https://doi.org/10.1038/s41586-019-1330-0

Immune evasion before tumour invasion in early lung squamous carcinogenesis

Céline Mascaux[1,2,3,4,14,15,18]*, Mihaela Angelova[5,6,7,8,16,18], Angela Vasaturo[5,6,7,8], Jennifer Beane[2], Kahkeshan Hijazi[2], Geraldine Anthoine[1], Bénédicte Buttard[5,6,7,8], Françoise Rothe[9], Karen Willard-Gallo[10], Annick Haller[11,17], Vincent Ninane[12], Arsène Burny[13], Jean-Paul Sculier[1], Avi Spira[2] & Jérôme Galon[5,6,7,8]*

Early detection and treatment are critical for improving the outcome of patients with cancer[1]. Understanding the largely uncharted biology of carcinogenesis requires deciphering molecular processes in premalignant lesions, and revealing the determinants of the intralesional immune reaction during cancer development. The adaptive immune response within tumours has previously been shown to be strongest at the earliest stage of carcinoma[2,3]. Here we show that immune activation and immune escape occur before tumour invasion, and reveal the relevant immune biomarkers of the pre-invasive stages of carcinogenesis in the lung. We used gene-expression profiling and multispectral imaging to analyse a dataset of 9 morphological stages of the development of lung squamous cell carcinoma, which includes 122 well-annotated biopsies from 77 patients. We identified evolutionary trajectories of cancer and immune pathways that comprise (1) a linear increase in proliferation and DNA repair from normal to cancerous tissue; (2) a transitory increase of metabolism and early immune sensing, through the activation of resident immune cells, in low-grade pre-invasive lesions; (3) the activation of immune responses and immune escape through immune checkpoints and suppressive interleukins from high-grade pre-invasive lesions; and, ultimately, (4) the activation of the epithelial–mesenchymal transition in the invasive stage of cancer. We propose that carcinogenesis in the lung involves a dynamic co-evolution of pre-invasive bronchial cells and the immune response. These findings highlight the need to develop immune biomarkers for early detection as well as immunotherapy-based chemopreventive approaches for individuals who are at high risk of developing lung cancer.

Despite developments in targeted therapies and immunotherapy, advanced lung cancer remains incurable[4]. Estimates suggest that early diagnosis and treatment could prevent more than 70,000 deaths from lung cancer in the United States per year[1]. The Nelson trials have recently shown that volume computed-tomography screening could reduce lung cancer mortality by 26% in men and 39–61% in women[5]. Beyond early detection, cancer prevention can considerably reduce the incidence of cancer[6]. Understanding the underlying mechanisms of carcinogenesis in the lung and the role of the microenvironment in early lesions may pave the way for personalized immunotherapy or other kinds of therapy for the prevention and interception of cancer[7]. Invasive lung squamous cell carcinoma (SCC) in smokers is preceded by a range of consecutive developmental stages[8], which makes it a convenient model for mechanistically studying the early evolution of cancer. Thus far, the rarity of pre-invasive lesion collections has limited our knowledge of their molecular and immune profiles[9]. Using gene-expression profiling and multispectral imaging, we sought to locate and time the changes in pre-invasive lesions and their microenvironment during the successive steps in the carcinogenesis of lung SCC.

We examined a dataset—comprising 122 carefully annotated biopsies from 77 patients—of 9 morphological stages of the carcinogenesis of lung SCC (stages 0–8) (Fig. 1a, Extended Data Fig. 1, Supplementary Tables 1, 2). Using gene-expression profiling, we first identified 4,734 genes that are associated with the 9 histological stages of development (linear mixed-effects model, false-discovery rate (FDR) < 0.001). Four distinct and successive molecular steps of progression were discerned by semi-supervised hierarchical clustering of the selected genes (Extended Data Fig. 2). The first step included bronchial mucosa with normal histology (stages 0 and 1, which had normal and low fluorescence, respectively) and hyperplasia (stage 2), which we subsumed under the category of 'normal bronchial tissue'; the second step comprised metaplasia (stage 3) and mild and moderate dysplasia (stages 4 and 5), which were grouped under 'low-grade' lesions; the third step combined severe dysplasia (stage 6) and in situ carcinoma (stage 7) into 'high-grade' lesions; and the fourth step segregated invasive (SCC, stage 8) from premalignant lesions (Extended Data Fig. 2).

Carcinogenesis has been described as the process of acquiring advantageous biological capabilities (the hallmarks of cancer) by abnormal cells. We identified modules of co-expressed genes with distinct expression patterns, and examined them for significant associations with cancer hallmarks (Fig. 1b, Extended Data Fig. 3a, b). We discerned seven evolutionary trajectories of gene expression as gene modules that were derived from weighted gene co-expression network analysis (Fig. 1b). The two largest modules exhibited linear evolution from normal tissue to cancer; the 'ascending' module ($n = 1,848$ genes) was associated with proliferation and the 'descending' module ($n = 939$ genes) was linked to genes that are downregulated in the DNA damage response. A module of 150 co-expressed genes displayed a late increase of expression in high-grade lesions that continued in SCC (the 'ascending from high-grade' module), and was highly enriched with genes that are involved in the immune response. The module of genes that remained unmodified until the onset of cancer ('SCC increase' module; $n = 51$ genes) was over-represented by genes that are involved in epithelial–mesenchymal transition, including a significant increase in expression of *CXCR4* but a low expression of *CXCL12* in SCC (Extended Data Fig. 3c, d). Two additional modules had biphasic gene-expression evolutions; both

[1]Department of Intensive Care and Thoracic Oncology, Jules Bordet Institute, Centre des Tumeurs de l'Université Libre de Bruxelles (ULB), ULB, Brussels, Belgium. [2]Section of Computational Biomedicine, Department of Medicine, Boston University School of Medicine, Boston, MA, USA. [3]INSERM UMR 1068, CNRS UMR 725, Centre de Recherche en Cancérologie de Marseille (CRCM), Aix-Marseille Université, Marseille, France. [4]Department of Multidisciplinary Oncology and Innovative Therapeutics, Assistance Publique-Hôpitaux de Marseille (AP-HM), Marseille, France. [5]INSERM, Laboratory of Integrative Cancer Immunology, Centre de Recherche des Cordeliers, Paris, France. [6]Equipe Labellisée Ligue Contre le Cancer, Paris, France. [7]Sorbonne Université, Paris, France. [8]Sorbonne Paris Cité, Université Paris Descartes, Université Paris Diderot, Université de Paris, Paris, France. [9]Breast Cancer Translational Research Laboratory, Jules Bordet Institute, Centre des Tumeurs de l'Université Libre de Bruxelles (ULB), ULB, Brussels, Belgium. [10]Laboratory of Molecular Immunology, Jules Bordet Institute, Centre des Tumeurs de l'Université Libre de Bruxelles (ULB), ULB, Brussels, Belgium. [11]Department of Pathology, Jules Bordet Institute, Centre des Tumeurs de l'Université Libre de Bruxelles (ULB), ULB, Brussels, Belgium. [12]Department of Pulmonary Medicine, CHU Saint-Pierre, Université Libre de Bruxelles (ULB), ULB, Brussels, Belgium. [13]Laboratory of Molecular and Cellular Biology, Faculté Universitaire des Sciences Agronomiques de Gembloux (FUSAGx), Gembloux, Belgium. [14]Present Address: Department of Pulmonology, Strasbourg University Hospital, Strasbourg, France. [15]Present Address: INSERM IRFAC UMR_S1113, Université de Strasbourg, Strasbourg, France. [16]Present address: Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK. [17]Present address: Pathology Centre, Strasbourg, France. [18]These authors contributed equally: Céline Mascaux, Mihaela Angelova. *e-mail: celine.mascaux@chru-strasbourg.fr; jerome.galon@crc.jussieu.fr

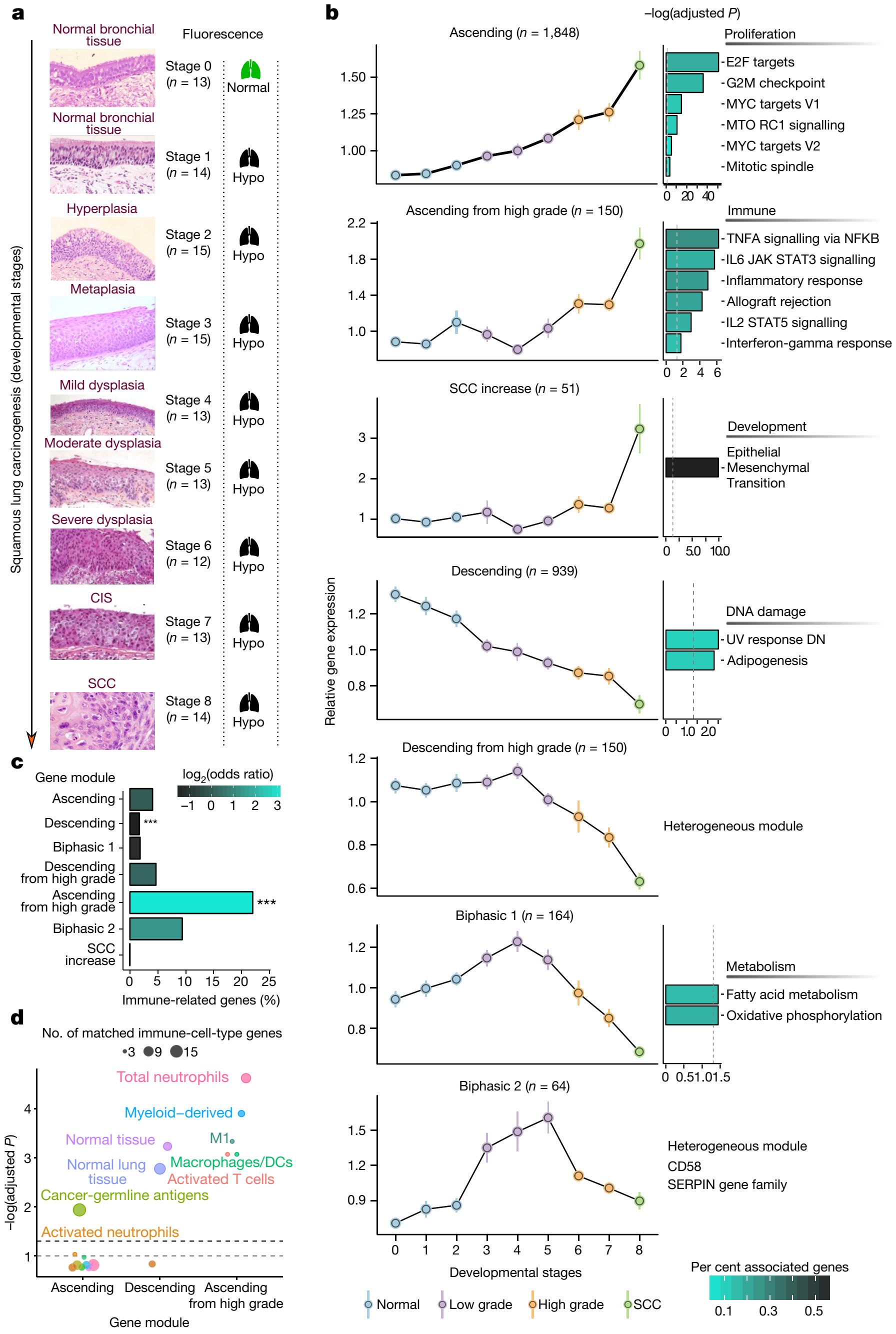


Fig. 1 | Temporal order of cancer hallmarks during carcinogenesis. **a**, Nine morphological stages of development. The normal tissues were split into stage 0 (normal fluorescence) and stage 1 (hypo-fluorescent; hypo) on the basis of fluorescence bronchoscopy. CIS, in situ carcinoma. **b**, Seven modules of co-expressed genes were identified with weighted gene-correlation network analysis. The gene-expression measurement represents the relative abundance of each gene compared to reference RNA from bronchial biopsies from 16 people who had never smoked, derived with two-colour gene-expression microarrays (mean ± s.e.m.). Over-representation analysis linked cancer hallmarks with several gene modules (hypergeometric test, FDR $\leq$ 0.05). Adjusted P values are shown as bar plots after $-\log_{10}(P)$ transformation. **c**, Genes representing immune-cell types were matched to the gene modules. Each module is illustrated with the corresponding fraction of immune-related genes. Odds ratios and P values were derived from Fisher's exact test. **d**, Over-representation analysis of immune, stromal and cancer cell-type gene signatures in gene modules (hypergeometric test, Benjamini–Hochberg correction), using the high definition (HD) immune signature (Supplementary Information). Dotted lines are drawn at $P = 0.10$ (grey) and $P = 0.05$ (black). See Extended Data Fig. 3 for further analysis.

of these modules reached a peak of expression in low-grade lesions (biphasic 1 module, $n = 164$ genes; biphasic 2 module, $n = 64$ genes). Indeed, metabolism regulation had a biphasic trajectory. Specifically, genes involved in fatty acid metabolism, oxidative phosphorylation and the citric acid cycle showed a transitory increase in expression in low-grade lesions (biphasic 1 module).

To analyse the evolutionary trajectory of the immune response, we compiled genes that represent specific immune-cell types, normal cells and cancer cells, and matched them to each gene module (Supplementary Information). We confirmed the highest percentage of immune-related genes in the 'ascending from high-grade' module, and observed a significant under-representation in the 'descending module' ($P < 0.001$) (Fig. 1c). Cancer-germline antigens were found in the 'ascending' module at a significantly higher frequency than expected by chance (FDR < 0.05), as were genes involved in neutrophil activation (FDR < 0.1) (Fig. 1d, Supplementary Table 3). Both observations

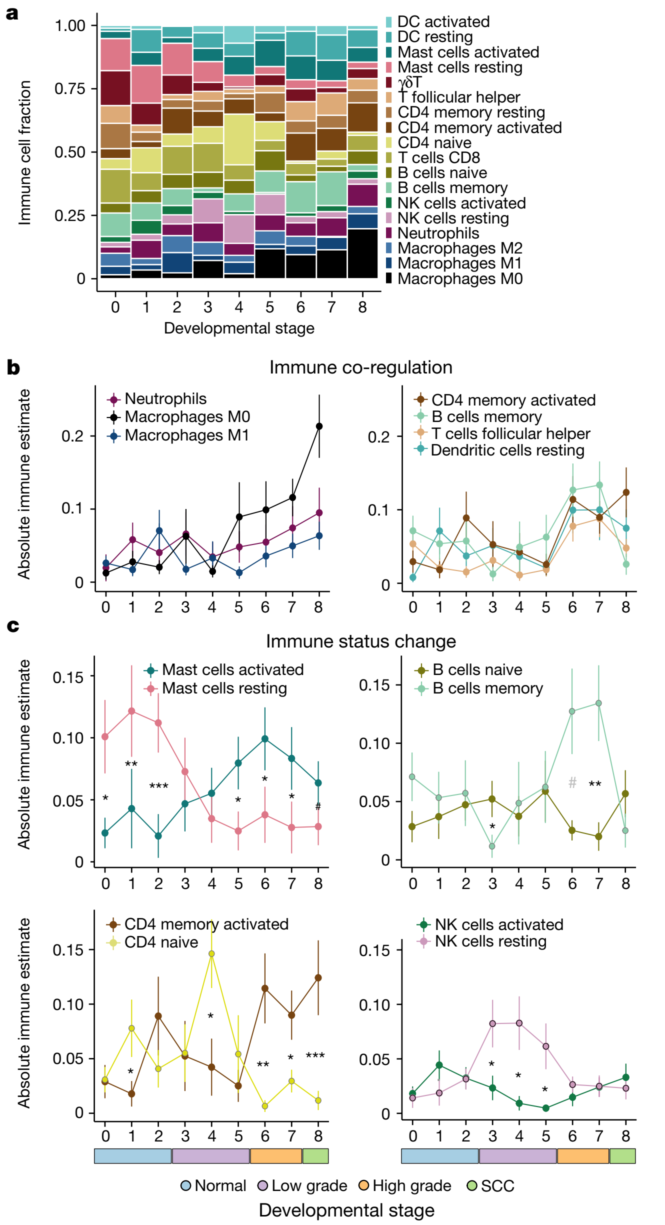


Fig. 2 | Evolving immune response during lung carcinogenesis. a, The estimation of immune-cell abundance from gene-expression profiles shows the evolving immune contexture for each developmental stage. DC, dendritic cell; NK, natural killer. **b**, Several immune-cell types from innate and adaptive immunity are co-regulated with increased abundance in late developmental stages (mean ± s.e.m.). **c**, Continuous shift of immune status for mast cells, memory B cells, CD4 T cells and natural killer cells (mean ± s.e.m.). Significant differences per stage are highlighted with black symbols at FDR <0.1, or otherwise with grey symbols. Mann–Whitney *U* test. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$, #$P < 0.1$, FDR. See Extended Data Fig. 4 for further analysis.

suggest that immune sensing occurs at the earliest steps of transformation. Markedly increased gene expression that represents activated T cells was detected in high-grade lesions before tumour invasion. The same pattern was observed in total neutrophils, M1 macrophages and the myeloid signature.

We then estimated the absolute abundance of different immune-cell types using a method for deconvolving the cell composition of complex tissues from gene expression (Fig. 2a, Extended Data Fig. 4). We confirmed an increase of myeloid-derived cells, neutrophils and macrophage subtypes in high-grade dysplasia (Fig. 2b). Additionally, we observed co-regulation of immune cells from both innate and adaptive immunity on the basis of correlation of immune-cell abundances (Fig. 2b, Extended Data Fig. 4b). Activated T cells (CD4 memory), macrophages (M0), memory B cells, follicular T-helper cells and dendritic cells followed the same abundance pattern. Of note, lesions within the same patient had different immune compositions at different developmental stages (Extended Data Fig. 5). We also detected a shift in the immune status, from resting to activated and from naive to memory (Fig. 2c, Extended Data Fig. 4c). Resting mast cells were more abundant in early developmental stages compared to late stages, whereas activated mast cells followed the opposite pattern (Fig. 2c, Supplementary Table 3). A drop in naive B cell abundance was accompanied by an increase in memory B cells. An influx of naive CD4 cells occurred at the stage of mild dysplasia (stage 4), which was followed by a sharp decline in their abundance and a concurrent increase of activated CD4 memory T cells in the successive stages (Fig. 2c, Supplementary Table 3).

To further elucidate the immune transition at each molecular step of transformation, we performed functional analysis of the differentially regulated genes in transformed compared to normal tissues. We identified Gene Ontology immune processes that were enriched among the differentially regulated genes in low-grade and high-grade lesions, and SCC (Fig. 3a, Supplementary Table 3). Few immune functions were specifically modulated for low-grade lesions—not only among upregulated genes ($n = 5$ functions) but also among downregulated genes ($n = 13$ functions) (for example, response to TGFβ). Unlike in low-grade lesions, a large number of immune functions were uniquely enriched among the upregulated genes in high-grade lesions ($n = 148$ functions) and SCC ($n = 240$ functions). Notably, negative regulation of the immune system, antigen processing and the presentation of peptide antigen were implicated in all developmental stages (Fig. 3a). Nevertheless, the genes associated with negative regulation were significantly downregulated in low-grade lesions, and were upregulated in high-grade lesions and SCC. Therefore, one of the earliest immune reactions is immune unleashing through the downregulation of genes that negatively regulate the immune system, among which we found *TNFRSF14* (also known as *HVEM*), *CD200*, *CD59*, *TGFB3* and *HLA-G* to be downregulated. Conversely, in high-grade lesions and SCC there was an upregulation of genes that are involved in immunosuppression.

Closer examination of immunomodulatory gene expression revealed that the average expression of co-inhibitory molecules and suppressive interleukins was significantly higher in severe dysplasia and the succeeding stages (Fig. 3b). Overall, many immunomodulatory molecules had a positive fold change in high-grade dysplasia compared to normal tissue (Fig. 3c, Extended Data Fig. 6a, b). As well as suppressive molecules such as *IDO1*, *PD-L1* (also known as *CD274*), *TIGIT*, *CTLA4*, *ICOS*, *IL10* and *IL6*, stimulatory molecules such as *TNFRSF9* (also known as *CD137*), *TNFRSF18* (also known as *GITR*), *ICOS*, *CD80*, *CD86*, *CD70*, *TNFSF9* (also known as *CD137L*) and *TNFRSF25* showed increased expression in high-grade dysplasia and, to a greater extent, at the invasive stage. The expression of the immune checkpoints IDO1, PD-L1, CTLA4, TIGIT and TIM3 was also confirmed at the protein level by immunohistochemistry (Extended Data Fig. 6a–c). Each of the tested markers showed an increase in SCC compared to normal tissue, which was significant for CTLA4, IDO1 and PD-L1 ($P < 0.05$) but not for TIGIT ($P = 0.14$) and TIM3 ($P = 0.095$) (Extended Data Fig. 6a–c). Collectively, immune escape occurred before tumour invasion, as shown by the fact that co-inhibitors and suppressive interleukins increased significantly from high-grade stages onwards.

For high-definition characterization of the microenvironment architecture, we used two seven-plex staining panels on the same bronchial epithelial lesions: a phenotype panel to discern immune-cell types and a functional panel that includes PD1, PD-L1, Ki67 and CD137 ($n = 110$ and 106 samples, respectively, Extended Data Fig. 6). First, we calculated immune-cell densities individually for the stromal and epithelial tissue category (Fig. 4a). Overall, we found a relatively large variation in the immune-cell densities. However, we observed significant differences among the four developmental stages in the stromal compartment, and the same trends in the epithelial compartment (Fig. 4a). CD4 T cells ($CD3^{+}CD8^{-}$) and $CD8^{+}$ lymphocytes both showed a transitory

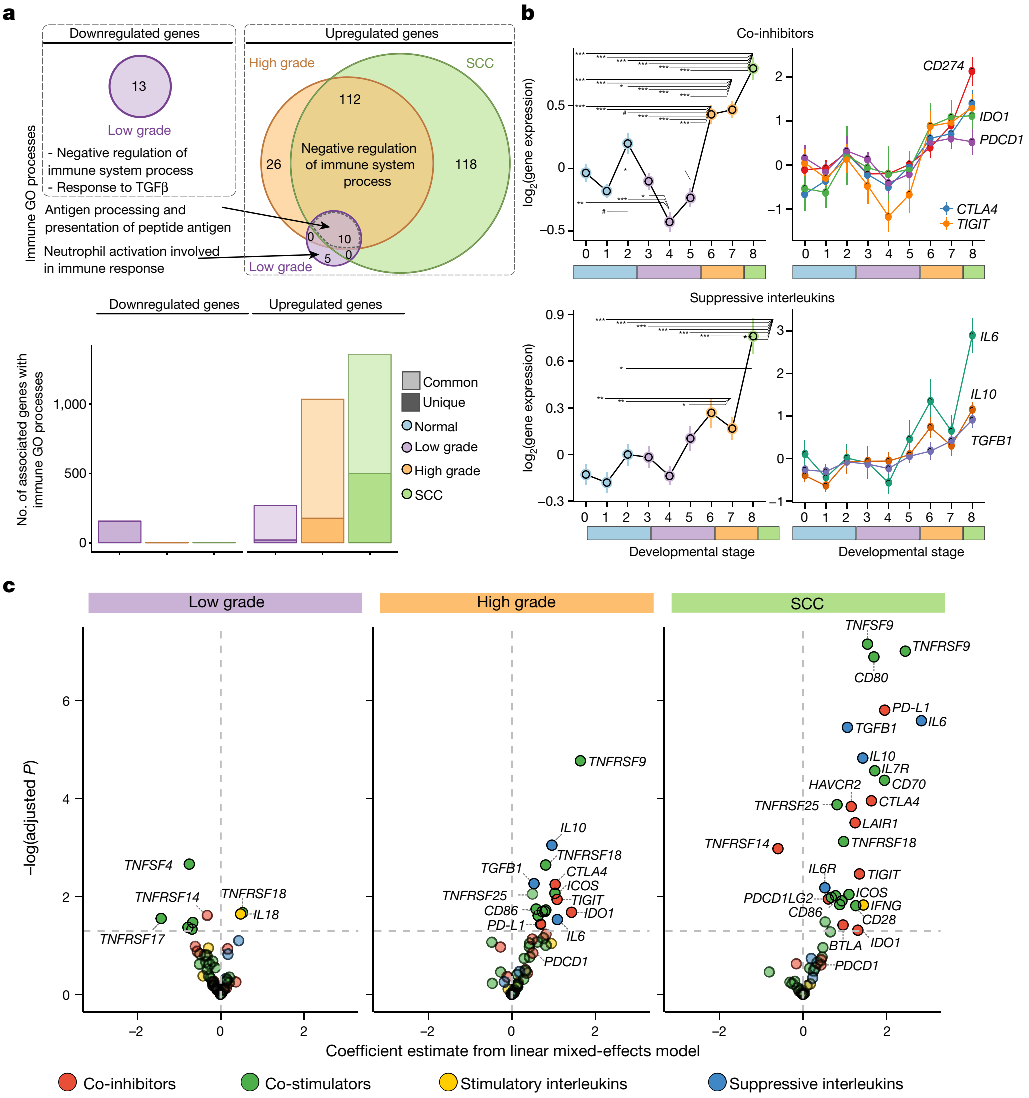


Fig. 3 | Immune evasion before tumour invasion in early lung squamous carcinogenesis. a, Gene Ontology (GO) immune functions that are significantly enriched in upregulated and downregulated genes were compared between low-grade and high-grade lesions, and SCC, accounting for smoking history, previous cancer status and inter-patient variability as confounding factors (linear mixed-effects model, FDR < 0.05). Top, Venn diagrams showed Gene Ontology immune functions that were significantly enriched in downregulated (left) and upregulated (right) genes, comparing low-grade and high-grade lesions, and SCC. Bottom, the number of genes associated with the immune functions is represented on a bar plot for each developmental stage (Supplementary Table 3). **b**, Average expression of co-inhibitory molecules and suppressive interleukins (mean ± s.e.m.). Rank-based test, Dunn's pairwise multiple comparison test. $***P < 0.001$, $**P < 0.01$, $*P < 0.05$, $\#P < 0.1$. **c**, Positive fold changes in high-grade lesions and SCC compared to the corresponding normal tissue expression observed for co-stimulatory, co-inhibitory and suppressive interleukins (linear mixed-effects model of the four molecular steps adjusted for smoking history and previous cancer status as fixed effects, and patient as random effect). See Extended Data Fig. 6a for further analysis. *TNFRSF17* is also known as *BCMA*, *PDCD1* as *PD1*, *PDCD1LG2* as *PD-L2* and *HAVCR2* as *TIM3*.

increase in high-grade pre-invasive lesions ($P < 0.01$). Consistent with the immune gene-expression evolution, myeloid, neutrophil and macrophage densities increased in the stroma ($P < 0.05$, FDR < 0.1) and the epithelium ($P < 0.1$ before Benjamini–Hochberg correction) of high-grade lesions. In accordance with the gene expression data, PD-L1 (PD-L1$^+$cytokeratin(CK)$^-$) densities significantly increased in high-grade lesions, and increased even more in SCC ($P < 0.05$) (Fig. 4a), similar to CD137 ($P < 0.1$). Cells with the CD137, PD-L1 and CD3$^+$FOXP3$^+$ phenotypes were rarely found in the epithelium at early stages of development (stages 0–5; that is, normal and low-grade lesions).

We next performed second-order spatial statistics and measured distances between each pair of cell phenotypes. On the basis of cross-type cumulative distribution of nearest neighbour distances ($G(r)$) (Fig. 4b), we detected segregation among epithelial cells (CK$^+$) and CD3 consistently in both panels ($P < 0.001$, FDR < 0.1) (Fig. 4c). In particular, we observed a lower number of epithelial cells than expected near CD3 cells in high-grade lesions (Fig. 4c). This pattern was observed for all CK$^+$ cells in the functional panel, total epithelial cells (all CK$^+$) and CK$^+$PD-L1$^+$ cells ($P < 0.01$, FDR < 0.1) (Extended Data Fig. 6f). Therefore, in high-grade lesions, we discerned the reconfiguration of the tumour microenvironment compared to the preceding stages of development, manifested by segregation of epithelial cells from CD3 cells.

Our data show that both immune activation and immune suppression occur at pre-invasive stages of cancer development, which supports the hypothesis of immune surveillance in pre-cancerous lesions, reinforces the use of immunotherapy at the earliest steps of treatment and underlines the potential role of immunotherapy in chemopreventive approaches. The prognostic effect of immune infiltrates has previously been demonstrated in various types of cancer[10–12] at early stages[13], including lung cancer[14] from stage I[15]. Recently, genomic

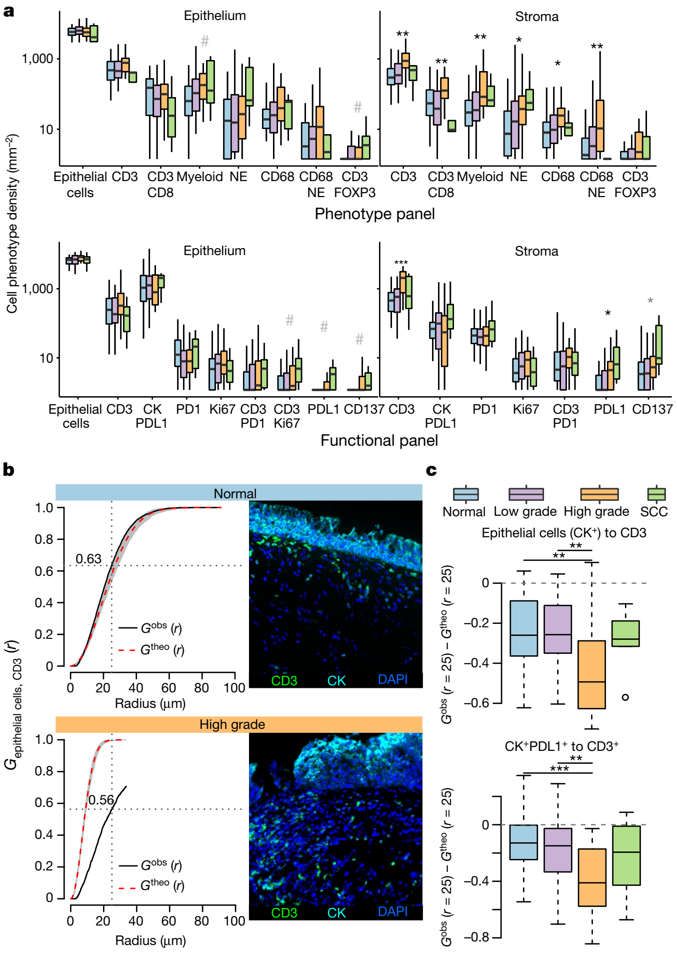


Fig. 4 | Spatial reconfiguration of the tumour microenvironment in high-grade lesions of lung squamous carcinogenesis. a, Using the multispectral imaging of the phenotype (top) and functional panel (bottom), we quantified the densities of each identified cell phenotype as the number of cells per tissue surface area. A cell phenotype with a single marker was single-positive only for that marker, and negative for all the other markers used. The four molecular groups were compared by immune-cell densities and the significant differences are highlighted in black at FDR <0.1, or otherwise in grey. Non-parametric, rank-based test, FDR. ***$P < 0.001$, **$P < 0.01$, *$P < 0.05$. **b**, Analysis and composite images from representative normal and low-grade samples. The function G^{obs} (r) for $G_{\text{epithelial cells, CD3}}$ (r) represents the cumulative distribution of the distances from a random epithelial cell to its nearest CD3 cell in an area of radius r. The theoretical curve G^{theo} (r) for $G_{\text{epithelial cells, CD3}}$ (r) shows a random sample distribution within its confidence envelope in grey. Deviations from the theoretical curve observed in the high-grade sample suggest the segregation of the epithelial cells from the CD3 cells. The differences between the observed and the expected number of epithelial cells for which the closest CD3 cell is within 25 μm were compared among the molecular groups. See Extended Data Fig. 6 for further analysis.

and epigenomic analyses of microdissected lung carcinoma in situ (high-grade lesion)—the pre-invasive precursor to SCC—have shown that progressive lesions have significantly more genomic alterations than regressive lesions, across base substitutions, indels, driver mutations, structural variants and copy-number changes[16]. However, no single cancer mutation perfectly discriminated between progressive and regressive lesions. Moreover, in situ carcinoma lesions, including spontaneously regressing lesions, contained genomic, epigenomic and transcriptomic hallmarks of advanced invasive SCC. Thus, the mechanism behind regression has so far remained unknown[16]. The contribution of tumour-intrinsic factors to the risk of carcinogenesis has previously been shown to be modest[17], as compared to extrinsic carcinogens[17] or dysregulation of the immune microenvironment[18]. It has previously been shown that the tumour microenvironment is a critical determinant of dissemination to distant metastasis[19] and of metastatic tumour development, in which tumour evolution could be traced back to immune-escaping clones[18]. These findings could also apply to pre-malignant transformation and the initiation of carcinoma. Furthermore, a major clinical benefit of checkpoint immunotherapy has been obtained in various settings of cancer treatment[20]. In non-small-cell lung cancer, checkpoint inhibitors are now standard as first-line[21,22] and second-line treatment options for advanced disease[23,24], and as maintenance after curative chemo-radiation of locally advanced stages[25]. However, early intervention remains the best opportunity for curing patients with lung cancer. The positive results of immune-checkpoint-blockade therapy in an adjuvant setting for melanoma[26] and in a neoadjuvant setting for lung cancer[27] reinforce the importance of using immunotherapy in the early steps of treatment strategies.

Our study has delineated the molecular pathways that are involved in four steps of the carcinogenesis of lung SCC (Extended Data Fig. 7), in which the earliest molecular changes affect proliferation and metabolism. The transient rise in metabolic pathways might reflect the shift in cellular function from secretory to protective keratinization, a pattern that has previously been described for micro RNA expression in a subset of the same pre-neoplastic lesions[28]. Similarly, a transient influx of naive T cells was observed in low-grade lesions. Collectively, the immune transition unfolds as follows: (1) immune sensing and immune unleashing are induced at the earliest step of transformation; (2) continual cell proliferation fosters the accumulation of somatic mutations, mounting an anti-tumour immune response; and (3) inherent immune-suppression mechanisms are triggered in high-grade pre-cancerous lesions. Previous studies have shown that the risk of cancer progression is much higher in high-grade lesions (32–87%), compared to low-grade lesions (2–9%)[29,30]. Our results suggest the need to assess the role of immunotherapy in chemoprevention approaches for individuals at a high risk of developing lung cancer.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-019-1330-0.

Received: 27 November 2018; Accepted: 24 May 2019;
Published online: 26 June 2019

1. Goldberg, S. W., Mulshine, J. L., Hagstrom, D. & Pyenson, B. S. An actuarial approach to comparing early stage and late stage lung cancer mortality and survival. *Popul. Health Manag.* **13**, 33–46 (2010).
2. Bindea, G. et al. Spatiotemporal dynamics of intratumoral immune cells reveal the immune landscape in human cancer. *Immunity* **39**, 782–795 (2013).
3. Mlecnik, B. et al. Histopathologic-based prognostic factors of colorectal cancers are associated with the state of the local immune reaction. *J. Clin. Oncol.* **29**, 610–618 (2011).
4. Herbst, R. S., Morgensztern, D. & Boshoff, C. The biology and management of non-small cell lung cancer. *Nature* **553**, 446–454 (2018).
5. De Koning, H., Van Der Aalst, C., Ten Haaf, K. & Oudkerk, M. PL02.05 effects of volume CT lung cancer screening: mortality results of the NELSON randomised-controlled population based trial. *J. Thorac. Oncol.* **13**, S185 (2018).
6. Umar, A., Dunn, B. K. & Greenwald, P. Future directions in cancer prevention. *Nat. Rev. Cancer* **12**, 835–848 (2012).
7. Kensler, T. W. et al. Transforming cancer prevention through precision medicine and immune-oncology. *Cancer Prev. Res.* **9**, 2–10 (2016).
8. Slaughter, D. P., Southwick, H. W. & Smejkal, W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. *Cancer* **6**, 963–968 (1953).
9. Kerr, K. M. Pulmonary preinvasive neoplasia. *J. Clin. Pathol.* **54**, 257–271 (2001).
10. Galon, J. et al. Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science* **313**, 1960–1964 (2006).
11. Galon, J., Angell, H. K., Bedognetti, D. & Marincola, F. M. The continuum of cancer immunosurveillance: prognostic, predictive, and mechanistic signatures. *Immunity* **39**, 11–26 (2013).
12. Pagès, F. et al. International validation of the consensus immunoscore for the classification of colon cancer: a prognostic and accuracy study. *Lancet* **391**, 2128–2139 (2018).
13. Pagès, F. et al. In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer. *J. Clin. Oncol.* **27**, 5944–5951 (2009).

14. Fridman, W. H., Pagès, F., Sautès-Fridman, C. & Galon, J. The immune contexture in human tumours: impact on clinical outcome. *Nat. Rev. Cancer* **12**, 298–306 (2012).
15. Gentles, A. J. et al. The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med*. **21**, 938–945 (2015).
16. Teixeira, V. H. et al. Deciphering the genomic, epigenomic, and transcriptomic landscapes of pre-invasive lung cancer lesions. *Nat. Med*. **25**, 517–525 (2019).
17. Wu, S., Powers, S., Zhu, W. & Hannun, Y. A. Substantial contribution of extrinsic risk factors to cancer development. *Nature* **529**, 43–47 (2016).
18. Angelova, M. et al. Evolution of metastases in space and time under immune selection. *Cell* **175**, 751–765 (2018).
19. Mlecnik, B. et al. The tumor microenvironment and immunoscore are critical determinants of dissemination to distant metastasis. *Sci. Transl. Med*. **8**, 327ra26 (2016).
20. Galon, J. & Bruni, D. Approaches to treat immune hot, altered and cold tumours with combination immunotherapies. *Nat. Rev. Drug Discov*. **18**, 197–218 (2019).
21. Gandhi, L. et al. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. *N. Engl. J. Med*. **378**, 2078–2092 (2018).
22. Reck, M. et al. Pembrolizumab versus chemotherapy for PD-L1-positive non-small-cell lung cancer. *N. Engl. J. Med*. **375**, 1823–1833 (2016).
23. Brahmer, J. et al. Nivolumab versus docetaxel in advanced squamous-cell non-small-cell lung cancer. *N. Engl. J. Med*. **373**, 123–135 (2015).
24. Herbst, R. S. et al. Pembrolizumab versus docetaxel for previously treated, PD-L1-positive, advanced non-small-cell lung cancer (KEYNOTE-010): a randomised controlled trial. *Lancet* **387**, 1540–1550 (2016).
25. Antonia, S. J. et al. Overall survival with durvalumab after chemoradiotherapy in stage III NSCLC. *N. Engl. J. Med*. **379**, 2342–2350 (2018).
26. Eggermont, A. M. M. et al. Adjuvant pembrolizumab versus placebo in resected stage III melanoma. *N. Engl. J. Med*. **378**, 1789–1801 (2018).
27. Forde, P. M. et al. Neoadjuvant PD-1 blockade in resectable lung cancer. *N. Engl. J. Med*. **378**, 1976–1986 (2018).
28. Mascaux, C. et al. Evolution of microRNA expression during human bronchial squamous carcinogenesis. *Eur. Respir. J*. **33**, 352–359 (2009).
29. Bota, S. et al. Follow-up of bronchial precancerous lesions and carcinoma in situ using fluorescence endoscopy. *Am. J. Respir. Crit. Care Med*. **164**, 1688–1693 (2001).
30. Breuer, R. H. et al. The natural course of preneoplastic lesions in bronchial epithelium. *Clin. Cancer Res*. **11**, 537–543 (2005).

Acknowledgements We thank P. Pierard, J. Faber, M. Bruyneel, S. Bensliman and O. Bauwens, who performed bronchoscopy sampling; A.-P. Meert and T. Berghmans, who recruited the patients from Jules Bordet Institute; C. Decoster, B. Martin and J.-F. Godart, who contributed to biopsy collection; Y. Cleuter, A. Vandenbroecke, J.-F. Laes and L. Lafontaine, who shared their technical expertise; B. Grisart, P. Martiat and C. Sotiriou, who opened access to their laboratory; and all the patients who generously participated in this study. This research was supported by the Fund of Medical Scientific Research (Belgium), the National Fund of Scientific Research-Télévie campaign and the ASBL 'Les Amis de l'Institut Jules Bordet'. C.M. was a research fellow from the National Fund of Scientific Research (Belgium) and was supported by a grant from the Louisiana Chapter of the National Lung Cancer Partnership (USA). J.B. was supported by a LUNGevity Foundation Career Development Award. This work was also supported by grants from the National Cancer Institute of France (INCa), the Canceropole Ile de France, INSERM, La Ligue Contre le Cancer, the Cancer research for personalized medicine (CARPEM), Paris Alliance of Cancer Research Institutes (PACRI) and LabEx Immuno-oncology.

Reviewer information *Nature* thanks Lynette Marie Sholl and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions C.M., A.B., J.-P.S. and J.G. designed the study. C.M., G.A., A.H. and V.N. collected, handled and classified the samples, and collected the clinical data. C.M., G.A., B.B. and F.R. performed the experiments. C.M., A.V., G.A. and B.B. acquired the data. M.A., C.M., A.V., J.B., K.H., K.W.-G., A.B., J.-P.S., A.S. and J.G. analysed and interpreted the data. M.A., C.M., J.B. and K.H. performed the statistical analyses. C.M., J.B., J.-P.S. and J.G. acquired the funding required for the study. C.M., A.B., J.-P.S, A.S. and J.G. supervised the project. C.M., M.A and J.G. drafted the manuscript. All authors performed a critical revision of the manuscript.

Competing interests The authors declare competing financial interests: J.G. is co-founder of HalioDx biotech company. A.S. is an employee of Johnson and Johnson. C.M. received honoraria for expertise and speaker roles, being on the advisory board or attending meetings, from Boehringher Ingelheim, Lilly, Bristol-Myers Squibb, AstraZeneca, Roche, Kephren and Novartis. European Patent Application EP19305434.3 was filed on 2 April 2019 by Inserm, Université Paris Descartes, Université Paris Diderot, Sorbonne Université, Université Aix-Marseille, APHM and Université Libre de Bruxelles, and European Patent Application EP19305535.7 was filed on 26 April 2019 by Inserm, Université Paris Descartes, Université Paris Diderot, Sorbonne Université, Université Aix-Marseille, APHM and Université Libre de Bruxelles. The inventors designation is in progress.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-019-1330-0.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1330-0.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.M. or J.G.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

A summary of the methods is shown in Extended Data Fig. 1.

Study population. Bronchial biopsies were collected between 2003 and 2007 at the Jules Bordet Institute, during fluorescence bronchoscopy of current and former smokers (defined as individuals who had quit smoking for more than six months), with a smoking exposure of ≥30 pack-years. The technique of fluorescence bronchoscopy is based on the autofluorescence of bronchial epithelium, which enables the discrimination of pre-invasive from normal tissue. When illuminated with a wavelength of 380–440 nm, the pre-invasive lesions exhibit much weaker green fluorescence than normal lung tissue (hypo-fluorescence). Therefore, lesions that are difficult to distinguish from normal epithelial tissue in white-light bronchoscopy can be detected under the 380–440-nm wavelength owing to their lack of green fluorescence (Extended Data Fig. 1b). This technique demonstrates high sensitivity at the cost of specificity. A lack of fluorescence can be observed also in some normal lesions, but it has not been shown whether hypo-fluorescent normal tissue is molecularly different from the normal tissue with normal fluorescence. Therefore, normal tissues that were characterized by normal fluorescence and hypo-fluorescence were considered, and analysed as separate stages of development, (stage 0 and stage 1, respectively). The study was approved by the ethics committee of the Jules Bordet Institute, and the patients gave informed consent. No statistical methods were used to predetermine sample size, but the number of samples per group (minimum 12) was based on previously published information (Supplementary Information). A total of 122 biopsies from 77 individuals (35 former and 42 current smokers) were studied (Supplementary Table 1). The median age was 62 years (range 42–78). The male-to-female ratio was 62 to 15. The 122 biopsies were distributed according to histology and fluorescence status as follows: 13 biopsies with normal histology and normal fluorescence (8 and 5), 14 biopsies with normal histology and hypo-fluorescence (8 and 6), 15 hyperplasia (7 and 8), 15 metaplasia (5 and 10), 13 mild dysplasia (8 and 5), 13 moderate dysplasia (7 and 6), 12 severe dysplasia (2 and 10), 13 carcinoma in situ (5 and 8) and 14 SCC (5 and 9) (parenthetical numbers refer to biopsies from former and current smokers, respectively). Among the 122 samples, matched formalin-fixed paraffin-embedded (FFPE) blocks were found for 110 of them (Supplementary Table 2).

Sample collection and RNA extraction. The process for handling and freezing the biopsies was carefully standardized (Supplementary Information). RNA was successfully extracted from 122 freshly frozen biopsies using previously described RNA extraction protocols[28]. The median yield of total RNA extracted from the biopsies was 1,275 ng (range 244–11,000 ng).

Acquisition and analysis of gene expression profiles. The cDNA was in vitro-transcribed into complementary (c)RNA and labelled using the dye Cy5 for the RNA derived from the 122 samples of interest, and Cy3 for the reference RNA. The reference RNA was pooled in equal amount from normal bronchial biopsies from 16 people who had never smoked (Agilent Technologies) (Supplementary Information). After amplification and labelling, cRNAs were hybridized on two Colours Whole Human Genome 4 × 44K arrays, according to the recommendation of the provider (Agilent Technologies) (Supplementary Information). Additional normalization steps were performed with Genespring GX version 7.3.1 software (Agilent Technologies): (1) per spot (divide by control channel), (2) per chip (normalize to the median expression value across chip) and (3) per gene (normalize to median expression value across patients). The gene-expression measurements reported in the Letter represent the relative abundance of each probe and gene—that is, the ratio between the red and green colour intensity (Cy5/Cy3) in the studied sample compared to normal biopsies from people who had never smoked. Several steps of data quality control were performed during data collection, generation and processing (Supplementary Information, Extended Data Fig. 8).

Identification of linear gene-expression changes and molecular phenotypes. Gene-expression alterations associated with developmental stages were identified using a linear model with mixed-effects. Each gene was modelled as a function of the developmental stage (factor variable), adjusting for smoking status, sex and history of cancer as fixed effects. Because patient-level observations are not independent, we considered the patient parameter as a random effect. Analysis of variance (ANOVA) tests compared the association of a gene and developmental stage to a null model. The FDR was calculated for each ANOVA P value using the Benjamini–Hochberg method. Genes that were significantly associated with developmental stages were determined by an ANOVA FDR <0.001. Semi-supervised hierarchical clustering of these genes was then used to compare the nine different developmental stages (Extended Data Fig. 1).

Definition and functional characterization of gene modules. To identify trajectories of gene expression during development, we applied weighted gene-correlation network analysis with the tool WGCNA to the genes that were significantly associated with developmental stages (Supplementary Information). A minimum cluster size of 50 genes was used to define a module. A P value ratio threshold of 0 was considered for reassigning genes across modules. We determined gene clusters (modules) of highly correlated genes with similar expression patterns across the nine developmental stages. We demonstrated stability and robustness of the detected modules using resampling techniques (Extended Data Fig. 9, Supplementary Information).

To functionally describe the gene modules, we used the cancer hallmark definitions from the mSigDB database (v.6.2) and applied the over-representation hypergeometric test using the R package clusterProfiler. The P values from the over-representation (hypergeometric) tests were adjusted for multiple comparison testing and significant associations were reported at adjusted P values of $P \leq 0.05$. The adjusted P values were then transformed as $-\log_{10}(P)$ and visualized as bar plots (Fig. 1b). Probes were mapped to unique Entrez gene identifiers. The genes were ranked by their z-score-transformed expression values. A minimum overlap of five genes with a given set of genes was required. The enrichment score represents the degrees to which the genes from a given cancer hallmark set of genes were upregulated or downregulated within a sample.

Immune-cell-type signatures. To explore a large number of different immune-cell subtypes and to examine their activation status, we compiled a large number of carefully annotated microarray gene-expression profiles from 1,769 publicly available microarrays normalized with the frozen robust multi-array averaging method (Extended Data Fig. 1, Supplementary Information, Supplementary Table 4).

Immune characterization from gene-expression profiles. The defined immune signatures were used to explore a large variety of immune-cell types from the gene-expression data at different histological stages of SCC development. First, we performed a hypergeometric test between the immune signatures and the gene modules to pinpoint potential evolutionary trajectories of specific immune-cell types (Fig. 1d).

We next applied the algorithm for absolute quantification implemented in CIBERSORT and deconvolved immune-cell-type expression from a mixed gene-expression signal, according to the predefined LM22 signature. LM22 is a validated gene signature matrix on 22 haematopoietic cell types that have significantly different expression in one leukocyte population compared to all other populations. The signature matrix was developed together with the method CIBERSORT, and used to deconvolve transcriptomes.

Last, we performed single-sample gene-set enrichment analysis using the in-house-defined immune-gene signature HD (Extended Data Fig. 4). We thereby obtained, for each immune cell type, an enrichment score per sample that indicated the extent of upregulation or downregulation of the associated genes. The probes identifiers were mapped to unique Entrez gene identifiers. A minimum overlap of five genes was required.

Immunohistochemistry. Four slides at 4-μm thickness were cut from FFPE blocks from SCC biopsies ($n = 7$) and biopsies of normal lung tissue ($n = 12$). Slides were then stained for the immune checkpoints CTLA-4 (1 h, 1 μg/ml, pH 9) (clone BSB88, BioSB), IDO1 (15 min, 0.25 μg/ml, pH 9) (clone V1NC3IDO, eBioscience), TIGIT (30 min, 1/200, pH 6) (clone BLR047F, Abcam) and TIM3 (30 min, 2 μg/ml, pH9) (clone 2321C, R&D system) with Leica Bond RX automate according to the classic IHC-F protocol. Optimizations of staining were performed on control tonsil tissue and on lung SCC tissue. The Bond Polymer Refine Detection kit (Leica) was used to obtain a chromogenic 3,3-diaminobenzidine (DAB) staining. All slides were enclosed in glycergel mounting medium (Dako) and scanned at 20× with the Nanozoomer 2.5 (Hamamatzu). The images were analysed for staining quantification on HALO software (Indica Labs), with Multiplex IHC 2.0 module. The cell densities were calculated as the cell count per tissue area (cell/mm^2). The clustering of density data was performed using Genesis software, hierarchical clustering, average dot product and average linkage.

Multiplex immunohistochemistry and multispectral image analysis. Matched FFPE blocks of the 122 freshly frozen samples were available for 110 samples. Two 4-μm-thick slides were cut from the FFPE blocks, deparaffinized in clarene, rehydrated through an ethanol gradient and fixed in NBF (10% neutral buffered formalin). Slides were then stained according to Opal 7-plex technology (PerkinElmer), enabling the simultaneous visualization of six markers on the same slide. Therefore, at each of the six cycles of staining, antigen retrieval was performed via microwave treatment in antigen retrieval solution pH 6 or pH 9 (AR6 or AR9) depending on the target; protein blocking was performed using Protein Block-Serum-free (Dako) for 15 min; and primary antibodies were then incubated for 30 min at room temperature. Next, incubation with HRP Labelled Polymer mouse or rabbit (Dako EnVision+ System- HRP Labelled Polymer) was performed at room temperature for 15 min followed by TSA opal fluorophores (Opal 520, Opal 540, Opal 570, Opal 620, Opal 650 or Opal 690) incubation for 10 min. Microwave treatment was performed at each cycle of staining to remove the antibody TSA complex with AR solution (pH 9 or pH 6). Finally, all slides were counterstained with DAPI for 5 min and enclosed in ProLong Diamond Antifade Mountant (Thermo Fisher). The slides were scanned using the PerkinElmer Vectra 3 System, and the multispectral images obtained were unmixed using spectral libraries that were previously built from images stained for each fluorophore (monoplex), using the inForm Advanced Image Analysis software (inForm 2.3.0 PerkinElmer).

The entire area of the biopsies was stained, scanned, scored and quantified. The scoring method consisted of several automated steps: tissue categorization, cell segmentation and cell phenotyping. Using the integrated InForm image analysis software, multispectral images that were representative of different samples were selected and used to train the InForm software for categorization of the tissue into epithelium and stroma. The settings learnt from the training on the representative images from different samples were saved within an algorithm, which enabled batch analysis of all the tissue slides. We designed two different seven-plex panels defined as phenotype and functional panels, which were used on two sequential slides to characterize the immune microenvironment of pre-cancer lesions of the lung, including activated and inactivated cells, activated and inactivated immune pathways, and immune-response type. The phenotype panel included CD3, CD8, FOXP3, CD68, neutrophil elastase (NE), DAPI and CK, and the functional panel included CD3, PD-L1, PD1, Ki67, CD137, DAPI and CK. The tumour (epithelial) and stromal regions were classified by the software using the CK^{+} staining. The classification was verified and approved by a certified pathologist (A.H.). Tumour heterogeneity analyses were performed (Extended Data Fig. 10, Supplementary Information).

Spatial statistics. We performed first- and second-order spatial analysis of multispectral imaging data, which enabled a high-definition characterization of the microenvironment architecture. First, we reconstructed whole slides rather than separately analyse each image (which introduces edge effects and leads to a loss of information) (Extended Data Fig. 6). We calculated immune-cell densities as the number of positive cells per unit of tissue surface area (mm^2). On the basis of the tissue categorization performed with the inForm software, the stroma and the epithelium compartments were annotated on the images, which enabled densities and spatial distribution to be calculated individually for the stromal and epithelial tissue categories (Extended Data Fig. 6). To compare the spatial localization of different immune-cell types, we calculated the distances to the nearest neighbours and their distribution by implementing edge corrections, $G(r)$. The function $G(r)$ is the cumulative distribution of the distance from a typical random cell (x) to its nearest cell (y), in which the argument r is the radius of the area in which $G(r)$ is evaluated. We expected a potential interaction when two cells were within a distance of 25 μm. By comparison of the observed empirical function $G_{x,y}(r)$ to the theoretical curve $G_{x,y}^{\mathrm{theo}}(r)$ that shows random sample distribution, deviations from the empirical and the theoretical $G(r)$ function indicate clustered and dispersed patterns. To demonstrate that the results do not depend on the distance cut-off, we calculated the area between these two curves and consistently confirmed that epithelial cells segregate from CD3 T cells in high-grade lesions (Extended Data Fig. 6g).

Statistics. R statistical software (v.3.3.3) was used for statistical analyses and graphical visualization. The null hypotheses were rejected at P values lower than 0.05, unless indicated otherwise. When comparing tumour-tissue to normal-tissue gene expression, a linear mixed-effects model was used to adjust for the confounding factors (smoking history, history of cancer, inter-patient variability, sex and age). The Benjamini–Hochberg method was applied for multiple testing correction. Post-hoc multiple testing correction was applied for pairwise comparison using Dunn's test.

Randomized principal component analysis was performed on $\log_2$-transformed gene expression, using the function rpca implemented in the R package rsvd, which allowed data centring and scaling by variance for all probes.

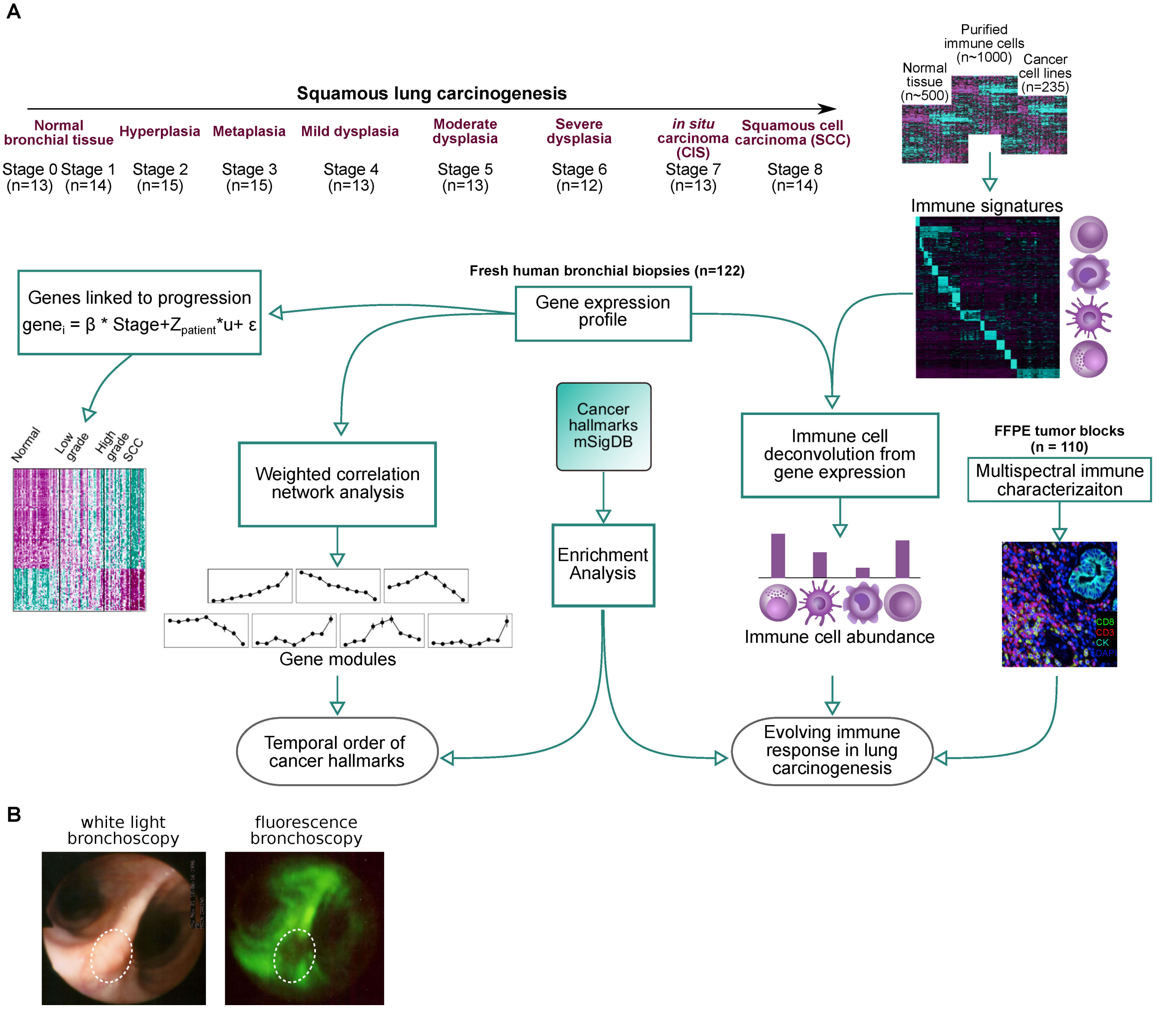
Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

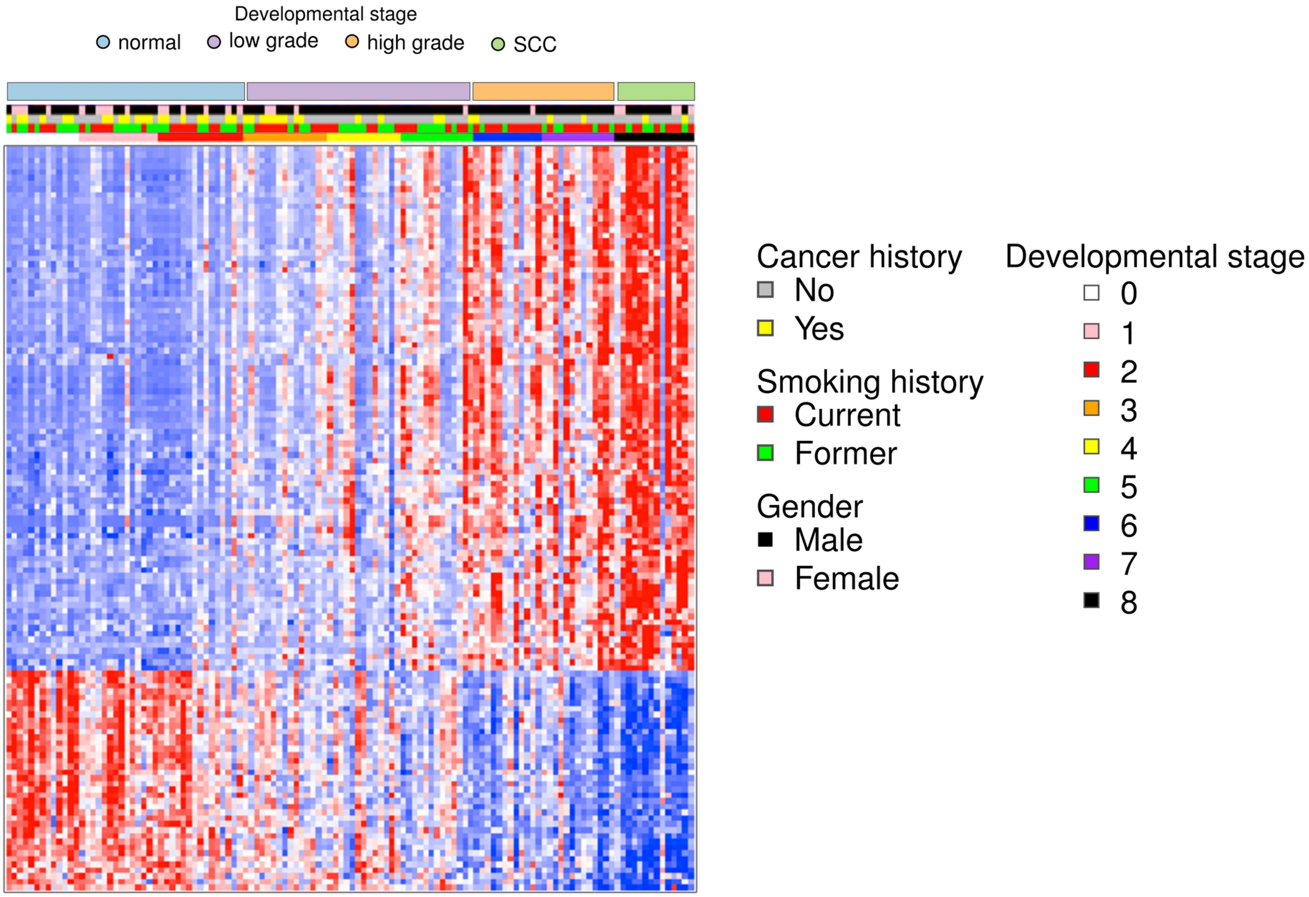
Gene expression data are available in the Gene Expression Omnibus database with accession number GSE33479.

Code availability

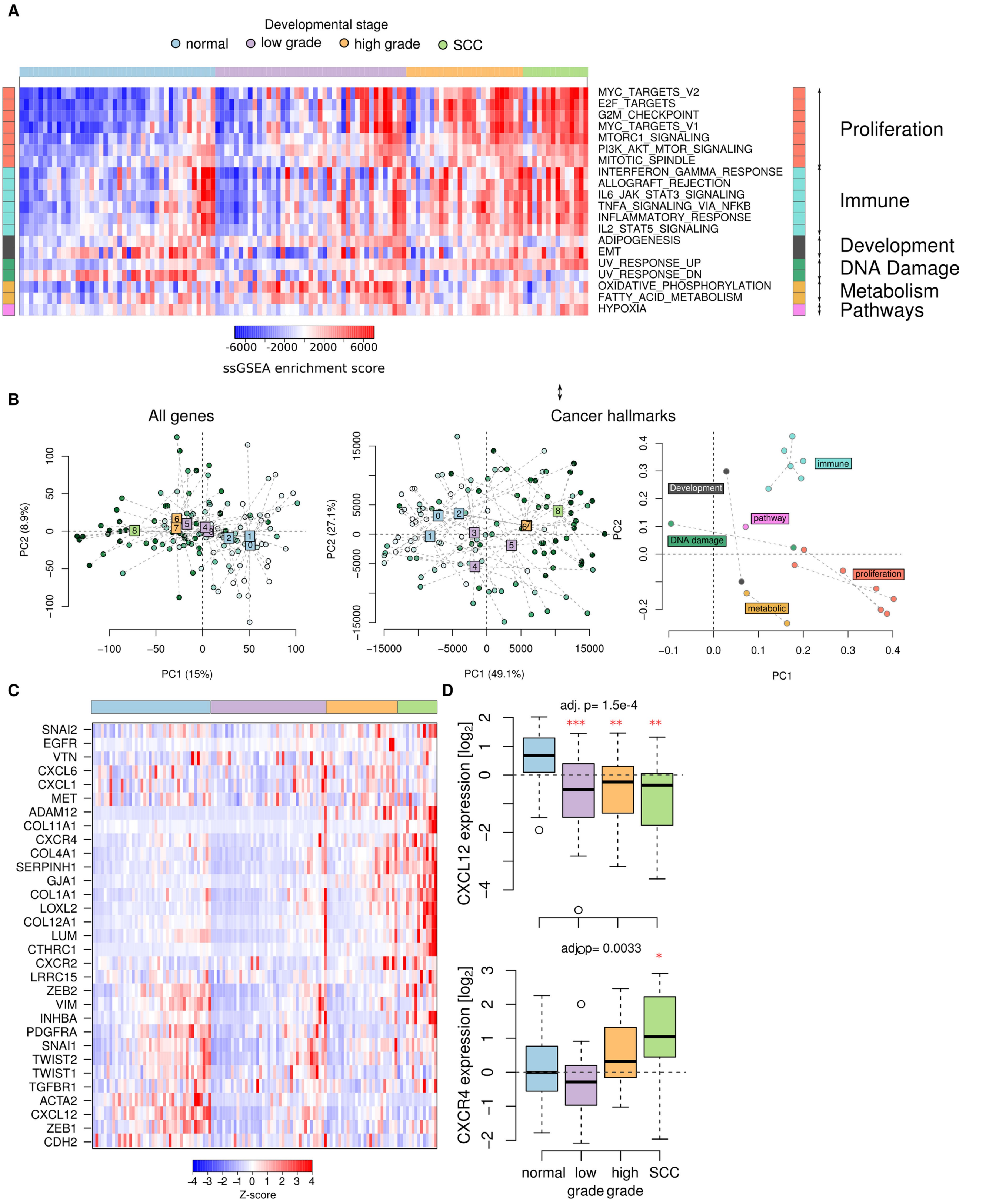
Code is available on Github at https://github.com/Precancer/SCC.


A
Squamous lung carcinogenesis
Normal bronchial tissue
Hyperplasia
Metaplasia
Mild dysplasia
Moderate dysplasia
Severe dysplasia
in situ carcinoma (CIS)
Squamous cell carcinoma (SCC)
Stage 0 (n=13)
Stage 1 (n=14)
Stage 2 (n=15)
Stage 3 (n=15)
Stage 4 (n=13)
Stage 5 (n=13)
Stage 6 (n=12)
Stage 7 (n=13)
Stage 8 (n=14)
Purified immune cells (n~1000)
Normal tissue (n~500)
Cancer cell lines (n=235)
Immune signatures
Fresh human bronchial biopsies (n=122)
Gene expression profile
Genes linked to progression
gene$_i$ = β * Stage+Z$_{patient}$*u+ ε
Normal
Low grade
High grade
SCC
Weighted correlation network analysis
Gene modules
Cancer hallmarks mSigDB
Enrichment Analysis
Immune cell deconvolution from gene expression
Immune cell abundance
FFPE tumor blocks (n = 110)
Multispectral immune characterizaiton
CD8
CD3
CK
DAPI
Temporal order of cancer hallmarks
Evolving immune response in lung carcinogenesis
B
white light bronchoscopy
fluorescence bronchoscopy

Extended Data Fig. 1 | Methodology for studying tumorigenesis. **a**, Across nine morphological stages of the development of lung SCC, freshly frozen samples were assayed for gene-expression profiling. The two methodological axes are visualized separately; one flow chart focuses on the detection of gene-expression patterns (left) and one flow chart focuses on the in-depth immune characterization (right) from gene expression and multispectral imaging. On the basis of gene co-expression, molecular phenotypes and gene modules were defined and functionally characterized. Immune-gene signatures and deconvolution methods were used for quantitative assessment of different immune-cell types. Relevant immune cells were investigated in more depth using multiplex immunohistochemistry and multispectral imaging. **b**, Bronchoscopy of in situ carcinoma tissue with white light (left) and with 400-nm wavelength illumination, under which the bronchial epithelium appears in green (right). Although it is difficult to distinguish the in situ carcinoma from normal tissue under white light, the in situ carcinoma displays a lack of green fluorescence compared to the normal epithelial tissue under fluorescence bronchoscopy.

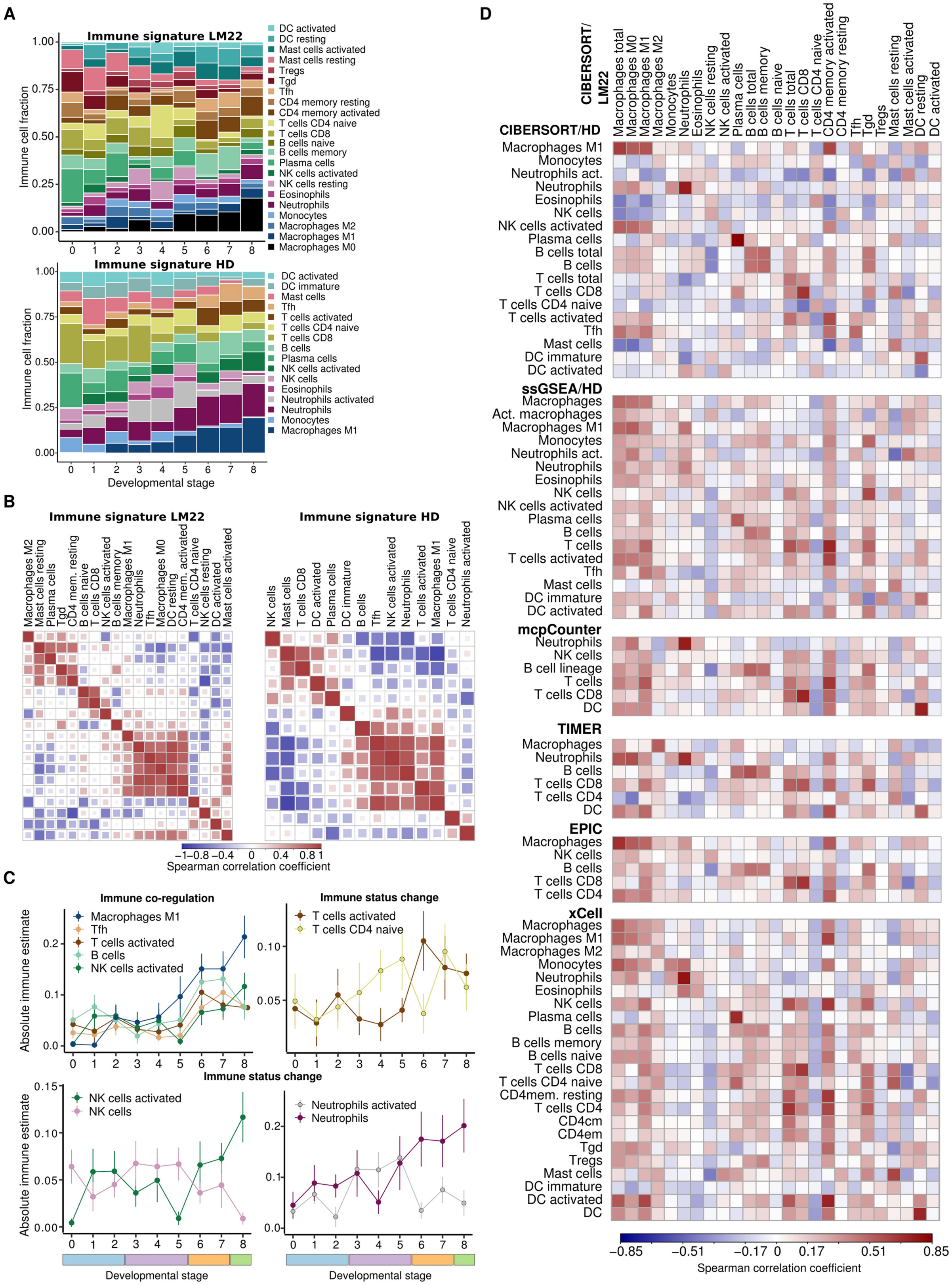


Extended Data Fig. 2 | Four molecular steps in carcinogenesis of lung SCC. The heat map shows genes that are associated with developmental stages identified using a linear mixed-effects model. Annotation bars are included for cancer history, sex and smoking history, all of which were used as fixed factors for the linear model, along with patient information (which was used as a random effect). Gene expression discerned four molecular groups: normal, low grade, high grade and SCC, on the basis of semi-supervised hierarchical clustering. Normal tissue with normal fluorescence, hypofluorescent normal tissue and hyperplasia lesions were subsumed under the category of normal tissue (stages 0, 1 and 2); metaplasia, mild dysplasia and moderate dysplasia were grouped as low grade (stages 3, 4 and 5); severe dysplasia and carcinoma in situ comprised the high-grade category; and the invasive stage was singled out as SCC.

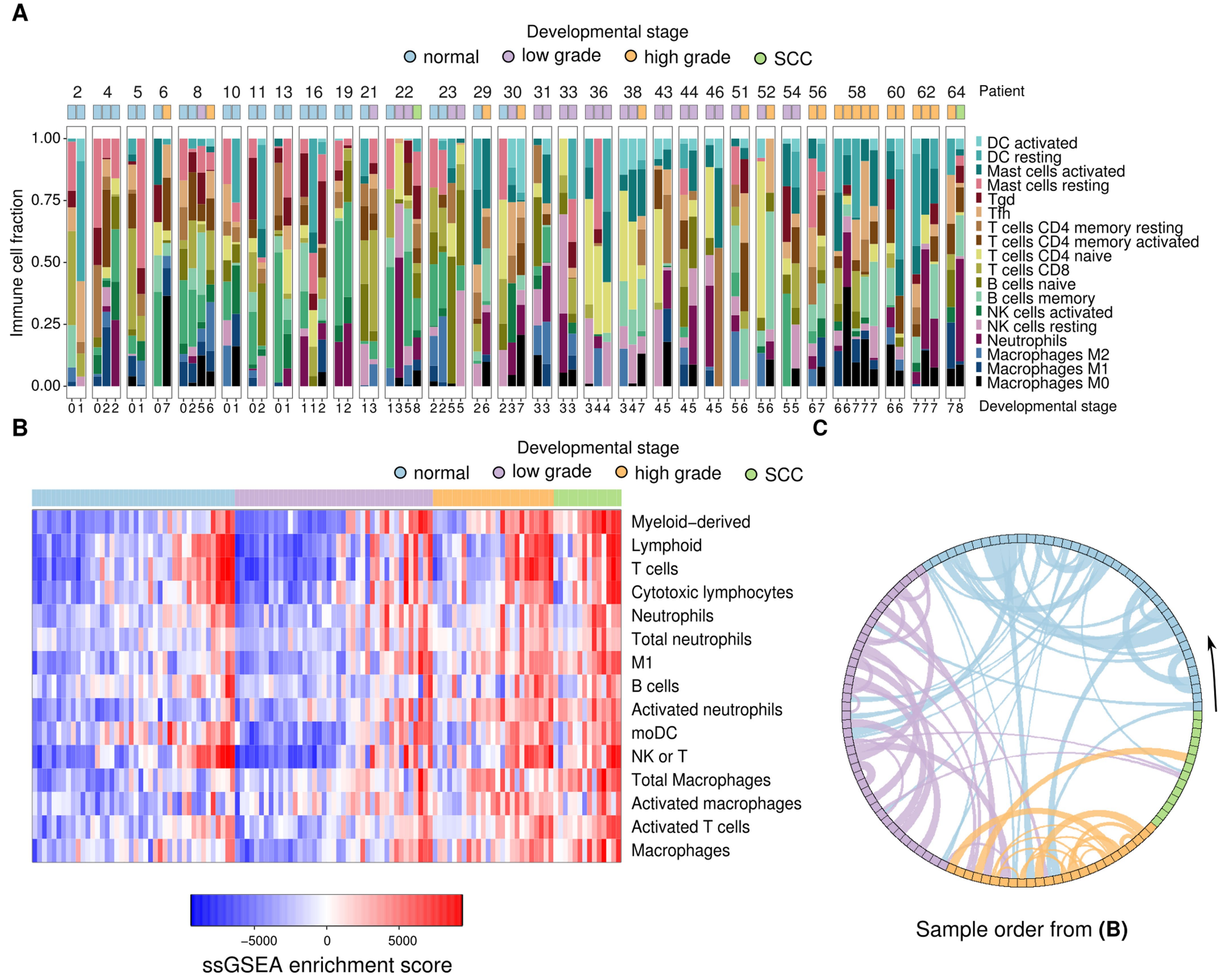


Extended Data Fig. 3 | See next page for caption.

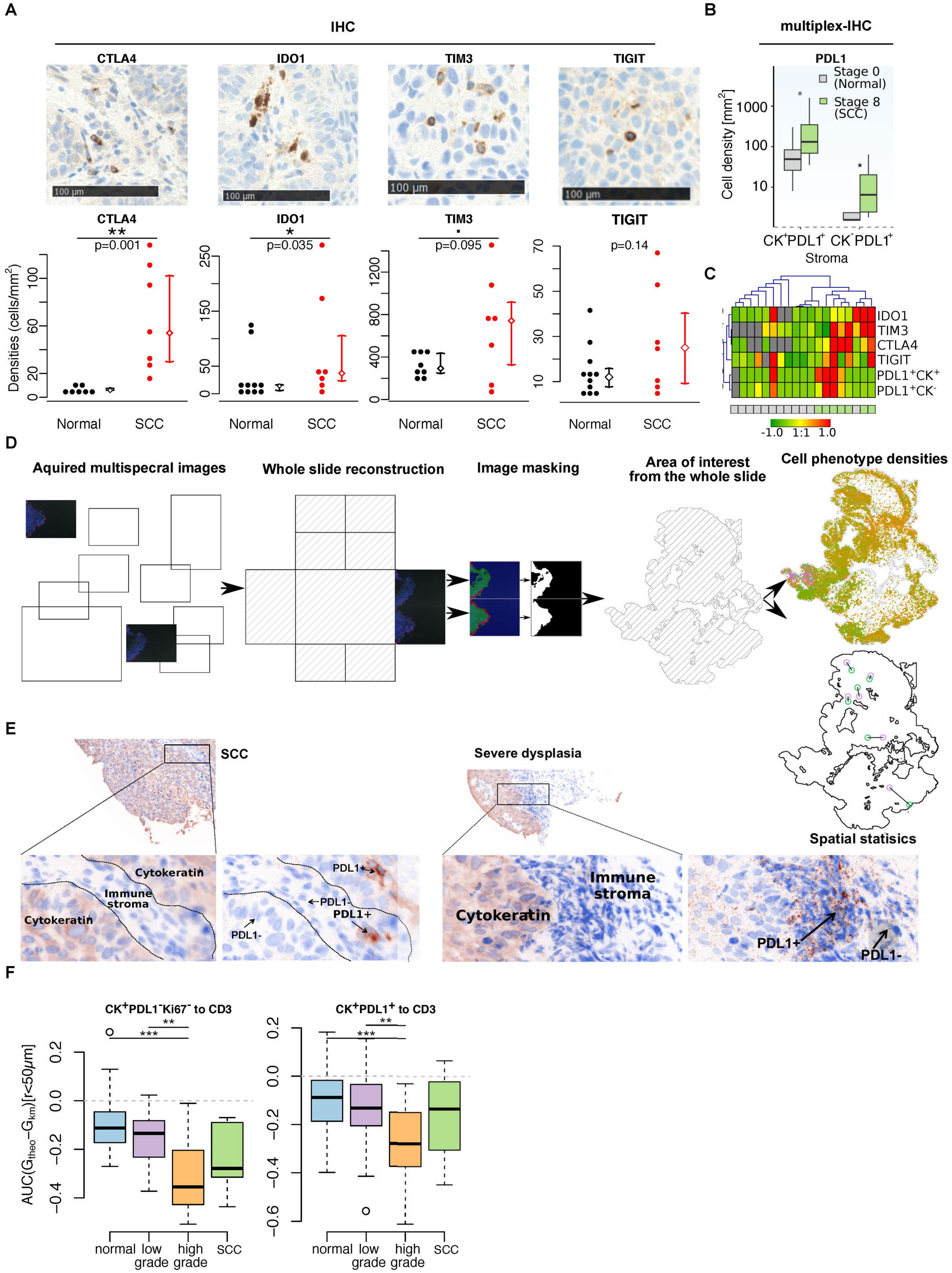
Extended Data Fig. 3 | Hallmarks of cancer. a, Single-sample gene-set enrichment analysis was performed on the full expression profile using cancer-hallmark gene definitions from mSigDb (v.6.2), independently of the gene modules. The heat map visualizes the enrichment scores from the single-sample gene-set enrichment analysis, in which the samples were ordered by their average enrichment scores for each molecular group individually. Only the cancer hallmarks that were significant with respect to the over-representation analysis of the gene modules in Fig. 1b are shown for validation. Three hallmark definitions associated with the ascending module at the highest adjusted *P* values ($P > 0.003$) are shown here (these are not shown in Fig. 1b): PI3K_AKT_MTOR_SIGNALING (proliferation), UV_RESPONSE_UP (DNA damage, confirmed by the UV_RESPONSE_DN in the descending module) and HYPOXIA (pathway). **b**, Left, randomized principal component (PC) analysis on the full expression profile shows a gradual continuum of expression changes from stage 0 to stage 8. Middle, randomized principal component analysis on enrichment scores for cancer hallmarks revealed distinct molecular steps. The cancer hallmarks explained up to 76.2% of the sample variability with the first two principal components (middle) as opposed to 24.9% variability explained by the full expression profile (left). Right, based on the principal component rotations, the hallmarks of proliferation, immune system, metabolism and the epithelial–mesenchymal transition each contribute to defining the developmental stages, as observed by their different directions of variability. **c**, Increase in expression of key genes and chemokines involved in the epithelial–mesenchymal transition, together with genes that overlap with the hallmark signature of the epithelial–mesenchymal transition and the 'SCC increase' gene module. The differential expression analysis of chemokines related to the epithelial–mesenchymal transition considered the confounding factors of smoking status, cancer history, sex, age and inter-patient variability. No differential expression across the developmental stages was found for CXCL1 and CXCL6, whereas the expression of CXCR2 was significantly increased in both the high-grade and SCC lesions. **d**, Only CXCR4 had a significant increase specific to SCC that was not observed in low- and high-grade lesions.



Extended Data Fig. 4 | Comparison of deconvolution approaches. **a**, Immune estimates derived from the CIBERSORT method, using the LM22 gene signature (all 22 cell types are presented) (top) or using on our in-house-developed immune signature (HD signature) (bottom). **b**, Symmetric correlation matrix of the average immune-cell abundance per developmental stage (Spearman correlation), estimated with CIBERSORT using the LM22 gene signature (left) and the HD signature (right). **c**, Immune co-regulation and immune-status shift derived from CIBERSORT estimates using the HD signature. **d**, Comparison of the mcpCounter, TIMER, EPIC and xCell methods for expression-based interrogation of the tumour immune infiltrates.

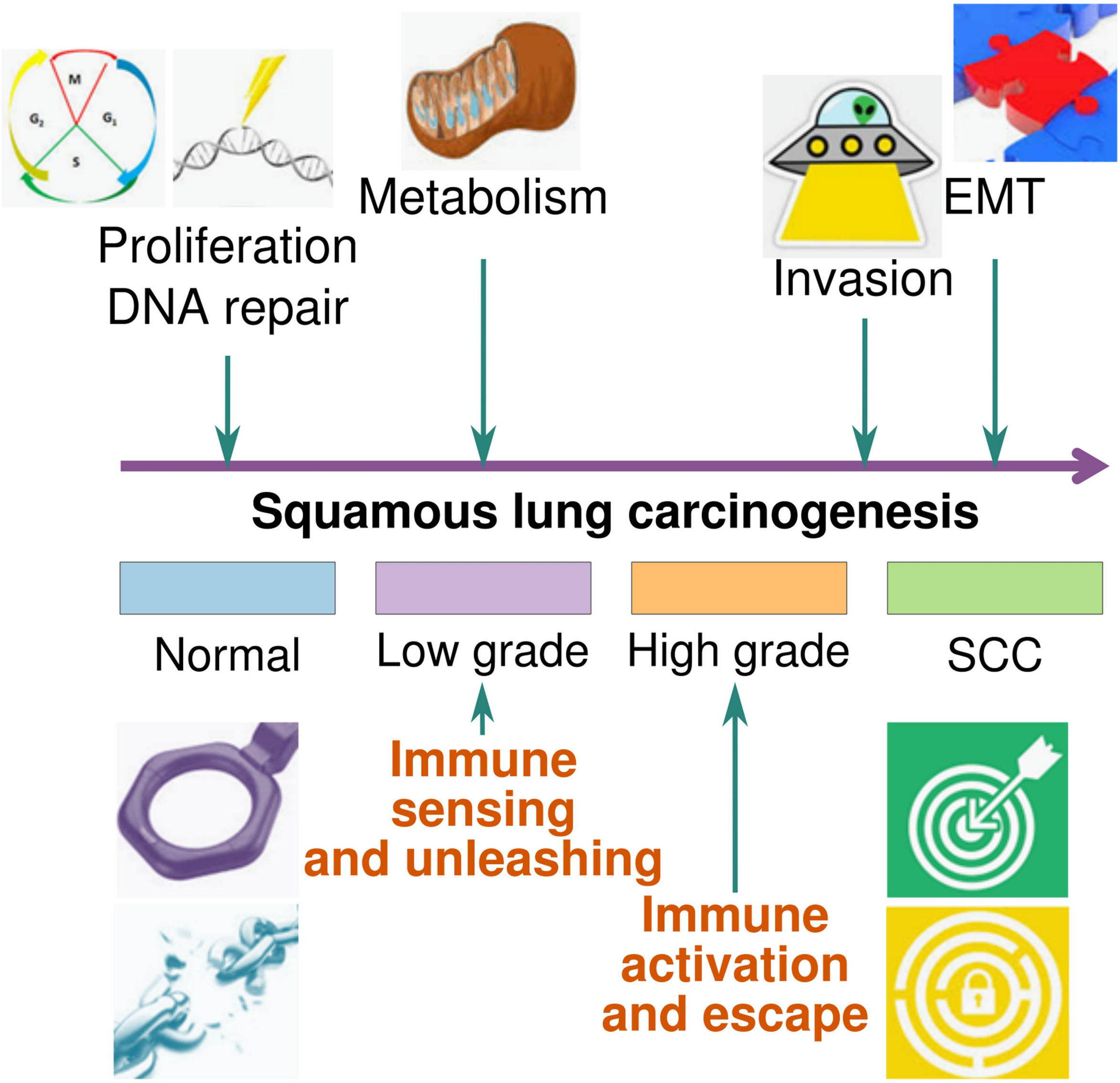


Extended Data Fig. 5 | Intrapatient immune heterogeneity. a, Stacked bars illustrate the relative cellular abundance of different immune-cell types estimated with CIBERSORT, in patients sampled for multiple grades. There are similar profiles for samples from the same grade, independent of the patient. **b**, Single-sample gene-set enrichment analysis was performed on the HD immune-cell signatures. The heat map represents one-dimensional clustering by immune-cell type, in which the samples of each molecular group were ordered by their average enrichment score. **c**, A chord diagram links the samples derived from the same patient. The order of the samples is preserved from **b**.

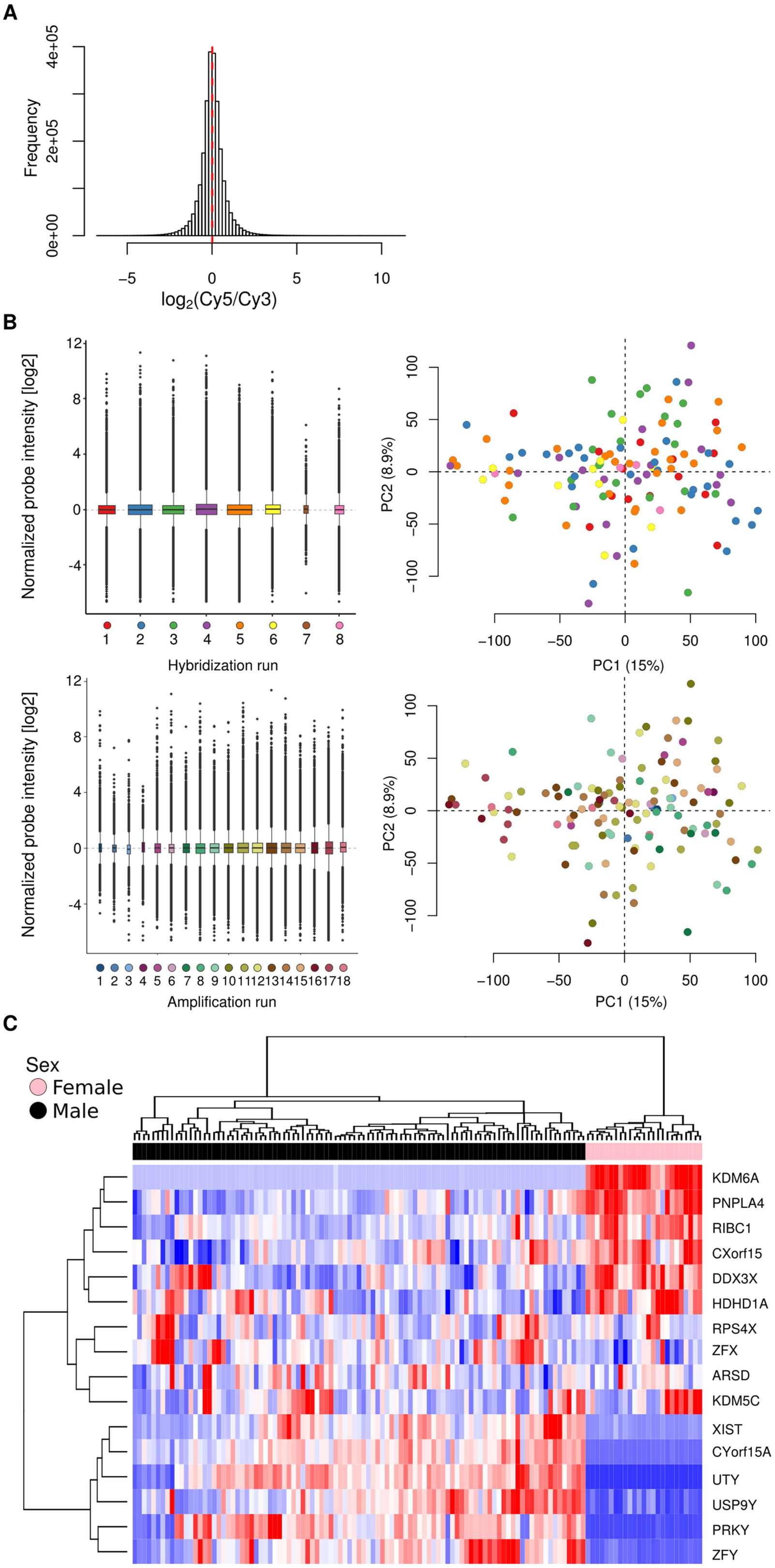


Extended Data Fig. 6 | See next page for caption.

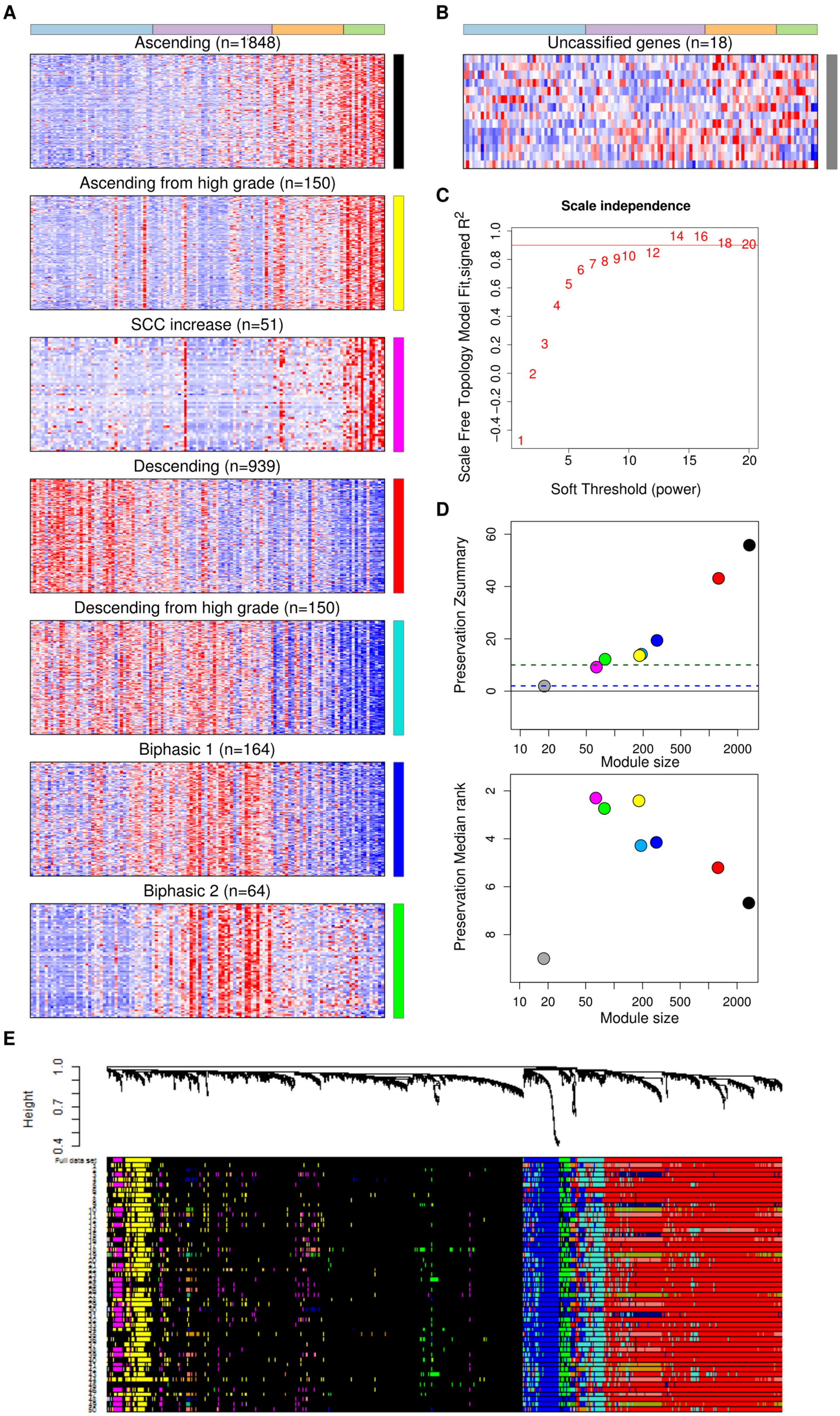
Extended Data Fig. 6 | Quantitative and spatial immune characterization through immunohistochemistry and multispectral imaging. a, Immunohistochemistry (IHC) quantification of the immune checkpoints CTLA4, IDO1, TIGIT and TIM3. Each of the tested markers was validated in SCC tissue (top). *P* values are derived from a non-parametric one-tailed Mann–Whitney *U* test used to validate increase in SCC compared to normal tissue. **b**, Comparison of PD-L1 densities between the stroma of normal tissue (stage 0) and SCC (stage 8), derived from multiplex immunohistochemistry. **c**, Clustering of normalized immunohistochemistry expression. **d**, A methodology for spatial analysis of multispectral imaging data. A whole slide is reconstructed from the individual images. On the basis of the tissue categorization, the images are masked to exclude the blank areas. Immune-cell densities are calculated as the number of cells per tissue area (m^2). Spatial localization is analysed within the selected region of interest. **e**, Representative examples of $CK^{-}PD\text{-}L1^{+}$ in both SCC and severe dysplasia. Single-positive PD-L1 cells ($CK^{-}PD\text{-}L1^{+}$) were generally immune cells that were located in the stroma, with morphological similarities to infiltrating macrophages.
f, We calculated the area between the theoretical and the empirical curve because deviations between the two can indicate clustering or segregation patterns (see Fig. 4b, bottom) to confirm that epithelial cells segregate from CD3 T cells in high-grade lesions, independently from the distance threshold of 25 μm.



Extended Data Fig. 7 | Graphical overview of lung carcinogenesis. Schematic illustrating the main stages of carcinogenesis for lung SCC.

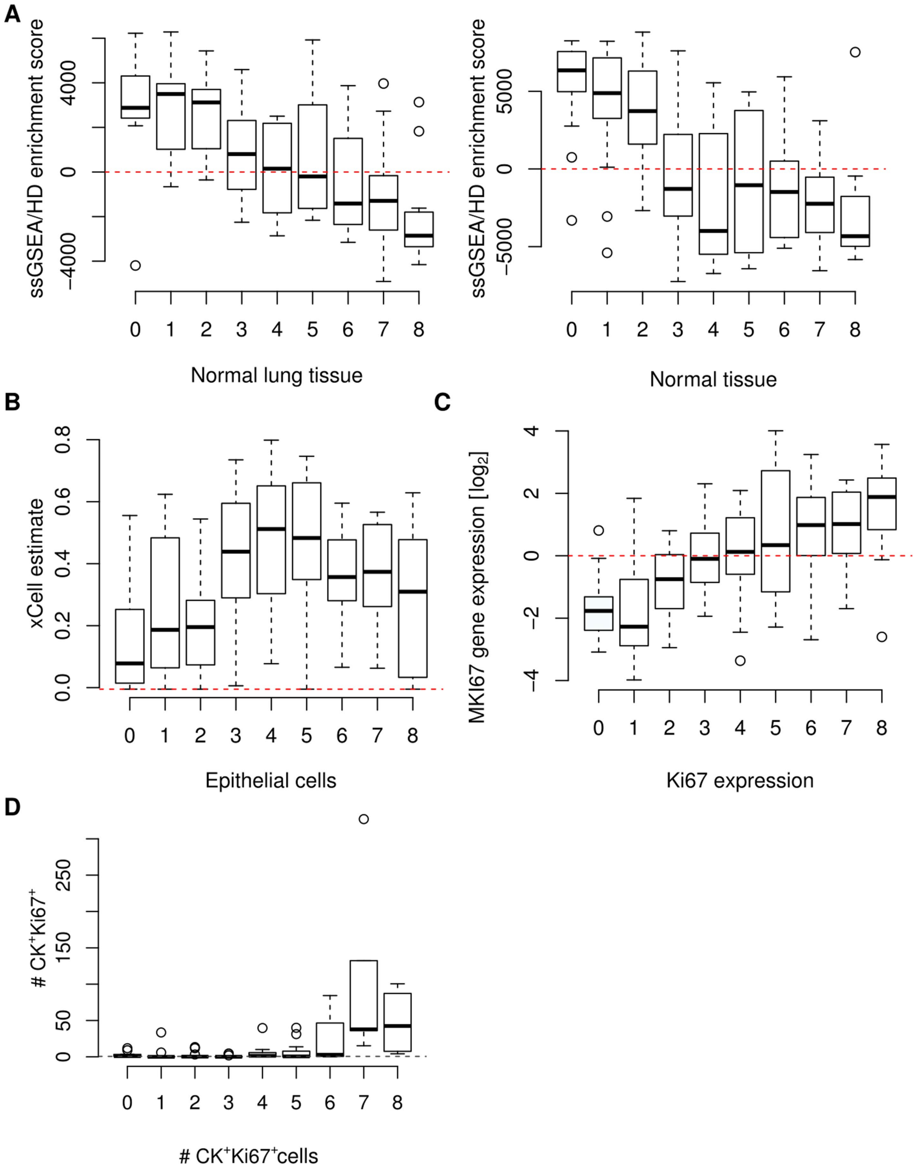


Extended Data Fig. 8 | Quality control of microarray data. a, Distribution of the relative abundance of each probe; that is, the ratio between the red and green colour intensity (Cy5/Cy3) for all probes across all patients ($\log_2$-transformed). **b**, Gene-expression distribution for each hybridization and amplification run (left). Using randomized principal component analysis, the samples were projected on the first two principal components and highlighted with different colours on the basis of their hybridization run (top right) and amplification run (bottom right). **c**, Classification of the samples is based on the expression of sex-chromosome genes, along with colour annotation for the sex of the patient.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Detection of gene co-expression modules. **a**, Expression of 4,734 genes that were associated with the nine developmental stages (linear mixed-C. effects model) is illustrated for each of the seven expression trajectories (that is, gene modules), detected using weighted network analysis of gene co-expression. **b**, Only 18 out of 4,734 genes were not assigned to a gene module and did not follow any of the seven illustrated expression patterns. **c**, A weighted network of genes is constructed by raising the adjacency matrix to a power. The value of the power for soft thresholding was chosen to be 12, as the lowest power term at which the network approximately fits a scale-free topology (red line $R^2 \leq 0.85$). The horizontal red line shows the squared correlation (R^2) cut-off of 0.85 recommended by the scale-free topology criterion. **d**, We randomly split the full dataset into a reference and test set, and evaluated the module preservation across the respective networks ($n = 50$ samples). The Z_{summary} statistic (top) provides evidence that the observed value of the preservation statistic is significantly higher than expected by chance (strong evidence if $Z_{\text{summary}} > 10$; weak-to-moderate if $Z_{\text{summary}} > 2$ and < 10; no evidence if $Z_{\text{summary}} < 2$). The grey module is unclassified, and expectedly showed no preservation ($Z_{\text{summary}} < 2$). All of the modules were preserved between the reference and the test datasets ($Z_{\text{summary}} > 10$), except in the SCC increase module ($Z_{\text{summary}} = 9.1$) which is also the smallest module with gene expression in small number of samples after resampling (an increase only in SCC). The median rank-preservation statistic showed, independently of the module size, that there is stronger preservation for all modules compared to the grey unclassified set of genes (bottom). **e**, The dendrogram derived from hierarchal clustering of the topology overlap matrix dissimilarity measure of the full dataset is shown in the top panel. The modules are defined on the basis of this dendrogram using a dynamic tree cut (top panel). We applied the same parameters for weighted gene-correlation network analysis on a resampled subset of the full dataset (a randomly selected two-thirds of the full dataset). The resampling was performed without replacement, which ensured proportional representation of each developmental stage. Colour rows indicate the module assignments obtained on the full dataset (first row) and on the resampled subsets of samples ($n = 50$ samples). All of the seven gene modules identified in the full dataset appear in almost every resampling, which indicates that the modules are stable.



Extended Data Fig. 10 | Tumour heterogeneity. a, Single-sample gene-set enrichment analysis using pannormal tissue, and normal lung tissue from HD gene signature. **b**, Estimates of epithelial-cell abundance derived from the xCell method (see Supplementary Information). **c**, Expression of the proliferation gene-marker *MKI67*. **d**, Expression of the proliferation gene-marker in CK$^+$ cells.

nature research

Corresponding author(s): Nature 2018-11-16768B

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Feature Extraction software version 9.5.3.1, Agilent Technologies
Data analysis	Software and algorithms Genespring GX, v7.3.1 software for two-color microarray data normalization (Agilent Technologies) Deconvolution of immune cell types from a mixed gene-expression signal CIBERSORT (Broad Institute) https://cibersort.stanford.edu/ single-sample Gene Set Enrichment Analysis (ssGSEA) R script (Broad Institute) http://software.broadinstitute.org/gsea clusterProfiler v3.2.14 (Bioconductor, Yu et al.) http://bioconductor.org/packages/release/bioc/html/clusterProfiler.html Weighted Gene Correlation Network Analysis (WGCNA) R package (CRAN) https://cran.r-project.org/web/packages/WGCNA/index.html Vectra 3.0 (PerkinElmer) perkinelmer.com Inform Cell Analysis (v2.3.0) (PerkinElmer) perkinelmer.com Phenochart 1.0 (PerkinElmer) perkinelmer.com HALO software (Indica Labs) http://www.indicalab.com/halo/ Spatstat v1.51 (CRAN) cran.r-project.org/web/packages/spatstat/ frozen Robust Multi-array Averaging (fRMA) (Bioconductor) http://bioconductor.org/packages/release/bioc/html/frma.html mcpCounter (Becht et al) https://github.com/ebecht/MCPcounter TIMER (Li et al.) http://cistrome.org/TIMER/ xCell (Aran et al.) http://xcell.ucsf.edu/

EPIC (Racle et al., GfellerLab) https://github.com/GfellerLab/EPIC
https://github.com/Precancer/SCC

Databases:
The Molecular Signatures Database (mSigDB, v6.2) http://software.broadinstitute.org/gsea/msigdb
Gene Expression Omnibus (GEO) https://www.ncbi.nlm.nih.gov/geo/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

https://www.ncbi.nlm.nih.gov/geo; access number GSE33479

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Based on the fact that high-grade lesions were rare and based on Dobbin et al.'s report 43, we included at least twelve biopsies from each histological stage.
Data exclusions	No data were excluded
Replication	The data could not be replicated or reproduced because the lesions are rare and because the amound of RNA available from the sample did not allow to duplicate the microarrays
Randomization	The samples were allocated to nine groups based on their histological classification
Blinding	As we were comparing the different histological categories, this information was known by the persons performing the statistical analyses. The identity of the patients was blinded by anonymisation of the samples.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
□	☒ Unique biological materials
□	☒ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology
☒	□ Animals and other organisms
□	☒ Human research participants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Unique biological materials

Policy information about availability of materials

Obtaining unique materials

The patients samples are unique biological materials that are not commercialy available : they were prospectively collected by the investigators for the purpose of this study

Antibodies

Antibodies used

Mouse monoclonal anti-CD3, Dako, Cat#M7254, clone F7.2.38
Mouse monoclonal anti-CD8, Dako, Cat#M7103, clone C8-144B P01732
Rabbit monoclonal anti-FoxP3, Cell Signaling Technology, Cat#98377S, clone D2W8E
Mouse monoclonal anti-Ki67, Dako, Cat#M7240, clone MIB-1 P46013
Rabbit monoclonal anti-PDL1, Biocare Medical, Cat#ACI3171A, C clone Cal10
Mouse monoclonal anti-CK, Dako, Cat#M3515 clone AE1/AE3
Mouse monoclonal anti-CD68, Dako, Cat#M0876, clone PG-M1
Rabbit policlonal anti-CD137, Abcam, Cat#ab197942
Mouse monoclonal anti-PD1, Ventana Roche, Cat#760-4448, Clone NAT105
Mouse monoclonal anti-NE, Dako, Cat#M0752, Clone NP57
Mouse monoconal anti-CTLA4, Bio SB, Cat#BSB 2884, Clone BSB-88
Mouse monoclonal anti-IDO1, Thermo Fisher Scientific, Cat# 14-9750-82, Clone V1NC3IDO
Rabbit monoclonal anti-TIGIT, Abcam, ab243903, Clone BLR047F
Rabbit monoclonal anti-TIM3, R&D Systems, Cat#MAB23652-100, Clone 2321C

Validation

All the primary antibodies are validated for immunohistochemistry on human tissues

Human research participants

Policy information about studies involving human research participants

Population characteristics

The developmental stage, the smoking status, the gender and the history of cancer were collected and used to perform the linear model. Other characteristics were potentialy available but not relevant for this study

Recruitment

Bronchial biopsies were collected between 2003 and 2007 at the Jules Bordet Institute, Brussels, Belgium, during fluorescence bronchoscopy in current or former smokers with a smoking exposure of ≥30 pack-years. Former smokers were defined as individuals who had quit smoking for more than 6 months.