De Novo Sequencing on the Ion Torrent PGM

Mate-pair library data improves genome assembly

- Highly accurate PGM data allows for de Novo Sequencing and Assembly
- For a draft assembly, generate contig assembly of N50 (on fragment data) of >119kb
- For a higher quality assembly, add Mate-pair data to generate an increase in contigs, N50 of >177kb (~50% increase) with only 22 scaffolds (N50 > 2.3Mb)

In order to carry out de novo sequencing large amounts of high quality sequence is required. The Ion Torrent PGM™ has a simple fast workflow that enables the rapid generation of large volumes of data. Rapid whole-genome sequencing of epidemiologically relevant organisms has been carried (Mellmann et al. (2011) PLoS One 6:e22751; Rhode et al. (2011) NEJM, 365:718-724.), but the final assembly inevitably contains gaps in the sequence. In this application note we show how the fragment library data can be augmented with mate-pair library data to produce a much higher quality assembled sequence (larger contigs, 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 can roughly be divided into two camps—Sanger shotgun sequencing and massively parallel high-throughput sequencing. Sanger sequencing yields very accurate reads up to 1,000 bases, but this methodology is more expensive and time consuming. The Ion Torrent PGM™ system produces read lengths of up to 200 bases, shorter than those of Sanger sequencing. Many researchers see this as a worthwhile tradeoff because next-gen sequencing 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 prevent 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 is sufficient. For example, when establishing contigs for many comparative genomics approaches, the extra time, effort, and expense required to fill sequence gaps is not warranted. When the sequencing strategy calls for complete or near-complete sequence, however, 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 base pairs each to 50 kbp—and package them in short fragments that are suitable for next-gen sequencing (Figure 1). This means that instead of getting positional information only for the stretch of sequence in each fragment library clone, you now get positional information 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 can be closed, which converts smaller contigs 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 allows you to generate a sequence that approaches “finished” status.
Fragment library—gives information over the length of the read only

A Adaptor | DNA Fragment | P1 Adaptor
---|---|---

≈230 bases representing one contiguous genomic piece

Mate-pair library—gives information over a much larger genomic distance

A Adaptor | End 1 ~80 bases | Internal adaptor 36 bases | End 2 ~80 | P1 Adaptor
---|---|---|---|---

≈200 bases comprising an adaptor flanked by two ~80-base segments taken from the extreme ends of a 3,000-base* genomic fragment

* Note that this example assumes 3,000 bp fragments were used to construct the mate-pair library. In practice lengths from a few hundred base pairs to 50 kbp can be used.

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

In order to demonstrate the benefits of Mate pairs, a model organism (E. coli MGG17655) is sequenced and assembled using fragment data and Mate Pairs.

Mate-pair library construction

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

Pairing analysis after acquiring sequencing data

After identifying the internal adapter sequence and then splitting the sequencing read into two tags, the individual tags are mapped to the reference genome and the distance between the reads is 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 demonstrates a constrained distribution around the expected size (Figure 3). [Note: a description of how to use SFF extract to identify the internal adaptor sequence can be found in the user bulletin, Using MIRA assembly with Ion Torrent PGM reads.docx at http://lifetech-it.hosted.jivesoftware.com/docs/DOC-2163.]
Assembling the sequence

From the *E. coli* MG1655 genome, three libraries are constructed and sequenced: one fragment library (mean length = 207 bp), 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 are 80 bp. Libraries are subsampled to yield combined coverage of approximately 40-fold and then assembled using MIRA (see [http://lifetech-it.hosted.jivesoftware.com/docs/DOC-2163](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 is performed using the standalone SSPACE software, which maps the ends of mate-paired reads to the set of contigs and identifies pairs that link adjacent contigs. Assembly statistics are described in Table 1.
Table 1. Improvements in contig and scaffold assemblies when mate-pair library data are used to augment fragment library data.

<table>
<thead>
<tr>
<th></th>
<th>Frag only</th>
<th>Frags plus 8.9 kb mates†</th>
<th>Frags plus 8.9 kb mates plus 3.5 kb mates†</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contig stats</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total consensus</td>
<td>4,607,823</td>
<td>4,615,388</td>
<td>4,613,542</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>83</td>
<td>75</td>
<td>79</td>
</tr>
<tr>
<td>Largest contig</td>
<td>357,570</td>
<td>327,207</td>
<td>344,146</td>
</tr>
<tr>
<td>Contig N50*</td>
<td>119,187</td>
<td>157,428</td>
<td>177,681</td>
</tr>
<tr>
<td>% reference genome covered</td>
<td>99.998%</td>
<td>99.997%</td>
<td>99.997%</td>
</tr>
<tr>
<td><strong>Scaffold stats</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total consensus</td>
<td>n/a</td>
<td>4,671,736</td>
<td>4,659,192</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>n/a</td>
<td>38</td>
<td>22</td>
</tr>
<tr>
<td>Largest scaffold</td>
<td>n/a</td>
<td>3,046,212</td>
<td>2,337,293</td>
</tr>
<tr>
<td>Scaffold N50*</td>
<td>n/a</td>
<td>3,046,212</td>
<td>2,337,293</td>
</tr>
</tbody>
</table>

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

The fragment library alone yields 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.

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 bp 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 is 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.
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 plus 8.9 kb MP assembly. The outer two alignments represent contigs and scaffolds from the assembly of fragment 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.

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* for creating mate-pair libraries and sequencing them on the Ion PGM™ System 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 MP data produces 22 scaffolds and an extra 52kb of consensus sequence. Most of the remaining breaks in the MG1655 assembly are largely due to natural repeat sequences in the genome.

**Mate-pair vs. paired-end**

There is some confusion regarding these two terms depending on the source. In this application note, a mate-pair is defined as a sequencing molecule that places the extreme ends of a longer genomic DNA fragment physically close together using molecular biology techniques so that the sequence for both ends of this long DNA fragment is 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 PGM with certain modification on the library and sequencing condition (see Paired End Sequencing application note).

**N50**

N50 is a weighted statistical measure of the median contig length in a set of sequences. Or, to put it another way, the N50 is length L (in base pairs) such that 50% of the bases are in contigs the size of L or greater. Larger N50 values correlate to more complete assemblies.

N50 was calculated by sorting all of the lengths of the contigs from largest to smallest and then determining the minimum set of contigs (starting with the largest contig on the list and adding successive contig lengths) until the cumulative sizes total 50% of the assembled genome. For example, for 5,000 bases of assembled sequence, the N50 length is fragment size you arrive 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 genome are intact; an N50 of 100 kb would mean that about 99% of the genes in that genome are intact.