Motivations
Assembling a diploid

- chrom. mom:
- chrom. dad:

Input = reads

What we want:
Assembling a diploid

• chrom. mom: ________________________________
• chrom. dad: ________________________________

• What we want: ________________________________

• What we got: ________________________________

Correctly separated

Homozygous zone
or
Crushed polymorphism

Who?
Mom and/or Dad

Where?
What happens in the “assembly graph”? 

mom: 

dad: 

Polymorphism
What happens in the “assembly graph”? 

mom: 

dad: 

assembly graph: 

- Polymorphism 

Size of the information used in graph: (read, kmer, long reads, read pairs….)
What happens in the “assembly graph”? 

Polymorphism

Size of the information used in graph:
(read, kmer, long reads, read pairs…)

“phased” polymorphism

“isolated” polymorphism

non-polymorphic regions
What happens in the “assembly graph”?

mom:

dad:

assembly graph:

contigs:

Polymorphism

Size of the information used in graph:
(read, kmer, long reads, read pairs…)

“phased” polymorphism

“isolated” polymorphism

non-polymorphic regions
size of the information used in graph:
(read, kmer, long reads, read pairs...)

With longer information, we could “phase” the polymorphism

DECISIVE
Assembly paradigm dichotomy

**string graph**
- read length info
- phase polymorphism

**de Bruijn graph**
- k-mer length info
- “bad” phasing
Assembly paradigm dichotomy

**String graph**
- Read length info
- Phase polymorphism
- Scaling issues
  - \(\approx\) quadratic

**De Bruijn graph**
- K-mer length info
- “Bad” phasing
- Scales large instances
  - \(\approx\) linear
Assembly paradigm dichotomy

**string graph**
- read length info
- phase polymorphism
- scaling issues

**de Bruijn graph**
- k-mer length info
- "bad" phasing
- scales large instances

**Our proposal, BWISE:**
- read length info (and beyond)
- phase polymorphism
- scales large instances
Algorithm overview
Main idea 1/2 – map reads on dBG

Super reads:
longer (or equal) than reads
exact paths from dBG.

Super Reads (SR) produced by the mapped reads:

- Shows the path 1,3,4: ACATGCATGCTAGCACTCATGCGCTATAGATATATATATATATATG
- Both show the path 2,3: CATGCACCTGACTTCACCTCATGCGTATG
- Shows the path 2,3,5: CATGCACCTGACTTCACCTCATGCGTGAGCAGCAGCAT
Main idea 2/2 – Super reads graph

- Connect super reads (exact prefix/suffix links)
- Remove redundancies
- Detect simple paths

Consequence: increases again SR size

Perfect overlap
Main idea –
Super reads graph ~ string graph

- Connect super reads (exact prefix/suffix links)
- Remove redundancies
- Detect simple paths

Output contigs
In practice

BLOOCOO

Corrected Reads

BGREAT + K2000

contigs

dBG

BCALM

Reads

In practice

Why a second iteration

- Super reads \( \rightarrow \) dBG
- Map back read pairs

Super reads:
- longer (or equal) than pairs of reads
- exact paths from dBG
In practice

BLOOCOO
Corrected Reads
BGREAT + K2000
BCALM
contigs
dBG

(preliminary) results
# Haploid simulations

<table>
<thead>
<tr>
<th></th>
<th>E. coli</th>
<th>C. elegans</th>
<th>Human</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPAdes</strong></td>
<td>contigs: 71</td>
<td>contigs: 6,550</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N50: 178,400</td>
<td>N50: 103,128</td>
<td></td>
</tr>
<tr>
<td><strong>Platanus</strong></td>
<td>contigs: 272</td>
<td>contigs: 21,413</td>
<td></td>
</tr>
<tr>
<td></td>
<td>N50: 133,252</td>
<td>N50: 103,128</td>
<td></td>
</tr>
<tr>
<td><strong>Bwise</strong></td>
<td>contigs: 56</td>
<td>contigs: 2,527</td>
<td>contigs:132,272</td>
</tr>
<tr>
<td></td>
<td>N50: 210,994</td>
<td>N50: 122,287</td>
<td>N50: 74,148</td>
</tr>
</tbody>
</table>

- Bwise & Platanus ≈ 2x more misassemblies than SPAdes

- Simulated 100x Miseq 2*250 reads 800 fragment size 1% error.

**Bwise parameters:**

- \( k_1 = 51 \)
- \( k_2 = 201 \)
**E. coli, simulated heterozygosity**

Simulated 100x Miseq 2*250 reads 800 fragment size 1% error.

Bwise parameters: $k_1=51$, $k_2=201$
Melinaea marsaeus Assembly

Platanus:
- N50: 8,129 bp

Bwise:
- N50: 15,093 bp

busco analyses
Future
&
(my) questions
Future

- Map long range information:
  - Long reads,
  - Mate pairs
  - 10x
  - HiC

- Polyploid & metagenomes

- Deal with low heterozygosity

- End user tool
  - Automatic parameters detection
  - Pipeline factorization
  - Output representation?
Thanks

Tool: https://github.com/Malfoy/BWISE

Contacts: bwise@inria.fr
**E. coli**, simulated heterozygosity

Simulated 100x Miseq 2*250 reads 800 fragment size.

<table>
<thead>
<tr>
<th></th>
<th>0.1%</th>
<th>0.5%</th>
<th>1%</th>
<th>3%</th>
<th>5%</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SPAdes</strong></td>
<td>contigs: 55</td>
<td>contigs: 973</td>
<td>contigs: 5,027</td>
<td>contigs: 121</td>
<td>contigs: 3</td>
</tr>
<tr>
<td></td>
<td>N50: 210,905</td>
<td>N50: 17,487</td>
<td>N50: 1,827</td>
<td>N50: 246,674</td>
<td>N50: 4,639,654</td>
</tr>
<tr>
<td><strong>Platanus</strong></td>
<td>contigs: 926</td>
<td>contigs: 2,224</td>
<td>contigs: 2,887</td>
<td>contigs: 1,614</td>
<td>contigs: 711</td>
</tr>
<tr>
<td><strong>Bwise</strong></td>
<td>contigs 3,835</td>
<td>contigs 1,370</td>
<td>contigs 307</td>
<td>contigs 8</td>
<td>contigs 2</td>
</tr>
<tr>
<td></td>
<td>N50: 4,520</td>
<td>N50: 10,390</td>
<td>N50: 64,217</td>
<td>N50: 2,610,736</td>
<td>N50: 4,639,654</td>
</tr>
</tbody>
</table>

BWISE : diploid assembly
SPAdes : haploid assembly

Bwise parameters : $k_1$=51, $k_2$=201