Targeting clinically significant dark regions of the human genome with high-accuracy, long-read sequencing

I. McLaughlin[1], J. Harting[1], Z. Kronenberg[1], Lori Aro[1], C. Heiner[1]
[1]Pacific Biosciences, Menlo Park, CA, USA

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.

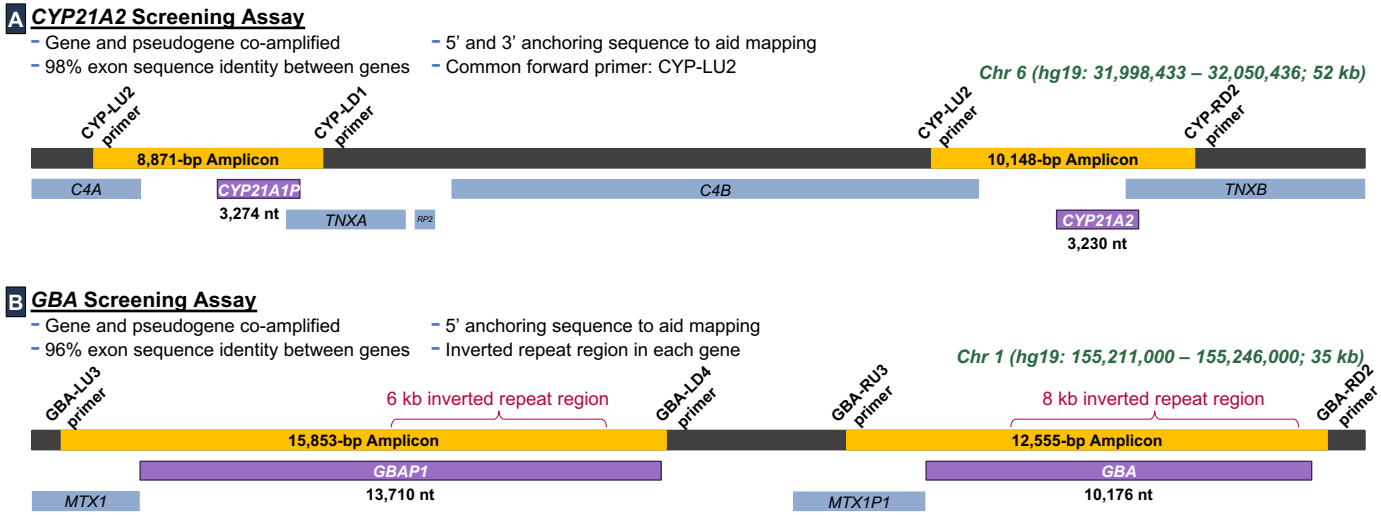

Figure 1. PCR target maps for *CYP21A2* (A) and *GBA* (B) screening assays. Each “dark” gene is difficult to accurately type with short-read sequencing due to proximity to and interaction with a highly homologous pseudogene. Inverted repeat regions in *GBA* pose additional target enrichment and analysis challenges.

METHODS

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.

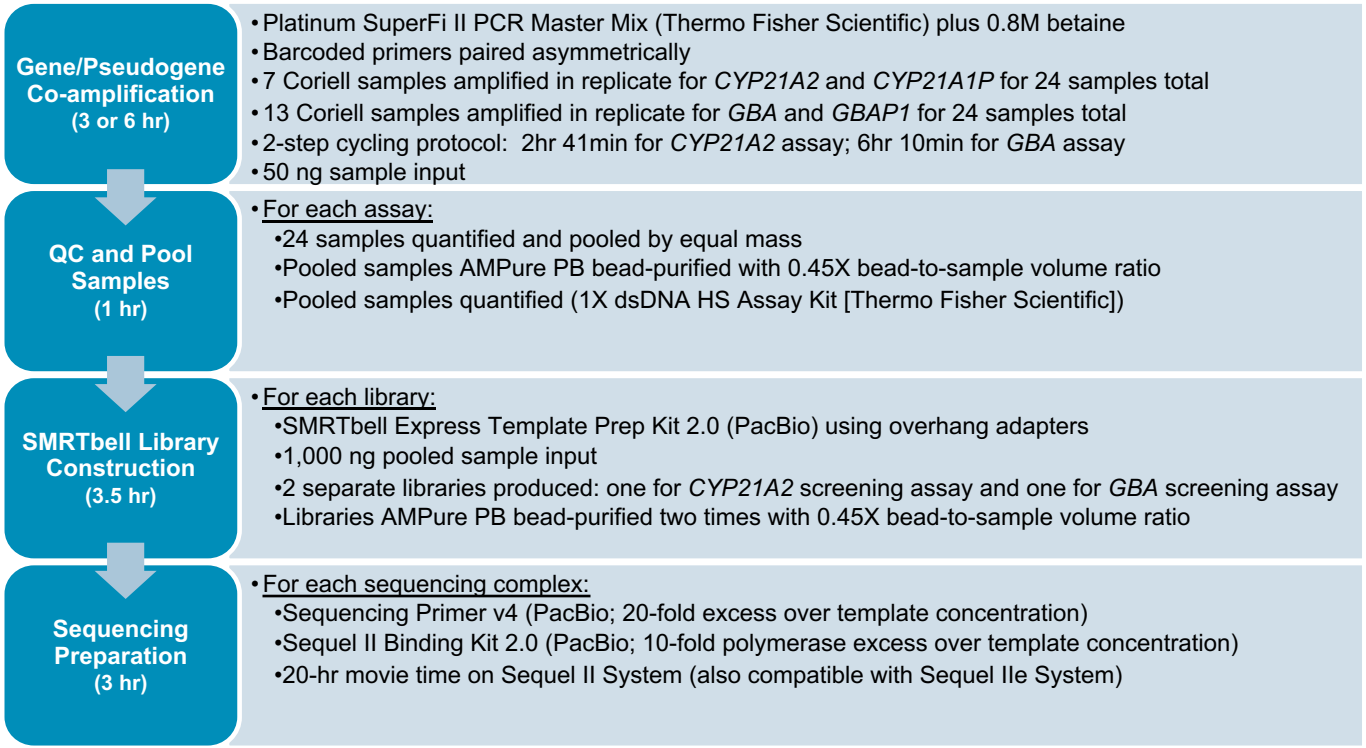

Figure 2. Workflow for *CYP21A2* and *GBA* screening assays. Two days are required starting from purified gDNA to have sequencing complexes ready to sequence. An additional two days are required to sequence the libraries and produce variant-typed results. Please see e-poster P17.028.C, Resolving complex 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 algorithm, **pbaa**).

RESULTS

Table 1. 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.

Assay	PCR Input	PCR Yield (mean)	Library Input	Library Yield	HiFi Reads (≥Q20)	%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. Mean number of aligned reads 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, reducing the number of reads for those samples.

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 for each sample in *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 for gene and pseudogene (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

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*		Outer Primers LU3-RD2	
	Observed Gene Alignment: *GBAP1*	*GBA*	*GBAP1*	*GBA*	*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

RESULTS

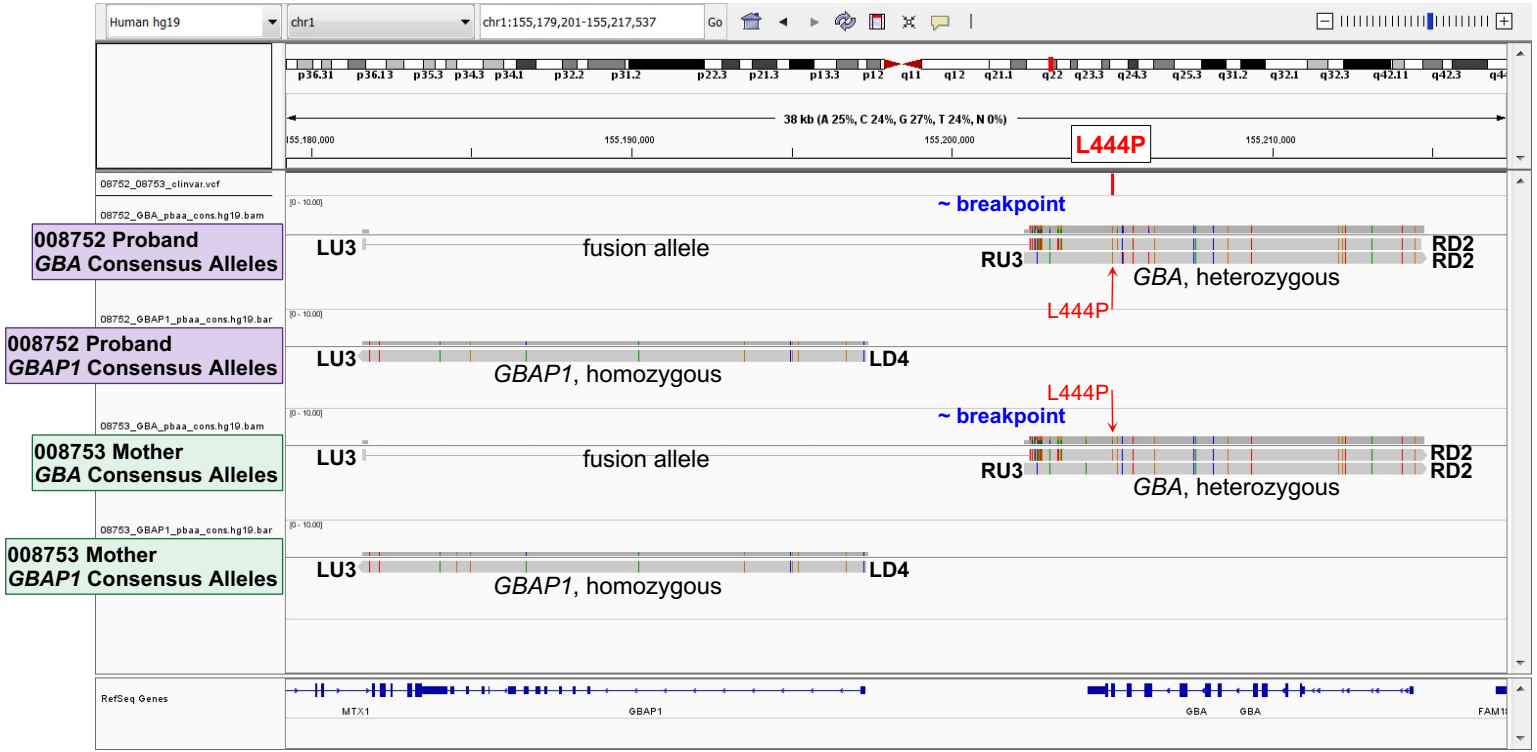

Figure 3. Detection of deletion alleles, gene fusions, and an SNV in related *GBA* compound heterozygote samples 08752 and 08753. The deletions and gene fusions were not described in the Coriell database, only the L444P pathogenic variant. Primers are identified at the ends of each aligned read, and 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.