JAZZ: Whole Genome Shotgun Assembler

Nik Putnam[1], Jarrod Chapman[1], Dan Rokhsar[1], Paul Winward[2], Serge Dusheyko[1], Craig Furman[1], Isaac Ho[1], Sara Ting[2]

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

[2]Lawrence 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

JAZZ : Whole Genome Shotgun Assembler

Authors: Nik Putnam[1], Jarrod Chapman[1], Dan Rokhsar[1] Presenters: Serge Dusheyko[1], Craig Furman[2], Isaac Ho[1], Sara Ting[2], Paul Winward[2]

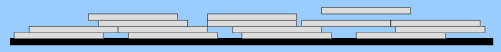
[1]Lawrence Berkeley National Laboratory, [2]Lawrence Livermore National Laboratory

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

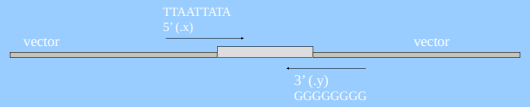
Whole Genome Shotgun: An Overview

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

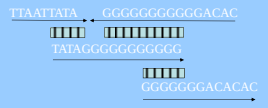
The Solution:
1.Shred many copies of the genome into randomly-distributed fragments



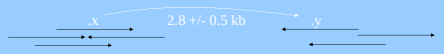
2. Sequence 6-10x worth of paired-end ***reads*** from the fragments



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



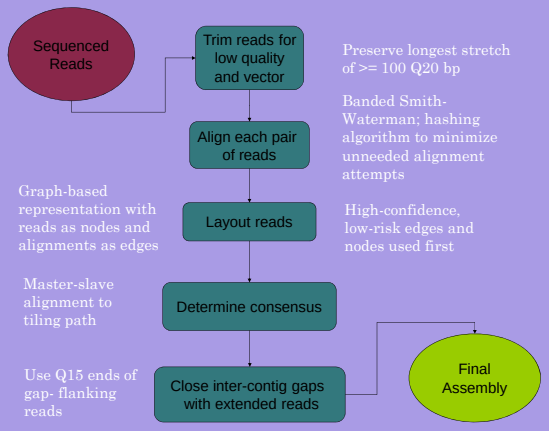
Why It's Hard

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

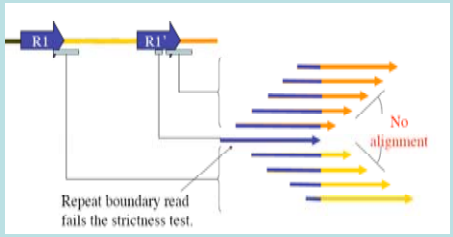
Imperfect data
Non-uniform shredding, contamination, poorly-distributed insert sizes defy theoretical models

The Jazz Workflow



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

Results

Fugu Rubripes

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

Ciona intestinalis

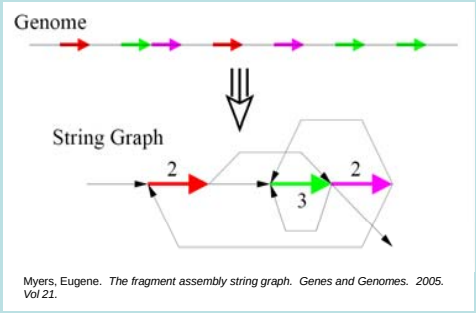
Size: 173.0 Mb
Est. depth: 11X
Scaffolds (> 2 Kb): 2242
Scaffold N50: 136 sc > 234Kb
Contigs: 1438

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

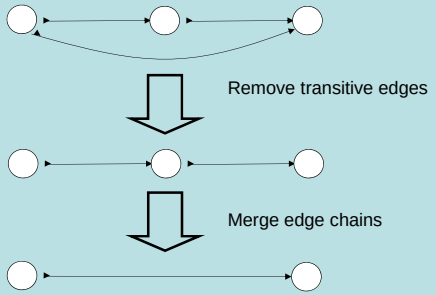
String Graph (Myers/BOA assembler)
- Collapse repetitive sequence
- Identify arcs of unique sequence
- Determine repeat counts by flow analysis
- Find Eulerian tour to spell out genome sequence



Myers, Eugene. *The fragment assembly string graph. Genes and Genomes. 2005. Vol 21.*

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 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