ion torrent
by life technologies™

de novo assembly using Ion semiconductor sequencing

Key findings:

- Ion semiconductor sequencing generates high-quality *de novo* assemblies using mate pair data
- A single 200 base run on the Ion PGM™ Sequencer produced high quality sequence with 50% of the bases in contigs of > 119 kb, (i.e. N50 > 119 kb)
- Combining fragment and long mate-pair data results in a higher quality assembly with longer contigs and fewer scaffolds

Ion semiconductor sequencing offers a simple, fast workflow enabling rapid generation of large volumes of high quality data. Fast whole-genome sequencing of epidemiologically relevant organisms is possible[1, 2], but the final assembly inevitably contains gaps in the sequence. In this application note we show how fragment library data can be augmented with mate-pair library data to produce a much higher-quality assembled sequence (larger contigs, fewer scaffolds, and better N50 lengths).

Improved sequence finishing using mate-pair library data

As sequencing technologies evolve, researchers are able to take advantage of increased throughput, higher accuracy, and longer reads. Current genome sequencing efforts divide into two camps—Sanger shotgun sequencing and massively parallel high-throughput sequencing (or next-generation sequencing (NGS)). Sanger sequencing yields very accurate reads up to 1,000 bases, but this methodology is more expensive and time-consuming. As of early 2012, Ion semiconductor sequencing produced read lengths of up to 200 bases, shorter than those of Sanger sequencing. Many researchers see this as a worthwhile tradeoff because NGS produces a very large number of reads in a very short time period, making data gathering with this strategy less expensive and faster. Even if 1,000-base reads were routinely achievable, accurately arranging a multitude of 1,000-base fragments into a genome is still fraught with difficulties—expansive sections of repeated sequence and variations in sequence quality in some regions may hinder accurate assembly. So at the end of all fragment library sequencing projects, the assembled genome comprises a set of overlapping reads (contigs) interspersed with gaps where the sequence is not reliably identified.

For some researchers, fragment library data alone are sufficient. For example, when establishing contigs for many comparative genomics approaches, the extra time, effort, and expense required to fill sequence gaps are not warranted. When the sequencing strategy calls for complete or nearly-complete sequence, many scientists opt for additional data from a different kind of library—a mate-pair (MP) library. MP libraries are created by using directed molecular biology steps to capture extreme ends of much longer DNA fragments—from hundreds of bases to 50 kb—and packages them in short fragments that are suitable for NGS (Figure 1). This means that instead of getting positional information only for the stretch of sequence in each fragment library clone, positional information is now obtained over much larger distances (Figure 1). When a fragment assembly is augmented with the data obtained from a mate-pair library, contigs can be ordered into scaffolds and many of the sequence gaps closed.

life technologies™

This allows smaller contigs to be converted into larger supercontigs and scaffolds (Figure 2) and results in longer N50 values. Adding one MP library to fragment library data generates a good draft assembly (small number of large scaffolds) with fairly accurate annotation. Adding a second MP library generates a sequence that approaches "finished" status.

Results

In order to demonstrate the benefits of mate pairs, a model organism (*E. coli* MG1655) was analyzed using Ion semiconductor sequencing, (i.e., on the Ion PGM™ system) and assembled using fragment data and mate pairs. Each sequencing run was completed in ~2.4 hours at a very affordable price.

Mate-pair library construction

In this study, genomic DNA was fragmented and size-selected on an agarose gel before ligation with mate-pair adaptors. These DNA fragments were then circularized by hybridization so that mate-pair adaptors formed an internal adaptor to connect both ends of DNA together. The DNA fragments containing mate-pair ends in the circularized DNA were released by a nick-translation reaction following with exonuclease treatment. The mate-pair DNA fragments were enriched and ligated with fragment library adaptors to form the mate-pair library. The detailed mate-pair library construction procedure is described in a user bulletin, Ion Mate-Paired Library Preparation (http://lifetech-it.hosted.jivesoftware.com/docs/DOC-1999). For this assembly, two separate libraries—with inserts of 3.5 kb and 8.9 kb—were constructed using this protocol.

Pairing analysis after acquiring sequencing data

After identifying the internal adaptor sequence and then splitting the sequencing read into two tags, the individual tags were mapped to the reference genome and the distance between the reads determined. This distance is the insert length of the library fragment from which these tags are derived. Plotting insert lengths of a 3.5 kb and a 8.9 kb mate-pair library

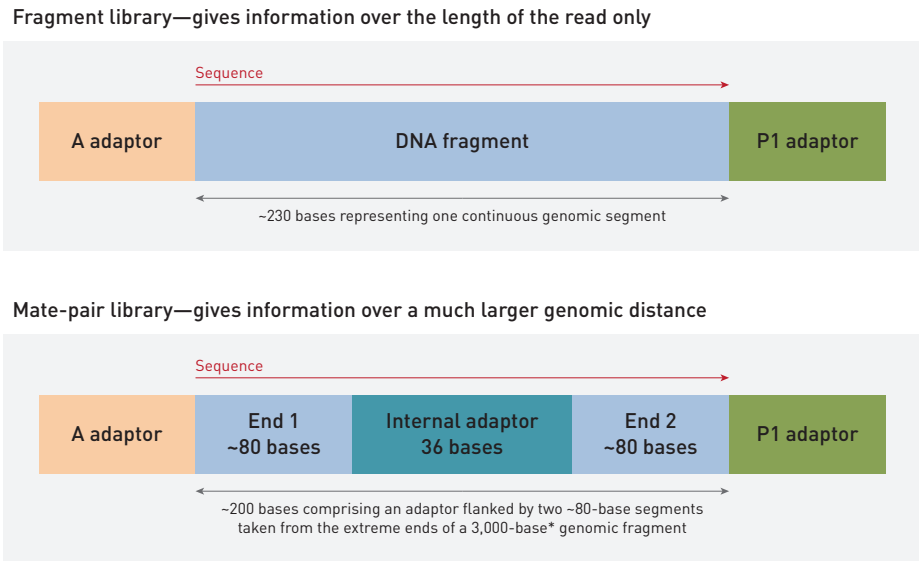

Figure 1. The mate-pair library delivers sequence position information over a much larger genomic distance than the fragment library.

demonstrated a constrained distribution around the expected size (Figure 3). [Note: A description of how to use sff_extract tool to identify the internal adaptor sequence can be found in a user bulletin, Using MIRA Assembly with Ion Torrent PGM Reads (http://lifetech-it.hosted.jivesoftware.com/docs/DOC-2163.)

Assembling the sequence

From the *E. coli* MG1655 genome, three libraries were constructed and sequenced: one fragment library (mean length = 207 bases), one 3.5 kb insert mate-pair library, and one 8.9 kb insert mate-pair library. After splitting the mate-pair reads into two tags and removing the internal adaptor sequence, the average tag lengths were 80 bases. Libraries were subsampled to yield combined coverage of approximately 40-fold and then assembled using MIRA (see http://lifetech-it.hosted.jivesoftware.com/docs/DOC-2163).

Because the MIRA assembler does not perform scaffolding, the output of this tool is a set of contigs. Subsequent joining of these contigs into scaffolds was performed using the stand-alone SSPACE software, which maps the ends of mate-pair reads to the set of contigs and identifies pairs that link adjacent contigs. Assembly statistics are described in Table 1.

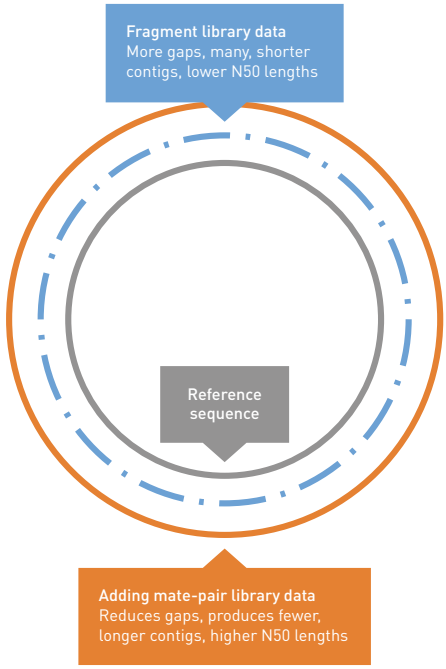

Figure 2. By augmenting the fragment library data with mate-pair library data, gaps in the assembly can be closed and contigs can be joined into scaffolds.

The fragment library alone yielded a contig assembly with an N50 of over 119 kb and reference genome coverage of 99.998%. For certain applications, this level of assembly may be adequate for answering the biological question of interest. For example, about 98% of protein-coding genes are expected to be contiguous, so genome content can be largely assessed.

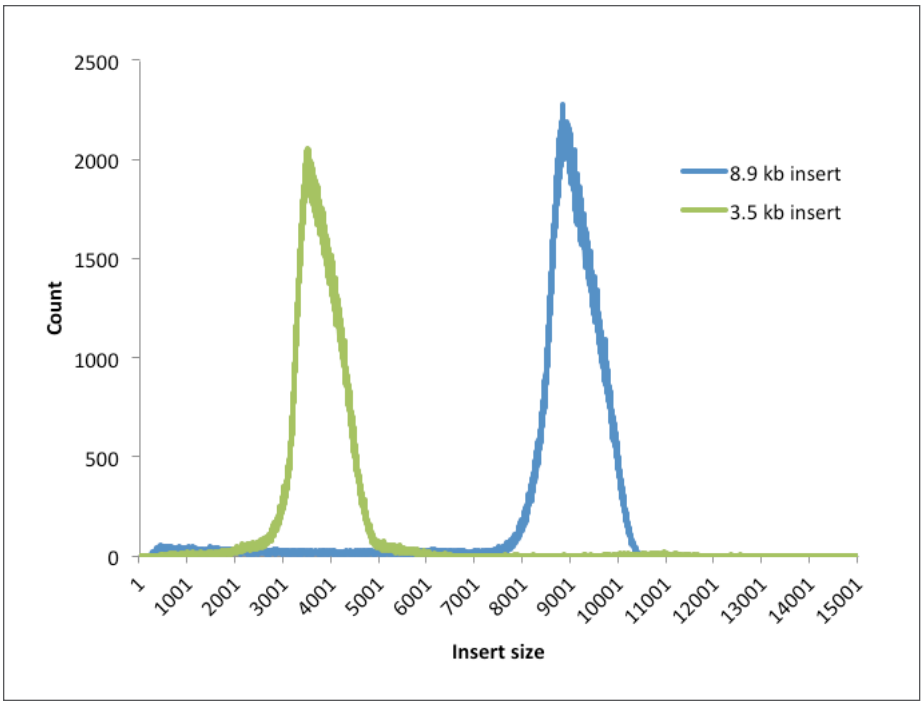

Figure 3. Insert length distribution of the mate-pair library.

Table 1. Improvements in contig and scaffold assemblies when mate-pair library data are used to augment fragment library data.

	Fragment only	Fragments plus 8.9 kb mates†	Fragments plus 8.9 kb mates plus 3.5 kb mates†
Contig			
Total consensus	4,607,823	4,615,388	4,613,542
Number of contigs	83	75	79
Largest contig	357,570	327,207	344,146
Contig N50*	119,187	157,428	177,681
% reference genome covered	99.998%	99.997%	99.997%
Scaffold			
Total consensus	NA	4,671,736	4,659,192
Number of scaffolds	NA	38	22
Largest scaffold	NA	3,046,212	2,337,293
Scaffold N50*	NA	3,046,212	2,337,293

* See glossary for definition of N50.
† Equal coverage of fragment reads and mate-pair reads.

For applications that require a more complete draft genome assembly, mate-pair data can be generated and combined with the fragment data. Adding the 8.9 kb insert, 2 x 80 base mate-pair data to the fragment data after subsampling each of the libraries down to 20-fold average coverage yielded further improvement in contig N50, with over half of the assembly in contigs larger than 157 kb. Scaffolding of these contigs using the 8.9 kb mate pairs further joins the contigs into 38 scaffolds with an N50 of 3.05 Mb. By adding a second mate-pair library with a 3.5 kb insert size to the assembly, further improvements in scaffolding are realized, with a total of 22 scaffolds and 99.96% of the genome in the four largest scaffolds.

Results of the three assemblies are depicted in Figure 4, in which contigs (blue) and scaffolds (green) are mapped to the MG1655 reference genome. Repeat sequences, including rRNA loci and transposons, are indicated around the outside of the circle. It should be noted that many gaps occur at these natural repeat sequences.

Conclusion

Based on the data from this MG1655 sequencing project, it is clear that the protocol described in the report Ion Mate-Paired Library Preparation (http://lifetech-it.hosted.jivesoftware.com/docs/DOC-1999) for creating mate-pair libraries and analyzing them using Ion semiconductor sequencing produces significantly improved assemblies, and when the reads are assembled using the MIRA assembler and the resulting data combined with fragment-read data, contig and scaffold assemblies are improved. Using SSPACE software to further join contigs based on mate-pair data results in extensive scaffolds that comprise the vast majority of the genome. The assembly improves significantly—83 contigs are seen with fragment-only assemblies, while using mate-pair data produces 22 scaffolds and an extra 52 kb of consensus sequence. Most of the remaining breaks in the MG1655 assembly are due to natural repeat sequences in the genome. After releasing this dataset to the Ion Community, another researcher has shown a complete single scaffold assembly using other software tools. Please see http://flxlexblog.wordpress.com/2012/03/02/ion-torrent-mate-pairs-and-a-single-scaffold-for-e-coli-k12-substr-mg1655/

Glossary

Mate pair (MP) vs. paired end (PE)

There is some confusion regarding these two terms depending on the reference. In this application note, a mate pair is defined as a sequencing construct that places the extreme ends of a longer genomic DNA fragment physically close together using molecular biology techniques so that the sequence of both ends of this long DNA fragment are contained in a much shorter fragment.

Paired-end sequencing describes a type of sequencing in which a single fragment is sequenced from both ends. Currently, paired-end reads are also enabled for the Ion PGM™ system with certain modifications to the library and sequencing conditions (see Paired-End Sequencing application note).

N50

N50 is a weighted statistical measure of the median contig length in a set of sequences. Larger N50 values correlate to more complete assemblies. All contigs are sorted from largest to smallest, and then the N50 is determined from the minimum set of contigs (starting with the largest contig and adding successive contig lengths) with cumulative size totaling 50% of the assembled genome. For example, for 5,000 bases of assembled sequence, the N50 length is the fragment size arrived at in the list when the cumulative size is at least 2,500 bases. For a microbial genome such as *E. coli* O104:H4 (assuming an average bacterial gene size of 1 kb), an N50 length of 10 kb would mean that about 90% of the genes in that asssembled genome sequence are intact; an N50 of 100 kb would mean that about 99% of the genes in that asssembled genome sequence are intact.

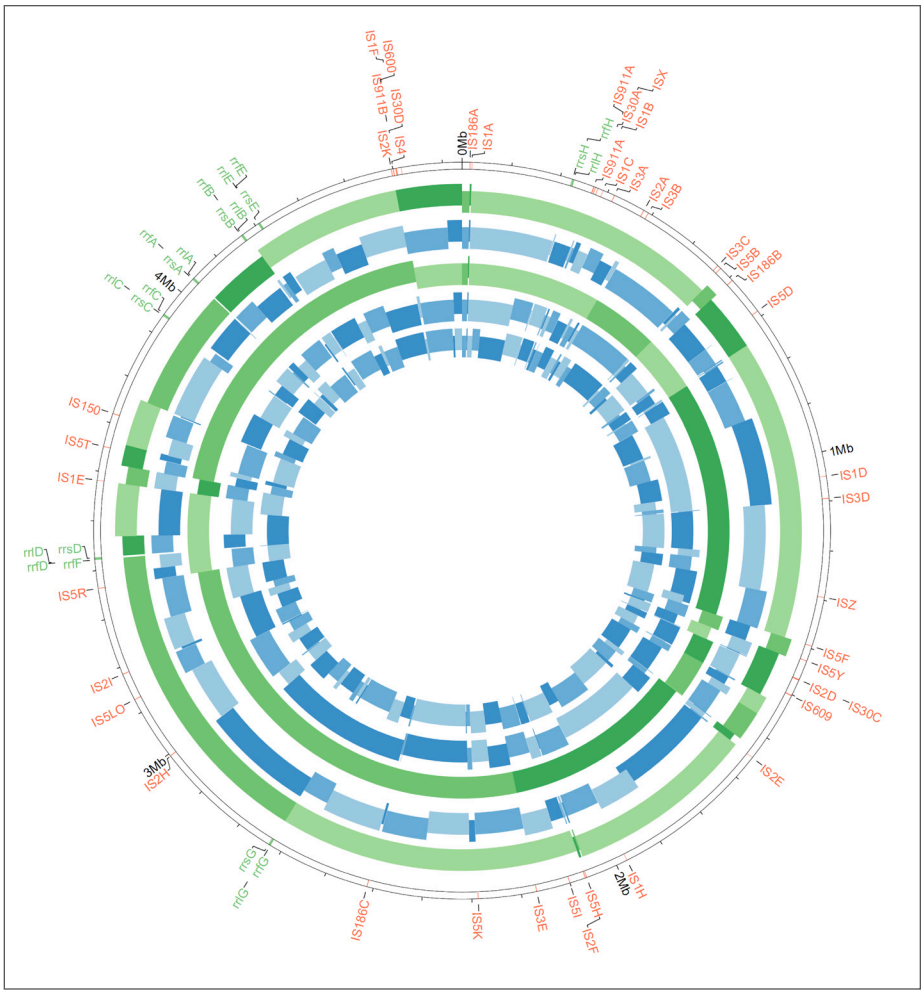

Figure 4. Circular plot of contigs and scaffold coverage. Assemblies are mapped against the *E. coli* MG1655 reference chromosome using the MUMmer software suite, and alignments are depicted using Circos. Contigs are in shades of blue and scaffolds are in shades of green. The inside circle represents contigs from the fragment library assembly. Moving outward, the next two circles are contigs and scaffolds from the fragment library plus 8.9 kb mate pair (MP) assembly. The outer two alignments represent contigs and scaffolds from the assembly of the fragment library plus both MP libraries. Each contig or scaffold alignment is depicted as a block, with individual contigs differentiated by stagger and shades of color. Repetitive sequences, including mobile elements (red) and rRNA loci (green), are indicated around the outside of the circle.

References

1. Mellmann A, Harmsen D, Cummings CA et al. (2011) Prospective genomic characterization of the German enterohemorrhagic Escherichia coli O104:H4 outbreak by rapid next generation sequencing technology. *PLoS One* 6:e22751.
2. Rohde H, Qin J, Cui Y et al. (2011) Open-source genomic analysis of Shiga-toxin-producing E. coli O104:H4. *N Engl J Med* 365:718–724.

Printed in the USA. **CO23721 0612**