Expanding the SPAdes Toolbox

Anton Korobeynikov

Center for Algorithmic Biotechnology
St. Petersburg State University, St. Petersburg, Russia
http://bioinf.spbau.ru/spades
SPAdes (Saint Petersburg Assembler)
SPAdes

- Originally designed as single-cell assembler
- Can deal with highly uneven coverage and MDA-imposed chimeric reads
- Turned out to work well for multi-cell isolate assemblies
- One of two best assemblers in GAGE-B study by Salzberg’s lab (Magoc et al., Bioinf., 2013)
- The best bacterial genome assembler in the recent poll by acgt.me
SPAdes 3.5

- Improved memory consumption at the repeat resolution step (more than 2x)
- Integrated support for Lucigen NxSeq Long Mate Pair libraries
- Rewritten mismatch correction module
- Support for Oxford Nanopore reads for hybrid assemblies
# Illumina + Nanopore Hybrid Assemblies

<table>
<thead>
<tr>
<th></th>
<th>Illumina only</th>
<th>Ilmn + Nanopore</th>
</tr>
</thead>
<tbody>
<tr>
<td>Contigs &gt; 500 bp</td>
<td>92</td>
<td>1</td>
</tr>
<tr>
<td>Largest Contig</td>
<td>285414</td>
<td>4649811</td>
</tr>
<tr>
<td>Total Length</td>
<td>4649811</td>
<td>4654532</td>
</tr>
<tr>
<td>Reference Length</td>
<td>4639675</td>
<td>4639675</td>
</tr>
<tr>
<td>NG50</td>
<td>133088</td>
<td><strong>4649811</strong></td>
</tr>
<tr>
<td>NG75</td>
<td>64475</td>
<td>4649811</td>
</tr>
<tr>
<td>Misassemblies*</td>
<td>0 (0)</td>
<td>6 (0)*</td>
</tr>
<tr>
<td>Genome fraction (%)</td>
<td>98.14</td>
<td>99.99</td>
</tr>
</tbody>
</table>

Illimina 2x100 bp E. coli K12 reads are available from [http://bioinf.spbau.ru/spades](http://bioinf.spbau.ru/spades)

Nanopore reads from Nick Loman

* Misassemblies are not real, this is the difference wrt the reference
BayesHammer improvements:

- Removed $2^{32}$ k-mer limit (bigger genomes!)
- Reduced memory consumption (2x–4x)
- Much faster (e.g. 36h → 8h)
- Completely rewritten read correction procedure: faster and more precise

SPAdes improvements:

- Significantly reworked repeat resolution and scaffolding module
Developed from the beginning as a set of modular and reusable parts
Different “stages” of an assembler can be stacked together and share common information
Allows one to assemble an assembler-like application from different building blocks
Developed from the beginning as a set of modular and reusable parts

Different “stages” of an assembler can be stacked together and share common information

Allows one to assemble an assembler-like application from different building blocks

And so we did!
dipSPAdes

The first de Bruijn graph assembler designed for highly polymorphic diploid genomes:

- **Fungus**
  - Heterozygosity up to 20%

- **Sea squirts**
  - Heterozygosity up to 12%

- **Plants**
  - Avg heterozygosity 7%

- **Insects**
  - Avg heterozygosity 9%

Conventional approaches assemble such genome as two highly repetitive sequences and construct very fragmented assemblies.

dipSPAdes constructs consensus for diploid haplomes and takes advantage of structure of de Bruijn graph for diploid genome to construct longer contigs.

Illumina TruSeq

DNA is shredded into 10Kb long fragments.

Fragments are distributed among 96 pools.

Pools are barcoded and sequenced by Illumina HiSeq.
Illumina TruSeq

Reads from each pool are assembled separately

Resulting in virtual TruSeq long reads

Which can be used as usual reads

long TruSeq reads

long read assembly
Why SPAdes?

- Complex repeat structure inherited from target genome
- Interstrand chimeric connections
- Fragmentation of barcode span
- Uneven coverage
truSPAdes

- SPAdes turned into assembler for pooled barcode data
- Tuning and refinements for TSLR data
- Accurate re-analysis of resulting contigs (virtual long reads) in order to reduce misassemblies
### truSPAdes

<table>
<thead>
<tr>
<th>Metric</th>
<th>Illumina assembler</th>
<th>Ray</th>
<th>SPAdes</th>
<th>truSPAdes</th>
<th>Ideal</th>
</tr>
</thead>
<tbody>
<tr>
<td>#contigs, pb*</td>
<td>419</td>
<td>414</td>
<td>677</td>
<td>430</td>
<td>≈300</td>
</tr>
<tr>
<td>#contigs (&gt;8000 bp), pb</td>
<td>106</td>
<td>83</td>
<td>108</td>
<td>126</td>
<td>≈300</td>
</tr>
<tr>
<td>Total length (Mb), pb</td>
<td>2.2</td>
<td>1.8</td>
<td>2.7</td>
<td>2.3</td>
<td>≈3</td>
</tr>
<tr>
<td>N50</td>
<td>7 579</td>
<td>6 222</td>
<td>6 235</td>
<td>8 250</td>
<td>≈10 000</td>
</tr>
<tr>
<td>NGA50</td>
<td>5 235</td>
<td>2 511</td>
<td>4 770</td>
<td>6 551</td>
<td>≈10 000</td>
</tr>
<tr>
<td>#N’s per 100 Kbp</td>
<td>0.9</td>
<td>3083</td>
<td>242</td>
<td>0.3</td>
<td>0</td>
</tr>
<tr>
<td>Misassemblies, pb</td>
<td>1.8</td>
<td>7</td>
<td>47</td>
<td>3.1</td>
<td>0</td>
</tr>
<tr>
<td>Mismatches per 100 Kbp</td>
<td><strong>75</strong></td>
<td>84</td>
<td>190</td>
<td>100</td>
<td>0</td>
</tr>
</tbody>
</table>

*pb - per barcode: average among all barcodes in dataset

Human TSLR dataset
Dip

SPAdes

Tru

SPAdes

SPAdes
metaSPAdes

Relative abundance of species in metagenome

Coverage of single-cell E. coli sample

Genome assembly of species with extremely different abundances is similar to assembly of MDA data

This is what SPAdes was designed for!
metaSPAdes

SRX024329 (HMP data) Nx plot
RNA-Seq assembly

- Trinity (Grabherr et al., Nat. Biotech., 2011)
- Oases (Schulz et al., Bioinf., 2012)

Who needs yet another RNA-Seq assembler?
RNA-Seq assembly

- Trinity (Grabherr et al., Nat. Biotech., 2011)
- Oases (Schulz et al., Bioinf., 2012)
- IDBA-tran (Peng et al., Bioinf., 2014)
- IDBA-MTP (Peng et al., RECOMB 2014)
- SOAPdenovo-Trans (Xie et al., Bioinf., 2014)
- StringTie (Pertea et al., Nat. Biotech., 2015)
- ....

Means there is a space for improving \textit{de novo} transcriptome assemblers
How does a *single-cell genome* assembler perform on a transcriptome dataset?
How does a single-cell genome assembler perform on a transcriptome dataset?

Quite well:

<table>
<thead>
<tr>
<th></th>
<th>IDBA-tran</th>
<th>SOAPdenovoTrans</th>
<th>Trinity</th>
<th>SPAdes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcripts</td>
<td>2872</td>
<td>2725</td>
<td>2171</td>
<td>3339</td>
</tr>
<tr>
<td>N50</td>
<td>312</td>
<td>213</td>
<td>309</td>
<td>370</td>
</tr>
<tr>
<td>Aligned</td>
<td>2845</td>
<td>2693</td>
<td>2150</td>
<td>3230</td>
</tr>
<tr>
<td>Unaligned</td>
<td>27</td>
<td>32</td>
<td>21</td>
<td>109</td>
</tr>
<tr>
<td>Avg. mismatches per transcript</td>
<td>0.447</td>
<td>0.456</td>
<td>0.341</td>
<td>0.57</td>
</tr>
<tr>
<td>Total annotation coverage</td>
<td>0.075</td>
<td>0.052</td>
<td>0.058</td>
<td>0.1</td>
</tr>
<tr>
<td>Partially-assembled isoforms (&gt;30%)</td>
<td>886</td>
<td>582</td>
<td>713</td>
<td>1119</td>
</tr>
<tr>
<td>Fully-assembled isoforms (&gt;90%)</td>
<td>96</td>
<td>53</td>
<td>91</td>
<td>234</td>
</tr>
<tr>
<td>Partially-annotated transcripts (&gt;30%)</td>
<td>2611</td>
<td>2493</td>
<td>2009</td>
<td>2967</td>
</tr>
<tr>
<td>Fully-annotated transcripts (&gt;90%)</td>
<td>1436</td>
<td>1449</td>
<td>1108</td>
<td>1553</td>
</tr>
</tbody>
</table>

Yeast RNA-Seq dataset
From SPAdes to rnaSPAdes

SPAdes

- De Bruijn Graph construction
  - De Bruijn graph
- Graph simplification
- Assembly graph
- Read alignment
  - Paired index
- Repeat resolution & scaffolding

BayesHammer

Raw RNA reads

Corrected reads

Contigs & scaffolds
From SPAdes to rnaSPAdes

- Raw RNA reads
- Corrected reads
- BayesHammer

**rnaSPAdes**

- De Bruijn Graph construction
- Graph simplification
- Assembly graph
- Read alignment
- Paired index
- Isoform identification

Transcripts
rnaQUAST

- One cannot develop an assembler without having an assembly quality assessment tool
- Based on our experience with SPAdes and QUAST, developing such a tool is not an easy task
- *Parallel* development of rnaSPAdes and rnaQUAST is crucial for the success of both tools
- rnaQUAST is tool for analysing assembled transcripts using various metrics (via the reference genome and/or genome annotation)
<table>
<thead>
<tr>
<th></th>
<th>IDBA-tran</th>
<th>SOAPdenovoTrans</th>
<th>Trinity</th>
<th>SPAdes</th>
<th>rnaSPAdes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcripts</td>
<td>2872</td>
<td>2725</td>
<td>2171</td>
<td>3339</td>
<td>6954</td>
</tr>
<tr>
<td>N50</td>
<td>312</td>
<td>213</td>
<td>309</td>
<td>370</td>
<td>303</td>
</tr>
<tr>
<td>Aligned</td>
<td>2845</td>
<td>2693</td>
<td>2150</td>
<td>3230</td>
<td>6692</td>
</tr>
<tr>
<td>Unaligned</td>
<td>27</td>
<td>32</td>
<td>21</td>
<td>109</td>
<td>262</td>
</tr>
<tr>
<td>Avg. mismatches per transcript</td>
<td>0.447</td>
<td>0.456</td>
<td>0.341</td>
<td>0.57</td>
<td>0.35</td>
</tr>
<tr>
<td>Total annotation coverage</td>
<td>0.075</td>
<td>0.052</td>
<td>0.058</td>
<td>0.1</td>
<td>0.105</td>
</tr>
<tr>
<td>Partially-assembled isoforms (&gt;30%)</td>
<td>886</td>
<td>582</td>
<td>713</td>
<td>1119</td>
<td>1135</td>
</tr>
<tr>
<td>Fully-assembled isoforms (&gt;90%)</td>
<td>96</td>
<td>53</td>
<td>91</td>
<td>234</td>
<td>188</td>
</tr>
<tr>
<td>Partially-annotated transcripts (&gt;30%)</td>
<td>2611</td>
<td>2493</td>
<td>2009</td>
<td>2967</td>
<td>6138</td>
</tr>
<tr>
<td>Fully-annotated transcripts (&gt;90%)</td>
<td>1436</td>
<td>1449</td>
<td>1108</td>
<td>1553</td>
<td>4094</td>
</tr>
</tbody>
</table>

Yeast RNA-Seq dataset
When?

- SPAdes 3.6: end June
- dipSPAdes: included into SPAdes
- rnaSPAdes: beta mid June, EAP
- truSPAdes: beta mid summer
- metaSPAdes: beta end summer
Acknowledgement

SPAdes team:
Dmitry Antipov
Anton Bankevich
Elena Bushmanova
Alexey Gurevich
Dmitry Meleshko
Sergey Nurk
Andrei Przhibelski
Yana Safonova

Alla Lapidus
Pavel Pevzner

This work was supported by the Russian Science Foundation (grant #14-50-00069)
Thank you!

http://bioinf.spbau.ru/spades