

High Throughput Genomic Sequencing of Higher Eukaryotes from Ultra-low Starting Concentrations

Viacheslav Y. Fofanov¹, Kent F. McCue², Maria Shin¹, Mark Rojas¹, Jeremy Morrell¹, Roger Kevin Stevens¹, Heather Koshinsky¹
¹ Eureka Genomics Corp, ² USDA Agricultural Research Service



Abstract

High Throughput Sequencing (HTS) technologies, such as Illumina's Genome Analyzer, produce gigabases of sequence data in a single run at a relatively low per-base cost. However, a significant limiting factor is the amount of starting material required; Illumina's Paired-End Sample Prep Guide (February 2010 publication) suggests using 1- 5 ug of isolated DNA for library preparation. Isolation of this amount of genomic material from unique (often irreplaceable) clinical, environmental, or forensic samples is not always feasible.

Eureka Genomics has developed a sample preparation approach, previously tested on bacteria, that allows successful library preparation and sequence data generation of samples with as little as 1 ng isolated genomic DNA. This 1 ng sample is NOT subject to whole genome or other amplification approaches, but used directly to generate a library for sequencing.

Methods

To illustrate the feasibility of our approach, paired end libraries were prepared with 1 ug, with 10 ng, and with 1 ng of starting genomic DNA isolated from *Solanum bulbocastanum*. Genomic DNAs were fragmented, end repaired, adenylated on 3' ends, adapter ligated, purified and enriched. Each of these libraries was used to generate 36 base long, single-end sequence reads in one lane of a flow cell. The libraries were evaluated on the basis of quantity and quality of reads, proportion of reads mapping to known scaffolds of *S. bulbocastanum*, and finally the length and quality of assembled contigs (fragments). Publicly available (velvet) and proprietary (Eureka Genomics) sequence analysis tools were used to perform various types of bioinformatic analysis (Table 1).

Table 1

Table 1. Analysis of sequence reads generated from 1 ug, 10 ng, and 1 ng starting genomic DNA from *S. bulbocastanum*.

	1 ug	10 ng	1 ng
ng of library generated	127.7	76.5	87.4
Instrument used	GAIi	GAIix	GAIix
Mb of sequence data generated	521	513	570
Average quality score of last cycle	23	28	28
% GC	38.50%	40.31%	40.03%
% unique reads	88.4%	47.23%	38.91%
% total reads mapped to known scaffolds of <i>S. bulbocastanum</i> with up to 1 mismatch	14.1%	15.5%	14.9%
% total reads mapped to 37.3 build of human genome	0.62%	1.68%	4.45%
% of SNPs in contigs mapped to known scaffolds of <i>S. bulbocastanum</i>	2.71%	2.69%	2.66%
# of contigs over 100 b	3889	2643	2915
Average contig length	193 b	244 b	243 b
Median contig length	142 b	159 b	158 b
Max contig length	2399 b	5100 b	8985 b

Discussion

The sequencing reads from each sample were analyzed using Eureka Genomics' QC pipeline to obtain summary statistics. These statistics include total number of reads generated, number of unique reads, and average GC content of reads. The resulting analysis revealed that the overall quality of reads was not significantly affected by the changes in the amount of starting genomic material. The number of unique reads in the 10 ng and 1 ng samples were 2 x less than the number of unique reads in the 10 ug sample. This means that the samples with less input DNA effectively generated half the unique sequence information compared to the sequence information generated with the 1 ug sample. In theory, this could be compensated for by generating more sequence data.

To make sure that the samples are still dominated by *S. bulbocastanum* and that no significant contamination occurred, the sequencing reads were mapped to known scaffolds of *S. bulbocastanum* (98Mbases, app 9% of haploid genome size) with up to 1 mismatch.

Results

The proportion of reads mapped to the reference was consistent in all three samples; 14.1% (1 ug sample) to 15.5% (10 ng sample). Despite the reduced number of unique reads in the 10 ng and 1 ng sample, the proportion of reads mapped to *S. bulbocastanum* remained comparable to the 1 ug sample. This suggests that distribution of reads across the reference was not drastically changed. The proportion of reads mapped to the 37.3 build of human genome increased in the 10 ng (1.67%) and 1 ng (4.45%) samples. This increase may be attributed to contamination by human genomic material and/or it may be *S. bulbocastanum* reads that map to the human genome.

The sequencing reads were also used to assemble contigs. The overall lengths of assemblies between samples were comparable. The max contig length increased in the 1 ng compared to the 1 ug sample. For each starting amount, the overall coverage of the *S. bulbocastanum* genome is less than 1x and thus the difference in max contig length likely reflects random chance of connecting reads and not an underlying bias. Finally the assembled contigs were mapped to known scaffolds of *S. bulbocastanum* (98Mbases). The proportion of the contigs identified as SNPs in all three samples was comparable.

Conclusion

Sequence data generation from 1 ng of unamplified starting genomic material is feasible and yields sequence reads of similar quality and utility compared to sequence reads generated from 1 ug of starting genomic material.

Contact information

Heather Koshinsky, Ph.D., CSO
heather@eurekagenomics.com
 510-299-9157
 Karen Roberts, Director of Sales and Marketing
karen@eurekagenomics.com
 415-990-7313