

1 **Deciphering the genomic, epigenomic and transcriptomic landscapes of pre-invasive**  
2 **lung cancer lesions.**

3  
4 Vitor H. Teixeira<sup>1\*</sup>, Christodoulos P. Pipinikas<sup>1,2\*</sup>, Adam Pennycuick<sup>1\*</sup>, Henry Lee-Six<sup>3</sup>,  
5 Deepak Chandrasekharan<sup>1</sup>, Jennifer Beane<sup>4</sup>, Tiffany J. Morris<sup>2</sup>, Anna Karpathakis<sup>2</sup>, Andrew  
6 Feber<sup>2</sup>, Charles E. Breeze<sup>2</sup>, Paschalis Ntoliou<sup>1</sup>, Robert E. Hynds<sup>1,5,6</sup>, Mary Falzon<sup>7</sup>, Arrigo  
7 Capitanio<sup>7</sup>, Bernadette Carroll<sup>8</sup>, Pascal F. Durrenberger<sup>9</sup>, Georgia Hardavella<sup>8</sup>, James M.  
8 Brown<sup>1</sup>, Andy G. Lynch<sup>10,11</sup>, Henry Farmery<sup>10</sup>, Dirk S. Paul<sup>2</sup>, Rachel C. Chambers<sup>9</sup>, Nicholas  
9 McGranahan<sup>5</sup>, Neal Navani<sup>1,8</sup>, Ricky M. Thakrar<sup>1,8</sup>, Charles Swanton<sup>5,6</sup>, Stephan Beck<sup>2</sup>,  
10 Phillip Jeremy George<sup>8</sup>, Avrum Spira<sup>4,12</sup>, Peter J. Campbell<sup>3</sup>, Christina Thirlwell<sup>2</sup>, Sam M.  
11 Janes<sup>1,8#</sup>

12  
13 <sup>1</sup> Lungs for Living Research Centre, UCL Respiratory, University College London, London,  
14 U.K.

15 <sup>2</sup> Research Department of Cancer Biology and Medical Genomics Laboratory, UCL Cancer  
16 Institute, University College London, London, U.K.

17 <sup>3</sup> The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, U.K.

18 <sup>4</sup> Department of Medicine, Boston University School of Medicine, Boston, MA, U.S.A.

19 <sup>5</sup> CRUK Lung Cancer Centre of Excellence, UCL Cancer Institute, London, U.K.

20 <sup>6</sup> Translational Cancer Therapeutics Laboratory, The Francis Crick Institute, London, U.K.

21 <sup>7</sup> Department of Pathology, University College London Hospitals NHS Trust, London, U.K.

22 <sup>8</sup> Department of Thoracic Medicine, University College London Hospital, London, U.K.

23 <sup>9</sup> Center for Inflammation and Tissue Repair, UCL Respiratory, University College London,  
24 London, U.K.

25 <sup>10</sup> Computational Biology and Statistics Laboratory, Cancer Research UK Cambridge  
26 Institute, Cambridge, U.K.

27 <sup>11</sup> School of Medicine/School of Mathematics and Statistics, University of St Andrews, St  
28 Andrews, U.K.

29 <sup>12</sup> Johnson and Johnson Innovation, Cambridge MA

30

31 \* Drs. Teixeira, Pipinikas and Pennycuick contributed equally to this study.

32

33 # **Corresponding author:**

34 **Professor Sam M. Janes**

35 **Address:** Lungs for Living Research Centre, UCL Respiratory, University College London, 5

36 University Street, London, WC1E 6JF, U.K.

37 **Phone:** (+44) 020 3549 5979

38 **E-mail:** [s.janes@ucl.ac.uk](mailto:s.janes@ucl.ac.uk)

39

40

41



42 **Abstract**

43

44       The molecular alterations that occur in cells before cancer is manifest are largely  
45 uncharted. Lung carcinoma-in-situ (CIS) lesions are the pre-invasive precursor to squamous  
46 cell carcinoma. While microscopically identical, their future is in equipoise with half  
47 progressing to invasive cancer and half regressing or remaining static. The cellular basis of  
48 this clinical observation is unknown. Here, we profile the genomic, transcriptomic and  
49 epigenomic landscape of CIS in a unique patient cohort with longitudinally monitored pre-  
50 invasive disease. Predictive modelling identifies which lesions will progress with remarkable  
51 accuracy. We identify progression-specific methylation changes on a background of  
52 widespread heterogeneity, alongside a strong chromosomal instability signature. We  
53 observe mutations and copy number changes characteristic of cancer and chart their  
54 emergence, offering a window into early carcinogenesis. We anticipate this new  
55 understanding of cancer precursor biology will improve early detection, reduce over-  
56 treatment and foster preventative therapies targeting early clonal events in lung cancer.

57

58

59

60

61

62

63 **Introduction**

64

65 Lung cancer is the commonest cause of cancer death worldwide with 1.5 million  
66 deaths per year<sup>1</sup>. Lung squamous cell carcinoma (LUSC) is the most common subtype in  
67 parts of Europe and second in the U.S.A.<sup>2</sup> Before progression to invasive LUSC, there is  
68 step-wise evolution of ever more disordered pre-invasive lesions, ranging from mild and  
69 moderate dysplasia (low-grade lesions) to severe dysplasia and carcinoma-in-situ (CIS;  
70 high-grade lesions).<sup>3</sup> The accessibility of the proximal airways allows detection and  
71 monitoring of these lesions using high-resolution diagnostic approaches such as  
72 autofluorescence bronchoscopy (AFB)<sup>4</sup>. This technique enables the acquisition of tissue  
73 throughout the natural history of LUSC, providing an excellent model to study early  
74 tumorigenesis in human patients.

75 Clinically, the optimal management of pre-invasive airway lesions remains unclear,  
76 despite the availability of surgery, radiotherapy and ablative techniques<sup>5</sup>. AFB with biopsy  
77 allows assessment of the size, gross morphology and histopathology of pre-invasive lesions  
78 (**Fig. 1a, b**) but cannot distinguish lesions that will ultimately progress to invasive tumours  
79 from those that will spontaneously regress. As such, indiscriminate surgical resection of pre-  
80 invasive lesions or external beam radiotherapy probably represent over-treatment: lesions  
81 will spontaneously regress in 30% of cases, patient co-morbidity and poor lung function  
82 impart considerable risk, and the presence of field cancerization means independent lung  
83 cancers frequently emerge at sites outside resection or therapy margins.<sup>6</sup>

84 We reasoned that information on the future clinical trajectory of a pre-invasive lung  
85 lesion might be encoded in the genetic and epigenetic profile present at diagnosis. We  
86 therefore undertook a prospective cohort study of patients with pre-invasive squamous  
87 airway lesions. Patients were managed conservatively, undergoing surveillance AFB with  
88 biopsy and CT scanning every 4 and 12 months, respectively, with definitive cancer  
89 treatment only performed at the earliest pathological evidence of progression to invasive  
90 tumours (**Fig. 1a, b**).<sup>7</sup> When a CIS lesion either progressed to invasive cancer or regressed

91 to normal epithelium/low-grade disease, molecular profiling was performed on the preceding  
92 CIS biopsy from the same lesion – the ‘index biopsy’ (**Fig. 1c**). Index biopsies all  
93 demonstrated histologically and morphologically indistinguishable CIS and were classified as  
94 either ‘progressive’ or ‘regressive’. All such index CIS biopsies were subjected to a  
95 predetermined combination of transcriptomic, epigenetic and finally genomic profiling  
96 depending on DNA/RNA availability (**Fig. 1d; Table 1; Extended Data Fig. 1;**  
97 **Supplementary Table 1**).

98 Whilst molecular techniques have revolutionized our understanding of cancer biology,  
99 the key steps from normal cell to the point of cancer (uncontrolled growth and invasion)  
100 remain unclear. This is, to our knowledge, a unique collection of high-grade pre-invasive  
101 lung lesions for which prospective follow-up under conservative management enabled their  
102 natural history to declare.

103

## 104 **Results**

105

### 106 **Patient Characteristics**

107

108 Patients with pre-invasive lung cancer lesions were recruited through University College  
109 London Hospitals (UCLH) Early Lung Cancer Surveillance Programme (ELCSP). Full details  
110 of the surveillance protocol including eligibility criteria for patient inclusion have been  
111 previously described<sup>7</sup>. Briefly, the programme has recruited 140 patients to date with pre-  
112 invasive lung cancer lesions of varying histological grades. 129 index CIS biopsies were  
113 obtained from 85 patients and subjected to molecular analysis (**Supplementary Table 1**).  
114 Dependent on stored tissue quantity, in total, 51 samples from 42 patients underwent gene  
115 expression profiling; 87 samples from 47 patients underwent methylation profiling; and 39  
116 samples from 29 patients underwent whole genome sequencing. Methylation and gene  
117 expression datasets were divided into independent discovery and validation groups.

118 Clinical characteristics within each analysis group are shown in Table 1. In comparing

119 progressive and regressive samples, we found that progressive samples were associated  
120 with a higher pack-year smoking history in the methylation discovery group only ( $p < 0.01$ )  
121 and with increased age in the WGS group ( $p = 0.01$ ). No clinical differences were  
122 consistently observed across the different analysis groups.

123

#### 124 **Characterization of CIS genomic profiles**

125 We believe that the 39 CIS lesions are the first pre-invasive LUSC lesions to be  
126 whole-genome sequenced, so we compared the burden and spectrum of mutations in CIS  
127 with publicly available LUSC exome sequencing data from The Cancer Genome Atlas  
128 (TCGA). Due to differences between whole-genome and exome sequencing, only broad  
129 comparisons can be made. We observe a similar mutation burden and copy number profile  
130 between CIS samples and TCGA LUSC tumours (**Fig. 2**). There is congruency of type and  
131 prevalence of potential driver mutations, broadly defined as any mutation in a gene  
132 previously implicated as a driver of lung cancer, between CIS and LUSC samples<sup>8</sup>. We  
133 observe frequent alterations in *TP53*, *CDKN2A*, *SOX2* and *AKT2*, and less frequent  
134 alterations in *FAT1*, *KMT2D*, *KEAP1*, *EGFR* and *NOTCH1* in CIS lesions (**Fig. 2**;  
135 **Supplementary Table 2**). CIS mutational signatures<sup>9,10</sup> showed a strong tobacco-  
136 associated signal and were similar to those found in LUSC (**Extended Data Fig. 2**).

137 Marked aneuploidy was observed in CIS lesions, with somatic copy number  
138 alterations (CNAs) present across the genome (**Fig. 2**; **Extended Data Fig. 3**). The most  
139 frequent changes were associated with gain and amplification of multiple locations on distal  
140 3q: this is known to be the most common genomic aberration in LUSC<sup>11</sup>. Other recognised  
141 copy number associations identified in our data include gain/amplification in 5p, 8q and 19q  
142 and regions of loss/deletion in 3p, 4q, 5q, 8p, 9p and 13q.<sup>12-18</sup>

143 Whilst most CIS samples have the genomic appearance of neoplasms, we observe  
144 six lesions which show markedly lower mutational load and fewer copy number alterations  
145 than the others (**Extended Data Fig. 3**; PD21884c, PD21885a, PD21885c, PD21904d,  
146 PD38317a, PD38319a). These samples have very few genomic changes, despite being CIS

147 histologically. All of these six samples regressed to normal epithelium or low-grade dysplasia  
148 on subsequent biopsy. Four further samples met this end-point for regression, despite  
149 widespread mutational and copy number changes. However, with longer follow up one of  
150 these cases developed CIS recurrence (**Extended Data Fig. 4a**; PD21893a), and two  
151 developed invasive cancer on further surveillance (**Extended Data Fig. 4b,c**; PD21884a,  
152 PD38326a). Only one sample, PD21908a, showed sustained clinical regression after 9  
153 years of follow up despite widespread molecular changes.

154 All but one progressive sample and all highly mutated regressive samples showed  
155 amplification in a small region of distal 3q (chr3:172516434-178440382). This region  
156 contains the gene *ECT2*, a regulator of cytokinesis which is associated with chromosomal  
157 instability. Progressive sample PD38320a had little change outside this region and did not  
158 harbour a *TP53* mutation, suggesting that this amplification may be a crucial early event in  
159 LUSC tumorigenesis.

160 We compared genomic features between the 29 progressive and 10 regressive  
161 lesions. The three samples which showed evidence of progression after meeting our end-  
162 point for regression were excluded from this analysis. Comparisons of mutation burden  
163 between progressive and regressive lesions were performed by mixed effects modelling,  
164 allowing us to account for samples that come from the same patient. Even after correcting  
165 for patient age, smoking history and sample purity, progressive lesions had more somatically  
166 acquired mutations than those from regressive lesions, across base substitutions ( $p < 0.001$ ),  
167 indels ( $p = 0.018$ ), structural variants ( $p < 0.001$ ) and copy number changes ( $p < 0.001$ )  
168 (**Extended Data Fig. 5a-d**). When the analysis was restricted only to substitutions that were  
169 fully clonal in each lesion, there were still substantially more substitutions in progressive than  
170 regressive lesions ( $p < 0.001$ ) (**Extended Data Fig. 5e**), suggesting that the increase in  
171 mutation burden is not due to recent subclonal diversification in progressive lesions. All the  
172 mutational processes (or signatures<sup>9,10</sup>) identified in the CIS lesions contribute to the excess  
173 of mutations in progressive compared to regressive samples; however, only tobacco-  
174 associated signature 4 showed proportionally more mutations ( $p = 0.017$ ) (**Extended Data**

175 **Fig. 2f-j).** Progressive lesions contained more putative driver mutations than regressive  
176 lesions ( $p=0.001$ ) (**Extended Data Fig. 5h; Supplementary Table 2**). Importantly, no single  
177 cancer mutation perfectly discriminated between progressive and regressive lesions.

178 Within the biopsied lesions, clonal architecture was similar between progressive and  
179 regressive lesions (**Extended Data Fig. 5e-g**). For four patients in whom we sequenced  
180 multiple progressive lesions, the lesions shared many somatic mutations despite their  
181 different locality in the bronchial tree, indicating their probable derivation from a common  
182 ancestral clone. By contrast, multiple regressive lesions from two further patients did not  
183 share common mutations and so are likely to have arisen independently (**Extended Data**  
184 **Fig. 6**). There were no differences in telomere lengths between progressive and regressive  
185 lesions ( $p=0.59$ ) (**Extended Data Fig. 5i**).

186

### 187 **CIS transcriptomic and epigenetic profiles**

188 Gene expression microarrays were performed on a discovery set of 17 progressive  
189 and 16 regressive CIS lesions. We identified 1335 genes with significant expression  
190 changes ( $FDR < 0.01$ ); 657 genes were up-regulated and 678 down-regulated in progressive  
191 CIS lesions (**Fig. 3a and Supplementary Table 3**).

192 Differential analysis of methylation profiles was performed on a discovery set of 26  
193 progressive, 11 regressive and 23 control samples. Widespread methylation changes were  
194 observed with 12,064 differentially methylated positions (DMPs), associated with 2,695  
195 genes, at which methylation was significantly different between progressive and regressive  
196 samples ( $FDR < 0.01$ ;  $|\Delta\beta| > 0.3$ ). 6,314 DMPs were hypermethylated and 5,750  
197 hypomethylated in progressive CIS (**Fig. 3b and Supplementary Table 3**). 260 differentially  
198 methylated regions (DMRs) were identified, of which 151 (58%) overlap with DMRs between  
199 TCGA cancer and control data (**Extended Data Fig. 7**). Finally, we identified 36,620  
200 differentially variable positions (DVPs) for which probe variance was markedly different  
201 between progressive and regressive groups.

202 Of the 1335 genes identified, *TPM3*, *PTPRB*, *SLC34A2*, *KEAP1*, *NKX2-1*, *SMAD4*

203 and *SMARCA4* have previously been implicated as potential lung cancer drivers  
204 (**Supplementary Table 4**). Regarding methylation, the potential driver genes *NKX2-1*,  
205 *TERT*, *DDR2*, *LRIG3*, *CUX1*, *EPHA3*, *CSMD3*, *MET*, *ZNF479*, *GRIN2A*, *PTPRD*, *NOTCH1*,  
206 *CD74*, *NSD1* and *CDKN2A* contain at least one significant DMP. Several genes which are  
207 significant in our gene expression analysis are also identified in our methylation data,  
208 including multiple genes in the homeobox family (*HOXC8*, *HOXC9*, *HOXC10*, *HOXD10*,  
209 *HOXA11AS*), previously implicated as an early epigenetic event in multiple cancers<sup>19</sup>. *NKX2-*  
210 *1* (*TTF-1*) is the only putative driver gene to be identified in both gene expression and  
211 methylation analyses, and is also a member of the homeobox family. It is hypermethylated  
212 and underexpressed in progressive samples compared to regressive. This gene is widely  
213 used in diagnosis of lung adenocarcinoma and both underexpression and hypermethylation  
214 have been implicated in the development of this disease<sup>20,21</sup>. *NKX2-1* loss has been shown  
215 to drive squamous cancer formation in combination with *SOX2* overexpression<sup>22</sup>; focal gains  
216 in the 3q region containing *SOX2* are commonly observed in progressive CIS (**Extended**  
217 **Data Fig. 4**).

218 Principal component analysis of all gene expression and methylation data showed a  
219 clear distinction between the progressive and regressive subgroups ( $p=0.0017$  and  
220  $p=6.8 \times 10^{-25}$ , respectively) (**Fig. 3c,d**). In the methylation dataset, the regressive lesions  
221 closely clustered with the control normal epithelial cells. A history of chronic obstructive  
222 pulmonary disease (COPD) had an effect on case segregation ( $p=1.2 \times 10^{-5}$ ) but all other  
223 clinical and technical variables analysed, including smoking status and history of lung  
224 cancer, had no effect (**Extended Data Fig. 8a-f**). This was also the case for PCA analysis of  
225 the gene expression data (**Extended Data Fig. 8g-k**).

226 For methylation, one control and four regressive cases clustered with the progressive  
227 cases (**Fig. 3d**). Three of the four mis-classified regressive cases were subjected to whole-  
228 genome sequencing and were found to have more copy number alterations than other  
229 regressive samples (PD21884a, PD21893a, PD21908a). Two of these correspond to the  
230 samples discussed above, which showed signs of progression after meeting the clinical end

231 point of regression (**Extended Data Fig. 4**). For the control bronchial epithelium sample that  
232 was classified with the progressive lesions, CIS was detected in a biopsy specimen 12  
233 months later from the same site. Thus, although we have formally treated these cases as  
234 mis-classifications, it is likely that the molecular data underpinning the apparent errors  
235 indicate a cellular phenotype that is not consistent with a straightforward regressive lesion.

236

237

### 238 **Molecular signatures predict CIS outcome**

239 The ability to predict if a pre-invasive lesion will progress to cancer has important  
240 clinical implications. For gene expression, we used the above pre-defined discovery set to  
241 define our classifier (n=33; 17 progressive, 16 regressive; 10-fold cross-validation applied).  
242 This was applied to a separate validation set (n=18; 10 progressive, 8 regressive). All  
243 samples in the validation set were classified correctly. When applied to external data from  
244 TCGA (n=551: 502 LUSC, 49 control), our 291-gene model was able to classify LUSC vs  
245 control samples with AUC=0.81 (**Fig. 4a-c; Extended Data Fig. 9**).

246 An analogous analysis was performed for methylation using a discovery set of 60  
247 samples and a validation set of 27 samples. This classified validation samples with  
248 AUC=0.99 and classified external TCGA samples (n=412: 370 LUSC, 42 controls) into  
249 LUSC vs controls with AUC=0.99, based on a 141-DMP classifier (**Extended Data Fig. 10a-**  
250 **i**).

251 We observed an increased number of methylation probes with intermediate  
252 methylation in TCGA LUSC cancer vs TCGA control samples (**Fig. 4d**), reflecting  
253 methylation heterogeneity in these samples. We therefore developed a methylation  
254 heterogeneity index (MHI), defined as the number of probes per sample with  $t_{lo} < \beta < t_{hi}$ .  
255 Optimization based on our discovery set of 26 progressive and 11 regressive samples  
256 defined values of  $t_{lo} = 0.26$  and  $t_{hi} = 0.88$ . Control samples were not used in this analysis.  
257 This model classified progressive vs regressive CIS samples in our validation set with  
258 AUC=0.74 and TCGA LUSC vs TCGA control samples with AUC=0.96 (**Fig. 4e; Extended**



259 **Data Fig. 10j-n).** Multivariate logistic regression in our CIS cohort demonstrated that this  
260 index was a predictor of progression status ( $p=0.017$ ); previous history of lung cancer was  
261 also significantly associated ( $p=0.02$ ), whereas smoking status, COPD status, age and  
262 gender were not.

263         Given the widespread nature of methylation changes, we hypothesised that this  
264 increase in heterogeneity may be a genome-wide process rather than specific to functional  
265 pathways. To test this theory, we assessed the predictive value of MHI calculated from a  
266 sample of 2,000 probes, randomly selected from across the genome. Running 10,000  
267 simulations with each using a different random sample of 2,000 probes gave a mean AUC  
268 for TCGA LUSC vs TCGA control of 0.95 (95% CI 0.92-0.98) (**Fig. 4f**), and for progressive  
269 vs regressive CIS of 0.75 (95% CI 0.69-0.82) (**Extended Data Fig. 10n**). These results are  
270 similar to those obtained using the entire set of 450,000 probes, suggesting that methylation  
271 heterogeneity is a genome-wide process. However, these AUC values are lower than those  
272 obtained from our predictive model based on just 141 differentially methylated positions,  
273 suggesting that specific methylation changes are also important, on this background of  
274 generalised change.

275         To build a predictive classifier based on copy number, we used copy number derived  
276 from methylation data to increase sample size and classified 46 of 54 samples correctly  
277 (**Extended Data Fig. 9g-i**). The 154 predictive cytogenetic bands that we identified overlap  
278 with, but are not limited to, a model previously proposed by van Boerdonk *et al.*. Our model  
279 replicated their results, classifying 24/24 regressive samples and 9/12 progressive samples  
280 correctly<sup>23</sup> (**Extended Data Fig. 9j-l**). When applied to external data from TCGA ( $n=763$ : 524  
281 LUSC, 239 control), our model was able to classify LUSC vs control samples with  $AUC=0.98$   
282 (**Extended Data Fig. 9m-o**).

283         We performed further analyses using only one sample per patient to demonstrate that  
284 our results are not dependent on multiple sampling. The first available sample for each  
285 patient was selected, with CIS samples prioritized over control samples for methylation data.  
286 Results are similar to our analysis above, validating our initial results (**data not shown**).

287           Although we cannot fully exclude that lesions meeting our end point for regression will  
288 progress in future, most patients in this cohort now have several years of follow up. Of 35  
289 regressive lesions undergoing molecular profiling (**Supplementary Table 1**), mean follow up  
290 was 67 months (median 57 months, range 11-150 months).

291

## 292 **CIN is an early marker of progression to cancer**

293           To investigate possible drivers of tumorigenic progression, we performed a  
294 differential analysis of gene expression data between the progressive and regressive  
295 groups. 5 of the top 100 genes identified have been previously associated with chromosomal  
296 instability (CIN)<sup>24</sup>, as defined by the previously published CIN70 signature<sup>25</sup> (*ACTL6A*,  
297 *ELAVL1*, *MAD2L1*, *NEK2*, *OIP5*). All five are up-regulated in progressive compared with  
298 regressive samples. CIN-related genes can predict progression (**Fig. 5a**); *NEK2* expression  
299 alone predicts progression with AUC=0.93 (**Fig. 5b**).

300           Pathway analysis was performed using the *gage* Bioconductor package<sup>26</sup> to compare  
301 our differentially expressed genes to KEGG gene sets. The CIN70 gene set was the most  
302 significant gene set identified (adjusted p value  $8.9 \times 10^{-32}$ ; up-regulated in progressive  
303 group), suggesting a role in early tumorigenesis. Cell cycle and DNA repair pathways were  
304 also implicated (**Fig. 5c**; **Supplementary Table 5**). Results were similar when cell-cycle  
305 associated genes were removed from the CIN70 signature, suggesting that this is a genuine  
306 CIN signal rather than a marker of proliferation.

307           Performing similar differential analysis of differentially methylated probes found  
308 widespread changes. The top probes identified were associated with cancer-associated cell  
309 signalling pathways, including TGF-beta, WNT and Hedgehog, as well as cell cycle and CIN-  
310 associated genes (**Fig 5d**).

311           This CIN signal is consistent with the observed pattern of widespread copy number  
312 change (**Fig. 2**). Overall copy number variation for a sample, as measured by Weighted  
313 Genome Integrity Index (wGII)<sup>27</sup>, correlates with mean CIN-associated gene expression of  
314 that sample (Pearson  $r^2=0.473$ ) (**Extended Data Fig. 5j**). We also observe a correlation

315 between local copy number of a gene and expression of that gene, consistent with previous  
316 results<sup>28,29</sup>.

317

318

319

320

## 321 **Discussion**

322 In summary, we have delineated changes in the genomic architecture, genome-wide  
323 gene expression and DNA methylation of pre-invasive cancers with known histological  
324 evidence of subsequent disease progression or regression. The CIS genome shares many  
325 of the hallmarks of advanced, invasive LUSC but marked genomic, transcriptomic and  
326 epigenetic differences exist between lesions that are benign and those that will progress to  
327 cancer. Our data demonstrate the potential use of these differences in predicting outcome  
328 over current clinical practice.

329 Among the strongest pathways associated with progression is chromosomal  
330 instability, defined as a high rate of gain or loss of whole (or parts of) chromosomes. CIN is  
331 implicated in many human cancers, including lung, and has been suggested both as a  
332 prognostic marker and therapeutic target<sup>30,31</sup>. Regressive lesions do not have the wholesale  
333 genomic instability of those that will progress and their epigenetic and transcriptional profiles  
334 more closely resemble normal bronchial epithelium than invasive cancers. Despite this, CIS  
335 lesions that spontaneously regress are genuine neoplasms; they harbour many somatic  
336 mutations, which can include known potential driver mutations. The mechanism of  
337 regression remains mysterious: it is unclear whether clones become exhausted and die out,  
338 potentially abetted by immune surveillance, or whether clones persist but phenotypically  
339 revert to an architecturally normal, physiological epithelium. Likewise the mechanisms of  
340 CIN are not well understood; our study paves the way for investigation of these CIN-  
341 associated genes in model systems to elucidate their role.

342           We present here the first major whole genome sequencing data of pre-invasive lung  
343 lesions. We acknowledge that, despite using the world's largest cohort of such lesions, the  
344 study remains underpowered to detect less common genomic alterations. Expanding our  
345 knowledge in this area will require a major international collaboration. Likewise we  
346 acknowledge that whilst our predictive signatures demonstrate the power of molecular data  
347 in guiding management decisions, a prospective clinical trial using predictors derived from  
348 our data will be required before clinical use. Again, international collaboration will be  
349 required to develop an appropriately powered trial.

350           Despite these limitations, our data offer the first insight into the molecular map of  
351 early lung squamous cancer pathogenesis, foretelling an era in which molecular profiling will  
352 enable personally tailored therapeutic decisions for patients with pre-invasive lung disease.

353

354

355

356 **Acknowledgements**

357 We thank all the patients who participated in this study and Kerra Pearce, George  
358 Chennel, David Chambers, Paul Mercer and Kate Gowers for technical help and proof  
359 reading. We thank Pamella Rabbitts, Anindo Banerjee and Cathy Read for their early  
360 development of the study. The results published here are in part based on data generated  
361 by a TCGA pilot project established by the National Cancer Institute and National Human  
362 Genome Research Institute. Information about TCGA and the investigators and institutions  
363 that constitute the TCGA research network can be found at <http://cancergenome.nih.gov>.

364 Grants support: S.M.J. and P.J.C. are Wellcome Trust Senior Fellows in Clinical  
365 Science. S.M.J. is also supported by the Rosetrees Trust, the Welton Trust, the Garfield  
366 Weston Trust , the Stoneygate Trust and UCLH Charitable Foundation. V.T., C.P., R.E.H.  
367 and S.M.J. have been funded by the Roy Castle Lung Cancer Foundation. A.P. is funded by  
368 a Wellcome Trust clinical PhD training fellowship. H.L.-S. is funded by the Wellcome Trust  
369 Sanger Institute non-clinical PhD studentship. C.T. was a CRUK Clinician Scientist. This  
370 work was partially undertaken at UCLH/UCL who received a proportion of funding from the  
371 Department of Health's NIHR Biomedical Research Centre's funding scheme (S.M.J. and  
372 N.N.). S.M.J. and C.S. are part of the CRUK Lung Cancer Centre of Excellence. A.S., C.S.  
373 and S.M.J. are supported by Stand Up to Cancer. The funders had no role in study design,  
374 data collection and analysis, decision to publish or preparation of the manuscript.

375

376

377 **Author Contributions**

378 V.H.T, C.P.P. and A.P. contributed equally to this work. S.M.J., P.J.C., V.H.T., A.P.,  
379 R.E.H., H.L.-S. and C.P.P. co-wrote the manuscript. S.M.J., P.J.C., C.T., V.H.T., and C.P.P.  
380 conceived the study design. S.M.J., P.J.C., C.T., V.H.T., C.P.P. and A.P. designed the study  
381 protocols. V.H.T. performed gene expression, qPCR and LCM experiments, analysed and  
382 integrated clinicopathological data and gene expression data. C.P.P. performed methylation  
383 and LCM experiments, analysed and integrated clinicopathological data and methylation

384 data. A.P. analysed and integrated clinicopathological data, WGS data, gene expression  
385 data and methylation data. H.L.-S., A.G.L. and H.F. analysed WGS data. D.C. and P.N.  
386 performed LCM experiments. J.B. analysed gene expression data. T.J.M., A.K., A.F., C.E.B.  
387 and D.S.P. analysed methylation data. M.F. and A.C. conducted the pathological review.  
388 P.J.G., B.C., N.N., G.H., J.M.B. and R.M.T. performed bronchoscopies and collected the CIS  
389 and control biopsies. P.F.D. performed histological experiments. R.E.H., R.C.C., N.M., C.S.,  
390 S.B. and A.S. gave advice and reviewed the manuscript. S.M.J. provided overall study  
391 oversight.

392

### 393 **Competing Interests Statement**

394 The authors declare the following competing interests:

395         A.S. is an employee of Johnson and Johnson. Discoveries within this manuscript  
396 have led S.M.J. to lead on Patent Applications 1819453.0 and 1819452.2 filed with the UK  
397 Intellectual Property Office through UCL Business PLC.

398

399

400

401

402

403

## References

- 405 1 Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Global cancer statistics, 2002. *CA Cancer*  
406 *J Clin* **55**, 74-108 (2005).
- 407 2 Torre, L. A., Siegel, R. L. & Jemal, A. Lung Cancer Statistics. *Advances in experimental*  
408 *medicine and biology* **893**, 1-19, doi:10.1007/978-3-319-24223-1\_1 (2016).
- 409 3 Nicholson, A. G. *et al.* Reproducibility of the WHO/IASLC grading system for pre-  
410 invasive squamous lesions of the bronchus: a study of inter-observer and intra-  
411 observer variation. *Histopathology* **38**, 202-208 (2001).
- 412 4 van der Heijden, E. H., Hoefsloot, W., van Hees, H. W. & Schuurbijs, O. C. High  
413 definition bronchoscopy: a randomized exploratory study of diagnostic value  
414 compared to standard white light bronchoscopy and autofluorescence  
415 bronchoscopy. *Respir Res* **16**, 33, doi:10.1186/s12931-015-0193-7 (2015).
- 416 5 Thakrar, R. M., Pennycuick, A., Borg, E. & Janes, S. M. Preinvasive disease of the  
417 airway. *Cancer Treat Rev* **58**, 77-90, doi:10.1016/j.ctrv.2017.05.009 (2017).
- 418 6 Pipinikas, C. P. *et al.* Cell migration leads to spatially distinct but clonally related  
419 airway cancer precursors. *Thorax* **69**, 548-557, doi:10.1136/thoraxjnl-2013-204198  
420 (2014).
- 421 7 Jeremy George, P. *et al.* Surveillance for the detection of early lung cancer in patients  
422 with bronchial dysplasia. *Thorax* **62**, 43-50, doi:10.1136/thx.2005.052191 (2007).
- 423 8 Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183,  
424 doi:10.1038/nrc1299 (2004).
- 425 9 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature*  
426 **500**, 415-421, doi:10.1038/nature12477 (2013).
- 427 10 Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic  
428 mutations hidden in cancer genomes. *Current opinion in genetics & development* **24**,  
429 52-60, doi:10.1016/j.gde.2013.11.014 (2014).
- 430 11 Cancer Genome Atlas Research, N. Comprehensive genomic characterization of  
431 squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).
- 432 12 Jiang, F., Yin, Z., Caraway, N. P., Li, R. & Katz, R. L. Genomic profiles in stage I primary  
433 non small cell lung cancer using comparative genomic hybridization analysis of cDNA  
434 microarrays. *Neoplasia* **6**, 623-635, doi:10.1593/neo.04142 (2004).
- 435 13 Chujo, M. *et al.* Comparative genomic hybridization analysis detected frequent  
436 overrepresentation of chromosome 3q in squamous cell carcinoma of the lung. *Lung*  
437 *Cancer* **38**, 23-29 (2002).
- 438 14 Tonon, G. *et al.* High-resolution genomic profiles of human lung cancer. *Proceedings*  
439 *of the National Academy of Sciences of the United States of America* **102**, 9625-9630,  
440 doi:10.1073/pnas.0504126102 (2005).
- 441 15 Petersen, I. *et al.* Patterns of chromosomal imbalances in adenocarcinoma and  
442 squamous cell carcinoma of the lung. *Cancer Res* **57**, 2331-2335 (1997).
- 443 16 Balsara, B. R. & Testa, J. R. Chromosomal imbalances in human lung cancer.  
444 *Oncogene* **21**, 6877-6883, doi:10.1038/sj.onc.1205836 (2002).
- 445 17 Massion, P. P. *et al.* Genomic copy number analysis of non-small cell lung cancer  
446 using array comparative genomic hybridization: implications of the  
447 phosphatidylinositol 3-kinase pathway. *Cancer Res* **62**, 3636-3640 (2002).
- 448 18 Ried, T. *et al.* Mapping of multiple DNA gains and losses in primary small cell lung  
449 carcinomas by comparative genomic hybridization. *Cancer Res* **54**, 1801-1806 (1994).

- 450 19 Rodrigues, M. F., Esteves, C. M., Xavier, F. C. & Nunes, F. D. Methylation status of  
451 homeobox genes in common human cancers. *Genomics* **108**, 185-193,  
452 doi:10.1016/j.ygeno.2016.11.001 (2016).
- 453 20 Matsubara, D. *et al.* Inactivating mutations and hypermethylation of the NKX2-  
454 1/TTF-1 gene in non-terminal respiratory unit-type lung adenocarcinomas. *Cancer Sci*  
455 **108**, 1888-1896, doi:10.1111/cas.13313 (2017).
- 456 21 Winslow, M. M. *et al.* Suppression of lung adenocarcinoma progression by Nkx2-1.  
457 *Nature* **473**, 101-104, doi:10.1038/nature09881 (2011).
- 458 22 Tata, P. R. *et al.* Developmental History Provides a Roadmap for the Emergence of  
459 Tumor Plasticity. *Dev Cell* **44**, 679-693 e675, doi:10.1016/j.devcel.2018.02.024  
460 (2018).
- 461 23 van Boerdonk, R. A. *et al.* DNA copy number aberrations in endobronchial lesions: a  
462 validated predictor for cancer. *Thorax* **69**, 451-457, doi:10.1136/thoraxjnl-2013-  
463 203821 (2014).
- 464 24 Lee, K., Kim, J. H. & Kwon, H. The Actin-Related Protein BAF53 Is Essential for  
465 Chromosomal Subdomain Integrity. *Mol Cells* **38**, 789-795,  
466 doi:10.14348/molcells.2015.0109 (2015).
- 467 25 Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of  
468 chromosomal instability inferred from gene expression profiles predicts clinical  
469 outcome in multiple human cancers. *Nat Genet* **38**, 1043-1048, doi:10.1038/ng1861  
470 (2006).
- 471 26 Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE:  
472 generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*  
473 **10**, 161, doi:10.1186/1471-2105-10-161 (2009).
- 474 27 Endesfelder, D. *et al.* Chromosomal instability selects gene copy-number variants  
475 encoding core regulators of proliferation in ER+ breast cancer. *Cancer Res* **74**, 4853-  
476 4863, doi:10.1158/0008-5472.CAN-13-2664 (2014).
- 477 28 Blackburn, A. *et al.* Effects of copy number variable regions on local gene expression  
478 in white blood cells of Mexican Americans. *Eur J Hum Genet* **23**, 1229-1235,  
479 doi:10.1038/ejhg.2014.280 (2015).
- 480 29 Mileyko, Y., Joh, R. I. & Weitz, J. S. Small-scale copy number variation and large-scale  
481 changes in gene expression. *Proceedings of the National Academy of Sciences of the*  
482 *United States of America* **105**, 16659-16664, doi:10.1073/pnas.0806239105 (2008).
- 483 30 McGranahan, N., Burrell, R. A., Endesfelder, D., Novelli, M. R. & Swanton, C. Cancer  
484 chromosomal instability: therapeutic and diagnostic challenges. *EMBO Rep* **13**, 528-  
485 538, doi:10.1038/embor.2012.61 (2012).
- 486 31 Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl*  
487 *J Med* **376**, 2109-2121, doi:10.1056/NEJMoa1616288 (2017).

488  
489



490

491

492 **Figure Legends**

493

494 **Figure 1. Analysis of pre-invasive lung carcinoma-in-situ (CIS) lesions.**

495 (a) Detection of bronchial pre-invasive CIS lesions by autofluorescence bronchoscopy. (b)  
496 Histological outcomes of bronchial pre-invasive lesions. (c) Overview of the study protocol.  
497 Patients with identified CIS lesions underwent repeat bronchoscopy and rebiopsy every 4  
498 months. Definitive cancer treatment was only performed if pathological evidence of  
499 progression to invasive cancer was detected. The 'index biopsy' profiled in this study refers  
500 to the biopsy immediately preceding progression to invasive cancer or regression to low-  
501 grade dysplasia or normal epithelium. (d) Venn diagram of different -omics analyses  
502 performed on laser capture microdissection (LCM)-captured CIS lesions. Due to the small  
503 size of bronchial biopsies, not all analyses were performed on all samples

504

505 **Figure 2. Genomic aberrations in pre-invasive lung carcinoma-in-situ (CIS) lesions.**

506 Circos diagram comparing CIS genomic profiles with TCGA LUSC data. The outer histogram  
507 (A), shows mutation frequencies of all genes in TCGA data. The inner histogram (D) shows  
508 mutation frequencies in our CIS data. Profiles appear similar and no statistically significant  
509 differences were identified between the two datasets. Genes previously identified as  
510 potential drivers of lung cancer are labelled. Between the two histograms, average copy  
511 number changes are shown for TCGA data (B) and CIS data (C). Copy number gains are  
512 shown in red, losses in blue. Although differences between whole-genome and whole-  
513 exome sequencing techniques makes these datasets difficult to compare, we observe many  
514 similar features between the two; for example, gains in 3q and 5p, which are well recognised  
515 features of squamous cell lung cancer. In the centre of the circos plot, 39 rings represent the  
516 copy number profiles of our 39 samples, illustrating the individual contribution of each  
517 sample to the average values presented (E).

518

519 **Figure 3. Altered methylation and gene expression in lung carcinoma-in-situ (CIS)**  
520 **lesions.**

521 (a) Hierarchical clustering of 1335 significantly differentially expressed genes in progressive  
522 (n=17) and regressive (n=16) CIS lesions, based on a discovery set. Biological and clinical  
523 factors including age at diagnosis, gender, smoking history (pack years) and COPD status  
524 had no effect on CIS lesion gene expression profile (high expression = purple, low  
525 expression = orange). (b) Hierarchical clustering of the top 1000 significantly differentially  
526 methylated positions (DMPs) between progressive (n=36) and regressive (n=18) CIS lesions  
527 and controls (n=33). Biological and clinical factors including age at diagnosis, gender and  
528 smoking history (pack years) status had no effect on the methylation profile (hypomethylated  
529 DMPs = blue, hypermethylated DMPs = orange). (c) Principle component analysis of all  
530 profiled genes in progressive (n=27) and regressive (n=24) CIS lesions showing a clear  
531 distinction between progressive and regressive groups (p=0.0017). (d) Principle component  
532 analysis of all methylation data in progressive (n=36), regressive (n=18) and control (n=33)  
533 CIS lesions showing a clear distinction between progressive and regressive groups  
534 (p=6.8x10<sup>-25</sup>). P values were calculated using multivariate ANOVA.

535

536 **Figure 4. Carcinoma-in-situ (CIS) gene expression and methylation profiles are**  
537 **predictive of progression to cancer.**

538 (a) Probability plot based on a 291-gene signature for correct class prediction (discovery set  
539 - red circles indicate progressive lesions, green circles indicate regressive lesions). (b)  
540 Challenging the 291-gene signature on a CIS validation set. Area under the curve (AUC) is 1  
541 using Receiver Operating Characteristic (ROC) analysis. (c) Application of the 291-gene  
542 signature to TCGA LUSC data. Our signature classified TCGA LUSC vs TCGA controls  
543 samples with AUC of 0.81 (green circles indicate TCGA controls, orange circles indicate  
544 TCGA LUSC). (d) Distribution of methylation beta values across the genome in TCGA  
545 controls, CIS regressive and progressive and TCGA LUSC samples. Most probes are

546 regulated at 0 or 1 in normal tissue but this regulation is reduced in both regressive and  
547 progressive CIS and TCGA LUSC samples. (e) Methylation Heterogeneity Index, defined as  
548 counts of methylation probes with  $0.26 < \beta < 0.88$ , for each sample. MHI is higher in  
549 regressive and progressive CIS and TCGA LUSC compared with TCGA controls and this  
550 can be used as an accurate predictor with AUC=0.96 for TCGA LUSC vs TCGA controls and  
551 AUC=0.74 for progressive vs regressive CIS. (f) Histogram of AUC values calculated by  
552 performing the same analysis used in (e) 10,000 times, with each run limited to a different  
553 random sample of 2,000 probes (AUC mean for TCGA LUSC vs TCGA controls is 0.95  
554 (95% CI 0.92–0.98)). This demonstrates that a random sample of methylation probes can be  
555 an accurate predictor using this method.

556

557 **Figure 5. Chromosomal instability is associated with progression to cancer.**

558 (a) Mean expression of CIN-associated genes in CIS samples. Progressive (n=27) and  
559 regressive (n=24) CIS samples are well differentiated with AUC=0.96. Green circles indicate  
560 regressive CIS lesions; red circles indicate progressive CIS. (b) Plot of NEK2 expression  
561 across CIS samples demonstrates increasing expression with progression to cancer.  
562 Expression of this gene alone classifies progressive vs regressive CIS with AUC=0.93. (c)  
563 Pathway analysis of gene expression data between progressive (n=17) and regressive  
564 (n=16) CIS shows a strong chromosomal instability (CIN) signal, based on a discovery set.  
565 This signal remains strong when cell cycle genes are removed from the CIN70 signature. (d)  
566 Pathway analysis of methylation data demonstrating several cancer-related pathways up-  
567 regulated in progressive CIS compared with regressive CIS. Quoted significance values in  
568 (c) and (d) are calculated using 2-sided t-tests adjusted for multiple testing using a False  
569 Discovery Rate method, as implemented in the *GAGE* Bioconductor package.

570

571 **Tables**

572

	Whole genome sequencing set (N=39)		Methylation discovery set (N=60)			Methylation validation set (N=27)			Gene expression discovery set (N=33)		Gene expression validation set (N=18)	
	Progression	Regression	Progression	Regression	Controls	Progression	Regression	Controls	Progression	Regression	Progression	Regression
<b>Clinical Characteristics</b>												
Patients	21	8	13	7	16	9	7	8	16	14	9	8
Lesions Profiled	29	10	26	11	23	10	7	10	17	16	10	8
<b>Gender</b>												
Male	18	8	11	7	15	7	7	7	14	10	7	4
Female	3	0	2	0	1	2	0	1	2	4	2	4
<b>Age at bronchoscopy (years)</b>												
Mean	71.1	63.1	69.81	63.27	65.96	70.2	69.86	64.3	69.29	66.56	69.4	68.125
Median	72	65.5	70	67	68	73	68	63	70	67.5	71.5	68
Range	58-81	52-71	52-79	53-79	44-77	58-78	64-76	56-77	55-80	53-81	56-82	57-84
<b>Smoking History (pack years)</b>												
Mean	54.4	54.9	58.08	31	41.95	57.3	62.14	37.71	57.07	47	49.125	59.2
Median	50	50	59.5	29	40	60	50	36	50	47.5	47.5	58
Range	30-100	9-141	32-141	5-88	20-65	40-75	30-141	20-60	22-141	5-141	30-75	30-96
<b>COPD status</b>												
Yes	12	3	9	3	14	5	1	7	4	8	3	7
No	9	5	4	4	1	4	6	1	12	6	1	0
<b>Previous History of Lung Cancer</b>												
Yes	12	2	6	2	9	7	4	3	5	4	3	4
No	9	6	7	5	7	2	3	5	11	10	6	4

573

574 **Table 1. Demographic and clinical characteristics.**

575 Table showing demographic and clinical characteristics of patients in the whole-genome  
 576 sequencing, methylation discovery and validation, and gene expression discovery and  
 577 validation datasets.

578

579

580 **Methods**

581

582 **Ethical approval**

583 All tissue and bronchial brushing samples were obtained under written informed  
584 patient consent and were fully anonymised. Study approval was provided by the UCL/UCLH  
585 Local Ethics Committee (REC references 06/Q0505/12 and 01/0148). All relevant ethical  
586 regulations were followed.

587

588 **Code availability**

589

590 All code used in our analysis will be made available at [http://github.com/ucl-](http://github.com/ucl-respiratory/preinvasive)  
591 [respiratory/preinvasive](http://github.com/ucl-respiratory/preinvasive) on publication. All software dependencies, full version information,  
592 and parameters used in our analysis can be found here.

593 Unless otherwise specified, all analyses were performed in an R statistical  
594 environment (v3.5.0; [www.r-project.org/](http://www.r-project.org/)) using Bioconductor<sup>1</sup> version 3.7.

595

596 **Biological samples**

597 All patients with pre-invasive lung cancer lesions were recruited through University  
598 College London Hospitals (UCLH) Early Lung Cancer Surveillance Programme (ELCSP).  
599 Full details of the surveillance protocol including eligibility criteria for patient inclusion have  
600 been previously described.<sup>2</sup> Briefly, the programme has recruited 140 patients to date with  
601 pre-invasive lung cancer lesions of varying histological grades. Patients undergo  
602 autofluorescence bronchoscopy (AFB) and CT/PET scans every four to six months during  
603 which multiple biopsy specimens are collected. This longitudinal sequential AFB procedure  
604 provides biopsies of the same lesion sampled repeatedly over time, allowing us to monitor  
605 whether the individual lesions have progressed, regressed or remained static<sup>2</sup>.

606 For a given CIS lesion under surveillance, when a biopsy from the same site showed  
607 evidence of progression to invasive cancer or regression to normal epithelium or low-grade

608 dysplasia, we define the preceding CIS biopsy as the 'index' lesion. An index lesion was  
609 defined as progressive if the subsequent biopsy at the same site showed invasive cancer, or  
610 as regressive if the subsequent biopsy showed normal epithelium or low-grade disease  
611 (metaplasia, mild or moderate dysplasia). Lesions which do not satisfy one of these end-  
612 points were excluded from this study. Patients with multiple fresh-frozen (FF) and formalin-  
613 fixed, paraffin-embedded (FFPE) tissue biopsies were identified for DNA methylation and  
614 gene expression analysis, respectively. Laser-capture micro-dissection (LCM) was used to  
615 selectively isolate CIS cells for molecular analysis, reducing the extent of contamination by  
616 stromal cells.

617 The following protocol was used to determine which profiling methods were applied to  
618 a given CIS lesion during our initial data collection phase:

- 619 • If FFPE samples were available, gene expression profiling was performed. For the first  
620 33 samples (17 progressive and 16 regressive), gene expression profiles were  
621 generated using Illumina microarrays. Our predictive models are trained on this  
622 discovery set. Subsequently, a further set of 10 progressive and 8 regressive  
623 samples from 18 patients were profiled using a different microarray platform  
624 (Affymetrix) to validate our findings on an independent platform.
- 625 • If FF samples were available, DNA from these samples was first used for methylation  
626 profiling. Samples with sufficient DNA after DNA profiling were additionally subjected  
627 to whole-genome sequencing. After acquisition of sufficient samples for our  
628 methylation dataset (54 samples; 36 progressive, 18 regressive), only 29 samples  
629 had sufficient DNA for WGS, therefore we prioritised WGS over methylation for the  
630 subsequent 10 samples.

631

### 632 **Tissue processing and laser-capture micro-dissection**

633 FF or FFPE tissue sections (7-10 $\mu$ M thickness) were mounted on a MembraneSlide  
634 1.0 PEN. Prior to cryosectioning, the slides were heat-treated for 4 h at 180°C in a drying  
635 cabinet to inactivate nucleases. To overcome the membrane's hydrophobic nature and to

636 allow better section adherence, the slides were then UV-treated for 30 min at 254nm. Prior  
637 to laser-capture micro-dissection (LCM), the slides containing the FF tissue sections for DNA  
638 extraction were washed in serial ethanol dilutions (50, 75, 100%) to remove the freezing  
639 medium (OCT) and to avoid any interference with the laser's efficiency. For RNA extraction,  
640 FFPE sections were dewaxed using the Arcturus® Paradise® PLUS Reagent System  
641 (Applied Biosystems, Foster City, CA, USA). For each case, epithelial areas of pre-invasive  
642 disease were identified by haematoxylin and eosin staining of the corresponding cryosection  
643 (~7 µM thick). The presence of epithelial areas of interest was confirmed by histological  
644 assessment of each case by two histopathologists. LCM to isolate the tissue area/cells of  
645 interest was performed with the PALM Microbeam™ system (Carl Zeiss MicroImaging,  
646 Munich, Germany) on unstained sections. The micro-dissected material was catapulted into  
647 a 500µl AdhesiveCap that allows capture of the isolated tissue without applying any liquid  
648 into the cap prior to LCM, thus minimizing the risk of nuclease activity. The captured cells  
649 were stored at -80°C until DNA extraction or processed immediately for RNA.

650

651

## 652 **DNA extraction**

653 DNA from the micro-dissected tissue and bronchial brushing samples was extracted  
654 using QIAGEN's QIAmp DNA Mini and Micro kits, respectively (Crawley, UK). Soluble carrier  
655 RNA was used to increase tissue DNA yield. Concentration was measured using the Qubit®  
656 dsDNA High-Sensitivity assay and Qubit® 2.0 Fluorometer (Life Technologies, Paisley, UK).  
657 Nucleic acid quality and purity was estimated based on the  $A_{260/280}$  absorbance ratio  
658 readings using the NanoDrop-8000 UV-spectrophotometer (Thermo Scientific, Hertfordshire,  
659 UK). Only samples with an  $A_{260/280}$  ratio of 1.7-1.9 were included in the study.

660

## 661 **RNA extraction**

662 RNA was extracted using the High Pure FFPE RNA Kit (Roche Applied Science, West  
663 Sussex, UK) according to manufacturer's protocol. Quantification was carried out using the

664 Quant-iT RNA assay kit and the Qubit<sup>®</sup> 2.0 fluorometer (Life Technologies, Paisley, UK).  
665 RNA integrity was analyzed using a BioAnalyzer 2100 (Agilent, Stockport, UK).

666

#### 667 **Bisulfite conversion**

668 For each sample undergoing methylation profiling, 200 ng of DNA were bisulfite  
669 converted using the EZ DNA methylation kit (Zymo Research Corp., Orange, CA, USA)  
670 according to the manufacturer's modified protocol for Illumina's Infinium 450K assay. This  
671 protocol incorporates a cyclic denaturation step to improve the conversion efficiency<sup>3</sup>. The  
672 10 µl final conversion reaction was concentrated down to 4 µl with a vacufuge plus vacuum  
673 concentrator (Eppendorf AG, Hamburg, Germany) and sent to UCL's Genomics Core  
674 Facility for hybridization on the 450K BeadArray according to Illumina's Infinium HD protocol  
675 (Illumina Inc., San Diego, CA, USA) as previously described.<sup>4</sup>

676

#### 677 **Infinium HumanMethylation450K raw data extraction and pre-processing**

678 Illumina's iScan fluorescent system was used to scan and image the arrays. DNA  
679 methylation data were extracted as raw intensity signals without any prior background  
680 subtraction or data normalization and were stored as IDAT files.

681 CpG-specific methylation levels ( $\beta$ -values; continuous value ranging from 0 to 1) for  
682 each sample were calculated as the ratio of the fluorescent signal intensity of the methylated  
683 (M) and unmethylated (U) alleles according to the following formula:

684

$$685 \beta = \frac{\text{intensity of methylated allele (M)}}{\text{intensity of [unmethylated (U) + methylated (M) allele] + 100}}$$

686

687 All subsequent raw  $\beta$ -value pre-processing, normalisation and down-stream analysis  
688 was performed using the Chip Analysis Methylation Pipeline (*ChAMP*) Bioconductor  
689 package with default settings.<sup>5</sup>



690 Analysis of differentially variable positions (DVP) was performed using iEVORA<sup>6</sup>. Beta  
691 values from ChAMP were used as input to iEVORA following normalization and batch  
692 correction.

693

#### 694 **Genome-wide gene expression array**

695 The extracted FFPE RNA used to generate the gene expression profiles on the  
696 discovery set was sent to UCL's Genomics Core Facility for hybridization on the Human  
697 Whole-Genome DASL (cDNA-mediated Annealing, Selection, extension and Ligation)  
698 beadarrays according to Illumina's protocol (Illumina Inc., San Diego, CA, USA).

699 The extracted FFPE RNA used to generate the gene expression profiles on the  
700 validation set was sent to UK Bioinformatics Limited for hybridization on the Clariom™ D  
701 Transcriptome Human Pico Assay 2.0 according to Affymetrix's protocol (Thermo Fisher  
702 Scientific Waltham, MA, USA).

703

#### 704 **Principal Component Analysis (PCA)**

705 In order to identify any potential factors of variability affecting sample/group  
706 segregation, we applied principal component analysis on all probes passing filters defined  
707 above (implemented in the *prcomp* method of the R *stats* package). Technical and biological  
708 variation was investigated for batch arrays, smoking (pack-years), age at initial diagnosis,  
709 gender and previous lung cancer history. The ability of these features to predict the first  
710 principal component was quantified using ANOVA analysis, implemented in the R *aov*  
711 method. p-values quoted are derived from this method.

712

#### 713 **Gene expression analysis**

714 Raw gene expression data were expressed as log<sub>2</sub> ratios of fluorescence intensities of  
715 the experimental samples. Quantile normalization was applied to Illumina data, using  
716 Illumina GenomeStudio Gene Expression Module v1.0 software. For Affymetrix data, RMA  
717 normalization was applied as defined in the *affy* Bioconductor package. For analyses

718 utilizing both data sets, only genes represented on both arrays were included and *ComBat*<sup>7</sup>  
719 was used to adjust for batch effects.

720 Differential expression analysis was performed using the *limma*<sup>8</sup> Bioconductor  
721 package. Raw p-values were adjusted by the Benjamini-Hochberg procedure to give a  
722 FDR.<sup>9</sup> A significance threshold of FDR < 0.01 was used to select differentially expressed  
723 genes. Cluster analysis and visualization was performed using the *pheatmap*<sup>10</sup> Bioconductor  
724 package.

725

### 726 **Real Time PCR Validation**

727 For microarray validation, total RNA from the 33 pre-invasive LUSC lesions  
728 undergoing Illumina gene expression profiling was reverse transcribed using qScript™  
729 cDNA Super-Mix (Quanta Biosciences, Lutterworth, UK) according to the manufacturer's  
730 protocol. Real-time quantitative PCR was carried out in eight genes using the SYBR-green  
731 master mix (Applied Biosystems, Bleiswijk, Netherlands) in an Eppendorf real-time PCR  
732 Machine (Eppendorf, Stevenage, UK). Findings were validated using quantitative PCR  
733 (qPCR) for four up-regulated (*GAGE5*, *GPNMB*, *MMP12* and *STC2*) and four down-  
734 regulated (*SPDEF*, *LMO7*, *OBSCN* and *MT1E*) genes. Gene-specific primers were designed  
735 inside or nearby the microarray sequence targeted, using Primer Express Software v2.0  
736 (Thermo Fisher Scientific). Relative gene expression was quantified using the threshold  
737 cycle (Ct) method and normalized to the amount of CTBL and CEP250, which met the  
738 criteria of less variation between samples and compatible expression level with the studied  
739 genes. Each sample was tested in triplicate and a sample without template was included in  
740 each run as a negative control. Correlations between microarrays and real time PCR data  
741 were measured using the Pearson coefficient. From microarray and real time PCR data, we  
742 calculated the progressive/regressive ratio for each gene expression. All eight genes tested  
743 were significant in our differential microarray analysis with FDR < 0.05. A high degree of  
744 correlation ( $r=0.982$ ) was observed between qPCR and array data.

745

746 **Predictive modelling**

747 For methylation, gene expression and copy number data we applied Prediction  
748 Analysis of Microarrays (PAM)<sup>11</sup> to predict whether a sample was progressive or regressive  
749 based on its molecular profile. The Bioconductor *pamr* package was used. In all presented  
750 analyses we select a threshold which minimizes the number of data inputs required whilst  
751 maintaining the minimum possible number of classification errors.

752 PAM calculates the probability of each sample being progressive. We describe this  
753 value as a 'Progression Score'. ROC analytics were performed on these progression scores  
754 to determine their value as a diagnostic test, using the *pROC*<sup>12</sup> and *PRROC*<sup>13</sup> Bioconductor  
755 packages.

756 For methylation and gene expression data a predictive model was trained on the  
757 training set and subsequently applied to an independent validation set. Regressive and  
758 control samples were grouped together for the methylation data analysis. ROC analytics  
759 were performed only on the validation set. Internal cross-validation was used for  
760 methylation-derived copy number data due to smaller sample size (control samples are used  
761 as a baseline to calculate copy number, therefore are excluded from predictive analysis).

762 When multiple lesions from one patient were included in an analysis, these were  
763 treated as independent events as they were always taken from different sites in the lung.  
764 The outcome of a lesion (whether it progressed or regressed) was determined on a per-  
765 lesion basis; the lesion was assigned to the progressive group only if cancer developed at  
766 the same site in the lung, and to the regressive group only if normal or low-grade dysplasia  
767 was obtained from the same site in the lung.

768 In some cases different technologies were used, for example our gene expression  
769 discovery set used Illumina microarrays whereas our validation set used Affymetrix. In such  
770 instances, both data sets were reduced to the subset of genes covered by probes in both  
771 platforms prior to creating a predictive model. The *ComBat* method from the *sva*  
772 Bioconductor package was used to correct for batch effects between the different platforms.  
773 In the case of RNAseq data, we used the *voom* transformation defined in the *limma*

774 Bioconductor package to derive data comparable to expression data prior to batch correction  
775 with *ComBat*.

776 A second predictive model based on methylation probe variation was also developed.  
777 For a given sample we defined Methylation Heterogeneity Index (MHI) by counting all  
778 probes with beta values between 0.26 and 0.88. These thresholds were optimized by  
779 calculating MHI for a range of different threshold values, and choosing those with the highest  
780 AUC for progressive vs regressive in our discovery cohort. We used ROC analytics to  
781 assess this model as a predictor of TCGA cancer vs control samples, and of progressive vs  
782 regressive samples in our validation cohort. We demonstrate in the main text that applying  
783 this method to a random sample of 2,000 probes performs similarly to using the entire array.  
784 We ran simulations using different sample sizes and found that performance with n=2000  
785 was similar to that of the entire array. To investigate potential confounding variables we use  
786 binomial logistic regression, implemented in the R *glm* method, to assess whether outcome  
787 (progression/regression) could be predicted by MHI, smoking status, COPD, previous history  
788 of lung cancer, age or gender. Control samples derived from brushings were excluded from  
789 these analyses.

790

### 791 **Copy number variation analysis**

792 For samples with whole-genome sequencing available we used ASCAT<sup>14</sup> to derive  
793 local copy number estimates as described below. To increase our sample size for  
794 comparative analyses, Copy number variation (CNV) data were obtained from non-  
795 normalised methylated and unmethylated signal intensities of probes in the 450K array as  
796 previously described<sup>15</sup> using the *ChAMP* Bioconductor package with default settings. Copy  
797 number (CN) profiles for progressive and regressive cases were obtained using the control  
798 cases for baseline normalisation. A previously defined threshold of  $\pm 0.3$  was used for the  
799 identification of single CNV. Probes associated with highly polymorphic regions (e.g. major  
800 histocompatibility complex) were removed from the analysis. The analysis generated group  
801 CN frequency plots and CN profiles for each sample. For samples with both methylation and

802 sequencing data available we observed good correlation between copy numbers derived  
803 from the two different methods (data not shown).

804 For comparison with previous results, the *ChAMP* pipeline was then modified to return  
805 CNV values per-probe. Probe locations were matched to cytogenetic bands using the  
806 Ensembl GRCh37 assembly, obtained from  
807 [http://grch37.rest.ensembl.org/info/assembly/homo\\_sapiens?content-](http://grch37.rest.ensembl.org/info/assembly/homo_sapiens?content-type=application/json&bands=1)  
808 [type=application/json&bands=1](http://grch37.rest.ensembl.org/info/assembly/homo_sapiens?content-type=application/json&bands=1), such that copy number variation could be assessed by  
809 cytogenetic band. The mean CNV value for each of 778 cytogenetic bands was calculated  
810 for each of our 54 samples. *Limma* analysis was used to identify bands that differed  
811 significantly between progressive and regressive samples with BH-adjusted p-value < 0.05.  
812 Predictive modelling was performed using *PAM* to find bands predictive of progression,  
813 using the same method as for gene expression data. Due to the low number of regressive  
814 samples, an internal cross-validation method was used rather than separate discovery and  
815 validation sets.

816 Following identification of predictive cytogenetic bands, *PAM* modelling was repeated  
817 with the dataset limited to only those bands identified by van Boerdonk et al: 3q26.2–29,  
818 3p26.3–p11.1 and 6p25.3–p24.3.<sup>16,17</sup> This model was also accurate.

819 Finally, we applied our model to the validation data set of 24 regressive and 12  
820 progressive samples used by van Boerdonk et al (GEO accession number GSE45287).  
821 These data were measured using a different microarray platform (arrayCGH). We assigned  
822 each probe to a cytogenetic band, and took the mean values to create a matrix of  
823 expression values by band. Our model was applied to the subset of chromosomal bands  
824 present in both data sets (760 of 778 bands). *ComBat* was used for batch correction  
825 between the two platforms. Our model correctly predicted 24/24 regressive samples and  
826 9/12 progressive samples, replicating the results of van Boerdonk et al.

827

## 828 **External validation using TCGA**

829 Lung cancer methylation datasets publically available through The Cancer Genome

830 Atlas (TCGA) were downloaded using *GenomicDataCommons* download tools<sup>18</sup>. We  
831 obtained the normalized  $\beta$ -values of 370 LUSC samples and 42 normal controls. *ComBat*  
832 was used to correct for batch effects between our data and TCGA data. These data were  
833 used as an external validation set to test our predictive models, and as input for our  
834 differential analysis of progression drivers from control through CIS to cancer.

835 Gene-expression microarray data sets comparable to our data were not publically  
836 available. RNAseq data was available from TCGA for 502 LUSC samples and 49 control  
837 samples. We applied a *voom* transformation<sup>19</sup> to these data, which uses normalized log-  
838 counts-per-million as an approximation for expression values, and hence allows comparison  
839 of RNAseq data with our gene expression pipeline. *ComBat* was used to correct for batch  
840 effects. The predictive model generated using *PAM* on our gene expression microarray data  
841 was applied to *voom*-transformed RNAseq data from TCGA and shown to be predictive (**Fig.**  
842 **4C**). We therefore demonstrate the applicability of our model to this fully independent data  
843 set. These data were again used as input to our differential analysis of progression drivers.

844

#### 845 **Pathway analysis**

846 For gene expression data, the *GAGE* Bioconductor package<sup>20</sup> was used with KEGG  
847 gene sets<sup>21-23</sup> to identify pathways associated with genes differentially expressed in our  
848 analysis of progression to cancer (BH-adjusted p-value <0.01). In addition to these pathways  
849 we use the CIN70 signature defined by Carter et al.<sup>24</sup> to assess for a chromosomal instability  
850 signal. We also use a subset of the CIN70 genes with cell-cycle associated genes<sup>25</sup>  
851 removed to ensure that our signal is genuinely CIN-related, rather than a measure of  
852 proliferation.

853 Methylation data was analysed in the same way, using beta values as input to *GAGE*.  
854 In cases where there are multiple methylation probes for a single gene we use the mean  
855 beta value over that gene as input to pathway analysis. We acknowledge that using mean  
856 signal may be insensitive to single-probe methylation changes, however given the scale of  
857 changes observed we believe it will identify areas of large methylation change.

858

### 859 **Genomic sequencing**

860 We created genome-wide shotgun libraries (insert size 331-367 bp) from native DNA  
861 using the Agilent Technologies Custom SureSelect Library Prep Kit library (cat no. 930075).  
862 150 bp paired-end sequence data were generated using the Illumina HiSeq X Ten system.  
863 Sequenced data were realigned to the human genome (NCBI build 37) using BWA-MEM.  
864 Unmapped reads and PCR duplicates were removed. A minimum sequencing depth of 40x  
865 was required.

866

### 867 **Somatic mutation calling and annotation**

868 Single base somatic substitutions were identified by our in-house algorithm Cancer  
869 Variants through Expectation Maximisation (CaVEMan:  
870 <https://github.com/cancerit/CaVEMan>)<sup>26</sup>. This algorithm compares the sequence data from  
871 each tumour sample to its matched normal and calculates a mutation probability at each  
872 locus. This calculation incorporates information from aberrant cell fraction and copy number  
873 estimates from the Allele-Specific Copy number Analysis of Tumours (ASCAT) algorithm  
874 (<https://www.crick.ac.uk/peter-van-loo/software/ASCAT>).<sup>14,27</sup> Additional post-processing as  
875 described previously<sup>28</sup> was implemented. Any putative driver mutations were visually  
876 inspected with Jbrowse.<sup>29</sup> For every substitution that passed all filters in at least one sample,  
877 we counted the number of wild-type and mutant reads at the same position in all other  
878 samples from the same patient to see if that mutation was also present in related samples  
879 but had not been called.

880

### 881 **Somatic small insertions and deletions**

882 These were identified using our in-house algorithm Pindel.<sup>30,31</sup> As with substitutions, all  
883 putative driver mutations were visualised with Jbrowse.

884

### 885 **Somatic structural variant detection”**

886 Abnormally paired read pairs were grouped using an in-house tool, “Brass”.<sup>32</sup> Read  
887 groups overlapping genomic repeats, reads from the matched normal, or from a panel of  
888 unmatched normals were ignored. Read pair clusters were then filtered by read remapping.  
889 Read pair clusters with >50% of the reads mapping to microbial sequences were removed.  
890 Finally, candidate SV breakpoints were matched to copy number breakpoints as defined by  
891 ASCAT within 10 kb. Candidate SVs that were not associated with copy number  
892 segmentation breakpoints and with a copy number change of at least 0.3 were removed. All  
893 putative driver rearrangements were visually inspected using IGV.<sup>33,34</sup>

894

### 895 **Somatic copy number events, ploidy, and stromal contamination**

896 Copy number changes were derived from whole-genome sequencing data using the  
897 ASCAT algorithm. This algorithm compares the relative representation of heterozygous  
898 SNPs and the total read depth at these positions to estimate the aberrant cell fraction and  
899 ploidy for each sample, and then to determine allele-specific copy number.

900

### 901 **Weighted Genome Integrity Index**

902 To estimate the overall chromosomal instability of a sample, we use the Weighted  
903 Genome Integrity Index (wGII) score<sup>35</sup>. This is calculated by measuring the percentage of  
904 the genome which is abnormal, corrected such that each chromosome is equally weighted.

905

### 906 **Mutation annotation**

907 Lung cancer driver genes were selected from the COSMIC Cancer Gene Census  
908 (CGC) v85 (cancer.sanger.ac.uk)<sup>36</sup>. CGC data was downloaded on 20<sup>th</sup> June 2018. Genes  
909 annotated in the CGC as potential drivers in lung cancer or NSCLC were included. Those  
910 specific to adenocarcinoma were excluded as our samples are precursors to squamous  
911 cancers. Genes identified in two large studies of squamous cell cancer, and some additional  
912 genes based on expert curation of the literature (*ARID1A*, *AKT2*, *FAT1*, *PTPRB*) were  
913 included if they were present in the CGC – even if they were not annotated explicitly as



914 implicated in lung cancer. Both Tier 1 and Tier 2 genes were included. A total of 96 genes  
915 were selected as putative lung squamous cell carcinoma drivers (**Supplementary Table 4**).

916 Mutations affecting these putative driver genes were annotated as driver mutations if  
917 they passed the following filters:

- 918 • The mutation type (e.g. missense, frameshift, amplification) must have been validated  
919 in the CGC for the affected gene.
- 920 • For genes annotated as tumour suppressors, mutations determined to have High or  
921 Moderate impact using Ensembl's Variant Effect Predictor<sup>37</sup> were classed as driver  
922 mutations.
- 923 • For genes annotated as oncogenes, we checked the specific mutation against  
924 COSMIC mutation data for lung carcinomas. If the specific mutation occurred 3 or  
925 more times in this dataset it was classed as a driver mutation.
- 926 • For genes annotated as fusion proteins, translocations with a translocation partner  
927 gene matching validated translocation partner genes in the CGC were classed as  
928 driver events.
- 929 • Copy number amplifications and deletions were all classed as driver events if  
930 amplifications/deletions in the affected gene have been previously validated in the  
931 CGC. We included homozygous deletions of tumour suppressor genes and  
932 amplifications to more than double the sample ploidy for oncogenes.

933

934 Driver mutation discovery was also attempted using *dndscv*<sup>38</sup>. This was  
935 underpowered, however, and only yielded *TP53* and *CDKN2A* as genes under positive  
936 selection. This package was also used to estimate the global dNdS for both progressive and  
937 regressive lesions.

938

### 939 **Subclonality analysis**

940 The number of subclones contributing to a sample and their relative contribution was

941 estimated by using a modified version of the *sciClone* Bioconductor package<sup>39</sup>. *sciClone*  
942 uses a Bayesian method to allocate mutations to clusters based on their variant allele  
943 frequency (VAF). By default, *sciClone* only considers regions that are copy number neutral  
944 and LOH-free. Given the significant aneuploidy in our data set we overcame this limitation by  
945 clustering on cancer cell fraction (CCF) rather than VAF. Briefly, cancer cell fraction  
946 represents the fraction of cancer cells in which a given mutation is present, therefore clonal  
947 mutations will have CCF=1. Following the method of McGranahan et al.<sup>40</sup>, we estimated the  
948 CCF for each mutation with a 95% confidence interval. Mutations for which 1 lay within this  
949 confidence interval were labelled as 'clonal', other mutations as 'subclonal'.

950 CCF values for each mutation were then used as input to *sciClone* in place of VAF  
951 values to quantify clusters present (divided by 2 such that clonal mutations have a value of  
952 0.5). As CCF corrects for local copy number, all regions were assumed to have copy number  
953 of 2, allowing *sciClone* to group mutations based only on their CCF estimates. A minimum  
954 tumour sequencing depth of 10 was required for each mutation.

955 Where more than one sample from a given patient was available, both one  
956 dimensional and multi-dimensional clustering were performed. Results from one dimensional  
957 clustering were used in the comparison of numbers of clones and proportion of clonal  
958 mutations between progressive and regressive lesions, in order to provide as fair a  
959 comparison as possible.

960

## 961 **Extraction of mutational signatures**

962 To obtain an approximate estimate of the contribution of different known mutational  
963 signatures to each sample, we used the *MutationalPatterns* Bioconductor package<sup>41</sup>. As a  
964 reference set of mutational signatures, we used a table with the relative frequency of each of  
965 the 96 trinucleotide substitutions across 30 known mutation signatures,<sup>42,43</sup> available through  
966 the COSMIC website (<http://cancer.sanger.ac.uk/cosmic/signatures>).

967 After a first run which indicated the most likely contribution of each signature, it  
968 seemed that the majority of substitutions were contributed by signatures 1, 2, 4, 5, and 13,

969 which have been described to be the strongest signatures in lung squamous cell cancer.<sup>44</sup>  
970 Some contribution was identified from signatures 16, 8, 18 and 3 in our initial analysis;  
971 however, in this context it is likely that these represent overfitting given that signature 16 is  
972 similar to signature 5, and signatures 8, 18 and 3 are similar to signature 4. We therefore ran  
973 the algorithm a second time, this time only using a 5x96 matrix of mutational signatures 1, 2,  
974 4, 5 and 13. All mutations were thus forced to belong to one of these five mutational  
975 signatures.

976 For a comparison of the clonal vs subclonal mutational processes in each sample,  
977 substitutions were annotated as clonal or subclonal based on CCF as described above.  
978 These were then run through the *MutationalPatterns* package.

979

#### 980 **Comparison of mutational burden and signatures with other cancer types**

981 Signatures of mutations in our CIS dataset were compared with mutational signatures  
982 found in lung squamous cell cancer. Raw whole-exome sequencing data for this cancer  
983 type was downloaded from TCGA, and run through our substitution-calling algorithm  
984 CAVEMaN as described above. We then looked at the total number of substitutions called,  
985 and estimated the contribution of each mutational signature using the methods described  
986 above. Only coding regions of the CIS whole-genome sequencing data were compared to  
987 these exomes.

988

#### 989 **Estimation of telomere lengths**

990 Telomere lengths were estimated using telomerecat<sup>45</sup>, and were compared in  
991 progressive and regressive groups. Telomerecat is a *de novo* method for the estimation of  
992 telomere length (TL) from whole-genome sequencing samples. The algorithm works by  
993 comparing the ratio of full telomere reads to reads on the boundary between telomere and  
994 subtelomere. This ratio is transformed to a measure of length by taking into account the  
995 fragment length distribution. Telomerecat also corrects for error in sequencing reads by  
996 modeling the observed distribution of phred scores associated with mismatches in the

997 telomere sequence. Samples were analysed in two groups corresponding to two separate  
998 sequencing batches, as per the telomerecat documentation.

999

#### 1000 **Data Availability Statement**

1001 Whole-genome sequencing data have been deposited at the European Genome  
1002 Phenome Archive (<https://www.ebi.ac.uk/ega/> at the EBI) with accession number  
1003 EGAD00001003883. All gene expression and methylation microarray data reported in this  
1004 study have been deposited in the National Center for Biotechnology Information Gene  
1005 Expression Omnibus (GEO, <http://www.ncbi.nlm.nih.gov/geo/>) public repository, and they  
1006 are accessible through GEO accession number GSE108124.

1007

1008

1009

1010

1011

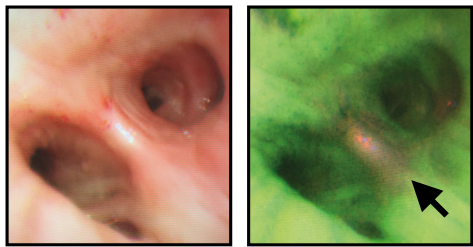
1012 **References (Methods-only)**

- 1013 1 Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor.  
1014 *Nat Methods* **12**, 115-121, doi:10.1038/nmeth.3252 (2015).
- 1015 2 Jeremy George, P. *et al.* Surveillance for the detection of early lung cancer in patients  
1016 with bronchial dysplasia. *Thorax* **62**, 43-50, doi:10.1136/thx.2005.052191 (2007).
- 1017 3 Bibikova, M. *et al.* High density DNA methylation array with single CpG site  
1018 resolution. *Genomics* **98**, 288-295, doi:10.1016/j.ygeno.2011.07.007 (2011).
- 1019 4 Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites  
1020 in the human genome. *Epigenetics* **6**, 692-702 (2011).
- 1021 5 Morris, T. J. *et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics*  
1022 **30**, 428-430, doi:10.1093/bioinformatics/btt684 (2014).
- 1023 6 Teschendorff, A. E. *et al.* DNA methylation outliers in normal breast tissue identify  
1024 field defects that are enriched in cancer. *Nat Commun* **7**, 10478,  
1025 doi:10.1038/ncomms10478 (2016).
- 1026 7 Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression  
1027 data using empirical Bayes methods. *Biostatistics* **8**, 118-127,  
1028 doi:10.1093/biostatistics/kxj037 (2007).
- 1029 8 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-  
1030 sequencing and microarray studies. *Nucleic acids research* **43**, e47,  
1031 doi:10.1093/nar/gkv007 (2015).
- 1032 9 Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false  
1033 discovery rate in behavior genetics research. *Behav Brain Res* **125**, 279-284 (2001).
- 1034 10 Kolde, R. Pheatmap: pretty heatmaps. *R package version* **61** (2012).
- 1035 11 Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types  
1036 by shrunken centroids of gene expression. *Proceedings of the National Academy of  
1037 Sciences of the United States of America* **99**, 6567-6572,  
1038 doi:10.1073/pnas.082099299 (2002).
- 1039 12 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare  
1040 ROC curves. *BMC Bioinformatics* **12**, 77, doi:10.1186/1471-2105-12-77 (2011).
- 1041 13 Keilwagen, J., Grosse, I. & Grau, J. Area under precision-recall curves for weighted  
1042 and unweighted data. *PLoS One* **9**, e92209, doi:10.1371/journal.pone.0092209  
1043 (2014).
- 1044 14 Raine, K. M. *et al.* ascatNgs: Identifying Somatic Acquired Copy-Number  
1045 Alterations from Whole-Genome Sequencing Data. *Current protocols in  
1046 bioinformatics* **56**, 15 19 11-15 19 17, doi:10.1002/cpbi.17 (2016).
- 1047 15 Feber, A. *et al.* Using high-density DNA methylation arrays to profile copy number  
1048 alterations. *Genome Biol* **15**, R30, doi:10.1186/gb-2014-15-2-r30 (2014).
- 1049 16 van Boerdonk, R. A. *et al.* DNA copy number alterations in endobronchial squamous  
1050 metaplastic lesions predict lung cancer. *American journal of respiratory and critical  
1051 care medicine* **184**, 948-956, doi:10.1164/rccm.201102-0218OC (2011).
- 1052 17 van Boerdonk, R. A. *et al.* DNA copy number aberrations in endobronchial lesions: a  
1053 validated predictor for cancer. *Thorax* **69**, 451-457, doi:10.1136/thoraxjnl-2013-  
1054 203821 (2014).
- 1055 18 Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med*  
1056 **375**, 1109-1112, doi:10.1056/NEJMp1607591 (2016).

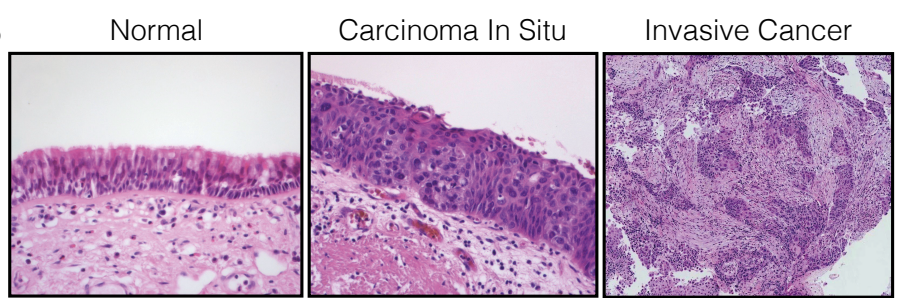
- 1057 19 Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear  
1058 model analysis tools for RNA-seq read counts. *Genome biology* **15**, R29,  
1059 doi:10.1186/gb-2014-15-2-r29 (2014).
- 1060 20 Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. & Woolf, P. J. GAGE:  
1061 generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics*  
1062 **10**, 161, doi:10.1186/1471-2105-10-161 (2009).
- 1063 21 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic*  
1064 *acids research* **28**, 27-30 (2000).
- 1065 22 Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a  
1066 reference resource for gene and protein annotation. *Nucleic acids research* **44**,  
1067 D457-462, doi:10.1093/nar/gkv1070 (2016).
- 1068 23 Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new  
1069 perspectives on genomes, pathways, diseases and drugs. *Nucleic acids research* **45**,  
1070 D353-D361, doi:10.1093/nar/gkw1092 (2017).
- 1071 24 Carter, S. L., Eklund, A. C., Kohane, I. S., Harris, L. N. & Szallasi, Z. A signature of  
1072 chromosomal instability inferred from gene expression profiles predicts clinical  
1073 outcome in multiple human cancers. *Nat Genet* **38**, 1043-1048, doi:10.1038/ng1861  
1074 (2006).
- 1075 25 Whitfield, M. L. *et al.* Identification of genes periodically expressed in the human cell  
1076 cycle and their expression in tumors. *Mol Biol Cell* **13**, 1977-2000,  
1077 doi:10.1091/mbc.02-02-0030. (2002).
- 1078 26 Jones, D. *et al.* cgpCaVEManWrapper: Simple Execution of CaVEMan in Order to  
1079 Detect Somatic Single Nucleotide Variants in NGS Data. *Current protocols in*  
1080 *bioinformatics* **56**, 15 10 11-15 10 18, doi:10.1002/cpbi.20 (2016).
- 1081 27 Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proc Natl Acad Sci U*  
1082 *S A* **107**, 16910-16915, doi:10.1073/pnas.1009843107 (2010).
- 1083 28 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007,  
1084 doi:10.1016/j.cell.2012.04.023 (2012).
- 1085 29 Skinner, M. E., Uzilov, A. V., Stein, L. D., Mungall, C. J. & Holmes, I. H. JBrowse: a  
1086 next-generation genome browser. *Genome Res* **19**, 1630-1638,  
1087 doi:10.1101/gr.094607.109 (2009).
- 1088 30 Raine, K. M. *et al.* cgpPindel: Identifying Somatically Acquired Insertion and Deletion  
1089 Events from Paired End Sequencing. *Curr Protoc Bioinformatics* **52**, 15 17 11-12,  
1090 doi:10.1002/0471250953.bi1507s52 (2015).
- 1091 31 Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth  
1092 approach to detect break points of large deletions and medium sized insertions from  
1093 paired-end short reads. *Bioinformatics* **25**, 2865-2871,  
1094 doi:10.1093/bioinformatics/btp394 (2009).
- 1095 32 Papaemmanuil, E. *et al.* RAG-mediated recombination is the predominant driver of  
1096 oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nature*  
1097 *genetics* **46**, 116-125, doi:10.1038/ng.2874 (2014).
- 1098 33 Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29**, 24-26,  
1099 doi:10.1038/nbt.1754 (2011).
- 1100 34 Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer  
1101 (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*  
1102 **14**, 178-192, doi:10.1093/bib/bbs017 (2013).

1103 35 Endesfelder, D. *et al.* Chromosomal instability selects gene copy-number variants  
1104 encoding core regulators of proliferation in ER+ breast cancer. *Cancer Res* **74**, 4853-  
1105 4863, doi:10.1158/0008-5472.CAN-13-2664 (2014).  
1106 36 Forbes, S. A. *et al.* COSMIC: somatic cancer genetics at high-resolution. *Nucleic acids*  
1107 *research* **45**, D777-D783, doi:10.1093/nar/gkw1121 (2017).  
1108 37 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* **17**, 122,  
1109 doi:10.1186/s13059-016-0974-4 (2016).  
1110 38 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues.  
1111 *Cell* **171**, 1029-1041 e1021, doi:10.1016/j.cell.2017.09.042 (2017).  
1112 39 Miller, C. A. *et al.* SciClone: inferring clonal architecture and tracking the spatial and  
1113 temporal patterns of tumor evolution. *PLoS Comput Biol* **10**, e1003665,  
1114 doi:10.1371/journal.pcbi.1003665 (2014).  
1115 40 McGranahan, N. *et al.* Clonal neoantigens elicit T cell immunoreactivity and  
1116 sensitivity to immune checkpoint blockade. *Science* **351**, 1463-1469,  
1117 doi:10.1126/science.aaf1490 (2016).  
1118 41 Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns:  
1119 comprehensive genome-wide analysis of mutational processes. *Genome Med* **10**, 33,  
1120 doi:10.1186/s13073-018-0539-0 (2018).  
1121 42 Alexandrov, L. B. *et al.* Clock-like mutational processes in human somatic cells. *Nat*  
1122 *Genet* **47**, 1402-1407, doi:10.1038/ng.3441 (2015).  
1123 43 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature*  
1124 **500**, 415-421, doi:10.1038/nature12477 (2013).  
1125 44 Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells.  
1126 *Science* **349**, 1483-1489, doi:10.1126/science.aab4082 (2015).  
1127 45 Farmery, J. H. R., Smith, M. L. & Lynch, A. G. Telomerecat: A Ploidy-Agnostic Method  
1128 For Estimating Telomere Length From Whole Genome Sequencing Data. *bioRxiv*,  
1129 doi:10.1101/139972 (2017).  
1130  
  
1131  
  
1132

A

White Light  
BronchoscopyAutofluorescence  
Bronchoscopy

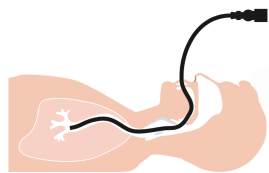
B



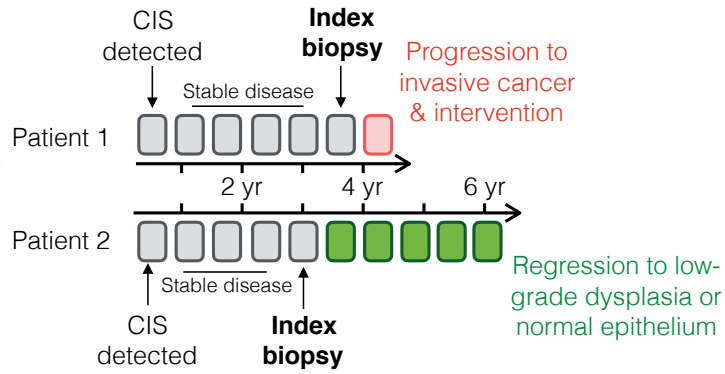
Regression (30%)

Progression (50%)

C

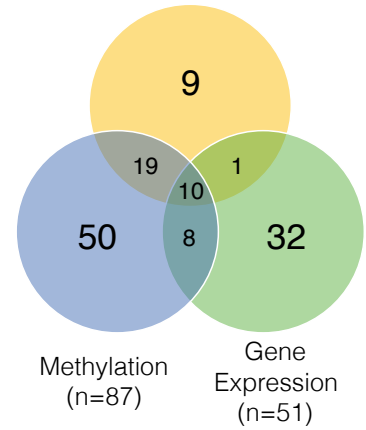


Surveillance Biopsies

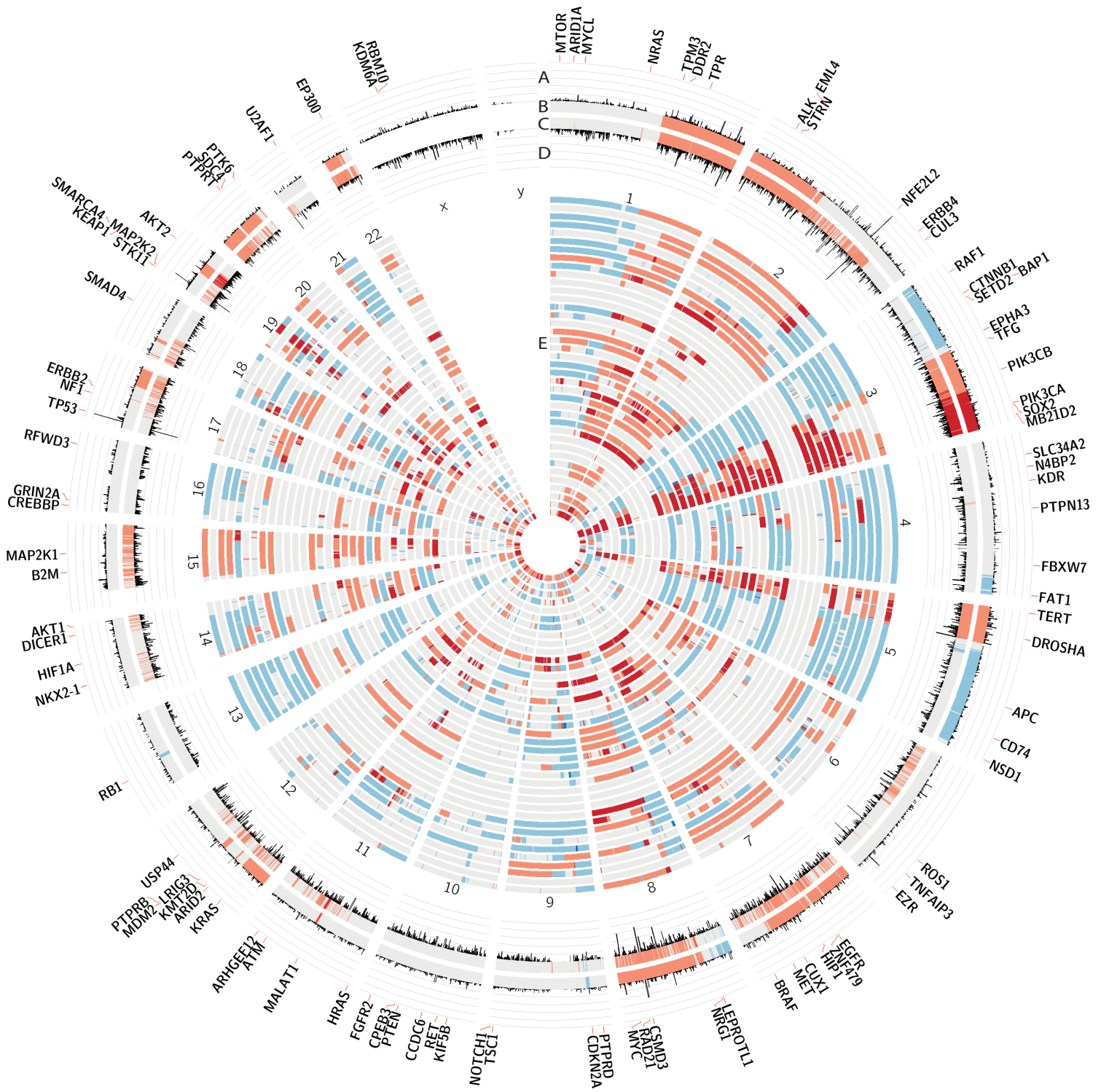


D

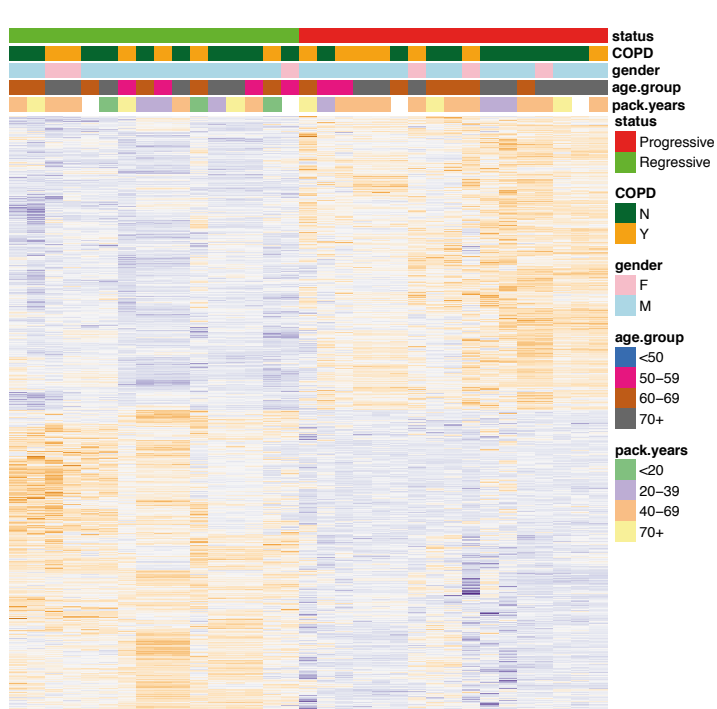
WGS (n=39)



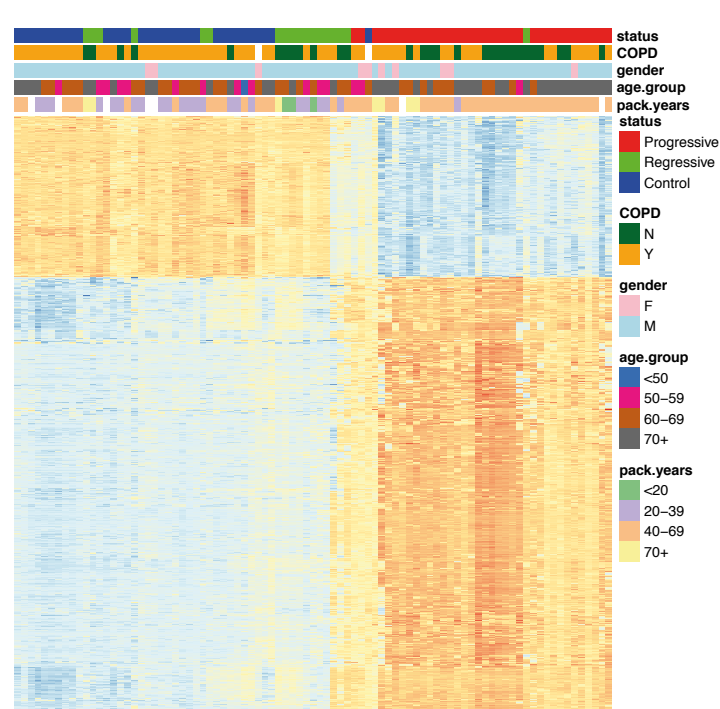




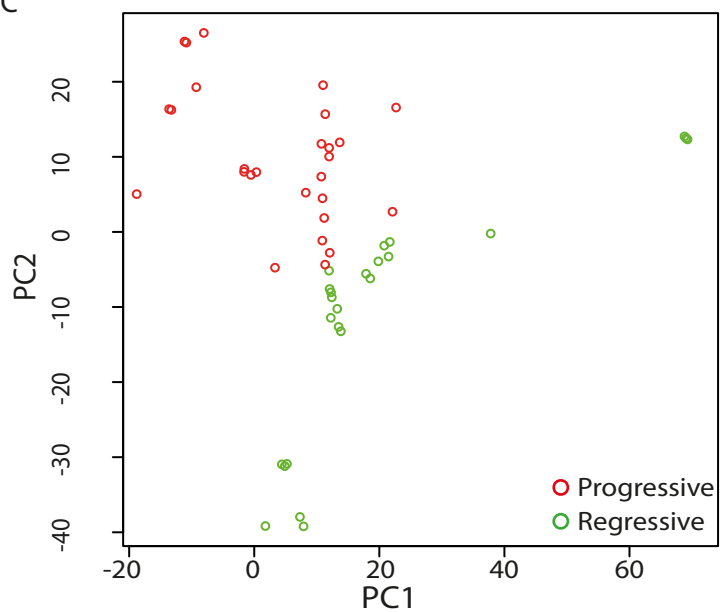
A



B



C



D

