An Effect of a Shearing Process on the Re-Sequencing of the Arabidopsis thaliana Genome

Introduction

Professor David Guttman is from the Centre for the Analysis of Genome Evolution and Function (CAGEF) at the University of Toronto. His laboratory uses whole genome sequencing and a range of genomic approaches to study the evolution of host specificity and virulence in plant pathogenic bacteria.

Application

- Genomics

Category

- Nucleic Acid Shearing

Oginization

- Centre for the Analysis of Genome Evolution and Function, University of Toronto, Toronto, Canada

Challenges

- Overall performance of DNA nebulization did not meet expectations and included DNA "smears"
- The lab's shift to sonication techniques solved some problems, but resulted in others, including a similar "smear"
- Failure to achieve adequate shearing and coverage through the use of so sonication became prohibitively expensive for the laboratory

Solution

- After acquiring a Covaris S2 System with Adaptive Focused Acoustics® (AFA®) for DNA shearing, immediate improvements in range of fragment size and precision of output fragment size were observed
- Gel "smears" of large to small fragments of DNA from nebulization and sonication now became tight bands of DNA
- Sample bias experienced using the previous methods was eliminated, enabling an additional 20% of the Arabidopsis genome (that had been previously missed) to be readily sequenced

Fragmentation Bias

During the early stages of Next Generation Sequencing, nebulization was the standard and accepted approach for DNA shearing. According to Pauline Wang (CAGEF Laboratory Facility Manager), their group initially employed nebulization to fragment the Arabidopsis thaliana plant genome for sequencing. Over time, they began to suspect there was bias in genome coverage using this approach, largely because of their inability to reproducibly shear samples.

The overall performance of nebulization did not meet expectations [1, 2]. For example, CAGEF observed a smear of fragment size ranges when they ran output DNA gels. As a result, a purification step needed to be added to the DNA preparation process to achieve the desired narrow fragment size range. This additional step, combined with the large sample losses typically experienced with nebulization, resulted in alarmingly low yields. Very little of the desired DNA fragment size remained, following this sample prep process.

Low DNA yields (as well as added processing times resulting from this problem) have been observed and reported by other labs performing Next-Gen or whole genome sequencing [1,2]. Due to their concerns about the potential for bias and low yields, the CAGEF lab switched to sonication using the Bioruptor® (Diagenode, Inc.). The Bioruptor sonicator was initially viewed as an affordable upgrade from nebulization, with the promise of improved shearing performance. Its technology is similar to other ordinary laboratory bath sonicators, in that it uses relatively long wavelengths and unfocused acoustic energy.

The Sonication Experience

CAGEF's shift to sonication brought its own set of challenges. Test shears needed to be performed to try to determine the best settings to achieve the desired DNA, and often the resulting output contained a "smear" with a range of large to small fragments.

While some size selectivity could be achieved by changing the shearing conditions, gaps in coverage (due to shearing biases) were

observed. It remained difficult for CAGEF to attain fragments of the desired size. At a cost of approximately $3,000 per Arabidopsis genome analyzed, any failure to achieve adequate and reproducible shearing and coverage through the use of conventional sonication became prohibitively expensive and prompted a study of the Covaris AFA system.

High variability in the read depth or coverage can be seen for the sample prepared using sonication (blue bars) while relatively uniform and consistent coverage is seen for the sample prepared using Covaris (orange line). Approximately 14 million mapped reads from each method were obtained from paired end, 38 base sequencing performed on an Illumina GAIIx. Total number of reads or read depth summed in 1Kb intervals is shown on the Y-axis. Position in base pairs along the chloroplast genome is shown on the X-axis. The Arabidopsis sequence was obtained from the TAIR9 release and mapped using the BWA aligner. Results were processed with SAMtools and plotted using R.

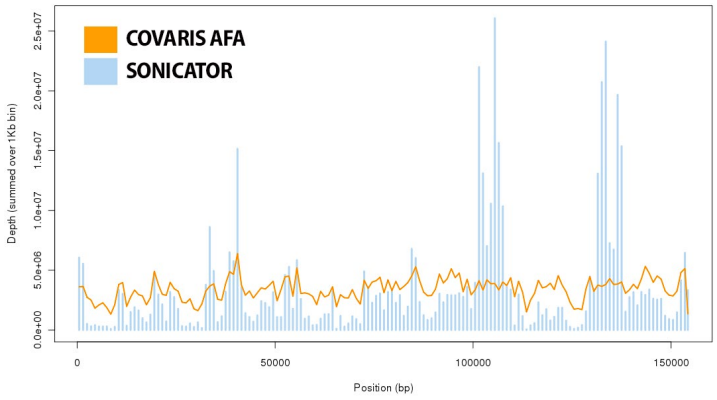

Figure 1: Histogram showing the number of reads mapped at each position across the chloroplast genome of Arabidopsis thaliana summed over 1 Kb intervals.

Analysis

There are several factors which can affect sequence depth and rate of coverage such as PCR amplification, GC/AT percentage, copy number, and shearing efficiency.

PCR Amplification

PCR amplification effectiveness can be impacted by GC content and 'repeats' which tend not to amplify as well as other parts of the genome. While the GAIIx library prep does involve amplification, the methodology has been optimized to not allow amplification to reach the exponential phase in order to avoid just such biases from occurring, making this an unlikely cause of the problem.

GC/AT Percentage

Higher GC content requires a higher melting temperature. Prevalence of GC rich regions of the genome could result in gaps and lower coverage. However, when CAGEF plotted content against the gaps in coverage, no relationship was seen.

Copy Number

The Chloroplast genome has a bacterial origin and is highly conserved, so it will lack copy number variations seen in higher organisms. Therefore, Copy Number Variation (CNV) is the least likely explanation for the problem.

Shearing Efficiency

Of all potential factors, lack of shearing efficiency is the most likely cause of the sequencing coverage problems experienced by CAGEF. This was demonstrated when identical sequencing experiments (performed using DNA fragmented with Covaris AFA technology) were compared to samples processed with the sonicator. As seen in Figure 2 below, under identical conditions the Covaris DNA shows little or no skewing in this experiment.

Sonication shows an exponential distribution skewed towards low coverage. The amount of mapped reads is shown above the graph. Reads were obtained from paired-end, 38 base sequencing performed on an Illumina GAIIx. The frequency of read depth is shown on the Y-axis. The read depth or total number of reads contributing to base calls for each position of the chloroplast is shown on the X-axis. The Arabidopsis sequence was obtained from the TAIR9 release and mapped using the BWA aligner. Results were processed with SAMtools and plotted using R.

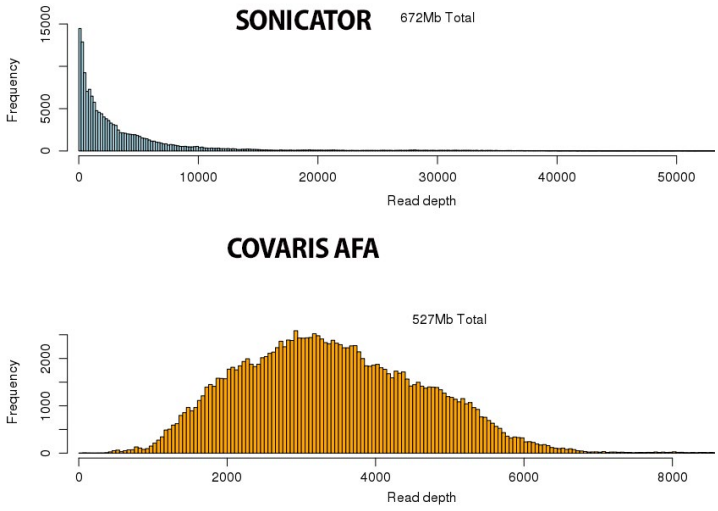

Figure 2: Histogram showing the overall distribution of read depths for each position of the chloroplast genome of Arabidopsis thaliana.

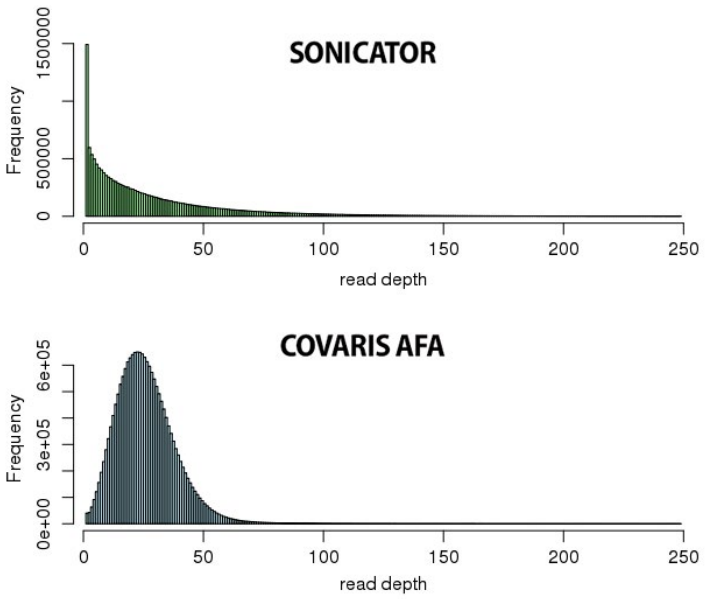

Figure 3: Histograms showing the overall distribution of read depths for each position along chromosome 4 of the Arabidopsis thaliana genome (the same procedure that was applied to Figure 2 was applied here).

The sample prepared using sonication shows an exponential distribution skewed towards low coverage (top graph) while the sample prepared using Covaris shows a normal distribution of read depths (lower graph). The same trends hold as were seen in Figure 2, but for a much larger genomic region. The Arabidopsis sequence was obtained from the TAIR9 release and mapped using the BWA aligner. Results were processed with SAMtools and plotted using R. Approximately 16 million reads were mapped for each method. Reads were obtained from paired-end, 38 base sequencing performed on an Illumina GAIIx. The frequency of each read depth is shown on the Y-axis. T he read depth or total number of reads contributing to base calls for each position across chromosome 4 is shown on the X-axis. Note the almost perfect Gaussian distribution of read depth/frequency for the Covaris treated samples.

Covaris, The Ultimate DNA Shearing Solution

In June of 2010, the CAGEF laboratory acquired a Covaris S2 System, which uses Adaptive Focused Acoustics (AFA) for DNA shearing. With its highly controlled shearing conditions, immediate improvements were observed in both the fragment size range and the precision of output fragment size achieved. What was previously seen on a gel as a "smear" of large to small DNA fragments became a tight band of DNA. Most significantly, the bias that had been observed with previous methods had been eradicated with AFA.

Summary

Considering all factors, CAGEF concluded that Covaris with AFA technology made the difference in coverage. Using the Covaris S2, CAGEF estimates that an additional 20% of the Arabidopsis genome that had been previously missed was now being readily and routinely sequenced (described in Figure 4 next page).
Figure 4: Histograms showing the number of positions across chromosome 4 of Arabidopsis thaliana that have no reads mapped to them. Each bin corresponds to the number of bases with no reads mapped across a 62Kb stretch, indicated by the frequency shown on the Y-axis. The basepair position along chromosome 4 is shown on the X-axis. T he total number of basepairs with no reads mapped (Total gaps) is shown for each method. In this experiment there are many more gaps present in the sequence data obtained from the sample using sonicator (top graph) versus Covaris (lower graph). The Arabidopsis sequence was obtained from the TAIR9 release and mapped using the BWA aligner. Results were processed with SAMtools and plotted using R. Approximately 16 million reads were mapped for each method. Reads were obtained from paired-end, 38 base sequencing performed on an Illumina GAIIx.

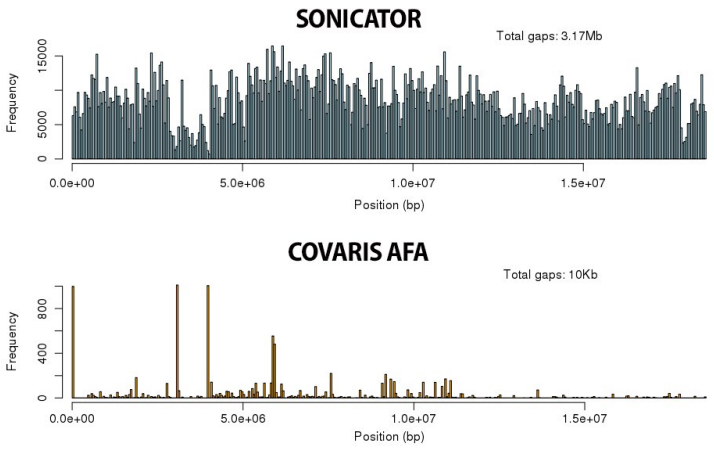

Figure 4: Histograms showing the number of positions across chromosome 4 of Arabidopsis thaliana that have no reads mapped to them.

The Ultimate DNA Shearing Solution: Because Sample Prep Matters:
The Covaris Adaptive Focused Acoustics (AFA) technology is used for a variety of sample prep processes and purposes, from DNA Shearing to Tissue Homogenization. For DNA Shearing, the AFA technology is the "Gold Standard" due to its technological advantages over other sample prep methods, such as sonication and nebulization. Key advantages of AFA technology include reproducibility, versatility (DNA output across a wide size range), uniformity in fragment size distribution and isothermal, non-contact methodology. Highly controlled AFA energy from Covaris delivers unsurpassed shearing consistency and reproducibility.

Acknowledgement

Our thanks for data submitted to us by: Wang, PW, Austin, RS & DS Guttman, University of Toronto Centre for the Analysis of Genome Evolution & Function, unpublished data

1. Quail, M, et al, "A Large Genome Center's Improvements to the Illumina Sequencing System", Nat. Meth., Vol 5, #12, Dec 2008
2. Fisher et al. Genome Biology 2011, 12:R1, "A scalable, fully automated process for construction of sequence-ready human exome targeted capture libraries"

USA: Covaris, Inc. | Tel: +1 781.932.3959 | Fax: +1 781.932.8705 | Email: customerservice@covaris.com
Europe: Covaris Ltd. | Tel: +44 (0)845 872 0100 | Fax: +44 (0)845 384 9160 | Email: emeacustomerservice@covaris.com
Web: www.covaris.com | Applications: applicationsupport@covaris.com | Service and Support: techsupport@covaris.com