Pipelines for Evaluating Prokaryotic References
Supplemental Material
Candidate Reference Material Sequencing and Genome Assembly

September 23, 2015

1   Material Development

Sequencing data for the NIST candidate reference material RM8376, genomic DNA from *Staphylococcus aureus* strain NRS100 isolate COL (Biosample SAMN02854573) was selected as a model system to demonstrate how PEPR is used to characterize a microbial genomic material. The strain was isolated from a clinical sample by Children's National Hospital (Biosample SAMN02700075). The genomic material represents a large homogeneous batch of extracted DNA, with 1500 vials with $3\mu g$ DNA in each. The extracted DNA was prepared from single colony of the initial culture stab was incubated overnight at 37 °C on an agar plate and a single colony was used to inoculate a new plate. One colony from the new plate was grown in 20 mL Luria-Bertani (LB) broth at 37 °C. The culture was used to inoculate 15 X 150 mm plates which were incubated at 37 °C for 16 hours. DNA was isolated by lysing the bacteria in a solution containing NaCl, Tris, EDTA and lysostaphin ($25\mu g/ml$) and SDS. Proteinase K and RNase A were used to treat protein and RNA. Ammonium acetate was used to remove protein. DNA was recovered by isopropanol precipitation as follows, the DNA was washed with 70% alcohol, drained, and dissolved in TE buffer (Tris 10 mM, EDTA 0.1 mM, pH8.0).

2   Sequencing Experimental Design

The *S. aureus* candidate reference materials were sequenced using three orthogonal sequencing platforms: Pacific Biosciences RSII (PacBio)[1], Ion Torrent PGM [2], and Illumina MiSeq [3]. For PacBio sequencing, the sequencing library was prepared using DNA Template Prep Kit 3.0 with pooled DNA from three randomly sampled vials of the candidate reference material RM8376. The resulting library was sequenced with the P6-C4 chemistry. For the Ion Torrent PGM [2] and Illumina MiSeq [3] sequencing, eight

---

[1]Pacific Biosciences of California Inc.http://www.pacificbiosciences.com/ 1380 Willow Rd. Menlo Park, CA 94025 USA

[2]Life Technologies Corp., http://www.iontorrent.com/ 7000 Shoreline Court # 201, South San Francisco, CA 94080 USA

[3]Illumina Inc., http://www.illumina.com/ 5200 Illumina Way San Diego, CA 92122 USA
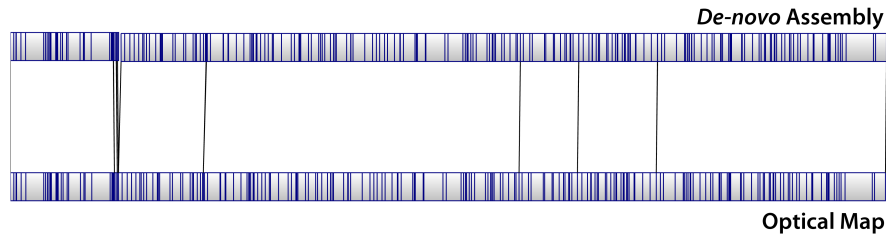
Figure 1: Comparison of optical map data to genome assembly. Alignment of *in-silico* genome map generated from the PacBio HGAP assembly to the OpGen optical map. Blue bars in map represent NocI restriction sites, black lines indicate co-linear regions.

vials were randomly sampled from the lot of 1500 vials. For MiSeq, two technical replicate libraries were prepared for each of the eight vials using the Nextera DNA Sample Prep Kit[3]; samples were barcoded using the Nextera Index Kit and sequenced using the MiSeq 600 cycle Reagent kit v3 for 2 X 300 bp reads. The 16 libraries were pooled and sequenced in a single run. Single barcoded 400 bp Ion Torrent PGM libraries were prepared for each of the eight vials using the Ion Xpress Plus kit [2]. The vials were barcoded using the IonXpress Kit[3], sequencing template was prepared using the Ion PGM Template OT2 400 kit and the IonPGM400 kit was used for sequencing on a 318C chip. The raw sequence data is archived in the Genbank Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra), see Table 1 of main paper for accession numbers. A genome map was obtained for the candidate reference material from optical mapping data generated using the Argus Optical Mapping System[4]. Agarose plugs generated from the same culture stock as that used to generate the DNA reference material were used for optical mapping measurement. Restriction enzyme NcoI was used for the restriction digest. These steps occured prior to our PEPR pipeline.

3 *S. aureus* Reference Assembly

PacBio long read data were used to generate a *de novo* genome assembly for use in the Genome Evaluation Pipeline. PacBio Reads were assembled using SMRTAnalysis software version 2.3 [1] to apply the HGAP assembly algorithm [1]. To identify potential errors in the assembly, an *in silico* digest of assembly was compared the genome map obtained from the optical mapping data using OpGen MapSolver software[4]. The *de novo* assembly was supported by Optical mapping data (Fig. 1). The plasmid is too small for OpGen optical mapping technology, and therefore the plasmid assembly was not validated with the optical mapping data.

---

[4]OpGen Inc. http://opgen.com 708 Quince Orchard Road Gaithersburg, MD 20878 USA

References

[1] Sergey Koren et al. "Reducing assembly complexity of microbial genomes with single-molecule sequencing." In: *Genome biology* 14.9 (2013), R101. ISSN: 1465-6914. DOI: 10.1186/gb-2013-14-9-r101. arXiv: 1304.3752. URL: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4053942\&tool=pmcentrez\&rendertype=abstract.