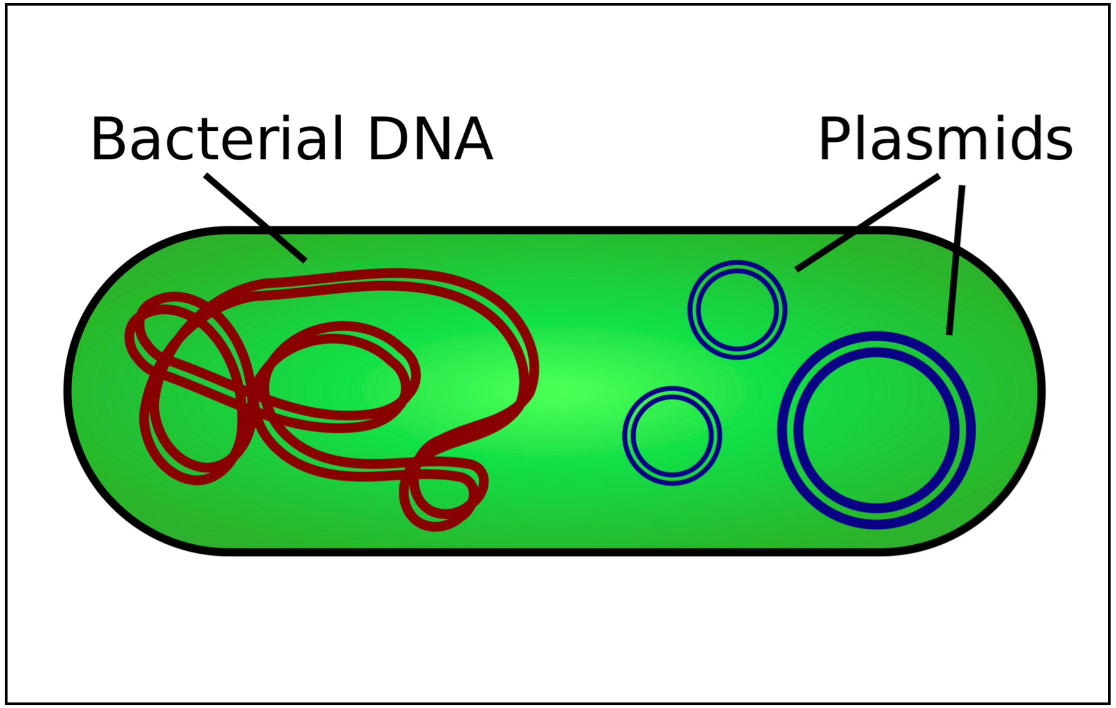
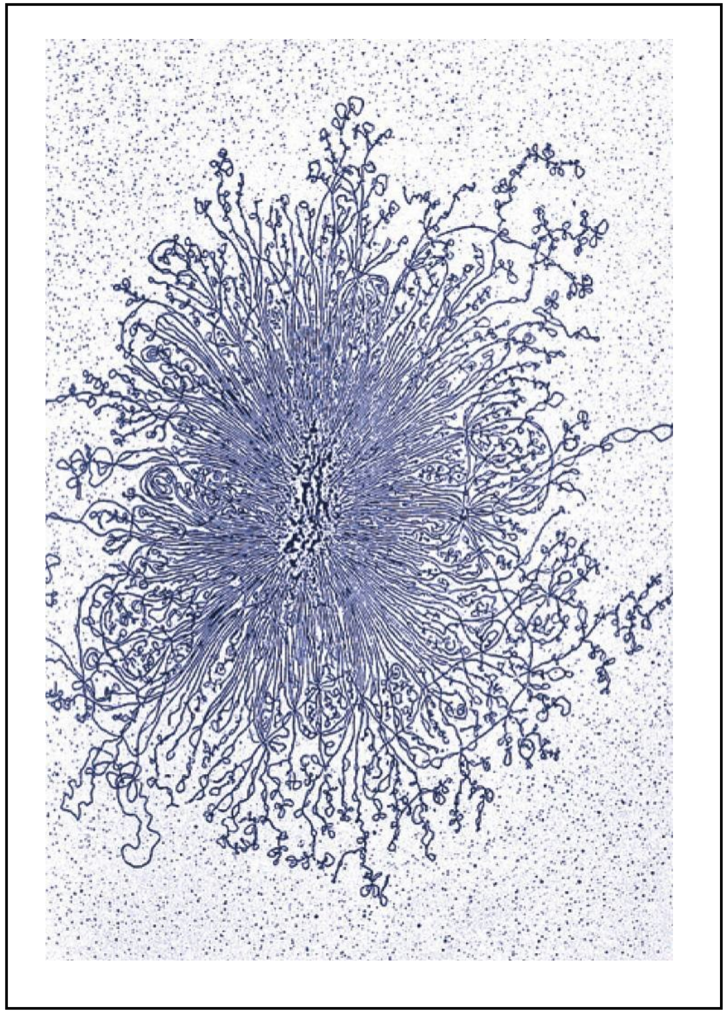
Closing
(some of the issues with circular)
Genomes

Jurassic Park, 1993 (28 years ago)

Erin Young, PhD
September 17, 2021

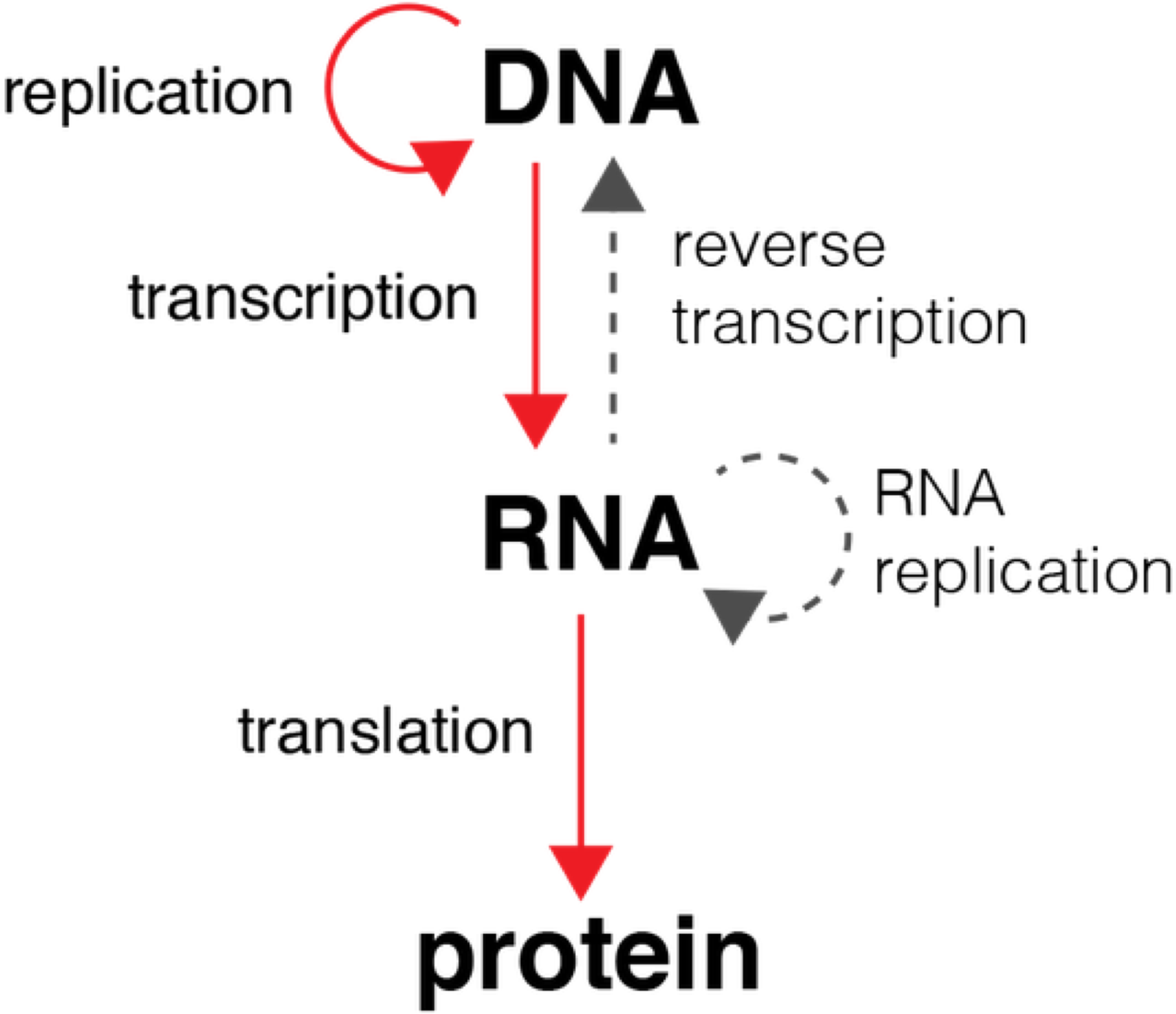
Bacterial genomes have a large circular chromosome made of DNA





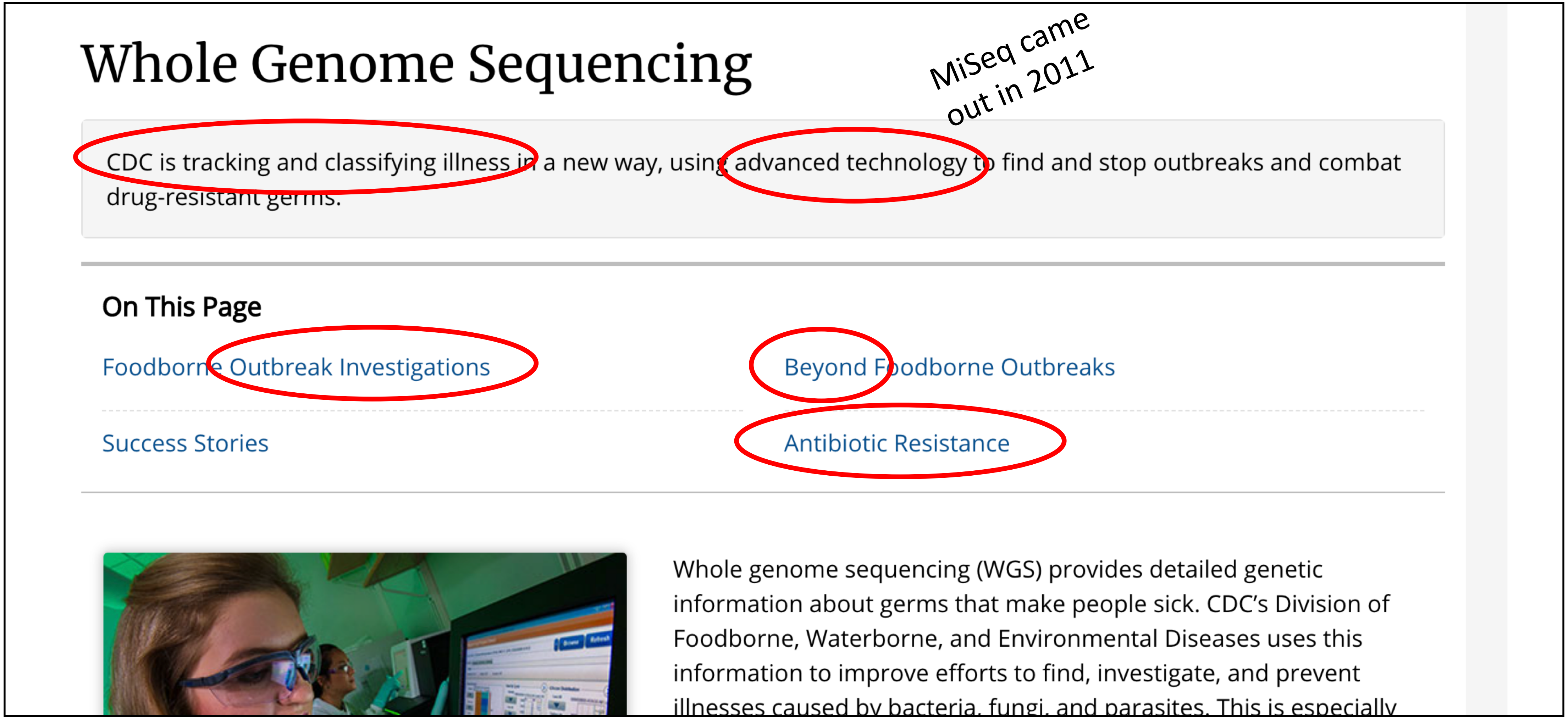
https://www.researchgate.net/publication/312376160 ; https://en.wikipedia.org/wiki/Plasmid

Central Dogma of Biology



Slide Stolen from Kelly Oakeson

We sequence bacterial DNA to improve public health



Whole Genome Sequencing

MiSeq came out in 2011

CDC is tracking and classifying illness in a new way, using advanced technology to find and stop outbreaks and combat drug-resistant germs.

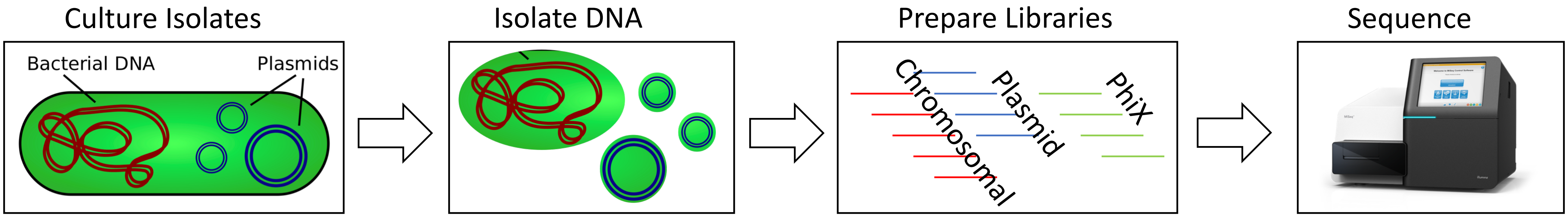
On This Page

- Foodborne Outbreak Investigations
- Beyond Foodborne Outbreaks
- Success Stories
- Antibiotic Resistance

Whole genome sequencing (WGS) provides detailed genetic information about germs that make people sick. CDC's Division of Foodborne, Waterborne, and Environmental Diseases uses this information to improve efforts to find, investigate, and prevent illnesses caused by bacteria, fungi, and parasites. This is especially

https://www.cdc.gov/ncezid/dfwed/keyprograms/tracking-foodborne-illness-wgs.html

Illumina WGS Sequencing

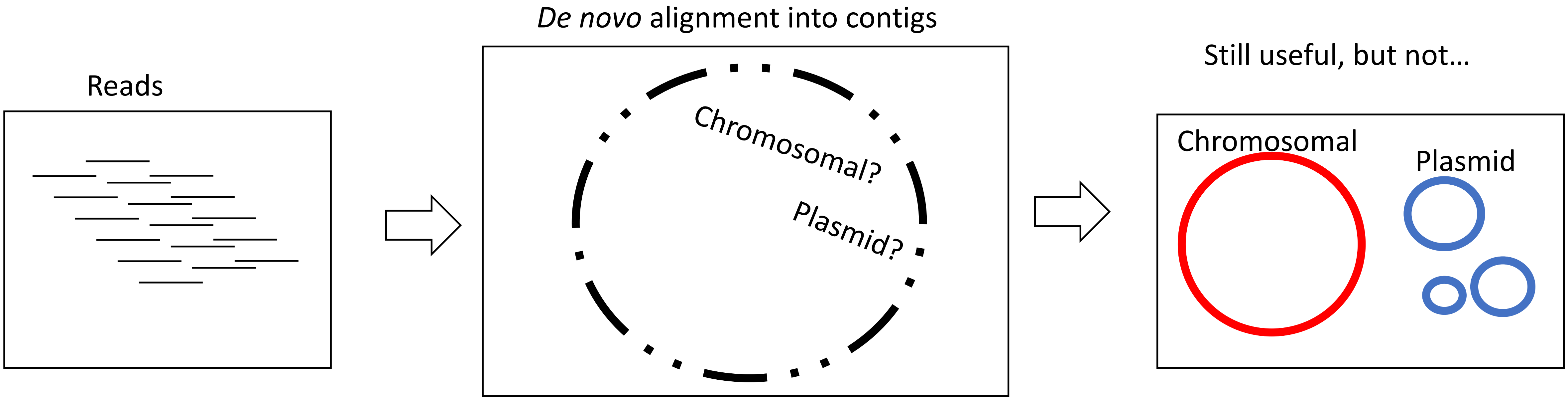


Culture Isolates

Isolate DNA

Prepare Libraries

Sequence



Reads

De novo alignment into contigs

Still useful, but not…

Long-range sequencing is less likely to have issues with troublesome regions

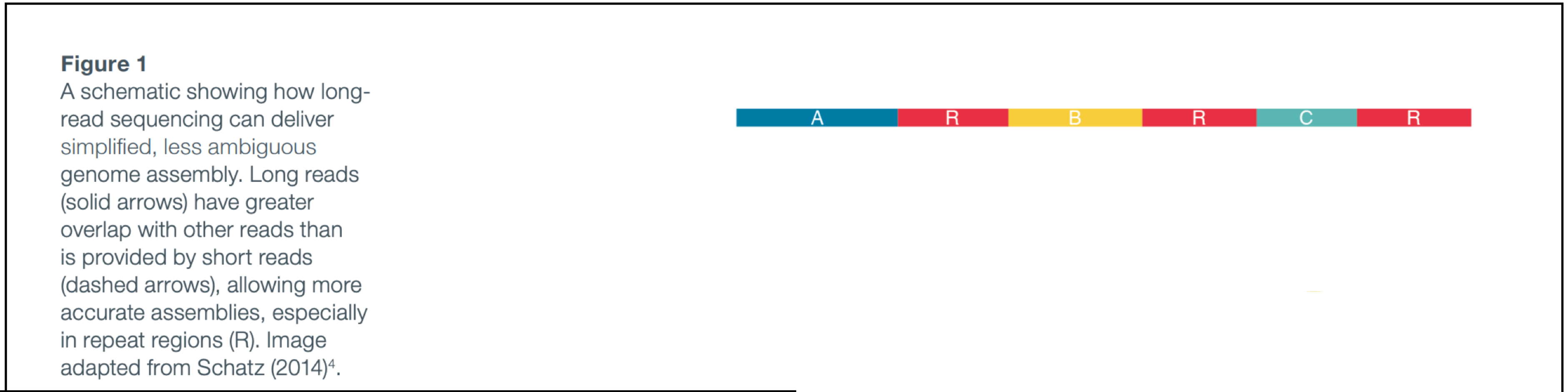


Figure 1
A schematic showing how long-read sequencing can deliver simplified, less ambiguous genome assembly. Long reads (solid arrows) have greater overlap with other reads than is provided by short reads (dashed arrows), allowing more accurate assemblies, especially in repeat regions (R). Image adapted from Schatz (2014)[4].

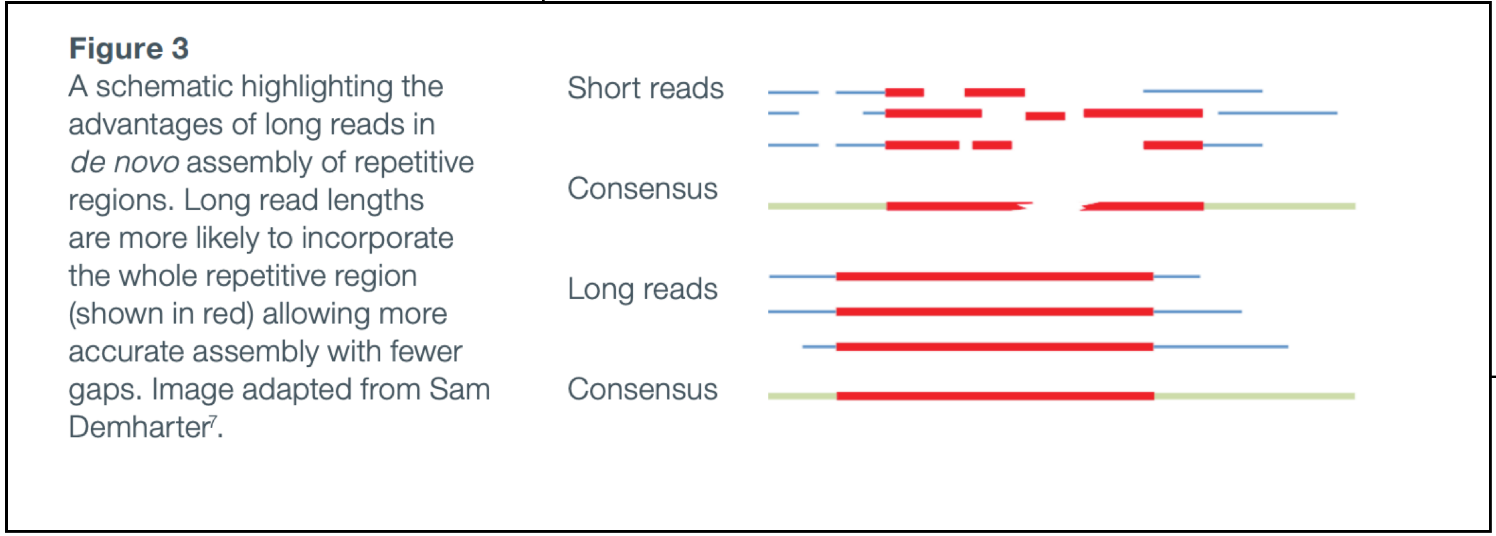
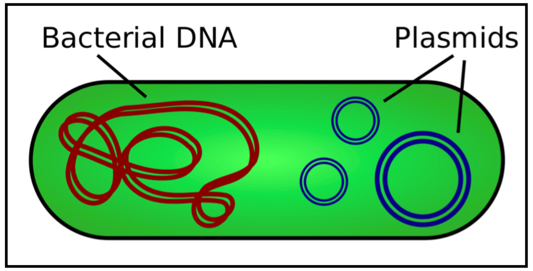


Figure 3
A schematic highlighting the advantages of long reads in *de novo* assembly of repetitive regions. Long read lengths are more likely to incorporate the whole repetitive region (shown in red) allowing more accurate assembly with fewer gaps. Image adapted from Sam Demharter[7].

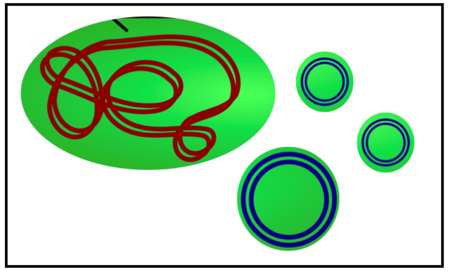
https://nanoporetech.com/sites/default/files/s3/white-papers/WGS_Assembly_white_paper.pdf

Oxford Nanopore WGS Sequencing

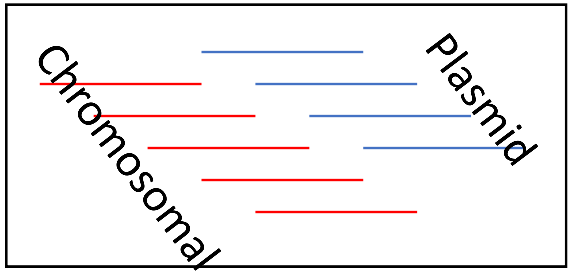
Culture Isolates



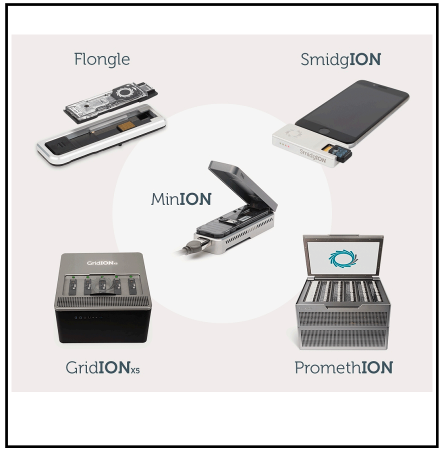
Isolate DNA



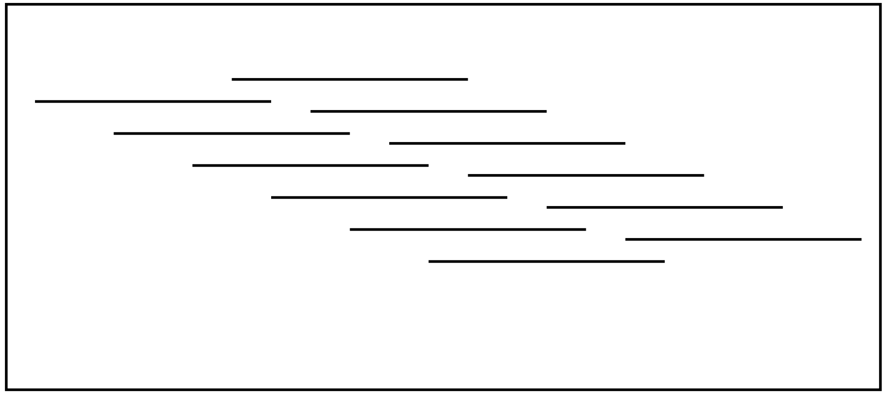
Prepare Libraries



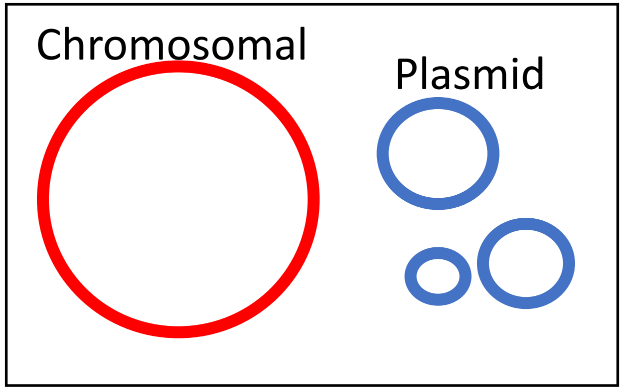
Sequence

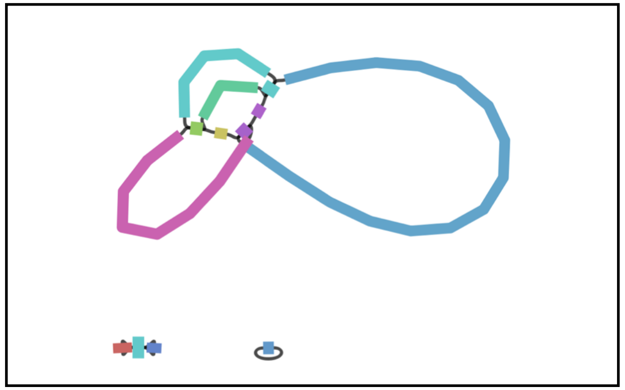


Reads



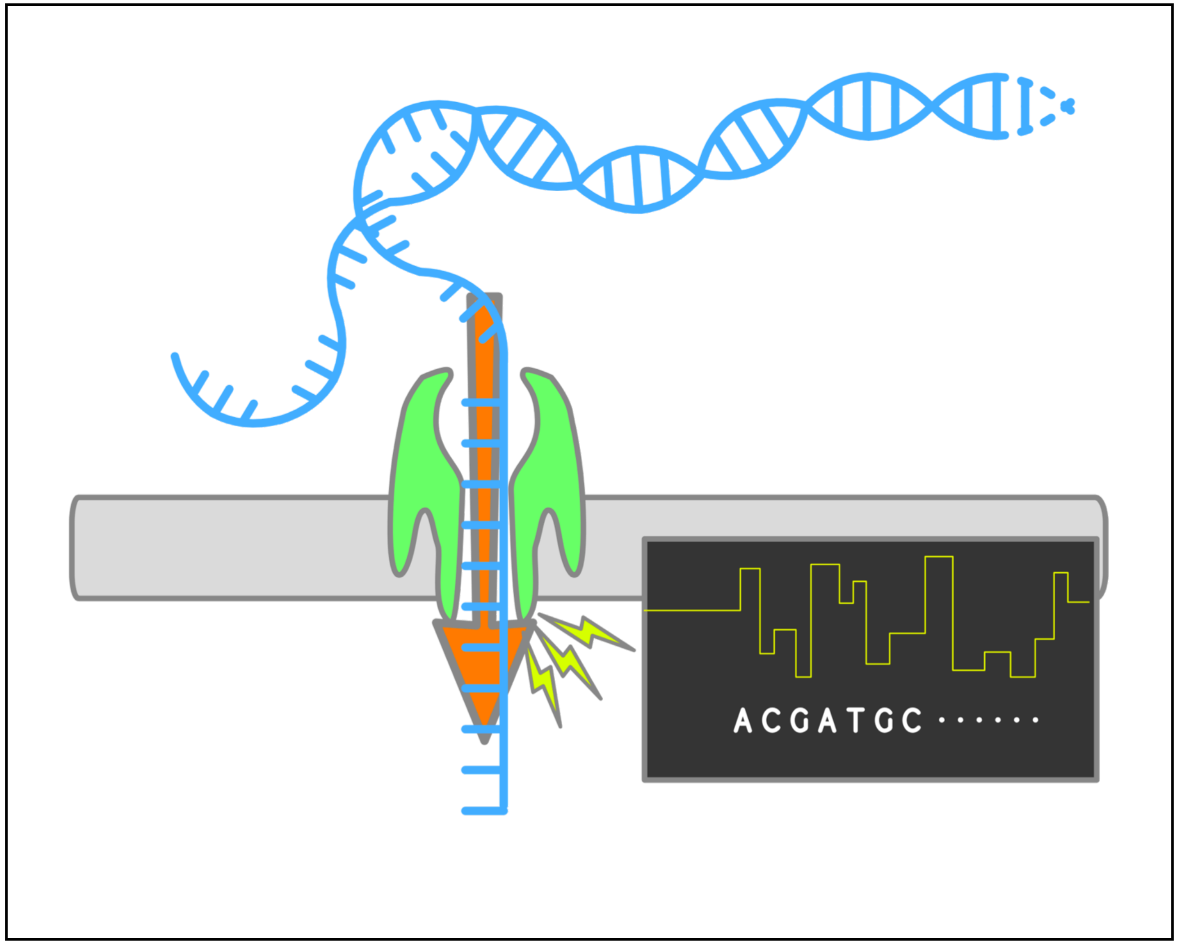
De novo alignment

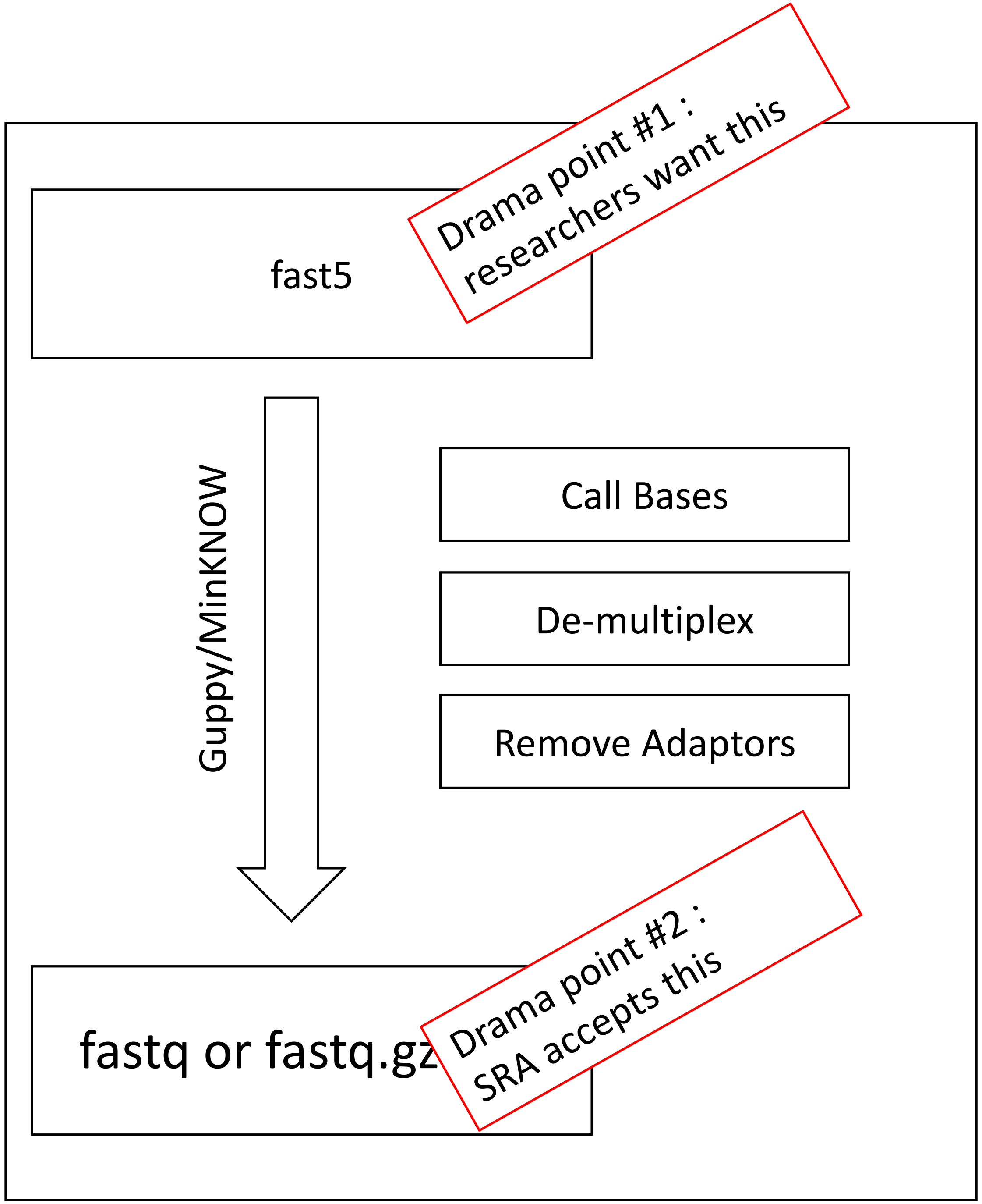




https://www.businesswire.com/news/home/20180319006158/en/Oxford-Nanopore-Announces-%C2%A3100-Million-140M-Fundraising-from-Global-Investors

Nanopore Output





fast5

Drama point #1 : researchers want this

Guppy/MinKNOW

Call Bases

De-multiplex

Remove Adaptors

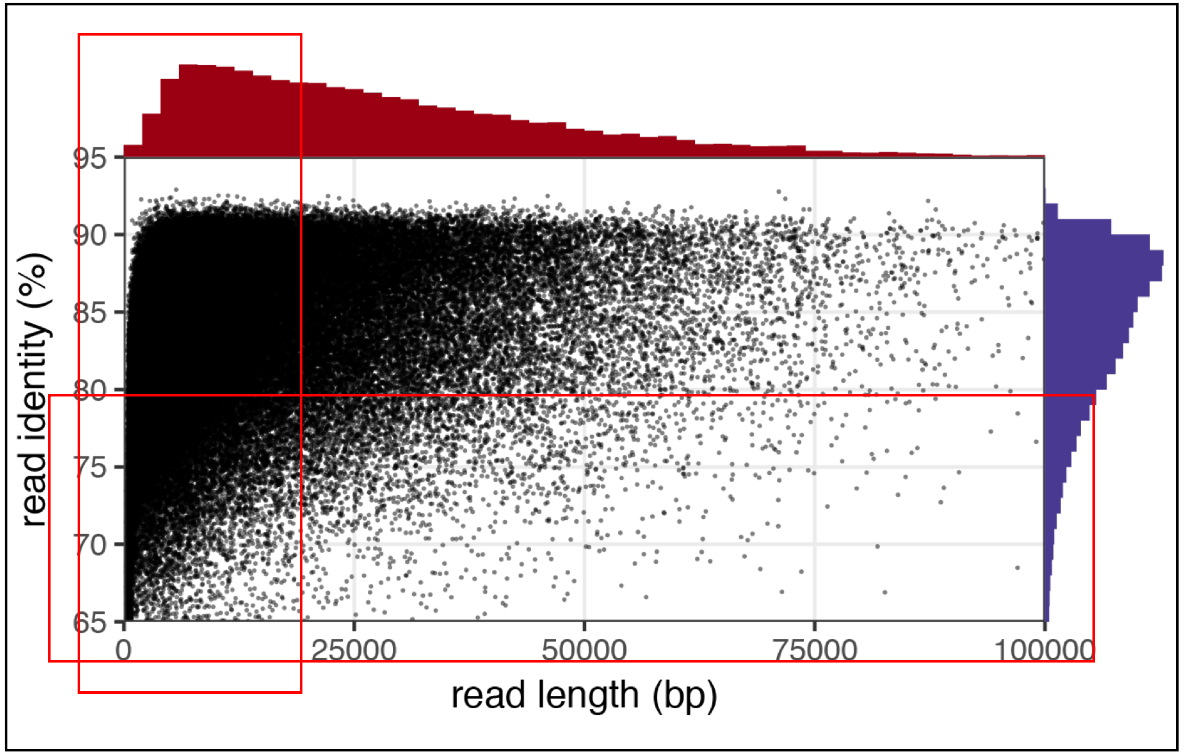
fastq or fastq.gz

Drama point #2 : SRA accepts this

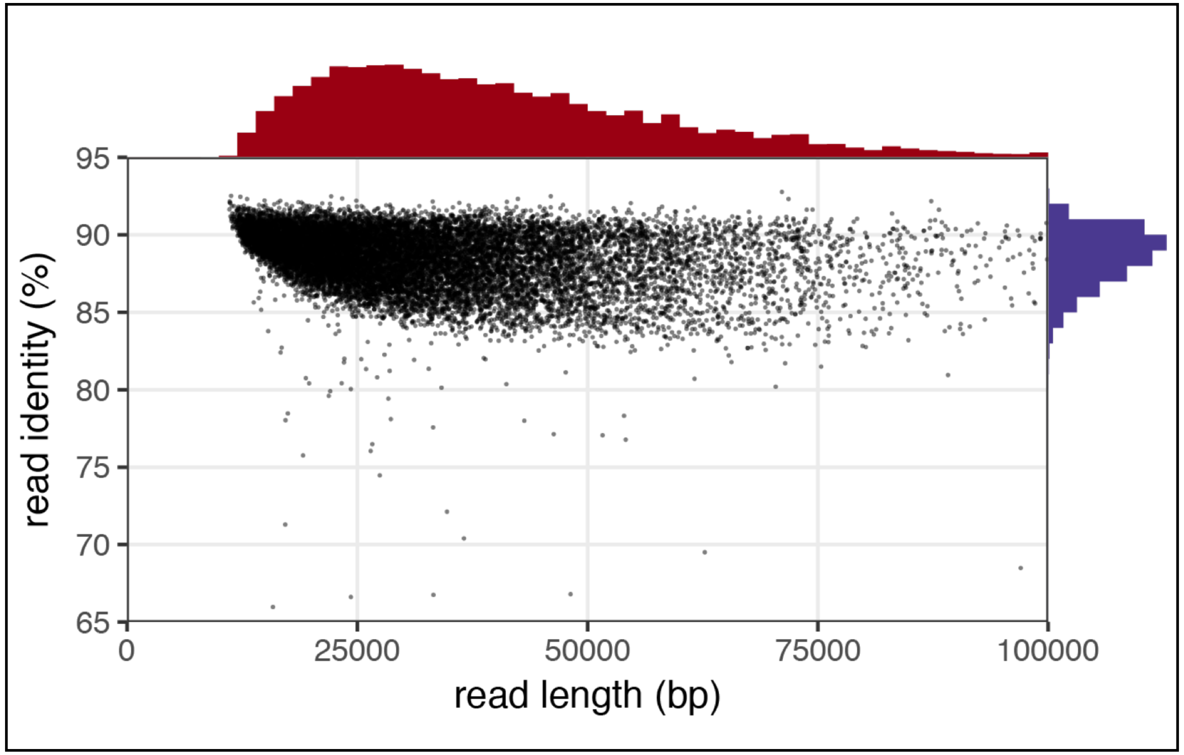
https://en.wikipedia.org/wiki/Nanopore_sequencing

Filtlong

You don’t want all the reads (trust me)



Before

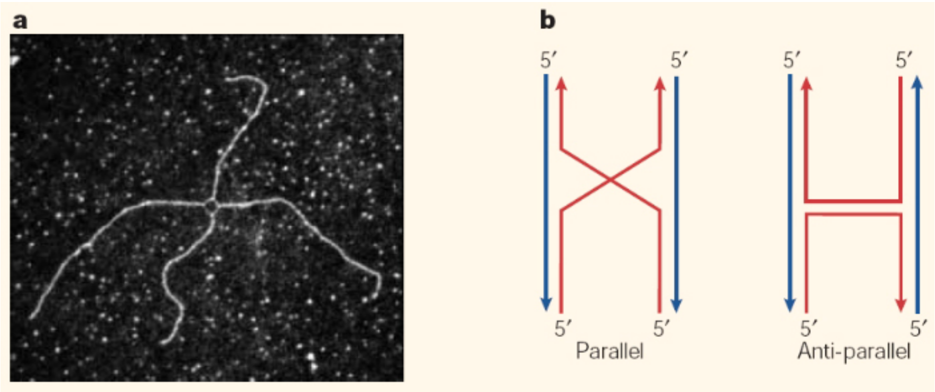


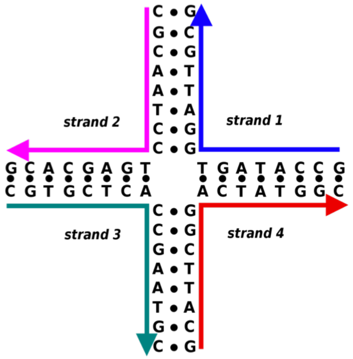
After

https://github.com/rrwick/Filtlong

Filtlong cannot filter out all issues

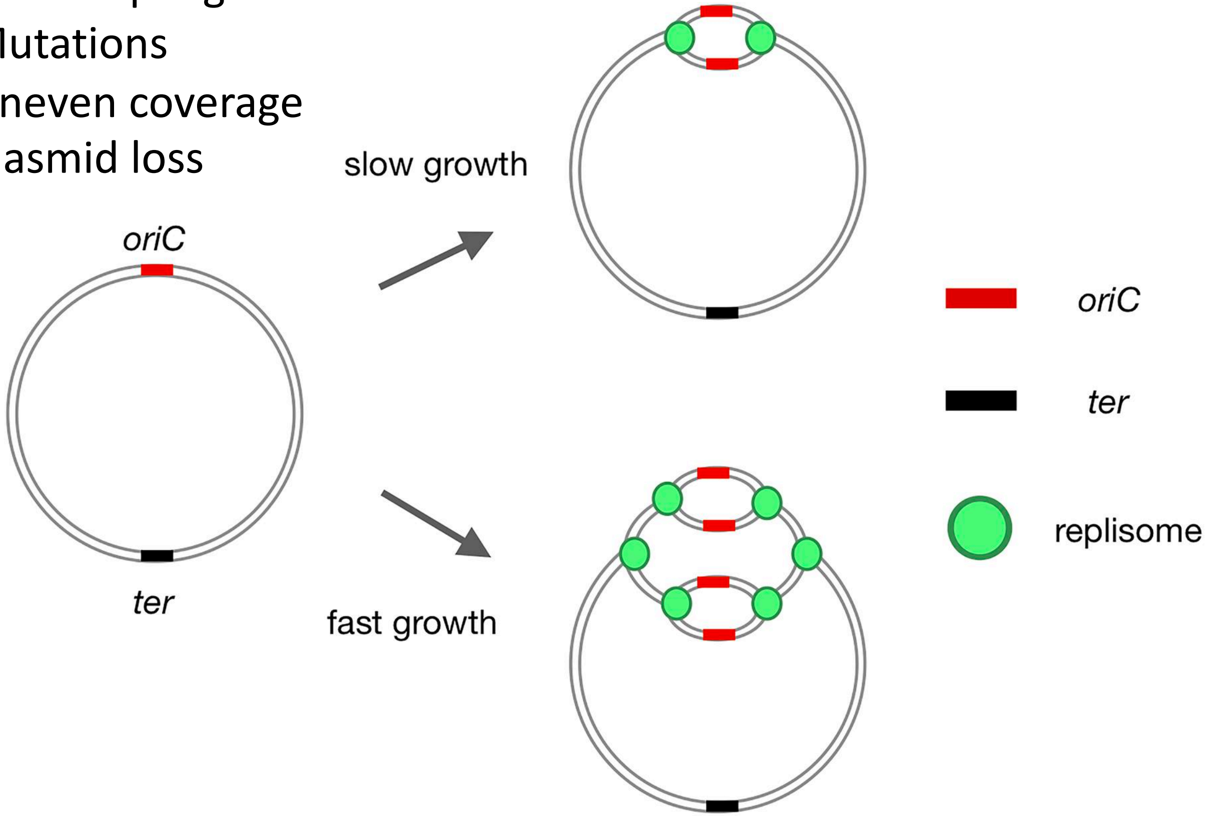
Holliday Junction artefacts





Evolution does not stop in the lab

- Bacteriophage
- Mutations
- Uneven coverage
- Plasmid loss



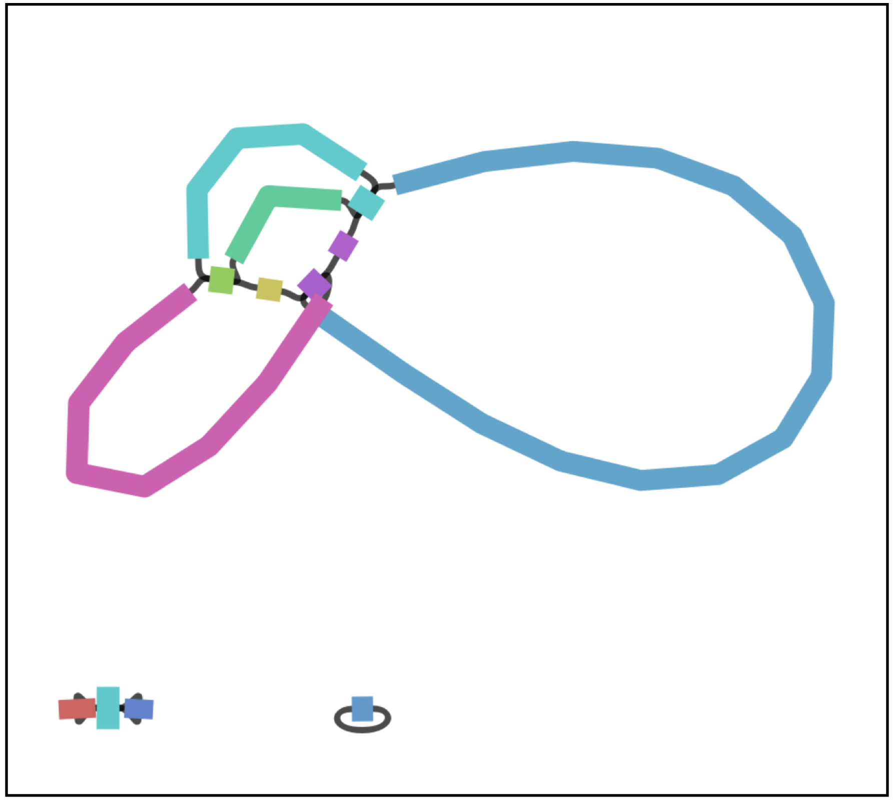
https://www.nature.com/scitable/content/structure-of-the-holliday-junction-24458/ ; https://en.wikipedia.org/wiki/Holliday_junction ; https://doi.org/10.3389/fmicb.2018.02819

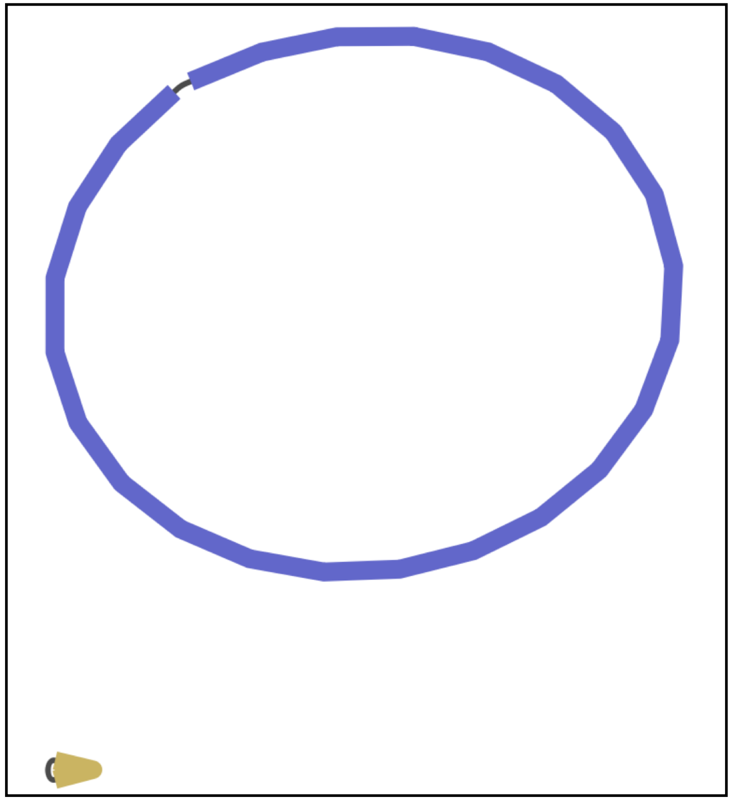
De novo long read assembly

Bandage

https://github.com/rrwick/Bandage

- Flye
- Miniasm/minipolish
- Raven
- Canu/Canu2
- RedBean
- Unicycler (hybrid)
- And more!





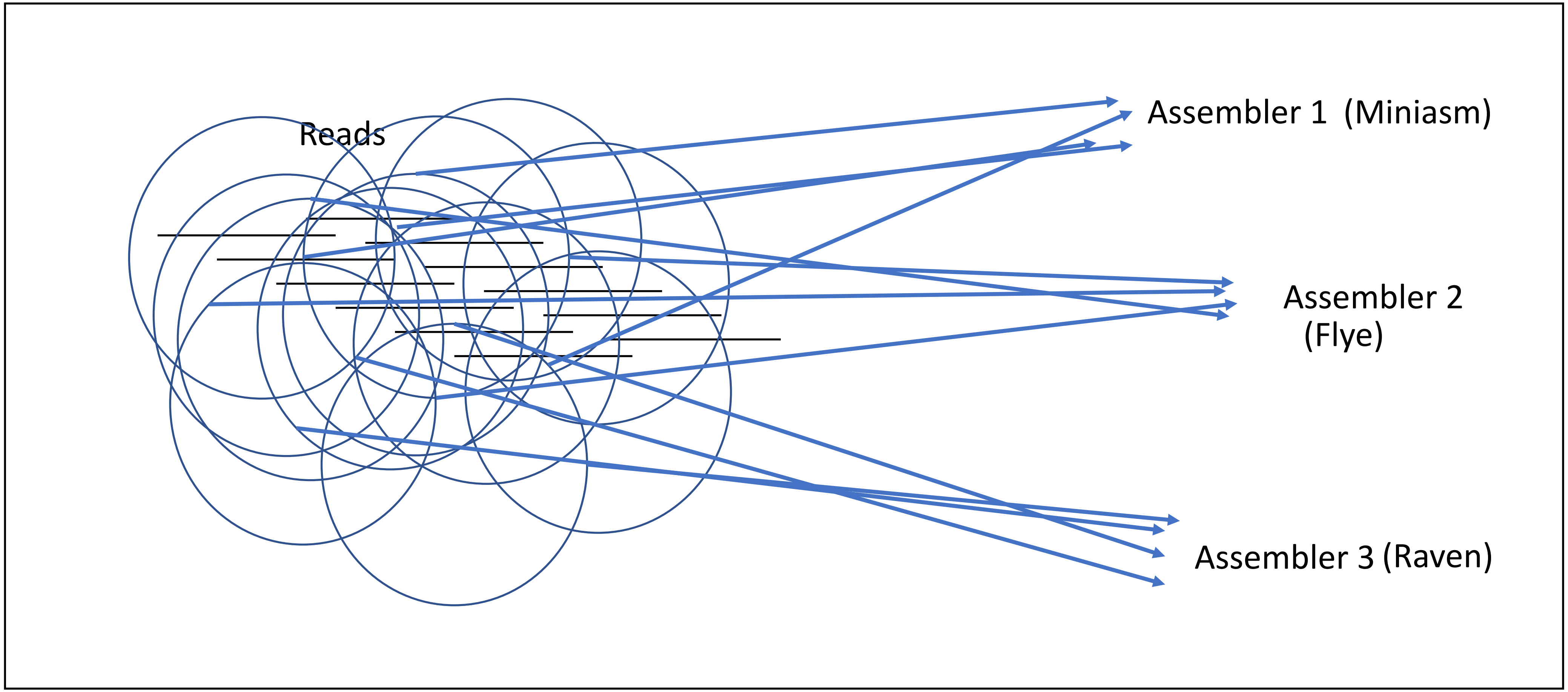
But … which one is BEST?

Trycycler: consensus by agreement

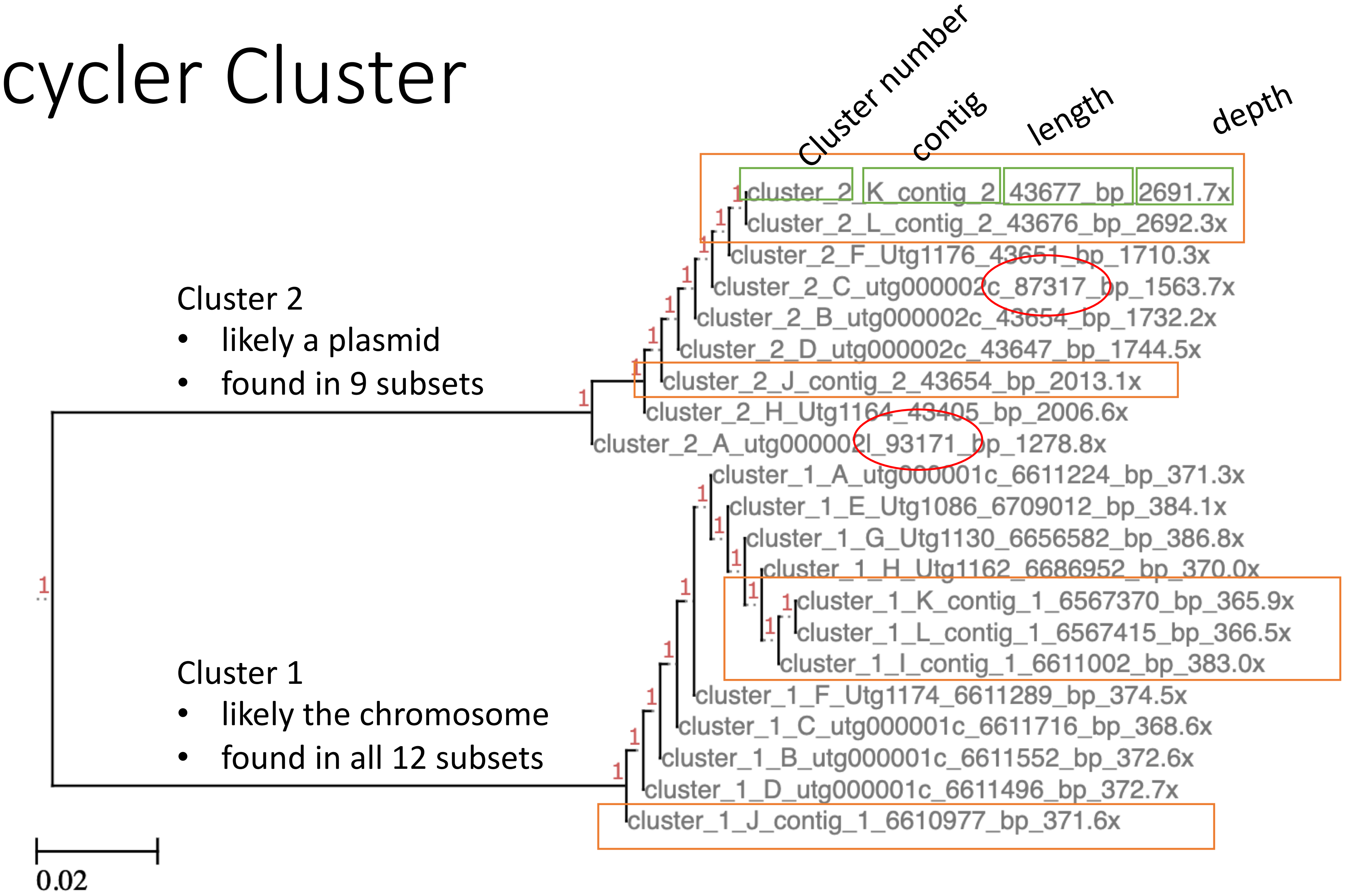
- Create multiple assemblies
- Resolve the differences between them

https://www.biorxiv.org/content/10.1101/2021.07.04.451066v1

Trycycler : SubSample & Assemble



Trycycler Cluster


Cluster number
contig
length
depth
cluster_2_K_contig_2_43677_bp_2691.7x
cluster_2_L_contig_2_43676_bp_2692.3x
cluster_2_F_Utg1176_43651_bp_1710.3x
cluster_2_C_utg000002c_87317_bp_1563.7x
cluster_2_B_utg000002c_43654_bp_1732.2x
cluster_2_D_utg000002c_43647_bp_1744.5x
cluster_2_J_contig_2_43654_bp_2013.1x
cluster_2_H_Utg1164_43405_bp_2006.6x
cluster_2_A_utg000002l_93171_bp_1278.8x
cluster_1_A_utg000001c_6611224_bp_371.3x
cluster_1_E_Utg1086_6709012_bp_384.1x
cluster_1_G_Utg1130_6656582_bp_386.8x
cluster_1_H_Utg1162_6686952_bp_370.0x
cluster_1_K_contig_1_6567370_bp_365.9x
cluster_1_L_contig_1_6567415_bp_366.5x
cluster_1_I_contig_1_6611002_bp_383.0x
cluster_1_F_Utg1174_6611289_bp_374.5x
cluster_1_C_utg000001c_6611716_bp_368.6x
cluster_1_B_utg000001c_6611552_bp_372.6x
cluster_1_D_utg000001c_6611496_bp_372.7x
cluster_1_J_contig_1_6610977_bp_371.6x
1
0.02

Cluster 2

- likely a plasmid
- found in 9 subsets

Cluster 1

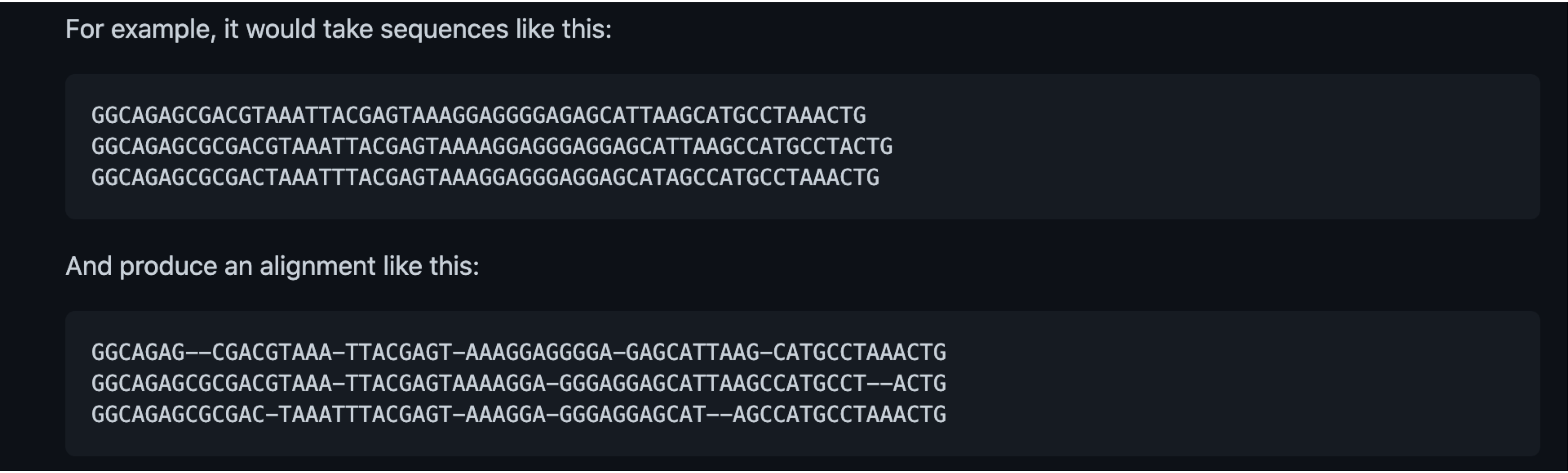
- likely the chromosome
- found in all 12 subsets

In theory, all contigs in a cluster will have similar lengths and depth

Trycycler reconcile

- All contigs in a cluster should have
  - Similar depth
  - Similar length
  - Similar sequence
- The end user must remove contigs that are not similar “enough”
- Reconcile
  - Ensure sequences on the same strand
  - Fix circularization
  - Rotate to common start

Trycycler MSA aligns sequences in a cluster



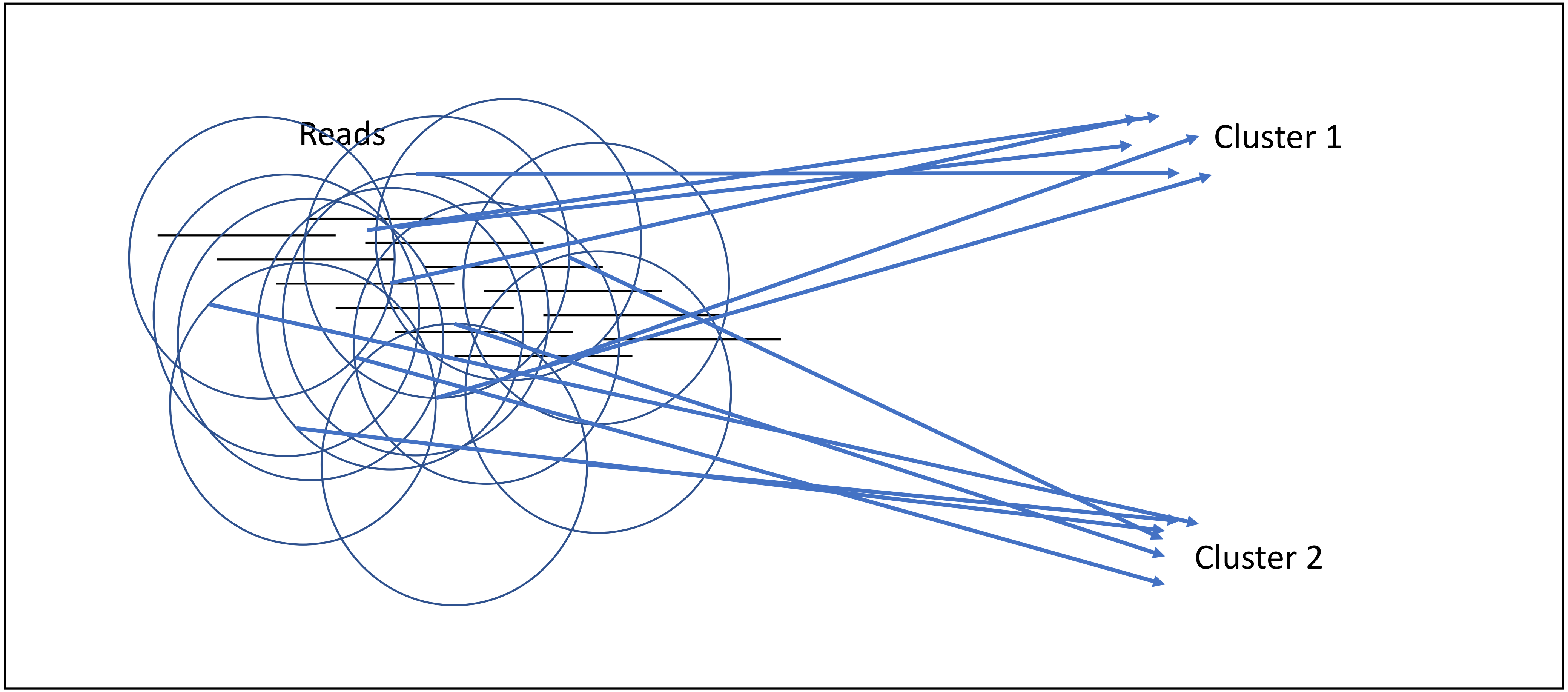
For example, it would take sequences like this:

```
GGCAGAGCGACGTAAATTACGAGTAAAGGAGGGGAGAGCATTAAGCATGCCTAAACTG
GGCAGAGCGCGACGTAAATTACGAGTAAAAGGAGGGAGGAGCATTAAGCCATGCCTACTG
GGCAGAGCGCGACTAAATTTACGAGTAAAGGAGGGAGGAGCATAGCCATGCCTAAACTG
```

And produce an alignment like this:

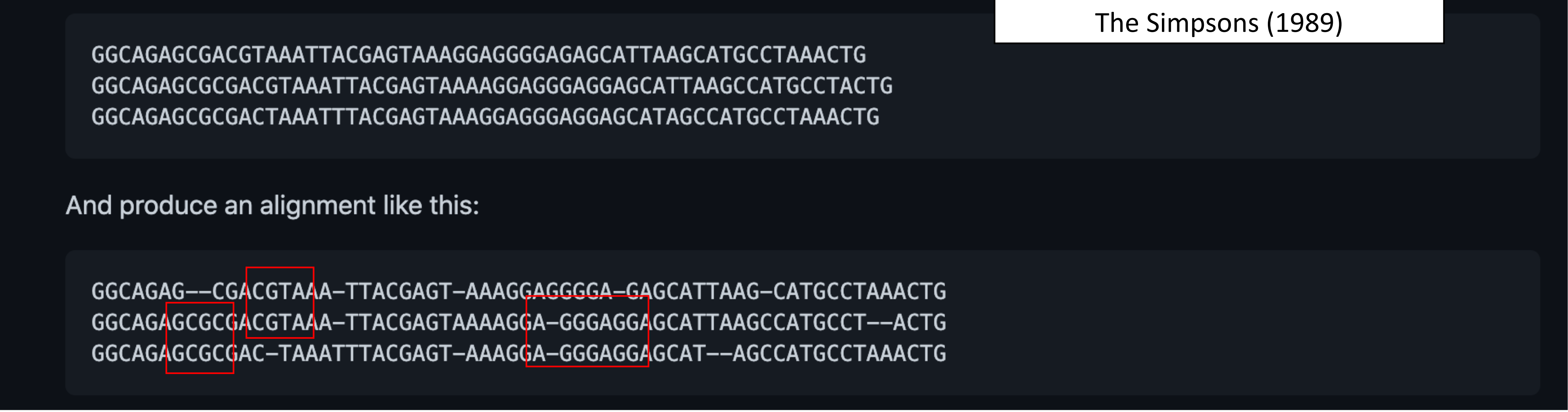
```
GGCAGAG--CGACGTAAA-TTACGAGT-AAAGGAGGGGA-GAGCATTAAG-CATGCCTAAACTG
GGCAGAGCGCGACGTAAA-TTACGAGTAAAAGGA-GGGAGGAGCATTAAGCCATGCCT--ACTG
GGCAGAGCGCGAC-TAAATTTACGAGT-AAAGGA-GGGAGGAGCAT--AGCCATGCCTAAACTG
```

Trycycler partition assigns reads to cluster



Trycycler consensus

The Simpsons (1989)



For example, it would take sequences like this:

```
GGCAGAGCGACGTAAATTACGAGTAAAGGAGGGGAGAGCATTAAGCATGCCTAAACTG
GGCAGAGCGCGACGTAAATTACGAGTAAAAGGAGGGAGGAGCATTAAGCCATGCCTACTG
GGCAGAGCGCGACTAAATTTACGAGTAAAGGAGGGAGGAGCATAGCCATGCCTAAACTG
```

And produce an alignment like this:

```
GGCAGAG--CGACGTAAA-TTACGAGT-AAAGGAGGGGA-GAGCATTAAG-CATGCCTAAACTG
GGCAGAGCGCGACGTAAA-TTACGAGTAAAAGGA-GGGAGGAGCATTAAGCCATGCCT--ACTG
GGCAGAGCGCGAC-TAAATTTACGAGT-AAAGGA-GGGAGGAGCAT--AGCCATGCCTAAACTG
```

Polishing : Because we are not done, yet

- Polishing is using prior reads to “correct” errors in the final assembly
    - Nanopolish : polishes raw ONT reads
    - Medaka : polishes assembly with ONT reads
    - Racon : polishes assembly with Illumina or ONT reads
    - Pilon : polishes assembly with Illumina reads
- Many assemblers include a polishing step
- Over-polishing is a thing

Donut falls : A Trycycler Nextflow Workflow

https://github.com/UPHL-BioNGS/Donut_Falls

Once guppy has called bases, removed adapters, and demultiplexed

- Create a sample key that links barcode and sample_id and Illumina fastq files
- Run phase 1 :

```
nextflow run Donut_Falls.nf -c configs/singularity.config
```

- Examine tree (at http://etetoolkit.org/treeview/)
- Remove problematic clusters
- Run Phase 2:

```
nextflow run Donut_Falls.nf -c configs/phase2_singularity.config
```

- Examine clusters with Bandage
- Find AMR genes, submit to repositories, etc.

https://www.theoutbound.com/utah/hiking/quick-hike-to-donut-falls/photos#photo-116242

Questions?

APHL

ARLABnetwork

AMD TRAINING LEAD
and BIOINFORMATICS
REGIONAL RESOURCE

State Public Health
StaPH-B
Bioinformatics

SEQUENCE ALL THE THINGS

UTAH PUBLIC HEALTH
LABORATORY

Erin Young, PhD

: https://twitter.com/ErinYoun

: https://github.com/erinyoung