

# Lawrence Berkeley National Laboratory

## Recent Work

**Title**

QD Extension for Microbial Finishing

**Permalink**

<https://escholarship.org/uc/item/7z82657w>

**Author**

Lapidus, Alla

**Publication Date**

2006-04-10

## QD Extension for Microbial Finishing

Benjamin Horowitz<sup>1</sup>, Stephan Trong<sup>1</sup>, Eugene Goltsman<sup>1</sup>, Alex Copeland<sup>1</sup>,  
Vasanth Singan<sup>1</sup>, Steve Lowry<sup>1</sup>, Stephanie Malfatti<sup>2</sup>, Patrick Chain<sup>2</sup>, Pat Kale<sup>1</sup>,  
Alla Lapidus<sup>1</sup>

<sup>1</sup> Joint Genome Institute Production Genomics Facility, Walnut Creek, CA

<sup>2</sup> Lawrence Livermore National Laboratory, Livermore, CA

The Production Quality Control (QC) produces a series of assemblies, screening shotgun reads for contamination by vector sequence, for low-quality reads, and, in the case of microbial projects, for contamination by eukaryotic sequence. In coordination with the Microbial Finishing group lead at the PGF, an assembly that is “good enough” is named the Quality-controlled Draft (QD) assembly. Once the QD assembly is produced, the project passes from Production QC to one of the finishing groups at the JGI. During the finishing process, misassemblies are resolved, gaps between contigs are closed, and confirming reads are gathered for thinly covered areas.

Judging whether a project is ready for finishing is a complex task, and requires answering numerous questions, such as: Is sufficient read and clone coverage present? Will existing contig linking information (scaffolding information) allow the closing of gaps between contigs? Has high or low GC content contributed to difficult-to-sequence or difficult-to-clone areas? Do the libraries have unusual features (e.g., being absent from certain areas of the genome, or having widely-spread insert size distributions)? Are contig GC content distributions consistent with project GC content distribution? How many repeats are present in the genome? Have repeats contributed to misassemblies?

In order to facilitate the answering of these questions, we undertook the QD Extension project. QD Extension enhances the existing microbial finishing pipeline with results from the Phrap, PGA and Arachne assemblers, with scaffold information, and with coherent, informative web reports. By providing such additional information in a browsable, web-based format, we help the Microbial Finishing group lead at PGF more accurately and quickly determine whether a project is ready for finishing. This project facilitated the identification of a set of metrics that help to shed light on the finishing task that lies ahead of the biologist assigned to finish a prokaryotic genome.

This work was performed under the auspices of the US DOE of Science, Biological and Environmental Research Program, and by the University of California, LLNL under Contract No. W-7405-Eng-48, LBNL under Contract No. DE-AC02-05CH11231 and LANL under Contract No. W-7405-ENG-36.