Targeting Clinically Significant Dark Regions of the Human Genome with High-Accuracy, Long-Read Sequencing

Abstract #: eP273
Ian McLaughlin, John Harting, Lori Aro, Cheryl Heiner
Pacific Biosciences, 1305 O'Brien Drive, Menlo Park, CA 94025

Introduction

There are many clinically important genes in "dark" regions of the human genome characterized by a paucity of NGS coverage as a result of short-read sequencing or mapping difficulties. Low NGS sequencing yield can arise in these regions due to the presence of various repeat elements or biased base composition while inaccurate mapping is attributable to segmental duplications. Long-read sequencing coupled with an optimized, robust enrichment method has the potential to illuminate these dark regions.

Methods and Materials

Long-range PCR targeted enrichment was combined with PacBio long-read, high-accuracy (HiFi) sequencing and a novel amplicon analysis tool (pbaa) to develop individual prototype screening assays for *CYP21A2* and *GBA*: genes associated with congenital adrenal hyperplasia and Gaucher disease, respectively. Each gene is difficult to accurately type with short-read sequencing due to proximity to and interaction with a highly homologous pseudogene.

Gene/Pseudogene Co-amplification (3 or 6 hr)
- Platinum SuperFi II PCR Master Mix (Thermo Fisher Scientific) plus 0.8M betaine
- Barcoded primers paired asymmetrically
- 7 Coriell samples amplified in replicate for *CYP21A2* and *CYP21A1P* for 24 samples total
- 13 Coriell samples amplified in replicate for *GBA* and *GBAP1* for 24 samples total
- 2-step cycling protocol: 2hr 41min for *CYP21A2* assay; 6hr 10min for *GBA* assay
- 50 ng sample input

QC and Pool Samples (1 hr)
- For each assay:
  - 24 samples quantified and pooled by equal mass
  - Pooled samples AMPure PB bead-purified with 0.45X bead-to-sample volume ratio
  - Pooled samples quantified (1X dsDNA HS Assay Kit [Thermo Fisher Scientific])

SMRTbell Library Construction (3.5 hr)
- For each library:
  - SMRTbell Express Template Prep Kit 2.0 (PacBio) using overhang adapters
  - 1,000 ng pooled sample input
  - 2 separate libraries produced: one for *CYP21A2* screening assay and one for *GBA* screening assay
  - Libraries AMPure PB bead-purified two times with 0.45X bead-to-sample volume ratio

Sequencing Preparation (3 hr)
- For each sequencing complex:
  - Sequencing Primer v4 (PacBio; 20-fold excess over template concentration)
  - Sequel II Binding Kit 2.0 (PacBio; 10-fold polymerase excess over template concentration)
  - 20-hr movie time on Sequel II System (also compatible with Sequel IIe System)

Figure 1. Workflow for *CYP21A2* and *GBA* screening assays. Two days are required starting from purified gDNA to have sequencing complexes ready to sequence. An additional two days are required to sequence the libraries and produce variant-typed results.

A *CYP21A2* Screening Assay

Cycle Step	Temperature	Time	Cycles
Initial Denaturation	98°C	30 sec	1
Denature	98°C	10 sec	26
Annealing/Extension	68°C	5 min	
Final extension	72°C	10 min	1

B *GBA* Screening Assay

Cycle Step	Temperature	Time	Cycles
Initial Denaturation	98°C	30 sec	1
Denature	98°C	10 sec	28
Annealing/Extension	68°C	12 min	
Final extension	72°C	12 min	1

Table 1. PCR cycling conditions for *CYP21A2* (A) and *GBA* (B) screening assays. Annealing/extension temperature and time as well as number of cycles were optimized for each assay to minimize PCR artifacts such as heteroduplex amplicons and self-primed amplicons.

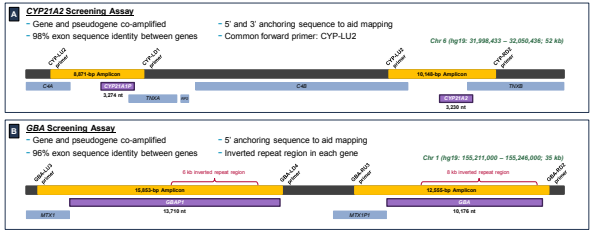

Figure 2. PCR target maps for *CYP21A2* (A) and *GBA* (B) screening assays. Primer sequences were developed in-house. Primer reaction concentrations were adjusted to achieve an approximately balanced yield of co-amplified amplicons in each assay system. Both systems benefited from supplemental betaine at a final concentration of 0.8M, especially the *GBA* system with a large tract of inverted repeats in each gene likely leading to stable secondary structures in the absence of supplemental betaine.

Results

Assay	PCR Input	PCR Yield (mean)	Library Input	Library Yield	HiFi Reads (gQ20)	%Barcoded HiFi Reads	%Full-length Aligned Reads
CYP21A2	50 ng	431 ± 131 ng	1 µg	671 ng	1,520,782	84% (1,279,425)	84% (1,276,879)
GBA	50 ng	1,086 ± 304 ng	1 µg	673 ng	1,045,191	77% (801,843)	81% (847, 538)

Table 2. Top-line performance metrics for *CYP21A2* and *GBA* screening assays. A high percentage of HiFi reads are barcoded and produce full-length alignments to the target genes.

Gene	Aligned Reads (mean)	CV	Min	Max
CYP21A1P	17,226	71%	526	41,817
CYP21A2	27,471	41%	985	45,696
GBAP1	8,065	51%	3,866	21,355
GBA	22,165	23%	10,341	31,652

Table 3. Mean number of aligned reads per sample for each target gene. The difference in yield between target genes in each assay reflects known PCR yield differences. The high CV level reflects the presence of samples with alleles containing deletions ultimately reducing the number of reads for those samples.

A

Sample	Expected Gene Amplification: *CYP21A1P*	*CYP21A2*	Observed Gene Alignment: *CYP21A1P*	*CYP21A2*	*CYP21A1P*	*CYP21A2*
02241	15,053	326	161	38,273		
02242	17,805	378	116	37,845		
11781	33,738	168	313	30,757		
12217	27,662	73	16,414	17,085		
14732	895	411	20,927	20,728		
14733	14,989	353	22,491	22,168		
14734	870	1	46,529	27		

B

Sample	Expected Gene Amplification: *GBAP1*	*GBA*	Observed Gene Alignment: *GBAP1*	*GBA*	Outer Primers LU3-RD2 *GBAP1*	*GBA*
00852	9,360	1	6	28,984	147	175
00877	6,145	9	16	23,416	474	453
00878	4,939	18	54	22,317	908	991
01031	5,324	20	33	25,102	706	699
01260	6,938	18	18	20,358	443	492
01607	5,805	10	30	24,768	815	921
02627	8,569	5	13	27,590	475	318
08752	5,839	8	50	19,438	20,249	20,260
08753	5,279	32	142	12,894	11,717	11,784
10873	7,119	13	31	21,249	865	961
10874	8,247	2	9	26,519	297	356
20270	10,090	4	5	24,745	0	0
20273	4,667	2	25	26,531	724	809

Table 4. Mean number of aligned reads per sample replicate for *CYP21A2* (A) and *GBA* (B) screening assays. Highlighted values in purple indicate numbers significantly different than others within their respective columns. *CYP21A2* samples 12217, 14732, 14733, and 14734 contain *CYP21A2* deletion alleles which alter the number of expected reads (e.g., sample 14734 contains a full deletion and partial deletion of *CYP21A2*, forming fusion alleles between *CYP21A1P* and partial *CYP21A2* or sequence downstream of *CYP21A2*). *GBA* samples 08752 and 08753 contain *GBA* deletion alleles but due to the priming strategy the resulting fusion product requires detection with the outer-most primers, LU3 and RD2. For each assay, the presence of a low number of mismatched aligned reads to expected amplicons is due to the high level of homology between gene and pseudogene target regions.

A *CYP21A2* Variants

Sample	Allele 1	Allele 2	Family	Relation
02241	IVS2AS,A/C>G,-13	-	85	proband
02242	IVS2AS,A/C>G,-13	-	85	proband
11781	IVS2AS,A/C>G,-13	IVS2AS,A/C>G,-13	-	proband
12217	30-KB DEL	ILE172ASN	-	proband
14732	Full DEL	-	1787	mother
14733	30-KB DEL	-	1787	father
14734	Full DEL	30-KB DEL	1787	proband

B *GBA* Variants

Sample	Allele 1	Allele 2	Family	Relation
00852	ASN370SER	1-BP INS, 84G	2976	proband
00877	LEU444PRO; ALA456PRO; VAL460VAL	LEU444PRO	57	proband
00878	LEU444PRO; ALA456PRO; VAL460VAL	-	57	mother
01031	IVS2DS, G>A, +1	-	250	parent
01260	LEU444PRO	PRO415ARG	-	proband
01607	ASN370SER	VAL394LEU	-	proband
02627	GLY325ARG	CYS342GLY	-	proband
08752	LEU444PRO	LEU444PRO	2089	proband
08753	LEU444PRO	-	2089	mother
10873	ASN370SER	ASN370SER	-	proband
10874	ASN370SER	ARG119GLN	-	proband
20270	LEU444PRO	IVS2DS, G>A, +1	3248	proband
20273	RECOMBINANT (Exon 9)	RECOMBINANT (Exon 9)	-	proband

Table 5. Coriell variant information for samples tested in the *CYP21A2* (A) and *GBA* (B) screening assays. All mutations above were identified, including single nucleotide variants, whole-gene deletions, and recombinant exons. In addition, unexpected gene duplications and gene fusions were identified, and phasing of compound heterozygotes was achieved. These results were made possible by long-read, HiFi sequencing (see poster eP240, Resolving Complex Pathogenic Alleles using HiFi Long-range Amplicon Data and a New Clustering Algorithm for additional information on variant detection for these samples and the new amplicon analysis tool, pbaa).

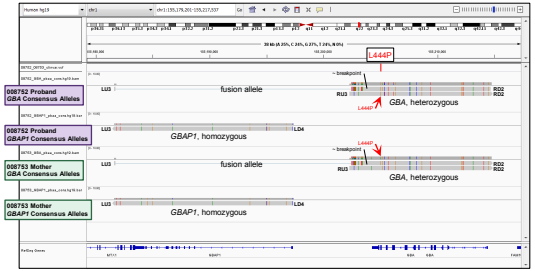

Figure 3. Detection of deletion alleles, gene fusions, and an SNV in *GBA* compound heterozygote samples 08752 and 08753. The deletions and gene fusions were not described in the Coriell database, only the L444P pathogenic variant. Primers are identified at the ends of each aligned read, and the L444P mutation is indicated with a red arrow. The fusion allele is created by a large deletion encompassing *GBA* and *GBAP1* genes.

Conclusions

- We demonstrate that HiFi sequencing provides new opportunities for sequencing clinically relevant but previously dark regions of the human genome that are underrepresented in short-read sequencing.
- The *CYP21A2* and *GBA* screening assays achieved accurate typing results for all samples and provided additional variant information due to the use of optimized and robust long-range PCR enrichment, HiFi sequencing, and newly developed PacBio amplicon analysis tool, pbaa.
- Accurate long reads provide important phasing information, identify structural variations, and avoid potential confusion with pseudogenes.
- The PCR, library preparation, and sequencing preparation steps of these assays are amenable to automation and a cost-effective workflow enabling high-throughput screening.
- HiFi sequencing of these regions enables a better understanding of the relationship between genetic factors and personal health and has the potential to ultimately help guide health-related decisions.