Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Thorac Oncol. Author manuscript; available in PMC 2021 October 07.

Published in final edited form as:

J Thorac Oncol. 2020 December ; 15(12): 1871–1879. doi:10.1016/j.jtho.2020.08.017.

Inherited rare, deleterious variants in *ATM* increase lung adenocarcinoma risk

Myvizhi Esai Selvan, PhD[a,b], **Marjorie G. Zauderer, MD**[c], **Charles M. Rudin, MD, PhD**[c], **Siân Jones, PhD**[d], **Semanti Mukherjee, PhD**[c], **Kenneth Offit, MD, MPH**[c], **Kenan Onel, MD, PhD**[a,b], **Gad Rennert, MD, PhD**[e], **Victor E. Velculescu, MD, PhD**[d], **Steven M. Lipkin, MD, PhD**[f], **Robert J. Klein, PhD**[a,b], **Zeynep H. Gümüş, PhD**[a,b,*]

[a]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

[b]Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

[c]Memorial Sloan Kettering Cancer Center, New York, New York, USA.

[d]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

[e]Department of Community Medicine and Epidemiology, Carmel Medical Center, Clalit National Israeli Cancer Control Center, Haifa, Israel.

[f]Weill Cornell Medical College, New York, New York, USA.

Abstract

Introduction: Lung cancer is the leading cause of cancer deaths in the world, and adenocarcinoma (LUAD) is its most prevalent subtype. Symptoms often appear in advanced disease when treatment options are limited. Identifying genetic risk factors will enable better identification of high-risk individuals.

Methods: To identify LUAD risk genes, we performed a case-control association study for gene-level burden of rare, deleterious variants (RDVs) in germline whole-exome sequencing (WES) data of 1,083 LUAD patients and 7,650 controls, split into discovery and validation cohorts. Of these, we performed WES on 97 patients and acquired the rest from multiple public databases. We annotated all rare variants for pathogenicity conservatively, using ACMG guidelines and ClinVar curation, and investigated gene-level RDV burden using penalized logistic regression. All statistical tests were two-sided.

[*]**Corresponding Author**. Zeynep H. Gümüş, zeynep.gumus@mssm.edu.

AUTHOR CONTRIBUTIONS

M.E., R.J.K. and Z.H.G. wrote the manuscript. M.Z., C.R., S.M.L. and G.R. recruited patients. V.V. and S.J. led sample sequencing at PGDx. R.J.K. and Z.H.G. performed germline sequence analysis (variant calling). M.E. performed burden and other data analyses. S.M and K.Of. analyzed MSK-IMPACT cohort data. K.On. interpreted results. Z.H.G. conceived, designed and supervised the study. All authors approved the final manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Results: We discovered and replicated the finding that the burden of germline *ATM* RDVs was significantly higher in LUAD patients versus controls ($OR_{combined}$=4.6; *p*=1.7e-04; 95% CI=2.2–9.5; 1.21% of cases; 0.24% of controls). Germline *ATM* RDVs were also enriched in an independent clinical cohort of 1,594 patients from the MSK-IMPACT study (0.63%). Additionally, we observed that an Ashkenazi Jewish (AJ) founder *ATM* variant, rs56009889, was statistically significantly more frequent in AJ cases versus AJ controls in our cohort ($OR_{combined, AJ}$=2.7, *p*=6.9e-03, 95% CI=1.3–5.3).

Conclusions: Our results indicate that *ATM* is a moderate-penetrance LUAD risk gene, and that LUAD may be part of the *ATM*-related cancer syndrome spectrum. Individuals with *ATM* RDVs are at elevated LUAD risk and can benefit from increased surveillance (particularly CT scanning), early detection and chemoprevention programs, improving prognosis.

Keywords

Lung adenocarcinoma; germline risk; whole-exome sequencing; burden analysis

INTRODUCTION

Lung cancer is the leading cause of cancer deaths in the USA[1] and worldwide. Although prognosis is substantially better in early-stage as opposed to late-stage disease, most patients are diagnosed at advanced stages, when treatment options are limited.[2] Understanding genetic risk factors will help identify high-risk individuals, who can then significantly benefit from intensive surveillance (particularly CT scanning),[3,4] early detection and precision prevention strategies.[5]

Although smoking is a primary risk factor for lung cancer, only ~15% of smokers develop the disease.[6] Non-small cell lung cancer (NSCLC) accounts for approximately 90% of all cases, and lung adenocarcinoma (LUAD) is the most prevalent subtype (40%).[7] Consistent with a genetic predisposition, some NSCLC patients have a positive family history, and are affected at a young age, although they have never smoked. In fact, a family history of lung cancer increases risk,[8] and heritability is estimated at 18%.[9] However, LUAD is not known to be a part of any cancer predisposition syndrome. Furthermore, genome-wide scans for common polymorphisms[10] have only explained a small fraction of overall heritability, and thus cannot distinguish high-risk individuals. As rare variants are expected to have larger effect size, as we observed for lung squamous cell carcinoma,[11] we reasoned that a similar approach focused on germline rare deleterious variants (RDVs) will provide novel insights into LUAD risk.

MATERIALS AND METHODS

Study design.

The study design for burden analysis is summarized in Figure 1 and described in detail below.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Sample collection.

We first recruited patients from three NYC institutes in the USA, including Memorial-Sloan Kettering Cancer Center (MSKCC; n=7, IRB #15–061), Weill Cornell Medical College (WCMC; n=14, IRB #1008011221) and Icahn School of Medicine at Mount Sinai (ISMMS; n=2, IRB #12–1072) and from the Lung Cancer in Northern Israel (LCINIS) study conducted at Carmel Medical Center and Clalit National Cancer Control Center (NICCC) in Israel (n=74). This patient cohort was enriched in individuals with familial lung cancer. We collected and processed 97 blood (WCMC; LCINIS) or spit (MSKCC; ISMMS) samples for whole-exome sequencing (WES) under IRB-approved protocols. Sample preparation and WES details are provided in Supplemental Method 1. We have deposited the sequenced data in European Genome-phenome Archive (EGA).

Data acquisition.

Pursuing a resource-conscious approach, we analyzed these WES data jointly with already existing germline LUAD WES datasets. Specifically, we added case-control WES data from the Transdisciplinary Research into Cancer of the Lung (TRICL) project, which we downloaded from the database of Genotypes and Phenotypes (dbGaP, www.ncbi.nlm.nih.gov/gap) (phs000876). We then used controls from three population-based studies in dbGaP (ClinSeq project (phs000971); Myocardial Infarction Genetics (MIGen) Exome Sequencing Consortium, U. of Leicester study (phs001000); and Malmo Diet and Cancer Study (phs001101)). We designated this case-control cohort as our *discovery cohort*. For *validation cohort*, we used cases from The Cancer Genome Atlas (TCGA) and controls from eight population-based studies in dbGaP listed in Supplemental Method 2. We downloaded the TCGA germline WES BAM files from National Cancer Institute Genomic Data Commons (GDC) data portal (https://gdc-portal.nci.nih.gov), and control samples from dbGaP.

Study Cohorts.

Despite having WES data for all individuals, to protect against false positives and to ensure reproducibility, we used our WES cases together with TRICL cases for discovery and the TCGA cases for replication cohorts with separate controls, and thereby divided our study into a discovery and a replication cohort (Supplemental Table 1). Overall, for discovery cohort, we analyzed 537 cases (97 we sequenced and 440 from TRICL study) and 3,697 controls (853 from TRICL and 2,844 from ClinSeq and MIGen studies). For validation cohort, we utilized 546 sporadic cases in TCGA and 3,953 controls from eight population-based studies in dbGaP. Together, the combined cohort included 1,083 cases and 7,650 controls. The clinical characteristics of all cohorts after sample QC are listed in Table 1. Notably after sample QC, there are 74.4% (351/472) samples from blood and 25.6% (121/472) samples from adjacent normal tissue in the TCGA LUAD cases from the validation cohort. While DNA used for identifying germline variants are often from blood or saliva, tumor-adjacent normal tissue is also a possible alternative even though they might harbor early genomic aberrations. To remove such potential variants in adjacent normal, as well as potential clonal hematopoiesis variants in blood, we have filtered out those with low read count (allele fraction < 0.3).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Joint variant calling.

We performed variant discovery by realignment and joint variant calling of all case and control germline samples using the GVCF-based best practices for the Genome Analysis Toolkit (GATK, https://www.broadinstitute.org/gatk/) as implemented in a custom pipeline at ISMMS.[12] Briefly, all samples were independently aligned to human genome build GRCh37 with BWA, subject to indel realignment, duplicate marking, and base quality score recalibration using GATK and Picard, and called to a GVCF file with HaplotypeCaller. Only samples for which over 75% of the exome was callable (depth ≥20, mapping quality ≥10, base quality ≥20) and for which there was no evidence for contamination (VerifyBamID < 3%) were included in the joint variant calling from the GVCF files and variant quality score recalibration with GATK.

Sample QC.

First, we removed samples with >15% of their data missing. We then identified duplicates and related individuals by first or second degree using KING software[13] and removed a sample from each inferred pair that had the higher fraction of missing data.

Next, we removed any bias that may arise due to systematic ancestry-based variations in allele frequency differences between cases and controls, by adjusting for population stratification using Principal Component Analysis (PCA). Briefly, to identify the population structure, we first removed indels and rare variants (defined by less than 5% of minor allele frequency), using 1000 Genomes[14] and The Ashkenazi Genome Consortium (TAGC) (https://ashkenazigenome.org) datasets as reference. Then, for the remaining variants, we performed Linkage Disequilibrium (LD) pruning, filtering for a call rate of at least 0.99, and PCA with smartpca using EIGENSOFT 5.0.1 software. Finally, to filter for the least ancestry-based variation in our downstream analyses, we focused on the largest cluster within the PCA plot by PCA gating, which corresponded to individuals of European ancestry. The PCA plots of the discovery, validation and combined case-control cohorts, and the gated regions are shown in Supplemental Figure 1.

Variant-level QC.

For samples that passed the PCA gating, to ensure high-quality genotype/variant calls, we first filtered for variants with: read genotype quality ≥20; read depth ≥10; allelic depth of alternate allele ≥4; variant sites with quality score >50; quality by depth score ≥2; mapping quality ≥40; read position rank sum >−3; mapping quality rank sum >−10 and variant tranche <99%. For heterozygous genotypes, we filtered for alternative allele ratio between 0.30 and 0.70. To reduce differences between case and control samples, we kept sites with differential missingness ≤0.05 between them. Finally, we kept sites with ≥88% of data (in both cases and controls), a threshold we chose empirically to balance eliminating sites with poor quality while not eliminating sites that were not found on the capture panel for a subset of the samples.

Variant filtering.

Next, among the variants that passed QC, we focused on rare, deleterious variants (RDVs) with known pathogenicity. Such variants have been shown to have high penetrance.[15] To

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

filter out common polymorphisms, we removed any variant present in both case and control cohorts at: minor allele frequencies (MAF) >2%; or in Exome Aggregation Consortium (ExAC) non-TCGA Non-Finnish European (NFE) population at MAF >1%; or in Genome Aggregation Database[16] (gnomAD) Ashkenazi Jewish population at MAF > 1%. We considered variants that pass these filters to be rare. We filtered the remaining variants for functional impact based on those present in the ClinVar database[17] using the Annovar tool (http://annovar.openbioinformatics.org). We considered a variant to be pathogenic if it is listed as pathogenic/likely pathogenic in ClinVar; or a frameshift or stopgain variant located 5' of a variant described to be a pathogenic LOF variant in ClinVar (nonsense and frameshift).

Statistical analysis

Background variation correction.

To test for possible background variation between cases and controls, we calculated the tally of rare autosomal synonymous variants per individual. We defined synonymous variants as rare at ExAC MAF ≤0.005% and MAF ≤0.05% in each case-control cohort. Supplemental Figure 2 shows the distribution and background variation statistics of genes with rare synonymous variants in all cohorts. We noted that the frequency of neutral variation varied between cases and controls (Supplemental Figure 2) and accounted for differences in background variation as described below.

Gene-level burden analyses.

To identify risk genes associated with LUAD, we performed an exome-wide gene-agnostic analysis. First, we filtered for genes above a minimal number of RDVs (cases >2 and controls > 2). In the discovery cohort, 1130 genes had at least one RDV, out of which only 176 passed this filter. In the validation cohort, 218 genes passed this filter. Next, we used aggregate rare, deleterious variant (RDV) burden per gene using Penalized Logistic Regression Analysis (PLRA), within the logistf package in R (https://cran.r-project.org/web/packages/logistf/index.html). To adjust for background variation among samples in terms of aggregate rare variant frequencies, we used the number of genes with rare synonymous variants as a covariate for each individual in gene-level burden analyses of rare variants. We deemed genes with p-value ≤0.05 and odds ratio >1 as statistically significant risk genes. All statistical tests were two-sided.

Enrichment of *ATM* RDVs in a third independent LUAD cohort.

Next, to check whether *ATM* is enriched in RDVs in a third independent cohort, we utilized targeted clinical germline sequencing data on *ATM* from 1,594 mostly advanced LUAD cases of European ancestry assayed using MSK-IMPACT (Integrated Mutation Profiling of Actionable Cancer Targets) (Supplemental Method 3).

Sex effects on gene-level RDV burden.

To test sex effects, we used PLRA with sex as second covariate, and background variation as the first covariate but did not observe statistically significant impact. The percentage of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

males and females in each cohort are listed in Table 1 (we did not include samples with missing sex data: 9 cases and 1 control).

Germline-somatic interactions.

To test for interactions of germline variants with somatic mutations, we downloaded somatic mutation data from the Comprehensive TCGA PanCanAtlas,[18] comprising 465 of the 472 LUAD TCGA cases in the validation cohort. To ascertain mutual exclusivity between *ATM* germline RDVs and somatic *TP53* mutations, we used CoMEt[19] at default settings.

RESULTS

Study Cohorts.

To identify genes associated with LUAD risk, we performed an exome-wide multi-stage case-control study, as visually summarized in Figure 1. Briefly, we first performed germline WES on familial-enriched 97 LUAD cases. Then, for increased sample size, we pursued a resource-conscious approach and combined our WES data with those available from cases and healthy controls in dbGaP, for a total discovery cohort of 537 cases and 3,697 controls (see Methods). For validation, we used an independent cohort of 546 sporadic LUAD TCGA cases and 3,953 dbGaP controls (see Methods). We first harmonized the data by realigning and jointly calling germline genetic variants (see Methods). After sample QC, we focused on the largest ancestry-based group for downstream analyses, which were individuals of European ancestry (Supplemental Figure 1). The final discovery dataset included 513 cases and 3,423 controls, while the validation dataset included 472 cases and 3,417 controls. We also considered the combined (discovery + validation) dataset, which after combined QC included 989 cases and 6,981 controls. Clinical characteristics of all cohorts are listed in Table 1.

ATM gene exhibits statistically significant burden of germline RDVs in cases versus controls.

Within the filtered cohorts, we focused on *Rare (see* Methods*) Deleterious Variants (RDVs)*, with deleterious defined as: i) being labeled pathogenic or likely pathogenic in ClinVar, or ii) a frameshift or stopgain variant located 5' of a pathogenic LOF variant in ClinVar (nonsense and frameshift). After QC, we identified at least one RDV in 1,130 genes (median: 1 RDV/gene; range: 1–28) in the discovery cohort. We performed gene-level tests for cumulative RDV burden in cases vs. controls for all genes with RDVs. Figure 2 shows the quantile-quantile (Q-Q) plots of all burden *p*-values. The complete set of genes with burden test $p<0.05$ in the combined cohort is in Supplemental Table 2.

From these analyses, we observed that only Ataxia-telangiectasia mutated (*ATM*), a DNA damage repair gene already known to contain moderate-penetrance RDVs predisposing to breast and other cancers,[20,21] was significantly associated with LUAD in all study cohorts with consistent direction of effect (discovery cohort OR=4.05, p=0.02, 95% CI=1.3–11.9; validation cohort OR=5.50, p=1.4e-03, 95% CI=2.0–14.4; combined cohort OR=4.58, p=1.7e-04, 95% CI=2.2–9.5) (Table 2), though this is not strictly significant when correcting for all 1,130 genes originally tested in the discovery cohort. We show *ATM* RDVs in Figure

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3. The clinical characteristics of the rare *ATM* RDV carriers are provided in Supplemental Table 3.

Notably, evidencing the importance of using validation cohorts, our rigorous approach enabled us to eliminate genes that were significant in only one cohort or had inconsistent direction of effect between the discovery and validation cohorts. For example, while we observed significant association with risk for *PRKRA* gene in both cohorts, the direction of effect was opposite (discovery cohort OR=0.13; validation cohort OR=3.71; see also Figure 2).

ATM RDVs are enriched in a third independent cohort of 1,594 LUAD cases.

We next tested enrichment for *ATM* RDVs in an independent set of 1,594 advanced stage LUAD individuals of European ancestry in whom *ATM* was sequenced as part of clinical care at MSKCC ("MSK-IMPACT panel"; see Supplemental Method 3 for clinical details). Consistent with our discovery (0.97% patients; 0.23% controls) and validation cohorts (1.48% patients; 0.26% controls), this clinical cohort also showed enrichment (0.63% patients) compared to NFE population-level controls from gnomAD[16] non-cancer dataset (Supplemental Table 4). Furthermore, we observed that two ultra RDVs from our case cohort were also in the MSK LUAD patients (p.Lys468fs and rs587776551 (p.Lys1192=)) (Supplemental Table 3).

ATM founder variant rs56009889 is more frequent in Ashkenazi Jewish (AJ) cases vs. AJ controls.

An *ATM* missense variant, rs56009889 (p.Leu2307Phe), was recently found to be associated with LUAD risk[22] in individuals of European descent, but we did not include it in our primary analyses due to conflicting pathogenicity information in ClinVar.[17] While this variant is rare in Europeans, it is relatively common in AJ (gnomAD MAF 3.0%). In our original combined cohort, we found the variant more frequent in cases (MAF 1.06%) than in controls (MAF 0.18%). We then investigated the association between this variant and LUAD in our combined AJ case-control cohort (See Supplemental Figure 3; 120 cases and 284 controls), and observed that it was statistically significantly more frequent in cases (MAF 7.92%) than in controls (MAF 3.17%) (OR=2.65, p=0.007, 95% CI=1.3–5.3) (Supplemental Table 5).

ATM germline RDV carriers with Loss of Heterozygosity (LOH).

Several recent studies suggest that in breast,[20,23] pancreatic[24] and prostate cancer patients with heterozygous germline pathogenic *ATM* variants, the remaining wild-type copy is frequently inactivated (40% to 79%). To determine whether LOH was also common in LUAD patients with *ATM* RDVs, we investigated the 7 individuals with *ATM* RDVs for whom tumor data was also available (3 males and 4 females in TCGA). We observed the same trend, with 3/7 (43%) patients exhibiting LOH (rs587779846 (p.Leu1764fs); rs587779866 (c.7630–2A>C, splice); and rs587782652 (p.Val2716Ala)). None had second somatic hits at other *ATM* coding loci. Of the four non-LOH patients, two had somatic mutations in the ATM-interacting protein, EphA5.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ATM germline RDV carriers have distinct somatic mutational patterns that are mutually exclusive of *TP53* mutations.

We next asked whether similar to recent studies on *ATM* germline variant carriers with breast cancer,[23] the 7 patients with *ATM* germline RDVs showed differences in their somatic mutation patterns compared to 465 non-carriers in TCGA data. In carriers, top recurrent somatically mutated genes were *LRP1B* (71.4%), *KRAS* (57.1%), *EPHA5* (28.6%), *PTPRS* (28.6%), *STK11* (28.6%) and *TRRAP* (28.6%). However, for non-carriers, the top somatically mutated gene was *TP53* (52.2%), followed by *LRP1B* (33.8%) and *KRAS* (29.5%). Notably, consistent with observations in breast cancer, only one *ATM* RDV carrier had a *TP53* somatic mutation (14.3%); in fact carrier status was mutually exclusive of somatic *TP53* mutations (p=0.03, Supplemental Figure 4). While our study population is relatively small, consistent with other studies,[21,23] it suggests germline *ATM* RDVs impact somatic mutational patterns in LUAD individuals, which could have clinical implications.

DISCUSSION

To identify genes associated with increased risk for LUAD, we have performed by far the largest population-based study on germline WES datasets to date, reporting results on 1,083 cases and 7,650 controls. Our approach has several unique advantages. First, we explored the genetic basis for LUAD predisposition in an unbiased exome-wide manner rather than performing a candidate-gene based approach that only investigates a few genes.[25] Second, using WES datasets enabled us to investigate *rare* variants, which can have higher penetrance than *common* variants typically discovered in GWAS studies[15,26] which we filtered for using a strict pathogenicity criteria based on ClinVar.[17] Third, unlike prior WES studies, we jointly analyzed case and control WES together, which enabled us to avoid biases associated with using population-level databases as controls (e.g. such databases do not allow us to correct for individuals with multiple rare variants). Finally, this is the first LUAD WES study that validates results in an independent case-control validation cohort, which we even checked in a third independent cohort within a clinical setting. Our results rigorously establish that, in addition to its known role as a predisposition gene for other cancers including pancreatic ductal adenocarcinoma,[27] breast[28] and prostate,[29] ATM is a LUAD predisposition gene.[30,31]

As we have used the TCGA LUAD cases in our validation cohort, it is worth mentioning that this cohort has been analyzed in recent other studies.[21,31,32] However, these studies were limited in scope. Parry *et al*[32] studied 8 DNA repair genes and observed that *ATM* had the highest number of rare pathogenic germline variants, while Lu *et al*[21] observed that *ATM* had the highest number of rare *truncating* germline mutations, both only studying TCGA LUAD cases. A recent study[31] suggested that *ATM* RDVs were enriched in LUAD cases, but only compared case variant frequencies to population database controls,[33] without validation in any independent case-control cohort. These findings are further complemented by a study on *ATM common* variants in a case-control cohort,[34] which identified significant association with lung cancer risk in never-smokers.

One germline *ATM* variant, rs56009889, that others recently associated with LUAD risk[22] in individuals of European ancestry, shows conflicting interpretation of its pathogenicity in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ClinVar.[17] Therefore, we did not include it in our primary analysis. To study this variant further, which is a founder variant in the AJ population, we focused on the AJ individuals in our combined cohort. Our results support the previous reports on this variant as a risk variant. Given its high frequency in a particular population, further studies are needed to assess its pathogenicity and evaluate its importance for inclusion in risk assessment clinical genetics testing.

Multiple studies, especially on breast cancer[23], have observed mutual exclusivity between germline *ATM* variants and somatic *TP53* aberrations. While we had a limited number of matching tumors, our LUAD results are consistent with these findings. These warrant future investigations on the distinct impact of ATM germline variants on somatic LUAD landscapes, and thereby patient selection for new therapies, and patient survival.

This study should be interpreted in the context of potential limitations. We were unable to perform risk stratification based on smoking history due to incomplete smoking information for most controls. As ~70% of individuals with LUAD smoke, well-annotated control samples will enable a better understanding of the effects of smoking and other environmental agents as confounders on germline risk for these distinct lung cancer subtypes. We anticipate that as WES databases such as dbGaP gets populated, and more data get published in large studies such as UK Biobank, future studies will address such limitations.

To conclude, while lung cancer has a dismal survival rate, it can be prevented, managed or treated by the timely detection of individuals at high-risk. Here, we used population-based sampling of case-control individuals of European ancestry to identify genetic markers of LUAD risk and demonstrated that individuals with *ATM* germline RDVs are at increased risk. As *ATM* is also a recognized risk gene for cancers of the pancreas, breast and prostate, this finding suggests that LUAD may be a part of the *ATM*-related cancer syndrome. Furthermore, as individuals with germline *ATM* variants have increased surveillance for these cancers to increase early inception, our findings have important implications for their additional surveillance for LUAD with low-dose CT (as is done for individuals with a history of heavy smoking).

Notably, LUAD individuals with *somatic ATM* variants have favorable treatment outcomes for local response to radiotherapy (RT)[35] and immunotherapy[36]. These strongly support future research efforts towards understanding the association of *ATM* germline RDVs with treatment outcomes, which would strongly impact the cost/benefit analyses for clinical genetics testing.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

This work was supported by a grant to Z.H.G from LUNGevity Foundation and in part through the computational resources and staff expertise provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai. We thank Dr. Charles Powell for help with MSSM IRB protocol and Dr. Brendon Stiles for WCMC samples.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2018. CA Cancer J Clin. 2018;68(1):7–30. [PubMed: 29313949]
2. Herbst RS, Morgensztern D, Boshoff C. The biology and management of non-small cell lung cancer. Nature. 2018;553(7689):446–454. [PubMed: 29364287]
3. National Lung Screening Trial Reserach Team. Reduced Lung-Cancer Mortality with Low-Dose Computed Tomographic Screening. N Engl J Med. 2011;365(5):395–409. [PubMed: 21714641]
4. Black WC. Computed tomography screening for lung cancer in the national lung screening trial a cost-effectiveness analysis. In: Journal of Thoracic Imaging. Vol 30. Lippincott Williams and Wilkins; 2015:79–87. [PubMed: 25635704]
5. Balata H, Fong KM, Hendriks LE, et al. Prevention and Early Detection for NSCLC: Advances in Thoracic Oncology 2018. J Thorac Oncol. 2019;14(9):1513–1527. [PubMed: 31228621]
6. Wu X, Zhao H, Suk R, Christiani DC. Genetic susceptibility to tobacco-related cancer. Oncogene. 2004;23(38):6500–6523. [PubMed: 15322521]
7. Zappa C, Mousa SA. Non-small cell lung cancer: Current treatment and future advances. Transl Lung Cancer Res. 2016;5(3):288–300. [PubMed: 27413711]
8. Gao Y, Goldstein AM, Consonni D, et al. Family history of cancer and nonmalignant lung diseases as risk factors for lung cancer. Int J Cancer. 2009;125(1):146–152. [PubMed: 19350630]
9. Mucci LA, Hjelmborg JB, Harris JR, et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. JAMA. 2016;315(1):68–76. [PubMed: 26746459]
10. Bossé Y, Amos CI. A Decade of GWAS Results in Lung Cancer. Cancer Epidemiol Biomarkers Prev. 2018;27(4):363–379. [PubMed: 28615365]
11. Esai Selvan M, Klein RJ, Gümüş ZH. Rare, Pathogenic Germline Variants in Fanconi Anemia Genes Increase Risk for Squamous Lung Cancer. Clin Cancer Res. 2019;25(5):1517–1525. [PubMed: 30425093]
12. Linderman MD, Brandt T, Edelmann L, et al. Analytical validation of whole exome and whole genome sequencing for clinical applications. BMC Med Genomics. 2014;7(1):20. [PubMed: 24758382]
13. Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. Robust relationship inference in genome-wide association studies. Bioinformatics. 2010;26(22):2867–2873. [PubMed: 20926424]
14. Auton A, Abecasis GR, Altshuler DM, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68–74. [PubMed: 26432245]
15. Stadler ZK, Thom P, Robson ME, et al. Genome-Wide Association Studies of Cancer. J Clin Oncol. 2010;28(27):4255–4267. [PubMed: 20585100]
16. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434–443. [PubMed: 32461654]
17. Landrum MJ, Lee JM, Riley GR, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res. 2014;42(Database issue):D980–5. [PubMed: 24234437]
18. Hoadley KA, Yau C, Hinoue T, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173(2):291–304.e6. [PubMed: 29625048]
19. Leiserson MDM, Wu HT, Vandin F, Raphael BJ. CoMEt: A statistical approach to identify combinations of mutually exclusive alterations in cancer. Genome Biol. 2015;16(160).
20. Goldgar DE, Healey S, Dowty JG, et al. Rare variants in the ATM gene and risk of breast cancer. Breast Cancer Res. 2011;13(4). doi:10.1186/bcr2919
21. Lu C, Xie M, Wendl MC, et al. Patterns and functional implications of rare germline variants across 12 cancer types. Nat Commun. 2015;6(1):10086. [PubMed: 26689913]
22. Ji X, Mukherjee S, Landi MT, et al. Protein-altering germline mutations implicate novel genes related to lung cancer development. Nat Commun. 2020;11(1):1–14. [PubMed: 31911652]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

23. Weigelt B, Bi R, Kumar R, et al. The Landscape of Somatic Genetic Alterations in Breast Cancers From ATM Germline Mutation Carriers. J Natl Cancer Inst. 2018;110(9):1030–1034. [PubMed: 29506079]
24. Lowery MA, Wong W, Jordan EJ, et al. Prospective evaluation of germline alterations in patients with exocrine pancreatic neoplasms. J Natl Cancer Inst. 2018;110(10). doi:10.1093/jnci/djy024
25. Cheng DT, Mitchell TN, Zehir A, et al. Memorial sloan kettering-integrated mutation profiling of actionable cancer targets (MSK-IMPACT): A hybridization capture-based next-generation sequencing clinical assay for solid tumor molecular oncology. J Mol Diagnostics. 2015.
26. Gibson G. Rare and common variants: Twenty arguments. Nat Rev Genet. 2012;13(2):135–145. [PubMed: 22251874]
27. Roberts NJ, Jiao Y, Yu J, et al. ATM Mutations in Patients with Hereditary Pancreatic Cancer. Cancer Discov. 2012;2(1):41–46. [PubMed: 22585167]
28. Renwick A, Thompson D, Seal S, et al. ATM mutations that cause ataxiatelangiectasia are breast cancer susceptibility alleles. Nat Genet. 2006;38(8):873–875. [PubMed: 16832357]
29. Angèle S, Falconer A, Edwards SM, et al. ATM polymorphisms as risk factors for prostate cancer development. Br J Cancer. 2004;91(4):783–787. [PubMed: 15280931]
30. Xu Y, Gao P, Lv X, Zhang L, Zhang J. The role of the ataxia telangiectasia mutated gene in lung cancer: recent advances in research. Ther Adv Respir Dis. 2017;11(9):375–380. [PubMed: 28825373]
31. Huang K, Mashl RJ, Wu Y, et al. Pathogenic Germline Variants in 10,389 Adult Cancers. Cell. 2018;173(2):355–370.e14. [PubMed: 29625052]
32. Parry EM, Gable DL, Stanley SE, et al. Germline Mutations in DNA Repair Genes in Lung Adenocarcinoma. J Thorac Oncol. 2017;12(11):1673–1678. [PubMed: 28843361]
33. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016.
34. Lo Y-L, Hsiao C-F, Jou Y-S, et al. ATM polymorphisms and risk of lung cancer among never smokers. Lung Cancer. 2010;69(2):148–154. [PubMed: 20004998]
35. Pitter KL, Casey DL, Setton J, et al. Pathogenic Mutations in ATM As Determinants of Local Control in Non-Small Cell Lung Cancers Treated with Radiation Therapy. Int J Radiat Oncol. 2018;102(3):S226.
36. Zhang Q, Green MD, Lang X, et al. Inhibition of ATM increases interferon signaling and sensitizes pancreatic cancer to immune checkpoint blockade therapy. Cancer Res. 2019;79(15):3940–3951. [PubMed: 31101760]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

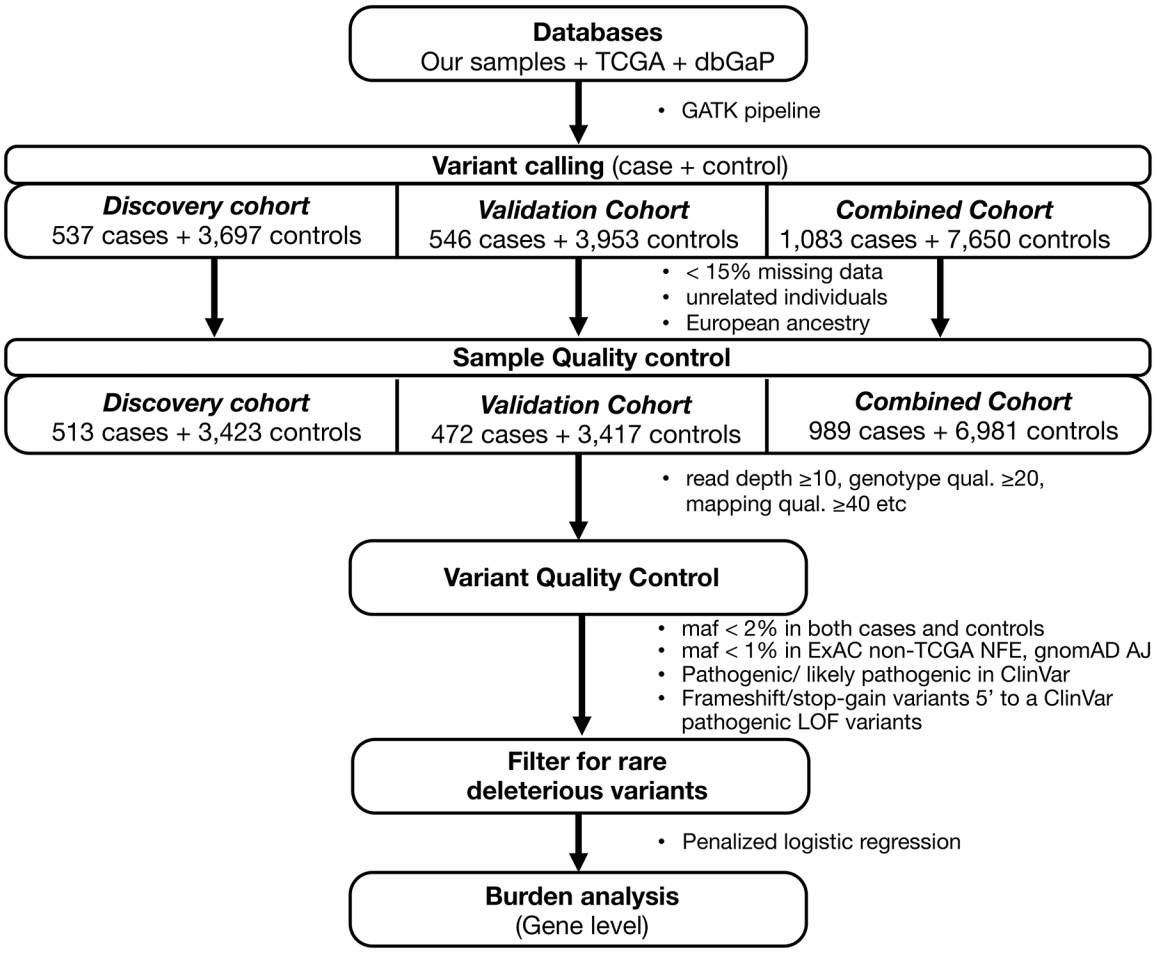

Figure 1:
Study design to perform burden analysis

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

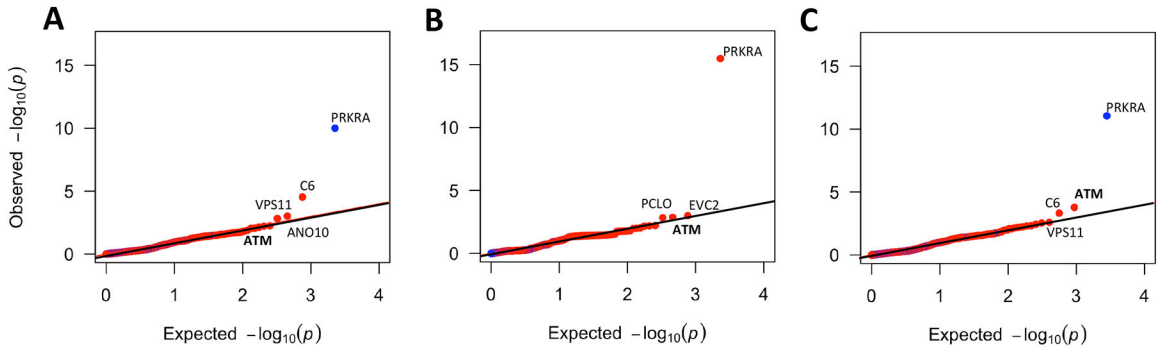

Figure 2: Quantile-Quantile (Q-Q) plots of *RDV* burden test p-values of all genes with RDVs in all study cohorts.

A) Discovery cohort, **B)** Validation cohort and **C)** Combined cohort. Red represents genes with odds ratio (OR) > 1 and blue represents genes with OR < 1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

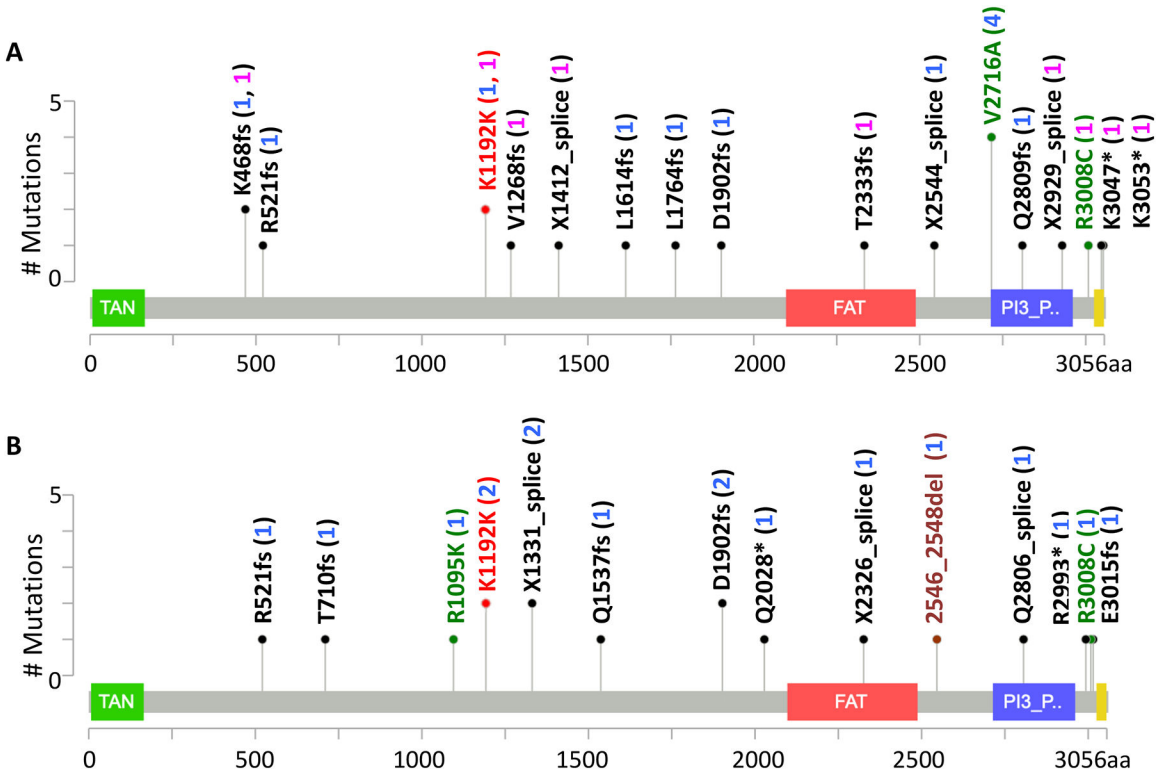

Figure 3: Rare, deleterious ATM variants in all study cohorts.

A) All cases in the combined cohort plus the MSK-IMPACT cohort (22/2,583) **B)** Controls in the combined cohort (17/6,981). Red: synonymous variants; green: missense variants; maroon: inframe variants; and black: frameshift, splicing and nonsense variants. The variant counts are given in brackets (blue: combined cohort; pink: MSK-IMPACT cohort). 1 intronic variant observed in MSK-IMPACT cohort is not displayed.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 1:

Characteristics of samples in the case-control study cohorts

Variables		Discovery cohort		Validation cohort		Combined cohort	
		Cases (513)	**Controls (3423)**	**Cases (472)**	**Controls (3417)**	**Cases (989)**	**Controls (6981)**
Gender	Male	216 (42.1%)	2341 (68.4%)	211 (44.7%)	1785 (52.2%)	431 (43.6%)	4201 (60.2%)
	Female	295 (57.5%)	1082 (31.6%)	254 (53.8%)	1632 (47.8%)	549 (55.5%)	2779 (39.8%)
	Unknown	2 (0.4%)	0	7 (1.5%)	0	9 (0.9%)	1 (0.01%)
Age	Mean (yrs)	62.0	59.1	65.8	57.4	63.8	58.4
	Unknown	30 (5.8%)	2066 (60.4%)	53 (11.2%)	2332 (68.2%)	83 (8.4%)	4536 (65.0%)
Smoking	Never	105 (20.5%)	282 (8.2%)	64 (13.6%)	102 (3.0%)	171 (17.3%)	384 (5.5%)
	Yes	373 (72.7%)	536 (15.7%)	359 (76.1%)	429 (12.6%)	734 (74.2%)	966 (13.8%)
	Unknown	35 (6.8%)	2605 (76.1%)	49 (10.4%)	2886 (84.5%)	84 (8.5%)	5631 (80.7%)

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2

Gene-level germline rare, deleterious variant (RDV) burden on *ATM* in all study cohorts.

	Discovery cohort		**Validation cohort**		**Combined cohort**	
	Case (513)	**Control (3423)**	**Case (472)**	**Control (3417)**	**Case (989)**	**Control (6981)**
	***ATM* Gene**					
# Variants	4	7	6	8	9	14
# Unique individuals	5 (0.97%)	8 (0.23%)	7 (1.48%)	9 (0.26%)	12 (1.21%)	17 (0.24%)
OR (*p*-val) [95% CI]	4.05 (0.02) [1.3–11.9]		5.50 (1.37e-03) [2.0–14.4]		4.58 (1.66e-04) [2.2–9.5]	