Exploring the Kinh Vietnamese genomic database for the polymorphisms of the P450 genes toward precision public health

Diep Thi Hoang[a], Tran Van Hiep[b], Thao Thi Phuong Nguyen[c], Hoang Thi My Nhung[b,d], Kien Trung Tran[d], Le Sy Vinh[a*]

[a]*VNU University of Engineering and Technology, Vietnam National University Hanoi, Ha Noi, Vietnam*

[b]*VNU University of Science, Vietnam National University Hanoi, 334 Nguyen Trai, Hanoi, Vietnam.*

[c]*Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam*

[d]*Vinmec Research Institute of Stem Cell and Gene Technology, Ha Noi, Vietnam*

[*]Correspondence

Le Sy Vinh, University of Engineering and Technology, Vietnam National University Hanoi

144 Xuan Thuy, Cau Giay, Hanoi, Vietnam

Email: vinhls@vnu.edu.vn

Exploring the Kinh Vietnamese genomic database for the polymorphisms of the P450 genes toward precision public health

Abstract

Background: Human cytochrome P450 (*CYP*s) genes are essential in metabolizing drugs. Due to their high polymorphism, population-specific studies are of great interest.

Aim: This research examined the six *CYP* genes, including *CYP2B6, CYP2C9, CYP2C19, CYP2D6, CYP3A5,* and *CYP4F2* in the Kinh Vietnamese (KHV) for population-scale precision medicine.

Subjects and methods: We processed data from a genomics database of 206 healthy and unrelated KHV individuals to calculate *CYP* allele frequencies. First, we compared the *CYP* genes of the KHV to six other populations retrieved from the 1000 Genomes Project. Second, we searched the PharmGBK database for drug-CYP interaction data to compile a drug dosage recommendation for KHV.

Results: We observed diverging trends in the genetic variations of *CYP2B6*, *CYP2D6,* and *CYP3A5* in KHV. In terms of the phenotypic drug responses in KHV, *CYP2C19* exhibited all of the metabolic phenotypes at a non-trivial frequency. *CYP3A5* metabolized drugs at a lower rate than the other five *CYP*s.

Conclusion: This is the first large-scale study to investigate multiple *CYP* genes in the KHV for precision medicine from a public health perspective. Differences found in the distributions of metabolizers for the KHV suggest careful prescriptions for *CYP2C19* and *CYP3A5*-metabolized drugs.

Keywords: *CYP450*; pharmacogenetics; KHV; drug dosing; next-generation sequencing.

Introduction

Precision public health is an emerging multidisciplinary field that leverages precision medicine resources and findings to help enhance population health (Khoury & Holt 2021; Roberts et al. 2021). The guidelines for medication prescribing must be integrated with

pharmacogenetic discoveries specific to the patient's ethnicity to maximize therapeutic efficacy while minimizing adverse drug reactions (Koopmans et al. 2021).

The well-known Pharmacogenomics Knowledge Base (PharmGKB; http://www.pharmgkb.org) maintains a list of human genes that are significantly involved in drug metabolism or response (Whirl-Carrillo et al. 2012), of which the six *CYP* genes have the strongest evidence to support their importance are *CYP2B6, CYP2C9, CYP2C19, CYP2D6, CYP3A5,* and *CYP4F2*. According to the Clinical Pharmacogenetic Implementation Consortium (CPIC) guidelines, about 30% of common drugs are metabolized by these genes (Relling & Klein 2011). *CYP* genes are well-known as highly polymorphic, or varying greatly inter-individually and across populations (Lakiotaki et al. 2017). Therefore, population-specific studies on the genetic diversity of *CYP* genes are necessary to explain the variability in drug responses across populations (Sivadas & Scaria 2019).

In this study, we investigated six *CYP* genes for the Kinh Vietnamese population, including *CYP2B6, CYP2C9, CYP2C19, CYP2D6, CYP3A5,* and *CYP4F2*. We focused on their interactions with the popular drug classes composed by the CPIC. Despite there have been some studies on the polymorphism of single *CYP* genes in the KHV population, e.g., *CYP2C9* (Lee et al. 2005), *CYP2C19* (Vu et al. 2019), or *CYP2D6* (Nguyen et al. 2019), there has been no systematic study to investigate the multiple *CYP* genes in this population for precision public health until now.

Materials and methods

Study populations

We collected genotypes from the KHV database of the whole genomes of 206 healthy and unrelated Kinh Vietnamese people (Le et al. 2019) and haplotypes of 1,008 samples from six populations in the 1000 Genomes Project Phase 3 release (The 1000 Genomes Project

Consortium 2015) (1KG) focusing on Asian populations (see Table S1). Afterward, we used Beagle 5.1 software (Browning et al. 2018) to perform haplotype phasing for the six *CYP* genes in the KHV samples. We performed subsequent *CYP* star allele calling by referencing the Pharmacogene Variation Consortium (PharmVar) at www.PharmVar.org (Gaedigk et al. 2019).

Statistical analyses of allele frequencies

Data processing and statistical analysis were performed in RStudio version 1.2 (RStudio Team 2020). From derived allele frequencies, we assessed the Hardy-Weinberg equilibrium (HWE) by the Chi-square test (Crow 1999) to determine the population stratification.

We measured genetic distance by *CYP* genes for population pairs using the G_{ST} statistic (Nei & Chakravarti 1977). We then used the Chi-square test with Bonferroni correction (Dunn 1959) to identify *CYP* alleles with significantly different frequencies in KHV. Furthermore, to visualize the comparison of *CYP* allele distribution across the study populations, we also conducted the principal component analysis (PCA) (Pearson 1901) and built a heatmap using the CompleteHeatmap package (Gu et al. 2016).

Drug metabolizer analysis

We searched the PharmGKB database for predicted drug metabolism rates for the genotypes of the six *CYP* genes. We next estimated the distribution of drug metabolism rates for each gene in the KHV population. Following that, we identified genes in which a substantial proportion of KHV samples did not have a normal drug metabolic rate.

Results and Discussion

Statistical analyses of allele frequencies

The *1 alleles (called wild-type alleles) of all genes, except for *CYP2D6* and *CYP3A5,* had a high frequency in KHV and other populations (Table 1). Some alleles with notable frequencies in the KHV population were *CYP2B6*6* (25.5%), *CYP2C19*2* (26.2%), *CYP2D6*2* (10.2%), *CYP2D6*10* (52.7%), *CYP3A5*3* (65.0%), and *CYP4F2*3* (15.3%). The Chi-square tests for HWE (see Table S2 for p-values obtained) showed that the Hardy-Weinberg principle (α=0.05) was satisfied in KHV for all *CYP* genes in the present study.

[Table 1 near here]

We conducted the gene-based comparisons between populations (see estimated G_{ST} distances in Table S3) because previous research showed that genome-wide genetic distance across populations does not always agree with the distance assessed on drug-metabolizing genes (Koopmans et al. 2021). The 1KG populations from closest to farthest from the KHV were CHB, CHS, JPT, SAS, CEU, and YRI, in that order. This finding was supported by graphical representations of the PCA plot based on all genes (Figure S1A) and the heatmap (Figure S2).

Considering only the *CYP2B6* gene, KHV was closest to JPT (see Table S3) because the frequencies of *CYP2B6*1* and *CYP2B6*6* in KHV differed significantly from those in CHB (see the p-values in Table S4). When only the *CYP3A5* gene was considered, KHV appeared closest to SAS because *CYP3A5*3* is less frequent in KHV than in the three East Asian populations. Although KHV is close to the East Asian populations, no Asian population studied had the *CYP2D6*17* allele, which was found in KHV at a frequency of more than 1%. From the perspective of population-level precision medicine, comparative analyses will help in the development of the drug guidelines for the KHV by utilizing ones for other populations of high genetic similarity (Nordling 2017).

KHV metabolizer distributions

We obtained the distributions of drug metabolic phenotypes in the KHV for each examined *CYP* gene and compiled the dosing recommendations for KHV adults on common drug classes (see Table 2 for details) referencing CPIC.

[Table 2 near here]

We note that *CYP2C19* had a variety of metabolic phenotypes. Rapid metabolizers (RMs), normal or extensive metabolizers (EMs), intermediate metabolizers (IMs), and poor metabolizers (PMs) had corresponding frequencies of 7.8%, 41.8%, 39.3%, and 8.3%. The percentage of RM phenotypes is considerably high. Hence, for *CYP2C19*-dependent drug classes, especially those associated with both CYP2C19 and CYP2D6 enzymes, e.g., tricyclic antidepressants (Hicks et al. 2015), careful prescriptions would be advised.

Figure S3 shows that *CYP3A5* in KHV differed from the rest. It metabolized drugs at a lower rate than the other 5 *CYP*s. The highest proportions in the distribution of *CYP3A5* were PM (45.6%) and IM (38.8%). The poor metabolic rate of *CYP3A5* results in a high trough concentration, which improves the chance of achieving target concentrations. For *CYP3A5*-dependent drugs such as tacrolimus, PMs might be considered as normal metabolizers and prescribed at a standard dose (Birdwell et al. 2015). For EMs and IMs, the recommended dose should be 1.5 to 2 times higher than the standard dose, followed by therapeutic drug monitoring for the risk of adverse drug reactions (Birdwell et al. 2015).

Further clinical research on *CYP*-dependent drugs in KHV will be conducted to verify our study, to determine the functions of unknown alleles (especially of *CYP2D6*), and to improve the strength of clinical guidelines.

Funding

DTH and TTPN were financially supported by the Vietnam Academy of Science and Technology (VAST) under grant number ĐLTE00.01/19-20.

Disclosure statement

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

Data availability statement

The data that support the findings of this study are openly available in KHV database (available at https://genomes.vn) at https://doi.org/10.1002/humu.23835.

References

Birdwell KA, Decker B, Barbarino JM, Peterson JF, Stein CM, Sadee W, Wang D, Vinks AA, He Y, Swen JJ, et al. 2015. Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for CYP3A5 genotype and tacrolimus dosing. Clin Pharmacol Ther. 98(1):19–24.

Brown JT, Bishop JR, Sangkuhl K, Nurmi EL, Mueller DJ, Dinh JC, Gaedigk A, Klein TE, Caudle KE, McCracken JT, et al. 2019. Clinical Pharmacogenetics Implementation Consortium guideline for cytochrome P450 (CYP) 2D6 genotype and atomoxetine therapy. Clin Pharmacol Ther. 106(1):94–102.

Browning BL, Zhou Y, Browning SR. 2018. A one-penny imputed genome from next-generation reference panels. Am J Hum Genet. 103(3):338–348.

Crow JF. 1999. Hardy, Weinberg and language impediments. Genetics. 152(3):821–825.

Desta Z, Gammal RS, Gong L, Whirl-Carrillo M, Gaur AH, Sukasem C, Hockings J, Myers A, Swart M, Tyndale RF, et al. 2019. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2B6 and efavirenz-containing antiretroviral therapy. Clin Pharmacol Ther. 106(4):726–733.

Dunn OJ. 1959. Estimation of the medians for dependent variables. Ann Math Stat. 30(1):192–197.

Gaedigk A, Sangkuhl K, Whirl-Carrillo M, Twist GP, Klein TE, Miller NA, the PharmVar Steering Committee. 2019. The evolution of PharmVar. Clin Pharmacol Ther. 105(1):29–32.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 32(18):2847–2849.

Hicks JK, Bishop JR, Sangkuhl K, Muller DJ, Ji Y, Leckband SG, Leeder JS, Graham RL, Chiulli DL, LLerena A, et al. 2015. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2D6 and CYP2C19 genotypes and dosing of selective serotonin reuptake inhibitors. Clin Pharmacol Ther. 98(2):127–134.

Hicks JK, Sangkuhl K, Swen JJ, Ellingrod VL, Muller DJ, Shimoda K, Bishop JR, Kharasch ED, Skaar TC, Gaedigk A, et al. 2017. Clinical Pharmacogenetics Implementation Consortium Guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of tricyclic antidepressants: 2016 update. Clin Pharmacol Ther. 102(1):37–44.

Johnson JA, Caudle KE, Gong L, Whirl-Carrillo M, Stein CM, Scott SA, Lee MT, Gage BF, Kimmel SE, Perera MA, et al. 2017. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for pharmacogenetics-guided warfarin dosing: 2017 update. Clin Pharmacol Ther. 102(3):397–404.

Khoury MJ, Holt KE. 2021. The impact of genomics on precision public health: beyond the pandemic. Genome Med. 13(1):67.

Koopmans AB, Braakman MH, Vinkers DJ, Hoek HW, van Harten PN. 2021. Meta-analysis of probability estimates of worldwide variation of CYP2D6 and CYP2C19. Transl Psychiatry. 11(1):141.

Lakiotaki K, Kanterakis A, Kartsaki E, Katsila T, Patrinos GP, Potamias G. 2017. Exploring public genomics data for population pharmacogenomics. PLoS One. 12(8):1–18.

Le VS, Tran KT, Bui HTP, Le HTT, Nguyen CD, Do DH, Ly HTT, Pham LTD, Dao LTM, Nguyen LT. 2019. A Vietnamese human genetic variation database. Hum Mutat. 40(10):1664–1675.

Lee SS, Kim K-M, Thi-Le H, Yea S-S, Cha I-J, Shin J-G. 2005. Genetic polymorphism of CYP2C9 in a Vietnamese Kinh population. Ther Drug Monit. 27(2).

Nei M, Chakravarti A. 1977. Drift variances of FST and GST statistics obtained from a finite number of isolated populations. Theor Popul Biol. 11(3):307–325.

Nguyen HH, Ma TTH, Vu NP, Bach QTN, Vu TH, Nguyen TD, Nong H Van. 2019. Single nucleotide and structural variants of CYP2D6 gene in Kinh Vietnamese population. Medicine (Baltimore). 98(22).

Nordling L. 2017. How the genomics revolution could finally help Africa. Nature. 544(7648):20–22.

Pearson K. 1901. On lines and planes of closest fit to systems of points in space. Philos Mag. 2(11):559–572.

Relling M V, Klein TE. 2011. CPIC: clinical pharmacogenetics implementation consortium of the pharmacogenomics research network. Clin Pharmacol Ther. 89(3):464–467.

Roberts MC, Fohner AE, Landry L, Olstad DL, Smit AK, Turbitt E, Allen CG. 2021. Advancing precision public health using human genomics: examples from the field and future research opportunities. Genome Med. 13(1):97.

RStudio Team. 2020. RStudio: Integrated development environment for R.

Sivadas A, Scaria V. 2019. Population-scale genomics—Enabling precision public health. Adv Genet. 103:119–161.

The 1000 Genomes Project Consortium. 2015. A global reference for human genetic variation. Nature. 526(7571):68–74.

Theken KN, Lee CR, Gong L, Caudle KE, Formea CM, Gaedigk A, Klein TE, Agúndez JAG, Grosser T. 2020. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2C9 and nonsteroidal anti-inflammatory drugs. Clin Pharmacol Ther. 108(2):191–200.

Vu NP, Nguyen HTT, Tran NTB, Nguyen TD, Huynh HTT, Nguyen XT, Nguyen DT, Nong H Van, Nguyen HH. 2019. CYP2C19 genetic polymorphism in the Vietnamese population. Ann Hum Biol. 46(6):491–497.

Whirl-Carrillo M, McDonagh EM, Hebert JM, Gong L, Sangkuhl K, Thorn CF, Altman RB, Klein TE. 2012. Pharmacogenomics knowledge for personalized medicine. Clin Pharmacol Ther. 92(4):414–417.

Tables

Table 1: Allele frequencies in KHV and other populations focusing on alleles with a frequency greater than 1% in KHV population.

Allele	CEU	CHB	CHS	JPT	KHV	SAS	YRI
CYP2B6							
*1	55.6	80.6	80.9	73.1	66.8	47.4	38.9
*2	4.0	2.4	3.3	2.9	5.8	4.1	3.7
*6	26.8	14.6	15.2	19.7	25.5	37.4	40.3
CYP2C9							
*1	76.8	94.7	93.8	97.6	94.4	83.2	77.8
*3	6.6	3.9	4.8	1.9	2.9	10.9	ND
CYP2C19							
*1	63.1	59.2	59.0	60.1	65.5	47.9	49.1
*2	13.1	33.5	35.2	32.2	26.2	35.8	16.7
*3	ND	4.4	4.8	7.2	1.5	1.2	ND
*17	22.2	2.4	1.0	0.5	4.1	13.6	24.5
CYP2D6							
*1	42.4	21.8	23.3	50.0	25.7	41.4	28.2
*2	13.6	12.6	8.6	13.0	10.2	22.6	15.3
*10	1.5	57.8	60.5	36.1	52.7	5.2	5.1
*14	ND	0.5	ND	0.5	1.2	ND	ND
*17	ND	ND	ND	ND	2.2	ND	25.5
*41	12.1	3.4	4.8	0.5	3.2	12.1	0.9
CYP3A5							
*1	4.0	31.1	27.1	25.5	31.3	33.2	51.4
*3	95.0	68.9	70.5	74.5	65.0	66.8	14.4
CYP4F2							
*1	75.3	78.2	80.0	76.9	77.9	58.3	70.4
*2	14.1	8.3	6.7	5.8	5.8	14.2	4.2
*3	10.6	13.6	13.3	17.3	15.3	27.1	1.4

ND, not detected

Table 2: Drug dosing recommendation summary for KHV adults. Recommended doses (when the genetic testing is not available) are highlighted in bold.

Gene	Drugs (drug class)	Phenotype	Frequency in KHV (%)	Trough concentration	Side effect	Recommendation
CYP2B6	efavirenz (antiretroviral) (Desta et al. 2019)	RM	1	lower plasma concentration		standard dose
		EM	**54.9**	**normal**		**standard dose**
		IM	36.4	higher plasma concentration	mild	reduced dose
		PM	7.3	higher plasma concentration	moderate /severe	reduced dose
CYP2C9	celecoxib, lornoxicam (NSAIDs) (Theken et al. 2020)	**EM**	**92.2**	**normal**		**standard dose**
		IM[b]	6.8	higher plasma concentration	mild	lowest effective dosage
		PM	1	higher plasma concentration	very high	reduced dosage[c] or alternative drug
CYP2C19	citalopram (selective serotonin reuptake inhibitors) (Hicks et al. 2015); amitriptyline (tricyclic antidepressants[a]) (Hicks et al. 2017)	RM	7.8	lower plasma concentration		alternative drug
		EM	**41.8**	**normal**		**standard dose**
		IM	39.3	higher plasma concentration	mild	standard dose
		PM	8.3	higher plasma concentration	severe	reduced dose or alternative drug
CYP2D6	atomoxetine (Brown et al. 2019)	**EM**	**51.0[d]**	**normal**		**dose increasing from 40 mg/day to 80 mg/day in 2 weeks [e]**
		IM/PM	42.7	possibly higher	mild	
CYP3A5	tacrolimus (cytostatic immunosuppressants drug class) (Birdwell et al. 2015)	EM	11.7	lower plasma concentration	moderate	0.3 mg/kg/day
		IM	38.8	lower plasma concentration	mild	0.21 mg/kg/day
		PM	**45.6**	**normal**		**0.15 mg/kg/day**
CYP4F2	warfarin (anticoagulants drug class)[f] (Johnson et al. 2017)	**EM**	**71.4**	**normal**		**standard dose**
		IM	23.3	possibly higher		increase by 8-11%
		PM	3.4			

[a]: CYP2D6 has the same effects on tricyclic antidepressants as CYP2C19 does.

[b]: Divided into 2 groups: mildly and moderately reduced metabolism. Their difference is still in controversy.

[c]: 25-50% of the lowest recommended starting dose.

[d]: Regardless of indeterminate genotypes.

[e]: With EM, the dose is increased rapidly after 3 days, or increased gradually after 2 weeks with IM/PM.

[f]: Regardless of *CYP2C9* and *VKORC1*. With carriers of no function *CYP2C9* alleles, the dose should be decreased by 15-30%.

Supplementary materials:

Exploring the Kinh Vietnamese genomics database for the polymorphisms of the P450 genes toward precision public health

Materials and methods

Study populations

[Table S1 near here]

Results and Discussion

Allele frequencies and Hardy-Weinberg equilibrium

[Table S2 near here]

Principal component analyses

Figure S1 shows the PCA plots for all genes as well as six individual genes. On *CYP2B6, CYP2D6,* and *CYP2C19* plots, the KHV was separate from other populations. On other genes, KHV, CHB, CHS, and JPT formed a cluster that was far away from the rest (CEU, SAS, and YRI).

[Figure S1 near here]

The total proportion of "variance explained" in PCA greatly varied among the 6 genes (Figure S1B-G). We observed that this proportion was related to the number of alleles in the gene. The number of alleles is the number of dimensions of the data space. Accordingly, with more alleles, data space has more dimensions, more complexity, and loses more information when converted into a two-dimensional space (Pearson 1901). Therefore, a PCA plot with a lower proportion of "variance explained" belongs to a gene with more alleles. According to this, *CYP2D6*, with the smallest proportion explained (73.4%), appeared to be the most diverse gene.

Heatmap visualization for the allele frequency patterns

A heatmap (see Figure S2) depicts how significantly frequent the *CYP* alleles were across the seven populations among the 20 common *CYP* alleles in the KHV. The algorithm of the CompleteHeatmap package organizes populations hierarchically based on their similarity in allele frequencies (Gu et al. 2016). There were three groupings, each with strong in-group similarities: CHB-CHS-KHV-JPT, CEU-SAS, and YRI.

Figure S2 demonstrates the allele frequency distribution in each population for alleles presented in Table 1 (Main Text). The red color represents alleles of frequency higher than or equal to 50%, and the blue color represents alleles of frequency below 50%. Each row is for an allele, while each column is for a population. The row hierarchy classified common alleles into 3 groups with high, moderate, and low frequencies in all populations.

[Figure S2 near here]

Genetic differences in terms of G_{ST}

In terms of the G_{ST} distances estimated pairwise between KHV and other populations (see Table S3), CHB had the shortest G_{ST} distances to KHV by the three genes *CYP2C9*, *CYP2C19*, and *CYP2D6*. JPT, on the other hand, was of the smallest G_{ST} distances to KHV by the two genes *CYP2B6* and *CYP4F2*. SAS had the shortest G_{ST} distance to KHV by *CYP3A5*. Based on the average G_{ST} distances over six genes, the populations sorted from the nearest to the most distant to KHV were CHB, CHS, JPT, SAS, CEU, and YRI.

[Table S3 near here]

The heatmap and G_{ST} on the *CYP2D6* gene both clearly show that the three closest populations to KHV were CHB, CHS, and JPT. Of the 6 *CYP* genes, *CYP2D6*

had the strongest impact on the ranking of 3 East Asian populations by distance to KHV.

Since *CYP4F2* in JPT was among the 6 cases that did not satisfy the HWE (Table S2), we do not discuss the implications of JPT yielding the smallest G_{ST} value by the *CYP4F2* gene in Table S3.

Allele frequency comparison

In Table S4, we summarized the p-values obtained by pairwise Chi-square tests to compare the frequency of a *CYP* allele between KHV and another population. The p-values are adjusted for the Bonferroni correction. We wanted to identify alleles that significantly separate KHV from others (i.e., a p-value < 0.05 indicates a significant difference between two populations). Most of the signals for the discrepancy were in the columns of YRI, SAS, and CEU. Meanwhile, in *CYP2C9*, *CYP3A5*, and *CYP4F2*, three other columns (CHB, CHS, and JPT) showed no statistically significant difference from KHV. In *CYP2B6*, only CEU and JPT were similar to KHV. In *CYP2C19* and *CYP2D6,* only CHB and CHS were similar to KHV.

[Table S4 near here]

The tests revealed nine alleles that KHV significantly differed from at least three other populations. For example, the frequency of *CYP2D6*17* in the KHV population significantly differed from all other populations studied.

KHV metabolizer distributions

The metabolizer distribution of six CYP enzymes in KHV was visualized in Figure S3 (see Table S5 for details). Intermediate metabolizers (IMs) and extensive metabolizers (EM, considered as the normal rate of metabolism) occupied predominantly.

[Figure S3 near here]

[Table S5 near here]

CYP2C9 had unique metabolic rate distributions, of which about 92 out of 100 persons were EM, and the rest were IM and PM. Therefore, *CYP2C9*-dependent drugs such as non-steroidal anti-inflammatory drugs should start with standard doses for the majority of Kinh Vietnamese patients.

CYP2B6, CYP2D6, and *CYP4F2* expressed the predominance of both EM and IM. These genes encode CYP enzymes that metabolize several important drug classes. For example, *CYP2B6* is involved in the metabolism of antiretroviral drugs (Desta et al. 2019), *CYP2D6* is involved in the metabolism of some psychotropic drugs (Stingl & Viviani 2015), and *CYP4F2* is involved in the metabolism of warfarin (Johnson et al. 2017). Therefore, for the majority of Kinh Vietnamese patients prescribed these drug classes, a plausible approach would be a standard initial dose, and follow-up monitoring and reduced dose adjustment would be recommended.

References

Desta Z, Gammal RS, Gong L, Whirl-Carrillo M, Gaur AH, Sukasem C, Hockings J, Myers A, Swart M, Tyndale RF, et al. 2019. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for CYP2B6 and efavirenz-containing antiretroviral therapy. Clin Pharmacol Ther. 106(4):726–733.

Gu Z, Eils R, Schlesner M. 2016. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics. 32(18):2847–2849.

Johnson JA, Caudle KE, Gong L, Whirl-Carrillo M, Stein CM, Scott SA, Lee MT, Gage BF, Kimmel SE, Perera MA, et al. 2017. Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline for pharmacogenetics-guided warfarin dosing: 2017 update. Clin Pharmacol Ther. 102(3):397–404.

Pearson K. 1901. On lines and planes of closest fit to systems of points in space. Philos Mag. 2(11):559–572.

Stingl J, Viviani R. 2015. Polymorphism in CYP2D6 and CYP2C19, members of the cytochrome P450 mixed-function oxidase system, in the metabolism of psychotropic drugs. J Intern Med. 277(2):167–177.

Tables

Table S1: The number of individuals in the 7 populations under the present study.

Source	Population	Number of individuals
KHV database	Kinh Vietnamese (KHV)	206
1KG	Utah residents with Northern and Western European ancestry (CEU)	99
	Han Chinese in Beijing, China (CHB)	103
	Southern Han Chinese (CHS)	105
	Japanese in Tokyo, Japan (JPT)	104
	South Asian Ancestry (SAS)	489
	Yoruba in Ibadan, Nigeria (YRI)	108

Table S2: P-values of the Chi-square tests for Hardy Weinberg equilibrium. Significant p-values are reported in bold.

	CEU	CHB	CHS	JPT	KHV	SAS	YRI
CYP2B6	0.372	0.719	0.095	0.341	0.674	0.998	0.223
CYP2C9	**0.019**	0.856	0.112	0.957	0.497	0.806	0.427
CYP2C19	0.563	0.062	0.061	0.581	0.597	0.914	0.067
CYP2D6	**0.003**	0.078	0.074	0.197	0.144	1.000	**0.019**
CYP3A5	0.881	0.351	0.922	0.163	0.132	0.644	**0.012**
CYP4F2	0.594	0.697	0.750	**0.007**	0.572	0.904	**0.017**

Table S3: Pairwise G_{ST} distances between KHV and other populations. Smallest observed values for each gene are reported in bold.

	CYP2B6	*CYP2C9*	*CYP2C19*	*CYP2D6*	*CYP3A5*	*CYP4F2*	Mean	SD
CEU	0.0059	0.0306	0.0239	0.0947	0.1245	0.0063	0.0433	0.0511
CHB	0.0194	**0.0002**	**0.0050**	**0.0021**	0.0008	0.0006	**0.0053**	0.0073
CHS	0.0190	0.0008	0.0070	0.0032	0.0026	0.0006	0.0061	**0.0067**
JPT	**0.0045**	0.0036	0.0054	0.0343	0.0071	**0.0003**	0.0085	0.0170
SAS	0.0226	0.0230	0.0215	0.0888	**0.0004**	0.0309	0.0285	0.0395
YRI	0.0409	0.0276	0.0323	0.0880	0.1107	0.0143	0.0518	0.0400

Table S4: Bonferroni adjusted p-values of pairwise Chi-square tests to compare allele frequencies between KHV and other populations. Significant p-values are reported in bold.

Allele	CEU	CHB	CHS	JPT	SAS	YRI
CYP2B6						
*1	0.1999	**0.0163**	**0.0102**	1	**< 10^{-4}**	**< 10^{-4}**
*2	1	1	1	1	1	1
*6	1	**0.0354**	0.1027	1	**0.0006**	**0.004**
CYP2C9						
*1	**< 10^{-4}**	1	1	1	**< 10^{-4}**	**< 10^{-4}**
*3	1	1	1	1	**< 10^{-4}**	ND
CYP2C19						
*1	1	1	1	1	**< 10^{-4}**	**0.0016**
*2	**0.0057**	1	0.4352	1	**0.0099**	0.2321
*3	ND	0.5585	0.2844	**0.0035**	1	ND
*17	**< 10^{-4}**	1	1	0.6001	**< 10^{-4}**	**< 10^{-4}**
CYP2D6						
*1	**0.0016**	1	1	**< 10^{-4}**	**< 10^{-4}**	1
*2	1	1	1	1	**< 10^{-4}**	1
*10	**< 10^{-4}**	1	1	**0.0027**	**< 10^{-4}**	**< 10^{-4}**
*14	ND	1	ND	1	ND	ND
*17	ND	ND	ND	ND	ND	**< 10^{-4}**
*41	**0.0007**	1	1	1	**< 10^{-4}**	1
CYP3A5						
*1	**< 10^{-4}**	1	1	1	1	**< 10^{-4}**
*3	**< 10^{-4}**	1	1	0.4503	1	**< 10^{-4}**
CYP4F2						
*1	1	1	1	1	**< 10^{-4}**	0.8746
*2	**0.0253**	1	1	1	**0.0002**	1
*3	1	1	1	1	**< 10^{-4}**	**< 10^{-4}**

ND, not detected

Table S5: Frequencies in KHV of *CYP* phenotypic drug responses: poor metabolizers (PMs), intermediate metabolizers (IMs), normal or extensive metabolizers (EMs), and rapid metabolizers (RMs). N/A means not applicable. Phenotypes with highest percentage are reported in bold.

Gene	PM	IM	EM	RM	N/A
CYP2B6	7.3	36.4	**54.9**	1.0	0.4
CYP2C9	1.0	6.8	**92.2**	0.0	0.0
CYP2C19	8.3	39.3	**41.8**	7.8	2.8
CYP2D6	0.5	42.2	**51.0**	0.0	6.3
CYP3A5	**45.6**	38.8	11.7	0.0	3.9
CYP4F2	3.4	23.3	**71.4**	0.0	1.9

Figures

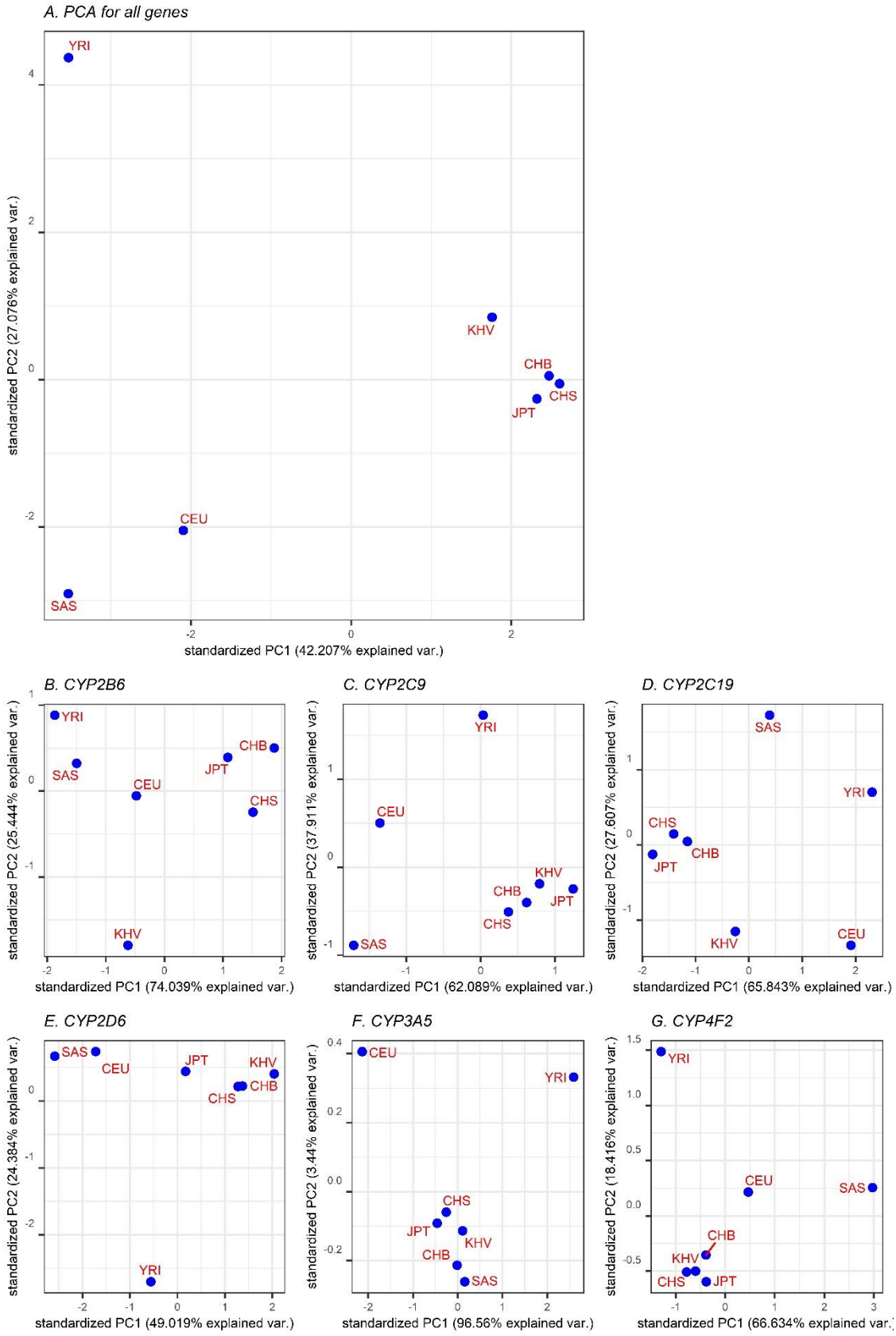

Figure S1: The principal component analyses of alleles for all genes as well as 6 individual genes.

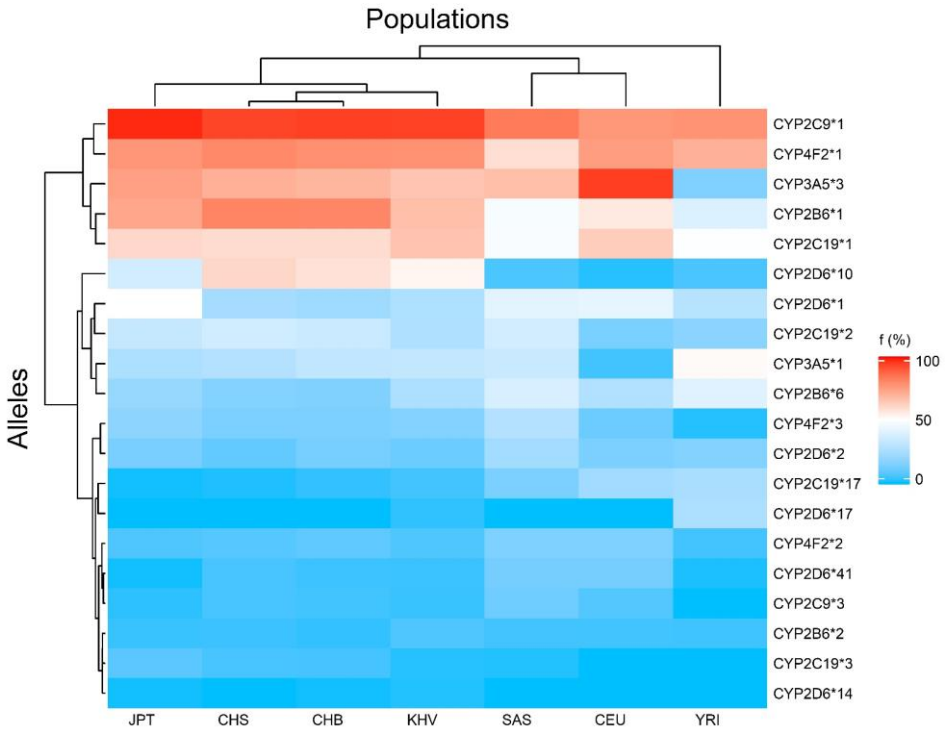

Figure S2: Heatmap for common alleles (with frequency higher than 1%) in the KHV. Red color denotes highly frequent alleles (frequency >50%). Blue color denotes alleles of frequency lower than 50%. The map showed that wild alleles were predominant. Besides, the horizontal dendrogram showed the hierarchical clustering of the seven populations, in which the KHV was placed closest to CHB and CHS.

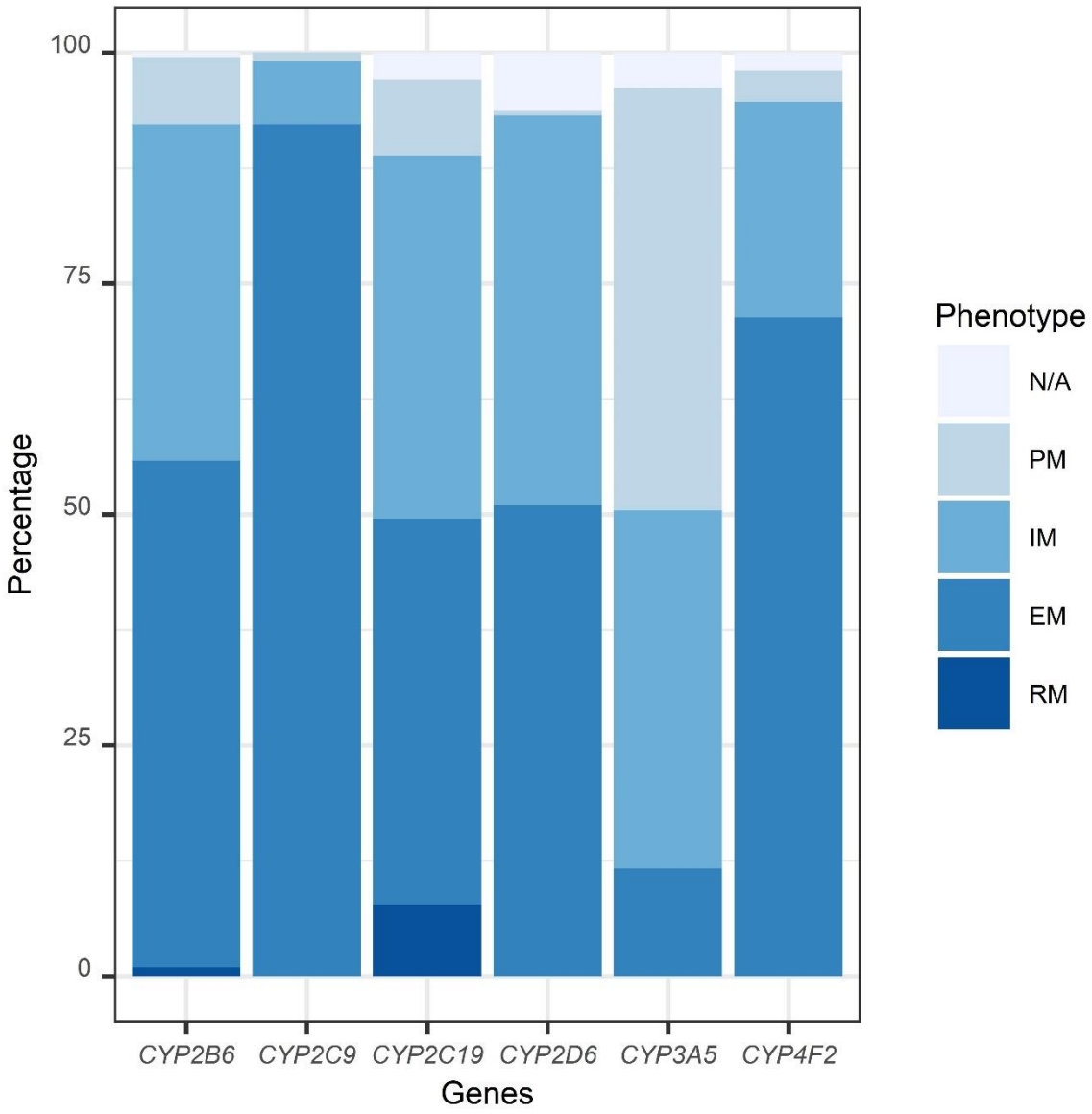

Figure S3: *CYP* metabolizer proportion in KHV. N/A means not applicable. In most of the genes, EM was the most frequent. PM and IM were of high proportions in *CYP2B6*, *CYP2C19* and *CYP3A5*. *CYP2C19* showed a non-trivial proportion of RM.