



High Throughput Sequencing of Ultra-low Genomic Sample Concentrations: *Pseudomonas aeruginosa*

Viacheslav Y. Fofanov, Maria Shin, Mark Rojas, Jeremy Morrell, Roger Kevin Stevens, Heather Koshinsky

Eureka Genomics Corp, www.eurekagenomics.com



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 library preparation approach, also tested on a higher eukaryote, 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, and with 1 ng of starting genomic DNA isolated from *Pseudomonas aeruginosa*. Genomic DNAs were fragmented, end repaired, adenylated on 3' ends, adapter ligated, purified and PCR enriched. Each of these libraries was used to generate 51 bases long, paired end sequence reads in one lane of a flow cell. The libraries were evaluated on the basis of quantity and quality of reads, mapping and SNP detection characteristics, and finally the length and quality of assembled contigs (fragments). Publicly available (Velvet) and proprietary (Eureka Genomics) sequence analysis tools were used to perform these various types of bioinformatic analysis.

Analysis of sequence reads generated from 1 ug and 1 ng starting genomic DNA from *P. aeruginosa*

		1 ug	1 ng
Sequencing read generation	Amount of library generated	170ng @ 30uM	29ng @ 5.4nM
	Amount of sequenced data (Mb)	1,297	1,110
	# of read pairs produced	12,970,159	10,889,356
	% unique reads	53	39
	% reads mapped to human (up to 3MM)	0.05	13.23
Mapping and SNP detection	Proportion of Ns	0.162%	0.052%
	% reads mapped to <i>Pseudomonas</i> (up to 3MM)	93.4	63.8
	Average coverage	180X	107X
	Total # of SNPs identified	37	38
Assembly	# of scaffolds (range, b)	140 (203 – 427,204)	265 (201 – 298,350)
	Average scaffold length (b)	44,147	23,450
	Median scaffold length (b)	13,022	6,044
	Total size of scaffolds (b)	6,180,624	6,214,363
	Total # (%) of errors (ins/del/sub) in scaffolds	2,253 (0.04%)	20,694 (0.34%)

Discussion

The feasibility of our library preparation approach was evaluated in three general ways: (1) sequencing read generation quality and quantity, (2) utility of sequencing reads in mapping and SNP identification applications, and (3) utility of sequencing reads in de-novo assembly applications.

Sequencing read generation. The ability of 1 ng library prep to generate enough genomic material for sequencing was evaluated on the basis of total number of paired-end sequencing reads, proportion of unique reads, Illumina quality scores for first and last bases of the read, level of contamination and proportion of N bases. While the total number of sequencing reads generated for each sample was similar, fewer of the reads from the 1 ng sample were unique. This likely indicates over-representation of some reads, lower proportion of Ns, and/or possible contamination with human genomic material in the 1 ng sample compared to the 1 ug sample. The quality scores for both samples were comparable (30-35 for first base and 15-20 for last base on average).

Mapping and SNP detection. Sequencing reads from both samples were mapped to the 6.3 Mb reference genome – *Pseudomonas aeruginosa* PAO1 (NC_002516). Reads from both samples covered > 99.999% of the reference genome, with the 1 ug sample achieving 180X coverage and 1 ng sample achieving 107X coverage. In the 1 ug sample, 30% more reads mapped to the reference genome compared to the 1 ng sample. The SNP profile between the two sample was not drastically altered with 34 concordant SNPs. Only one (1) False Positive SNP was identified in the 1 ng sample.

Assembly The 1 ug sample generated better scaffolds than the 1 ng sample in terms of total number of scaffolds, average, median, longer scaffold sizes and scaffold errors. Scaffolds were mapped to the reference genome to identify errors (insertion / deletions / substitutions) in the assembly and, assemblies from either sample had scaffold accuracy > 99.5%. It is unclear whether the quality of the assembly was more affected by increased genome coverage for 1 ug sample (180X for 1 ug vs 107X for 1 ng) or by the contaminating sequencing reads in the 1 ng sample.

Summary

Use of 1 ng sample for read generation
Total number of sequencing reads NOT significantly affected by low starting material.
Higher level of contamination detected in the 1 ng sample.
Smaller proportion of Ns detected in the 1 ng sample.

Use of 1 ng sample for SNP detection
Proportion of the genome covered was NOT affected.
Average coverage was 40% lower in the 1 ng sample.
SNP detection was similar (only 1 extra false positive SNP in the 1 ng sample).

Use of 1 ng sample for de-novo assembly
The number of scaffolds of 200+ bases was roughly doubled in the 1 ng sample.
While the accuracy of the assembly was higher in the 1 ug sample, the accuracy of the assembly in the 1 ng sample was greater than 99.5%.

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