Associating *CYP2A6* structural variants with ovarian and lung cancer risk in the UK Biobank: replication and extension

Alec W.R. Langlois[1,2]; Jennie G. Pouget[2,3]; Jo Knight[4]; Meghan J. Chenoweth[1,2,3] & Rachel F. Tyndale[†1,2,3]

[1]Department of Pharmacology & Toxicology, University of Toronto; 1 King's College Circle, Toronto, ON, M5S 1A8, Canada.

[2]Campbell Family Mental Health Research Institute, Centre for Addiction and Mental Health 100 Stokes Street, Toronto, ON, M6J 1H4, Canada

[3]Department of Psychiatry, University of Toronto; 250 College Street, Toronto, ON, M5T 1R8, Canada

[4]Data Science Institute and Medical School, Lancaster University

†Corresponding author.

Correspondence: Rachel F. Tyndale, Department of Pharmacology and Toxicology, Room 4326, Medical Sciences Building, 1 King's College Circle, University of Toronto, ON M5S 1A8, Canada.

Email: r.tyndale@utoronto.ca

Funding: This work was funded by a Canadian Institutes of Health Research (CIHR) Project grant (PJY-159710) and Foundation grant (FDN-154294), National Institutes of Health (NIH) Grant PGRN DA020830, and a Canada Research Chair in Pharmacogenomics (Tyndale).

Abstract

CYP2A6 is a polymorphic enzyme that inactivates nicotine; structural variants (SVs) include gene deletions and hybrids with the neighbouring pseudogene *CYP2A7*. Two studies found that *CYP2A7* deletions were associated with ovarian cancer risk. Using their methodology, we aimed to characterize *CYP2A6* SVs (which may be misidentified by prediction software as *CYP2A7* SVs), then assess *CYP2A6* SV-associated risk for ovarian cancer, and extend analyses to lung cancer.

An updated reference panel was created to impute *CYP2A6* SVs from UK Biobank array data. Logistic regression models analyzed the association between *CYP2A6* SVs and cancer risk, adjusting for covariates.

Software-predicted *CYP2A7* deletions were concordant with known *CYP2A6* SVs. Deleterious *CYP2A6* SVs were not associated with ovarian cancer (OR=1.06; 95% CI: 0.80-1.37; p=0.7) but did reduce the risk of lung cancer (OR=0.44; 95% CI: 0.29-0.64; p<0.0001), and a lung cancer subtype. Replication of known lung cancer associations indicates the validity of array-based SV analyses.

Keywords: CYP2A6; ovarian cancer; lung cancer; structural variants; pharmacogenetics; UK Biobank

Introduction

CYP2A6 is the primary nicotine-inactivating enzyme; it also metabolizes other drugs (e.g. efavirenz and tegafur) (1). The gene encoding CYP2A6 is highly polymorphic (2). Genetic variation in CYP2A6 alters the rate of nicotine inactivation which alters cigarette smoking behaviours, cessation and risk for tobacco-related diseases including lung cancer (LC) (3-6).

CYP2A6, located on chromosome 19q13.2, is 30 Kb downstream of *CYP2A7*, an *inactive* homologue sharing 95% nucleotide identity (7). Structural variants (SV) in *CYP2A6* and *CYP2A7* arise from unequal cross-over events involving their homologous regions, resulting in full gene deletions, duplications, and hybrids (7). *CYP2A6*4*, a common *CYP2A6* deletion variant, was associated with a lower risk of LC among current smokers in a meta-analysis of case-control studies (n=4385 cases, 4142 controls) (8).

Recent papers investigated whether ovarian cancer (OC) in European-ancestry individuals (EUR) was associated with genome-wide deletions and duplications, predicted based on signal intensity from single nucleotide polymorphism (SNP) array data using PennCNV and similar SV prediction programs (9-11). Among females with pathogenic *BRCA1* variants, Walker et al. found that there was an association between *CYP2A7* deletions and a *decreased* risk of OC (9). Among all females, Reid et al. found that there was an association between *CYP2A7* deletions and an *increased* risk of epithelial OC (10). Disruption of a nearby *EGLN2* enhancer was proposed as an explanation for the association (10).

Both papers used gene deletion and duplication prediction programs including PennCNV that use SNP array signal intensity data as input (9-11). We determined whether the *CYP2A7* deletions identified (9, 10) represent known *CYP2A6* SVs (Figure 1), as all known deletion SVs in this region affect both genes, by evaluating PennCNV performance in an internal dataset with known *CYP2A6* SV diplotypes. Next, we imputed *CYP2A6* SVs from SNP array genotype data available in the UK Biobank (UKB) using a validated SV reference panel (>70% sensitivity, ~99% specificity (7)), and analyzed the association between *CYP2A6*

deletion SVs and risk for OC and LC (confirming our method through replication and extension of the LC risk).

Methods

Reference panel and internal PennCNV validation

Previously, we developed a reference panel (n=935 EUR individuals) with known CYP2A6 SV diplotypes for use in imputing *CYP2A6* SVs from SNP array data (7)). Individuals (n=209) from the reference panel underwent next-gen sequencing (NGS) (GRCh37 chr19:41322500-41615000) (12). Reference panel participants underwent genome-wide SV prediction with PennCNV, using QC and CNV merging (CNVruler) (13), following the approach of Reid et al (10).

Updated SV imputation panel validation

The original reference panel (n=935) was developed using Illumina-array-genotyped SNPs in a ~4 Mb genomic region surrounding *CYP2A6* (SNPs within *CYP2A6* were excluded as they are disrupted by SVs, described in (14))(Figure 2). Within this ~4 Mb region, the overlap of original reference panel genotyped SNPs with those genotyped in UKB Axiom arrays was minimal (i.e. n=243/1659; 24% of reference panel genotyped SNPs overlapped with the Axiom Array) (Figure 1). Thus, an updated reference panel was created using imputed SNPs overlapping with UKB Illumina-array-genotyped SNPs (GRCh37 chr19:39000000-43000000). Genotyped plus imputed SNPs in the updated reference panel overlapped more substantially (i.e. N=1386/1659, 84% of SNPs genotyped on the Axiom Array overlapped with the genotyped and imputed updated reference panel SNPs)(Figure 2).

Genotype calls from NGS versus imputation were compared at overlapping positions (n=5047; minimum read depth=20).

Leave-one-out cross validation was used to estimate the accuracy of the updated reference panel, and accuracy was compared to the original reference panel using only genotyped SNPs.

SV Imputation

VCF files with SNP genotypes extracted from GRCh37 chr19:39000000-43000000 were created for UKB EUR (n=409522) (15), who shared similar genetic ancestry based on principal components analysis (UKB data-field 22006). These were then used as target files for SV imputation using Beagle 5.2, with our updated reference panel as the reference (16).

Case-control analyses

Cases were selected using ovarian (184.1 and 184.11) and lung (165.1) cancer phecodes. OC analyses were limited to females and adjusted for smoking status (current, former, or never smokers). LC case-control analyses were within current smokers, and adjusted for sex; further analyses were performed in the subset of LC cases with "squamous cell carcinoma" histology (UKB data-field 40011), a subtype of LC where *CYP2A6* deletions were strongly protective in a recent study (17). Logistic regression analyses, where having at least one deleterious *CYP2A6* SV (*CYP2A6*4, *12, *34,* or **53*) was the exposure, tested for an association with case status (coded as 1 = case, 0 = control). Analyses controlled for age and the first ten principal components.

Results

Results – Internal PennCNV validation

PennCNV and CNVruler software identified a deletion region (19:41341589-41386033) encompassing *CYP2A6* and *CYP2A7* (Figure 1). All individuals predicted by PennCNV to have deletions in the region (n=34) had *CYP2A6* SV diplotypes *CYP2A6*1/*12* (n=27), *CYP2A6*1/*4* (n=4), *CYP2A6*1x2/*12* (n=2), or *CYP2A6*12/*12* (n=1).

Updated SV imputation panel validation

To validate the use of imputed SNPs as proxies for genotyped SNPs in our updated reference panel, we examined concordance of imputed SNP genotypes with NGS genotypes within the n=209 subset. Reference panel SNPs overlapped with n=5047 sequenced positions; on average, n=4598 positions per sample were sequenced at a depth of >20 reads. Concordance was 99.7% (4586/4598 concordant calls per sample, Figure 2), indicating the validity of using imputed SNPs as a proxy for genotyped SNPs in our updated reference panel.

Leave-one-out cross validation of the updated reference panel (n=935 participants) was performed. Overall, 70% (52/74 SV alleles) of SV alleles were accurately imputed; this included duplication (*CYP2A6*1x2:* 1/15) and deleterious (*CYP2A6*4*: 0/6; *CYP2A6*12*: 42/43; *CYP2A6*53*: 9/10) SVs. False positives were rare, occurring for <1% of non-SV alleles (3 called SV alleles/1796 total non-SV alleles). These data were consistent with previous data using the original reference panel with genotyped SNPs (Figure 2) (7).

SV imputation in UKB and case-control analyses

Demographic characteristics of the genetically-confirmed EUR are found in Supplementary Table 1. SV diplotype was imputed for all participants (n=409277). Among females (n=1097 cases, n=201390 controls) the risk of OC among those with, relative to without, at least one deleterious SV allele was not significantly different (OR=1.06; 95% CI: 0.80-1.37; p=0.7)(Figure 3A).

Among current smokers (n=1040 cases, n=40211 controls) the risk of LC among those with, relative to without, at least one deleterious SV allele was significantly lower (OR=0.44; 95% CI: 0.29-0.64; p<0.0001). In a sub-analysis, the risk of SCC (n=270/1040 LC cases) among those with, relative to without, at least one deleterious SV allele was also significantly lower (OR=0.25; 95% CI: 0.08-0.58; p<0.01)(Figure 3B).

Discussion

Our findings suggest that the *CYP2A7* gene deletions detected in previous analyses of OC (9, 10) are actually *CYP2A6*4* and **12* (Figure 1). The deletion region inferred by Reid et al. using PennCNV includes both *CYP2A6* and *CYP2A7* (19:41341589-41433931), similar to the region detected using PennCNV in our reference panel participants (10). The approach used by Walker et al. merged results from PennCNV and three additional CNV prediction algorithms (these algorithms were not replicated due to difficulties running on modern Linux/Java (9)). Nevertheless, considering the overlap of inferred deletion regions (Figure 1), and similar frequencies of deletions in Reid et al. (3.4%), Walker et al. (2.9%), and in our reference panel participants (by Taqman CNV genotyping: **12* and **4* combined 2.6%), we have provided evidence that the *CYP2A7* deletions identified using CNV prediction software are known *CYP2A6* SVs.

We found no association between deleterious *CYP2A6* SVs and risk for OC. These results contrast with Reid et al. and Walker et al. who found an association between *CYP2A7* deletions (likely *CYP2A6 SVs*) identified using *in silico* deletion prediction software and significantly increased risk and decreased risk, respectively, of OC (10, 18). Reid et al. restricted analyses to epithelial OC cases; while our study investigated all OC cases together (due to limited histological data available). However, most OC cases are epithelial (~90%) (19). Walker et al. included only *BRCA1* pathogenic variant carriers; since only 10-15% of OC cases carry *BRCA1* pathogenic variants, a UKB sub-analysis (n=1097 OC cases total) was unfeasible (18). Thus, the association between *CYP2A6/CYP2A7* SV and risk for OC selectively within females with *BRCA1* mutations remains to be clarified. Recently rare SVs were examined using a method similar to PennCNV with no *CYP2A6* association with OC risk found; common SVs were analyzed using tag SNPs, but *CYP2A6* SVs were not captured within these analyses as there were no SNPs tagging common *CYP2A6* SVs for EUR (20).

In contrast to OC, we found an association between deleterious *CYP2A6* SV and reduced risk of LC among current smokers. These results extend previous associations of deleterious *CYP2A6* SNPs as protective for LC (6), add to the body of literature examining *CYP2A6* SNP associations with LC risk in EUR, and serve as a validation of the updated *CYP2A6* SV reference panel's use in the UKB.

Overall, we did not detect an association between *CYP2A6*/*CYP2A7* SVs and OC risk. Our study extends previous findings of a role for *CYP2A6* SV in reducing risk for LC among smokers and demonstrates the utility of SV imputation of array data in large publicly available biobanks.

Data availability statement: Data from participants is accessible in the UK Biobank (datafields: 20116, 21022, 22001, 22006, 22418, 41270, 41271); reference panel data is not publicly available due to individual privacy concerns.

Code availability statement: Available upon request.

References

1. McDonagh EM, Wassenaar C, David SP, Tyndale RF, Altman RB, Whirl-Carrillo M, et al. PharmGKB summary: very important pharmacogene information for cytochrome P-450, family 2, subfamily A, polypeptide 6. Pharmacogenet Genomics. 2012;22(9):695-708.
2. El-Boraie A, Taghavi T, Chenoweth MJ, Fukunaga K, Mushiroda T, Kubo M, et al. Evaluation of a weighted genetic risk score for the prediction of biomarkers of CYP2A6 activity. Addict Biol. 2020;25(1):e12741.
3. Benowitz NL, Pomerleau OF, Pomerleau CS, Jacob P. Nicotine metabolite ratio as a predictor of cigarette consumption. Nicotine Tob Res. 2003;5(5):621-4.
4. Wassenaar CA, Ye Y, Cai Q, Aldrich MC, Knight J, Spitz MR, et al. CYP2A6 reduced activity gene variants confer reduction in lung cancer risk in African American smokers--findings from two independent populations. Carcinogenesis. 2015;36(1):99-103.

5. Liu T, David SP, Tyndale RF, Wang H, Zhou Q, Ding P, et al. Associations of CYP2A6 genotype with smoking behaviors in southern China. Addiction. 2011;106(5):985-94.

6. Wassenaar CA, Dong Q, Wei Q, Amos CI, Spitz MR, Tyndale RF. Relationship between CYP2A6 and CHRNA5-CHRNA3-CHRNB4 variation and smoking behaviors and lung cancer risk. J Natl Cancer Inst. 2011;103(17):1342-6.

7. Langlois AWR, El-Boraie A, Pouget JG, Cox LS, Ahluwalia JS, Fukunaga K, et al. Genotyping, characterization, and imputation of known and novel CYP2A6 structural variants using SNP array data. J Hum Genet. 2023.

8. Johani FH, Majid MSA, Azme MH, Nawi AM. Cytochrome P450 2A6 whole-gene deletion (CYP2A6*4) polymorphism reduces risk of lung cancer: A meta-analysis. Tob Induc Dis. 2020;18:50.

9. Walker LC, Marquart L, Pearson JF, Wiggins GA, O'Mara TA, Parsons MT, et al. Evaluation of copy-number variants as modifiers of breast and ovarian cancer risk for BRCA1 pathogenic variant carriers. Eur J Hum Genet. 2017;25(4):432-8.

10. Reid BM, Permuth JB, Chen YA, Fridley BL, Iversen ES, Chen Z, et al. Genome-wide Analysis of Common Copy Number Variation and Epithelial Ovarian Cancer Risk. Cancer Epidemiol Biomarkers Prev. 2019;28(7):1117-26.

11. Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665-74.

12. Tanner JA, Zhu AZ, Claw KG, Prasad B, Korchina V, Hu J, et al. Novel CYP2A6 diplotypes identified through next-generation sequencing are associated with in-vitro and in-vivo nicotine metabolism. Pharmacogenet Genomics. 2018;28(1):7-16.

13. Kim JH, Hu HJ, Yim SH, Bae JS, Kim SY, Chung YJ. CNVRuler: a copy number variation-based case-control association analysis tool. Bioinformatics. 2012;28(13):1790-2.

14. Chenoweth MJ, Ware JJ, Zhu AZX, Cole CB, Cox LS, Nollen N, et al. Genome-wide association study of a nicotine metabolism biomarker in African American smokers: impact of chromosome 19 genetic influences. Addiction. 2018;113(3):509-23.

15. Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, Sharp K, et al. The UK Biobank resource with deep phenotyping and genomic data. Nature. 2018;562(7726):203-9.

16. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. Am J Hum Genet. 2007;81(5):1084-97.

17. Ohnami S, Naruoka A, Isaka M, Mizuguchi M, Nakatani S, Kamada F, et al. Comparison of genetic susceptibility to lung adenocarcinoma and squamous cell carcinoma in Japanese patients using a novel panel for cancer-related drug-metabolizing enzyme genes. Sci Rep. 2022;12(1):17928.

18. Ramus SJ, Gayther SA. The contribution of BRCA1 and BRCA2 to ovarian cancer. Mol Oncol. 2009;3(2):138-50.

19. Reid BM, Permuth JB, Sellers TA. Epidemiology of ovarian cancer: a review. Cancer Biol Med. 2017;14(1):9-32.

20. DeVries AA, Dennis J, Tyrer JP, Peng PC, Coetzee SG, Reyes AL, et al. Copy Number Variants Are Ovarian Cancer Risk Alleles at Known and Novel Risk Loci. J Natl Cancer Inst. 2022;114(11):1533-1544.

Figure legends

Figure 1. **Schematic of computationally-inferred deletion regions and comparison to known *CYP2A6* SVs.** (A, B) Bars in the top panel indicate deletion regions computationally inferred from SNP array signal intensity data in (A) Reid et al. [10]; and (B) Walker et al. (the central gray bar represents the deletion region inferred in the majority of participants; white bars with a dotted border indicate the range of

other regions) [9]. (C) We inferred deletions for reference panel participants with *CYP2A6*4* or *CYP2A6*12* SVs using PennCNV and CNVruler, identifying 34 participants with predicted deletions in the region indicated (n=4 true *CYP2A6*4*; n=30 true *CYP2A6*12*). (D, E) Illustrations of the known deletion regions and resulting gene locus for (D) *CYP2A6*4* and (E) *CYP2A6*12* SVs. (F) *CYP2A7-CYP2A6* gene locus without SVs (i.e. *CYP2A6*1*). For detailed descriptions of the gene locus and structural variants, see PharmVar structural variant document (https://www.pharmvar.org/gene/CYP2A6).

Figure 2. **SV imputation reference panel creation flowchart**. (A) The original reference panel [7] included only 243 SNPs (of 1021 total reference panel SNPs) that overlapped with SNPs on the UK Biobank array (of 1659 total UK Biobank array SNPs)(GRCh37 chr19:39000000-43000000). Cross-validation of the reference panel limited to the 243 SNPs available in the UK Biobank resulted in 58% of SV alleles being positively identified (vs. 70% when all 1021 originally genotyped SNPs are included). (B) An updated reference panel including imputed SNPs was developed. This resulted in considerably more SNPs on the updated reference panel (1386 vs 243) overlapping with SNPs on the UK Biobank array (GRCh37 chr19:39000000-43000000). Cross-validation of the updated imputed SNP reference panel resulted in the recovery of the 70% positive identification rate of SV alleles.

Figure 3. ***CYP2A6* SV alleles and risk for ovarian or lung cancer.** (A) *CYP2A6* SV deleterious alleles were not associated with the risk of OC (OR=1.1; 95%CI: 0.80-1.37), where the frequency of having one or more *CYP2A6* SV alleles was not significantly different in controls (n=201390) vs. cases (n=1097). (B) *CYP2A6* SV alleles were associated with a lower risk of LC (OR=0.4; 95%CI: 0.29-0.64), where the frequency of having one or more *CYP2A6* SV alleles was significantly lower in LC cases (n=1040) vs. controls (n=40211). In SCC cases (n=270; a subset of LC cases), *CYP2A6* SV alleles were also associated with a lower risk of LC (vs. LC controls)(OR=0.2; 95%CI: 0.08-0.58). OC analyses restricted to females; LC analyses restricted to current smokers.

Acknowledgements: We acknowledge the work of Haidy Giratallah in obtaining and formatting UK Biobank data for analysis.

Author contributions: AWRL performed analyses and drafted the manuscript; AWRL, JGP, JK, MJC, and RFT conceived of the research and reviewed the manuscript.

Funding: This work was funded by a Canadian Institutes of Health Research (CIHR) Project grant (PJY-159710), National Institutes of Health (NIH) Grant PGRN DA020830, and a Canada Research Chair in Pharmacogenomics (Tyndale).

Ethical approval: Use of genetic data from imputation reference panel participants was approved at the University of Toronto and clinical trial sites where genetic material was collected.

Competing interests: The authors declare no conflicts of interest.