Accurate *CYP2D6* star (*) allele diplotyping for long-read PacBio HiFi sequencing

J. Harting[1], N. Gonzaludo[1], L. Zhu[1], X. Chen[1], A. Gaedigk[2], S. A. Scott[3], P. Baybayan[1]

1. PacBio, Menlo Park, CA; 2. Children's Mercy, Kansas City, MO; 3. Stanford University, Stanford, CA

Introduction

The *CYP2D6* locus is known for its importance to pharmacogenetics as well as for its high diversity and complexity.

SNPs, indels, gene duplications, hybrid genes, gene conversion, and large deletions are common at this locus.

Resolving and phasing individual alleles without imputation requires long and highly accurate reads.

We demonstrate a novel *CYP2D6* typing tool, Pangu, for accurately assigning star allele diplotypes from PacBio HiFi reads.

Unambiguous calls are made using multiple input data types from HiFi long-read sequencing, including whole genome sequencing (WGS), hybridization-based target enrichment, and long-range amplicon sequencing.

Validation datasets:

WGS (read length ~15 kb)

- 32 HPRC[1] samples with orthogonal short-read calls using Cyrius[2]

Target enrichment[3] (read length ~5.5 kb)

- 23 Coriell samples selected from the GeT-RM[4] reference set

Long-range amplicons[5] (read length 9–12 kb)

- 22 Coriell samples selected from the GeT-RM[4] reference set

Star-allele typing analysis

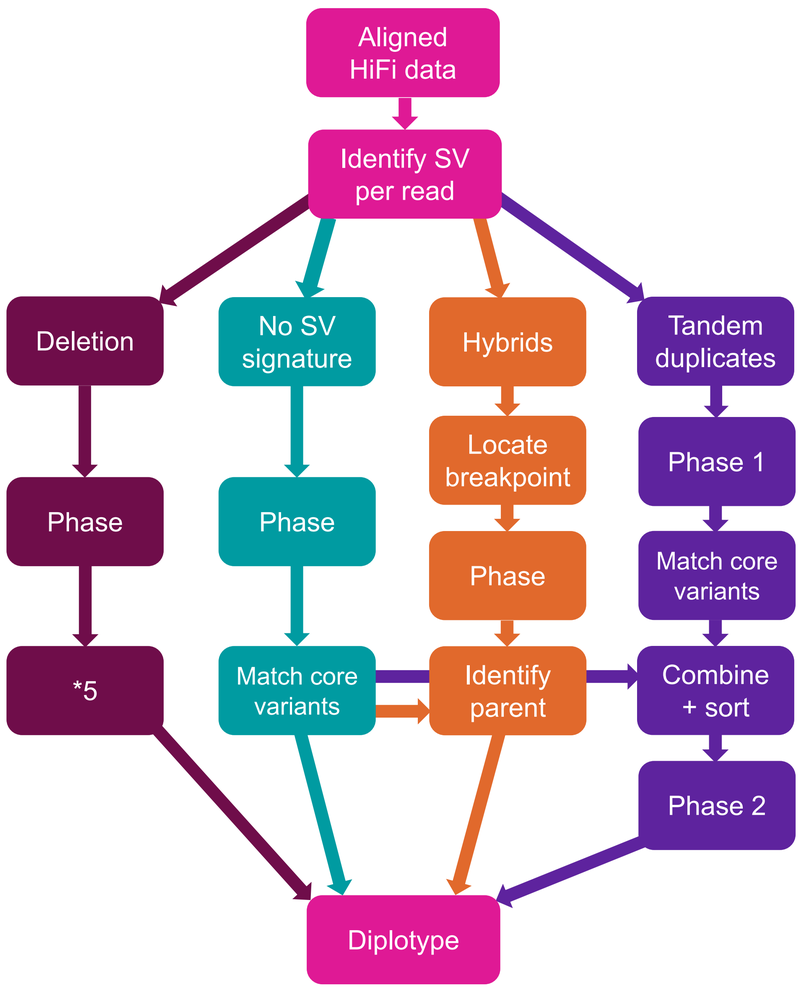

Figure 1. Typing workflow. HiFi reads are sorted into structural variant (SV) subclasses and phased where possible prior to recombining into accurate haplotypes.

Whole genome HiFi sequencing

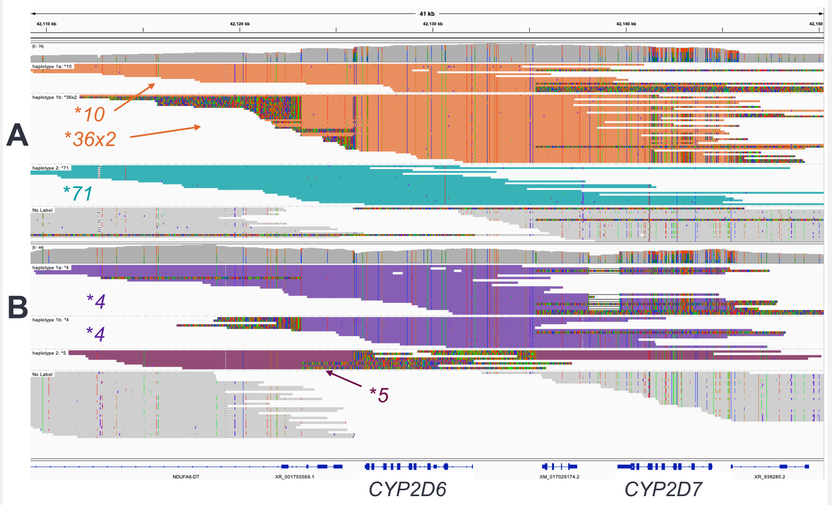

Figure 2. Unambiguous haplotypes. Long and accurate HiFi reads allow for unambiguous copy-number assignment to alleles. **(A)** Sample **HG00438** (**71/*36x2+*10*). **(B) NA20129** (**4x2/*5*) incorrectly called "**4/*4*" by short-read callers. Individual HiFi reads are long enough to span multiple tandem copies. Gray reads are unlabeled.

HiFi target enrichment

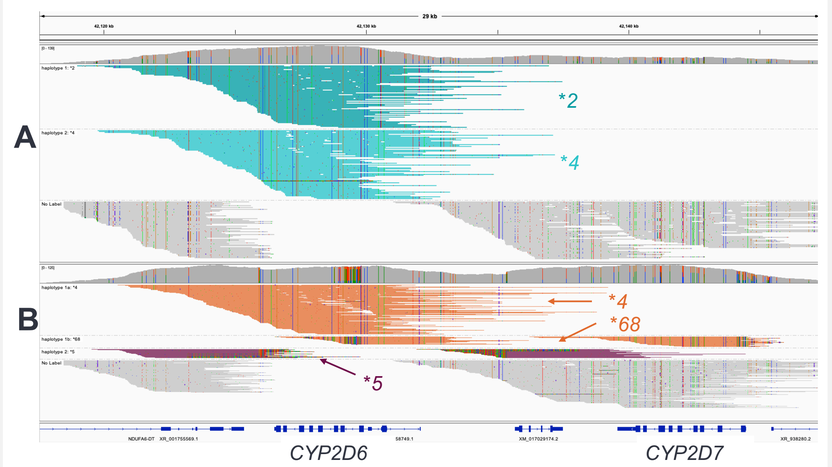

Figure 3. Efficient and comprehensive for panels. Efficiently capture and call *CYP2D6* as part of a targeted pharmacogenomics panel. **(A)** Typical sample without SV, **HG002** (**2/*4*). **(B)** Positive evidence for whole-gene deletion and hybrids, **HG01190** (**5/*68+*4*). Gray reads are unlabeled.

HiFi amplicon sequencing

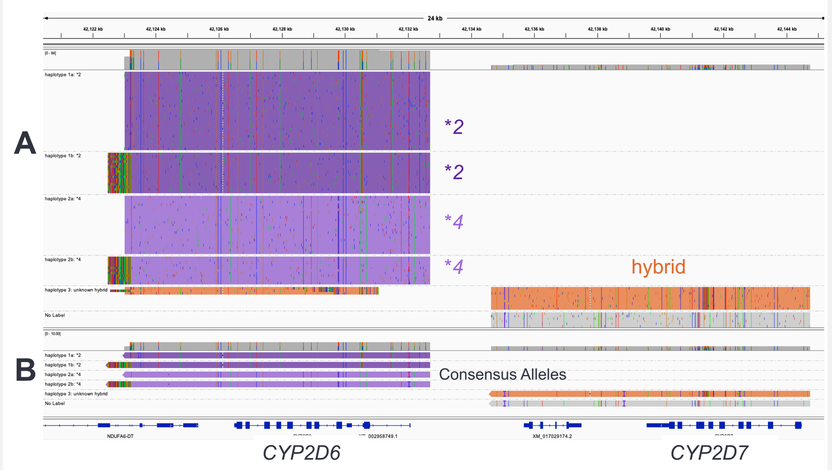

Figure 4. HiFi reads or consensus as input. Amplicon data can be typed from **(A)** HiFi reads directly or **(B)** after clustering and consensus using pbaa[6]. Diplotype "**2x2/*4x2 + hybrid*" for **NA17244** includes four unique *CYP2D6* alleles and an unnamed hybrid with introgression of *CYP2D6* in the middle of *CYP2D7*. Gray reads/consensus are unlabeled.

Acknowledgements

The authors would like to thank everyone who helped generate data for the poster.

Results

Sample	CYP2D6 reference	PacBio HiFi call	Sample	CYP2D6 reference	PacBio HiFi call
Whole genome sequencing					
HG00438	**36+*36+*10/*71*	**36x2+*10/*71*	HG02148	**1/*2*	**1/*2*
HG00621	**5/*49*	**5/*49*	HG02257	**2/*35*	**2/*35*
HG00673	**1/*10*	**1/*10*	HG02572	**1/*1*	**1/*1*
HG00733	**68+*4/*41*	**68+*4/*41*	HG02622	**17/*46*	**17/*46*
HG00735	**1/*4*	**1/*4*	HG02717	**1/*1*	**1/*1*
HG00741	**1/*2*	**1/*2*	HG02723	**2/*17*	**2/*17*
HG01071	**4/*41*	**4/*41*	HG02818	**1/*2*	**1/*2*
HG01106	**4/*35*	**4/*35*	HG02886	**1/*4*	**1/*4*
HG01175	**41/*41*	**41/*41*	HG03453	**1/*29*	**1/*29*
HG01258	**2/*68+*68+*4*	**2/*68x2+*4*	HG03486	**1/*17*	**1/*17*
HG01358	**1/*1*	**1/*1*	HG03516	**2/*17*	**2/*17*
HG01361	**1/*4*	***1/*4*	HG03540	**2x2/*17*	**2x2/*17*
HG01891	**5/*29*	**5/*29*	HG03579	**1/*2*	**1/*2*
HG01928	**1/*1*	**1/*1*	NA18906	**17/*125*	**17/*125*
HG01952	**1/*4*	**1/*4*	NA19240	**15/*17*	**15/*17*
HG01978	**1/*1*	**1/*1*	**NA20129**	**4/*4*	**4x2/*5*
Target enrichment sequencing					
HG002	**2/*4*	**2/*4*	NA18552	**1/*14*	**1/*14*
HG00276	**4/*5*	**4/*5*	NA18564	**2/*36+*10*	**2/*36+*10*
HG01190	**68+*4/*5*	**68+*4/*5*	NA18855	**1/*5*	**1/(*5)*
NA07019	**1/*4*	**1/*4*	NA18868	**2/*5*	**2/*5*
NA07348	**1/*6*	**1/*6*	NA19109	**2x2/*29*	**2x2/*29*
NA10831	**4/*5*	**4/*5*	NA19122	**2/*17*	**2/*17*
NA11832	**1/*68+*4*	**1/*68+*4*	NA19147	**17/*29*	**17/*29*
NA18484	**1/*17*	**1/*17*	NA19174	**4/*40*	**4/*40*
NA18509	**2/*17*	**2/*17*	NA19207	**2x2/*10*	**2x2/*10*
NA18518	**17/*29*	**17/*29*	NA19226	**2/*2x2*	**2/*2x2*
NA18519	**29/*1*	**29/*106*	NA19239	**15/*17*	**15/*17*
NA18540	**32x2+*10/*41*	**36+*10/*41*			
Amplicon sequencing					
NA02016	**2xN/*17*	**2x2/*17*	NA17211	**2/*4*	**2/*4*
NA07439	**4x2/*41*	**4x2/*41*	NA17214	**2/*2*	**2/*2*
NA09301	**1/*2x2*	**1/*2x2*	NA17215	**4/*41*	**4/*41*
NA12244	**35/*41*	**35/*41*	NA17217	**1/*41*	**33/*41*
NA16654	**10/*10*	**36+*10/*10*	NA17226	**4/*4*	**4/*4.013+*4*
NA16688	**2/*10*	**2/*36+*10*	NA17227	**1/*9*	**1/*9*
NA17020	**1/*10*	**1/*10*	NA17232	**2/*2xN*	**2x2/*35*
NA17039	**2/*17*	**2/*17*	**NA17244**	**2x2/*4x2 + hyb*	**2x2/*4x2 + hyb*
NA17073	**1/*17*	**1/*17*	NA17276	**2/*5*	**2/*5*
NA17114	**1/*5*	**1/*5*	NA17282	**41/*41*	**41/*41*
NA17209	**1/*4*	**1/*4.013+*4*	NA17300	**1/*6*	**1/*6*

Table 2. HiFi calling accuracy. Comparison of reference *CYP2D6* calls (GeT-RM and/or short-read) to calls made with HiFi reads and the PacBio caller. Magenta calls indicate HiFi corrections or improvements over purple reference calls. Target enrichment was not able to resolve a double tandem hybrid in NA18540. **Bolded** samples are displayed in figures 2–4.

Conclusion

- Generalized software designed for long, highly accurate PacBio HiFi data
- Multi-mode star-allele calling for all HiFi sequencing data types
- Comprehensive *CYP2D6* diplotyping from a single assay
- Fully phased and labeled reads for easy validation
- Find and characterize novel genotypes

References

1. https://github.com/human-pangenomics/HPP_Year1_Data_Freeze_v1.0
2. Chen, X., et al. (2021). Cyrius: accurate *CYP2D6* genotyping using whole-genome sequencing data. *Pharmacogenomics J. 2021; 21(2): 251–261.*
3. https://www.pacb.com/wp-content/uploads/Application-Brief-HiFi-Target-Enrichment-Best-Practices.pdf
4. https://www.cdc.gov/labquality/get-rm/index.html
5. https://www.pacb.com/wp-content/uploads/Application-note-HiFi-amplicon-sequencing-for-pharmacogenetics-CYP2D6.pdf
6. https://github.com/PacificBiosciences/pbAA

Code availability:
https://github.com/pacificbiosciences/pangu/