the number of somatic mutations and patterns of mutated genes (for all genes with mutation frequencies > 3%) between any two groups separated by thresholds of each parameter with relevant clinical implications.

The mutation patterns were first correlated with some basic clinical features such as gender, smoking condition, and pathology (**Figure 4A(a)(b)(c)**). Driver mutations of SSNs, including *EGFR*, *RBM10*, *TP53*, *STK11*, displayed roughly the same mutation frequencies between genders, an observation which differed from previous studies [3, 5]. *KRAS* mutations were more commonly observed in males, although the difference was not statistically significant. We also found that *KTRAP1-3* mutations were significantly enriched in males, although *FLG* mutations were enriched in female patients ($P < 0.05$, Fisher's exact test). Compared with non-smokers and light smokers (< 20 pack-years), medium-heavy smokers (≥ 20 pack-years) had significantly greater numbers of *KRAS*, *KTRAP1-3*, and *ARID1B* mutations ($P < 0.01$ or $P < 0.001$, Fisher's exact test) along with a higher somatic mutation burden ($P = 0.01$, Wilcoxon rank sum test). When considering the pathological characteristics of the SSNs (**Supplementary Table S4**), the mutation frequencies of *EGFR*, *RBM10*, *TP53*, *ARID1B*, and *FBXO11* were significantly greater in SSNs that pathologically presented as IAC compared to those that presented as AAH, AIS, and MIA ($P < 0.05$, Fisher's exact test). We also found that IAC SSNs possessed a much higher somatic mutation load ($P < 0.001$, Wilcoxon rank sum test).

We next investigated the relationship between the radiological predictors of invasiveness and the genomic profiles of SSNs. Similar to the results found for pathology, PSNs in our cohort were enriched for *EGFR*, *RBM10*, *TP53*, and *ARID1B* mutations ($P < 0.05$, Fisher's exact test) and displayed significantly higher numbers of somatic mutations ($P < 0.001$, Wilcoxon rank sum test)

than did pGGNs (**Figure 4A(d)**). We then compared the patterns of mutation of the subgroups separated by total size and solid size (**Figure 4A(e)(f)**). Only mutation in *EGFR* was significantly enriched in SSNs with a total size >20 mm when compared to smaller SSNs ($P < 0.05$, Fisher's exact test). However, when considering the size of the solid component, in addition to *EGFR*, *TP53* was also significantly enriched in SSNs with a solid size >5 mm ($P < 0.01$, Fisher's exact test). This finding underscored the importance of mutation of *TP53* in driving the growth of the invasive tumor component (solid size) rather than the total SSN growth (total size). In contrast, mutation in *EGFR* may play an import role in both SSN growth and SSN invasiveness. Both groups (total size >20 mm and solid size >5 mm) exhibited much higher somatic mutation burdens than did other groups ($P < 0.01$, Wilcoxon rank sum test) (**Figure 4A(e)(f)**).

The consolidation tumor ratio (CTR), solid volume percentage, and SSN mass are additional predictive radiological parameters that are used to evaluate the tumor solid components (**Figure 4A(g)(h)(i)**). We found that the number of somatic mutations was significantly higher in groups with a CTR >0.25, a solid volume percentage >50%, and an SSN mass >250 mg, which suggested that increased solid component in SSNs is correlated with increased mutation events. *EGFR*, *RBM10*, and *TP53* mutations were significantly enriched in SSNs with a CTR ≥ 0.25 ($P < 0.01$, Fisher's exact test). Similar enrichments were also found when comparing subgroups classified by solid volume percentage and mass, though without statistical significance. However, these two classification groups both displayed a higher frequency of *ARID1B* mutations ($P < 0.01$, Fisher's exact test).

Discussion

In this study, we provided a comprehensive genomic landscape of Chinese SSNs. We highlighted that mutations in genes including *EGFR*, *RBM10*, *TP53*, *STK11*, *BIRC6*, and *KRAS* play important roles in driving SSN tumorigenesis and discovered mutational signatures associated with smoking and defective DNA mismatch repair, the latter of which is often identified in colorectal cancer [16] but is rare in LUADs [19]. Consistent with this observation, genes associated with DNA repair, such as *ATM* [20], *ASXL1* [21], *ATRX* [22], and *ARID1B* [23], were also mutated in SSNs as well as most early lesions of LUADs [15]. By comparing the genomic features of SSNs to those of advanced-stage LUADs, we were able to delineate the potential order of driver mutations during the progression of SSNs to invasive LUADs. These findings may explain why SSNs are able to remain indolent for many years.

Our study confirmed that most multifocal SSNs are tumors of multiple synchronous occurrence. However, we identified two cases with metastatic multifocal SSNs. Together with the two cases we reported previously [24], we found that SSNs can disseminate metastatic lesions while the metastatic lesions can remain as SSNs. The exact metastatic routes remain unknown, but are possibly through hematogenous, lymphatic, or airway metastasis. We demonstrated that genomic heterogeneity and branched evolution are present even in the early stages of LUADs as shown in a recent study [25], highlighting the complex evolutionary history of lung cancers and the therapeutic challenges we face.

It is widely accepted that radiological features of SSNs can serve as important clinical parameters for the diagnosis and treatment of SSNs. The total size and solid size of SSNs can predict their clinical behaviors. A more aggressive surveillance strategy should be applied to

patients with pGGNs with a total size >20 mm [26]. CTR, which takes both total size and solid size into account, also plays a critical role in the evaluation and treatment of SSNs [27]. Meanwhile, tumor volume and mass are novel indexes that reflect the invasiveness of SSNs [28]. By radiogenomic analysis, we revealed the potential genomic grounds supporting the use of these parameters as important clinical references for diagnosis and treatment of SSNs. To summarize, when considering the total size of SSNs, only mutation in *EGFR* was remarkably different between subgroups stratified by this parameter. However, when we took the presence (pGGN or PSN), the size (solid size), and the weight (CTR, solid volume percentage, mass) of solid components into consideration, we found that more solid components were accompanied by significantly higher mutation frequencies in *EGFR*, *TP53*, *RBM10*, and *ARID1B*. We revealed that the formation of the solid components of SSNs is potentially driven by additional mutations in cancer-related genes. Together with the relationship between pathological characteristics and genomic profiles of SSNs, higher mutation frequencies in *EGFR*, *TP53*, *RBM10*, and *ARID1B* were proven to relate to a more invasive behavior radiologically and pathologically.

Acknowledgements

This work was supported by the National Key Research and Development Program [2016YFC0900102 to F.B.]; the National Science and Technology Major Project [2018ZX10302205 to F.B.]; and the National Natural Science Foundation of China [81772469 to F.Y.].

We thank Weiwei Zhai from Institute of Zoology, Chinese Academy of Sciences for helpful discussions. We thank all the patients for their participation in this study.

Conflict of interest

FY, FB and JW designed and supervised the study. YFZ and YML carried out the experiments. YML, ZYL and RYL analyzed WES and clinical data. QYQ and BA performed the radiologic analyses of SSNs. KKS, XZ and DHS performed the histological examination. TRL and JHM collected the clinical data. YML compiled graphs and figures. YML, XL, HL and FB wrote the manuscript. All authors have read and agreed to the final version of the manuscript.

The authors declare that they have no competing interests.

Support statement

The WES data generated and analysed during the current study are available in the Genome Sequence Archive (Genomics, Proteomics & Bioinformatics 2017) in BIG Data Center (Nucleic Acids Res 2018), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession numbers CRA001568 (publicly accessible at http://bigd.big.ac.cn/gsa).

References

1. Naidich, D.P., et al., *Recommendations for the management of subsolid pulmonary nodules detected at CT: a statement from the Fleischner Society.* Radiology, 2013. **266**(1): p. 304-17.
2. Campbell, J.D., et al., *Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas.* Nat Genet, 2016. **48**(6): p. 607-16.
3. Cancer Genome Atlas Research, N., *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-50.
4. Imielinski, M., et al., *Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing.* Cell, 2012. **150**(6): p. 1107-20.
5. Wu, K., et al., *Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas.* Nat Commun, 2015. **6**: p. 10131.
6. de Bruin, E.C., et al., *Spatial and temporal diversity in genomic instability processes defines lung cancer evolution.* Science, 2014. **346**(6206): p. 251-6.
7. de Bruin, E.C., N. McGranahan, and C. Swanton, *Analysis of intratumor heterogeneity unravels lung cancer evolution.* Mol Cell Oncol, 2015. **2**(3): p. e985549.
8. Inamura, K., *Clinicopathological Characteristics and Mutations Driving Development of Early Lung Adenocarcinoma: Tumor Initiation and Progression.* Int J Mol Sci, 2018. **19**(4).
9. Kobayashi, Y., et al., *Genetic features of pulmonary adenocarcinoma presenting with ground-glass nodules: the differences between nodules with and without growth.* Ann

Oncol, 2015. **26**(1): p. 156-61.

10. Kakinuma, R., et al., *Natural History of Pulmonary Subsolid Nodules: A Prospective Multicenter Study.* J Thorac Oncol, 2016. **11**(7): p. 1012-28.
11. MacMahon, H., et al., *Guidelines for Management of Incidental Pulmonary Nodules Detected on CT Images: From the Fleischner Society 2017.* Radiology, 2017. **284**(1): p. 228-243.
12. Sun, F., et al., *Ground glass opacities: Imaging, pathology, and gene mutations.* J Thorac Cardiovasc Surg, 2018. **156**(2): p. 808-813.
13. Kadara, H., et al., *Whole-exome sequencing and immune profiling of early-stage lung adenocarcinoma with fully annotated clinical follow-up.* Ann Oncol, 2017. **28**(1): p. 75-82.
14. Nahar, R., et al., *Elucidating the genomic architecture of Asian EGFR-mutant lung adenocarcinoma through multi-region exome sequencing.* Nat Commun, 2018. **9**(1): p. 216.
15. Izumchenko, E., et al., *Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA.* Nat Commun, 2015. **6**: p. 8258.
16. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-421.
17. Sanchez-Vega, F., et al., *Oncogenic Signaling Pathways in The Cancer Genome Atlas.* Cell, 2018. **173**(2): p. 321-+.
18. Vinayanuwattikun, C., et al., *Elucidating Genomic Characteristics of Lung Cancer Progression from In Situ to Invasive Adenocarcinoma.* Sci Rep, 2016. **6**: p. 31628.

19. Takamochi, K., et al., *DNA mismatch repair deficiency in surgically resected lung adenocarcinoma: Microsatellite instability analysis using the Promega panel.* Lung Cancer, 2017. **110**: p. 26-31.

20. Li, G.M., *Mechanisms and functions of DNA mismatch repair.* Cell Res, 2008. **18**(1): p. 85-98.

21. Ismail, I.H., et al., *Germline mutations in BAP1 impair its function in DNA double-strand break repair.* Cancer Res, 2014. **74**(16): p. 4282-94.

22. Clynes, D., D.R. Higgs, and R.J. Gibbons, *The chromatin remodeller ATRX: a repeat offender in human disease.* Trends Biochem Sci, 2013. **38**(9): p. 461-6.

23. Watanabe, R., et al., *SWI/SNF factors required for cellular resistance to DNA damage include ARID1A and ARID1B and show interdependent protein stability.* Cancer Res, 2014. **74**(9): p. 2465-75.

24. Li, R., et al., *Early metastasis detected in patients with multifocal pulmonary ground-glass opacities (GGOs).* Thorax, 2018. **73**(3): p. 290-292.

25. Zhang, C., et al., *Genomic Landscape and Immune Microenvironment Features of Preinvasive and Early-Invasive Lung Adenocarcinoma.* J Thorac Oncol, 2019.

26. *National Comprehensive Cancer Network. NCCN guide-lines for NSCLC. http://www.nccn.org/professionals/physician_gls/pdf/nsclc.pdf.* Accessed April 2, 2019.

27. Yano, M., et al., *Survival of 1737 lobectomy-tolerable patients who underwent limited resection for cStage IA non-small-cell lung cancer.* Eur J Cardiothorac Surg, 2015. **47**(1): p. 135-42.

28. Wang, X.W., et al., *CT features differentiating pre- and minimally invasive from invasive*

adenocarcinoma appearing as mixed ground-glass nodules: mass is a potential imaging biomarker. Clin Radiol, 2018. **73**(6): p. 549-554.

Table 1. Clinical characteristics of 120 patients and 154 SSNs.

Characteristics	Value
Age (years)	
Median	59
Range	29-82
Gender	
Female	68 (56.7%)
Male	52 (43.3%)
Smoking status	
Smoker	32 (26.7%)
Non-smoker	88 (73.3%)
Lesion maximum diameter (mm)	
≤10.0	41 (26.6%)
10.1-20.0	68 (44.2%)
>20.0	45 (29.2%)
SSN type	
Pure GGN	66 (42.9%)
Part-solid Nodule	87 (56.5%)
NA	1 (0.6%)
Solid maximum diameter (mm)	
⩽5	79 (51.3%)
>5	74 (48.1%)
NA	1 (0.6%)
CTR	
0-0.25	68 (44.2%)
>0.25	85 (55.2%)
NA	1 (0.6%)
Solid volume percentage (%)	
⩽50	120 (77.9%)
>50	22 (14.3%)
NA	12 (7.8%)
SSN mass (mg)	
⩽250	53 (34.4%)
>250	89 (57.8%)
NA	12 (7.8%)
Histologic type	
AAH	14 (9.1%)
AIS	19 (12.3%)
MIA	30 (19.5%)
IAC	91 (59.1%)
Tumor stage*	
NA (AAH)	14 (9.1%)
0	19 (12.3%)

IA1	74 (48.1%)
IA2	33 (21.4%)
IA3	3 (1.9%)
IB	8 (5.2%)
IIA	1 (0.6%)
IIB	2 (1.3%)

* According to the International Union Against Cancer (UICC) eighth TNM staging.

NA, not applicable; AAH, atypical adenomatous hyperplasia; AIS, adenocarcinoma *in situ*; MIA, minimally invasive adenocarcinoma; IAC, invasive adenocarcinoma; SSN, subsolid nodule; GGN, ground glass nodule.

Figure legends

Figure 1. Study overview and sample information for 154 SSNs.

(A) Schematic of the study design.

(B) Detailed information about the number of patients and samples assigned to different research strategies.

(C) Schematic drawings of a pGGN and a PSN with corresponding CT images under the lung window setting. Red arrows indicate sites of SSNs.

(D) Histologic features (hematoxylin and eosin staining) of SSNs including AAH, AIS, MIA, and IAC. Scale bar, 100 μm.

Figure 2. The mutational landscape and signatures of Chinese SSNs

(A) Mutational landscape of nonsynonymous somatic mutations, including SNVs (single nucleotide variants) and INDELs (insertions and deletions). Samples displayed as columns are ordered by the number of nonsynonymous somatic mutations, as shown in the upper panel. Clinical features are annotated in the middle panel. The heatmap below displays the somatic mutated genes ordered by their respective mutation frequencies. Twelve significant genes with a corrected q value < 0.1 were identified by the MutSigCV algorithm and are labeled (*).

(B) 'Lego' plots displaying the frequency of base substitutions within specific trinucleotide mutational contexts.

(C) Mutational signatures of Chinese SSNs.

(D) Concurrent and mutually exclusive somatic mutation patterns of significantly mutated genes. Significance was calculated using Fisher's exact test (*: $p < 0.05$).

(E) Comparison of the mutation frequencies of important genes in cancer-associated pathways between Chinese SSNs and Chinese advanced-stage LUADs (stage Ⅱ-Ⅳ LUADs) [5]. Oncogenes labeled in red and tumor

suppressor genes labeled in blue are listed on the horizontal axis ordered by the corresponding pathways. The vertical axis represents the mutation frequencies obtained from different cohorts. Significance was calculated using Fisher's exact test (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$).

Figure 3. Inter-tumor and intra-tumor heterogeneity of SSNs.

Heatmaps on the left of each panel depict the regional distribution of all of the somatic mutations among different sequencing sites of multiregion SSNs and multifocal SSNs. Phylogenetic trees on the right of each panel were constructed for each patient using the maximum parsimony algorithm. The color of each line corresponds to the categories of mutations shown in the heatmaps, and the lengths of the trunks and branches are proportional to the number of mutations in each sample. Spatial location information of multi-focal samples in P114(B) and P119(A) is represented in the three-dimensional reconstructions with the corresponding CT images on the side. L: left lobe. R: right lobe.

Figure 4. Clinical enrichment analysis of mutated genes by different clinical criterion groups.

(A) Both barplots, which indicate the mutation frequencies of significantly mutated genes as well as additional non-significantly enriched genes (for all genes with mutation frequencies > 3%), and violin plots, which show the somatic mutation number, demonstrate the comparisons between sample groups classified by (a) gender, (b) smoking condition with a cutoff of 20 pack-years, (c) pathology, (d) radiological SSN classification (pGGN or PSN), (e) total size with a cutoff of 20 mm, (f) solid component size with a cutoff of 5 mm, (g) CTR with a cutoff of 0.25, (h) solid volume percentage with a cutoff of 50%, and (i) SSN mass with a cutoff of 250 mg. For the barplots, significance was calculated by Fisher's exact test (*: $p < 0.05$, **: $p < 0.01$, ***: $p < 0.001$); for the violin plots, the p value was calculated by Wilcoxon rank sum test.

(B) Schematic diagrams and formulas for calculation of several of the clinical criteria mentioned in (A).

Supplementary figure legends

Supplementary Figure S1. Comparison of the number of nonsynonymous somatic mutations between Chinese SSNs (*N*=154) and Chinese advanced-stage LUADs (*N*=75).

Supplementary Figure S2. Comparison of the frequencies of nonsynonymous mutations in key genes in cancer-associated pathways between Chinese SSNs (*N*=154) and advanced-stage LUADs (TCGA stage Ⅱ-ⅣLUADs, *N*=154; Chinese stage Ⅱ-ⅣLUADs, *N*=75).

Each gene box includes three percentages representing the nonsynonymous mutation frequencies of activation and inactivation of three cohorts. The data layout is described in the legend. Genes are grouped by their signaling pathways. Interactions between genes are indicated.

Supplementary Figure S3. Lollipop plots for amino acid changes resulting from *EGFR* and *RBM10* mutations.

Locations of *EGFR* and *RBM10* mutations on the protein structure. Mutation types and protein domains are indicated below. The vertical axis represents the number of occurrences of each mutation site.

Supplementary Figure S4. Regional distribution of all somatic mutations in multifocal tumors (including SSNs and solid tumors) in the same patient.

The regional distribution of all of the somatic mutations of 66 tumors from 26 patients is depicted in the heatmaps. Each column represents a single mutation site. Grey represents wild type, while blue represents mutation of a certain site in a certain gene.

Supplementary Figure S5. Regional distribution of all somatic mutations in multiregions of the same SSN.

The regional distribution of all of the somatic mutations of 35 spatially separate regions sampled from 11 large SSNs is depicted in the heatmaps. Each column represents a single mutation site. Grey represents wild type, while blue represents mutation of a certain site in a certain gene.

Table legends

Table 1. Clinical characteristics of 120 patients and 154 SSNs.

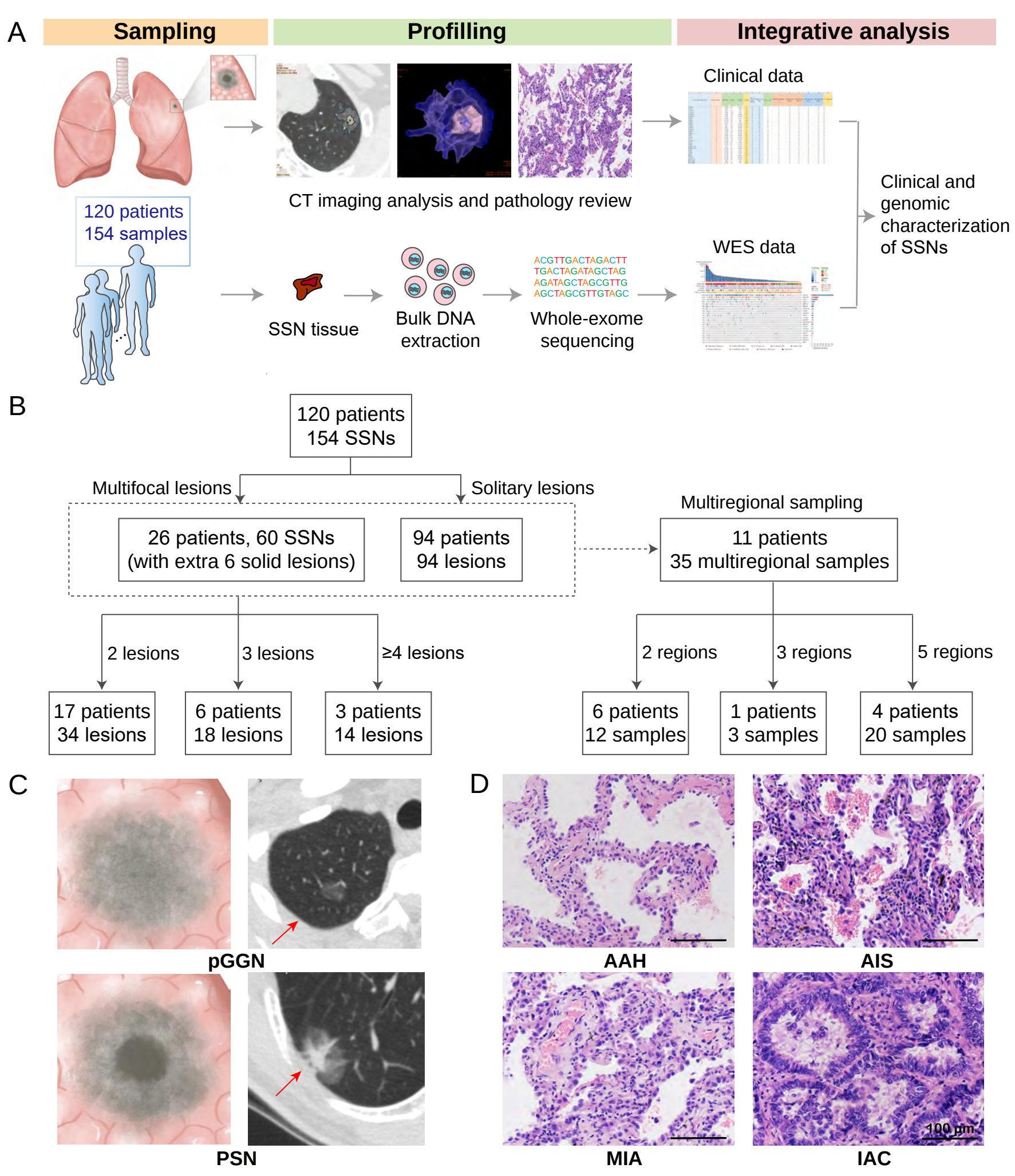
Supplementary Table legends

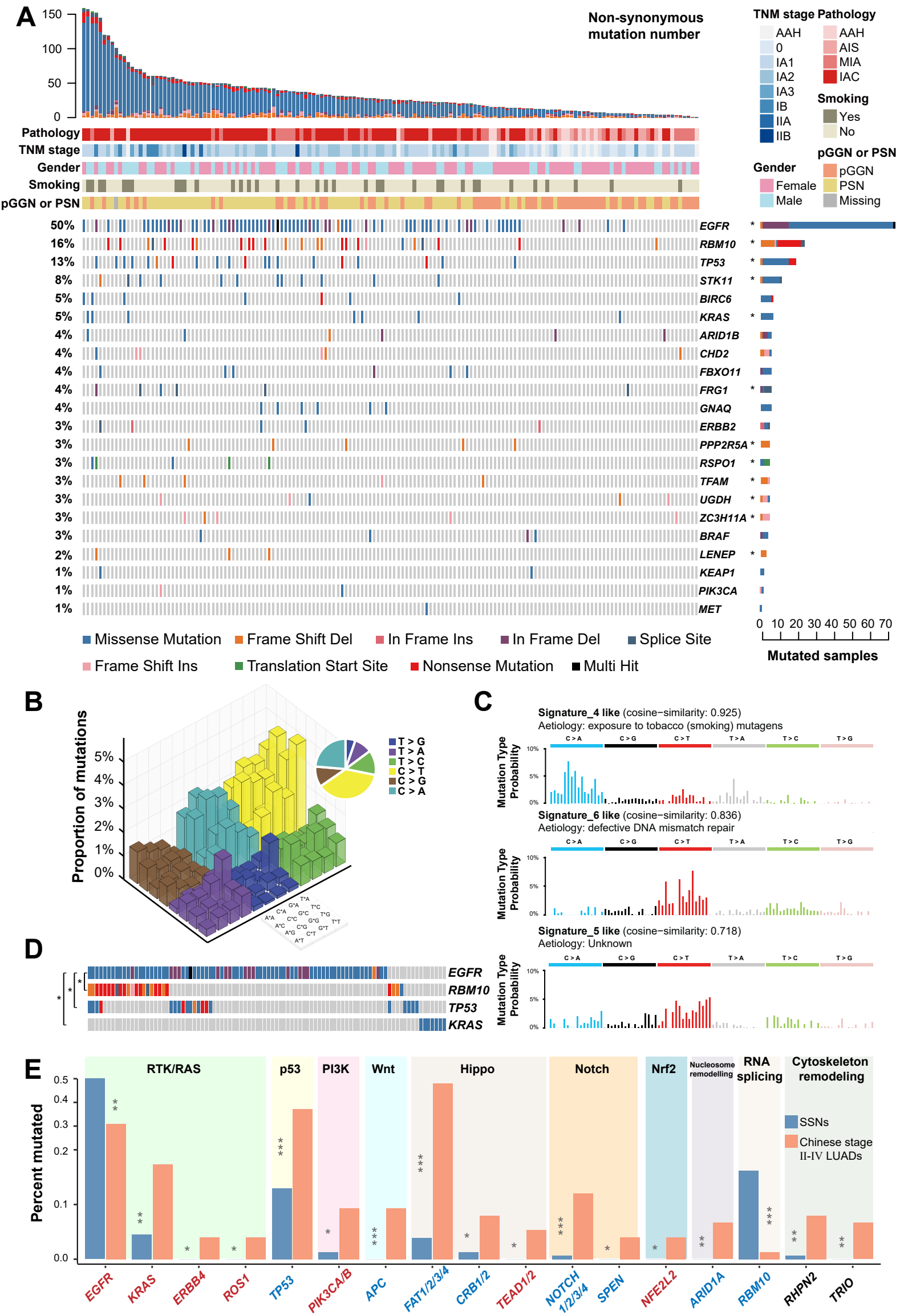
Supplementary Table S1. Detailed clinical information for 120 patients and 154 SSNs

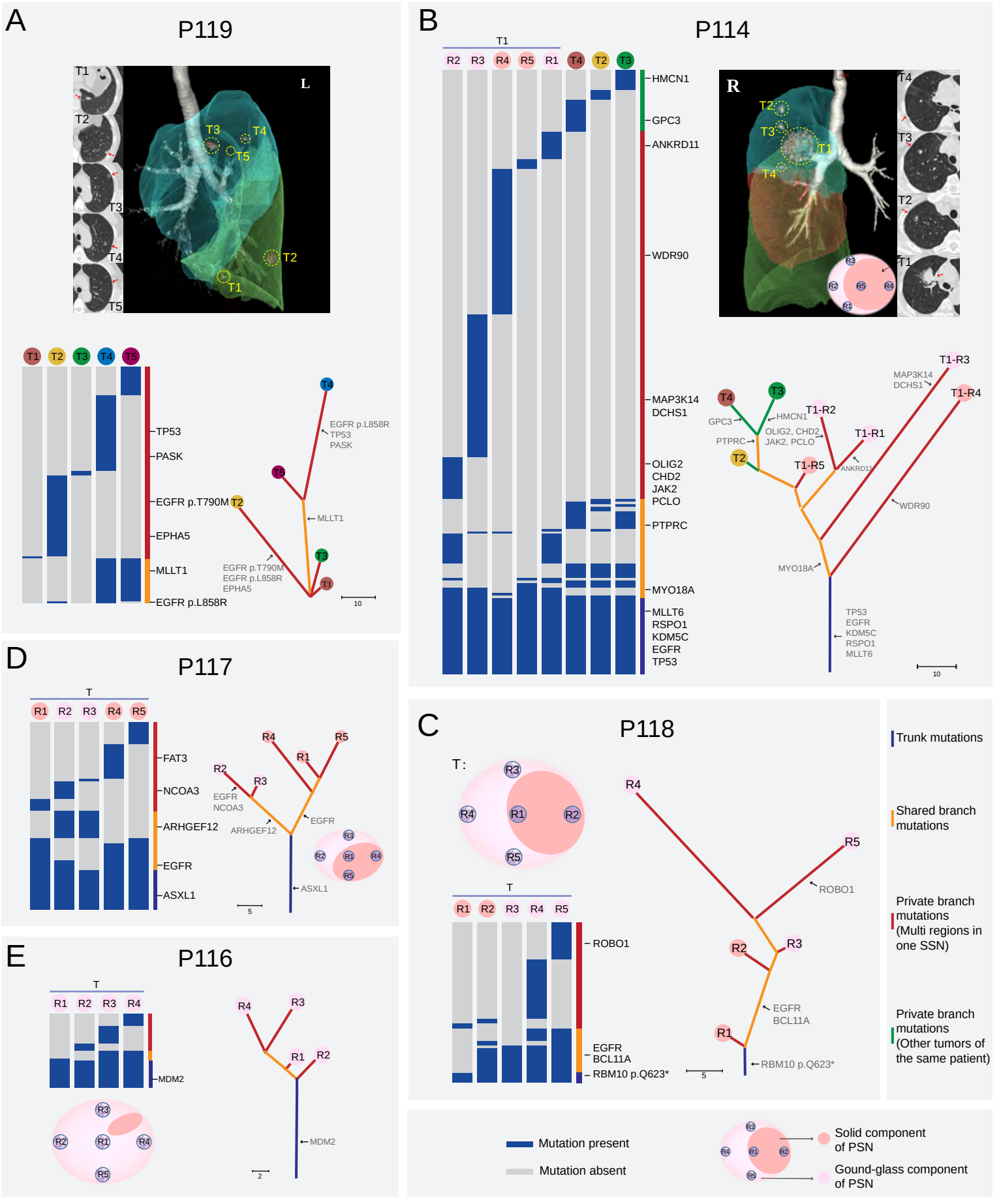
Supplementary Table S2. Detailed information for multi-region and multi-focal tumors

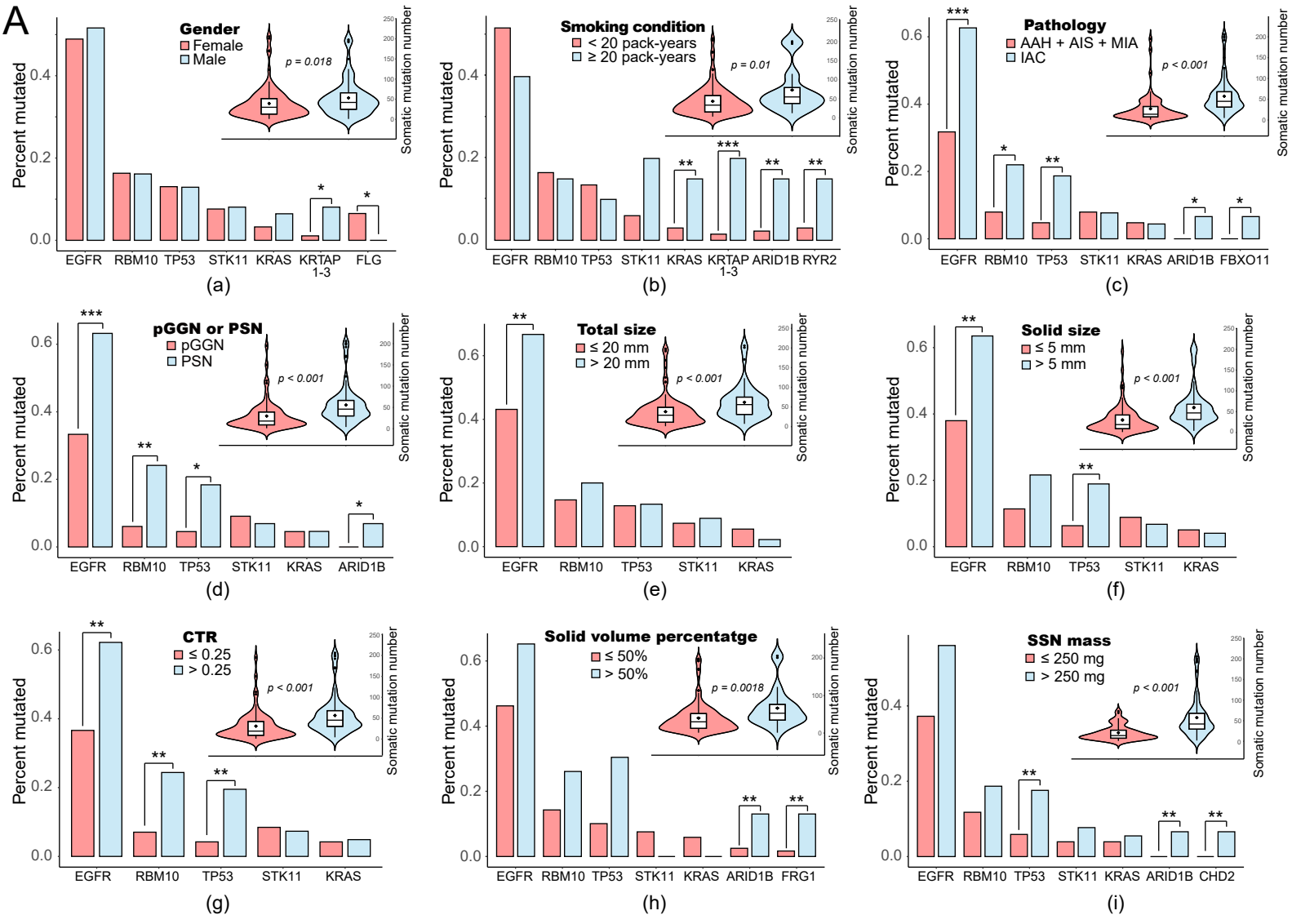
Supplementary Table S3. Comparison of altered pathway frequencies between Chinese SSNs, Chinese advanced-stage LUADs, and TCGA advanced-stage LUADs

Supplementary Table S4. Comparison of mutation rates of some important genes of each pathological group (AAH and AIS, MIA, IAC)

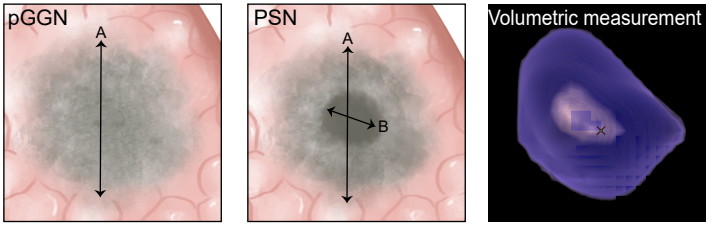

A
Sampling
Profilling
Integrative analysis
120 patients
154 samples
CT imaging analysis and pathology review
Clinical data
Clinical and genomic characterization of SSNs
SSN tissue
Bulk DNA extraction
Whole-exome sequencing
WES data
B
120 patients
154 SSNs
Multifocal lesions
Solitary lesions
26 patients, 60 SSNs
(with extra 6 solid lesions)
94 patients
94 lesions
Multiregional sampling
11 patients
35 multiregional samples
2 lesions
3 lesions
≥4 lesions
17 patients
34 lesions
6 patients
18 lesions
3 patients
14 lesions
2 regions
3 regions
5 regions
6 patients
12 samples
1 patients
3 samples
4 patients
20 samples
C
pGGN
PSN
D
AAH
AIS
MIA
IAC
100 μm


A
Non-synonymous mutation number
150
100
50
0
TNM stage
Pathology
AAH
0
IA1
IA2
IA3
IB
IIA
IIB
AAH
AIS
MIA
IAC
Smoking
Yes
No
pGGN or PSN
pGGN
PSN
Missing
Gender
Female
Male
Pathology
TNM stage
Gender
Smoking
pGGN or PSN
50% EGFR *
16% RBM10 *
13% TP53 *
8% STK11 *
5% BIRC6
5% KRAS *
4% ARID1B
4% CHD2
4% FBXO11
4% FRG1 *
4% GNAQ
3% ERBB2
3% PPP2R5A *
3% RSPO1 *
3% TFAM *
3% UGDH *
3% ZC3H11A *
3% BRAF
2% LENEP *
1% KEAP1
1% PIK3CA
1% MET
0 10 20 30 40 50 60 70
Mutated samples
Missense Mutation
Frame Shift Del
In Frame Ins
In Frame Del
Splice Site
Frame Shift Ins
Translation Start Site
Nonsense Mutation
Multi Hit
B
Proportion of mutations
5%
4%
3%
2%
1%
0%
T > G
T > A
T > C
C > T
C > G
C > A
C
Signature_4 like (cosine−similarity: 0.925)
Aetiology: exposure to tobacco (smoking) mutagens
Signature_6 like (cosine−similarity: 0.836)
Aetiology: defective DNA mismatch repair
Signature_5 like (cosine−similarity: 0.718)
Aetiology: Unknown
Mutation Type Probability
C > A
C > G
C > T
T > A
T > C
T > G
D
EGFR
RBM10
TP53
KRAS
E
Percent mutated
0.5
0.4
0.3
0.2
0.1
0.0
RTK/RAS
p53
PI3K
Wnt
Hippo
Notch
Nrf2
Nucleosome remodelling
RNA splicing
Cytoskeleton remodeling
SSNs
Chinese stage II-IV LUADs
EGFR
KRAS
ERBB4
ROS1
TP53
PIK3CA/B
APC
FAT1/2/3/4
CRB1/2
TEAD1/2
NOTCH 1/2/3/4
SPEN
NFE2L2
ARID1A
RBM10
RHPN2
TRIO


A
P119
T1
T2
T3
T4
T5
L
TP53
PASK
EGFR p.T790M
EPHA5
MLLT1
EGFR p.L858R
EGFR p.L858R
TP53
PASK
MLLT1
EGFR p.T790M
EGFR p.L858R
EPHA5
10
B
P114
T1
R2 R3 R4 R5 R1 T4 T2 T3
HMCN1
GPC3
ANKRD11
WDR90
MAP3K14
DCHS1
OLIG2
CHD2
JAK2
PCLO
PTPRC
MYO18A
MLLT6
RSPO1
KDM5C
EGFR
TP53
R
T2
T3
T1
T4
T1-R3
MAP3K14
DCHS1
T1-R4
T4
T3
GPC3
HMCN1
T1-R2
PTPRC
OLIG2, CHD2
JAK2, PCLO
T1-R1
T2
T1-R5
ANKRD11
WDR90
MYO18A
TP53
EGFR
KDM5C
RSPO1
MLLT6
10
C
P118
T:
R3
R4
R1
R2
R5
T
R1 R2 R3 R4 R5
ROBO1
EGFR
BCL11A
RBM10 p.Q623*
R4
R5
ROBO1
R2
R3
EGFR
BCL11A
R1
RBM10 p.Q623*
5
Trunk mutations
Shared branch mutations
Private branch mutations (Multi regions in one SSN)
Private branch mutations (Other tumors of the same patient)
D
P117
T
R1 R2 R3 R4 R5
FAT3
NCOA3
ARHGEF12
EGFR
ASXL1
R4
R5
R1
R2
R3
EGFR
NCOA3
ARHGEF12
EGFR
R3
R2
R1
R4
R5
ASXL1
5
E
P116
T
R1 R2 R3 R4
MDM2
R3
R2
R1
R4
R5
R4
R3
R1
R2
MDM2
2
Mutation present
Mutation absent
R3
R4
R1
R2
R5
Solid component of PSN
Gound-glass component of PSN



B



A: Total size of SSN

B: Solid size of PSN

CTR (The consolidation tumor ratio) = Solid size (B) / Total size (A)

Solid volume percentage = Solid component volume / Tumor total volume

SSN mass = Volume * [(Mean CT attenuation+1000) * 0.01]

SUPPLEMENTARY MATERIAL

SUPPLEMENTARY METHODS

SSN cohort

For this study, we selected 120 patients with 154 pulmonary SSNs from patients diagnosed with lung adenocarcinoma (including preinvasive tumors) between January 2015 and December 2018 at Peking University People's Hospital who underwent surgical resection of their tumors prior to receiving any form of therapy. Ninety-four patients had solitary SSN lesions, and 26 patients had multifocal pulmonary lesions with 6 extra solid lesions. Samples of all 154 SSNs and 6 solid lesions from these patients were subjected to whole-exome sequencing (WES). Radiological and clinicopathologic features were reviewed. All patients were free of extra thoracic metastasis. Detailed clinical characteristics are provided in **Table 1**. Informed consent was obtained from all participating patients. The institutional review boards (IRB) of both Peking University People's Hospital and Peking University approved this study.

Radiological evaluation

All standard dose non-contrast chest CT imaging was carried out within 4 weeks prior to surgery. Results were evaluated by two thoracic radiologists. Twelve patients were excluded from imaging analysis due to the lack of chest CT at Peking University People's Hospital. CT scans of the other 108 patients were reconstructed using the standard kernel with a slice thickness of 1.0 mm or 1.25 mm to assure adequate volumetric analysis. Two thoracic radiologists independently assessed the CT images under the lung window setting (window level, 1500 HU; width, -700 HU) using a picture archiving and communication system (PACS). A consensus was reached after mutual discussion and/or consultation with a third radiologist, in case of disagreements.

The maximum long-axis of the largest cross-sectional area of the lesions was measured as the total size, while the maximum long-axis of the solid component was measured separately as the solid size [1]. The consolidation tumor ratio (CTR) was defined as the ratio of the maximum diameters of consolidation to the tumor size on a thin-section CT scan **(Fig 3B)** [2]. The nodule volume, solid portion volume, solid volume percentage, and mean CT attenuation were semi-automatically obtained using the CT lung analysis workstation Lung VCAR (GE Healthcare) by a single thoracic radiologist. If the segmentation and the nodule were visually mismatched, the radiologist adjusted the nodule margin and solid component margin manually. It should be noted that measurement of nodule diameter and volume, both semi-automatically and manually, is subject to inter- and intra-reader variability to some extent.

The mean CT attenuation was recorded as lung density. SSN mass is a parameter which integrates volume and density, and is thought to enable early detection of SSN growth [3]. Therefore, we calculated the mass of SSNs using the following equation: mass = volume ×[(mean CT attenuation+1000)×0.01].

Sample processing and DNA isolation

Specimens underwent gross pathologic examination by experienced pathologists to estimate their clinicopathological subtypes. Adjacent normal lung tissue or blood was used as the matched normal control for all patients. Formalin-fixed, paraffin-embedded and fresh frozen specimens were used for genomic DNA extraction. All the FFPE samples subjected to WES and data analysis were of high quality and passed the strict quality control of genomic DNA isolation and whole exome library preparation. For the paraffin-embedded group, specimens were cut into 10-15 consecutive tissue slides with 10 μm in thickness, and genomic DNA extraction was performed using the GeneRead DNA FFPE Kit (Qiagen, Germany). For the fresh frozen group, tumor tissue samples approximately 5 × 5 × 5 mm in size were used for genomic DNA extraction using QIAamp DNA Micro Kit (Qiagen, Germany). The size of the DNA was checked using the Fragment AnalyzerTM platform (Advanced Analytical Technologies, Inc).

Of note, for 11 large PSN tumors from 11 patients, specimens were divided into 2-5 multiregion sections according to gross appearance of the resected tumor and the radiological features. For other large PSN tumors that were not divided into different parts, we cut a 3-mm tumor slice through the section of the tumor with the largest diameter, cut the slice into small pieces, and selected several pieces for DNA extraction in order to eliminate tumor heterogeneity.

Whole exome library preparation and sequencing

For each sample, 200 ng -1 μg of genomic DNA was fragmented into pieces approximately 300 bp in size using the Covaris ultrasonic system (Covaris). Libraries were constructed using NEBNext Ultra DNA Library Prep Kit for Illumina (New England Biolabs) according to the manufacturer's protocol. The fragmented DNA was blunted with 5'-phosphorylated ends using the NEB Quick Blunting Kit and ligated to adaptors using NEBNext Quick Ligation Module. Purification was conducted using Agencourt AMPure XP SPRI beads (Beckman Coulter), and DNA fragment size was assessed using the Fragment AnalyzerTM platform (Advanced Analytical Technologies, Inc). The exonic regions of each sample were captured with the SureSelect All Exon V6 kit (Agilent Technologies) according to the manufacturer's protocol. The amplified library was further purified and quality checked. Sequencing was performed on the NovaSeq™ platform (Illumina, San Diego) to generate 2×150 bp paired-end (PE) reads.

Processing of raw sequencing reads

The adapter sequences introduced during library preparation were trimmed from the raw paired-end sequencing reads using Cutadapt (version 1.14) [4]. If either read of a read pair was shorter than 35 base pairs after adapter removal, the read pair was removed. Several criteria were further applied to the trimmed reads to remove low quality reads: 1) Greater than half of the bases in the read had a base quality score < 20; and 2) Greater than 10% of the bases were undetermined bases (N). If either read of a read pair met one of the above criteria, the read pair was removed.

WES data analysis to identify SNVs and indels

For all of the following analyses, default parameters were used unless otherwise indicated. Clean paired-end reads were aligned to the human genome build hg19 (UCSC) using the Burrows-Wheeler Aligner (version 0.7.15) [5]. Aligned SAM files of each library were sorted and formatted to BAM files and were merged using Picard (version 2.11.0) (http://broadinstitute.github.io/picard/). Duplicate reads were identified using Picard, and the aligned reads were realigned to hg19 and the base quality scores were recalibrated using the Genome Analysis Toolkit (GATK, version 3.8.0).

SNVs and small indels were called using MuTect (version 1.1.4) [6] and Strelka (version 2.8.4) [7], respectively. A series of filtering criteria were applied to the raw variant candidates to identify the credible SNVs and indels: For SNVs, 1) At least 10X total coverage and 3X variant coverage was required in tumor samples. 2) At least 10X coverage was required for the matched normal sample of each patient. 3) The variant allele frequency was equal or greater than 5%. For indels, 1) All calls marked with PASS in the VCF file were retained. 2) At least 10X total coverage and 3X variant coverage was required in tumor samples. 3) At least 10X coverage was required for the matched normal sample of each patient and no mutated read presented. 4) The variant allele frequency was equal or greater than 5%.

The filtered candidate variants were annotated using SnpEff (version 3.0) [8]. All variants listed in the National Heart, Lung, and Blood Institute Exome Sequencing Project were removed. All variants listed in dbSNP 138 (https://www.ncbi.nlm.nih.gov/snp) were removed unless they were also documented in the Catalog of Somatic Mutations in Cancers (COSMIC) databases (version 68). All mutations that passed the filtering were manually reviewed using "Samtools tview" to further eliminate mutations in poorly mapped reads. Of note, for multiregional samples, we used the union set of mutations identified in all regions of the same lesion for further analyses and statistics.

Identification of significantly mutated genes

To identify significantly mutated genes, MutSigCV (version 1.41) algorithm [9] was applied using default parameters. The algorithm estimates the background mutation rate for each gene–patient–category combination based on the observed silent mutations in the gene and non-coding mutations in the surrounding regions.

Mutational signature analysis

The mutational signatures of our cohort were analyzed using non-negative matrix factorization as implemented in the R package NMF [10]. Base substitutions (silent and non-silent mutations included) can be divided into 6 directions, namely, C > T, C > A, C > G, T > C, T > G, and T > A. We constructed a matrix of nucleotide substitutions classified into 96 classes based on the mutation direction and the immediate bases surrounding the mutated base, as extracted from the hg19 (UCSC) reference genome [11]. The NMF algorithm was then applied to deconstruct the mutational signatures. We used cosine similarity as a metric to compare the similarities of our signatures to the 30 signatures in COSMIC (https://cancer.sanger.ac.uk/cosmic/signatures).

Identification of potential driver mutations

To identify potential driver genes in our cohort, we assessed three collections of genes for each patient sample: 1) significantly mutated genes reported by other studies of lung adenocarcinoma [12-14]; 2) all cancer genes documented in the oncoKB cancer gene list (http://oncokb.org/cancerGenes, last update 1/24/2019); and 3) genes documented in the oncogenic signaling pathways in The Cancer Genome Atlas [15]. Mutations that met one of the above criteria were considered as potential driver mutations and were labelled in the phylogenetic trees.

Phylogenetic tree construction

All qualified mutations (including silent and non-silent mutations) of all multifocal and multiregion samples from patients (P114, P116, P117, P118, and P119) were included in the phylogenetic tree construction. Sequences encompassing the mutation sites (total length of 21 bp) were extracted to infer the phylogeny among the samples of each patient. Trees were constructed using the maximum parsimony algorithm in MEGA X with default parameters [16]. All phylogenetic trees were further optimized in Adobe Illustrator. Potential driver mutations were labeled on the trees according to their regional distributions.

Comparisons of the mutation frequencies of oncogenic genes and oncogenic pathways

To unravel the somatic mutation events involved in maintaining the indolent clinical phenotype of SSNs (rather than progressing to malignant late stage LUAD), we compared the mutation frequencies of the selected genes between SSNs and advanced stage LUADs (The Cancer Genome Atlas [TCGA] 106 stage II-IV LUADs [13]; Chinese 75 stage II-IV LUADs [14]). Genes were selected from: 1) those documented in the 10 oncogenic signaling pathways in TCGA [15] and 2) genes in LUAD-related pathways reported in other LUAD cohorts [13, 14].

We defined the samples that had one more mutation on the genes listed above of each pathway are altered in this specific pathway. Altered frequency of each pathway of three cohorts was calculated by the number of samples with the altered pathway divided by the total number of the samples in each cohort.

Statistical analyses

All statistical analyses were performed in the R statistical environment (version 3.4.4).

References

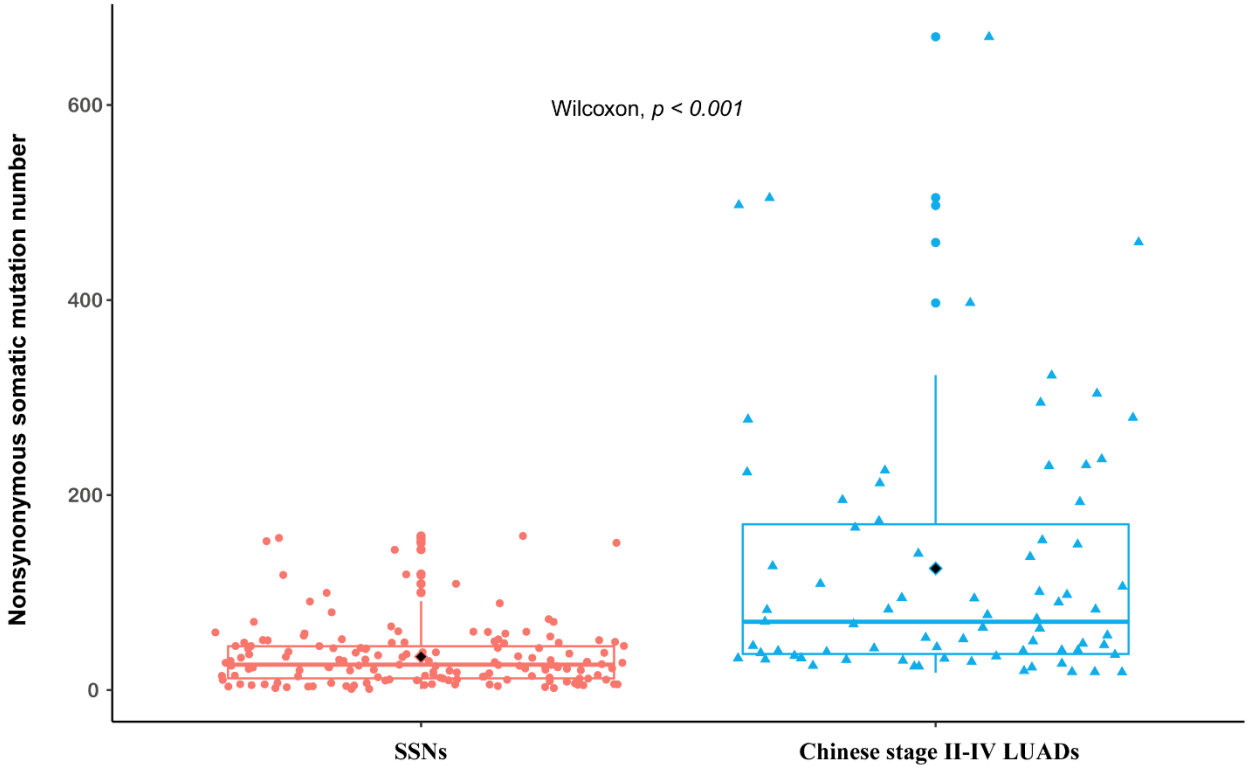
1. Travis, W.D., et al., *The IASLC Lung Cancer Staging Project: Proposals for Coding T Categories for Subsolid*

Nodules and Assessment of Tumor Size in Part-Solid Tumors in the Forthcoming Eighth Edition of the TNM Classification of Lung Cancer. J Thorac Oncol, 2016. **11**(8): p. 1204-1223.

2. Suzuki, K., et al., *A prospective radiological study of thin-section computed tomography to predict pathological noninvasiveness in peripheral clinical IA lung cancer (Japan Clinical Oncology Group 0201).* J Thorac Oncol, 2011. **6**(4): p. 751-6.
3. de Hoop, B., et al., *Pulmonary ground-glass nodules: increase in mass as an early indicator of growth.* Radiology, 2010. **255**(1): p. 199-206.
4. Martin, M., *Cutadapt removes adapter sequences from high-throughput sequencing reads.* Embnet Journal, 2011. **17(1)**.
5. Li, H. and R. Durbin, *Fast and accurate long-read alignment with Burrows-Wheeler transform.* Bioinformatics, 2010. **26**(5): p. 589-95.
6. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.
7. McKenna, A., et al., *The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data.* Genome Res, 2010. **20**(9): p. 1297-303.
8. Cingolani, P., et al., *A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3.* Fly (Austin), 2012. **6**(2): p. 80-92.
9. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-218.
10. Gaujoux, R. and C. Seoighe, *A flexible R package for nonnegative matrix factorization.* BMC Bioinformatics, 2010. **11**: p. 367.
11. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
12. Campbell, J.D., et al., *Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas.* Nat Genet, 2016. **48**(6): p. 607-16.
13. Cancer Genome Atlas Research, N., *Comprehensive molecular profiling of lung adenocarcinoma.* Nature, 2014. **511**(7511): p. 543-50.
14. Wu, K., et al., *Frequent alterations in cytoskeleton remodelling genes in primary and metastatic lung adenocarcinomas.* Nat Commun, 2015. **6**: p. 10131.
15. Sanchez-Vega, F., et al., *Oncogenic Signaling Pathways in The Cancer Genome Atlas.* Cell, 2018. **173**(2): p. 321-+.
16. Kumar, S., et al., *MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms.* Mol Biol Evol, 2018. **35**(6): p. 1547-1549.

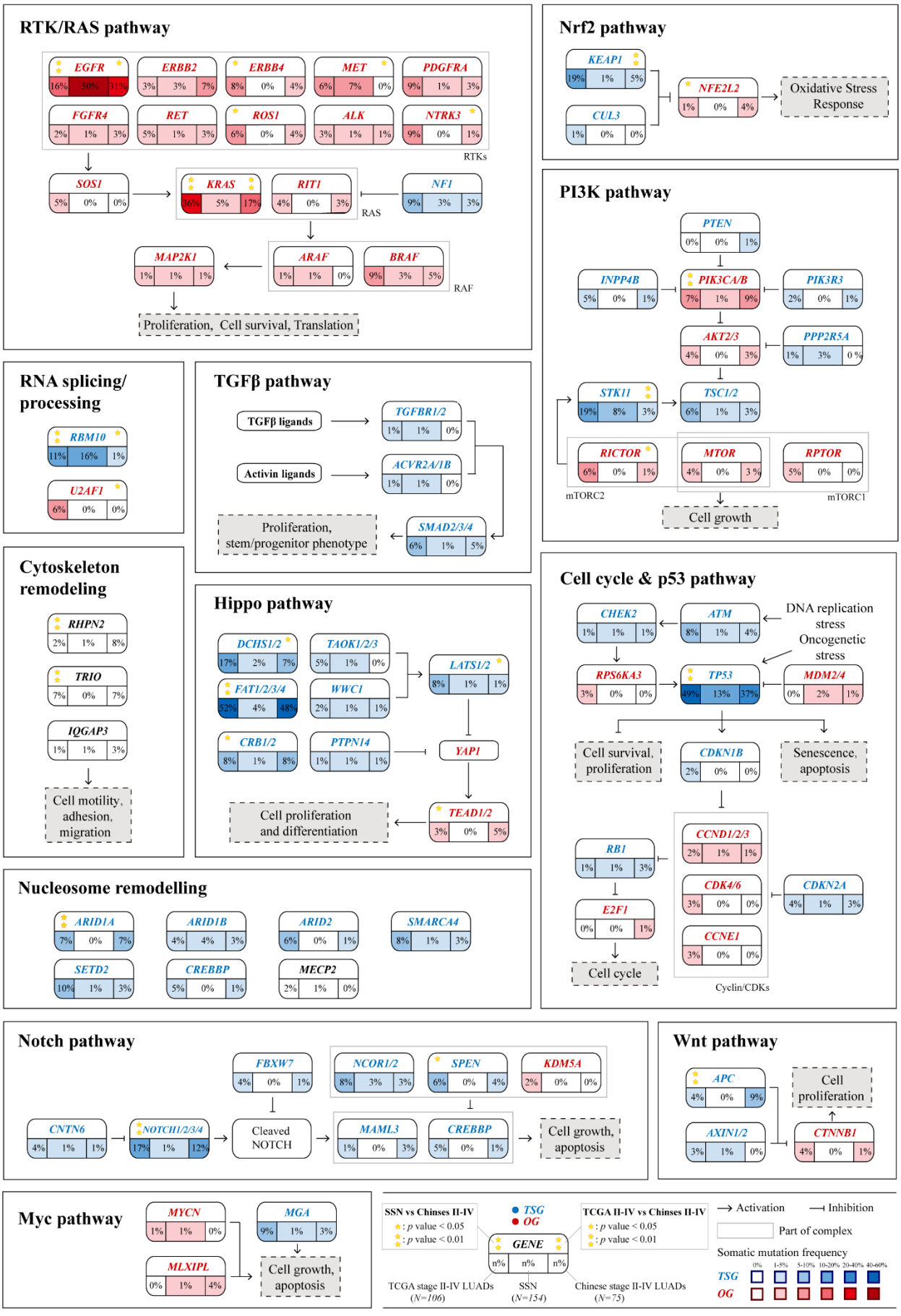
SUPPLEMENTARY FIGURES

Supplementary Figure S1



Supplementary Figure S1. Comparison of the number of nonsynonymous somatic mutations between Chinese SSNs (*N*=154) and Chinese advanced-stage LUADs (*N*=75).

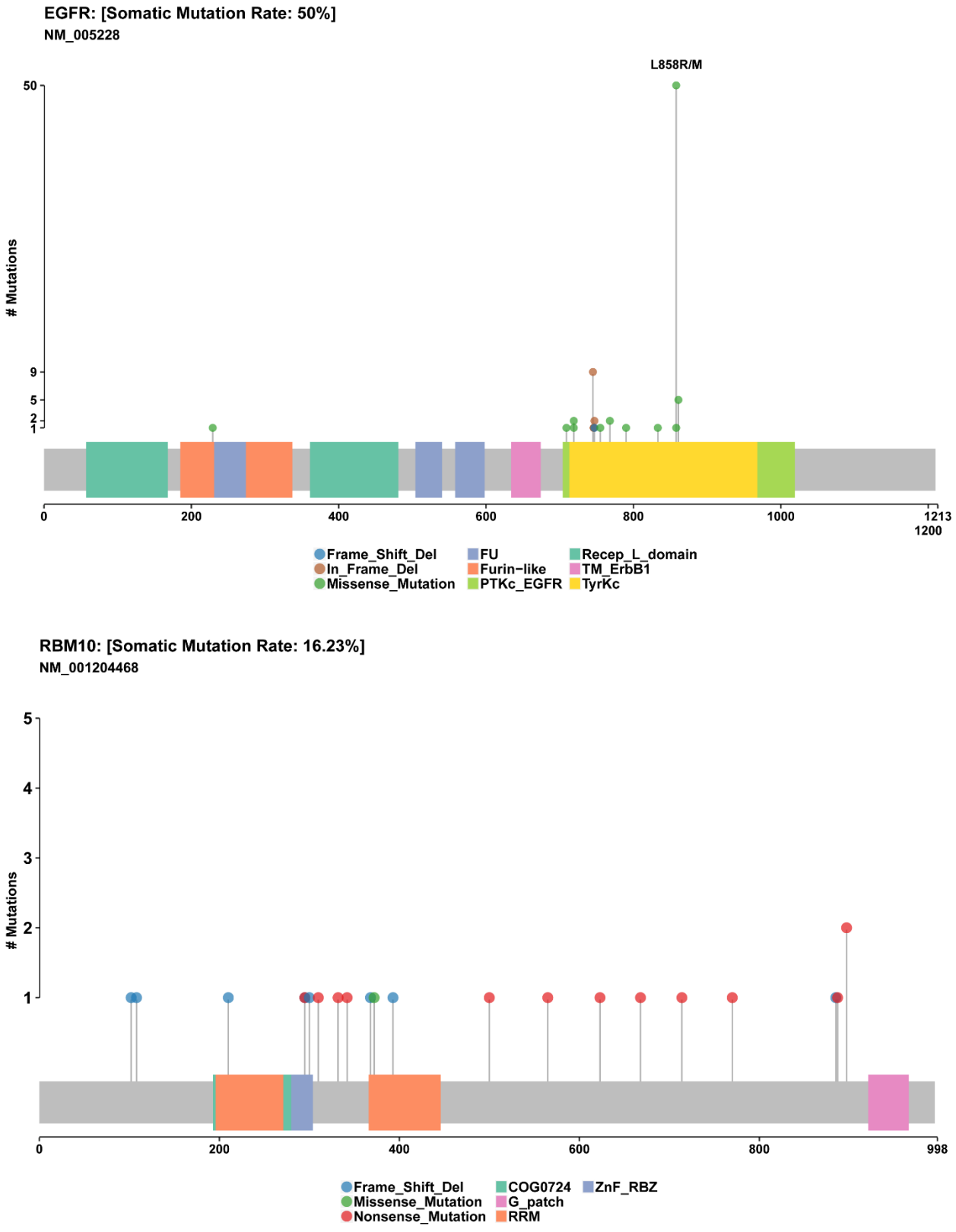
Supplementary Figure S2


RTK/RAS pathway
EGFR 16% 50% 31%
ERBB2 3% 3% 7%
ERBB4 8% 0% 4%
MET 6% 7% 0%
PDGFRA 9% 1% 3%
FGFR4 2% 1% 3%
RET 5% 1% 3%
ROS1 6% 0% 4%
ALK 3% 1% 1%
NTRK3 9% 0% 1%
RTKs
SOS1 5% 0% 0%
KRAS 36% 5% 17%
RIT1 4% 0% 3%
NF1 9% 3% 3%
RAS
MAP2K1 1% 1% 1%
ARAF 1% 1% 0%
BRAF 9% 3% 5%
RAF
Proliferation, Cell survival, Translation
Nrf2 pathway
KEAP1 19% 1% 5%
CUL3 1% 0% 0%
NFE2L2 1% 0% 4%
Oxidative Stress Response
PI3K pathway
PTEN 0% 0% 1%
INPP4B 5% 0% 1%
PIK3CA/B 7% 1% 9%
PIK3R3 2% 0% 1%
AKT2/3 4% 0% 3%
PPP2R5A 1% 3% 0 %
STK11 19% 8% 3%
TSC1/2 6% 1% 3%
RICTOR 6% 0% 1%
MTOR 4% 0% 3 %
RPTOR 5% 0% 0%
mTORC2
mTORC1
Cell growth
RNA splicing/ processing
RBM10 11% 16% 1%
U2AF1 6% 0% 0%
TGFβ pathway
TGFβ ligands
Activin ligands
TGFBR1/2 1% 1% 0%
ACVR2A/1B 1% 1% 0%
SMAD2/3/4 6% 1% 5%
Proliferation, stem/progenitor phenotype
Cytoskeleton remodeling
RHPN2 2% 1% 8%
TRIO 7% 0% 7%
IQGAP3 1% 1% 3%
Cell motility, adhesion, migration
Hippo pathway
DCHS1/2 17% 2% 7%
FAT1/2/3/4 52% 4% 48%
CRB1/2 8% 1% 8%
TAOK1/2/3 5% 1% 0%
WWC1 2% 1% 1%
PTPN14 1% 1% 1%
LATS1/2 8% 1% 1%
YAP1
TEAD1/2 3% 0% 5%
Cell proliferation and differentiation
Cell cycle & p53 pathway
CHEK2 1% 1% 1%
ATM 8% 1% 4%
DNA replication stress
Oncogenetic stress
RPS6KA3 3% 0% 0%
TP53 49% 13% 37%
MDM2/4 0% 2% 1%
Cell survival, proliferation
CDKN1B 2% 0% 0%
Senescence, apoptosis
RB1 1% 1% 3%
E2F1 0% 0% 1%
Cell cycle
CCND1/2/3 2% 1% 1%
CDK4/6 3% 0% 0%
CCNE1 3% 0% 0%
CDKN2A 4% 1% 3%
Cyclin/CDKs
Nucleosome remodelling
ARID1A 7% 0% 7%
ARID1B 4% 4% 3%
ARID2 6% 0% 1%
SMARCA4 8% 1% 3%
SETD2 10% 1% 3%
CREBBP 5% 0% 1%
MECP2 2% 1% 0%
Notch pathway
FBXW7 4% 0% 1%
NCOR1/2 8% 3% 3%
SPEN 6% 0% 4%
KDM5A 2% 0% 0%
CNTN6 4% 1% 1%
NOTCH1/2/3/4 17% 1% 12%
Cleaved NOTCH
MAML3 1% 0% 3%
CREBBP 5% 0% 1%
Cell growth, apoptosis
Wnt pathway
APC 4% 0% 9%
AXIN1/2 3% 1% 0%
Cell proliferation
CTNNB1 4% 0% 1%
Myc pathway
MYCN 1% 1% 0%
MLXIPL 0% 1% 4%
MGA 9% 1% 3%
Cell growth, apoptosis
SSN vs Chinses II-IV
★ : p value < 0.05
★ : p value < 0.01
TSG
OG
GENE
n% n% n%
TCGA II-IV vs Chinses II-IV
★ : p value < 0.05
★ : p value < 0.01
TCGA stage II-IV LUADs (N=106)
SSN (N=154)
Chinese stage II-IV LUADs (N=75)
Activation
Inhibition
Part of complex
Somatic mutation frequency
0% 1-5% 5-10% 10-20% 20-40% 40-60%
TSG
OG

Supplementary Figure S2. Comparison of the frequencies of nonsynonymous mutations in key genes in cancer-associated pathways between Chinese SSNs (*N*=154) and advanced-stage LUADs (TCGA stage II-IV LUADs, *N*=154; Chinese stage II-IV LUADs, *N*=75).

Each gene box includes three percentages representing the nonsynonymous mutation frequencies of activation and inactivation of three cohorts. The data layout is described in the legend. Genes are grouped by their signaling pathways. Interactions between genes are indicated.

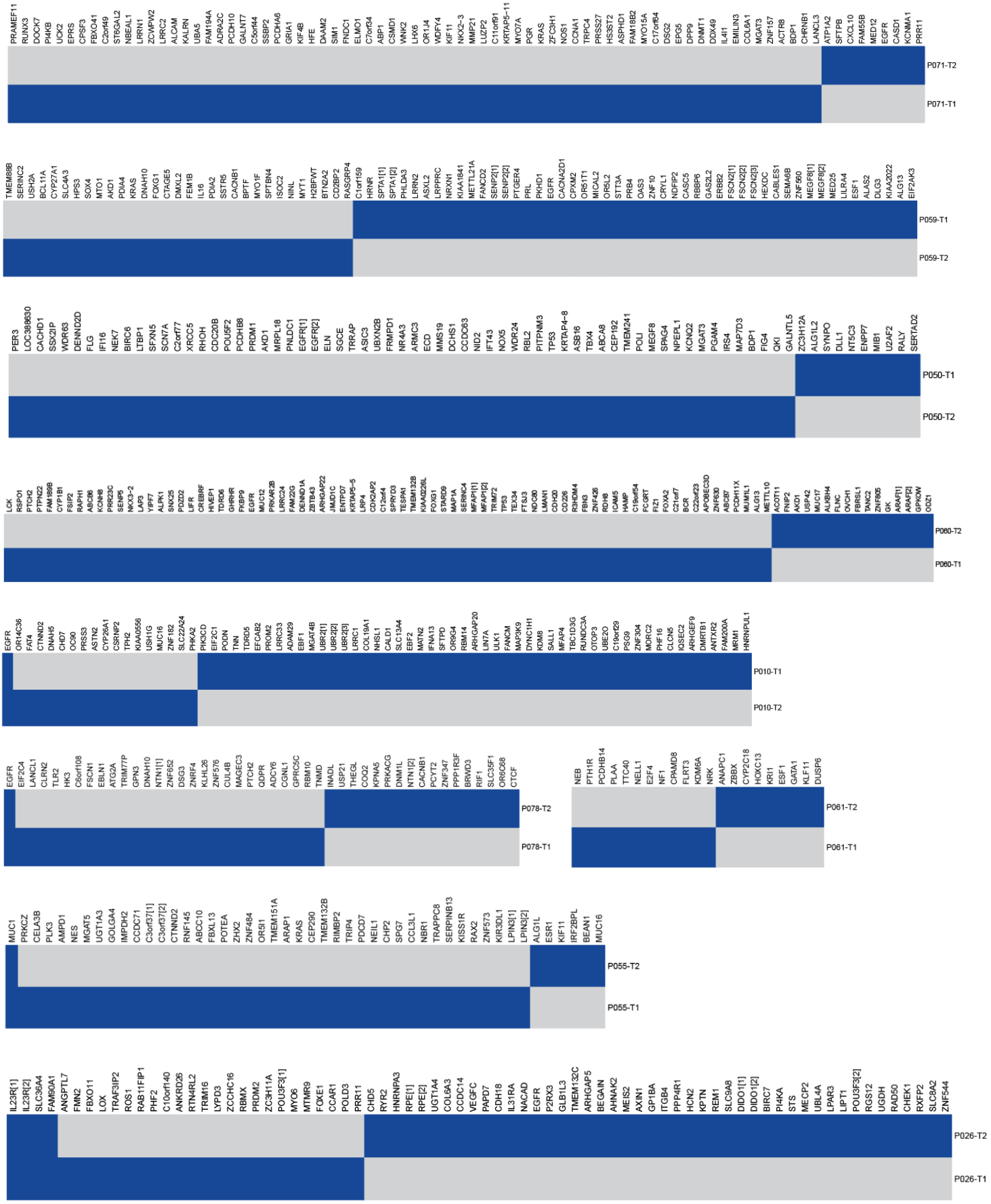
Supplementary Figure S3

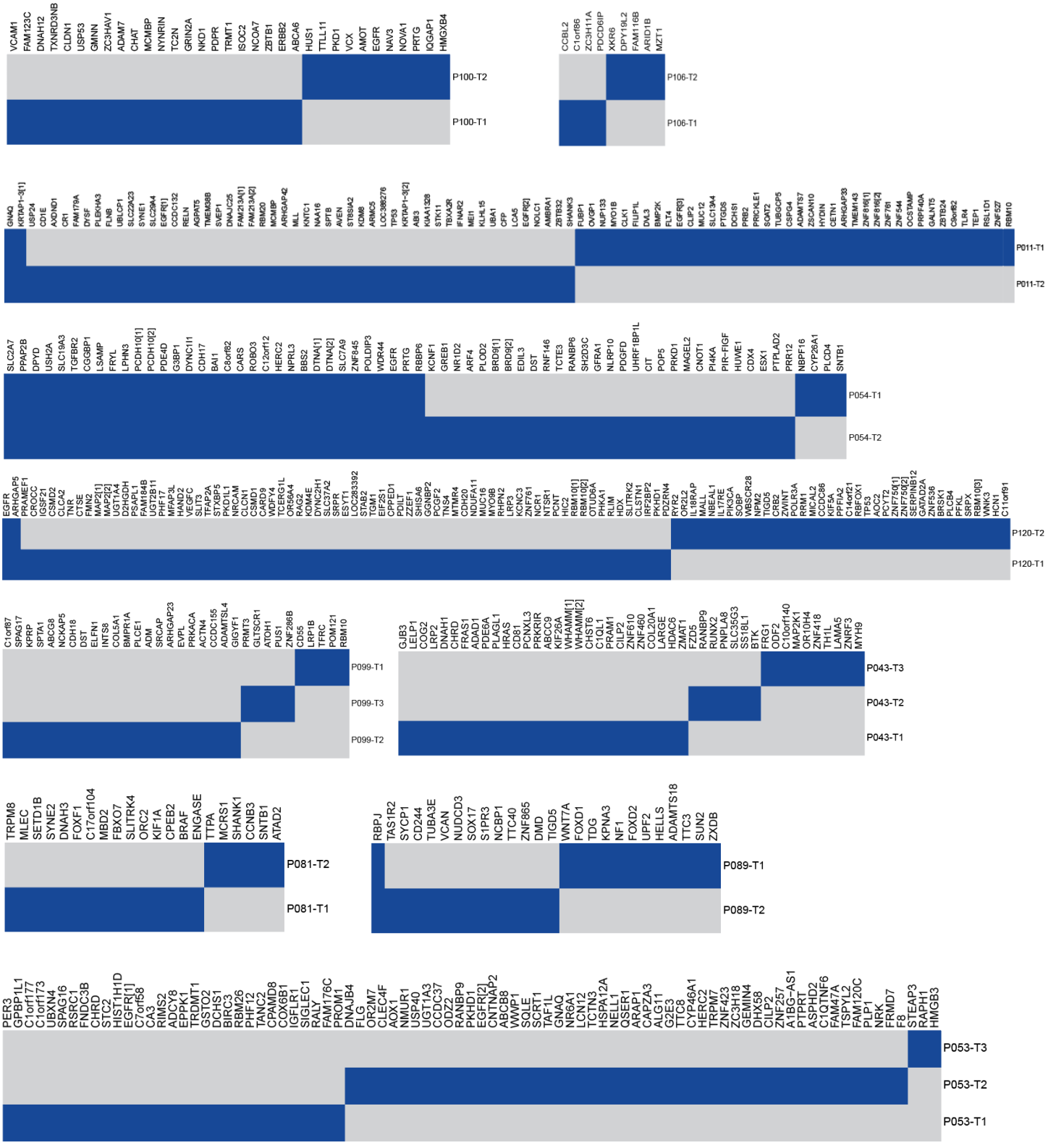


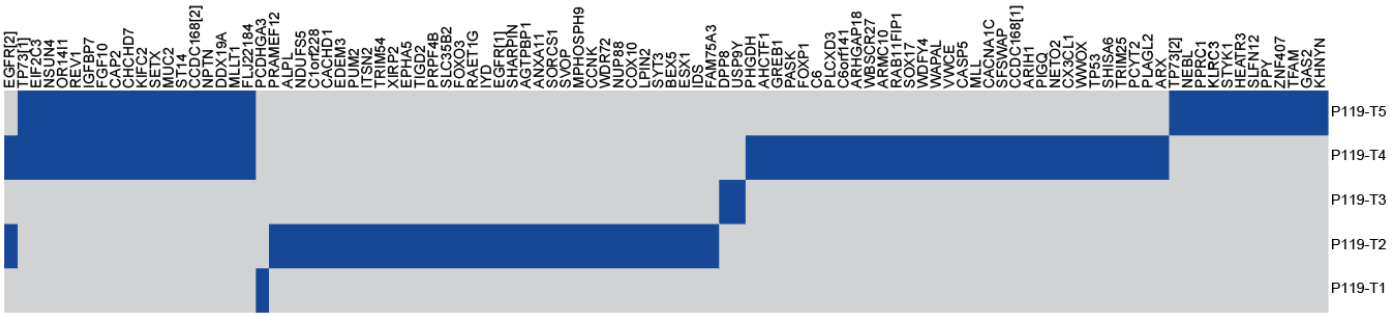
Supplementary Figure S3. Lollipop plots for amino acid changes resulting from *EGFR* and *RBM10* mutations.

Locations of *EGFR* and *RBM10* mutations on the protein structure. Mutation types and protein domains are indicated below. The vertical axis represents the number of occurrences of each mutation site.

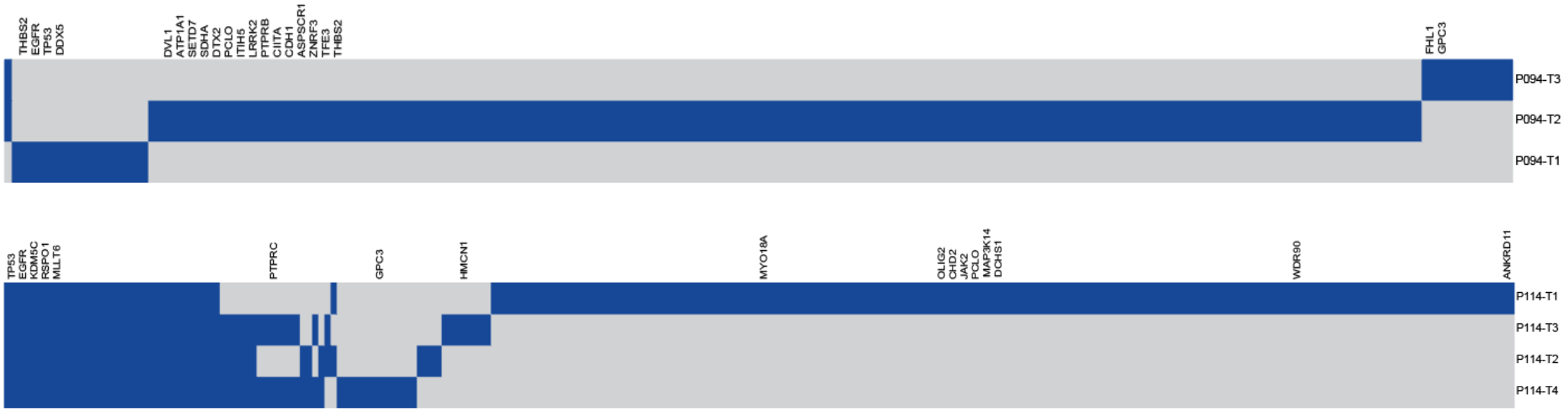
Supplementary Figure S4




VCAM1 FAM123C DNAH12 TXNRD3NB CLDN1 USP53 GMNN ZC3HAV1 ADAM7 CHAT MCMBP NYNRIN TC2N GRIN2A NKD1 PDPR TRMT1 ISOC2 NCOA7 ZBTB1 ERBB2 ABCA6 HUS1 TTLL11 PKD1 VCX AMOT EGFR NAV3 NOVA1 PRTG IQGAP1 HMGXB4
P100-T2
P100-T1
CCBL2 C1orf86 ZC3H11A PDCD6IP XKR6 DPY19L2 FAM116B ARID1B MZT1
P106-T2
P106-T1
GNAQ KRTAP1-3[1] USP24 CD1E AXDND1 CR1 FAM179A DYSF PLEKHA3 FLNB UBLCP1 SLC22A23 SYNE1 SLC2A4 EGFR[1] CCDC132 RELN AGPAT5 TMEM38B SVEP1 DNAJC25 FAM213A[1] FAM213A[2] RBM20 MCMBP ARHGAP42 MLL KNTC1 NAA16 SPTB AVEN ST8SIA2 KDM8 ARMC5 LOC388276 TP53 KRTAP1-3[2] ABI3 KIAA1328 STK11 TBX3R IFNAR2 MEI1 KLHL15 UBA1 CFP LCA5 EGFR[2] NOLC1 AMBRA1 ZBTB32 SHANK3 FUBP1 OVGP1 NUP133 MYO1B CLK1 FIUP1L DVL3 BMP2K FLT4 EGFR[3] CLIP2 MUC12 SLC13A4 PTGDS DCHS1 PRB2 PRICKLE1 SOAT2 TUBGCP5 CSPG4 ADAMTS7 ZSCAN10 HYDIN CETN1 ARHGAP33 TMEM143 ZNF816[1] ZNF816[2] ZNF761 ZNF544 OCSTAMP PRPF40A GALNT5 ZBTB24 C9orf62 TLR4 TEP1 RSL1D1 ZNF527 RBM10
P011-T1
P011-T2
SLC2A7 PPAP2B DPYD USH2A SLC19A3 TGFBR2 CGGBP1 LSAMP FRYL LPHN3 PCDH10[1] PCDH10[2] PDE4D G3BP1 DYNC1I1 CDH17 BA11 C8orf62 CARS ROBO3 C12orf12 HERC2 NPRL3 BBS2 DTNA[1] DTNA[2] SLC7A9 ZNF845 POLDIP3 WDR44 EGFR PRTG RBBP6 KCNF1 GREB1 NR1D2 ARF4 PLOD2 BRD9[1] BRD9[2] EDIL3 DST RNF146 TCTE3 RANBP6 SH2D3C GFRA1 NLRP10 PDGFD UHRF1BP1L CIT POP5 PRKD1 MAGEL2 CNOT1 PI4KA PIR-FIGF HUWE1 CDX4 ESX1 PTPLAD2 PRR12 NBPF16 CYP26A1 PLCD4 SNTB1
P054-T1
P054-T2
EGFR ARHGAP5 PRAMEF1 CROCC IGSF21 CSMD2 CA2 TNR CTSE FMN2 MAP2[1] MAP2[2] UGT1A4 D2HGDH PSAPL1 FAM184B UGT8 PHF17 MFAP3L HAND2 VEGFC SLIT3 TFAP2A STXBP5 PKDL1 NRCAM CDH1 CSDC1 CARD9 WDFY4 TCERG1L OR56A4 RAG2 KDM4E DYNC2H1 SLC37A2 SRPR SYT1 LOC283392 STAB2 TGM1 EIF2S1 CPPED1 PDILT ZZEF1 SHISA6 GGNBP2 PCGF2 TNIK MTMR4 CDH20 NDUFA11 MUC16 MYO9B RHPN2 LRP3 KCNC3 ZNF761 ZNF521 NTSR1 PCNT HIC2 RBM10[1] RBM10[2] OTUD6A PHKA1 RLIM HDX SLITRK2 CLSTN1 IRF2BP2 PKHD1 PDZRN4 RYR2 OR2L2 IL16RAP MALL NBEAL1 IL17RE HIVEP3 SOBP WBSCR28 NPM2 TIGD5 CRB2 ZWINT POLR3A RRM1 MICAL2 CEP68 KIF26B PPFIA2 C14orf21 RBFOX1 TP53 AOC2 PCYT2 ZNF750[1] ZNF750[2] SERINC1 GRIK1-AS2 ZNF536 BRSK1 PLCB4 PFKL SRPX RBM10[3] WNK3 HCN1 C11orf91
P120-T2
P120-T1
C1orf87 SPAG17 KPRP SFTA1 ABCG8 NCKAP5 CDH18 DST ELFN1 INTS8 COL5A1 BMPR1A PLCE1 ADM SRCAP ARHGAP23 EVPL PRKACA ACTN4 CCDC155 ADAMTSL4 GIGYF1 PRMT3 GLTSCR1 ATOH1 PUS1 ZNF286B CD55 LRP1B TFRC POM121 RBM10
P099-T1
P099-T3
P099-T2
GJB3 LELP1 COG2 LRP2 DNAH1 CHRD FRAS1 ADAD1 PDE6A PLAGL1 HRAS CD81 PCNXL3 PRKRIR ABCC9 KIF26A WHAMM[1] WHAMM[2] CHST6 C1QL1 PRAM1 CILP2 ZNF610 ZNF460 COL20A1 LARGE HDAC6 ZMAT1 FZD5 RANBP9 RUNX2 PNPLA8 SLC35G3 SS18L1 BTK FRG1 ODF2 C10orf140 MAP2K1 OR10H4 ZNF418 TH1L LAMA5 ZNRF3 MYH9
P043-T3
P043-T2
P043-T1
TRPM8 MLEC SETD1B SYNE2 DNAH3 FOXF1 C17orf104 MBD2 FBXO7 SLITRK4 ORC2 KIF1A CPEB2 BRAF ENGASE TTPA MCRS1 SHANK1 CCNB3 SNTB1 ATAD2
P081-T2
P081-T1
RBPJ TAS1R2 SYCP1 CD244 TUBA3E VCAN NUDCD3 SOX17 S1PR3 NCBP1 TTC40 ZNF865 DMD TIGD5 WNT7A FOXD1 TDG KPNA3 NF1 FOXD2 UPF2 HELLS ADAMTS18 TTC3 SUN2 ZXDB
P089-T1
P089-T2
PER3 GPBP1L1 C1orf177 C1orf173 UBXN4 SPAG16 RSRC1 FNDC3B CHRD STC2 HIST1H1D EGFR[1] C7orf58 CA3 RIMS2 ADCY8 EPPK1 TRDMT1 GSTO2 DCHS1 BIRC3 RBM26 PHF12 TANC2 CPAMD8 COX6B1 IGFLR1 SIGLEC1 RALY FAM176C PROM1 DNAJB4 FLG OR2M7 CLEC4F AOX1 NMUR1 USP40 UGT1A3 CCDC37 ODZ2 RANBP9 PKHD1 EGFR[2] CNTNAP2 ABCB8 WWP1 SQLE SCRT1 TAF1L GNAQ NR6A1 LCN12 TCTN3 HSPA12A NELL1 QSER1 ARAP1 CAPZA3 ALG11 G2E3 TTC8 CYP46A1 HERC2 TRPM7 ZNF423 ZC3H18 GEMIN4 DHX58 CILP2 ZNF257 A1BG-AS1 PTPRT ASPHD2 C1QTNF6 FAM47A TSPYL2 FAM120C PLP1 NRK FRMD7 F8 STEAP3 RAPH1 HMGB3
P053-T3
P053-T2
P053-T1


EGFR[2]
TP73[1]
EIF2C3
NSUN4
OR14I1
REV1
IGFBP7
FGF10
CAP2
CHCHD7
KIFC2
SETX
MUC2
ST14
CCDC168[2]
NPTN
DDX19A
MLLT1
FLJ22184
PCDHGA3
PRAMEF12
ALPL
NDUFS5
C1orf228
CACHD1
EDEM3
PUM2
ITSN2
TRIM64
XIRP2
EPHA5
TGIF2
RP4-6B
SC35B2
FOXO3
RAET1G
IYD
EGFR[1]
SHARPIN
AGTPBP1
ANXA11
SORCS1
SVOP
MPHOSPH9
CCNK
WDR72
NUP88
COX10
LPIN2
SYT3
BEX5
IDS
FAM75A3
DPP8
USP9Y
PHGDH
AHCTF1
GREB1
PASK
FOXP1
C6
PLCXD3
C6orf141
ARHGAP18
WBSCR27
ARMC10
RAB11FIP1
SOX17
WAPAL
VWCE
CASP5
MLL
CACNA1C
SFSWAP
CCDC168[1]
ARIH1
PIGQ
NETO2
CX3CL1
WWOX
TP53
SHISA6
TRIM25
PCYT2
ARAP3
TP73[2]
NEB
PPRC1
KLRC3
STYK1
HEATR3
SLFN12
PPY
ZNF407
TFAM
GAS2
KHNYN
P119-T5
P119-T4
P119-T3
P119-T2
P119-T1

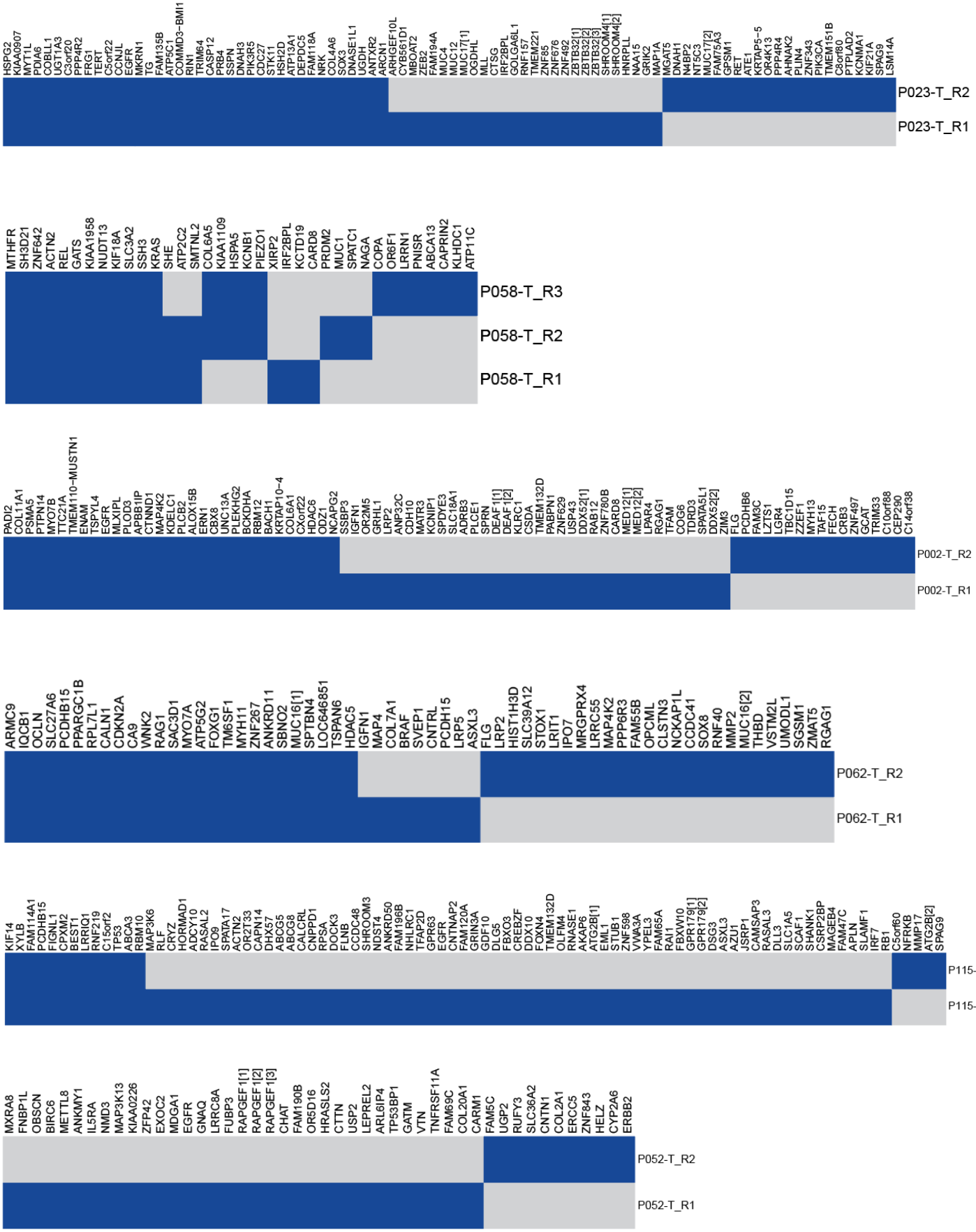
TTK
BCR
PO13
CACHD1
NTRK1
MARC2
MAP1LC3C
SFTPB
LYPD6
PABPC4L
PCDHB4
PCDHGA3
ADRB2
CCZ1B
DPP6
TMEM215
ASTN2
AGAP6
PRF1
C10orf57
PSMA1
CAPRIN1
DIAPH3
TRDN
NRG1
MAN2C1
PAPD5
PRDX1
CTSK[1]
CTSK[2]
CALM2
MAT2A
TBC1D23
HNRPDL
HNRNPA2B1[1]
HNRNPA2B1[2]
ACTA2[1]
ACTA2[2]
EIF4G2[1]
EIF4G2[2]
ITM2B[1]
TDP1
SERF2−C15ORF63
SERF2
COL1A1
DDX5
S100A7L2
SMEK2
DNAH6
ITM2B[2]
LOC100131094
UHRF2
UBAP1
ERP44
DGKZ
LOC100505679
PREX1
MED14
ZBED6
C10orf140
RORC
HMCN1
LRP1B
PPIG
ABCA12
UGT1A3
TGM4
MME
CWH43
RNF44
GPRIN1
EYS
TYW1
TRPM6
FAM149B1
PAMR1
PCDH17
AKAP13
ADCY9
MUC16
MED12
GIGYF2
SLC9B1
C4orf21
ZNF292
SNTB1
CIT
HMGXB4
P077-T3
P077-T5
P077-T4
P077-T2
P077-T1
BACH2
BCR
ERBB2
EWSR1
FAT4
FBXO11
GOPC
KEAP1
LIFR
PDE4DIP
PGR
PSIP1
RELN
STK11
TLE4
BCL11A
KRAS
LRP1B
LYN
PDGFRA
SOX17
ZNF384
P05-T2
P05-T1
CHD8
EP300
TET3
EPHA3
EGFR[1]
ACSL3
BAP1
CTNNB1
EGFR[2]
ELN
MDM2
EXT1
NT5C2
RBM10
ARID1B
KAT6B
NCOR1
ATM
EGFR[3]
P073-T3
P073-T1
P073-T2
EGFR
CDC42
FLT4
MET
TP53
EPAS1
FZD8
MRE11A
PIGA
PPP1CB
PTPN14
PTPRK
TGFBR1
RBM10[1]
BCL11A
RBM10[2]
ROBO1
P118-T1
P118-T3
P118-T2

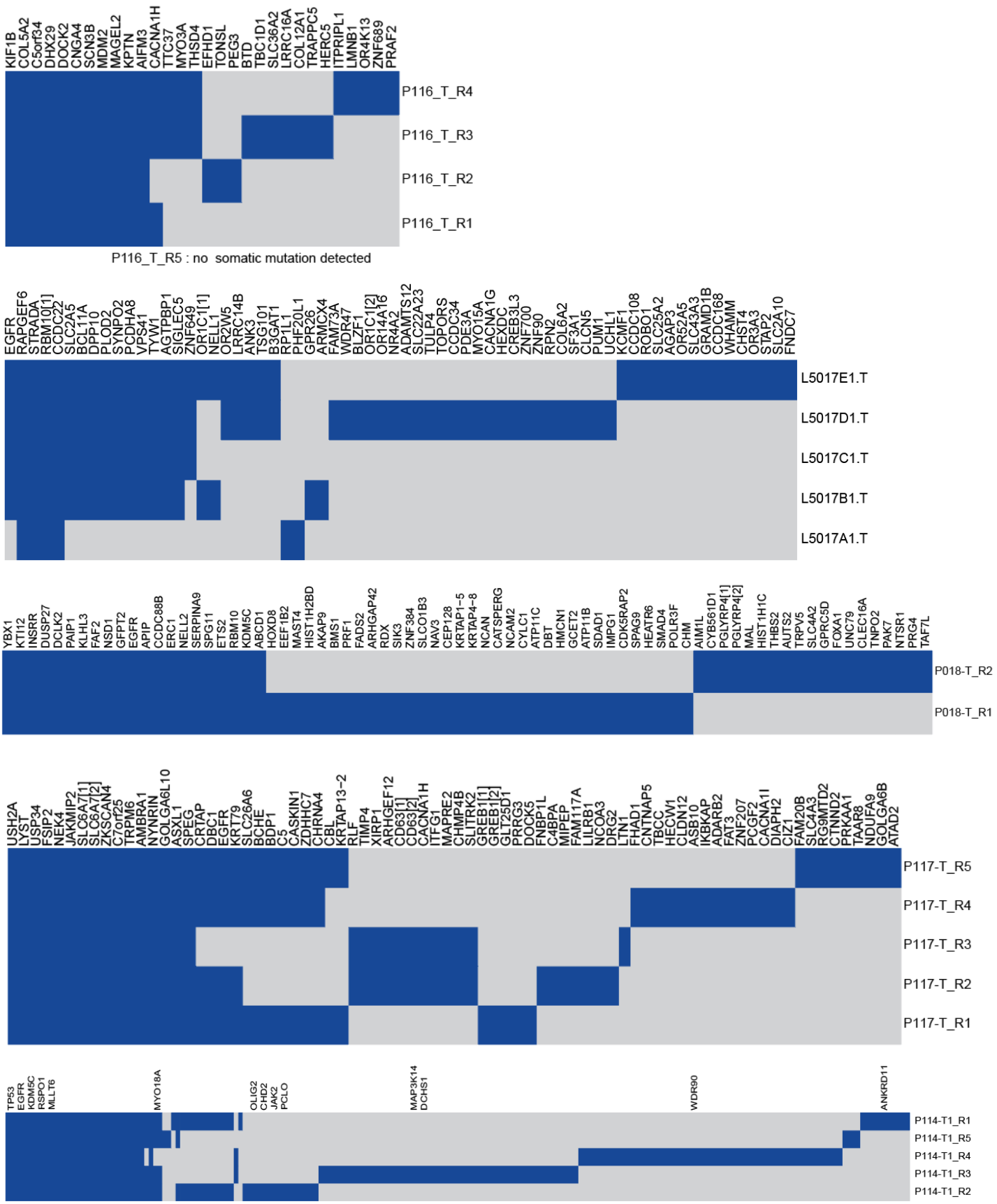


Supplementary Figure S4. Regional distribution of all somatic mutations in multifocal tumors (including SSNs and solid tumors) in the same patient.

The regional distribution of all of the somatic mutations of 66 tumors from 26 patients is depicted in the heatmaps. Each column represents a single mutation site. Grey represents wild type, while blue represents mutation of a certain site in a certain gene.

Supplementary Figure S5





Supplementary Figure S5. Regional distribution of all somatic mutations in multiregions of the same SSN.

The regional distribution of all of the somatic mutations of 35 spatially separate regions sampled from 11 large SSNs is depicted in the heatmaps. Each column represents a single mutation site. Grey represents wild type, while blue represents mutation of a certain site in a certain gene.

SUPPLEMENTARY TABLES

Supplementary Table S1

Supplementary Table S1. Detailed clinical information for 120 patients and 154 SSNs.

Patient No	Tumor No	Tumor_Sample_Barcode	Pathology	Stage	Gender	Pure/Mix	Pack-year	Total size	Solid size	Total volume	Solid_volume
P001	P001-T	PBW01_F9T	LUAD	IA1	male	M	0	16.1	8.1	1830	617
P002	P002-T	PBW02_GG012T	LUAD	IA1	female	P	0	36.0	0.0	10657	231
P003	P003-T	PBW03_F20T	LUAD	IA2	female	M	0	15.4	14.1	3879	2010
P004	P004-T	PBW04_F21T	LUAD	IA1	female	M	0	22.8	10.0	2794	1632
P005	P005-T1	PBW05_F23_1T	LUAD	IA1	male	M	80	19.2	8.4	3943	722
P005	P005-T2	PBW05_F23_2T	AAH	N/A	male	P	80	12.3	0.0	2010	0
P006	P006-T	PBW06_F24T	LUAD	IA1	male	P	0	14.0	0.0	N/A	N/A
P007	P007-T	PBW07_F34T	LUAD	IA2	female	M	0	17.8	14.0	3394	1092
P008	P008-T	PBW08_F38T	MIA	IA1	female	P	0	13.7	0.0	904	0
P009	P009-T	PBW09_GG037T	LUAD	IA2	male	N/A	0	17.0	N/A	N/A	N/A
P010	P010-T2	PBW10_L3720shang	LUAD	IA2	female	M	0	28.7	18.7	14008	1688
P011	P011-T2	PBW11_GG061_2T	MIA	IA1	male	P	20	14.7	0.0	466	0
P011	P011-T1	PBW11_GG061T	LUAD	IA2	male	M	20	14.3	11.7	1167	575
P012	P012-T	PBW12_GG067T	LUAD	IA2	female	M	0	28.6	15.5	3765	1250
P013	P013-T	PBW13_GG069T	LUAD	IA2	female	M	0	20.0	17.0	N/A	N/A
P014	P014-T	PBW14_GG073T	LUAD	IA1	male	M	45	18.0	7.0	N/A	N/A
P015	P015-T	PBW15_F17T	MIA	IA1	male	P	0	12.0	0.0	N/A	N/A
P016	P016-T	PBW16_F18T	LUAD	IA1	female	M	0	15.4	6.3	756	324
P017	P017-T	PBW17_G27T	LUAD	IA2	female	M	0	15.0	11.0	1383	374
P018	P018-T	PBW18_F41T	LUAD	IB	male	M	0	32.8	19.1	10135	6952
P019	P019-T	PBW19_GG074T	LUAD	IA1	male	P	0	23.5	0.0	2833	0
P020	P020-T	PBW20_GG075T	LUAD	IA2	female	M	0	18.2	18.2	2457	1495
P021	P021-T	PBW21_GG078T	LUAD	IB	female	M	0	40.5	28.3	22913	11653

P022	P022-T	PBW23_GG080T	LUAD	IA2	female	M	0	28.6	12.6	5632	1835
P023	P023-T	PBW24_GG081T	LUAD	IA2	female	M	0	34.0	12.0	N/A	N/A
P024	P024-T	PBW25_GG083T	MIA	IA1	female	M	0	10.0	4.0	N/A	N/A
P025	P025-T	PBW26_GG085T	LUAD	IA1	female	P	0	13.7	0.0	1170	34.6
P026	P026-T1	PBW27_GG089_1T	LUAD	IA2	female	M	0	22.4	11.5	4828	2069
P026	P026-T2	PBW27_GG089_2T	MIA	IA1	female	P	0	9.1	0.0	63	0
P027	P027-T	PBW28_GG090T	MIA	IA1	male	P	0	21.0	0.0	N/A	N/A
P028	P028-T	PBW29_GG096T	MIA	IA1	female	P	0	10.0	0.0	466	0
P029	P029-T	PBW30_GG0104T	LUAD	IB	female	M	0	49.5	30.8	3071	368
P030	P030-T	PBW31_GG0106T	LUAD	IA1	female	P	0	16.0	0.0	982	0
P031	P031-T	PBW33_GG073T	LUAD	IA2	female	M	0	17.3	17.3	1780	551
P032	P032-T	PBW35_GG099T	LUAD	IA1	male	M	50	28.0	9.0	3275	1374
P033	P033-T	PBW36_GG0105T	LUAD	IA1	male	M	20	22.7	9.4	2064	160
P034	P034-T	PBW37_GG0113T	MIA	IA1	male	P	0	10.8	0.0	297	0
P035	P035-T	PBW38_GG0114T	AIS	0	female	P	0	8.6	0.0	227	0
P036	P036-T	PBW39_GG0120T	LUAD	IA1	female	M	0	9.0	4.0	N/A	N/A
P037	P037-T	PBW40_GG0121T	LUAD	IA1	male	P	3	15.4	0.0	762	0
P038	P038-T	PBW41_GG0124T	MIA	IA1	female	P	0	10.6	0.0	292	0
P039	P039-T	PBW42_GG0125T	LUAD	IA2	female	M	0	15.1	10.7	947	557
P040	P040-T	PBW43_GG0126T	LUAD	IA2	male	M	40	25.0	12.4	3010	1282
P041	P041-T	PBW44_GG0127T	LUAD	IA1	male	M	10	17.4	5.8	781	134
P042	P042-T	PBW45_GG0128T	MIA	IA1	female	M	0	9.7	3.4	143	3
P043	P043-T1	PBW46_GG0129_1T	LUAD	IB	female	M	0	32.2	32.2	8466	5342
P043	P043-T2	PBW46_GG0129_2T	LUAD	IA1	female	M	0	15.8	7.2	714	273
P043	P043-T3	PBW46_GG0129_3T	AIS	0	female	P	0	12.4	0.0	1017	0
P044	P044-T	PBW47_GG0131T	LUAD	IA2	male	M	20	30.9	10.5	6300	1292
P045	P045-T	PBW48_GG0134T	LUAD	IA2	male	M	5	25.1	15.6	5943	3280
P046	P046-T	PBW49_GG0135T	MIA	IA1	female	P	0	8.0	0.0	N/A	N/A
P047	P047-T	PBW50_GG0137T	MIA	IA1	male	P	20	16.2	0.0	952	0

P048	P048-T	PBW51_GG0138T	LUAD	IA2	male	M	0	22.4	10.3	2768	502
P049	P049-T	PBW52_GG0140T	LUAD	IA1	male	M	0	19.0	9.0	N/A	N/A
P050	P050-T1	PBW53_GG0141-1T	LUAD	IA2	female	M	0	16.0	12.1	2261	1026
P050	P050-T2	PBW53_GG0141-2T	AIS	0	female	P	0	6.3	0.0	122	0
P051	P051-T	PBW54_GG0142T	MIA	IA1	male	M	0	18.1	5.5	404	47
P052	P052-T	PBW55_GG0144T	LUAD	IA1	male	M	0	28.2	6.5	2492	315
P053	P053-T1	PBW56_GG0145-1T	LUAD	IA1	female	M	0	21.9	7.8	4616	936
P053	P053-T2	PBW56_GG0145-2T	MIA	IA1	female	P	0	8.0	0.0	107	0
P053	P053-T3	PBW56_GG0145-6T	AAH	N/A	female	P	0	6.0	0.0	43	0
P054	P054-T1	PBW57_GG0146_4185T	LUAD	IA2	female	M	0	29.6	12.3	8467	2022
P055	P055-T1	PBW58_GG0147-1T	MIA	IA1	male	P	0	10.9	0.0	330	0
P055	P055-T2	PBW58_GG0147-2T	AAH	N/A	male	P	0	4.0	0.0	62	0
P056	P056-T	PBW59_GG0148T	MIA	IA1	male	P	100	14.9	0.0	1217	0
P057	P057-T	PBW60_GG0149T	LUAD	IA1	female	M	0	15.8	4.5	3250	120
P058	P058-T	PBW61_GG0150T	LUAD	IA2	female	M	0	27.8	15.5	6157	2678
P059	P059-T1	PBW62_GGO155_xia	LUAD	IA1	female	P	0	11.2	0.0	475	0
P059	P059-T2	PBW62_GGO155shang	LUAD	IA1	female	M	0	13.8	8.0	1212	246
P060	P060-T1	PBW63_GGO157T_big	LUAD	IA1	female	M	0	18.3	5.8	1171	91
P060	P060-T2	PBW63_GGO157T_small	AAH	N/A	female	P	0	5.5	0.0	69	0
P061	P061-T1	PBW64_GGO160_shang	AIS	0	female	P	0	7.2	0.0	153	0
P061	P061-T2	PBW64_GGO160_xia	MIA	IA1	female	P	0	10.8	0.0	315	7
P062	P062-T	PBW65_GGO162	LUAD	IIA	female	M	0	44.7	40.3	31088	10046
P063	P063-T	P01_2449T	AIS	0	male	P	30	7.1	0.0	328	0
P064	P064-T	P02_L2288T	LUAD	IA1	male	P	12.5	14.9	0.0	753	0
P065	P065-T	P03_L2557T	AIS	0	female	P	0	9.4	0.0	301	0
P066	P066-T	P04_3345T	LUAD	IA2	male	M	40	28.8	16.6	8441	2464
P067	P067-T	P05_L2863T	LUAD	IA1	male	M	30	14.4	5.5	2624	627
P068	P068-T	P07_L2351T	LUAD	IIB	male	M	7.5	34.8	22.5	20895	13542
P069	P069-T	P08_L2430T-SY	AIS	0	female	P	0	14.1	0.0	1775	0

P070	P070-T	P09_L2495T	LUAD	IA3	male	M	20	27.9	25.0	7590	4393
P071	P071-T2	P10_L2505T-XXY	AAH	N/A	male	P	0	7.3	0.0	603	0
P072	P072-T	P11_L2510T	AIS	0	male	P	20	5.9	0.0	120	0
P073	P073-T2	P12_L2526T-D	LUAD	IA2	male	M	0	28.4	16.2	4984	1984
P073	P073-T3	P12_L2526T-Z.1	LUAD	IA2	male	M	0	16.9	10.9	1308	673
P074	P074-T	P13_L2565T	MIA	IA1	male	P	0	8.4	0.0	232	0
P075	P075-T	P14_L2589T	MIA	IA1	female	P	0	12.0	0.0	282	0
P076	P076-T	P15_L2667T	LUAD	IA1	male	P	N/A	18.0	0.0	1400	0
P077	P077-T2	P16_L2763-S2	AIS	0	female	P	0	5.9	0.0	68	0
P077	P077-T4	P16_L2763-S4	AAH	N/A	female	P	0	5.5	0.0	48	0
P077	P077-T5	P16_L2763-Z	AAH	N/A	female	P	0	5.5	0.0	79	0
P077	P077-T3	P16_S3	AAH	N/A	female	M	0	19.0	7.0	599	87.8
P078	P078-T1	P17_L2881T-ZS	LUAD	IA1	female	M	11.25	18.4	3.9	963	103
P078	P078-T2	P17_L962T	LUAD	IA3	female	M	11.25	33.4	21.1	9140	5279
P079	P079-T	P18_L3010T	LUAD	IA1	male	M	5	16.7	4.8	1276	173
P080	P080-T	P19_L3038T	LUAD	IA1	female	M	0	12.8	4.3	594	81
P081	P081-T1	P20_L3089T-SY	AIS	0	female	P	0	9.8	0.0	312	0
P081	P081-T2	P20_L3089T-XY	AIS	0	female	P	0	7.3	0.0	282	0
P082	P082-T	P21_L3116T	AIS	0	male	P	15	9.7	0.0	234	0
P083	P083-T	P22_L3117T	MIA	IA1	female	P	0	8.6	0.0	502	0
P084	P084-T	P23_L3126T-YS	MIA	IA1	male	M	1.5	13.5	5.7	639	205
P085	P085-T	P24_L3130T	MIA	IA1	male	P	0	22.4	0.0	1443	289
P086	P086-T	P25_L3190T-HD	LUAD	IA2	female	M	0	35.2	15.9	5074	1946
P087	P087-T	P26_L3194T	LUAD	IA2	male	M	10	25.2	15.3	7372	2400
P088	P088-T	P27_L3220T	MIA	IA1	male	P	15	9.5	0.0	422	0
P089	P089-T2	P28_L3255-2X	AIS	0	female	P	0	6.0	0.0	132	0
P089	P089-T1	P28_L3255-D	AIS	0	female	P	0	6.7	0.0	131	0
P090	P090-T	P29_L3285-2TX	AAH	N/A	female	P	0	5.3	0.0	52	0

P091	P091-T	P30_L3389T	AIS	0	female	P	0	12.7	0.0	243	0
P092	P092-T	P31_L4253T	LUAD	IA1	male	M	60	13.2	4.1	988	247
P093	P093-T	P32_L4255T	LUAD	IA1	female	M	0	31.2	7.9	4005	1324
P094	P094-T1	P33_GGO166-1	LUAD	IA2	female	M	0	10.9	10.9	1689	817
P094	P094-T2	P33_GGO166-2	AIS	0	female	M	0	11.6	5.5	1070	90.4
P094	P094-T3	P33_GGO166-3	AIS	0	female	P	0	6.6	0.0	36	0
P095	P095-T	P34_GGO158	LUAD	IA2	female	M	0	28.7	12.1	2693	607
P096	P096-T	P35_GGO159	MIA	IA1	male	P	0	17.5	0.0	N/A	N/A
P097	P097-T	P36_GGO164	LUAD	IA1	female	P	0	12.4	0.0	604	0
P098	P098-T	L2442T	LUAD	IA1	female	M	0	9.1	4.0	178	89
P099	P099-T1	L2771T	AIS	0	male	P	0	12.2	0.0	253	0
P099	P099-T2	L2771T2	AAH	N/A	male	P	0	5.2	0.0	44	0
P099	P099-T3	L2771T4	AAH	N/A	male	P	0	4.3	0.0	71	0
P100	P100-T1	L2793TX	AAH	N/A	female	P	0	4.5	0.0	124	0
P100	P100-T2	L2793TXX	AAH	N/A	female	P	0	5.9	0.0	133	0
P101	P101-T	L2829T	LUAD	IA2	male	M	40	30.1	15.8	4836	2536
P102	P102-T	L2853T	LUAD	IA1	male	M	7.5	13.4	6.5	759	126
P103	P103-T	L2876T	LUAD	IA2	male	M	100	17.9	11.8	1016	577
P104	P104-T	L2949T	MIA	IA1	female	P	0	10.8	0.0	180	0
P105	P105-T	L3087T	MIA	IA1	female	M	0	10.9	6.6	505	239
P106	P106-T1	L3224TS	MIA	IA1	female	P	0	5.0	0.0	322	60
P106	P106-T2	L3224TX	LUAD	IA1	female	M	0	17.4	4.4	1751	105
P107	P107-T	L4685	LUAD	IA3	female	M	0	40.0	27.0	1720	835
P108	P108-T	L4693T-T	LUAD	IB	female	M	0	38.2	31.7	22107	16591
P109	P109-T	L4727	LUAD	IA2	female	M	0	17.6	11.4	1926	1062
P110	P110-T	L4738	LUAD	IA1	female	M	0	18.9	6.6	2969	612
P111	P111-T	L4773	LUAD	IA1	female	M	0	32.7	8.8	9862	272
P112	P112-T	L4783T-T	LUAD	IIB	male	M	50	28.1	25.8	12457	7281
P113	P113-T	L4791	LUAD	IA1	female	M	0	24.7	9.8	2513	335

P114	P114-T1	L4935	LUAD	IB	female	M	0	41.9	36.3	30604	20989
P114	P114-T2	L4935T-ZY1	AIS	0	female	M	0	11.1	4.3	491	151
P114	P114-T3	L4935T-ZY2	LUAD	IA1	female	M	0	8.5	5.6	361	149
P115	P115-T	L4978	LUAD	IB	male	M	0	23.5	12.7	3830	2101
P116	P116-T	L4982	LUAD	IA2	female	M	0	27.2	11.0	7407	1480
P117	P117-T	L5007	LUAD	IA2	male	M	19.5	35.4	19.5	10890	5411
P118	P118-T1	L5017	LUAD	IB	female	M	0	51.4	32.3	26109	6095
P118	P118-T2	L5017G-T	LUAD	IA1	female	P	0	4.1	0.0	35	0
P118	P118-T3	L5017J1-T	LUAD	IA1	female	M	0	12.2	3.0	248	0
P119	P119-T1	PE1M_BF_XUEL4-index1_C1_lane7	AAH	N/A	male	P	0	7.5	0.0	130	67
P119	P119-T2	PE1M_BF_XUEL4-index1_C2_lane7	LUAD	IA1	male	M	0	14.0	8.2	1350	889
P119	P119-T3	PE1M_BF_XUEL4-index1_C3_lane7	MIA	IA1	male	P	0	17.9	0.0	969	75
P119	P119-T4	PE1M_BF_XUEL4-index1_C4_lane7	MIA	IA1	male	P	0	13.0	0.0	425	58
P119	P119-T5	PE1M_BF_XUEL4-index1_D1_lane7	MIA	IA1	male	P	0	5.8	0.0	107	0
P120	P120-T1	WES-P3-1	LUAD	IA1	female	M	0	5.7	3.0	123	71
P120	P120-T2	WES-P3-2	LUAD	IA1	female	M	0	18.7	7.5	4360	291

Supplementary Table S2

Supplementary Table S2. Detailed information for multi-region and multi-focal tumors

Patient No	Tumor No	Tumor_Sample_Barcode	GGO or solid
Multi-region tumor information			
1	P002-T_R1	PBW02_GG012_1T	N/A
	P002-T_R2	PBW02_GG012_2T	N/A
2	P018-T_R1	PBW18_F41_1T	Solid
	P018-T_R2	PBW18_F41_2T	GGO
3	P023-T_R1	PBW24_GG081_1T	Solid
	P023-T_R2	PBW24_GG081_2T	GGO
4	P052-T_R1	PBW55_GG0144-1T	N/A
	P052-T_R2	PBW55_GG0144-2T	N/A
5	P058-T_R1	PBW61_GG0150-1T	Solid
	P058-T_R2	PBW61_GG0150-2T	N/A
	P058-T_R3	PBW61_GG0150-3T	GGO
6	P062-T_R1	PBW65_GGO162-1T	Solid
	P062-T_R2	PBW65_GGO162-2T	GGO
7	P115-T_R1	L4978A-T	GGO
	P115-T_R2	L4978B-T	Solid
8	P116_T_R1	L4982A	N/A
	P116_T_R2	L4982B	N/A
	P116_T_R3	L4982C	N/A
	P116_T_R4	L4982D	N/A
	P116_T_R5	L4982E	N/A
9	P117-T_R1	L5007A	Solid
	P117-T_R2	L5007B	GGO
	P117-T_R3	L5007C	GGO
	P117-T_R4	L5007D	Solid
	P117-T_R5	L5007E	Solid
10	P114-T1_R1	L4935A-T	GGO
	P114-T1_R2	L4935B-T	GGO
	P114-T1_R3	L4935C-T	GGO
	P114-T1_R4	L4935D-T	Solid
	P114-T1_R5	L4935E-T	Solid
11	P118-T1_R1	L5017A-T	solid
	P118-T1_R2	L5017B-T	solid
	P118-T1_R3	L5017C-T	GGO
	P118-T1_R4	L5017D-T	GGO
	P118-T1_R5	L5017E-T	GGO
Multi-focal tumor information			
1	P099-T1	L2771T	pGGO
	P099-T2	L2771T2	pGGO
	P099-T3	L2771T4	pGGO
2	P100-T1	L2793TX	pGGO

	P100-T2	L2793TXX	pGGO
3	P106-T1	L3224TS	pGGO
	P106-T2	L3224TX	mGGO
4	P114-T1	L4935	mGGO
	P114-T2	L4935T-ZY1	mGGO
	P114-T3	L4935T-ZY2	mGGO
	P114-T4	L4935T-XMX	solid
5	P118-T1	L5017	mGGO
	P118-T2	L5017G-T	pGGO
	P118-T3	L5017J1-T	mGGO
6	P071-T2	P10_L2505T-XXY	pGGO
	P071-T1	P10_L2505T1	solid
7	P073-T2	P12_L2526T-D	mGGO
	P073-T3	P12_L2526T-Z.1	mGGO
	P073-T1	P12_L2526T-X	solid
8	P077-T1	P16_L2763-S1	solid
	P077-T2	P16_L2763-S2	pGGO
	P077-T4	P16_L2763-S4	pGGO
	P077-T5	P16_L2763-Z	pGGO
	P077-T3	P16_S3	mGGO
9	P078-T1	P17_L2881T-ZS	mGGO
	P078-T2	P17_L962T	mGGO
10	P081-T1	P20_L3089T-SY	pGGO
	P081-T2	P20_L3089T-XY	pGGO
11	P089-T2	P28_L3255-2X	pGGO
	P089-T1	P28_L3255-D	pGGO
12	P094-T1	P33_GGO166-1	mGGO
	P094-T2	P33_GGO166-2	mGGO
	P094-T3	P33_GGO166-3	pGGO
13	P005-T1	PBW05_F23_1T	mGGO
	P005-T2	PBW05_F23_2T	pGGO
14	P010-T2	PBW10_L3720shang	mGGO
	P010-T1	PBW10_L3720xia	solid
15	P011-T2	PBW11_GG061_2T	pGGO
	P011-T1	PBW11_GG061T	mGGO
16	P026-T1	PBW27_GG089_1T	mGGO
	P026-T2	PBW27_GG089_2T	pGGO
17	P043-T1	PBW46_GG0129_1T	mGGO
	P043-T2	PBW46_GG0129_2T	mGGO
	P043-T3	PBW46_GG0129_3T	pGGO
18	P050-T1	PBW53_GG0141-1T	mGGO
	P050-T2	PBW53_GG0141-2T	pGGO
19	P053-T1	PBW56_GG0145-1T	mGGO
	P053-T2	PBW56_GG0145-2T	pGGO
	P053-T3	PBW56_GG0145-6T	pGGO
20	P054-T1	PBW57_GG0146_4185T	mGGO

	P054-T2	PBW57_GG0146-2T	solid
21	P055-T1	PBW58_GG0147-1T	pGGO
	P055-T2	PBW58_GG0147-2T	pGGO
22	P059-T1	PBW62_GGO155_xia	pGGO
	P059-T2	PBW62_GGO155shang	mGGO
23	P060-T1	PBW63_GGO157T_big	mGGO
	P060-T2	PBW63_GGO157T_small	pGGO
24	P061-T1	PBW64_GGO160_shang	pGGO
	P061-T2	PBW64_GGO160_xia	pGGO
25	P119-T1	PE1M_BF_XUEL4-index1_C1_lane7	pGGO
	P119-T2	PE1M_BF_XUEL4-index1_C2_lane7	mGGO
	P119-T3	PE1M_BF_XUEL4-index1_C3_lane7	pGGO
	P119-T4	PE1M_BF_XUEL4-index1_C4_lane7	pGGO
	P119-T5	PE1M_BF_XUEL4-index1_D1_lane7	pGGO
26	P120-T1	WES-P3-1	mGGO
	P120-T2	WES-P3-2	mGGO

Supplementary Table S3

Supplementary Table S3. Comparison of altered pathway frequencies between Chinese SSNs, Chinese advanced-stage LUADs, and TCGA advanced-stage LUADs

Pathway	TCGA II-IV LUADs			SSNs			BGI II-IV LUADs			*p value* (SSNs&BGI) by Fisher's exact test	*p value* (TCGA&BGI) Fisher's exact test
	mut_n	sum	mut_per	mut_n	sum	mut_per	mut_n	sum	mut_per		
01.cell_cycle	15	106	14.15%	3	154	1.95%	6	75	8.00%	0.06194	0.244
02.HIPPO	62	106	58.49%	23	154	14.94%	37	75	49.33%	6.78E-08	0.23
03.MYC	16	106	15.09%	6	154	3.90%	5	75	6.67%	0.3463	0.1007
04.NOTCH	55	106	51.89%	17	154	11.04%	28	75	37.33%	6.03E-06	0.06904
05.NRF2	22	106	20.75%	2	154	1.30%	7	75	9.33%	0.006414	0.0419
06.PI3K	55	106	51.89%	16	154	10.39%	18	75	24.00%	0.009557	0.0002047
07.TGF	8	106	7.55%	5	154	3.25%	3	75	4.00%	0.719	0.3669
08.RTK_RAS	96	106	90.57%	101	154	65.58%	57	75	76.00%	0.1287	0.01144
09.TP53	60	106	56.60%	26	154	16.88%	32	75	42.67%	4.42E-05	0.07167
10.WNT	48	106	45.28%	23	154	14.94%	29	75	38.67%	9.78E-05	0.4458
11.Nucleosome_remodelling	27	106	25.47%	4	154	2.60%	5	75	6.67%	0.1575	0.001276
12.Cytoskeleton_remodelling	10	106	9.43%	3	154	1.95%	12	75	16.00%	0.0001436	0.2481
13.RNA_splicing_processing	17	106	16.04%	25	154	16.23%	1	75	1.33%	0.0003064	0.000724

Supplementary Table S4

Supplementary Table S4. Comparison of mutation rates of some important genes of each pathological group (AAH and AIS, MIA, IAC)

	AAH_AIS			MIA			IAC		
	Total number	Mutation number	Mutation rate	Total number	Mutation number	Mutation rate	Total number	Mutation number	Mutation rate
EGFR	33	5	15.15%	30	15	50.00%	91	57	62.64%
RBM10	33	1	3.03%	30	4	13.33%	91	20	21.98%
TP53	33	1	3.03%	30	2	6.67%	91	17	18.68%
STK11	33	1	3.03%	30	4	13.33%	91	7	7.69%
KRAS	33	1	3.03%	30	2	6.67%	91	4	4.40%
ARID1B	33	0	0.00%	30	0	0.00%	91	6	6.59%
FBXO11	33	0	0.00%	30	0	0.00%	91	6	6.59%