# *De novo* Bacterial Sequencing on the MiSeq® System

High-quality, paired-end reads ensure superior *de novo* genome assembly.

## Introduction

The MiSeq System uses Illumina's proven TruSeq® sequencing by synthesis (SBS) reversible terminator chemistry, matching the data quality of a HiSeq® 2000 run. This application note describes sequencing the same library prepared from a reference strain of *Escherichia coli* on both HiSeq 2000 and MiSeq platforms, and *de novo* assembly. Bacterial genome sequencing output from the MiSeq instrument is compared with recent *E. coli* data sets from the Ion Torrent Personal Genome Machine (PGM), demonstrating the importance of high-quality, paired-end reads for *de novo* genome assembly.

## Sequencing and *De novo* Assembly Workflow

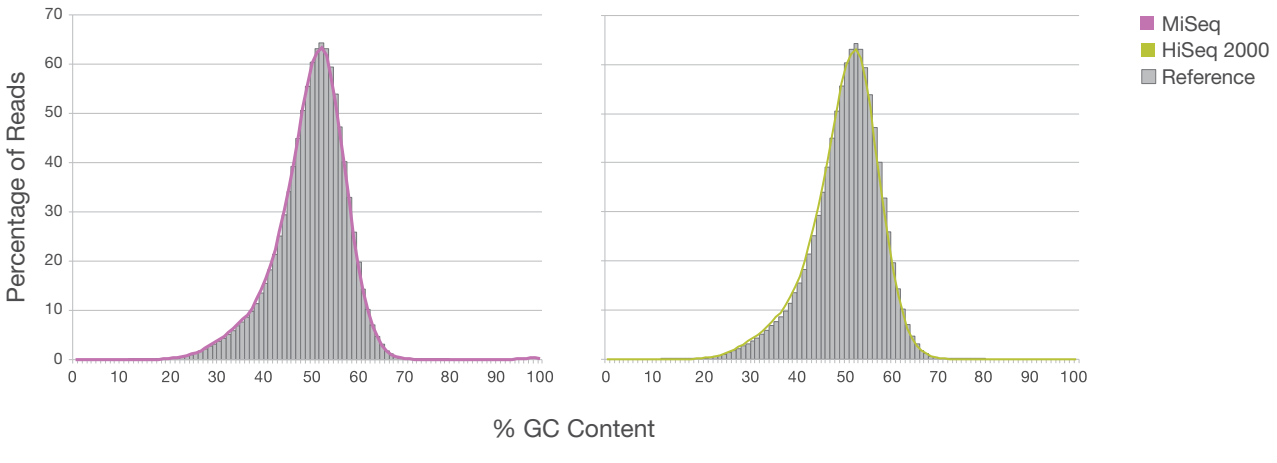
Genomic DNA isolated from the well-characterized *E. coli* strain MG1655 was used to prepare a sequencing library using Illumina's TruSeq library preparation reagents. For sequencing on the MiSeq instrument, samples were placed in the reagent cartridge and loaded on the instrument along with the flow cell. All subsequent steps were performed on the instrument, including cluster generation and 2 × 150 paired-end sequencing, in less than 27 hours. The MG1655 library was also used to prepare clusters on a HiSeq flow cell using a cBot™ and TruSeq v3 clustering reagents, followed by 2 × 100 bp paired-end sequencing on the HiSeq 2000, which was completed in 10.5 days.

For resequencing, data analysis was performed directly on the MiSeq integrated computer, requiring no specialized servers or computing facilities. Both sets of basecall files from HiSeq and MiSeq were

**Table 1: MiSeq and HiSeq 2000 Run and Assembly Metrics for *E. coli* MG1655 Library**

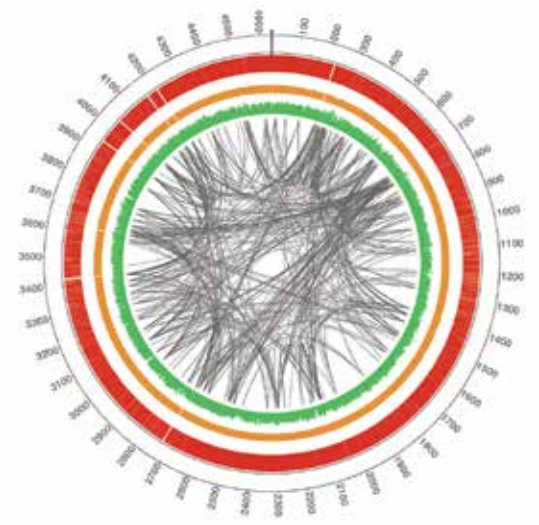
| Metric | MiSeq | HiSeq 2000 |
|---|---|---|
| Raw Clusters | 840/mm² | 794/mm² |
| Clusters Passing Filter | 799/mm² | 726/mm² |
| % Bases ≥ Q30 | Read 1: 91.9<br>Read 2: 87.5 | Read 1: 89.3<br>Read 2: 86.1 |
| % Total Bases ≥ Q30 | 89.7 | 87.7 |
| Assembly Size | 4,573,128 | 4,571,539 |
| Number of Contigs | 125 | 125 |
| N50 (bp (Contigs)) | 148,627 (11) | 148,770 (12) |
| Avg Contig Size (bp) | 36,585 | 36,571 |
| % Gaps in Scaffolds | 0.07 | 0.07 |
| Max Contig (bp) | 311,761 | 305,482 |
| % GC | 50.71 | 50.71 |

**Figure 1: Even Coverage Across GC Content**




Percentage of reads passing filter for MiSeq (purple, left) and HiSeq 2000 (green, right), over % GC content compared to the reference sequence, showing even genome coverage for both MiSeq and HiSeq 2000.

**Figure 2: Circos Plot of MiSeq *E. coli* Assembly**



Circos plot showing *E. coli* MG1655 assembly from MiSeq data. Genomic position is indicated on the white outermost ring. Genes are indicated in red (numbers are in thousands). Coverage from the *de novo* assembly is represented in orange, aligned to the complete reference using BLAST and filtered for 99% identity. Percent GC content is indicated in green, Grey lines show repetitive sequences and their relationships across the genome.

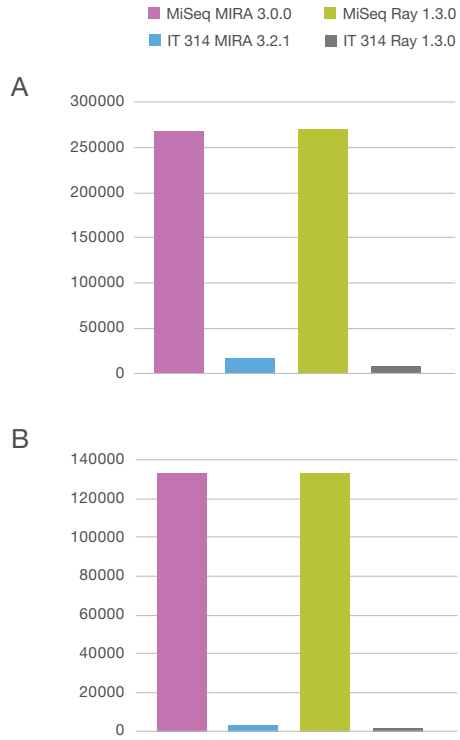
analyzed using CASAVA 1.8a5, and *de novo* assembly was completed using Velvet. For the *de novo* assembly comparison between MiSeq and Ion Torrent, the open access assemblers MIRA[1] and Ray[2] were used on MiSeq data down-sampled to 50× and compared to the entire data set from Ion Torrent reads[3]. These open source assembly tools are reported to work well with both Illumina and Ion Torrent data[3], and produced results comparable to Velvet for the MiSeq data.

## Results and Data Analysis

Data generated from the MiSeq and HiSeq systems showed similar cluster density and numbers of clusters passing filter. *De novo* assembly metrics from the HiSeq and MiSeq reads are very similar (Table 1). Comparison of HiSeq and MiSeq data with the reference sequence illustrates equivalent coverage over a range of GC content (Figure 1). Data from the MiSeq assembly overlaying the *E. coli* reference sequence are shown in a Circos plot (Figure 2), demonstrating excellent coverage over the entire genome.

*De novo* assembly data from the 2 × 150 bp MiSeq run was compared with Ion Torrent data[3]. To make an equal comparison, MiSeq data was down-sampled to 50× coverage, comprising 231 Mb, or approximately 1/7th of the data. Both the max contig length and N50 values were vastly superior in the down-sampled MiSeq data compared to the entire Ion Torrent data set (Figures 3A and B).

**Figure 3: *De novo* Assembly Contigs and N50 Length**




A. Max contig length in bp for MiSeq assemblies (purple) using MIRA v3.0.0 and Ray v1.3.0 (green) compared to Ion Torrent assemblies (blue) using MIRA v3.2 and Ray v1.3.0 (grey). B. N50 is the length of the smallest contig in the set that contains the fewest (largest) contigs whose combined length represents at least 50% of the assembly.

## Conclusions

Using the same library preparation from bacterial DNA, sequencing on MiSeq was shown to be very comparable to HiSeq; both platforms yield high-quality data with > 85% bases above Q30 with even GC coverage. *De novo* assembly with these data also produce similar results, with excellent coverage of the reference sequence. Sequencing results generated on the MiSeq System are highly predictive of those delivered by the high-throughput HiSeq 2000 sequencing platform, making MiSeq ideal for piloting larger studies or performing independent experiments requiring speed and accuracy. For *de novo* assembly, the importance of high-quality, paired-end MiSeq reads is readily apparent compared to Ion Torrent. The high quality assembly produced from MiSeq paired-end reads show that better data give a more accurate picture of the genome.

## References

1. http://www.chevreux.org/projects_mira.html
2. http://sourceforge.net/projects/denovoassembler/files
3. http://pathogenomics.bham.ac.uk/blog/2011/05/first-look-at-ion-torrent-data-de-novo-assembly

**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. 770-2011-009 Current as of 10 November 2014

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA