An Optimal Long-Read Workflow for Human Genome Sequencing

Justine Lenhart, Sukhinder Sandhu, Laurie Kurihara, Vladimir Makarov, Timothy Harkins
Swift Biosciences, 58 Parkland Plaza, Suite 100, Ann Arbor, MI 48103

Abstract

The goal of precision medicine is to treat disease with targeted therapies practiced on highly accurate diagnostic sequencing. Short-read sequencing technology offers low error-rates, high throughput capabilities, and low sequencing costs; however, short-read sequencing has a limited ability to map reads in regions banded by low complexity (i.e. repetitive elements) or high polymorphic density while having minimal ability in identifying large insertions or deletions. An alternate long-read sequencing technology, Single Molecule Real-Time (SMRT) offered by Pacific Biosciences® (PacBio®) provides continuous, low bias reads upwards of 60 kb in length. SMRT technology is ideally suited to produce readily mapped reads in regions that would typically be undetected with short-read technologies. These capabilities ideally position the long-range sequencing PacBio platform to provide enhanced coverage and assembly metrics for whole-genome sequencing and potentially de novo assembly of human genomes, bringing the goal of precision medicine within reach. However, standard workflows are laborious and time consuming.

Here, we provide a complete workflow with the Swift Accell-NGS™ XL Library Kit for the acquisition of ultra-long (>30 kb) subreads for the PacBio sequencing platforms. The Accell-NGS XL Library Kit for PacBio combines Swift’s proprietary repair and ligation chemistry to offer a significant improvement in library conversion efficiency. Furthermore, we highlight several strategies to further optimize average human genome subread length (>10 kb), from producing a high quality initial sheared using the Diagenode Megaruptor® 2 and selection of ultra-long library molecules with the Sage BluePippin™, to improving integrity of the final library with the provided Swift terminal repair module. These long-length libraries can result in increased coverage of hard-to-sequence regions of the human genome, facilitating detection of both small and large structural variants, improved phasing of distant SNPs, along with precise elucidation of repetitive repeat number. Ultimately, by enabling an improved, complete genomic landscape for precision medicine, better understanding of variants can be realized.

DNA Shear

An underappreciated variable in maximizing data output is the shear length of the library. With the Accell-NGS XL kit, a short library molecule (~8 kb) will not produce as much unique data as a long library molecule. Conversely, longer library molecules (~60 kb) can provide more unique data; however, may be more difficult to both handle during the library prep and to load into the ZMW. Here, we present data evaluating various shearing strategies and lengths of libraries made with the Accell-NGS XL Kit and run on the PacBio RSII.

Best Practices Workflow

To produce a high performing library that maximizes unique DNA output, we recommend this workflow.

- The researcher needs to produce DNA that is both free of metabolites that can poison SMRT-sequencing and that is >100 kb in size for optimal performance.
- Shear the DNA with the Megaruptor to a mode size of 35 kb long. This size can produce long average sub-reads on the PacBio RSII, while facilitating gentle handling and efficient loading when combined with a sufficiently stringent size selection.
- After making your Accell-NGS XL libraries, we recommend using the BluePippin Size Selection (BPSS) to remove 50 kb size selection lower and upper threshold settings. Use the PacBio Cassette with the cassette definition of 0.75% DF Master 51 high-pass 15-20 kb. Recommended library inputs and expected outputs are shown in Table 1.

- This library will be effectively loaded at an on plate concentration of 150-250 pM using the Megabead loading protocol detailed by PacBio.

Table 1: Considerations for determination of library amount entering into the BluePippin

<table>
<thead>
<tr>
<th>Percentage Recovery</th>
<th>On-Plate Loading Concentration</th>
<th>Minimum Input</th>
</tr>
</thead>
<tbody>
<tr>
<td>50%</td>
<td>100 pM</td>
<td>150 pM</td>
</tr>
<tr>
<td>5%</td>
<td>150 pM</td>
<td>250 pM</td>
</tr>
</tbody>
</table>

Blue Pippin Size Selection

Size selection using the Sage BluePippin (BPSS) of Accell-NGS XL libraries impacts the acquisition of longer reads, providing more unique data per run.

Specific Shear Size (30-40kb) vs. BPSS

Evaluated whether larger Megaruptor2 shear lengths could add unique data when incorporating a stringent (20-50 kb) BPSS.

Conclusion

- Swift Biosciences’ Accell-NGS XL Library Kit for PacBio offers an advantage for whole genome assembly and application elsewhere where long reads are imperative for success.
- The Accell-NGS XL Library Kit produces complete de novo assemblies with the HapGood pipeline (data not shown, visit swiftbiosci.com).
- Our best practice workflow provides unique advantages in a shorter assay time (~5 hours) with single-tube capability and no overnight incubations.
- A better ligation efficiency and improvements in read length help the user gather higher quality data with less time and cost.
- Currently, the Accell-NGS XL Library Kit is validated for PacBio RSII.