Identification and Resolution of Repetitive Regions in Microbial Genomes

Bradley Toms, Yasmin Mohamoud, Diana Radune, Heather Forberger, and Nadia Fedorova
Finishing Group at J. Craig Venter Institute, Rockville, MD, USA

Abstract

There are clear theoretical reasons and many well-documented examples which show that repetitive sequences are essential for genome function. Repetitive signals are necessary to regulate expression of coding and non-coding sequences. In prokaryotes, repetitive elements are often associated with transposable elements. In bacteria, these transposable elements are often DNA in nature. Another category of repeats found in bacteria are those that are part of the genome and are not associated with transposable elements. These repeats are often referred to as repetitive DNA. They are found in a wide variety of bacteria and are often associated with genome plasticity. In this poster, we will discuss the process by which repetitive regions are resolved in the microbial genome. We will also discuss the tools and techniques used to resolve these regions.

Introduction

In closure, a repetitive region is defined as a region that has high sequence similarity with another region in the genome. There are two general types of repeats: low (<70%) identity repeats and high (>97%) identity repeats. The high identity repeats can be further classified as: large repeats (>400 bp) or small repeats (<400 bp).

During the microbial random shotgun process, thousands of random sequences are assembled into contigs using CELERA ASSEMBLER program. However, the initial assembly process may fail to assemble the repeats correctly. This may lead to gaps, misassembled contigs, and/or collapsed repeats. The assembler considers the sequence similarity and the clone size constraints; however in many instances it has difficulty resolving large repeats. This may lead to gaps, misassembled contigs, and/or collapsed repeats, which leaves these regions to be resolved and finished manually. Even the novel Pyrosequencing technology using 454 20/364 machines does not help in the resolution of repetitive regions because the generated reads are too short. In this poster, we will discuss the process by which the JCVI microbial finishing group identifies, confirms, and sequences repeats. We will also display many interesting examples of large difficult repeats, which we have successfully resolved.

Uncharacteristically High Coverage in Repetitive Regions May be a Sign of Misassembled

Presence of Non Unique Unitigs (NUUs) Suggest that the Region Needs to be Examinied Closely for Assembly errors

A New Computational Tool that May Help in Resolving Collapsed Repeats with Less Laboratory Work

The closure team verifies that each repeat in the genome is assembled correctly. This is a difficult and labor intensive process and there are several clues which lead us to believe that a repeat is misassembled. These clues include: Sequence discrepancies seen during the editing of sequence reads, clone linking information around the repetitive region, abnormal coverage depth, and presence of NUUs.

Collapsd Tandem Repeat in Bdellovibrio bivora

This is the collapsed tandem repeat area as seen in Assembly Viewer. The repeat wasn’t identified as a repeat because it was not unique in the genome. The repeat region was solved by reassembling the assembly with free mates and outputting a new assembly containing all the sequences to be included in the analysis. This new assembly contained 100% of the repeat.

In the graph above, green shows a collapsed tandem repeat with assembly 135. Restriction digest of a clone spanning the area confirmed that we were missing about 2300 bases. The area was resolved by reassembling the assembly with free mates and outputting a new assembly containing all the sequences to be included in the analysis. This new assembly contained 100% of the repeat.

References

5. GeneRacer Human Kit: A 5’-rapid amplification of cDNA ends (RACE) method for generating products that are 5’-rapidly amplifiable. PNAS 93:1716-1720, 1996.