Development of SureSelect Target Capture Methods for Sequencing on the PacBio RS

Scott Happe1, Moraima Guadalupe1, Swati Ranade2, Lawrence Lee1, Julia Barbosa1, Angelica Giuffre1, Barbara Novak2, Carlos Pabon-Peña2, and Emily Leproust2

Agilent Technologies ¦ Cedar Creek, TX 78612, USA and ¦ Santa Clara, CA 95051, USA; ¦ Pacific Biosciences, Menlo Park, CA 94025, USA

Abstract

PacBio’s PacBio RS is a powerful new technology for single molecule, real-time sequencing, featuring long reads, fast turnaround times, and capabilities for both de novo sequencing and resequencing applications. To determine the efficacy of the Agilent SureSelect Target Enrichment System for targeted resequencing on the PacBio RS, we developed methods to prepare generic DNA libraries and analyzed sequencing performance. First, to take advantage of the long read capability, we tested different shearing conditions to create inserts of varying sizes. Second, we modified the Agilent SureSelect® kit such that unique short adapter sequences were ligated onto sheared DNA, thereby allowing manipulation of the library throughout the target enrichment process. Following hybrid capture with SureSelect® index-labeled RNA libraries and appropriate blocking reagents, PacBio SMRT® templates were constructed using the enriched DNA and subsequently sequenced on the PacBio RS. Sequencing results revealed high specificity for capturing targets of interest, excellent uniformity, and deep coverage across the targeted regions. High performance was achieved across different sample types captured with SureSelect™ libraries of various sizes, complexity, and content. Data obtained was used to efficiently identify SNPs and indels, and showed high correlation with previously-determined genotypes. Long insert libraries retained the advantages of long reads while focusing data on regions of interest. These findings illustrate the utility of the SureSelect method for target enrichment on the PacBio RS and provide a path forward for targeted variation discovery and profiling.

Materials & Methods

Figure 1. SureSelect® workflow for PacBio sequencing. A) Agilent 2100 Bioanalyzer with either the DNA1000 or High-Sensitivity DNA assay. B) Agilent 2420Bioanalyzer with the DNA-1000 chip. Library size determination and quantification.

Materials & Methods

Figure 2. A) Capture efficiency of SureSelect libraries as tested on the 2 Kb and 250 bp inserts. B) Agilent SureSelect Target Enrichment System for targeted resequencing on the PacBio RS.

Figure 3. A) Average sequencing read length for the true library insert sizes. B) Uniformity (% of reads uniquely mapped).

Figure 4. Improved coverage with long reads. The design of effective capture probes for targeted resequencing using short insert technologies is limited to regions that are highly repetitive, prime to secondary structure, or high in GC content. Coverage within such regions can be recovered by capturing long gDNA fragments and reading into the difficult regions that flank the capture regions. Shown is a region of chromosome 10 targeted by the 0.5 Mb SureSelect library. Capture probes are shown in red. Buffers genes to black, and PacBio RS read depth is indicated in gray. Coverage across non-probed regions increase dramatically when using 2 Kb inserts. Coverage across SNP sites (inserted here) is also dramatically increased, thereby potentially improving the quality of SNP calls.

Figure 5. Improved coverage across contiguous chromosomal regions using SureSelect coupled with PacBio RS long reads. The design of effective capture probes for targeted resequencing using short insert technologies is limited to regions that are highly repetitive, prime to secondary structure, or high in GC content. Coverage within such regions can be recovered by capturing long gDNA fragments and reading into the difficult regions that flank the capture regions. Shown is a region of chromosome 10 targeted by the 0.5 Mb SureSelect library. Capture probes are shown in red. Buffers genes to black, and PacBio RS read depth is indicated in gray. Coverage across non-probed regions increase dramatically when using 2 Kb inserts. Coverage across SNP sites (inserted here) is also dramatically increased, thereby potentially improving the quality of SNP calls.

Materials & Methods

Figure 6. Improved SNP detection within intronic regions. Similar to Figure 5 shown, a region of the X-chromosome targeted by the 0.5 Mb design is shown. A SNP within a paired-end read (blue) is effectively identified with both long and short-insert libraries, but a SNP within a non-probed intron (red) is only detected in the long-insert data set. Therefore, intronic regions can be effectively targeted using SureSelect non-designs coupled with PacBio RS long reads.

Materials & Methods

Table 1. SNP detection.

<table>
<thead>
<tr>
<th>Sample ID</th>
<th>gDNA Sample</th>
<th>Insert Size</th>
<th>Capture Library</th>
<th>% of reads on-target</th>
<th>% of reads uniquely mapped</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample1, 2Kb</td>
<td>165</td>
<td>156</td>
<td>40</td>
<td>156</td>
<td>100</td>
</tr>
<tr>
<td>sample2, 2Kb</td>
<td>196</td>
<td>156</td>
<td>40</td>
<td>156</td>
<td>100</td>
</tr>
<tr>
<td>sample3, 2Kb</td>
<td>196</td>
<td>156</td>
<td>40</td>
<td>156</td>
<td>100</td>
</tr>
<tr>
<td>sample4, 2Kb</td>
<td>196</td>
<td>156</td>
<td>40</td>
<td>156</td>
<td>100</td>
</tr>
<tr>
<td>sample5, 2Kb</td>
<td>196</td>
<td>156</td>
<td>40</td>
<td>156</td>
<td>100</td>
</tr>
<tr>
<td>sample6, 2Kb</td>
<td>196</td>
<td>156</td>
<td>40</td>
<td>156</td>
<td>100</td>
</tr>
</tbody>
</table>

Materials & Methods

Table 2. SNP detection. Similar to Figure 5 above, a region of the X-chromosome targeted by the 0.5 Mb SureSelect library. Capture probes are shown in red. Buffers genes to black, and PacBio RS read depth is indicated in gray. Coverage across non-probed regions increase dramatically when using 2 Kb inserts. Coverage across SNP sites (inserted here) is also dramatically increased, thereby potentially improving the quality of SNP calls.

Materials & Methods

Table 3. SNP detection. Similar to Figure 5 above, a region of the X-chromosome targeted by the 0.5 Mb SureSelect library. Capture probes are shown in red. Buffers genes to black, and PacBio RS read depth is indicated in gray. Coverage across non-probed regions increase dramatically when using 2 Kb inserts. Coverage across SNP sites (inserted here) is also dramatically increased, thereby potentially improving the quality of SNP calls.

Materials & Methods

Conclusion

• SureSelect efficiently captures both short and long inserts for effective targeted resequencing on the PacBio RS system

www.agilent.com/genomics/sureselect

For research use only. Not for use with diagnostics procedures.

©Agilent Technologies, Inc., 2012

Published in USA, January 4, 2012
Publication Number 5990-9636EN