International Journal of Legal Medicine (2019) 133:807–810
https://doi.org/10.1007/s00414-018-1874-9

LETTER TO THE EDITOR

Exploring the 1000 Genomes Project haplotype reporting for the *CYP2D6* pharmacogene

Frank R. Wendt[1,2] · August E. Woerner[1,2] · Antti Sajantila[3] · Rodrigo S. Moura-Neto[4] · Bruce Budowle[1,2]

Received: 12 March 2018 / Accepted: 28 May 2018 / Published online: 2 June 2018

The Gaedigk et al. article "A perspective by PharmVar: Are the hundreds of *CYP2D6* haplotypes predicted by Wendt and colleagues real?" describes shortcomings of the 2017 Wendt et al. article "Full-gene haplotypes refine *CYP2D6* metabolizer phenotype inferences" [1]. To summarize, they discuss (1) the lack of submission of novel variants to www.PharmVar.org; (2) inaccurate activity score reporting, namely for those haplotypes containing the 843T>G SNP; (3) use of 1000 Genomes Project (1kGP) data from the inaccessible regions of the database; and (4) lack of sequence and structural validation for any of the described haplotypes.

We thank Gaedigk and colleagues for their review of the Wendt et al. 2017 findings and in many ways share their concerns. In general, the authors' letter raises valid concerns for the data presented in the original Wendt et al. study and many pharmacogenomics studies utilizing publically available data. However, the authors' appear to overstate our reported findings and seem to ignore where we already transparently discuss the major limitations of using such a database for this type of data exploration.

We summarize our responses to their concerns below. In general, we urge PharmVar to actively update its nomenclature table as to reflect most recent submitted findings. Additionally, we encourage PharmVar, its affiliates, and other pharmacogenomics researchers to release full-gene information as it becomes available, rather than only those sites relevant to the PharmVar nomenclature table(s) or the repository of knowledge for their respective gene(s) of interest. In doing so, the initiative described by Gaedigk and colleagues will continue to thrive.

The Wendt et al. paper was intended to explore the use of full-gene *CYP2D6* haplotype diversity in publically available data. A number of single-nucleotide polymorphisms from the 1kGP and the Wendt et al. study are not found on www.PharmVar.com. Gaedigk and colleagues note that PharmVar accepts submission of high-quality haplotype data. The Pilot Criteria of the 1kGP Phase3 Paired-end Accessible Regions are quite stringent, requiring "a depth of coverage between 8,960 and 35,840 inclusive (between one-half and twice the average depth) and that no more than 20% of covering reads have mapping quality zero" [2]. Indeed, read depth is a limiting factor for using data such as those of the 1kGP; however, Wendt et al. never recommended or even suggested that the data were high quality and be considered for submission to PharmVar. Such a recommendation would have been inappropriate and misleading to the community. Indeed, we stress in our paper that "empirical data are required to confirm their enzyme activity [of the resulting haplotypes]" and thus share similar concerns. Wendt et al. indicated that the relatively low sequencing read depth of the 1kGP is a major limitation of their findings. However, low read depth of pharmaco- and immunogenes does not warrant ignoring the public availability of 1kGP short-read data for exploratory purposes. It is a great resource used by many scientists for developing hypotheses and addressing probing questions.

There appears to be some confusion by Gaedigk et al. regarding the methods of Wendt et al. in which 1kGP haplotypes were characterized first using only those sites recognized and published on the PharmVar website (Human Cytochrome p450 Allele Nomenclature Database at the time of Wendt et al. analyses). Here, the consortium defines *CYP2D6*

✉ Frank R. Wendt
frw5010@gmail.com

[1] Center for Human Identification, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA

[2] Graduate School of Biomedical Sciences, University of North Texas Health Science Center, 3500 Camp Bowie Blvd., Fort Worth, TX 76107, USA

[3] Laboratory of Forensic Biology, Department of Forensic Medicine, University of Helsinki, P.O Box 40, 00014 Helsinki, Finland

[4] Instituto de Biologia, Universidade Federal do Rio de Janeiro, Rio de Janeiro 21941, Brazil

haplotypes and the functional consequences of each. These data were used to assign an activity score to each 1kGP sample. Second, Wendt et al. used multiple variant effect prediction algorithms to predict the function of all single-nucleotide variants including those recognized and empirically evaluated by PharmVar [3]. We used the activity score assigned to each participant, consistent with PharmVar, and evaluated the inclusion of additional variants in the haplotype, producing best- and worst-case predictions of functional impact on the enzyme.

The haplotypes containing the 843T>G SNP certainly confound the data presented in Wendt et al. and contradict existing literature. It should be noted that pathogenicity is quite difficult to predict and variant effect prediction algorithms have demonstrable error rates [4]. In fact, we urged caution of those haplotypes containing 843T>G as "the 843G SNP was incorrectly identified as damaging [by the variant effect predictors], emphasizing the importance of using [multiple] variant effect predictors with caution." Thus, Wendt et al. do not mislead the community regarding the functional consequence of 843T>G and emphasized caution in their discussion to infer functional consequences based on variant effect predictions. Indeed, Wang et al. [5] were cited who suggested revisiting definitions of certain *CYP2D6** alleles following interrogation of new haplotypes. As high-quality full-gene haplotype data are generated beyond the Wang et al. pediatric cohort, some previous definitions will likely evolve [5, 6].

The 3384A>C polymorphism was not detected in any of our haplotypes due to lack of reporting this locus by the 1kGP (Fig. 1). The 1kGP states that lack of genotype data for a locus indicates lack of detectable variation at the locus considering application of the strict-level stringency criteria. Personal communication with the HelpDesk suggests that there is no way to extract variant calling information for these sites. Gaedigk and colleagues raise an important concern that the alternate allele was not reported. This observation is also true for the genotype data at many other PharmVar loci, including −1109C>T, −960G>C, −629A>G, −98C>T, and 14C>T [3]. As described above and by Gaedigk and colleagues, the lack of variation at these loci also may be a result of low read depth at many sites in the genome. However, Gaedigk and colleagues allow us herein again to caution that the data are exploratory, and such databases do have limitations of which we were fully aware.

The SNP rs267608275 C-deletion is not indicated as a defining SNP in the PharmVar *CYP2D6* webpage [3] and therefore was not considered in the analysis using PharmVar recognized loci even though Gaedigk and colleagues indicate in vivo data for enzyme functionality for *CYP2D6**29. We use *29 as an example in our study but likely rs267608275delC may be observed in non-*29 haplotype conditions that have not been observed to date. Based on variant effect prediction for the locus, it is indicated as a damaging polymorphism. While in vivo data may contradict this finding, Wendt et al. reported best and worst-case enzyme function predictions. In doing so, the best-case scenario treats rs267608275 as lacking any functional consequence (as Gaedigk and colleagues discuss and the literature currently support). In contrast, it may be possible that in certain populations, currently unobserved or non-*29 haplotypes, or following exposure to emerging drug(s) that this locus may result in decreased activity of the enzyme [6], warranting a worst-case activity score of 0. Wendt et al. had no intention of mitigating the work that has gone into functional studies of *CYP2D6**29 or rs267608275 C-deletion, but instead incorporated additional functional possibilities, that may not have been observed yet by PharmVar, by using models generated from well-established, but certainly limited, variant effect predictors.

The M33388 reference sequence may not truly exist due to sequencing errors; however, this also may be due to lack of sampling enough individuals or enough individuals of the appropriate population(s) to observe the M33388 haplotype. Additionally, the *CYP2D6* community utilizes positional information from this reference to align variants [3, 7]. As such, it seems appropriate to report similarities to this reference sequence. It also should be noted that Wendt et al. aligned haplotypes to the M33388 and AY545216 reference sequences and the GRCh37 and GRCh38 reference genomes (Wendt et al. [1] Supplemental Table 2) to minimize

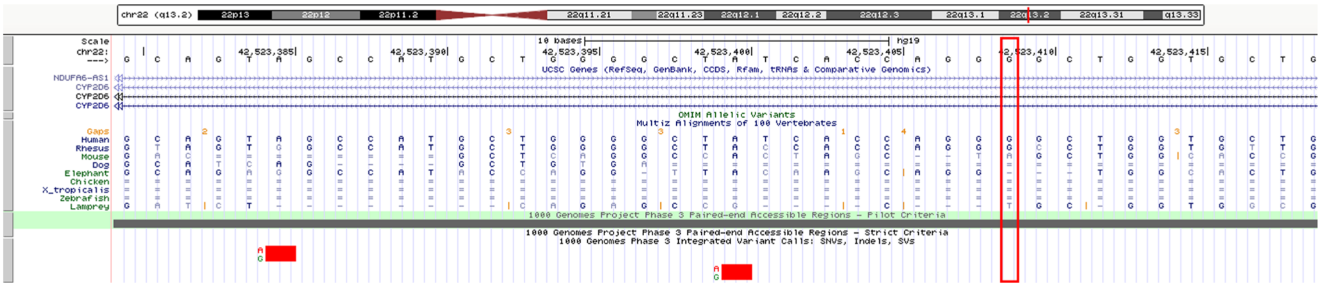
chr22 (q13.2)
22p13
22p12
22p11.2
22q11.21
22q11.23
22q12.1
22q12.2
22q12.3
22q13.1
22q13.2
22q13.31
q13.33
Scale
chr22:
10 bases
hg19
42,523,385
42,523,390
42,523,395
42,523,400
42,523,405
42,523,410
42,523,415
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
NDUFA6-AS1
CYP2D6
CYP2D6
CYP2D6
OMIM Allelic Variants
Multiz Alignments of 100 Vertebrates
Gaps
Human
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
1000 Genomes Project Phase 3 Paired-end Accessible Regions - Pilot Criteria
1000 Genomes Project Phase 3 Paired-end Accessible Regions - Strict Criteria
1000 Genomes Phase 3 Integrated Variant Calls: SNVs, Indels, SVs
A
G
A
G

Fig. 1 University of California Santa Cruz (UCSC) Table Browser screenshot showing the two single-nucleotide variants rs57175590 and rs28578778 (M33388:3408A>G and M33388:3393A>G, respectively) but not rs1985842 (M33388:3384T>G). The vertical red box indicates the position where one would expect to observe the rs1985842 locus

community confusion and maximize comparisons of 1kGP data with other reference sequences.

Gaedigk and colleagues are correct that the inaccessible regions of the genome are problematic. However, we disagree that they should be ignored or wholesale rejected. Ideally, the 1kGP should contain continuous reads and/or high read depth to provide substantially greater confidence in haplotype calls for pharmaco- and immunogenomic targets (e.g., CYP family enzymes, human leukocyte antigen, toll-like receptors, estrogen receptors, etc.). However, again from an exploratory perspective, there is no justification for rejecting the 1kGP "inaccessible regions" on the basis of low genotyping confidence. As described above, relatively stringent quality thresholds have been applied to the data, and certainly all of *CYP2D6* does not meet them. In fact, only 38/417 loci used to define our haplotypes fall below the pilot-level stringency application. In the discussion, Wendt et al. note that their findings should be used with caution, while emphasizing that poor genotyping and computational phase confidence of the 1kGP are issues. Our initial study does not ignore these limitations but instead is quite transparent about them.

Lastly, Wendt et al. state that "Although empirical data are required to confirm their enzyme activity, approximately 11% of healthy individuals *may* be wrongly identified as NMs according to traditional *CYP2D6* genotyping and activity score predictions." We stressed that empirical studies on full-gene haplotype interrogation are needed as opposed to recommending the findings be applied to immediate clinical and case-work applications. We agree that validating haplotype observations is essential to making strong claims about their clinical influence, and indeed the community is working towards such endeavors (which is indicated by Gaedigk and colleagues). Gaedigk et al.'s recommendation for validation by Sanger sequencing to verify the 1kGP haplotypes is beyond the scope of our 1kGP data exploration and does not overcome many of the issues presented by Gaedigk and colleagues related to *CYP2D6* structural variants or appropriate phase estimates. While Sanger sequencing certainly may be one avenue, there are a number of next generation sequencing technologies that can provide phasing and high read depth with much higher throughput. Moving forward, gene sequencing with massively parallel sequencing will more likely become the standard.

Gaedigk and colleagues highlight that copy number variation and adjacent gene rearrangements must be taken into consideration for the locus, and we agree with and did address this limitation of the 1kGP in stating that "Copy number variations (CNV) of some *CYP2D6** alleles and *CYP2D7* pseudogene conversion do occur in some individuals, namely UMs, and may influence the HWE and LD results [of this study]. It is likely that some 1000 Genomes Project individuals from the AFR super-population carry CNVs based on deviations from HWE expectations, but the project does not explore CNV in detail due to limitations of short-read sequencing. The data presented have been analyzed as though only two copies of *CYP2D6* are present in each individual…" we continued by stating that "A number of unique haplotypes have been identified that may be true haplotype observations but may also be attributed to duplication of two common haplotypes and/or *CYP2D7* pseudogene conversion. This is particularly true for the African populations which exhibit relatively frequent gene duplications." To indicate that "none of the latter were taken into account by Wendt and colleagues" is a gross misrepresentation of the discussions provided by Wendt et al.

Gaedigk and colleagues, on behalf of PharmVar, raised a number of concerns in which we are in full agreement and discussed in our paper. Their position about recommendations and use by Wendt et al. are unfounded and are contraindicative with the language and cautions raised within our study. It is quite possible that many of the haplotypes reported are not real due to limitations of the sequence data in the database, the computational phase reportedly provided by the 1kGP, and the variant effect prediction algorithms used by Wendt et al. to make predictions. The field and capabilities are expanding and we note that Shah and Gaedigk 2018 recently reinforced the claim that "pharmacogenetics was broadened by the observation that multifactorial genetic influences, in conjunction with environmental factors, usually determine drug responses," and that "it is wise to expect that, even after we have reached the goal to establish personalized medicine, we will not have eliminated all uncertainties" [8]. We presented the findings of Wendt et al. as exploratory, and they should be considered with the same caution and skepticism as any scientist would use to evaluate the merit of other articles. However, it is likely that substantially more *CYP2D6* haplotypes exist than are currently reported, especially considering the incorporation of full-gene data currently not reported by PharmVar. In fact, a recent report by Pratt et al. [6] discusses this phenomenon in regard to the *CYP2C19* locus. The community should continue to look at various avenues to expand knowledge of the *CYP2D6* pharmacogene.

References

1. Wendt FR, Sajantila A, Moura-Neto RS, Woerner AE, Budowle B (2017) Full-gene haplotypes refine CYP2D6 metabolizer phenotype inferences. Int J Legal Med. https://doi.org/10.1007/s00414-017-1709-0
2. 1000 Genomes Project Table Browser Track Settings. http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=658730751_1OUZr59IBjgjG7BMQr2qIad4nSf0&c=chr22&g=tgpPhase3Accessibility
3. Pharmacogene Variation Consortium (PharmVar) *CYP2D6* allele nomenclature. https://www.pharmvar.org/gene/CYP2D6
4. Mahmood K, Jung CH, Philip G, Georgeson P, Chung J, Pope BJ, Park DJ (2017) Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery

and diagnostics. Hum Genomics 11(1):10. https://doi.org/10.1186/s40246-017-0104-8

5. Wang D, Poi MJ, Sun X, Gaedigk A, Leeder JS, Sadee W (2014) Common CYP2D6 polymorphisms affecting alternative splicing and transcription: long-range haplotypes with two regulatory variants modulate CYP2D6 activity. Hum Mol Genet 23(1):268–278
6. Pratt VM, Del Tredici AL, Hachad H, Ji Y, Kalman LV, Scott SA, Weck KE (2018) Recommendations for clinical CYP2C19 genotyping allele selection: a report of the Association for Molecular Pathology. J Mol Diagn 20:269–276. https://doi.org/10.1016/j.jmoldx.2018.01.011
7. Qiao W, Wang J, Pullman BS, Chen R, Yang Y, Scott SA (2017) The CYP2D6 VCF Translator. Pharmacogenomics J 17(4):301–303. https://doi.org/10.1038/tpj.2016.14
8. Shah RR, Gaedigk A (2018) Precision medicine: does ethnicity information complement genotype-based prescribing decisions? Ther Adv Drug Saf 9(1):45–62. https://doi.org/10.1177/2042098617743393