Rare germline variants in *Fanconi Anemia* genes increase risk for squamous lung cancer

Myvizhi Esai Selvan[1,2], Robert J. Klein[1,2] & Zeynep H. Gümüş[1,2]

[1]Department of Genetics and Genomic Sciences and [2]Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, New York, USA.

Lung cancer is the leading cause of cancer death worldwide, with substantially better prognosis in early stage as opposed to late stage disease. Identification of high-risk individuals could enable personalized intensive surveillance, early detection, and prevention strategies. To identify genetic risk factors for lung cancer that could be used in risk stratification, we analyzed whole-exome sequencing datasets of 444 patients and 3,479 controls and then validated our findings in an independent cohort of 318 cases and 814 controls, all of European descent. Our results show that rare germline variants in *Fanconi Anemia* genes increase risk for squamous lung cancer, with a high mutational burden of rare, disruptive variants in disease individuals. These findings can be used towards the development of a genetic diagnostic test in the clinic to identify high-risk individuals, who can benefit from personalized diagnostic programs, improving prognosis.

Corresponding Author: Zeynep H. Gümüş (zeynep.gumus@mssm.edu)

Lung cancer is the leading cause of cancer mortality in the United States[1]. Although the disease is significantly more common in individuals with a positive family history (odds ratio 1.57-5.52)[2,3] and has an estimated heritability of 18%[4], genome-wide scans have only identified common loci associated with modest risk[5]. Identifying individuals at increased genetic risk for lung cancer would lead to personalized surveillance programs for diagnosis and improve prognosis. Recent studies have identified rare variants in *BRCA2* and *CHEK2* genes with large effects in genome-wide scans of lung cancer patients[6]. Interestingly, seemingly sporadic cancer patients also carry a significant number of rare germline variants of unknown significance[7,8].

Here, we utilized the untapped resources of exome sequencing data on lung cancer cases from The Cancer Genome Atlas (TCGA) (http://cancergenome.nih.gov) and controls from various studies in database of Genotypes and Phenotypes (dbGaP) (http://www.ncbi.nlm.nih.gov/gap) to examine the role of rare coding variants in risk for lung squamous cell carcinoma (SqCC), and validated our results in an independent case-control cohort from Transdisciplinary Research in Cancer of the Lung (TRICL) (http://u19tricl.org). We confirm previous observations regarding the role of *BRCA2* gene in lung cancer risk and expand this observation to identify an overall burden of rare, deleterious variants in *Fanconi Anemia (FA)* genes in lung SqCC. While FA genes have been previously implicated in cancer risk[9], this is the first report of a role for them in lung cancer susceptibility.

Results

Discovery and Validation Cohorts. For Discovery, we realigned and jointly called 494 sporadic (non-familial) lung SqCC exomes from TCGA together with 3,954 control

exomes from dbGaP (Discovery cohort). We performed QC and relatedness analyses, removing all duplicates and first- and second- degree relatives (see Methods Section). We identified genetic ancestry using principal component analysis (PCA) (Supp. Fig.1) and focused on the single largest cluster within the PCA, which represented predominantly European ancestry (Supp. Table 1). We also examined the distribution of genes with rare synonymous variants per individual between the cases and controls and corrected for any background variation (see Methods Section and Supp. Fig. 2). For validation cohort, we performed the same analyses (including realignment and variant calling) on 333 lung SqCC exomes enriched in familial cases and 853 control exomes from the TRICL study in dbGaP (phs000876) (Validation cohort). Finally, we combined both cohorts for a joint analysis, in a combined cohort of 827 lung SqCC cases and 4,807 controls.

BRCA2 showed enrichment of rare, deleterious variants in cases compared to controls. We asked if any gene showed evidence for an enrichment of rare, deleterious variants in cases compared to controls. While numerous computational approaches exist for predicting the functional impact of a missense mutation in protein-coding sequence[10], these predictions are not ready for use in the clinic. Therefore, we focused on those genes present in the ClinVar database[11] and considered a variant to be pathogenic if: (i) it is listed as pathogenic/likely pathogenic in ClinVar; or (ii) it is a frameshift or stopgain variant located 5' of a variant described to be a pathogenic Loss of Function (LOF) variant in ClinVar (nonsense and frameshift). When we applied these criteria to rare variants (Minor Allele Frequency (MAF) <2% in either cases or controls; MAF <1% in Exome Aggregation Consortium (ExAC) non-TCGA Non-Finnish European population), we were able to test 1032 genes in the discovery cohort. The most significant gene was *BRCA2* (p=8.0e-05; odds ratio [OR]=3.0; 95% confidence interval [CI]=1.8–5.0). To

confirm this finding, we applied the same analysis approach to the validation cohort. We observed that rare, clinically significant mutations in *BRCA2* are associated with lung SqCC in this cohort as well (*p*=1.5e-03; OR=3.2; 95% CI=1.6–6.4) and in a joint analysis of both cohorts together (*p*=8.7e-08; OR=3.2; 95% CI=2.1–4.7).

Cancer predisposition gene sets showed enrichment of rare, deleterious variants in cases compared to controls. We next studied whether clinically significant variants in other cancer predisposition genes are correlated with increased risk of lung SqCC. We considered three sets of cancer predisposition genes: (i) 22 *FA* complementation group genes[12,13] (of which *BRCA2* is a member); (ii) 114 cancer predisposition (CPD) genes reported by Rahman[14]; and (iii) 20 DNA repair genes (DRG)[15] associated with autosomal dominant CPD syndromes along with 94 genes used in the clinic for genetic testing for CPD with the TruSight Cancer Gene panel[16]. We observed that in these genes, rare, pathogenic germline variants were more frequent in cases than in controls for the three gene sets (Table 1). We present the OR and *p*-values in Figure 1a and list the identified variants in Supp. Table 2.

We noted that one *BRCA2* stopgain variant in our analysis, rs11571833 (p.Lys3326Ter), has previously been reported to confer risk for lung SqCC[6]. While there are conflicting reports of its pathogenicity in ClinVar database[11], our results support the previous findings on the association between rs11571833 and lung SqCC[6]. In our study, we observed this variant is statistically significantly different between cases and controls (OR=3.3, *p*=4.4e-05, 95% CI=1.9–5.5 in discovery cohort and OR=3.0, *p*=5.3e-03, 95% CI=1.4–6.4 in validation cohort). To test the extent to which our findings could have been influenced by this one variant, we removed it and repeated our analyses. Without this variant, no significant association is observed for the CPD gene sets based on the

Rahman paper or the TruSight panel (Figure 1b). However, even when this variant is removed there is still a significantly higher rate of pathogenic germline variants in *FA* genes in cases compared to controls (OR=2.8; *p*=7.0e-04, 95% CI=1.6–4.7 in combined cohort).

Effect of smoking. Next, we tested the possible confounding effects of smoking. This analysis was restricted to the validation cohort, since the discovery cohort controls did not include complete smoking information. There was no noticeable change in the results for any gene set, after adjusting for smoking (Supp. Fig. 3). Thus, the genetic effects appear to be independent of smoking history.

Discussion

In this study, we have utilized existing Whole Exome Sequencing (WES) datasets to discover and validate genes with high burden of rare mutations that confer increased risk for lung SqCC. Similar WES methods have led to the identification of causative germline variants in patients with hereditary syndromes[17]. For this purpose, we have pursued a resource-conscious approach, leveraging all publicly available germline WES datasets on lung SqCC in TCGA and dbGaP, demonstrating the utility of these databases and especially the importance of analyzing germline DNA from cancer genomics studies. This arguably represents the first successful use of the TCGA exome data to identify a novel set of inherited risk genes for a specific cancer.

Our study further demonstrates that by using the ClinVar database to restrict analyses of WES data to only those variants known to have a clinical impact, even on other diseases, we can identify important genes and gene groups for a disease of interest. Our

results support *BRCA2* as a lung cancer risk gene, and we discovered that rare, pathogenic variants in *FA* genes also increase lung SqCC risk. Given that lung SqCC is strongly associated with a history of smoking (e.g. 95% of the cases in this study smoked), it is plausible that in these individuals, if the inter-strand cross-links caused by some of the chemicals in tobacco smoke (e.g. 1,2,3,4-diepoxybutane (DEB) -that are repaired by *FA* genes) are left untreated due to *FA* defects, this may lead to tumorigenesis. Our results are consistent with a prior study that has shown that lung adenocarcinomas are associated with common variants in 18 *FA* genes and their combined effects with smoking in a Taiwanese population[18]. Within our analyses, we note that these inherited genetic risk factors in the unselected (sporadic) cohort from TCGA were also significant. These observations are in agreement with literature that suggests that genetic predisposition can play a substantial role in the disease onset in sporadic cases, comparable to genetically enriched cohorts[19].

Overall, our results suggest that *FA* genes should be included in efforts towards the development of genetic diagnostic panel tests for lung SqCC risk. These findings increase our understanding of tumorigenesis, reveal actionable chemoprevention targets and can impact important clinical decisions regarding precision medicine based surveillance for both the affected individuals and their families. Such high-risk individuals would benefit from intensive surveillance and early detection with CT scanning.

METHODS

Data acquisition. We downloaded data for two case-control cohorts. The Discovery cohort has cases from The Cancer Genome Atlas (TCGA) and controls from eight

population-based studies in dbGaP. We downloaded the TCGA germline WES BAM files from National Cancer Institute Genomic Data Commons (GDC) data portal (https://gdc-portal.nci.nih.gov). We downloaded control samples from the following datasets in dbGaP database (www.ncbi.nlm.nih.gov/gap): 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 (MIGen_ExS) Consortium: Ottawa heart study (phs000806). The Validation case-control cohort was from the Transdisciplinary Research in Cancer of the Lung (TRICL) project, which we downloaded from the dbGaP database (phs000876).

Study cohorts. The cohorts used in the study and their sample sizes are listed in Supp. Table 1. We analyzed 494 sporadic lung squamous cancer cases in TCGA and 3954 samples from the eight population-based studies in dbGaP as the Discovery cohort. We utilized TRICL dataset (enriched in familial cases) with 333 individuals as cases and 853 individuals as controls as the Validation cohort. The Combined cohort obtained by combining the discovery and validation cohort had 827 cases and 4807 controls. After the sample QC, the Discovery dataset had 444 cases and 3479 controls, Validation dataset had 318 cases and 814 controls and Combined dataset had 765 cases and 4344 controls. Among the 444 TCGA cases, 2.7% are never smokers, 68.0% are ex-smokers and 27.0% are current smokers. Among the 318 TRICL cases, 4.1% are never smokers, 35.2% are former smokers and 60.7% are current smokers. Among the 814

TRICL controls, 34.8% are never smokers, 43.0% are former smokers and 22.1% are current smokers.

Joint calling. We performed 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 the Icahn School of Medicine at Mt. Sinai.

Sample QC. We removed samples that do not have at least 85% of their data. We identified duplicates and related individuals by first or second degree based on analysis with KING software.[20] Then we removed a sample from each inferred pair that had the higher level of data missingness between the two samples. We also removed any bias that may arise due to systematic ancestry-based variations in allele frequency differences between cases and controls (i.e. population stratification) by using samples only of European ancestry.

Population stratification. Principle Component Analysis (PCA) is a proven approach to remove spurious associations and to adjust for population stratification in large-scale studies. To identify the population structure, we removed indels and rare variants (defined by less than 5% of minor allele frequency), using 1000 Genomes dataset[21] and TAGC as reference (https://ashkenazigenome.org). For the remaining variants, we performed Linkage Disequilibrium (LD) pruning, filtered for a call rate of at least 0.99, and did PCA with smartpca in EIGENSOFT 5.0.1 software. We filtered for the least ancestry-based variation by focusing our downstream analyses on the largest set of individuals clustered within the PCA plot by PCA gating. The PCA plots of the three

case-control cohorts along with the gated regions are shown in Supp. Fig.1. As shown in the figure, these correspond to European ancestry.

Variant-Level QC. Within the individuals that passed the PCA gating, we focused on ensuring high-quality genotype/variant calls for analysis. For this purpose, we 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. Finally, we kept sites with ≥ 88% of data (in both cases and controls).

Variant filtering. Next, among the variants that passed QC, we focused on identifying rare variants with known pathogenic effects. Such causal variants are proven to significantly affect the function of the encoded protein and occur at a very low level in the control population. To filter out common polymorphisms, we removed any variant that is present in both case and control cohorts at minor allele frequencies (MAF) > 2% or present in Exome Aggregation Consortium (ExAC) non-TCGA Non-Finnish European population at MAF > 1%. We considered variants that pass these filters to be rare. We then filtered the remaining variants for functional impact based on those present in the ClinVar database[11] using the Annovar tool (http://annovar.openbioinformatics.org) (version: 2016Feb01, clinvar_20170905). We considered a variant to be pathogenic if: (i) it is listed as pathogenic/likely pathogenic in ClinVar; or (ii) it is a frameshift or stopgain variant located 5' of a variant described to be a pathogenic LOF variant in ClinVar (nonsense and frameshift). We did not consider missense variants whose functional impact we could not clearly define due to conflicting information on pathogenicity in ClinVar (with some submissions stating pathogenic/likely pathogenic and others stating

benign/likely benign/uncertain significance). To reduce any differences between samples in cases and controls, we kept sites that have differential missingness ≤ 0.05 between the cases and controls.

Background variation correction. To test for possible background variation between case and control cohorts, we calculated the tally of rare autosomal synonymous variants per individual between the cases and controls. For this purpose, we removed common variants (Exac with MAF > 0.005%) and considered rare synonymous variants with MAF ≤ 0.05% in cases and controls combined. Supp. Fig. 2 shows the distribution and background variation statistics of genes with rare synonymous variants in each of the three case-control cohorts. We accounted for differences in background variation as described below.

Gene-burden analysis. We identified risk genes associated with lung squamous cancer based on aggregate rare, pathogenic variant burden for each gene using Penalized Logistic Regression Analysis (PLRA). For this purpose we used the logistf package in R (https://cran.r-project.org/web/packages/logistf/index.html). To adjust for background variation among samples, we used the number of genes with rare synonymous variants as a covariate for each individual. We filtered out genes without pathogenic variants in cases as well as those with minimal number of variants (cases <= 2 and controls <= 1). Genes with *P*-value ≤ 0.05 were deemed statistically significant risk genes. For gene-set level burden analysis, instead of considering the number of rare pathogenic variants in each individual, we considered if an individual has a pathogenic variant in a gene only when using PLRA.

Gene sets. We considered the following gene sets associated with DNA repair and cancer predisposition for the gene-set level burden of rare, pathogenic variants associated with lung SqCC risk: (i) 22 *FA* genes involved in the recognition and repair of damaged DNA, particularly the inter-strand crosslink (ICL) that inhibits DNA replication and transcription and that arises from exposure to chemicals found in cigarette smoke. We considered all known 22 *FA/ FA*-like genes[12,13] (*FANC-A, -B, -C, -D1 (BRCA2), -D2, -E, -F, -G, -I, -J (BRIP1), -L, –M, -N (PALB2), -O (RAD51C), -P (SLX4), -Q (ERCC4), -R (RAD51), -S (BRCA1), -T, -U (XRCC2), -V (REV7) AND –W (RFWD3).* (ii) 114 CPD genes identified by Rahman[14], which were restricted to those genes in which rare mutations confer high or moderate cancer risk (Relative Risk > 2-fold). These are predominantly tumor suppressors involved in DNA repair and cell-cyle regulation. (iii) 94 genes associated with cancer predisposition (CPD), based on the Trusight Clinical Genetic Testing Panel[16], together with 20 DNA-repair genes (DRGs) associated with autosomal dominant cancer-predisposition syndromes[15]. This gene set has 100 unique genes, which overlap with 76 genes from (ii).

Smoking effects. To account for the effects of smoking on gene burden, we used PLRA with pack-years smoked a second covariate (with background variation as the first covariate). In the validation cohort, a small percentage of smokers were missing pack-year values, which were imputed as average number of pack-years of other smokers (1.9% (6/318) of cases and 1.1% (9/814) of controls). Resulting ORs and *p*-values of the three gene sets are shown in Supp. Fig. 3.

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 would also like to thank Dr. Paolo Boffetta for comments on the manuscript.

AUTHOR CONTRIBUTIONS

M.E., R.J.K. and Z.H.G. wrote the manuscript. Z.H.G supervised the study. R.J.K. performed germline sequence analysis (joint calling). M.E. performed gene burden analysis and other data analyses. R.J.K and Z.H.G. conceived and designed the study.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

REFERENCES

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2016. *CA. Cancer J. Clin.* **66,** 7–30 (2016).
2. Gao, Y. *et al.* Family history of cancer and nonmalignant lung diseases as risk factors for lung cancer. *Int. J. Cancer* **125,** 146–52 (2009).
3. Gorlova, O. Y., Weng, S.-F., Zhang, Y., Amos, C. I. & Spitz, M. R. Aggregation of cancer among relatives of never-smoking lung cancer patients. *Int. J. Cancer* **121,** 111–8 (2007).
4. Mucci, L. A. *et al.* Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *JAMA* **315,** 68–76 (2016).
5. Timofeeva, M. N. *et al.* Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Hum. Mol. Genet.* **21,** 4980–95 (2012).
6. Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of

lung cancer. *Nat. Genet.* **46,** 736–41 (2014).

7. Cheng, D. T. *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* **10,** 33 (2017).
8. Zhang, J. *et al.* Germline Mutations in Predisposition Genes in Pediatric Cancer. *N. Engl. J. Med.* **373,** 2336–2346 (2015).
9. Mathew, C. G. Fanconi anaemia genes and susceptibility to cancer. *Oncogene* **25,** 5875–5884 (2006).
10. Gonzalez-Perez, A. *et al.* Computational approaches to identify functional genetic variants in cancer genomes. *Nat. Methods* **10,** 723–9 (2013).
11. Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42,** D980-5 (2014).
12. Ceccaldi, R., Sarangi, P. & D'Andrea, A. D. The Fanconi anaemia pathway: new players and new functions. *Nat. Rev. Mol. Cell Biol.* **17,** 337–349 (2016).
13. Knies, K. *et al.* Biallelic mutations in the ubiquitin ligase RFWD3 cause Fanconi anemia. *J. Clin. Invest.* **127,** 3013–3027 (2017).
14. Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505,** 302–308 (2014).
15. Pritchard, C. C. *et al.* Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. *N. Engl. J. Med.* **375,** 443–453 (2016).
16. Illumina. www.illumina.com/clinical/translational_ genomics/panels/kits.html.
17. Bellido, F. *et al.* Association Between Germline Mutations in BRF1 , a Subunit of the RNA Polymerase III Transcription Complex, and Hereditary Colorectal Cancer. *Gastroenterology* **154,** 181–194.e20 (2018).
18. Yang, S.-Y. *et al.* Fanconi anemia genes in lung adenocarcinoma- a pathway-wide study on cancer susceptibility. *J. Biomed. Sci.* **23,** 23 (2016).

19. Artomov, M. *et al.* A Strategy for Large-Scale Systematic Pan-Cancer Germline Rare Variation Analysis. *bioRxiv* 162479 (2017). doi:10.1101/162479

20. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26,** 2867–2873 (2010).

21. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

Table 1: Pathogenic mutations identified by our ClinVar analysis of the following case-control cohorts: i) Discovery cohort (TCGA-dbGap) ii) Validation cohort (TRICL) and iii) Combined cohort (TCGA+TRICL-dbGaP), for the following gene sets: 1) 22 Fanconi Anemia genes; 2) 114 cancer predisposition genes from ref.14; and 3) 20 DNA repair genes from ref.15 and 94 cancer predisposition genes from ref.16.

	TCGA-dbGaP dataset		**TRICL dataset**		**TCGA+TRICL-dbGaP dataset**	
	case	**control**	**case**	**control**	**case**	**control**
	Fanconi Anemia Complementation Group					
# Mutations	9	27	10	10	18	35
# Genes	6	11	7	5	10	11
# Individuals	30	97	24	23	53	107
# Unique individuals	30 (6.76%)	96 (2.76%)	24 (7.55%)	22 (2.70%)	53 (6.93%)	105 (2.42%)
	114 Cancer Predisposition Genes					
# Mutations	21	108	28	56	38	132
# Genes	15	40	17	26	22	42
# Individuals	51	235	46	97	91	306
# Unique individuals	49 (11.04%)	228 (6.55%)	42 (13.21%)	91 (11.18%)	87 (11.37%)	295 (6.79%)
	20 DNA Repair Genes + 94 Cancer Predisposition Genes					
# Mutations	20	85	18	43	31	108
# Genes	15	33	12	20	20	37
# Individuals	45	185	33	68	73	224
# Unique individuals	43 (9.68%)	182 (5.23%)	31 (9.75%)	64 (7.86%)	71 (9.28%)	217 (5.00%)

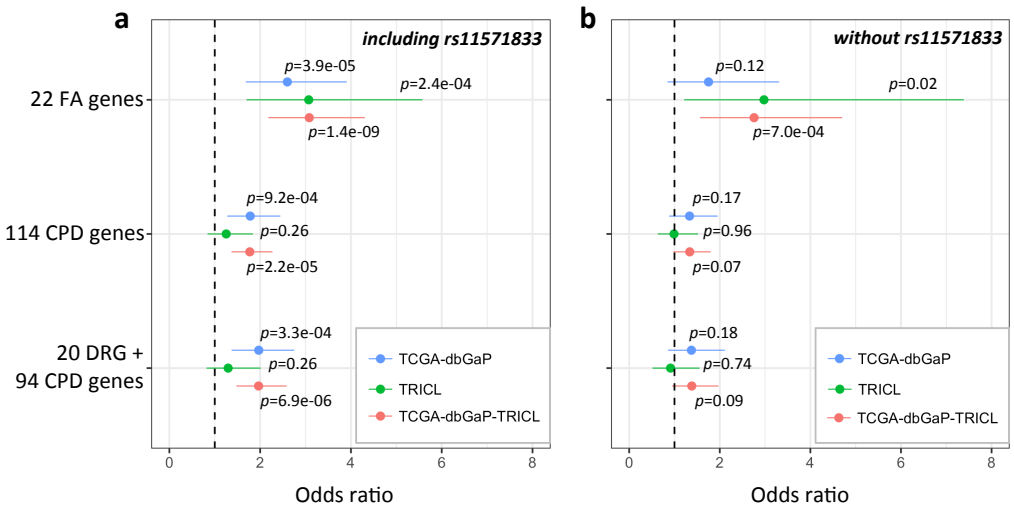

Figure 1: Geneset-level burden of rare, pathogenic variant p-values and Odds Ratios (ORs). For each gene set, (22 FA genes; 114 cancer predisposition (CPD) genes; 20 DNA repair genes (DRG) with 94 CPD genes) geneset-level burden of rare, pathogenic variants for the three cohorts i) Discovery cohort (TCGA-dbGaP) – blue ii) Validation cohort (TRICL) – green iii) Combined cohort (TCGA-dbGaP-TRICL) – red **a)** including all variants; **b)** without BRCA2 rs11571833 stopgain variant. The whiskers span the 95% confidence interval for OR values.