Title
Enhancing Microbial Genome Finishing Through Frameshift Targeting

Authors
Foster, Brian
Goltsman, Eugene
LaButti, Kurt
et al.

Publication Date
2008-05-22

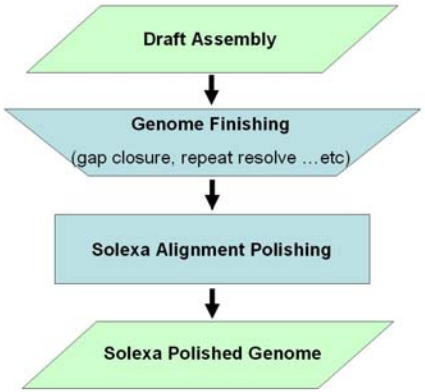
Enhancing Microbial Genome Finishing through Frameshift Targeting

B. Foster, E. Goltsman, K. LaButti, S. Lowry, and A. L. Lapidus
Joint Genome Institute Production Genomics Facility, Walnut Creek, CA

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Abstract
The JGI's finishing standards require every nucleotide in the final microbial consensus to be supported by at least two Sanger reads, and to be of at least Q30 quality. Additionally, areas covered by pyrosequence only (454-only) should be inspected for potential errors and re-sequenced. The polishing process is a very important but time consuming step. Our current polishing strategy developed for Microbial genomes incorporates the use of Solexa data to bring the final consensus quality up to our pre-defined standard (see Kurt LaButti poster).

Although this process can eliminate up to 97 percent of our polishing targets automatically, some regions still remain. These regions still require manual analysis based on our traditional (Sanger based) polishing methods. An addition to our polishing process automation includes selective targeting of the remaining regions. This automation increases efficiency by only polishing targets within proximity to predicted frameshifts (for details of frameshift detection see Andrey Kislyuk poster). This approach automates the remaining polishing process without wasting resources on improving regions which may not be biologically relevant. Our quality improving approach (polishing) will be presented in detail.

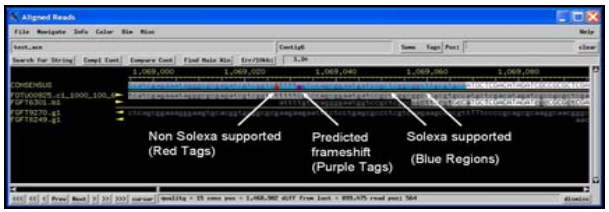


The Polishing Process
Traditionally, using only Sanger based sequencing, several rounds of primer walking and/or PCR were needed to bring some areas of the genome up to standard. The Polisher, a tool developed at JGI, is used to add coverage cheaply and efficiently with as little manual effort as possible. The polisher leverages the high sequence coverage generated by the Solexa platform to correct areas in the assembly and support areas of low sequence coverage. The process begins with the alignment of Solexa reads to the consensus sequence using Arachne's QueryLookupTable aligner. The alignments are screened for the best hit based on percent identity. Reads aligning to multiple consensus positions with the same percent id are randomly distributed to such positions. The reference is then examined. A disagreement is marked for correction if the Solexa coverage at that disagreement is at least 10 x and the majority disagreeing Solexa calls account for at least 70% of the coverage. The corrections for base mismatches, deletions, and insertions can then be automatically made to the reference sequence in the .ace file.

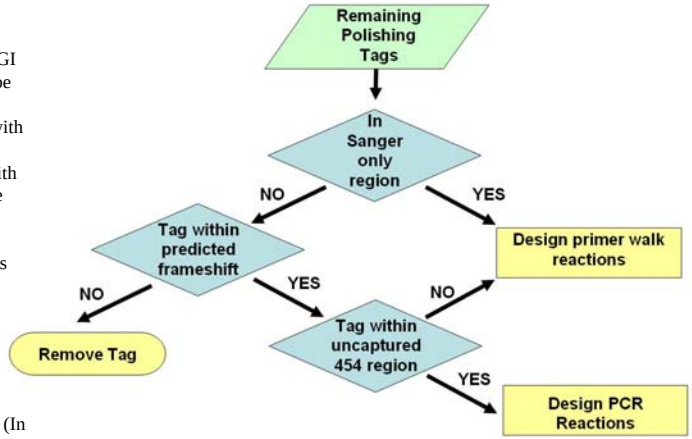
Our corrected reference file is then screened for areas that do not meet JGI criteria for polished final sequence. Low quality consensus bases must be addressed if the quality value of the consensus base is less than Q30. Additionally, areas which do not meet our criteria include: areas with 1 x coverage and 2 x coverage areas with reads coming from the same clone. Areas also identified as needing additional support are regions with 2x coverage but the reads at the region belong to different clones and the read sequences have conflicting base calls. 454 sequence only areas are also addressed. The Solexa alignments generated previously are used to potentially support the low quality, single subclone, and 454 only regions identified. If the aligned Solexa coverage is at least 10x and 70% of the coverage supports the sequence in these areas, the areas are changed to Solexa supported regions and do not need any additional confirming coverage.

Targeting remaining regions
Although the polisher typically eliminates over 95% of polishing targets (In the Methlyocella example 99%), some polishing targets still remain. These can arise when Solexa criteria are not met or when Solexa reads fail to align due to errors in the consensus sequence. A fraction of the remaining polishing target areas are found to lie in areas which are covered with only Sanger sequencing regions (no 454 reads). These regions are automatically targeted for Sanger sequencing based confirmation using Consed's autofinish functionality.

There are however polishing targets that fall within areas that have 454 data underlying them. It is well known that the main sequencing errors produced by the 454 platform are insertions and deletions of bases in homopolymer regions. These insertions and deletions can be a major factor in the introduction of frameshift errors in the final sequencing product. These 454 covered regions can be selectively targeted for verification utilizing the GeneMark based frame shift finding tool developed by Andrey Kislyuk and Mark Borodovsky.



Frameshifts are predicted for the Solexa polished reference sequence utilizing this tool. The remaining polishing targets are examined to determine if they fall within 500 bases of a predicted frameshift. If they do, then the tags remain for polishing. The polishing tags are then examined to determine if there are any Sanger clones spanning the target. If there are no clones spanning the polishing target and hence no template to be used for Sanger based sequencing, the area is identified for PCR primer design to generate a template. The remaining polishing regions are targeted using Consed's autofinish.



Methylocella example
Methylocella silvestris, a microbial finishing project, was polished using the frameshift targeting described. The Solexa polishing removed more than 99% of the polishing targets generated. With the addition of frameshift targeting, ~60% of the remaining targets are potentially eliminated by not falling within proximity to a predicted frameshift.

Methylocella Project Stage	Bases with Polishing Tags	Oligos suggested by Autofinish
Finished Draft Pre Solexa polishing	245455	1629
Post Solexa polishing	1738	130
Post Frameshift	648	43
Net reduction after frameshift targeting	**1090**	**87**

Conclusions
Frameshift detection has the potential to reduce the amount of polishing targets remaining after Solexa alignment polishing. Furthermore, the automatic identification of frameshift errors within 454-only regions with no clone templates, coupled with automatic PCR primer design of these regions will become an increasingly valuable step as the trend towards lower clone based coverage continues. Automation of the remaining polishing steps seeks to reduce resources spent inspecting the targets remaining after Solexa polishing, as well as time spent inspecting frame shifts at the annotation stage.

Further work
The automation of frameshift targeting is still in prototype stage. Further work would involve testing the software on a variety of microbial genomes at the post Solexa polished stage and analyzing the results. The number of frameshifts fixed by Sanger re-sequencing using the frameshift targeting method should be tracked. It would be of additional interest to note the fraction of frameshifts identified by the annotation group which were missed by the frameshift detection tool.

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 Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.