Day 1
Genome Assembly

Arun Gonzales Decano
Senior Bioinformatician
NDM Experimental Medicine
University of Oxford, UK

GENOME ASSEMBLY

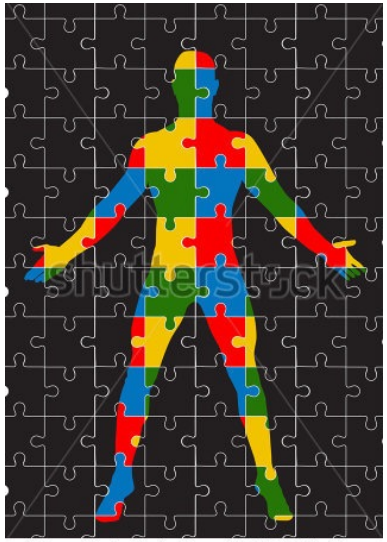
1. What is Genome Assembly?
2. Reference-guided and *De novo* Genome Assembly
3. Assembly Algorithms
4. Short and Long Read Assembly
5. SR Assembly with *Velvet, Unicycler (Spades) and Shovill Pipeline*
6. Assembly Improvement
7. Assembly QC using Quast

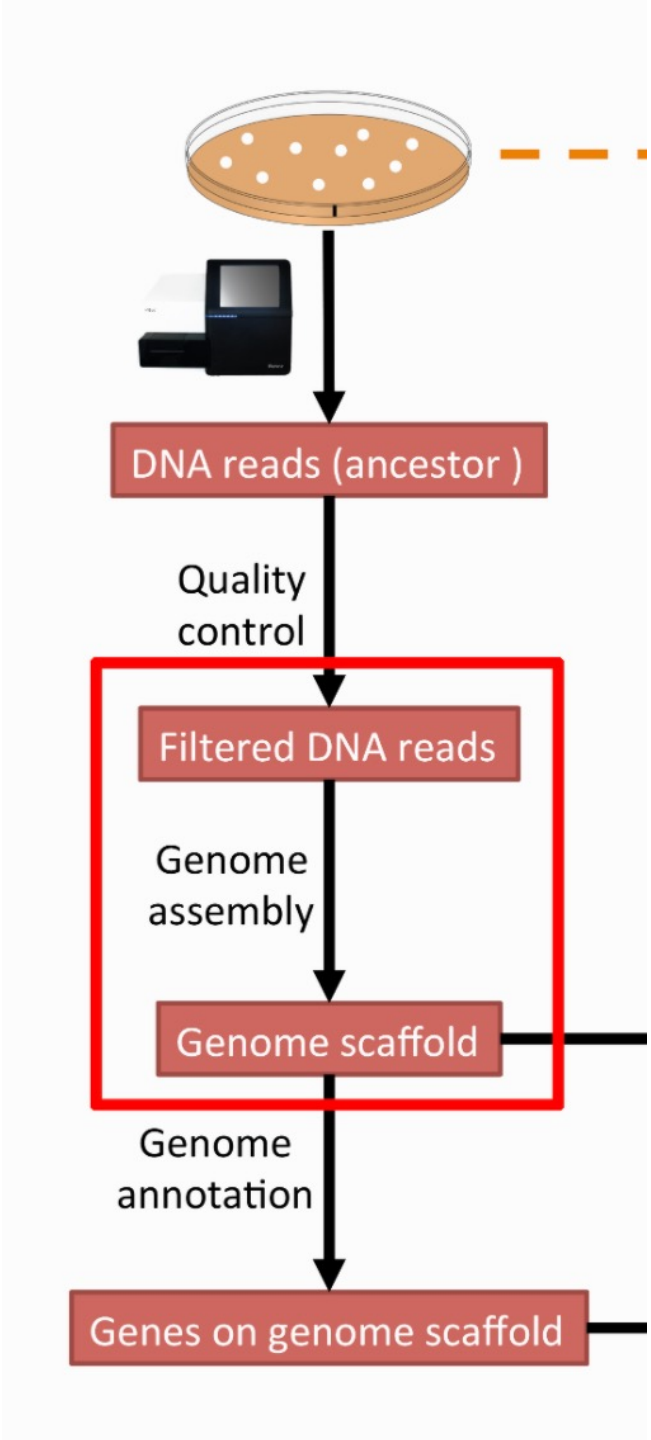
Bonus: Visualizing Genome Assemblies

What is Genome Assembly?

Puzzle: read library

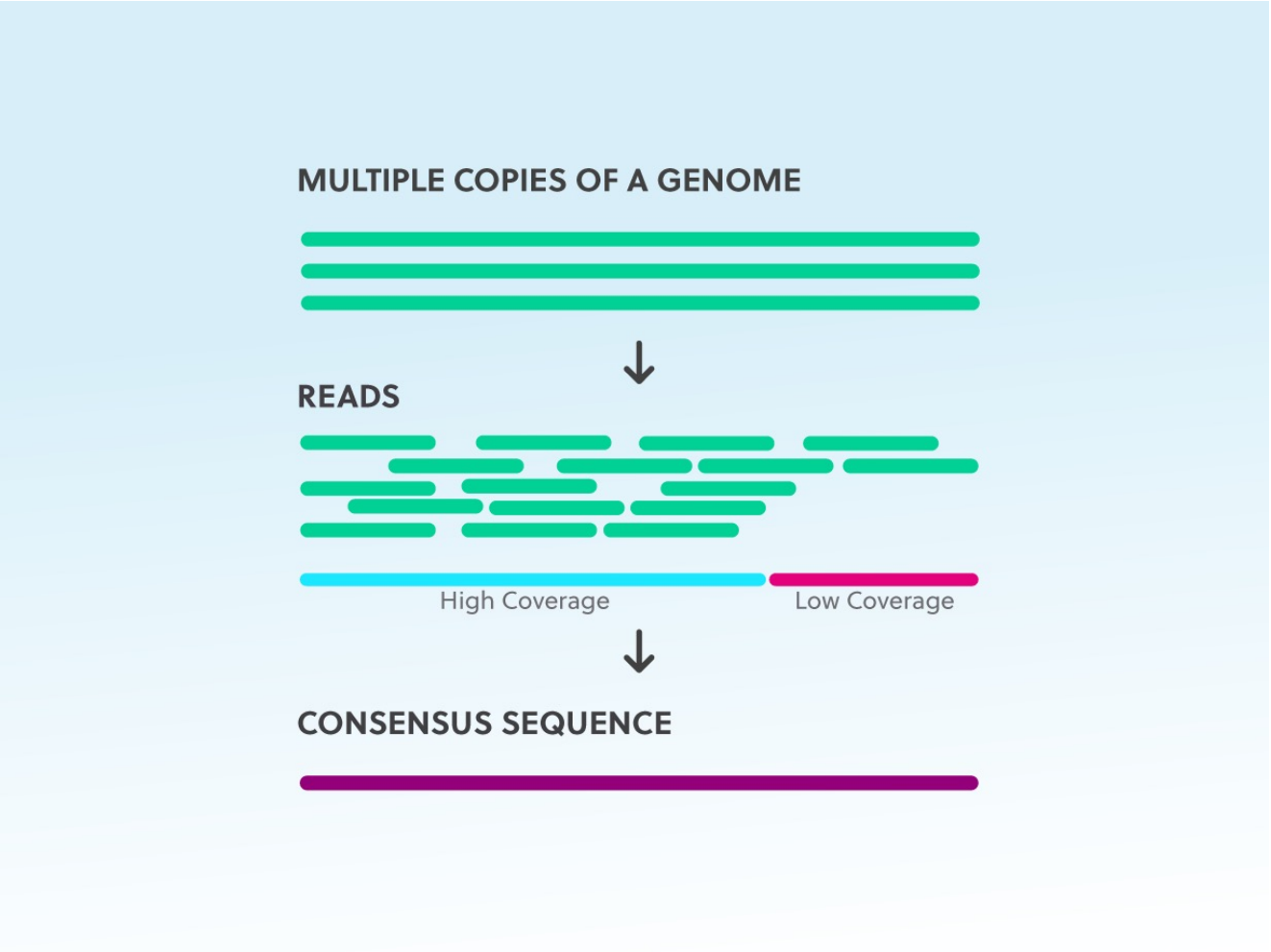
Genome Assembly = joining up overlapping regions (reads) into a continuous sequence known as a 'contig'





https://genomics.sschmeier.com/index.html

Genome assembly = creating min set of consensus sequences



irepertoire.com

Coverage/Depth (x)

Consensus Sequence

ATTGTCGTAAGTACAGTAGACGATAGCAGTTGACGATTGAGCCCCCATGCTAT

ATTGTCGTAAGTACAGTAGA TAGCAGTTGACGATTGAGCC

TTGTCGTAAGTATAGTAGAC ATAGCAGTTGACGATTGAGC

TGTCGTAAGTACAGTAGACG ATAGCAGTTGACGATTGAGC

GTCGTAAGTATAGAAGACGA TAGCAGTTGACGATTGAG

TCGTAAGTATAGTAGACGAT TTGACGATTGAGCCCCCATG

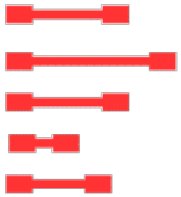
ACAGTAGACGATAGCAGTTG CGATTGAGCCCCCATGCTAT

TTGTCGTAAGTATAGTAGAC ATAGCAGTTGACGATTGAGC

Coverage / Depth (x)

How do we compute for Coverage/Depth (x)?

(A)

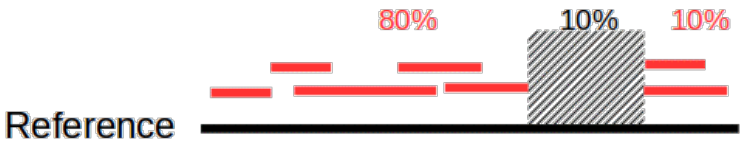


Reads

Reference

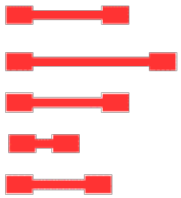
$$C = \frac{\#\ \text{sequenced bases}^{1}\ (= \#\ \text{bases of all mapped reads})}{\#\ \text{bases of reference}}$$

(B)



$$C = \frac{\#\ \text{area covered by reads}}{\#\ \text{reference area}}$$

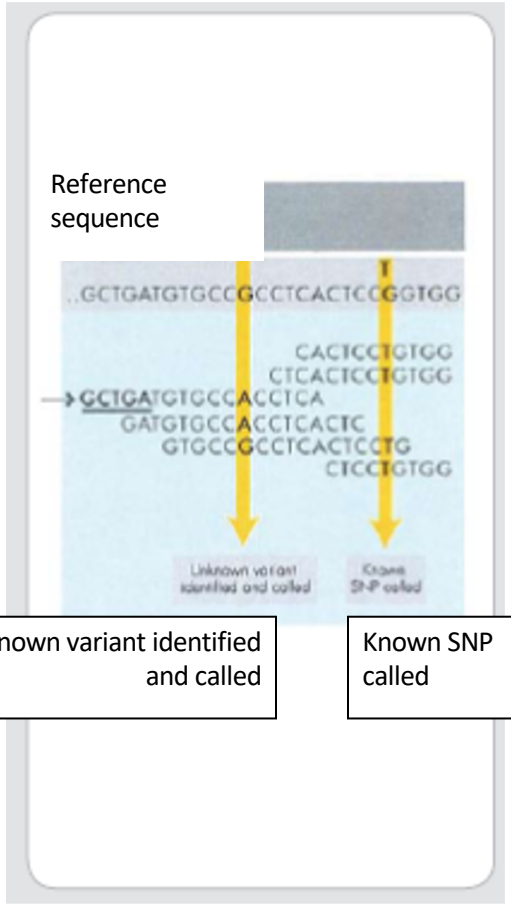
(C)



Reads for mapping

Sequencing depth= total read number

Reference-guided Genome Assembly



7. Determine first base

8. Image first base

9. Determine second base

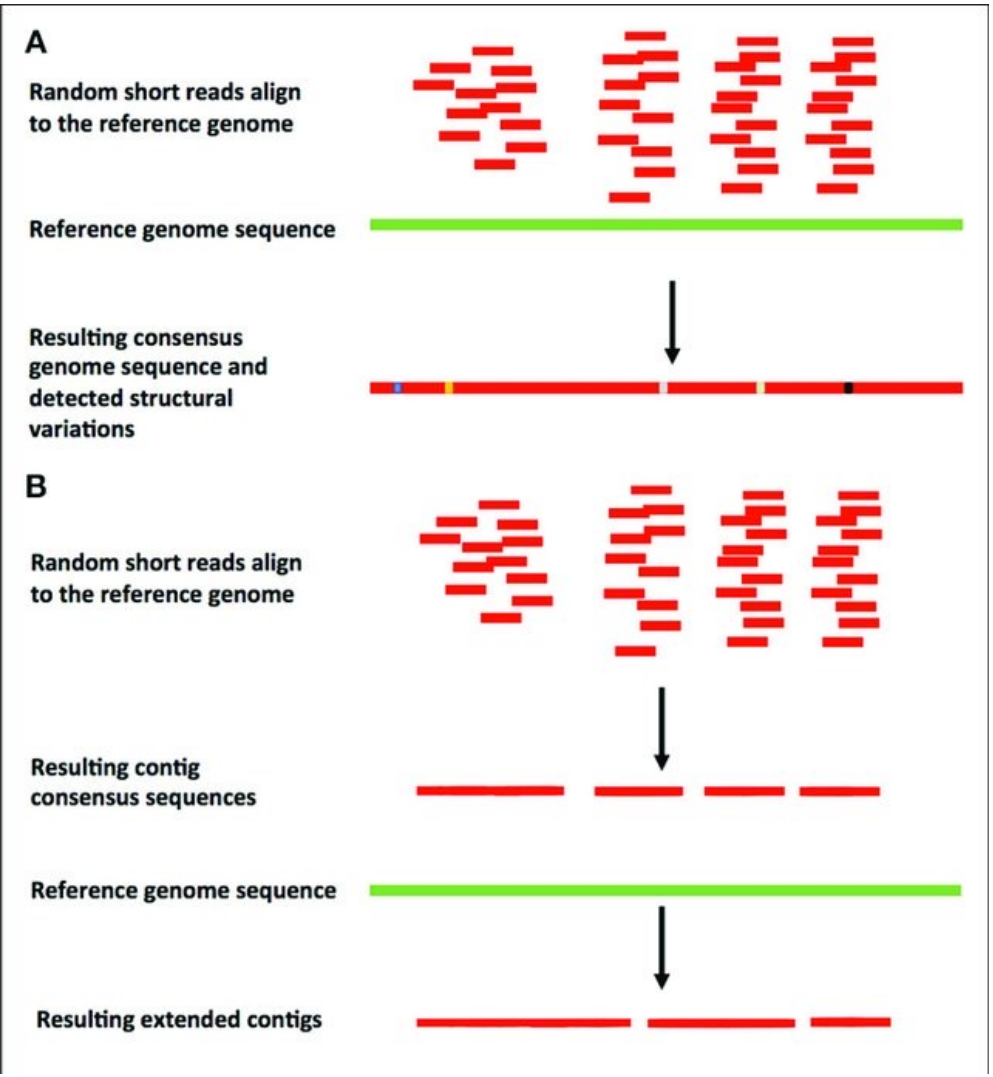
10. Image second chemistry cycle

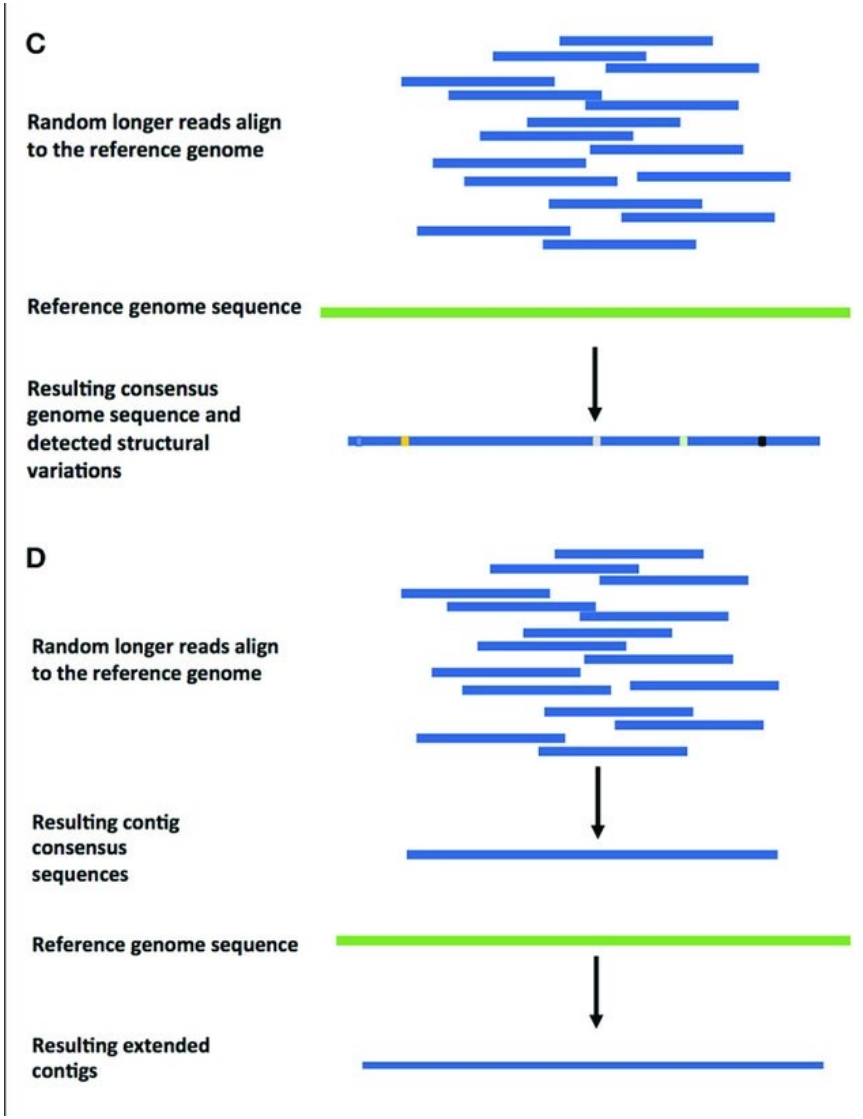
11. Sequencing over multiple chemistry cycles

12. Align data

The data are aligned and compared to a reference and sequencing differences are identified.

Reference-guided Genome Assembly





Kyriakidou et al. (2018)

Where to download reference microbial sequences from

- NCBI
  - Raw sequencing reads: https://www.ncbi.nlm.nih.gov/sra
  - Assemblies (Nucleotide): https://www.ncbi.nlm.nih.gov/nuccore
- Ensembl: http://www.ensembl.org
  - Wide range of annotated genomes
  - Non-vertebrates: http://ensemblgenomes.org
- ENA: http://www.ebi.ac.uk/ena

Where to download reference microbial sequences from

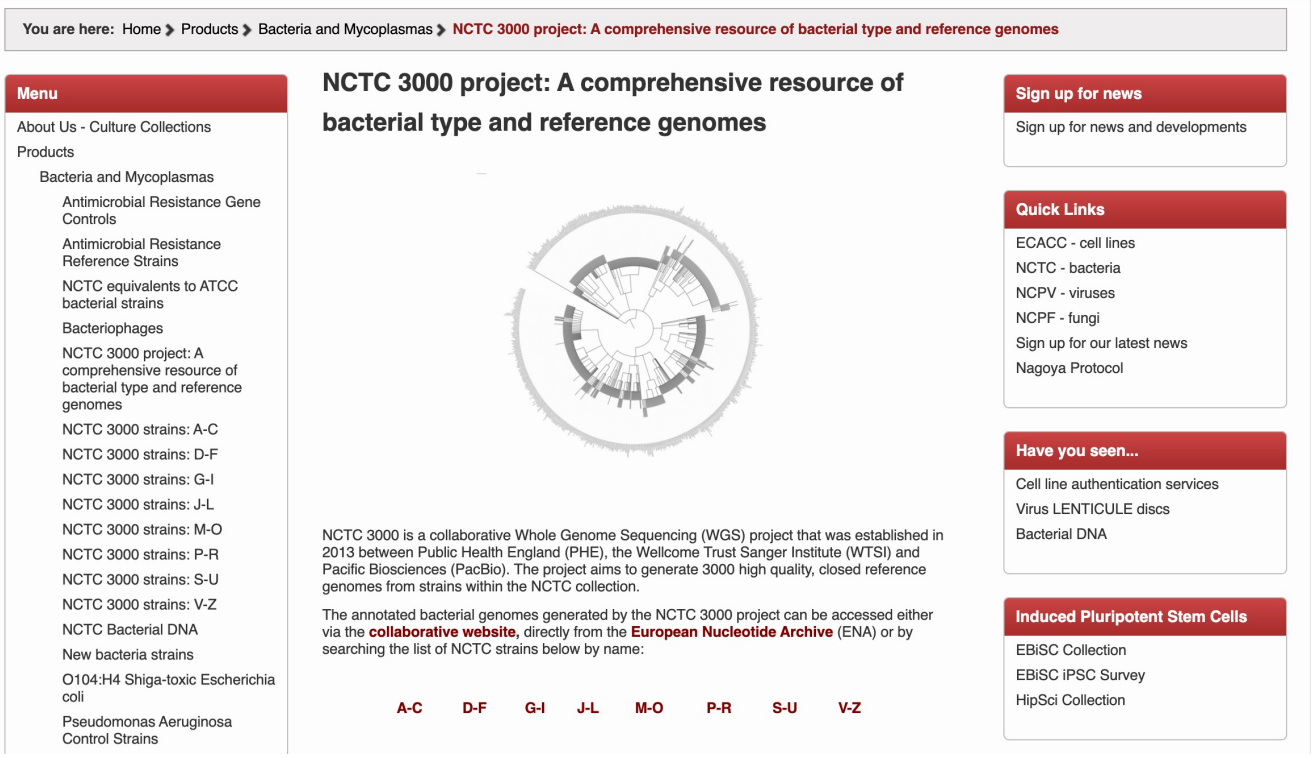
NCTC 3000 Project

What is the NCTC 3000 project?

- it is a five year joint project to be delivered by Public Health England's (PHE) Culture Collections, the **Wellcome Trust Sanger Institute** and Pacific Biosciences
- it aims to produce a unique website for the clinical and research microbiology community

https://www.phe-culturecollections.org.uk/collections/nctc-3000-project.aspx
https://www.phe-culturecollections.org.uk/products/bacteria/index.jsp

Where to download reference microbial sequences from



https://www.phe-culturecollections.org.uk/products/bacteria/nctc-3000-project-a-comprehensive-resource-of-bacterial-type-and-reference-genomes.aspx
https://www.phe-culturecollections.org.uk/products/bacteria/browse.jsp

De novo Genome Assembly

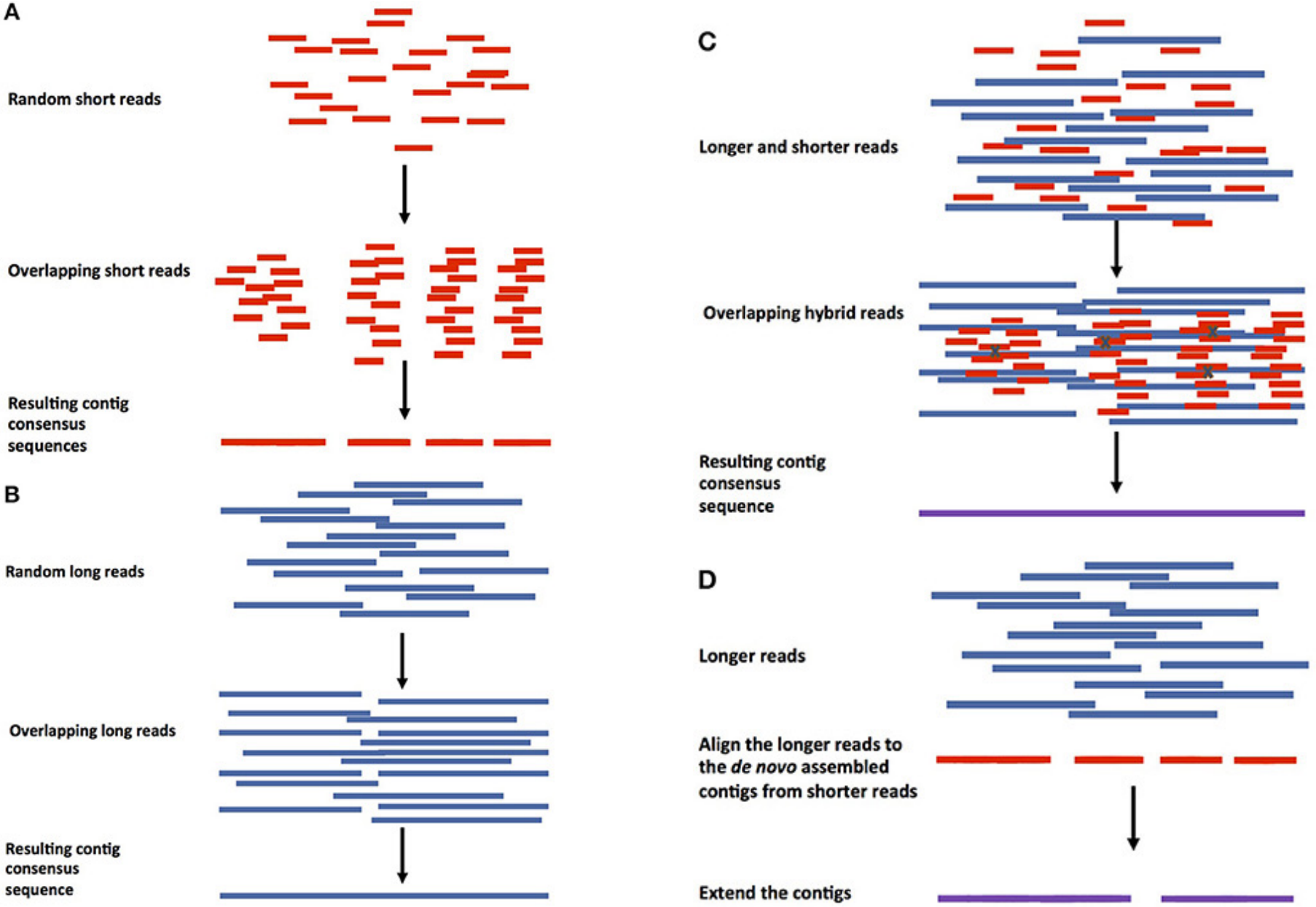
Why do we de novo *assemble?*

To discover highly variable or divergent loci eg HGT in microbes

Our reference genome(s) may be incomplete

More accurate variant finding uses local information

De novo Genome Assembly



Kyriakidou et al. (2018)

Assembly issues

Repeats

Low coverage (GI-GO)

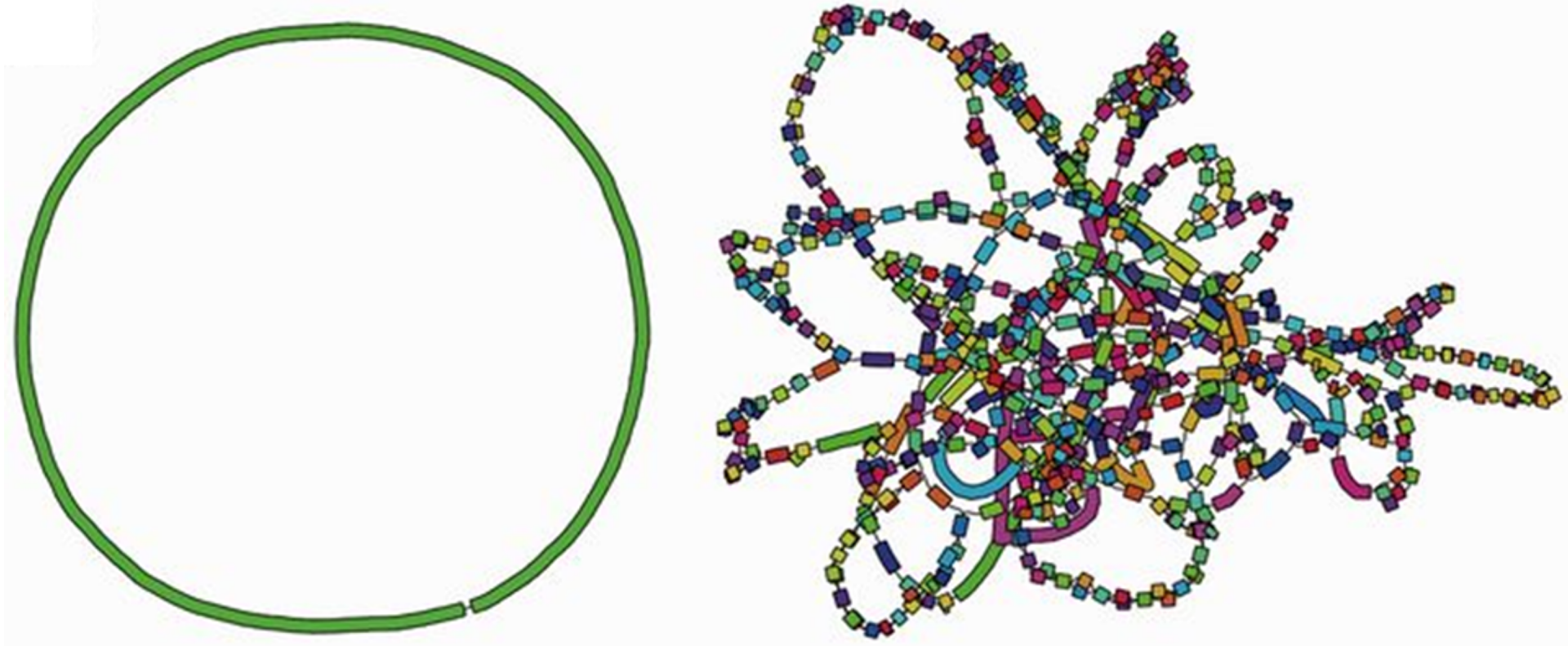
PCR amplification bias in preparation

Sequencing adapters

Contamination

Wick et al "Bandage: interactive visualization of *de novo* genome assemblies"

left = ideal bacterial assembly = single contig
right = poor assembly with many short contigs



Methodology Article | Open Access | Published: 10 November 2017

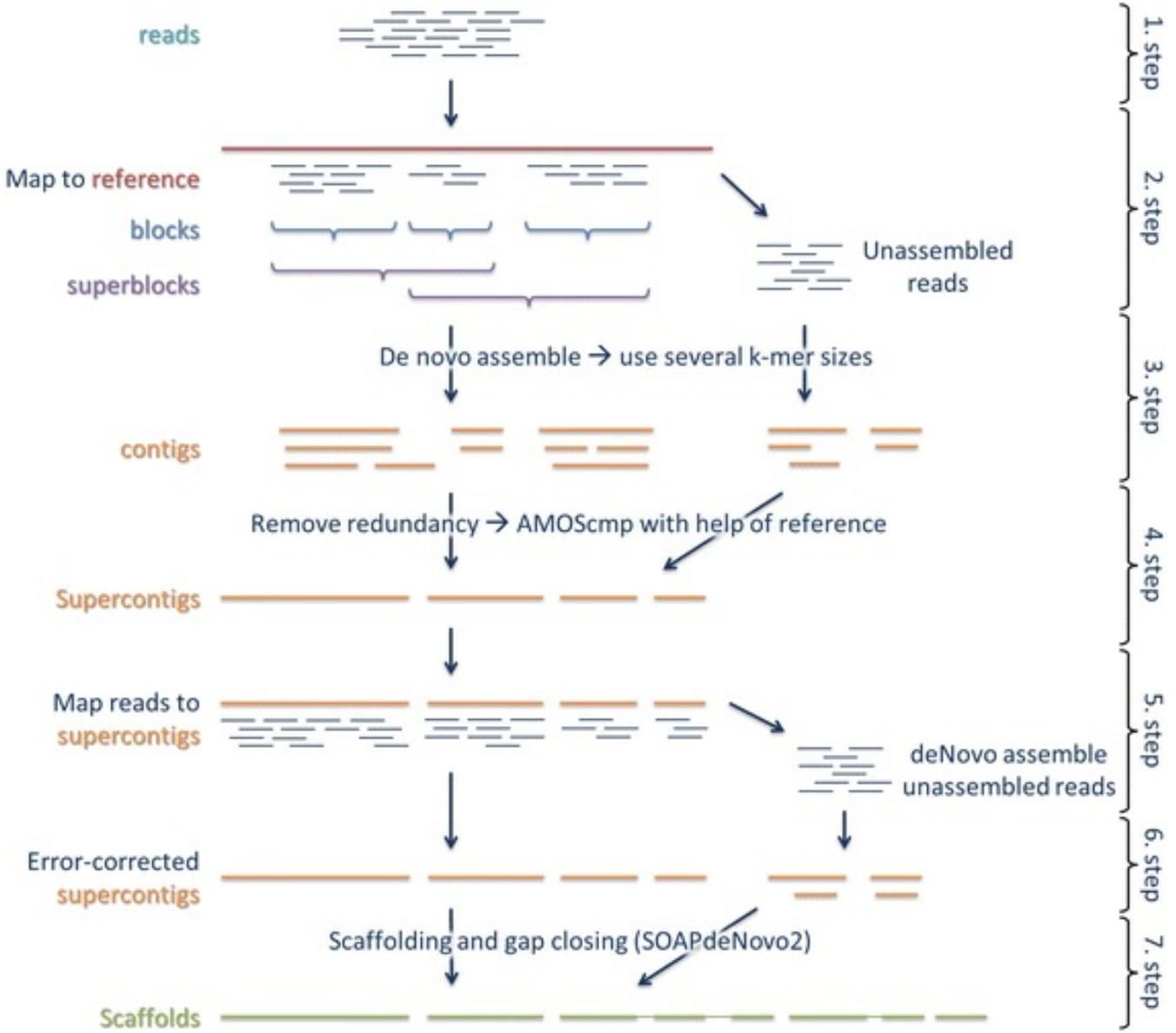
Reference-guided *de novo* assembly approach improves genome reconstruction for related species

Heidi E. L. Lischer & Kentaro K. Shimizu

BMC Bioinformatics **18**, Article number: 474 (2017) | Cite this article

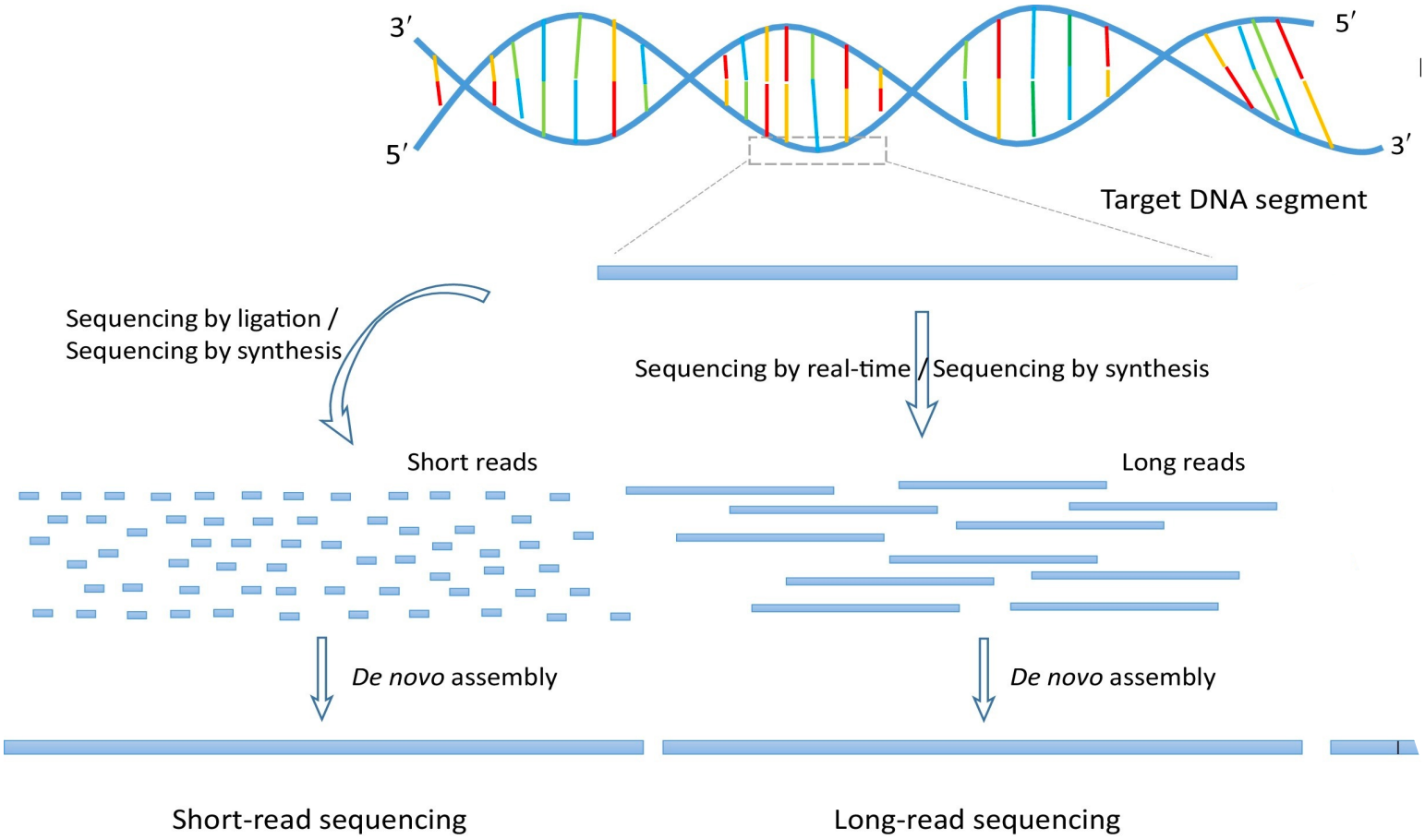
Example: ALLPATHS-LG

Reference-guided *De novo* Genome Assembly

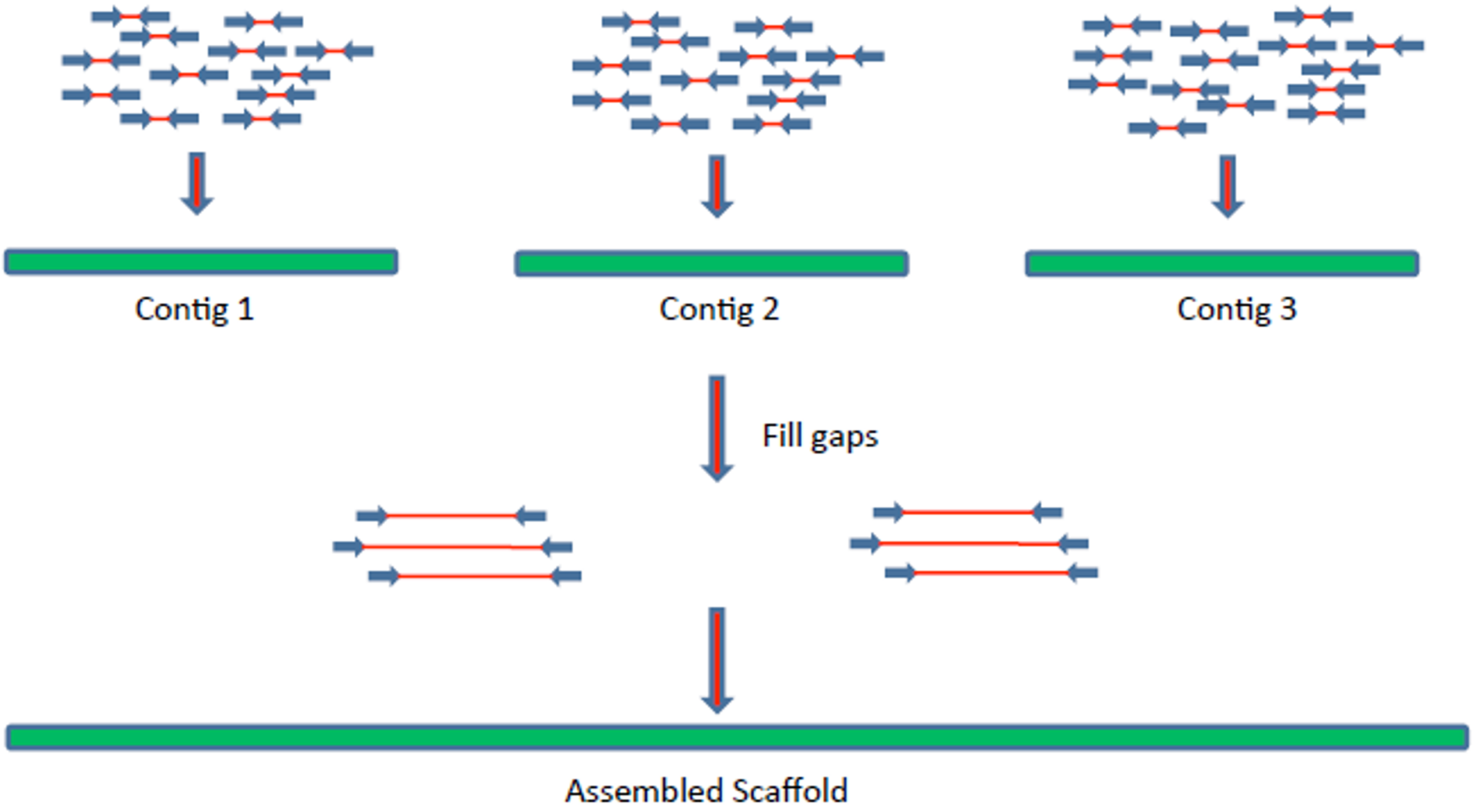


https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-017-1911-6

Assembly Process



Assembly process: Short/PE reads



Assembly Algorithms

Assembly by greedy algorithms

Traditional methods for small (capilliary) data

Makes locally optimal choices (eg join pairs of reads with best overlap)

Examples: phrap, TIGR Assembler (old)

Assembly by overlap-layout-consensus

OLC = overlap-layout-consensus (eg Celera)

Nodes = reads or contigs
Overlap = edges joining nodes

Assembly by overlap-layout-consensus

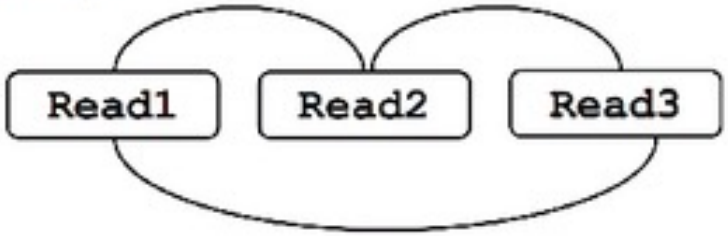
(i) **OVERLAP**: Local/global alignment of overlap b/w seqs computed by seed-and-extend approach

(ii) **LAYOUT**: Alignment scores = distances on graph or tree => merge and make initial sets of contigs

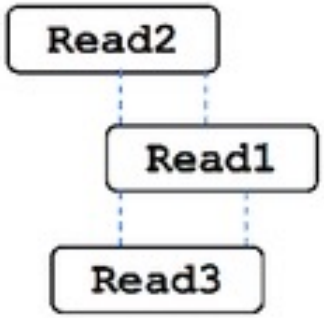
(iii) **CONSENSUS**: contigs iteratively aligned to minimise redundancy

(a) Overlap, Layout, Consensus assembly

(i) Find overlaps



(ii) Layout reads



(iii) Build consensus

```
CGATTCTA
    TTCTAAGT
 GATTGTAA
-----------
CGATTCTAAGT
```

Ayling M, Clark MD, Leggett RM. New approaches for metagenome assembly with short reads. Brief Bioinform. 2020 Mar 23;21(2):584-594. doi: 10.1093/bib/bbz020. PMID: 30815668; PMCID: PMC7299287.

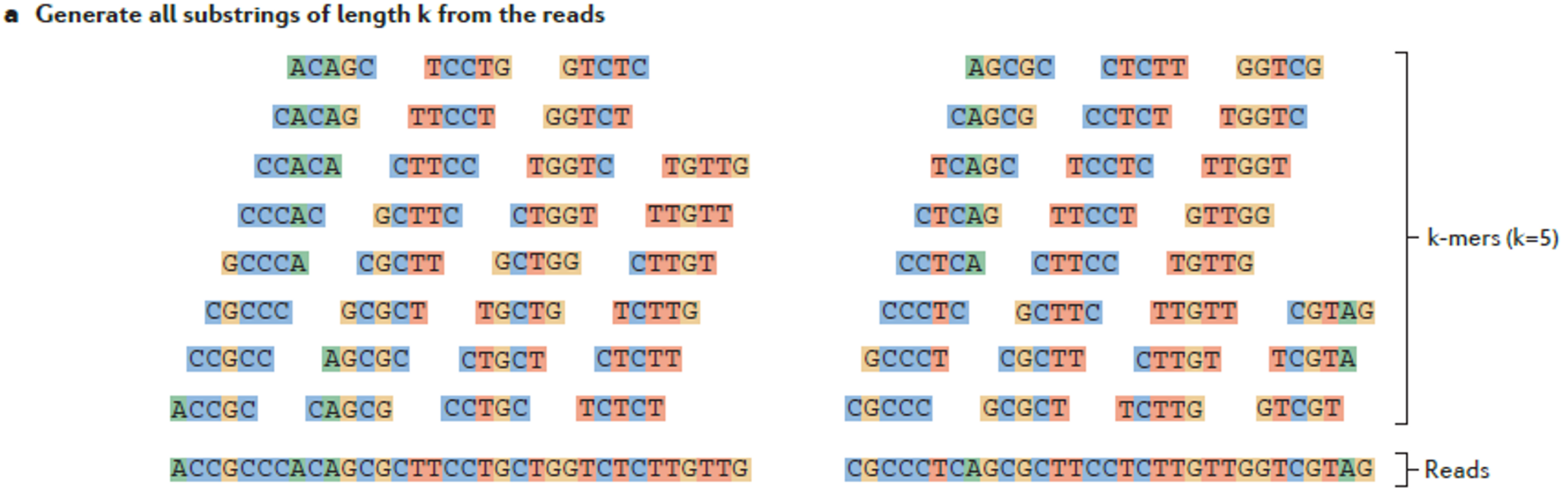
Assembly by de Bruijn graphs

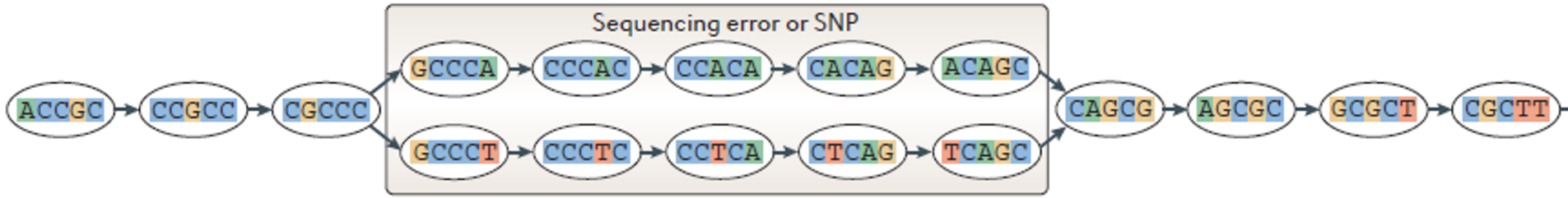
"Hashing" speeds up search tasks by splitting genome into seqs of length k ("k-mers")

De Bruijn graphs align k-mers (faster) not reads

As k->bigger/higher, seqs more unique, yielding more precise mapping but longer compute time

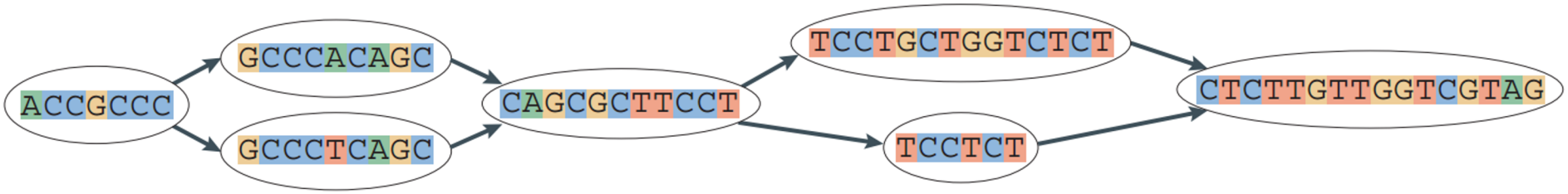
Assembly by de Bruijn graphs





Convert reads to 5-mers and graph in 5-base overlapping blocks

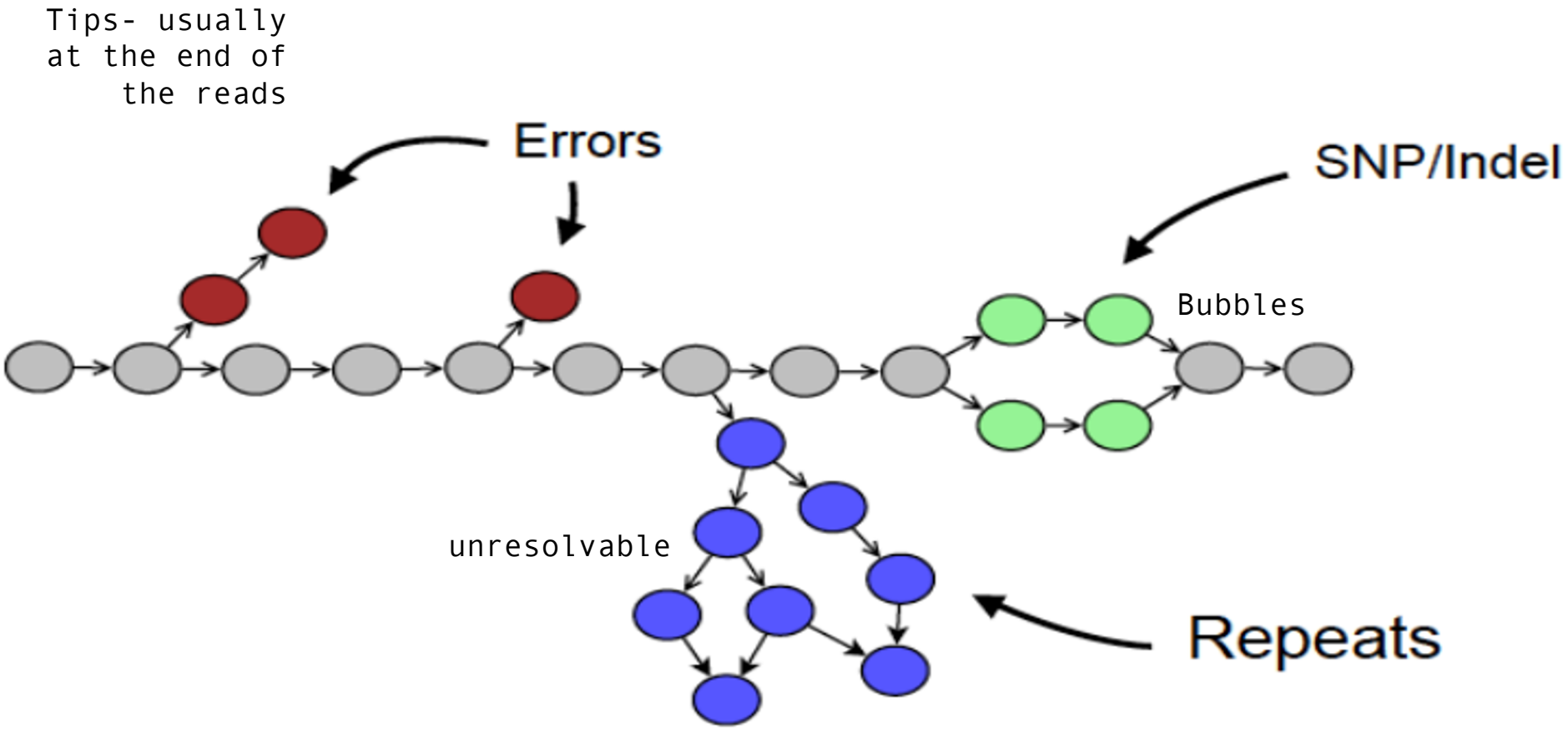
Assembly by de Bruijn graphs



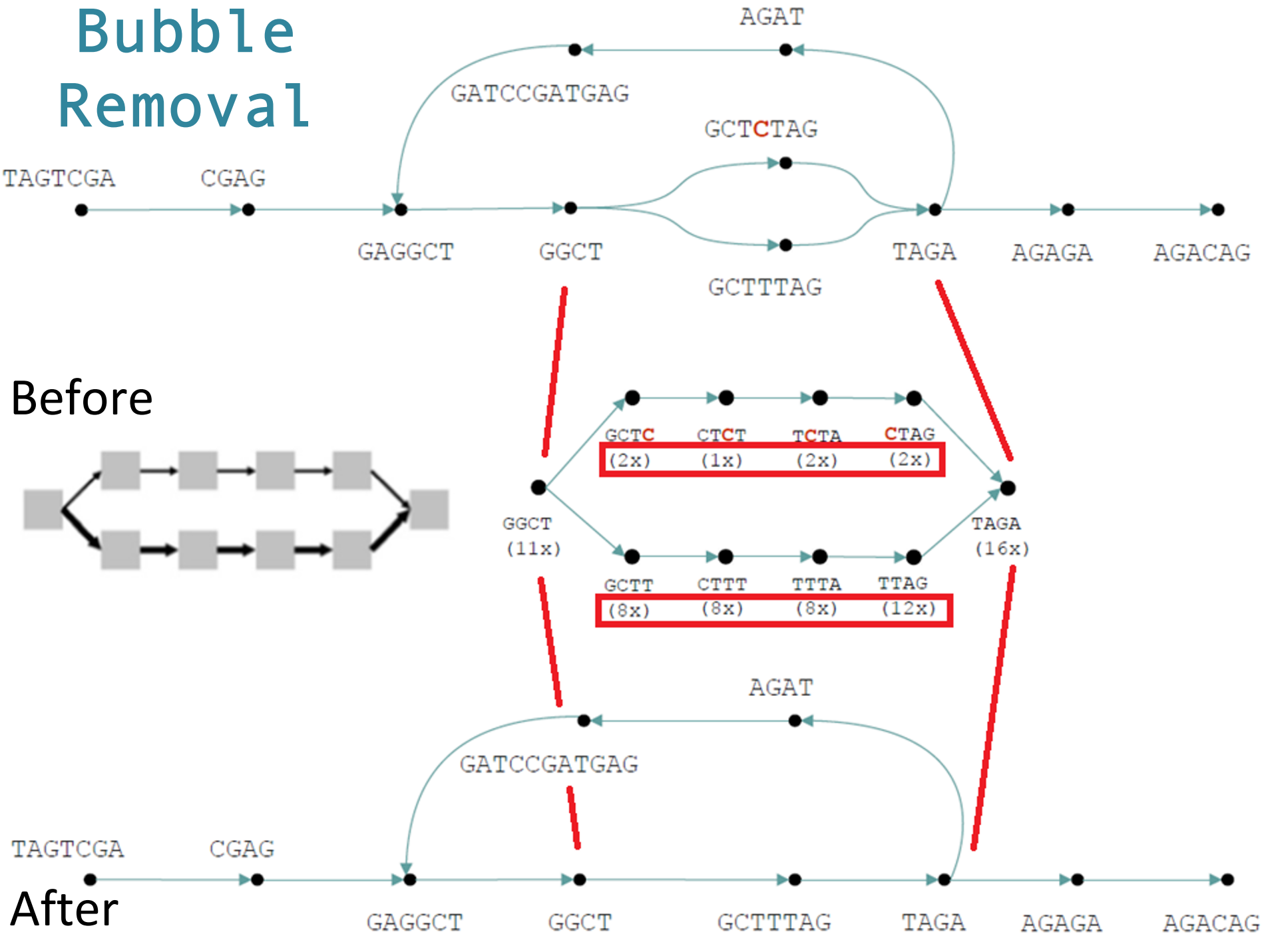
Collapse 5-base overlapping blocks into consensus seqs

These are mini-contigs

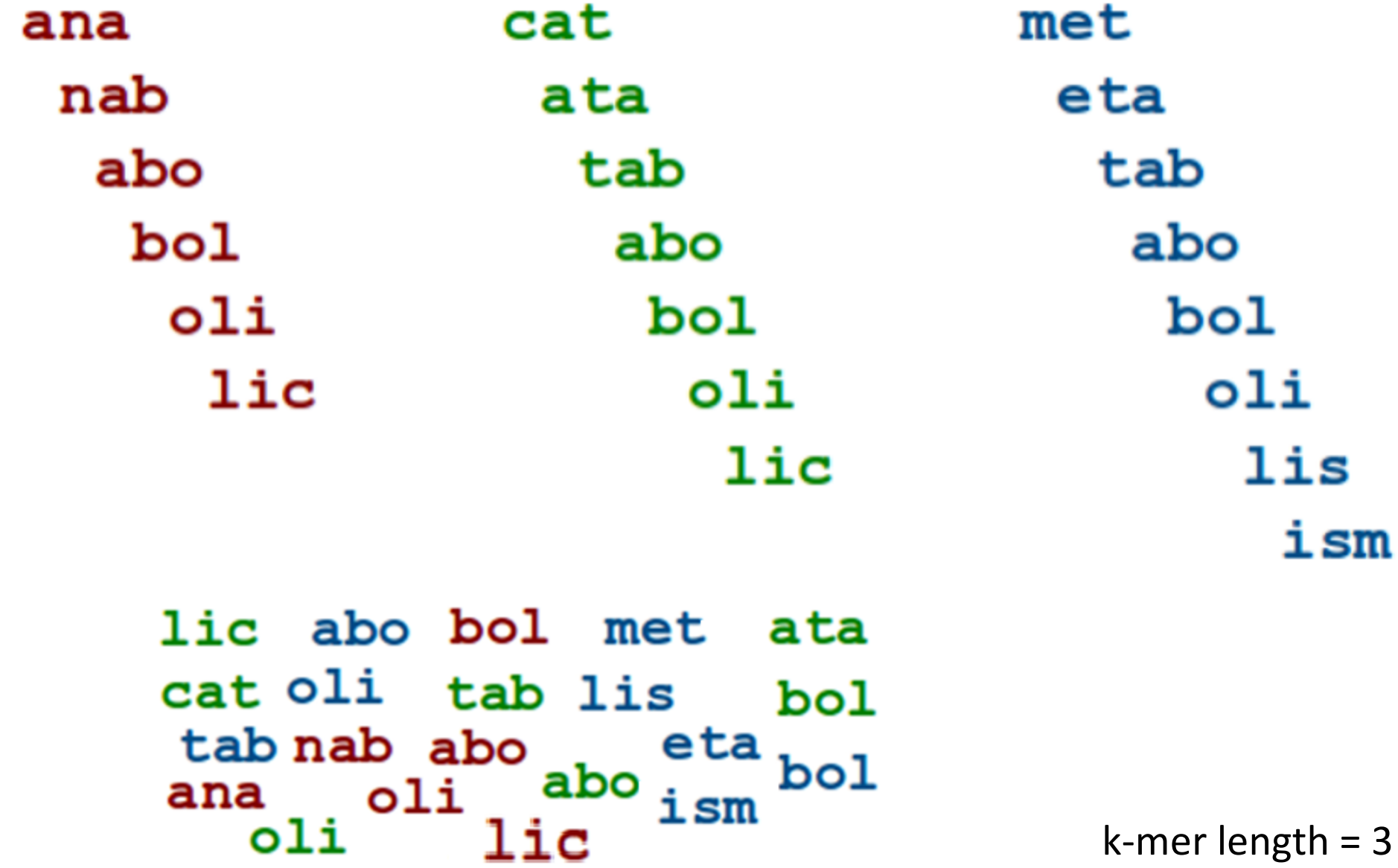
Graph Artefacts



Bubble Removal



Make your own de Bruijn graphs!



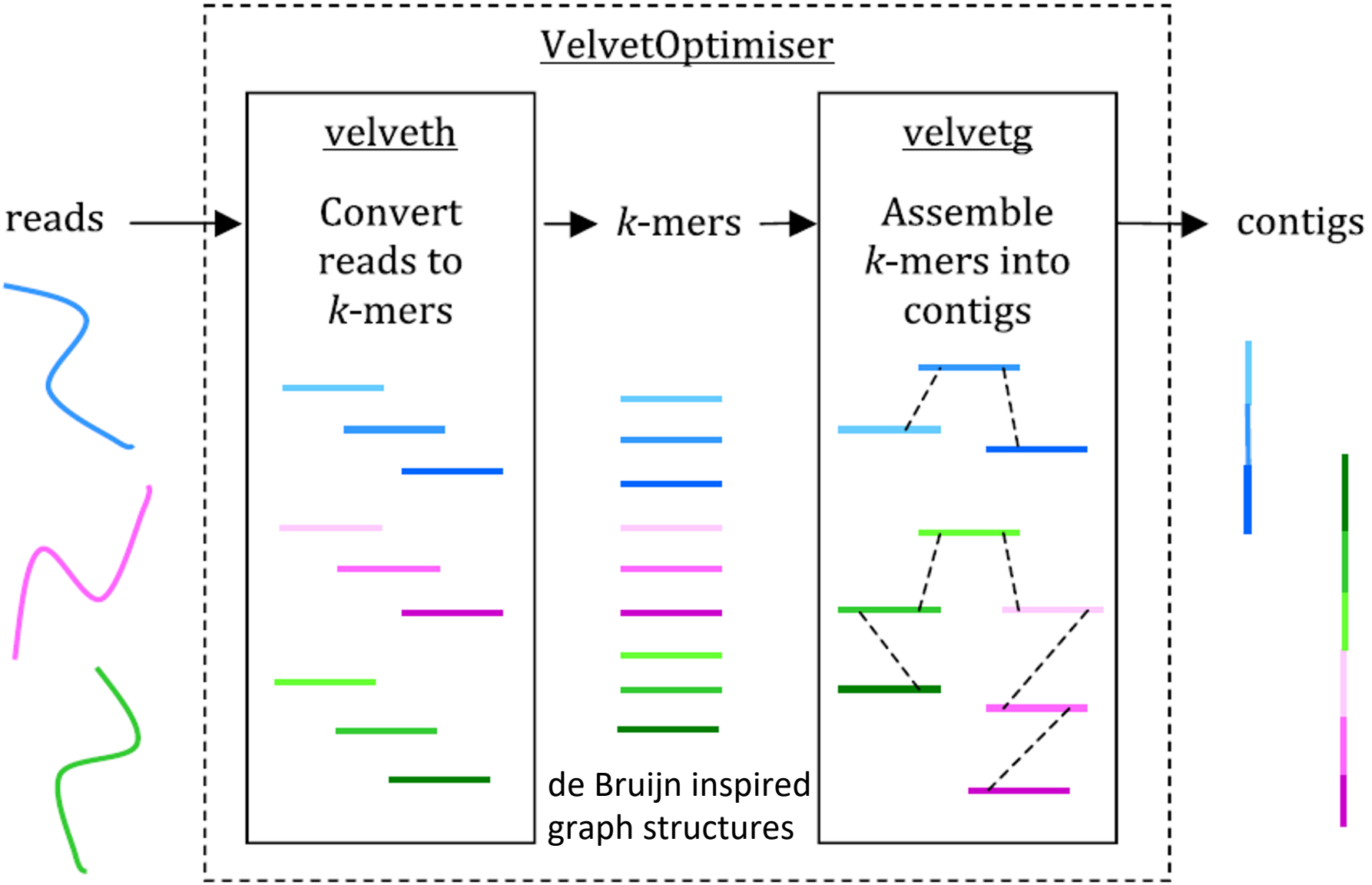
Assembly by string graphs

de Bruijn align k-mers with computational speed ~ N^2 (N = sum all read lengths)

R_1 ACATACGATACA
R_2 TACGATACAGTT
R_3 GATACAGTTGCA

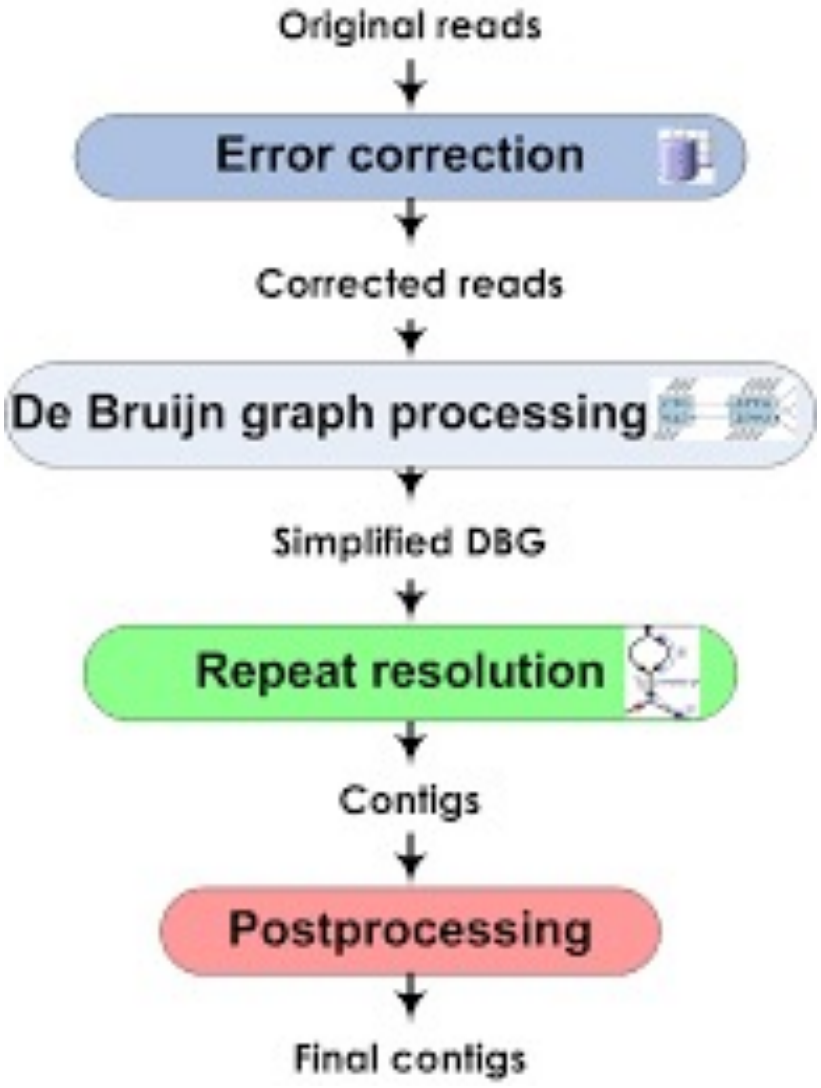
Each overlap sequence is represented once
"suffix arrays" eg by Simpson and Durbin

Assembly with Velvet



D. Zerbino (2010): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2952100/

SPAdes - St. Petersburg genome assembler - is an assembly toolkit containing various assembly pipelines

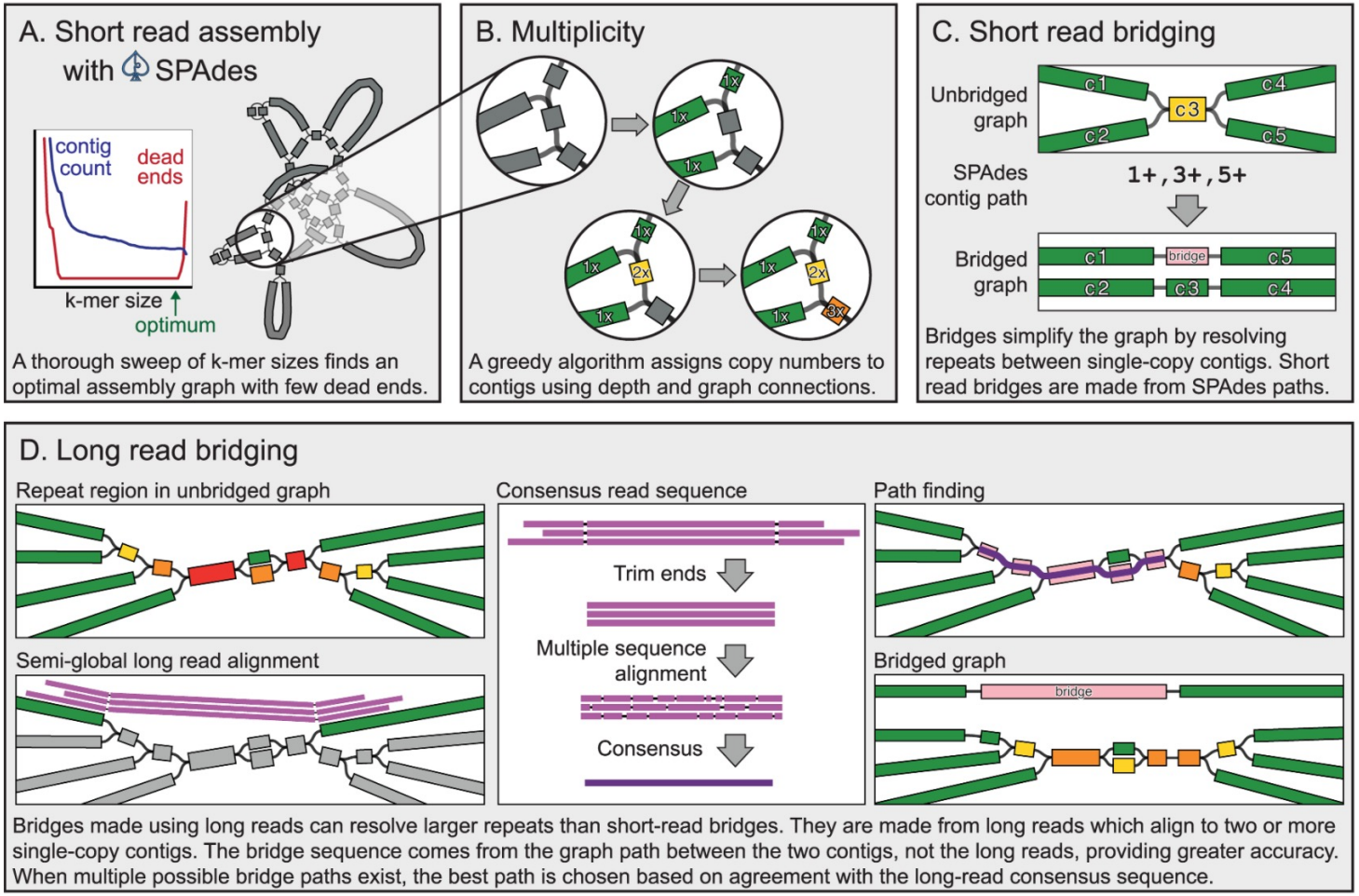


Bankevich et. al. (2012): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3342519/; https://logic.pdmi.ras.ru/~sergey/slides/N12_SPAdesRECOMB.pdf; https://github.com/ablab/spades

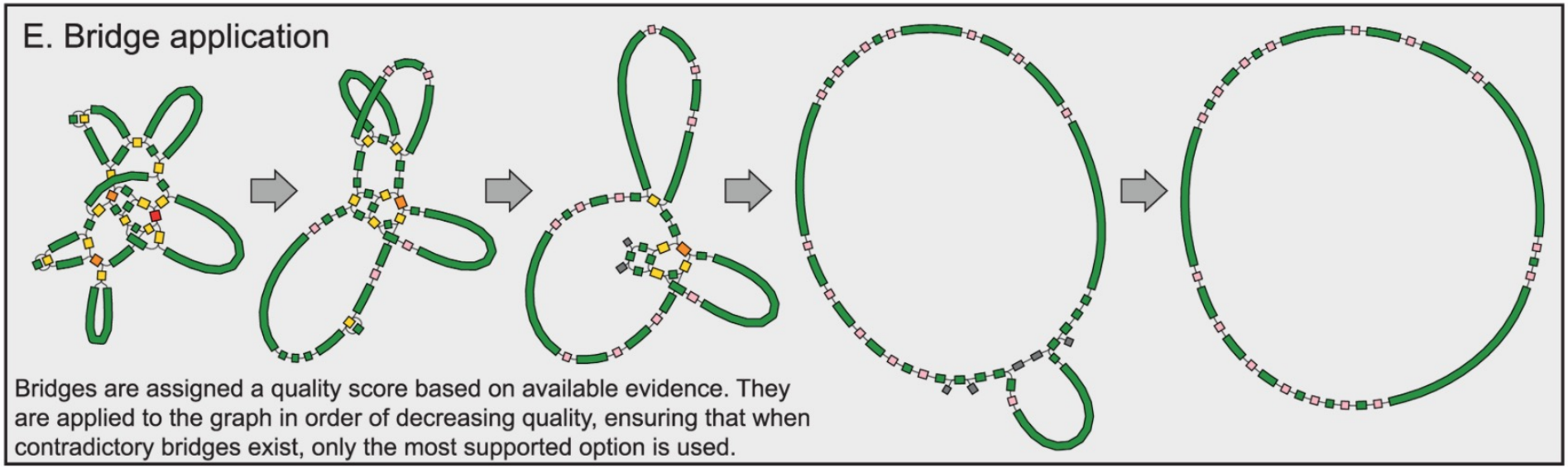
Assembly using Unicycler

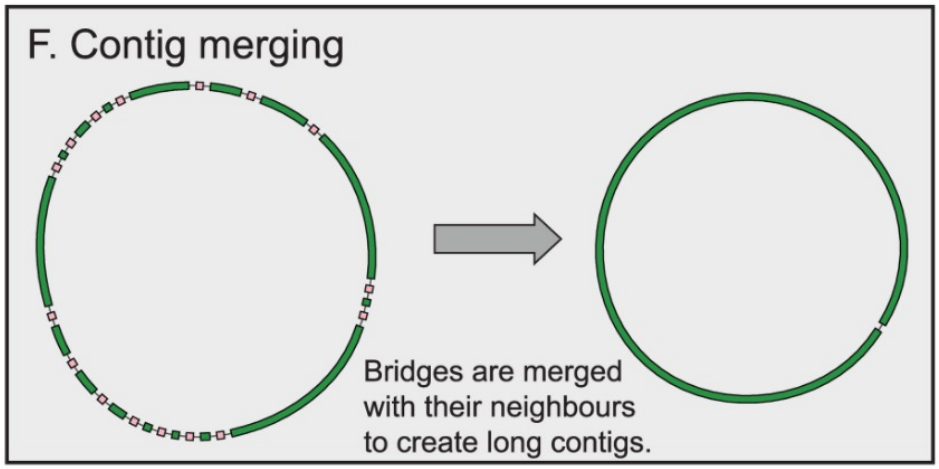
Github: https://github.com/rrwick/Unicycler

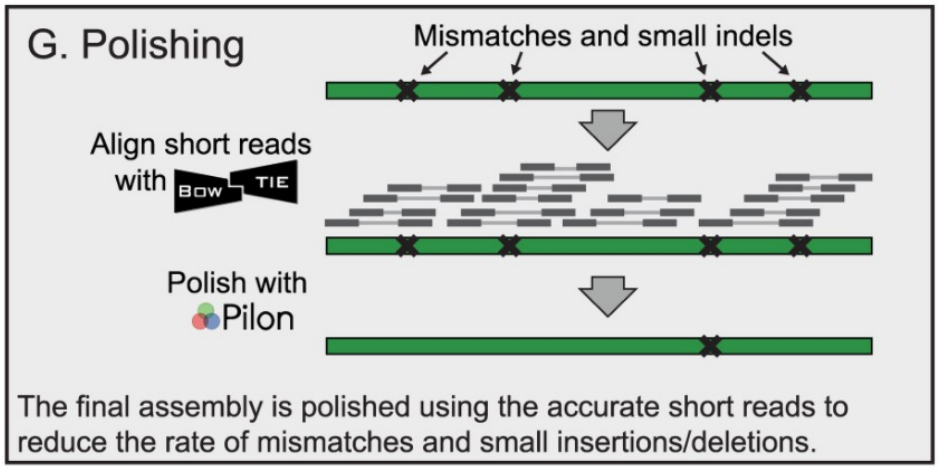
Simplified view of the Unicycler assembly process (From Wick *et al.* 2017).In short, Unicycler uses SPAdes (see below) to produce an assembly graph, which is then bridged (simplified) using long reads to produce the longest possible set of contigs.



Wick *et al.* 2017







These are then polished by aligning the original short reads against contigs and feeding these alignments to Pilon - an assembly improvement tool.

Wick *et al.* 2017

Shovill

Assemble bacterial isolate genomes from Illumina paired-end reads

Introduction

The SPAdes genome assembler has become the *de facto* standard *de novo* genome assembler for Illumina whole genome sequencing data of bacteria and other small microbes. SPAdes was a major improvement over previous assemblers like Velvet, but some of its components can be slow and it traditionally did not handle overlapping paired-end reads well.

Shovill is a pipeline which uses SPAdes at its core, but alters the steps before and after the primary assembly step to get similar results in less time. Shovill also supports other assemblers like SKESA, Velvet and Megahit, so you can take advantage of the pre- and post-processing the Shovill provides with those too.

⚠️ Shovill is for isolate data only, primarily small haploid organisms. It will *NOT* work on metagenomes or larger genomes. Please use Megahit directly instead.

Main steps

1. Estimate genome size and read length from reads (unless `--gsize` provided)
2. Reduce FASTQ files to a sensible depth (default `--depth 100`)
3. Trim adapters from reads (with `--trim` only)
4. Conservatively correct sequencing errors in reads
5. Pre-overlap ("stitch") paired-end reads
6. Assemble with SPAdes/SKESA/Megahit with modified kmer range and PE + long SE reads
7. Correct minor assembly errors by mapping reads back to contigs
8. Remove contigs that are too short, too low coverage, or pure homopolymers
9. Produce final FASTA with nicer names and parseable annotations

https://github.com/tseemann/shovill

Anatomy of an assembled genome

File format: “.fasta” or ”.fa” or “.fna”

Command: grep “>” BN76.fasta | more

```
>BN26_1_1_cov_65
>BN26_1_2_cov_64
>BN26_1_3_cov_61
>BN26_1_4_cov_167
>BN26_1_5_cov_63
>BN26_1_6_cov_62
>BN26_1_7_cov_63
>BN26_1_8_cov_316
>BN26_1_9_cov_509
>BN26_1_10_cov_64
--More--
```

Assembly Improvement by Polishing

Pilon is a software tool which can be used to:

- Automatically improve draft assemblies
- Find variation among strains, including large event detection

Pilon requires as input a FASTA file of the genome along with one or more BAM files of reads aligned to the input FASTA file. Pilon uses read alignment analysis to identify inconsistencies between the input genome and the evidence in the reads. It then attempts to make improvements to the input genome, including:

- Single base differences
- Small indels
- Larger indel or block substitution events
- Gap filling
- Identification of local misassemblies, including optional opening of new gaps

Pilon then outputs a FASTA file containing an improved representation of the genome from the read data and an optional VCF file detailing variation seen between the read data and the input genome.

https://github.com/broadinstitute/pilon/wiki

Assembly QC

Contig stats

- *No. of contigs*: The total number of contigs in the assembly.
- *Largest contig*: The length of the largest contig in the assembly.
- *Total length*: The total number of bases in the assembly.
- *Nx (where* $0 \leq x \leq 100$): The largest contig length, L, such that using contigs of length $\geq L$ accounts for at least x% of the bases of the assembly.
- *NGx, Genome Nx*: The contig length such that using equal or longer length contigs produces x% of the length of the reference genome, rather than x% of the assembly length.

N50: Length of contig that contains ≥50% of the genome information.
****Higher N50 value = More complete assembly****

Assembly QC

Misassemblies and structural variations

- *No. of misassembled contigs*: The number of contigs that contain misassembly breakpoints.
- *Misassembled contigs length*: The total number of bases contained in all contigs that have one or more misassemblies.
- *No. of unaligned contigs*: The number of contigs that have no alignment to the reference sequence.
- *No. of ambiguously mapped contigs*: The number of contigs that have high-scoring reference alignments of equal quality in multiple locations on the reference genome.
- In addition to these summary statistics, QUAST also generates reports with detailed information about each contig, including whether the contig is unaligned, ambiguously mapped, misassembled or correct.

Assembly QC using Quast

Quast

QUality ASsessment Tool [GUREVICH2013], evaluates genome assemblies by computing various metrics, including:

- N50: length for which the collection of all contigs of that length or longer covers at least 50% of assembly length
- NG50: where length of the reference genome is being covered
- NA50 and NGA50: where aligned blocks instead of contigs are taken
- miss-assemblies: miss-assembled and unaligned contigs or contigs bases
- genes and operons covered

Report Example

E. coli single-cell assemblies

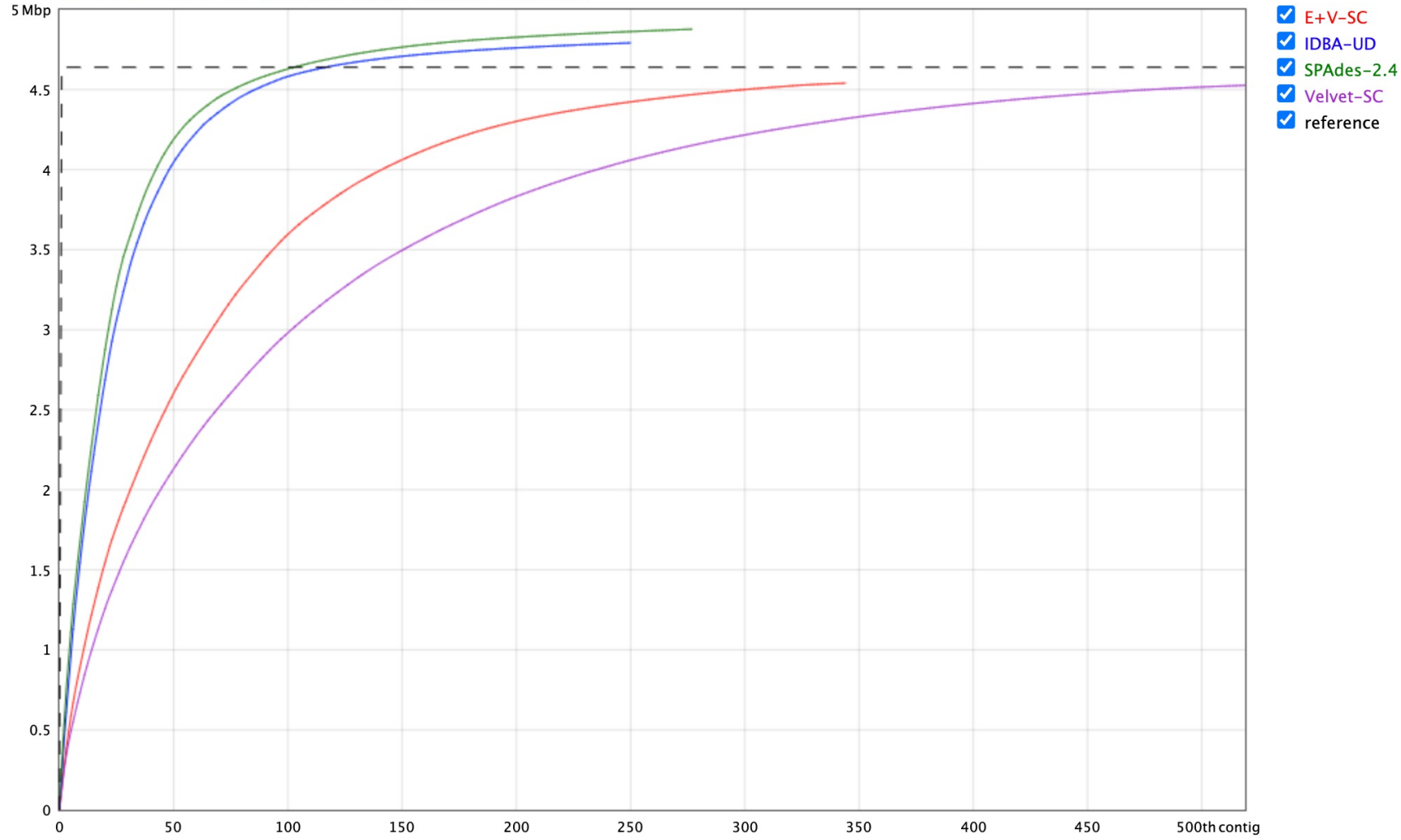
Aligned to "e.coli_reference" | 4 639 675 bp | 50.79% G+C | 884 operons

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

Worst Median Best ☑ Show heatmap

Statistics without reference	E+V-SC	IDBA-UD	SPAdes-2.4	Velvet-SC
# contigs	344	250	277	519
Largest contig	132 865	224 018	269 177	121 367
Total length	4 540 286	4 791 744	4 877 521	4 526 656
N50	33 616	96 947	106 927	20 445
Misassemblies				
# misassemblies	2	9	2	2
Misassembled contigs length	23 485	66 335	26 551	22 359
Mismatches				
# mismatches per 100 kbp	2.26	3.65	5.06	1.77
# indels per 100 kbp	0.7	0.2	0.7	0.92
# N's per 100 kbp	0	0	4.86	0
Genome statistics				
Genome fraction (%)	91.727	94.943	95.759	91.43
Duplication ratio	1.001	1.001	1.004	1.002
# genes	3767 + 160 part	4026 + 80 part	4046 + 102 part	3630 + 288 part
# operons	723 + 87 part	802 + 40 part	809 + 48 part	650 + 158 part
NGA50	32 051	96 947	110 539	19 791
Predicted genes				
# predicted genes (unique)	4258	4394	4417	4331
# predicted genes (>= 0 bp)	4258	4394	4490	4331
# predicted genes (>= 300 bp)	3643	3736	3784	3666
# predicted genes (>= 1500 bp)	524	559	559	515
# predicted genes (>= 3000 bp)	44	49	48	39

Plots: Cumulative length | Nx | NAx | NGx | NGAx | Operons | GC content



Assembly QC using Quast

Open ”report.html” in your browser

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

22 May 2022, Sunday, 19:31:11

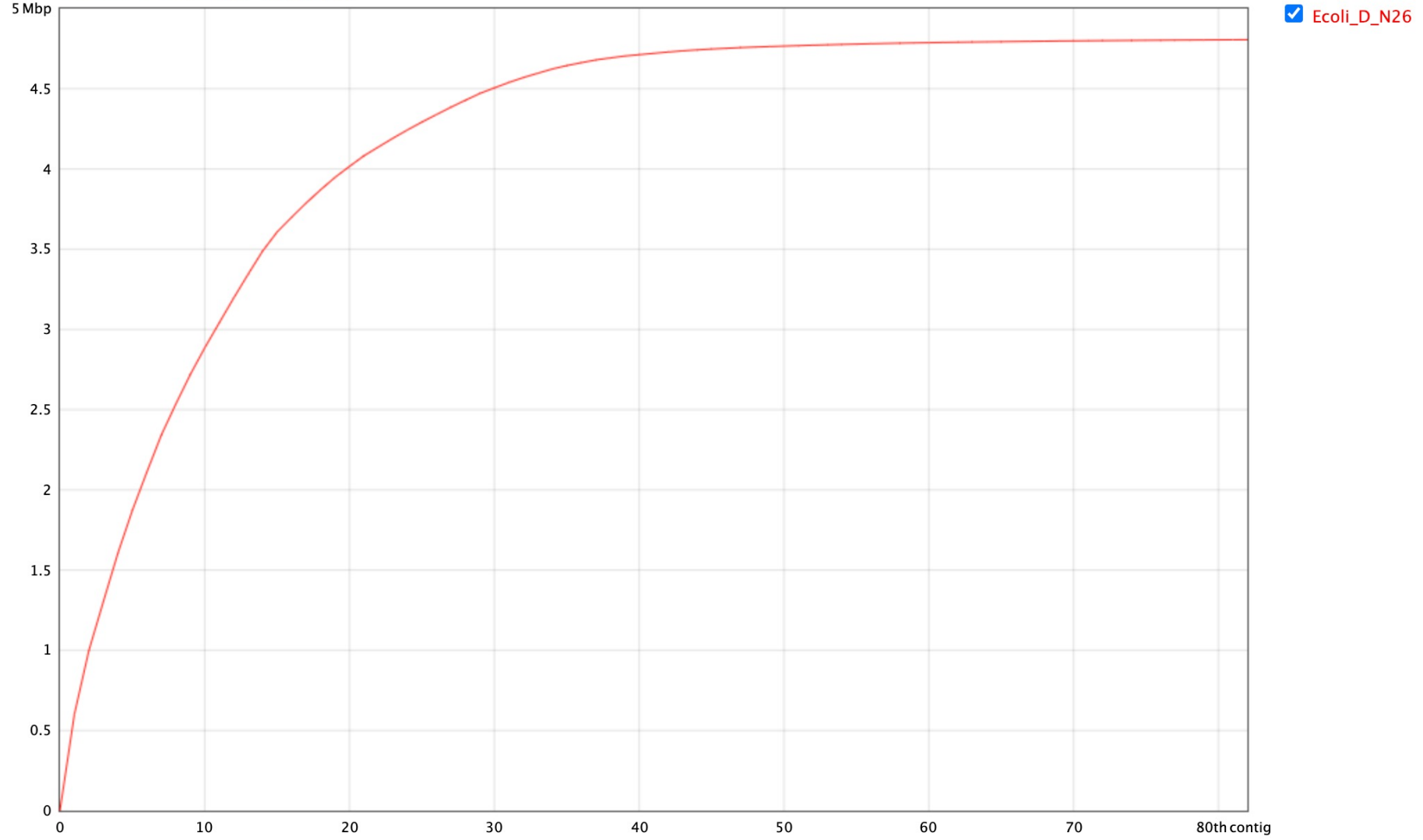
View in Icarus contig browser

All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

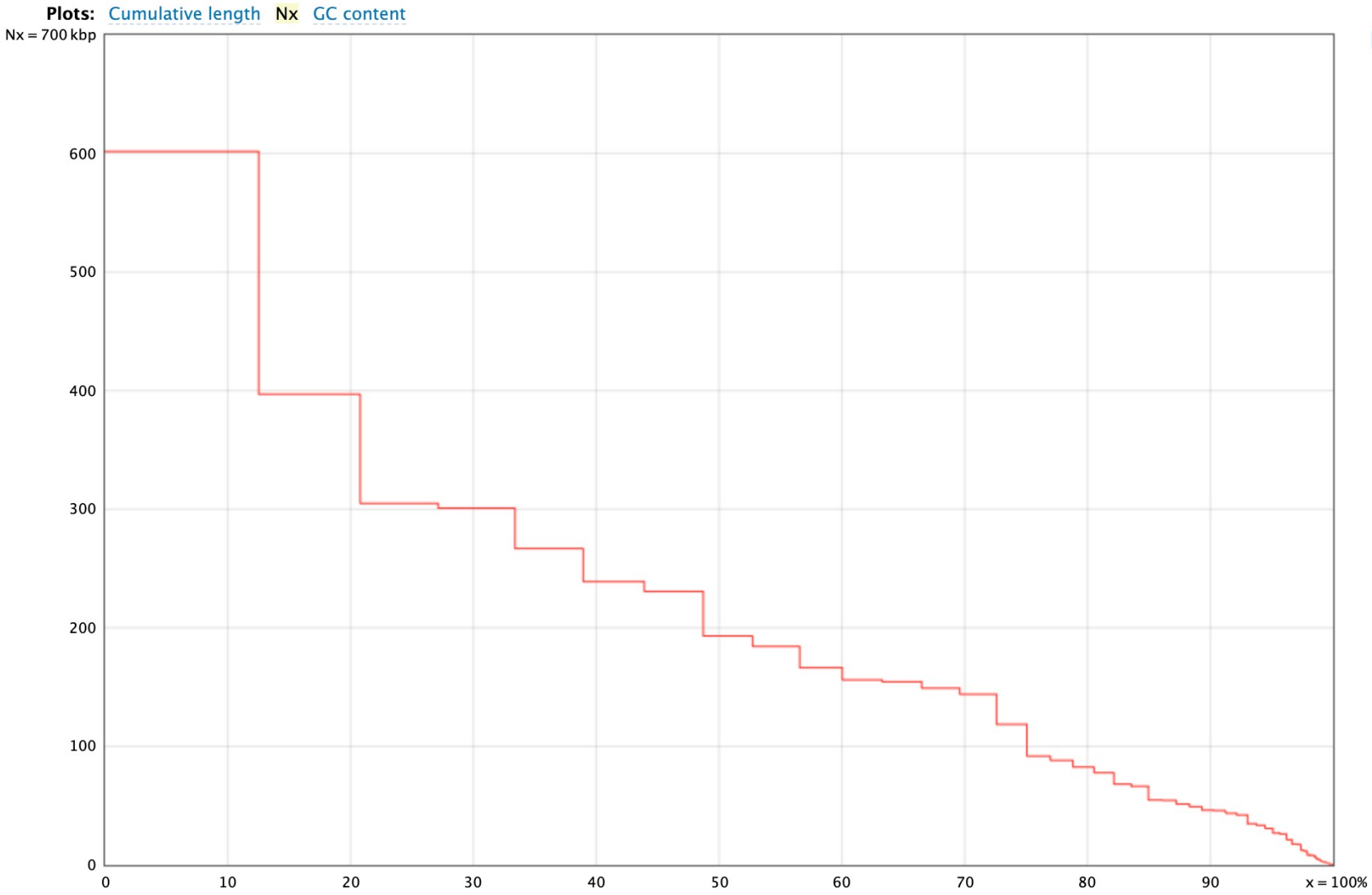
Statistics without reference	Ecoli_D_N26
# contigs	82
# contigs (>= 0 bp)	159
# contigs (>= 1000 bp)	69
# contigs (>= 5000 bp)	45
# contigs (>= 10000 bp)	39
# contigs (>= 25000 bp)	34
# contigs (>= 50000 bp)	24
Largest contig	601 691
Total length	4 806 061
Total length (>= 0 bp)	4 823 512
Total length (>= 1000 bp)	4 797 788
Total length (>= 5000 bp)	4 747 706
Total length (>= 10000 bp)	4 704 036
Total length (>= 25000 bp)	4 623 235
Total length (>= 50000 bp)	4 243 509
N50	193 217
N75	118 658
L50	8
L75	15
GC (%)	50.77
Mismatches	
# N's	0
# N's per 100 kbp	0

Quast web interface: http://cab.cc.spbu.ru/quast/

Plots: Cumulative length | Nx | GC content

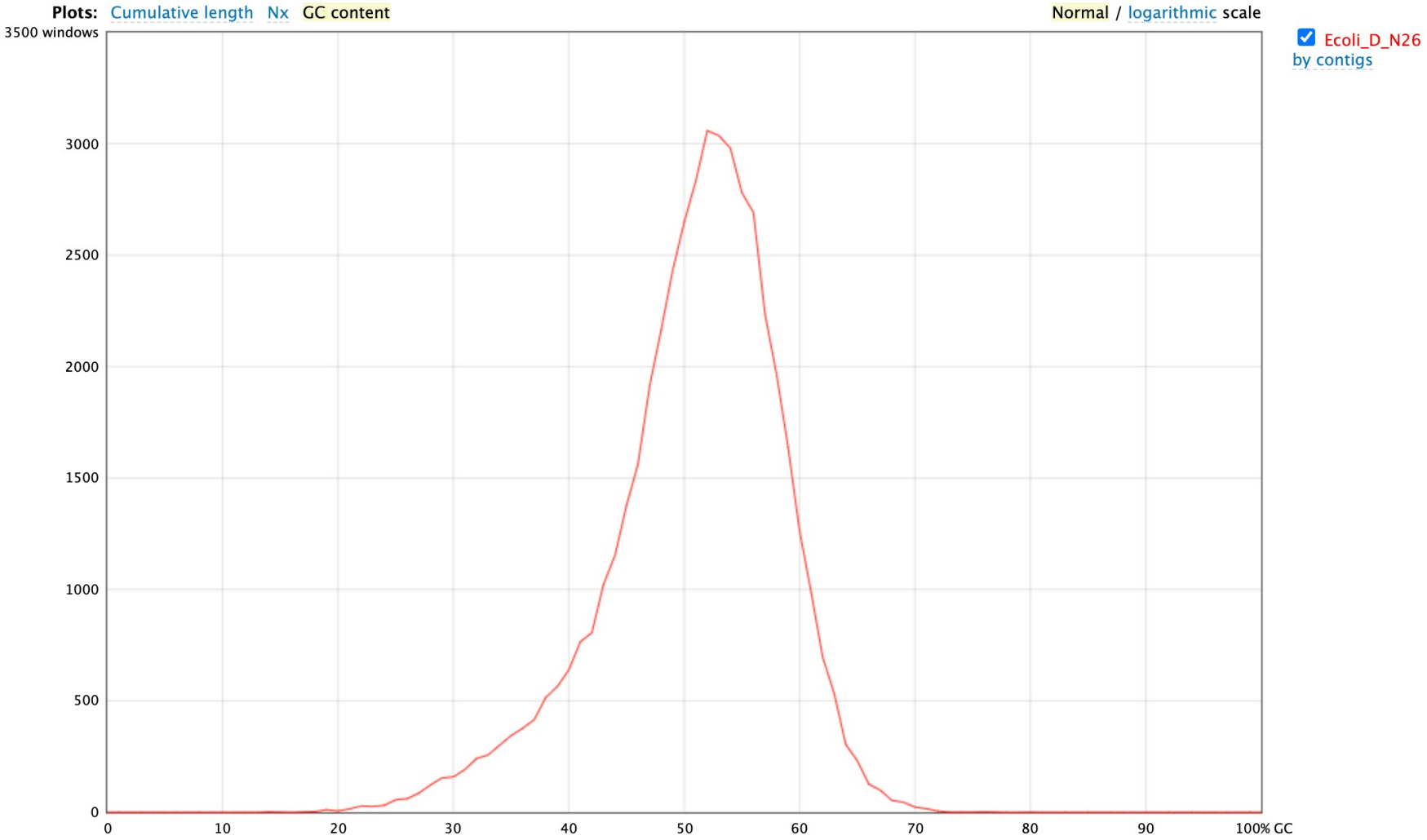


Contigs are ordered from largest (contig #1) to smallest.



- *Nx (where* $0 \leq x \leq 100$*)*: The largest contig length, L, such that using contigs of length $\geq L$ accounts for at least x% of the bases of the assembly.

`Nx distributions = cumulative length`



Contigs are broken into nonoverlapping 100 bp windows. Plot shows number of windows for each GC percentage.

Assembly QC using **Quast**

Command: more report.tsv

```
Assembly Ecoli_D_N26
# contigs (>= 0 bp) 159
# contigs (>= 1000 bp) 69
# contigs (>= 5000 bp) 45
# contigs (>= 10000 bp) 39
# contigs (>= 25000 bp) 34
# contigs (>= 50000 bp) 24
Total length (>= 0 bp) 4823512
Total length (>= 1000 bp) 4797788
Total length (>= 5000 bp) 4747706
Total length (>= 10000 bp) 4704036
Total length (>= 25000 bp) 4623235
Total length (>= 50000 bp) 4243509
# contigs 82
Largest contig 601691
Total length 4806061
GC (%) 50.77
N50 193217
N75 118658
L50 8
L75 15
# N's per 100 kbp 0.00
```

Example of a Bad Assembly: Ecoli-A_38843.fasta

QUAST

Quality Assessment Tool for Genome Assemblies by CAB

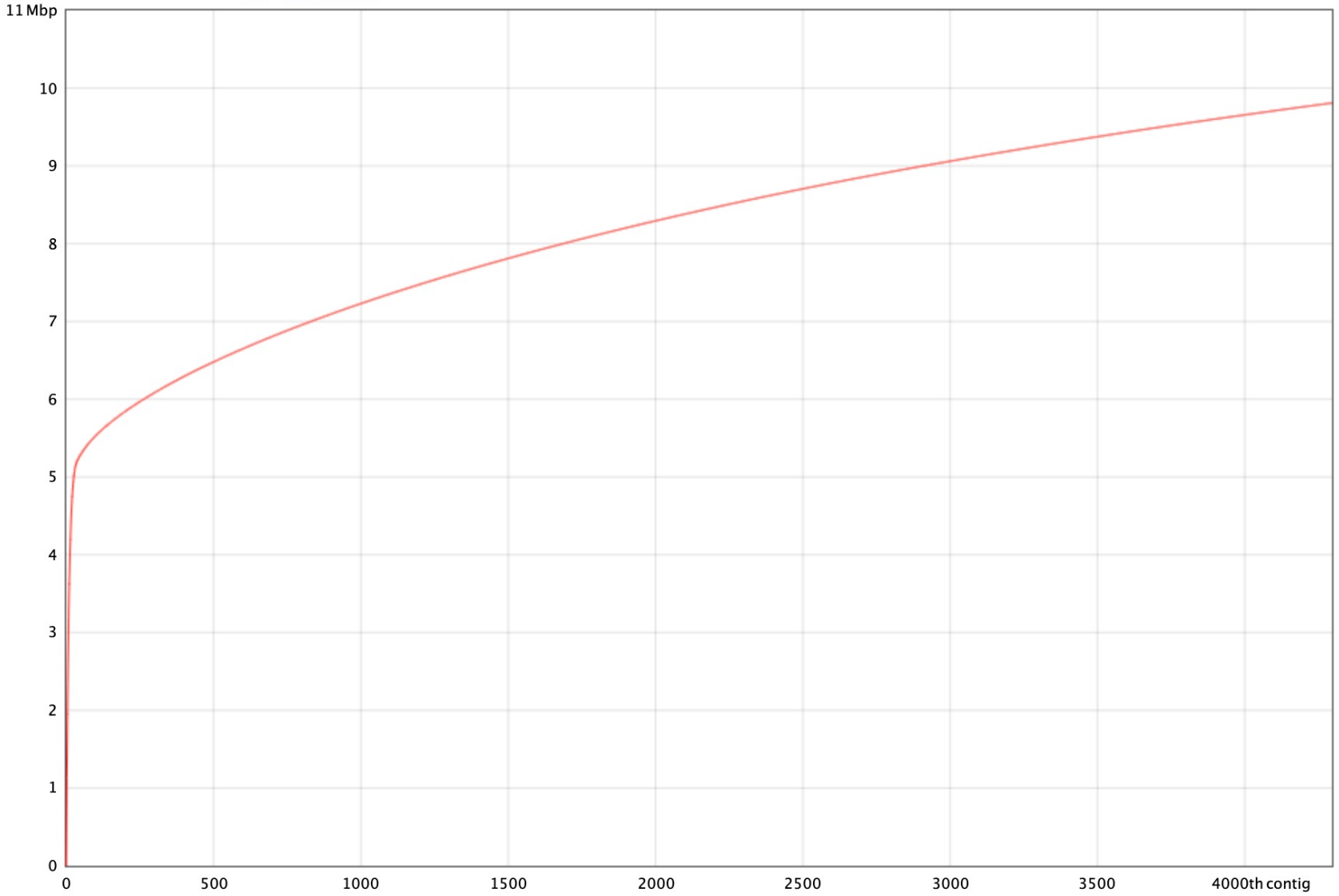
22 May 2022, Sunday, 19:30:18

View in Icarus contig browser

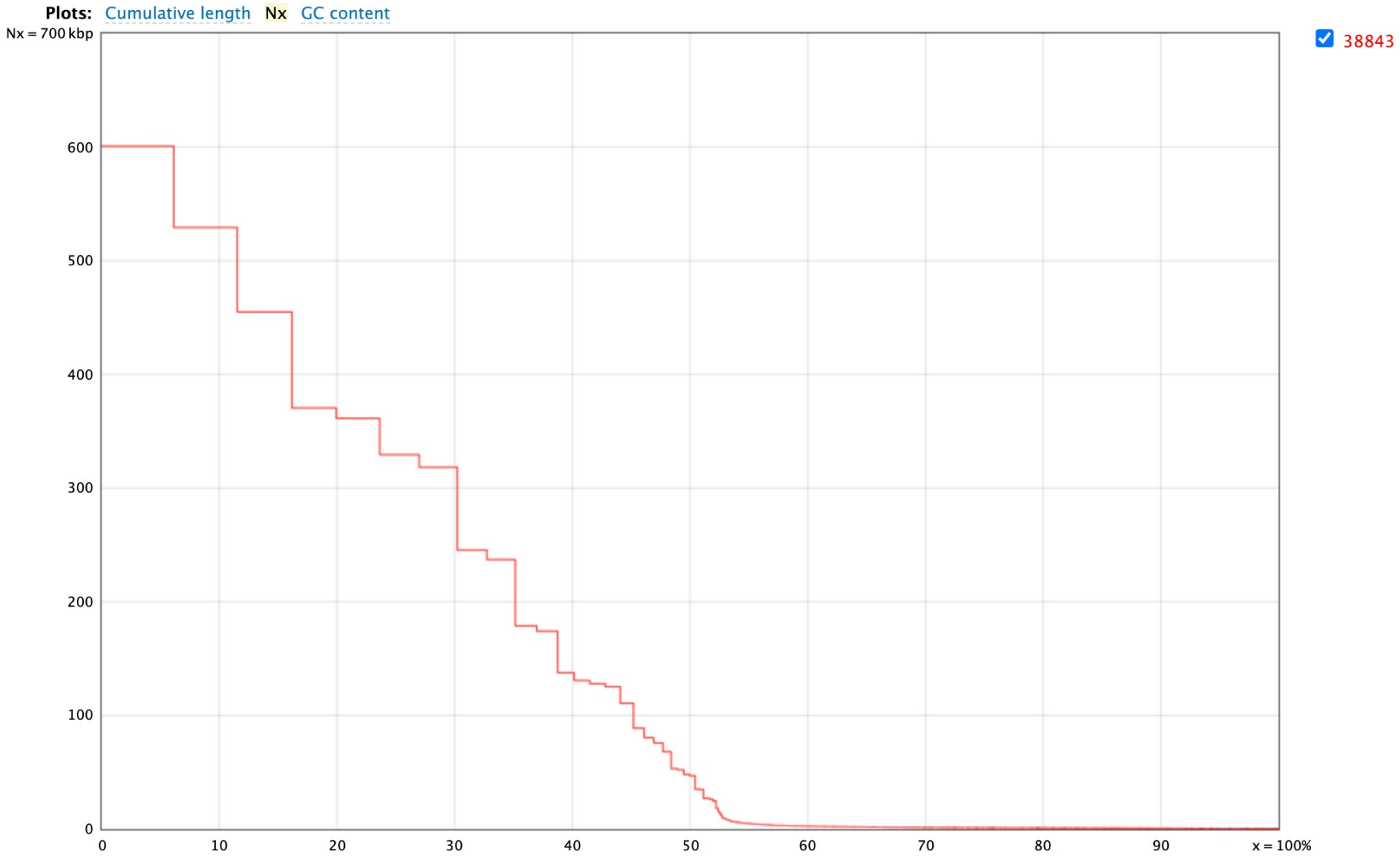
All statistics are based on contigs of size >= 500 bp, unless otherwise noted (e.g., "# contigs (>= 0 bp)" and "Total length (>= 0 bp)" include all contigs).

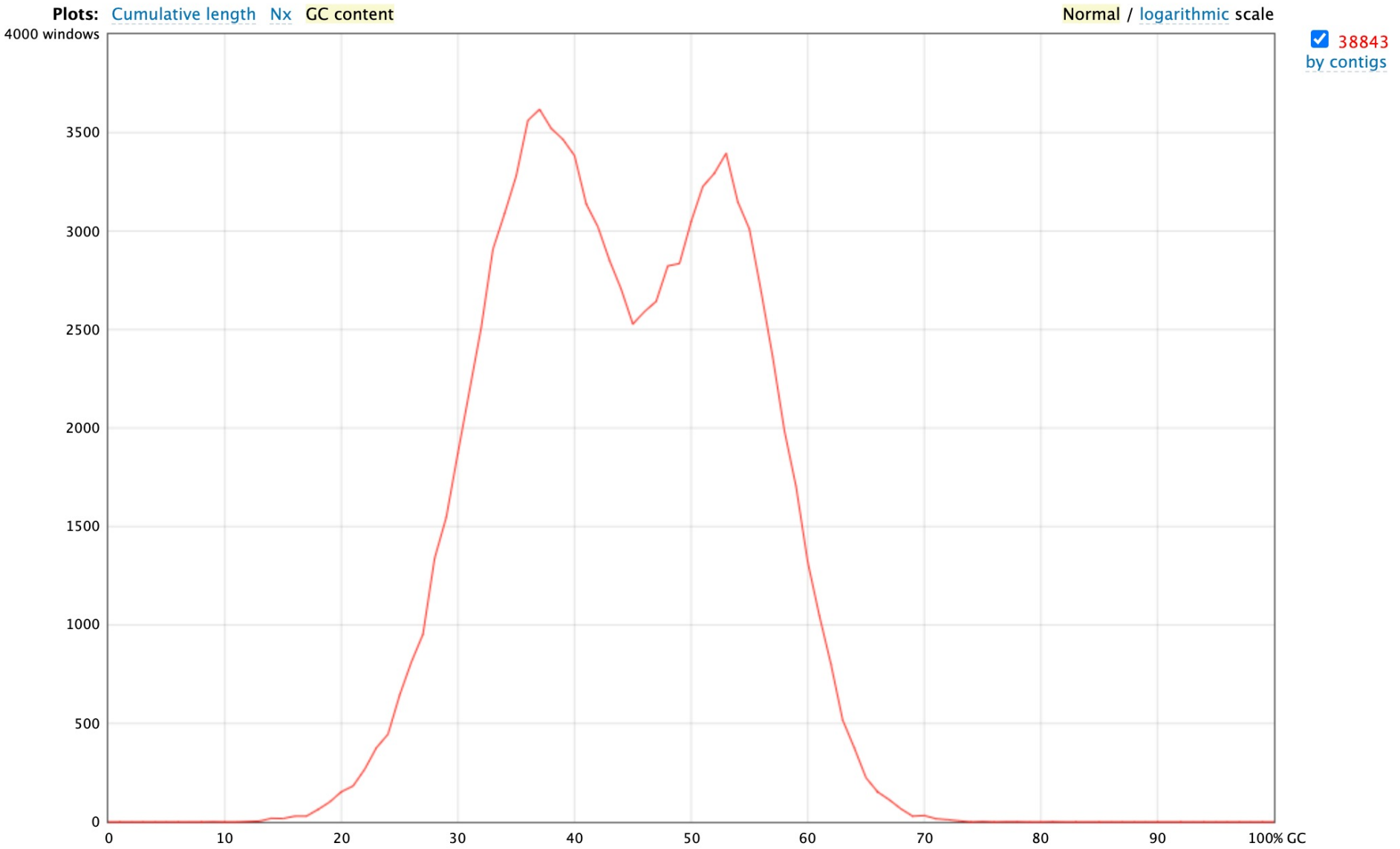
Statistics without reference	**38843**
# contigs	4297
# contigs (>= 0 bp)	6317
# contigs (>= 1000 bp)	1642
# contigs (>= 5000 bp)	66
# contigs (>= 10000 bp)	34
# contigs (>= 25000 bp)	29
# contigs (>= 50000 bp)	22
Largest contig	600 783
Total length	9 808 727
Total length (>= 0 bp)	10 672 464
Total length (>= 1000 bp)	7 955 898
Total length (>= 5000 bp)	5 384 113
Total length (>= 10000 bp)	5 177 649
Total length (>= 25000 bp)	5 096 465
Total length (>= 50000 bp)	4 851 041
N50	47 013
N75	1234
L50	24
L75	1101
GC (%)	43.79
Mismatches	
# N's	0
# N's per 100 kbp	0

Plots: Cumulative length | Nx | GC content



Contigs are ordered from largest (contig #1) to smallest.


Plots: Cumulative length Nx GC content
Nx = 700 kbp
38843
600
500
400
300
200
100
0
0
10
20
30
40
50
60
70
80
90
x = 100%

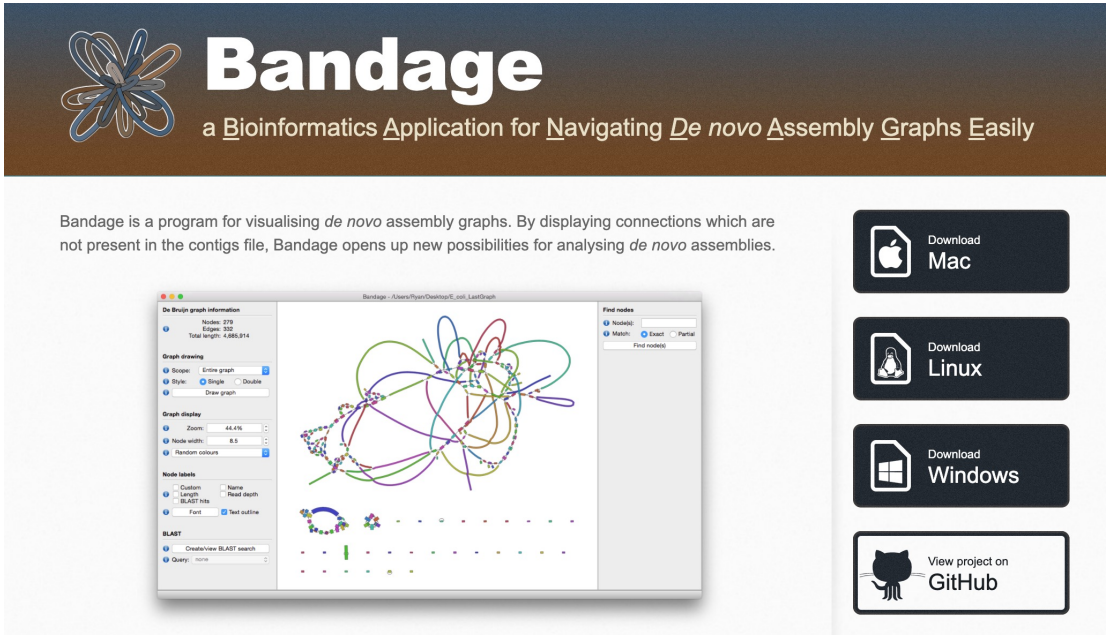


Contigs are broken into nonoverlapping 100 bp windows. Plot shows number of windows for each GC percentage.

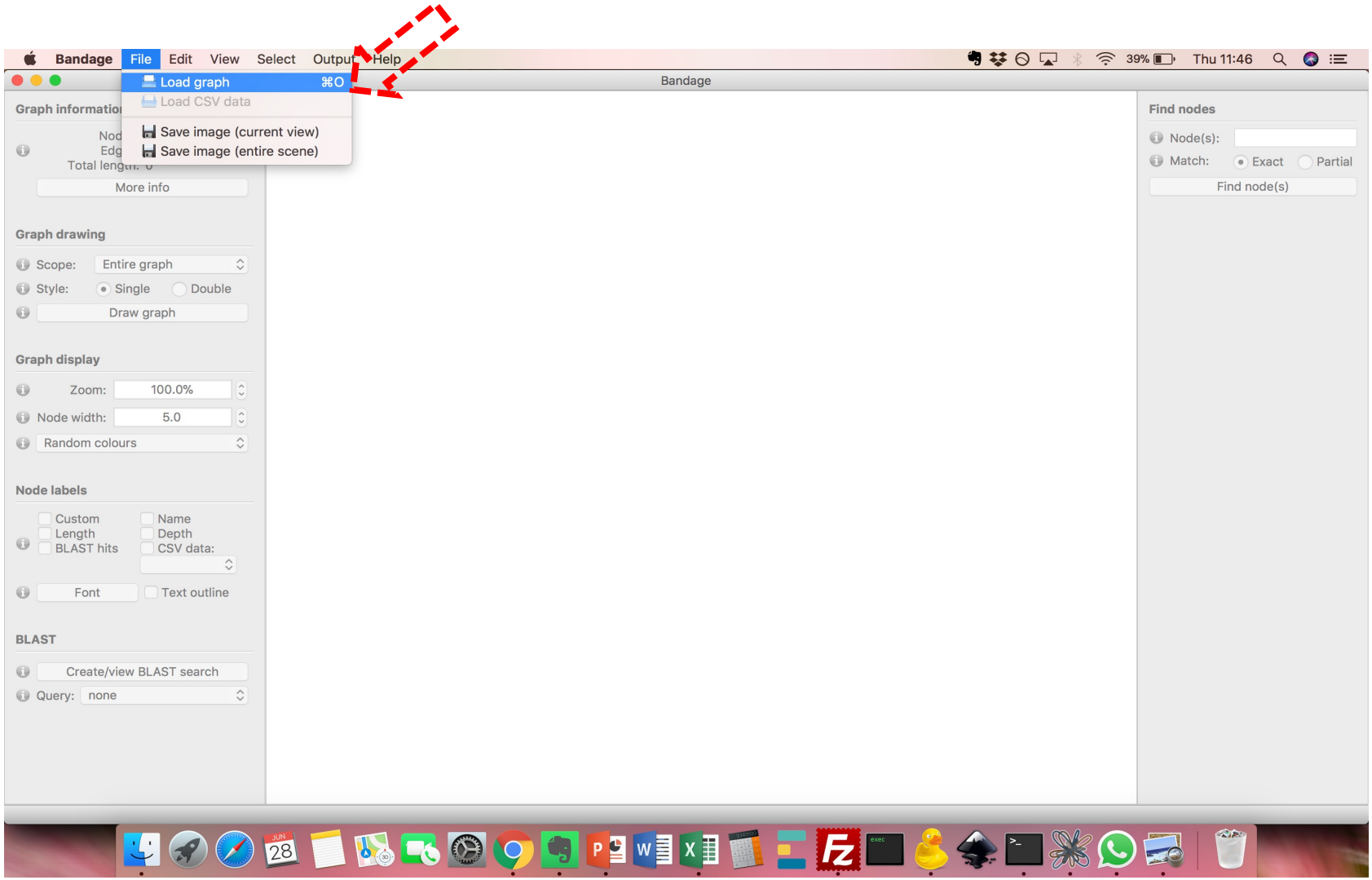
Genome Visualization:

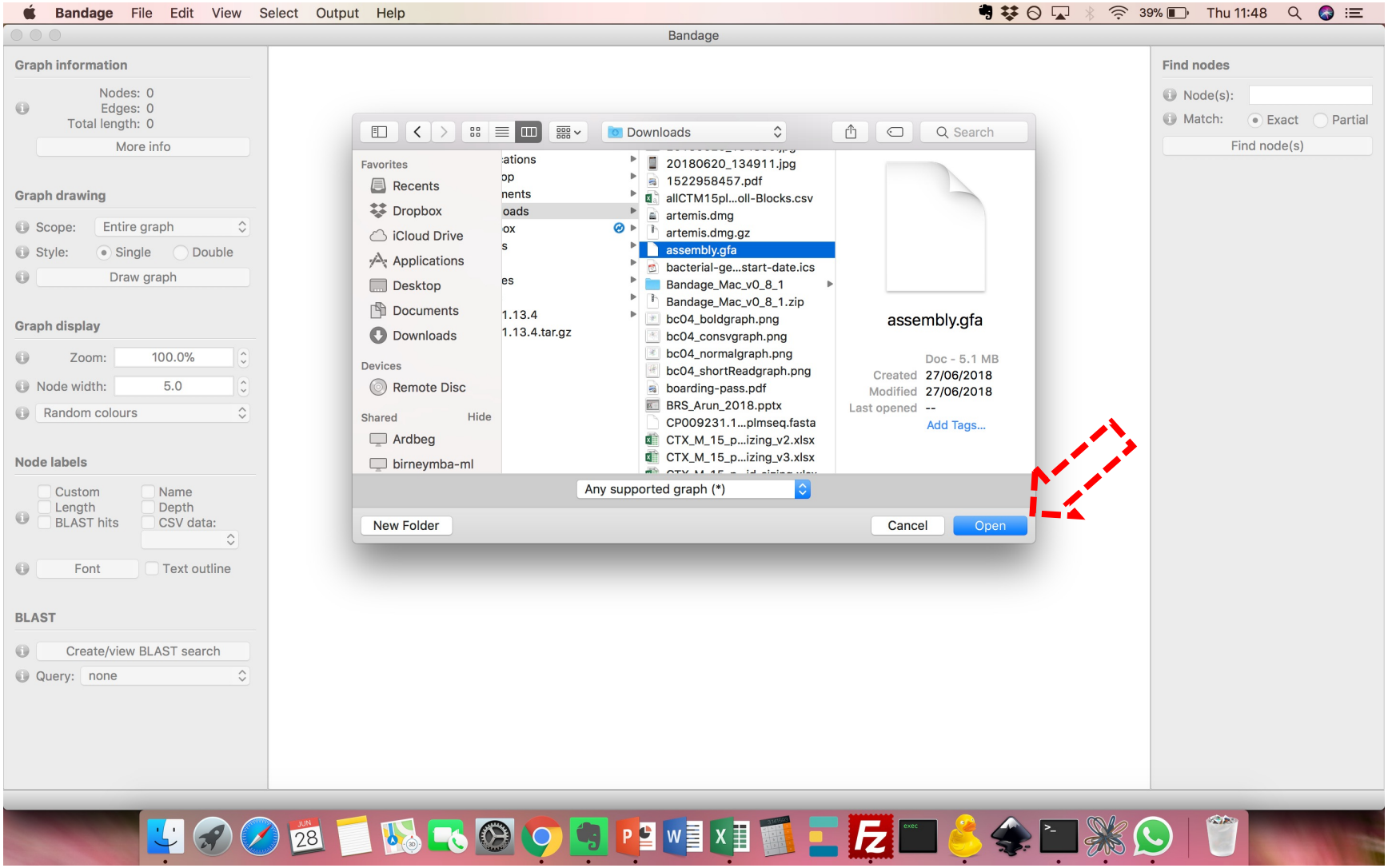
Bandage

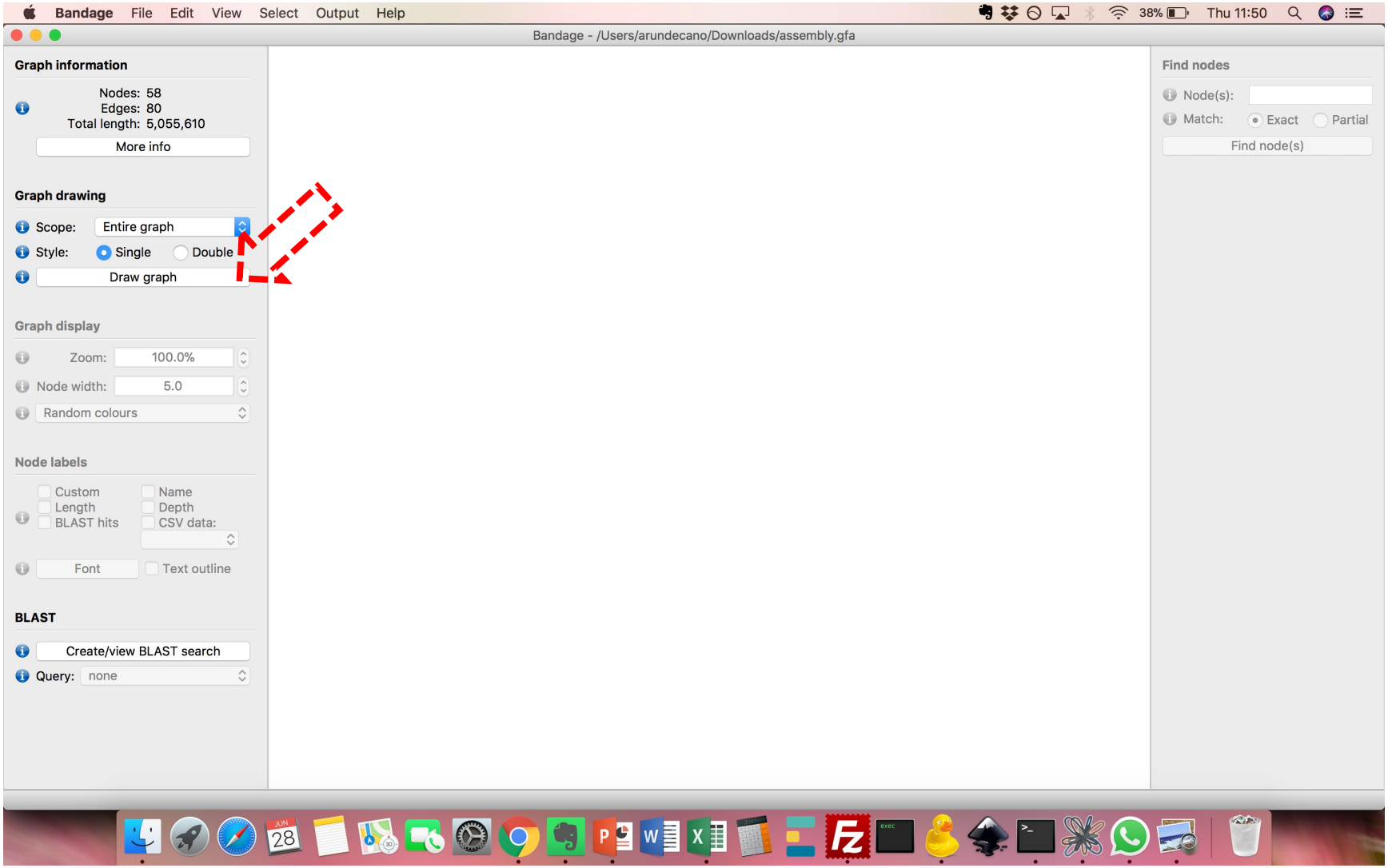
(Bioinformatics Application for Navigating *De novo* Assembly Graphs Easily)

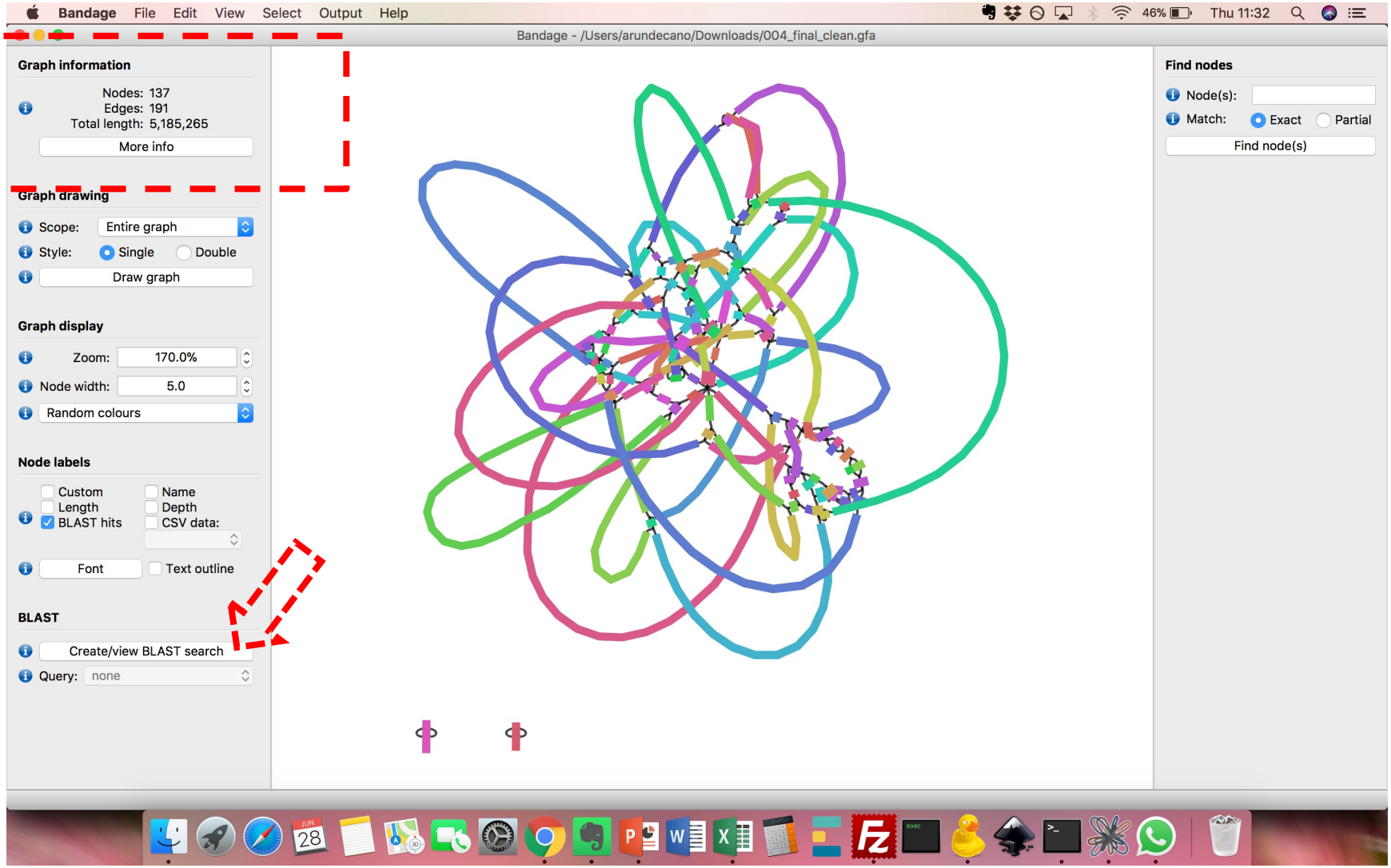


https://rrwick.github.io/Bandage/


Bandage
File
Edit
View
Select
Output
Help
Load graph ⌘O
Load CSV data
Save image (current view)
Save image (entire scene)
39%
Thu 11:46
Bandage
Graph information
More info
Graph drawing
Scope:
Entire graph
Style:
Single
Double
Draw graph
Graph display
Zoom:
100.0%
Node width:
5.0
Random colours
Node labels
Custom
Length
BLAST hits
Name
Depth
CSV data:
Font
Text outline
BLAST
Create/view BLAST search
Query:
none
Find nodes
Node(s):
Match:
Exact
Partial
Find node(s)


Bandage File Edit View Select Output Help
39% Thu 11:48
Bandage
Graph information
Nodes: 0
Edges: 0
Total length: 0
More info
Graph drawing
Scope: Entire graph
Style: Single Double
Draw graph
Graph display
Zoom: 100.0%
Node width: 5.0
Random colours
Node labels
Custom
Length
BLAST hits
Name
Depth
CSV data:
Font
Text outline
BLAST
Create/view BLAST search
Query: none
Find nodes
Node(s):
Match: Exact Partial
Find node(s)
Downloads
Search
Favorites
Recents
Dropbox
iCloud Drive
Applications
Desktop
Documents
Downloads
Devices
Remote Disc
Shared
Hide
Ardbeg
birneymba-ml
20180620_134911.jpg
1522958457.pdf
allCTM15pl...oll-Blocks.csv
artemis.dmg
artemis.dmg.gz
assembly.gfa
bacterial-ge...start-date.ics
Bandage_Mac_v0_8_1
Bandage_Mac_v0_8_1.zip
bc04_boldgraph.png
bc04_consvgraph.png
bc04_normalgraph.png
bc04_shortReadgraph.png
boarding-pass.pdf
BRS_Arun_2018.pptx
CP009231.1...plmseq.fasta
CTX_M_15_p...izing_v2.xlsx
CTX_M_15_p...izing_v3.xlsx
assembly.gfa
Doc - 5.1 MB
Created 27/06/2018
Modified 27/06/2018
Last opened --
Add Tags...
Any supported graph (*)
New Folder
Cancel
Open


Bandage File Edit View Select Output Help
38% Thu 11:50
Bandage - /Users/arundecano/Downloads/assembly.gfa
Graph information
Nodes: 58
Edges: 80
Total length: 5,055,610
More info
Graph drawing
Scope: Entire graph
Style: Single Double
Draw graph
Graph display
Zoom: 100.0%
Node width: 5.0
Random colours
Node labels
Custom
Length
BLAST hits
Name
Depth
CSV data:
Font
Text outline
BLAST
Create/view BLAST search
Query: none
Find nodes
Node(s):
Match: Exact Partial
Find node(s)


Bandage File Edit View Select Output Help
46% Thu 11:32
Bandage - /Users/arundecano/Downloads/004_final_clean.gfa
Graph information
Nodes: 137
Edges: 191
Total length: 5,185,265
More info
Graph drawing
Scope: Entire graph
Style: Single Double
Draw graph
Graph display
Zoom: 170.0%
Node width: 5.0
Random colours
Node labels
Custom
Length
BLAST hits
Name
Depth
CSV data:
Font
Text outline
BLAST
Create/view BLAST search
Query: none
Find nodes
Node(s):
Match: Exact Partial
Find node(s)

WORKSHEET

***De novo* Genome Assembly using Shovill**

(1) Select 2 *Escherichia coli* and 2 *Salmonella typhi* paired read libraries (from the ~/Raw_FQs/ directory) you want to work with.

(2) *De novo* assemble your chosen genomes by running the Shovill pipeline as below.

```
shovill --outdir [output_directory_name] --R1 [/path/to/R1.fq.gz/] --R2 [/path/to/R2.fq.gz/]
```

(3) Inspect the output files at the end of each Shovill run.

```
ls [output_directory_name]
```

Assessing the quality of assembled genomes using Quast

(1) Run Quast on your newly assembled genomes. The final assembled genomes from the Shovill run is be labelled as “contigs.fa” in the output directory. Rename the genome with the original sample name.

(2) From any of the summary report files, take note of the following parameters for each assembled genome:

 # contigs (>= 0 bp)
 # contigs (>= 1000 bp)
 # contigs (>= 5000 bp)
 # contigs (>= 10000 bp)
 # contigs (>= 25000 bp)
 # contigs (>= 50000 bp)
Total length (>= 0 bp)
Total length (>= 1000 bp)
Total length (>= 5000 bp)
Total length (>= 10000 bp)
Total length (>= 25000 bp)
Total length (>= 50000 bp)
 # contigs
Largest contig
Total length
GC (%)
N50
N75
L50
L75
 # N's per 100 kbp

(3) Which of the two *E. coli* genomes you assembled is more complete? Why so? Use and cite relevant assembly parameters to compare.

(4) Which of the two *S. typhi* genomes you assembled is more complete? Why so? Use and cite relevant assembly parameters to compare.