HYBRID ASSEMBLY

Welcome!

Today

Assemble bacterial genome using two methods

Large genome strategy (eukaryotic organisms)
Small genome strategy (bacterial / protist / fungal organisms)

By the end of the session

Understand role of long and short reads
Aware of challenges
Troubleshoot your own assembly

First 20 mins

Long read technology

Nanopore, PacBio, use cases

Genome assembly

Genome, exome, transcriptome, mitochondria, metagenomic assembly
Mammals, plants, protists, bacteria

Hybrid genome assembly

Repeats
Base-level accuracy & structural accuracy

Technology Comparison

The relationship between NGS and TGS

Technology comparison

	Nanopore	PacBio	Illumina
Read length	1+ Mbp	25 Kbp	100 - 250 bp
Per-base accuracy	90 - 95%	88 - 99% (HiFi)	> 99.9%
Cheapest hardware	$1k - 200k	$500k	$25k - 800k
Cost per bp	$50 - 100	$50 - 100	$25 - 50

Technology comparison

Best at resolving structural information

Great for amplicon / transcript isoform identification

Best for sequence variant detection

	Nanopore	PacBio	Illumina
Read length	1+ Mbp	25 Kbp	100 - 250 bp
Per-base accuracy	90 - 95%	88 - 99% (HiFi)	> 99.9%
Cheapest hardware	$1k - 200k	$500k	$25k - 800k
Cost per bp	$50 - 100	$50 - 100	$25 - 50

Accessible to small organisations

Technology comparison

Short read niche

- Variant detection
  (small INDEL, SNP)
- Transcriptomics
  (RNA, scRNA seq)

Long read niche

- Structural variant detection
- Transcript isoform detection
- Pathogen detection
- Epigenetics

Technology comparison

Short read niche

- Variant detection
  (small INDEL, SNP)
- Transcriptomics
  (RNA, scRNA seq)

Base-level information

Long read niche

- Structural variant detection
- Transcript isoform detection
- Pathogen detection
- Epigenetics

Structural information

Nanopore

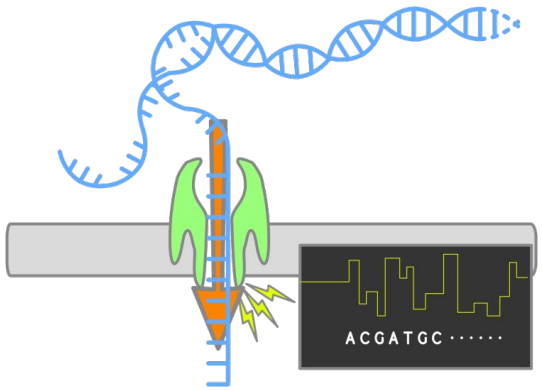
Low cost
$1000 USD for minion + 2 flow cells + reagents

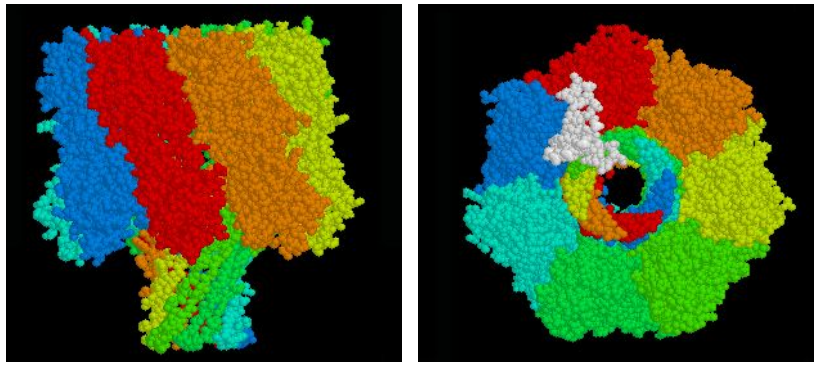
Portable
500g

Simple
library preparation < 15 mins (no PCR required)

The MinION sequencer. Image credit: Oxford Nanopore Technologies

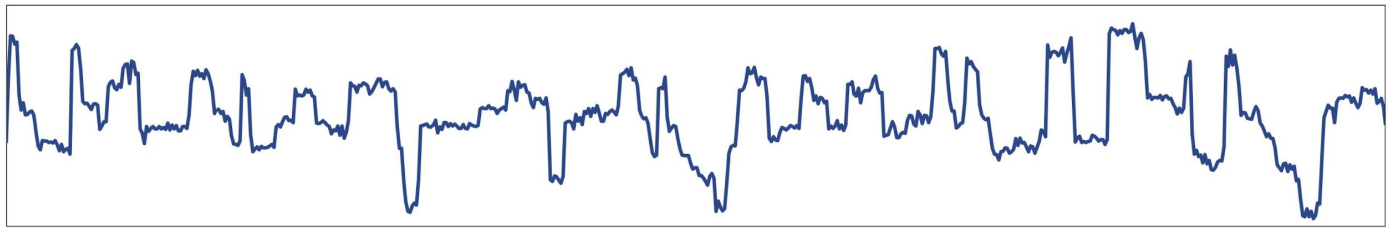
Nanopore





Raw ‘squiggle’ data FAST5

FASTQ



PacBio

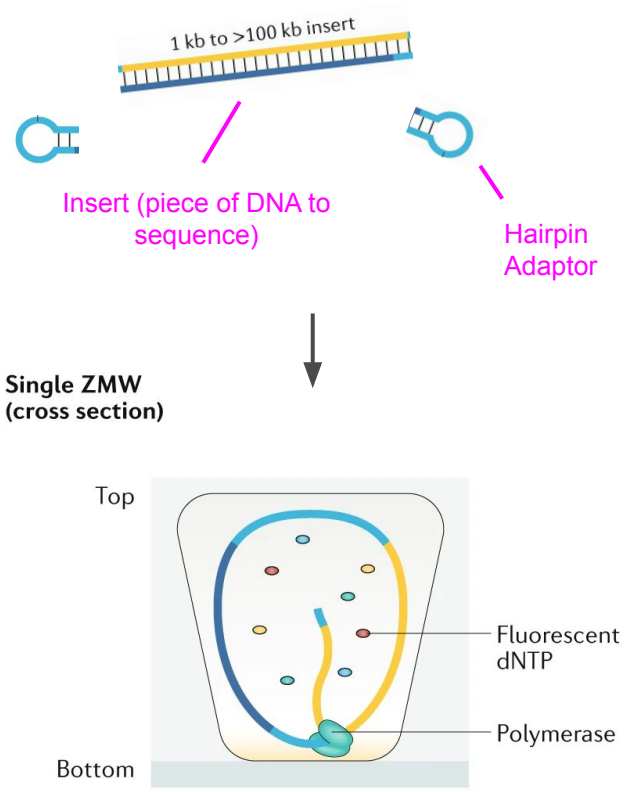
High-accuracy long reads
> 99% accuracy (HiFi reads)

Multiple run methods
Choose between throughput or accuracy

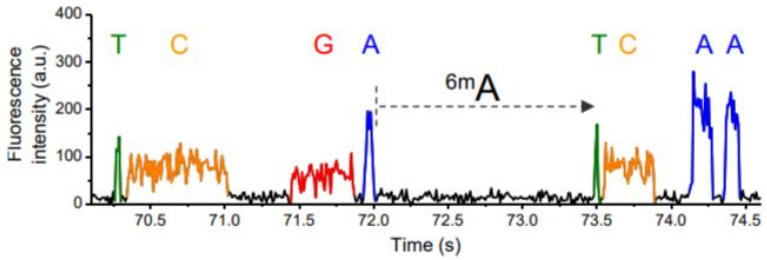
Heading
Transcript isoform detection

Image credit:
Pacific Biosciences of California, Inc

PacBio

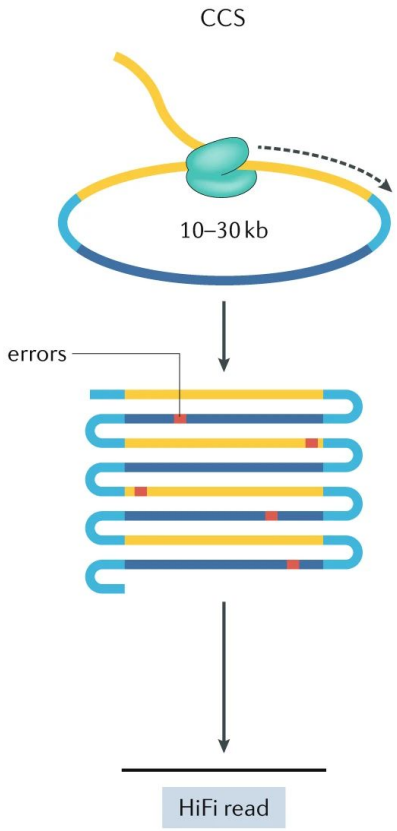


Continuous Long Read (CLR)

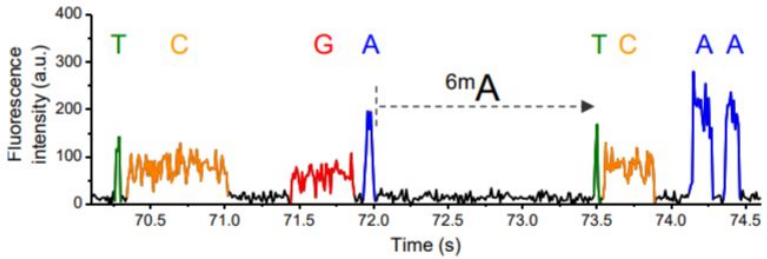


Logsdon, Vollger, Eichler 2020
https://doi.org/10.1038/s41576-020-0236-x

PacBio



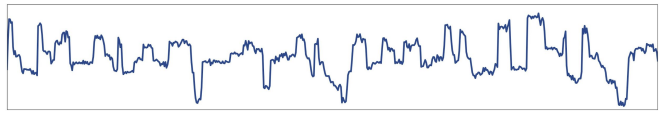
Circular Consensus Sequencing (CCS)



Logsdon, Vollger, Eichler 2020
https://doi.org/10.1038/s41576-020-0236-x

Base calling

Raw ‘squiggle’ data FAST5

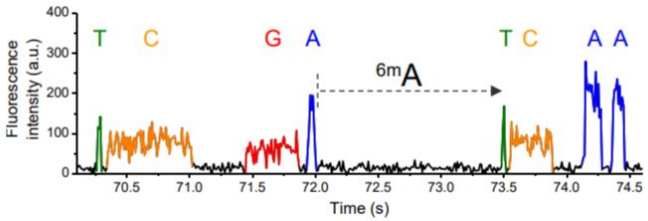


Current basecaller:
Guppy (RNN)

FASTQ

```
@ERR3152364.1.196571 3f3fb634-ac4e
TACGGTAGCCCACTTTCCCGTTCAGTTACGTATT
```

Fluorescence



FASTQ

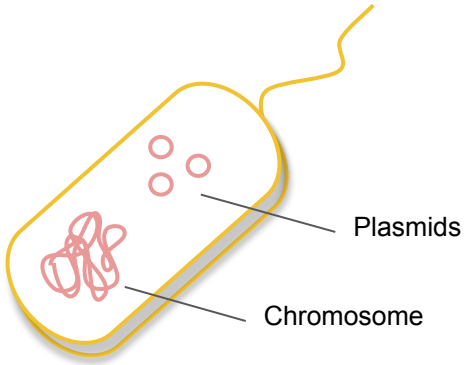
```
@ERR3152364.1.196571 3f3fb634-ac4e
TACGGTAGCCCACTTTCCCGTTCAGTTACGTATT
```

Genome Assembly

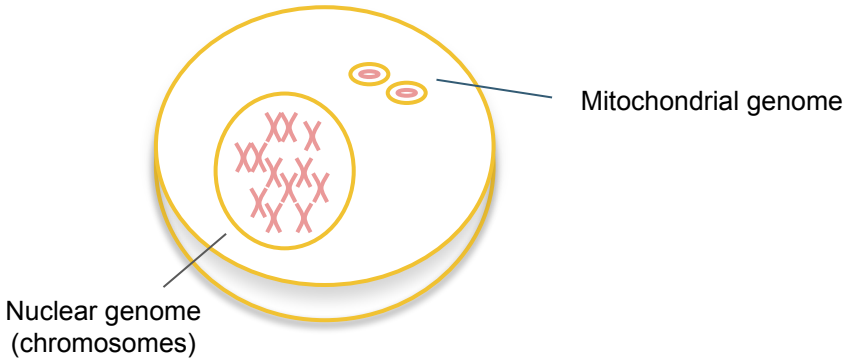
Different organism
Different method

The Genome

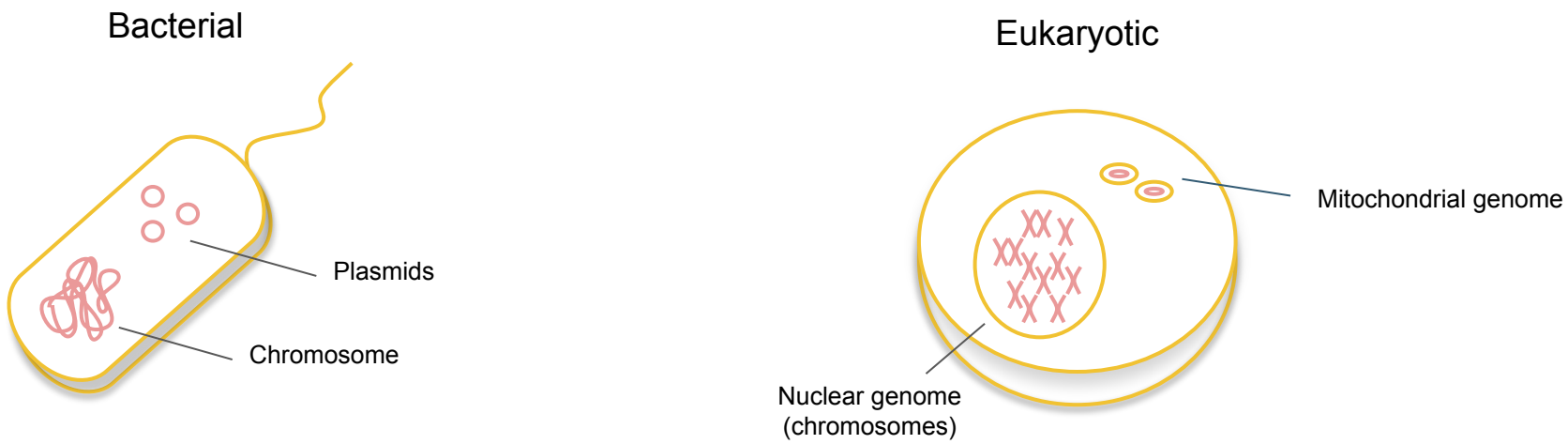
Bacterial



Eukaryotic



The Genome



Be careful about DNA source

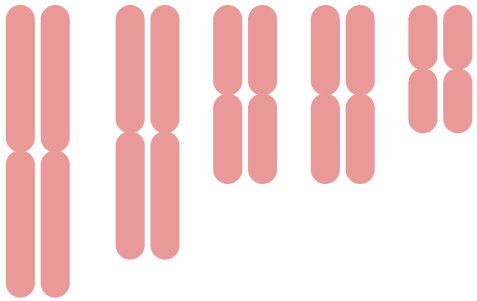
Bacteria: sample a single clonal colony from culture media
Human: cultured cell line

If anything else, there will be foreign DNA contamination.
Microbiome from human cheek swab, multiple organisms in any environmental sample

Can use software filtering methods to identify and remove contaminating DNA

Eukaryotic genomes

Nuclear genome



23 chromosomes
Linear chromosomes
Large genome ~ 6,400,000,000bp (2000x)
~ 30000 genes
Often diploid (2 copy of each chromosome), can be polyploid

Coding sparse: < 2% genome
Repetitive DNA 40% - 80% of genome
Hard to assemble

Mitochondrial genome

16kbp
37 genes
Haploid
Hundreds - thousands of mitochondria per cell

Prokaryotic genomes

Chromosome

Single chromosome (usually)
Chromosomes are circular
Small genome ~ 3,000,000bp
~ 4000 genes
Haploid (1 copy of chromosome)
Relatively easy to assemble

Plasmids

Often 10 - 200 kb
Many copies of each plasmid per cell
Can have multiple different plasmids in single organism (1 - 4 common)
Some convey anti-microbial resistance & virulence

What is genome assembly

Process of transforming sequence reads into a more cohesive picture

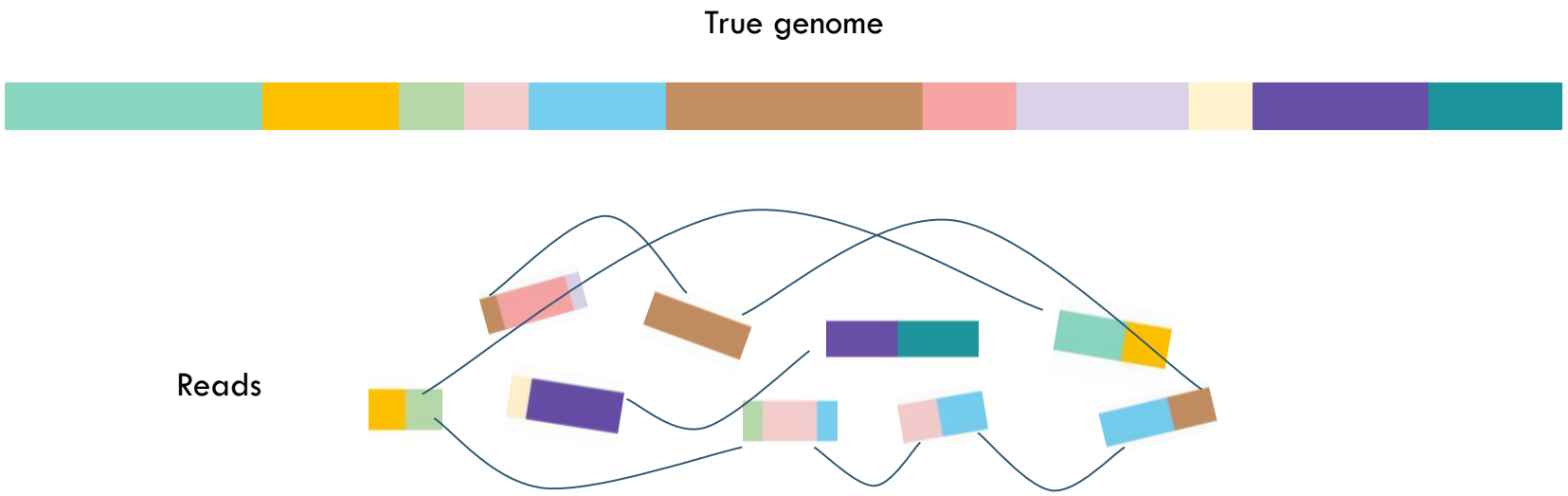
True genome

What is genome assembly

Process of transforming sequence reads into a more cohesive picture

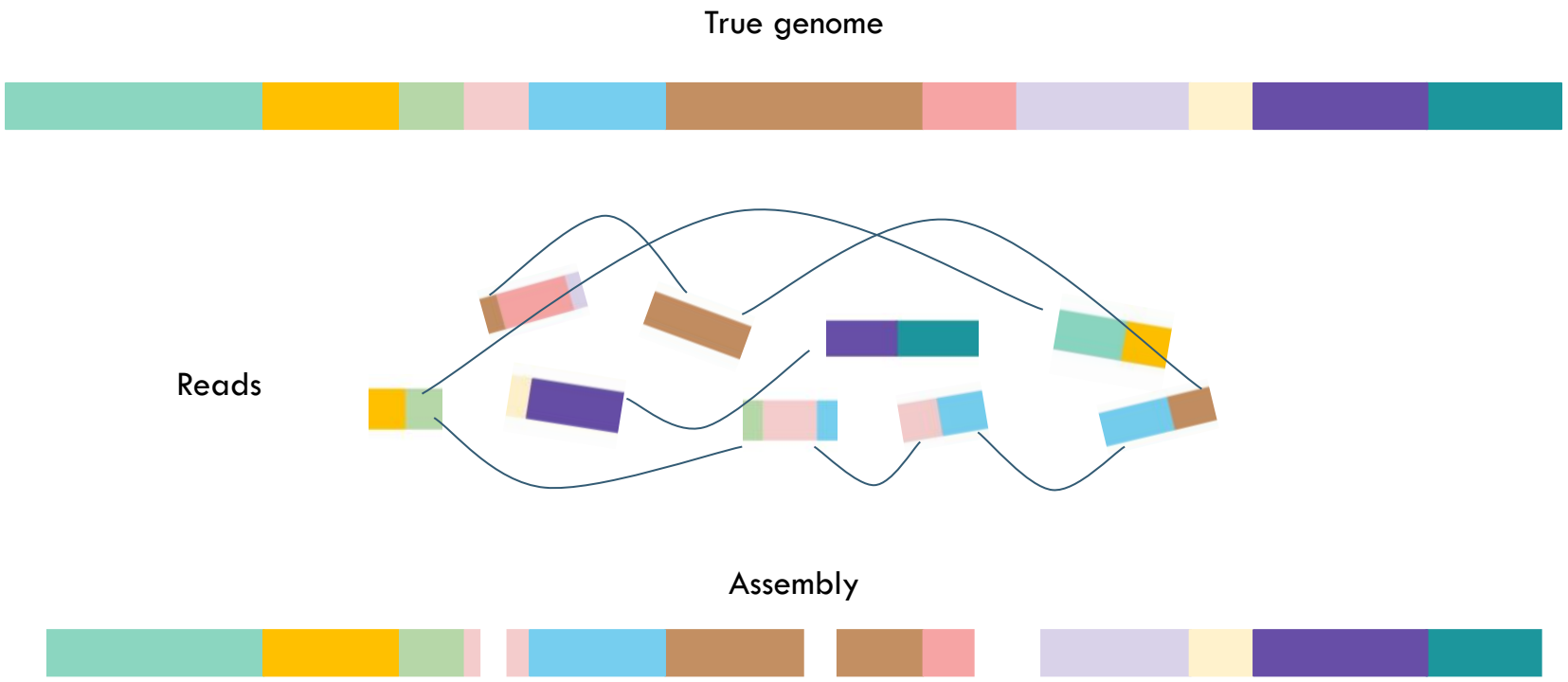
What is genome assembly

Process of transforming sequence reads into a more cohesive picture



What is genome assembly

Process of transforming sequence reads into a more cohesive picture



Technical terms

Reads

DNA fragments sequenced by platform

DNA fragments

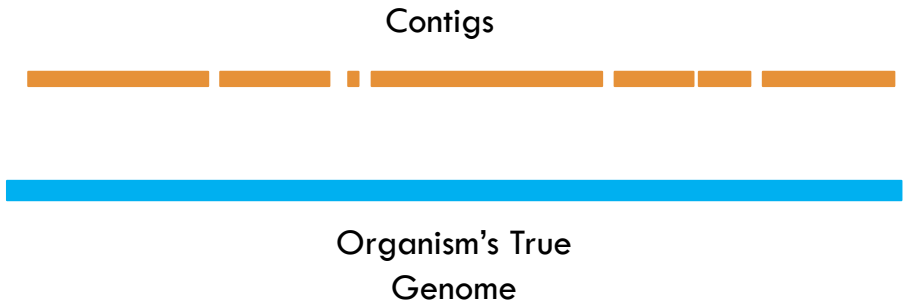
@ERR2137M1.10.2
GTTCGTACGGGCGGGACTG

Nanopore reads

Technical terms

Reads
DNA fragments sequenced by platform

Contig
Unbroken assembled piece of genome sequence

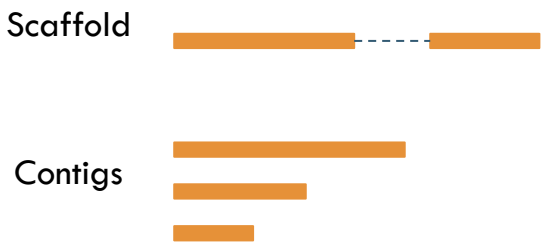


Technical terms

Reads
DNA fragments sequenced by platform

Contig
Unbroken assembled piece of genome sequence

Scaffold
2 or more contigs on same chromosome with known relative position



Technical terms

Reads
DNA fragments sequenced by platform

Contig
Unbroken assembled piece of genome sequence

Scaffold
2 or more contigs on same chromosome with known relative position

Coverage
Ratio of sequenced base pairs to genome length. 30x coverage of human genome (3.2 Gbp) would require 100Gb sequence data.

Technical terms

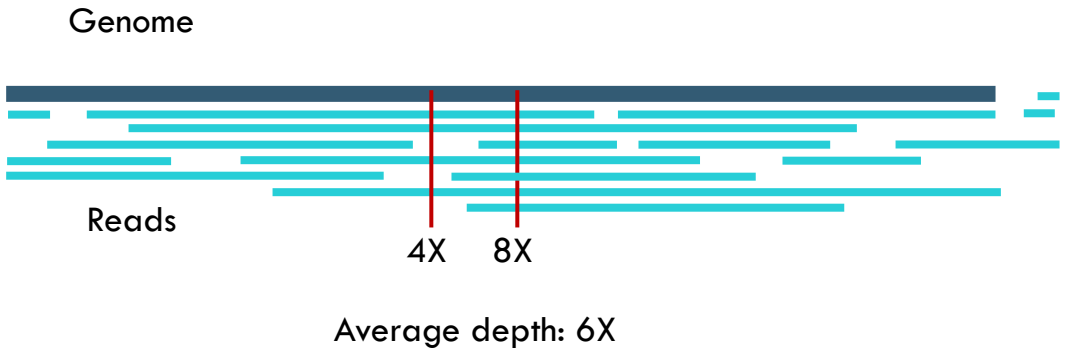
Reads
DNA fragments sequenced by platform

Contig
Unbroken assembled piece of genome sequence

Scaffold
2 or more contigs on same chromosome with known relative position

Coverage
Ratio of sequenced base pairs to genome length. 30x coverage of human genome (3.2 Gbp) would require 100Gb sequence data.

Depth
Number of reads sampling a given nucleotide



Technical terms

Reads
DNA fragments sequenced by platform

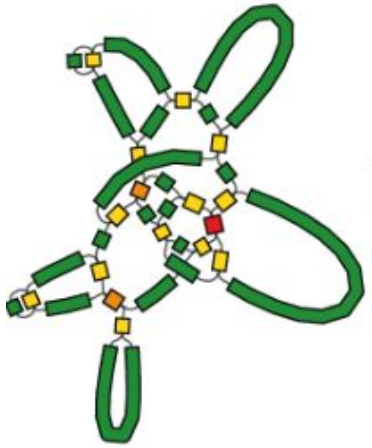
Contig
Unbroken assembled piece of genome sequence

Scaffold
2 or more contigs on same chromosome with known relative position

Coverage
Ratio of sequenced base pairs to genome length. 30x coverage of human genome (3.2 Gbp) would require 100Gb sequence data.

Depth
Number of reads sampling a given nucleotide

Assembly Graph
Contigs and the connections between contigs



Assembly completeness levels

Complete

All chromosomes are gapless and have no runs of 10 or more ambiguous bases

Plasmids and organelles may or may not be included in the assembly but if present then the sequences are gapless

Chromosome

There is sequence for one or more chromosomes

May be a chromosome without gaps or a chromosome containing scaffolds or contigs with gaps between them

May also be unplaced or unlocalized scaffolds

Scaffold

Some sequence contigs have been connected across gaps to create scaffolds

Ie know the relative placement of two contigs

Scaffolds are all unplaced or unlocalized

Contig

Nothing is assembled beyond the level of sequence contigs.

No understanding of relative positions of contigs.

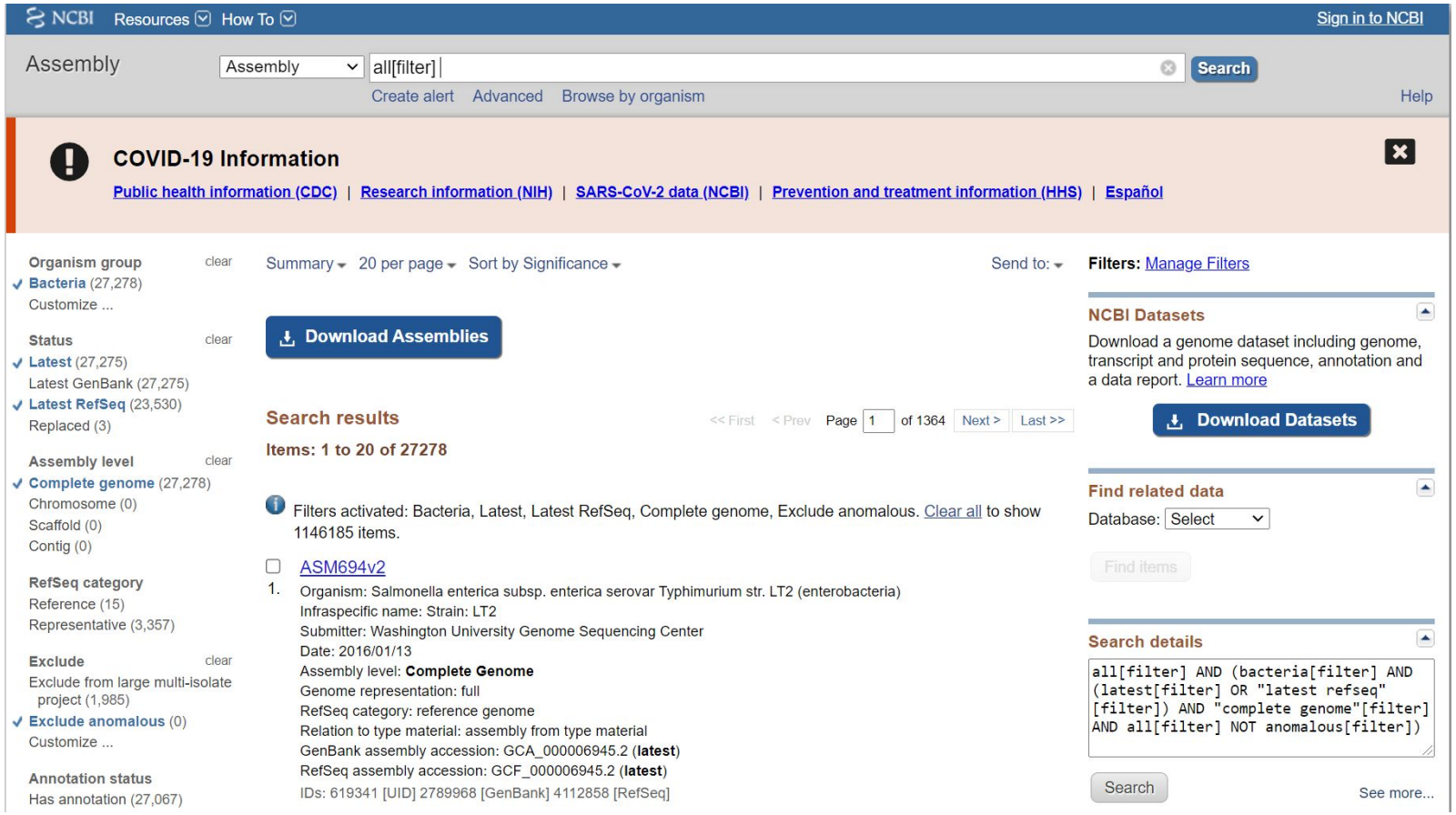
Contigs do not represent complete chromosomes or plasmids

Representing a genome

Keep in mind:

- Haploid representation (single sequence) even in diploid or polypoid organisms
- Assembly may be derived from a single individual or (usually) a group
  - Bacteria - DNA extracted from thousands of individual cells
  - Hg38 consists of sequence from > 50 individual people
- Genetic variation exists within any population
  - For very good assemblies (like Hg38), each position ideally represents the most common allele among the population
  - Even for Hg38 this isn't always true

Depositing and accessing genomes



Repeats

Why do we need long reads anyway?

ROLE OF LONG READS

Interspersed Repeat:

Section of DNA
Occurs in multiple places

Extra copy of gene
More protein produced

New version of gene
New function (some mutation)

ROLE OF LONG READS

Interspersed Repeat:

Section of DNA
Occurs in multiple places

Extra copy of gene
More protein produced

New version of gene
New function (some mutation)

45% of Human Genome

ROLE OF LONG READS

Interspersed Repeat:

Section of DNA
Occurs in multiple places

Extra copy of gene
More protein produced

New version of gene
New function (some mutation)

45% of Human Genome

Tandem Repeat:

Short Repeating Sequence
Side by side

ROLE OF LONG READS

ROLE OF LONG READS

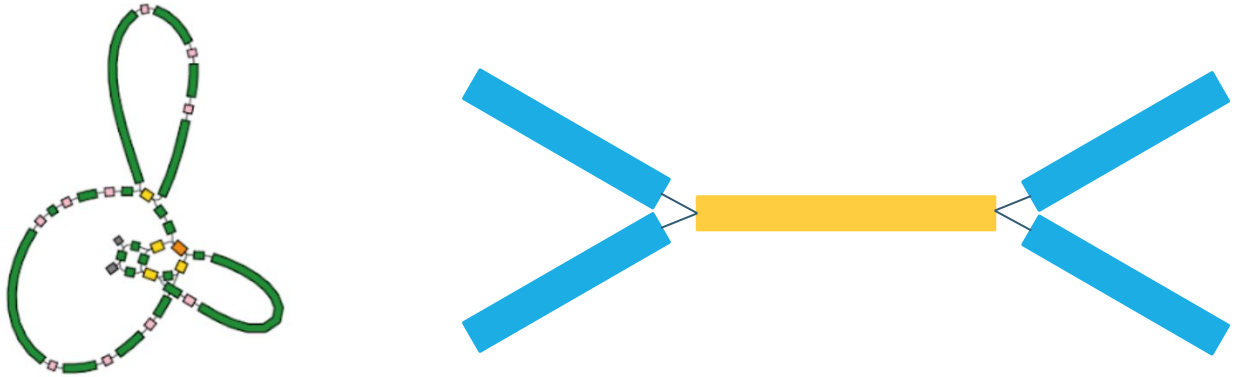
Illumina reads

ROLE OF LONG READS

ROLE OF LONG READS

ROLE OF LONG READS

ROLE OF LONG READS



ROLE OF LONG READS

ROLE OF LONG READS

ROLE OF LONG READS

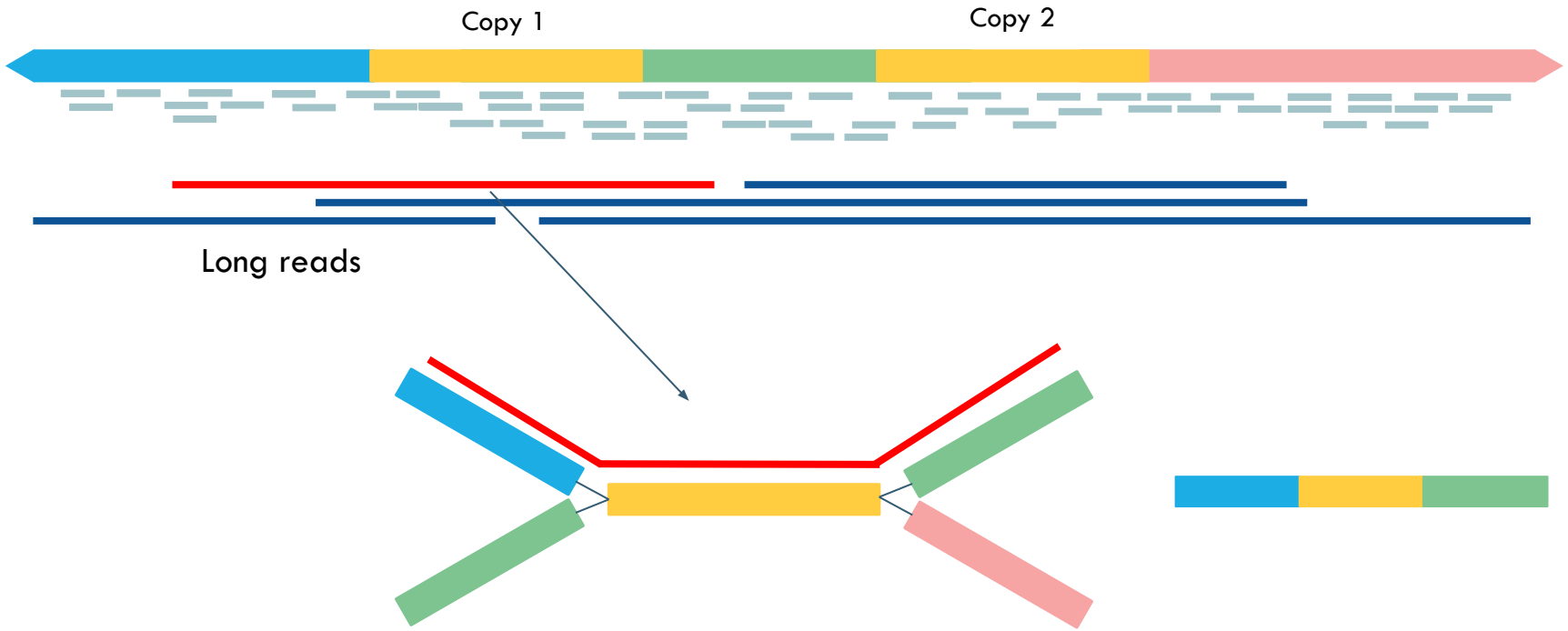
ROLE OF LONG READS

ROLE OF LONG READS

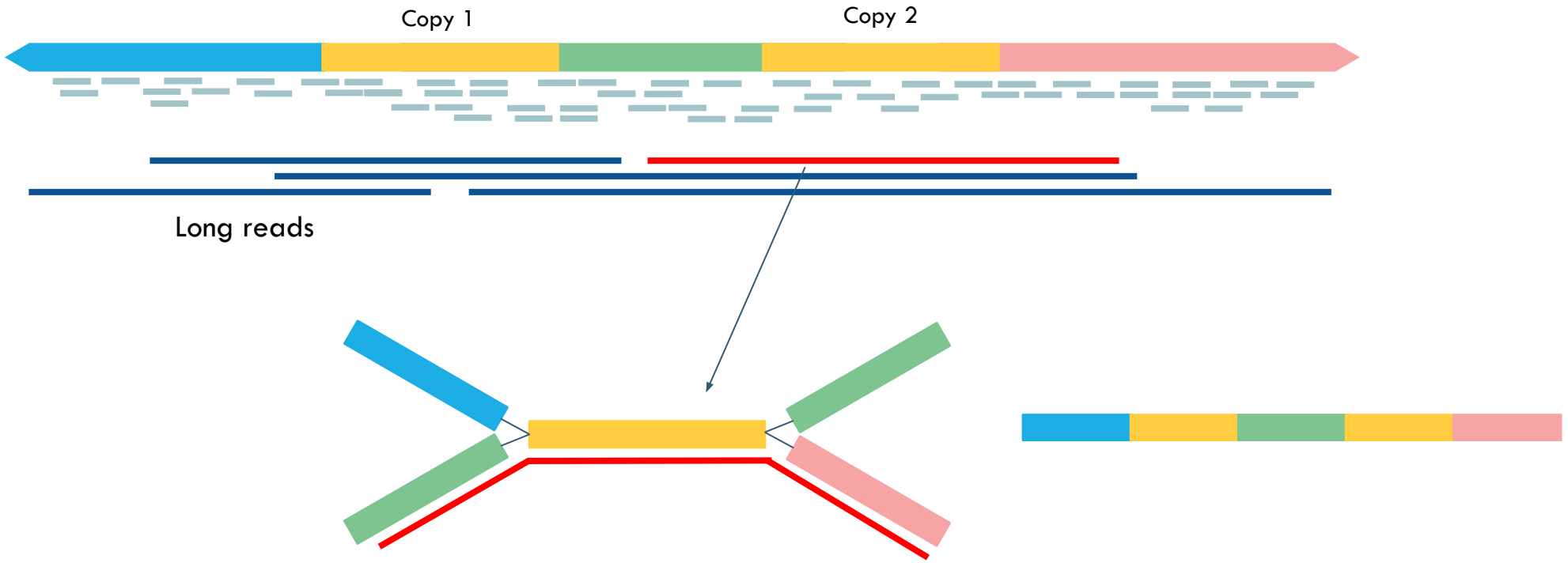
ROLE OF LONG READS

ROLE OF LONG READS

ROLE OF LONG READS



ROLE OF LONG READS



ROLE OF LONG READS

ROLE OF LONG READS

ROLE OF LONG READS

ROLE OF LONG READS

ROLE OF LONG READS

ROLE OF LONG READS

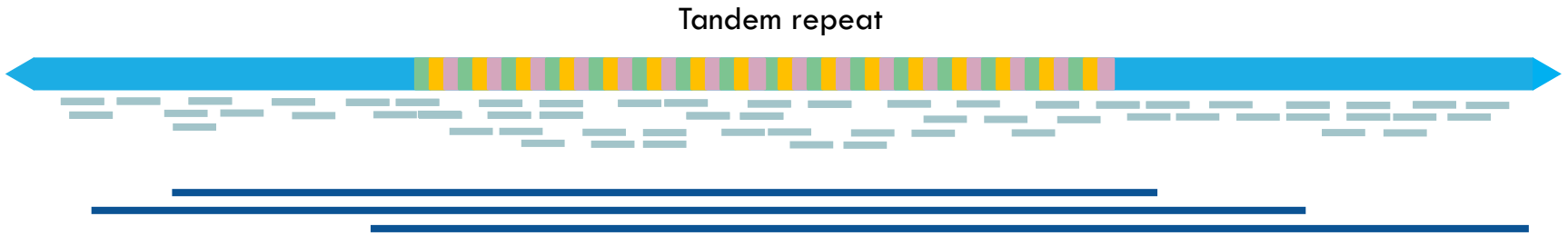
ROLE OF LONG READS

ROLE OF LONG READS

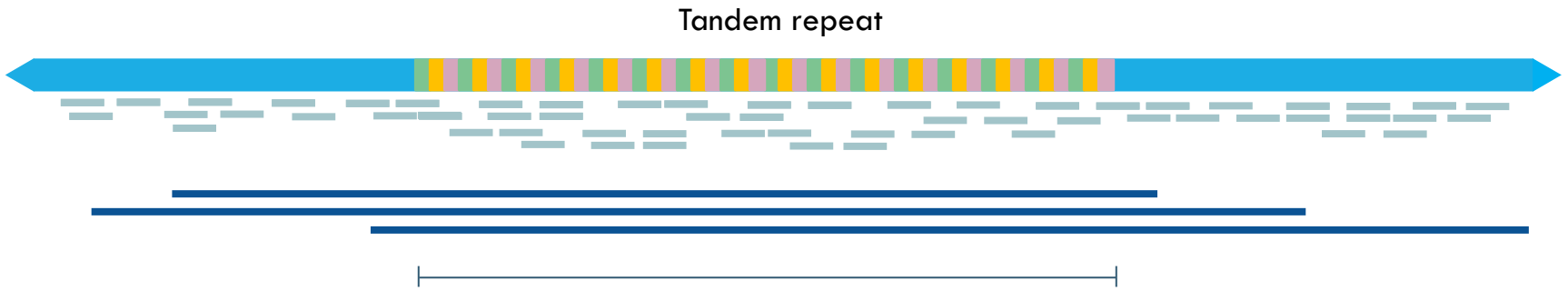
Tandem repeat

1x? 10x? 100x?

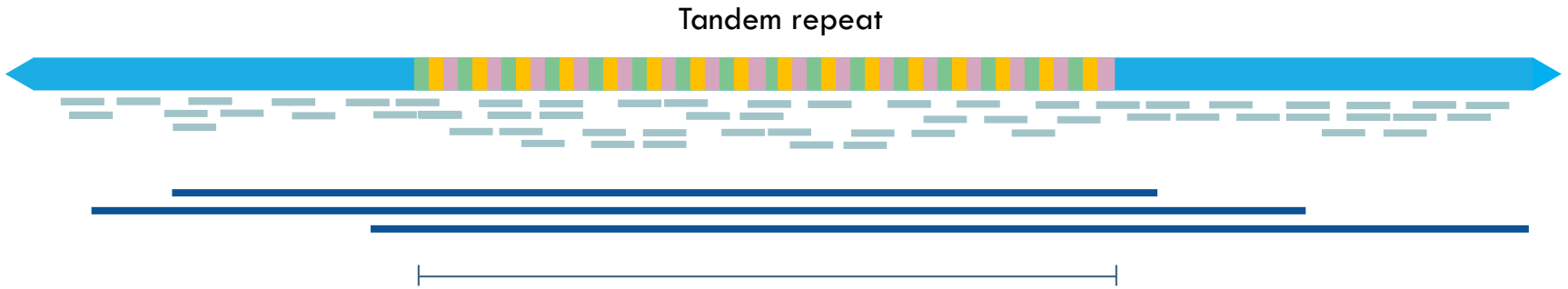
ROLE OF LONG READS



ROLE OF LONG READS



ROLE OF LONG READS



When long reads matter

Long reads provide information about the structural layout of the genome which short sequences cannot.
For small, repeat light genomes (bacterial), not as important to use long reads
For large genomes with repeats (human, plant, mammal, eukaryotes in general) highly needed.

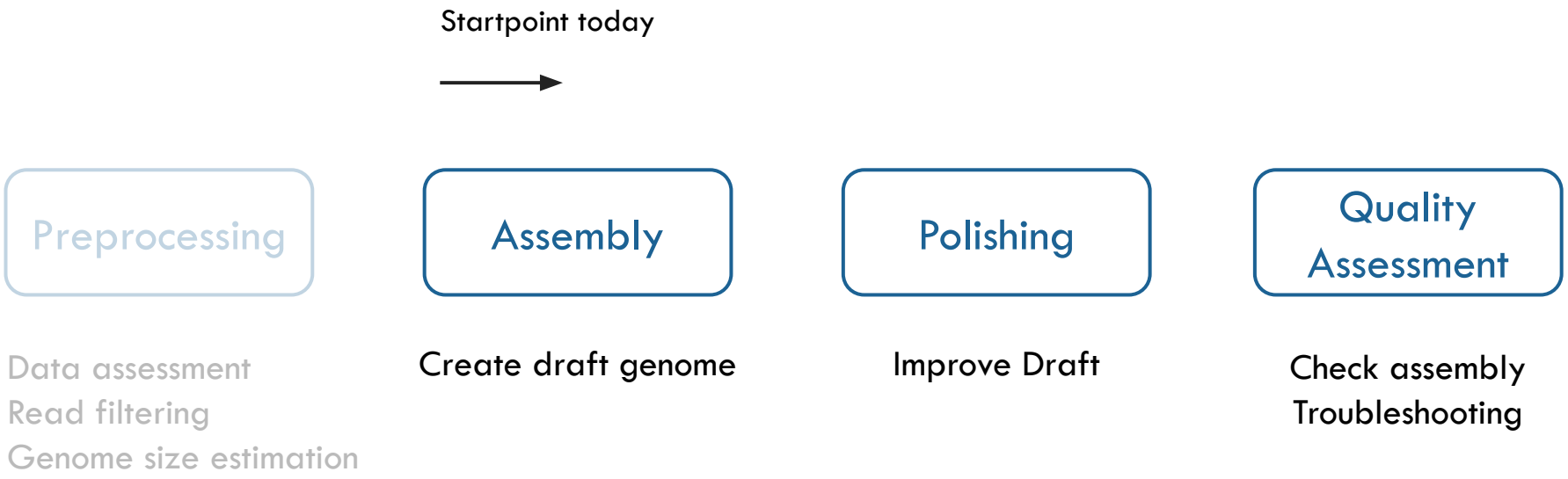
Hybrid Assembly

The beautiful combination of long and short reads

Key steps

Preprocessing	Assembly	Polishing	Quality Assessment
Data assessment Read filtering Genome size estimation	Create draft genome	Improve Draft	Check assembly Troubleshooting

Key steps


Startpoint today
Preprocessing
Assembly
Polishing
Quality Assessment
Data assessment
Read filtering
Genome size estimation
Create draft genome
Improve Draft
Check assembly
Troubleshooting

1. Preprocessing

Data assessment

Was sequencing successful? Is read quality adequate?
Nanoplot / fastqc + multiqc

Read filtering

Remove lowest quality reads from our pool
Long reads - filtlong
Short reads - fastp / trimmomatic

Genome size estimation

Helps some assembly tools by providing estimate
Jellyfish / Meryl + GenomeScope

2. Assembly

3. Polishing

2 Main hybrid methods

- Large genomes: Long-reads-first
- Small genomes: Short-reads-first

2. Assembly

3. Polishing

2 Main hybrid methods

- Large genomes: Long-reads-first
- Small genomes: Short-reads-first

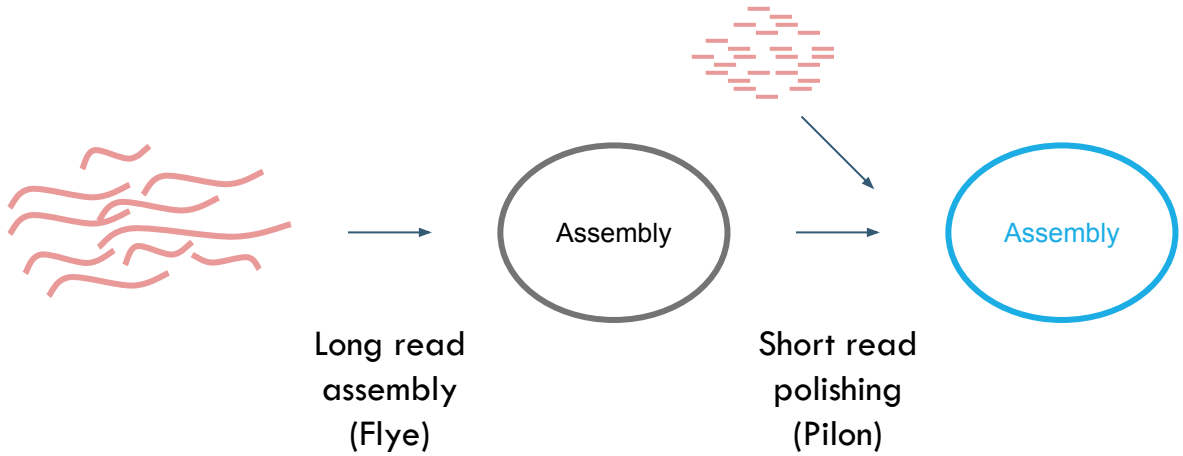
2. Assembly

3. Polishing

2 Main hybrid methods

Large genomes: Long-reads-first

Small genomes: Short-reads-first



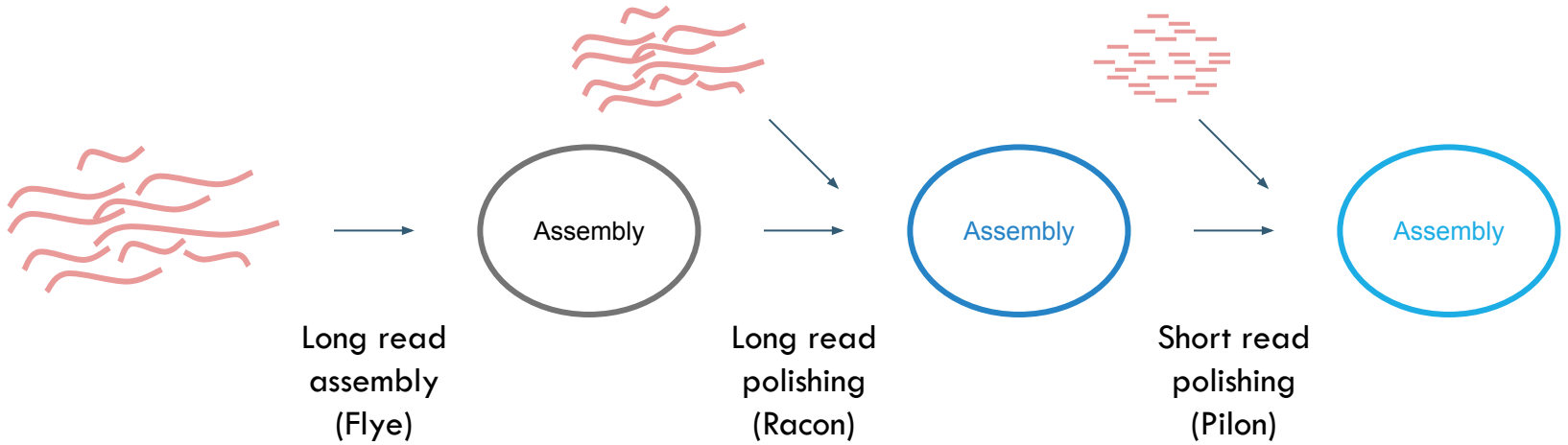
2. Assembly

3. Polishing

2 Main hybrid methods

Large genomes: Long-reads-first

Small genomes: Short-reads-first



2. Assembly

3. Polishing

2 Main hybrid methods

Large genomes: Long-reads-first

Small genomes: Short-reads-first

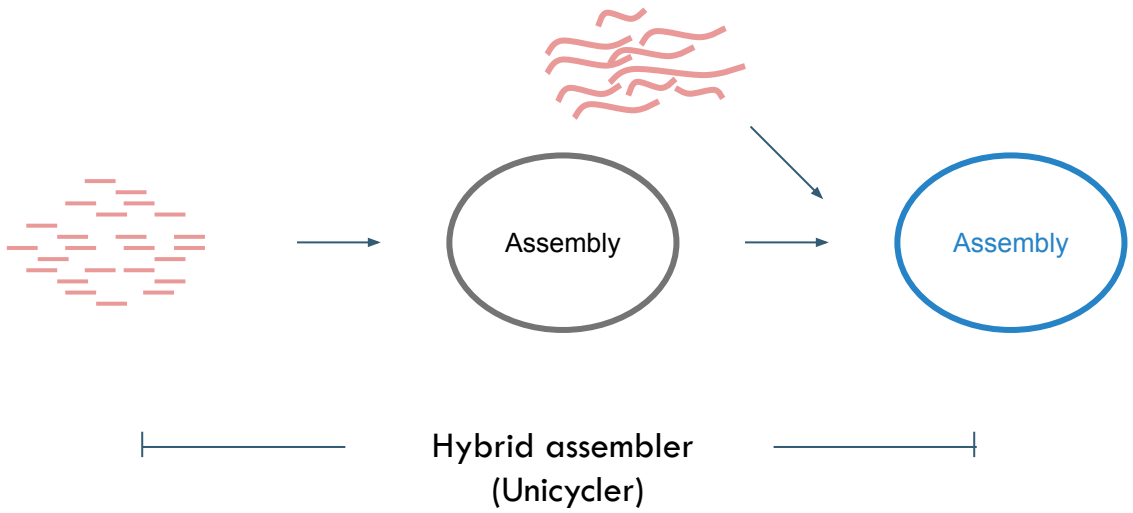
2. Assembly

3. Polishing

2 Main hybrid methods

Large genomes: Long-reads-first

Small genomes: Short-reads-first



4. Quality Assessment

Analyse the quality of assembly produced

- Compare to a similar genome already sequenced: QUAST
- Attempt to locate genes we believe should be present: BUSCO

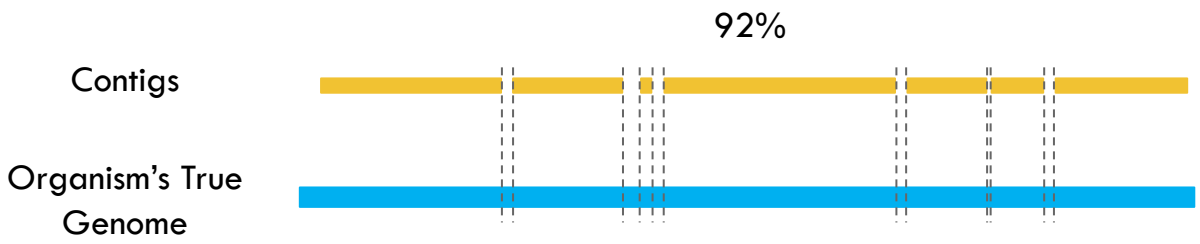
Let’s do it

Assessing Assemblies

Terms

Genome fraction

Proportion of true genome spanned by assembled contigs



Terms

Genome fraction

Proportion of true genome spanned by assembled contigs

Mismatches

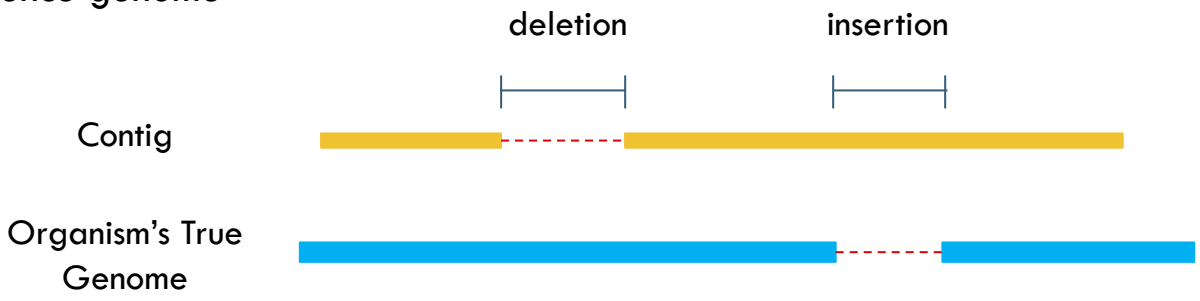
Differing base at position between assembly and reference genome

Terms

Genome fraction
Proportion of true genome spanned by assembled contigs

Mismatches
Differing base at position between assembly and reference genome

Indels
Runs of added or deleted bases relative to a reference



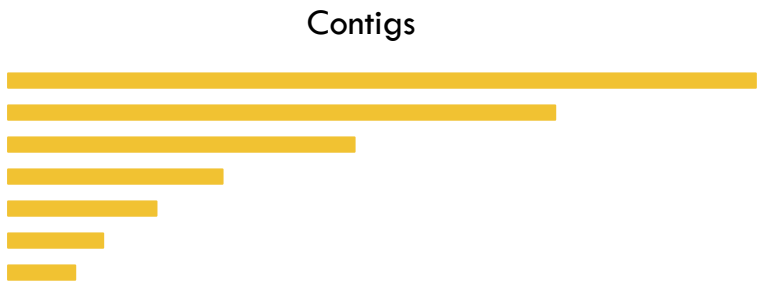
Terms

Genome fraction
Proportion of true genome spanned by assembled contigs

Mismatches
Differing base at position between assembly and reference genome

Indels
Runs of added or deleted bases
relative to a reference

N50
50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value



Terms

Genome fraction
Proportion of true genome spanned by assembled contigs

Mismatches
Differing base at position between assembly and reference genome

Indels
Runs of added or deleted bases
relative to a reference

N50
50% of the entire assembly is contained in contigs or scaffolds equal to or larger than this value

Misassemblies
Structural variation between assembly and reference genome

BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Ortholog

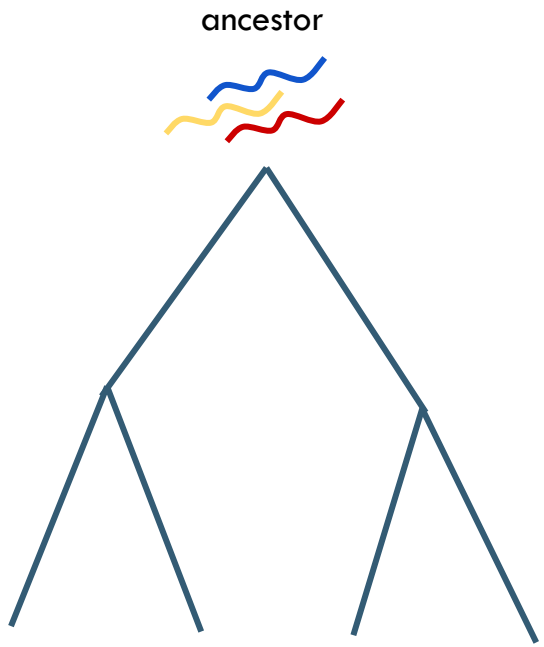
- Gene present in multiple organisms
- Evolved from common ancestor
- All living organisms - **DNA polymerase**

BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Ortholog

- Gene present in multiple organisms
- Evolved from common ancestor
- All living organisms - **DNA polymerase**

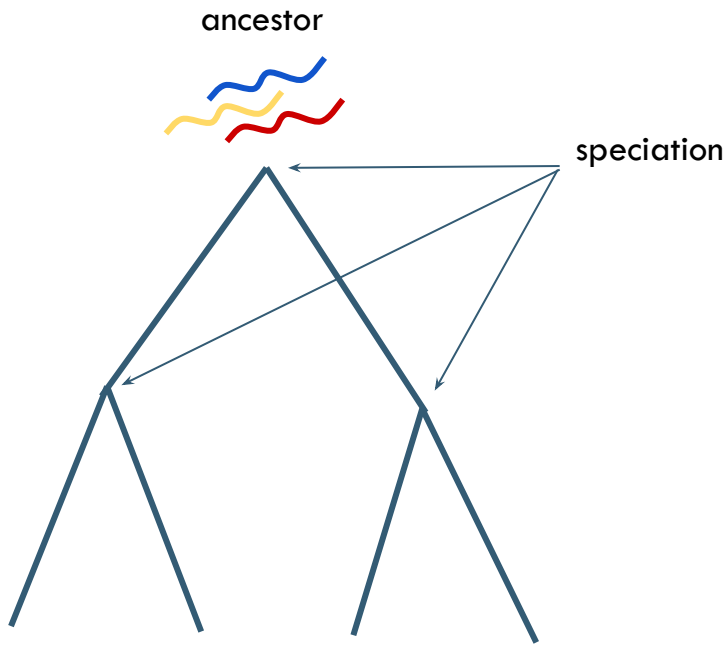


BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Ortholog

- Gene present in multiple organisms
- Evolved from common ancestor
- All living organisms - **DNA polymerase**

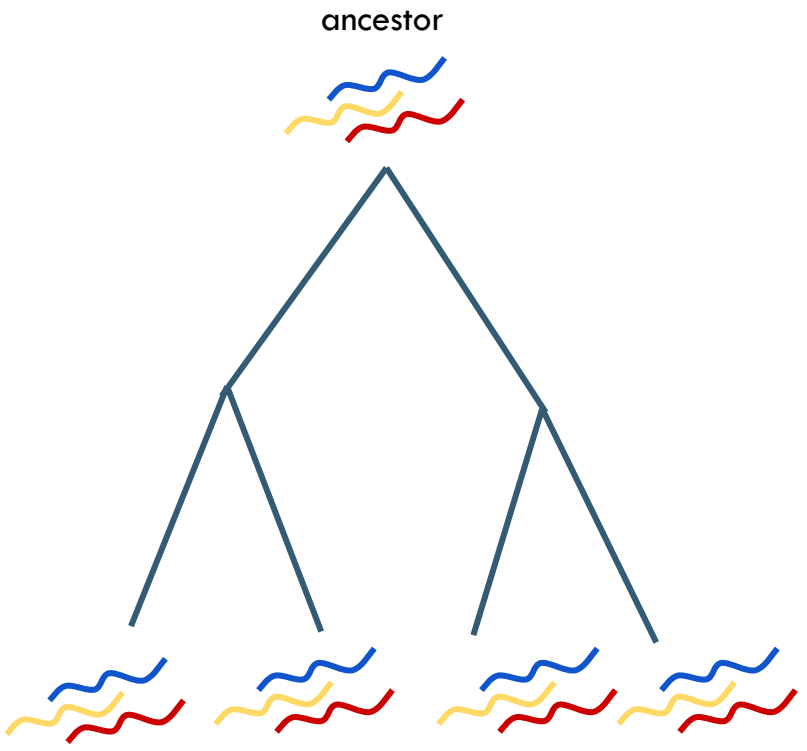


BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Ortholog

- Gene present in multiple organisms
- Evolved from common ancestor
- All living organisms - **DNA polymerase**

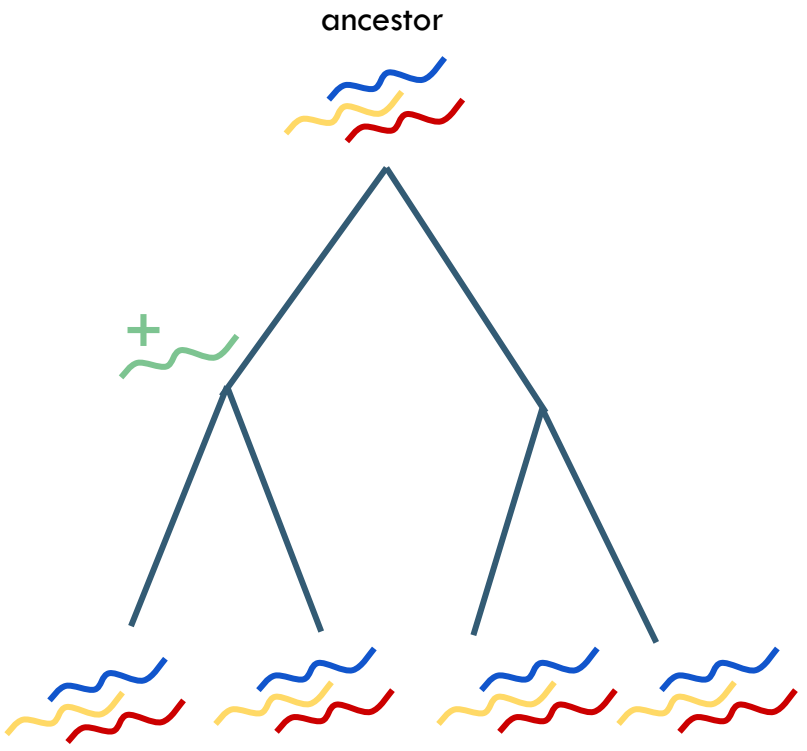


BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Ortholog

- Gene present in multiple organisms
- Evolved from common ancestor
- All living organisms - **DNA polymerase**



BUSCO

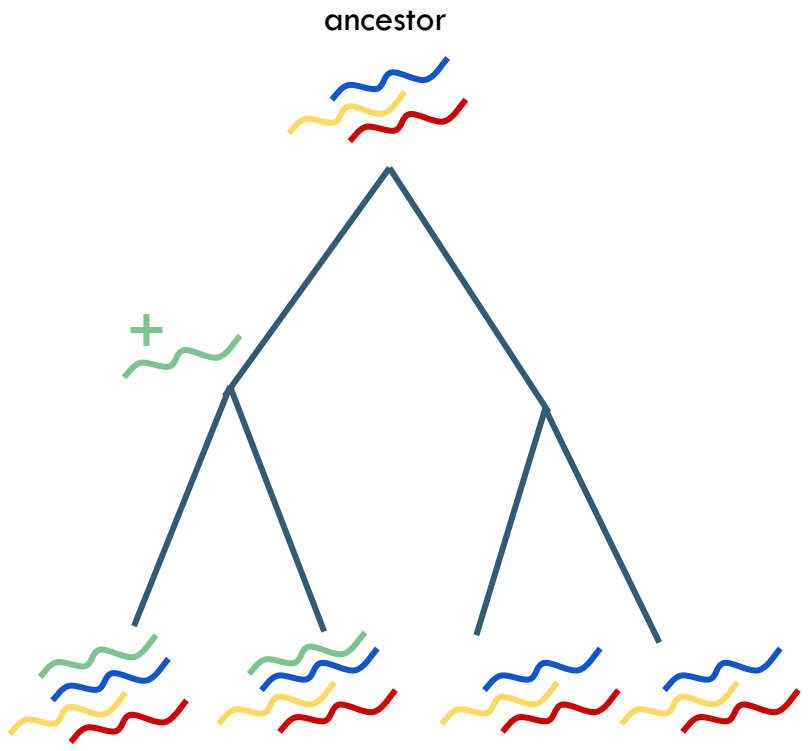
Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Ortholog

Gene present in multiple organisms

Evolved from common ancestor

All living organisms - **DNA polymerase**

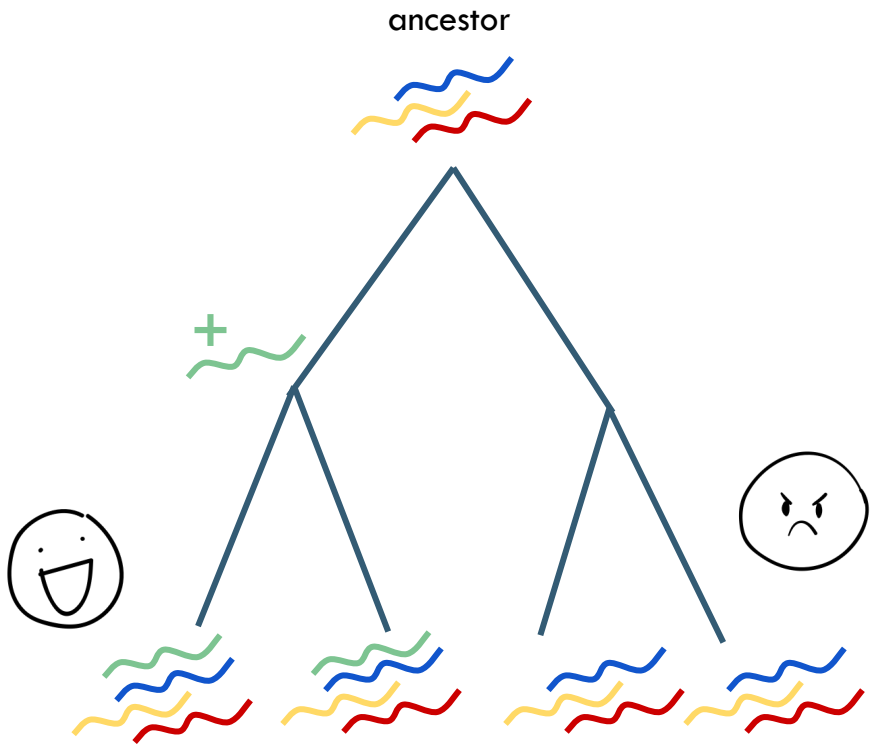


BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Ortholog

- Gene present in multiple organisms
- Evolved from common ancestor
- All living organisms - **DNA polymerase**



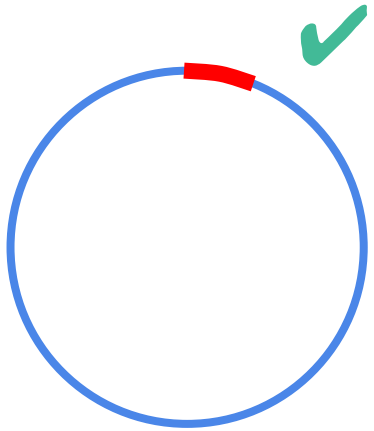
BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Single-Copy

Appears once in organisms

No duplicates allowed

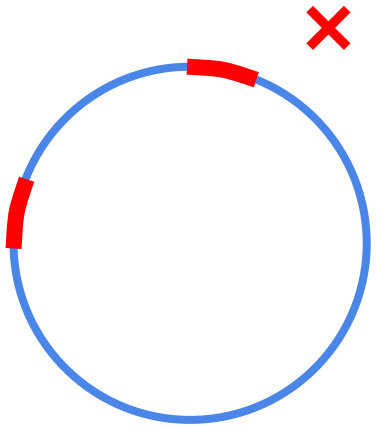


BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Single-Copy

- Appears once in organisms
- No duplicates allowed



BUSCO

Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

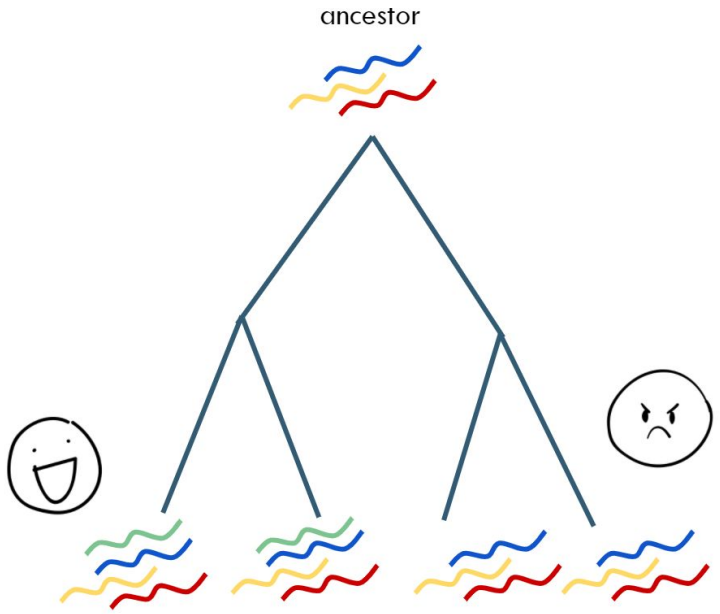
Universal

You define the **clade**

Bacillus subtilis:

- Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus subtilis group

BUSCO



Benchmarking **U**niversal **S**ingle-**C**opy **O**rthologs

Universal

You define the **clade**

Bacillus subtilis:

? Bacteria; Terrabacteria group; Firmicutes; Bacilli; Bacillales; Bacillaceae; Bacillus; Bacillus subtilis group

124 BUSCOs

450 BUSCOs

Results

Quast

Genome statistics	nanopore_draft_assembly
Genome fraction (%)	97.598
Duplication ratio	1.009
Largest alignment	692 020
Total aligned length	3 980 483
NGA50	252 801
LGA50	5
Misassemblies	
# misassemblies	0
Misassembled contigs length	0
Mismatches	
# mismatches per 100 kbp	77.65
# indels per 100 kbp	550.99
# N's per 100 kbp	0
Statistics without reference	
# contigs	29
Largest contig	692 065
Total length	3 986 877
Total length (>= 1000 bp)	3 986 877

BUSCO

```
C:14.4%[S:14.4%,D:0.0%],F:41.1%,M:44.5%,n:450
65      Complete BUSCOs (C)
65      Complete and single-copy BUSCOs (S)
0       Complete and duplicated BUSCOs (D)
185     Fragmented BUSCOs (F)
200     Missing BUSCOs (M)
450     Total BUSCO groups searched
```

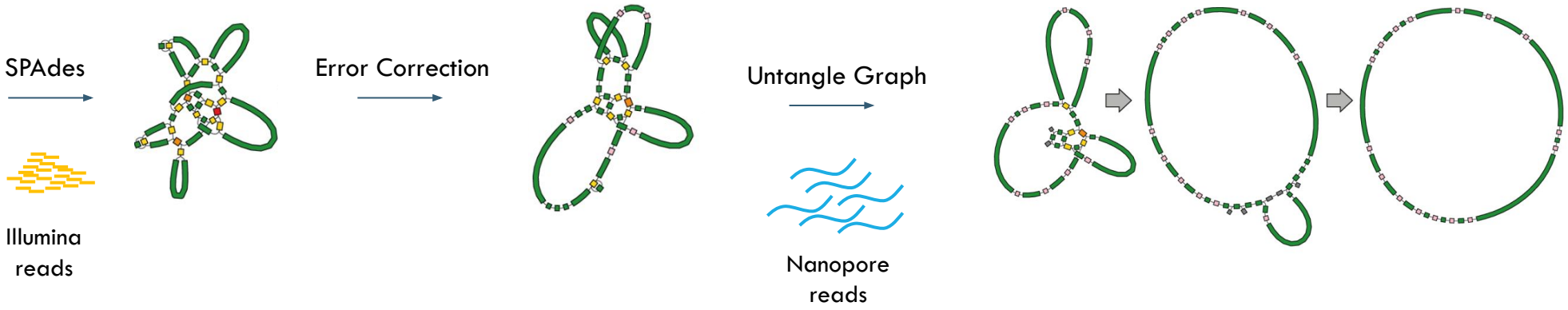
Unicycler

Unicycler

Short-reads-first

Unicycler

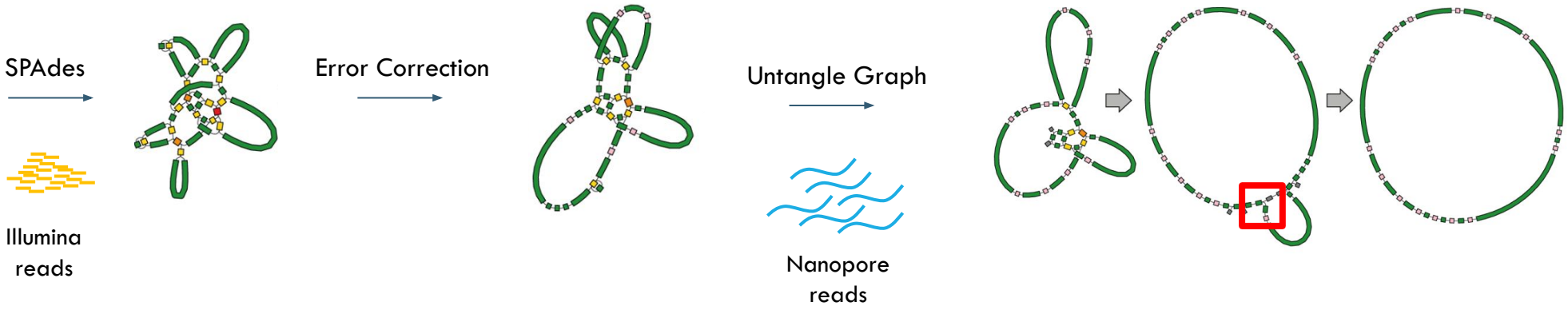
Short-reads-first



Ryan Wick 2017

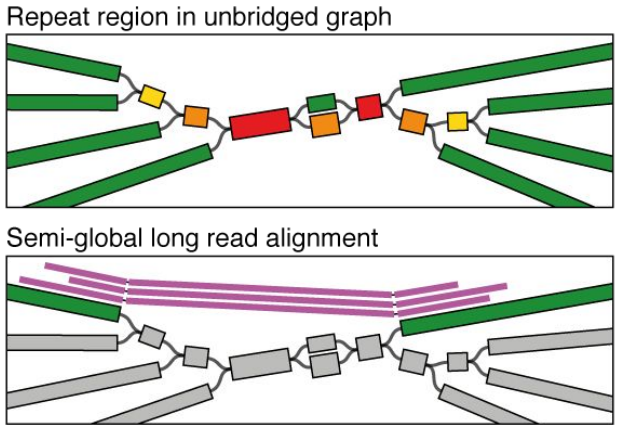
Unicycler

Short-reads-first



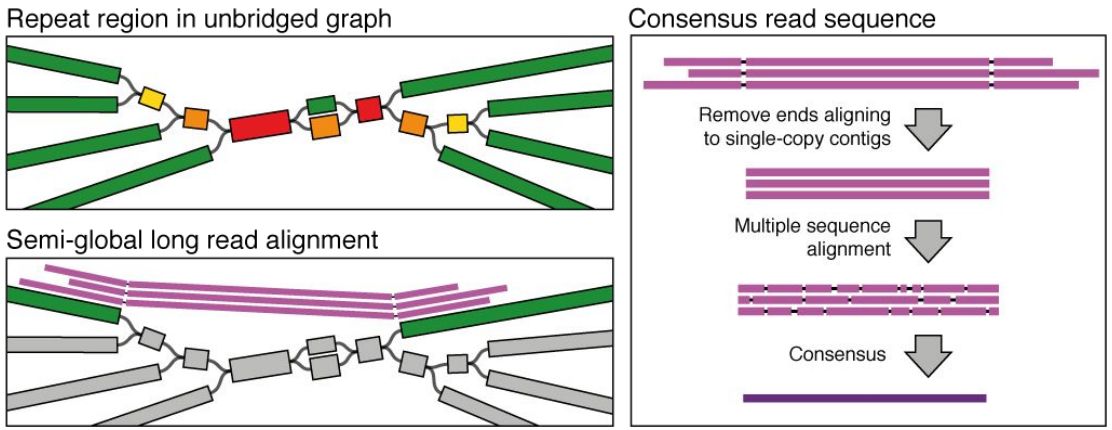
Ryan Wick 2017

Unicycler



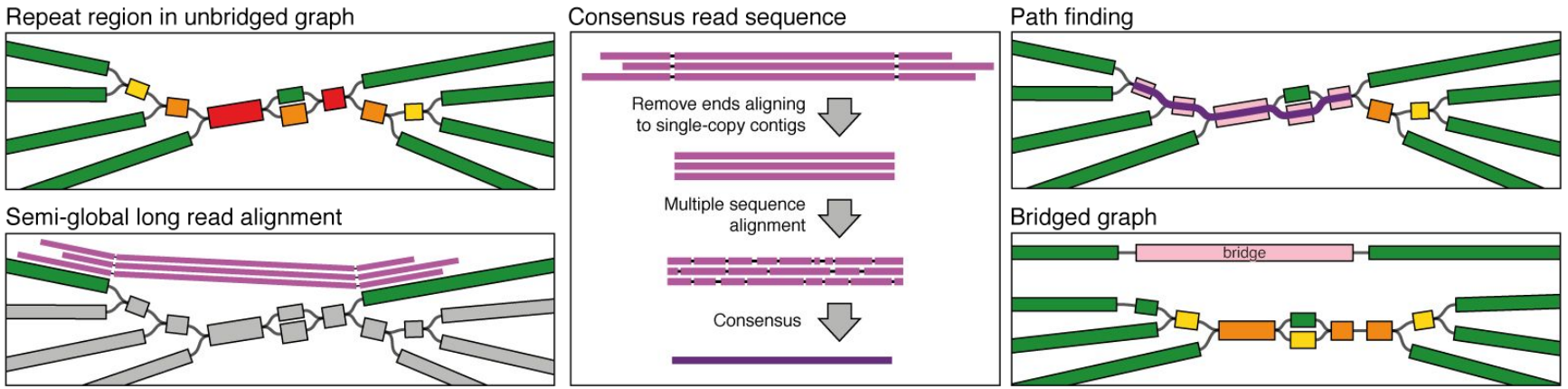
Ryan Wick 2017

Unicycler



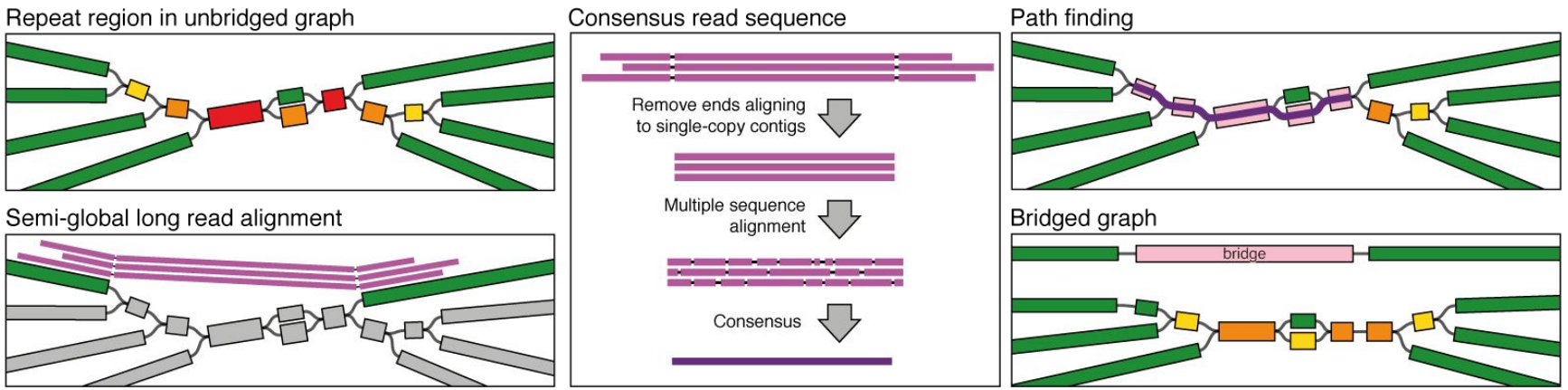
Ryan Wick 2017

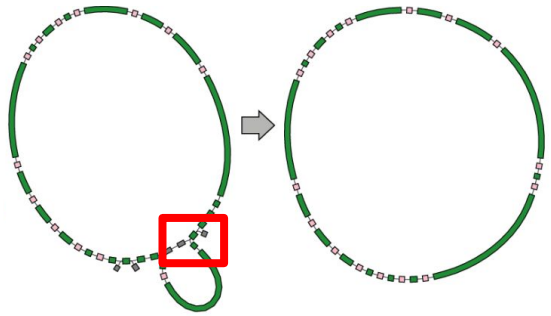
Unicycler



Ryan Wick 2017

Unicycler





Ryan Wick 2017

Fine-tuning & technical considerations

Why do we need long reads anyway?

Assembly stages

Quast

	Nanopore Draft	Nanopore Draft + Pilon Polishing	Unicycler
genome fraction (%)	97.6	97.602	97.92
# contigs	30	30	10
# mismatches per 100kb	80.59	9.75	3.28
# indels per 100kb	552.98	35.25	1.74

BUSCO

Nanopore Draft

```
C:15.1%[S:15.1%,D:0.0%],F:42.0%,M:42.9%,n:450
68      Complete BUSCOs (C)
68      Complete and single-copy BUSCOs (S)
0       Complete and duplicated BUSCOs (D)
189     Fragmented BUSCOs (F)
193     Missing BUSCOs (M)
450     Total BUSCO groups searched
```

Nanopore Draft + Pilon Polishing

```
C:90.9%[S:90.7%,D:0.2%],F:5.3%,M:3.8%,n:450
409     Complete BUSCOs (C)
408     Complete and single-copy BUSCOs (S)
1       Complete and duplicated BUSCOs (D)
24      Fragmented BUSCOs (F)
17      Missing BUSCOs (M)
450     Total BUSCO groups searched
```

Unicycler

```
C:98.4%[S:98.2%,D:0.2%],F:0.4%,M:1.2%,n:450
443     Complete BUSCOs (C)
442     Complete and single-copy BUSCOs (S)
1       Complete and duplicated BUSCOs (D)
2       Fragmented BUSCOs (F)
5       Missing BUSCOs (M)
450     Total BUSCO groups searched
```

Tool parameters

Unicycler Bridging Modes

Quast

	Conservative	Normal	Bold
genome fraction (%)	97.916	97.920	97.952
# contigs	11	10	8
# mismatches per 100kb	3.23	3.28	3.33
# indels per 100kb	1.64	1.74	1.94

BUSCO

Conservative

```
C:98.4%[S:98.2%,D:0.2%],F:0.4%,M:1.2%,n:450
443     Complete BUSCOs (C)
442     Complete and single-copy BUSCOs (S)
1       Complete and duplicated BUSCOs (D)
2       Fragmented BUSCOs (F)
5       Missing BUSCOs (M)
450     Total BUSCO groups searched
```

Normal

```
C:98.4%[S:98.2%,D:0.2%],F:0.4%,M:1.2%,n:450
443     Complete BUSCOs (C)
442     Complete and single-copy BUSCOs (S)
1       Complete and duplicated BUSCOs (D)
2       Fragmented BUSCOs (F)
5       Missing BUSCOs (M)
450     Total BUSCO groups searched
```

Bold

```
C:98.4%[S:98.2%,D:0.2%],F:0.4%,M:1.2%,n:450
443     Complete BUSCOs (C)
442     Complete and single-copy BUSCOs (S)
1       Complete and duplicated BUSCOs (D)
2       Fragmented BUSCOs (F)
5       Missing BUSCOs (M)
450     Total BUSCO groups searched
```

Run times & resources

FLYE

Genome	Genome size	Input data	CPU time	RAM
E.coli	5 Mbp	250 Mb	2 h	2 Gb
C.elegans	100 Mbp	4 Gb	100 h	31 Gb
A.thaliana	135 Mbp	10 Gb	100 h	59 Gb
D.melanogaster	140 Mbp	4 Gb	130 h	33 Gb
D.melanogaster	140 Mbp	17 Gbp	150 h	70 Gb
Human NA12878	3200 Mbp	112000 Gbp	3000 h	394 Gb

FLYE scaling:

CPU time - 1hr per Mb genome size
RAM - log relationship

Multithreading usually available:

3000 cpu hours = 48 real hours on machine with 64 cores

Ensure your compute has enough RAM to handle the genome!

Run times & resources

Genome	Genome size	Assembler	Genome fraction	Complete Single copy BUSCOs (%)	Mismatch + indel rate (per 100 kbp)	Time (h)	Memory (Gb)
E. coli	5 Mbp	Canu	99.6	4	1354.2	0.5	4
		Flye	99.9	15	1089.3	0.2	12
		SPAdes	98.3	98	1.3	0.5	114
		Unicycler	99.9	99.5	3.3	0.5	22
C. elegans	100 Mbp	Canu	99.7	96.8	125.7	4	14
		Flye	99.6	98.0	103.6	1	90
		SPAdes	92.0	90.8	16.8	2	75
		Unicycler	97.0	97.1	90.2	24	105
H. sapiens	3200 Mbp	Canu	95.1	94.6	165.4	562	59
		Flye	95.5	89.7	284.9	120	400
		SPAdes	NA	NA	NA	∞	∞
		Unicycler	NA	NA	NA	∞	∞

Machine has 64 cores and 720 Gb RAM. Figures are approximate.

Many tools!

Pick the one which suits your requirements

Long read assembly

Tool	Properties
Flye	Best overall
Canu	Low RAM requirement
Miniasm+	Good contiguity & plasmid assembly
Shasta	Low resource usage, low runtime

Polishing

Tools	Reads type
Racon	Long read polishing
Medaka	Long read polishing (only ONT)
Pilon	Short reads
NextPolish	Short reads

Special cases: If reads are PacBio Hifi, can use HiCanu (purpose-built)

Thank you!