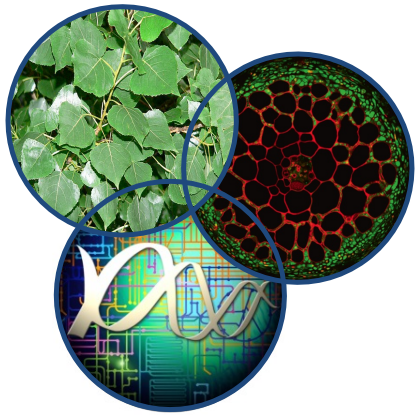
Deriving High Quality Genome Sequences



Background

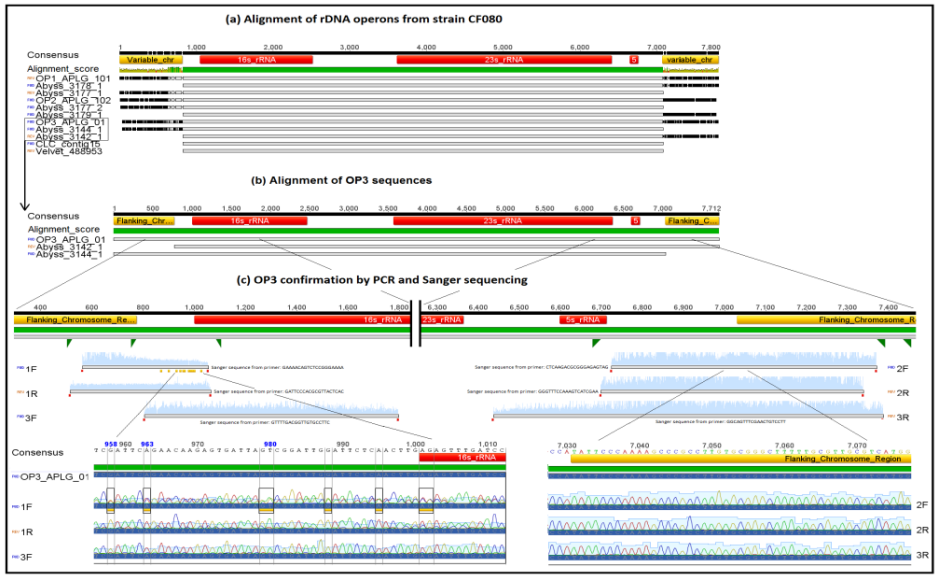
- Genome data containing a large number of contigs reduces its usefulness and can prevent publication of the information. Can combined sequencing from different technology platforms improve assembly and quality?

Approach

- Illumina, 454 and PacBio sequencing technologies were used to generate up to 11 different *de novo* and hybrid genome assemblies for four different bacteria, which were assessed for quality using summary statistics (e.g. number of contigs, N50) and using *in silico* evaluation tools.
- Differences in predictions of multiple copies of rRNA operons were evaluated by PCR/Sanger sequencing and then the validated results were applied as an additional criterion to rank the assemblies.

Outcomes

- Assemblies employing longer PacBio reads were better able to resolve repetitive regions.
- In this study, the combination of Illumina and PacBio sequence data assembled through the ALLPATHS-LG algorithm gave the best summary statistics and most accurate rRNA operon number predictions.



Significance

- This comprehensive comparison of different technologies, library types and assembly algorithms will aid others looking to improve existing genome assemblies.

Utturkar S.M., *et al.* 2014. Evaluation and validation of *de novo* and hybrid assembly techniques to derive high quality genome sequences. Bioinformatics. In press.

U.S. DEPARTMENT OF ENERGY

http://pmi.ornl.gov/

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY