An Open Source Tool for Determination of Pharmacogenomics Haplotypes from Diverse and Complex Data

S. M. Adams[1,2]; J. Larusch[2]; M. Ellison[2]; M. Haupt[2]; E. Orlova[2,3,4]; D.C. Whitcomb[2,5,6]; J. Gibson[2]

[1]Department of Pharmacogenomics, Shenandoah University School of Pharmacy, Fairfax, VA, USA; [2]Ariel Precision Medicine, Pittsburgh, PA, USA; [3]Department of Human Genetics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA; [4]Center for Craniofacial and Dental Genetics, University of Pittsburgh, Pittsburgh, PA; [5]Departments of Medicine, Cell Biology & Molcular Physiology and Human Genetics, University of Pittsburgh, Pittsburgh, PA; [6]Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA

1 Introduction

1.1 Background

- Growing availability of genomic data has increased the opportunity for precision medicine with pharmacogenomics (PGx).
- Variation in methods drives challenge in data interpretation for PGx, creating a need for accessible and configurable bioinformatics tools for determination of clinically relevant haplotypes and structural variations (SV) within pharmacogenes.
- Efforts by existing authorities (CPIC, PharmVar, PharmGKB) to standardize definitions are sporadically implemented.
- An existing open source solution for pharmacogenomics haplotyping and structural variation determination does not exist.

1.2 Objectives

Create an open source software tool that could determine pharmacogenomics haplotypes based on standardized inputs from pharmacogenomics authority organizations and provide estimation of breakpoints + copy gain/loss for genes with known structural variations (e.g. *CYP2D6*).

2 Development of hiMoon PGx

2.1 Implementation

- Written in Python, requires Python 3.6+.
- Heavy utilization of Pysam for VCF and BAM parsing.
- CNV detection optimized with Numba.

2.2 Haplotype Identification

- Considers all possible haplotypes H (from library) in each gene.
- Tests adherence of each subject S variant v (AA=0, AB=1, BB=2) to candidate haplotype.
- Determines all possible 2-way haplotype combinations (S_{H_1}, S_{H_2}) such that:

1. $f(S_H, 0) = 1$
2. $f(S_{H_1} \cap S_{H_2}, 2) = 1$

Where:

$$f(S_H, x) = \frac{|\{S_H | v > x\}|}{|S_H|}$$

- Reports diplotypes where the number of used variants in both haplotypes is maximized.
- Suggests that novel haplotype exists if additional variants that are not used in reported diplotype.

2.3 Haplotype Libraries

- Haplotype libraries are provided by the user, and follow the same format as the tab-delimited files available from PharmVar.
  - Users may also create their own files in this format for genes not annotated in PharmVar.

3 Development of hiMoon PGx Continued

3.1 Structural Variation Determination

Method 1: Maximum penalized likelihood estimation method (MPLE, below), adapted from *SeqCNV* method published by Chen, et al (DOI: 10.1186/s12859-017-1566-3).

- Seeks to maximize the following, where p_i is the probability of a given read originating from the unknown sample, t_i is the number of reads from the unknown sample for a given region, c_i is the number of reads from the control sample, and λ is the Bayesian information criterion (BIC).

$$PL = \sum_i (t_i ln(p_i) + c_i ln(1 - p_i)) - 2\lambda$$

- User defines regions to estimate breakpoints and copy number in an unknown sample relative to a control sample.
- Predicts copy gain if coverage ratio in highest likelihood region is > 1.4, or copy loss if < 0.6.

Method 2: Control region ratio determination (quick CNV).

- Copy number is estimated based on a highly similar region in the same sample (e.g. CYP2D6 = Target; CYP2D8P = Comparator).
- Depth of coverage is normalized to the length of the region, and a log ratio of the case:control is calculated.
- Usually a ratio <= 0 copy loss and ratio > 0.5 = copy gain.

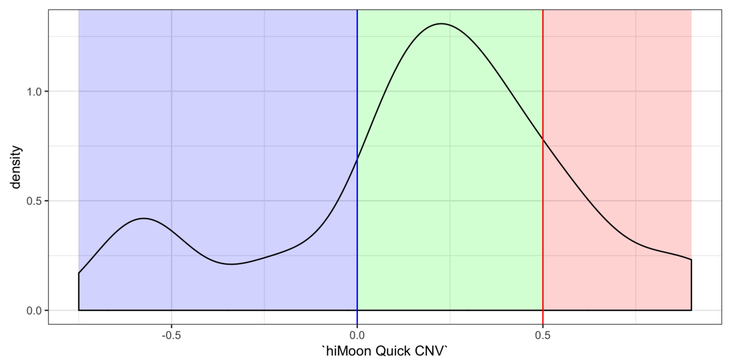

Figure 3.1: Bimodal distribution of average read depth shows delineation between copy loss and normal/gain.

3.2 Testing

- Whole exome sequencing BAM files were obtained from fifteen 1000 genomes samples.
- hiMoon was used to estimate star alleles and copy number with quick CNV.
  - Note that poor coverage in many areas leads to erroneous star allele calls relative to the known calls generated using higher fidelity genotyping methods.
- Haplotype and structural variation carried out in 100 samples from 1000X + coverage of *CYP2D6* with ~200bp flanks.

4 Results

4.1 Quick CNV and *CYP2D6* haplotype calling with 1000 Genomes WES data

- Whole exome data applied to the full *CYP2D6* translation table from PharmVar provides sufficient coverage for calling most star alleles.
- Inadequate coverage of intronic regions and gene flanks leads to ambiguities.
- hiMoon successfully called star alleles in 10/15 samples, and noted coverage ambiguity in 5/5 cases of the miscalled samples.
- Copy number was correctly called in all samples.

4.2 MPLE CNV and allele calls for 100 high coverage samples

- Evidence of structural variation (partial/complete gain or loss) in > 15% of samples.
- Non-ambiguous call obtained in 42/100 samples.
- Prevalence of ambiguous star allele calls highlights a possible need to subset translation tables due to the need for extensive gene coverage.

Table 4.1: Concordance with star alleles as reported by Pratt, et al. (DOI: 10.1016/j.jmoldx.2015.08.005)

ID	Expected	hiMoon Quick CNV	hiMoon Called
NA18945	*1/*5	-0.5878	*1/*5
NA19035	*2/*5	-0.5697	*2/*5
HG01190	*4/*5	-0.2199	*4/*5
NA07357	*1/*6	0.0536	*1/*6
NA18544	*10/*41	0.0950	*2/*10
NA07000	*9/*2 (*35)	0.1314	*2/*9
NA12717	*1/*1	0.1550	*1/*1
NA20509	*4/*35	0.2490	*2/*4
NA19239	*15/*17	0.2751	*2/*15
NA12006	*4/*41	0.2948	*10/*119
NA19007	*1/*1	0.3619	*1/*1
NA12878	*3/*4	0.4406	*3/*4
NA18565	*10/(*10 [*36])	0.5197	*10/*10xN
NA19785	*1/*2(XN)	0.6282	*34/*39(xN)
NA18959	*2/(*10 [*36])	0.8827	*2/*10(xN)

4.3 Future Directions

- Ongoing development to determine maximum resolution for breakpoints, currently >100bp.
- Further validation with other clinical relevant genes with common structural variation.
- Potential development of clinical outflow to allow integration with clinical practice guidelines.
- Transition Numba optimized CNV functions to Cython to improve speed with high coverage samples, currently up to 2 minutes per samples.

5 Conclusion

- Accessible, fast estimation of star alleles, easy to use, and tunable for data from sequencing, array, or targeted genotyping.
- Standardized library (inputs directly from PharmVar or formatted per PharmVar specifications).
- Reference agnostic (accepts VCF/BAM aligned to any reference, as long as library matches).
- Accurate calling of haplotypes + structural variation for complex pharmacogenes (e.g. CYP2D6).
- Conservative reporting with acknowledgement of possible novel discoveries.

6 Disclosures

All authors are affiliated with Ariel Precision Medicine. hiMoon is developed in coordination of resources from Shenandoah University and Ariel Precision Medicine.