

Multi-ancestry meta-analyses of lung cancer in the Million Veteran Program reveal novel risk loci and elucidate smoking-independent genetic risk

Bryan R. Gorman^{1,2}, Sun-Gou Ji^{1,15}, Michael Francis^{1,2}, Anoop K. Sendamarai^{1,16}, Yunling Shi¹, Poornima Devineni¹, Uma Saxena¹, Elizabeth Partan¹, Andrea K. DeVito^{1,2}, Jinyoung Byun^{11,12}, Younghun Han^{11,12}, Xiangjun Xiao^{11,12}, Don D. Sin³, Wim Timens^{4,5}, Jennifer Moser⁶, Sumitra Muralidhar⁶, Rachel Ramoni⁶, Rayjean J. Hung⁷, James D. McKay⁸, Yohan Bossé⁹, Ryan Sun¹⁰, Christopher I. Amos^{11,12,13}, VA Million Veteran Program, Saiju Pyarajan^{1,14,‡}

¹Center for Data and Computational Sciences (C-DACS), VA Boston Healthcare System, Boston, MA, USA, ²Booz Allen Hamilton, McLean, VA, USA, ³The University of British Columbia Centre for Heart Lung Innovation, St Paul's Hospital, Vancouver, British Columbia, Canada, ⁴University Medical Centre Groningen, GRIAC (Groningen Research Institute for Asthma and COPD), University of Groningen, Groningen, Netherlands, ⁵Department of Pathology & Medical Biology, University Medical Centre Groningen, University of Groningen, Groningen, Netherlands, ⁶Office of Research and Development, Department of Veterans Affairs, Washington, DC, USA. ⁷Lunenfeld-Tanenbaum Research Institute, Sinai Health System, University of Toronto, Toronto, Ontario, Canada, ⁸Section of Genetics, International Agency for Research on Cancer, World Health Organization, Lyon, France, ⁹Institut universitaire de cardiologie et de pneumologie de Québec, Department of Molecular Medicine, Laval University, Quebec City, Quebec, Canada, ¹⁰Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA, ¹¹Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA, ¹²Department of Medicine, Section of Epidemiology and Population Sciences, Baylor College of Medicine, Houston, TX, USA, ¹³Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA, ¹⁴Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA, ¹⁵Present address: Sun-Gou Ji, BridgeBio Pharma, Palo Alto, CA, USA, ¹⁶Present address: Carbone Cancer Center, University of Wisconsin, Madison, WI, USA. [‡]Saiju Pyarajan: saiju.pyarajan@va.gov

34 **Abstract**

35 Lung cancer remains the leading cause of cancer mortality, despite declines in smoking
36 rates. Previous lung cancer genome-wide association studies (GWAS) have identified
37 numerous loci, but separating the genetic risks of lung cancer and smoking behavioral
38 susceptibility remains challenging. We performed multi-ancestry GWAS meta-analyses
39 of lung cancer using the Million Veteran Program (MVP) cohort and a previous study of
40 European-ancestry individuals, comprising 42,102 cases and 181,270 controls, followed
41 by replication in an independent cohort of 19,404 cases and 17,378 controls. We further
42 performed conditional meta-analyses on cigarettes per day and identified two novel,
43 replicated loci, including the 19p13.11 pleiotropic cancer locus in LUSC. Overall, we
44 report twelve novel risk loci for overall lung cancer, lung adenocarcinoma (LUAD), and
45 squamous cell lung carcinoma (LUSC), nine of which were externally replicated. Finally,
46 we performed phenome-wide association studies (PheWAS) on polygenic risk scores
47 (PRS) for lung cancer, with and without conditioning on smoking. The unconditioned
48 lung cancer PRS was associated with smoking status in controls, illustrating reduced
49 predictive utility in non-smokers. Additionally, our PRS demonstrates smoking-
50 independent pleiotropy of lung cancer risk across neoplasms and metabolic traits.

51

52 Introduction

53 Lung cancer remains the leading cause of overall cancer mortality, as the most
54 prevalent cancer type in men, and the second highest in women after breast cancer¹⁻³.
55 Despite declines in smoking rates in the US since the 1980s⁴, tobacco use is currently
56 implicated in upwards of 80% of lung cancer diagnoses¹. Even in those who have never
57 smoked, nor had meaningful exposure to environmental carcinogens^{1,5}, there exists a
58 heritable risk component of lung cancer conferred by genetic factors⁶⁻⁸. Differentiating
59 the mutations which directly predispose an individual to lung cancer from those whose
60 effect is mediated through environmental components remains challenging.

61 Genome-wide association studies (GWAS) have identified lung cancer risk
62 variants associated with oncogenic processes such as immune response⁷, cell cycle
63 regulation⁹, and those affecting DNA damage response and genomic stability⁸. Several
64 lung cancer GWAS have also reported strong effects of genes such as *CHRNA* nicotine
65 receptor genes which putatively increase the risk of lung cancer through behavioral
66 predisposition towards smoking⁵. Characteristic molecular markers and genetic risk
67 factors in smokers and never-smokers have been identified^{10,11}, though fewer variants
68 have been found in GWAS performed exclusively in never-smokers¹².

69 Lung cancer has a heterogeneous genetic architecture across ancestral
70 groups^{13,14}. In the two most well-studied ancestries, European (EA) and East Asian
71 (EAS), the majority of genome-wide significant loci are not shared^{15,16}; this is in
72 agreement with molecular studies showing differences in tumor characteristics between
73 EA and EAS¹⁷. Smaller African ancestry (AA) cohorts have replicated known loci from
74 EA or EAS^{8,18}, though no AA-specific GWAS loci have been reported.

75 In this study, we examined lung cancer genetic variation in EA as well as in the
76 largest AA cohort to-date. Our discovery analysis is performed in an older cohort of
77 mostly male US veterans in the Department of Veterans Affairs Million Veteran Program
78 (MVP)¹⁹. Lung cancer incidence is approximately twice as high in men than in women²,
79 and additionally MVP contains a large number of cigarette smokers, positioning this
80 biobank as particularly valuable for this analysis. We performed GWAS in overall cases
81 of lung cancer as well as two non-small cell lung cancer (NSCLC) subtypes,
82 adenocarcinoma (LUAD) and squamous cell lung carcinoma (LUSC).

83

84 **Results**

85 *Genome-wide association studies for lung cancer*

86 We performed a GWAS on overall lung cancer within EA participants in MVP
87 (10,398 lung cancer cases and 62,708 controls; Supplementary Data 1), followed by a
88 meta-analysis with the EA International Lung Cancer Consortium OncoArray study
89 (ILCCO; McKay et al., 2017)⁷, for a total of 39,781 cases and 119,158 controls
90 (Supplementary Fig. 1). The EA meta-analysis for overall lung cancer identified 26
91 conditionally independent SNPs within 17 genome-wide significant loci ($P < 5 \times 10^{-8}$;
92 Supplementary Fig. 2a; Supplementary Data 2). All 12 loci reported by ILCCO⁷ were
93 confirmed, with consistent direction of effect on all single nucleotide polymorphisms
94 (SNPs) with $P < 1 \times 10^{-5}$, as well as high correlation of effect sizes and allele frequency
95 (Supplementary Fig. 3). Of the 17 genome-wide significant loci for overall lung cancer,
96 four were novel with respect to the broader literature: neuronal growth regulator

97 *LSAMP*, WNT signaling regulator *NMUR2*, DNA damage repair protein *XCL2*, and
98 hedgehog signaling regulator *TULP3*, (Table 1; Supplementary Fig. 4a-d).

99 Further association tests stratified by cancer subtypes LUAD and LUSC in MVP
100 EA (Supplementary Fig. 2bc; Supplementary Data 3-4) replicated associations reported
101 by ILCCO⁷ (Supplementary Fig. 3) and identified additional loci. Two novel EA meta-
102 analysis loci were identified for LUAD, proto-oncogene *MYC* and Wnt signaling inhibitor
103 *TLE3* (Table 1; Supplementary Fig. 4e-h). For LUSC, we identified one novel locus at
104 10q24.31 near NFκB inhibitor *CHUK* and *BLOC1S2*. Across all subtypes for EA meta-
105 analysis index variants, the MVP cohort had associations with $P < 0.05$ in all but one in
106 overall lung cancer, five in LUAD, including approximately nominal significance at
107 rs67824503 (*MYC*; $P = 0.057$), and one in LUSC (Supplementary Data 2-4).

108 We investigated expression quantitative trait loci (eQTL) relationships between
109 top SNPs from the EA meta-analysis across all lung cancer GWAS in GTEx v8 Lung²⁰
110 and the Lung eQTL Consortium²¹ (Supplementary Data 2-4). This analysis showed that
111 the LUSC index SNP rs36229791 on 10q24.31 was associated with the mRNA
112 expression levels of *BLOC1S2* (Fig. 1a-d), consistent with previous TWAS²². *BLOC1S2*
113 is an oncogene whose gene product is associated with centrosome function;
114 centrosomal abnormalities have previously been observed *in vitro* in LUSC^{23,24}.

115 We improved our variant selection by fine-mapping and estimating credible sets
116 of candidate causal variants in EA meta analysis using sum of single effects (SuSiE)^{25,26}
117 modeling. For overall lung cancer, LUAD, and LUSC, we identified 23, 23, and 9 high
118 quality credible sets, respectively, containing 370, 246, and 192 total SNPs
119 (Supplementary Data 5).

120

121 *GWAS in AA*

122 We analyzed overall lung cancer risk in 2,438 cases and 62,112 controls of
123 African ancestry (AA), the largest AA GWAS discovery cohort to date (Supplementary
124 Fig. 5a). Two loci reached genome-wide significance in our discovery scan: 15q25,
125 replicating the association in *CHRNA5* for AA populations reported by an earlier
126 GWAS¹⁸, and a putative novel locus at 12q23 with index SNP rs78994068 (Table 1; Fig.
127 1e). We further performed GWAS in AA within LUAD and LUSC subtypes but found no
128 genome-wide significant associations (Supplementary Fig. 5b-c).

129 The putative AA locus at 12q23 is driven by six SNPs in high linkage
130 disequilibrium (LD; $R^2 > 0.8$) found in long non-coding RNAs *LINC00943* and *LINC00944*
131 (Fig. 1e). These imputed SNPs all had odds ratios (ORs) close to 2, with 1.3%
132 frequency in AA and 0% in EA, consistent with gnomAD v3. *LINC00944* is highly
133 expressed in immune cells and blood, and enriched in T cell pathways in lung tissue
134 and cancer²⁷⁻³⁰. We fine-mapped this locus to define a 95% credible set
135 (Supplementary Data 6), and annotated the functional consequence of the variants
136 using the Variant Effect Predictor (VEP)³¹. Two variants, rs78994068 and rs115962601,
137 were in a known enhancer regulatory region (ENSR00000974920) and thus may involve
138 regulatory changes. However, this locus was directionally consistent but not significant
139 in our AA replication cohort (discussed below); therefore, larger-scale AA analyses are
140 needed to confirm this finding.

141

142 *GWAS multi-ancestry meta-analysis*

143 We conducted a fixed-effect inverse variance-weighted multi-ancestry meta-
144 analysis, combining the EA meta-analysis and the MVP AA GWAS for overall lung
145 cancer, LUAD, and LUSC (Supplementary Data 7-9; Supplementary Fig. 6a-c). This
146 analysis identified two additional novel genome-wide significant loci in overall lung
147 cancer (Table 1; Supplementary Fig. 4i-j): ubiquitin ligase *JADE2*, previously associated
148 with smoking initiation³², and RNA polymerase-associated *RPAP3*. Neither of these
149 novel multi-ancestry meta-analysis loci were reported in a recent multi-ancestry analysis
150 by Byun et al.⁸ that included fewer AA and more Asian ancestry samples, indicating the
151 value our larger AA sample provided for novel discovery. All genome-wide significant
152 EA meta-analysis associations reached genome-wide significance in the multi-ancestry
153 meta-analysis except rs11855650 (*TLE3*) in LUAD ($P=6.19\times 10^{-8}$). We additionally
154 performed random effects meta-analyses using the Han-Eskin method (RE2)³³, and
155 observed similar *P*-values to the fixed effect meta-analysis, with all index variants
156 $P_{RE2}<5\times 10^{-8}$ (Supplementary Data 7-9).

157

158 *Polygenic risk scoring*

159 To gain an understanding of the penetrance and pleiotropy of lung cancer risk,
160 we constructed PRSs based on the ILCCO summary statistics⁷ for every EA subject in
161 MVP. As expected, the PRS was highly associated with both lung cancer risk as well as
162 smoking behavior (Supplementary Fig. 7a-b). Even after removing individuals with any
163 history of lung cancer risk to prevent enrichment of risk factors and comorbidities, the
164 association with smoking behavior remained, suggesting that the PRS is partially

165 capturing genetic smoking behavioral risk factors (Supplementary Fig. 7c). In all groups,
166 individuals at the top decile of the PRS were at significantly higher risk of lung cancer
167 than those in the lowest decile.

168 *Multi-trait conditional analysis for smoking status*

169 Despite adjusting for smoking status, both in MVP EA and ILCCO⁷, a significant
170 genetic correlation was observed between all subsets of lung cancer GWAS and a
171 recently published GWAS of smoking behaviors³⁴ (Fig. 2a, Supplementary Data 10). In
172 order to remove all residual effects of smoking on lung cancer susceptibility, we
173 conducted a multi-trait-based conditional and joint analysis (mtCOJO)^{35,36}, conditioning
174 on a GWAS for cigarettes per day³⁴, which was the smoking trait most strongly
175 correlated with overall lung cancer and subtype GWAS from the EA meta-analysis.
176 Because lung cancer case selection also preferentially selects smokers, conventional
177 adjustment for smoking may inadvertently cause selection bias, which functions as a
178 collider to induce biased genetic effects³⁷. mtCOJO is considered more robust to
179 potential collider bias than conventional covariate adjustment^{35,36}. The total observed-
180 scale SNP-heritability³⁸ of lung cancer risk decreased substantially after conditioning on
181 cigarettes per day, from 5.4% to 3.1% in overall LC, from 6.7% to 5.5% in LUAD, and
182 from 5.8% to 3.8% in LUSC (Fig. 2b; Supplementary Data 11).

183 Significant loci from the conditional analyses are shown in Supplementary Fig. 8-
184 9 and Supplementary Data 12-14. As expected, the statistical significance of loci
185 harboring smoking-related genes (e.g., *CHRNA5*, *CYP2A6*, *CHRNA4*) dropped to below
186 genome-wide significance after conditioning (Fig. 3). Conversely, five signals (four loci)
187 became significant only after conditioning, including novel signals at *MMS22L* in overall

188 lung cancer and 19p13.1 (*ABHD8*) in LUSC. *MMS22L* is a novel GWAS signal but was
189 previously identified as overexpressed in lung cancer in genome-wide gene expression
190 scan³⁹. These may represent biological lung cancer signals partially masked by
191 countervailing genetic effects on smoking behavior. We performed fine-mapping to
192 identify candidate causal variants in the conditioned EA meta-analysis summary
193 statistics, and for overall lung cancer, LUAD, and LUSC, we identified 11, 15, and 6 high
194 quality credible sets, respectively, containing a total of 243, 277, and 78 SNPs
195 (Supplementary Data 5).

196 We constructed PRS based on mtCOJO-conditioned ILCCO summary statistics⁷
197 to directly compare the predictive performance of PRS derived from the conditioned and
198 non-conditioned GWAS in MVP EA. While the PRS based on the non-conditioned
199 overall lung cancer GWAS exhibited reduced performance in never-smokers compared
200 to ever-smokers, the PRS based on the conditional analysis resulted in similar
201 performance across smoking status (Fig. 2c; Supplementary Data 15).

202

203 *Replication of novel variants in OncoArray and combined meta-analysis*

204 We queried the OncoArray Consortium Lung Study (OncoArray) as an external
205 non-overlapping replication dataset for our significant GWAS signals (Supplementary
206 Data 16-17). For GWAS in EA meta-analysis for overall lung cancer, LUAD, and LUSC,
207 we replicated five of seven novel loci ($P < 0.01$) in an OncoArray European ancestry
208 cohort: *XCL2* and *TLE3* in overall lung cancer, *MYC* and *TLE3* in LUAD, and *BLOC1S2*
209 in LUSC. The novel African ancestry association for overall lung cancer at *LINC00944*
210 was not replicated. We meta-analyzed OncoArray European and African ancestry

211 participants to replicate our multi-ancestry meta-analysis signals for overall lung cancer
212 at *RPAP3* ($P=0.0044$) and *JADE2* which bordered on nominal significance (rs329122;
213 $P=0.053$). For the two novel loci which were identified in EA meta-analysis conditioned
214 on cigarettes per day, we included smoking as a covariate for association analysis in
215 the OncoArray European ancestry cohort. These association signals were replicated for
216 overall lung cancer at *MMS22L* ($P=0.006$) and LUSC at *ABHD8* ($P=0.003$). In a variant-
217 level replication of 137 conditionally independent discovery associations which fell
218 within ≤ 1 Mb of a previously reported lung cancer GWAS signal, 134 had $P < 0.05$ in
219 OncoArray, and 42 had $P < 5 \times 10^{-8}$ (Supplementary Data 18).

220 We then performed a combined meta-analysis of our discovery results with
221 OncoArray replication results (Supplementary Data 18). We considered a conservative
222 threshold of $P=4.17 \times 10^{-9}$ ($P=5 \times 10^{-8}/12$ total GWAS analyses) to be significant, which
223 was met by 9 of the 12 loci. Because rs329122 in *JADE2* achieved the more
224 conservative significance threshold ($P=3.69 \times 10^{-9}$), and has also been associated with
225 smoking behavior³² and identified as a splicing-related variant associated with lung
226 cancer⁴⁰, we considered this locus to be replicated. In the combined meta-analysis we
227 observed similar P -values in fixed effects and random effects (RE2) models.

228 Next, for all previously reported lung cancer and subtype loci in this study, we
229 identified lung cancer associations from GWAS Catalog which fell within the same loci
230 as our index variants (Supplementary Data 19). We confirmed two loci that previously
231 had been reported only in a recent genome-wide association by proxy (GWAX) of lung
232 cancer⁴¹: *CENPC* (rs75675343) in overall lung cancer in the EA meta-analysis
233 ($P=2.40 \times 10^{-8}$) and the multi ancestry meta-analysis, and *TP53BP1* in overall lung

234 cancer in the multi-ancestry meta-analysis (rs9920763; $P=1.63\times 10^{-8}$). Our multi-
235 ancestry meta-analysis for overall lung cancer also confirmed a recently reported locus
236 at 4q32.2 (*NAF1*)¹⁵ in East Asian ancestry.

237

238 *Multi-trait analysis with breast cancer*

239 At 19p13.1, a known pleiotropic cancer locus^{42,43}, the index SNP of LUSC
240 conditioned on smoking (rs61494113) sits in a gene-rich region where a recent fine-
241 mapping effort of breast cancer risk loci⁴⁴ proposed two independent associations, one
242 affecting the regulation of *ABHD8* and *MRPL34*, and another causing a coding mutation
243 in *ANKLE1*. Here, we used the increased power provided by a multi-trait analysis of
244 GWAS (MTAG)⁴⁵ of LUSC and estrogen receptor negative (ER-) breast cancer⁴⁶ to
245 disentangle the complex relationships between cancer risk and the genes in this locus
246 (Fig. 4a). Overexpression of *ABHD8* has been shown to significantly reduce cell
247 migration^{42,43}. Similar odds ratios at rs61494113 were observed across LUSC and
248 breast cancer, and MTAG enhanced the GWAS signal at this locus (Fig. 4b).

249 We used the coloc-SuSiE method⁴⁷ to assess colocalized associations between
250 pairs of credible sets in this locus underlying the risk of LUSC and ER- breast cancer,
251 allowing for multiple causal signals. We found evidence for a shared causal signal
252 between credible sets in the LUSC conditional meta-analysis and ER- breast cancer
253 (97.7% posterior probability; Supplementary Data 20). The index SNPs for the credible
254 sets of LUSC conditioned on smoking and ER- breast cancer (rs61494113 and
255 rs56069439, respectively) have $r^2=0.99$.

256 The eQTL effect of *ABHD8* was replicated in multiple tissues of GTEx v8,
257 including Lung (Fig. 4c). Interestingly, the group of SNPs in the LUSC-BC credible set
258 did not have the most significant eQTL effect, suggesting a complex relationship
259 between the multiple causal variants at the locus and gene expression (Fig. 4d). For
260 instance, a recent splice variant analysis⁴⁸ implicated splicing of *BABAM1* (a BRCA1-
261 interacting protein) as a culprit of the associations observed in 19p13.1. Consistent with
262 previous reports^{42,43}, the cancer risk-increasing haplotype was correlated with increased
263 expression of *ABHD8* and alternative splicing of *BABAM1*. However, there was no
264 overlap between the 95% eQTL credible sets of *ABHD8* and *BABAM1*, and neither of
265 the credible sets included rs61494113.

266

267 *Phenome-wide association study*

268 Finally, to investigate the pleiotropy of lung cancer genetic risk in the absence of
269 the overwhelming effect of smoking behavior, we performed PheWAS in MVP using the
270 PRS scores constructed from the ILCCO summary statistics⁷ for overall lung cancer,
271 both based on the standard GWAS (“unconditioned PRS”; Fig. 5a; Supplementary Data
272 21) and the GWAS conditioned on cigarettes per day using mtCOJO (“conditioned
273 PRS”; Fig. 5b; Supplementary Data 22). Each PRS was tested for association with
274 1,772 phecode-based phenotypes. Overall, 240 phenotypes were associated with the
275 unconditioned PRS and 112 were associated with the conditioned PRS at a Bonferroni-
276 corrected significance threshold ($P < 0.05/1,772$). Although lung cancer remained a top
277 association with the conditioned PRS, the association with tobacco use disorder was
278 greatly reduced, from an OR associated with a standard deviation increase in the PRS

279 of 1.151 [1.142-1.160] ($P=2.32\times 10^{-237}$) in the unconditioned PRS to OR=1.046 [1.038-
280 1.053] ($P=1.05\times 10^{-32}$) in the conditioned PRS. However, the effect on alcohol use
281 disorder was only modestly attenuated between the unconditioned (OR=1.098 [1.089-
282 1.108]; $P=1.05\times 10^{-87}$) and conditioned LC (OR=1.078 [1.069-1.088], $P=4.41\times 10^{-60}$)
283 PRSs. Whether a role for alcohol in lung cancer exists independently of smoking is
284 controversial^{49,50}; this analysis suggests that may be the case. Other putatively
285 smoking-related associations, such as chronic obstructive pulmonary disease,
286 pneumonia, and peripheral vascular disease were greatly diminished with the
287 conditioned PRS. Mood disorders, depression, and post-traumatic stress disorder, were
288 also significantly associated with the unconditioned PRS but no longer significantly
289 associated with the conditioned PRS, reflecting neuropsychiatric correlates of smoking
290 behavior.

291 Intriguingly, a category of metabolic traits that were not associated with the
292 unconditioned PRS were highly associated with the conditioned PRS and in a negative
293 effect direction. We observed protective associations of the conditioned PRS with
294 metabolic traits such as type 2 diabetes (OR=0.945 [0.938-0.952], $P=9.46\times 10^{-52}$) and
295 obesity (OR=0.952 [0.945-0.959], $P=2.48\times 10^{-41}$). Neither were associated with the
296 unconditioned PRS (OR=1.006 [0.999-1.014]; $P=0.092$, and OR=1.005 [0.998-1.012];
297 $P=0.183$, respectively). Other traits in this category included sleep apnea and
298 hyperlipidemia. These findings are consistent with prior observational findings of an
299 inverse relationship between BMI and lung cancer⁵¹ and illustrate the extent to which
300 smoking may be a major confounder of this relationship.

301 Finally, we observed strong associations of the lung cancer PRS with skin cancer
302 and related traits, such as actinic keratitis. In basal cell carcinoma, the OR increased
303 from 1.087 [1.072-1.102] ($P=6.06\times 10^{-32}$) with the unconditioned PRS to 1.105 [1.090-
304 1.120] ($P=1.82\times 10^{-47}$) with the conditioned PRS. As a sensitivity analysis, we tested the
305 strength of this association after removing the *TERT* locus, which is prominently
306 associated with both traits. Doing so only modestly reduced the effect of the conditioned
307 PRS to OR=1.092 [1.077-1.107] ($P=4.08\times 10^{-36}$). Thus, our results are consistent with a
308 genome-wide genetic correlation between lung cancer and basal cell carcinoma that is
309 strengthened when the effect of smoking is removed. Overall, our results suggest that
310 the biology underlying lung cancer risk may be partially masked by the residual genetic
311 load of smoking.

312

313 **Discussion**

314 We identified novel lung cancer-associated loci in a new cohort of EA and AA
315 participants, including the largest AA cohort analyzed to-date. We also show that,
316 despite studies on the genetic basis of lung cancer risk taking smoking status into
317 account, the effects of smoking continue to obfuscate our understanding of lung cancer
318 genetics. In particular, we report two novel loci, at *MMS22L* (overall) and *ABHD8*
319 (*LUSC*), which may be partially masked by countervailing genetic effects on smoking.
320 Our replication analysis which adjusted for smoking pack-years confirmed these loci.
321 Additionally, our analyses demonstrated that PRSs for lung cancer contain large
322 uncorrected genetic loading for smoking behavioral factors. Our results indicate that
323 controlling for these factors can improve risk assessment models, potentially improving

324 lung cancer screening even for non-smokers. Finally, our phenomic scans comparing
325 PRSs derived from GWAS with and without genomic conditioning on smoking showed
326 divergent associations across numerous traits, especially metabolic phenotypes.

327 The increased sample size in this study enabled the interpretation of multiple
328 causal variants underlying the gene-rich *ADHL8-BABAM1* region, synthesizing prior
329 observations into a clearer understanding of this locus. Our other novel loci strengthen
330 established lung cancer mechanisms. We identify for the first time a susceptibility locus
331 at *MYC*, a well-known oncogene and master immune regulator. *XCL2* is involved in
332 cellular response to inflammatory cytokines⁵². *LSAMP* is a tumor suppressor gene in
333 osteosarcoma⁵³, and 3q13.31 homozygous deletions have been implicated in
334 tumorigenesis⁵⁴. *TLE3* is a transcriptional corepressor involved in tumorigenesis and
335 immune function⁵⁵. The transcription factor *TULP3* has been implicated in pancreatic
336 ductal adenocarcinoma and colorectal cancer⁵⁶. *XCL2*, *NMUR2*, and *TULP3* may also
337 be related to cancer progression via G-protein-coupled receptor (GPCR) signaling
338 pathways⁵⁷. *JADE2* expression has been experimentally linked to NSCLC⁵⁸, and has
339 been identified in GWAS of smoking behavior³⁴. Finally, DNA damage repair
340 mechanisms emerge, including *RPAP3*, an RNA polymerase that may be involved in
341 DNA damage repair regulation⁵⁹, and *MMS22L* which repairs double strand breaks⁶⁰.

342 Although smoking is the major risk factor for lung cancer, it is important to clearly
343 disentangle the effect of smoking to fully understand the complex genetic and
344 environmental causes of lung cancer. Our approach enables the development of new
345 polygenic scores, which can improve precision medicine applications for lung cancer in
346 both smokers and nonsmokers.

347 **Author contributions statement**

348 Drafted the manuscript: B.R.G., M.F., S.-G. J., A.K.S., E.P., A.K.D., S.P.

349 Acquired the data: B.R.G., S.-G. J., A.K.S., Y.S., P.D., U.S., D.D.S., W.T., J.M., S.M.,
350 R.R., R.J.H., J.D.M., Y.B., C.I.A., S.P.

351 Analyzed the data: B.R.G., S.-G. J., M.F., A.K.S., Y.S., P.D., U.S., Y.B., R.S.

352 Critically revised the manuscript for important intellectual content: all authors.

353 **Acknowledgements**

354 This work was supported by award #MVP000 from the United States Department
355 of Veterans Affairs (VA) Million Veteran Program. The contents of this publication are
356 the sole responsibility of the authors and do not necessarily represent the views of VA
357 or the United States Government. Where authors are identified as personnel of the
358 International Agency for Research on Cancer/World Health Organization, the authors
359 alone are responsible for the views expressed in this article, and they do not necessarily
360 represent the decisions, policy, or views of the International Agency for Research on
361 Cancer/World Health Organization. Full consortium acknowledgements for MVP and the
362 ILCCO OncoArray study⁷ are provided in Supplementary Information.

363

364 **Subject terms and techniques**

365 Biological sciences > Cancer > Lung cancer

366 Biological sciences > Genetics > Genetic association study > Genome-wide association
367 studies

368 **Data Availability**

369 The full summary level association data from the individual population analyses in MVP
370 will be available upon publication via the dbGaP study accession number phs001672.

371

372 **Competing interests**

373 S.-G.J. is an employee and shareholder of BridgeBio Pharma. The other authors
374 declare no competing interests.

375 **Methods**

376 *Cohort definition*

377 Patients were identified from MVP participants¹⁹ utilizing clinical information
378 available through the United States Department of Veterans Affairs (VA) Corporate Data
379 Warehouse (CDW) with ICD codes for primary lung cancer. Occurrences of the ICD-9
380 codes 162.3, 162.4, 162.5, 162.8, and 162.9 or the ICD-10 codes C34.10, C34.11,
381 C34.12, C34.2, C34.30, C34.31, C34.32, C34.80, C34.81, C34.82, C34.90, C34.91, and
382 C34.92 were used in case identification. Patients with secondary lung cancer were
383 excluded from the cohort using ICD-9/10 codes 197.x, C78.00, C78.01, and C78.02.
384 Additional patients were identified in the VA Cancer Registry using ICD-O site, including
385 lung/bronchus, other respiratory system or intrathoracic organs, or trachea. The Cancer
386 Registry was also used to determine the lung cancer subtypes LUAD and LUSC among
387 cases.

388 Preliminary totals of 18,633 and 10,845 patients with MVP participation were
389 identified from the VA CDW and Cancer Registry, respectively. A combined cohort of
390 20,631 unique patients was generated for further analysis. The cohort was
391 predominantly male (~95%) with a median age of 64–68 for sub-cohorts, depending on
392 ancestry assignments and cancer subtypes. The cohort was curated further to remove
393 any participant with missing data. The final cohorts are described in Supplementary
394 Data 1.

395 Once patients were identified from VA's CDW and Cancer Registry, cases were
396 used to gather records related to age, sex, smoking status, and ancestry. Smoking
397 status included former, current, and never, based on the MVP survey at the time of

398 enrollment and on electronic medical records. Ancestry was defined using a machine
399 learning algorithm that harmonizes self-reported ethnicity and genetic ancestry
400 (HARE)⁶¹. All analyses described here were performed on patients of EA or AA ancestry
401 in ancestry-stratified cohorts. Additionally, the cohorts were further stratified by lung
402 cancer subtypes for analysis. Matched controls were selected based on age, gender,
403 smoking status, and HARE assignments. Age was binned into 5-year intervals for this
404 purpose.

405 *Array genotyping, genotype quality control, and principal component analysis*

406 Genotyping and quality control were conducted as described previously⁶². Briefly,
407 we removed all samples with excess heterozygosity (F statistic < -0.1), excess
408 relatedness (kinship coefficient ≥ 0.1 with 7 or more MVP samples), and samples with
409 call rates $< 98.5\%$. Additional samples with a mismatch between self-reported sex and
410 genetic sex were removed.

411 Principal component (PC) analysis was conducted using PLINK 2.0⁶³
412 (v2.00a3LM), on a pruned set of SNPs (window size 1Mb, step size 80, $r^2 < 0.1$, minor
413 allele frequency (MAF) < 0.01 , Hardy-Weinberg equilibrium $P < 1 \times 10^{-10}$, missingness
414 rate $< 10\%$) within European ancestry (EA) and African ancestry (AA) on unrelated
415 individuals, where unrelated individuals were defined as greater than third-degree
416 relatives as previously described⁶². PCs were then projected onto related individuals in
417 EA.

418 *Imputation*

419 Prior to imputation, a within-cohort pre-phasing procedure was applied across the
420 whole cohort by chromosome using Eagle2⁶⁴. Imputation was then conducted on pre-

421 phased genotypes using Minimac4⁶⁵ and the 1000 Genomes Phase 3 (v5) reference
422 panel⁶⁶ in 20Mb chunks and 3Mb flanking regions. Quality of imputation (Minimac Rsq
423 or INFO) was then re-computed in EA and AA separately to be used as filters for
424 respective GWAS. Imputed loci reaching genome-wide significance were tested for
425 deviation from Hardy-Weinberg equilibrium (HWE) in 61,538 EA controls
426 (Supplementary Data 23). Of the 93 conditionally independent SNPs across the GWAS
427 analyses, 6 SNPs had a significant ($P < 1 \times 10^{-6}$) HWE signal; unsurprisingly, the
428 strongest HWE signal was from SNPs in the Major Histocompatibility Complex region.
429 However, none of the 12 novel loci reported in Table 1 significantly deviated from HWE.

430 *Association analyses*

431 For the EA lung cancer overall and subtype GWAS, we performed standard
432 logistic regression using PLINK 2.0 (v2.00a2LM)⁶³ with a matched control design. EA
433 GWAS was performed in unrelated individuals, defined as greater than third-degree
434 relatives. For the AA lung cancer overall and subtype analyses, because the case
435 numbers were smaller, we performed a mixed-model logistic regression using
436 REGENIE (v1.0.6.7)⁶⁷; REGENIE applies a whole genome regression model to control
437 for relatedness and population structure, and includes a Firth correction to control for
438 bias in rare SNPs as well as case-control imbalance. GWAS covariates for each
439 ancestry included age, age-squared, sex, and smoking status as a categorical variable
440 (current, former, never), and the first ten principal components. Participants with missing
441 smoking status (n=786) were removed.

442 *EA meta-analysis*

443 We performed inverse-variance weighted meta-analyses of MVP-EA summary
444 statistics and summary statistics previously reported by ILCCO⁷ using METAL
445 (v20100505)⁶⁸ with scheme STDERR. Significant inflation across GWAS and meta-
446 analyses was not observed (all genomic control values (λ) for GWAS in this study
447 ≤ 1.15). Only variants present in both studies were meta-analyzed. We further performed
448 a sensitivity analysis using the Han-Eskin random effects model (RE2) in METASOFT
449 v2.0.1³³.

450 *Lung eQTL consortium*

451 The lung tissues used for eQTL analyses were from human subjects who
452 underwent lung surgery at three academic sites: Laval University, University of British
453 Columbia (UBC), and University of Groningen. Genotyping was carried out using the
454 Illumina Human1M-Duo BeadChip. Expression profiling was performed using an
455 Affymetrix custom array (see GEO platform GPL10379). Only samples that passed
456 genotyping and gene expression quality controls were considered for eQTL analysis,
457 leaving sample sizes of 409 for Laval, 287 for UBC, and 342 for Groningen. Within each
458 set, genotypes were imputed in each cohort with the Michigan Imputation Server⁶⁵ using
459 the Haplotype Reference Consortium⁶⁹ version 1 (HRC.r1-1) data as a reference set,
460 and gene expression values were adjusted for age, sex, and smoking status.
461 Normalized gene expression values from each set were then combined with ComBat⁷⁰.
462 eQTLs were calculated using a linear regression model and additive genotype effects
463 as implemented in the Matrix eQTL package in R⁷¹. Cis-eQTLs were defined by a 2 Mb
464 window, i.e., 1 Mb distance on either side of lung cancer-associated SNPs. Pre-

465 computed lung eQTLs were also obtained from the Genotype-Tissue Expression
466 (GTEx) Portal²⁰. Lung eQTLs in GTEx (version 8) are based on 515 individuals and
467 calculated using FastQTL⁷².

468 *Fine-mapping*

469 We performed Bayesian fine-mapping the genome-wide significant loci from EA
470 meta-analysis and AA using the FinnGen fine-mapping pipeline⁷³
471 (<https://github.com/FINNGEN/finemapping-pipeline>) and SuSiE^{25,26}. Pairwise SNP
472 correlations were calculated directly from imputed dosages on European-ancestry MVP
473 samples from this analysis using LDSTORE 2.0⁷³. The maximum number of allowed
474 causal SNPs at each locus was set to 10. Fine-mapping regions which overlapped the
475 major histocompatibility complex (MHC; chr6:25,000,000-34,000,000) were excluded.
476 High quality credible sets were defined as those with minimum $r^2 < 0.5$ between variants.
477 The functional consequences of the AA credible set variants were annotated using the
478 Variant Effect Predictor (VEP)³¹.

479 *Replication analysis*

480 External replication was performed for all genome-wide significant associations in
481 overall lung cancer, LUAD, and LUSC in OncoArray Consortium Lung Study
482 (OncoArray)^{8,74}. Replication for genome-wide significant multi-ancestry associations
483 was performed in a fixed effects meta-analysis of OncoArray CEU Europeans for
484 significant EA meta-analysis associations, and in an YRI AA meta-analysis composed of
485 5 studies⁸ for significant MVP AA associations. Meta-analysis associations from this
486 study were replicated against a meta-analysis of these OncoArray groups. To replicate
487 significant variants from EA analysis conditioned on smoking, pack-years was

488 additionally included as a covariate in replication cohorts. There was no participant
489 overlap between the replication cohorts and the ILCCO study⁷ used in the discovery
490 scan. Covariates included the first five genetic principal components and participant
491 study sites. Proxy SNPs were used to replicate known associations at rs75675343
492 (rs2318539/4:67831628:C:A; $R^2_{EUR}=1$) and rs4586884 (rs4435699/4:164019500:C:G;
493 $R^2_{EUR}=0.999$).

494

495 *Multi-ancestry meta-analysis*

496 A multi-ancestry meta-analysis of MVP EA and AA cohorts with summary
497 statistics previously reported by ILCCO⁷ was conducted in METAL⁶⁸ using an inverse
498 variance-weighted fixed effects scheme. Only variants present in two or more cohorts
499 were meta-analyzed. Index variants were defined using the two-stage “clumping”
500 procedure implemented in the Functional Mapping and Annotation (FUMA) platform⁷⁵.
501 In this process, genome-wide significant variants are collapsed into LD blocks ($r^2>0.6$)
502 and subsequently re-clumped to yield approximately independent ($r^2<0.1$) signals;
503 adjacent signals separated by <250kb are ligated to form independent loci. Novel
504 variants are defined as meta-analysis index variants located >1Mb from previously
505 reported lung cancer associations. We additionally performed a sensitivity analysis
506 using the random effects model (RE2) in METASOFT v2.0.1³³.

507 *Polygenic risk score (PRS) calculation*

508 We used PRS-CS⁷⁶ to generate effect size estimates under a Bayesian
509 shrinkage framework, and then used PLINK 2.0 (v2.00a3LM)⁶³ to linearly combine
510 weights into a risk score using a global shrinkage prior of 1×10^{-4} , which is

511 recommended for less polygenic traits. Finally, scores were normalized to a mean of 0
512 and a standard deviation of 1.

513 *Multi-trait analyses*

514 In order to remove all residual effects of smoking on lung cancer susceptibility,
515 we conducted a multi-trait meta-analysis³⁵ conditioned on cigarettes per day, which was
516 shown to be most significantly correlated with all lung cancer GWAS³⁴. The meta-
517 analysis was performed on the EA meta-analysis summary statistics using mtCOJO,
518 part of the GCTA software package⁷⁷. An LD reference was constructed from 50,000
519 MVP EA samples.

520 Multi-trait analysis of GWAS (MTAG)⁴⁵ (v0.9.0) was applied using genome-wide
521 LUSC summary statistics after conditioning on cigarettes per day, and estrogen
522 receptor negative (ER-) breast cancer summary statistics⁴⁶ which were munged using
523 LDSC (v1.01)³⁸. Single causal variant colocalization between LUSC conditioned on
524 cigarettes per day and ER- breast cancer was performed using Coloc (R; version 4)⁷⁸
525 for variants at *ABHD8* (chr19: 17,350,000 to 17,475,000). A posterior probability > 0.9
526 for Hypothesis 4 (both traits are associated and share a single causal variant) was used
527 as the criteria for colocalization.

528 *Heritability and genetic correlations*

529 Linkage Disequilibrium score regression (LDSC) v1.0.1 was used to calculate
530 observed-scale SNP-heritability³⁸ using lung cancer and subtypes summary statistics,
531 before and after conditioning on cigarettes per day. Pairwise genetic correlations were
532 estimated between lung cancer and subtypes from MVP, ILCCO⁷, and EA meta-

533 analysis, and four smoking traits (smoking initiation, cigarettes per day, smoking
534 cessation, and age of initiation)³⁴.

535 *Conditional and joint SNP analysis*

536 To find independently associated genome-wide significant SNPs at each locus in
537 a stepwise fashion, we used GCTA-COJO using the --cojo-slct option. An LD reference
538 was constructed from 50,000 MVP EA samples. Variants with MAF<0.01 in the COJO
539 reference panel were not included in identification of independent signals. LDTrait⁷⁹ was
540 queried to identify previously published significant GWAS variants within 1Mb of our
541 index variants in all populations. Novel loci were defined as those at which the index
542 variant was not within ± 500 kb of previously reported genome-wide significant lead
543 SNPs for lung cancer or its subtypes in any ancestry.

544 *Phenome-wide association study (PheWAS)*

545 We conducted a PheWAS of electronic health record-derived phenotypes and lab
546 results in EA subjects using either the normalized PRS as the predictor or
547 independently associated genome-wide significant SNPs. Comparison of unconditioned
548 PRS PheWAS and conditioned PRS PheWAS were based on ILCCO summary
549 statistics⁷ and used MVP EA as the out-of-sample test set. Associations were tested
550 using the R PheWAS package⁸⁰ version 0.1 with QC procedures described previously⁸¹.
551 Control and sex-based exclusion criteria were applied.

Main Tables

Table 1: Novel genome-wide significant loci and their respective index variants associated with lung cancer risk in European-ancestry meta-analyses from MVP and ILCCO⁷ cohorts, MVP African ancestry, multi-ancestry meta-analyses, and in European-ancestry meta-analyses after conditioning on cigarettes per day. LUAD, adenocarcinoma; LUSC, squamous cell carcinoma; EA, effect allele; NEA, non-effect allele; EAF, effect allele frequency in the given population; OR (95% CI), odds ratio and 95% confidence interval.

Lung cancer subtype	rsID	Cytoband	Position (hg19)	Candidate gene	EA	NEA	EAF	Discovery OR (95% CI)	Discovery <i>P</i>	Replication OR (95% CI)	Replication <i>P</i>	Combined meta-analysis OR (95% CI)	Combined meta-analysis <i>P</i>
Novel loci from the European ancestry GWAS meta-analysis													
Overall	rs77045810	1q24.2	168,505,017	<i>XCL2</i>	A	C	0.89	1.10 (1.07, 1.13)	1.43×10 ⁻¹⁰	1.07 (1.02, 1.13)	0.0057	1.09 (1.07, 1.12)	3.94×10 ⁻¹²
Overall	rs144840030	3q13.31	117,147,326	<i>LSAMP</i>	T	G	0.01	1.31 (1.19, 1.44)	1.09×10 ⁻⁸	1.07 (0.88, 1.30)	0.49	1.26 (1.16, 1.37)	5.01×10 ⁻⁸
Overall	rs62400619	5q33.1	152,343,053	<i>NMUR2</i>	T	C	0.68	1.06 (1.04, 1.08)	6.33×10 ⁻⁹	1.03 (0.99, 1.06)	0.16	1.05 (1.03, 1.07)	1.10×10 ⁻⁸
Overall	rs9988980	12p13.33	3,038,917	<i>TULP3</i>	T	C	0.39	1.05 (1.04, 1.08)	5.34×10 ⁻⁸	1.05 (1.02, 1.09)	0.0022	1.05 (1.04, 1.07)	3.72×10 ⁻¹⁰
LUAD	rs67824503	8q24.21	129,535,264	<i>MYC</i>	T	C	0.75	1.10 (1.07, 1.14)	1.81×10 ⁻⁸	1.11 (1.05, 1.16)	5.05×10 ⁻⁵	1.10 (1.07, 1.14)	4.09×10 ⁻¹²
LUAD	rs11855650	15q23	70,431,773	<i>TLE3</i>	T	G	0.38	1.09 (1.06, 1.12)	1.12×10 ⁻⁸	1.12 (1.07, 1.17)	1.22×10 ⁻⁷	1.10 (1.07, 1.13)	1.15×10 ⁻¹⁴
LUSC	rs36229791	10q24.31	101,991,135	<i>BLOC1S2</i>	A	T	0.04	1.27 (1.17, 1.38)	4.04×10 ⁻⁸	1.25 (1.12, 1.41)	1.49×10 ⁻⁴	1.26 (1.18, 1.35)	2.48×10 ⁻¹¹
Novel loci from the African ancestry GWAS													
Overall	rs78994068	12q24.32	127,225,803	<i>LINC00944</i>	C	A	0.01	2.13 (1.66, 2.72)	1.87×10 ⁻⁹	1.026 (0.681, 1.548)	0.90	1.76 (1.42, 2.17)	1.81×10 ⁻⁷
Novel loci from the multi-ancestry meta-analysis (not genome-wide significant in the European meta-analysis)													
Overall	rs329122	5q31.1	133,864,599	<i>JADE2</i>	A	G	0.43	0.95 (0.93, 0.97)	1.12×10 ⁻⁸	0.97 (0.94, 1.00)	0.053	0.96 (0.94, 0.97)	3.69×10 ⁻⁹
Overall	rs7300571	12q13.11	47,857,826	<i>RPAP3</i>	T	C	0.11	1.08 (1.05, 1.12)	3.47×10 ⁻⁸	1.07 (1.02, 1.13)	0.0044	1.08 (1.06, 1.11)	6.48×10 ⁻¹⁰
Novel loci after conditioning on cigarettes per day from the European ancestry GWAS meta-analysis													
Overall	rs1124241	6q16.1	97,722,453	<i>MMS22L</i>	A	G	0.22	1.08 (1.05, 1.11)	1.26×10 ⁻⁸	1.06 (1.02, 1.11)	0.0062	1.08 (1.05, 1.10)	3.39×10 ⁻¹⁰
LUSC	rs61494113	19p13.11	17,401,859	<i>ABHD8</i>	A	G	0.29	1.12 (1.07, 1.16)	4.90×10 ⁻⁸	1.10 (1.03, 1.17)	0.0031	1.11 (1.08, 1.15)	6.39×10 ⁻¹⁰

560 **Main Figure captions**

561 **Figure 1. Highlighted novel GWAS loci. a-d)** The meta-analysis of squamous cell
562 lung carcinoma (LUSC) in European ancestry (EA) identifies a novel locus at 10q24.31.
563 **a)** Odds ratios for rs36229791 in LUSC compared to lung adenocarcinoma (LUAD) and
564 overall lung cancer. **b)** *BLOC1S2* expression varies by genotype at rs36229791. **c)**
565 *BLOC1S2* eQTL t statistic vs LUSC z statistic. **d)** Regional association plot showing
566 SNP significance and genes around lead SNP rs36229791. **e)** The African ancestry
567 GWAS highlights a putatively novel locus on chr12 at *LINC00944*. The risk allele has
568 effectively 0% frequency in EA.

569
570 **Figure 2. Association of lung cancer GWAS with smoking behaviors. a)** Genetic
571 correlations (with 95% confidence interval) between the lung cancer GWAS and
572 smoking behaviors, including smoking initiation, cigarettes per day, smoking cessation,
573 and age of initiation. **b)** SNP heritability for the meta-analysis and conditional meta-
574 analysis. The heritability decreases in the conditional analysis for overall lung cancer as
575 well as both subtypes, suggesting that some portion of the heritability of lung cancer is
576 due to smoking behavior. **c)** Polygenic risk scores (PRS) based on standard lung
577 cancer GWAS (blue) performs worse in never-smokers than former or current smokers,
578 while conditioning on smoking behavior (orange) results in similar performance.

579
580 **Figure 3. Forest plot of genome-wide significant associations.** Within each cancer
581 subtype, changes in effect size and significance are shown before and after conditioning

582 on cigarettes per day. Novel loci are indicated by an asterisk after the gene name (*).
583 Loci that became significant after conditioning ($P < 5 \times 10^{-8}$) are in red.

584

585 **Figure 4. Significant locus after conditioning on smoking behavior, 19p13.11, has**
586 **pleiotropic associations with ER-negative breast cancer. a)** Regional association
587 plot of the 19p13.11 multi-trait analysis of GWAS (MTAG) locus. **b)** Odds ratios for lead
588 SNP rs61494113 in squamous cell lung carcinoma (LUSC), before and after
589 conditioning, and MTAG analysis, compared to lung adenocarcinoma and overall lung
590 cancer. **c)** *ABHD8* expression varies by genotype at rs61494113. **d)** *ABHD8* eQTL t
591 statistic vs LUSC z statistic; red X's indicate the 95% credible set.

592

593 **Figure 5. Phenome-wide association study (PheWAS) of polygenic risk scores**
594 **(PRS) of lung cancer and lung cancer conditioned on cigarettes per day. a)**
595 PheWAS of PRS on lung cancer is mostly confounded with smoking associations. **b)**
596 PheWAS of the conditional meta-analysis PRS shows associations with skin cancer and
597 metabolic traits.

598 **Supplementary Figure captions**

599

600 **Supplementary Fig. 1. Study overview.** Genome-wide association studies were
601 performed in Million Veteran Program (MVP) European and African ancestry (AA)
602 cohorts for overall lung cancer, adenocarcinoma, and squamous cell carcinoma. MVP
603 and International Lung Cancer Consortium OncoArray (ILCCO) European cohorts were
604 meta-analyzed, and further meta-analyzed with AA for multi-ancestry meta-analysis.
605 Multi-trait conditional meta-analysis was performed on EA using average cigarettes per
606 day from Liu et al. (2019). Replication and combined meta-analysis was performed
607 using external OncoArray cohorts.

608

609 **Supplementary Fig. 2. Manhattan plots and quantile-quantile (QQ) plots for**
610 **European meta-analyses.** Manhattan and QQ plots are shown for **a)** overall lung
611 cancer; **b)** lung adenocarcinoma (LUAD); and **c)** squamous cell lung carcinoma (LUSC).
612 Cytoband positions for significant loci are noted in each Manhattan plot; putatively novel
613 loci identified in this study are in red; externally replicated novel loci are indicated by a
614 box. Genomic control (λ) values, LDSC intercepts, and sample sizes are inset in QQ
615 plots.

616

617 **Supplementary Fig. 3. Effect allele frequency concordance between International**
618 **Lung Cancer Consortium OncoArray (ILCCO) and Million Veteran Program**
619 **European ancestry (EA) GWAS. (a-c)** Effect allele frequency concordance for all
620 variants tested in both studies with $P < 1 \times 10^{-5}$ in ILCCO for **a)** overall lung cancer, **b)**

621 lung adenocarcinoma, and **c**) squamous cell lung carcinoma. Points are styled based
622 on significance level in MVP. **(d-f)** Effect size concordance for genome-wide significant
623 variants in **d)** overall lung cancer, **e)** lung adenocarcinoma, and **f)** squamous cell lung
624 carcinoma. One-to-one concordance is shown as a dashed line. Index variants from the
625 EA meta-analysis between ILCCO and MVP are annotated by locus. Novel significant
626 loci after meta-analysis are annotated in red.

627

628 **Supplementary Fig. 4. Genome-wide significant novel lung cancer loci.** Forest
629 plots (left) and regional Manhattan plots (right) for novel loci from European meta-
630 analysis: **a)** *XCL2*, **b)** *LSAMP*, **c)** *NMUR2*, **d)** *TUPL3*, **e)** *MYC*, **f)** *TLE3*, and **g)**
631 *BLOC1S2*; and from multi-ancestry meta-analysis: **h)** *JADE2*; **i)** *RPAP3*.

632

633 **Supplementary Fig. 5. Manhattan plots and quantile-quantile (QQ) plots for MVP**
634 **African ancestry.** Manhattan and QQ plots are shown for **a)** African ancestry overall
635 lung cancer; **b)** lung adenocarcinoma (LUAD); and **c)** squamous cell lung carcinoma
636 (LUSC). Cytoband positions for significant loci are noted in each Manhattan plot;
637 putatively novel loci identified in this study are in red. Genomic control (λ) values and
638 sample sizes are inset in QQ plots.

639

640 **Supplementary Fig. 6. Manhattan plots and quantile-quantile (QQ) plots for multi-**
641 **ancestry meta-analyses.** Manhattan and QQ plots are shown for **a)** the multi-ancestry
642 meta-analysis in overall lung cancer; **b)** lung adenocarcinoma (LUAD); and **c)**
643 squamous cell lung carcinoma (LUSC). Cytoband positions for significant loci are noted

644 in each Manhattan plot; novel loci not identified in the European meta-analysis are in
645 red; externally replicated novel loci are indicated by a box. Genomic control (λ) values
646 and sample sizes are inset in QQ plots.

647

648 **Supplementary Fig. 7. Association of the lung cancer polygenic risk score (PRS)**
649 **with lung cancer by smoking status. a)** Association of the lung cancer PRS with
650 overall lung cancer risk. The risk of lung cancer reached an odds ratio (OR) of 2.51
651 (95% confidence interval: 1.80, 3.51) in the top decile. **b)** Association of the lung cancer
652 PRS with lung cancer risk in never-smokers. Among never-smokers, lung cancer risk
653 reached an OR of 2.67 (2.40, 2.98) in the top decile. **c)** Association of the lung cancer
654 PRS with lung cancer risk in ever-smokers with no history of lung cancer. The top PRS
655 decile was associated with an OR of 1.25 (1.18, 1.32).

656

657 **Supplementary Fig. 8. Manhattan plots and quantile-quantile (QQ) plots for**
658 **European meta-analyses conditioned on cigarettes per day.** Manhattan and QQ
659 plots for **a)** overall lung cancer conditioned on cigarettes per day; **b)** lung
660 adenocarcinoma (LUAD) conditioned on cigarettes per day; and **c)** squamous cell lung
661 carcinoma (LUSC) conditioned on cigarettes per day. Cytoband positions for significant
662 loci are noted in each Manhattan plot; novel loci not identified in the European meta-
663 analysis are in red; externally replicated novel loci are indicated by a box. Genomic
664 control (λ) values, LDSC intercepts, and sample sizes are inset in QQ plots.

665

666 **Supplementary Fig. 9. Novel loci for overall lung cancer and squamous cell**
667 **carcinoma conditioned on smoking.** Forest plots (left) and regional Manhattan plots
668 (right) for novel loci identified in the European meta-analysis conditioned on cigarettes
669 per day: a) *MMS22L* in overall lung cancer and b) *ABHD8* in squamous cell lung cancer.

670

671

672

673

674

675

676

677

678

679

680

681 **References**

- 682 1. Schabath, M. B. & Cote, M. L. Cancer Progress and Priorities: Lung Cancer. *Cancer*
683 *Epidemiol. Biomarkers Prev.* **28**, 1563–1579 (2019).
- 684 2. Leiter, A., Veluswamy, R. R. & Wisnivesky, J. P. The global burden of lung cancer: current
685 status and future trends. *Nat. Rev. Clin. Oncol.* **20**, 624–639 (2023).
- 686 3. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and
687 Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249
688 (2021).
- 689 4. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer statistics, 2022. *CA Cancer J.*
690 *Clin.* **72**, 7–33 (2022).
- 691 5. Bossé, Y. & Amos, C. I. A Decade of GWAS Results in Lung Cancer. *Cancer Epidemiol.*
692 *Biomarkers Prev.* **27**, 363–379 (2018).
- 693 6. Timofeeva, M. N. *et al.* Influence of common genetic variation on lung cancer risk: meta-
694 analysis of 14 900 cases and 29 485 controls. *Hum. Mol. Genet.* **21**, 4980–4995 (2012).
- 695 7. McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility
696 loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat. Genet.* **49**,
697 1126–1132 (2017).
- 698 8. Byun, J. *et al.* Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237
699 controls identifies new susceptibility loci contributing to lung cancer. *Nat. Genet.* **54**, 1167–
700 1177 (2022).
- 701 9. Wang, Y. *et al.* SNP rs17079281 decreases lung cancer risk through creating an YY1-
702 binding site to suppress DCBLD1 expression. *Oncogene* **39**, 4092–4102 (2020).
- 703 10. Zhang, T. *et al.* Genomic and evolutionary classification of lung cancer in never smokers.
704 *Nat. Genet.* **53**, 1348–1359 (2021).
- 705 11. Govindan, R. *et al.* Genomic landscape of non-small cell lung cancer in smokers and never-

- 706 smokers. *Cell* **150**, 1121–1134 (2012).
- 707 12. Wang, Z. *et al.* Meta-analysis of genome-wide association studies identifies multiple lung
708 cancer susceptibility loci in never-smoking Asian women. *Hum. Mol. Genet.* **25**, 620–629
709 (2016).
- 710 13. Schabath, M. B., Cress, D. & Munoz-Antonia, T. Racial and Ethnic Differences in the
711 Epidemiology and Genomics of Lung Cancer. *Cancer Control* **23**, 338–346 (2016).
- 712 14. Long, E., Patel, H., Byun, J., Amos, C. I. & Choi, J. Functional studies of lung cancer
713 GWAS beyond association. *Hum. Mol. Genet.* **31**, R22–R36 (2022).
- 714 15. Shi, J. *et al.* Genome-wide association study of lung adenocarcinoma in East Asia and
715 comparison with a European population. *Nat. Commun.* **14**, 3043 (2023).
- 716 16. Dai, J. *et al.* Identification of risk loci and a polygenic risk score for lung cancer: a large-
717 scale prospective cohort study in Chinese populations. *Lancet Respir Med* **7**, 881–891
718 (2019).
- 719 17. Nahar, R. *et al.* Elucidating the genomic architecture of Asian EGFR-mutant lung
720 adenocarcinoma through multi-region exome sequencing. *Nat. Commun.* **9**, 216 (2018).
- 721 18. Zanetti, K. A. *et al.* Genome-wide association study confirms lung cancer susceptibility loci
722 on chromosomes 5p15 and 15q25 in an African-American population. *Lung Cancer* **98**, 33–
723 42 (2016).
- 724 19. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences
725 on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).
- 726 20. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human
727 tissues. *Science* **369**, 1318–1330 (2020).
- 728 21. Hao, K. *et al.* Lung eQTLs to help reveal the molecular underpinnings of asthma. *PLoS*
729 *Genet.* **8**, e1003029 (2012).
- 730 22. Bossé, Y. *et al.* Transcriptome-wide association study reveals candidate causal genes for
731 lung cancer. *Int. J. Cancer* **146**, 1862–1878 (2020).

- 732 23. Koutsami, M. K. *et al.* Centrosome abnormalities are frequently observed in non-small-cell
733 lung cancer and are associated with aneuploidy and cyclin E overexpression. *J. Pathol.*
734 **209**, 512–521 (2006).
- 735 24. Chan, J. Y. A clinical overview of centrosome amplification in human cancers. *Int. J. Biol.*
736 *Sci.* **7**, 1122–1144 (2011).
- 737 25. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable
738 selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B*
739 *Stat. Methodol.* **82**, 1273–1300 (2020).
- 740 26. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with
741 the ‘Sum of Single Effects’ model. *PLoS Genet.* **18**, e1010299 (2022).
- 742 27. de Goede, O. M. *et al.* Population-scale tissue transcriptomics maps long non-coding RNAs
743 to complex disease. *Cell* **184**, 2633–2648.e19 (2021).
- 744 28. Li, Y. *et al.* Pan-cancer characterization of immune-related lncRNAs identifies potential
745 oncogenic biomarkers. *Nat. Commun.* **11**, 1000 (2020).
- 746 29. de Santiago, P. R. *et al.* Immune-related lncRNA LINC00944 responds to variations in
747 ADAR1 levels and it is associated with breast cancer prognosis. *Life Sci.* **268**, 118956
748 (2021).
- 749 30. Chen, D. *et al.* Genome-wide analysis of long noncoding RNA (lncRNA) expression in
750 colorectal cancer tissues from patients with liver metastasis. *Cancer Med.* **5**, 1629–1639
751 (2016).
- 752 31. McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol.* **17**, 122 (2016).
- 753 32. Saunders, G. R. B. *et al.* Genetic diversity fuels gene discovery for tobacco and alcohol
754 use. *Nature* **612**, 720–724 (2022).
- 755 33. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-
756 analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
- 757 34. Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the

- 758 genetic etiology of tobacco and alcohol use. *Nat. Genet.* **51**, 237–244 (2019).
- 759 35. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred from
760 GWAS summary data. *Nat. Commun.* **9**, 1–12 (2018).
- 761 36. Xue, A. *et al.* Genome-wide analyses of behavioural traits are subject to bias by misreports
762 and longitudinal changes. *Nat. Commun.* **12**, 20211 (2021).
- 763 37. Munafò, M. R., Tilling, K., Taylor, A. E., Evans, D. M. & Davey Smith, G. Collider scope:
764 when selection bias can substantially influence observed associations. *Int. J. Epidemiol.* **47**,
765 226–235 (2018).
- 766 38. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity
767 in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
- 768 39. Nguyen, M.-H., Ueda, K., Nakamura, Y. & Daigo, Y. Identification of a novel oncogene,
769 MMS22L, involved in lung and esophageal carcinogenesis. *Int. J. Oncol.* **41**, 1285–1296
770 (2012).
- 771 40. Yang, W. *et al.* Deciphering associations between three RNA splicing-related genetic
772 variants and lung cancer risk. *NPJ Precis Oncol* **6**, 48 (2022).
- 773 41. Gabriel, A. A. G. *et al.* Genetic Analysis of Lung Cancer and the Germline Impact on
774 Somatic Mutation Burden. *J. Natl. Cancer Inst.* **114**, 1159–1166 (2022).
- 775 42. Lawrenson, K. *et al.* Functional mechanisms underlying pleiotropic risk alleles at the
776 19p13.1 breast-ovarian cancer susceptibility locus. *Nat. Commun.* **7**, 12675 (2016).
- 777 43. Lesseur, C. *et al.* Genome-wide association meta-analysis identifies pleiotropic risk loci for
778 aerodigestive squamous cell cancers. *PLoS Genet.* **17**, e1009254 (2021).
- 779 44. Fachal, L. *et al.* Fine-mapping of 150 breast cancer risk regions identifies 191 likely target
780 genes. *Nat. Genet.* **52**, 56–73 (2020).
- 781 45. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using
782 MTAG. *Nat. Genet.* **50**, 229–237 (2018).
- 783 46. Michailidou, K. *et al.* Association analysis identifies 65 new breast cancer risk loci. *Nature*

- 784 **551**, 92–94 (2017).
- 785 47. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal
786 variants. *PLoS Genet.* **17**, e1009440 (2021).
- 787 48. Gusev, A. *et al.* A transcriptome-wide association study of high-grade serous epithelial
788 ovarian cancer identifies new susceptibility genes and splice variants. *Nat. Genet.* **51**, 815–
789 823 (2019).
- 790 49. Brenner, D. R. *et al.* Alcohol consumption and lung cancer risk: A pooled analysis from the
791 International Lung Cancer Consortium and the SYNERGY study. *Cancer Epidemiol.* **58**,
792 25–32 (2019).
- 793 50. Larsson, S. C. *et al.* Smoking, alcohol consumption, and cancer: A mendelian
794 randomisation study in UK Biobank and international genetic consortia participants. *PLoS*
795 *Med.* **17**, e1003178 (2020).
- 796 51. Petrelli, F. *et al.* Association of Obesity With Survival Outcomes in Patients With Cancer: A
797 Systematic Review and Meta-analysis. *JAMA Netw Open* **4**, e213520 (2021).
- 798 52. Lan, T., Chen, L. & Wei, X. Inflammatory Cytokines in Cancer: Comprehensive
799 Understanding and Clinical Progress in Gene Therapy. *Cells* **10**, (2021).
- 800 53. Kresse, S. H. *et al.* LSAMP, a novel candidate tumor suppressor gene in human
801 osteosarcomas, identified by array comparative genomic hybridization. *Genes*
802 *Chromosomes Cancer* **48**, 679–693 (2009).
- 803 54. Xie, J. *et al.* Copy number analysis identifies tumor suppressive lncRNAs in human
804 osteosarcoma. *Int. J. Oncol.* **50**, 863–872 (2017).
- 805 55. Yu, G. *et al.* Roles of transducin-like enhancer of split (TLE) family proteins in
806 tumorigenesis and immune regulation. *Front Cell Dev Biol* **10**, 1010639 (2022).
- 807 56. Sartor, I. T. S., Recamonde-Mendoza, M. & Ashton-Prolla, P. TULP3: A potential biomarker
808 in colorectal cancer? *PLoS One* **14**, e0210762 (2019).
- 809 57. Chaudhary, P. K. & Kim, S. An Insight into GPCR and G-Proteins as Cancer Drivers. *Cells*

- 810 **10**, (2021).
- 811 58. Murphy, C. *et al.* An Analysis of JADE2 in Non-Small Cell Lung Cancer (NSCLC).
812 *Biomedicines* **11**, (2023).
- 813 59. Ni, L. *et al.* RPAP3 interacts with Reptin to regulate UV-induced phosphorylation of H2AX
814 and DNA damage. *J. Cell. Biochem.* **106**, 920–928 (2009).
- 815 60. Saredi, G. *et al.* H4K20me0 marks post-replicative chromatin and recruits the TONSL–
816 MMS22L DNA repair complex. *Nature* **534**, 714–718 (2016).
- 817 61. Fang, H. *et al.* Harmonizing Genetic Ancestry and Self-identified Race/Ethnicity in Genome-
818 wide Association Studies. *Am. J. Hum. Genet.* **105**, 763–772 (2019).
- 819 62. Hunter-Zinck, H. *et al.* Genotyping Array Design and Data Quality Control in the Million
820 Veteran Program. *Am. J. Hum. Genet.* **106**, 535–548 (2020).
- 821 63. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer
822 datasets. *Gigascience* **4**, 7 (2015).
- 823 64. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium
824 panel. *Nat. Genet.* **48**, 1443–1448 (2016).
- 825 65. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**,
826 1284–1287 (2016).
- 827 66. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation.
828 *Nature* **526**, 68–74 (2015).
- 829 67. Mbatchou, J. *et al.* Computationally efficient whole-genome regression for quantitative and
830 binary traits. *Nat. Genet.* **53**, 1097–1103 (2021).
- 831 68. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of
832 genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
- 833 69. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat.*
834 *Genet.* **48**, 1279–1283 (2016).
- 835 70. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data

- 836 using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
- 837 71. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations.
838 *Bioinformatics* **28**, 1353–1358 (2012).
- 839 72. Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL
840 mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479–1485 (2016).
- 841 73. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-
842 wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
- 843 74. Amos, C. I. *et al.* The OncoArray Consortium: A Network for Understanding the Genetic
844 Architecture of Common Cancers. *Cancer Epidemiol. Biomarkers Prev.* **26**, 126–135
845 (2017).
- 846 75. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and
847 annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
- 848 76. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via
849 Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
- 850 77. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide
851 complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
- 852 78. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic
853 association studies using summary statistics. *PLoS Genet.* **10**, e1004383 (2014).
- 854 79. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-
855 specific haplotype structure and linking correlated alleles of possible functional variants.
856 *Bioinformatics* **31**, 3555–3557 (2015).
- 857 80. Carroll, R. J., Bastarache, L. & Denny, J. C. R PheWAS: data analysis and plotting tools for
858 phenome-wide association studies in the R environment. *Bioinformatics* **30**, 2375–2376
859 (2014).
- 860 81. Klarin, D. *et al.* Genetics of blood lipids among ~300,000 multi-ethnic participants of the
861 Million Veteran Program. *Nat. Genet.* **50**, 1514–1523 (2018).

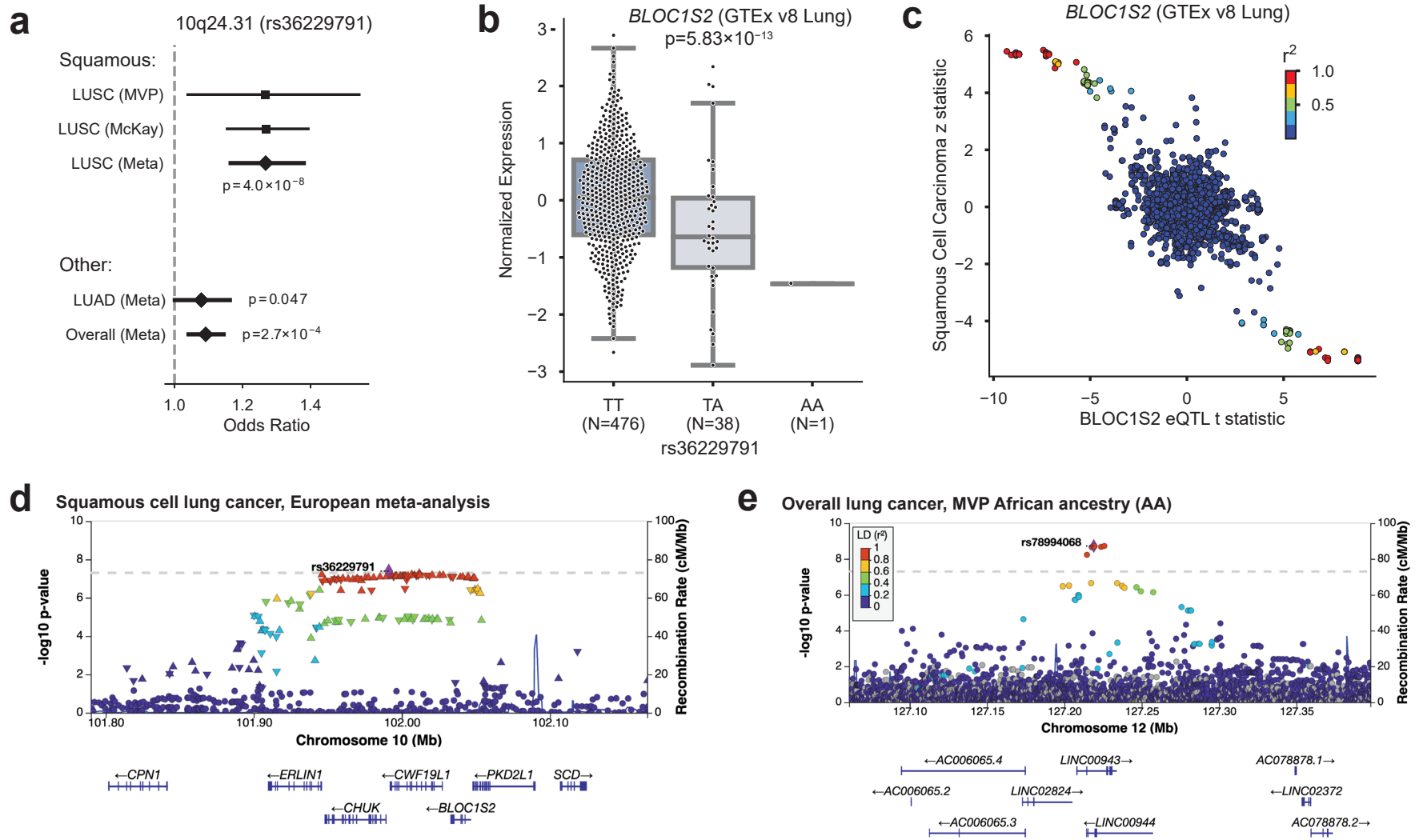


Figure 1. Highlighted novel GWAS loci. a-d) The meta-analysis of squamous cell lung carcinoma (LUSC) in European ancestry (EA) identifies a novel locus at 10q24.31. **a)** Odds ratios for rs36229791 in LUSC compared to lung adenocarcinoma (LUAD) and overall lung cancer. **b)** *BLOC1S2* expression varies by genotype at rs36229791. **c)** *BLOC1S2* eQTL t statistic vs LUSC z statistic. **d)** Regional association plot showing SNP significance and genes around lead SNP rs36229791. **e)** The African ancestry GWAS highlights a putatively novel locus on chr12 at *LINC00943/LINC00944*. The risk allele has effectively 0% frequency in EA.

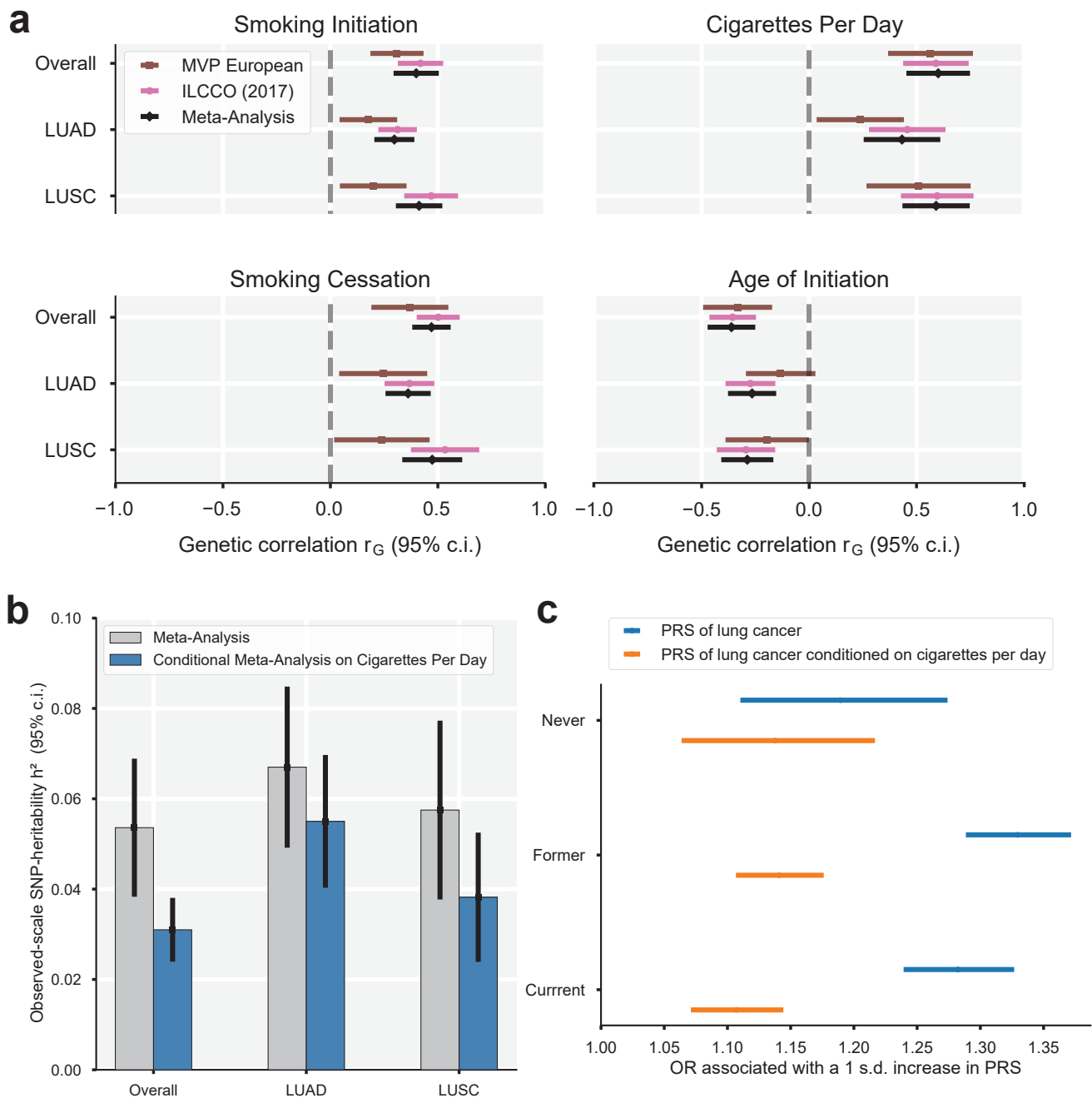


Figure 2. Association of lung cancer GWAS with smoking behaviors. **a)** Genetic correlations (with 95% confidence interval) between the lung cancer GWAS and smoking behaviors, including smoking initiation, cigarettes per day, smoking cessation, and age of initiation. **b)** SNP heritability for the meta-analysis and conditional meta-analysis. The heritability decreases in the conditional analysis for overall lung cancer as well as both subtypes, suggesting that some portion of the heritability of lung cancer is due to smoking behavior. **c)** Polygenic risk scores (PRS) based on standard lung cancer GWAS (blue) performs worse in never-smokers than former or current smokers, while conditioning on smoking behavior (orange) results in similar performance.

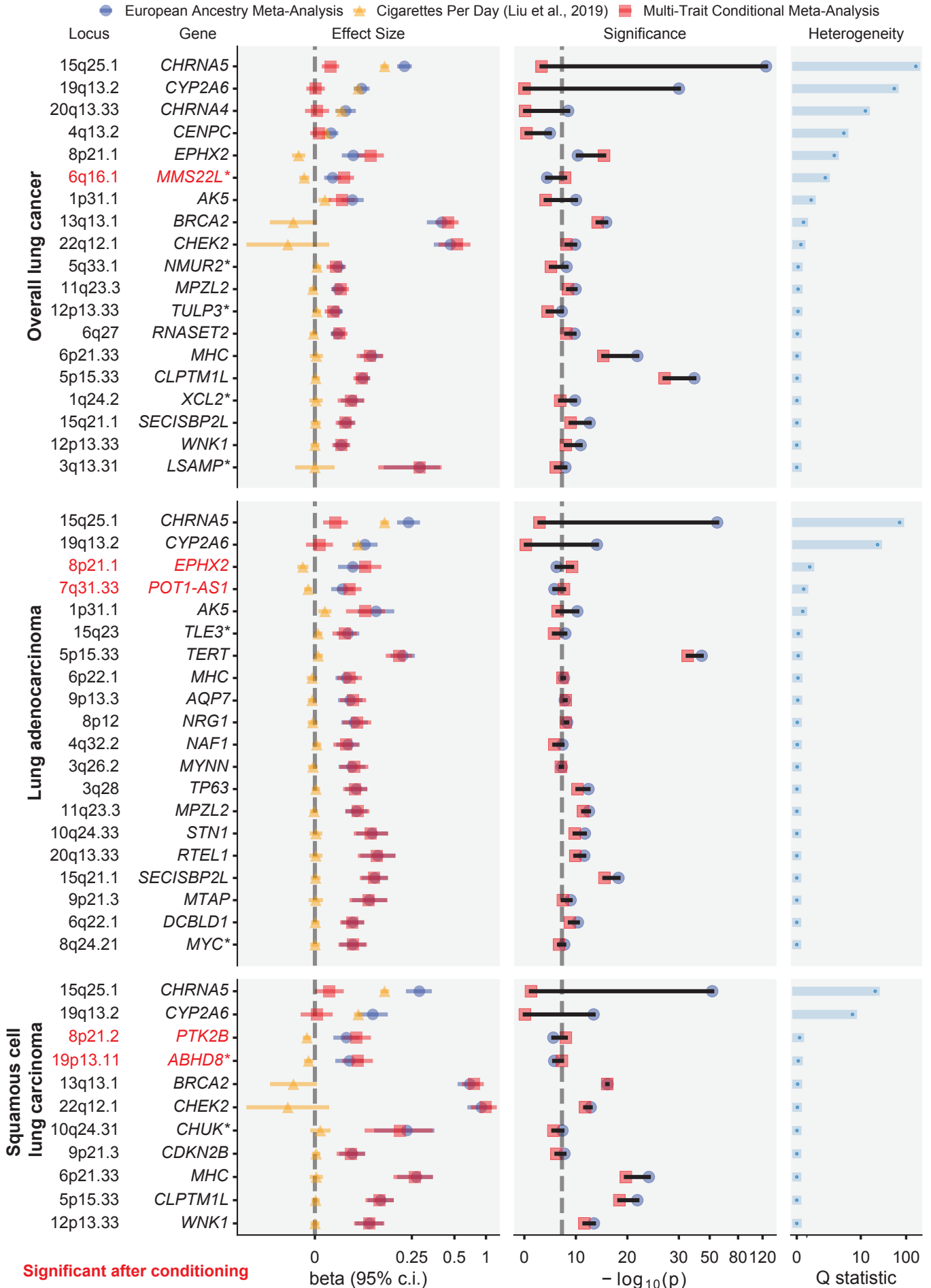


Figure 3. Forest plot of genome-wide significant associations. Within each cancer subtype, changes in effect size and significance are shown before and after conditioning on cigarettes per day. Novel loci are indicated by an asterisk after the gene name (*). Loci that became significant after conditioning ($P < 5 \times 10^{-8}$) are in red.

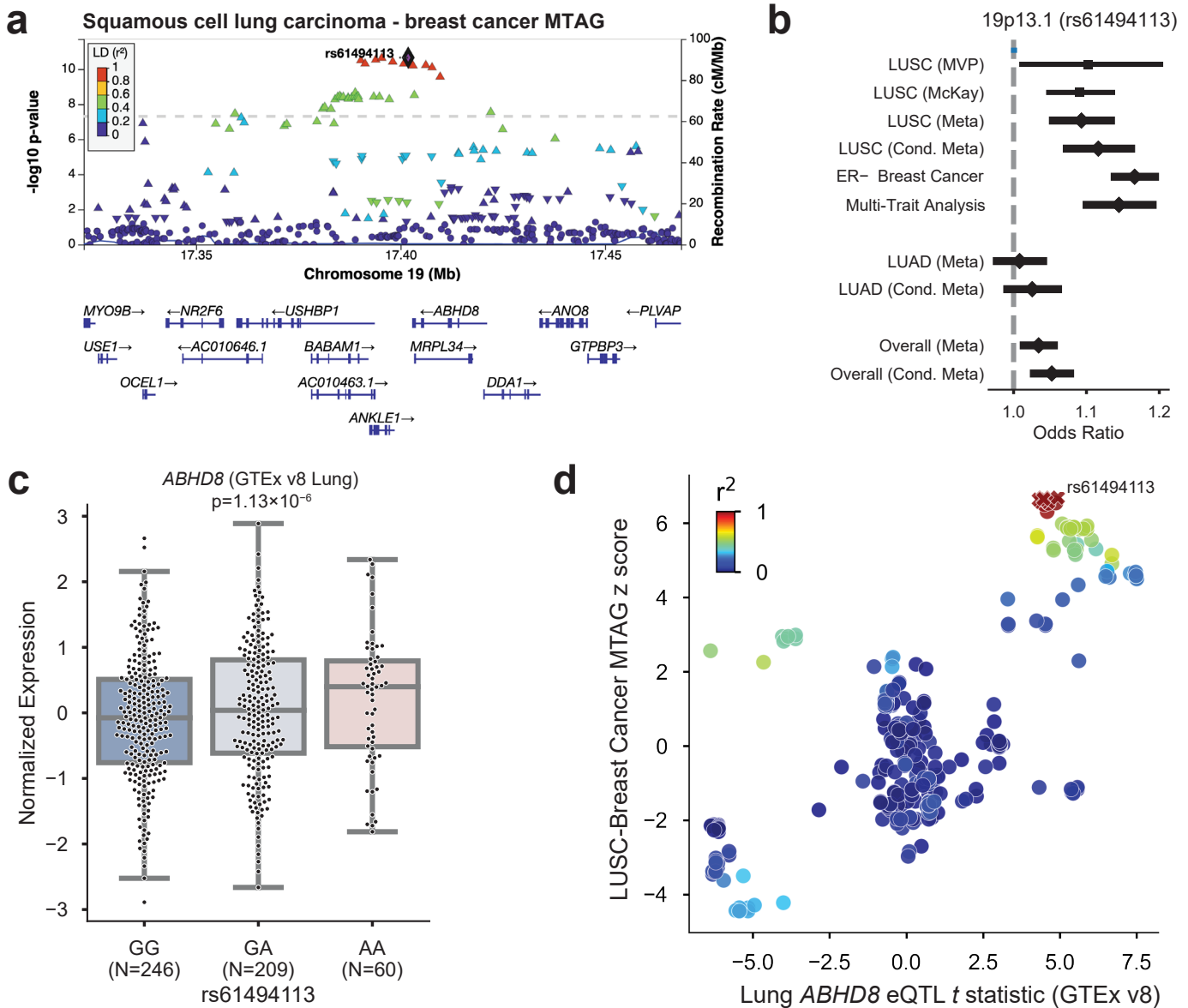


Figure 4. Significant locus after conditioning on smoking behavior, 19p13.11, has pleiotropic associations with ER-negative breast cancer. **a)** Regional association plot of the 19p13.11 multi-trait analysis of GWAS (MTAG) locus. **b)** Odds ratios for lead SNP rs61494113 in squamous cell lung carcinoma (LUSC), before and after conditioning on cigarettes per day, and MTAG analysis, compared to lung adenocarcinoma and overall lung cancer. **c)** *ABHD8* expression varies by genotype at rs61494113. **d)** *ABHD8* eQTL t statistic vs LUSC z statistic; red X's indicate the 95% credible set.

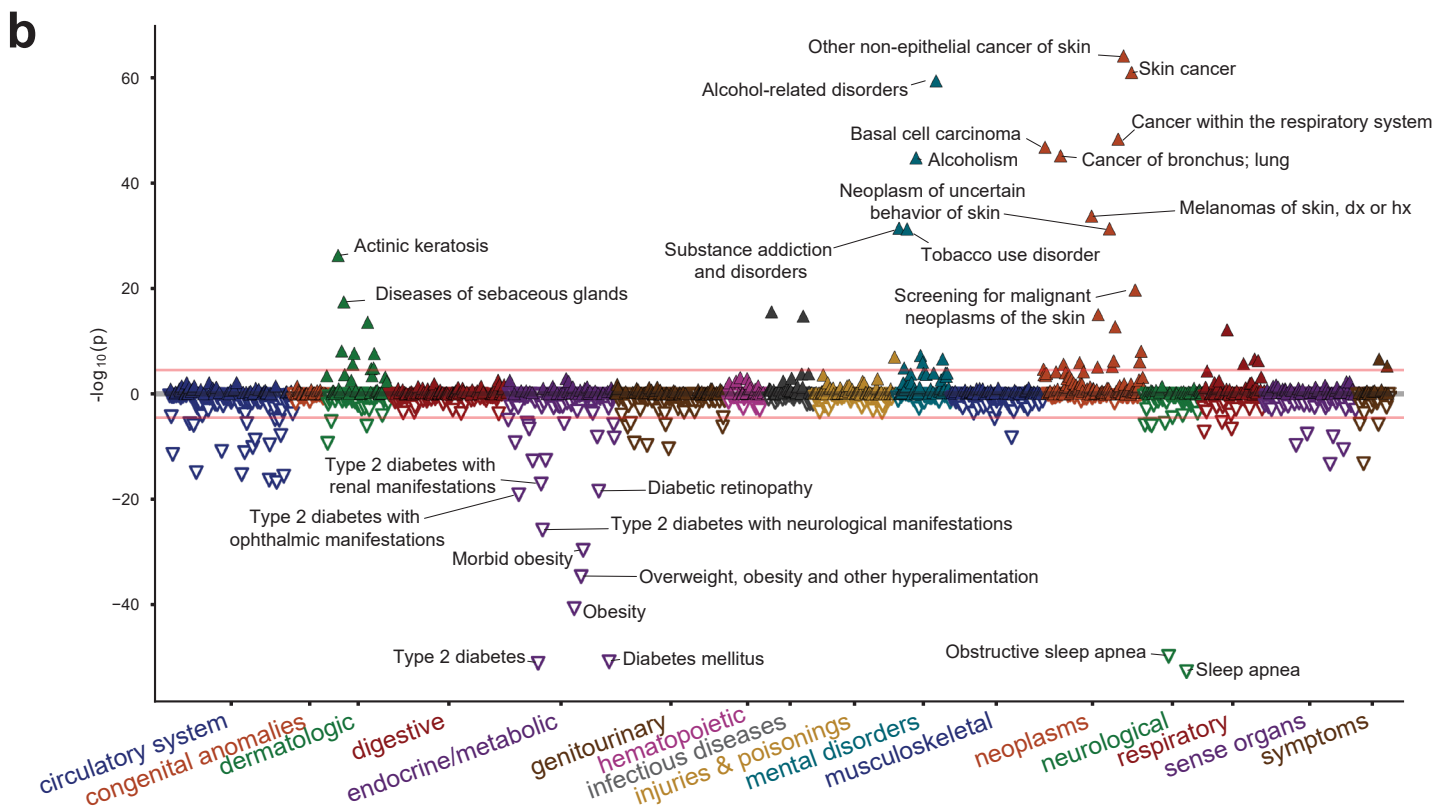
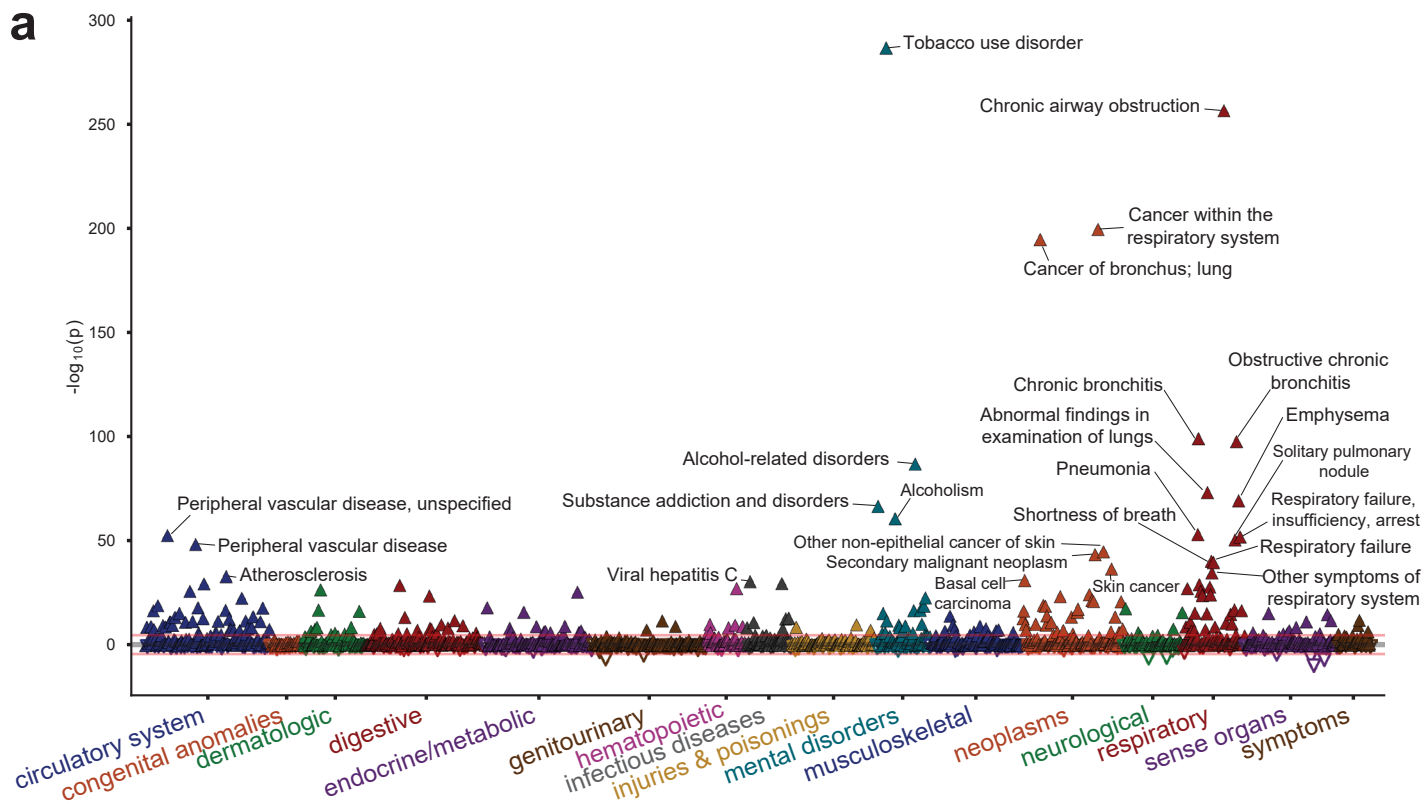


Figure 5. Phenome-wide association study (PheWAS) of polygenic risk scores (PRS) of lung cancer and lung cancer conditioned on cigarettes per day. a) PheWAS of PRS on lung cancer is mostly confounded with smoking associations. b) PheWAS of the conditional meta-analysis PRS shows associations with skin cancer and metabolic traits.