



Published in final edited form as:

Genet Med. 2019 February ; 21(2): 361–372. doi:10.1038/s41436-018-0054-0.

Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model

Seung-been Lee, B.S.¹, Marsha M. Wheeler, Ph.D.¹, Karynne Patterson, B.S.¹, Sean McGee, Ph.D.¹, Rachel Dalton, Pharm.D.³, Erica L. Woodahl, Ph.D.³, Andrea Gaedigk, Ph.D.⁴, Kenneth E. Thummel, Ph.D.⁵, and Deborah A. Nickerson, Ph.D.^{1,2}

¹Department of Genome Sciences, School of Medicine, University of Washington, Seattle, WA, USA.

²Brotman Baty Institute for Precision Medicine, Seattle, WA, USA.

³Department of Biomedical and Pharmaceutical Sciences, University of Montana, Missoula, MT, USA.

⁴Division of Clinical Pharmacology, Toxicology and Therapeutic Innovation, Children's Mercy Kansas City, Kansas City, MO, USA.

⁵Department of Pharmaceutics, School of Pharmacy, University of Washington, Seattle, WA, USA.

Abstract

Purpose—Genotyping *CYP2D6* is important for precision drug therapy because it metabolizes approximately 25% of drugs and its activity varies considerably among individuals. Genotype analysis of *CYP2D6* is challenging due to its highly polymorphic nature. Over 100 haplotypes (star alleles) have been defined for *CYP2D6*, some involving a gene conversion with its nearby non-functional but highly homologous paralog *CYP2D7*. We present Stargazer, a new bioinformatics tool that uses next-generation sequencing (NGS) data to call star alleles for *CYP2D6* (<https://stargazer.gs.washington.edu/stargazerweb/>). Stargazer is currently being extended for other pharmacogenes.

Methods—Stargazer identifies star alleles from NGS data by detecting single nucleotide variants, insertion-deletion variants, and structural variants. Stargazer detects structural variation including gene deletions, duplications, and conversions by calculating paralog-specific copy number from read depth.

Results—We applied Stargazer to NGS data of 32 ethnically diverse HapMap trios that were genotyped by TaqMan assays, long-range PCR, quantitative multiplex PCR, High Resolution Melt analysis, and/or Sanger sequencing. *CYP2D6* genotyping by Stargazer was 99.0% concordant with

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use:http://www.nature.com/authors/editorial_policies/license.html#terms

CORRESPONDING AUTHOR, Author: Deborah A. Nickerson, Ph.D., debnick@uw.edu, Phone: (206) 685-7387, Fax: (206) 221-6498, Address: William H. Foege Hall, 3720 15th Ave NE, Seattle, WA 98195.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

data obtained by these methods and showed 28.1% of the samples had structural variation including *CYP2D6/CYP2D7* hybrids.

Conclusion—Accurate genotyping of pharmacogenes with NGS and subsequent allele calling with Stargazer will aid the implementation of precision drug therapy.

Keywords

star alleles; structural variation; *CYP2D6* genotyping; pharmacogenomics; next-generation sequencing

INTRODUCTION

Many cytochrome P450 enzymes play a role in pharmacological responses by contributing to the metabolism of numerous drugs. Among these, cytochrome P450 2D6 (CYP2D6) is considered one of the most important because it contributes to the metabolism of about 25% of drugs.¹ Drugs metabolized by CYP2D6 include opioids, chemotherapeutic agents, antidepressants, and antipsychotics, among others.²

The activity of CYP2D6 varies considerably between individuals due to the high level of polymorphisms observed in the *CYP2D6* gene. There are more than 100 haplotypes defined for *CYP2D6* by the Pharmacogene Variation Consortium.³ These are called star alleles (e.g., *CYP2D6*1*, **2*, etc.) and are characterized by single nucleotide variants (SNVs), insertion-deletion variants (indels), structural variants (SVs), or a combination of these. Among these are fully functional, decreased function, and non-functional alleles, which provide a wide spectrum of CYP2D6 enzymatic activity ranging from ultrarapid to poor metabolism. Different ethnic groups have distinct frequencies of star alleles and metabolic phenotypes.⁴ Further studies are warranted, however, for individuals of African or Asian ancestry because these populations are underrepresented in the estimation of *CYP2D6* genetic diversity.⁵

Drug therapy without preemptive knowledge of a patient's CYP2D6 phenotype status can lead to severe adverse reactions or a loss of efficacy due to inappropriate drug choice and/or dosing. For example, codeine is one of the most common and widely used opioids whose analgesic effect is elicited by CYP2D6 through the formation of morphine. Patients who are CYP2D6 poor metabolizers exhibit very low plasma concentrations of morphine following codeine administration, which can complicate their pain management because the affinity of morphine to the μ -opioid receptor is 200-fold stronger compared to that of codeine.⁶ Conversely, a patient can experience life-threatening morphine intoxication after receiving a small dose of codeine if they are a CYP2D6 ultrarapid metabolizer.⁷ Similarly, a breastfed infant died from morphine poisoning because the mother was prescribed a normal dose of codeine for childbirth-related pain. In this case, the mother was a CYP2D6 ultrarapid metabolizer passing a toxic amount of morphine to her newborn through her breast milk.⁸

For these reasons, there is considerable interest in genotyping *CYP2D6*. Genotype analysis of *CYP2D6* is complex, however, because a large fraction of its existing variation cannot be accurately assessed with a single approach. SVs in *CYP2D6* such as gene deletions, duplications, and conversions are particularly challenging to detect due to high sequence

homology (> 95%) with a non-functional paralog, *CYP2D7*, located upstream of *CYP2D6*.⁹ Therefore, *CYP2D6* is prone to genotype misclassification and incorrect phenotype prediction.¹⁰ In laboratory settings, several orthogonal genotyping methods such as TaqMan assays, long-range PCR, quantitative multiplex PCR, High Resolution Melt analysis, and Sanger sequencing are employed to call star alleles. Many of these methods are time-consuming and heavily biased toward the detection of known variants, however. In clinical settings, due to practical limitations, only a handful of major star alleles, if any, are tested. Hence, a new approach for genotyping *CYP2D6* is needed that is more robust and capable of higher throughput.

In this study, we developed Stargazer, a new bioinformatics tool for calling star alleles in *CYP2D6* from next-generation sequencing (NGS) data. NGS is a powerful platform for variant detection because of its high-throughput data generation, comprehensive genotyping capabilities, and ever-decreasing cost. Additionally, NGS does not require prior knowledge about the variants of interest and can uncover novel functional variants, which is not possible for many of the aforementioned genotyping methods. Furthermore, its cost-effectiveness can be increased for variant discovery by applying custom capture panels. To assess the accuracy of Stargazer, we applied it to the NGS data of 32 ethnically diverse HapMap trios. We report a correlation of 99.0% between *CYP2D6* genotype calls determined with Stargazer and orthogonal methods. We are now extending Stargazer to call star alleles for other clinically important pharmacogenes. Accurate diplotype calls from NGS data using Stargazer provide a promising approach for precision medicine to maximize drug efficacy and minimize toxicity for individual patients. Stargazer is publicly accessible through <https://stargazer.gs.washington.edu/stargazerweb/>.

MATERIALS AND METHODS

Samples

We built Stargazer using NGS data from a total of 32 ethnically diverse trios. These trios were selected from the International HapMap Project, and they are comprised of 13 European, five Yoruban, four African American, three Han Chinese, three Mexican American, two Peruvian, and two Puerto Rican families (Table 1). These trios were originally sequenced to assess the performance of PGRNseq, a recently developed custom capture panel of key pharmacogenes including *CYP2D6*.¹¹ These trios were specifically chosen for this study because they are a genetically diverse set of samples, in which we would likely encounter a wide range of *CYP2D6* variants, including SVs, to test Stargazer's genotyping abilities and limitations. In addition, these trios allow for the analysis of Mendelian inheritance patterns to further the validation of Stargazer's star allele calls. These trios were also previously genotyped for *CYP2D6* by a variety of orthogonal methods (see below), allowing us to assess the accuracy of Stargazer's diplotype calls.

Orthogonal genotyping methods

HapMap trios were genotyped for *CYP2D6* according to procedures described elsewhere.^{12–15} Briefly, SNVs/indels were detected with TaqMan assays. Gene deletions, duplications, and multiplications were assessed by long-range PCR and quantitative multiplex PCR.

CYP2D6/CYP2D7 hybrids were identified with quantitative multiplex PCR, High Resolution Melt analysis, and/or Sanger sequencing.

Custom capture panel and NGS

HapMap trios were sequenced twice, once with PGRNseq v1.1 and once with PGRNseq v2.0 to a mean coverage of ~400X and ~160X, respectively. Both sequencing runs were performed with Illumina HiSeq 2500 machines using 100bp paired-end reads. Three samples failed during one of the two sequencing runs: NA19835 in PGRNseq v1.1 and NA19686 and NA11834 in PGRNseq v2.0. Note that the probes designed to capture *CYP2D6* and *CYP2D7* were more specific and extensive in PGRNseq v2.0 compared to PGRNseq v1.1; however, both versions generated reads that mapped to all the exons, introns, UTRs, and promoters of *CYP2D6* and *CYP2D7* (Figure S1). Two samples, NA12878 and NA19238, were also previously whole genome sequenced to a mean coverage of ~30X with Illumina HiSeq X instruments using 150bp paired-end reads. These data were used to test Stargazer's generalizability to whole genome sequencing (WGS) data.

Input and output data of Stargazer

The Stargazer *CYP2D6* genotyping pipeline is outlined in Figure 1. The pipeline uses BAM files comprised of sequence reads aligned with BWA-MEM to human reference genome assembly GRCh37.¹⁶ BAM files are then used to generate a VCF file with GATK-HaplotypeCaller (v3.4),¹⁷ from which Stargazer extracts all SNVs/indels located within 3kb from either end of *CYP2D6*. More specifically, Stargazer stores the genomic position of each variant, the reference allele, the alternate allele(s), the genotype status (homozygous or heterozygous), and the allelic depth for each sample. Stargazer uses the variant information from the VCF file to call star alleles based on SNVs/indels.

BAM files are also used to calculate read depth for *CYP2D6* and *CYP2D7* with GATK-DepthOfCoverage (v3.4).¹⁷ For convenience, we will refer to this output as a target GDF (GATK-DepthOfCoverage format) file. Since the high homology between *CYP2D6* and *CYP2D7* can cause reads to align to erroneous or multiple locations, only uniquely mapping reads with a mapping quality ≥ 20 are counted. Similarly, a control GDF file is produced from a user-chosen locus which serves as a read depth normalization factor. Stargazer computes paralog-specific copy number using read depth from the target and control GDF files in order to detect SVs.

In the initial development of Stargazer, three genes, *VDR*, *RYR1*, and *EGFR*, were evaluated as control loci. These genes are covered by PGRNseq and are 63kb, 154kb, and 188kb in size, respectively. They are also reported to exhibit low rates of whole gene deletion and/or duplication according to the Database of Genomic Variants.¹⁸ All three genes produced the same copy number results for *CYP2D6* and *CYP2D7*. The analyses shown in the results section were all performed using *RYR1* as the control locus.

The output data of Stargazer include each sample's *CYP2D6* diplotype, predicted phenotype, and plots to visually inspect copy number for *CYP2D6* and *CYP2D7* (Figure 2). Note that when calling diploypes, Stargazer only considers those variants that are currently used by the Pharmacogene Variation Consortium. Stargazer also returns all detected SNVs/

indels including those that are novel as well as those that are known but are not currently used to define any star allele. As follow up, these variants can be functionally annotated using variant annotation tools such as SeattleSeq Annotation (<http://snp.gs.washington.edu/SeattleSeqAnnotation>).

Prediction of star alleles

From a VCF file, Stargazer uses Beagle (v4.1)¹⁹ to haplotype phase heterozygous variants for *CYP2D6* with over 2,500 reference samples from the 1000 Genomes Project. Stargazer then matches phased haplotypes to star alleles with a translation table built from publicly available data (<https://www.pharmvar.org>). The table contains information on more than 90 star alleles and 185 SNVs/indels including variant positions and nucleotide changes in relation to the reference *CYP2D6*1* allele and human reference genome assembly GRCh37.

Detection of SVs

From a target GDF file, Stargazer converts read depth for *CYP2D6* and *CYP2D7* to copy number by performing intra- and inter-sample normalizations. The former accounts for individual variation in sequencing efficiency using read depth from a control GDF file, while the latter considers heterogeneity in coverage across all samples. Stargazer then automates detection of SVs with changepoint (v2.2.2), an R package that approximates one or more points at which the statistical properties of a sequence of observations change.²⁰ Here, the sequence is DNA, the observation is per-base copy number, and the statistical property is the mean copy number. If there is a significant shift in the mean copy number (e.g., from two to one), the algorithm returns the change point location and the two mean values (e.g., two and one).

Identification of diplotypes

For samples without SVs, Stargazer determines *CYP2D6* diplotypes by combining the star allele used to assign each phased haplotype. For samples with a whole gene deletion, the affected haplotype is assigned the *CYP2D6*5* deletion allele, which is then combined with the star allele assigned to the other haplotype to form a diplotype. For samples with a whole gene duplication, the affected haplotype will be assigned “x2” (e.g., *CYP2D6*1x2*, **2x2*, etc.) because it has two gene copies of *CYP2D6*. For samples with more complex SVs such as *CYP2D6/CYP2D7* hybrids, individual algorithms have been developed to determine diplotypes. More details on identification of diplotypes are discussed in the context of HapMap trios in the results section.

Assignment of predicted phenotypes

There are four *CYP2D6* metabolizer classes: poor, intermediate, normal, and ultrarapid. To predict these phenotypes, Stargazer first translates *CYP2D6* diplotypes into a standard unit of enzyme activity known as an activity score.²¹ The fully functional reference *CYP2D6*1* allele is assigned a value of 1, decreased function alleles such as *CYP2D6*10* and **17* receive a value of 0.5, and non-functional alleles including *CYP2D6*4* and **5* have a value of 0. The sum of values assigned to both alleles constitutes the activity score of a diplotype. Consequently, subjects with *CYP2D6*1/*1*, **1/*4* and **4/*5* diplotypes will have an activity

score of 2, 1 and 0, respectively. These activity scores are used to predict the four metabolizer classes as following: poor, 0; intermediate, 0.5; normal, 1 to 2; ultrarapid, greater than 2.

RESULTS

Identification of diplotypes by Stargazer for HapMap trios

We used Stargazer (v1.0.0) to call *CYP2D6* diplotypes for 32 HapMap trios sequenced with PGRNseq (Table 1). Data from PGRNseq v1.1 and PGRNseq v2.0 served as technical validation and produced the same diplotype calls for all samples. Moreover, all diplotype calls were inherited in a predictable manner, as exemplified in Figure 3. Diplotypes were identified using the following algorithms:

For samples without SVs, diplotypes were determined by combining the star allele assigned to each of the two haplotypes. For example, phasing algorithms estimated from subject NA12805 two haplotypes, of which one matched *CYP2D6**2 and the other *4 to form a *CYP2D6**2/*4 diplotype (Figure 2A).

For samples with a whole gene deletion, diplotypes were determined such that one haplotype contained the *CYP2D6**5 deletion allele while the other haplotype was assigned a star allele based on detected SNVs/indels. For example, subject NA18508 had only one *CYP2D6* gene copy and all detected SNVs were hemizygous and matched *CYP2D6**2. Stargazer called this sample as having a *CYP2D6**2/*5 diplotype (Figure 2B).

For samples with a whole gene duplication, Stargazer resolved the identity of the extra *CYP2D6* gene copy in the affected haplotypes. For example, Stargazer detected three gene copies in subject NA19685 with a *CYP2D6**1/*2 diplotype (Figure 2C); the sample could tentatively have a duplication on the *CYP2D6**1 or *2 allele, or in other words could have a *CYP2D6**1x2/*2 or *1/*2x2 diplotype. Stargazer used the allelic depth ratios of the SNVs defining the *CYP2D6**2 allele to determine which allele carried the duplication. If the *CYP2D6**2 allele carried the extra copy, the sample would have a read ratio of 2:1 reads for the respective SNVs. Indeed, most samples that were heterozygous for these SNVs had read ratios close to 1, whereas NA19685 was a significant outlier with an average ratio of 2.4 from the PGRNseq 1.1 data. Stargazer called the sample as having a *CYP2D6**1/*2x2. The read ratio approach can also be used to distinguish between diplotypes having duplications on both alleles and those having multiple copies on one allele (e.g., *CYP2D6**1x2/*2x2 vs. *1/*2x3).

For samples with complex structural variation, diplotypes were called using individual algorithms. For example, *CYP2D6**68+*4 is a tandem duplication where the *CYP2D6**4 gene copy is defined by only one SNV while *CYP2D6**68 is a hybrid gene featuring a *CYP2D7* sequence from intron 1 onward and four *CYP2D6* SNVs before the breakpoint. Stargazer called one of the two haplotypes of HG01979 as having this tandem structure because all five SNVs were detected, they were haplotype phased together, and the conversion to *CYP2D7* was also observed (Figure 2D). The other haplotype was matched to *CYP2D6**2 and Stargazer called this sample as having a *CYP2D6**2/*68+*4 diplotype.

Similar approaches were employed to determine diplotypes involving other tandem duplications such as *CYP2D6**36+*10 (Figure 2E) or *78+*2 (Figure 2F) where the *CYP2D6**36 and *78 alleles each contain a gene conversion to *CYP2D7*.

For samples with more than one SV, Stargazer tested all possible pairwise combinations of SVs to determine diplotypes. More specifically, Stargazer first fits every combination of SVs against a sample's observed *CYP2D6* and *CYP2D7* copy number profiles and then selects the combination that produces the least deviance. For example, HG00463 and HG00465 (family ID SH021) carry both a gene duplication and a gene conversion on each of their chromosomes and their profiles are best explained by having a *CYP2D6**36+*10 tandem arrangement on each chromosome (Figure S2A–B). There was one additional sample with multiple SVs, HG01190, whose profile was best explained by a combination of a *CYP2D6**5 deletion and a *CYP2D6**68+*4 (Figure S2C).

Summary of genotype and phenotype calls by Stargazer for HapMap trios

Stargazer called a total of 20 unique haplotypes from the 32 HapMap trios. The frequencies of these haplotypes in the 64 unrelated parents are shown in Table S1. As expected, fully functional *CYP2D6**1 and *2 alleles had the highest frequencies (31.3% and 18.8%, respectively) amongst the parents, followed by the non-functional *CYP2D6**4 allele (8.6%). Stargazer also detected *CYP2D6**17 (7.0%), which is commonly found in subjects of African ancestry including African Americans, and *CYP2D6**10 (2.3%), which occurs most frequently in East Asians. Both *CYP2D6**17 and *10 are decreased function alleles. In addition, Stargazer identified many haplotypes with structural variation in the parents: 4.7% with a gene deletion (*CYP2D6**5), 11.7% with a gene duplication (*CYP2D6**2x2, *4x2, *4N+*4, *10x2, *36+*10, *68+*4, and *78+*2), and 8.6% with a gene conversion (*CYP2D6**4N, *36, *68, and *78). This translates to 9.4%, 21.9%, and 15.6% of the parents having at least one gene deletion, duplication, or conversion, respectively. Based on the diplotype calls, 4.7%, 10.9%, 82.8%, and 1.6% of the parents were predicted to be poor, intermediate, normal, and ultrarapid metabolizers, respectively.

Comparison between Stargazer and orthogonal genotyping methods

When the *CYP2D6* haplotype calls of Stargazer were compared to those determined with orthogonal methods, the concordance rate was 99.0% (190 out of 192 haplotypes; Table 1). The two discordant haplotypes were found in NA19200 and NA19202 from the Y045 trio. For these samples, the orthogonal methods called *CYP2D6**5/*76+*1 and *CYP2D6**1/*76+*1 diplotypes, respectively, while Stargazer called *CYP2D6**1/*5 and *CYP2D6**1/*1 diplotypes. The non-functional *CYP2D6**76 allele, a *CYP2D6*/*CYP2D7* hybrid, was identified using long-range PCR and Sanger sequencing. The allele is essentially a *CYP2D7* gene that has a *CYP2D6* downstream sequence with a switch region to *CYP2D6* past exon 9, and lacks a *CYP2D7*-specific sequence also referred to as 'spacer'.²² Since this allele has a *CYP2D6*-specific sequence it may produce positive results with some long-range PCR reactions that are deemed diagnostic for the presence of a gene duplication. Nonetheless, Stargazer did not detect any significant change in copy number either at the switch region or in the spacer, so the program did not call *CYP2D6**76 (Figure S3). Note that the phenotype prediction is the same whether the allele is detected or not.

Stargazer calls using WGS data

To assess the performance of Stargazer on WGS data, we evaluated two samples NA19238 and NA12878 that were sequenced with PGRNseq v1.1, PGRNseq v2.0, and WGS. Although NA12878 carried a *CYP2D6**68+*4 tandem duplication, Stargazer called the same diplotypes regardless of sequencing platform (Figure 4).

Testing Stargazer at various sequencing coverages

We applied Stargazer to simulated datasets generated by randomly downsampling sequence reads from PGRNseq v2.0 data (Table S2). When 15% of reads were used (corresponding to 23.7X coverage), 186 out of 188 haplotypes were correctly called; the two misclassified haplotypes carried a SV. The misclassifications did not affect the phenotype prediction, however. Based on these results, our recommendation is to use Stargazer for datasets with a mean read coverage greater than 20X.

SNVs/indels detected by NGS

From the PGRNseq v1.1 and PGRNseq v2.0 data, 142 SNVs/indels were detected at 138 loci within 3kb from either end of *CYP2D6*, 86 of which are not currently used to define star alleles (Table S3). Among those, according to SeattleSeq Annotation, five are missense mutations while the remaining ones are either synonymous or within the 3' or 5' UTR, downstream or upstream of the gene, or within an intron. We did not find any novel variants that are obviously detrimental to *CYP2D6* function such as nonsense, frameshift, or splice site mutations.

DISCUSSION

We developed Stargazer, a new software tool for calling star alleles in various polymorphic pharmacogenes from NGS data. When building Stargazer, we used *CYP2D6* as a model for detection and interpretation of SVs in the context of other observed SNVs/indels. We purposely chose *CYP2D6* as a starting point because it is one of the most complex genetic loci to genotype in the human genome. Two other programs, Cypiripi and Astrolabe, have been published to genotype *CYP2D6* from NGS data.^{23–24} Although both tools can reliably call simple diplotypes, they have difficulties with detection of complex SVs such as *CYP2D6/CYP2D7* hybrids. We show that Stargazer can reliably detect those hybrids from targeted or WGS data.

More specifically, we show Stargazer correctly genotyped *CYP2D6* for 32 ethnically diverse HapMap trios. These trios were previously validated by a variety of orthogonal methods and comparisons show Stargazer is 99.0% concordant to these methods. In the future, we will test additional verified samples in order to further validate Stargazer's performance. All diplotype calls by Stargazer were inherited according to expectations including population-specific star alleles such as *CYP2D6**10 and *17. Stargazer also produced the same diplotype calls for all samples from the two independent PGRNseq v1.1 and PGRNseq v2.0 datasets.

We plan to extend Stargazer to *CYP2A6*, another highly polymorphic pharmacogene displaying many SNVs/indels as well as SVs. *CYP2A6* metabolizes nicotine and sequence variation in *CYP2A6* has been linked to nicotine dependence and withdrawal symptoms upon smoking cessation.²⁵ Similar to *CYP2D6*, *CYP2A6* has several star alleles with a gene conversion to its nearby paralog *CYP2A7*.²⁶ We also plan to develop Stargazer for other CYP genes.

As larger genomic datasets become available, several aspects of Stargazer will improve. This includes the statistical estimation of phased haplotypes, primarily haplotypes based on rare variants. In the current version of Stargazer, we incorporated a large panel of reference samples from the 1000 Genomes Project. This approach performed well for our dataset, but we are aware that in further applications, rare variants may have frequencies that are too low to be phased reliably. To ameliorate this issue, we plan to merge multiple large reference panels to obtain additional haplotype information. Novel variants will require physical phasing backed by sequence reads. When short reads cannot provide adequate phasing, long-read sequencing from Oxford Nanopore Technologies or Pacific Biosciences can be used to generate reference haplotype information. Recently, both technologies have been successfully applied to sequence *CYP2D6*.^{27–28}

Certain features of Stargazer are specific for targeted sequencing such as PGRNseq. For example, for the purposes of normalization, Stargazer requires multiple samples to be analyzed at a time. This is because sequencing with custom capture typically yields uneven coverage across the genes of interest and Stargazer's copy number estimation is based on population statistics. If the sample size is too small or if a large fraction of samples shares the same type of structural variation, population statistics can be shifted dramatically, generating biased copy number data. However, this problem can be addressed by including reference samples with known copy number. For WGS data, where coverage is usually distributed more evenly, the inter-sample normalization may be skipped allowing Stargazer to analyze a single sample.

We reported five missense variants that are not currently used to define any star allele. However, interpretation of these variants is difficult without functional characterization. In fact, the same is true for many variants in existing star alleles (e.g., *CYP2D6**22 is defined by a nonsynonymous SNV in exon 9 with unknown effect). Therefore, there is clearly a need to more rigorously characterize the function of the rapidly increasing number of haplotypes to facilitate phenotype prediction. In the future, it is possible that data from deep mutational scanning for the *CYP2D6* enzyme could be incorporated into Stargazer to aid in the characterization of the functional consequences of all possible single mutations of this protein.²⁹

The HapMap trios used in this study consist of seven distinct ethnic groups and therefore represent a sampling of the global distribution of *CYP2D6* genotypes. Characterized by multiple genotyping platforms including NGS, these trios can serve as a reference resource for other *CYP2D6* genotyping projects.

There is a growing awareness of individual variation in drug response. For example, in March 2013, the Food and Drug Administration (FDA) cautioned against the use of codeine in children of any age to treat pain after surgery to remove the tonsils or adenoids.³⁰ Shortly after, a prospective study showed that children who were CYP2D6 ultrarapid metabolizers and taking codeine after those surgeries were at a higher risk for toxicity and death.³¹ In April 2017, the FDA issued the agency's strongest warning against codeine alerting that the medication should not be used to treat pain or cough in children younger than 12 years. While limiting the therapeutic use of codeine addresses the concern for patient safety, tailoring codeine or other drug treatments based on an individual's *CYP2D6* genotype could achieve the same goal. There may also be therapeutic settings where alternative treatments are not fully interchangeable and health outcomes could suffer from restrictions in drug choice. For example, national guidelines recommend codeine as a front-line drug for the treatment of pain in patients with sickle cell disease and many hematologists prefer codeine to other analgesics that have comparable efficacy but higher potential for abuse and physical dependence.³² With additional validation Stargazer may offer an alternative approach for optimizing treatment response in all patients.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors acknowledge the Pharmacogenomics Research Network for supporting the development of PGRNseq. This work was supported by NIH grants HL069757, GM092676, GM116691, GM115318, and GM115277, S10OD021553, and by the University of Washington's Graduate School Fund for Excellence and Innovation.

References

1. Zhou SF. Polymorphism of human cytochrome P450 2D6 and its clinical significance: Part I. *Clin Pharmacokinet.* 2009; 48(11):689–723. [PubMed: 19817501]
2. Crews KR, Gaedigk A, Dunnenberger HM, et al. Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450 2D6 genotype and codeine therapy: 2014 update. *Clin Pharmacol Ther.* 2014; 95(4):376–382. [PubMed: 24458010]
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. [PubMed: 29134625]
4. 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. [PubMed: 27388693]
5. LLerena A, Naranjo ME, Rodrigues-Soares F, Penas-LLedó EM, Fariñas H, Tarazona-Santos E. Interethnic variability of CYP2D6 alleles and of predicted and measured metabolic phenotypes across world populations. *Expert Opin Drug Metab Toxicol.* 2014; 10(11):1569–1583. [PubMed: 25316321]
6. Kirchheiner J, Schmidt H, Tzvetkov M, et al. Pharmacokinetics of codeine and its metabolite morphine in ultra-rapid metabolizers due to CYP2D6 duplication. *Pharmacogenomics J.* 2007; 7(4): 257–265. [PubMed: 16819548]
7. Gasche Y, Daali Y, Fathi M, et al. Codeine intoxication associated with ultrarapid CYP2D6 metabolism. *N Engl J Med.* 2004; 351(27):2827–2831. [PubMed: 15625333]
8. Koren G, Cairns J, Chitayat D, Gaedigk A, Leeder SJ. Pharmacogenetics of morphine poisoning in a breastfed neonate of a codeine-prescribed mother. *Lancet.* 2006; 368(9536):704. [PubMed: 16920476]

9. Gaedigk A. Complexities of CYP2D6 gene analysis and interpretation. *Int Rev Psychiatry*. 2013; 25(5):534–553. [PubMed: 24151800]
10. Naranjo ME, de Andrés F, Delgado A, Cobaleda J, Peñas-Lledó EM, Llerena A. High frequency of CYP2D6 ultrarapid metabolizers in Spain: controversy about their misclassification in worldwide population studies. *Pharmacogenomics J*. 2016; 16(5):485–490. [PubMed: 27272044]
11. 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. [PubMed: 26736087]
12. Gaedigk A, Bradford LD, Alander SW, Leeder JS. CYP2D6*36 gene arrangements within the cyp2d6 locus: association of CYP2D6*36 with poor metabolizer status. *Drug Metab Dispos*. 2006; 34(4):563–569. [PubMed: 16415111]
13. Gaedigk A, Ndjountché L, Divakaran K, et al. Cytochrome P4502D6 (CYP2D6) gene locus heterogeneity: characterization of gene duplication events. *Clin Pharmacol Ther*. 2007; 81(2):242–251. [PubMed: 17259947]
14. Gaedigk A, Fuhr U, Johnson C, Bérard LA, Bradford D, Leeder JS. CYP2D7-2D6 hybrid tandems: identification of novel CYP2D6 duplication arrangements and implications for phenotype prediction. *Pharmacogenomics*. 2010; 11(1):43–53. [PubMed: 20017671]
15. Gaedigk A, Twist GP, Leeder JS. CYP2D6, SULT1A1 and UGT2B17 copy number variation: quantitative detection by multiplex PCR. *Pharmacogenomics*. 2012; 13(1):91–111. [PubMed: 22111604]
16. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25(14):1754–1760. [PubMed: 19451168]
17. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010; 20(9):1297–1303. [PubMed: 20644199]
18. MacDonald JR, Ziman R, Yuen RK, Feuk L, Scherer SW. The Database of Genomic Variants: a curated collection of structural variation in the human genome. *Nucleic Acids Res*. 2014; 42(Database issue):D986–992. [PubMed: 24174537]
19. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet*. 2007; 81(5):1084–1097. [PubMed: 17924348]
20. Killick R, Eckley IA. changepoint: An R Package for Changepoint Analysis. *J Stat Softw*. 2014; 58(3):1–19.
21. 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. [PubMed: 17971818]
22. Gaedigk A, Jaime LK, Bertino JS, et al. Identification of Novel CYP2D7-2D6 Hybrids: Non-Functional and Functional Variants. *Front Pharmacol*. 2010; 1:121. [PubMed: 21833166]
23. Numanagi I, Maliki S, Pratt VM, Skaar TC, Flockhart DA, Sahinalp SC. Cypiripi: exact genotyping of CYP2D6 using high-throughput sequencing data. *Bioinformatics*. 2015; 31(12):i27–34. [PubMed: 26072492]
24. 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 Genom Med*. 2016; 1:15007. [PubMed: 29263805]
25. Kubota T, Nakajima-Taniguchi C, Fukuda T, et al. CYP2A6 polymorphisms are associated with nicotine dependence and influence withdrawal symptoms in smoking cessation. *Pharmacogenomics J*. 2006; 6(2):115–119. [PubMed: 16402086]
26. Fukami T, Nakajima M, Sakai H, McLeod HL, Yokoi T. CYP2A7 polymorphic alleles confound the genotyping of CYP2A6*4A allele. *Pharmacogenomics J*. 2006; 6(6):401–412. [PubMed: 16636685]
27. Ammar R, Paton TA, Torti D, Shlien A, Bader GD. Long read nanopore sequencing for detection of HLA and CYP2D6 variants and haplotypes. *F1000Res*. 2015; 4:17. [PubMed: 25901276]
28. Qiao W, Yang Y, Sebra R, et al. Long-Read Single Molecule Real-Time Full Gene Sequencing of Cytochrome P450-2D6. *Hum Mutat*. 2016; 37(3):315–323. [PubMed: 26602992]

29. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. *Nat Methods*. 2014; 11(8):801–807. [PubMed: 25075907]
30. Kuehn BM. FDA: No codeine after tonsillectomy for children. *JAMA*. 2013; 309(11):1100. [PubMed: 23512037]
31. Prows CA, Zhang X, Huth MM, et al. Codeine-related adverse drug reactions in children following tonsillectomy: a prospective study. *Laryngoscope*. 2014; 124(5):1242–1250. [PubMed: 24122716]
32. Gammal RS, Crews KR, Haidar CE, et al. Pharmacogenetics for Safe Codeine Use in Sickle Cell Disease. *Pediatrics*. 2016; 138(1)

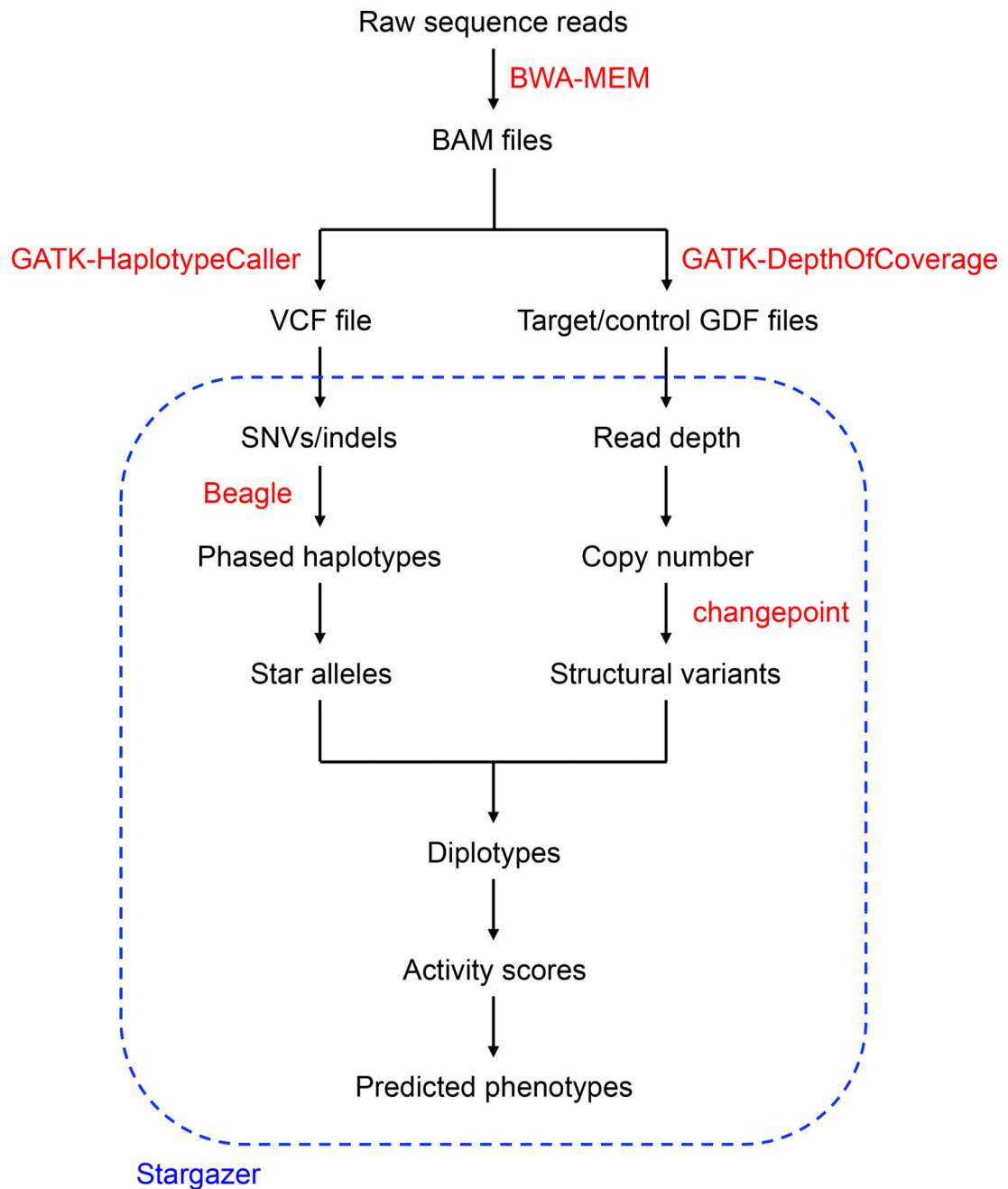


Figure 1.

A schematic diagram of the Stargazer *CYP2D6* genotyping pipeline. Stargazer takes as input a VCF file, a target GDF file, and a control GDF file. Stargazer uses the variant information from the VCF file to call star alleles based on SNVs/indels. Using the target and control GDF files, Stargazer converts read depth to copy number for detection of structural variation. The output data of Stargazer include each sample's *CYP2D6* diplotype and plots to visually inspect copy number for *CYP2D6* and *CYP2D7*. Based on called *CYP2D6* diploypes, the program outputs predicted phenotypes as well. Several external software tools, shown in red, are used within and outside of Stargazer.

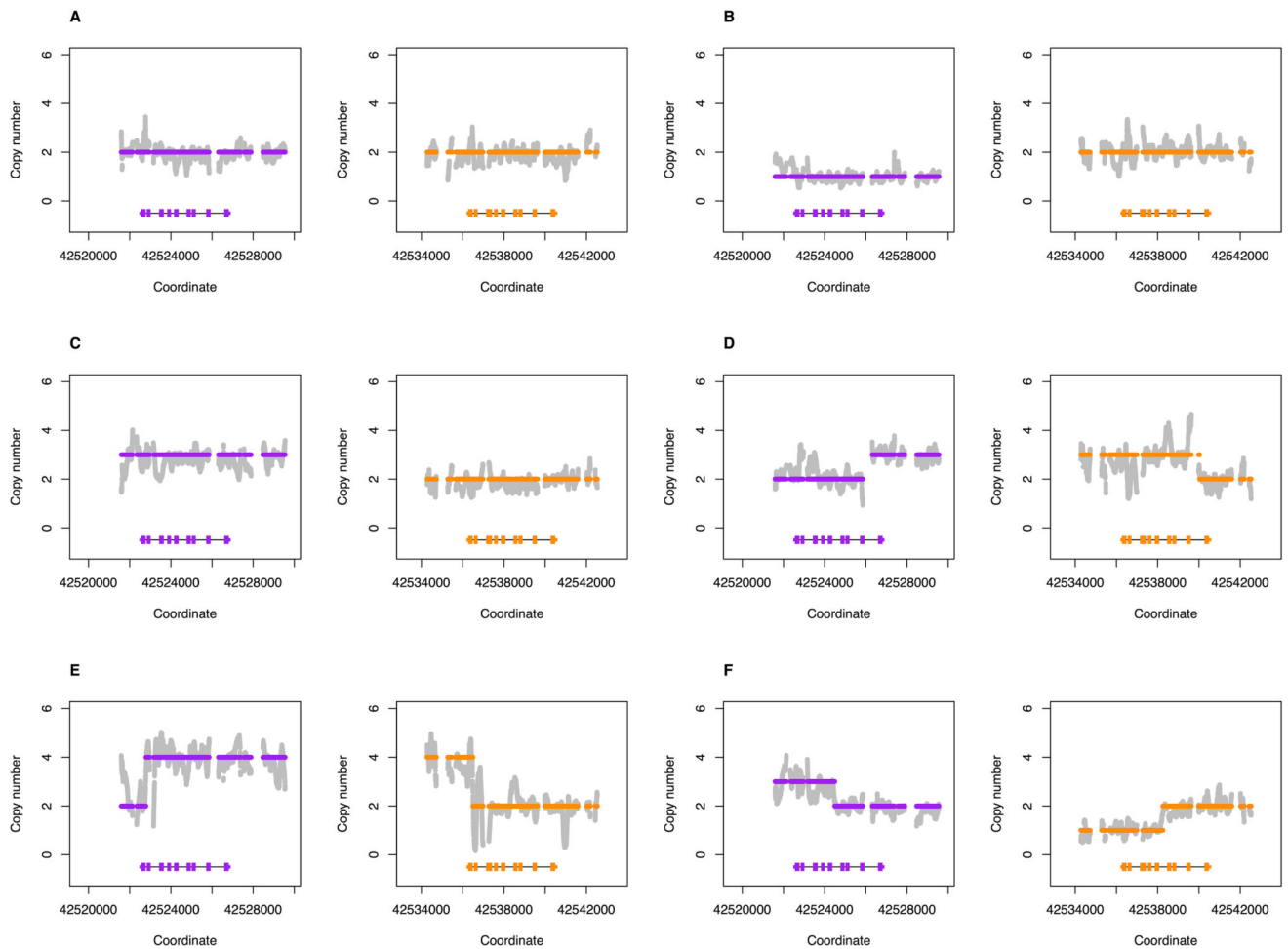


Figure 2.

Examples of structural variation detected by Stargazer in HapMap trios. Grey dots are copy number calculated from read depth. Dots colored purple and orange are the mean copy number for *CYP2D6* and *CYP2D7*, respectively, determined by the changepoint algorithm. Each panel contains scaled *CYP2D6* and *CYP2D7* gene models, in which the exons and introns are depicted with boxes and lines, respectively. All panels were generated from PGRNseq v2.0 data. (A) shows European sample NA12805 that has a *CYP2D6**2/*4 diplotype without structural variation; this sample was included for comparison. (B) shows a gene deletion in Yoruban sample NA18508 with a *CYP2D6**2/*5 diplotype. (C) shows a gene duplication in Mexican American sample NA19685 with a *CYP2D6**1/*2x2 diplotype. (D) shows a complex structural variation involving a gene duplication and a gene conversion in Peruvian sample HG01979 genotyped as *CYP2D6**2/*68+*4. (E) shows a complex structural variation involving multiple gene duplications and gene conversions in Han Chinese sample HG00465 genotyped as *CYP2D6**36+*10/*36+*10. (F) shows a complex structural variation involving a gene conversion in Mexican American sample NA19790 genotyped as *CYP2D6**1/*78+*2.

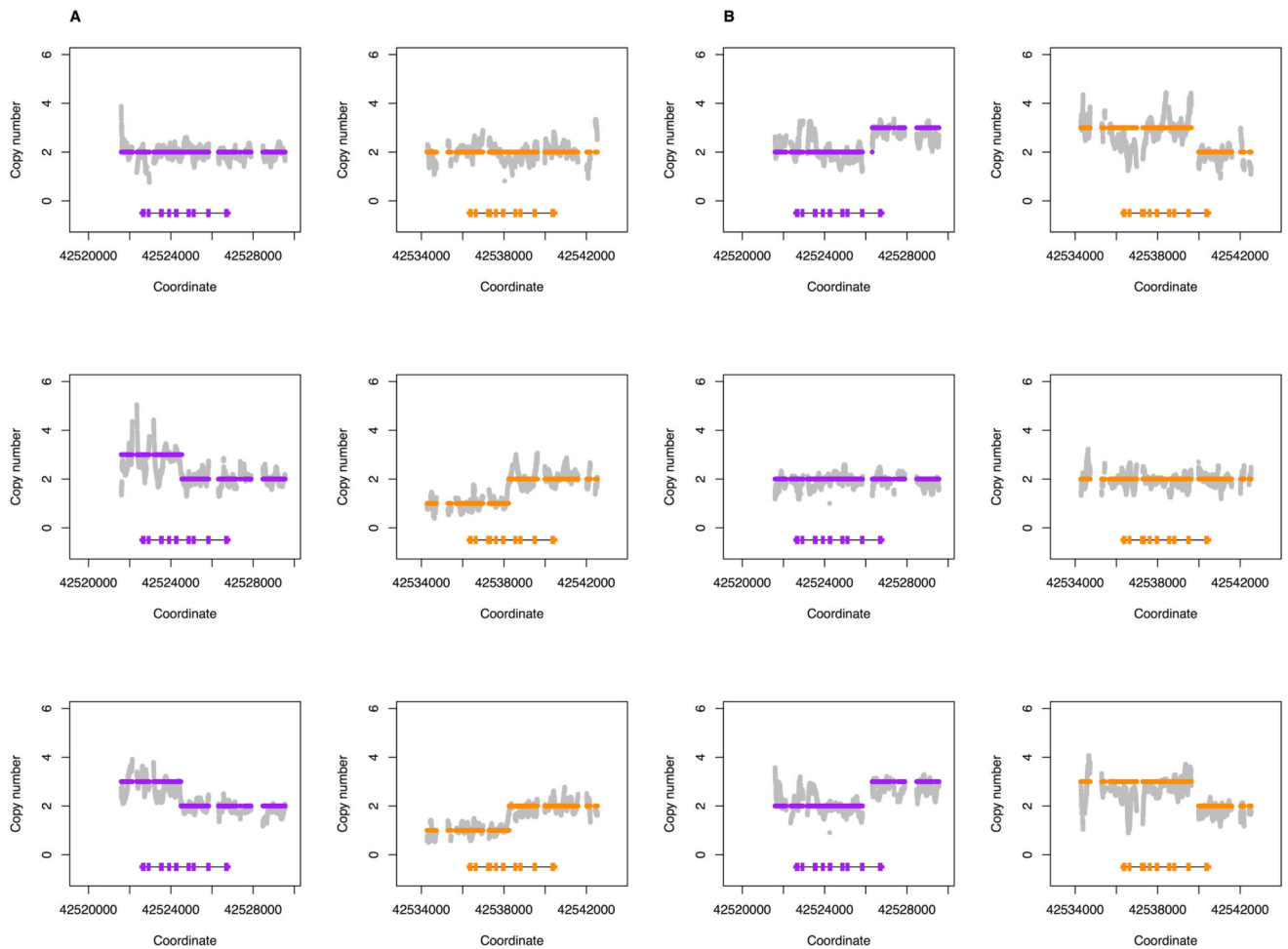


Figure 3. Segregation of complex structural variations detected by Stargazer in two HapMap trios. For each trio, data from the father is shown in the top panel, data from the mother is shown in the middle panel, and data from the child is shown in the bottom panel. Grey dots are copy number calculated from read depth. Dots colored purple and orange are the mean copy number for *CYP2D6* and *CYP2D7*, respectively, determined by the changepoint algorithm. Each panel contains scaled *CYP2D6* and *CYP2D7* gene models, in which the exons and introns are depicted with boxes and lines, respectively. All panels were generated from PGRNseq v2.0 data. (A) shows segregation of *CYP2D6**78+*2 in the Mexican American M037 family. (B) shows segregation of *CYP2D6**68+*4 in the European 1463 family.

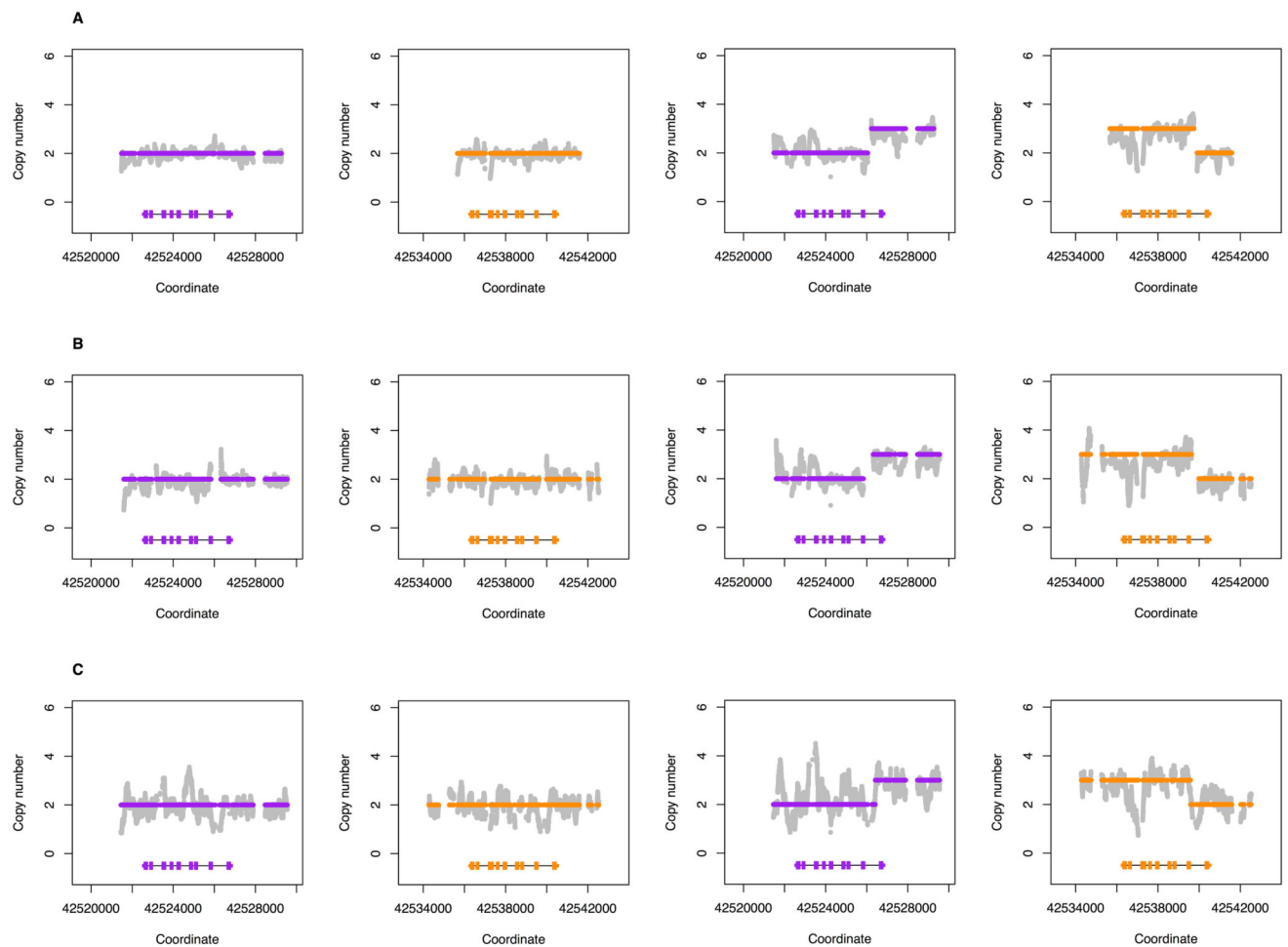


Figure 4.

Comparison of custom capture and whole genome sequencing. Grey dots are copy number calculated from read depth. Dots colored purple and orange are the mean copy number for *CYP2D6* and *CYP2D7*, respectively, determined by the changepoint algorithm. Each panel contains scaled *CYP2D6* and *CYP2D7* gene models, in which the exons and introns are depicted with boxes and lines, respectively. Two subjects NA19238 and NA12878 were sequenced with (A) PGRNseq v1.1 at ~400X coverage with 100bp paired-end reads, (B) PGRNseq v2.0 at ~160X coverage with 100bp paired-end reads, and (C) whole genome sequencing at ~30X coverage with 150bp paired-end reads. In all three cases, Stargazer called the correct diplotypes *CYP2D6**1/*17 and *3/*68+*4, respectively.

Table 1

Comparison of *CYP2D6* genotype calls for 32 HapMap trios between orthogonal methods and Stargazer using PGRNseq v1.1 or v2.0 data. Three samples failed during one of the two sequencing runs ('-'). When the *CYP2D6* haplotype calls of Stargazer were compared to those determined with the orthogonal methods, the concordance rate was 99.0% (190 out of 192 haplotypes). The two discordant haplotypes were found in samples NA19200 and NA19202 of the Y045 trio ('[]'). Predicted phenotypes were assigned based on the activity score and Clinical Pharmacogenetics Implementation Consortium guidelines.

#	Sample	Ethnicity	Family	Relation	Orthogonal Methods	PGRNseq v1.1	PGRNseq v2.0	Activity Score	Phenotype
1	NA12801	European	1454	Father	*4/*6	*4/*6	*4/*6	0	Poor
2	NA12802	European	1454	Mother	*2/*41	*2/*41	*2/*41	1.5	Normal
3	NA12805	European	1454	Child	*2/*4	*2/*4	*2/*4	1	Normal
4	NA12891	European	1463	Father	*41/*68+*4	*41/*68+*4	*41/*68+*4	0.5	Intermediate
5	NA12892	European	1463	Mother	*2/*3	*2/*3	*2/*3	1	Normal
6	NA12878	European	1463	Child	*3/*68+*4	*3/*68+*4	*3/*68+*4	0	Poor
7	NA19834	African American	2424	Father	*2/*2	*2/*2	*2/*2	2	Normal
8	NA19835	African American	2424	Mother	*1/*2	-	*1/*2	2	Normal
9	NA19836	African American	2424	Child	*1/*2	*1/*2	*1/*2	2	Normal
10	NA19239	Yoruban	Y117	Father	*15/*17	*15/*17	*15/*17	0.5	Intermediate
11	NA19238	Yoruban	Y117	Mother	*1/*17	*1/*17	*1/*17	1.5	Normal
12	NA19240	Yoruban	Y117	Child	*15/*17	*15/*17	*15/*17	0.5	Intermediate
13	NA12750	European	1444	Father	*2/*2	*2/*2	*2/*2	2	Normal
14	NA12751	European	1444	Mother	*1/*2	*1/*2	*1/*2	2	Normal
15	NA12740	European	1444	Child	*1/*2	*1/*2	*1/*2	2	Normal
16	NA19685	Mexican American	M011	Father	*1/*2x2	*1/*2x2	*1/*2x2	3	Ultrarapid
17	NA19684	Mexican American	M011	Mother	*1/*4	*1/*4	*1/*4	1	Normal
18	NA19686	Mexican American	M011	Child	*1/*1	*1/*1	-	2	Normal
19	HG00421	Han Chinese	SH007	Father	*2/*10x2	*2/*10x2	*2/*10x2	2	Normal
20	HG00422	Han Chinese	SH007	Mother	*2/*10	*2/*10	*2/*10	1.5	Normal
21	HG00423	Han Chinese	SH007	Child	*10/*10x2	*10/*10x2	*10/*10x2	1.5	Normal
22	HG01979	Peruvian	PEL027	Father	*2/*68+*4	*2/*68+*4	*2/*68+*4	1	Normal
23	HG01980	Peruvian	PEL027	Mother	*1/*2	*1/*2	*1/*2	2	Normal
24	HG01981	Peruvian	PEL027	Child	*1/*2	*1/*2	*1/*2	2	Normal

#	Sample	Ethnicity	Family	Relation	Orthogonal Methods	PGRNseq v1.1	PGRNseq v2.0	Activity Score	Phenotype
25	NA12003	European	1420	Father	*4/*35	*4/*35	*4/*35	1	Normal
26	NA12004	European	1420	Mother	*2/*41	*2/*41	*2/*41	1.5	Normal
27	NA10838	European	1420	Child	*2/*4	*2/*4	*2/*4	1	Normal
28	NA12155	European	1408	Father	*1/*5	*1/*5	*1/*5	1	Normal
29	NA12156	European	1408	Mother	*1/*4	*1/*4	*1/*4	1	Normal
30	NA10831	European	1408	Child	*4/*5	*4/*5	*4/*5	0	Poor
31	NA19128	Yoruban	Y077	Father	*17/*17	*17/*17	*17/*17	1	Normal
32	NA19127	Yoruban	Y077	Mother	*2/*17	*2/*17	*2/*17	1.5	Normal
33	NA19129	Yoruban	Y077	Child	*17/*17	*17/*17	*17/*17	1	Normal
34	NA19700	African American	2367	Father	*4/*29	*4/*29	*4/*29	0.5	Intermediate
35	NA19701	African American	2367	Mother	*1/*17	*1/*17	*1/*17	1.5	Normal
36	NA19702	African American	2367	Child	*4/*17	*4/*17	*4/*17	0.5	Intermediate
37	NA19771	Mexican American	M031	Father	*2/*4	*2/*4	*2/*4	1	Normal
38	NA19770	Mexican American	M031	Mother	*1/*2	*1/*2	*1/*2	2	Normal
39	NA19772	Mexican American	M031	Child	*2/*4	*2/*4	*2/*4	1	Normal
40	HG01060	Puerto Rican	PR14	Father	*1/*41	*1/*41	*1/*41	1.5	Normal
41	HG01061	Puerto Rican	PR14	Mother	*1/*4	*1/*4	*1/*4	1	Normal
42	HG01062	Puerto Rican	PR14	Child	*1/*4	*1/*4	*1/*4	1	Normal
43	NA10860	European	1362	Father	*1/*4N+4	*1/*4N+4	*1/*4N+4	1	Normal
44	NA10861	European	1362	Mother	*4/*35	*4/*35	*4/*35	1	Normal
45	NA11984	European	1362	Child	*1/*35	*1/*35	*1/*35	2	Normal
46	HG02259	Peruvian	PEL042	Father	*1/*2	*1/*2	*1/*2	2	Normal
47	HG02260	Peruvian	PEL042	Mother	*1/*1	*1/*1	*1/*1	2	Normal
48	HG02261	Peruvian	PEL042	Child	*1/*2	*1/*2	*1/*2	2	Normal
49	NA07357	European	1345	Father	*1/*6	*1/*6	*1/*6	1	Normal
50	NA07345	European	1345	Mother	*1/*1	*1/*1	*1/*1	2	Normal
51	NA07348	European	1345	Child	*1/*6	*1/*6	*1/*6	1	Normal
52	NA06984	European	1328	Father	*4/*68+*4	*4/*68+*4	*4/*68+*4	0	Poor
53	NA06989	European	1328	Mother	*9/*9	*9/*9	*9/*9	1	Normal
54	NA12331	European	1328	Child	*4/*9	*4/*9	*4/*9	0.5	Intermediate
55	NA18507	Yoruban	Y009	Father	*2/*4x2	*2/*4x2	*2/*4x2	1	Normal

#	Sample	Ethnicity	Family	Relation	Orthogonal Methods	PGRNseq v1.1	PGRNseq v2.0	Activity Score	Phenotype
56	NA18508	Yoruban	Y009	Mother	*2/*5	*2/*5	*2/*5	1	Normal
57	NA18506	Yoruban	Y009	Child	*2/*5	*2/*5	*2/*5	1	Normal
58	NA19900	African American	2425	Father	*3/*29	*3/*29	*3/*29	0.5	Intermediate
59	NA19901	African American	2425	Mother	*1/*1	*1/*1	*1/*1	2	Normal
60	NA19902	African American	2425	Child	*1/*29	*1/*29	*1/*29	1.5	Normal
61	NA19789	Mexican American	M037	Father	*1/*1	*1/*1	*1/*1	2	Normal
62	NA19788	Mexican American	M037	Mother	*2/*78+*2	*2/*78+*2	*2/*78+*2	2	Normal
63	NA19790	Mexican American	M037	Child	*1/*78+*2	*1/*78+*2	*1/*78+*2	2	Normal
64	HG00463	Han Chinese	SH021	Father	*36+*10*36+*10	*36+*10*36+*10	*36+*10*36+*10	1	Normal
65	HG00464	Han Chinese	SH021	Mother	*1/*36+*10	*1/*36+*10	*1/*36+*10	1.5	Normal
66	HG00465	Han Chinese	SH021	Child	*36+*10*36+*10	*36+*10*36+*10	*36+*10*36+*10	1	Normal
67	HG00592	Han Chinese	SH057	Father	*1/*10	*1/*10	*1/*10	1.5	Normal
68	HG00593	Han Chinese	SH057	Mother	*2/*36+*10	*2/*36+*10	*2/*36+*10	1.5	Normal
69	HG00594	Han Chinese	SH057	Child	*1/*2	*1/*2	*1/*2	2	Normal
70	HG01190	Puerto Rican	PR40	Father	*5/*68+*4	*5/*68+*4	*5/*68+*4	0	Poor
71	HG01191	Puerto Rican	PR40	Mother	*2/*41	*2/*41	*2/*41	1.5	Normal
72	HG01192	Puerto Rican	PR40	Child	*5/*41	*5/*41	*5/*41	0.5	Intermediate
73	NA12342	European	1330	Father	*4/*41	*4/*41	*4/*41	0.5	Intermediate
74	NA12343	European	1330	Mother	*1/*5	*1/*5	*1/*5	1	Normal
75	NA12336	European	1330	Child	*5/*41	*5/*41	*5/*41	0.5	Intermediate
76	NA10853	European	1349	Father	*2/*41	*2/*41	*2/*41	1.5	Normal
77	NA10854	European	1349	Mother	*1/*4	*1/*4	*1/*4	1	Normal
78	NA11834	European	1349	Child	*2/*4	*2/*4	-	1	Normal
79	NA19200	Yoruban	Y045	Father	*5/*76+*11	[*1]/*5	[*1]/*5	1	Normal
80	NA19201	Yoruban	Y045	Mother	*1/*17	*1/*17	*1/*17	1.5	Normal
81	NA19202	Yoruban	Y045	Child	*1/*76+*11	[*1]/*1	[*1]/*1	2	Normal
82	NA18516	Yoruban	Y013	Father	*1/*17	*1/*17	*1/*17	1.5	Normal
83	NA18517	Yoruban	Y013	Mother	*5/*10	*5/*10	*5/*10	0.5	Intermediate
84	NA18515	Yoruban	Y013	Child	*1/*10	*1/*10	*1/*10	1.5	Normal
85	NA19818	African American	2418	Father	*1/*17	*1/*17	*1/*17	1.5	Normal
86	NA19819	African American	2418	Mother	*2/*4x2	*2/*4x2	*2/*4x2	1	Normal

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

#	Sample	Ethnicity	Family	Relation	Orthogonal Methods	PGRNseq v1.1	PGRNseq v2.0	Activity Score	Phenotype
87	NA19828	African American	2418	Child	*2/*17	*2/*17	*2/*17	1.5	Normal
88	NA12399	European	1354	Father	*1/*1	*1/*1	*1/*1	2	Normal
89	NA12400	European	1354	Mother	*1/*68+*4	*1/*68+*4	*1/*68+*4	1	Normal
90	NA12386	European	1354	Child	*1/*1	*1/*1	*1/*1	2	Normal
91	NA11891	European	1377	Father	*1/*1	*1/*1	*1/*1	2	Normal
92	NA11892	European	1377	Mother	*6/*41	*6/*41	*6/*41	0.5	Intermediate
93	NA10865	European	1377	Child	*1/*41	*1/*41	*1/*41	1.5	Normal
94	NA12272	European	1418	Father	*1/*1	*1/*1	*1/*1	2	Normal
95	NA12273	European	1418	Mother	*1/*1	*1/*1	*1/*1	2	Normal
96	NA10837	European	1418	Child	*1/*1	*1/*1	*1/*1	2	Normal