Resolving Complex Pathogenic Alleles using HiFi Long-range Amplicon Data and a New Clustering Algorithm

Abstract #: eP240

John Harting, Cheryl Heiner, Ian McLaughlin, Zev Kronenberg
PacBio, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

Genetic diseases caused by pathogenic variants in genes with highly homologous pseudogenes commonly include structural rearrangements and distantly separated heterozygous SNV that are difficult to call. We leverage highly accurate long-range single-molecule HiFi reads to accurately and consistently call a wide range of complex disease-causing variants in two gene-pseudogene systems by applying a single multiplexed long-range paired assay to each target.

We amplified 7 samples with complex pathogenic genotypes causing adrenal hyperplasia (*CYP21A2/CYP21A1P*) and 13 samples with genotypes causing Gaucher disease (*GBA/GBAP1*). We use a new amplicon clustering method, pbaa, which is tailored to HiFi data and designed to deconvolve amplicon mixtures. 100% of expected variants for all samples and replicates were called and phased, including large structural variants such as gene fusions and gene deletions, as well as complex heterozygous SNV.

Methods

Targets and Sequencing

Samples with known difficult genotypes were obtained from Coriell.

Co-Amplified Targets (including gene flanking regions):

7 samples: CYP21A2 (10kb)/*CYP21A1P* (8kb)

13 samples: GBA (12kb)/*GBAP1* (15kb)

Amplicon Libraries replicated to obtain 24-plex for each target.

Primers selected to include full-length genes and pseudogenes for comprehensive SNV detection.

Large deletions occurring between primers generate unique amplicons with the outermost primer pairs.

Sequencing on PacBio Sequel and Sequel II instruments.

See poster #eP273 Targeting Clinically Significant Dark Regions of the Human Genome with High-Accuracy, Long-Read Sequencing for details on experimental design and sequencing.

Analysis

Single-molecule Circular Consensus Sequence (CCS) or HiFi reads were generated and demultiplexed using SMRT Link version 9.0.

Primers for each amplicon are identified for deletion events.

pbaa Amplicon Analysis workflow for clustering and consensus

Assign to locus using *guide* kmers

Reference-free variant detection

Cluster and consensus

Quality filtering

Custom annotation and variant calling to generate reports and VCF.

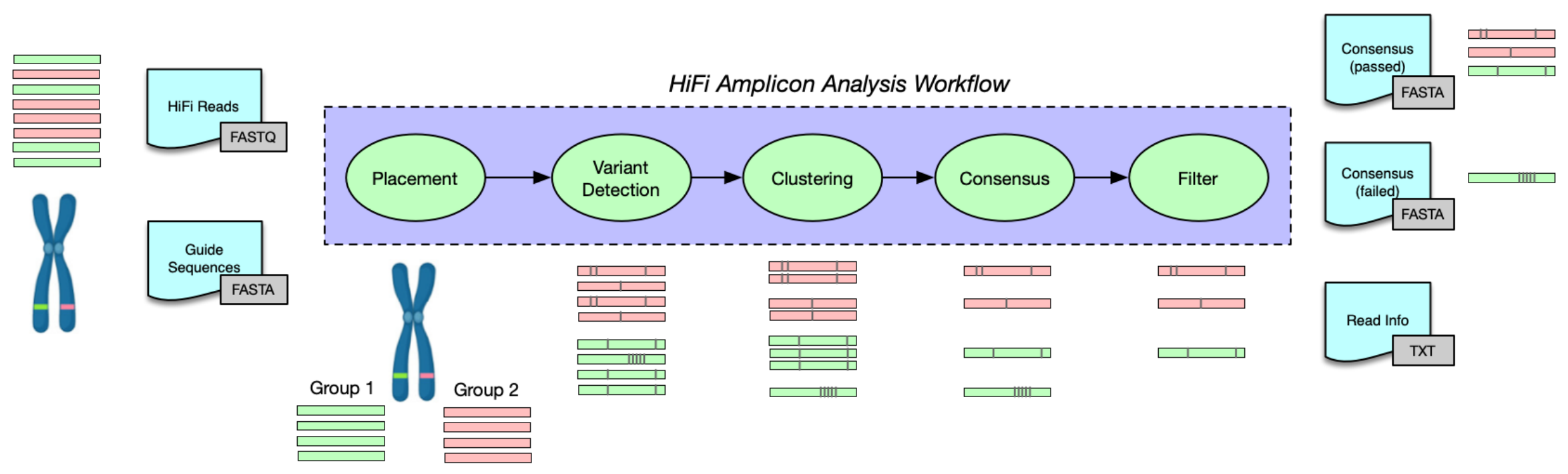

Figure 1. pbaa Workflow. Fast and sensitive separation of complex mixtures of HiFi reads from amplicon targets. pbaa generates high-quality consensus sequence and includes tools for evaluating clustered datasets.

Results

> 1,700,000 HiFi reads 8 kb - 10 kb for *CYP21A2 / CYP21A1P*

> 1,000,000 HiFi reads 12 kb - 15 kb *GBA / GBAP1*

One SMRT Cell 8M each with a median HiFi accuracy >99.8%

No sample drop-outs for any replicate. See #eP273 for sequencing details.

See figure 2 for an example of clustered HiFi reads subsampled to recommended 100 reads / amplicon target.

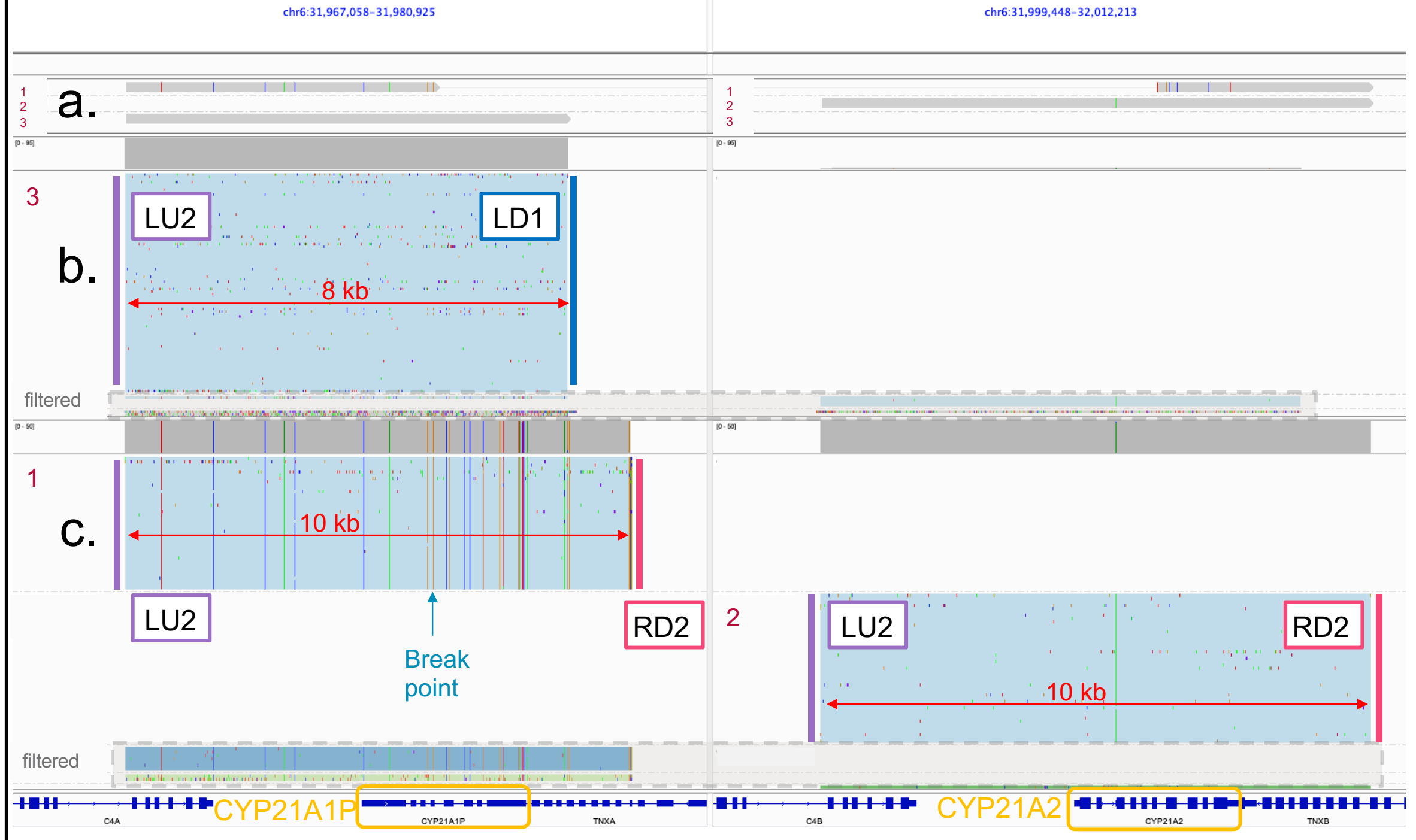

Figure 2. HiFi Read Clustering. Clustered and painted *CYP21A2* and *CYP21A1P* HiFi reads subsampled to 100 reads/locus for sample NA14733 with 30 kb deletion. Red numerals indicate clusters and corresponding consensus sequences. Primers are indicated by colored bars. **a.** pbaa consensus sequences. **b.** Primers LU2/LD1. **c.** Primers LU2/RD2.

CYP21A2 / CYP21A1P

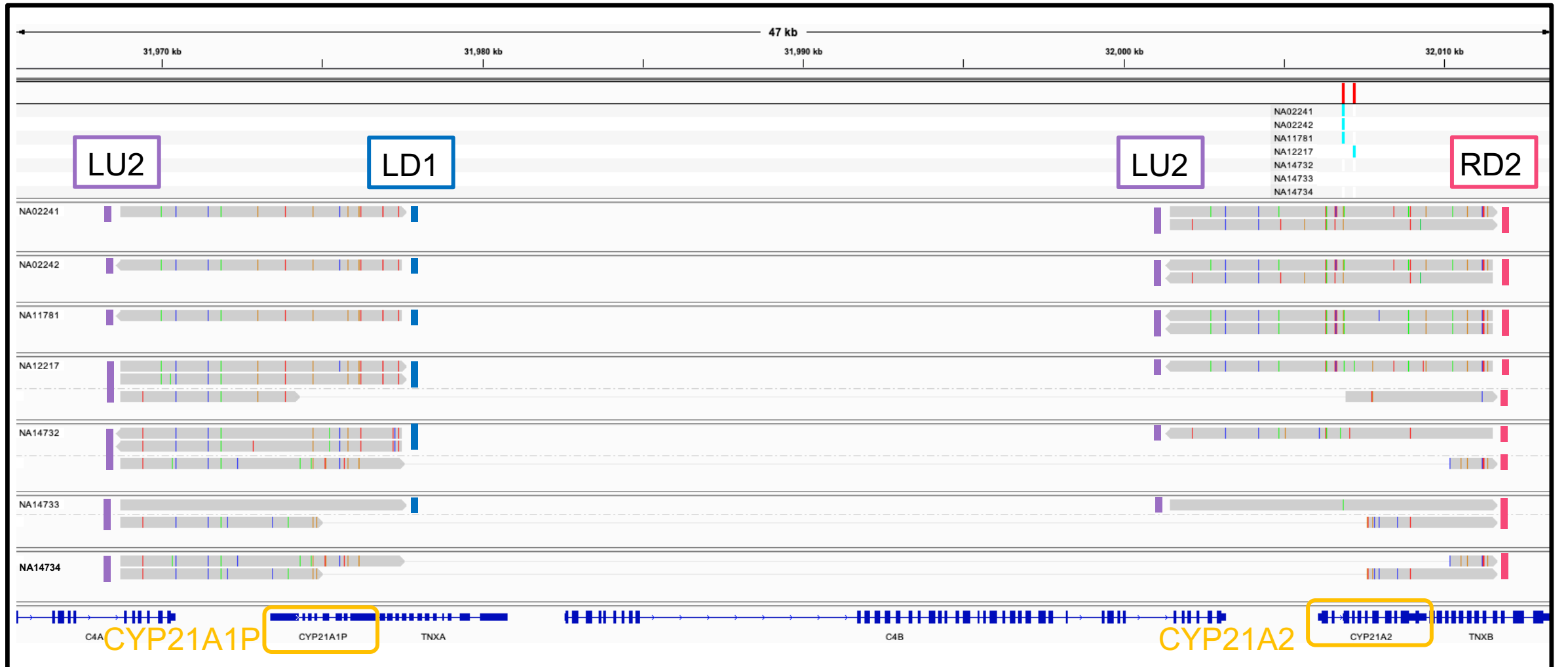

Figure 3. *CYP21A2* and *CYP21A1P* pbaa Consensus. Colored bars indicate amplifying primers. Expected pathogenic variants from ClinVar are indicated at the top in red with observed calls in light blue. Note 3 different examples of the common ~30 kb deletion genotype with unique breakpoints as shown.

	Sample	Status	Correctly typed	CYP21A2 Alleles	Pathogenic Variant Calls	CYP21A1P Alleles
	NA02241	Affected	✓	2	Homozygous SNV	1
	NA02242	Affected	✓	2	Homozygous SNV	1
	NA11781	Affected	✓	2	Homozygous SNV	1
	NA12217	Affected	✓	1	Homozygous SNV & Gene Fusion	3*
trio	NA14732	Carrier	✓	1	CYP21A2 Deleted	3
	NA14733	Carrier	✓	1	Gene Fusion	2**
	NA14734	Affected	✓	0	Deletion & Gene Fusion	2**

Table 2. CYP21A2 Call Summary. Phased calls from pbaa consensus sequences match all expected calls for all samples and replicates and include additional information about large structural variants. */** Count includes hybrid *CYP21A1P-CYP21A2* gene*.*

Results

GBA / GBAP1

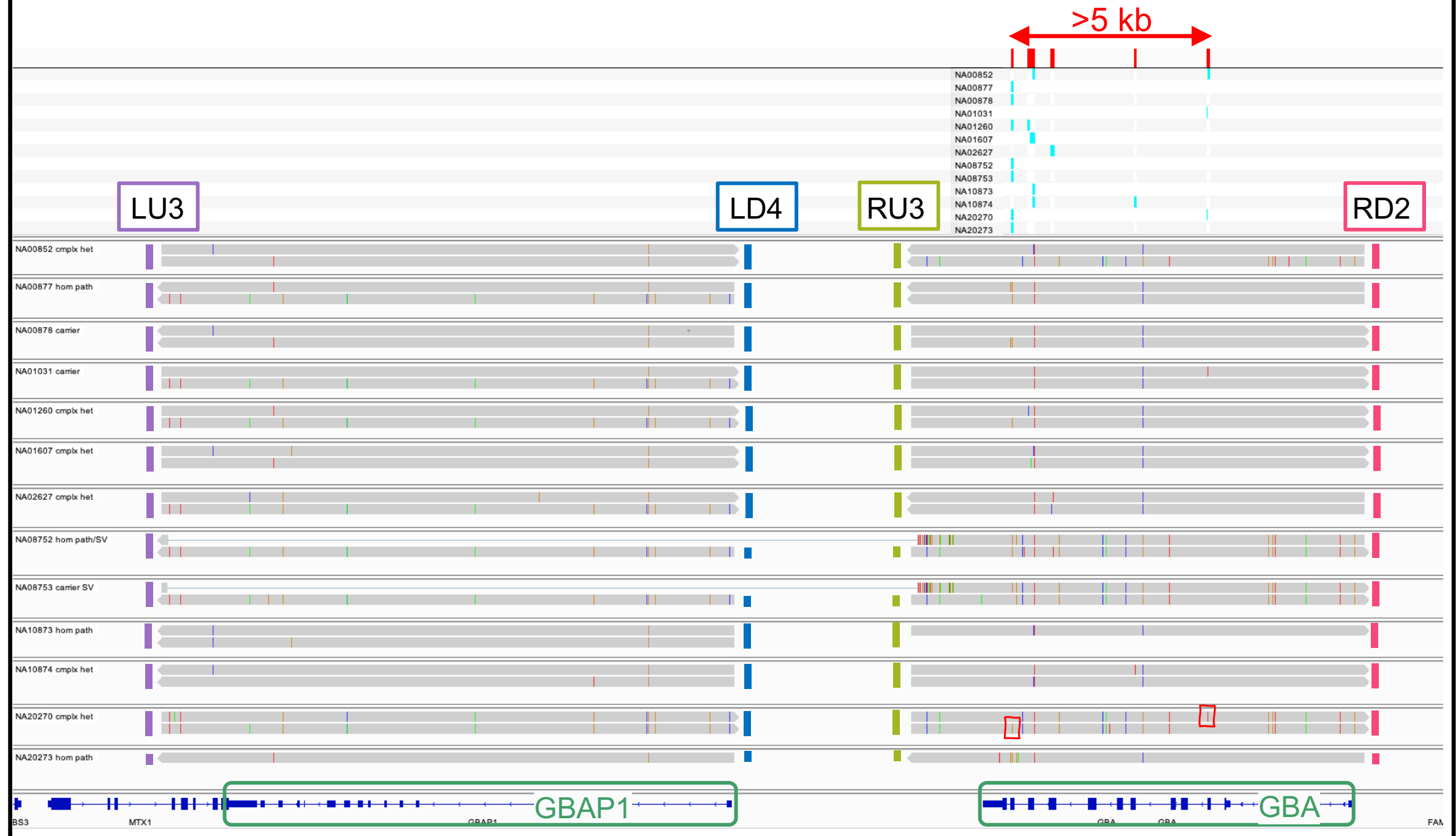

Figure 4. *GBA* and *GBAP1* pbaa Consensus. Colored bars indicate amplifying primers. Expected pathogenic variants from ClinVar indicated at top in red with observed calls in blue. Outer primers (purple and pink) generate amplicons from samples with large deletions. Complex heterozygotes in *GBA* include phased variants >5 kb apart in a single sample.

Sample	Status	Correctly typed	GBA Alleles	Pathogenic Variant Calls	Variant Separation	GBAP1 Alleles
NA00852	Affected	✓	2	Complex Heterozygous SNV	4817 bp	2
NA00877	Affected	✓	2	Homozygous SNV	-	2
NA00878	Carrier	✓	2	Heterozygous SNV	-	2
NA01031	Carrier	✓	2	Heterozygous SNV	-	2
NA01260	Affected	✓	2	Complex Heterozygous SNV	456 bp	2
NA01607	Affected	✓	2	Complex Heterozygous SNV	71 bp	2
NA02627	Affected	✓	2	Complex Heterozygous SNV	51 bp	2
NA08752	Affected	✓	2	Homozygote SNV & Fusion	-	1
NA08753	Carrier	✓	2	Heterozygous SNV	-	1
NA10873	Affected	✓	1	Homozygous SNV	-	2
NA10874	Affected	✓	2	Complex Heterozygous SNV	2786 bp	2
NA20270	Affected	✓	2	Complex Heterozygous SNV	5377 bp	2
NA20273	Affected	✓	1	Homozygous SNV	-	1

Table 1. GBA Call Summary. Phased calls from pbaa consensus sequences match all expected calls for all samples and replicates.

Discussion

CYP21A2 / CYP21A1P

- Common large (~30 kb) deletions occur in as many as 30% of some populations[1].
- 3 unique deletions identified in NA12217, NA14732, NA14733; 2 were confirmed in proband sample NA14734
- Deletion breakpoints determined by custom analysis are consistent with references[2].
- Copy number variation (CNV) for both gene and pseudogene are common[1].
- CNV is identified for *CYP21A1P* in NA12217 and NA14732
- Long-range amplicons confirm 2 unique alleles for homozygous calls in NA02241, NA02242, NA11781

GBA / GBAP1

- Phased complex heterozygous calls span genomic distances > 5 kb
- Homozygous SNV in NA00877 confirmed with 2 unique long-range alleles
- Homozygous SNV in NA08752 linked to deletion of *GBAP1* from the mother, NA08753. Paternal allele separately identified.

Conclusions

- Long-range PacBio HiFi reads for genetic assays:
  - Comprehensive variant detection
    - Including multi-kb structural variants
  - Accurate and reproducible for all variant classes
  - Phased results
  - Uniquely map-able to gene or pseudogene
  - Single targeted assay
- Paired amplicon assays
  - Efficient for capture of common deletion and fusion events
- pbaa clustering and consensus
  - Robust separation of mixtures of complex alleles
  - Accurate variant calling from consensus

Studies and databases utilizing targeted long-range and highly accurate HiFi reads have the potential to greatly increase resolution in difficult regions of the genome for all types of genetic variation.

References

Analysis:

- pbaa: https://github.com/PacificBiosciences/pbAA
- Variant calling script: https://github.com/jrharting/CoSA. See vcf/consensusVariants.py

Samples and Variants

- Coriell https://www.coriell.org/
- Clinvar https://www.ncbi.nlm.nih.gov/clinvar/

References

[1] Baumgartner-Parzer, S., Witsch-Baumgartner, M. & Hoeppner, W. EMQN best practice guidelines for molecular genetic testing and reporting of 21-hydroxylase deficiency. *Eur J Hum Genet* **28**, 1341–1367 (2020). https://doi.org/10.1038/s41431-020-0653-5

[2] Chen W, Xu Z, Sullivan A, et al. Junction site analysis of chimeric CYP21A1P/CYP21A2 genes in 21-hydroxylase deficiency. *Clin Chem*. 2012;58(2):421-430. doi:10.1373/clinchem.2011.174037