Assembling the genome of *Porphyromonas gingivalis* (gingivitis) from a single cell

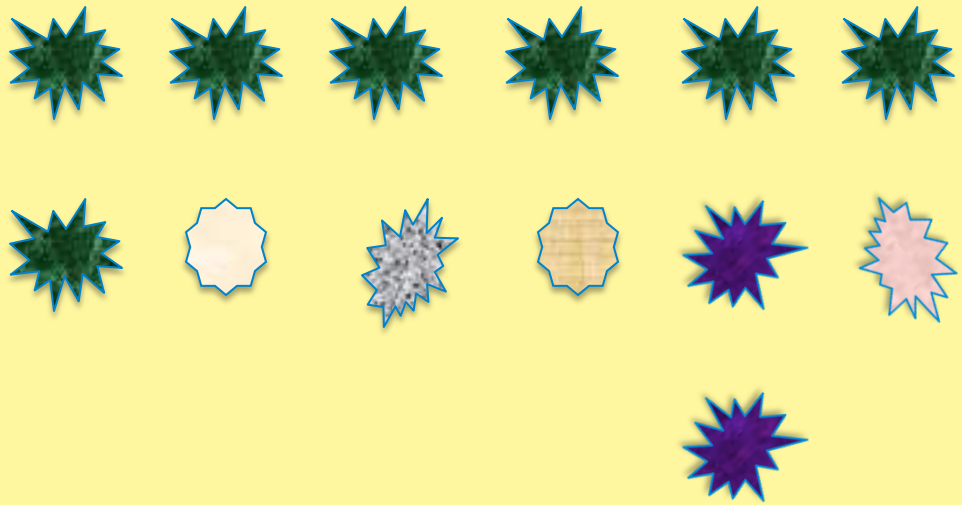
Glenn Tesler

University of California, San Diego
Department of Mathematics

Joint work with
Jeff McLean and Roger Lasken's labs at JCVI
Pavel Pevzner's labs at UCSD and St. Petersburg

Outline

- Genome sequencing
  - Conventional
  - Metagenomics
  - Single Cell



- De Bruijn graphs & SPAdes genome assembler
- *P. gingivalis* found in a hospital sink drain

DNA Sequence of a Genome

- The *E. coli* genome is ~ 4.6 million nucleotides long.
  Represent it as a (circular) string over the alphabet {A, C, G, T}:

  ***E. coli* K-12 substr. MG1655**

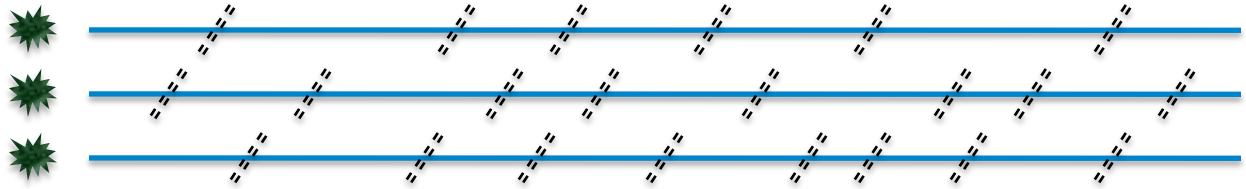
```
  1-50   AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAA
 51-100  AAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAAT
101-150  TAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAATATAGGCATA
151-200  GCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCAT
 . . .    . . . . . . . . . . . . . . . . . . . . . . . . .
4639651-4639675
         AAAAACGCCTTAGTAAGTATTTTTC
```

- The human genome is ~ 3 billion nucleotides long, split into chromosomes represented as linear strings over {A, C, G, T}.

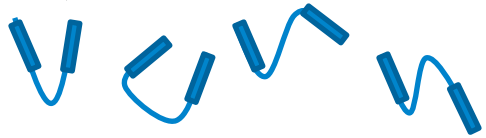
- Current technologies read ~ 25 – 10000 consecutive nucleotides.
  We focus on the popular Illumina GAIIx, with 100 nt **reads**.

Whole Genome Shotgun Sequencing

Fragment many copies of same genome. Lose positional information.



Sequence reads (25 to 10000 nt) at one or both ends of fragments.

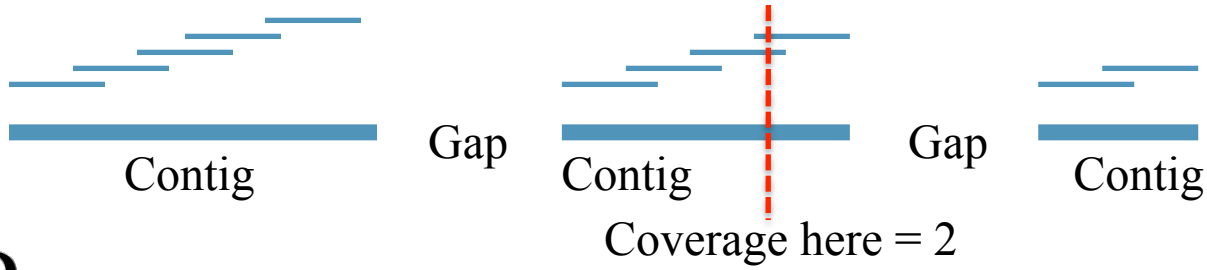


Find overlapping reads.

```
                    ACGTAGAATCGACCATG...
...AACATAGTTGACGTAGAATC
```

Merge overlapping reads into contigs.

```
...AACATAGTTGACGTAGAATCGACCATG...
```



Computational Challenges in Genome Assembly

- *Problem:* Given a collection of *reads* (short substrings of the genome sequence in the alphabet A, C, G, T),
  reconstruct the genome from which the reads are derived.

- *Challenges:*
  - Repeats in the genome

    …ACCCAGTT*GACTGGGAT*CCTTTTTAAA*GACTGGGAT*TTTAACGCGTAAG…

    CAGTT*GACTG*

    *ACTGGGAT*CC

    GACTGGGATT

    Sample reads
  - Sequencing errors (vary by platform and protocol), including:

CCTTTTTATAGACTG	Substitution
CCTTTTTA-AGACTGG	Deletion
CCTTTCTTAAAGACT	Insertion
CCTTTTTTTTTAAAGA	Homopolymer
CCTTTTTTCGCGTAA	Chimeric read

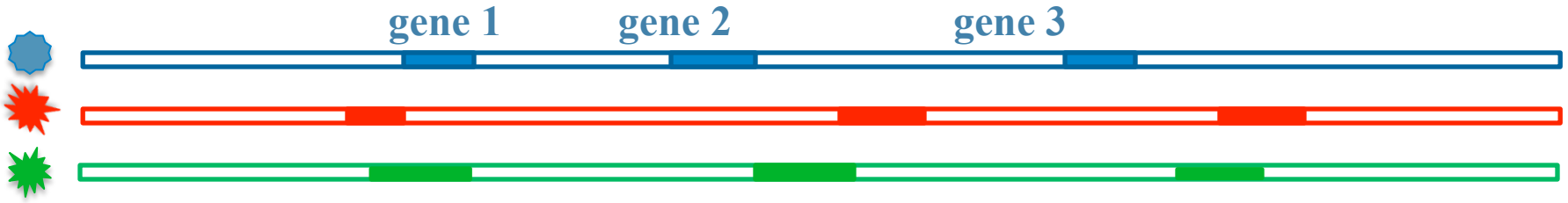
  - Size of the data, e.g. 30 million reads of length 100 nt in a 7 GB file.

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), and/or information about the dominant 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.

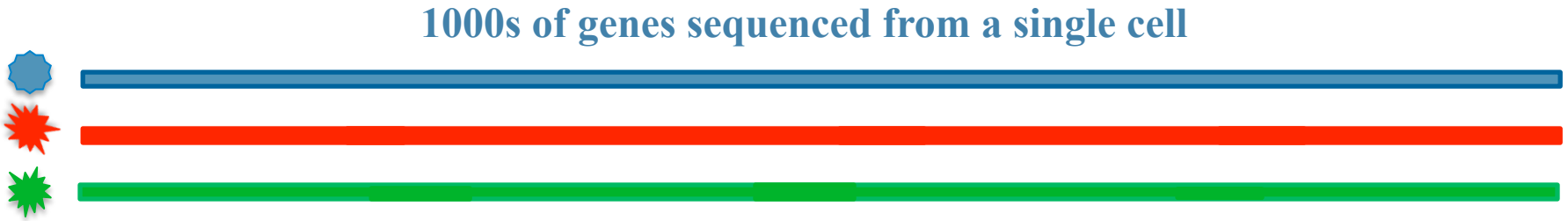
- Until recently, **metagenomics** was the only option for studies of microbial communities. However, metagenomics provides information about only a **few genes** (across many species), and/or information about the dominant species.

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-centri**c metagenomics data with **whole-genom**e assembly of uncultivated organisms.



Single Cell Sequencing via MDA:

Multiple Displacement Amplification

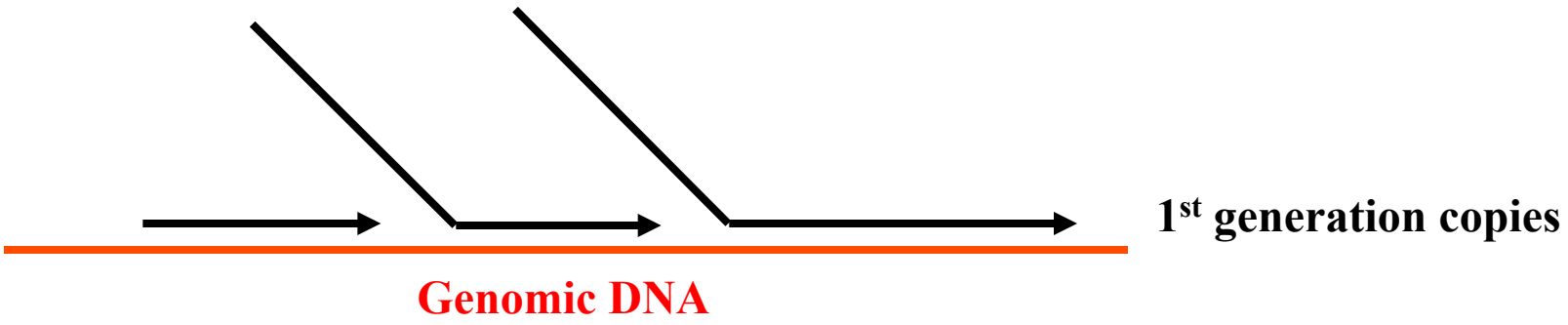
Genomic DNA

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.
- Commercially available kits:
  TempliPhi and GenomiPhi (GE Healthcare) and REPLI-g (Qiagen).
- REPLI-g: fragments ~ 2 – 100 kb; usually > 10 kb on average.

Single Cell Sequencing via MDA:

Multiple Displacement Amplification

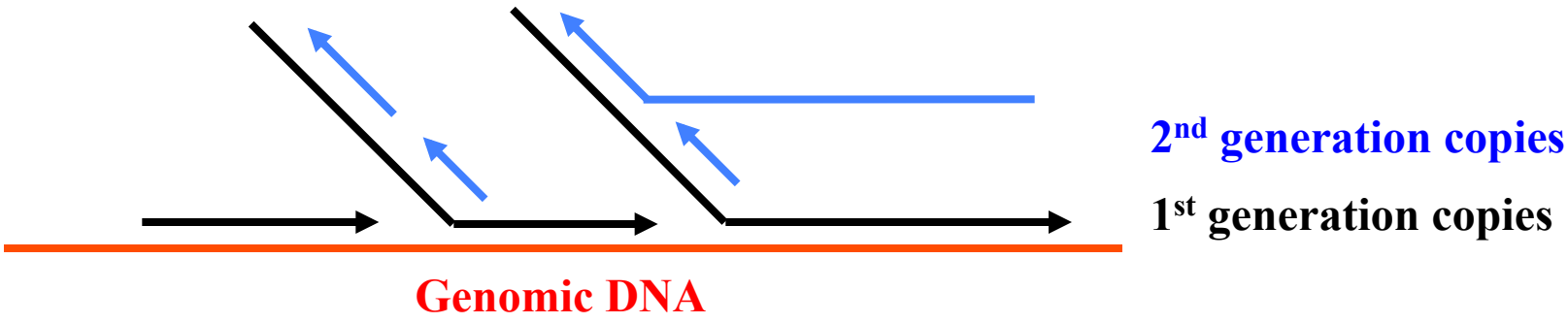


F.B. Dean, J.R. Nelson, T.L. Giesler, R.S. Lasken (2001). *Genome Res.* 11:1095-9
F.B. Dean, S. Hosono, L. Fang, et al. (2002). *PNAS* 99:5261-6

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

Single Cell Sequencing via MDA:

Multiple Displacement Amplification

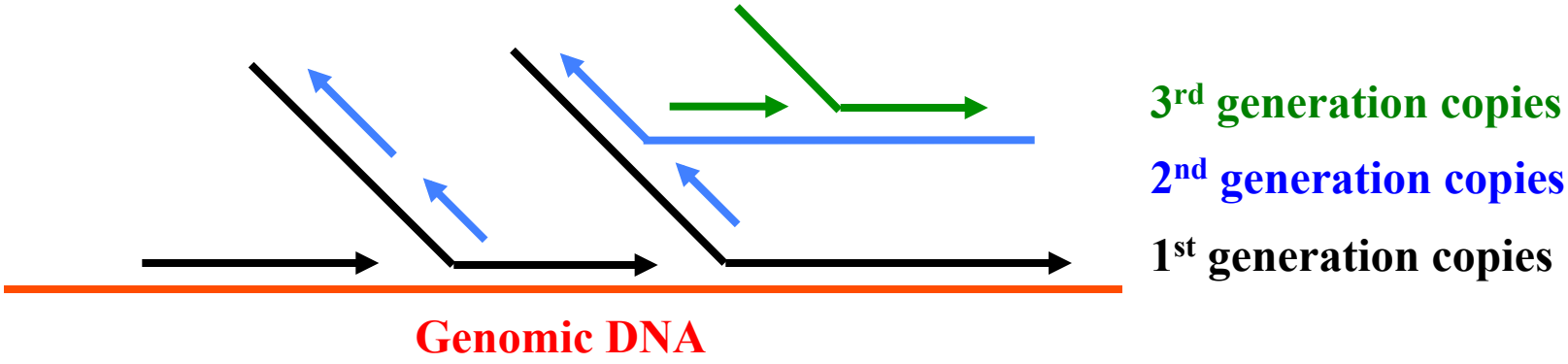


F.B. Dean, J.R. Nelson, T.L. Giesler, R.S. Lasken (2001). *Genome Res.* 11:1095-9
F.B. Dean, S. Hosono, L. Fang, et al. (2002). *PNAS* 99:5261-6

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

Single Cell Sequencing via MDA:

Multiple Displacement Amplification

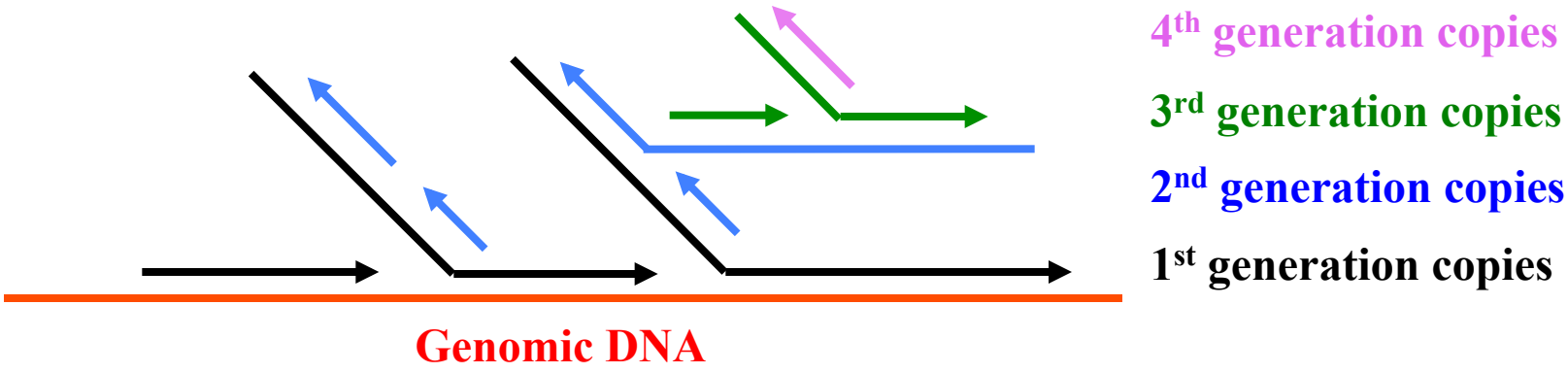


F.B. Dean, J.R. Nelson, T.L. Giesler, R.S. Lasken (2001). *Genome Res.* 11:1095-9
F.B. Dean, S. Hosono, L. Fang, et al. (2002). *PNAS* 99:5261-6

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

Single Cell Sequencing via MDA:

Multiple Displacement Amplification



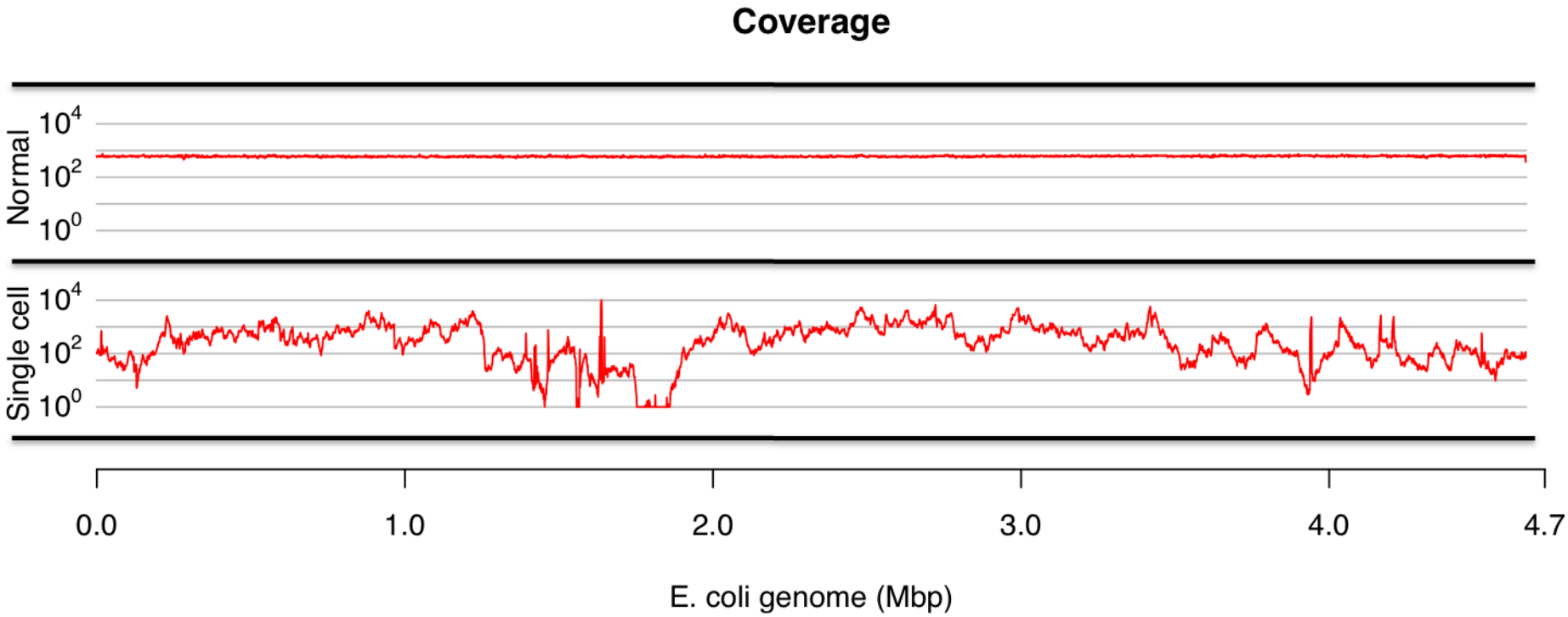
F.B. Dean, J.R. Nelson, T.L. Giesler, R.S. Lasken (2001). *Genome Res.* 11:1095-9
F.B. Dean, S. Hosono, L. Fang, et al. (2002). *PNAS* 99:5261-6

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

Sequencing Coverage

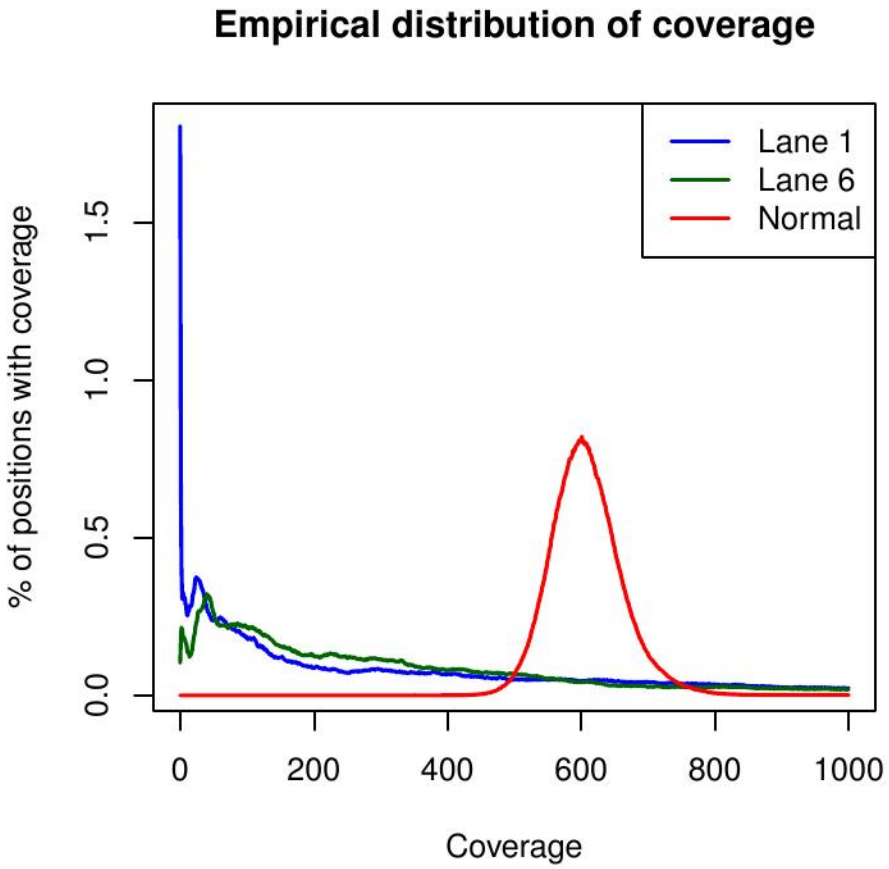
Normal multicell vs. single cell *E. coli* via MDA

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

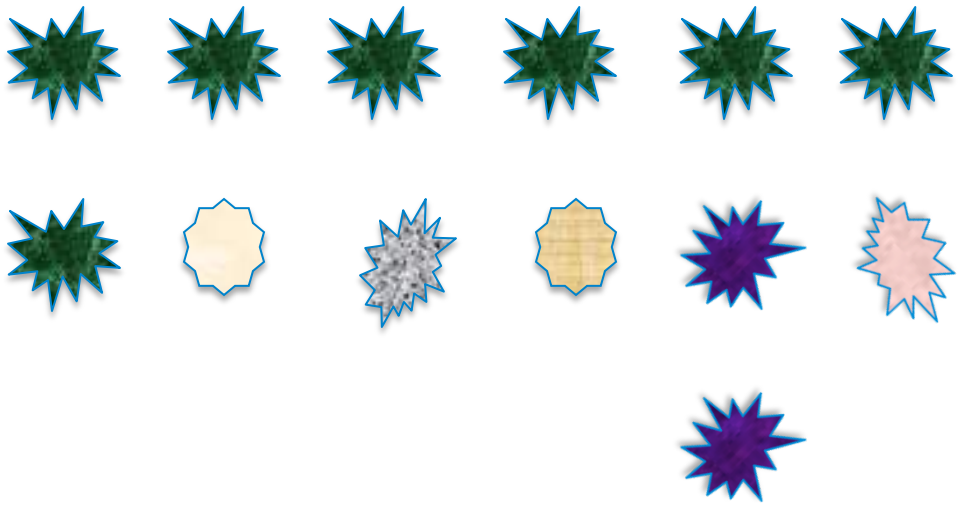


A cutoff threshold will eliminate about 25% of valid data in the single cell case, whereas it eliminates noise in the normal multicell case.

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

Outline

- Genome sequencing
  - Conventional
  - Metagenomics
  - Single Cell



- De Bruijn graphs & SPAdes genome assembler

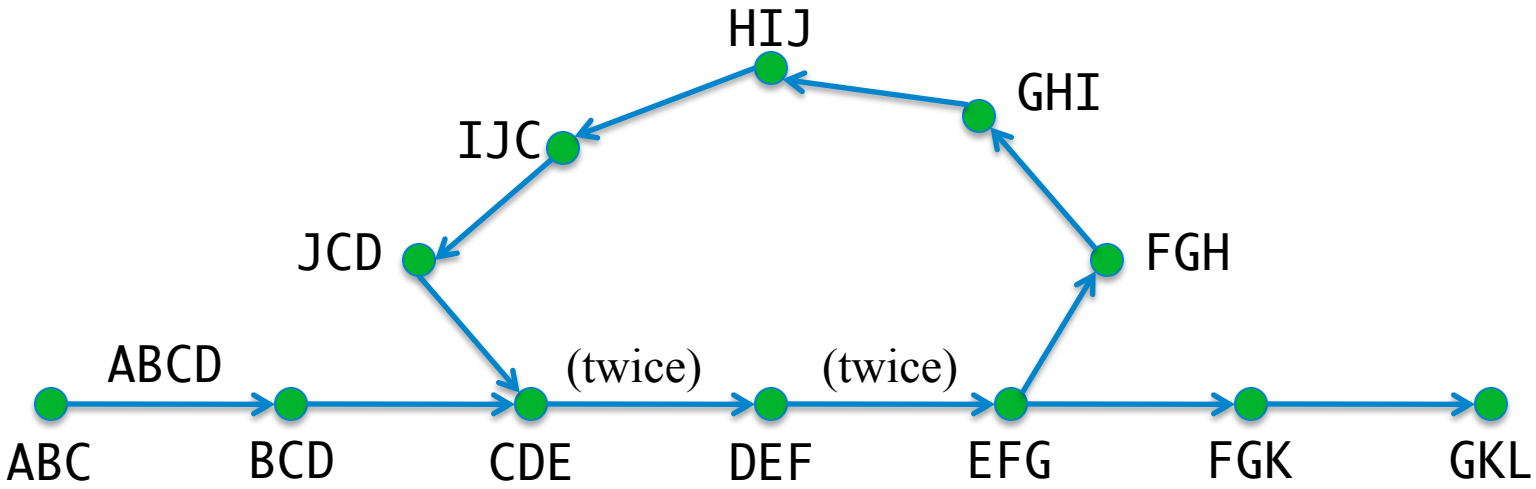
- *P. gingivalis* found in a hospital sink drain

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* → *xyz*
Genome: Eulerian path through graph (using edge multiplicities)

Genome: ABCDEFGHIJCDEFGKL



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*

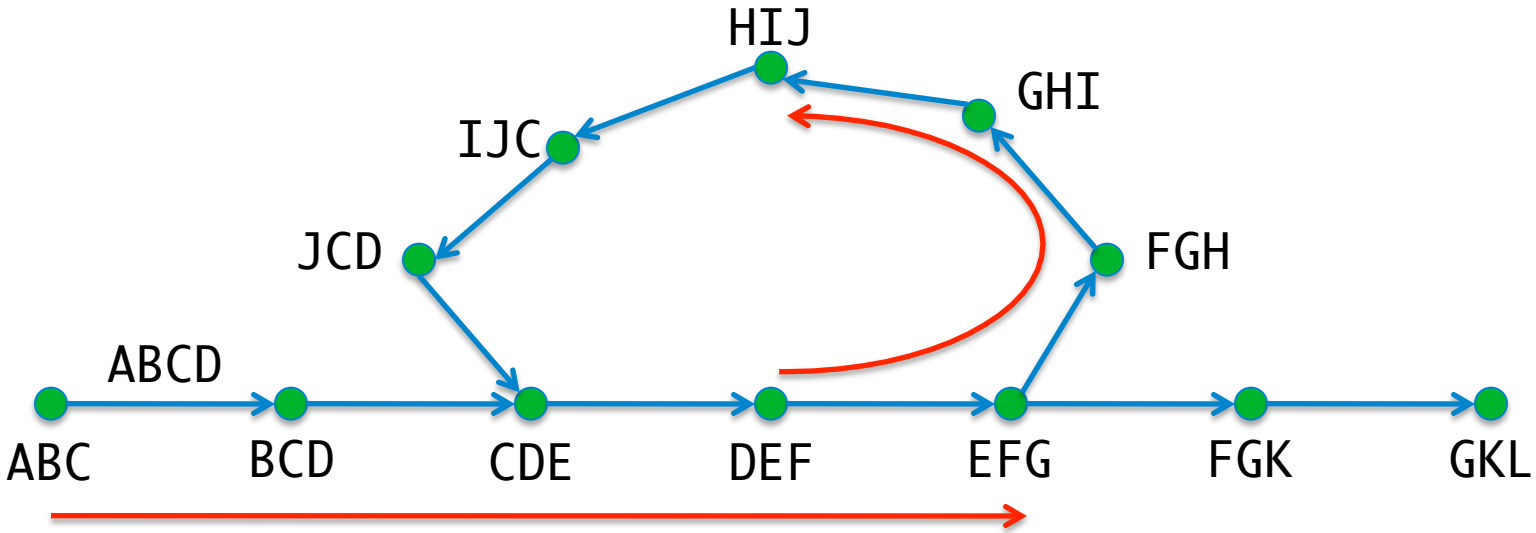
Same De Bruijn Graph from Perfect Reads

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* → *xyz*
Reads: short walks through graph (red)
Genome: long walk through graph
We lose exact repeat multiplicities

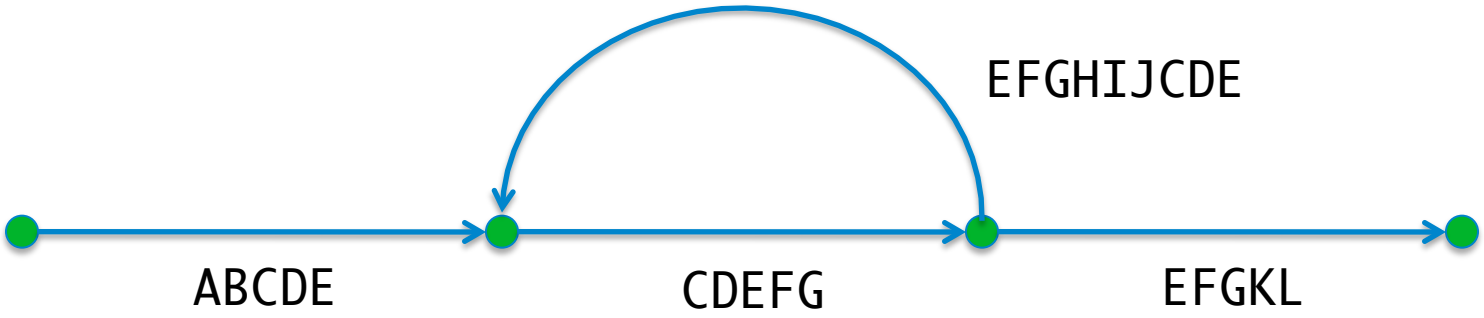
Reads (but order would be random in real data):

```
ABCDEFG
   DEFGHIJ
      GHIJCDE
         IJCDEFG
            CDEFGKL
```

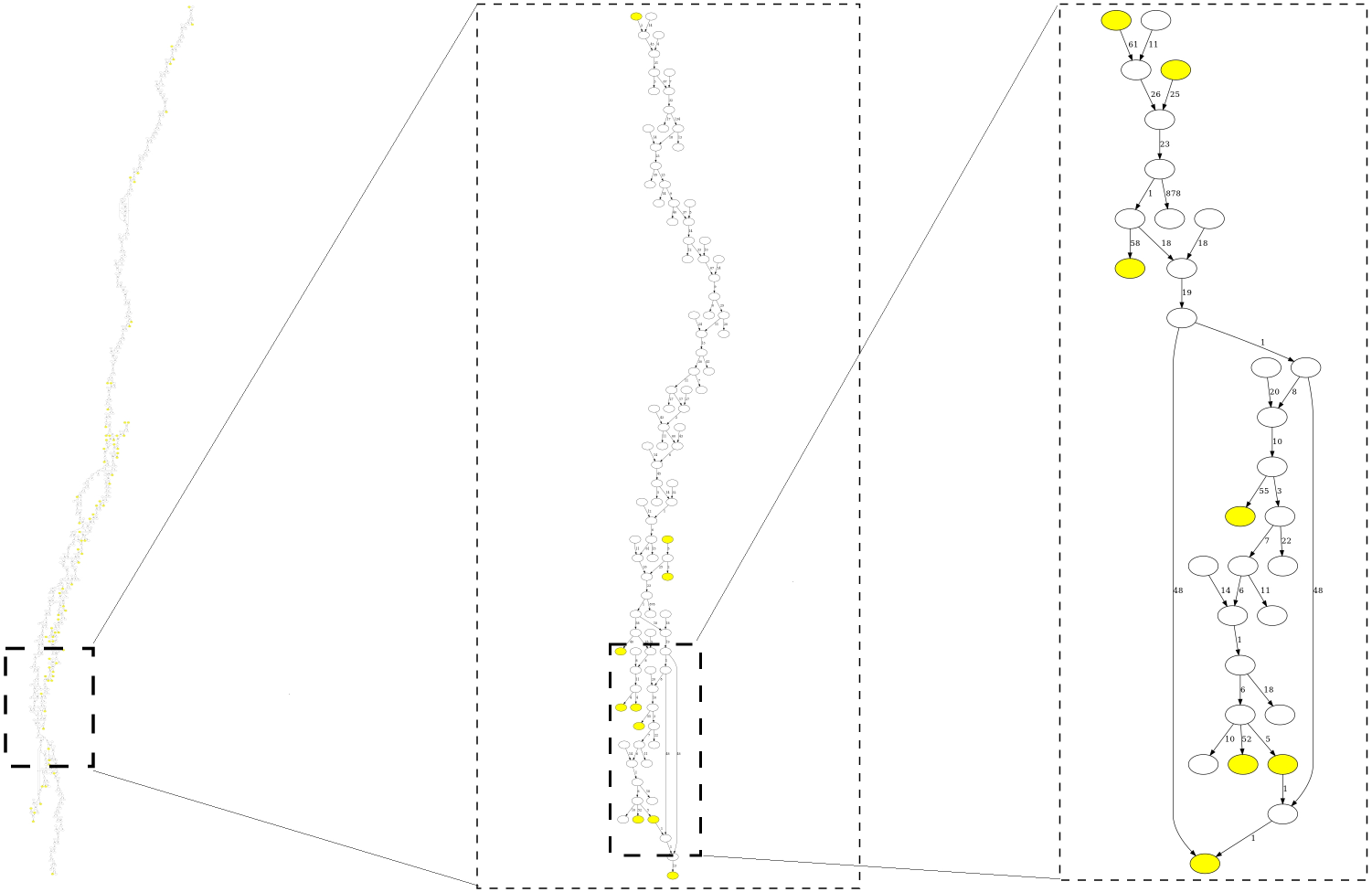


Condensed graph

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



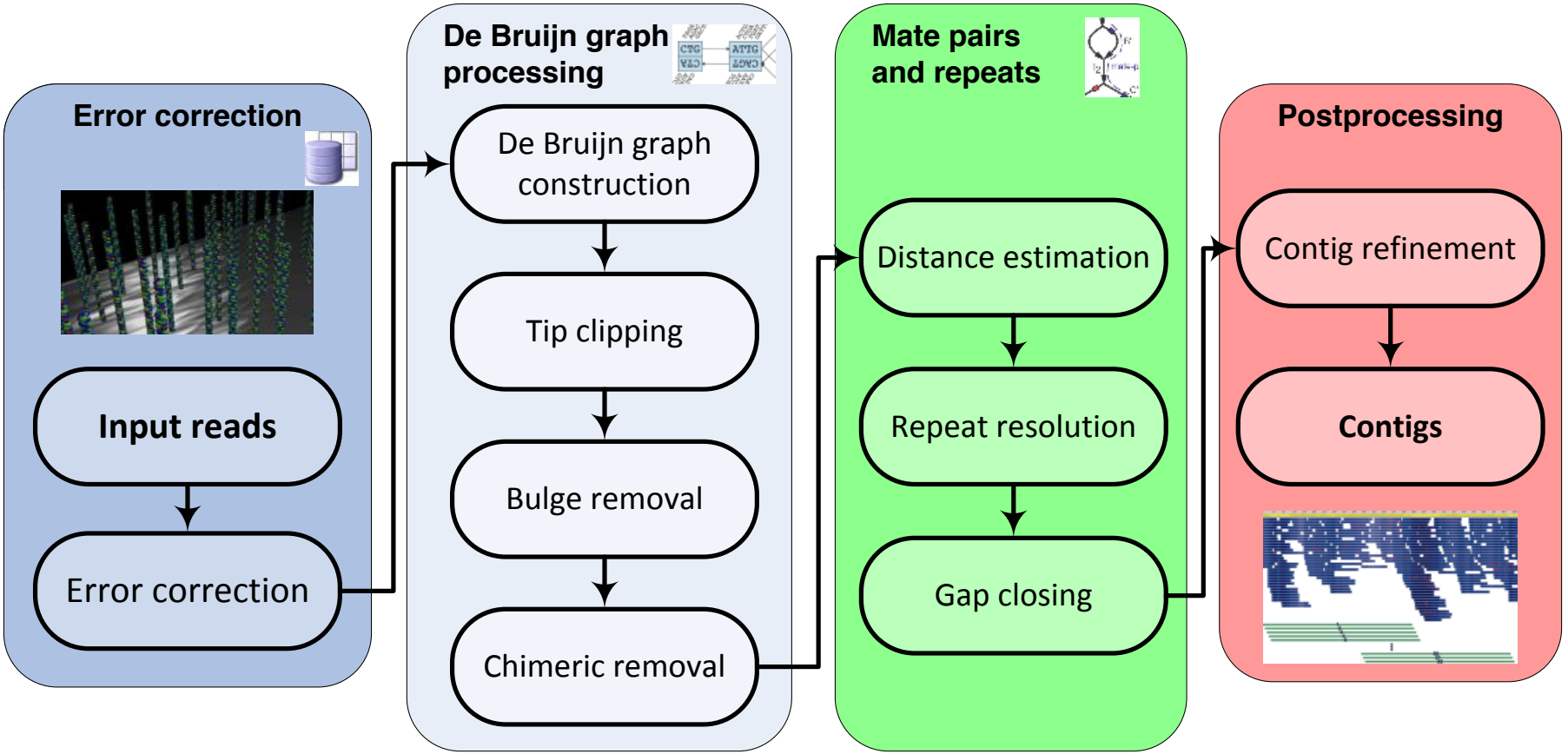
Read errors and imperfect repeats lead to a complicated graph



De Bruijn Graph of *E. coli*

Genome length	4.6 million bases
Reads	Illumina GA IIx platform, paired end sequencing 100 bases/read Reads are in pairs spanning ~ 250 bases (varies) ~ 30 million reads (15 million read pairs) ~ 600x coverage ~ 7 GB FASTQ file
De Bruin Graph parameters	Can set k between ~ 25 – 70. We used 55-mer vertices 56-mer edges
Graph size	Initially: ~ 200 million vertices (55-mers) Output: ~ 200 – 2000 contigs (varies by assembler) ~ 4.6 million bases

SPAdes genome assembler



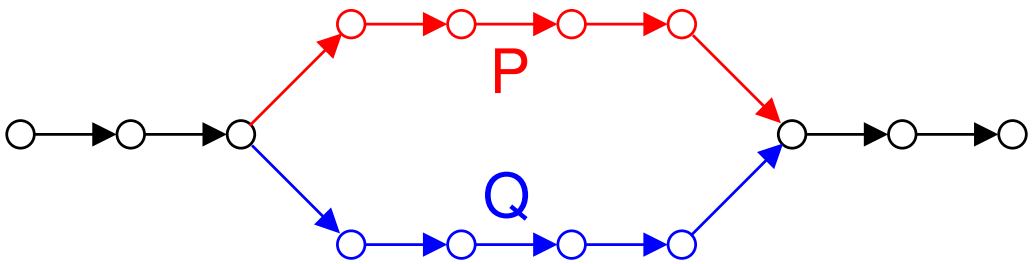
- De Bruijn graph assembler.
- Adapted to handle conventional and single cell datasets.
- Instead of global thresholds, uses local coverage, topology, and lengths to decide how to process the assembly graph.

Graph Simplification in SPAdes

P: Erroneous edges Q: correct alternative

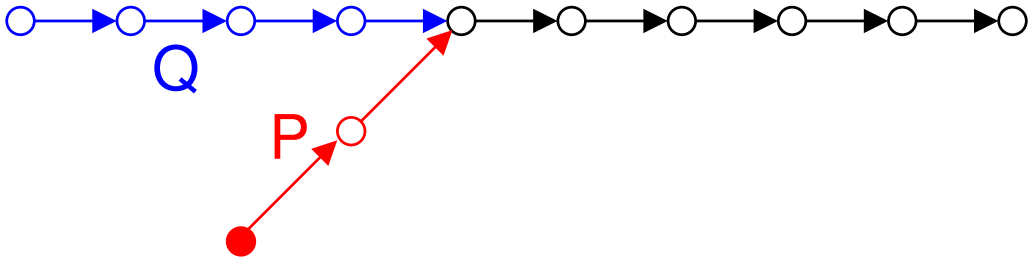
Bulge from error in middle of read

```
TCGGTGAAAGAGCTTT
 CGGTGAACGAGCTTTG
  GGTGAAAGAGCTTTGA
   GTGAAAGAGCTTTGAT
```



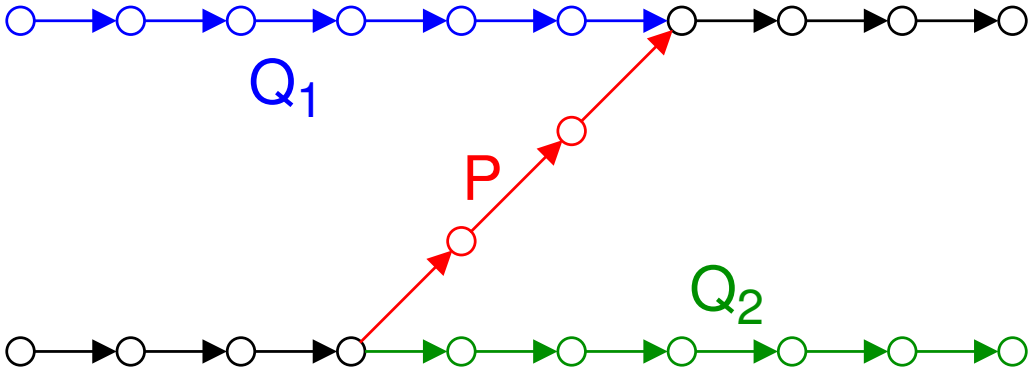
Tip from error near start/end of read

```
TCGGTGAAAGAGCTTT
 CGCTGAAAGAGCTTTG
  GGTGAAAGAGCTTTGA
   GTGAAAGAGCTTTGAT
```

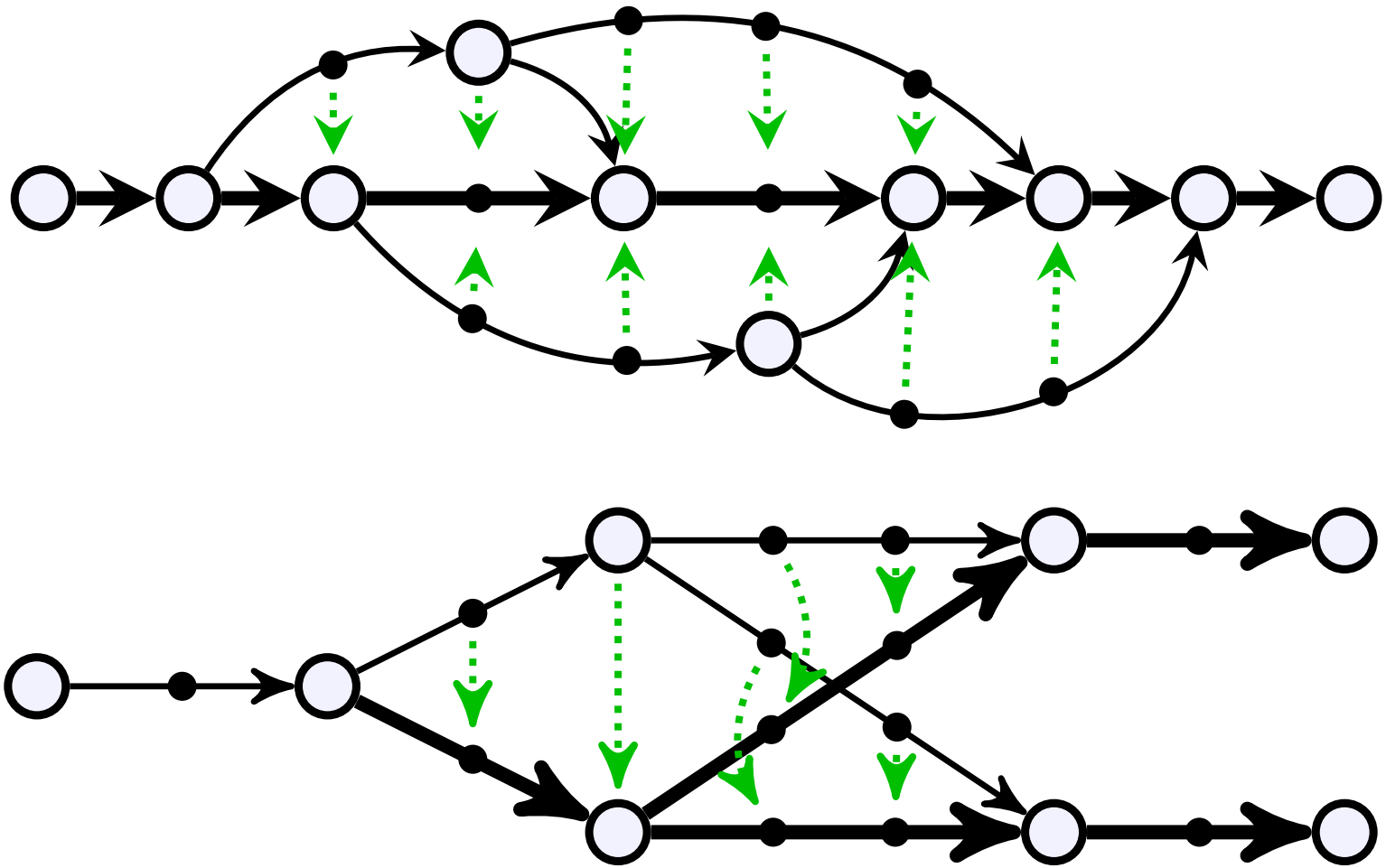


Chimeric connection joining two distant parts of genome

```
TCGGTGAAAGAGCTTT
 CGGTGAAAGAGCTTTG
  ACATCGTAAGCTTTGC
     TCGTAGTAGCCGATTC
      CGTAGTAGCCGATTCG
```



Bulges can be more complex



Nurk et al (2013), *Journal of Computational Biology*

Graph Simplification in SPAdes

- To clean up these problems, we consider local coverage, topology, and lengths.

- **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.

Graph Simplification complications

- These configurations also arise from repeats.
  In other contexts, they can arise from diploid variations and polymorphic samples.

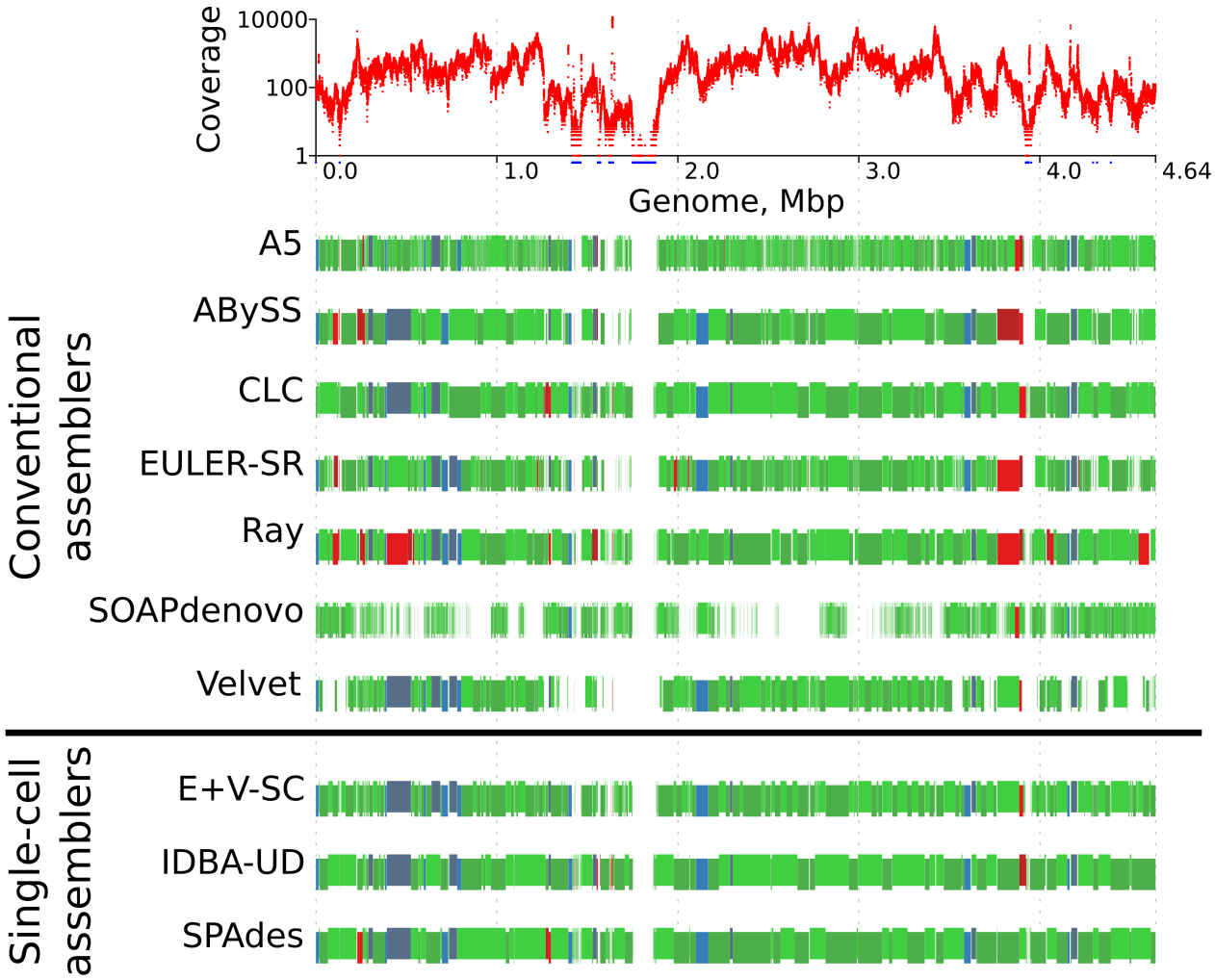
- Most de Bruijn assemblers use a fixed global coverage cutoff to eliminate many erroneous edges, and then use heuristics for further simplifications. This doesn't work well for single cell MDA.

- Error correcting reads before assembly reduces the number of erroneous edges. A global vs. local coverage cutoff issue also applies to the error correction stage. BayesHammer (in SPAdes) does error correction for single-cell data.

- Velvet-SC (Chitsaz et al, 2011) and SPAdes (Bankevich et al, 2012), both from Pevzner's group, and recently IDBA-UD (Peng et al, 2012), make better use of local coverage and variable thresholds in performing simplifications.

E. coli mapped contigs (single cell)

QUAST plot



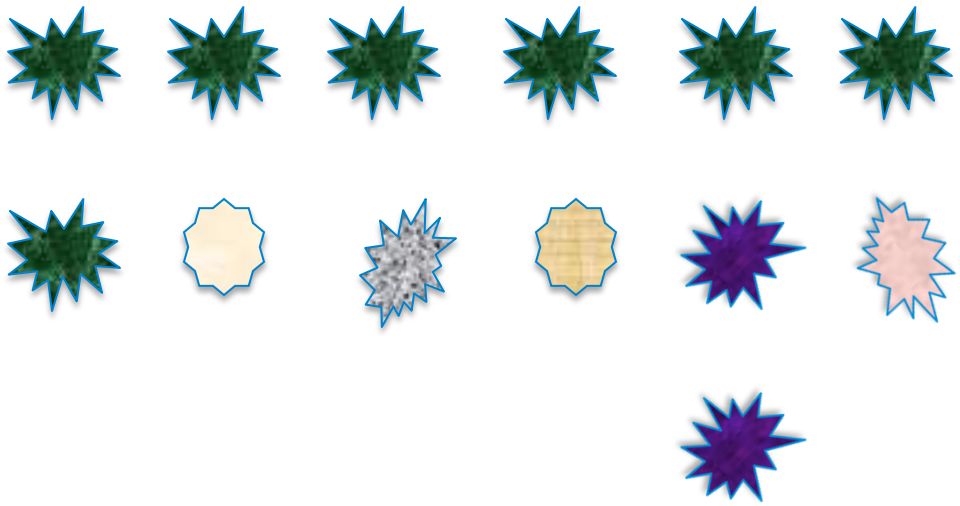
Correctly assembled.
Blue: similar boundaries in at least half of the assemblers.

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

UCSD

Outline

- Genome sequencing
  - Conventional
  - Metagenomics
  - Single Cell



- De Bruijn graphs & SPAdes genome assembler

- *P. gingivalis* found in a hospital sink drain

Genome of the pathogen *Porphyromonas gingivalis* recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform

Jeffrey S. McLean,[1,2,7] Mary-Jane Lombardo,[1] Michael G. Ziegler,[3] Mark Novotny,[1] Joyclyn Yee-Greenbaum,[1] Jonathan H. Badger,[1] Glenn Tesler,[4] Sergey Nurk,[5] Valery Lesin,[5] Daniel Brami,[1] Adam P. Hall,[1] Anna Edlund,[1] Lisa Z. Allen,[1] Scott Durkin,[1] Sharon Reed,[3] Francesca Torriani,[3] Kenneth H. Nealson,[1,2] Pavel A. Pevzner,[5,6] Robert Friedman,[1] J. Craig Venter,[1] and Roger S. Lasken[1,7]

Genome Research (2013) 23: 867-877

Gingivitis found in a hospital sink biofilm

McLean et al. (2013), Genome Research

- Single-cell genomics is becoming an accepted method to capture novel genomes, particularly in marine and soil environments, and in hosts (human, termite gut, and others).

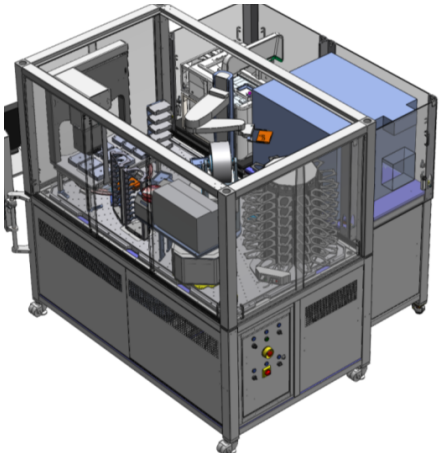
Binga et al (2008), *The ISME Journal*, 2:233-241
Chitsaz et al (2011), *Nature Biotechnology,* 29:915-921
Stepanauskas (2012), *Current Opinion in Microbiology*, 15:613-620
Lasken (2012), *Nature Reviews Microbiology*, 10:631-640
Blainey (2013), *FEMS Microbiol Rev*, 37:407-427

- Here we show for the first time that it also enables comparative analysis of strain variations in a pathogen captured in a hospital biofilm.

- Single-cell assemblies enable sequence-level comparisons previously only possible with cultivated organisms, including gene annotations, and variations in genes, repeats, and virulence factors.

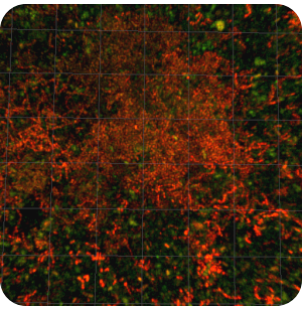
Automated Single Cell Genomics Pipeline

Automated single cell genomics pipeline



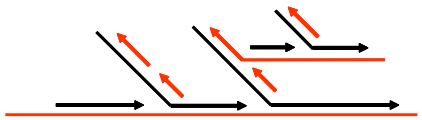
- Sink drain biofilm samples collected from a public restroom at UCSD Medical Center emergency room.
- Analyzed with a new robotic platform developed by Lasken & McLean labs at JCVI.
- Platform flow sorts single cells from a sample onto 384 well plates, amplifies them via MDA, and classifies them based on 16S typing.
- Throughput: 5000 wells/week from sort to 16S data.

Disrupt Biofilm Into Single Cells



Flow sort single cells 384-well format

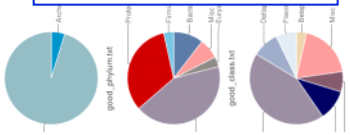
MDA 10^9-fold DNA amplification



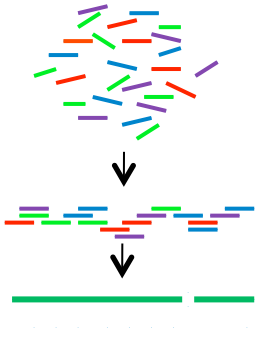
16S PCR

Cycle sequencing

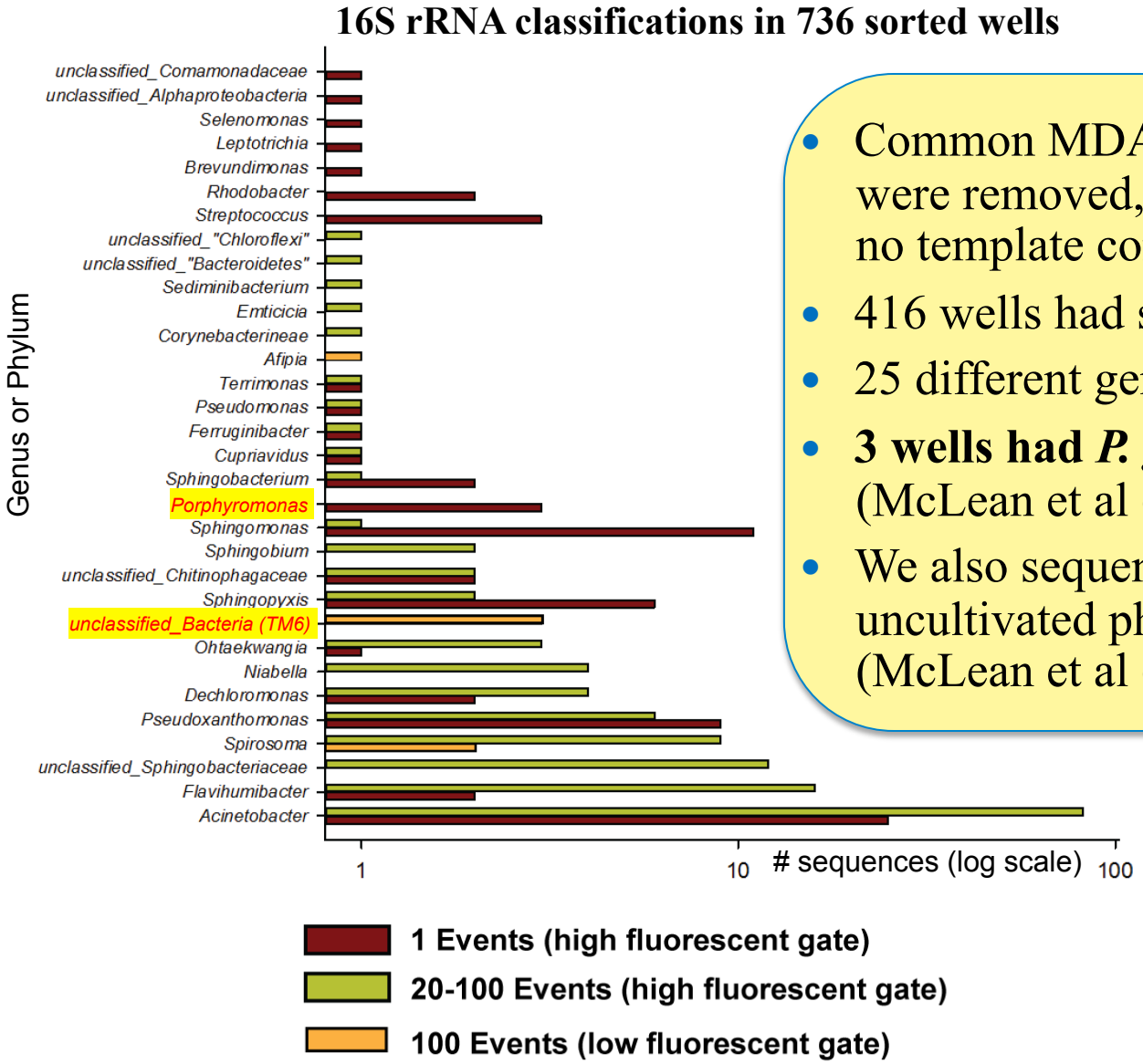
Classification



Whole Genome Sequencing and Assembly



16S rRNA classifications in hospital biofilm

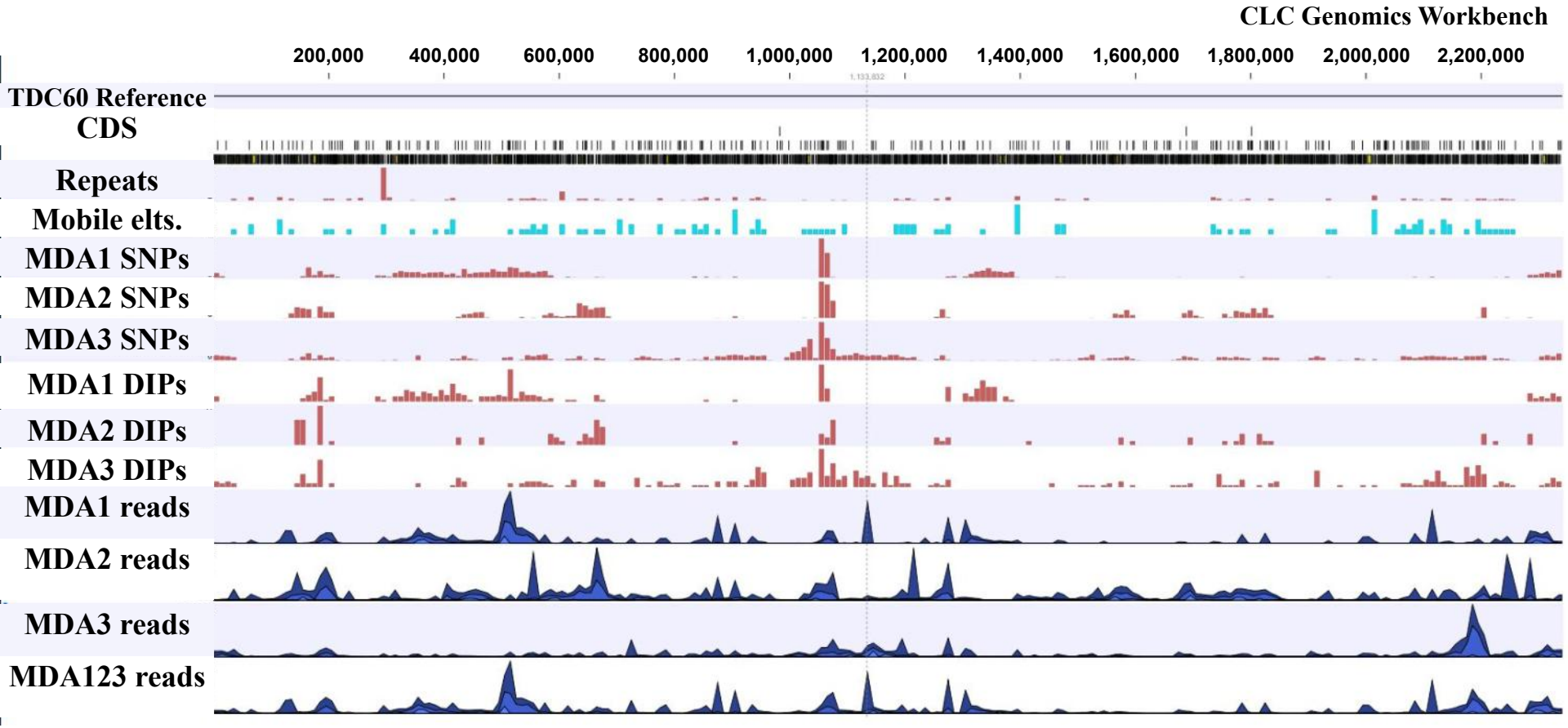


- Common MDA contaminant sequences were removed, and those sequences within no template control plates.
- 416 wells had single events.
- 25 different genera found.
- **3 wells had *P. gingivalis*.** (McLean et al (2013), *Genome Research*)
- We also sequenced bacteria from TM6, an uncultivated phylum (McLean et al (2013), *PNAS*).

UCSD

Single-Cell Sequencing allows base-level analysis of the individual cells

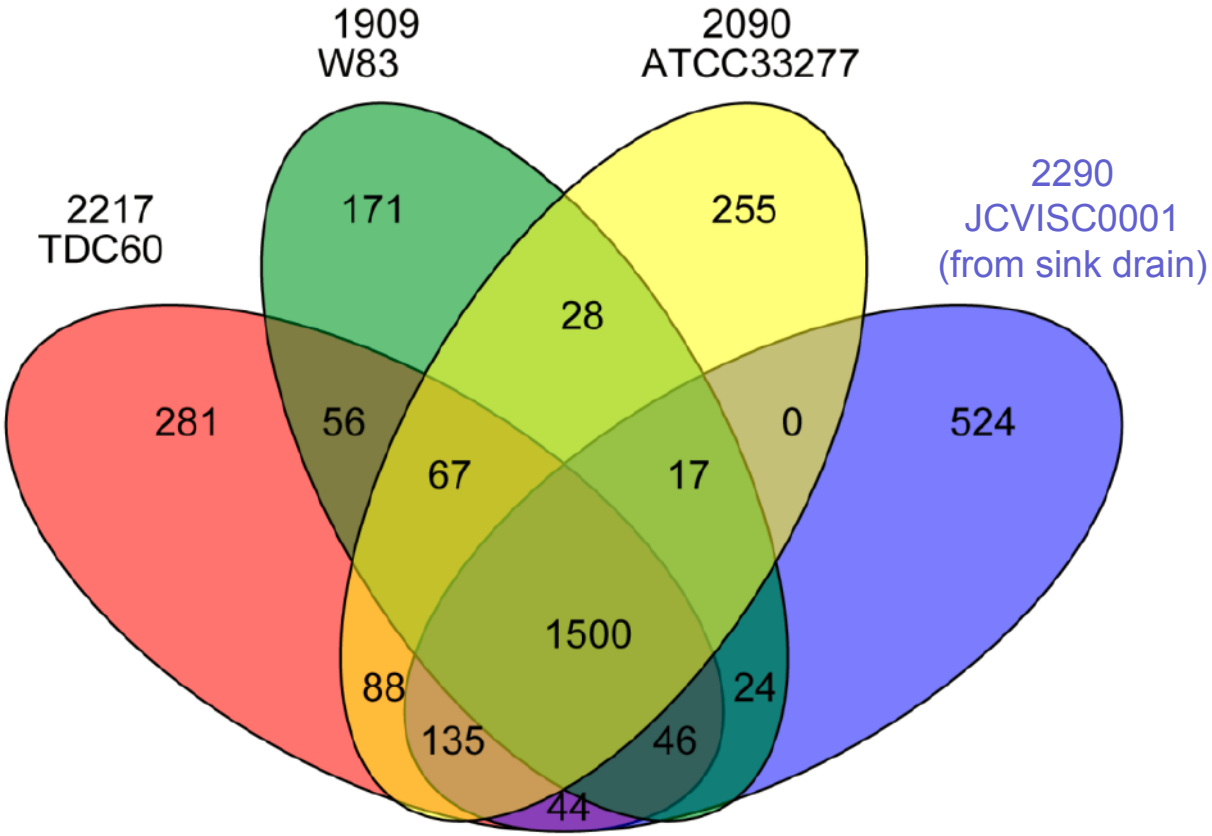
Comparing Coverage, SNPs, and DIPs in 3 MDAs vs. reference strain TDC60



Shared	Shared Total	Within CDS	Missense
SNPS	847	791	202
DIPS	75	44	31

UCSD

Unique Genes Found in Novel Strain of *P. gingivalis* from de novo assembly



New strain (JCVISC0001) annotated with the JCVI Prokaryotic Annotation Pipeline, which is based on Glimmer

Publications

SPAdes

Bankevich et al. (2012), **SPAdes: a New Genome Assembly Algorithm and its Applications to Single-Cell Sequencing,** *Journal of Computational Biology*, 19(5):455-477.

Nurk et al. (2013), **Assembling Genomes and Mini-Metagenomes From Chimeric MDA Products,** *Journal of Computational Biology,* 20(10):714-737.

Gurevich et al. (2013), **QUAST: QUality ASsessment Tool for genome assemblies,** *Bioinformatics*, 29(8):1072-1075.

Hospital Biofilm

McLean et al. (2013), **Genome of the pathogen Porphyromonas gingivalis recovered from a biofilm in a hospital sink using a high-throughput single-cell genomics platform,** *Genome Research*, 23(5):867-877. Published online ahead of print April 5, 2013.

McLean et al., (2013), **Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum,** *PNAS*, 110(26):E2390-E2399.

Acknowledgements

St. Petersburg Academic University

Dmitry Antipov
Anton Bankevich
Alexey Gurevich
Anton Korobeynikov
Alla Lapidus
Sergey Nurk
Andrey Prjibelsky

and former lab members
Mikhail Dvorkin
Alexander Kulikov
Valery Lesin
Sergey Nikolenko
Alexey Pyshkin
Alexander Sirotkin
Yakov Sirotkin
Nikolay Vyahhi

University of South Carolina

Max Alekseyev

University of California, San Diego

Glenn Tesler
Son Pham
Pavel Pevzner

UCSD Dept. of Medicine
Michael Ziegler
Sharon Reed
Francesca Torriani

J. Craig Venter Institute

Jeffrey S. McLean
Mary-Jane Lombardo
Mark Novotny
Joyclyn L. Yee-Greenbaum
Johnathan H. Badger
Daniel Brami
Adam Hall
Anna Edlund
Lisa Z. Allen
Scott Durkin
Kenneth H. Nealson
Robert Friedman
J. Craig Venter
Roger S. Lasken

UCSD

Funding

- Alfred P. Sloan Foundation (Sloan Foundation-2007-10-19)
- National Institutes of Health (1R01GM095373, 1R01DE020102, 3P41RR024851-02S1, UL1TR000100)
- Government of the Russian Federation (grant 11.G34.31.0018)
- National Science Foundation (grant CCF-1115206)

Software

- http://bioinf.spbau.ru/spades

Double-Stranded DNA

One strand, *s*: ACAATGAG

Complement: Pair up A↔T and C↔G

Double-stranded DNA:

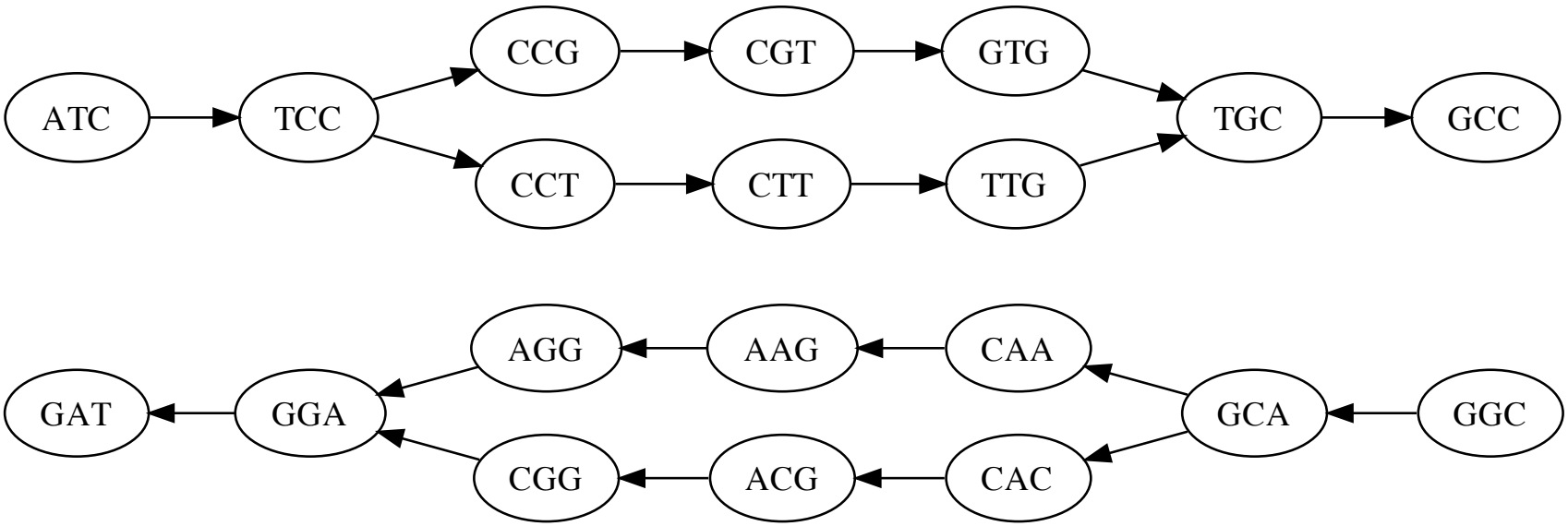
```
5' — A C A A T G A G — 3'
     | | | | | | | |
3' — T G T T A C T C — 5'
```

Complement of *s*: TGTTACTC

Reverse complement of *s*: CTCATTGT

Double-Stranded DNA

- Reads may come from either strand.
- Assembler throws in the reverse complement of each read to detect this, creating dual vertices and dual edges in graph.
- Reads ATCC{G,T}TGCC and reverse complements
  GGCA{C,A}GGAT



Double-Stranded DNA

- We also may use a *bidirected graph.*
- Each edge has arrowheads on **both** ends.
  Each arrowhead may point in or out of the node.
- Enter node on an ‘in’ / exit on an ‘out’: use sequence as-is
  Enter node on an ‘out’ / exit on an ‘in’: use rev. complement
  Not allowed to enter/exit on ‘in’/‘in’ or ‘out’/‘out’