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

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

[1] Lungs for Living Research Centre, UCL Respiratory, University College London, London, U.K.

[2] Research Department of Cancer Biology and Medical Genomics Laboratory, UCL Cancer Institute, University College London, London, U.K.

[3] The Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, U.K.

[4] Department of Medicine, Boston University School of Medicine, Boston, MA, U.S.A.

[5] CRUK Lung Cancer Centre of Excellence, UCL Cancer Institute, London, U.K.

[6] Translational Cancer Therapeutics Laboratory, The Francis Crick Institute, London, U.K.

[7] Department of Pathology, University College London Hospitals NHS Trust, London, U.K.

[8] Department of Thoracic Medicine, University College London Hospital, London, U.K.

[9] Center for Inflammation and Tissue Repair, UCL Respiratory, University College London, London, U.K.

[10] Computational Biology and Statistics Laboratory, Cancer Research UK Cambridge Institute, Cambridge, U.K.

[11] School of Medicine/School of Mathematics and Statistics, University of St Andrews, St Andrews, U.K.

[12] Johnson and Johnson Innovation, Cambridge MA

[*] Drs. Teixeira, Pipinikas and Pennycuick contributed equally to this study.

[#] **Corresponding author:**

Professor Sam M. Janes

Address: Lungs for Living Research Centre, UCL Respiratory, University College London, 5 University Street, London, WC1E 6JF, U.K.

Phone: (+44) 020 3549 5979

E-mail: s.janes@ucl.ac.uk

Abstract

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

Introduction

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

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

We reasoned that information on the future clinical trajectory of a pre-invasive lung lesion might be encoded in the genetic and epigenetic profile present at diagnosis. We therefore undertook a prospective cohort study of patients with pre-invasive squamous airway lesions. Patients were managed conservatively, undergoing surveillance AFB with biopsy and CT scanning every 4 and 12 months, respectively, with definitive cancer treatment only performed at the earliest pathological evidence of progression to invasive tumours (**Fig. 1a, b**).[7] When a CIS lesion either progressed to invasive cancer or regressed to normal epithelium/low-grade disease, molecular profiling was performed on the preceding CIS biopsy from the same lesion

– the ‘index biopsy’ (**Fig. 1c**). Index biopsies all demonstrated histologically and morphologically indistinguishable CIS and were classified as either ‘progressive’ or ‘regressive’. All such index CIS biopsies were subjected to a predetermined combination of transcriptomic, epigenetic and finally genomic profiling depending on DNA/RNA availability (**Fig. 1d; Table 1; Extended Data Fig. 1; Supplementary Table 1**).

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

Results

Patient Characteristics

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

Clinical characteristics within each analysis group are shown in Table 1. In comparing progressive and regressive samples, we found that progressive samples were associated with a higher pack-year smoking history in the methylation discovery group only ($p < 0.01$) and with

increased age in the WGS group (p = 0.01). No clinical differences were consistently observed across the different analysis groups.

Characterization of CIS genomic profiles

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

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

Whilst most CIS samples have the genomic appearance of neoplasms, we observe six lesions which show markedly lower mutational load and fewer copy number alterations than the others (**Extended Data Fig. 3;** PD21884c, PD21885a, PD21885c, PD21904d, PD38317a, PD38319a). These samples have very few genomic changes, despite being CIS histologically. All of these six samples regressed to normal epithelium or low-grade dysplasia on subsequent biopsy. Four further samples met this end-point for regression, despite widespread mutational

and copy number changes. However, with longer follow up one of these cases developed CIS recurrence (**Extended Data Fig. 4a**; PD21893a), and two developed invasive cancer on further surveillance (**Extended Data Fig. 4b,c**; PD21884a, PD38326a). Only one sample, PD21908a, showed sustained clinical regression after 9 years of follow up despite widespread molecular changes.

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

We compared genomic features between the 29 progressive and 10 regressive lesions. The three samples which showed evidence of progression after meeting our end-point for regression were excluded from this analysis. Comparisons of mutation burden between progressive and regressive lesions were performed by mixed effects modelling, allowing us to account for samples that come from the same patient. Even after correcting for patient age, smoking history and sample purity, progressive lesions had more somatically acquired mutations than those from regressive lesions, across base substitutions ($p<0.001$)**,** indels ($p=0.018$), structural variants ($p<0.001$) and copy number changes ($p<0.001$) (**Extended Data Fig. 5a-d**). When the analysis was restricted only to substitutions that were fully clonal in each lesion, there were still substantially more substitutions in progressive than regressive lesions ($p<0.001$) (**Extended Data Fig. 5e**), suggesting that the increase in mutation burden is not due to recent subclonal diversification in progressive lesions. All the mutational processes (or signatures[9,10]) identified in the CIS lesions contribute to the excess of mutations in progressive compared to regressive samples; however, only tobacco-associated signature 4 showed proportionally more mutations ($p=0.017$) (**Extended Data Fig. 2f-j**). Progressive lesions contained more putative driver mutations than regressive lesions ($p=0.001$) (**Extended Data**

Fig. 5h; **Supplementary Table 2**). Importantly, no single cancer mutation perfectly discriminated between progressive and regressive lesions.

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

CIS transcriptomic and epigenetic profiles

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

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

Of the 1335 genes identified, *TPM3, PTPRB, SLC34A2, KEAP1, NKX2-1, SMAD4* and *SMARCA4* have previously been implicated as potential lung cancer drivers (**Supplementary**

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

Principal component analysis of all gene expression and methylation data showed a clear distinction between the progressive and regressive subgroups ($p=0.0017$ and $p=6.8\times10^{-25}$, respectively) **(Fig. 3c,d)**. In the methylation dataset, the regressive lesions closely clustered with the control normal epithelial cells. A history of chronic obstructive pulmonary disease (COPD) had an effect on case segregation ($p=1.2\times10^{-5}$) but all other clinical and technical variables analysed, including smoking status and history of lung cancer, had no effect (**Extended Data Fig. 8a-f**). This was also the case for PCA analysis of the gene expression data (**Extended Data Fig. 8g-k)**.

For methylation, one control and four regressive cases clustered with the progressive cases (**Fig. 3d**). Three of the four mis-classified regressive cases were subjected to whole-genome sequencing and were found to have more copy number alterations than other regressive samples (PD21884a, PD21893a, PD21908a). Two of these correspond to the samples discussed above, which showed signs of progression after meeting the clinical end point of regression (**Extended Data Fig. 4**). For the control bronchial epithelium sample that was classified with the progressive lesions, CIS was detected in a biopsy specimen 12 months

later from the same site. Thus, although we have formally treated these cases as mis-classifications, it is likely that the molecular data underpinning the apparent errors indicate a cellular phenotype that is not consistent with a straightforward regressive lesion.

Molecular signatures predict CIS outcome

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

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

We observed an increased number of methylation probes with intermediate methylation in TCGA LUSC cancer vs TCGA control samples (**Fig. 4d**), reflecting methylation heterogeneity in these samples. We therefore developed a methylation heterogeneity index (MHI), defined as the number of probes per sample with $t_{lo} < ß < t_{hi}$. Optimization based on our discovery set of 26 progressive and 11 regressive samples defined values of $t_{lo} = 0.26$ and $t_{hi} = 0.88$. Control samples were not used in this analysis. This model classified progressive vs regressive CIS samples in our validation set with AUC=0.74 and TCGA LUSC vs TCGA control samples with AUC=0.96 (**Fig. 4e; Extended Data Fig. 10j-n**). Multivariate logistic regression in our CIS cohort demonstrated that this index was a predictor of progression status (p=0.017); previous history of lung cancer was also significantly associated (p=0.02), whereas smoking status, COPD status, age and gender were not.

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

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

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

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

CIN is an early marker of progression to cancer

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

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

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

This CIN signal is consistent with the observed pattern of widespread copy number change (**Fig. 2**). Overall copy number variation for a sample, as measured by Weighted Genome Integrity Index (wGII)[27], correlates with mean CIN-associated gene expression of that sample (Pearson r^2=0.473) (**Extended Data Fig. 5j**). We also observe a correlation between local copy number of a gene and expression of that gene, consistent with previous results[28,29].

Discussion

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

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

We present here the first major whole genome sequencing data of pre-invasive lung lesions. We acknowledge that, despite using the world's largest cohort of such lesions, the study remains underpowered to detect less common genomic alterations. Expanding our knowledge in this area will require a major international collaboration. Likewise we acknowledge that whilst our predictive signatures demonstrate the power of molecular data in guiding management decisions, a prospective clinical trial using predictors derived from our

data will be required before clinical use. Again, international collaboration will be required to develop an appropriately powered trial.

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

Acknowledgements

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

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

Author Contributions

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., 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. conceived the study design. S.M.J., P.J.C., C.T., V.H.T., C.P.P. and A.P. designed the study protocols. V.H.T. performed gene expression, qPCR and LCM experiments, analysed and integrated clinicopathological data and gene expression data. C.P.P. performed methylation and LCM experiments, analysed and integrated clinicopathological data and methylation data.

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

Competing Interests Statement

The authors declare the following competing interests:

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

.

References

1 Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. Global cancer statistics, 2002. *CA Cancer J Clin* **55**, 74-108 (2005).

2 Torre, L. A., Siegel, R. L. & Jemal, A. Lung Cancer Statistics. *Advances in experimental medicine and biology* **893**, 1-19, doi:10.1007/978-3-319-24223-1_1 (2016).

3 Nicholson, A. G. *et al.* Reproducibility of the WHO/IASLC grading system for pre-invasive squamous lesions of the bronchus: a study of inter-observer and intra-observer variation. *Histopathology* **38**, 202-208 (2001).

4 van der Heijden, E. H., Hoefsloot, W., van Hees, H. W. & Schuurbiers, O. C. High definition bronchoscopy: a randomized exploratory study of diagnostic value compared to standard white light bronchoscopy and autofluorescence bronchoscopy. *Respir Res* **16**, 33, doi:10.1186/s12931-015-0193-7 (2015).

5 Thakrar, R. M., Pennycuick, A., Borg, E. & Janes, S. M. Preinvasive disease of the airway. *Cancer Treat Rev* **58**, 77-90, doi:10.1016/j.ctrv.2017.05.009 (2017).

6 Pipinikas, C. P. *et al.* Cell migration leads to spatially distinct but clonally related airway cancer precursors. *Thorax* **69**, 548-557, doi:10.1136/thoraxjnl-2013-204198 (2014).

7 Jeremy George, P. *et al.* Surveillance for the detection of early lung cancer in patients with bronchial dysplasia. *Thorax* **62**, 43-50, doi:10.1136/thx.2005.052191 (2007).

8 Futreal, P. A. *et al.* A census of human cancer genes. *Nat Rev Cancer* **4**, 177-183, doi:10.1038/nrc1299 (2004).

9 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

10 Alexandrov, L. B. & Stratton, M. R. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. *Current opinion in genetics & development* **24**, 52-60, doi:10.1016/j.gde.2013.11.014 (2014).

11 Cancer Genome Atlas Research, N. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).

12 Jiang, F., Yin, Z., Caraway, N. P., Li, R. & Katz, R. L. Genomic profiles in stage I primary non small cell lung cancer using comparative genomic hybridization analysis of cDNA microarrays. *Neoplasia* **6**, 623-635, doi:10.1593/neo.04142 (2004).

13 Chujo, M. *et al.* Comparative genomic hybridization analysis detected frequent overrepresentation of chromosome 3q in squamous cell carcinoma of the lung. *Lung Cancer* **38**, 23-29 (2002).

14 Tonon, G. *et al.* High-resolution genomic profiles of human lung cancer. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9625-9630, doi:10.1073/pnas.0504126102 (2005).

15 Petersen, I. *et al.* Patterns of chromosomal imbalances in adenocarcinoma and squamous cell carcinoma of the lung. *Cancer Res* **57**, 2331-2335 (1997).

16 Balsara, B. R. & Testa, J. R. Chromosomal imbalances in human lung cancer. *Oncogene* **21**, 6877-6883, doi:10.1038/sj.onc.1205836 (2002).

17 Massion, P. P. *et al.* Genomic copy number analysis of non-small cell lung cancer using array comparative genomic hybridization: implications of the phosphatidylinositol 3-kinase pathway. *Cancer Res* **62**, 3636-3640 (2002).

18 Ried, T. *et al.* Mapping of multiple DNA gains and losses in primary small cell lung carcinomas by comparative genomic hybridization. *Cancer Res* **54**, 1801-1806 (1994).

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

Figure Legends

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

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

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

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

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

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

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

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

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

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

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

Tables

	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)	
Clinical Characteristics	**Progression**	**Regression**	**Progression**	**Regression**	**Controls**	**Progression**	**Regression**	**Controls**	**Progression**	**Regression**	**Progression**	**Regression**
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
Gender												
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
Age at bronchoscopy (years)												
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
Smoking History (pack years)												
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
COPD status												
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
Previous History of Lung Cancer												
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

Table 1. Demographic and clinical characteristics.

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

Methods

Ethical approval

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

Code availability

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

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

Biological samples

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

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

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

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

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

Tissue processing and laser-capture micro-dissection

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

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

DNA extraction

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

RNA extraction

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

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

Bisulfite conversion

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

Infinium HumanMethylation450K raw data extraction and pre-processing

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

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

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

All subsequent raw *β*-value pre-processing, normalisation and down-stream analysis was performed using the Chip Analysis Methylation Pipeline (*ChAMP*) Bioconductor package with default settings.[5]

Analysis of differentially variable positions (DVP) was performed using iEVORA[6]. Beta values from ChAMP were used as input to iEVORA following normalization and batch correction.

Genome-wide gene expression array

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

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

Principal Component Analysis (PCA)

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

Gene expression analysis

Raw gene expression data were expressed as $\log_2$ ratios of fluorescence intensities of the experimental samples. Quantile normalization was applied to Illumina data, using Illumina GenomeStudio Gene Expression Module v1.0 software. For Affymetrix data, RMA normalization was applied as defined in the *affy* Bioconductor package. For analyses utilizing

both data sets, only genes represented on both arrays were included and *ComBat*[7] was used to adjust for batch effects.

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

Real Time PCR Validation

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

Predictive modelling

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

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

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

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

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

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

Copy number variation analysis

For samples with whole-genome sequencing available we used ASCAT[14] to derive local copy number estimates as described below. To increase our sample size for comparative analyses, Copy number variation (CNV) data were obtained from non-normalised methylated and unmethylated signal intensities of probes in the 450K array as previously described[15] using the *ChAMP* Bioconductor package with default settings. Copy number (CN) profiles for progressive and regressive cases were obtained using the control cases for baseline normalisation. A previously defined threshold of ±0.3 was used for the identification of single CNV. Probes associated with highly polymorphic regions (e.g. major histocompatibility complex) were removed from the analysis. The analysis generated group CN frequency plots and CN profiles for each sample. For samples with both methylation and sequencing data available we observed good correlation between copy numbers derived from the two different methods (data not shown).

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

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

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

External validation using TCGA

Lung cancer methylation datasets publically available through The Cancer Genome Atlas (TCGA) were downloaded using *GenomicDataCommons* download tools[18]. We obtained the normalized *β*-values of 370 LUSC samples and 42 normal controls. *ComBat* was used to correct for batch effects between our data and TCGA data. These data were used as an

external validation set to test our predictive models, and as input for our differential analysis of progression drivers from control through CIS to cancer.

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

Pathway analysis

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

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

Genomic sequencing

We created genome-wide shotgun libraries (insert size 331-367 bp) from native DNA using the Agilent Technologies Custom SureSelect Library Prep Kit library (cat no. 930075).

150 bp paired-end sequence data were generated using the Illumina HiSeq X Ten system. Sequenced data were realigned to the human genome (NCBI build 37) using BWA-MEM. Unmapped reads and PCR duplicates were removed. A minimum sequencing depth of 40x was required.

Somatic mutation calling and annotation

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

Somatic small insertions and deletions

These were identified using our in-house algorithm Pindel.[30,31] As with substitutions, all putative driver mutations were visualised with Jbrowse.

Somatic structural variant detection"

Abnormally paired read pairs were grouped using an in-house tool, "Brass".[32] Read groups overlapping genomic repeats, reads from the matched normal, or from a panel of unmatched normals were ignored. Read pair clusters were then filtered by read remapping. Read pair clusters with >50% of the reads mapping to microbial sequences were removed.

Finally, candidate SV breakpoints were matched to copy number breakpoints as defined by ASCAT within 10 kb. Candidate SVs that were not associated with copy number segmentation breakpoints and with a copy number change of at least 0.3 were removed. All putative driver rearrangements were visually inspected using IGV.[33,34]

Somatic copy number events, ploidy, and stromal contamination

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

Weighted Genome Integrity Index

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

Mutation annotation

Lung cancer driver genes were selected from the COSMIC Cancer Gene Census (CGC) v85 (cancer.sanger.ac.uk)[36]. CGC data was downloaded on 20th June 2018. Genes annotated in the CGC as potential drivers in lung cancer or NSCLC were included. Those specific to adenocarcinoma were excluded as our samples are precursors to squamous cancers. Genes identified in two large studies of squamous cell cancer, and some additional genes based on expert curation of the literature (*ARID1A, AKT2, FAT1, PTPRB*) were included if they were present in the CGC – even if they were not annotated explicitly as implicated in lung cancer. Both Tier 1 and Tier 2 genes were included. A total of 96 genes were selected as putative lung squamous cell carcinoma drivers (**Supplementary Table 4**).

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

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

Driver mutation discovery was also attempted using *dndscv*[38]. This was underpowered, however, and only yielded *TP53* and *CDKN2A* as genes under positive selection. This package was also used to estimate the global dNdS for both progressive and regressive lesions.

Subclonality analysis

The number of subclones contributing to a sample and their relative contribution was estimated by using a modified version of the *sciClone* Bioconductor package[39]. *sciClone* uses a Bayesian method to allocate mutations to clusters based on their variant allele frequency (VAF). By default, *sciClone* only considers regions that are copy number neutral and LOH-free. Given the significant aneuploidy in our data set we overcame this limitation by clustering

on cancer cell fraction (CCF) rather than VAF. Briefly, cancer cell fraction represents the fraction of cancer cells in which a given mutation is present, therefore clonal mutations will have CCF=1. Following the method of McGranahan et al.[40], we estimated the CCF for each mutation with a 95% confidence interval. Mutations for which 1 lay within this confidence interval were labelled as 'clonal', other mutations as 'subclonal'.

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

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

Extraction of mutational signatures

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

After a first run which indicated the most likely contribution of each signature, it seemed that the majority of substitutions were contributed by signatures 1, 2, 4, 5, and 13, which have been described to be the strongest signatures in lung squamous cell cancer.[44] Some contribution was identified from signatures 16, 8, 18 and 3 in our initial analysis; however, in this context it is likely that these represent overfitting given that signature 16 is similar to signature 5, and signatures 8, 18 and 3 are similar to signature 4. We therefore ran the algorithm a second time, this time only using a 5x96 matrix of mutational signatures 1, 2, 4, 5

and 13. All mutations were thus forced to belong to one of these five mutational signatures.

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

Comparison of mutational burden and signatures with other cancer types

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

Estimation of telomere lengths

Telomere lengths were estimated using telomerecat[45], and were compared in progressive and regressive groups. Telomerecat is a *de novo* method for the estimation of telomere length (TL) from whole-genome sequencing samples. The algorithm works by comparing the ratio of full telomere reads to reads on the boundary between telomere and subtelomere. This ratio is transformed to a measure of length by taking into account the fragment length distribution. Telomerecat also corrects for error in sequencing reads by modeling the observed distribution of phred scores associated with mismatches in the telomere sequence. Samples were analysed in two groups corresponding to two separate sequencing batches, as per the telomerecat documentation.

Data Availability Statement

Whole-genome sequencing data have been deposited at the European Genome Phenome Archive (https://www.ebi.ac.uk/ega/ at the EBI) with accession number EGAD00001003883. All gene expression and methylation microarray data reported in this

study have been deposited in the National Center for Biotechnology Information Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) public repository, and they are accessible through GEO accession number GSE108124.

References (Methods-only)

1 Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat Methods* **12**, 115-121, doi:10.1038/nmeth.3252 (2015).

2 Jeremy George, P. *et al.* Surveillance for the detection of early lung cancer in patients with bronchial dysplasia. *Thorax* **62**, 43-50, doi:10.1136/thx.2005.052191 (2007).

3 Bibikova, M. *et al.* High density DNA methylation array with single CpG site resolution. *Genomics* **98**, 288-295, doi:10.1016/j.ygeno.2011.07.007 (2011).

4 Sandoval, J. *et al.* Validation of a DNA methylation microarray for 450,000 CpG sites in the human genome. *Epigenetics* **6**, 692-702 (2011).

5 Morris, T. J. *et al.* ChAMP: 450k Chip Analysis Methylation Pipeline. *Bioinformatics* **30**, 428-430, doi:10.1093/bioinformatics/btt684 (2014).

6 Teschendorff, A. E. *et al.* DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. *Nat Commun* **7**, 10478, doi:10.1038/ncomms10478 (2016).

7 Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118-127, doi:10.1093/biostatistics/kxj037 (2007).

8 Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research* **43**, e47, doi:10.1093/nar/gkv007 (2015).

9 Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behav Brain Res* **125**, 279-284 (2001).

10 Kolde, R. Pheatmap: pretty heatmaps. *R package version* **61** (2012).

11 Tibshirani, R., Hastie, T., Narasimhan, B. & Chu, G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6567-6572, doi:10.1073/pnas.082099299 (2002).

12 Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* **12**, 77, doi:10.1186/1471-2105-12-77 (2011).

13 Keilwagen, J., Grosse, I. & Grau, J. Area under precision-recall curves for weighted and unweighted data. *PLoS One* **9**, e92209, doi:10.1371/journal.pone.0092209 (2014).

14 Raine, K. M. *et al.* ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. *Current protocols in bioinformatics* **56**, 15 19 11-15 19 17, doi:10.1002/cpbi.17 (2016).

15 Feber, A. *et al.* Using high-density DNA methylation arrays to profile copy number alterations. *Genome Biol* **15**, R30, doi:10.1186/gb-2014-15-2-r30 (2014).

16 van Boerdonk, R. A. *et al.* DNA copy number alterations in endobronchial squamous metaplastic lesions predict lung cancer. *American journal of respiratory and critical care medicine* **184**, 948-956, doi:10.1164/rccm.201102-0218OC (2011).

17 van Boerdonk, R. A. *et al.* DNA copy number aberrations in endobronchial lesions: a validated predictor for cancer. *Thorax* **69**, 451-457, doi:10.1136/thoraxjnl-2013-203821 (2014).

18 Grossman, R. L. *et al.* Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med* **375**, 1109-1112, doi:10.1056/NEJMp1607591 (2016).

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

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