Introduction to single-cell genome assembly using SPAdes

Kasia (Katarzyna) Zaremba-Niedzwiedzka
Jimmy Saw

Uppsala University

Outline: practical part

- Assembly basics
- Assembly comparisons
- Sample
- Today's exercise

Outline: theory

- Assembly basics
- Assembly comparisons
- Sample
- Today's exercise

- Single-cell data specific problems
- Available assemblers
- How SPAdes works

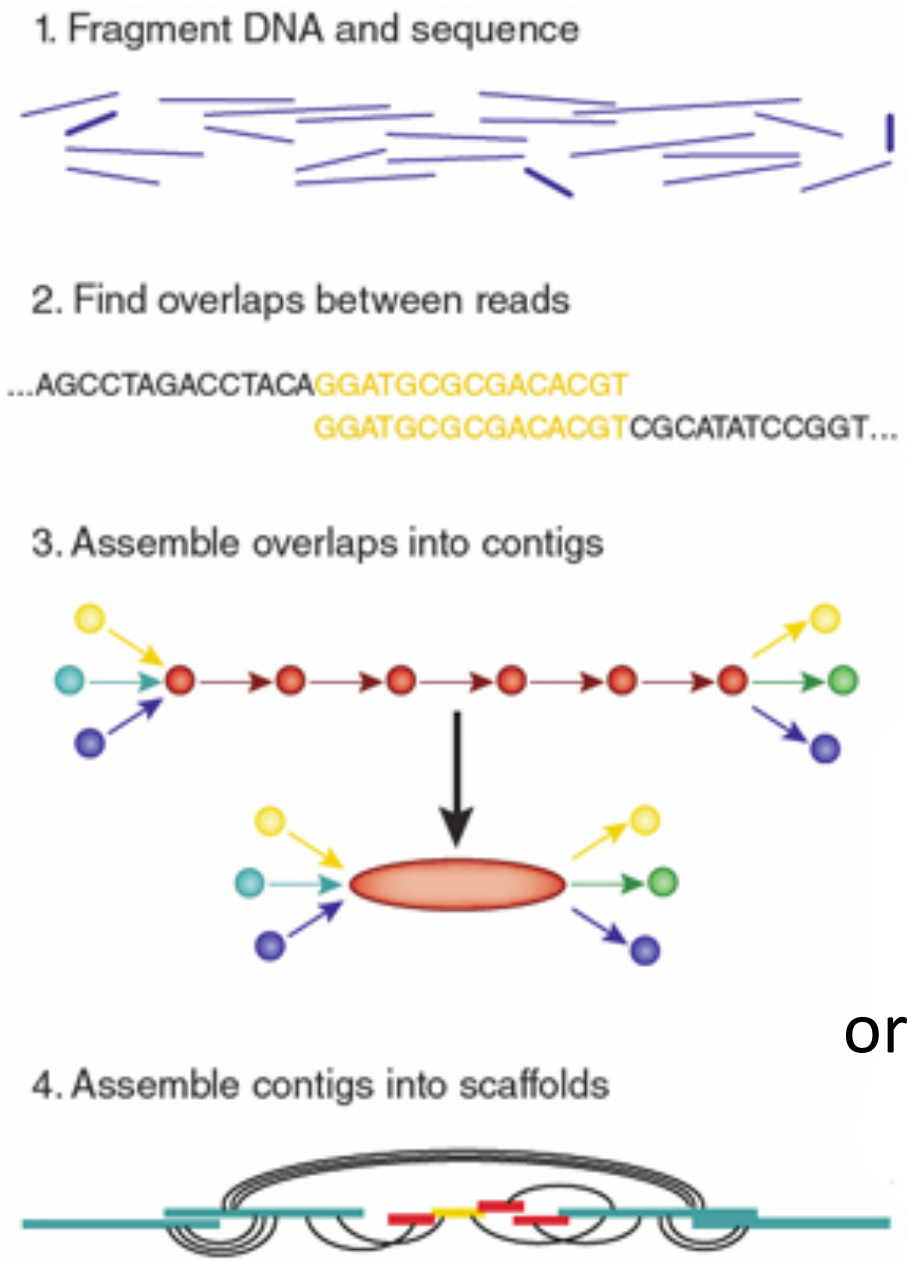
De novo genome assembly: what every biologist should know **Monya Baker**
Nature Methods **9**, 333–337 (2012) doi:10.1038/nmeth.1935

Assembly puzzle

http://www.scienceinschool.org

Assembly puzzle

http://www.scienceinschool.org

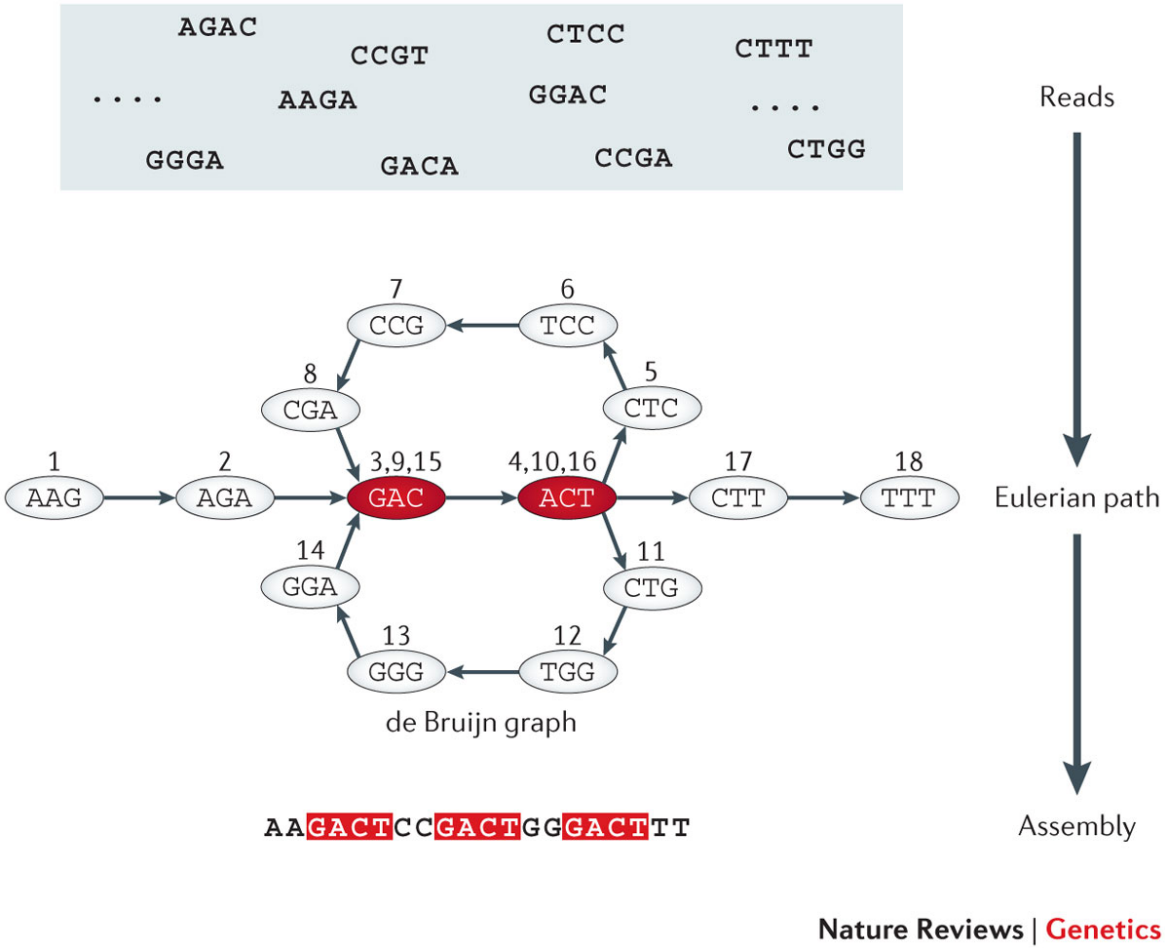


Contigs = continuous sequence

Scaffolds = ordered contigs with gaps

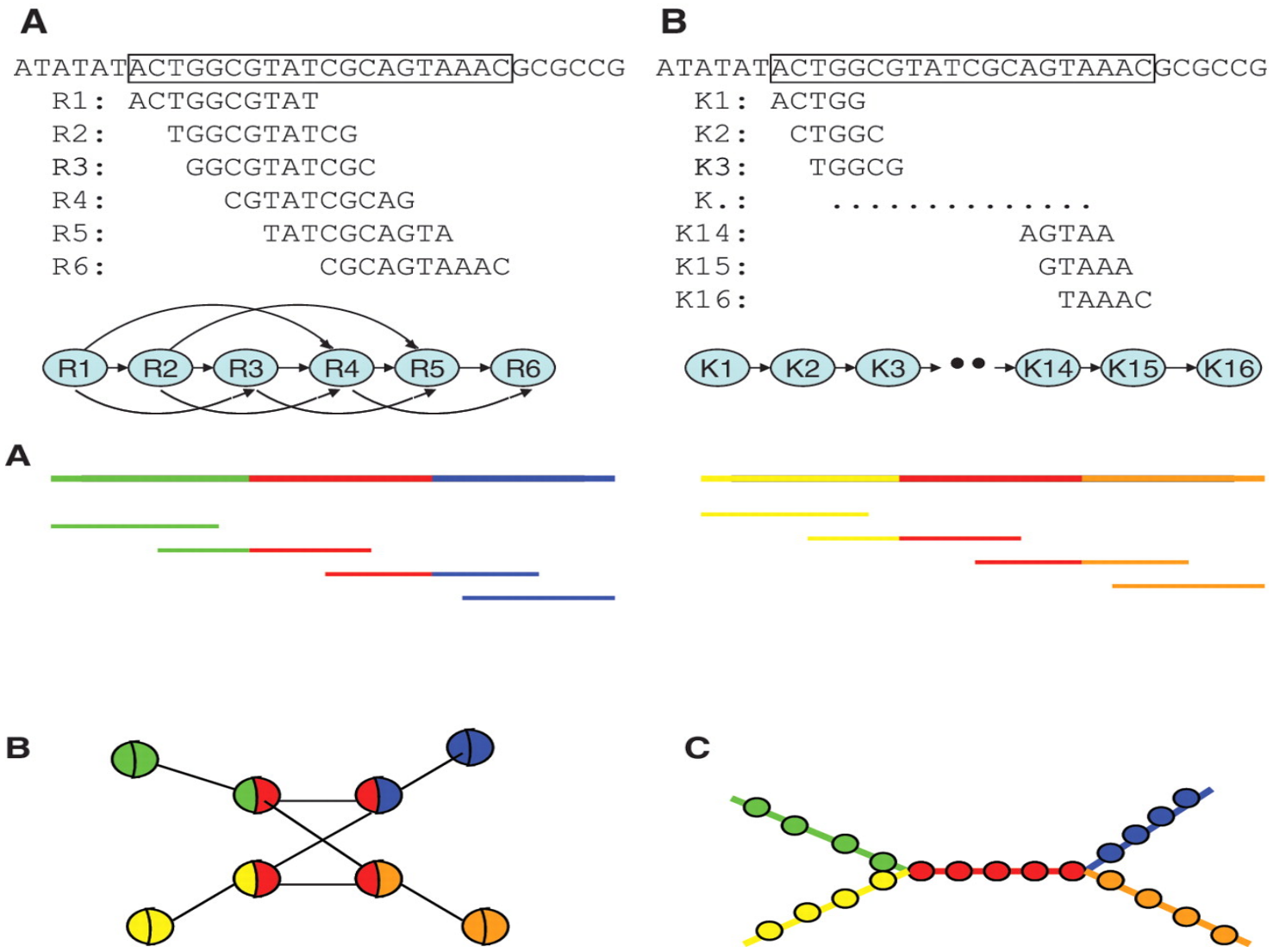
De novo genome assembly: what every biologist should know **Monya Baker** *Nature Methods* **9**, 333–337 (2012) doi:10.1038/nmeth.1935

de Bruijn graph assembly



Nat Rev Genet. 2013 May;14(5):333-46. doi: 10.1038/nrg3433.
Computational solutions for omics data. Berger B1, Peng J, Singh M.

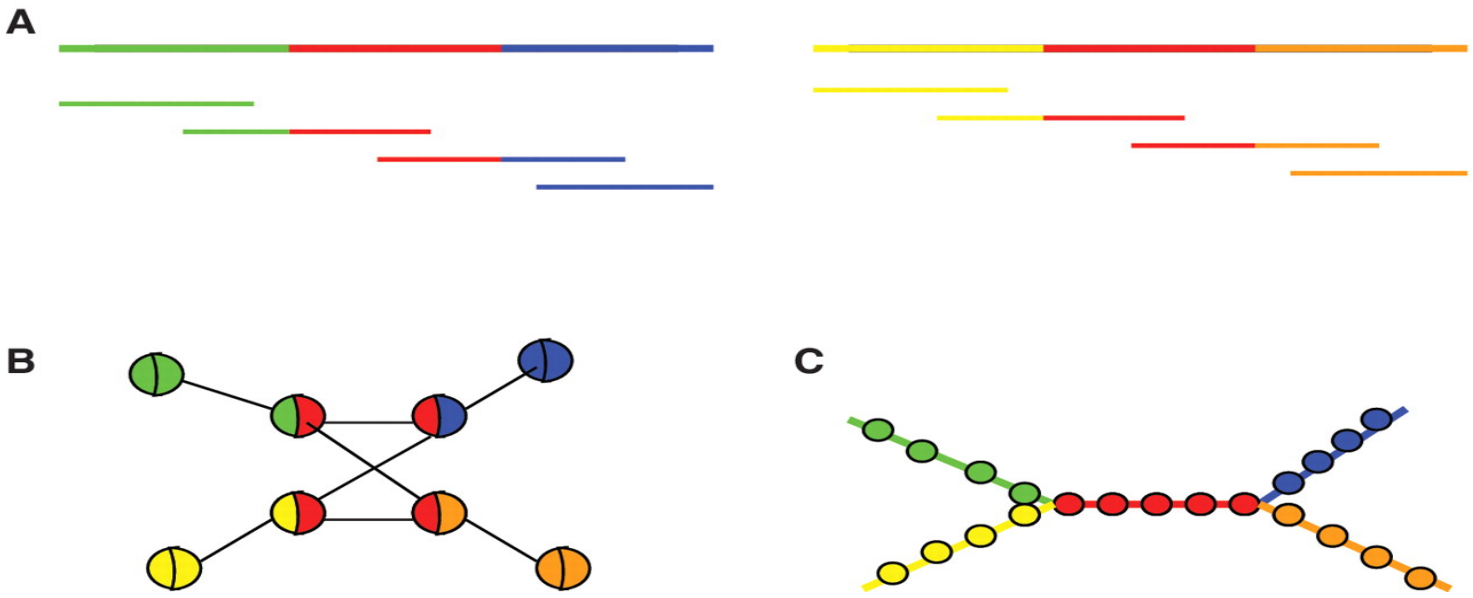
Overlap vs kmer graphs



Slide courtesy of Francesco Vezzi, SciLife Lab

Assembly difficulties

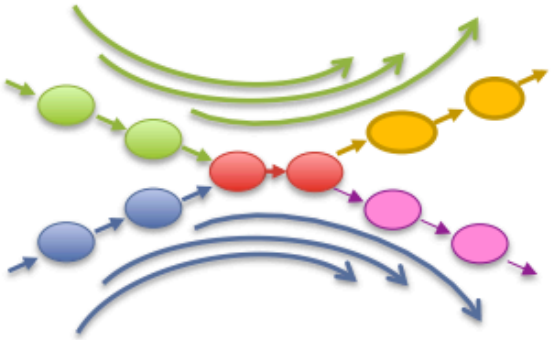
REPEATS



Slide courtesy of Francesco Vezzi, SciLife Lab

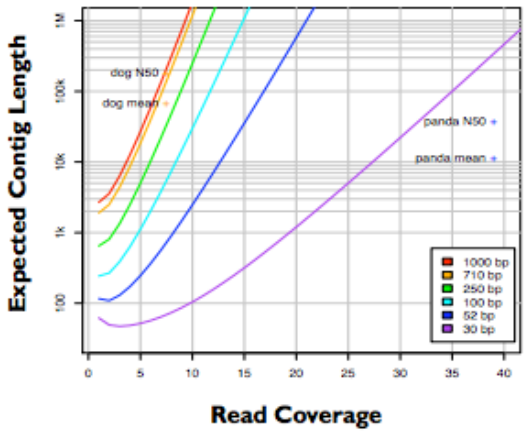
Ingredients for a good assembly

Read Length



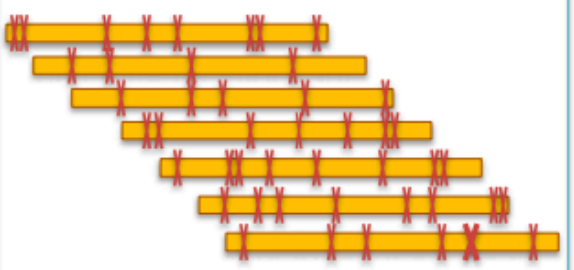
Reads & mates must be longer than the repeats

Coverage



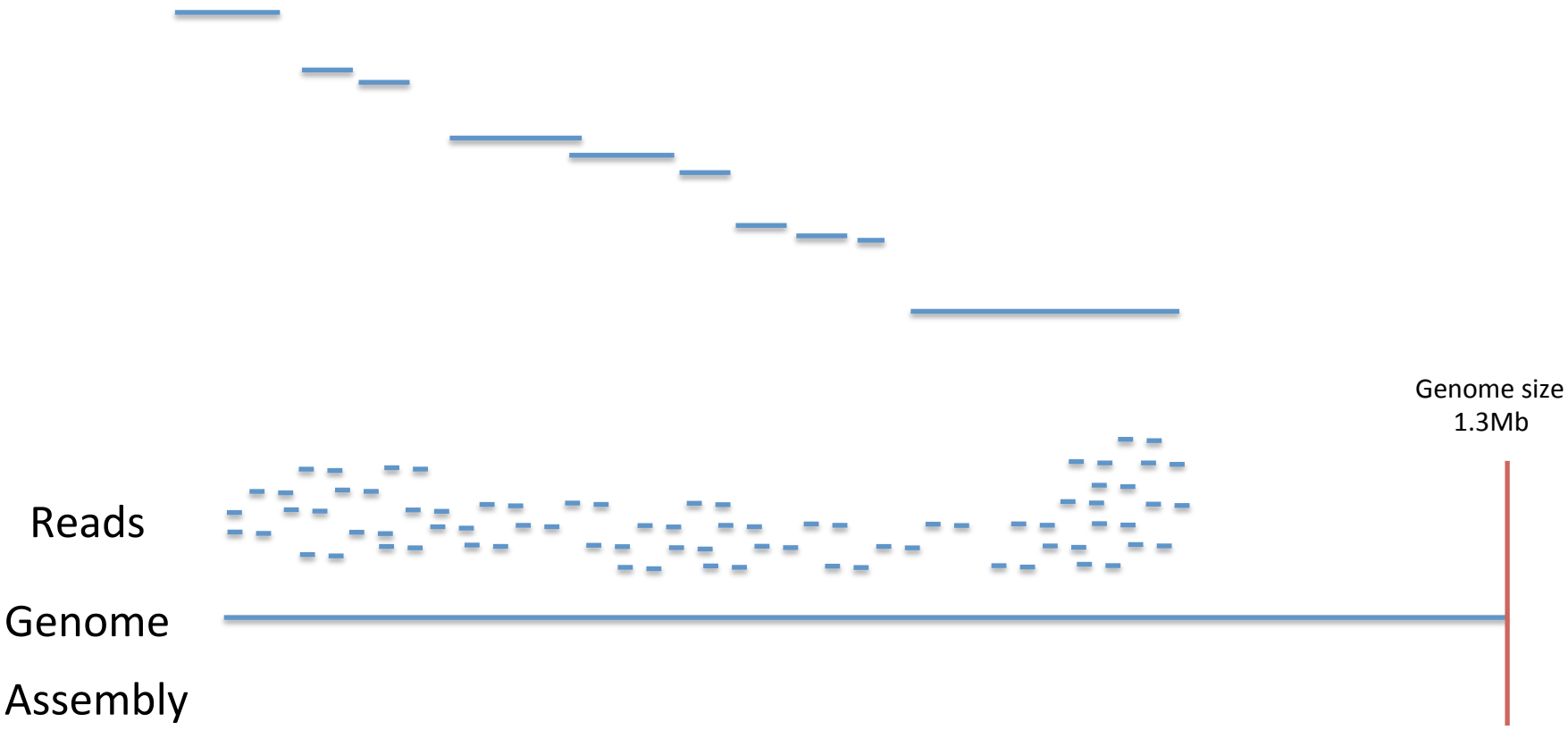
High coverage is required

Quality



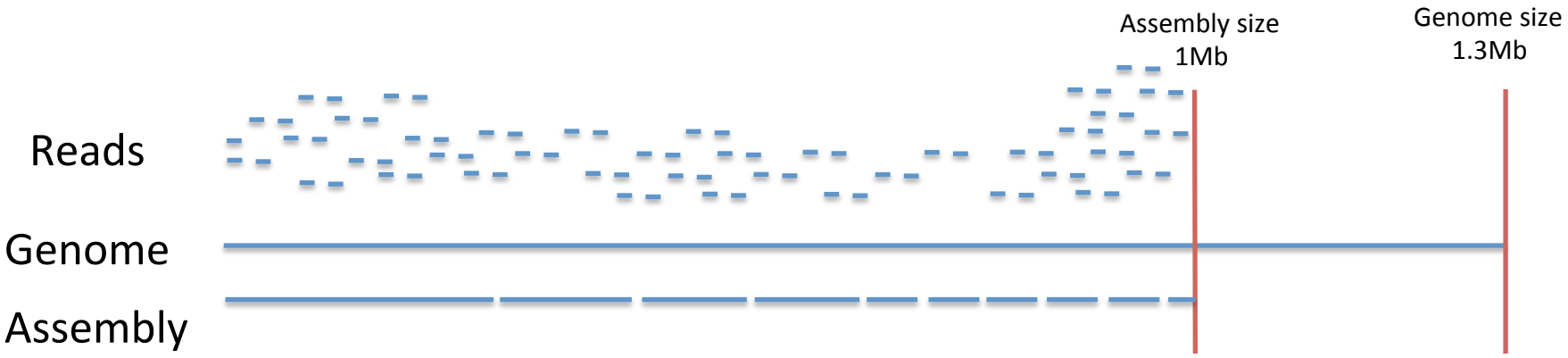
Errors obscure overlaps

Slide courtesy of Francesco Vezzi, SciLife Lab


Genome size
1.3Mb
Reads
Genome
Assembly

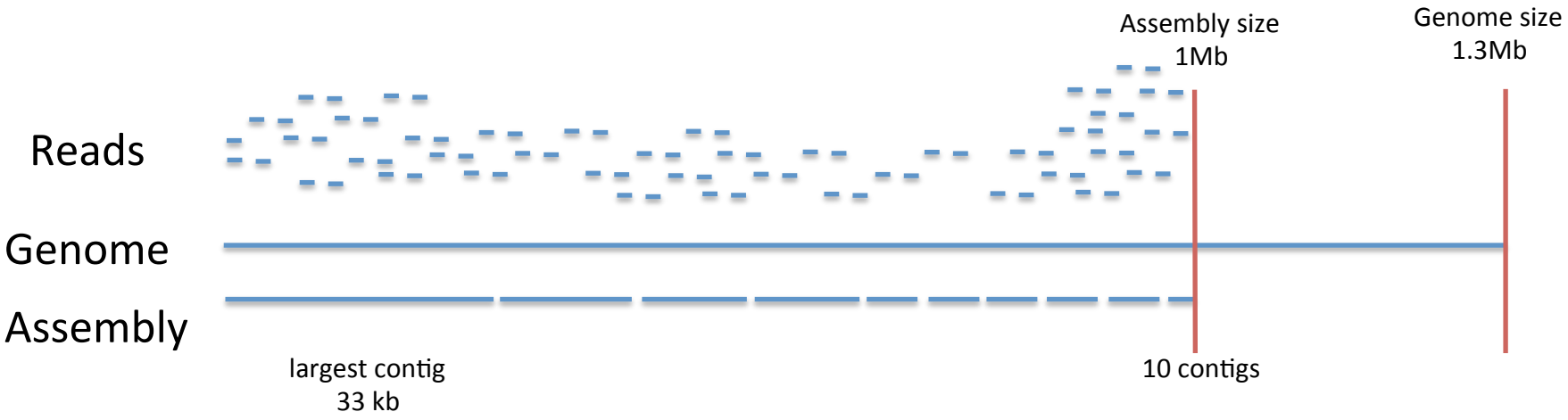
Assembly metrics

- assembly size
- number of contigs, largest contig
- N50



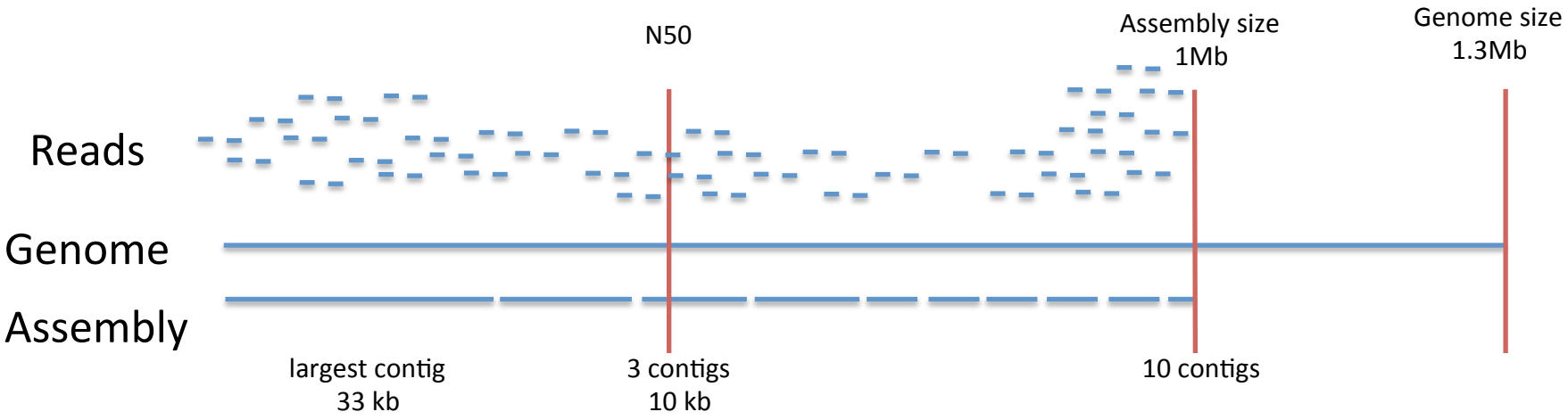
Assembly metrics

- assembly size
- number of contigs, largest contig
- N50



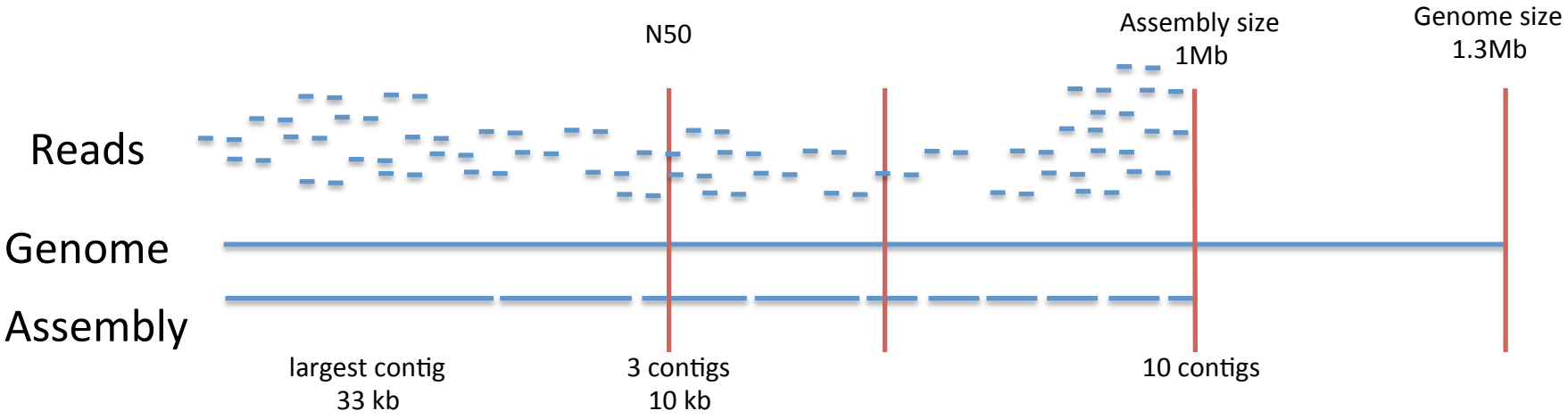
Assembly metrics

- assembly size
- number of contigs, largest contig
- N50



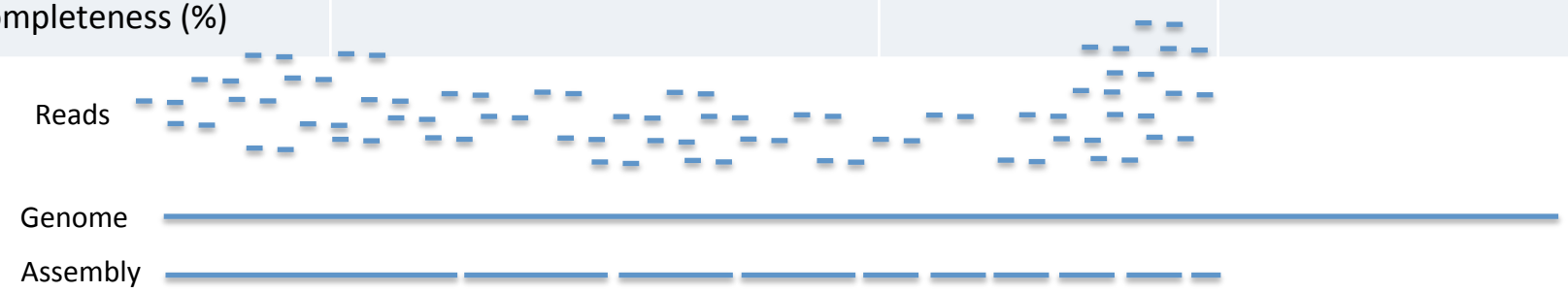
Assembly metrics

- assembly size
- number of contigs, largest contig
- N50



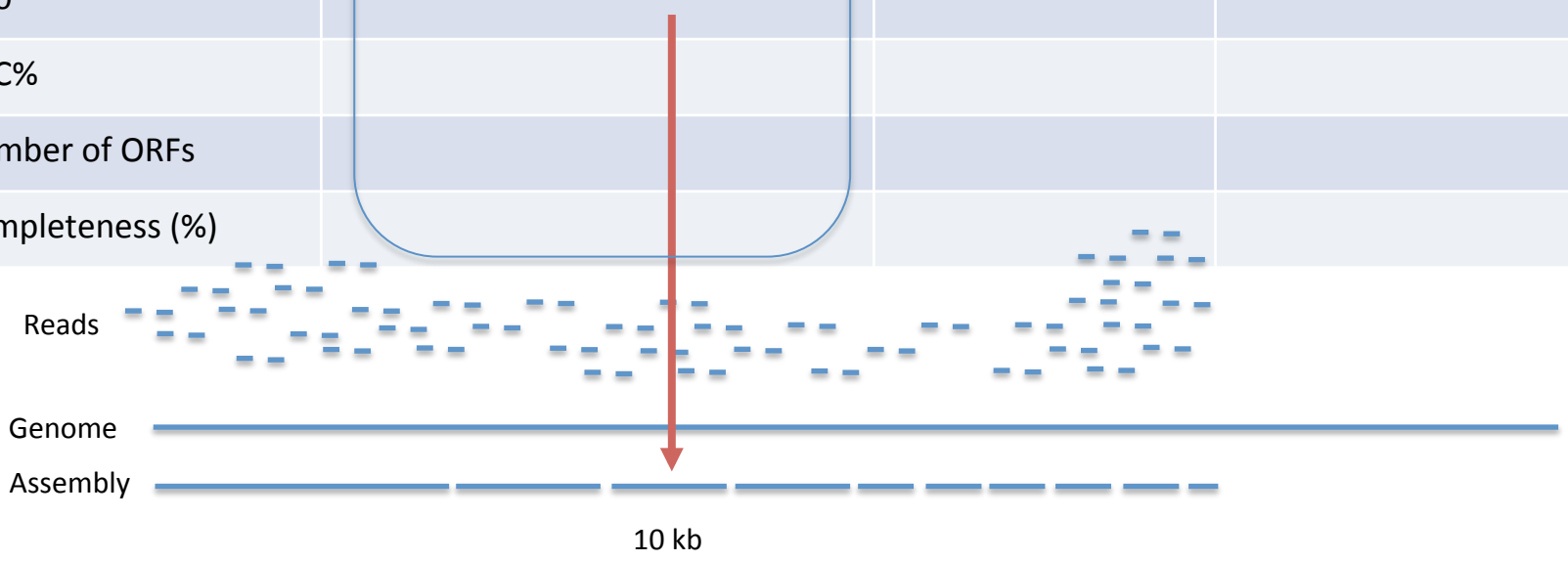
Exercise: compare assemblies

Table 1	HiSeq data (Original data)		
	With --sc and --careful flags	Without --sc flag	Without --careful flag
Number of reads			
Assembly time			
Number of contigs			
Total assembly size			
Largest contig			
N50			
G+C%			
Number of ORFs			
Completeness (%)			



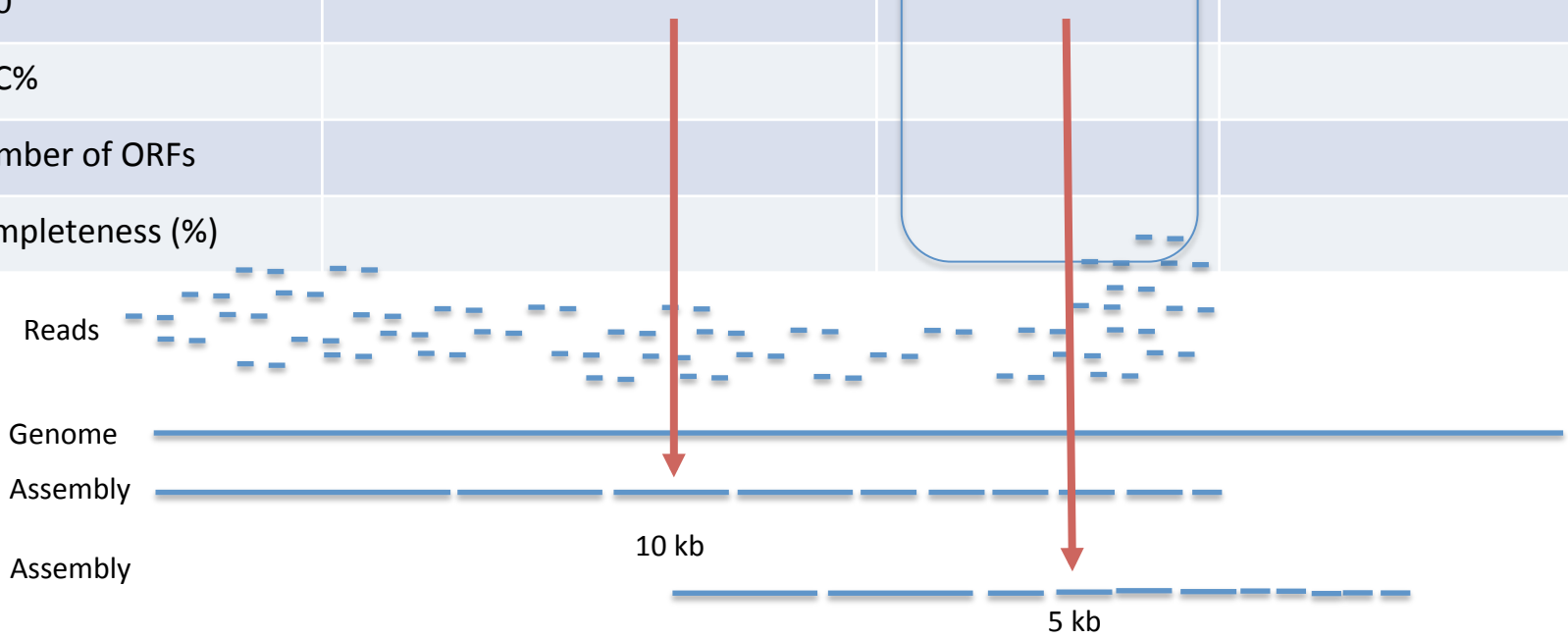
Exercise: compare assemblies

Table 1	HiSeq data (Original data)		
	With --sc and --careful flags	Without --sc flag	Without --careful flag
Number of reads			
Assembly time			
Number of contigs			
Total assembly size			
Largest contig			
N50			
G+C%			
Number of ORFs			
Completeness (%)			

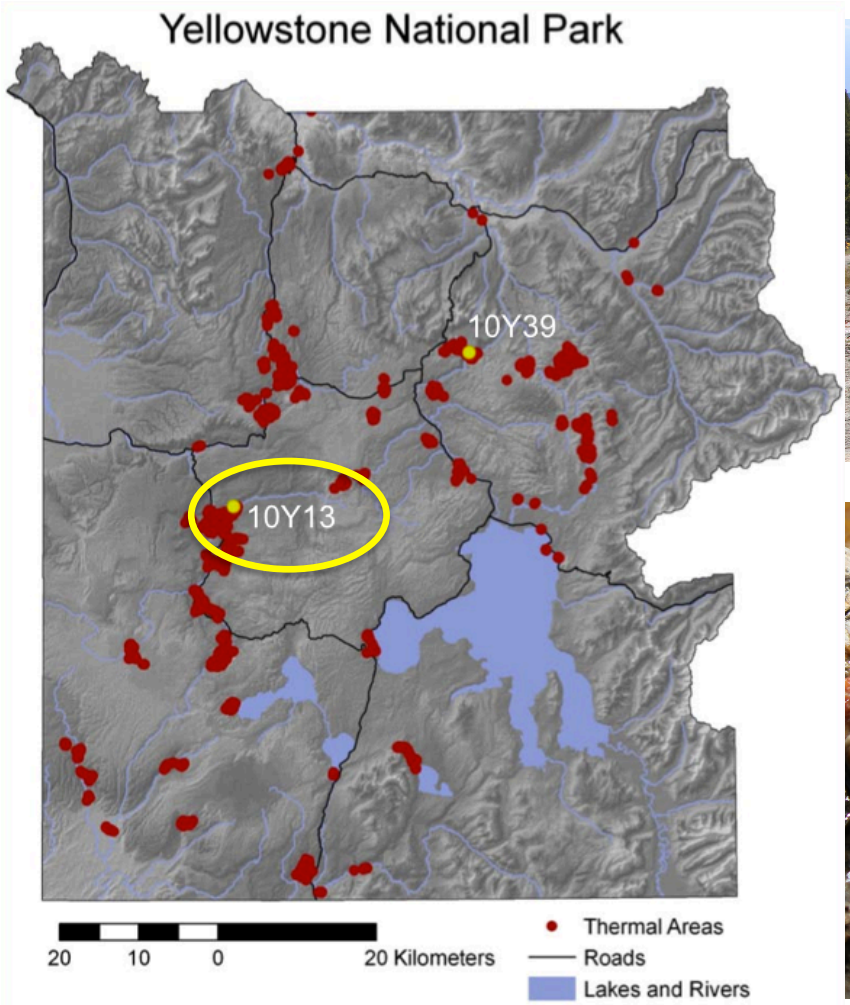


Exercise: compare assemblies

Table 1	HiSeq data (Original data)		
	With --sc and --careful flags	Without --sc flag	Without --careful flag
Number of reads			
Assembly time			
Number of contigs			
Total assembly size			
Largest contig			
N50			
G+C%			
Number of ORFs			
Completeness (%)			



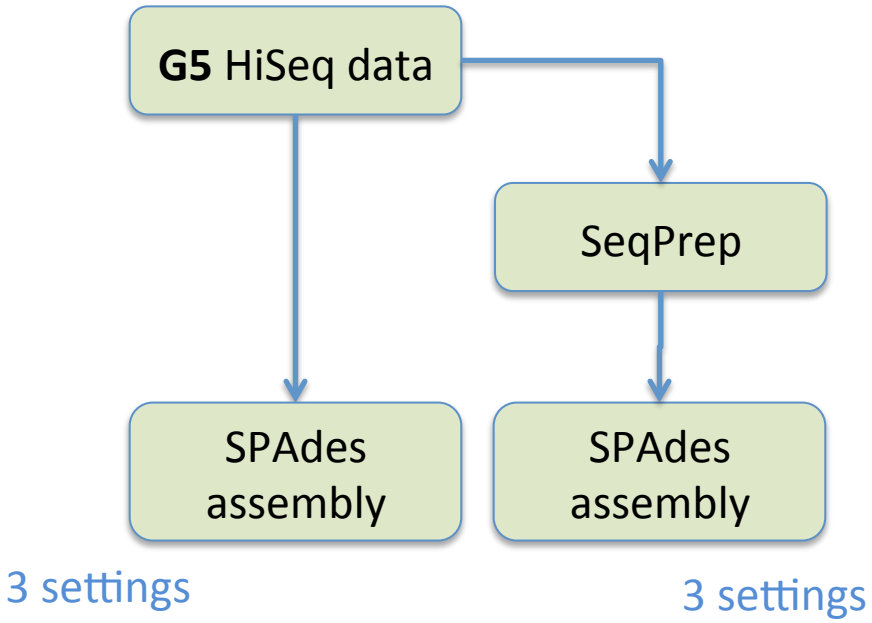
Sample



Culex Basin
pH 8.6, T=68.8°C

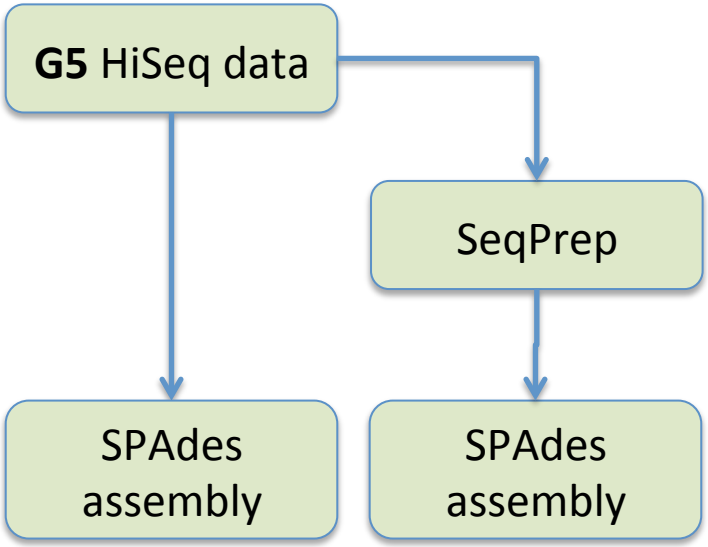
Images on courtesy of Cristina Takacs-Vesbach and Dan Coleman

Overview of exercises today



1. General instructions
2. Familiarizing with data (QC)
3. Single-cell genome assemblies using SPAdes (HiSeq data)

Overview of exercises today

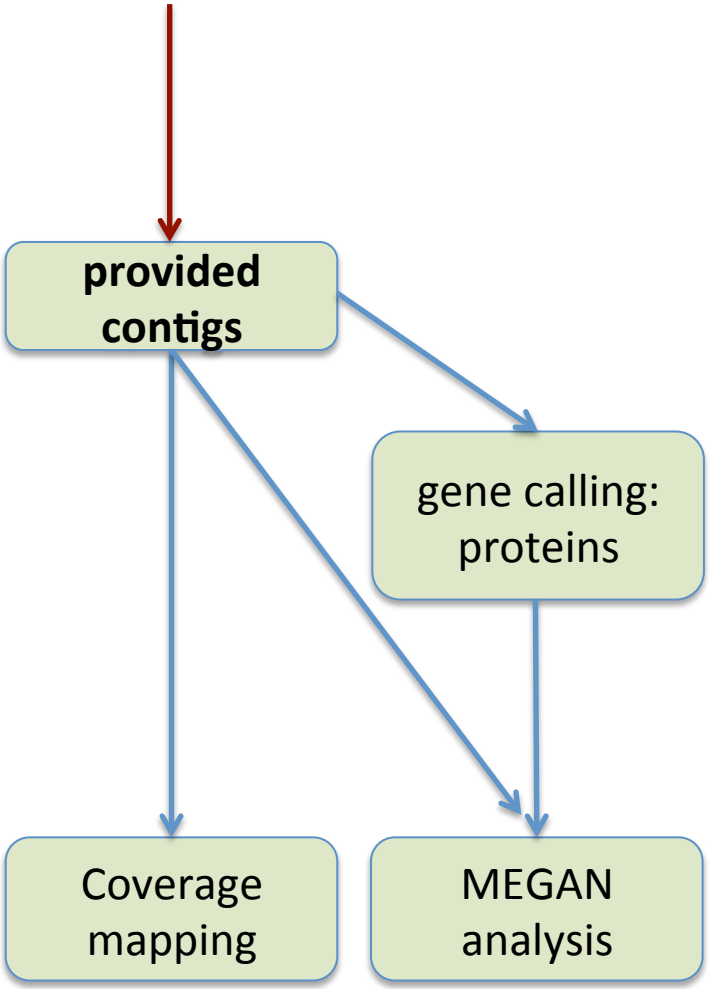


1. General instructions
2. Familiarizing with data (QC)
3. Single-cell genome assemblies using SPAdes (HiSeq data)

3 settings:

1 both --sc and --careful flags
2 only--sc flag
3 only--careful flag

Overview of exercises today

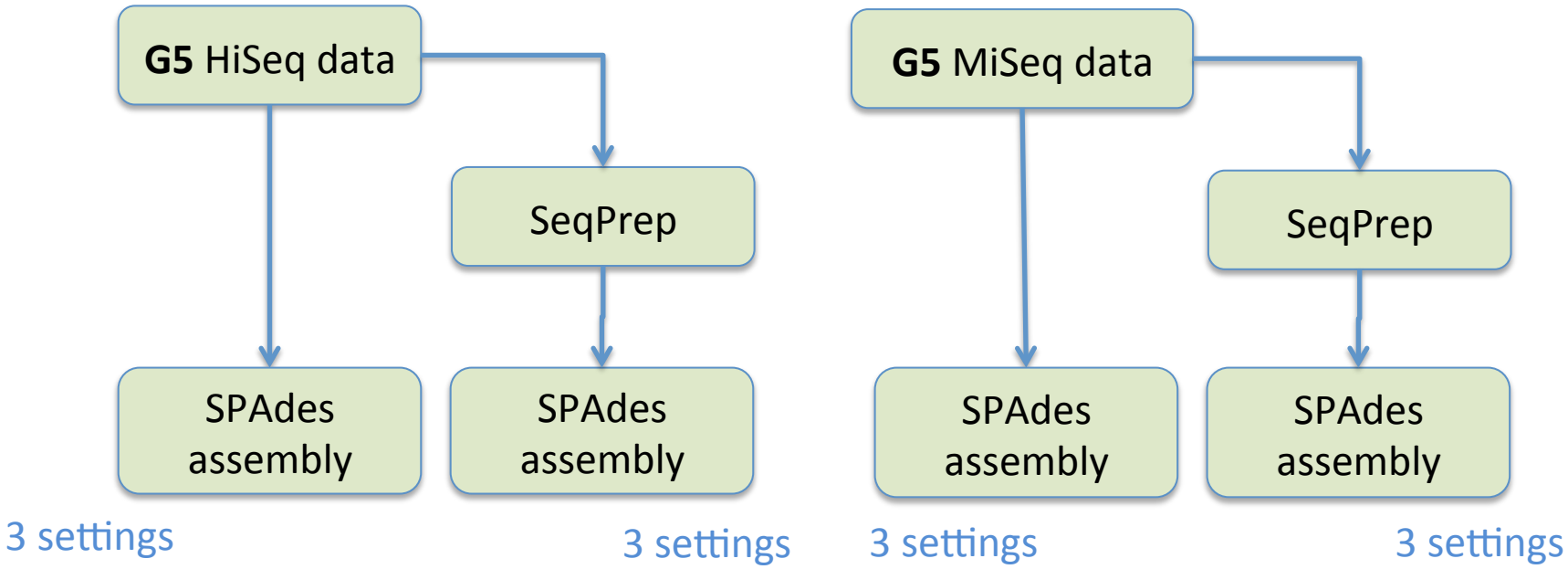


1. General instructions
2. Familiarizing with data (QC)
3. Single-cell genome assemblies using SPAdes (HiSeq data)

PROVIDED CONTIGS

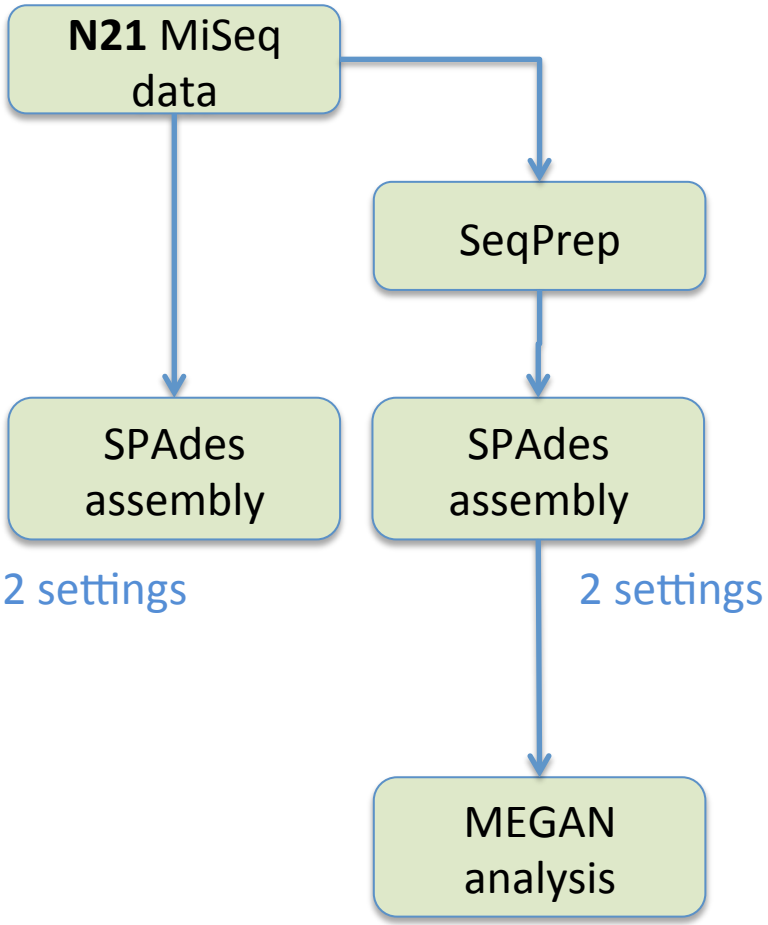
4. Assessing read coverage and chimera checking (with Artemis)
5. Checking for contaminants (with MEGAN)

Overview of exercises today



6. Single-cell genome assembly using SPAdes (MiSeq data) and comparison between HiSeq vs. MiSeq data

Extra exercise
(if there's enough time)



7. Analysis of a novel single-cell genome

Overview of the exercises today

1. General instructions
2. Familiarizing with data
3. Single-cell genome assemblies using SPAdes (HiSeq data)
4. Assessing read coverage and chimera checking (with Artemis)
5. Checking for contaminants (with MEGAN)
6. Single-cell genome assembly using SPAdes (MiSeq data) and comparison between HiSeq vs. MiSeq data
7. Analysis of a novel single-cell genome (bonus exercise)

Part 1-3: Morning session

Part 4-7: Afternoon session

Datasets to be used

From same SAG

- **Dataset1**
  - Paired end **HiSeq** data for **G5**
  - G5_Hiseq_1.fastq
  - G5_Hiseq_2.fastq
- **Dataset2**
  - Paired end **MiSeq** data for **G5**
  - G5_Miseq_1.fastq
  - G5_Miseq_2.fastq
- **Dataset3** (extra exercise)
  - Paired end MiSeq data for **N21**
  - N21_Miseq_1.fastq
  - N21_Miseq_2.fastq

- **6 assemblies**
  - 3 assemblies with original data
  - 3 assemblies with processed data
- **6 assemblies**
  - 3 assemblies with original data
  - 3 assemblies with processed data
- **4 assemblies**
  - 2 assemblies with original data
  - 2 assemblies with processed data

Organization into groups

- Groups
  - 7 groups of 4 people
  - 1 group (of 3?)
- Morning session
  - Playing with the data individually
  - Each person run 3-4 assemblies
- Afternoon session
  - Each person run 3-4 assemblies
  - 4 extra assemblies (bonus exercise) if there's enough time
- Coverage and chimera checking analyses
  - individually
- MEGAN analysis
  - individually

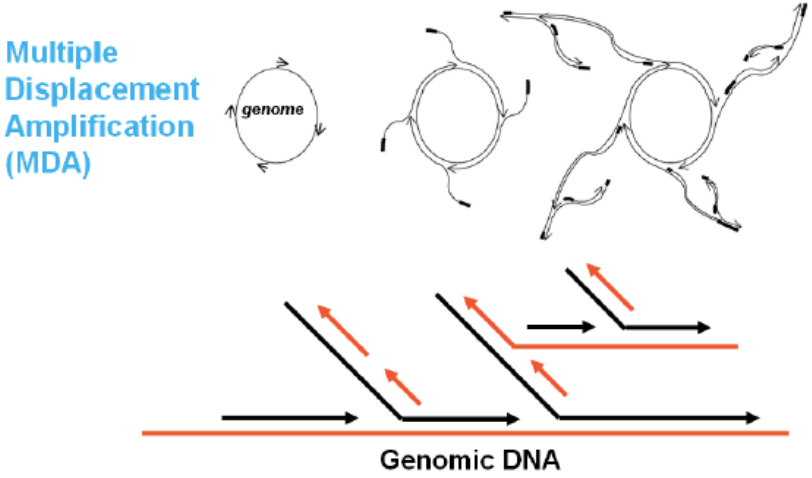
Outline: some theory

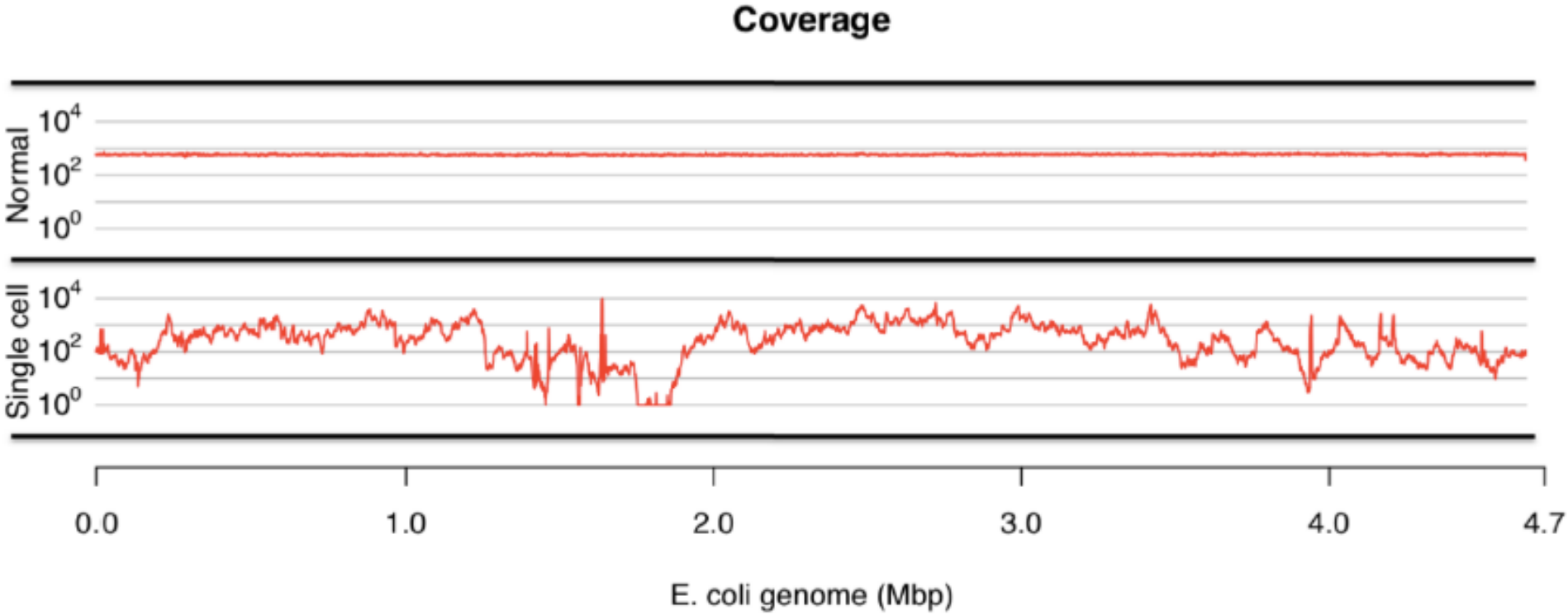
- Assembly basics (contig, scaffold, N50)
- Today's exercise – assembly comparisons
- Background information about the data
- Organization into groups (to get all comparisons)
- Single-cell data - differences compared to multi-cell and metagenomic assembly problems
- Single-cell genome assemblers available
- How SPAdes works

Problems with single-cell data

MDA artefacts

- Chimeras
- Uneven coverage





How does this affect assembly?

- de Bruijn graph sensitive to k-mer quality
- Bad quality k-mers from low-coverage regions
  - Erroneous graph connections → misassemblies
  - Or gaps due to removal of low-coverage areas
- Specialized single-cell genome assemblers are needed

Single-cell genome assemblers available currently

- E+V-SC (Euler+Velvet-SC) (2011)
  - Euler and Velvet modification
  - Not for pairs
  - single k-mer
- IDBA-UD (2012)
  - Error correction
  - Multiple k-mers
  - paired-end reads
- SPAdes (2012)
  - Error correction
  - Multiple k-mers
  - paired-end reads
  - Also tries to solve chimera problems

Why use SPAdes?

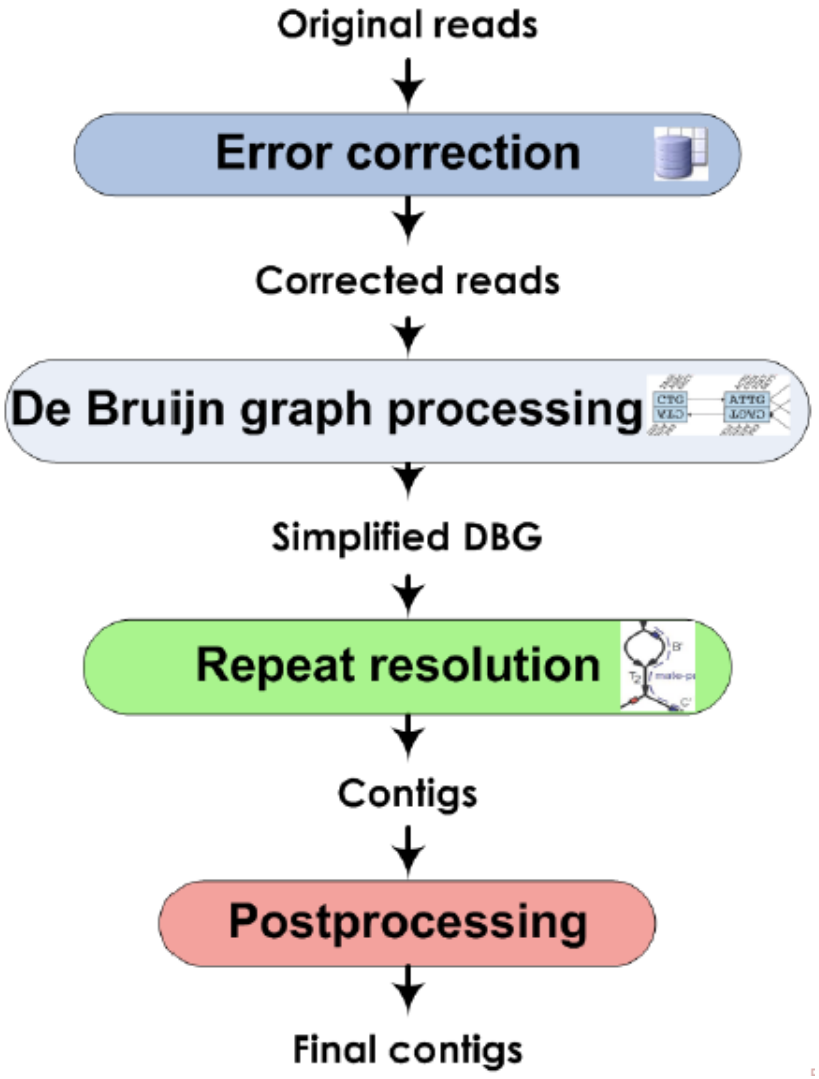
(better assembly results)

Assembly	NG50	# of contigs	Largest contig	Total length	Misassemled contigs	mismatch (bp per 100kbp)	indels (bp per 100kbp)	Mapped genome (%)	# genes
A5	14399	745	101584	4441145	8	12.01	0.17	89.88	3444
ABySS	68534	179	178720	4345617	6	3.32	1.68	88.268	3704
CLC	32506	503	113285	4656964	2	5.53	1.42	92.291	3768
EULER-SR	26662	429	140518	4248713	17	10.87	35.67	84.898	3416
Ray	45448	361	210820	4379139	17	6.29	2.83	88.372	3636
SOAPdenovo	1540	1166	51517	2958144	1	1.87	0.11	57.672	1766
Velvet	22648	261	132865	3501984	2	2.19	1.23	73.765	3080
E+V-SC	**32051**	**344**	**132865**	**4540286**	**2**	**2.33**	**0.73**	**91.744**	**3771**
IDBA-UD contigs	**98306**	**244**	**284464**	**4814043**	**8**	**5.09**	**0.27**	**95.21**	**4045**
IDBA-UD scaffolds	**109057**	**229**	**284464**	**4813609**	**8**	**5.14**	**0.77**	**95.199**	**4052**
SPAdes3.1 contigs	**109059**	**238**	**268493**	**4797090**	**1**	**3.29**	**0.45**	**94.936**	**4036**
SPAdes1.1 scaffolds	**110081**	**233**	**268493**	**4799481**	**1**	**4.02**	**0.64**	**94.959**	**4041**

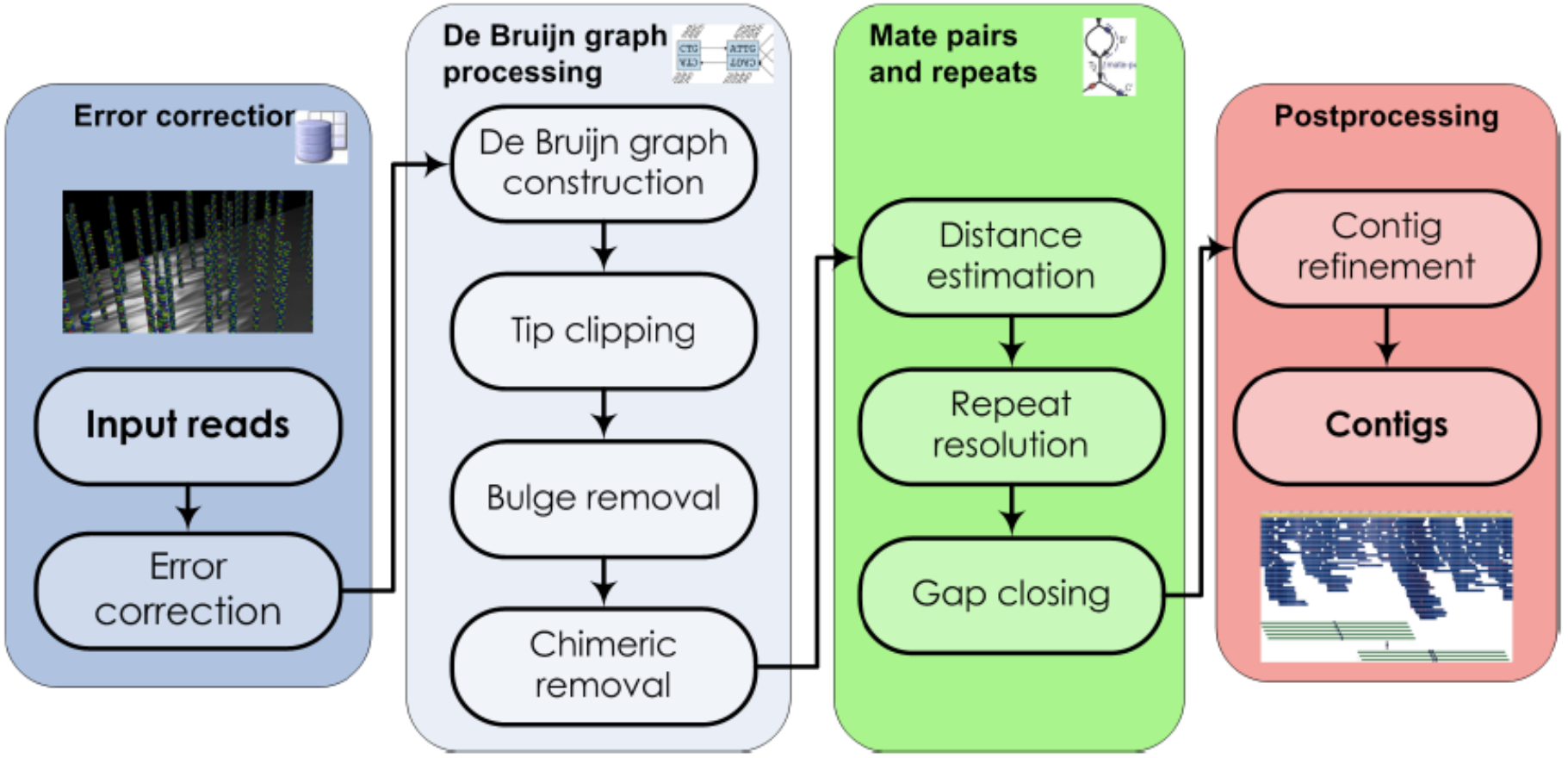
Using *E. coli* single-cell

How does SPAdes achieve this?

- Error correction of reads before assembly
  - *Uses novel algorithm: BayesHammer*
  - *This reduces erroneous k-mers that could mess up assembly*
- Use of multiple k-mers to construct assembly graph
  - *Improved resolution of assembly graphs*
- Uses mate pairs to improve de Bruijn graph construction
  - *Paired de Bruijn graphs ("Rectangle Graphs")*
    - *helps to resolve repeats*
    - *Helps with contig scaffolding*
- Removal of chimeric connections in graph
  - *Less mis-assemblies in the contigs*
- Final correction of errors in contigs (using bwa)
  - *Improved contig quality*
- All these steps in a single command
  - *Other tools need multiple tools to do same procedures*



Details of each step



A few things to consider when using SPAdes

- SPAdes currently only works on Illumina data
  - Other NGS data won't work
- HiSeq data
  - 100-150 bp paired end reads
    - Shorter k-mers
    - Faster assembly
- MiSeq data
  - 250-300 bp paired end reads (longer)
    - Larger k-mers
      - assembly takes longer if smaller k-mers are used
    - User may need to optimize k-mer selection to produce optimal assembly
- In general, it works better with short, high quality reads
- Can also be used for multi-cell genomic data

Acknowledgements

- Jimmy Saw
- Anders Lind (Coverage/chimera checks)
- Joran Martijn (MEGAN analysis)
- Lionel Guy (Genome completeness estimates)
- Thijs Ettema
- Henrik Lantz
- Wenner-Gren Stiftelserna

Why use SPAdes?
(better genome coverage)