AMP 2017: G19

Detecting Pharmacogenomic Variants Using Long- and Short-Read Next Generation Sequencing Platforms

Cassie Schumacher, Ashley Wood, Sukhinder Sandhu, Justin Lenhart, Laurie Kurihara, Timothy Harkins, Vladimir Makarov

Swift Biosciences, 58 Parkland Plaza, Suite 100, Ann Arbor, MI 48103

Abstract

Introduction: The key processes underlying the field of pharmacokinetics are absorption, distribution, metabolism, and excretion (ADME). Screening for pharmacogenomic markers is often required to ensure safe and effective disease treatment. Despite this requirement, problems including high homology with known pseudogenes and difficult to sequence motifs arise when trying to sequence ADME genes, like CYP2D6. Most next generation sequencing (NGS) assays utilize short-read (SR) chemistry that enables genotyping of known biomarkers in accessible parts of the genome. Newer long-read (LR) assays provide comprehensive sequencing of an entire gene, providing insight into copy number variations, but are limited to fewer genes per assay. Massively parallel screening using either NGS assay is cost effective, especially compared with traditional methods such as real-time PCR. Here we show the utility of these two assays to characterize ADME biomarkers. The SR sequencing assay detected hotspots in CYP2D6 as well as 150 additional genes for germline SNP genotyping. The LR sequencing assay surveyed the entire CYP2D6 gene for a full and detailed look at CYP2D6 and its associated variants.

Methods: 24 DNA samples from the Coriell Institute with known CYP2D6 variants were used in each assay. In the SR assay, an amplicon-based panel targeting 364 ADME-specific targets was performed. 10ng of each DNA was used as input for a multiplexed PCR, and the products were subsequently adapted for Illumina sequencing. These libraries were sequenced on an Illumina® MiniSeq® using a paired end read length of 151bp. In the LR assay, less than 100ng of each DNA was used as the input for long range PCR, and the 6.5Kb PCR products were subsequently barcoded for sample multiplexing and adapted for Pacific Biosciences® (PacBio) sequencing. These libraries were then sequenced on a PacBio RSII.

Results: Libraries from the SR assay were sequenced with >90% on-target and coverage uniformity. 92% of known CYP2D6 variants, among other ADME variants, were detected using this technique. Using the LR technique, all known CYP2D6 variants could be detected. Importantly, 100% of the CYP2D6 gene could be covered by this assay.

Conclusions: Having the right tool to address questions concerning pharmacogenomic profiles is critical to ensuring proper, personalized treatments. Using the short-read assay, we demonstrate an ability to screen for a wide range of germline ADME targets. Such an assay takes advantage of the current knowledge bank and is informative for immediate and broad use. With the long-range assay, we show the ability to comprehensively sequence the entire CYP2D6 gene, thereby gaining more insight into CYP2D6 variants and phenotypes and further informing personalized treatment options.

Dual Platform Approach

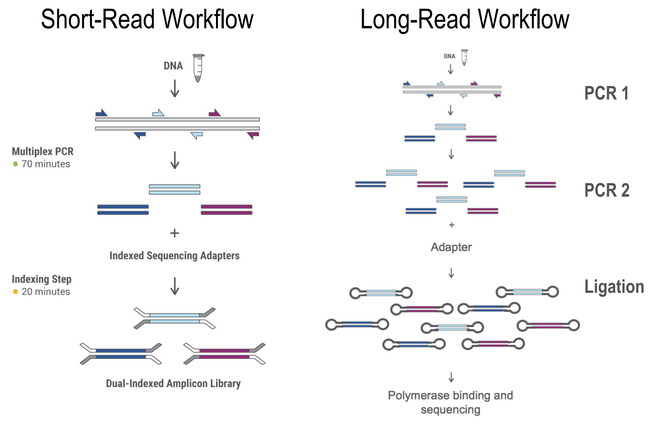

Figure 1. Short- and long-read workflows. (Left) The short-read workflow consists of two steps: multiplexed PCR followed by an indexing step to barcode and adapt the PCR products for Illumina sequencing. (Right) The long-read workflow consists of three steps, two PCR steps to amplify the target sequences and a ligation step to barcode and adapt the PCR products for PacBio sequencing.

Panel Specifications

C

Panel	# of Amplicons	Amplicon Size (kb)	Panel Size (kb)	% CYP2D6 Covered Exonic/Total	Tiling?
ADME - Short	364	0.15	40	73/39	Yes
CYP2D6- Long	1	6.5	6.5	100/100	No

Figure 2. A) Integrative Genome Viewer (IGV, Broad Institute) genome-wide view of the 364 short- panel targets. B) IGV view of short- and long-read panel target coverage of the CYP2D6 gene. C) Short- and long-read panel specifications.

Sequencing Metrics

CYP2D6						
	# Aligned Reads		% On Target		Coverage Uniformity	
DNA	Short-Read	Long-Read	Short-Read	Long-Read	Short-Read	Long-Read
NA17084	1604933	715	94.1	99.8	91.5	100
NA17221	1786760	535	94.3	99.7	91.5	100
NA17205	1361074	632	93.6	99.8	92.2	100
NA17293	1372737	507	94.3	99.7	91.9	100
NA17230	1752193	631	93.3	99.8	92.3	100
NA12244	1728736	538	93.9	99.7	92.0	100
NA17272	1829503	663	93.0	99.8	91.8	100
NA17039	2421000	334	93.3	99.7	93.1	100
NA17276	2024712	254	93.0	99.8	92.2	100
NA17281	1854698	285	93.8	99.8	91.2	100
NA17204	1925657	593	93.3	99.7	93.1	100
NA17300	1840144	441	92.8	99.8	92.7	100
NA17245	1593169	360	93.3	99.8	91.5	100
NA17252	1830937	694	93.2	99.8	93.4	100
NA17280	1776664	332	94.2	99.7	91.6	100
NA02016	9744941	329	94.5	99.7	92.4	100
NA16688	1772670	616	93.7	99.8	93.7	100
NA17291	1666463	646	93.8	99.8	92.2	100
NA17130	4065337	258	94.4	99.7	91.6	100
NA17227	1537362	268	94.4	99.8	90.5	100
NA17246	3661637	528	93.3	99.8	92.4	100
NA17114	1945540	478	92.8	99.8	92.9	100
NA10005	2373805	307	93.9	99.8	91.5	100

Figure 3. Sequencing metrics for short- and long-read. Tables list the number of aligned reads, percent of reads on target, and the coverage uniformity in both assay types. All DNA samples were obtained from the Coriell Institute. All short-read samples were sequenced on an Illumina MiniSeq; while long-read samples were sequenced on a PacBio RSII. Aligned reads for all long-read samples are derived from CCS read numbers.

Sequencing Challenges with Pseudogenes

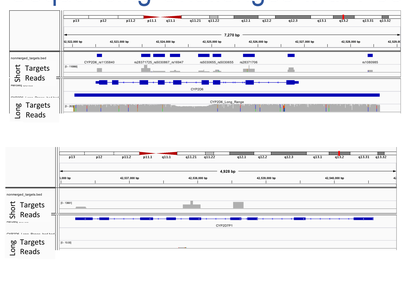

Figure 4. Alignment challenges with pseudogenes. Short-read sequencing poses a significant challenge when it comes to genes with known pseudogenes. Even when care has been taken to design primers as specifically to unique regions as possible, reads can still align to the pseudogene. Long-read technology has the ability to include much more unique content and additional bases, therefore all of the reads align on target. The figures show short-read and long-read targets and aligned reads to CYP2D6 (top) and the CYP2D7 pseudogene (bottom) for NA02016.

Variant Analysis Using Short- and Long-Read Technology

CYP2D6					
				Observed AF	
DNA	Haplo-type	Variant	Expected AF	Short-Read	Long-Read
NA17084	*1/*10	P34S	0.5	0.68	0.51
		S486T	0.5	0.52	0.57
		R296C	0.5	0.69	0.47
NA17205	*1/*41	R296C	0.5	0.45	0.53
		S486T	0.5	0.49	0.55
NA17230	*4/*41	P34S	0.5	0.54	0.47
		L91M	0.5	0.5	0.48
		H94R	0.5	0.23	0.4
		S486T	1	1	1
		R296C	0.5	0.49	0.53
NA17272	*4/*10	P34S;	1	1.00	0.96
		L91M	0.5	0.5	0.5
		H94R	0.5	0.34	0.45
		S486T	1	1	1
NA17269	*2/*41	R296C	1	1	1
		S486T	1	1	1
NA17204	*1/*35	V11M	0.5	0.51	0.46
		R296C	0.5	0.57	0.47
		S486T	0.5	0.52	0.49
NA17300	*1/*6	118FS	0.5	0.54	0.45
NA17252	*4/*5	S486T	1	1	0.98
		P34S	1	0.97	0.95
NA16688	*2/*10	R296C	0.5	0.66	0.45
		S486T	1	1	1
		P34S	0.5	0.66	0.53
NA17291	*1/*4	P34S	0.5	0.50	0.46
		L91M	0.5	0.54	0.45
		H94R	0.5	0.27	0.38
		S486T	0.5	0.50	0.44
NA17246	*4/*35	P34S	0.5	0.65	0.5
		L91M	0.5	0.50	0.52
		H94R	0.5	0.49	0.45
		S486T	1	1	0.99
		V11M	0.5	0.35	0.47
		R296C	0.5	0.53	0.47
NA17114	*1/*5	WT	-	-	-

CYP2D6					
				Observed AF	
DNA	Haplo-type	Variant	Expected AF	Short-Read	Long-Read
NA17039	*2/*17	R296C	1	1	0.99
		S486T	1	1	1
		T107I	0.5	0.51	0.54
NA17276	*2/*5	R296C	1	1	1
		S486T	1	1	1
NA17281	*5/*9	K281del	1	1	0.98
NA17245	*2/*4	R296C	0.5	0.55	0.45
		S486T	1	1	1
		P34S	0.5	0.69	0.56
		L91M	0.5	0	0.57
		H94R	0.5	0.24	0.5
NA17280	*2/*3	R296C	0.5	0.49	0.5
		S486T	0.5	0.51	0.54
		259FS	0.5	0.50	0.48
NA02016	*2XN/*17	R296C	1	1	1
		S486T	1	1	0.99
		T107I	0.5	0.37	0.36
NA17130	*1/*2	R296C	0.5	0.48	0.44
		S486T	0.5	0.54	0.48
NA17227	*1/*9	K281del	0.5	0.51	0.56
NA10005	*17/*29	T107I	0.5	0.50	0.49
		R296C	1	1	1
		S486T	1	1	1
		V136I	0.5	0.54	0.48
		V338M	0.5	0.54	0.52
NA17293	*2/*9	R296C	0.5	0.49	0.45
		S486T	0.5	0.55	0.49
		K281del	0.5	0.56	0.55
NA17221	*1XN/*2	R296C	0.5	0.65	0.34
		S486T	0.5	0.61	0.39
NA12244	*35/*41	V11M	0.5	0.50	0.5
		R296C	1	1	1
		S486T	1	1	1

Figure 5. Variant calling across panels and platforms. Known variants for all 24 samples are detailed. The expected allele frequency (AF) is given, along with the observed AF in both the short-read and long-read assay. Despite the off-target alignment issues identified in Figure 4, only one known variant was missed in one sample using the short-read assay. All known variants were called using the long-read assay.

Conclusion

- Accel-Amplicon™ panels from Swift Biosciences can be used on short- and long-read sequencing technologies for variant detection.
- Short-read amplicon panels are useful tools to interrogate variants of known significance in coding regions and intron/exon boundaries.
- Long-read amplicon panels are useful for full gene coverage to not only analyze variants in the coding region, but to also probe neighboring introns which can be difficult to target with short-read amplicons due to repetitive regions and low complexity motifs.
- Long-read technology provides more accurate alignment to identify structural variants and can overcome pseudogene alignment artifacts.
- Long-read technology enables phasing of the CYP2D6 gene within the 6.5 kb single amplicon.
- Both short- and long-read amplicon panel workflows can be completed in a single day and have the ability to process samples in a high-throughput manner.

Swift Biosciences, Inc.
58 Parkland Plaza, Suite 100 • Ann Arbor, MI 48103 wwwswiftbiosci.com

www.swiftbiosci.com