

Microbial Assembly

Experimental Design

Introduction

Complete microbial genomes can be sequenced for comparative genomic analyses using the PacBio® RS system. Existing assemblies can be refined using Pacific Biosciences® Continuous Long Read (“CLR”) data to scaffold contigs. Importantly, microbial genomes can also be *de novo* assembled using PacBio RS sequencing alone, using 1 kb to 2 kb insert SMRTbell™ libraries and Circular Consensus Sequencing (“CCS”) data, to provide error correction for the CLRs prior to assembly.

In this document, we explain the experimental design methods and analysis protocols for sequencing microbial genomes, including recommendations for understanding the microbial genome, experimental goals, estimating the coverage needed, preparing the samples, PacBio RS sequencing, and subsequent analyses.

Finished Genomes

A comprehensive understanding of the complete microbial genome is critical to understanding its pathogenicity, infectivity, and virulence factors. A typical microbial genome can range from two to five megabases in size, and may contain numerous repetitive regions, mobile elements (e.g., plasmids, phages), and antibiotic resistance-associated genes or regions. Effective genomic assembly for variant and strain characterization requires read lengths long enough to span these repetitive regions and high uniform coverage through GC-rich regions.

The PacBio RS can successfully sequence the complete genomes, of hard-to-sequence microbial genomes, due to the long readlengths and relatively unbiased coverage generated.

With second generation, short-read sequencing technologies, a reference strain is typically needed for genome assembly, introducing a “reference bias” into the resulting assembly. In contrast, no such references are required when using PacBio data in *de novo* assembly, thereby allowing for more accurate, reference-agnostic characterization, identification, and discovery of new microbial strains.

The PacBio RS and SMRT® sequencing protocols can be used to generate both extremely long reads, or shorter (<2 kb) high accuracy reads, both of which are useful for genome assembly. Large inserts allow for CLRs that can help span gaps or larger repetitive regions. Shorter inserts yield more CCS data, which take full advantage of the circularized SMRTbell construct to sequence a single insert multiple times (to sample a given insert more than once) and result in higher intra-molecular accuracy; these CCS sequences can be useful for error correction of the longer CLRs.

Assembly Methods

Assembly methods can be generalized into three categories:

- *SMRT® de novo*: refers to the assembly of PacBio CLR reads only
- *SMRT Hybrid*: refers to the hybrid *de novo* assembly of PacBio CLR and a second high accuracy data type (either PacBio CCS reads or second generation short-read data)
- *SMRT Scaffolding*: refers to using PacBio CLR to scaffold existing contigs generated from short-read data


 Experiment Design

Isolate DNA

 Template
Preparation

Sequencing

DATA Analysis

For *de novo* assembly, CLR accuracy is equally as important as long readlengths for accurately calculating overlaps and improving the overall assembly. In hybrid assembly, highly accurate CCS reads are needed, in addition to CLR, for error correction. Note that short-read data from other platforms can also be used for this purpose; however, short-read error corrected PacBio reads will still contain sequencing biases inherent with second generation platforms and subject to the

same coverage problems. To ensure proper mapping for the error correction step, CCS data requires sufficient intra-molecular coverage for high QV circular consensus basecalls.

The following table describes the current assembly programs compatible with PacBio data and the types of assembly each can perform.

Table 1: Assembly Programs Compatible with PacBio Data and Assembly Methods

	Description	SMRT de novo	SMRT Hybrid	SMRT Scaffolding
AHA (SMRTAnalysis)	Assemble short reads into high-confidence contigs and scaffold with PacBio CLR.			✓
ALLORA (SMRTAnalysis)	Assemble PacBio CLR and short read or CCS data. The P_ErrorCorrection module has to be run manually to error correct the CLR reads prior to assembly with ALLORA.	✓	✓	
ALLPATHS-LG	Error correct and scaffold PacBio CLR in a multistage process using different types of short read data. Optimized for single node high-memory computation.		✓	
Celera Assembler	Error correct PacBio CLR with accurate short reads and assemble. Optimized for cluster computation.		✓	
MIRA	Assemble error corrected PacBio CLR generated by another error correction pipeline, e.g. Celera Assembler.		✓	

Estimating the Coverage Needed to “Finish” a Genome

The amount of coverage required to “finish” a genome depends on the study objective, complexity of the genome, and choice of assembly software. For example, resolving a genome with long repetitive elements may require more coverage with CLR using libraries whose insert sizes span the gaps.

Table 2 provides some general recommendations for coverage estimates based on choice of assembly software.

Table 2: Coverage Estimates

	SMRT <i>de novo</i>	SMRT Hybrid	SMRT Scaffolding
AHA (SMRTAnalysis)			10X PacBio CLR (in addition to high confidence contigs)
ALLORA (SMRTAnalysis 1.3)	50X PacBio CLR	50X PacBio CCS or short reads 20X PacBio CLR	
ALLPATHS-LG		50X PacBio 3kb CLR 100X Illumina PE 100X Illumina Jumping Libraries	
Celera Assembler		50X PacBio CCS or short reads 20X PacBio CLR	
MIRA		15-20X PacBio Error Corrected CLR	

Additional Factors to Consider when Designing an Assembly Experiment

It should be noted that longer reads increase the probability of bridging “gaps” between contigs in a given genome assembly. Most of these gaps come from repetitive elements in the genome (tandem repeats, etc.) that cannot be resolved with reads shorter than the repeat length. A higher depth of coverage maximizes the probability of spanning various gaps in the genome. Larger insert sizes (>5 Kb) and higher PacBio long read coverage (30-50X) will generally result in more complete and accurate assemblies.

Also, error correction of PacBio CLR, with high accuracy short reads, are affected by the sequencing characteristics of the short reads. For example, the Celera Assembler error correction pipeline currently splits PacBio CLR at regions of low short-read coverage, which limits the potential of PacBio CLR to bridge assembly gaps. Lack of short-read coverage for a particular region can be explained by:

- An inability to align the short-read to the long-read due to accuracy differences in the long read
- Ambiguity in aligning short-reads from repetitive regions to the true long-read target
- Lack of short-read data due to platform sequencing biases such as uneven or lack of coverage through AT- or GC-rich sequences

A remedy for this is to include PacBio CCS reads with up to 2 kb inserts as they will have more uniform coverage across all genome contexts. If CCS data is available, their inclusion in error correction assembly pipelines has great value in overcoming the short-read coverage deficiencies.

Preparing the Samples

SMRT sequencing is compatible with a variety of genomic DNA isolation methods, several of which are outlined in Pacific Biosciences *Technical Note, Microbial Analysis, E.coli Genome Assembly*³.

DNA quality is extremely important for successful SMRT sequencing, particularly for the sequencing of large-insert (e.g., 10 kb) libraries. With any DNA isolation procedure, it is critical to limit damage to the gDNA by keeping all incubation steps at low temperatures (<60°C), buffering genomic DNA (“gDNA”) from the effects of nucleases, and minimizing freeze-thaws of extracts. SMRT sequencing does not require PCR or cloning during any step of the library preparation process: the input DNA is sequenced directly. For large-insert libraries, we recommend using either the Repair Mix provided in the Pacific Biosciences DNA Template Prep Kit 2.0 or the New England BioLabs® Inc. PreCR® Repair Mix (see www.neb.com).

For more information on template preparation methods, see the recommendations provided in the *Pacific Biosciences Template Preparation and Sequencing Guide*.

Target Insert Size

Fragments for Large Insert Libraries (5-10 kb)

DNA should be sheared to a target size of 5 kb –10 kb to maximize sub-read (the long contiguous regions of a sequence) length. However, the desired insert length can be limited by the amount of available sample. Low amounts of sample DNA may necessitate the construction of smaller libraries to achieve desired coverage targets. The current protocols recommend 2 µg input for 5 kb – 6 kb libraries and 5 µg input for 10 kb libraries sheared DNA into the End Repair reaction. With high-quality starting material and minimal up-front losses during the DNA fragmentation and cleanup steps, we typically recommend 3.5 µg and 7.5 µg of DNA for a 5 kb and 10 kb library, respectively.

It should be noted that shearing methods can also be sources of high variability in yields. Currently, we recommend using a Hydroshear® Plus Shearing system (from Digilab®) for large assemblies. Various Hydroshear speed code settings (near the manufacturer's recommended parameters) should be tested to find the optimal setting for a given gDNA. The Covaris® g-Tube, a centrifugation based DNA fragmentation device, is another viable shearing option. DNA shearing protocols for three different types of centrifuges have been developed for 6 kb and 10 kb target shears. Analytical gels should be run to examine the sizes and characteristics of test shears before deciding on final conditions.

SMRTbell libraries should be constructed using the Pacific Biosciences Template Prep Kit and accompanying procedures. Note that the reaction volumes should be scaled appropriately with increasing DNA input amounts.

Refer to the Pacific Biosciences *Technical Note, Microbial Analysis, E.coli Genome Assembly* for more information on shearing recommendations and SMRTbell™ library preparation methods used in that study.³

Fragments for Short Insert Libraries for CCS (500-2 kb)

If CCS data is going to be used in the assembly, a second library should be created using the *Pacific Biosciences Template Preparation and Sequencing Guide* and accompanying procedures. DNA should be

sheared to 500 bp to 2 kb, using the appropriate shearing devices from Covaris and their recommended protocols for fragmentation.

Similar to the large insert libraries, SMRTbell libraries should be constructed using the Pacific Biosciences Template Prep Kit and accompanying procedures. Note that the reaction volumes should be scaled with increasing DNA input amounts.

Refer to the Pacific Biosciences *Technical Note, Microbial Analysis, E.coli Genome Assembly* for more information on shearing recommendations and SMRTbell™ library preparation methods used in that study.³

Sequencing

Sequencing Recommendations

To optimize loading concentrations and maximize throughput without compromising accuracy, conduct a loading titration. For the large insert libraries, 1X90 movies should be collected to maximize read lengths. For medium insert libraries, 1X75 movies can be used to increase the number of SMRT Cells that can be run per day. For insert libraries less than 2 kb, 2X45 movies can be run to maximize the number of CCS reads collected.

Conclusion

Improving and finishing microbial genomes has previously been a costly venture. PacBio CLR and CCS data now make it possible to sequence microbial genomes with read lengths capable of spanning large repetitive regions and assembly gaps - often in excess of 7 kb in length - without complicated primer design or Sanger sequencing. PacBio reads also have even coverage across all GC contexts, even in extreme GC-rich and poor genomes where second generation technologies typically struggle. The combination of PacBio CLR and CCS data, or second generation short reads, can generate extremely accurate assemblies in short periods of time. In addition, open-source assembly programs are now able to integrate PacBio data and second generation data with great success.

References

1. Gnirke, A., et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnology* Feb 27 (2009):182-9.
2. Travers, K., et al. Flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Research* (2010), e159.
3. Technical Note, Pacific Biosciences, Microbial Analysis, E.coli Genome Assembly (2011).

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2012, Pacific Biosciences of California, Inc. All rights reserved. Information in this document is subject to change without notice. Pacific Biosciences assumes no responsibility for any errors or omissions in this document. Certain notices, terms, conditions and/or use restrictions may pertain to your use of Pacific Biosciences products and/or third party products. Please refer to the applicable Pacific Biosciences Terms and Conditions of Sale and to the applicable license terms at <http://www.pacificbiosciences.com/licenses.html>.

Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT and SMRTbell are trademarks of Pacific Biosciences in the United States and/or certain other countries. All other trademarks are the sole property of their respective owners.

PN 100-092-300-01