Resolving Highly Diverse HLA and CYP2D6 Alleles using HiFi Sequencing for Long-Range Amplicon Data with a New Clustering Algorithm

Abstract #: B14
John Harting, Lei Zhu, Ian McLaughlin, Zev Kronenberg
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

HLA and CYP2D6 loci are highly diverse genes important to pharmacogenetics and immunology.

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

We demonstrate and benchmark the accuracy of PacBio HiFi reads and the pbaa clustering algorithm for resolving these important loci.

Methods

HLA

- 8 Coriell samples
  HG001,HG002,HG003,HG004,HG005,HG007, 06986-3,C1-218
- 6 Loci
  HLA-A/ -B / -C / -DPB1 / -DQB1 / -DRB1
- GenDx NGSgo-MX6-1 kit
- Replicate samples barcoded and pooled at 96 plex
- HiFi reads analyzed by pbaa
- Typing results validated with NGSEngine

CYP2D6

- 22 Coriell samples (see table 3)
- 3 Amplicon primer design (see ASHG Poster #3540)
- Barcoded and pooled at 22 plex
- HiFi reads analyzed by pbaa and pbCYP2D6typer
- Typing results validated against GeT RM pharmacogenetics panel

➡ *5 allele primers: 5.1 kb amplicon
➡ Upstream dup primers: 8.6 or 10.2 kb amplicon
⇨ Downstream primers: 8.2 kb amplicon

Qiao et al., 2019; Fukuda et al., 2005

Figure 1. CYP2D6 Primer Design. Three amplicon design captures duplicates and deletion alleles in one assay.

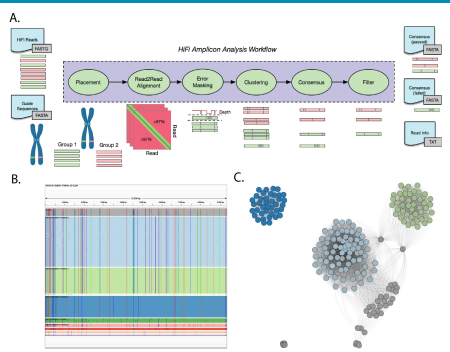

Figure 2. Pbaa Workflow and Visualization. (A) Clustering workflow. HiFi reads are assigned to guides and errors are masked within groups. Corrected reads are clustered and consensuses are generated. Post process filters separate pass/fail clusters. **(B)** Clustered and painted aligned HiFi reads in IGV. **(C)** Corrected HiFi read graph, colors match alignments with passing clusters in image B.

Results

HLA

Pbaa results are highly accurate at the recommended 100-fold coverage per locus.

	10x	20x	30x	40x	50x	75x	100x	300x	500x
TP	615	1028	1067	1075	1075	1085	1088	1092	1092
FN (filtered)	140	25	23	15	17	7	4	0	0
FN (missing)	337	39	2	2	0	0	0	0	0
FP	0	4	5	3	2	1	0	0	0
Accuracy	0.56	0.94	0.97	0.98	0.98	0.99	1.00	1.00	1.00
Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Recall	0.56	0.94	0.98	0.98	0.98	0.99	1.00	1.00	1.00
Avg. edit distance	1.06	0.21	0.19	0.16	0.15	0.13	0.13	0.14	0.14
Avg. PHRED QV	35.3	43	43.7	44.4	45.2	45.5	45.2	45.2	45.1

Table 1. HLA Accuracy Titration. Pbaa results compared against the truth set. Shaded rows are presence/absence benchmark statistics. True positive (TP) is a pbaa consensus sequence that has a best match with an allele in the truth set. False negatives (FN) are alleles filtered by pbaa or missing completely from pbaa result set. False positives (FP) are additional clusters generated for expected truth alleles.

Results

CYP2D6

Pbaa results are highly accurate at the recommended minimum 100-fold coverage.

	100x	200x	300x	400x	500x	1000x
TP	53	53	53	53	53	53
FN (filtered)	0	0	0	0	0	0
FN (missing)	0	0	0	0	0	0
FP	1	0	0	0	0	0
Accuracy	0.98	1.00	1.00	1.00	1.00	1.00
Precision	0.98	1.00	1.00	1.00	1.00	1.00
Recall	1.00	1.00	1.00	1.00	1.00	1.00
Avg. edit distance	0.02	0	0	0	0	0
Avg. PHRED QV	56	>56	>56	>56	>56	>56

Table 2. CYP2D6 Accuracy Titration. Pbaa results compared against truth set.

Improved calls:
- NA09301 Duplication resolved
- NA17217 Missed variant in reference
- NA17232 Phased variants improve call
- Multiple Hybrid alleles (*36) identified

Sample	*CYP2D6* Reference	HiFi + pbaa Calling	Sample	*CYP2D6* Reference	HiFi + pbaa Calling
NA02016	*2×N/*17	*2×2/*17	NA17211	*2/*4	*2/*4
NA07439	*4×N/*41	*4×2/*41	NA17214	*2/*2	*2/*2
NA09301	Duplication	*1/*2×2	NA17215	*4/*41	*4/*41
NA12244	*35/*41	*35/*41	NA17217	*1/*41	*33/*41
NA16654	*10/*10	*10 + *36	NA17226	*4/*4	*4 + *36
NA16688	*2/*10	*2/*10 + *36	NA17227	*1/*9	*1/*9
NA17020	*1/*10	*1/*10	NA17232	*2/*2×N	*2×2/*35
NA17039	*2/*17	*2/*17	NA17244	DUP *4/*2A	*2×2/*4
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 + *36	NA17300	*1/*6	*1/*6

Table 3. HiFi CYP2D6 *-Allele Calls. Published calls compared to calls generated from long read HiFi amplicons. Calls in red are improved with respect to published results.

Discussion

- The long read lengths and high accuracy of PacBio HiFi reads allow for unprecedented precision when sequencing complex and diverse loci such as HLA and CYP2D6.
- The new pbaa algorithm for HiFi reads generates highly accurate consensus sequences as benchmarked against 6 HLA loci.
- Typing CYP2D6 samples with structural variants (SV), pseudogenes, and pooled amplicons can be problematic with current assays.
- Long highly accurate reads map uniquely for consistent type calls
- Pbaa can resolve CYP2D6 alleles into easily-typed, fully phased consensus sequences.

Conclusion

The pbaa clustering algorithm was developed as a HiFi successor to the previous PacBio long amplicon analysis tool for clustering long targeted reads.

The application of PacBio HiFi reads and pbaa to the HLA and CYP2D6 targets provides a demonstration of the utility of these tools.

- **Comprehensive** variant detection, including SV
- **Robust** read clustering with fully **phased** results
- Uniquely **map-able** to gene or pseudogene
- Highly accurate ***-allele** calls

References

Pbaa: https://github.com/PacificBiosciences/pbAA
CYP2D6 typing: https://github.com/PacificBiosciences/apps-scripts/tree/master/CYP2D6tools
HLA Benchmark Data: https://downloads.pacbcloud.com/public/dataset/pbAmpliconAnalysis_HLA/
GenDx kit: https://www.gendx.com/product_line/ngsgo-mx6-1/
NGSEngine: https://www.gendx.com/product_line/ngsengine/
GeT RM: GeT_RM_pharmacogenetics