bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Accurate *CYP2D6* genotyping using whole genome sequencing data

Xiao Chen[1], Fei Shen[1], Nina Gonzaludo[1], Alka Malhotra[1], Cande Rogert[1], Ryan J Taft[1], David R Bentley[2], Michael A Eberle[1]
1. Illumina Inc., 5200 Illumina Way, San Diego, CA, USA
2. Illumina Cambridge Ltd., Illumina Centre 19 Granta Park, Great Abington, Cambridge, UK

Corresponding author: Michael Eberle, meberle@illumina.com

Abstract

Responsible for the metabolism of 25% of all drugs, *CYP2D6* is a critical component of personalized medicine initiatives. Genotyping *CYP2D6* is challenging due to sequence similarity with its pseudogene paralog *CYP2D7* and a high number and variety of common structural variants (SVs). Here we describe a software tool, Cyrius, that accurately genotypes *CYP2D6* using whole-genome sequencing. We show that Cyrius has superior performance (96.5% concordance with truth genotypes) compared to existing methods (83.8-86.6%). Using Cyrius, we built a haplotype frequency database from 2504 ethnically diverse samples and estimate that SV-containing star alleles are more frequent than previously reported.

Keywords

Bioinformatics, whole-genome sequencing, pharmacogenomics, precision medicine, population studies, *CYP2D6*

Background

There is significant variation in the response of individuals to a large number of clinically prescribed drugs. A strong contributing factor to this differential drug response is the genetic composition of the drug-metabolizing genes, and thus genotyping pharmacogenes is important for personalized medicine[1]. Cytochrome P450 2D6 (*CYP2D6*) is one of the most important drug-metabolizing genes and is responsible for the metabolism of 25% of drugs[2]. The *CYP2D6* gene is highly polymorphic, with 131 star alleles defined by the Pharmacogene Variation (PharmVar) Consortium (https://www.pharmvar.org/gene/*CYP2D6*)[3]. Star alleles are *CYP2D6* gene copies defined by a combination of small variants (SNVs and indels) and structural variants (SVs), and correspond to different levels of *CYP2D6* enzymatic activity, i.e. poor, intermediate, normal, or ultrarapid metabolizer[4,5].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Genotyping *CYP2D6* is challenged by common deletions and duplications of *CYP2D6* and fusions between *CYP2D6* and its pseudogene paralog, *CYP2D7*, that is upstream of *CYP2D6* [6,7]. An additional difficulty is that *CYP2D7* shares 94% sequence similarity, with a few near-identical regions[6,8]. Traditionally, *CYP2D6* genotyping is done with arrays or polymerase chain reaction (PCR) based methods such as TaqMan assays, ddPCR and long-range PCR. These assays differ in the number of star alleles (variants) they interrogate, leading to variability in genotyping results across assays[9]. For example, a different allele is reported when the variant defining the true star allele is not interrogated[6,9,10], e.g. *2 is reported when *45 (with one more variant than *2) is not interrogated and the wild-type allele *1 is reported when none of the interrogated variants are detected. In addition, many of these assays are low throughput and often have difficulty in accurately detecting SVs[6].

With recent advances in next-generation sequencing (NGS), it is now possible to profile the entire genome at high-throughput and in a clinically-relevant timeframe. Driven by these advances, many countries are undertaking large scale population sequencing projects[11–13] wherein pharmacogenomics testing will greatly increase the clinical utility of these efforts. There exists a few informatics genotyping methods for *CYP2D6* (Cypiripi[14], Astrolabe (formerly Constellation)[15], Aldy[16] and Stargazer[17,18]) that can be applied to targeted (PGRNseq[19]) and/or whole genome sequencing (WGS) data. Among these, Cypiripi and Astrolabe (both developed more than four years ago) were not designed to detect complex SVs and have been shown to have much lower performance than the more recent methods[16]. The two most recent *CYP2D6* callers, Aldy and Stargazer, work by detecting SVs based on depth and deriving the haplotype configurations based on the observed small variants and SVs. However, they rely on accurate read alignments, which is often not possible at many positions throughout the gene as the sequence is highly similar or even indistinguishable with *CYP2D7*. As a result, this often leads to ambiguous depth patterns or false positive/negative small variant calls. Another limitation of both Aldy and Stargazer is that, currently, neither method supports hg38 so many studies will require a re-alignment to hg37 to use these tools.

The availability of a panel of reference samples by the CDC Genetic Testing Reference Material Program (GeT-RM)[9,20], where the consensus genotypes of major pharmacogenetic genes are derived using multiple genotyping platforms, has enabled assessment of genotyping accuracy for newly developed methods. The GeT-RM samples cover 43 *CYP2D6* star alleles. In addition, the availability of high quality long reads can provide a complete picture of *CYP2D6* for improved validation of complicated variants and haplotypes[9,20]. Here we describe Cyrius, a novel WGS-specific *CYP2D6* genotyping method that overcomes the challenges with the homology between *CYP2D6* and *CYP2D7* (referred to as *CYP2D6/7* hereafter). We demonstrate superior genotyping accuracy compared to other methods in 138 GeT-RM reference samples and 8 samples with PacBio HiFi data, covering 41 known star alleles. Finally, we applied this method to high depth sequence data on 2504 unrelated samples from the 1000 Genomes Project[21] (1kGP) to report on the distribution of star alleles across five ethnic populations. This analysis demonstrates differences with frequencies in PharmGKB, highlighting the potential errors associated with merging limited star allele calls made using diverse technologies designed to identify specific subsets of the known star alleles. This analysis

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

expands the current understanding on *CYP2D6*'s genetic diversity, particularly on complex star alleles with SVs.

Methods

Samples

We analyzed WGS data for 138 GeT-RM reference samples, including 96 samples that were genotyped in the initial GeT-RM study[9] and updated in the latest GeT-RM release[20], as well as 42 additional samples that were newly added in the latest GeT-RM release. For the first batch of 96 samples, WGS was performed with TruSeq DNA PCR-free sample preparation with 150bp paired reads sequenced on Illumina HiSeq X instruments. Genome build GRCh37 was used for read alignment. The WGS data for the second batch of 42 samples were downloaded from NYGC as part of the 1000 Genomes Project (see below).

For population studies, we used the 1000 Genomes Project (1kGP) data, for which WGS BAMs for 2504 samples were downloaded from https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31736/. These BAMs were generated by sequencing 2x150bp reads on Illumina NovaSeq 6000 instruments from PCR-free libraries sequenced to an averaged depth of at least 30X and aligned to the human reference, hs38DH.

PacBio sequencing data for 8 samples were downloaded from the Genome in a Bottle (GIAB) Consortium (available in SRA under PRJNA540705, PRJNA529679, and PRJNA540706), and http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/.

CYP2D6 genotyping method

Cyrius first calls the sum of the copy number (CN) of *CYP2D6/7*, following a similar method as previously described[22]. In brief, read counts are calculated directly from the WGS aligned BAM file using all reads mapped to either *CYP2D6* or *CYP2D7*. The summed read count is normalized and GC corrected. CNs of *CYP2D6*+*CYP2D7* are called from a Gaussian mixture model built on the normalized depth values. We use the same approach to call the CN of the 1.5kb spacer region between the repeat REP7 and *CYP2D7* (Figure 1). Subtracting the spacer CN from total *CYP2D6*+*CYP2D7* CN gives the CN of genes that are *CYP2D6*-derived downstream of the gene and contain REP6 (can be complete *CYP2D6* or fusion gene ending in *CYP2D6*).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

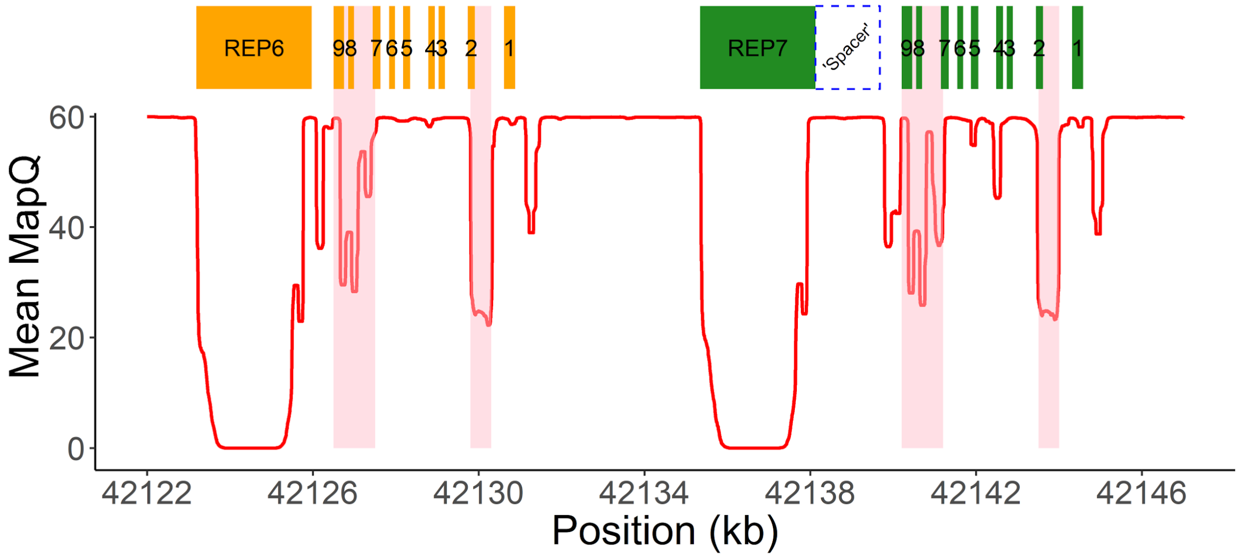


Figure 1. WGS data quality in *CYP2D6/7* region

Mean mapping quality across 1kGP samples are plotted for each position in the *CYP2D6/7* region (hg38). A median filter is applied in a 200bp window. The 9 exons of *CYP2D6/7* are shown as orange (*CYP2D6*) and green (*CYP2D7*) boxes. Two 2.8kb repeat regions downstream of *CYP2D6* (REP6) and *CYP2D7* (REP7) are identical and essentially unalignable. The purple dashed line box denotes the spacer region between *CYP2D7* and REP7. Two major homology regions within the genes are shaded in pink.

We identified 117 *CYP2D6/7* differentiating bases (see Supplementary information, Figure S1). Cyrius estimates the number of copies of *CYP2D6* at each of these differentiating base positions. Based on the called total *CYP2D6*+*CYP2D7* CN, Cyrius calls the combination of *CYP2D6* CN and *CYP2D7* CN that produces the highest posterior probability for the observed number of reads supporting *CYP2D6*- and *CYP2D7*-specific bases, as described previously[22]. Gene fusions are identified when the CN of *CYP2D6* changes within the gene (Figure 2).

Cyrius parses the read alignments to identify the small variants that define star alleles and call their CNs. These variants are divided into two classes: 1) variants that fall in *CYP2D6/7* homology regions, i.e. the shaded low mapping quality regions in Figure 1, and 2) variants that occur in unique regions of *CYP2D6*. For the former, Cyrius looks for variant reads in *CYP2D6* and its corresponding site in *CYP2D7*. For the latter, Cyrius only uses the reads aligned to *CYP2D6*. *CYP2D6* CNs at the variant sites are taken into account during small variant calling so that a variant can be called at one copy, two copies or any CN less than or equal to the *CYP2D6* CN at that site.

Finally, Cyrius matches the called SVs and small variants against the definition of star alleles (downloaded and parsed from PharmVar, https://www.pharmvar.org/gene/*CYP2D6*, last accessed on 5/1/2020) to call star alleles. These star alleles are then grouped into haplotypes

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

when, for example, there are more than two copies of *CYP2D6*. For this we include prior information to define the exact haplotypes, e.g. *68 is on the same haplotype as *4, and *36 is on the same haplotype as *10. These priors are made based on the tandem arrangement patterns described in PharmVar and are also supported by our truth data. We also provide an option to only match against star alleles with known functions.

Of the 131 star alleles defined in PharmVar, 25 are still awaiting curation, so we excluded those and focused on the 106 curated ones (another option is provided in Cyrius to include those uncurated ones, see Discussion). Of these 106 star alleles, we removed four from our target list, none of which are in GeT-RM. They include *61 and *63 (both classified as “unknown function” by PharmVar), which are *CYP2D6/7* hybrid genes very similar to *36 with the fusion breakpoint slightly upstream. Since we cannot distinguish the Exon7-Exon8 region between *CYP2D6/7* (Figure S1), these two star alleles cannot be distinguished from *36, and they will be called as *36 by Cyrius. Additionally, we removed *27 (normal function) and *32 (unknown function), which share g.42126938C>T, a gene conversion variant in a high homology region (a variant read will align to *CYP2D7* perfectly, leading to reduced accuracy to call the CN of this variant). As a result, *27 will be called as *1 and *32 will be called as *41.

Validating against truth from GeT-RM and long reads

When comparing the *CYP2D6* calls made by Cyrius, Aldy and Stargazer against the consensus genotypes provided by GeT-RM, a genotype is considered a match as long as all of the star alleles in the truth genotype are present, even if the haplotype assignment is different. An example of this occurs in several samples listed by GeT-RM as *1/*10+*36+*36 but called by Aldy as *1+*36/*10+*36.

When validating genotype calls against the PacBio data, PacBio reads that cover the entire *CYP2D6* gene were analyzed to identify small variants known to define the star alleles. Long (~10kb) reads allow us to fully phase these variants into haplotypes and these haplotypes are matched against the star allele definitions to determine the star allele. Reads carrying SVs were determined by aligning reads against a set of reference contigs that were constructed to represent known SVs (*5/*13/*36/*68/duplications). Visualization in Figure 3 was done using the software tool sv-viz2[23].

Running Aldy and Stargazer

Aldy v2.2.5 was run using the command “aldy genotype -p illumina -g *CYP2D6*”.

Stargazer v1.0.7 was run to genotype *CYP2D6* using VDR as the control gene, with GDF and VCF files as input.

The 1kGP GeT-RM samples were originally aligned against hg38. As Aldy and Stargazer only support GRCh37, for comparison between methods, these samples were realigned against GRCh37 using Isaac[24].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Results

Validation and performance comparison

We compared the *CYP2D6* calls made by Cyrius, Aldy and Stargazer against 144 samples where we were able to obtain high quality ground truth. These 144 samples included 138 GeT-RM samples and 8 samples with truth generated using PacBio HiFi sequence reads (1, two samples overlap between GeT-RM and PacBio). Samples with SVs show distinct depth signals that allow us to call SVs accurately (Figure 2, see Methods). The long reads allowed us to locate and visualize breakpoints of the common SVs in the region (Figure 3) and thus serve as a valuable resource for studying complex star alleles and confirm the phasing of the variants for the star alleles.

Comparing against the GeT-RM samples we found three samples where the calls of all three software methods agree with each other but disagree with the GeT-RM consensus (Table S1). First, for NA18519, the WGS-based genotype is *106/*29 with reads carrying the variant defining *106 shown in Figure S2. The GeT-RM consensus is *1/*29 but, excluding sequencing, none of the GeT-RM assays interrogate *106. Both samples in GeT-RM that have *106 were detected by Sanger sequencing or NGS (see Table S2), while no sequencing was done for NA18519. Therefore, the genotype of NA18519 should be updated to *106/*29. The remaining two samples have the *68 fusion that is not represented in the GeT-RM consensus. For these, the depth profiles show a CN gain in Exon 1 (Figure S3) and are highly similar to NA12878 (*3/*4+*68, Figure 2), where the *68 fusion is confirmed by PacBio reads (Figure 3). Comparing these two samples with 6 samples in GeT-RM that have *68 in their consensus genotypes, they only have results from two assays that have the lowest accuracy for *68 - PharmacoScan and iPLEX *CYP2D6* v1.1 (without custom panel and VeriDose) (Table S2). In addition, no TaqMan CNV result is available for Exon 1, which is the region affected by *68 fusion. Therefore, the truth genotypes are likely to be GeT-RM errors and we removed these two samples from the concordance calculation for all three callers.

Among 142 truth samples, Cyrius initially made five discordant calls, showing a concordance of 96.5% (Table 1). Included amongst these discrepancies is the sample NA19908 (GeT-RM defined *1/*46), where Cyrius called *1/*46 and *43/*45 as two possible diplotypes. Both of these two star allele combinations result in the same set of variants and neither read phasing or population frequency analysis could rule out either genotype combination. The genotyping results from various assays that generated the GeT-RM consensus for this sample also showed disagreement between *1/*46 and *43/*45, highlighting the difficulty of these combinations (Table S2). Future sequencing of more samples of either diplotypes could help identify new variants that distinguish the two.

In the remaining four samples where Cyrius is discordant with the truth, we were able to identify the causes and improve Cyrius to correctly call these star alleles. First, in NA23275 (*1/*40), the 18bp insertion defining *40 was originally missed as the insertion-containing reads were often not aligned as having an insertion but as soft-clips. Second, in HG03225 (*5/*56), *CYP2D7*-

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

derived reads were mis-aligned to *CYP2D6*, preventing the *56 defining variant from being called. Third, in NA18565, we miscalled *10/*36x2 to be *10/*10+*36, as did the other two callers. The depth profile shows a gene-conversion version of *36 (with REP6) in addition to the fusion version of *36 (with REP7) (Figure S4). Lastly, in HG00421 (*10x2/*2), we miscalled the fusion to be *36, as did the other two callers. The depth profile shows a different SV, *10D, with the duplication breakpoint located downstream of Exon 9 (Figure S4). The last two samples have rare SV patterns so it is challenging to call such genotypes without seeing a known sample, as suggested by the fact that all three callers made the same wrong calls. While we treat these four samples as miscalls for this study, we made improvements to Cyrius after seeing these samples, allowing us to call them accurately and reach a concordance of 99.3% (141 out of 142 samples). This highlights how more truth data and more population data will identify limitations that can enable improvements to the caller for subsequent samples.

In contrast, both of the other *CYP2D6* callers had concordance less than 90% when compared against the 142 samples. Aldy has a concordance of 86.6%. In particular, it overcalled several *CYP2D6/7* fusions such as *61, *63, *78 and *83 (called in 7 out of 19 discordant samples, Table S1), even in samples without SV. Stargazer has a concordance of 83.8% and is most prone to errors when SVs are present. The concordance in samples with SVs is only 75.0%, and 13 out of the 23 discordant calls are in samples with SVs (Table 1). Notably, Stargazer has a high error rate with the *36 fusion (7 wrong calls out of 17 total samples with *36), and miscalled all 4 samples that have more than one copy of *36 on one haplotype (Table S1).

Together, the validation samples used in this study confirmed our *CYP2D6* calling accuracy in 47 distinct haplotypes (Table 2), including 41 star alleles as well as several SV structures, such as duplications, *2+*13, *4+*68, *10+*36 and *10+*36+*36. These 41 star alleles represent 38.7% of the 106 curated star alleles currently listed in PharmVar and 53.4% (31 out of 58) of the ones with known function.

Table 1. Summary of benchmarking results against truth in 142 samples (after removing 2 samples that are likely GeT-RM errors, see Figure S3)

Caller	Total concordant	Concordance	Deletion N=15	Duplication N=13	Fusion N=24	No SV N=90	Concordance, samples with SV	Concordance, samples without SV
Cyrius	137*	96.5%	14	12	23	88	94.2%	97.8%
Aldy	123	86.6%	13	11	21	78	86.5%	86.7%
Stargazer	119	83.8%	14	10	15	80	75.0%	88.9%

*Cyrius has since been improved and can correctly call 141 (99.3%) out of 142 of these samples.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

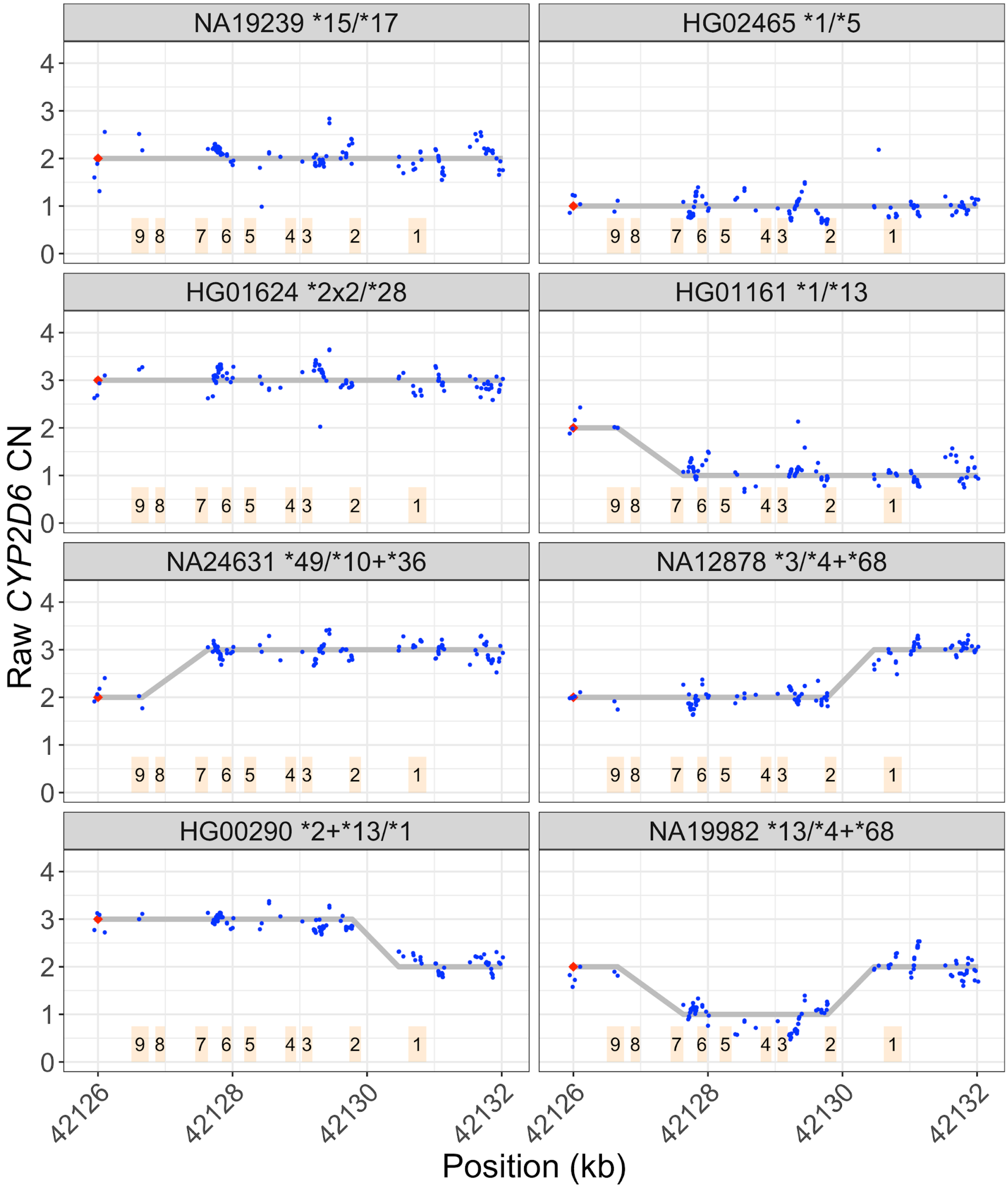


Figure 2. Depth patterns in samples with different types of SVs.

Blue dots denote raw *CYP2D6* CNs across *CYP2D6/7* differentiating sites. Raw *CYP2D6* CN is calculated as the total *CYP2D6*+*CYP2D7* CN multiplies the ratio of *CYP2D6* supporting reads out of *CYP2D6* and *CYP2D7* supporting reads. The red diamond denotes the CN of genes that

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

are *CYP2D6*-derived downstream of the gene (can be complete *CYP2D6* or fusion gene ending in *CYP2D6*), calculated as the total *CYP2D6*+*CYP2D7* CN minus the CN of the *CYP2D7* spacer region (see Figure 1). To detect SVs, a *CYP2D6* CN is called at each *CYP2D6/7* differentiating site (see Methods) and a change in *CYP2D6* CN within the gene indicates the presence of a fusion. For example, in HG01161, the *CYP2D6* CN changes from 2 to 1 between Exon 9 and Exon 7, indicating a *CYP2D7*-*CYP2D6* hybrid gene. In NA12878, the *CYP2D6* CN changes from 2 to 3 between Exon 2 and Exon 1, indicating a *CYP2D6*-*CYP2D7* hybrid gene. The depth profiles for different SV patterns are shown in NA19239 (no SV), HG02465 (complete deletion), HG01624 (complete duplication), HG01161 (fusion deletion), NA24631 (fusion duplication, *36), NA12878 (fusion duplication, *68), HG00290 (tandem arrangement *2+*13), and NA19982 (two different fusions, *13 and *68, on two haplotypes). The fusions in NA24631 and NA12878 are confirmed with PacBio reads in Figure 3.

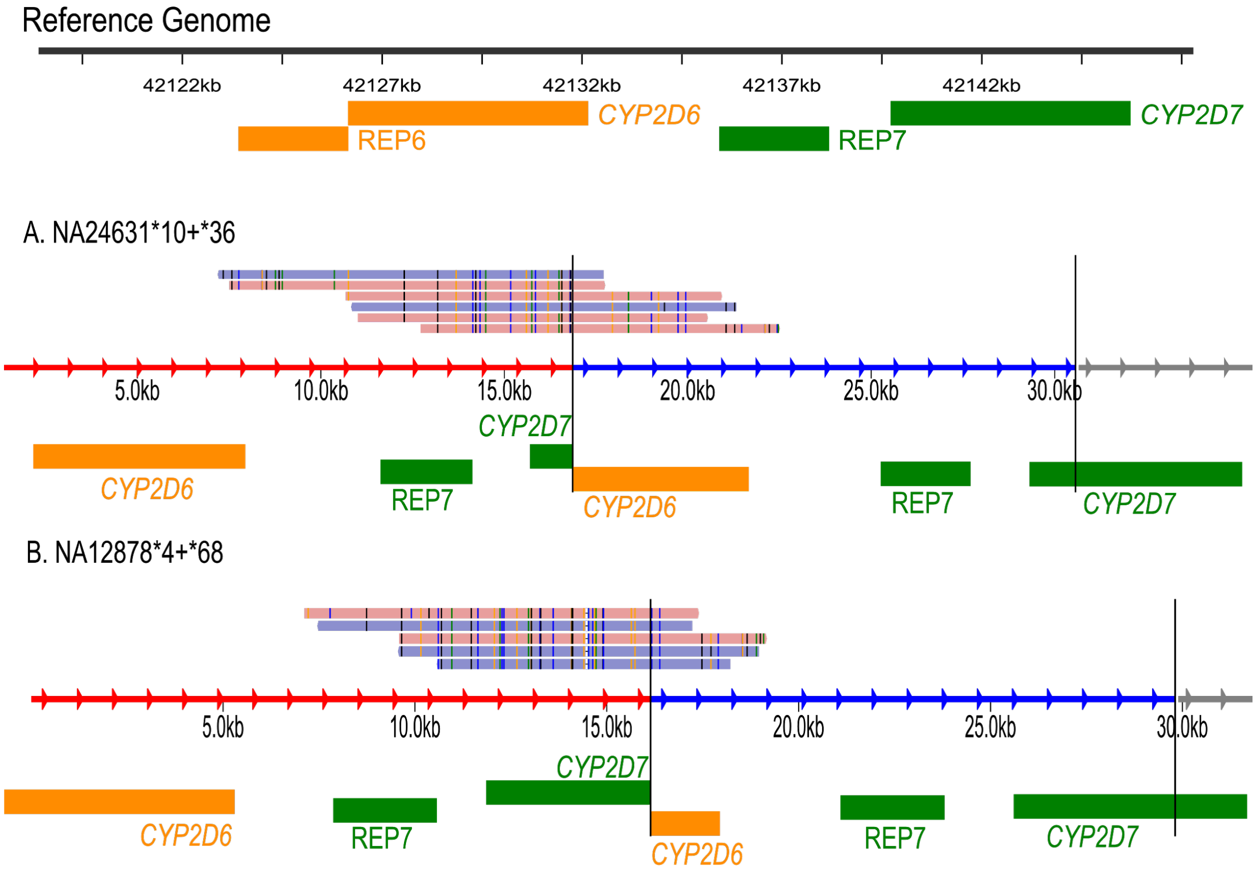


Figure 3. Structural variants validated by PacBio HiFi reads.

PacBio reads supporting fusion duplications *36 and *68, confirming SVs called in NA24631 and NA12878 (third row, Figure 2). PacBio reads were realigned against sequence contigs representing the fusions and plots were generated using sv-viz2[22]. The black vertical lines mark the boundaries of the duplicated sequences, represented by the blue region (the original copy inside the red region). The red and gray regions represent sequences upstream and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

downstream, respectively. The genotypes are *49/*10+*36 (NA24631) and *3/*4+*68 (NA12878).

CYP2D6 haplotype frequencies across five ethnic populations

Given the high accuracy demonstrated in the previous section, we next looked beyond the validation samples to study *CYP2D6* in the global population. For this, we analyzed the haplotype distribution by population (Europeans, Africans, East Asians, South Asians and admixed Americans consisting of Colombians, Mexican-Americans, Peruvians and Puerto Ricans) in 2504 1kGP samples (Figure 4A, Table 2, Table S3). Cyrius made definitive diplotype calls in 2445 (97.6%) out of 2504 samples calling 46 distinct star alleles. Of these 46 star alleles, 41 overlapped star alleles that had been included in the validation data. These 41 star alleles that had been tested in the validation data represent 96.1% of all the star alleles called in the 1kGP samples (Table 2).

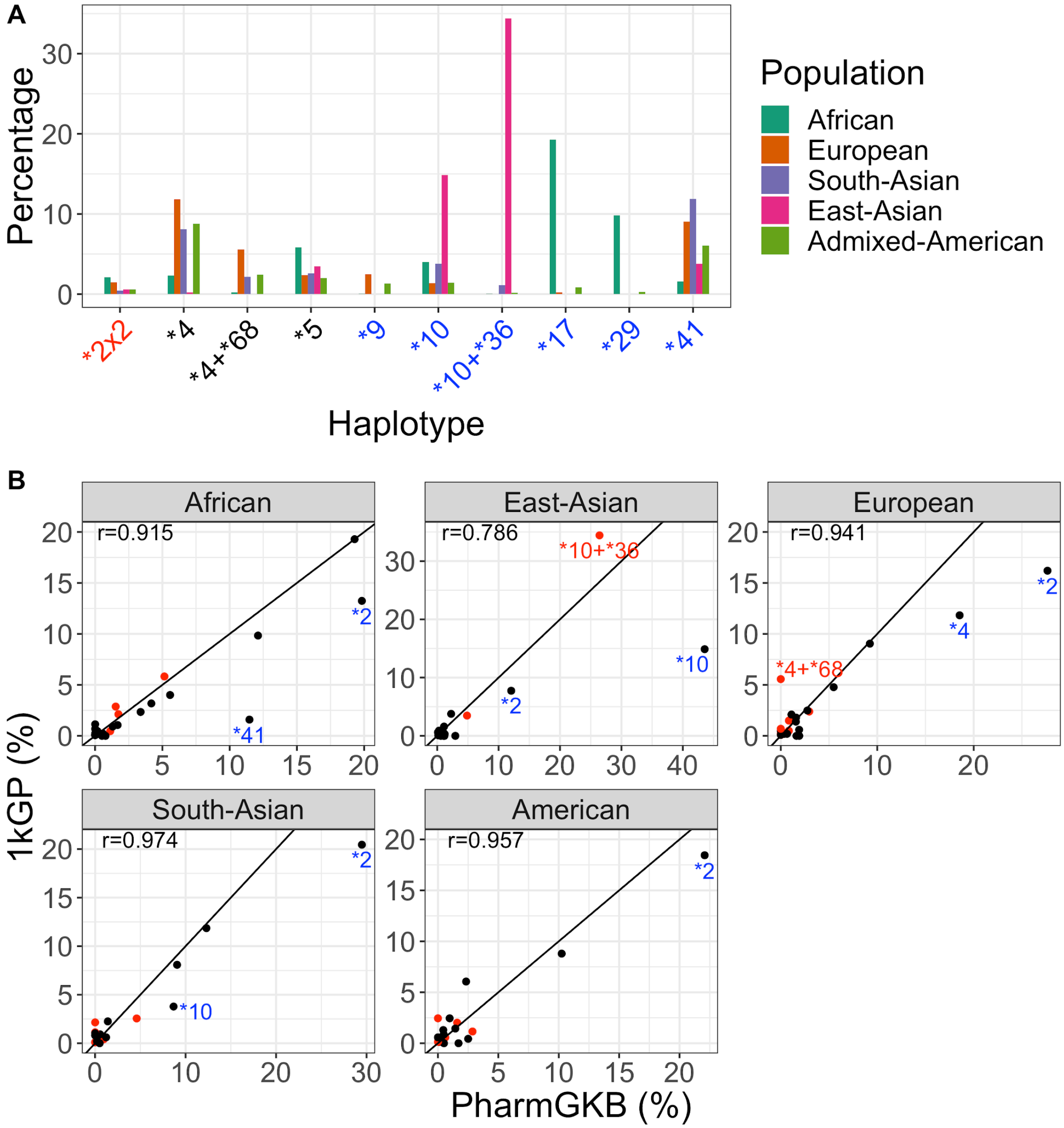
In the 59 samples where Cyrius did not call a definitive diplotype, 10 samples had a non-definitive SV call, 30 samples had variant calls that did not match any of the known star alleles, 4 samples have the same ambiguity between *1/*46 and *43/*45 as described in the validation sample NA19908 above, and 15 samples have definitive star allele calls that Cyrius was not able to unambiguously phase into haplotypes.

Mostly, the haplotype frequencies agree with pharmGKB[5,25] (Figure 4B, Figure S5, Table S4, pharmGKB last accessed on 5/1/2020). For example, Africans have a high frequency of *17 and *29, South-Asians have a high frequency of *41, Europeans have a high frequency of *4, and East Asians have a high frequency of *10 (Figure 4A). With the improved sensitivity for SVs by Cyrius, we are able to provide a more comprehensive picture on the frequencies of SVs across populations. Among those, the fusion-containing haplotype *10+*36 is very common in East-Asians, and another fusion-containing haplotype *4+*68 is also quite common in Europeans. We report a higher frequency than PharmGKB for both of these SV-containing haplotypes (Figure 4B, red annotated dots). Previously reported frequencies of *10+*36 in East-Asians fall into a wide range (10-35%)[26–31], indicating that different assays have different sensitivity for this haplotype. Additionally, *68 is often not interrogated in many studies, and it has been suggested that >20% of reported *4 alleles are actually in tandem with *68[32]. Together, we estimate that the frequencies of haplotypes involving SVs are ~6% higher than reported in PharmGKB in East-Asians and Europeans (Table S5).

There are a few other star alleles for which we report a lower frequency than PharmGKB (Figure 4B), highlighting the difficulty of merging data from multiple studies using different technologies. These include *2 in all five populations. Since *2 is often reported if some other star alleles are not interrogated, its frequency could be overestimated in PharmGKB. Similarly, *10 is overestimated in PharmGKB in East-Asians and South-Asians, as *10 is reported when some other star alleles, particularly *36 and *10+*36, are not tested. Additionally, *4 is overestimated in PharmGKB in Europeans, as *4 is reported when *4+*68 is not tested. Finally, we report a

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

lower frequency for *41 in Africans. It is known that *41 has not been consistently determined by its defining SNP across studies, leading to an overestimation of its frequency, especially in those of African ancestry (quoted from PharmGKB).



bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4. *CYP2D6* allele frequencies across five ethnic populations.

A. Ten most common haplotypes with altered *CYP2D6* function. Those with increased function are labeled in red, those with no function in black and those with decreased function in blue. B. Comparison between 1kGP and PharmGKB frequencies. Each dot represents a haplotype with a frequency >=0.5% in either 1kGP or pharmGKB. SV-related haplotypes are marked in red, including the two haplotypes with the largest deviation (*10+*36 in East-Asians and *4+*68 in Europeans). Other haplotypes with deviated values are annotated in blue. A diagonal line is drawn for each panel. Correlation coefficients are listed for each population.

Discussion

We presented a new software tool, Cyrius, that can accurately genotype the difficult *CYP2D6* region. In this study we used long read data to validate both the haplotypes and the SVs. Long reads provide a unique opportunity to confirm the breakpoint regions for common SVs and confirm the phasing of the *CYP2D6* gene. Using 144 samples, including 8 with long read data, as an orthogonal validation dataset, we showed that Cyrius outperforms other *CYP2D6* genotypers, achieving 96.5% concordance versus 86.6% for Aldy and 83.8% for Stargazer. In particular, compared to these existing *CYP2D6* callers, Cyrius allows for the possibility that reads may be misaligned in the regions where *CYP2D6/7* have high-similarity. Ambiguous read alignments in these regions can lead to incorrect copy number estimation and errors in small variant calling. By accounting for possible mis-aligned reads and selecting a set of reliable *CYP2D6/7* differentiating sites, Cyrius is able to do a much better job identifying star alleles with SVs, achieving 94.2% concordance compared to 86.5% for Aldy and 75.0% for Stargazer.

Across these 144 validation samples, we were able to validate genotype calls that included 41 different star alleles. These 41 star alleles represent 38.7% of all curated star alleles listed in PharmVar and 53.4% of the ones with a known functional status. Even though our validation set included only 38.7% of the total known star alleles, based on our analysis of the 1kGP samples, we estimate that they represent roughly 96% of the star alleles in the pangenomic population. In general, the allele frequencies we calculated for the 2504 1kGP samples from five ethnic populations agree with previous studies for single copy star alleles. Conversely, for some of the star alleles that were defined by the presence of SVs, we identified higher frequencies, likely because SV-impacted star alleles, particularly *36 and *68, are difficult to resolve with conventional assays. In addition, a star allele is often reported when another star allele is not interrogated by an assay[9,10] (e.g. *10 in place of *36 or *10+*36 and *4 in place of *4+*68), and thus our calculated frequencies for these alleles are lower than PharmGKB. This highlights the inherent difficulties of merging results from studies that used a variety of different *CYP2D6* assays that are each designed to call just a subset of star alleles. For example, of the 5 assays used to generate the GeT-RM consensus genotypes, the individual concordance ranged from 49.3% to 75.2% when compared against the consensus (Table S6). A single method that is able to resolve all of the known star alleles from a single assay is a better choice to build a population-level database. A previous study used small variant calls from whole exome and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

whole genome sequencing data to report allele frequencies for single copy star alleles only[33]. This study presents the first comprehensive *CYP2D6* haplotype frequency database, including SV-containing haplotypes, built upon NGS data by a targeted genotyper designed for *CYP2D6*.

In our analysis of the 1kGP samples, Cyrius is able to call a definitive genotype in 97.6% of the samples. A future direction is to better understand the 2.4% of the samples that were not genotyped and improve our algorithm so that it can also resolve these genotypes. For example, in samples where multiple haplotype configurations are possible, it could be useful to take a probabilistic approach to derive the most likely genotype given the observed variants. In addition, continuing to sequence and test more samples will help confirm our ability to genotype rare star alleles and will also identify new variants that can be used to distinguish ambiguous diplotypes. This process was demonstrated in this study where we made improvements to better call four star alleles that were initially mis-called in the 144 validation samples.

As new star alleles are identified, we will continue to incorporate them into the Cyrius database. A possible problem with adding new star alleles that are defined by new variants is that these variants are unlikely to be considered in the previous star allele definitions. As a result, there could exist novel combinations of new and existing variants that could not match any of the known combinations, leading to no-calls. For example, we include an option in Cyrius to genotype against 25 new uncurated star alleles added in PharmVar v4 (not included in GeT-RM, Aldy or Stargazer). However, 5 (*119, *122, *135, *136, *139) of the 25 new star alleles have new variants that, when included, lead to no-calls in samples that we are currently able to call. This suggests that there exist common novel star alleles with variant combinations not captured in PharmVar. As a result, we removed these 5 star alleles together with two others (*127, with a gene conversion variant in homology region, and *131, with a variant at a noisy site), keeping the remaining 18. For future studies, special attention should be paid to the possibility of novel star alleles as new variants/star alleles are identified. Public WGS datasets like the 2504 1kGP samples analyzed here will be an important component of integrating new variants into the star allele definitions because this data will allow variants to be rapidly assessed across many samples with diverse genotypes.

WGS provides a unique opportunity to profile all genetic variations for the entire genome but many of regions/variants that are clinically important are beyond the ability of most secondary analysis pipelines. *CYP2D6* is among the difficult regions in the genome that are both clinically important and also require specialized informatics solutions in addition to normal WGS pipelines. Such targeted methods have already been applied successfully to some difficult regions, such as repeat expansions[28] and the *SMN1* gene[20] responsible for spinal muscular atrophy. With the continued development of more targeted methods like Cyrius, we can help accelerate pharmacogenomics and move one step closer towards personalized medicine.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Conclusions

Cyrius is an accurate *CYP2D6* genotyper based on whole-genome sequencing data. Cyrius outperforms existing *CYP2D6* callers particularly for star alleles involving SVs. Cyrius will be a useful tool for pharmacogenomics applications with WGS and help bring the promise of precision medicine one step closer to reality.

Figure legends

Figure 1. WGS data quality in *CYP2D6/7* region

Mean mapping quality across 1kGP samples are plotted for each position in the *CYP2D6/7* region (hg38). A median filter is applied in a 200bp window. The 9 exons of *CYP2D6/7* are shown as orange (*CYP2D6*) and green (*CYP2D7*) boxes. Two 2.8kb repeat regions downstream of *CYP2D6* (REP6) and *CYP2D7* (REP7) are identical and essentially unalignable. The purple dashed line box denotes the spacer region between *CYP2D7* and REP7. Two major homology regions within the genes are shaded in pink.

Figure 2. Depth patterns in samples with different types of SVs.

Blue dots denote raw *CYP2D6* CNs across *CYP2D6/7* differentiating sites. Raw *CYP2D6* CN is calculated as the total *CYP2D6+CYP2D7* CN multiplies the ratio of *CYP2D6* supporting reads out of *CYP2D6* and *CYP2D7* supporting reads. The red diamond denotes the CN of genes that are *CYP2D6*-derived downstream of the gene (can be complete *CYP2D6* or fusion gene ending in *CYP2D6*), calculated as the total *CYP2D6+CYP2D7* CN minus the CN of the *CYP2D7* spacer region (see Figure 1). To detect SVs, a *CYP2D6* CN is called at each *CYP2D6/7* differentiating site (see Methods) and a change in *CYP2D6* CN within the gene indicates the presence of a fusion. For example, in HG01161, the *CYP2D6* CN changes from 2 to 1 between Exon 9 and Exon 7, indicating a *CYP2D7-CYP2D6* hybrid gene. In NA12878, the *CYP2D6* CN changes from 2 to 3 between Exon 2 and Exon 1, indicating a *CYP2D6-CYP2D7* hybrid gene. The depth profiles for different SV patterns are shown in NA19239 (no SV), HG02465 (complete deletion), HG01624 (complete duplication), HG01161 (fusion deletion), NA24631 (fusion duplication, *36), NA12878 (fusion duplication, *68), HG00290 (tandem arrangement *2+*13), and NA19982 (two different fusions, *13 and *68, on two haplotypes). The fusions in NA24631 and NA12878 are confirmed with PacBio reads in Figure 3.

Figure 3. Structural variants validated by PacBio HiFi reads.

PacBio reads supporting fusion duplications *36 and *68, confirming SVs called in NA24631 and NA12878 (third row, Figure 2). PacBio reads were realigned against sequence contigs representing the fusions and plots were generated using sv-viz2[22]. The black vertical lines mark

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

the boundaries of the duplicated sequences, represented by the blue region (the original copy inside the red region). The red and gray regions represent sequences upstream and downstream, respectively. The genotypes are *49/*10+*36 (NA24631) and *3/*4+*68 (NA12878).

Figure 4. *CYP2D6* allele frequencies across five ethnic populations.

A. Ten most common haplotypes with altered *CYP2D6* function. Those with increased function are labeled in red, those with no function in black and those with decreased function in blue. B. Comparison between 1kGP and PharmGKB frequencies. Each dot represents a haplotype with a frequency >=0.5% in either 1kGP or pharmGKB. SV-related haplotypes are marked in red, including the two haplotypes with the largest deviation (*10+*36 in East-Asians and *4+*68 in Europeans). Other haplotypes with deviated values are annotated in blue. A diagonal line is drawn for each panel. Correlation coefficients are listed for each population.

Table 2. Haplotypes validated in this study and their frequencies in 1kGP

Haplotype	Pan-ethnic	European	Admixed-American	East-Asian	African	South-Asian	Validated in this study	In GeT-RM full set	Function
*1	33.39	35.69	45.97	26.19	26.25	39.16	x	x	Normal
*2	14.86	16.2	18.44	7.74	13.24	20.45	x	x	Normal
*3	0.54	1.79	0.58	0	0.23	0.2	x	x	None
*4	5.83	11.83	8.79	0.2	2.34	8.08	x	x	None
*5	3.49	2.39	2.02	3.47	5.82	2.56	x	x	None
*6	0.5	2.09	0.29	0	0.08	0.1	x	x	None
*7	0.18	0	0	0	0	0.92	x	x	None
*9	0.7	2.49	1.3	0	0.08	0	x	x	Decreased
*10	5.27	1.39	1.44	14.88	4.01	3.78	x	x	Decreased
*11	0.02	0	0	0	0.08	0	x	x	None
*13	0.1	0.2	0.14	0	0.08	0.1	x	x	None
*14	0.18	0	0	0.89	0	0	x	x	Decreased
*15	0.06	0	0	0	0.23	0	x	x	None
*17	5.25	0.2	0.86	0	19.29	0	x	x	Decreased

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

*21	0.1	0	0	0.5	0	0	x	x	None
*22	0.06	0.3	0	0	0	0	x	x	Unknown
*28	0.12	0.5	0.14	0	0	0	x	x	Unknown
*29	2.64	0	0.29	0	9.83	0	x	x	Decreased
*31	0.12	0.2	0.58	0	0	0	x	x	None
*33	0.18	0.6	0.29	0	0	0.1	x	x	Normal
*34	0.02	0	0	0	0.08	0			Normal
*35	1.48	4.77	2.45	0	0.23	0.61	x	x	Normal
*36	0.24	0	0	0.3	0.68	0			None
*39	0.08	0	0.14	0	0.08	0.2		x	Normal
*40	0.24	0	0	0	0.91	0	x	x	None
*41	6.15	9.05	6.05	3.77	1.59	11.86	x	x	Decreased
*43	0.5	0.1	0	0	1.06	1.02	x	x	Unknown
*45	0.88	0	0.29	0	3.18	0	x	x	Normal
*46	0.14	0	0.14	0	0.45	0	x	x	Normal
*49	0.1	0	0	0.5	0	0	x		Decreased
*52	0.02	0	0	0.1	0	0	x	x	Unknown
*56	0.02	0	0	0	0.08	0	x	x	None
*59	0.06	0.2	0.14	0	0	0	x	x	Decreased
*71	0.12	0	0	0.6	0	0	x	x	Unknown
*82	0.06	0	0.43	0	0	0	x	x	Unknown
*83	0.04	0.1	0	0	0	0.1			Unknown
*84	0.02	0	0	0	0.08	0			Decreased
*86	0.44	0	0	0	0	2.25			Unknown
*99	0.04	0	0	0	0	0.2	x	x	None
*106	0.32	0	0.14	0	1.13	0	x	x	Unknown
*108	0.06	0.3	0	0	0	0		x	Unknown
*111	0.16	0	0	0	0	0.82	x	x	Unknown
*112	0.04	0	0	0	0	0.2	x	x	Unknown
*113	0.16	0	0	0	0	0.82	x	x	Unknown
*1x2	0.5	0.5	1.15	0.1	0.45	0.51	x	x	Increased

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

*1x3	0.02	0	0	0	0.08	0			Increased
*2x2	1.14	1.49	0.58	0.6	2.12	0.41	x	x	Increased
*2x3	0.04	0.1	0	0	0.08	0			Increased
*4x2	0.84	0.3	0.14	0	2.87	0	x	x	None
*4x3	0.04	0	0	0	0.15	0			None
*9x2	0.02	0.1	0	0	0	0			Normal
*10x2	0.06	0	0	0.3	0	0	x	x	Decreased
*17x2	0.02	0	0	0	0.08	0		x	Normal
*29x2	0.1	0	0	0	0.38	0			Normal
*35x2	0.02	0	0.14	0	0	0			Increased
*43x2	0.04	0	0.14	0	0.08	0			Unknown
*45x3	0.02	0	0	0	0.08	0			Increased
*10+*36	7.19	0	0.14	34.42	0.08	1.12	x	x	Decreased
*36+*36	0.04	0	0	0.2	0	0			None
*4+*68	1.94	5.57	2.45	0	0.23	2.15	x	x	None
*4+*68+*68	0.08	0.1	0.43	0	0	0			None
*10+*36+*36	0.32	0	0	1.59	0	0	x	x	Decreased
*10+*36+*36+*36	0.02	0	0	0.1	0	0	x	x	Decreased
*2+*13	0.06	0.2	0.14	0	0	0	x	x	Normal
*4+*4N	0.14	0.7	0	0	0	0		x	None
*1+*90	0.02	0	0	0.1	0	0	x	x	Unknown
*10+*36+*36+*83	0.02	0	0	0.1	0	0			Decreased
Unknown	2.36	0.6	3.75	3.37	2.27	2.25			
% haplotypes overlapping the validation set	96.1	98.0	95.4	96.0	95.9	95.2			

Availability of data and materials

Cyrius can be downloaded from: https://github.com/Illumina/Cyrius

The 1kGP data can be downloaded from https://www.ncbi.nlm.nih.gov/bioproject/PRJEB31736/.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

WGS data for 70 GeT-RM samples can be downloaded from: https://www.ebi.ac.uk/ena/data/view/PRJEB19931.

For NA12878, NA24385, and NA24631, the PacBio Sequel II data is available in SRA under PRJNA540705, PRJNA529679, and PRJNA540706, and the Illumina data is available in ENA under PRJEB35491. For the remaining 5 samples with PacBio truth, the PacBio Sequel II data is available from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/HGSVC2/working/.

Acknowledgements

We thank the New York Genome Center (supported by National Human Genome Research Institute grant 3UM1HG008901–03S1), and the Coriell Institute for Medical Research for generating and releasing the 1kGP WGS data.

References

1. Evans WE, Relling MV. Moving towards individualized medicine with pharmacogenomics. *Nature*. 2004;429(6990):464-468. doi:10.1038/nature02626
2. Zhou S-F. Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I. *Clin Pharmacokinet*. 2009;48(11):689-723. doi:10.2165/11318030-000000000-00000
3. Gaedigk A, Ingelman-Sundberg M, Miller NA, et al. The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database. *Clin Pharmacol Ther*. 2018;103(3):399-401. doi:10.1002/cpt.910
4. Gaedigk A, Simon SD, Pearce RE, Bradford LD, Kennedy MJ, Leeder JS. The *CYP2D6* activity score: translating genotype information into a qualitative measure of phenotype. *Clin Pharmacol Ther*. 2008;83(2):234-242. doi:10.1038/sj.clpt.6100406
5. Gaedigk A, Sangkuhl K, Whirl-Carrillo M, Klein T, Leeder JS. Prediction of *CYP2D6* phenotype from genotype across world populations. *Genet Med*. 2017;19(1):69-76. doi:10.1038/gim.2016.80
6. Nofziger C, Paulmichl M. Accurately genotyping *CYP2D6*: not for the faint of heart. *Pharmacogenomics*. 2018;19(13):999-1002. doi:10.2217/pgs-2018-0105
7. Yang Y, Botton MR, Scott ER, Scott SA. Sequencing the *CYP2D6* gene: from variant allele discovery to clinical pharmacogenetic testing. *Pharmacogenomics*. 2017;18(7):673-685. doi:10.2217/pgs-2017-0033
8. Gaedigk A. Complexities of *CYP2D6* gene analysis and interpretation. *Int Rev Psychiatry Abingdon Engl*. 2013;25(5):534-553. doi:10.3109/09540261.2013.825581
9. Pratt VM, Everts RE, Aggarwal P, et al. Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project. *J Mol Diagn JMD*. 2016;18(1):109-123. doi:10.1016/j.jmoldx.2015.08.005
10. Kalman LV, Agúndez J, Appell ML, et al. Pharmacogenetic allele nomenclature: International workgroup recommendations for test result reporting. *Clin Pharmacol Ther*. 2016;99(2):172-185. doi:10.1002/cpt.280
11. Ashley EA. The Precision Medicine Initiative: A New National Effort. *JAMA*. 2015;313(21):2119-2120. doi:10.1001/jama.2015.3595
12. The Genome of the Netherlands Consortium, Francioli LC, Menelaou A, et al. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014;46(8):818-825. doi:10.1038/ng.3021

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

13. Turnbull C, Scott RH, Thomas E, et al. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*. 2018;361:k1687. doi:10.1136/bmj.k1687
14. Numanagić I, Malikić S, Pratt VM, Skaar TC, Flockhart DA, Sahinalp SC. Cypiripi: exact genotyping of *CYP2D6* using high-throughput sequencing data. *Bioinformatics*. 2015;31(12):i27-i34. doi:10.1093/bioinformatics/btv232
15. Twist GP, Gaedigk A, Miller NA, et al. Constellation: a tool for rapid, automated phenotype assignment of a highly polymorphic pharmacogene, *CYP2D6*, from whole-genome sequences. *NPJ Genomic Med*. 2016;1:15007. doi:10.1038/npjgenmed.2015.7
16. Numanagić I, Malikić S, Ford M, et al. Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. *Nat Commun*. 2018;9(1):1-11. doi:10.1038/s41467-018-03273-1
17. Lee S, Wheeler MM, Patterson K, et al. Stargazer: a software tool for calling star alleles from next-generation sequencing data using *CYP2D6* as a model. *Genet Med*. 2019;21(2):361. doi:10.1038/s41436-018-0054-0
18. Lee S-B, Wheeler MM, Thummel KE, Nickerson DA. Calling Star Alleles With Stargazer in 28 Pharmacogenes With Whole Genome Sequences. *Clin Pharmacol Ther*. 2019;106(6):1328-1337. doi:10.1002/cpt.1552
19. Gordon AS, Fulton RS, Qin X, Mardis ER, Nickerson DA, Scherer S. PGRNseq: a targeted capture sequencing panel for pharmacogenetic research and implementation. *Pharmacogenet Genomics*. 2016;26(4):161-168. doi:10.1097/FPC.0000000000000202
20. Gaedigk A, Turner A, Everts RE, et al. Characterization of Reference Materials for Genetic Testing of *CYP2D6* Alleles: A GeT-RM Collaborative Project. *J Mol Diagn JMD*. Published online August 9, 2019. doi:10.1016/j.jmoldx.2019.06.007
21. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393
22. Chen X, Sanchis-Juan A, French CE, et al. Spinal muscular atrophy diagnosis and carrier screening from genome sequencing data. *Genet Med*. Published online February 18, 2020:1-9. doi:10.1038/s41436-020-0754-0
23. Spies N, Zook JM, Salit M, Sidow A. svviz: a read viewer for validating structural variants. *Bioinformatics*. 2015;31(24):3994-3996. doi:10.1093/bioinformatics/btv478
24. Raczy C, Petrovski R, Saunders CT, et al. Isaac: ultra-fast whole-genome secondary analysis on Illumina sequencing platforms. *Bioinformatics*. 2013;29(16):2041-2043. doi:10.1093/bioinformatics/btt314
25. Whirl-Carrillo M, McDonagh EM, Hebert JM, et al. Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther*. 2012;92(4):414-417. doi:10.1038/clpt.2012.96
26. Hosono N, Kato M, Kiyotani K, et al. *CYP2D6* genotyping for functional-gene dosage analysis by allele copy number detection. *Clin Chem*. 2009;55(8):1546-1554. doi:10.1373/clinchem.2009.123620
27. Kiyotani K, Shimizu M, Kumai T, Kamataki T, Kobayashi S, Yamazaki H. Limited effects of frequent *CYP2D6**36-*10 tandem duplication allele on in vivo dextromethorphan metabolism in a Japanese population. *Eur J Clin Pharmacol*. 2010;66(10):1065-1068. doi:10.1007/s00228-010-0876-4
28. Kim J, Lee S-Y, Lee K-A. Copy number variation and gene rearrangements in *CYP2D6* genotyping using multiplex ligation-dependent probe amplification in Koreans. *Pharmacogenomics*. 2012;13(8):963-973. doi:10.2217/pgs.12.58
29. Qiao W, Martis S, Mendiratta G, et al. Integrated *CYP2D6* interrogation for multiethnic copy number and tandem allele detection. *Pharmacogenomics*. 2019;20(1):9-20. doi:10.2217/pgs-2018-0135
30. Del Tredici AL, Malhotra A, Dedek M, et al. Frequency of *CYP2D6* Alleles Including Structural Variants in the United States. *Front Pharmacol*. 2018;9. doi:10.3389/fphar.2018.00305

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.05.077966; this version posted May 6, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

31. Chan W, Li MS, Sundaram SK, Tomlinson B, Cheung PY, Tzang CH. *CYP2D6* allele frequencies, copy number variants, and tandems in the population of Hong Kong. *J Clin Lab Anal*. 2019;33(1):e22634. doi:10.1002/jcla.22634
32. Black JL, Walker DL, O'Kane DJ, Harmandayan M. Frequency of Undetected *CYP2D6* Hybrid Genes in Clinical Samples: Impact on Phenotype Prediction. *Drug Metab Dispos*. 2012;40(1):111-119. doi:10.1124/dmd.111.040832
33. Zhou Y, Ingelman-Sundberg M, Lauschke VM. Worldwide Distribution of Cytochrome P450 Alleles: A Meta-analysis of Population-scale Sequencing Projects. *Clin Pharmacol Ther*. 2017;102(4):688-700. doi:10.1002/cpt.690