A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies

Contact: Steven D. Brown (brownsd@ornl.gov), (865) 576-2368
Funding Source: BioEnergy Science Center & Plant-Microbe Interfaces SFA, DOE Office of BER, Genomic Sciences Program

Background

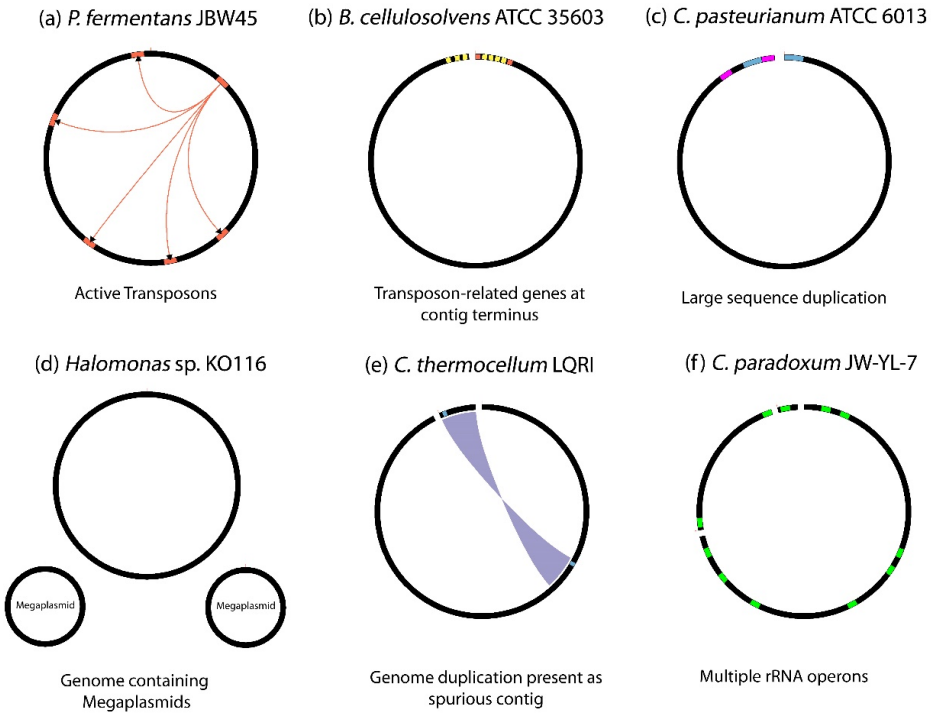
- Information is lacking on the nature of unassembled DNA regions or gaps within unfinished PacBio assemblies, necessitating a systematic evaluation of draft, near-finished and finished genome assemblies to reveal the features and properties of the unassembled DNA regions from both PacBio and Illumina platforms.

Science

- This study characterized regions of DNA which remained unassembled by both PacBio and Illumina sequencing technologies for seven bacterial genomes.
- Our hypothesis that strong secondary DNA structures blocked DNA polymerases and contributed to gap sequences was not accepted. PacBio assemblies had few limitations overall and gaps were explained as cumulative effect of lower than average sequence coverage and repetitive sequences at contig termini.

Significance

- This study offers insights into the nature of gaps associated with Illumina and PacBio assemblies of microbial genomes, describes bioinformatics and manual steps for assembly improvement and underlines the importance of post-assembly polishing steps for genome refinement. .
- The targeted genome finishing approach and systematic evaluation of the unassembled DNA will be useful for others looking to close, finish and polish microbial genome sequences.



Biological features with potential to interfere with the assembly process. (a) Active transposon elements (b) repetitive transposon sequences at contig termini (c) large sequence duplications (d) megaplsmids (e) genome duplication assembled as spurious (f) multiple rRNA operons copies.

Utturkar, S. M. *et al.*, "A case study into microbial genome assembly gap sequences and finishing strategies" *Front. Microbiol. (*2017) 8:1272.

U.S. DEPARTMENT OF ENERGY | Office of Science

A Case Study into Microbial Genome Assembly Gap Sequences and Finishing Strategies

Sagar M. Utturkar[1], Dawn M. Klingeman[2,3], Richard A. Hurt Jr[2], and Steven D. Brown[1,2,3]

[1]Graduate School of Genome Science and Technology, University of Tennessee, Knoxville, TN; [2]Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN; [3]Bioenergy Science Center, Oak Ridge, TN.

Abstract

This study characterized regions of DNA which remained unassembled by either PacBio and Illumina sequencing technologies for seven bacterial genomes. Two genomes were manually finished using bioinformatics and PCR/Sanger sequencing approaches and regions not assembled by automated software were analyzed. Gaps present within Illumina assemblies mostly correspond to repetitive DNA regions such as multiple rRNA operon sequences. PacBio gap sequences were evaluated for several properties such as GC content, read coverage, gap length, ability to form strong secondary structures, and corresponding annotations. Our hypothesis that strong secondary DNA structures blocked DNA polymerases and contributed to gap sequences was not accepted. PacBio assemblies had few limitations overall and gaps were explained as cumulative effect of lower than average sequence coverage and repetitive sequences at contig termini. An important aspect of the present study is the compilation of biological features that interfered with assembly and included active transposons, multiple plasmid sequences, phage DNA integration, and large sequence duplication. Our targeted genome finishing approach and systematic evaluation of the unassembled DNA will be useful for others looking to close, finish, and polish microbial genome sequences.

U.S. DEPARTMENT OF ENERGY | Office of Science

Utturkar, S. M. *et al.*, "A case study into microbial genome assembly gap sequences and finishing strategies" *Front. Microbiol. (*2017) 8:1272.