FF00023

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

J. Craig Venter INSTITUTE

Abstract

There are clear theoretical reasons and many well-documented examples which show that repetitive regions are essential for genome function[1]. Repetitive signals are necessary to regulate expression of coding sequences and to organize functions essential for accurate genome replication[1]. Additionally, repeat elements have an evolutionarily significant role in genome architecture and reorganization[1]. Studies done on prokaryote genomes that contain direct repeats suggest that recombination between direct repeats is a widely conserved mechanism to promote genome diversification[2]. Due to the functional importance of these regions to the organism, the JCVI finishing group dedicates a lot of resources to accurately assembling of repetitive areas in the genome. During the microbial random shotgun process, thousands of random sequences are assembled into contigs using Celera Assembler program. The assembler considers the sequence similarities 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 (GS 20) 454 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.

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 (<97%) identity repeats and high (>97%) identity repeats. The high identity repeats can be further classified as: large repeats (>600 bp) or small repeats (<600 bp).

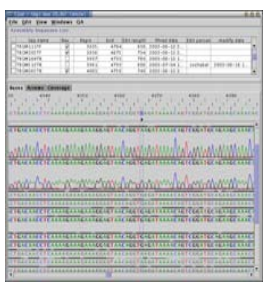
During the microbial random shotgun process, thousands of random sequences are assembled into contigs using CELERA ASSEMBLER program[3]. However, the initial assembly process may fail to assemble the repeats correctly. This may lead to gaps, misassembled contigs, and/or collapsed repeats. The assembler considers the sequence similarities and the clone size constraints. If the sequence falls within the repeat or is a low quality sequence, the repeat may not be assembled correctly.

After the initial CA assembly, repeats are detected using a computational tool called *repeatFinder*[4]. *repeatFinder* finds repeats in the genome or in a list of contigs. It analyzes the genome contigs for repeats greater than 200bp and then groups these repeats into classes. The results are uploaded to the asm_feature table in the genome database and can be viewed graphically using TIGR developed Closure Assembly Viewer tool and textually using the Feature page website.

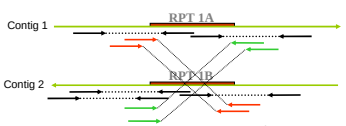
The closure team verifies that each repeat in the genome is assembled correctly. This is a difficult and labor intensive process and there are several clues which lead us to believe that a repeat is misassembled. These clues include: Sequence discrepancies seen during the editing of sequence reads, clone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

Ambiguous Base Calls May be an Indication of Misassembled Repeats

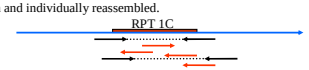
Discrepancies between the base calls of several of the underlying clones may be an indication that there is a collapsed repeat in this region. The consensus of the ambiguous region may include any of the following IUB codes; k, m, s, r, w, y, n.



Linking Clones that Span the Repeat from One Unique End to Another Are Vital to Confirm the Integrity of the Region

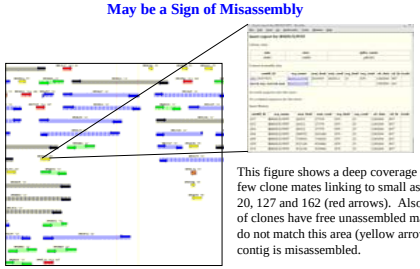


Here both repeats RPT1A and RPT1B are not spanned from one unique end to another. There are misassembled linking clones that start outside the repeat. To resolve this area both contigs will need to be manually broken and individually reassembled.



In this example repeat RPT 1C is spanned by few linking clones. However, the rest of the underlying clones are unlinked. The clone mates are found in other assemblies, or they failed to sequence. This may lead to incorrect consensus sequence.

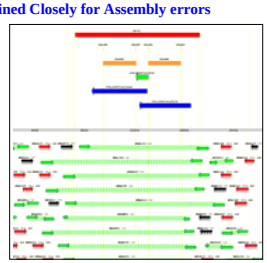
Uncharacteristically High Coverage in Repetitive Regions May be a Sign of Misassembly



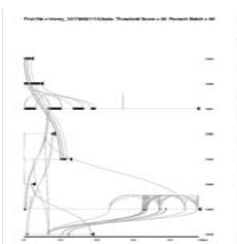
This figure shows a deep coverage area with few clone mates linking to small assemblies 20, 127 and 162 (red arrows). Also a number of clones have free unassembled mates that do not match this area (yellow arrows). This contig is misassembled.

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

NUU (blue arrows) are low-scoring unitigs that could be placed in multiple locations. The CELERA scaffolder notes where they might go but keeps them separate. NUUs may contain reads from collapsed repeats[3].



printrepeats is a Tool that Can be Very Helpful in Identifying Repeats and their Orientation

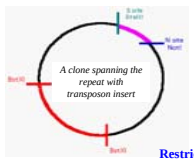


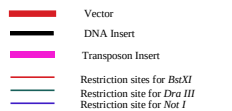
This tool will analyze a set of assemblies for repeats and output a list of repeat coordinates and percentages as well as a graphical representation of the repeats. The user must provide a multifasta file containing all the sequences to be included in the analysis.

Tandem Repeat in *Streptococcus*

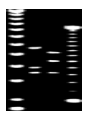
Tandem repeats are areas of a genome where multiple copies of a similar sequence are in close proximity to each other. Often times these areas are large in size spanning several kilobases where there is no unique sequence between copies. There are also cases where repeat units are in tandem and inverted in orientation. Repeat areas of this size and nature, especially those larger than a clone size, present a challenge in the finishing process because of the difficulties in assembling the sequences correctly. Conventional sequencing techniques and PCR are ineffective due to the repetitive nature of the areas. With the aid of *repeatFinder* software[4], restriction digest mapping, PCR and transposon libraries, we have been successful in identifying the true DNA sequence of all these regions. The following is an example of a tandem repeat region in *Streprococcus*.

- Repetitive region consists of 13 full repeat units (237bp in length each) and two truncated units at each end.
- *repeatFinder* represents this area of the genome smaller than it is.
- Red and black arrows represent one of the large insert clones that span the area.
- Due to 100% similarity between repeat units, this region is impossible to walk because any primer designed inside the region will anneal in a number of places. In addition, random sequences can not be properly assembled by Celera Assembler Software (CA).
- Random insertion of transposable elements (transposons) is helpful in this case because it introduces unique priming sites. Transposon insertion was performed using pGPS Transposon Kit (*NEBiolabs*)[5].
- Transposon primed sequences fail to assemble correctly because of absolute identity of the repeat units.
- Restriction digest mapping of large insert shotgun clones containing transposon insertions provided an accurate picture of the repetitive area.
- Restriction sites were chosen and the spanning clones were checked for undesirable cutting sites using DNA Strider[6].
- Double digestion was performed first with *BstX* I to cut out the insert and then with *Not I* or *Dra III* to cut one side of transposon insert.





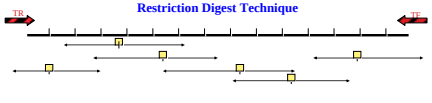
Restriction Digest Results



1 - 1Kb DNA Ladder
2 - Not I digestion
3 - Dra III digestion
4 - 500bp DNA Ladder

Using the resulting restriction digest data, the locations of transposon insertions were elucidated. Transposon primed sequences were assembled using TIGR Assembler[7].

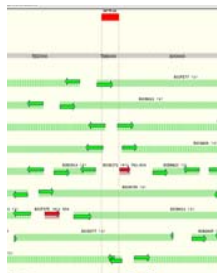
Map of Transposon Primed Sequences Obtained Using Restriction Digest Technique



References

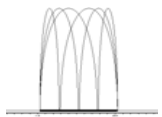
1. Shapiro J., et al. *Why Repetitive DNA is Essential to Genome Function*. Biological Reviews. 80: 227-250. 2005.
2. Aras R. A., et al. *Extensive Repetitive DNA Facilitates Prokaryotic Genome Plasticity*. PNAS. 100(23): 13579-13584. 2003.
3. Myers et al. *A Whole-Genome Assembly of Drosophila*. Science. 287: 2196-2204. 2000.
4. Kurtz et al. *REPuter: The manifold Application of repeat Analysis on a Genomic Scale*. Nucleic Acids Res. 29(22): 4633-4642. 2001.
5. GPS-1 Genome priming System Instruction manual. New England Biolabs. Cat# E7100S.
4. Christian Marck. *'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers*. Nucleic Acids Research. 16(5): 1829-1836. 1988.
5. Sutton G.G., et al. *TIGR Assembler: A new tool for assembling large shotgun sequencing projects*. Genome Science and Technology. 1(1): 9-19. 1995.
6. *Contig Stitching Tool*. Mike Schatz from CBCB at University of Maryland. 2006. Under development.

Collapsed Tandem Repeat in *Babesia bovis*



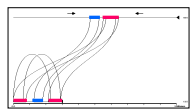
An area that contained <10bp 1X coverage was discovered. Attempts to resolve the 1X region by direct sequencing off a clone failed.

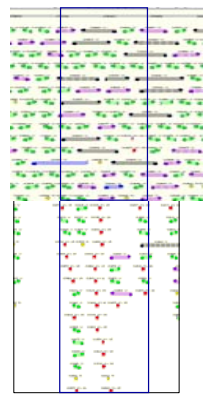
Some clones with free, unassembled mates were discovered in this region. A large clone, BSCIG66, that spanned the area was selected and transposed. The resolved repeat has 4 Copies of a ~250bp sequence. This can be seen in the *printrepeat* graphical output shown below.



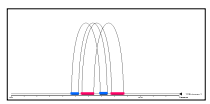
Unflagged Repetitive Region in *Colwellia sp. MT41*

This is the collapsed area as seen in Assembly Viewer. *repeatFinder* did not flag this area as a repeat[4]. Few single base ambiguities were found on the sequence level, this was the first hint. The second was deep coverage with a lot of red mates linking to a small assembly 135. This assembly did not grasta 100% to this area.

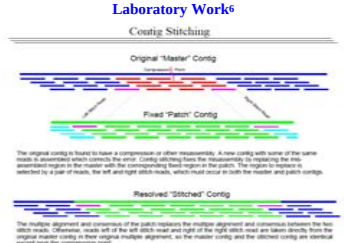




In the graphic above, *printrepeats* shows a collapsed tandem repeat with assembly 135. Restriction digest of a clone spanning the area confirmed that we were missing about 2300 bases. The area was resolved by reassembling the assembly with free mates and assembly 135 using TIGR Assembler[7]. A picture of the correctly assembled region is seen in the display below.



A New Computational Tool that May Help in Resolving Collapsed Repeats with Less Laboratory Work[6]



Using Contig Stitching a tandem repeat in *Xanthomonas oryzae pv. oryzicola BLS256* was resolved by local reassembly at higher stringency[6]. The spanning clone was 3725 bases in the collapsed region. When resolved the clone became 4890 bases. On the left is a view of the collapsed area in both assembly viewer and *printrepeats*. On the right is the resolved version of this repeat.