BAC Combinatorial Pooling Strategy for Complex Genome Assembly

Simon, Serenity Banden, Laura Lambiase & Andrew Barry

Broad Institute Genome Sequencing Platform, Cambridge, MA, USA

Cambridge Rindge & Latin

Introduction

Sequencing and assembly of the complete DNA sequence of large and complex genomes has traditionally been performed using Sanger Sequencing. This technology has been used for both "BAC by BAC" and shotgun genome assembly approaches. The "BAC by BAC" approach involves sequencing many BACs, or Bacterial Artificial Chromosomes, each of which contains a 200kb DNA insert. Sanger sequencing technology, invented some 30 years ago, has been recently replaced by so-called "next-generation" sequencing technologies. These technologies, which utilize a "massively parallel" approach, analyze hundreds of thousands of molecules per detection event, whereas Sanger sequencing analyzes a single molecule per detection event. The result is a dramatic decrease in the cost per base pair of sequence data.

One tradeoff to these new technologies is the length of the read. While Sanger sequencing produced read lengths of 750 to 1,000 base pairs, these new technologies are only capable of producing read lengths of 100 base pairs. This poses a challenge to traditionally used assembly algorithms, which utilize overlapping reads to build regions of contiguous sequence.

The goal of this project is to test a method for sequencing complex genomes on next-generation technology by using a combined "BAC by BAC" and shotgun sequencing approach. The guinea pig, *C. porcellus*, was the model organism for this experiment. The method utilizes an indexed combinatorial pooling strategy, and will be used to assemble individual BAC inserts of 200kb, that will then be assembled into complete genome sequences. To accomplish this, we performed the combinatorial pooling strategy on 3,840 BAC inserts, and created indexed libraries out of each pool to be sequenced using Illumina sequencing technology.

Sample Pooling Diagram

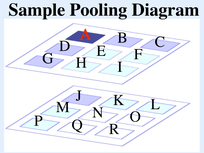
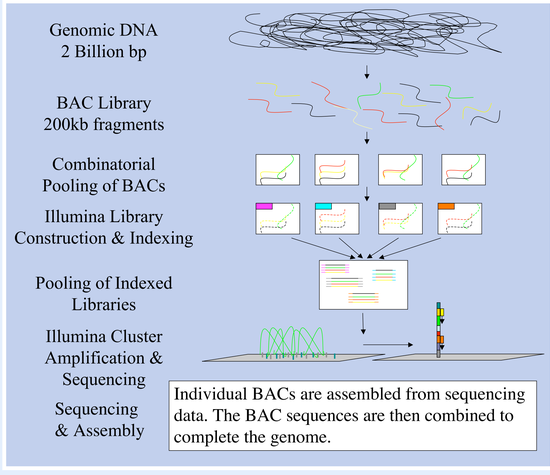


Plate 1: ABCDEFGHI
Plate 2: JKLMNOPQR

Row 1: ABCJKL
Row 2: DEFMNO
Row 3: GHIPQR

Column 1: ADGJMP
Column 2: BEHKNQ
Column 3: CFILOR

A unique sequence is added to each of the eight pools. "A" can always be identified because it is the only sequence present in plate 1, row 1 and column 1.

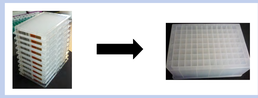


Individual BACs are assembled from sequencing data. The BAC sequences are then combined to complete the genome.

Materials and Methods

BAC Preparation - Individual clones pooled from glycerol plates by column, row and plate

Step 1: Pool all individual clone wells together into Deep Well plate. (ex. A1 from all plates into A1 of Deep Well)

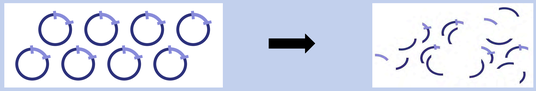


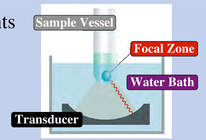
Step 2: Create Row and Column Pools from Deep Block (ex: All wells in Row 1 pooled together)

Multimek- used to pool individual clones together by plate

GSSR - Genome Sequencing Sample Repository: Samples receive ID number for tracking

Shearing - Covaris uses ultrasonic acoustics to shear DNA into 150 base pair (bp) fragments

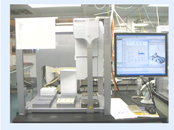




Covaris E210- used to shear DNA into variously sized fragments

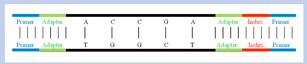
Library Construction - Repair DNA fragments and add Illumina indexing adapters for each library

150bp DNA fragment | Blunt end 5'-phosphorylated | 3' A-base overhangs | Indexing adapters attached to ends



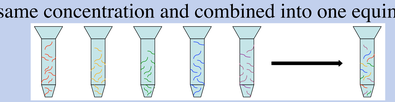
Bravo- liquid handler used during automated library construction (LC) and enriched indexed library pooling

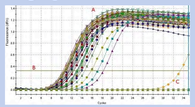
Indexing Enrichment - Addition of synthetic indexing barcodes (or "tags") allows for identification of an individual library. Primers added to ends of the DNA. Polymerase Chain Reaction (PCR) amplifies DNA.

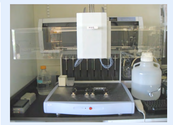


Thermocycler- used during LC enzyme steps and for indexed enrichment

Quantification and Pooling - DNA concentration of libraries is determined using qPCR. All indexed libraries are normalized to the same concentration and combined into one equimolar pool.

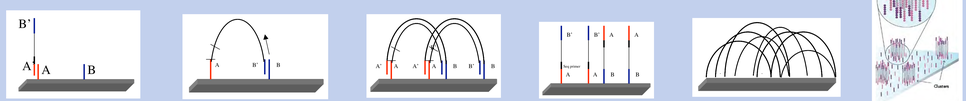






Multiprobe- used to pool individual clones together by column and row

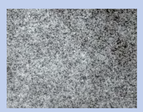
Cluster Amplification - Samples are bound to complementary primers on the surface of the flowcell and amplified to optimal density (200,000 clusters per tile)



Cluster Station- used to hybridize DNA to the flowcell

Illumina Sequencing - 36bp indexing paired end run

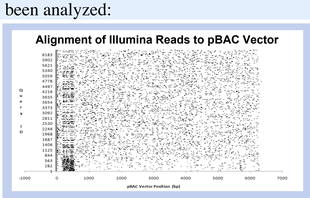
Illumina sequencing uses reversible terminators to add one fluorescently-tagged base each cycle. Each base fluoresces under a different combination of lasers and filters. An optics system detects which base is added. The index is also sequenced to determine which sample each fragment originated from.



Illumina Sequencer- used to analyze base pair composition of DNA molecules

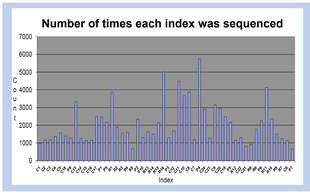
Results and Discussion

The Illumina sequencing run on our guinea pig DNA library has been performed. Data from the sequencing of the pBAC1 vector we used has been analyzed:



This graph shows the Illumina reads assembled to the reference sequence for the pBAC1 vector. Each data point is a 36-bp read.

Data on sequencing of the index tags we used have also been analyzed:



This graph shows that we met our goal in having a ~5-fold difference between the most frequent tag and least frequent tag. The data shown are from one representative tile.

Analysis of the guinea pig DNA is ongoing according to plan below:

Analysis Plan -

- Sort Sequencing Data by Index
- Search for matching ends within other index pools
- Find BAC ends within a pool
- Create an index key of which indices are found in each well
- Assemble individual BACs
- Assemble whole genome

This study shows that combinatorial pooling of indexed BACs is a feasible way to circumvent the problem of short read lengths from Illumina sequencing, and may provide a low cost method for complex genome assembly.

Future Directions

Although "BAC by BAC" sequencing already has proved to be a useful tool for assembly using long reads, a number of process changes must occur before BAC sequencing can be used to its full potential on Next Generation Sequencing technologies. A streamlined messaging system must be developed for tracking samples throughout the process. Additionally, further automation is necessary before BAC sequencing can transition to a large scale production process. With these alterations made, "BAC by BAC" sequencing on Illumina could become the primary method of *de novo* assembly of genomes.

Acknowledgements

I would like to thank my mentors Serenity Banden, Laura Lambiase and Andrew Barry, as well as Ed Kelliher, Adam Navidi and the Technology Development Group. Thank you Megan Rokop, Kate MacSwain, Allison Martino for the internship opportunity; it's been an experience I will never forget.