deCYPher: Star Allele-Resolution Computational Framework of Pharmacogenes for Haplotype-Resolved Long-Read Assemblies

Ting-Yu Chang[1], Yu-Shi Liu[1], Hong-Sheng Lai[2], Tsung-Kai Hung[1], Hsin-Fu Lin[1], Yu-Hui Lin[3], Chia-Lang Hsu[1,4], Ya-Chien Yang[5,6], Chein-Yu Chen[2], Pei-Lung Chen[1,3,7,8,9], Jacob Shu-Jui Hsu[1*]

Graduate Institute of Medical Genomics and Proteomics, College of Medicine, National Taiwan University, Taipei, Taiwan[1]

Department of Biomechatronics Engineering, National Taiwan University, Taipei, Taiwan[2]

Department of Medical Genetics, National Taiwan University Hospital, Taipei, Taiwan[3]

Department of Medical Research, National Taiwan University Hospital, Taipei, Taiwan[4]

Department of Clinical Laboratory Sciences and Medical Biotechnology, College of Medicine, National Taiwan University, Taipei, Taiwan[5]

Department of Laboratory Medicine, National Taiwan University Hospital, Taipei, Taiwan[6]

Genome and Systems Biology Degree Program, Academia Sinica and National Taiwan University, Taipei, Taiwan[7]

Graduate Institute of Clinical Medicine, College of Medicine, National Taiwan University, Taipei, Taiwan[8]

Division of Endocrinology and Metabolism, Department of Internal Medicine, National Taiwan University Hospital, Taipei, Taiwan[9]

* Corresponding authors

J.S.H: jacobhsu@ntu.edu.tw

Running title

Pharmacogene Allele Profiling for all HPRC Genomes

Abstract

Although existing next-generation sequencing (NGS) tools, such as Aldy and Cyrius, have been applied for allele typing, they cannot achieve complete accuracy due to various genomic challenges including pseudogenes, structural variations, hybrid genes, copy number variations, and gene deletions. These complexities make accurate pharmacogene interpretation more challenging, despite the crucial role pharmacogenomics plays in precision medicine. We developed **deCYPher**, a tool that generates personalized pharmacogenomic reports from haplotype-resolved assemblies. The tool enables analysis of all PharmVar 1A level genes, such as *CYP2B6, CYP2C9, CYP2C19, CYP2D6, CYP3A5, CYP4F2, DPYD, NUDT15,* and *SLCO1B1*. Applied to all HPRC haplotypes (including both release 1 and release 2 data), deCYPher demonstrated high accuracy in resolving complex gene structures. In the case of CYP2D6, release 1 identified 6% gene multiplications, 6% full gene deletions, and 4% CYP2D6/CYP2D7 hybrids. By contrast, release 2 demonstrated an increased prevalence of multiplications (14%) and hybrids (11%), while the frequency of full gene deletions remained comparable at 5%. Comparison with pb-StarPhase revealed discrepancies in 12 of 94 assemblies in the release 1 dataset. For instance, in sample HG02257, Aldy, Cyrius, and deCYPher consistently identified the genotype as *2/*35, whereas pb-StarPhase reported *2/*2. Notably, the *35-defining variants were present in the BAM and VCF files in the pb-StarPhase pipeline, but the local read depth over the *35-specific region was only 5x in HG02257-p, suggesting that the misclassification likely resulted from insufficient coverage - a known limitation of pb-StarPhase under low-depth conditions.

53 **Introduction**

54 Pharmacogenomics investigates the genetic factors influencing drug efficacy and toxicity,
55 with pharmacogenes identified as key contributors to therapeutic outcomes. The Pharmacogene
56 Variation (PharmVar) Consortium curates star-allele definitions and structural-variant
57 annotations to enable reproducible allele calling (Gaedigk et al., 2018), while CPIC and
58 PharmGKB provide evidence-based drug-dosing and reporting guidance that depends on
59 accurate genotype calls (Relling & Klein, 2011; Whirl-Carrillo et al., 2021). PharmVar
60 currently catalogues pharmacogenes such as *CYP1A2, CYP2A6, CYP2A13, CYP2B6, CYP2C8,*
61 *CYP2C9, CYP2C19, CYP2D6, CYP3A4, CYP3A5, CYP4F2, DPYD, NUDT15, NAT2*, and
62 *SLCO1B1*, offering standardized allele definitions across major drug-metabolizing enzymes
63 and transporters. Concurrently, CPIC has published clinical guidelines for an expanded array
64 of genes—including *ABCG2, CACNA1S, CFTR, CYP2B6, CYP2C8, CYP2C9, CYP2C19,*
65 *CYP2D6, CYP3A4, CYP3A5, CYP4F2, DPYD, G6PD, HLA-A, HLA-B, HLA-C, IFNL3, IFNL4,*
66 *MT-RNR1, NUDT15, RYR1, SLCO1B1, TPMT, UGT1A1*, and *VKORC1*. Collectively, these
67 complementary resources establish a molecular basis for the integration of pharmacogenomics
68 into clinical practice.

69 This tool enables comprehensive analysis of all PharmVar level 1A genes. Within this
70 group, cytochrome P450 (CYP) enzymes play a significant role, as they are involved in phase
71 I oxidative metabolism for many commonly prescribed drugs and contribute to interindividual
72 differences in pharmacokinetics and drug response (Zanger & Schwab, 2013). Reviews of the
73 CYP superfamily summarize how genetic variation and regulation of expression in these
74 enzymes shape drug metabolism phenotypes and influence efficacy, adverse drug reactions,
75 and drug–drug interactions.

76 Among CYP and related pharmacogenes, loci such as *CYP2D6*, *CYP2B6, CYP2C9,*
77 *CYP2C19*, and others are highly polymorphic. They are well-established predictors of clinically
78 actionable metabolic phenotypes (for example, poor, intermediate, normal, rapid, or ultrarapid
79 metabolizers) used in therapeutic decision-making. These genotype–phenotype translations
80 underpin numerous Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines
81 and PharmGKB clinical annotations that convert molecular test results into prescribing
82 recommendations.

83 Accurate star-allele assignment for these genes is complicated by multiple genetic
84 variation types. In addition to single-nucleotide variants (SNVs) and small indels, clinically
85 important structural variation occurs frequently — including whole-gene deletions and
86 duplications, hybrid alleles created by gene–pseudogene conversions, and complex
87 rearrangements — all of which can have large effects on enzyme activity. *CYP2D6* serves as a
88 classic example; it is located in a region with related pseudogenes such as *CYP2D7* and features
89 numerous structural alterations. Approximately 5-10% of people have copy-number variants
90 like gene deletions or duplications, while *CYP2D6/CYP2D7* hybrid alleles are found in about
91 2-6%. Altogether, these variations make up more than 10% of clinically significant differences
92 in *CYP2D6* (Hicks et al., 2015; Relling & Klein, 2011; Whirl-Carrillo et al., 2021).

93 Conventional short-read sequencing and basic variant-calling pipelines frequently
94 misidentify structural variants, leading to inaccurate star-allele calls and incorrect downstream
95 enzyme activity assignments. Consequently, substantial inter-laboratory and inter-algorithm
96 discrepancies in phenotype assignment often originate from differences in the underlying
97 genotype calls, particularly when structural variation is present (Del Tredici et al., 2018;
98 Gaedigk et al., 2018). A host of computational tools and experimental workflows have been
99 developed to tackle pharmacogene genotyping. Short-read–based star-allele callers have
100 achieved high concordance for common alleles (Chen et al., 2021; Hari et al., 2023; Lee et al.,

101 2019; Twesigomwe et al., 2020), but their effectiveness is limited when complex structural
102 rearrangements or high sequence homology to pseudogenes are involved. Recent benchmarking
103 studies indicate that sequence ensemble approaches combined with specialized algorithms
104 perform well. However, complete concordance among all allele classes—including rare
105 structural variants and hybrids—has not been consistently attained (Chen et al., 2021; Qiao et
106 al., 2016). Integrating multiple sequencing techniques, including PacBio HiFi, ONT ultra-long
107 reads, and Illumina Hi-C reads, to construct haplotype-resolved assemblies has demonstrated
108 the ability to resolve complex loci, detect novel structural variants, and provide contiguous
109 haplotypes that are essential for unambiguous allele definition (Cheng et al., 2021; Liao et al.,
110 2023; Nurk et al., 2022).

111 The pb-StarPhase demonstrates a phase-aware pipeline diploid typing of twenty-one
112 clinically important pharmacogenes utilizing PacBio HiFi reads, with the ability to correct or
113 refine historical genotype calls (Holt et al., 2024). Although pb-StarPhase have advanced allele
114 resolution for complex pharmacogenes, their performance can still be affected by sequencing
115 coverage and read length, which may lead to incomplete or uncertain genotype calls under
116 suboptimal data quality. These limitations highlight the need for a targeted and systematic
117 framework that operates consistently across clinically relevant pharmacogenes, integrates
118 PharmVar allele definitions with CPIC and PharmGKB conventions, and delivers high-
119 confidence SNV, CNV, and hybrid allele calls directly from haplotype-resolved genome
120 assemblies.

121 We therefore developed **deCYPher**, a computational method for star-allelic resolution of
122 pharmacogenes listed in CPIC/PharmGKB/PharmVar using haplotype-resolved long-read
123 assemblies, particularly in genes with complex structural variations. We benchmarked
124 deCYPher on nine PharmVar 1A-level genes (*CYP2D6, CYP2B6, CYP2C9, CYP2C19,*
125 *CYP3A5, CYP4F2, DPYD, NUDT15,* and *SLCO1B1*) and compared its results with those of

126 Aldy, Cyrius, pb-StarPhase, and TypeAssembly. Compared to existing short-read or long-read
127 mapping–based approaches, deCYPher leverages fully assembled haplotypes to achieve high
128 accuracy and unambiguous allele definitions, thereby enhancing the reliability of clinical
129 pharmacogenomic reporting.

130 **Results**

131 Method Overview

132 The 47 phased diploid genome assemblies from the Human Pangenome Reference
133 Consortium (HPRC), were analyzed with our **deCYPher** pipeline (Fig.1). The deCYPher was
134 designed to achieve the highest resolution for allele assignments defined by the PharmVar
135 Consortium. The standardized CYP nomenclature — comprising the *CYP* root, family number,
136 subfamily letter, gene number, and a star allele designation — was used to ensure consistency
137 with established pharmacogenomic conventions.

138 For each haplotype-resolved assembly, deCYPher first identified the intervals of target
139 pharmacogenes and mapped curated reference sequences representing star alleles ("star-one"
140 references) to the phased contigs. Alignments were performed at full-sequence resolution to
141 capture both single-nucleotide variants (SNVs) and small indels. Alleles matching known
142 PharmVar definitions were directly assigned. When variant patterns did not match any existing
143 alleles, deCYPher flagged them as potential novel alleles, with particular attention to previously
144 unreported structural variants (SVs) such as novel hybrid genes or deletion genes.

145 We validated deCYPher-identified structural variants by comparing them to genotypes
146 from Aldy, Cyrius, and pb-StarPhase. Cross-comparison confirmed the structural variants.

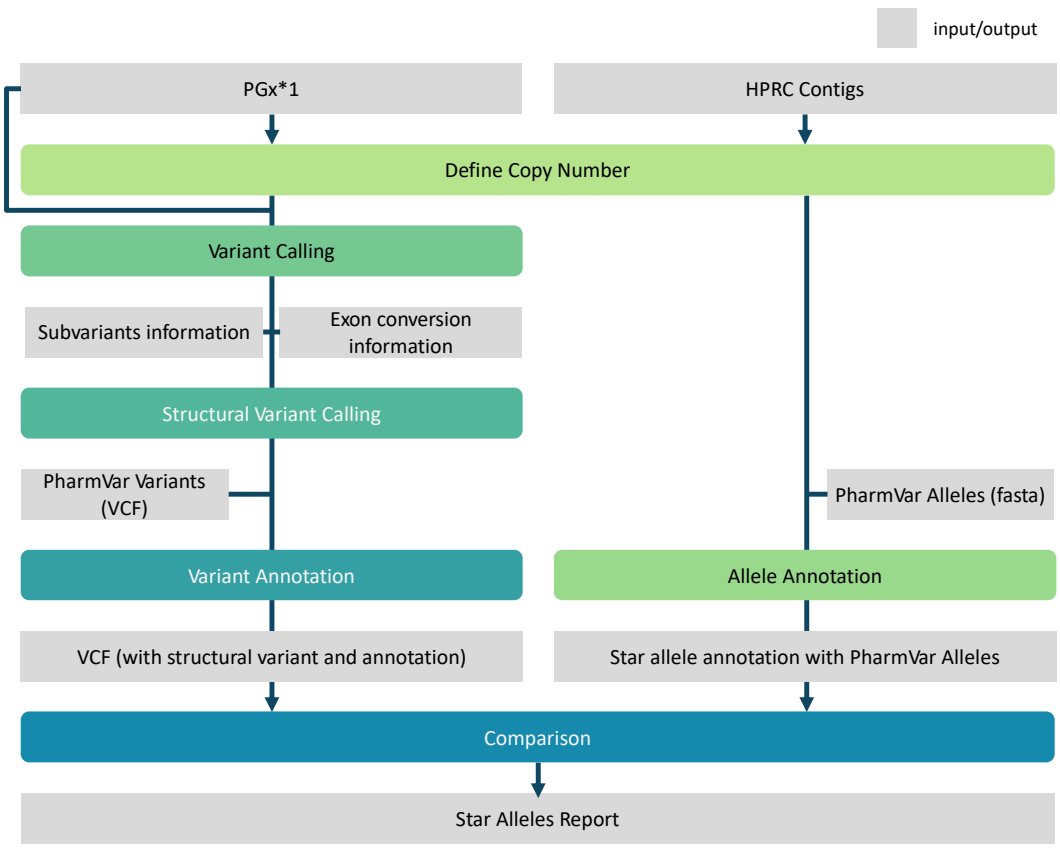


Figure 1. Workflow of the deCYPher pipeline for CYP allele annotation in diploid genome assemblies.

The pipeline integrates the long-read alignment tool minimap2 (Li, 2018) with custom Python and Shell scripts to map allele sequences from the Pharmacogene Variation Consortium (PharmVar) database onto genome assemblies, including the HPRC Year-1 and Year-2 releases. This approach enables accurate identification of CYP gene alleles on phased contigs, achieving up to 95% accuracy in high-resolution allele assignment.

147

148 Overview of the HPRC CYP Year-1 and Year-2 releases haplotypes and phenotypes

149 We fully characterized the high-resolution alleles of CYP genes for all known CYP alleles
150 in Year-1 47 HPRC human genome assemblies and Year-2 232 HPRC human genome
151 assemblies (Fig.2A-B). According to the CYP genotype definitions from the Pharmacogene
152 Variation Consortium (PharmVar; https://www.pharmvar.org/), the CYP2D6 gene includes

185star alleles, encompassing numerous hybrid and deletion alleles, as well as copy number variants (CNVs) (Taylor et al., 2020). Our results show that, among the 94 CYP2D6 haplotypes in the HPRC Year-1 dataset, 6 haplotypes (6.38%, 6/94) harbor one or more CNVs. Additionally, 6 haplotypes (6.38%, 6/94) were identified as full gene deletions, as illustrated in Figure 2C. Similarly, in the HPRC Year-2 dataset, 66 out of 464 CYP2D6 haplotypes (14.2%, 66/464) contained one or more CNVs, while 22 haplotypes (4.74%, 22/464) were classified as full gene deletions, as shown in Figure 2C.

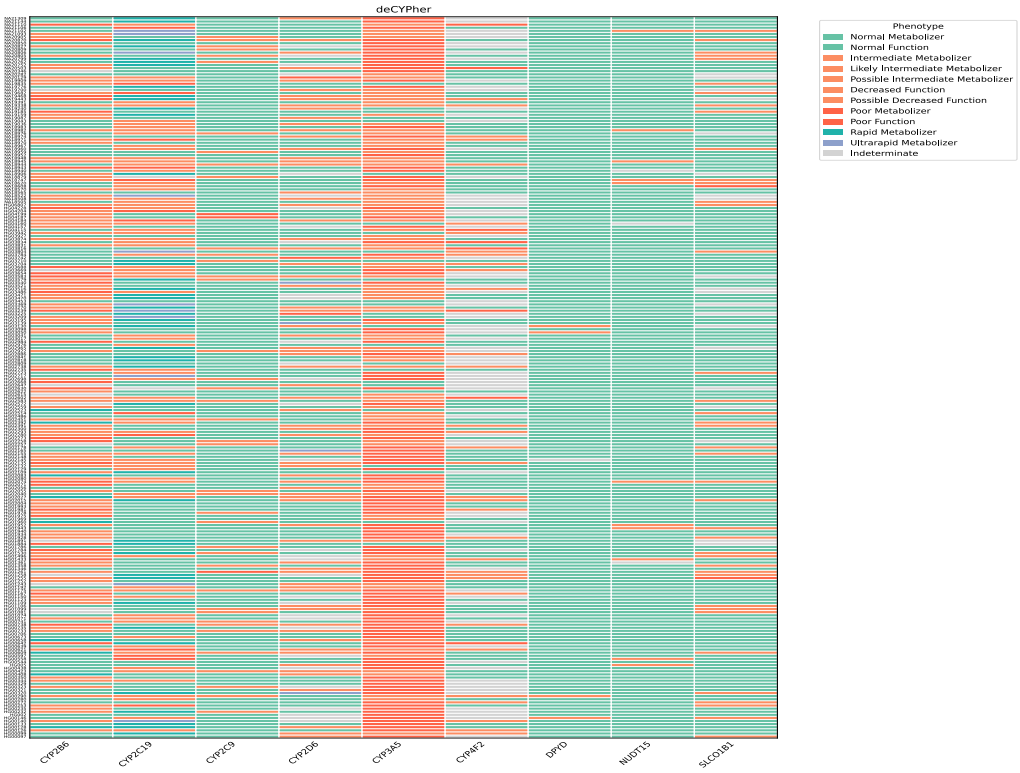
Regarding phenotype predictions, 24 samples (51.06%, 24/47) in the HPRC Year-1 assemblies were classified as Normal Metabolizers, whereas in the HPRC Year-2 assemblies, 121 samples (52.16%, 121/232) fell into this category (detailed data are provided in Supplemental Tables 1 and 2). These findings address the limitations of previous technologies, which could not resolve CYP gene structural variations at base-pair resolution across whole-genome sequences, highlighting the importance of deCYPher for the comprehensive characterization of pharmacogene genetic variation.

In this study, we denote paternal haplotypes as SampleID-p and maternal haplotypes as SampleID-m, where SampleID corresponds to the individual sample. For example, HG002-p and HG002-m represent the paternal and maternal haplotypes of sample HG002, respectively.

A.

deCYPher
NA21309
NA20129
NA19240
NA18906
HG03579
HG03540
HG03516
HG03492
HG03486
HG03453
HG03098
HG02886
HG02818
HG02723
HG02717
HG02630
HG02622
HG02572
HG02559
HG02486
HG02257
HG02148
HG02145
HG02109
HG02080
HG02055
HG01978
HG01952
HG01928
HG01891
HG01361
HG01358
HG01258
HG01243
HG01175
HG01123
HG01109
HG01106
HG01071
HG00741
HG00735
HG00733
HG00673
HG00621
HG005
HG00438
HG002
CYP2B6
CYP2C19
CYP2C9
CYP2D6
CYP3A5
CYP4F2
DPYD
NUDT15
SLCO1B1
Phenotype
Normal Metabolizer
Normal Function
Intermediate Metabolizer
Likely Intermediate Metabolizer
Possible Intermediate Metabolizer
Decreased Function
Possible Decreased Function
Poor Metabolizer
Poor Function
Rapid Metabolizer
Ultrarapid Metabolizer
Indeterminate

B.


deCYPher
CYP2B6
CYP2C19
CYP2C9
CYP2D6
CYP3A5
CYP4F2
DPYD
NUDT15
SLCO1B1
Phenotype
Normal Metabolizer
Normal Function
Intermediate Metabolizer
Likely Intermediate Metabolizer
Possible Intermediate Metabolizer
Decreased Function
Possible Decreased Function
Poor Metabolizer
Poor Function
Rapid Metabolizer
Ultrarapid Metabolizer
Indeterminate

C.

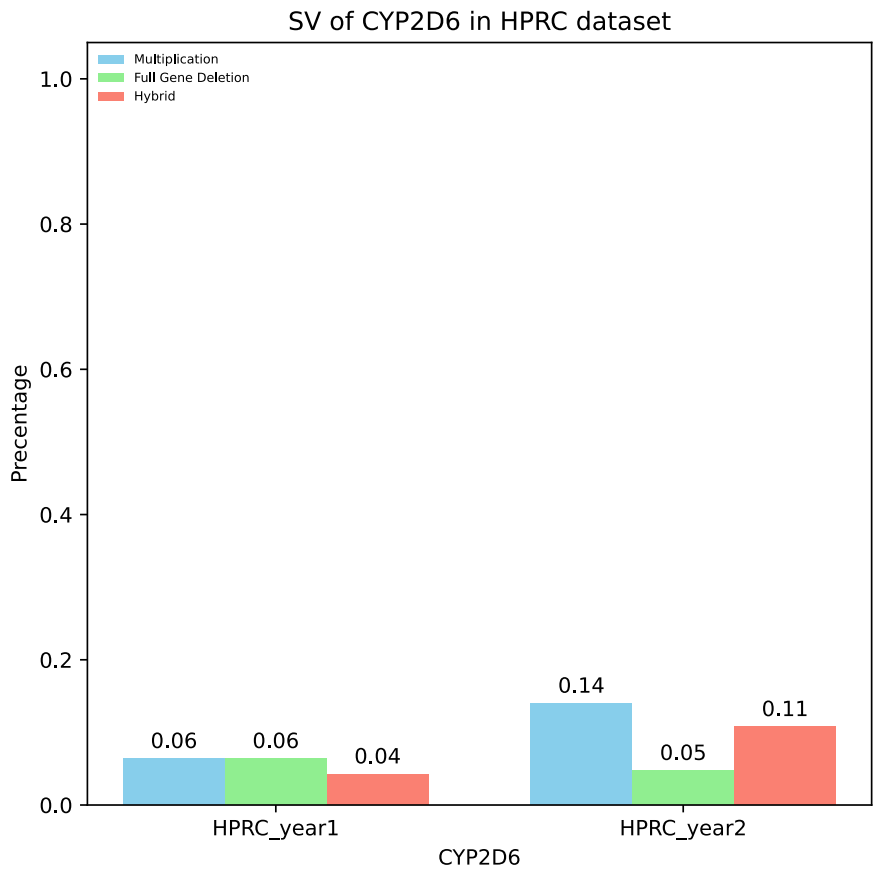


Figure 2. Phenotypic profiles of pharmacogenes and structural variants (SVs) of *CYP2D6* identified in the 47 phased assemblies and 232 phased assemblies analyzed by deCYPher.

(A) Heatmap of predicted pharmacogene phenotypes across the HPRC Year-1 cohort.

(B) Heatmap of predicted pharmacogene phenotypes across the HPRC Year-2 cohort.

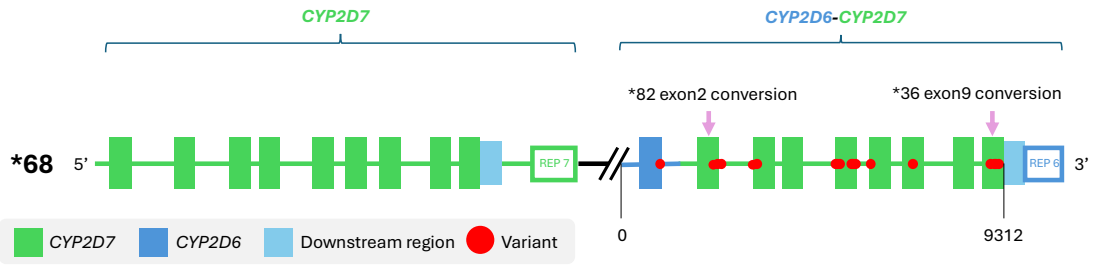
(C) Structural variant configurations of *CYP2D6*, including full-gene deletions, duplications, and gene–pseudogene hybrid arrangements, as detected in the HPRC Year-1,-2 assemblies.

Within the HPRC assemblies, in addition to *CYP2D6*, we also identified structural variants in other pharmacogenes. For example, in HG01993-m we detected a *CYP2B6* hybrid gene, while in HG00423-m we observed a full-gene deletion of *CYP3A5* — a variant that has never been previously reported. Furthermore, in HG02615-m we identified a CYP4F2 copy number variation (CNV), which is also novel, and in HG02615-p we found a full-gene deletion

181 of CYP4F2. Notably, all of these non-*CYP2D6* structural variants were discovered in the HPRC
182 Year-2 phased assemblies. This suggests that expanding the size of cohorts with PacBio HiFi
183 phased assemblies may enable the discovery of additional low allele frequency and novel
184 structural variants, further underscoring the importance of **deCYPher** for structural variant
185 analysis in pharmacogenes (Supplemental Tables 3 and 4).
186
187 Distinctive variant positions defining the CYP2D6*68 hybrid allele

188 We identified a *CYP2D6* hybrid allele, CYP2D6 *68, in five phased assemblies from the
189 HPRC dataset: HG00733-p, HG01258-p, HG01784-m, NA20805-p, and NA20827-m.
190 According to PharmVar, *68 is defined by a CYP2D7-to-CYP2D6 gene conversion spanning
191 exons 2 to 9. By contrast, *36 is characterized by a conversion limited to exon 9, and *82 by a
192 conversion restricted to exon 2. Consequently, any haplotype carrying the exon 2–9 conversion
193 of *68 will inherently meet the defining criteria of both *36 (exon 9 conversion) and *82 (exon
194 2 conversion). Based on this relationship, our initial strategy was to classify any assembly
195 carrying both *36 and *82 as *68 (Fig. 3A). Using this approach, we detected three *68 alleles
196 in the HPRC Year-1 phased assemblies—one in HG00733-p and two in HG01258-p—yielding
197 results that were fully concordant with calls from Aldy and Cyrius.

198 Although carrying both *36 and *82 as a *68 criterion successfully identified *68, it may
199 not capture all instances of this hybrid allele, as *68 involves a continuous exon 2–9 conversion.
200 To refine detection, we sought variant positions that could serve as a more complete signature
201 for *68. We first compiled all star alleles present across confirmed *68 assemblies and found
202 consistent co-occurrence of the following alleles: *10, *36, *39, *49, *53, *74, *82, *83, *86,
203 *88 (with one variant site optionally absent), *131, *132, *139, and *176. We then recorded
204 the variant positions corresponding to these alleles, excluding NG_008376.4:6032 T>C, to
205 generate a representative *68 variant profile (VCF) (Fig. 3B and Supplemental Table 5).



POS	REF	ALT	POS	REF	ALT	POS	REF	ALT	POS	REF	ALT	POS	REF	ALT
5119	C	T	6046	A	G	7630	T	A	9151	A	G	9183	C	CG
5992	C	A	6631	T	A	7838	A	G	9153	T	C	9185	T	C
6002	A	G	6637	G	T	8285	G	A	9175	C	T	9186	T	A
6040	A	T	7486	T	C	9144	G	C	9176	A	C	9191	TC	T
6041	C	A	7626	G	A	9148	C	G	9178	G	C	9200	G	C

Figure 3. Detection and variant profile of CYP2D6*68.

CYP2D6*68 is characterized by a continuous CYP2D7-to-CYP2D6 conversion spanning exons 2–9, in contrast to *36 (exon 9 only) and *82 (exon 2 only). Assemblies harboring both *36 and *82 were classified as *68. Variant positions defining the *68 hybrid allele across exons 2–9 are shown, with each marker indicating a characteristic nucleotide change.

Fusion of *CYP2B6*

Previous studies have shown that the **CYP2B6*29** allele shares complete sequence identity with exons 1–4 of ***CYP2B7*** and exons 5–9 of ***CYP2B6***, and is associated with reduced enzymatic activity. In the PharmVar database, CYP2B6*29 is annotated as a hybrid gene between *CYP2B7* and *CYP2B6*. Furthermore, the GenBank database contains FASTA sequences representing the *CYP2B7/CYP2B6* breakpoint junction, as documented in earlier reports (Rotger *et al.* 2007; Martis *et al.* 2013; Gaedigk *et al.* 2018).

Although the genomic characterization of **CYP2B6*29** is already well documented, current genotyping tools such as **Cyrius**, **Aldy**, and **pb-StarPhase** perform suboptimally in detecting this allele. In both the HPRC Year-1 and Year-2 phased assemblies, no samples were

classified as CYP2B6*29 by these tools. However, when applying **deCYPher** to the HPRC Year-2 phased assemblies, we identified **HG01993-m** as carrying the CYP2B6*29 allele.

deCYPher recognized this allele based on two key observations. First, we detected abnormal mapping patterns when aligning the HG01993-m contigs to the **CYP2B6*1** reference: read mapping with generally poor mapping quality, except for the primary mapping segment spanning **exons 5–9**, where mapping quality was higher, whereas **exons 1–4** were predominantly soft-clipped (Fig. 4A). This mapping profile is consistent with the known structure of CYP2B6*29. Second, when aligning HG01993-m to the *CYP2B7/CYP2B6* breakpoint junction sequence, we observed almost perfect alignment across the junction with no evidence of soft or hard clipping, and only a few single-nucleotide variants (SNVs) (Fig. 4B). These Integrative Genomics Viewer (IGV)-based visualizations (Robinson et al., 2011) provide orthogonal evidence that deCYPher correctly identified HG01993-m as CYP2B6*29. Existing algorithms missed this call. This demonstrates deCYPher's unique ability to resolve hybrid alleles in pharmacogenes that are often misclassified or overlooked by short-read–based tools.

A.

B.

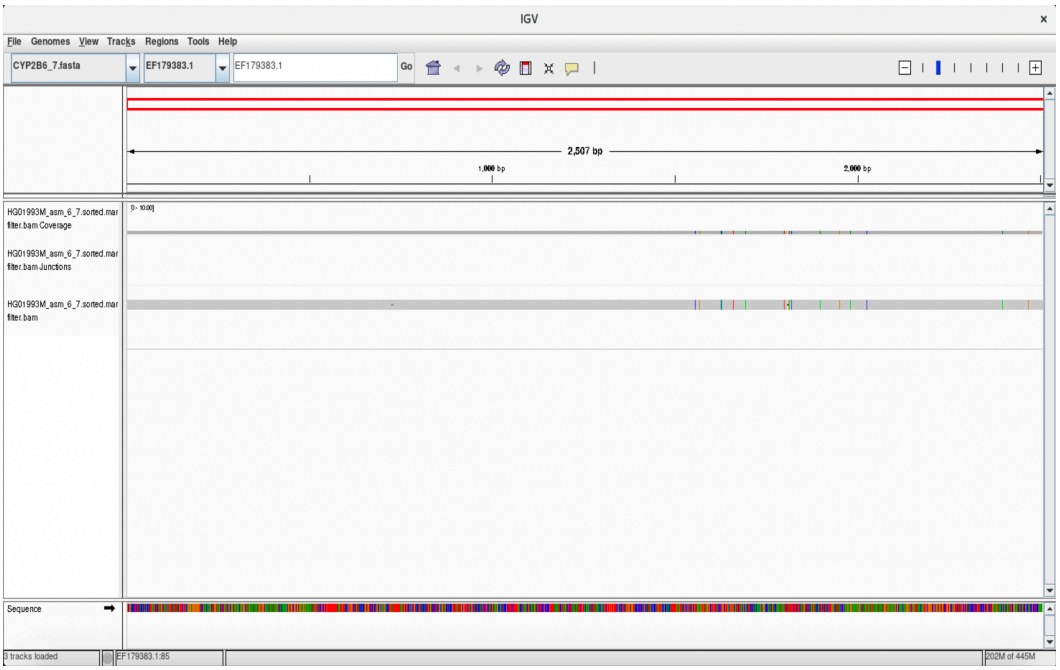


Figure 4. IGV visualization of CYP2B6*29 in HG01993-m.

(A) IGV alignment view of the HG01993-m contigs against the CYP2B6*1 reference. The region corresponding to exons 1–4 displayed extensive soft clipping and poor mapping quality, whereas alignment improved across exons 5–9, consistent with the CYP2B6*29 hybrid structure.

238

239 Comparisons with pb-StarPhase in HPRC genomes

240 pb-StarPhase (Holt et al., 2024) provides an innovative approach for pharmacogene
241 genotyping using PacBio long-read whole-genome sequencing. To directly compare its
242 performance with deCYPher, we applied the recommended pb-StarPhase pipeline to the 47
243 HPRC Year-1 samples, starting from randomly selected long-read raw data. We then generated
244 diplotype calls for all samples and compared them against deCYPher results.

245 During this comparison, we observed that pb-StarPhase frequently produced
246 "NO_MATCH" calls. For the purpose of evaluation, such calls were treated as discordances
247 and provisionally considered errors on the pb-StarPhase side. For other discordant cases, we
248 examined upstream evidence (i.e., BAM and VCF files) to adjudicate correctness. Specifically,
249 if a critical variant was clearly present in the alignments and variant calls but missing from the
250 output, we classified the call as incorrect. Conversely, if a critical variant was clearly present
251 in the alignments and variant calls and also present in the output, we classified the call as correct.
252 In cases where both tools produced plausible but differing outputs supported by partial evidence,
253 we labeled the discrepancy as "Unknown", reflecting unresolved classification.

254 As an illustrative example, in HG03516, both the pbmm2-aligned bam file and the
255 Deepvariant vcf file contained the variant chr22:42129770 G>A, a defining variant of
256 CYP2D6*17 (Fig. 5). Nevertheless, pb-StarPhase assigned this sample as CYP2D6*2/*2,
257 missing the variant, whereas deCYPher correctly identified the diplotype as CYP2D6*17/*2.

258 Overall, deCYPher demonstrated superior accuracy for *CYP2D6*, achieving a precision of
259 94.68%, compared to 87.23% for pb-StarPhase — the largest discrepancy observed among the
260 nine pharmacogenes analyzed. Notably, both tools produced identical results for *CYP2C9* and
261 *NUDT15* (Table 1).

These findings highlight the limitations of pb-StarPhase when allele-defining, while underscoring the advantage of deCYPher in leveraging haplotype-resolved assemblies to deliver unambiguous star-allele assignments. This accuracy is particularly critical for complex loci such as *CYP2D6*, where reliable genotype-to-phenotype translation is essential for clinical pharmacogenomics.

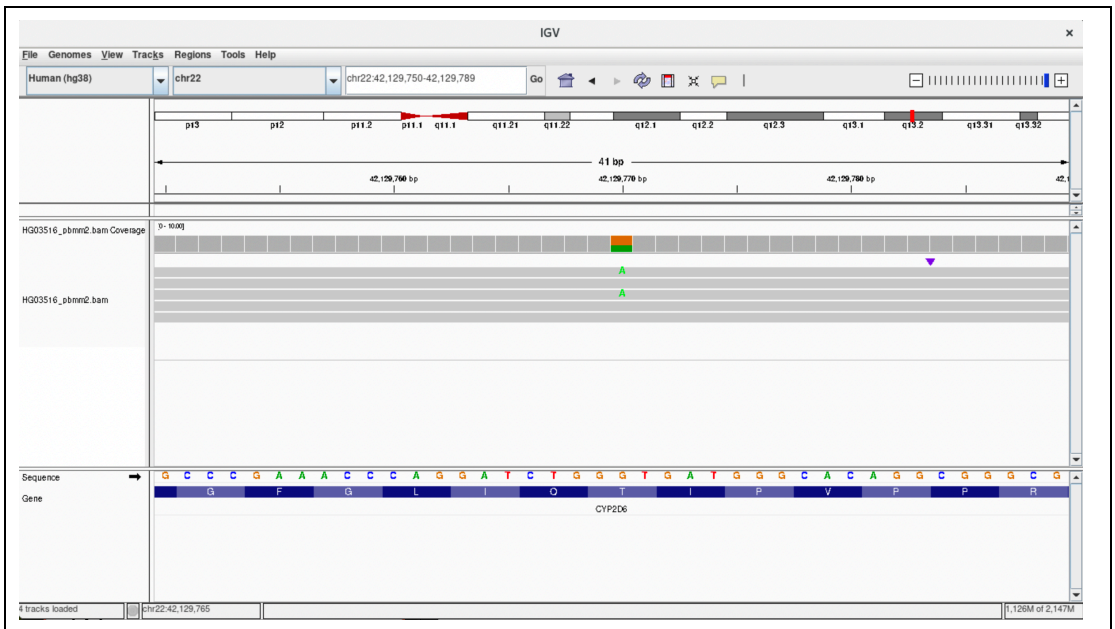


Figure 5. IGV visualization of the CYP2D*17-defining variant in HG03516.

This track shows clear evidence of the chr22:42129770 G>A substitution. Despite this, pb-StarPhase assigned the sample as CYP2D6*2/*2, whereas deCYPher correctly identified it as CYP2D6*17/*2.

Table 1. Comparison between deCYPher and pb-StarPhase.

	Total Differences	deCYPher Correct	Unknown	deCYPher Precision	pbStarPhase Precision
CYP2B6	6	6/6(100%)	0(0%)	94/94(100%)	84/94(89.36%)
CYP2C19	3	1/3(33.3%)	2/3(66.7%)	93/94(98.94%)	89/94(94.68%)
CYP2D6	12	7/12(58.3%)	5/12(41.7%)	89/94(94.68%)	82/94(87.23%)
CYP3A5	4	2/4(50%)	2/4(50%)	92/94(97.87%)	90/94(95.74%)
CYP4F2	2	2/2(100%)	0(0%)	94/94(100%)	92/94(97.87%)
DPYD	3	3/3(100%)	0(0%)	94/94(100%)	91/94(96.81%)
SLCO1B1	2	1/2(50%)	1/2(50%)	93/94(98.94%)	92/94(97.87%)
CYP2C9	0			94/94(100%)	94/94(100%)
NUDT15	0			94/94(100%)	94/94(100%)

Discussion

In this study, we developed and applied **deCYPher**, a computational framework for star allele typing of pharmacogenes using haplotype-resolved assemblies. By leveraging long-read contiguity and phasing, deCYPher enables accurate assignment of PharmVar-defined alleles, detection of structural variants, and standardized diplotype-to-phenotype translation consistent with CPIC conventions. Compared to existing genotyping tools such as Aldy, Cyrius, and pb-StarPhase (Chen et al., 2021; Hari et al., 2023; Holt et al., 2024), deCYPher demonstrated improved accuracy, particularly in loci with complex structural variation.

Hybrid alleles of *CYP2D6*, such as *36 and *82, have been well characterized in previous studies (Gaedigk et al., 2018; Gaedigk et al., 2017). In line with these reports, we identified CYP2D6*68, a complex hybrid allele defined in PharmVar as a gene conversion event spanning exons 2–9 from *CYP2D7* (Gaedigk et al., 2018). Importantly, our analysis also revealed structural variants outside of *CYP2D6*, including a CYP2B6*29 hybrid in HG01993-m, which, to our knowledge, has not been reported in other tools before. These findings highlight the strength of deCYPher in resolving both known and unclear structural variants across pharmacogenes, particularly when complex scenarios involving CNVs, hybrid genes, or large deletions complicate haplotype assignment.

These findings highlight a broader limitation of short-read pipelines. Previous work has shown that genotype-to-phenotype assignments can differ substantially across laboratories and algorithms, especially in the presence of structural variants (Twesigomwe et al., 2020). In such cases, ambiguity often arises not because phenotype rules differ, but because the underlying genotype calls are inconsistent. By integrating haplotype-resolved assemblies with PharmVar

293 allele definitions, deCYPher substantially reduces this discordance and produces more reliable
294 haplotype calls.

295 Our benchmarking against HPRC assemblies further supports this advantage. While pb-
296 StarPhase achieved a precision of 8% for *CYP2D6* calls, deCYPher achieved 94.7%,
297 representing the largest performance gap across the nine genes evaluated. In two genes
298 (*CYP2C9* and *NUDT15*), both tools produced identical results, indicating that the primary
299 benefit of deCYPher emerges in genes with extensive structural complexity. These results are
300 consistent with prior observations that structural variation is a major source of error in
301 pharmacogene genotyping (Twesigomwe et al., 2020; Zhou et al., 2017).

302 Finally, our work parallels advances in other structurally complex genomic loci, such as
303 the KIR region. The SKIRT pipeline has demonstrated how leveraging high-quality assemblies
304 enables the discovery of novel gene fusions and CNVs in KIR genes (Hung et al., 2024; Zhou
305 et al., 2024). By analogy, deCYPher provides a foundation for high-resolution pharmacogene
306 genotyping, paving the way for future studies in larger HiFi-based cohorts and the systematic
307 discovery of rare or novel pharmacogene SVs.

308

309

310 **Methods**

311 deCYPher - Algorithms of the Pharmacogenes annotation Pipeline

312 *Read Mapping and Variant Detection*

313 The deCYPher pipeline was developed to resolve PharmVar-defined star alleles and
314 structural variation, including gene duplications and deletions. To capture multiple
315 polymorphic alleles at each locus, we aligned PharmVar reference allele sequences to

haplotype-resolved genome assemblies using **minimap2** (Li, 2018) optimized for PacBio HiFi data. Parameter settings and alignment configurations are provided in **Supplementary Note 2**.

To validate structural variants identified by deCYPher, we compared the results with genotypes obtained from established tools, including Aldy, Cyrius, and pb-StarPhase. The input data consisted of short-read Illumina whole-genome sequencing (WGS) from the 1KGP and HPRC cohorts, as well as long-read PacBio HiFi sequencing. After cross-comparing these results, we evaluated and confirmed the presence of structural variants.

Output and Visualization

deCYPher reports haplotype-level star allele calls in a structured JSON format, providing a standardized and machine-readable representation of nine PharmVar/CPIC-prioritized pharmacogenes. This format encodes PharmVar-defined star alleles together with copy number information and annotations of putative novel structural variants, thereby supporting downstream analysis, integration into pharmacogenomic pipelines, and clinical interpretation. Further details regarding the output structure and visualization are provided in **Supplementary Note 1**.

Data access

All data generated or analyzed during this study are included in this article, its Supplemental files, and Zenodo (https://doi.org/10.5281/zenodo.8094803, https://zenodo.org/records/14854401). deCYPher source code is available on GitHub (https://github.com/HolidayChang/deCYPher).

Competing interest statement

The authors declare no competing interests.

Acknowledgment

This work was supported in part by computational and storage resources provided by the National Center for High-performance Computing (NCHC), National Applied Research Laboratories (NARLabs), Taiwan. Funding informaiton

Author contributions

J.S.H., P-L.C., Y-S.L., and T-K.H. proposed the initial idea of annotating pharmacogene alleles using high-quality assemblies, initiated the project, and provided input during discussions.

T-Y.C. conceived the study, designed the analysis pipeline, acquired the HPRC data, performed the analyses, interpreted the results, and generated the figures and tables. T-Y.C. also managed the writing and organization of the manuscript, with suggestions and proofreading from J.S.H.

J.S.H. supervised the overall project.

All authors read and approved the final manuscript.

Supplemental material

Supplemental File1: decepher.xlsx

361 **Supplemental Table 1.** Predicted phenotypes for pharmacogenes in the HPRC Year-1 phased
362 assemblies as determined by **deCYPher**.

HPRC Year1	Ultrarapid Metabolizer, Rapid Metabolizer	Normal Metabolizer, Normal Function	Intermediate Metabolizer, Likely Intermediate Metabolizer, Possible Intermediate Metabolizer, Decreased Function, Possible Decreased Function	Poor Metabolizer, Poor Function	Indeterminate
CYP2B6	0/47(0%)	15/47(31.91%)	24/47(51.06%)	3/47(6.38%)	5/47(10.64%)
CYP2C9	0/47(0%)	38/47(33.3%)	12/47(66.7%)	0/47(0%)	0/47(0%)
CYP2C19	14/47(29.79%)	20/47(42.55%)	12/47(25.53%)	1/47(2.13%)	0/47(0%)
CYP2D6	1/47(2.13%)	24/47(51.06%)	17/47(36.17%)	1/47(2.13%)	4/47(8.51%)
CYP3A5	0/47(0%)	6/47(12.77%)	11/47(0%)	30/47(100%)	0/47(0%)
CYP4F2	0/47(0%)	25/47(53.19%)	5/47(10.64%)	0/47(0%)	17/47(36.17%)
DPYD	0/47(93.62%)	44/47(93.62%)	0/47(0%)	0/47(0%)	3/47(6.38%)
SLCO1B1	0/47(0%)	35/47(74.47%)	5/47(10.64%)	0/47(0%)	7/47(14.89%)
NUDT15	0/47(0%)	44/47(93.62%)	2/47(4.26%)	0/47(0%)	1/47(2.13%)

363

364 **Supplemental Table 2.** Predicted phenotypes for pharmacogenes in the HPRC Year-2 phased
365 assemblies as determined by **deCYPher**.

HPRC Year2	Ultrarapid Metabolizer, Rapid Metabolizer	Normal Metabolizer, Normal Function	Intermediate Metabolizer, Likely Intermediate Metabolizer, Possible Intermediate Metabolizer, Decreased Function, Possible Decreased Function	Poor Metabolizer, Poor Function	Indeterminate
CYP2B6	10/232(4.31%)	83/232(35.78%)	96/232(41.38%)	31/232(13.36%)	12/232(5.17%)
CYP2C9	0/232(0%)	189/232(81.47%)	40/232(17.24%)	3/232(1.29%)	0/232(0%)
CYP2C19	54/232(23.28%)	89/232(38.36%)	73/232(31.47%)	14/232(6.03%)	2/232(0.86%)
CYP2D6	4/232(1.72%)	121/232(52.16%)	72/232(31.03%)	3/232(1.29%)	32/232(13.79%)
CYP3A5	0/232(0%)	27/232(11.64%)	87/232(37.5%)	118/232(50.86%)	0/232(0%)
CYP4F2	0/232(0%)	108/232(46.55%)	34/232(14.66%)	8/232(3.45%)	82/232(35.34%)
DPYD	0/232(0%)	227/232(97.84%)	4/232(1.72%)	0/232(0%)	1/232(0.43%)
SLCO1B1	0/232(0%)	170/232(73.28%)	40/232(17.24%)	2/232(0.86%)	20/232(8.62%)
NUDT15	0/232(0%)	218/232(93.97%)	11/232(4.74%)	0/232(0%)	3/232(1.29%)

366

367 **Supplemental Table 3,4.** Predicted structural variants for pharmacogenes in the HPRC Year-
368 1 and Year-2 phased assemblies as determined by **deCYPher**.

HPRC Year1	CYP2B6	CYP2C9	CYP2C19	CYP2D6	CYP3A5	CYP4F2	DPYD	SLCO1B1	NUDT15
CNV	0/94(0%)	0/94(0%)	0/94(0%)	6/94(6.38%)	0/94(0%)	0/94(0%)	0/94(0%)	0/94(0%)	0/94(0%)
Full Gene Deletion	0/94(0%)	0/94(0%)	0/94(0%)	6/94(6.38%)	0/94(0%)	0/94(0%)	0/94(0%)	0/94(0%)	0/94(0%)
Hybrid	0/94(0%)	0/94(0%)	0/94(0%)	4/94(4.26%)	0/94(0%)	0/94(0%)	0/94(0%)	0/94(0%)	0/94(0%)
SNV	94/94(100%)	94/94(100%)	94/94(100%)	82/94(87.23%)	94/94(100%)	94/94(100%)	94/94(100%)	94/94(100%)	94/94(100%)

HPRC Year2	CYP2B6	CYP2C9	CYP2C19	CYP2D6	CYP3A5	CYP4F2	DPYD	SLCO1B1	NUDT15
CNV	0/464(0%)	0/464(0%)	0/464(0%)	65/464(14%)	0/464(0%)	1/464(0.22%)	0/464(0%)	0/464(0%)	0/464(0%)
Full Gene Deletion	0/464(0%)	0/464(0%)	0/464(0%)	22/464(4.74%)	1/464(0.22%)	1/464(0.22%)	0/464(0%)	0/464(0%)	0/464(0%)
Hybrid	1/464(0.22%)	0/464(0%)	0/464(0%)	50/464(10.78%)	0/464(0%)	0/464(0%)	0/464(0%)	0/464(0%)	0/464(0%)
SNV	463/464(99.78%)	464/464(100%)	464/464(100%)	375/464(80.82%)	463/464(99.78%)	462/464(99.57%)	464/464(100%)	464/464(100%)	464/464(100%)

369

370 **Reference**

371 Chen, X., Shen, F., Gonzaludo, N., Malhotra, A., Rogert, C., Taft, R. J.,

Bentley, D. R., & Eberle, M. A. (2021). Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing data. *Pharmacogenomics J*, *21*(2), 251–261. https://doi.org/10.1038/s41397-020-00205-5

Cheng, H., Concepcion, G. T., Feng, X., Zhang, H., & Li, H. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Del Tredici, A. L., Malhotra, A., Dedek, M., Espin, F., Roach, D., Zhu, G. D., Voland, J., & Moreno, T. A. (2018). Frequency of CYP2D6 Alleles Including Structural Variants in the United States. *Front Pharmacol*, *9*, 305. https://doi.org/10.3389/fphar.2018.00305

Gaedigk, A., Ingelman-Sundberg, M., Miller, N. A., Leeder, J. S., Whirl-Carrillo, M., & Klein, T. E. (2018). The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clin Pharmacol Ther*, *103*(3), 399–401. https://doi.org/10.1002/cpt.910

Gaedigk, A., Sangkuhl, K., Whirl-Carrillo, M., Klein, T., & Leeder, J. S. (2017). Prediction of CYP2D6 phenotype from genotype across world populations. *Genet Med*, *19*(1), 69–76. https://doi.org/10.1038/gim.2016.80

Hari, A., Zhou, Q., Gonzaludo, N., Harting, J., Scott, S. A., Qin, X., Scherer, S., Sahinalp, S. C., & Numanagić, I. (2023). An efficient genotyper and star-allele caller for pharmacogenomics. *Genome Res*, *33*(1), 61–70. https://doi.org/10.1101/gr.277075.122

Hicks, J. K., Bishop, J. R., Sangkuhl, K., Müller, D. J., Ji, Y., Leckband, S. G., Leeder, J. S., Graham, R. L., Chiulli, D. L., A, L. L., Skaar, T. C., Scott, S. A., Stingl, J. C., Klein, T. E., Caudle, K. E., & Gaedigk, A. (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. https://doi.org/10.1002/cpt.147

Holt, J. M., Harting, J., Chen, X., Baker, D., Saunders, C. T., Kronenberg, Z., Gonzaludo, N., Yoo, B., Hudjashov, G., Jõeloo, M., Lawlor, J. M. J., Lim, W. K., Jamuar, S. S., Cooper, G. M., Milani, L., Pastinen, T., & Eberle, M. A. (2024). StarPhase: Comprehensive Phase-Aware Pharmacogenomic Diplotyper for Long-Read Sequencing Data. *bioRxiv*. https://doi.org/10.1101/2024.12.10.627527

Hung, T. K., Liu, W. C., Lai, S. K., Chuang, H. W., Lee, Y. C., Lin, H. Y., Hsu, C. L., Chen, C. Y., Yang, Y. C., Hsu, J. S., & Chen, P. L. (2024).

410 Genetic complexity of killer-cell immunoglobulin-like receptor genes in
411 human pangenome assemblies. *Genome Res*, *34*(8), 1211–1223.
412 https://doi.org/10.1101/gr.278358.123
413 Lee, S. B., Wheeler, M. M., Thummel, K. E., & Nickerson, D. A. (2019).
414 Calling Star Alleles With Stargazer in 28 Pharmacogenes With Whole
415 Genome Sequences. *Clin Pharmacol Ther*, *106*(6), 1328–1337.
416 https://doi.org/10.1002/cpt.1552
417 Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences.
418 *Bioinformatics*, *34*(18), 3094–3100.
419 https://doi.org/10.1093/bioinformatics/bty191
420 Liao, S., Wang, L., & Wei, X. (2023). Pharmacogenetics and
421 pharmacogenomics in glaucoma therapeutics: the way to personalized therapy.
422 *Chin Med J (Engl)*, *136*(21), 2573–2575.
423 https://doi.org/10.1097/cm9.0000000000002419
424 Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V.,
425 Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A.,
426 Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M.,
427 Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y.,…Phillippy,
428 A. M. (2022). The complete sequence of a human genome. *Science*,
429 *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987
430 Qiao, W., Yang, Y., Sebra, R., Mendiratta, G., Gaedigk, A., Desnick, R.
431 J., & Scott, S. A. (2016). Long-Read Single Molecule Real-Time Full Gene
432 Sequencing of Cytochrome P450-2D6. *Hum Mutat*, *37*(3), 315–323.
433 https://doi.org/10.1002/humu.22936
434 Relling, M. V., & Klein, T. E. (2011). CPIC: Clinical Pharmacogenetics
435 Implementation Consortium of the Pharmacogenomics Research Network.
436 *Clin Pharmacol Ther*, *89*(3), 464–467. https://doi.org/10.1038/clpt.2010.279
437 Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander,
438 E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat*
439 *Biotechnol*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754
440 Taylor, C., Crosby, I., Yip, V., Maguire, P., Pirmohamed, M., & Turner,
441 R. M. (2020). A Review of the Important Role of CYP2D6 in
442 Pharmacogenomics. *Genes (Basel)*, *11*(11).
443 https://doi.org/10.3390/genes11111295
444 Twesigomwe, D., Wright, G. E. B., Drögemöller, B. I., da Rocha, J.,
445 Lombard, Z., & Hazelhurst, S. (2020). A systematic comparison of
446 pharmacogene star allele calling bioinformatics algorithms: a focus on
447 CYP2D6 genotyping. *NPJ Genom Med*, *5*, 30. https://doi.org/10.1038/s41525-

448 020-0135-2
449 Whirl-Carrillo, M., Huddart, R., Gong, L., Sangkuhl, K., Thorn, C. F.,
450 Whaley, R., & Klein, T. E. (2021). An Evidence-Based Framework for
451 Evaluating Pharmacogenomics Knowledge for Personalized Medicine. *Clin*
452 *Pharmacol Ther*, *110*(3), 563−572. https://doi.org/10.1002/cpt.2350
453 Zanger, U. M., & Schwab, M. (2013). Cytochrome P450 enzymes in drug
454 metabolism: regulation of gene expression, enzyme activities, and impact of
455 genetic variation. *Pharmacol Ther*, *138*(1), 103–141.
456 https://doi.org/10.1016/j.pharmthera.2012.12.007
457 Zhou, Y., Ingelman-Sundberg, M., & Lauschke, V. M. (2017).
458 Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of
459 Population-scale Sequencing Projects. *Clin Pharmacol Ther*, *102*(4), 688–700.
460 https://doi.org/10.1002/cpt.690
461 Zhou, Y., Song, L., & Li, H. (2024). Full resolution HLA and KIR genes
462 annotation for human genome assemblies. *bioRxiv*.
463 https://doi.org/10.1101/2024.01.20.576452
464