



SPAdes family of tools for genome assembly and analysis

Dmitry Antipov*, Anton Bankevich*, Elena Bushmanova*, Aleksey Gurevich*, Anton Korobeynikov*,
Dmitry Meleshko*, Sergey Nurk*, Andrei Przhibelski*, Yana Safonova*, Alla Lapidus*, Pavel Pevzner*,**

*Centre for Algorithmic Biotechnology, Saint Petersburg State University, Saint Petersburg, Russia

**University of California San Diego, San Diego, USA

Dip

dipSPAdes:
The first de Bruijn graph assembler designed for highly polymorphic diploid genomes (heterozygosity > 1%)

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 haplotypes and takes advantage of structure of de Bruijn graph for diploid genome to construct longer contigs

Yana Safonova, Anton Bankevich, Pavel A. Pevzner.
dipSPAdes: an assembler for highly polymorphic diploid genomes. *J. of Comp. Biol.*, 2015

RNA

rnaSPAdes:
Yet another transcriptome assembler. Brings the wisdom of single-cell assembler into RNA-seq world.

How well a *genome* assembler would perform on a transcriptome dataset?.. Quite well! And we can only make it better:

| | IDBA-tran | SOAPdenovoTrans | Trinity | SPAdes | rnaSPAdes |
|--|-----------|-----------------|--------------|-------------|--------------|
| Transcripts | 2872 | 2725 | 2171 | 3339 | 6954 |
| N50 | 312 | 213 | 309 | 370 | 303 |
| Aligned | 2845 | 2693 | 2150 | 3230 | 6692 |
| Unaligned | 27 | 32 | 21 | 109 | 262 |
| Avg. mismatches per transcript | 0.447 | 0.456 | 0.341 | 0.57 | 0.35 |
| Total annotation coverage | 0.075 | 0.052 | 0.058 | 0.1 | 0.105 |
| Partially-assembled isoforms (>30%) | 886 | 582 | 713 | 1119 | 1135 |
| Fully-assembled isoforms (>90%) | 96 | 53 | 91 | 234 | 188 |
| Partially-annotated transcripts (>30%) | 2611 | 2493 | 2009 | 2967 | 6138 |
| Fully-annotated transcripts (>90%) | 1436 | 1449 | 1108 | 1553 | 4094 |

Yeast RNA-seq data

RNA-seq toppings for SPAdes:

- RNA-seq-aware graph simplification
- Isoform detection
- rnaQUAST
- and many more...

Tru

truSPAdes:
An assembler for Illumina TruSeq Synthetic Long-Read barcoded data

Barcode assembly challenges:

All the issues SPAdes was designed to deal with!

Human TSLR assembly

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

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



Download SPAdes at <http://bioinf.spbau.ru/spades>

Meta

metaSPAdes:
SPAdes tuned for the complexity and sizes of metagenomic datasets

Genome assembly of species with extremely different abundances is similar to genome assembly of MDA data. metaSPAdes borrows some ideas from dipSPAdes in order to separate close species.

The tool is in the early development phase, however already shows promising results as compared to SPAdes and IDBA-UD

SRX024329 (HMP) Nx plot