

JAZZ: Whole Genome Shotgun Assembler

Nik Putnam¹, Jarrod Chapman¹, Dan Rokhsar¹, Paul Winward², Serge Dusheyko¹, Craig Furman¹, Isaac Ho¹, Sara Ting²

¹Lawrence Berkeley National Laboratory / DOE Joint Genome Institute
2800 Mitchell Drive, Walnut Creek, CA 94598

²Lawrence Livermore National Laboratory / DOE Joint Genome Institute
2800 Mitchell Drive, Walnut Creek, CA 94598

March 26, 2007

ACKNOWLEDGMENTS:

The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

DISCLAIMER:

[LBNL] This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California

[LLNL] This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

Whole Genome Shotgun: An Overview

The Problem:

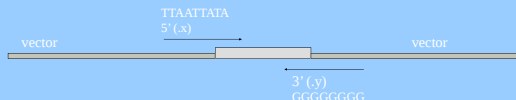
Determine the sequence of multibillion bp genomes when machines can only read 700bp accurately

The Solution:

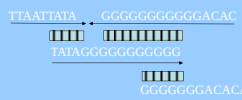
1. Shred many copies of the genome into randomly-distributed fragments



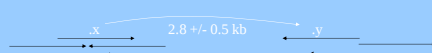
2. Sequence 6-10x worth of paired-end reads from the fragments



3. Assemble reads into **contigs** using overlapping ends of reads, like a jigsaw puzzle



4. Order and orient contigs into **scaffolds**, using paired-end insert info



Why It's Hard

Repeats

Because repeats play a central role in genome evolution, two reads with shared sequence may not be from the same place

Polymorphism

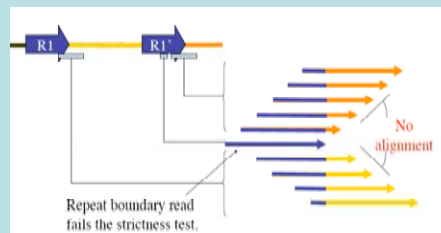
If multiple haplotypes are present, two reads with different sequences might be from the same place

Imperfect data

Non-uniform shredding, contamination, poorly-distributed insert sizes defy theoretical models

Algorithmic Details

Strictness – alignments should be presented between all the reads in a neighborhood if a given region is unique



Rectangles involve paired-end linking info early in the layout process



Alignments scored in conjunction with local depth

Spring-scaffold model for complex topology resolution



A scripting control language allows for custom recipes without recompilation

Most stages are *parallelized*

A *rollback mechanism* allows starting from an intermediate point in a previous assembly

Current Work

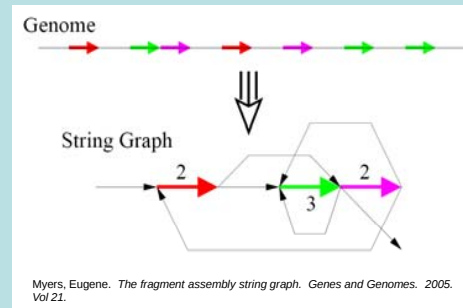
A suite of data structures for rapid R&D which leverages string graph and traditional graph paradigms

Applications

- Short reads
- Mapping data
- Polymorphism
- Hybrid assembly

String Graph (Myers/BOA assembler)

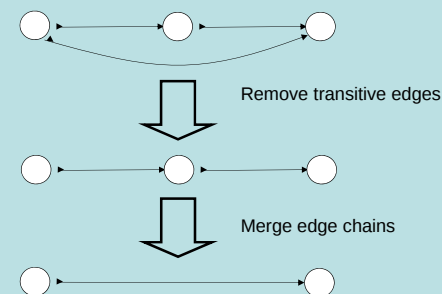
- Collapse repetitive sequence
- Identify arcs of unique sequence
- Determine repeat counts by flow analysis
- Find Eulerian tour to spell out genome sequence



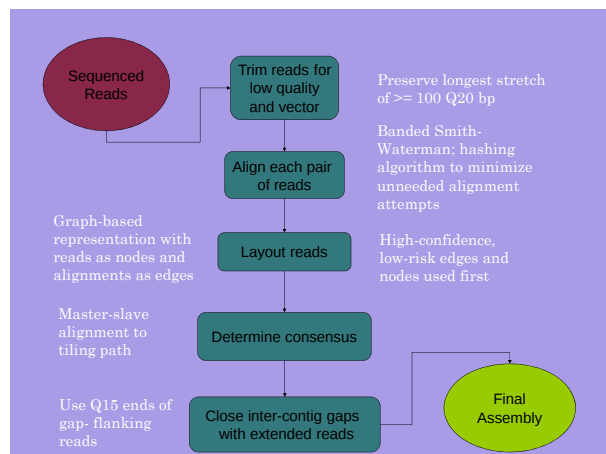
Benefits

- Highly-distilled graph representation compacts the # of nodes and edges
- Transitive reduction, edge chain merging and contained read removal reduces space and complexity
- Eulerian tours are solvable in polynomial time

Simplifications



The Jazz Workflow



Results



Fugu rubripes

Size: 393.0 Mb
Est. depth: 8.5X
Contigs: 1438
Scaffolds (>2 Kb): 7213



Ciona intestinalis

Size: 173.0 Mb
Est. depth: 11X
Scaffolds (>2 Kb): 2242
Scaffold N50: 136 sc > 234Kb
Contigs: 1438