Diagnostics

Genome Sequencer System

Application Note No. 1/July 2006

Whole Genome Assembly using Paired End Reads in *E. coli*, *B. licheniformis*, and *S. cerevisiae*

www.roche-applied-science.com

Whole Genome Assembly using Paired End Reads in *E. coli*, *B. licheniformis*, and *S. cerevisiae*

Corresponding author: Tom Jarvie, 454 Life Sciences Corporation, Branford, CT, USA, Email: tjarvie@454.com

Introduction

The Genome Sequencer System is the first-generation instrument employing the 454 sequencing technology [1]. This system enables researchers to perform rapid and comprehensive "whole genome shotgun sequencing" – the determination of the nucleotide sequence of entire genomes. Using this technology platform, scientists have been able to generate whole genome sequences for a wide variety of organisms, from viruses to fungi. The whole genome sequencing application employs shotgun sequencing of whole genomic DNA, typically to a 20-fold level of oversampling for a given genome. The Genome Sequencer 20 Instrument is capable of generating 20 million bases, or 200,000 reads of 100 bp, of sequence data per 5.5-hour run, enabling fast and cost-effective generation of the required level of data. The standard whole genome *de novo* sequencing application utilizes the Newbler Assembler software to assemble the reads into a number of unordered and unoriented contigs that typically cover >99% of the non-repeat regions of the genome. In contrast to the Sanger method of sequencing that employs Paired End reads in the original assembly process, 454 has developed a new protocol to generate a library of Paired End reads that are used to determine the orientation and relative positions of contigs produced by the *de novo* shotgun sequencing and assembly.

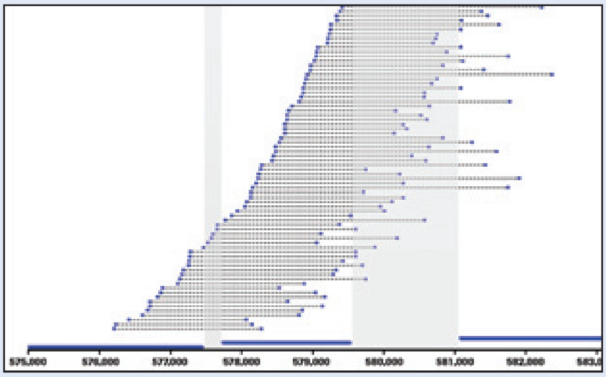

Figure 1: An illustration of the Paired End assembly process. Paired End reads are used to order and orient the contigs derived from the Newbler assembly. The large blue lines represent contigs generated from the whole genome shotgun sequencing and assembly. The multiple blue and grey lines represent Paired End information. The blue segments represent the two 20 nucleotide regions that were sequenced while the dotted grey line represents the distance between those two sequenced regions.

These Paired End reads are approximately 84-nucleotide DNA fragments that have a 44-mer adaptor sequence in the middle flanked by a 20-mer sequence on each side. The two flanking 20-mers are segments of DNA that were originally located approximately 2.5 kb apart in the genome of interest. The ordering and orienting of contigs generates scaffolds which provide a high-quality draft sequence of the genome. The 454 method, as compared to the Sanger method, dramatically reduces the time involved to generate a high-quality draft from weeks (or months) to days.

Materials and Methods

Materials

Equipment:

Genome Sequencer 20 Instrument (Software 1.0.53)

Reagents:

Sample Preparation: GS Paired End Adaptor Kit;
GS emPCR Kit II (Amplicon A, Paired End)
Sequencing: GS 20 Sequencing Kit; GS PicoTiterPlate Kit

A detailed list of all required equipment, reagents and methods is provided in the GS Guide to Paired End Sequencing.

Methods

For complete details on the Paired End DNA library preparation procedure and sequencing, please refer to the GS Guide to Paired End Sequencing, the GS emPCR Kit User's Manual, and the Genome Sequencer 20 Operator's Manual.

Procedure

Library Generation

Generation of a Paired End library: Intact genomic DNA is fragmented to yield an average length of 2.5 kb. The fragmented genomic DNA is methylated with EcoRI methylase to protect the EcoRI restriction sites. The ends of the fragments are blunt-ended, polished, and an adaptor DNA oligo is blunt-end ligated onto both ends of the digested DNA fragments. Subsequent digestion with EcoRI cleaves a portion of the adaptor DNA, leaving sticky ends. The fragments are circularized and ligated, resulting in 2.5 kb circular fragments. The adaptor DNA contains two MmeI restriction sites, and after treatment with MmeI the circularized DNA is cleaved 20 nucleotides away from the restriction sites in the adaptor DNA. This digestion generates small DNA fragments that have the adaptor DNA in the middle and 20 nucleotides of genomic DNA that were once approximately 2.5 kb apart on each end. These small, biotinylated DNA fragments are purified from the rest of the genomic DNA by using streptavidin beads. The purified Paired End fragments are processed through the standard library preparation protocol for the Genome Sequencer System.

emPCR and Sequencing

The Paired End library is processed through the standard emPCR process to generate hundreds of thousands of beads, each containing millions of copies of clonally amplified DNA. This bead-bound library is then deposited onto the PicoTiterPlate for sequencing with the standard protocol.

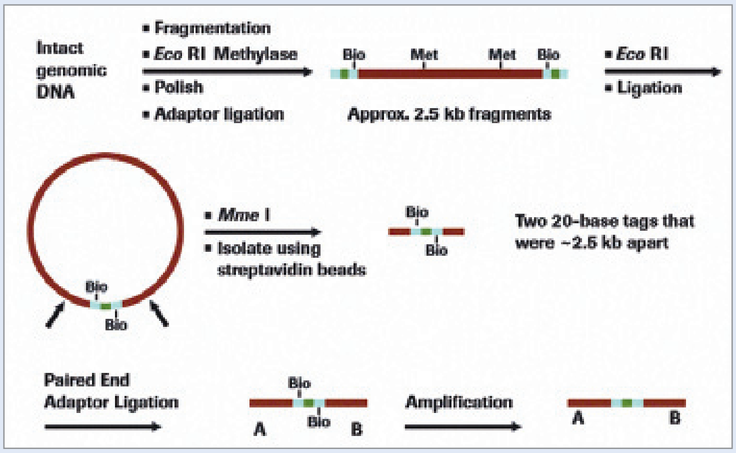

Figure 2: Illustration of the steps required to generate a Paired End library for sequencing.

The Paired End data is combined with standard whole genome shotgun sequencing data at an oversampling level of approximately 20-fold. Shotgun sequencing is performed using the standard whole genome preparation protocols for the Genome Sequencer System. An updated version of the Newbler Assembler, which incorporates both standard 100 bp reads and Paired End reads into the assembly, is used to generate contigs and scaffolds for the genome of interest.

Bioinformatics

Post-Run Analysis: Flow data was assembled using the Assembly Software (Newbler Assembler Software) of the Genome Sequencer 20 Software Version 1.0.53, as described in the GS 20 Data Processing Software Manual.

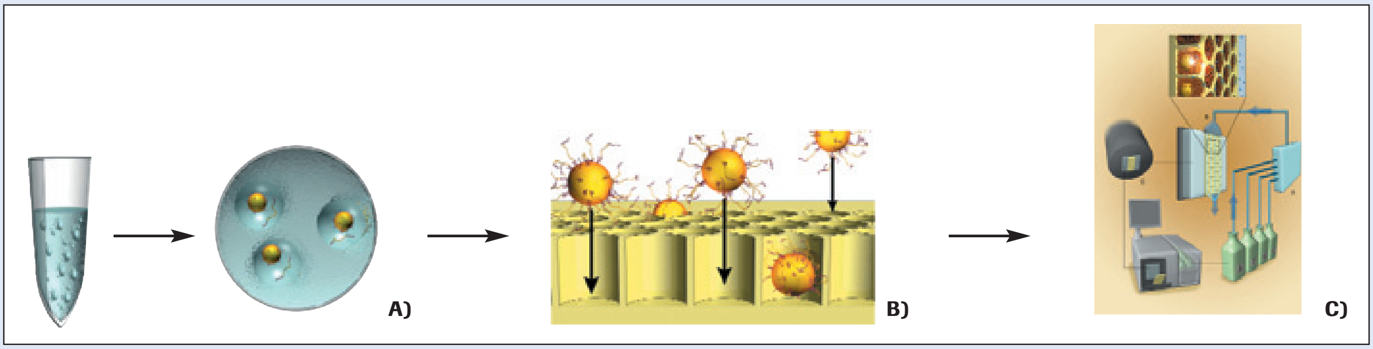

Figure 3:
A) The Paired End library is amplified onto beads using the emPCR process.
B) The bead-bound library is deposited onto the PicoTiterPlate.
C) The Paired End reads are sequenced on the Genome Sequencer 20 Instrument.

Results

Assembly of Escherichia coli *K12*
The 4.6 Mb genome of *E. coli* K12 was shotgun sequenced in three sequencing runs, yielding approximately 22-fold oversampling. Assembly performed with the Newbler Assembler software resulted in 140 unoriented contigs. One additional sequencing run of a Paired End library yielded approximately 112,000 reads. The Paired End data reduced the assembly to 24 scaffolds covering 98.6% of the genome.

Assembly of Bacillus licheniformis
The 4.2 Mb genome of *Bacillus licheniformis* ATCC14580 (DSM 13) was shotgun sequenced in three sequencing runs, yielding approximately 27-fold oversampling. Assembly performed with the Newbler Assembler software resulted in 98 unoriented contigs. One additional sequencing run of a Paired End library yielded approximately 255,000 reads. The Paired End data reduced the assembly to 9 scaffolds covering 99.2% of the genome.

Assembly of Saccharomyces cerevisiae
The 12.2 Mb genome of *Saccharomyces cerevisiae* S288C (containing 16 haploid chromosomes and one 86 kb mitochondrion) was shotgun sequenced in nine sequencing runs, yielding approximately 23-fold oversampling. The assembly performed with the Newbler Assembler software resulted in 821 unoriented contigs. Two additional sequencing runs of a Paired End library yielded approximately 395,000 reads. The Paired End data reduced the assembly to 153 scaffolds covering 93.2% of the genome.

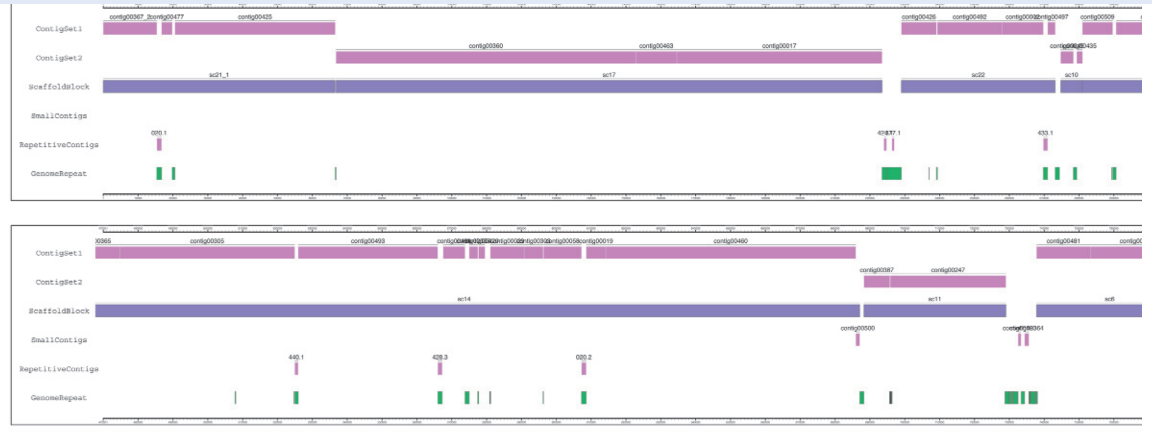

Figure 4: *De novo* assembly results for *E. coli* K12 aligned against a reference genome. The reference genome is represented by the top black line. The standard whole genome shotgun sequence and assembly is represented by the pink bars. Repeat regions of the genome are represented by the green bars at the bottom of the figure. Spaces between the pink bars are typically a result of repeat regions that cannot be uniquely assigned to a region in the genome. With the addition of one sequencing run of Paired End reads (represented by the purple bars) using the Genome Sequencer 20 Instrument, the genome sequence becomes much more complete.

NOTICE TO PURCHASER

RESTRICTION ON USE: Purchaser is only authorized to use the Genome Sequencer 20 Instrument with PicoTiterPlate devices supplied by 454 Life Sciences Corporation and in conformity with the procedures contained in the Operator's Manual.

Trademarks

PICOTITERPLATE is a trademark of 454 Life Sciences Corporation, Branford, CT, USA.
Other brands or product names are trademarks of their respective holders.

References

[1] Margulies, M (2005) Nature 437:376-380

For more information, visit
www.genome-sequencing.com

454 LIFE SCIENCES

Roche Diagnostics

Roche Diagnostics GmbH
Roche Applied Science
68298 Mannheim
Germany
www.roche-applied-science.com