Towards closed genomes with scaffolding and gap closure

Marten Boetzer
BaseClear – Leiden

WWW.BASECLEAR.COM
My Background

- **2008**: B.Sc. Bioinformatics, Hogeschool Leiden - Netherlands

- **2010**: M.Sc. Bioinformatics, University of Leiden

- **from 2010**: After Master Thesis, working at BaseClear in Leiden
  - BaseClear’s is a commercial sequencing provider
  - Wide-range of bioinformatics services, depending on the customers needs, mainly focused on NGS
Basic strategy next-gen sequencing

- Multiple copies of genome
- Sheared random fragments
- Size fractionated fragments
- Reads
Library types

- Single read:
  - Length: ~200-1000bp

- Paired-end read:
  - Length: ~1000-10000 bp

- Mate pair read:
  - Length: ~1000-10000 bp
De novo assembly

Reconstruction of original sequence with de novo assembly
Problems assembly

- Regardless of the assembly method all assemblies are subject to some common issues:
  - Lack of coverage;
  - Repeated elements in the genome;
  - Sequencing errors;

- These issues causes the production of several contigs instead of only one
Lack of coverage

- Lack of coverage of the reads on the original genome causes multiple contigs
  - due to randomness of shearing process there is a chance that some regions of the genome are sequenced significantly lower than other regions
Repeated elements

- Repeats in the genome can cause two major problems to the assembler;
  - 1) May be collapsed in two separate contigs
Result: many unconnected contigs

Most WGS submissions to the NCBI lack of a significant number of nucleotides!
Paired-end sequencing

- **Paired-end strategy:** sequencing from two ends of a longer fragment (cut from gel)

  \[ \text{CGTAGAGA} \quad \text{ATGATCTTC} \]

  \[ \sim 200-500\text{bp} \]

- **We do not know the full length sequence, but we can use the known interspacing size to determine distance between contigs**

  \[ \text{Count} \]

  \[ \text{Fragment Size} \]
Paired-end sequencing for scaffolding

- **Example scaffolding:**
  - Pairs:
  - Contig:
  - Alignment:

- Based on alignment we can determine:
  1) **Order & orientation:**
  2) **Distance estimation:**
  3) The more pairs link the contigs, stronger the evidence:
Paired-end sequencing

- So we can link contigs using the paired-end information …

- Good data, but limited fragment size only solves small repeats
Paired-end scaffolding

- With paired-end only, we cannot cross over large repeats
- Example:
  - Tig1 -> tig3 -> tig4 or
  - Tig1 -> tig3 -> tig5???
Matepair sequencing

- Similar to paired-end sequencing, but allows for larger fragment sizes (up to 10 kb)

- Allows to jump over even larger repeats … but it is more difficult to make a good library and is more expensive!

- The larger the insert, the harder to make a good library.
Matepair scaffolding

- With matepairs, we can cross over large repeats

- With this information we can solve the large repeat
  - Tig1 -> tig3 -> tig4
  - Tig2 -> tig3 -> tig5
Scaffolding with paired-end and matepair data

- Help to find the links between contigs
- Merge them into scaffolds
- Shed light on the contig order, orientation, distance
Tools for scaffolding

- **BAMBUS** (Pop *et al.*, 2004)
  - [http://www.cbcb.umd.edu/software/bambus/](http://www.cbcb.umd.edu/software/bambus/)

- **SOPRA** (Dayaran *et al.*, 2010)
  - [http://www.physics.rutgers.edu/~anirvans/SOPRA/](http://www.physics.rutgers.edu/~anirvans/SOPRA/)

- **SSPACE** (Boetzer *et al.*, 2010)
  - BaseClear’s scaffolder!

- **MIP scaffolder** (Salmela *et al.*, 2011)

- **SCARPA** (Donmez *et al.*, 2012)
  - [http://compbio.cs.toronto.edu/scarpa/](http://compbio.cs.toronto.edu/scarpa/)

- Most of today’s de novo assemblers have an internal scaffolder

- Independent control of scaffolding process, dedicated solution for particular problem
SSPACE is a stand-alone program for scaffolding pre-assembled contigs using paired-end / matepair data.

Published in *Bioinformatics* (2011)
> 74 citations > 500 downloads > 23,000 hits

SeqAnswers (used for assembly of butterfly, eel, cobra, crow, oyster ...)

Unique in its capability to:
- control the scaffolding in an independent manner
- easiness to use
- one step method (internal alignment)
- fast and reliable
unique paired-end / matepair reads

+ contigs

align pairs to contigs (Bowtie/BWA)

combine contigs into scaffolds if linked by a specified number of pairs
Evidence of contigs on scaffolds:

- FASTA file with scaffolds

```
>scaffold2 size1086523 tigs84
f_tig201 size9884 links322 gaps-27 merged25
r_tig405 size1811 links116 gaps68
f_tig40 size24735 links258 gaps-20 merged20
r_tig69 size5388 links159 gaps52
f_tig141 size10362 links221 gaps-27 merged24
```
We initially tested (2010) the performance of stand-alone scaffolder against internal scaffolder of assemblers.

We have 3 datasets for *E. coli* (genome size of 4.63 Mb):

1. Paired-end reads (10.4M reads, 36 cycles, 200 insert)
2. Paired-end reads (7.0M reads, 36 cycles, 500 insert)
3. Matepair reads (5.4M reads, 26 cycles, 3 kb insert)
## Denovo assembly with paired-end

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Contigs (len &gt;100)</th>
<th>N50</th>
<th>Genome size (MB)</th>
<th>Genome coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLCbio</td>
<td>374</td>
<td>30.922</td>
<td>4.542</td>
<td>97.9 %</td>
</tr>
<tr>
<td>ABYSS</td>
<td>577</td>
<td>21.181</td>
<td>4.545</td>
<td>97.9 %</td>
</tr>
<tr>
<td>Velvet</td>
<td>181</td>
<td>96.357</td>
<td>4.562</td>
<td>98.3 %</td>
</tr>
</tbody>
</table>
Adding the matepair data for scaffolding

<table>
<thead>
<tr>
<th>Assembler</th>
<th>Scaffolder</th>
<th>Scaffolds (len &gt;100)</th>
<th>N50</th>
<th>Genome size (MB)</th>
<th>Genome coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLCbio *</td>
<td>SSPACE</td>
<td>16</td>
<td>826.702</td>
<td>4.582</td>
<td>98,8 %</td>
</tr>
<tr>
<td>Abyss</td>
<td>Internal</td>
<td>213</td>
<td>30.469</td>
<td>6.490</td>
<td>139,9 %</td>
</tr>
<tr>
<td>Abyss</td>
<td>SSPACE</td>
<td>21</td>
<td>649.719</td>
<td>4.579</td>
<td>98,7 %</td>
</tr>
<tr>
<td>Velvet</td>
<td>Internal</td>
<td>132</td>
<td>178.290</td>
<td>4.569</td>
<td>98,5 %</td>
</tr>
<tr>
<td>Velvet</td>
<td>SSPACE</td>
<td>44</td>
<td>1.227.351</td>
<td>4.594</td>
<td>99,0 %</td>
</tr>
</tbody>
</table>

* Note that CLCbio has recently released a scaffolder. Currently not tested.
Scaffolder comparison - datasets

- Comparison performed on two bacterial genomes

<table>
<thead>
<tr>
<th>Genome</th>
<th>No. of reads</th>
<th>Sequence Coverage</th>
<th>Read length (bp)</th>
<th>Insert size (bp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>4.6</td>
<td>2×10.4m</td>
<td>160x</td>
<td>36</td>
</tr>
<tr>
<td>P. syringae</td>
<td>6.1</td>
<td>2×3.5m</td>
<td>40x</td>
<td>36</td>
</tr>
</tbody>
</table>

- Performed by developers of SCARPA (Donmez et al., 2012)
## Scaffolder comparison – *E. coli*

<table>
<thead>
<tr>
<th></th>
<th>No. of sequences</th>
<th>Accuracy at 3k (%)</th>
<th>Coverage (%)</th>
<th>Largest (bp)</th>
<th>N50 (bp)</th>
<th>NG50 (bp)</th>
<th>Total (mbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>371</td>
<td>100.00</td>
<td>98.08</td>
<td>82,795</td>
<td>23,195</td>
<td>22,156</td>
<td>4.57</td>
</tr>
<tr>
<td>ScaRPA</td>
<td>140</td>
<td>99.76</td>
<td>98.21</td>
<td>248,493</td>
<td>74,796</td>
<td>72,387</td>
<td>4.56</td>
</tr>
<tr>
<td>SSPACE</td>
<td>142</td>
<td>99.61</td>
<td>98.17</td>
<td>178,229</td>
<td>74,044</td>
<td>74,044</td>
<td>4.56</td>
</tr>
<tr>
<td>MIP</td>
<td>134</td>
<td>99.47</td>
<td>98.28</td>
<td>236,583</td>
<td>65,283</td>
<td>62,344</td>
<td>4.56</td>
</tr>
<tr>
<td>SOPRA</td>
<td>171</td>
<td>99.71</td>
<td>98.10</td>
<td>180,052</td>
<td>71,485</td>
<td>67,663</td>
<td>4.57</td>
</tr>
<tr>
<td>Opera</td>
<td>312</td>
<td>99.20</td>
<td>98.11</td>
<td>82,805</td>
<td>24,859</td>
<td>24,833</td>
<td>4.57</td>
</tr>
</tbody>
</table>
Scaffolder comparison – *P. syringae*

<table>
<thead>
<tr>
<th></th>
<th>No. of sequences</th>
<th>Accuracy at 3k (%)</th>
<th>Coverage (%)</th>
<th>Largest (bp)</th>
<th>N50 (bp)</th>
<th>NG50 (bp)</th>
<th>Total (mbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>876</td>
<td>99.70</td>
<td>98.17</td>
<td>51,433</td>
<td>12,947</td>
<td>12,866</td>
<td>6.03</td>
</tr>
<tr>
<td>ScaRPA</td>
<td>244</td>
<td>98.90</td>
<td>98.23</td>
<td>273,039</td>
<td>77,009</td>
<td>74,955</td>
<td>6.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>272,420</td>
<td>76,699</td>
<td>74,804</td>
<td>6.01</td>
</tr>
<tr>
<td>SSPACE</td>
<td>237</td>
<td>98.83</td>
<td>98.20</td>
<td>227,759</td>
<td>59,665</td>
<td>59,665</td>
<td>6.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>227,712</td>
<td>59,728</td>
<td>59,728</td>
<td>6.02</td>
</tr>
<tr>
<td>MIP</td>
<td>244</td>
<td>98.67</td>
<td>98.32</td>
<td>252,801</td>
<td>67,675</td>
<td>67,442</td>
<td>6.02</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>252,586</td>
<td>67,333</td>
<td>67,299</td>
<td>6.01</td>
</tr>
<tr>
<td>SOPRA</td>
<td>387</td>
<td>98.69</td>
<td>98.19</td>
<td>262,140</td>
<td>53,635</td>
<td>53,635</td>
<td>6.03</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>261,935</td>
<td>53,533</td>
<td>53,533</td>
<td>6.03</td>
</tr>
<tr>
<td>Opera</td>
<td>407</td>
<td>97.56</td>
<td>98.34</td>
<td>137,453</td>
<td>30,290</td>
<td>30,283</td>
<td>6.05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>137,274</td>
<td>30,263</td>
<td>30,250</td>
<td>6.03</td>
</tr>
</tbody>
</table>
Scaffolder comparison: quantity vs quality

- **E. coli**
  - N50 (bp)
  - Accuracy

- **P. syringae**
  - N50 (bp)
  - Accuracy

Data points for different scaffolding tools:
- ScaRPA
- SSPACE
- MIP
- SOPRA
- Opera
**Scaffolder comparison**

**Speed**

- **Note:** SSPACE has an internal aligner (Bowtie or BWA) while other scaffolders require the reads to be aligned by the user themself.
Genome finishing - Gapclosure

- Scaffolding leads to undefined nucleotides (N’s) in the assembly, which do not contain information.

- Gapclosure is to replace the N’s with nucleotides.

- Manual gap closure with Sanger sequencing can require a fair amount of runs ( $$$ ).

- Automated gap closure can save a lot of time and money, as long as the output is reliable!

- Gapclosure makes use of paired-read information to fill in the gaps.
Gapclosure – manual vs automated

Manual with Sanger:

Automated with paired-reads:

Design sequencing reactions to close gap

AGTCGATAGCNNNNNNNNNNNNTAGCTCGTAGTA

AGTCGATAGCAAAGTAGAGCGTAGCTCGTAGTA
Tools for gap closure

- IMAGE (Iterative Mapping and Assembly for Gap Elimination)
  [http://sourceforge.net/apps/mediawiki/image2/](http://sourceforge.net/apps/mediawiki/image2/)

- SOAP’s GapClosure

- GapFiller
  - BaseClear’s gap-closure tool!
Contig A

1. align the paired end reads onto draft sequence

2. local assembly of the aligned reads; new contigs are produced

3. gaps are now shortened. Repeat the whole procedure in a few iterations

4. The gap is now closed

Gap

Contig B

New contigs

New reads can be aligned with the presence of extended reference

Merged contig
GapFiller strategy

A) scaffolds + R1 paired-reads

B) gap

C) gap scaffolds paired-reads

D) reads in the gap → k-mers

E) extend gap edges with k-mers
Finishing genomes in practice – a bacterial case (*E. coli*)

• A very “simple” case: *E. coli*

• A full reference is available to check the outcomes!

• We take just 1 dataset: Paired-end reads (10.4M reads, 36 cycles, 200 insert)

• Quality assessment using GAGE assembly comparison parameters (Salzberg et al., 2012):
  o Detect erroneous SNPs, INDELS, Rearrangements
### Assembly results *E. coli* (1 PE library)

<table>
<thead>
<tr>
<th>Assembly</th>
<th>SOAPdenovo assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>1.111</td>
</tr>
<tr>
<td>Genome size (bp, no N’s)</td>
<td>4.486.667</td>
</tr>
<tr>
<td>Gap count</td>
<td>0</td>
</tr>
<tr>
<td>Gap length (bp)</td>
<td>0</td>
</tr>
<tr>
<td>N50</td>
<td>6.710</td>
</tr>
<tr>
<td>Errors (SNPs)</td>
<td>12</td>
</tr>
<tr>
<td>Errors (Indels)</td>
<td>4</td>
</tr>
<tr>
<td>Errors (Misjoins)</td>
<td>1</td>
</tr>
</tbody>
</table>
Assembly results *E. coli* (1 PE library)

<table>
<thead>
<tr>
<th>Assembly</th>
<th>SOAPdenovo assembly</th>
<th>SSPACE scaffolding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>1.111</td>
<td>179</td>
</tr>
<tr>
<td>Genome size (bp, no N’s)</td>
<td>4.486.667</td>
<td>4.478.287</td>
</tr>
<tr>
<td>Gap count</td>
<td>0</td>
<td>544</td>
</tr>
<tr>
<td>Gap length (bp)</td>
<td>0</td>
<td>12.516</td>
</tr>
<tr>
<td>N50</td>
<td>6.710</td>
<td>50.557</td>
</tr>
<tr>
<td>Errors (SNPs)</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>Errors (Indels)</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Errors (Misjoins)</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Assembly results *E. coli* (1 PE library)

<table>
<thead>
<tr>
<th>Assembly</th>
<th>SOAPdenovo assembly</th>
<th>SSPACE scaffolding</th>
<th>IMAGE Gapcloser</th>
<th>SOAPdenovo gapcloser</th>
<th>GapFiller gap closure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs</td>
<td>1.111</td>
<td>179</td>
<td>179</td>
<td>179</td>
<td>179</td>
</tr>
<tr>
<td>Genome size (bp, no N's)</td>
<td>4.486.667</td>
<td>4.478.287</td>
<td>4,530,961</td>
<td>4,490.973</td>
<td>4,490.638</td>
</tr>
<tr>
<td>Gap count</td>
<td>0</td>
<td>544</td>
<td>291</td>
<td>16</td>
<td>11</td>
</tr>
<tr>
<td>Gap length (bp)</td>
<td>0</td>
<td>12.516</td>
<td>2.861</td>
<td>16</td>
<td>130</td>
</tr>
<tr>
<td>N50</td>
<td>6.710</td>
<td>50.557</td>
<td>50.558</td>
<td>50.558</td>
<td>50.558</td>
</tr>
<tr>
<td>Errors (SNPs)</td>
<td>12</td>
<td>12</td>
<td>40</td>
<td>33</td>
<td>22</td>
</tr>
<tr>
<td>Errors (Indels)</td>
<td>4</td>
<td>4</td>
<td>17</td>
<td>25</td>
<td>9</td>
</tr>
<tr>
<td>Errors (Misjoins)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

SOAP has fewest number of undefined nucleotides

GapFiller has fewest number of errors

IMAGE has most gaps remaining, and most errors
Another example: *S. coelicolor*

- A more difficult case, larger genome size, high GC%
- 2 Paired-end datasets, 70 cycles, 200 and 500 bp insert
- Draft assembly with CLCbio, scaffolding with SSPACE
- Contigs: 430 – Scaffolds: 115
Another example: *S. coelicolor*

- A more difficult case, larger genome size, high GC%
- Data taken from GAGE assembly comparison:
  - Paired-end set, 88 cycles, 180bp insert,
  - Matepair dataset, 37 cycles, 3.5 kb insert
  - Extensive quality-filtering with Quake
- Best assembly taken from the comparison: ALLPATHS
  - [http://gage.cbcb.umd.edu/results/index.html](http://gage.cbcb.umd.edu/results/index.html)
### S. coelicolor, results on state-of-the-art assembly

<table>
<thead>
<tr>
<th>Assembly</th>
<th>Assembly</th>
<th>SOAPdenovo gapcloser</th>
<th>GapFiller gap closure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genome size (bp)</td>
<td>8,558,275</td>
<td>8,557,720</td>
<td>8,558,333</td>
</tr>
<tr>
<td>Scaffolds</td>
<td>115</td>
<td>115</td>
<td>115</td>
</tr>
<tr>
<td>Gap count</td>
<td>158</td>
<td>60</td>
<td>23</td>
</tr>
<tr>
<td>Gap length (bp)</td>
<td>9,221</td>
<td>1,288</td>
<td>806</td>
</tr>
<tr>
<td>Errors (SNPs)</td>
<td>299</td>
<td>406</td>
<td>280</td>
</tr>
<tr>
<td>Errors (Indels)</td>
<td>664</td>
<td>769</td>
<td>686</td>
</tr>
<tr>
<td>Errors (Misjoins)</td>
<td>12</td>
<td>18</td>
<td>18</td>
</tr>
</tbody>
</table>
De novo assembly pipeline
Future

• Use long contigs to scaffold, e.g. with PacBio: