



454
SEQUENCING

3K Long-Tag Paired End sequencing with the Genome Sequencer FLX System

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 (Fig. 1) can also be used to identify genomic structural variations¹ and their associated breakpoints. Structural variation of the genome, involving large, kilobase- to megabase-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 comprise 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.

Traditional approaches to the sequencing of paired-end reads rely upon inserting a DNA fragment into a vector, such as a bacterial artificial chromosome or a fosmid, cloning this into bacteria and subsequently generating two sequences, one from each end of the vector. These methods entailed 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 4 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; the size distribution of the fragments (on average 3 kb) determines the distance between the paired-end sequencing tags. The fragments are methylated to prevent *EcoRI* cleavage, 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 remaining long insert fragments are circularized by digestion with *EcoRI* to remove the terminal hairpin structures, providing cohesive 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. Long Paired End Adaptors 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. This library is ready for

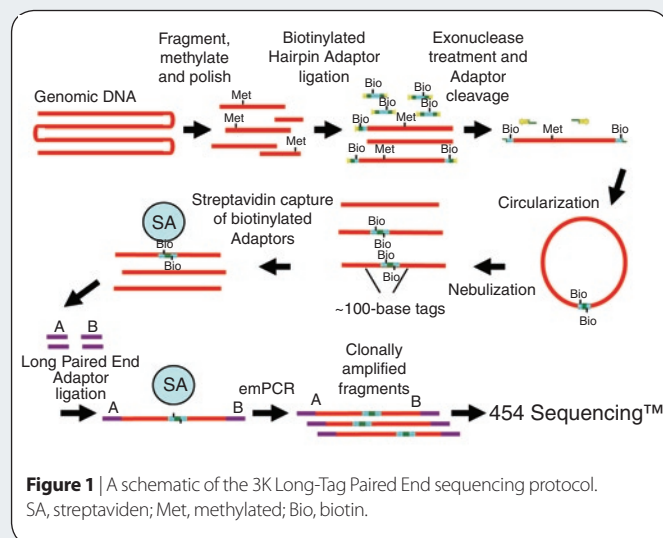


Figure 1 | A schematic of the 3K Long-Tag Paired End sequencing protocol. SA, streptavidin; Met, methylated; Bio, biotin.

Thomas Jarvie¹ & Timothy Harkins²

¹454 Life Sciences, 20 Commercial Street, Branford, Connecticut 06405, USA. ²Roche Diagnostics, Roche Applied Science, 9115 Hague Road, Indianapolis, Indiana 46250, USA. Correspondence should be addressed to T.J. (thomas.jarvie@roche.com).

APPLICATION NOTES

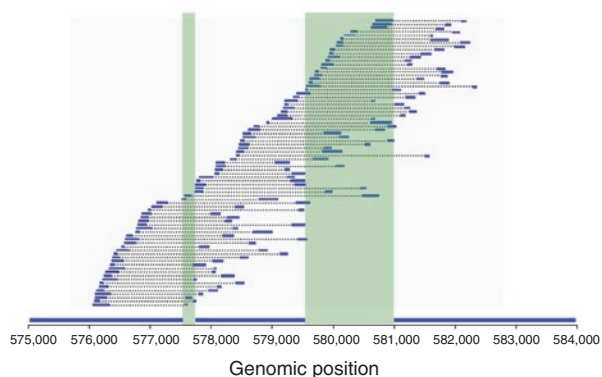


Figure 2 | Sequencing results from a typical 3K Long-Tag Paired End library prep of *E. coli* K-12. A region of the *de novo* assembly of *E. coli* K-12, with the *de novo*-assembled contigs covering the region shown in blue along the bottom axis. The paired-end reads generated with this protocol are capable of bridging the 0.2-kb and 1.5-kb gaps between the contigs, highlighted in green.

emulsion-based clonal amplification (emPCRTM) using the GS emPCR Kit II (Amplicon A, Paired End) and for sequencing using appropriate GS Sequencing and GS PicoTiterPlate kits and the Genome Sequencer FLX instrument.

Assemblies

Examples of assemblies resulting from the 3K Long-Tag Paired End protocol are shown in **Table 1**. We assembled three bacterial genomes, *Escherichia coli* K-12, *Thermus thermophilus* and *Campylobacter jejuni*, by three different sequencing methods: only shotgun sequencing reads (250–300 bases in length); a 50:50 mix of shotgun and 3K Long-Tag Paired End sequencing reads; and only 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 of reads generated using the 3K Long-Tag Paired End protocol.

The GS *de novo* Assembler Software (version 1.1.03) identifies the reads as either linker positive or linker negative. 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 places the shotgun reads into contigs, the linker-positive reads (Long-Tag Paired End reads) are used to orient the contigs into scaffolds (**Fig. 2**). This assembly method is used with 3K Long-Tag Paired End read data alone and when the 3K Long-Tag Paired End data are mixed with shotgun data.

All three methods of assembly generate comprehensive, highly accurate assemblies (**Table 1**). The choice of which experimental approach and assembly method to use depends on the goals of the research. If a quick view of the genome (for example, to identify which genes are present) is desired, 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.

Table 1 | Comparison of data from three methods of *de novo* assembly

	Shotgun	50:50 mix ^a	3K LT PE ^a
<i>E. coli</i> K-12			
Coverage depth	20×	20×	20×
Assembly contigs	147	113	121
Assembly coverage	97.62%	97.67%	97.49%
Overall accuracy	100.000%	99.999%	99.999%
Average contig (kb)	31.1	40.3	37.5
Largest contig (kb)	268.0	268.0	209.3
Scaffolds	–	11	10
<i>T. thermophilus</i>			
Coverage depth	20×	20×	20×
Assembly contigs	52	65	278
Assembly coverage	98.01%	98.08%	96.36%
Overall accuracy	99.997%	99.998%	99.998%
Average contig (kb)	40.6	32.5	7.5
Largest contig (kb)	482.3	291.8	49.8
Scaffolds	–	7	15
<i>C. jejuni</i>			
Coverage depth	23×	20×	20×
Assembly contigs	32	33	39
Assembly coverage	97.54%	97.59%	97.59%
Overall accuracy	99.993%	99.998%	99.997%
Average contig (kb)	50.4	48.9	41.5
Largest contig (kb)	304.5	304.5	304.5
Scaffolds	–	4	5

The 'assembly coverage' represents the non-repeat portions of the genome. 'Overall accuracy' was determined by mapping the 'assembly contigs' against the reference genome and reporting discrepancies. Inclusion of paired-end data into the assemblies aligns the assembly contigs into scaffolds.

^a50:50 mix, 50:50 mix of shotgun and 3K Long-Tag Paired End sequencing data; 3K LT PE, pure 3K Long-Tag Paired End sequencing data.

Summary

The sequencing of kilobase-sized inserts is quite valuable for a number of applications, including improved *de novo* assembly and identification of genomic structural variations. The 3K Long-Tag Paired End protocol provides a quick, 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 size 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 (<http://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 use only. Not for use in diagnostic procedures. License disclaimer information is available online (<http://www.genome-sequencing.com>).

1. Korbel, J.O. *et al.* Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).

This article was submitted to *Nature Methods* by a commercial organization and has not been peer reviewed. *Nature Methods* takes no responsibility for the accuracy or otherwise of the information provided.