E. coli Sequencing on the MiSeq® System and Ion Torrent PGM System

Introduction

The MiSeq personal sequencing system uses Illumina's proven TruSeq® sequencing by synthesis (SBS) reversible terminator chemistry to generate highly accurate base-by-base sequence. This results in robust base calls across the genome, including within repetitive regions and homopolymer stretches. This application note compares sequencing output from the MiSeq system with publicly available data from the Ion Torrent Personal Genome Machine (PGM) for the same laboratory strain of *Escherichia coli*. Quality metrics including Q scores, percent error-free reads, homopolymer-associated read-level indels, putative false positive indel calls, and normalized GC coverage are compared between the two sequencing platforms.

Results

High Read Quality and Yield

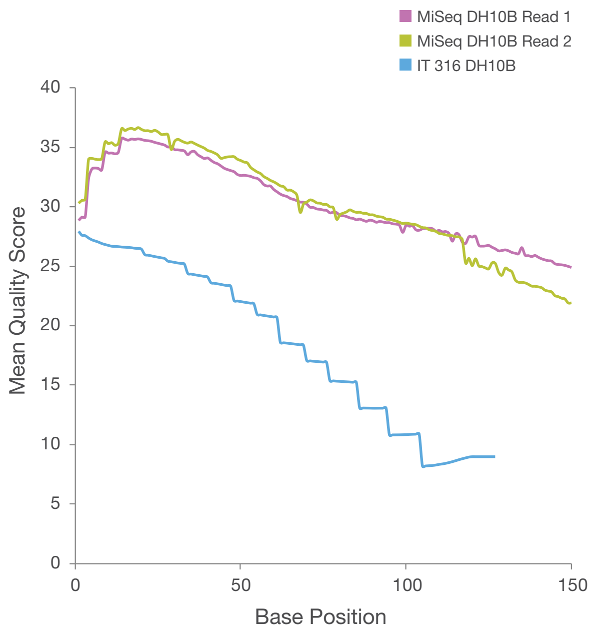
Quality scores measure the probability that a base is called correctly, and are assigned for each base in a read using a Phred-like algorithm[1]. Higher quality scores indicate a higher probability that a base is correctly called. Figure 1 shows the mean quality scores at each base position for a MiSeq run at 2 × 150 bp. Quality scores for both read 1 and read 2 are > Q20 (99% accuracy), and are highly comparable to one another. By comparison, the quality of Ion Torrent data run on the 316 chip falls dramatically after 50 bp, yielding exceedingly poor quality data (< Q10, < 90% accuracy) around 100 bp, which is close to the average length of a typical PGM run[2].

Data generated from a single 2 × 150 bp MiSeq run yield ~70–80× more data than that generated using the Ion Torrent PGM 314 chip, and > 10× more data than the Ion Torrent PGM 316 chip (Table 1). More high quality data can be generated on the MiSeq system, eliminating the need to perform multiple rounds of sequencing.

Error-Free Reads

Error rates are determined by comparing sequences to a well-characterized reference genome, and the percentage of error-free reads is indicative of the overall error rate from a sequencing run and a good indicator of the amount of usable data[3]. Using the same *E. coli* strain DH10B, sequence from the MiSeq run has higher numbers of error-free (perfect) reads than sequence generated on the Ion Torrent 316 chip (Figure 2). Results using a different *E. coli* laboratory strain (MG1655), show that the percentage of error-free reads are similar for 2 × 150 bp and down-sampled 2 × 100 bp read lengths on the MiSeq system[4], suggesting that differences between strains or different library preparations do not have an effect on error rate.

Figure 1: Quality Scores by Base Position

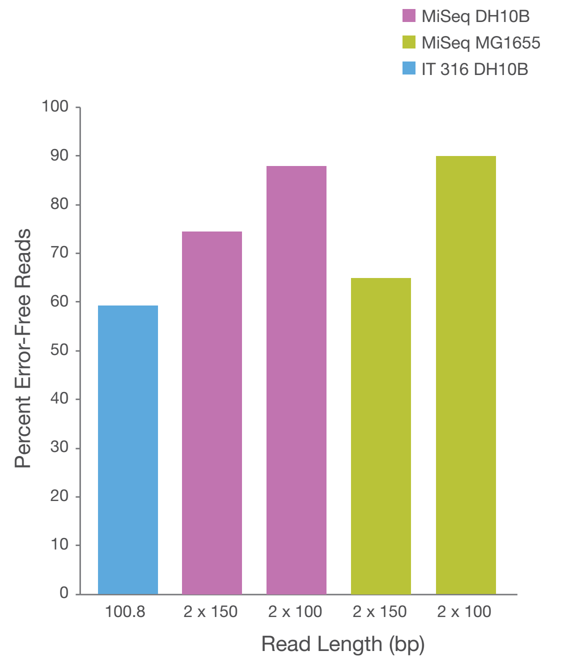


Predicted quality scores at corresponding base positions for DH10B read 1 and read 2 of a 2 × 150 bp MiSeq run, and for an Ion Torrent PGM run using the 316 chip.

Table 1: Run Summary

Source	Instrument	Number of Runs	*E. coli* Strain	Avg Yield/Run	Avg Depth/Run
Illumina	MiSeq System	1	K-12 DH10B	2.0 Gb	421×
Illumina	MiSeq System	1	K-12 MG1655	1.7 Gb	393×
LIFE Technologies[8]	Ion Torrent PGM (316)	1	K-12 DH10B	0.175 Gb	37×
EdgeBio[7]	Ion Torrent PGM (314)	6	K-12 DH10B	0.024 Gb	5×

Figure 2: Percent of Error-Free Reads

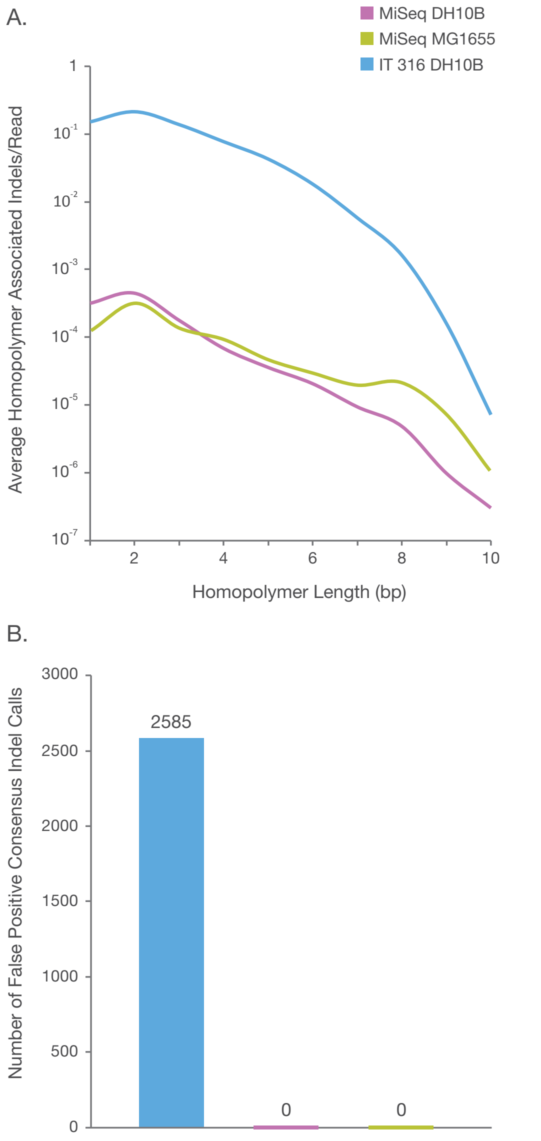


Regardless of bacteria strain, MiSeq sequences have a higher percentage of perfect reads than sequence of average length generated from the Ion Torrent PGM 316 chip.

Homopolymer-Associated Indels and False Positive Calls

Homopolymers are sequence regions of identical bases, and base calling inaccuracies within homopolymer regions lead to false positive insertions and deletions (indels) upon alignment. Although long homopolymers (> 6 bp) occur rarely within most genomes, even homopolymers of short length (i.e., 2 bp) can have deleterious effects on alignment accuracy. Unlike sequencing methods that use flow-based procedures, Illumina SBS technology uses natural competition from all four reversible terminator nucleotides and single-base detection to ensure accurate basecalls within homopolymer stretches. Figure 3A shows that sequence data from the Ion Torrent 316 chip contains more homopolymer-associated indels per read compared to MiSeq reads. Figure 3B shows that Ion Torrent 316 data results in a larger number of putative false positive consensus indel calls with Q > 20. No false positive indels were detected using MiSeq data from same DH10B strain or from a different *E. coli* strain, MG1655.

Figure 3: Homopolymer-Associated Indels and Putative False Positive Calls



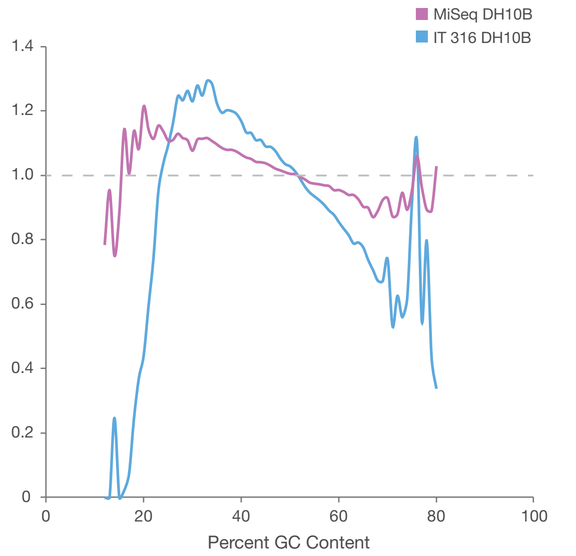
A. MiSeq sequence data contain far fewer homopolymer-associated read level indels than reads generated from the Ion Torrent 316 chip.

B. Ion Torrent data results in a large number of putative false positive consensus indel calls, compared to zero MiSeq false positive indel calls.

Normalized GC Coverage

Coverage uniformity is an important measure of data comprehensiveness. Lack of coverage uniformity in areas of the genome that are AT or GC rich means that any variation in those regions will be poorly covered or omitted. Figure 4 shows mean normalized coverage over percent GC content for MiSeq and Ion Torrent 316 sequence reads. The distribution of MiSeq reads is more uniform (closer to normalized coverage = 1.0) over the entire range of GC content. In particular, the Ion Torrent reads show dramatic drops at both the low and high ends of GC coverage (0–20% and ~70–80%, respectively), suggesting a lack of coverage uniformity across the DH10B genome.

Figure 4: Normalized GC Coverage



Mean normalized coverage as a function of % GC for *E. coli* strain DH10B sequenced on the MiSeq system and the Ion Torrent 316 chip. An ideal GC curve is represented by the dashed line at coverage = 1.0

Conclusions

Sequence quality has a direct impact on the usefulness and biological relevance of your data. Several variables account for overall sequence quality, including quality scoring, read error rate, alignment, and coverage. Comparing the same *E. coli* laboratory strain demonstrates that sequence data from the MiSeq system is of better quality and higher yield than comparable data generated on the Ion Torrent PGM using either the 314 or 316 chip.

Methods

Bacterial Genome Sequencing on MiSeq System

Genomic DNA isolated from the well-characterized *E. coli* K-12 strains DH10B and MG1655 were used to prepare sequencing libraries 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.

Data Analysis

Primary data analysis was performed directly on the MiSeq integrated computer, requiring no specialized servers or computing facilities. Ion Torrent data was used from publicly-available sources (See Data Sources for details).

Averaged quality score comparisons were generated by loading FASTQ files into Picard[5] (MeanQualityByCycle.jar) to generate mean quality scores.

Perfect reads were analyzed by generating a samtools calmd[6] (samtools calmd –f reference.fa in.bam > out.sam) and then parsing the out.sam file and determining the number of reads in which the NM field is equal to zero.

Homopolymer-associated indels were determined by generating a samtools mpileup (samtools mpileup –f reference.fa in.bam > out.txt) and then parsing the out.txt file for read level indels. For each read level indel, the ten bases on either side of the indel were determined from the reference genome and used to determine the largest homopolymer (two or more identical bases) adjacent to the indel.

Putative consensus indel calls were determined by the following procedure:

```
samtools mpileup -uf reference.fa in.bam | bcftools view
-bvcg - > var.raw.bcf

bcftools view var.raw.bcf | vcfutils.pl varFilter > var.flt.
vcf  (keeping only indels w/ quality >= 20).
```

Mean normalized coverage as a function of %GC was calculated using CollectGcBiasMetrics.jar from Picard tools

All input bam files were filtered to keep reads having a mapping quality score > 0 and remove reads that did not pass platform quality controls.

```
samtools view -bq 0 -F 512 –o filtered.bam in.bam
```

Data Sources

Analysis was performed using four sets of publicly-available data from the *E. coli* DH10B laboratory strain sequenced on Ion Torrent PGM 314 and 316 chips.

- 6 runs sequenced by EdgeBio on 314 chips[7]
- 1 run sequenced by LIFE Technologies on 316 chips[8]

References

1. www.illumina.com/truseq/quality_101/quality_scores.ilmn
2. www.iontorrent.com/lib/images/PDFs/performance_overview_application_note_041211.pdf
3. http://www.illumina.com/truseq/quality_101/error_free_reads.ilmn
4. www.illumina.com/systems/miseq/ecoli.ilmn
5. http://picard.sourceforge.net/
6. http://samtools.sourceforge.net/mpileup.shtml
7. www.edgebio.com/data/ion/delivery/Auto_1EB-16-0039010CA/Default_Report.php
8. https://iontorrent.box.net/shared/dgjscpc8o9ic1u8uky4f

Learn More

Visit www.illumina.com/miseq to learn more about the next revolution in personal sequencing.

FOR RESEARCH USE ONLY

Pub. No. 770-2011-012 Current as of 10 November 2014