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] and Zeynep H. Gümüş, PhD[a,b,*]

[a]Department of Genetics and Genomic Sciences and
[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.

Supplemental Method 1: Sample Preparation and whole-exome sequencing details

Supplemental Method 2: List of controls from eight population-based studies in dbGaP in the validation cohort.

Supplemental Method 3: Sequencing and processing details of MSK-IMPACT cohort.

Supplemental Table 1: Study cohorts.

Supplemental Table 2: List of genes that have significant ($p < 0.05$) burden of rare deleterious variants in cases compared to controls in the combined cohort.

Supplemental Table 3: List of rare deleterious variants observed in *ATM* gene in the combined cohort.

Supplemental Table 4: Comparison of allele frequencies of rare deleterious variants observed in *ATM* gene in the MSK-IMPACT cohort with gnomAD non-Finnish European (NFE) population non-cancer dataset

Supplemental Table 5: Burden analysis of *ATM* rs56009889 variant in the AJ population of the combined cohort.

Supplemental Figure 1: Principal Component Analyses (PCA) of all study cohorts and all gated study cohorts.

Supplemental Figure 2: Tally of genes with per-sample rare synonymous variants between cases and controls in all cohorts.

Supplemental Figure 3: Gating for Ashkenazi Jewish population in the combined cohort

Supplemental Figure 4: Mutual exclusivity analysis of rare deleterious germline mutation in *ATM* (n=7) with somatic *TP53* mutation (n=241) in the LUAD TCGA cases of the validation cohort.

Supplemental Method 1: Sample Preparation and whole-exome sequencing.
Sequencing was performed at multiple centers (ISMMS, MSK and WCMC sequencing cores or at Personal Genome Diagnostics (PGDx, Baltimore, MD)) using Illumina HiSeq2500 instrumentation (Illumina, San Diego, CA) with 100bp paired-end reads at mean 100X coverage. At PGDx, DNA from 50 whole blood samples collected by LCINIS study were extracted using the Qiagen DNA blood mini kit (Qiagen, CA). Genomic DNA was fragmented using a Covaris sonicator (Covaris, Woburn, MA) to a size of 150-450bp and libraries were prepared using the Illumina TruSeq library kit (Illumina, San Diego, CA) according to the manufacturer's instructions. All DNA purification steps during library preparation were performed using Agencourt AMPure XP beads (Beckman Coulter, IN) and the NucleoSpin Extract II purification kit (Macherey-Nagel, PA) following the manufacturer's instructions. Exonic regions were captured in solution using the Agilent SureSelect v.4 kit according to the manufacturer's instructions (Agilent, Santa Clara, CA). The captured library was then purified using Qiagen MinElute column purification kit. The captured DNA library was amplified and PCR products were purified using NucleoSpin Extract II purification kit (Macherey-Nagel, PA), following the manufacturer's instructions. The 24 LCINIS and 2 MSSM samples sequenced at MSSM core were prepped using whole exome library human SureSelect v5-CRE, multiplexing 3 samples per lane. For 7 samples sequenced at MSK core, genomic DNA was sheared using the Covaris E220 (Covaris, Woburn, MA). Size selection was done using AMPure beads. The sheared DNA was processed into amplified indexed adapter ligated fragments using the Agilent SureSelect XT prep kit. All processing was done in 96 well plate formats using robotics (Beckman FXp, Agilent Bravo). Sample cleanups were performed following shearing and adapter ligation. Amplified libraries were pooled prior to enrichment following the SureSelect protocol (24 hour hybridization). Post-enrichment PCR was performed according to the manufacturer's protocol, with the adjustment of PCR cycles. For 14 samples sequenced at Weill Cornell Sequencing Core, germline DNA extracted from peripheral blood was used for whole exome capture using Agilent SureSelect 38 Mb paired-end sequencing and ran on Illumina HiSeq 2000s/2500s. Downstream processing of all raw sequencing files was performed using a standardized pipeline at Mt. Sinai developed for this purpose.

Supplemental Method 2: Controls from eight population-based studies in dbGaP in the validation cohort.
For the validation cohort, we included controls from the following datasets in the database of Genotypes and Phenotypes (dbGaP): Multiethnic Study of Atherosclerosis (MESA) cohort (phs000209), STAMPEED study: Northern Finland Birth Cohort (NFBC) 1966 (phs000276), NHLBI GO-ESP: Lung Cohorts Exome Sequencing Project (COPDGene) (phs000296), Common Fund (CF) Genotype-Tissue Expression (GTEx) (phs000424), Genetic Analyses in Epileptic Encephalopathies: A sub-study of Epi4K - Gene Discovery in 4,000 Epilepsy Genomes: (phs000654), ARRA Autism Sequencing Collaboration (phs000298), Bulgarian schizophrenia trio sequencing study (phs000687) and Myocardial Infarction Genetics Exome Sequencing Consortium: Ottawa heart study (phs000806).

Supplemental Method 3: Sequencing and processing details of MSK-IMPACT cohort.
Tumor and blood (matched normal) DNA from patients were sequenced by MSK-IMPACT that is New York state –approved for clinical use. This assay captures the coding exons and select introns of 468 cancer associated genes. Germline variant calling was performed as previously described[1] in anonymized data. Variants were

filtered to exclude clonal hematopoiesis and circulating tumor DNA associated variants as described earlier (Srinivasan P. et al Science under revision). Variants present in gnomAD at MAF above 2% were considered common variants and excluded from downstream analysis. An IRB (#12-245) protocol facilitated this prospective genomic analysis and the return of results to patients. All variants with <1% population frequency in the ExAC database were interpreted and pathogenicity assessment for germline variants was determined according to American College of Medical Genetics (ACMG).

References

1. Cheng DT, Prasad M, Chekaluk Y, et al. Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. *BMC Med Genomics*. 2017;10(1):33.

Supplemental Table 1: Study Cohorts. Total number of samples in each study cohort before and after sample QC.

Set	Cohort	Initial			Final		
		Cases	Controls	Total	Cases	Controls	Total
1	Discovery cohort	537	3,697	4,234	513	3,423	3,936
2	Validation cohort	546	3,953	4,499	472	3,417	3,889
3	Combined cohort	1,083	7,650	8,733	989	6,981	7,970

Supplemental Table 2: List of genes that have significant ($p < 0.05$) burden of rare deleterious variants in cases compared to controls in the combined cohort.

	Gene	Cases (989)		Controls (6981)		OR	*p*-value	95% CI Lower Limit	95% CI Upper Limit
		# RDV	Freq	# RDV	Freq				
1	ATM	9	1.21E-02	14	2.44E-03	4.58	1.66E-04	2.15	9.49
2	C6	4	1.92E-02	8	6.73E-03	2.81	4.78E-04	1.61	4.72
3	CHRNE	1	3.03E-03	1	2.86E-04	13.67	2.85E-03	2.64	82.67
4	PCLO	1	7.08E-03	1	1.43E-03	4.55	3.59E-03	1.70	11.63
5	TGM5	2	1.82E-02	2	8.88E-03	2.26	4.76E-03	1.30	3.75
6	RBBP8	1	3.03E-03	1	1.43E-04	13.96	5.03E-03	2.27	145.20
7	ANO10	2	5.06E-03	2	1.15E-03	5.31	6.27E-03	1.67	15.50
8	MMACHC	3	9.10E-03	5	2.86E-03	3.21	7.12E-03	1.40	6.82
9	SERAC1	2	5.06E-03	2	1.00E-03	5.10	8.13E-03	1.58	15.46
10	SERPINA6	1	4.04E-03	1	7.16E-04	6.33	8.17E-03	1.68	22.57
11	EVC2	4	5.06E-03	7	1.29E-03	4.56	1.06E-02	1.47	12.79
12	SLC22A5	2	4.04E-03	4	7.16E-04	5.71	1.15E-02	1.53	20.20
13	PPOX	2	3.13E-02	1	1.92E-02	1.66	1.73E-02	1.10	2.43
14	SMPD1	4	4.04E-03	5	1.00E-03	4.39	2.47E-02	1.23	13.97
15	ALOX12B	5	5.06E-03	5	1.43E-03	3.58	2.72E-02	1.17	9.76
16	PYGM	4	1.31E-02	8	6.45E-03	2.09	2.75E-02	1.09	3.76
17	TSHR	2	6.07E-03	4	1.86E-03	3.15	2.90E-02	1.14	7.86
18	QARS	2	3.03E-03	4	5.73E-04	5.49	3.01E-02	1.20	23.10
19	TMPRSS3	3	8.09E-03	6	3.44E-03	2.55	3.19E-02	1.09	5.38
20	LRP2	2	6.07E-03	2	1.86E-03	3.05	3.32E-02	1.10	7.62
21	RNASEH2B	1	7.08E-03	1	2.29E-03	2.78	3.40E-02	1.09	6.47
22	SLC24A1	2	3.03E-03	1	5.73E-04	5.12	3.45E-02	1.14	21.08
23	ATP8A2	1	4.04E-03	1	8.59E-04	3.99	3.76E-02	1.09	13.32
24	DLD	1	3.03E-03	4	1.00E-03	4.57	3.79E-02	1.10	15.54
25	OAT	2	2.02E-03	2	4.30E-04	6.82	3.81E-02	1.13	35.29
26	DUOX2	2	3.03E-03	3	4.30E-04	5.39	3.90E-02	1.10	26.35
27	CYP27A1	5	6.07E-03	9	1.86E-03	2.89	4.09E-02	1.05	7.20
28	PEX7	3	5.06E-03	2	1.58E-03	3.09	4.69E-02	1.02	8.27
29	TYMP	3	4.04E-03	3	1.43E-03	3.46	4.88E-02	1.01	10.12

Supplemental Table 3: List of rare deleterious variants observed in *ATM* gene in the combined cohort.

Pos	Id	Ref	Alt	Type	Protein Change	#Case (Gender/ Age/ Smoker)	#Ctrl (Gender)	Case Allele Freq	Ctrl Allele Freq	Replicated in MSK cohort
11:108121593	.	CAA	C	frameshift deletion	K468fs	1 (F/72/S)	0	5.06E-04	0.00E+00	yes
11:108121752	.	CAG	C	frameshift deletion	R521fs	1 (M/52/S)	1 (F)	5.06E-04	7.18E-05	
11:108126946	.	C	CAA	frameshift insertion	T710fs	0	1 (M)	0.00E+00	7.44E-05	
11:108143579	.	G	A	nonsynonymous SNV	R1095K	0	1 (M)	0.00E+00	7.24E-05	
11:108151895	rs587776551	G	A	synonymous SNV	K1192K	1 (F/70/S)	2 (F) (M)	5.14E-04	1.45E-04	yes
11:108155201	rs200196781	G	A	splicing	X1331_splice	0	2 (F) (F)	0.00E+00	1.49E-04	
11:108163518	.	CAG	C	frameshift deletion	Q1537fs	0	1 (M)	0.00E+00	7.17E-05	
11:108165719	.	A	ACT	frameshift insertion	L1614fs	1 (F/68/S)	0	5.06E-04	0.00E+00	
11:108172486	rs587779846	TC	T	frameshift deletion	L1764fs	1 (F/65/S)	0	5.46E-04	0.00E+00	
11:108178655	.	T	TA	frameshift insertion	D1902fs	1 (F/51/NS)	2 (M) (M)	5.13E-04	1.44E-04	
11:108186625	.	C	T	stopgain	Q2028X	0	1 (M)	0.00E+00	7.21E-05	
11:108198370	.	A	C	splicing	X2326_splice	0	1 (M)	0.00E+00	7.28E-05	
11:108202604	rs587779866	A	C	splicing	X2544_splice	1 (M/70/S)	0	5.07E-04	0.00E+00	
11:108202611	.	CTCTAGAATT	C	nonframeshift deletion	2546_2548del	0	1 (M)	0.00E+00	7.21E-05	
11:108205832	rs587782652	T	C	nonsynonymous SNV	V2716A	4 (M/44/S) (M/48/S) (F/74/NS) (M/67/S)	0	2.02E-03	0.00E+00	
11:108214098	.	GGTGA	G	splicing	X2806_splice	0	1 (M)	0.00E+00	7.27E-05	
11:108216476	.	CA	C	frameshift deletion	Q2809fs	1 (F/58/S)	0	5.34E-04	0.00E+00	
11:108235935	.	C	T	stopgain	R2993X	0	1 (F)	0.00E+00	7.17E-05	
11:108236086	rs587782292	C	T	nonsynonymous SNV	R3008C	0	1 (M)	0.00E+00	7.17E-05	yes
11:108236109	.	G	GA	frameshift insertion	E3015fs	0	1 (F)	0.00E+00	7.16E-05	

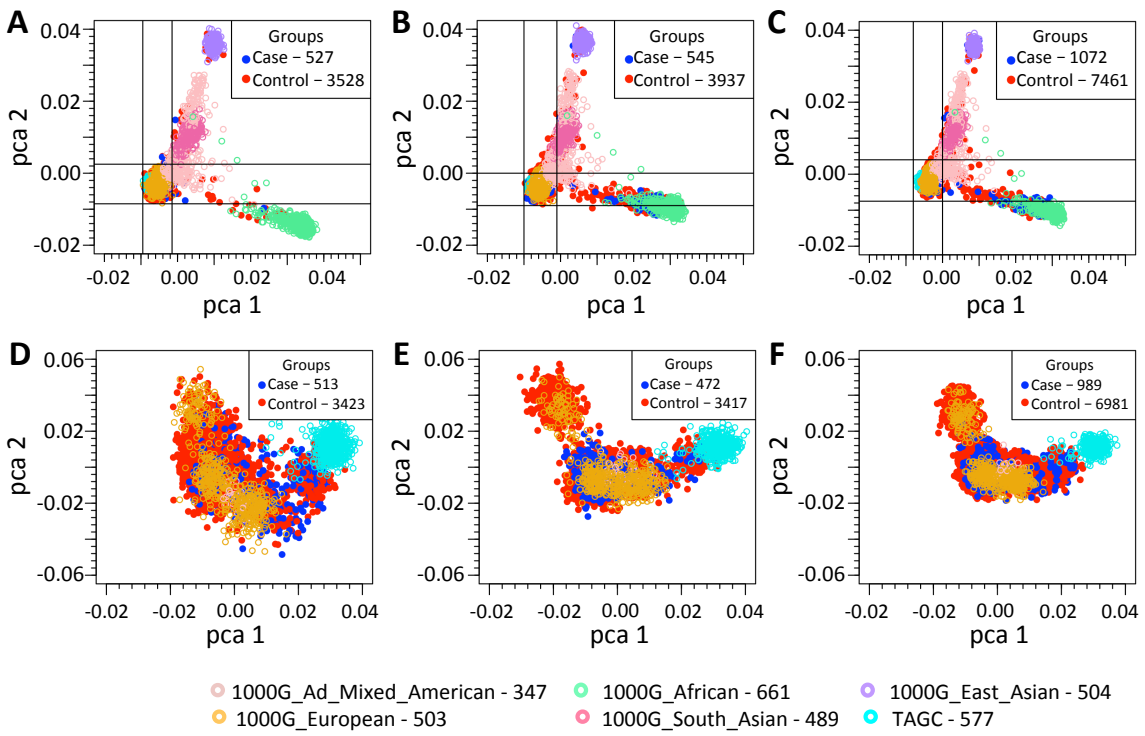
**Smoking status : S- Smoker, NS – never smoker.*

Supplemental Table 4: Comparison of allele frequencies of rare deleterious variants observed in *ATM* gene in the MSK-IMPACT cohort with gnomAD non-Finnish European (NFE) population non-cancer dataset.

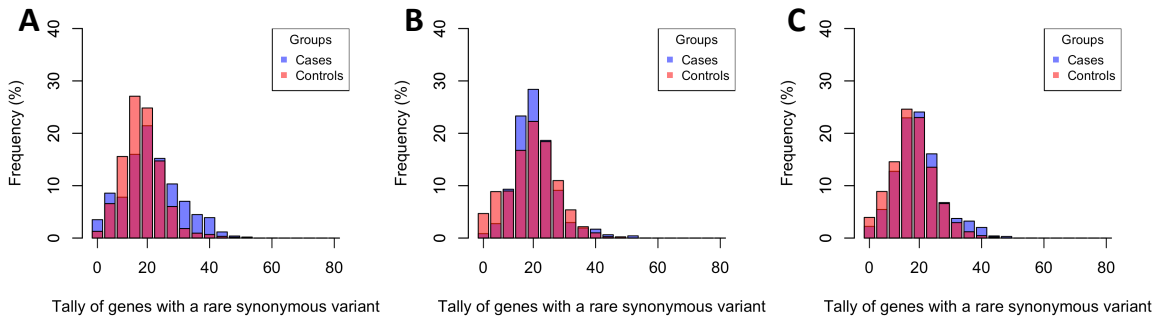
Position	HGVSc	Protein Change	MSK-IMPACT Cohort			gnomAD NFE population (non-cancer)		
			Allele count	Allele number	Allele frequency	Allele count	Allele number	Allele frequency
11:108121593	c.1402_1403del	p.Lys468fs	1	3188	3.14E-04	4	118116	3.39E-05
11:108151895	c.3576G>A	p.Lys1192=	1	3188	3.14E-04	3	102564	2.93E-05
11:108155008	c.3802delG	p.Glu1267_Val1268insTer	1	3188	3.14E-04	6	117974	5.09E-05
11:108159831	c.4236+1G>T	splice donor	1	3188	3.14E-04			
11:108198392	c.6997dupA	p.Thr2333AsnfsTer40	1	3188	3.14E-04	2	102410	1.95E-05
11:108214099	c.8418+5_8418+8delGTGA	splicing	1	3188	3.14E-04	2	117834	1.70E-05
11:108224608	c.8786+1G>A	splicing	1	3188	3.14E-04	2	102722	1.95E-05
11:108236086	c.9022C>T	p.Arg3008Cys	1	3188	3.14E-04	2	102698	1.95E-05
11:108236203	c.9139C>T	p.Arg3047Ter	1	3188	3.14E-04	1	118162	8.46E-06
11:108236221	c.9157A>T	p.Lys3053Ter	1	3188	3.14E-04			

Supplemental Table 5: Burden analysis of *ATM* rs56009889 variant in the AJ population of the combined cohort.

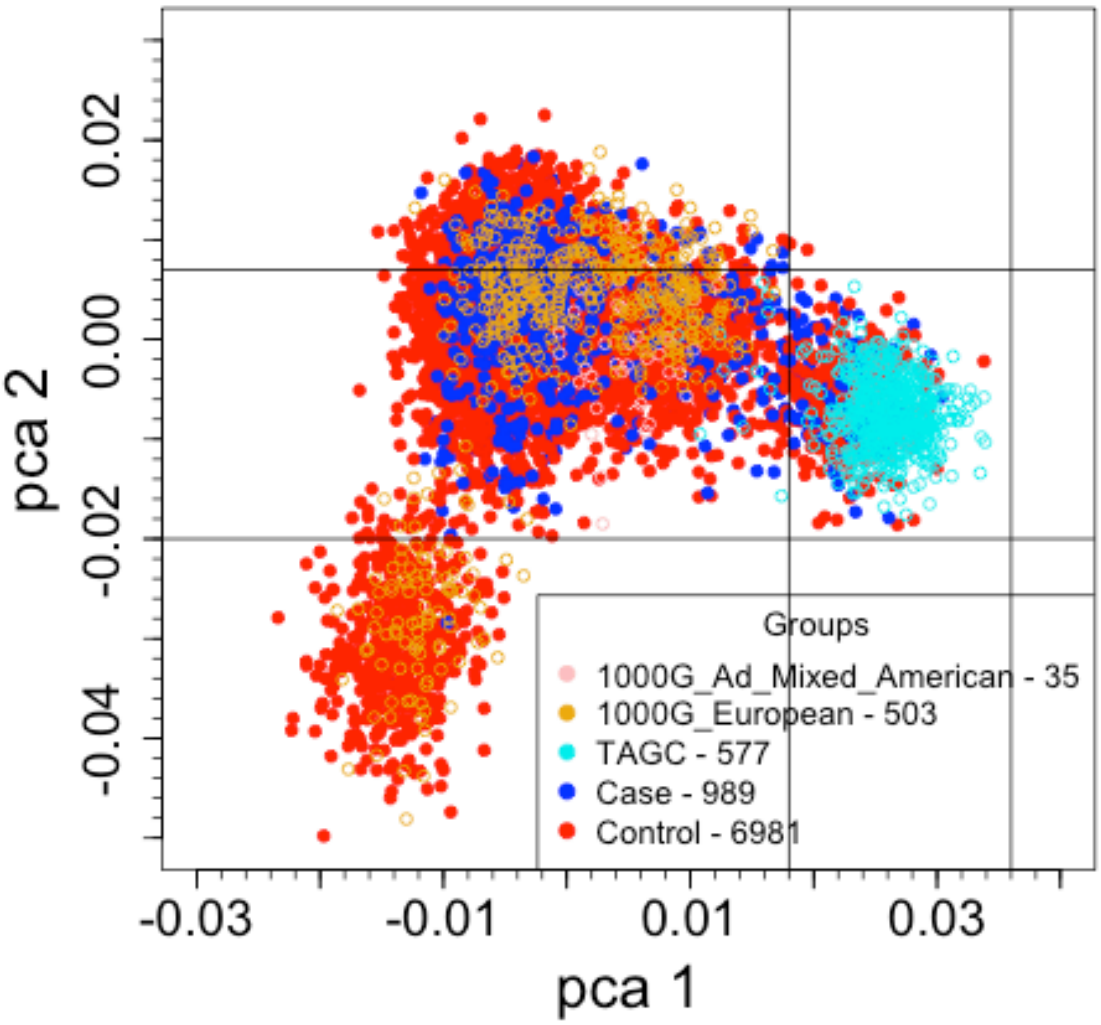
	AJ population in combined cohort	
	Case (120) Male: 33 (28%) Female: 83 (69%)	**Control (284) Male: 165 (58%) Female: 119 (42%)**
# Individuals with mutation (MAF)	17 (7.92%)	18 (3.17%)
# Male with mutation	4 (24%)	11 (61%)
# Female with mutation	13 (76%)	7 (39%)
OR (*p*-val) [95% CI]	2.65 (0.007) [1.31–5.34]	



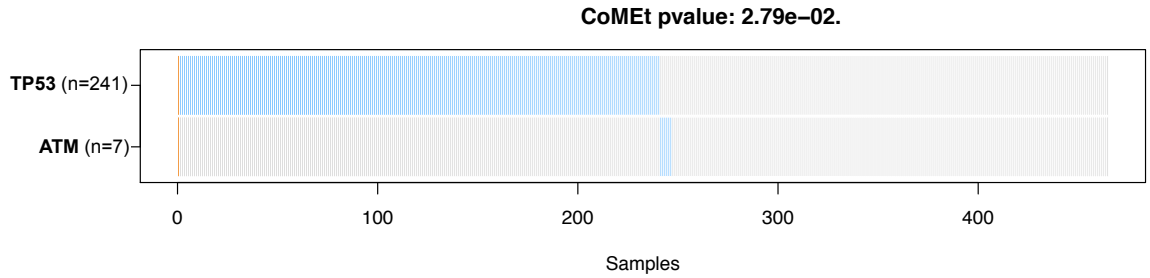
Supplemental Figure 1: Principal Component Analyses (PCA) of all study cohorts and all gated study cohorts. PCA based on common SNPs (MAF > 0.05) showing the top two principal components of (i) the study cohorts together with 1000 Genomes and TAGC samples (A-C) and of (ii) the gated samples from the study cohorts with European ancestry (D-F). **A)** Discovery cohort; **B)** Validation cohort; **C)** Combined cohort; **D)** Gated samples of discovery cohort (513 cases and 3,423 controls); **E)** Gated samples of validation cohort (472 cases and 3417 controls); **F)** Gated samples of combined cohort (989 cases and 6,981 controls).



Supplemental Figure 2: Tally of genes with per-sample rare synonymous variants between cases and controls in all study cohorts. **A)** Discovery cohort; cases: average 19.8±9.1 genes, controls: average 16.9±6.4 genes, Mann-Whitney U test *p*-value: 8.5e-14) **B)** Validation cohort; cases: average 19.2±6.9 genes, controls: average 18.2±7.9 genes, Mann-Whitney U test *p*-value: 0.13) **C)** Combined cohort; cases: average 18.3±7.6 genes, controls: average 16.4±6.9 genes, Mann-Whitney U test *p*-value: 4.8e-11.



Supplemental Figure 3: Gating for Ashkenazi Jewish population in the combined cohort. Top two principal components from Principal Component Analysis (PCA) of the gated samples of European ancestry from 1000 Genomes, TAGC and combined cohort.



Supplemental Figure 4: Mutual exclusivity analysis of rare deleterious germline mutation in *ATM* (n=7) with somatic *TP53* mutation (n=241) in the LUAD TCGA cases of the validation cohort. Red (co-occurrence) and blue (exclusive) represents mutations in the sample.