Ragout – a reference-assisted assembly tool for bacterial genomes

Mikhail Kolmogorov¹, Brian Raney², Benedict Paten² and Son Pham³

¹St. Petersburg University of the Russian Academy of Sciences,
²University of California Santa Cruz, ³University of California San Diego

ISMB 2014, Boston
Outline

1. Introduction
2. Quick Overview
3. Algorithm Description
4. Results
5. Further plans
Trans-Siberian Railway

- The longest railroad in the world
- 9248 km
- ~ 15 000 000 railroad ties
A Secret Story

 крышка 

 Once Aliens came:
Once Aliens came:

And they have painted the ties in different colors:
After, they took a lot of pictures:
And after they had been gone, rain has vanished all dyes from the railroad :(
And after they had been gone, rain has vanished all dyes from the railroad :(

Can we now reconstruct the original coloring using those pictures?
And after they had been gone, rain has wanished all dyes from the railroad :(

Can we now reconstruct the original coloring using those pictures?

This is exactly a problem that genome assemblers solve!

- SPAdes
- ABySS
- Velvet
- SOAPdenovo
- SGA
- ...

A Secret Story III
Genome Assembly

Join short overlapping reads into chromosomes

Expectation:
Genome Assembly

🎉 Join short overlapping reads into chromosomes
🎉 Expectation:

🎉 Reality:
Complete Sequence?

Jumping libraries:

Long reads:

Still expensive and not as reliable as short reads

Is there any alternative?
Reference-assisted Assembly

- Using a complete genome of another closely-related organism
- Contigs are being aligned on that *reference* genome
Reference-assisted Assembly

- Using a complete genome of another closely-related organism
- Contigs are being aligned on that *reference* genome

- Structural variations?
Rearrangement Approaches

  - Tries to minimize number of structural variations between two genomes

- Kim et. al. "Reference-assisted Chromosome Assembly", *PNAS*, 2013
  - First attempt to use multiple genomes simultaneously
  - One *reference* and multiple *outgroups*
  - Still heavily rely on that reference

- Both approaches may introduce errors
Rearrangement Approaches

- Tries to minimize number of structural variations between two genomes

Kim et. al. "Reference-assisted Chromosome Assembly", *PNAS*, 2013
- First attempt to use multiple genomes simultaneously
- One *reference* and multiple *outgroups*
- Still heavily rely on that reference

Both approaches may introduce errors

So maybe we need multiple references?
Outline

1. Introduction
2. Quick Overview
3. Algorithm Description
4. Results
5. Further plans
Ragout Recipe

Ragout – Reference-Assisted Genome Ordering UTility

Written in Python/C++

Ingredients:
- Multiple references (in FASTA format)
- Contigs/scaffolds from short-read assembly
- Phylogenetic tree

Output: scaffolds
Outline

1. Introduction

2. Quick Overview

3. Algorithm Description

4. Results

5. Further plans
Comparing nucleotide by nucleotide is expensive
- Extract conserved segments (synteny blocks)
- Assumption: each block is represented exactly once in each genome
Genome as Synteny Blocks and Adjacencies

- Chromosome is represented as an alternating cycle of **directed black** and **undirected red** edges.
- **Black** edges correspond to synteny blocks.
- **Red** edges connect ends of adjacent synteny blocks.
Breakpoint Graphs Are Simple!

P

Q

a b c d

a c b d

a b c d
Breakpoint Graphs Are Simple!

![Diagram showing two graphs P and Q, with labeled vertices and edges.]
Breakpoint Graphs Are Simple!
Breakpoint Graphs Are Simple!

Diagram showing two graphs labeled P and Q with vertices a, b, c, and d, and edges connecting these vertices. The diagram also shows a transformation between the two graphs.
Breakpoint Graphs Are Simple!
Breakpoint Graphs Are Simple!

Each color defines a perfect matching
Breakpoint Graphs Are Simple!

Each color defines a perfect matching
Each color defines a perfect matching
Incomplete Breakpoint Graph

\[
\begin{array}{c}
T \\
\end{array}
\]

\[
\begin{array}{c}
a \rightarrow b \\
\rightarrow c \\
\rightarrow d \\
\rightarrow a \\
\end{array}
\]
Incomplete Breakpoint Graph

Some adjacencies are missing
Find missing edges

= Recover perfect matching

There are multiple variants of such matching

How to find the correct one?
States of Adjacencies

- **State** = adjacent vertex
- **State of** \( c^t \): \( d^t \rightarrow a^h \)
- Rearrangements change states of adjacencies
Choose an arbitrary perfect matching
Choose an arbitrary perfect matching
Pick a vertex from the graph
Choose an arbitrary perfect matching
Pick a vertex from the graph
Label tree nodes as *states* of chosen vertex in genomes
The tree represents evolution of breakpoint states
Parsimony Procedure

Find scenario with minimum number of changes

Associated cost for graph vertex $u$ and tree $T$:

$$P(u, T) = \sum_{branch (i, j), i \neq j} W(branchlength)$$
Cost for a complete graph $G$: $\sum_{u \in G} P(u, T)$

Want a prefect matching which minimizes this cost.

An efficient solution:

- Node weight $\rightarrow$ edge weight
- Find minimum weight perfect matching
- Blossom algorithm in $O(n^4)$
Iterative Assembly

- Solve the dilemma about choice of synteny block size
- Merge scaffolds with different precision into one assembly
Reﬁnement with Assembly Graph

- Incorporate very small/repetitive contigs
- Analogously to repeat resolution in short-read assembly
## Toy Test – One *E. Coli* Reference

<table>
<thead>
<tr>
<th></th>
<th>Ragout</th>
<th>MCM</th>
<th>OSLay</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffolds</td>
<td>1</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Contigs (coverage)</td>
<td>129 (97.9%)</td>
<td>77 (97.6%)</td>
<td>80 (96.7%)</td>
</tr>
<tr>
<td>Miss-ordered</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

- One *E. Coli* reference without rearrangements
- #Contigs – 156 (98.18%)
### Assembly with Rearrangements – Four *H. Pylori* References

<table>
<thead>
<tr>
<th>#References</th>
<th>Scaffolds</th>
<th>Contigs (cov.)</th>
<th>Miss-ordered</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ragout</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>91 (97.7%)</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>95 (97.8%)</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>95 (97.8%)</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>93 (97.6%)</td>
<td>0</td>
</tr>
<tr>
<td><strong>RACA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>35 (83.6%)</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>35 (83.6%)</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>35 (83.8%)</td>
<td>1</td>
</tr>
</tbody>
</table>

🔍 Four *H. Pylori* references with rearrangements

🔍 #Contigs – 183 (98.57%)
Long Reads or ...

A hybrid approach for the automated finishing of bacterial genomes

Ali Bashir1,2,7, Aaron A Klammer1,7, William P Robins3, Chen-Shan Chin1, Dale Webster4, Ellen Paxinos1, David Hsu1, Meredith Ashby1, Susana Wang1, Paul Peluso1, Robert Sebra1, Jon Sorenson1, James Bullard1, Jackie Yen4, Marie Valdovino1, Emilia Mollova1, Khai Luong1, Steven Lin1, Brianna LaMay1, Amruta Joshi1, Lori Rowe4, Michael Frace4, Cheryl L Tarr4, Maryann Turnsek4, Brigid M Davis5,6, Andrew Kasarskis1, John J Mekalanos8, Matthew K Waldor3,5,6 & Eric E Schadt1,3

- 40 bp non-paired Illumina reads
- Roche 454 reads
- PacBio reads
Long Reads or ...

- 40 bp non-paired Illumina reads
- Roche 454 reads?
- PacBio reads?
- Can we replace long reads with Ragout here?
# References Scaffolds Contigs (cov.) Miss-ordered

<table>
<thead>
<tr>
<th>#References</th>
<th>Scaffolds</th>
<th>Contigs (cov.)</th>
<th>Miss-ordered</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ragout</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>185 (94.8%)</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>179 (94.7%)</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>174 (94.7%)</td>
<td>0</td>
</tr>
<tr>
<td><strong>RACA</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>124 (85.8%)</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>127 (90.0%)</td>
<td>0</td>
</tr>
</tbody>
</table>

️ Three *V. Cholerae* references with rearrangements
️ #Contigs – 1407 (96.89%)
️ Results are shown without refinement (poor assembly quality)
New results & further plans

Assembly of *Drosophila yakuba* with three other *Drosophila* species:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scaffolds</td>
<td>10</td>
</tr>
<tr>
<td>Contigs</td>
<td>1538 (94.92%)</td>
</tr>
<tr>
<td>Miss-ordered</td>
<td>26</td>
</tr>
<tr>
<td>Contigs N50</td>
<td>162 216</td>
</tr>
<tr>
<td>Scaffolds N50</td>
<td>30 316 814</td>
</tr>
</tbody>
</table>

Assembly of multiple mouse lines

Capturing rearrangements with assembly graph

Illumina BaseSpace integration
Acknowledgements

Travel funding was generously provided by Akamai Technologies

http://fenderglass.github.io/Ragout