De novo transcriptome assembly

Does anybody even need it?

Andrey Prjibelski

Center for Algorithmic Biotechnology
SPbSU
1. Who assembles transcriptomes *de novo*?
2. How do assemblers work?
3. What is SPAdes and how does it work on RNA-Seq?
4. How do we evaluate transcriptome assemblies?
Reference based RNA-Seq analysis

- Spliced alignment
  - TopHat (Trapnell et al., Bioinf., 2009, 3684)
  - TopHat2 (Kim et al., Gen. Biol., 2013, 1052)
  - STAR (Dobin et al., Bioinf., 2013, 558)

- Expression analysis
  - DeSeq (Anders et al., Gen. Biol., 2010, 3265)
  - DeSeq2 (Love et al., Gen. Biol., 2014, 310)
  - Cuffdiff (Trapnell et al., Nat. Biotech., 2013, 578)
A lot of species do not have finished reference
De novo transcriptome assemblers

- **Trans-ABySS** (Robertson et al., Nat. Met., 2010)
- **Trinity** (Grabherr et al., Nat. Biotech., 2011)
- **Oases** (Schulz et al., Bioinf., 2012)
- **SOAPdenovo-Trans** (Xie et al., Bioinf., 2014)
- **IDBA-tran** (Peng et al., Bioinf., 2014)
- ...
*De novo* transcriptome assemblers citations

- **Trans-ABySS** (Robertson et al., 387)
- **Trinity** (Grabherr et al., 2333)
- **Oases** (Schulz et al., 537)
- **SOAPdenovo-Trans** (Xie et al., 102)
- **IDBA-tran** (Peng et al., 18)
**De novo RNA-Seq studies**

- Reference genome is unsequenced, poorly assembled or not annotated
- Reference genome is too large to sequence
- The goal is to find fusion/novel genes
Ethanol production

- Sequenced brown algae *A. cruciatus*
- Created synthetic yeast platform for ethanol production from mannitol and DEHU
- Achieved up to 83% of the maximum theoretical yield from consumed sugars

Associative transcriptomics

- Markers in crops are hard to identify due to high ploidy
- Used Tetraploid *Brassica napus* sequencing data to identify deleted gene which controls aliphatic glucosinolate biosynthesis

Harper et al., Nat. biotech., 2012.
Cistanche deserticola assembly

Identified key enzymes for synthesis of

- Lignin
- Phenylethananoid glycosides

Li et al., PLoS one, 2015
Biosynthesis of capsaicinoids

- Assembly and annotation of Capsicum frutescens
- Several SSR and SNP markers are predicted
- 3 novel genes are predicted which filled gaps of the capsaicinoid biosynthetic pathway

Shaoqun Liu et al., PLoS one, 2013
1. Who assembles transcriptomes *de novo*?
2. How do assemblers work?
3. What is SPAdes and how does it work on RNA-Seq?
4. How do we evaluate transcriptome assemblies?
De Bruijn graph

ACGTCCGTAA
De Bruijn graph

\[ \text{AC} \quad \text{GTCCG} \quad \text{GTAA} \]

\[ k=2 \]
De Bruijn graph

ACGTCGGTAA

k=2

AC
CG
De Bruijn graph

ACG\textcolor{red}{GT}\textcolor{black}{CCG}\textcolor{red}{GT}\textcolor{black}{AA}

\[ k=2 \]

\begin{itemize}
  \item AC
  \item CG
  \item GT
\end{itemize}
De Bruijn graph

\[ \text{ACGTCGCCGTA} \]

\[ k=2 \]

AC \quad CG \quad GT \quad TC
De Bruijn graph

ACGTCCGTTAA

k=2
De Bruijn graph

ACGTCCTCCCGTAA

k=2

AC
CG
GT
CC
TC
De Bruijn graph

ACGTCCGTAA

k=2
De Bruijn graph

ACGTCCCGTA

$k=2$

AC
CG
GT
TA
CC
TC
De Bruijn graph

ACGTC
CCGTA

k=2

AC
CG
GT
TA
CC
TC
AA
De Bruijn graph

ACG

TCCGTAA

\( k=2 \)

\[
\begin{align*}
\text{AC} & \rightarrow \text{CG} \\
\text{GT} & \\
\text{TA} & \\
\text{CC} & \\
\text{TC} & \\
\text{AA} & 
\end{align*}
\]
De Bruijn graph

ACG TCCGTA

k=2

AC -- ACG --> CG

CG -- CGT --> GT

GT -- TA

CC

TC

AA
De Bruijn graph

ACG TCC GAA

k=2

AC -> CG
CG -> GT

ACG
CGT

CC -> TC

TA

AA
De Bruijn graph

ACG TCC CG GT AA

k = 2

AC → ACG → CG → CGT → GT → GTC → TCC → TC → AA
De Bruijn graph

ACGTC

CCG

GTA

k=2

AC

CG

GT

TA

CC

TC

AA

ACG

CGT

CCG

GTC

TCC
De Bruijn graph

ACGTC

CGT

AA

k=2

ACG

CG

GT

TCC

AA

TA

CCG

CGT

GTC

AC

CG

GT

TC

ACG

CGT

TCC
De Bruijn graph

ACG TCC GTAA

$k=2$
De Bruijn graph

ACGTCGCCGTAA

k=2

![De Bruijn graph diagram](image-url)
De Bruijn graph

ACGTCG

k=2

ACG → CG → GT → TA
CCG → CGT → GTA
CC → TC → AA
TCC → GTC → TAA
Condensed de Bruijn graph

**ACGTCCCGTAA**

$k=2$

Diagram:

- Node 1: AC, outgoing edges to CG (ACG), CC (CCG)
- Node 2: CG, outgoing edges to GT (CGT), TC (TCC)
- Node 3: GT, outgoing edges to TA (GTA), TC (TCC)
- Node 4: TA, outgoing edge to AA (TAA)
- Node 5: CC, outgoing edge to TC (TCC)
- Node 6: TC, outgoing edges to TA (GTA), AA (TAA)
Condensed de Bruijn graph

ACGTCGGTAA

k=2
Condensed de Bruijn graph

ACGTCGCGTGAA

k=2
Condensed de Bruijn graph

ACGTCCCGTAA

k=2
Condensed de Bruijn graph

ACGTCCGTA

k=2
Condensed de Bruijn graph

ACGTCCGCGTAA

k=2
Condensed de Bruijn graph

ACGTCGGTGAA

\( k=2 \)
ACGTTCGCGTAA

k=2
So, what’s the difference?
Heuristics

- Sequencing errors
Sequencing errors

CCGTTG
CGTTAC
GTTGCA
TGCAGG
Sequencing errors

CCGTTTG
CGTTCAC
GTTGCA
TGCAGG

CCG
CGT
GTT
TTG
TTA
TAC
TGC
GCA
CAG
AGG
What about sequencing errors?

CCGTT TG
CGT TAC
GTTGCA
TGCAGG
Sequencing errors

CCGTTTG
CGTTACAG
GTTGCA
TGCAGG

CCG
CGT
GTT
TTG
TTA
TAC
ACA
TGC
GCA
CAG
AGG
Sequencing errors

CCGTTTGC
CGTTACAG
GTTGCA
TGCAGG

CCG
CCGTT
GTT
GTTACAG
GTTGCAG
CAG
CAGG
AGG
Real life
Heuristics

● Sequencing errors
● Isoform detection
Isoform detection
Isoform detection
Isoform detection
Isoform detection
Isoform detection
Isoform detection
Isoform detection
Isoform detection
De Bruijn graph
De Bruijn graph
De Bruijn graph

- Ideal case: 1 gene — 1 component
De Bruijn graph

- Ideal case: 1 gene — 1 component
- Reality
  - Short repeats
  - Paralogous genes
  - Introns
  - Intergenic sequences
  - Poly-A tails (caused by polyadenylation)
De Bruijn graph

- **Ideal case:** 1 gene — 1 component
- **Reality**
  - Short repeats
  - Paralogous genes
  - Introns
  - Intergenic sequences
  - Poly-A tails (caused by polyadenylation)
- **Varying expression levels result in different coverage**
  - Harder to detect sequencing errors
1. Who assembles transcriptomes *de novo*?
2. How do assemblers work?
3. What is SPAdes and how does it work on RNA-Seq?
4. How do we evaluate transcriptome assemblies?
SPAdes now

● > 600 citations (Bankevich et al., JCB, 2012)
● ~ 3.5K unique users downloaded the latest SPAdes 3.6
● One of the most popular apps on BaseSpace
● 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
Why to create new assembler?

- Conventional bacterial sequencing
- Metagenomics
Conventional bacterial sequencing

Multiple (Unsequenced) Genome Copies

Read Generation

Reads

Fragment Assembly

Sequenced Genome

...GGCATGCAGTCAGAAACTATCATAGCTAGATCGTACGTTAGCC...
Metagenomics

- >99% bacteria cannot be cultured
- Metagenomics: sequencing of whole bacterial community
  - Reads from dozens of different genomes mixed in one data set
  - Different coverage for different bacteria
  - Presence of different strains
  - Conservative genomic regions
- Hard to assemble and classify resulting sequences
Single-cell sequencing via MDA

- Random hexamer primers
- Phi29 DNA polymerase strand displacement
Challenges in single-cell assembly

- *E. coli* isolate dataset
- *E. coli* single-cell dataset
SPAdes and its offsprings

• Originally designed as a single-cell assembler
• Can deal with highly uneven coverage and MDA-imposed chimeric reads
• Became the base for other assembly projects:
  o hybridSPAdes (for Illumina + PacBio assembly)
  o dipSPAdes (for highly polymorphic genomes)
  o truSPAdes (for Illumina True Synthetic Long Reads)
  o metaSPAdes (for metagenome assemblies)
  o rnaSPAdes (for RNA-Seq data)
De novo transcriptome assemblers

- Trans-ABySS (Robertson et al., Nat. Met., 2010)
- Trinity (Grabherr et al., Nat. Biotech., 2011)
- Oases (Schulz et al., Bioinf., 2012)
- SOAPdenovo-Trans (Xie et al., Bioinf., 2014)
- IDBA-tran (Peng et al., Bioinf., 2014)

Who needs yet another?
De novo transcriptome assemblers

- *De novo* assemblers are important
- However, reference-based assemblers perform better than *de novo* assemblers (Martin & Zhang, Nat. Rev. Genetics., 2011)
  - Means there is space for improving *de novo* transcriptome assemblers
New de novo transcriptome assemblers

- IDBA-tran (Peng et al., Bioinf., 2014)
- IDBA-MTP (Peng et al., RECOMB 2014)
- SOAPdenovo-Trans (Xie et al., Bioinf., 2014)
- Fu et al., ICCABS, 2014
- StringTie (Pertea et al., Nat. Biotech., 2015)
- Bermuda (Tang et al., ACM, 2015)
**SPAdes on RNA-Seq?**

- SPAdes is a single-cell assembler
  - Single-cell data sets have highly uneven coverage
  - So does **RNA-Seq** data sets (different expression levels)
### SPAdes on RNA-Seq

**M. musculus** RNA-Seq dataset SRX648736
39179 (22740) genes, 94929 (47618) isoforms

<table>
<thead>
<tr>
<th></th>
<th>Trans-ABYSS</th>
<th>IDBA-tran</th>
<th>SOAPdenovo Trans</th>
<th>Trinity</th>
<th>SPAdes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcripts</td>
<td>61508</td>
<td>38294</td>
<td>47025</td>
<td>51245</td>
<td>48706</td>
</tr>
<tr>
<td>Aligned</td>
<td><strong>59666</strong></td>
<td>38201</td>
<td>46891</td>
<td>51121</td>
<td>47979</td>
</tr>
<tr>
<td>Unaligned</td>
<td>1842</td>
<td><strong>93</strong></td>
<td>134</td>
<td>124</td>
<td>727</td>
</tr>
<tr>
<td>Gene database coverage, %</td>
<td>15.2</td>
<td>16.8</td>
<td>13.1</td>
<td><strong>18.2</strong></td>
<td>17.6</td>
</tr>
<tr>
<td>Partially-assembled isoforms (&gt;50%)</td>
<td>5824</td>
<td>6804</td>
<td>4608</td>
<td><strong>7089</strong></td>
<td>6998</td>
</tr>
<tr>
<td>Fully-assembled isoforms (&gt;95%)</td>
<td>1552</td>
<td>1599</td>
<td>877</td>
<td>2053</td>
<td><strong>2315</strong></td>
</tr>
<tr>
<td>Misassemblies</td>
<td>692</td>
<td>378</td>
<td>21</td>
<td>320</td>
<td>817</td>
</tr>
<tr>
<td>Avg. mismatches per transcript</td>
<td>0.5</td>
<td>0.9</td>
<td><strong>0.4</strong></td>
<td>1.1</td>
<td>0.9</td>
</tr>
</tbody>
</table>
From SPAdes to rnaSPAdes

- Raw reads
  - Bayes-Hammer
  - Corrected reads

SPAdes:
- De Bruijn Graph construction
  - De Bruijn graph
  - Graph simplification
    - Assembly graph
  - Paired read alignment
    - Paired index
  - Repeat resolution & scaffolding

Output: Contigs & scaffolds
From SPAdes to rnaSPAdes

Raw RNA-Seq reads

Bayes-Hammer

Corrected reads

rnaSPAdes

De Bruijn Graph construction

De Bruijn graph

Graph simplification

Assembly graph

Paired read alignment

Paired index

Isoform identification

Transcripts
## Results

*M. musculus* RNA-Seq dataset SRX648736

39179 (22740) genes, 94929 (47618) isoforms

<table>
<thead>
<tr>
<th></th>
<th>Trans-ABySS</th>
<th>IDBA-tran</th>
<th>SOAPdenovo Trans</th>
<th>Trinity</th>
<th>SPAdes</th>
<th>rnaSPAdes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcripts</td>
<td>61508</td>
<td>38294</td>
<td>47025</td>
<td>51245</td>
<td>48706</td>
<td>34615</td>
</tr>
<tr>
<td>Aligned</td>
<td>59666</td>
<td>38201</td>
<td>46891</td>
<td>51121</td>
<td>47979</td>
<td>34381</td>
</tr>
<tr>
<td>Unaligned</td>
<td>1842</td>
<td>93</td>
<td>134</td>
<td>124</td>
<td>727</td>
<td>234</td>
</tr>
<tr>
<td>Gene database coverage, %</td>
<td>15.2</td>
<td>16.8</td>
<td>13.1</td>
<td><strong>18.2</strong></td>
<td>17.6</td>
<td>16.7</td>
</tr>
<tr>
<td>Partially-assembled isoforms (&gt;50%)</td>
<td>5824</td>
<td>6804</td>
<td>4608</td>
<td><strong>7089</strong></td>
<td>6998</td>
<td>6916</td>
</tr>
<tr>
<td>Fully-assembled isoforms (&gt;95%)</td>
<td>1552</td>
<td>1599</td>
<td>877</td>
<td>2053</td>
<td>2315</td>
<td><strong>2344</strong></td>
</tr>
<tr>
<td>Misassemblies</td>
<td>692</td>
<td>378</td>
<td><strong>21</strong></td>
<td>320</td>
<td>817</td>
<td>527</td>
</tr>
<tr>
<td>Avg. mismatches per transcript</td>
<td>0.5</td>
<td>0.9</td>
<td><strong>0.4</strong></td>
<td>1.1</td>
<td>0.9</td>
<td>1.1</td>
</tr>
</tbody>
</table>
Transcripts with low coverage

- genome
- exon
Transcripts with low coverage

genome

exon
Transcripts with low coverage

ASSEMBLY
### Results on strand-specific data

**S.cerevisiae** RNA-Seq dataset SRR1920959  
7126 genes, 7126 isoforms

<table>
<thead>
<tr>
<th></th>
<th>Trans-ABysS</th>
<th>IDBA-tran</th>
<th>SOAPdenovo Trans</th>
<th>Trinity</th>
<th>rnaSPAdes (k=55)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcripts</td>
<td>117032</td>
<td>6914</td>
<td>18198</td>
<td>49133</td>
<td>23181</td>
</tr>
<tr>
<td>Aligned</td>
<td>108075</td>
<td>6885</td>
<td>17914</td>
<td>49002</td>
<td>20932</td>
</tr>
<tr>
<td>Unaligned</td>
<td>8957</td>
<td>29</td>
<td>284</td>
<td>131</td>
<td>2249</td>
</tr>
<tr>
<td>Gene database coverage, %</td>
<td>93.5</td>
<td>76.8</td>
<td>86.1</td>
<td>92.1</td>
<td>79.0</td>
</tr>
<tr>
<td>Partially-assembled isoforms (&gt;50%)</td>
<td>5747</td>
<td>4219</td>
<td>4721</td>
<td><strong>5788</strong></td>
<td>4379</td>
</tr>
<tr>
<td>Fully-assembled isoforms (&gt;95%)</td>
<td>4541</td>
<td>3152</td>
<td>4721</td>
<td><strong>4766</strong></td>
<td>3146</td>
</tr>
<tr>
<td>Misassemblies</td>
<td>183</td>
<td>35</td>
<td>13</td>
<td>365</td>
<td>35</td>
</tr>
<tr>
<td>Avg. mismatches per transcript</td>
<td><strong>0.3</strong></td>
<td>1.4</td>
<td>0.6</td>
<td>1.3</td>
<td>0.8</td>
</tr>
</tbody>
</table>
### Results on strand-specific data

**Z. mays** RNA-Seq dataset SRR1588569

39625 genes, 63391 isoforms

<table>
<thead>
<tr>
<th></th>
<th>Trans-ABYSS</th>
<th>IDBA-tran</th>
<th>SOAPdenovo Trans</th>
<th>Trinity</th>
<th>rnaSPAdes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcripts</td>
<td>174880</td>
<td>123295</td>
<td><strong>170645</strong></td>
<td>159058</td>
<td>91261</td>
</tr>
<tr>
<td>Aligned</td>
<td>164161</td>
<td>123293</td>
<td><strong>170581</strong></td>
<td>159055</td>
<td>91236</td>
</tr>
<tr>
<td>Unaligned</td>
<td>35</td>
<td>2</td>
<td>64</td>
<td>3</td>
<td>25</td>
</tr>
<tr>
<td>Gene database coverage, %</td>
<td>35.9</td>
<td>32.2</td>
<td>30.6</td>
<td>35.5</td>
<td>30.9</td>
</tr>
<tr>
<td>Partially-assembled isoforms (&gt;50%)</td>
<td>15529</td>
<td>9459</td>
<td>9980</td>
<td><strong>16365</strong></td>
<td>13556</td>
</tr>
<tr>
<td>Fully-assembled isoforms (&gt;95%)</td>
<td>3549</td>
<td>843</td>
<td>1499</td>
<td><strong>4387</strong></td>
<td>3047</td>
</tr>
<tr>
<td>Misassemblies</td>
<td>452</td>
<td>265</td>
<td>47</td>
<td>1741</td>
<td>1583</td>
</tr>
<tr>
<td>Avg. mismatches per transcript</td>
<td>1.4</td>
<td>1.9</td>
<td><strong>0.3</strong></td>
<td>3.2</td>
<td>1.4</td>
</tr>
</tbody>
</table>
RNA-Seq QC

**RSeQC** (Wang et al., Bioinf., 2012) — a tool for reference based RNA-Seq QC

- Strand-specificity
- Insert size
- Reads distribution
- Reads duplication
- and many more
Outline

1. Who assembles transcriptomes *de novo*?
2. How do assemblers work?
3. What is SPAdes and how does it work on RNA-Seq?
4. How do we evaluate transcriptome assemblies?
You cannot develop an assembler without having an assembly quality assessment tool
You cannot develop an assembler without having an assembly quality assessment tool.

Based on our experience with SPAdes and QUAST, developing such tool is not an easy task
(Gurevich et al., Bioinf., 2013)
But why?
RNA-Seq assembly evaluation

- RNA-Seq assemblers benchmark (Martin & Zhang, Nat. Rev. Genetics., 2011)
- Evaluating 454 RNA-Seq assemblies (Mundry et al., PloS one, 2012)
- RGASP-consortium (Steijger et al., Nat. Met., 2013)
- Metrics assessment (O’Neil et al., BMC, 2013)
- Detonate RSEM-EVAL/REF-EVAL (Li et al., Gen. Biol., 2014)
RNA-Seq assembly evaluation

Every paper describing new transcriptome assembler uses its own metrics and methods.

No tool yet have set a standard for assessing quality of transcriptome assemblies.
Misassembly search

- \( \{a_1, \ldots, a_n\} \) — alignments of the transcript \( T \)
- \( A_j = (a_{k1}, a_{k2}, \ldots, a_j) \) — alignment union ending at \( a_j \)
- Best alignment union is detected via dynamic programming algorithm
  - \( S(A_j) = \max \{ \text{score}(A_i, a_j) \} \) for all \( i \) such that
    - last position of \( a_i \) - first position of \( a_j < \Delta \)
  - Score includes mismatches, indels and alignment locality
- A transcript with 2 or more discordant partial alignments from best alignment union is considered as misassembly candidate
BLAT alignment

Genome

A  B  C

Transcript

A’  B’  C’

A”  B”  C”

Alignment to the reference:

Expectation

A”  B”

Reality

A”  B”  C”
BLAST alignment

Genome

Transcript:

Expectation

Reality

Alignment to the reference
De novo evaluation

- Core gene detection
  - CEGMA (Parra et al., Bioinf., 2007, no longer supported)
  - BUSCO (Simão et al., Bioinf., 2015)
De novo evaluation

● Core gene detection
  o CEGMA (Parra et al., Bioinf., 2007, no longer supported)
  o BUSCO (Simão et al., Bioinf., 2015)

● De novo gene finding
  o GeneMarkS-T (Tang et al., Nucl. ac. res., 2015)
De novo evaluation

- Core gene detection
  - CEGMA (Parra et al., Bioinf., 2007, no longer supported)
  - BUSCO (Simão et al., Bioinf., 2015)

- De novo gene finding
  - GeneMarkS-T (Tang et al., Nucl. ac. res., 2015)

- Quality assessment using initial reads
  - Detonate RSEM-EVAL (Li et al., Gen. Bio., 2014)
Acknowledgements

Dmitry Antipov
Elena Bushmanova
SPAdes team
Pavel Pevzner
Alla Lapidus
Vladimir Suvorov