Constructing a De Bruijn Graph for De Novo Assembly

John Tyler
Seung-Jong Park
The Project

- Genome Assembly
- Construct a De Bruijn Graph
- Intel GraphBuilder and Hadoop
The Algorithm

• Focused on Edges
The Algorithm

- Focused on Edges
- K-mer Length:
The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
The Algorithm

• Focused on Edges
• K-mer Length:
  • 3-mer
• Short Read:
The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA
The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA
- Extract k+1-mers:
The Algorithm

• Focused on Edges
• K-mer Length:
  • 3-mer
• Short Read:
  • GTTACA
• Extract k+1-mers:
  • GTTA  TTAC  TACA
The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA
- Extract k+1-mers:
  - GTTA   TTAC   TACA
- Each k+1-mer represents two vertices and the edge between them:
The Algorithm

- Focused on Edges
- K-mer Length:
  - 3-mer
- Short Read:
  - GTTACA
- Extract k+1-mers:
  - GTTA  TTAC  TACA
- Each k+1-mer represents two vertices and the edge between them:
  - GTT  ➔ TTA  ➔ TAC  ➔ ACA
The Algorithm

• Focused on Edges
• K-mer Length:
  • 3-mer
• Short Read:
  • GTTACA
• Extract k+1-mers:
  • GTTA TTAC TACA
• Each k+1-mer represents two vertices and the edge between them:
  • GTT TTA TAC ACA
• Converted to a form GraphBuilder understands
Results

- Brute Force Algorithm
- Edge Focused Algorithm
- Contrail
- 10,000 Short Reads
Results

Time to Construct a De Bruijn Graph

Brute Force Algorithm Intel GraphBuilder

Edge Focused Algorithm Intel GraphBuilder Programs

Contrail
Conclusions

Performance
• Brute Force Algorithm ~ 4 Days
• Edge Focused Algorithm ~ 2 Hours
• Contrail ~ 8 Minutes

Intel GraphBuilder
• Some Significant Bugs
• Documentation Incomplete
• Runtime
• Early in Development