Sequencing and *De Novo* Assembly of the 17q21.31 Disease Associated Region Using Long Reads Generated by Pacific Biosciences® SMRT® Sequencing Technology

Swati Ranade[1], Can Alkan[2], Francesca Antonacci[2], Aaron Klammer[1], Lawrence Hon[1], Eric Schadt[1] and Evan Eichler[2]
[1]Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025, USA; [2] University of Washington, Seattle, WA 98105,USA

Summary

Assessment of genome-wide variation revealed regions of the genome with complex, structurally diverse haplotypes that are insufficiently represented in the human reference genome. The 17q21.31 region is one of the most dynamic and complex regions of the human genome. Different haplotypes exist, in direct and inverted orientation, showing evidence of positive selection and predisposing to microdeletion associated with mental retardation. Sequencing of different haplotypes is extremely important to characterize the spectrum of structural variation at this locus. However, *de novo* assembly with second-generation sequencing reads is still problematic.

Using PacBio® technology we have sequenced and *de novo* assembled a tiling path of eight BAC clones (~1.6 Mb region) across this medically relevant region from the library of a hydatidiform mole. Complete hydatidiform moles arise from the fertilization of an enucleated egg from a single sperm and therefore carry a haploid complement of the human genome, eliminating allelic variation that may confound mapping and assembly.

The PacBio® *RS* system enables single molecule real time sequencing, featuring long reads and fast turnaround times. With deep sequencing, PacBio reads were able to generate a very uniform sequencing coverage with close to 100% coverage of most of the target interval regions covered. Due to long read lengths, the PacBio *RS* data could be accurately assembled.

Methods

SMRTbell™ template preparation and sequencing using PacBio® RS:

- Genomic DNA for each of the eight BACs (~200 kb) was prepared using Roche® high pure plasmid isolation kit.
- DNA was fragmented using HydroShear™ Large Shearing Assembly to prepare 7 to 8 kb SMRTbell™ libraries[1].
- Libraries were sequenced using the C1 chemistry and 45 minute movies.
- Larger insert size was used to generate long subread lengths (1.1 kb), with 95th percentile length approaching 6-8 kb.
- Library preparation and sequencing was performed as recommended in the PacBio Template Preparation and Sequencing guide.

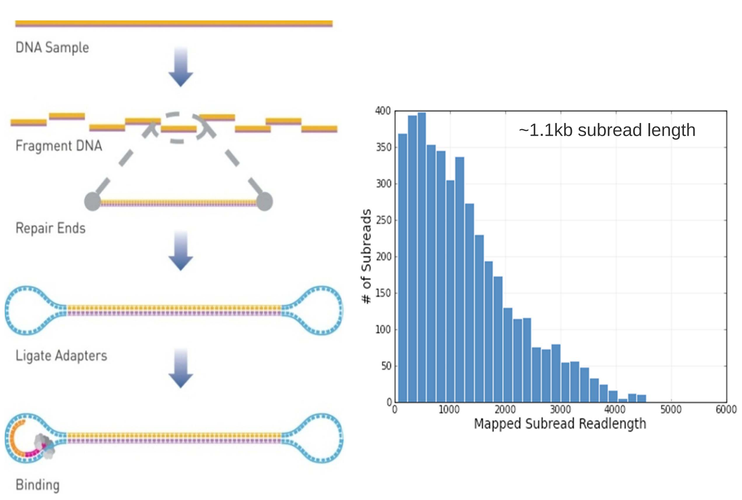

Figure 1. SMRTbell™ template preparation and sequencing using C1 chemistry per PacBio recommended protocol for large-insert libraries.

Analysis:

- PacBio® data was *de novo* assembled using ALLORA.
- MaxIterations were set to 15.
- A post processing Phrap alignment was added to the ALLORA contigs to further merge and scaffold the data.
- Per Base accuracy of *de novo* assembled ALLORA contigs was improved using Illumina® reads
- Efforts are ongoing to use Celera® assembler for an error correction based hybrid approach to combine Illumina® and PacBio® data.

Sequence Coverage

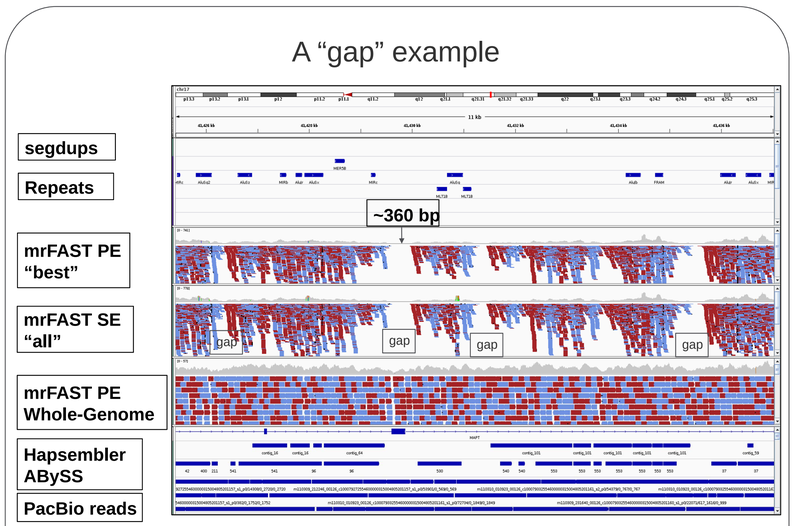

Figure 2. Comparison of sequencing coverage for CH17-170H8 using PacBio long reads. High-GC regions have very low (and sometimes zero) coverage in Illumina® Nextera™ libraries while PacBio and Illumina® WGS libraries seem to do fine.

De Novo Assembly

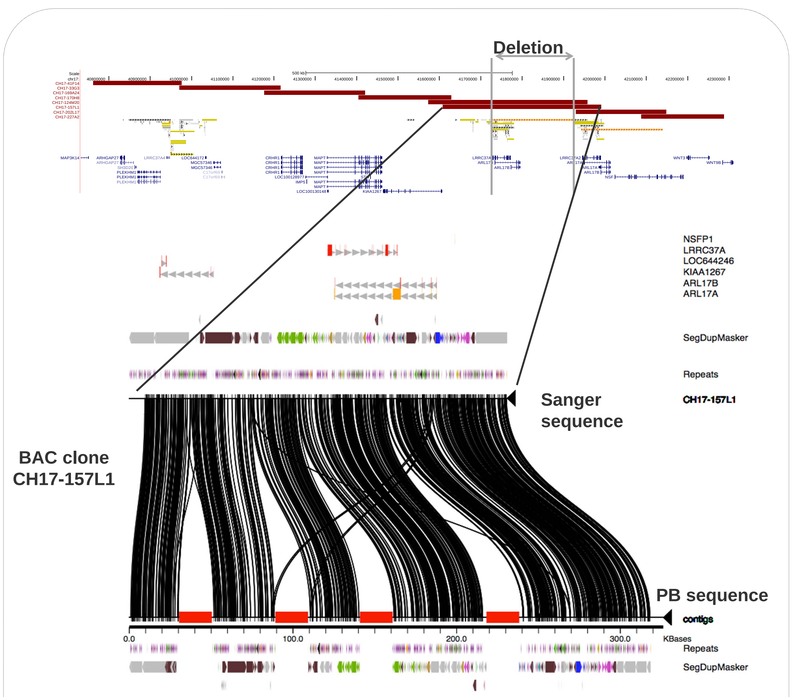

Figure 3. A tiling path of clones across the 17q21.31 region was selected based on end-sequence mapping and sequenced using SMRT technology. A 200 kb clone (CH17-157L1, mapping within segmental duplications and spanning a 210 kb deletion in the Hydatidiform mole was sequenced and de novo assembled into six contigs.

Assembly Comparison

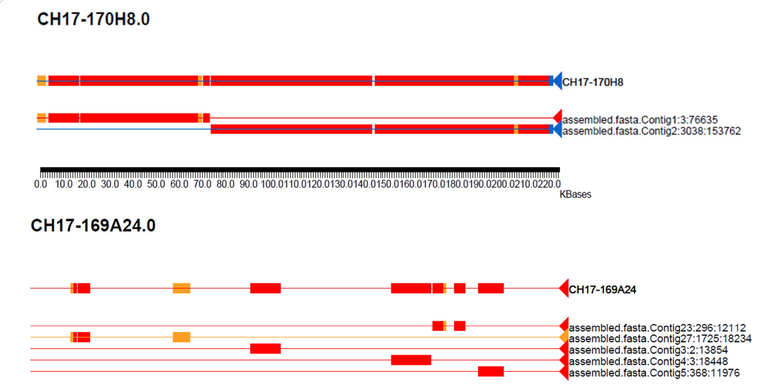

Figure 4. MVP plot of Sanger vs. PacBio assembly. The least fragmented / most fragmented clones.

Assembly Summary

Clone	Finished Length	Length (Allora + phrap)	PB only contigs	% aligned	Gaps	Gap size	Illumina only contigs	PB seq cove-rage
CH17-124M20	202,892	242,991	7	99.38%	1	1,261	110	500X
CH17-157L1	230,865	265,854	6	96.24%	1	8,673	118	650X
CH17-169A24	243,129	323,053	27	63%	23	89,936	162	610X
CH17-202L17	217,579	229,344	4	100%	0	0	51	550X
CH17-227A2	200,190	229,344	3	80.25%	3	39,544	56	250X
CH17-33G3	244,867	262,004	8	99.08%	5	2,247	111	450X
CH17-170H8	223,520	244,333	2	99.88%	1	258	120	230X

Table 1. BAC assembly using PacBio and Illumina® reads.

Conclusion

In the current work, we have sequenced and *de novo* assembled one of the most dynamic and complex regions of the human genome. Upon *de novo* assembly and mapping, most of the eight clones could be correctly assembled and aligned well, even within complex regions of the genome enriched in repeats and segmental duplications. One of the clones CH17-41F14 *de* novo assembled into 3 contigs, however, aligned poorly to the Sanger reference and is being investigated. After error correction of the PacBio reads with Illumina® reads, the overall consensus accuracy of the PacBio reads jumped to high 99% from the low 98%. This data was generated using C1 chemistry and shorter movie collection time. With the recently released C2 chemistry and V 1.3 software upgrade, longer reads and better accuracies will be able to improve BAC sequencing and assembly capability of SMRT® technology even further. Overall, unlike other short-read technologies, SMRT sequencing provides long reads that can be used to sequence and assemble alternative complex structural haplotypes not present in the currently available human reference genome.

References

1. Travers, K., et al. (2010) . Flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research*, e159.

Acknowledgements

The authors would like to thank John Huddleston who helped generate MVP plots for the Sanger vs. PacBio® data for the poster.