Microbial Finishing at DOE Joint Genome Institute (JGI): Sequencing Difficult DNA Templates

Michele Martinez, Paul Richardson, and Alla Lapidus

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

The US DOE Joint Genome Institute (JGI) mission is to provide the scientific community with high-quality finished genomes. Approximately 300 microbial genomes are currently in the JGI pipeline and to date, 65 have been completed. The objective of the Microbial Finishing laboratory is to process sequencing reactions in order to close physical gaps, sequence gaps, and to increase quality of reads. Since most of the genomes contain complex regions which are difficult to sequence with standard protocols, the lab must use a multitude of techniques specialized for each project. Problematic regions, for example, can be GC-rich or contain hairpin loops, have long homopolymer stretches, can be AT-rich, and/or contain tandem repeats of variable length. Gap closer in such regions is expensive as well as time-consuming, since it requires extensive troubleshooting strategies. Approaches include, optimizing reaction conditions, applying various sequencing chemistries, sequencing the opposite strand, and additional manual editing. For genomes with ≥ 65% GC content, we use a four step approach to sequence through difficult regions: DMSO, Sequence Finishing Kit (SFK), PCR, and shatter libraries. This strategy has allowed JGI's Microbial Genome Finishing Group to complete a number of complex microbial projects, such as, Frankia (~75% GC-rich) and Thermobifida fusca (~68% GC rich).

Future Development:

3% of the genomes in the JGI pipeline have greater than 60% AT content. These genomes are more difficult to clone leading to higher number of uncaptured gaps when compared to those with lower AT content. Also, physical gaps and the polishing has proven to be difficult with standard sequencing strategies. For example, Prochloroccocus sp 9215 (~70% AT) content, has benefited from 454 data. However, confirming consensus (with 454 only reads) has not been completely successful. Therefore, it is necessary to research and develop new methods of approach in this area.

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. W-7405-ENG-36.
LBNL-59314 Poster

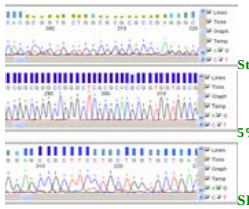
Problematic regions include, but are not limited to:

1. GC rich
2. Tandem Repeats
3. AT rich
4. Homopolymer Stretches

Mycobacterium sp MCS ~ 69% GC content

Identify problematic region

For genomes with ≥ 65 % GC content, all RCA and sequencing chemistry reactions are processed with 5% DMSO. As all projects move toward the polishing stage, usually during the second round, areas are evaluated at an individual level. This allows for the finisher and the laboratory to construct a "plan of attack." Each situation is different, but in general: (1) if the region is GC the process begins with using 5% DMSO in reactions, (2) if those reactions fail then Sequence Finishing Kit is used on the same reactions (3) if the region is repetitive, PCR followed by sequencing to confirm consensus (4) if AT rich, 5% DMSO does not negatively effect the reaction and is used (5) if homopolymer stretches, generally begin with PCR. All of these finishing reactions strongly rely on primer design and placement. In general, primers should be at least 21 bp in length, about 100 bp back from region, and with a Tm higher than 55°C.



Standard RCA and Sequencing Chemistries

5% DMSO in RCA and in Sequencing Chemistries

SFK and standard Sequencing Chemistries

Four Step Approach

5% DMSO added to RCA and Sequencing Chemistry

↓

Sequence Finishing Kit (SFK)

↓

PCR

↓

Shatter Library

Other options to use duringsequencing include:

- 98°C soak for 10 minutes
- 98°C denaturation temperature
- Increase number of cycles to 35

Microbe	Overall GC content of genome	Average Read Length	
		Standard Reaction	5% DMSO
Bacillus cereus NVH391-98	~35% GC	508	591
Chlamydia trachomatis G/9768	~41% GC	541	546
Mycobacterium vanbaalenii Pyr-1	~68% GC	599	594

A small sample set was tested to determine how 5% DMSO in sequencing reactions would effect genomes with different GC content. This experiment needs to be conducted on a larger scale, but preliminary results indicate that 5% DMSO can be used on high AT rich genomes as well. The read lengths are not increased, however, for high-throughput production this will allow, in the future, for all projects to be processed with DMSO. In addition, within high AT rich genomes, there are still areas with secondary structures, so this may help reduce the amount of finishing.

PCR product spanning region or Clone

Subclone into 2 kb pieces:
Psychrobacter had a ~15kb tandem repeat. A fosmid clone spanning the region was sheared into 2kb fragments and subcloned.

Sonicate PCR Product:
Thermobifida fusca had GC hard stops. Sonicated PCR into 500 bp pieces and cloned into pMCL200 vector. Smaller fragments are necessary in order to "break" the secondary structure

Amplify using SFK and sequence with 98°C deanturation temperature:
With Psychrobacter readlengths of 1000 untrimmed, were achieved and tandem was solved. In the case of Thermobifida fusca, the final gap was closed.

If all else fails, shatter!

PCR products, in general, will produce "cleaner" sequence and can melt secondary structures. However, there are cases, where a PCR product is obtained, but good quality sequence is not.

PCR product fails to sequence

Clone PCR product into pCR4-TOPO vector with TOPO TA Cloning Kit

Use Sequence Finishing Kit and standard sequencing chemistries

Standard sequencing of PCR product failed, after cloning and use of SFK, BHXU14 and BHXU15 closed the gap.

If sequencing reactions fail

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. W-7405-ENG-36. LBNL-59314 Poster