Closing and Finishing Microbial Genomes with Short Read Sequencers

David Mead, Svetlana Jasinovica, Erin Ferguson, Richard Davis, Peter Panizzi,* Mark Liles,* Michael Lodes, Scott Monsma*

The vast majority of sequenced microbial genomes are incomplete, resulting in hundreds of DNA fragments, misassembled and missing genes, and unordered pathways. Multiple repetitive DNA elements abundant in most species are too long for short read sequencers to span, which prevents bioinformatics software from being able to accurately assemble these multiple identical choices. Lucigen has developed an NGS library technology that generates 20 kb mate pair reads that jump across repetitive elements found in microbial genomes (www.lucigen.com/matepair), which dramatically improves the success of assembly *de novo* by resolving repeat elements and ordering contigs. Six microbial genomes have been assembled *de novo* and completely closed and finished using Lucigen's NxSeq® 20 kb mate pair library kit, a companion fragment library, and SPAdes assembly software.

20 kb Mate Pair Libraries Enable De Novo Genome Assembly and Closing

Determining the genetic basis of microbial genomes is relatively easy and economical since the advent of next generation sequencing (NGS) technologies. Nevertheless, more than 87% of the microbial genomes in GenBank are incomplete, or draft quality (29,617 draft vs. 3,886 finished, as of April 2015), and have an average of 190 contigs compared to approximately 5 contigs for genomes defined as finished [1]. The accuracy of these draft genomes is in question, due to their incomplete genome size, the number of predicted genes, number of repeats, GC content, status of plasmids, and missing genes. The effort and cost to close and finish microbial genomes using short read sequencers is not practical using conventional library construction technologies. Alternative methods such as PCR amplification across gaps with subsequent Sanger sequencing or bridging with PacBio long reads have been described. A simple method for generating long range genomic information from short read sequencers could revolutionize the ability to complete genomes using the most readily available Illumina or Ion Torrent platforms. This application note describes how long span mate pairs can be used to close and finish microbial genomes with minimal manual effort using short read sequencing instruments.

The first step in sequencing a microbial genome consists of shearing the DNA to small fragments ~300-600 bp long, attaching platform-specific adapters, and generating millions of reads using Illumina or Ion Torrent instruments. These short reads are then assembled into longer contiguous stretches using software that aligns the sequences, typically generating ~100 "contigs" from the original full length genome. To link these contigs in the correct order and orientation, long range sequence information, such as that obtained from mate pair (MP) library data is required. MP libraries are created by capturing the extreme ends of DNA fragments separated by a span of multiple kb, and manipulating them into close proximity for short read sequencers. When the sequences of the two mate pairs are aligned to a conventional fragment library data set, their orientation and span distance can be used to assemble the numerous contigs into a small number of scaffolds spanning over repeats. Contig building and scaffolding can be accomplished by use of the SPAdes genome assembly software [2], which has been updated to version 3.5.0 (http://bioinf.spbau.ru/spades). This latest version includes error correction for Ion Torrent and Illumina reads, as well as automatic handling of Lucigen's Chimera Code™ and Junction Code™ sequences, which have been built into the new mate pair library kit to facilitate accurate genome closing. Manual inspection of mate pair linkages, gap filling, and curation of the scaffolds generated by SPAdes can often result in a single contig, as shown for the seven genomes described here.

The NxSeq Long Mate Pair Kit was optimized using *Escherichia coli* K12 (*E. coli* K12) [3] genomic DNA as a high quality reference standard. Genomic DNA was sheared to 20 kb using a g–TUBE™ (Covaris) centrifugal fragmentation device or Megaruptor™ shearing apparatus (Diagenode). The DNA was end repaired, A-tailed, ligated to adaptors, proteinase K-treated and precipitated prior to size selection using an Elutrap™ (Whatman) or BluePippin™ (Sage Science) electroelution system. Eluted DNA (20 kb) was processed with Lucigen's NxSeq Kit (http://lucigen.com/NxSeq-Long-Mate-Pair-Library-Kit/) by ligation to the coupler, and exonuclease treatment was followed by restriction

* Auburn University, Auburn, AL

endonuclease digestion, circularization with the junction adaptor and PCR amplification. The MP library was sequenced using 2 × 250 paired end reads on an Illumina MiSeq platform and raw data was processed automatically by the SPAdes software. True mate pair reads (detected by Chimera Code™ sequences and split by detection of Junction Code™ sequences) were used in conjunction with a fragment library to assemble and close the *E. coli* K12 genome. Mapping the 20 kb MP data onto the *E. coli* reference genome reveals a nearly symmetric distribution of mate pair distances (Figure 1) with the peak at ~20 kb with some mate pairs approaching 32 kb. This genome can be readily closed (Table 1) and the *de novo* assembly is perfect as judged by comparing a Quast analysis [4] of the data generated here against the high quality reference genome [3] (data not shown).

The Joint Genome Institute was unable to close a *Thermus aquaticus* genome using existing technology, leaving it in permanent draft status in 2008 (http://www.ncbi.nlm.nih.gov/nuccore/218297404). This genome is challenging due to its high GC content and moderate level of repetitive elements. Nevertheless, a 20 kb MP library, in conjunction with a conventional fragment library enabled the first ever closed genome for this organism, as well as the delineation of four separate plasmids contained within the bacterium (Table 1). In order to confirm the quality of the newly closed genome and compare it with the 22 contig draft submission, 18 sets of primers were designed to amplify across and sequence the gaps using conventional PCR and Sanger sequencing. Although the original draft sequence was assembled correctly, the number of misassembled and split genes was high at the termini of the contig junctions (data not shown). A complete and finished *T. aquaticus* genome has been submitted to NCBI/GenBank and for peer-reviewed publication.

Based on these results we attempted to close and finish five additional microbial genomes (Table 1). A variety of microbes were chosen to span a broad range of GC contents, genome sizes, and Gram classifications. In all cases the genomes had been sequenced to adequate coverage (30-100X) using paired end reads from conventional fragment libraries. However, none of them could be closed, even using early-generation PacBio reads for one of them (data not shown). As described previously, 20 kb mate pair libraries were constructed from the *Staphylococcus, Streptomyces, Nonomuraea, Bacillus*, and *Aeromonas* genomes show in Table 1. In conjunction with a conventional fragment libray and the SPAdes genome assembly software, all five genomes were closed using the new tools described here. Three of the five genomes were finished via manual inspection and curation, with no additional PCR or sequencing required. A small number of PCR products were sequenced for gap closure and confirmation for two of these five genomes (*Streptomyces* and *Nonomuraea*).

Conclusion

Closing and finishing microbial genomes to high standards of accuracy has been complicated by the limited number of tools for generating long range sequence information from short read sequencers. In conjunction with a conventional fragment library, the NxSeq® Long Mate Pair Library kit and SPAdes genome assembly software enabled the conversion of six draft microbial genomes into closed and finished projects. The NxSeq Long Mate Pair Library kit described here generates PacBio-like "long span reads" from Illumina and Ion Torrent instruments. In the same amount of time required to construct a draft genome, it is now possible to close and finish microbial genomes with minimal effort and resources. Genome closure de novo now requires dramatically less manual work, reducing the number of unresolved repeats, unordered contigs, and split and missing genes. Long span mate pair libraries with pair distances up to 100 kb have also been constructed (data not shown), including libraries from plants and animals. These methods provide access to long range linkage information previously unavailable, enabling the analysis of structural variation, phasing and haplotype information from complex animal and plant genomes.

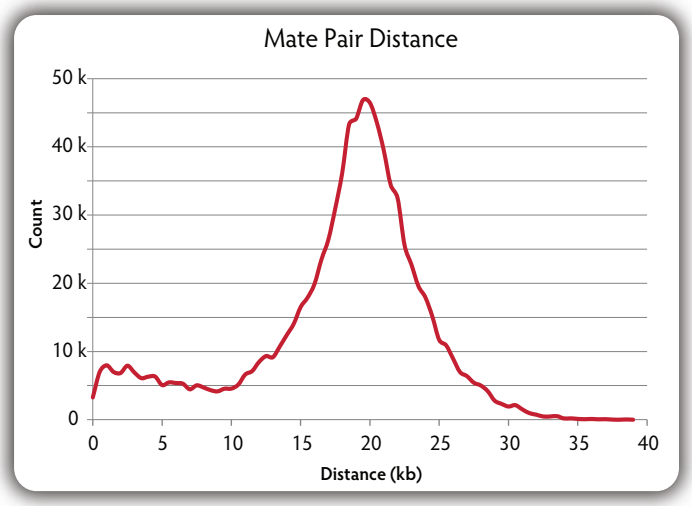

Figure 1. Mate pair distances of 20 kb *E. coli* K12 library generated using the NxSeq kit. *E. coli* 10 - 30 kb mate pairs were mapped against the closed and finished reference genome and the pair distance distribution was plotted in CLC Genomics Workbench.

References

1. Land ML et al. 2014. Quality scores for 32,000 genomes. *Stand Genomic Sci.* **9:**20.
2. Bankevich A et al. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455-77.
3. Blattner FR et al. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277:**1453-62.
4. Gurevich A et al. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29:**1072-5.

Acknowledgements Supported in part by NHGRI grants to ML and DM.

Microbial Species	Genome Size	%GC	Repeats	SPAdes Scaffolds	Manual Curation Scaffolds	Finishing Requirements
Escherichia coli K12 MG1665	4,641,652	50.8	75 (21 types) Max: 5,340 bp Highest copy# 11	2	1	manual curation
Thermus aquaticus Y51MC23 Chr. plus 4 plasmids	Chr: 2,158,963 4 Plasmids: 78,727; 69,906 16,597; 14,448	68.1	44 (14 types) Max: 3,544 bp Highest copy# 8	4 Chr + 5 Plasmid	1 Chr + 4 Plasmid	18 PCRs + Sanger
Staphylococcus aureus Tager 104	2,820,837	32.8	99 (27 types) Max: 16,174 bp Highest copy# 9	1	1	manual curation
Streptomyces sp. Strain A115	8,714,963 (linear)	71.0	26 (10 types) Max: 37,886 bp Highest copy# 8	2	1	4 PCRs + Sanger
Nonomuraea sp. Strain F4	10,464,236	70.7	96 (34 types) Max: 5,717 bp Highest copy# 8	1	1	1 PCR + Sanger
Bacillus amyloliquefaciens AP183	4,005,352	46.5	24 (7 types) Max: 12,647 bp Highest copy# 6	2	1	manual curation
Aeromonas hydrophila S14-451	5,090,173	60.7	32 (10 types) Max: 5,771 bp Highest copy# 8	2	1	manual curation

Table 1. Closing and finishing microbial genomes using NxSeq MP plus fragment libraries with SPAdes software.

Materials used include NxSeq® Long Mate Pair Library Kit (catalog number13000-1) and NxSeq Long Mate Pair Index Kit (catalog number 13100-1). The NxSeq Long Mate Library kit is provided with two protocols: 20 kb gel-based and 8 kb gel-free, bead-based. To close a microbial genome, the 20 kb Mate Pair protocol is recommended. The 8 kb Mate Pair protocol is recommended for spanning repeats <8 kb. For complex genomes, such as plant or animal genomes, we recommend constructing multiple mate pair libraries ranging from 2 - 20 kb, which would use both protocols.