Genome Assembly

CHRIS FIELDS, HPCBIO

MAYO-ILLINOIS COMPUTATIONAL GENOMICS WORKSHOP
JUNE 23, 2025

Ideal World!

I have this joke slide (thx to Torsten Seemann) on all my past talks…

Human DNA

Sample

iSequencer™

AGTCTAGGATTCGCTACAGAT
TCAGGCTCTGAAGCTAGATCG
CTATGCTATGATCTAGATCTC
GAGATTCGTATAAGTCTAGGA
TTCGCTATAGATTCAGGCTCT
GATATAT

46 complete, haplotype-resolved, chromosome sequences

T. Seemann

Ideal World!

We may not be too far from this now.

HOME > SCIENCE > VOL. 376, NO. 6588 > THE COMPLETE SEQUENCE OF A HUMAN GENOME

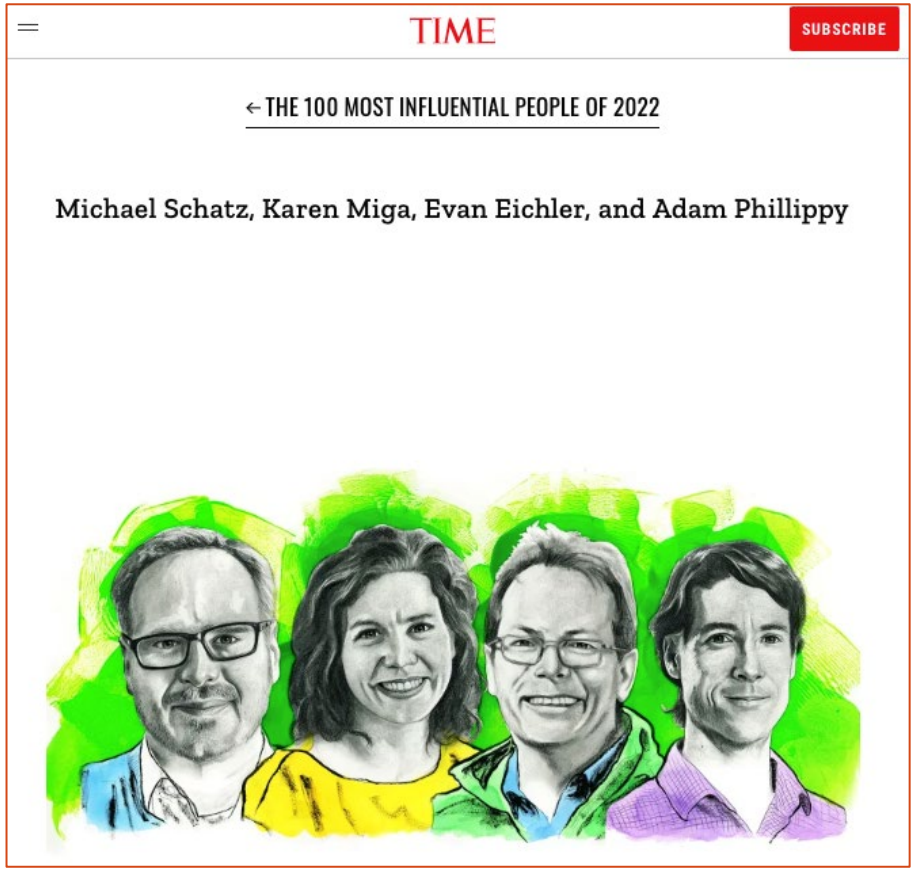
SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

The complete sequence of a human genome

SERGEY NURK, SERGEY KOREN, ARANG RHIE, MIKKO RAUTIAINEN, ANDREY V. BZIKADZE, ALLA MIKHEENKO, MITCHELL R. VOLLGER, NICOLAS ALTEMOSE, LEV URALSKY, [...] ADAM M. PHILLIPPY +91 authors Authors Info & Affiliations

SCIENCE · 31 Mar 2022 · Vol 376,Issue 6588 · pp.44-53 · DOI: 10.1126/science.abj6987

Science, March 2022



TIME SUBSCRIBE

← THE 100 MOST INFLUENTIAL PEOPLE OF 2022

Michael Schatz, Karen Miga, Evan Eichler, and Adam Phillippy

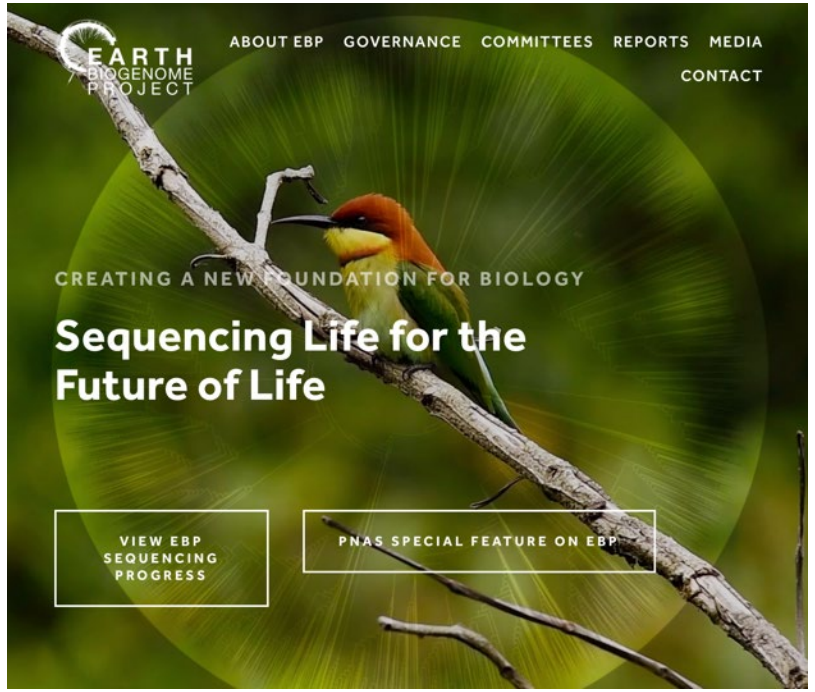
Time, May 2022

Ideal World!

Earth Biogenome Project

Pangenomics

Announced 2018,
started early 2022



EBP website

Nature,
May 2023

HPRC Nature issue

Current Sequencing Technologies

Illumina

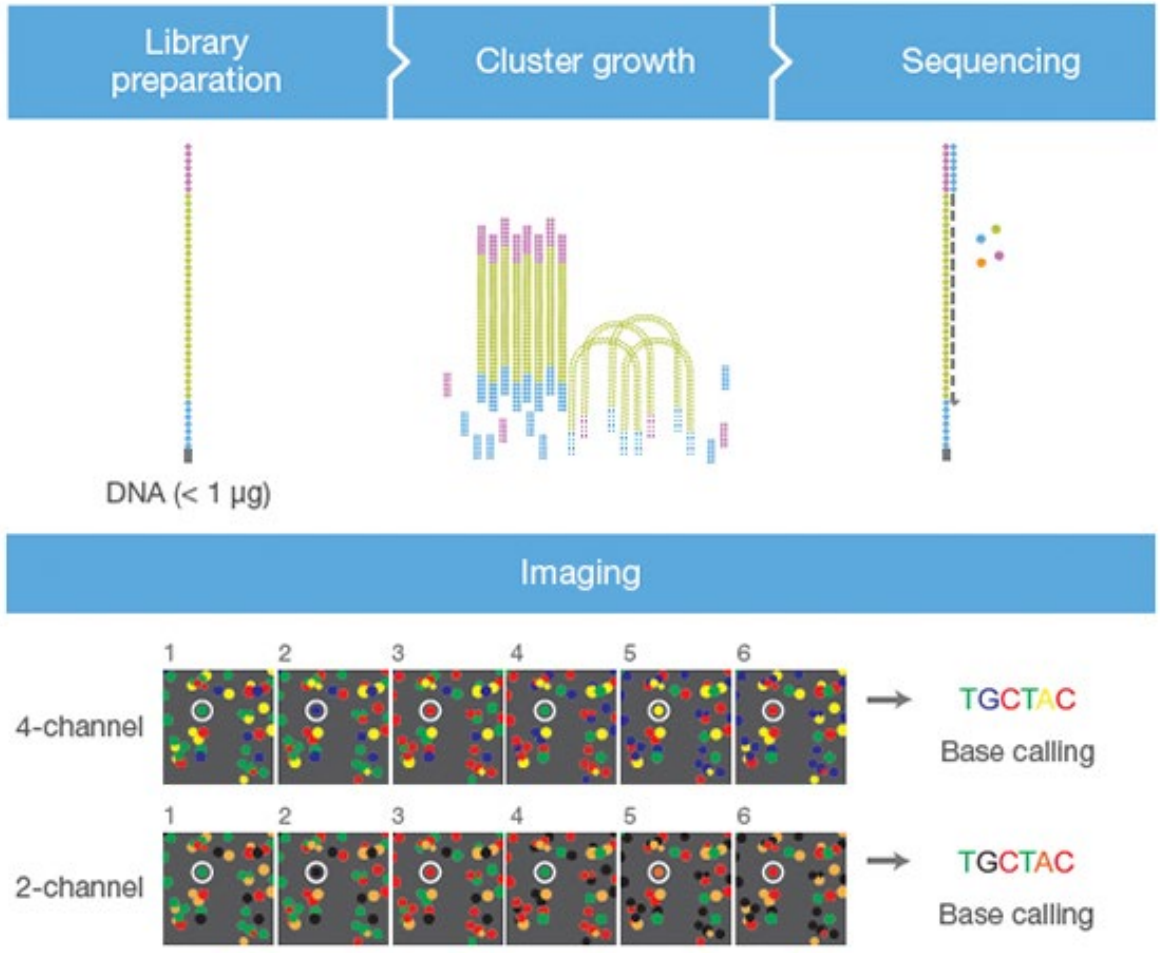
Millions to billions of short but highly accurate reads (>99.9%)

Can be paired-end (sequence ends of fragments)

Flow-cell based

Advantages

- *Highly accurate* (~99.9%)
- Relatively even coverage of the genome
- Well-vetted technology
- Most cost-effective, as low as $10 per **billion** bases
- Robust to sample issues



Illumina information

Illumina

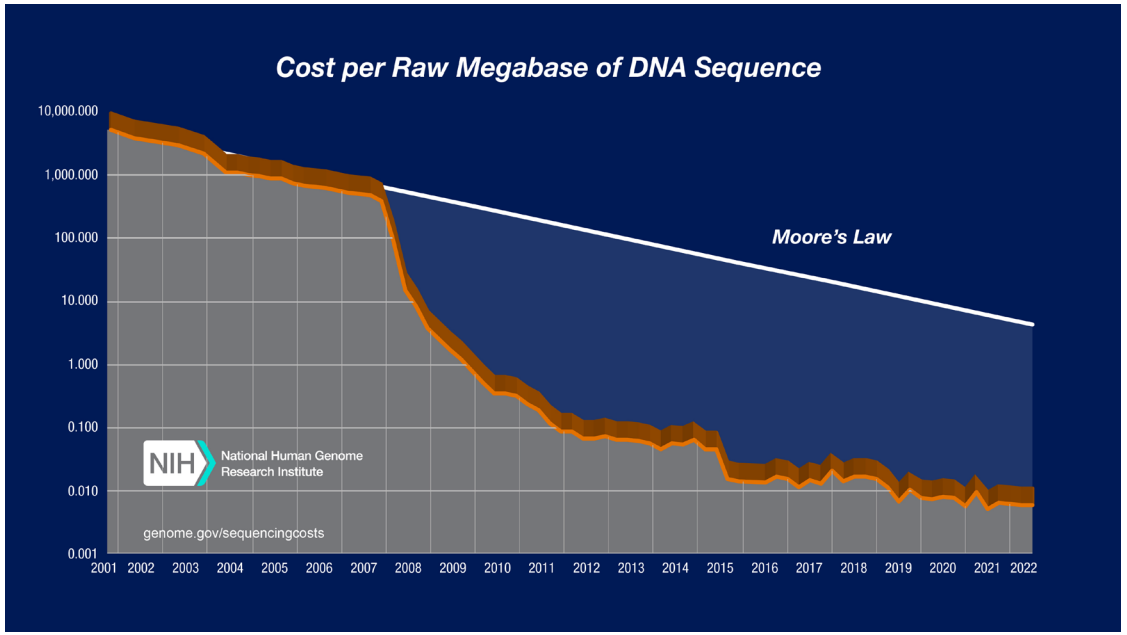
Millions to billions of short but highly accurate reads (>99.9%)

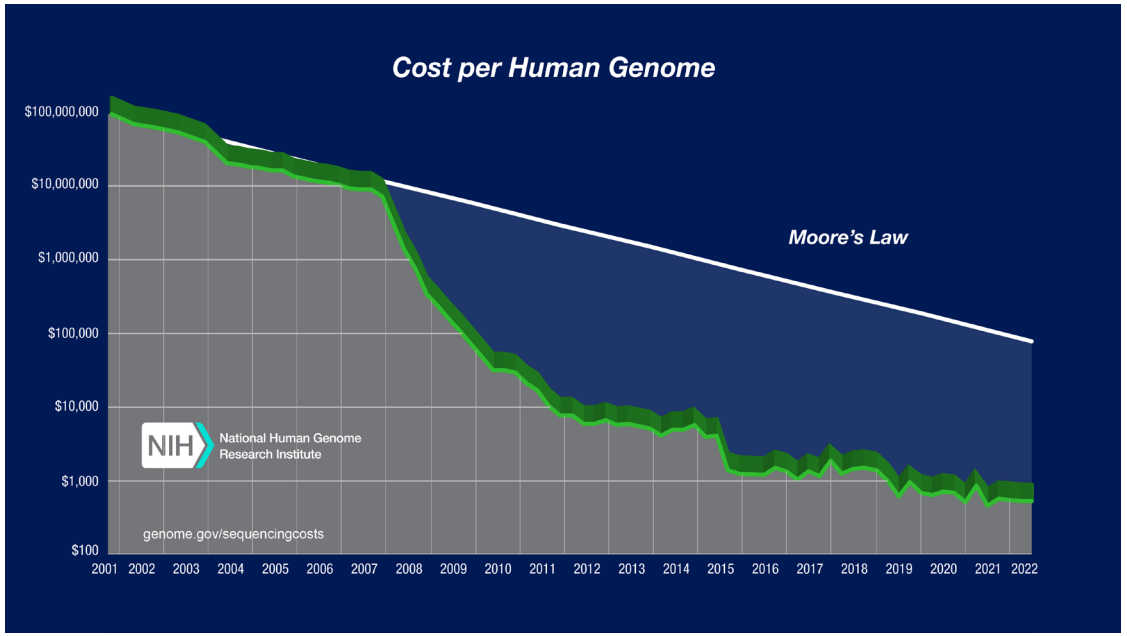
Can be paired-end (sequence ends of fragments)

Flow-cell based

Disadvantages

- Sequence length is ~150-300nt
- Requires high depth (>50x)
- Fragment size is a problem (<1000nt)

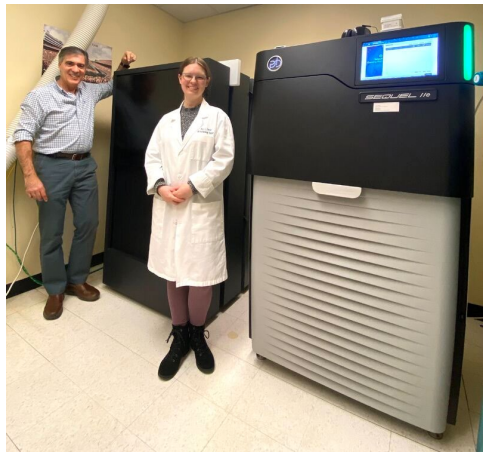




BINGO!

‘Long reads’

Pacific Biosciences (PacBio)

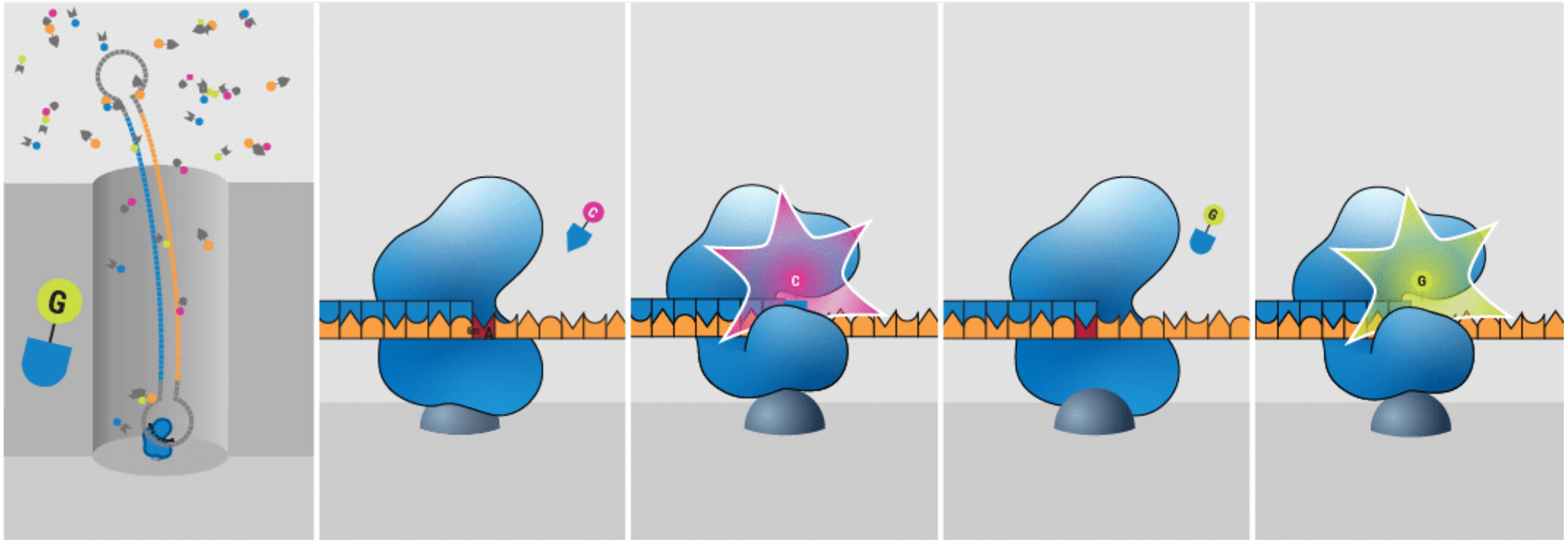


Oxford Nanopore (ONT)

MinION

Pacific Biosciences

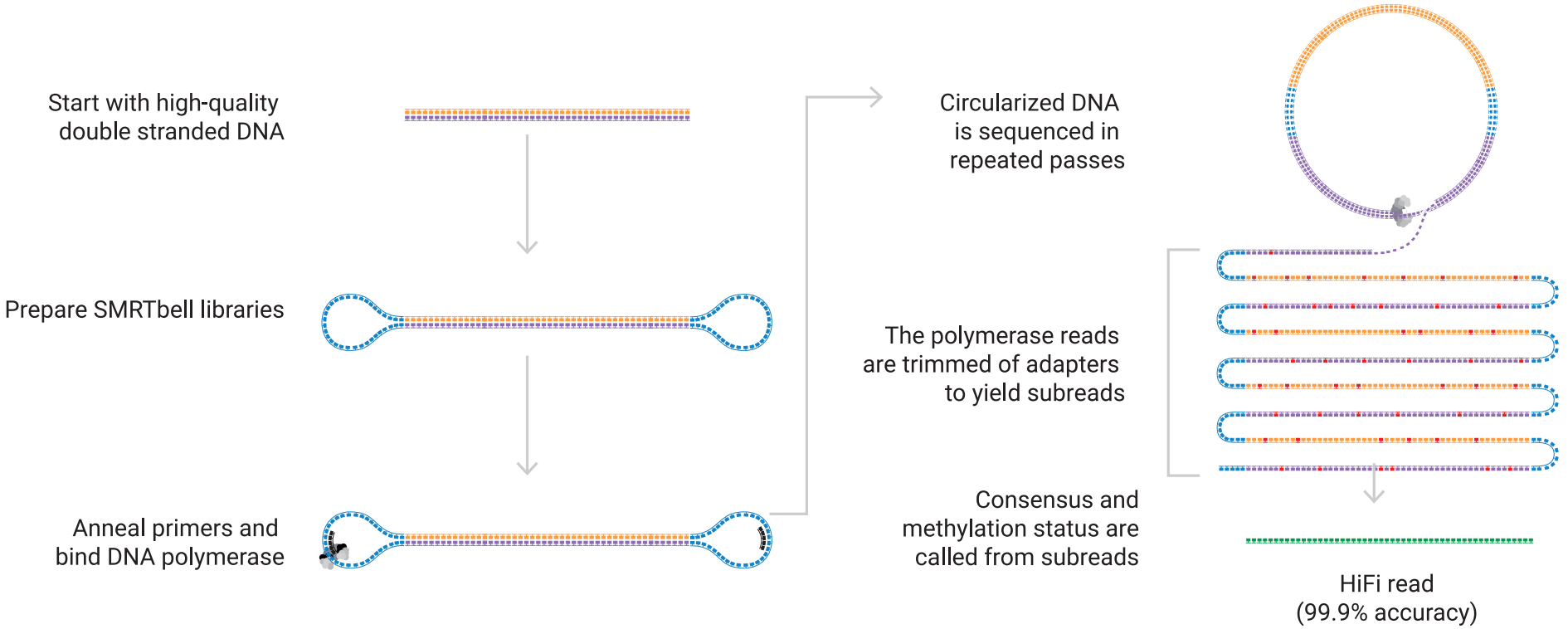
Oxford Nanopore



Pacific Biosciences

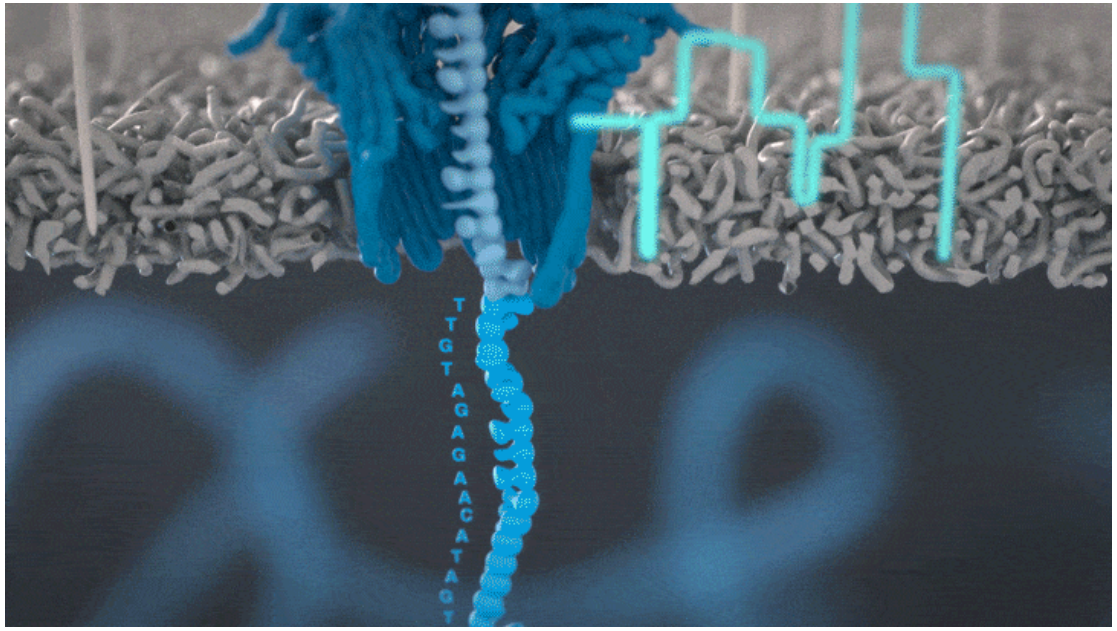
RHOADS AND AU, GENOMICS, PROTEOMICS & BIOINFORMATICS, 13(5), OCT 2015

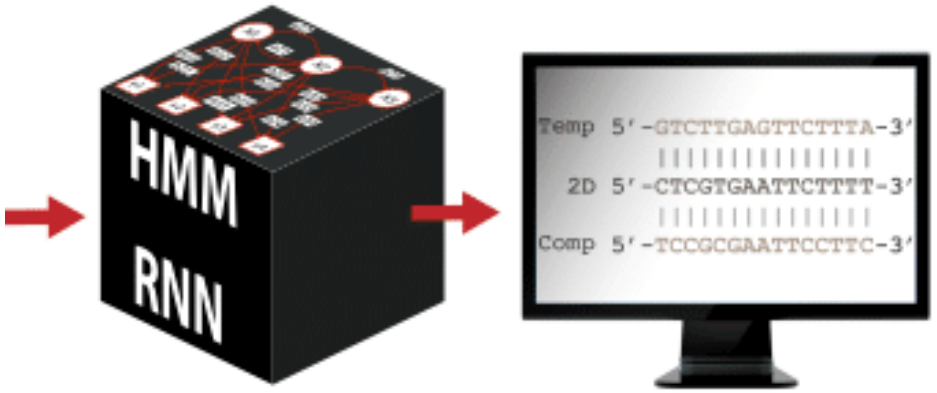
Pacific Biosciences



PacBio Circular Consensus Sequencing (CCS) - aka PacBio HiFi

Pacific Biosciences





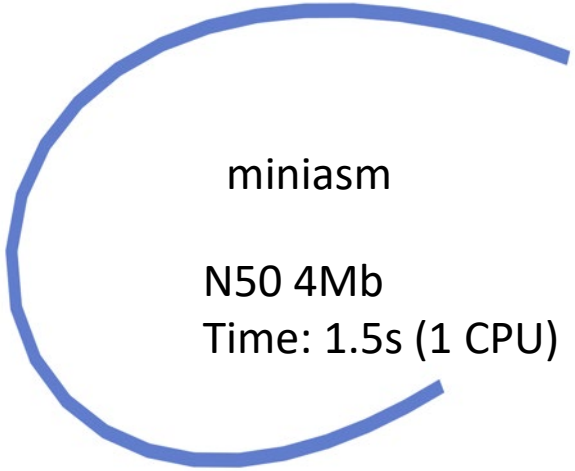
ONT

Alberto Magi *et al*, *Briefings in Bioinformatics*, Volume 19, Issue 6, November 2018

Oxford Nanopore

E. coli: genome assembly in 8 reads

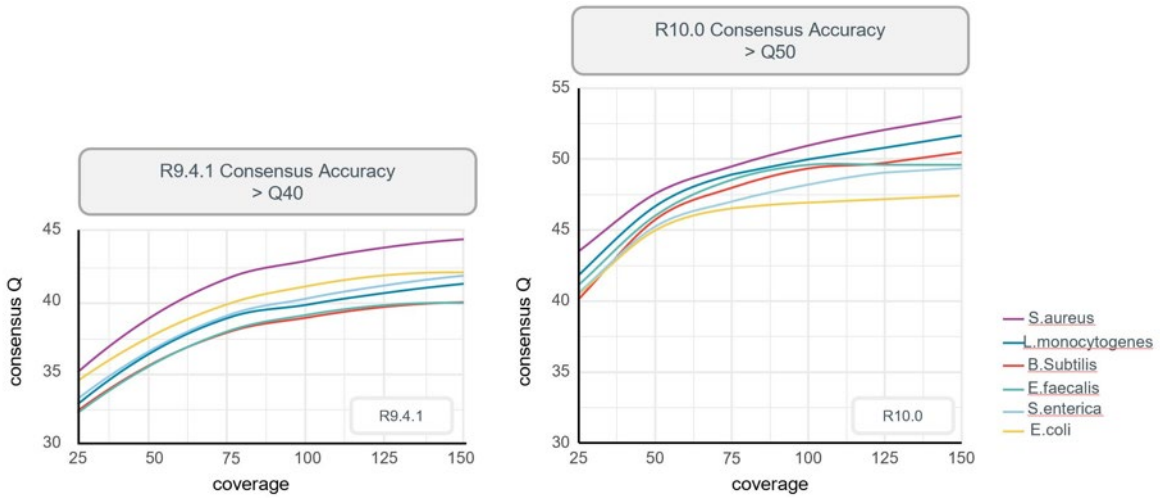
Read	Length	Ref start	Ref end	Time (m)
1	876991	4398844	634183	32.48
2	696402	470003	1166405	25.79
3	799047	1137438	1936485	29.59
4	642071	1759431	2401502	23.78
5	826662	2106227	2932889	30.61
6	883962	2699626	3583588	32.73
7	825191	3285196	4110387	30.56
8	463341	3995967	4459308	17.16



1x coverage!

Oxford Nanopore

2021 – New flow cells (R10), kits



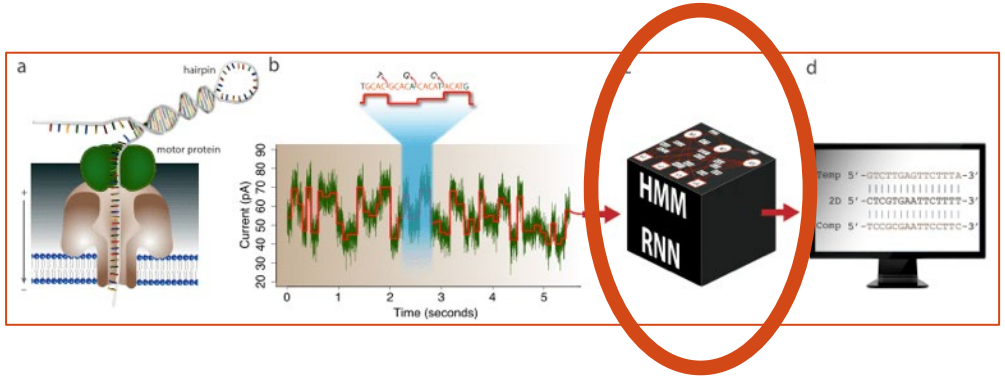
Methodology | Open Access | Published: 14 December 2022

Species-specific basecallers improve actual accuracy of nanopore sequencing in plants

Scott Ferguson, Todd McLay, Rose L. Andrew, Jeremy J. Bruhl, Benjamin Schwessinger, Justin Borevitz & Ashley Jones

Plant Methods **18**, Article number: 137 (2022) | Cite this article

2609 Accesses | **2** Citations | **8** Altmetric | Metrics



Oxford Nanopore

Long Reads

Advantages

- Dependent on technology; can be very long (1kb – 100kb)
- Relatively even coverage of the genome
- PacBio HiFi, ONT using latest release (duplex) - Highly accurate (99%)
- PacBio HiFi, ONT - DNA modifications (RNA mods for ONT)
- ONT - real-time sequencing; portable; direct RNA seq

Disadvantages

- Expensive compared to Illumina short reads
- Need very high quality, high MW DNA samples
- Depending on technology, can have *systematic errors* (homopolymer issues)

Regardless of the disadvantages: ***Use long reads for genome assembly***

Genome assembly steps

(a) **Collect DNA** – samples are fragmented and sequenced.

(b) **Sequence** – generate millions/billions of unordered DNA fragments from random positions in the genome.

(c) **Compare** – how do sequence fragments connect with one another

(d) **Graph** – capture relationships in a large *assembly graph*

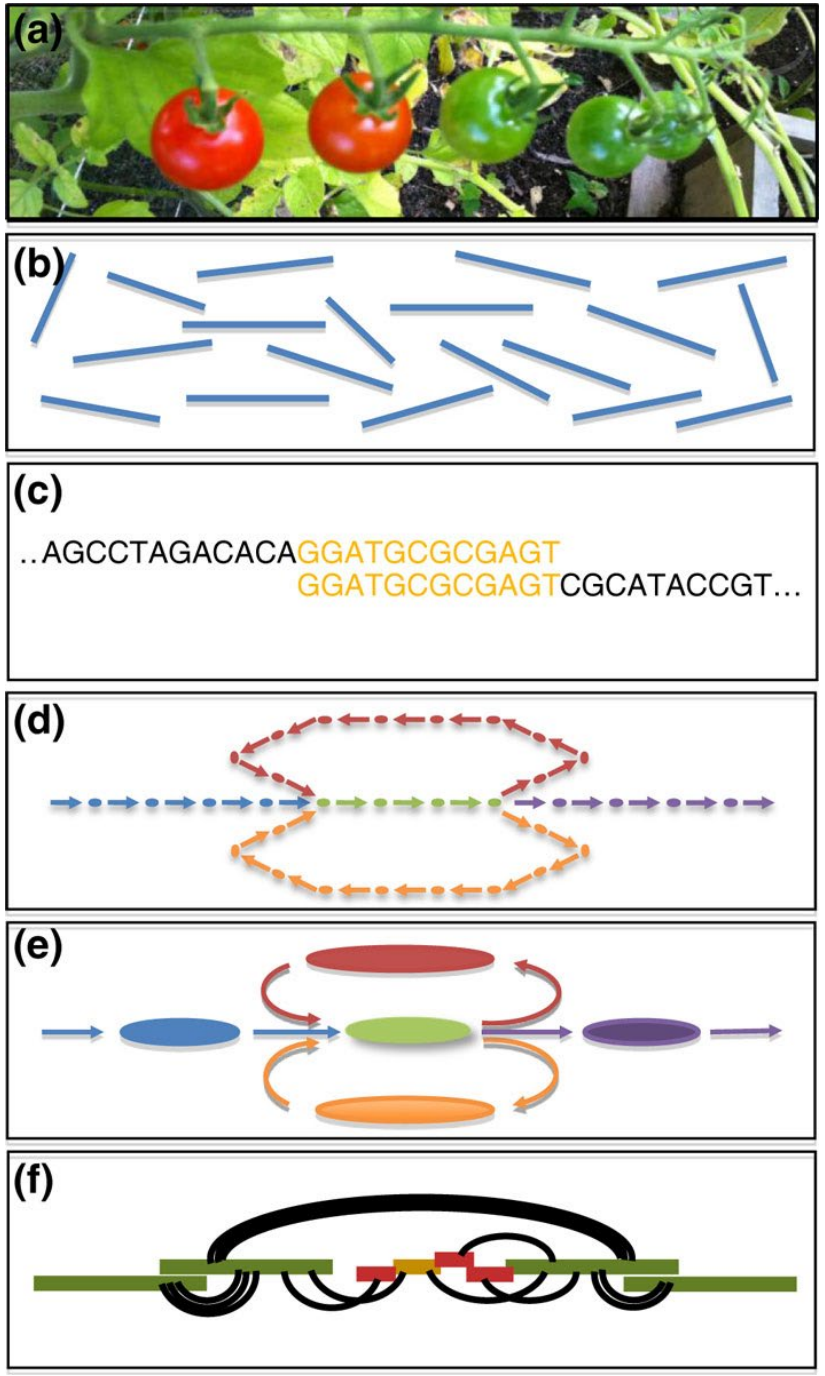
(e) **Simplify-** The assembly graph is refined to correct errors and simplify

(f) **Scaffold** – Use long reads, mates, markers, other long-range information to order/orient assembly (**contigs**) into large **scaffolds**

(g) **Clean** – resolve artifacts, remove contaminants, check gene completeness, contiguity, etc

(h) **Annotate –** Add features to the genome. Don't forget RNA if you want to predict genes, preferably from a broad range of tissues/conditions

Schatz et al. Genome Biology 2012 13:243



Steps

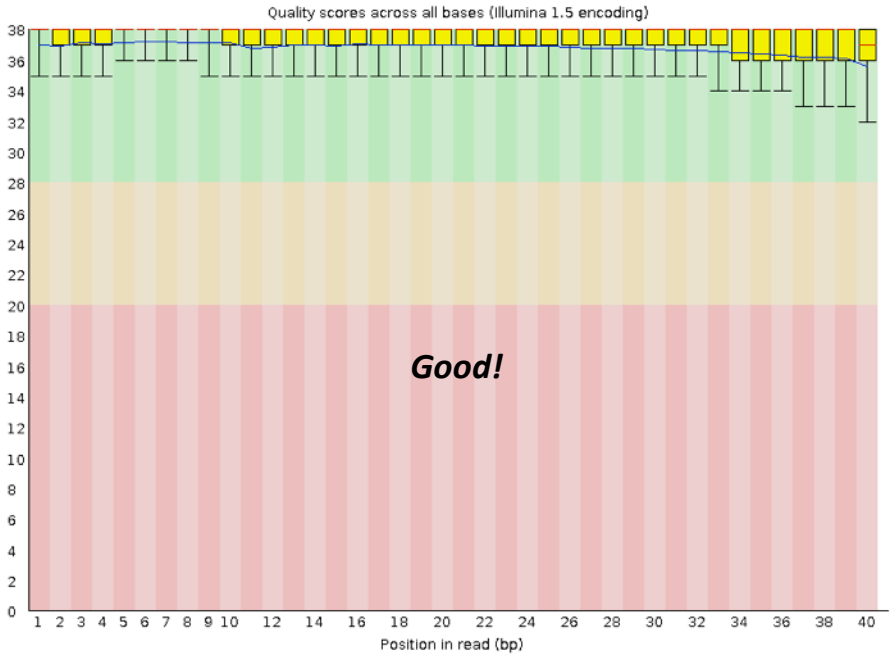
- Basic DNA sequence cleanup and evaluation (pre-assembly)
- Contig building
- Scaffolding
- Post-assembly processing and analyses

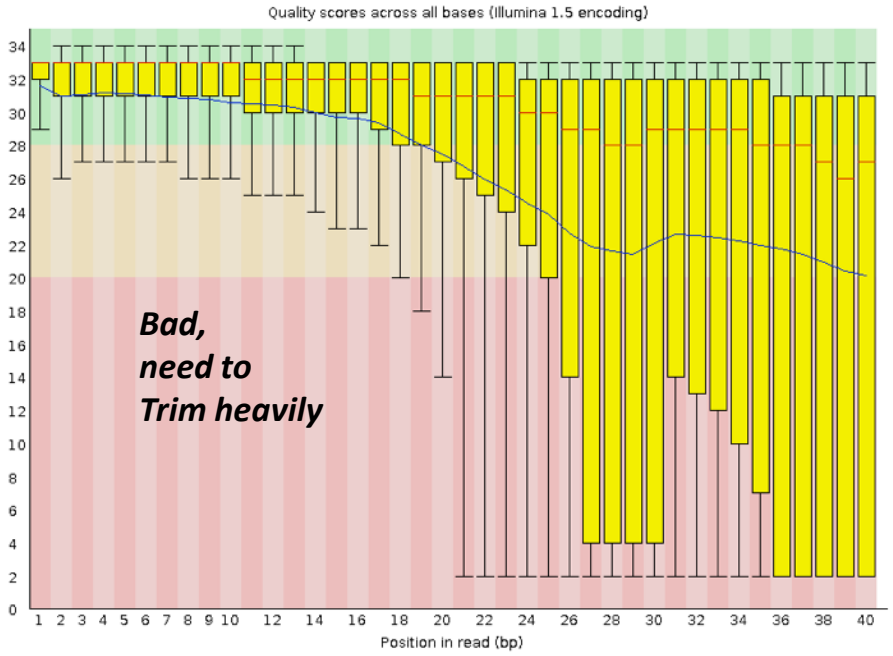
Basic cleanup and evaluation

- Is the DNA sequence high quality?
- Does it need to be trimmed?
- Evaluate libraries for read ‘coverage’
- Any additional sequence preparation steps

DNA Quality (FASTQC)

Illumina Data





https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Adapters

Overrepresented sequences

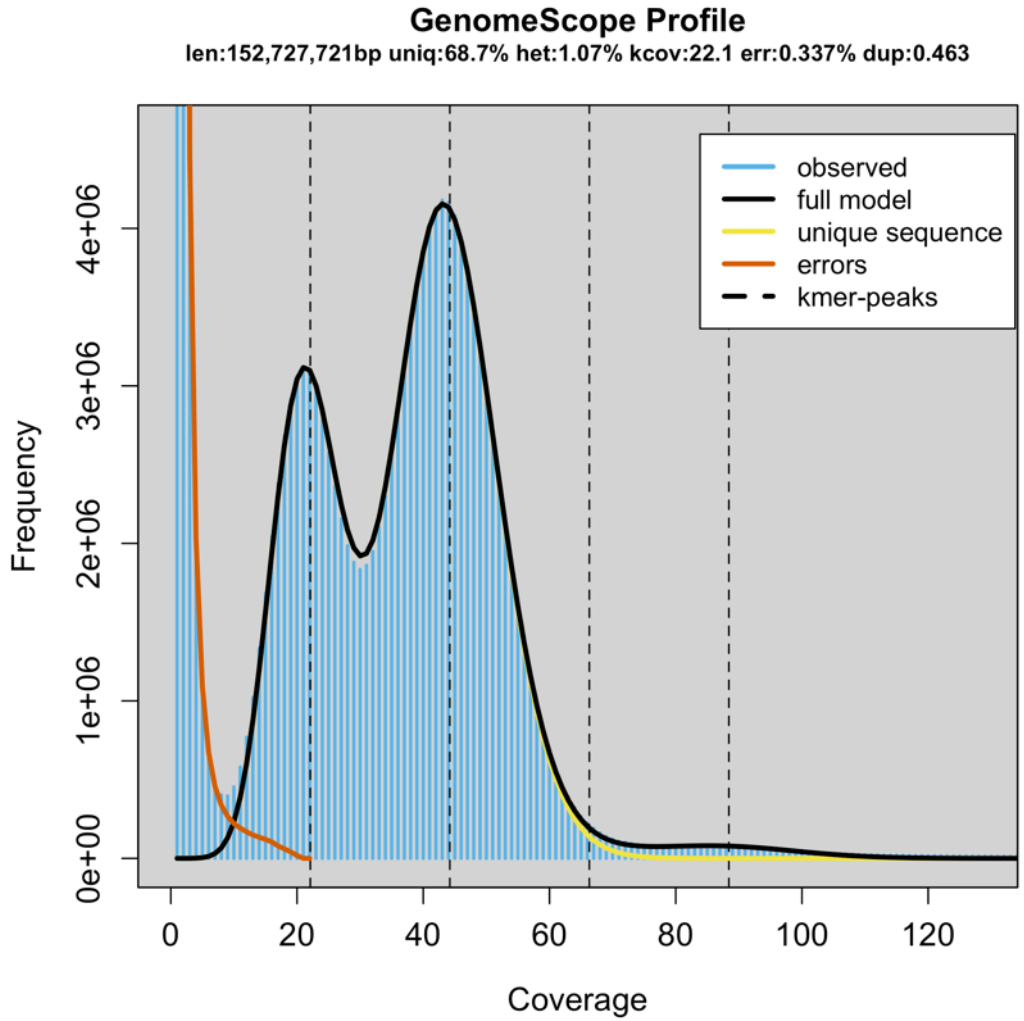
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	8122	8.122	Illumina Paired End PCR Primer 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGATCGGAAG	5086	5.086	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	1085	1.085	Illumina Single End PCR Primer 1 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGAAG	508	0.508	Illumina Paired End PCR Primer 2 (97% over 36bp)
AATTATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	242	0.242	Illumina Single End PCR Primer 1 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAAGATCGGAA	235	0.23500000000000001	Illumina Paired End Adapter 2 (96% over 31bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCGAGATCGGAAGA	228	0.22799999999999998	Illumina Paired End Adapter 2 (96% over 28bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACG	205	0.20500000000000002	Illumina Paired End PCR Primer 2 (97% over 36bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGATCGGAA	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGGTCGGAAG	183	0.183	Illumina Paired End Adapter 2 (100% over 32bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGAACT	164	0.164	Illumina Paired End PCR Primer 2 (97% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGTCT	129	0.129	Illumina Paired End PCR Primer 2 (97% over 40bp)
AATTATACTTCTACCACCTATATCTACACTCTTTCCCTAC	123	0.123	No Hit
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGGACT	122	0.122	Illumina Paired End PCR Primer 2 (97% over 36bp)
CGGTTCAGCAGGAATGCCGAGATCGGAAGAGCGGTTCAGC	113	0.11299999999999999	Illumina Paired End PCR Primer 2 (96% over 25bp)

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

Coverage

- Requires highly accurate reads
  - Illumina
  - PacBio HiFi
- Kmer read distribution

Arabidopsis F1 cross



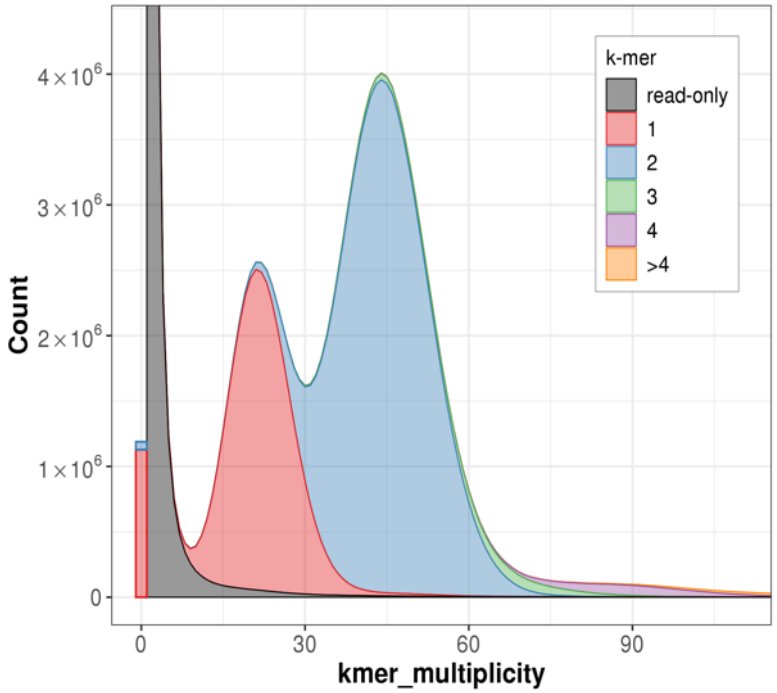
Genomescope

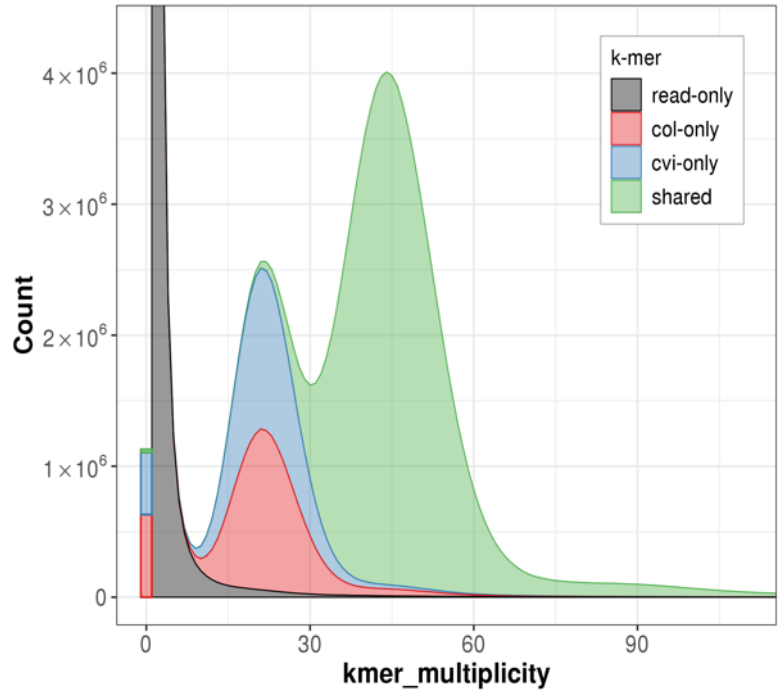
Rannallo T, Jaron K, Schatz M, Nature Comm, 1432(2020)

Coverage

- Requires highly accurate reads
  - Illumina
  - PacBio HiFi
- Kmer read distribution

Arabidopsis trio-binning assembly





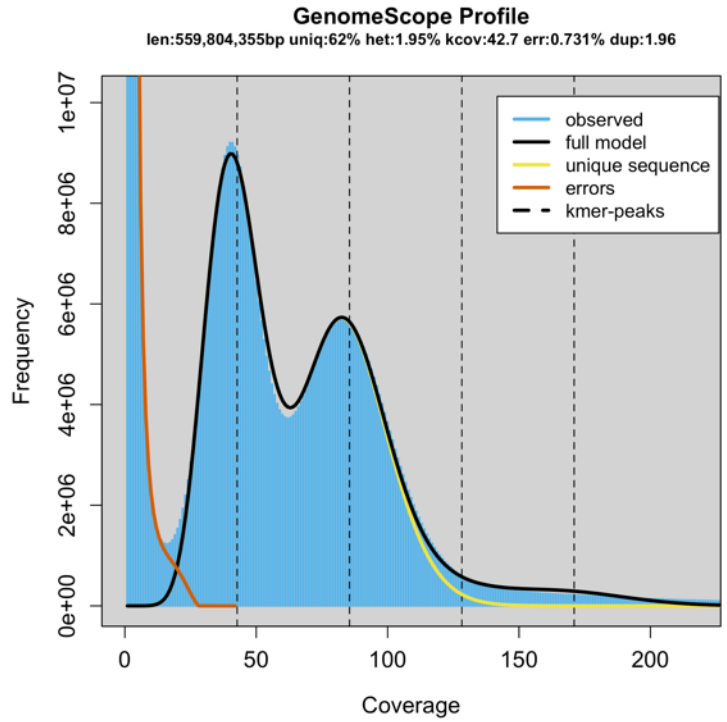
Merqury

Rhie, Walenz, Koren, Phillipy, Genome Biology (2020)

Other pre-assembly steps

Depending on the assembler and technology you use, you may want to:

- **Assess reads for contaminants**
- Join paired-end reads into longer reads
- Error correction of reads (e.g. fix sequencing errors)



Starting the assembly

Assembly recipe

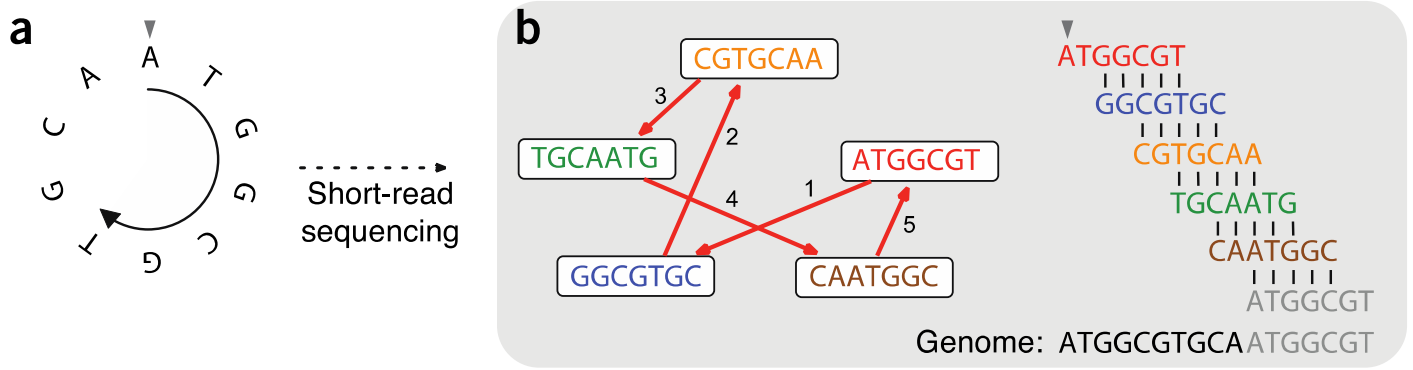
- Find all overlaps between reads
  - hmm, sounds like a lot of work…
- Build a graph
  - a picture of read connections
- Simplify the graph
  - sequencing errors will mess it up a lot
- Traverse the graph
  - trace a sensible path to produce a consensus

Graph

Review: A structure where objects are related to one another somehow

Nodes/Vertices = objects (sequence)

Edges = relationship (overlap)



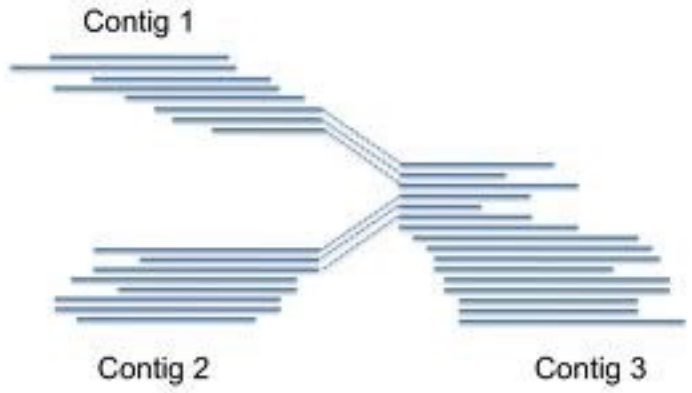
Compeau *et al,* Nature Biotech, 29(11), 2011; https://en.wikipedia.org/wiki/Graph_(discrete_mathematics)

Contigs

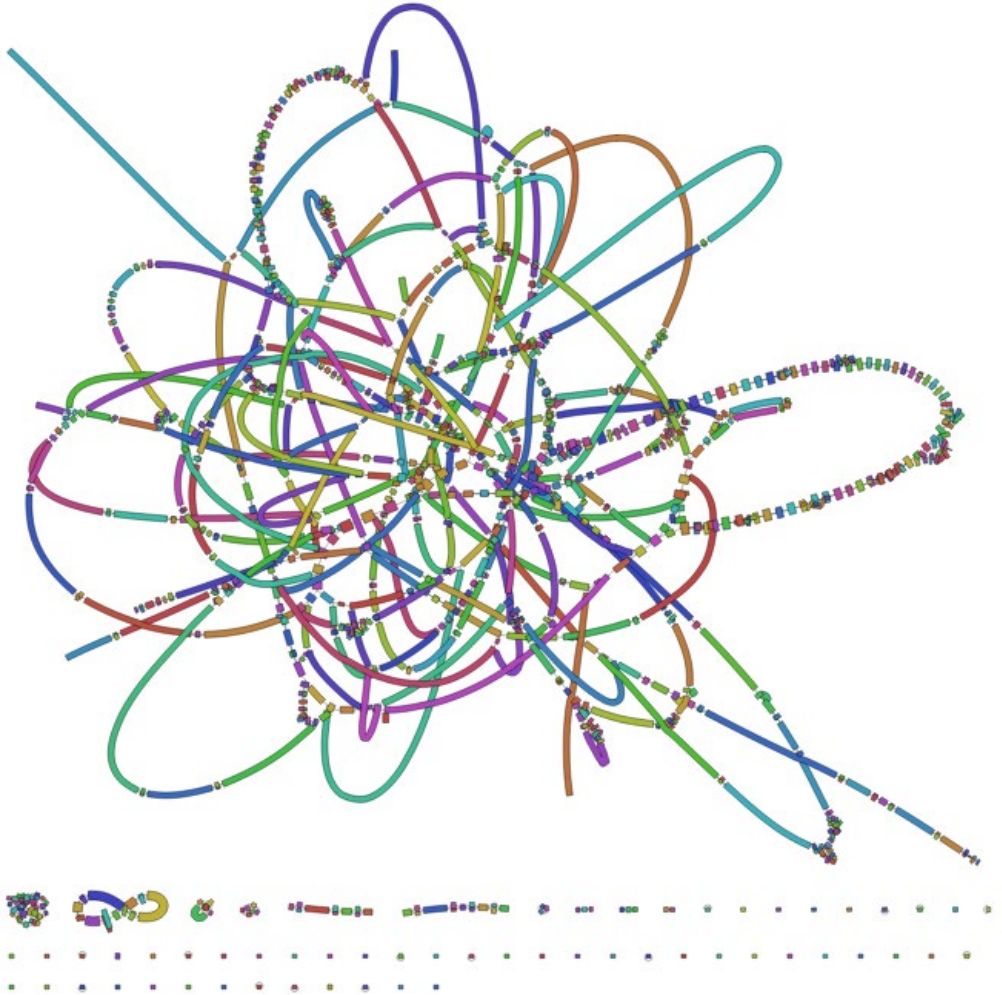
Contiguous, unambiguous stretches of assembled DNA sequence

Contigs ends correspond to

- Real ends (for linear DNA molecules)
- Dead ends (missing sequence)
- Decision points (forks in the road)



Simple?



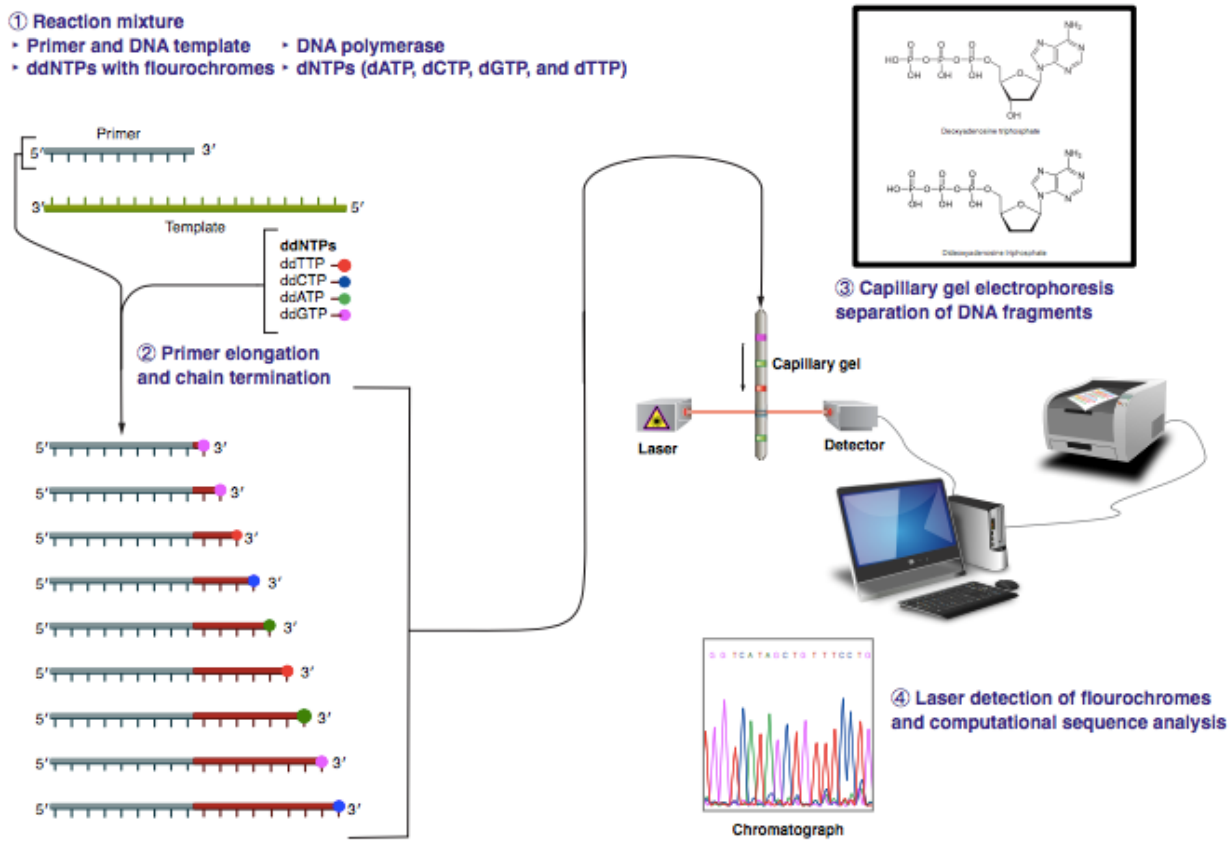
https://github.com/rrwick/Bandage/wiki/Effect-of-kmer-size

Overlap Layout Consensus Assembly

Used for longer read data

Sanger

Newer variants for PacBio and Oxford Nanopore

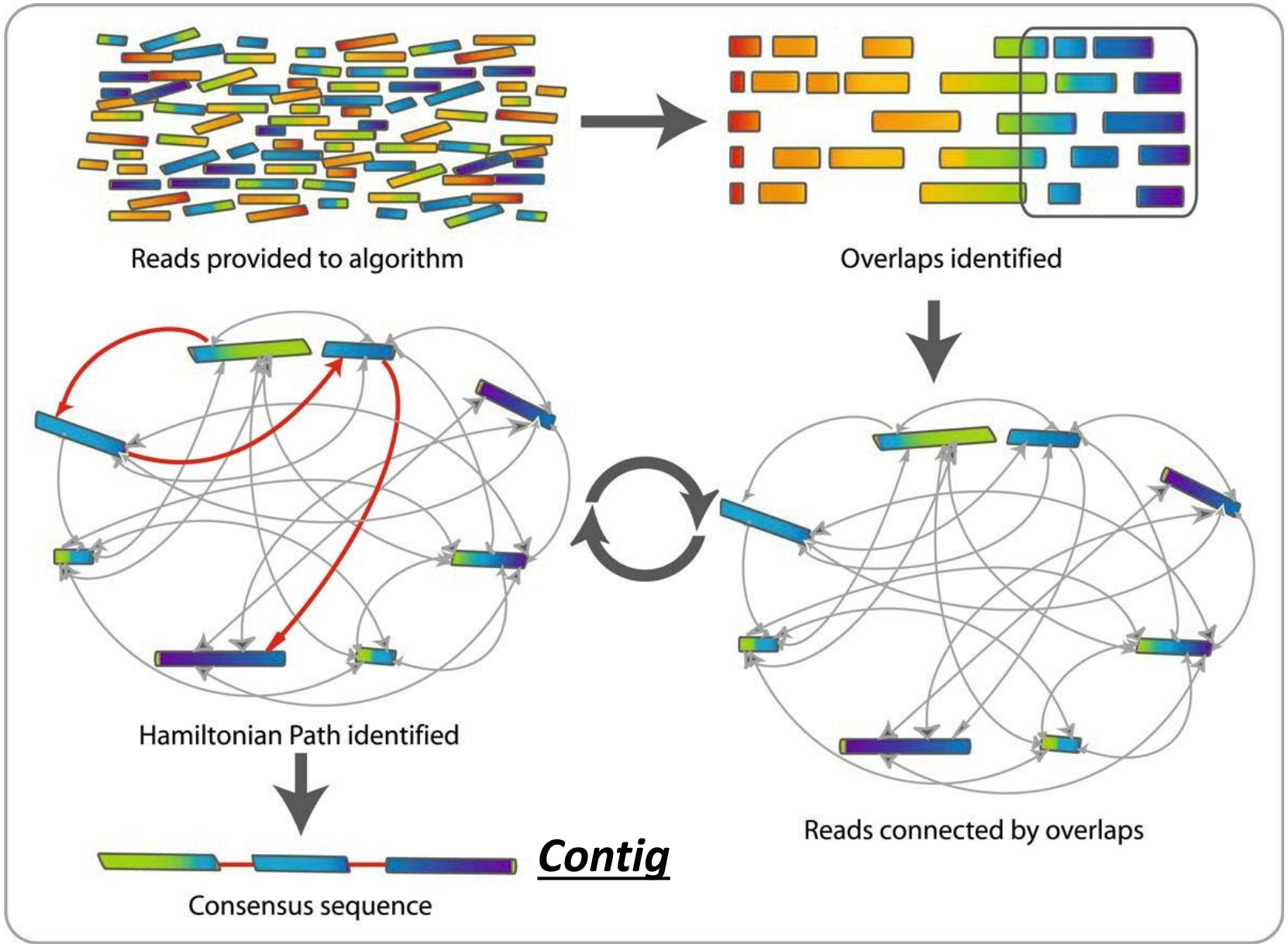


By Estevezj - Own work, CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=23264166

For each unconnected graph, at least one per replicon in original sample

Find a path which visits each node once

Form consensus Sequences from paths



Some OLC-based assemblers

HiCanu – PacBio HiFi assembler

Hifiasm – a hybrid *diploid* assembler (phasing)

Verkko – version of Canu that can incorporate additional information to generate diploid (phased) assemblies

Older ones (currently not maintained)…

Canu – Fork of the Celera Assembler designed for older high-noise single-molecule sequencing (PacBio, Oxford Nanopore)

Newbler – designed for Roche 454 sequences

Falcon – PacBio-focused assembler, capable of phasing.

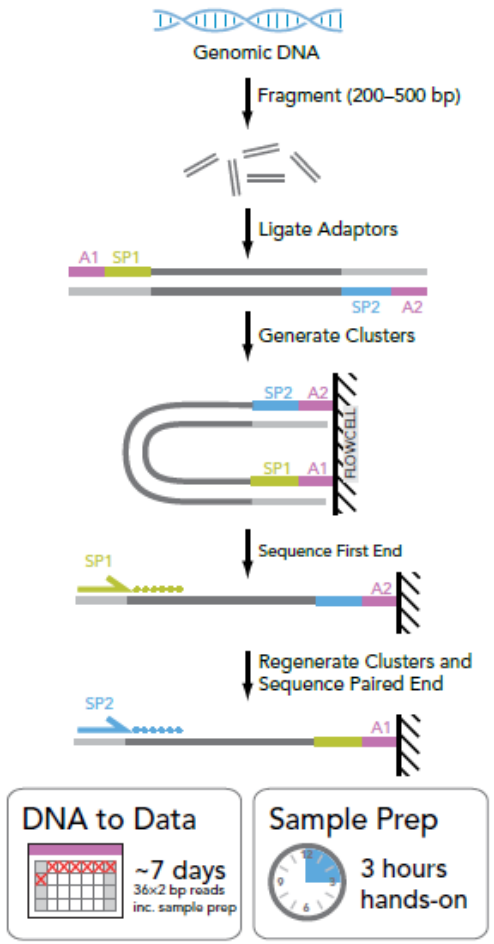
De Bruijn graph assemblers

Developed to deal with high-throughput highly accurate short-read data

Uses shotgun data (generally paired-end fragments of 300-500nt)

Illumina

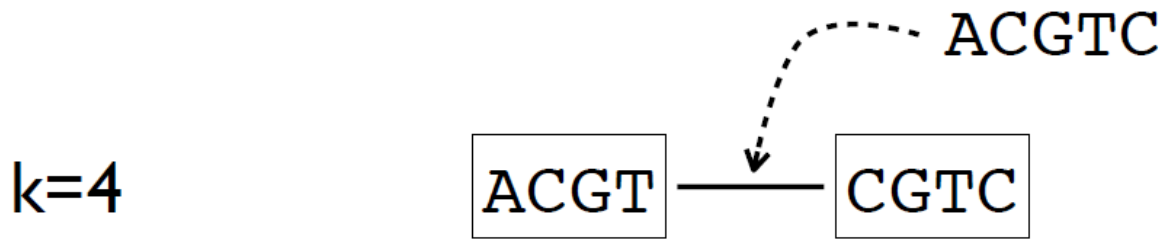
Figure 6B: Paired-End Sequencing



Adapters containing attachment sequences (A1 & A2) and sequencing primer sites (SP1 & SP2) are ligated onto DNA fragments (e.g., genomic DNA). The resulting library of single molecules is attached to a flow cell. Each end of every template is read sequentially.

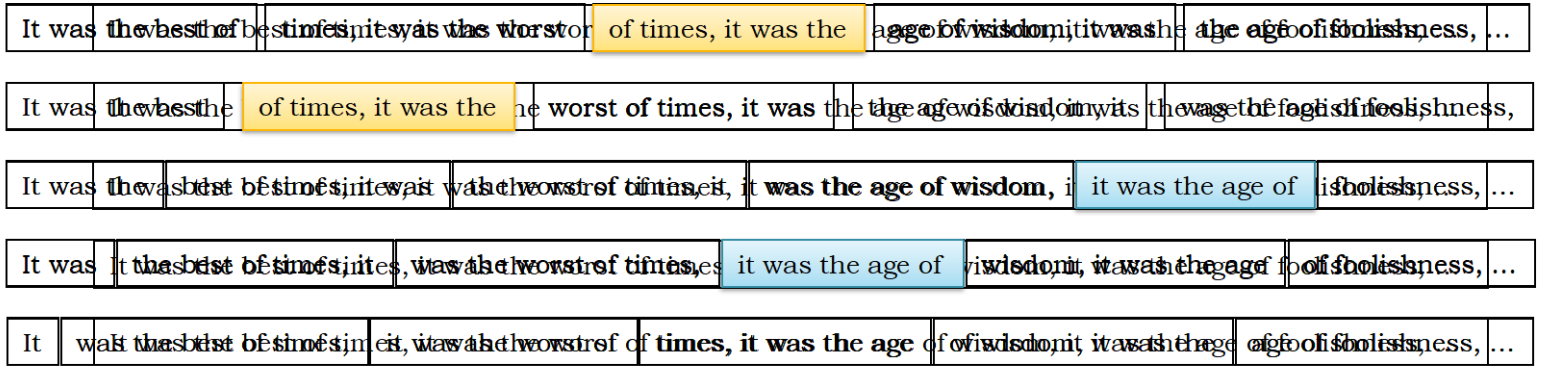
De Bruijn graphs - concept

- de Bruijn graph
  - k-dimensional graph over four symbols {A, C, G, T}
  - vertex: k-mer -- a string of k nucleotides
  - edge: (k+1)-mer



Shredded Book Reconstruction

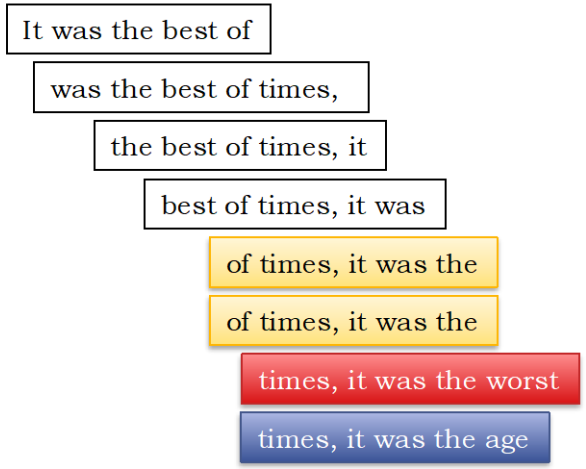
- Dickens accidentally shreds the first printing of A Tale of Two Cities
  - Text printed on 5 long spools



- How can he reconstruct the text?
  - 5 copies x 138, 656 words / 5 words per fragment = 138k fragments
  - The short fragments from every copy are mixed together
  - Some fragments are identical

Greedy Reconstruction

It was the best of
age of wisdom, it was
best of times, it was
it was the age of
it was the age of
it was the worst of
of times, it was the
of times, it was the
of wisdom, it was the
the age of wisdom, it
the best of times, it
the worst of times, it
times, it was the age
times, it was the worst
was the age of wisdom,
was the age of foolishness,
was the best of times,
was the worst of times,
wisdom, it was the age
worst of times, it was



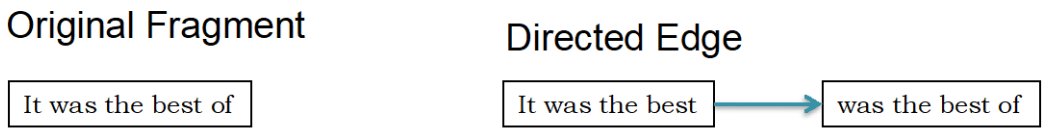
The repeated sequence make the correct reconstruction ambiguous

- It was the best of times, it was the [worst/age]

Model the assembly problem as a graph problem

de Bruijn Graph Construction

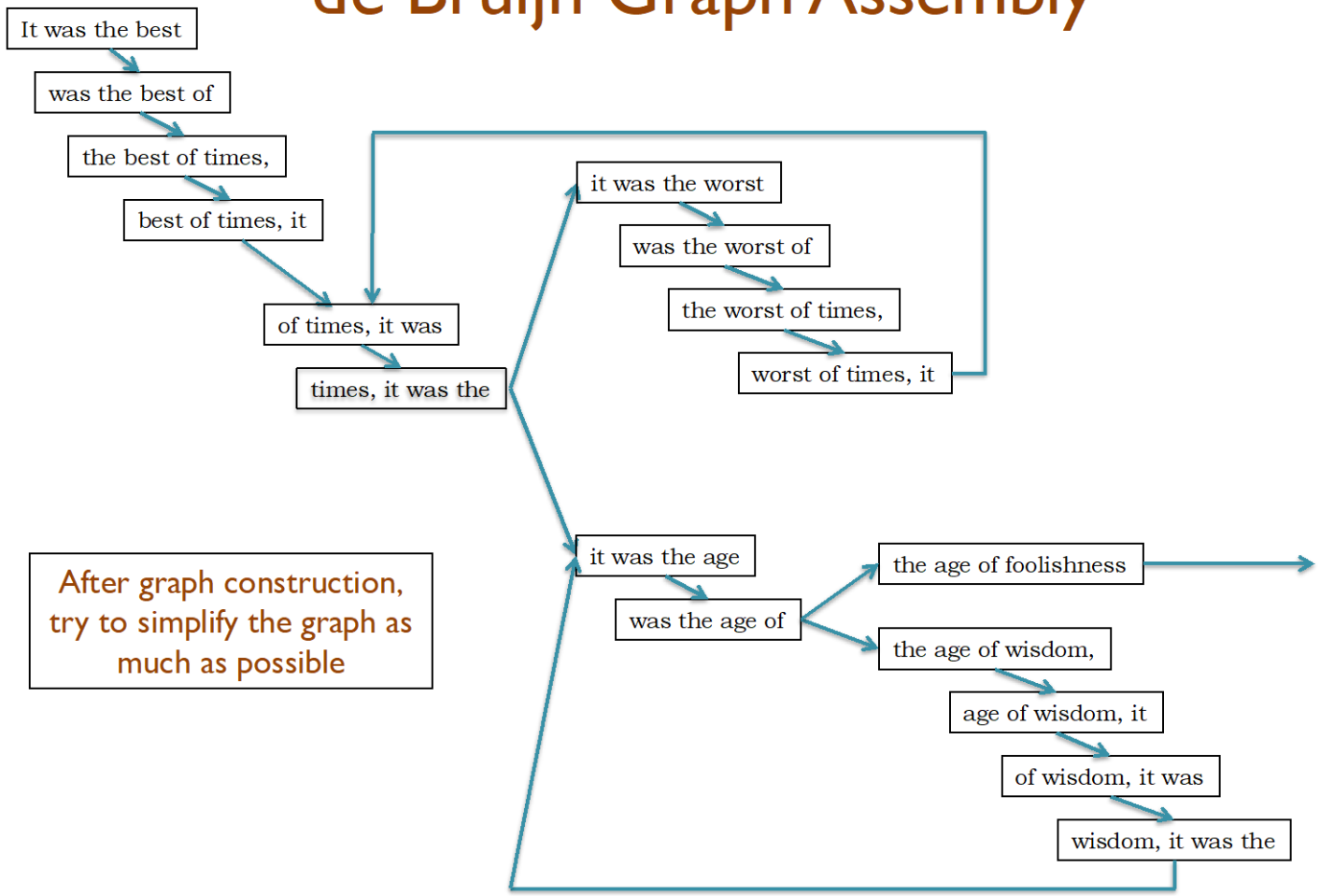
- D_k = (V,E)
  - V = All length-k subfragments (k < l)
  - E = Directed edges between consecutive subfragments
    - Nodes overlap by k-1 words



- Locally constructed graph reveals the global sequence structure
  - Overlaps between sequences implicitly computed

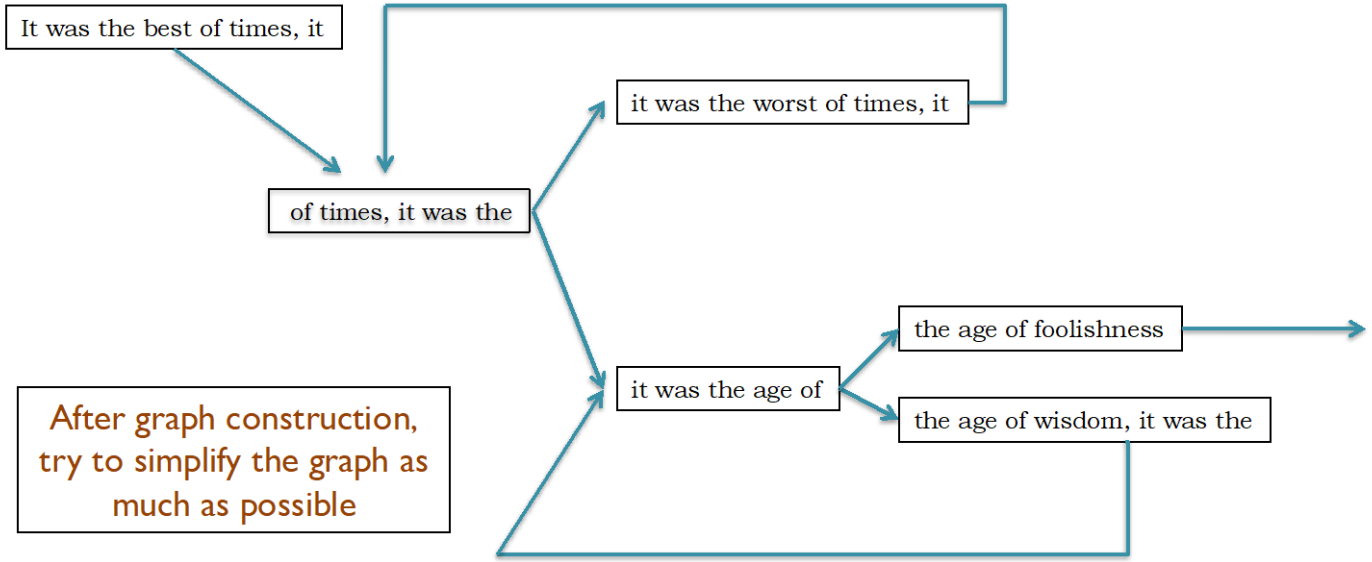
de Bruijn, 1946
Idury and Waterman, 1995
Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly



After graph construction, try to simplify the graph as much as possible

de Bruijn Graph Assembly



The full tale

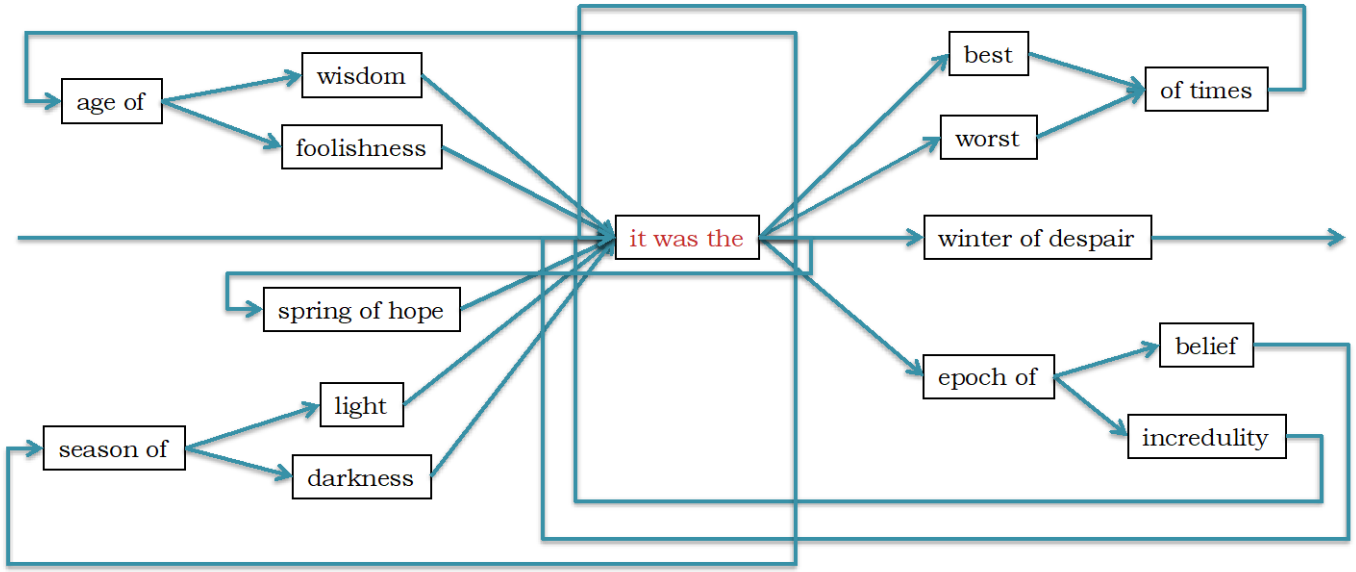
… it was the best of times it was the worst of times …

… it was the age of wisdom it was the age of foolishness …

… it was the epoch of belief it was the epoch of incredulity …

… it was the season of light it was the season of darkness …

… it was the spring of hope it was the winder of despair …



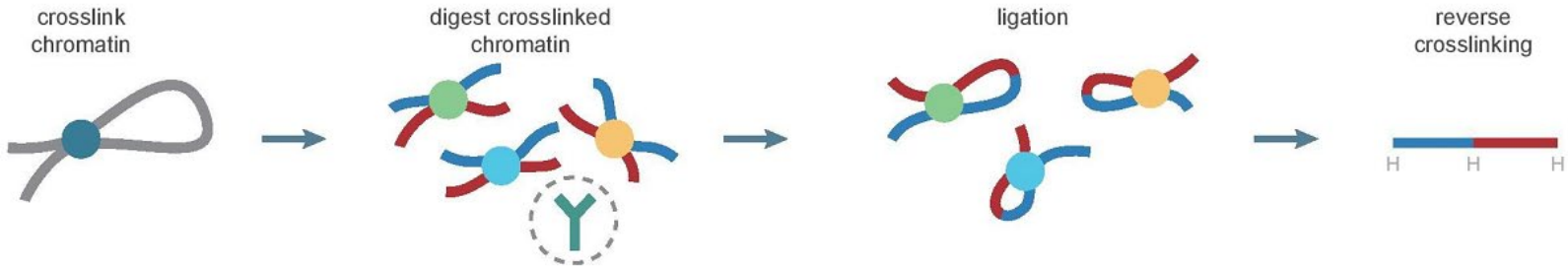
Scaffolding

- Now, you have a huge pile of contigs but you want to make them larger. How?
- Add context!
- Link together contigs using ***other*** genomic information
  - Infer contigs position on the genome relative to one another

Linking Contigs via DNA Seq

HiC (Chromosome Conformation Capture)

Wikipedia



PacBio/ONT long-reads

10-100 kb+

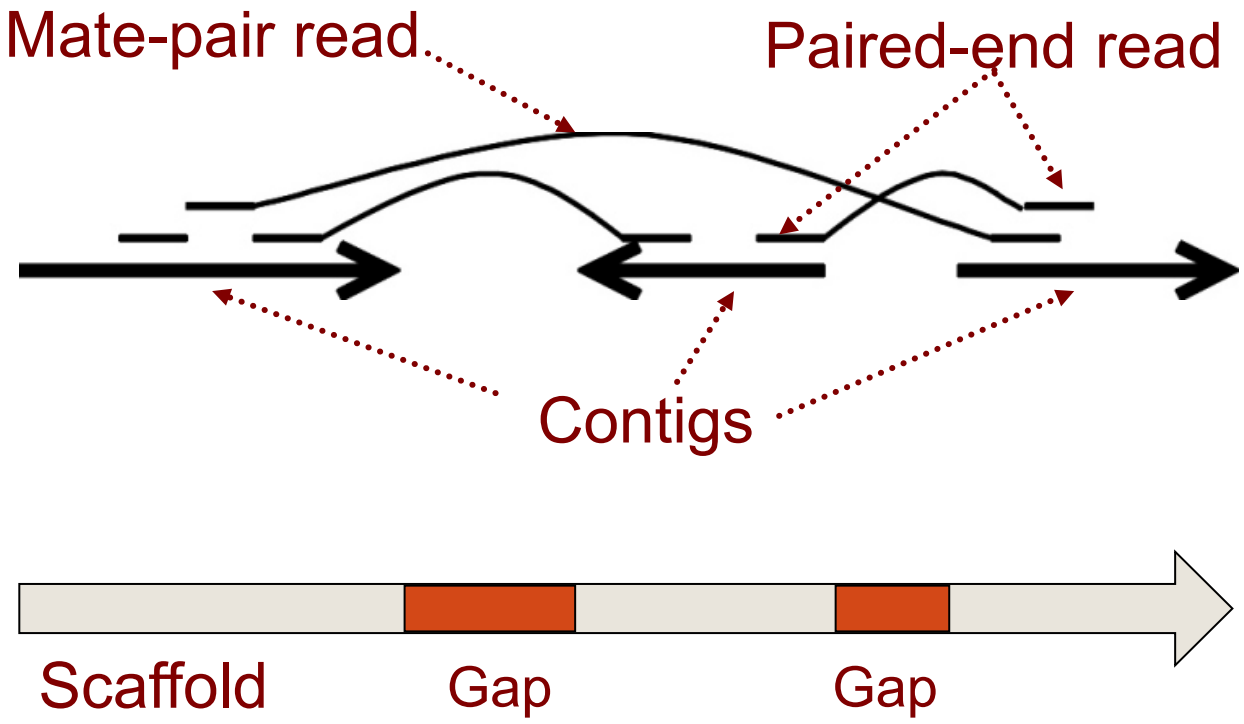
Illumina sequencing

Paired-end reads

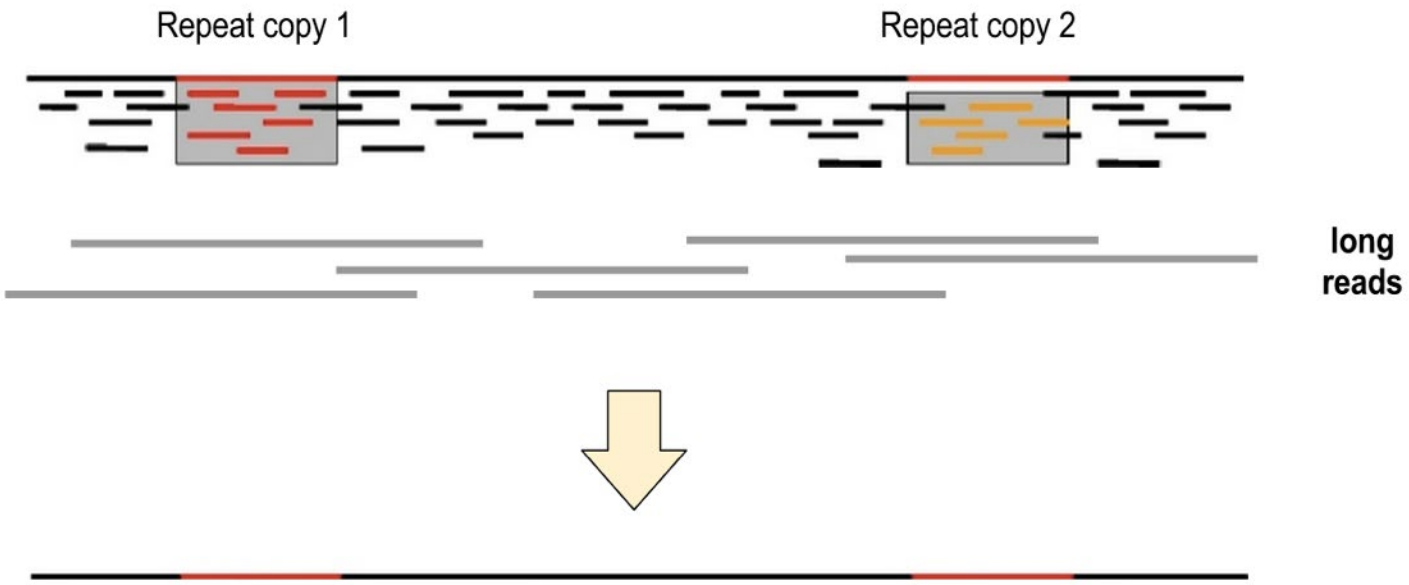
Mate pair reads

Linked reads

Contigs to scaffolds

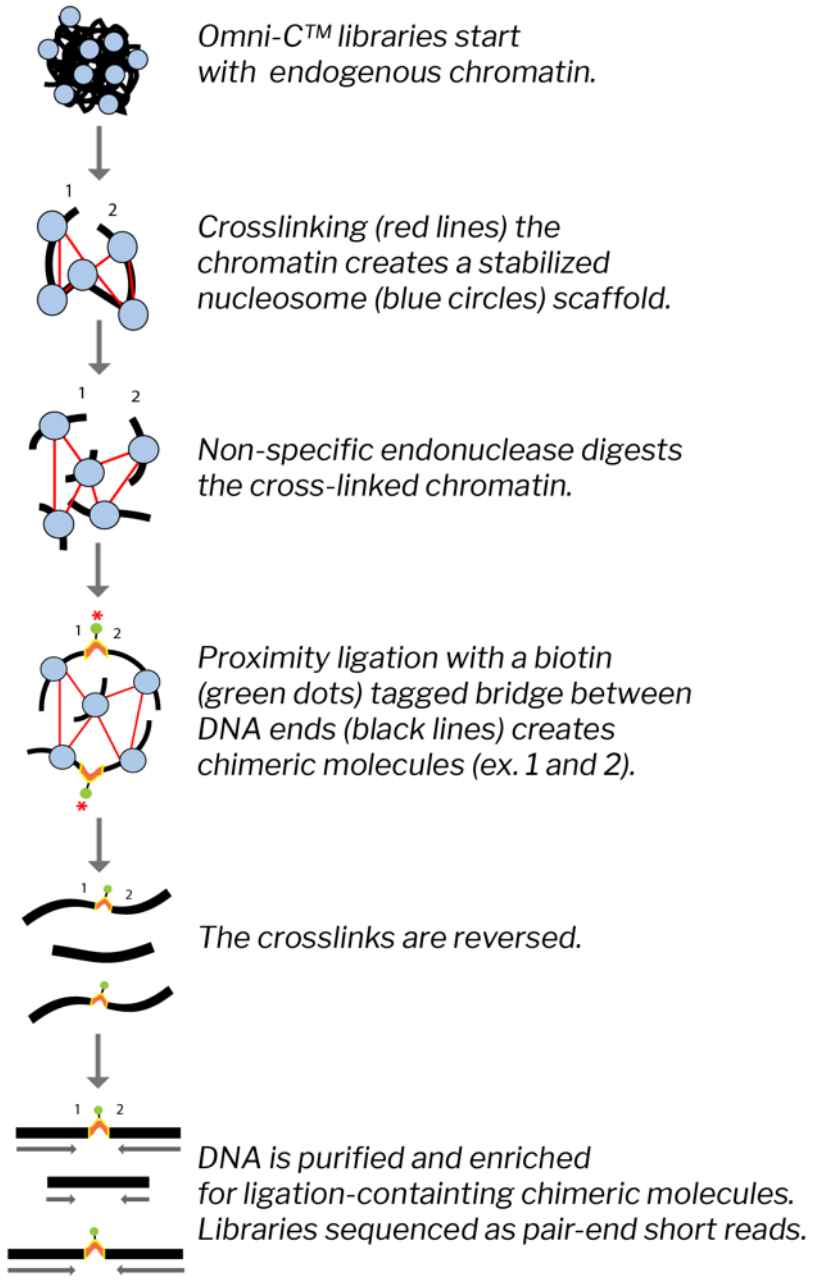


Long reads

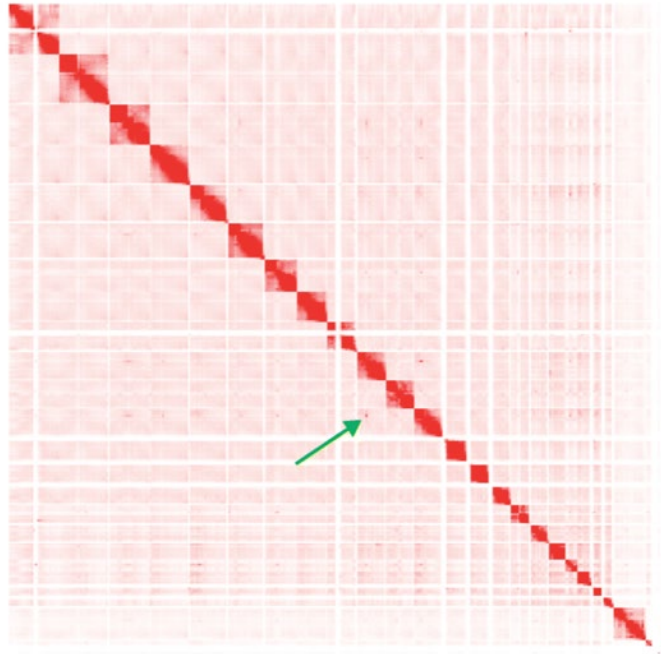


HiC

Chromosome Conformation Technology



Dovetail Genomics

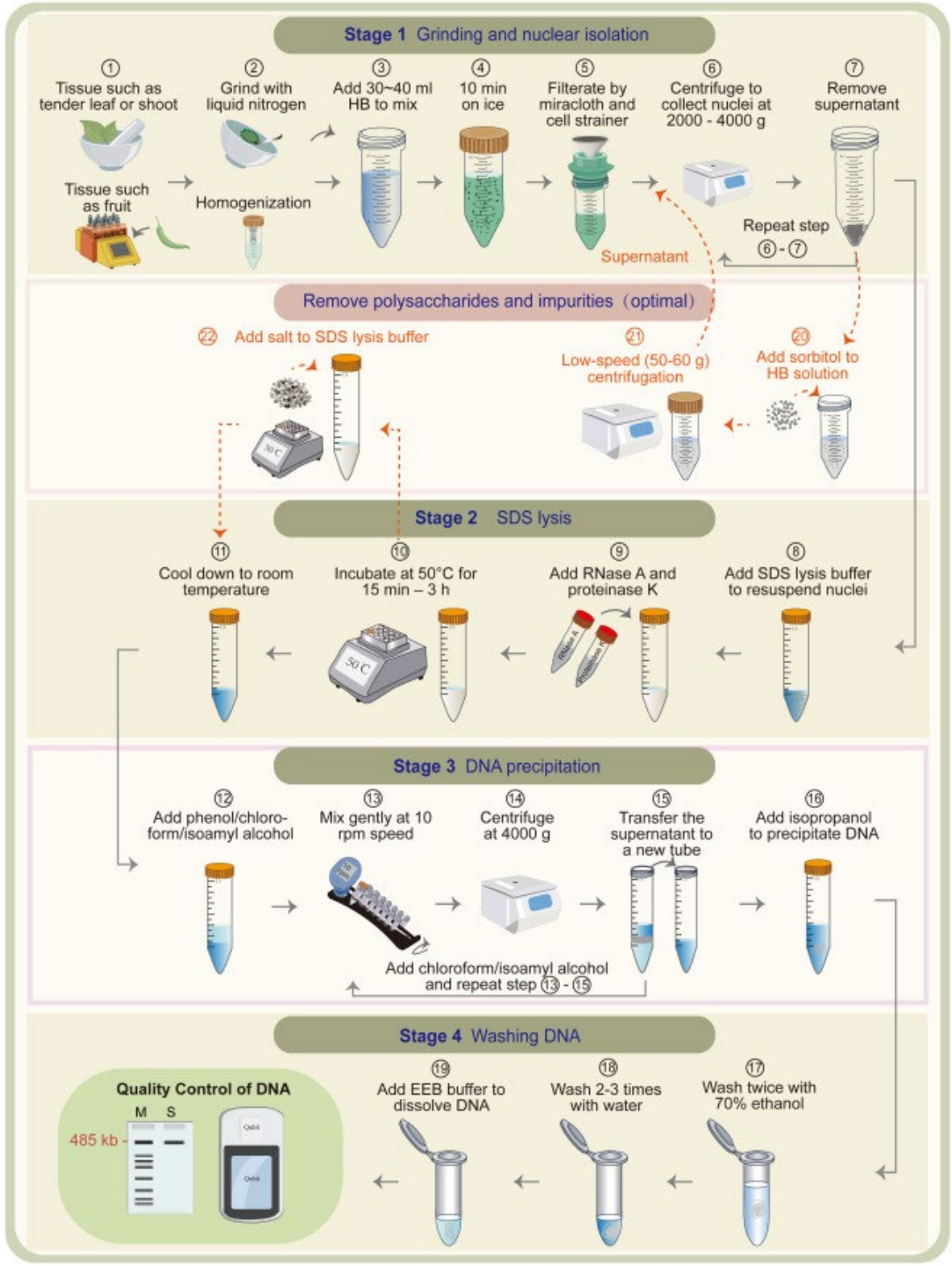


Wikipedia

ONT Scaffolding and Phasing

ONT Ultralong (>100kb)

ONT PoreC



Dong *et al*, Molecular Plant, 17(11): 2024

Starting a new assembly project

Planning a genome sequencing project?

BUDGET!!!

- *Technological costs*
- *Computational costs*
- *Person costs (time)!*

Biology!

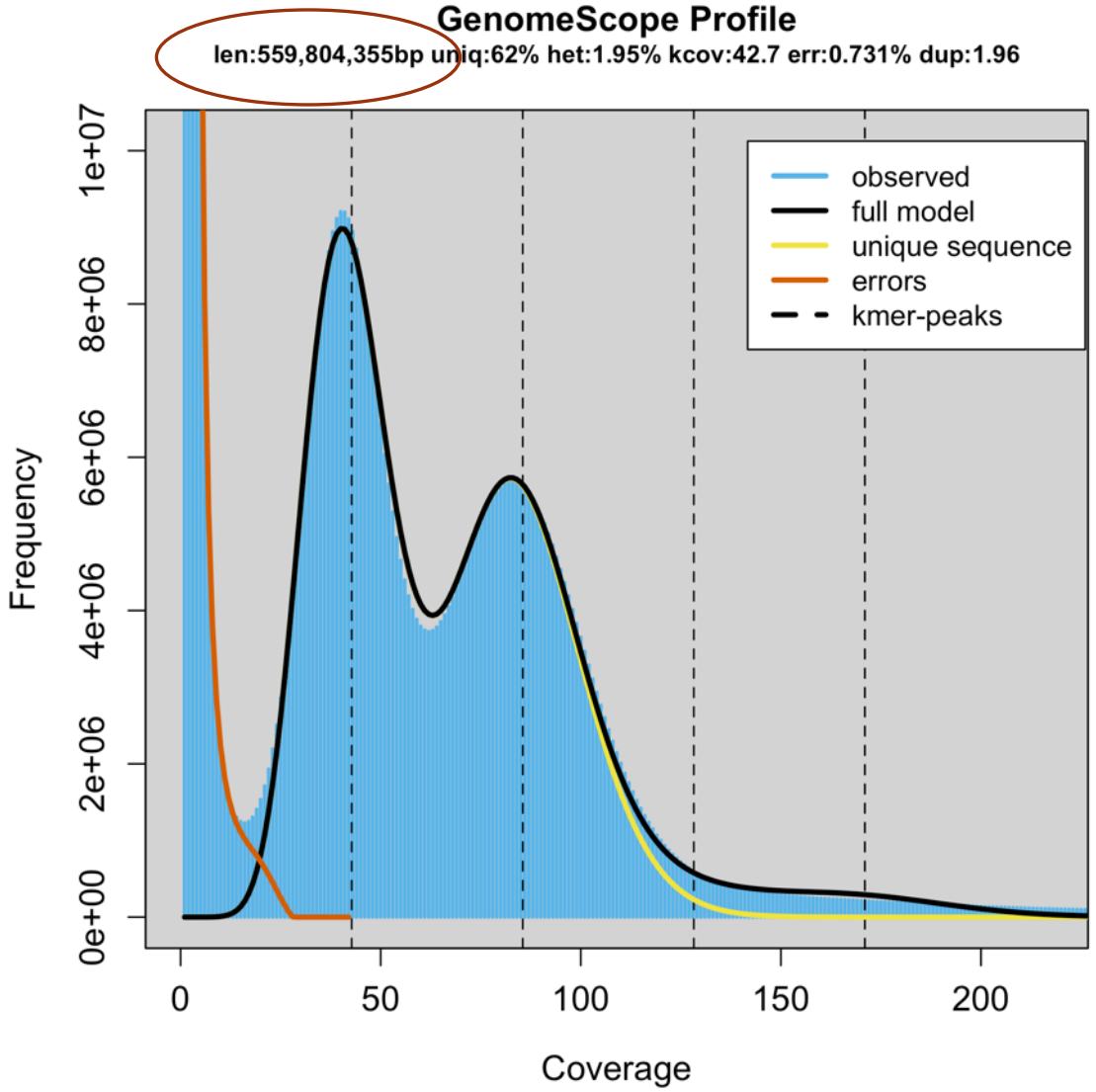
- **Size**: how large and/or complex is my genome?
- **Ploidy:** number of sets of chromosomes of the genome?
- **Multinucleated:** can cells have more than one nucleus?
- **Repetitive:** How much of the genome is repetitive? Repeat size distribution?
- **Heterozygosity:** Is my genome highly heterozygous? Inbred (homozygous)?
- **Public data:** Is a good quality genome of a related species available?

How large is my genome?

The size and complexity of the genome can be estimated from the ploidy of the organism and the DNA content per cell

This will affect:

- How many reads will be required to attain sufficient coverage (typically 10x to 100x, depending on read length)
- What sequencing technology to use (short vs. long reads)
- What computational resources will be needed (generally amount of memory needed and length of time resources will be used)



Oyster (GenomeScope)

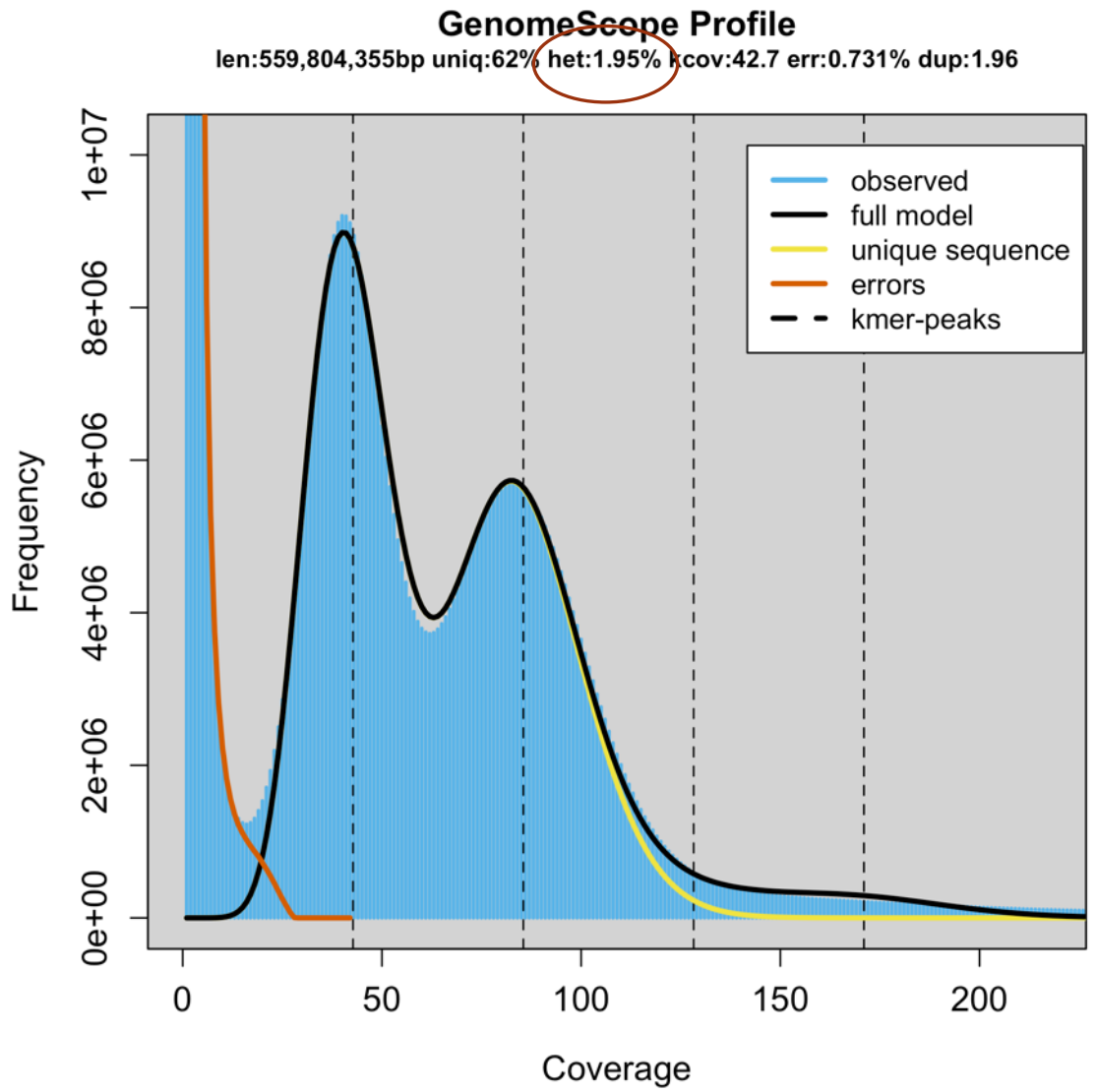
Heterozygosity

Heterozygous – Locus-specific; diploid organism has two different alleles at the same locus.

Heterozygosity is a metric used to denote the probability an individual will be heterozygous at a given allele.

Higher heterozygosity == more diverse == harder to assemble

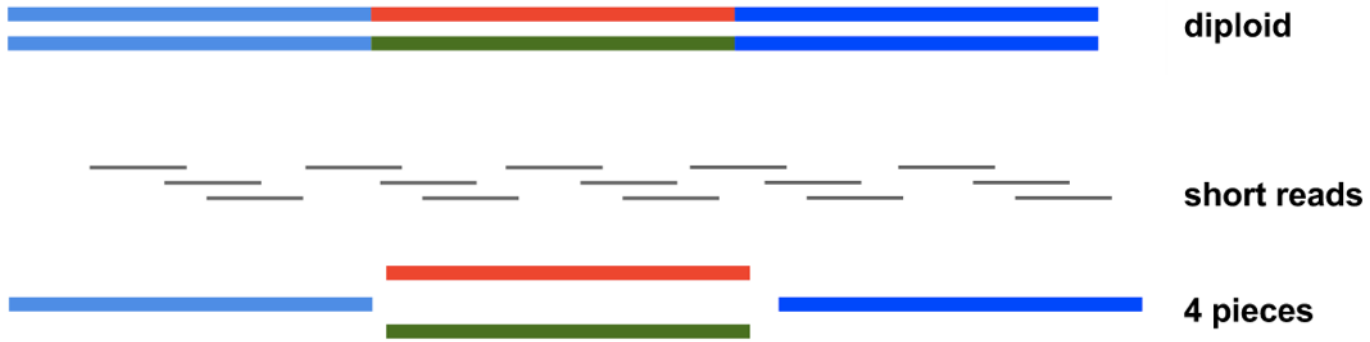
Unfortunately, assemblies are represented (for now) as haploid. So this is a major problem!



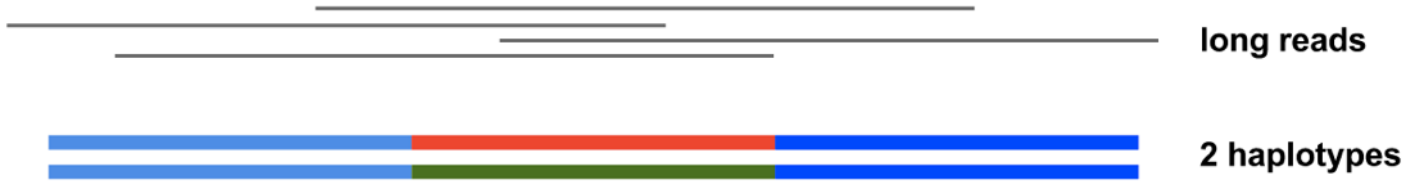
Oyster: http://qb.cshl.edu/genomescope/genomescope2.0/

Heterozygosity

- **Short reads** - initial assembly has mix of homozygous and heterozygous regions
- **Long reads** – can get partial to fully phased diploid assemblies
  - May need multiple technologies to do this



Unphased haploid assembly
Haplotypes are separate contigs (**haplotigs**)



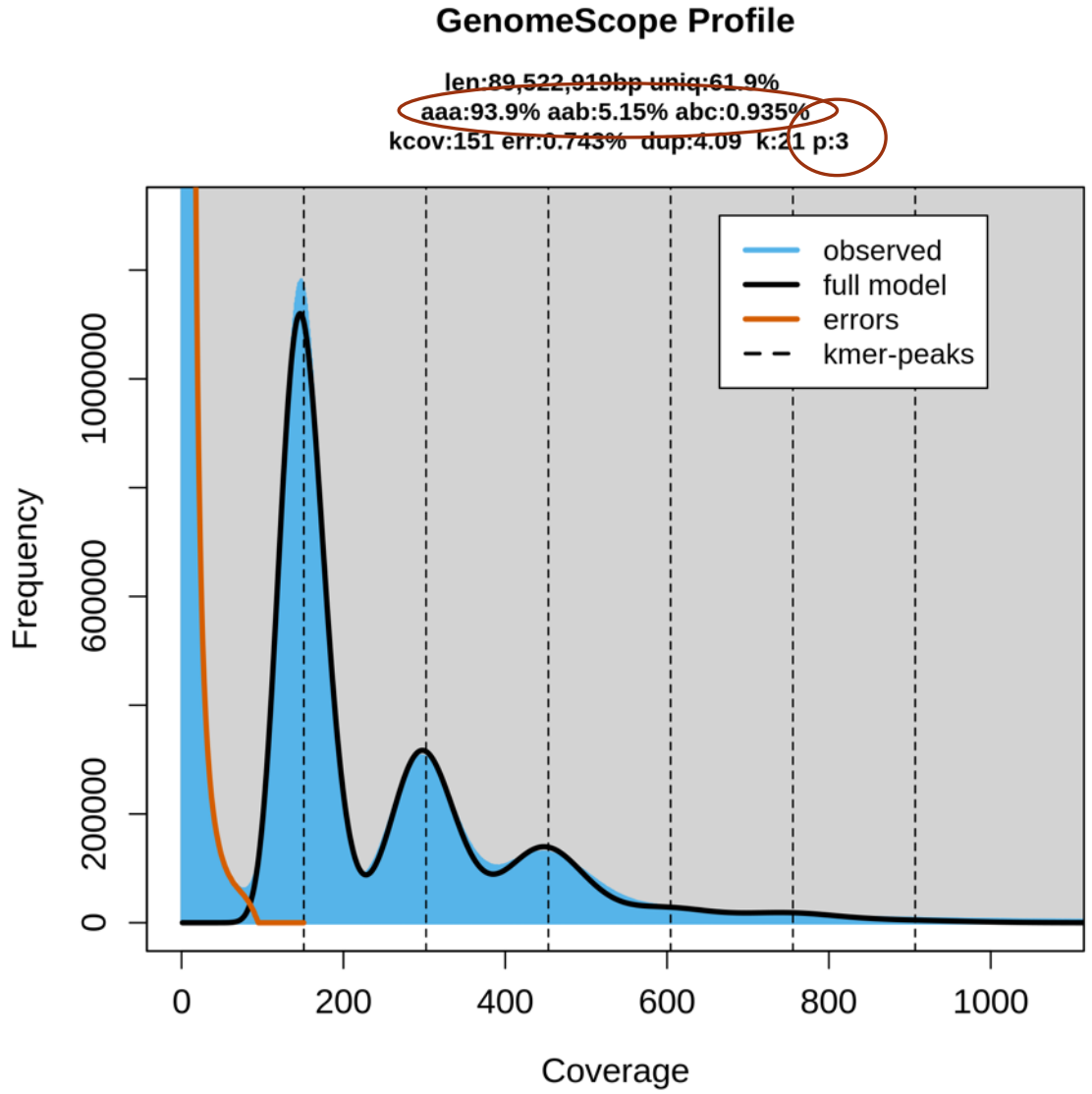
Phased diploid assembly

T. Seemann

Ploidy

Number of sets of chromosomes in a cell (N)

- Bacteria – 1N
- Vertebrates – 2N (human, mouse, rat)
- Amphibians – 2N to 12N
- Plants – 2N to ??? (wheat is 6N)



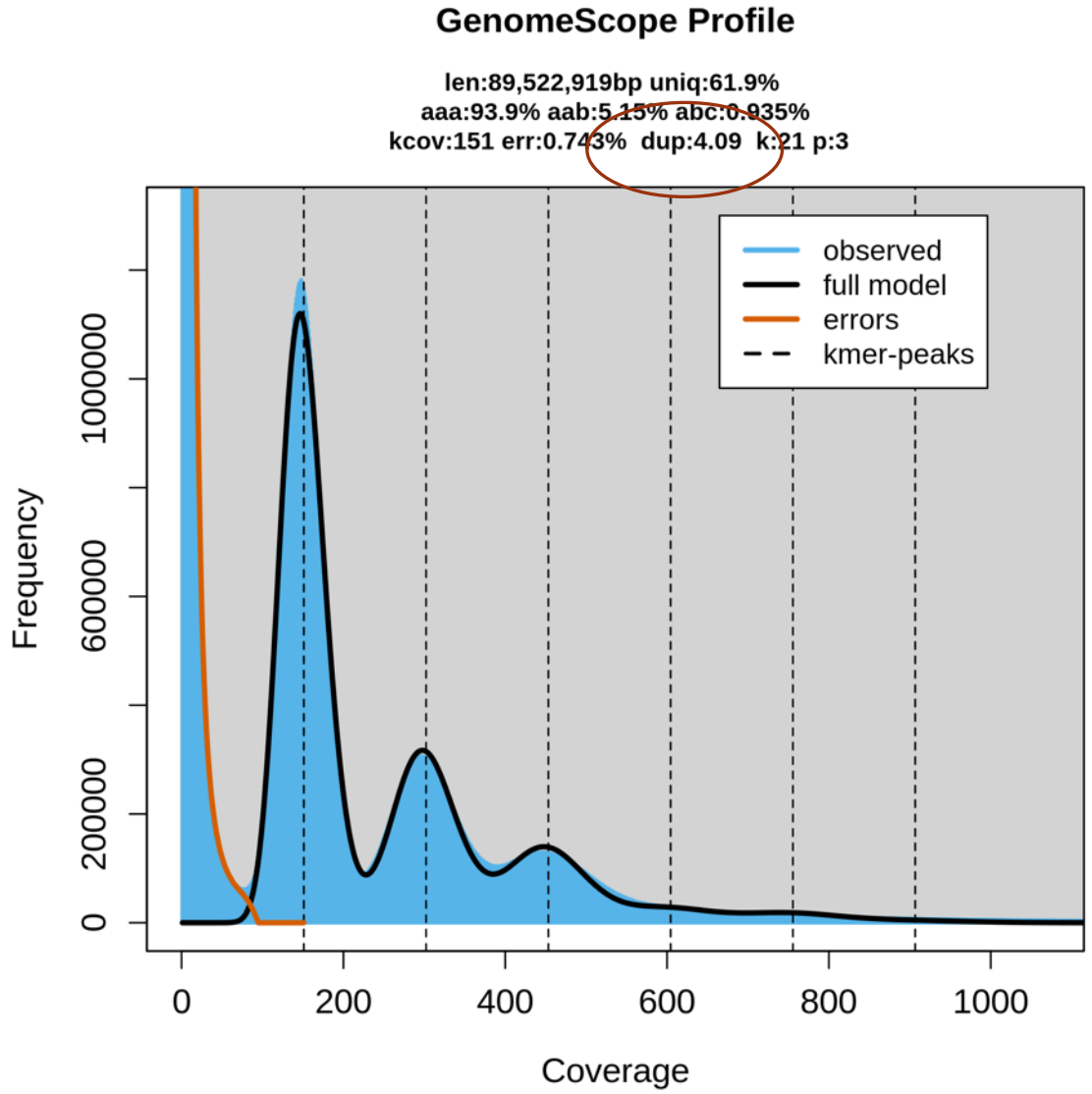
Root knot nematode (GenomeScope)

Repetitive sequences

Most common source of assembly errors

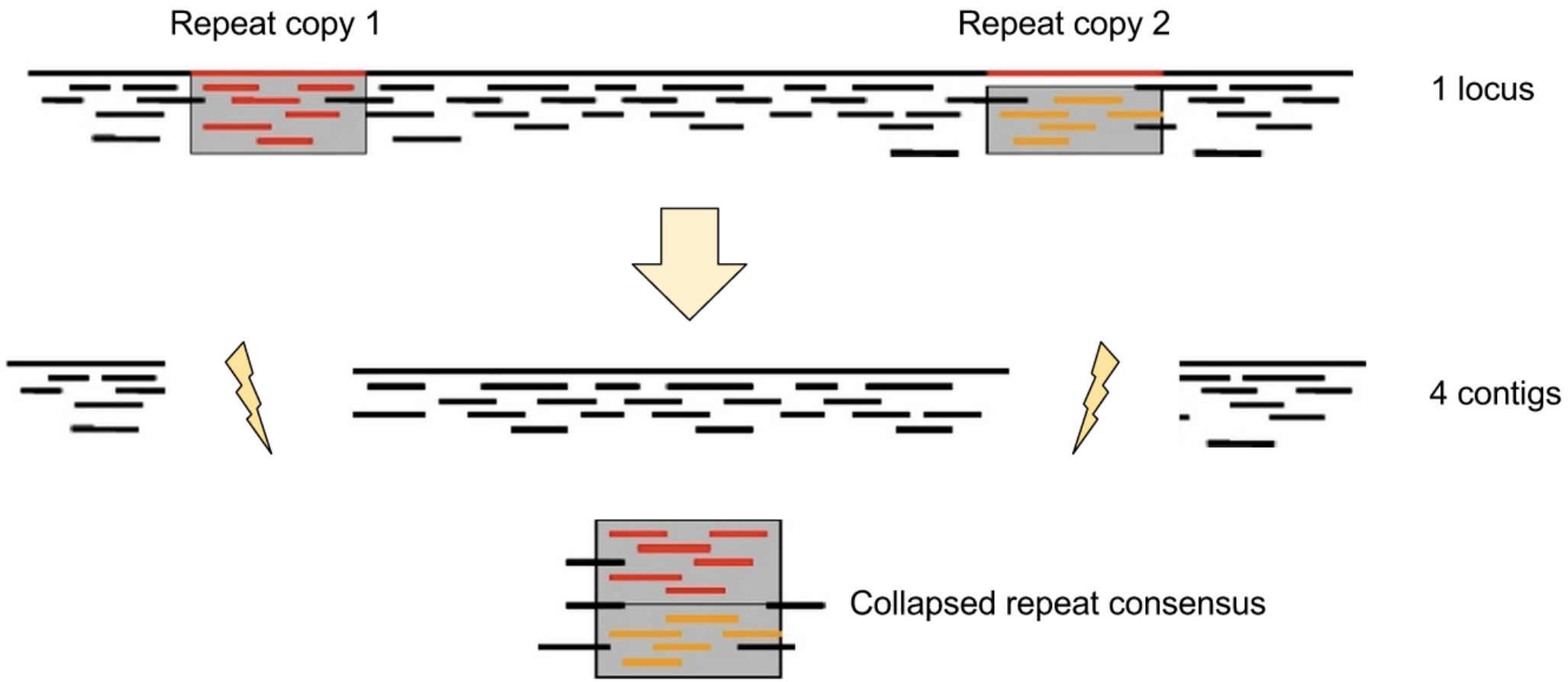
If sequencing technology produces reads > repeat size, impact is much smaller

Most common solution: generate reads or mate pairs with spacing > largest known repeat

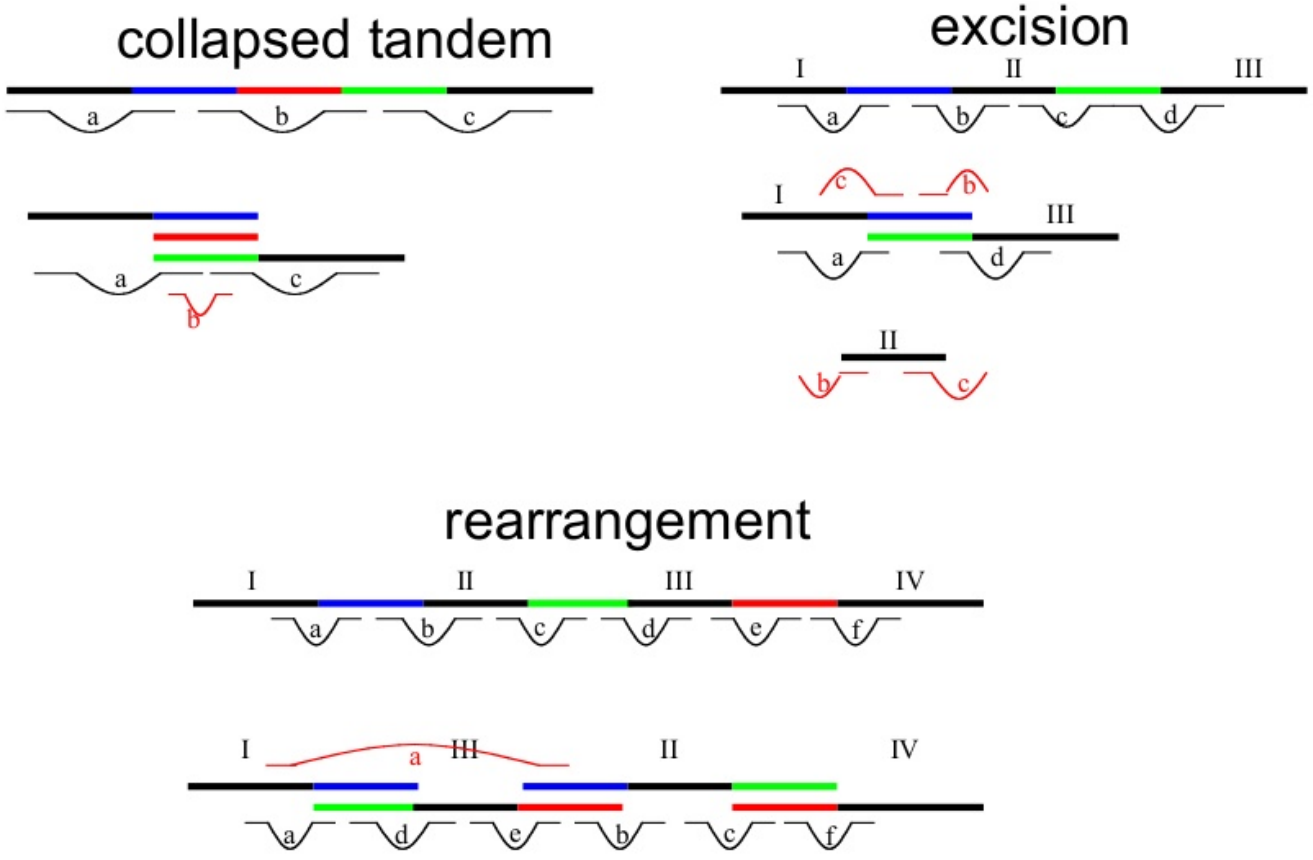


Root knot nematode (GenomeScope)

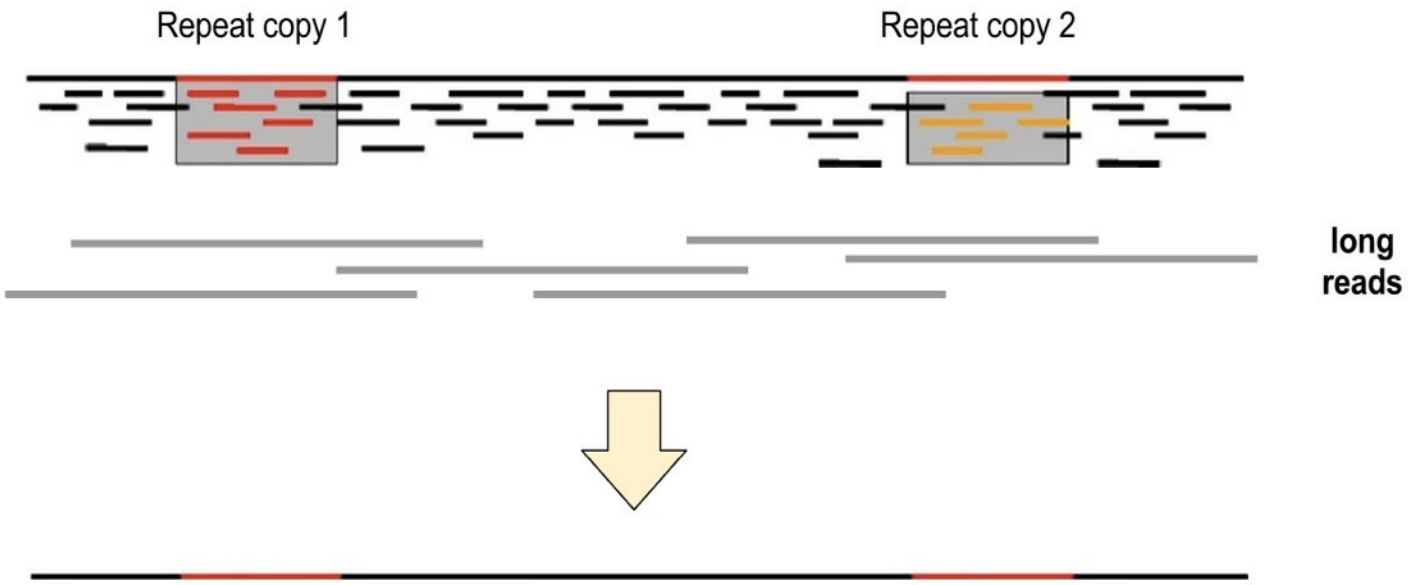
Assembling repeats

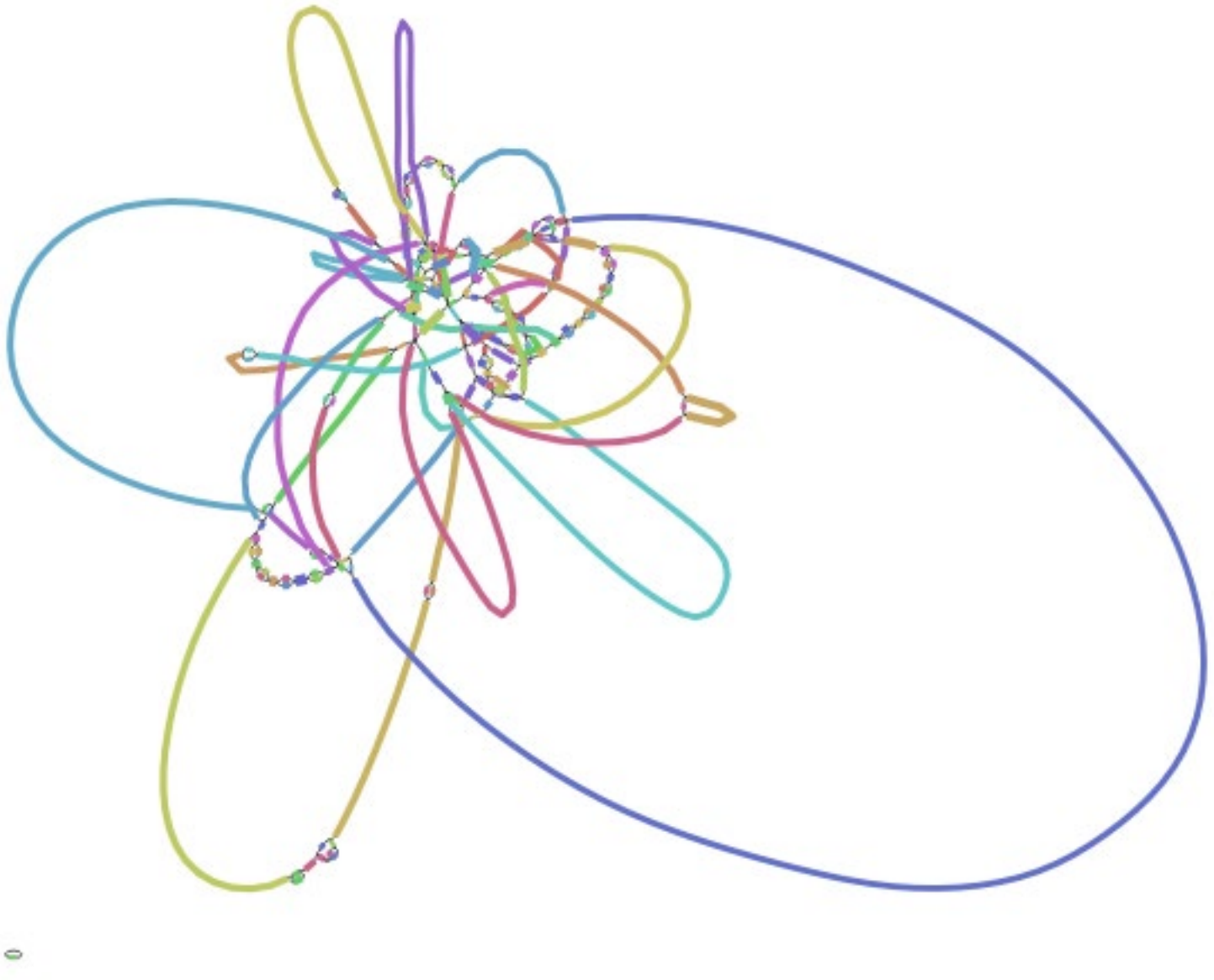


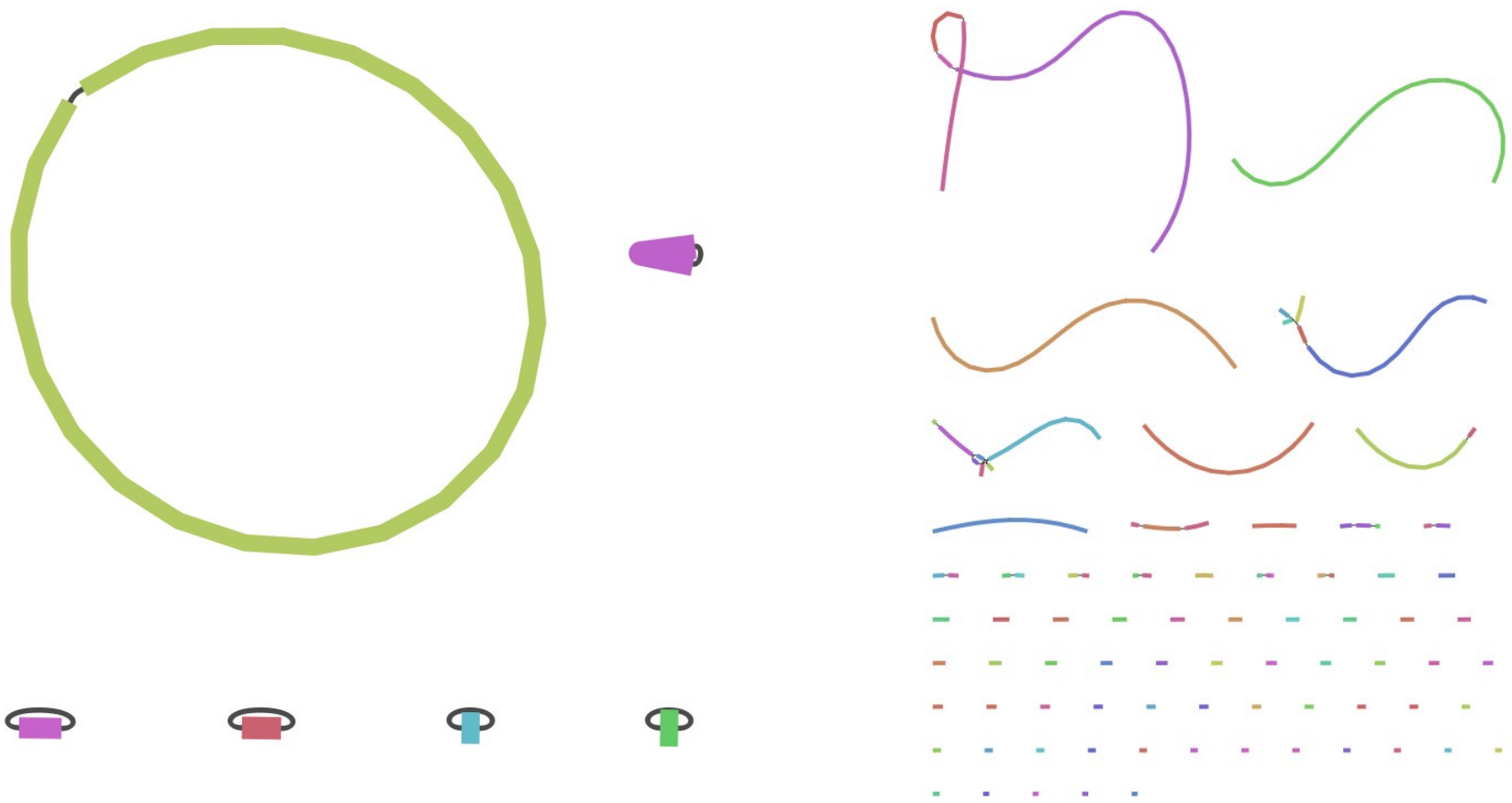
Repeat mis-assembly



Long reads







Typical sequencing strategies

You want a high-quality assembly (scaffolds, not chromosome level, partial phasing)

- PacBio HiFi - ~30-40-fold coverage per chromosome copy if highly heterozygous
- Inbred samples can be ~30-40-fold
- Oxford Nanopore - 40-50-fold coverage
- Consider linked reads (TELL-Seq) or HiC

You want chromosome-level and/or phased ("diploid") assembly

- The above + HiC for scaffolding (40-50x)
- Possibly optical mapping?

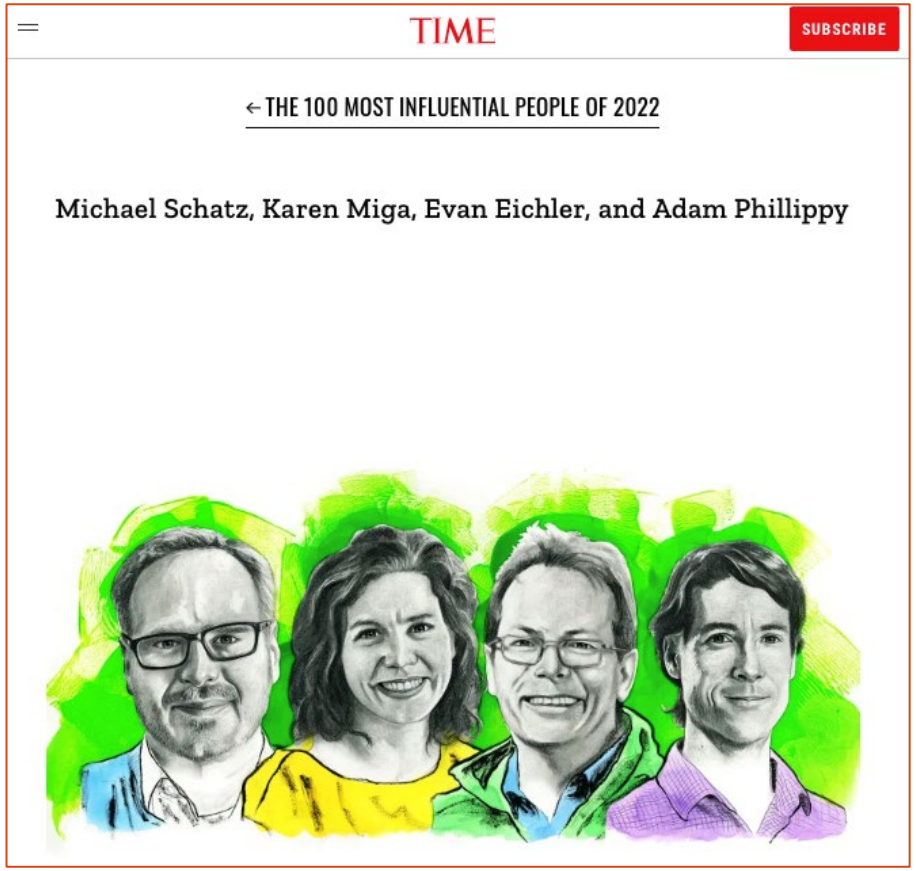
Do you want chromosomes fully resolved, phased (T2T)

- The above + ONT (ultralong prep) – 30-fold
- Possibly ONT PoreC – 30-fold
- A student, postdoc, analyst to manually curate the genome

T2T strategy

- Human assemblies
- HMW DNA preps
- **Min 50x-fold PacBio HiFi reads**
- **15-30x Oxford ultralong reads (>100kb)**
- ‘Near T2T’ - (mostly) automated assembles
- Can involve manual effort to resolve
- Oxford only protocol now available!

Science, March 2022



Time, May 2022

Assembly strategies and algorithms

For long reads (>500 nt), Overlap/Layout/Consensus (OLC) algorithms work best.

- **Examples: hifiasm (PacBio HiFi only), Verrko, HiCanu, Redbean, Flye, Shasta**
- **Hifiasm is generally recommended for PacBio HiFi data**
- **Verrko is pretty good too, but works best w/ alternate libraries (HiC, ONT)**

For short reads, De Bruijn graph-based assemblers are most widely used

- **Examples: MEGAHIT, SPAdes**

Key points:

- There is no simple solution, best to try different assemblers and strategies
- Use simple metrics to gauge quality of assembly
- The field is rapidly evolving, like the sequencing technology

NEXT YEAR THIS PRESENTATION WILL CHANGE AGAIN!

Assessing your assembly

How good is my assembly?

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

How many pieces, and what is their size distribution?

Are the contigs assembled correctly?

Are the scaffolds connected in the right order / orientation?

How were the repeats handled?

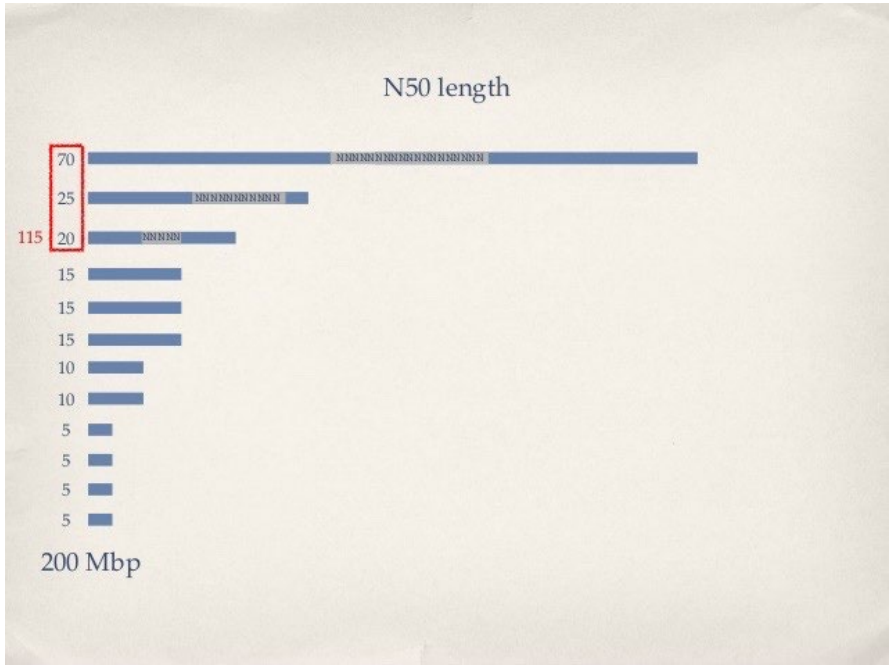
Are all the genes I expected in the assembly?

N50: the most common measure of assembly quality

N50 = length of the shortest contig in a set making up 50% of the total assembly length (***Larger is better***)

NG50 = length of the shortest contig in a set making up 50% of the **estimated genome size**

NG50 is generally better



Comparative analysis

Compare against

- A close reference genome
- Results from another assembler
- Self-comparison
- Versions of the same assembly

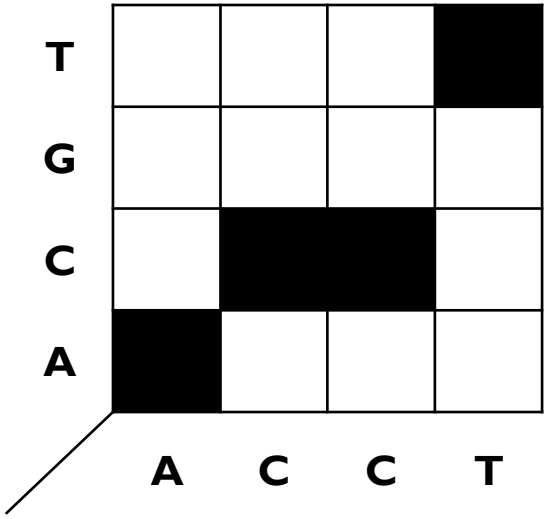
Whole genome alignment

- ***minimap2***
- ***MUMmer***
- ***Lastz***

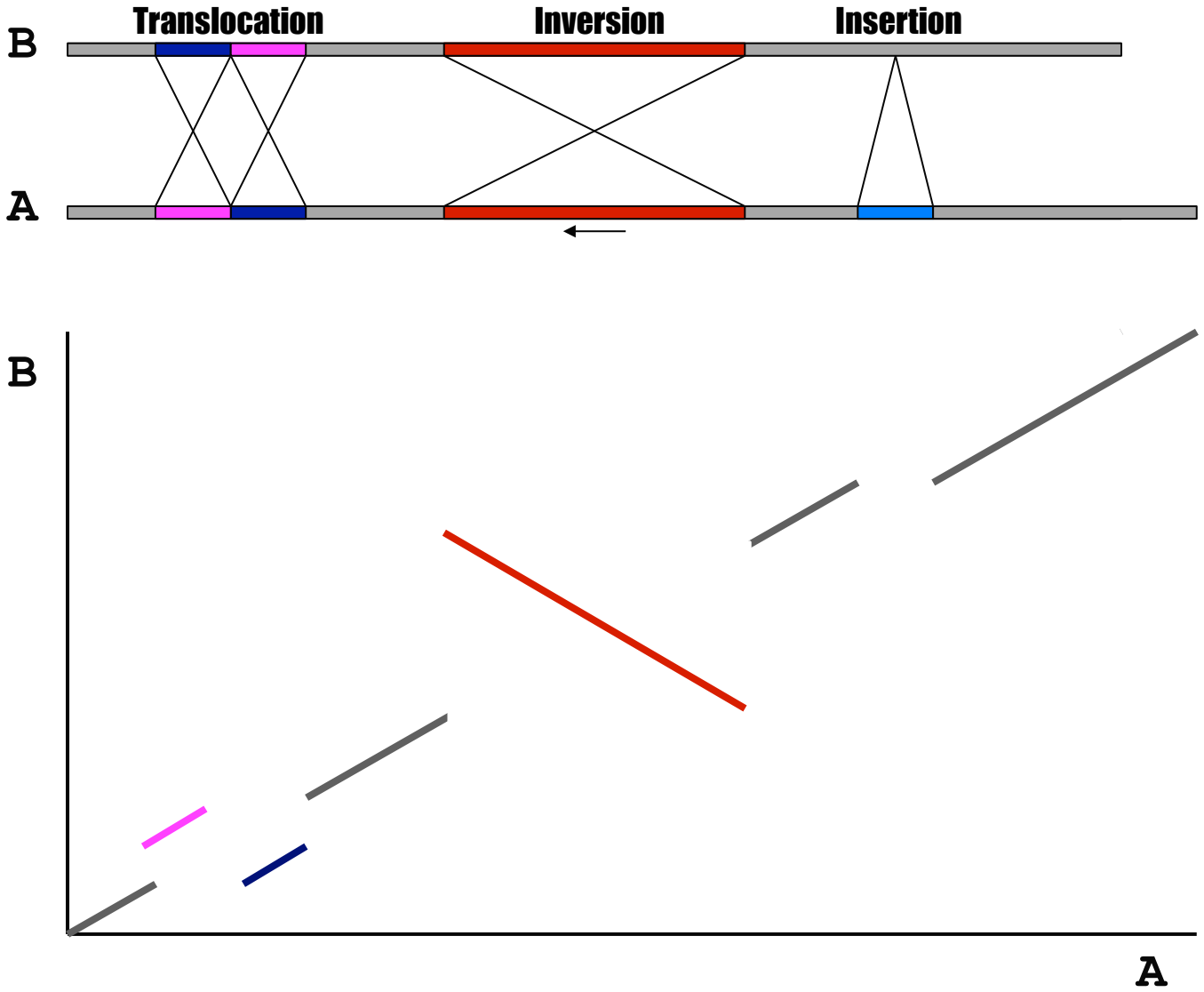
Generate an alignment and a ***dot plot***

Dot Plot

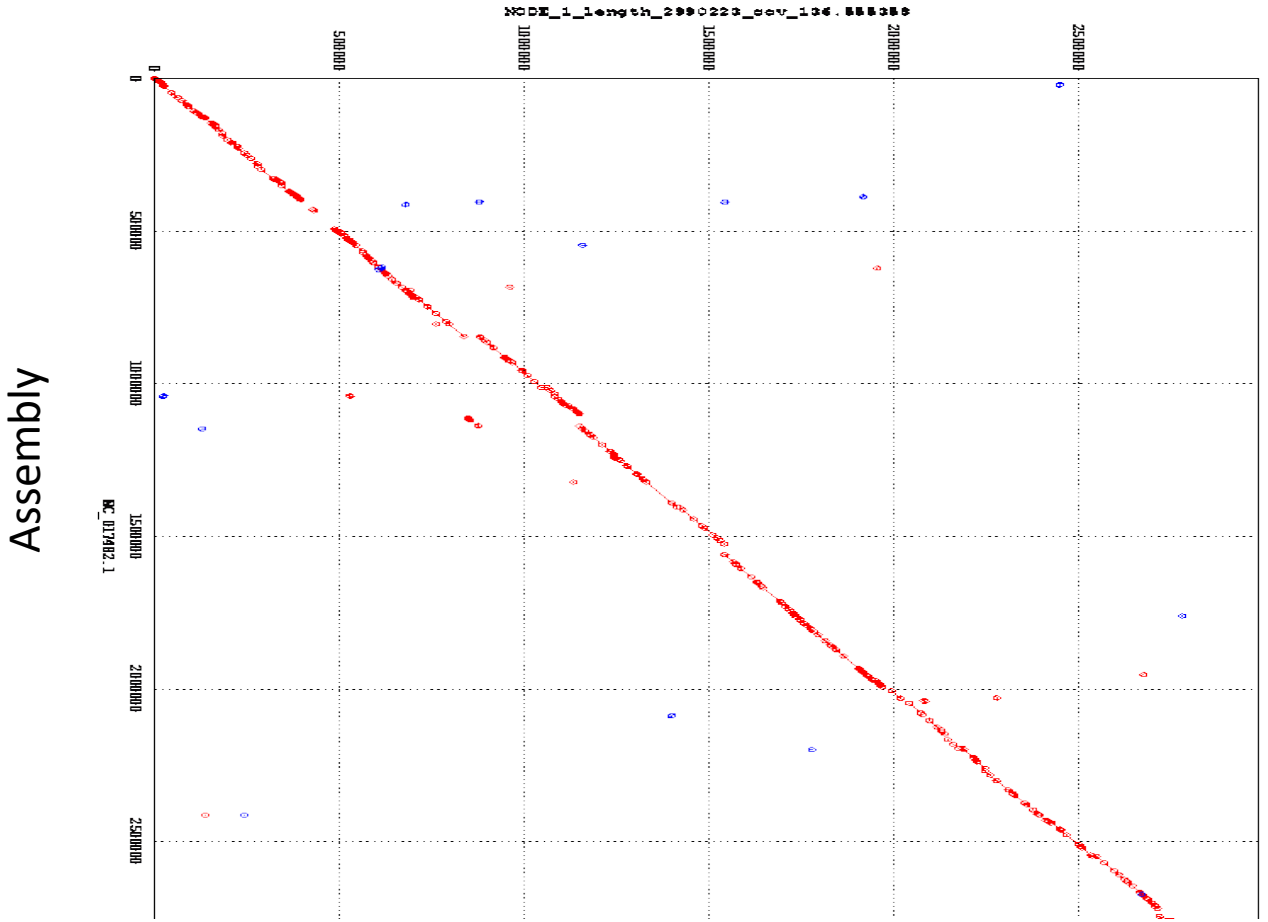
- How can we visualize *whole* genome alignments?
- With an alignment dot plot
  - *N* x *M* matrix
    - Let *i* = position in genome *A*
    - Let *j* = position in genome *B*
    - Fill cell *(i,j)* if A_i shows similarity to B_j

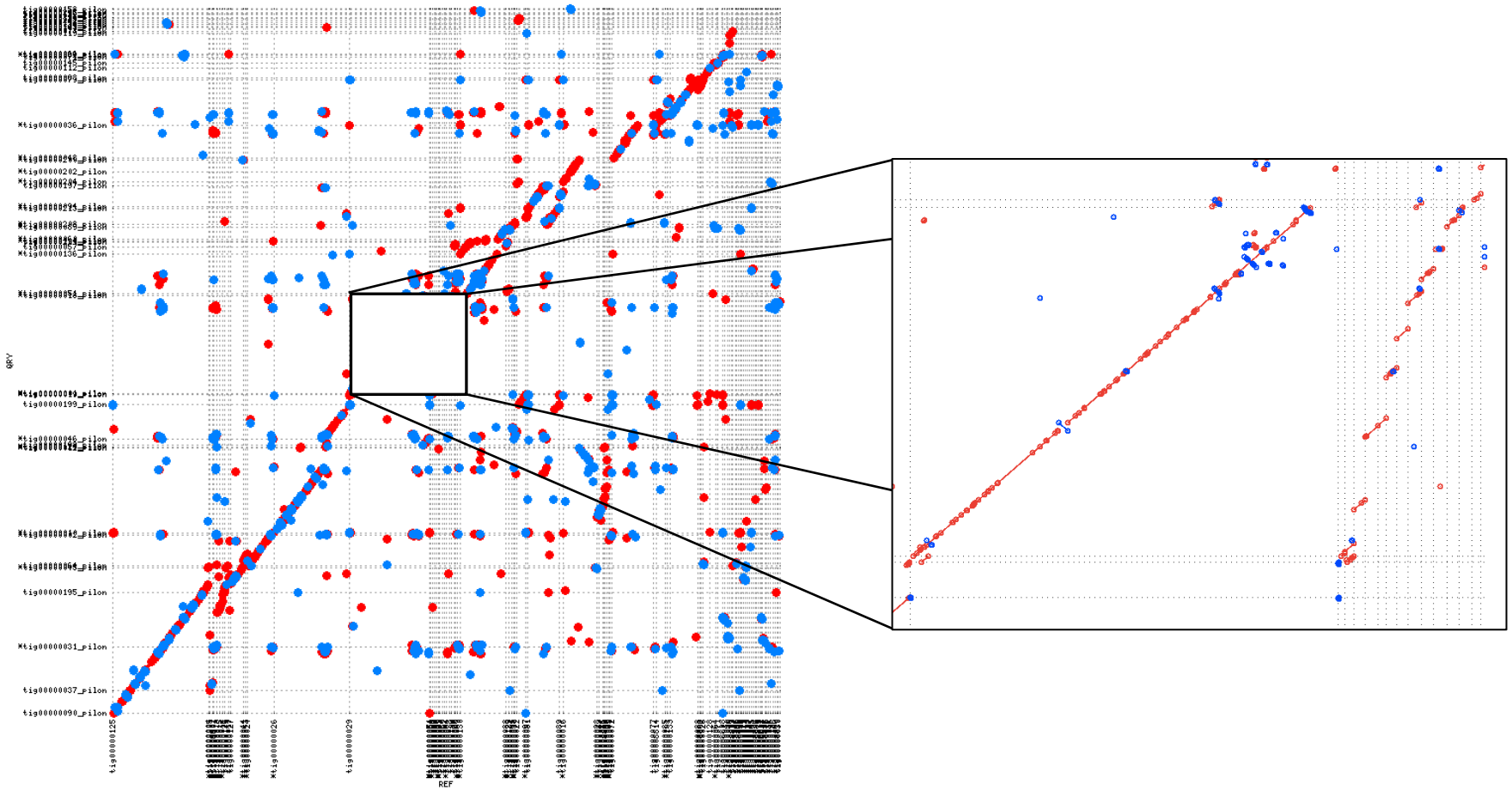


  - A perfect alignment between *A* and *B* would completely fill the positive diagonal



http://mummer.sourceforge.net/manual/AlignmentTypes.pdf

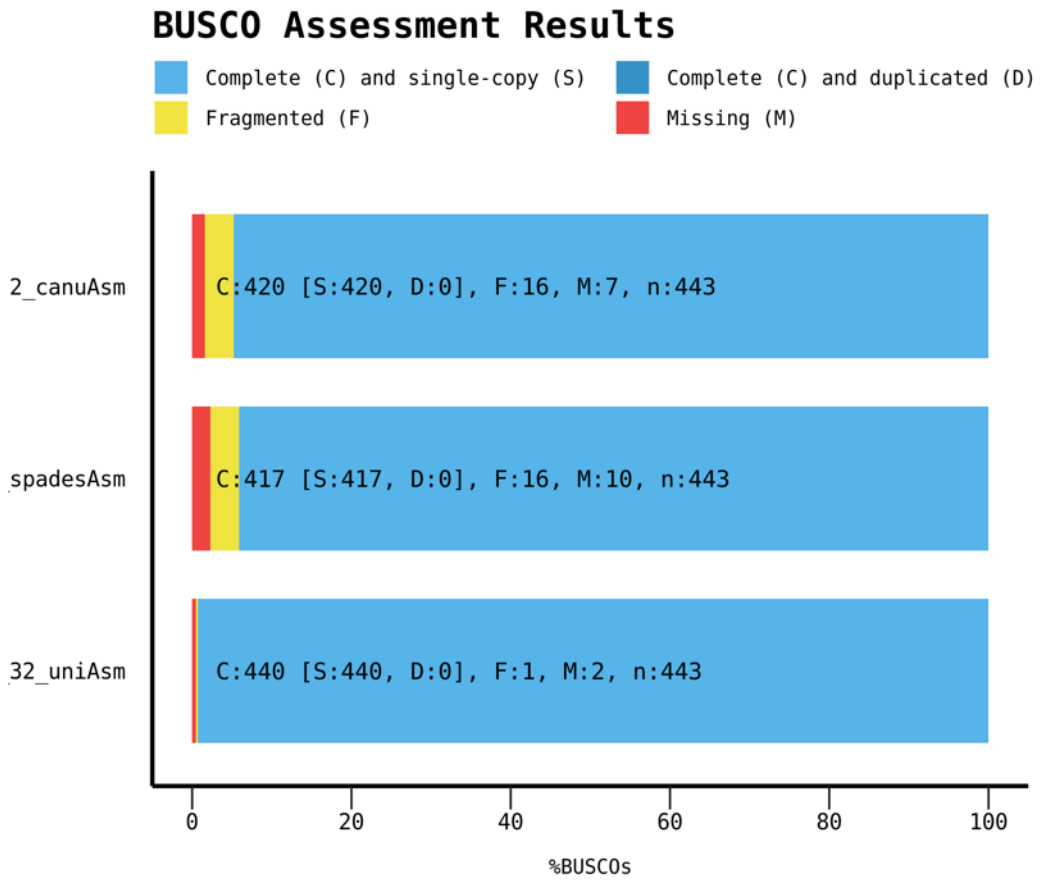

Assembly
Reference
NODE_1_length_2990223_cov_136.555358
0
500000
1000000
1500000
2000000
2500000
NC_017402.1


QRY
REF

BUSCO: conserved gene sets

BUSCO: From Evgeny Zdobnov's group, University of Geneva

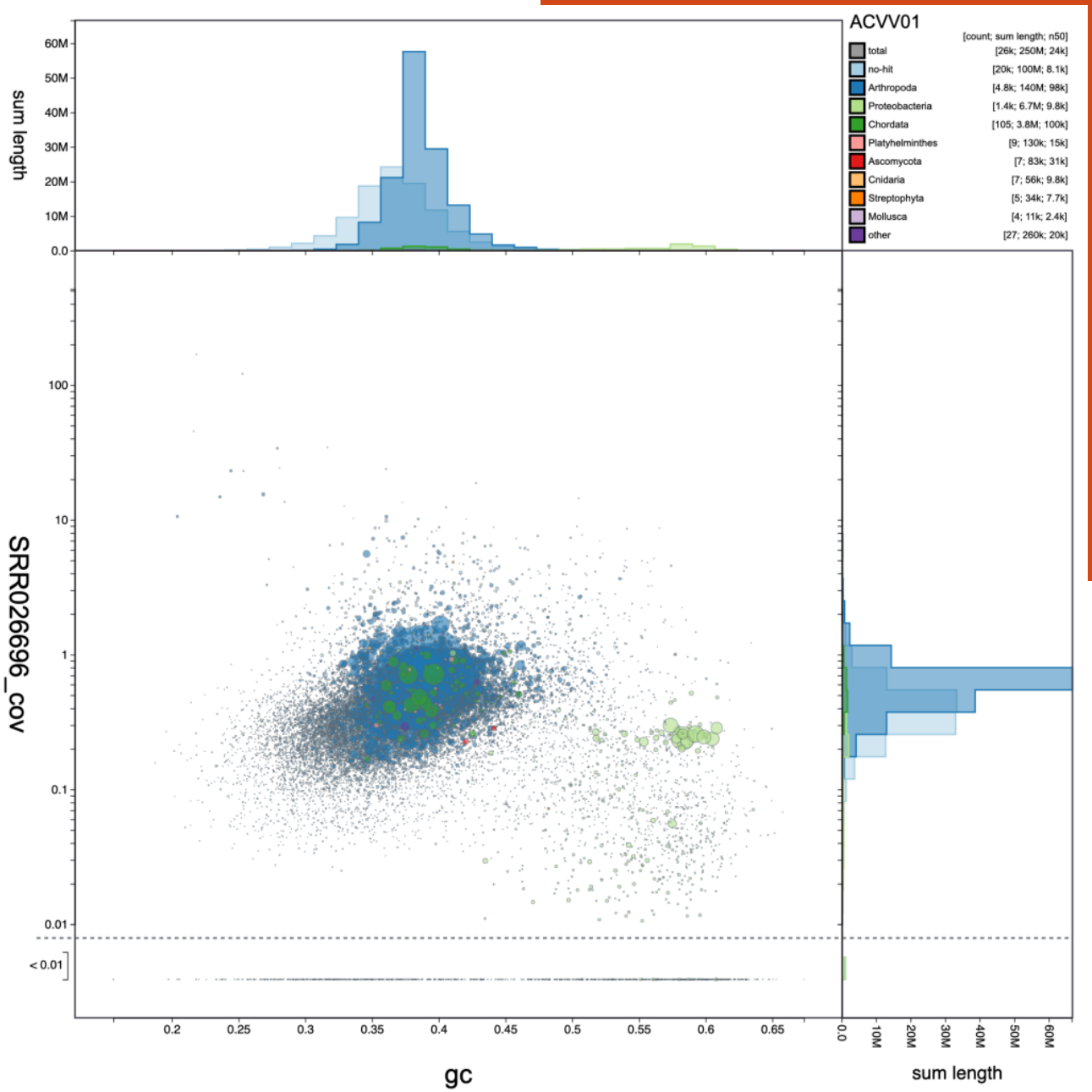
Coverage is indicative of quality and completeness of assembly



Blob plots

Analyses checking for contaminants, endosymbionts, etc.

Interactive version: BlobToolKit



Other tools

Flagger - identify potential mis-assembled regions by alternative libraries (primarily human data)

Klumpy (Catchen lab – UIUC) – identify mis-assemblies in genic regions

New tools being developed for this constantly!

Acknowledgements

Materials from this slide deck include figures and slides from many publications, Web pages and presentations by:

- Carver Biotechnology Center (HPCBio, DNA Sequencing Core)
- M. Schatz, A. Phillipy, T. Seemann, S. Salzberg, K. Bradnam, D. Zerbino, M. Pop, G. Sutton, Nick Loman, Carson Holt, Ryan Wick.
- I highly recommend Ben Langmead's teaching materials; he has a ton fabulous (and much more in-depth) notes on his lab page: http://www.langmead-lab.org/teaching-materials/
- Thank you!