Structural variant detection with long-read sequencing reveals driver and passenger mutations in a melanoma cell line

Aaron Wenger1, Meredith Ashby1, Marcel Nelen2, Wizard Kooistra3
1. PacBio, 1305 O’Brien Drive, Menlo Park, CA 94025; 2. Radboud University Medical Center, 6500 HB Nijmegen, Netherlands; 3. University Medical Center Utrecht, 3584 CG Utrecht, Netherlands

Abstract #5515

Structural variation and long-read sequencing

<table>
<thead>
<tr>
<th>Technology</th>
<th>Precision (%)</th>
<th>Recall (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>short-read sequencing</td>
<td>5 Mb</td>
<td>83.23</td>
</tr>
<tr>
<td>PacBio SMRT sequencing</td>
<td>5 Mb</td>
<td>96.13</td>
</tr>
<tr>
<td>PacBio SMRT sequencing</td>
<td>3 MB</td>
<td>83.76</td>
</tr>
<tr>
<td>PacBio SMRT sequencing</td>
<td>10 Mb</td>
<td>81.79</td>
</tr>
</tbody>
</table>

Figure 1. Variation in a typical germline human genome1. Most of the base pairs that differ between two human genomes are indels and structural variants (SVs), differences ≥50 base pairs. Short-read sequencing has limited sensitivity for indels and SVs, while PacBio long-read sequencing comprehensively detects variant sizes of all kinds.

Long Reads average up to 30 kb
High Consensus Accuracy random errors produce Q50 consensus Uniform, Unbiased Coverage no GCs or sequence complexity bias
Epigenetic Characterization simultaneous detection of DNA methylation Single-Molecule Resolution directly measure individual DNA molecules

Figure 2. Advantages of sequencing on the PacBio Sequel Systems.

The Genome in a Bottle consortium2 has developed a benchmark set of insertion and deletion structural variants (https://tinyurl.com/GIABSV06) in a human male, GRCh37 NA24385. Comparing technologies against this benchmark, PacBio has the highest precision and recall across the structural variant size range, and particularly for insertions.

Figure 3. Variant calling performance against the GIAB HG002 v6.0 benchmark. Histograms indicate the number of variants and lines show the precision (blue) and recall (orange) at each variant size for call sets from different technologies.

Structural variant size range, and particularly for insertions.

Figure 4. SVs and SV calls from PacBio in a typical germline human genome shows that PacBio has the highest precision and recall across the structural variant size range, and particularly for insertions. For Research Use Only. Not for use in diagnostic procedures. © Copyright 2019 by Pacific Biosciences of California, Inc. All Rights Reserved. PacBio, PacBio, SMRT, SMRT sequencing, and Single-Molecule Resolution are trademarks of Pacific Biosciences, Inc. The term IGNitetm is a trademark of Sage Science, Inc. and/or its licensors. The term Genomic Link is a trademark of Sage Science, Inc. and/or its licensors. All other trademarks are the property of their respective owners.

References

To evaluate the coverage required to detect somatic variants at different tumor sample purity levels, a stringent (≥4 variant reads in tumor, 0 in normal) set of 46 somatic variants from the full-coverage, pure input sample were treated as a benchmark truth set. Tumor purity was titrated by mixing sequences from the tumor and normal samples. Coverage was titrated by subsampling (same in tumor and normal), and putative somatic variants were called as variants present in the tumor but not normal (≥2 variant reads in tumor, 0 in normal). Variant calls in the truth set were considered true positives; others were considered false positives. Precision=TP/(TP+FP), Recall=TP/(TP+FN).

Conclusions
• PacBio long-read sequencing has the highest precision and recall for structural variants of evaluated sequencing platforms.
• PacBio long-read sequencing of the cancer cell line COLO829 and its matched normal COLO829BL shows large-scale copy-number changes and tens of somatic structural variants, including a rearrangement of chr10 and chr19 and 12 kb deletions that impacts PTEN.
• Even with permissive criteria for calling somatic structural variants, precision is high across a range of tumor purity and coverage. Recall is high at 20-fold coverage in samples that are at least half tumor and nearly saturates by 30-fold coverage at all purity levels.

References

The authors acknowledge support from the GIAB, Sage Science, Inc., and Pacific Biosciences, Inc. For Research Use Only. Not for use in diagnostic procedures. © Copyright 2019 by Pacific Biosciences of California, Inc. All Rights Reserved. PacBio, PacBio, SMRT, SMRT sequencing, and Single-Molecule Resolution are trademarks of Pacific Biosciences, Inc. The term IGNitetm is a trademark of Sage Science, Inc. and/or its licensors. The term Genomic Link is a trademark of Sage Science, Inc. and/or its licensors. All other trademarks are the property of their respective owners.