sequencing_project

Clint Valentine

October 6, 2016

Background

The Illumina HiSeq 2500 High Output platform will be utilized to sequence the 430 MB genome of our target organism to 100x coverage. The sequencing will be performed at the Harvard Bauer Core. After sequencing, data analysis will begin with quality control and end with assembly of the sequencing information into a novel assembled genome of our organism. The estimated expense of this project is on the following page in Table 1[1]. Three libraries will be prepared following these protocols:

Paired-end Library

Our lab will extract genomic DNA (gDNA) from our target organism and then proceed with an enthanol precipitation to purify the sample.

Library preparation, as described here, will be carried out by the personnel at the Harvard Bauer Core. The Bauer core will accept our sample and then use a Covaris sonicator to fragment the DNA to a mean size of 200 bp. All intial sonication protocols will be quality verified with one run *per* sample on the Agilent TapeStation. After passing the first QC check, the library will be repaired and ligated with Illumina sequencing barcodes. The molecules in the library, at this point, will be approximately 300 bp in length. The library will then be size selected using Aline magnetic beads and quality verified with the Agilent 2100 Bioanalyzer.

The Bauer Core will sequence this library using the Illumina HiSeq 2500 with High Ouput setting. The Bauer Core is currently using v4 Illumina reagents and has estimates that one sample should fill at least 70% of a flow cell lane. Each flow cell, with 2 lanes, can thus hold a maximum of 2 samples[2]. We intend to sequence the paired-end library using one lane of the flow cell.

Mate-Pair Libraries

A portion of each sample will be submitted to the Bauer Core for two mate-pair libraries. The library preparation protocol will be carried out by the personnel at the Bauer Core and will follow a similar strategy to that of the paired-end library. gDNA will be sonicated to a quality verified 5 kb and 10 kb for the two mate-pair libraries. As we are sequencing a novel organism of large size we hope that, minimally, two mate-pair libraries, one with a maximum insert size, will aid in the initial assembly of the genome.

Briefly, the mate-pair libraries will be circularized and sonicated. The fragments containing the original join of the molecules will be selected for and Illumina barcodes will then be ligated. The mate-pair libraries will be sequenced using the same pipeline as the above paired-end reads since they are effectively similar in size and genome coverage. These two libraries will each use a lane to complete a flow cell.

Assembly of Genome

The assembler `ALLPATHS-LG` was chosen to assemble our target organism's genome. As of 2012 the Genome Assembly Gold-Standard Evalutations (GAGE) has concluded that for a genome size of 88 Mb (the nearest to our target organism's genome size) `ALLPATHS-LG` was the most successful with a resultant 34 scaffolds with an N50 score of 3,192[3]. The assembler relies on advanced computational techniques and methods for developing genome graphs to assemble medium to large scale genomes[4].

Table 1: Cost Estimates

Line Item	Cost per Sample	Quantity	Total Cost
DNA Library Preparation Includes QC —-	$304	3	$912
Illumina HiSeq 2500 High Output v4 Chemistry 2 x 125	$2,631	3	$7,893
		Total	$8,805

References

[1] "Fees." *Bauer Core Facility.* N.p., n.d. Web. 06 Oct. 2016.

[2] "Sequencing Coverage Calculator." *Sequencing Support - Coverage Calculator.* Illumina, n.d. Web. 06 Oct. 2016.

[3] "GAGE Assembly Results." *GAGE Assembly Results.* N.p., n.d. Web. 07 Oct. 2016.

[4] Gnerre, S., I. Maccallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe. "High-quality Draft Assemblies of Mammalian Genomes from Massively Parallel Sequence Data." *Proceedings of the National Academy of Sciences* 108.4 (2010): 1513-518. Web.