TLA & Long-Read Sequencing: Efficient Targeted Sequencing and Phasing of the CFTR Gene

Max van Min[1], Erik Splinter[1], Vera Boersma[1], Marieke Simonis[1], Karima Hajo[1], John Harting[2], Lori Aro[2], Janet Ziegle[2], and Cheryl Heiner[2]

[1]Cergentis, Utrecht, The Netherlands, and [2]Pacific Biosciences, Menlo Park, CA,

Abstract

Background

The sequencing and haplotype phasing of entire gene sequences improves the understanding of the genetic basis of disease and drug response. One example is cystic fibrosis (CF). Cystic fibrosis transmembrane conductance regulator (CFTR) modulator therapies have revolutionized CF treatment, but only in a minority of CF subjects. Observed heterogeneity in CFTR modulator efficacy is related to the range of CFTR mutations; revertant mutations can modify the response to CFTR modulators, and other intronic variations in the ~200 kb CFTR gene have been linked to disease severity. Heterogeneity in the CFTR gene may also be linked to differential responses to CFTR modulators.

The Targeted Locus Amplification (TLA) technology from Cergentis can be used to selectively amplify, sequence and phase the entire CFTR gene. With PacBio long-read SMRT Sequencing, TLA amplicons are sequenced intact and long-range phasing information of all fragments in entire amplicons is retrieved.

Experimental Design and Methods

The TLA process produces amplicons consisting of 5-10 proximity ligated DNA fragments. TLA was performed on cell line and genomic DNA from Coriell GM12878, which has few heterozygous SNVs in CFTR, and the IB3 cell line, with known haplotypes but heterozygous for the delta508 mutation. All sample types were prepared with high and low density TLA primer sets, targeting coverage of >100 kb of the CFTR gene.

Conclusion

We have demonstrated the power and utility of TLA with long-read SMRT Sequencing as a valuable research tool in sequencing and phasing across very long regions of the human genome. This process can be done in an efficient manner, multiplexing multiple genes and samples per SMRT Cell in a process amenable to high-throughput sequencing.

TLA Technology

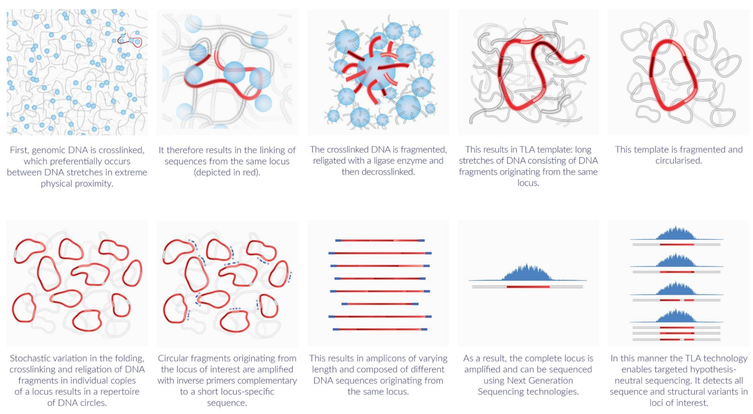


Figure 1. Targeted Locus Amplification (TLA) workflow

Workflow

Figure 2. TLA amplicon pool

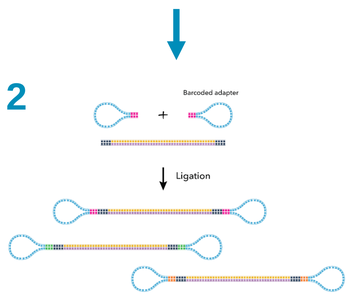
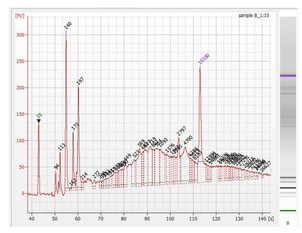


Figure 3. SMRTbell library preparation

TLA amplicons of varying length were generated for SMRTbell library preparation. Each amplicon included several different DNA sequences in the gene of interest, originating from the same allele.

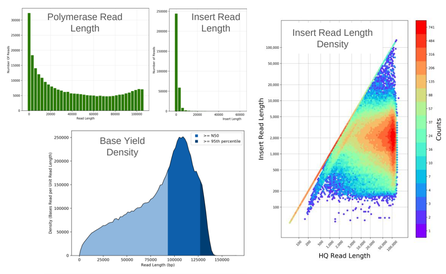
Bioanalyzer QC of TLA samples indicates a broad range of fragment sizes, some >10 kb.



One SMRTbell library was prepared from 10 samples using barcoded adapters, following standard protocols.

Long-read sequencing on Sequel System generated >300,000 reads

Results from 1 SMRT Cell 1M	
Total Bases	16.5 Gbases
Unique Molecular Yield	610 Mbases
Polymerase Read Length	52,406 (average); 93,784 (N50)
Subread Length	2,380 (average); 3,094 (N50)
Primary Sequencing Reads	320,343



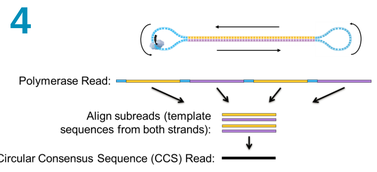
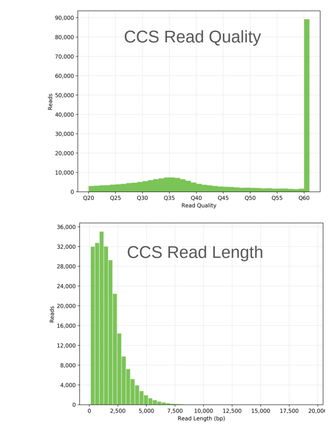


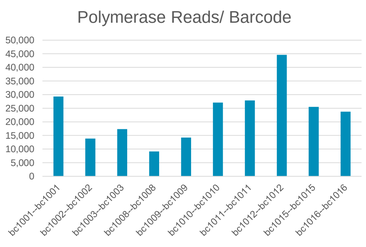
Figure 4. Generation of CCS Reads

SMRT Link analysis produced 232,000 barcoded, high accuracy single molecule (CCS) reads, followed by demultiplexing.



CCS and demultiplexing results from one Sequel SMRT Cell 1M

CCS Analysis	
≥Q20 Reads	232,788
≥Q20 Yield (bp)	396,276,923
≥Q20 Read Length Average (bp)	1,702
≥Q20 Read Quality (median)	Q44



Barcodes	
Unique Barcodes	10
Barcoded Reads	232,599
Average Reads / barcode	23,259
Maximum Reads / barcode	44,610
Minimum Reads / barcode	9,127
Mean Read Length	1,670
Unbarcoded Reads	189

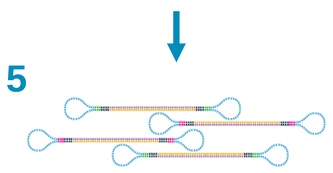


Figure 5. IB3 cell line library pool

Additional data was gathered on IB3 cell line samples. IB3 cell 29plex data was used for the subsequent haplotype analysis.

Barcode	Sample	Total CCS Reads
BC1011	IB3 cell 29plex	242,116
BC1012	IB3 cell 6plex	318,076

Results

From IB3 cell line data, phasing information was generated for 307 SNVs across a 711 kb region spanning the entire CFTR gene and surrounding sequences. Within the CFTR gene, 95% of the 190 detected SNVs were phased.

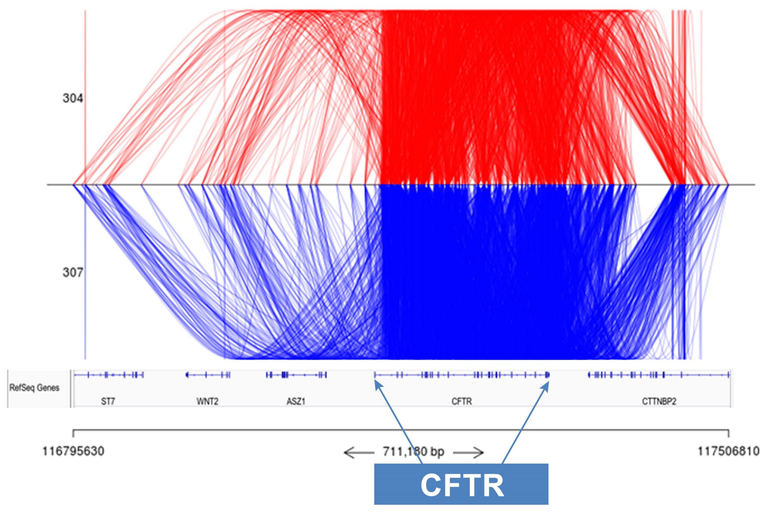


Figure 6. TLA haplotyping of a 711 kb region around the CFTR gene: results from IB3 cell 29-plex data, with phasing of 307 SNVs.

Based on the obtained haplotypes, CCS reads were assigned to alleles, resulting in **complete sequencing information of both individual alleles of the entire 200 kb CFTR gene**.

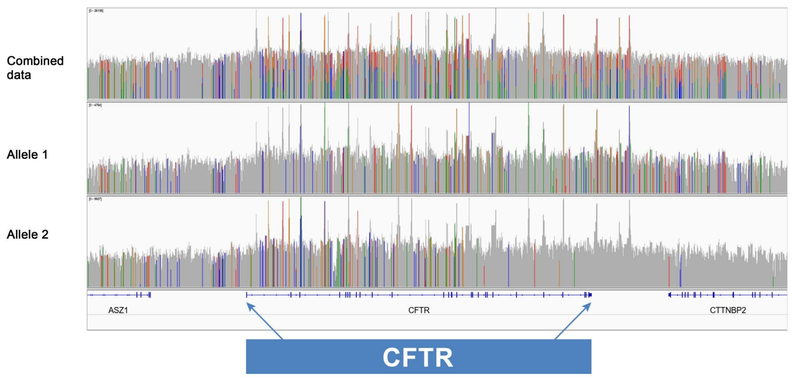


Figure 7. Complete sequence for each allele across the CFTR gene in the IB3 cell sample, based on 29-plex TLA results. Of 107 heterozygous SNVs detected in the CFTR gene by *Vermeulen et al*[1], all are present in the data. Additional SNVs not present in dbSNP were also detected.

[1]Vermeulen et al., Sensitive Monogenic Noninvasive Prenatal Diagnosis by Targeted Haplotyping, The American Journal of Human Genetics (2017), http://dx.doi.org/10.1016/j.ajhg.2017.07.012

CFTR exon and whole gene coverage per allele

Allele	Genomic Region	Median Coverage	>1-Fold Coverage	>10-Fold Coverage
Allele 1	CFTR exons	27x	100%	95%
Allele 1	CFTR whole-gene	28x	99%	89%
Allele 2	CFTR exons	66x	100%	100%
Allele 2	CFTR whole-gene	53x	100%	97%

CCS read coverage for each allele averaged >25-fold, with ≥99% of the entire 200 kb gene region covered more than once per allele, and ≥89% of the whole gene showing >10-fold coverage per allele.

Conclusions

By combining Cergentis' TLA technology with PacBio's long-read sequencing, we achieved targeted sequencing with phasing of an entire 200 kb gene.

With >200,000 high-accuracy CCS reads from TLA amplicons on one Sequel SMRT Cell 1M, having an average read quality of Q44, this process can achieve efficient phasing across long targeted physical distances.