Title

In-silico Methods for Troubleshooting Genomic Shotgun Data

Authors
Goltsman, Eugene
Singan, Vasanth
Trong, Stephan
et al.

Publication Date
2007-03-01

LBNL-62519 Abs UCRL-ABS-228701

In-silico Methods for Troubleshooting Genomic Shotgun Data

Eugene Goltsman, Vasanth Singan, Stephan Trong, Alex Copeland, Alla Lapidus

End-sequencing of shotgun libraries of small genomic inserts is today the most popular approach to Whole Genome Sequencing (WGS). Irregularities in WGS datasets present assembly problems that are expensive and time-consuming to solve, with cloning bias, contamination and long repeats posing the biggest challenges. Shotgun assembly data exhibit well recognizable patterns that follow certain statistical models, and deviations from these models usually stem from flaws and anomalies in the input data, which in turn reflect problems in the cloning protocol, chemistries, or the DNA being sequenced. We developed several statistical and bioinformatic methods for detecting cloning bias, DNA contamination and high repeat content at early stages of the WGS project. These methods are based on analyses of i) depth of coverage distributions, ii) dynamics of iterarative assemblies and iii) GC profiles of real and simulated shotgun datasets. We identify and describe relationships between read coverage and the Poisson function and demonstrate ways to routinely identify cloning bias and contamination through these relationships. Identifying abnormal patterns in the dataset's GC profile at various levels (s.a. genome, library, plate) provided a convenient method for catching suspected contamination. Routine automated application is also discussed.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. DE-AC52-06NA25396.