

Topic 5: Genome Assembly

de novo genome assembly

- The basic concept
- Complications for genome assembly
- Evaluating assembly quality
- The dominant paradigm for genome assembly

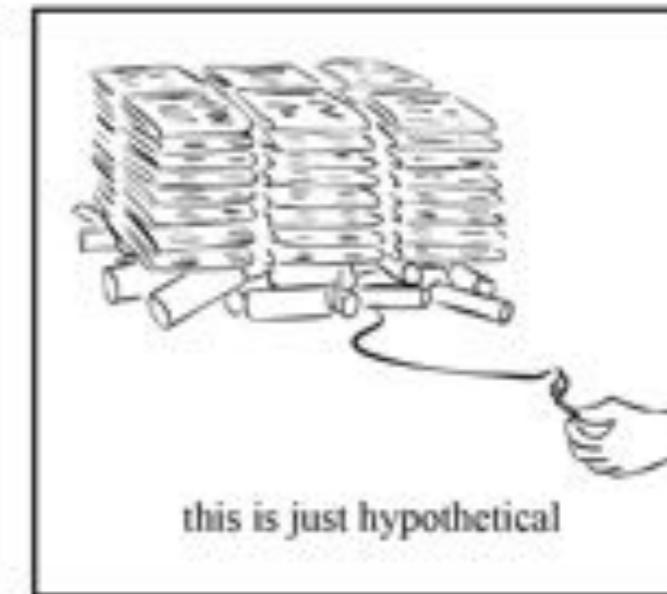
The central challenge...



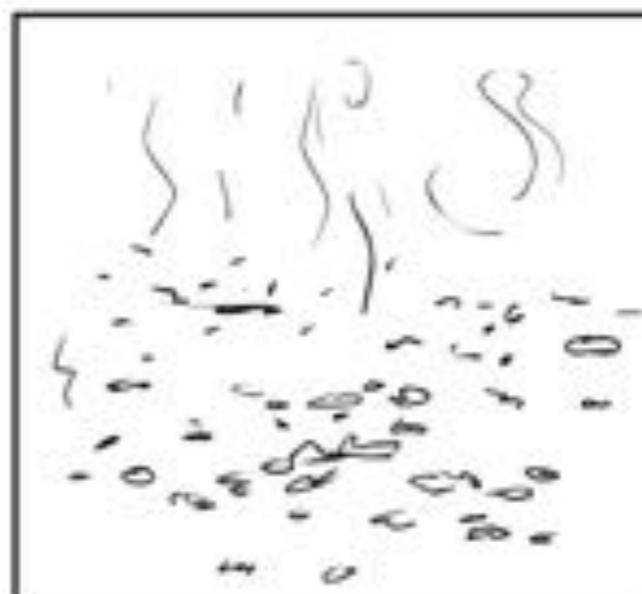
stack of NY Times, June 27, 2000



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



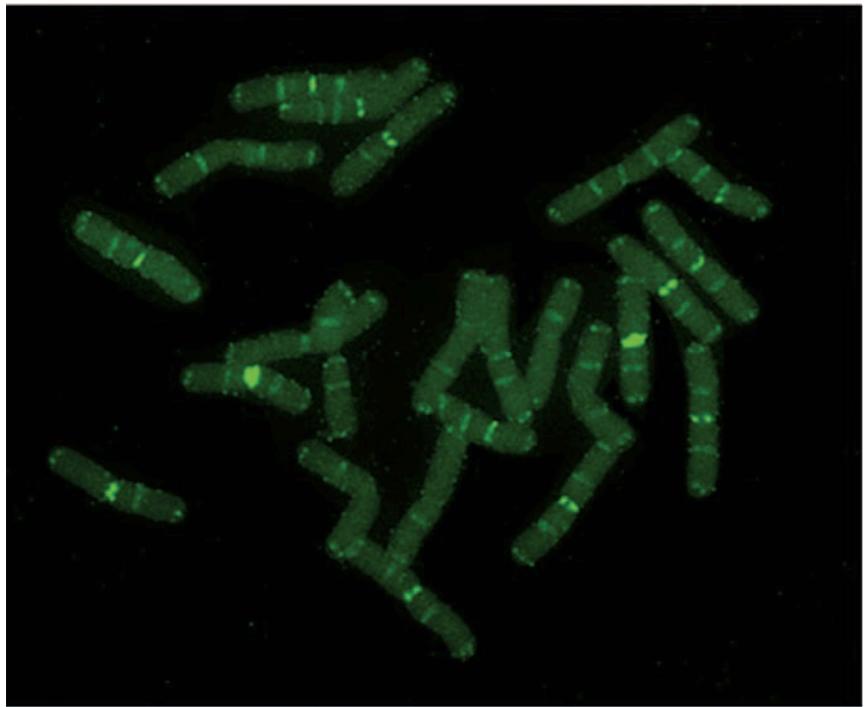
this is just hypothetical



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

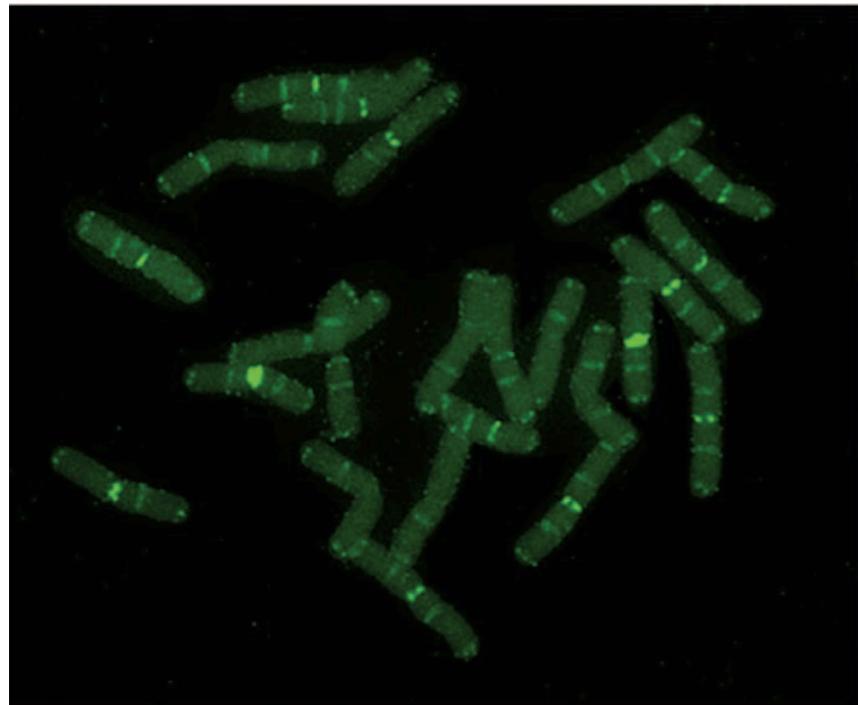
Genome sequencing

Loblolly pine chromosomes

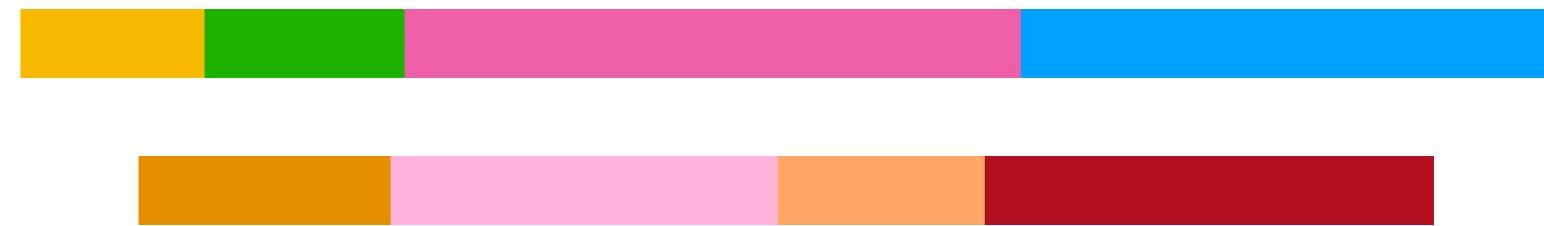


Genome sequencing

Loblolly pine chromosomes

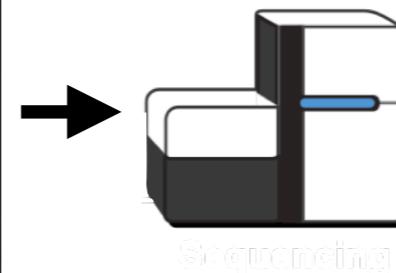
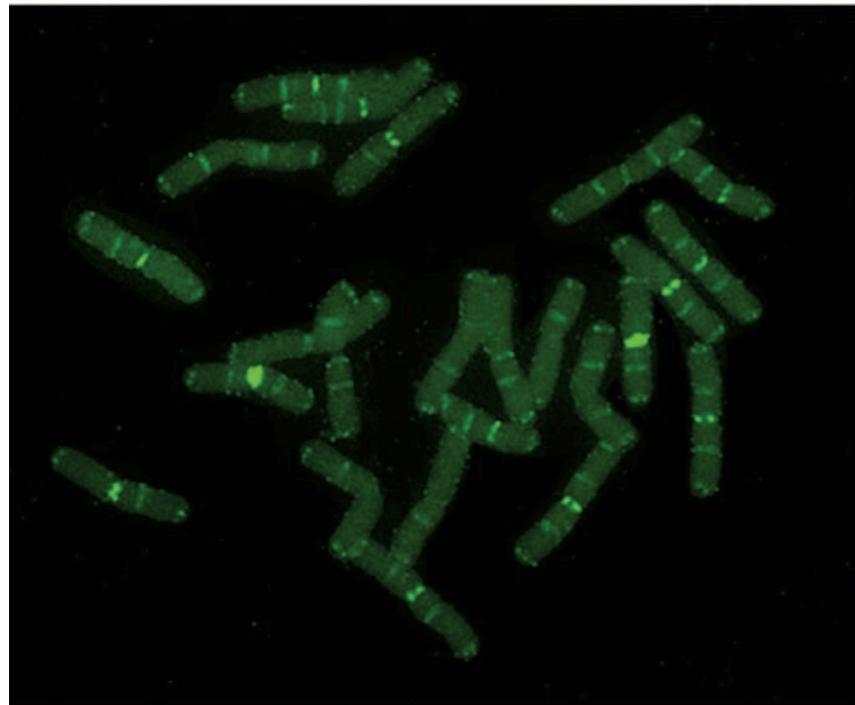


Cartoon chromosomes



Genome sequencing

Loblolly pine chromosomes

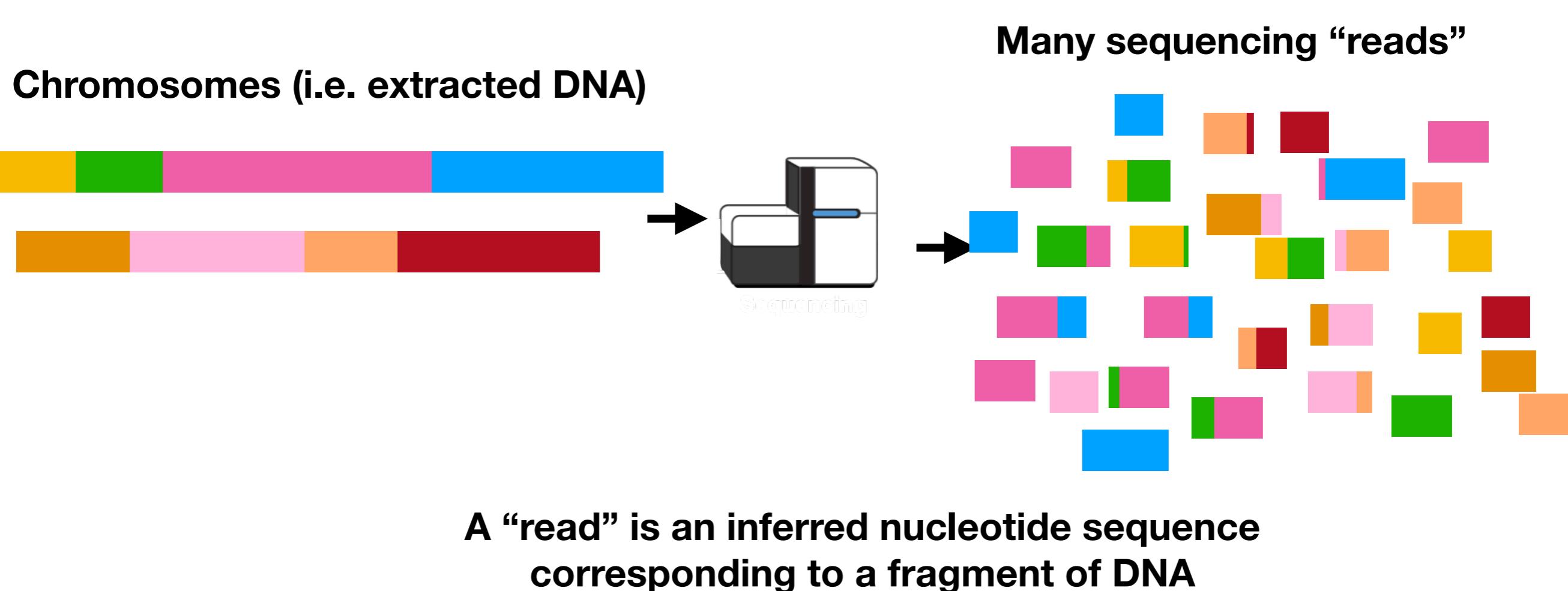


DNA sequencing machine

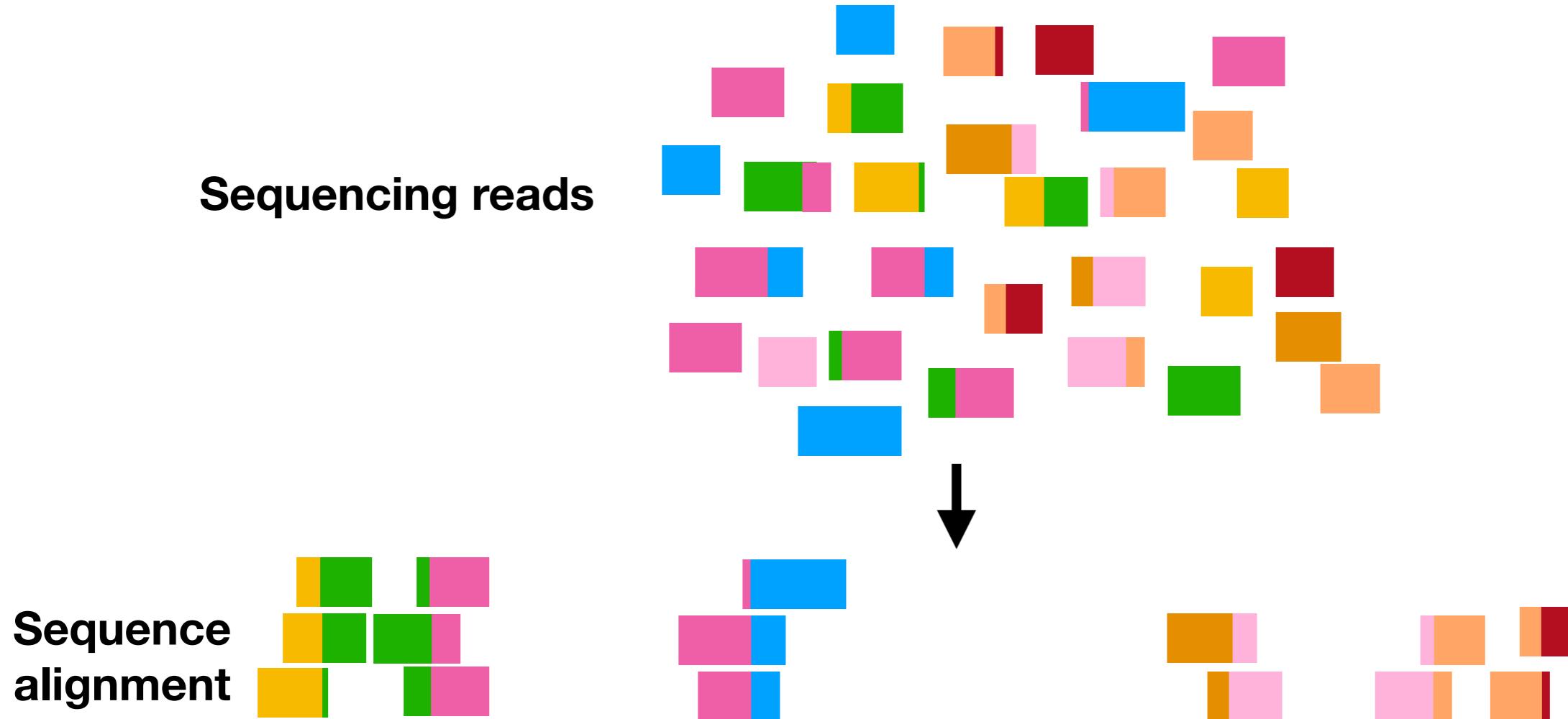
- There are numerous technologies available
- The various technologies have different attributes

Genome sequencing

What does DNA sequence data actually look like?

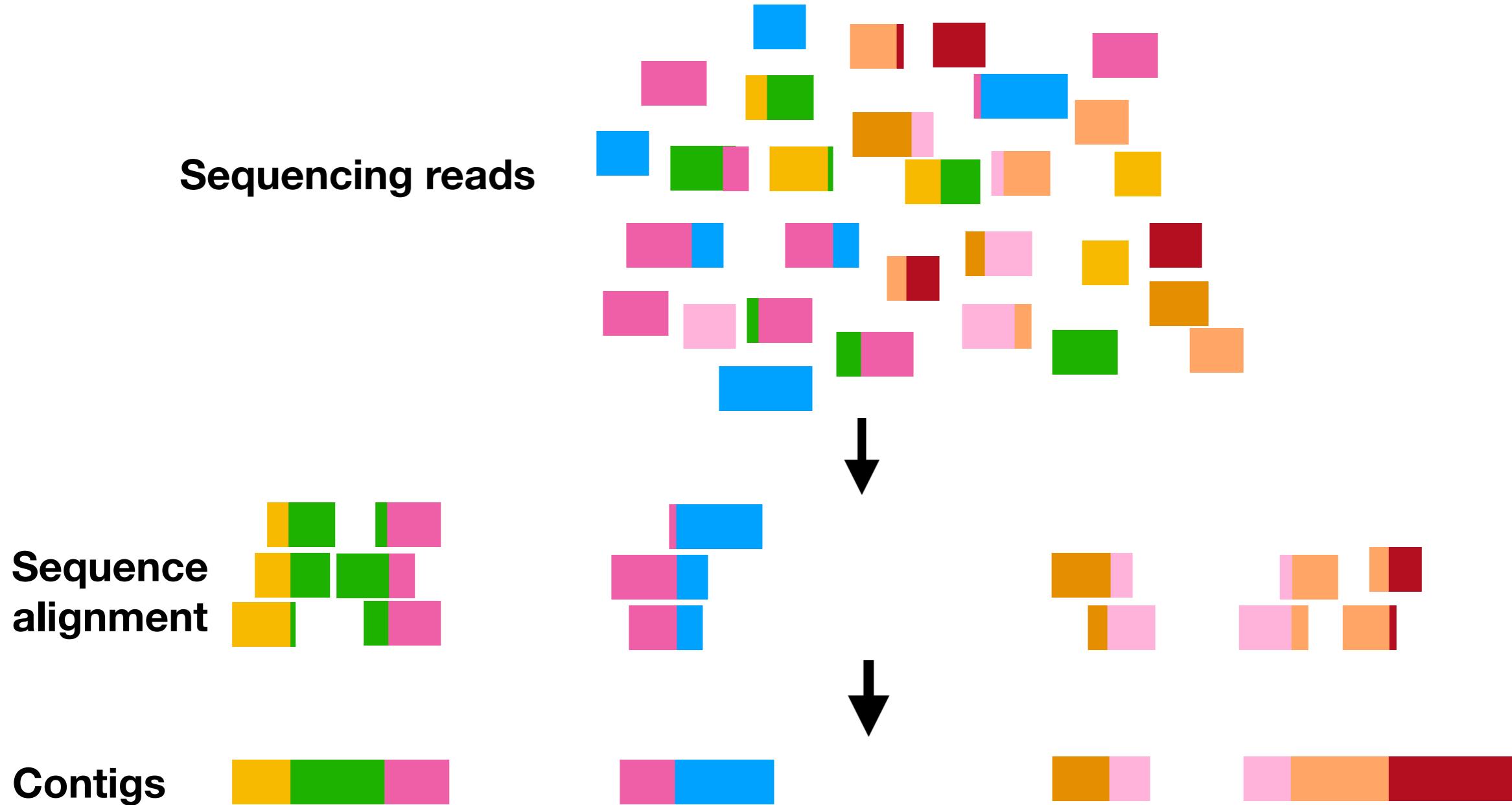


Genome assembly



Genome assembly is the process of reconstructing an individual's genome from sequence data

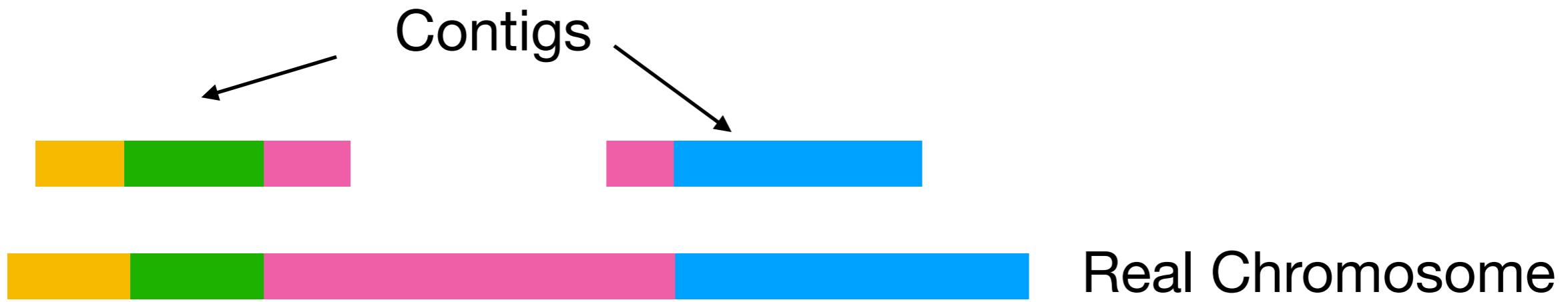
Genome assembly



Genome assembly is the process of reconstructing an individual's genome from sequence data

Genome assembly

Contig = Contiguous sequence



With additional knowledge we can combine contigs into scaffolds, connected with gaps (typically Ns)



Reference genome

A reference genome is a representation of the average genome for a species/population

- Used as the template against which to evaluate genetic variation (see next lecture)
- Reference genomes are often incomplete pictures

What a reference genome looks like



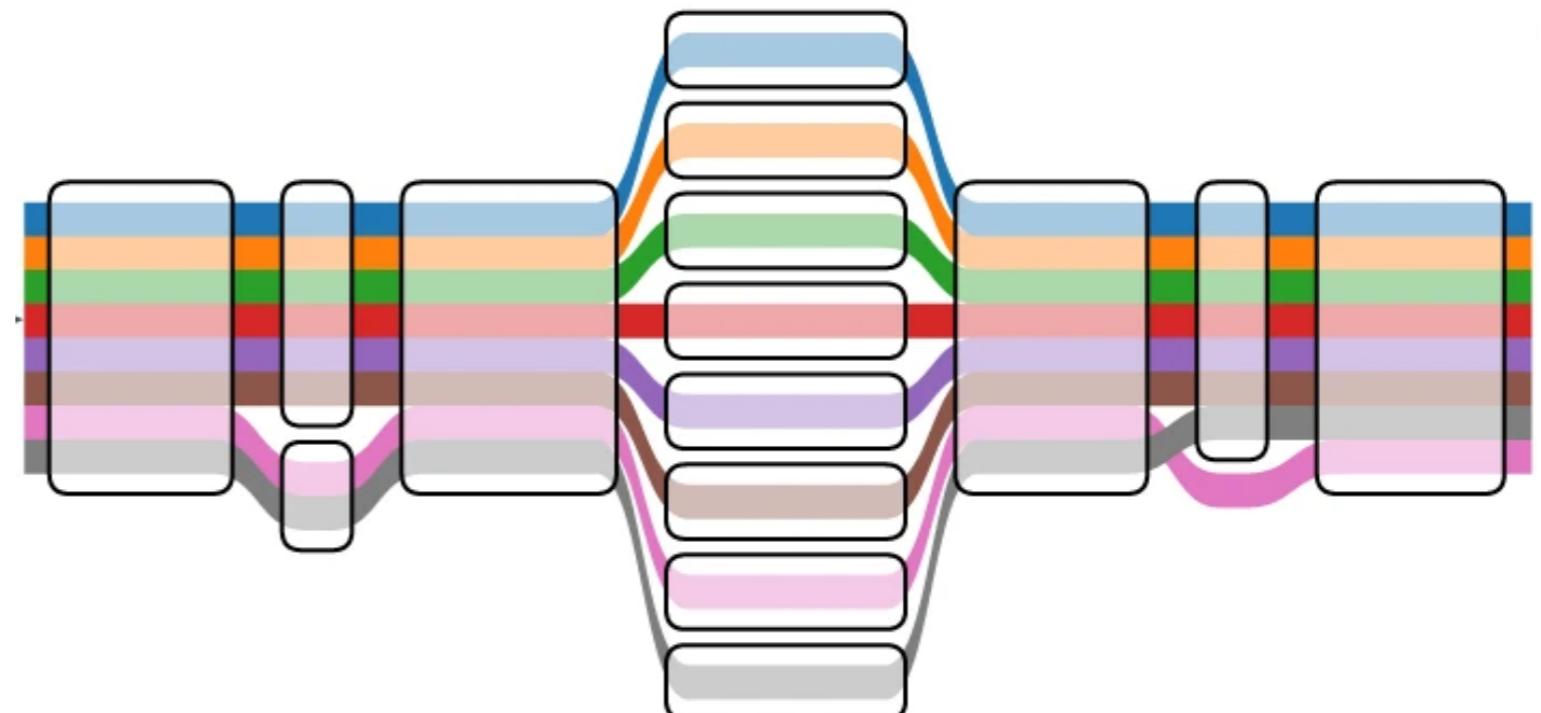
```
>chr_1
TGGGCAAGGCTGATGAACAGCAGCTGCATAAATTCTCCCTAATTATATTGTAAATAGCT
GCAGCACACAATAAAGCTTGTTAGAGACATCTAGAGAATCACACACTGCATCTGTTCT
GCCGCTCTCCCTCTTGCTCTGTTCTGAGAACGACTTGTTCACTGATTCTGGTTGTATT
TGTGTTTTCATGCTAACATTGTTATTGTTGCCTAGAAAGTTCTTGATTGGGCCAA
ATTAGTCGATTTAAAGAGTCACCTCTCTAGTCATGTAATCTATGTGGACATCTCAAT
AGCTGCTTAATTGTTAGTGGTAATCTCCTCTGAACAGAGAGAAAGGCCTACATGCAGC
CCTCAGAGGAGAGGTGTCAATCTCTCTTGTATTCTCTTGTGTTCCCTCAGAAGAATC
ATTCTAACATGGTATTGTACAAGAGGAAATAATGGGACTAAAACCAGGCATGCACCATC
TGATAGATTACATCCCTAGAAGACTTTGTTGTGTTCAAGTGGAGAGCCTGCTG
```

Pangenomes

A pan genome is the collection of genes present in a species (i.e. including copy number and structural variants)

- A richer picture of the genome than a single collapsed FASTA file
- Represented as a graph of genetic
- Currently limited analysis options

What a reference
pan genome looks like



From Hickey et al 2024

Choices, choices, choices

When setting out to perform genome assembly, one has a LOT of options to consider:

Sequencing type(s)

Assembly software

Sequencing depth

Metrics to assess quality

Choices, choices, choices

When setting out to perform genome assembly, one has a LOT of options to consider:

Sequencing type(s)

Assembly software

Sequencing depth

Metrics to assess quality

Short reads, long reads, DNA/RNA
Illumina, PacBio, Oxford Nanopore
Source of DNA (i.e. tissue)
Library preparation method (e.g.
HMW, HiC)

Choices, choices, choices

When setting out to perform genome assembly, one has a LOT of options to consider:

Sequencing type(s)

Assembly software

Sequencing depth

Metrics to assess quality

Alice-asm, Flye, HiCanu, hifiasm,
IPA, LJA, mdBG, MBG

(These are assemblers just for hifi reads; see full list: https://github.com/nadegeguiglielmoni/genome_assembly_tools)

Choices, choices, choices

When setting out to perform genome assembly, one has a LOT of options to consider:

Sequencing type(s)

Assembly software

Sequencing depth

Metrics to assess quality

How much do you sequence?

Largely determined by your project budget

Strategically allocating resources onto different sequencing types will likely be determined by your organism and question

Choices, choices, choices

When setting out to perform genome assembly, one has a LOT of options to consider:

Sequencing type(s)

Assembly software

Sequencing depth

Metrics to assess quality

How do you know that you did a good job?

How to compare the quality of your assembly to that of another?

Factors that complicate genome assembly

What are some factors that would complicate genome assembly?

Approaches to genome assembly

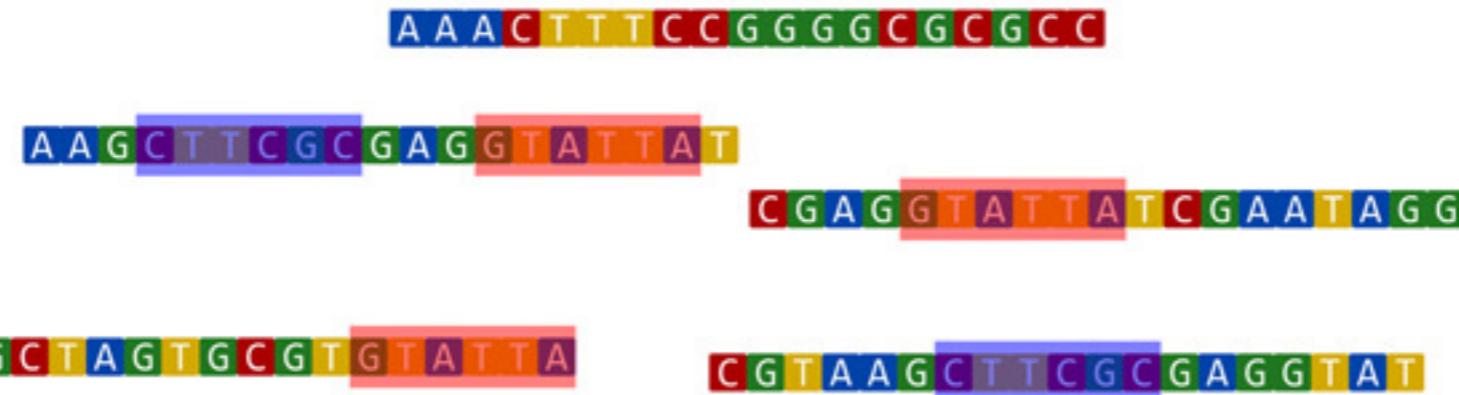
There are three main algorithms at the heart of genome assemblers

- Overlap-Layout-Consensus
- De-Brujin Graph
- String Graph

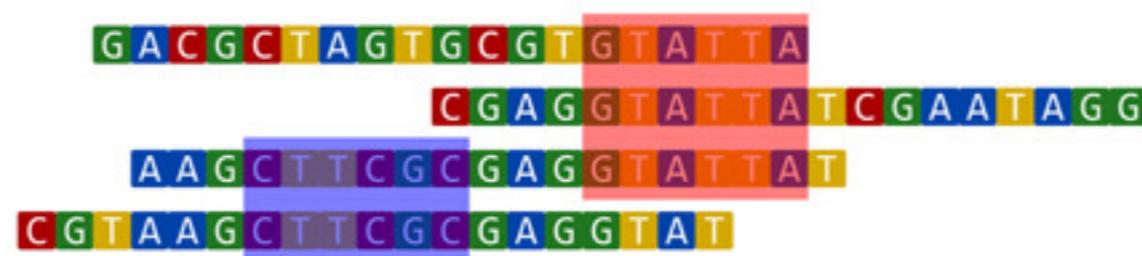
There are three main algorithms at the heart of genome assemblers

- Overlap-Layout-Consensus

Overlap



Layout

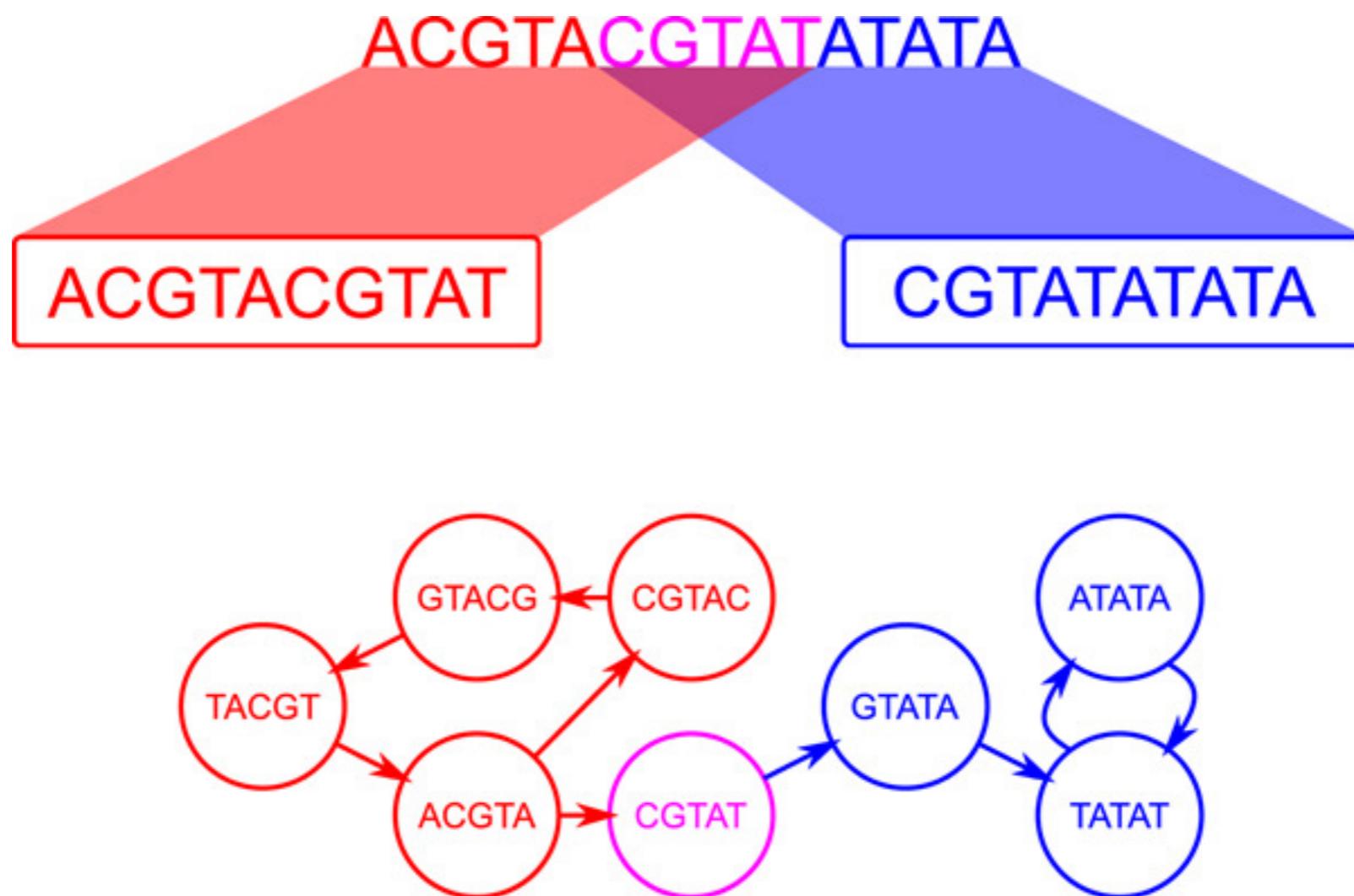


Consensus

- CGTAAGCTTCGCGAGGTATTATCGAATAGG

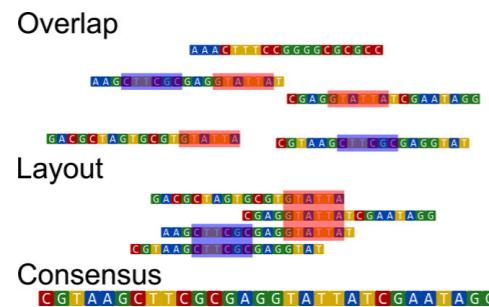
There are three main algorithms at the heart of genome assemblers

- De-Brujin Graph



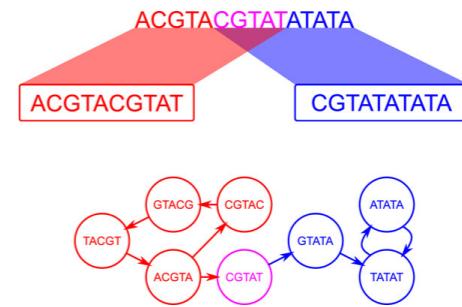
There are three main algorithms at the heart of genome assemblers

Overlap-Layout-Consensus



better with long reads
can tolerate some mismatch (e.g. heterozygosity/sequencing error)
high coverage causes memory issues
repeats can cause memory issues

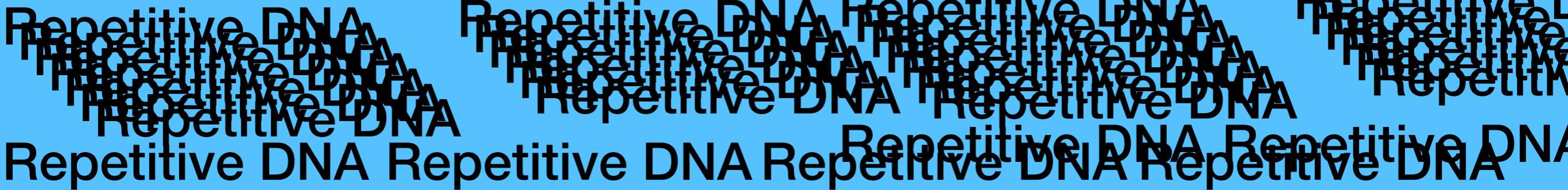
De-Bruijn Graph



better for short reads
repeats are collapsed in the graph
no overlap step
shorter contigs
loses information

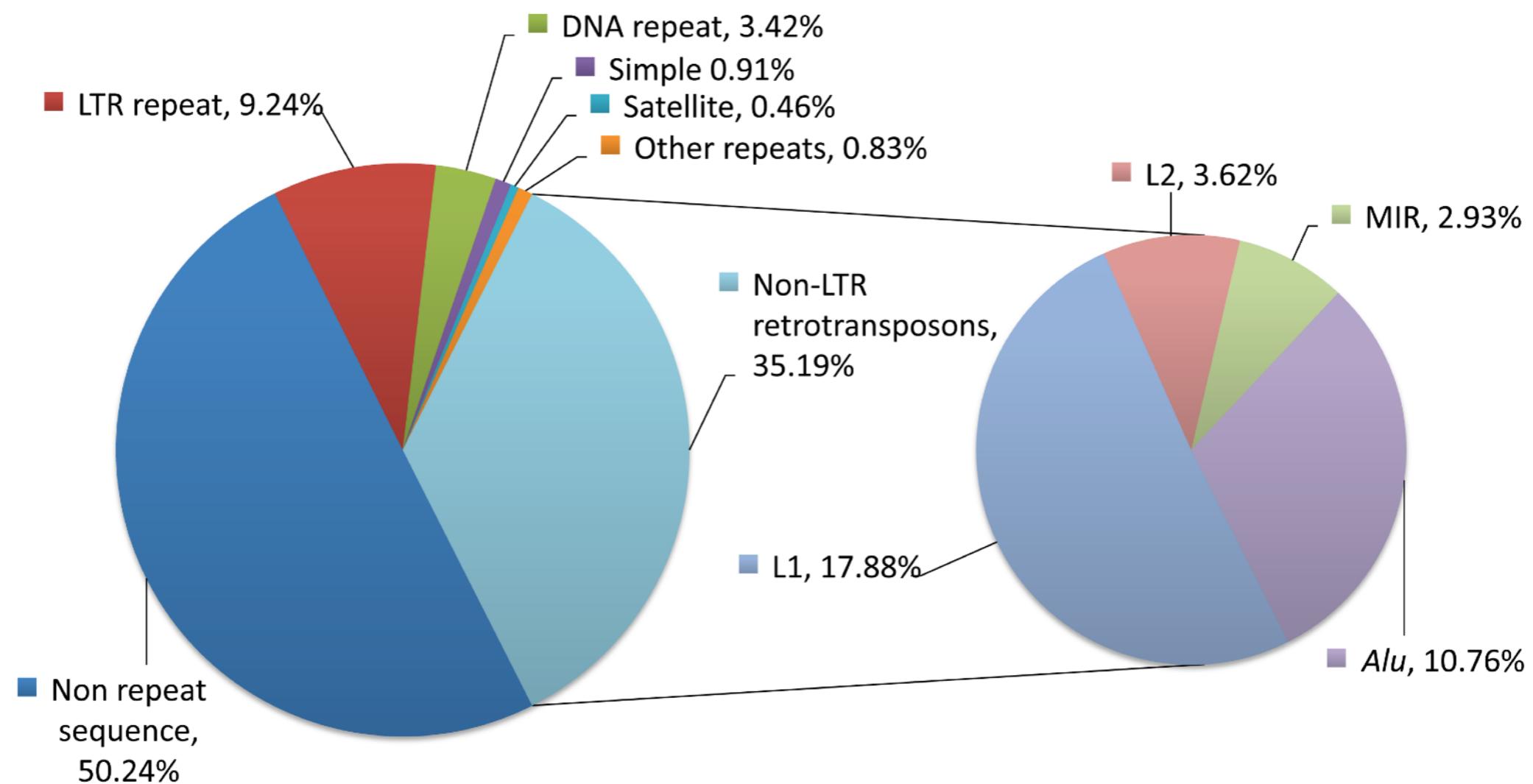
Factors that complicate genome assembly

- Sequencing error
- Repetitive DNA
- Genome Size
- Polyploidy
- Heterozygosity
- Sequencing artefacts (e.g. intact adaptor sequences)



Repetitive DNA is a feature of eukaryotic genomes

The human genome, for example, is approximately 50% repetitive



Repetitive DNA

Repetitive DNA breaks up the assembly and can obscure the order and orientation of contigs

Even well studied model organisms can have poorly assembled regions of their genomes

For example, the human Y-chromosome (enriched for repeats) was only fully assembled in 2023

The dominant paradigm for genome assembly

In the last 5 years or so, approaches to genome assembly have largely coalesced around a small number of approaches

The dominant paradigm for genome assembly

In the last 5 years or so, approaches to genome assembly have largely coalesced around a small number of approaches



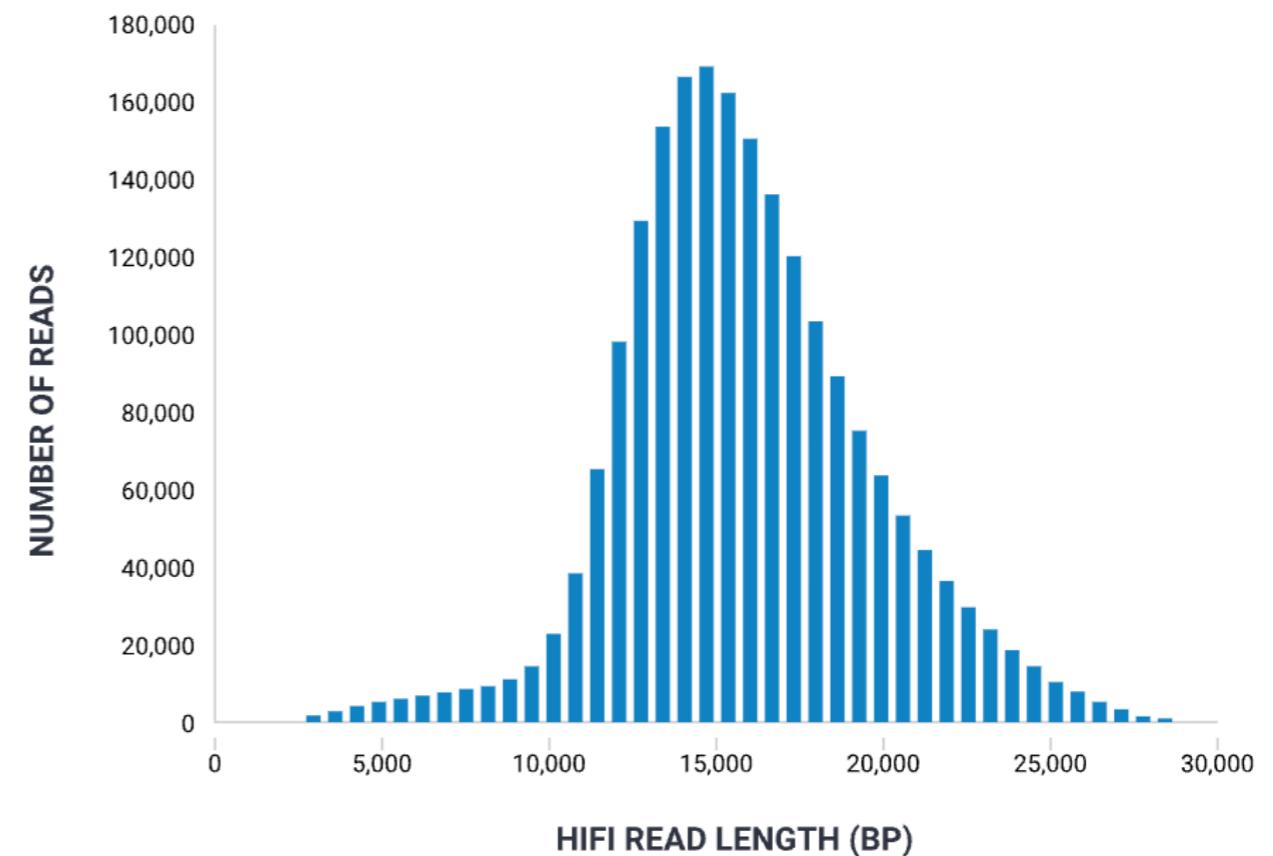
Subsidiaries include:

The Canada Biogenome Project
The Darwin Tree of Life Project (UK)

The dominant paradigm for genome assembly

In the last 5 years or so, approaches to genome assembly have largely coalesced around a small number of approaches

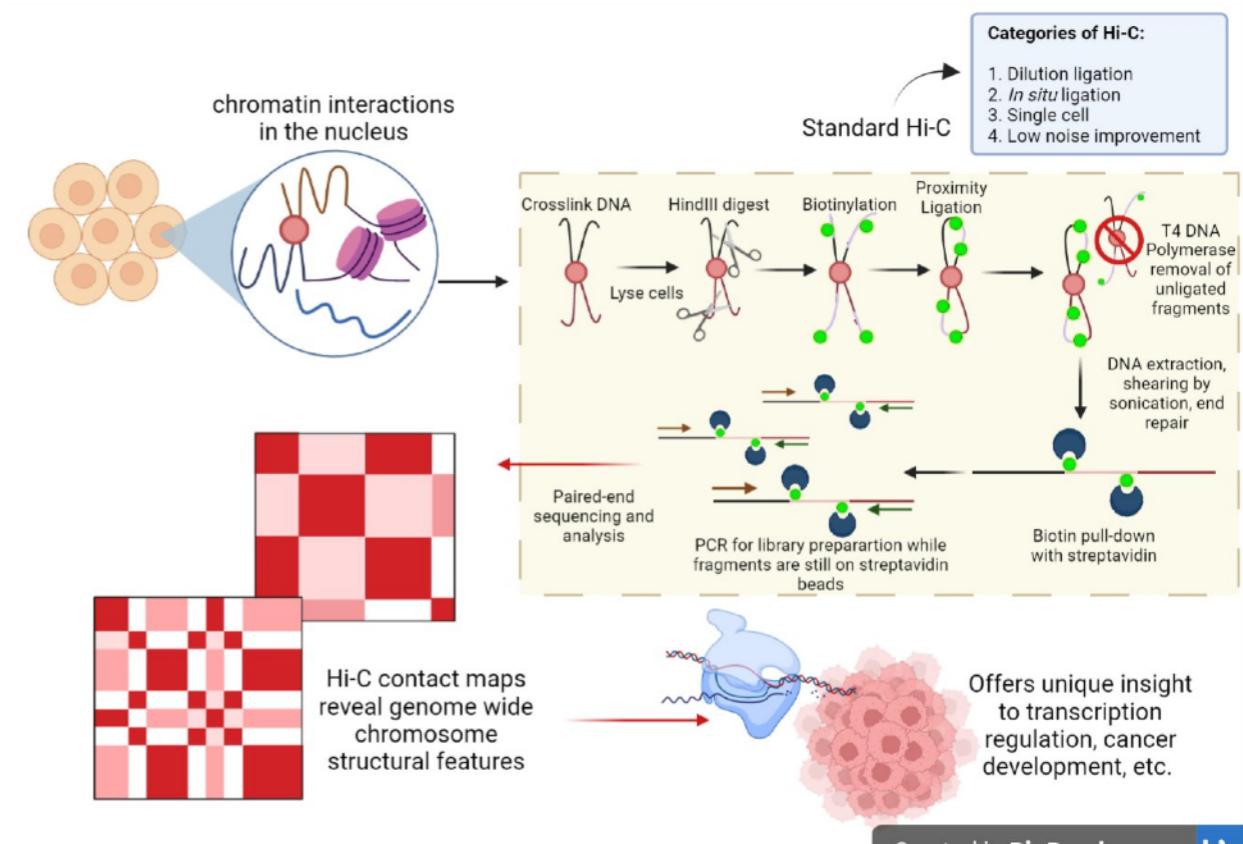
PacBio HiFi reads
*99.99% accurate
median length 17,000bp*



The dominant paradigm for genome assembly

In the last 5 years or so, approaches to genome assembly have largely coalesced around a small number of approaches

HiC short reads (>60x)
For contig scaffolding



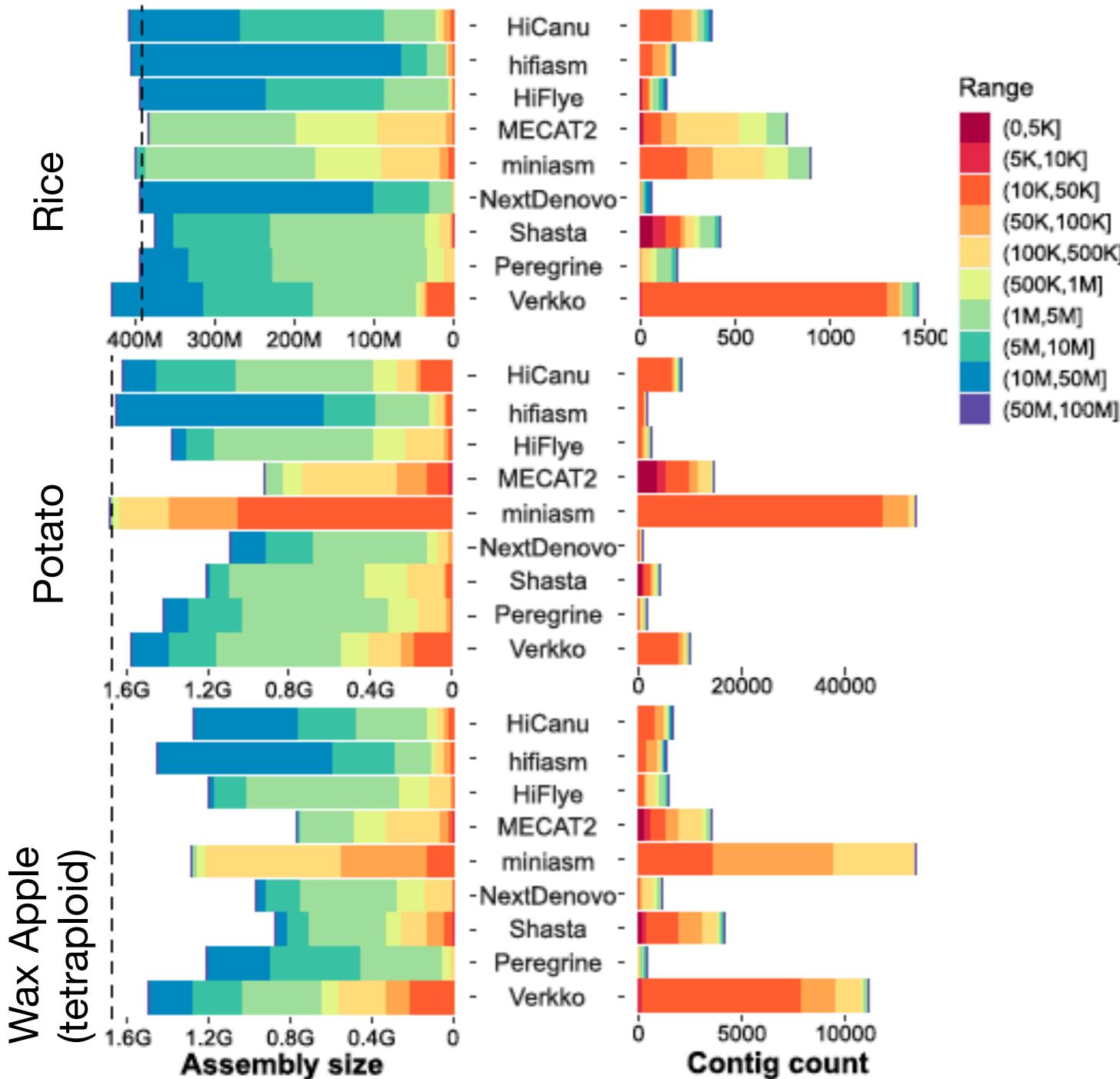
The dominant paradigm for genome assembly

In the last 5 years or so, approaches to genome assembly have largely coalesced around a small number of approaches

Phased Genome Assembly with HiFiasm

The screenshot shows the homepage of **nature methods**. At the top, there are navigation links: "View all journals", "Search", and "Log in". Below these are three dropdown menus: "Explore content", "About the journal", and "Publish with us". A red horizontal bar separates the header from the main content area. In the main area, the URL "nature > nature methods > articles > article" is shown. Below it, the text "Article | Published: 01 February 2021" is displayed. The title of the article is "**Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm**". The authors listed are [Haoyu Cheng](#), [Gregory T. Concepcion](#), [Xiaowen Feng](#), [Haowen Zhang](#) & [Heng Li](#) (with an envelope icon). The citation information is "Nature Methods 18, 170–175 (2021) | [Cite this article](#)". At the bottom, metrics are shown: "36k Accesses | 2100 Citations | 188 Altmetric | [Metrics](#)".

Why HiFiasm?



HiFiasm outperforms
produces genome
assemblies with longer,
more contiguous contigs

Resource

Comprehensive assessment of *de novo* HiFi assemblers on complex eukaryotic genomes and metagenomes

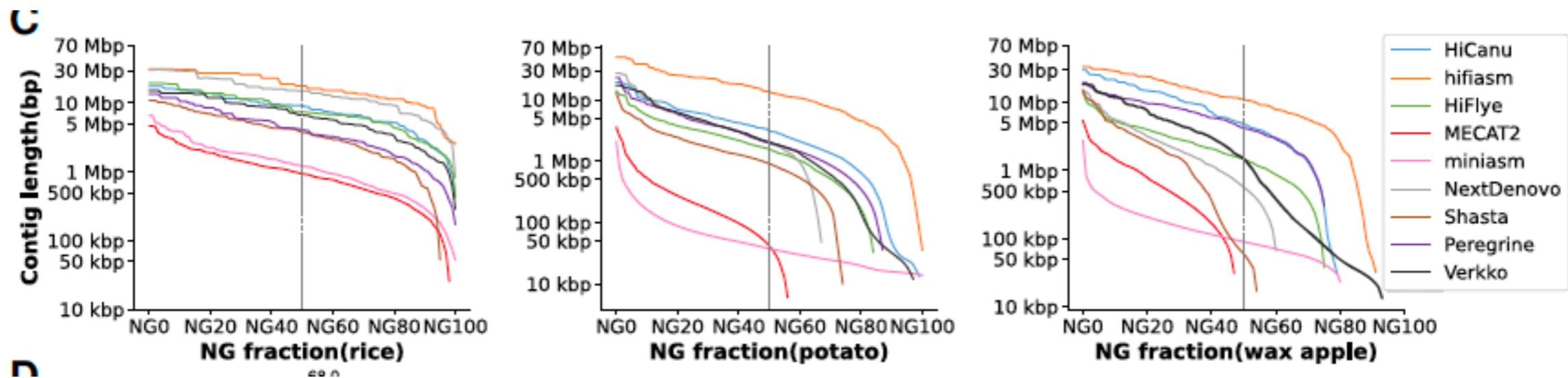
Wenjuan Yu,^{1,6} Haohui Luo,^{1,6} Jinbao Yang,^{1,5,6} Shengchen Zhang,^{1,5,6} Heling Jiang,^{1,6} Xianjia Zhao,^{1,4} Xingqi Hui,^{1,4} Da Sun,¹ Liang Li,² Xiu-qing Wei,² Stefano Lonardi,³ and Weihua Pan¹

¹Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China; ²Fruit Research Institute, Fujian Academy of Agricultural Sciences, Fuzhou, Fujian 350002, China; ³Department of Computer Science and Engineering, University of California, Riverside, California 92521, USA; ⁴School of Agricultural Sciences, Zhengzhou University, Zhengzhou, Henan 450001, China; ⁵College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

Modified Figure 1A

Why HiFiasm?

HiFiasm outperforms produces genome assemblies with longer, more contiguous contigs



Modified Figure 1C

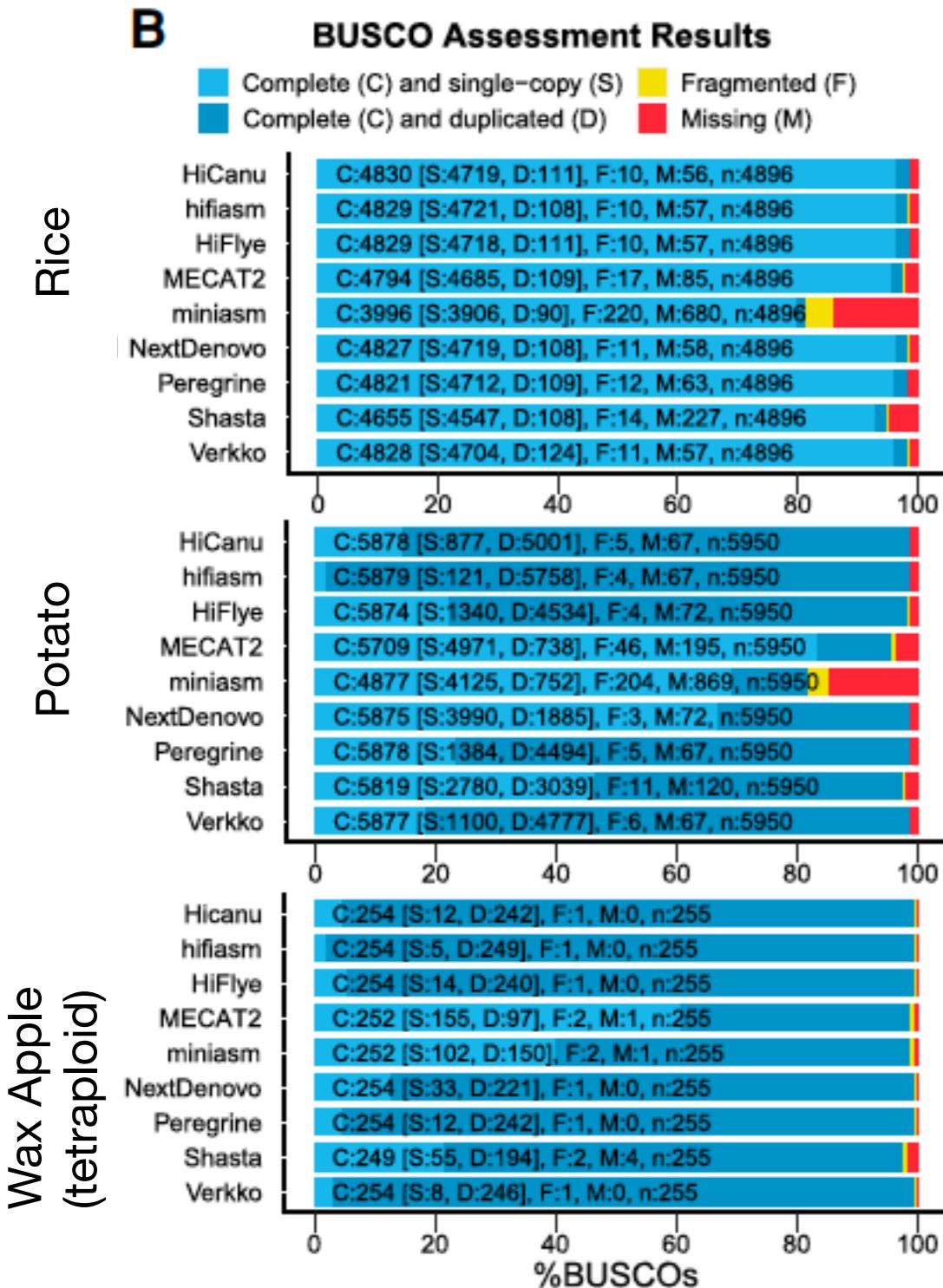
Resource

Comprehensive assessment of 11 de novo HiFi assemblers on complex eukaryotic genomes and metagenomes

Wenjuan Yu,^{1,6} Haohui Luo,^{1,6} Jinbao Yang,^{1,5,6} Shengchen Zhang,^{1,5,6} Heling Jiang,^{1,6} Xianjia Zhao,^{1,4} Xingqi Hui,^{1,4} Da Sun,¹ Liang Li,² Xiu-qing Wei,² Stefano Lonardi,³ and Weihua Pan¹

¹Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China; ²Fruit Research Institute, Fujian Academy of Agricultural Sciences, Fuzhou, Fujian 350002, China; ³Department of Computer Science and Engineering, University of California, Riverside, California 92521, USA; ⁴School of Agricultural Sciences, Zhengzhou University, Zhengzhou, Henan 450001, China; ⁵College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

Why HiFiasm?



HiFiasm outperforms
produces very complete
and accurate assemblies

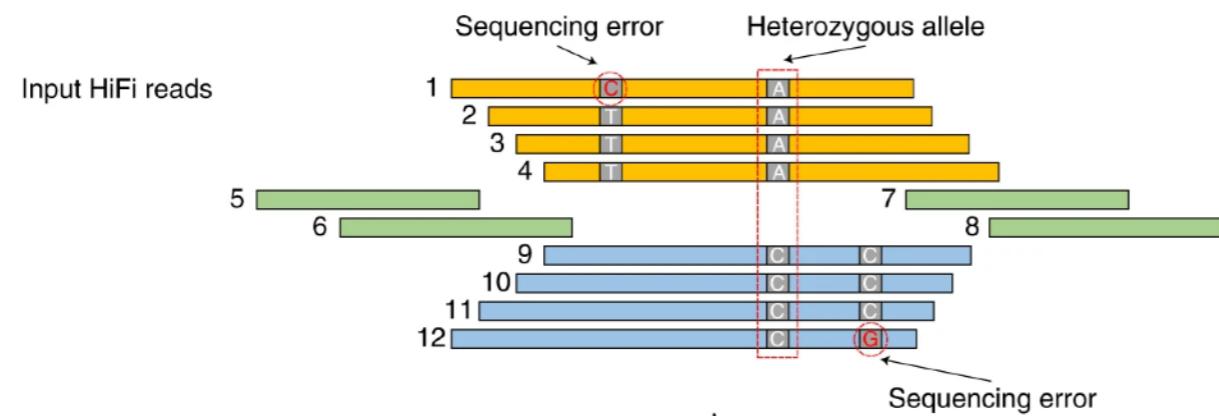
Resource

Comprehensive assessment of *de novo* HiFi assemblers on complex eukaryotic genomes and metagenomes

Wenjuan Yu,^{1,6} Haohui Luo,^{1,6} Jinbao Yang,^{1,5,6} Shengchen Zhang,^{1,5,6} Heling Jiang,^{1,6} Xianjia Zhao,^{1,4} Xingqi Hui,^{1,4} Da Sun,¹ Liang Li,² Xiu-qing Wei,² Stefano Lonardi,³ and Weihua Pan¹

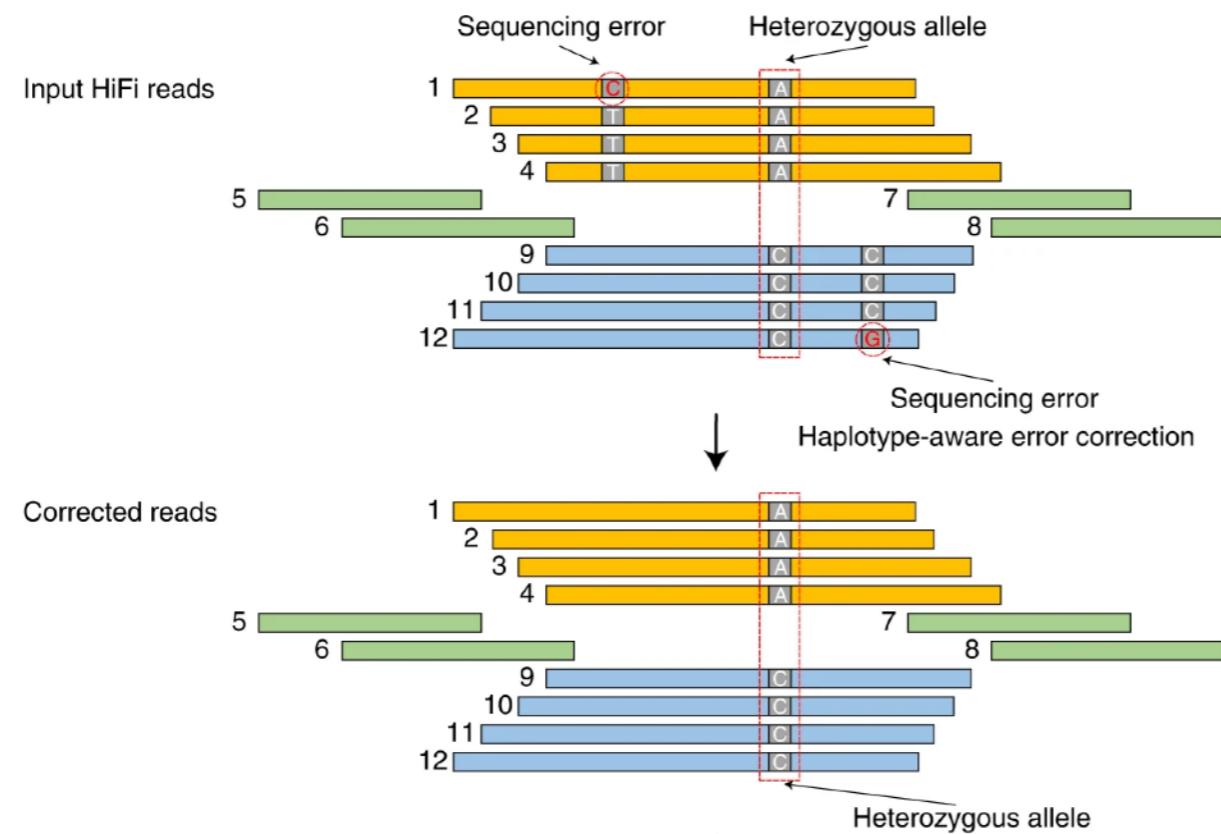
¹Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China; ²Fruit Research Institute, Fujian Academy of Agricultural Sciences, Fuzhou, Fujian 350002, China; ³Department of Computer Science and Engineering, University of California, Riverside, California 92521, USA; ⁴School of Agricultural Sciences, Zhengzhou University, Zhengzhou, Henan 450001, China; ⁵College of Informatics, Huazhong Agricultural University, Wuhan 430070, China

Overview of HiFiAsm

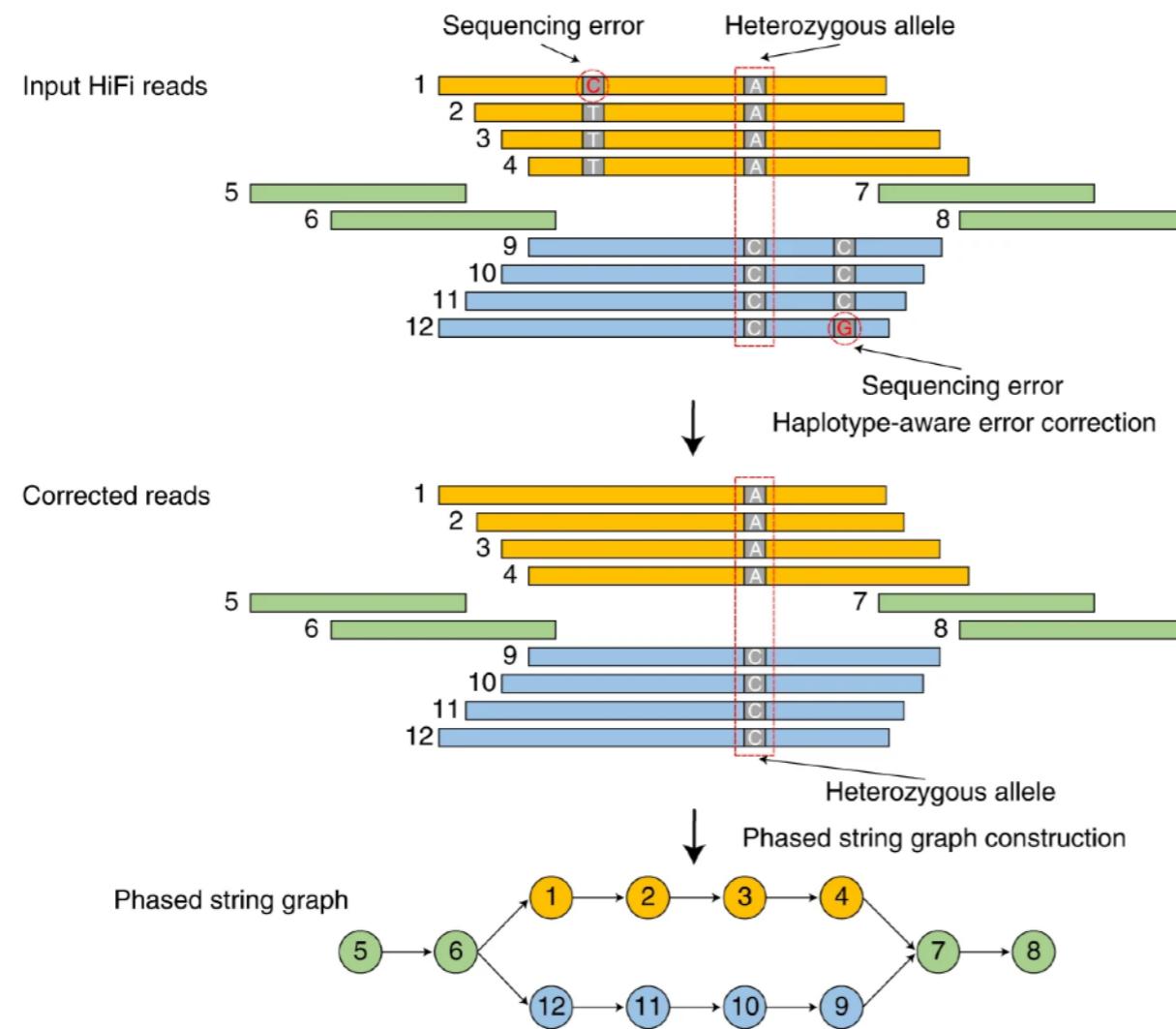


Align reads to themselves (*à la OLC*)

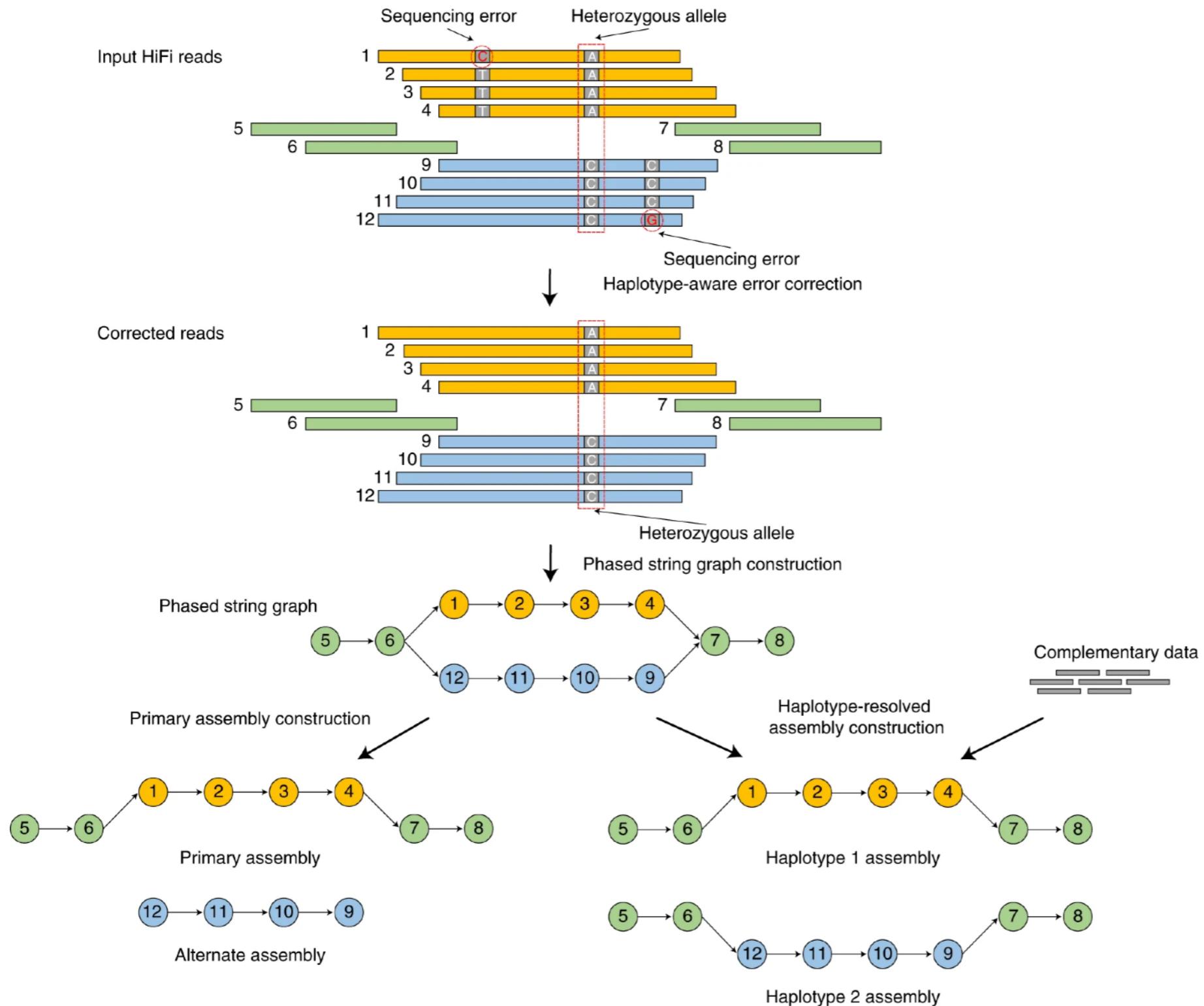
Overview of HiFiAsm



Overview of HiFiAsm



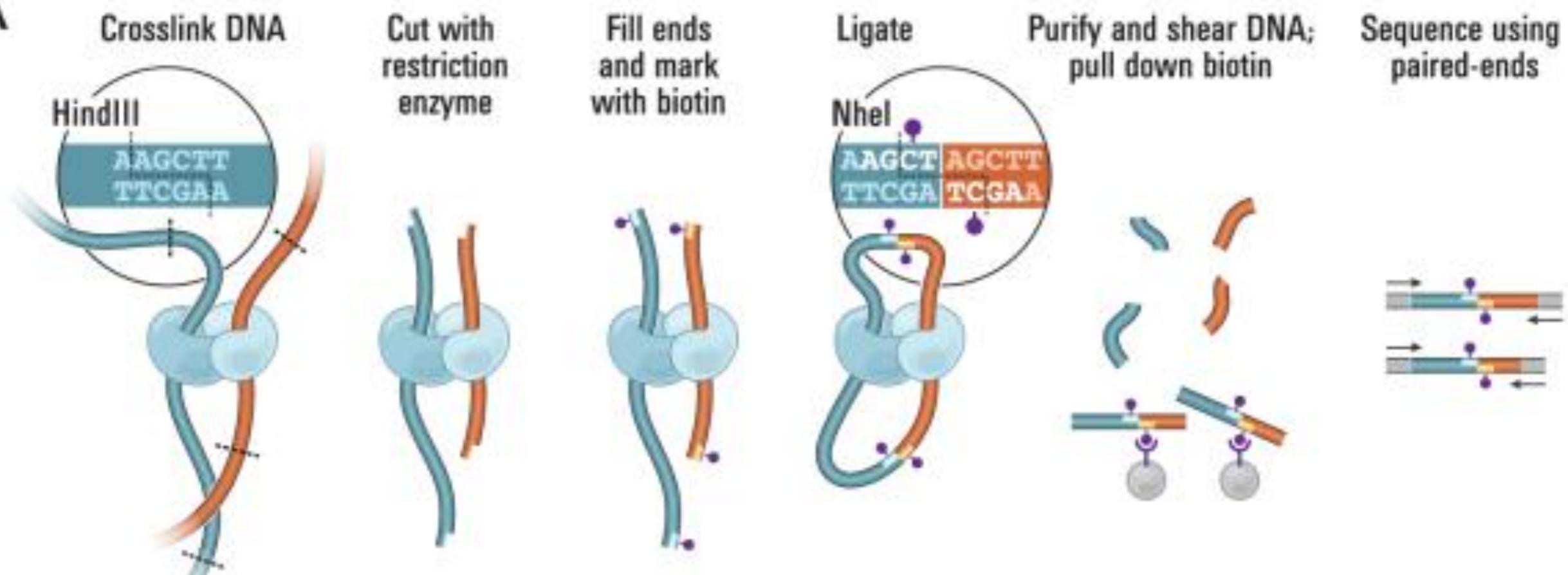
Overview of HiFiAsm



Scaffolding

Why HiC?

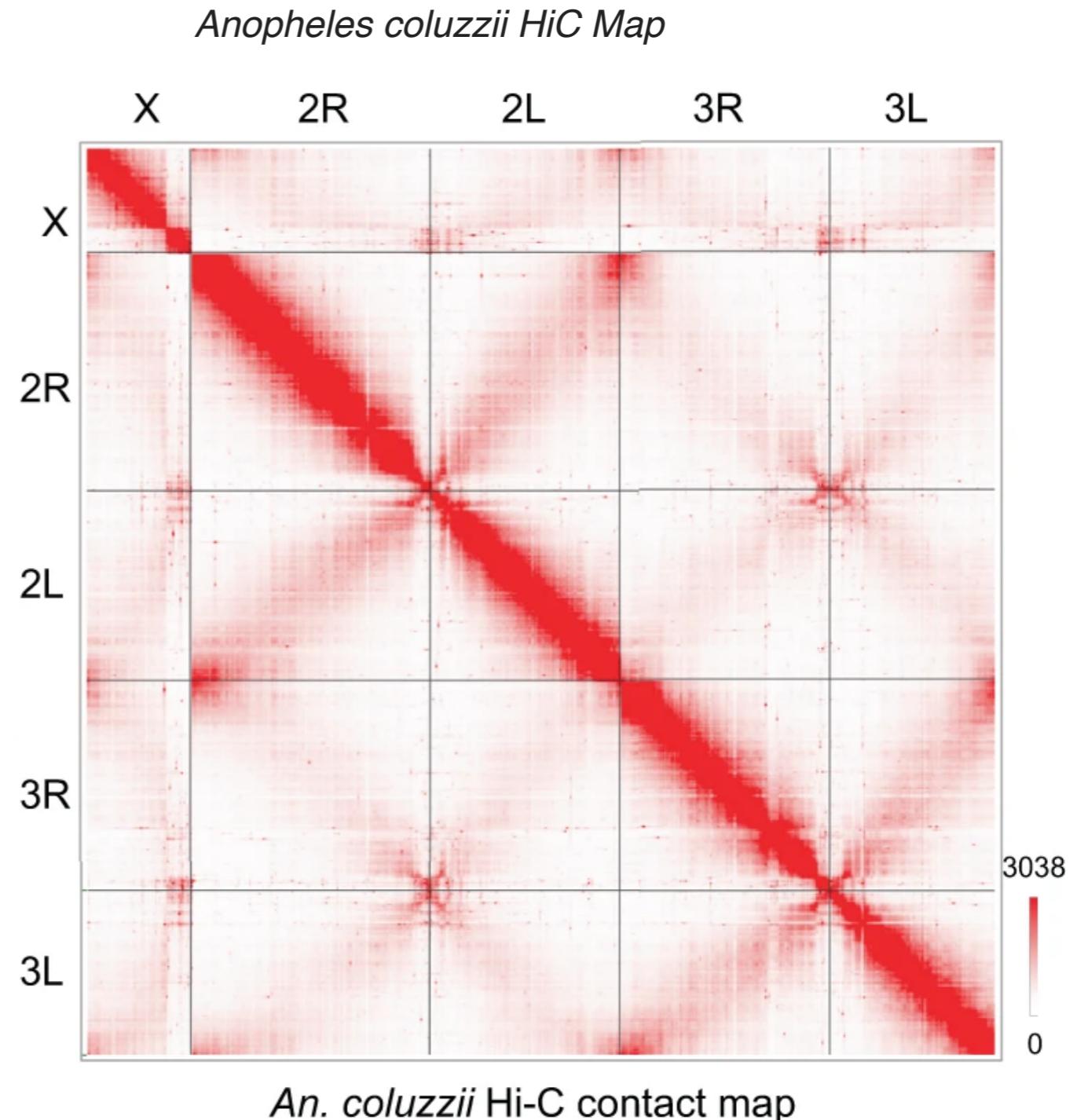
A



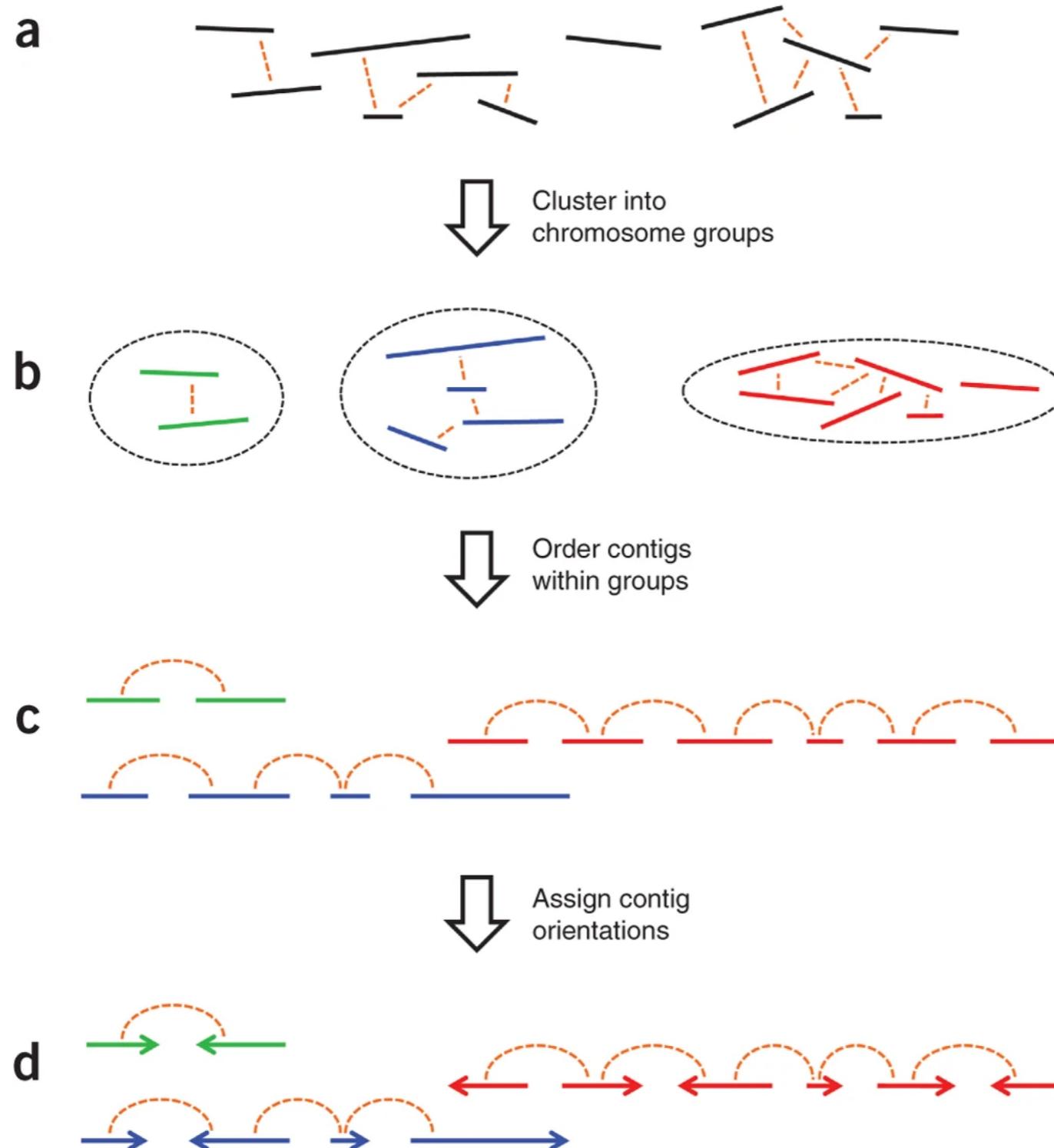
Why HiC?

HiC captures the 3D structure of chromosomes

Long range interactions between regions on the same chromosome can be identified



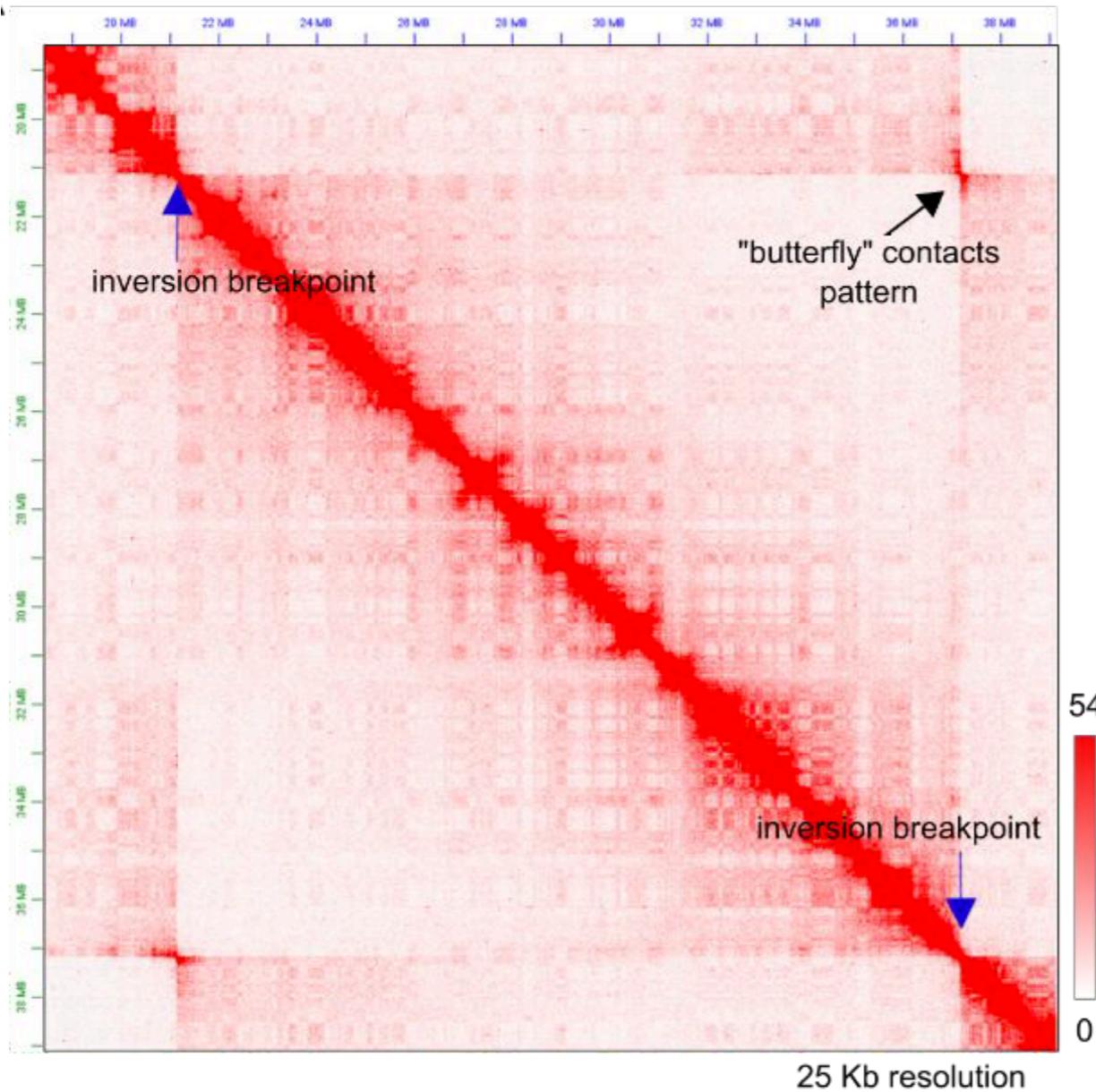
HiC for scaffolding



The long range information in HiC data can be used to assign contigs to chromosomes, place them in order and orient them

Ultimately allowing contigs to be placed into scaffolds

Why HiC?

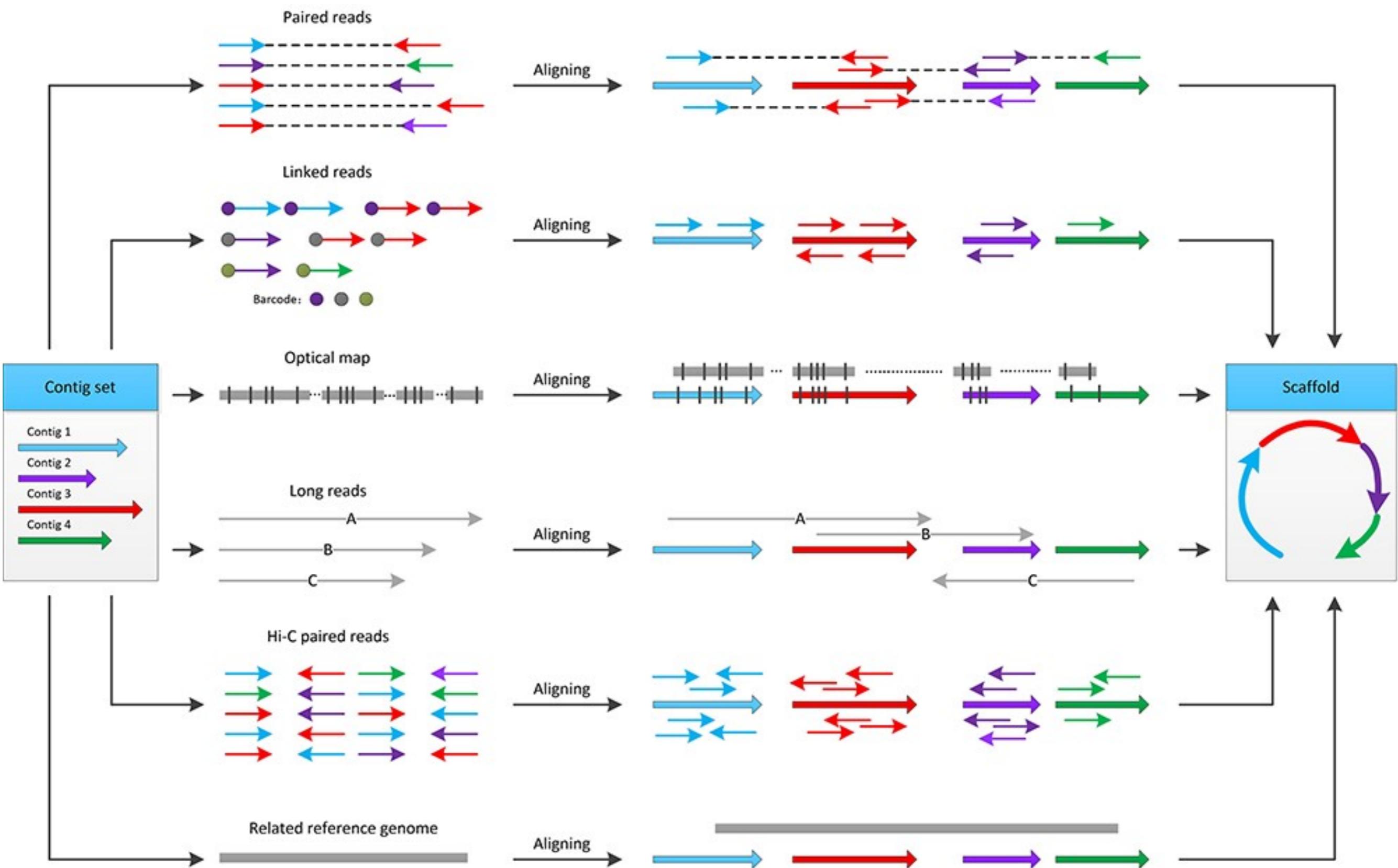


Anopheles stephensi, Chr 2R

HiC can also be used to identify chromosomal inversions!

(*If they are heterozygous in the sample*)

Other methods for scaffolding



How good is my assembly?

Important Questions

How much total sequence is in the assembly relative to the estimated genome size?

How many pieces did I assemble, and what is their size distribution?

Are the contigs assembled correctly?

Are the contigs scaffolded in the right order and orientation?

How were repeats handled?

Are all the genes I expected in my assembly?

Evaluating Genome Assembly

Many statistics have been devised to assess assembly quality

The EBP uses the metrics in the following table...

Table 1. | Proposed standards and metrics for defining genome assembly quality

Quality Category	Quality Metric	Finished	7.C.Q50	6.7.Q40	4.5.Q30	VGP
Continuity	Contig (NG50)	= Chr. NG50	>10 Mbp	>1 Mbp	>10 kbp	1-25 Mbp
	Scaffolds (NG50)	= Chr. NG50	= Chr. NG50	>10 Mbp	>100 kbp	23-480 Mbp
	Gaps / Gbp	No gaps	<200	<1,000	<10,000	75-1500
Structural accuracy	False duplications	0%	<1%	<5%	<10%	0.2-5.0%
	Reliable blocks	= Chr. NG50	>90% of Scaffold NG50	>75% of Scaffold NG50	>50% of Scaffold NG50	2-75%
	Curation improvements	All conflicts resolved	Automated + Manual	Automated	No requirement	Automated + Manual
Base accuracy	Base pair QV	>60	>50	>40	>30	39-43
	k-mer completeness	100% complete	>95%	>90%	>80%	87-98%
Haplotype phasing	Phased block (NG50)	= Chr. NG50	>1 Mbp	>100 kbp	No requirement	1.6 Mbp*
Functional completeness	Genes	>98% complete	>95% complete	>90%	>80%	82-98%
	Transcript mappability	98%	>90%	>80%	>70%	96%
Chromosome status	Assigned %	98%	>90%	>80%	No requirement	94.4-99.9%
	Sex chromosomes	Right order, no gaps	Localized homo pairs	At least 1 shared (e.g. X or Z)	Fragmented	At least 1 shared
	Organelles (e.g. MT)	1 Complete allele	1 Complete allele	Fragmented	No requirement	1 Complete allele

Evaluating Genome Assembly

The EBP outlines a target of 6.C.Q40 for all new genome assemblies - this means...

Contig NG50>10Mbp

Scaffold NG50 = ChrNG50

Base pair QV>50

To achieve this, they recommend a combination of longread data and long range scaffolding information

For example PacBio HiFi reads + HiC data

Tutorial:
Work through the tutorial associated with this session