

Genome Assembly

CHRIS FIELDS

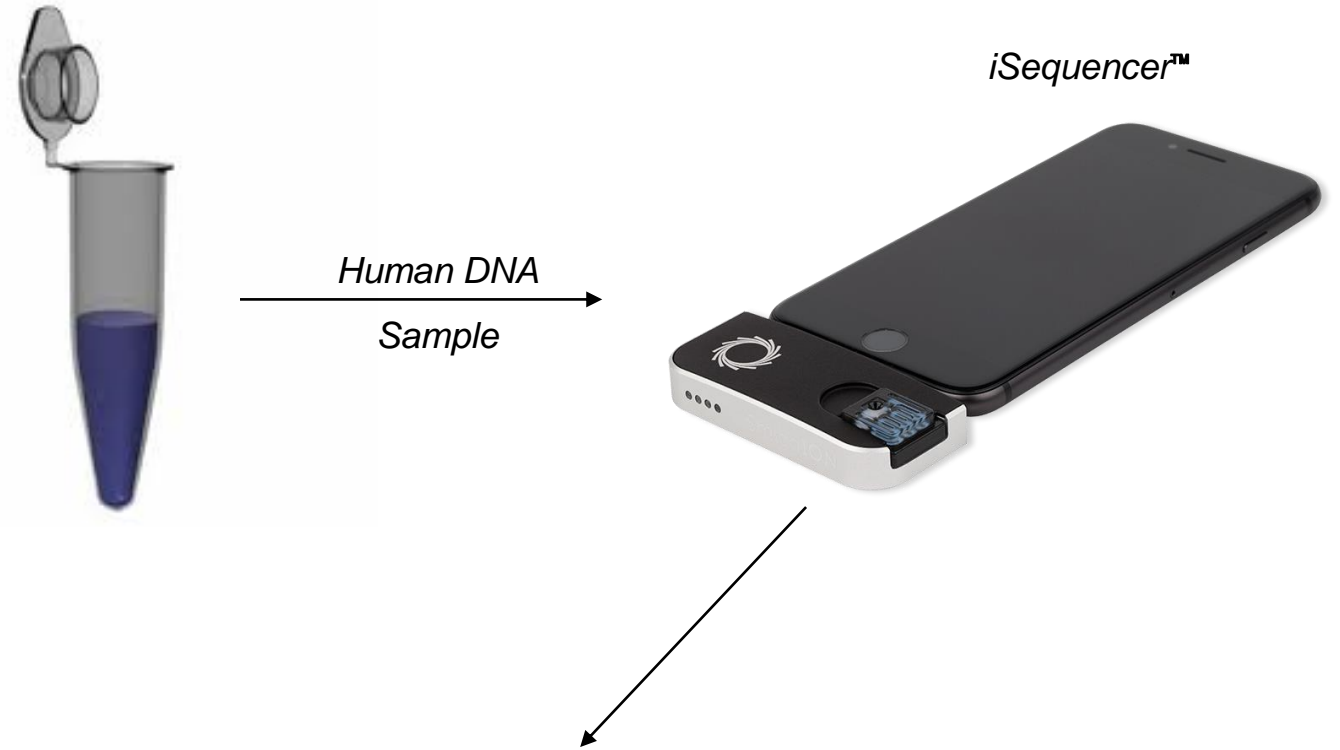
MAYO-ILLINOIS COMPUTATIONAL GENOMICS WORKSHOP
JUNE 6, 2022

Overview

- *What is a genome assembly?*
- *Sequencing technologies (2022)*
- *General steps in a genome assembly*
- *Planning an assembly project*
- *Assembly assessment*
- *Annotation*

Ideal World!

I wouldn't need to give this talk!



```
AGTCTAGGATTCGCTACAGAT
TCAGGCTCTGAAGCTAGATCG
CTATGCTATGATCTAGATCTC
GAGATTCGTATAAGTCTAGGA
TTCGCTATAGATTCAGGCTCT
GATATAT
```

**46 complete,
haplotype-
resolved,
chromosome
sequences**

Ideal World!

But we may not be too far from this!

Science,
March 2022

Time, May 2022

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

🔒 | SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

f t in r s e

The complete sequence of a human genome

SERGEY NURK , SERGEY KOREN , ARANG RHIE , MIKKO RAUTIAINEN , ANDREY V. BZIKADZE , ALLA MIKHEENKO, MITCHELL R. VOLLGER , NICOLAS ALTE-MOSE , 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


☰

TIME

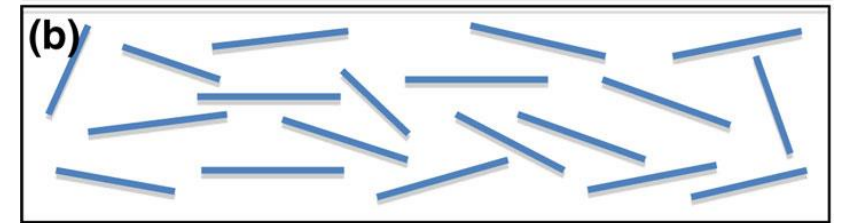
SUBSCRIBE

← THE 100 MOST INFLUENTIAL PEOPLE OF 2022

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



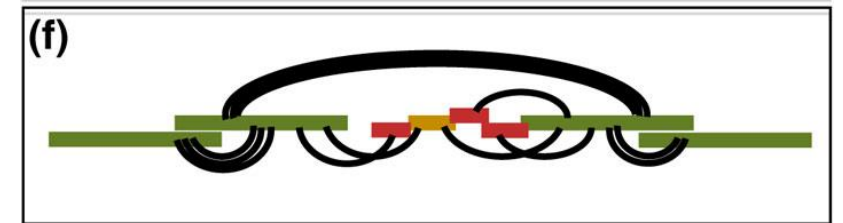
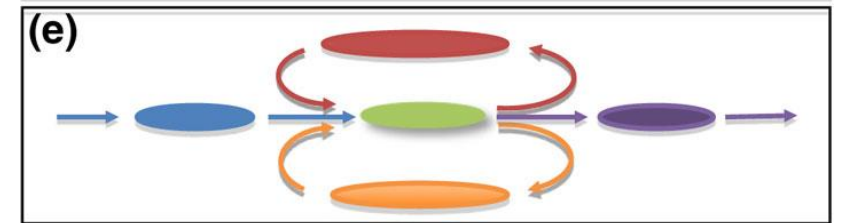
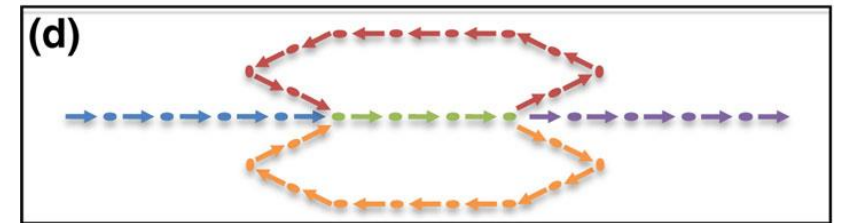
- (a) DNA is collected from the biological sample, fragmented, and sequenced.
- (b) Output from the sequencer consists of many millions/billions of (possibly short) unordered DNA fragments from random positions in the genome.
- (c) Fragments are compared with each other in some way to discover how they overlap.
- (d) The overlap relationships are captured in a large assembly graph
- (e) The graph is refined to correct errors and simplify
- (f) Finally, additional information such as mates, markers and other long-range information can be used to order and orient the initial assembly (contigs) into large scaffolds



(c)

```

..AGCCTAGACACAGGATGCGCGAGT
                GGATGCGCGAGTTCGCATACCGT...
  
```



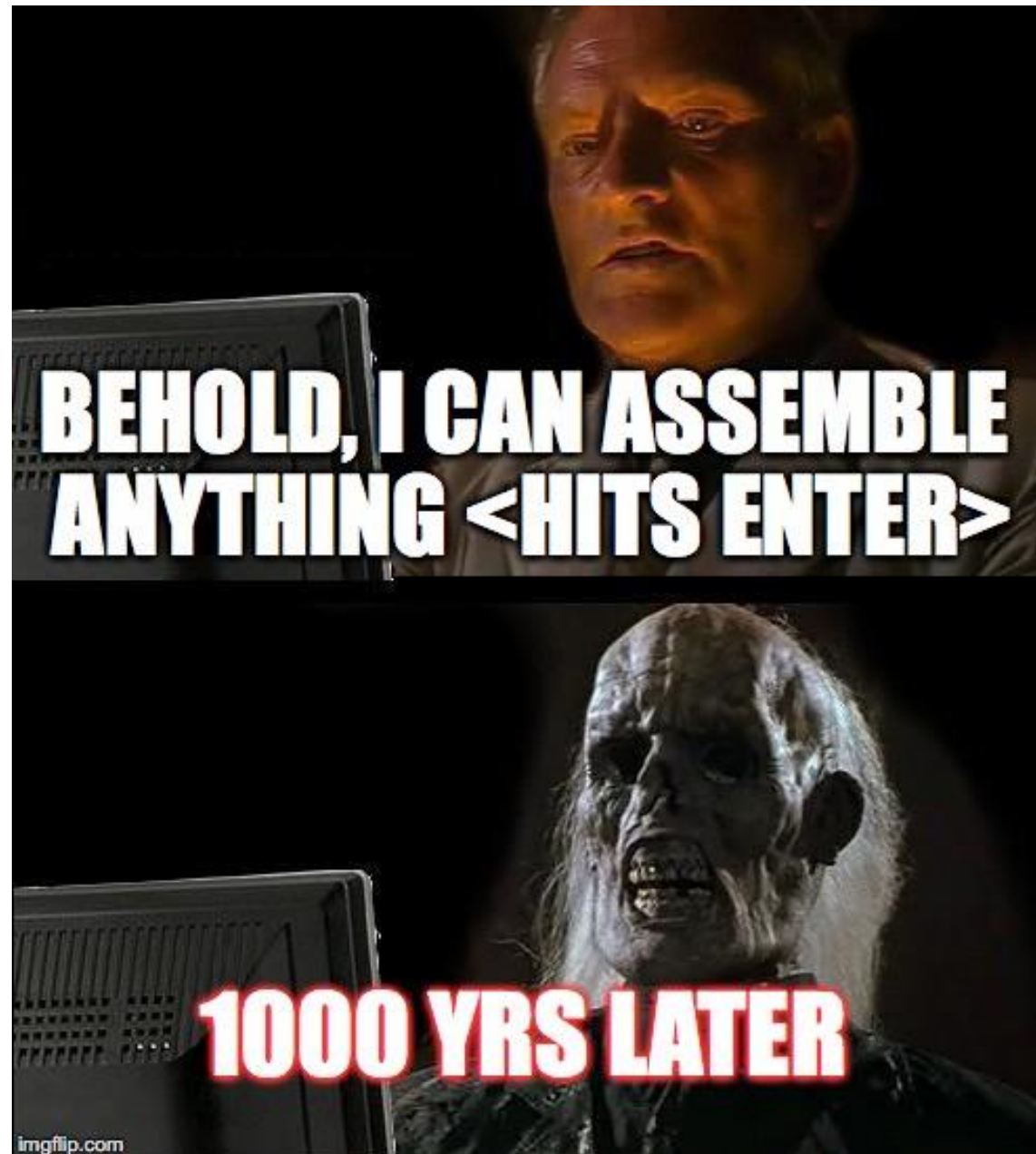
Let's Do a Genome Assembly!

- Sequence a sample, and have the computer do the rest?
- How do you find overlaps between sequences (when you have millions to billions of them)?
 - *You compare them all (overlapping pieces)*
 - *You find shorter perfectly overlapping segments*
 - *Faster but has a lot of assumptions!!!*
- How do you store all this information?
- How long does it take?

Resource needs

- Technology dependent!
- Memory + CPU
- Short reads (billions of reads)
 - Sequencing costs - \$\$
 - Compute costs - \$\$\$\$\$
 - Results – fragmented, requires significant 'cleanup'
- Long error-prone reads
 - Sequencing costs - \$\$\$\$
 - Compute costs - \$\$\$\$
 - Results – better quality, but can't easily phase
- Long accurate reads
 - Sequencing costs - \$\$\$\$\$
 - Compute costs - \$\$\$
 - Results – (partly) phased diploid assembly***

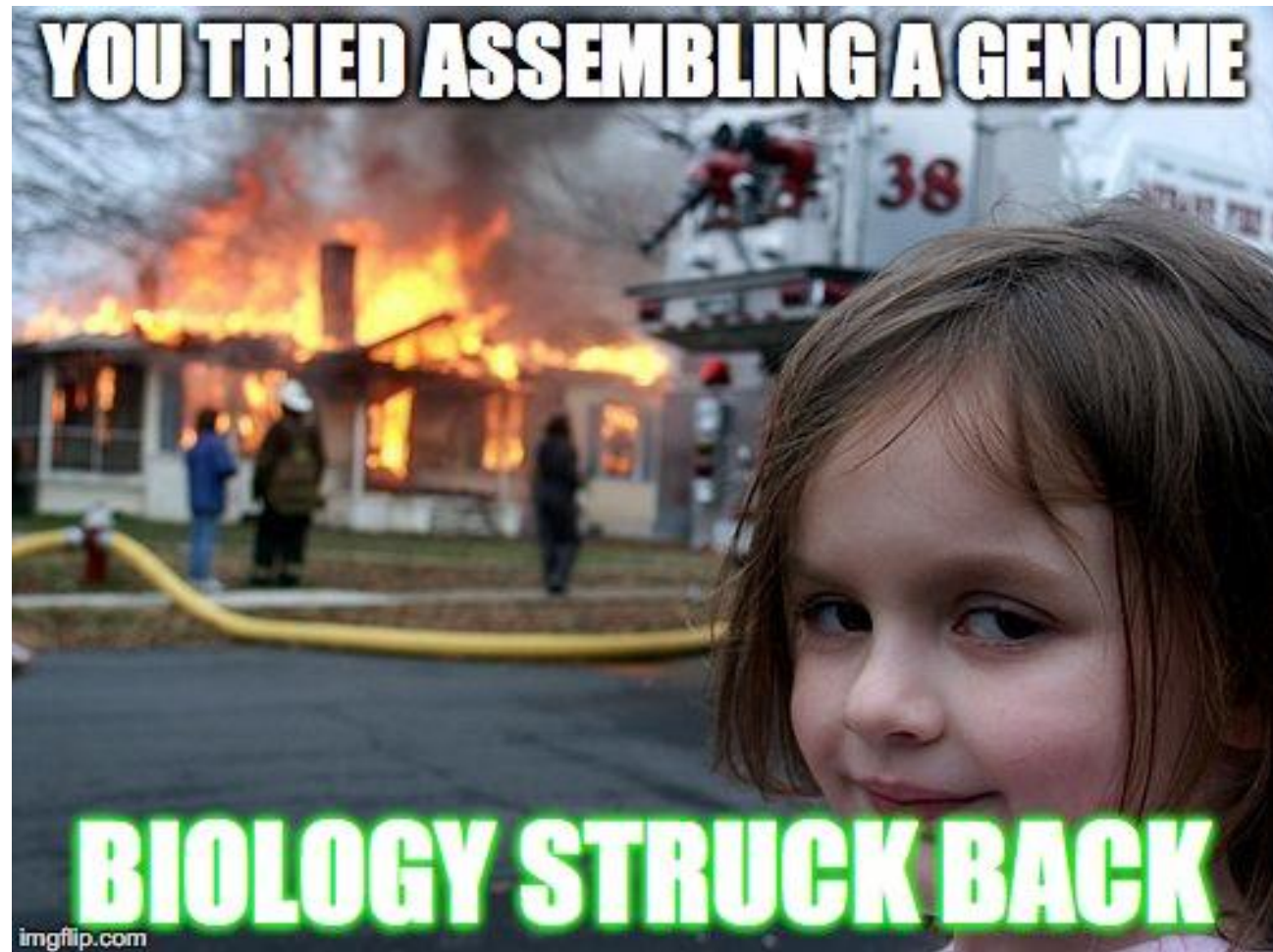
*** - doesn't help much if you have higher ploidy! (though this will likely change)



Results

- You spent your entire grant on getting sequence data and buy a monster multi-core high-memory server
- You assemble your genome with your favorite genome assembly tool
- You waited a week to a month and you now have results!
- Wait, why do I have a million scaffolds? And why is my server on fire?!?

Biology



Current Sequencing Technologies

Illumina

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

Can be paired-end (sequence ends of fragments)

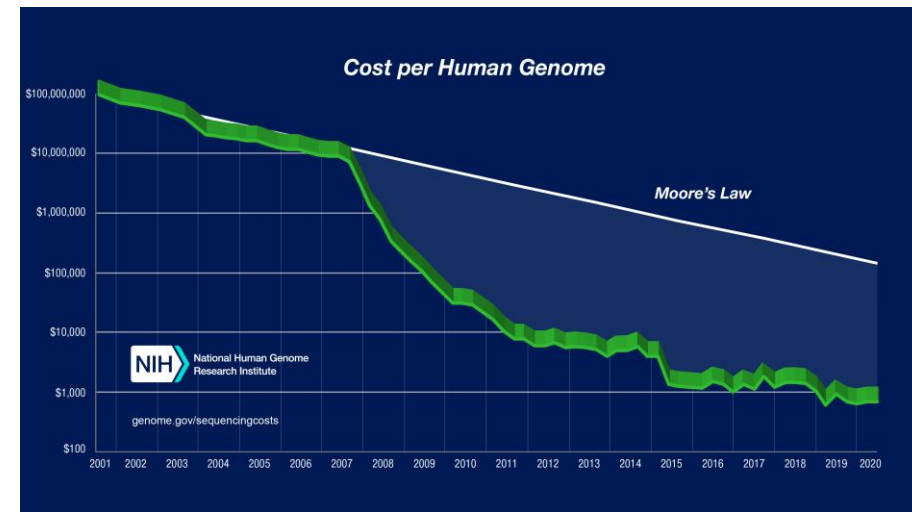
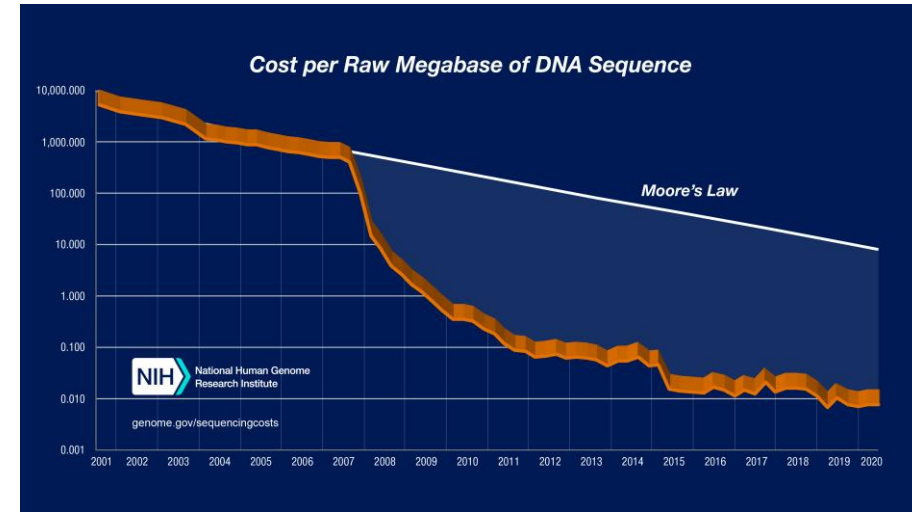
Advantages

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

Disadvantages

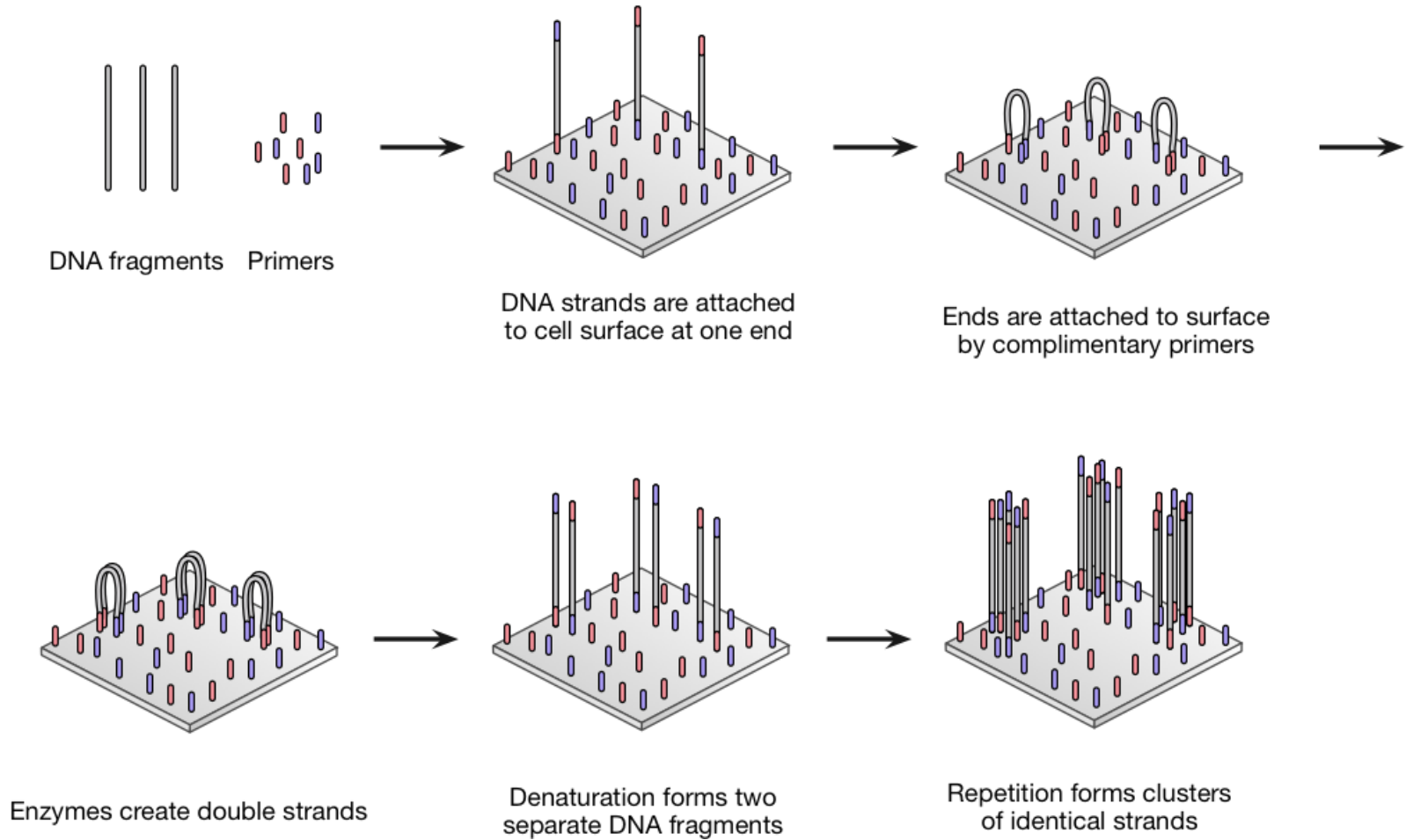
- Requires high depth for many applications (**50x + for assembly**)
- Sequence length (100-150nt reads) problematic for repeats
- Maximum fragment length (<800bp) is an issue

<https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>



Illumina

<https://www.atdbio.com/content/58/N-ext-generation-sequencing>



'Long reads'

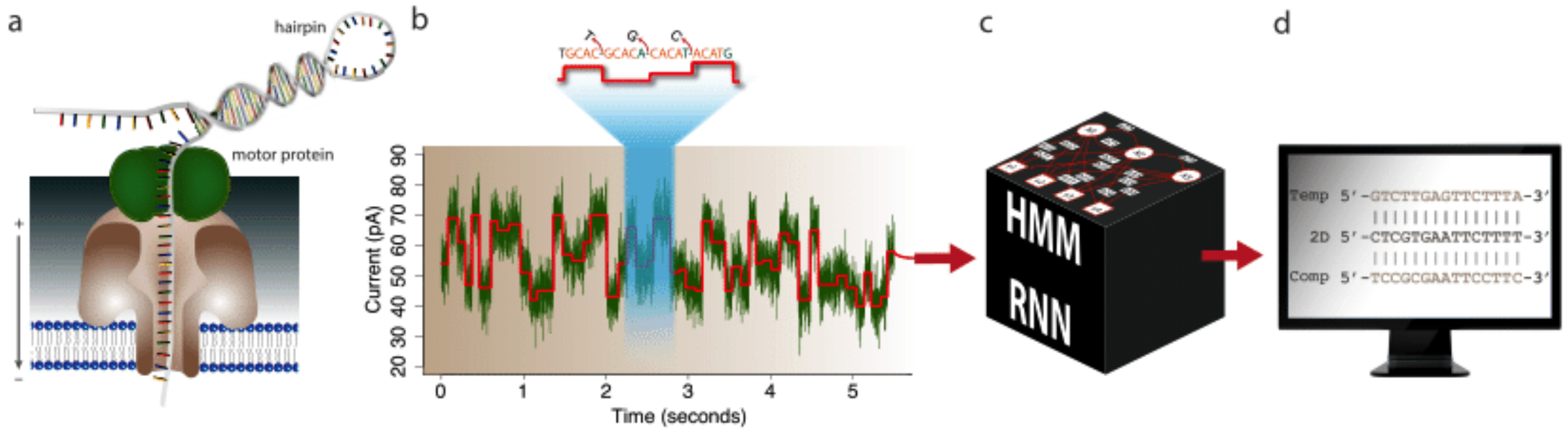
Pacific
Biosciences
(PacBio)



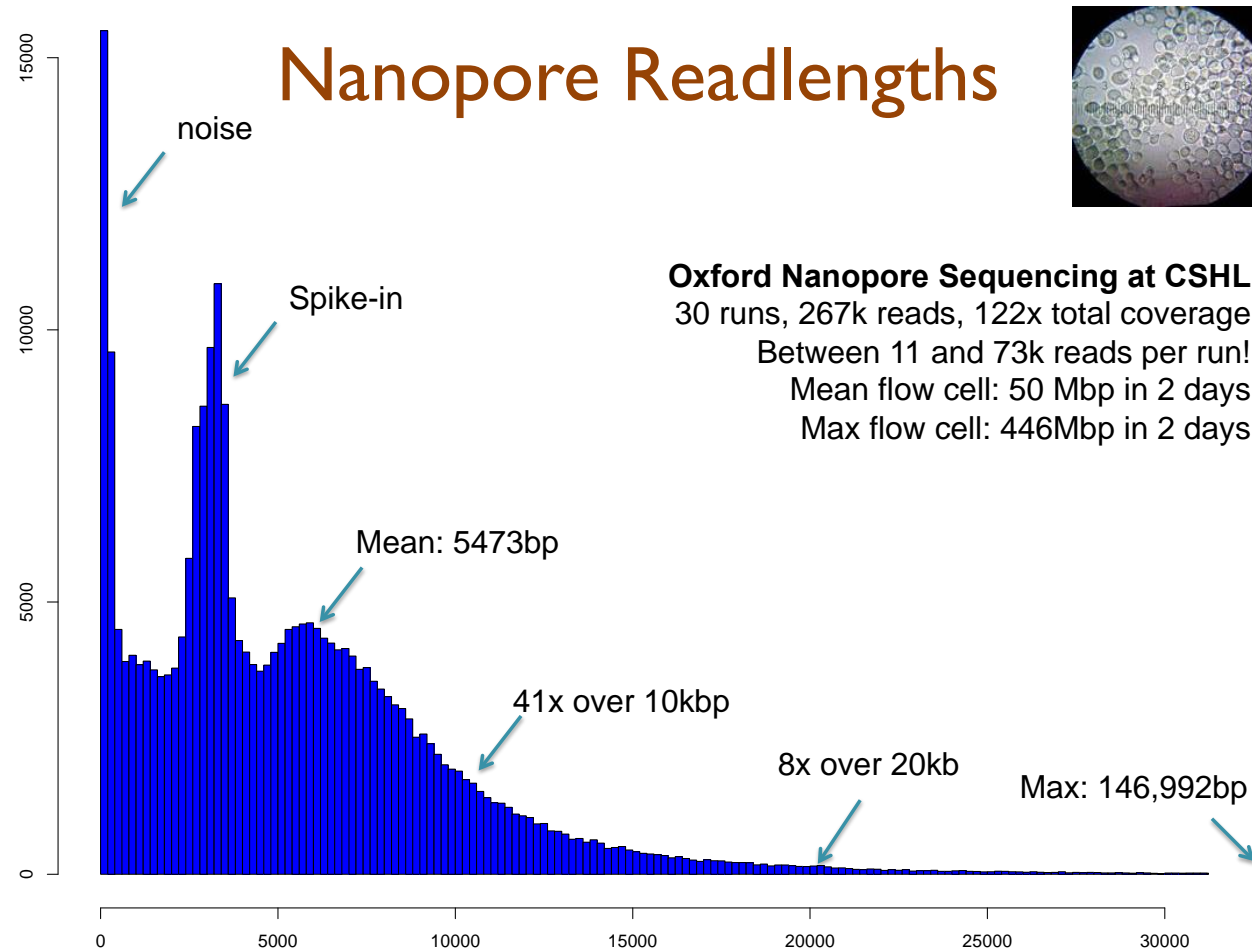
Oxford
Nanopore
(ONT)



MinION

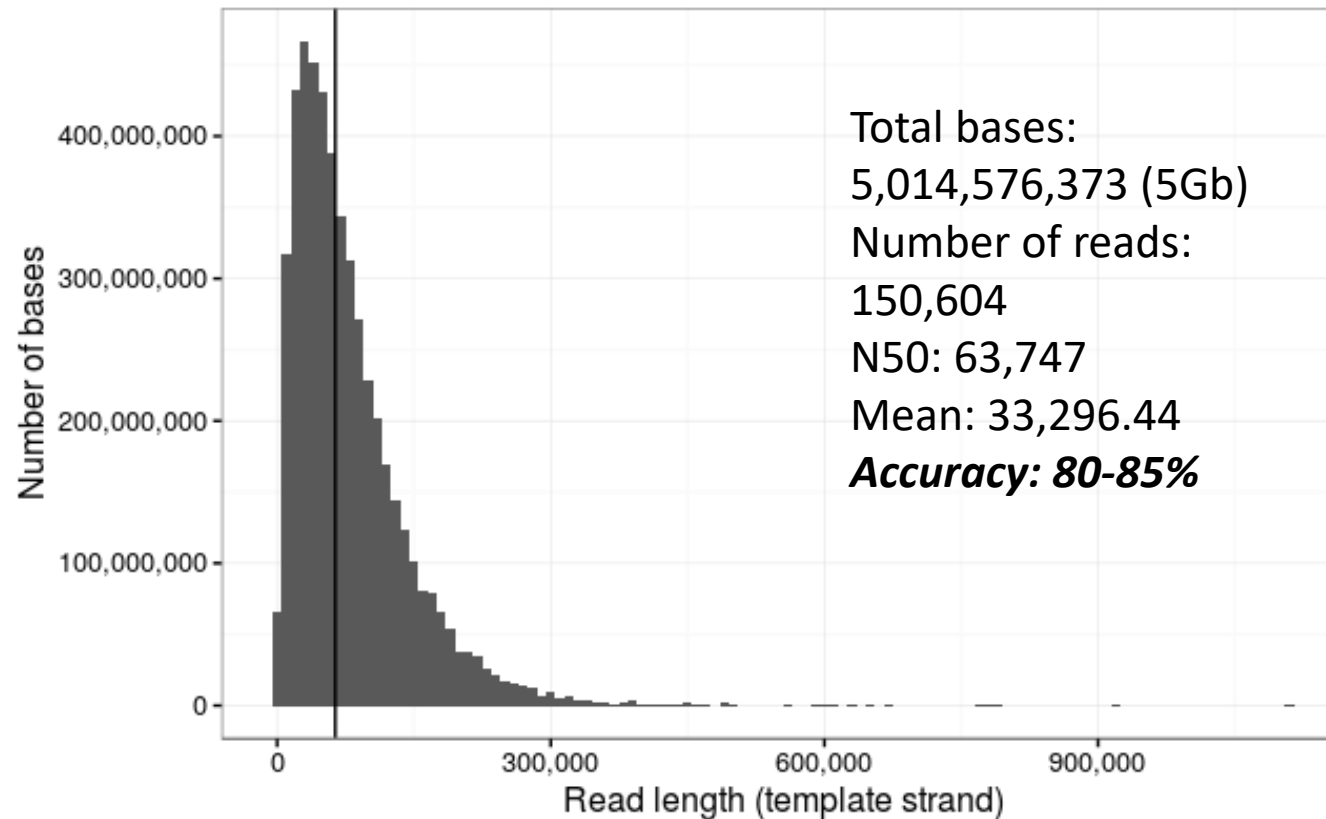


Oxford Nanopore

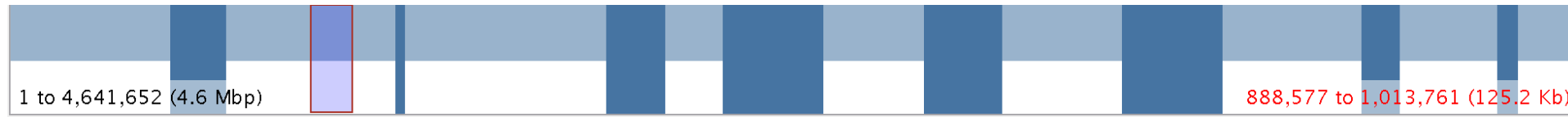


Whale watching: *E. coli*

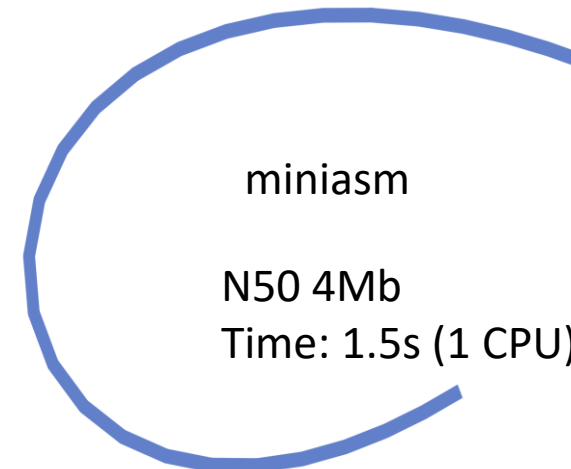
2017



E. coli: genome assembly in 8 reads



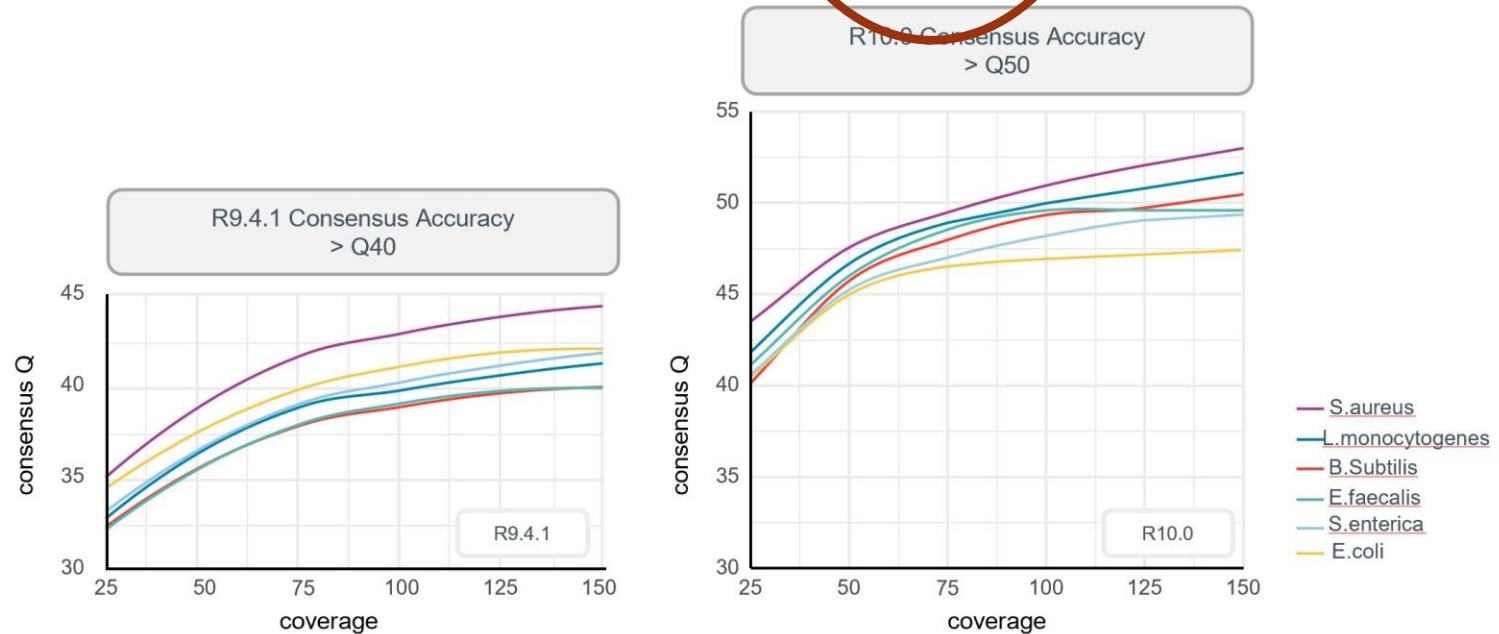
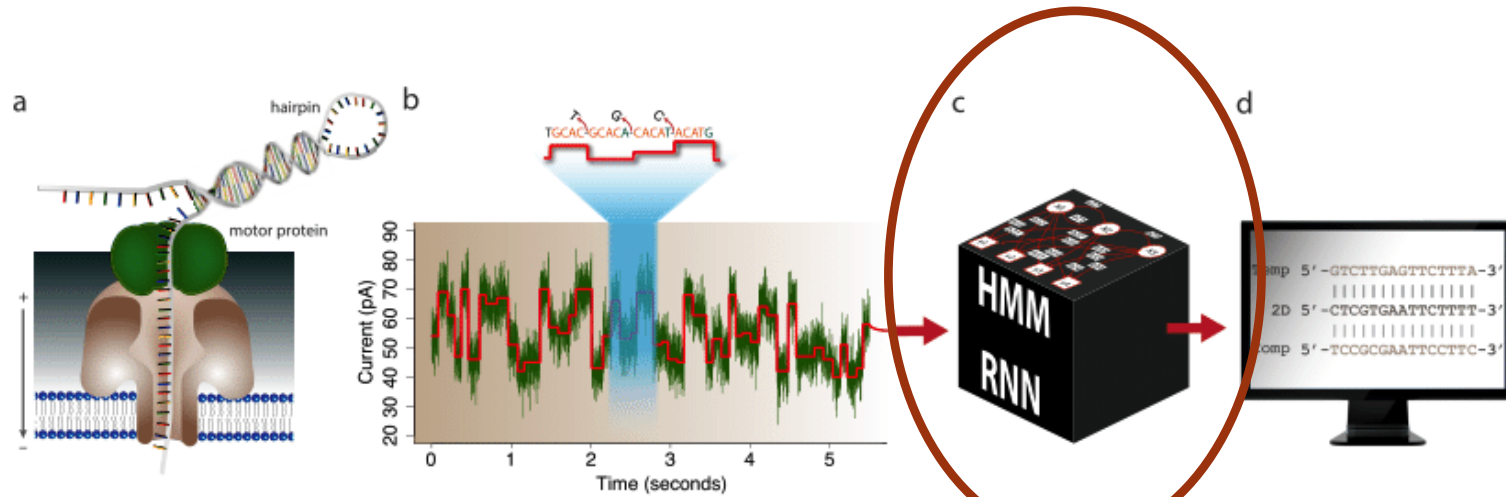
Read	Length	Ref start	Ref end	Time (m)
1	876991	4398844	634183	32.48
2	696402	470003	1166405	25.79
3	799047	1137438	1936485	29.59
4	642071	1759431	2401502	23.78
5	826662	2106227	2932889	30.61
6	883962	2699626	3583588	32.73
7	825191	3285196	4110387	30.56
8	463341	3995967	4459308	17.16



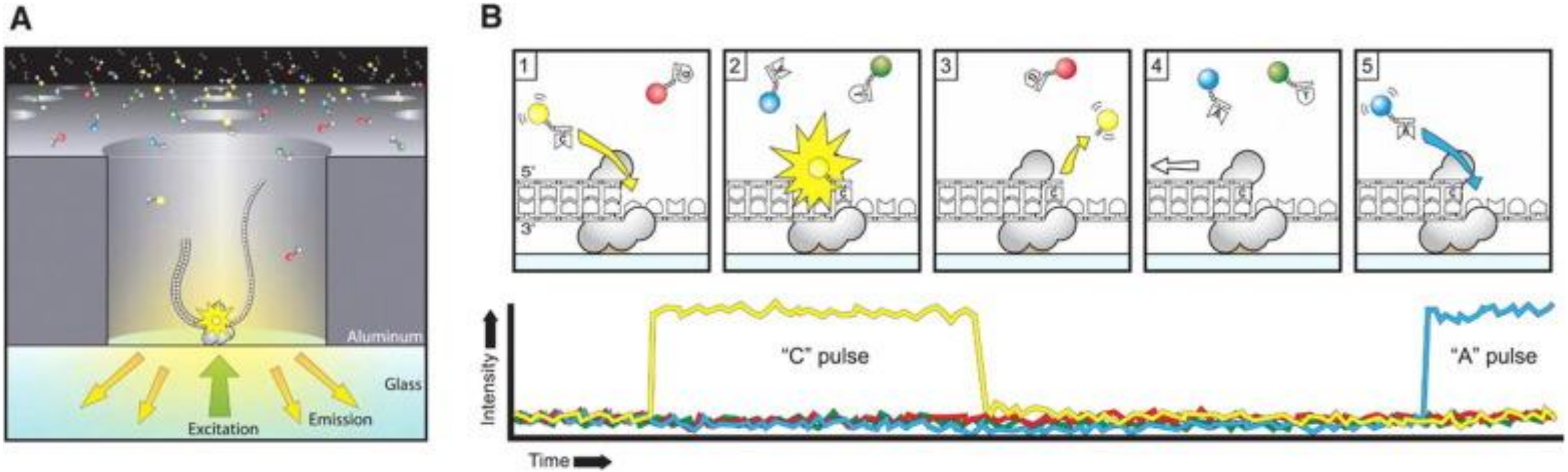
1x coverage!

Oxford Nanopore

2022



Note that accuracy is based on comparison to human data!



<https://www.sciencedirect.com/science/article/pii/S1672022915001345?via%3Dihub>

Pacific Biosciences

PacBio Continuous Long Read Sequencing (aka PacBio CLR)

Optimized for length

25-50kb long reads

90% accuracy

Yields of ~125Gb+ per SMRT cell

Need ~50-90x coverage

Needs error correction, polishing

1-2 SMRT cells per human sample

Start with high-quality
double stranded DNA

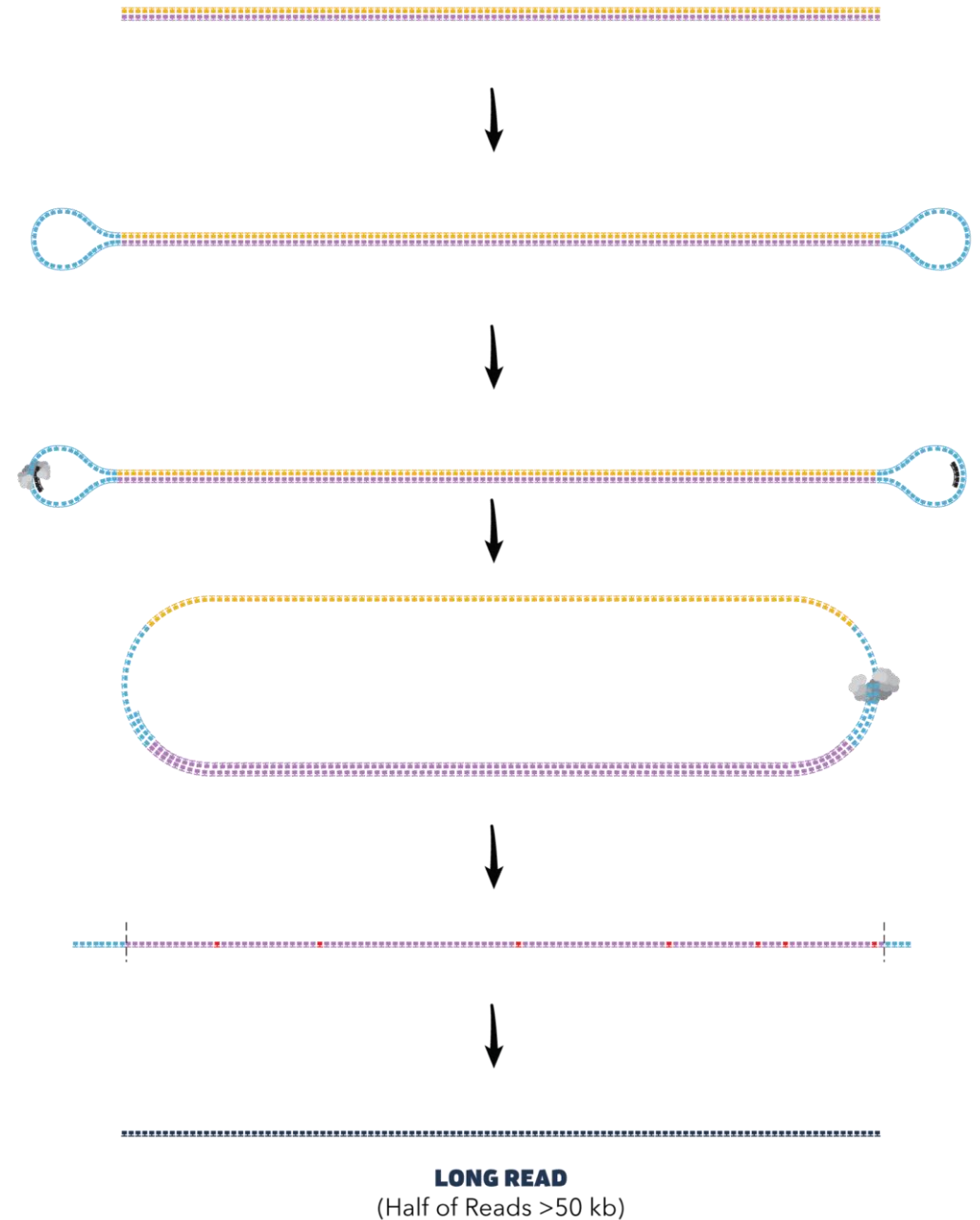
Ligate SMRTbell
adapters and size select

Anneal primers and
bind DNA polymerase

Circularized DNA
is sequenced in a
single pass

The polymerase reads
are trimmed of adapters
to yield subread

During assembly,
consensus is called from
multiple molecules



PacBio Circular Consensus Sequencing (aka PacBio HiFi)

Optimized for accuracy

10-15kb long reads

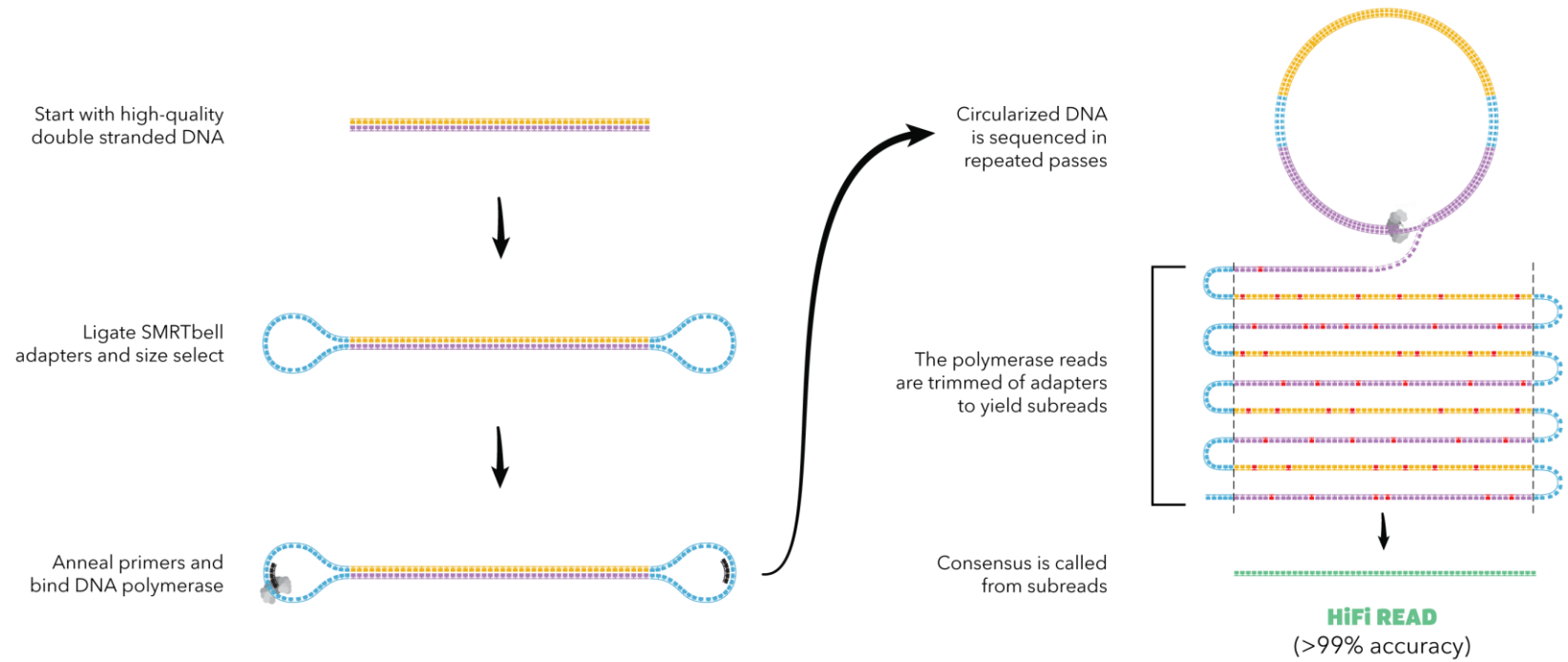
99% accuracy

Yields of ~25Gb per SMRT cell

Need ~25-50x coverage

No error correction/polishing required

~2-3 SMRT cells per human sample



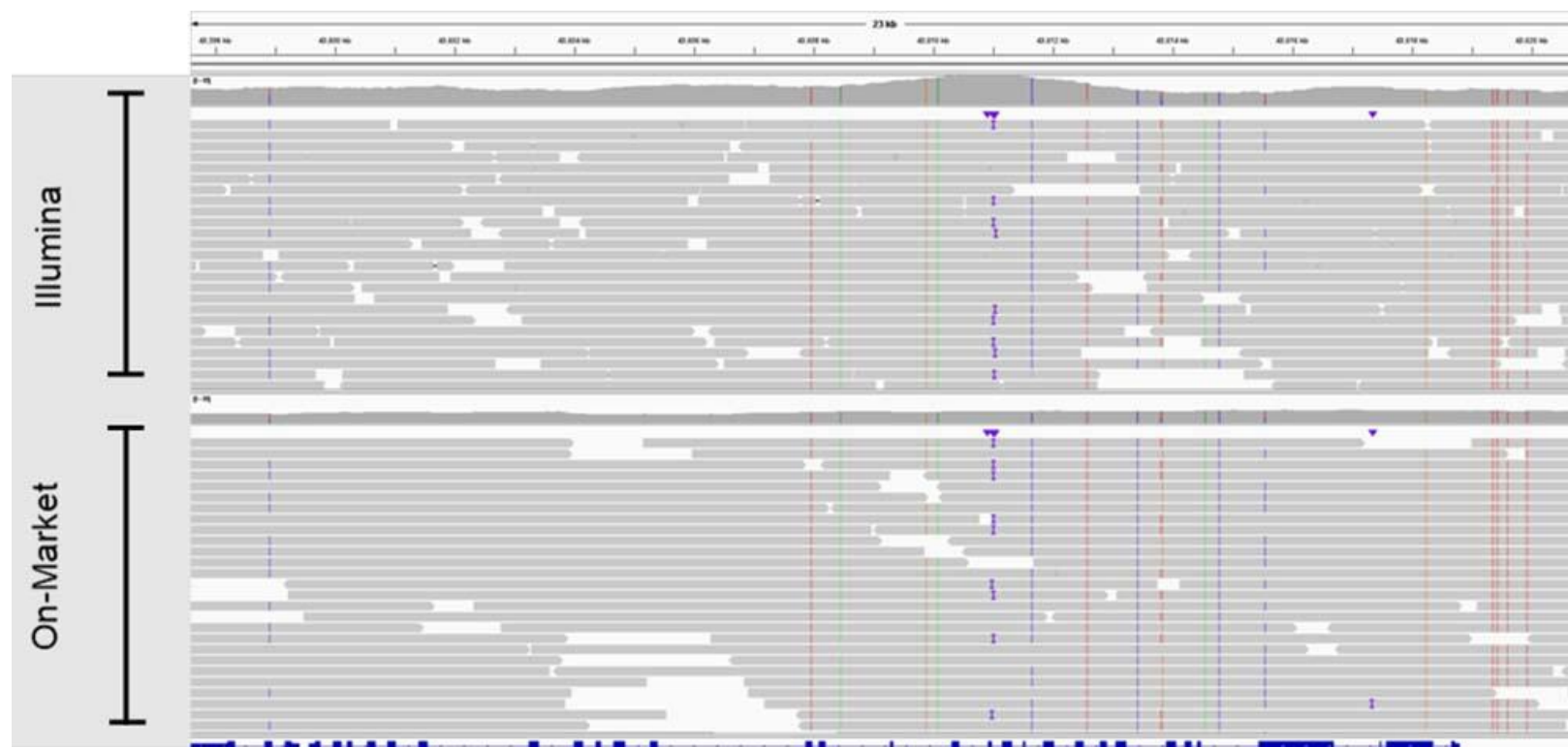
Illumina Infinity

Announced early 2022: "... we are developing a novel, high performance long read assay, code named 'Infinity' that will accelerate access to the remaining ~5% of genic regions that are challenging to map"

- Contiguous reads up to 10 kb
- ~10x the throughput compared to traditional long read technologies
- 90% less DNA input compared to current Long Read methods
- Fully automatable workflow

Early access **2H of 2022**

Very little known about this one so far



[Illumina Infinity announcement](#)

'Long Reads'

Advantages

- Reads can be very long (1kb – 100kb)
- Relatively even coverage of the genome
- (PacBio HiFi) Highly accurate (99%)
- (Oxford) real-time sequencing
- (Oxford) portable

Disadvantages

- Expensive compared to Illumina short reads
- Need very high quality, high MW DNA samples
- Least expensive options are *error-prone*
- Depending on technology, can have *systematic errors* (homopolymer issues), but getting better

Basic Steps for Genome Assembly

Steps

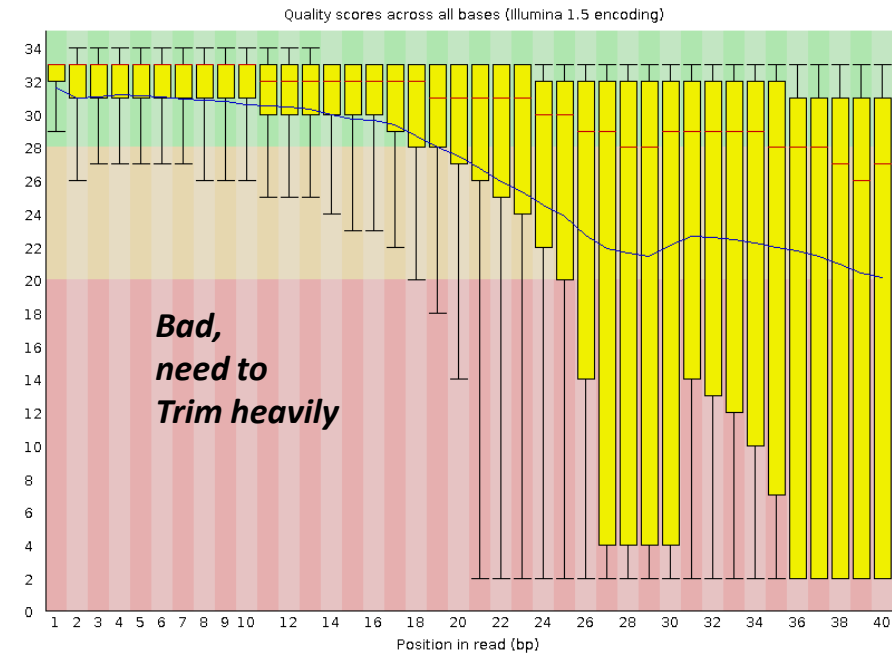
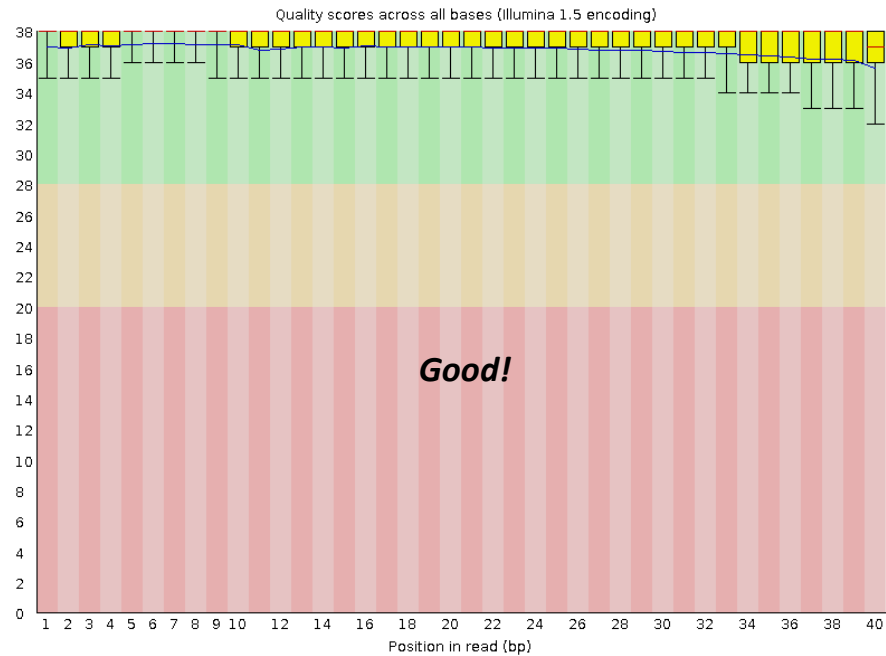
1. Basic DNA sequence cleanup and evaluation (pre-assembly)
2. Contig building
3. Scaffolding
4. Post-assembly processing and analyses

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



Adapters

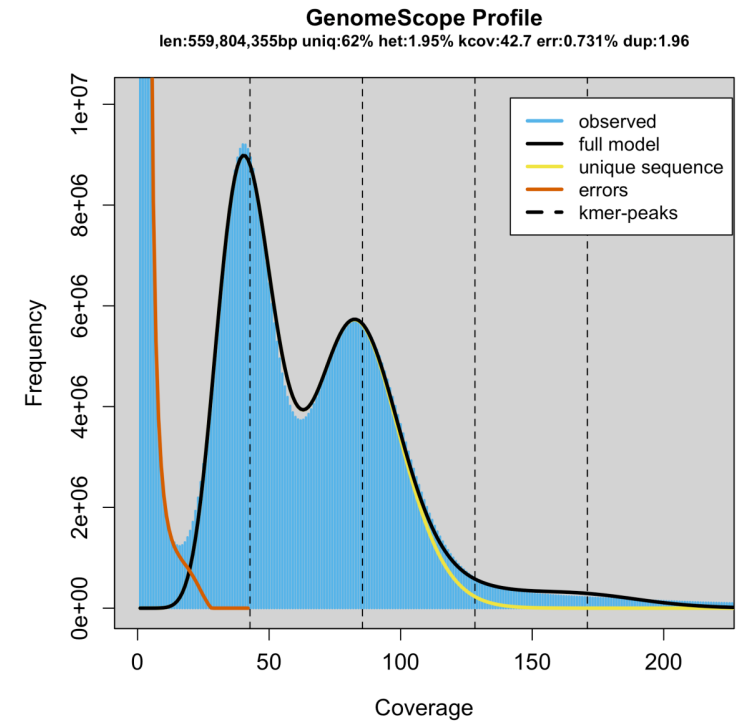
Overrepresented sequences

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)

Other pre-assembly steps

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

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



Starting the assembly

Contig building

Greedy assembly

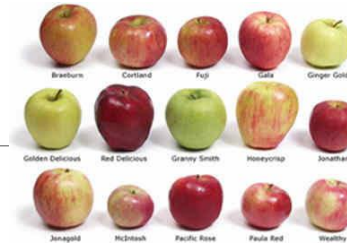
Seed and extend

Overlap graph

de Bruijn graphs

String graphs

..etc etc



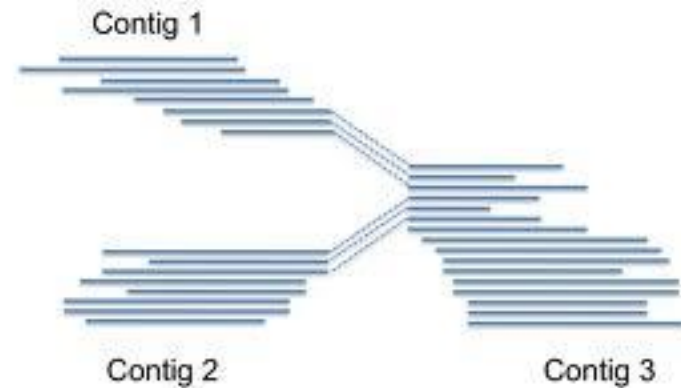
*... all essentially doing similar things,
but taking different 'shortcuts' based on
needs*

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)



Assembly recipe



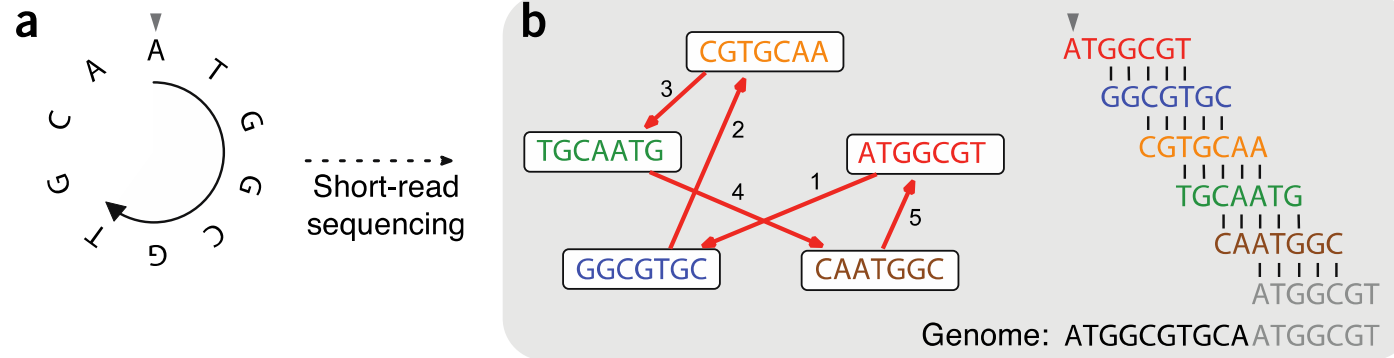
- Find all overlaps between reads
 - hmm, sounds like a lot of work...
- Build a graph
 - a picture of read connections
- Simplify the graph
 - sequencing errors will mess it up a lot
- Traverse the graph
 - trace a sensible path to produce a consensus

Graph

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

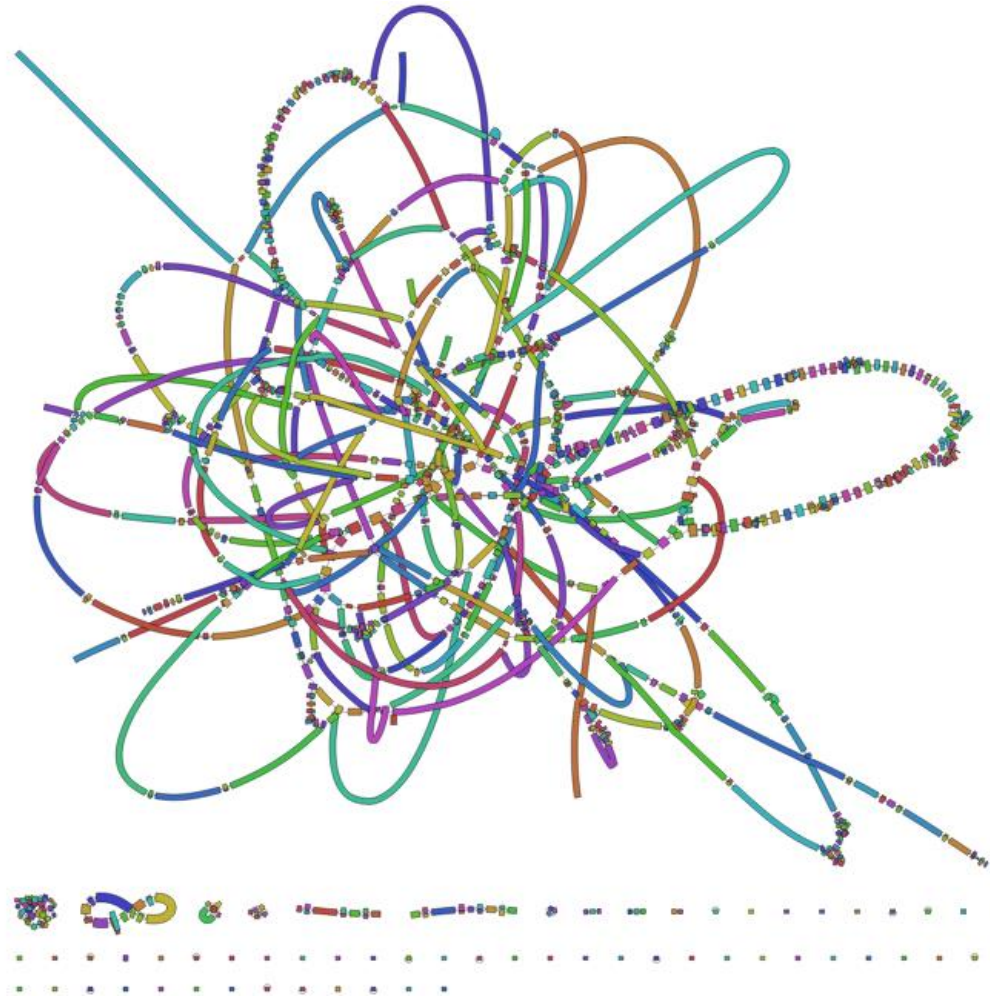
Nodes/Vertices = objects (sequence)

Edges = relationship (overlap)



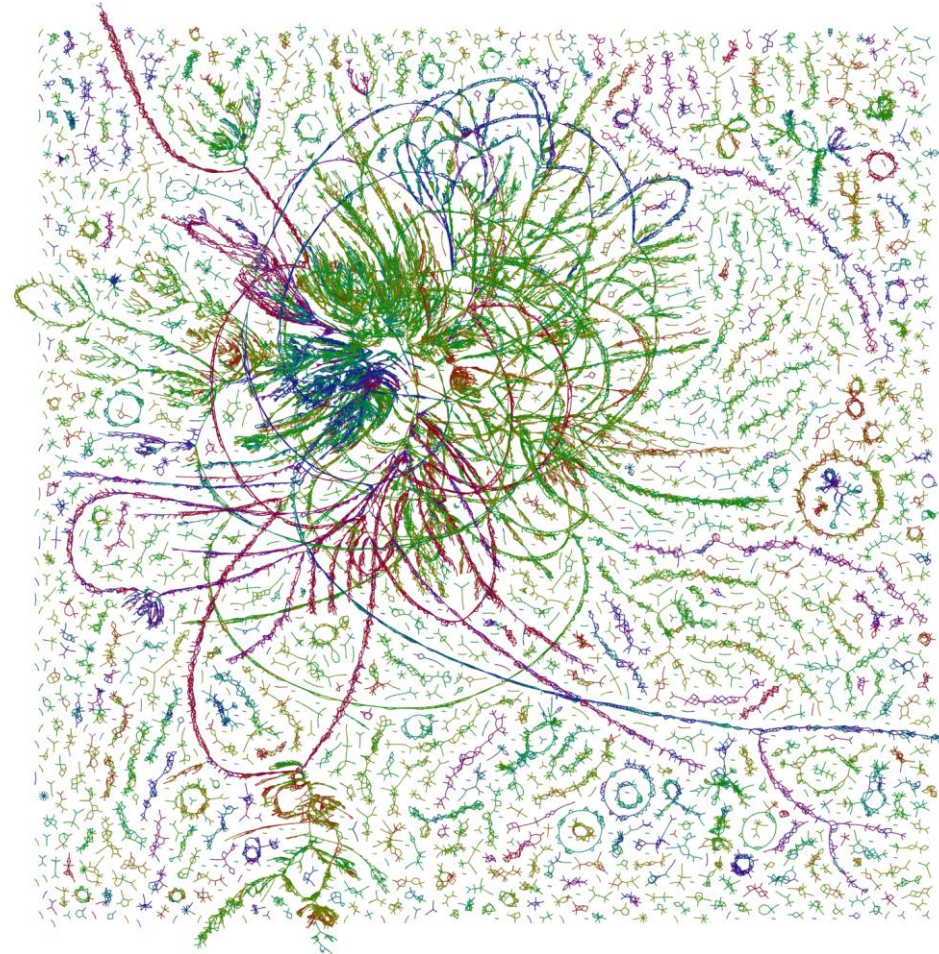
Compeau *et al*, Nature Biotech, 29(11), 2011; [https://en.wikipedia.org/wiki/Graph_\(discrete_mathematics\)](https://en.wikipedia.org/wiki/Graph_(discrete_mathematics))

Simple?



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

Erm...

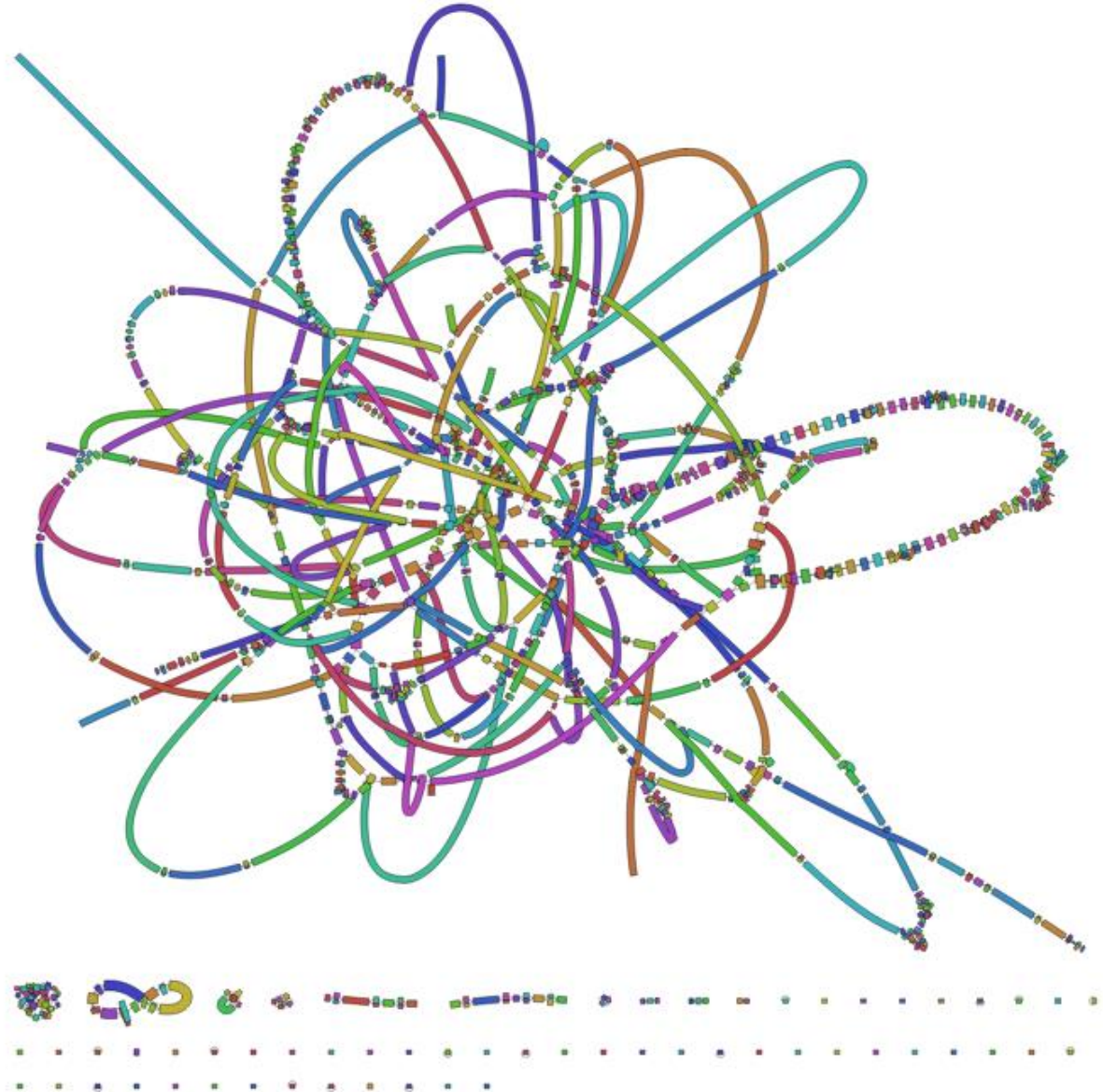


<http://armbrustlab.ocean.washington.edu/seastar>

In essence...

For each unconnected graph:

- **Find a path** which visits each node once
 - This is referred to as a **Hamiltonian path/cycle**
- **Form consensus sequences** from paths
 - use all the overlap alignments
 - each of these collapsed paths is a ***contig***



Overlap Layout Consensus Assembly

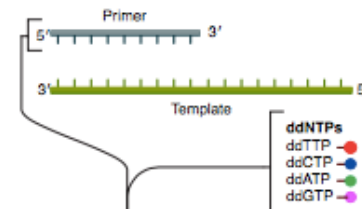
Used for longer read data

Sanger

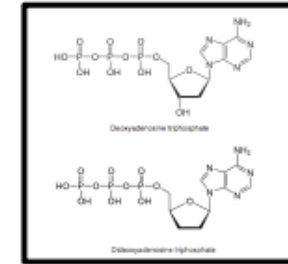
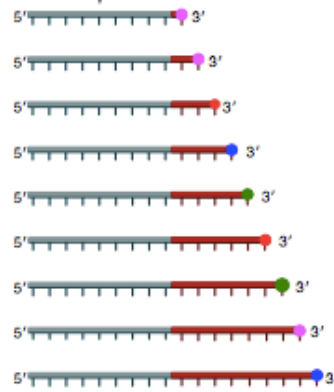
Newer variants for PacBio and Oxford Nanopore

① Reaction mixture

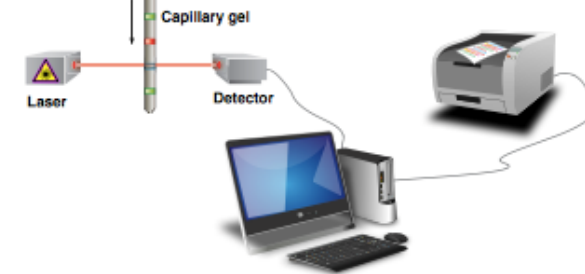
- Primer and DNA template
- DNA polymerase
- ddNTPs with flouochromes
- dNTPs (dATP, dCTP, dGTP, and dTTP)



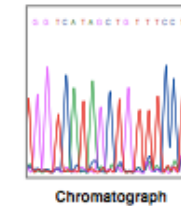
② Primer elongation and chain termination



③ Capillary gel electrophoresis separation of DNA fragments



④ Laser detection of flouochromes and computational sequence analysis

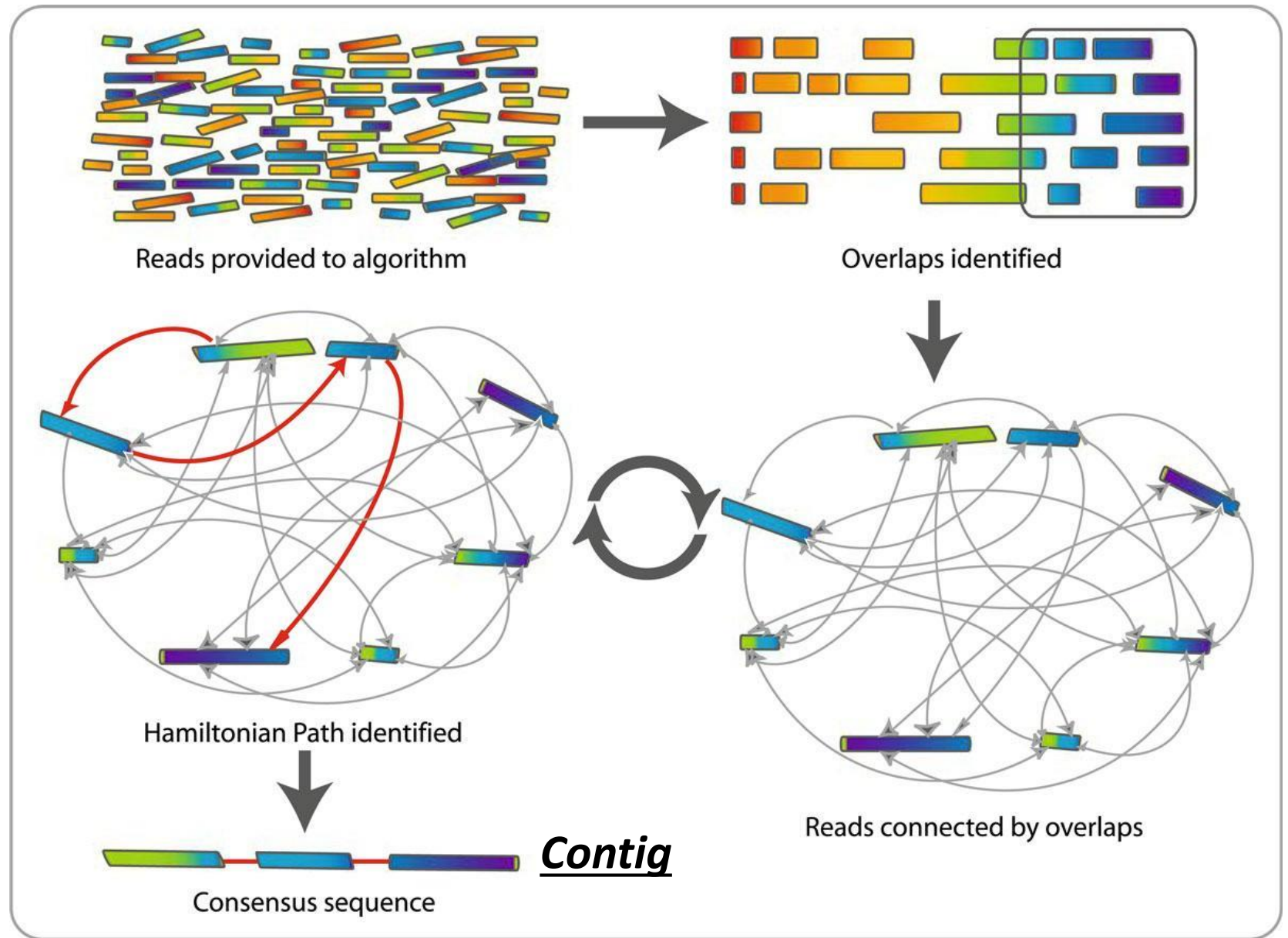


By Estevezj - Own work, CC BY-SA 3.0,
<https://commons.wikimedia.org/w/index.php?curid=23264166>

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



OLC assembly steps

Calculate *overlays*

- Can use BLAST-like methods, but finding common strings (**k-mers**) more efficient

Assemble *layout* graph, try to simplify graph and remove nodes (reads) – find Hamiltonian path

Generate *consensus* from the alignments between reads (overlays)

Some OLC-based assemblers

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

HiCanu – PacBio HiFi assembler

Newbler, a.k.a. GS de novo Assembler - designed for 454 sequences, but works with Sanger reads

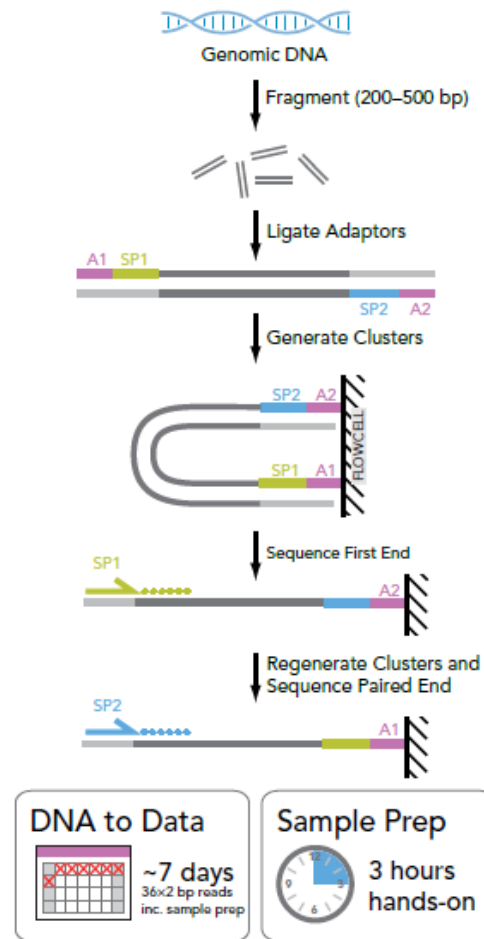
Hifiasm – a hybrid *diploid* assembler

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)

Figure 6B: Paired-End Sequencing



Adaptors 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.

Shredded Book Reconstruction

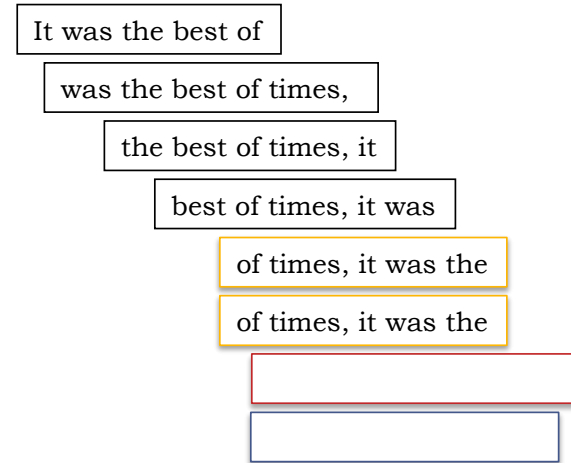
- Dickens accidentally shreds the first printing of A Tale of Two Cities
 - Text printed on 5 long spools

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

- 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 $k-l$ words

Original Fragment

It was the best of

Directed Edge

It was the best → was the best of

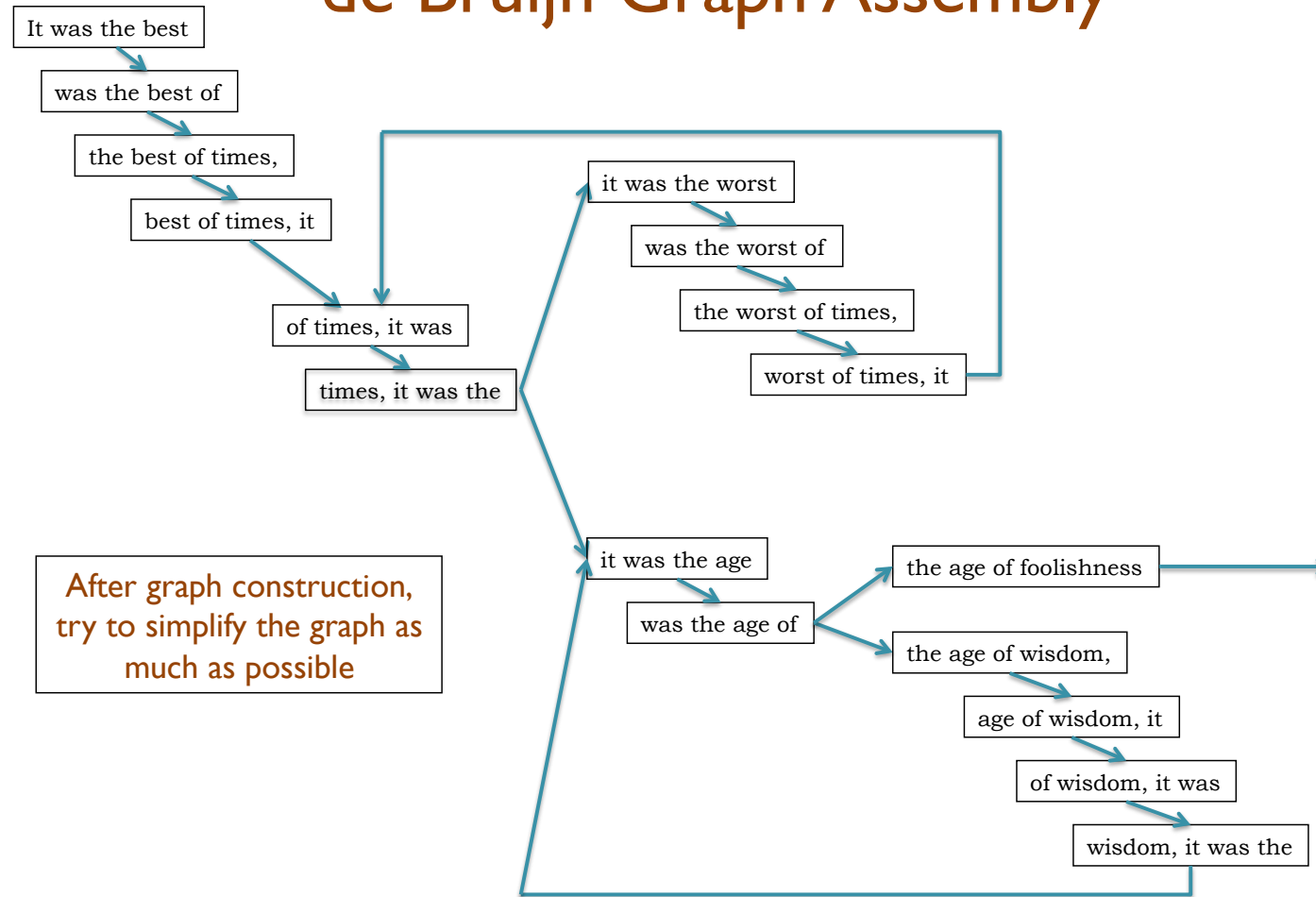
- Locally constructed graph reveals the global sequence structure
 - Overlaps between sequences implicitly computed

de Bruijn, 1946

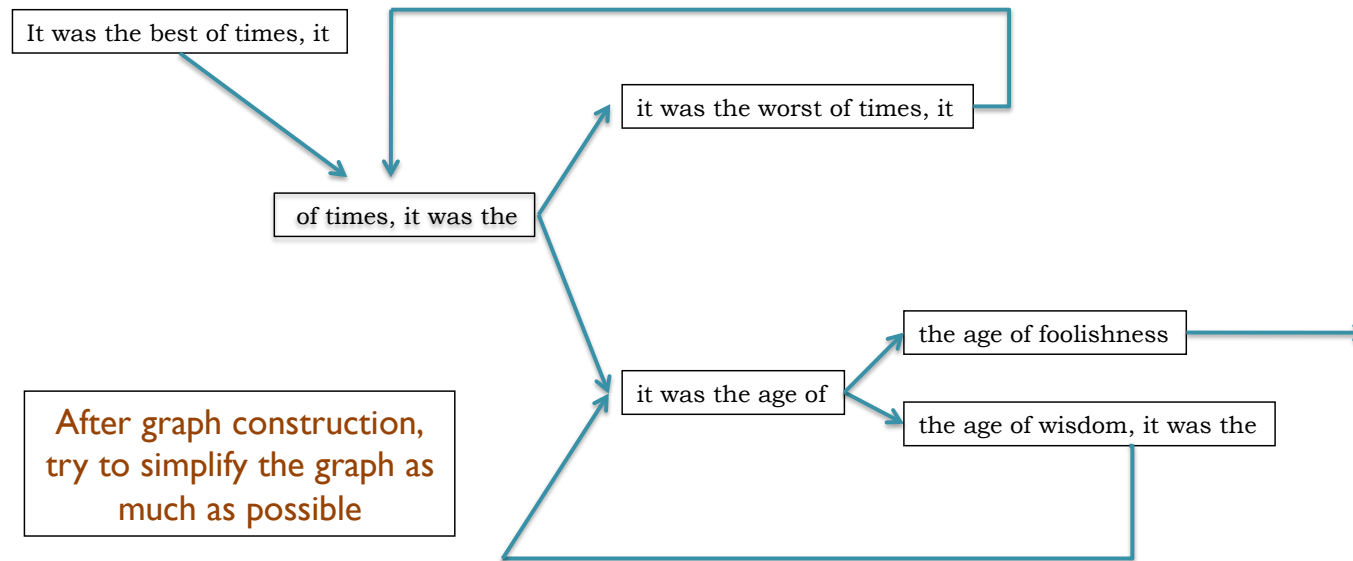
Idury and Waterman, 1995

Pevzner, Tang, Waterman, 2001

de Bruijn Graph Assembly

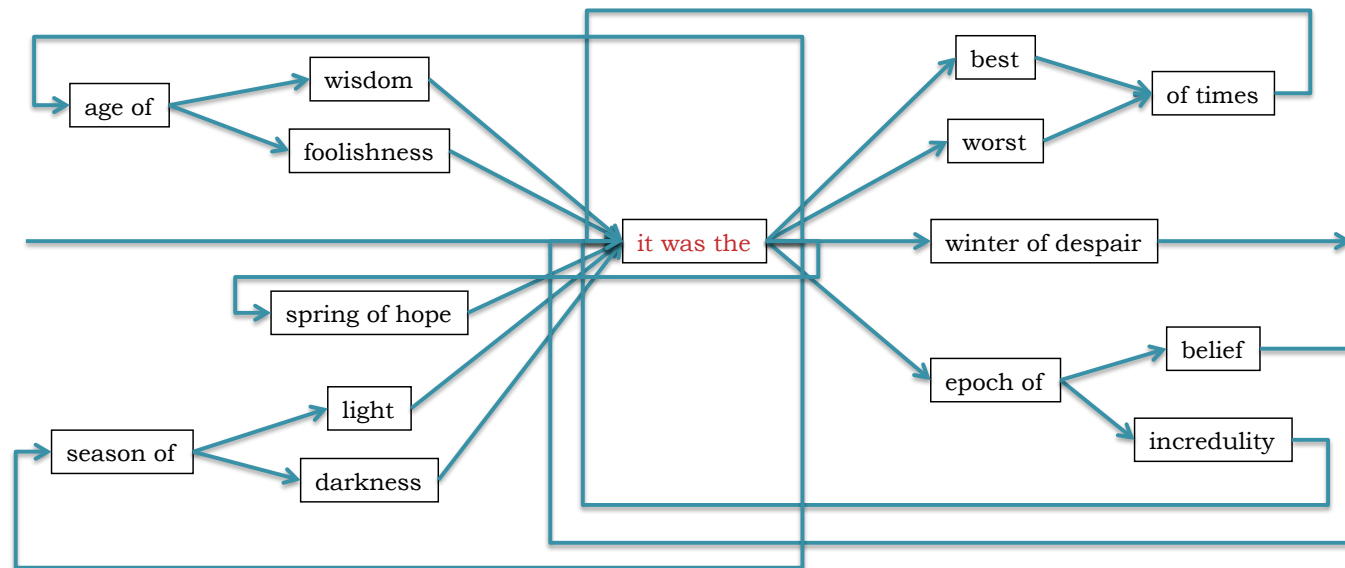


de Bruijn Graph Assembly



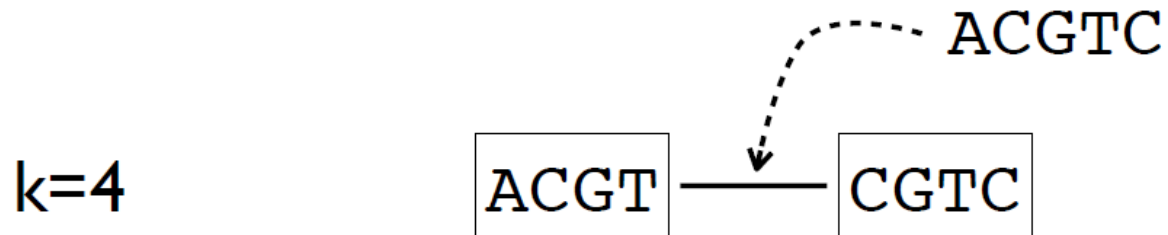
The full tale

... 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 winter of despair ...



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

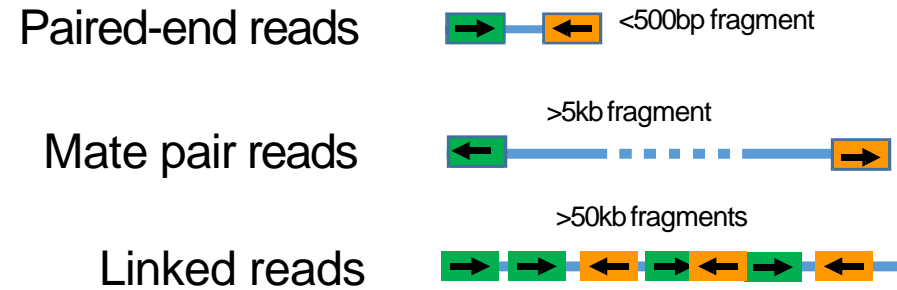


Scaffolding

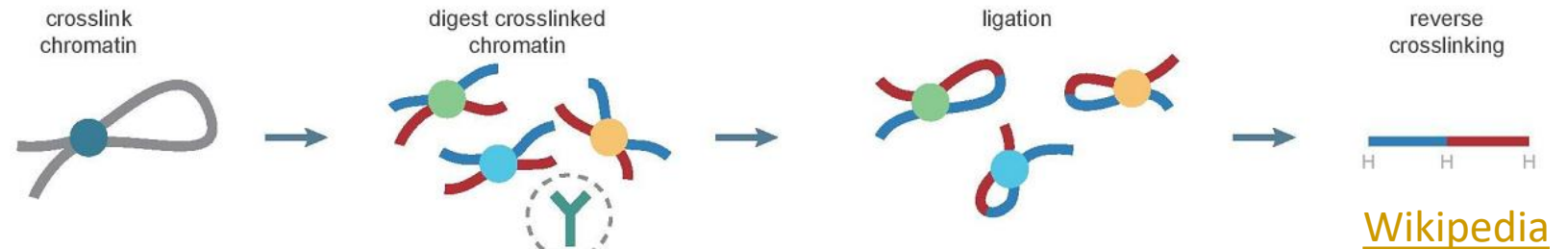
- Now, you have a huge pile of contigs but you want to make them larger. How?
- Add context!
- Link together contigs using *other* genomic information
 - Infer contigs position on the genome relative to one another

Linking Contigs via DNA Seq

Illumina sequencing



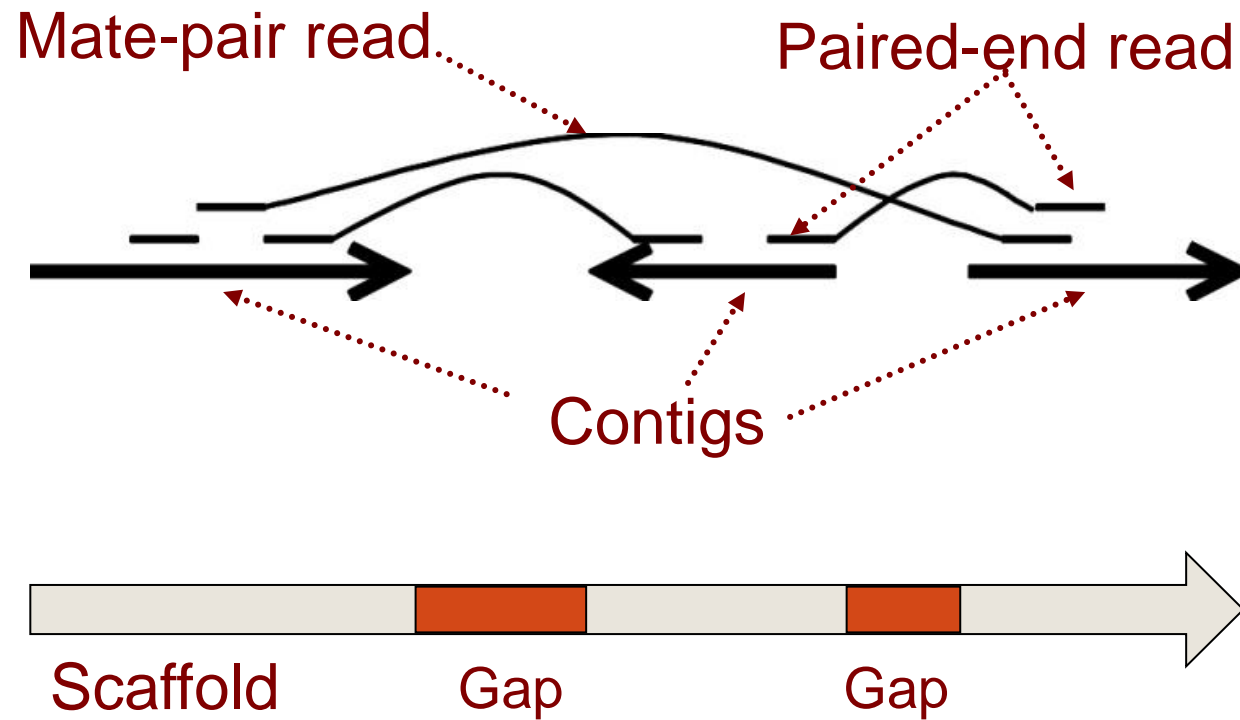
HiC (Chromosome Conformation Capture)



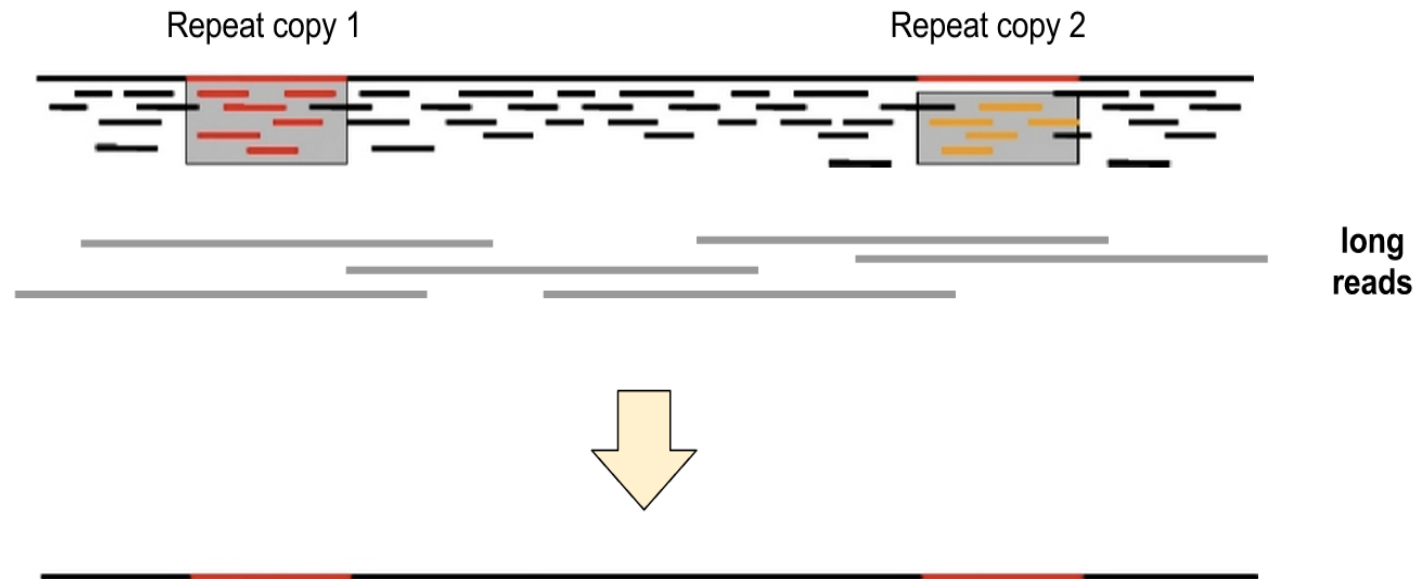
PacBio/ONT long-reads



Contigs to scaffolds



Long reads

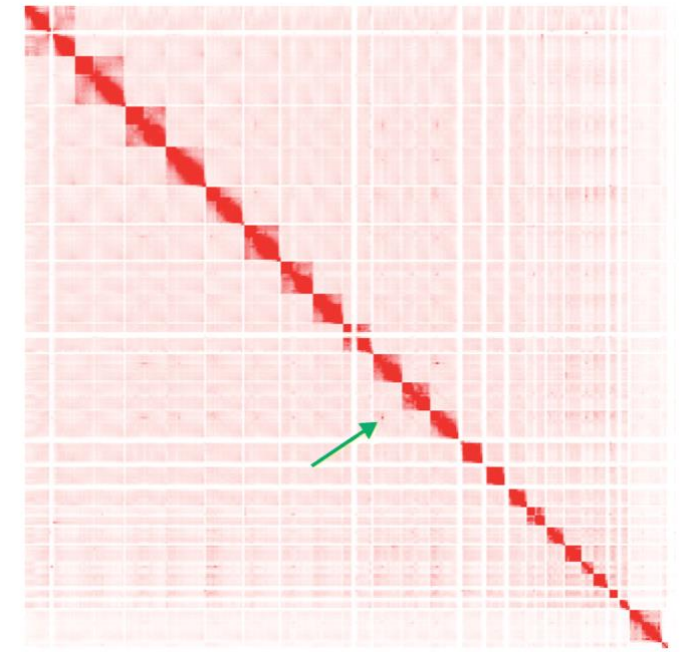
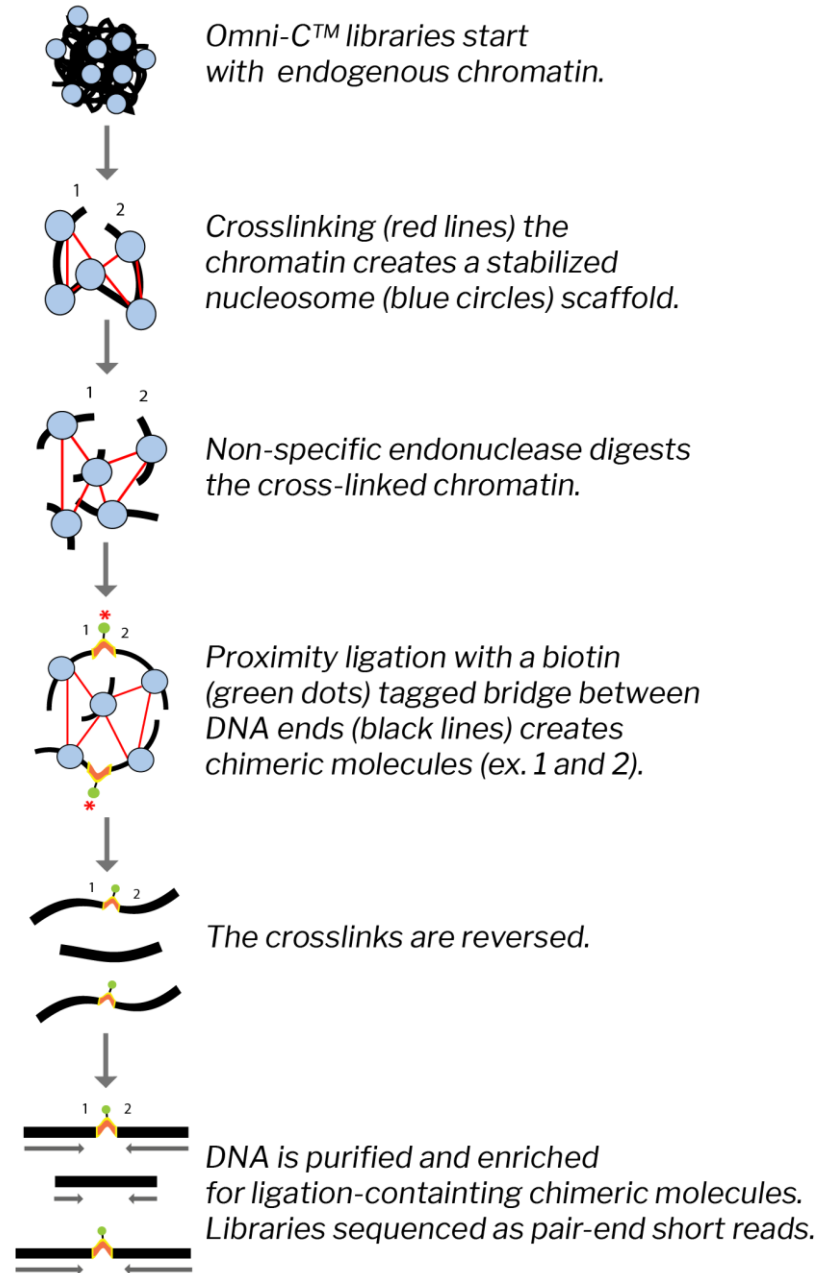


HiC

Chromosome Conformation Technology

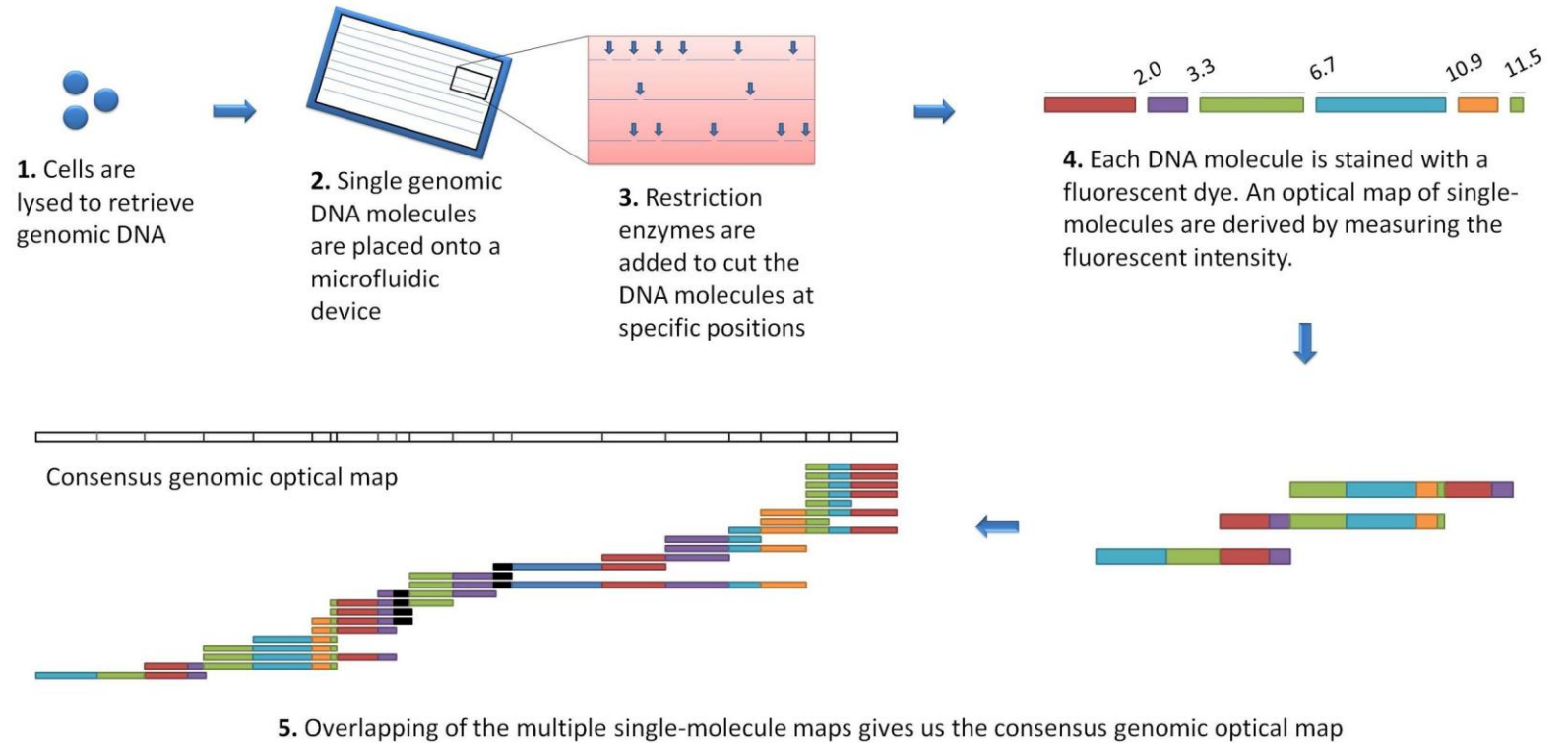
<https://dovetailgenomics.com/omni-c/>

[Wikipedia](#)



Optical Mapping

Using high resolution single-molecule restriction mapping combined with fluorescent dyes and fluorescence microscopy to produce a genomic map



Starting a new assembly project

Planning a genome sequencing project?

BUDGET!!!

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

Biology!

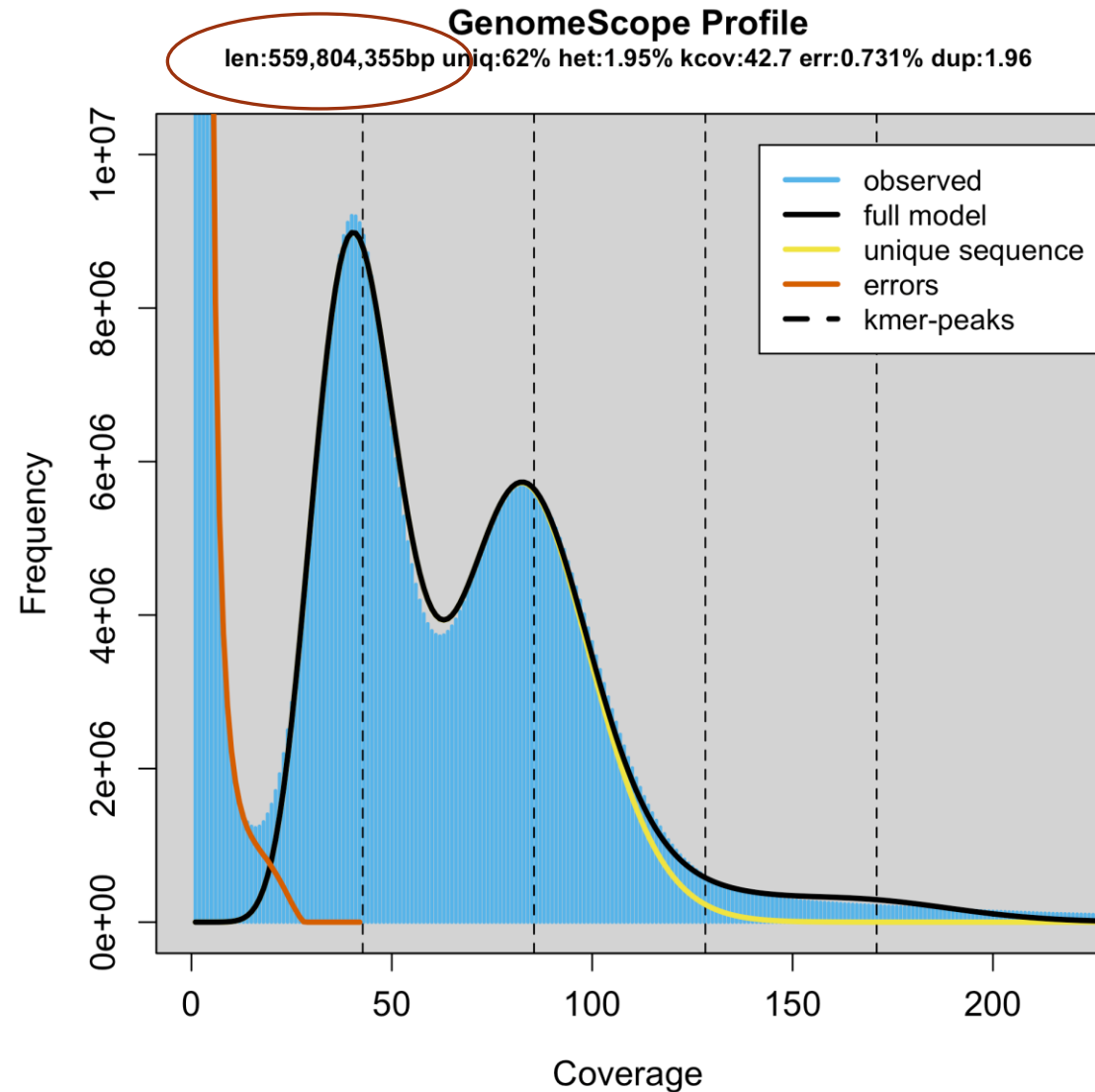
- **Size:** how large and/or complex is my genome?
- **Ploidy:** number of sets of chromosomes of the genome?
- **Multinucleated:** can cells have more than one nucleus?
- **Repetitive:** How much of the genome is repetitive? Repeat size distribution?
- **Heterozygosity:** Is my genome highly heterozygous? 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

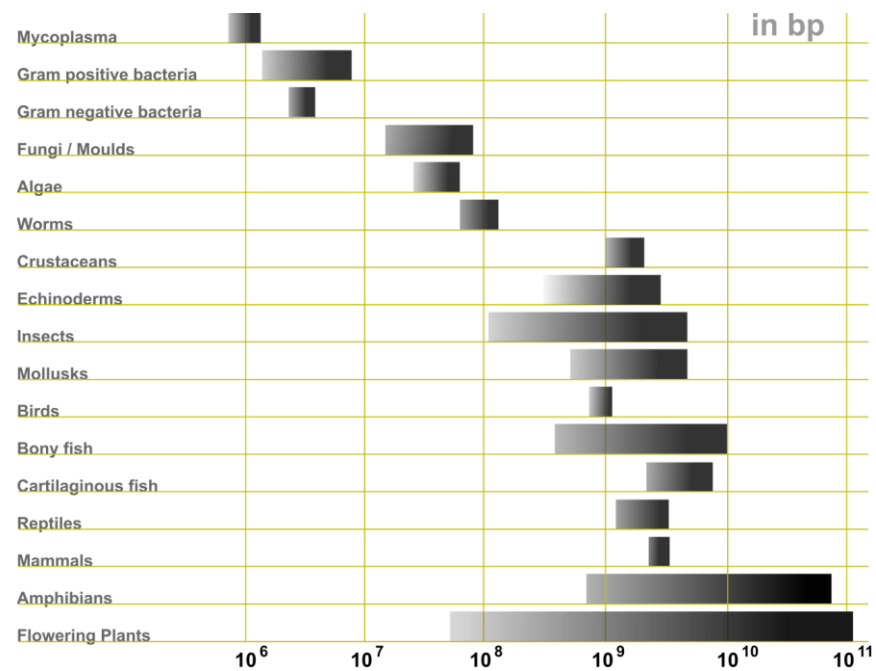
This will affect:

- 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: <http://qb.cshl.edu/genomescope/genomescope2.0/>

Genome size/complexity



By Abizar at English Wikipedia, CC BY-SA 3.0, <https://commons.wikimedia.org/w/index.php?curid=19537795>

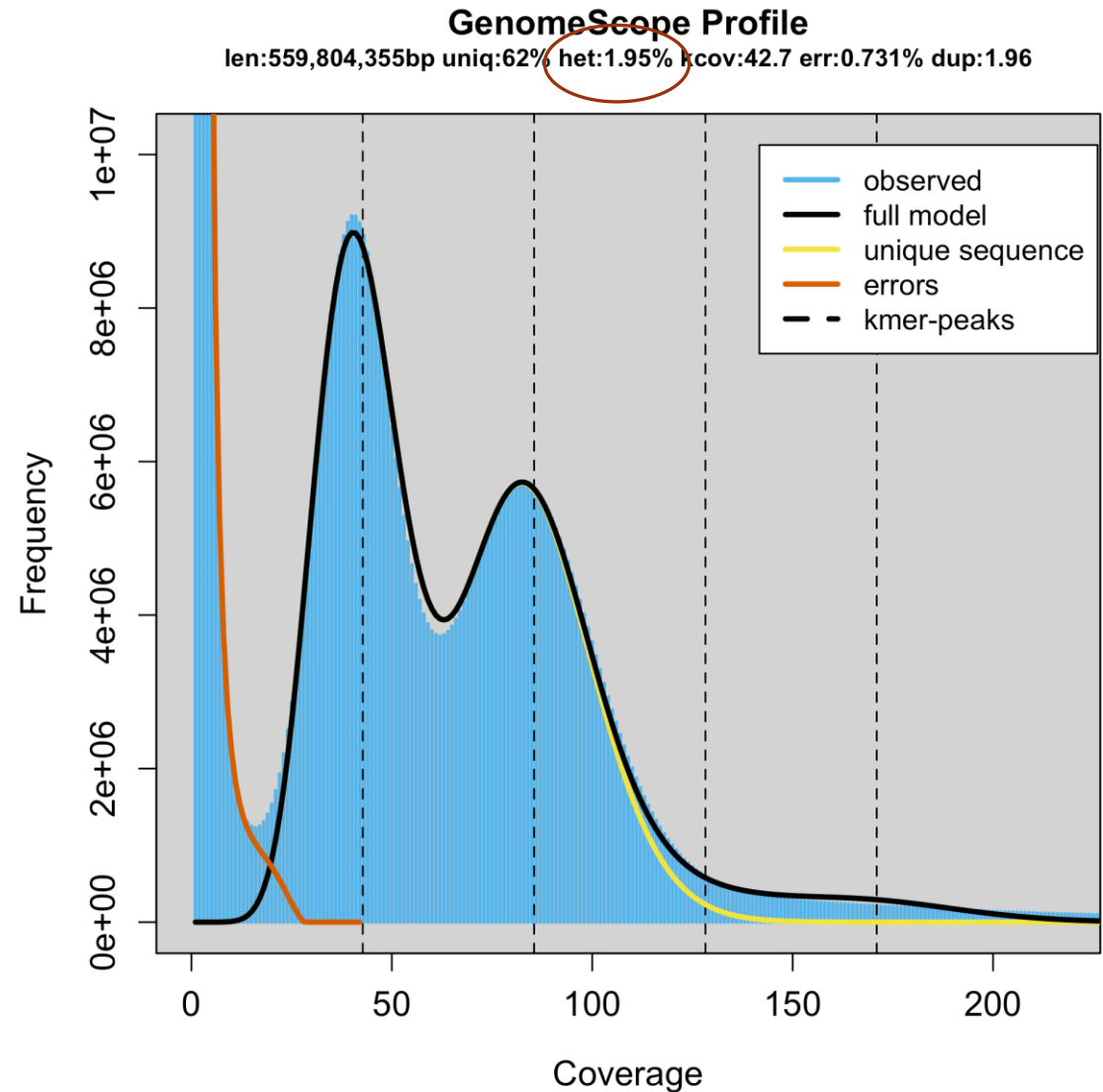
Heterozygosity

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

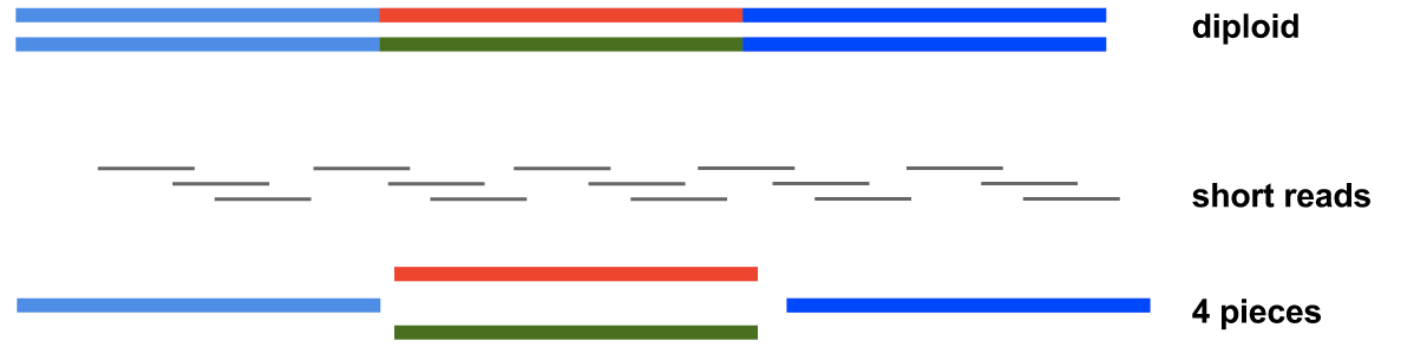
Unfortunately, assemblies are represented (for now) as haploid. So this is a major problem!



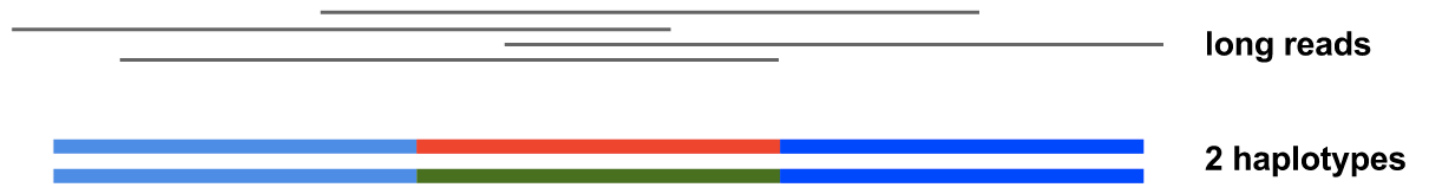
Oyster: <http://qb.cshl.edu/genomescope/genomescope2.0/>

Heterozygosity

- Short reads - initial assembly has mix of homozygous and heterozygous regions



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



Phased diploid assembly

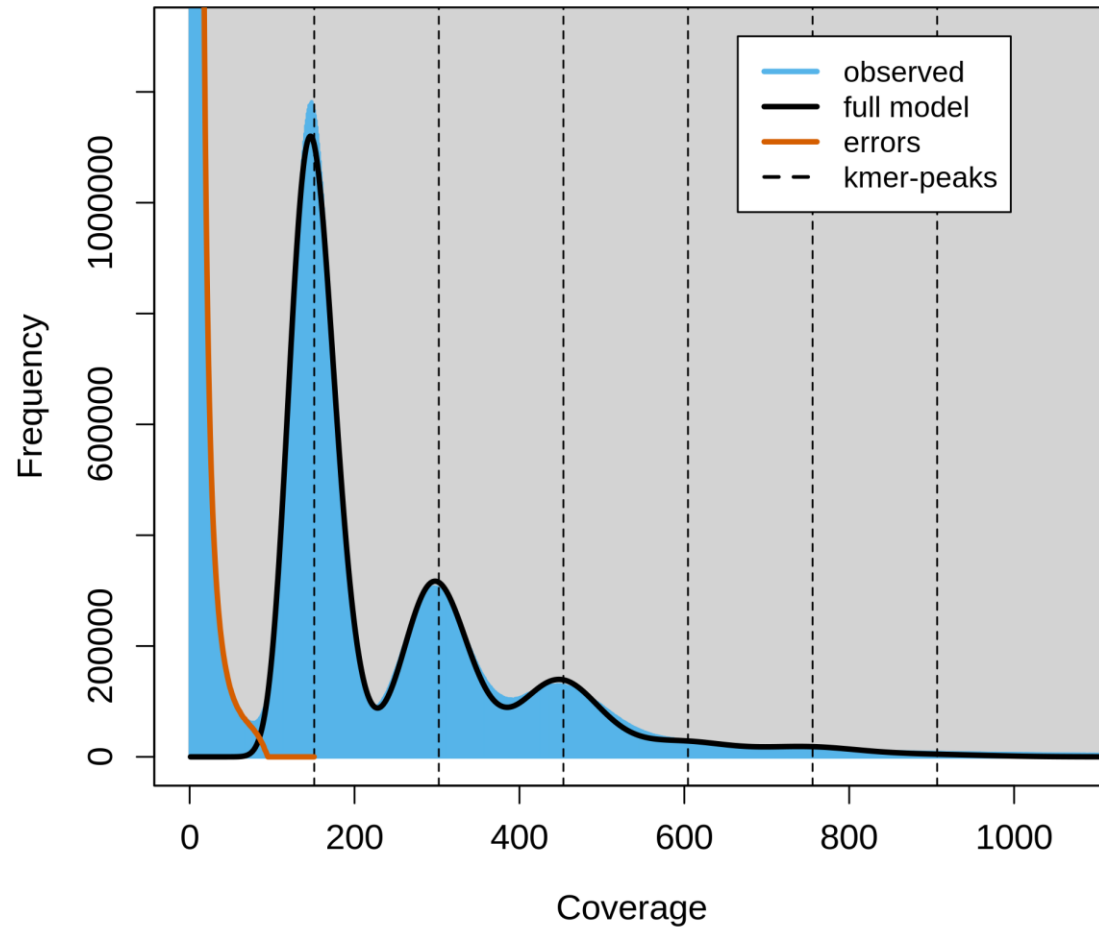
Ploidy

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)

GenomeScope Profile

len:89,522,919bp uniq:61.9%
aaa:93.9% aab:5.15% abc:0.935%
kcov:151 err:0.743% dup:4.09 k:21 p:3



Root knot nematode: <http://qb.cshl.edu/genomescope/genomescope2.0/>

Repetitive sequences

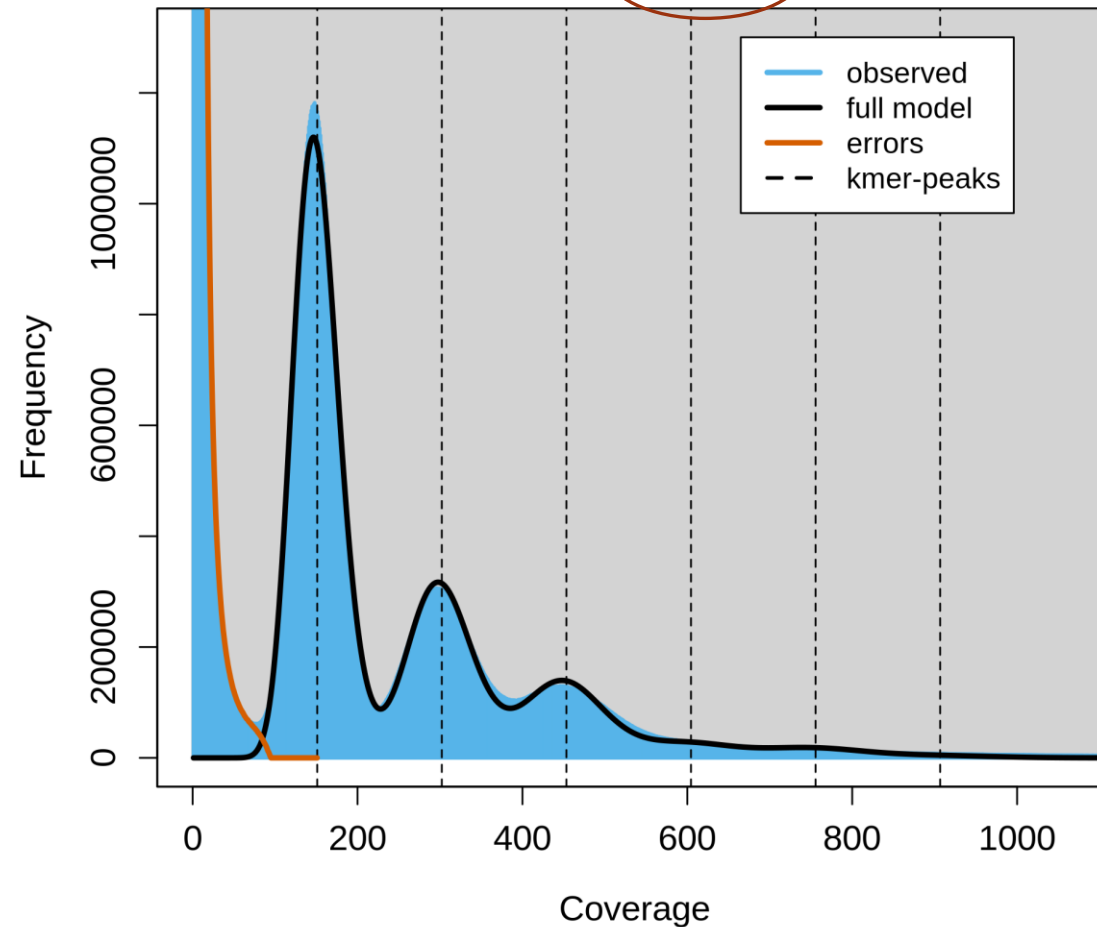
Most common source of assembly errors

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

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

GenomeScope Profile

len:89,522,919bp uniq:61.9%
aaa:93.9% aab:5.15% abc:0.935%
kcov:151 err:0.743% dup:4.09 k:21 p:3



Root knot nematode: <http://qb.cshl.edu/genomescope/genomescope2.0/>

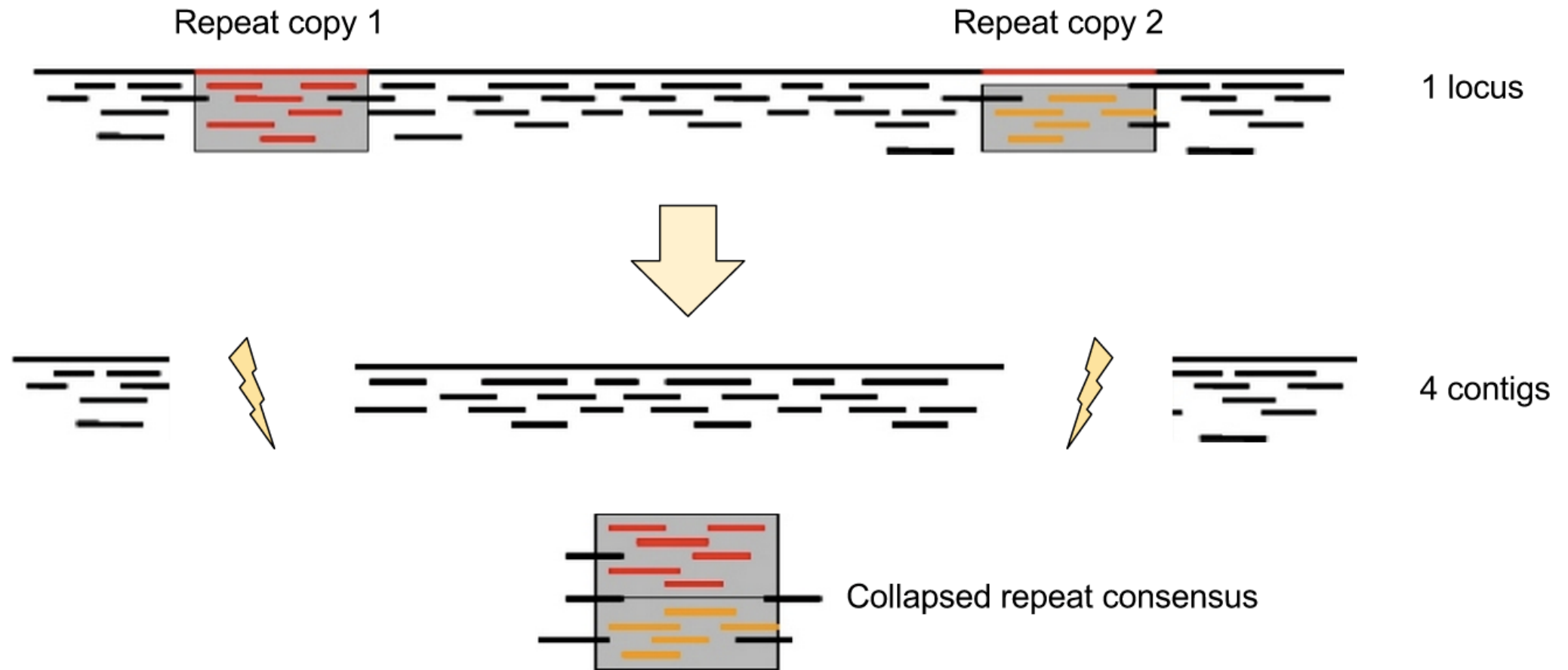
What is a repeat?

*A segment of DNA
which occurs more than once
in the genome sequence*

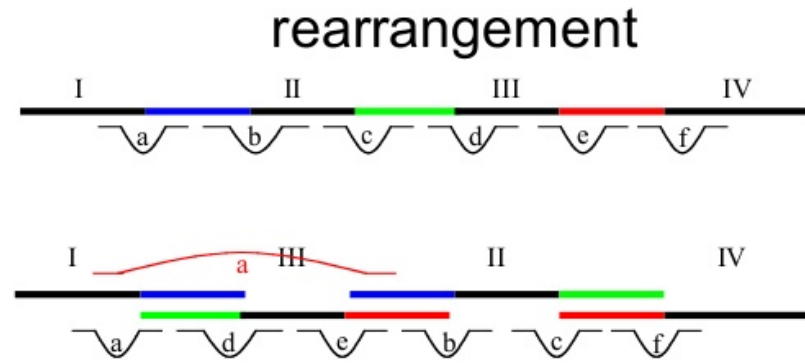
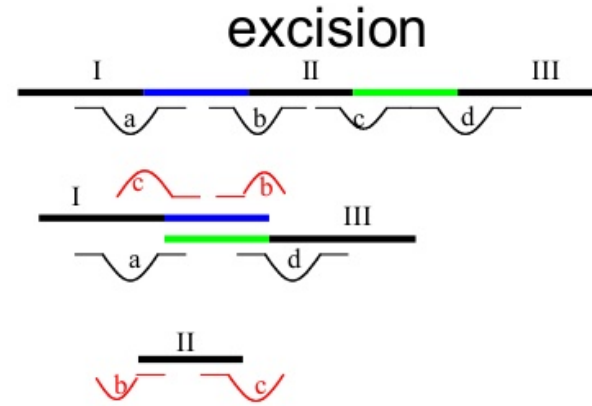
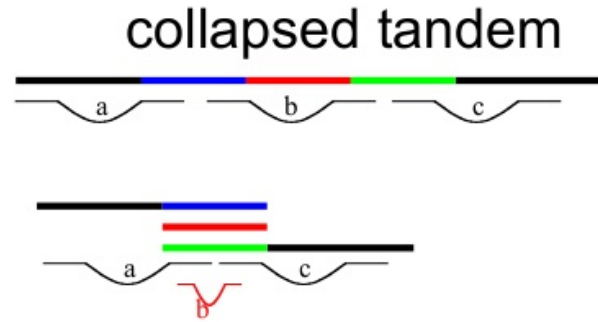


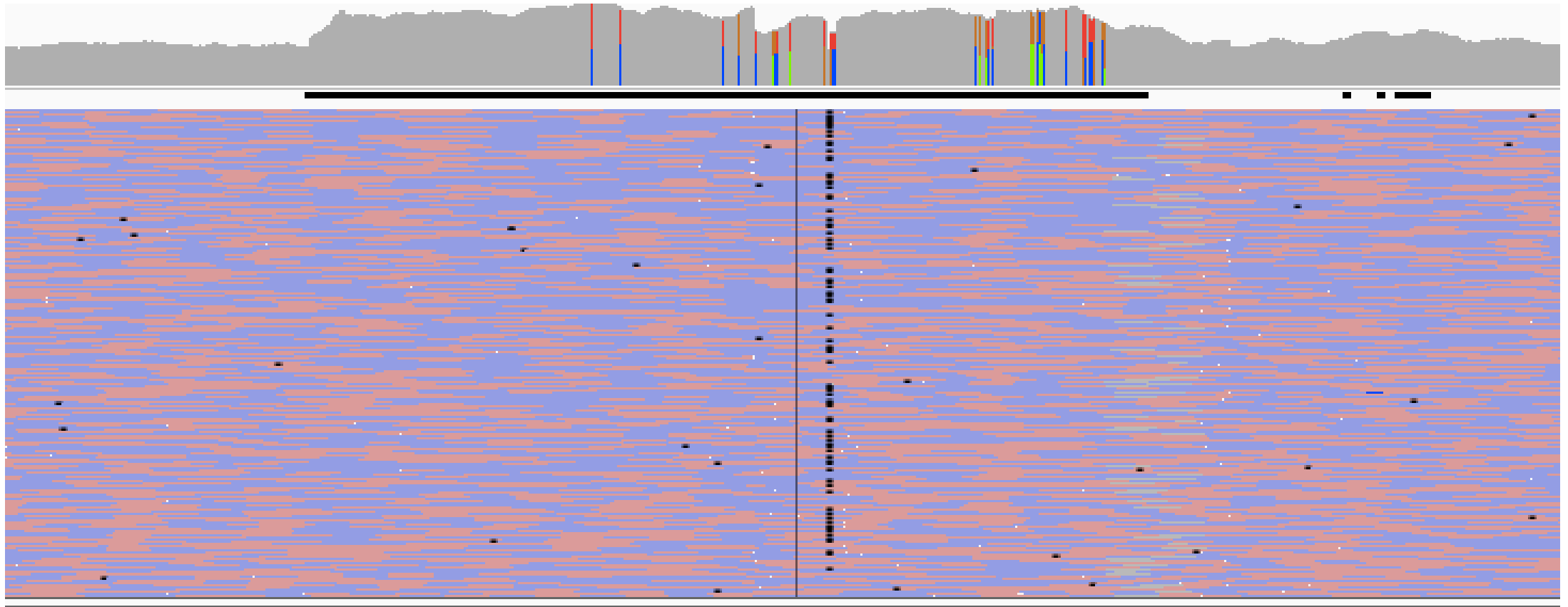
- Very common
 - Transposons (self replicating genes)
 - Satellites (repetitive adjacent patterns)
 - Gene duplications (paralogs)

Assembling repeats

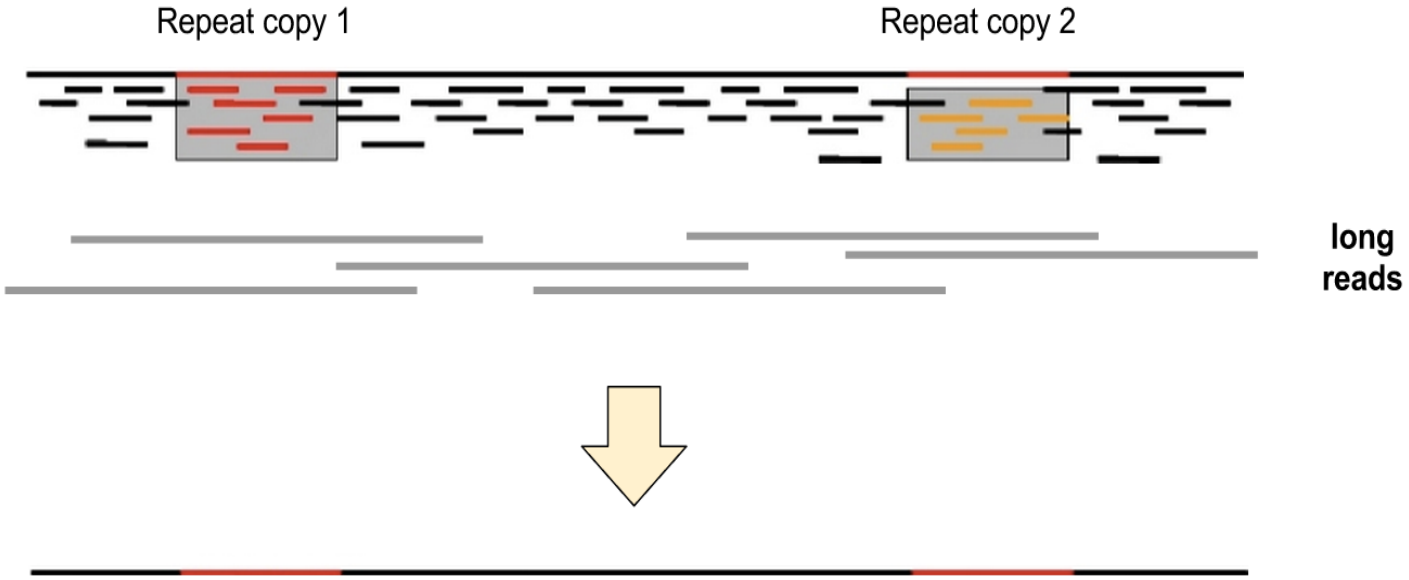


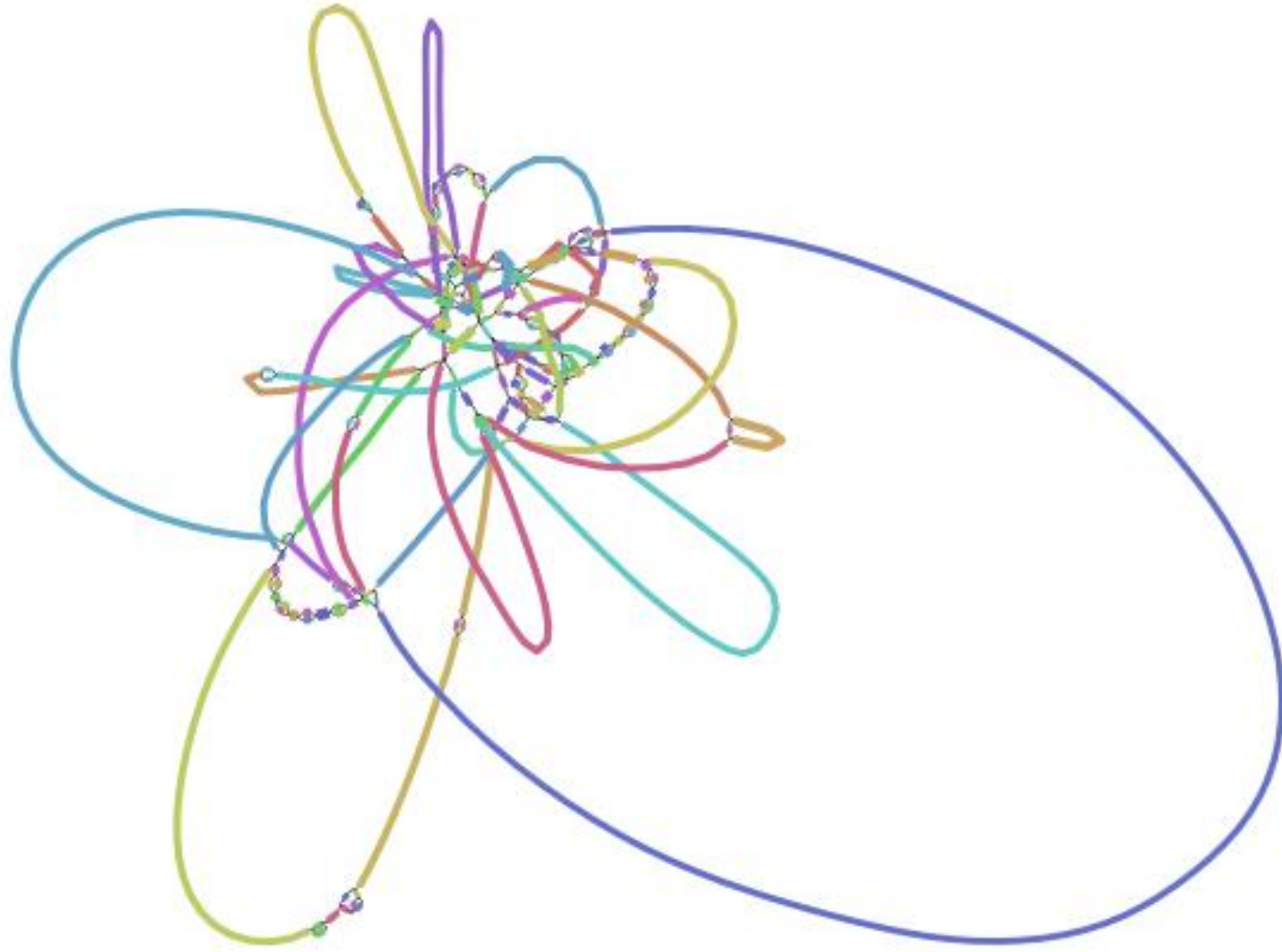
Repeat mis-assembly



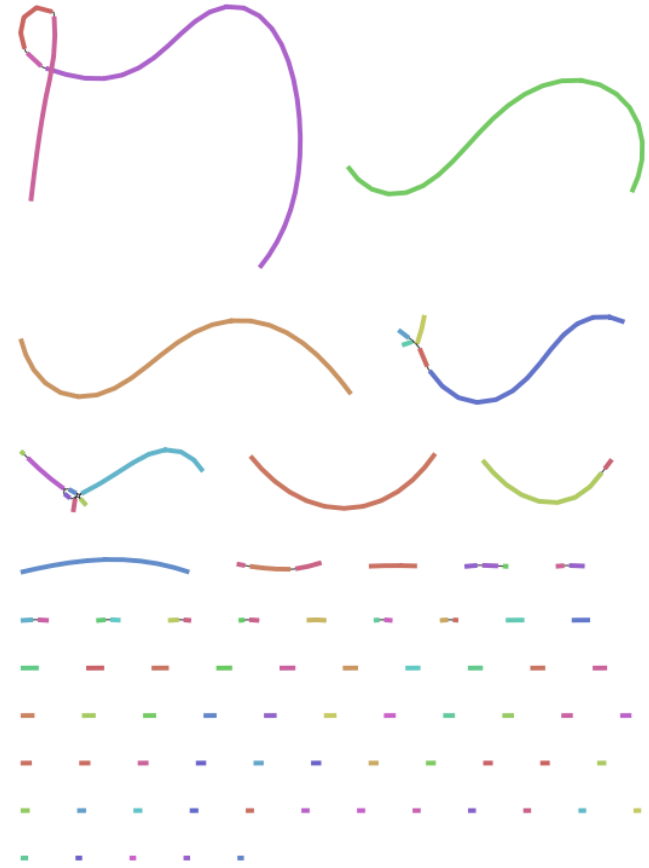
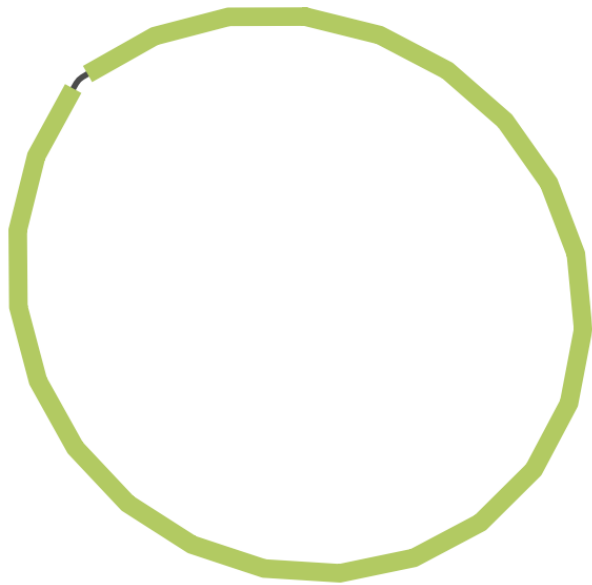


Long reads





0



Genome(s) from related species

Preferably of good quality, with large reliable scaffolds

Help verifying the completeness of the assembly

Can themselves be improved in some cases

Help guiding the assembly of the target species

- **But to be used with caution** – can cause errors when genome architecture is different!
- *Large-scale genomic rearrangement in particular is a problem*

Typical sequencing strategies

Small genomes (bacteria, fungal)

- If you can can get HMW DNA!
 - **PacBio HiFi**
 - **Oxford Nanopore** sequences at 40-50x coverage, 'polish' with hybrid correction (using Illumina data) and assembly using Unicycler, Canu, Flye
 - This may be changing with newer flow cells (R10.4.1 + 'kit14', as of May 2022)
- 2 x 300bp overlapping paired-end reads from Illumina MiSeq works okay but will get fragments

Larger genomes

- If you can afford it and can get HMW DNA
 - **PacBio HiFi**
 - **HiC** for scaffolding

T2T strategy

- Human assemblies
- HMW DNA preps
- 50x PacBio HiFi reads or higher
- 15-30x Oxford ultralong reads (>100kb)
- This is also in flux!
- \$\$\$\$\$\$\$\$\$

Science,
March 2022

Time, May 2022

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

🔒 | SPECIAL ISSUE RESEARCH ARTICLE | HUMAN GENOMICS

f t in r s e

The complete sequence of a human genome


SERGEY NURK , SERGEY KOREN , ARANG RHIE , MIKKO RAUTIAINEN , ANDREY V. BZIKADZE , ALLA MIKHEENKO, MITCHELL R. VOLLGER , NICOLAS ALTE-MOSE , 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

TIME SUBSCRIBE

← THE 100 MOST INFLUENTIAL PEOPLE OF 2022

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



Assembly strategies and algorithms

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

- **Examples: hifiasm (PacBio HiFi only), Canu, Redbean, Flye, Shasta**
- **Hifiasm is generally recommended for PacBio HiFi data**

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?

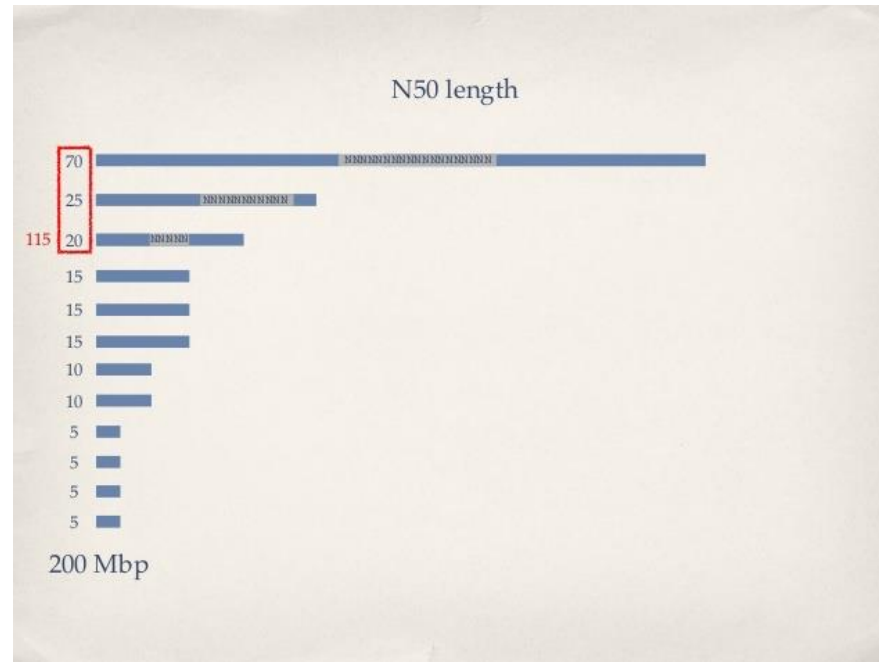
Are all the genes I expected in the assembly?

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



N50 concerns



Optimizing for N50

- Encourages mis-assemblies!
- Encourages 'gaming' the stats

An aggressive assembler may over-join:

- 1,1,3,5,**8,12**,20 (previous)
- 1,1,3,5,**20**,20 (now)
- $1+1+3+5+20+20 = 50$ (unchanged)

N50 is the "halfway sum" (still 25)

- $1+1+3+5+20 = 30 (\geq 25)$ so **N50 is 20** (was 12)

You can also filter contigs below a certain (arbitrary) size, which lowers overall assembly size (and increases N50)

Comparative analysis

Compare against

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

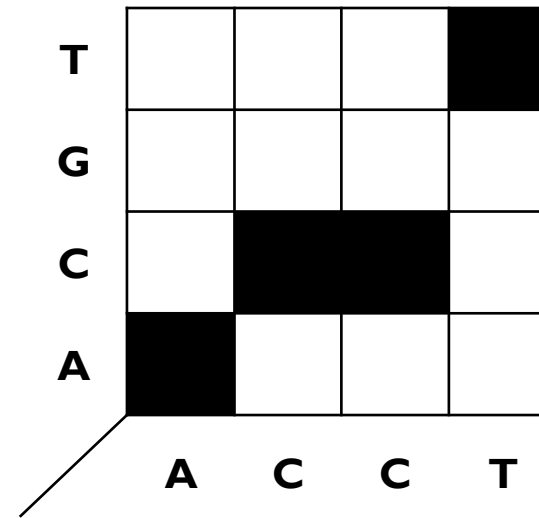
Whole genome alignment

- *MUMmer*
- *Lastz*

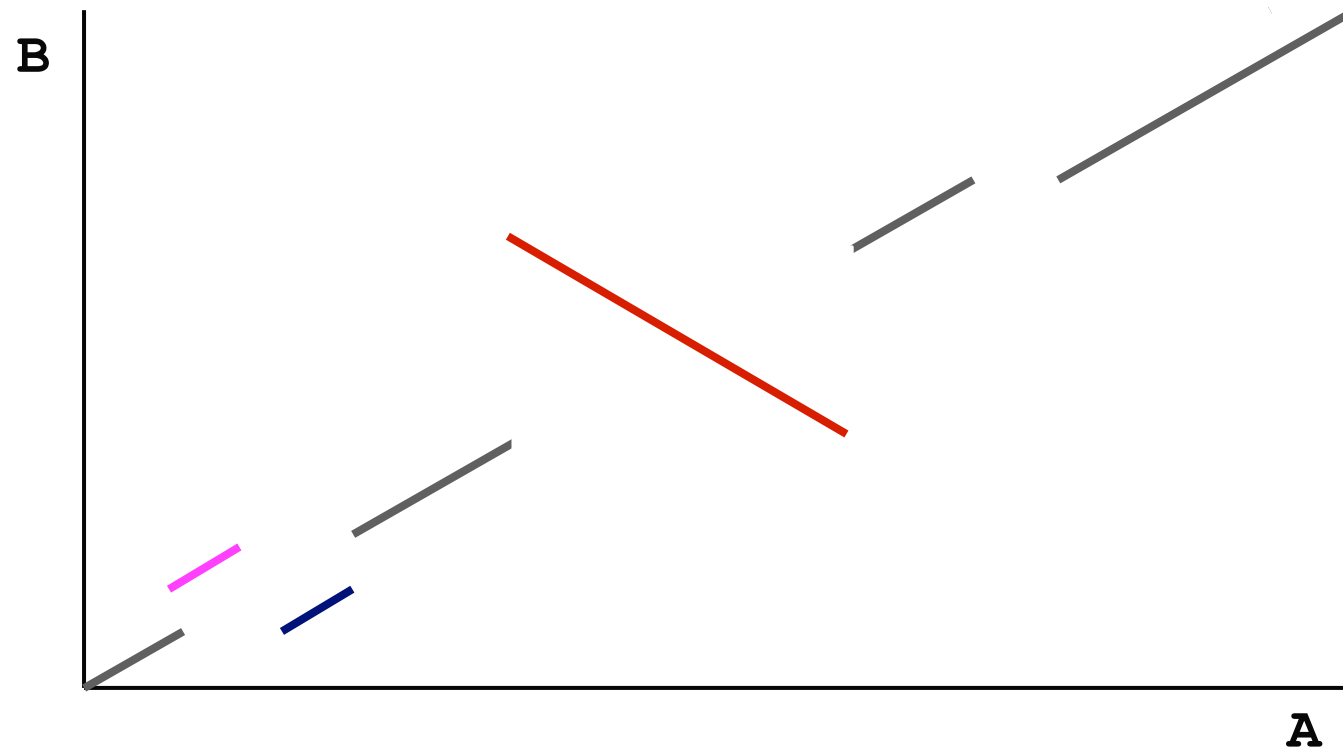
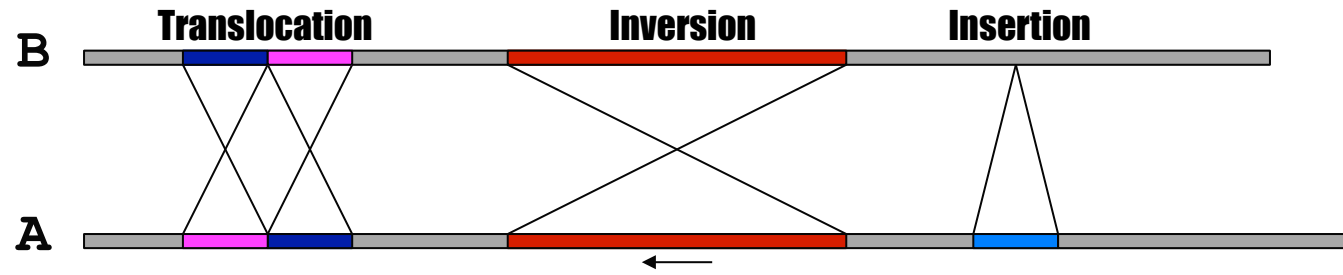
Generates an alignment and a *dot plot*

Dot Plot

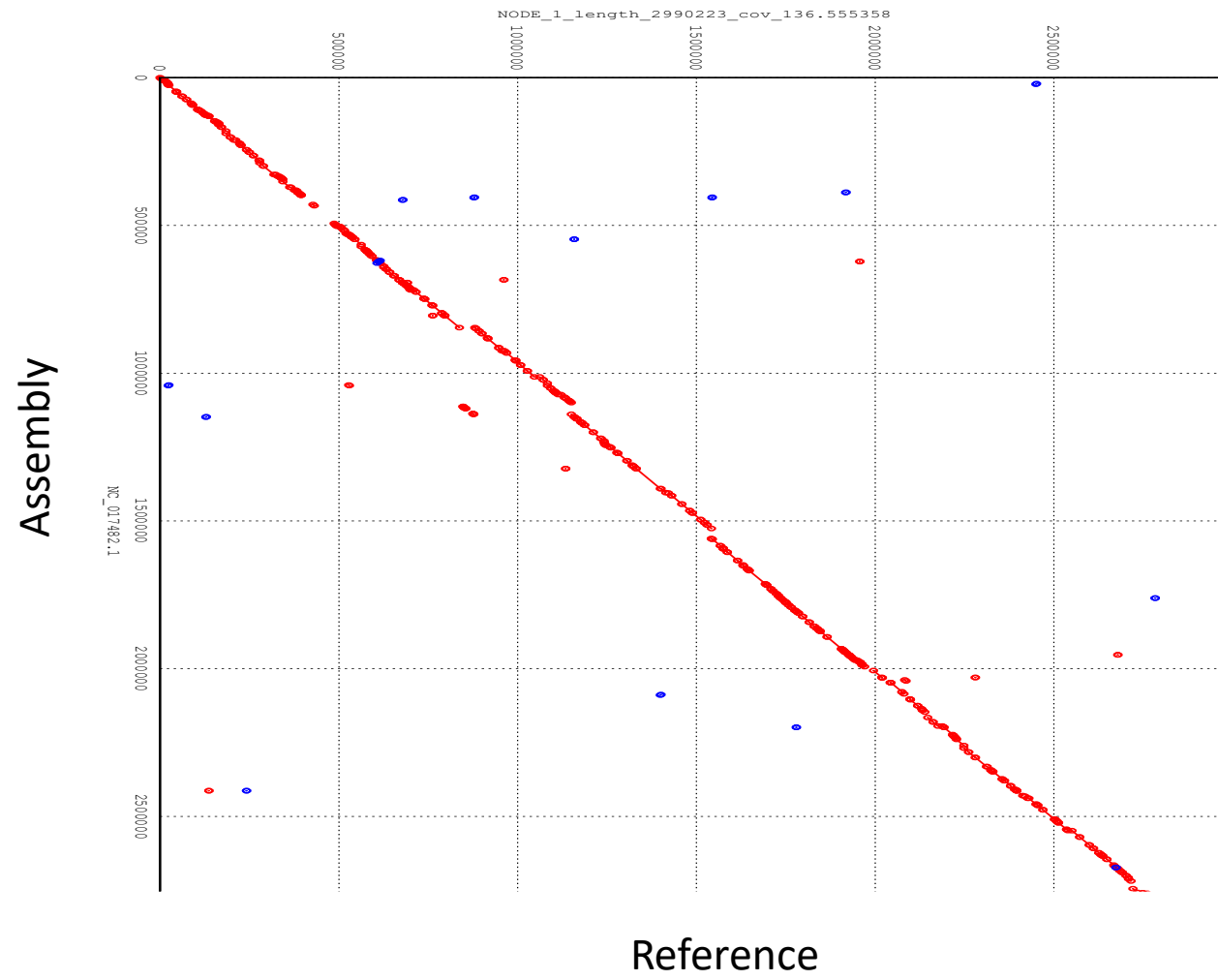
- How can we visualize *whole* genome alignments?
- With an alignment dot plot
 - $N \times M$ matrix
 - Let i = position in genome A
 - Let j = position in genome B
 - Fill cell (i,j) if A_i shows similarity to B_j



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



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

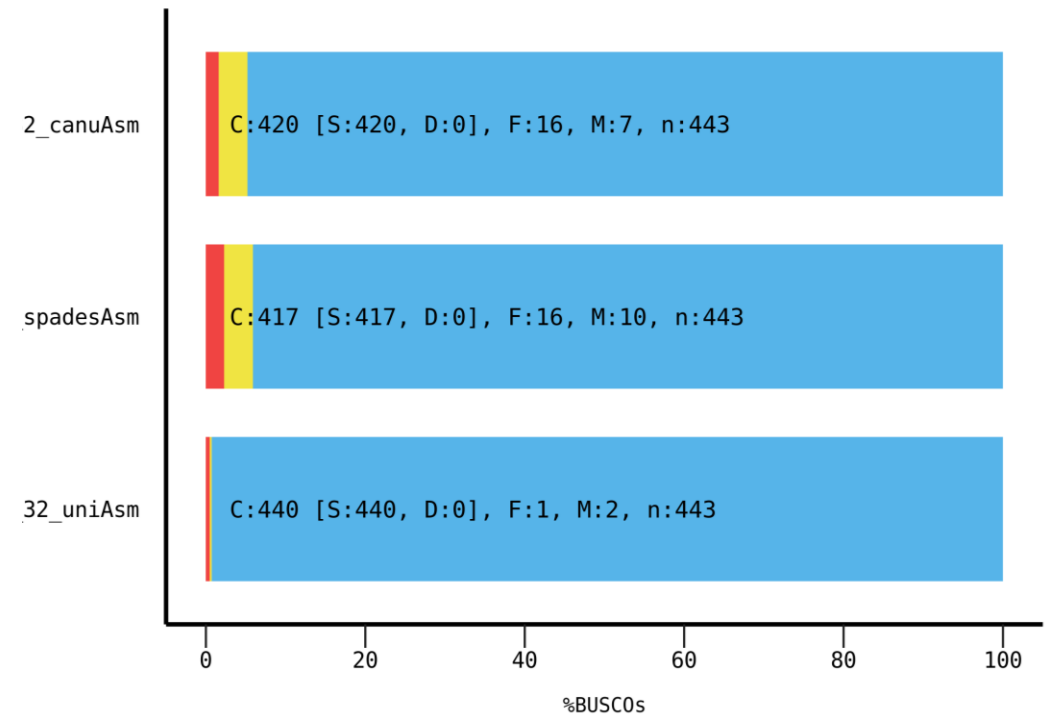


BUSCO: conserved gene sets

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

Coverage is indicative of quality
and completeness of assembly

BUSCO Assessment Results



QUAST

QUality ASsessment Tool

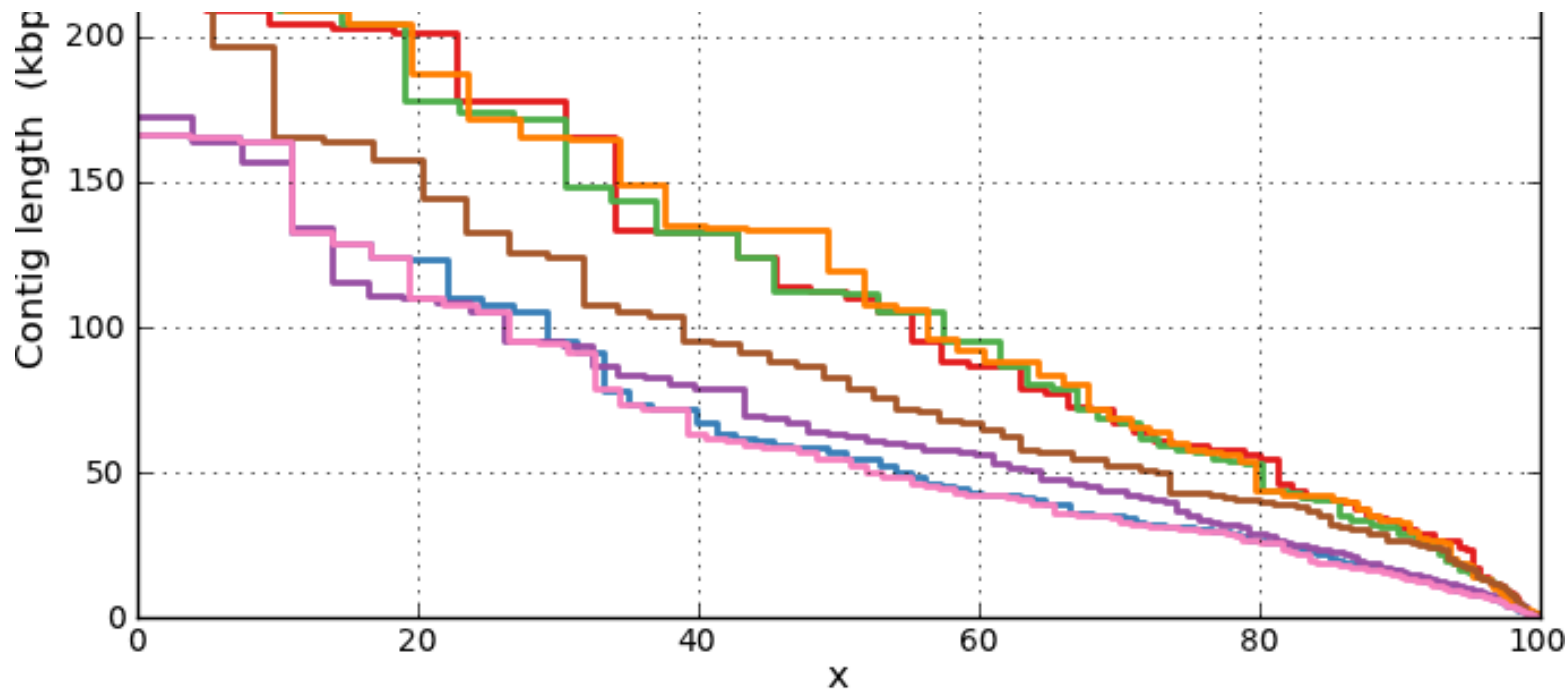
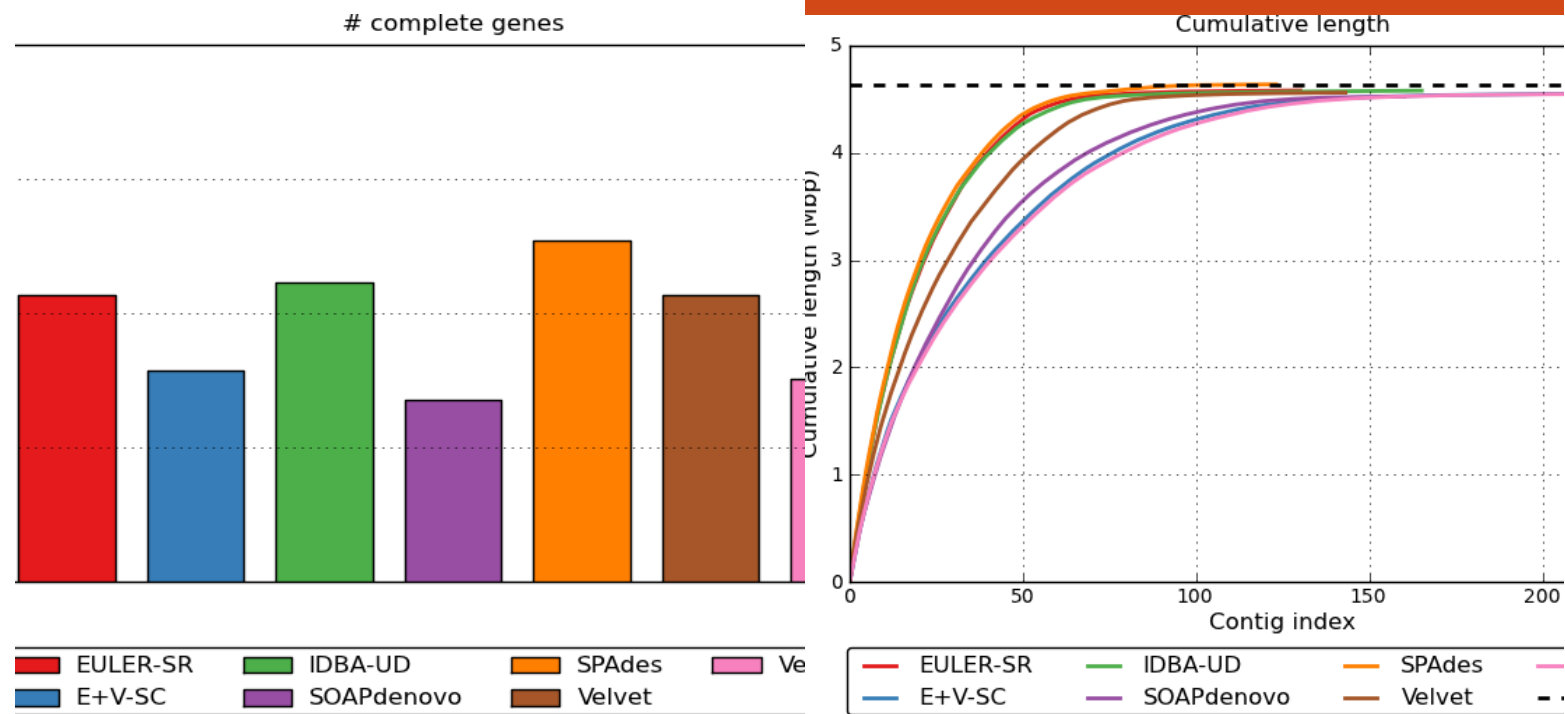
- Small (bacterial, fungal) and large (eukaryotic) genomes
- Metagenomes
- Icarus for contig alignment visualization

Can compare multiple assemblies against one another

Compare against a known (or close) reference

Optional: Predict genes or include annotations (checks for odd issues like frameshifts)

Generates a summary HTML report



Even the best genomes are not perfect

nature

Explore content ▾ Journal information ▾ Publish with us ▾ | [Subscribe](#)

nature > news > article

NEWS | 04 June 2021

A complete human genome sequence is close: how scientists filled in the gaps

Researchers added 200 million DNA base pairs and 115 protein-coding genes – but they've yet to entirely sequence the Y chromosome.

Genome graphs

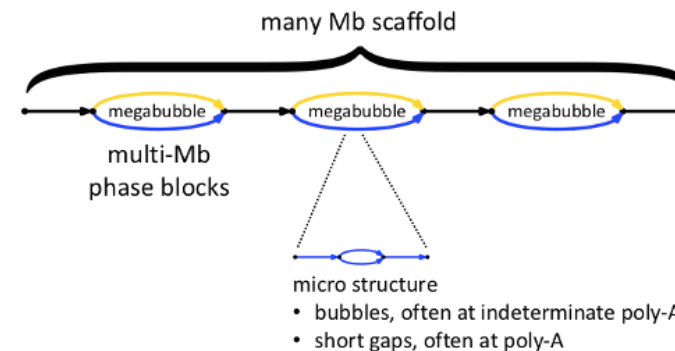
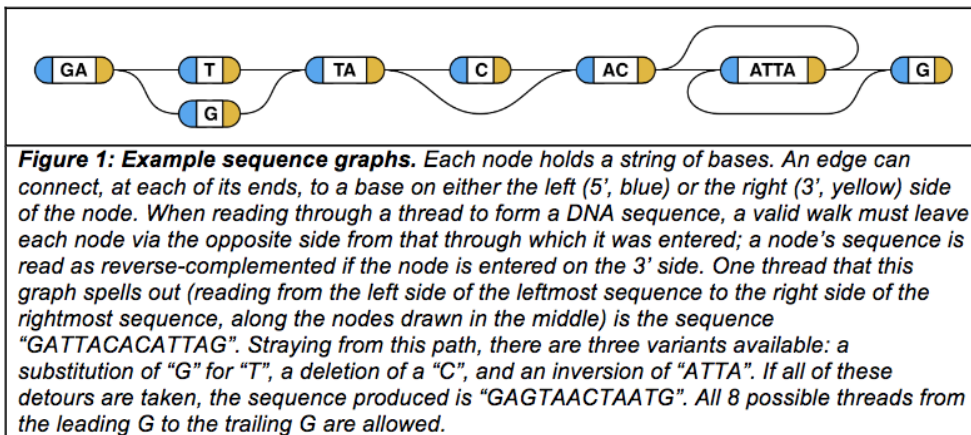
Genome graphs

With the release of the latest human genome reference, there is more pressure to represent more data with a genome.

Current representations are mainly **haploid** (one copy)

Newer representations are **genome graphs**, where variant information is retained (e.g. heterozygosity)

Tools are still catching up, but many new assemblers (e.g. hifiasm) generate a *diploid* assembly now



Genome graphs

One interesting application of graphs: *trio assembly*

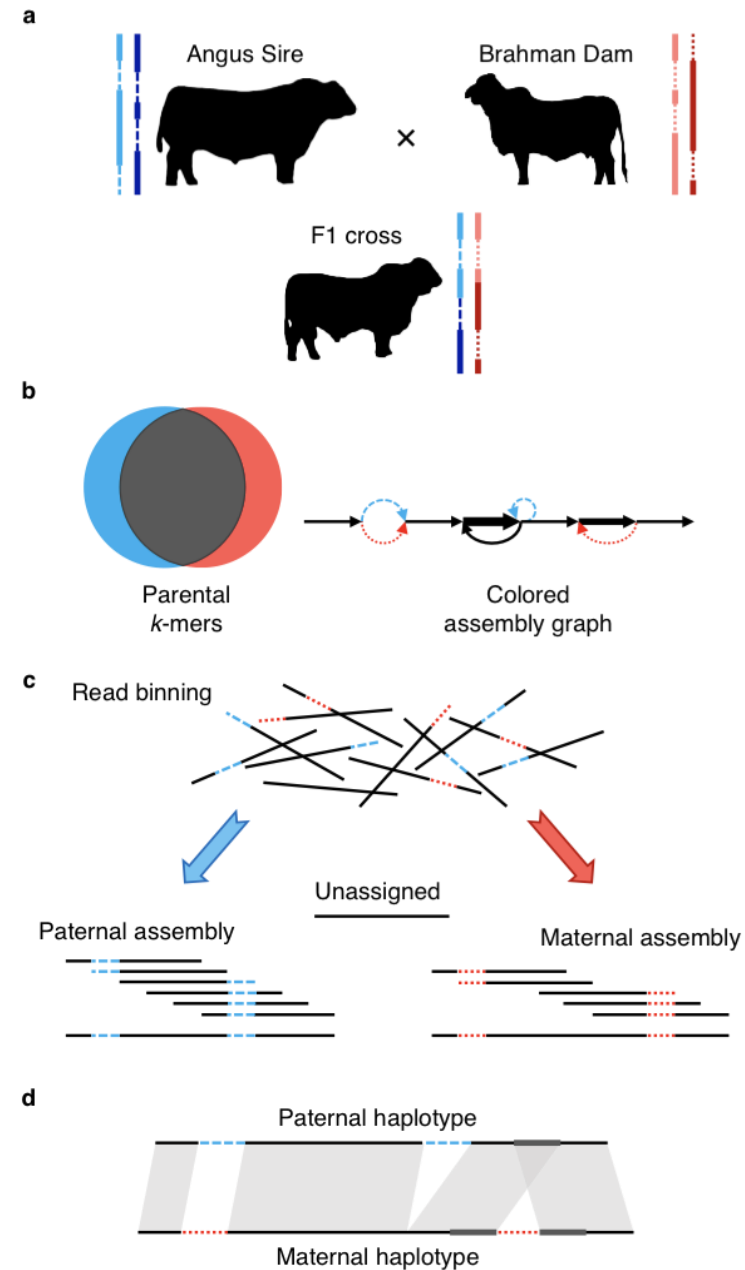
Short- and long-read sequencing

Advantage:

Better assembly

Phased variants

Structural variants!



Genome Annotation

Methods for genome annotation

Ab initio

- i.e. based on sequence alone
- INFERNAL/rFAM (RNA genes), miRBase (miRNAs), RepeatMasker (repeat families), many gene prediction algorithms (e.g. AUGUSTUS, Glimmer, GeneMark, ...)

Evidence-based

- Transcriptome data for the target organism (the more the better)
- Proteins of interest
- Align trx/protein sequences to assembly, generate gene models

Combined approaches

- ***Most common***

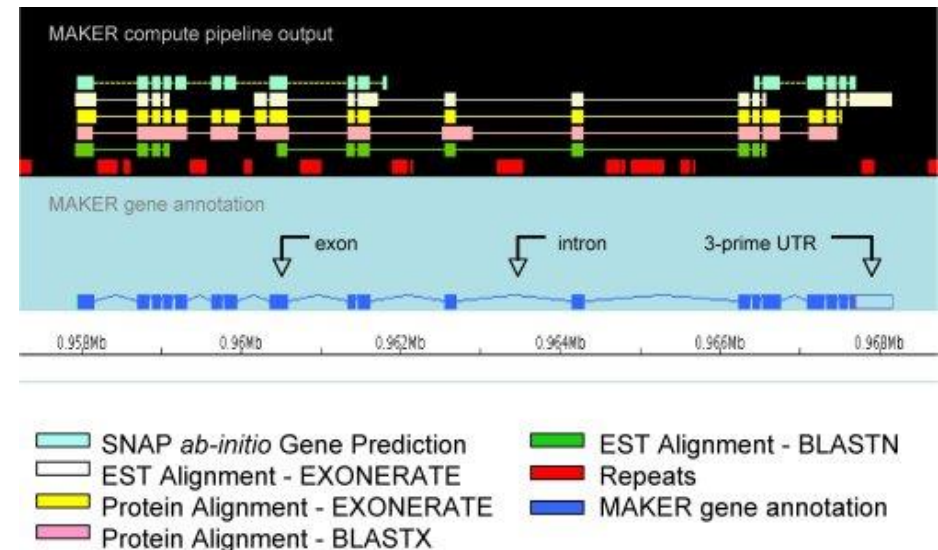
General steps for biological annotation

1. Predict gene models using *ab initio*-based tools
 - May require considerable tuning and a bootstrapping step
2. Using closely-related protein/transcripts, BLAST against assembly to find locations
3. Find potential splice junctions of BLAST hits
4. Combine all evidence and make consensus gene (and possibly transcript/isoform) predictions, with annotation metrics for confidence of matches
 - Can include UTR regions if RNA-Seq is included
5. Assess completeness of annotation (run BUSCO, but on proteins/transcripts)
6. Run InterProScan of predicted proteins against databases of protein domains (Pfam, Prosite, HAMAP, PANTHER, ...)

MAKER, integration framework for genome annotation

MAKER runs many software tools on the assembled genome and collates the outputs

See <http://gmod.org/wiki/MAKER>



Example Pipelines

Bacterial

- **Prokka (bacterial/archaeal/viral)**
- **NCBI Prokaryotic Genome Annotation Pipeline (PGAP)**
- **Joint Genome Institute – IMG/ER (Integrated Microbial Genome Expert Review)**
 - Online only

Eukaryotic

- **NCBI – RefSeq pipeline**
 - Have to submit to NCBI (and make public) to use
 - Requires RNA-Seq
- MAKER
- Braker2

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!