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 both 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).

**Figure 3.** The HGAP method. By first aligning long reads against the very long used 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 variations 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 a separate contig.

### De Novo Assembly

### 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.

<table>
<thead>
<tr>
<th>Method</th>
<th>DQA1</th>
<th>DOB1</th>
<th>HLA-A</th>
<th>HLA-B</th>
<th>HLA-C</th>
</tr>
</thead>
<tbody>
<tr>
<td>PacBio de novo Assembly</td>
<td>Not analyzed</td>
<td>Not analyzed</td>
<td>Not analyzed</td>
<td>Not analyzed</td>
<td>Not analyzed</td>
</tr>
<tr>
<td>NimbleGen + PacBio</td>
<td>02:01</td>
<td>01:01</td>
<td>02:01</td>
<td>01:01</td>
<td>02:01</td>
</tr>
<tr>
<td>GenDx NGS-go + PacBio</td>
<td>02:01</td>
<td>01:01</td>
<td>02:01</td>
<td>01:01</td>
<td>02:01</td>
</tr>
</tbody>
</table>

| Table 3. Summary of All Typing Results. Both the de novo assembly and SeqCap EZ enrichment methods produce identical MHC class I gene typing results. *One 7C6C discordance located in an intron.*

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

**Figure 6.** Structural Variation in the MHC Region. The *assembly string graph contains many ‘bubbles’ in the 5 Mb contig encompassing the MHC region. Falcon identified 20 significant structural variations between the homologous chromosomes during de novo assembly.

**Figure 7.** In an orthogonal approach, GenDx NGS-go primers were used to amplify simplified HLA genotyping library. Long amplicon analysis in GenDx Analysis was used to cluster and generate consensus sequences for all alleles.

**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 variants were phased using SMRTtools 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 haplotype sequences from this genome are different from the haplotyped sequences from the de novo assembly by only one base pair.

---

**Table 1.** Comparison of PacBio vs. Sanger HuRef Assemblies. The long reads, high consensus accuracy, and comprehensive coverage of PacBio data allows for the rapid assembly of gold standard genomes at reasonable cost.

![Table 1](http://example.com/table1.png)

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

<table>
<thead>
<tr>
<th>Fragment Size</th>
<th>% Reads on Target</th>
<th>Enrichment Factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>6,000 bp</td>
<td>4</td>
<td>48.4</td>
</tr>
</tbody>
</table>