Calling Star Alleles With Stargazer in 28 Pharmacogenes With Whole Genome Sequences

Seung-been Lee[1], Marsha M. Wheeler[1], Kenneth E. Thummel[2,3] and Deborah A. Nickerson[1,3,*]

Variation in the enzymatic activity of pharmacogenes is defined by star alleles (haplotypes) comprised of single-nucleotide variants, small insertion-deletions, and large structural variants. We recently developed Stargazer, a next-generation sequencing-based tool to call star alleles for the clinically important *CYP2D6* gene. Here, we present the utility of extending Stargazer to call star alleles for 28 pharmacogenes using whole genome sequencing (WGS) data. We applied Stargazer to WGS data from 70 ethnically diverse samples from the Genetic Testing Reference Materials Coordination Program (GeT-RM). These reference samples were extensively characterized by GeT-RM using multiple pharmacogenetic testing assays. In all 28 genes, Stargazer recalled 100% of star alleles (*N* = 92) present in GeT-RM's consensus genotypes (*N* = 1,559). Stargazer also detected star alleles not previously reported by GeT-RM, including complex structural variants. Our results demonstrate that combining WGS data and Stargazer enables automated, accurate, and comprehensive genotyping of pharmacogenes in the human genome.

Study Highlights

WHAT IS THE CURRENT KNOWLEDGE ON THE TOPIC?

☑ As the cost of next-generation sequencing continues to decrease and as the clinical value of whole genome (or targeted panel) sequencing expands, there is an increasing demand for tools to automatically and accurately genotype pharmacogenes from sequence data.

WHAT QUESTION DID THIS STUDY ADDRESS?

☑ Our study compares conventionally obtained genotypes from Genetic Testing Reference Materials Coordination Program with those obtained by interpreting whole genome sequence (WGS) data using an expanded bioinformatics tool known as Stargazer. We also show the utility of WGS and Stargazer to identify additional genetic variants not identified by existing assays.

WHAT DOES THIS STUDY ADD TO OUR KNOWLEDGE?

☑ This study shows that WGS data coupled with Stargazer can provide not only accurate but also a comprehensive platform for pharmacogenetic testing compared with multiple standard approaches.

HOW MIGHT THIS CHANGE CLINICAL PHARMACOLOGY OR TRANSLATIONAL SCIENCE?

☑ By allowing automated, accurate, and comprehensive genotyping of pharmacogenes, the combination between whole genome (or targeted) sequencing and Stargazer offers a feasible path for broad implementation of PGx testing and the optimization of individual drug treatment responses.

Genetic variation contributes significantly to the wide interindividual variability in pharmacological responses and gives rise to differences in systemic drug exposure, safety, and efficacy.[1] Not accounting for this genetic variation can lead to severe adverse reactions or a loss of efficacy, due to inappropriate drug choice and/or dosing.[2,3] For example, multiple loss-of-function variants in the *CYP2C9* gene can greatly diminish drug metabolism by blocking enzyme synthesis or reducing its catalytic function.[4,5] Individuals who are homozygous for these variants are called CYP2C9 poor metabolizers and are at risk of abnormal bleeding if prescribed the average dose of the anticoagulant warfarin.[6]

Pharmacogenetic (PGx) testing offers the potential for precision drug therapy, through the combination of genetic information and corresponding drug response phenotypes. Optimal pharmacotherapy can be determined by PGx testing to increase the overall efficacy and prevent adverse drug reactions.[7] The US Food and Drug Administration provides additional guidance by requiring applicable PGx test information be included in the drug labeling.[8]

[1]Department of Genome Sciences, School of Medicine, University of Washington, Seattle, Washington, USA; [2]Department of Pharmaceutics, School of Pharmacy, University of Washington, Seattle, Washington, USA; [3]Brotman Baty Institute for Precision Medicine, Seattle, Washington, USA.
*Correspondence: Deborah A. Nickerson (debnick@uw.edu)

Received March 12, 2019; accepted May 30, 2019. doi:10.1002/cpt.1552

However, to date, broad implementation of PGx testing has met several challenges, and only a few PGx tests are currently routinely used in the clinic.[9]

A major barrier to broad implementation has been the complexity of many pharmacogenes. Several genes require that PGx testing include a large number of genetic variants to provide accurate predictions of enzymatic activity.[10] For example, the clinically important *CYP2D6* gene has >100 star alleles (haplotypes) defined by single-nucleotide variants (SNVs), small insertion-deletions (indels), and/or large structural variants (SVs).[11] These *CYP2D6* alleles encode enzymes with normal, decreased, increased, or no function, which translate to inferred clinical phenotypes that range from ultrarapid to poor metabolism.[12] Importantly, the frequency of star alleles and phenotypes can vary across different populations,[13] highlighting the need for comprehensive variant testing.

Another major challenge has been that a large fraction of existing star alleles cannot be accurately assessed with a single methodology. As stated above, *CYP2D6* alleles include SVs, such as deletions, duplications, and complex gene hybrids. Many of these SVs are difficult to detect due to high sequence homology (>95%) with a nearby nonfunctional paralog.[14] Thus, several orthogonal genotyping methods, including TaqMan assays, long-range polymerase chain reaction (PCR), quantitative multiplex PCR, High Resolution Melt analysis, and Sanger sequencing are required to accurately call all SVs in *CYP2D6*.[15] These methods do reliably detect the star alleles needed for clinical application but can be time-consuming and biased toward the detection of known SVs.

The Centers for Disease Control and Prevention–based Genetic Testing Reference Materials Coordination Program (GeT-RM) has established genomic DNA reference materials to help the genetic testing community obtain characterized reference materials.[16] A GeT-RM collaborative project recently published genotyping results for 137 ethnically diverse Coriell DNA samples and 28 pharmacogenes.[17,18] These samples were

Table 1 Star alleles previously reported by GeT-RM and assessed by Stargazer's analysis of whole genome sequencing data

Gene	Reference allele	Star alleles found in consensus GeT-RM genotypes (*N* = 92)	Star alleles only found in nonconsensus GeT-RM genotypes (*N* = 31)
CYP1A1	**1*	**2, *4, *5*	None
CYP1A2	**1A*	**1C, *1F, *1L*	None
CYP2A6	**1*	**2, *4 (del), *9, *17, *20*	**8*
CYP2B6	**1*	**2, *6, *7, *18*	**4, *5, *15, *20, *22, *27*
CYP2C8	**1*	**2, *3, *4*	None
CYP2C9	**1*	**2, *3, *5, *6, *8, *9, *11*	**18*
CYP2C19	**1*	**2, *3, *4, *8, *13, *15, *17*	**6, *27*
CYP2D6	**1*	**2, *2x2 (dup), *4, *5 (del), *6, *9, *10, *14, *15, *17, *29, *35, *41, *xN (dup)*	**21, *36 + *10 (hyb), *40*
CYP2E1	**1*	**7*	**4, *5*
CYP3A4	**1*	**1B, *2, *3, *22*	**15, *16*
CYP3A5	**1*	**3, *6, *7*	None
CYP4F2	**1*	**2, *3*	None
DPYD	**1*	**9*	**4*
GSTM1	**A*	**B, *0 (del)*	None
GSTP1	**A*	**B, *C, *D*	None
GSTT1	**A*	**0 (del)*	**AxN (dup), *B*
NAT1	**4*	**11, *14, *17*	None
NAT2	**4*	**5, *6, *7, *14*	**12, *13*
SLC15A2	**1*	**2*	None
SLC22A2	**1*	**3, *6, *7*	**2, *K432Q*
SLCO1B1	**1A*	**1B, *5, *14, *15, *17, *21*	None
SLCO2B1	**1*	None	**S464F*
TPMT	**1*	**3C, *8*	None
UGT1A1	**1*	**6, *28, *60*	**7, *27, *36, *37*
UGT2B7	**1*	**2*	**3*
UGT2B15	**1*	**2, *5*	**4*
UGT2B17	**1*	**2 (del)*	None
VKORC1	**1*	**2, *3, *4*	None

Structural variant–defined alleles are indicated by "del" (deletion), "dup" (duplication), and "hyb" (hybrid).
GeT-RM, Genetic Testing Reference Materials Coordination Program.

genotyped using several commercial and laboratory-developed PGx testing assays.[17] More recently, GeT-RM has made whole genome sequencing (WGS) data for 70 of the 137 reference samples publicly available.[18]

In this study, we utilized the genotyping results from GeT-RM and the available WGS data to continue the development of a new next-generation sequencing-based tool. We extended the SV-aware algorithm of Stargazer[19] to assess star alleles in 28 pharmacogenes (**Table 1**). Among these genes, *CYP2A6*, *GSTM1*, *GSTT1*, and *UGT2B17* are known to display extensive gene-deletion polymorphisms.[20] Additionally, *CYP2A6*, *CYP2B6*, and *CYP2D6* have been shown to frequently exhibit complex SVs, which include gene hybrids with their paralogs *CYP2A7*, *CYP2B7*, and *CYP2D7*, respectively.[21–23] To evaluate the accuracy of this algorithm, we compared star alleles detected by Stargazer to those previously reported by GeT-RM. In addition, we provide an in-depth characterization of the WGS data of these 70 reference samples, which includes the identification of star alleles not tested in previous genotyping efforts. In order to verify Stargazer's SV calls, we explored the Database of Genomic Variants (DGV)[24] for variant reports submitted by various studies, including the 1000 Genomes Project (1KGP).[25]

RESULTS

Evaluating Stargazer's genotyping accuracy

We applied Stargazer to assess 1,960 genotypes in 28 pharmacogenes in 70 WGS samples from GeT-RM. To estimate the accuracy of Stargazer, we compared these genotypes with those previously published by GeT-RM.[17] For these samples, GeT-RM reported a total of 1,559 consensus genotypes comprised of 92 star alleles (**Table 1**). These consensus genotypes were verified by two or more PGx testing assays.[17] In all 28 genes, Stargazer recalled 100% of star alleles present in GeT-RM's consensus genotypes (**Table S1**).

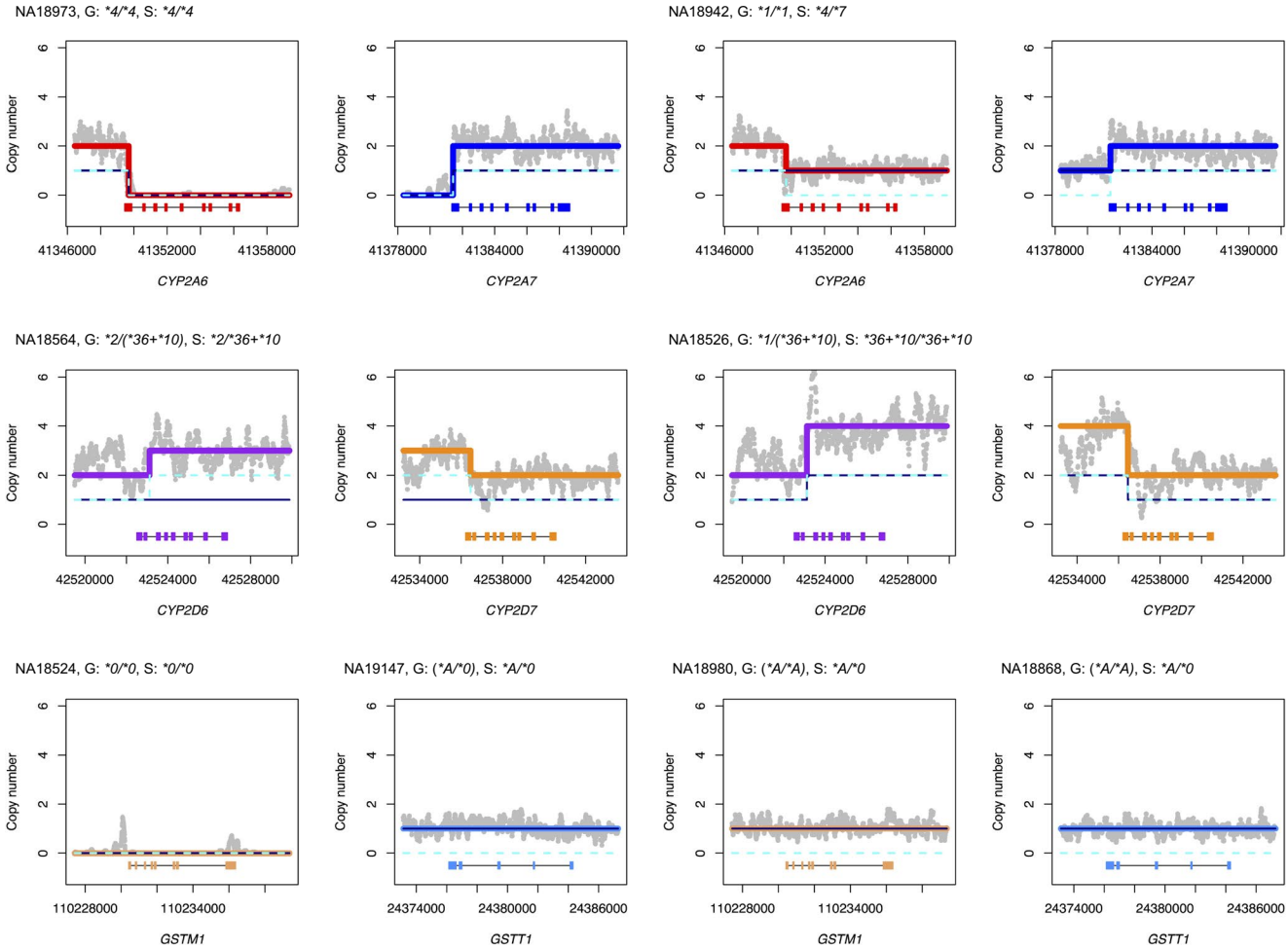

Figure 1 Examples of star alleles with structural variation previously undercalled by Genetic Testing Reference Materials Coordination Program (GeT-RM). Panels display Stargazer's result for copy number analysis for individual samples ($N = 8$). Genotypes from GeT-RM and Stargazer (abbreviated as "G" and "S" for brevity) are also shown, with "()" indicating nonconsensus genotypes. The left and right panels exhibit samples whose structural variant calls are matched and not matched, respectively. Gray dots in each panel indicate the sample's copy number estimates computed from read depth. The navy solid line and the cyan dashed line represent copy number profiles for each haplotype. Thick colored lines represent copy number profiles for different genes for both haplotypes combined. Each panel contains gene names and scaled gene models, in which exons and introns are depicted with colored boxes and black lines, respectively. Reports in the Database of Genomic Variants supported Stargazer's gene deletion calls in NA18942, NA18980, and NA18868.

Table 2 Star alleles with structural variation previously undercalled by GeT-RM

Star allele	Assays[a]	*N* of heterozygotes from GeT-RM	*N* of homozygotes from GeT-RM	*N* of heterozygotes from Stargazer	*N* of homozygotes from Stargazer
*CYP2A6*4*	[1, 2]	4	2	7	2
*CYP2D6*36 + *10*	[2, 3]	7	0	4	4
*GSTM1*0*	[1, 2]	0	32	21	32
*GSTT1*0*	[1, 2]	16	18	36	18

GeT-RM, Genetic Testing Reference Materials Coordination Program.
[a][1] Affymetrix DMET Plus Array (Affymetrix, Santa Clara, CA); [2] Agena Bioscience iPLEX ADME PGx Pro Panel (Agena Bioscience, San Diego, CA); [3] Agena Bioscience iPLEX ADME CYP2D6 Panel (Agena Bioscience, San Diego, CA).

WGS confirmation of star alleles present in "non-consensus" GeT-RM genotypes

A subset of GeT-RM genotypes ($N = 401$) could not be verified by multiple PGx testing assays (**Table S1**). This is either because certain star alleles were tested by a single method or multiple test results disagreed.[17] A total of 31 star alleles were only found in these "nonconsensus" genotypes (**Table 1**). Stargazer's output confirmed the presence of most of these star alleles with the exception of four alleles: *CYP2A6*8*, *CYP2B6*27*, *CYP2C9*18*, and *GSTT1*AxN*. These four alleles were not present in the WGS data (**Table S2**). For example, eight samples were predicted to contain a *GSTT1* duplication (*GSTT1*AxN*) using the Agena Bioscience iPLEX ADME PGx Pro Panel (Agena Bioscience, San Diego, CA). However, Stargazer's output showed that three of these samples had a normal copy number of two, and the remaining five samples had a deletion (*GSTT1*0*) instead (**Figure S1**). To provide validation for these deletion events, we searched DGV and found that four of these five samples were also previously shown by 1KGP to have copy number loss in *GSTT1* (**Table S3**).

SV-defined alleles previously undercalled by GeT-RM

Stargazer's output showed that three gene deletions (*CYP2A6*4*, *GSTM1*0*, and *GSTT1*0*) and one *CYP2D6/CYP2D7* hybrid (*CYP2D6*36+*10*) were previously under-reported by GeT-RM (**Figure 1 and Table 2**). For example, GeT-RM tested gene deletions in *GSTM1* using the Affymetrix DMET Plus Array (Affymetrix, Santa Clara, CA) and the Agena Bioscience iPLEX ADME PGx Pro Panel. Both assays identified 32 samples with homozygous deletions but found no samples with heterozygous deletions. In contrast, Stargazer detected both heterozygous and homozygous deletions in 21 and 32 samples, respectively. By cross-referencing to DGV reports from 1KGP, we validated copy number loss in *GSTM1* for 13 of the 21 samples with heterozygous deletions (**Table S3**). Similarly, we used 1KGP's DGV reports to verify Stargazer's genotype calls for samples with *CYP2A6*4* ($N = 2$) and *GSTT1*0* ($N = 13$; **Table S3**).

WGS identification of additional star alleles not previously reported by GeT-RM

Using the WGS data, Stargazer detected 38 additional star alleles not previously reported by GeT-RM (**Table 3**). These alleles were found in 127 of 1,960 genotypes assessed (**Table S1**). Seven of these alleles contained SVs and were comprised of five gene duplications, one *CYP2B6/CYP2B7* hybrid, and one *CYP2D6/CYP2D7* hybrid (**Figure 2**). Of the five duplications, only *CYP2A6*1x2*, *CYP2D6*2x2*, and *CYP2D6*4x2* are currently listed in existing PGx databases, suggesting that the remaining two, *CYP2D6*34x2* and *CYP2E1*7x2*, may be novel. *CYP2D6*34x2* was identified from a single African sample (NA19207), and *CYP2E1*7x2* was identified from three African samples (NA19095, NA19226, and NA19908); note that NA19908 was homozygous for *CYP2E1*7x2* (**Figure 2**). DGV reports support the presence of *CYP2A6*1x2* in NA18861 (DGV gold standard; copy number gain observed by multiple studies) and the presence of *CYP2E1*7x2* in NA19908 (copy number gain observed by 1KGP; **Table S3**).

Surprisingly, Stargazer detected three gene copies for the *CYP3A4*, *CYP3A5*, *UGT2B7*, and *UGT2B15* genes in a single sample (NA18540) (**Figure S2**). *CYP3A4* and *CYP3A5* are 77 kbp apart on chromosome 7, whereas *UGT2B7* and *UGT2B15* are 426 kbp apart on chromosome 4. GeT-RM did not report any SVs in these genes for this sample or other samples tested.[17] Copy number analyses using Stargazer showed no breakpoints in flanking genomic regions (**Figure S2**),

Table 3 Star alleles identified by Stargazer's analysis of whole genome sequencing and not previously reported by GeT-RM

Gene	Star alleles (*N* = 38)
CYP1A1	**2A, *2B, *13*
CYP2A6	**1x2 (dup), *7, *15, *18, *19, *21, *22, *23, *24, *25, *35*
CYP2B6	**17, *23, *29 (hyb)*
CYP2C19	**35*
CYP2D6	**4x2 (dup), *34, *34x2 (dup), *39, *46, *68 + *4 (hyb)*
CYP2E1	**7x2 (dup)*
DPYD	**5, *6*
GSTM1	**Ax2 (dup)*
NAT1	**3, *10, *26*
SLC22A2	**4*
SLCO1B1	**24, *27, *30, *31, *35*
TPMT	**16*

Structural variant-defined alleles are indicated by "dup" (duplication) and "hyb" (hybrid).
GeT-RM, Genetic Testing Reference Materials Coordination Program.

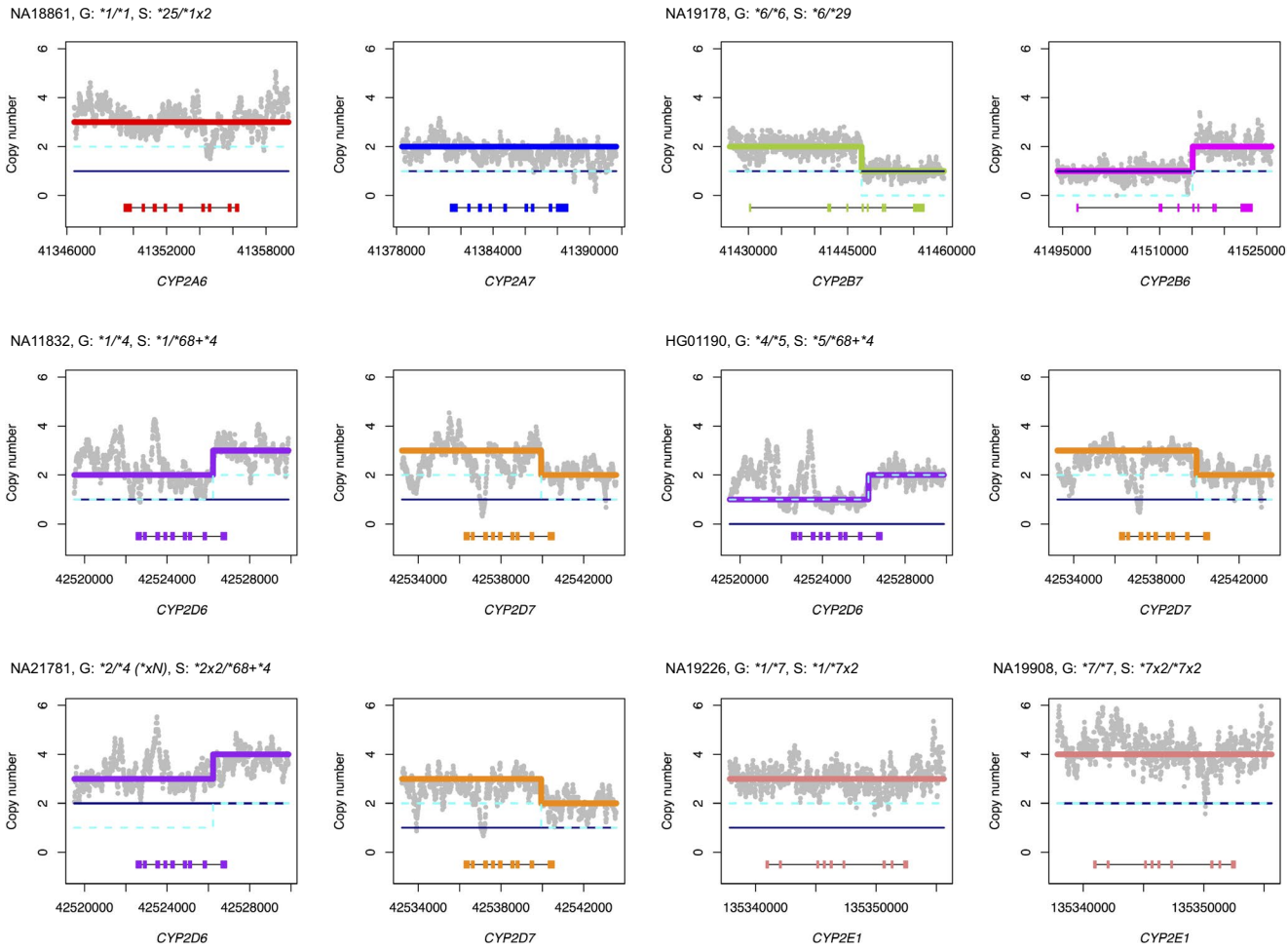

Figure 2 Examples of star alleles with structural variation not previously reported by Genetic Testing Reference Materials Coordination Program (GeT-RM). Panels display Stargazer's results for copy number analysis for individual samples ($N = 7$). Genotypes from GeT-RM and Stargazer (abbreviated as "G" and "S" for brevity) are also shown, with "()" indicating nonconsensus genotypes. Gray dots in each panel indicate the sample's copy number estimates computed from read depth. The navy solid line and the cyan dashed line represent copy number profiles for each haplotype. Thick colored lines represent copy number profiles for different genes for both haplotypes combined. Each panel contains gene names and scaled gene models, in which exons and introns are depicted with colored boxes and black lines, respectively. Stargazer's genotype call for HG01190 involving two different structural variants (a *CYP2D6* deletion and a *CYP2D6/CYP2D7* hybrid) was independently verified previously.[19] Reports in the Database of Genomic Variants supported Stargazer's gene duplication calls in NA18861 and NA11908.

indicating that this sample likely has chromosomal trisomy for chromosomes 4 and 7 (i.e., *CYP3A4*1/*1/*1*, *CYP3A5*1/*1/*3*, *UGT2B7*1/*1/*2*, and *UGT2B15*2/*2/*4*). This result has been independently confirmed through karyotyping by Redon *et al.*,[26] which has additionally revealed trisomy in chromosomes 9, 14, and 21. This aberrant karyotype most likely arose during cell immortalization.

Statistical phasing of SNVs/indels for star alleles

Using statistical phasing[27] with the 1KGP haplotype reference panel,[28] Stargazer revised a total of 64 GeT-RM genotypes (**Table S1**). For instance, both GeT-RM and Stargazer found 4 heterozygous SNVs in 14 samples that were indicative of the *CYP1A2*1A/*1L* or **1C/*1F* genotype. GeT-RM reported both genotypes as equally likely, whereas the phasing algorithm indicated the *CYP1A2*1A/*1L* genotype to be more likely. In addition, Stargazer revised GeT-RM genotypes in two related samples, NA12156 (mother) and NA10831 (child), to follow expected inheritance patterns. As an example, GeT-RM and Stargazer genotyped the mother as *UGT1A1*28/*60* and **1/*28*, **60*, respectively. The mother's correct genotype should be the latter because the child was genotyped as *UGT1A1*28*, **60/*28*, **60* by both GeT-RM and Stargazer.

Resolving ambiguous *CYP2D6* duplications using WGS allelic depth

Both GeT-RM and Stargazer found seven samples with a gene duplication in the *CYP2D6* gene. For four of these samples, GeT-RM reported genotypes containing gene duplications of an unspecified star allele (*CYP2D6*xN*), whereas Stargazer resolved these ambiguous duplications using allelic depth of WGS reads (**Table S1**). For instance, GeT-RM genotyped the sample

Table 4 New star alleles discovered by Stargazer's analysis of whole genome sequencing data

Star allele	Description[a]	*N* of African samples	*N* of East Asian samples	*N* of European samples
*CYP2A6*1 + *S6*	Gene duplication of *CYP2A7* (chr19:41358125-41389907)	0	1	0
*CYP2E1*S1*	Gene duplication in exons 7-9 (chr10:135350465-135439323)	4	0	0
*SLC22A2*S1*	Gene deletion in intron 9 (chr6:160649735-160654861)	3	0	0
*SLC22A2*S2*	Gene deletion affecting 3′-UTR (chr6:160627751-160638068)	2	0	0
*UGT2B15*S1*	Gene deletion affecting exons 4-6 (chr4:69510975-69528283)	0	0	1
*CYP2C9*S1*	Nonsense (chr10:96741125A>T; no rsID)	0	1	0
*SLCO1B1*S1*	Nonsense (chr12:21349909C>T; rs183501729)	0	1	0
*SLCO1B1*S2*	Splice site (chr12:21329832G>T; rs77271279)	2	0	0
*SLCO2B1*S1*	In-frame deletion (chr11:74873754GCACAGAAAA>G; rs72408262)	0	4	1

[a]Genomic coordinates and nucleotide changes are according to Human Genome version 19.

NA19819 as *CYP2D6*2/*4/*xN* because the sample contained three gene copies as well as the *CYP2D6*2* (normal function) and **4* (no function) alleles. In contrast, Stargazer called the sample as *CYP2D6*2/*4x2*, a genotype that was previously independently verified for this sample.[19]

New star alleles defined with WGS findings

Stargazer identified SNVs, indels, and SVs not present in existing haplotype translation tables. These variants represent 9 new star alleles and were found in a total of 20 Stargazer genotypes in a population-specific manner (**Table 4 and Table S1**). More specifically, functional annotation of SNVs/indels added four new star alleles defined by two nonsense variants (*CYP2C9*S1* and *SLCO1B1*S1*), a splice site variant (*SLCO1B1*S2*), and an in-frame deletion variant (*SLCO2B1*S1*). All of these variants have an rsID except for *CYP2C9*S1*. Concordant with our data, 1KGP previously found *SLCO2B1*S1* predominantly in East Asians with allele frequency of 0.105. In addition, 1KGP observed *SLCO1B1*S1* and *SLCO1B1*S2* exclusively in East Asians and Africans, respectively, which is in agreement with our findings. As shown in **Figure 3**, copy number analyses with Stargazer also identified three partial gene deletions (*SLC22A2*S1*, *SLC22A2*S2*, and *UGT2B15*S1*), a partial gene duplication (*CYP2E1*S1*), and a whole gene duplication of *CYP2A7* (*CYP2A6*1+*S6*). To enable automated detection of these SVs, we defined and included five new star alleles as part of the Stargazer algorithm. DGV records confirm the presence of *CYP2E1*S1* in NA19143 (copy number gain observed by Wang *et al.*[29]) and the presence of *SLC22A2*S2* in NA19226 and NA19819 (DGV gold standard; copy number loss observed by multiple studies) (**Table S3**).

DISCUSSION

Here, we present an extension of the Stargazer algorithm to call star alleles in 28 pharmacogenes from next-generation sequencing data. Stargazer is one of the first bioinformatics tools that enable systematic identification of star alleles (e.g., Cypiripi,[30] Astrolabe,[31] PharmCAT,[32] and Aldy[33]). Stargazer is the only tool that uses statistical haplotype phasing,[27] which is informed by population haplotype frequencies to call star alleles more accurately. In addition, other tools tend to have difficulties with the detection of complex SVs, such as *CYP2D6/CYP2D7* hybrids.[30–32] Lastly, to our knowledge, Stargazer is the most comprehensive tool available and assesses star alleles in more genes than has previously been reported.[32,33]

To evaluate the performance of Stargazer, we utilized public WGS data from 70 genotyping reference samples from GeT-RM. These samples were extensively characterized using multiple standard methods (e.g., allele-specific PCR, molecular inversion probes, hybridization-based arrays, and TaqMan assays).[17] To verify Stargazer's SV calls, we explored DGV reports from various sources, including 1KGP. In all 28 genes, Stargazer recalled 100% of star alleles present in GeT-RM's consensus genotypes. Stargazer also identified additional star alleles not previously reported by GeT-RM, including both known and novel SVs, and correctly found trisomies of the chromosomes 4 and 7 in the sample NA18540. Altogether, these results demonstrate that Stargazer has high sensitivity for the detection of SVs and can accurately assess star alleles in these 28 genes.

With statistical phasing, Stargazer revised star alleles in reference samples that previously had ambiguous or incorrect GeT-RM genotypes. In the current version of Stargazer, we incorporated the 1KGP haplotype reference panel to increase sample size and improve phase accuracy.[34] This approach performed well for our dataset, but we are aware that further applications may be challenged by low-frequency variants and limited by the magnitude and extent of linkage disequilibrium.[35] To ameliorate this issue, we plan to merge multiple large, high-quality reference panels to obtain additional haplotype information.

CYP2D6 duplications have been reported for normal function, decreased function, and nonfunctional alleles.[36–38] GeT-RM previously reported ambiguous *CYP2D6* genotypes involving duplication of an unspecified allele (*CYP2D6*xN*). For instance, the sample NA19819 was genotyped as *CYP2D6*2/*4/*xN* by GeT-RM where *CYP2D6*2* and **4* are normal function and nonfunctional alleles, respectively. Therefore, the sample could tentatively be *CYP2D6*2x2/*4* or **2/*4x2*, which would predict two completely different phenotypes—normal metabolizer and

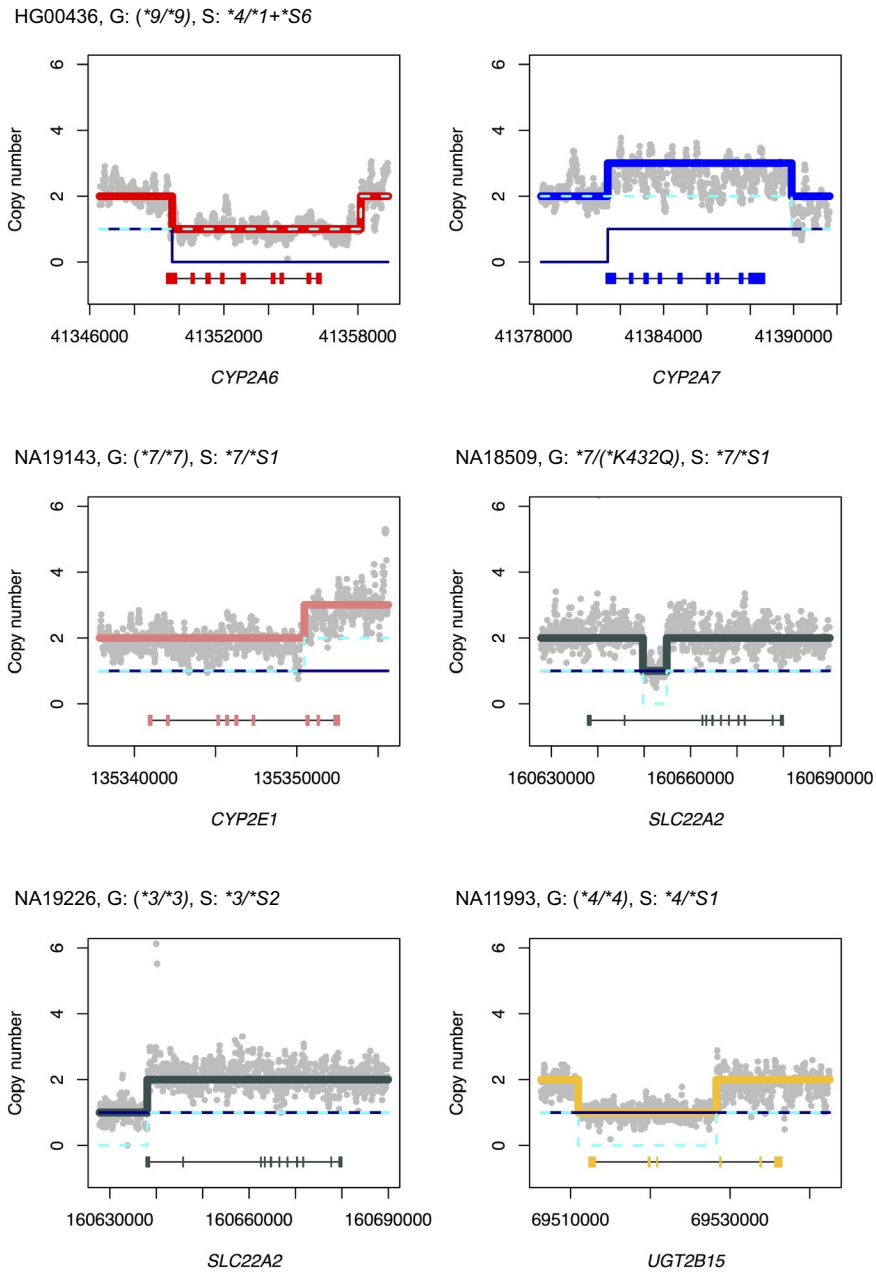

Figure 3 Examples of new star alleles with structural variation. Panels display Stargazer's results for copy number analysis for individual samples (*N* = 5). Genotypes from Genetic Testing Reference Materials Coordination Program (GeT-RM) and Stargazer (abbreviated as "G" and "S" for brevity) are also shown, with "()" indicating nonconsensus genotypes. Gray dots in each panel indicate the sample's copy number estimates computed from read depth. The navy solid line and the cyan dashed line represent copy number profiles for each haplotype. Thick colored lines represent copy number profiles for different genes for both haplotypes combined. Each panel contains gene names and scaled gene models, in which exons and introns are depicted with colored boxes and black lines, respectively. *CYP2A6*1 + *S6* was identified in only one sample (HG00436) exhibiting both a *CYP2A6* deletion (*CYP2A6*4*) and a duplication in the *CYP2A7* paralog (*CYP2A6*1 + *S6*). Reports in the Database of Genomic Variants supported Stargazer's partial duplication call in NA19143 and partial deletion call in NA19226.

intermediate metabolizer, respectively. By using allelic depth of WGS reads, Stargazer correctly called the *CYP2D6*2/*4x2* genotype for the sample[19] and resolved other genotypes with ambiguous *CYP2D6* duplications. Furthermore, both GeT-RM and Stargazer predicted the samples HG00436, NA19109, and NA19226 to be ultrarapid metabolizers (*CYP2D6*1/*2x2*, **29/*2x2*, and **2/*2x2*, respectively), which is a major phenotypic consequence of carrying *CYP2D6* gene duplications. Collectively, these results highlight that allelic decomposition performed by Stargazer enables accurate phenotype prediction for samples with gene duplications.

We report nine new star alleles defined by variants discovered in WGS data. Although the enzymatic activity of these alleles remains to be functionally characterized, seven alleles likely have an impact on enzyme activity. *UGT2B15*S1*, for instance, is likely nonfunctional because it includes deletion of the last three exons of the *UGT2B15* gene. Another new allele, *CYP2A6*1 + *S6*,

contains a gene duplication in the paralog *CYP2A7*. The duplication does not directly affect the *CYP2A6* sequence, but it could still change CYP2A6 activity because CYP2A7 transcript level has been shown to alter CYP2A6 expression via competition for miRNA binding.[39] Conversely, the partial gene deletions in the *SLC22A2*S1* and **S2* alleles do not affect the translated region of the *SLC22A2* gene and, thus, are unlikely to have a functional consequence.

As of April 2019, there are 359 gene/drug pairs (e.g., *CYP2D6*/codeine) described by the Clinical Pharmacogenetics Implementation Consortium with accompanying levels of evidence for changing drug choice and dosing decisions.[40] The assigned levels (A, B, C, and D) are subject to change, and only levels A and B gene/drug pairs ($N = 144$) have sufficient evidence for at least one prescribing action to be recommended.[40] The 28 pharmacogenes currently targeted by Stargazer include 132 of these gene/drug pairs, 67 of which have level A or B. We plan to further extend Stargazer to additional pharmacogenes, including 26 PGx loci whose consensus genotypes could not be determined in Pratt *et al*.[17] because they were only characterized by one laboratory.

In summary, by leveraging WGS data, we confirmed the consensus results reported by GeT-RM and expanded the current PGx variation catalogs for the 70 important reference samples. Therefore, our WGS characterization can be added to this public reference resource for other PGx genotyping projects. As sequencing costs continue to decline and as the clinical value of whole genome (or targeted panel) sequencing continues to emerge, there will be an increasing need for systematic and highly efficient analysis algorithms. Our results show that WGS data combined with Stargazer offers a feasible path for accurate PGx testing and the optimization of individual drug treatment responses.

METHODS

PGx genotypes and WGS data from GeT-RM

We accessed PGx genotypes and WGS data for 70 ethnically diverse Coriell DNA samples from GeT-RM (**Tables S1 and S4**). Both PGx genotypes and WGS data are publically available through the GeT-RM website.[18] GeT-RM generated consensus and nonconsensus genotypes for 28 pharmacogenes using a variety of testing platforms, as detailed in Pratt *et al*.[17] WGS was performed to a depth of >30X using paired-end 150-bp sequence reads on the Illumina HiSeq X (Illumina, San Diego, CA). WGS data were downloaded in the BAM file format, which contains sequence reads aligned to Human Genome version 19 with the program ISAAC.[41]

Extension of Stargazer to 28 pharmacogenes

The first version of Stargazer (version 1.0.0) included a haplotype translation table for >100 star alleles in the *CYP2D6* gene.[19] Extension of Stargazer (version 1.0.4) involved construction of 27 additional haplotype translation tables for > 500 star alleles. Star allele information was compiled from several public PGx databases: the Pharmacogene Variation Consortium,[42] the Pharmacogenomics Knowledgebase,[43] the *UGT* database,[44] the *NAT* database,[45] and the *TPMT* database.[46] Generation of haplotype translation tables involved lifting cDNA coordinates of PGx variants to genomic coordinates from Human Genome version 19. The new version of Stargazer and all haplotype translation tables are available for download: https://stargazer.gs.washington.edu/stargazerweb/.

Description of Stargazer's algorithm

Stargazer's algorithm has been described previously.[19] Briefly, for this extended version, SNVs/indels in each gene were assessed from a Variant Call Format (VCF) file generated from BAM files using GATK-HaplotypeCaller.[47] The VCF file was phased using the program Beagle[27] with the 1KGP haplotype reference panel.[28] Phased SNVs/indels were then matched to star alleles in each gene's haplotype translation table. BAM files were also used to calculate read depth using GATK-DepthOfCoverage.[47] Read depth was converted to copy number by intrasample normalization. Following normalization, SVs were detected by testing pairwise combinations of expected haplotype copy number profiles against the sample's observed copy number profile for both haplotypes. SV results were incorporated to inform the final star allele assignment. Output data of Stargazer included individual genotypes and copy number plots to visually inspect SVs calls (see **Figure 1** for examples of these plots).

Assessment of differences in genotype calls between GeT-RM and Stargazer

Differences in genotype calls between GeT-RM and Stargazer were carefully evaluated by considering the consistency of star allele assignment across individual PGx testing assays as well as assessment of WGS reads. WGS reads were assessed through visual inspection using Integrative Genomics Viewer,[48] as exemplified in **Table S2**. For validation of Stargazer's results involving SVs, we used existing reports available in DGV. A total of 33 Stargazer genotypes were verified this way (26 samples overlapped with DGV; **Table S3**).

Assessment of novel variation by Stargazer

Stargazer's output includes a VCF file of detected SNVs/indels not present in existing PGx databases. This VCF file is functionally annotated using SeattleSeq Annotation.[49] Functional annotation enables prediction of nonsynonymous variants, which may impact enzyme activity. For samples with SVs that do not match expected copy number profiles, Stargazer performs change point analysis.[50] This analysis identifies approximate breakpoints, which are then used to identify subsequent SVs with similar breakpoints in copy number profiles.

SUPPORTING INFORMATION

Supplementary information accompanies this paper on the *Clinical Pharmacology & Therapeutics* website (www.cpt-journal.com).

Figure S1. Whole genome sequencing data for samples previously reported by GeT-RM to have more than two gene copies of *GSTT1* (*GSTT1*AxN*).
Figure S2. Three gene copies detected by Stargazer for the *CYP3A4*, *CYP3A5*, *UGT2B7*, and *UGT2B15* genes in a single sample (NA18540).
Table S1. Genotypes for 70 reference samples and 28 pharmacogenes identified by Stargazer's analysis of whole genome sequencing data.
Table S2. Star alleles previously reported by the Genetic Testing Reference Materials Coordination Program and not identified by Stargazer's analysis of whole genome sequencing data.
Table S3. Structural variant reports in the Database of Genomic Variants supporting genotype calls from Stargazer over those from the Genetic Testing Reference Materials Coordination Program.
Table S4. Demographic and sequencing information for 70 reference samples.

ACKNOWLEDGMENTS

The authors acknowledge the Genetic Testing Reference Materials Coordination Program (GeT-RM) for their generous contribution of pharmacogenetic genotypes and whole genome sequencing data.

FUNDING

This work was supported by the National Institute of General Medical Sciences (NIGMS) (R24GM115277, P50GM115318, and P01GM116691). S.B.L. is a recipient of Macrogen PhD Fellowship.

CONFLICT OF INTEREST

The authors declared no competing interests for this work.

AUTHOR CONTRIBUTIONS

S.B.L., M.M.W., K.E.T., and D.A.N. wrote the manuscript. S.B.L., K.E.T., and D.A.N. designed the research. S.B.L. performed the research. S.B.L. and M.M.W. analyzed the data.

1. Evans, W.E. & Relling, M.V. Moving towards individualized medicine with pharmacogenomics. *Nature* **429**, 464–468 (2004).
2. Koren, G., Cairns, J., Chitayat, D., Gaedigk, A. & Leeder, S.J. Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother. *Lancet* **368**, 704 (2006).
3. Pirmohamed, M., Kamali, F., Daly, A.K. & Wadelius, M. Oral anticoagulation: a critique of recent advances and controversies. *Trends Pharmacol. Sci.* **36**, 153–163 (2015).
4. Dai, D.P. *et al*. CYP2C9 polymorphism analysis in Han Chinese populations: building the largest allele frequency database. *Pharmacogenomics J.* **14**, 85–92 (2014).
5. Haining, R.L., Hunter, A.P., Veronese, M.E., Trager, W.F. & Rettie, A.E. Allelic variants of human cytochrome P450 2C9: baculovirus-mediated expression, purification, structural characterization, substrate stereoselectivity, and prochiral selectivity of the wild-type and I359L mutant forms. *Arch. Biochem. Biophys.* **333**, 447–458 (1996).
6. Sanderson, S., Emery, J. & Higgins, J. CYP2C9 gene variants, drug dose, and bleeding risk in warfarin-treated patients: a HuGEnet systematic review and meta-analysis. *Genet. Med.* **7**, 97–104 (2005).
7. Relling, M.V. & Evans, W.E. Pharmacogenomics in the clinic. *Nature* **526**, 343–350 (2015).
8. US Food and Drug Administration. Table of pharmacogenomic biomarkers in drug labeling. <https://www.fda.gov/Drugs/ScienceResearch/ucm572698.htm>. Accessed April 18, 2019.
9. Daly, A.K. & Cascorbi, I. Opportunities and limitations: the value of pharmacogenetics in clinical practice. *Br. J. Clin. Pharmacol.* **77**, 583–586 (2014).
10. Swen, J.J. *et al*. Pharmacogenetics: from bench to byte–an update of guidelines. *Clin. Pharmacol. Ther.* **89**, 662–673 (2011).
11. Zhou, S.F. Polymorphism of human cytochrome P450 2D6 and its clinical significance: part I. *Clin. Pharmacokinet.* **48**, 689–723 (2009).
12. Gaedigk, A., Simon, S.D., Pearce, R.E., Bradford, L.D., Kennedy, M.J. & Leeder, J.S. The CYP2D6 activity score: translating genotype information into a qualitative measure of phenotype. *Clin. Pharmacol. Ther.* **83**, 234–242 (2008).
13. 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).
14. Gaedigk, A. Complexities of CYP2D6 gene analysis and interpretation. *Int. Rev. Psychiatry.* **25**, 534–553 (2013).
15. Gaedigk, A. *et al*. Cytochrome P4502D6 (CYP2D6) gene locus heterogeneity: characterization of gene duplication events. *Clin. Pharmacol. Ther.* **81**, 242–251 (2007).
16. Kalman, L.V., Datta, V., Williams, M., Zook, J.M., Salit, M.L. & Han, J.Y. Development and characterization of reference materials for genetic testing: focus on public partnerships. *Ann. Lab. Med.* **36**, 513–520 (2016).
17. Pratt, V.M. *et al*. Characterization of 137 genomic DNA reference materials for 28 pharmacogenetic genes: a GeT-RM collaborative project. *J. Mol. Diagn.* **18**, 109–123 (2016).
18. Centers for Disease Control and Prevention. RM materials - material availability. <https://wwwn.cdc.gov/clia/Resources/GetRM/MaterialsAvailability.aspx>. Accessed April 18, 2019.
19. Lee, S.B. *et al*. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. *Genet. Med.* **21**, 361–372 (2019).
20. McCarroll, S.A. *et al*. Common deletion polymorphisms in the human genome. *Nat. Genet.* **38**, 86–92 (2006).
21. Oscarson, M. *et al*. Characterization of a novel CYP2A7/CYP2A6 hybrid allele (CYP2A6*12) that causes reduced CYP2A6 activity. *Hum. Mutat.* **20**, 275–283 (2002).
22. Rotger, M. *et al*. Partial deletion of CYP2B6 owing to unequal crossover with CYP2B7. *Pharmacogenet. Genomics* **17**, 885–890 (2007).
23. Black, J.L., Walker, D.L., O'Kane, D.J. & Harmandayan, M. Frequency of undetected CYP2D6 hybrid genes in clinical samples: impact on phenotype prediction. *Drug Metab. Dispos.* **40**, 111–119 (2012).
24. MacDonald, J.R., Ziman, R., Yuen, R.K., Feuk, L. & Scherer, S.W. The database of genomic variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
25. Mills, R.E. *et al*. Mapping copy number variation by population-scale genome sequencing. *Nature* **470**, 59–65 (2011).
26. Redon, R. *et al*. Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
27. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am. J. Hum. Genet.* **81**, 1084–1097 (2007).
28. 1000 Genomes Project Consortium *et al*. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
29. Wang, K. *et al*. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res.* **17**, 1665–1674 (2007).
30. Numanagić, I., Malikić, S., Pratt, V.M., Skaar, T.C., Flockhart, D.A. & Sahinalp, S.C. Cypiripi: exact genotyping of CYP2D6 using high-throughput sequencing data. *Bioinformatics* **31**, i27–i34 (2015).
31. Twist, G.P. *et al*. Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, CYP2D6, from whole-genome sequences. *NPJ Genom. Med.* **1**, 15007 (2016).
32. Klein, T.E. & Ritchie, M.D. PharmCAT: a pharmacogenomics clinical annotation tool. *Clin. Pharmacol. Ther.* **104**, 19–22 (2018).
33. Numanagić, I. *et al*. Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. *Nat. Commun.* **9**, 828 (2018).
34. Browning, S.R. & Browning, B.L. Haplotype phasing: existing methods and new developments. *Nat. Rev. Genet.* **12**, 703–714 (2011).
35. Snyder, M.W., Adey, A., Kitzman, J.O. & Shendure, J. Haplotype-resolved genome sequencing: experimental methods and applications. *Nat. Rev. Genet.* **16**, 344–358 (2015).
36. Johansson, I., Lundqvist, E., Bertilsson, L., Dahl, M.L., Sjöqvist, F. & Ingelman-Sundberg, M. Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proc. Natl. Acad. Sci. USA* **90**, 11825–11829 (1993).
37. Dahl, M.L., Johansson, I., Bertilsson, L., Ingelman-Sundberg, M. & Sjöqvist, F. Ultrarapid hydroxylation of debrisoquine in a Swedish population. Analysis of the molecular genetic basis. *J. Pharmacol. Exp. Ther.* **274**, 516–520 (1995).
38. Chida, M. *et al*. New allelic arrangement CYP2D6*36x2 found in a Japanese poor metabolizer of debrisoquine. *Pharmacogenetics* **12**, 659–662 (2002).
39. Nakano, M. *et al*. CYP2A7 pseudogene transcript affects CYP2A6 expression in human liver by acting as a decoy for miR-126. *Drug Metab. Dispos.* **43**, 703–712 (2015).
40. Clinical Pharmacogenetics Implementation Consortium. Genes-drugs. <https://cpicpgx.org/genes-drugs/>. Accessed April 18, 2019.
41. Raczy, C. *et al*. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics* **29**, 2041–2043 (2013).

42. Gaedigk, A. *et al*. The Pharmacogene Variation (PharmVar) Consortium: incorporation of the human cytochrome P450 (CYP) allele nomenclature database. *Clin. Pharmacol. Ther.* **103**, 399–401 (2018).
43. Whirl-Carrillo, M. *et al*. Pharmacogenomics knowledge for personalized medicine. *Clin. Pharmacol. Ther.* **92**, 414–417 (2012).
44. UGT Nomenclature Committee. UGT alleles nomenclature home page. <https://www.pharmacogenomics.pha.ulaval.ca/ugt-alleles-nomenclature/>. Accessed April 18, 2019.
45. NAT Nomenclature Committee. NAT alleles nomenclature home page. <http://nat.mbg.duth.gr>. April 18, 2019.
46. TPMT Nomenclature Committee. TPMT alleles nomenclature home page. <https://www.imh.liu.se/tpmtalleles/tabell-over-tpmt-alleler?l=en>. Accessed April 18, 2019.
47. McKenna, A. *et al*. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
48. Robinson, J.T. *et al*. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
49. Ng, S.B. *et al*. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
50. Killick, R. & Eckley, I.A. Changepoint: an R package for changepoint analysis. *J. Stat. Softw.* **58**, 1–19 (2014).