ursaPGx: a new R package to annotate pharmacogenetic star alleles using phased whole genome sequencing data

Gennaro Calendo[1], Dara Kusic[1], Jozef Madzo[1], Neda Gharani[1,2], Laura 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

Long-read sequencing technologies offer new opportunities to generate high confidence phased whole genome sequencing data for robust pharmacogenetic annotation. Here we describe a new user-friendly R package, ursaPGx, designed to accept multi-sample phased whole genome sequencing data VCF input files and output star allele annotations for pharmacogenes annotated in PharmVar.

Background

Pharmacogenomics (PGx) benefits medication management [1-7], however, pharmacogenetic annotation is often quite complex. Functional PGx annotation and corresponding clinical PGx recommendations rely on star (*) allele annotation [8, 9]; star alleles are often defined by more than one genetic variant [10-12]; and when the star allele defining variants are heterozygous, phased haplotype information is needed to resolve the annotation. In addition, annotations may change over time as new variants are characterized and incorporated into clinical PGx recommendations. Many resources and off the shelf tools are available to support researchers and clinicians interested in PGx annotation. Several tools are well suited for PGx annotation of unphased data (e.g., StellarPGx, Stargazer [13, 14]), and tools such as PharmCAT, while not computationally streamlined for multi-sample annotation, go a step further to incorporate clinical recommendations into the software output [15].

New long-read sequencing technologies offer opportunities to generate high confidence phased whole genome sequencing (WGS) data for robust PGx annotation. Here we describe ursaPGx, an R package designed to complement existing tools that leverages phased whole genome sequencing data for PGx annotation. ursaPGx is designed to run on a typical laptop computer using multi-sample, phased, WGS VCF files and provides an output table of star allele annotations for selected pharmacogenes annotated in PharmVar.

Materials and Methods

Samples

Phased multi-sample VCF files were downloaded for each of the star allele containing chromosomes from the 1000 Genomes Project. These VCF files were generated by the New York Genome Center for 3,202 1000 Genomes Project samples by aligning the 30x WGS reads to GRCh38 and performing SNV and INDEL variant calling as described in [16].

Benchmark data

The accuracy of the star allele calling algorithm of ursaPGx was benchmarked against the next generation sequencing consensus calls generated by the Genetic Reference and Testing Material Coordination Program (GeT-RM) for *CYP2C8*, *CYP2C9,* and *CYP2C19* which combined the output of Astrolabe, Stargazer, and Aldy across investigator groups to generate a uniform diplotype call for each of the 137 samples included in their study [17], of which 87 also have 30x WGS data [16]. *CYP2D6* calls generated by ursaPGx's implementation of Cyrius were benchmarked against calls generated by Chen et al. [18].

Implementation and algorithm description

Users may choose any phased WGS VCF file of interest for use as input to ursaPGx. ursaPGx assigns phased diplotype calls from single-sample or multi-sample indexed VCF files using publicly available star allele definitions from PharmVar [10-12]. First, for a given pharmacogene, star allele defining positions are used to extract genotype data for all samples in the VCF. Next, the extracted positions are checked against each PharmVar haplotype definition in order to determine 'callable' alleles. In this context, a callable allele is defined as a haplotype definition where all allele defining variants are present in the sample VCF. Downstream analysis is then limited to the set of callable alleles. The set of callable alleles is then used to generate a genomic position by haplotype definition reference matrix. The cells of the reference matrix contain the nucleotide which defines the given haplotype for each of the positions present in the sample VCF. Positions that are not part of a given haplotype definition are filled with the reference nucleotide for the position. Using this reference matrix allows ursaPGx to disambiguate star allele definitions such as *CYP2C19*2* and *CYP2C19*35,* which share the same core allele definitions (CYP2C19*2, non-reference alleles for rs4244285, rs12769205, rs3758581; CYP2C19*35, non-reference alleles for rs12769205, rs3758581) and therefore must be distinguished by using a SNV unique to *CYP2C19*2* (rs4244285). After constructing the reference matrix, genotype calls are converted to their nucleotide representation and split into haplotype strings for each sample. For each sample, each haplotype string is checked for exact matches against all columns of the reference matrix. All exact matches to the reference for each sample haplotype string are reported for each sample. If no exact matches occur, then the haplotype call for that sample is reported as ambiguous (*Amb). Haplotype calls for each sample are then combined to form a single diplotype call for the given pharmacogene for each sample included in the VCF.

CYP2D6 star allele calling in ursaPGx is performed with a modified version of Illumina *CYP2D6* star allele caller Cyrius, designed to function in R. The *CYP2D6* haplotype calling algorithm implemented in Cyrius is fully described in [18]. Briefly, Cyrius uses WGS BAM files to estimate the total number of copies of *CYP2D6* and *CYP2D7,* determines the number of complete *CYP2D6* and hybrid genes and uses these to estimate SVs impacting *CYP2D6* annotation. Cyrius then performs small variant calling for star allele defining positions and derives an estimate of their copy number, and then matches these calls and SVs against star allele definitions from PharmVar (7/15/2020) in order to produce final diplotype calls for each sample.

Software

ursaPGx is a freely available and open source package implemented in the R programming language [19] and utilizes the *VariantAnnotation* package [20] from the Bioconductor project to provide a consistent interface with existing R packages for the analysis of genetic variant data.

Star allele definitions in VCF format are downloaded from PharmVar (current version 5.2.13) and parsed into R objects. All package code and analysis scripts are hosted on GitHub (https://github.com/coriell-research/ursaPGx).

Requirements

ursaPGx is designed to run on a personal laptop. Star allele calling for all 3,202 1000 Genomes Project samples for all 12 pharmacogenes takes ~45 seconds on a 3.7 GHz 6-Core Intel Core i5 iMac. Cyrius *CYP2D6* calling implemented in ursaPGx takes ~4 seconds per sample BAM.

Results

CYP2C8, *CYP2C9*, and *CYP2C19* concordance was assessed for samples with matching IDs from the 30x WGS data in the GeT-RM benchmarking data sets (87/137) [17]. *CYP2D6* concordance was tested against diplotype calls from Chen et al. [18] in order to ensure accuracy of the Cyrius implementation within ursaPGx. Diplotype calls produced by ursaPGx were found to be highly consistent with those generated by GeT-RM for all four benchmarked pharmacogenes (**Table 1, Table S1**). For the 87 samples with matching IDs between the 1000 Genomes Project 30x WGS data and the GeT-RM NGS consensus benchmarking data, *CYP2C8* was found to be perfectly concordant. For *CYP2C19,* one subject sample (NA19122) was reported as *2|*Amb according to ursaPGx whereas the GeT-RM consensus call for this sample was reported as *2/*35. In the phased 30x WGS dataset, one haplotype was an exact match for *CYP2C19*2* but the other haplotype had no exact match to any PharmVar definition. Assuming accurate phasing of the input 30x WGS dataset, ursaPGx reports the inexact match as ambiguous for this sample.

For *CYP2C9,* three samples were found to be discordant between ursaPGx and GeT-RM reported consensus calls. Two of the subject samples with discordant *CYP2C9* calls, NA19143 and NA19213, were annotated as *1/*6 by GeT-RM whereas ursaPGx assigned these samples as *1|*1. Because the *CYP2C9*6* defining variant (rs9332131) is not present in the phased 30x WGS dataset, *CYP2C9*6* is not included as a callable allele by ursaPGx and thus is not reported for these samples. One subject sample, HG01190, was assigned as *61|*1 by ursaPGx whereas GeT-RM reported the diplotype as *2/*61. However, this sample was found to be inconsistently annotated across labs in the GeT-RM benchmarking data with a minority subset of three of the annotation approaches assigning *1/*61. Additionally, in the 30x WGS dataset, rs1799853 and rs202201137 are both heterozygous, and the non-reference allele for rs1799853 (*CYP2C9*2*) is on the same phased chromosome as the rs202201137 non-reference allele (presence of both non-reference alleles on the same haplotype defines the *61 variant according to PharmVar). Given the phase information from the 30x WGS, *61|*1 is the diplotype most consistent with the observed data for this sample.

Since Cyrius has already been shown to produce highly accurate *CYP2D6* star allele calls [18], we benchmarked ursaPGx's implementation of Cyrius against the 2,504 Phase 3 1000 Genomes Project samples analyzed in the Cyrius publication in order to ensure that changes made to Cyrius, which were needed to port the software to R, were consistent with the original Cyrius implementation. 2,502 of the 2,504 samples were found to be exact matches with the Cyrius reported results. For the two discordant samples, NA18611 and HG02490, ursaPGx reported diplotype calls for these samples (*10/*2 and *2/*33, respectively) whereas the Cyrius benchmark did not assign a diplotype for these samples. This discrepancy is likely due to

differences in BAM file input and downstream processing used in the 1000 Genomes Project NYGC 30x WGS data versus the WGS dataset used in the Cyrius publication [18].

Discussion

Here we describe a new pharmacogenetic annotation tool, ursaPGx, that is designed to complement existing tools by leveraging multi-sample phased WGS data and PharmVar annotations. ursaPGx is implemented as an efficient and user-friendly R package that provides a simple interface for assigning star allele diplotypes to samples for PharmVar annotated genes including *CYP2D6,* by integrating the Cyrius *CYP2D6* star allele caller. ursaPGx is especially well suited to long-read WGS datasets (e.g., PacBio HiFi) where phasing confidence is high.

Our benchmarking analysis demonstrated high concordance, 100%, 97% and 99%, respectively for the three overlapping pharmacogenes, *CYP2C8*, *CYP2C9*, and *CYP2C19* included in the most recent GeT-RM report [17]. Two of the discordant samples for *CYP2C9* result from a star allele defining variant (*6) that is present in the GeT-RM dataset but not occurring in the 30x WGS 1000 Genomes Project dataset used to benchmark ursaPGx. The third discordant *CYP2C9* sample (HG01190) results presumably from differences in phasing and variant calling results. Finally, as detailed in the methods section above, when no perfect match to any PharmVar defined haplotype occurs, the ursaPGx output will be '*Amb', and this implementation approach explains the single discordant *CYP2C19* sample, NA19122.

As with any annotation approach, ursaPGx includes several limitations. First and foremost, any error or missing variants in the input VCF file will propagate into errors in annotation. Similarly, any errors or uncertainty in phase will propagate into annotation errors, particularly when heterozygotes are phased incorrectly. In addition, our annotation approach is limited to the pharmacogenes annotated in PharmVar [10-12] and requires already phased input data. This annotation choice is specifically designed to take advantage of increasingly common long-read WGS datasets, such as the data being generated by the Human Pangenome Reference Consortium [21].

Conclusion

New long-read sequencing technologies offer opportunities to generate high confidence phased whole genome sequencing data for robust PGx annotation. Here we describe ursaPGx, a user-friendly R package that leverages multi-sample phased whole genome sequencing data for star allele annotation.

Availability and requirements

Project name: ursaPGx
Project home page: https://github.com/coriell-research/ursaPGx
Operating system(s): Platform independent
Programming language: R
Other requirements: None
License: GC will check Cyrius
Any restrictions to use by non-academics: GC will check Cyrius

Declarations

Ethics approval and consent to participate

All human data used in this study is publicly available through the 1000 Genomes Project [16].

Availability of data and materials

NYGC WGS data (VCF files) are available through the www.internationalgenome.org website (https://www.internationalgenome.org/data-portal/data-collection/30x-grch38), and can be accessed from the following website link (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20220422_3202_phased_SNV_INDEL_SV/). The version we used for the current study were last modified on 2022-11-14 08:33.

All package code and analysis scripts are hosted on GitHub (https://github.com/coriell-research/ursaPGx).

Competing interests

The authors declare that they have no competing interests

Funding

This study was funded by NHGRI 5U24HG008736 to LS.

Author's Contributions

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

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. Caudle KE, Klein TE, Hoffman JM, Muller DJ, Whirl-Carrillo M, Gong L, McDonagh EM, Sangkuhl K, Thorn CF, Schwab M *et al*: **Incorporation of pharmacogenomics into routine clinical practice: the Clinical Pharmacogenetics Implementation Consortium (CPIC) guideline development process**. *Curr Drug Metab* 2014, **15**(2):209-217.
9. Kalman LV, Agundez J, Appell ML, Black JL, Bell GC, Boukouvala S, Bruckner C, Bruford E, Caudle K, Coulthard SA *et al*: **Pharmacogenetic allele nomenclature: International workgroup recommendations for test result reporting**. *Clin Pharmacol Ther* 2016, **99**(2):172-185.
10. 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.
11. 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.
12. 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.
13. Twesigomwe D, Drogemoller BI, Wright GEB, Siddiqui A, da Rocha J, Lombard Z, Hazelhurst S: **StellarPGx: A Nextflow Pipeline for Calling Star Alleles in Cytochrome P450 Genes**. *Clin Pharmacol Ther* 2021, **110**(3):741-749.
14. Lee SB, Wheeler MM, Patterson K, McGee S, Dalton R, Woodahl EL, Gaedigk A, Thummel KE, Nickerson DA: **Stargazer: a software tool for calling star alleles from**

next-generation sequencing data using CYP2D6 as a model. *Genet Med* 2019, **21**(2):361-372.

15. Sangkuhl K, Whirl-Carrillo M, Whaley RM, Woon M, Lavertu A, Altman RB, Carter L, Verma A, Ritchie MD, Klein TE: **Pharmacogenomics Clinical Annotation Tool (PharmCAT)**. *Clin Pharmacol Ther* 2020, **107**(1):203-210.
16. 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.
17. 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.
18. 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.
19. Team RC: **R: A language and environment for statistical computing. R Foundation for Statistical Computing**. Vienna, Austria; 2020.
20. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M: **VariantAnnotation: a Bioconductor package for exploration and annotation of genetic variants**. *Bioinformatics* 2014, **30**(14):2076-2078.
21. Liao WW, Asri M, Ebler J, Doerr D, Haukness M, Hickey G, Lu S, Lucas JK, Monlong J, Abel HJ *et al*: **A draft human pangenome reference**. *Nature* 2023, **617**(7960):312-324.

Table 1. Concordance of ursaPGx diplotype calls with benchmarking datasets.

Gene	Concordance	Benchmarking Data
CYP2C8	1.00 (87/87)	GeT-RM [17]
CYP2C9	0.97 (84/87)	GeT-RM [17]
CYP2C19	0.99 (86/87)	GeT-RM [17]
CYP2D6	0.99 (2502/2504)	Cyrius [18]