ASSEMBLING SANGER AND NON-SANGER SEQUENCING DATA

Darren Platt [1,3,] Paul Richardson[2,3]

1 Lawrence Livermore National Laboratory,
2 Lawrence Berkeley Laboratory
3 D.O.E Joint Genome Institute

New sequencing technologies promise cheaper and faster sequencing of DNA but with the current drawback of producing data that are much more difficult to assemble due to smaller read lengths and unpaired reads. We tested the **Forge** genome assembler unmodified with ~58Mb of short 100-bp **454** sequencing data *Prochlorococcus MIT9215* species and were able to obtain significant contigs without modifying code or parameters. We also separately generated ~2x coverage of Sanger sequencing data. In the absence of paired data, accurate quantitative detection of short overlaps becomes much more critical. We discuss the calibration framework used in the assembler that allows this adaptation. The much larger number of input reads also requires a very efficient system for determining overlapped quality. We describe a set of heuristics that can quantify overlapped quality without performing an alignment. As we expect in the near term that hybrid assemblies of Sanger and "short" reads will be necessary for producing finished genomes, we also describe the statistical modifications required to produce denovo assemblies of such hybrid data and discuss the properties of the regions that confounded one technology or the other in this organism.