Analysis of Inaccuracies in Ion Torrent’s Long Read Application Note

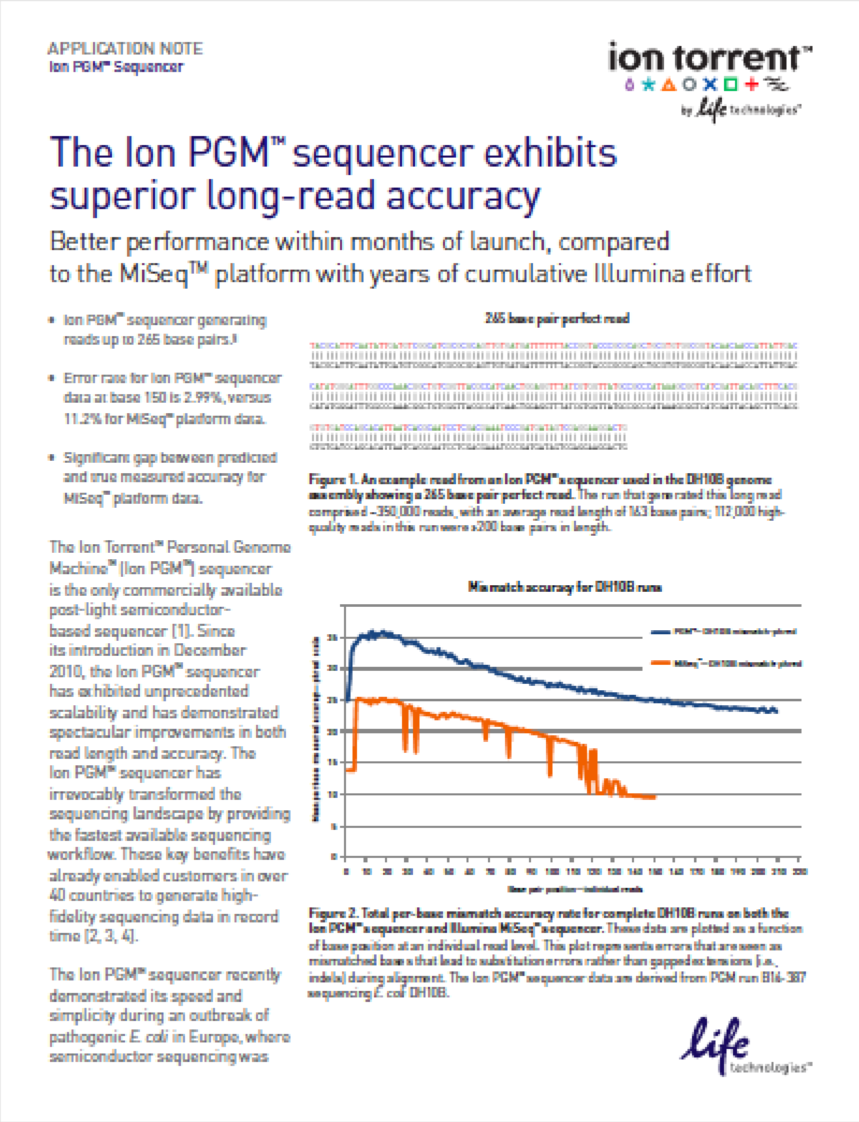
August 30, 2011

Information Contained in this Slide Deck will:

- Present Ion Torrent application note claims
- Provide accurate representation of MiSeq system data
- Propose directly comparable alternative analysis approaches

Ion Torrent Application Note Overview

- **The Ion PGM™ sequencer exhibits superior long-read accuracy**
  - **Better performance within months of launch, compared to the MiSeq™ platform with years of cumulative Illumina effort**
- Ion PGM™ sequencer generating reads up to 265 base pairs.
- Error rate for Ion PGM™ sequencer data at base 150 is 2.99%, versus 11.2% for MiSeq platform data.
- Significant gap between predicted and true measured accuracy for MiSeq platform data.
- http://www.iontorrent.com/applications-pgm-accuracy/



APPLICATION NOTE
Ion PGM™ Sequencer

ion torrent™
by life technologies™

The Ion PGM™ sequencer exhibits superior long-read accuracy

Better performance within months of launch, compared to the MiSeq™ platform with years of cumulative Illumina effort

- Ion PGM™ sequencer generating reads up to 265 base pairs.[1]
- Error rate for Ion PGM™ sequencer data at base 150 is 2.99%, versus 11.2% for MiSeq™ platform data.
- Significant gap between predicted and true measured accuracy for MiSeq™ platform data.

The Ion Torrent™ Personal Genome Machine™ (Ion PGM™) sequencer is the only commercially available post-light semiconductor-based sequencer [1]. Since its introduction in December 2010, the Ion PGM™ sequencer has exhibited unprecedented scalability and has demonstrated spectacular improvements in both read length and accuracy. The Ion PGM™ sequencer has irrevocably transformed the sequencing landscape by providing the fastest available sequencing workflow. These key benefits have already enabled customers in over 40 countries to generate high-fidelity sequencing data in record time [2, 3, 4].

The Ion PGM™ sequencer recently demonstrated its speed and simplicity during an outbreak of pathogenic *E. coli* in Europe, where semiconductor sequencing was

Figure 1. An example read from an Ion PGM™ sequencer used in the DH10B genome assembly showing a 265 base pair perfect read. The run that generated this long read comprised ~350,000 reads, with an average read length of 163 base pairs; 112,000 high-quality reads in this run were ≥200 base pairs in length.

Mismatch accuracy for DH10B runs
PGM™ – DH10B mismatch phred
MiSeq™ – DH10B mismatch phred
0
5
10
15
20
25
30
35
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
Base pair position – individual reads

Figure 2. Total per-base mismatch accuracy rate for complete DH10B runs on both the Ion PGM™ sequencer and Illumina MiSeq™ sequencer. These data are plotted as a function of base position at an individual read level. This plot represents errors that are seen as mismatched bases that lead to substitution errors rather than gapped extensions (i.e., indels) during alignment. The Ion PGM™ sequencer data are derived from PGM run B16-387 sequencing *E. coli* DH10B.

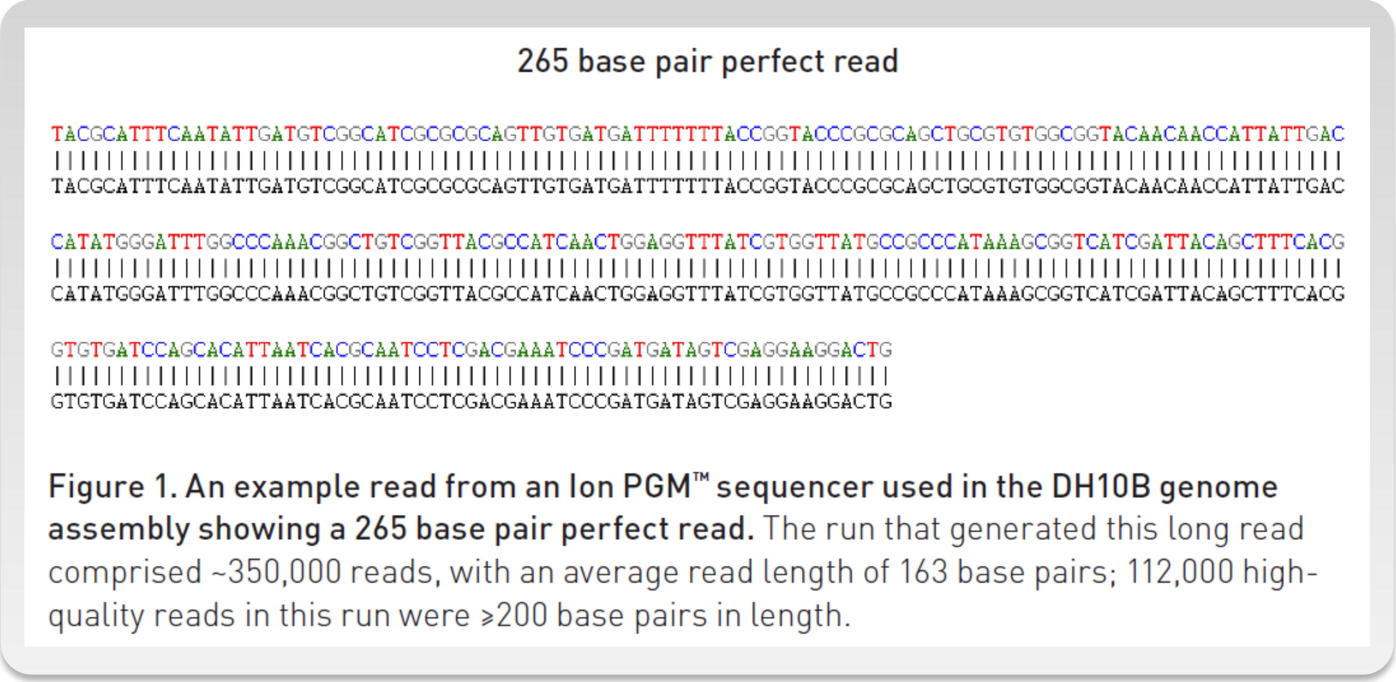
life technologies™

Ion Torrent Claims

1. The PGM exhibits superior long-read accuracy
2. The PGM shows superior measured mismatch accuracy at all base positions
3. The MiSeq system shows an order of magnitude difference between predicted and measured quality values
4. The MiSeq system significantly underperforms compared to PGM for consensus mismatch quality
5. The PGM has higher genome coverage than the MiSeq system

Claim #1: Superior Long-read Accuracy

- Important but omitted information
  - Serial nucleotide addition chemistries suffer from indel errors caused by homopolymeric regions
  - Indel errors were not included in analyses
  - False-positive indel calls can't be removed without also losing true positive indel calls
  - Less that 1/3 of reads were 200 bp with "long read" chemistry



265 base pair perfect read

```
TACGCATTTCAATATTGATGTCGGCATCGCGCGCAGTTGTGATGATTTTTTTACCGGTACCCGCGCAGCTGCGTGTGGCGGTACAACAACCATTATTGAC
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
TACGCATTTCAATATTGATGTCGGCATCGCGCGCAGTTGTGATGATTTTTTTACCGGTACCCGCGCAGCTGCGTGTGGCGGTACAACAACCATTATTGAC

CATATGGGATTTGGCCCAAACGGCTGTCGGTTACGCCATCAACTGGAGGTTTATCGTGGTTATGCCGCCCATAAAGCGGTCATCGATTACAGCTTTCACG
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
CATATGGGATTTGGCCCAAACGGCTGTCGGTTACGCCATCAACTGGAGGTTTATCGTGGTTATGCCGCCCATAAAGCGGTCATCGATTACAGCTTTCACG

GTGTGATCCAGCACATTAATCACGCAATCCTCGACGAAATCCCGATGATAGTCGAGGAAGGACTG
|||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
GTGTGATCCAGCACATTAATCACGCAATCCTCGACGAAATCCCGATGATAGTCGAGGAAGGACTG
```

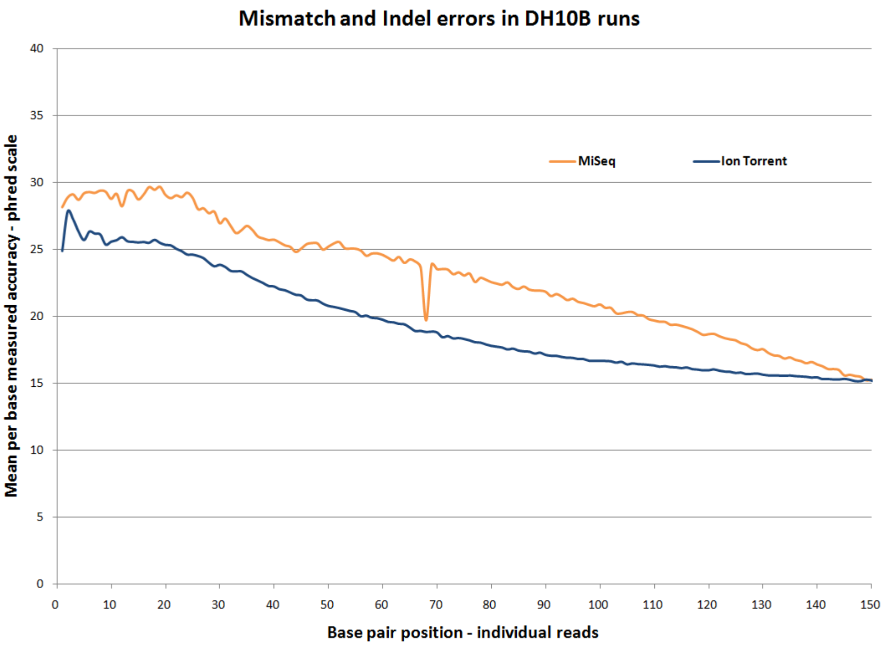
Figure 1. An example read from an Ion PGM™ sequencer used in the DH10B genome assembly showing a 265 base pair perfect read. The run that generated this long read comprised ~350,000 reads, with an average read length of 163 base pairs; 112,000 high-quality reads in this run were ≥200 base pairs in length.

MiSeq TOTAL Accuracy Outperforms PGM

- Ion Torrent inaccurately represents error rates by leaving indels out of analysis

- Recommended analysis approach:
  - Accuracy comparisons should include both mismatches and indel errors

- Result:
  - **With indel errors included, the MiSeq system total accuracy outperforms PGM**



1.11% of unmapped reads were removed from Ion Torrent data

Claim #2: Superior Mismatch Accuracy

- Important but omitted information
  - Ion Torrent compared extensively trimmed PGM data to untrimmed MiSeq data
  - Analysis included MiSeq data properly flagged as not passing quality filter (non-PF reads)
  - Because PGM data is variably trimmed and MiSeq data is not, comparing accuracy at any particular base position (such as position 150) is misleading

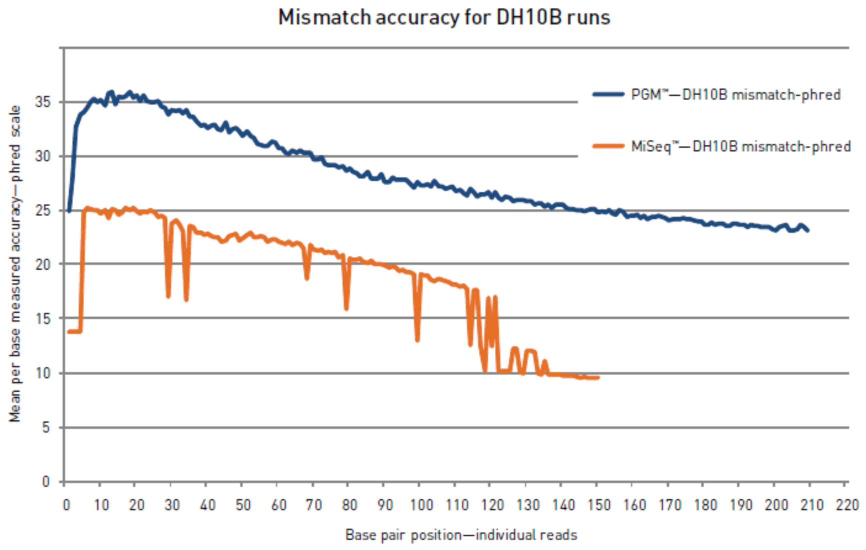
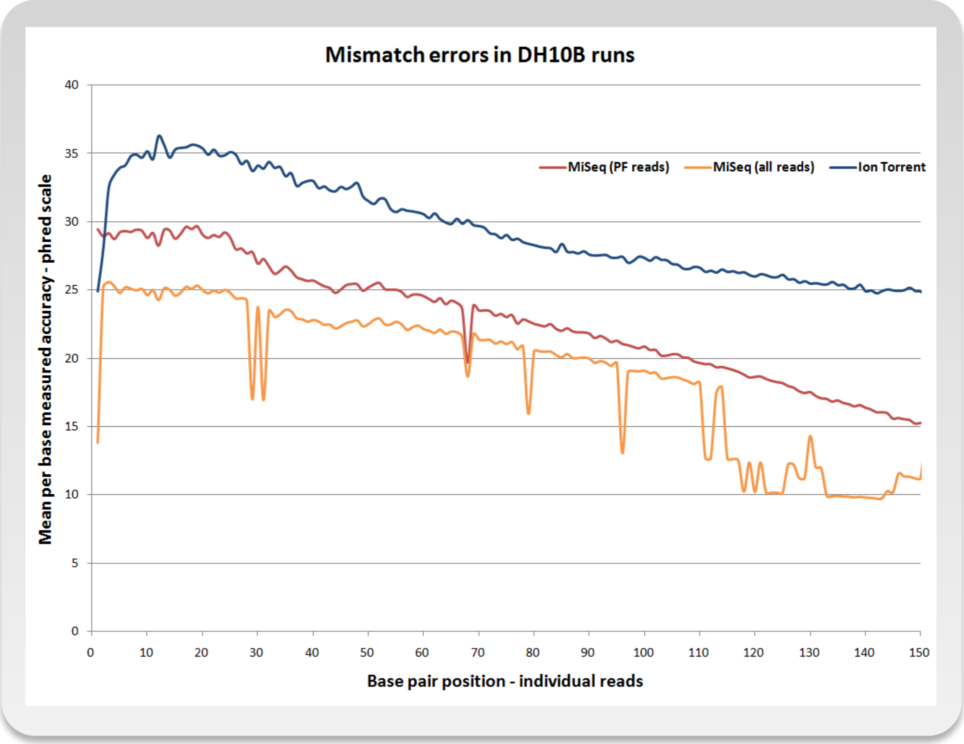


Figure 2. Total per-base mismatch accuracy rate for complete DH10B runs on both the Ion PGM™ sequencer and Illumina MiSeq™ sequencer.

REAL MiSeq Mismatch Accuracy is Higher

- Ion Torrent inaccurately represents mismatch accuracy by including MiSeq data that was properly flagged as not passing quality filter (non-PF reads)
- Recommended analysis approach:
  - Compare mismatch accuracy on similarly processed data sets
    - e.g., either raw vs. raw or filtered & trimmed vs. filtered & trimmed
- Result:
  - **MiSeq mismatch accuracy is higher when non-PF reads are properly excluded**



1. Reproduced IT's Figure 2 for IT data of clipped reads (blue) and MiSeq data of all reads (orange)
2. MiSeq PF reads (red) shows improved measured mismatch accuracy
3. 1.11% of unmapped reads were removed from Ion Torrent data

Claim #3: MiSeq Shows Difference in Predicted and Measured Quality Values

- Important but omitted information
  - Figure plots different types of data on same y-axis
    - Blue curve plots average predicted quality score
    - Orange curve plots average mismatch error rate converted to a phred scale
  - Orange curve is expected to be lower for any data set, including PGM data

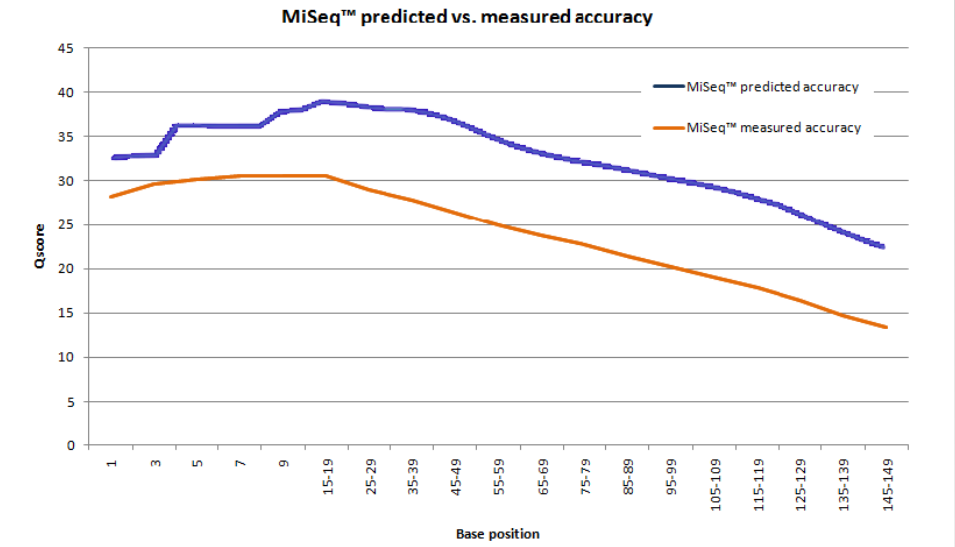
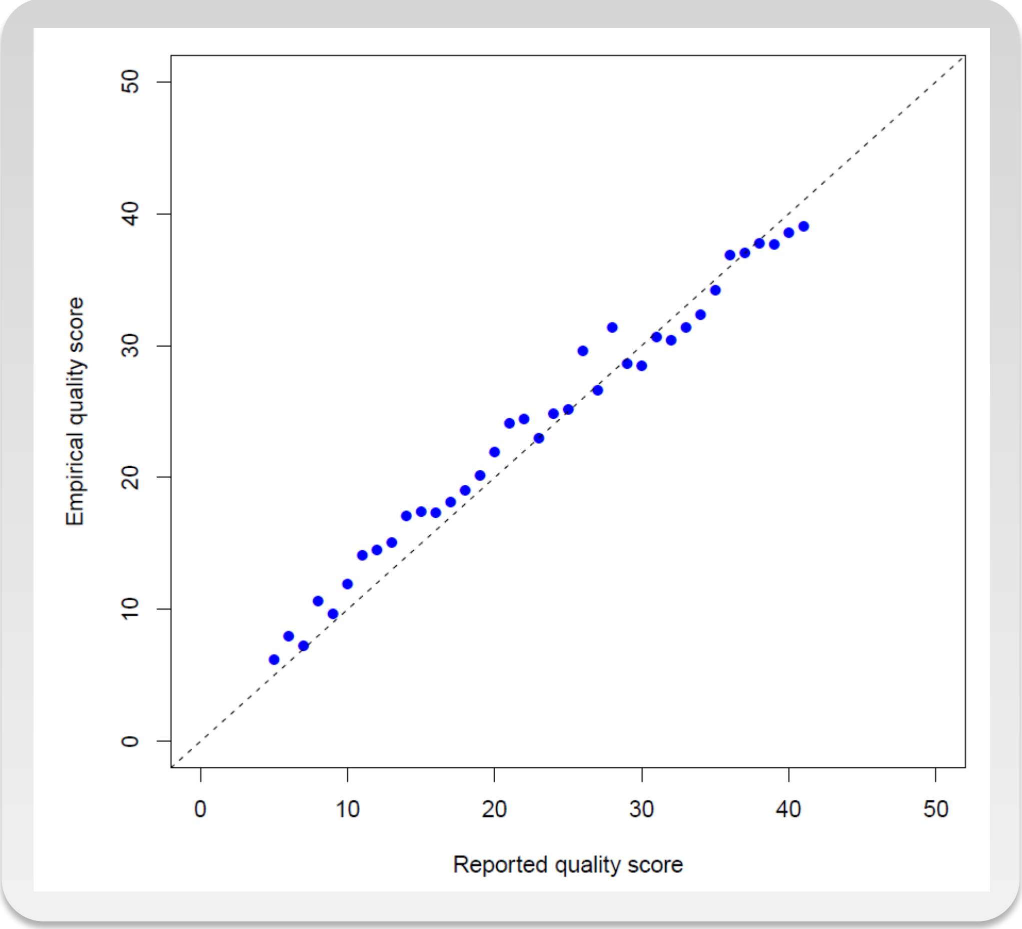


Figure 3. Illumina MiSeq™ platform predicted accuracy versus measured accuracy.

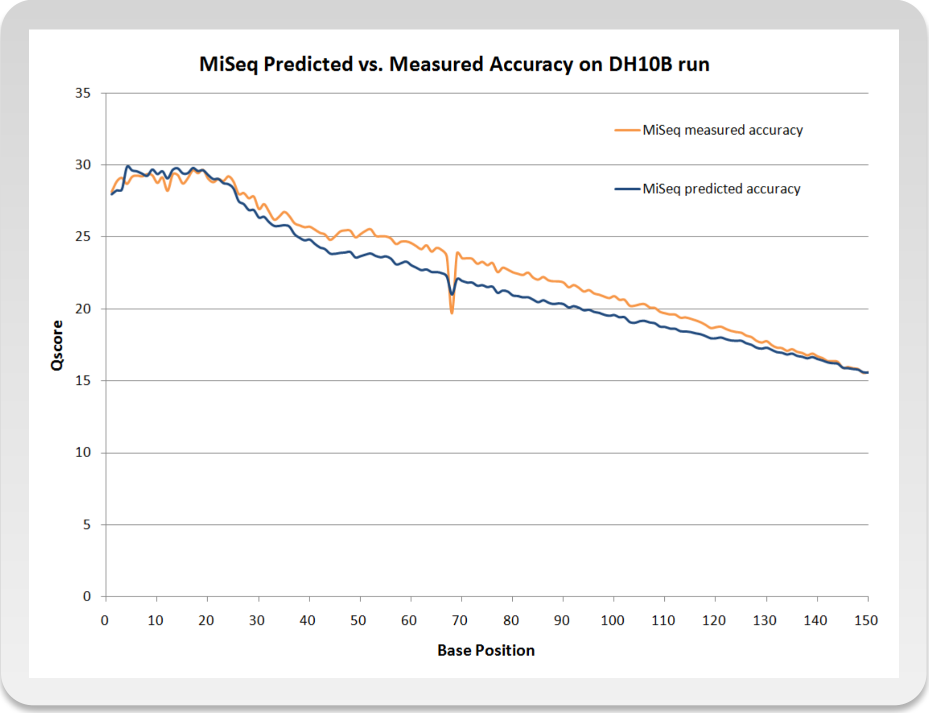
Match in MiSeq Predicted vs. Measured Quality Scores

- Q-Q plot from the Broad's GATK (Genome Analysis Toolkit) shows the MiSeq system's predicted quality score accurately reflects measured mismatch rate
- Ion Torrent would have created this plot during data analysis



Match in MiSeq Predicted vs. Measured Quality Scores

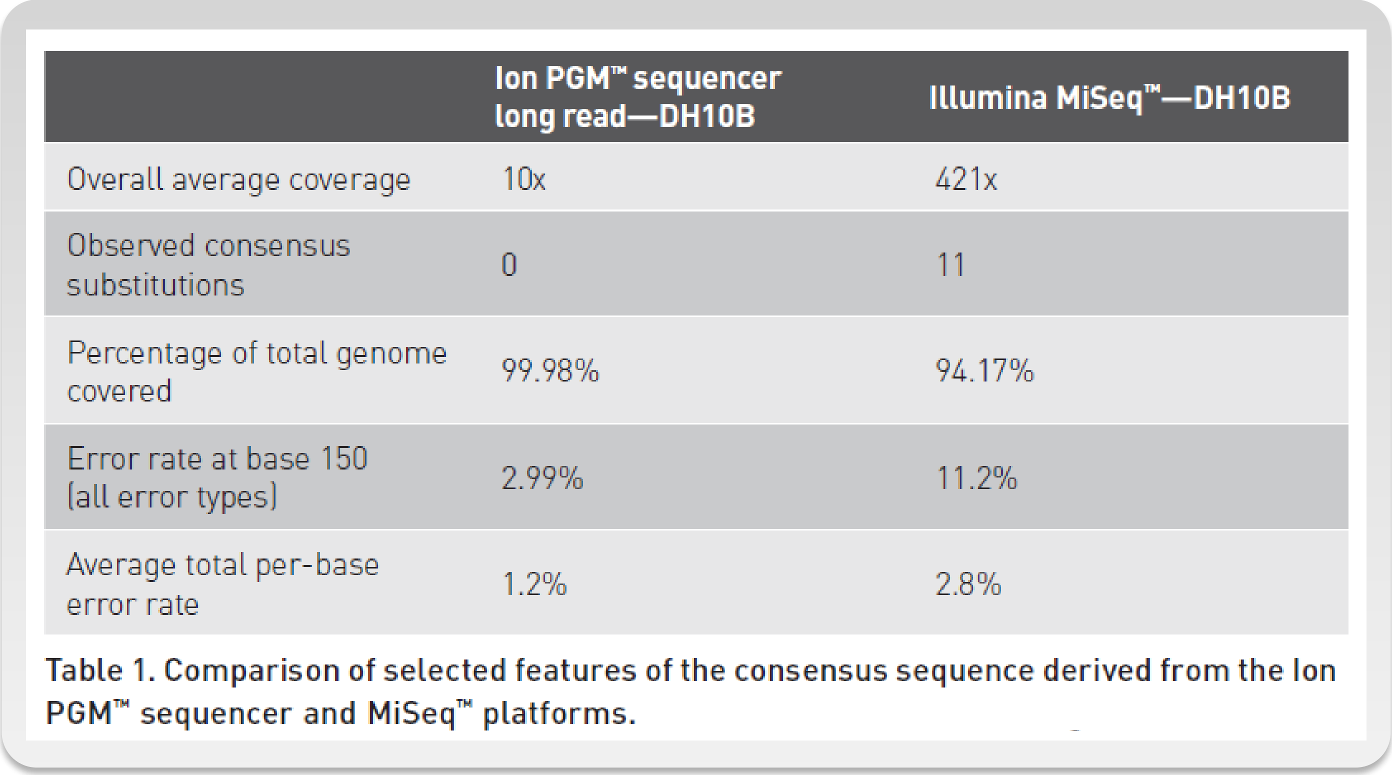
- Ion Torrent inaccurately represents MiSeq predicted vs. measured quality by incorrectly plotting two different types of data on the same y-axis
- Recommended analysis approach:
  - Compare predicted vs. empirical quality scores using standard tools or approaches
- Results:
  - **Broad's GATK shows MiSeq predicted quality score accurately reflects measured mismatch rate**



Re-plotted from per position empirical versus reported QV values found in the *.PositionCovariate.dat file from Broad GATK AnalyzeCovariates tool

Claim #4: MiSeq Underperforms in Consensus Quality

- Important but omitted information
  - MiSeq data was evaluated with protocols optimized for PGM
  - Using standard default tool settings results in fewer mismatches for MiSeq vs. PGM
  - Adding additional quality filter settings reduces MiSeq consensus mismatches to zero



	Ion PGM™ sequencer long read—DH10B	Illumina MiSeq™—DH10B
Overall average coverage	10x	421x
Observed consensus substitutions	0	11
Percentage of total genome covered	99.98%	94.17%
Error rate at base 150 (all error types)	2.99%	11.2%
Average total per-base error rate	1.2%	2.8%

Table 1. Comparison of selected features of the consensus sequence derived from the Ion PGM™ sequencer and MiSeq™ platforms.

MiSeq Data Evaluated with Protocols Optimized for PGM

- Ion Torrent inaccurately represents MiSeq performance in consensus quality by evaluating MiSeq data with protocols optimized for PGM data

- Recommended analysis approach:
  - Use standard tools and settings in calculating consensus error rates

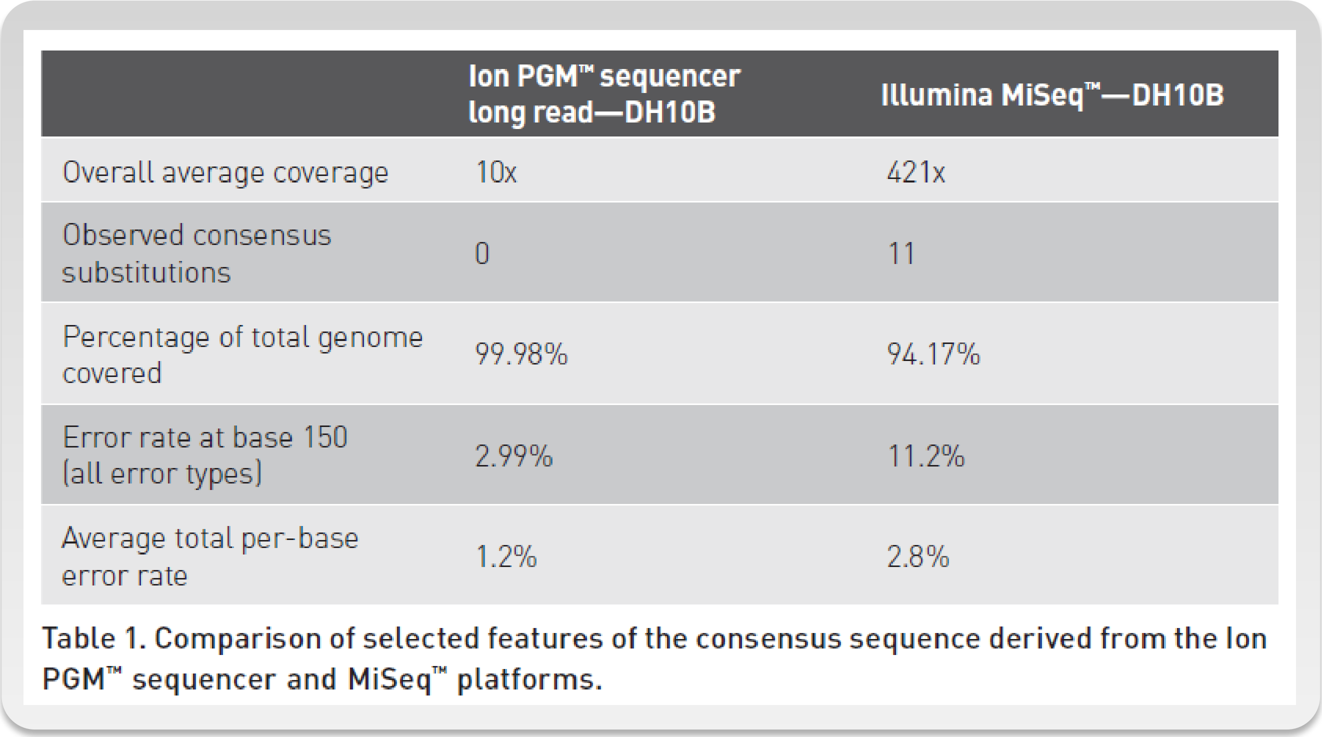
- Result
  - **MiSeq outperforms PGM in consensus indel error rate using standard tools and settings**

MiSeq Outperforms PGM with Standard Tools and Settings

Consensus Accuracy	Ion Torrent		MiSeq System	
	Claimed	Evaluated	Claimed	Evaluated
Substitutions				
Observed consensus substitutions - IT's mpileup parameters[1] - strandness[2]	0	0	11	10 of Q(SNP)>0 2 of Q(SNP)>=20[3]
Observed consensus substitutions - default mpileup parameters - strandness[2]		0		4 of Q(SNP)>0 0 of Q(SNP)>=20[3]
Observed consensus substitutions - default mpileup parameters		21 of Q(SNP)>0 10 of Q(SNP)>=20[3]		12 of Q(SNP)>0 5 of Q(SNP)>=20[3]
Indels				
Observed consensus indels - IT's mpileup parameters - strandness[2]	32	32		0
Observed consensus indels - default mpileup parameters[4] - strandness[2]		882 of Q(INDEL)>0 288 of Q(INDEL)>=20[3]		0
Observed consensus indels - default mpileup parameters		7774 of Q(INDEL)>0 2758 of Q(INDEL)>=20[3]		0

Claim #5: Higher Genome Coverage

- Important but omitted information
  - Illumina aligner discards non-uniquely mapped reads
  - Ion Torrent aligner randomly distributes non-uniquely mapped reads across multiple mapping sites while assigning mapping quality score of zero
  - DH10B genome has large segmental duplications that neither MiSeq nor Ion Torrent aligners can uniquely place reads



	Ion PGM™ sequencer long read—DH10B	Illumina MiSeq™—DH10B
Overall average coverage	10x	421x
Observed consensus substitutions	0	11
Percentage of total genome covered	99.98%	94.17%
Error rate at base 150 (all error types)	2.99%	11.2%
Average total per-base error rate	1.2%	2.8%

Table 1. Comparison of selected features of the consensus sequence derived from the Ion PGM™ sequencer and MiSeq™ platforms.

MiSeq Offers Higher Genomic Coverage than PGM

- Ion Torrent inaccurately represents genomic coverage by including non-uniquely mapping reads in final analysis

- Recommended analysis approach:
  - Genomic coverage comparisons should treat non-uniquely mapping reads the same for both platforms

- Result:
  - **MiSeq's genomic coverage is higher than PGM coverage, when using an unbiased analysis approach**

Coverage	Ion Torrent		MiSeq System	
	Claimed	Evaluated	Claimed	Evaluated
Overall average coverage	10x	13.6x	421x	421.7x
Percentage of total genome covered	99.98%	99.99% from all reads 93.75% from uniquely mapped reads	94.17%	94.17%

Summary

- Ion Torrent's Application Note is **not** an **accurate** representation of the **current** performance of either instrument
- Ion Torrent does not include indel errors in most of their accuracy comparisons
  - When indel errors are included, the MiSeq total error rate is substantially lower than the PGM total error rate
- Ion Torrent's data is extensively trimmed and they perform their comparisons against untrimmed MiSeq data that includes non-PF reads
  - This heavily distorts comparisons of mismatch rates between the platforms
- Ion Torrent claims that MiSeq has an order of magnitude difference between predicted and empirical quality scores
  - To support this claim they show plots of two similar, but ultimately very different metrics on the same graph, and ignore the Q-Q plot from Broad's GATK
- Ion Torrent's analysis approaches to the consensus mismatch rate and % genome coverage comparisons were biased
  - Standard comparison approaches shows that the MiSeq system performs as well or better than Ion Torrent for these metrics

We Stand by Our Quality and Performance Specifications

- Do the analysis for yourself:
  - Compare apples-to-apples
    - Ion Torrent data (last accessed on 8/26/11)
      http://lifetech-it.hosted.jivesoftware.com/docs/DOC-1848
    - MiSeq data
      http://www.illumina.com/downloads/Data/SequencingRuns/DH10B/MiSeq_Ecoli_DH10B_110721.bam

- Look at independent sources of information:
  - Nick Loman's blog at http://pathogenomics.bham.ac.uk/blog/author/nick/
  - Keith Robison's blog at http://omicsomics.blogspot.com/

- Our proof is in the publications:
  - The Illumina sequencing technology utilized in the MiSeq has enabled over 1,850 peer-reviewed publications