The Pharmacogenomics Journal (2016), 1–3

www.nature.com/tpj

LETTER TO THE EDITOR

The *CYP2D6* VCF Translator

The Pharmacogenomics Journal advance online publication, 15 March 2016; doi:10.1038/tpj.2016.14

The cytochrome P450-2D6 (CYP2D6) hepatic oxidase is directly involved in the metabolism of ~25% of all commonly used drugs,[1–3] making it one of the most well-studied genes in the field of pharmacogenetics. As is likely well known by the readership of *The Pharmacogenomics Journal*, the *CYP2D6* gene is highly polymorphic, with over 100 variant star (*) alleles currently cataloged by the Human Cytochrome P450 (CYP) Allele Nomenclature Committee.[4] This committee was formed in 1999 to manage *CYP450* allele nomenclature using established sequence variation guidelines and to provide a summary of curated star (*) alleles and their effects when known;[5] however, the field of pharmacogenetics has been transformed by the recent advances in high-throughput sequencing technologies. Although next-generation sequencing has greatly facilitated pharmacogenomic discovery,[6–8] one of the resultant challenges is now reconciling the historical star (*) allele nomenclature with current genome reference assemblies and a more informed understanding of the extent of variation in the human genome.[9]

Interrogating *CYP2D6* by short-read sequencing technologies is challenging due to pseudogene sequence homology and the existence of copy number variant *CYP2D6* alleles.[10] In addition, one of the most cumbersome informatics issues is converting identified *CYP2D6* sequence variants from the genome assembly used for sequence analysis (for example, GRCh37/hg19) to the M33388.1 or AY545216.1 GenBank reference sequences used to define *CYP2D6* star (*) alleles (http://www.cypalleles.ki.se/cyp2d6.htm). This tedious but necessary process typically is accomplished by manual curation, but recently assisted by the *CYP2D6* haplotype tables available at PharmGKB (https://www.pharmgkb.org).

In an effort to simplify and automate this important step, we developed a *CYP2D6* variant call format (VCF) Translator tool that takes a standard VCF file (.vcf) limited to the *CYP2D6* coordinates from human reference hg19, including flanking sequences (chr22:42522071–42528563), and converts identified sequence variant coordinates to the M33388.1 and the AY545216.1 reference coordinates used to define *CYP2D6* star (*) alleles. Importantly, the reference coordinates in the *CYP2D6* VCF Translator have already been modified to include the necessary nucleotide corrections indicated by the Nomenclature Committee. Moreover, the *CYP2D6* VCF Translator corrects for the fact that the hg19 reference sequence actually contains *CYP2D6*2* and other common variants (for example, intron 1 conversion, 1661G>C, 2850C>T, 4180G>C and so on), so these nucleotides are automatically interrogated as if being compared with wild-type *CYP2D6*1*. In addition, it converts variants to their reverse complement (as *CYP2D6* is encoded on the negative strand), annotates variants with gene location (for example, exon and intron), reference/alternate alleles, dbSNP identifier (when available) and genotype calls based on VCF metrics (for example, wild-type, heterozygous and mutant), which can all be downloaded as a tab-delimited text (.txt.) file after executing the program (Figure 1).

The script was written in Python and implemented on a web server with PHP, and was designed to focus on the *CYP2D6* variants cataloged on the *CYP2D6* Nomenclature Committee website to facilitate subsequent star (*) allele conversion. To accomplish this, the *CYP2D6* VCF Translator extracts variant coordinates from an uploaded VCF and compares them to an annotated reference file of *CYP2D6* sequence variants with hg19, M33388.1 and AY545216.1 coordinates (ATG start codon = nucleotide 1). The *CYP2D6*2* correction is accomplished by reversing the *CYP2D6* nucleotides that are incorrectly variant in hg19 before making genotype calls from the VCF. Accordingly, any *CYP2D6*2* variants not called in a VCF because they are homozygous reference in hg19 are annotated in the output file as 'Mutant or Not Sequenced' (Figure 1). The *CYP2D6* VCF Translator also annotates novel sequence variants not listed on the Nomenclature Committee website.

The *CYP2D6* VCF Translator is run on the high-performance computer cluster at the Icahn School of Medicine at Mount Sinai (Minerva) and is freely available for investigators to use at http://stuartscottlab.org/vcf. Future iterations of the *CYP2D6* VCF Translator aim to infer star (*) allele diplotypes based on genotype data; however, in the interim we have found it to have substantial utility when converting high-throughput targeted *CYP2D6* sequencing data to star (*) alleles, as well as translating *CYP2D6* coordinates extracted from exome and whole-genome sequencing VCFs that were derived using hg19. As such, this letter is submitted to *The Pharmacogenomics Journal* to increase visibility of this online tool and, therefore, facilitate more simplified *CYP2D6* sequencing analyses for those investigators studying this very important pharmacogenetic gene.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGMENTS

This research was supported in part by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health (NIH) through grant K23 GM104401 (SAS), and the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

W Qiao[1], J Wang[1,2], BS Pullman[3], R Chen[1,2], Y Yang[1] and SA Scott[1]
[1]*Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA;*
[2]*Icahn Institute for Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, USA and*
[3]*Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA*
E-mail: stuart.scott@mssm.edu

REFERENCES

1 Gonzalez FJ, Skoda RC, Kimura S, Umeno M, Zanger UM, Nebert DW *et al.* Characterization of the common genetic defect in humans deficient in debrisoquine metabolism. *Nature* 1988; **331**: 442–446.
2 Gough AC, Miles JS, Spurr NK, Moss JE, Gaedigk A, Eichelbaum M *et al.* Identification of the primary gene defect at the cytochrome P450 CYP2D locus. *Nature* 1990; **347**: 773–776.

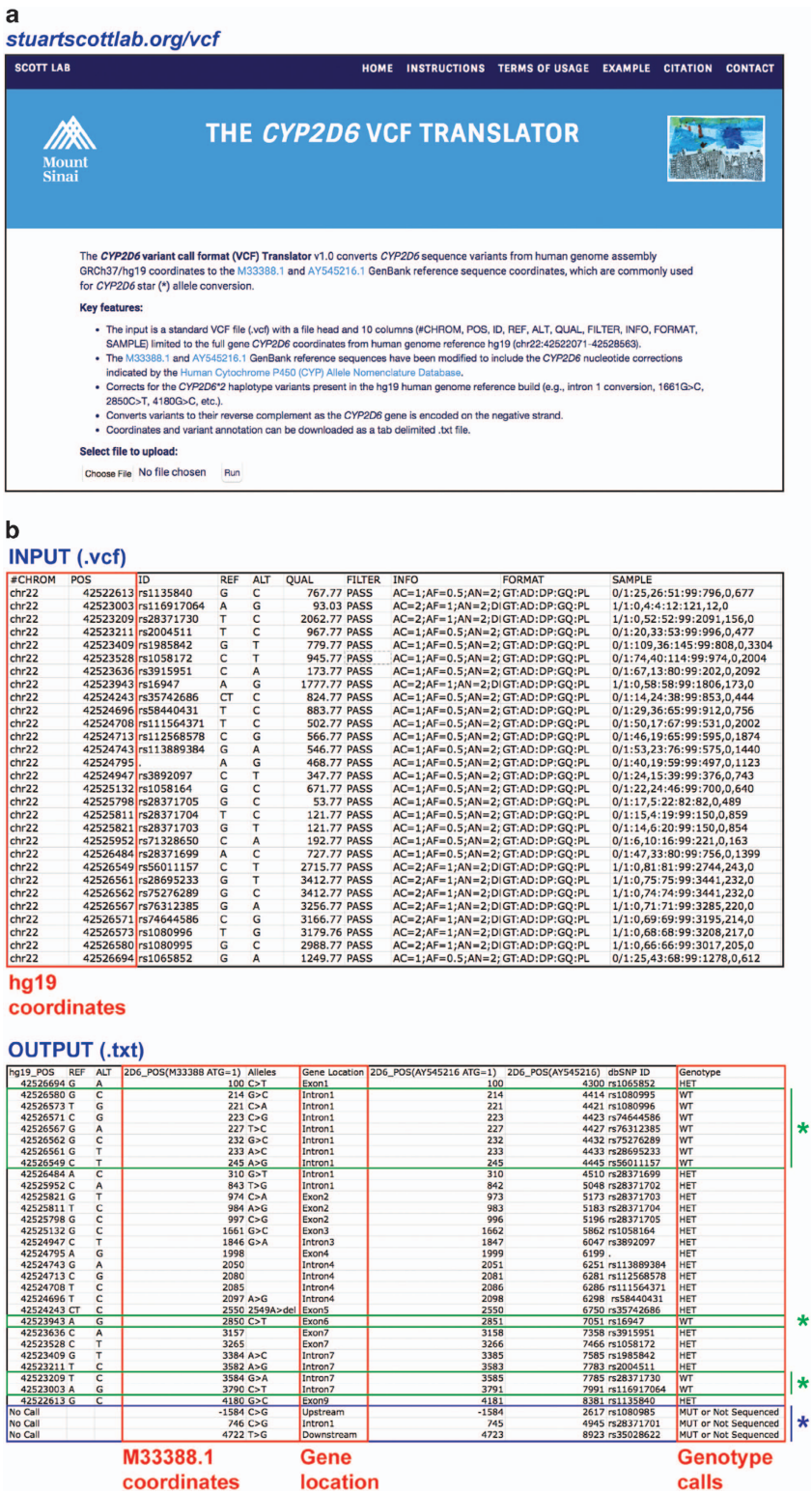

Figure 1. An overview of the *CYP2D6* variant call format (VCF) Translator with (**a**) a screenshot of the online homepage (http://stuartscottlab.org/vcf) and (**b**) example input (.vcf) and output (.txt) files. Highlighted in red boxes are some of the key components of the input and output files of an example VCF using DNA from the NA12878 HapMap cell line with a *CYP2D6*3/*4* diplotype sequenced through chr22:42522044–42527019 (hg19), or −225 to 4752 using M33388.1 coordinates. Note the variants in the output file annotated as 'Mutant or Not Sequenced': −1584C>G was not called in the VCF as it is outside of the sequenced region, and the common intronic 746C>G and downstream 4722T>G variants were not called in the VCF as they are found on both the **3* and **4* haplotypes, as well as in hg19. As such, these three variants were classified as 'Mutant or Not Sequenced' by the *CYP2D6* VCF Translator (blue asterisk). Similarly, the original VCF shows homozygous variants for the *CYP2D6* intron 1 conversion, 2850C>T, 3584G>A and 3790C>T; however, this is actually a reflection of those variants being present in hg19 and not in NA12878. As such, these variants were corrected to wild-type (WT) by the *CYP2D6* VCF Translator (green asterisks).

3 Owen RP, Sangkuhl K, Klein TE, Altman RB. Cytochrome P450 2D6. *Pharmacogenet Genomics* 2009; **19**: 559–562.
4 Sim SC, Ingelman-Sundberg M. The Human Cytochrome P450 (CYP) Allele Nomenclature website: a peer-reviewed database of CYP variants and their associated effects. *Hum Genomics* 2010; **4**: 278–281.
5 Nebert DW. Suggestions for the nomenclature of human alleles: relevance to ecogenetics, pharmacogenetics and molecular epidemiology. *Pharmacogenetics* 2000; **10**: 279–290.
6 Nelson MR, Wegmann D, Ehm MG, Kessner D St, Jean P, Verzilli C *et al.* An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. *Science* 2012; **337**: 100–104.
7 Daneshjou R, Gamazon ER, Burkley B, Cavallari LH, Johnson JA, Klein TE *et al.* Genetic variant in folate homeostasis is associated with lower warfarin dose in African Americans. *Blood* 2014; **124**: 2298–2305.
8 Gordon AS, Tabor HK, Johnson AD, Snively BM, Assimes TL, Auer PL *et al.* Quantifying rare, deleterious variation in 12 human cytochrome P450 drug-metabolism genes in a large-scale exome dataset. *Hum Mol Genet* 2014; **23**: 1957–1963.
9 Robarge JD, Li L, Desta Z, Nguyen A, Flockhart DA. The star-allele nomenclature: retooling for translational genomics. *Clin Pharmacol Ther* 2007; **82**: 244–248.
10 Gaedigk A. Complexities of CYP2D6 gene analysis and interpretation. *Int Rev Psychiatry* 2013; **25**: 534–553.