Highly Contiguous *de novo* Human Genome Assembly and Long-range Haplotype Phasing Using SMRT Sequencing

Meredith Ashby, Jason Chin, Lawrence Hon, John Harting, Paul Peluso, David Rank, Steve Kujawa, Luke Hickey, Jenny Ekholm, Jonas Korlach
Pacific Biosciences, 1380 Willow Road, Menlo Park, CA 94025

Introduction

The long reads, random error, and unbiased sampling of SMRT® Sequencing enables high quality, *de novo* assembly of the human genome. PacBio® long reads are capable of resolving genomic variations at all size scales, including SNPs, insertions, deletions, inversions, translocations, and repeat expansions, all of which are important in understanding the genetic basis for human disease and difficult to access via other technologies. In demonstration of this, we report a new high-quality, diploid aware *de novo* assembly of Craig Venter's well-studied genome.

Data Collection

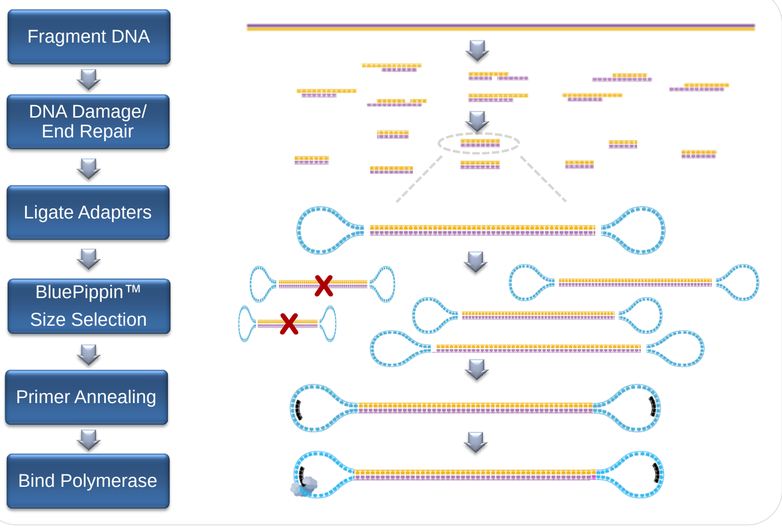


Figure 1. SMRTbell™ Library Prep Workflow. *De novo* human genome sequencing is done with one large-insert, size-selected library. BluePippin size selection was used to select inserts > 20 kb.

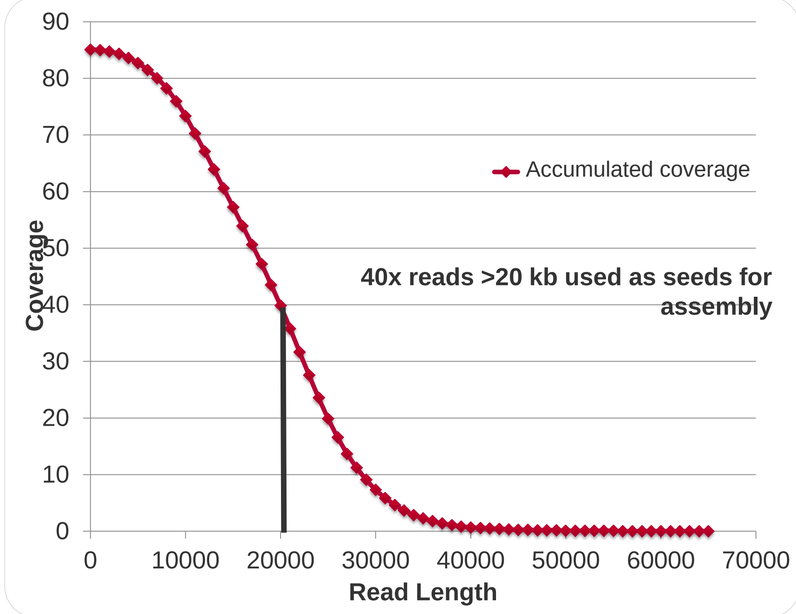
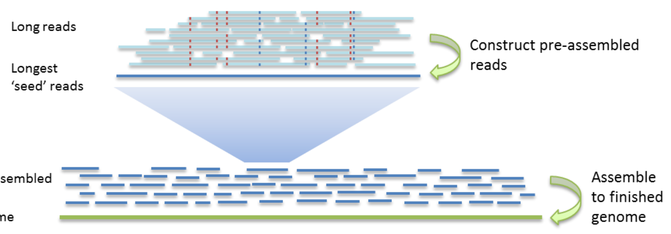


Figure 2. SMRT Sequencing Data. A total of 85x sequencing coverage was collected, with half of the data contained in reads >20 kb. These longest reads were used as alignment seeds for the multi-molecule consensus error correction step of the hierarchical genome assembly process (HGAP).

De Novo Assembly

Figure 3.The HGAP method. By first aligning long reads against the very long seed reads, randomly distributed errors are easily removed to generate highly accurate pre-assembled consensus reads, which are fed into an overlap layout consensus assembler.



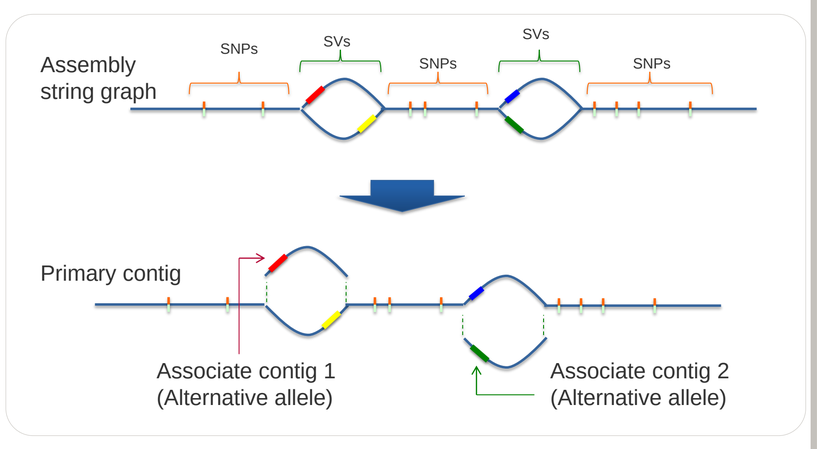


Figure 4. Diploid Aware Contig Representation With Falcon. During the assembly process, structural variants between haplotypes appear as 'bubbles' in the string graph. Falcon retains the long range information, while maintaining the relationship between the alternative alleles in the form of associate contigs.

HuRef[a]	PacBio	Sanger
#Seqs	3,004	71,343
Total	2.89 Gb	2.84 Gb
Max	34.6 Mb	-
N50	10.4 Mb	0.11 Mb

Table 1. Comparison of PacBio vs. Sanger HuRef Assemblies. The long read lengths, high consensus accuracy, and bias-free nature of PacBio data allows for the rapid assembly of gold standard genomes at reasonable cost. [a]http://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0050254

Phasing the MHC Region During Assembly

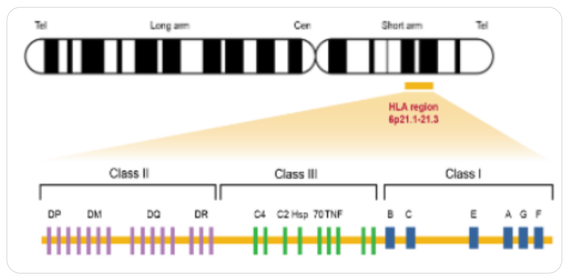


Figure 5. The HLA genes within the MHC Region of chromosome 6 are critical for distinguishing self from non-self antigens.

Of interest in any assembly is the degree of success in resolving and phasing repetitive and highly polymorphic genes implicated in human disease. One such region of interest is the 4 Mb MHC region on chromosome 6. Human MHC class I genes HLA A, -B, -C, and class II genes HLA -DR, -DP and -DQ, play a critical role in the immune system, with some alleles linked to increased risk for autoimmune disease.

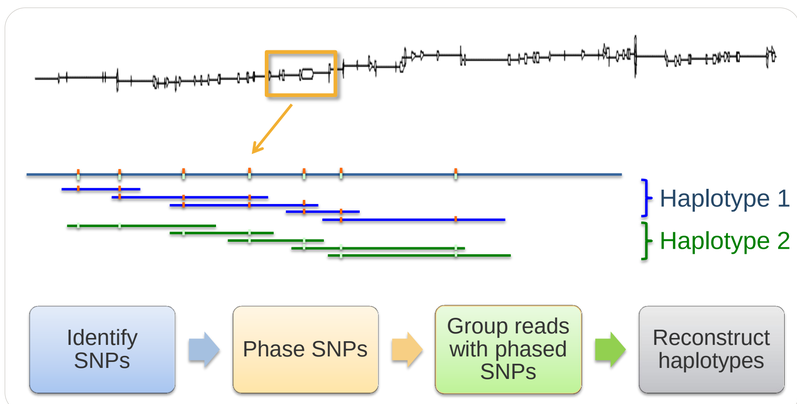
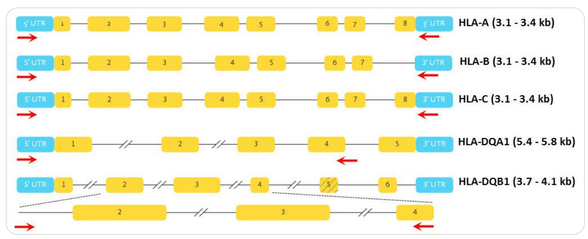


Figure 6. Structural Variation in the MHC Region. The assembly string graph contains many 'bubbles' in the 9 Mb contig encompassing the MHC region. Falcon identified 20 significant structural variations between the homologous chromosomes during *de novo* assembly.

MHC Phasing by Full-Length PCR

Figure 7. In an orthogonal approach, GenDx NGS-go® primers were used to amplify full-length HLA genes, which were converted into SMRTbell libraries. Long amplicon analysis in SMRT Analysis was used to cluster and generate consensus sequences for all alleles.



MHC Phasing By Targeted Capture

In yet another approach, the DQA1 variant calls were validated using capture technology, a cost-effective alternative to whole-genome human sequencing. Combining Roche NimbleGen's SeqCap EZ enrichment technology with Pacific Biosciences' SMRT Sequencing provides a more comprehensive view of variants and haplotype information over multi-kilobase regions. While the SeqCap EZ technology is typically used to capture 200 bp fragments, we demonstrate that capture of 6 kb fragments, when combined with the long reads of SMRT Sequencing, allows phasing of multi-kilobase regions of the human genome including exons, introns and intergenic regions.

Average Fragment Size	# SMRT Cells	Mapped Reads of Insert Mean	% Reads on Target	Enrichment Factor
6,000 bp	4	4,788 bp	48.4	600x

Table 2. Results of SeqCap EZ Enrichment of 6 kb fragments

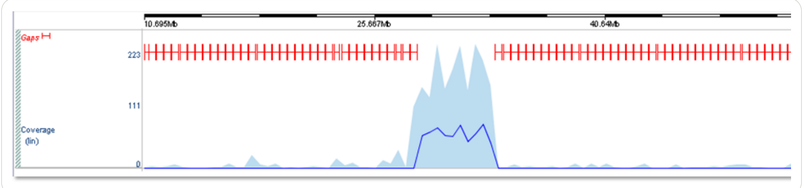


Figure 8. PacBio Sequencing Coverage of the MHC Region. The plot shows solid coverage of the MHC region targeted by the SeqCap EZ Human MHC design, with minimal off-target coverage in the flanking regions.

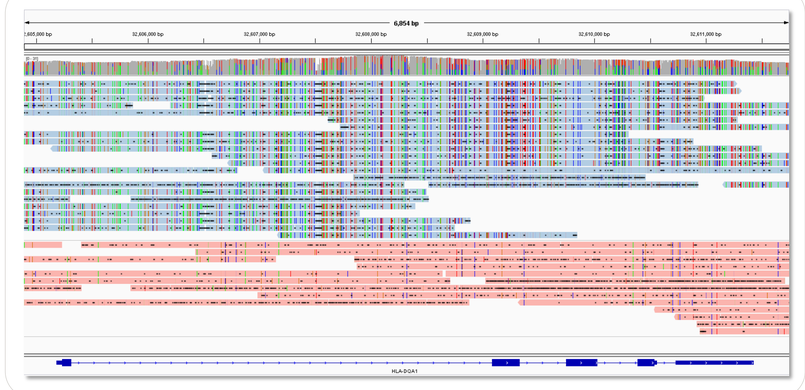


Figure 9. PacBio Phasing of the HLA-DQA1 Gene. Phased PacBio sequencing reads aligned to hg19. These reads were phased using SAMtools and then separated and grouped by haplotype (blue for one haplotype, pink for the other). For clearer visualization of the variants, reads of insert with a predicted accuracy of >97% were used. Quiver was then used to generate a consensus sequence for each haplotype. The haplotyped sequences from this gene differed from the haplotyped sequences from the *de novo* assembly by only one base pair.

Method	DQA1	DQB1	HLA-A	HLA-B	HLA-C
PacBio *de novo* Assembly	02:01 01:05	02:02:01 02:02:01	02:01:01:01 01:01:01:01[a]	13:02:01 37:01:01	06:02:01:01 06:02:01:01
NimbleGen + PacBio	02:01 01:05	Not analyzed	Not analyzed	Not analyzed	Not analyzed
GenDx NGS-go + PacBio	02:01 01:05	02:02:01 02:02:01	02:01:01:01 01:01:01:01	13:02:01 37:01:01	06:02:01:01 06:02:01:01

Table 3. Summary of All Typing Results. Both the *de novo* assembly and SeqCap EZ enrichment methods produced identical MHC Class I gene typing results. [a]One 7C>6C discordance located in an intron.

PacBio *de novo* Assembly Information Available Here:
http://www.pacb.com/applications/denovo/

PacBio Targeted Sequencing Information Available Here:
http://www.pacb.com/applications/target/

NimblGen Capture Data & Analysis Available Here:
https://github.com/lhon/targeted-phasing-consensus