Improving Sequence Assemblies Using High-Quality Overlaps

University of Maryland Genome Group [Mike Roberts,
James Yorke, Brian Hunt, Wayne Hayes, Cevat Ustun, Aleksey Zimin]
and
Paul Havlak, Baylor College of Medicine.

April 6, 2003

Abstract

Finishing a genome costs about as much as the initial assembly, with most of that cost directed towards filling gaps (Celniker et al, Genome Biology 2002-03-12). Since initial assemblies typically get 95-99% of the sequence, any improvement in quality and amount of sequence to bring us closer to 100%, no matter how small, translates into an enormous cost savings for the finishing step.

Recall that one of the first steps in genome sequence assembly is determining which reads overlap. In this talk we will present recent results from a collaboration between the University of Maryland and the Baylor College of Medicine which measures the effect on assembly of various techniques for computing overlaps, while the remainder of the assembly process remains unchanged. The efficacy of some of the Maryland techniques have already been demonstrated last year in collaboration with Celera Genomics in their assembly of Drosophila melanogaster; here we study their effect on the assembly of the genome of Rattus norvegicus. As a basis for comparison, we test our assemblies against a small amount of independently finished sequence which exists for R. norvegicus.

The Atlas assembly at Baylor has already produced a high-quality draft sequence for R. norvegicus. Nonetheless, this still leaves some five percent of the mapped scaffolds in gaps. We find that when the set of overlaps are more carefully selected before being fed to Atlas, the quality of the scaffolds improves over the already high quality assembly. Specifically, the total amount of sequence produced, correctness of individual bases, and contig length improve.

Read Extension. Trimmed reads have far fewer bases than untrimmed reads. Making use of some of the low quality region is of considerable value since the U.S. government alone spends roughly $100 million generating these sequences annually. We use multi-read-comparison based error correction to generate a consensus sequence across long stretches of low-quality bases. We find that several moderately low-quality overlapping sequences can give us as much information as a single high-quality sequence.