Introduction to genome assembly

Nadège Guiglielmoni

About me

PhD student at ULB and in the ITN IGNITE www.itn-ignite.eu

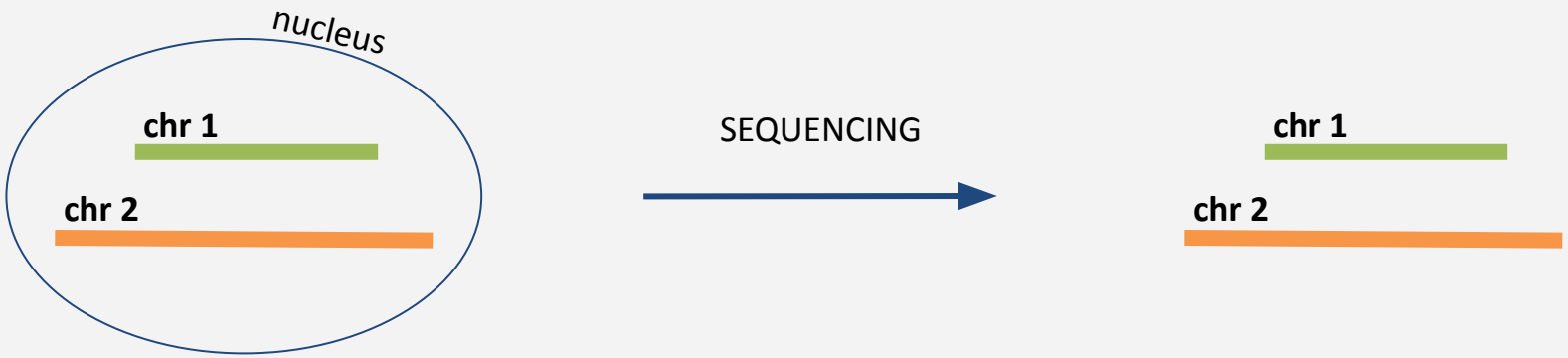
euraxess.ec.europa.eu/jobs/search

- Long-read assembly
- Hi-C scaffolding
- Genomics of non-vertebrates

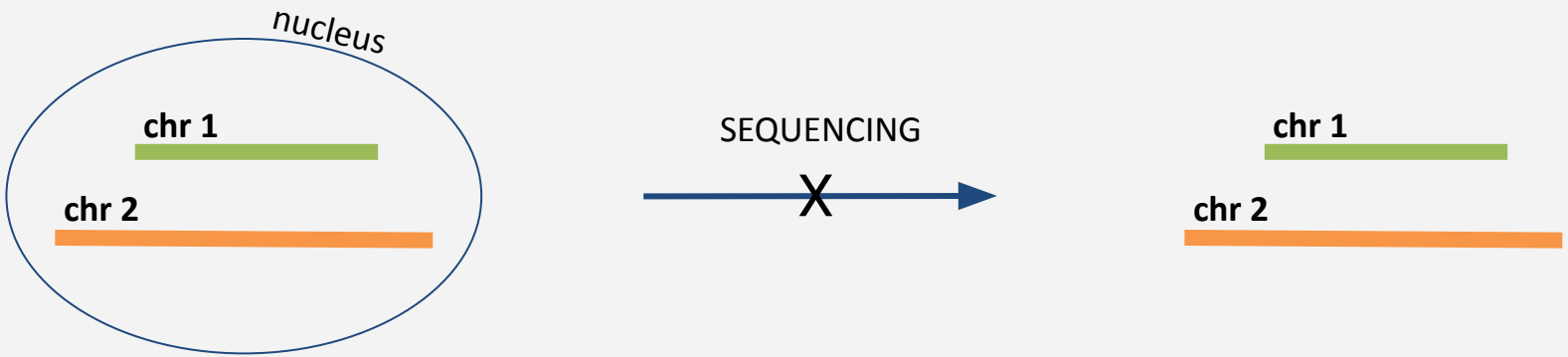
nadege.guiglielmoni@ulb.be

@NGuiglielmoni

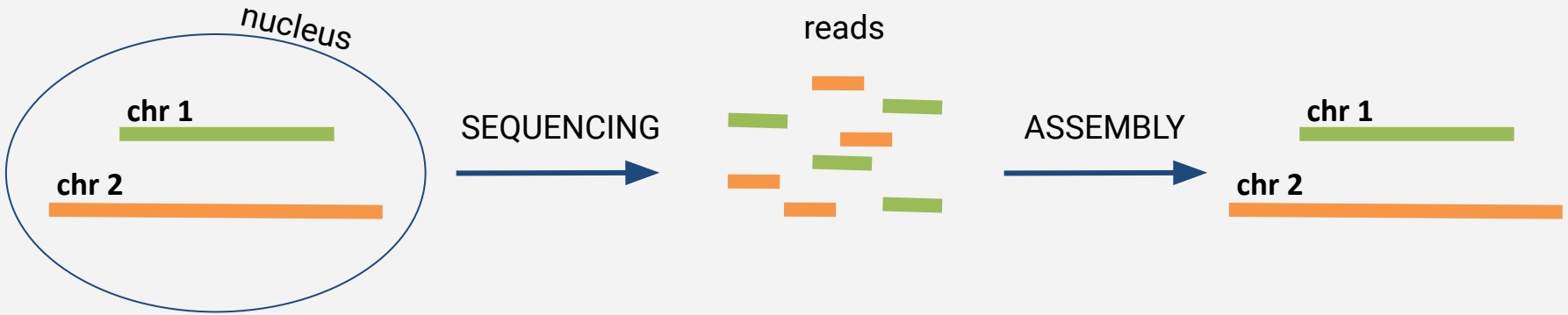
Introduction



Introduction



Introduction

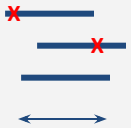


Introduction

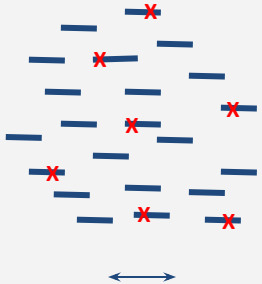
1977
First generation sequencing / Sanger

2005
Next Generation Sequencing

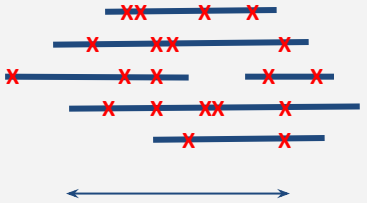
2009
Third Generation Sequencing



1 kb

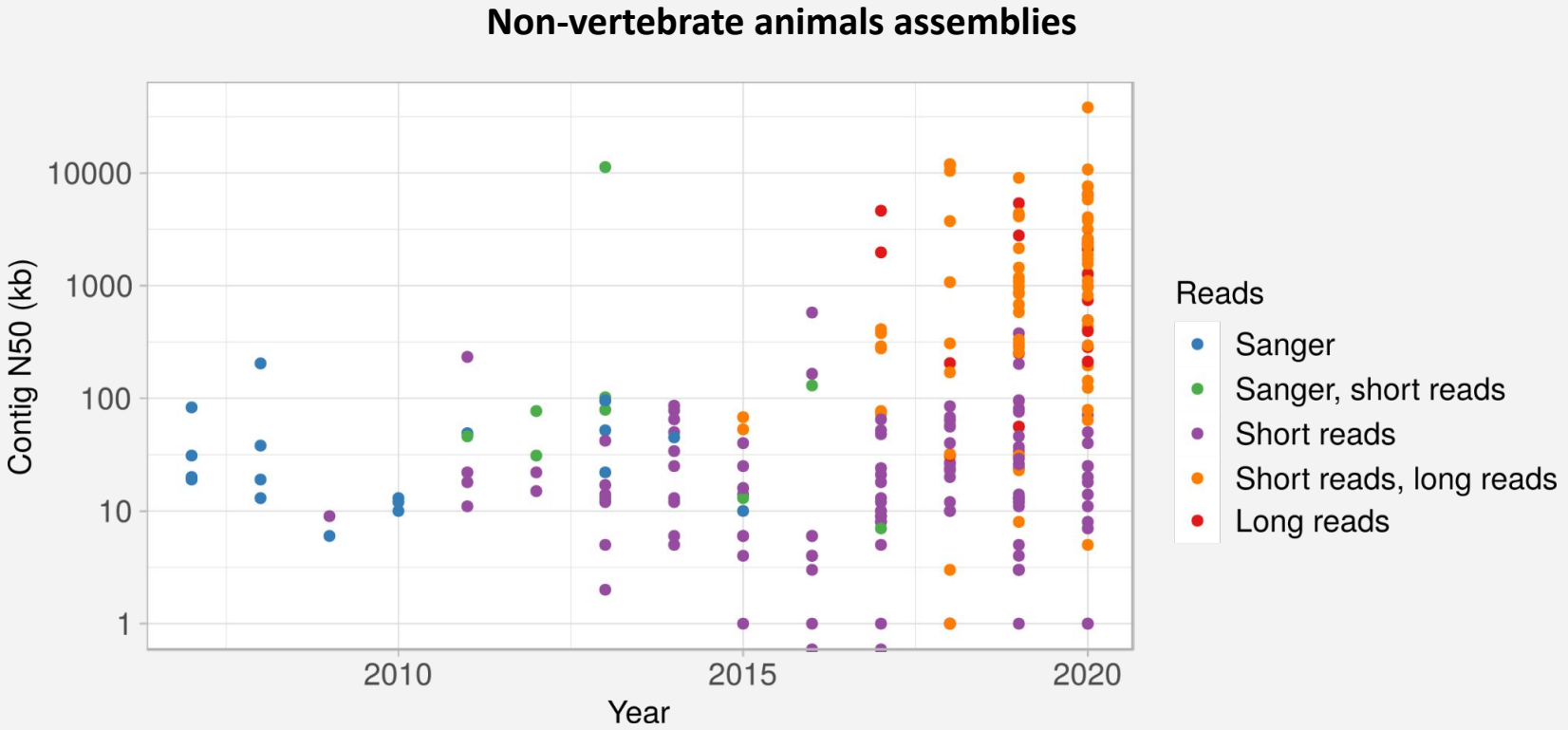


35 - 300 bp



10 - 100+ kb

Introduction



Introduction

- ➔ Assembly algorithms
- ➔ Reads pre-processing
- ➔ Assembly post-processing
- ➔ Scaffolding approaches
- ➔ Phasing assemblies
- ➔ Assembly evaluation

Assembly algorithms

Assemble the reads: reconstitute, from the set of reads, the suite of bases (A,T,G and C) characteristic of this genome

De novo assembly: no reference

Sources: http://www.langmead-lab.org/teaching-materials/

class of Jean-François Flot

Assembly algorithms: overlap graphs

Overlap: length-*l* suffix of X matches length-*l* prefix of Y, where *l* is given

l = 3

X: CTCTAGGCC
Y: TAGGCCCTC

Assembly algorithms: overlap graphs

Overlap: length-*l* suffix of X matches length-*l* prefix of Y, where *l* is given

l = 3

X: CTCTAGGCC
Y: TAGGCCCTC

X: CTCTAGGCC
Y: TAGGCCCTC

Assembly algorithms: overlap graphs

Overlap: length-*l* suffix of X matches length-*l* prefix of Y, where *l* is given

l = 3

X: CTCTAGGCC
Y: TAGGCCCTC

X: CTCTAGGCC
Y: TAGGCCCTC

X: CTCTAGGCC
Y: TAGGCCCTC

Assembly algorithms: overlap graphs

Overlap: length-*l* suffix of X matches length-*l* prefix of Y, where *l* is given

l = 3

X: CTCTAGGCC
Y: TAGGCCCTC

X: CTCTAGGCC
Y: TAGGCCCTC

X: CTCTAGGCC
Y: TAGGCCCTC

Assembly algorithms: overlap graphs

Overlap: length-*l* suffix of X matches length-*l* prefix of Y, where *l* is given

l = 3

X: CTCTAGGCC
Y: TAGGCCCTC

X: CTCTAGGCC
Y: TAGGCCCTC

a length-6 prefix of Y matches a suffix of X

X: CTCTAGGCC
Y: TAGGCCCTC

Assembly algorithms: overlap graphs

Graph vocabulary:

- **vertex/node**
- **edge/arc**: connects vertices
- **directed graph** (digraph): set of vertices and directed edges

Overlap graph = directed graph

- nodes = reads
- edges = overlaps betwen reads

Assembly algorithms: overlap graphs

Vertices (reads): { *a:* CTCTAGGCC, *b:* GCCCTCAAT, *c:* CAATTTTT }

Edges (overlaps): { (*a*, *b*), (*b*, *c*) }


```
CTCTAGGCC          GCCCTCAAT
|||                ||||
GCCCTCAAT          CAATTTTT
```

Assembly algorithms: overlap graphs

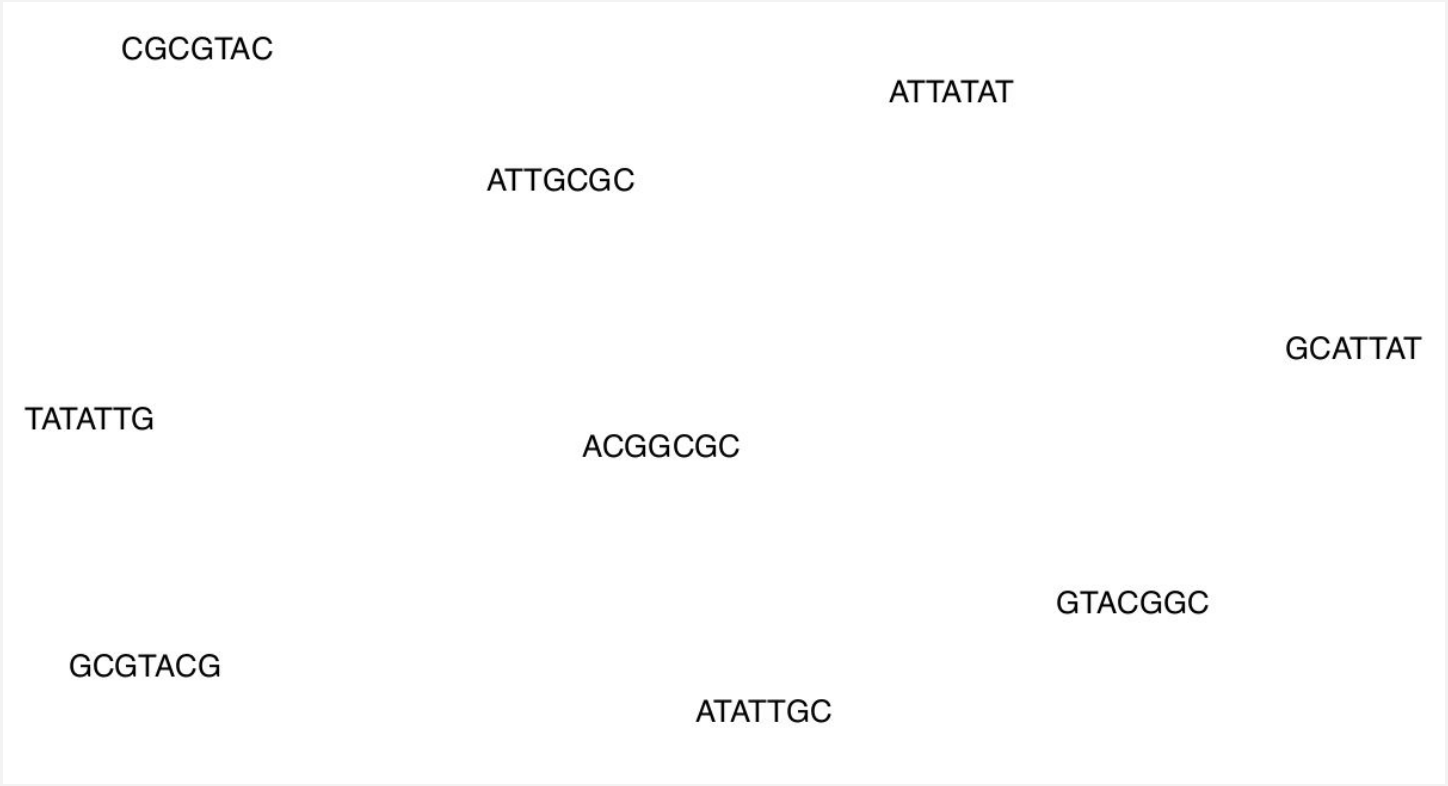
Exercise: find all the overlaps of minimum 3 bases between the following reads

How many overlaps do you find?

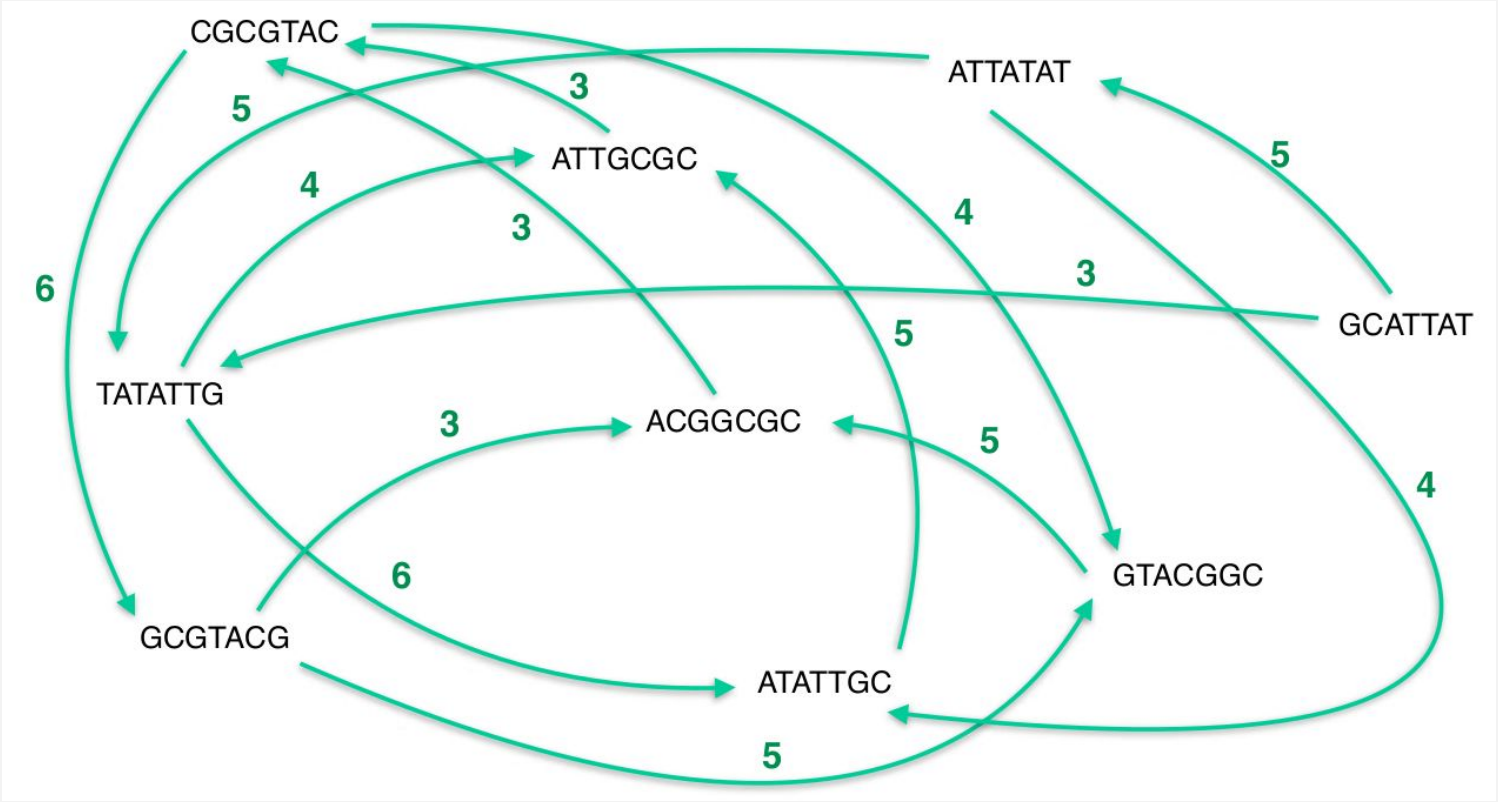
Go to www.menti.com to give your answer, code: 3123 7622

Assembly algorithms: overlap graphs

menti: 3123 7622



Assembly algorithms: overlap graphs

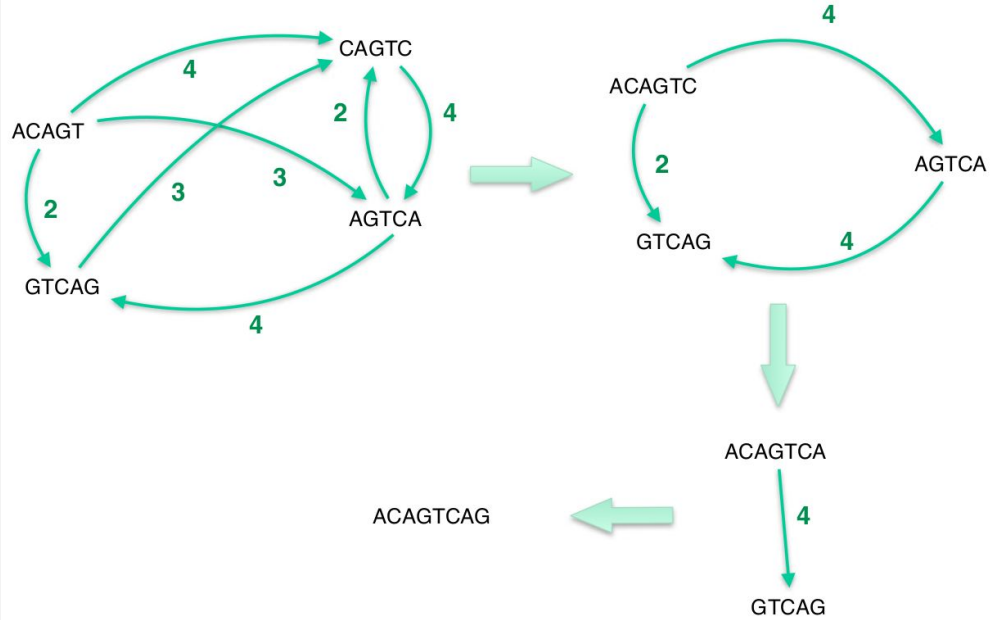


Assembly algorithms: greedy assemblers

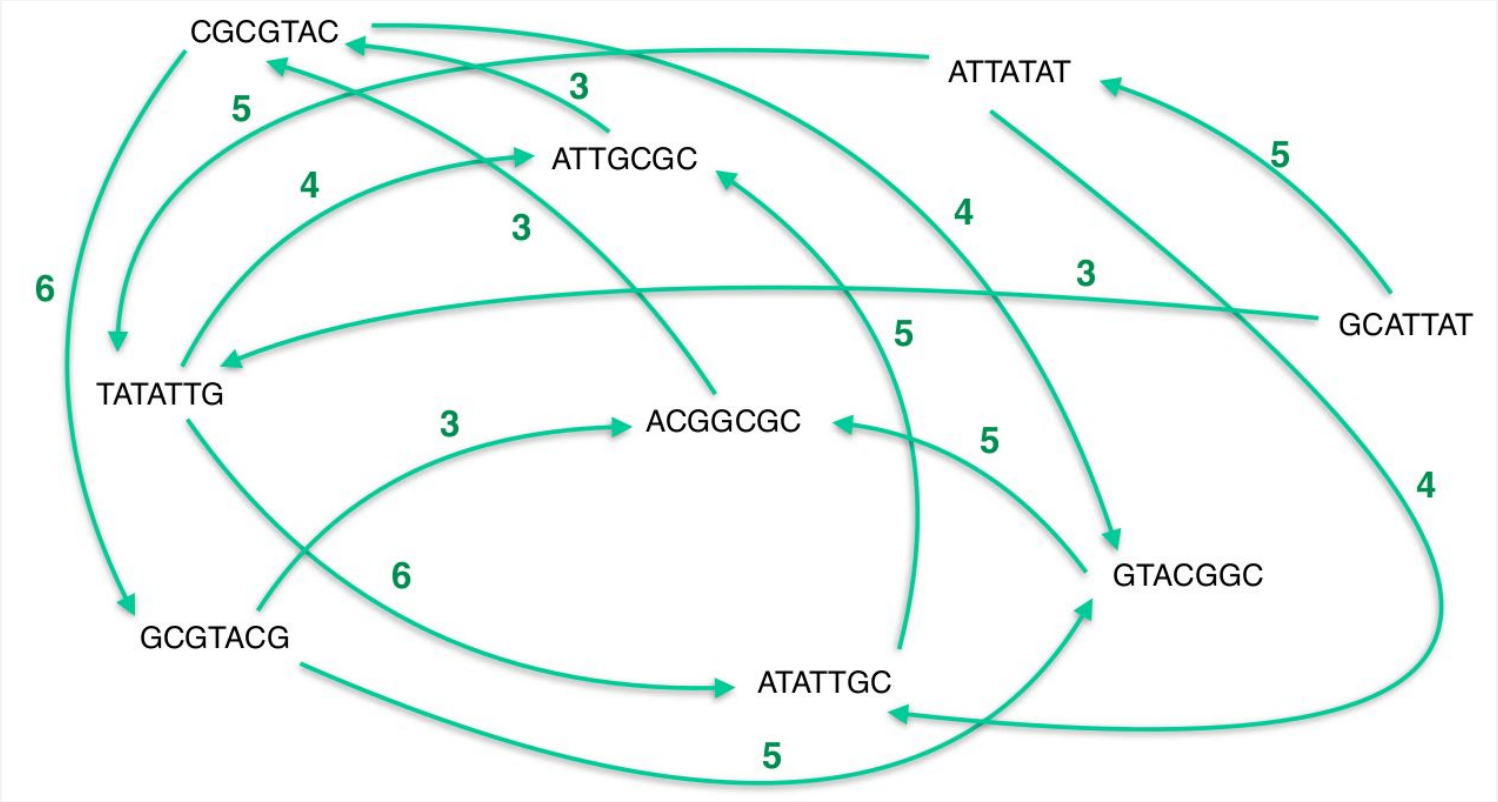
Greedy approach: find the shortest common superstring

→ merge the best overlapping reads

ex: CAP3, phrap, TIGR



Assembly algorithms: greedy assemblers menti: 3123 7622



Assembly algorithms: greedy assemblers

Solution:

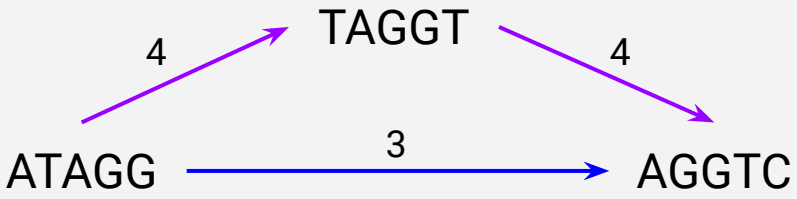
GCATTATATTGCGCGTACGGCGC

Assembly algorithms: Overlap-Layout-Consensus

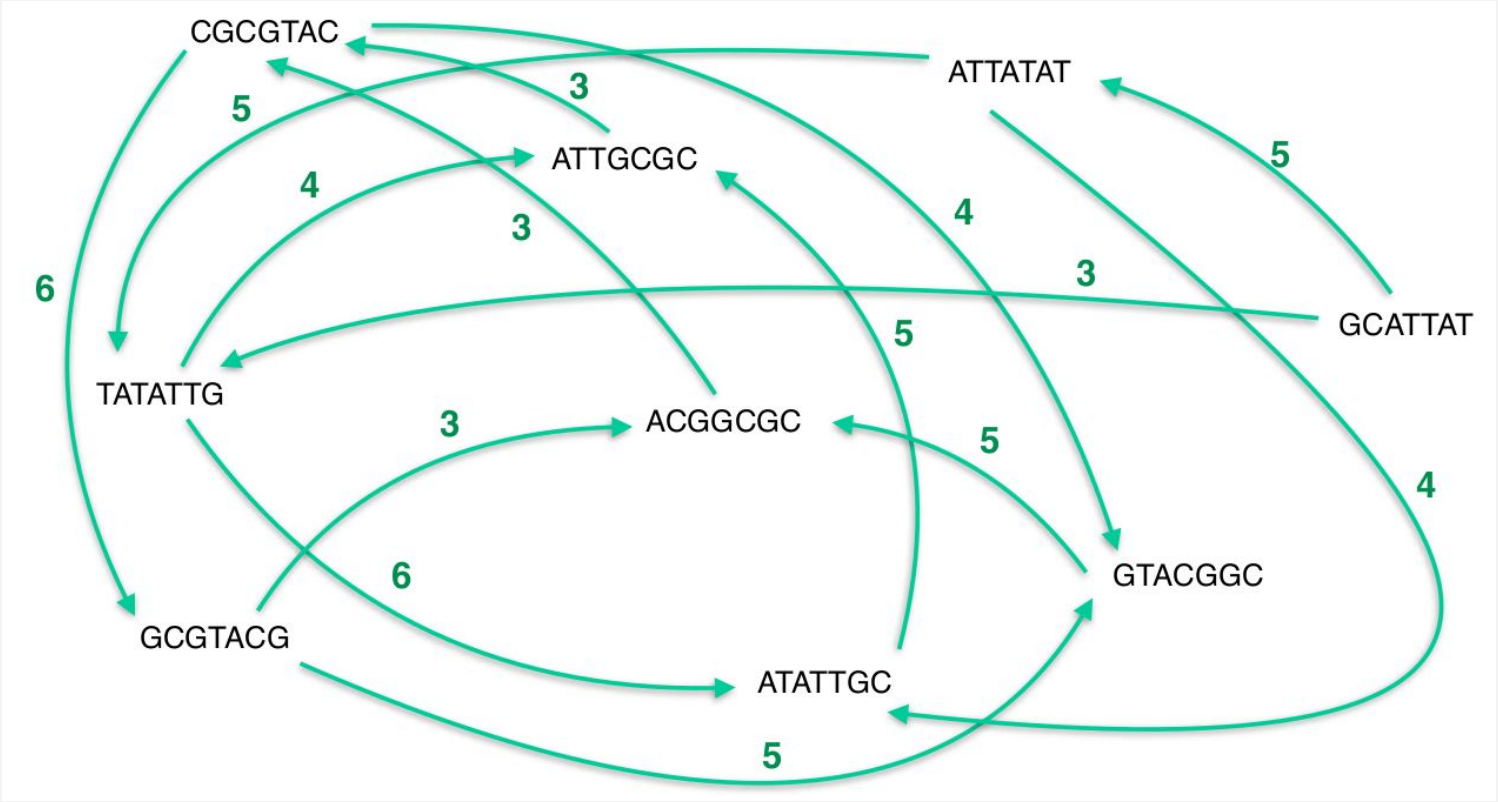
Overlap-Layout-Consensus (OLC): disantangle repeats by looking for the shortest generalised Hamiltonian path in the overlap graph

Hamiltonian path: goes through each node once

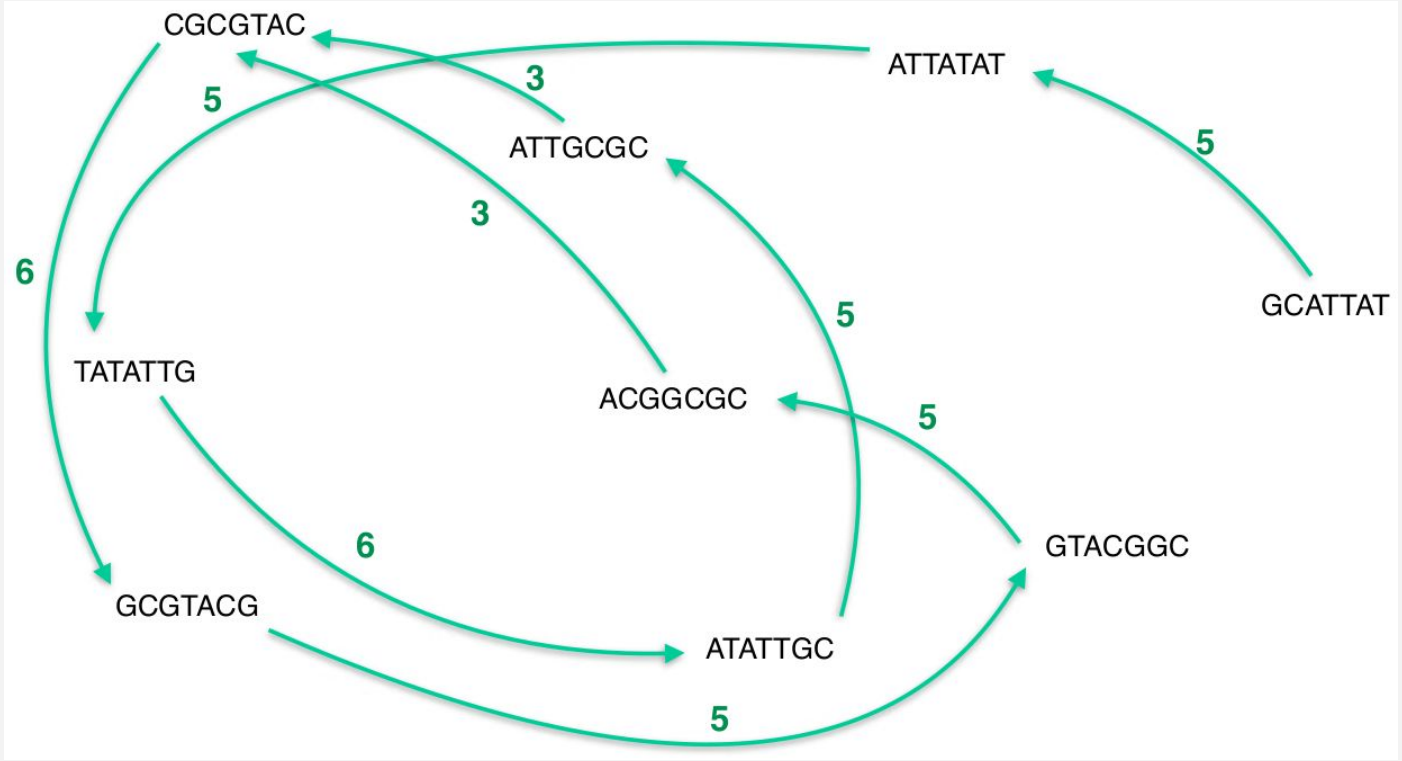
Layout: simplify the graph by removing redundant edges



Assembly algorithms: Overlap-Layout-Consensus



Assembly algorithms: Overlap-Layout-Consensus



Assembly algorithms: Overlap-Layout-Consensus

Consensus step

Assembly algorithms: Overlap-Layout-Consensus

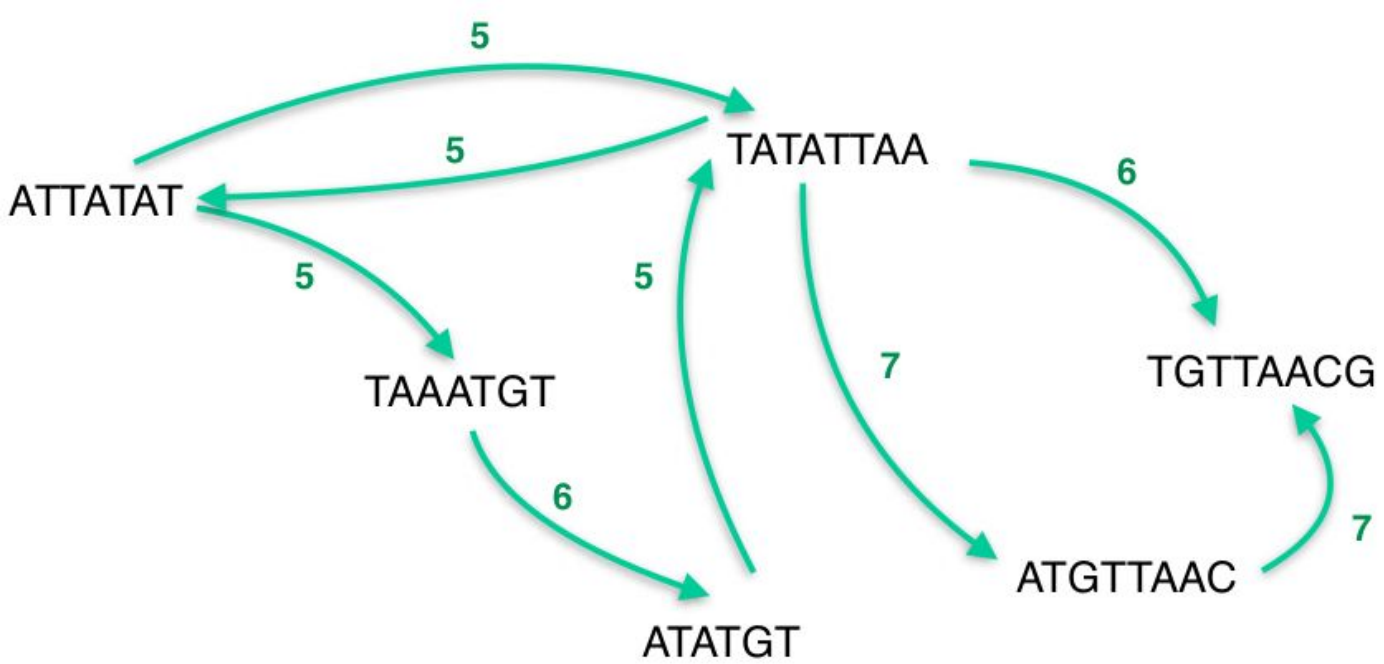
Build a graph representing all the overlaps of a minimum of 5 bases with at most one mismatch between the following reads, then assemble them in the OLC way:

TATATTAA ATTATAT ATGTTAAC

TAAATGT ATATGT TGTTAACG

menti: 7627 3441

Assembly algorithms: Overlap-Layout-Consensus




```
ATTATAT
  TAAATGT
   ATATGT
    TATATTAA
     ATGTTAAC
      TGTTAACG
---------------
ATTATATGTTAACG
```

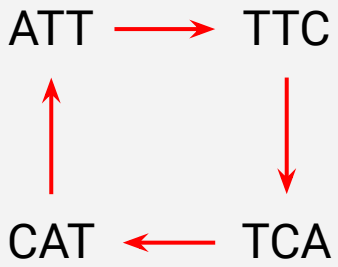
Assembly algorithms: de Bruijn graphs

de Bruijn graphs (DBG): connects words of a *k* length with (*k*-1)-long overlaps

k-mers: *k*-length words in a highly accurate genomic dataset (Illumina, HiFi)

node centric DBG: *k*-mers are nodes, overlaps are edges

edge centric DBG: overlaps are nodes, *k*-mers are edges



Assembly algorithms: de Bruijn graphs

Eulerian path: goes through every edge once

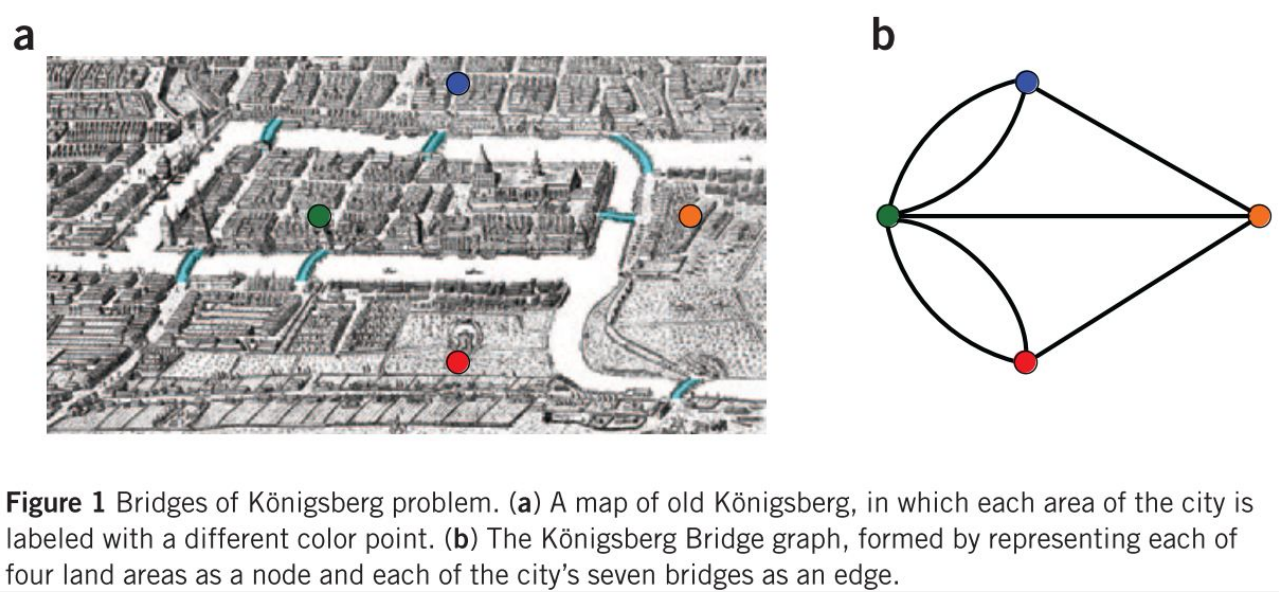


Figure 1 Bridges of Königsberg problem. (**a**) A map of old Königsberg, in which each area of the city is labeled with a different color point. (**b**) The Königsberg Bridge graph, formed by representing each of four land areas as a node and each of the city's seven bridges as an edge.

How to apply de Bruijn graphs to genome assembly, Compeau et *al.*, 2011

Assembly algorithms: de Bruijn graphs

k = 4 ATTATAT CGCGTAC ATTGCGC GCATTAT ACGGCGC

TATATTG GTACGGC GCGTACG ATATTGC

menti: 7627 3441

Assembly algorithms: de Bruijn graphs

k = 4 ATTATAT CGCGTAC ATTGCGC GCATTAT ACGGCGC

TATATTG GTACGGC GCGTACG ATATTGC

ATTA→TTAT→TATA→ATAT

CGCG→GCGT→CGTA→GTAC

ATTG→TTGC→TGCG→GCGC

GCAT→CATT→ATTA→TTAT

ACGG→CGGC→GGCG→GCGC

TATA→ATAT→TATT→ATTG

GTAC→TACG→ACGG→CGGC

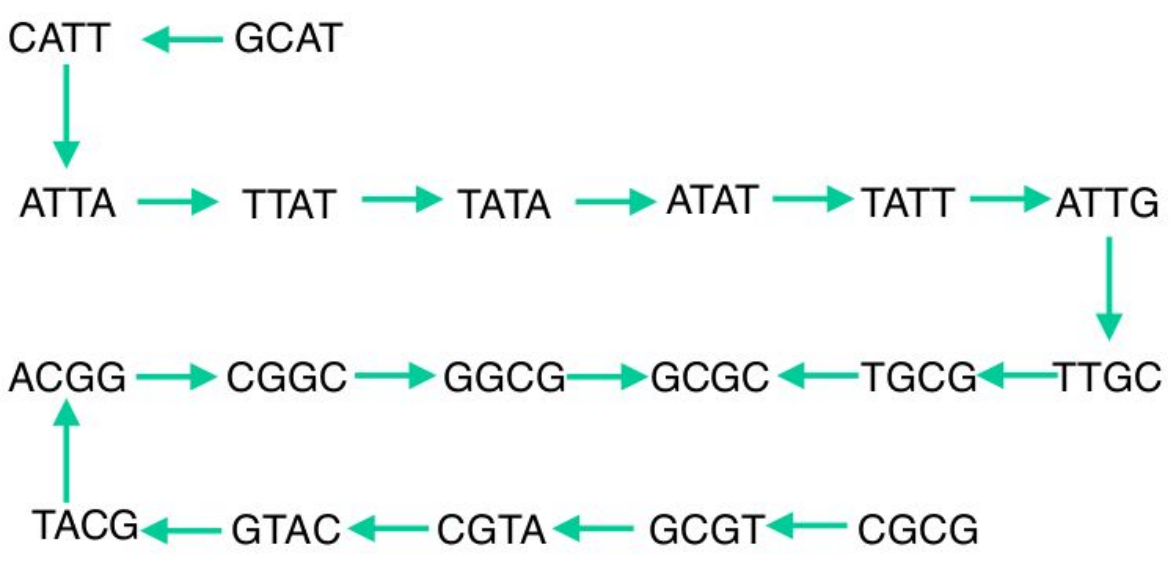
GCGT→CGTA→GTAC→TACG

ATAT→TATT→ATTG→TTGC

Assembly algorithms: de Bruijn graphs

k = 4 ATTATAT CGCGTAC ATTGCGC GCATTAT ACGGCGC

TATATTG GTACGGC GCGTACG ATATTGC



Assembly algorithms: de Bruijn graphs

k = 4 ATTATAT CGCGTAC ATTGCGC GCATTAT ACGGCGC

TATATTG GTACGGC GCGTACG ATATTGC

After compaction:

Two unitigs: GCATTATATTGCGC, CGCGTACGGCGC

Assembly algorithms

DBG assemblers: high accuracy reads → short reads

ABySS, IDBA, SOAPdenovo, SPAdes, VELVET

OLC assemblers:

Sanger reads: Celera

Long reads: Canu, Flye, NextDenovo, Ra, Raven, Shasta, wtdbg2...

Reads pre-processing

Remove adaptors

- **Illumina**: fastqc, cutadapt, Trimmomatic…
- **Nanopore**: Nanopore tools, Porechop

Filtering

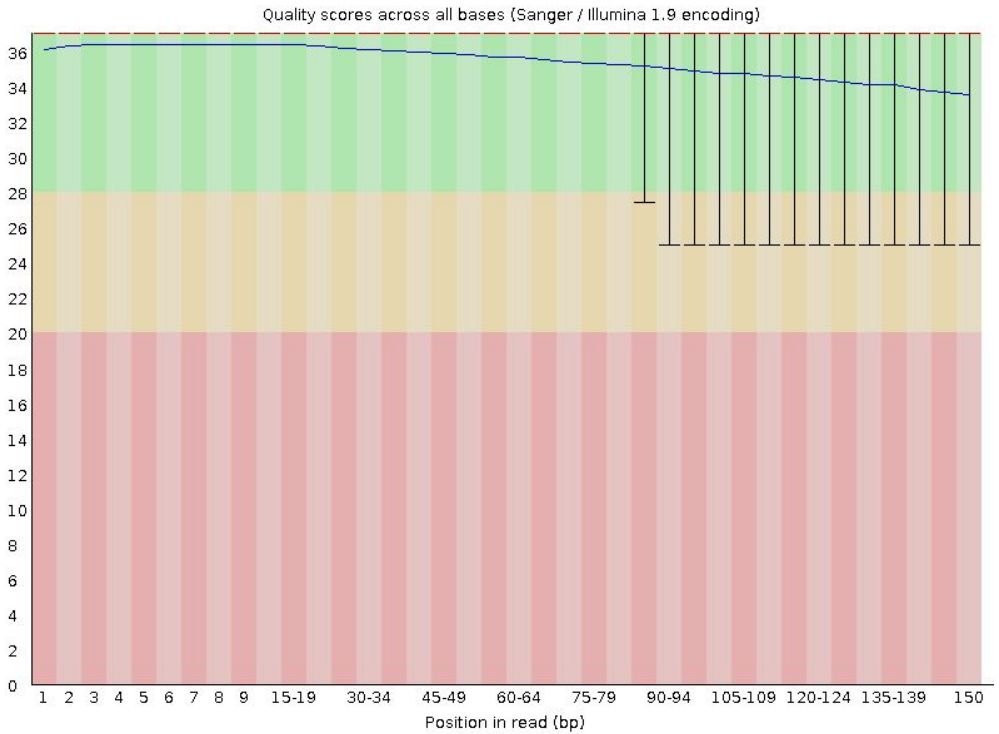
- **Long reads:** Filtlong

Reads pre-processing

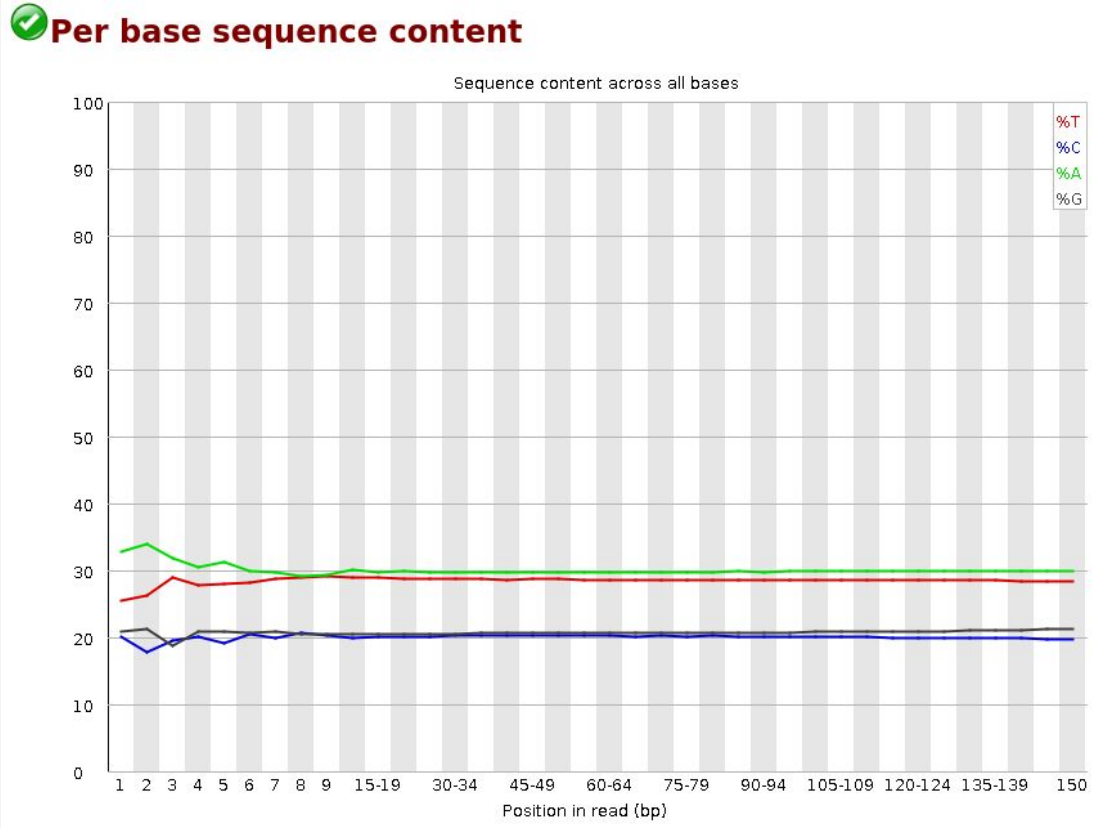
Summary

- Basic Statistics
- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content
- Kmer Content

Per base sequence quality



Reads pre-processing

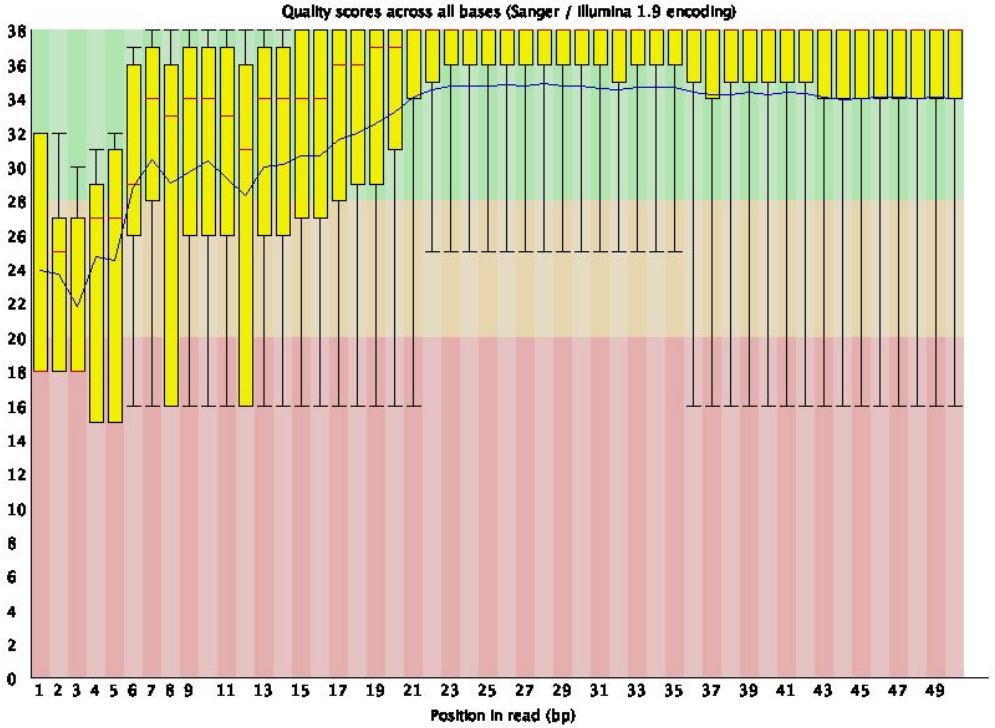


Reads pre-processing

Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Per base sequence quality



Reads pre-processing

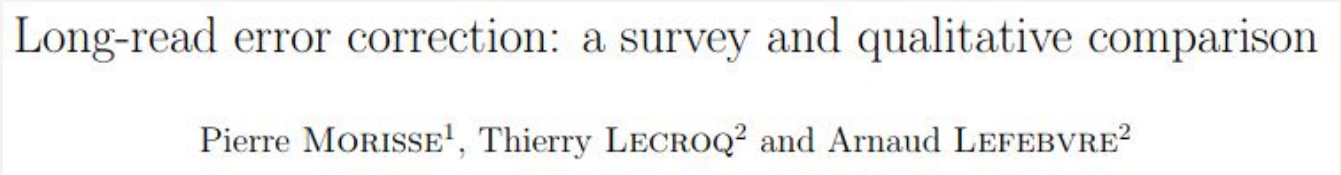
Reads correction: reduce error rate of long reads

- **self correction:** long reads only

  Canu, NextDenovo, Daccord, CONSENT...

- **hybrid correction:** long reads & short reads

  Ratatosk, LoRDEC, CoLoRMAP, proovread...


Long-read error correction: a survey and qualitative comparison

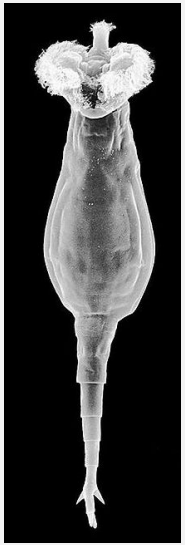
Pierre Morisse[1], Thierry Lecroq[2] and Arnaud Lefebvre[2]

Assembly post-processing

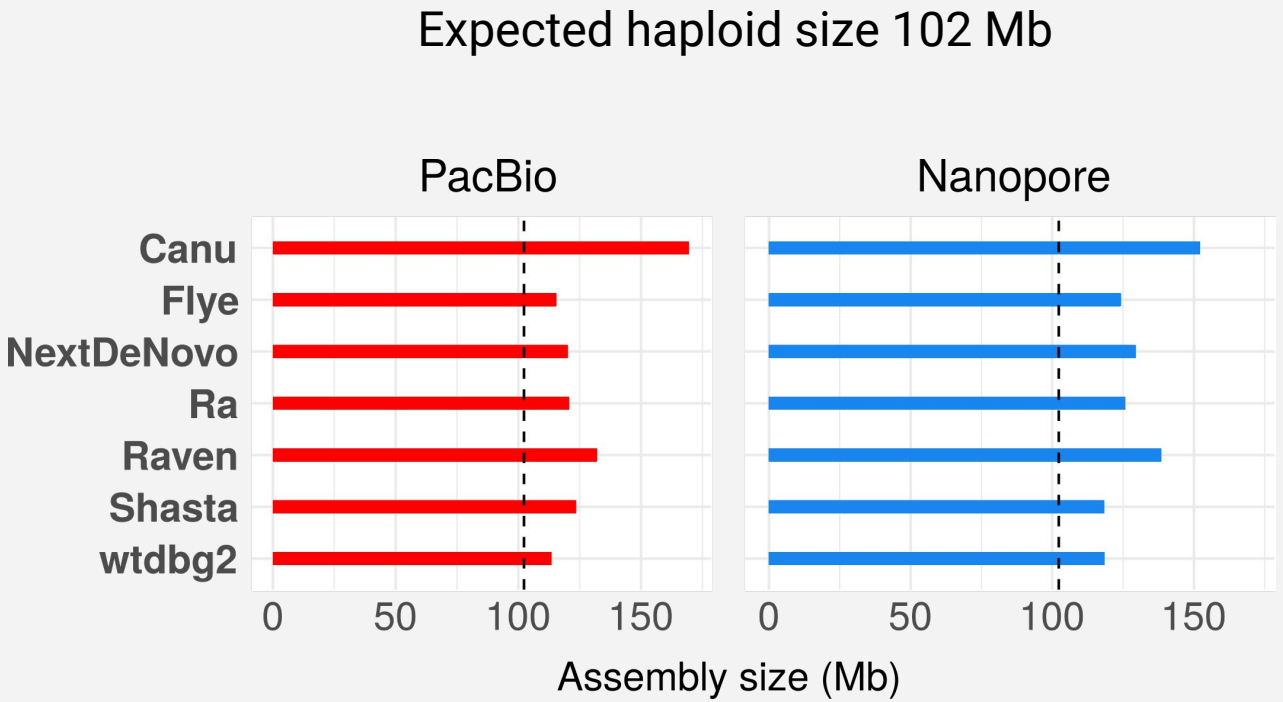
- **Polishing**: reduce errors
- **Haplotig purging**: remove uncollapsed haplotypes
- **Scaffolding**: increase contiguity
- **Gap filling**: find missing sequences

Assembly post-processing: haplotig purging

Adineta vaga



Who Needs Sex (or Males) Anyway?
Liza Gross, PloS Biology, 2007



Assembly post-processing: haplotig purging

Haplotype 1 ATTACCAGTCTCAA**TGGATGGCTACTC**TTTGACGATAGCT

Haplotype 2 ATTACCAGTCTCAA**AGGCTGCTAGTG**TTTGACGATAGCT

Assembly process

Assembly output

Good haploid assemblies

contig 1 ✓

OR

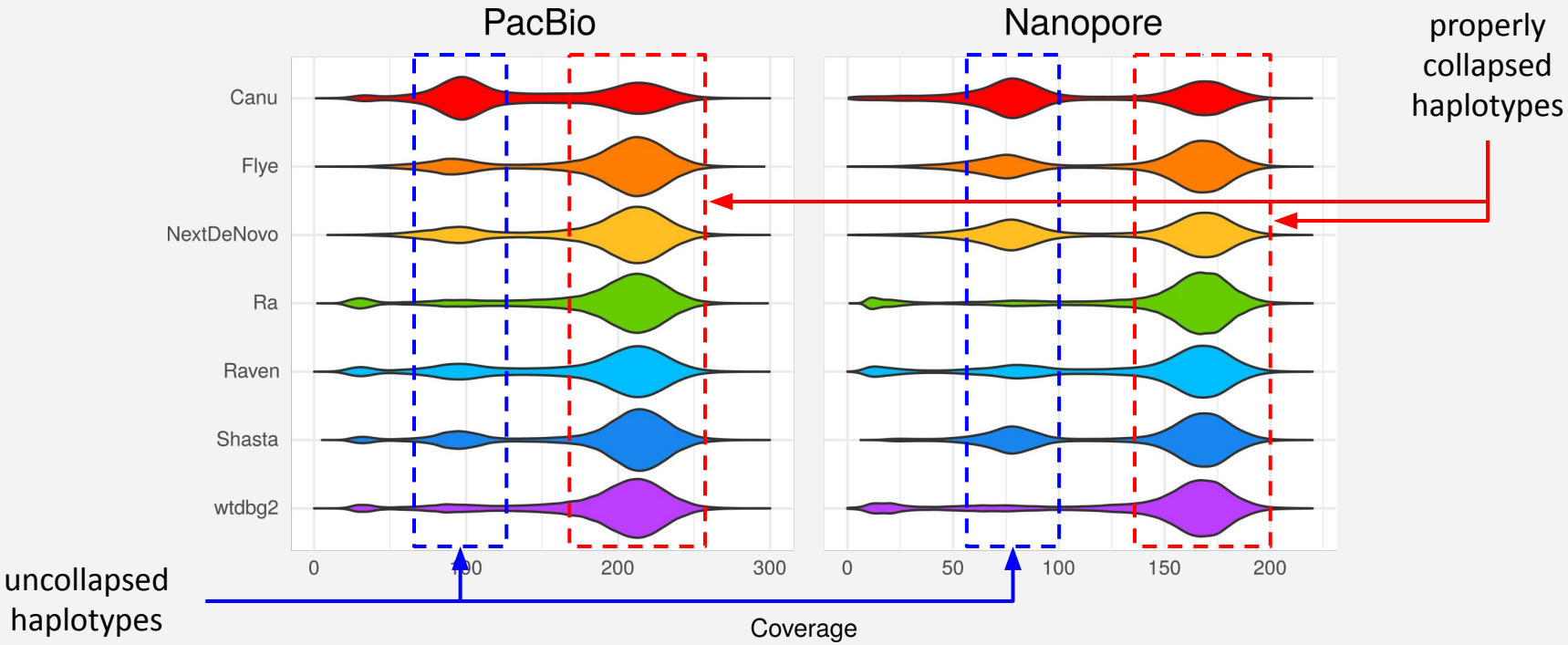
contig 1 ✓

Problematic assembly

contig 1
contig 2
contig 3
contig 4

✗

Assembly post-processing: haplotig purging



Assembly post-processing: haplotig purging

HaploMerger2

HaploMerger2: rebuilding both haploid sub-assemblies from high-heterozygosity diploid genome assembly

Shengfeng Huang*, Mingjing Kang and Anlong Xu

Identifying and removing haplotypic duplication in primary genome assemblies

Dengfeng Guan[1,2], Shane A. McCarthy [2], Jonathan Wood[3], Kerstin Howe [3], Yadong Wang[1,*] and Richard Durbin [2,3,*]

purge_dups

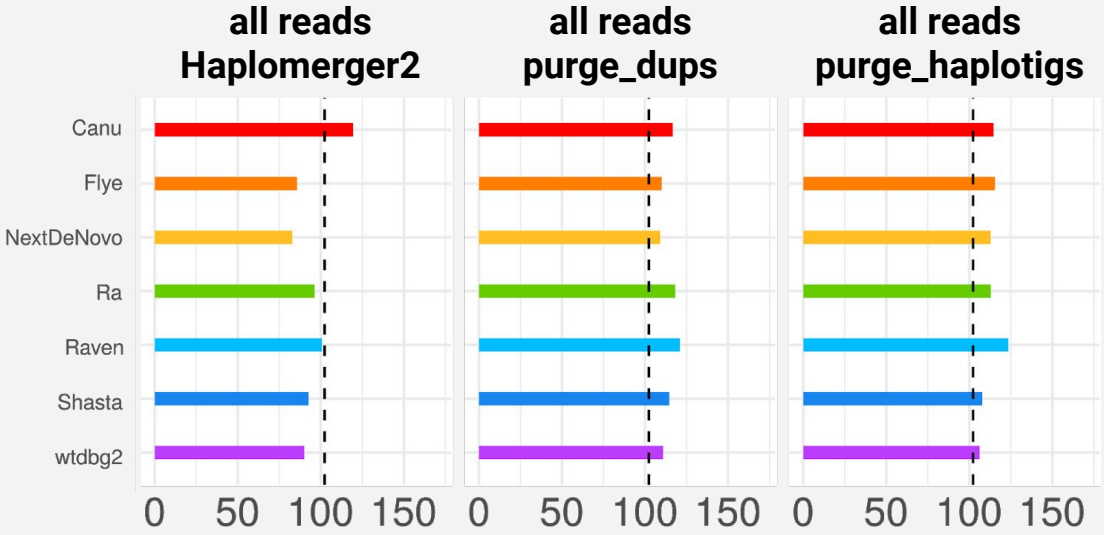
Purge Haplotigs

Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies

Michael J. Roach*, Simon A. Schmidt and Anthony R. Borneman

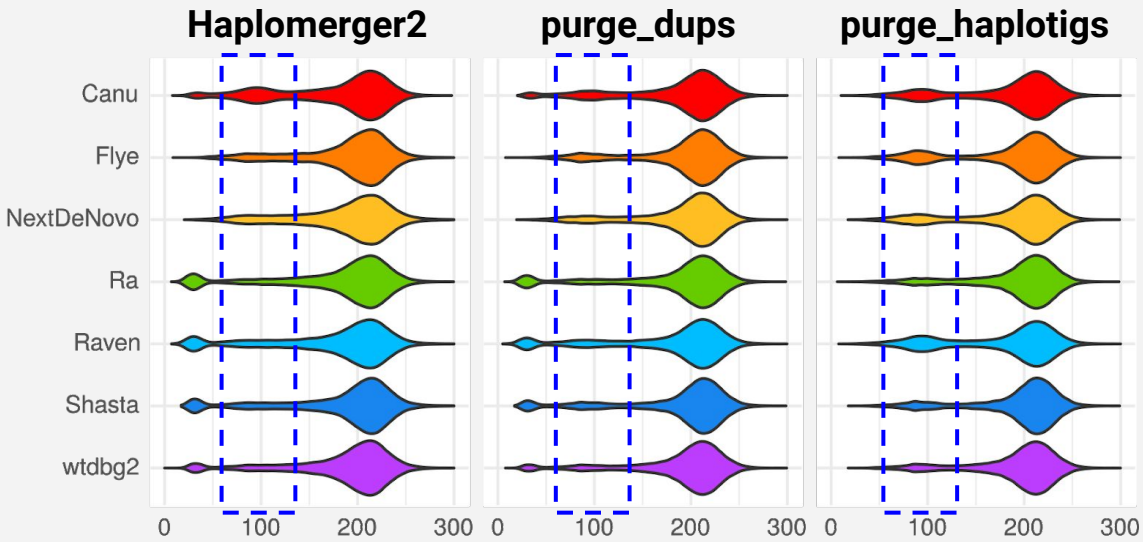
Assembly post-processing: haplotig purging

PacBio assemblies



Assembly post-processing: haplotig purging

PacBio assemblies

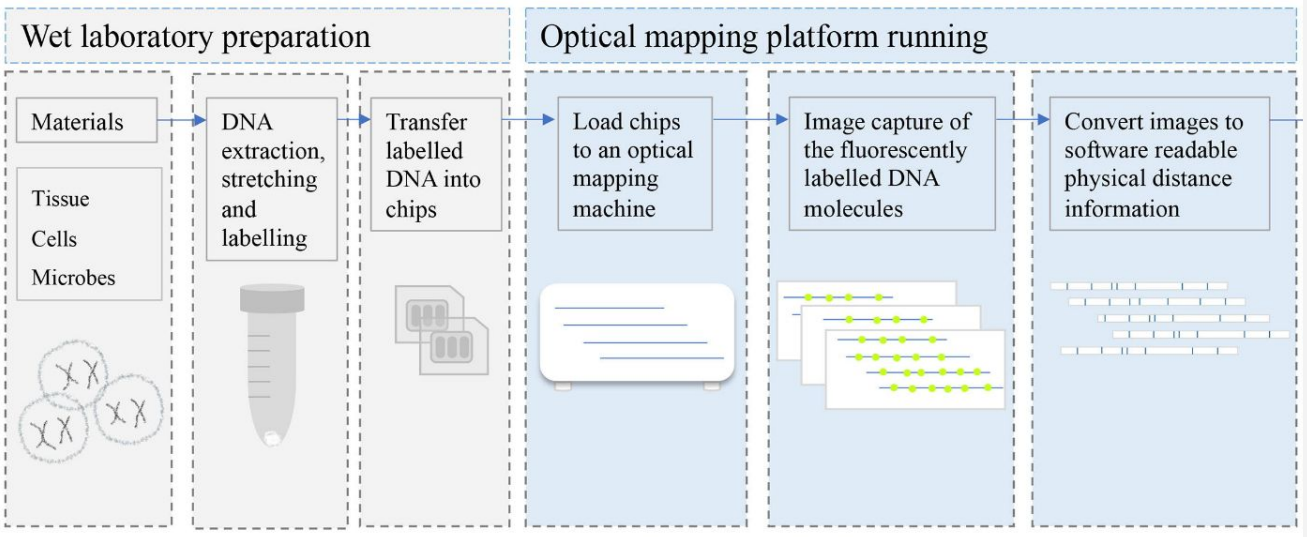


Scaffolding approaches

Scaffolding: grouping and orienting contigs to build chromosome-level scaffolds

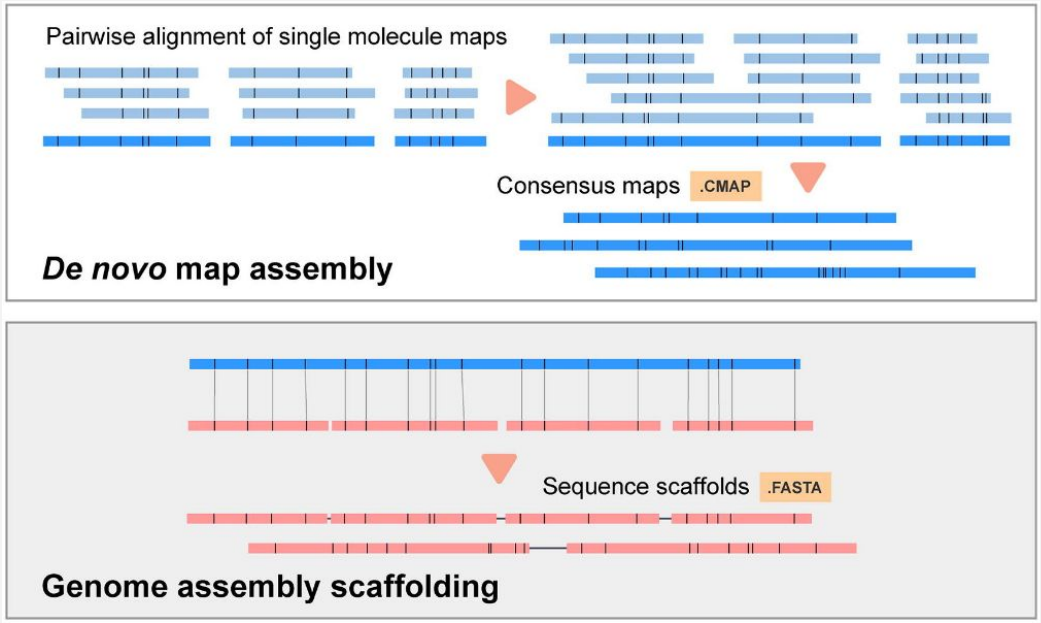
- **Mate-pairs**: short reads with a long insert
- **Long reads**
- **Genetic maps**: ordered markers
- **Optical maps**: ordered markers
- **Linked reads:** barcoded short reads
- **Hi-C/3C/Proximity ligation**

Scaffolding approaches: optical maps



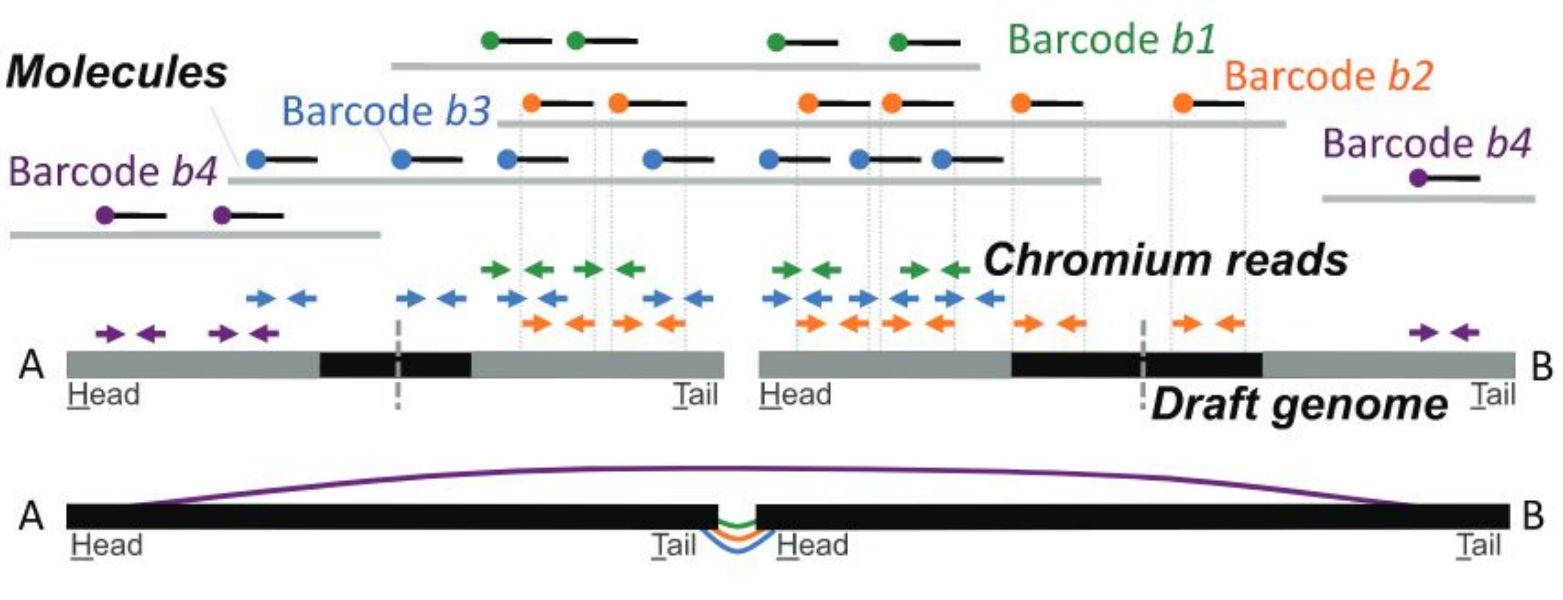
Advances in optical mapping for genomic research, Yuan et *al.*, 2020

Scaffolding approaches: optical maps



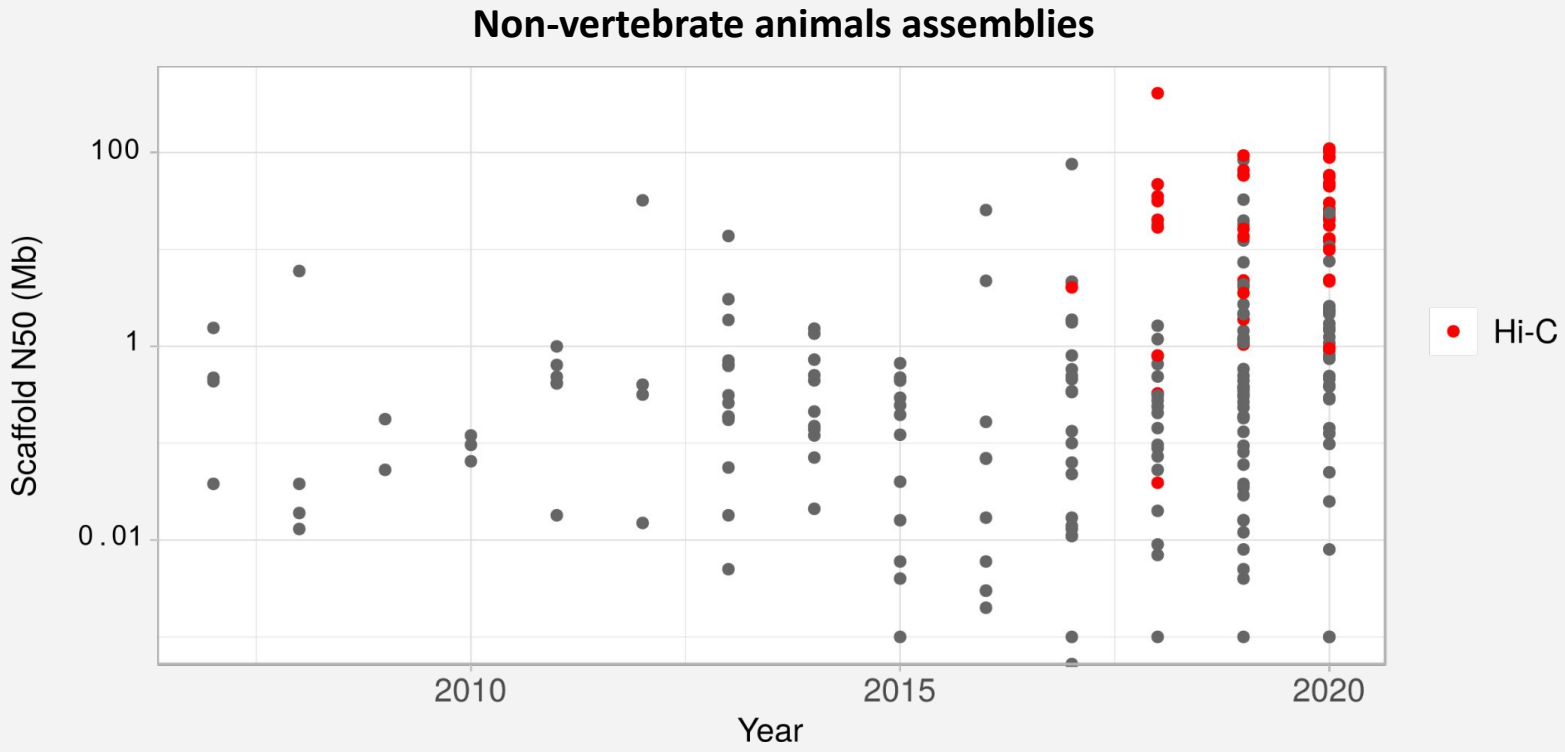
Advances in optical mapping for genomic research, Yuan et *al.*, 2020

Scaffolding approaches: linked reads



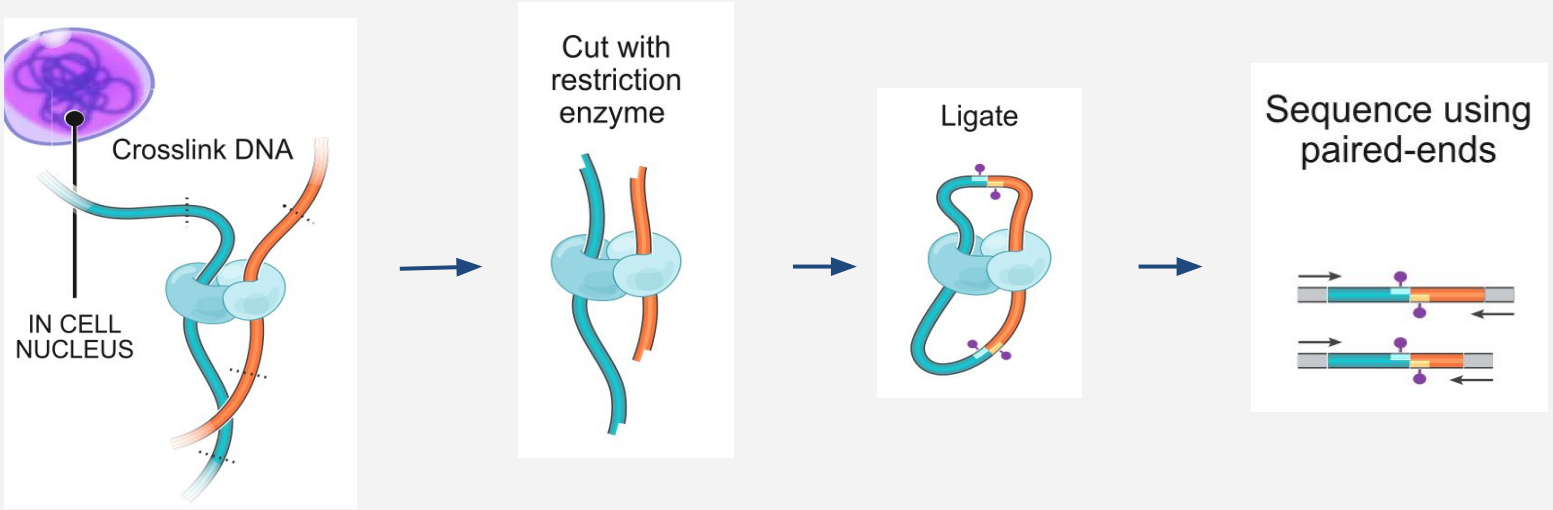
ARCS: scaffolding genome drafts with linked reads, Yeo et *al.*, 2018

Scaffolding approaches: Hi-C scaffolding



Scaffolding approaches: Hi-C scaffolding

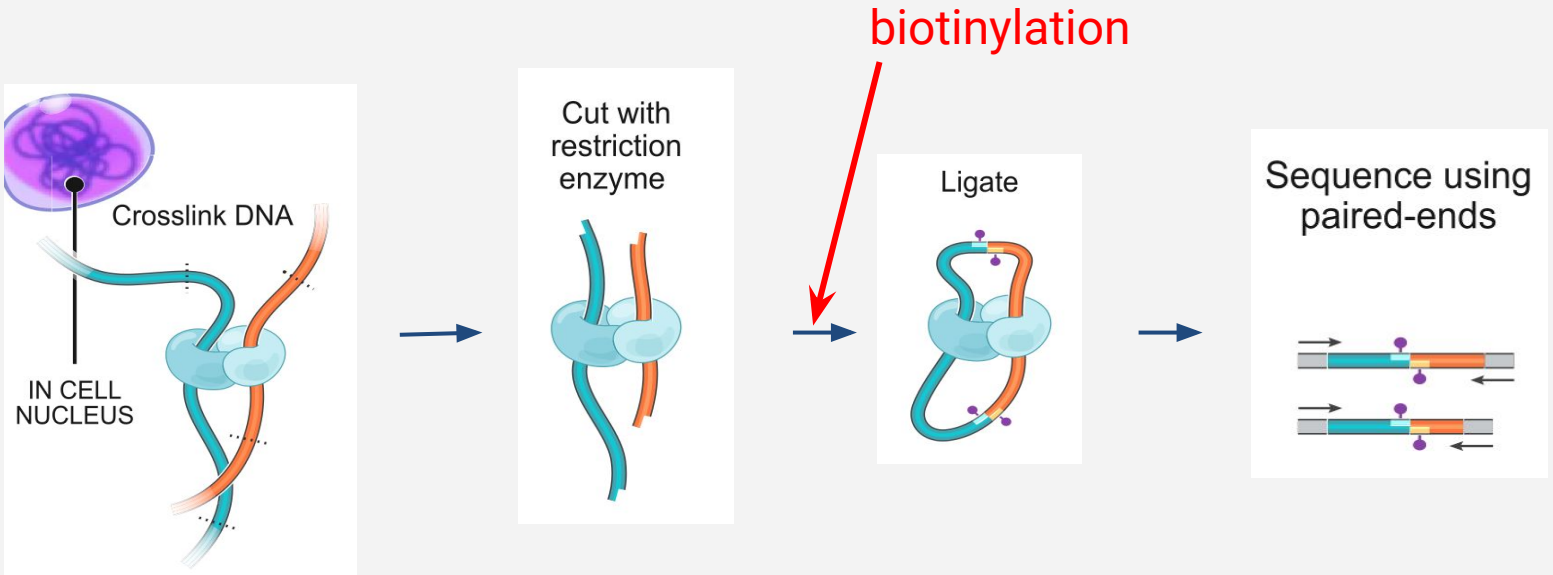
3C



A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Rao et *al.*, 2014

Scaffolding approaches: Hi-C scaffolding

Hi-C



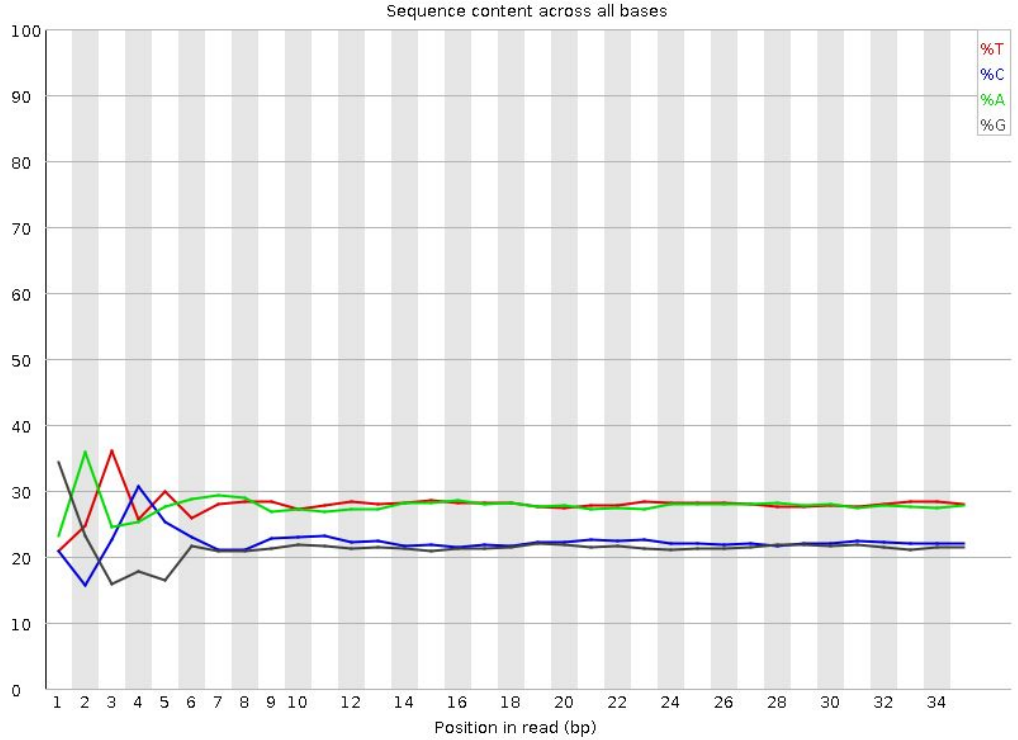
A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Rao et *al.*, 2014

Scaffolding approaches: Hi-C scaffolding

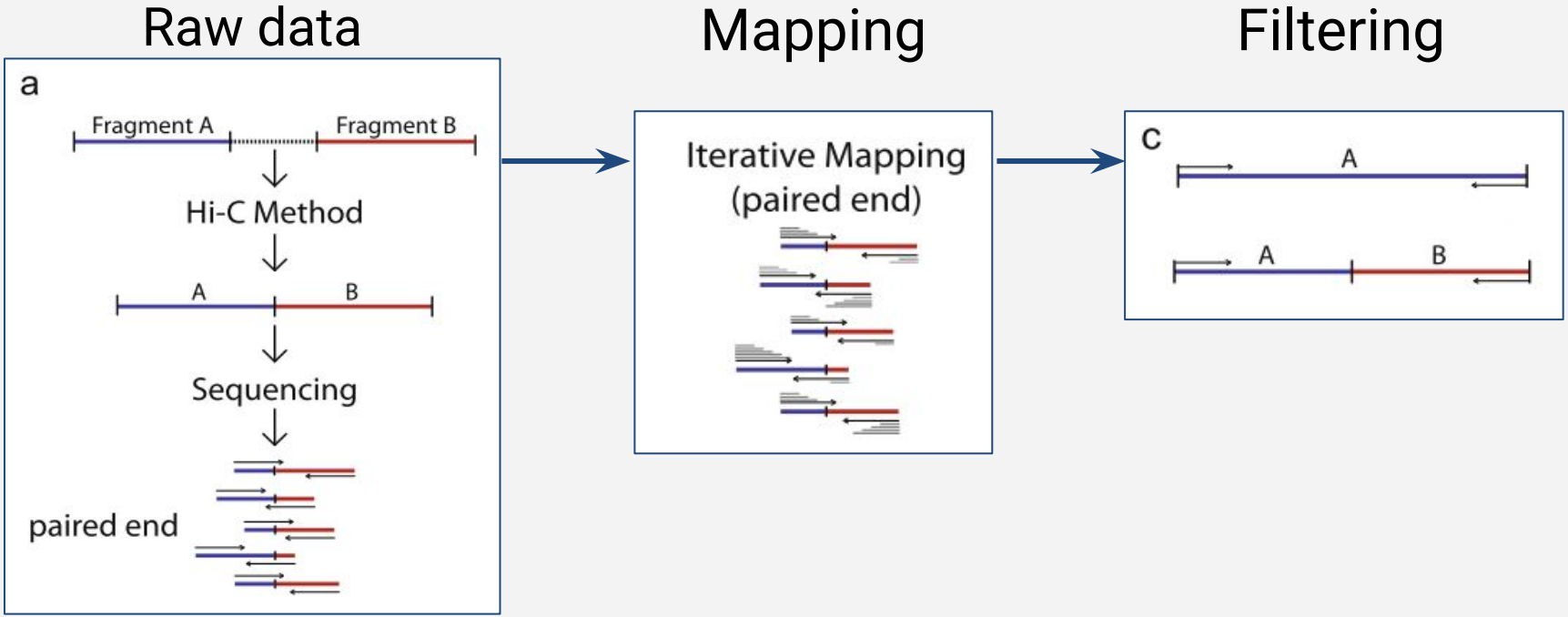
Summary

- Basic Statistics
- Per base sequence quality
- Per tile sequence quality
- Per sequence quality scores
- Per base sequence content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Adapter Content

Per base sequence content

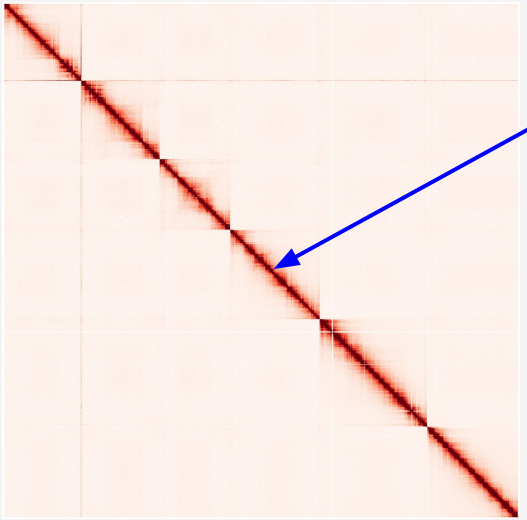


Scaffolding approaches: Hi-C scaffolding



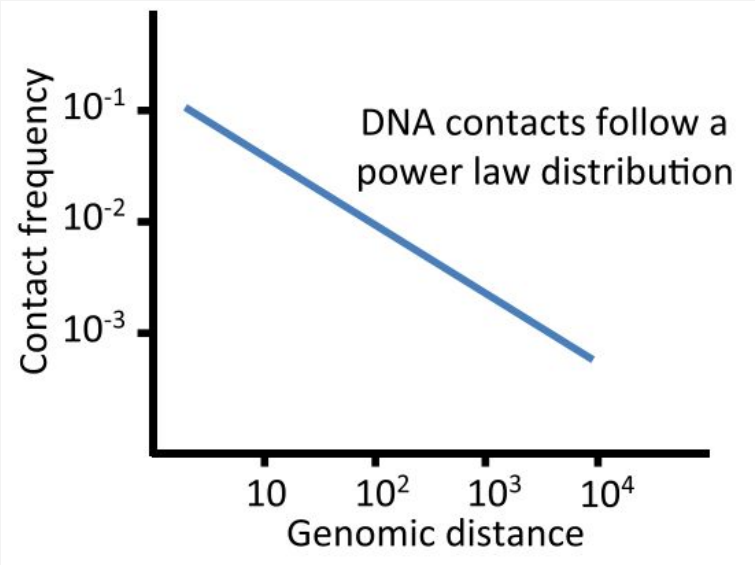
The Hitchhiker's guide to Hi-C analysis: Practical guidelines. Lajoie et *al.*, 2015

Scaffolding approaches: Hi-C scaffolding



Contact map of
Caenorhabditis elegans

contact frequency = f(genomic distance)



Contact genomics: scaffolding and phasing (meta)genomes using chromosome 3D physical signatures. Flot et *al.*, 2015

Scaffolding approaches: Hi-C scaffolding

High-throughput genome scaffolding from *in vivo* DNA interaction frequency

Noam Kaplan & Job Dekker

dnaTri

Lachesis

Chromosome-scale scaffolding of *de novo* genome assemblies based on chromatin interactions

Joshua N Burton, Andrew Adey, Rupali P Patwardhan, Ruolan Qiu, Jacob O Kitzman & Jay Shendure

High-quality genome (re)assembly using chromosomal contact data

Hervé Marie-Nelly, Martial Marbouty, Axel Cournac, Jean-François Flot, Gianni Liti, Dante Poggi Parodi, Sylvie Syan, Nancy Guillén, Antoine Margeot, Christophe Zimmer & Romain Koszul

GRAAL

Scaffolding approaches: Hi-C scaffolding

De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds

Olga Dudchenko[1,2,3,4], Sanjit S. Batra[1,2,3,*], Arina D. Omer[1,2,3,*], Sarah K. Nyquist[1,3], Marie Hoeger[1,3], Neva C. Durand[1,...]

3D-DNA

SALSA2

Integrating Hi-C links with assembly graphs for chromosome-scale assembly

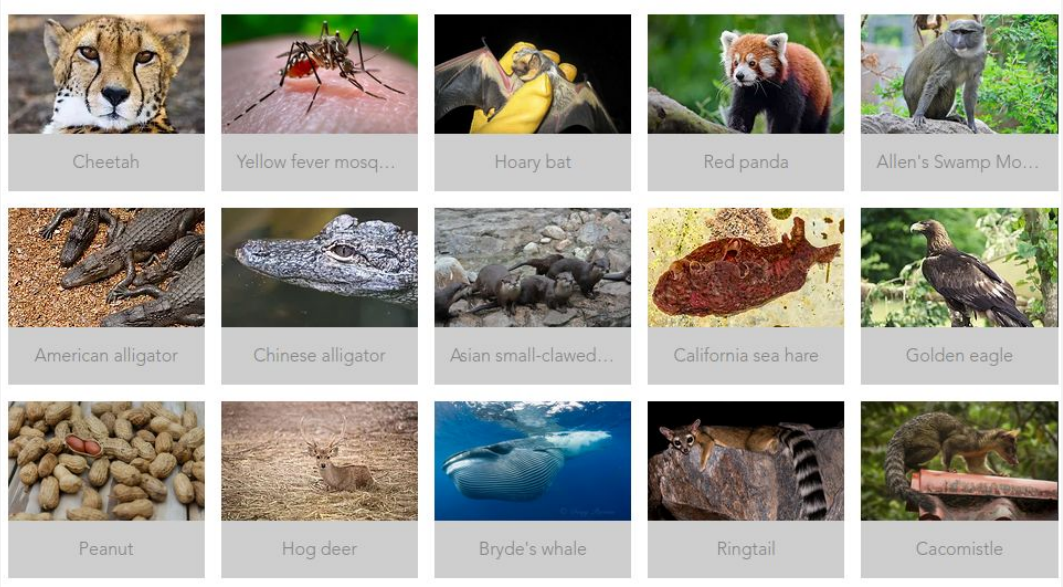
Jay Ghurye, Arang Rhie, Brian P. Walenz, Anthony Schmitt, Siddarth Selvaraj, Mihai Pop, Adam M. Phillippy, Sergey Koren

instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder

Lyam Baudry, Nadège Guiglielmoni, Hervé Marie-Nelly, Alexandre Cormier, Martial Marbouty, Komlan Avia, Yann Loe Mie, Olivier Godfroy, Lieven Sterck, J. Mark Cock, Christophe Zimmer, Susana M. Coelho & Romain Koszul

instaGRAAL

Scaffolding approaches: Hi-C scaffolding



www.dnazoo.org

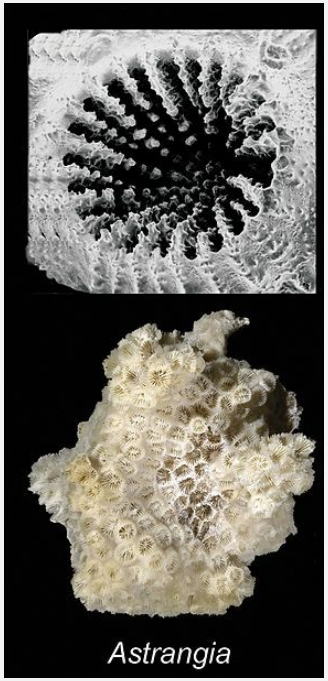
Scaffolding approaches: Hi-C scaffolding

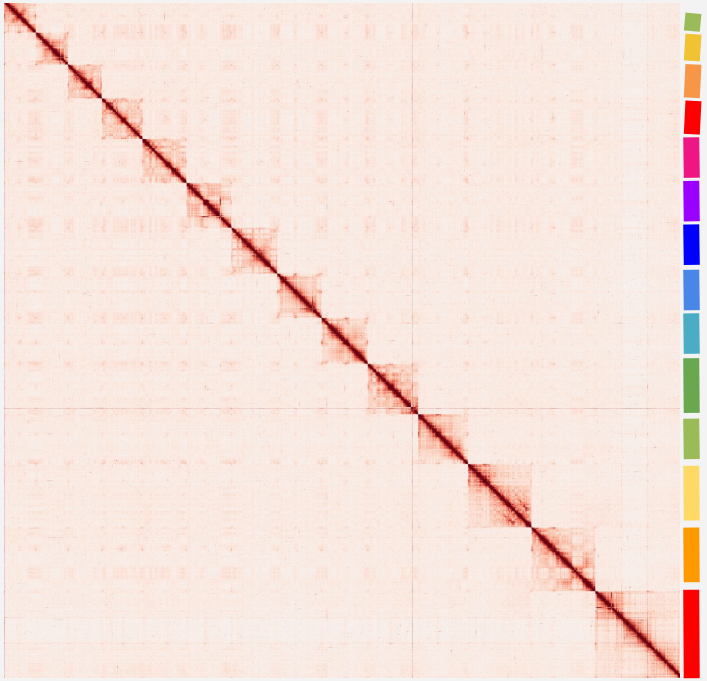
Astrangia poculata

(coral)

14 scaffolds

455 Mb



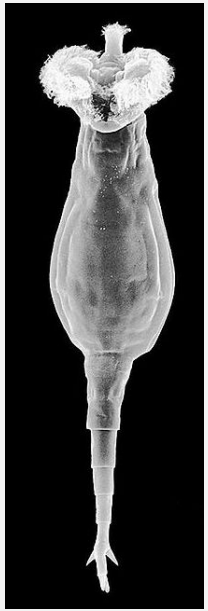


Hi-C contact map of *Astrangia poculata*

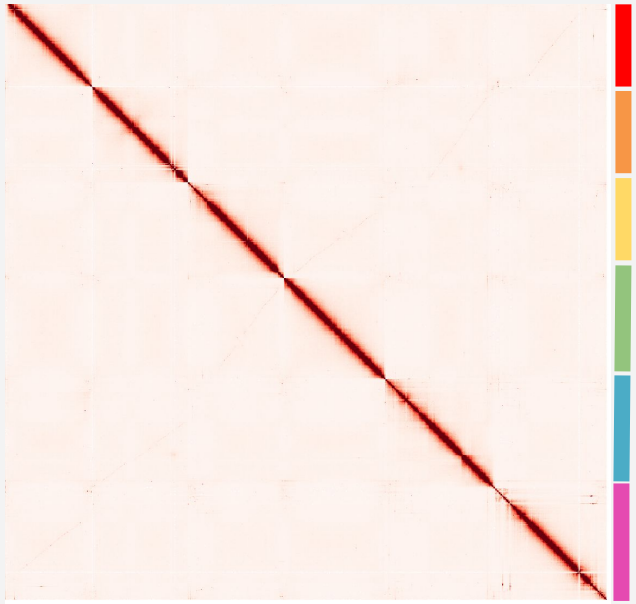
Scaffolding approaches: Hi-C scaffolding

Adineta vaga (rotifer)

6 scaffolds



Who Needs Sex (or Males) Anyway?
Liza Gross, PloS Biology, 2007



Hi-C contact map of *Adineta vaga*

Scaffolding approaches: Hi-C scaffolding

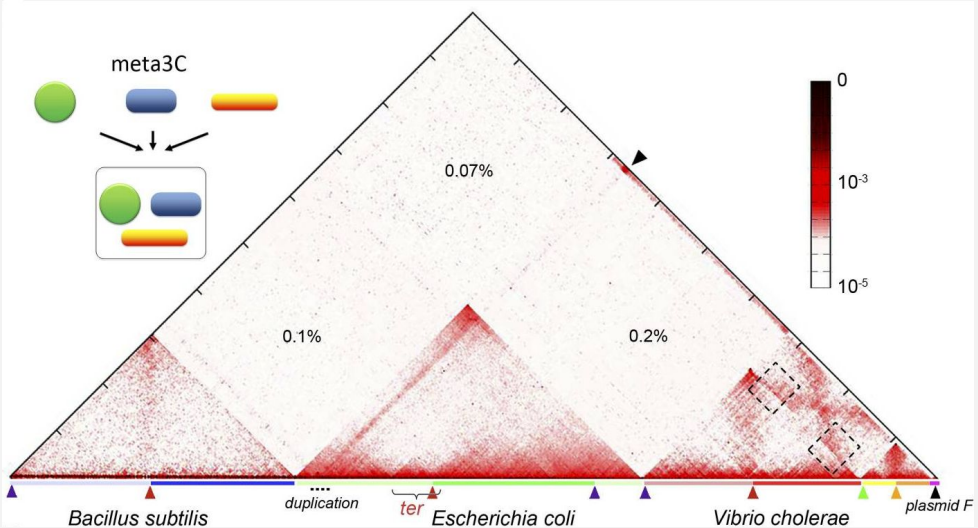
“What coverage should I get?”

➔ Arima recommends 200 millions pairs per Gb

Species	**Size**	**# fragments**	**# Hi-C pairs**	**Hi-C mapping**
Adineta vaga	101 Mb	30	55 millions	83%
Astrangia poculata	455 Mb	2995	723 millions	67%
Flaccisagitta enflata	929 Mb	6612	489 millions	37%
Mercenaria mercenaria	1.86 Gb	5118	455 millions	55%

Scaffolding approaches: Hi-C scaffolding

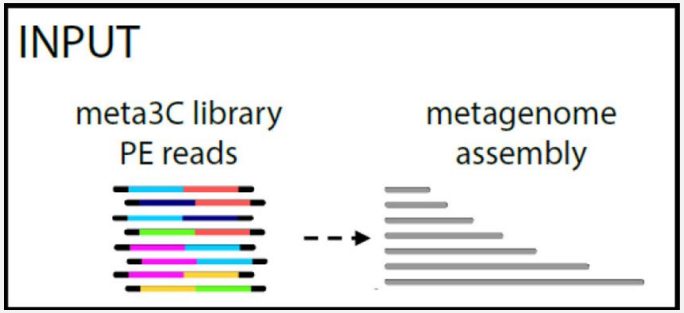
Mixture of organisms, but different nuclei = no 3D contacts

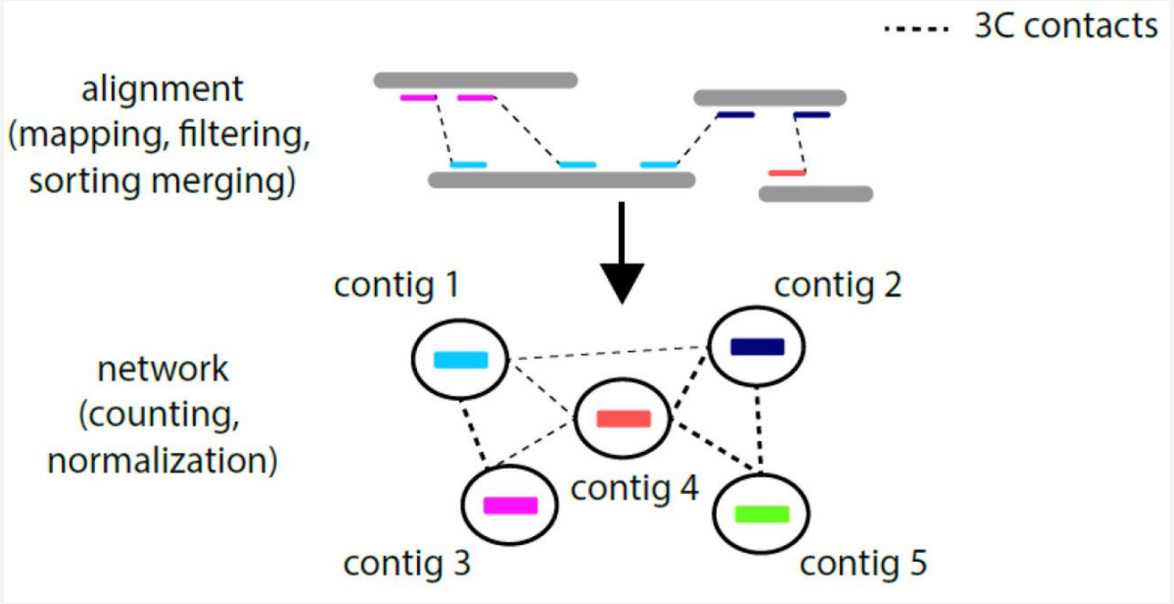


Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms, Marbouty et *al.*, 2014

Scaffolding approaches: Hi-C scaffolding

Metator

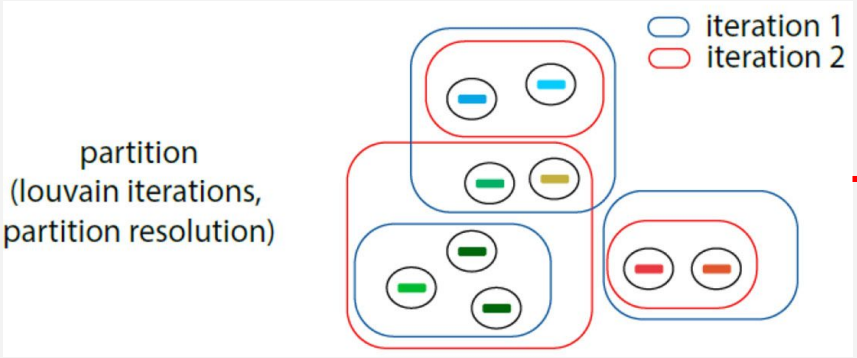


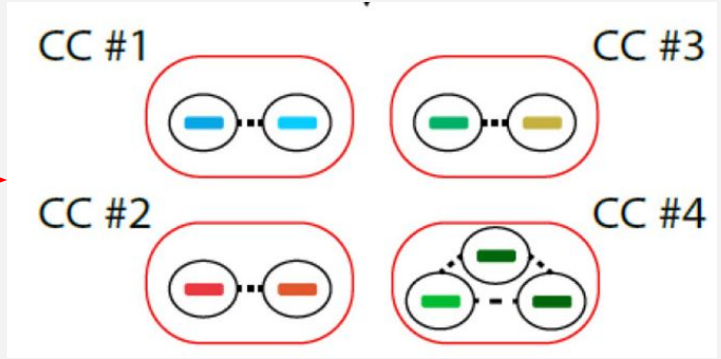


align

Scaffolding approaches: Hi-C scaffolding

Metator





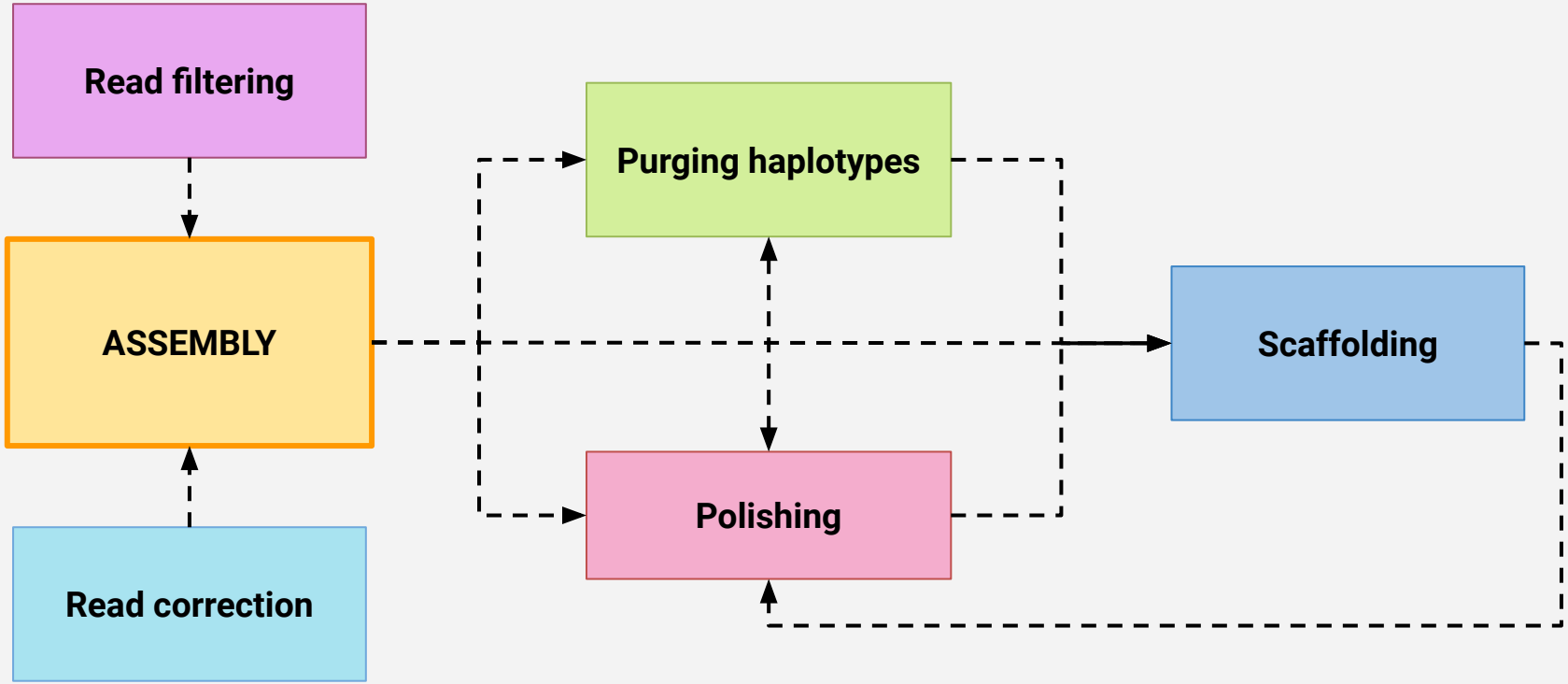
partition **binning**

Scaffolding approaches: gap filling

GATTCCGGAGACCTANNNNNNNNNNNNNNNNNNNNNNNNNATATTTGTCAGAC

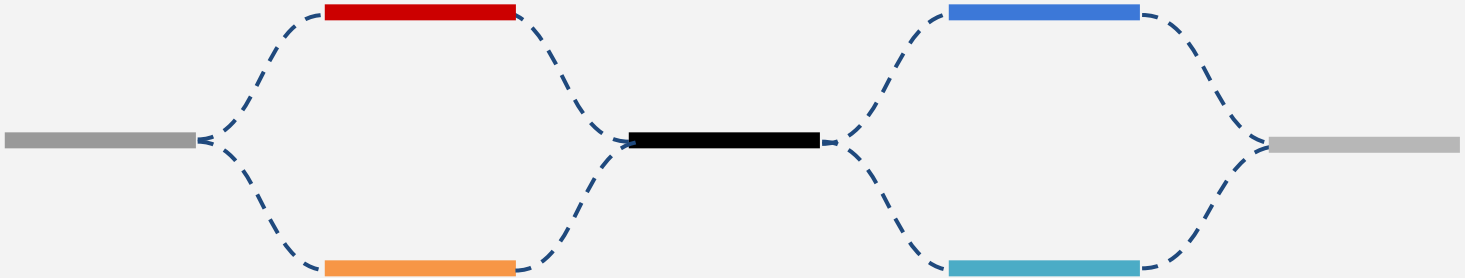
- Short reads: GapFiller, GAPPadder, Sealer
- Long reads: FGAP, GMCloser, LR_Gapcloser, PBJelly, PGcloser, TGS-GapCloser

Assembly pipeline

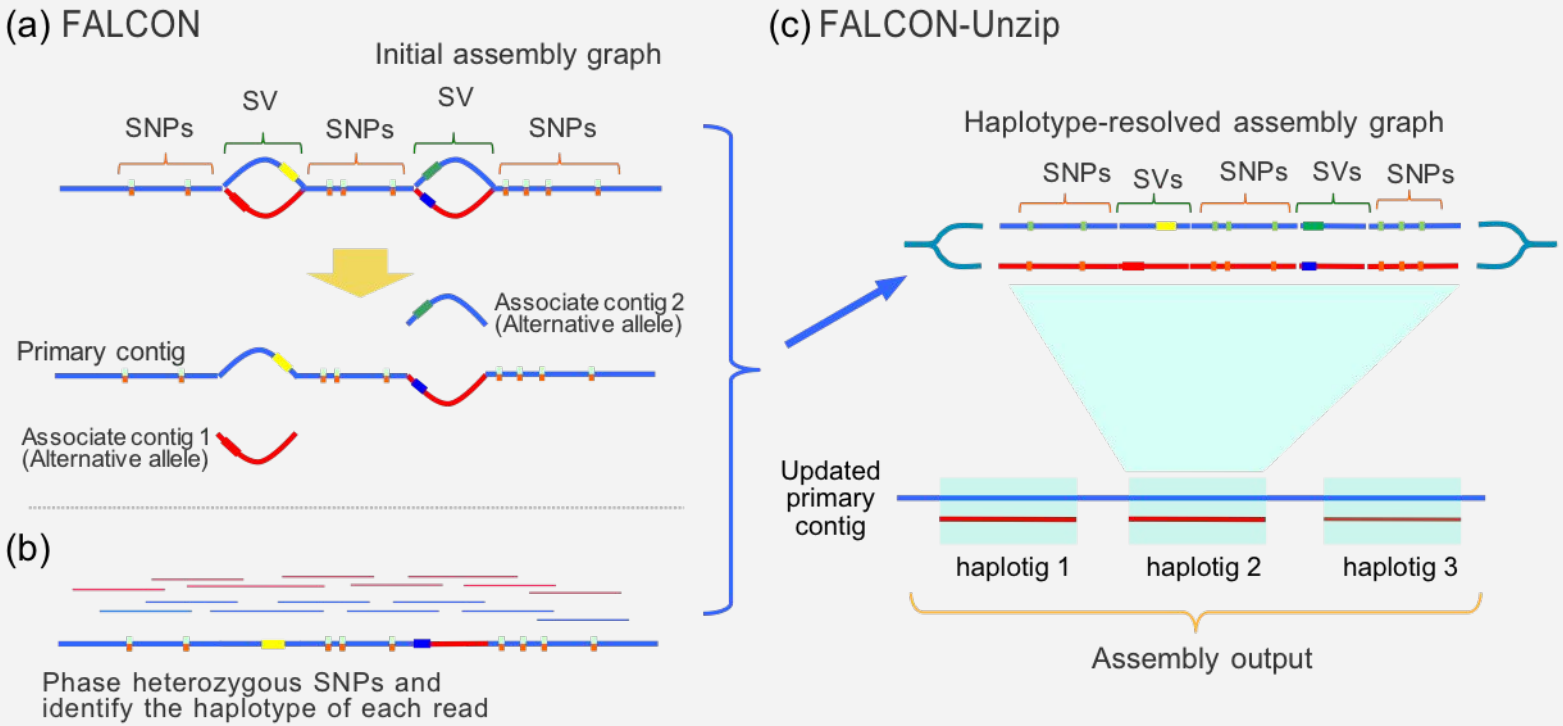


Phasing assemblies

Diploid genome
= 2 haplotypes



Phasing assemblies

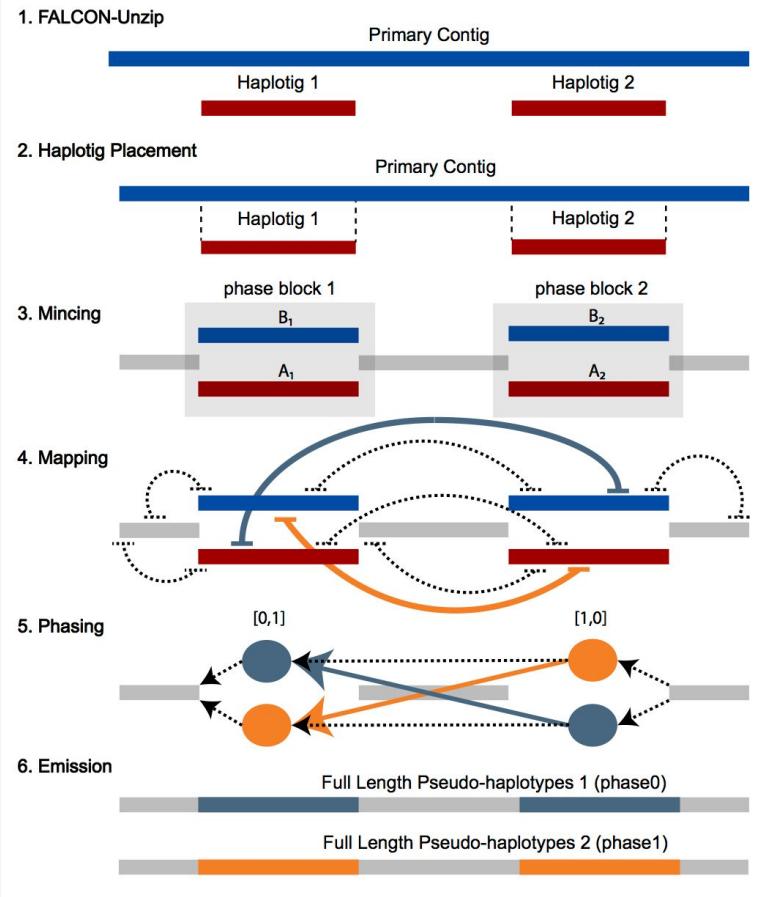


Phased diploid genome assembly with single-molecule real-time sequencing, Chin et *al.*, 2016

Phasing assemblies

FALCON-Phase

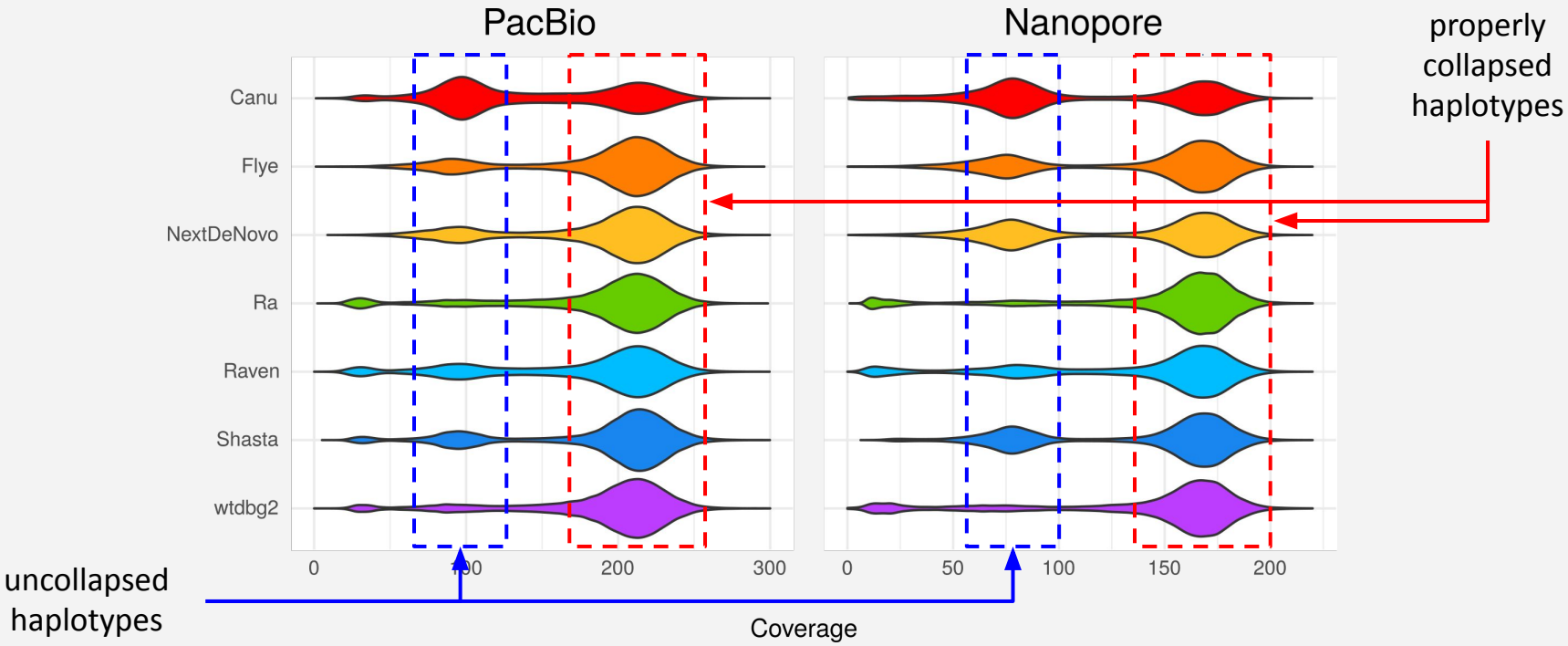
Extended haplotype phasing of *de novo* genome assemblies with FALCON-Phase, Kronenberg et al., 2019



Assembly evaluation

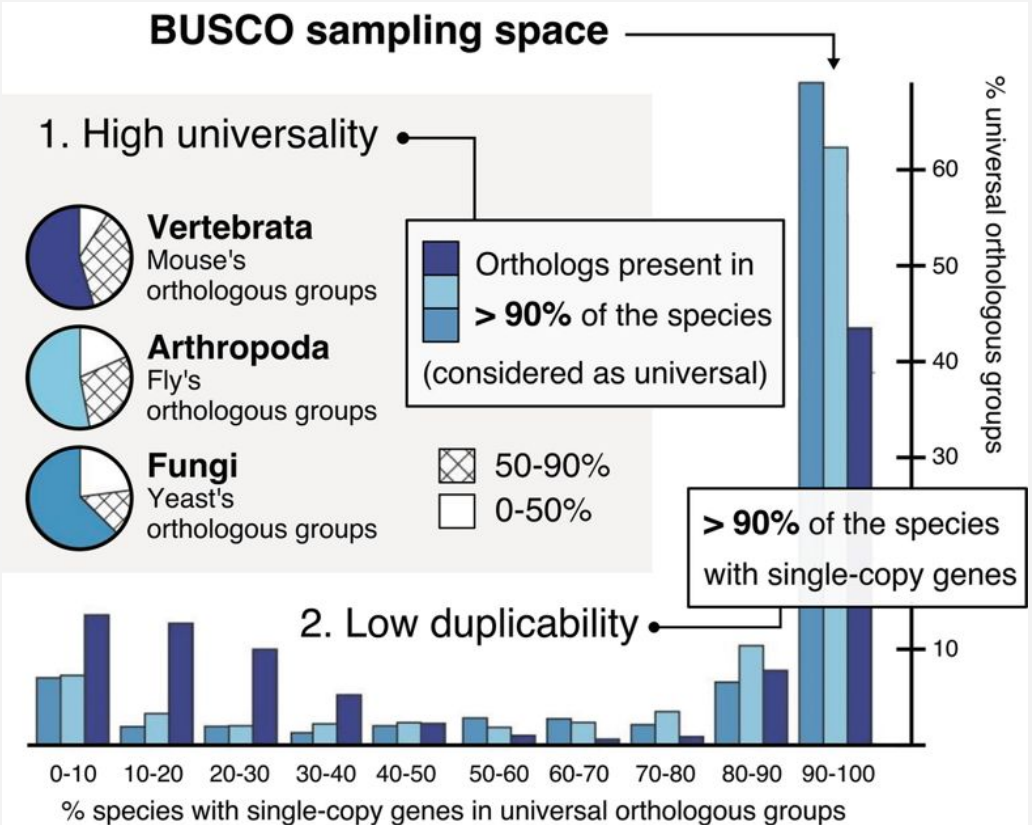
- **Continuity**: N50, NG50
- **Correctness**: mapping reads
- **Completeness**: BUSCO, KAT
- **Structure**: contact map

Assembly evaluation



Assembly evaluation

BUSCO



Assembly evaluation

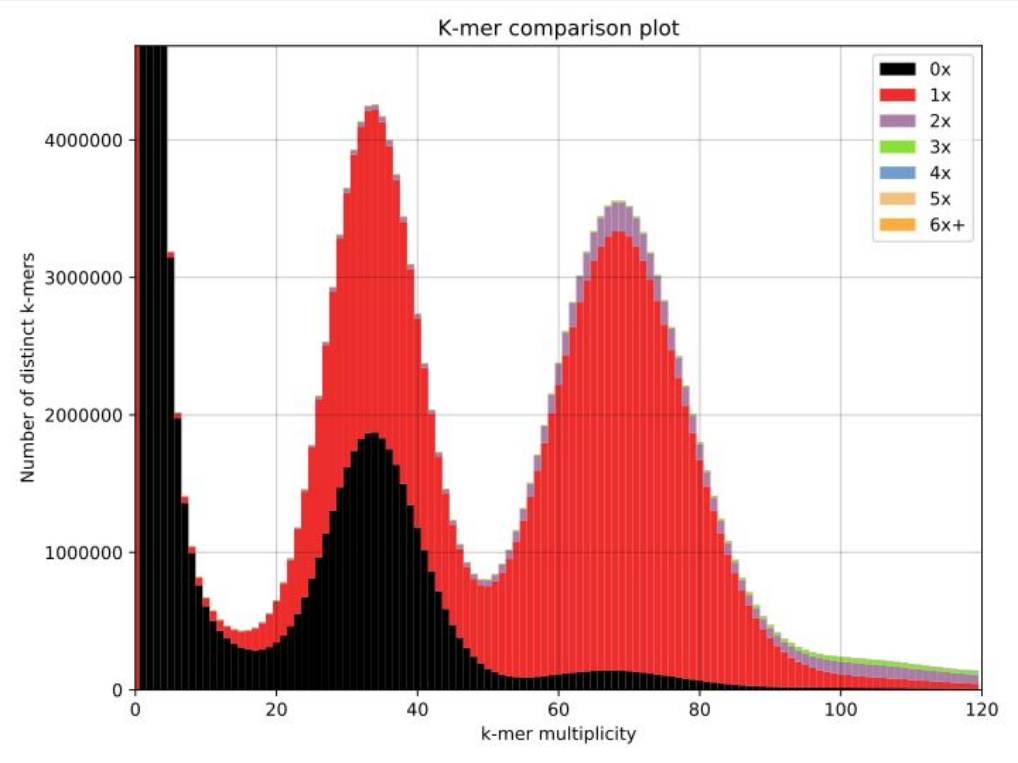
BUSCO: assemblies of *Adineta vaga*

Reads	Assembly	Complete features	Complete single-copy	Complete duplicated	Fragmented	Missing
Illumina	BWISE	82.6%	16.4%	66.2%	3.2%	14.2%
PacBio	NextDenovo	84.1%	72.0%	12.1%	2.8%	13.1%
Nanopore	NextDenovo	44.7%	39.6%	5.1%	21.1%	34.2%
Nanopore + Illumina	NextDenovo	86.2%	72.0%	14.2%	1.9%	11.9%
HiFi	hifiasm	85.8%	18.7%	67.1%	1.8%	12.4%

Assembly evaluation

KAT

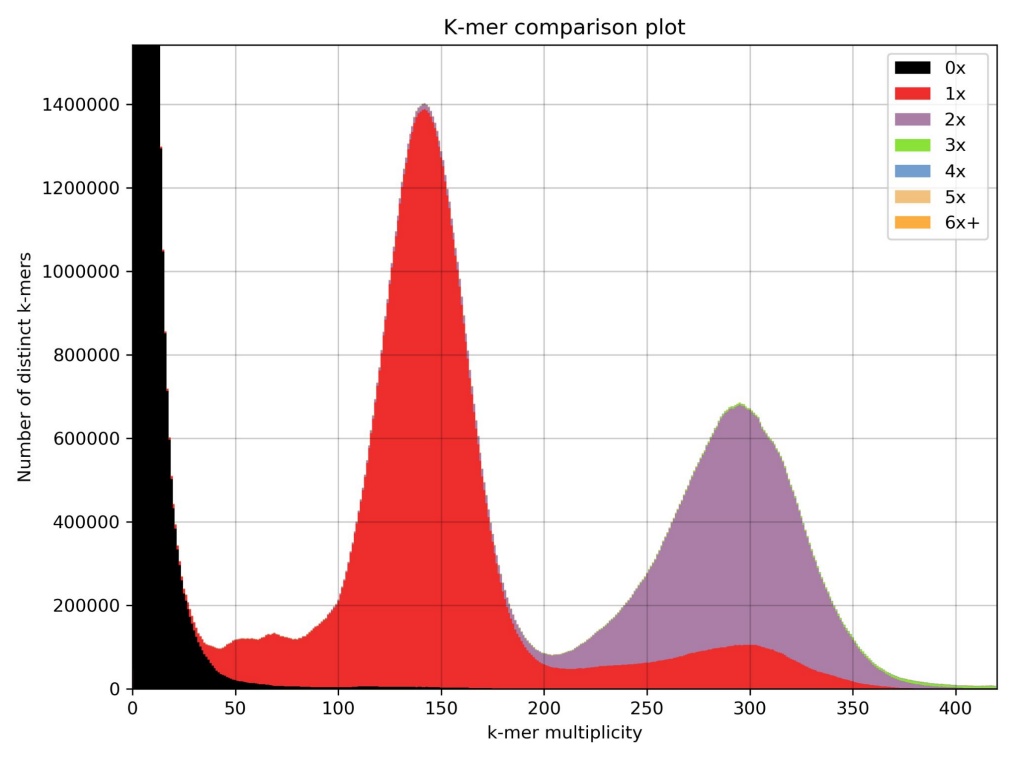
Xenia sp.



Assembly evaluation

KAT

Adineta vaga

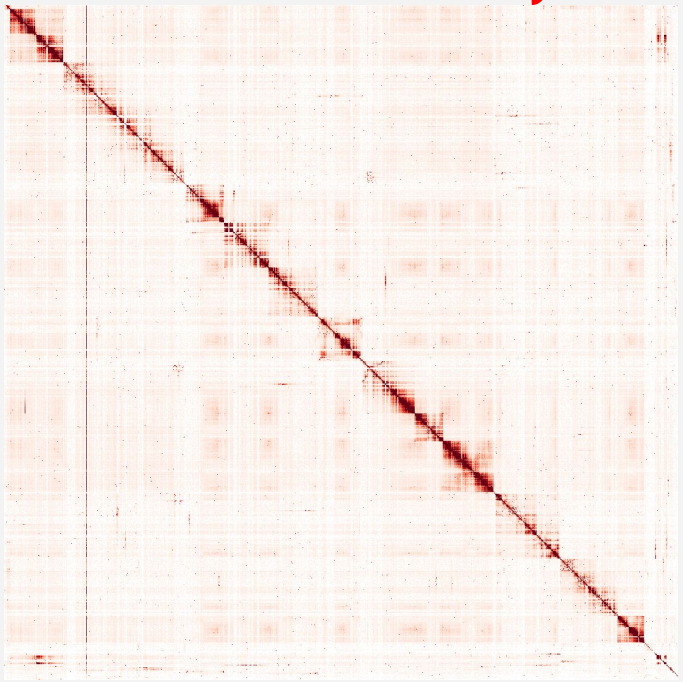


Assembly evaluation

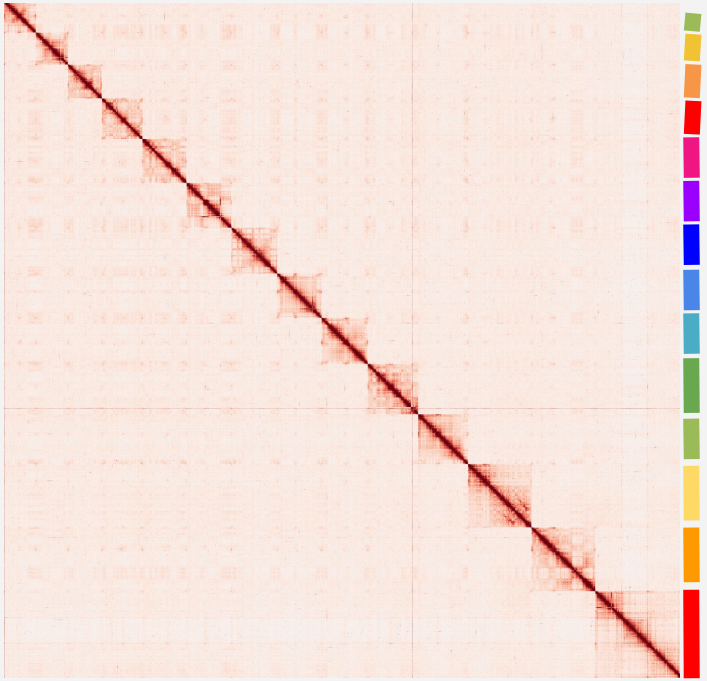
Lineage dynamics of the endosymbiotic cell type in the soft coral *Xenia*

Minjie Hu, Xiaobin Zheng, Chen-Ming Fan & Yixian Zheng

Nature **582**, 534–538(2020) | Cite this article



Hi-C contact map of *Xenia sp.*



Hi-C contact map of *Astrangia poculata*

Thank you for your attention!
Questions?

https://github.com/nadegeguiglielmoni/presentations