A Probabilistic Method for Comprehensive Allele Typing of CYP2D6 Applied to 31K Exomes

Ruomu Jiang, Ph.D.[1], Shishi Luo, Ph.D.[1], Jasmine Dhaliwal, B.A.[1], Sharoni Jacobs, Ph.D.[1], William Lee, Ph.D.[1]

[1]Helix LLC, San Carlos, CA

Summary

We present a probabilistic method capable of calling most known CYP2D6 alleles using exome sequencing data.

Key improvement over previous NGS based methods[1,2]:

- account for experimental variation in sequencing processes.
- provide meaningful probability scores for QC purpose.
- the ability to flag potentially novel alleles.

We validate the method in 2 data sets and then apply it to 31K exomes sequenced from saliva samples.

- 74 different alleles are observed, with most being rare.
- 22% of samples carry a structural variant.
- up to 1.5% of samples carry a potentially novel allele.

Methods

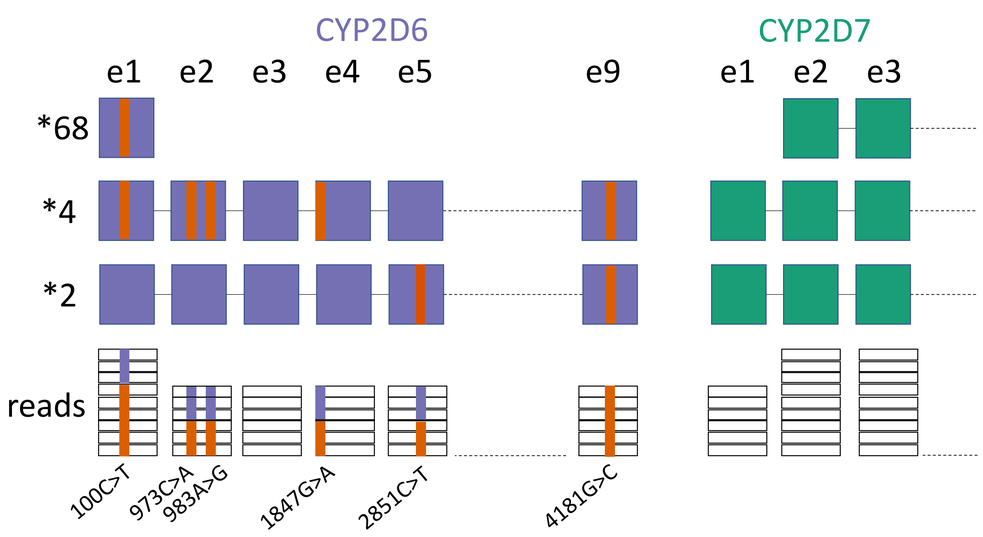
For each possible candidate allele combination (S), calculate the likelihood of it giving rise to the observed data, $\{CN_i\}$ $\{AD_j\}$. Find the combination with the highest likelihood.

CN_i: Per exon read count, $i \in$ all CYP2D6/7 exons

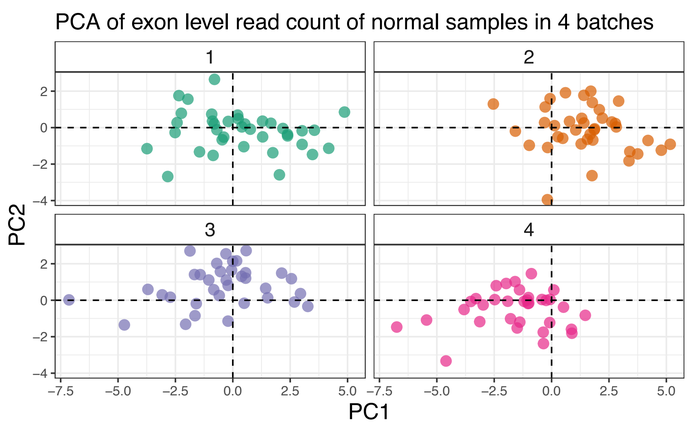
AD_j: REF/ALT allele depth, $j \in$ all defining mutations

$$Likelihood(S) = \prod_i^{exons} \mathrm{P}(CN_i|S) \times \prod_j^{mutations} \mathrm{P}(AD_j|S)$$

$$Quality(S_{1st}) = Likelihood\ (S_{1st}) \,/\, Likelihood(S_{2nd})$$



Exon level read count exhibits both intra- and inter- batch variability. A naive normalization method using a "normal" sample as reference does not account for such variability and can lead to biased estimate of $\mathrm{P}(CN_i|S)$.



Results

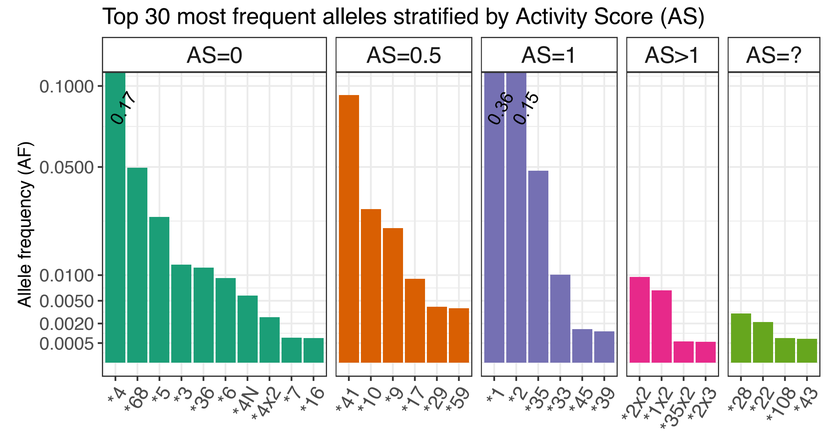
Evaluate the accuracy of allele typing by stratifying alleles into:

- Simple cases, alleles defined by SNPs/INDELs and *5.
- Complex cases, copy number amplification and hybrid alleles.

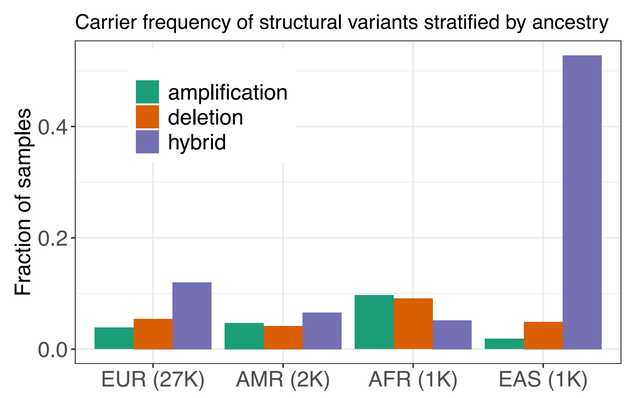
In two data sets, our method exhibits high accuracy and reproducibility. The incorrect calls are due to confusion between *10 vs *36, and amplification events such as *41x2 vs *41x3.

	Simple	Complex
Concordance with known allele calls in 270 replicates of 145 unique samples	398/399	105/108
Concordant calls in 550 replicates of 148 unique samples	703/703	166/167

Applying the method to 31K exomes (85% with mostly European ancestry), we observe 74 different alleles (98 if counting amplification as distinct alleles), most of which are rare with AF<0.0005. The top 17 most common alleles make up 97% of all observed alleles, while the top 30 make up 99.2%.



On average, 12.5% samples carry a hybrid allele, 5.5% a deletion and 4% an amplification. The distribution varies across ancestry, e.g. African Americans have more amplification (*4 and *2) and deletion while more than half of East Asians carry a hybrid allele (*36).



In 1.5% of samples, our method couldn't find a combination of known alleles that fully explains the observed data. We attribute this at least in part due to the presence of novel alleles, which consists of a novel combination of known defining mutations. For example we recurrently observe:

- An allele that has a single defining mutation, 983A>G.
- An allele that has a single defining mutation, 2611T>A.

We also observe novel putative loss-of-function variants in 0.1% of samples. For example:

- 7 samples harbor g.42126980delGA/L395fs in exon 8.
- 3 samples harbor g.42128326insC/L231fs in exon 5.

Reference

[1] Numanagic, I. *et al*. Allelic decomposition and exact genotyping of highly polymorphic and structurally variant genes. Nat. Commum. 9(1), 828 (2018)

[2] Lee, S. *et al*. Stargazer: a software tool for calling star alleles from next-generation sequencing data using CYP2D6 as a model. Genet Med. 21(2), 361 (2018)