Materials Design Institute

Research Summer School

Summer 2009

DNA Sequencing of Plants and Microbes for the Optimization of Biofuel Production

Final Report

August 5, 2009

Krysta Biniek

Mentor:

Lance Green

LA-UR 09-06704

Abstract

The United States currently imports approximately 65% of the petroleum consumed domestically, and disruptions in the supply of this energy source represent a threat to national security and economic growth. The deleterious effect of the burning of fossil fuels on the health of the environment has also recently been under scrutiny. Biofuels derived from cellulosic biomass are a renewable, clean alternative source of energy, but for biofuels to compete successfully with petroleum, the process of turning feedstock into fuel must be optimized. To help achieve this goal, the Department of Energy (DOE) Joint Genome Institute at LANL (JGI-LANL) in the Bioscience Division performs high throughput genome finishing using traditional Sanger sequencing as well as next generation sequencing platforms, including the Roche 454. The Virus Interactions Project, one project in the vast sea of biofuels research, seeks to determine how viruses may influence biofuel production from grass species communities. LANL's role in the project is to isolate and analyze virus-like particles (VLPs) that may represent unidentified viruses in order to evaluate the potential for lesser-known viruses to emerge as significant pathogens when biofuels plantings increase.

1.0 Introduction

The United States is dangerously dependent on oil, much of which comes from foreign soil. In April 2009 the U.S. consumed 18.5 million barrels of petroleum products per day, 65% of which was imported [1]. Due to the fact that much of the imported oil comes from traditionally unstable regions of the world, supply disruptions are a legitimate concern. Because the U.S. relies so heavily on petroleum for transportation and industry, these supply disruptions represent a threat to both the U.S. economy and national security.

The effect of the burning of fossil fuels on the environment has also recently been under considerable scrutiny. Temperatures have increased significantly during the last half of the 20th century. According to the Intergovernmental Panel on Climate Change, this recent increase can most likely be attributed to the vast amount of carbon dioxide and other greenhouse gases that are released into the atmosphere due to human activities, the most significant of which is the burning of fossil fuels [2].

Due to the economic, security, and climate concerns caused by fossil fuels, President Bush, in his 2007 State of the Union address, set forth his “20 in 10” goal: reduce gasoline usage by 20% in the next 10 years. A reduction of 5% will come from gas mileage improvements, while the other 15% will come from alternative fuel sources such as biofuels [3]. Biofuels are a clean, renewable alternative to fossil fuels, but to be competitive with petroleum, the process of converting biomass feedstock into fuel must be optimized. As a result, a substantial amount of research is being conducted on biofuel crops and how to efficiently convert cellulose into ethanol in order to achieve former President Bush’s ambitious plan.

2.0 The Department of Energy Joint Genome Institute

The Department of Energy Joint Genome Institute (DOE-JGI) unites six institutions, including five national labs, to advance genomics in support of DOE missions related to clean energy. The JGI is hugely productive in the area of sequencing. In October 2008 alone, DNA sequence data equivalent to six human genomes was generated [4]. As a member of the JGI, Los Alamos National Lab (LANL), specifically B-6, performs high throughput genome finishing and analysis, using traditional Sanger sequencing as well as the Roche 454 and Illumina Genome Analyzer next generation sequencing platforms. To date LANL has finished nearly 300 genomes.

3.0 Virus Interactions Project

3.1 Introduction

The Virus Interactions Project, proposed by Michigan State University, seeks to determine how viruses may influence biofuel production from grass species communities and to evaluate any spillover influence of biofuels plantings on virus and vector populations in nearby plantings of other susceptible crop species, such as wheat, oats, and maize. Many viruses can cause stunting and yield loss in annual cereals and may have similar effects on some perennial species. LANL's role in the project is to use Roche 454 sequencing to analyze virus-like particles (VLPs) that may represent unidentified viruses in order to determine which specific viruses should be investigated more in depth and to evaluate the potential for lesser-known viruses to emerge as significant pathogens when biofuels plantings increase [5]. Sequencing viruses on the Roche 454 is a new technique that Roche does not yet have a protocol for. Sequencing viruses differs from sequencing bacteria in that the starting material is RNA instead of DNA, and the amount of material is significantly reduced.

3.2 Experimental Procedure

The isolation, amplification, and barcoding of VLPs was performed by the Department of Biochemistry and Molecular Biology at Oklahoma State University. RNA was isolated from *Andropogon gerardii* (big bluestem), *Schizachyrium scoparium* (little bluestem), and *Panicum virgatum* (switchgrass). Differential centrifugation of plant homogenates was used to produce a pellet containing VLPs. Contaminating plant DNA was removed by adding DNase I, which was then deactivated for further processing. Nucleic acid (specifically RNA) from the VLPs was extracted using phenol and diethyl ether. The samples underwent reverse transcription using a primer with a degenerate 3' end to form complimentary DNA (cDNA) and second strands were formed using Sequenase. Products were then amplified using PCR in a 96 well format. Gel electrophoresis was performed to confirm amplification occurred. Five microliters from each amplification were pooled and shipped to LANL.

LANL received the amplified, barcoded DNA fragments and prepared the sample, which was given the code ATB, to be sequenced on the Roche 454. An electrophoresis gel was run to select properly sized fragments. The library was prepared by ligating A and B adaptors onto the end of the fragments and then denaturing the DNA into single strands. The fragments were attached to capture beads and underwent emulsion PCR amplification. The clonally amplified fragments were enriched, denatured, and loaded onto a PicoTiterPlate (PTP) device. The PTP device was loaded into the DNA Sequencer FLX instrument and sequenced. The detailed library preparation, emPCR, and sequencing procedures are available in manuals provided by Roche. The raw data was sent to finishing, where assembly was attempted using 454 Newbler assembly software. The sequence was BLASTed against the NCBI (National Center for Biotechnology Information) nucleotide database.

3.3 Results

3.3.1 Electrophoresis Gel

An electrophoresis gel was run before the library was prepared in order to select properly sized fragments between 500 and 800 base pairs in length. The gel showed, however, that the majority of the fragments were smaller than this optimal size, with an average size of 237 base pairs. Because sequencing viruses on the Roche 454 is a new technique for which there is not yet a protocol, it was important to run the sample to evaluate how the 454 process could be adapted to RNA. Therefore, it was decided that the majority of the fragments would be harvested for further preparation (Figure 1). It was hypothesized that smaller fragments would still yield high quality reads, but the full read length potential of the 454 would not be utilized.

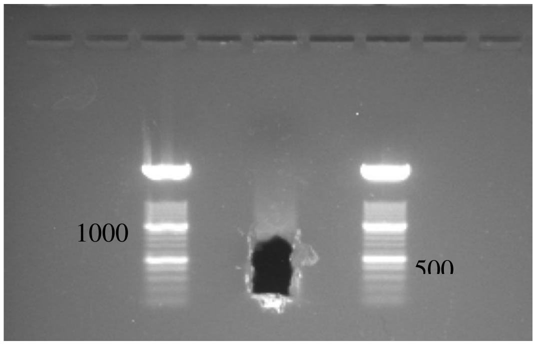

Figure 1: Electrophoresis gel of sample ATB illustrating the large range of fragment sizes that was removed for further sequencing. Bands on either side of hole are ladders with known sizes of DNA. Bright bands are 500 and 1000 base pairs in length.

3.3.2 Sequencing Run

Sample ATB was successfully sequenced on the Roche 454 GS FLX instrument. The results of the run are given in Table 1. The low average length of 182 base pairs was expected from the initial electrophoresis gel. This also explains why the percentage of short reads was higher than usually desired. Although the total number of bases sequenced (37,738,504) is significantly lower than the goal of 60,000,000 to 110,000,000 bases, it was estimated that only 10,000,000 bases were needed to construct the draft assembly due to the small size of viral genomes. Thus, more than enough data was obtained.

Table 1: Summary of 454 sequence data for sample ATB

	ATB Sample	Desired Values
Raw Wells	425,967	-
Key Pass Wells	420,653	-
Passed Filter Wells	207,405	-
% Dot + Mix	17.43	< 20
% Short	33.27	< 20
% Passed Filter	49.31	> 60
Length Average	182	400 - 500
Total Bases	37,738,504	60,000,000 – 110,000,000

3.3.3 Finishing

Assembly was attempted with the 454 Newbler assembly software. Very few reads were assembled, but thirty-four contigs were produced, each consisting of between five and two hundred reads and all between one hundred and five hundred base pairs in length. Thus, only a very small fraction of the data produced from sequencing could be assembled.

3.3.4 BLAST results

BLAST (Basic Local Alignment Search Tool) is a publicly available algorithm offered by the NCBI (National Center for Biotechnology Information) that locates regions of local similarity between sequences. The program compares nucleotide sequences to sequence databases and calculates the statistical significance of the matches. BLAST is commonly used to infer functional and evolutionary relationships between sequences, as well as help identify members of gene families [6].

The thirty-four contigs were each BLASTed against the nucleotide database. The majority of the contigs matched 16S and 23S ribosomal RNA from various kinds of bacteria. Three contigs had no hits in the database, but these contigs consisted of very few reads and were of poor quality. None of the contigs matched viral sequences.

3.4 Discussion

It is not definitively known why no viruses were found in the sample. However, it is hypothesized that the plants used to create the sample harbored a great diversity of species, only a small percentage of which was viruses. Thus, when the RNA was collected and randomly amplified, the viral RNA, if any was present, was drowned out by the much more abundant bacterial ribosomal RNA. Although approximately half of the sample remains, it must be discussed with the collaborator whether the sample should be sequenced again or if the process to isolate viral RNA must be altered.

3.5 Conclusions

Although the ultimate goal of identifying VLPs from biofuel crops was not completed, the new technique of sequencing RNA on the Roche 454 did prove to be successful. In order to complete the project, the hypothesis that bacterial RNA is overwhelmingly more abundant and thus drowning out the viral RNA must be verified and, if proven correct, a method for isolating viral RNA will need to be devised. This will allow for the evaluation of which lesser-known pathogens might become a greater threat when biofuel crop plantings increase. One project in the vast body of research that is being conducted on biofuels by the DOE-JGI, the Virus Interactions Project will aid the process of making biofuels a viable alternative to petroleum.

WORKS CITED

[1] "U.S. Imports by Country of Origin," *U.S. Total Crude Oil and Products Imports*, Energy Information Administration, http://www.eia.doe.gov/oil_gas/petroleum/info_glance/petroleum.html, Accessed July 14, 2009.

[2] S. Solomon, D. Qin, M. Manning, Z. Chen, M. Marquis, K.B. Averyt, M.Tignor and H.L. Miller (eds.), IPCC, 2007: *Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

[3] "DOE Mission Focus: Biofuels," Genomics:GTL, http://genomicsgtl.energy.gov/biofuels, Accessed July 14, 2009.

[4] "Bioenergy at DOE JGI," DOE Joint Genome Institute, www.jgi.doe.gov, Accessed July 14, 2009.

[5] "Virus Interactions with Biofuel Host Crops," Statement of Work for JGI Sequencing Projects.

[6] BLAST: Basic Local Alignment Search Tool, http://blast.ncbi.nlm.nih.gov/Blast.cgi.