Lawrence Berkeley National Laboratory
Recent Work

Title
Comparison of Whole Genome Shotgun Assemblers

Permalink
https://escholarship.org/uc/item/0vv9784f

Authors
Trong, Stephan
Horowitz, Benjamin
Singan, Vasanth
et al.

Publication Date
2005-09-27
Comparison of Whole Genome Shotgun Assemblers
Stephan Trong¹, Benjamin Horowitz¹, Vasanth Singan², Eugene Goltsman², Alex Copeland² and Alla Lapidus²

¹ EEBI Division
Lawrence Livermore National Laboratory

Whole genome shotgun assembly involves randomly breaking genomic DNA into small fragments, sequencing each fragment, then piecing the fragments together to recreate the original DNA sequence. Software programs called assemblers have been developed to reconstruct the original DNA using the overlapping fragments. Although diverse algorithmic approaches are applied by different assemblers, no assembler consistently pieces together the original DNA sequence correctly. We analyze some of the most popular assemblers.

Introduction:

– How can we reduce the time and labor of finishing a genome?
– Most assemblers follow overlap-layout-consensus approach
  – Overlap: determine which reads share common sequence
  – Layout: place reads, given overlap and pairing information
  – Consensus: generate DNA sequence
– The main errors include misassemblies and gaps
  – We need to replace our current assembler (Phrap) with one that joins fragments more accurately
– Does one assembler produce fewer misassemblies and gaps?

Results:

– Arachne produces few misassemblies, but many gaps, and has short N50 contig length
– PGA has long N50 contig length, and a moderate number of misassemblies and gaps
– Phrap 4 does not produce fewer misassemblies than Phrap 3
  – Phrap 4 uses paired read information whereas Phrap 3 does not

Discussion:

– Based on our current findings, we conclude that both Arachne and PGA join sequence fragments more accurately, and produce fewer misassemblies than Phrap
– Further work is needed to determine which is better for reducing both cost and time during the production of finished sequences
  – Need to measure number of gaps that can be closed by primer walks for each assembler and organism
– Future work: find misassemblies prior to finishing
  – Use mate-pair information to find breakpoints
  – Use cost model or regression analysis to determine least-cost assembly

Methods:

– Choose 5 organisms – a cross-section of microbes
– Run 4 assemblers on the 5 organisms
– Evaluate reconstructed sequence using 3 criteria

Criteria

– Number of misassemblies
  – Adjacent sequence separated or transposed during assembly
– Number of gaps
  – Genome not covered by assembly
– N50 contig length
  – Length L such that contigs of length ≥ L contain 50% of genome