

SPAdes: a New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing

St. Petersburg Academic University

Anton Bankevich
Sergey Nurk
Dmitry Antipov
Alexey Gurevich
Mikhail Dvorkin
Alexander Kulikov
Valery Lesin
Sergey Nikolenko
Andrey Prjibelski
Alexey Pyshkin
Alexander Sirotkin
Nikolay Vyahhi

University of California, San Diego

Pavel Pevzner
Son Pham
Glenn Tesler

University of South Carolina

Max Alekseyev

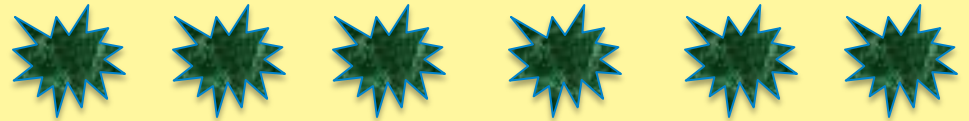
Funding

Russian Federation grant
11.G34.31.0018
NIH 3P41RR024851-02S1

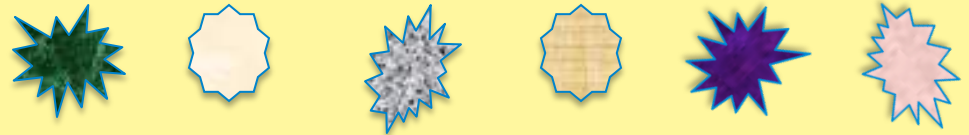
Outline

- Genome sequencing

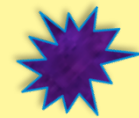
- Conventional



- Metagenomics



- Single Cell



- De Bruijn graphs and SPAdes

- Results on *E. coli* and an uncultivated marine genome

Whole Genome Shotgun Sequencing

Multiple (Unsequenced) Genome Copies



Read Generation

Reads



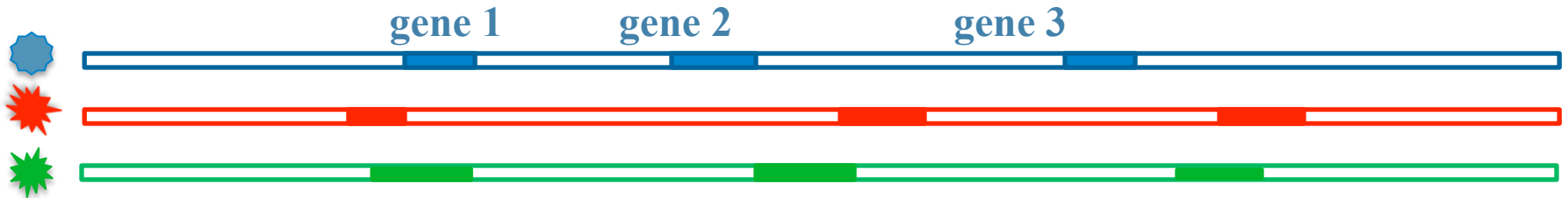
Fragment Assembly

Sequenced Genome

...GGCATGCGTCAGAACTATCATAGCTAGATCGTACGTAGCC...

From Metagenomics to Single Cell Sequencing

- Traditional microbial genome sequencing requires isolating a pure strain and reproducing it in a ‘culture’ under controlled laboratory conditions. But >99% of bacteria cannot be cultured.
- *Metagenomics* enables studies of organisms not easily cultured in a laboratory. It uses collective sequencing of non-identical cells.
- Until recently, **metagenomics** was the only option for studies of microbial communities. However, metagenomics provides information about only a **few genes** (across many species).



From Metagenomics to Single Cell Sequencing

- Traditional microbial genome sequencing requires isolating a pure strain and reproducing it in a ‘culture’ under controlled laboratory conditions. But >99% of bacteria cannot be cultured.
- *Metagenomics* enables studies of organisms not easily cultured in a laboratory. It uses collective sequencing of non-identical cells.
- Until recently, **metagenomics** was the only option for studies of microbial communities. However, metagenomics provides information about only a **few genes** (across many species).



From Metagenomics to Single Cell Sequencing

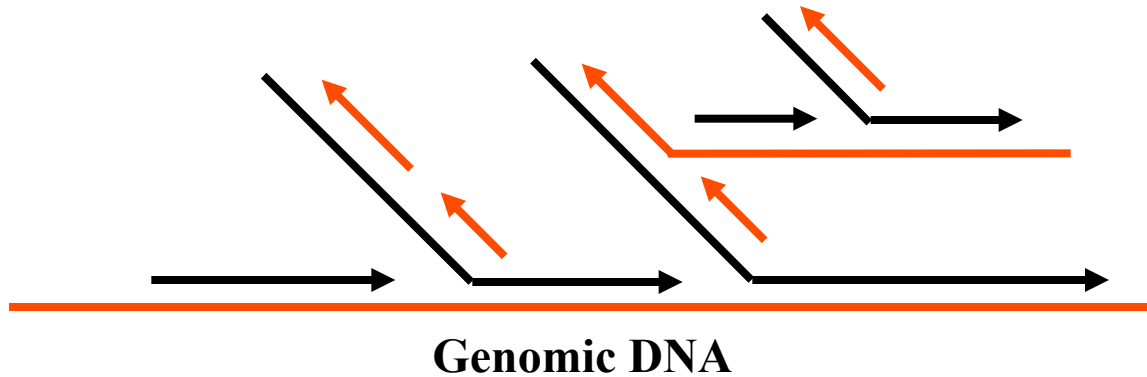
- Traditional microbial genome sequencing requires isolating a pure strain and reproducing it in a ‘culture’ under controlled laboratory conditions. But >99% of bacteria cannot be cultured.
- *Metagenomics* enables studies of organisms not easily cultured in a laboratory. It uses collective sequencing of non-identical cells.
- **Single Cell Bacterial Genomics:** Complementing **gene-centric** metagenomics data with **whole-genome** assembly of uncultivated organisms.

1000s of genes sequenced from a single cell



Single Cell Sequencing via MDA:

Multiple Displacement Amplification



1. Random hexamer primers
2. Phi29 DNA polymerase Strand displacing
3. Isothermal reaction (30°C)

F.B. Dean, J.R. Nelson, T.L. Giesler, R.S. Lasken (2001). *Genome Res.* 11:1095-9
F.B. Dean, S. Hosono, L. Fang, et al. (2002). *PNAS* 99:5261-6

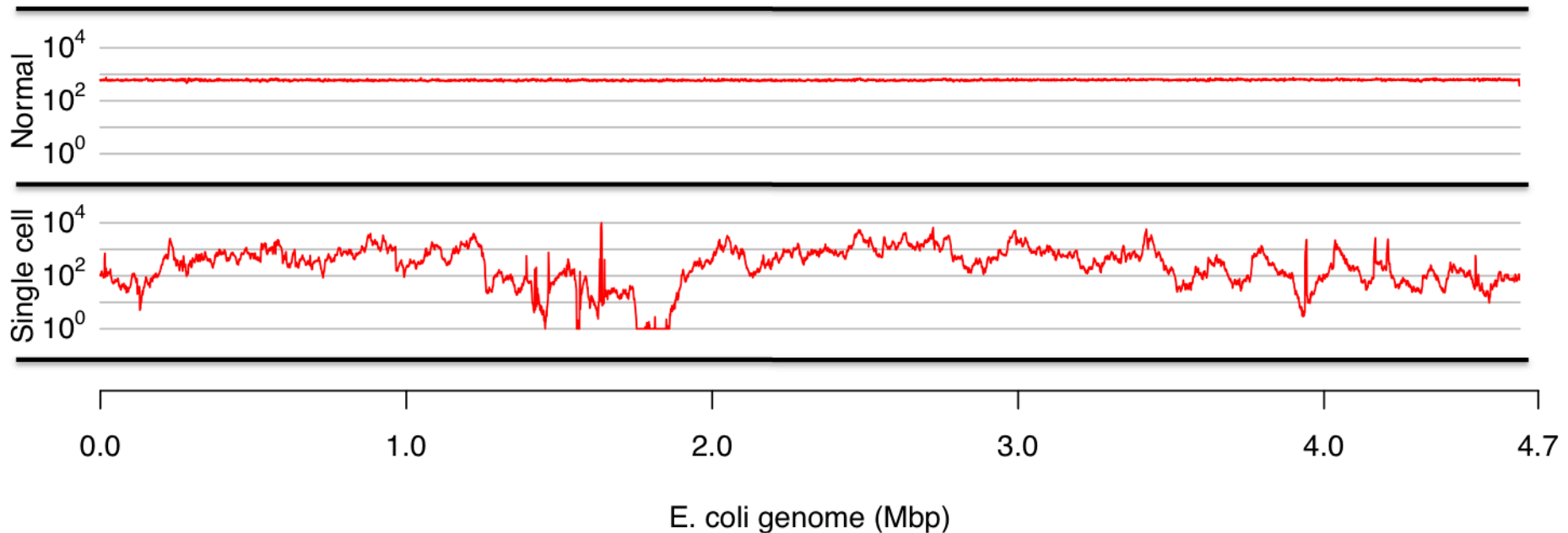
- Roger Lasken's lab developed *Multiple Displacement Amplification* (MDA).
- More effective than PCR for amplification of a single cell.
- TempliPhi and GenomiPhi (GE Healthcare) and REPLI-g (Qiagen).
- REPLI-g: fragments ~ 2 – 100 kb; usually > 10 kb on average.

Sequencing Coverage

Normal multicell vs. single cell *E. coli*

Illumina GA IIX paired-end sequencing, 100 bp reads, ~ 600x coverage

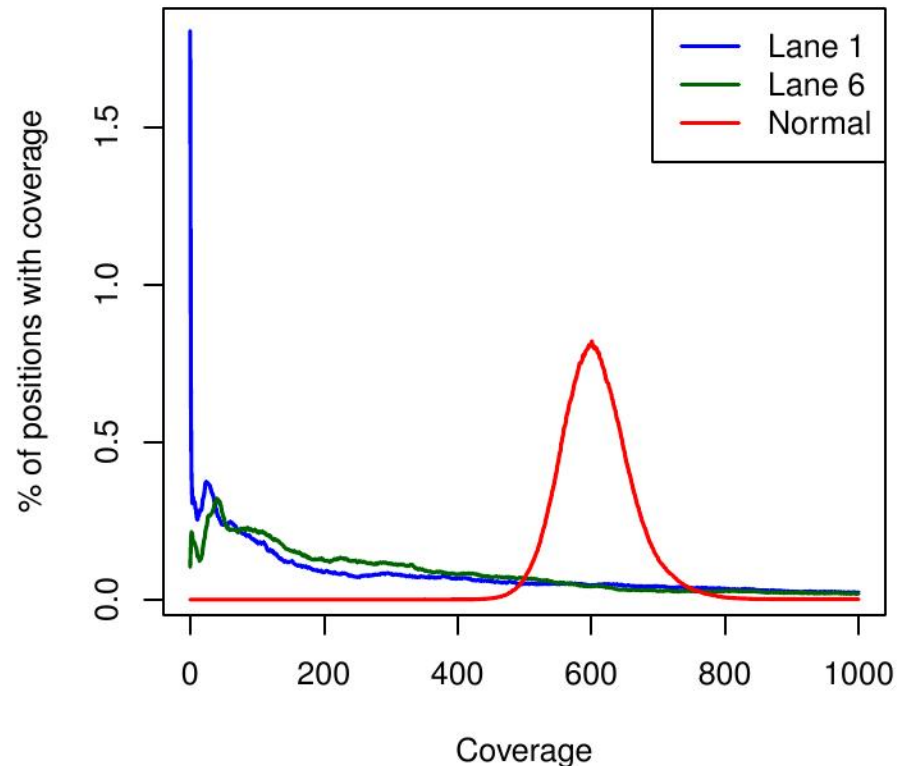
Coverage



- Lander-Waterman model predicts ~15x coverage needed for complete *E. coli* assembly.
- Assumes uniform coverage; error-free reads; and no repeats in genome.
- For our single cell *E. coli* assembly, 600x average coverage still has some gaps since there are positions with no reads.

Distribution of Coverage

Empirical distribution of coverage

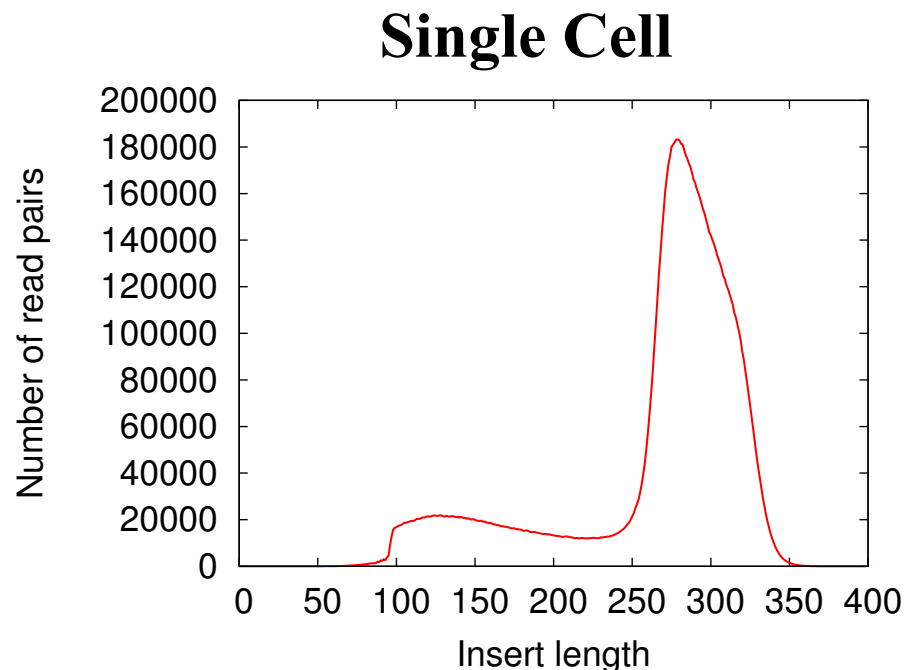
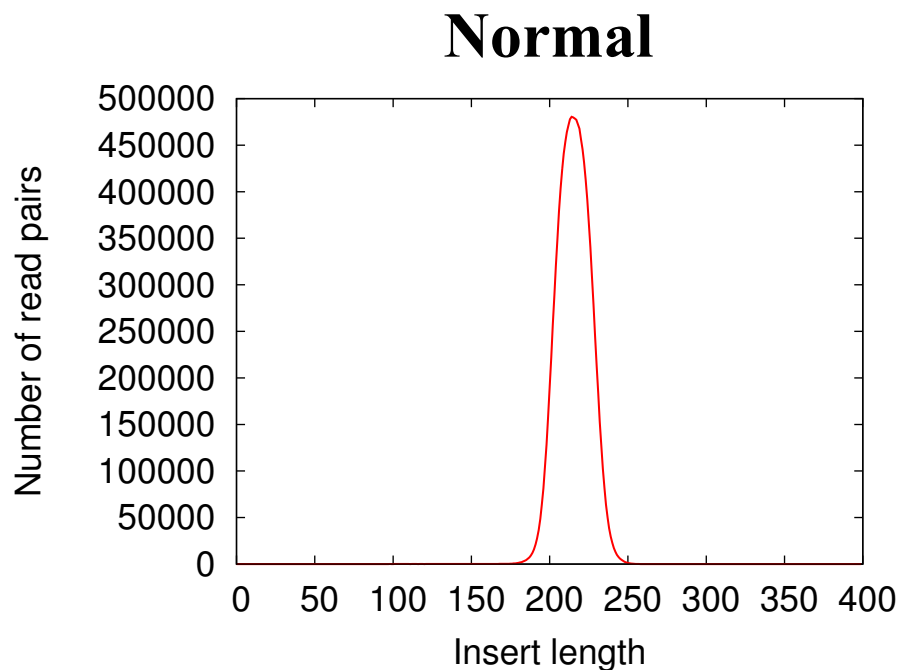


A cutoff threshold will eliminate about 25% of valid data in the single cell case, whereas it eliminates noise in the normal multicell case.

Chitsaz, et al., *Nat. Biotechnol.* (2011).

Insert Size Distribution

Illumina GA IIx sequencing of *E. coli*, 600x coverage

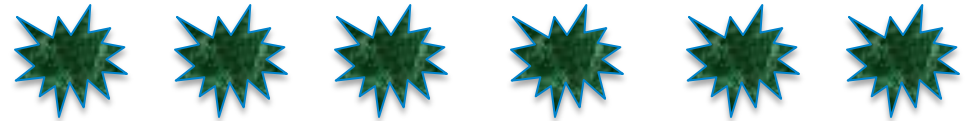


Chitsaz, et al., *Nat. Biotechnol.* (2011).

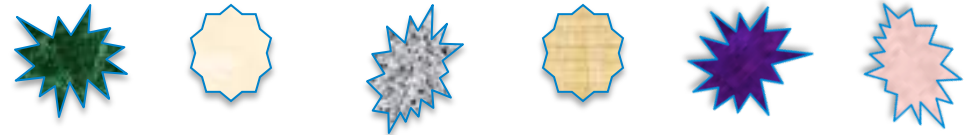
Outline

- Genome sequencing

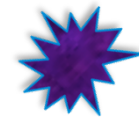
- Conventional



- Metagenomics



- Single Cell



- De Bruijn graphs and SPAdes

- Results on *E. coli* and an uncultivated marine genome

De Bruijn Graph for Genome Assembly

- Introduced by Pavel Pevzner in 1989 (based on DNA arrays)

P.A. Pevzner, *J Biomol Struct Dyn* (1989) 7:63–73.

R. Idury, M. Waterman, *J Comput Biol* (1995) 2:291–306.

- Adapted to Sanger sequencing (EULER) and 2nd generation sequencing (EULER-SR).

P.A. Pevzner, H. Tang, & M.S. Waterman, *PNAS* (2001) 98(17):9748-9753.

P.A. Pevzner, H. Tang, H. & G. Tesler, *Genome Res.* (2004) 14:1786-1796.

M.J. Chaisson & P.A. Pevzner, *Genome Res.* (2008) 18:324-330.

- Used in many other short-read assemblers.

Velvet: D.R. Zerbino, & E. Birney, *Genome Res.* (2008) 18:821-829.

ALLPATHS: Butler et al, *Genome Res.* (2008) 18(5):810-820.

ABYSS: Simpson et al, *Genome Res.* (2009) 19:1117-1123.

SOAPdenovo: Li et al, *Genome Res.* (2010) 20(2): 265-272.

De Bruijn Graph of a Genome

Toy example: shred genome into 3-mer vertices, 4-mer edges

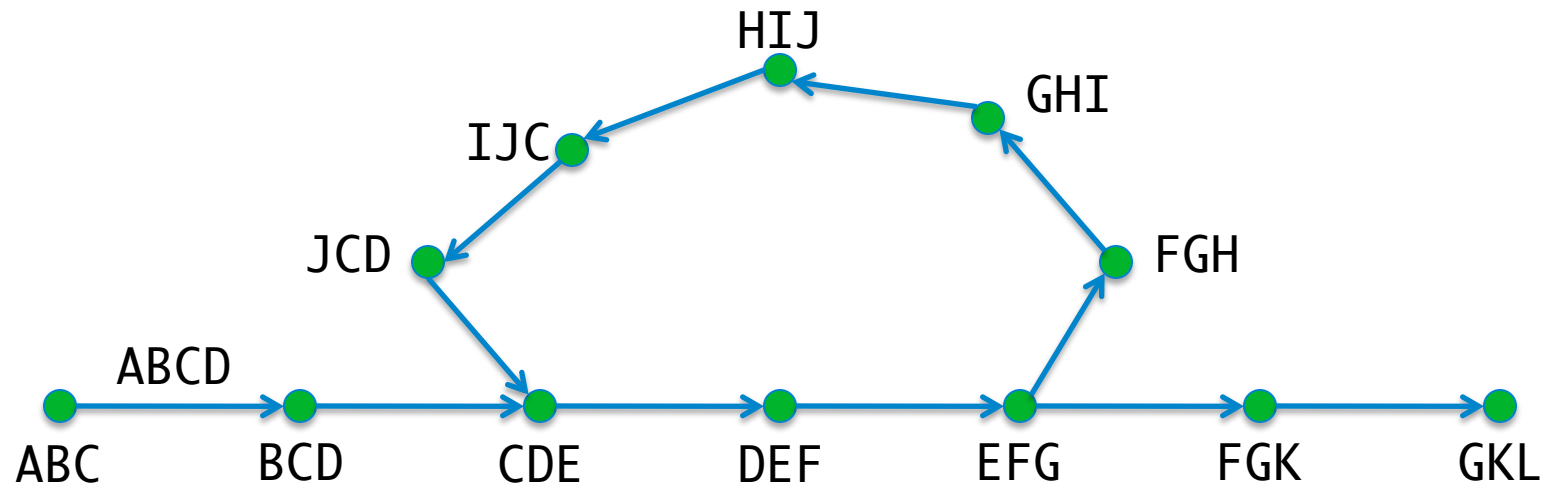
Vertices: k -mers from the sequence

Edges: $(k+1)$ -mers from the sequence

$k=3$: 4-mer $wxyz$ gives $wxy \rightarrow xyz$

Genome: Eulerian path through graph
(using edge multiplicities)

Genome: **ABCDEF**GHIJ**CDEF**GKL



P. Pevzner, *J Biomol Struct Dyn* (1989) 7:63–73

R. Idury, M. Waterman, *J Comput Biol* (1995) 2:291–306

P. Pevzner, H. Tang, M. Waterman, *PNAS* (2001) 98(17):9748–53

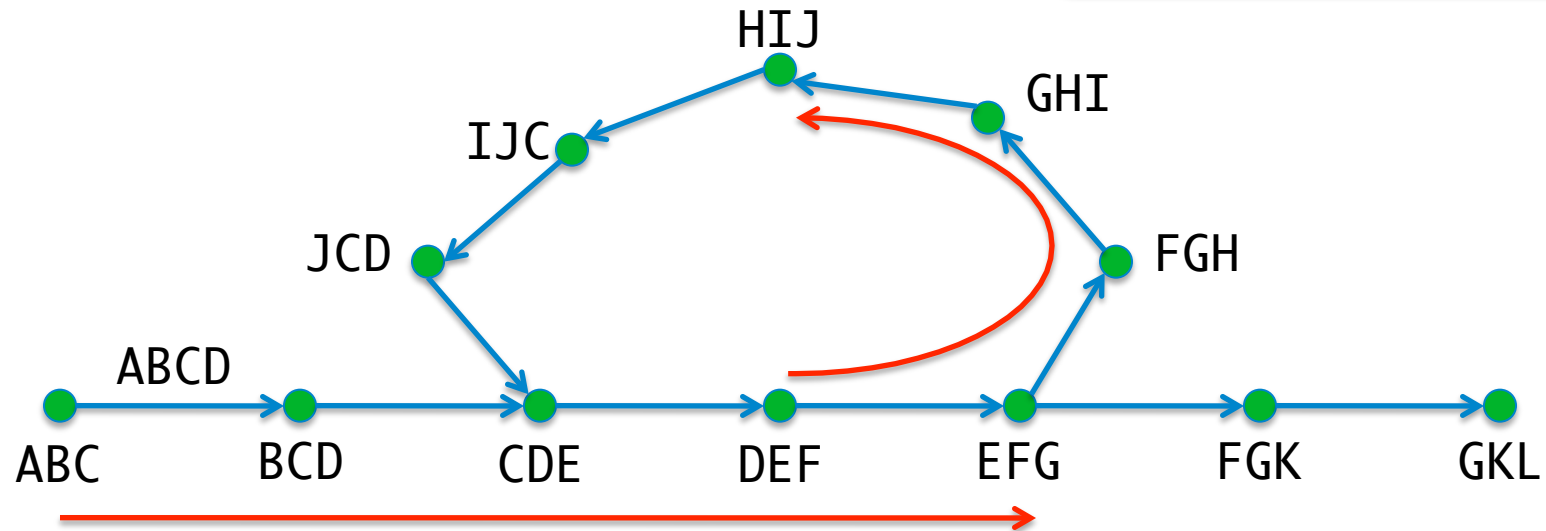
Same De Bruijn Graph from Perfect Reads

Toy example: shred reads into 3-mer vertices, 4-mer edges

Vertices: k -mers from the reads
Edges: $(k+1)$ -mers from the reads
 $k=3$: 4-mer $wxyz$ gives $wxy \rightarrow xyz$
Reads: short paths through graph (red)
Genome: long path through graph

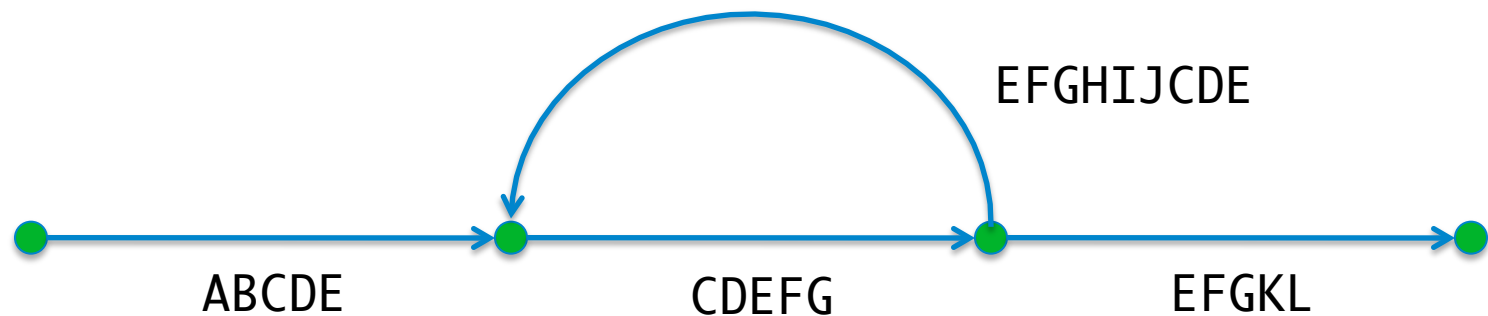
Reads (but order would be random in real data):

ABCDEF
DEFGHIJ
GHIJCDE
IJCDEFG
CDEFGKL



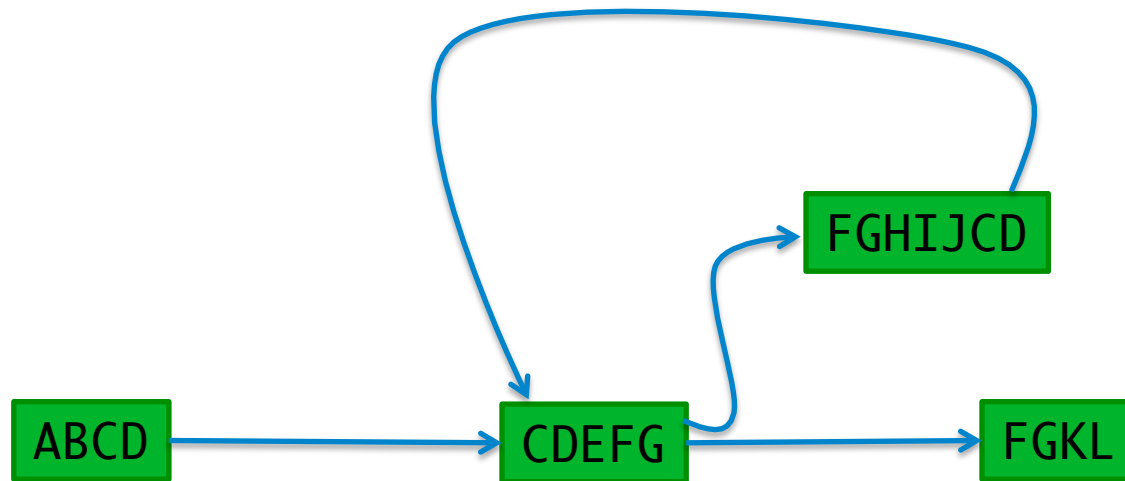
Condensed graph

Toy example: 3-mer vertices, long edges=contigs

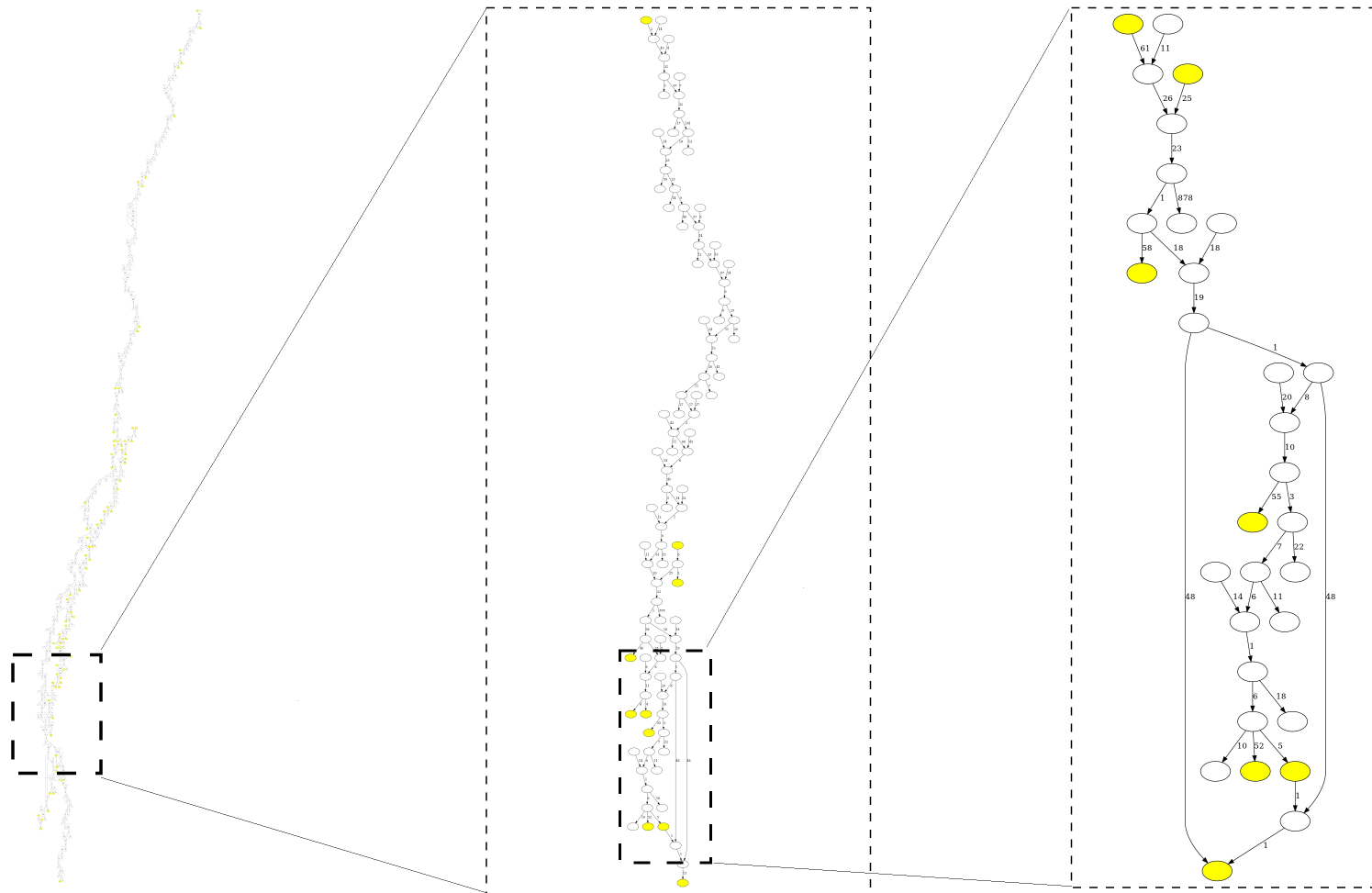


Condensed graph

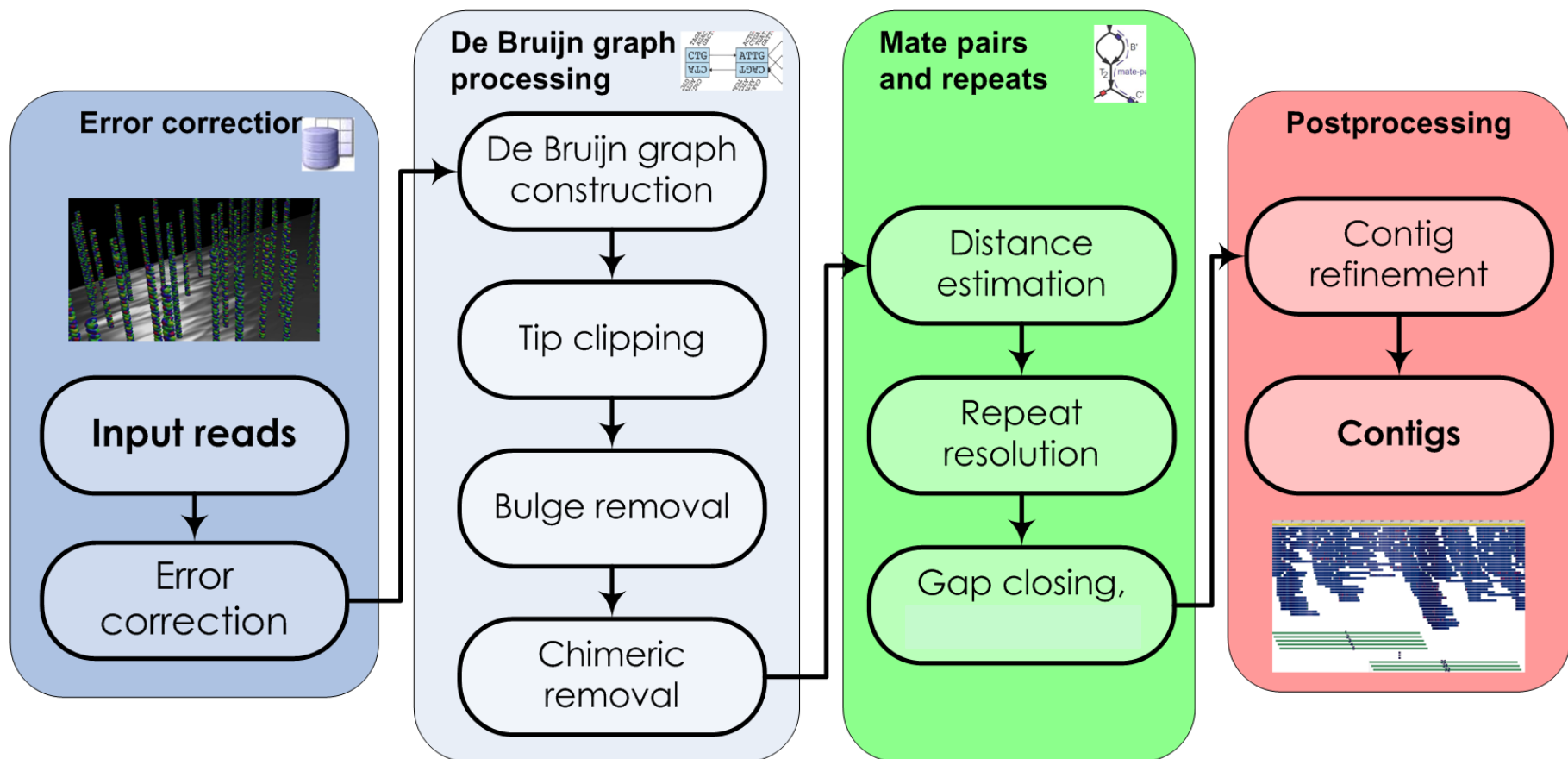
Toy example: vertices=contigs, edges=2-mer overlaps



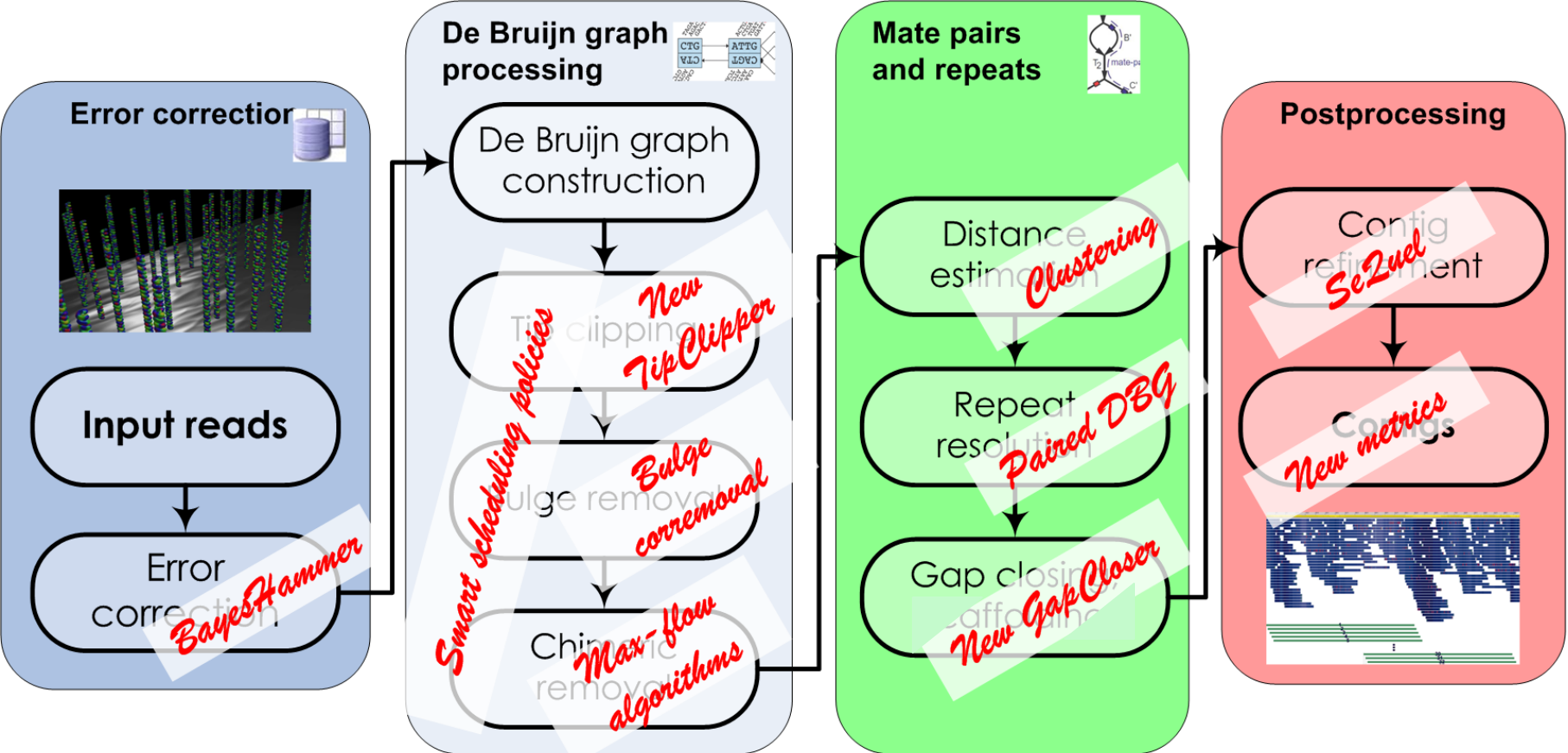
Read errors and imperfect repeats lead to a complicated graph



SPAdes genome assembler



SPAdes genome assembler

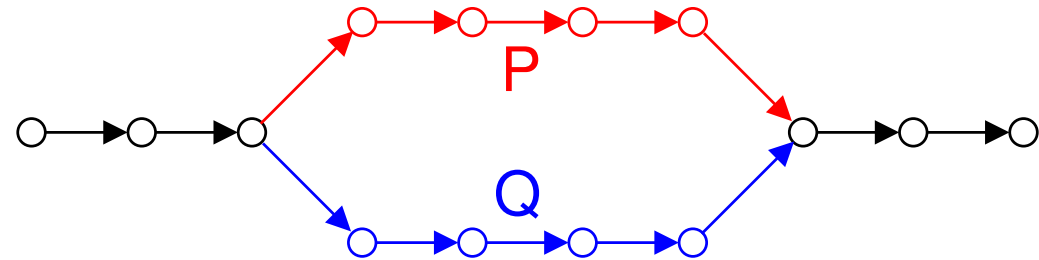


Graph Simplification in SPAdes

Bulge from error in middle of read

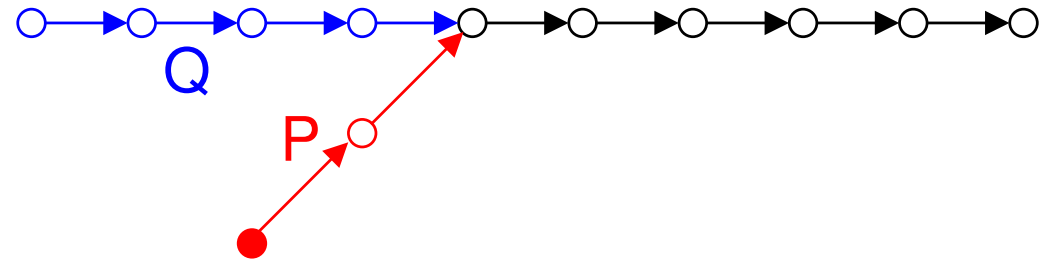
```
TCGGTGAAAGAGCTTT
CGGTGAACGAGCTTTG
GGTGAAAGAGCTTTGA
GTGAAAGAGCTTTGAT
```

P: Erroneous edges Q: correct alternative



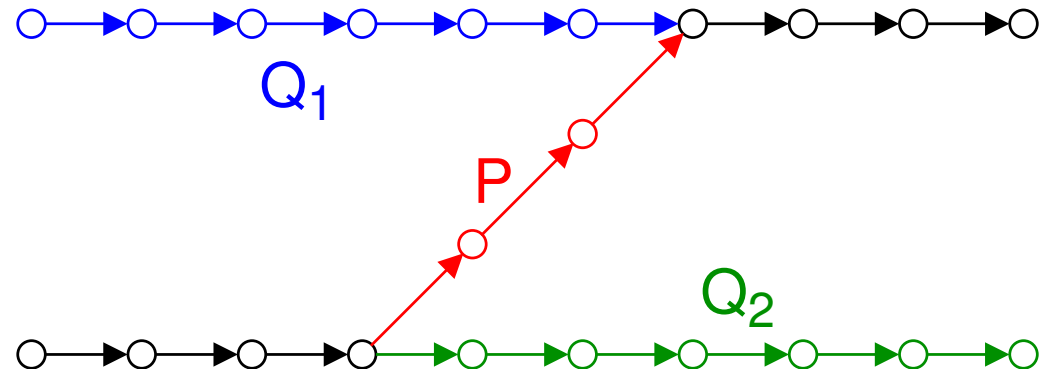
Tip from error near start/end of read

```
TCGGTGAAAGAGCTTT
CGCTGAAAGAGCTTTG
GGTGAAAGAGCTTTGA
GTGAAAGAGCTTTGAT
```



Chimeric connection joining two distant parts of genome

```
TCGGTGAAAGAGCTTT
CGGTGAAAGAGCTTTG
ACATCGTAAGCTTTGC
TCGTAGTAGCCGATTC
CGTAGTAGCCGATTCG
```

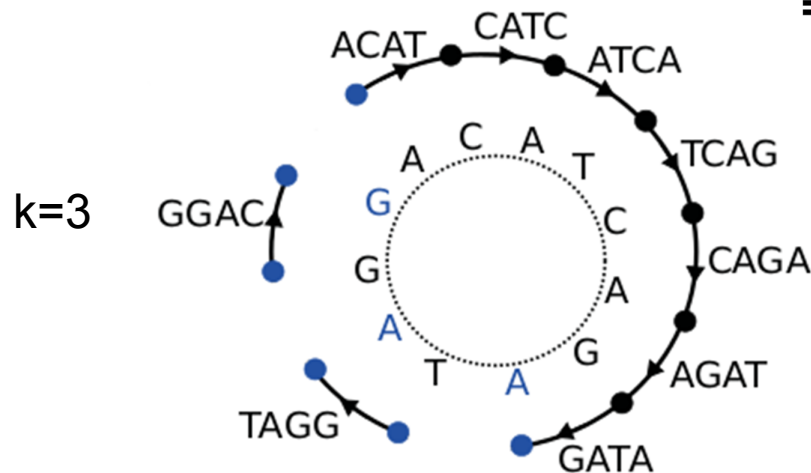
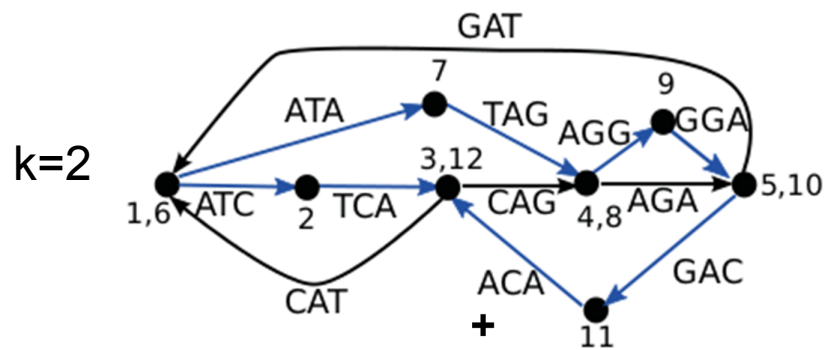


Graph Simplification in SPAdes

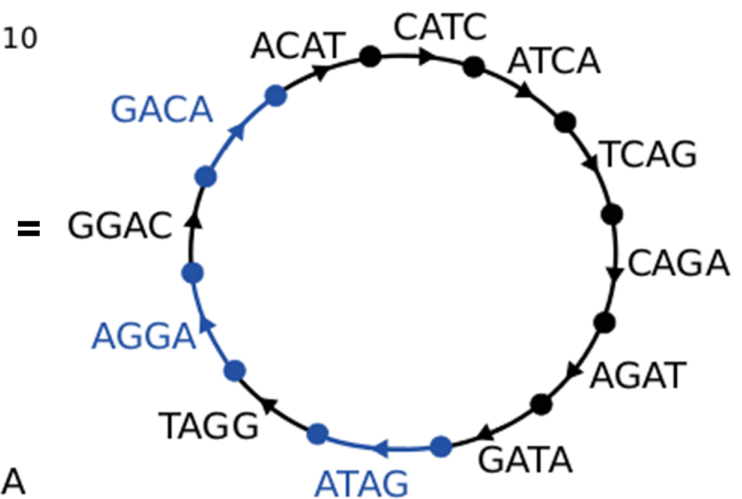
- We use local coverage, topology, and lengths to decide how to simplify the graph.
- **Smart scheduling:** For bulges and chimeric connections, SPAdes examines all edges in order from lowest to highest coverage. For tips, we go in order by length. This is inspired by, but improves upon, E+V-SC (Chitsaz et al, 2011), which used a gradually increasing threshold to discard low-coverage k-mers.
- **Efficient bookkeeping** allows us to map all reads to the final contigs using the actual logic of graph simplification, and produce an accurate SAM file placing reads onto contigs, instead of relying on external alignment tools to guess how the reads were mapped.

Multisized de Bruijn graph

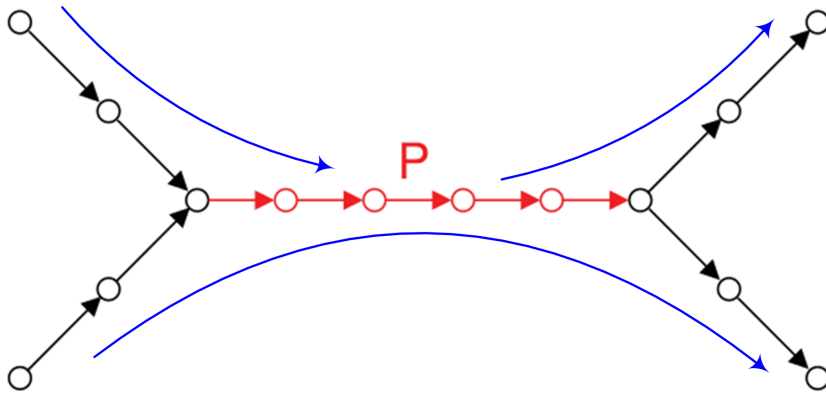
- Smaller values of k make the graph more connected but more tangled.
- Larger values of k make the graph less tangled but less connected.
- SPAdes combines multiple values of k to get the best of all worlds.
- Also see IDBA (Peng et al., 2010).



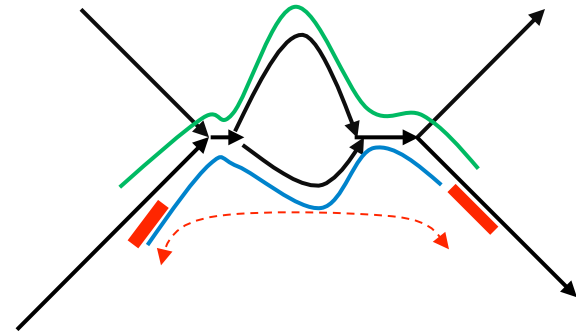
Multisized de Bruijn graph for $k=2,3$



Repeats



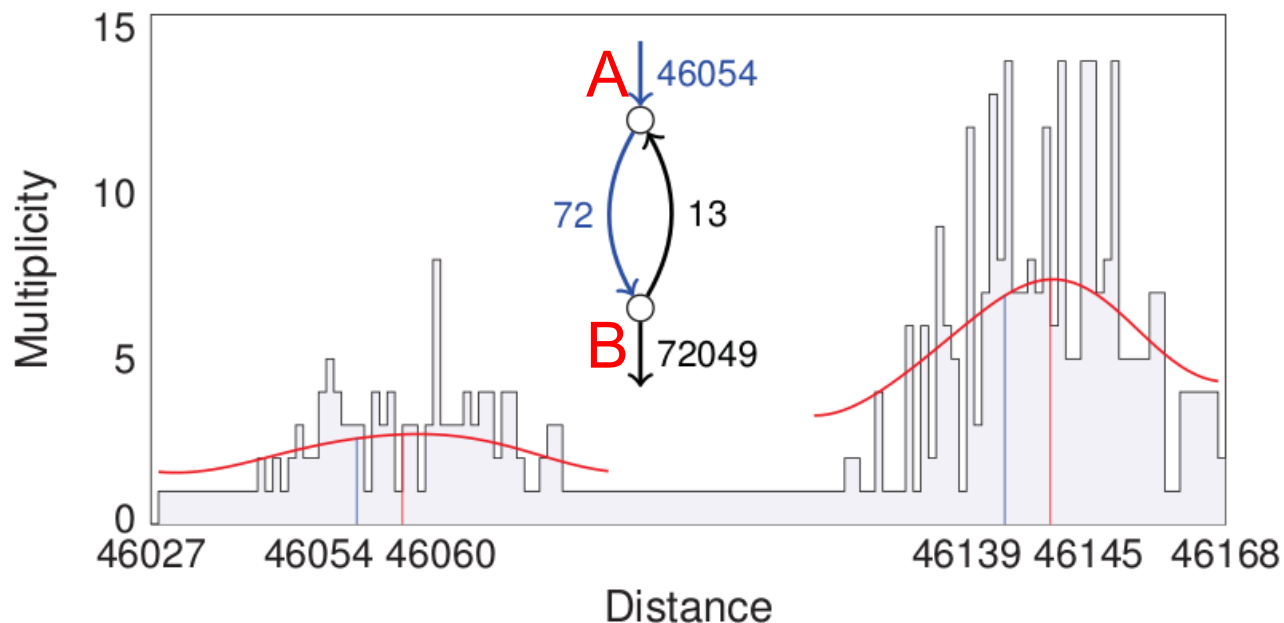
Many repeats can be resolved using either single reads (bottom) or paired reads (top), but it depends on repeat length, read length, and insert size.



Is the correct path between red reads **short** (passing through lower edge) or **long** (passing through upper edge)?

Paired reads

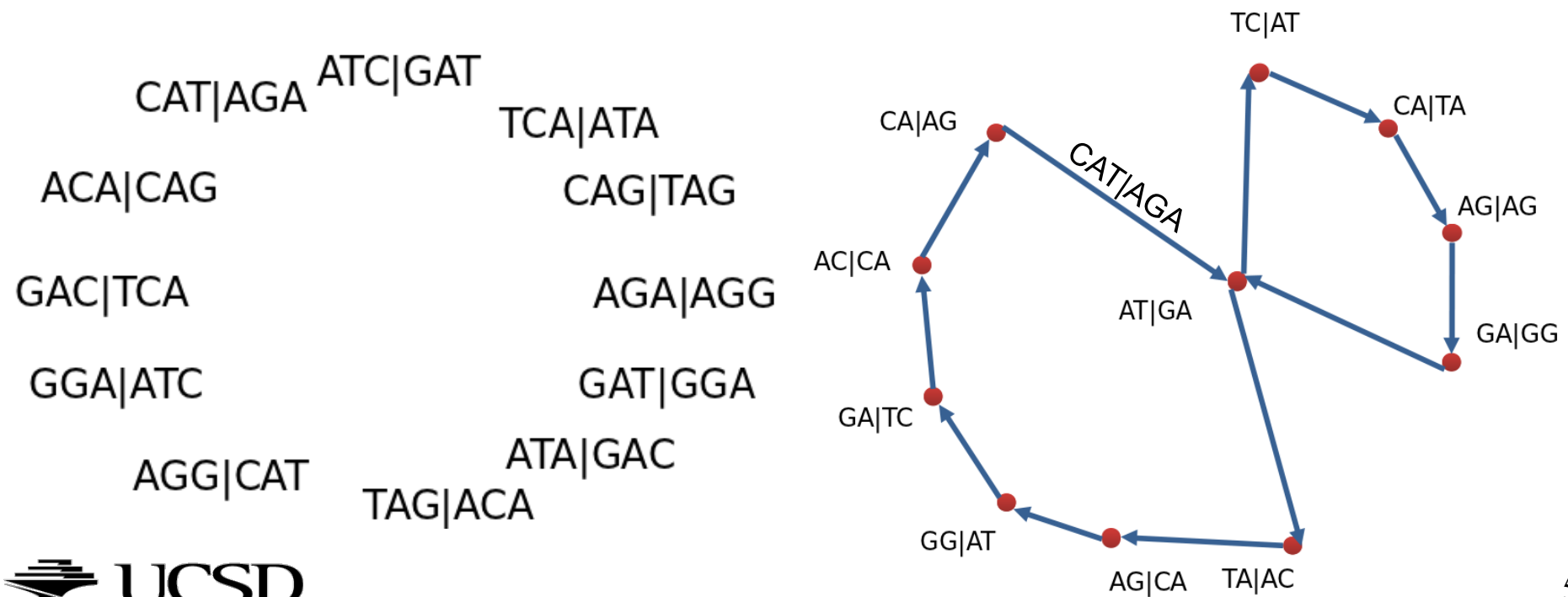
- Insert size varies in different read pairs.
- The genomic distance between two edges can be estimated when they are linked by many read pairs.
- Edges A & B could be separated by 72 bp, or by $72 + (13+72)$, etc. A distance histogram (using nominal insert sizes) reveals the actual distances and lets us correct for insert size variation.



Repeats and paired de Bruijn Graph

P. Medvedev, S. Pham, M. Chaisson, G. Tesler, P. Pevzner,
J Comput Biol (2011) 18(11):1625-1634

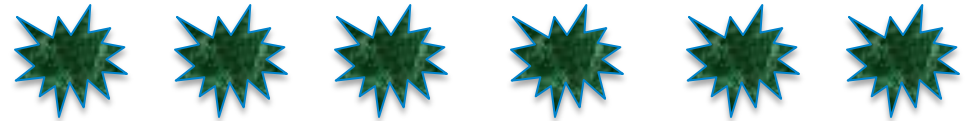
- The *paired de Bruijn graph* generalizes de Bruijn graphs to paired reads.
- Vertices are pairs of k-mers at a fixed distance (after adjusting for small variations in insert size).
- Graph is much sparser than the normal de Bruijn graph, which helps resolve repeats.
- SPAdes implements a generalization of this.



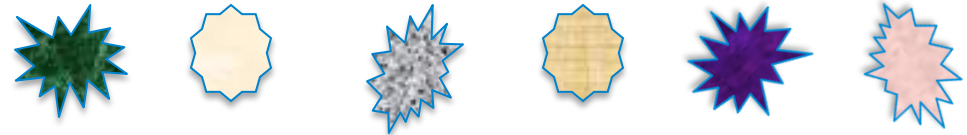
Outline

- Genome sequencing

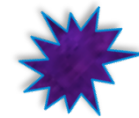
- Conventional



- Metagenomics



- Single Cell



- De Bruijn graphs and SPAdes

- Results on *E. coli* and an uncultivated marine genome

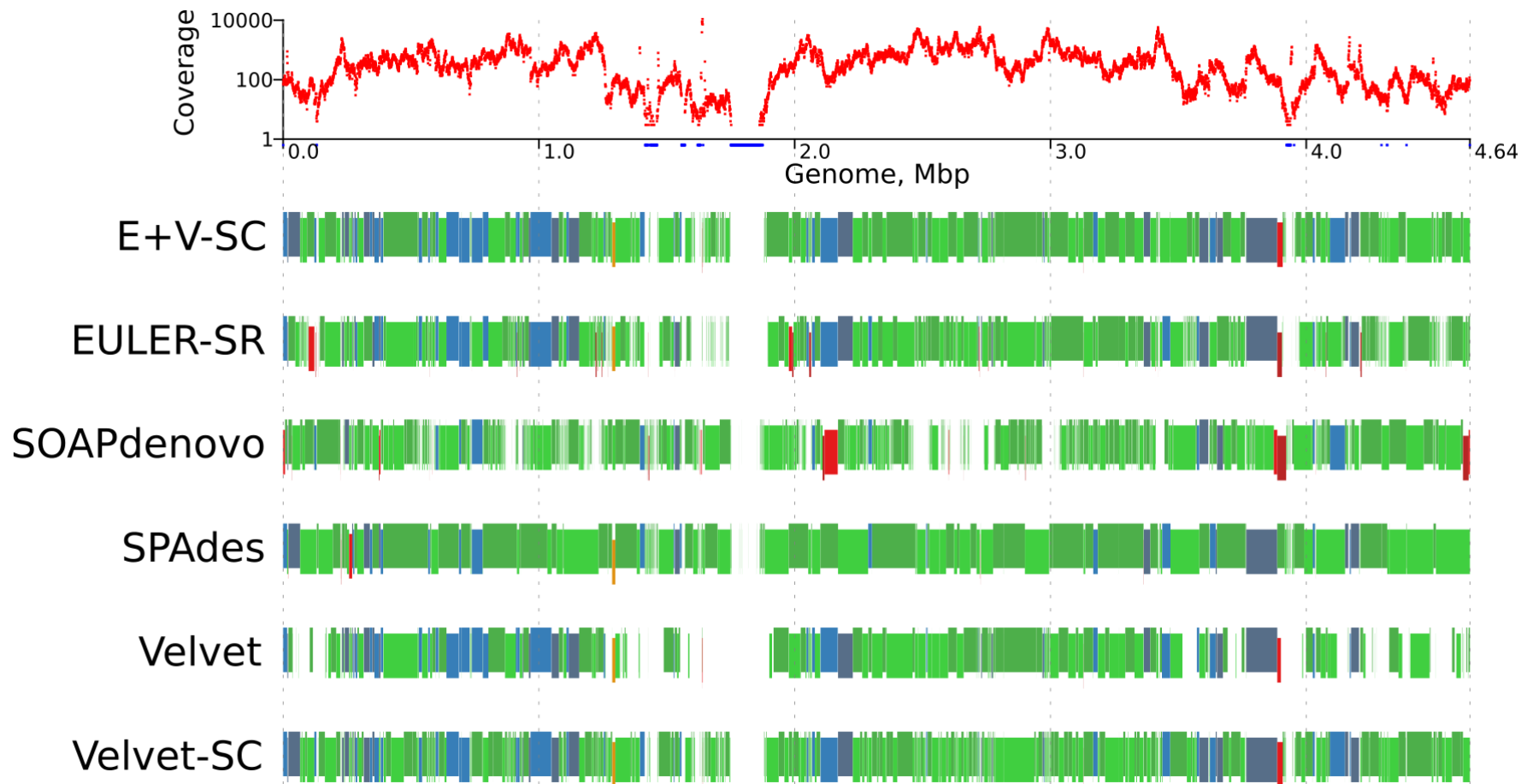
Benchmarking SPAdes:



90% of *E. coli* genes fully captured from single cell data



Table 1. Comparison of assemblies for single-cell (ECOLI-SC) and standard (ECOLI-MC) datasets.

Assembler*	# contigs	N50 (bp)	Largest (bp) [†]	Total (bp) [‡]	Covered (%) [§]	Misassemblies [¶]	Mismatches (per 100 kbp)	Complete genes (out of 4324)
Single-cell <i>E. coli</i> (ECOLI-SC)								
EULER-SR	1344	26662	126616	4369634	87.8	21	11.0	3457
SOAPdenovo	1240	18468	87533	4237595	82.5	13	99.5	3059
Velvet	428	22648	132865	3533351	75.8	2	1.9	3117
Velvet-SC	872	19791	121367	4589603	93.8	2	1.9	3654
E+V-SC	501	32051	132865	4570583	93.8	2	6.7	3809
SPAdes-single reads	1164	42492	166117	4781576	96.1	1	6.2	3888
SPAdes	1024	49623	177944	4790509	96.1	1	5.2	3911
Normal multicell sample of <i>E. coli</i> (ECOLI-MC)								
EULER-SR	295	110153	221409	4598020	99.5	10	5.2	4232
IDBA	191	50818	164392	4566786	99.5	4	1.0	4201
SOAPdenovo	192	62512	172567	4529677	97.7	1	26.1	4141
Velvet	198	78602	196677	4570131	99.9	4	1.2	4223
Velvet-SC	350	52522	166115	4571760	99.9	0	1.3	4165
E+V-SC	339	54856	166115	4571406	99.9	0	2.9	4172
SPAdes-single reads	445	59666	166117	4578486	99.9	0	0.7	4246
SPAdes	195	86590	222950	4608505	99.9	2	3.7	4268

E. coli mapped contigs (single cell)



  Correctly assembled.
Blue: similar boundaries in at least half of the assemblers.

  Misassembled.
Orange: similar boundaries in at least half of the assemblers.

New Genome assembled with E+V-SC

Deltaproteobacteria (marine bacteria) single cell assembly (2011)

ARTICLES

nature
biotechnology

Efficient *de novo* assembly of single-cell bacterial genomes from short-read data sets

Hamidreza Chitsaz^{1,6}, Joyclyn L Yee-Greenbaum^{2,6}, Glenn Tesler³, Mary-Jane Lombardo², Christopher L Dupont², Jonathan H Badger², Mark Novotny², Douglas B Rusch⁴, Louise J Fraser⁵, Niall A Gormley⁵, Ole Schulz-Trieglaff⁵, Geoffrey P Smith⁵, Dirk J Evers⁵, Pavel A Pevzner¹ & Roger S Lasken²

Whole genome amplification by the multiple displacement amplification (MDA) method allows sequencing of DNA from single cells of bacteria that cannot be cultured. Assembling a genome is challenging, however, because MDA generates highly nonuniform coverage of the genome. Here we describe an algorithm tailored for short-read data from single cells that improves assembly through the use of a progressively increasing coverage cutoff. Assembly of reads from single *Escherichia coli* and *Staphylococcus aureus* cells captures >91% of genes within contigs, approaching the 95% captured from an assembly based on many *E. coli* cells. We apply this method to assemble a genome from a single cell of an uncultivated SAR324 clade of Deltaproteobacteria, a cosmopolitan bacterial lineage in the global ocean. Metabolic reconstruction suggests that SAR324 is aerobic, motile and chemotactic. Our approach enables acquisition of genome assemblies for individual uncultivated bacteria using only short reads, providing cell-specific genetic information absent from metagenomic studies.

Chitsaz et al, *Nat. Biotech.* (2011) 29:915-921.

Collaboration between UCSD, JCVI, Illumina.

E+V-SC: EULER-SR error correction + we modified Velvet for single cell coverage issues

New Genome assembled with E+V-SC

Deltaproteobacteria (marine bacteria) single cell assembly results

Uncultivated bacteria from a seawater sample.

Assembler	# of contigs	N50 (bp)	Length (bp)	# Conserved single copy genes
Velvet	1,856	11,531	3,921,396	55/111 (46%)
E+V-SC	823	30,293	4,282,110	75/111 (67%)

N50 = the contig length at which longer contigs represent half of the total *assembly* length.

Chitsaz, et al., *Nat. Biotechnol.* (2011)

New Genome assembled with E+V-SC

Deltaproteobacteria (marine bacteria) single cell assembly features

Assembly size	4.3 Mb
Estimated genome size	4.9-6.4 Mb
# genes	3811

Chitsaz, et al., *Nat. Biotechnol.* (2011)

New Genome assembled with E+V-SC

Deltaproteobacteria single cell assembly completeness

- JCVI annotated assembly with their standard pipeline.
- Comparison to other microbial genomes using metrics from Nelson et al., *Science* (2010) 328: 994-999 shows similar completeness to other draft microbial genomes

# tRNA genes	20 out of 20 types
# tRNA synthetases	17 of 21 types
# rRNAs	1 each of 5S, 16S, 23S
# conserved single copy genes	75 out of 111 (67%)
# conserved single copy gene clusters	58 out of 66 (87%)

Chitsaz, et al., *Nat. Biotechnol.* (2011)

New Genome

Deltaproteobacteria assembly comparison

Assembler	N50 (bp)	Length (bp) (contigs > 200 bp)	# Long ORFs (> 600 bp)
E+V-SC	30,293	4,255,983	2,377
SPAdes	75,366	4,826,160	2,600

Ongoing SPAdes Collaborations

- Human Microbiome Project
(**Ashlee Earl, Broad Institute**)
- Sequencing uncultivated bacteria representing gray matter of life
(**Roger Lasken, Venter Institute**)
- Sequencing pathogens isolated from hospital environment
(**Jeff McLean, Venter Institute**)
- Sequencing antibiotics producing bacteria
(**Bill Gerwick, Scripps Institute of Oceanography**)
- Sequencing drug-resistant pathogens
(**Nik Schork, Scripps Translational Medicine**)

Publication

Anton Bankevich, Sergey Nurk, Dmitry Antipov, Alexey Gurevich, Mikhail Dvorkin, Alexander Kulikov, Valery Lesin, Sergey Nikolenko, Son Pham, Andrey Prjibelski, Alexey Pyshkin, Alexander Sirotkin, Nikolay Vyahhi, Glenn Tesler, Max Alekseyev and Pavel Pevzner.

SPAdes: a New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing,

Journal of Computational Biology, 19(5): 455-477 (2012)

Website

<http://bioinf.spbau.ru/spades>

Funding

Russian Federation grant
11.G34.31.0018

NIH 3P41RR024851-02S1

Acknowledgments: SPAdes Assembler

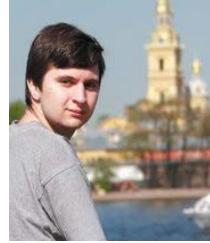
Saint Petersburg Academic University, Russian Academy of Sciences



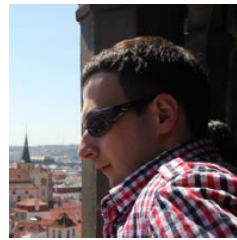
Dmitry
Antipov



Anton
Bankevich



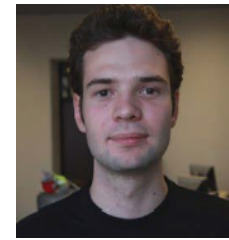
Mikhail
Dvorkin



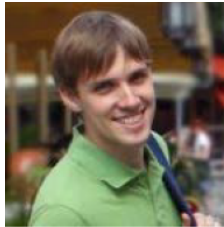
Valery
Lesin



Alexander
Kulikov



Sergey
Nurk



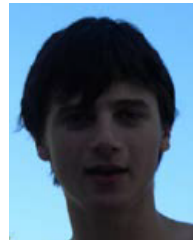
Nikolay
Vyahhi



Alexander
Sirotkin



Alexey
Gurevich



Alexey
Pyshkin



Andrey
Przhibelsky



Sergey
Nikolenko

University of South Carolina

University of California, San Diego



Max Alekseyev



Glenn Tesler



Son Pham



Pavel Pevzner