ion torrent
by life technologies™

Ion semiconductor sequencing uniquely enables both accurate long reads and paired-end sequencing

Summary

- Ion Torrent recently launched a long read kit for the Ion PGM™ sequencer with modal high-quality read lengths of 225 bases.
- Read lengths greater than 500 bases are feasible, as demonstrated by the generation of a perfect 525 bases read.
- A novel paired-end sequencing (PES) protocol for the Ion PGM™ system has been demonstrated. Ion PES further enhances accuracy. The number of indel errors was reduced 5-fold, and the total number of consensus errors was reduced 10-fold.

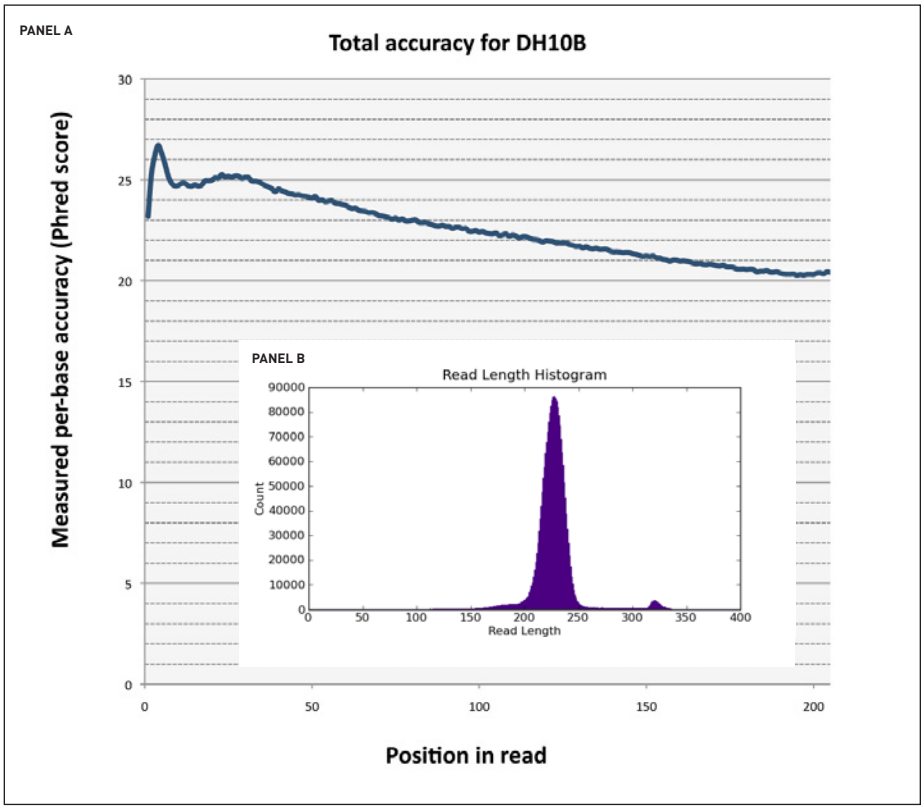

Figure 1. Ion semiconductor sequencing utilizes natural nucleotides and simple sequencing chemistry, producing accurate reads that are at least twice as long as reads generated with fluorescent sequencing-by-synthesis (SBS) chemistry. Sequencing accuracy generally decreases as read length increases, but this slope is relatively flat for the Ion PGM™ sequencer when compared with fluorescent SBS chemistry. The Ion PGM™ sequencer demonstrates modal read lengths of 225 bases (panel B). Measured per-base accuracy (Phred score) is calculated after aligning reads to the DH10B reference genome and is then plotted as a function of position in read, (panel A). A Phred score of 20 is equivalent to 99% accuracy, or 1 error per 100 bases.

Next-generation sequencing — based applications and methodologies are extensively used to investigate genome biology. The lengths of next-generation sequencing reads are limited by accuracy—as read length increases, accuracy decreases. This fundamental inverse relationship between read length and accuracy governs sequencing-based applications.

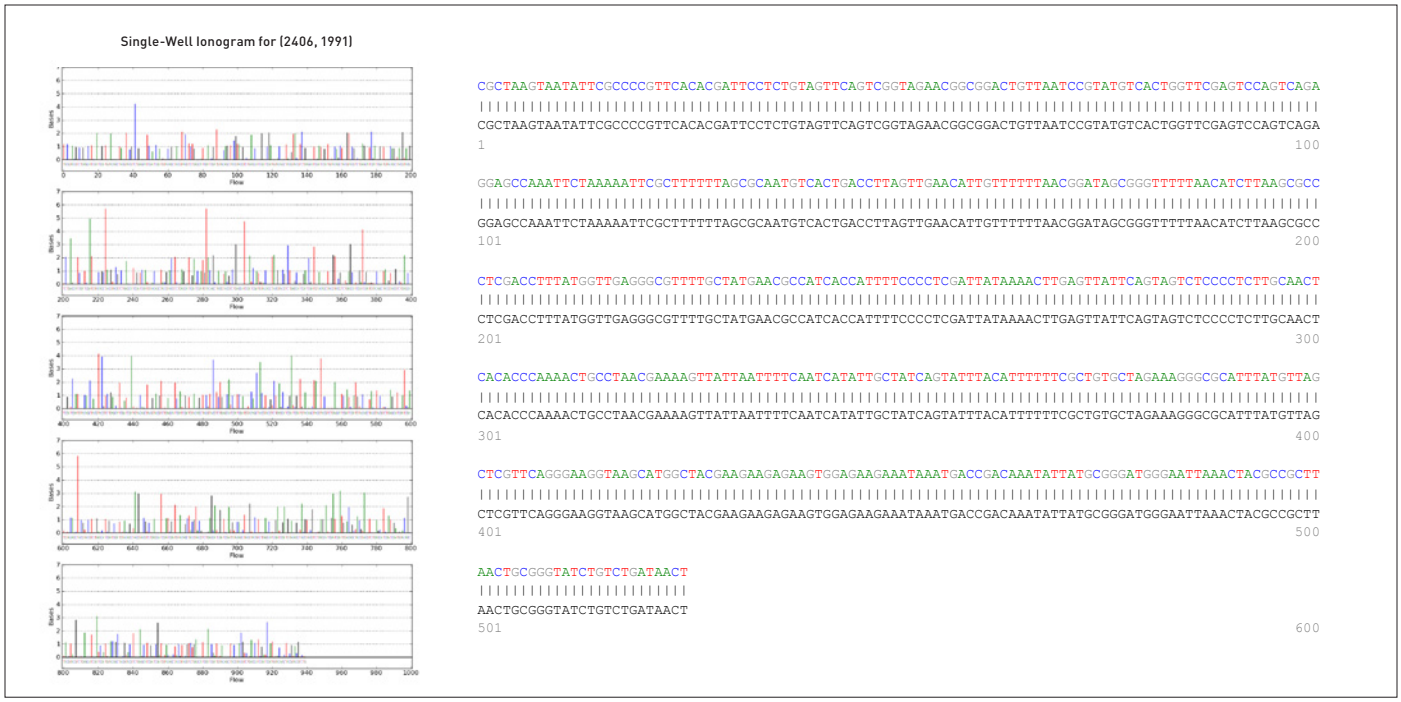

Figure 2. Demonstration of a 525 base perfect read on the Ion PGM™ Sequencer. The "ionogram" (left) displays the number of each base (A, T, C, or G) called at each nucleotide flow during an Ion PGM™ sequencing run. The sequence alignment (right) displays the read generated by the Ion PGM™ sequencer in the top strand and the reference genome sequence in the bottom strand.

In October 2011, Ion Torrent launched a long read kit for the PGM™ sequencer, which can provide modal high-quality read lengths of 225 bases (Figure 1). Mean raw accuracy across the length of the read is 99%, (Q20) and consensus accuracy is 99.999%, (Q50). **This data is available for download at ioncommunity.iontorrent.com — run name STO-409.**

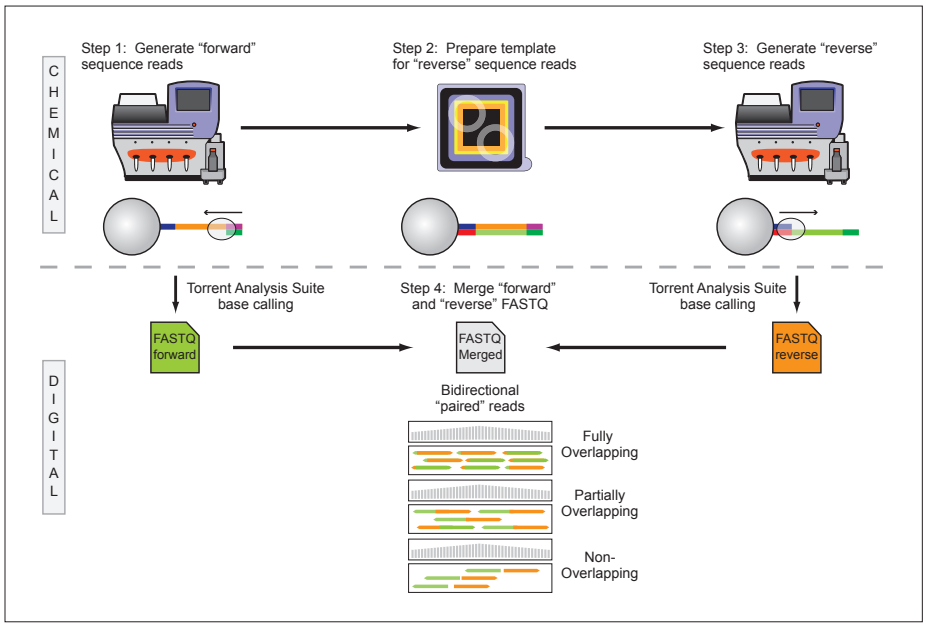

Figure 3. Only Ion Torrent Personal Genome Machine™ (PGM™) semiconductor sequencing technology delivers the unique combination of long reads and PES. This figure depicts the chemical and digital workflow of the novel PES method developed by Ion Torrent. Forward and reverse reads are denoted by green and orange colors, respectively. The degree of paired-read overlap is shown aligned to a reference genome.

Reads longer than 500 bases are achievable, as demonstrated by the generation of a perfect read spanning 525 contiguous bases (Figure 2).

Paired-end sequencing (PES) was initially developed for short-read next-generation sequencing technologies to increase both coverage of the human genome and the fraction of generated reads aligning to the genome. Fortuitously, PES has become useful for applications that demand ever longer reads. Increasing the length of accurately called bases can obviate the need for PES in many instances, allowing more of the genome to be accurately sequenced, and for some especially demanding applications there is no substitute for long reads. Given the utility of both accurate long reads and PES, Ion Torrent has recently developed solutions for both methods.

Using the Ion sequencing kit launched in July 2011, which supports read lengths of 100 bases, a novel PES method was developed for the Ion PGM™ sequencer that produces paired reads, each 100 bases in length (2 x 100). We anticipate that applying recent improvements in read length to the Ion PES protocol should produce paired reads of 200 bases in length (2 x 200), with the possibility of paired reads up to 400 bases in length (2 x 400).

Until recently, no sequencing technology could simultaneously offer the benefits conferred by both long reads and PES. The Ion Torrent Personal Genome Machine™ (PGM) semiconductor sequencing technology is now the only platform that provides this unique combination of long reads and PES.

Ion PES workflow

The simplicity of semiconductor sequencing allows a template to be sequenced in both directions using a single Ion Chip. The first read results from the polymerase extending the primer towards the Ion Sphere™ particle, in the "forward" direction (Figure 3, step 1), the standard operating procedure for the Ion PGM™ Sequencer. Then, off instrument, directly within the Ion Chip, the template is prepared for the second read (Figure 3, step 2) via a series of simple enzymatic steps. The forward primer is fully extended to the Ion Sphere™ particle, and then the original template is nicked and degraded to produce a primer for the second read. The same Ion Chip is reintroduced to the Ion PGM™ Sequencer, and the second read is generated by the polymerase extending the primer away from the Ion Sphere™ particle, in the "reverse" direction (Figure 3, step 3).

Each microwell on the Ion Chip, and thus, each template is in perfect register with each transistor, ensuring that the linkage between forward and reverse reads is unbreakable even when the Ion Chip is removed from the Ion PGM™ sequencer between the forward and reverse sequencing runs. Furthermore, because the Ion Chip directly translates biochemical signals into digital signals, via perfectly arrayed transistors, only a particle-immobilized template, primer and polymerase are required in each Ion Chip microwell. This simplicity facilitates the occurrence of nucleic acid biochemistry and enzymology directly in each microwell, enabling PES as well as other novel sequencing methods.

Ion PES data analysis

Forward and reverse sequencing produces two FASTQ-formatted files. Merging the two FASTQ files produces both unidirectional (singleton) and bidirectional reads. Unidirectional reads are defined as Ion Chip microwell positions that do not intersect in the forward and reverse FASTQ files, whereas bidirectional reads are defined as the intersection of the Ion Chip microwell position in the forward and reverse FASTQ files. After alignment to a reference genome, the bidirectional or "paired" reads can be further subclassified into "improper pairs", those with improper orientation or improper distance between forward and reverse reads, and "proper pairs", those with proper orientation and distance between forward and reverse reads (Figure 3).

PES generates differing degrees of overlap between the forward and reverse reads, depending on the size of the library fragments and read lengths. With a library insert size of 100–200 bases in length, the Ion PES methodology can produce fully overlapping forward and reverse reads with either a 2 x 100 or 2 x 200 Ion semiconductor sequencing run. Fully overlapping reads are useful for further enhancing accuracy. In counting-based applications such as RNA-Seq or ChIP-Seq, counting accuracy is improved by partially overlapping reads (Figure 3), which increase the number of uniquely mapped reads and identify redundant reads by demarcating the start and end of the sequence reads. Non-overlapping reads (Figure 3) are useful for detecting genomic rearrangements or *de novo* assembly because the distance between the forward and reverse reads can be longer than the length of a single read and the distance is known and uniformly distributed.

To compare the accuracy of fully overlapping paired reads to individual reads, the proper pairs were either aligned as individual reads or as a single consolidated read to the DH10B reference genome using the TMAP alignment software in the Torrent Analysis Suite. A total of 977,458 reads were paired and

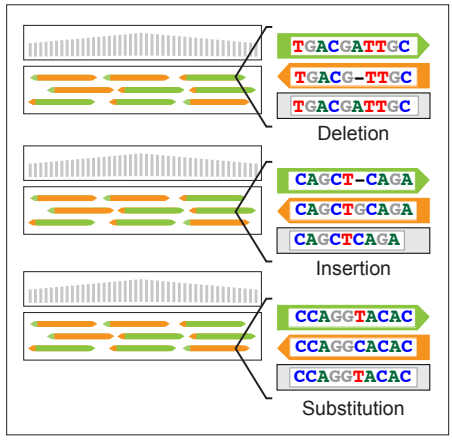

Figure 4. Depiction of a paired-read error-correction schema. Green and orange bars represent forward and reverse reads derived from fully overlapping paired reads, aligned to a reference genome. Information from paired strands at the same nucleotide as well as per-position quality scores can be used to develop improved bioinformatics methods for error correction

aligned to the DH10B reference genome using the Ion 314™ Chip, and of these 466,654 reads were considered proper pairs, to produce 53 Mb of consolidated data, giving greater than 10x coverage on *E. coli*. To consolidate forward and reverse reads, errors were handled via a simple schema. Specifically, if one of the reads from the pair did not exhibit the same substitution, insertion, or deletion at the same position, then the substitution, insertion, or deletion was removed (Figure 4). Via this methodology, raw insertion and deletion errors were reduced ~5-fold to 0.19%. Consensus accuracy was reduced over 10x, to only 17 total errors with pairing.

Conclusion

The combination of long reads and paired-end sequencing delivers a unique and powerful capability that is only available by employing Ion semiconductor sequencing. Perfect reads exceeding 500 bases have now been demonstrated using the Ion PGM™ system. When these increasingly longer reads are coupled with paired-end sequencing the outcome is enhanced accuracy.

The difference between "paired-end sequencing" (PES) and "mate-pair sequencing" (MPS)

PES and MPS are similar in that both methods produce reads from the ends of the templates, yet differ in that MPS produces a single read of the template ends that were paired during library construction—whereas PES produces two sequence reads, one from each end of the same template. MPS is generally used for longer inserts (2–10 kp), and PES is employed for shorter inserts (100–800 bases). Ion Torrent recently released a long MPS protocol on the Ion Community (ioncommunity.iontorrent.com) for sequencing the ends of 10 kb inserts, which have important applications in cancer rearrangements and *de novo* genome assembly.

The non-overlapping reads produced by MPS and PES enable applications such as interrogation of genomic rearrangement or *de novo* genome assembly because the distance between the forward and reverse reads can be longer than the length of a single read and the distance is known and uniformly distributed. Beyond assembly and interrogation of genome structural variation, PES can improve the number of uniquely mapped reads, the ability to identify unique (different start and end points) and redundant reads (same start and end point), and, as we demonstrate here, accuracy.

Bioinformatics methods

Raw voltage data were processed using Torrent Suite v1.5 to generate FASTQ files. FASTQ files were aligned against *E. coli* DH10B using TMAP v0.1.3-1 to produce BAM files. Variants for STO-409 were called using SAMTOOLS mpileup using default parameters as wrapped with Torrent Suite v1.5 in the Germ Line Variant Caller Plug-In. Variants for PES analysis were called from BAM files using both SAMTOOLS mpileup (-Q 7 -h 50 -o 10 -e 17 -m4) and GATK v1.0.5777 with the following commands.

```
##- Making indel calls
/usr/java/default/bin/java -jar /home/sunya/tool/GATK/GenomeAnalysisTK.
jar \
-T IndelGenotyperV2 \
-R /data/output/pgm/pe/reference/dh10b.fa \
-I /data/output/pgm/pe/consolid/output.bam \
-bed /data/output/pgm/pe/consolid/norecal/consolid_indel.bed \
-o /data/output/pgm/pe/consolid/norecal/consolid_indel.vcf \
--minFraction 0.5 \
--minIndelCount 5 \
--window_size 300

##- Making SNP calls
/usr/java/default/bin/java -jar /home/sunya/tool/GATK/GenomeAnalysisTK.
jar \
 -T UnifiedGenotyper -nt 6 \
 -R /data/output/pgm/pe/reference/dh10b.fa \
 -I /data/output/pgm/pe/consolid/output.bam \
 -o /data/output/pgm/pe/consolid/norecal/consolid_snp.vcf \
 -stand_call_conf 30.0 -stand_emit_conf 10.0
```

For Research Use Only. Not intended for any animal or human therapeutic or diagnostic use.

© 2011 Life Technologies Corporation. All rights reserved. The trademarks mentioned herein are the property of Life Technologies Corporation or their respective owners.
Printed in the USA. **CO24084 1011**

lifetechnologies.com