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 (PharmGkb, 2023; Relling & Klein, 2011). PharmVar currently
60 catalogues pharmacogenes such as *CYP1A2, CYP2A6, CYP2A13, CYP2B6, CYP2C8, CYP2C9,*
61 *CYP2C19, CYP2D6, CYP3A4, CYP3A5, CYP4F2, DPYD, NUDT15, NAT2*, and *SLCO1B1*,
62 offering standardized allele definitions across major drug-metabolizing enzymes and
63 transporters. Concurrently, CPIC has published clinical guidelines for an expanded array of
64 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.

Among CYP and related pharmacogenes, loci such as *CYP2D6*, *CYP2B6, CYP2C9, CYP2C19*, and others are highly polymorphic. They are well-established predictors of clinically actionable metabolic phenotypes (for example, poor, intermediate, normal, rapid, or ultrarapid metabolizers) used in therapeutic decision-making. These genotype–phenotype translations underpin numerous Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines and PharmGKB clinical annotations that convert molecular test results into prescribing recommendations. (Hicks et al., 2015; PharmGkb, 2023; Relling & Klein, 2011).

Accurate star-allele assignment for these genes is complicated by multiple genetic variation types. In addition to single-nucleotide variants (SNVs) and small indels, clinically important structural variation occurs frequently — including whole-gene deletions and duplications, hybrid alleles created by gene–pseudogene conversions, and complex rearrangements — all of which can have large effects on enzyme activity. *CYP2D6* serves as a classic example; it is located in a region with related pseudogenes such as *CYP2D7* and features numerous structural alterations. Approximately 5-10% of people have copy-number variants like gene deletions or duplications, while *CYP2D6/CYP2D7* hybrid alleles are found in about 2-6%. Altogether, these variations make up more than 10% of clinically significant differences in *CYP2D6* (Del Tredici et al., 2018; Gaedigk et al., 2017; Qiao et al., 2020).

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

101 Hari et al., 2023; Lee et al., 2019; Twesigomwe et al., 2020), but their effectiveness is limited
102 when complex structural rearrangements or high sequence homology to pseudogenes are
103 involved. Recent benchmarking studies indicate that sequence ensemble approaches combined
104 with specialized algorithms perform well. However, complete concordance among all allele
105 classes—including rare structural variants and hybrids—has not been consistently attained
106 (Chen et al., 2021; Qiao et al., 2020). Integrating multiple sequencing techniques, including
107 PacBio HiFi, ONT ultra-long reads, and Illumina Hi-C reads, to construct haplotype-resolved
108 assemblies has demonstrated the ability to resolve complex loci, detect novel structural variants,
109 and provide contiguous haplotypes that are essential for unambiguous allele definition (Cheng
110 et al., 2021; Liao et al., 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 (Song et al., 2023; Wang et al., 2021). Although pb-StarPhase
114 have advanced allele resolution for complex pharmacogenes, their performance can still be
115 affected by sequencing coverage and read length, which may lead to incomplete or uncertain
116 genotype calls under suboptimal data quality. These limitations highlight the need for a targeted
117 and systematic framework that operates consistently across clinically relevant pharmacogenes,
118 integrates PharmVar allele definitions with CPIC and PharmGKB conventions, and delivers
119 high-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

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

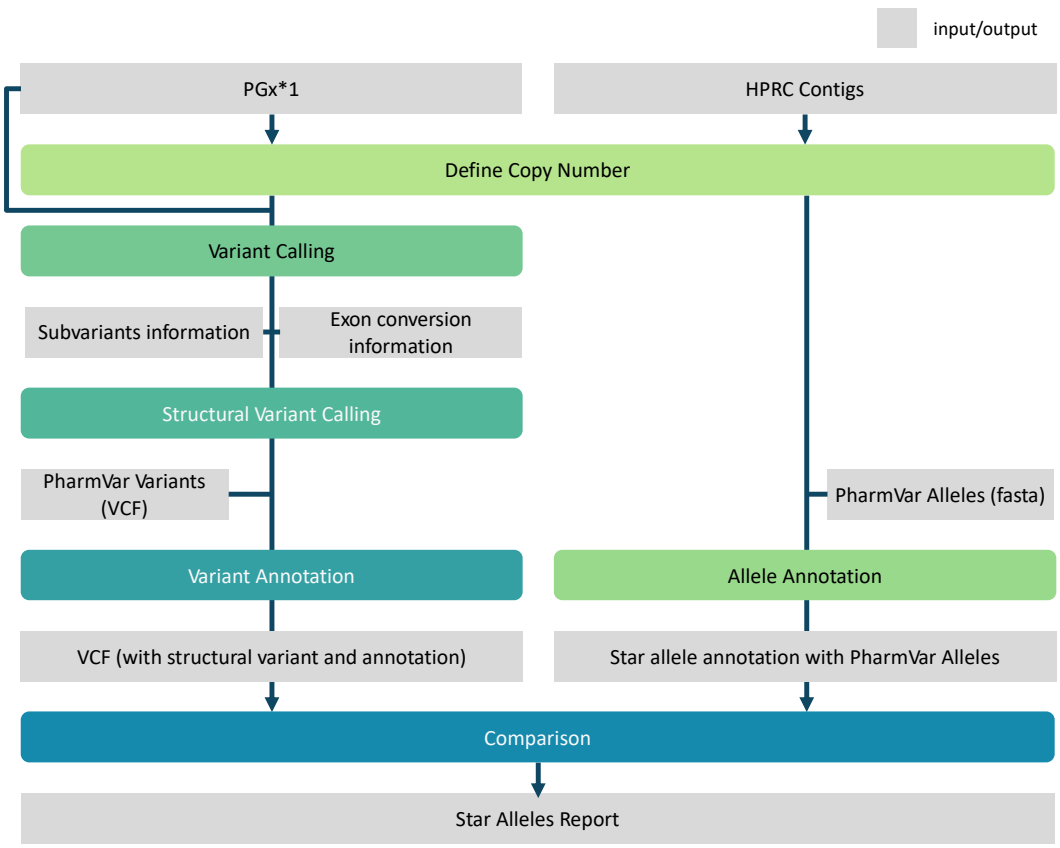


Figure 1. Workflow of the deCYPher pipeline for CYP allele annotation in diploid genome assemblies.

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.

148

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

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

154 185 star alleles, encompassing numerous hybrid and deletion alleles, as well as copy number
155 variants (CNVs). (Taylor et al., 2020). Our results show that, among the 94 CYP2D6 haplotypes
156 in the HPRC Year-1 dataset, 6 haplotypes (6.38%, 6/94) harbor one or more CNVs.
157 Additionally, 6 haplotypes (6.38%, 6/94) were identified as full gene deletions, as illustrated in
158 Figure 2C. Similarly, in the HPRC Year-2 dataset, 66 out of 464 CYP2D6 haplotypes (14.2%,
159 66/464) contained one or more CNVs, while 22 haplotypes (4.74%, 22/464) were classified as
160 full gene deletions, as shown in Figure 2C.

161 Regarding phenotype predictions, 24 samples (51.06%, 24/47) in the HPRC Year-1
162 assemblies were classified as Normal Metabolizers, whereas in the HPRC Year-2 assemblies,
163 121 samples (52.16%, 121/232) fell into this category (detailed data are provided in
164 Supplemental Tables 1 and 2). These findings address the limitations of previous technologies,
165 which could not resolve CYP gene structural variations at base-pair resolution across whole-
166 genome sequences, highlighting the importance of deCYPher for the comprehensive
167 characterization of pharmacogene genetic variation.

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

171

172

173

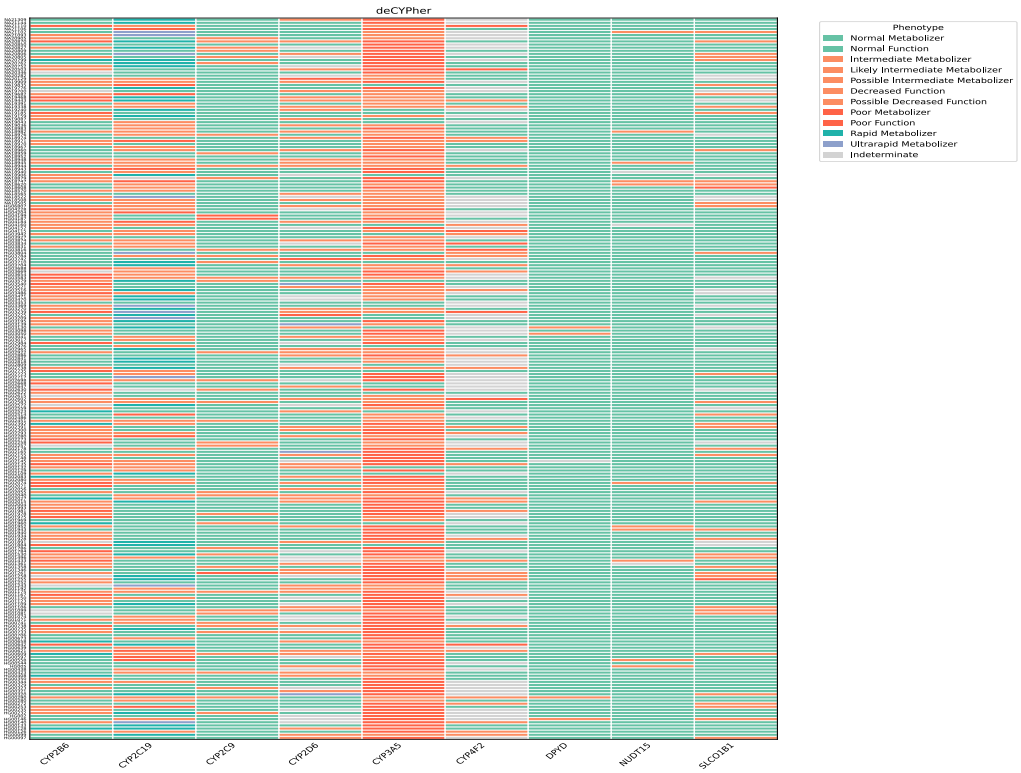
174

175

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
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
CYP2B6
CYP2C19
CYP2C9
CYP2D6
CYP3A5
CYP4F2
DPYD
NUDT15
SLCO1B1

C.

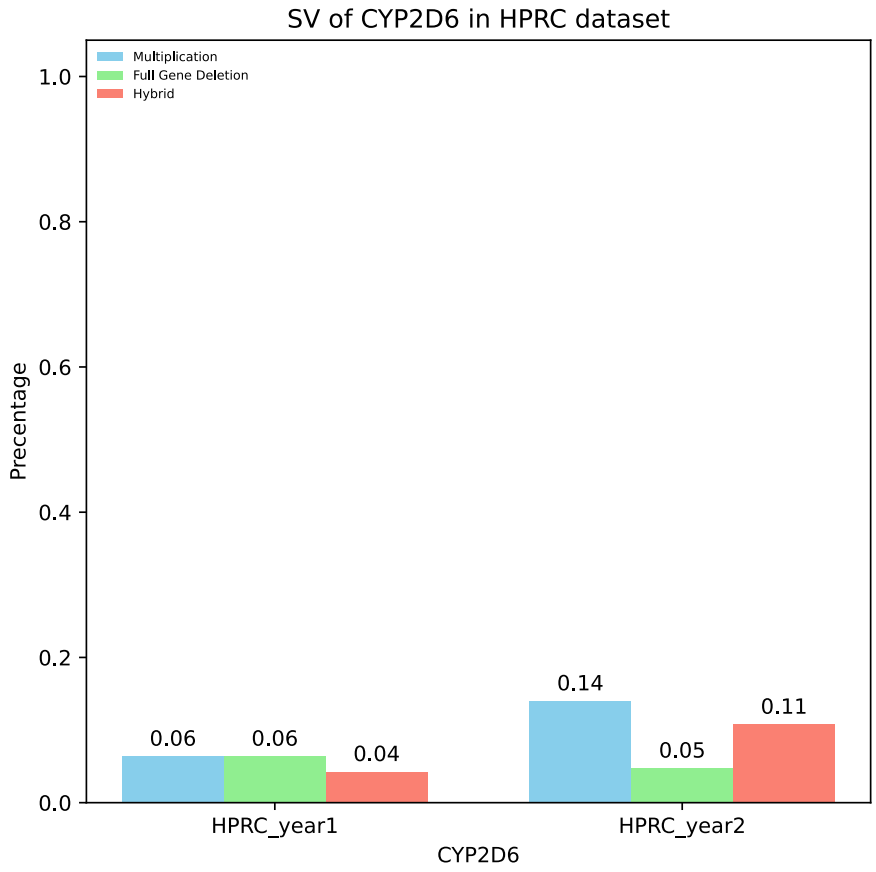

SV of CYP2D6 in HPRC dataset
Multiplication
Full Gene Deletion
Hybrid
Precentage
1.0
0.8
0.6
0.4
0.2
0.0
0.06
0.06
0.04
0.14
0.05
0.11
HPRC_year1
HPRC_year2
CYP2D6

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.

176

177 Within the HPRC assemblies, in addition to *CYP2D6*, we also identified structural
178 variants in other pharmacogenes. For example, in HG01993-m we detected a *CYP2B6* hybrid
179 gene, while in HG00423-m we observed a full-gene deletion of *CYP3A5* — a variant that has
180 never been previously reported. Furthermore, in HG02615-m we identified a CYP4F2 copy
181 number variation (CNV), which is also novel, and in HG02615-p we found a full-gene deletion
182 of CYP4F2. Notably, all of these non-*CYP2D6* structural variants were discovered in the HPRC
183 Year-2 phased assemblies. This suggests that expanding the size of cohorts with PacBio HiFi
184 phased assemblies may enable the discovery of additional low allele frequency and novel
185 structural variants, further underscoring the importance of **deCYPher** for structural variant
186 analysis in pharmacogenes (Supplemental Tables 3 and 4).

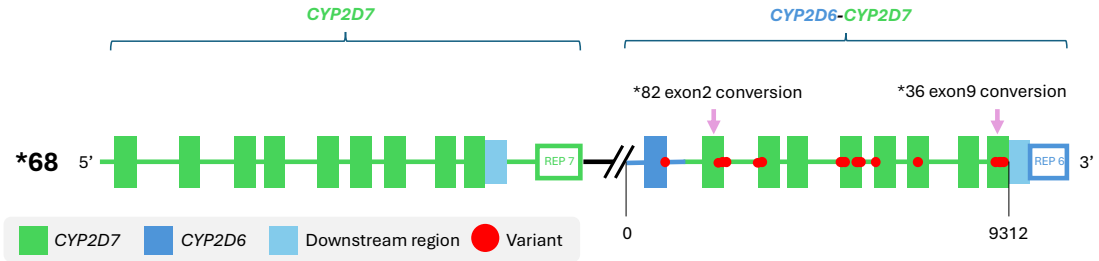
187

188 <u>Distinctive variant positions defining the CYP2D6*68 hybrid allele</u>

189 We identified a *CYP2D6* hybrid allele, CYP2D6 *68, in five phased assemblies from the
190 HPRC dataset: HG00733-p, HG01258-p, HG01784-m, NA20805-p, and NA20827-m.
191 According to PharmVar, *68 is defined by a CYP2D7-to-CYP2D6 gene conversion spanning
192 exons 2 to 9. By contrast, *36 is characterized by a conversion limited to exon 9, and *82 by a
193 conversion restricted to exon 2. Consequently, any haplotype carrying the exon 2–9 conversion

194 of *68 will inherently meet the defining criteria of both *36 (exon 9 conversion) and *82 (exon
195 2 conversion). Based on this relationship, our initial strategy was to classify any assembly
196 carrying both *36 and *82 as *68 (Fig. 3A). Using this approach, we detected three *68 alleles
197 in the HPRC Year-1 phased assemblies—one in HG00733-p and two in HG01258-p—yielding
198 results that were fully concordant with calls from Aldy and Cyrius.

199 Although carrying both *36 and *82 as a *68 criterion successfully identified *68, it may
200 not capture all instances of this hybrid allele, as *68 involves a continuous exon 2–9 conversion.
201 To refine detection, we sought variant positions that could serve as a more complete signature
202 for *68. We first compiled all star alleles present across confirmed *68 assemblies and found
203 consistent co-occurrence of the following alleles: *10, *36, *39, *49, *53, *74, *82, *83, *86,
204 *88 (with one variant site optionally absent), *131, *132, *139, and *176. We then recorded
205 the variant positions corresponding to these alleles, excluding NG_008376.4:6032 T>C, to
206 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.

207

208 Fusion of *CYP2B6*

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

215 Although the genomic characterization of **CYP2B6*29** is already well documented,
216 current genotyping tools such as **Cyrius**, **Aldy**, and **pb-StarPhase** perform suboptimally in
217 detecting this allele. In both the HPRC Year-1 and Year-2 phased assemblies, no samples were
218 classified as CYP2B6*29 by these tools. However, when applying **deCYPher** to the HPRC
219 Year-2 phased assemblies, we identified **HG01993-m** as carrying the CYP2B6*29 allele.

220 deCYPher recognized this allele based on two key observations. First, we detected
221 abnormal mapping patterns when aligning the HG01993-m contigs to the **CYP2B6*1** reference:
222 read mapping with generally poor mapping quality, except for the primary mapping segment
223 spanning **exons 5–9**, where mapping quality was higher, whereas **exons 1–4** were
224 predominantly soft-clipped (Fig. 4A). This mapping profile is consistent with the known
225 structure of CYP2B6*29. Second, when aligning HG01993-m to the *CYP2B7/CYP2B6*
226 breakpoint junction sequence, we observed almost perfect alignment across the junction with
227 no evidence of soft or hard clipping, and only a few single-nucleotide variants (SNVs) (Fig.
228 4B). These Integrative Genomics Viewer(IGV)(Robinson et al., 2011)-based visualizations

229 provide orthogonal evidence that deCYPher correctly identified HG01993-m as CYP2B6*29.

230 Existing algorithms missed this call. This demonstrates deCYPher's unique ability to resolve

231 hybrid alleles in pharmacogenes that are often misclassified or overlooked by short-read–based

232 tools.

A.

B.

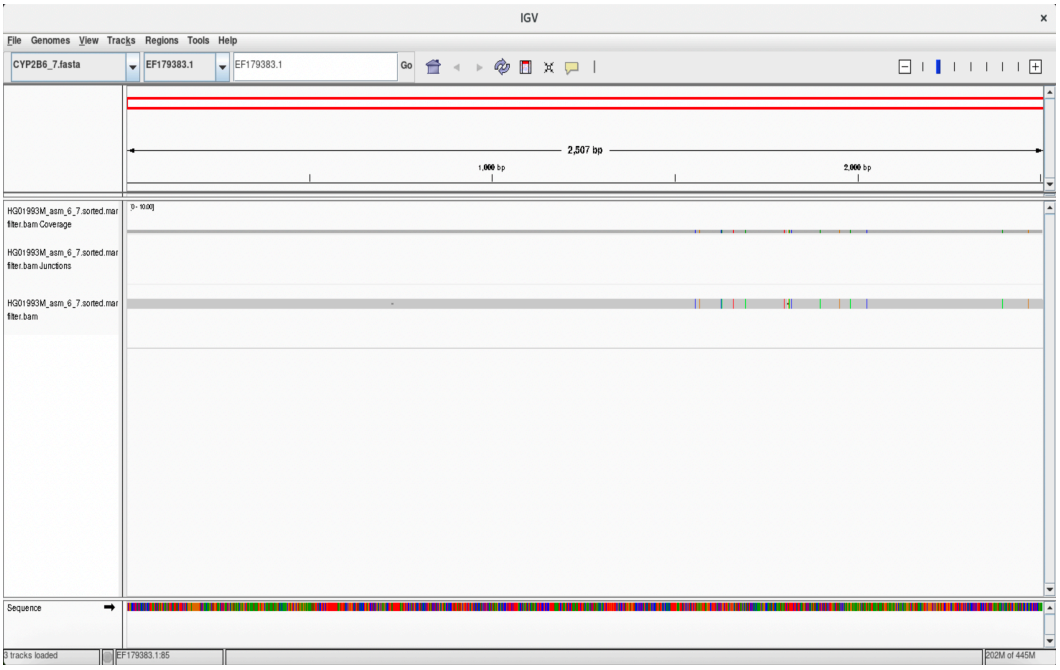


Figure 4. IGV visualization of CYP2B6*29 in HG01993-m.

(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.

233

234 Comparisons with pb-StarPhase in HPRC genomes

235 pb-StarPhase (Wang & et al., 2021) provides an innovative approach for pharmacogene
236 genotyping using PacBio long-read whole-genome sequencing. To directly compare its
237 performance with deCYPher, we applied the recommended pb-StarPhase pipeline to the 47
238 HPRC Year-1 samples, starting from randomly selected long-read raw data. We then generated
239 diplotype calls for all samples and compared them against deCYPher results.

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

249 As an illustrative example, in HG03516, both the pbmm2-aligned bam file and the
250 Deepvariant vcf file contained the variant chr22:42129770 G>A, a defining variant of

251 CYP2D6*17 (Fig. 5). Nevertheless, pb-StarPhase assigned this sample as CYP2D6*2/*2,
252 missing the variant, whereas deCYPher correctly identified the diplotype as CYP2D6*17/*2.

253 Overall, deCYPher demonstrated superior accuracy for *CYP2D6*, achieving a precision of
254 94.68%, compared to 87.23% for pb-StarPhase — the largest discrepancy observed among the
255 nine pharmacogenes analyzed. Notably, both tools produced identical results for *CYP2C9* and
256 *NUDT15* (Table 1).

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

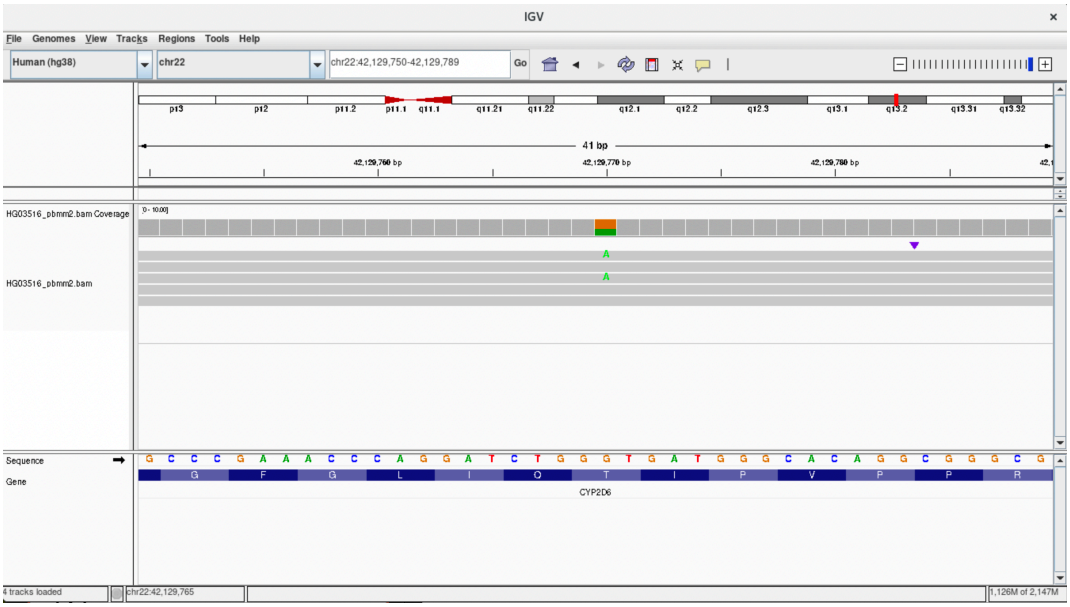


Figure 5. IGV visualization of the CYP2D*17-defining variant in HG03516.

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.

262

263 **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%)

264

265

266 **Discussion**

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

274 Hybrid alleles of *CYP2D6*, such as *36 and *82, have been well characterized in previous
275 studies (Gaedigk et al., 2010; Gaedigk et al., 2018). In line with these reports, we identified
276 CYP2D6*68, a complex hybrid allele defined in PharmVar as a gene conversion event spanning
277 exons 2–9 from *CYP2D7* (Gaedigk et al., 2018). Importantly, our analysis also revealed
278 structural variants outside of *CYP2D6*, including a CYP2B6*29 hybrid in HG01993-m, which,
279 to our knowledge, has not been reported in other tools before. These findings highlight the
280 strength of deCYPher in resolving both known and unclear structural variants across

281 pharmacogenes, particularly when complex scenarios involving CNVs, hybrid genes, or large
282 deletions complicate haplotype assignment.

283 These findings highlight a broader limitation of short-read pipelines. Previous work has
284 shown that genotype-to-phenotype assignments can differ substantially across laboratories and
285 algorithms, especially in the presence of structural variants(Twesigomwe et al., 2020). In such
286 cases, ambiguity often arises not because phenotype rules differ, but because the underlying
287 genotype calls are inconsistent. By integrating haplotype-resolved assemblies with PharmVar
288 allele definitions, deCYPher substantially reduces this discordance and produces more reliable
289 haplotype calls.

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

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

303

304

305 **Methods**

306 deCYPher - Algorithms of the Pharmacogenes annotation Pipeline

307 *Read Mapping and Variant Detection*

308 The deCYPher pipeline was developed to resolve PharmVar-defined star alleles and
309 structural variation, including gene duplications and deletions. To capture multiple
310 polymorphic alleles at each locus, we aligned PharmVar reference allele sequences to
311 haplotype-resolved genome assemblies using **minimap2** (Li, 2018) optimized for PacBio HiFi
312 data. Parameter settings and alignment configurations are provided in **Supplementary Note 2**.

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

318

319 *Output and Visualization*

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

327

328

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.

Reference

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., Faris, J. G., & et al. (2021). Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nature Methods*, *18*(2), 170–175. https://doi.org/10.1038/s41592-020-01056-5

Del Tredici, A. L., Malhotra, A., Dedek, M., & et al. (2018). Frequency of CYP2D6 alleles including structural variants in the United States. *Frontiers in Pharmacology*, *9*, 305. https://doi.org/10.3389/fphar.2018.00305

Gaedigk, A., Ingelman-Sundberg, M., Miller, N. A., & et al. (2018). The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clinical Pharmacology & Therapeutics*, *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., Sangkuhl, K., Swen, J. J., & et al. (2015). Clinical Pharmacogenetics Implementation Consortium guideline (CPIC) for CYP2D6 and CYP2C19 genotypes and dosing of selective serotonin reuptake inhibitors. *Clinical Pharmacology & Therapeutics*, *98*(2), 127–134. https://doi.org/10.1002/cpt.147

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). Genetic complexity of killer-cell immunoglobulin-like receptor genes in human pangenome assemblies. *Genome Res*, *34*(8), 1211–1223. https://doi.org/10.1101/gr.278358.123

Lee, S. B., Wheeler, M. M., Patterson, K., McGee, S., Dalton, R., Woodahl, E. L., Gaedigk, A., Thummel, K. E., & Nickerson, D. A. (2019). Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genet Med*, *21*(2), 361–372. https://doi.org/10.1038/s41436-018-0054-0

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Liao, W. W., Asri, M., Wang, T., & et al. (2023). A draft human pangenome reference. *Nature*, *617*(7960), 312–324. https://doi.org/10.1038/s41586-023-05896-x

Nurk, S., Koren, S., Rhie, A., & et al. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

PharmGkb. (2023). Gene-Specific Information Tables.

Qiao, W., Yang, Y., Sebra, R., & et al. (2020). Long-read single molecule real-time full gene sequencing of cytochrome P450-2B6. *Pharmacogenetics and Genomics*, *30*(2), 37–46. https://doi.org/10.1097/FPC.0000000000000388

Relling, M. V., & Klein, T. E. (2011). CPIC: Clinical Pharmacogenetics Implementation Consortium of the Pharmacogenomics Research Network. *Clinical Pharmacology & Therapeutics*, *89*(3), 464–467. https://doi.org/10.1038/clpt.2010.279

Robinson, J. T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E. S.,

Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nature Biotechnology*, *29*(1), 24–26. https://doi.org/10.1038/nbt.1754

Song, W., & et al. (2023). Benchmarking pharmacogene genotyping accuracy across sequencing technologies. *Pharmacogenomics Journal*. https://doi.org/10.1038/s41397-023-00310-2

Taylor, C., Crosby, I., Yip, V., Maguire, P., Pirmohamed, M., & Turner, R. M. (2020). A Review of the Important Role of CYP2D6 in Pharmacogenomics. *Genes (Basel)*, *11*(11). https://doi.org/10.3390/genes11111295

Twesigomwe, D., Wright, G. E. B., Drögemöller, B. I., da Rocha, J., Lombard, Z., & Hazelhurst, S. (2020). A systematic comparison of pharmacogene star allele calling bioinformatics algorithms: a focus on CYP2D6 genotyping. *NPJ Genom Med*, *5*, 30. https://doi.org/10.1038/s41525-020-0135-2

Wang, J., & et al. (2021). StarPhase: haplotype-aware variant calling and diplotype assignment from long-read assemblies. *bioRxiv*. https://doi.org/10.1101/2021.12.15.472864

Zanger, U. M., & Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacological Therapeutics*, *138*(1), 103–141. https://doi.org/10.1016/j.pharmthera.2012.12.007

Zhou, Y., Ingelman-Sundberg, M., & Lauschke, V. M. (2017). Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of Population-scale Sequencing Projects. *Clin Pharmacol Ther*, *102*(4), 688–700. https://doi.org/10.1002/cpt.690

Zhou, Y., Song, L., & Li, H. (2024). Full resolution HLA and KIR genes annotation for human genome assemblies. *bioRxiv*. https://doi.org/10.1101/2024.01.20.576452