illumina®

# *De Novo* Assembly of Bacterial Genomes

Using mate pair technology and the MiSeq® System to generate high-quality genome assemblies.

## Introduction

With recent advances in next-generation sequencing technology and bioinformatics tools, the possibilities for microbial genomics continue to grow, especially for *de novo* sequencing applications. Conventional approaches to assembling complete genomes include long-read technologies, which can be time-consuming and costly. Mate pair sequencing offers a more cost-effective strategy, delivering long contigs, high genomic coverage, and high-quality genome assemblies quickly and at a lower cost per sample.

The Nextera® Mate Pair Library Preparation Kit[1] provides an affordable, easy-to-use solution for mate pair sequencing, generating DNA libraries from as little as 1 µg of DNA. This application note describes how to use the Nextera kit, the MiSeq desktop sequencer, and the BaseSpace® Velvet *De Novo* Assembly App[2] for assembly of nine bacterial specimens in less than 4 days (Figure 1).

## Methods

### Library Construction

Genomic DNA samples from nine bacterial specimens were obtained from ATCC[3] (Table 1). Using the gel-free workflow of the Nextera Mate Pair Library Preparation Kit, DNA libraries were constructed from 1 µg genomic DNA in less than 2 days. The protocol used an initial "tagmentation" reaction to simultaneously fragment the DNA to > 1 kb and add biotinylated adapters to the ends of the molecules. Next, the tagmented DNA molecules were circularized and the ends of each fragment were joined. Circularized molecules were fragmented again and enriched by streptavidin beads, yielding smaller fragments (< 1 kb) containing the mate pair ends. Adapters were then added to these fragments for cluster generation and sequencing (Figure 2). Sequencing these smaller mate pair fragments provided information about genomic regions separated by large distances, enabling complete genome assembly.

**Table 1: Bacterial Species Sequenced**

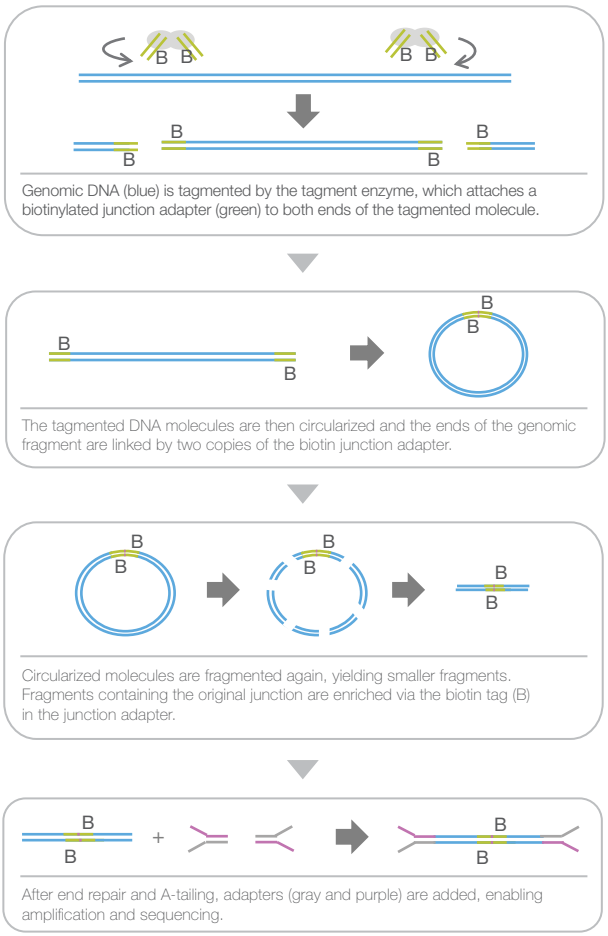
| Strain | NCBI Reference Genome ID |
|---|---|
| *Listeria monocytogenes* EGDe | 61583 |
| *Meiothermus ruber* DSM 1279 | 46661 |
| *Pedobacter heparinus* DSM 2366 | 59111 |
| *Klebsiella pneumoniae* MGH 78578 | 57619 |
| *Bacillus cereus* ATCC-10987 | 57673 |
| *Rhodobacter sphaeroides* 2.4.1 | 57653 |
| *Escherichia coli* K-12 substr. DH10B | 58979 |
| *Mycobacterium tuberculosis* H37 Ra | 58853 |
| *Escherichia coli* K-12 substr. MG1655 | 57779 |

### Sequencing

All bacterial libraries were pooled together for cluster generation and sequencing. The pool contained 2 replicates of each of the 9 libraries, yielding a total of 18 genomes. Libraries were loaded onto a MiSeq reagent cartridge and clustered on the MiSeq System. Paired-end sequencing was performed using MiSeq v2 reagents and 2 × 150 bp reads. Although longer read lengths are possible with the MiSeq System, 2 × 150 bp was determined to be the optimal read length for these samples. Sequencing generated 13 million reads passing filter, with a total run time of 32 hours.

**Figure 1: Mate Pair Sequencing Workflow for Small Genome Assembly**


**Figure 2: Nextera Mate Pair Workflow**




The Nextera Mate Pair Library Preparation Kit features a simple workflow that enables library construction in less than 2 days.

### Data Analysis

Primary analysis was performed in the BaseSpace environment. Base call (*.bcl) files generated by the MiSeq System were converted to FASTQ files[4]. Bacterial genomes were assembled using the Velvet *De Novo* Assembly App. The Quality Assessment Tool for Genome Assemblies (QUAST 2.3)[5] was used to evaluate assembly quality.

For best results, the adapters in the input FASTQ files should not be trimmed before using the Velvet *De Novo* Assembly App because the app contains a modified trimming algorithm. Lines in the sample sheet that specify adapter sequences should be removed before uploading data to BaseSpace. When using the Illumina Experiment Manager[6], navigate to the Assembly Workflow-Specific Settings:

1. Select **Create Sample Sheet**.
2. Select **MiSeq**.
3. Select **Small Genome Sequencing**.
4. Select **Assembly**.

To prevent adapter trimming, uncheck the **Use Adapter Trimming** and **Use Adapter Trimming Read 2** checkboxes. If reads stored in BaseSpace have already been processed by the on-instrument trimming algorithm, they can still be analyzed with the Velvet *De Novo* Assembly App; however, Illumina recommends using the trimming function within the app.

## Results

Sequencing generated approximately 13 million reads that passed quality filtering, corresponding to 0.6–0.9 million reads per sample. Coverage depth ranged from 24.24× to 68.02× across the 18 samples, with an average depth of 40.40×. Assembly metrics for each sample are shown in Table 2. The optimal k-mer size for each sample was chosen by maximizing the contig N50 length, which is performed automatically by the Velvet *De Novo* Assembly App. The observed N50 values demonstrated that high-quality assemblies were obtained even at low sequencing depth, and in some cases depended on the size of the genome (Table 2).

To assess assembly quality, each *de novo* assembly was aligned to its respective reference genome (Figure 3). Each assembly showed high correlation to the corresponding reference, demonstrating that the Nextera Mate Pair Library Preparation Kit enables accurate, nearly complete genome assemblies.

Genomic coverage was 96.59–99.87% across the 18 samples (Table 3), with an average coverage of 98.57%. Average gene coverage was 98.58%.

## Conclusions

The method outlined in this application note demonstrates the advantages of combining Nextera technology with the output and data analysis capabilities of the MiSeq System. With this workflow, 18 bacterial libraries can be prepared, sequenced, and assembled accurately in 4 days using a single sequencing run. The ability to sequence multiple samples simultaneously drastically reduced the cost per assembly. Depending on the degree of completion of a given microbial genome, it is possible for more than 18 libraries to be sequenced in one run. This application note demonstrates that the Nextera Mate Pair Library Preparation Kit and the MiSeq desktop sequencer provide an accurate, cost-effective method for *de novo* assembly of microbial genomes.

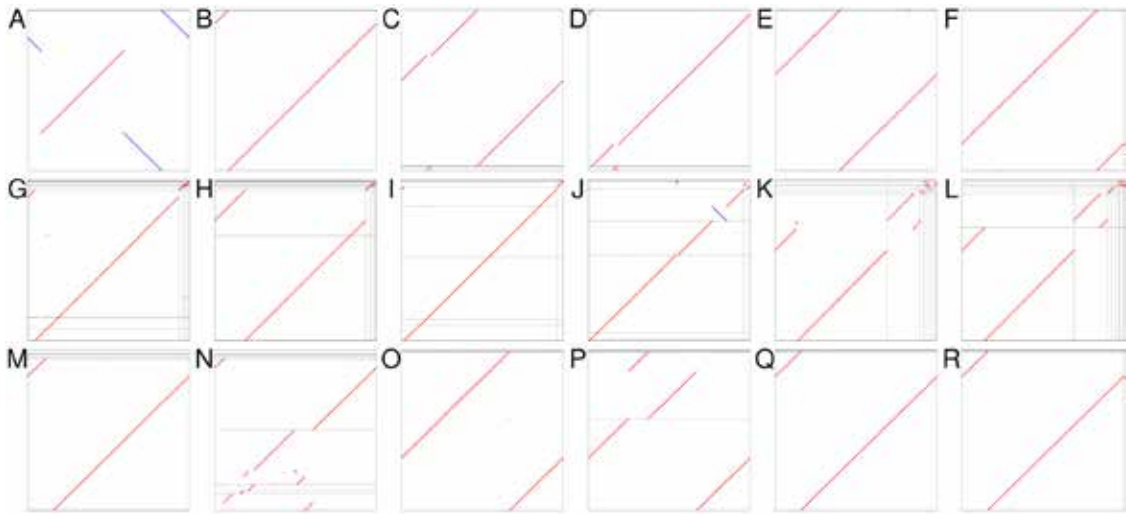
## Table 2: *De Novo* Sequence Assemblies

| Figure 3 Plot | Sample* | % GC | Reference Size (bp) | Coverage Depth (×) | K-mer Size | Contig N50 Length (bp) | Number of Contigs | Scaffold N50 Length (bp) | Number of Scaffolds | Assembly Length (bp) |
|---|---|---|---|---|---|---|---|---|---|---|
| A | *L. monocytogenes* (1) | 37.98 | 2,944,528 | 68.02 | 75 | 1,620,614 | 7 | 2,932,577 | 4 | 2,938,359 |
| B | *L. monocytogenes* (2) | 37.98 | 2,944,528 | 53.62 | 75 | 691,775 | 11 | 2,928,935 | 5 | 2,933,419 |
| C | *M. ruber* (1) | 63.38 | 3,097,457 | 55.43 | 75 | 203,917 | 26 | 3,001,248 | 9 | 3,100,592 |
| D | *M. ruber* (2) | 63.38 | 3,097,457 | 47.96 | 69 | 145,226 | 35 | 2,998,617 | 14 | 3,105,938 |
| E | *P. heparinus* (1) | 42.05 | 5,167,383 | 36.07 | 51 | 297,483 | 57 | 5,144,140 | 16 | 5,160,459 |
| F | *P. heparinus* (2) | 42.05 | 5,167,383 | 27.48 | 55 | 289,124 | 65 | 5,150,646 | 14 | 5,158,523 |
| G | *K. pneumoniae* (1) | 57.15 | 5,694,894 | 32.28 | 59 | 142,641 | 141 | 4,455,493 | 40 | 5,575,733 |
| H | *K. pneumoniae* (2) | 57.15 | 5,694,894 | 29.37 | 51 | 159,999 | 144 | 3,651,106 | 46 | 5,579,858 |
| I | *B. cereus* (1) | 35.50 | 5,432,652 | 24.24 | 43 | 90,460 | 122 | 1,710,854 | 24 | 5,384,726 |
| J | *B. cereus* (2) | 35.50 | 5,432,652 | 27.36 | 51 | 100,734 | 106 | 1,149,845 | 20 | 5,398,264 |
| K | *R. sphaeroides* (1) | 68.79 | 4,602,977 | 39.17 | 61 | 191,856 | 88 | 4,129,505 | 14 | 4,503,303 |
| L | *R. sphaeroides* (2) | 68.79 | 4,602,977 | 45.22 | 69 | 210,411 | 78 | 3,186,951 | 17 | 4,513,006 |
| M | *E. coli* DH10B (1) | 50.78 | 4,686,137 | 48.38 | 57 | 234,160 | 65 | 4,422,317 | 26 | 4,545,800 |
| N | *E. coli* DH10B (2) | 50.78 | 4,686,137 | 35.31 | 57 | 178,404 | 76 | 1,591,696 | 25 | 4,537,054 |
| O | *M. tuberculosis* (1) | 65.61 | 4,419,977 | 47.19 | 73 | 92,267 | 97 | 4,354,886 | 23 | 4,362,803 |
| P | *M. tuberculosis* (2) | 65.61 | 4,419,977 | 38.99 | 53 | 88,733 | 117 | 2,527,876 | 27 | 4,359,302 |
| Q | *E. coli* MG1655 (1) | 50.79 | 4,639,675 | 35.60 | 61 | 179,369 | 69 | 4,590,510 | 22 | 4,598,448 |
| R | *E. coli* MG1655 (2) | 50.79 | 4,639,675 | 35.43 | 55 | 198,090 | 64 | 4,595,933 | 18 | 4,603,074 |

* The replicate number for each strain is denoted in parentheses.

Reference sizes include plasmids. Coverage depth denotes the coverage after adapter trimming, excluding reads < 25 bp in length after trimming. The assembly data presented correspond to contigs > 500 bp in length except where "scaffold" is denoted. Reference genomes were obtained from the NCBI index of genomes[7].

## Figure 3: *De Novo* Assemblies Compared to Reference Genomes




These plots demonstrate the concordance between the bacterial genome assemblies and the respective reference genomes. Table 2 defines the strains associated with plots A–R. Reference genomes are shown on the X-axes and assemblies are shown on the Y-axes. A red line represents a consistent segment between the assembly and the reference. A blue line represents a consistent (inverted) segment. The vertical lines separate chromosomes and plasmids within the reference. The horizontal lines separate scaffolds in the assemblies. The misassemblies shown in plots A, J, and N could be refined with further analysis. These plots were generated using MUMmer[8].

**Table 3: Gene Coverage**

| Sample | Genomic Coverage (%) | Average Gap Size (bp) | Maximum Gap Size (bp) | Number of Gaps | Total Gap Length (bp) | Full Genes† | Partial Genes‡ | % Genes Covered§ |
|---|---|---|---|---|---|---|---|---|
| *L. monocytogenes* (1) | 99.84 | 935 | 2,088 | 5 | 4,674 | 2,928 | 3 | 99.09 |
| *L. monocytogenes* (2) | 99.58 | 1,175 | 4,979 | 7 | 12,423 | 2,937 | 3 | 99.39 |
| *M. ruber* (1) | 99.87 | 342 | 819 | 11 | 3,757 | 3,079 | 25 | 99.97 |
| *M. ruber* (2) | 99.78 | 418 | 2,576 | 16 | 6,682 | 3,072 | 27 | 99.81 |
| *P. heparinus* (1) | 99.67 | 644 | 6,086 | 26 | 16,746 | 4,295 | 36 | 99.82 |
| *P. heparinus* (2) | 99.67 | 456 | 4,274 | 36 | 16,427 | 4,279 | 51 | 99.79 |
| *K. pneumoniae* (1) | 97.71 | 1,369 | 24,956 | 95 | 130,092 | 5,111 | 87 | 98.00 |
| *K. pneumoniae* (2) | 97.70 | 1,431 | 22,095 | 91 | 130,261 | 5,103 | 84 | 97.79 |
| *B. cereus* (1) | 98.92 | 1,086 | 6,882 | 53 | 57,572 | 5,833 | 88 | 98.45 |
| *B. cereus* (2) | 99.06 | 879 | 6,681 | 57 | 50,122 | 5,846 | 83 | 98.59 |
| *R. sphaeroides* (1) | 97.64 | 2,078 | 65,543 | 52 | 108,070 | 4,299 | 69 | 97.63 |
| *R. sphaeroides* (2) | 97.82 | 2,372 | 65,543 | 42 | 99,643 | 4,305 | 69 | 97.76 |
| *E. coli* DH10B (1) | 96.72 | 3,576 | 114,350 | 43 | 153,750 | 4,181 | 25 | 96.65 |
| *E. coli* DH10B (2) | 96.59 | 3,191 | 114,350 | 50 | 159,572 | 4,164 | 37 | 96.53 |
| *M. tuberculosis* (1) | 97.92 | 1,129 | 10,372 | 81 | 91,486 | 3,946 | 93 | 98.90 |
| *M. tuberculosis* (2) | 97.94 | 917 | 10,372 | 99 | 90,786 | 3,932 | 107 | 98.90 |
| *E. coli* MG1655 (1) | 98.86 | 1,169 | 5,402 | 45 | 52,584 | 4,425 | 31 | 98.67 |
| *E. coli* MG1655 (2) | 98.94 | 1,116 | 5,612 | 44 | 49,096 | 4,430 | 28 | 98.72 |

† The number of genes covered at 100%

‡ The number of genes covered at less than 100% but longer than 100 bp

§ The total number of genes covered in both full and part

All metrics were calculated for contigs > 500 bp in length. A gap denotes a gap in coverage that intersects with at least one gene. Genomic coverage is defined as the ratio of the total number of aligned base pairs in the assembly to the genome size. Gene annotations were obtained from the NCBI index of genomes. The comparisons presented here include plasmids.

## Learn More

Visit www.illumina.com/miseq to learn more about the MiSeq System. For more information about the use of Illumina sequencing technology in microbial genomics, visit www.illumina.com/microbiology.

## References


1. Nextera Mate Pair Library Preparation Kit (www.illumina.com/products/nextera-mate-pair-sample-prep-kit.ilmn)
2. BaseSpace (basespace.illumina.com)
3. ATCC (www.atcc.org)
4. Data Processing of Nextera Mate Pair Reads on Illumina Sequencing Platforms Technical Note (www.illumina.com/documents/products/technotes/technote_nextera_matepair_data_processing.pdf)
5. Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29: 1072–5.
6. Illumina Experiment Manager (support.illumina.com/sequencing/sequencing_software/experiment_manager.ilmn)
7. NCBI FTP Index (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/)
8. Kurtz S, Phillippy A, Delcher AL, Smoot M, Schumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biol 5: R12.


**Illumina** • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY



Pub. No. 1270-2013-006 Current as of 11 November 2014

illumina®