De Novo Assembly of Small Genome Nextera® Mate Pair Libraries with a Single MiSeq® System Run

Obtain high-quality, low-cost genome assemblies with the Nextera Mate Pair Library Prep Kit and a single, multiplexed run on the MiSeq System.

Introduction

The presence of repeats can complicate the *de novo* assembly of genomes from short read data. To overcome this challenge, researchers have supplemented paired end read libraries with either long reads[1] or mate pair reads.[2] The latter refers to reads generated from the ends of a long fragment (eg, ≥ 3 kb) of genomic DNA. Although the strategies of combining several library types showed some success, the cost and complexity of multiple preparations can be high. The Nextera Mate Pair Library Preparation Kit offers a much simpler, cost-effective approach: it can be used to prepare up to 48 libraries and requires only 1 µg of DNA per organism.

A recent paper by Vasilinetc et al.,demonstrates that a single Nextera Mate Pair library can be used to generate high-quality assemblies of bacterial genomes without the cost, complexity, and high DNA input associated with combining multiple libraries.[3] In this application note, we extend the results of Vasilinetc et al. in several important ways. First, we report 12 high-quality assemblies that were generated from Nextera Mate Pair libraries on a single MiSeq System run. Second, we describe the complete end-to-end workflow, including data analysis using the BaseSpace® Informatics Suite (Figure 1). We also provide the per organism cost of generating the assembly using the complete Illumina workflow.

Methods

Library Preparation and Sequencing

Libraries were prepared with the Nextera Mate Pair Library Prep Kit (Illumina, Catalog No. FC-132-1001) with DNA samples from *H. Pylori*, *B. cereus*, *C. jejuni*, *E. coli* MG1655, *M. tuberculosis*, *R. sphaeroides*, *L. monocytogenes*, and *S. enterica* (strains obtained from ATCC, www.atcc.org/). The libraries were indexed, pooled together, and sequenced to 2 × 250 read length on a MiSeq System with MiSeq Reagents v2 (Illumina, Catalog No. MS-102-2003).

Data Analysis

Genome assembly was performed with the BaseSpace SPAdes App, which is designed to assemble small genomes from single-cell and standard bacterial data sets.[4] This App also utilizes the algorithm described in Vasilinetc et al. The data quality was assessed with QUAST.[5] The genomes were annotated with the BaseSpace Prokka App, which quantifies and identifies genomic features such as genes, transfer RNAs, and ribosomal RNAs.[6] The annotated sequence data was then uploaded to NCBI using the BaseSpace short read archive (SRA) Submission App.[7] All data sets and applications are freely available on the BaseSpace website.*

Results

Data sets from all 12 bacterial strains were assembled with the SPAdes genome assembler (Table 1). Due to high coverage results, the data were down sampled to 70× before assembly, to simulate the case of larger genomes, higher multiplexing, or lower output. Down sampling with the FASTQ Toolkit, is an optional step that can be used if more than 1 gigabase is generated for any organism. This keeps the data range within the memory limitations of the SPAdes App. Genome lengths ranged from 1641–5224 kb, the fraction of the genome covered ranged from 98.0–99.9%, and misassemblies ranged from 0–64. One replicate of *M. tuberculosis* likely produced high

* To access the Nextera Mate Pair data in BaseSpace 1) login to an existing account on basespace.illumina.com 2) at the top of the home page, click the Public Data icon 3) in the right-hand navigation panel, under Categories, click the De Novo Assembly button 4) click the "MiSeq v3: Nextera Mate Pair (12 Bacterial Samples)" link.

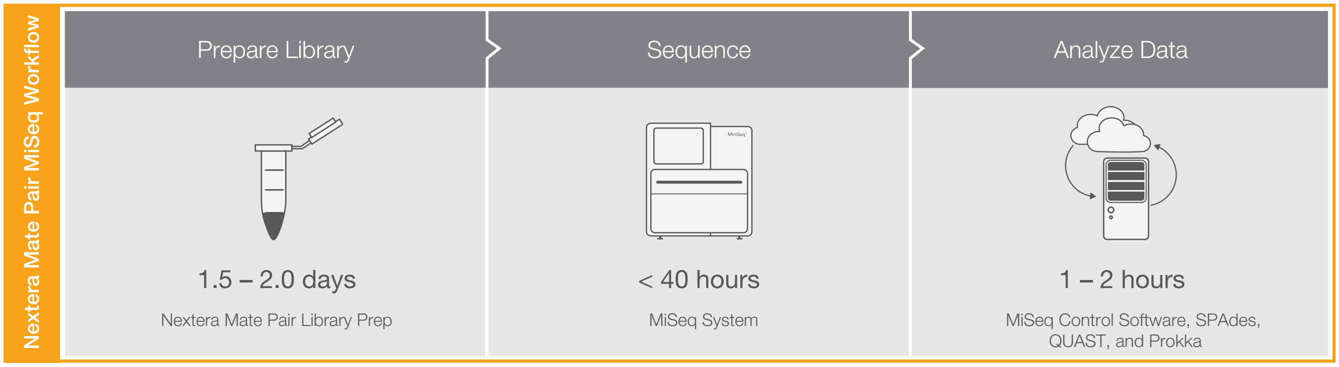
Nextera Mate Pair MiSeq Workflow
Prepare Library
Sequence
Analyze Data
1.5 – 2.0 days
Nextera Mate Pair Library Prep
< 40 hours
MiSeq System
1 – 2 hours
MiSeq Control Software, SPAdes, QUAST, and Prokka

Figure 1: End-to-End Illumina Workflow—This figure outlines the full workflow from library preparation through annotation. SPAdes and Prokka are available as easy-to-use Apps in BaseSpace.

Table 1: Assembly Results for 12 Multiplexed Strains Sequenced in a Single Run with the MiSeq System

Organism[a]	Genome Length (kb)	# Contigs (> 1 kb)	Contig NGA50 (kb)[b]	Misassemblies	% Covered
B. cereus	5224	16	1705	0	98.9
C. jejuni	1641	4	355	4	99.6
E. coli MG1655	4642	14	662	1	99.3
E. coli MG1655	4642	14	696	1	99.3
E. coli MG1655	4642	12	696	0	99.3
H. pylori	1668	7	318	3	99.9
L. monocytogenes	2945	6	1495	1	99.3
L. monocytogenes	2945	8	1495	0	99.1
M. tuberculosis[c]	4412	60	121	64	98.0
M. tuberculosis	4412	24	686	19	99.5
R. sphaeroides	4131	21	827	3	99.4
S. enterica	4857	18	767	0	99.2

a. Data sets for each organism were down sampled to 70×.
b. The NGA50 size is defined as the value N such that 50% of the finished sequence is contained in contigs whose alignments to the finished sequence are of size N or larger.[8]
c. This replicate of *M. tuberculosis* likely produced high misassemblies due to a library preparation issue.

misassemblies (64) and a lower coverage (98.0%) due to a library preparation issue. This sample showed an uncharacteristic coverage distribution and a second attempt to sequence the library also yielded poor results. The remaining 11 strains had low numbers of misassemblies ranging from 0–19 and high coverage values ranging from 98.9–99.9%. The data also show excellent reproducibility across the replicate samples.

Conclusions

The Vasilinetc et al. publication demonstrated that high-quality assembly of small organisms is possible by sequencing a single Nextera Mate Pair library. We have extended their results to show that 12 assemblies can be generated from a single MiSeq System run. All the assemblies are high quality as measured by the NGA50 values, the number of contigs, few misassemblies, and high genome coverage percentages. Additional analysis, including annotation and submission to the NCBI SRA, can be performed without any bioinformatics training with the BaseSpace Suite. We have also calculated the per organism costs and report that the Nextera Mate Pair Library Prep Kit, combined with a MiSeq 2 × 250 sequencing run, can sequence 12 libraries for under $250 USD per organism.† With the Nextera Mate Pair Library Prep Kit, multiple small genomes can be sequenced on a single, MiSeq System run—saving significant time, sample costs, and labor costs.

Learn More

To learn more about the Nextera Mate Pair Library Prep Kit, visit:
www.illumina.com/products/nextera-mate-pair-sample-prep-kit.html

For more on the MiSeq System, visit:
www.illumina.com/systems/miseq.html

Ordering Information

System and Reagents	Catalog No.
MiSeq System	SY-410-1003
MiSeq v2 (300-cycles)	MS-102-2003
Library Preparation Kit	
Nextera Mate Pair Library Prep Kit	FC-132-1001

References

1. Bashir A, Klammer AA, Robins WP, et al. A hybrid approach for the automated finishing of bacterial genomes. *Nat Biotechnol.* 2012;30:701–707.
2. Earl D, Bradnam K, John St. John, et al. Assemblathon 1: a competitive assessment of *de novo* short read assembly methods. *Genome Res.* 2011;21:2224–2241.
3. Vasilinetc I, Prjibelski AD, Gurevich A, Korobeynikov A, and Pevzner PA. Assembling short reads from jumping libraries with large insert sizes. *Bioinformatics.* 2015;31:3262–3268.
4. BaseSpace SPAdes App (www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps/spades-genome-assembler-1989988.html). Accessed 14 Jan 2016.
5. Gurevich A, Saveliev V, Vyahhi N, and Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics*. 2013;29:1072–1075.
6. BaseSpace Prokka App (www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps/prokka-genome-annotation-590590.html). Accessed 14 Jan 2016.
7. BaseSpace SRA Submission App (www.illumina.com/informatics/research/sequencing-data-analysis-management/basespace/basespace-apps/sra-submission-147147.html). Accessed 14 Jan 2016.
8. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, and Yorke JA. The MaSuRCA genome assembler. *Bioinformatics*. 2013;21:2669–2677.

† Costs calculated based on 2015 price lists. Pricing subject to change.

For Research Use Only. Not for use in diagnostic procedures.