Genome assembly post-processing

Nadège Guiglielmoni

Read pre-processing

- **Adapter trimming**

  especially when PCR amplification is involved: Lima

- **Read filtering**

  to select the longest/highest quality reads: Filtlong

- **Read correction**

  self correction: long reads only

  hybrid correction: long reads & short reads

Assembly post-processing

- **Polishing**

  reduce errors using high-accuracy reads: HyPo, Pilon, Hapo-G

- **Haplotig purging**

  remove uncollapsed haplotypes: purge_dups, Purge Haplotigs

- **Scaffolding**

  increase contiguity

  <u>using long reads:</u> LINKS

  <u>using Hi-C:</u> YaHS, instaGRAAL

- **Gap filling**

  find missing sequences: TGS-GapCloser

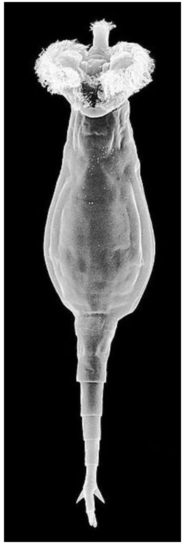
Haplotig purging

Goal: obtain a collapsed assembly, where each set of chromosomes is represented by a single sequence

≠ phasing: sequencing data from multiple individuals, limited sequencing data…

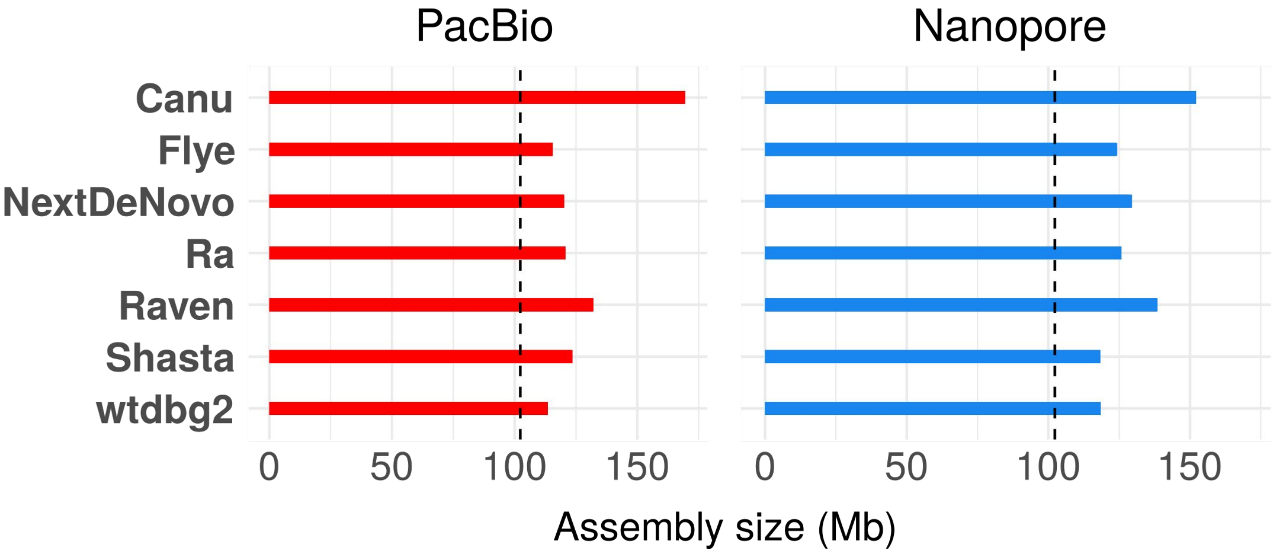
Haplotig purging

Adineta vaga

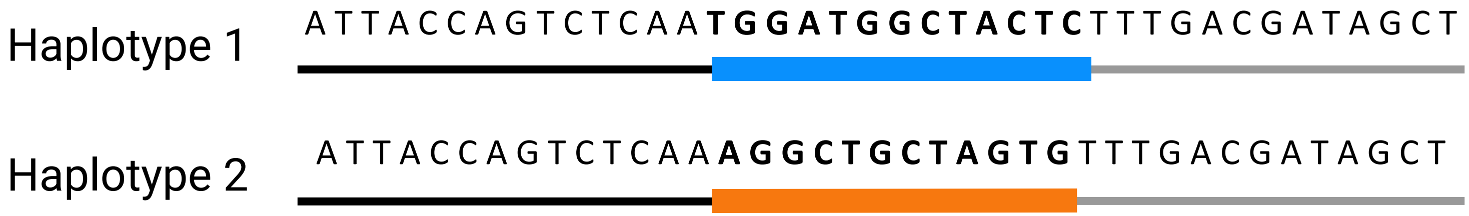


Who Needs Sex (or Males) Anyway?
Liza Gross, PloS Biology, 2007

Expected haploid size 102 Mb



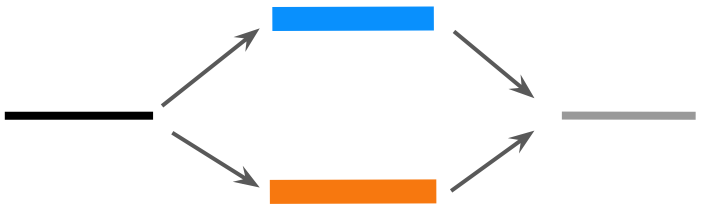
Haplotig purging



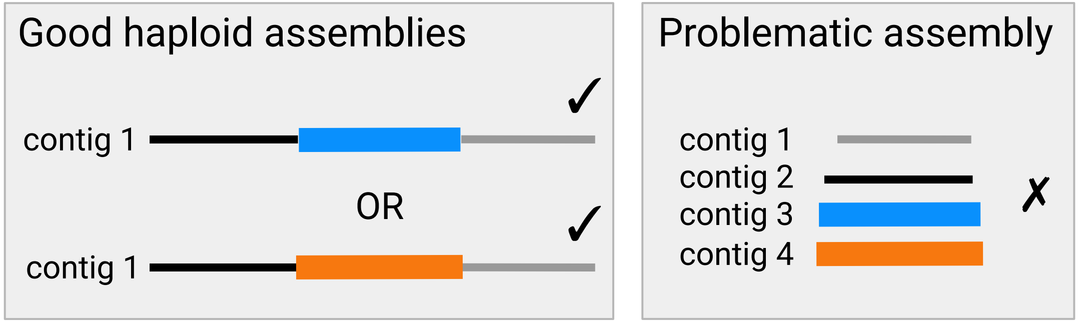
Haplotype 1 ATTACCAGTCTCAA**TGGATGGCTACTC**TTTGACGATAGCT

Haplotype 2 ATTACCAGTCTCAA**AGGCTGCTAGTG**TTTGACGATAGCT

Assembly process



Assembly output



Good haploid assemblies

contig 1 ✓

OR

contig 1 ✓

Problematic assembly

contig 1

contig 2

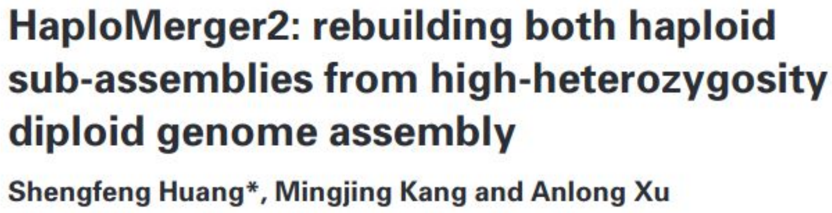
contig 3

contig 4

✗

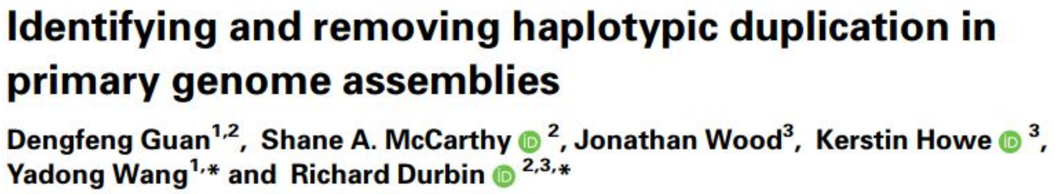
Haplotig purging

HaploMerger2



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

Shengfeng Huang*, Mingjing Kang and Anlong Xu

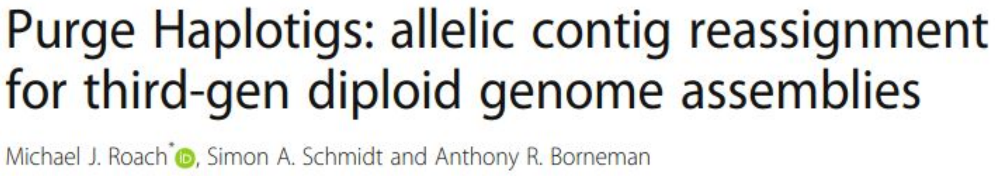


Identifying and removing haplotypic duplication in primary genome assemblies

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

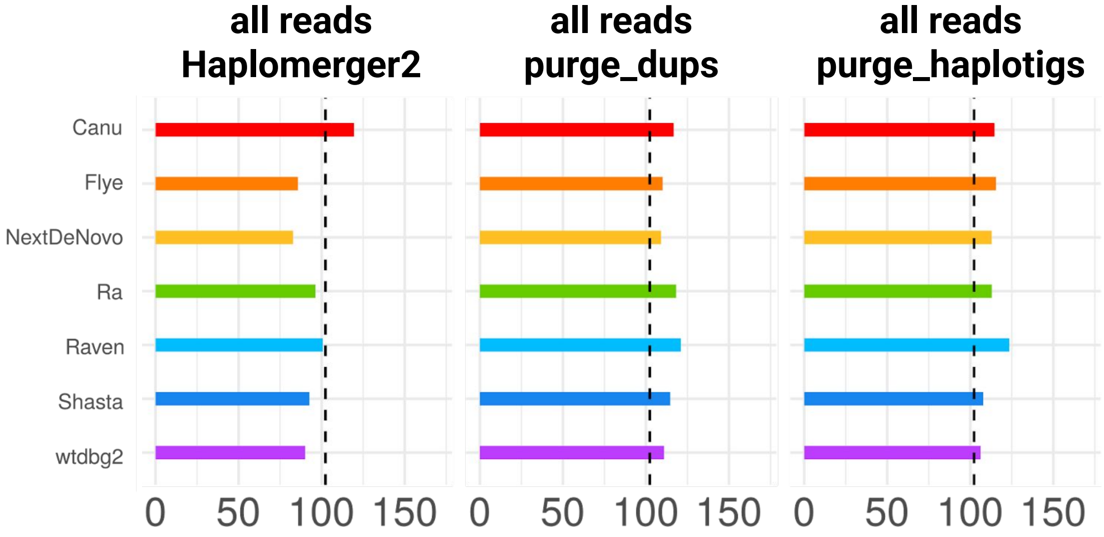


Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies

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

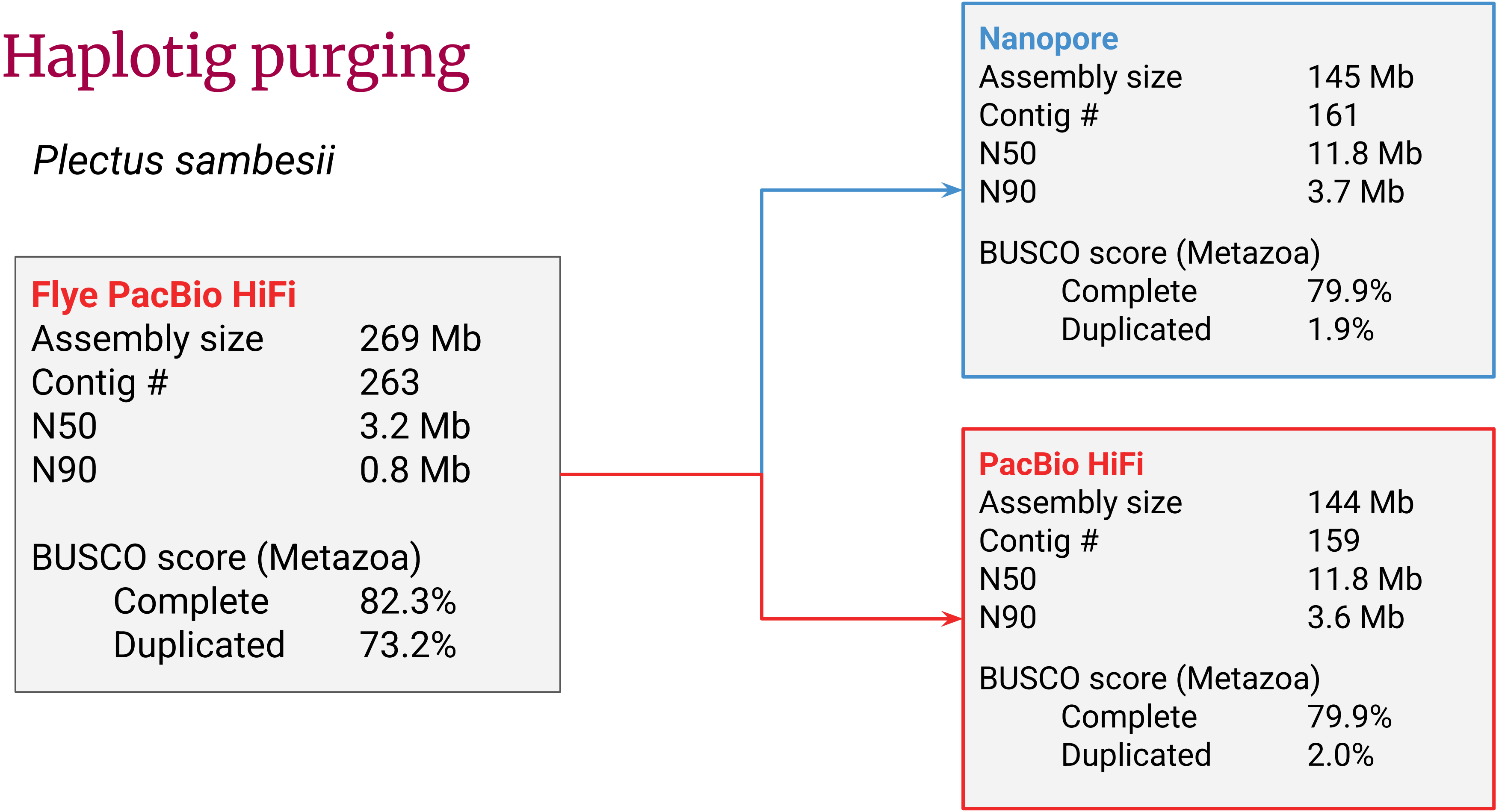
Haplotig purging

PacBio assemblies



Haplotig purging

Plectus sambesii



Flye PacBio HiFi

Assembly size	269 Mb
Contig #	263
N50	3.2 Mb
N90	0.8 Mb
BUSCO score (Metazoa)	
Complete	82.3%
Duplicated	73.2%

Nanopore

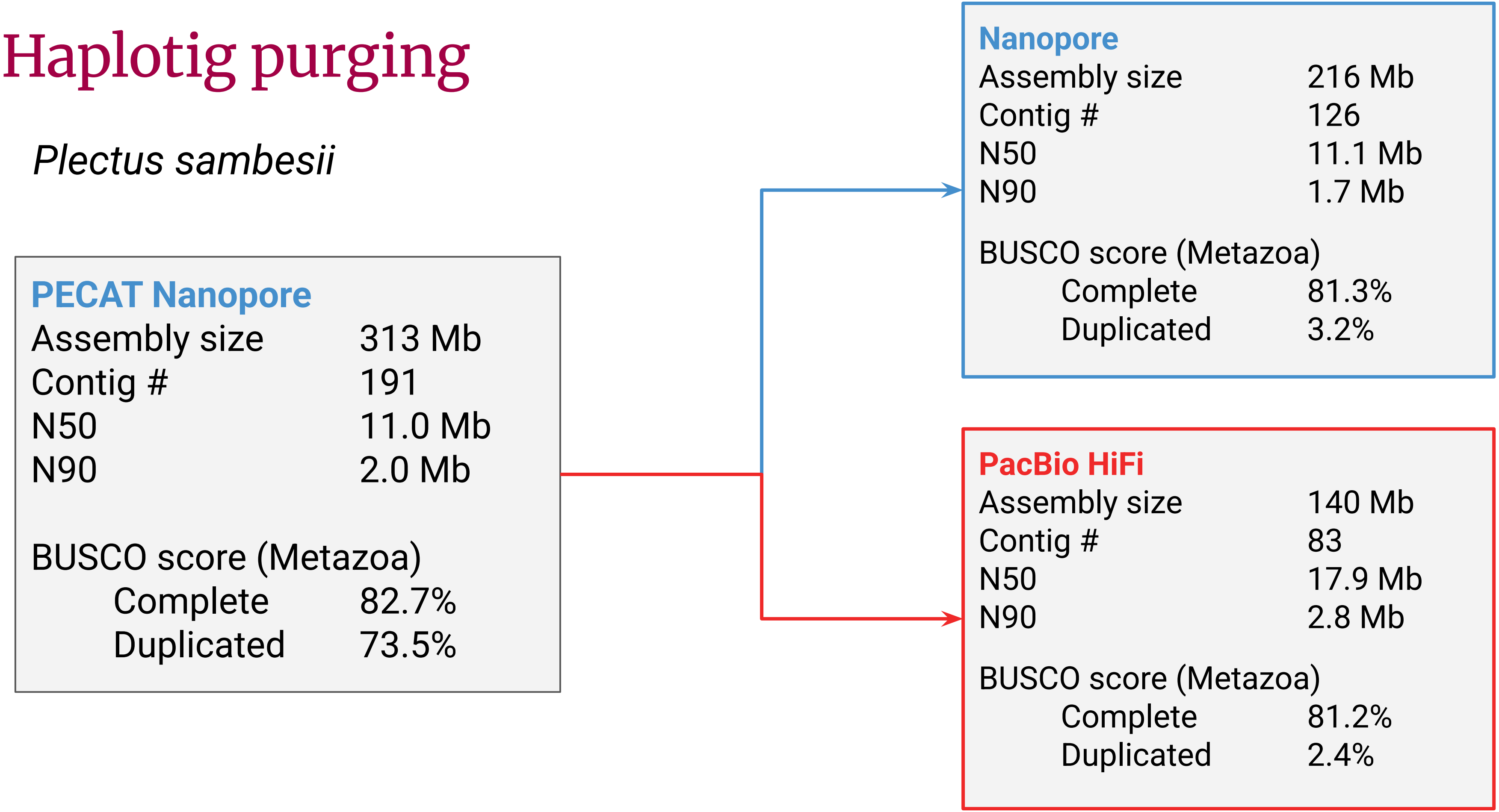
Assembly size	145 Mb
Contig #	161
N50	11.8 Mb
N90	3.7 Mb
BUSCO score (Metazoa)	
Complete	79.9%
Duplicated	1.9%

PacBio HiFi

Assembly size	144 Mb
Contig #	159
N50	11.8 Mb
N90	3.6 Mb
BUSCO score (Metazoa)	
Complete	79.9%
Duplicated	2.0%

Haplotig purging

Plectus sambesii



PECAT Nanopore

Assembly size	313 Mb
Contig #	191
N50	11.0 Mb
N90	2.0 Mb

BUSCO score (Metazoa)

Complete	82.7%
Duplicated	73.5%

Nanopore

Assembly size	216 Mb
Contig #	126
N50	11.1 Mb
N90	1.7 Mb

BUSCO score (Metazoa)

Complete	81.3%
Duplicated	3.2%

PacBio HiFi

Assembly size	140 Mb
Contig #	83
N50	17.9 Mb
N90	2.8 Mb

BUSCO score (Metazoa)

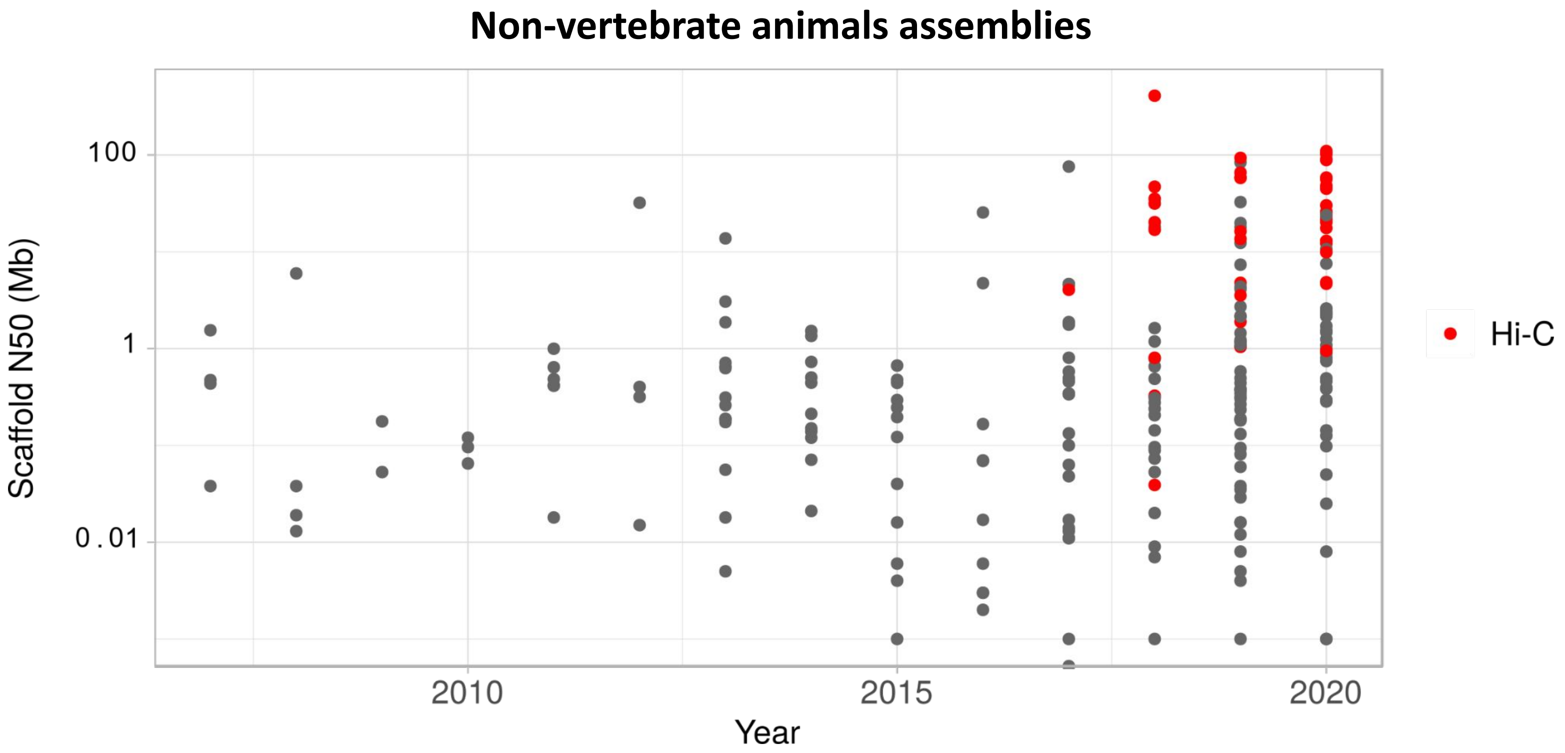
Complete	81.2%
Duplicated	2.4%

Scaffolding approaches

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

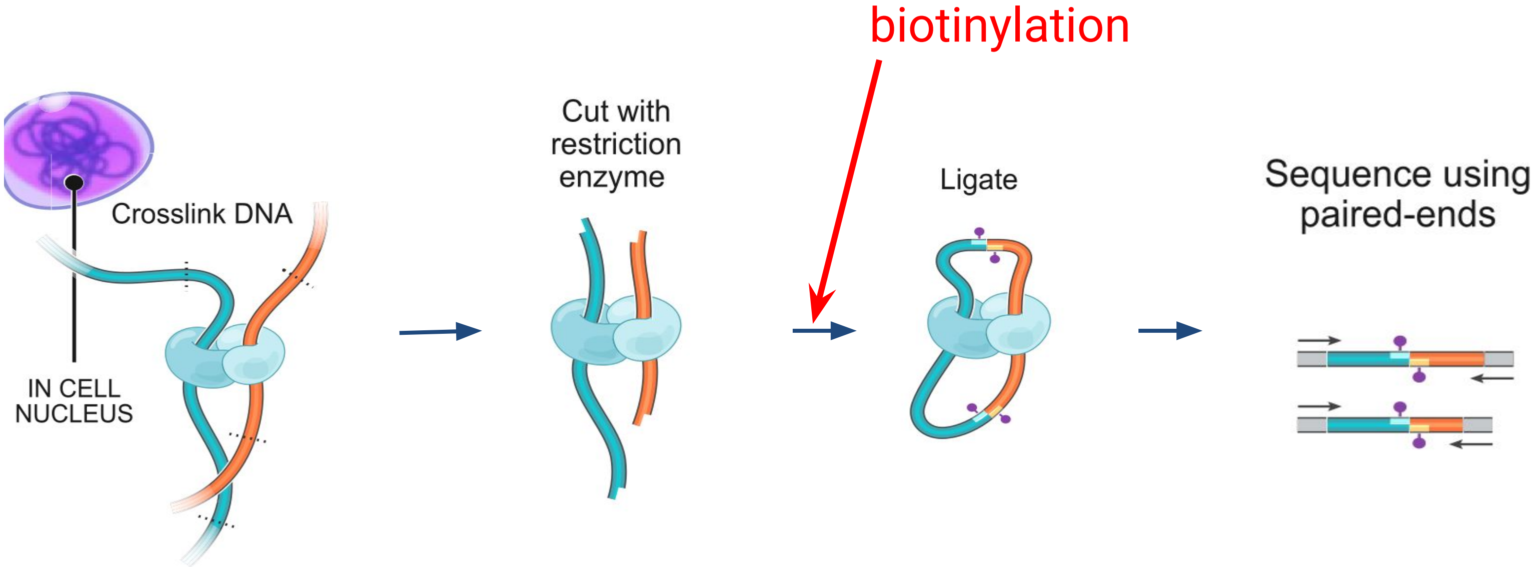
- **Long reads**
- **Genetic maps**
- **Optical maps**
- **Linked reads**
- **Hi-C**

Hi-C scaffolding



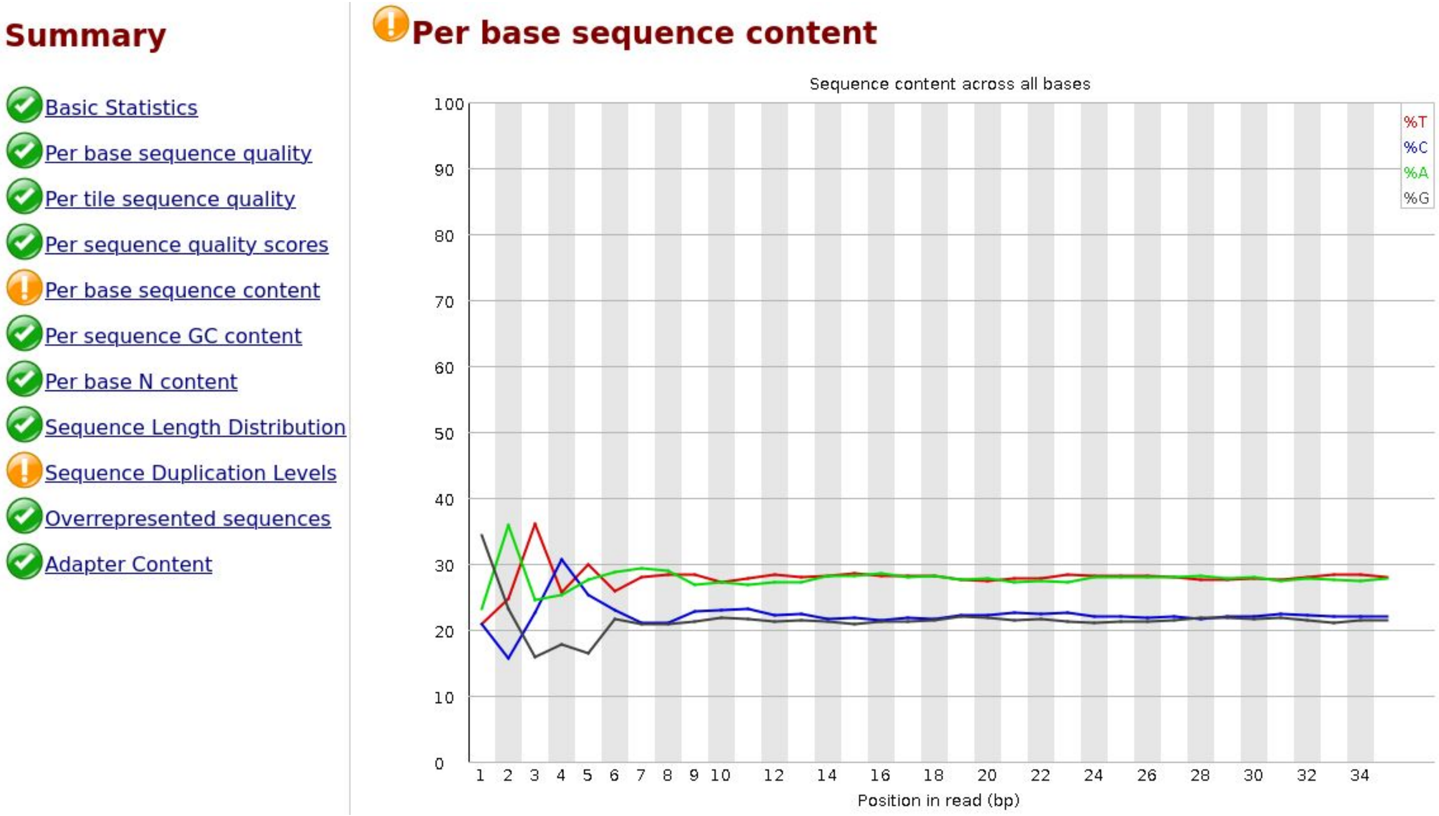
Hi–C scaffolding

Hi-C



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

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

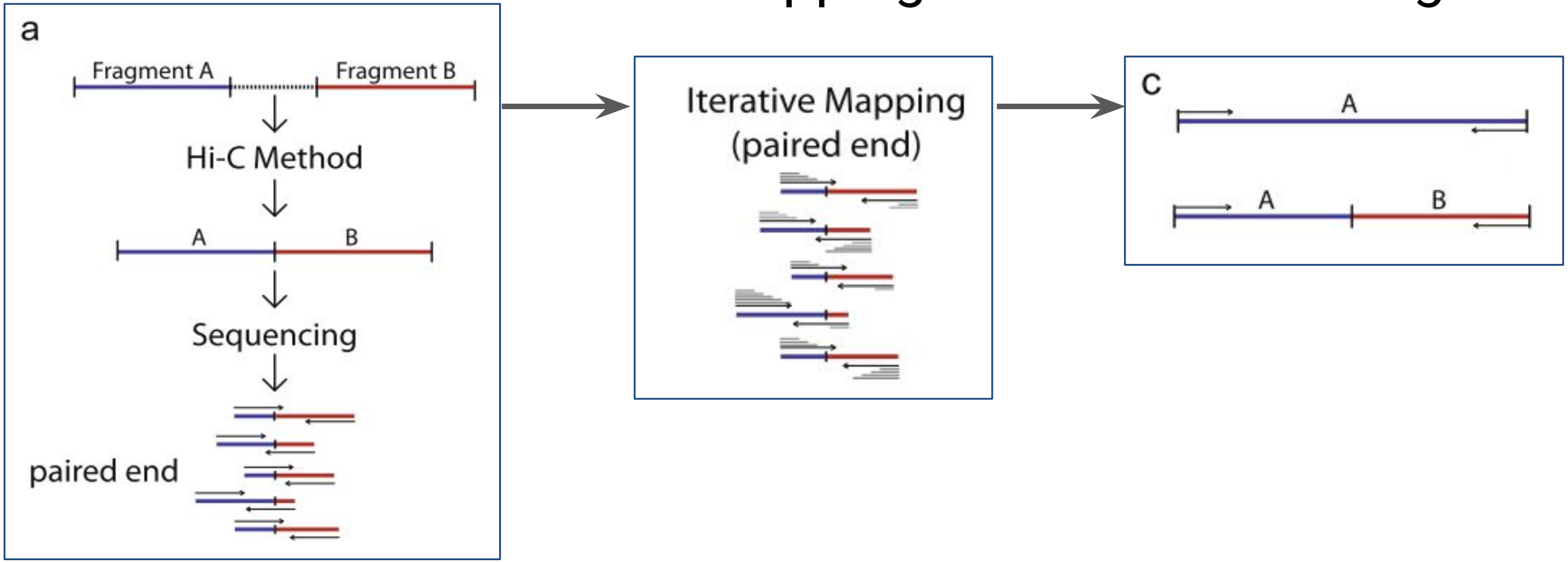
! Per base sequence content

Hi-C scaffolding

Raw data

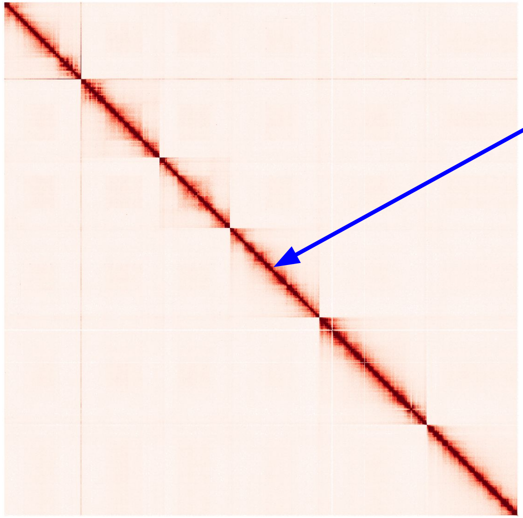
Mapping

Filtering



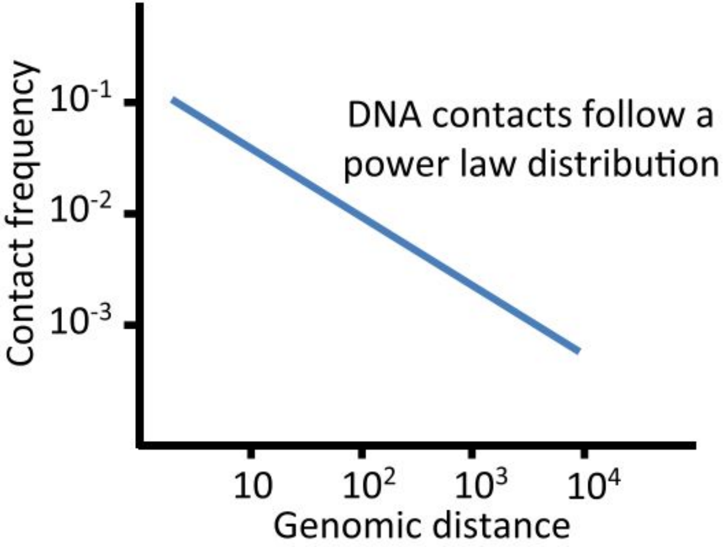
The Hitchhiker’s guide to Hi-C analysis: Practical guidelines. Lajoie et *al.*, 2015

Hi-C scaffolding



Contact map of
Caenorhabditis elegans

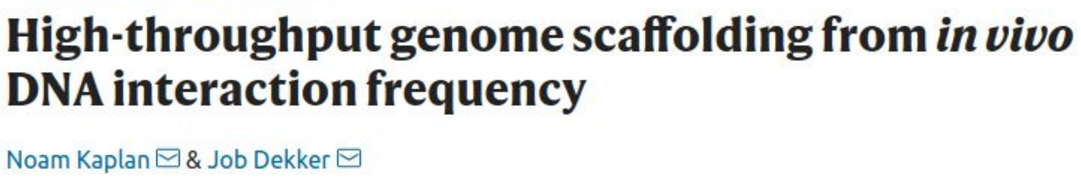
contact frequency = f(genomic distance)



DNA contacts follow a power law distribution

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

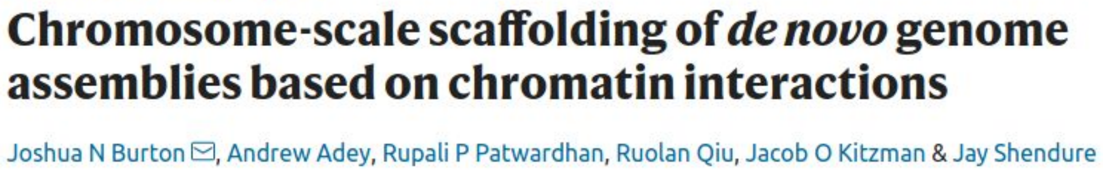
Hi-C scaffolding



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

Noam Kaplan ✉ & Job Dekker ✉

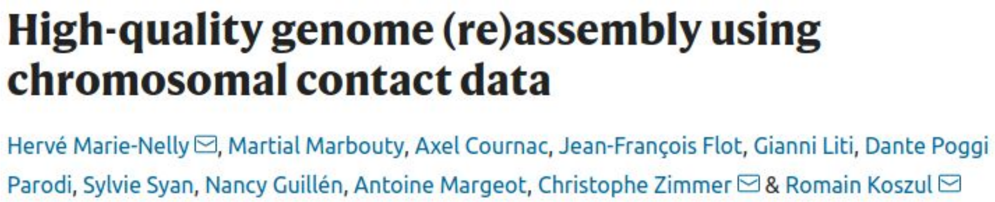
dnaTri



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

Lachesis

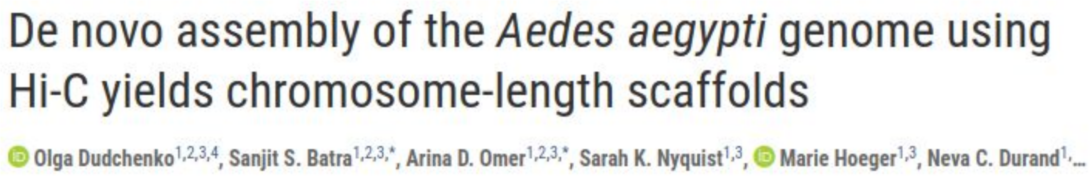


High-quality genome (re)assembly using chromosomal contact data

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

Hi–C scaffolding

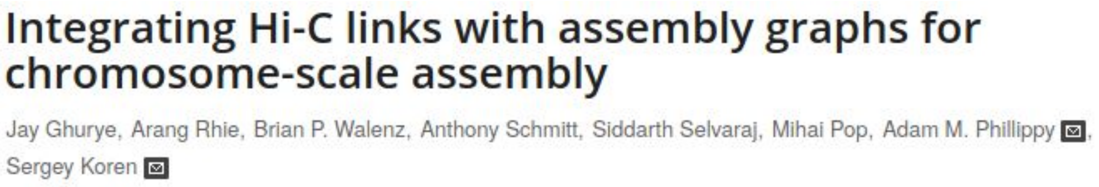


De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds

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,...]

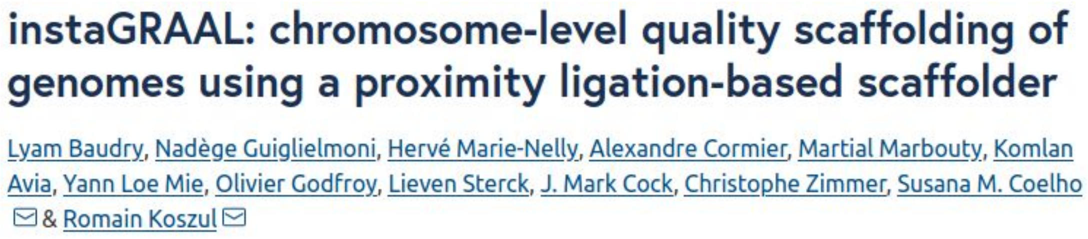
3D-DNA

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 ✉



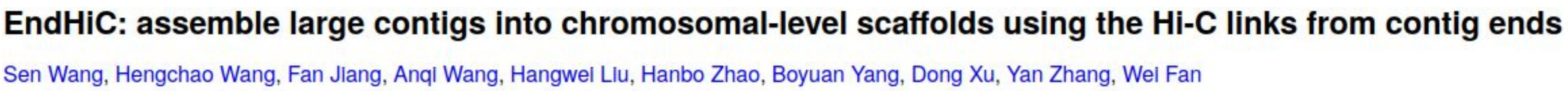
instaGRAAL: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder

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

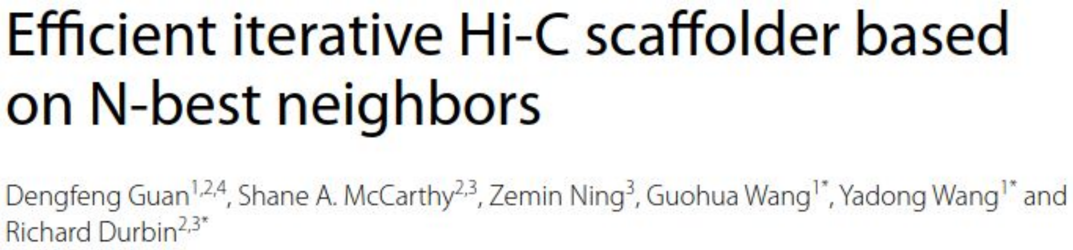
Hi-C scaffolding

And in 2021



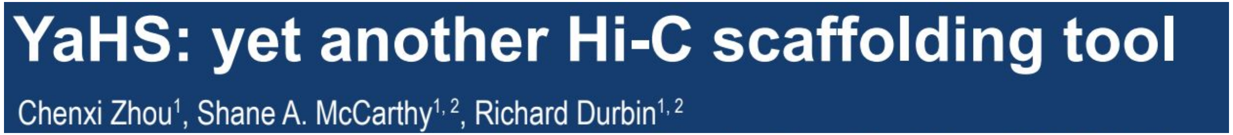
EndHiC: assemble large contigs into chromosomal-level scaffolds using the Hi-C links from contig ends

Sen Wang, Hengchao Wang, Fan Jiang, Anqi Wang, Hangwei Liu, Hanbo Zhao, Boyuan Yang, Dong Xu, Yan Zhang, Wei Fan



Efficient iterative Hi-C scaffolder based on N-best neighbors

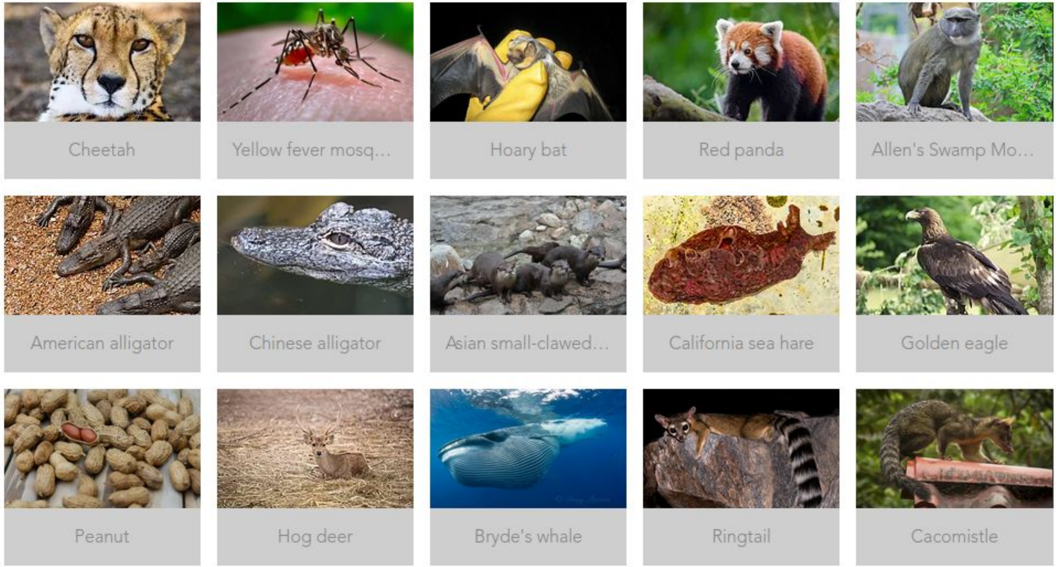
Dengfeng Guan[1,2,4], Shane A. McCarthy[2,3], Zemin Ning[3], Guohua Wang[1*], Yadong Wang[1*] and Richard Durbin[2,3*]



YaHS: yet another Hi-C scaffolding tool

Chenxi Zhou[1], Shane A. McCarthy[1, 2], Richard Durbin[1, 2]

Hi-C scaffolding

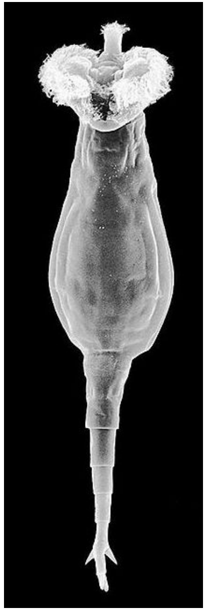


www.dnazoo.org

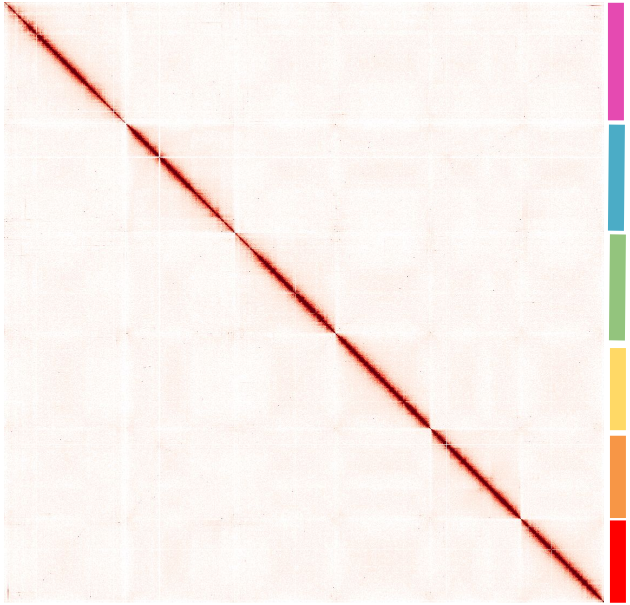
Hi-C scaffolding

Adineta vaga (rotifer)

6 scaffolds

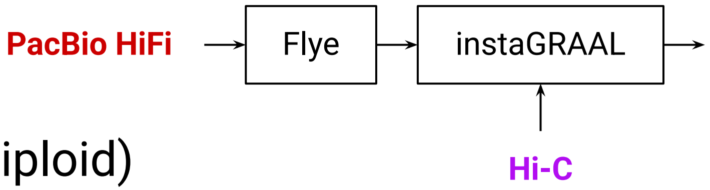


Who Needs Sex (or Males) Anyway?
Liza Gross, PloS Biology, 2007



Hi-C contact map of *Adineta vaga*

Hi–C scaffolding



PacBio HiFi

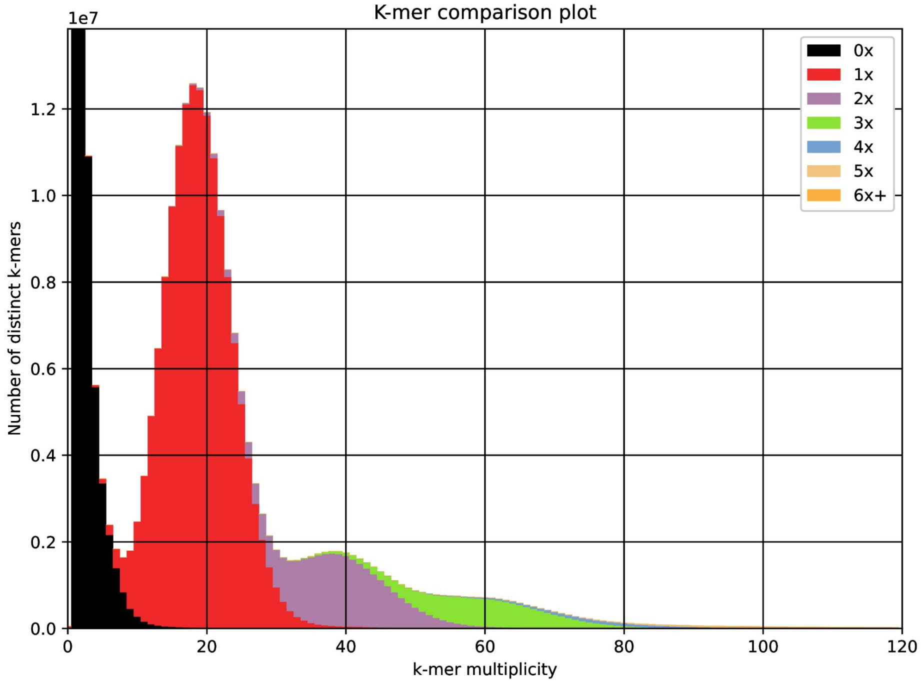
Flye

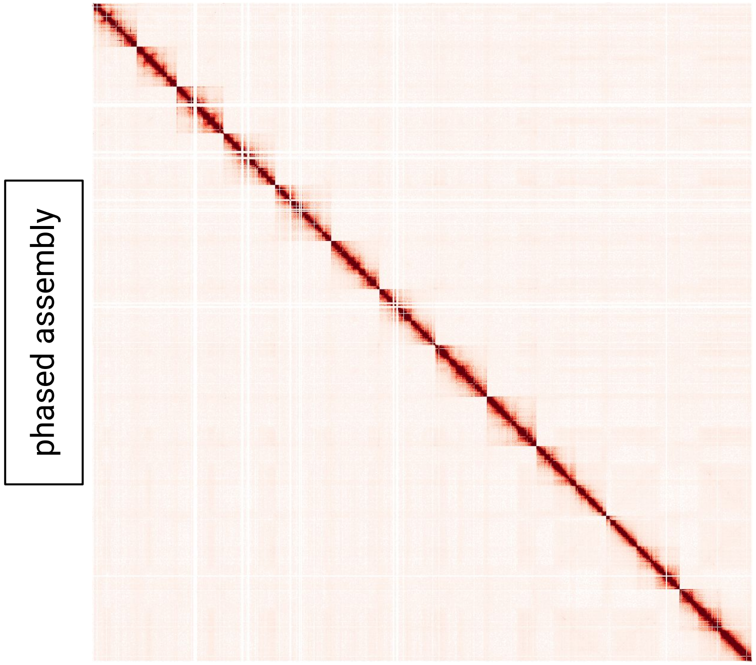
instaGRAAL

Hi-C

Panagrolaimus sp. PS1579
Size 314 Mb
12 chromosomes
19.8 - 35.1 Mb

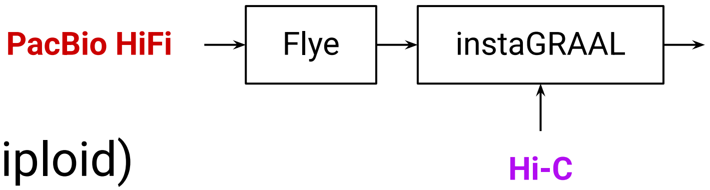
Panagrolaimus sp. PS1579 (triploid)





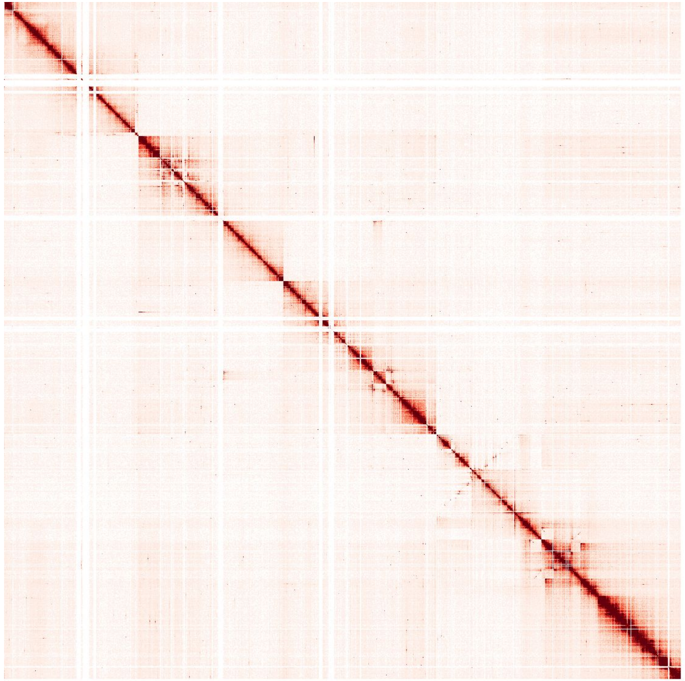
Hi-C scaffolding

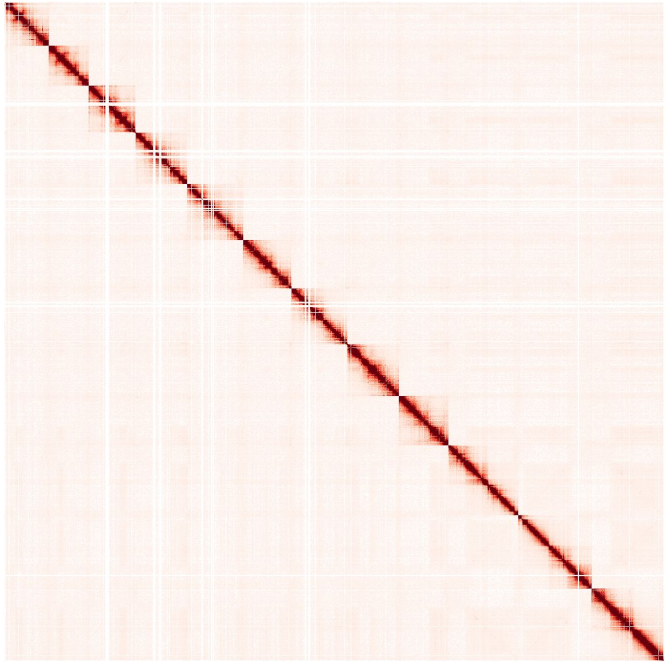
Panagrolaimus sp. PS1579 (triploid)


PacBio HiFi → Flye → instaGRAAL →

Hi-C

Panagrolaimus sp. PS1579
Size 314 Mb
12 chromosomes
19.8 - 35.1 Mb


purged assembly


phased assembly

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%

And then...

- **Gap filling**: TGS-GapCloser...
- **Polishing**: using high-accuracy reads, HyPo, Racon...

Gap filling & Polishing

	Scaffolds	**After TGS-Gapcloser**	**After HyPo**
Flaccisagitta enflata	9,239	3,694	1,476
Norana najaformis	860	748	632
Lucinoma borealis	24,786	5,093	2,135

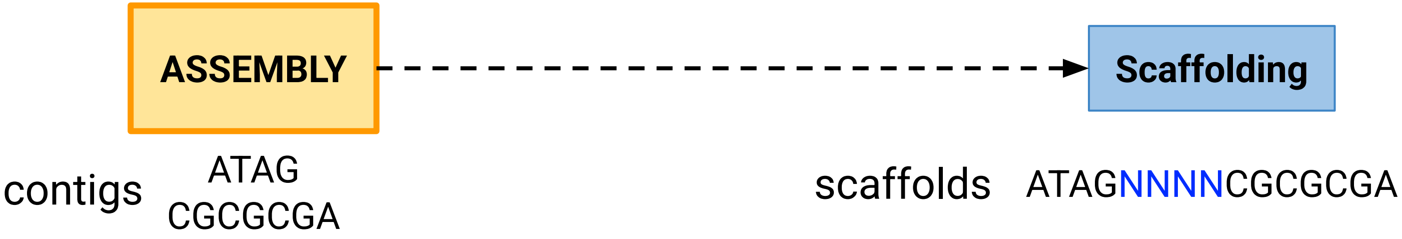
Assembly pipeline

ASSEMBLY

```
        ATTTGTACG
reads       GTACGGACA
                GGACATAGTA

contig  ATTTGTACGGACATAGTA
```

Assembly pipeline



ASSEMBLY

Scaffolding

contigs

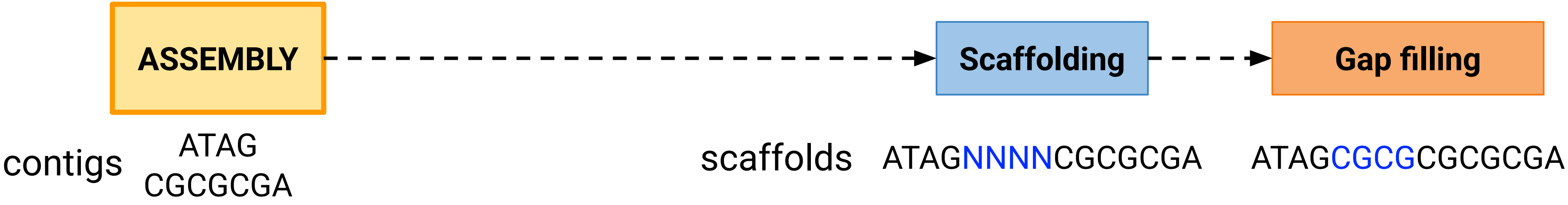
ATAG

CGCGCGA

scaffolds

ATAGNNNNCGCGCGA

Assembly pipeline



ASSEMBLY

Scaffolding

Gap filling

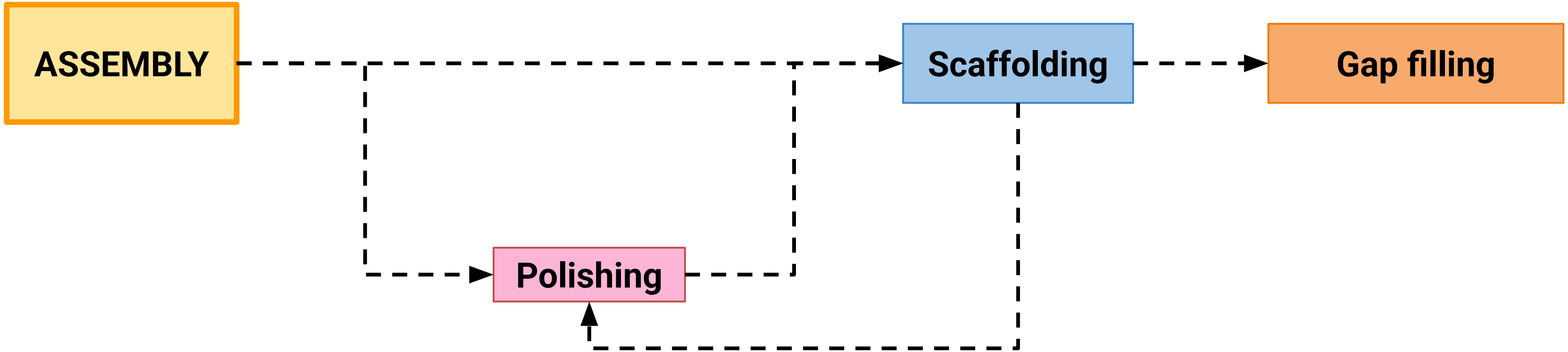
contigs ATAG

CGCGCGA

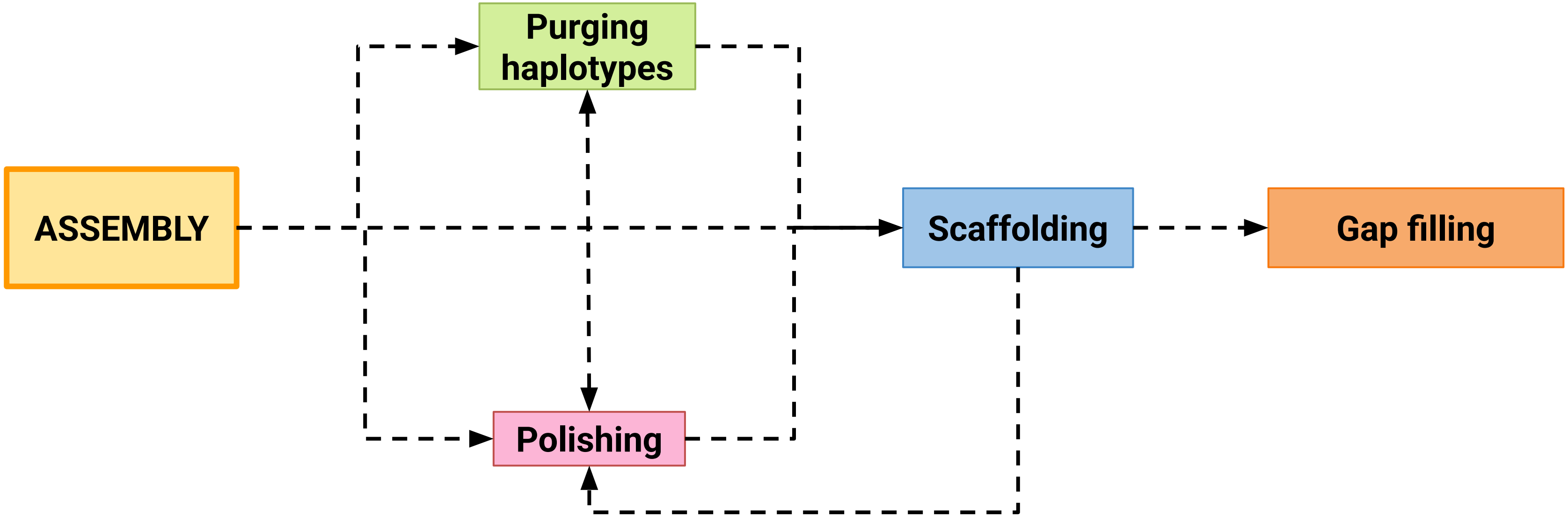
scaffolds ATAGNNNNCGCGCGA

ATAGCGCGCGCGCGA

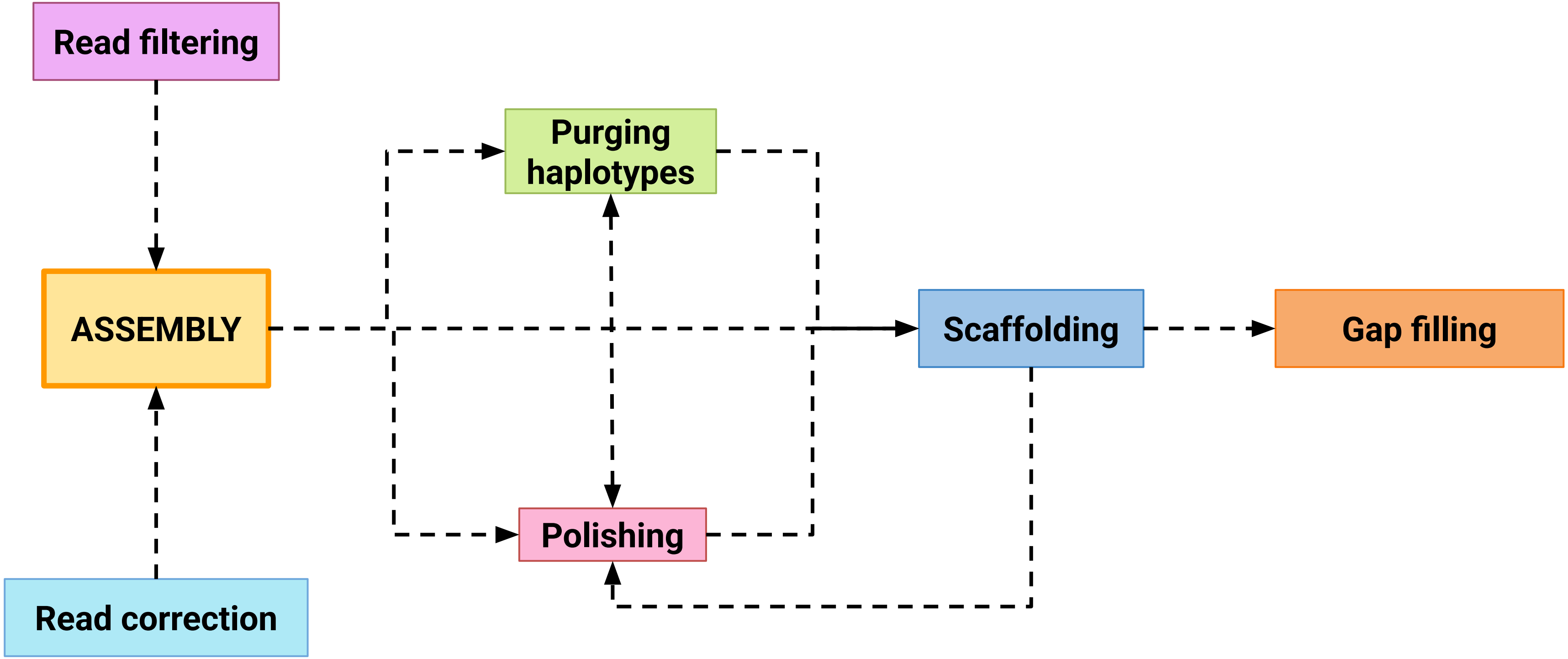
Assembly pipeline



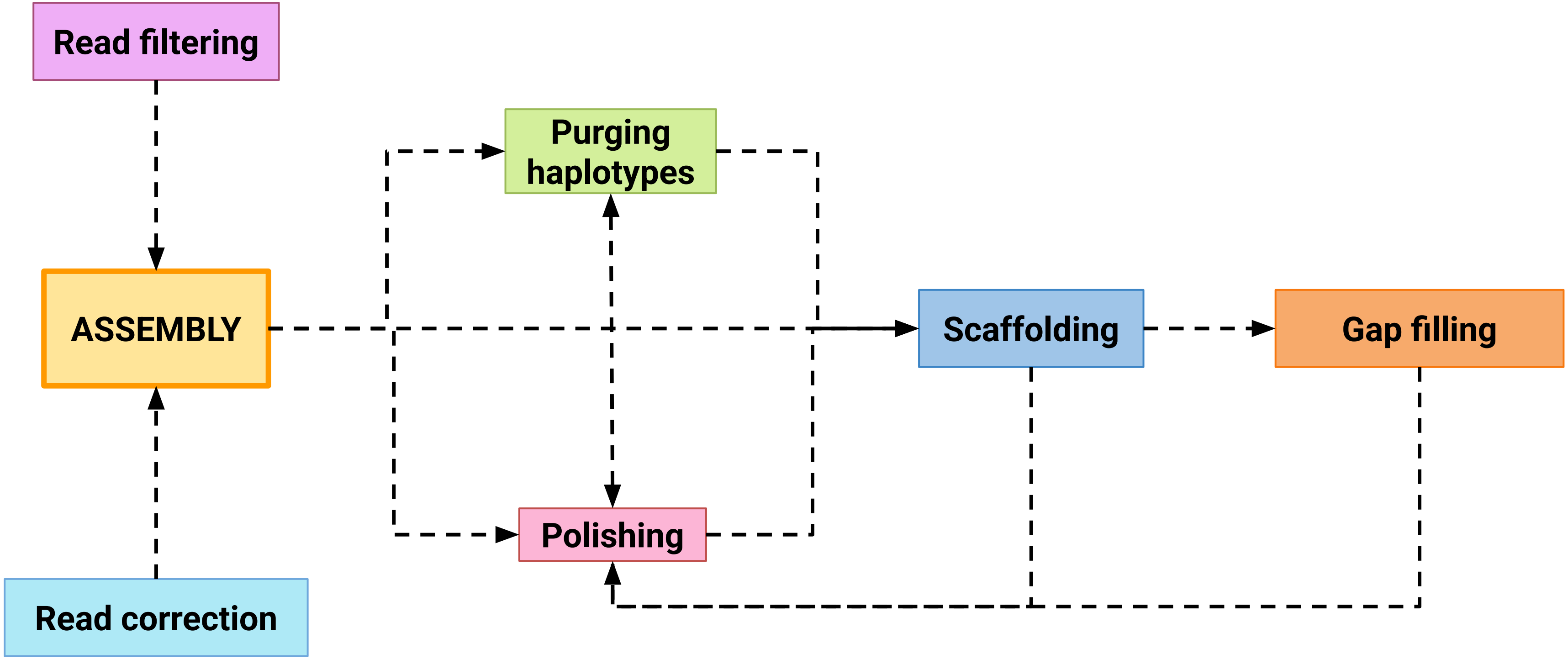
Assembly pipeline



Assembly pipeline



Assembly pipeline


Read filtering
Purging haplotypes
ASSEMBLY
Scaffolding
Gap filling
Polishing
Read correction