nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-019-13460-3 OPEN

# Genomic and immune profiling of pre-invasive lung adenocarcinoma

Haiquan Chen[1,2,3,4,12,13*], Jian Carrot-Zhang[5,6,7,12,13*], Yue Zhao[1,2,13], Haichuan Hu[1,2], Samuel S. Freeman[5,6,7], Su Yu[1,2], Gavin Ha[8], Alison M. Taylor[5,6,7], Ashton C. Berger[6], Lindsay Westlake[6], Yuanting Zheng[3,9], Jiyang Zhang[3,9], Aruna Ramachandran[5,6], Qiang Zheng[2,10], Yunjian Pan[1,2], Difan Zheng[1,2], Shanbo Zheng[1,2], Chao Cheng[1,2], Muyu Kuang[1,2], Xiaoyan Zhou[2,10], Yang Zhang[1,2], Hang Li[1,2], Ting Ye[1,2], Yuan Ma[1,2], Zhendong Gao[1,2], Xiaoting Tao[1,2], Han Han[1,2], Jun Shang[3,9], Ying Yu[3,9], Ding Bao[3,9], Yechao Huang[3,9], Xiangnan Li[3,9], Yawei Zhang[1,2], Jiaqing Xiang[1,2], Yihua Sun[1,2], Yuan Li[2,10], Andrew D. Cherniack[5,6], Joshua D. Campbell[11], Leming Shi[3,9] & Matthew Meyerson[5,6,7,12*]

Adenocarcinoma in situ and minimally invasive adenocarcinoma are the pre-invasive forms of lung adenocarcinoma. The genomic and immune profiles of these lesions are poorly understood. Here we report exome and transcriptome sequencing of 98 lung adenocarcinoma precursor lesions and 99 invasive adenocarcinomas. We have identified *EGFR*, *RBM10*, *BRAF*, *ERBB2*, *TP53*, *KRAS*, *MAP2K1* and *MET* as significantly mutated genes in the pre/minimally invasive group. Classes of genome alterations that increase in frequency during the progression to malignancy are revealed. These include mutations in *TP53*, arm-level copy number alterations, and HLA loss of heterozygosity. Immune infiltration is correlated with copy number alterations of chromosome arm 6p, suggesting a link between arm-level events and the tumor immune environment.

[1] Department of Thoracic Surgery, Fudan University Shanghai Cancer Center, Shanghai, China. [2] Department of Oncology, Shanghai Medical College, Fudan University, Shanghai, China. [3] State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai, China. [4] Institute of Thoracic Oncology, Fudan University, Shanghai, China. [5] Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA, USA. [6] Broad Institute of MIT and Harvard, Cambridge, MA, USA. [7] Harvard Medical School, Boston, MA, USA. [8] Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [9] Human Phenome Institute, Fudan University, Shanghai, China. [10] Department of Pathology, Fudan University Shanghai Cancer Center, Shanghai, China. [11] Division of Computational Biomedicine, Department of Medicine, Boston University, Boston, MA, USA. [12] These authors jointly supervised this work: Haiquan Chen, Jian Carrot-Zhang, Matthew Meyerson. [13] These authors contributed equally: Haiquan Chen, Jian Carrot-Zhang, Yue Zhao. *email: hqchen1@yahoo.com; zhangj@broadinstitute.org; matthew_meyerson@dfci.harvard.edu

Lung adenocarcinoma (LUAD) is the most common histological subtype of lung cancer, with an average 5-year survival rate of 15%[1,2]. In contrast, the pre-invasive stages of LUAD, such as adenocarcinoma in situ (AIS) and minimally invasive adenocarcinoma (MIA), are associated with a nearly 100% survival rate, after surgical resection[3–5]. AIS is defined as a ≤3 cm adenocarcinoma lacking invasion, while MIA is a ≤3 cm adenocarcinoma with ≤5 mm invasion[6]. Although some focused studies have identified mutations in lung cancer drivers in AIS and MIA[7–10], there remains a lack of deep insight into the molecular events driving progression of these lesions to invasive LUAD. To address this gap in our knowledge of AIS/MIA pathogenesis, we undertook a systematic investigation of the genomic and immune profiles of pre/minimally invasive lung lesions. Known driver mutations are present in the lung precursors. T cell and B cell responses to the AIS/MIA samples are observed. By comparing the genomic landscapes of the pre-invasive and invasive samples, we suggest the potential molecular events underlying the invasiveness of LUAD.

## Results

**The landscape of somatic alterations in AIS and MIA**. We performed whole-exome sequencing (WES) and RNA-sequencing (RNA-seq) on tumor and matched adjacent normal tissue of 24 AIS, 74 MIA, and 99 invasive LUAD samples (Supplementary Table 1), obtained from patients who underwent surgery at Fudan University Shanghai Cancer Center (FUSCC). We identified eight significantly mutated genes in AIS and MIA specimens, including *EGFR*, *RBM10*, *BRAF*, *ERBB2*, *TP53*, *KRAS*, *MAP2K1*, and *MET*, all previously reported as recurrently mutated in LUAD from The Cancer Genome Atlas (TCGA) cohort[11,12]. *EGFR*, *TP53*, *RB1*, and *KRAS* were significantly mutated in the tested LUAD cases (Fig. 1a, b). Amplified regions that included *MDM2*, *MYC*, *TERT*, *KRAS*, *NKX2-1*, and *CDK6* were observed in the AIS or MIA samples (Fig. 1c). Novel amplifications of *RIT1* were identified in the FUSCC LUAD cohort (Supplementary Fig. 1). RNA-seq analysis revealed a *RET* fusion in an MIA sample (Fig. 1a), and *ALK* and *ROS1* fusions in LUAD (Fig. 1b). When testing significantly mutated genes, *TP53* mutations were the most enriched alteration in the invasive stage (38%) compared to pre/minimally invasive stages (6%), followed by *EGFR* and *RB1* mutations (Fig. 1d). When testing all mutated genes in the pre/minimally invasive lung lesions, only *TP53* mutations significantly increased in frequency through malignancy, after false discovery rate correction.

**The relatively simpler genomes in AIS and MIA than LUAD**. Tumor mutation burden (TMB) was significantly lower in AIS and MIA, compared to stage I LUAD (Supplementary Fig. 2a). Mutational signature analysis identified aging, smoking, APOBEC, and DNA mismatch repair signatures in our cohort. The APOBEC signature was higher in MIA compared to LUAD, although the smoking signature activity did not differ among the three groups (Supplementary Fig. 2b, c). Arm-level copy-number alteration (CNA) was less common in the pre/minimally invasive stages, with a median of 5, 11, and 26 events in AIS, MIA, and LUAD, respectively (Supplementary Fig. 3a). Similarly, focal CNA increased from MIA to LUAD (Supplementary Fig. 3b). TMB, arm-level CNA and focal CNA were all correlated with advancing malignant potential, controlling for specimen purity (linear regression, $p < 0.001$, Methods, Supplementary Fig. 4a, b).

**Molecular mechanism underlying the invasive progression**. Next, we tested the association of genes with increased alteration frequency from AIS/MIA to LUAD and genomic features that distinguish LUAD from AIS/MIA (increased TMB, APOBEC signature, and focal and arm-level CNAs). Notably, *TP53* mutations were strongly correlated with arm-level and TMB, but marginally correlated with focal CNA events (Fig. 2a, b). These data suggest that, in contrast to oncogenic mutations, which occurred frequently in pre/minimally invasive lung tumors, *TP53* mutations were highly involved in the invasiveness during tumor development.

**Immune characterization of AIS and MIA**. In the analysis of T cell receptor (TCR) repertoire and B cell receptor (BCR) repertoire, we observed a tendency that the highest-frequency T cell clones or B cell clones in the tumors were represented as lower frequency clones in the matched normal tissues (Supplementary Fig. 5a, b). However, neither T cell nor B cell clonality was increased from normal samples to AIS/MIA or LUAD (Supplementary Fig. 6a, b).

Loss of human leukocyte antigen (HLA) alleles has been identified as a potential immune escape mechanism in lung cancers[13,14] and can be observed as a subclonal event in LUADs[14]. In our study, we noted HLA loss of heterozygosity (LOH) in 3.1% of AIS/MIA and 16.7% of LUAD specimens (Fig. 3a). The significantly increased frequency of HLA LOH in the invasive group compared to the pre-invasive group (Fisher's exact test, $p < 0.01$) suggested the potential role of loss of HLA alleles during tumor development. The frequency of germline HLA homozygosity, however, was similar in all three stages (Supplementary Fig. 7a). Approximately 60% of the HLA LOH events in LUAD were related to loss of chromosome 6p. Interestingly, we found that 6p gain was significantly anti-correlated with T cell abundance (Mann–Whitney *U* test, $p = 0.038$, Fig. 3b), and this trend was also observed when analyzing B cell infiltration in correlation with 6p CNA (Supplementary Fig. 7b–d). We subsequently tested the correlation of immune infiltration with large-scale chromosome alterations, using samples from the TCGA LUAD cohort. We observed the most significant correlation of leukocyte fraction[15] with chromosome 6p CNA ($p = 0.0030$, coef. = −0.74, 95% CI: −1.23 to −0.25), followed by 1q ($p = 0.0033$, coef. = −0.60, 95% CI: −1 to −0.2) and 19p CNA ($p = 0.0047$, coef. = 0.53, 95% CI: 0.16 to 0.9), after controlling for TMB and the degree of overall aneuploidy (see Methods, Fig. 3c, d). 6p and 1q CNA showed significantly increased frequency from AIS/MIA to LUAD in the FUSCC cohort (Fisher's exact test, $p < 0.001$, Supplementary Fig. 7e).

## Discussion

We have interrogated the genomic and immune features of pre/minimally invasive lung cancers. Seventy-one percent of AIS and MIA patients carried at least one mutation in previously identified cancer genes in the RTK/RAS/RAF pathway, similar to the oncogenic driver events found in LUAD. In addition, we showed an overall high frequency of *EGFR* mutations (65% in LUAD), which may reflect the enrichment of never smoking patients with East Asian origin in our cohort. APOBEC-related mutations are contributors to lung cancer heterogeneity[16], and might be involved in the progression from AIS/MIA to LUAD[10]. We found that genomic aberrations including TMB, APOBEC signature, and arm and focal CNA were increased from the pre-invasive to invasive stage. Mutations in *TP53* and HLA LOH also increased in frequency in the aggressive stage .

Our work reveals *TP53* as a key mediator in the invasiveness of lung cancer. Previous studies in Barrett's esophagus suggested that *TP53* occurred early in esophageal adenocarcinoma precursors followed by oncogenic amplifications[17]. *TP53* was also frequently mutated in lung carcinoma in situ, which is the

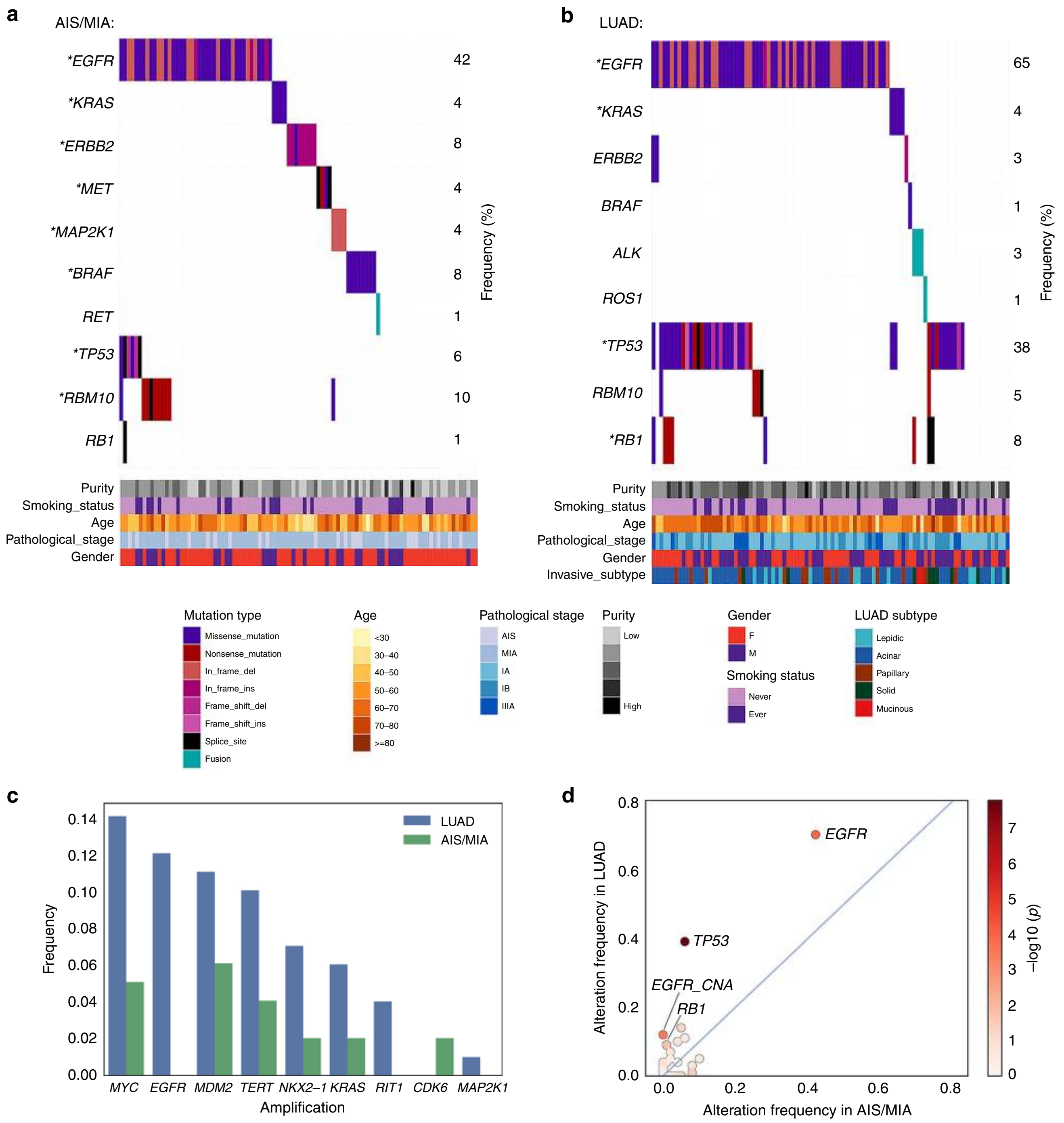

**Fig. 1** Somatic alterations in pre-invasive and invasive lung adenocarcinomas. **a** Co-mutation plots for AIS/MIA and **b** LUAD. Stars indicate significantly mutated genes in each group. **c** Lung cancer genes with focal amplification in AIS/MIA and LUAD. **d** Somatic alterations with higher frequencies in LUAD, compared to AIS and MIA. Color bar represents $\log_{10}$-transformed *p* value calculated from two-sided Fisher's exact test. Source data are provided as a source data file.

precursor form of squamous cell carcinoma[18]. We have shown the high frequency of oncogenic driver mutations, but low frequency of *TP53* mutations in the LUAD precursors. Previous studies have suggested the functional association of *TP53* mutations with invasive potential in cancers[19]. Our findings also demonstrate a strong association of *TP53* mutations with aneuploidy, in line with recent work from TGCA[20]. Given previous reports of aneuploidy in association with decreased immune infiltration[20,21], our data raise the possibility that copy-number changes in specific chromosomes may influence the tumor microenvironment. Our work provides new insights into the biology of lung pre-malignancy, with implications for disease monitoring and prognosis, and future therapeutic intervention.

## Methods

**Patient cohort and pathological review**. One hundred and ninety-seven patients who underwent surgery between September 2011 and May 2016 at the Department of Thoracic Surgery, Fudan University Shanghai Cancer Center were enrolled in this study. No patient received neoadjuvant therapy. Preoperative tests, including contrast-enhanced chest computed tomography (CT) scanning, were performed to determine the clinical stage of the disease. Fiber optic bronchoscopy was routinely

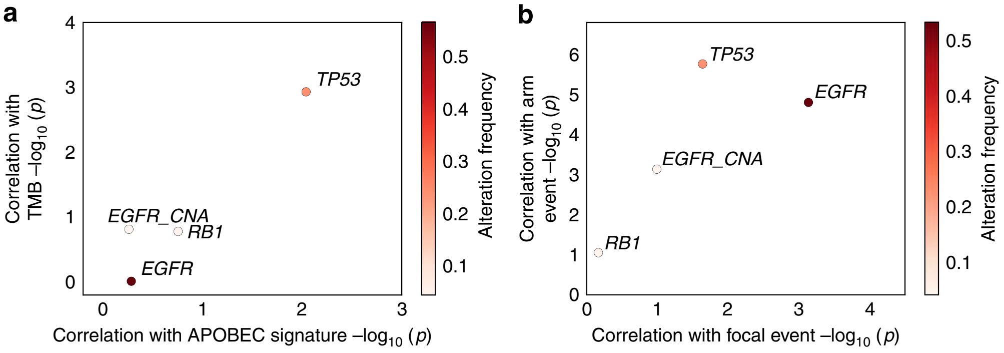

**Fig. 2** Correlation of somatic alterations with genomic features. *TP53*, *EGFR*, *RB1* mutations and *EGFR* amplification in correlation with **a** TMB and APOBEC signature, and **b** arm and focal CNA. Student's *t* test was used to calculate the $\log_{10}$-transformed *p* value. Samples in all stages were included to calculate the alteration frequency. Source data are provided as a source data file.

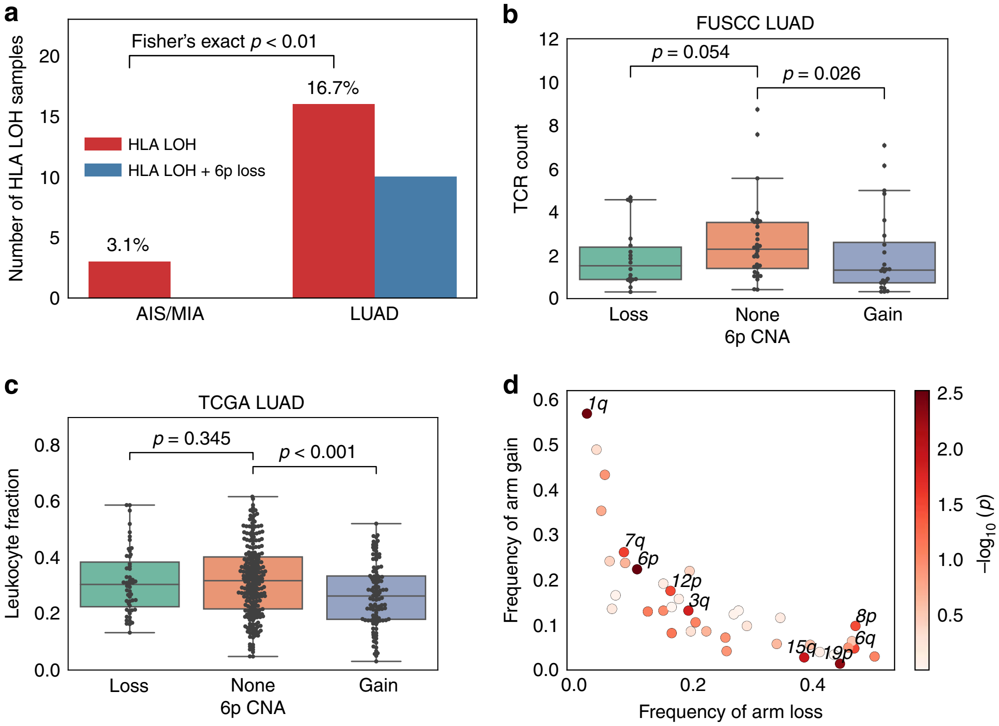

**Fig. 3** Tumor immune environment in association with arm-level CNA. **a** Frequency of loss of HLA heterozygosity and the co-occurrence of HLA LOH with 6p loss. Significantly more HLA LOH events are found in the LUAD group compared to the AIS/MIA group. **b** Comparison of inferred T cell infiltration in FUSCC LUAD samples and **c** leukocyte infiltration[15] in TCGA LUAD samples with 6p CNA loss, gain, or no change. *P* values are calculated from Mann–Whitney *U* test. Significantly decreased level of T cell or leukocyte infiltrations are found in 6p gain samples compared to 6p neutral samples. In the box plots, the upper and lower hinges represent the first and third quartile, the whiskers span the first and third quartile, and center lines represent the median. **d** Correlation of arm-level CNA with leukocyte infiltration for the TCGA LUAD samples. *P* values are calculated from multivariate linear regression, while each arm is assigned 1 if gained, −1 if lost and 0 if unchanged, and adding the aneuploidy score[18] and TMB as covariates. Source data are provided as a source data file.

performed. When necessary, CT-guided hook-wire localization was performed before surgery, to define the resection area. Tumor specimens were initially sent for intraoperative frozen section diagnosis after they were removed. The specimen was sliced at the largest diameter of the tumor for sampling. Usually two sections of each specimen were made for intraoperative diagnosis. After surgery, the tumor specimens were sent to be reviewed by two pathologists independently to confirm the clinical stage and determine the histological classification. Stage IIIA patients in this study cohort were those with initial clinical stage I diagnosis, but mediastinal lymph node metastasis was found by postsurgical pathological review. Usually 3–5 sections of each specimen were used to determine the final pathological diagnosis. Tumors were classified into AIS, MIA, and invasive adenocarcinoma, according to the LUAD classification of the International Association for the Study of Lung Cancer, American Thoracic Society, and European Respiratory Society[1]. For invasive adenocarcinomas, the occupancy of each one of these several patterns, namely, lepidic, acinar, papillary, micropapillary, solid, and invasive mucinous adenocarcinoma, was recorded in a 5% increment, and the subtype with the highest percentage was considered as the predominant subtype. This study was approved by the Committee for Ethical Review of Research (Fudan University Shanghai Cancer Center Institutional Review Board No. 090977-1). Informed consents of all patients for donating their samples to the tissue bank of Fudan University Shanghai Cancer Center were obtained from patients themselves or their relatives. Source data are provided as a source data file.

**Whole-exome sequencing**. Genomic DNA from tumors and paired adjacent normal tissues was extracted and prepared using the QIAamp DNA Mini Kit (Qiagen) following the manufacturer's instructions. Exon libraries were constructed using the SureSelect XT Target Enrichment System. A total amount of 1–3 μg genomic DNA for each sample was fragmented into an average size of ~200 bp. DNA was captured using SureSelect XT reagents and protocols to generate indexed, target-enriched library amplicons. Constructed libraries were then sequenced on the Illumina HiSeq X Ten platform and 150 bp paired-end reads were generated.

**RNA-sequencing**. Total RNA from tumors and paired adjacent normal tissues was extracted and prepared using NucleoZOL (Macherey-Nagel) and NucleoSpin RNA Set for NucleoZOL (Macherey-Nagel) following the manufacturer's instructions. A total amount of 3 μg RNA per sample was used as initial material for RNA sample preparations. Ribosomal RNA was removed using Epicenter Ribo-Zero Gold Kits (Epicenter, USA). Subsequently, the sequencing libraries were generated using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (NEB, Ipswich, USA) according to manufacturer's instructions. Libraries were then sequenced on the Illumina HiSeq X Ten platform and 150 bp paired-end reads were generated.

**Alignment and mutation calling**. Sequencing reads from the exome capture libraries were aligned to the reference human genome (hg19) using BWA-MEM[22]. The Picard tools (https://broadinstitute.github.io/picard/) was used for marking PCR duplicates. The Genome Analysis Toolkit[23] was used to perform base quality recalibration and local indel re-alignments. SNVs were called using MuTect and MuTect2[24]. Indels were called using MuTect2 and Strelka v2.0.13[25]. Variants were filtered if called by only one tool. Oncotator v1.9.1[26] was used for annotating somatic mutations. Significantly mutated genes were identified using MutSig2CV[27]. TMB was calculated as the total number of nonsynonyous SNVs and indels per sample divided by 30, given coverage of ~30 MB. Linear regression was used to test the correlation of TMB with disease stages, while coding AIS, MIA, and LUAD as 0, 1, and 2, respectively, and adding purity as a covariate.

**Mutational signature and copy-number changes**. Mutational signature was called using SignatureAnalyzer[28] with SNVs classified by 96 tri-nucleotide mutation. Read coverage was calculated at 50 kb bins across the genome and was corrected for GC content and mappability biases using ichorCNA v0.1.0[29]. The copy-number analysis was performed using TitanCNA v1.17.1[30]. GISTIC 2.0.22[31] was used to identify amplification peaks and to separate arm and focal level CNA using ichorCNA generated segments. Arm-level event was defined by $\log_2$-transformed copy-number ratio >0.1 or <−0.1. Focal level events were defined by $\log_2$-transformed copy-number ratios of >1 or <−1. For *EGFR* and *KRAS* in the AIS/MIA samples, we lowered the amplification threshold to 0.8, and did not detect additional events. Purity and ploidy were calculated by the ABSOLUTE algorithm[32]. Linear regression was used to test the correlation of focal and arm-level CNA with disease stages, while coding AIS, MIA, and LUAD coded as 0, 1, and 2, respectively, and adding purity as a covariate.

**Analysis of expression and fusion**. RNA-seq reads were aligned to the reference human genome (hg19) with STAR v2.5.3[33]. Expression values were normalized to the transcripts per million (TPM) estimates using RSEM v1.3.0[34]. The $\log_2$-transformed TPM values were used to measure gene expression. Fusion events were called using STAR-fusion[35]. We focused on known lung cancer fusions (*ALK*, *ROS1*, *NTRK2*, *RET*, and *MET*) with read count supporting the fusion event >10, and visually inspected the BAM files to ensure accuracy.

**TCR, BCR, and HLA analysis**. TCR or BCR sequences were analyzed using MiXCR 2.1.11[36] based on the RNA-seq data. The reads per million (RPM) value was used to normalize the total TCR or BCR count to the total reads aligned in sample. Infiltration was inferred by the RPM of TCR or BCR count. T cell or B cell diversity is inferred by the Shannon entropy score. Samples that have at least 10 clones with clone count >5 were used in the entropy test. For each sample, we calculated the entropy score based on the top 10 clones. Samples with purity <0.2 and >0.8 were excluded. Samples with possible contamination (top clones found in more than one samples) were excluded. HLA types were called with POLYSOLVER[37]. Loss of HLA heterozygosity was called by LOHHLA[14]. An event of the copy number calculated with binned B-allele frequency <0.5 and the *p* value (Pval_unique) of allelic imbalance <0.1 was considered as HLA LOH for AIS or MIA, and 0.05 for LUAD. For the analyses with TCGA samples, we obtained the fraction of leukocytes, TMB, aneuploidy score, and arm-level CNA from Taylor et al.[20]. Linear regression was used to test the correlation of arm CNA with the leukocyte fraction, while coding loss, gain, and none as −1, 1, and 0, respectively, and adding TMB and aneuploidy score as covariates.

**Reporting summary**. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

## Data availability

Raw data from WES and RNA-seq of AIS/MIA and LUAD have been deposited at European Genome-phenome Archive (EGA) under the accession code EGAS00001004006. Source data underlying all figures are provided as a Source Data file.

## Code availability

All custom code used in the analyses is available at https://github.com/jcarrotzhang/Code-for-preinvasive.

Received: 30 November 2018; Accepted: 31 October 2019;
Published online: 29 November 2019

## References

1. Siegel, R. L. et al. Cancer statistics, 2018. *CA Cancer J. Clin.* **68**, 7–30 (2018).
2. Chen, W. et al. Cancer statistics in China, 2015. *CA Cancer J. Clin.* **66**, 115–132 (2016).
3. Yim, J. et al. Histologic features are important prognostic indicators in early stage lung adenocarcinomas. *Mod. Pathol.* **20**, 233–241 (2007).
4. Borczuk, A. C. et al. Invasive size is an independent predictor of survival in pulmonary adenocarcinoma. *Am. J. Surg. Pathol.* **33**, 462–469 (2009).
5. Maeshima, A. M. et al. Histological scoring for small lung adenocarcinomas 2 cm or less in diameter: a reliable prognostic indicator. *J. Thorac. Oncol.* **5**, 333–339 (2010).
6. Travis, W. D. et al. International association for the study of lung cancer/American thoracic society/European respiratory society international multidisciplinary classification of lung adenocarcinoma. *J. Thorac. Oncol.* **6**, 244–285 (2011).
7. Murphy, S. J. et al. Genomic rearrangements define lineage relationships between adjacent lepidic and invasive components in lung adenocarcinoma. *Cancer Res.* **74**, 3157–3167 (2014).
8. Izumchenko, E. et al. Targeted sequencing reveals clonal genetic changes in the progression of early lung neoplasms and paired circulating DNA. *Nat. Commun.* **6**, 8258 (2015).
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.* **26**, 156–161 (2015).
10. Vinayanuwattikun, C. et al. Elucidating genomic characteristics of lung cancer progression from in situ to invasive adenocarcinoma. *Sci. Rep.* **6**, 31628 (2016).
11. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
12. Campbell, J. D. et al. Distinct patterns of somatic genome alterations in lung adenocarcinomas and squamous cell carcinomas. *Nat. Genet.* **48**, 607–616 (2016).
13. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519–525 (2012).
14. McGranahan, N. et al. Allele-specific HLA loss and immune escape in lung cancer evolution. *Cell* **171**, 1259–1271 (2017).
15. Thorsson, V. et al. The immune landscape of cancer. *Immunity* **48**, 812–830 (2018).
16. de Bruin, E. C. et al. Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251–256 (2014).
17. Stachler, M. D. et al. Paired exome analysis of Barrett's esophagus and adenocarcinoma. *Nat. Genet.* **47**, 1047–1055 (2015).
18. Teixeira, V. H. et al. Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. *Nat. Med.* **25**, 517–525 (2019).
19. Goh, A. M. et al. The role of mutant p53 in human cancer. *J. Pathol.* **223**, 116–126 (2011).
20. Taylor, A. M. et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676–689 (2018).
21. Davoli, T. et al. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy. *Science* **355**, eaaf8399 (2017).
22. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
23. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
24. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
25. Saunders, C. T. et al. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. *Bioinformatics* 28, 1811–1817 (2012).
26. Ramos, A. H. et al. Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).
27. Lawrence, M. S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
28. Kim, J. et al. *Somatic ERCC2 mutations are associated with a distinct genomic signature in urothelial tumors. Nat. Genet.* **48**, 600–606 (2016).

29. Adalsteinsson, V. A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tuomrs. *Nat. Commun.* **8**, 1324 (2017).
30. Ha, G. et al. TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
31. Mermel, C. H. et al. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
32. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
33. Dobin, A. et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
34. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
35. Brian, J. H. et al. STAR-Fusion: fast and accurate fusion transcript detection from RNA-seq. Preprint at https://www.biorxiv.org/content/early/2017/03/24/120295 (2017).
36. Bolotin, D. A. et al. MiXCR: software for comprehensive adaptive immunity profiling. *Nat. Methods* **12**, 380–381 (2015).
37. Shukla, S. A. et al. Comprehensive analysis of cancer-associated somatic mutations in class I HLA genes. *Nat. Biotechnol.* **33**, 1152–1158 (2015).

## Acknowledgements

We would like to first acknowledge the patients for their participation in this study. All patients had signed informed consent for donating their samples to the tissue bank of Fudan University Shanghai Cancer Center. This study is supported by the National Natural Science Foundation of China (81330056, 81930073, 81572253, 31720103909, 31471239, and 31671368), the National Human Genetic Resources Sharing Service Platform (2005DKA21300), National Key R&D Program of China (2017YFC1311004, 2016YFC1201701, and 2016YFC0902302), Shanghai R&D Public Service Platform Project (12DZ2295100), Shanghai Shen Kang Hospital Development Center City Hospital Emerging Cutting-edge Technology Joint Research Project (SHDC12017102), National Key Research and Development Plan (2016YFC0902302), Chinese Minister of Science and Technology grant (2016YFA0501800 and 2017YFA0505501), the National Key R&D Project of China (2016YFC0901704, 2017YFC0907502, and 2017YFF0204600), Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), and Shanghai Municipal Health Commission Key Discipline Project (2017ZZ02025 and 2017ZZ01019). M.M. receives a grant from Stand Up to Cancer (SU2C-AACR-DT23-17) and the Pre-Cancer Genome Atlas 2.0 (1U2CCA233238-01). J.C.-Z. has a Canadian Institutes of Health Research (CIHR) fellowship. J.D.C. is funded by the LUNGevity Career Development award. We thank Galen Gao and Kar-Tong Tan for their helpful suggestions.

## Author contributions

H.C. and M.M. designed the study. J.C.-Z. led the computational analyses. Y.Z. led the experiments. H. Hu, S.Y., Yawei Zhang, J.X. and Y.S. collected samples. Y.Z., S.S.F., G.H., A.M.T., A.C.B., L.W., A.R., A.D.C., and J.D.C. provided critical input for the analyses. Q.Z, X.Z. and Y.L. performed pathological review. Y.Y., D.B., Y.H., X.L., and L.S. performed quality control of data. T.Y., Y.P, D.Z., S.Z., C.C., M.K., Yang Zhang, H.L., Y.M., Z.G., X.T. and H. Han collected clinical information. Y. Zheng, J.Z., J.S., and L.S. aided in study design. J.C.-Z wrote the original draft of the manuscript. Y.Z. provided text and figures. H.C., M.M., A.R., S.S.F, A.M.T., A.D.C. and J.D.C. edited the manuscript.

## Competing interests

M.M. is the scientific advisory board chair of OrigiMed; an inventor of a patent licensed to LabCorp for *EGFR* mutation diagnosis; and receives research funding from Bayer. M. M. and A.M.T. receive research funding from Ono Pharmaceutical.

## Additional information

**Supplementary information** is available for this paper at https://doi.org/10.1038/s41467-019-13460-3.

**Correspondence** and requests for materials should be addressed to H.C., J.C.-Z. or M.M.

**Peer review information** *Nature Communications* thanks the anonymous reviewer(s) for their contribution to the peer review of this work.