

## Protein-altering germline mutations implicate novel genes related to lung cancer development

### Abstract

Germline mutations cause 3–10% of cancers diagnosed yearly<sup>1</sup>. Identifying cancer-related mutations can provide targets for personalized cancer screening, prevention<sup>2,3</sup>, treatment and drug development<sup>4</sup>. However, only a few germline mutations responsible for lung cancer etiology have been identified<sup>5</sup>. Here, by performing whole-exome analyses of mutations in two independent cohorts with 39,146 individuals of European ancestry and investigating gene expression levels in 7,773 samples, we found large-effect associations for lung adenocarcinoma with the L2307F mutation of *ATM* for lung cancer with a stop-gain mutation (Q4X) of a previously uncharacterized gene *KIAA0930*, which gene was found to be significantly over-expressed in lung cancer and most carcinomas in our study. All *ATM*-L2307F homozygotes had lung adenocarcinoma ( $P=0.004$ ) and all *KIAA0930*-Q4X homozygotes had lung cancer ( $P=2.29 \times 10^{-8}$ ). Despite being very rare in the general European population, L2307F was much more common in the Israeli population (MAF=0.023), especially in Ashkenazi Jews (47/640). L2307F was predicted to enhance ATM protein aggregation and is predicted to have a pathogenic effect on the ATM protein, but has not previously been associated with lung cancer or ataxia telangiectasia. *KIAA0930*-Q4X completely abolishes *KIAA0930* transcription. Our results demonstrate *ATM*-L2307F and *KIAA0930*-Q4X increased lung cancer risk with large effects among heterozygotes and high risk for developing lung cancer among homozygotes indicating the importance of *ATM* and *KIAA0930* for lung cancer etiology. Selected rare mutations in *ATM* have also been identified through the TCGA as associating with lung adenocarcinoma risk. Individuals with mutations of *ATM* and *KIAA0930* may benefit from targeted screening and the findings may suggest targets for treatment of lung cancer in selected patients. Our results offer novel genes to include in cancer panels for effective lung cancer screening and suggest new targets to advance our understanding of cancer pathogenesis.

## Main

Lung cancer is a leading cause of cancer death in the U.S. and around the world and represents a major public health problem<sup>6</sup>. Hereditary factors also play a crucial role in lung cancer etiology with an estimated 18% of lung cancer risk attributed to hereditary factors<sup>7</sup>. The first wave of genome-wide association studies identified susceptibility regions and common variants for lung cancer risk. However, few previous studies investigated the association between germline mutations and lung cancer risk because this type of research requires large sample sizes and high-coverage genome sequencing or large scale genotyping. Although only less than 1% of most populations are carriers of a germline mutation that drive cancers, these mutations may confer as much as an 80% lifetime risk for developing cancer<sup>8</sup> and such mutations cause between 3–10% of cancers diagnosed yearly<sup>1</sup>. In addition, identification of cancer-related mutations provided potential targets for cancer treatment and drug development. The rare inherited T790M mutation of EGFR is associated with greatly increased risk for lung cancer in nonsmokers<sup>9</sup>. Individuals with this mutation do not respond well to first-line EGFR therapy<sup>10</sup> and a targeted approach to risk management and therapy is required for individuals carrying this mutation<sup>11</sup>. Similarly, identification of mutations in *BRCA1* and *BRCA2* led to the development of PARP inhibition therapy for breast cancer. Defining driver germline mutations for lung cancer may also help with targeted prevention and early detection, similar to the benefit conveyed in screening for harmful *BRCA1* and *BRCA2* germline mutations<sup>2,3</sup>. Therefore, we investigated germline mutations with a large effect on lung cancer pathogenesis to offer insights into understanding cancer mechanisms.

To discover driver germline mutations with high effect on lung cancer risk, we performed association analyses for all the mutations within the exome using a discovery cohort of 28,878 individuals of European ancestry (**Supplementary Table 1**). Three mutations, including rs56009889, rs150665432, and rs61816761 had association *P* values of less than  $5.0 \times 10^{-8}$  and OR values of more than 2.0 in the discovery cohort (**Supplementary Table 2**).

We then validated these findings in an independent cohort consisting of 10,268 individuals of European ancestry, in which repeat genotyping was performed for 5742 subjects along with the discovery cohort to investigate the genotyping fidelity of the mutations for both cohorts. The variants rs56009889 and rs150665432 had excellent concordances of 99.95% and 99.08% for overall genotypes and 89.66% and 92.31% for the rare alleles, respectively, which confirmed their genotyping fidelity in both cohorts. However, genotyping for rs61816761 showed poor concordance, so it was dropped from further analysis (**Supplementary Table 3**). Additionally, among unaffected individuals, the Minor Allele Frequencies (MAFs) of rs56009889 and rs150665432 in both cohorts were comparable to the MAFs found for European populations in the Exome Aggregation Consortium<sup>12</sup>, and were in agreement with MAFs in the NHLBI GO Exome Sequencing Project and the Trans-Omics for Precision Medicine Program<sup>13,14</sup> (**Supplementary Table 4**), which supported the reliability of the genotyping data of rs56009889 and rs150665432 in both cohorts. Therefore, we further investigated rs56009889 and rs150665432.

Although rs56009889, mapping within *ATM* gene (**Fig. 1a**) is not reported as causing Ataxia Telangiectasia<sup>15</sup>, its mutation results in L2307F missense mutation in the FAT domain which regulates ATM activity<sup>16</sup> and thereby possibly affects its function (**Supplementary Fig. 1a**). Compared to non-carriers (C/C), L2307F carriers (T/C + T/T) had an increased risk of lung cancer with statistical significance in the discovery cohort (adjusted odds ratios (ORs)=4.19,  $P=3.56 \times 10^{-7}$ ), though the increased risk did not reach significance in the replication cohort (**Table 1**). Among females, L2307F was significantly associated with lung cancer risk with ORs being 7.76 ( $P=0.0002$ ) in the discovery cohort and 3.22 ( $P=0.03$ ) in the replication cohorts. Among males, L2307F showed a weakly significant association with lung cancer risk in the discovery cohort and no association in the replication cohort (**Fig. 2a** and **Supplementary Table 5**). Stratification analysis by histology indicated that L2307F carriers had a significant 5.23-fold increased risk for lung adenocarcinoma (LAD) in the discovery cohort ( $P=6.47 \times 10^{-9}$ ) and a 2.48-fold increased risk in the replication cohorts ( $P=0.01$ ), and exhibited no association with the risk of lung squamous cell carcinoma (LSQ) or of small cell lung cancer (SCLC) in either cohort (**Fig. 2a** and **Supplementary Table 5**). Females

carrying L2307F showed an 8.05-fold ( $P=0.0001$ ), 4.69-fold ( $P=0.004$ ) and 6.10-fold ( $P=2.14\times 10^{-6}$ ) greater risk of LAD in the discovery, replication and meta-analysis, respectively (**Fig. 2b** and **Supplementary Table 6**). All the L2307F homozygotes, no matter what age, gender, and smoking status, developed LAD in this study ( $P=0.004$ ) (**Fig. 2c**). Moreover, the association exhibited a dose-response relationship between the number of mutated alleles and the LAD risk in the discovery cohort ( $P_{\text{trend}}=5.44\times 10^{-9}$ ). A more significant role for L2307F in LAD etiology than in LSQ or SCLC is reflected in the no observed mutational frequency in LSQ and SCLC in the replication cohort. These results suggested the association between rs56009889 and lung cancer risk was restricted to LAD, especially in women. Although previous studies have failed to demonstrate *ATM*-L2307F affects lung cancer risk, since ATM protein is responsible for repairing the broken DNA strands and maintaining the stability of genes<sup>17</sup>, not surprisingly there is variability in genetic effects on lung cancer risk by histology with subtype-specific associations at *ATM*-L2307F.

Interestingly, the *ATM*-L2307F was found in 4.43% (MAF=0.023) individuals from Israel, especially in 7.34% (47/640) of Ashkenazi Jews, even though rs56009889 was almost monomorphic in other European countries (**Supplementary Table 7**). Concordant with such results is the finding that Ashkenazi Jews also had high prevalence of harmful BRCA founder mutations<sup>18</sup>. In addition, North Americans were observed to have a slightly higher prevalence for this specific harmful *ATM*-L2307F (MAF=0.002) than general Europeans. We therefore investigated if the association of rs56009889 and lung cancer risk was affected by country of origin. In both Israeli and North Americans, rs56009889 was significantly associated with the risk of lung cancer, of LAD in general and in women, respectively, in the discovery cohort. However, the association was substantially stronger and higher in the Israeli population than in North Americans (**Fig. 2d**). The ORs for LAD risk among L2307F carriers were 3.36 in North Americans ( $P=0.004$ ) and 6.74 in the Israeli population ( $P=3.85\times 10^{-6}$ ). The female carriage of L2307F conferred LAD risk with an OR of 3.81 in North Americans ( $P=0.04$ ) and 17.15 for the Israeli ( $P=0.006$ ). There were no Israeli samples

available in the replication data. Our results by geographic populations validated the association and significance of L2307F for LAD, especially in women.

rs150665432, a premature stop codon, is mapped within *KIAA0930* (**Fig. 1b**) locating at 22q13.31 that is associated with 22q13 deletion syndrome and lung cancer risk<sup>19,20</sup>. Mutation in rs150665432 is responsible for Q4X and truncate the protein length from 409 to 3 amino acids<sup>21</sup> (**Supplementary Fig. 1b**). Compared to non-carriers (G/G), Q4X carriers (A/G+A/A) had an increased lung cancer risk in both discovery (adjusted OR=2.59;  $P=1.15 \times 10^{-18}$ ) and replication cohorts (adjusted OR=1.69;  $P=0.03$ ) (**Fig. 3a**). Additionally, all Q4X homozygotes developed lung cancer ( $P=2.29 \times 10^{-8}$ ) (**Fig. 3b**), and the number of mutated alleles showed a dose-response relationship with lung cancer risk ( $P_{\text{trend}}=1.51 \times 10^{-19}$ ) in the discovery cohort (**Table 1**). No homozygotes were found in the replication cohort. Stratification analysis showed that Q4X had a significant risk among females, males, smokers, non-smokers (**Supplementary Table 8**), and of LAD, LSQ and SCLC (**Supplementary Table 9**) in the discovery cohort and in the meta-analysis. In the replication, none of the strata reached significance, reflecting the small number of individuals with this uncommon variant in the subset analyses. Although the frequency of rs150665432 in controls varies non-significantly between geographic populations, ORs of the association between Q4X and lung cancer risk were higher in North American Countries (adjusted OR=4.19;  $P=3.27 \times 10^{-16}$ ) than in European Countries (adjusted OR=1.65;  $P=0.0003$ , **Supplementary Table 10**). Although no previous study demonstrated Q4X affected the risk of any disease, the possible involvement of Q4X in the complex genetics of a severe congenital heart disease<sup>22</sup> may indicate its pathogenicity. Our findings raise the possibility *KIAA0930*-Q4X may have a direct effect on lung cancer risk.

We also investigated whether the mutations were associated with the onset age of lung cancer and found that mean onset age for female LAD cases harboring ATM-L2307F was later than that for cases among non-carriers with significance in the discovery cohort ( $69.37 \pm 10.71$  vs  $62.78 \pm 11.06$ ,  $P=0.0007$ ) and borderline significance because of small number of individuals in the replication cohort ( $68.74 \pm 10.49$  vs

63.69±10.31,  $P=0.09$ , **Supplementary Table 11** and **Fig. 2e**). The difference of age for lung cancer in general and other strata was not validated in two cohorts. We also found significant differences of onset age between various genotypes was shown in cases among North Americans, but not among Israeli or in controls in the discovery data (**Supplementary Table 12 and 13**). A significant multiplicative interaction on the onset age was also found between rs56009889 and countries ( $P=0.01$ ), and rs56009889 was associated with onset age adjusted for countries ( $P<0.0001$ ) in the discovery cohort, supporting L2307F affected onset age of lung cancer. Although it is counter-intuitive that L2307F may be associated with late-onset LAD in females because germline mutations are associated with early-onset cancers in general, concordant with our results was that TP53 germline mutations can be related to late-onset common cancers<sup>23</sup> and that familial cancers can exist in cancers of advanced ages<sup>24</sup>. A possible explanation for our observations is that *ATM* previously is more active at elderly age<sup>25</sup>. No consistent variation in age by genotypes of rs150665432 was found overall or in subgroups (**Supplementary Table 14**).

In order to verify the effect of the mutations on lung cancer risk, we also performed structure-based prediction to determine its functional significance. Using SNPeffect 4.0<sup>26</sup>, we found that L2307F had a 2.2-fold greater TANGO score, which defined protein aggregation based on the physics-chemical principles of  $\beta$ -sheet formation<sup>27</sup>, than the wildtype protein, suggesting L2307F contributed to protein aggregation. By applying PolyPhen-2<sup>28</sup>, L2307F was categorized as being deleterious with a probability of 0.98 and a false discovery rate of 0.09, implying it damaged the stability and function of ATM protein<sup>29</sup>. Although it has not been biochemically characterized, Fathmm-XF<sup>30</sup> labeled L2307F ascribed to pathogenic mutations<sup>31</sup>. Therefore, it is possible that L2307F could influence cancer risk. Q4X truncates KIAA0930 to 3 amino acids<sup>21</sup> and therefore completely abolishes KIAA0930 protein and destroys its function.

Because KIAA0930 is an uncharacterized protein<sup>32</sup>, we elucidated its role in lung cancer pathogenesis by investigating whether its expression was regulated during lung cancer development, as long as comparing its significance to ATM which is cancer suppressor protein. The surprise is that KIAA0930 was

significantly over-expressed in LAD ( $P=0.004$ ) and LSQ ( $P=1.62 \times 10^{-12}$ ) in TCGA database (**Fig. 3c and 3d**), while ATM was up-regulated with a borderline significance in LAD and LSQ (**Supplementary fig. 2**). Harvard expression data<sup>33</sup> confirmed that KIAA0930 showed significant over-expression in lung cancer than in normal lung samples ( $P=0.0005$ ), while ATM had different significance in exon-Level expression ( $P=0.01$  for exon 2 and  $P=0.15$  exon 5) (**Supplementary Table 15**). Additionally, we observed KIAA0930 expression was significantly upregulated in the majority of carcinomas, but not in other cancer types (**Supplementary Table 16**), meaning that KIAA0930 is a carcinoma-associated candidate gene, which was suggested by the data in The Human Protein Atlas showing that KIAA0930 expression significantly affect survival in patients with carcinomas<sup>34</sup>.

In conclusion, this whole-exome study documents that among ~19,000 genes examined, ATM and KIAA0930 show the strongest evidence to be the candidate gene that were responsible for lung cancer etiology. The most common germline mutation reported in the lung adenocarcinomas in the Cancer Genome Atlas were heterozygous mutations in ATM occurring in aggregate among 1.2% of cases<sup>35</sup>. However, the L2307F mutation was not identified, perhaps because of its rarity in many North American populations. The new cohorts, in which we investigated our mutations, had large sample sizes and have not been previously used to search for uncommon high-risk germline mutations, which improved our ability to identify novel germline mutations and genes contributing to lung cancer. Elements of our study design, such as validating results in two independent cohorts, analyzing the data by geographic populations, confirming the genotyping fidelity, comparing MAFs of the mutations in our cohorts to those in public sequencing datasets and investigating gene expression, ensure the reliability for our results. Notably the identification of the novel lung cancer related germline mutations and genes could greatly advance our understanding of lung cancer etiology.

## **Online Methods**

### **Study subjects**

The OncoArray consortium, which was used in the discovery phase, is a network created to increase understanding of the genetic architecture of common cancers. The Dartmouth component of the Oncoarray consortium included GWAS data from 57,776 samples, obtained from 29 lung cancer studies across North America and Europe, as well as Asia<sup>36,37</sup>, along with additional samples from head and neck cancer patients that were included to improve genotype calling for rare variants. The OncoArray consortium participants who were lacking disease status (because they were not part of the lung cancer related studies), who were close relatives (second-degree relatives or closer) or who were duplicate individuals or other subjects, or who had a low call rate of genotype data, or who did not pass quality control, or who were non-European, were excluded from the current study. There were 5742 participants in the OncoArray consortium who had duplicate samples for the Affymetrix study that was used in the replication phase, and therefore these samples were also excluded from the analysis in the discovery phase. Finally, a total of 28,878 European-descent participants from 26 studies<sup>37</sup>, including 15,851 lung cancer cases and 13,027 healthy controls, were included in the discovery cohort of the current case-control study.

The 26 studies in the current discovery cohort included the Alpha-Tocopherol Beta-Carotene Cancer Prevention Study (ATBC), Canadian screening study (Canada), Cancer de Pulmon en Asturias (CAPUA), Copenhagen lung cancer study (COPENHAGEN), Environment and Genetics in Lung Cancer Study Etiology (EAGLE), The Carotene and Retinol Efficacy Trial (FHCRC), Liverpool Lung Cancer Project (FIELD), German lung cancer study (GERMANY), Harvard Lung Cancer Study (HSPH), The IARC L2 Study (IARC), Israel study (ISRAEL), The Kentucky Lung Cancer Research (KENTUCKY), MD Anderson Cancer Center Study (MDACC), The Malmö Diet and Cancer Study (MDCS), Multiethnic Cohort Study (MEC), New England Lung Cancer Study (NELCS), new samples from Harvard (NEWSAMPLE-600), The Nijmegen Lung Cancer Study (NIJMEGEN), Norway Lung Cancer Study (NORWAY), Northern Sweden Health and Disease Study (NSHDC), The Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial (PLCO), RESOLUCENT study (RESOLUCENT), Tampa Lung Cancer

Study (TAMPA), Total Lung Cancer: Molecular Epidemiology of Lung Cancer Survival (TLC), The Mount-Sinai Hospital-Princess Margaret Study (TORONTO), The Vanderbilt Lung Cancer Study (VANDERBILT). Among the 26 studies, 13 studies, including ATBC, CAPUA, COPENHAGEN, EAGLE, FIELD, IARC, MDCS, NIJMEGEN, NORWAY, ISRAEL, NSHDC, GERMANY and RESOLUCENT, obtained samples from Europe. Another 13 studies, including CANADA, FHCRC, HSPH, KENTUCKY, MDACC, MEC, NELCS, PLCO, TAMPA, TLC, TORONTO, and VANDERBILT, recruited subjects from North America.

We used Affymetrix Axiome array study<sup>38</sup> from the Transdisciplinary Research in Cancer of the Lung consortium in the replication phase. The Affymetrix Axiome array study was a large pooled sample, assembled from 10 independent case-control studies, including Mount-Sinai Hospital-Princess Margaret (MSH-PMH), Multiethnic Cohort, Liverpool Lung Project, Nurses' Health Study and National Physicians Health Study, the European Prospective Investigation into Cancer and Nutrition (EPIC) Lung, the Prostate, Lung and Ovarian Cancer Screening Trial, Carotene and Retinol Efficacy Trial, Russian Multi-Cancer Case-Control Study, Melbourne Collaborative Cohort Study and Harvard Lung Cancer Study<sup>38</sup>. Of the 12651 subjects in the Affymetrix Axiome array study, the participants who were lacking disease status, or who were non-European, or whose samples did not pass quality control, were excluded. Finally, the replication cohort of the current case-control study comprised a total of 10,268 European-descent participants, including 4,916 lung cancer cases and 5,352 healthy controls.

All studies were reviewed and approved by institutional ethics review committees at the involved institutions.

### **Genotyping**

A novel technology developed by Illumina to facilitate efficient genotyping was used to genotype OncoArray samples<sup>37,39</sup>. The lists of genotyped SNPs were designed and generated to include about

230,00 SNPs in a GWAS backbone; a dense set of SNPs within genes that were associated with pharmacogenetic traits relevant to cancer; SNPs within a susceptibility regions for common cancer types (breast, prostate, colon, ovarian or lung); within 1 Mb of known variants relevant to cancer-associated traits, risk variants from epigenetic datasets; variants in genome-wide regulatory profiling data for lung, breast, prostate and ovarian cancers. This process resulted in a total of 568,712 SNPs submitted to Illumina for manufacturing. Finally, 533,631 variants passed quality control procedures and were included as valid markers<sup>39</sup>, of which 105,736 variants were rare variants. Genotyping 395,745 SNPs from samples of the Affymetrix study was performed using a custom Affymetrix Axiom Array (Affymetrix, Santa Clara, CA, USA), which contains a comprehensive panel of key GWAS markers, rare and low-frequency variants and indels<sup>38</sup>.

### **Statistical analyses for demographic characteristics**

To discover the driver germline mutations with high effect on lung cancer risk, we performed association analyses for all the mutations within the exome having Minor Allele Frequencies (MAF) less than 0.01.

Descriptive statistical analyses were conducted to characterize the study population of lung cancer cases and controls in both discovery and replication cohorts. The difference between cases and controls in the distribution of age at diagnosis, gender and smoking status were evaluated using the  $\chi^2$  test. Statistical analyses were performed with Statistical Analysis System software (Version 9.3). Principal component analysis was performed based on GWAS data with the EIGENSTRAT program for both discovery and replication cohorts, respectively. To calculate these principal components, we analyzed GWAS data after excluding the sex chromosomes, variants with MAF less than 0.05 and after sampling SNPs that were uncorrelated with each other.

### **Genotyping concordance between OncoArray genotyping and Affymetrix genotyping**

We confirmed the genotyping fidelity of the selected germline mutations in the OncoArray platform and the Affymetrix platform, respectively, by considering the concordance of these genotypes between the 2 platforms. A total of 5,742 subjects in the OncoArray consortium were duplicate individuals in the Affymetrix data. Even though the 5,742 subjects were excluded in the discovery cohort and included in the replication cohort, we calculated the concordance of genotyping between the OncoArray consortium and the Affymetrix study for the selected germline variants in the 5,742 individuals whose genotyping results were available for both platforms.

The concordance rate was based on the agreement between OncoArray genotyping and Affymetrix genotyping, and we considered the general concordance and also concordance between the rare alleles only<sup>39</sup>. Supplementary Note Table 1 below describes the genotype frequencies in different situations of agreement between OncoArray genotyping and Affymetrix genotyping.

**Supplementary Note Table 1.** Genotype frequencies based on the agreement between OncoArray genotyping and Affymetrix genotyping.

Affymetrix Genotyping	OncoArray Genotyping		
	major/major	major/minor	minor/minor
major/major	a	b	c
major/minor	d	e	f
minor/minor	g	h	i

The general concordance rate was estimated using the genotype frequencies, which were in agreement between OncoArray genotyping and Affymetrix genotyping, incorporating all genotype frequencies (n).

$$\text{General Concordance} = \frac{a + e + i}{n}$$

The concordance of rare alleles was estimated using the genotype frequencies of the minor/minor and major/minor, which were in agreement between Oncoarray genotyping and Affymetrix genotyping, incorporating all genotype frequencies other than the genotype frequency of major/major.

$$\text{Concordance of Rare Allele} = \frac{i + e}{n - a}$$

### **Validation of the reliability of genotyping data**

In order to further validate the reliability of genotyping data, we compared the MAFs of the selected germline mutations in unaffected individuals of the discovery and the replication cohort, respectively, to those in public sequencing projects or datasets including the Exome Aggregation Consortium (ExAC)<sup>12</sup>, the NHLBI GO Exome Sequencing Project (GO-ESP)<sup>40</sup> and the Trans-Omics for Precision Medicine (TOPMed) Program. ExAC is a released public exome sequencing dataset with variations on 60,706 unrelated individuals that included 33,370 individuals from a non-Finnish European population and 3,307 Finnish European population. GO-ESP is an exome sequencing project that included European American and African American participants<sup>40</sup>. TOPMed sequenced the DNA of people from diverse ethnic backgrounds, with 50% being of European descent and 30% of African descent<sup>41</sup>.

### **Association analysis**

Case-control association tests for genotyped data were conducted using 1-degree-of-freedom Cochran-Mantel-Haenszel tests with the application of PLINK version 1.9 to discover the germline mutations with large effects on lung cancer risk. We estimated the association between the risk of lung cancer and the selected germline mutations by computing the odds ratios (ORs) and 95% confidence intervals (CIs) in univariate and multivariate logistic regression analyses in both cohorts. In the multivariate logistic regression model, OR and 95% CI were adjusted by age, gender, smoking status (never and ever) and significant principal components. We further stratified the association of the selected germline mutations and lung cancer risk by gender and smoking status. We also estimated the association between the

selected SNP variants and the risk of lung adenocarcinoma, lung squamous cell carcinoma or small cell lung cancer, respectively, in univariate and multivariate logistic regression analyses.

We further stratified the association in the discovery cohort by geographic populations in univariate and multivariate logistic regression analyses. Based on the MAF of rs56009889 and the location of the study sites in the discovery cohort, we categorized all the studies to 3 subgroups, including Israeli among which rs56009889 had the highest MAF, population in other European countries, and North Americans. Since the frequency of rs150665432 in controls varies non-significantly between geographic populations, we categorized all the studies to 2 subgroups, including population in European countries and North American countries, to calculate the associations of rs150665432 and lung cancer risk in different geographic populations.

Statistical analyses were performed with SAS 9.3 in both discovery and replication phase; a p-value of less than 0.05 was considered to be significant.

### **Meta-analysis**

Case-control association results of both cohorts from PLINK version 1.9 and the results of multivariate logistic regression analysis in both cohorts were performed using meta-analyses with the application of R package ‘meta’ (<http://www.imbi.uni-freiburg.de/lehre/lehrbuecher/meta-analysis-with-r>) that combined test statistics and standard errors across studies. A fixed effect model was used to combine the studies in meta-analysis.

### **Analysis of the differences in age**

Student t-test was used to evaluate the differences in age at onset of lung cancer between different genotypes of the selected germline mutations in cases. We then evaluated the differences in age at onset with stratified by gender, smoking status and histology of lung cancer. In addition, we stratified the

differences in age at onset of lung cancer by the continental origins. For rs56009889, the types of countries were categorized as follows: (a) Israel, which had the highest MAF, (b) Countries with youngers only, including studies in Europe for which selection was based on early age of onset that included 2 substudies from GERMANY and RESOLUCENT from the UK (mean age = 44.83 years for GERMANY and 54.74 years for RESOLUCENT), (c) Other European Countries or (d) North American countries which included the studies located in North America. These categories were formulated based on locations and MAFs of the selected variants, as well as on the age limitation of recruitment.

To support the reliability of the differences in age at onset of lung cancer between different genotypes of rs56009889 in cases, we then compared the age at study registration between different genotypes of rs56009889 in controls.

To investigate whether or not rs56009889 was associated with onset age, we used a multivariable logistic regression model adjusted for the types of countries. To explore the effect of the interaction between the genotypes of the selected mutations and the types of countries on the onset age of lung cancer, we built a multivariate logistic regression model with adjustment for the types of countries and the genotypes of the selected mutations. In both models, age at onset of lung cancer in cases was categorized as follows: (a) 65 years of age or older, or (b) <65 years of age.

### **Structure-based prediction**

Structure-based prediction is a determinant of the functional significance of missense variants<sup>42</sup>. With using SNPeff 4.0<sup>26</sup>, we explored TANGO<sup>43</sup> that is a statistical mechanics algorithm to predict protein aggregation based on the physics-chemical principles of  $\beta$ -sheet formation<sup>27</sup>. PolyPhen-2<sup>28</sup> was applied to predict the functional effects of the germline mutations. We used Fathmm-XF<sup>30</sup> to perform accurate prediction of the functional consequences of the mutations with applying machine learning method.

## Gene Expression Data

The Cancer Genome Atlas (TCGA, <https://tcgadata.nci.nih.gov/tcga/>) level 3 RNA-seq data and clinical patient data related to 19 cancer types, composing of lung adenocarcinoma (LUAD) that included 515 tumor samples and 59 normal samples, lung squamous cell carcinoma (LUSC) that included 503 tumor samples and 52 normal samples, bladder urothelial carcinoma (BLCA), breast invasive carcinoma (BRCA),cervical squamous cell carcinoma (CESC),cholangiocarcinoma (CHOL),colon adenocarcinoma (COAD), esophageal carcinoma (ESCA),glioblastoma multiforme (GBM), head and neck squamous cell carcinoma (HNSC), kidney renal clear cell carcinoma (KIRC), kidney renal papillary cell carcinoma (KIRP), liver hepatocellular carcinoma (LIHC),pancreatic adenocarcinoma (PAAD), prostate adenocarcinoma (PRAD), rectal adenocarcinoma (READ), Sarcoma, Thymoma, thyroid carcinoma (THCA), were used to investigate whether or not the expression of ATM or KIAA0930 were associated with the primary cancer. A total of 7570 samples, including 6930 tumor samples and 640 normal samples, were consisted in the analysis. portal was employed for estimating The significance of difference in gene expression levels between tumor and normal samples was estimated by comparing generating Transcripts per million (TPM) expression values with employing UALCAN to perform t-test<sup>44</sup>.

Harvard lung expression data<sup>33</sup> included the mRNA expression values for 12,600 genes that was rescaled and normalized from the raw expression data by a rank-invariant scaling method, in order to removing the batch differences. A total of 203 samples, including 127 lung adenocarcinomas, 21 lung squamous cell carcinomas, 20 lung carcinoids, 6 small cell lung cancer and 17 normal lung specimens, were consisted in the study and performed with microarray analysis. Of the 12,600 genes, both ATM and KIAA0930 were included. Additionally, exon 2 and 5 of ATM were analyzed because isoform changes without corresponding whole gene expression changes were showed to be associated with tumor phenotype<sup>45</sup>. We used t-test to evaluate the differences in gene expression levels of ATM exon 2 and 5 and KIAA0930, respectively, between lung cancer and normal lung samples, in general and within each histologic type.

## **Data availability**

The data that support the findings of this study are available. The access numbers are “phs001273” for Oncoarray study and “phs000876.v2.p1.” for Affymetrix study in dbGAP. The Cancer Genome Atlas (TCGA, <https://tcgadata.nci.nih.gov/tcga/>) level 3 RNA-seq data are publicly available. Harvard lung cancer mRNA expression data were available at [www.pnas.org](http://www.pnas.org).

## **Acknowledgements**

This work was supported by Cancer Prevention Research Interest of Texas award RR170048 and National Institutes of Health (NIH) for the research of lung cancer (grant P30CA023108 and P20GM103534); Transdisciplinary Research in Cancer of the Lung (TRICL) (grant U19CA148127); UICC American Cancer Society Beginning Investigators Fellowship funded by the Union for International Cancer Control (UICC) (to X.Ji).

CAPUA study was supported by FIS-FEDER/Spain grant numbers FIS-01/310, FIS-PI03-0365, and FIS-07-BI060604, FICYT/Asturias grant numbers FICYT PB02-67 and FICYT IB09-133, and the University Institute of Oncology (IUOPA), of the University of Oviedo and the Ciber de Epidemiologia y Salud Pública. CIBERESP, SPAIN.CARET study was supported by NIH awards UM1 CA167462, UO1-CA6367307, CA111703 and R01 CA151989. The Liverpool Lung project is supported by the Roy Castle Lung Cancer Foundation. The Harvard Lung Cancer Study was supported by NIH grants CA092824, CA090578, CA074386. The Multiethnic Cohort Study was partially supported by NIH Grants CA164973, CA033619, CA63464 and CA148127. The MSH-PMH study was supported by The Canadian Cancer Society Research Institute (020214), Ontario Institute of Cancer and Cancer Care Ontario Chair Award to R.J.H. and G.L. and the Alan Brown Chair and Lusi Wong Programs at the Princess Margaret Hospital Foundation. R.F.T is supported by a Canada Research Chair in Pharmacogenomics, CIHR grant FDN-154294) and CAMH. NJLCS was funded by the State Key

Program of National Natural Science of China (81230067), the National Key Basic Research Program Grant (2011CB503805), the Major Program of the National Natural Science Foundation of China (81390543). The Norway study was supported by Norwegian Cancer Society, Norwegian Research Council. The Shanghai Cohort Study was supported by NIH R01 CA144034 and UM1 CA182876. The Singapore Chinese Health Study (SCHS) was supported by NIH R01 CA144034 and UM1 CA182876. The TLC study has been supported in part the James & Esther King Biomedical Research Program (09KN-15), NIH grant P50 CA119997 and P30CA76292. The Vanderbilt Lung Cancer Study – BioVU dataset used for the analyses described was obtained from Vanderbilt University Medical Center’s BioVU, which is supported by institutional funding, the 1S10RR025141-01 instrumentation award, and by the Vanderbilt CTSA grant UL1TR000445 from NCATS/NIH, K07CA172294 and U01HG004798. The Copenhagen General Population Study was supported by the Chief Physician Johan Boserup and Lise Boserup Fund, the Danish Medical Research Council and Herlev Hospital. The NELCS study was supported by NIH grant P20RR018787. The MDACC study was supported in part by grants from the NIH (P50 CA070907, R01 CA176568), Cancer Prevention & Research Institute of Texas (RP130502), and The University of Texas MD Anderson Cancer Center institutional support for the Center for Translational and Public Health Genomics. The study in Lodz center was partially funded by Nofer Institute of Occupational Medicine, under task NIOM 10.13: Predictors of mortality from non-small cell lung cancer - field study. Kentucky Lung Cancer Research Initiative was supported by the Department of Defense [Congressionally Directed Medical Research Program, U.S. Army Medical Research and Materiel Command Program] under award number: 10153006 (W81XWH-11-1-0781). Views and opinions of, and endorsements by the author(s) do not reflect those of the US Army or the Department of Defense. It also supported by NIH grant UL1TR000117 and P30 CA177558 using Shared Resource Facilities: Cancer Research Informatics, Biospecimen and Tissue Procurement, and Biostatistics and Bioinformatics. The Resource for the Study of Lung Cancer Epidemiology in North Trent (ReSoLuCENT) study was funded by the Sheffield Hospitals Charity, Sheffield Experimental Cancer

Medicine Centre and Weston Park Hospital Cancer Charity. FT was supported by a clinical PhD fellowship funded by the Yorkshire Cancer Research/Cancer Research UK Sheffield Cancer Centre. LVW held a GSK / British Lung Foundation Chair in Respiratory Research and MDT was supported by a Wellcome Trust Investigator Award (WT202849/Z/16/Z).

The authors at Laval would like to thank the staff at the Respiratory Health Network Tissue Bank of the FRQS for their valuable assistance with the lung eQTL dataset at Laval University. The lung eQTL study at Laval University was supported by the Fondation de l'Institut universitaire de cardiologie et de pneumologie de Québec, the Respiratory Health Network of the FRQS, the Canadian Institutes of Health Research (MOP - 123369). Y.B. holds a Canada Research Chair in Genomics of Heart and Lung Diseases. The research undertaken by M.D.T., L.V.W. and M.S.A. was partly funded by the National Institute for Health Research (NIHR). The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health. M.D.T. holds a Medical Research Council Senior Clinical Fellowship (G0902313).

### **Author contributions**

C.A. and X.J. designed research. J.M. contributed to the design of replication study in the research. C.A. edited the article. X.J. analyzed data and wrote the article. All authors conducted data preparation, discussed the results and revised the manuscript.

### **Conflict of interest statement**

The authors declare no competing interests.

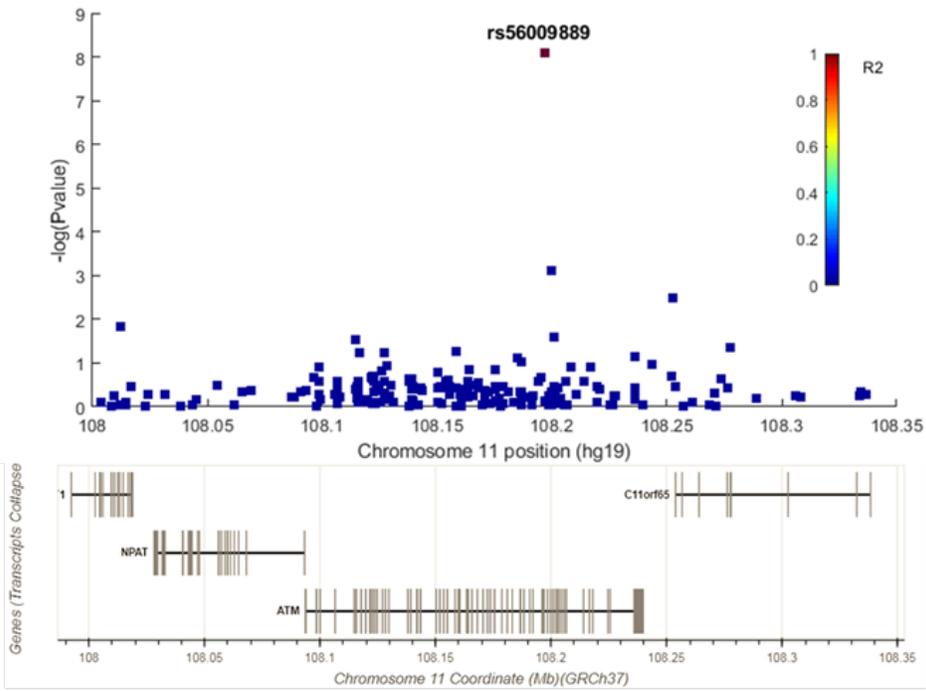
### **Reference**

1. Roukos, D.H. Genome-wide association studies: how predictable is a person's cancer risk? *Expert Rev Anticancer Ther* **9**, 389-92 (2009).

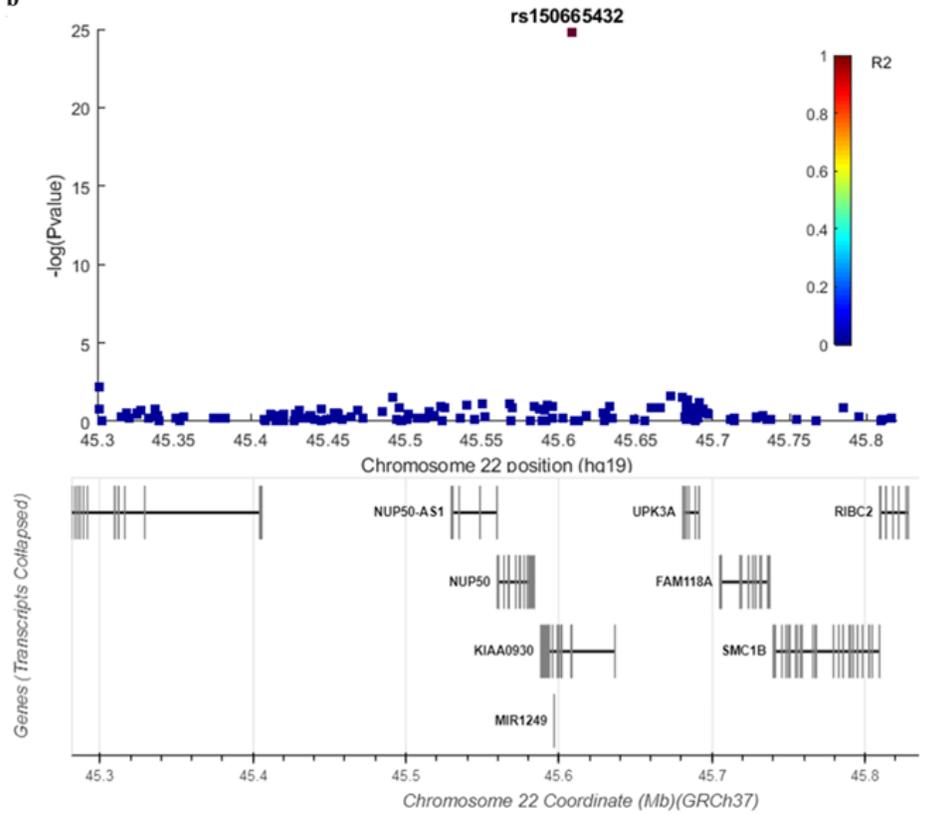
2. Hoskins, P.J. & Gotlieb, W.H. Missed therapeutic and prevention opportunities in women with BRCA-mutated epithelial ovarian cancer and their families due to low referral rates for genetic counseling and BRCA testing: A review of the literature. *CA Cancer J Clin* **67**, 493-506 (2017).
3. Turnbull, C., Sud, A. & Houlston, R.S. Cancer genetics, precision prevention and a call to action. *Nat Genet* **50**, 1212-1218 (2018).
4. Gridelli, C. *et al.* Non-small-cell lung cancer. *Nat Rev Dis Primers* **1**, 15009 (2015).
5. Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nat Genet* **46**, 736-41 (2014).
6. Siegel, R.L., Miller, K.D. & Jemal, A. Cancer Statistics, 2017. *CA Cancer J Clin* **67**, 7-30 (2017).
7. Mucci, L.A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* **315**, 68-76 (2016).
8. Roukos, D.H., Murray, S. & Briasoulis, E. Molecular genetic tools shape a roadmap towards a more accurate prognostic prediction and personalized management of cancer. *Cancer Biol Ther* **6**, 308-12 (2007).
9. (!!! INVALID CITATION !!! {}).
10. Yu, H.A. *et al.* Poor response to erlotinib in patients with tumors containing baseline EGFR T790M mutations found by routine clinical molecular testing. *Ann Oncol* **25**, 423-8 (2014).
11. Mok, T.S. *et al.* Osimertinib or Platinum-Pemetrexed in EGFR T790M-Positive Lung Cancer. *N Engl J Med* **376**, 629-640 (2017).
12. <http://exac.broadinstitute.org/>.
13. [https://www.ncbi.nlm.nih.gov/projects/SNP/snp\\_ref.cgi?rs=150665432](https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=150665432).
14. [https://www.ncbi.nlm.nih.gov/projects/SNP/snp\\_ref.cgi?rs=56009889](https://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=56009889).
15. <https://www.ncbi.nlm.nih.gov/clinvar/variation/127430/>.
16. Marechal, A. & Zou, L. DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harb Perspect Biol* **5**(2013).
17. <https://ghr.nlm.nih.gov/gene/ATM>.
18. Rennert, G. *et al.* Clinical outcomes of breast cancer in carriers of BRCA1 and BRCA2 mutations. *N Engl J Med* **357**, 115-23 (2007).
19. Takeuchi, T. *et al.* Characteristics of loss of heterozygosity in large cell neuroendocrine carcinomas of the lung and small cell lung carcinomas. *Pathol Int* **56**, 434-9 (2006).
20. Liu, C.Y. *et al.* Genome-wide Gene-Asbestos Exposure Interaction Association Study Identifies a Common Susceptibility Variant on 22q13.31 Associated with Lung Cancer Risk. *Cancer Epidemiol Biomarkers Prev* **24**, 1564-73 (2015).
21. [https://www.ncbi.nlm.nih.gov/protein/NP\\_056079.1?report=graph](https://www.ncbi.nlm.nih.gov/protein/NP_056079.1?report=graph).
22. Liu, X. *et al.* The complex genetics of hypoplastic left heart syndrome. *Nat Genet* **49**, 1152-1159 (2017).
23. Ruijs, M.W. *et al.* Late-onset common cancers in a kindred with an Arg213Gln TP53 germline mutation. *Fam Cancer* **5**, 169-74 (2006).
24. Kharazmi, E., Fallah, M., Sundquist, K. & Hemminki, K. Familial risk of early and late onset cancer: nationwide prospective cohort study. *BMJ* **345**, e8076 (2012).
25. Begam, N., Jamil, K. & Raju, S.G. Promoter Hypermethylation of the ATM Gene as a Novel Biomarker for Breast Cancer. *Asian Pac J Cancer Prev* **18**, 3003-3009 (2017).
26. De Baets, G. *et al.* SNPeff 4.0: on-line prediction of molecular and structural effects of protein-coding variants. *Nucleic Acids Res* **40**, D935-9 (2012).
27. Fernandez-Escamilla, A.M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat Biotechnol* **22**, 1302-6 (2004).

28. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248-9 (2010).
29. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20 (2013).
30. <http://fathmm.biocompute.org.uk/>.
31. <https://cancer.sanger.ac.uk/cosmic/mutation/overview?id=5019801>.
32. <http://www.uniprot.org/uniprot/Q6ICG6>.
33. Bhattacharjee, A. *et al.* Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. *Proc Natl Acad Sci U S A* **98**, 13790-5 (2001).
34. <https://www.proteinatlas.org/ENSG00000100364-KIAA0930/pathology>.
35. Huang, K.L. *et al.* Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**, 355-370 e14 (2018).
36. Ji, X. *et al.* Identification of susceptibility pathways for the role of chromosome 15q25.1 in modifying lung cancer risk. *Nature Communications* **9**, 3221 (2018).
37. McKay, J.D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* **49**, 1126-1132 (2017).
38. Kachuri, L. *et al.* Fine mapping of chromosome 5p15.33 based on a targeted deep sequencing and high density genotyping identifies novel lung cancer susceptibility loci. *Carcinogenesis* **37**, 96-105 (2016).
39. Amos, C.I. *et al.* The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* **26**, 126-135 (2017).
40. [https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study\\_id=phs000290.v1.p1](https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000290.v1.p1).
41. <https://www.genome.wustl.edu/items/topmed/>.
42. Ponzoni, L. & Bahar, I. Structural dynamics is a determinant of the functional significance of missense variants. *Proc Natl Acad Sci U S A* **115**, 4164-4169 (2018).
43. <http://tango.crg.es/>.
44. Chandrashekar, D.S. *et al.* UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* **19**, 649-658 (2017).
45. Bemmo, A. *et al.* Exon-level transcriptome profiling in murine breast cancer reveals splicing changes specific to tumors with different metastatic abilities. *PLoS One* **5**, e11981 (2010).

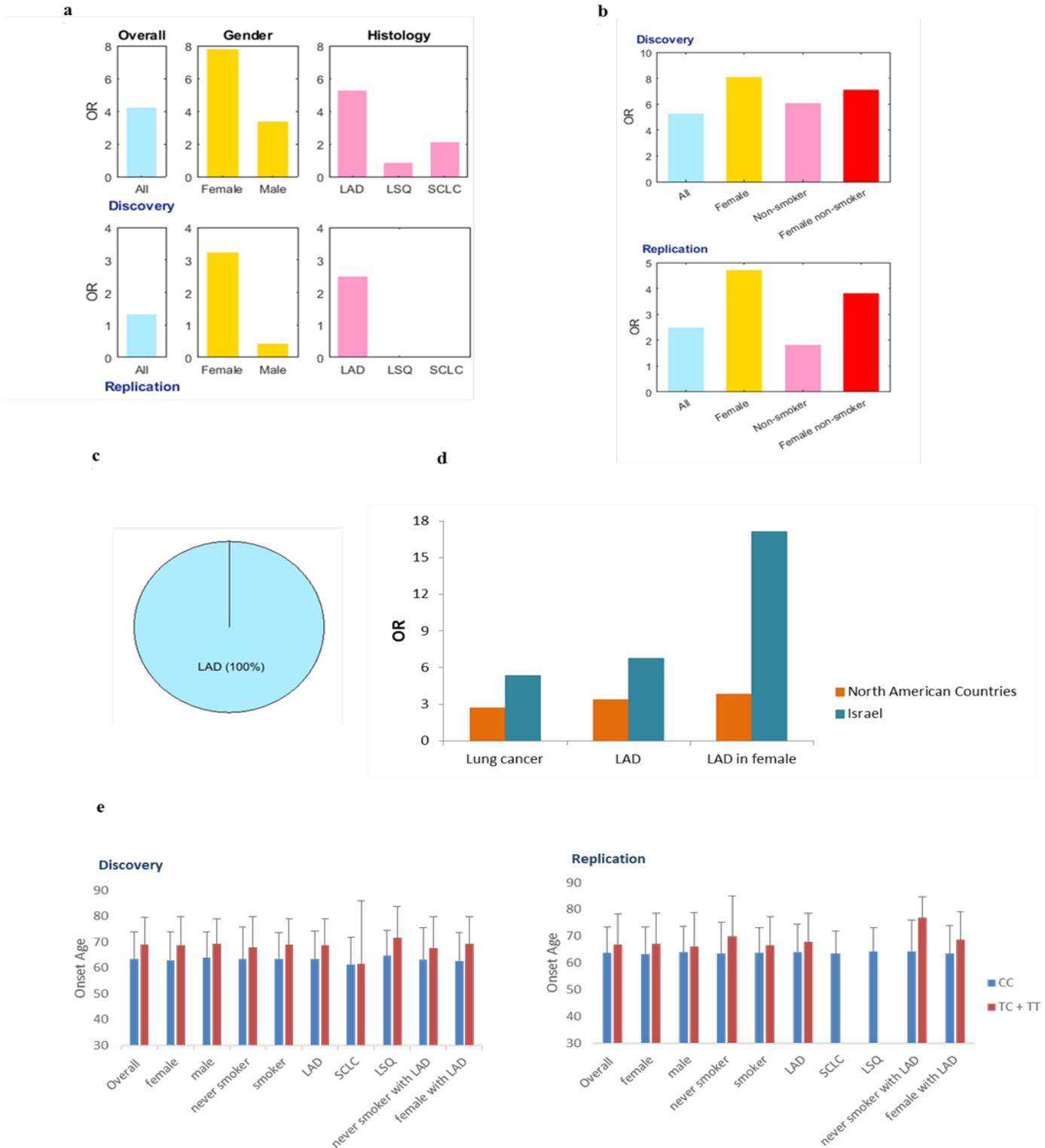
**a**



**b**

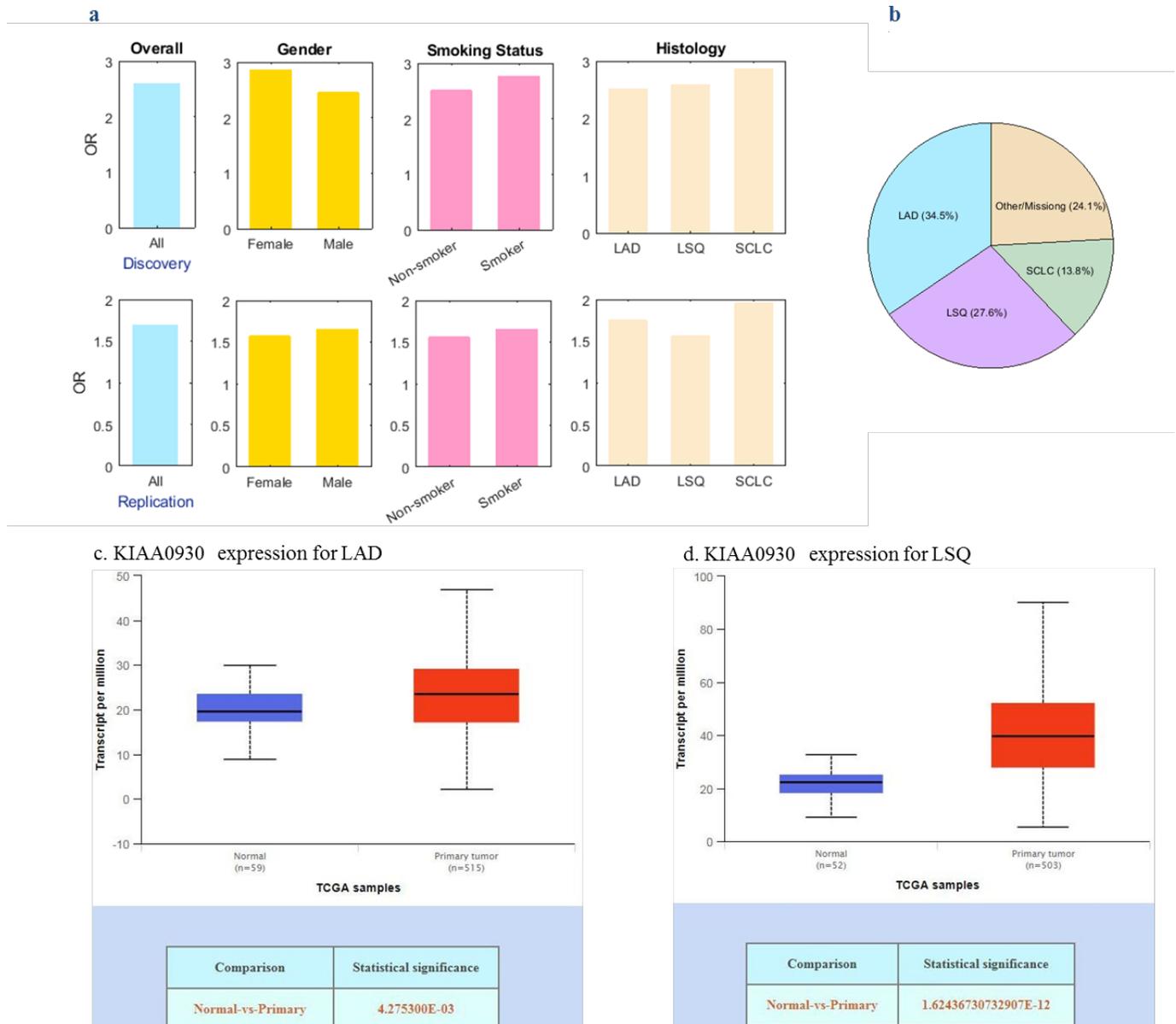


**Fig. 1: Regional lung cancer association plots for the *ATM* and *KIAA0930* risk loci.** **a**, *ATM* region for lung cancer risk. rs56009889, localizing to chromosome 11 and mapping within *ATM*, is not in linkage disequilibrium (LD) with any SNPs that have been identified before; **b**, *KIAA0930* region for lung cancer risk. rs150665432 localizes to chromosome 22 and maps within uncharacterized *KIAA0930*, which is not in LD with any SNPs that have been identified before. For each plot,  $-\log_{10} P$  values (y-axis) of the SNPs are shown according to their chromosomal positions (x-axis). The top genotyped SNP in each analysis is labeled by its rs number. The color intensity of each symbol reflects the extent of LD with the top lung cancer-associated SNP in the discovery data: blue ( $r^2=0$ ) through to red ( $r^2=1.0$ ). Physical positions are based on NCBI build 37 of the human genome. The relative positions of genes are also shown.



**Fig. 2: *ATM* rs56009889 association with lung cancer risk by histological subtypes, sex, and age of onset. a,** Compared to non-carriers, L2307F carriers had an increased risk of lung cancer with ORs being

4.19 in the discovery cohort ( $P=3.56 \times 10^{-7}$ ) and 1.31 in the replication cohort ( $P=0.45$ ). In females, L2307F carriers had a lung cancer risk with ORs being 7.76 in the discovery cohort ( $P=0.0002$ ) and 3.22 in the replication cohort ( $P=0.03$ ). In males, L2307F carriers had a lung cancer risk with ORs being 3.34 in the discovery cohort ( $P=0.0006$ ) and 0.40 in the replication cohort ( $P=0.13$ ). L2307F carriers had a significant 5.2-fold increased risk for LAD in the discovery cohort ( $P=6.47 \times 10^{-9}$ ) and a 2.5-fold increased risk in the replication cohort ( $P=0.002$ ). No associations of L2307F with the risk of LSQ or SCLC were observed. **b**, Stratified analyses of the association between rs56009889 and LAD. Females who carried L2307F had a >8-fold greater risk of LAD in the discovery cohort ( $P=0.0001$ ) and a 4.7-fold risk of LAD in the replication cohort ( $P=0.004$ ). In never smokers, L2307F exhibited a >6-fold greater risk of LAD in the discovery cohort, but we did not confirm this finding in the replication cohort. Never smoking females who harbored L2307F had a 7-fold greater risk of LAD in the discovery cohort ( $P=0.01$ ) and a 3.8-fold risk of LAD in the replication cohort ( $P=0.07$ ). **c**, Distribution of L2307F homozygotes. All the homozygotes of L2307F in the discovery cohort, no matter what age, gender, and smoking status, developed LAD in the discovery cohort ( $P=0.004$ ). No homozygotes were found in the replication cohort. **d**, Higher ORs of association between rs56009889 and the risk of lung cancer, of LAD in overall and in females were found in Israeli than in North Americas. All of the associations have reached significant. **e**, rs56009889 affects the age of onset. In the discovery cohort, the mean age of onset for lung cancer cases carrying L2307F was significantly higher than cases of non-carriers. This was observed for overall lung cancer, females, males, smokers, LAD and females with LAD. In the replication cohort, a borderline significant difference in the age of onset was observed only in females with LAD and non-smoker with LAD though the sample size is small. No carrier of the T allele developed LSQ and SCLC in the replication cohort.



**Fig. 3: KIAA0930 rs150665432 associated with lung cancer risk. a,** Stratified analyses of the association between Q4X and lung cancer risk. Compared to non-carriers, Q4X carriers had a significantly increased lung cancer risk with ORs being 2.59 in the discovery ( $P=1.15 \times 10^{-18}$ ) and 1.69 in the replication cohort ( $P=0.03$ ). Stratified analysis showed that Q4X carriers had an increased risk for lung cancer among females, males, smokers and non-smokers. We also found Q4X were associated with an enhanced risk of LAD, of SCLC, and of LSQ in both discovery and replication cohorts. The risk

associated with rs150665432 appeared to have no difference between males and females, or between smokers and non-smokers, or for various histology types. **b**, Distribution of Q4X homozygotes. In the discovery cohort, all homozygotes of the mutated allele in rs150665432 were developed to lung cancer in the discovery cohort ( $P=2.29 \times 10^{-8}$ ), no matter what age, gender and smoking status. No Q4X homozygotes were found in the replication cohort. **c**, Box-whisker plot of KIAA0930 expression in LAD. KIAA0930 is significantly over-expressed in LAD than in normal lung tissue. **d**, Box-whisker plot of KIAA0930 expression in LSQ. KIAA0930 is significantly over-expressed in LSQ than in normal lung tissue.

**Table 1.** Associations for LAD with ATM-L2307F (rs56009889) and for lung cancer with KIAA0930-Q4X(rs150665432)

Outcome	Population	Gene	Genotype	Discovery Cohort						Replication Cohort						Meta-analysis <sup>#</sup>				
				No.		Adjusted <sup>a</sup>			P	No.		Adjusted <sup>a</sup>			P	OR (95% CI)		P		
				Control	Case	OR	(95%CI)	Control		Case	OR	(95%CI)	OR (95% CI)	P						
lung cancer	All	ATM	CC	13005	15767	1				5331	4891	1								
			TC	18	77	3.79	2.2	6.6	2.57E-06	15	19	1.31	0.65	2.65	0.45	2.52	1.63	3.91	3.2E-05	
			TT	0	5	Inf	0.8	Inf	0.068*	0	0	-	-	-	-	-	-	-	-	-
			TC+TT	18	82	4.19	2.4	7.3	3.56E-07	15	19	1.31	0.65	2.65	0.45	2.7	1.75	4.16	7.8E-06	
			Trend						2.45E-07											-
lung cancer	Female	ATM	CC	5096	5777	1				2475	2203	1								
			TC	4	41	7.67	2.6	22	0.0002	5	15	3.22	1.12	9.21	0.03	4.94	2.34	10.5	2.9E-05	
			TT	0	1	Inf	0	Inf	0.49*	0	0	-	-	-	-	-	-	-	-	
			TC+TT	4	42	7.76	2.7	22	0.0002	5	15	3.22	1.12	9.21	0.03	4.97	2.35	10.5	2.7E-05	
			Trend						0.0002											-
LAD	All	ATM	CC	13005	6267	1				5331	2139	1								
			TC	18	61	4.68	2.7	8.2	7.92E-08	15	18	2.48	1.22	5.04	0.01	3.66	2.36	5.69	7.9E-09	
			TT	0	5	Inf	1.9	Inf	0.004*	0	0	-	-	-	-	-	-	-	-	
			TC+TT	18	66	5.23	3	9.2	6.47E-09	15	18	2.48	1.22	5.04	0.01	3.93	2.53	6.1	1E-09	
			Trend						5.44E-09											-
LAD	Female	ATM	CC	5096	2923	1				2475	1186	1								
			TC	4	32	7.91	2.7	23	0.0002	5	14	4.69	1.65	13.4	0.004	6.05	2.86	12.8	2.5E-06	
			TT	0	1	Inf	0	Inf	0.36*	0	0	-	-	-	-	-	-	-	-	
			TC+TT	4	33	8.05	2.8	23	0.0001	5	14	4.69	1.65	13.4	0.004	6.1	2.89	12.9	2.1E-06	
			Trend						0.0001											-
lung cancer	All	KIAA0930	GG	12642	14814	1				5308	4861	1								
			AG	126	355	2.41	2	3	7.83E-16	32	47	1.69	1.05	2.7	0.03	2.27	1.87	2.75	1.9E-16	
			AA	0	29	Inf	6.3	Inf	2.29E-08*	0	0	-	-	-	-	-	-	-	-	
			AG+AA	126	384	2.59	2.1	3.2	1.15E-18	32	47	1.69	1.05	2.7	0.03	2.41	1.99	2.92	3.9E-19	
			Trend						1.51E-19											-

<sup>a</sup>Adjusted for age at diagnosis/interview, gender, smoking status and PCs.<sup>#</sup>Fixed-effects meta-analysis adjusted for age at diagnosis/interview, gender, smoking status and PCs.

\*Values generated from Fisher's Exact Test.