De Novo Assembly and Genomic Structural Variation Analysis with Genome Sequencer FLX 3K Long-Tag Paired End Reads

Thomas Jarvie¹ and Timothy Harkins²
¹454 Life Sciences, 20 Commercial Street, Branford, CT 06405, USA
²Roche Diagnostics, Roche Applied Science, 9115 Hague Road, Indianapolis, IN 46250, USA
Correspondence should be addressed to Thomas Jarvie (thomas.jarvie@roche.com)

The Genome Sequencer FLX System from Roche and 454 Life Sciences™ is a versatile sequencing platform suitable for a wide range of applications, including de novo sequencing and assembly of genomic DNA, transcriptome sequencing, metagenomics analysis, and amplicon sequencing. The Genome Sequencer FLX enables long sequence reads separated by kilobase distances of genomic DNA. These Long-Tag Paired End reads enable improved de novo assemblies and genomic structural variation studies.

454 Life Sciences has developed and commercially released a new protocol for generating a library of paired-end fragments to determine the orientation and relative positions of contigs produced by de novo shotgun sequencing and assembly. This 3K Long-Tag Paired End protocol (Figure 1) can also be used to identify genomic structural variations (1) and their associated breakpoints. Structural variation of the genome, involving large, kilo- to mega-base-sized deletions, duplications, insertions, inversions, and complex combinations of rearrangements, is widespread in humans and is presumably responsible for a considerable amount of phenotypic variation. The 3K Long-Tag Paired End library DNA fragments contain an approximately 250-bp fragment with a 44-mer adaptor sequence in the middle, flanked by 100-mer sequences, on average. The two flanking 100-bp sequences are segments of DNA that were originally located approximately 3 kb apart in the genome of interest. In addition to the 3-kb paired-end protocol, initial results from an unreleased protocol that generates flanking reads separated by 16 kb are presented (Figure 2). The 16-kb protocol utilizes a different chemistry than the 3-kb protocol described here.

Traditional approaches to the sequencing of paired-end reads rely upon inserting a DNA fragment into a vector, such as a BAC or fosmid, cloning into bacteria, and subsequently generating two sequences, one from each end of the vector. These methods entail weeks of laboratory work and could cost several hundred thousand dollars to prepare the libraries needed for Sanger sequencing. The Genome Sequencer FLX method presented here, which requires no cloning, generates up to 200,000 paired-end reads from a single Genome Sequencer FLX instrument run with a total elapsed time — from genomic DNA to result — of less than four days.

Sample Preparation Protocol

The preparation of a 3K Long-Tag Paired End library is depicted schematically in Figure 1.

The protocol begins with fragmentation of the high molecular weight DNA sample by hydrodynamic shearing (HydroShear, Genomic Solutions); the size distribution of the fragments (on average 3 kb) will determine the distance between the Paired End sequencing tags. After purification using size exclusion beads (AMPure, Agencourt) the fragments are protected from EcoRI cleavage by methylation with EcoRI Methylase, and their ends are polished (blunted and 5′-phosphorylated). These ends are made blunt with T4 DNA polymerase and T4 polynucleotide kinase (T4 PNK).

Hairpin Adaptors (biotinylated, and containing non-methylated EcoRI recognition sites; provided in the GS Paired End Adaptor Kit) are ligated onto both ends, and all DNA species that are not protected by hairpins are removed by exonuclease digestion. The small, unwanted molecular species are removed with AMPure beads. The remaining long insert fragments are circularized by digestion with EcoRI to remove the terminal hairpin structures, providing cohe-

Figure 1. Schematic overview of the 3K Long-Tag Paired End Sequencing protocol. Genomic DNA is randomly fragmented and subsequently circularized via adaptors. After circularization, the DNA is randomly fragmented, followed by enrichment for linker-positive fragments via streptavidin (SA) capture. Long Paired End Adaptors (A and B) are then ligated onto the linker-positive fragment ends. Fragments that contain the A and B Adaptors are captured via streptavidin beads for emPCR amplification and subsequent 454 Sequencing™. Met, methylated; Bio, biotin.
sive ends for ligation. The resulting 3-kb circular fragments contain the 44-bp linker (the remainder of the two Hairpin Adaptors), joining the two ends of the fragmented DNA.

The DNA circles are then fractionated by nebulization, generating molecules that are a few hundred base pairs in length, with random-sized paired ends flanking the linker (plus other random fragments from the circles). After polishing the fragment ends (as above with T4 DNA polymerase and T4 polynucleotide kinase), the Paired End library fragments are immobilized onto streptavidin beads (Dynal M-270, Invitrogen) using the biotin tags incorporated into the 44-linker, resulting in the enrichment of linker-positive fragments. The Long Paired End Adaptors (sequences shown below) are ligated to the ends of the linker-positive fragments. The Adaptors provide priming sequences for both amplification and sequencing of the Paired End library fragments, as well as the “sequencing key,” a short sequence of four nucleotides that the Genome Sequencer System software uses for base calling and to recognize legitimate library reads.

The Adaptors are combined with the Long Paired End library fragments in a ligation reaction containing a large molar excess of Adaptors. The excess Adaptors maximize utilization of the Paired End library fragments and minimize the formation of fragment concatamers. Because the oligonucleotides that make up the Long Paired End Adaptors are non-phosphorylated to minimize concatamerization of the Adaptors, a subsequent step of nick repair with Bst DNA polymerase is required. Following nick repair, the double-stranded library is moderately amplified by PCR to generate sufficient material for accurate quantitation. One of the amplification primers is biotinylated, allowing streptavidin capture of the double-stranded library and subsequent isolation of the library of (single-stranded) Long Paired End template DNA fragments. This single-stranded library is ready for emulsion-based clonal amplification (emPCR™) using the GS emPCR Kit II (Amplicon A, Paired End) and subsequent sequencing using appropriate GS Sequencing and GS PicoTiterPlate Kits and the Genome Sequencer FLX Instrument.

The 3K Long-Tag Paired End protocol can be used to generate approximately 400,000 reads per sequencing run; 150,000 to 200,000 of these reads are paired ends. The reads will be either linker positive (input DNA–44-mer linker–input DNA) or linker negative (only input DNA). Only reads that are linker positive and contain at least 15 bp of input DNA sequence on either side of the linker are considered paired-end reads. The other reads can be treated as traditional sequencing reads.

The Adaptors are combined with the Long Paired End library fragments in a ligation reaction containing a large molar excess of Adaptors. The excess Adaptors maximize utilization of the Paired End library fragments and minimize the formation of fragment concatamers. Because the oligonucleotides that make up the Long Paired End Adaptors are non-phosphorylated to minimize concatamerization of the Adaptors, a subsequent step of nick repair with Bst DNA polymerase is required. Following nick repair, the double-stranded library is moderately amplified by PCR to generate sufficient material for accurate quantitation. One of the amplification primers is biotinylated, allowing streptavidin capture of the double-stranded library and subsequent isolation of the library of (single-stranded) Long Paired End template DNA fragments. This single-stranded library is ready for emulsion-based clonal amplification (emPCR™) using the GS emPCR Kit II (Amplicon A, Paired End) and subsequent sequencing using appropriate GS Sequencing and GS PicoTiterPlate Kits and the Genome Sequencer FLX Instrument.

The 3K Long-Tag Paired End protocol can be used to generate approximately 400,000 reads per sequencing run; 150,000 to 200,000 of these reads are paired ends. The reads will be either linker positive (input DNA–44-mer linker–input DNA) or linker negative (only input DNA). Only reads that are linker positive and contain at least 15 bp of input DNA sequence on either side of the linker are considered paired-end reads. The other reads can be treated as traditional sequencing reads.

**ASSEMBLY RESULTS**

The 3K Long-Tag Paired End protocol enhances the de novo assembly of shotgun-fragmented DNA by aligning the assembly contigs into scaffolds. Examples of assemblies resulting from the 3K Long-Tag Paired End protocol are shown in Table 1. Three bacterial genomes, *E. coli* K-12, *T. thermophilus*, and *C. jejuni*, were assembled with three different sequencing methods: pure shotgun sequencing...
reads (250 to 300 bases in length); a 50:50 mix of shotgun and 3K Long-Tag Paired End sequencing reads; and pure 3K Long-Tag Paired End sequencing reads. In all of the assemblies the number of reads in the data sets was minimized to an approximately 20× depth of coverage by randomly discarding sequence reads. The data used in the 50:50 mix were a 10× depth of shotgun reads and 10× depth reads from the 3K Long-Tag Paired End protocol. All assemblies were performed with the GS de novo Assembler Software version 1.1.03.

The GS de novo Assembler Software identifies the reads as either linker positive or linker negative. Linker-positive reads that contain 15 bp or less of input DNA sequence on either side of the linker are considered to be linker negative by the assembler. The initial step in the assembly is the generation of a de novo shotgun assembly using the linker-negative reads and the DNA reads on either side of the linker. Once the de novo assembler puts the shotgun reads into contigs, the linker-positive reads (Long-Tag Paired Ends) are used to orient the contigs into scaffolds. This method of assembly is used with 3K Long-Tag Paired End read data alone and when the 3K Long-Tag Paired End data is mixed with shotgun data.

As the results in Table 1 demonstrate, all three methods of assembly generate comprehensive, highly accurate assemblies. The Assembly Coverage in all assemblies is less than 100% as the portions of the genome with repeats longer than the 250- to 300-base read lengths are excluded. Overall accuracy was determined by mapping the Assembly Contigs against the appropriate reference genome and reporting discrepancies. The source of errors is due to one or more of the following: inaccuracy in the DNA sequence; inaccuracy of the reference sequence; or, the clone used to generate the reference sequence is slightly different than the clone used to generate the sequence data reported here. Inclusion of paired-end data into the assemblies aligns the Assembly Contigs into Scaffolds. The choice of experimental approach and assembly method will depend upon the goals of the research. If a quick view of the genome (for example, to identify which genes are present) is desired, then a shotgun-only approach is suitable. If the research goal is to generate a high-quality draft of the target genome, then the inclusion of Long-Tag Paired End data is the best option.

### SUMMARY

The sequencing of kilobase-size inserts is quite valuable for a number of applications, including improved de novo assembly and the identification of genomic structural variations. The 3K Long-Tag Paired End protocol provides a rapid, efficient, and cost-effective method for generating hundreds of thousands of sequence reads, each containing a pair of ~100-bp reads separated by 3-kb inserts. Future development plans include a protocol for sequencing tags separated by 15- to 20-kb distances. The combination of both 3-kb and longer paired-end spacing will better enable the assembly of larger and more complex genomes.

Additional information about the Genome Sequencer System is available from Roche Applied Science (www.genome-sequencing.com).

454, 454 Life Sciences, 454 Sequencing, emPCR, and PicoTiterPlate are trademarks of 454 Life Sciences Corporation, Branford, Connecticut, USA.

**For life science research only. Not for use in diagnostic procedures.** License disclaimer information is available online (www.genome-sequencing.com).

### REFERENCES