

UBC Bioinformatics Class

Topic 5: de novo assembly

Outcomes

- Identify the difference between de novo assembly and reference guided alignment
- Evaluate two different approaches to de novo genome assembly
- Describe how repetitive elements can hamper proper assembly and compare approaches that can overcome this problem
- Describe approaches for transcriptome/ GBS de novo assembly

Introduction



stack of NY Times, June 27, 2000



stack of NY Times, June 27, 2000
on a pile of dynamite



this is just hypothetical



so, what did the June 27, 2000 NY
Times say?

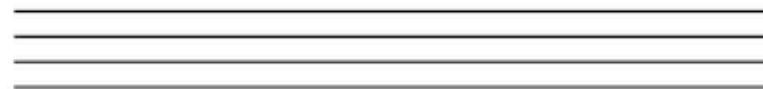
Introduction

atshirt, appre
e have not yet named a
mation is welc

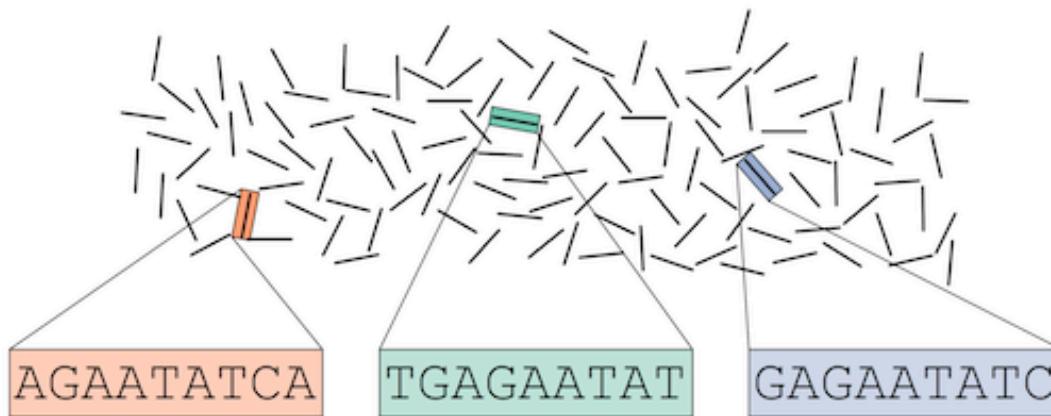
shirt, approximately 6'2" 18
t yet named any suspects
is welcomed. Please ca

Introduction

Multiple identical
copies of a genome



Shatter the genome
into reads



Sequence the reads

Assemble the
genome using
overlapping reads

AGAATATCA
GAGAATATC
TGAGAATAT
... TGAGAATATCA ...

Alignment vs assembly

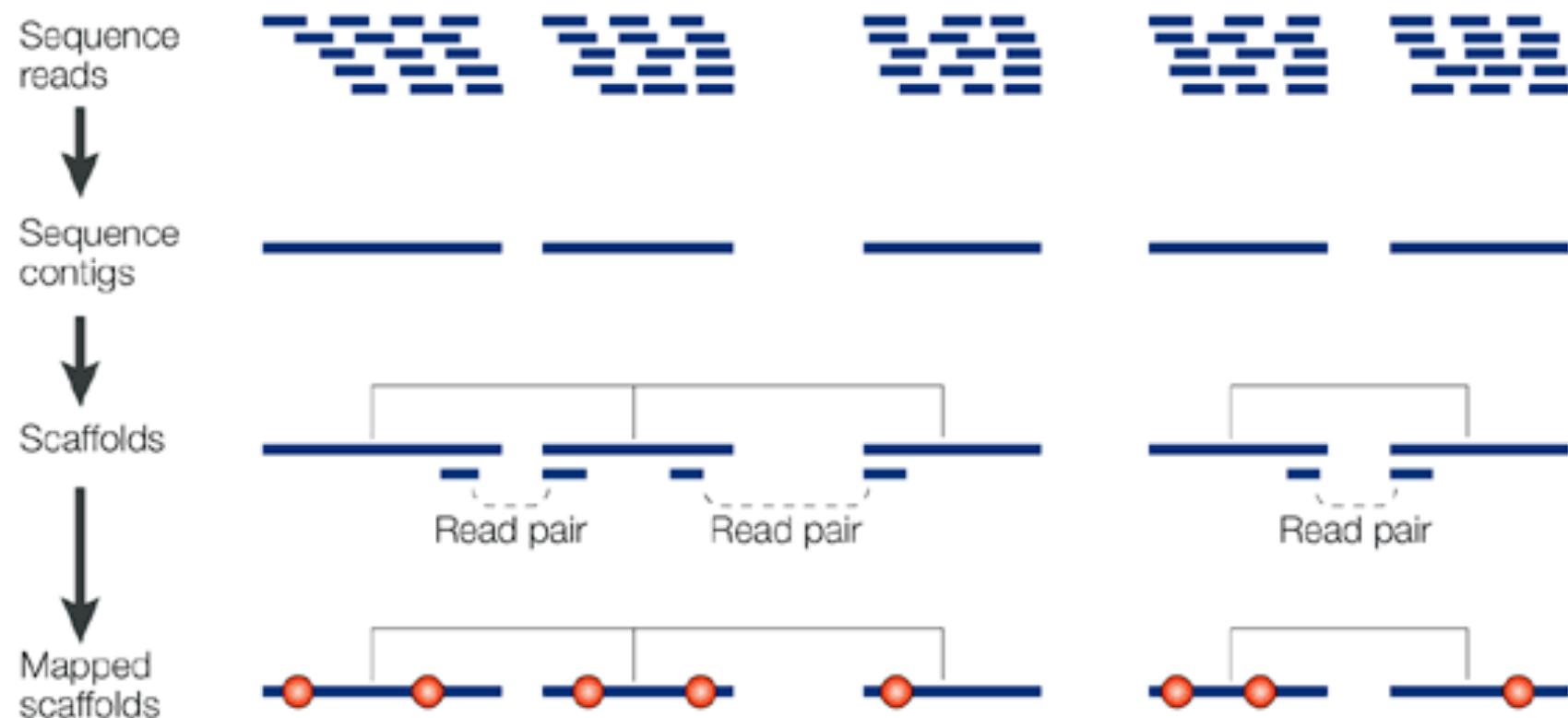
Aligning to a reference:

- Reference guided alignments: align the reads to a reference genome and look for differences

Building a reference:

- De novo assembly: no previous genome assembly is used
- Comparative genome assembly: assemble a newly sequenced genome by mapping it on to a reference
- Hybrid approach: reference-guided and de novo for unused reads or de novo and then reference guided alignments

Introduction



Adams 2008

Introduction

Original sequence

GATAGAAGGGTCCGCTCGCTCAGCTACCGGTTTTATAGATCTA

GATAGAAGGGTCCGCT
AGAAGGGTCCGCTC
GGGTCCGCTCGCTCA
CCGCTCGCTCAGC
CTCGCTCAGCTACC
TCAGCTACCGGTTT
CTACCGGTTTT
AGCTACCGGTTTTAT
TTTTTATAGATCTA



fragmented sequences
from sequencer
(reads)

Introduction

assembled

fragmented sequences
from sequencer
(reads)

TTTTTATAGATCTA
AGCTACCGGTTTTAT
CAGCTACCGGTTTT
TCAGCTACCGGTTT
CTCGCTCAGCTACC
CCGCTCGCTCAGC

GGGTCCGCTCGCTCA
AGAAGGGTCCGCTC
GATAGAAGGGTCCGCT

GATAGAAGGGTCCGCTCGCTCAGCTACCGGTTTTATAGATCTA

We want to reconstruct this from the reads

Introduction

Simplified scenario

- Single strand
- Error free
- Complete coverage

TTTTTATAGATCTA
AGCTACCGGTTTTAT
CAGCTACCGGTTTT
TCAGCTACCGGTT
CTCGCTCAGCTACC
CCGCTCGCTCAGC
GGGTCCGCTCGCTCA
AGAAGGGTCCGCTC
GATAGAAGGGTCCGCT

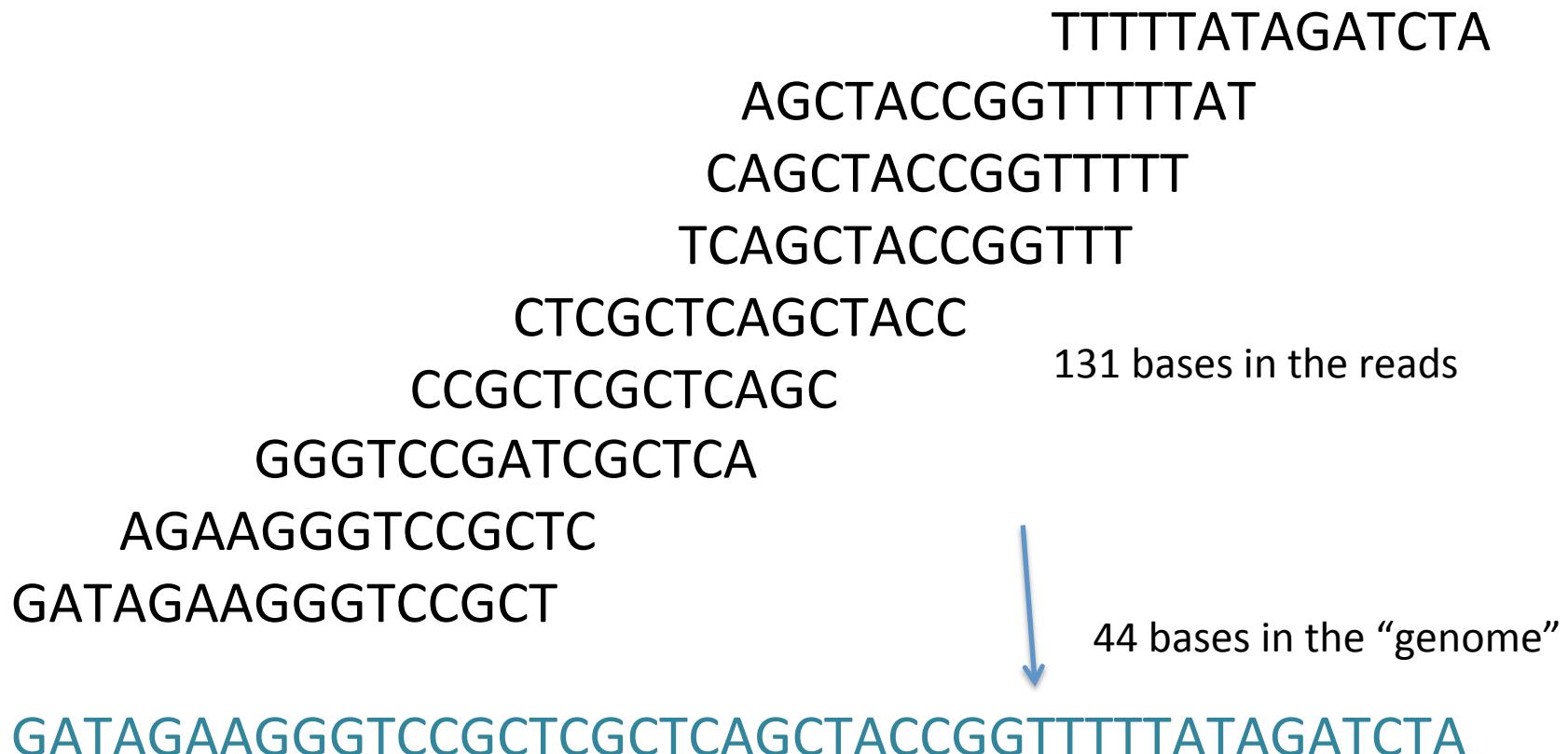
GATAGAAGGGTCCGCTCGCTCAGCTACCGGTTTTATAGATCTA

Introduction

Coverage: reads “covering” a position in the genome
(average or at a single base or region)

TTTTTATAGATCTA
AGCTACCGGTTTTAT
CAGCTACCGGTTTT
TCAGCTACCGGTTT
CTCGCTCAGCTACC
CCGCTCGCTCAGC 131 bases in the reads
GGGTCCGATCGCTCA
AGAAGGGTCCGCTC
GATAGAAGGGTCCGCT

GATAGAAGGGTCCGCTCGCTCAGCTACCGGTTTTATAGATCTA



44 bases in the “genome”

What is our average coverage?
What is the coverage at the arrow?

Introduction

TTTTTATAGATCTA
AGCTACCGGTTTTAT
CAGCTACCGGTTTT
TCAG**G**TACCGGTT
CTCGCTCAGCTACC
CCGCTCGCTCAGC
GGGTCCGATTGCTCA
AGAAGGGTCCGCTC
GATAGAAGGGTCCGCT

GATAGAAGGGTCCGCTCGCTCAGCTACCGGTTTTATAGATCTA

Why might there be differences among reads covering the same position?

Introduction

CCGCTCGCTCAGC
TCAGCTACCGGTTT
CTCGCTCAGCTACC
CAGCTACCGGTTTT
AGAAGGGTCCGCTC
GATAGAAGGGTCCGCT
AGCTACCGGTTTTAT
TTTTTATAGATCTA
GGGTCCGCTCGCTCA

How would you go about “assembling” these reads when you have no reference?

Code break

Write a one liner to find all the overlaps exactly 4 bp in length between **CTCTAGGCC** and a list of other sequences in the file
~/data/Topic5/overlaps.fa

Overlap-layout-consensus

Overlap: make an overlap graph

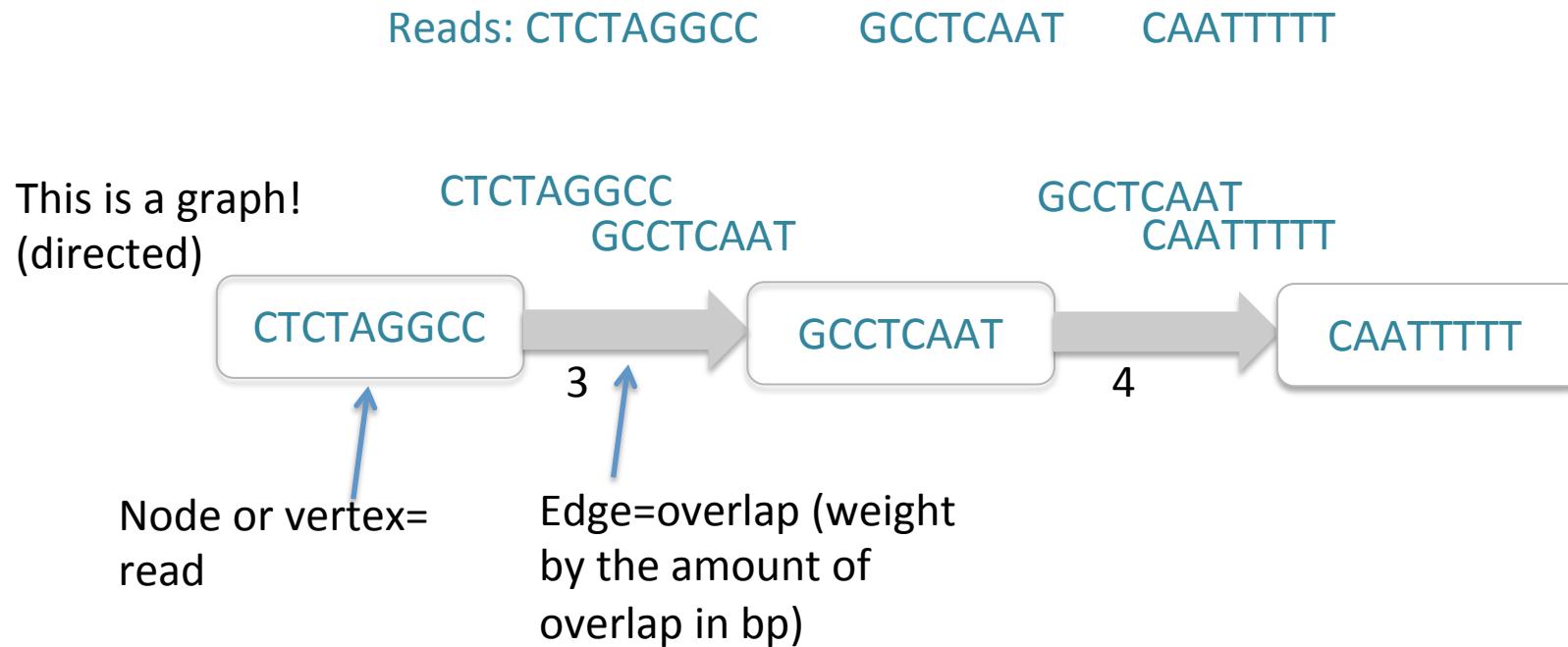
Layout: find the path through the graph

Consensus: find the most likely contig sequence

OLC programs:

ARACHNE, PHRAP, CAP, TIGR, CELERA

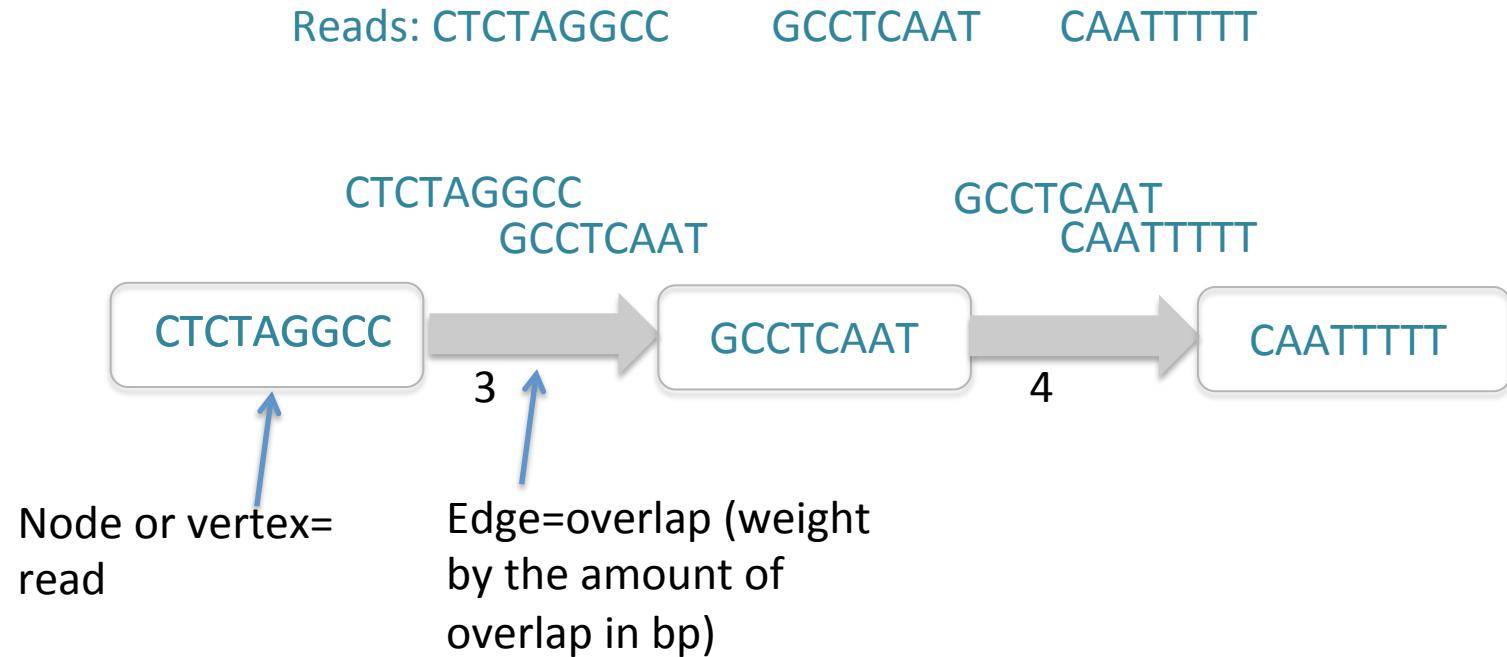
Overlap-layout-consensus



Can pick a minimum overlap length (e.g. 3 bp)

Finding overlaps can be computationally challenging when you have millions of reads!

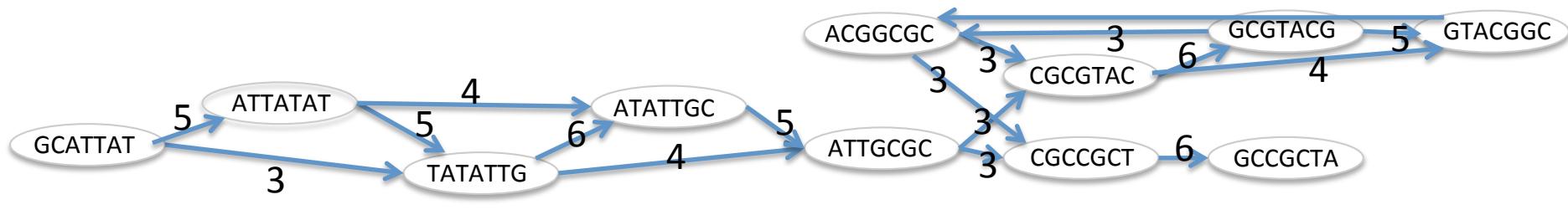
Overlap-layout-consensus



Here we have only one path through the graph

Overlap-layout-consensus

These graphs get complicated!



Minimum
overlap = 3
Read length = 7

GCATTATATATTGCGCGTACGGCGCCGCTACA

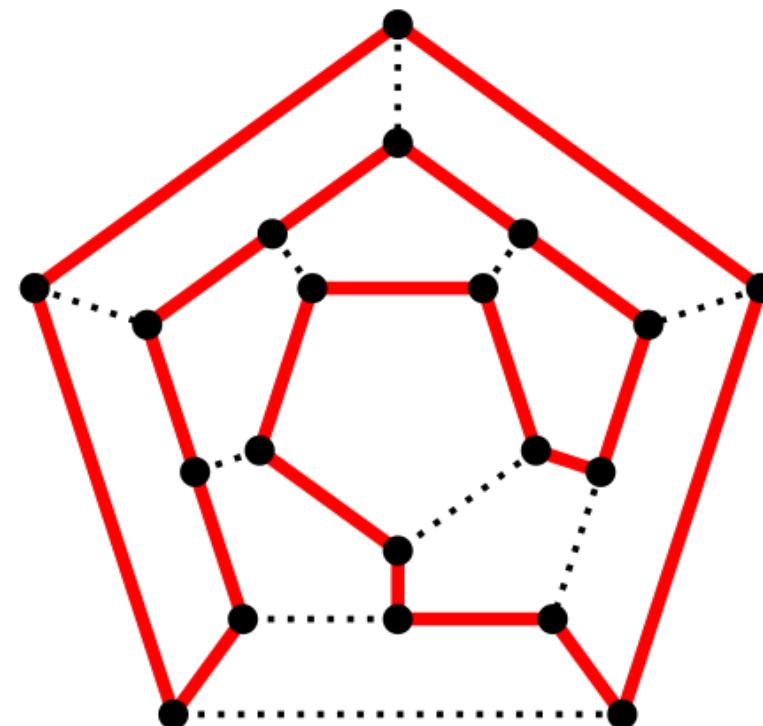
Original sequence

How can we find the best path?

Overlap-layout-consensus

Hamiltonian path: hit each node (read) once

- no quick way to figure it out (NP-complete)
- not practical and not implemented



Overlap-layout-consensus

Shortest superstring: find the shortest final sequence (greatest overlap between reads)

- hit each node (read) once
- NP-hard

Overlap-layout-consensus

Greedy algorithm (example)

- 1) Pairwise alignments between all fragments
- 2) Pick the two with the largest overlap
- 3) Merge chosen fragments
- 4) Repeat



the greedy one

Overlap-layout-consensus

Join sequences together into one sequence

Reads: CTCTAGGCC GCCTCAAT CAATTTTT

CTCTAGGCC
GCCTCAAT

GCCTCAAT
CAATTTTT

CTCTAGGCC

GCCTCAAT

CAATTTTT

3

4



CTCTAGGCC
GCCTCAAT
CAATTTTT
CTCTAGGCCCTCAATT

Overlap-layout-**consensus**

Limitations of OLC

- require overlaps to be scored between all possible pairs of reads. This is a problem when you have millions of reads
- finding the best path through the graph with a huge number of nodes (reads) is computationally challenging

Is there a faster way to assemble many short reads?

De Bruijn graphs

What are all the 5-mers (5 bp fragments) in these reads?

2 reads of 9 bp

read 1

ATGGGGAAC

read 2

GGGAACCCC

ATGGG

GGGAA

TGGGG

GGAAC

GGGGA

GAACC

GGGAA

AACCC

GGAAC

ACCCC

If a read is L bp long, how many kmers of size k can you make?

Code break

Find all the unique 9mers in a fasta sequence and sort them alphabetically `~/data/Topic5/kmer.fa`

1. Find all the kmers in this fasta sequence.

Hints: test out the following commands

`cut -c1- kmer.fa`

`cut -c1-4 kmer.fa`

```
for num in {1..10}  
do  
echo $num >> file.txt  
done
```

2. Sort them and keep the unique ones

Hint: try `sort`

De Bruijn graphs

- Join up all the k-mers (length = k bp) into a graph with an overlap of k-1 (here k=5)



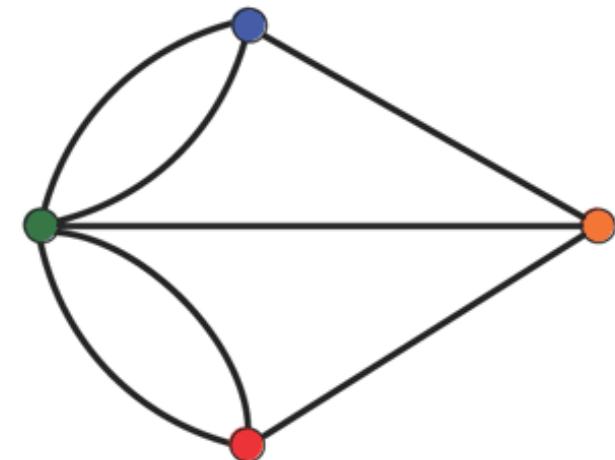
- Traverse through the graph
- The first base of each node spells out the sequence

De Bruijn graphs

a

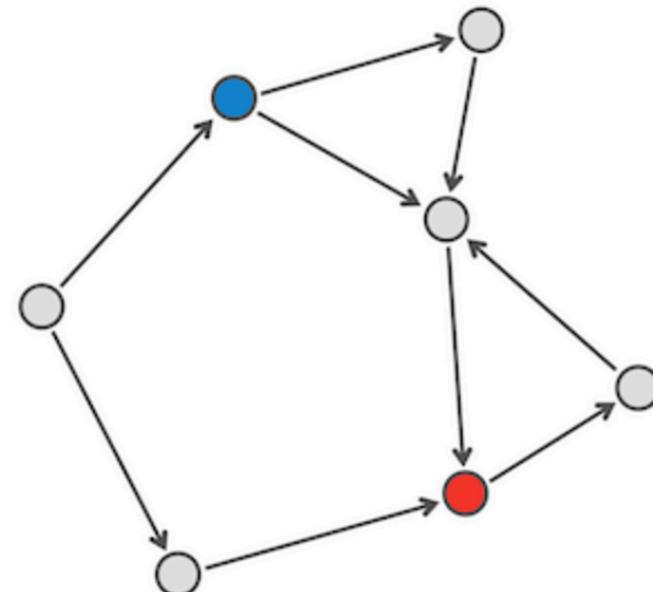
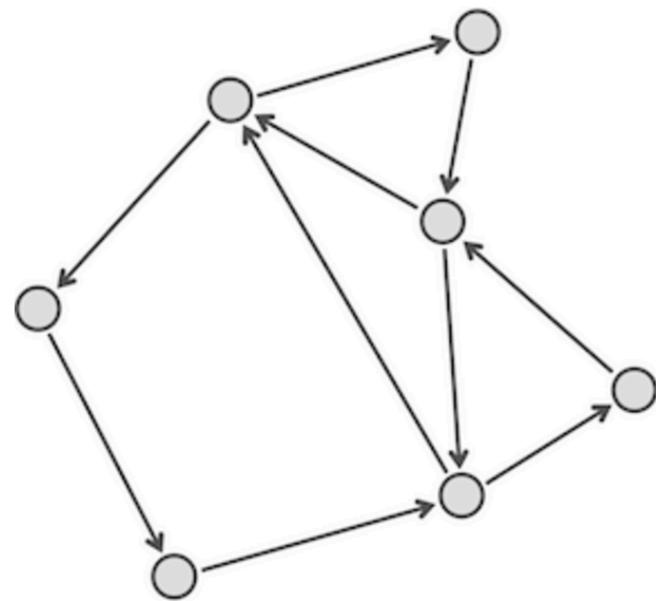


b

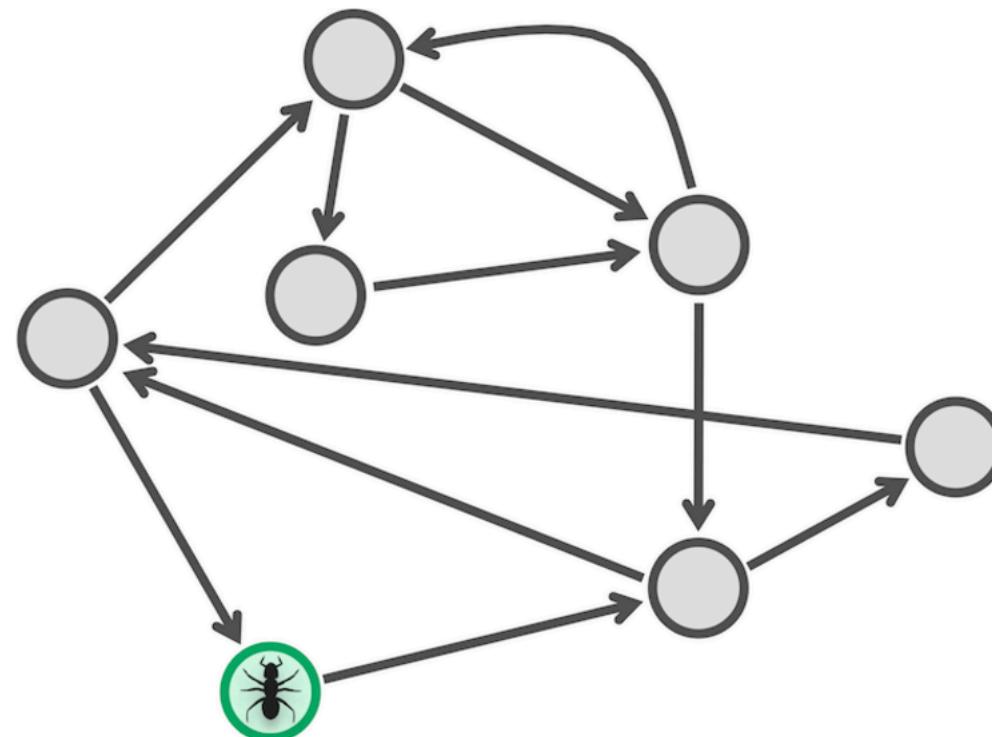


De Bruijn graphs

Eulerian graph must be both balanced and strongly connected

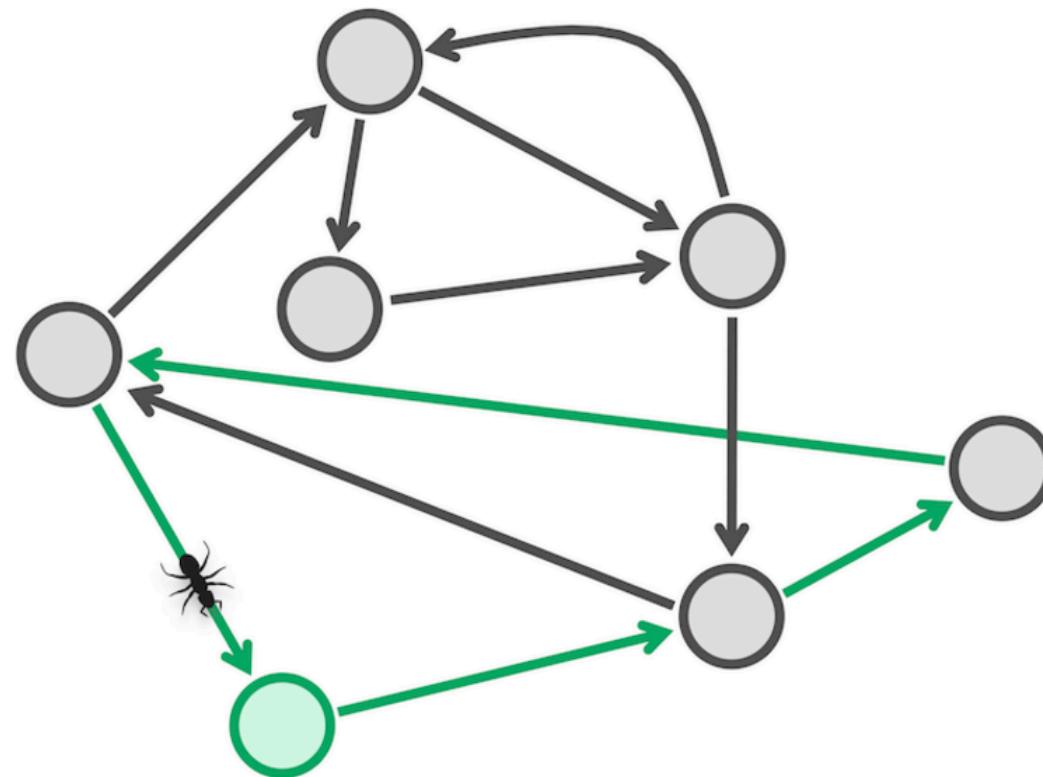


De Bruijn graphs



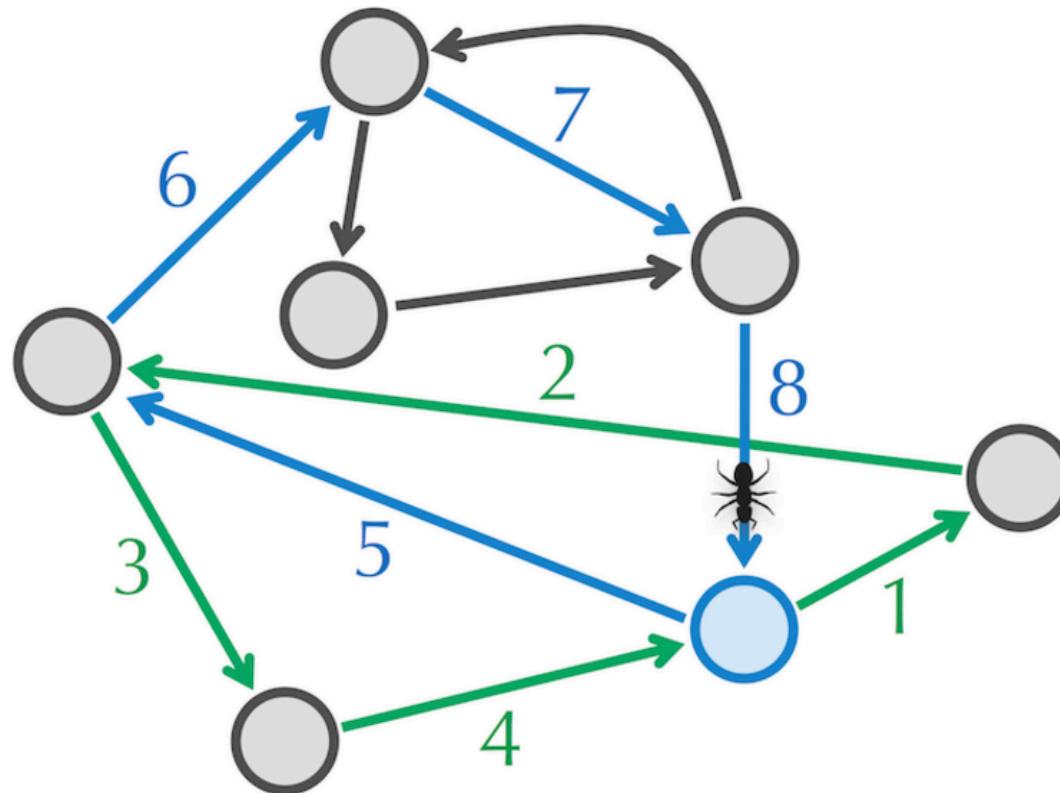
Algorithm to find a path through an Eulerian graph

De Bruijn graphs



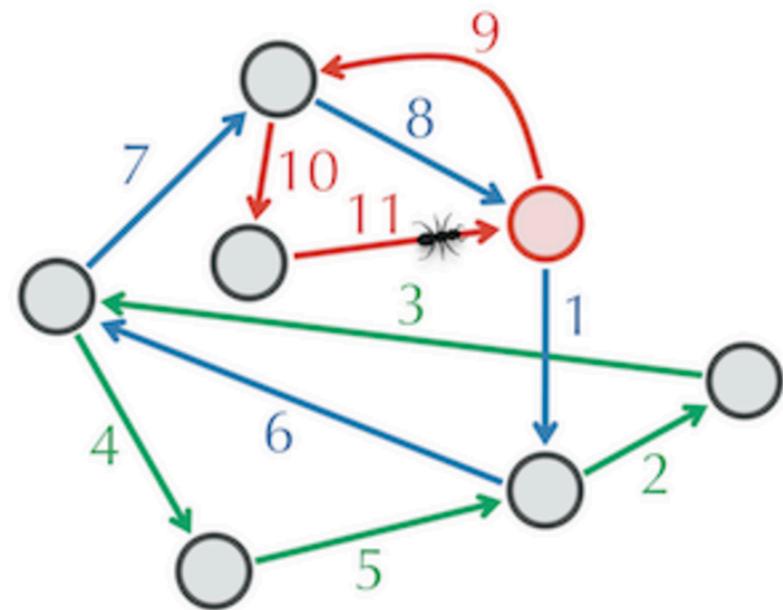
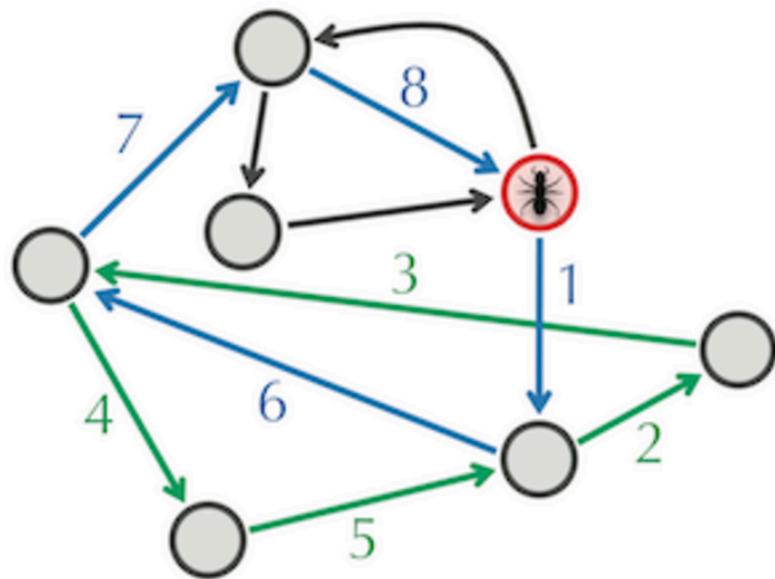
Algorithm to find a path through an Eulerian graph

De Bruijn graphs



Algorithm to find a path through an Eulerian graph

De Bruijn graphs



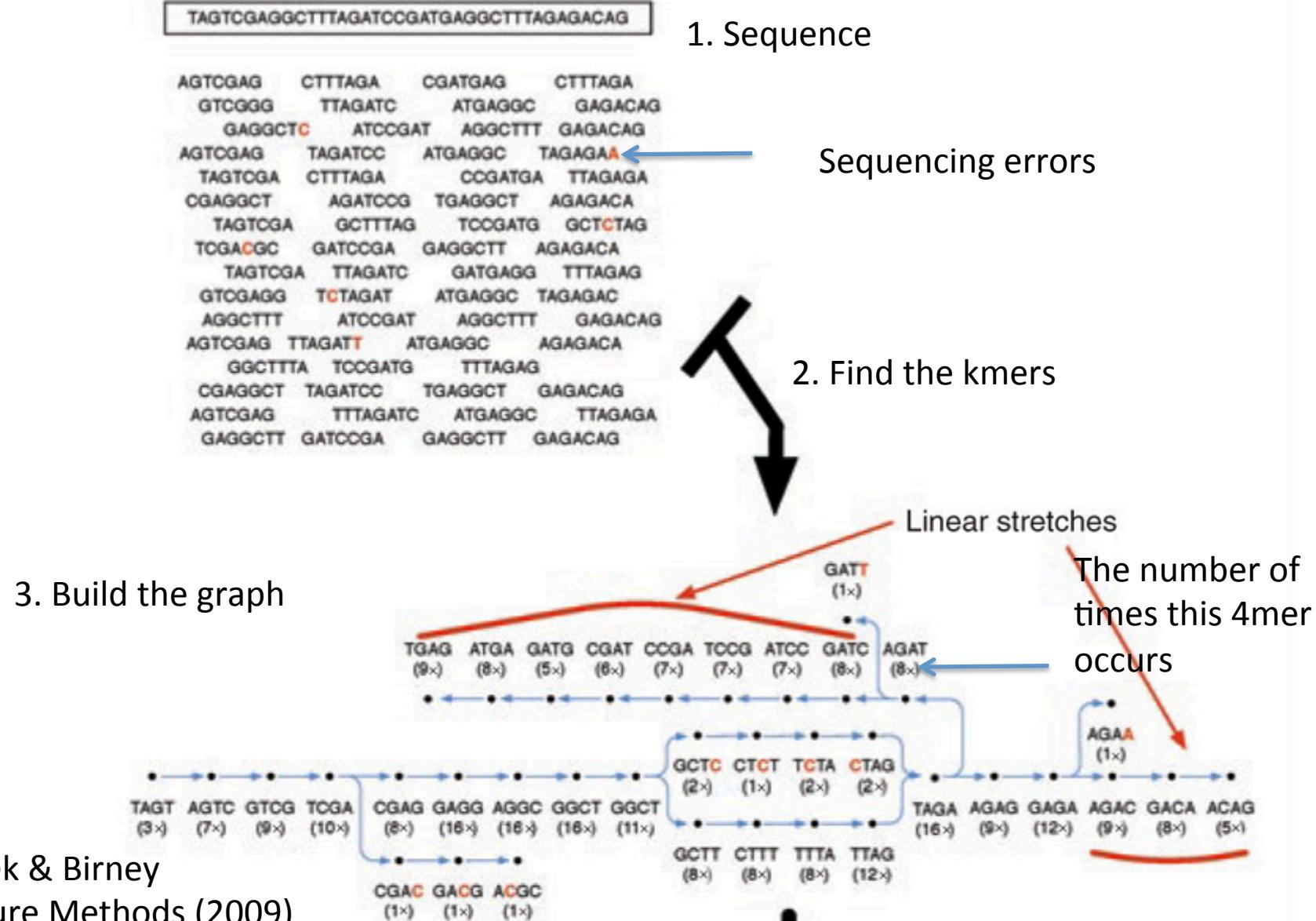
Algorithm to find a path through an Eulerian graph

De Bruijn graphs

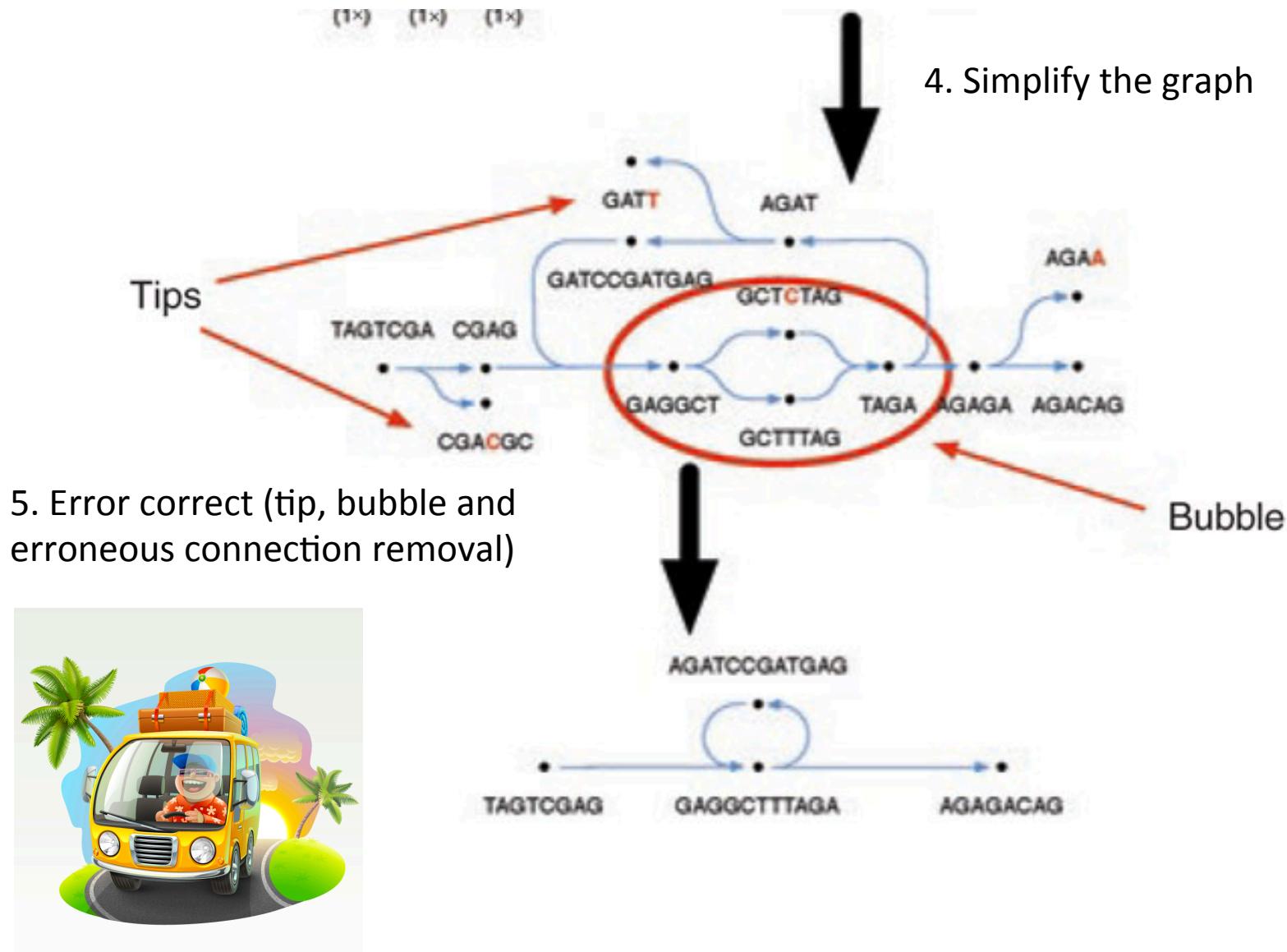
Limitations of the Eulerian path:

- With “perfect” genomic data there are usually many Eulerian tours
- Data is not perfect (areas of low coverage, errors, repeats, etc.)

De Bruijn graphs



De Bruijn graphs



Flicek & Birney Nature Methods (2009)

De Bruijn graphs

Advantages:

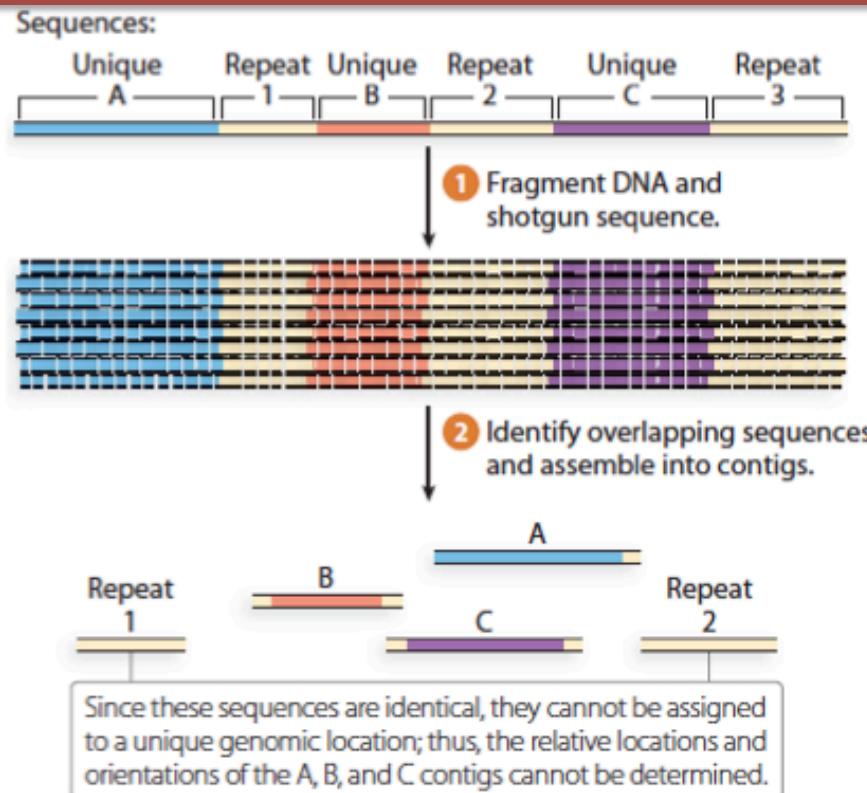
- 1) Set node length (no overlap algorithm)
- 2) Easy approaches for traversing through the graph
- 3) Simpler representation of repeats in the graph

Disadvantages:

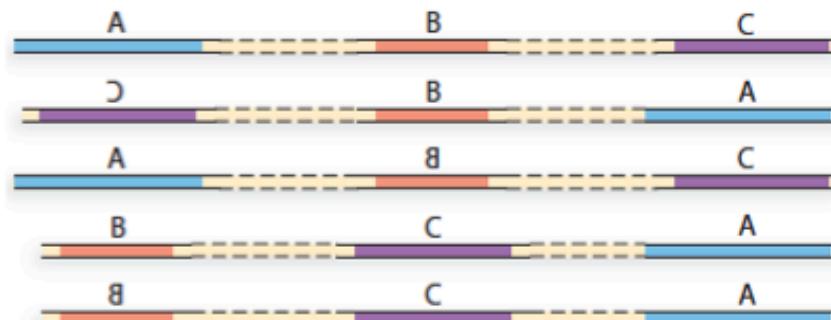
- 1) Lose information
- 2) Shorter contigs

For PacBio and other long read sequences, what type of assembly strategy would you use?

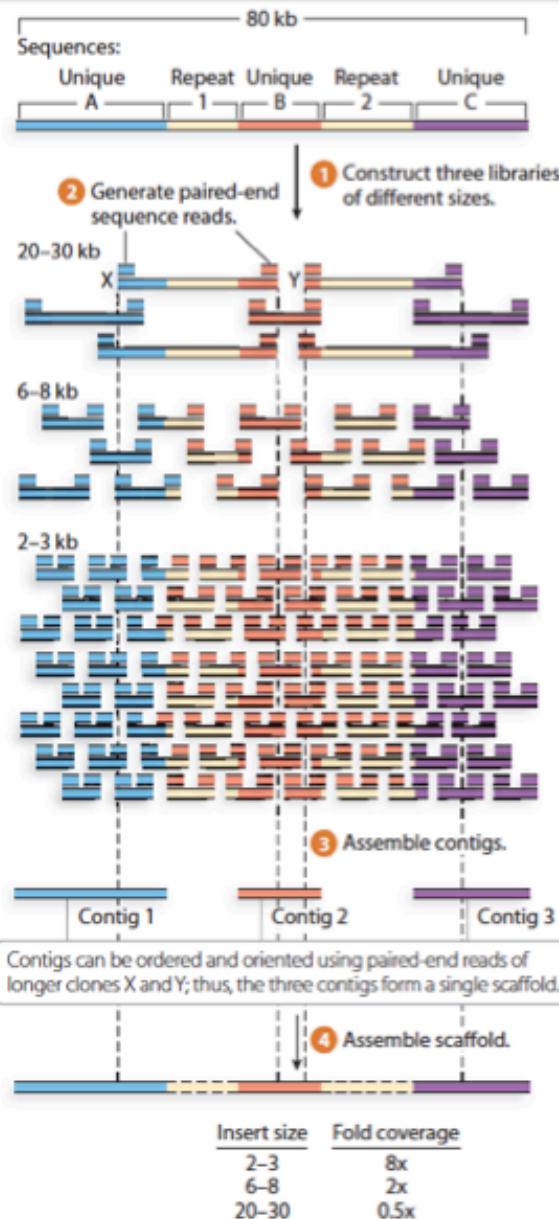
Repetitive regions



Some possible assemblies:



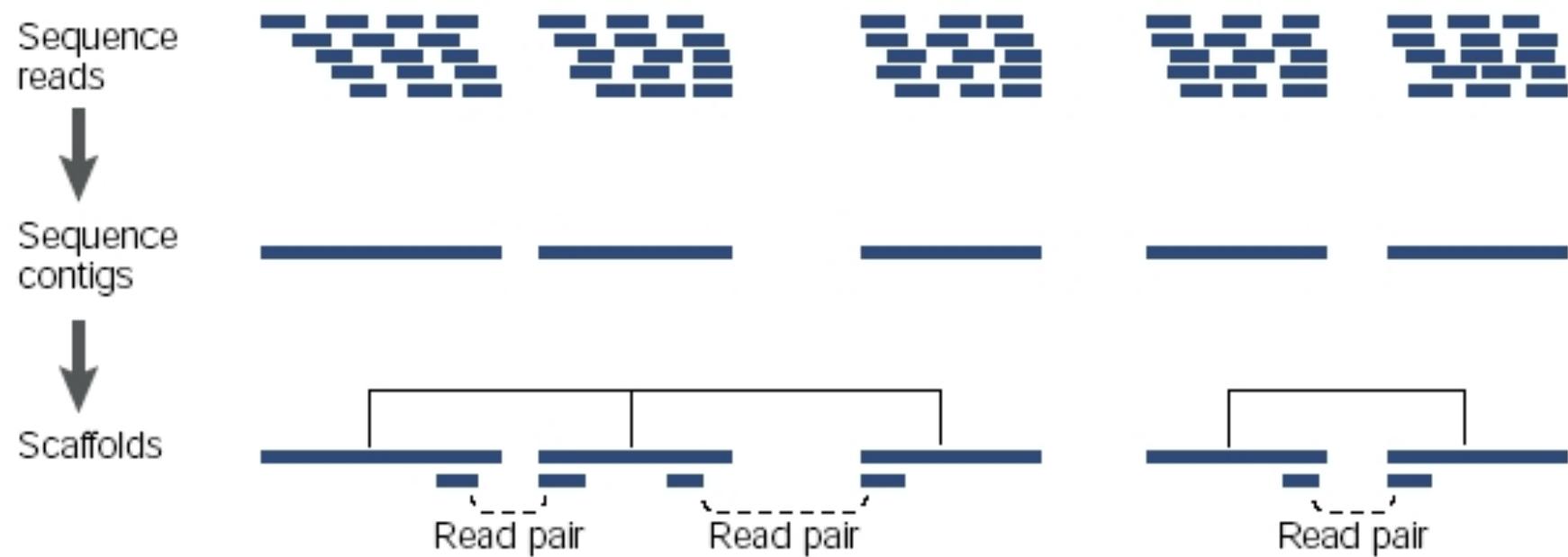
Repetitive regions



Finishing genomes

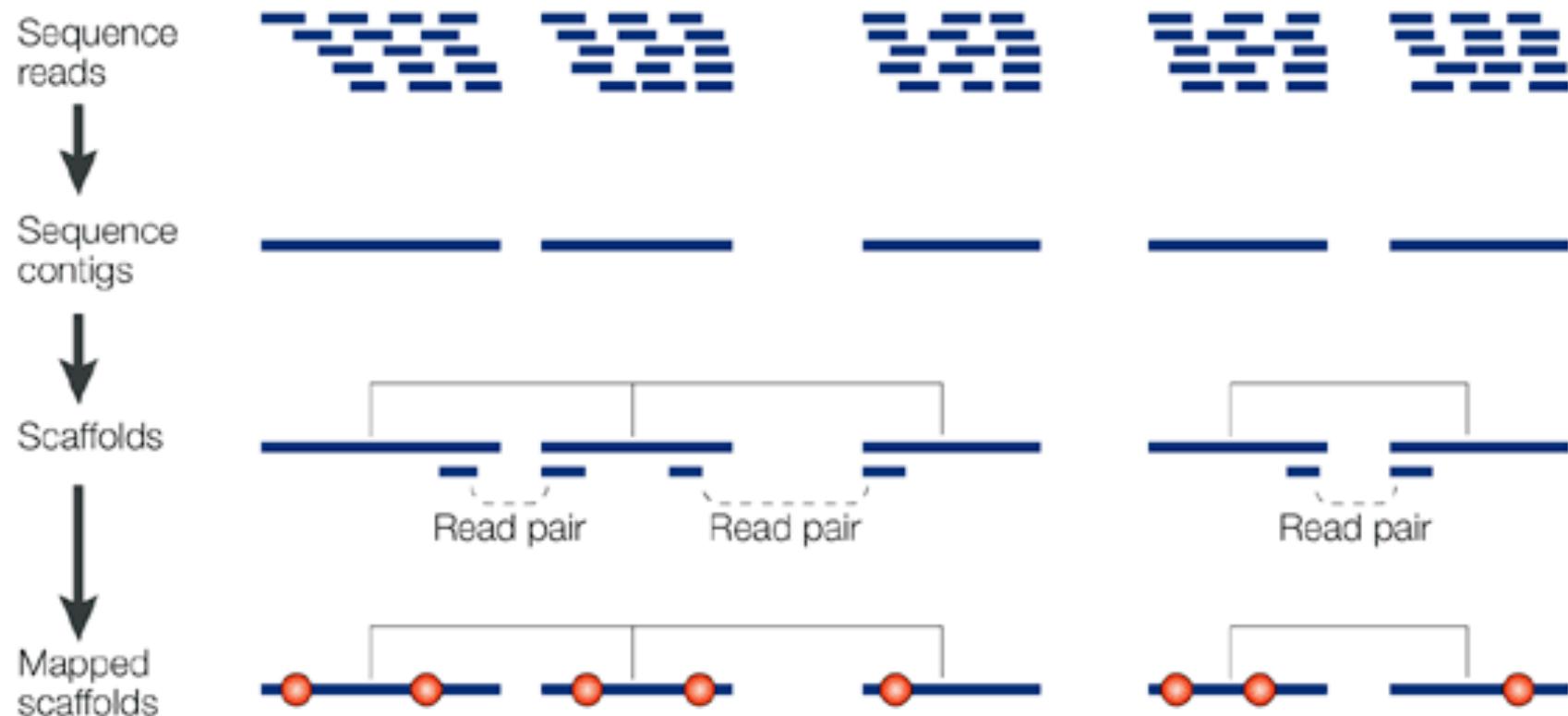
- Finishing eukaryotic genome assemblies can be challenging because much of the genome is repetitive
- This repetitive DNA breaks up the assembly and obscures the order and orientation of the assembled contigs
- Even well studied model organisms can have poorly assembled regions of their genome

Finishing genomes



Adams 2008

Finishing genomes



Adams 2008

Finishing genomes

1. Long read sequencing (e.g. PacBio)
2. Optical mapping

[https://www.youtube.com/watch?
v=jQ5BeOjbWk](https://www.youtube.com/watch?v=jQ5BeOjbWk)

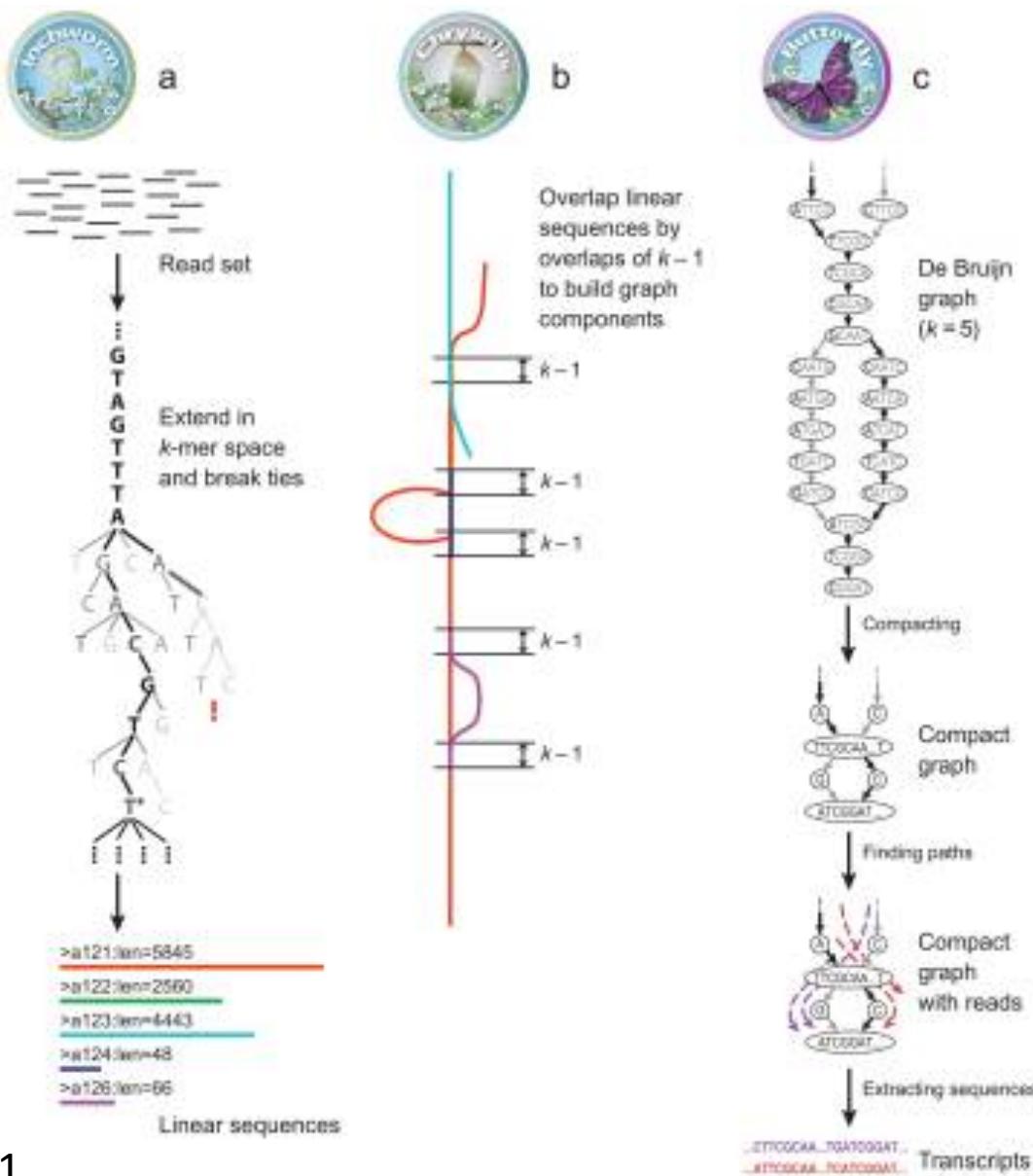
3. Genetic and physical maps
4. Jumping libraries

Other types of de novo assembly

Transcriptome

- Variable coverage among genes/isoforms
- Alternative splicing
promotors, exons, and poly(A)

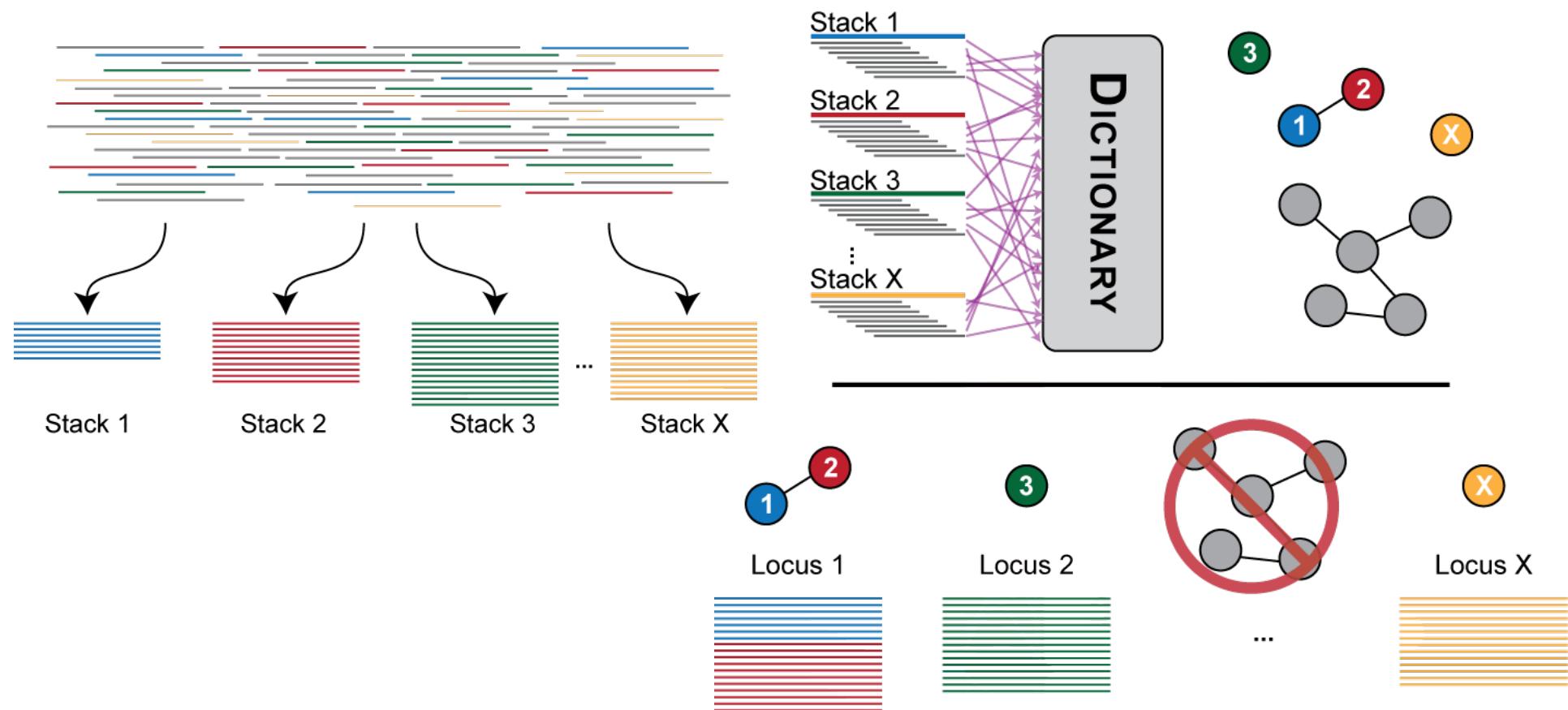
Other types of de novo assembly



Grabherr et al 2011

Other types of de novo assembly

De novo assembly of GBS reads: Stacks



Further Reading

Flicek, P., & Birney, E. (2009). Sense from sequence reads: methods for alignment and assembly. *Nature methods*, 6, S6-S12.

Zerbino DR, Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*. 2008;18(5):821-829. doi:10.1101/gr.074492.107.

<http://computing.bio.cam.ac.uk/local/doc/velvet.pdf>

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., ... & Yang, B. (2012). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Briefings in functional genomics*, 11(1), 25-37.

Grabherr MG, Haas BJ, Yassour M, et al. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*. 2011;29(7):644-652. doi:10.1038/nbt.1883.

<https://github.com/trinityrnaseq/trinityrnaseq/wiki>

J. Catchen, A. Amores, P. Hohenlohe, W. Cresko, and J. Postlethwait. Stacks: building and genotyping loci de novo from short-read sequences. *G3: Genes, Genomes, Genetics*, 1:171-182, 2011.

<http://catchenlab.life.illinois.edu/stacks/>

Tutorial

Today you will use the genome assembly program Velvet to assemble a bacterial genome.

Velvet overview:

1. Hash k-mers
2. Construct the graph
3. Correct for errors
4. Resolve the repeats

Open ~/data/Topic5/README_assembly.txt and follow the instructions

Tutorial

- 1) Given the above information, what is the expected coverage?
- 2) For a k-mer of 21 what would the k-mer coverage be for this genome assembly?
- 3) Can you think of other ways to assess assembly quality? What might be the trouble with only focusing on maximizing N50? Discuss this with your group.
- 4) Quantify the assembly metrics for your first assembly that you ran without any options. In your group of four, each person should pick different sets of parameters to run. Compare the resulting assemblies with one another and discuss which ones seemed to have improved the assembly and why that might be. Be prepared to share your findings with the class.