

Significantly improved *de novo* assembly of small genomes using the Ion TrueMate™ Plus Library Kit on the Ion PGM™ System

Key findings:

- A single scaffold covering the genome of *E. coli* (K-12/MG1655) was produced by the combination of mate-pair and fragment reads assembled using DNASTAR Lasergene® Genomics Suite
- The use of mate-pair and fragment library data improved contig N50 values by 65%, while decreasing the number of contigs by 61% when compared to fragment-only assemblies (Figure 1)
- The new Ion PGM™ Hi-Q™ sequencing chemistry reduced insertion/deletion (indel) errors by 67% and substitution errors by 25% compared to the previous version of the Ion PGM™ 400 base pair chemistry

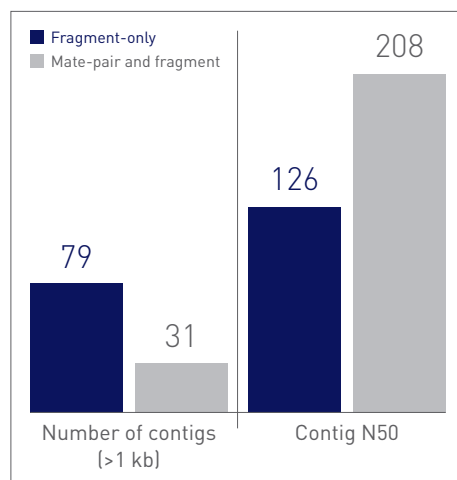


Figure 1. The addition of mate-pair reads to fragment reads improves *de novo* assembly. Assembly using SPAdes v3.1 of *E. coli* (K-12/MG1655) with fragment-only reads (dark blue) or mate-pair combined with fragment sequencing reads (grey) decreases the number of contigs (79 to 31) with concomitant increase in N50 contig lengths (126 kb to 208 kb).

Mate-pair libraries improve microbial genome *de novo* assembly

Mate-pair libraries span larger regions of the genome supporting consistent mapping of repeats, mobile elements, paralogous regions, and regions of low complexity or high homology, resulting in superior *de novo* assemblies.

A mate-pair library is a specific type of DNA fragment library derived from fragments separated by several kilobases of intervening sequence. The resulting single-end sequencing read consists of two juxtaposed sequence tags (a pair) from opposite ends of the same long, user-defined DNA insert of approximately known size (Figure 2).

The output of *de novo* assemblies incorporating fragment and mate-pair sequencing data is a set of scaffolds composed of contigs and gaps. A contig, short for contiguous sequence, is a length of genomic sequence assembled from overlapping sequence reads so that the order of bases is known with reasonably high confidence. Ideally a single contig would result from an assembly but repeats, mobile elements, and other regions of low-complexity sequence hinder the assembly process resulting in gaps and unordered contigs.

Since the length of the intervening sequence (library insert size) between mate-pair reads is roughly known,

combining fragment library assembly data with the mate-pair reads that overlap with other reads in two different contigs results in the ordering of adjacent contigs into scaffolds with gaps of estimated length but unknown sequence.

Mate-pair sequencing identifies novel genomes and genomic features

Accurate assemblies are critical for the identification of novel structural and functional genomic arrangements in newly sequenced microbial strains. For example, using the Ion PGM™ System, sequencing of mate-pair and fragment libraries improved the assembly of a *Corynebacterium pseudotuberculosis* genome allowing the identification of a phage and nitrate-reductase operon that were absent from the reference genome [1]. By integrating Ion Torrent™ mate-pair and fragment library data, novel insight into the genomic evolution of the important human pathogen *Mycobacterium tuberculosis* was revealed through sequencing of the SP21 strain [2].



As part of a study investigating microbial control of methane from anoxic marine sediments, the assembly and scaffolding of the Candidatus *Methanoperedens nitroreducens* genome was improved using a 3 kb insert mate-pair library [3]. Mate-pair sequencing on the Ion PGM™ System was also used to scaffold and improve the genome assembly of a novel archaeal methogen, Candidatus *Methanoflorens stordalenmirensis*, that dominates a thawed permafrost site and is possibly a major contributor to global methane production and climate change [4].

The Ion TrueMate™ Plus Library Kit

The Ion TrueMate™ Plus Library Kit enables highly efficient construction of mate-pair libraries with insert size range of 2–8 kb from any genomic DNA. Optimized over current demonstrated protocols for Ion Torrent™ sequencing, the Ion TrueMate™ Plus Library Kit requires only 2–10 µg total input DNA and importantly, does not involve agarose gels (“gel-free”). With ≥70% of paired true mate-pair reads used for assembly, successful genome scaffolding results with minimal bias and low redundancy. Encrypted Chimera Code™ detection in the Ion TrueMate™ Plus Library Kit minimizes false mate-pair fragment configurations (i.e., junk reads) while Junction Code™ sequence identifies junctions of true mate-pair fragments (Figure 2).

Uniformity and mate-pair yield for the Ion TrueMate™ Plus Library Kit

The *E. coli* (K12/MG1655) genome was used to characterize the performance of Ion TrueMate™ Plus Library Kit. Sequencing mate-pair libraries resulted in an average mate-pair distance of ~8,300 bp, demonstrating excellent agreement with the targeted

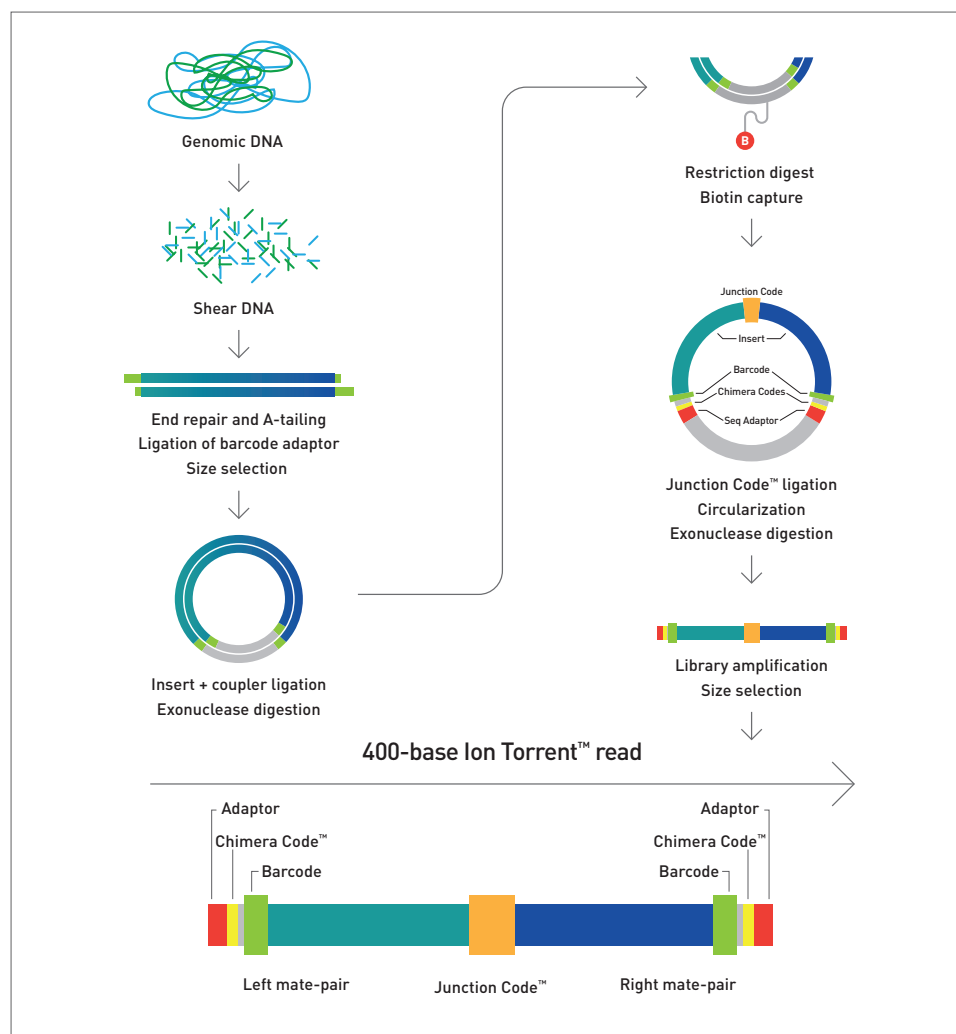


Figure 2. Ion TrueMate™ Plus library preparation workflow. Genomic DNA was physically sheared to generate 8 kb fragments which were end-repaired, A-tailed, and adapted with Ion Torrent™ sequencing adaptors. Following size selection, fragments were circularized with the coupler sequence and digested with restriction endonucleases to yield linear molecules containing short genomic sequence tags separated by the intervening coupler adaptor (shown in grey). These linear fragments were then purified from chimeras and were re-circularized by ligating a Junction Code™ adaptor between the tags. An exonuclease digestion was used to digest remaining linear DNA and a final amplification step yielded the mate-pair library for Ion PGM™ sequencing.

8 kb insert size and superb uniformity with 41% of mate-pair read distances within 20% of target insert size. Sequencing reads with matching left and right chimera codes are defined as true mate-pair reads while paired true mate sequencing reads have read 1 and read 2 split at junction codes (Figure 2). Supporting the improved sequencing and assembly efficiency through use of these codes, true mate-pair reads accounted for up to 57% of the total sequencing reads generated whereas assembled mate-pairs accounted for up to 81% of the paired true mate-pair reads.

Improved accuracy with the new Ion PGM™ Hi-Q™ sequencing chemistry

Fragment and mate-pair libraries representing the *E. coli* (K12/MG1655) genome were sequenced with the new Ion PGM™ Hi-Q™ Sequencing Kit and these runs were compared to the same libraries sequenced using the Ion PGM™ Sequencing 400 Kit. As a result of enzyme evolution and sequencing chemistry improvements, both the substitution and insertion/deletion (indel) error rates per 100 bp were improved with a read-level reduction of between 23%–64% and 16%–25%, respectively.

To assess the improvement of the Ion PGM™ Hi-Q™ sequencing chemistry on assembled reads, *E. coli* (K-12/MG1655) fragment libraries reads were *de novo* assembled using SPAdes v3.1 (<http://bioinf.spbau.ru/spades>). Evaluation of the fragment-only assemblies using QUAST (<http://bioinf.spbau.ru/en/quast>) indicated a consensus-level reduction in the indel error rate per 100 kb of 59%.

Improved *de novo* assembly with fragment and mate-pair library data

Important metrics for assessing the quality of an assembly are contig N50 and the total number of contigs. Contig N50 is the size of the smallest contig, in an ordered set of the largest contigs, which cover 50% of the total size of all the contigs comprising an assembled genome. Consequently, larger N50 values and fewer contigs correlate to more complete and less fragmented assembly.

Inclusion of mate-pair library data substantially improved *E. coli* (K12/MG1655) genome assembly metrics. Using the SPAdes v3.1 assembler, the N50 contig sizes increased from 126 kb to 208 kb (65%) and the number of contigs decreased from 79 to 31 (61%) for fragment and mate-pair data when compared to fragment data alone (Table 1 and Figure 1). Similarly, using the DNASTAR assembler, mate-pair and fragment library data improved contig N50 values by 45% (from 96 to 139), while decreasing the number of contigs by 37% (from 95 to 60) when compared to fragment-only assemblies.

Both assemblers produced scaffolds covering nearly the entire genome with 4.57 megabases (Mbp) and 4.61 Mbp using SPAdes v3.1 and DNASTAR, respectively (Table 1). Fragment and mate-pair libraries sequenced with the Ion PGM™ Hi-Q™

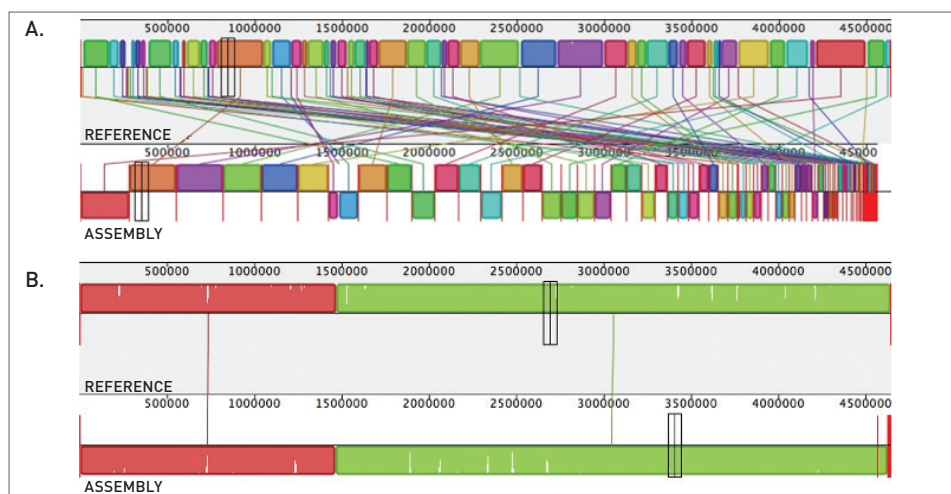


Figure 3. Scaffold assemblies produced using SPAdes v3.1 assembler using fragment and mate-pair libraries sequenced using the Ion Torrent PGM™ System. A) *De novo* assembly of fragment-only reads result in unordered contigs with coverage of ~98% of the reference. B) 8 kb Ion TrueMate™ Plus Library reads assembled with fragment library reads resulted in a single scaffold of 4.57 mb that is collinear with *E. coli* (K-12/MG1655) reference and a total ordered assembly that covers 99.32% of the reference.

Sequencing Kit were ordered into a single scaffold using DNASTAR and 4 scaffolds with a scaffold N50 of 4,570 kb using SPAdes v3.1 (Figure 3).

High reference coverage and improved assembly quality

Contig N50 values, number of scaffolds and other assembly metrics suggest the completeness of an assembly, but not the accuracy nor coverage of that genome. To assess coverage, the contigs from each assembly were aligned to the reference sequence with QUAST. Using either SPAdes or DNASTAR, the assembly of Ion PGM™ Hi-Q™ runs resulted in 99.21–99.32% coverage of the reference genome with excellent representation of 4,450–4,451 full genes in the assemblies and coverage of 98.91–99.80% of the genes encoded in the *E. coli* (K-12/MG1655) genome (Table 1).

Conclusions

The combination of the Ion TrueMate™ Plus Library Kit, the Ion Xpress™ Plus Fragment Library Kit, the Ion PGM™ System, and assembly using DNASTAR Lasergene® Genomics Suite (Figure 4) enabled a single scaffold assembly of *E. coli* (K-12/

MG1655). Indicative of improved, less fragmented assembly, mate-pair and fragment libraries reads—assembled using the SPAdes v3.1—delivered better contig N50 values, with a 65% increase compared to fragment-only assemblies, and an associated 61% decrease in number of contigs. In addition, sequencing chemistry improvements on the Ion PGM™ System, enabled by the Ion PGM™ Hi-Q™ sequencing polymerase, reduced error frequencies at the read level for fragment and mate-pair sequencing data ranging from 23–67% and 17–25% for indel and substitution errors, respectively. Ideal for gap closure and genome finishing, assembly of mate-pair and fragment read data resulted in a more complete assembly that covered ≥99.21% of the reference genome leading to an assembled sequence covering 4,450 full genes and ≥98.91% of the genes encoded in *E. coli* (K-12/MG1655) genome. In summary, Thermo Fisher Scientific supplies an easy-to-implement and cost-effective microbial genome sequencing workflow for improved *de novo* assembly of sequence reads from the Ion PGM™ System.

Table 1. Mate-pair and fragment library *de novo* assembly metrics and quality assessment from runs sequenced with the new Ion PGM™ Hi-Q™ sequencing chemistry.

Assembler	SPAdes v3.1		DNASTAR	
Libraries	Fragment-only library	Fragment and mate-pair	Fragment-only library	Fragment and mate-pair
Contigs				
Number of contigs (>1 kb)	79	31	95	60
Contig N50 (kb)	126	208	96	139
Assembly length (bp)	4,551,315	4,629,323	4,622,427	4,605,469
Scaffold				
Number of scaffolds	-	4	-	1
Largest scaffold (mb)	-	4.57	-	4.61
Scaffold N50 (kb)	-	4,570	-	4,605
Percent reference coverage	-	99.32	-	99.21
Percent genes covered (number)	-	98.91 (4,467)	-	99.80 (4,507)
Number of full genes	-	4,450	-	4,451

References

1. Ramos RTJ, Carneiro AR, de Castro Soares S, et al. (2013). High efficiency application of a mate-paired library from next-generation sequencing to postlight sequencing: *Corynebacterium pseudotuberculosis* as a case study for microbial *de novo* genome assembly. *J Microbiol Methods* 95:441–447.
2. Shitikov EA, Bespyatykh JA, Ischenko DS, et al. (2014). Unusual large-scale chromosomal rearrangements in *Mycobacterium tuberculosis* Beijing B0/W148 cluster isolates. *PLoS ONE* 9: e84971.
3. Haroon MF, Hu S, Shi Y, et al. (2013). Anaerobic oxidation of methane coupled to nitrate reduction in a novel archaeal lineage. *Nature* 500: 567–570.
4. Mondav R, Woodcroft BJ, Kim EH, et al. (2014). Discovery of a novel methanogen prevalent in thawing permafrost. *Nat Commun* 5:3212.

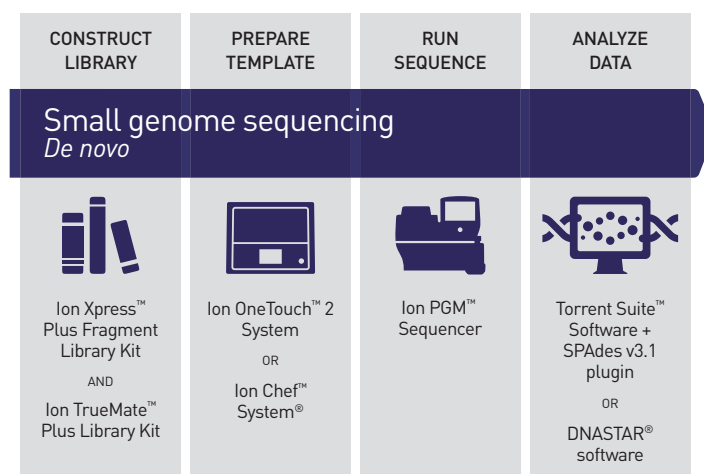


Figure 4. Workflow for small genome *de novo* sequencing. Genomic DNA from *E. coli* (K-12/MG1655) was used to prepare a fragment library using the Ion Xpress™ Plus Fragment Library Kit and an 8 kb fragment insert mate-pair library using the Ion TrueMate™ Plus Library Kit (using a shearing target of 10 kb with the Covaris g-TUBE™ device). Template preparation and enrichment for both library types was accomplished using the Ion PGM™ Template OneTouch™ 2 400 Kit and the Ion OneTouch™ 2 System. Sequencing was performed using the Ion 318™ Chip Kit v2 and the Ion PGM™ Sequencing 400 Kit or the new Ion PGM™ Hi-Q™ sequencing chemistry. Primary data analysis was performed using Torrent Suite™ software v4.2 and *de novo* assembly accomplished using SPAdes v3.1 plugin on the Torrent Server or with the DNASTAR Lasergene® Genomics Suite software package.

Find out more at lifetechnologies.com/truemate

