scientific reports

OPEN

The variation landscape of *CYP2D6* in a multi-ethnic Asian population

Yusuf Maulana[1], Rodrigo Toro Jimenez[2], David Twesigomwe[3,4], Levana Sani[1], Astrid Irwanto[1], Nicolas Bertin[2,5]✉ & Mar Gonzalez-Porta[1,5]✉

Cytochrome P450 2D6 (CYP2D6) plays a crucial role in metabolizing approximately 20% of medications prescribed clinically. This enzyme is encoded by the *CYP2D6* gene, known for its extensive polymorphism with over 170 catalogued haplotypes or star alleles, which can have a profound impact on drug efficacy and safety. Despite its importance, a gap exists in the global genomic databases, which are predominantly representative of European ancestries, thereby limiting comprehensive knowledge of *CYP2D6* variation in ethnically diverse populations. In an effort to bridge this knowledge gap, we focused on elucidating the *CYP2D6* variation landscape within a multi-ethnic Asian cohort, encompassing individuals of Chinese, Malay, and Indian descent. Our study comprised data analysis of 1850 whole genomes from the SG10K_Health dataset using an in-house consensus algorithm, which integrates the capabilities of Cyrius, Aldy, and StellarPGx. This analysis unveiled distinct population-specific star-allele distribution trends, highlighting the unique genetic makeup of the Singaporean population. Significantly, 46% of our cohort harbored actionable *CYP2D6* variants—those with direct implications for drug dosing and treatment strategies. Furthermore, we identified 14 potential novel *CYP2D6* star-alleles, of which 7 were observed in multiple individuals, suggesting their broader relevance. Overall, our study contributes novel data on *CYP2D6* genetic variations specific to the Southeast Asian context. The findings are instrumental for the advancement of pharmacogenomics and personalized medicine, not only in Southeast Asia but also in other regions with comparable genetic diversity.

Pharmacogenomics entails understanding how genetic variation affects individual responses to drugs and it stands as a cornerstone in the advancement of personalized medicine. Within the landscape of drug-metabolizing enzymes, cytochrome P450 2D6, encoded by the *CYP2D6* gene, is well-recognized as a pivotal player. While it represents only 2–4% of the total CYP enzymes in the liver, CYP2D6 is responsible for metabolizing approximately 20% of clinically prescribed drugs, including a wide array of medications such as beta blockers, antiarrhythmics, antidepressants, antipsychotics, anticancer agents like tamoxifen, and opioids[1]. The significance of *CYP2D6* in pharmacogenomics is further emphasized by its extensive genetic diversity. To date, the Pharmacogene Variation (PharmVar) Consortium catalogue contains over 170 *CYP2D6* haplotypes, commonly referred to as star alleles[2,3]. The majority of these star alleles are defined by specific combinations of single nucleotide polymorphisms (SNPs) and/or small insertions and deletions (indels). However, known variations in the *CYP2D6* locus also encompass complex structural variants (SVs), such as full gene deletions, duplications, multiplications, and hybrid tandem rearrangements involving the closely related *CYP2D7* and *CYP2D8* pseudogenes.

The vast genetic heterogeneity of *CYP2D6* profoundly influences the effectiveness and safety of drugs metabolized by the CYP2D6 enzyme[1]. For instance, poor metabolizers carry two non-functional alleles and may not effectively metabolize or bioactivate drugs through the CYP2D6 pathway. Conversely, ultra-rapid metabolizers, who possess at least one increased function allele in addition to a normal-function allele, are at heightened risk of experiencing dose-related adverse events or treatment ineffectiveness. Thus, genetic variations can serve as pharmacogenetic biomarkers to guide drug dosing decisions, improve the effectiveness of therapies, and prevent adverse reactions for drugs metabolized by CYP2D6[4].

In this context, the Clinical Pharmacogenetics Implementation Consortium (CPIC) has developed guidelines to aid in translating pharmacogenetic laboratory test results into actionable prescribing decisions for specific

[1]Nalagenetics Pte Ltd, Singapore, Singapore. [2]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore. [3]Sydney Brenner Institute for Molecular Bioscience, Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, South Africa. [4]Division of Human Genetics, Faculty of Health Sciences, National Health Laboratory Service, and School of Pathology, University of the Witwatersrand, Johannesburg, South Africa. [5]These authors jointly supervised this work: Nicolas Bertin and Mar Gonzalez-Porta. ✉email: Nicolas_Bertin@gis.a-star.edu.sg; mar@nalagenetics.com

drugs, including three dedicated to CYP2D6[5]. However, the frequency of *CYP2D6* alleles exhibits considerable variability across different global populations, with some allelic variants universally present and others varying significantly in frequency or being exclusive to certain ethnic groups[4,6–9]. This diversity highlights the critical need to understand the variation landscape of *CYP2D6* for effective pharmacogenomic applications, particularly in the context of large-scale population genomic studies that can help inform drug policies. Nonetheless, there remains a noticeable gap in the data concerning variant frequencies in pharmacogenes, especially within Asian populations, since the majority of genetic variation data in public databases primarily originate from Western populations.

Singapore, with its diverse ethnic population, presents a unique opportunity for inclusive and comprehensive pharmacogenomic research. In addition, the advent of next-generation sequencing (NGS) technologies has made it feasible to sequence large cohorts, and with the aid of dedicated callers, NGS has demonstrated its capability to accurately characterize haplotypes in the *CYP2D6* gene, even in the presence of the aforementioned complexities[10]. In this context, our study aims to leverage Singapore's highly diverse genetic landscape, coupled with the capabilities of whole genome short-read sequencing, to provide insights into the variation landscape of the *CYP2D6* gene within Asian groups. To achieve this, we introduce a newly developed multi-tool bioinformatics workflow, and we apply it to a unique cohort comprising over 1,400 samples that have been sequenced at high depth using short reads[11]. We use the results of these analyses to investigate the distribution of known and potential novel *CYP2D6* star alleles in the three majority Asian ethnicities in Singapore (Chinese, Indian and Malay), and to characterize the prevalence of different metabolizer profiles in the population. The insights gathered from this research hold the potential for significant implications in the realm of clinical pharmacogenetics implementation strategies, not only within Singapore but also across Southeast Asia.

Results

Development and evaluation of a multi-tool pipeline for *CYP2D6* star allele calling

We developed a multi-tool analysis pipeline (Fig. 1a) to resolve *CYP2D6* diplotypes using three publicly available callers: Cyrius[12], Aldy[13], and StellarPGx[14]. Our pipeline, implemented in Nextflow (see "Methods"), accepts whole genome sequencing CRAM files as input. It initially generates individual calls from each of the three callers, which are then subjected to a consensus algorithm. This algorithm reports a consensus call when at least two out of the three callers concur on the diplotype. Typically, ambiguous calls are assigned for samples where this consensus criteria is not met. However, we manually inspected cases where the presence of potential novel alleles had been predicted by StellarPGx to further characterize and report the novel core variant combinations.

To characterize the variation landscape of *CYP2D6* in a multi-ethnic population, we executed our analysis pipeline on a dataset comprising 1850 samples from SG10K_Health, consisting of unrelated and healthy study participants (Fig. 1b). This dataset represents a highly diverse Southeast Asian cohort, including individuals of Chinese (74.6%), Malay (14.9%), and Indian (10.5%) ancestries, and was generated via short-read Whole Genome Sequencing (WGS) at 20–30 × coverage. PCA analysis of these data confirmed the existence of three clusters that align with the self-reported ancestries of study participants and overlap with East Asian (for Chinese and Malay) and South Asian (for Indian) populations in the HGDP and 1000 Genomes callset (N = 3942 samples), as expected (Supplementary Fig. 1). We first examined the call rates of each individual caller and observed that Aldy and StellarPGx exhibited the highest call rates at 99.9% and 99. 8%, respectively, while Cyrius employed a more conservative approach with a call rate of 89.6% (Fig. 1c). Next, we assessed the concordance of diplotype calls across the three tools and found that 56.6% (N = 1,040) of calls were supported by all three callers, 80.4% (N = 1487) were supported by at least two callers, and 19.6% (N = 363) were either supported by only one caller or harbor potential novel alleles. In pairwise comparisons, Cyrius and Aldy demonstrated the highest level of agreement, sharing 1266 haplotype calls (including 1040 shared among the three callers), followed by StellarPGx and Cyrius with 1165 shared calls, and StellarPGx and Aldy with 1136 shared calls (Fig. 1d). Notably, StellarPGx exhibited the ability to resolve the greatest number of unique diplotypes (N = 585), followed by Aldy (N = 487) and Cyrius (N = 267), which could be attributed to StellarPGx's capacity to detect novel alleles. In total, we identified 93 samples that potentially contained novel alleles, representing approximately 5% of our study population, which were further refined to a smaller set of 28 samples (1.5%) following manual curation (see "Methods").

CYP2D6 star allele frequencies and correlation with PharmGKB and 1KGP

In our subsequent analysis, we aimed to characterize the prevalence of star alleles within the different ancestral groups in our study cohort. Figure 2a presents the frequencies of the ten most common haplotypes, categorized by genetic ancestry, and reveals distinct patterns in star allele frequencies across the various ethnic groups. For example, the Chinese and Malay populations predominantly exhibit the **36 + *10*, **10*, and **36* alleles. In contrast, the Indian population displays a higher frequency of the **2*, **41*, **5*, and **4* alleles. Notably, 80% of the top ten most prevalent haplotypes contain actionable variants, characterized as either non-functional or reduced function *CYP2D6* alleles. In addition, we have detected 11 haplotypes that are only present in one individual (**112, *52, *133, *7, *75, *82, *17, *9, *15, *4 + *4, *69*) (Table 1).

Furthermore, we conducted a comparison of haplotype frequencies between the SG10K_Health and PharmGKB datasets, revealing a strong correlation with coefficients between 0.90 and 0.93 (Fig. 2b). Importantly, this trend remains consistent across individual ethnicities and when correlating frequencies with the 1KGP dataset (Supplementary Fig. 2). We noted discrepancies in the frequency of several non-reference alleles compared to the average frequencies estimated in PharmGKB. Specifically, when analyzed in aggregate, the SG10K_Health dataset showed a lower frequency of the *10 allele. This discrepancy may be due to overestimation in PharmGKB, potentially because *36 + *10 tandems are not always accurately detected, leading to *10 being reported instead. Additionally, we detected outliers for *36 + *10 tandems in Malays and *2 alleles in Indians,

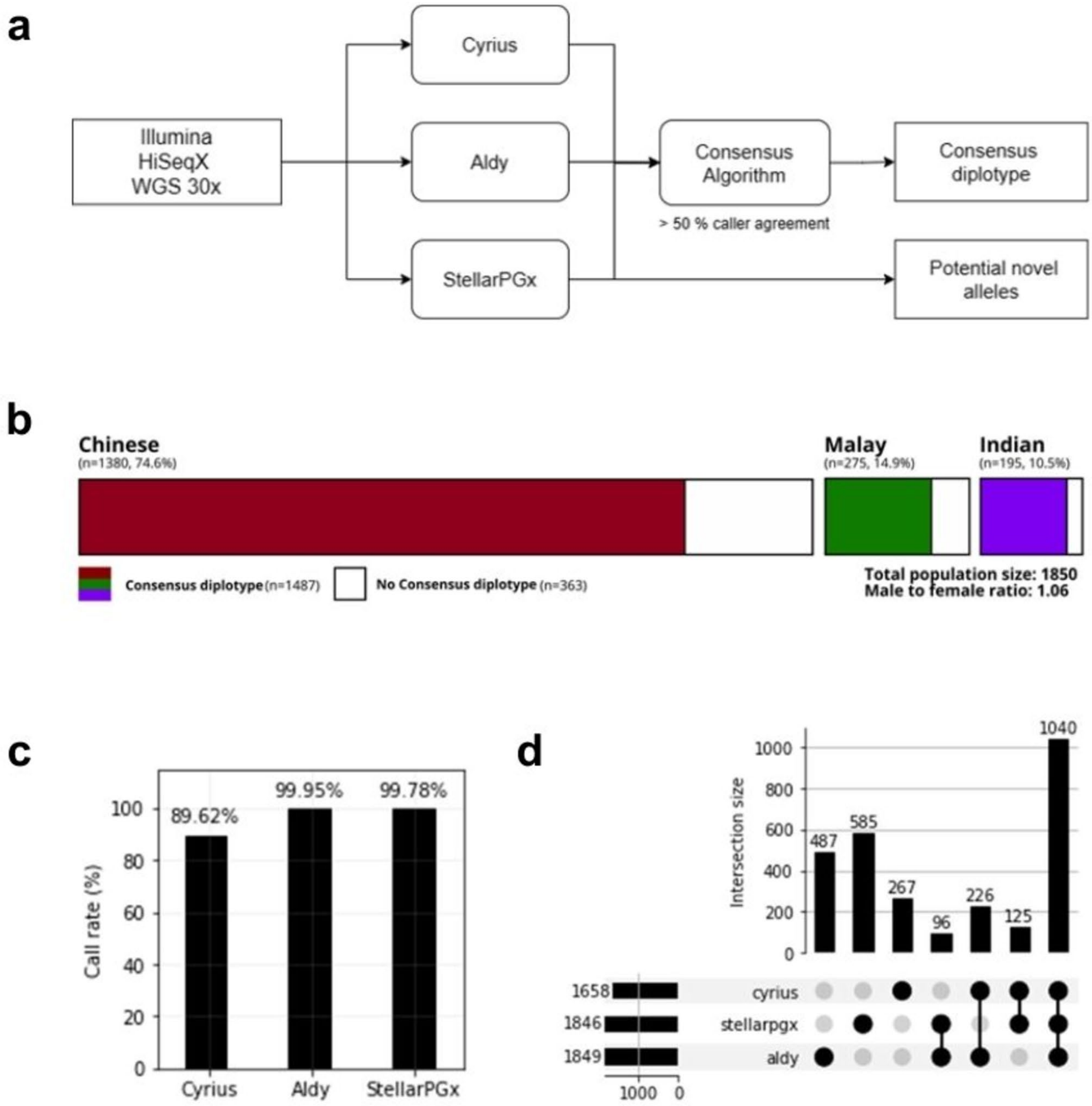

Figure 1. Overview of data analysis workflow and study design. (**a**) Data analysis pipeline. The diagram illustrates the workflow for analyzing WGS 30X CRAM files. Three tools —Cyrius, Aldy, and StellarPGx— are utilized for *CYP2D6* genotyping. A consensus algorithm determines the final diplotypes based on agreement from at least two of the three tools. StellarPGx is employed to identify potential novel alleles. (**b**) SG10K_Health cohort. The cohort for the study comprises individuals from the three major ethnic groups in Singapore, namely Chinese, Malay, and Indian. Colored bars represent the count and proportion of consensus diplotypes within the total cohort of 1,850 participants. (**c**) Comparison of call rates. The bar graph compares the successful calling rates of Cyrius, Aldy, and StellarPGx in generating diplotype calls. (**d**) UpSet plot of CYP2D6 diplotype calls. The plot displays the overlap in diplotype calls among the three callers. Numbers atop the bars show the size of each intersection set. The lower left side of the plot indicates the total successful calls per tool. The unique diplotypes identified by each tool are shown in the three leftmost bars (Aldy with 487, StellarPGx with 585 and Cyrius with 267 unique diplotype calls). The middle bars represent diplotypes identified by tool pairs, and the far-right bar shows 1040 diplotypes identified by all three tools.

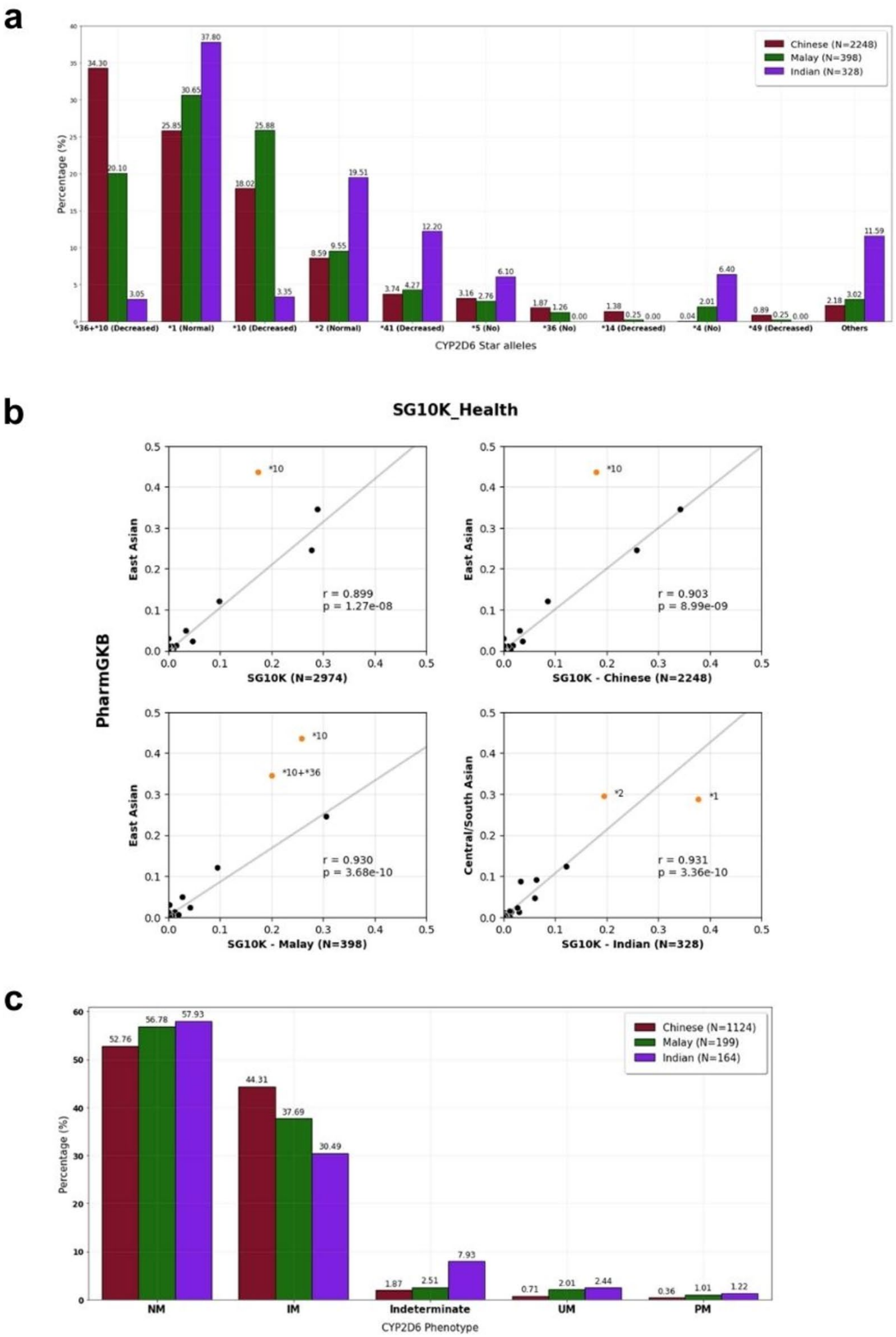

Figure 2. Distribution of key *CYP2D6* star alleles and phenotypes across different genetic ancestries. (**a**) Frequency distribution of the top 10 *CYP2D6* star alleles in the SG10K_Health dataset, categorized by genetic ancestries. The x-axis depicts the star alleles, each annotated with their respective functional classification (normal, decreased, or no function). The y-axis shows the frequency distribution (%) of these star alleles within each genetic ancestry group. (**b**) Comparative analysis of *CYP2D6* star allele frequencies between the SG10K_Health dataset and PharmGKB. The x-axis displays the frequency of star alleles observed in the SG10K_Health dataset. The y-axis indicates the frequency of star alleles as reported in PharmGKB for East Asian and Central/South Asian populations. The Pearson correlation coefficient (*r*) is represented for each comparison. (**c**) Frequency distribution of CYP2D6 phenotypes among different genetic ancestries. The x-axis lists all the possible CYP2D6 phenotypes, including normal metabolizers (NM), intermediate metabolizers (IM), indeterminate, ultrarapid metabolizers (UM), and poor metabolizers (PM). The y-axis shows the frequency distribution (%) of these phenotypes across each genetic ancestry.

Star Allele	Defining variants	CPIC function	Frequency in SG10K_Health (%)				Frequency in 1KGP (%)	
			Overall (N = 2974)	Chinese (N = 2248)	Indian (N = 328)	Malay (N = 398)	EAS (1KGP)	SAS (1KGP)
*1	Reference	Normal function	27.81	25.85	37.80	30.65	26	39.5
*2	2851C > T (rs16947, R296C) + 4181G > C (rs1135840, S486T)	Normal function	9.92	8.59	19.51	9.55	7.8	20.8
*35	31G > A (rs769258, V11M) + 2851C > T (rs16947, R296C) + 4181G > C (rs1135840, S486T)	Normal function	0.50	0.44	1.52	0	0	0.6
*39	4181G > C (rs1135840, S486T)	Normal function	0.13	0.13	0	0.25	0.3	0.4
*33	2484G > T (rs28371717, A237S)	Normal function	0.07	0	0.61	0	0	0.2
*2 × 2	Duplication of *2	Increased function	0.34	0.31	0.30	0.50	0.5	0.4
*1 × 2	Duplication of *1	Increased function	0.30	0.09	1.22	0.75	0	0
*36 + *10	Tandem duplication with exon 9 conversion (*36)	Decreased function	28.95	34.30	3.05	20.10	34.5	1.2
*10	100C > T (rs1065852, P34S) + 4181G > C (rs1135840, S486T)	Decreased function	17.45	18.02	3.35	25.88	15.2	3.5
*41	2851C > T (rs16947, R296C) + 2989G > A (rs28371725, splice defect) + 4181G > C (rs1135840, S486T)	Decreased function	4.74	3.74	12.20	4.27	3.4	11.4
*14	1759G > A (rs5030865, G169R) + 2851C > T (rs16947, R296C) + 4181G > C (rs1135840, S486T)	Decreased function	1.08	1.38	0	0.25	1	0
*49	100C > T (rs1065852, P34S) + 1612 T > A (rs1135822, F120I) + 4181G > C (rs1135840, S486T)	Decreased function	0.71	0.89	0	0.25	0.5	0
*17	1022C > T (rs28371706, T107I) + 2851C > T (rs16947, R296C) + 4181G > C (rs1135840, S486T)	Decreased function	0.03	0	0.30	0	0	0
*9	2616delAAG (rs5030656, K281del)	Decreased function	0.03	0	0.30	0	0.3	0.4
*5	Full gene deletion	No function	3.43	3.16	6.10	2.76	0	0.6
*36	100C > T (rs1065852, P34S) + Ex9-conv (P469A, T470A, H478S, G479R, F481V, A482S) + 4181G > C (rs1135840, S486T)	No function	1.58	1.87	0	1.26	0.2	0
*4	1847G > A (rs3892097, splice defect)	No function	1.01	0.04	6.40	2.01	0	0
*68 + *4	CYP2D6-2D7 hybrid (intron 1 switch region) + *4	No function	0.34	0.04	2.74	0	0	2.2
*21	2580_2581insC (rs72549352, R269fs) + 2851C > T (rs16947, R296C) + 4181G > C (rs1135840, S486T)	No function	0.13	0.18	0	0	0.5	0
*7	2936A > C (rs5030867, H324P)	No function	0.03	0	0.30	0	0	0
*15	137_138insT (rs774671100, L47fs)	No function	0.03	0	0	0.25	0	0
*4 × 2	Duplication of *4	No function	0.03	0	0.30	0	0	0
*69	100C > T (rs1065852, P34S) + 2851C > T (rs16947, R296C) + 2989G > A (rs28371725, splice defect) + 4181G > C (rs1135840, S486T)	No function	0.03	0.04	0	0	0.3	0.2
*71	125G > A (rs118203758, G42E)	Uncertain function	0.44	0.44	0	0.75	0.6	0
*86	2607G > A (rs77913725, E278K) + 2611 T > A (rs1135828, M279K)	Uncertain function	0.20	0.09	1.22	0	0	2.3
*43	77G > A (rs28371696, R26H)	Uncertain function	0.20	0.09	1.22	0	0	1

Continued

Star Allele	Defining variants	CPIC function	Frequency in SG10K_Health (%)				Frequency in 1KGP (%)	
			Overall (N = 2974)	Chinese (N = 2248)	Indian (N = 328)	Malay (N = 398)	EAS (1KGP)	SAS (1KGP)
*94	100C > T (rs1065852, P34S) + 3182A > G (rs748712690, D337G) + 4181G > C (rs1135840, S486T)	Uncertain function	0.10	0.13	0	0	0	0
*65	100C > T (rs1065852, P34S) + 2851C > T (rs16947, R296C) + 4181G > C (rs1135840, S486T)	Uncertain function	0.07	0.04	0	0.25	0	0
*113	4040G > A (rs569439709, G439D)	Uncertain function	0.07	0	0.61	0	0	0.8
*111	1033G > A (rs535642512, G111S) + 2851C > T (rs16947, R296C) + 4181G > C (rs1135840, S486T)	Uncertain function	0.07	0	0.30	0.25	0	0.8
*112	4187C > T (rs568495591, S488F)	Uncertain function	0.03	0	0.30	0	0	0.2
*52	100C > T (rs1065852, P34S) + 3878G > A (rs28371733, E418K) + 4181G > C (rs1135840, S486T)	Uncertain function	0.03	0.04	0	0	0.1	0
*133	2851C > T (rs16947, R296C) + 3190G > A (G340R) + 4181G > C (rs1135840, S486T)	Uncertain function	0.03	0.04	0	0	0	0
*75	4046G > A (rs532668079, R441H)	Uncertain function	0.03	0.04	0	0	0	0
*82	973C > A (rs28371703, L91M) + 983A > G (rs28371704, H94R) + 1013 T > C (rs76187628, V104A) + 1021A > T + 1022C > A (rs74802369, rs28371706, T107Y) + 1027A > G (rs78459009, I109V)	Uncertain function	0.03	0	0.30	0	0	0

Table 1. Frequencies of *CYP2D6* star alleles in the SG10K_Health and 1KGP datasets, categorized by genetic ancestries. Star alleles are sorted by CPIC function and overall frequency in SG10K_Health.

both over-represented in PharmGKB compared to our dataset. These ancestry-specific discrepancies are likely due to the smaller sample sizes of these two groups; however, since *2 is considered a "backbone" allele, this trend could also reflect mis-assignments if additional haplotype-defining variants are not accurately detected.

Distribution of CYP2D6 metabolizer profiles

Among the 1487 samples that met the criteria of our consensus algorithm, approximately 46.1% exhibited actionable CYP2D6 phenotypes affecting drug metabolism (i.e., intermediate metabolizer (IM), poor metabolizer (PM), and ultrarapid metabolizer (UM) phenotypes), emphasizing the role that pharmacogenomics can play in clinical practice. Further analysis of phenotype distribution in our cohort revealed that 53.9% of the study samples could be classified as normal metabolizer (NM), followed by 41.9% as IM, then 1.1% as UM, and 0.5% as PM (Supplementary Fig. 3). IMs and PMs could be attributed to the higher prevalence of alleles and nonfunctional alleles such as **10*, **36*, **41* and **5*. Additionally, 2.6% of the samples had indeterminate phenotypes, representing variants with unknown effects on metabolism. We also identified three instances of complete *CYP2D6* deletions (**5/*5*).

Our study also unveiled distinct patterns of metabolizer profiles within the SG10K_Health cohort based on ancestry (Fig. 2c). Notably, a lower incidence of NM was observed in the Chinese and Malay participants, possibly linked to the higher prevalence of **36* + **10* hybrids. In contrast, individuals of Indian descent exhibited a higher prevalence of indeterminate phenotypes (7.9% compared to 2.5% and 1.9% for Malay and Chinese, respectively), associated with the prevalence of star alleles with unknown function (**43*, **86*, **113*, **82*, **111*, and **112*).

Distribution of *CYP2D6* structural variants

Subsequently, we aimed to characterize *CYP2D6* Structural Variants (SVs) within the SG10K_Health cohort. As expected, the most prominent SV-containing star allele was **36* + **10*, which displayed an average frequency of 29% in the total population, primarily contributed by Chinese and Malays (Fig. 3a). This was followed by *CYP2D6*5* (full-gene deletion), albeit at a much lower frequency of 3.4% overall, and with a two-fold higher prevalence in Indians. Additionally, we also detected **2* × *2* and **1* × *2* haplotypes, indicative of gene duplications, with frequencies of 0.3% for both across the total population. These haplotypes showed increased function due to the multiplication of normal function alleles (**2* and **1*). Lastly, the **68* + **4* hybrid and **4* + **4* allele, both characterized by having no function, were less common in the overall cohort (0.4%) and were predominantly found in the Indian subgroup.

Notably, the majority of participants (55.6%) harbored at least one star allele containing an SV, while 44.38% harbored no observable SVs (Fig. 3b). However, these results may constitute an under-representation of the prevalence of SVs in our study population, given the stringency of the consensus algorithm that we employed and the inherent limitations of short-reads. Stratification based on the presence or absence of SVs revealed significant differences in metabolizer profiles (Fig. 3c). As expected, in samples without SVs, NM made up 72.1% of the cases, followed by IM at 23.9%, and a smaller proportion of indeterminate samples at 3.9%. In contrast, the group with SVs predominantly consisted of IM (56.2%). NM constituted the second most common category, with 39.3% frequency, highlighting the importance of detecting the exact nature of the *CYP2D6* SVs present in a given sample. We also identified UM and PM, though they were less prevalent, with frequencies of 1.9% and 1%, respectively.

Potentially novel *CYP2D6* haplotypes

Lastly, we proceeded to examine the set of potentially novel star alleles detected by StellarPGx. Initially, this tool flagged 93 samples as containing potential novel alleles; however, following manual review, we refined the dataset to 13 computationally-resolved potential novel haplotypes and 1 potential novel allele in a diplotype with unresolved phasing (Table 2 and Fig. 4). The remaining samples did not exhibit sufficient evidence to be considered as harboring novel alleles, with the majority presenting no additional core variants after manual curation (N = 37, see "Methods") while the rest are characterized by low quality or potential miscalls indicative of existing combination of star alleles. Among the 14 potential novel alleles, the majority were associated with the **10* and **2* haplotype backbones (9 out of 14). Notably, these novel alleles were observed in 28 study samples, with 7 of these alleles detected in multiple individuals, and 4 samples from different ethnic backgrounds, consistently observed in both Chinese and Malay individuals. In contrast, among single-population occurrences, 7 unique events were exclusive to the Chinese group, while 2 were unique to the Indian group. All detected novel alleles in our study were found to contain functional variants, including missense and frameshift mutations, which potentially alter protein functionality. While these novel core variants are cataloged in the dbSNP database, they are novel in the context of our study as they appear in combinations not previously observed with *CYP2D6* alleles in the PharmVar database. Noteworthy, 4 of the allele-defining variants—rs746803316, rs3915951, rs1135830, and rs759234339—have not been previously catalogued as *CYP2D6* core variants in the PharmVar database.

Discussion

While extensive research has been conducted on *CYP2D6* across various ethnicities, a gap still exists in understanding the extent of *CYP2D6* pharmacogenetic diversity within Southeast Asian populations. To address this gap, we conducted a study to characterize the distribution of *CYP2D6* star alleles and their associated phenotypes using a genetically diverse cohort from Singapore[11]. This cohort comprises individuals representing the three major ethnicities in the country: Chinese, Malay, and Indian, and includes high-coverage short-read whole genome sequences from over 1800 participants. To the best of our knowledge, this study represents the most comprehensive examination of *CYP2D6* genetic variation in the Singaporean population, and given the country's rich diversity, it provides an ideal platform for comprehensively exploring *CYP2D6* variation within Southeast Asia.

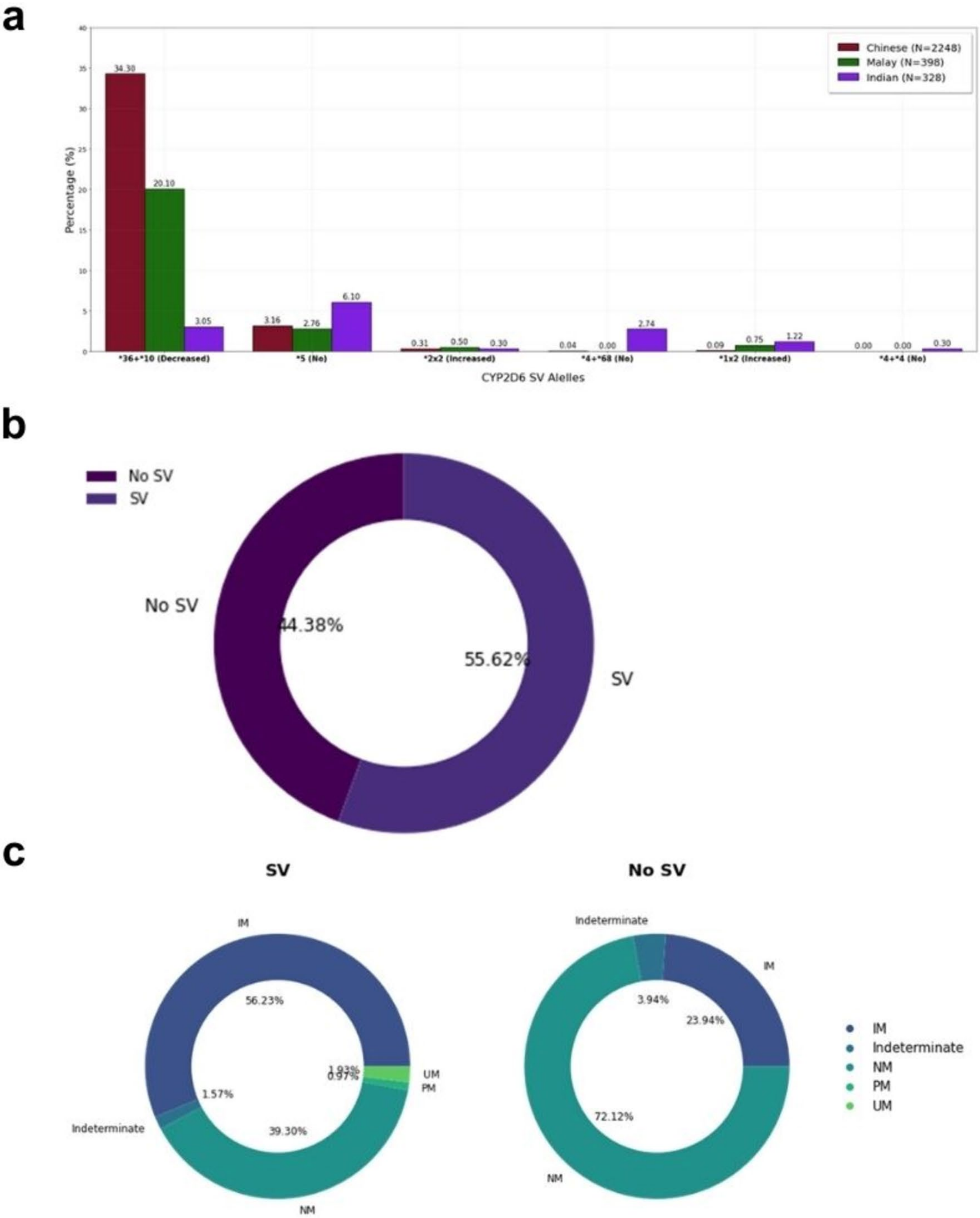

Figure 3. Frequency distribution of *CYP2D6* structural variations (SVs). (**a**) Frequency distribution of SV-containing *CYP2D6* star alleles in the SG10K_Health dataset, categorized by genetic ancestries. The x-axis depicts the star alleles, each annotated with their respective functional classification (increased, decreased, or no function). The y-axis shows the frequency distribution (%) of these star alleles within each genetic ancestry group. (**b**) Comparison of the proportion of samples with SV and samples with no SV in the SG10K_Health dataset (N = 1,487). (**c**) Distribution of CYP2D6 phenotypes in the SG10K_Health dataset, categorized by SV presence. The left side illustrates the proportion in samples with SV (N = 827), and the right side shows the proportion in samples with no SV (N = 660).

We developed a bioinformatics workflow using three distinct tools to mitigate inaccuracies in identifying *CYP2D6* variants and diplotypes, given the limitations presented by the short-read data we used in this study. Our workflow includes a consensus algorithm, which reports a diplotype call only when at least two out of three

Haplotype	Background allele	Additional core variant(s) GRCh38 (NC_000022.11)	 Rsid	Variant type	Frequency in gnomAD Global	 East Asian	Count
1	*1	42128341C > T	rs746803316	Missense(A226T)†	2.73.E − 05	8.27.E − 04	3
2	*10	42127634C > A	rs3915951	Missense(R329L)†	2.66.E − 03	7.56.E − 03	6
3	*10	42127853G > A	rs140513104	Missense(P325L)	1.53.E − 04	4.69.E − 04	4
4	*10	42128812A > G, 42128818 T > TC	rs199535154, rs72549354	Missense(L213P), Frameshift(L213fs)	2.54.E − 04, 0	5.64.E − 04, 0	2
5	*10	42127895G > A	rs1135830	Missense(S311L)†	6.77.E − 05	6.71.E − 05	2
6	*10	42128864C > T	rs755518310	Missense(E196K)	8.11.E − 06	2.25.E − 05	1
7	*10	42128818 T > TC	rs72549354	Frameshift(L213fs)	0	0	1
8	*2	42129042 T > C	rs1135824	Missense(N166D)	1.58.E − 04	1.79.E − 04	2
9	*2	42129819G > T	rs28371703	Missense(L91M)	1.50.E − 01	8.06.E − 04	2
10	*2	42129809 T > C	rs28371704	Missense(H94R)	1.59.E − 01	7.39.E − 04	1
11	*41	42126747G > A	rs730882251	Missense(R441C)	1.21.E − 04	0	1
12	*86	42130715C > T	rs28371696	Missense(R26H)	1.52.E − 03	3.15.E − 04	1
13	*86	42130692G > A	rs1065852	Missense(P34S)	2.12.E − 01	4.88.E − 01	1
Total							**27**
Unresolved diplotypes with potentially novel star alleles							
1	*1/*2	42129807C > T	rs759234339	Missense(G283S)†	8.72.E-06	0	1
Total							**1**

Table 2. Potentially novel *CYP2D6* star alleles. The 14 potentially novel haplotypes identified following manual curation of StellarPGx outputs. †Variants not currently listed in PharmGKB, for which interpretation was conducted using VEP.

tools concur; however, cases where potential novel star alleles were predicted by StellarPGx were subjected to additional manual inspection to ascertain the final diplotype calls. This approach, while conservative, was chosen to prioritize the accuracy of star allele assignments. Additionally, our workflow includes steps to interpret predicted diplotypes into metabolizer profiles and to identify potential novel alleles. Upon applying the workflow to the 1850 samples in our cohort, we successfully determined consensus diplotype calls for 1487 samples, encompassing over 80% of the population. In contrast, around 20% of samples remained uncharacterized. Notably, the majority of samples that did not reach consensus contained haplotypes that included SVs and potential novel alleles (90.4%), thus underscoring the challenges associated with star-allele analysis of the *CYP2D6* locus via short-read sequencing. Such challenges arise from the gene's significant homology with the *CYP2D7* and *CYP2D8* pseudogenes, hinting at the potential for future research utilizing novel technologies such as long-read sequencing. In fact, technology providers such as Pacific Biosciences and Oxford Nanopore Technologies have begun releasing dedicated workflows for PGx, and dedicated software tools like pangu are already demonstrating improvements over reference calls previously established with short-read sequencing and long-range PCR[15].

In our study, we observed significant patterns in the distribution of *CYP2D6* alleles among the Southeast Asian populations we examined. Among the most prevalent star alleles, we noted a predominance of alleles associated with reduced or absent function, except for *1 and *2, which are associated with normal function. *1 emerged as the most prevalent *CYP2D6* allele with normal function, followed by *2, aligning with previously reported trends in Southeast Asian populations[9]. Additionally, we observed variations in allele frequencies among the Singaporean populations included in our study. *36 + *10, *10, and *36 were more prevalent in the Chinese and Malay populations, with the Chinese and Malays exhibiting approximately six-fold higher allele frequencies of *10 compared to Indians. This trend aligns with well-documented findings, which consistently report a high prevalence of the *36 + *10 tandem in East Asian populations, including Japanese, Korean, and Chinese[9]. Furthermore, it supports the common identification of the *10 allele as a reduced-function variant in Asian populations[8,9]. Notably, *36 was not observed in the Indian participants in this study. In contrast, *2, *41, *5, and *4 alleles exhibited higher prevalence in the Indian population, reaffirming previous research highlighting the prominence of these alleles in Indian samples[16]. We also observed 11 haplotypes (*112, *52, *133, *7, *75, *82, *17, *9, *15, *4 + *4, *69) that are only present in one individual which expands the previous findings of rare stare alleles in our study population[16]. Of particular interest was the identification of three individuals with no detectable copies of *CYP2D6* (*5/*5*), constituting approximately 0.2% of our population. This aligns with previous reports indicating that the *5/*5 diplotype occurs at a very low frequency, ranging from 0% to 1.9% in Southeast Asian populations[9].

The frequencies of common star alleles in our population, such as *10, *36 + *10, and *2, exhibited significant differences from the average frequencies estimated by the PharmGKB and 1000 Genomes Project (1KGP) for the equivalent populations. Whilst PharmGKB may include additional ethnicities to the ones in SG10K_Health, these discrepancies may originate from variations in data generation methods (e.g. WGS versus genotyping) and star allele calling approaches employed in various studies, leading to variability in the range of detectable star alleles. It is also possible that the frequency of *10 could be overestimated when *36 + *10 tandems and/or *36 alleles are not reported. Additionally, *2 alleles are considered "backbone" alleles since their defining Single

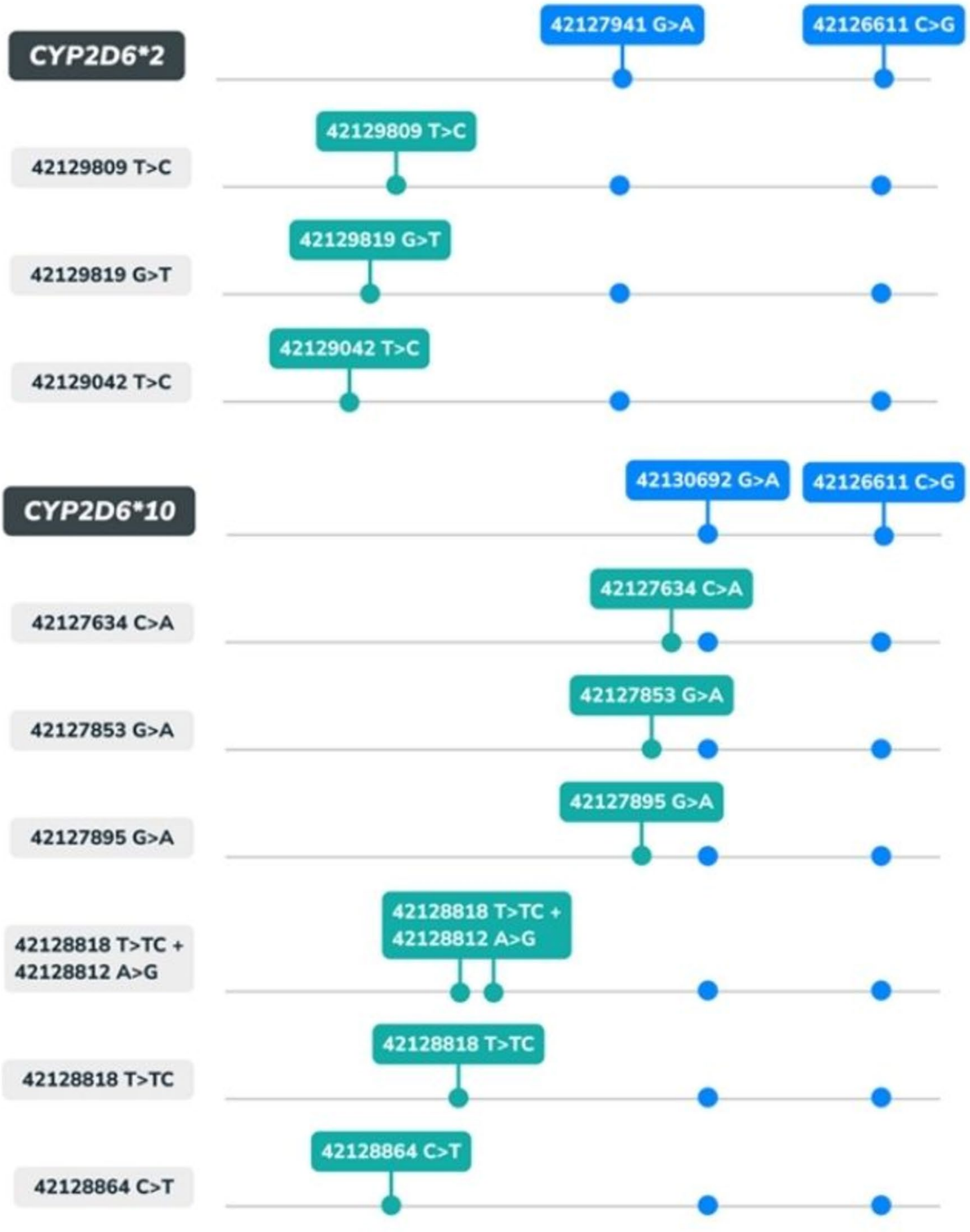

Figure 4. Visual representation of potential novel alleles with **2* and **10* background alleles. The core variants of the background alleles are marked in blue, and the unique novel variants are marked in green.

Nucleotide Variants (SNVs) occur in multiple other haplotypes, which may introduce potential mis-assignments if any of these additional haplotype-defining variants cannot be accurately detected.

After translating diplotypes into metabolizer profiles, we identified actionable variants in over 46% of the population, increasing to over 80% when focusing on the top ten most common haplotypes. This underscores the significant impact of implementing pharmacogenomics on a large scale. In our study population, normal metabolizers (NMs) were the most prevalent phenotype at 53.9%, ultra-rapid metabolizers (UMs) accounted for 1.1%, while poor metabolizers (PMs) represented 0.5%. The highest frequencies of diplotypes predicting PM were found in Indian subjects (1.2%), followed by Malays (1%) and Chinese (0.4%). Interestingly, we detected PMs in all three ethnicities (Chinese, Indian, and Malay), a difference from previous studies in the same population that did not identify PMs in Chinese individuals[16], likely due to our larger sample size and higher-depth sequencing approach. Compared to global trends, our cohort showed a slightly higher incidence of intermediate metabolizers, around 42%, surpassing the previously reported 34%[17]. This coincided with a decrease in the prevalence of normal metabolizers, typically observed at 64–68% in global populations. This trend may be attributed to the higher prevalence of diplotypes associated with reduced or no function, including **36 + *10, *10, *41*, and **36*.

Both PMs and UMs exhibit altered capacity to metabolize CYP2D6 substrates, including codeine, certain antidepressants, and antipsychotics[18]. UMs face an elevated risk of toxicity due to increased morphine formation after codeine administration, while individuals with non-functional alleles are at risk of inadequate pain relief due to reduced efficacy. In our study, the proportion of UMs exceeded that of PMs by approximately than two-fold. Unlike Caucasians, where the **4* allele predominates and accounts for 70–90% of poor metabolizer status, its low frequency in Asians may explain the lower proportion of poor metabolizers in our population[17]. Lastly, we observed a relatively high proportion of Indian participants (7.9%) with an indeterminate CYP2D6 metabolizer phenotype, highlighting the limitations of current CPIC guidelines for genotype–phenotype translation. Most of these individuals carried alleles with uncertain functions, such as **43, *86, *113, *82, *111*, and **112*. This underscores the need for extensive allele characterization and phenotypic studies to develop effective precision medicine strategies, particularly for medications metabolized by CYP2D6.

We further inspected the prevalence of structural variants (SVs) in CYP2D6, as it remains underexplored in Asian populations[8]. Our analysis indicated that the majority of study participants (55.6%) had at least one SV-containing star allele, with *36 + *10 hybrids being the most prevalent overall. This percentage exceeds the prevalence previously reported for the same population by Chan et al.[8]. All three tools used in our consensus diplotype calling approach demonstrated > 90% recall for known CYP2D6 structural variants, including key star alleles such as *36 + *10, *36 × 2 + *10, and *36 + *10 × 2, when using high-depth short-read WGS as input in previous benchmarks (aldy, Cyrius, and StellarPGx). Based on these previous validations and the single-tool deficiencies resolved by our consensus approach, we are confident in having minimal false-positive calls. However, there might be cases where novel SVs were missed or where complex diplotypes (e.g., individuals having an SV on each haplotype) might have been miscalled. As more data from long-read technologies is generated, we expect an overall improvement in accurately characterizing SV-driven CYP2D6 variation, particularly in understudied populations. When stratifying metabolizer profiles based on the presence or absence of SVs, we detected a higher incidence of IMs (56.2%), UMs (1.9%) and PMs (1%) among participants with SV-containing alleles compared to individuals with no *CYP2D6* SVs, as expected. Interestingly, we still detected a high prevalence of NMs among the first group (39.3%), highlighting the importance of detecting the exact nature of the *CYP2D6* SVs present in each sample, and emphasizing the limitations of relying on copy number or SV information alone to make predictions on phenotypic outcomes.

Lastly, our study also provides an initial assessment of the extent of genetic variation that remains undocumented in public databases. We identified 14 potential novel haplotypes for *CYP2D6*, based on a carefully curated subset. This group includes both shared variants observed across multiple individuals (N = 7) and private events (N = 7). Although these novel haplotypes are individually rare, collectively, they appear in 28 individuals, accounting for 1.5% of our study population. This underscores the significant, yet often overlooked, impact that rare allelic variations could have on precision medicine strategies, both in Asia and globally. The majority of these 13 novel haplotypes are variations of the **10* allele (N = 6), followed by **2* (N = 3). This distribution aligns with the prevalence of these alleles, which are among the top ten star alleles frequently identified in our study populations. All the novel haplotypes we detected include potentially functional variants previously catalogued in PharmVar in combination with other haplotypes, except for 5, which we inferred using the Variant Effect Predictor. They encompass a range of genetic changes, including missense and frameshift mutations, which could significantly alter protein function. However, given the complexities of genotyping and star allele calling in *CYP2D6*, caution is advised when interpreting computationally inferred novel haplotypes, particularly singletons. The predictions can be influenced by limitations in the genomic datasets used, such as short-read sequencing and software. While these computational predictions are informative, they are not a substitute for actual experimental validation. Although we could not validate these haplotypes due to lack of DNA access, our findings offer a valuable foundation for future research, especially as Singapore's National Precision Medicine program progresses in characterizing a larger portion of its population. A prevalent practice in the field is to report variants based on their backbone allele, which could significantly influence drug dosage recommendations. Unexamined variants with potential functional significance, like those uncovered in our study, could affect phenotype assignments and clinical decisions, potentially leading to adverse pharmacological outcomes for patients.

Overall, our research represents a significant step towards enhancing the understanding of *CYP2D6* variations in under-represented populations. Gaining a thorough insight into the allelic diversity within the populations in this study is key for precisely predicting drug responses and successfully applying pharmacogenetics in Singaporean and global clinical settings, where targeted assays are still the norm. Therefore, understanding the genetic variability of the various populations is essential to ensure that the most prevalent alleles are effectively identified and included in the list of targets. While a broader test such as whole genome sequencing would be ideal to continuously explore genetic variation and guide future clinical applications, cost constraints currently limit this approach. However, the ongoing reduction in sequencing costs and the advanced discovery potential of long-read sequencing techniques point to a promising future in this research area. Despite the technological limitations in characterizing our dataset, we believe our analysis presents a valuable contribution to the Singapore and global scientific community. We advocate for the consensus *CYP2D6* star allele calling method used in our study for similar analyses, to address the challenges of short-read sequencing. Future studies focusing on the definitive characterization of the novel haplotypes not validated in this research, along with functional studies assessing their clinical significance, will be crucial for enhancing clinical pharmacogenomics implementation strategies. In the meantime, we believe that the detailed mapping of *CYP2D6* star allele distributions in Southeast Asian populations, as presented in our study, can serve as a resource in advancing precision medicine strategies and fostering the adoption of proactive pharmacogenetic testing in diverse clinical environments across Asia and worldwide.

Methods

Study population

The study made use of the sequencing data provided by the SG10K_Health dataset[11]. This dataset was acquired in compliance with the ethical principles set forth by the National University Hospital Singapore (CIRB/E/2019/2655) and the SingHealth Centralised Institutional Review Board (2013/605/C). Participants in the study cohorts provided signed informed consent, or in the case of minors, consent was obtained from a parent or guardian. The processes of library preparation and sequencing adhered to applicable guidelines and regulations. The study evaluated 1850 samples, each undergoing short-read whole genome sequencing at a 30 × depth. The ethnic composition showed a significant number of Chinese participants, comprising 74.6% (N = 1380) of the total. The Malay and Indian participants followed, contributing to 14.9% (N = 275) and 10.5% (N = 195) respectively. Participants' ethnicities were determined based on their self-reported information.

Consensus algorithm

Aldy[13], Cyrius[12], and StellarPGx[14] are three well-established *CYP2D6* star allele callers designed for use with short-read data. We compared the star allele calls from these three tools to reach a consensus on the diplotypes assigned to each sample in our dataset, aiming to increase confidence in our callset.

The three tools were executed in parallel from CRAM inputs using a Nextflow workflow developed in-house. A custom parser was used to read each sample's output from Aldy, Cyrius, and StellarPGx. For ease of comparison across tools, parsed outputs were initially stored as separate haplotypes (e.g., **5/*36 + *10* was split into **5* and **36 + *10*). For callers that reported minor star alleles, these were excluded from the analysis since they are not functionally different from their corresponding major star alleles (e.g., **4C* was simplified to **4, *2.005* to **2*, and **2.ALDY* to **2*). Furthermore, the major star alleles of each haplotype were sorted in numerically ascending order and merged into diplotypes (e.g., **10 + *36/*5* was rearranged to **5/*10 + *36*) to facilitate downstream consensus comparison. Star-allele definitions were based on PharmVar version 5.2.14, and are provided in Supplementary Table 1 for reference.

Processed diplotypes for each sample were then compared against each other. A consensus was considered achieved when more than 50% of calls supported a particular diplotype. Additionally, if the samples are found to have potential novel alleles, they were excluded from the mainstream consensus diplotype call-set and subjected to further manual inspections.

PCA analysis

Principal Component Analysis (PCA) was conducted to investigate the genetic structure of the SG10K_Health dataset and compare it with the HGDP and 1 KG projects. The genotyping callset for all 3942 samples from HGDP-1 KG was retrieved from gnomAD (https://gnomad.broadinstitute.org/downloads#v3-hgdp-1kg). Additionally, the genotyping callset for the 1850 SG10K_Health samples, which were subjected to CYP2D6 haplotyping, was obtained from the SG10K_Health release 5.3 (https://www.nature.com/articles/s41588-022-01274-x). The PCA was performed using the "hwe_normalized_pca" function provided by Hail (https://hail.is/docs/0.2/_modules/hail/methods/pca.html). This analysis incorporated 49832 single nucleotide variant (SNV) loci common to both the HGDP-1 KG and SG10K_Health callsets. To ensure accurate ancestry assignment, the VerifyBAMID-defined 1000 g.phase3.100 k reference panel was employed (https://github.com/Griffan/VerifyBamID/blob/master/resource/1000g.phase3.100k.b38.vcf.gz.dat).

Ancestry labeling for the HGDP-1 KG samples was based on gnomAD-defined superpopulations, while the SG10K_Health samples were labeled according to the participants' self-reported ancestry. Specifically, the "SAS" (South Asian) superpopulation corresponds closely to SG-Indian ancestry, while the "EAS" (East Asian) superpopulation aligns with SG-Chinese and SG-Malay ancestries.

Calculation of star allele frequencies and correlation with PharmGKB

The frequency of *CYP2D6* haplotypes was derived from the consensus diplotypes. We conducted a comparative analysis of the observed frequencies in the SG10K_Health dataset and those in PharmGKB[19]. For this comparison, we matched subsets of the population in each dataset, aligning East Asian with overall, Chinese, and Malay populations, and South Asian with the Indian population, as detailed in the PharmGKB *CYP2D6* frequency table (https://www.pharmgkb.org/page/cyp2d6RefMaterials). A notable limitation in our approach was the absence of frequency data for tandem variants (e.g. **36 + *10* and **68 + *4*) in the PharmGKB dataset. To address this, we supplemented our analysis with data from the 1000 Genomes Project (1KGP)[6]. Outliers in allele frequency between the two datasets were identified by calculating the frequency differences and considering a z-score greater than 3 as indicative of an outlier. Additionally, we assessed the correlation between the two datasets using Pearson's correlation coefficient.

Interpretation of star alleles into metabolizer profiles

Metabolizer profiles are categorized based on activity scores derived from the *CYP2D6* diplotypes. This scoring is based on the Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines[20]. The scoring system assigns function values to the star alleles (e.g., increased, normal, decreased, or no function). For each allele, an "activity value" ranging from 0 to 1 is assigned, such as 0 for no function, 0.5 for decreased function, and 1.0 for normal function. The activity score (AS) of a *CYP2D6* diplotype is then the sum of these values. Additionally, in cases where the *CYP2D6* allele has variable copy numbers, the activity value of an allele is multiplied by the number of gene copies. Within CPIC guidelines, metabolizer phenotypes are classified based on their total AS: individuals with an AS of 0 are poor metabolizers (PMs), those with a score between 0 and 1.25 are intermediate metabolizers (IMs), scores ranging from 1.25 to 2.25 indicate normal metabolizers (NMs), and scores above 2.5 signify ultrarapid metabolizers (UMs).

Identification and curation of novel haplotypes

The initial flagging of samples with potential novel haplotypes was conducted using StellarPGx. A total of 93 samples were flagged for detailed investigation to infer potential novel alleles. The first step involved identifying additional variants apart from the core variants reported by StellarPGx and ensuring that the combination of variants was not already documented in PharmVar. For samples that passed this filter, we manually inspected BAM files using the Integrative Genomics Viewer (IGV) to verify the expected number of reads supporting the reference (REF) and alternative (ALT) alleles for the additional core variant (e.g., allelic depth in the 0.4–0.6 range for heterozygous variants). To determine the background allele, phasing was achieved through two primary methods. Firstly, if the same background allele was present on both chromosomes, that allele was immediately assigned (e.g., *2 reported as the background allele if the predicted diplotype is *2/*2). Alternatively, if occurrences of the same additional core variant were detected in multiple samples, the reported diplotypes were inspected to identify common ones (e.g. *2 reported as the background allele if one participant had a diplotype of *2/*10 and another had *1/*2, and both shared the same additional core variant). Following this curation process, the original sample set was refined to 28 study samples containing 14 distinct novel alleles. For additional variants that are not documented in PharmGKB we inferred the variant classification using Variant Effect Predictor (VEP)[21].

Data availability

The authors of this study do not own the rights to the SG10K_Health dataset, and this dataset is under controlled access to ensure good data governance, responsible data use, and that the dataset is only used for the intended research purposes in compliance with SG10K_Health study cohort IRB and ethics approval. Users interested in accessing the SG10K_Health individual-level data (WGS and VCF files) are required to submit a Data Access Request outlining the proposed research for approval by the NPM Data Access Committee (DAC), which convenes monthly. The forms and data access policy can be downloaded via the SG10K_Health portal (https://npm.a-star.edu.sg/help/NPM) upon registration with an institutional email address. For more information, users can contact the National Precision Medicine Programme Coordinating Office, A*STAR (contact_npco@gis.a-star.edu.sg). The average turnaround timeframe for a request is 4–6 weeks from receipt of request to receiving a notification outcome from the NPM DAC on whether the application is accepted/rejected/requires amendments. The approved requestor will be asked to sign a non-negotiable data access agreement to ensure that (1) the data are used only for the proposed research purpose, (2) no attempt is made to re-identify the participants, (3) there is no onward sharing of the data to a third party and (4) a standard acknowledgement statement is included in the manuscript. PharmGKB *CYP2D6* frequency database can be accessed online (https://www.pharmgkb.org/page/cyp2d6RefMaterials). Pharmvar *CYP2D6* alleles database can be accessed online (https://www.pharmvar.org/gene/CYP2D6). *CYP2D6* alleles frequency data used in the study is provided within the supplementary information.

Code availability

All custom codes used to perform the analyses reported here are hosted on GitHub (https://github.com/c-BIG/SG10K-CYP2D6.git).

Received: 24 January 2024; Accepted: 15 July 2024
Published online: 20 July 2024

References

1. Taylor, C. *et al.* A review of the important role of CYP2D6 in pharmacogenomics. *Genes-Basel* **11**, 1295 (2020).
2. Gaedigk, A., Casey, S. T., Whirl-Carrillo, M., Miller, N. A. & Klein, T. E. Pharmacogene variation consortium: A global resource and repository for pharmacogene variation. *Clin. Pharmacol. Ther.* **110**, 542–545 (2021).
3. PharmVar - CYP2D6. https://www.pharmvar.org/gene/CYP2D6.
4. Gaedigk, A., Sangkuhl, K., Whirl-Carrillo, M., Klein, T. & Leeder, J. S. Prediction of CYP2D6 phenotype from genotype across world populations. *Genet. Med.* **19**, 69–76 (2017).
5. Guidelines – CPIC. https://cpicpgx.org/%20guidelines/.
6. Twesigomwe, D. *et al.* Characterisation of CYP2D6 pharmacogenetic variation in sub-Saharan African populations. *Clin. Pharmacol. Ther.* https://doi.org/10.1002/cpt.2749 (2022).
7. Fukunaga, K. *et al.* Determination of novel CYP2D6 haplotype using the targeted sequencing followed by the long-read sequencing and the functional characterization in the Japanese population. *J. Hum. Genet.* https://doi.org/10.1038/s10038-020-0815-x (2020).
8. Chan, W. *et al.* CYP2D6 allele frequencies, copy number variants, and tandems in the population of Hong Kong. *J. Clin. Lab. Anal.* **33**, e22634 (2019).
9. Dorji, P. W., Tshering, G. & Na-Bangchang, K. CYP2C9, CYP2C19, CYP2D6 and CYP3A5 polymorphisms in South-East and East Asian populations: A systematic review. *J. Clin. Pharm. Ther.* **44**, 508–524 (2019).
10. Twesigomwe, D. *et al.* A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: A focus on CYP2D6 genotyping. *NPJ Genom. Med.* **5**, 30 (2020).
11. Wong, E. *et al.* The Singapore National Precision Medicine Strategy. *Nat. Genet.* **55**, 178–186 (2023).
12. Chen, X. *et al.* Cyrius: Accurate CYP2D6 genotyping using whole-genome sequencing data. *Pharmacogenomics J.* **21**, 251–261 (2021).
13. Hari, A. *et al.* An efficient genotyper and star-allele caller for pharmacogenomics. *Genome Res.* **33**, 61–70 (2023).
14. Twesigomwe, D. *et al.* StellarPGx: A nextflow pipeline for calling star alleles in cytochrome P450 genes. *Clin. Pharmacol. Ther.* https://doi.org/10.1002/cpt.2173 (2021).

15. Accurate CYP2D6 star (*) allele diplotyping for long-read PacBio HiFi sequencing. https://pacbio.cn/wp-content/uploads/poster_harting.pdf.
16. Goh, L. L., Lim, C. W., Sim, W. C., Toh, L. X. & Leong, K. P. Analysis of genetic variation in CYP450 genes for clinical implementation. *Plos One* **12**, e0169233 (2017).
17. CYP2D6 Overview: Allele and Phenotype Frequencies - Medical Genetics Summaries—NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK574601/.
18. Clinical Pharmacogenetics Implementation Consortium Guideline for CYP2D6, OPRM1, and COMT Genotypes and Select Opioid Therapy. https://files.cpicpgx.org/data/guideline/publication/opioids/2020/33387367.pdf.
19. Whirl-Carrillo, M. *et al.* An evidence-based framework for evaluating pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **110**, 563–572 (2021).
20. Relling, M. V. *et al.* The clinical pharmacogenetics implementation consortium: 10 years later. *Clin. Pharmacol. Ther.* **107**, 171–175 (2020).
21. McLaren, W. *et al.* The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).

Acknowledgements

This study made use of data generated as part of the Singapore National Precision Medicine program funded by the Industry Alignment Fund (Pre-Positioning) (IAF-PP: H17/01/a0/007). This study made use of data/samples collected in the following cohorts in Singapore: The Health for Life in Singapore (HELIOS) study at the Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore (supported by grants from a Strategic Initiative at Lee Kong Chian School of Medicine, the Singapore Ministry of Health (MOH) under its Singapore Translational Research Investigator Award (NMRC/STaR/0028/2017) and the IAF-PP:H18/01/a0/016); The Growing up in Singapore Towards Healthy Outcomes (GUSTO) study, which is jointly hosted by the National University Hospital (NUH), KK Women's and Children's Hospital (KKH), the National University of Singapore (NUS) and the Singapore Institute for Clinical Sciences (SICS), Agency for Science Technology and Research (A*STAR) (supported by the Singapore National Research Foundation under its Translational and Clinical Research (TCR) Flagship Programme and administered by the Singapore Ministry of Health's National Medical Research Council (NMRC), Singapore -NMRC/TCR/004- NUS/2008; NMRC/TCR/012-NUHS/2014. Additional funding is provided by SICS and IAF-PPH17/01/a0/005); The Singapore Epidemiology of Eye Diseases (SEED) cohort at Singapore Eye Research Institute (SERI) (supported by NMRC/CIRG/1417/2015; NMRC/CIRG/1488/2018; NMRC/OFLCG/004/2018); The Multi-Ethnic Cohort (MEC) cohort (supported by NMRC grant 0838/2004; BMRC grant 03/1/27/18/216; 05/1/21/19/425;11/1/21/19/678, Ministry of Health, Singapore, National University of Singapore and National University Health System, Singapore); The SingHealth Duke-NUS Institute of Precision Medicine (PRISM) cohort (supported by NMRC/CG/M006/2017_NHCS; NMRC/STaR/0011/2012, NMRC/STaR/ 0026/2015, Lee Foundation and Tanoto Foundation); The TTSH Personalised Medicine Normal Controls (TTSH) cohort funded (supported by NMRC/CG12AUG17 and CGAug16M012). The views expressed are those of the author(s) are not necessarily those of the National Precision Medicine investigators, or institutional partners. We thank all investigators, staff members and study participants who made the National Precision Medicine Project possible.

Author contributions

N.B. and M.G.P. designed the research; N.B. and R.T.J. acquired the data; Y.M., M.G.P. and R.T.J. analyzed the data and wrote the manuscript; Y.M., M.G.P., R.T.J., N.B. and D.T. interpreted the data; L.S. and A.I supervised the research; D.T. revised the critical revision of the manuscript.

Competing interests

Y.M., M.G.P., L.S. and A.I. are employees of NalaGenetics; R.T.J., N.B. and D.T. declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-67691-6.

Correspondence and requests for materials should be addressed to N.B. or M.G.-P.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.