Resolving complex pathogenic alleles using HiFi long-range amplicon data and a new clustering algorithm



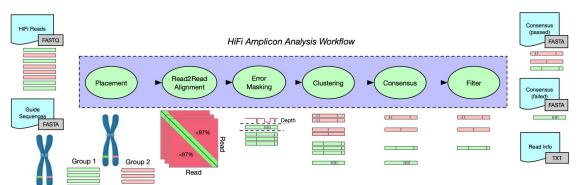
J. Harting¹, C. Heiner¹, I. McLaughlin¹, Z. Kronenberg¹

¹Pacific Biosciences, Menlo Park, CA, USA

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2021 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. All other trademarks are the sole property of their respective owners.

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 call a wide range of complex disease-causing variants in two gene-pseudogene systems, CYP21A2 and GBA, by applying a single multiplexed long-range paired assay to each target.

https://github.com/PacificBiosciences/pbAA





TARGETS AND ANALYSIS METHODS

Targets

Co-Amplified Targets

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

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

- Samples with known difficult genotypes from Coriell
- Primers designed to include full-length genes, pseudogenes, and flanking regions for comprehensive variant detection.
- Sequencing on PacBio Sequel Systems.

 See P15.015.C - Targeting clinically significant dark regions of the human genome with high-accuracy, longread sequencing for details on experimental design and sequencing.

pbAA

pbaa consensus

Unique key for each amplicon (uuid)

Sample/Amplicon metadata

Variants

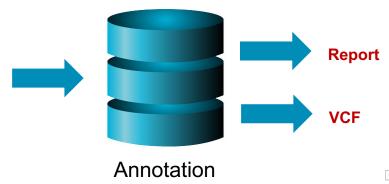
Key: uuid

CHR,POS,VAR,
Support

Variant Calling

Analysis

- Demultiplex and identify primers
- Deconvolute and generate consensus with pbaa
- Call variants from consensus
- Import to annotation database
 - Gene features (e.g. gene and exons coordinates)
 - ClinVar annotations
 - Call large SV
- Export Report and/or VCF





RESULTS: CYP21A2 / CYP21A1P

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	Homo SNV & Fusion	3
NA14732	Carrier	✓	1	CYP21A2 Deleted	3
NA14733	Carrier	Tio V	1	Gene Fusion	2
NA14734	Affected	✓	0	Deletion & Fusion	2

	31,970 Nb	31,989 No	47 kb	32,000 kb	32,910 kb
[LU2	LD1		N602341 N60264 N60262 N11791 N41277 N14732 N14732 N14732	RD2
NA02241					
NA02242					II
NA11781					II II I
NA12217					
NA14732					
NA14733				l .	
NA14734				·	
B		21A1P TNOVA	C48	CYP21A2	CYP21A2 TNXB

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





Called ClinVar Pathogenic SNV

RESULTS: GBA / GBAP1

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

							4		
[LU3					LD4	NA00973 NA00977 NA00977 NA00977 NA00977 NA00973 NA0097		RD2
NA00852 cmplx het								1 1 1 1	
NA00877 hom path									
NA00878 carrier	1=								
NA01031 carrier	11							<u> </u>	
NA01260 empty het									
NA01607 emptx het									
NA02627 cmplx het					1				
NA08752 hom path/8V	I in				-11				
NA08753 carrier SV									
NA10873 hom path									
NA10674 cmplx het									
NA20270 cmplx het									
NA20273 hom path	1				ļ.		•		
₽ ■	MTX1	 	GBAP1	GBAP'	1		-11	GBA GBA	GBA

- 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.







CONCLUSION

PacBio HiFi reads, with the features shown here, have the potential to greatly increase resolution in difficult regions of the genome for all types of genetic variation:

- Comprehensive variant detection
- SNV and multi-kb structural variants
- Accurate and reproducible

- Phased results
- Uniquely map-able to gene or pseudogene
- Single targeted assay

HiFi reads and pbaa capture 3 unique SV in a single family, including a duplicate pseudogene that differs by a single SNV in an 8kb amplicon

