Evolution of DNA methylome from precancer to invasive lung adenocarcinoma

Xin Hu[1,#], Marcos Roberto Estecio[2,13,#], Runzhe Chen[3], Alexandre Reuben[3], Linghua Wang[1], Junya Fujimoto[4], Jian Carrot-Zhang[15,16,17], Lisha Ying[5,6], Junya Fukuoka[7], Chi-Wan Chow[4], Nicholas McGranahan[8], Hoa Pham[7], Myrna C.B. Godoy[9], Brett W. Carter[9], Carmen Behrens[3,4], Jianhua Zhang[1], Ara A. Vaporciyan[10], John V. Heymach[3], Ignacio I. Wistuba[4], Yue Lu[2,13], Harvey Pass[14], Humam Kadara[4], Paul Scheet[11], Jack J. Lee[12], P. Andrew Futreal[1,&], Dan Su[6,18,&], Jean-Pierre Issa[19,&], Jianjun Zhang[1,3,&]

Affiliations: Departments of [1]Genomic Medicine, [2]Epigenetics and Molecular Carcinogenesis, [3]Thoracic/Head and Neck Medical Oncology, [4]Translational Molecular Pathology, [9]Thoracic Imaging, [10]Thoracic and Cardiovascular Surgery, [11]Epidemiology, [12]Biostatistics, [13]Center of Cancer Epigenetics, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.

[5]Zhejiang Cancer Research Institute, [18]Department of Pathology, Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), Hangzhou, China.

[6]Institute of Cancer and Basic Medicine (IBMC), Chinese Academy of Sciences, Hangzhou, China.

[7]Department of Pathology, Nagasaki University Graduate School of Biomedical Sciences, Nagasaki, Japan.

[8]Cancer Research United Kingdom-University College London Lung Cancer Centre of Excellence, London, UK.

[14]Department of Cardiothoracic Surgery, New York University Langone Medical Center, New York, NY 10016, USA

[15]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[16]Department of Medical Oncology, Dana-Farber Cancer Institute, [17]Harvard Medical School, Boston, MA, USA.

[19]Coriell Institute for Medical Research, Camden, New Jersey, USA.

[#]Co-first authors.

[&]Correspondence should be addressed to Jianjun Zhang (jzhang20@mdanderson.org), Jean-Pierre Issa (jpissa@coriell.org), Dan Su (sudan@zjcc.org.cn), or P. Andrew Futreal (afutreal@mdanderson.org).

Key words: epigenetic intratumor heterogeneity, methylome evolution, immune evasion, lung precancers

ABSTRACT

The evolution of DNA methylation at genome level and methylation intra-tumor heterogeneity (ITH) during early lung carcinogenesis has not been systematically studied. We performed multiregional reduced representation bisulfite sequencing (RRBS) of 62 resected lung nodules from 39 patients including atypical adenomatous hyperplasia (AAH, n=14), adenocarcinoma in situ (AIS, n=15), minimally invasive adenocarcinoma (MIA, n=22), and invasive adenocarcinoma (ADC, n=11). We observed significantly higher level of methylation ITH in later-stage lesions and gradual increase in both hyper- and hypomethylation compared to matched normal lung tissues over the course of neoplastic progression. The phyloepigenetic patterns inferred from methylation aberrations resembled those based on somatic mutations suggesting parallel methylation and genetic evolution during the early lung carcinogenesis. De-convolution of transcriptomic profiles from a previously published cohort and RBBS data from the current cohort demonstrated higher ratios of T regulatory cells (Tregs) versus CD8+ T cells in later-stage diseases implying progressive immunosuppression with neoplastic progression. Furthermore, increased global hypomethylation was associated with higher mutation burden, higher copy number aberration burden, higher allelic imbalance burden as well as higher Treg/CD8 ratio highlighting the potential impact of methylation states on chromosomal instability, mutagenesis and the tumor immune microenvironment.

INTRODUCTION

Lung cancer remains the leading cause of cancer-related death worldwide, yet it is curable if treated early. Many cancers, including lung cancers, are preceded by precancers. Treating precancers to prevent invasive lung cancer is theoretically an attractive approach to reduce lung cancer–associated morbidity and mortality. However, developing strategies for lung cancer prevention has been challenging owing to our

limited understanding of neoplastic progression from precancers to invasive lung cancers[1]. Lung adenocarcinoma (ADC) is the most common histologic subtype, accounting for more than 50% of all lung cancers[2]. It has been proposed that invasive lung ADC evolves from atypical adenomatous hyperplasia (AAH), the only recognized precancer for lung ADC, which could progress to ADC *in situ* (AIS), a pre-invasive lung cancer, then to minimally invasive ADC (MIA), and eventually frankly invasive ADC[3]. AAH, AIS, MIA and some ADC often present as pulmonary nodules with a unique radiologic feature termed ground-glass opacity (GGO, hazy nodular opacity with the preservation of underlying bronchial and vascular margins) on CT scans[4]. The biology and clinical course of these lesions are not well defined and their management is controversial. Surgical resection is not the standard of care for treating these lung nodules and the diagnostic yield of biopsy is often low, particularly for GGO-dominant nodules[5]. Therefore, these lung nodules are often referred to as indeterminate pulmonary nodules (IPN) [6]. However, as surgical resection is not often offered, obtaining adequate tissue for comprehensive profiling of these IPNs is difficult, hindering our understanding of the biology underlying these lesions.

We have initiated an international collaboration for the collection and characterization of these lung precancers, pre-invasive and early invasive ADC presenting as IPNs. We recently reported the genomic landscape, including the genomic intra-tumor heterogeneity (ITH), of 116 IPNs from 53 patients and revealed evidence of progressive evolution from AAH to AIS, MIA, and ADC at the single-nucleotide level and macroevolution at the transitions from AAH to AIS, and from AIS to MIA associated with somatic copy number aberration (SCNA) and allelic imbalance (AI), respectively[7]. In addition, we found that genomic ITH is more pronounced in AAH than in ADC, which suggests a clonal sweeping model during the neoplastic progression of AAH[7].

In addition to mutations, somatic epigenetic alterations such as DNA methylation can also impact neoplastic transformation and fitness[8-11]. Recent genome-wide methylation profiling studies have revealed that certain alterations, such as the silencing of tumor suppressor genes (TSG) and the activation of genes in stem-like cellular programs[12-14] may contribute to carcinogenesis. Our previous study has demonstrated that complex methylation ITH was associated with larger tumor size and increased risk of postsurgical recurrence in patients with invasive lung ADC[15]. Methylation aberrations have been reported in AAH lesions and tumor-adjacent lung tissues suggesting that somatic methylation changes may be early molecular events[16]. However, these pioneer studies only analyzed small numbers of genes implicated in lung carcinogenesis. The dynamic changes in the methylome at the genome level and the evolutionary trajectories of methylation ITH during the initiation and progression of lung precancers have not been studied systematically. In the current study, using a unique cohort of resected IPNs of different histologic stages, we delineated the evolution of the methylation landscape and methylation ITH architecture during progression from precancers to invasive lung ADC.

RESULTS

DNA methylation aberrations increase with the progression of precancers

We performed reduced representation bisulfite sequencing (RRBS) of 62 resected IPNs (14 AAH, 15 AIS, 22 MIA, and 11 invasive ADC) and their paired normal lung tissues from 39 patients (**Supplementary Table 1**). There was no significant difference regarding age (p= 0.6288, Kruskal-Wallis H test), sex (p= 0.6482, Chi-squared test) or smoking status (p= 0.5696, Chi-squared test) between different histologic groups (**Supplementary Table 2**). The mean RRBS sequencing coverage was 58.98 reads. With a minimum of 10 reads shared across all samples, the RRBS profiling allowed the

quantification of methylation status at 751,462 CpG sites mapped to 15,761 known genes. Principal component analysis (PCA) demonstrated that the DNA methylome of AAH was more similar to that of normal lung tissue, whereas those of AIS, MIA, and ADC were clearly different from that of normal lung (**Fig. 1A**). Similarly, unsupervised hierarchical clustering identified two separate clusters: one comprising normal lung and AAH, and the other including AIS, MIA, and ADC (**Supplementary Fig. 1**) indicating a major change in the global DNA methylation landscape (i.e., macroevolution) at the transition from AAH to AIS. In addition, different regions from the same IPNs tended to cluster together, suggesting that inter-lesion heterogeneity was more prevalent than intra-lesion heterogeneity. Similarly, the methylation profiles of different regions from the same IPNs were correlated (median coefficient r, 0.884 [range, 0.736-0.969], $p < 2.2 \times 10^{-16}$, **Supplementary Fig. 2**). There was a progressive increase in both hypomethylation and hypermethylation (as compared with matched normal lung tissues from the same patients) from AAH to AIS, MIA, and invasive ADC; hypermethylation emerged as early as AAH, whereas hypomethylation was evident in only AIS, MIA, and ADC. Meanwhile, the correlation between the methylation profiles of IPNs and those of their paired normal tissues progressively decreased ($r = 0.885$ for AAH–normal, 0.824 for AIS–normal, 0.796 for MIA–normal, and 0.740 for ADC–normal, $p < 2.2 \times 10^{-16}$ for all 4 comparisons; **Fig. 1B**). Further quantification of the IPNs' differentially methylated regions (DMRs) compared with their paired normal lung tissues revealed that later-stage IPNs had more CpG sites with hypermethylation ($p = 0.3635$, Kruskal-Wallis test) and significantly more CpG sites with hypomethylation ($p = 0.000287$, Kruskal-Wallis test) (**Fig. 1C**).

Enrichment of transcription factor binding sites at DMRs

DNA methylation can impact the DNA binding of transcription factors (TFs) to their target sequences, termed "motifs"[17,18]. We searched the motifs that were covered by the DMRs in the IPNs and identified 50 motifs in AAH, 57 in AIS, 64 in MIA, and 66 in ADC that were significantly enriched over all DMRs in each histologic stage (**Supplementary Table 3**). Some of these motifs were significantly aligned with known motifs recognized by different TFs, such as *MYC, ITF2, KLF4,* and *WT1* (**Supplementary Fig. 3A-C, Supplementary Table 4**). These TFs are involved in critical biological processes, including cell cycle progression, cell proliferation, apoptosis, tumor metastasis, angiogenesis and immune response[19-22], indicating that the epigenetic regulation of these processes, by modifying TF binding capacity, may have potential roles during the early carcinogenesis of lung ADC.

Later-stage disease has higher epigenetic ITH

Molecular ITH can have a profound impact on cancer biology[23,24], and our previous work demonstrated that methylation ITH is associated with clinical outcome in patients with invasive ADC[15]. To assess the evolution of methylation ITH during the progression of lung precancers, we first calculated the "epiallele shift"[25], the combinatory difference in epiallele status in all captured loci (>60 reads), for each IPN specimen compared to paired normal lung tissues from the same patients. MIA and invasive ADC had more pronounced epiallele shifts than AAH or AIS did, indicating a higher level of methylation ITH in IPNs of later histologic stages (**Fig. 2A**). In addition, the abundance of eloci (loci with distinct epiallele shifts) progressively increased in later-stage IPNs ($p = 0.004567$, Kruskal–Wallis test) (**Fig. 2B**). When focusing on the 15 patients with multiple IPNs of different stages (IPNs of different stages with the identical genetic background and exposure history), the abundance of eloci was also higher in later-stage IPNs than in

early-stage IPNs in 13 of the 15 patients (**Supplementary Fig. 4**), further suggesting that later-stage IPNs have a higher level of methylation ITH.

To determine whether methylation ITH has the potential to impact transcriptional dynamics, we examined the relative distance of eloci to the nearest transcription start sites (TSSs). Interestingly, the vast majority of the eloci were much closer to TSSs in invasive ADC than in AAH, AIS, or MIA (**Supplementary Fig. 5**). To elucidate the potential impact of methylation ITH on chromatin remodeling, we performed Locus Overlap Analysis[26] of the genomic regions identified as eloci to evaluate the overlap between genomic regions with methylation ITH and genomic regions targeted by diverse histone posttranslational modifications previously profiled in lung cancer cell lines[27] and normal lung epithelial cells[28]. Compared with AAH or AIS, MIA and ADC had higher incidences of eloci significantly enriched in genomic regions occupied by H3K27me3, H3K9me3, and H3K9me2 (**Supplementary Table 5**), epigenetic modifications strongly associated with transcriptional repression[29-31], supporting the concept that epigenetic marks co-operate with DNA methylation alterations along the evolution of lung precancers[32,33].

Furthermore, we calculated the frequency of 16 combinatory DNA methylation patterns at four consecutive CpG sites covered by the same RRBS reads (termed “epipolymorphisms”) in each specimen. Again, later-stage IPNs had higher epiallele diversity than early-stage IPNs did (**Supplementary Fig. 6**). Moreover, ADCs had higher levels of epipolymorphism at eloci than MIA, AIS, or AAH did (**Fig. 2C**), indicating that the epigenetic states had undergone a greater extent of drifting in late-stage IPNs than in early-stage IPNs.

Genomic and methylation evolution occur in parallel during early lung carcinogenesis

To dissect the evolutionary relationship between the methylome and genome in lung ADC, we constructed phyloepigenetic trees for patients with more than four spatially separated specimens available. The overall structure of the methylation-based phyloepigenetic trees was similar to that of trees based on mutations of the same multi-regional specimens[7] (**Fig. 3A, Supplementary Fig. 7**). Furthermore, the genomic distance based on all somatic mutations was positively correlated with the methylation distance between any pair of samples from the same IPNs ($\rho = 0.7283$, $p < 2.2 \times 10^{-16}$, Spearman's rank correlation test) (**Fig. 3B**). Taken together, these results suggest that genomic and methylation evolution occurred in parallel during early carcinogenesis in this cohort of lung ADC. Interestingly, in patient C10, promoter hypermethylation of *TSC2*, a candidate TSG known to inhibit cell growth in lung[34,35], was identified in AIS specimens, whereas copy number loss was identified in AAH lesions from the same patient. Taken together, these data suggested convergent evolution, whereby the same genes or pathways are activated or inactivated by different mechanisms in different cancer cell clones during lung cancer development and progression.

Global hypomethylation was associated increased chromosomal instability in IPNs

As an essential chemical modification, the methylation states can directly impact the chromosomal structure and DNA mutagenesis. It has been well documented that global hypomethylation is associated with chromosomal instability (CIN) and increased mutational rates in cancers[36-38]. To further depict the interaction between genome and epigenome during early carcinogenesis of lung ADC, we sought to delineate the association between global hypomethylation and genomic changes of these IPNs. As

RRBS is designed to be overrepresented by CpG islands but low in repetitive DNA sequences[39], and therefore for suboptimal quantification of global hypomethylation, we used long interspersed transposable elements-1 (LINE-1), a widely used surrogate marker for global DNA methylation[40,41] to assess these IPNs. There was a significant decrease of LINE-1 methylation in AIS, MIA and ADC compared[42] to normal lung tissues or AAH (**Fig. 4A**) indicating increased global hypomethylation in IPNs of later histologic stages. Importantly, methylation level of LINE-1 was inversely associated with SCNA burden (**Fig. 4B**), AI burden (**Fig. 4C**) and total mutation burden (**Fig. 4D**) indicating global hypomethylation is associated with higher level of CIN. Interestingly, LINE-1 methylation status was also inversely associated with the proportion of clonal mutations in each specimen (**Fig. 4E**).

Global hypomethylation was associated with suppressed T cell infiltration

Cancer evolution is shaped by interaction between cancer cells and host factors, particularly the host immune response. Given T cells' central role in anti-tumor immune surveillance[24,43], we depicted the T cell infiltration in AIS, MIA and invasive ADC by de-convoluting RNA sequencing data from an independent dataset recently published[44]. Our analysis demonstrated a progressive increase of CD4+ T regulatory cells (Tregs) ($p<2.2e-16$) and progressive decrease of CD8+ T cells (although the difference was not significant, $p=0.1374$) from normal lung tissues to AIS/MIA and invasive ADC leading to significantly higher Treg/CD8 ratio in invasive ADC ($p<2.2e-16$) (**Supplementary Fig. 8A-C**). We next applied MethylCIBERSORT[45] to the RRBS data to delineate the T cell infiltration of IPNs in our cohort. Similarly, with progression from normal lung to AAH, AIS, MIA and ADC, there was a progressive increase in Tregs (**Fig 5A,** $p=0.009758$), while CD8+ T cell infiltration was lower in later-stage IPNs (although the difference did not reach statistical significance (**Fig 5B**, $p=0.1472$), leading to significantly higher

Treg/CD8 ratio in later-stage IPNs (**Fig. 5C**, p=0.00194). As increased Treg/CD8 ratio is known to associate with suppressed anti-tumor immune surveillance[46], these results indicated dynamic changes of anti-tumor immune response during the neoplastic progression from precancers to invasive lung ADC with potentially more suppressive immune microenvironment in later stage IPNs, consistent with our previous findings[47,48]. Interestingly, Treg/CD8 ratio was inversely correlated with LINE-1 methylation level (**Fig. 5D**) implying the potential association between global hypomethylation and immunosuppression.

DISCUSSION

The methylation landscape of invasive lung cancers has been studied extensively[49,50]. However, somatic methylation aberrations in precancers and preinvasive lung cancers are less defined, largely because of a lack of appropriate specimens, as surgery is not the standard of care for treating lung precancers. Using small panels of genes implicated in lung carcinogenesis, previous studies have demonstrated gradual change in DNA methylation in AAH, AIS, and ADC[51]. Leveraging a unique collection of resected IPNs of different stages, we for the first time revealed evolution of comprehensive DNA methylome as lung precancers progress to preinvasive lung ADC and eventually invasive lung ADC. Our results demonstrated progressive changes in both hypermethylation and hypomethylation along with neoplastic evolution from AAH to AIS, MIA, and ADC. Interestingly, compared to normal lung tissues, somatic hypermethylation alterations appeared to emerge as early as AAH, whereas hypomethylation only became obvious after AIS (**Fig. 1B**), implying that somatic hypermethylation may have preceded hypomethylation during early carcinogenesis of lung ADC.

Different cells within the same tumor can exhibit different molecular and phenotypic features, a phenomenon termed ITH. ITH may foster tumor evolution by providing diverse cell populations, and the dynamics of ITH architecture may evolve with neoplastic progression[15,52]. Methylation ITH has been observed in various malignancies, including colorectal cancer[53], lung cancer[15], and chronic lymphocytic leukemia[54]. High levels of methylation ITH have been reported to associate with inferior clinical outcomes[15,54,55]. In addition, methylation ITH has also been reported in precancers; for example, Merlo et al. reported that methylation ITH in Barrett esophagus, a precursor to esophageal ADC, was associated with the risk of malignant transformation[55]. In the current study, we provided the first evidence of methylation ITH in lung precancers and its dynamic changes during neoplastic progression. Later-stage IPNs demonstrated more complex methylation ITH than early-stage IPNs did, with more divergent epiallele shift and higher fraction of epipolymorphisms (**Fig. 2**). Importantly, eloci were significantly more abundant around TSSs in ADC than in AAH, AIS, or MIA. One plausible explanation is that although somatic methylation ITH may be stochastic during early lung carcinogenesis, some of the methylation aberrations (particularly those that are close to TSSs and potentially impact gene expression) may convey survival and/or growth advantages, resulting in selection of cells with higher densities of eloci around TSSs.

Cancer evolution is associated with the accumulation of genetic and epigenetic aberrations. Our previous work has revealed evidence that epigenetic evolution and genetic evolution take place in parallel in invasive lung ADC[15]. Our current study demonstrated similar phylogenetic patterns and correlated genetic and methylation distances in IPNs of different stages (**Fig. 3**), suggesting genetic alterations and DNA methylation also evolve in parallel during the early carcinogenesis of lung ADC. Meanwhile, promoter hypermethylation and copy number loss of TSG *TSC2* were

identified in two independent IPNs respectively from the same patient. These results were reminiscent of previous findings showing that distinct mutations of the same cancer genes were present in different regions of the same tumors[23,56,57] or in different primary tumors from the same patients[58], indicating convergent evolution. Although most genetic and methylation changes were occurring in parallel, and genetic events (e.g., copy number loss of TSGs) or methylation changes (e.g. promoter hypermethylation of TSGs) may independently offer proliferation or survival advantages to cells, these processes may be constrained around certain genes or pathways (e.g., inactivation of *TSC2* in the case of patient C10) that are essential to carcinogenesis in certain patients.

In addition to independently and/or cooperatively promoting cell expansion by affecting different genes or pathways, DNA methylation states can directly impact on vulnerability of DNA for genomic aberrations. It is well-known that global hypomethylation is associated with CIN[36] and increased rate of somatic mutations[38]. In the current cohort, global hypomethylation assessed by LINE-1, was associated with significantly increased mutation burden, SCNA burden as well as AI burden (**Fig. 4B-D**). Importantly, LINE-1 methylation level in AAH was similar to normal lung, but significantly decreased in AIS, MIA and ADC (**Fig. 4A**). Genomic analysis of the same cohort of IPNs in our previous study have revealed progressive increase in mutation burden from AAH to ADC, while SCNA only became obvious after stage of AIS and AI events were rare in AAH and AIS[7]. These data suggest that methylation aberrations have not only evolved in parallel with genomic aberrations, but may have also facilitated accumulating genomic aberrations that may have led to more drastic phenotypic changes in IPNs of later stages. Interestingly, higher-level of global hypomethylation was associated with higher proportion of clonal mutations (**Fig. 4E**). As the progression of lung precancers into invasive lung ADC predominantly follows a clonal sweeping model with selective

outgrowth of fit subclones[7], one plausible explanation for this association is that cells within the precancers with high level of global hypomethylation may be prone to accumulate genomic aberrations, which subsequently provide growth advantages to these cell clones to develop into major clones in invasive lung ADC.

We have previously reported that the immune microenvironment was suppressed in invasive lung cancers compared to preinvasive cancers or precancers[47,48,59]. We de-convoluted previously published transcriptomic data of an independent cohort of AIS/MIA and invasive ADC[44] and RRBS data of IPNs in the current cohort. Both analyses demonstrated higher Treg/CD8 ratio (**Fig. 5C and Supplemental Fig. 8C**) implying more suppressed T cell repertoire in later-stage diseases, in line with a concomitant immune profiling study on the same cohort of IPNs (Dejima et al, BioRix, 2020). These findings are consistent with the concept of immune-editing, whereby the immunogenicity of cancer cells evolves under the selection pressure from anti-tumor immune response, resulting in the emergence of immune-resistant cancer clone variants in later stage diseases[60]. Interestingly, increased global hypomethylation (lower LINE-1 methylation) was associated with higher Treg/CD8 ratio (**Fig. 5D**). Methylation aberrations may affect anti-tumor immune surveillance directly by regulating the expression of immune related genes[61] and/or potential neoantigens (mutated genes that can be recognized by T cells)[62] or indirectly via modifying chromosomal vulnerability for SCNA and mutations, both of which are well known to influence tumor immune microenvironment[63,64]. However, these impacts are complicated as many processes can affect anti-tumor immune surveillance positively and negatively. For example, high level of global hypomethylation may lead to a high SCNA burden known to associate with a cold tumor immune microenvironment; meanwhile global hypomethylation is also associated with increased mutation rate, which may facilitate recruitment of anti-tumor immune cells[65]. In the end,

selection of cancer cell clones under immune pressure is determined by the accumulative effects of these molecular aberrations and only the cells with the best combination of molecular features including aberrant methylation status, mutation burdens and SCNA will survive and develop into the dominant clones in invasive cancers.

Compared with invasive cancers, precancers and preinvasive cancers may have less complex molecular landscapes, as well as better preserved immune contextures, and thus may be easier to eradicate. There has been increasing enthusiasm toward moving interventions successfully applied to metastatic cancers to early stage cancers and even precancers, a concept called interception[66,67]. For example, we have launched the IMPRINT-Lung clinical trial (NCT03634241), in which patients with high-risk IPNs (many of which may be AAH or AIS) are treated with immune checkpoint inhibitors. In the current study, we demonstrated that methylation aberrations are less complex in precancers and preinvasive lung cancers than in invasive cancers. Therefore, novel therapeutic agents that can modulate methylation by targeting aberrant methylations and potentially reprogram the immune microenvironment[68] may also have potential in treating precancers and preinvasive cancers to prevent invasive lung cancers.

To the best of our knowledge, this is the first study on the evolution of genome-wide DNA methylation during the early carcinogenesis of lung ADC. Due to the scarcity of adequate materials for comprehensive profiling of IPNs, our study has several inevitable limitations. First, the sample size was relatively small, so our data has to be interpreted with caution. Second, because we did not have enough materials for transcriptomic profiling, the biological impact of the methylation changes remains to be determined; however, as a substantial number of DMRs were associated with TSSs, promoter binding sites, and TF binding sites, many of these changes could potentially regulate the

transcriptome of the IPNs. Third, all patients in our cohort were from China or Japan, and whether our findings can be applied to patients in other ethnic groups remains to be investigated. Fourth, the follow-up times for all patients in this study were relatively short, so we were not able to investigate the impact of these methylation changes on recurrence or survival. Finally, the resected specimens in this study could only provide a single molecular snapshot of the evolutionary process of IPNs. Whether all AAHs will evolve into AIS, MIA, and ADCs, and whether all ADCs evolve from AAH, is still unknown. Although we used a multi-regional sequencing approach to de-convolute the methylation evolution, deciphering the temporal evolution of the methylome during neoplastic progression will require specimens obtained over the course of disease progression. Clinical trials collecting longitudinal biopsy specimens, such as IMPRINT-Lung (NCT03634241), may provide such opportunities in the future.

METHODS

Sample acquisition

A total of 53 resected pulmonary nodules and paired normal lung tissues from 39 patients treated at Nagasaki University Hospital or Zhejiang Cancer Hospital between 2014 and 2017 were used in the study. None of the patients received chemotherapy or radiotherapy before surgery. 29 lung nodules from 15 patients had multiregional specimens for spatial heterogeneity assessment (**Supplementary Table 1**). Whole-exome sequencing (WES) data was available for all specimens (EGAS00001004960)[7]. Written informed consent was obtained from all patients. The study was approved by the Institutional Review Boards of MD Anderson Cancer Center, Nagasaki University Graduate School of Biomedical Sciences, and Zhejiang Cancer Hospital.

DNA methylation profiling by RRBS and data processing

DNA was extracted using the QIAamp DNA FFPE Tissue Kit (QIAGEN), and 200ng–1μg of DNA was subjected to RRBS for genome-wide DNA methylation profiling as previously described[69,70]. Raw RRBS data were processed using the Bismark pipeline[71]. Briefly, TrimGalore v. 0.4.3 was used to trim the Illumina adapter sequences (a minimum of 5 bp in a read was required to overlap with the adapter sequence); FastQC v. 0.11.7 was used for quality control; and then bowtie2 v. 2.2.3 was used to align the trimmed reads to the GRCh37 assembly of the human genome. The DNA methylation levels for individual CpGs were calculated using methyKit[72]. Methylated reads (containing Cs) and unmethylated reads (containing Ts) at each cytosine site were counted, and the percentage of methylated reads among total reads covering the corresponding cytosine was calculated to quantify the DNA methylome for each sample at the single-base resolution. The CpG sites, DMRs, and loci of known genes, as well as genomic features, including CpG islands, were annotated using the R package "ChIPSeeker"[73] and the "genomation" toolkit[74], which is based on the "TxDb.Hsapiens.UCSC.hg19.knownGene" annotation database and the UCSC Genome Browser CpG islands table. To avoid bias, we kept CpG sites mapped to the autosome and removed CpG sites overlapping with the single-nucleotide polymorphism positions in dbSNP137. DNA methylation was analyzed at either a single-CpG resolution or at genomic region bins, in which DNA methylation values were averaged across 5-kb regions. Promoter methylation was calculated as the averaged DNA methylation values based on GENCODE promoter regions (i.e., 1 kb upstream to 500 bp downstream of the annotated TSS).

Comparison of methylation profiles between different specimens

To assess the DNA methylome profiles of distinct IPNs of different pathological stages and examine the heterogeneity between IPN samples, we first aggregated the DNA methylation levels of 5-kb tiling regions across the genome in each sample by retaining

only the CpG sites with ≥10 sequencing reads. We then applied principal component analysis to identify global DNA methylation patterns between samples. To evaluate consistent clusters of these DNA methylation profiles, we performed unsupervised hierarchical agglomerative analysis of CpG sites covered by ≥50 reads across all samples; these reads were based on single CpG methylation calls without any binning. To evaluate the correlation between overall DNA methylation in all samples from each patient, we used a pairwise approach to compare distance and similarity matrices on the basis for all CpGs with a coverage of ≥10 reads.

Differential DNA methylation analysis

Differentially methylated regions (DMRs) encompassing the differentially methylated CpGs between paired disease and normal tissue samples were identified using a triangular kernel to smooth the number of methylated reads and total number of reads by applying the "noise filter" function in the "DMR caller" package[75]. The differentially methylated CpG positions of paired samples were identified by mapping CpG sites to DMRs; only CpG sites covered by ≥10 reads in paired samples were included.

Motif identification and prediction of TF binding sites

The de novo methylated DNA motifs in IPNs of each stage were identified by mEpigram, which discovers motifs by using position-specific weight matrices from the k-mers that are most enriched in the positive sequences compared with the negative sequences as "seeds" and extending the motifs in both directions[76]. The Tomtom tool from the MEME suite was used to select significantly enriched methylated DNA motifs to the database of known transcript factors (HOCOMOCO_v11)[77].

Estimation of DNA methylation ITH

To estimate DNA methylation ITH, we applied “methclone” to identify epigenetic loci whose distributions of epigenetic allele (“epiallele”)[25] clonality differed between paired tumor and normal samples by quantifying the degree to which the compositions of epialleles at given loci in the tumor samples were distinct from those in the normal tissue samples. An epiallele was defined by setting 60 reads in four consecutive CpG sites as the threshold to consider the epigenetic allele composition of the locus. We then calculated the differences in epiallele entropy between each IPN sample and its matched normal tissue sample. Loci with combinatorial entropy changes below –60 between each IPN sample and its paired normal tissue sample were defined as epigenetic shift loci (termed “eloci”). To reduce the bias due to the different coverage for each sample, we then calculated relative epiallele shifts (i.e., the normalized number of eloci) by dividing the number of eloci by the total number of assessed loci in each sample and then multiplying that ratio by the average number of total loci across all samples[25]. To evaluate the dynamics of methylation change in paired tumor and normal epigenomes, we also assessed epigenetic polymorphism (“epipolymorphism”) to measure the epiallelic diversity in each IPN sample as described previously[78] by calculating the frequency of each specific epiallele from multiple stochastic changes in the frequencies of many epialleles. We calculated 16 epiallele statuses (0000, 0100, 0010, 0001, 1000, 1100, 0110, 0011, 0101, 1010, 1001, 0111, 1011, 1101, 1110, and 1111, where 1 represents a methylated CpG site, and 0 represents an unmethylated CpG site).

Locus Overlap Analysis

We applied Locus Overlap Analysis (LOLA) [26] to the genomic regions identified as eloci ($\Delta S < -60$) in all samples of each stage to evaluate the overlap between genomic regions with methylation ITH and chromatin marks. The genomic regions of all loci with ≥ 60 reads were used as background genomic regions, and the selected genomic regions

were mapped to a compendium of publicly available histone mark profiles, including CTCF, H2AZ, H3K4me1, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K27me3, H3K27ac, H3K36me3, H3K79me2, H4K20me1 in the A549 lung adenocarcinoma cell line (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeBroadHistone) and H3K27me3, H3K4me3, H3K9me2, H3K9ac, and CTCF in an immortalized human bronchial epithelial cell line (BEAS-2B; GSE56053)[28]. P-values in the enrichment analyses were calculated using a one-sided Fisher exact test. Adjustment for multiple testing was performed using the Benjamini–Yekutieli method.

Construction of phylogenetic trees

To construct epiphylogenetic trees from the methylation data, we used promoter CpG sites (≥50 reads per CpG site selected for all samples from each patient) with the most variable methylation values (mean absolute deviation >10%) shared by all samples, including a normal tissue sample used as the tree root, to build a Euclidean distance matrix. We built the phylogeny trees by applying a neighbor-joining algorithm from the "ape" package to independently infer phylogenetic relationships between IPN specimens for each patient from the mutation and methylation profiles. To assess the phylogenetic similarity between the genetic and epigenetic profiles, we employed an independent but parallel distance matrix construction from the mutation and methylation profiles for each patient and calculated the Spearman correlation coefficient for all pairwise samples grouped by lesion.

Estimation of global hypomethylation

To determine global methylation levels, we chose CpG sites (covered by ≥10 aligned reads) mapped to evolutionarily young subfamilies of LINE-1 repeat elements (L1HS and L1PA). LINE-1 family annotation were obtained from the Repeat-Masker of the UCSC

genome browser[79]. We averaged the methylation values of the chosen CpG sites, to represent the global methylation level of each sample.

Deconvolution of T cell profiles

To derive tumor-infiltrating T cell subtypes from transcriptomic data previously published[44], processed RNAseq dataset (normalized and log2 transformed) comprise of sequencing of 197 normal lung tissues, 98 AIS/MIA and 99 ADC samples were retrieved from EGAS00001004006. ImmuCellAI[80] was applied to infer immune cell components for each sample.

To derive T cell repertoire from RRBS data, we first obtained a reference methylation signature by retrieving the methylation calls of whole-genome bisulfite datasets (WGBS) from the BLUEPRINT epigenome project, including those for samples of regulatory T cells (EGAX00001343016/EGAX00001236257 in EGAD00001002492) and CD8+ T cells (EGAX00001195937/EGAX00001195943 in EGAD00001002486). Only promoter CpG sites that were covered by ≥5 reads in all samples were retained and binned with a 50-bp window by mean methylation values, then only 50-bp binned regions that overlapped with 50-bp binned regions of CpG sites covered by the RRBS methylation profiles of preneoplastic lesions in all samples were used for reference signature extraction by non-negative matrix factorization (NMF) implemented in the "MethylCIBERSORT" package. We then performed DNA methylation deconvolution (mean methylation by 50-bp bin) using the aforementioned BLUEPRINT signature and the CIBERSORT webserver (https://cibersort.stanford.edu). We performed T cell deconvolution in relative mode, running 100 permutations with quantile normalization disabled[81]. The resulting cellular fraction tables were used to compare samples of different pathological stages from the same patient. The averaged values of immune cell

infiltration for each sample inferred by whole-genome bisulfite analysis from two patients independently were used to calculate Treg/CD8 ratio.

Statistical analysis

Violin plots were created using the “geom_violin” function in the R statistical software package “ggplot2” to represent data point density along the y axis, and the “stat_summary” function was used to calculate the median as the center point. Differences in DMR numbers, eloci numbers, and immune cell infiltration between lesions of different stages were assessed using the Kruskal–Wallis chi-square test. We used a two-sided Pearson correlation coefficient to compare methylation profiles between two samples and between groups of samples of different stages. We used a two-sided Spearman correlation coefficient to determine the extent to which distance matrices were correlated with DNA methylation profiles and somatic mutation profiles.

Data availability

The raw sequencing data are available from the European Bioinformatics Institute European Genome–phenome Archive (EGA) (accession number: EGASXXX0000XX) through controlled access. To protect patient privacy, interested researchers need to apply via a data access committee (DAC), which will grant all reasonable requests by bona fide researchers.

AUTHOR CONTRIBUTIONS

X.H. and J.J.Z. conceived the study and wrote the manuscript. J.J.Z. and P.A.F. jointly supervised and financially supported the study. X.H. performed all bioinformatics and statistical data analyses in consultation with M.E.R., J.P.I., L.H.W., J.J.L., and Y.L.; J.F.J. supervised pathological assessments and the preparation of multi-region samples. L.Y.,

J.F.K., and H.P. collected resected specimens and clinical data. C.W.C., L.D.L., C.B, and R.Z.C. prepared DNA samples. M.C.G. and B.W.C. performed radiological assessment. M.E.R. supervised RRBS profiling. X.H., M.E.R., R.Z.C., A.R., N.M., L.H.W., J.C.Z., J.H.Z., A.A.V., J.V.H., I.I.W., H.K., P.S., J.C.Z., M.M., J.P.I, D.S., P.A.F., and J.J.Z. interpreted the data. All authors reviewed the manuscript.

COMPETING OF INTERESTS

Dr. Zhang reports research funding from Merck, Johnson and Johnson, and consultant fees from BMS, Johnson and Johnson, AstraZeneca, Geneplus, OrigMed, Innovent outside the submitted work. The other authors declare no competing financial interests.

ACKNOWLEDGMENTS

This study was supported by the MD Anderson Khalifa Scholar Award, the National Cancer Institute of the National Institute of Health Research Project Grant (R01CA234629-01), the AACR-Johnson & Johnson Lung Cancer Innovation Science Grant (18-90-52-ZHAN), the MD Anderson Physician Scientist Program, the MD Anderson Lung Cancer Moon Shot Program, Sabin Family Foundation Award, Duncan Family Institute Cancer Prevention Research Seed Funding Program. We thank MD Anderson Cancer Center's Epigenomics Profiling Core and its Science Park Next-Generation Sequencing Core (supported by CPRIT Core Facility Support Award #RP120348) for performing RRBS profiling. We thank Sally Boyd, Jinzhen Chen, and Rong Yao, Stan Bujnowski for providing technical support for high-performance cluster resource; we thank Joe Munch in Scientific Publications in MD Anderson's Research Medical Library for editing the manuscript.

References

1. Kadara H, Scheet P, Wistuba, II, Spira AE. Early Events in the Molecular Pathogenesis of Lung Cancer. *Cancer Prev Res (Phila).* 2016;9(7):518-527.
2. Chen Z, Fillmore CM, Hammerman PS, Kim CF, Wong KK. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nat Rev Cancer.* 2014;14(8):535-546.
3. Lee HJ, Lee CH, Jeong YJ, et al. IASLC/ATS/ERS International Multidisciplinary Classification of Lung Adenocarcinoma: novel concepts and radiologic implications. *J Thorac Imaging.* 2012;27(6):340-353.
4. Hiramatsu M, Inagaki T, Inagaki T, et al. Pulmonary ground-glass opacity (GGO) lesions-large size and a history of lung cancer are risk factors for growth. *J Thorac Oncol.* 2008;3(11):1245-1250.
5. Chen D, Dai C, Kadeer X, Mao R, Chen Y, Chen C. New horizons in surgical treatment of ground-glass nodules of the lung: experience and controversies. *Ther Clin Risk Manag.* 2018;14:203-211.
6. Maiga AW, Deppen SA, Massion PP, et al. Communication About the Probability of Cancer in Indeterminate Pulmonary Nodules. *JAMA Surg.* 2018;153(4):353-357.
7. Hu X, Fujimoto J, Ying L, et al. Multi-region exome sequencing reveals genomic evolution from preneoplasia to lung adenocarcinoma. *Nat Commun.* 2019;10(1):2978.
8. Baylin SB. DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol.* 2005;2 Suppl 1:S4-11.
9. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. *Nat Rev Cancer.* 2011;11(10):726-734.
10. Belinsky SA. Unmasking the lung cancer epigenome. *Annu Rev Physiol.* 2015;77:453-474.
11. Teneng I, Tellez CS, Picchi MA, et al. Global identification of genes targeted by DNMT3b for epigenetic silencing in lung cancer. *Oncogene.* 2015;34(5):621-630.
12. Jones PA, Baylin SB. The epigenomics of cancer. *Cell.* 2007;128(4):683-692.
13. Widschwendter M, Fiegl H, Egle D, et al. Epigenetic stem cell signature in cancer. *Nat Genet.* 2007;39(2):157-158.
14. Reed MD, Tellez CS, Grimes MJ, et al. Aerosolised 5-azacytidine suppresses tumour growth and reprogrammes the epigenome in an orthotopic lung cancer model. *Br J Cancer.* 2013;109(7):1775-1781.
15. Quek K, Li J, Estecio M, et al. DNA methylation intratumor heterogeneity in localized lung adenocarcinomas. *Oncotarget.* 2017;8(13):21994-22002.
16. Kerr KM, Galler JS, Hagen JA, Laird PW, Laird-Offringa IA. The role of DNA methylation in the development and progression of lung adenocarcinoma. *Dis Markers.* 2007;23(1-2):5-30.
17. Yin Y, Morgunova E, Jolma A, et al. Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science.* 2017;356(6337).
18. Kribelbauer JF, Lu XJ, Rohs R, Mann RS, Bussemaker HJ. Toward a Mechanistic Understanding of DNA Methylation Readout by Transcription Factors. *J Mol Biol.* 2019.
19. Casey SC, Baylot V, Felsher DW. The MYC oncogene is a global regulator of the immune response. *Blood.* 2018;131(18):2007-2015.
20. Wils LJ, Bijlsma MF. Epigenetic regulation of the Hedgehog and Wnt pathways in cancer. *Crit Rev Oncol Hematol.* 2018;121:23-44.
21. Yu T, Chen X, Zhang W, et al. KLF4 regulates adult lung tumor-initiating cells and represses K-Ras-mediated lung cancer. *Cell Death Differ.* 2016;23(2):207-215.

22. Wu C, Zhu W, Qian J, et al. WT1 promotes invasion of NSCLC via suppression of CDH1. *J Thorac Oncol.* 2013;8(9):1163-1169.
23. Zhang J, Fujimoto J, Zhang J, et al. Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science.* 2014;346(6206):256-259.
24. Reuben A, Zhang J, Chiou SH, et al. Comprehensive T cell repertoire characterization of non-small cell lung cancer. *Nat Commun.* 2020;11(1):603.
25. Li S, Garrett-Bakelman F, Perl AE, et al. Dynamic evolution of clonal epialleles revealed by methclone. *Genome Biol.* 2014;15(9):472.
26. Sheffield NC, Bock C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics.* 2016;32(4):587-589.
27. Consortium EP. An integrated encyclopedia of DNA elements in the human genome. *Nature.* 2012;489(7414):57-74.
28. Jose CC, Xu B, Jagannathan L, et al. Epigenetic dysregulation by nickel through repressive chromatin domain disruption. *Proc Natl Acad Sci U S A.* 2014;111(40):14631-14636.
29. Lienert F, Mohn F, Tiwari VK, et al. Genomic prevalence of heterochromatic H3K9me2 and transcription do not discriminate pluripotent from terminally differentiated cells. *PLoS Genet.* 2011;7(6):e1002090.
30. Kondo Y, Shen L, Cheng AS, et al. Gene silencing in cancer by histone H3 lysine 27 trimethylation independent of promoter DNA methylation. *Nat Genet.* 2008;40(6):741-750.
31. Hon GC, Hawkins RD, Caballero OL, et al. Global DNA hypomethylation coupled to repressive chromatin domain formation and gene silencing in breast cancer. *Genome Res.* 2012;22(2):246-258.
32. Ohm JE, McGarvey KM, Yu X, et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat Genet.* 2007;39(2):237-242.
33. Schlesinger Y, Straussman R, Keshet I, et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat Genet.* 2007;39(2):232-236.
34. Astrinidis A, Cash TP, Hunter DS, Walker CL, Chernoff J, Henske EP. Tuberin, the tuberous sclerosis complex 2 tumor suppressor gene product, regulates Rho activation, cell adhesion and migration. *Oncogene.* 2002;21(55):8470-8476.
35. Carsillo T, Astrinidis A, Henske EP. Mutations in the tuberous sclerosis complex gene TSC2 are a cause of sporadic pulmonary lymphangioleiomyomatosis. *Proc Natl Acad Sci U S A.* 2000;97(11):6085-6090.
36. Eden A, Gaudet F, Waghmare A, Jaenisch R. Chromosomal instability and tumors promoted by DNA hypomethylation. *Science.* 2003;300(5618):455.
37. Lee ST, Wiemels JL. Genome-wide CpG island methylation and intergenic demethylation propensities vary among different tumor sites. *Nucleic Acids Res.* 2016;44(3):1105-1117.
38. Chen RZ, Pettersson U, Beard C, Jackson-Grusby L, Jaenisch R. DNA hypomethylation leads to elevated mutation rates. *Nature.* 1998;395(6697):89-93.
39. Sun Z, Cunningham J, Slager S, Kocher JP. Base resolution methylome profiling: considerations in platform selection, data preprocessing and analysis. *Epigenomics.* 2015;7(5):813-828.

40. Saito K, Kawakami K, Matsumoto I, Oda M, Watanabe G, Minamoto T. Long interspersed nuclear element 1 hypomethylation is a marker of poor prognosis in stage IA non-small cell lung cancer. *Clin Cancer Res.* 2010;16(8):2418-2426.
41. Zheng Y, Joyce BT, Liu L, et al. Prediction of genome-wide DNA methylation in repetitive elements. *Nucleic Acids Res.* 2017;45(15):8697-8711.
42. Venkatesan S, Birkbak NJ, Swanton C. Constraints in cancer evolution. *Biochem Soc Trans.* 2017;45(1):1-13.
43. Reuben A, Gittelman R, Gao J, et al. TCR Repertoire Intratumor Heterogeneity in Localized Lung Adenocarcinomas: An Association with Predicted Neoantigen Heterogeneity and Postsurgical Recurrence. *Cancer Discov.* 2017;7(10):1088-1097.
44. Chen H, Carrot-Zhang J, Zhao Y, et al. Genomic and immune profiling of pre-invasive lung adenocarcinoma. *Nat Commun.* 2019;10(1):5472.
45. Chakravarthy A, Furness A, Joshi K, et al. Pan-cancer deconvolution of tumour composition using DNA methylation. *Nat Commun.* 2018;9(1):3220.
46. Peng GL, Li L, Guo YW, et al. CD8(+) cytotoxic and FoxP3(+) regulatory T lymphocytes serve as prognostic factors in breast cancer. *Am J Transl Res.* 2019;11(8):5039-5053.
47. Hu X, Fujimoto J, Ying L, et al. Investigation of Genomic and TCR Repertoire Evolution of AAH, AIS, MIA to Invasive Lung Adenocarcinoma by Multiregion Exome and TCR Sequencing. *Journal of Thoracic Oncology.* 2017;12(11):S2102-S2102.
48. Chen RZ, Fujimoto J, Reuben A, et al. T cell repertoire evolution from the normal lung to invasive lung adenocarcinoma. *Cancer Research.* 2018;78(13).
49. Horie M, Kaczkowski B, Ohshima M, et al. Integrative CAGE and DNA Methylation Profiling Identify Epigenetically Regulated Genes in NSCLC. *Mol Cancer Res.* 2017;15(10):1354-1365.
50. Cancer Genome Atlas Research N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* 2014;511(7511):543-550.
51. Selamat SA, Galler JS, Joshi AD, et al. DNA methylation changes in atypical adenomatous hyperplasia, adenocarcinoma in situ, and lung adenocarcinoma. *PLoS One.* 2011;6(6):e21443.
52. Prasetyanti PR, Medema JP. Intra-tumor heterogeneity from a cancer stem cell perspective. *Mol Cancer.* 2017;16(1):41.
53. Kreso A, O'Brien CA, van Galen P, et al. Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. *Science.* 2013;339(6119):543-548.
54. Landau DA, Clement K, Ziller MJ, et al. Locally disordered methylation forms the basis of intratumor methylome variation in chronic lymphocytic leukemia. *Cancer Cell.* 2014;26(6):813-825.
55. Merlo LM, Shah NA, Li X, et al. A comprehensive survey of clonal diversity measures in Barrett's esophagus as biomarkers of progression to esophageal adenocarcinoma. *Cancer prevention research.* 2010;3(11):1388-1397.
56. Jamal-Hanjani M, Wilson GA, McGranahan N, et al. Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med.* 2017;376(22):2109-2121.
57. consortium TRR. TRACERx Renal: tracking renal cancer evolution through therapy. *Nat Rev Urol.* 2017;14(10):575-576.
58. Liu Y, Zhang J, Li L, et al. Genomic heterogeneity of multiple synchronous lung cancer. *Nat Commun.* 2016;7:13200.

59. Sivakumar S, Lucas FAS, McDowell TL, et al. Genomic Landscape of Atypical Adenomatous Hyperplasia Reveals Divergent Modes to Lung Adenocarcinoma. *Cancer Res.* 2017;77(22):6119-6130.
60. Kim R, Emi M, Tanabe K. Cancer immunoediting from immune surveillance to immune escape. *Immunology.* 2007;121(1):1-14.
61. Liu M, Zhou J, Chen Z, Cheng AS. Understanding the epigenetic regulation of tumours and their microenvironments: opportunities and problems for epigenetic therapy. *J Pathol.* 2017;241(1):10-24.
62. Serrano A, Castro-Vega I, Redondo M. Role of gene methylation in antitumor immune response: implication for tumor progression. *Cancers (Basel).* 2011;3(2):1672-1690.
63. Bakhoum SF, Cantley LC. The Multifaceted Role of Chromosomal Instability in Cancer and Its Microenvironment. *Cell.* 2018;174(6):1347-1360.
64. Porta-Pardo E, Godzik A. Mutation Drivers of Immunological Responses to Cancer. *Cancer Immunol Res.* 2016;4(9):789-798.
65. Jung H, Kim HS, Kim JY, et al. DNA methylation loss promotes immune evasion of tumours with high mutation and copy number load. *Nat Commun.* 2019;10(1):4278.
66. Blackburn EH. Cancer interception. *Cancer Prev Res (Phila).* 2011;4(6):787-792.
67. Gold KA, Kim ES, Lee JJ, Wistuba, II, Farhangfar CJ, Hong WK. The BATTLE to personalize lung cancer prevention through reverse migration. *Cancer Prev Res (Phila).* 2011;4(7):962-972.
68. Forde PM, Brahmer JR, Kelly RJ. New strategies in lung cancer: epigenetic therapy for non-small cell lung cancer. *Clin Cancer Res.* 2014;20(9):2244-2248.
69. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. *Nat Protoc.* 2011;6(4):468-481.
70. Meissner A, Mikkelsen TS, Gu H, et al. Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature.* 2008;454(7205):766-770.
71. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics.* 2011;27(11):1571-1572.
72. Akalin A, Kormaksson M, Li S, et al. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol.* 2012;13(10):R87.
73. Yu G, Wang LG, He QY. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics.* 2015;31(14):2382-2383.
74. Akalin A, Franke V, Vlahovicek K, Mason CE, Schubeler D. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics.* 2015;31(7):1127-1129.
75. Catoni M, Tsang JM, Greco AP, Zabet NR. DMRcaller: a versatile R/Bioconductor package for detection and visualization of differentially methylated regions in CpG and non-CpG contexts. *Nucleic Acids Res.* 2018;46(19):e114.
76. Ngo V, Wang M, Wang W. Finding de novo methylated DNA motifs. *Bioinformatics.* 2019;35(18):3287-3293.
77. Bailey TL, Boden M, Buske FA, et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* 2009;37(Web Server issue):W202-208.
78. Landan G, Cohen NM, Mukamel Z, et al. Epigenetic polymorphism and the stochastic formation of differentially methylated regions in normal and cancerous tissues. *Nat Genet.* 2012;44(11):1207-1214.
79. Rosenbloom KR, Armstrong J, Barber GP, et al. The UCSC Genome Browser database: 2015 update. *Nucleic Acids Res.* 2015;43(Database issue):D670-681.

80. Miao YR, Zhang Q, Lei Q, et al. ImmuCellAI: A Unique Method for Comprehensive T-Cell Subsets Abundance Prediction and its Application in Cancer Immunotherapy. *Adv Sci (Weinh).* 2020;7(7):1902880.
81. Chen B, Khodadoust MS, Liu CL, Newman AM, Alizadeh AA. Profiling Tumor Infiltrating Immune Cells with CIBERSORT. *Methods Mol Biol.* 2018;1711:243-259.

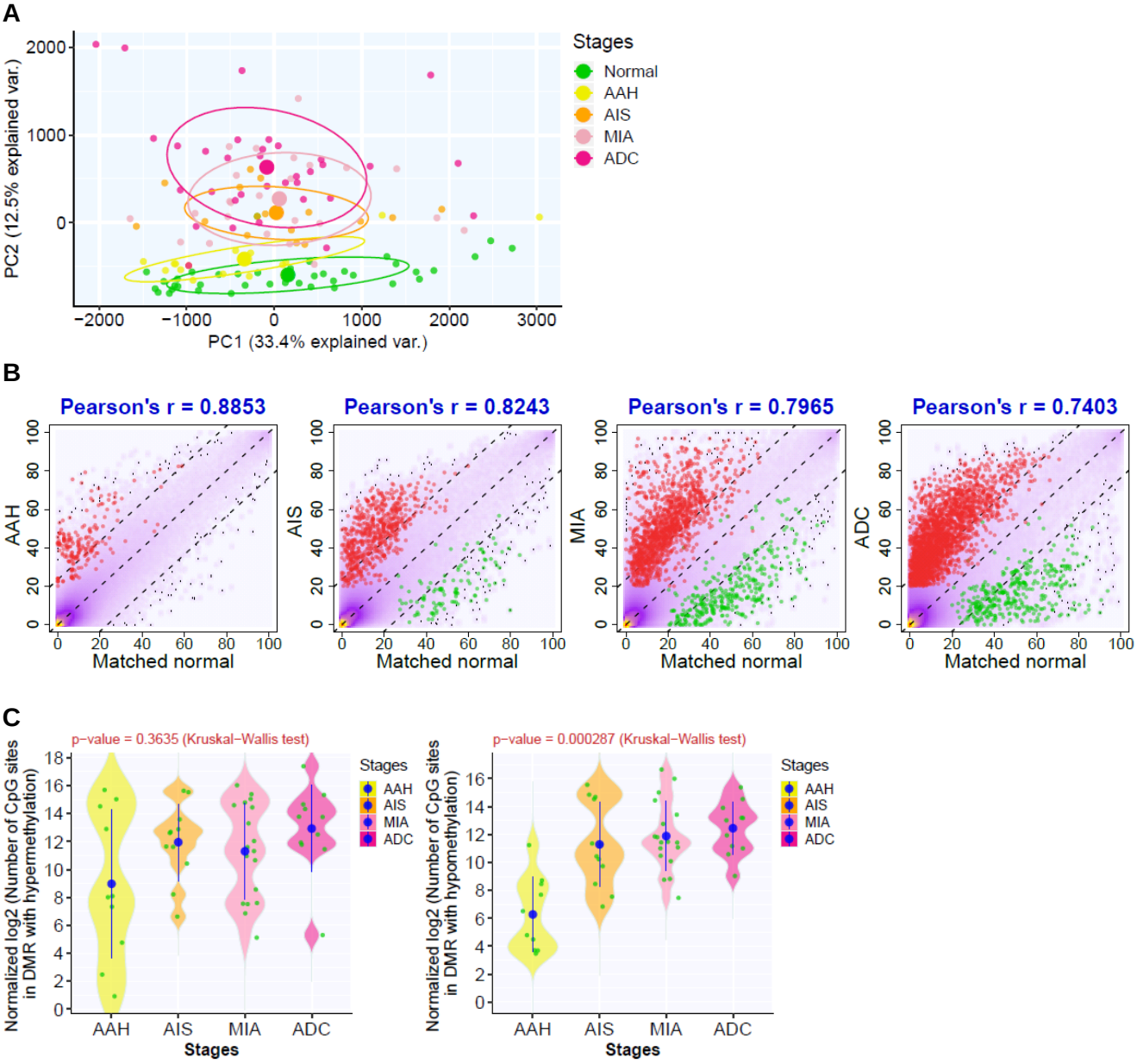


Figure 1. Somatic DNA methylation aberrations increase with progression of precancers. (**A**) Principal component analysis of the methylation profiles of the average methylation values in 39,148 5-kb tiling regions (common in all samples) composed of autosomal, non-polymorphism CpG sites supported by at least 10X read coverage in IPNs of different stages. (**B**) Overall genome-wide DNA methylation correlation between IPNs of different stages and matched normal tissue from the same patients. The yellow-purple clouded dots in the smooth scatter plot represent the CpG sites covered by IPN specimens (n = 14 for AAH, n = 11 for AIS, n = 18 for MIA, n = 10 for ADC) and paired normal lung. Pearson correlation coefficient *r* values are shown on the top. P-values are < $p < 2.2 \times 10^{-16}$ in all 4 histologic stages. The red dots represent CpG sites with hypermethylation in IPNs (within DMRs of methylation gain in IPNs and methylation ≤20% of CpG sites in the corresponding normal lung) and the green dots represent CpG sites with hypomethylation (within DMRs of methylation loss in IPNs and methylation ≥20% of CpG sites in the corresponding normal lung). (**C**) The number of CpG sites overlapping with DMRs showing hypermethylation (left) or hypomethylation (right) in IPNs of different

histologic stages. The green dots represent the mean numbers of CpG sites overlapping with DMRs in each IPN. The blue dots represent the mean numbers of CpG sites overlapping with DMRs in IPNs of different histologic stages. Differences among all stages were assessed using the Kruskal-Wallis H test.

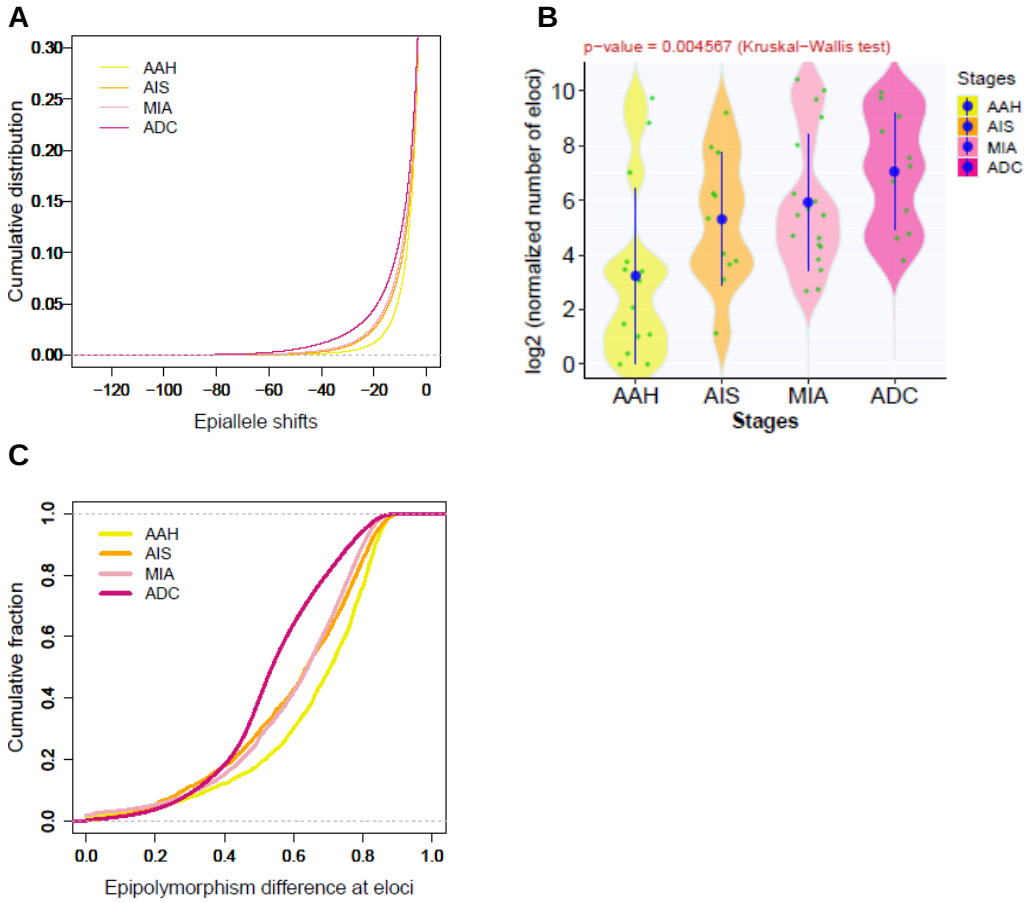
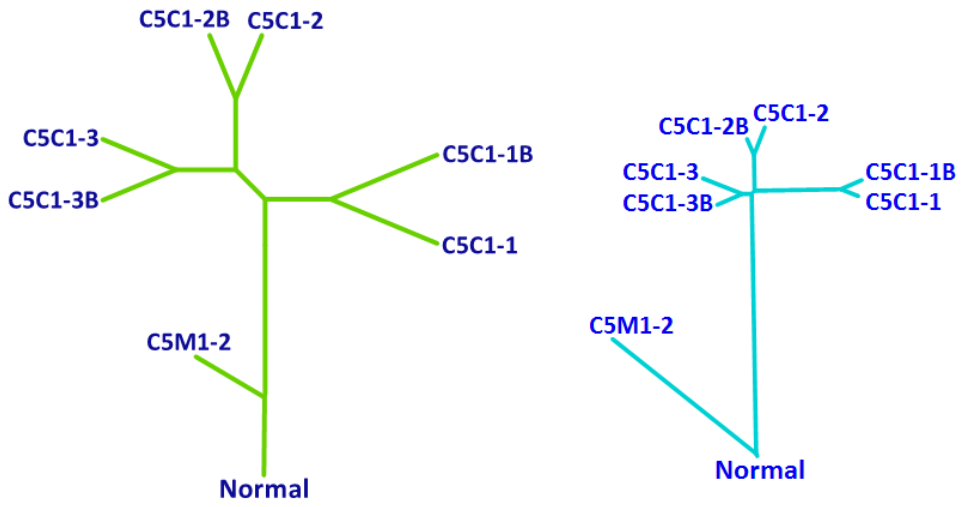


Figure 2. Increased DNA methylation ITH in later-stage IPNs. (**A**) Cumulative distribution curves of epiallele shifts of the DNA methylome in AAH, AIS, MIA, and ADC specimens compared to normal lung. (**B**) The number of log2-transformed eloci. Each green dot represents the average number of eloci in each IPN. The blue dots represent the mean numbers of eloci in the IPN of each histologic stage. Error bars represent 95% confidence intervals. Differences among all stages were assessed using the Kruskal-Wallis H test. (**C**) The cumulative distribution curves of epipolymorphism difference for loci with significant epiallele shifts (ΔS<-60) in AAH, AIS, MIA, and ADC compared with their paired normal lung tissue. X-axis denotes the difference of epipolymorphism between each IPN specimen and paired normal lung at each locus. Y-axis denotes normalized cumulative fraction of epipolymorphism difference from all IPN specimens of each stage.

A



B

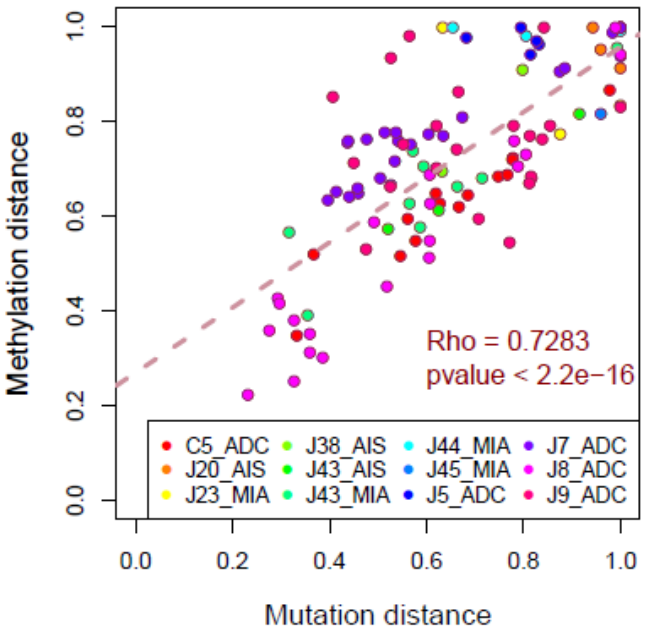


Figure 3. The evolutionary relationship between genomic and methylation landscape. (**A**) Example phylogenetic trees based on mutations (right) and methylation values (left) in patient C5. The length of each branch indicates the similarity of mutational or methylation profiles between any pair of two spatially separated tumor specimens from patient C5. (**B**) Correlation of genetic distance (Hamming distance based on all mutations) and methylation distance (Euclidean distance based on methylation values of all CpG sites) between different spatially separated regions from the same IPNs assessed by Spearman correlation analysis. Each dot represents the normalized distance between each pair of specimens from the same IPN.

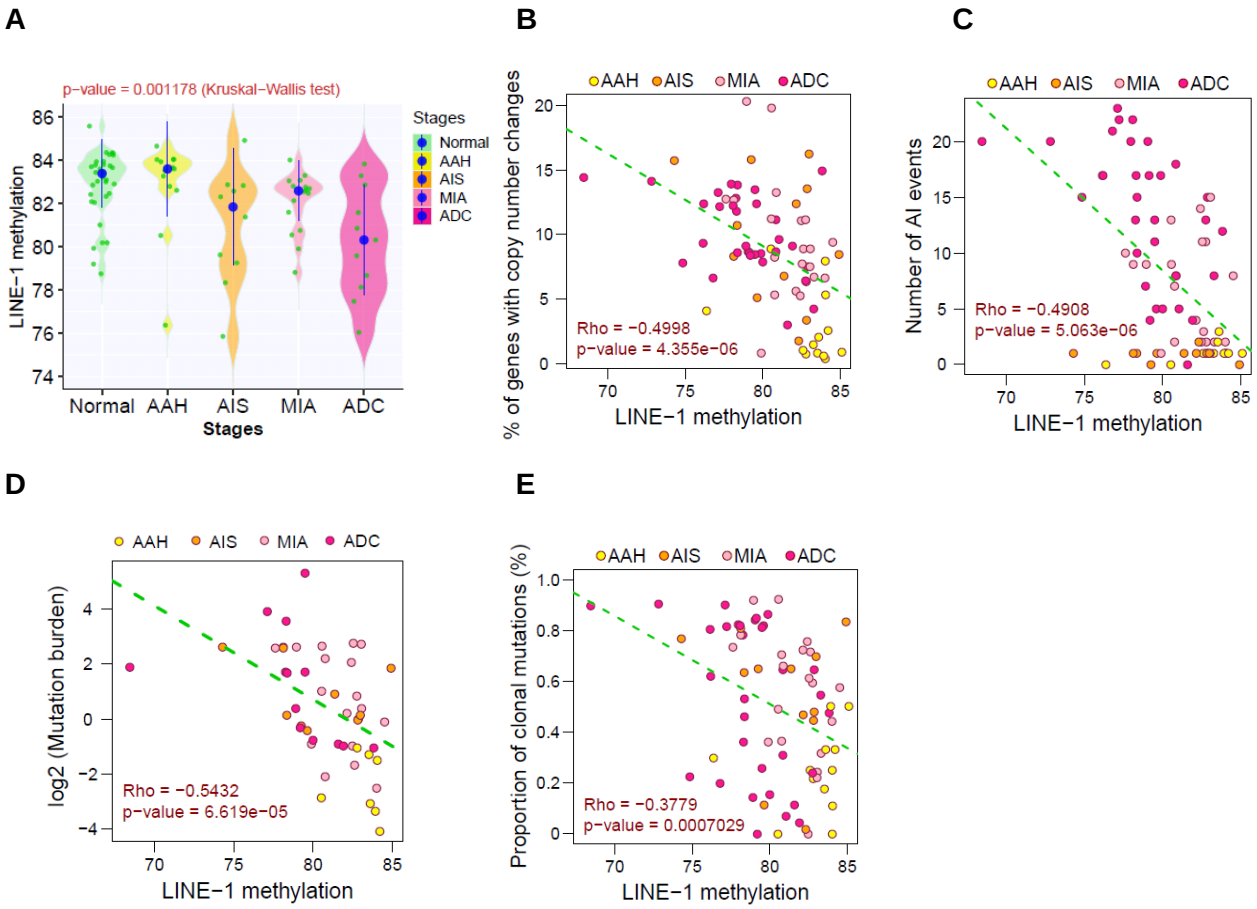


Figure 4. Correlation of LINE-1 methylation with genomic features in IPNs of different stages. (**A**) LINE-1 methylation level in IPNs of different stages. Each green dot represents each IPN specimen and the blue dots represent the mean. Error bars indicate 95% confidence intervals. The differences among all stages were assessed using the Kruskal-Wallis H test. Correlation between LINE-1 methylation levels and percent of genes with copy number changes (**B**), number of events with allelic imbalance (AI) (**C**), mutational burden (log2 transformed) (**D**), proportion of clonal mutations (**E**), assessed by Spearman correlation analysis. Each dot represents each IPN specimen.

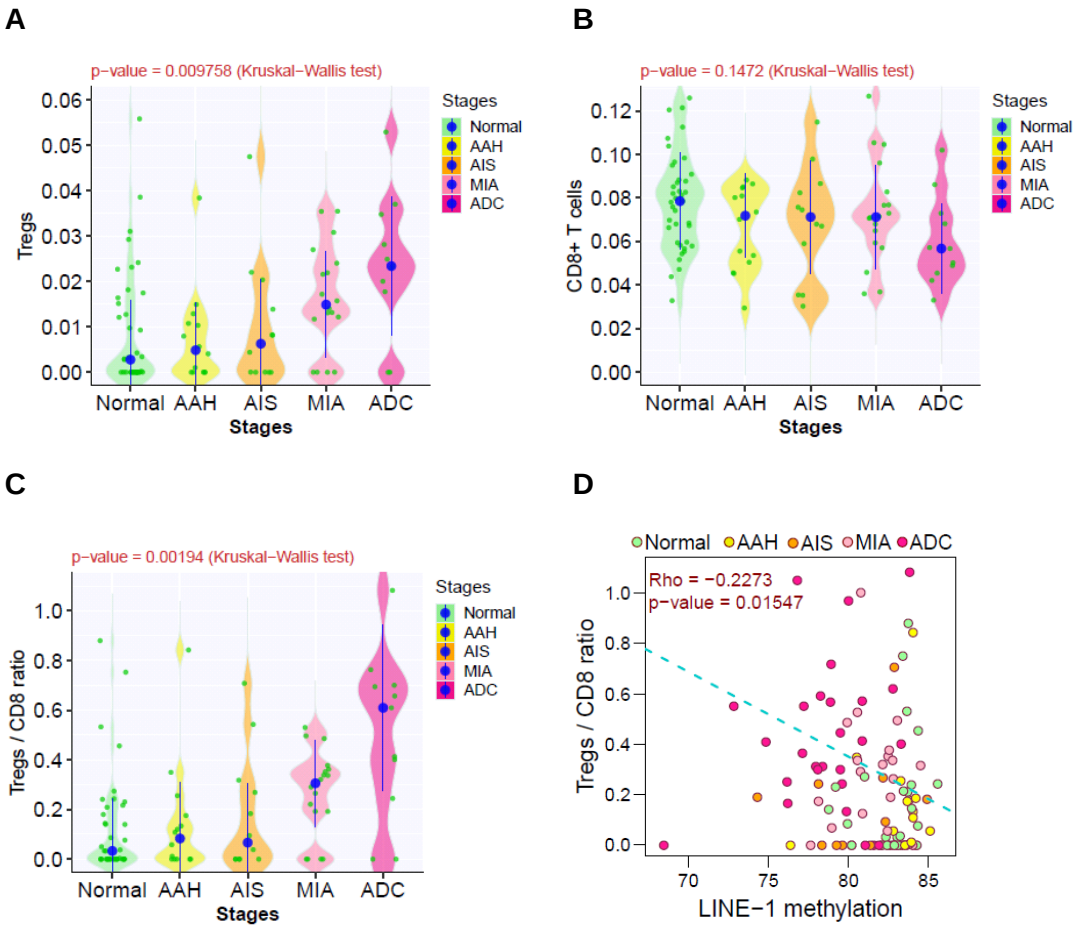
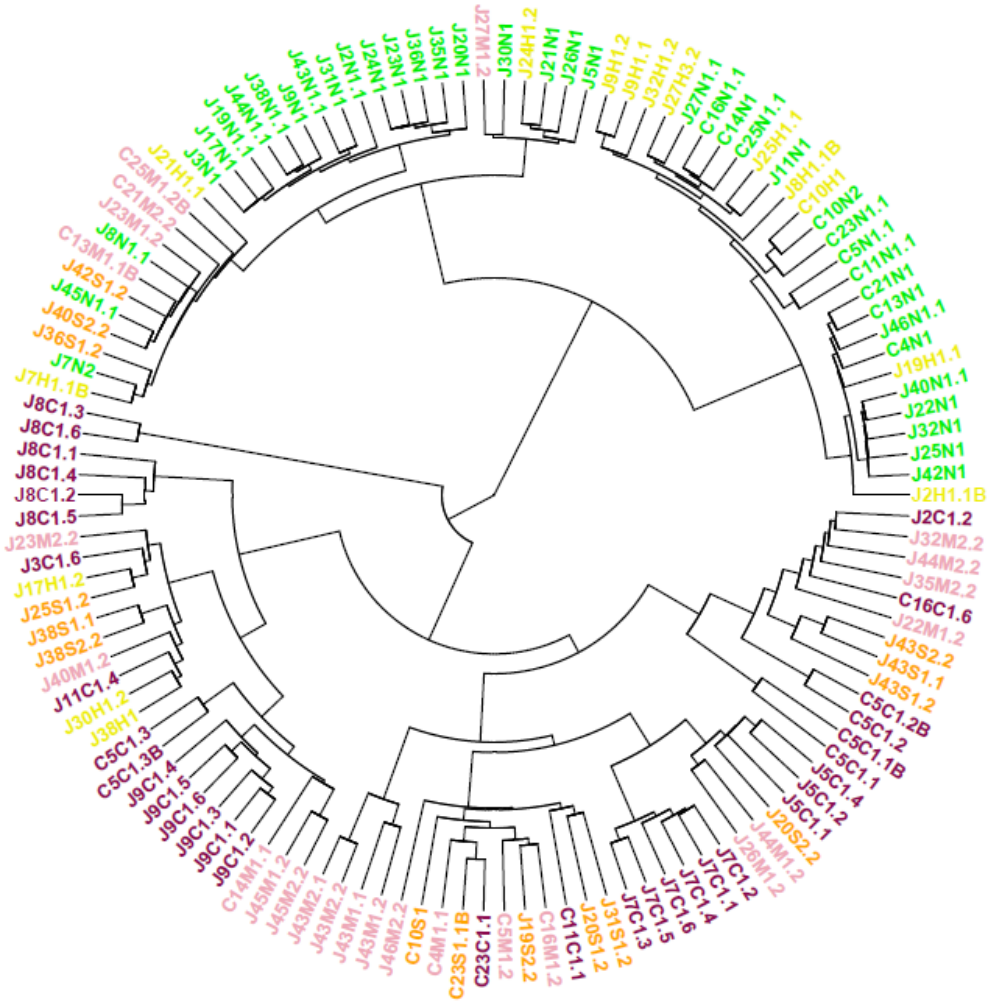
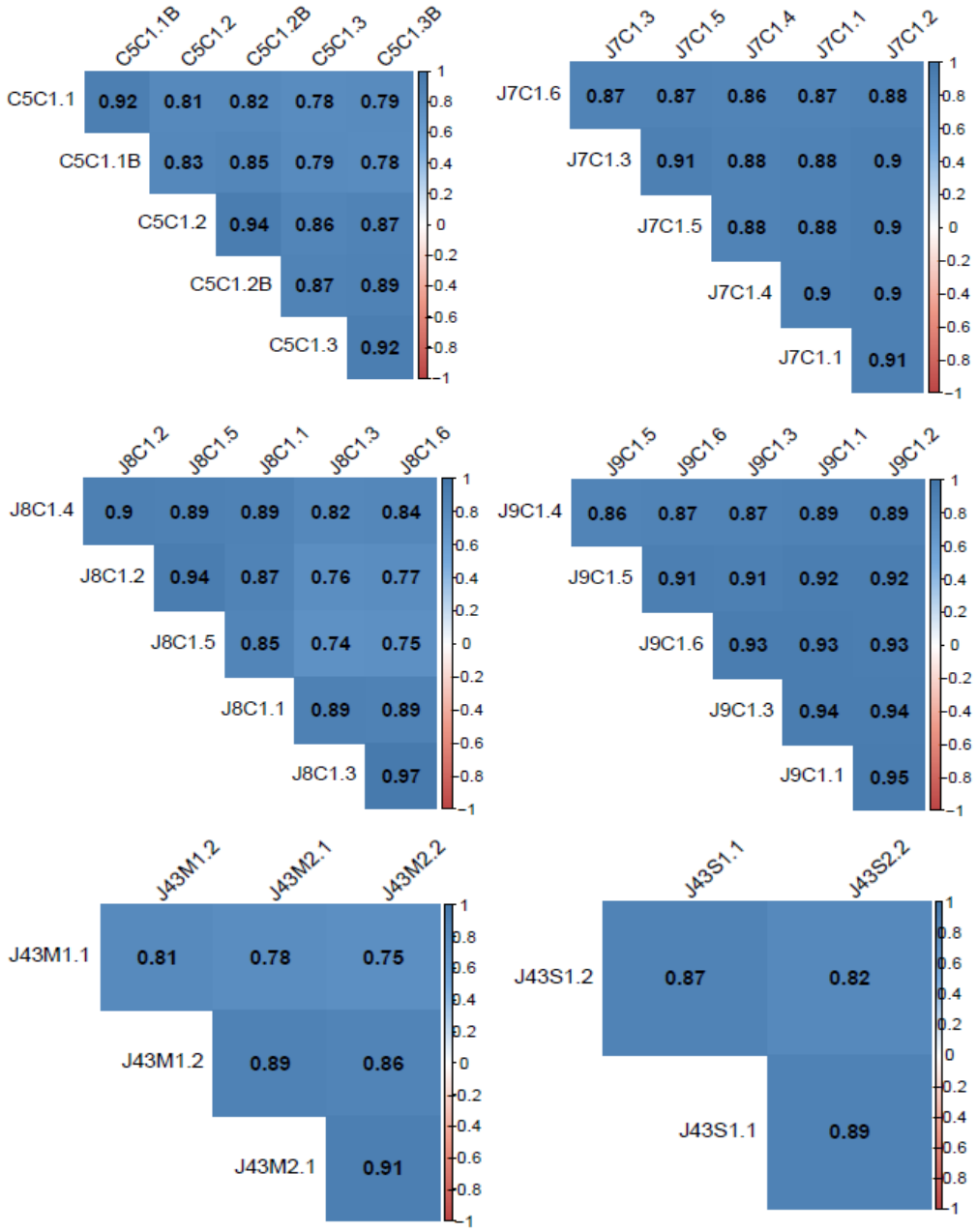


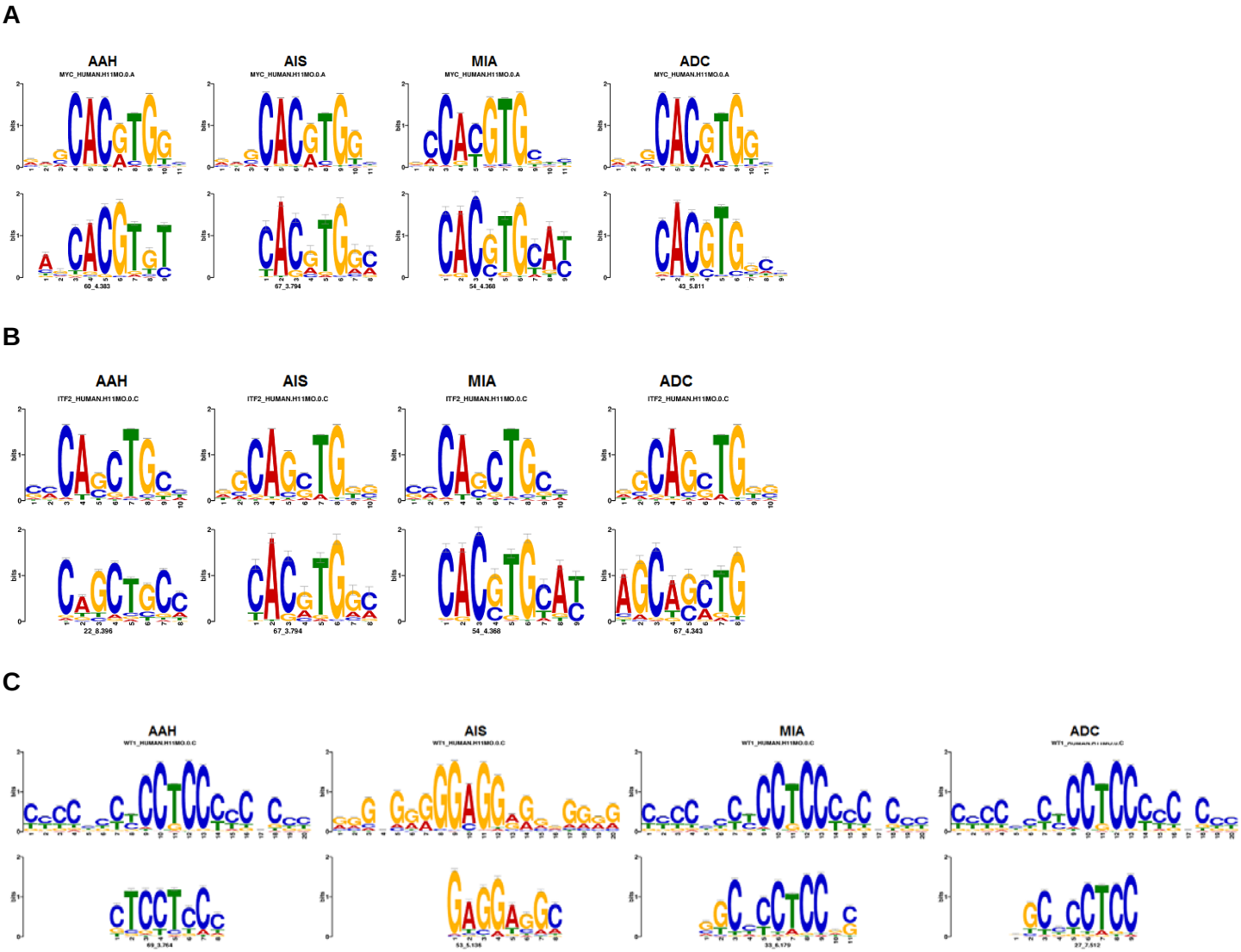
Figure 5. Dynamic changes of T cell infiltration in IPNs of different stages. The immune cell fraction of T regulatory cells (Tregs) (**A**), CD8+ T-cells (**B**), Treg/CD8 ratio (**C**) in IPNs of different stages. Each green dot represents each specimen and the blue dots represent the means. Error bars indicate 95% confidence intervals. The differences among all stages were assessed using the Kruskal-Wallis H test. Infiltration of T lymphocytes was inferred by MethylCIBERSORT. (**D**) Correlation between LINE-1 methylation and Treg/CD8 ratio assessed by Spearman correlation analysis. Each dot represents each specimen.



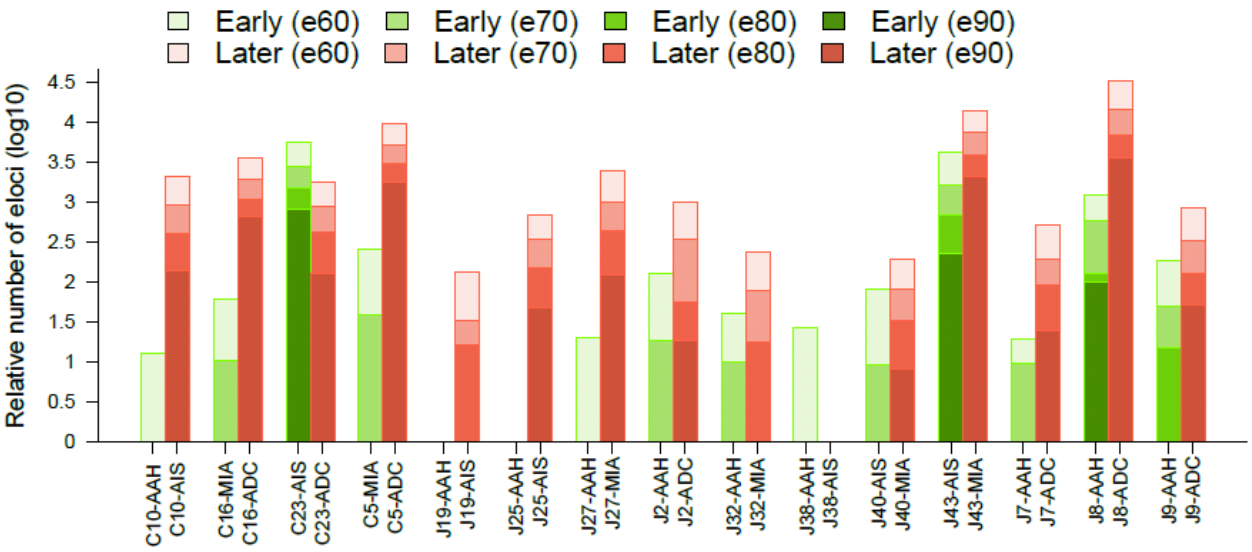
Supplementary Fig. 1. Unsupervised hierarchical clustering of promoter DNA methylation. Unsupervised hierarchical clustering was performed using 2,261 CpG sites commonly shared across all samples with minimum coverage of 50 reads per CpG site with a median absolute deviation (MAD) of >50. The colors of the sample names denote samples originating from normal tissue (green), AAH (yellow), AIS (orange), MIA (pink), and ADC (rose).



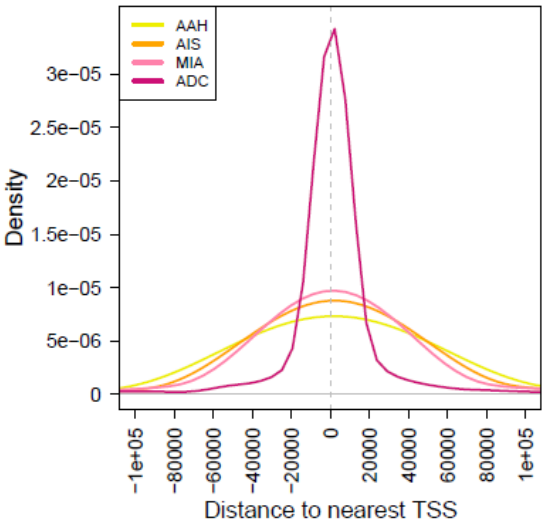
Supplementary Fig. 2. Comparison of global DNA methylation patterns between different regions from the same IPNs. The heat-maps display the Pearson correlation coefficients for pairwise comparisons of all samples from each IPN.



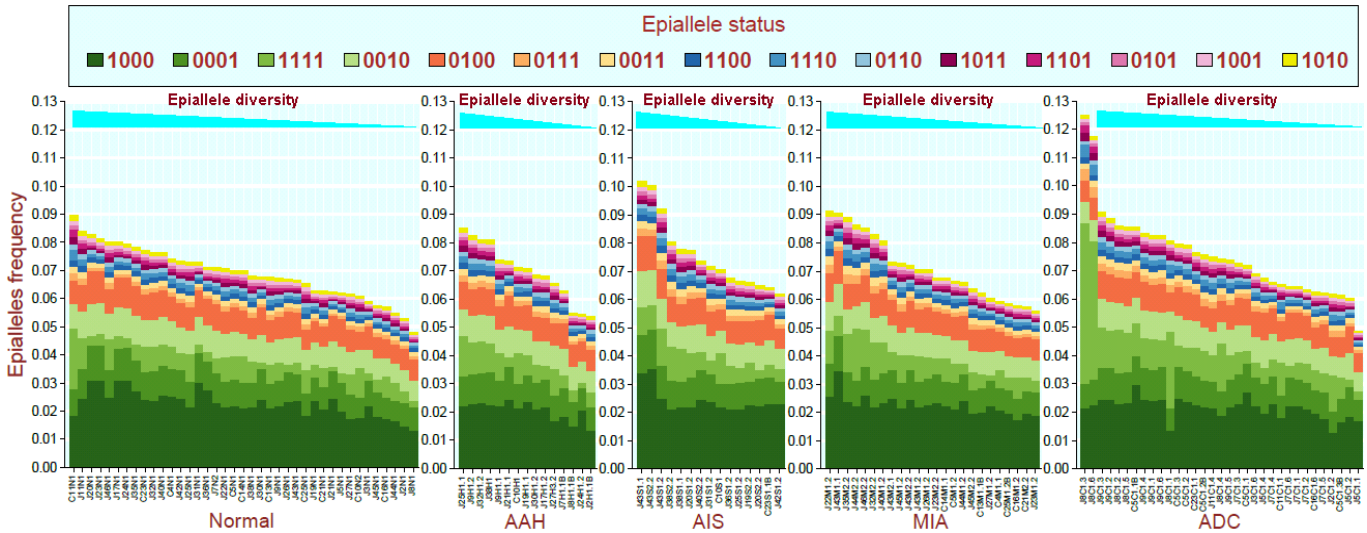
Supplementary Fig. 3. Example motifs associated with transcription factor (TF) binding sites enriched at DMRs for *MYC* (A), *ITF2* (B) and *WT1* (C). Top: The motifs associated with DNA binding sites of known TFs. Bottom: Matched DNA motifs significantly enriched in DMRs combined from all samples of each stage.



Supplementary Fig. 4. Comparison of relative numbers of eloci in IPNs of different stages within the same patients. Each bar represents the relative number (log10-transformed) of eloci composed of 4 adjacent CpG sites for each IPN. Pairwise comparison of an early-stage IPN and a later-stage IPN from the same patient was performed using four different combinatorial entropy difference cutoffs (e90: -90, e80: -80, e70: -70, and e60: -60). Early stages: green, later stage: brown.

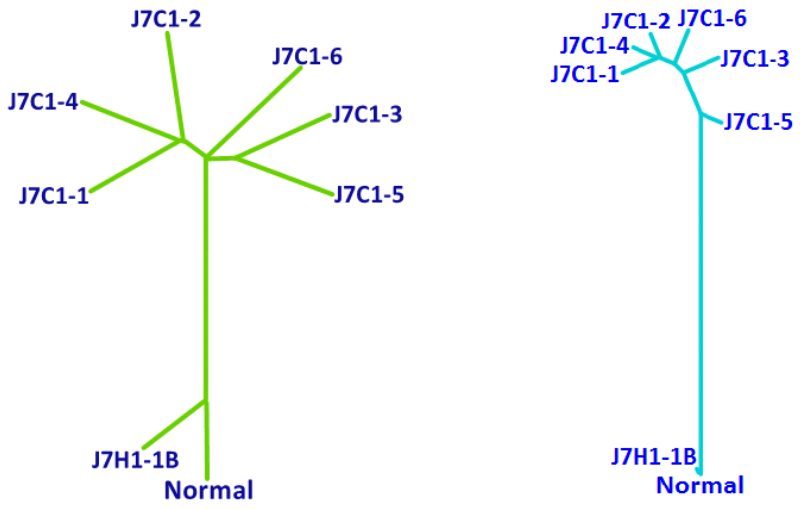


Supplementary Fig. 5. The distances of eloci to nearest transcription start sites (TSSs) in IPNs of different stages. The density plot shows the distances of eloci (ΔS<-60) to the nearest TSSs all AAH, AIS, MIA, and ADC samples.

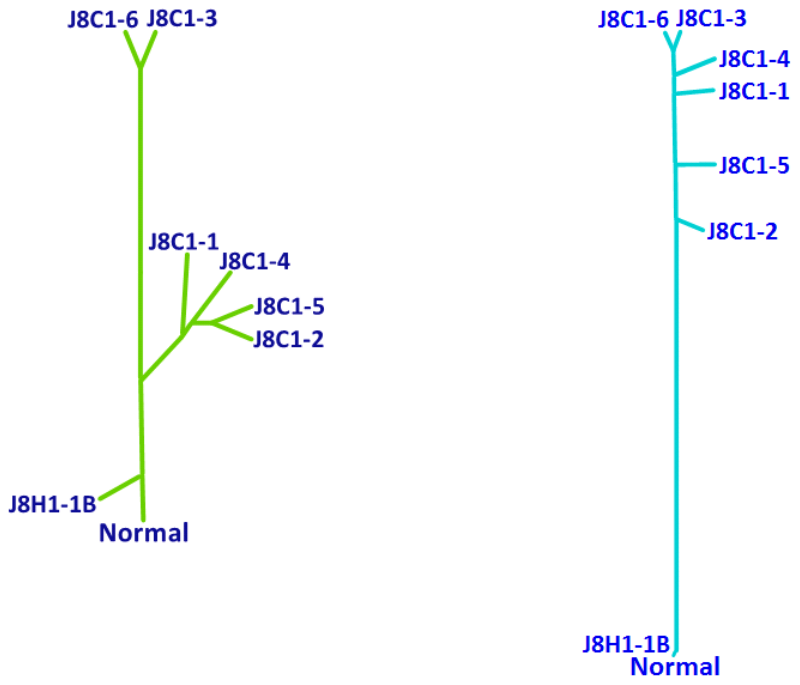


Supplementary Fig. 6. The distribution of epigenetic status. The epiallele frequencies across autosomal regions for all IPN specimens of each histologic stage. Different combinations of four consecutive CpG sites at epiallele loci are color-coded. The 0s indicate unmethylated CpG sites and the 1s indicate methylated CpG sites. The “0000”, accounting for ~80%-90% of epialleles is not displayed.

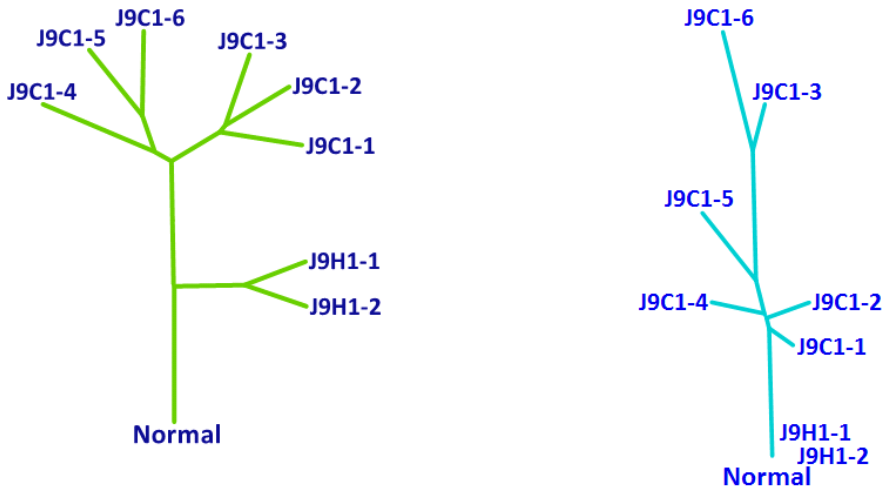
A


J7C1-2
J7C1-6
J7C1-4
J7C1-3
J7C1-1
J7C1-5
J7H1-1B
Normal
J7C1-2
J7C1-6
J7C1-4
J7C1-3
J7C1-1
J7C1-5
J7H1-1B
Normal

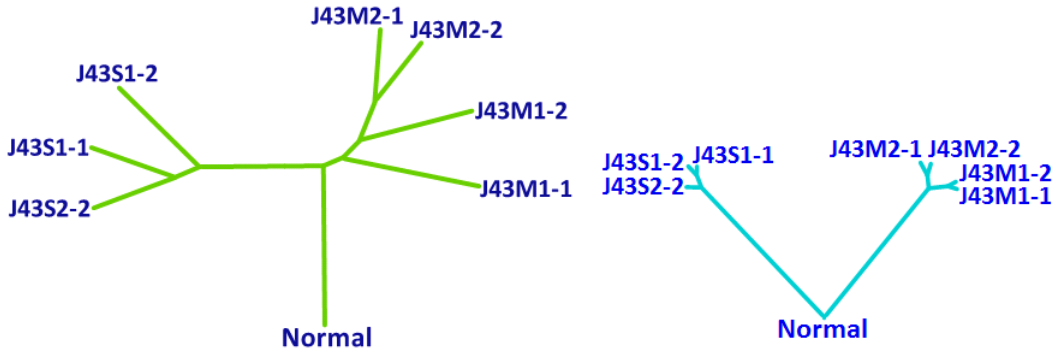
B


J8C1-6
J8C1-3
J8C1-1
J8C1-4
J8C1-5
J8C1-2
J8H1-1B
Normal
J8C1-6
J8C1-3
J8C1-4
J8C1-1
J8C1-5
J8C1-2
J8H1-1B
Normal

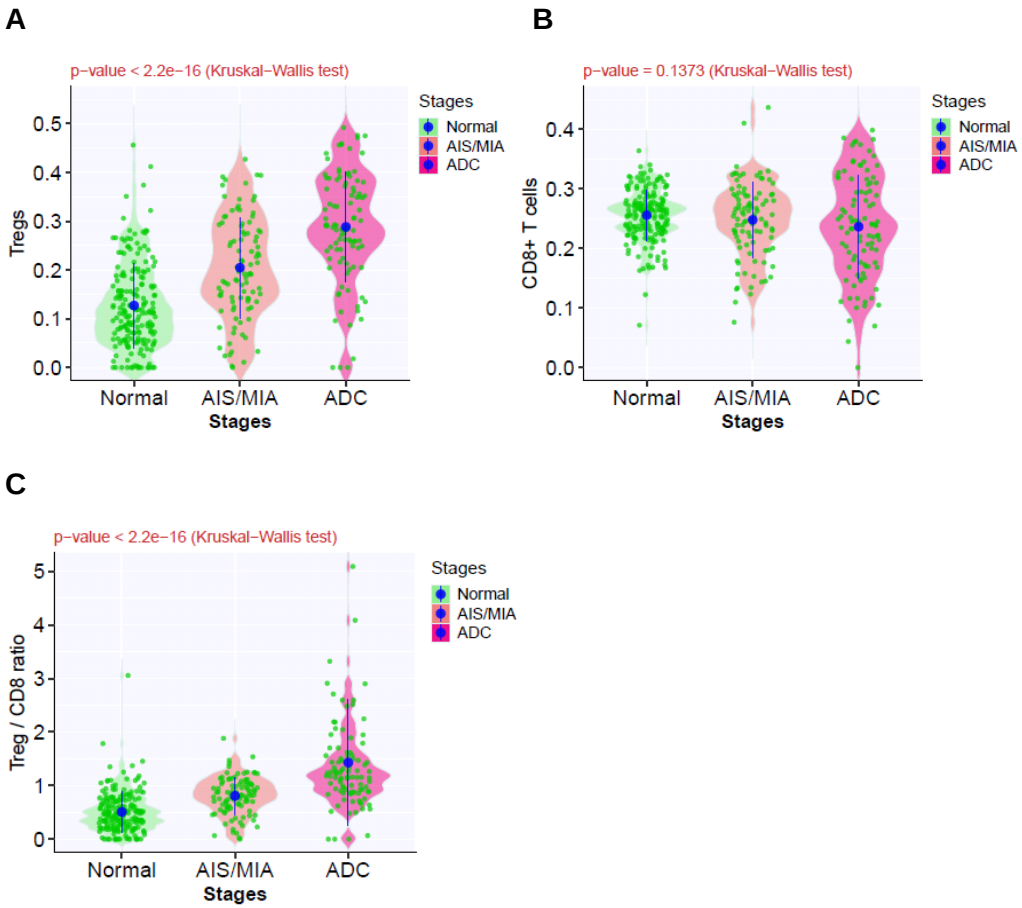
C



D



Supplementary Fig. 7. Genomic and methylation evolution in patients with multifocal IPNs. Phylogenetic trees based on mutations (right) and methylation values (left) in patients with multifocal IPNs. The length of each branch indicates the similarity of mutational or methylation profiles between any pair of two spatially separated specimens.



Supplementary Fig. 8. Validation for dynamic changes of T cell infiltration in IPNs of different stages. The immune cell fraction of T regulatory cells (**A**), CD8 T-cells (**B**), Treg/CD8 ratio (**C**) in normal lung, AIS/MIA and invasive ADC inferred by deconvolution of transcriptomic data from a previously published cohort using ImmuCellAI. Each green dot represents each specimen and the blue dots represent the means. Error bars indicate 95% confidence intervals. The differences among all stages were assessed using the Kruskal-Wallis H test.