Exploring Size Selection Methods to Optimize Read Length in 454 Sequencing

Lisa, Danielle McCarthy, and Niall Lennon

Broad Institute Genome Sequencing Platform, Cambridge, MA, USA

Introduction

Background:

Sequencing of complex genomes was originally done using Sanger sequencing. Developed about 30 years ago, Sanger Sequencing can produce up to 384 individual sequence reads per run. Recently, 'next generation' sequencing technologies have replaced Sanger sequencing. 454 sequencing, a next generation method, is able to produce around a million sequence reads per run, resulting in a drastic decrease in the cost of sequencing runs.

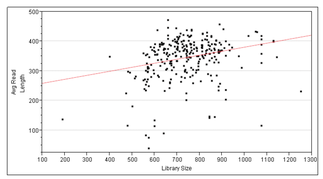
One downfall of newer technologies is a sacrifice in read length. Sanger sequencing produced read lengths of 750-1000 base pairs (bp), whereas new technologies such as 454 can only produce read lengths of up to 500 bp. This creates a challenge for traditional assembly algorithms, which work by utilizing overlapping reads to build a continuous genome sequence.

Goal:

Explore variations of the 454's Library Construction process in order to find conditions that will produce optimal read lengths

Justification:

Our analysis of historical data showed a positive correlation between the sizes of fragments sent to sequencing (library size) and average read length.



$r^2 = 0.07$

Approach:

We chose to test variations of the DNA Shearing and Size Selection steps of the Library Construction protocol. These steps have the greatest impact on library size.

For some conditions, samples went through the entire 454 sequencing process, and actual sequencing data, including average read length, was examined.

For other conditions, electropherograms of samples were analyzed to determine what conditions produce libraries with:

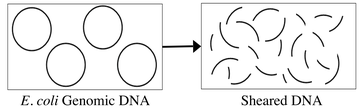
- optimal **yield**
- optimal **selectivity** of fragments between 500 and 800 bp
- optimal **omission** of fragments under 500 bp

Conditions that produce the best selectivity and omission should produce the longest average read lengths.

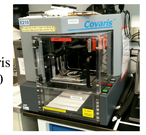
Materials

SHEARING: Using Acoustics to Shear DNA

Covaris: Uses ultrasonic acoustics to shear DNA into fragments of various sizes. Shearing for longer times will produce shorter fragments.



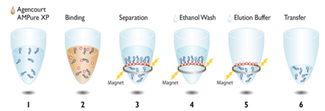
E. coli Genomic DNA → Sheared DNA



Covaris E210

SIZE SELECTION CLEAN-UP: Comparing 4 Methods

✱**Solid-Phase Reversible Immobilization (SPRI)**: Under conditions of high polyethylene glycol (PEG) and salt concentrations, DNA will bind to the surface of carboxyl coated magnetic particles (beads). Longer fragments are preferentially bound. Desired size ranges are selected for by manipulating the concentration of PEG in solution. Adding smaller amounts of PEG is increasingly selective against small fragments. Undesired fragments are discarded with supernatant, and those bound to beads are eluted in buffer or water to produce purified DNA.

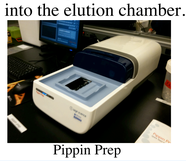


1. DNA fragments of various size
2. Larger fragments bind to SPRI beads
3. Separation of longer bound fragments from unwanted shorter fragments
4. Removal of unwanted fragments and ethanol wash
5. Elution in buffer
6. Transfer to a clean tube

Gel Box

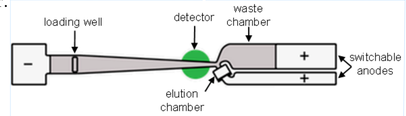
✱**Gel Cut**: Sheared DNA is run on a 2% agarose gel alongside a 100 base pair ladder. The area of the gel containing the desired size range is cut out and DNA is extracted using a QiaQuick Gel Extraction Kit.

✱**Pippin Prep**: Samples are run through the instrument on a gel cassette. Detectors use a ladder to determine the timing of sizes running through the cassette. Anodes switch at the right time to pull the DNA in the desired size range into the elution chamber.



Pippin Prep

✱**Caliper LabChip XT**: Samples are run through the instrument on a chip. Sizes are detected and the instrument utilizes a 'current switch' that drives fragments in the desired size range into the collection well.



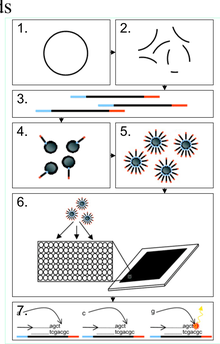
Caliper LabChip XT

ANALYSIS: Utilizing 2 Methods

✱**Agilent 2100 Bioanalyzer**: A 1ul aliquot of each sample is run on a DNA chip. The Agilent software produces an electropherogram and virtual gel image.

DNA Chip

Agilent

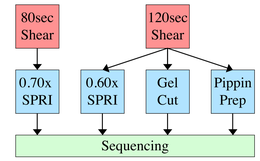


✱**454 Sequencing Run:**

1. Genomic DNA
2. Shearing into fragments
3. Adaptor ligation
4. Immobilization onto DNA capture beads
5. Emulsion PCR amplification
6. Beads and enzymes loaded onto a PicoTiter Plate
7. Sequencing by synthesis

Methods

4 Conditions Analyzed With 454 Sequencing:

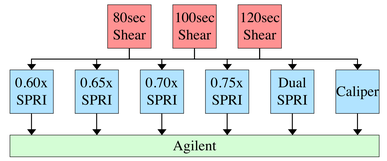


SHEARING

CLEAN-UP

ANALYSIS

18 Conditions Analyzed With Agilent:



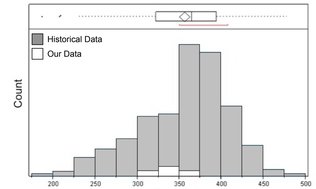
3 shear times
x
6 clean-ups
=
18 conditions

Results

Conditions Analyzed With 454 Sequencing:

Condition	Library Size	Average Read Length
120/Pippin Prep #1	630	354.1
120/Pippin Prep #2	662	353.1
120/Gel Cut #1	675	347.3
120/Gel Cut #2	650	340.1
80/0.70 #2	924	334.1
80/0.70 #1	900	330.4
120/0.60 #1	923	321.9
120/0.60 #2	945	321.1

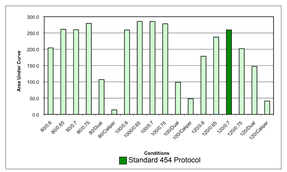
Sequencing results do not show a significant difference between the conditions tested.



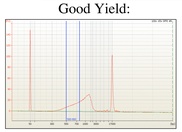
Our sequencing data falls within the normal distribution of past runs.

Conditions Analyzed With Agilent:

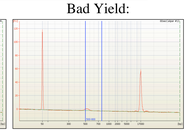
Total **yield** of DNA was found by analyzing the area under the curve on the electropherogram.



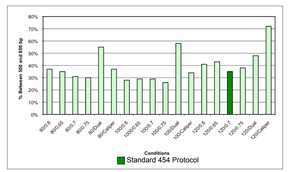
Good Yield:



Bad Yield:

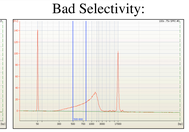


Selectivity was found by calculating the percent of the area under the curve on the electropherogram that is between 500 and 800 bp.

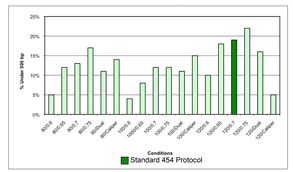


Good Selectivity:

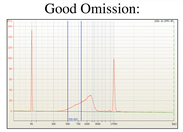
Bad Selectivity:



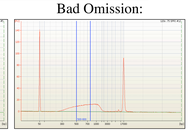
Omission was found by calculating the percent of the area under the curve on the electropherogram that is under 500 bp.



Good Omission:



Bad Omission:



Conclusions

There was no significant difference between the average read lengths of the conditions analyzed by sequencing. This indicates that a gel cut, shorter shearing time, smaller SPRI concentration, or use of the Pippin Prep all produce read lengths similar to those achieved by the standard 454 protocol.

The conditions analyzed with Agilent show that single SPRI protocols provide the best yields. If yield can be sacrificed, shearing for 120 seconds then using the Caliper for clean up will provide better selectivity than the standard 454 protocol. If higher yield is necessary, lower shearing times and a dual SPRI also give better selectivity. Shearing for less time and performing an 0.60x SPRI is better than the standard 454 protocol to eliminate fragments under 500 bp. Presumably, conditions that provide better selectivity and/or omission will lead to longer read lengths.

Future Directions

The conditions analyzed with Agilent should be sequenced to test their impact on average read length.

Historical data should be analyzed further to find correlations between other areas/conditions and average read length.

Acknowledgements

I would like to thank Megan Rokop, Allison Martino, and Rachel Woodruff for making this internship possible. I would also like to thank my mentor Danielle McCarthy, and everyone on the 454 team who were so helpful during this summer.

Image of 454 Sequencer courtesy of www.roche.com.