Genotyping, characterization, and imputation of known and novel *CYP2A6* structural variants using SNP array data

Alec W.R. Langlois[1,2]; Ahmed El-Boraie[1,2]; Jennie G. Pouget[2,3]; Lisa Sanderson Cox[4]; Jasjit S. Ahluwalia[5]; Koya Fukunaga[6]; Taisei Mushiroda[6]; Jo Knight[7]; 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]Department of Population Health, University of Kansas School of Medicine, Kansas City, KS 66160, USA

[5]Departments of Behavioral and Social Sciences and Medicine, Brown University School of Public Health, Providence, RI 02912, USA.

[6]Center for Integrative Medical Sciences, RIKEN; 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa, 230-0045, Japan

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

Corresponding Author: 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, E-mail: r.tyndale@utoronto.ca Phone: +1 (416) 978-6374

Conflicts of Interest: Some of the authors declared Financial and Non-Financial Relationships and Activities, and Conflicts of Interest regarding this manuscript as indicated in the supplementary materials.

Abstract

CYP2A6 metabolically inactivates nicotine. Faster CYP2A6 activity is associated with heavier smoking and higher lung cancer risk. The *CYP2A6* gene is polymorphic, including functional structural variants (SV) such as gene deletions (*CYP2A6*4*), duplications (*CYP2A6*1x2*), and hybrids with the *CYP2A7* pseudogene (*CYP2A6*12*, *CYP2A6*34*). SVs are challenging to genotype due to their complex genetic architecture. Our aims were to develop a reliable protocol for SV genotyping, functionally phenotype known and novel SVs, and investigate the feasibility of *CYP2A6* SV imputation from SNP array data in two ancestry populations.

European- (EUR; n=935) and African- (AFR; n=964) ancestry individuals from smoking cessation trials were genotyped for SNPs using an Illumina array and for *CYP2A6* SVs using Taqman copy number (CN) assays. SV-specific PCR amplification and Sanger sequencing was used to characterize a novel SV. Individuals with SVs were phenotyped using the nicotine metabolite ratio, a biomarker of CYP2A6 activity. SV diplotype and SNP array data were integrated and phased to generate ancestry-specific SV reference panels. Leave-one-out cross-validation was used to investigate the feasibility of *CYP2A6* SV imputation.

A minimal protocol requiring three Taqman CN assays for *CYP2A6* SV genotyping was developed and known SV associations with activity were replicated. The first domain swap *CYP2A6-CYP2A7* hybrid SV, *CYP2A6*53*, was identified, sequenced, and associated with lower CYP2A6 activity. In both EURs and AFRs, most SV alleles were identified using imputation (>70% and >60%, respectively); importantly, false positive rates were <1%. These results confirm that CYP2A6 SV imputation can identify most SV alleles, including a novel SV.

Introduction

Tobacco smoking remains a global problem, causing approximately eight million deaths annually(1). CYP2A6 is a genetically polymorphic enzyme metabolizing cancer (e.g. letrozole, tegafur) and HIV (e.g. efavirenz) therapeutics(2). CYP2A6 is responsible for the inactivation of nicotine to cotinine, and cotinine to 3'-hydroxycotinine (3'-HC)(3, 4). The nicotine metabolite ratio (NMR, the ratio of 3'-HC to cotinine) is a well-established biomarker of CYP2A6 activity in regular smokers(4). Faster CYP2A6 activity is associated with more cigarettes smoked per day, lower cessation rates, and higher risk of lung cancer(5,6,7,8,9,10).

CYP2A6 has important structural variants (SV), generated by non-allelic homologous recombination events with the pseudogene *CYP2A7*. SVs include gene deletions (*CYP2A6*4*), duplications (*CYP2A6*1x2*), and *CYP2A7-CYP2A6* hybrids (*CYP2A6*12, CYP2A6*34*)(Figure 1). SVs impact function, such that deletion, hybrid, and duplication SVs were associated with null, decreased, or increased function, respectively(11,12,13,14,15,16). SV allele frequencies differ substantially by ancestry(17).

In a weighted genetic risk score (wGRS) for CYP2A6 activity (i.e. the NMR) developed in a European-ancestry population (EUR), SV alleles contributed the strongest effects (vs. SNPs), particularly the common decreased-function *CYP2A6*12* hybrid variant(11). In a wGRS developed in an African-ancestry population (AFR), SVs also contributed strong effects, including the common *CYP2A6*4* deletion and *CYP2A6*1x2* duplication variants(12). Another polygenic risk score for the NMR in EUR, based on GWAS data alone, captured less NMR variation (<20%)(18) compared to the EUR and AFR wGRSs (>30%)(11, 12).

Due to high frequencies and functional impacts, SV genotyping should be performed when investigating *CYP2A6* genetics, but can be challenging. First, PCR genotyping assays are low-throughput, generally require two consecutive PCR reactions (gene-specific then SV-specific), and are limited by genetic

variation where primers anneal(19). Second, next-generation sequencing with short reads, while high-throughput, is of limited utility due to high DNA sequence identity between *CYP2A6* and *CYP2A7*, leading to read misalignment(20). Finally, GWAS genotyping arrays only detect SNPs and small indels. For example, large GWASs of lung cancer do not capture the variation contributed by SVs, thus underestimating the degree of risk associated with *CYP2A6*(ref. 8).

Taqman copy number (CN) assays targeting *CYP2A6* introns 1 and 7 produce SV calls highly concordant with expected effects on CYP2A6 activity(11), but cannot distinguish between certain SV diplotypes (Supplementary Table 1)(e.g. *CYP2A6*12* or *CYP2A6*34* hybrid SVs). Here, we characterized *CYP2A6* SVs using six available Taqman CN assays (*CYP2A6* introns 1, 2 and 7; *CYP2A7* exon 1, introns 2 and 7) in a large dataset of EUR and AFR participants in two smoking cessation clinical trials(21, 22). Both trials included baseline NMR, SNP array, SV genotype (from both PCR-based and Taqman CN assays), and deep exon sequencing data.

While Taqman CN assays can be used to genotype SVs, they remain low-throughput (24 samples in quadruplicate per 96-well plate, including controls) and expensive. An alternative approach is SV genotype imputation. Phased SNP array data can be integrated with SV genotypes, forming a reference panel that can be used to predict SV genotype in targets *with* SNP array data but *without* SV genotype data using imputation software. This approach has been used to accurately predict SV genotypes for several genes (e.g. *C4A/C4B*, *GYPA/GYPB, CYP2D6*, *HP*) in large clinical datasets, and associate those SVs with health outcomes(23,24,25,26). Large biobanks can be used to investigate associations between *CYP2A6* variants and a variety of health conditions. However, impactful SVs are not captured in SNP array or short-read next-generation sequencing data in biobanks, attenuating associations. Our first objective was to characterize CN patterns for all known (and potential novel) *CYP2A6* SVs and determine the minimum number of assays necessary for unambiguous genotyping of SV diplotypes. Second, we tested the association of established and newly identified *CYP2A6* SVs with CYP2A6 activity. Finally, we

investigated the feasibility of *CYP2A6* SV genotype imputation from SNP array data within two ancestry populations (EUR and AFR) with differing linkage disequilibrium (LD) structure.

Materials and Methods

Taqman CN Assay Genotyping

PNAT2 (EUR n=935; AFR n=506; NCT01314001)(27) and KIS3 (AFR n=458; NCT00666978)(28) participants previously underwent genotyping for *CYP2A6* SVs and SNPs using PCR assays (11, 12, 19), SNP array genotyping(29), and deep exon sequencing (PNAT2 only)(20), and were genotyped for *CYP2A6* SVs using Taqman CN assays (Thermo Fisher Scientific Inc., Waltham, MA) measuring *CYP2A6* CN at Introns 1 (Hs07545274_cn) and 7 (Hs07545275_cn).

An Applied Biosystems Viia 7 real-time PCR system (Thermo Fisher Scientific Inc., Waltham, MA) was used to carry out qPCR reactions. Individual samples were run in quadruplicate on fast 96-well microplates. Reaction volumes were 5µL, composed of 2.5µL GTXpress master mix (Thermo Fisher Scientific Inc., Waltham, MA; catalog number: 4401857), 1µL water, 1µL gDNA (5ng/µL), 0.25µL internal control Taqman CN reference assay (Thermo Fisher Scientific Inc., Waltham, MA*; RPPH1* (catalog number: 4403326) or *TERT* (catalog number: 4403316)*,* see Results), and 0.25µL target CN assay. Assay qPCR reactions used a 30sec denaturation step at 95°C, followed by 50 cycles of denaturation for 7sec at 95°C and annealing/extension for 30sec at 60°C.

CopyCaller v2.1 (Thermo Fisher Scientific Inc., Waltham, MA) was used to assign CN calls from qPCR results. Calls were made relative to controls with CN=2 (e.g. *CYP2A6*1/*1*). CN=1 controls (e.g. *CYP2A6*1/*4*) were used as a qualitative check for intra-plate reliability.

Following this initial screen, a subset of participants was characterized for *CYP2A6* SVs using additional Taqman CN assays in *CYP2A6, CYP2A7,* and *CYP2A13*. Additional assays targeted *CYP2A6* intron 2

(Hs04488984_cn), *CYP2A7* exon 1 and introns 2 and 7 (Hs07545276_cn, Hs04488016_cn, Hs07545277_cn, respectively), and *CYP2A13* intron 5/exon 6 (Hs03069103_cn). Individuals assessed were: (1) previously genotyped as having an SV (using Taqman CN or PCR assays); (2) identified as potentially having an SV through analysis of deep exon sequencing data using CoNVaDING v1.1.6(ref. 30); and (3) "false positives" for SVs during initial rounds of imputation cross-validation (see below).

Characterization of a novel SV (*CYP2A6*53*)

A novel SV was identified, designated *CYP2A6*53* by PharmVar(31). A *CYP2A6*4/**53 individual was selected for characterization of *CYP2A6*53* using nested PCR and Sanger sequencing. The first PCR reaction amplified *CYP2A6* from exon 3 to the 3' flanking region (NC_000019.9:g.41347784_41354528) using *CYP2A6*-specific primers (Supplementary Table 2). Two subsequent nested PCR reactions were specific for *CYP2A6*53*: first, using a *CYP2A6*-specific forward primer in exon 4 and *CYP2A7*-specific reverse primer in Intron 7 (generating the intron 4 product); second, using a *CYP2A7*-specific forward primer in Intron 7 and *CYP2A6*-specific reverse primer in the 3' flanking region (generating the 3' product)(Figure 1). Amplification of both products was observed in the *CYP2A6*4/*53* target, while no amplification of *CYP2A6*53* was observed in a *CYP2A6*4/*4* control.

The intron 4 and 3' products were extracted from 0.8% agarose gels with Midori Green DNA stain (Nippon Genetics Co., Ltd., Tokyo, Japan) using a QIAquick gel extraction kit (QIAGEN, Hilden, Germany). The intron 4 product was sequenced for intron 4 and exons 5-7, while the 3' product was sequenced for exons 8-9, and the 3'-UTR and flanking region. Sequencing primers were designed to be *CYP2A6*-, *CYP2A7*- or *CYP2A6/7*-specific as described in Supplementary Table 2.

Association between SVs and rate of nicotine metabolism

To examine associations between SVs and the rate of nicotine metabolism, we excluded individuals with other *CYP2A6* star alleles or non-synonymous SNPs. Star alleles and non-synonymous SNPs were

determined in PNAT2 participants (EUR and AFR) using deep exon sequencing; KIS3 participants (AFR) were genotyped for common star alleles in AFR (*CYP2A6*2, *9, *17, *20, *23, *25, *26, *27, *28,* and **35*) using PCR-based assays. Differences in mean log-transformed NMR between individuals with (e.g. *CYP2A6*1/*4, CYP2A6*1/*12)* and without (i.e. *CYP2A6*1/*1* reference group) SVs were assessed using ANOVA and post-hoc Dunnett tests for multiple comparisons. An additional analysis was performed comparing individuals with SVs to non-SV individuals, including individuals with other star alleles or non-synonymous SNPs. All statistical analyses were performed in GraphPad Prism 9.0.0 (GraphPad Software Inc., San Diego, CA).

Imputation reference panel creation

Participants in this study had previously undergone SNP genotyping using the Illumina HumanOmniExpressExome-8 v1.2 microarray, with 2688 custom SNP markers, as previously described(29). Separate EUR and AFR reference panels were created as SNP markers and LD differed by ancestry, and QC was performed separately(29).

Input VCF files included markers in a ±2 Mb window surrounding *CYP2A6* (NC_000019.9:g.39352000_43352000). Since SNP array genotyping algorithms assume three clusters of possible SNP genotype calls (homozygous reference, heterozygous, or homozygous variant)(32), they can produce unreliable results in individuals with SVs (i.e. an individual hemizygous for the reference allele would likely be genotyped as homozygous for the reference allele). Thus, we removed markers within the region which could be deleted or duplicated in *CYP2A6* SVs (NC_000019.9:g.41345000_41387000), as done for similar SV reference panels(25). SV genotypes were integrated with SNP array VCF files by adding a multiallelic entry representing SV genotype at an arbitrary position within *CYP2A6* (NC_000019.9:g.41352000). Integrated VCFs were then phased using Beagle 5.2, forming the reference panels(33).

CYP2A6 SV imputation reference panel cross-validation

The performance of the *CYP2A6* SV imputation reference panel was evaluated using leave-one-out cross validation(23) within EUR and AFR. A target individual was removed from the reference panel; the panel was then re-phased, followed by phasing and imputation of the target's SV diplotype in Beagle 5.2 with default settings except for "burnin=10" and "iterations=15". This was repeated until all individuals were tested as targets. Imputed SV diplotype calls were compared to "true" calls determined through Taqman CN assays. For each SV, the rate of positively identified alleles was calculated; overall false positive rates were also calculated (i.e. the proportion of non-SV alleles imputed as having an SV). Minor variation occurs across repeated trials; therefore, cross-validation was repeated 10 times and results were averaged across repeats.

Results

Spurious gene duplication calls due to *RPPH1* Taqman CN reference assay

Among all individuals genotyped for *CYP2A6* SVs, a subset of AFR individuals (n=18) appeared to have 3 copies (CN=3) for all assays (i.e. duplications at *CYP2A6, CYP2A7*, *and CYP2A13*); at face value, these results either suggest that all three genes had duplications or indicate an issue with the Taqman CNV assay. There are no known *CYP2A13* SVs*,* and recent literature has described high rates of spurious CN=3 calls in AFR resulting from common SNPs in *RPPH1*, the internal control targeted by the default Taqman CN reference assay(34). Thus, these individuals were re-genotyped using an alternative reference assay targeting *TERT*; calls at all seven assays for all individuals were revised to CN=2 (i.e. no SVs). In addition, 100% concordance of calls was confirmed between the *RPPH1* and *TERT* reference assays in other participants; thus, all subsequent genotyping used the *TERT* reference assay.

CYP2A6 SV genotyping

Known and novel *CYP2A6* SVs were characterized according to CN at three positions in both *CYP2A6* and *CYP2A7* using Taqman CN assays (Supplementary Table 1). It was determined that most known *CYP2A6* SV diplotypes can be genotyped unambiguously using the three *CYP2A6* assays (Hs07545274_cn, Hs07545275_cn, and Hs04488984_cn) with the exception of two pairs of diplotypes: *CYP2A6*1/*4* cannot be distinguished from *CYP2A6*34/*53,* and *CYP2A6*1/*1* cannot be distinguished from *CYP2A6*4/*1x2* (Table 1). Use of the *CYP2A7* assays in addition to the three *CYP2A6* assays does not aid in discriminating between these two ambiguous pairs (Supplementary Table 1).

To assign an SV diplotype in these cases, we calculated the expected frequencies of the indistinguishable diplotypes using observed SV allele frequencies (p^2 for *CYP2A6*1/*1;* 2pq for the rest). The expected frequency of *CYP2A6*1/*4* (0.04 in AFR; 0.006 in EUR) was higher than *CYP2A6*34/*53* (0.000002 in AFR; 0 in EUR), and the expected frequency of *CYP2A6*1/*1* (0.91 in AFR; 0.92 in EUR) was higher than *CYP2A6*4/*1x2* (0.0006 in AFR; 0.00005 in EUR). As *CYP2A6*1/**4 (>20000x) and *CYP2A6*1/*1* (>1500x) were much more likely, individuals with the two ambiguous CN patterns were coded as *CYP2A6*1/*4* and *CYP2A6*1/*1,* respectively.

SV genotyping in a large sample (EUR n=935; AFR n=964) allowed us to revise *CYP2A6* SV allele frequencies vs. frequencies previously determined using PCR-based genotyping assays (19) (Table 2). In particular, our updated data find a frequency of 0.013 for *CYP2A6*1x2* in AFR vs. 0.022 using the *RPPH1* reference assay(12).

<u>*CYP2A6*53* characterization</u>

During Taqman CN genotyping, a CN pattern was identified in some individuals which did not match any known *CYP2A6* SV (Supplementary Table 1). Analysis of deep exon sequencing using CoNVaDING v1.1.6(30) suggested that this SV resulted in the deletion of *CYP2A6* exons 5-9, and Taqman CN assay results suggested that *CYP2A6* was deleted at intron 7, while *CYP2A7* was duplicated at Intron 7.

Additionally, results from a PCR genotyping assay for the *CYP2A6*4H* allele(19) carried out in all participants indicated that the novel SV was *CYP2A7*-derived in Intron 7, but *CYP2A6*-derived in the 3' flanking region. The novel SV has been designated *CYP2A6*53* by PharmVar(31).

Characterization of *CYP2A6*53* was carried out in a *CYP2A6*4/*53* target participant. As *CYP2A6*4* is a full gene deletion, deep exon sequencing data in the target represented the *CYP2A6*53* allele. Reads from exons 1-4 were used; exons 5-9 had no reads, consistent with their expected deletion in *CYP2A6*53* and known deletion in *CYP2A6*4*. Within exons 1-4, there are 31 positions where nucleotide bases differ between *CYP2A6* and *CYP2A7;* in *CYP2A6*53,* 30 (97%) were identical to *CYP2A6*, demonstrating that exons 1-4 are *CYP2A6*-derived (Figure 2)(full sequence in Supplementary Figure 1).

Two PCR-amplified regions (the intron 4 and 3' products) were Sanger sequenced, including intron 4, exons 5-9, and the 3'-UTR plus flanking region of the *CYP2A6*53* allele. Among *CYP2A6* and *CYP2A7*, there are long segments of 100% nucleotide identity in intron 4 and the 3' flanking region where SV breakpoints were expected (bolded in Supplementary Figure 1). Sanger sequencing of Intron 4 of the target's *CYP2A6*53* allele indicated that the region 5' of the intron 4 segment with 100% identity (i.e. proximal to Exon 4) was nearly identical to *CYP2A6*, and the region 3' (i.e. proximal to Exon 5) was identical to *CYP2A7* (Figure 2).

Within exons 5-9, there are 22 positions where nucleotide bases differ between *CYP2A6* and *CYP2A7*. In *CYP2A6*53*, 18 (82%) of these positions were identical to *CYP2A7* and four (18%) were identical to *CYP2A6*. Thus, it was concluded that exons 5-9 are derived from *CYP2A7* in *CYP2A6*53* (Figure 2).

In the 3' flanking region, the region 5' of the segment with 100% identity (i.e. proximal to the 3'-UTR and Exon 9) was nearly identical to *CYP2A7* and the region 3' (i.e. proximal to the intergenic region) was identical to *CYP2A6*. Five additional unique SNPs were identified; all were synonymous or intronic (indicated by yellow lines in Figure 2).

Finally, Taqman CN assay data from individuals with the *CYP2A6*53* SV allele (including the *CYP2A6*4/*53* target) suggested the presence of a full copy of the *CYP2A7* gene in addition to the hybrid *CYP2A6*53* gene. This is a unique feature of *CYP2A6*53*, as other known hybrid SVs (*CYP2A6*12*, *CYP2A6*34*) are characterized by a single hybrid gene replacing *CYP2A6* and *CYP2A7* (Figure 1, Supplementary Table 1).

Overall, these data suggest that *CYP2A6*53* is a unique "domain swap" SV, where exons 1-4 are *CYP2A6*-derived, exons 5-9 are *CYP2A7*-derived, and the 3' flanking region is *CYP2A6*-derived. SV breakpoints are located in Intron 4 (*CYP2A6* to *CYP2A7*) and the 3' flanking region (*CYP2A7* back to *CYP2A6*), within the regions of 100% identity between *CYP2A6* and *CYP2A7* (Figures 1, 2). When translated, the target's *CYP2A6*53* allele would have eight (8/494; 2%) amino acid sequence changes relative to *CYP2A6*1,* and 25 (25/494; 5%) amino acid sequence changes relative to *CYP2A7*.

Association between *CYP2A6* SV and rate of nicotine metabolism

To examine the association between *CYP2A6* SVs and CYP2A6 activity (measured using the NMR), individuals without SVs (*CYP2A6*1/*1*) were compared to heterozygotes for each *CYP2A6* SV. Analyses were stratified by ancestry.

In EUR, SV heterozygote genotype was associated with logNMR ($p<0.0001$). Post-hoc tests found that logNMR in *CYP2A6*1/*4* (n=4; mean NMR: 0.20; $p<0.0001$), *CYP2A6*1/*12* (n=24; mean NMR: 0.25; $p<0.0001$), and *CYP2A6*1/*53* (n=5; mean NMR: 0.22; $p<0.001$) individuals was significantly different from *CYP2A6*1/*1* individuals (n=593; mean NMR: 0.47)(Figure 3A). *CYP2A6*1/*1x2* individuals' logNMR was not significantly different (n=4; mean NMR: 0.71; $p>0.05$).

In AFR, SV heterozygote genotype was associated with logNMR ($p<0.0001$). Post-hoc tests found that logNMR in *CYP2A6*1/*4* (n=27; mean NMR: 0.25; $p<0.0001$) and *CYP2A6*1/*12* (n=4; mean NMR: 0.16; $p<0.001$) individuals was significantly different from *CYP2A6*1/*1* individuals (n=395; mean NMR:

0.45)(Figure 3B). *CYP2A6*1/*1x2* individuals' logNMR was not significantly different (n=12; mean NMR: 0.62; p>0.05). Additional analyses in EUR and AFR focused exclusively on the impact of the SV (i.e. including individuals with other star alleles and non-synonymous SNPs) replicated significant main associations and most post-hoc test findings above (Supplementary Figure 2A and B).

CYP2A6 SV reference panel and imputation cross-validation

Overall, the EUR *CYP2A6* SV reference panel includes 1870 phased haplotypes and the AFR panel contains 1928 phased haplotypes (SV allele frequencies are given in Table 2). Visual representations of SNP haplotypes (surrounding *CYP2A6*) for the SV alleles included in the reference panel are shown in the upper panel of Figure 4, in addition to haplotype subgroups (i.e. subgroups of SV alleles algorithmically determined to be associated with each other) identified for *CYP2A6*1x2* and *CYP2A6*4.*

Leave-one-out cross-validation, repeated 10 times, was used to evaluate the utility of the reference panels for SV diplotype prediction. In EUR, 70% (52.4/74 SV alleles) of SV alleles were accurately imputed from array data (*CYP2A6*1x2:* 2.1/15; *CYP2A6*4*: 0.1/6; *CYP2A6*12*: 41.2/43; *CYP2A6*53*: 9/10). False positives were rare, occurring for <1% of *CYP2A6*1* alleles (4.9 called SV alleles/1796)(Figure 4A).

In AFR, 63% (53.5/85 SV alleles) of SV alleles were accurately imputed from array data (*CYP2A6*1x2:* 15.8/26; *CYP2A6*4*: 24.7/44; *CYP2A6*12*: 11/11; *CYP2A6*34:* 1.9/2; *CYP2A6*53*: 0.1/2). False positives were rare, occurring for <1% of *CYP2A6*1* alleles (7.9 called SV alleles/1843)(Figure 4B).

Discussion

In this study, we developed a protocol for reliable *CYP2A6* SV genotyping, corrected SV allele frequencies (i.e. by using the Taqman *TERT* CN reference assay), and sequenced and characterized the novel functional SV *CYP2A6*53.* We also corrected SV alleles accordingly (i.e. some *CYP2A6*4* calls

revised to *CYP2A6*53*), established associations between known and novel SVs with CYP2A6 activity, and determined that *CYP2A6* SV imputation from SNP array genotype data was feasible.

First, we established a Taqman CN assay-based protocol for unambiguous genotyping of all known *CYP2A6* SVs (in addition to a novel SV). This is important as these SVs have a major impact on CYP2A6 function (Figure 3). Taqman CN assays allow for considerably higher throughput and consume less gDNA than PCR-based genotyping approaches, which are SV- or SV sub-allele specific(19). We determined that use of the default Taqman *RPPH1* CN reference assay leads to over-calling of gene duplications among AFR populations, as seen before, due to high frequency of heterozygosity in *RPPH1*(ref. 34). As a result of switching from *RPPH1* to *TERT* in our study, SV diplotype calls were changed for 41% (n=18/44) of individuals previously identified as having a *CYP2A6*1x2* allele (see Table 2 for revised allele frequencies). Inaccurate SV genotyping rates of this magnitude in AFR populations will affect weighting in genetic risk scores and associations with phenotypes of interest (e.g. NMR, smoking, disease). Thus, studies using Taqman CN assays in an AFR population should use the *TERT* reference assay.

A novel SV, named *CYP2A6*53,* was discovered during SV genotyping. We found that the PCR genotyping assay for *CYP2A6*4H*(19) consistently mis-identifies *CYP2A6*53* alleles as *CYP2A6*4* alleles. This occurs because most *CYP2A6*4* sub-alleles represent a gene locus with a virtually intact copy of *CYP2A7* followed by a *CYP2A6*-derived 3'-flanking region (*CYP2A6* itself is deleted); thus, primers target *CYP2A7* Intron 7 and the *CYP2A6* 3'-flanking region(19). *CYP2A6*53* also has a *CYP2A7*-derived Intron 7 and *CYP2A6*-derived 3'-flanking region, causing spurious calls. Our new assessments indicate that *CYP2A6*53* (MAF=0.005) is more common than *CYP2A6*4* (MAF=0.003) in EUR (Table 2). Thus, prior studies using PCR assays likely overestimated the frequency of *CYP2A6*4,* especially in EUR.

Within individuals without other known functional variants, we found that CYP2A6 activity in *CYP2A6*1/*53* individuals was significantly lower than *CYP2A6*1/*1* individuals; mean NMR was similar

298 to *CYP2A6*1/*4* individuals (Figure 3). This suggests that the *CYP2A6*53* allele results in an unexpressed
299 or inactive enzyme. Thus, overestimation of *CYP2A6*4* frequency in prior studies likely had no impact on
300 findings. Furthermore, we replicated known associations of *CYP2A6*4* and *CYP2A6*12* with lower
301 CYP2A6 activity in EUR and AFR. Our additional analysis comparing SV heterozygotes while including
302 those with other star alleles or non-synonymous SNPs also replicated these associations (Supplementary
303 Figure 2), suggesting that, without SNP genotyping, SVs alone still have substantive impact on activity.

304 *CYP2A6*53* characterization revealed it to be a structurally unique domain swap SV, consisting of a
305 hybrid *CYP2A6-CYP2A7* gene and a full copy of *CYP2A7*. Other known *CYP2A6-CYP2A7* hybrids involve
306 the replacement of both *CYP2A6* and *CYP2A7* with a single hybrid gene. While *CYP2A7* does not encode
307 an active protein, its transcript is associated with increased expression of CYP2A6 *in vitro* achieved
308 through decreased binding of miR-126* to *CYP2A6*(ref. 35). Thus, *CYP2A6*53* may be associated with
309 higher expression of a heterozygote's intact *CYP2A6* allele and higher mean NMR than *CYP2A6*1/*34*
310 individuals (who lack *CYP2A7* in their *CYP2A6*34* allele).

311 Further SV genotyping for *CYP2A6* and *CYP2A7* should be performed, particularly in understudied
312 populations, to identify other potential novel SVs. *CYP2D6* is known to have highly complex structural
313 variation with domain swap hybrids similar to *CYP2A6*53* (i.e. with two apparent recombination
314 breakpoints)(36). Given that complex variation in *CYP2D6* arises from the presence of *CYP2D7* (an
315 inactive gene analogous to *CYP2A7*), similarly complex variation could exist in *CYP2A6*. *CYP2A6* SV
316 frequency varies between ancestry populations, so 1) more complex alleles and 2) different frequencies
317 of alleles may be found in non-EUR/non-AFR populations.

318 We developed EUR- and AFR-specific *CYP2A6* SV reference panels by combining SNP array genotypes
319 with the SV genotypes we obtained through Taqman CN assays; subsequent cross-validation indicated
320 that imputation was relatively accurate for SV prediction overall in EUR and AFR populations (>70% of SV

alleles identified in EUR; >60% in AFR). While not all SV alleles were identifiable (30% were not), low false positive rates (<1%) suggest that variant genotypes identified using this approach are usable in subsequent studies. The panel performed particularly well for more frequent SVs in EUR (*CYP2A6*12*) and AFR (*CYP2A6*1x2* and *CYP2A6*4*). However, some *CYP2A6*1x2* and *CYP2A6*4* alleles, which are rare in EUR, were not identified. The low rate of *CYP2A6*1x2* prediction is consistent with prior work suggesting that proximal SNPs tend to be in lower LD with duplications vs. deletions(37).

Investigation of Beagle's hidden Markov model parameters revealed which panel SV alleles were used to genotype targets. While *CYP2A6*12* targets tended to be associated with most other *CYP2A6*12* alleles on the panel, distinct allele clusters were identified for *CYP2A6*1x2* and *CYP2A6*4* (Figure 4). Thus, *CYP2A6*1x2* and *CYP2A6*4* may have been generated from multiple non-allelic homologous recombination events on different SNP haplotype backgrounds. This is consistent with the existence of several distinct *CYP2A6*4* sub-alleles(17) and is a potential explanation for poorer reference panel performance for these SVs. Furthermore, leave-one-out cross validation likely underestimates accuracy in external use, particularly in *CYP2A6*1x2* and *CYP2A6*4* where the removal of a single allele from the panel can represent the loss of a substantial portion of a cluster.

Reference panel expansion may improve accuracy, and thus simplify SV genotyping for prediction of CYP2A6 activity, particularly for clinical use of wGRS. Additionally, imputation may be useful for predicting *CYP2A6* SVs in large biobanks and external clinical trials to replicate *CYP2A6* SV associations with ovarian cancer risk(38, 39) and investigate probable associations with lung cancer and COPD. These approaches could also be applied more widely to other genes with SVs.

Given that *CYP2A6* SVs are known to occur in other populations, including Japanese, Chinese, Turkish, Alaska Native and American Indian, and Indian populations(40, 41, 42, 43, 44), additional reference panels for these populations are necessary. In particular, the CYP2A6*4 deletion is very frequent in East

Asian populations(40), where SV imputation could help capture a substantial portion of overall variation in CYP2A6 activity

While our *CYP2A6* SV imputation reference panel leverages proximal SNP and Taqman CN assay genotypes to impute SVs, array signal intensity data can also be used to predict SVs. PennCNV interprets signal intensity data using a Hidden Markov model without a reference panel, so SVs can be predicted using only SNP array data(45). However, PennCNV is limited in its ability to discriminate between SVs, as they are reported simply as deletions or duplications. This limits its use for accurate associations of SVs and phenotypes of interest. For example, *CYP2A6*12* is a decreased function variant and not a loss of function deletion like *CYP2A6*4*; PennCNV cannot specifically distinguish these variants.

Overall, we determined a new, minimal protocol for unambiguous *CYP2A6* SV genotyping using three commercial Taqman CN assays. During this work, a novel domain swap hybrid allele (*CYP2A6*53*) was identified, characterized, and determined to be inactive. Finally, we validated the use of SNP array data for imputation of the majority of CYP2A6 SV alleles in EUR and AFR.

Acknowledgements

We acknowledge the contributions of Dr. Caryn Lerman in the design and administration of the Pharmacogenetics of Nicotine Addiction Treatment 2 clinical trial (NCT01314001), of Dr. Michiaki Kubo in contributing to the deep exon sequencing and initial Taqman CNV assay genotyping of participants, and of Dr. Andrea Gaedigk and the *CYP2A6* Pharmvar expert panel in naming the novel structural variant, *CYP2A6*53*, and inspiring the *CYP2A6* structural variant schematics in Figure 1.

This research was undertaken, in part, thanks to funding from the Canada Research Chairs program (Tyndale, the Canada Research Chair in Pharmacogenomics), Canadian Institutes of Health Research (Tyndale, FDN-154294; Tyndale, Knight, and Chenoweth, PJY-159710); National Institute on Drug Abuse (Tyndale, PGRN U01-DA20830), and National Institutes of Cancer (Cox, NCI CA091912) as well as the

Centre for Addiction and Mental Health and the CAMH Foundation. Ahluwalia funded in part by P20GM130414, a NIH funded Center of Biomedical Research Excellence (COBRE).

Conflict of Interest

Some of the authors declared Financial and Non-Financial Relationships and Activities, and Conflicts of Interest regarding this manuscript as indicated in the supplementary materials.

Ethical Approval

Ethical approval for these analyses was obtained from the University of Toronto (PNAT2 REB number: 25510; KIS3 REB number: 38361) and at all original participating sites of the trials.

Data Availability Statement

Full exon and breakpoint DNA sequences for the *CYP2A6*53* SV are available in Supplementary Figure 1.

References

1. Lung Cancer Fact Sheet | American Lung Association 2019 [Available from: https://www.lung.org/lung-health-and-diseases/lung-disease-lookup/lung-cancer/resource-library/lung-cancer-fact-sheet.html.

2. 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.

3. Benowitz NL. Clinical pharmacology of nicotine: implications for understanding, preventing, and treating tobacco addiction. Clin Pharmacol Ther. 2008;83(4):531-41.

4. Dempsey D, Tutka P, Jacob P, Allen F, Schoedel K, Tyndale RF, et al. Nicotine metabolite ratio as an index of cytochrome P450 2A6 metabolic activity. Clin Pharmacol Ther. 2004;76(1):64-72.

5. Lerman C, Tyndale R, Patterson F, Wileyto EP, Shields PG, Pinto A, et al. Nicotine metabolite ratio predicts efficacy of transdermal nicotine for smoking cessation. Clin Pharmacol Ther. 2006;79(6):600-8.

6. 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.

7. 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.

8. Byun J, Han Y, Li Y, Xia J, Long E, Choi J, et al. Cross-ancestry genome-wide meta-analysis of 61,047 cases and 947,237 controls identifies new susceptibility loci contributing to lung cancer. Nat Genet. 2022;54(8):1167-77.

9. 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.

10. 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.

11. 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.

12. El-Boraie A, Chenoweth MJ, Pouget JG, Benowitz NL, Fukunaga K, Mushiroda T, et al. Transferability of Ancestry-Specific and Cross-Ancestry CYP2A6 Activity Genetic Risk Scores in African and European Populations. Clin Pharmacol Ther. 2020.

13. Nunoya K, Yokoi T, Kimura K, Inoue K, Kodama T, Funayama M, et al. A new deleted allele in the human cytochrome P450 2A6 (CYP2A6) gene found in individuals showing poor metabolic capacity to

412 coumarin and (+)-cis-3,5-dimethyl-2-(3-pyridyl)thiazolidin-4-one hydrochloride (SM-12502).
413 Pharmacogenetics. 1998;8(3):239-49.
414 14. Oscarson M, McLellan RA, Asp V, Ledesma M, Bernal Ruiz ML, Sinues B, et al. Characterization of
415 a novel CYP2A7/CYP2A6 hybrid allele (CYP2A6*12) that causes reduced CYP2A6 activity. Hum Mutat.
416 2002;20(4):275-83.
417 15. Rao Y, Hoffmann E, Zia M, Bodin L, Zeman M, Sellers EM, et al. Duplications and defects in the
418 CYP2A6 gene: identification, genotyping, and in vivo effects on smoking. Mol Pharmacol.
419 2000;58(4):747-55.
420 16. di Iulio J, Fayet A, Arab-Alameddine M, Rotger M, Lubomirov R, Cavassini M, et al. In vivo
421 analysis of efavirenz metabolism in individuals with impaired CYP2A6 function. Pharmacogenet
422 Genomics. 2009;19(4):300-9.
423 17. CYP2A6 allele nomenclature: Pharmacogene Variation Consortium; 2014 [Available from:
424 https://www.pharmvar.org/gene/CYP2A6.
425 18. Chen LS, Hartz SM, Baker TB, Ma Y, L Saccone N, Bierut LJ. Use of polygenic risk scores of
426 nicotine metabolism in predicting smoking behaviors. Pharmacogenomics. 2018;19(18):1383-94.
427 19. Wassenaar CA, Zhou Q, Tyndale RF. CYP2A6 genotyping methods and strategies using real-time
428 and end point PCR platforms. Pharmacogenomics. 2016;17(2):147-62.
429 20. Langlois AWR, El-Boraie A, Fukunaga K, Mushiroda T, Kubo M, Lerman C, et al. Accuracy and
430 applications of sequencing and genotyping approaches for CYP2A6 and homologous genes.
431 Pharmacogenet Genomics. 2022;32(4):159-72.
432 21. Cox LS, Faseru B, Mayo MS, Krebill R, Snow TS, Bronars CA, et al. Design, baseline characteristics,
433 and retention of African American light smokers into a randomized trial involving biological data. Trials.
434 2011;12:22.

22. Nollen NL, Cox LS, Yu Q, Ellerbeck EF, Scheuermann TS, Benowitz NL, et al. A clinical trial to examine disparities in quitting between African-American and White adult smokers: Design, accrual, and baseline characteristics. Contemp Clin Trials. 2016;47:12-21.

23. Sekar A, Bialas AR, de Rivera H, Davis A, Hammond TR, Kamitaki N, et al. Schizophrenia risk from complex variation of complement component 4. Nature. 2016;530(7589):177-83.

24. Leffler EM, Band G, Busby GBJ, Kivinen K, Le QS, Clarke GM, et al. Resistance to malaria through structural variation of red blood cell invasion receptors. Science. 2017;356(6343).

25. Häkkinen K, Kiiski JI, Lähteenvuo M, Jukuri T, Suokas K, Niemi-Pynttäri J, et al. Implementation of CYP2D6 copy-number imputation panel and frequency of key pharmacogenetic variants in Finnish individuals with a psychotic disorder. Pharmacogenomics J. 2022;22(3):166-72.

26. Boettger LM, Salem RM, Handsaker RE, Peloso GM, Kathiresan S, Hirschhorn JN, et al. Recurring exon deletions in the HP (haptoglobin) gene contribute to lower blood cholesterol levels. Nat Genet. 2016;48(4):359-66.

27. Lerman C, Schnoll RA, Hawk LW, Cinciripini P, George TP, Wileyto EP, et al. Use of the nicotine metabolite ratio as a genetically informed biomarker of response to nicotine patch or varenicline for smoking cessation: a randomised, double-blind placebo-controlled trial. Lancet Respir Med. 2015;3(2):131-8.

28. Cox LS, Nollen NL, Mayo MS, Choi WS, Faseru B, Benowitz NL, et al. Bupropion for smoking cessation in African American light smokers: a randomized controlled trial. J Natl Cancer Inst. 2012;104(4):290-8.

29. 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.

30. Johansson LF, van Dijk F, de Boer EN, van Dijk-Bos KK, Jongbloed JD, van der Hout AH, et al. CoNVaDING: Single Exon Variation Detection in Targeted NGS Data. Hum Mutat. 2016;37(5):457-64.

31. Gaedigk A, Casey ST, Whirl-Carrillo M, Miller NA, Klein TE. Pharmacogene Variation Consortium: A Global Resource and Repository for Pharmacogene Variation. Clin Pharmacol Ther. 2021;110(3):542-5.

32. Zhao S, Jing W, Samuels DC, Sheng Q, Shyr Y, Guo Y. Strategies for processing and quality control of Illumina genotyping arrays. Brief Bioinform. 2018;19(5):765-75.

33. 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.

34. Sicko RJ, Romitti PA, Browne ML, Brody LC, Stevens CF, Mills JL, et al. Rare Variants in RPPH1 Real-Time Quantitative PCR Control Assay Binding Sites Result in Incorrect Copy Number Calls. J Mol Diagn. 2022;24(1):33-40.

35. Nakano M, Fukushima Y, Yokota S, Fukami T, Takamiya M, Aoki Y, et al. CYP2A7 pseudogene transcript affects CYP2A6 expression in human liver by acting as a decoy for miR-126. Drug Metab Dispos. 2015;43(5):703-12.

36. Nofziger C, Turner AJ, Sangkuhl K, Whirl-Carrillo M, Agúndez JAG, Black JL, et al. PharmVar GeneFocus: CYP2D6. Clin Pharmacol Ther. 2020;107(1):154-70.

37. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, et al. Origins and functional impact of copy number variation in the human genome. Nature. 2010;464(7289):704-12.

38. 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.

39. 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.

40. Haberl M, Anwald B, Klein K, Weil R, Fuss C, Gepdiremen A, et al. Three haplotypes associated with CYP2A6 phenotypes in Caucasians. Pharmacogenet Genomics. 2005;15(9):609-24.

41. Pang C, Liu JH, Xu YS, Chen C, Dai PG. The allele frequency of CYP2A6*4 in four ethnic groups of China. Exp Mol Pathol. 2015;98(3):546-8.

42. Claw KG, Beans JA, Lee SB, Avey JP, Stapleton PA, Scherer SE, et al. Pharmacogenomics of Nicotine Metabolism: Novel CYP2A6 and CYP2B6 Genetic Variation Patterns in Alaska Native and American Indian Populations. Nicotine Tob Res. 2020;22(6):910-8.

43. Yadav VK, Katiyar T, Ruwali M, Yadav S, Singh S, Hadi R, et al. Polymorphism in cytochrome P4502A6 reduces the risk to head and neck cancer and modifies the treatment outcome. Environ Mol Mutagen. 2021;62(9):502-11.

44. Takeshita H, Hieda Y, Fujihara J, Xue Y, Nakagami N, Takayama K, et al. CYP2A6 polymorphism reveals differences in Japan and the existence of a specific variant in Ovambo and Turk populations. Hum Biol. 2006;78(2):235-42.

45. 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.

Titles and legends to figures

Figure 1. Schema of known *CYP2A6* SVs and proposed mechanism for generation of *CYP2A6*53*. Numbers in boxes indicate exons composing gene loci where blue boxes represent *CYP2A6* exons and red boxes represent *CYP2A7* exons; blue or red lines between exons indicate intronic, UTR, or flanking region sequence, and diagonal doubled black lines indicate shortened intergenic regions. Large X symbols indicate approximate breakpoints for *CYP2A6**53. *CYP2A6*53* is a novel SV composed of a normal copy of *CYP2A7* and a "domain swap" *CYP2A6-CYP2A7* hybrid (NC_000019.9:g.(41353205_41353410)_(41348573_41349232)delins(41385090_41385295)_(41380456_41381115); primers used for PCR amplification of the intron 4 and 3' products in *CYP2A6*53* are shown. The hypothesized reciprocal of produced by *CYP2A6*53* generation is shown but has not been observed. CYP2A6**1x2* is a duplication of *CYP2A6*; *CYP2A6*4* is a full deletion of *CYP2A6*; *CYP2A6*12* and *CYP2A6*34* are *CYP2A7-CYP2A6* hybrids. Schematics adapted from Pharmvar *CYP2A6* structural variation document (www.pharmvar.org/gene/CYP2A6). WT: wild-type.

Figure 2. Characterization of *CYP2A6*53* exon and breakpoint sequences. All exons were sequenced; exons 1-4 were sequenced using targeted deep exon sequencing, while exons 5-9 were PCR-amplified and Sanger sequenced. In addition, intron 4 and the proximal 3' flanking region were PCR-amplified and Sanger sequenced in order to resolve structural variant breakpoints. Regions derived from *CYP2A6* are shaded in blue, while those derived from *CYP2A7* are shaded in pink. Hatched pink and blue regions indicate regions of identity between *CYP2A6* and *CYP2A7* (i.e. origin cannot be determined). Vertical lines indicate relative positions where *CYP2A6* and *CYP2A7* sequences differ, where dark blue lines indicate identity with *CYP2A6*, dark red lines indicate identity with *CYP2A7*, and yellow lines indicate unique SNPs (i.e. no identity with *CYP2A6* or *CYP2A7*). Sequence alignment panels present evidence of *CYP2A6* or *CYP2A7* origin flanking homologous regions where breakpoints occur (shown in hashed pink and blue).

Figure 3. Comparison of NMR for individuals with no structural variants *(*1/*1*) against heterozygotes with structural variants in A. European-ancestry B. African-ancestry. Individual NMR values are plotted as points, and the mean NMR is indicated by horizontal grey line. Individuals with other non-SV star alleles and non-synonymous SNPs were excluded. ANOVA tests with post-hoc Dunnett test for multiple comparisons (vs. *CYP2A6*1/*1*) were run using log-transformed NMR (logNMR). ***p<0.001, ****p<0.0001

Figure 4. Structural variant allele SNP haplotype plots and reference panel leave-one-out cross-validation results. Plots of **A. European-ancestry** and **B. African-ancestry** individuals are heatmaps where the rows represent individual alleles, and columns represent SNPs (the first 100 SNP markers down- and upstream of *CYP2A6*); grey cells represent the reference allele at a SNP marker, while black cells represent the variant allele. Multiple distinct haplotype clusters were identified for *CYP2A6*4* and *CYP2A6*1x2*, and are numbered; "Other" represents non-clustered alleles. In pie charts, dark slices represent the proportion of SV alleles accurately identified through imputation; light slices represent the proportion of misidentified SV alleles (i.e., false negatives or prediction of another SV). Total sample size for each SV allele is indicated above pie charts and chart sizes are relative to the number of SV alleles.