1   **The mutagenic forces shaping the genomic landscape of lung cancer in never**

2   **smokers**

3

4   Marcos Díaz-Gay[1-3,^], Tongwu Zhang[4,^], Phuc H. Hoang[4], Azhar Khandekar[1-4], Wei Zhao[4],

5   Christopher D. Steele[1-3], Burçak Otlu[1-3,5], Shuvro P. Nandi[1-3], Raviteja Vangara[1-3], Erik N.

6   Bergstrom[1-3], Mariya Kazachkova[1-3], Oriol Pich[6], Charles Swanton[6,7], Chao Agnes Hsiung[8], I-

7   Shou Chang[9], Maria Pik Wong[10], Kin Chung Leung[11], Jian Sang[4], John McElderry[4], Lixing

8   Yang[12], Martin A Nowak[13], Jianxin Shi[4], Nathaniel Rothman[4], David C. Wedge[14,15], Robert

9   Homer[16], Soo-Ryum Yang[17], Qing Lan[4], Bin Zhu[4], Stephen J. Chanock[4], Ludmil B.

10 Alexandrov[1-3,18,*], Maria Teresa Landi[4,*]

11

12 [1]Department of Cellular and Molecular Medicine, University of California San Diego, La Jolla,

13 CA, USA

14 [2]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA,

15 [3]Moores Cancer Center, University of California San Diego, La Jolla, CA, USA

16 [4]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA

17 [5]Department of Health Informatics, Graduate School of Informatics, Middle East Technical

18 University, Ankara, Turkey

19 [6]Cancer Evolution and Genome Instability Laboratory, The Francis Crick Institute, London, UK

20 [7]Cancer Research UK Lung Cancer Centre of Excellence, University College London Cancer

21 Institute, London, UK

22 [8]Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan

23 [9]National Institute of Cancer Research, National Health Research Institutes, Zhunan, Taiwan

24 [10]Queen Mary Hospital, The University of Hong Kong, Hong Kong, China

25 [11]Department of Pathology, The University of Hong Kong, Hong Kong, China

26 [12]Ben May Department for Cancer Research, Department of Human Genetics, Comprehensive

27 Cancer Center, The University of Chicago, Chicago, IL, USA

28 [13]Department of Mathematics, Harvard University, Cambridge, MA, USA

29 [14]Manchester Cancer Research Centre, The University of Manchester, Manchester, UK

30 [15]Manchester NIHR Biomedical Research Centre, Manchester, UK

31 [16]Yale Surgery Pathology Department, Yale University, New Haven, CT, USA

32 [17]Department of Pathology and Laboratory Medicine, Memorial Sloan Kettering Cancer Center,

33 New York, USA

34 [18]Sanford Stem Cell Institute, University of California San Diego, La Jolla, CA, USA

35

36 ^These authors contributed equally

37 *Correspondence should be addressed to L2alexandrov@health.ucsd.edu and

38 landim@mail.nih.gov

39

1

ABSTRACT

Lung cancer in never smokers (LCINS) accounts for up to 25% of all lung cancers and has been associated with exposure to secondhand tobacco smoke and air pollution in observational studies. Here, we evaluate the mutagenic exposures in LCINS by examining deep whole-genome sequencing data from a large international cohort of 871 treatment-naïve LCINS recruited from 28 geographical locations within the Sherlock-*Lung* study. *KRAS* mutations were 3.8-fold more common in adenocarcinomas of never smokers from North America and Europe, while a 1.6-fold higher prevalence of *EGFR* and *TP53* mutations was observed in adenocarcinomas from East Asia. Signature SBS40a, with unknown cause, was found in most samples and accounted for the largest proportion of single base substitutions in adenocarcinomas, being enriched in *EGFR*-mutated cases. Conversely, the aristolochic acid signature SBS22a was almost exclusively observed in patients from Taipei. Even though LCINS exposed to secondhand smoke had an 8.3% higher mutational burden and 5.4% shorter telomeres, passive smoking was not associated with driver mutations in cancer driver genes or the activities of individual mutational signatures. In contrast, patients from regions with high levels of air pollution were more likely to have *TP53* mutations while exhibiting shorter telomeres and an increase in most types of somatic mutations, including a 3.9-fold elevation of signature SBS4 (q-value=3.1×10^{-5}), previously linked mainly to tobacco smoking, and a 76% increase of clock-like signature SBS5 (q-value=5.0×10^{-5}). A positive dose-response effect was observed with air pollution levels, which correlated with both a decrease in telomere length and an elevation in somatic mutations, notably attributed to signatures SBS4 and SBS5. Our results elucidate the diversity of mutational processes shaping the genomic landscape of lung cancer in never smokers.

63 **INTRODUCTION**

64 While cancers of the lungs are commonly associated with tobacco smoking[1], prior studies have

65 indicated that between 10 and 25% of the 2.2 million lung cancer cases worldwide are found within

66 individuals that have never smoked tobacco cigarettes[2,3]. The prevalence of lung cancer in never

67 smokers (LCINS) varies by several notable factors, namely enrichment in females[4], Asian

68 populations[5], and individuals with a family history of lung cancer[6]. The occurrence of LCINS can

69 also differ by geographic regions, with higher rates reported in East Asia[7,8] and Eastern Europe[9]

70 when compared to countries in North America and Western Europe[10]. Epidemiological studies

71 have also identified environmental exposures that can increase LCINS risk, including exposure to

72 secondhand smoke[11] and air pollution[12,13].

73

74 Elucidating the mutational signatures operative within a cancer genome allows the understanding

75 of mutational processes implicated in cancer development[14]. Prior analyses have revealed more

76 than 100 characteristic signatures across the spectrum of human neoplasia, with putative etiology

77 assigned to approximately one third of known signatures[15]. However, there has been no

78 comprehensive examination of the mutational signatures operative in LCINS as prior lung cancer

79 whole-genome sequencing studies have focused almost exclusively on tobacco smokers from

80 European descent[16,17] and they have included only a relative small number of LCINS, mostly from

81 East Asian descent[18]. To the best of our knowledge, our prior whole-genome sequencing study

82 encompassing 232 lung cancers from never-smokers[19], which was focused on evolutionary

83 classification and its potential for personalized treatment, has been the largest examination of the

84 mutational landscape in LCINS.

85

86   To understand the mutational processes shaping the lung cancer genomes of never smokers, here,

87   we generated and analyzed deep whole-genome sequencing data from 871 treatment-naïve LCINS,

88   including lung cancer histologies both commonly and rarely found in never-smokers. Our large

89   international cohort of LCINS allowed a comprehensive evaluation of the mutagenic role of

90   secondhand smoke, as well as the examination and comparison of mutational signatures between

91   biological sexes, ancestries, and 28 geographic regions with different levels of air pollution.

RESULTS

The Sherlock-*Lung* never-smoking lung cancer cohort

As part of the Sherlock-*Lung* study[20], a total of 871 treatment-naïve never-smoking lung cancer cases were whole-genome sequenced with a mean coverage of 88x and 35x for tumor and matched germline samples, respectively. Patients were recruited from 28 different locations across four continents (**Fig. 1a**) and were predominantly female (*n*=688; 79.0%). Patients from North America and Europe (NA/EU; *n*=541; 62.1%) mostly clustered with the EUR super-sample from the 1,000 Genomes Project[21] (472/541; 87.2%), and patients from East Asia (AS; *n*=309; 35.5%) clustered exclusively with the EAS super-sample (**Fig. 1b**). Patients of European descent were almost exclusively recruited from Europe, Russia, USA, and Canada, while patients from Asian descent were recruited mainly in Hong Kong, Taipei, Korea, and Canada (**Supplementary Table 1**). Most patients were diagnosed with adenocarcinomas (*n*=737; 84.6%) or carcinoid tumors (*n*=61; 7.0%). In addition, we report the genomic landscapes of 31 squamous cell carcinomas and additional rarer histologies, including 13 adenosquamous carcinomas, 5 large cell carcinomas and 24 tumors with uncertain histological subtype (**Supplementary Table 1**). Information on passive smoking was collected for 458 patients, with 250 exposed and 208 not exposed to secondhand tobacco smoke. No subjects included in the study reported occupational exposure to mutagenic agents.

Mutational patterns of lung cancer in never smokers

The median tumor mutational burden (TMB) was 5,068 for single base substitutions (SBSs; range 393-151,385), 279 for small insertions and deletions (IDs; range 14-10,474), and 105 for structural variants (SV; range 1-1,103 **Fig. 1c**), with differences across histologies and geographical locations. Similarly, the median percentage of genome aberration was 55.3%, with clear variances

5

115   across histologies and geographical areas (**Fig. 1*c***). To identify the mutational processes operative

116   in LCINS, we performed *de novo* extraction[22] of mutational signatures for SBS, ID, doublet base

117   substitutions (DBS), copy number (CN) alterations, and structural variants (SV; **Supplementary**

118   **Tables 2-6**). Eleven *de novo* SBS signatures were extracted and further decomposed into 18

119   previously identified signatures from the Catalogue of Somatic Mutations in Cancer

120   (COSMICv3.4) reference signature database[23] (**Supplementary Fig. 1**; **Supplementary Table 7**).

121   The unprecedented size of the cohort allowed us to refine the prevalence and intensity of the

122   signatures involved in LCINS, with several signatures identified for the first time in LCINS,

123   including signature SBS40a, recently extracted in clear cell renal cell carcinomas but shown to be

124   active in multiple cancer types[15,24]. Other SBS signatures not previously seen in LCINS included

125   unknown etiology signatures SBS12, SBS33, and SBS39 as well as the aristolochic acid-related

126   signature SBS22a and tobacco smoking-associated signatures SBS4 and SBS92 (**Fig. 1*d***).

127   Additionally, the mismatch repair deficiency-associated signatures SBS21 and SBS44 were

128   detected in one lung adenocarcinoma case, which had 10,474 indels and was confirmed by

129   MMRDetect[25] as microsatellite unstable cancer. Analysis of indels and doublet base substitutions

130   revealed a new compendium of mutational signatures operative in LCINS (**Supplementary Figs.**

131   **2-3**; **Supplementary Tables 8-9**). Specifically, a novel ID signature was identified (termed, ID24),

132   characterized by short microhomology deletions, whereas several ID and DBS signatures were

133   identified for the first time in LCINS, including ID6, ID10, ID11, ID14, ID19, ID23, DBS5, DBS6,

134   and DBS13. *De novo* extraction of signatures of large mutational events revealed four CN

135   (**Supplementary Fig. 4**) and four SV (**Supplementary Fig. 5**) signatures, which were

136   subsequently decomposed into the recently described COSMICv3.4 reference signatures[26-29]

137   (**Supplementary Tables 10-11**).

Mutational landscape of lung adenocarcinoma in never smokers

The genomic landscape of the 737 lung adenocarcinomas from never smokers showed a remarkable heterogeneity for both SBS and CN alterations (**Fig. 2a**). Approximately 4.9% of adenocarcinomas (36/737) harbored tobacco-associated signature SBS4, while the most prevalent signature was SBS40a, of unknown etiology, which contributed 28.2% of all substitutions. Interestingly, a cluster of 25 hypermutated samples (>25,000 mutations) with high proportions of SBS4 and APOBEC-associated signatures SBS2 and SBS13 was identified (**Fig. 2a-b**). Additionally, a smaller subset of 11 cases harboring signatures SBS3 and ID6, both linked to homologous recombination deficiency (HRD), was also identified. The presence of HRD was further evaluated using CHORD[30] and HRDetect[31], which jointly predicted six of the cases as HRD, with two additional cases predicted as HRD exclusively by CHORD (**Supplementary Fig. 6**).

Comparisons between patients from different regions revealed differences in somatic mutagenesis (**Fig. 2c-d**). Notably, signature SBS22a (previously known as SBS22, and recently renamed[24]) was enriched in patients from East Asia (**Fig. 2c**). This enrichment was also observed after removing the patients from Canada with EAS ancestry (**Supplementary Fig. 7**; **Methods**), which encompassed most of the patients with EAS ancestry not residing in Asia (35/40, 87.5%; **Fig. 1b**). Further, SBS22a was found almost exclusively in patients from Taipei (32/36 SBS22a-positive cases, 88.9%). Although SBS22a has been associated with aristolochic acid exposure in liver, bladder, and kidney cancers[32,33], this is the first evidence of aristolochic acid causing mutations in lung cancer. Moreover, SBS22a also showed a significant co-occurrence with SBS12 in patients from East Asia (OR=4.55; p-value=1.2×10^{-4}). SBS12 is commonly found in gastrointestinal

7

161 cancers, especially cancers of the livers[15,34] and kidneys[24], with a consistent enrichment of the

162 signature in patients from East Asia[24,34].

163

164 Several signatures of indels and large genomics alterations differed across regions (**Extended**

165 **Data Fig. 1*a***), with ID3, CN20, SV2, SV4, SV6, and SV9 being more prevalent in patients from

166 East Asia, whereas three additional SV signatures, all linked primarily with non-clustered

167 alterations, SV5, SV7, and SV10, were enriched in North American and European patients.

168 Interestingly, CN20 was previously found elevated in Black TCGA patients but not in Asian

169 patients when compared to White patients[26]. Additionally, signatures SBS4 and ID3 showed a

170 difference between males and females across the spectrum of variant classes, with a notably higher

171 prevalence in males, whereas ID9 was enriched in females (**Extended Data Fig. 1*b***).

172

173 As in prior LCINS studies[10], *EGFR* and *TP53* harbored the most driver mutations, with 52.2% and

174 30.5% mutated in adenocarcinomas, respectively, whereas *KRAS* was only mutated in 6.5% of

175 samples. *EGFR* and *TP53* mutations were enriched in patients from East Asia, as previously

176 reported[35], and *KRAS* mutations were elevated in North American and European patients (**Fig. 2*e-***

177 ***f*** and **Extended Data Fig. 1*c-e***). No differences were found for driver mutations between males

178 and females for adenocarcinomas (**Extended Data Fig. 1*f***). Interestingly, for *EGFR*-mutated

179 tumors, a significant increase in TMB was observed for those *TP53*-wild-type (q-

180 value=1.2×10^{-4}), whereas a decrease was found for *TP53*-mutated cases (q-value=0.15;

181 **Extended Data Fig. 1*g***). In addition, while analyzing *EGFR*-mutated cases specifically, we

182 identified a high enrichment in the prevalence of signature SBS40a, along with other signatures of

183 multiple variant types (**Extended Data Fig. 1*h***).

184   Most driver mutations in *EGFR*, *TP53*, and *KRAS* were probabilistically assigned to signatures

185   SBS5 and SBS40a (**Fig. 2g**), with the proportion of driver mutations significantly higher than

186   expected according to their prevalence in the overall adenocarcinoma cohort (OR=2.55; p-

187   value=9.8×10^{-22}; **Fig. 2d**). Prior analyses have shown that SBS4 generates the majority of driver

188   mutations in *KRAS* in tobacco smokers[36]. In our samples, SBS4 contributed only ~10% of KRAS

189   mutations. Most driver indels, especially in *EGFR* (mostly exon 19 deletions; 150/173, 86.7%),

190   were probabilistically assigned to signature ID8 (**Fig. 2h**), which has been previously related to

191   radiation exposure[37]. Despite contributing a significant number of mutations in adenocarcinomas

192   (**Fig. 2d**), APOBEC-associated signatures SBS2 and SBS13 generate a significantly lower number

193   of driver mutations in *EGFR*, *TP53*, and *KRAS* (OR=0.28; p-value=2.8×10^{-14}) in comparison to

194   their prevalence in the overall adenocarcinoma cohort, with most APOBEC-associated mutations

195   being *TP53* point mutations in a small number of samples (**Fig. 2g**).

196

197   **Mutational landscape of lung carcinoids in never smokers**

198   Carcinoid tumors were the second most common histology, with all samples originating from

199   North American and European patients except one case from Taipei (**Supplementary Table 1**),

200   and only one case harboring signature SBS4. Carcinoids presented a lower number of single base

201   substitutions (p-value=3.3×10^{-29}), copy number (p-value=2.6×10^{-32}), and structural variants (p-

202   value=4.2×10^{-26}) compared to adenocarcinomas, as well as longer telomeres (p-value=2.0×10^{-6};

203   **Extended Data Fig. 2**). SBS5 accounted for 55.7% of all mutations in carcinoids in contrast to

204   27.0% in the adenocarcinomas (p-value<2.2×10^{-16}; **Fig. 1d** and **Extended Data Fig. 3a**). In

205   comparison to adenocarcinomas, carcinoids showed 3.39-fold lower TMB (**Extended Data Fig.**

206   **2a**), with depletion of signatures SBS2, SBS5, SBS12, and SBS13 (**Extended Data Fig. 4a**).

9

207   Nevertheless, we observed an enrichment in signature SBS8, previously linked to nucleotide

208   excision repair[38] and late replicating regions[39]. Interestingly, SBS8 was a relatively minor

209   signature in adenocarcinomas (**Fig. 1*d***) but present in the majority of carcinoids (35/61), where it

210   contributed 11.0% of all mutations (**Extended Data Fig. 3*a***). As lung carcinoids are thought to

211   originate from neuroendocrine cells producing hormones and hormone-like substances[40], the

212   presence of SBS8 is perhaps not surprising as this mutational signature is commonly observed in

213   hormone-dependent cancers such as breast[41] and prostate[15] adenocarcinomas.

214

215   Additionally, three ID signatures and the diploid CN1 signature were found enriched in carcinoids,

216   whereas several different signatures were enriched in adenocarcinomas, including two DBS, four

217   ID, and most CN and SV signatures (**Extended Data Fig. 3*b*** and **4*a***), as expected due to the lower

218   number of large genomic aberrations of this histological subtype of LCINS (**Extended Data Fig.**

219   **2*a***). No *EGFR* or *KRAS* driver mutations were found in carcinoid histology, whereas an

220   enrichment was found for driver mutations in *ARID1A* (**Extended Data Fig. 3c** and **4*b-d***).

221

222   **Mutational landscape of lung squamous cell carcinomas in never smokers**

223   Squamous cell carcinomas accounted for 3.6% of LCINS (31/871), with 13 samples harboring

224   tobacco-associated signature SBS4. In comparison with adenocarcinomas, squamous cell

225   carcinomas showed an elevated burden of SBS, DBS, and ID, in contrast to a similar landscape of

226   large genomic alterations and telomere lengths (**Extended Data Fig. 2**). Several signatures showed

227   higher prevalence compared to adenocarcinomas, including SBS3, SBS4, SBS92, and SV7

228   (**Extended Data Fig. 5*a-c*** and **6*a***). *TP53* was the most prevalent driver mutation (58.1% of cases;

229   **Extended Data Fig. 5*d***) and was found enriched in LCINS squamous cell carcinomas compared

230   to adenocarcinomas, along with mutations in *LRP1B*, *PIK3CA*, and *PTEN* (**Extended Data Fig.**

231   **6b-g**). In contrast, *EGFR* mutations were rarely observed in squamous cell carcinomas, as

232   previously reported[42] (**Extended Data Fig. 6d**).

233

234   **SBS4 in lung tumors from never smokers**

235   Tobacco-associated signature SBS4 was found active in 56 LCINS tumors (6.4%), and was the

236   predominant signature in 39 of them (69.6%) and in many hypermutated cases (5.51-fold higher

237   median SBS burden in SBS4+ cases; **Extended Data Fig. 7a-b**). Similarly, signature SBS92,

238   recently linked to tobacco smoking in bladder[22,43] and other cancer types[44], was also active in four

239   SBS4+ tumors, and found significantly enriched in SBS4+ cases (p-value=2.1×10^{-4}; **Extended**

240   **Data Fig. 7b**). Tumors presenting active SBS4 also showed a significant enrichment of previous

241   ID and DBS signatures linked to tobacco smoking, namely ID3 (p-value=1.1×10^{-22}) and DBS2

242   (p-value=5.0×10^{-7}; **Extended Data Fig. 7c-d**). Additionally, SBS4 in LCINS exhibited the same

243   genome topography characteristics as SBS4 in smokers[45], including enrichment in late replicating

244   regions, periodicity in the vicinity of nucleosomes, and strong genic and transcriptional strand

245   asymmetries (**Extended Data Fig. 8**).

246

247   In principle, detecting SBS4 in never-smokers can be due to misannotation of tobacco smoking

248   status. The observed SBS4 prevalence increase in males and lung squamous cell carcinomas

249   indicates that some samples might have been incorrectly annotated as never smokers.

250   Nevertheless, the 56 LCINS tumors with SBS4 exhibited statistically different driver mutations

251   when compared to previously generated lung cancers from tobacco smokers[36] (**Extended Data**

252   **Fig. 7e**). Specifically, the LCINS samples had fewer mutations in *KRAS* (OR=0.41; p-

11

253   value=0.021) and an enrichment of *EGFR* p.L858R hotspot mutations (OR=29.35; p-

254   value=4.1×10^{-4}). These differences indicate that, in addition to possible misannotation, other

255   exogenous processes may be generating SBS4 in these never smokers. As prior work has shown

256   that SBS4 can be found in lifelong non-smokers exposed to different environmental mutagens,

257   including occupational history of coal tar work[46] and indoor air pollution[47], we evaluated whether

258   SBS4 in this cohort, as well as mutations in other samples, can be generated by either secondhand

259   exposure to tobacco smoking or by high levels of air pollution.

260

261   **Passive smoking has low mutagenicity in LCINS**

262   To understand the mutational processes activated by secondhand tobacco smoking, we compared

263   the genomic landscapes of 250 LCINS from passive smokers to 208 LCINS from individuals who

264   were not exposed to secondhand smoke. We observed an increase of SBS mutations in passive

265   smokers (**Fig. 3*a***), as well as a decrease in tumor/normal telomere length ratio (**Fig. 3*b***). Although

266   not significant, the directionality of these associations with passive smoking was consistent after

267   adjustment for other covariates, encompassing age, sex, genetic ancestry, histology, and tumor

268   purity – 8.3% increase and 5.4% decrease in the magnitude of regression coefficients after

269   covariate correction, respectively (**Fig. 3*c-d***). Almost identical results were observed when we

270   restricted the analyses to the lung adenocarcinomas from never smokers (**Supplementary Fig. 8**),

271   and when the principal components from the ancestry analysis were used instead of the ancestry

272   labels for adjustment (**Supplementary Table 12**; **Methods**). However, this increase in mutation

273   burden was not specifically associated with any mutational signature (**Fig. 3*e***) or mutation type

274   (**Fig. 3*f***). In addition, no significant differences were found for mutations in any cancer driver

275   genes (**Fig. 3*g***), nor for the burden of ID, DBS, SV, or CN segments (**Extended Data Fig. 9*a-e***)

276   or the presence of signatures derived from these mutation types (**Extended Data Fig. 9f**). Amongst

277   the 250 cases identified as exposed to secondhand smoke, only three (1.2%) displayed signature

278   SBS4 (OR=0.62; p-value=0.71), each with a contribution exceeding 20% and at least 500

279   mutations attributed to SBS4. Moreover, only two of the 281 driver mutations found in secondhand

280   smoke-exposed samples were assigned to SBS4 with a probability above 50%, including a *TP53*

281   missense mutation (p.Val157Phe) and a missense mutation (p.Val409Leu) in *NF1*.

282

283   To evaluate if we were unable to assign SBS4 to other secondhand smoke-exposed cases due to

284   low levels of the signature, we synthetically injected SBS4 mutations using simulations at different

285   levels into the mutational profiles of the 247 passive smoker cases that lacked SBS4 and assessed

286   the number of samples where SBS4 was detected (**Supplementary Fig. 9**). We were able to detect

287   SBS4 in around a quarter of simulated samples, if it contributed at least 5% of all mutations within

288   a sample. Moreover, SBS4 was detectable in almost every sample when the signature accounted

289   for more than 10% of mutations. These simulation results agree with the levels of SBS4 observed

290   in the set of SBS4+ samples, where SBS4 is contributing above 10%. Thus, although SBS4 could

291   have been missed in some secondhand smoke exposed LCINS, this signature would be likely

292   contributing less than 5% of mutations in these cancers. Overall, our simulations demonstrate that

293   it is unlikely that exposure to secondhand smoke accounts for SBS4, whereas the modest increase

294   in overall SBS mutations suggests that passive smoking has low mutagenicity.

295

296   **Air pollution is associated with increased somatic mutations in LCINS**

297   Given the recent evidence of the role of air pollution in lung carcinogenesis[12,48], we assessed

298   whether atmospheric air pollution, quantified by the environmental particulate matter measuring

13

299 ≤2.5 μm (PM$_{2.5}$), had an effect on the accumulation of somatic mutations in LCINS. Yearly PM$_{2.5}$

300 estimates from 1998–2021 were obtained for the 853 LCINS cases with annotated country of

301 residence using a hybrid model combining satellite-based measurements of aerosol optical depth

302 with chemical transport modeling and ground-based observations (**Methods**)[49]. Considering the

303 distribution of PM$_{2.5}$ estimates across the cohort, a threshold of 20 μg/m^3 was used to separate

304 patients diagnosed in regions with high and low levels of pollution (**Supplementary Fig. 10**).

305 LCINS originating from areas with high PM$_{2.5}$ levels exhibited increased burdens of SBS (13.0%),

306 DBS (8.3%), and ID mutations (7.6%; **Fig. 4a**), along with telomere shortening (**Fig. 4b**), even

307 after accounting for age, sex, genetic ancestry, histology, and tumor purity (**Fig. 4c-d**;

308 **Supplementary Table 12**). No enrichments were identified for the numbers of SV or CN segments

309 after corrections (**Extended Data Fig. 10a-b**). In order to determine whether higher levels of

310 pollution lead to an increase in somatic mutagenesis (*i.e.*, a dose-response effect), we calculated

311 individual estimates of PM$_{2.5}$ per lung cancer patient (**Methods**). Consistent with a dose-response

312 effect, we found statistically significant positive correlations of PM$_{2.5}$ with the burden of SBS,

313 DBS, and ID (**Fig. 4e-g**), as well as a negative correlation with telomere length (**Fig. 4h**), further

314 corroborating the dose-response relationship of outdoor air pollution and somatic mutations in

315 LCINS.

316

317 Several mutational signature-specific associations with individual PM$_{2.5}$ estimates were found

318 after adjusting for covariates (**Fig. 5a**). These included clock-like signature SBS5 (OR for 10 μg/m^3

319 of PM$_{2.5}$=1.76; q-value=5.0 × 10^{-5}) as well as signatures SBS4 (OR=3.86; q-value=3.1 × 10^{-5}) and

320 ID3 (OR=1.63; q-value=5.1 × 10^{-4}; **Fig. 5b**; **Supplementary Table 12**). Furthermore, dose-

321 response effects were also observed for these signatures, with a significant positive correlation

14

322   between the total numbers of mutations assigned to a given signature and the individual $PM_{2.5}$

323   estimates (**Fig. 5*c-e***). Specifically, a unit ($1\mu g/m^3$) increase of $PM_{2.5}$ was associated with a 2.3%

324   increase in SBS5-associated mutations, 12.0% in SBS4-associated mutations, and 6.0% in ID3-

325   associated mutations, independent of other covariates. No associations were found with DBS, CN,

326   or SV signatures (**Fig. 5*a***; **Extended Data Fig. 10*c***). Patients in regions with high $PM_{2.5}$ exposure

327   were 1.6 times more likely to have *TP53* mutations and 2.5 times less likely to have *CTNNB1*

328   mutations (**Fig. 5*f-h***). Overall, these results indicate that elevated air pollution levels are associated

329   with an increase in both somatic mutations contributing to specific mutational signatures as well

330   as an increased prevalence of *TP53* mutations.

331

332 **DISCUSSION**

333   This study presents the most extensive mutational analyses of whole-genome sequenced lung

334   cancers in never smokers. Our results reveal multiple mutational signatures and mutated cancer

335   driver genes whose frequencies differ across lung cancer histologies, regions, and exposures.

336   Particularly noteworthy is our finding of the aristolochic acid signature SBS22a in patients from

337   Taipei, possibly shedding light on one of the previous unappreciated environmental factors

338   contributing to the accumulation of mutations in LCINS from East Asia. Similarly, multiple

339   signatures of indels and large genomics alterations, most with unknown etiology, were found to

340   differ across regions, further highlighting the likely existence of additional population-specific

341   exposures and other factors contributing to LCINS. Notably, SBS40a, with an unknown cause, and

342   previously associated with kidney cancer[24], accounted for the largest proportion of substitutions in

343   adenocarcinomas, thus indicating the existence of a globally widespread and previously

344   unappreciated mutagenic process, likely of endogenous origin.

345

346   Prior epidemiological research[11] has established that passive smoking can be associated with a

347   modest heightened risk for developing LCINS. In line with the observational research, our analysis

348   of 458 patients with known exposure to passive smoking uncovered only minimal indications of

349   elevated mutagenesis amongst individuals exposed to secondhand smoke. As in previous

350   epidemiological studies, the data about second-hand tobacco smoking exposure was derived from

351   questionnaires and clinical datasets, and varied across individuals, going from very detailed

352   (during childhood, at work, or with a smoking spouse) to just binary yes/no exposure status. Thus,

353   low exposure levels or very remote exposure (during childhood) may contribute to our findings of

354   low mutagenicity. Nevertheless, our results are also consistent with a recent study in 291 Japanese

16

355   LCINS patients with extensive information on passive smoking[50]. Additional non-mutagenic

356   carcinogenesis could be contributing to the observed rise in lung cancer in those exposed to

357   secondhand smoke.

358

359   Our investigation of the mutagenic role of outdoor air pollution relied on an average country- and

360   state/province-level quantification of $PM_{2.5}$ that did not provide granularity within these

361   geographical regions or consideration for seasonal variabilities or indoor air pollution levels. The

362   exposure was linked to the subjects' residence at the time of lung cancer diagnosis and may not

363   reflect their lifetime exposure, and the assigned exposures for the patients recruited to this study

364   may systematically differ from the average country, state, province, or city level pollution levels.

365   As such, these associations should be interpreted with caution. Despite these limitations, our

366   findings suggest that LCINS in individuals from areas with elevated air pollution have a higher

367   prevalence of *TP53* mutations while exhibiting shorter telomeres and heightened mutagenesis,

368   notably attributed to signatures SBS4 and SBS5. Consistent with our findings, a previous study

369   comparing Chinese regions with different levels of air pollution reported an increase in *TP53* and

370   total somatic mutations in the highly polluted region of Xuanwei, with a C>A enriched mutational

371   pattern resembling signature SBS4[51].

372

373   In principle, mutations linked to clock-like signature SBS5 that accumulate in LCINS with the

374   increase of $PM_{2.5}$ could be due to additional DNA damage that accumulates in people living in

375   more polluted areas. In addition, considering the significant telomere shortening observed with

376   high levels of $PM_{2.5}$ exposure, the clock-like somatic mutations could also be a readout of a

377   promotion mechanism where lung cells are undergoing more cell divisions in individuals residing

17

378   in highly polluted areas. Consistently, a recent experimental study[48] showed that air pollution acts

379   as a tumor-promoting inflammatory agent, demonstrating that short-term exposure to $PM_{2.5}$ is not

380   strongly mutagenic in mice, although leading to an elevation of signature SBS5 and reduced

381   telomeres (**Extended Data Fig. 11**). Future studies with long-term $PM_{2.5}$ exposures and higher

382   number of replicates will be required to experimentally validate the role of outdoor air pollution

383   as a driver of both mutations and inflammation leading to tumorigenesis.

384

385   While the association between SBS4 and $PM_{2.5}$ potentially encompasses the effects of atmospheric

386   pollution exposure in LCINS, the small proportion of cases affected by SBS4 (6.4% of all samples)

387   and the high number of mutations in SBS4+ cancers (21,785 on average) could indicate extreme

388   events of outdoor, indoor, or occupational air pollution. Indeed, such events have been previously

389   reported, for example, SBS4 was found in the lung cancers of never-smoking Chinese women that

390   often cook with smoky coal in poorly ventilated houses[47], and it is likely that severe levels of

391   pollution are more common in countries with overall higher air pollution levels. Lastly, it is also

392   possible that the self-reported smoking status was inaccurate, as previously shown in other cohorts

393   after biochemical verification[52], and that some SBS4 cases from smokers were misannotated as

394   never smokers, albeit it is unlikely that such misannotation will exhibit any correlation with air

395   pollution. Notwithstanding, the reported associations between air pollution and somatic mutations

396   remained unaffected when all SBS4 mutations were excluded from our analysis (**Extended Data**

397   **Fig. 12**).

398

399   Overall, the unprecedented size of our global cohort allowed us to refine the prevalence and

400   intensity of the mutational signatures and driver mutations involved in LCINS, and their variability

18

401   across histologies, genetic ancestries, and geographical locations. Although LCINS appear to be

402   dominated by endogenous processes, our results indicate exposure to aristolochic acids in Taipei's

403   samples, a low mutagenicity of passive smoking, and a role of air pollution as both a mutagenic

404   initiator and promoter of neoplastic expansion in LCINS.

405

19

FIGURE LEGENDS

Fig. 1. Overview of the Sherlock-*Lung* cohort of lung cancers in never smokers. a, Geographical distribution of the 871 patients across four continents and 28 geographic locations. **b**, Clinical characterization based on histology, genetic ancestry, geographical region, biological sex, and passive smoking status. **c**, Prevalence of mutations, percentage of genome altered, and structural variants across geographic locations and stratified based on histology. Left panel, dots represent median values for the three genomic alterations individually per country and histology. Right panel, dots represent individual tumors, colors different histology types, and horizontal purple lines median values across all histologies. **d**, Landscape of mutational signatures across histologies and somatic variant classes, including single base substitutions (SBS), doublet base substitutions (DBS), small insertions and deletions (ID), copy number alterations (CN), and structural variants (SV). The size and color of the dots represent the percentage of mutations contributed by the signature in all samples sharing the same histology (top panel) or the percentage of samples of a histological type where a particular signature is active (bottom panel). AS: East Asian geographical regions, EAS: East Asian genetic ancestry super-sample, EUR: European genetic ancestry super-sample, LUAD: lung adenocarcinomas, LUSC: lung squamous cell carcinomas, NA/EU: North America and Europe geographical regions.

Fig. 2 Repertoire of mutational signatures and driver mutations in LCINS adenocarcinomas.
a, Distribution of the most prevalent signatures per sample according to the number of single base substitutions (SBS) and percentage of genome aberrated. **b**, Activity of SBS mutational signatures across samples, representing the total number of mutations attributed to each signature in a given sample. Dots represent individual samples and purple horizontal bars median values. The numbers

20

429   on the bottom indicate the total number of samples where a particular signature was found active

430   (blue) and the total number of LCINS adenocarcinoma samples (green). **c**, Regional differences

431   across signatures. Volcano plot indicating enrichment of SBS signatures in patients from East

432   Asian (AS) and North American/European regions (NA/EU) in LCINS adenocarcinomas (top

433   panel) and bar plot indicating prevalence by geographical region (bottom panel). Horizontal lines

434   marking statistically significant thresholds were included at 0.05 (dashed orange line) and 0.01

435   FDR levels (dashed red line). **d**, Total number of mutations assigned to specific SBS signatures

436   for patients from East Asia or North America and Europe. **e**, Volcano plot indicating enrichment

437   of mutations in driver genes affecting specific LCINS adenocarcinomas. Blue-colored genes were

438   enriched in patients from North America and Europe, whereas red-colored genes were enriched in

439   patients from East Asia. **f**, Frequency of LCINS adenocarcinoma cases harboring driver mutations

440   in the driver genes significantly differently mutated between geographical regions (*EGFR*, *TP53*,

441   and *KRAS*) **g**, Proportion of driver mutations affecting *EGFR*, *TP53*, and *KRAS* probabilistically

442   assigned to each of the SBS mutational signatures identified in the LCINS adenocarcinoma cohort.

443   The numbers indicate the total number of driver single base substitutions found in each of the

444   genes.

445

446   **Fig. 3. Passive smoking influence in the genomic landscape of LCINS. a**-**d**, Differences in base

447   substitution burden and tumor-to-normal telomere length ratio using univariate comparisons (**a**

448   and **b**) as well as multivariable linear regressions considering clinical and epidemiological

449   covariates (**c** and **d**), including age, sex, genetic ancestry, histology, and tumor purity. **e**, Volcano

450   plot indicating enrichment of mutational signatures derived from SBS mutations in passive vs.

451   non-passive smokers. Horizontal lines marking statistically significant thresholds were included

452   at 0.05 (dashed orange line) and 0.01 FDR levels (dashed red line). **f**, Comparison of the mutations

453   belonging to each of the six main SBS mutation subtypes in passive vs. non-passive smokers. **g**,

454   Volcano plot indicating enrichment of mutations in driver genes affecting specific LCINS tumors.

455

456   **Fig. 4. Mutagenic effects of PM$_{2.5}$ exposure in LCINS. a**, Quantification of tumor mutational

457   burden according to different mutation types, including SBS, DBS, and ID, for patients living in

458   geographical regions with high and low PM$_{2.5}$ exposure levels (threshold defined at 20 μg/m^3; only

459   samples for which the country of origin was known (*n*=853) are included). **b**, Quantification of the

460   ratio of telomere lengths for tumor and normal samples across high and low PM$_{2.5}$ exposed cases.

461   **c-d**, Forest plots corresponding to multivariable linear regressions considering high/low PM$_{2.5}$

462   exposure group, age, sex, genetic ancestry, histology, and tumor sample purity as covariates and

463   tumor mutational burden for the specific mutation type, SBS, DBS, ID (**c**), or telomere length ratio

464   (**d**) as independent variables. **e-h**, Scatter plots showing significant correlations between individual

465   sample estimates of PM$_{2.5}$ exposure and tumor mutational burden for SBS (**e**), DBS (**f**), ID (**g**), and

466   telomere length ratio (**h**).

467

468   **Fig. 5. Associations between PM$_{2.5}$ exposure and specific mutational signatures affecting**

469   **LCINS tumors. a**, Enrichment analysis of the presence of mutational signatures derived from

470   SBS, DBS, and ID with PM$_{2.5}$ exposure levels for all samples for which the country of origin was

471   known (*n*=853). Horizontal lines marking statistically significant thresholds were included at 0.05

472   (dashed orange line) and 0.01 FDR levels (dashed red line). The odds ratios for the

473   increase/decrease of 10 μg/m^3 of PM$_{2.5}$ estimates are shown. **b**, Detailed forest plots for the logistic

474   regression models corresponding to signatures SBS4, SBS5, and ID3. **c-e**, Scatter plots assessing

22

475   dose-response effect between individual sample estimates of PM$_{2.5}$ exposure and mutations

476   assigned to signatures SBS4 (**c**), SBS5 (**d**), and ID3 (**e**). **f**, Enrichment analysis of the presence of

477   mutations in driver genes with PM$_{2.5}$ exposure levels. **g-h**, Detail of the enrichment of *TP53* (**g**)

478   and *CTNNB1* (**h**) driver mutations in high and low pollution regions, respectively.

479

480　**EXTENDED DATA FIGURE LEGENDS**

481　**Extended Data Fig. 1. Association of mutational signature prevalence and driver mutations**

482　**with geographical regions, biological sexes, and *EGFR* mutation status in LCINS**

483　**adenocarcinoma cases**. **a**, DBS, ID, CN, and SV mutational signatures enrichment analysis with

484　geographical regions. Horizontal lines marking statistically significant thresholds were included

485　at 0.05 (dashed orange line) and 0.01 FDR value levels (dashed red line). Blue-colored signatures

486　were enriched in North American and European patients, whereas red-colored signatures were

487　enriched in East Asian patients. **b,** SBS, DBS, ID, CN, and SV mutational signatures enrichment

488　analysis with biological sexes. Blue-colored signatures were enriched in males, whereas red-

489　colored signatures were enriched in females. **c-e**, Detail of the enrichment of *EGFR* (**c**), *TP53* (**d**),

490　and *KRAS* (**e**) driver mutations in North American and European vs. East Asian LCINS

491　adenocarcinoma cases. **f**, Driver mutations enrichment analysis with biological sexes. Blue-

492　colored genes were enriched in males, whereas red-colored genes were enriched in females. **h**,

493　SBS, DBS, ID, CN, and SV mutational signatures enrichment analysis with *EGFR* mutation status.

494　Blue-colored signatures were enriched in *EGFR* mutant tumors, whereas red-colored signatures

495　were enriched in *EGFR* wild-type tumors.

496

497　**Extended Data Fig. 2. Tumor mutational burden differences in LCINS across histologies. a-**

498　**b**, Quantification of tumor mutational burden according to different mutation types, including SBS,

499　DBS, ID, number of copy number segments, and structural variants (**a**), as well as telomere length

500　ratios between tumor and normal samples (**b**) across histologies.

501

502   **Extended Data Fig. 3. Repertoire of mutational signatures and driver mutations in LCINS**

503   **carcinoids. a-b**, Mutational signature landscape for SBS (**a**) and ID (**b**) mutation types, including

504   absolute and relative number of mutations assigned to each mutational signature, unsupervised

505   clustering based on the signature contributions, and sample-level annotations of sex, genetic

506   ancestry, and accuracy of signature reconstruction based on cosine similarity. **c**, Driver mutations

507   landscape, including different types of genomic alterations, as well as sample-level annotations of

508   sex, genetic ancestry, histology, and tumor purity.

509

510   **Extended Data Fig. 4. Association of mutational signature prevalence and driver mutations**

511   **with adenocarcinoma and carcinoid histology in LCINS cases**. **a-b** SBS, DBS, ID, CN, and SV

512   mutational signatures, and driver mutations (**b**) enrichment analysis. Horizontal lines marking

513   statistically significant thresholds were included at 0.05 (dashed orange line) and 0.01 FDR value

514   levels (dashed red line). Blue-colored signatures/genes were enriched in adenocarcinomas,

515   whereas green-colored signatures/genes were enriched in carcinoids. **c-d**, Detail of the enrichment

516   of *TP53* (**c**) and *ARID1A* (**d**) driver mutations in carcinoid vs. adenocarcinoma LCINS.

517

518   **Extended Data Fig. 5. Genomic landscape in LCINS squamous cell carcinomas. a-c**,

519   Mutational signature landscape for SBS (**a**), DBS (**b**), and ID (**c**) mutation types, including

520   absolute and relative number of mutations assigned to each mutational signature, unsupervised

521   clustering based on the signature contributions, and sample-level annotations of sex, genetic

522   ancestry, and accuracy of signature reconstruction based on cosine similarity. **d**, Driver mutations

523   landscape, including different types of genomic alterations, as well as sample-level annotations of

524   sex, genetic ancestry, histology, and tumor purity.

525

526 **Extended Data Fig. 6. Association of mutational signature prevalence and driver mutations**

527 **with adenocarcinoma and squamous cell carcinoma histology in LCINS cases**. **a-b** SBS, DBS,

528 ID, CN, and SV (**a**) mutational signatures, and driver mutations (**b**) enrichment analysis.

529 Horizontal lines marking statistically significant thresholds were included at 0.05 (dashed orange

530 line) and 0.01 FDR value levels (dashed red line). Blue-colored signatures/genes were enriched in

531 adenocarcinomas, whereas red-colored signatures/genes were enriched in squamous cell

532 carcinomas. **c-f**, Detail of the enrichment of *TP53* (**c**), *EGFR* (**d**), *LRP1B* (**e**), *PTEN* (**f**) and

533 *PIK3CA* (**g**) driver mutations in squamous cell carcinoma vs. adenocarcinoma LCINS.

534

535 **Extended Data Fig. 7. Genomic landscape of 56 LCINS tumors presenting SBS4 activity. a**,

536 Tumor mutational burden differences between SBS4 positive and negative LCINS tumors for SBS,

537 DBS, ID, CN segments, and SV events. **b-d**, Mutational signature landscape for SBS (**b**), DBS (**c**),

538 and ID (**d**) mutation types, including absolute and relative number of mutations assigned to each

539 mutational signature, unsupervised clustering based on the signature contributions, and sample-

540 level annotations of sex, genetic ancestry, passive smoking, and accuracy of signature

541 reconstruction based on cosine similarity. **e**, Driver mutations landscape, including different types

542 of genomic alterations, as well as sample-level annotations of sex, genetic ancestry, histology, and

543 tumor purity.

544

545 **Extended Data Fig. 8. Topographical characteristics of 56 LCINS and 68 lung cancer from**

546 **smokers presenting SBS4 activity. a-b**, Distribution of SBS4 mutations with replication timing

547 in our cohort of never smokers (**a**) and in the smokers from the PCAWG cohort (**b**). Data are

548 separated into deciles, with each segment harboring 10% of the observed replication time signal

549 in the x-axis, and the normalized mutational density displayed in the y-axis. **c-d**, Association of

550 SBS4 mutations with nucleosome occupancy in never smokers (**c**) and smokers (**d**). The solid blue

551 line represents real somatic mutations, whereas the dashed grey line indicates the distribution of

552 simulated mutations. Both lines show the average nucleosome signal in the y-axis, using a genomic

553 window of 2 kilobases centered around the SBS4-associated mutations in the x-axis. **e-f**, Strand

554 asymmetry of SBS4-associated mutations in comparison to simulations and considering lagging

555 and leading DNA strands, transcribed and untranscribed DNA regions and genic and intergenic

556 genomic locations in never smokers (**e**) and smokers (**f**).

557

558 **Extended Data Fig. 9. Passive smoking influence in the landscape of ID, DBS, CN, and SV in**

559 **LCINS. a-e**, Differences in DBS, ID, CN, and SV burden using univariate comparisons (**a**) as well

560 as multivariable linear regressions considering clinical and epidemiological covariates (**b-e**),

561 including age, sex, genetic ancestry, and tumor purity. **f**, Volcano plots indicating enrichment of

562 mutational signatures derived from DBS, ID, CN, and SV alterations. Horizontal lines marking

563 statistically significant thresholds were included at 0.05 (dashed orange line) and 0.01 FDR value

564 levels (dashed red line).

565

566 **Extended Data Fig. 10. Effects of PM$_{2.5}$ exposure in large genomic alterations in LCINS. a-**

567 **b**, Differences in the number of CN segments and SV events using univariate comparisons (**a**) as

568 well as multivariable linear regressions, considering clinical and epidemiological covariates (**b**),

569 including age, sex, genetic ancestry, histology, and tumor purity, for patients diagnosed in

570 geographical regions with high and low PM$_{2.5}$ exposure levels (threshold defined at 20 μg/m^3; only

27

571 samples for which the country of origin was known, *n*=853, were included). **c**, Volcano plots

572 indicating enrichment of mutational signatures derived from CN and SV alterations. Horizontal

573 lines marking statistically significant thresholds were included at 0.05 (dashed orange line) and

574 0.01 FDR value levels (dashed red line).

575

576 **Extended Data Fig. 11. Assignment of mutational signatures and estimation of telomere**

577 **length using data from control (*n*=5) and PM$_{2.5}$-exposed mice (*n*=5) from Ref.[48]. a-b**, Boxplots

578 comparing the mutations assigned to SBS5 (**a**) and the estimations for the telomere length ratio

579 between the tumor and normal samples (**b**). Student's t-tests were used to calculate statistical

580 significance.

581

582 **Extended Data Fig. 12. Mutagenic effects of PM$_{2.5}$ exposure in LCINS cases excluding SBS4**

583 **contributions. a**, Quantification of SBS burden excluding SBS4 mutations for patients living in

584 geographical regions with high and low PM$_{2.5}$ exposure levels (threshold defined at 20 µg/m^3; only

585 samples for which the country of origin was known, *n*=853, were included). **b**, Forest plot

586 corresponding to a multivariable linear regression considering high/low PM$_{2.5}$ exposure group,

587 age, sex, genetic ancestry, histology, and tumor sample purity as covariates and SBS burden as

588 independent variable. **c**, Scatter plot showing a significant correlation between individual sample

589 estimates of PM$_{2.5}$ exposure and SBS burden.

590   **SUPPLEMENTARY MATERIAL LEGENDS**

591   **Supplementary Table 1. Clinical characteristics of the Sherlock-*Lung* never-smoking lung**

592   **cancer cohort by histology.**

593

594   **Supplementary Table 2. Mutational profiles for SBS *de novo* extracted mutational**

595   **signatures using the SBS-288 mutational context.**

596

597   **Supplementary Table 3. Mutational profiles for ID *de novo* extracted mutational signatures**

598   **using the ID-83 mutational context.**

599

600   **Supplementary Table 4. Mutational profiles for DBS *de novo* extracted mutational**

601   **signatures using the DBS-78 mutational context.**

602

603   **Supplementary Table 5. Mutational profiles for CN *de novo* extracted mutational**

604   **signatures using the CN-68 mutational context.**

605

606   **Supplementary Table 6. Mutational profiles for SV *de novo* extracted mutational**

607   **signatures using the SV-38 mutational context.**

608

609   **Supplementary Table 7. Decomposition of SBS *de novo* mutational signatures into**

610   **COSMICv3.4 reference signatures.**

611

29

612   **Supplementary Table 8. Decomposition of ID *de novo* mutational signatures into**

613   **COSMICv3.4 reference signatures.**

614

615   **Supplementary Table 9. Decomposition of DBS *de novo* mutational signatures into**

616   **COSMICv3.4 reference signatures.**

617

618   **Supplementary Table 10. Decomposition of CN *de novo* mutational signatures into**

619   **COSMICv3.4 reference signatures.**

620

621   **Supplementary Table 11. Decomposition of SV *de novo* mutational signatures into**

622   **COSMICv3.4 reference signatures.**

623

624   **Supplementary Table 12. Associations of passive smoking and pollution with genomic**

625   **features adjusted by genetic ancestry principal components.**

626

627   **Supplementary Table 13. dbGaP and EGA unique identifiers for the publicly available**

628   **datasets included as part of the analyzed LCINS cohort.**

629

630   **Supplementary Fig. 1. *De novo* mutational signatures extracted using the SBS-288**

631   **mutational context**. **a**, Mutational profiles of the *de novo* extracted signatures, with indication of

632   the cosine similarity of the decomposition into COSMICv3.4 reference signatures. **b**, Contribution

633   of the different COSMICv3.4 reference mutational signatures after decomposition of the *de novo*

634   extracted signatures. **c**, Activity of *de novo* mutational signatures across samples, representing the

635 total number of substitutions attributed to each signature in a given sample. Dots represent

636 individual samples, colors different histology types, and purple horizontal bars median values

637 across all histologies. The numbers on top indicate the total number of samples where a particular

638 signature was found active (blue) and the total number of samples of the assessed cohort (green).

639

640 **Supplementary Fig. 2.** *De novo* **mutational signatures extracted using the ID-83 mutational**

641 **context**. **a**, Mutational profiles of the *de novo* extracted signatures. **b**, Contribution of the different

642 COSMICv3.4 reference mutational signatures after decomposition of the *de novo* extracted

643 signatures. **c**, Activity of *de novo* mutational signatures across samples, representing the total

644 number of indels attributed to each signature in a given sample. Dots represent individual samples,

645 colors represent different histology types, and purple horizontal bars represent median values

646 across all histologies. The numbers on top indicate the total number of samples where a particular

647 signature was found active (blue) and the total number of samples of the assessed cohort (green).

648

649 **Supplementary Fig. 3.** *De novo* **mutational signatures extracted using the DBS-78 mutational**

650 **context**. **a**, Mutational profiles of the *de novo* extracted signatures. **b**, Contribution of the different

651 COSMIC v3.4 reference mutational signatures after decomposition of the *de novo* extracted

652 signatures. **c**, Activity of *de novo* mutational signatures across samples, representing the total

653 number of doublets attributed to each signature in a given sample. Dots represent individual

654 samples, colors represent different histology types, and purple horizontal bars represent median

655 values across all histologies. The numbers on top indicate the total number of samples where a

656 particular signature was found active (blue) and the total number of samples of the assessed cohort

657 (green).

658

659   **Supplementary Fig. 4.** *De novo* **mutational signatures extracted using the CN-68 mutational**

660   **context**. **a**, Mutational profiles of the *de novo* extracted signatures. **b**, Contribution of the different

661   COSMICv3.4 reference mutational signatures after decomposition of the *de novo* extracted

662   signatures. **c**, Activity of *de novo* mutational signatures across samples, representing the total

663   number of copy number segments attributed to each signature in a given sample. Dots represent

664   individual samples, colors represent different histology types, and purple horizontal bars represent

665   median values across all histologies. The numbers on top indicate the total number of samples

666   where a particular signature was found active (blue) and the total number of samples of the

667   assessed cohort (green).

668

669   **Supplementary Fig. 5.** *De novo* **mutational signatures extracted using the SV-38 mutational**

670   **context**. **a**, Mutational profiles of the *de novo* extracted signatures. **b**, Contribution of the different

671   COSMIC v3.4 reference mutational signatures after decomposition of the *de novo* extracted

672   signatures. **c**, Activity of *de novo* mutational signatures across samples, representing the total

673   number of structural variants attributed to each signature in a given sample. Dots represent

674   individual samples, colors represent different histology types, and purple horizontal bars represent

675   median values across all histologies. The numbers on top indicate the total number of samples

676   where a particular signature was found active (blue) and the total number of samples of the

677   assessed cohort (green).

678

679   **Supplementary Fig. 6. Validation of homologous recombination deficient cases in LCINS**

680   **using computational predictors**. The x-axis shows the probability prediction scores for all the

681   LCINS cases in the cohort (categorized by histology and presence of signature SBS3) using

682   HRDetect[31], whereas the y-axis shows the probability prediction scores for CHORD[30]. Dashed

683   lines showed the thresholds proposed by both computational tools for considering a sample as

684   homologous recombination deficient or proficient.

685

686   **Supplementary Fig. 7. Sensitivity analysis of the associations of mutational signatures with**

687   **geographical regions excluding Canadian patients with EAS genetic ancestry.** Volcano plot

688   indicating enrichment of SBS signatures in patients from East Asian (AS) and North

689   American/European regions (NA/EU) in LCINS adenocarcinomas (top panel) and bar plot

690   indicating prevalence by geographical region (bottom panel). Horizontal lines marking

691   statistically significant thresholds were included at 0.05 (dashed orange line) and 0.01 FDR

692   levels (dashed red line).

693

694   **Supplementary Fig. 8. Passive smoking influence in the genomic landscape of lung**

695   **adenocarcinomas from never smokers. a-d**, Differences in base substitution burden and

696   telomere length ratio between tumor and normal samples using univariate comparisons (**a** and **b**)

697   as well as multivariable linear regressions considering clinical and epidemiological covariates (**c**

698   and **d**), including age, sex, genetic ancestry, and tumor purity.

699

700   **Supplementary Fig. 9. Assessment of the resolution to assign signature SBS4 in cases exposed**

701   **to secondhand smoke.** COSMICv3.4 mutational signatures obtained for the whole cohort of

702   LCINS patients were assigned to 247 tumors from SBS4-negative patients exposed to secondhand

703   smoke after the synthetic injection of SBS4 at different levels (1%, 2%, 5%, 10%, 15%, and 20%;

704   100 simulations per injection level). **a**, Boxplots showing the distribution of the proportion of

705   samples where SBS4 was detected for different simulations within a synthetic injection level. **b**,

706   Detail of the SBS4 contributions observed in the synthetic samples, where each row corresponds

707   to a sample, each column to a simulation corresponding to a specific injection level, and the color

708   represents the contribution level of SBS4 to the overall mutational profile.

709

710   **Supplementary Fig. 10. Split of samples in pollution exposure groups according to their levels**

711   **of estimated population weighted PM$_{2.5}$.** Limit used for the two-group classification (high-low;

712   20 μg/m^3; 440 high exposed samples vs. 413 low exposed samples).

713 **ONLINE METHODS**

714 **Ethics declarations**

715 Since the National Cancer Institute only received de-identified samples and data from

716 collaborating centers, had no direct contact or interaction with the study participants, and did not

717 use or generate identifiable private information, Sherlock-*Lung* has been determined to constitute

718 "Not Human Subject Research (NHSR)" based on the federal Common Rule (45 CFR 46;

719 https://www.ecfr.gov/cgi-bin/ECFR?page=browse).

720

721 **Collection of lung cancer samples**

722 Fresh-frozen tumor tissue and matched germline DNA from whole-blood samples or fresh-frozen

723 normal lung tissue sampled approximately 3 cm from the tumor were obtained from 871 treatment-

724 naïve lung cancer patients from 14 institutions/centers across the world after sample and

725 sequencing quality control. Among these patients, 114 were from the Institut universitaire de

726 cardiologie et de pneumologie de Québec – Université Laval (IUCPQ-UL), Quebec, Canada; 54

727 from Université Côte d'Azur, Nice, France; 25 from the EAGLE study from Italy[53]; 22 from Yale

728 University, New Haven, Connecticut, USA; 11 from H. Lee Moffitt Cancer Center & Research

729 Institute, Tampa, Florida, USA; 27 from Harvard University, Cambridge, Massachusetts, USA;

730 113 from Hong-Kong; 192 from the International Agency for Research on Cancer, Lyon, France;

731 13 from Mayo Clinic, Rochester, Minnesota, USA; 13 from Roswell Park Comprehensive Cancer

732 Center, Buffalo, New York, USA; 185 from Taipei; 68 from Toronto, Canada; 5 from Valencia,

733 Spain; and 29 from previously published and publicly available whole-genome sequenced (WGS)

734 lung cancers from never smokers (**Supplementary Table 13**). For these 871 individuals, the mean

735 age at lung cancer diagnosis was 64.1 years (range: 21–92) and 79.0% of patients were female.

35

736

737   Similar to our prior publication[19], we utilized seven rigorous criteria for sample inclusion: *(i)*

738   *Sequencing coverage.* We maintained a minimum average sequencing coverage of >40x for tumor

739   samples and >25x for normal samples; *(ii) Contamination and Relatedness.* Cross-sample

740   contamination was limited to <1% by Conpair[54], and detected relatedness was maintained <0.2 by

741   Somalier[55]; *(iii) Copy number analysis.* Subjects with abnormal copy number profiles in normal

742   samples were excluded, as determined by Battenberg[56]; *(iv) Mutational signatures.* Tumor samples

743   exhibiting mutational signatures SBS7 (associated with ultraviolet light exposure[41]) and SBS31

744   (associated with platinum chemotherapy[57]) were removed from the analysis. *(v) Tumor type*

745   *validation.* Tumor samples reported as non-lung cancer or not originating from primary lung

746   cancer were excluded. *(vi) WGS quality control.* Tumor samples with a total genomic alteration

747   count of <100 or <1000 along with NRPCC (the number of reads per clonal copy)[58] <10 were

748   excluded as low-quality samples. *(vii) Multiple-region sequencing.* In rare cases where multiple

749   regions of a tumor were sequenced, only one high-purity tumor sample was included to avoid

750   redundancy. These stringent criteria were applied consistently to ensure the robustness and

751   reliability of the data collected for the Sherlock-*Lung* study.

752

753   All 871 matched tumor and germline samples underwent DNA WGS. Except for the 29 previously

754   generated WGS cancers[18,59-62], sequencing was performed the same for all other samples.

755   Specifically, frozen tumor tissue with matched blood or normal tissue samples were immediately

756   put into 1ml of 0.2 mg/ml Proteinase K (Qiagen) in DNA lysis buffer (10 mMTris-Cl (pH 8.0),

757   0.1 M EDTA (pH 8.0), and 0.5% (w/v) SDS) for 24 hours at 56°C with shaking at 850 rpm in

758   Thermomixer R (Eppendorf) until the tissue was completely lysed. Genomic DNA was extracted

759   from fresh frozen tissue using the QIAmp DNA Mini Kit (Qiagen) according to the manufacturer's

760   instructions. Each sample was eluted in 200 μl AE buffer and DNA concentration was determined

761   by Nanodrop spectrophotometer. All DNA samples were aliquoted and stored at −80°C until use.

762

763   DNA was quantified using the QuantiFluor® dsDNA System (Promega Corporation, USA). DNA

764   was normalized to 25ng/ul and underwent fragment analysis via AmpFLSTR™ Identifiler™ PCR

765   Amplification Kit (ThermoFisher Scientific, USA). DNA samples are required to meet minimum

766   mass and concentration thresholds for each assay, as well as show no evidence of contamination

767   or profile discordance in the Identifiler assay. Samples meeting these requirements were aliquoted

768   at the appropriate mass needed for downstream assay processing.

769

770   **Whole-genome sequencing (WGS)**

771   The resulting post-capture enriched multiplexed sequencing libraries were used in cluster

772   formation on an Illumina cBOT (Illumina, San Diego, CA, USA). Paired-end sequencing was

773   performed by the Broad Institute (https://www.broadinstitute.org) using the Illumina HiSeq X

774   system following Illumina-provided protocols for 2x151bp paired-end sequencing. FASTQ files

775   were generated after Illumina base-calling. Next, paired FASTQ were converted to unmapped

776   BAM using the GATK pipeline (https://github.com/gatk-workflows/seq-format-conversion). The

777   unmapped BAM files were then processed using GATK on the cloud-based platform TERRA

778   workspaces (https://app.terra.bio). The sequence data were aligned to the human reference genome

779   GRCh38, and the aligned BAM files were transferred to the NIH HPC system (https://hpc.nih.gov)

780   for downstream analyses.

781

782   For the public WGS data, the preprocessed aligned BAM/CRAM files (including unmapped reads)

783   were first converted back to FASTQ files using Bazam (v.1.0.1)[63] to retain the sequencing lane

784   and read group information and then processed using the same pipeline as for the Sherlock-*Lung*

785   WGS dataset.

786

787   **Somatic variant calling**

788   The somatic variant calling was performed using our established bioinformatics pipeline as

789   previously described[19]. The analysis-ready BAM files were processed using four different

790   mutation calling algorithms for tumor-normal paired analysis, including MuTect[64], MuTect2,

791   Strelka v.2.9.10[65] and TNscope[66], implemented in the Sentieon's genomics software (v202010.01).

792   We employed an ensemble method to merge the results from these different callers followed by

793   additional filtering to reduce false positive calling. The final mutation calls for both single base

794   substitutions (SBSs) and small insertions and deletions (indels) were required to meet the

795   following criteria: *(i)* read depth >12 in tumor samples and >6 in normal samples; *(ii)* variant allele

796   frequency <0.02 in the matched-normal sample; and *(iii)* overall allele frequency (AF) <0.001 in

797   multiple genetics databases including 1000 Genomes (phase 3 v5), gnomAD exomes (v2.1.1), and

798   gnomAD genomes (v3.0)[67]. The filtered variants were annotated with Oncotator v.1.9.1.0[68] and

799   ANNOVAR v.2019-10-2495. For the indel calling, only variants called by at least three algorithms

800   were kept (MuTect2, Strelka, and TNscope). The UPS-indel[69] algorithm was used to compare and

801   combine different indel call sets. Similar filtering steps as those used for SNV calling were also

802   applied to indel calling. The final set of indels were left-normalized (left-aligned and trimmed) for

803   the downstream analysis.

804

805 **Ancestry estimation**

806 To confirm the genetic ancestry of the patients, we calculated the principal component (PC)

807 coordinates based on WGS data. This analysis was performed using the VerifyBamID (v.2.0.1)

808 algorithm[70] in conjunction with samples from the 1000 Genomes Project[21].

809

810 **Driver gene discovery**

811 The IntOGen pipeline v2020.02.0123[71], which combines seven state-of-the-art computational

812 methods, was employed to detect signals of positive selection in the mutational pattern of driver

813 genes across the cohort. The 65 genes identified as drivers with combination q-value<0.1 in the

814 cohort were classified according to their mode of action in tumorigenesis (*i.e.*, tumor suppressor

815 genes or oncogenes) based on the relationship between the excess of observed nonsynonymous

816 and truncating mutations computed by dNdScv[72] and their annotations in the Cancer Gene

817 Census[73]. Genes with conflicting computed and annotated modes of action were labeled

818 ambiguous. To identify potential driver mutations across the 65 cancer driver genes annotated in

819 the Cancer Gene Census, we selected mutations that fulfilled any of the following criteria: *(i)*

820 truncating mutations in genes annotated as tumor suppressors; *(ii)* recurrent missense mutations

821 (seen in at least three independent tumors); *(iii)* mutations classified as "Likely Drivers" by

822 boostDM (score >0.5)[74]; *(iv)* mutations classified as "Oncogenic" or "Likely Oncogenic" by

823 OncoKB[75]; *(v)* mutations classified as drivers in the TCGA MC3 drivers study[76]; *(vi)* missense

824 mutations classified as "Likely Pathogenic" by AlphaMissense[77] in genes annotated as tumor

825 suppressors.

826

827 **Estimation of tumor purity, ploidy, and allele-specific copy numbers**

39

828   We used the Battenberg algorithm (v2.2.9)[56] to conduct analyses of somatic copy number

829   alterations (SCNA). Initial SCNA profiles were generated, followed by an assessment of the

830   clonality of each segment, purity, and ploidy. Any SCNA profile determined to have low-quality

831   after manual inspection underwent a refitting process using the Battenberg algorithm. This process

832   required new tumor purity and ploidy inputs, either estimated by ccube (v1.0)[78] or recalculated

833   from local copy number status. The Battenberg refitting procedures were iteratively executed until

834   the final SCNA profile was established and met the criteria of manual validation check. GISTIC

835   (v2.0)[79] was used to identify the recurrent copy number alterations at the gene level based on the

836   major clonal copy number for each segmentation.

837

838   **Structural variants calling**

839   Meerkat (v.0.189)[80] and Manta (v.1.6.0)[81] were applied with recommended filtering for identifying

840   structural variants (SVs), and the union set of these two callers was merged as the final SV dataset.

841

842   **Telomere length**

843   We estimated telomere length in kb using TelSeq (v.0.0.2)[82] for all 871 LCINS samples as well as

844   mouse data from a recent experimental study, including control and $PM_{2.5}$-exposed mice[48]. We

845   used seven as the threshold for the number of TTAGGG/CCCTAA repeats in a read for the read

846   to be considered telomeric. The TelSeq calculation was done individually for each read group

847   within a sample, and the total number of reads in each read group was used as weight to calculate

848   the average TL for each sample.

849

850

851 **Mutational signature analysis**

852 *Extraction of de novo mutational signatures*

853 *De novo* mutational signatures for single base substitutions (SBS), doublet base substitutions

854 (DBS), and indels (ID) were extracted using SigProfilerExtractor[22] v1.1.21 with default parameters

855 and normalization set to 10,000 mutations, in order to limit the effect of hypermutators in the

856 signature extraction process. For SBSs, *de novo* signatures were extracted using the SBS-288 and

857 SBS-1536 high-definition mutational contexts, which, beyond the common SBS-96 trinucleotide

858 context using the mutated base and the 5' and 3' adjacent nucleotides[83,84], also consider the

859 transcriptional strand bias and the pentanucleotide context (two 5' and 3' adjacent nucleotides),

860 respectively[84]. Given the high similarity obtained for both mutational contexts as well as the

861 additional separation of mutational processes obtained by the SBS-288 mutational context (11 *de*

862 *novo* signatures using SBS-288 vs. 10 *de novo* signatures using SBS-1536; average cosine

863 similarity 0.97), the results using the SBS-288 context were used for further analysis

864 (**Supplementary Table 2**). Previously established mutational contexts DBS-78 and ID-83[15,84]

865 were used for the extraction of DBS and ID signatures (**Supplementary Tables 3-4**). Copy number

866 signatures were extracted *de novo* following an updated context definition benefitting from deep

867 WGS data (CN-68) (**Supplementary Table 5**), which allowed to further characterize CN segments

868 below 100kbp in length (in contrast to current COSMICv3.4 reference signatures using the CN-48

869 context, which are based on SNP6 microarray data and therefore without the resolution to

870 characterize short CN segments)[26]. SV signatures were extracted using a similarly refined context,

871 with an in-depth characterization of short SV alterations below 1kbp (SV-38 context;

872 **Supplementary Table 6**).

873

Decomposition and assignment of mutational signatures to individual tumors

After *de novo* extraction was completed, SigProfilerAssignment[85] v0.1.1 was used to decompose the *de novo* extracted SBS, ID, DBS, CN, and SV mutational signatures into COSMICv3.4[23] reference signatures based on the GRCh38 reference genome (**Supplementary Tables 7-11**) as well as to assign signatures to individual samples obtaining signature activities, based on the forward stagewise algorithm for sparse regression and nonnegative least squares for numerical optimization. Hierarchical clustering of the activities of mutational signatures was performed using Euclidean distance and Ward's minimum-variance clustering. For the SBS signatures, the 11 *de novo* extracted signatures were originally decomposed into 15 COSMICv3.4 reference signatures. However, two of the COSMICv3.4 signatures, enriched in C>T substitutions, were removed from the decomposition to avoid misassignment, namely SBS23 and SBS32, as the patterns of strong transcriptional strand bias generated by these two signatures[15] were not observed in our LCINS dataset, and their individual contribution to the decomposition of *de novo* signatures was minimal. On the other hand, five additional COSMICv3.4 SBS signatures were included for the assignment to individual samples, including SBS3, SBS21, SBS33, SBS44, and SBS92. SBS3 and SBS92 were included considering their previously reported strong associations with indel signatures ID6 and ID3, both of which were observed in our indel signature analysis after decomposition to COSMICv3.4 ID signatures[15,22,43]. Thus, for those tumors harboring ID6 (n=41), we allowed assigning SBS3 (assigned only to 14 of the cases), whereas for those tumors with presence of ID3 (n=300), SBS92 was allowed in the assignment (assigned to 6 tumors). Suspected homologous recombination deficient (HRD) tumors exhibiting signatures SBS3 and ID6 were tested using two independent computational algorithms, CHORD[30] and HRDetect[31]. Lastly, two samples showing a low cosine similarity for the mutational profile reconstruction (NSLC-0477 and NSLC-0637)

42

897   showed high similarities in their mutational profiles with signatures linked to microsatellite

898   instability (MSI) and COSMIC SBS33, respectively. Considering this, we allowed assignment of

899   SBS33 and all MSI-associated signatures (SBS6, SBS14, SBS15, SBS20, SBS21, SBS26, and

900   SBS44), respectively, in each of this specific samples, with SBS21 and SBS44 being finally

901   assigned to sample NSLC-0477. The MSI phenotype of this sample was independently confirmed

902   by MMRDetect[25], and further evidence was collected by exploring the indel landscape, with

903   10,474 alterations, 29-fold more than the average number of indels in lung adenocarcinomas, and

904   a strong enrichment of signature ID2, consistent with previously reported MSI cases[15]. The

905   decomposed COSMICv3.4 SBS signatures obtained in the whole LCINS cohort were used for the

906   assignment of signatures to the mouse data obtained from a recent publication[48], after their

907   renormalization to the mouse genome (mm10 build).

908

909   ***Assignment of mutational signatures to individual somatic mutations***

910   Signatures   were   probabilistically   assigned   to   individual   somatic   mutations   using

911   SigProfilerAssignment based on Bayes' rule and the specific mutational context for the mutation,

912   as previously described[85]. Briefly, to calculate the probability of a specific mutational signature

913   being responsible for a mutation in a given mutational context and in a particular sample, we

914   multiplied the general probability of the signature causing mutations in a specific mutational

915   context (obtained from the mutational signature profile) by the activity of the signature in the

916   sample (obtained from the signature activities), and then normalized this value dividing by the

917   total number of mutations corresponding to the specific mutational context (obtained from the

918   reconstructed mutational profile of the sample).

919

Assessment of statistical power to assign mutational signatures in passive smokers

920 To assess the statistical power to assign SBS4 in passive smokers, we performed simulations where

922 SBS4 was injected at different average levels (1%, 2%, 5%, 10%, 15%, and 20% of total mutations

923 in each sample; 100 simulations for each injection level) in all 247 tumors from passive smoker

924 patients lacking SBS4. For each sample, prior to the injection of SBS4, the number of mutations

925 to be injected into the sample was randomly subtracted from the sample while ensuring all

926 mutation counts were still non-negative. Next, SBS4 mutations were injected at the current level

927 being tested. After, 10% Gaussian noise was added to the resulting mutational profile.

928 Subsequently, mutational signatures were re-assigned in each sample using

929 SigProfilerAssignment[85] as well as the 18 SBS COSMICv3.4 reference signatures considered for

930 the original data using default parameters except for a relaxed addition penalty

931 (nnls_add_penalty=0.01) in order to increase the sensitivity of the signature assignment analysis.

932

Evaluating the topography of mutational signatures

934 Topography analyses specific for the 56 SBS4 positive samples as well as 68 lung cancer

935 samples from smoker patients from the Pan-Cancer Analysis of Whole Genomes (PCAWG)

936 cohort[45] were carried out with SigProfilerTopography[86], which evaluates the effect of DNA

937 replication, DNA transcription, chromatin organization, histone modifications, and transcription

938 factor binding on the activities of different mutational processes. SigProfilerTopography

939 examines the distribution of topographical features and narrows down the analyses by calculating

940 the average signal of each feature in the close vicinity of the somatic mutations. Next, all the

941 results of somatic mutations are statistically compared with those from simulated mutations that

44

942   account for the patterns of all operative mutational signatures within an examined sample to

943   elucidate statistically significant differences.

944   **Estimating air pollution**

945   Annual country-level population-weighted mean concentration estimates of the environmental

946   particulate matter measuring $\leq 2.5\,\mu m$ ($PM_{2.5}$) ($\mu g/m^3$) from 1998–2021 were obtained from a

947   hybrid model combining measurements of aerosol optical depth from different satellites (MODIS,

948   VIIRS, MISR, and SeaWiFS) with chemical transport modeling (GEOS-Chem) and ground-based

949   photometer (AERONET) observations[49], and downloaded from

950   https://sites.wustl.edu/acag/datasets/surface-pm2-5/. Additional state-level and provincial-level

951   yearly mean estimates were also used for the patients from the United States and Canada,

952   respectively. An individual estimate of the $PM_{2.5}$ exposure for each patient was calculated by

953   averaging the annual mean concentration values from the year of lung cancer diagnosis (calculated

954   by adding the age of diagnosis to the year of birth) until the earliest year with data available, i.e.,

955   1998. For two patients diagnosed before 1998, as well as for 28 patients whose age of diagnosis

956   or year of birth was unknown, the $PM_{2.5}$ estimates for 1998 were considered. $PM_{2.5}$ estimates were

957   not available for 18 cases for which the country of residence was unknown. For the dichotomized

958   analysis between high and low polluted regions according to the levels of individual estimates of

959   the $PM_{2.5}$ exposure, a threshold of 20 $\mu g/m^3$ was used, considering the distribution of $PM_{2.5}$

960   estimates across the whole cohort (**Supplementary Fig. 10**).

961

962   **Association of geographical regions, histologies, *EGFR* driver mutation status, air**

963   **pollution, and passive smoking with genomic features**

964   In order to quantify the influence of passive smoking and air pollution on the number of total

965   somatic mutations and the ratio of telomere length between tumor and normal samples, we fitted

966   a series of multivariable linear models, including adjustments for several confounding covariates,

967   namely age, sex, genetic ancestry, histology, and tumor purity. The total number of mutations

968   ($\log_{10}$ scale) or telomere length ratio ($\log_2$ scale) served as dependent variables for the linear

969   regressions, respectively, whereas passive smoking/air pollution groups and covariates were used

970   as independent variables. In addition, for assessing the dose-response effect of air pollution, we

971   fitted multivariable linear regressions using the individual $PM_{2.5}$ estimates per sample, considering

972   similar adjustments as in the analysis by pollution groups. Genetic ancestry was primarily

973   considered for the adjustments based on the super-samples from the 1,000 Genomes Project[21],

974   classifying the samples into EUR, EAS, or Other genetic ancestry (**Fig. 1b**). Additionally, we also

975   considered the genetic ancestry principal components derived from the WGS data for the

976   adjustments, with no differences in the associations compared to the ancestry label-based

977   adjustments (**Supplementary Table 12**)

978

979   Similarly, the presence of specific mutational signatures or mutations in specific driver genes was

980   modeled based on the influence of geographical region/histology/*EGFR* mutation status/passive

981   smoking/air pollution and the additional covariates previously considered, including age, sex,

982   genetic ancestry, tumor purity, and histology (if applicable), by using multivariable logistic

983   regressions. For this purpose, signature activities were dichotomized into the presence or absence

984   of a particular signature, and these binary activities served as dependent variables for the logistic

985   regressions, with geographical regions, *EGFR* mutation status, histology subtypes, passive

986   smoking groups, or individual $PM_{2.5}$ estimates, respectively, along with covariates, being used as

46

987   independent variables. For signatures present in more than 50% of the cases, dichotomization was

988   done above and below the median of assigned mutations to the overall cohort of LCINS patients

989   (*n*=871). Signatures SBS21, SBS33, and SBS44 were excluded from the analysis as they were only

990   found in individual tumors. For the enrichment analysis of mutations in driver genes, the presence

991   or absence of mutations in specific genes was used as the dependent variable for the logistic

992   regressions. Only driver genes having driver mutations in more than 2% of cases of the assessed

993   cohort were considered.

994

995   In addition, to assess the dose-response effect of air pollution, we fitted multivariable linear

996   regressions using the individual $PM_{2.5}$ estimates per sample, considering similar corrections for

997   covariates and the tumor mutational burden ($\log_{10}$ scale), the telomere length ratio ($\log_2$ scale), or

998   the number of mutations contributed by a given signature ($\log_{10}$ scale) as the dependent variable.

999   Only samples where a particular mutational signature was present were considered for the

1000   multivariable linear regressions. Univariate linear regressions for the average tumor mutational

1001   burden ($\log_{10}$ scale), the telomere length ratio ($\log_2$ scale), or the number of mutations contributed

1002   by a given signature ($\log_{10}$ scale) vs. the average $PM_{2.5}$ per geographical region were used for

1003   visualization.

1004

1005   In all cases, p-values were corrected according to the different signatures from the same variant

1006   type or driver genes considered by using a false-discovery rate correction based on the Benjamini-

1007   Hochberg method[87] and reported as FDR. FDR<0.05 were considered statistically significant.

1008

1009   **Statistical analysis**

1010   All statistical analyses and graphic displays were performed using the R software v4.2.3

1011   (https://www.r-project.org/). Two-sided Fisher's exact tests were used for the enrichment analyses

1012   of categorical variables. For the comparison of numerical variables across groups, we used non-

1013   parametric Mann-Whitney (Wilcoxon rank sum) tests. P-values<0.05 were considered statistically

1014   significant. If multiple hypothesis testing was required, we used a false-discovery rate correction

1015   based on the Benjamini-Hochberg method[87] and reported FDR. FDR<0.05 were considered

1016   statistically significant.

1017

1018   **DATA AVAILABILITY**

1019   Normal and tumor-paired CRAM files for the 871 WGS subjects of the Sherlock-*Lung* study have

1020   been deposited in dbGaP under the accession numbers phs001697.v1.p1. Detailed access

1021   information for the publicly available datasets can be found in **Supplementary Table 13**.

1022

1023   **CODE AVAILABILITY**

1024   The WGS bioinformatics pipelines can be accessed at https://github.com/xtmgah/Sherlock-Lung.

1025   Battenberg SCNA calling algorithm can be found at https://github.com/Wedge-lab/battenberg.

1026

1040

COMPETING INTERESTS

1042 LBA is a co-founder, CSO, scientific advisory member, and consultant for io9, has equity and

1043 receives income. The terms of this arrangement have been reviewed and approved by the

1044 University of California, San Diego in accordance with its conflict of interest policies. LBA is also

1045 a compensated member of the scientific advisory board of Inocras. LBA's spouse is an employee

1046 of Biotheranostics. ENB and LBA declare U.S. provisional patent application filed with UCSD

1047 with serial numbers 63/269,033. LBA also declares U.S. provisional applications filed with UCSD

1048 with serial numbers: 63/366,392; 63/289,601; 63/483,237; 63/412,835; and 63/492,348. LBA is

1049 also an inventor of a US Patent 10,776,718 for source identification by non-negative matrix

49

1050   factorization. SRY has received consulting fees from AstraZeneca, Sanofi, Amgen, AbbVie, and

1051   Sanofi; received speaking fees from AstraZeneca, Medscape, PRIME Education, and Medical

1052   Learning Institute. All other authors declare that they have no competing interests.

REFERENCES

1   Proctor, R. N. Tobacco and the global lung cancer epidemic. *Nature Reviews Cancer* **1**, 82-86 (2001). https://doi.org/10.1038/35094091

2   Sun, S., Schiller, J. H. & Gazdar, A. F. Lung cancer in never smokers — a different disease. *Nature Reviews Cancer* **7**, 778-790 (2007). https://doi.org/10.1038/nrc2190

3   Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin* **71**, 209-249 (2021). https://doi.org/10.3322/caac.21660

4   Siegel, D. A., Fedewa, S. A., Henley, S. J., Pollack, L. A. & Jemal, A. Proportion of Never Smokers Among Men and Women With Lung Cancer in 7 US States. *JAMA Oncol* **7**, 302-304 (2021). https://doi.org/10.1001/jamaoncol.2020.6362

5   Lui, N. S. *et al.* Sub-solid lung adenocarcinoma in Asian versus Caucasian patients: different biology but similar outcomes. *J Thorac Dis* **12**, 2161-2171 (2020). https://doi.org/10.21037/jtd.2020.04.37

6   Gaughan, E. M., Cryer, S. K., Yeap, B. Y., Jackman, D. M. & Costa, D. B. Family history of lung cancer in never smokers with non-small-cell lung cancer and its association with tumors harboring EGFR mutations. *Lung Cancer* **79**, 193-197 (2013). https://doi.org/10.1016/j.lungcan.2012.12.002

7   Toh, C. K. *et al.* Never-smokers with lung cancer: epidemiologic evidence of a distinct disease entity. *J Clin Oncol* **24**, 2245-2251 (2006). https://doi.org/10.1200/JCO.2005.04.8033

8   Yano, T. *et al.* Never-smoking nonsmall cell lung cancer as a separate entity: clinicopathologic features and survival. *Cancer* **113**, 1012-1018 (2008). https://doi.org/10.1002/cncr.23679

9   Brennan, P. *et al.* High cumulative risk of lung cancer death among smokers and nonsmokers in Central and Eastern Europe. *Am J Epidemiol* **164**, 1233-1241 (2006). https://doi.org/10.1093/aje/kwj340

10   Wang, P., Sun, S., Lam, S. & Lockwood, W. W. New insights into the biology and development of lung cancer in never smokers-implications for early detection and treatment. *J Transl Med* **21**, 585 (2023). https://doi.org/10.1186/s12967-023-04430-x

11   Tobacco Smoke and Involuntary Smoking. *IARC Monographs on the Evaluation of Carcinogenic Risks to Humans*. Vol. 83 1-1438 (2004).

12   Turner, M. C. *et al.* Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. *CA Cancer J Clin* **70**, 460-479 (2020). https://doi.org/10.3322/caac.21632

13   Ciabattini, M., Rizzello, E., Lucaroni, F., Palombi, L. & Boffetta, P. Systematic review and meta-analysis of recent high-quality studies on exposure to particulate matter and risk of lung cancer. *Environ Res* **196**, 110440 (2021). https://doi.org/10.1016/j.envres.2020.110440

14   Koh, G., Degasperi, A., Zou, X., Momen, S. & Nik-Zainal, S. Mutational signatures: emerging concepts, caveats and clinical applications. *Nat Rev Cancer* **21**, 619-637 (2021). https://doi.org/10.1038/s41568-021-00377-7

15   Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020). https://doi.org/10.1038/s41586-020-1943-3

16   ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium, T. Pan-cancer analysis of whole genomes. *Nature* **578**, 82-93 (2020). https://doi.org/10.1038/s41586-020-1969-6

17   Wang, X. *et al.* Association between Smoking History and Tumor Mutation Burden in Advanced Non-Small Cell Lung Cancer. *Cancer Res* **81**, 2566-2573 (2021). https://doi.org/10.1158/0008-5472.CAN-20-3991

18   Lee, J. J. *et al.* Tracing Oncogene Rearrangements in the Mutational History of Lung Adenocarcinoma. *Cell* **177**, 1842-1857 e1821 (2019). https://doi.org/10.1016/j.cell.2019.05.013

19   Zhang, T. *et al.* Genomic and evolutionary classification of lung cancer in never smokers. *Nature Genetics* **53**, 1348-1359 (2021). https://doi.org/10.1038/s41588-021-00920-0

20   Landi, M. T. *et al.* Tracing Lung Cancer Risk Factors Through Mutational Signatures in Never-Smokers : The Sherlock-Lung Study. *American Journal of Epidemiology* **190**, 962-976 (2020). https://doi.org/10.1093/aje/kwaa234

21   Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68-74 (2015). https://doi.org/10.1038/nature15393

22   Islam, S. M. A. *et al.* Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genom* **2**, 100179 (2022). https://doi.org/10.1016/j.xgen.2022.100179

23   Sondka, Z. *et al.* COSMIC: a curated database of somatic variants and clinical data for cancer. *Nucleic Acids Res* **52**, D1210-D1217 (2024). https://doi.org/10.1093/nar/gkad986

24   Senkin, S. *et al.* Geographic variation of mutagenic exposures in kidney cancer genomes. *Nature* (2024). https://doi.org/10.1038/s41586-024-07368-2

25   Zou, X. *et al.* A systematic CRISPR screen defines mutational mechanisms underpinning signatures caused by replication errors and endogenous DNA damage. *Nature Cancer* **2**, 643-657 (2021). https://doi.org/10.1038/s43018-021-00200-0

26   Steele, C. D. *et al.* Signatures of copy number alterations in human cancer. *Nature* **606**, 984-991 (2022). https://doi.org/10.1038/s41586-022-04738-6

27   Everall, A. *et al.* Comprehensive repertoire of the chromosomal alteration and mutational signatures across 16 cancer types from 10,983 cancer patients. *medRxiv*, 2023.2006.2007.23290970 (2023). https://doi.org/10.1101/2023.06.07.23290970

28   Nik-Zainal, S. *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47-54 (2016). https://doi.org/10.1038/nature17676

29   Degasperi, A. *et al.* A practical framework and online tool for mutational signature analyses show inter-tissue variation and driver dependencies. *Nat Cancer* **1**, 249-263 (2020). https://doi.org/10.1038/s43018-020-0027-5

30   Nguyen, L., W. M. Martens, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. *Nature Communications* **11**, 5584 (2020). https://doi.org/10.1038/s41467-020-19406-4

31   Davies, H. *et al.* HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat Med* **23**, 517-525 (2017). https://doi.org/10.1038/nm.4292

32   Poon, S. L. *et al.* Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci Transl Med* **5**, 197ra101 (2013). https://doi.org/10.1126/scitranslmed.3006086

33   Hoang, M. L. *et al.* Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci Transl Med* **5**, 197ra102 (2013). https://doi.org/10.1126/scitranslmed.3006200

34   Letouze, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat Commun* **8**, 1315 (2017). https://doi.org/10.1038/s41467-017-01358-x

35   Chen, Y.-J. *et al.* Proteogenomics of Non-smoking Lung Cancer in East Asia Delineates Molecular Signatures of Pathogenesis and Progression. *Cell* **182**, 226-244.e217 (2020). https://doi.org/10.1016/j.cell.2020.06.012

36   Zhang, T. *et al.* APOBEC shapes tumor evolution and age at onset of lung cancer in smokers. *bioRxiv*, 2024.2004.2002.587805 (2024). https://doi.org/10.1101/2024.04.02.587805

37   Morton, L. M. *et al.* Radiation-related genomic profile of papillary thyroid carcinoma after the Chernobyl accident. *Science* **372**, eabg2538 (2021). https://doi.org/10.1126/science.abg2538

38   Jager, M. *et al.* Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. *Genome Res* **29**, 1067-1077 (2019). https://doi.org/10.1101/gr.246223.118

39   Singh, V. K., Rastogi, A., Hu, X., Wang, Y. & De, S. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. *Commun Biol* **3**, 421 (2020). https://doi.org/10.1038/s42003-020-01119-5

40   Caplin, M. E. *et al.* Pulmonary neuroendocrine (carcinoid) tumors: European Neuroendocrine Tumor Society expert consensus and recommendations for best practice for typical and atypical pulmonary carcinoids. *Ann Oncol* **26**, 1604-1620 (2015). https://doi.org/10.1093/annonc/mdv041

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

42   Lindeman, N. I. *et al.* Molecular testing guideline for selection of lung cancer patients for EGFR and ALK tyrosine kinase inhibitors: guideline from the College of American Pathologists, International Association for the Study of Lung Cancer, and Association for Molecular Pathology. *J Thorac Oncol* **8**, 823-859 (2013). https://doi.org/10.1097/JTO.0b013e318290868f

43   Lawson, A. R. J. *et al.* Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75-82 (2020). https://doi.org/10.1126/science.aba8347

44   Degasperi, A. *et al.* Substitution mutational signatures in whole-genome-sequenced cancers in the UK population. *Science* **376**, abl9283 (2022). https://doi.org/10.1126/science.abl9283

45   Otlu, B. *et al.* Topography of mutational signatures in human cancer. *Cell Rep* **42**, 112930 (2023). https://doi.org/10.1016/j.celrep.2023.112930

46   Jamal-Hanjani, M. *et al.* Tracking the Evolution of Non–Small-Cell Lung Cancer. *New England Journal of Medicine* **376**, 2109-2121 (2017). https://doi.org/10.1056/NEJMoa1616288

47   Zhang, T. *et al.* Distinct Genomic Landscape of Lung Adenocarcinoma from Household Use of Smoky Coal. *American Journal of Respiratory and Critical Care Medicine* **208**, 733-736 (2023). https://doi.org/10.1164/rccm.202302-0340LE

48   Hill, W. *et al.* Lung adenocarcinoma promotion by air pollutants. *Nature* **616**, 159-167 (2023). https://doi.org/10.1038/s41586-023-05874-3

49   van Donkelaar, A. *et al.* Monthly Global Estimates of Fine Particulate Matter and Their Uncertainty. *Environmental Science & Technology* **55**, 15287-15300 (2021). https://doi.org/10.1021/acs.est.1c05309

50   Mochizuki, A. *et al.* Passive smoking-induced mutagenesis as a promoter of lung carcinogenesis. *J Thorac Oncol*, S1556-0864(1524)00074-00071 (2024). https://doi.org/10.1016/j.jtho.2024.02.006

51   Yu, X. J. *et al.* Characterization of Somatic Mutations in Air Pollution-Related Lung Cancer. *EBioMedicine* **2**, 583-590 (2015). https://doi.org/10.1016/j.ebiom.2015.04.003

52   Chan, W.-H. *et al.* Verifying the accuracy of self-reported smoking behavior in female volunteer soldiers. *Scientific Reports* **13**, 3438 (2023). https://doi.org/10.1038/s41598-023-29699-2

53   Landi, M. T. *et al.* Environment And Genetics in Lung cancer Etiology (EAGLE) study: an integrative population-based case-control study of lung cancer. *BMC Public Health* **8**, 203 (2008). https://doi.org/10.1186/1471-2458-8-203

54   Bergmann, E. A., Chen, B. J., Arora, K., Vacic, V. & Zody, M. C. Conpair: concordance and contamination estimator for matched tumor-normal pairs. *Bioinformatics* **32**, 3196-3198 (2016). https://doi.org/10.1093/bioinformatics/btw389

55   Pedersen, B. S. *et al.* Somalier: rapid relatedness estimation for cancer and germline studies using efficient genome sketches. *Genome Medicine* **12**, 62 (2020). https://doi.org/10.1186/s13073-020-00761-2

56   Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012). https://doi.org/10.1016/j.cell.2012.04.023

57   Boot, A. *et al.* In-depth characterization of the cisplatin mutational signature in human cell lines and in esophageal and liver tumors. *Genome Res* **28**, 654-665 (2018). https://doi.org/10.1101/gr.230219.117

58   Dentro, S. C. *et al.* Characterizing genetic intra-tumor heterogeneity across 2,658 human cancer genomes. *Cell* **184**, 2239-2254 (2021). https://doi.org/10.1016/j.cell.2021.03.009

59   Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-1120 (2012). https://doi.org/10.1016/j.cell.2012.08.029

60   Lee, J. K. *et al.* Clonal History and Genetic Predictors of Transformation Into Small-Cell Carcinomas From Lung Adenocarcinomas. *J Clin Oncol* **35**, 3065-3074 (2017). https://doi.org/10.1200/JCO.2016.71.9096

61   Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550 (2014). https://doi.org/10.1038/nature13385

62   Carrot-Zhang, J. *et al.* Whole-genome characterization of lung adenocarcinomas lacking the RTK/RAS/RAF pathway. *Cell Rep* **34**, 108707 (2021). https://doi.org/10.1016/j.celrep.2021.108707

63   Sadedin, S. P. & Oshlack, A. Bazam: a rapid method for read extraction and realignment of high-throughput sequencing data. *Genome Biol* **20**, 78 (2019). https://doi.org/10.1186/s13059-019-1688-1

64   Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-219 (2013). https://doi.org/10.1038/nbt.2514

1232 65 Kim, S. *et al.* Strelka2: fast and accurate calling of germline and somatic variants. *Nat*
1233   *Methods* **15**, 591-594 (2018). https://doi.org/10.1038/s41592-018-0051-x
1234 66 Freed, D., Pan, R. & Aldana, R. TNscope: Accurate Detection of Somatic Mutations with
1235   Haplotype-based Variant Candidate Detection and Machine Learning Filtering. *bioRxiv*,
1236   250647 (2018). https://doi.org/10.1101/250647
1237 67 Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in
1238   141,456 humans. *Nature* **581**, 434-443 (2020). https://doi.org/10.1038/s41586-020-2308-
1239   7
1240 68 Ramos, A. H. *et al.* Oncotator: cancer variant annotation tool. *Hum Mutat* **36**, E2423-
1241   2429 (2015). https://doi.org/10.1002/humu.22771
1242 69 Hasan, M. S., Wu, X., Watson, L. T. & Zhang, L. UPS-indel: a Universal Positioning
1243   System for Indels. *Sci Rep* **7**, 14106 (2017). https://doi.org/10.1038/s41598-017-14400-1
1244 70 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in
1245   sequencing and array-based genotype data. *Am J Hum Genet* **91**, 839-848 (2012).
1246   https://doi.org/10.1016/j.ajhg.2012.09.004
1247 71 Martinez-Jimenez, F. *et al.* A compendium of mutational cancer driver genes. *Nat Rev*
1248   *Cancer* **20**, 555-572 (2020). https://doi.org/10.1038/s41568-020-0290-x
1249 72 Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues.
1250   *Cell* **171**, 1029-1041 (2017). https://doi.org/10.1016/j.cell.2017.09.042
1251 73 Sondka, Z. *et al.* The COSMIC Cancer Gene Census: describing genetic dysfunction
1252   across all human cancers. *Nat Rev Cancer* **18**, 696-705 (2018).
1253   https://doi.org/10.1038/s41568-018-0060-1
1254 74 Muiños, F., Martinez-Jimenez, F., Pich, O., Gonzalez-Perez, A. & Lopez-Bigas, N. In
1255   silico saturation mutagenesis of cancer genes. *Nature* **596**, 428-432 (2021).
1256   https://doi.org/10.1038/s41586-021-03771-1
1257 75 Chakravarty, D. *et al.* OncoKB: A Precision Oncology Knowledge Base. *JCO Precis*
1258   *Oncol* **2017**, 1-16 (2017). https://doi.org/10.1200/PO.17.00011
1259 76 Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and
1260   Mutations. *Cell* **173**, 371-385 e318 (2018). https://doi.org/10.1016/j.cell.2018.02.060
1261 77 Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with
1262   AlphaMissense. *Science* **381**, eadg7492 (2023). https://doi.org/10.1126/science.adg7492
1263 78 Yuan, K., Macintyre, G., Liu, W. & Markowetz, F. Ccube: A fast and robust method for
1264   estimating cancer cell fractions. *bioRxiv*, 484402 (2018). https://doi.org/10.1101/484402
1265 79 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the
1266   targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41
1267   (2011). https://doi.org/10.1186/gb-2011-12-4-r41
1268 80 Yang, L. *et al.* Diverse mechanisms of somatic structural variations in human cancer
1269   genomes. *Cell* **153**, 919-929 (2013). https://doi.org/10.1016/j.cell.2013.04.010
1270 81 Chen, X. *et al.* Manta: rapid detection of structural variants and indels for germline and
1271   cancer sequencing applications. *Bioinformatics* **32**, 1220-1222 (2016).
1272   https://doi.org/10.1093/bioinformatics/btv710
1273 82 Ding, Z. *et al.* Estimating telomere length from whole genome sequence data. *Nucleic*
1274   *Acids Res* **42**, e75 (2014). https://doi.org/10.1093/nar/gku181
1275 83 Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R.
1276   Deciphering signatures of mutational processes operative in human cancer. *Cell Rep* **3**,
1277   246-259 (2013). https://doi.org/10.1016/j.celrep.2012.12.008

1278 84   Bergstrom, E. N. *et al.* SigProfilerMatrixGenerator: a tool for visualizing and exploring
1279   patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
1280   https://doi.org/10.1186/s12864-019-6041-2
1281 85   Diaz-Gay, M. *et al.* Assigning mutational signatures to individual samples and individual
1282   somatic mutations with SigProfilerAssignment. *Bioinformatics* **39**, btad756 (2023).
1283   https://doi.org/10.1093/bioinformatics/btad756
1284 86   Otlu, B. & Alexandrov, L. B. Evaluating topography of mutational signatures with
1285   SigProfilerTopography. *bioRxiv*, 2024.2001.2008.574683 (2024).
1286   https://doi.org/10.1101/2024.01.08.574683
1287 87   Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and
1288   Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-*
1289   *Statistical Methodology* **57**, 289-300 (1995). https://doi.org/10.1111/j.2517-
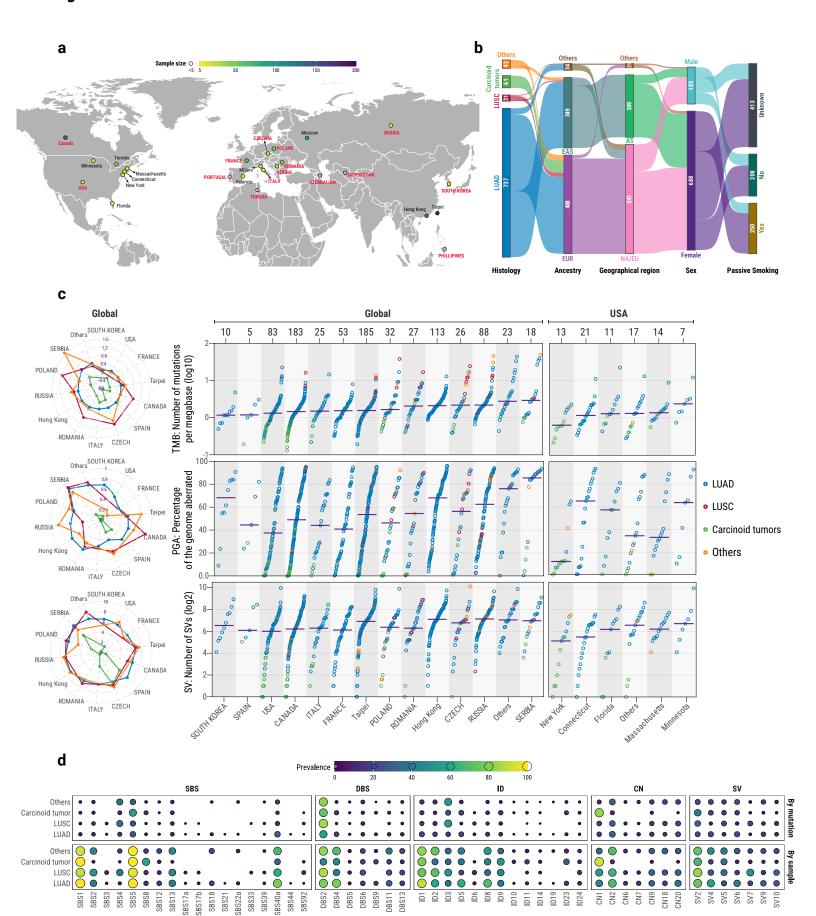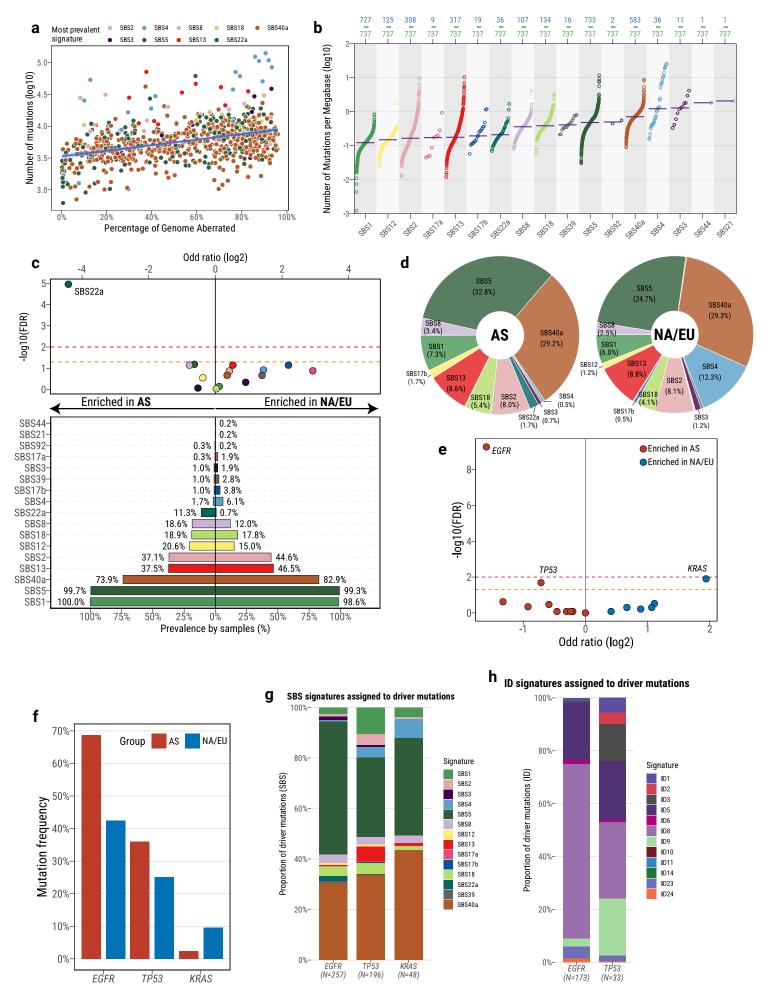1290   6161.1995.tb02031.x
1291

Fig. 1

Fig. 2

Fig. 3

Fig. 4

Fig. 5