ion torrent
by life technologies™

Ion Torrent

Ion Personal Genome Machine™ Performance Overview

Summary

Rapidly increasing throughput
Read numbers and quality metrics have scaled

Uniform coverage
Simple and natural chemistry yields exceptionally uniform coverage

Excellent accuracy
Rapid improvements have yielded 99+% raw accuracy on single reads and over 99.97% consensus accuracy

Introduction

IIon Torrent released semiconductor sequencing and the Personal Genome Machine™ sequencer at the end of 2010. The PGM™ sequencer has brought scalability, simplicity, and speed to the high throughput sequencing field. The PGM™ sequencer can use multiple Ion sequencing chips and produce from 10 Mb to more than 1 Gb of sequence. The simplicity of the Ion Torrent sequencing chemistry results in minimal bias and unmatched uniformity of coverage. The PGM™ sequencer produces highly accurate long-read sequences in about 2 hours, the fastest Next Generation sequencing run time available. This application note summarizes the performance of the Ion Torrent technology. Understanding this performance over time shows us the incredible rate of innovation achieved with semiconductor sequencing.

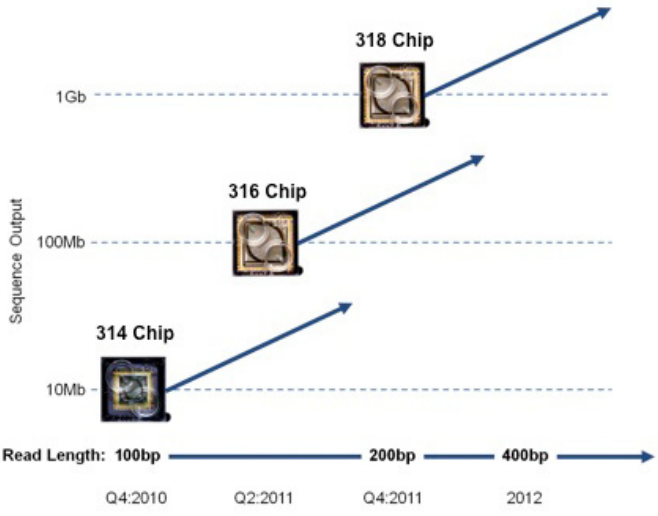

Figure 1. The PGM™ Sequencer provides the foundation for a series of scalable semiconductor sequencing chips.

Chip Type	314	316	318
Wells (in millions)	1.2	6.2	11.1
Bases (in Mb)	>10	>100	>1000

Table 1. The PGM™ Sequencer provides the foundation for a series of scalable semiconductor sequencing chips.

Rapidly Increasing Throughput

The detection of a specific nucleotide on a growing DNA strand occurs inside a fabricated well of an Ion Torrent semiconductor chip. The Ion sequencing chip captures voltage measurements from the direct release of hydrogen ions following DNA polymerization. The total number of independent measurements, or sequence reads, is a function of the number of sensors and fabricated wells that a chip contains. The semiconductor industry has increased the number of sensors within the confines of a microprocessor. This evolution is leveraged by the Ion Torrent sequencing technology to offer different sequencing chip densities, hence different sequence throughput (Figure 1 and Table 1). The parallels in scalability of semiconductor sequencing and personal computer are worth noting. Because the Ion Torrent chips are made in the same fabrication facilities as the semiconductors that run our computers or mobile phones, semiconductor sequencing has the potential to follow a similar innovation curve. Together, the first three chips released by Ion Torrent will deliver a 100X improvement in base yield (10 Mb to 1 Gb) in a single year. This lays a path for semiconductor sequencing to deliver scalable, simple, and fast DNA sequencing to the research and clinical research com-

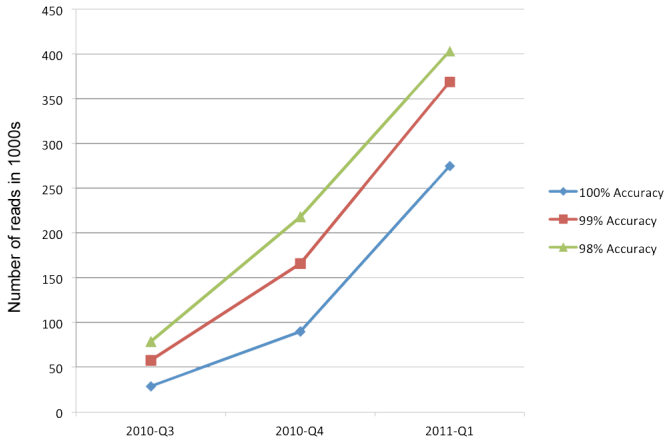

Figure 2. Increasing yield of number of reads across the last 3 quarters. X axis is the number of 100 base reads in thousands.

munity. Literally, the Chip is the Machine™.

The Ion 314 sequencing chip produces 10 Mb of data at greater than 99% raw accuracy (table 1). The Ion 316 sequencing chip is being released during the second quarter of 2011. The 316 chip contains five times the number of wells but produces more than 100 Mb of sequencing data. The Ion 318 sequencing chip will be available in the second half of 2011 and will produce over 1 Gb of data. All three chips are fully compatible with the PGM™ sequencer.

The improvements in sample preparation and primary data processing have also had an impact on the resulting sequencing data. The throughput from a single run of an Ion 314 chip shows dramatic improvements over the last 9 months. Figure 2 shows the number of reads from single sequencing runs of *E. coli* on 314 Ion sequencing chips. The number of 100 base reads with 0 or 1 errors (99% accuracy) has increased from 58K to 166K to 369K over the last three quarters of development effort. It is also important to notice that the number of perfect reads of 100 bases or more is greater than 275K.

Read length will get to 200 bases in 2011 reaching 400 bases in 2012 with the consequent improvement in throughput.

The combination of increased chip density, read length and improvements in molecular biology protocols and data processing has resulted in an impressive rate of gains in sequencing throughput. Development programs focused on these areas are in place and will bring further throughput improvements in the near future.

Uniform coverage

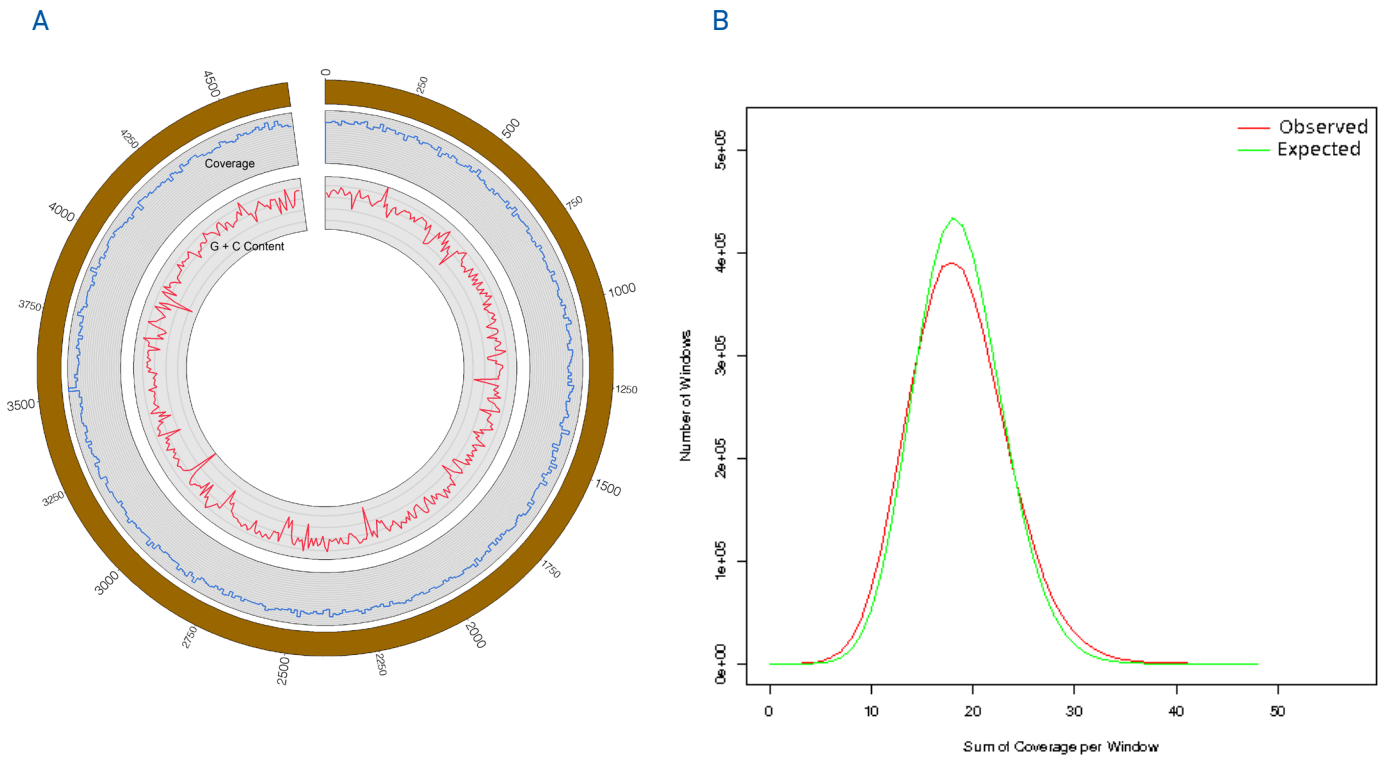

Figure 3. A& B – Uniformity of Sequencing Coverage
Data generated from a single run of 314 using E. coli DH10B. Base calls generated with v1.3.0 of the Torrent Suite software. Any filtering that occurs is completely independent of alignment and based solely on the inherent signal properties of the data.
A - Left panel –Circular plot shows percentage of G+C content in the inner red circle and regional coverage level in the outer blue circle.
B - Right panel –Histogram of expected versus actual coverage. X axis is the sum of the number of reads (coverage) per genomic window. The Y axis is the number of genomic windows that have the level of coverage. The green line (Possion) shows what coverage would be if sequencing occurs in a truly unbiased manner. The red line represents the actual sampling of the Ion Torrent reads. The high degree of overlap between the red and green lines illustrates the unbiased nature of semiconductor sequencing.

The Ion Torrent technology relies on simple biochemistry – just DNA polymerase and natural nucleotides. This chemistry is the product of more than a billion years of evolution. The detection process is equally simple –hydrogen ions released when known nucleotides are incorporated into the DNA molecules are detected by the sensors on the semiconductor surface. Unlike other technologies, Ion Torrent does not use modified nucleotides, enzyme cascades or light-based detection methods that lead to a significant bias in coverage.

The simplicity of both synthesis and detection translates into exceptionally uniform sequencing coverage. Figure 3A shows a plot of depth of coverage and G+C content from a single 314 run of *E. coli* DH10B. The coverage (shown by the blue line in the outer track) is very even across the bacterial genome regardless of the changes in G+C content (red line in the inner track).

If a genome were divided into 100 base windows and those windows were randomly sampled for coverage, the sampling would show a perfect Poisson distribution (Figure 3B). To measure if there was any sequence bias in the semiconductor sequencing reads, the reads in a single 314 run were each aligned to 100 base windows of the E. coli genome. The resulting distribution of the reads maps very closely to the random Poisson distribution demonstrating that there is very little bias in coverage generated with the PGM™ sequencer.

Broad, even coverage provides access to regions of the genome that other technologies cannot sequence and reduces the amount of sequencing needed to have confidence in the data.

Excellent accuracy

The overall quality of semiconductor sequencing data can be examined by monitoring the raw accuracy rate across the length of the sequence reads. Figure 4 shows the raw accuracy rate across the cumulative length of the collection of 100 base reads from a single Ion 314 chip sequencing run of *E.coli* DH10B. After additional filtering and trimming, over 44 Mb of data were considered for this analysis, well above the 10 Mb of data that corresponds to the specification for the Ion 314 sequencing chip. The raw accuracy at the start of

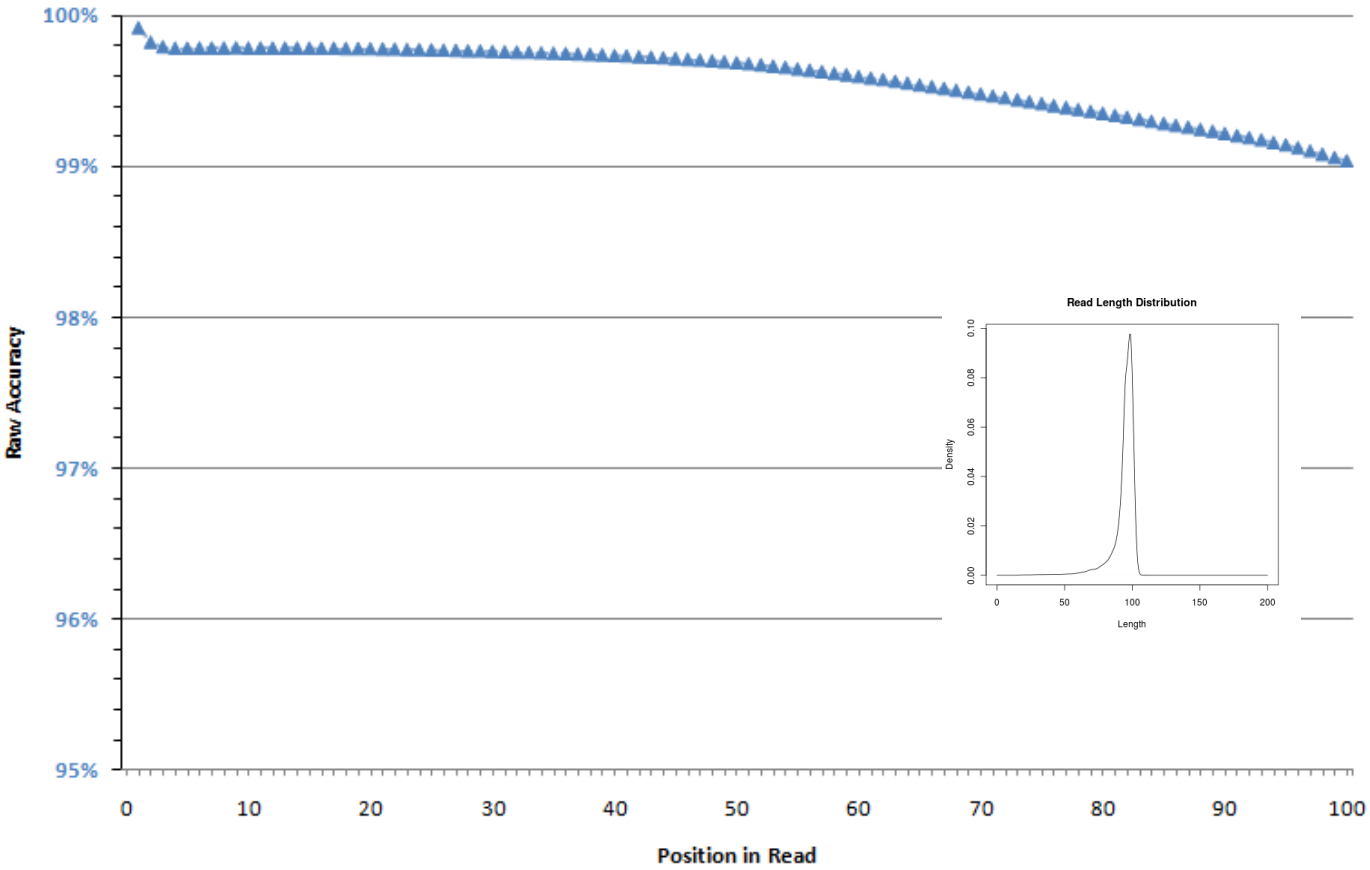

Figure 4. Accuracy versus total position from a single run of 314 using E. coli DH10B. Y axis represents raw accuracy. X axis is the base position within the read. Base calls were generated with default parameters in v1.3.0 of the Torrent Suite software, changing CR filter set to 0.06 and trimming an additional 10 bases from the end of each read. After filtering based solely on the inherent signal properties of the data 475K reads were produced. Additional filtering (but not trimming) is performed to identify reads that map across 50 bases minimum at Q10 or better. This removed only 11K of the reads (464K retained out of 475K) meaning that almost 98% of the reads were mappable. Figure 4 Insert – Distribution of read length for the data. X axis shows length of individual reads in base pairs. Y axis is the number of reads for a given length (most reads approach 100 bases read-length).

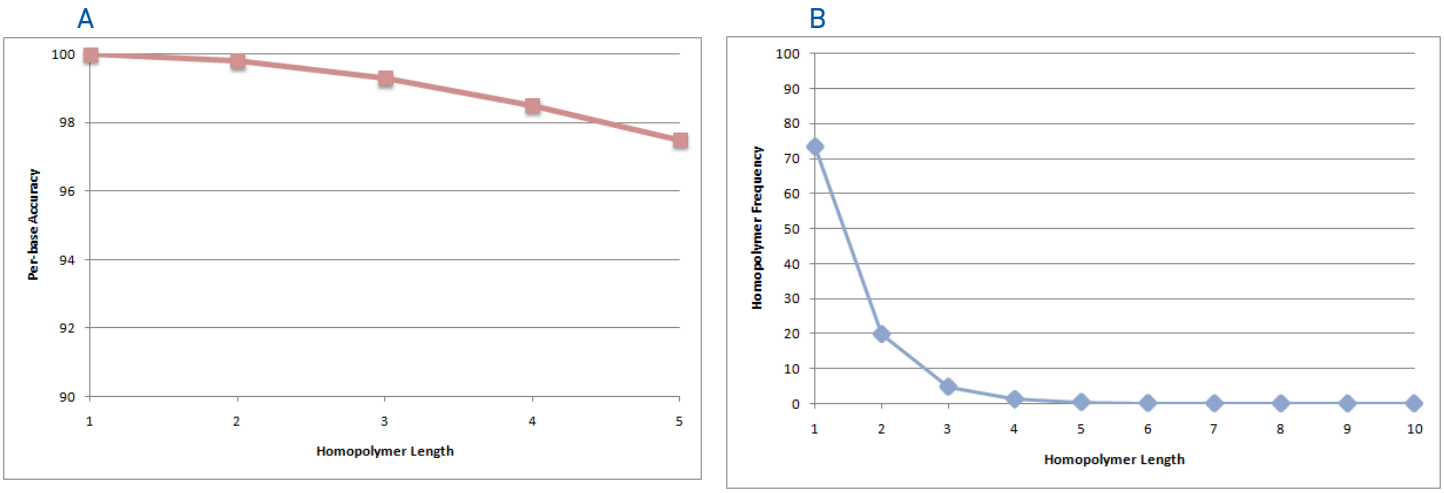

Figure 5. A & B - Homopolymer Accuracy
Panel A - RED LINE- Data generated from a single run of an Ion 314 chip using E. coli DH10B.
The same primary data used in Figure 4 was analyzed to measure raw per base accuracy in homopolymer stretches with Torrent Suite software v1.3.0. Raw accuracy rates are listed on the Y-axis and the length of the homopolymer stretch is represented on the X-axis.
Panel B - BLUE LINE – Incident rate of homopolymers of various lengths in E. coli DH10B.

Question: How does Ion Torrent measure accuracy?

Answer: Accuracy metrics are calculated either through prediction algorithms or through actual alignment to a known reference genome. Predicted quality scores are derived from algorithms that look at the inherent properties of the input signal and make fairly accurate estimates regarding if that one base will align. Predicted quality is useful to filter and remove lower quality reads prior to downstream alignment. However predicted Q scores are only as good as their prediction algorithm. Ion Torrent also calculates accuracy based on proper alignment using a reference genomic sequence. This measurement is what we refer on this document as "raw accuracy. This is single pass accuracy implying that we are measuring the true per base error associated with a single read, not consensus accuracy, which measures the error rate from the consensus sequence which is the result of multiple reads. Within the Analysis Report, raw accuracy is labeled with "AQ" for aligned quality. The Phred-like Q score measures accuracy on logarithmic scale that: Q10 = 90%, Q20= 99%, Q30 = 99.9%, Q40 = 99.99%, and Q50 = 99.999%

the read is 99.9% (Q30). As the read length progresses, the raw accuracy is kept at 99.7% near the middle of the 100 bases read and is nicely maintained over 99% (Q20) at the end of the 100 bases read. The same single sequencing run yielded more than 99.97% of consensus accuracy and covered more than 99.99% of the *E.coli* genome. The insert within Figure 4 shows the length distribution of the reads demonstrating the predominance of full length reads.

Stretches of the same nucleotide sequence, also known as homopolymer stretches, are also detected at a very high accuracy. A 5-mer is currently called with greater 97.5% per base accuracy (Figure 5A). The rarity of homopolymers of 6 bases of longer makes statistical analysis difficult (Figure 5B). An example of a single read containing a homopolymer 8 bases long shows a signal intensity very close to an integer value of 8 and as a consequence, the sequence was called correctly.

Improvements in software algorithms, molecular biology protocols, and manufacturing processes will continue driving improvements in the accuracy of semiconductor sequencing.

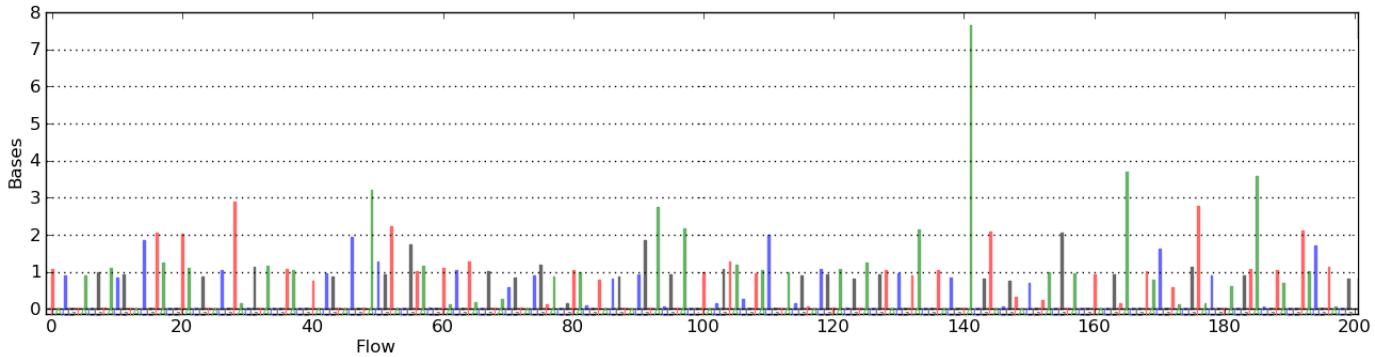

Figure 6: Ionogram showing a single perfect read containing a homopolymer of length 8 near the 140 flow within the semiconductor sequencing run. The X axis shows the flow number. The Y axis displays the intensity of the processed voltage signal against the integer values of increasing base pairs.

Summary

Semiconductor sequencing using the Ion semiconductor chips and the Ion PGM™ Sequencer has excited the scientific community since its announcement and its commercial launch. Several key factors stand out:

Increasing Throughput - The commercialization of three semiconductor sequencing chips, the Ion 314, 316, and 318 demonstrate a 100X scalability path in 2011 to move from more than 10 Mb of output to over 1 Gb. Going from read-length of 100 to 400 bases in 2012 will also have a major impact on the PGM™ sequencer throughput.

Uniform Coverage – The simple and natural biochemistry of DNA synthesis is leveraged in semiconductor sequencing. This simplicity provides minimum bias in coverage resulting in highly uniform sequencing representation. Keeping the technology as close to native molecular processes allows a billion years of evolution to determine the best chemistry to sequence any region of any genome with unprecedented quality.

Excellent Accuracy – Rapid improvements in molecular biology protocols and software innovations have resulted in excellent sequencing accuracy. Currently, semiconductor sequencing produces raw accuracy rate of over 99%. Semiconductor sequencing is expected to continue this rapid pace of innovation and improvements into the near term future.

Data supporting the figures presented in this document can be obtained within the Ion Community. Please register at http://ioncommunity.iontorrent.com.

Ion Torrent by Life Technologies | 7000 Shoreline Court | Suite 201 |South San Francisco, CA 94080 USA

Phone +1-203-458-8552 | Toll Free in North America 1-87-SEQUENCE (1-877-378-3623)

www.lifetechnologies.com | www.iontorrent.com | http://ioncommunity.com CO31484

ion torrent

by life technologies™