Star allele search: a pharmacogenetic annotation database and user-friendly search tool of publicly available 1000 Genomes Project biospecimens

N. Gharani[1,2], G. Calendo[1], D. Kusic[1], J. Madzo[1], L. Scheinfeldt[1]*
[1] Coriell Institute for Medical Research, 403 Haddon Ave, Camden, NJ 08103, USA
[2] Gharani Consulting Limited, 272 Regents Park Road, London, N3 3HN, UK
*Corresponding author

Abstract

Here we describe a new public pharmacogenetic (PGx) annotation database of a large (n=3202) and diverse biospecimen collection of 1000 Genomes Project cell lines and DNAs. The database is searchable with a user friendly, web-based tool (www.coriell.org/StarAllele/Search). This resource leverages existing whole genome sequencing data and PharmVar annotations to characterize *alleles for each biospecimen in the collection. This new tool is designed to facilitate in vitro functional characterization of *allele haplotypes and diplotypes as well as support clinical PGx assay development, validation, and implementation.

Keywords
Pharmacogenetic, pharmacogenomic, reference material

Background

Pharmacogenomics (PGx) holds the potential to improve medication management by increasing efficacy and by reducing toxicity [1-7]. Translating pharmacogenomic research into clinical care, however, requires a robust inter-disciplinary infrastructure [8, 9]. Characterizing the full range of functionally relevant human pharmacogenetic variation is limited by the documented underrepresentation of many communities living in the United States and around the world [10-16], and this effort would benefit from a large and diverse collection of publicly available and well-characterized cell lines. Such a resource would facilitate a more comprehensive understanding of pharmacogene variation and in vitro drug response [17-23]. Moreover, a well-characterized and diverse set of publicly available and renewable DNA samples would benefit the clinical communities that require positive and negative controls for assay development, validation, implementation and proficiency testing for robust PGx testing.

The Genetic Reference and Testing Materials Coordination Program (GeT-RM) has used a variety of clinical testing methods to characterize lymphoblastoid cell line (LCL) DNAs for 28 pharmacogenes [24], and more recently has incorporated next generation sequencing data for the characterization of *CYP2D6* [25] (n=179), as well as *CYP2C8*, *CYP2C9* and *CYP2C19* (n=137) [26]. Here we describe a complementary PGx annotation resource that includes a significantly larger set (n=3,202) of renewable and publicly available 1000 Genomes Project LCLs and DNAs available through the NHGRI Sample Repository for Human Genetic Research (https://catalog.coriell.org/1/NHGRI) and the NIGMS Human Genetic Cell Repository (https://catalog.coriell.org/1/NIGMS). This new annotation resource leverages 30x whole genome sequencing (WGS) data [27], is downloadable (**Table S1**) and may be searched with a user-friendly, web-based tool (www.coriell.org/StarAllele/Search).

Construction and content

Table 1 includes a summary of all of the publicly available 1000 Genomes Project biospecimens included in the star allele annotation database. The majority of the samples (n=3023) are available through the NHGRI Sample Repository for Human Genetic Research (https://catalog.coriell.org/1/NHGRI), and the collection of Utah Residents (CEPH) with Northern and Western European Ancestry biospecimens (n=179) are available through the NIGMS Human Genetic Cell Repository (https://catalog.coriell.org/1/NIGMS). **Table S1** includes each individual NHGRI Sample Repository for Human Genetic Research and NIGMS Human Genetic Cell Repository identifier for all of the 1000 Genomes Project biospecimens annotated in the star allele annotation database.

We leveraged existing publicly available 30x coverage WGS data generated from 3202 samples and phased by the New York Genome Center (NYGC) [27]. The detailed description of the data collection and analysis can be found in Byrska-Bishop et al. [27]. Briefly, 3,202 samples from the 1000 Genomes Project collection were selected for inclusion [27] in the WGS data collection (**Table 1**). The sample set includes 2,504 unrelated individuals as well as 698 relatives (that together complete 602 complete trios) [27], and the WGS data were collected with an Illumina NovaSeq 6000 System [27]. The raw WGS data were aligned to the GRCh38 reference genome, and variant calling was performed with GATK [27, 28]. The WGS variant information was additional phased into haplotypes; autosomal single nucleotide variants (SNVs) and insertion / deletions (INDELs) were statistically phased using SHAPEIT-duoHMM with pedigree-based correction [27, 29, 30].

We used the phased NYGC WGS VCF files [27] identified through the www.internationalgenome.org website (https://www.internationalgenome.org/data-portal/data-collection/30x-grch38), and accessed from the following publicly available FTP site (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/), which were last modified on 2022-11-14 08:33. This dataset was filtered prior to phasing according to the following criteria as described in the NYGC README file: 1) missing genotype rate >5%; 2) Hardy Weinberg test P-Value >1e-10; Mendelian error rate ≤5%; and 3) minor allele count ≥2. We did not perform any additional data post processing. We compared the variants in the phased VCF files against PGx annotations for 12 of the 13 pharmacogenes annotated in PharmVar [31-33] version 5.2.13 using ursaPGx [34]which implements Cyrius for *CYP2D6* calling using raw WGS BAM files [35]. The detailed description of the ursaPGx annotation can be found here [34]. Briefly, for each pharmacogene, the star allele defining variants according to PharmVar are extracted from the phased VCF file, and the annotation is assigned when all star allele defining variants are present for a given VCF haplotype. In cases where no complete match between the phased haplotype and the PharmVar star allele occurs, the haplotype is annotated as ambiguous.

Utility and discussion

Here we describe a new public PGx annotation database with a user friendly, web-based search tool of associated lymphoblastoid cell line and DNA biospecimens. This new resource complements existing databases generated by GeT-RM; while GeT-RM works directly with clinical laboratories to develop robust PGx annotated biospecimens designed to serve as reference materials for genetic testing, this effort is extremely involved and not easily scalable to larger collections of biospecimens. This new annotation database therefore offers a slightly less robust characterization of a significantly larger collection of diverse biospecimens (**Table 1**) to support PGx related research efforts and to serve as a starting point for clinical testing communities to identify potentially relevant reference materials for their testing needs.

To assess the quality and accuracy of the PGx annotation database, we compared overlapping samples that were already characterized by GeT-RM using next-generation sequencing data, which is available for *CYP2C8*, *CYP2C9*, *CYP2C19*, and *CYP2D6* [25, 26]. In total, we identified 87 overlapping samples between GeT-RM and the current annotation dataset [25, 26]. We found 100%, 99% and 97% concordance, respectively between our annotation and the GeT-RM NGS consensus annotation for *CYP2C8, CYP2C19, and CYP2C9* [26], and we found 94% concordance between our annotation and the GeT-RM NGS consensus annotation for *CYP2D6* [25].

Our *CYP2C19* comparison identified a discrepancy for a single sample (NA19122). The GeT-RM NGS consensus is *2/*35 [26], while our annotation is *2|*Amb. We note that as described above, our approach requires a complete match between a given phased haplotype and all of the PharmVar defining variants for a given star allele. For NA19122, the first haplotype included all of the variants required to annotate *2 (non-reference alleles for rs12769205, rs4244285, rs3758581), consistent with GeT-RM [26]; however, the second haplotype in our phased VCF file includes both variants required to annotate *35 (non-reference alleles for rs12769205, rs3758581) as well as a non-reference allele at rs17882687, which in our approach precludes it from an unambiguous call of *35 or *15.

We identified discordant *CYP2C9* star allele annotations for three samples. Our approach annotated two samples (NA19143, NA19213) as *1|*1 while the GeT-RM NGS consensus is *1/*6 [26]. This discrepancy is due to the limitation of the WGS phased VCF file we used which unfortunately does not contain rs9332131, which is the single base deletion that defines *6. We annotated the third discordant sample (HG01190) to be *61|*1, whereas the GeT-RM NGS consensus is *2/*61 [26]. We believe this difference is due to differences in variant calling and phasing approaches. In the phased VCF we used, this sample is heterozygous for both variants required to annotate *61 (rs1799853 and rs202201137), and both of these variants occur on the first haplotype of the sample. Here we also note that while the consensus annotation is *2/*61, a minority of the groups participating in the study annotated this sample as *1/*61 [26].

For *CYP2D6*, we identified a perfect match for 82 overlapping samples, with an additional two samples (NA07000, NA19143) concordant between the ursaPGx implementation of Cyrius and the tentative GeT-RM assignment designated with parentheses [25]. For NA18519, the ursaPGx implementation of Cyrius annotated *106/*29, while GeT-RM annotated *1/*29. As far as we can tell from the detail included in Table S2 of the publication's supplementary materials [25], the *106 defining variant (rs28371733) was not included in the NA18519 annotation assessment; *106 was not detected by the assays used for the full set of included samples (n=179), including genotyping, PharmacoScan, iPLEX V1.1, CYP2D6 V1.1, a custom panel, and VeriDose, but rather sequencing (Sanger, NGS or SMRT) appears to have been used for a subset of 50 samples that did not include NA18519. The ursaPGx implementation of Cyrius was not able to resolve a single diplotype for NA19908, but rather annotated two possible diplotypes (*1/*46;*43/*45) due to short read phase uncertainty [35]. GeT-RM utilized long-range PCR to resolve the phase uncertainty and correct diplotype for this sample [25]. Similarly, the ursaPGx implementation of Cyrius was not able to fully resolve the diplotype for NA18565 using short read WGS data beyond *36/*36+*10, while GeT-RM was able to fully resolve the diplotype to *10/*36x2 (one *10 allele and a second allele with two copies of *36).

In addition to the GeT-RM annotation benchmarking, we compared PGx annotation using the newest NYGC 30x WGS dataset available [27] against the older Phase 3 10x coverage WGS dataset available for a subset of 2504 unrelated individuals [36] for *CYP2C9* and *CYP2C19*. Several *allele-defining variants were present in the Phase 3 10x dataset but absent from the NYGC phased 30x VCF files (**Table S2**). In particular *CYP2C9* *6 (rs9332131, A deletion), *7 (rs67807361, A allele), *16 (rs72558192, G allele), *33 (rs200183364, A allele), *36 (rs114071557, G allele), *45 (rs199523631, T allele), *63 (rs141489852, A allele), *68 (rs542577750 A allele), and *73 (rs17847037, T allele) and *CYP2C19* *16 (rs192154563, T allele), *24 (rs118203757, A allele), and *30 (rs145328984, T allele).

In total, star allele search includes 956 diplotypes across 14 pharmacogenes (**Table 2, Table S1**), excluding diplotypes with one or two ambiguous (i.e., Amb) allele calls. Each unique diplotype and associated diplotype frequency in the database is detailed in **Table S3,** and each unique *allele haplotype and associated haplotype allele frequency in the database is detailed in **Table S4**. To determine the contribution of the larger sample set included in the database, we identified 3, 3, 7, and 10 new *alleles, respectively in the dataset relative to GeT-RM [25, 26] for *CYP2C19* (*11, *22, *34), *CYP2C8* (*6, *11, *14), *CYP2C9* (*12, *13, *14, *29, *31, *44, *66), and *CYP2D6* (*27, *32, *34, *49, *84, *86, *117, *121, *125, *139) (**Figure 1, Table S5**). We performed a similar comparison for unique pairs of *alleles (diplotype combinations). We chose to conservatively exclude ambiguous calls, copy number variants and complex *CYP2D6* structural variants and identified 12, 17, 23, and 132 new diplotypes, respectively for *CYP2C8*, *CYP2C19*, *CYP2C9*, and *CYP2D6* (**Figure 1, Table S5**).

This new star allele annotated biospecimen database is of use for a wide range of applications. For example, researchers interested in functionally characterizing *alleles of interest can use the resource to choose LCLs with the most relevant diplotype combinations; researchers interested in developing new PGx assays can use the resource to benchmark performance; and clinical laboratories can use the resource to minimize the number of positive and negative control DNAs needed for a given PGx test.

We have additionally developed Star Allele Search (**Figure 2**), which is a web-based search tool of the new PGx biospecimen annotation database to facilitate these types of research and clinical applications. In addition to this new database and search tool, users can choose to search existing public data one variant at a time, up to one hundred variants at a time, or by gene (https://www.coriell.org/SNPSearch/WGS; [27]). Users can also search gene expression data collected from a subset of the *allele annotated LCLs (n=462) (http://omicdata.coriell.org/geuv-expression-browser/; [37]).

All of these genomic data search tools are designed to complement each other to ensure researchers have a simple way to search a large collection of biospecimen genetic, genomic, and transcriptomic profiles with a web-based interface that does not require bioinformatic skill or experience. For example, a researcher interested in developing a *CYP2C19* assay could first view, sort, filter and/or download a .CSV file of all of the *CYP2C19* variants included in the WGS dataset with a single HUGO symbol search (https://www.coriell.org/SNPSearch/WGS) to confirm the variants of interest are present in the data; then view, sort, filter and/or download a .CSV of the annotated *CYP2C19* *alleles for the entire sample set with Star Allele Search (www.coriell.org/StarAllele/Search) to identify the biospecimens with the relevant diplotypes; and if an alternate annotation scheme is needed (i.e., not PharmVar), the researcher can view, sort, filter and/or download a .CSV of up to 100 individual *CYP2C19* variants at a time to investigate any alternate combination of variants needed for the alternative annotation scheme (https://www.coriell.org/SNPSearch/WGS).

It is important to note all of the limitations of our approach. Our database annotations are based on short read, 30x coverage WGS and computational phasing [27]. Any error in variant calling or missing variation as well as any error in phasing in the input VCF will propagate into annotation errors. In addition, any error or missing variation in the input BAM files used for *CYP2D6* annotation will similarly produce annotation errors. While this is the most robust dataset available for our sample set at present, we anticipate that as long-read sequencing becomes more affordable and more accessible, that phase uncertainty (particularly for rare variants) will significantly go down and structural variation resolution will significantly improve. We also employed PharmVar annotation for our database and chose a strict matching requirement for each *allele annotation. This choice resulted in several ambiguous biospecimen calls in cases where one or both phased haplotypes were not an exact match to any PharmVar defined *allele.

Conclusion

We have developed a public resource of PGx annotation for a large (n=3202) and diverse set of 1000 Genomes Project LCLs and DNAs that are available for general research use. This new resource includes a database of star allele annotation for each biospecimen and an accompanying web-based search tool (www.coriell.org/StarAllele/Search). This new tool is especially relevant to researchers interested in in vitro functional characterization of *alleles as well as for use in support of clinical PGx assay development, validation, and implementation.

Declarations

Ethics approval and consent to participate

All human data used in this study is publicly available through the 1000 Genomes Project [27], and all associated biospecimens belong to the NIGMS Human Genetic Cell Repository or the NHGRI Sample Repository for Human Genetic Research. All NHGRI Sample Repository for Human Genetic Research biospecimens have been consented for general research use and associated public genomic data sharing. All NIGMS Human Genetic Cell Repository biospecimens included in the 1000 Genomes Project and thereby the current study have been consented for general research use and associated public genomic data sharing.

Consent for publication

Not applicable

Availability of data and materials

The detailed database content is available in **Table S1** and is available through a web-based search tool (www.coriell.org/StarAllele/Search). The majority of associated 1000 Genomes Project biospecimens (LCLs and DNA) are available through the NHGRI Sample Repository for Human Genetic Research (https://catalog.coriell.org/1/NHGRI), and the collection of Utah Residents (CEPH) with Northern and Western European Ancestry biospecimens are available through the NIGMS Human Genetic Cell Repository (https://catalog.coriell.org/1/NIGMS).

Competing interests

The authors declare that they have no competing interests

Funding

This study was funded by NHGRI 5U24HG008736 to LS.

Author's Contributions

LS designed and implemented the project and contributed to writing and editing the manuscript; NG designed the project and contributed to writing and editing the manuscript. DK, JM, and GC contributed to the design of the project and contributed to writing and editing the manuscript.

Acknowledgments

We thank Coriell's IT team, and John Witherspoon in particular, for supporting the implementation of Star Allele Search.

References

1. Zhang G, Zhang Y, Ling Y, Jia J: **Web resources for pharmacogenomics**. *Genomics Proteomics Bioinformatics* 2015, **13**(1):51-54.
2. Bank PCD, Swen JJ, Guchelaar HJ: **Implementation of Pharmacogenomics in Everyday Clinical Settings**. *Adv Pharmacol* 2018, **83**:219-246.
3. Dunnenberger HM, Crews KR, Hoffman JM, Caudle KE, Broeckel U, Howard SC, Hunkler RJ, Klein TE, Evans WE, Relling MV: **Preemptive clinical pharmacogenetics implementation: current programs in five US medical centers**. *Annu Rev Pharmacol Toxicol* 2015, **55**:89-106.
4. Gharani N, Keller MA, Stack CB, Hodges LM, Schmidlen TJ, Lynch DE, Gordon ES, Christman MF: **The Coriell personalized medicine collaborative pharmacogenomics appraisal, evidence scoring and interpretation system**. *Genome Med* 2013, **5**(10):93.
5. Relling MV, Krauss RM, Roden DM, Klein TE, Fowler DM, Terada N, Lin L, Riel-Mehan M, Do TP, Kubo M *et al*: **New Pharmacogenomics Research Network: An Open Community Catalyzing Research and Translation in Precision Medicine**. *Clin Pharmacol Ther* 2017, **102**(6):897-902.
6. Relling MV, Evans WE: **Pharmacogenomics in the clinic**. *Nature* 2015, **526**(7573):343-350.
7. Bush WS, Crosslin DR, Owusu-Obeng A, Wallace J, Almoguera B, Basford MA, Bielinski SJ, Carrell DS, Connolly JJ, Crawford D *et al*: **Genetic variation among 82 pharmacogenes: The PGRNseq data from the eMERGE network**. *Clin Pharmacol Ther* 2016, **100**(2):160-169.
8. Schmidlen T, Sturm AC, Scheinfeldt LB: **Pharmacogenomic (PGx) Counseling: Exploring Participant Questions about PGx Test Results**. *J Pers Med* 2020, **10**(2).
9. Herr TM, Peterson JF, Rasmussen LV, Caraballo PJ, Peissig PL, Starren JB: **Pharmacogenomic clinical decision support design and multi-site process outcomes analysis in the eMERGE Network**. *J Am Med Inform Assoc* 2019, **26**(2):143-148.
10. Scheinfeldt LB, Brangan A, Kusic DM, Kumar S, Gharani N: **Common Treatment, Common Variant: Evolutionary Prediction of Functional Pharmacogenomic Variants**. *J Pers Med* 2021, **11**(2).
11. Bentley AR, Callier S, Rotimi CN: **Diversity and inclusion in genomic research: why the uneven progress?** *J Community Genet* 2017, **8**(4):255-266.
12. Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE: **Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations**. *Am J Hum Genet* 2017, **100**(4):635-649.
13. Popejoy AB, Fullerton SM: **Genomics is failing on diversity**. *Nature* 2016, **538**(7624):161-164.
14. Scheinfeldt LB, Tishkoff SA: **Recent human adaptation: genomic approaches, interpretation and insights**. *Nat Rev Genet* 2013, **14**(10):692-702.
15. Wojcik GL, Graff M, Nishimura KK, Tao R, Haessler J, Gignoux CR, Highland HM, Patel YM, Sorokin EP, Avery CL *et al*: **Genetic analyses of diverse populations improves discovery for complex traits**. *Nature* 2019, **570**(7762):514-518.
16. Sirugo G, Williams SM, Tishkoff SA: **The Missing Diversity in Human Genetic Studies**. *Cell* 2019, **177**(1):26-31.

17. Wheeler HE, Dolan ME: **Lymphoblastoid cell lines in pharmacogenomic discovery and clinical translation**. *Pharmacogenomics* 2012, **13**(1):55-70.
18. Zhang W, Dolan ME: **Use of cell lines in the investigation of pharmacogenetic loci**. *Curr Pharm Des* 2009, **15**(32):3782-3795.
19. Choy E, Yelensky R, Bonakdar S, Plenge RM, Saxena R, De Jager PL, Shaw SY, Wolfish CS, Slavik JM, Cotsapas C *et al*: **Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines**. *PLoS Genet* 2008, **4**(11):e1000287.
20. Jack J, Rotroff D, Motsinger-Reif A: **Lymphoblastoid cell lines models of drug response: successes and lessons from this pharmacogenomic model**. *Curr Mol Med* 2014, **14**(7):833-840.
21. Morag A, Kirchheiner J, Rehavi M, Gurwitz D: **Human lymphoblastoid cell line panels: novel tools for assessing shared drug pathways**. *Pharmacogenomics* 2010, **11**(3):327-340.
22. Green AJ, Anchang B, Akhtari FS, Reif DM, Motsinger-Reif A: **Extending the lymphoblastoid cell line model for drug combination pharmacogenomics**. *Pharmacogenomics* 2021, **22**(9):543-551.
23. Shukla SJ, Dolan ME: **Use of CEPH and non-CEPH lymphoblast cell lines in pharmacogenetic studies**. *Pharmacogenomics* 2005, **6**(3):303-310.
24. Pratt VM, Everts RE, Aggarwal P, Beyer BN, Broeckel U, Epstein-Baak R, Hujsak P, Kornreich R, Liao J, Lorier R *et al*: **Characterization of 137 Genomic DNA Reference Materials for 28 Pharmacogenetic Genes: A GeT-RM Collaborative Project**. *J Mol Diagn* 2016, **18**(1):109-123.
25. Gaedigk A, Turner A, Everts RE, Scott SA, Aggarwal P, Broeckel U, McMillin GA, Melis R, Boone EC, Pratt VM *et al*: **Characterization of Reference Materials for Genetic Testing of CYP2D6 Alleles: A GeT-RM Collaborative Project**. *J Mol Diagn* 2019, **21**(6):1034-1052.
26. Gaedigk A, Boone EC, Scherer SE, Lee SB, Numanagic I, Sahinalp C, Smith JD, McGee S, Radhakrishnan A, Qin X *et al*: **CYP2C8, CYP2C9, and CYP2C19 Characterization Using Next-Generation Sequencing and Haplotype Analysis: A GeT-RM Collaborative Project**. *J Mol Diagn* 2022, **24**(4):337-350.
27. Byrska-Bishop M, Evani US, Zhao X, Basile AO, Abel HJ, Regier AA, Corvelo A, Clarke WE, Musunuri R, Nagulapalli K *et al*: **High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios**. *Cell* 2022, **185**(18):3426-3440 e3419.
28. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M *et al*: **The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data**. *Genome Res* 2010, **20**(9):1297-1303.
29. Delaneau O, Marchini J, Zagury JF: **A linear complexity phasing method for thousands of genomes**. *Nat Methods* 2011, **9**(2):179-181.
30. O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, Traglia M, Huang J, Huffman JE, Rudan I *et al*: **A general approach for haplotype phasing across the full spectrum of relatedness**. *PLoS Genet* 2014, **10**(4):e1004234.

31. Gaedigk A, Casey ST, Whirl-Carrillo M, Miller NA, Klein TE: **Pharmacogene Variation Consortium: A Global Resource and Repository for Pharmacogene Variation**. *Clin Pharmacol Ther* 2021, **110**(3):542-545.
32. Gaedigk A, Whirl-Carrillo M, Pratt VM, Miller NA, Klein TE: **PharmVar and the Landscape of Pharmacogenetic Resources**. *Clin Pharmacol Ther* 2020, **107**(1):43-46.
33. Gaedigk A, Ingelman-Sundberg M, Miller NA, Leeder JS, Whirl-Carrillo M, Klein TE, PharmVar Steering C: **The Pharmacogene Variation (PharmVar) Consortium: Incorporation of the Human Cytochrome P450 (CYP) Allele Nomenclature Database**. *Clin Pharmacol Ther* 2018, **103**(3):399-401.
34. Gennaro C, Dara K, Jozef M, Neda G, Laura S: **ursaPGx: a new R package to annotate pharmacogenetic star alleles using phased whole genome sequencing data**. *bioRxiv* 2023:2023.2007.2024.550372.
35. Chen X, Shen F, Gonzaludo N, Malhotra A, Rogert C, Taft RJ, Bentley DR, Eberle MA: **Cyrius: accurate CYP2D6 genotyping using whole-genome sequencing data**. *Pharmacogenomics J* 2021, **21**(2):251-261.
36. Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA *et al*: **A global reference for human genetic variation**. *Nature* 2015, **526**(7571):68-74.
37. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, Rivas MA, Gonzalez-Porta M, Kurbatova N, Griebel T, Ferreira PG *et al*: **Transcriptome and genome sequencing uncovers functional variation in humans**. *Nature* 2013, **501**(7468):506-511.

Table 1. List of Included Populations.

Population Descriptor	N samples with NYGC WGS data
African Caribbean in Barbados	116
Bengali in Bangladesh	131
British from England and Scotland	91
Chinese Dai in Xishuangbanna, China	93
Colombian in Medellin, Colombia	132
Esan in Nigeria	149
Finnish in Finland	99
Gambian in Western Division - Mandinka	178
Han Chinese South, China	163
Iberian Populations in Spain	157
Indian Telugu in the UK	107
Kinh in Ho Chi Minh City, Vietnam	122
Mende in Sierra Leone	99
Peruvian in Lima, Peru	122
Puerto Rican in Puerto Rico	139
Punjabi in Lahore, Pakistan	146
Sri Lankan Tamil in the UK	114
African Ancestry in Southwest USA	74
Gujarati Indians in Houston, Texas, USA	103
Han Chinese in Beijing, China	103
Japanese in Tokyo, Japan	104
Luhya in Webuye, Kenya	99
Mexican Ancestry in Los Angeles, California, USA	97
Toscani in Italia	107
Yoruba in Ibadan, Nigeria	178
Utah Residents (CEPH) with Northern and Western European Ancestry*	179
Total Samples	3202

* samples available through the NIGMS Human Genetic Cell Repository

Table 2. List of PharmVar annotated pharmacogenes, number of diplotypes and *alleles included in database.

Pharmacogene	N Unique Diplotypes	N Unique *Alleles
CYP2A13	21	6
CYP2A6	68	14
CYP2B6	60	13
CYP2C19	63	15
CYP2C8	34	12
CYP2C9	49	16
CYP2D6	316	316
CYP3A4	53	19
CYP3A5	16	4
CYP4F2	16	4
DPYD	111	26
NUDT15	11	5
SLCO1B1	138	23

Figure Legend

Figure 1. New *alleles and diplotypes identified in sample set (n=3,202) relative to GeT-RM. The top panel of Figure 1 displays the total number of new star alleles (Y-axis) for *CYP2C19* (n=3)*, CYP2C8* (n=3)*, CYP2C9* (n=7) *and CYP2D6* (n=10), respectively along the X-axis in magenta. The bottom panel of Figure 1 displays the total number of new diplotypes (Y-axis) for *CYP2C19* (n=17)*, CYP2C8* (n=12)*, CYP2C9* (n=23) *and CYP2D6* (n=132), respectively along the X-axis in purple.

Figure 2. Star Allele database search results example for *CYP2C19*. Figure 2 displays a screen shot of the web-based Star Allele Search. This example is displaying results for *CYP2C19,* chosen from the dropdown search on the top, left-hand side of the page. The user may choose to view the list of PharmVar annotated pharmacogenes, the NCBI entry for the selected gene, the associated Gene Search page (which will display all of the variants included in the 30x WGS dataset for the selected gene), or to return to the general search page. The user may choose to export the Star Allele search results to a CSV file by clicking the green button on the right-hand side of the page. The user may additionally choose to filter by a given Star Allele diplotype, and this filtered drop down also displays the number of samples with each corresponding diplotypes. Figure 2 displays results after filtering for *2|*2 diplotypes in the database.

Figure 1. New *alleles and diplotypes identified in sample set (n=3,202) relative to GeT-RM.

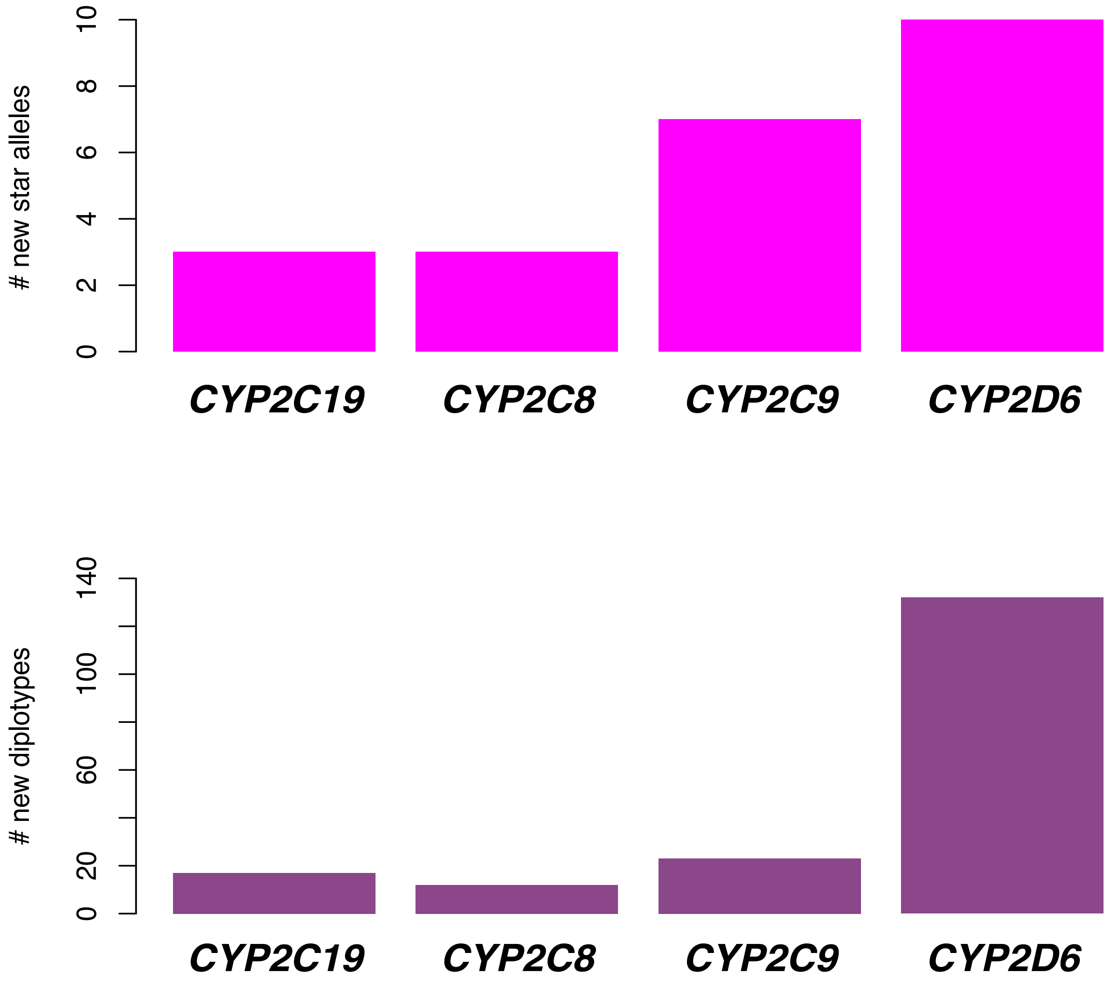
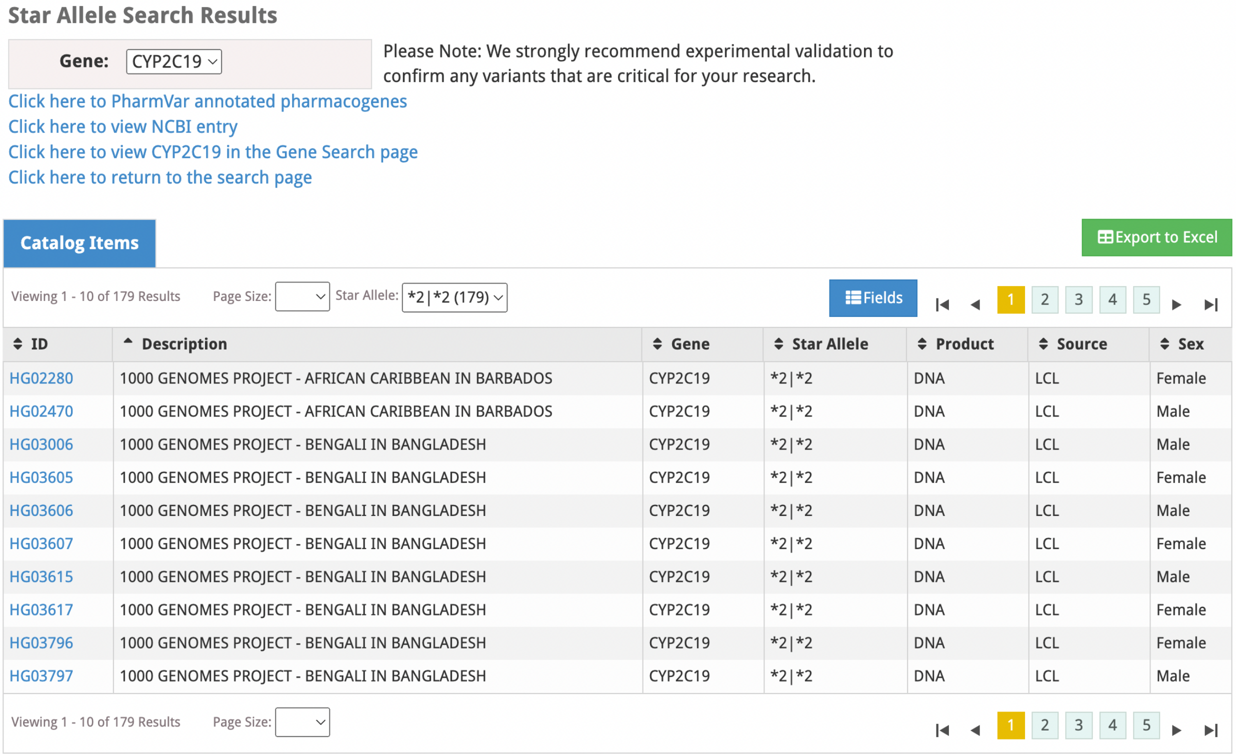


Figure 2. Star Allele database search results example for *CYP2C19*.



Star Allele Search Results

Gene: CYP2C19

Please Note: We strongly recommend experimental validation to confirm any variants that are critical for your research.

Click here to PharmVar annotated pharmacogenes
Click here to view NCBI entry
Click here to view CYP2C19 in the Gene Search page
Click here to return to the search page

Viewing 1 - 10 of 179 Results Page Size: Star Allele: *2|*2 (179)

ID	Description	Gene	Star Allele	Product	Source	Sex
HG02280	1000 GENOMES PROJECT - AFRICAN CARIBBEAN IN BARBADOS	CYP2C19	*2|*2	DNA	LCL	Female
HG02470	1000 GENOMES PROJECT - AFRICAN CARIBBEAN IN BARBADOS	CYP2C19	*2|*2	DNA	LCL	Male
HG03006	1000 GENOMES PROJECT - BENGALI IN BANGLADESH	CYP2C19	*2|*2	DNA	LCL	Male
HG03605	1000 GENOMES PROJECT - BENGALI IN BANGLADESH	CYP2C19	*2|*2	DNA	LCL	Female
HG03606	1000 GENOMES PROJECT - BENGALI IN BANGLADESH	CYP2C19	*2|*2	DNA	LCL	Male
HG03607	1000 GENOMES PROJECT - BENGALI IN BANGLADESH	CYP2C19	*2|*2	DNA	LCL	Female
HG03615	1000 GENOMES PROJECT - BENGALI IN BANGLADESH	CYP2C19	*2|*2	DNA	LCL	Male
HG03617	1000 GENOMES PROJECT - BENGALI IN BANGLADESH	CYP2C19	*2|*2	DNA	LCL	Female
HG03796	1000 GENOMES PROJECT - BENGALI IN BANGLADESH	CYP2C19	*2|*2	DNA	LCL	Female
HG03797	1000 GENOMES PROJECT - BENGALI IN BANGLADESH	CYP2C19	*2|*2	DNA	LCL	Male

Viewing 1 - 10 of 179 Results Page Size: