Assembly

Tung Trinh

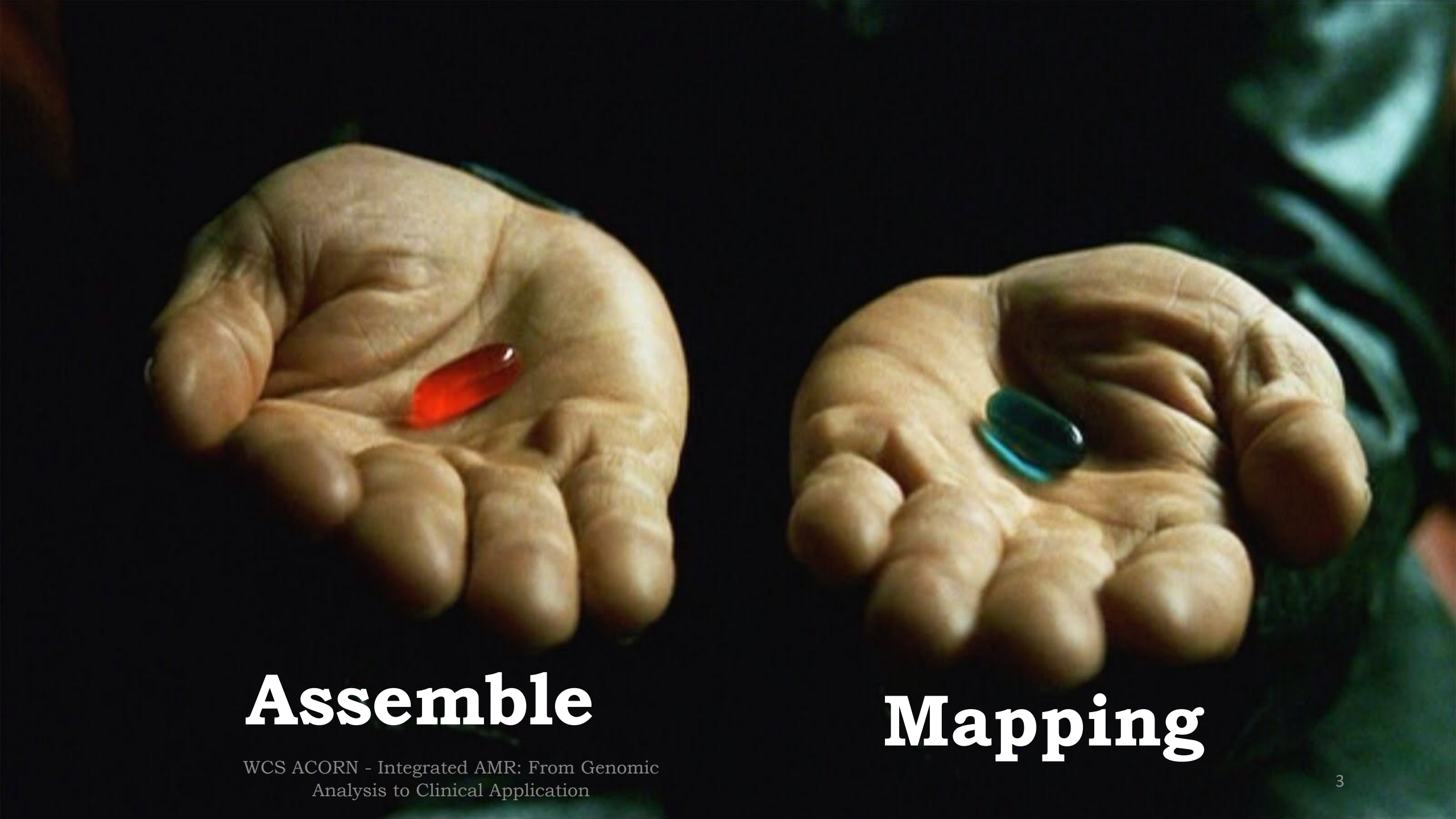
Bioinformatician – Oxford University Clinical Research

16/09/2024

The first analysis step

The first step of downstream analysis is either

- *De novo* assembly
  - Reconstruct the original sequences from raw reads alone
  - Like a jigsaw puzzle but ambiguous
- Read mapping (align to the reference)
  - Find where reads fits on a known sequence
  - Can not always uniquely placed


Assemble
Mapping

When and why we should do *de novo* assembly

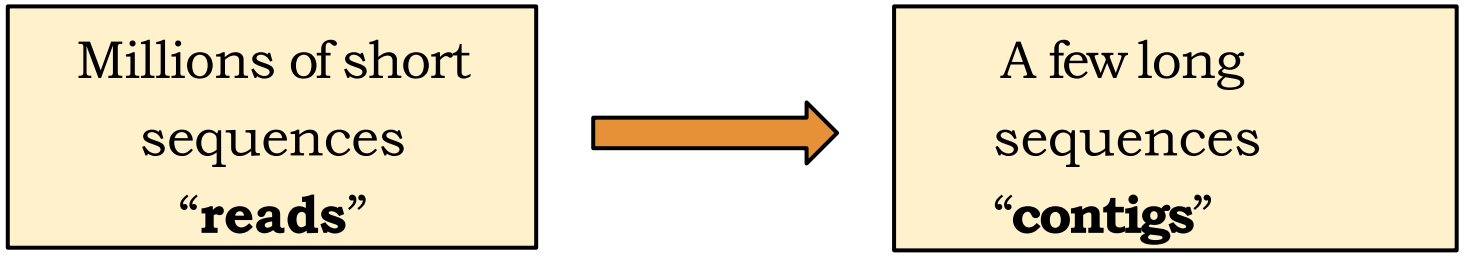
De novo : expression of smt “from the very beginning”

When is *de novo* assembly required?

- new "non-model" organisms (when we still called SARS-CoV-2 as nCoV)
  - no sufficiently related reference genome
- novel DNA segments (plasmid, horizontal gene transferring)
- novel RNA transcripts and splice variants
- discover fusion genes
- identify contamination

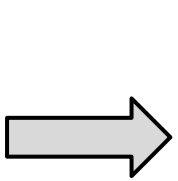
Curial step for bacterial genomic analysis !!!

De novo assembly



Ideally, one sequence per replicon.

De novo assembly



- Sequencing a population of cells
- PCR amplification steps

Another example is this

not. Look

s not. Look at

size, do you?

Size matt

e. Jud

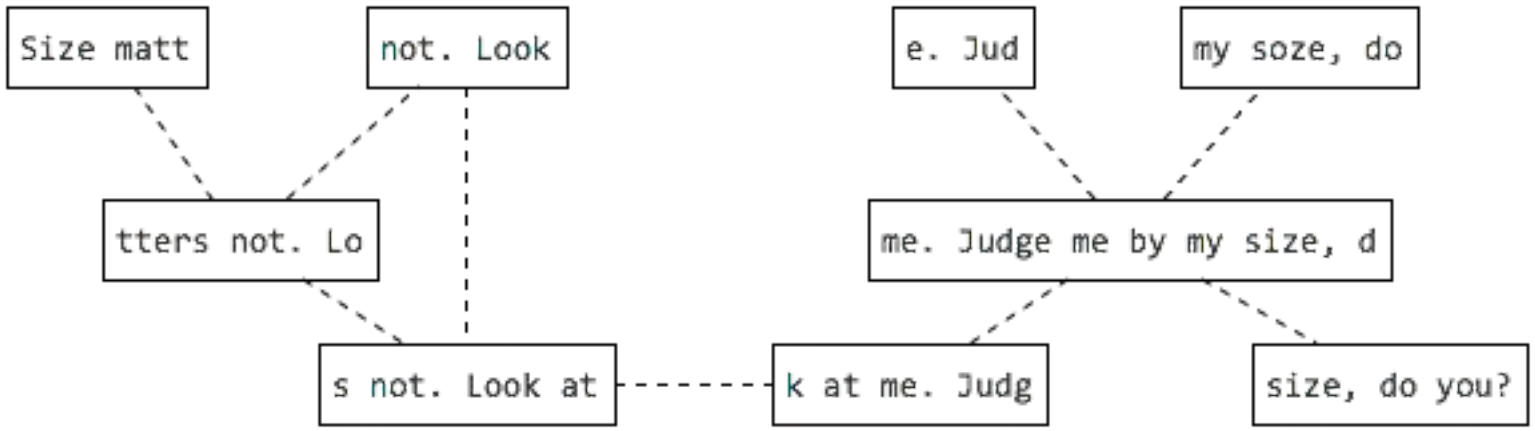
my soze, do

tters not. Lo

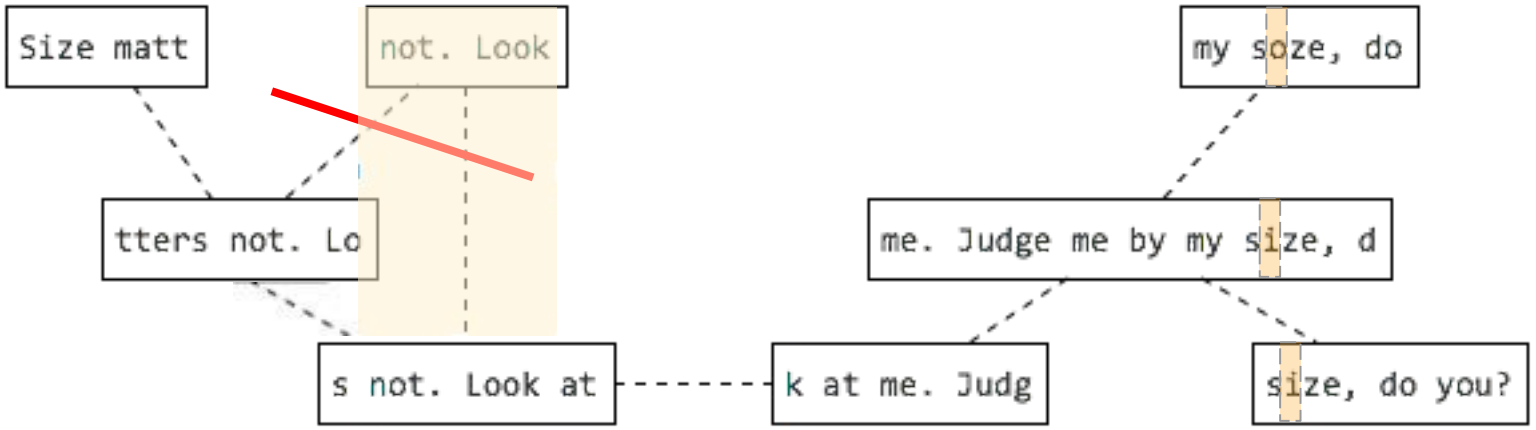
k at me. Judg

me. Judge me by my size, d

Overlaps find

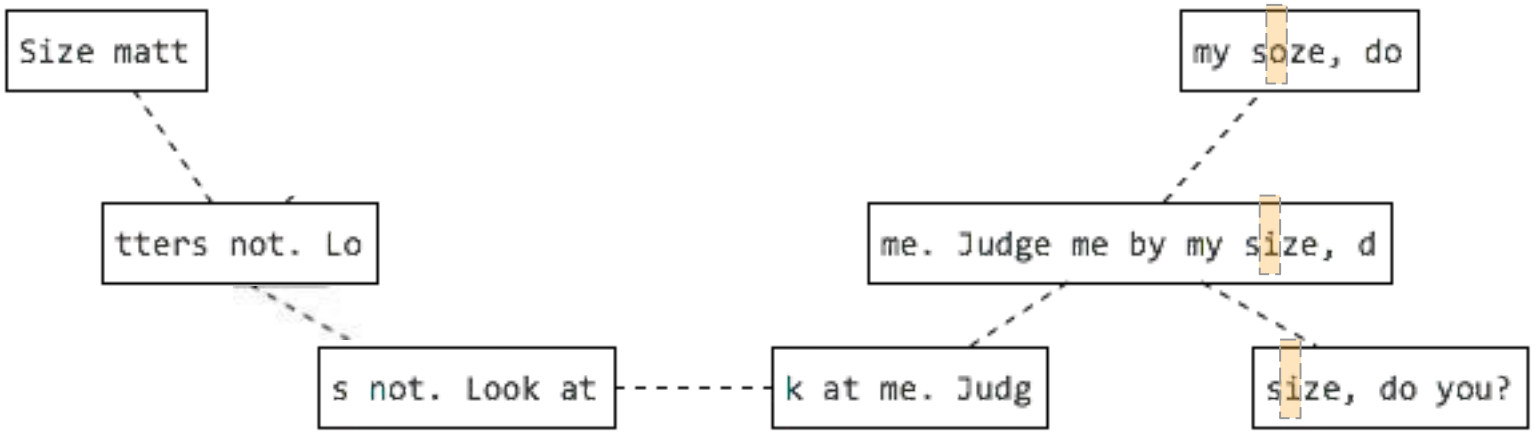


The graph one can simplify



“not, look” is fully contained within the other read, and can be removed.

Do the graph traverse



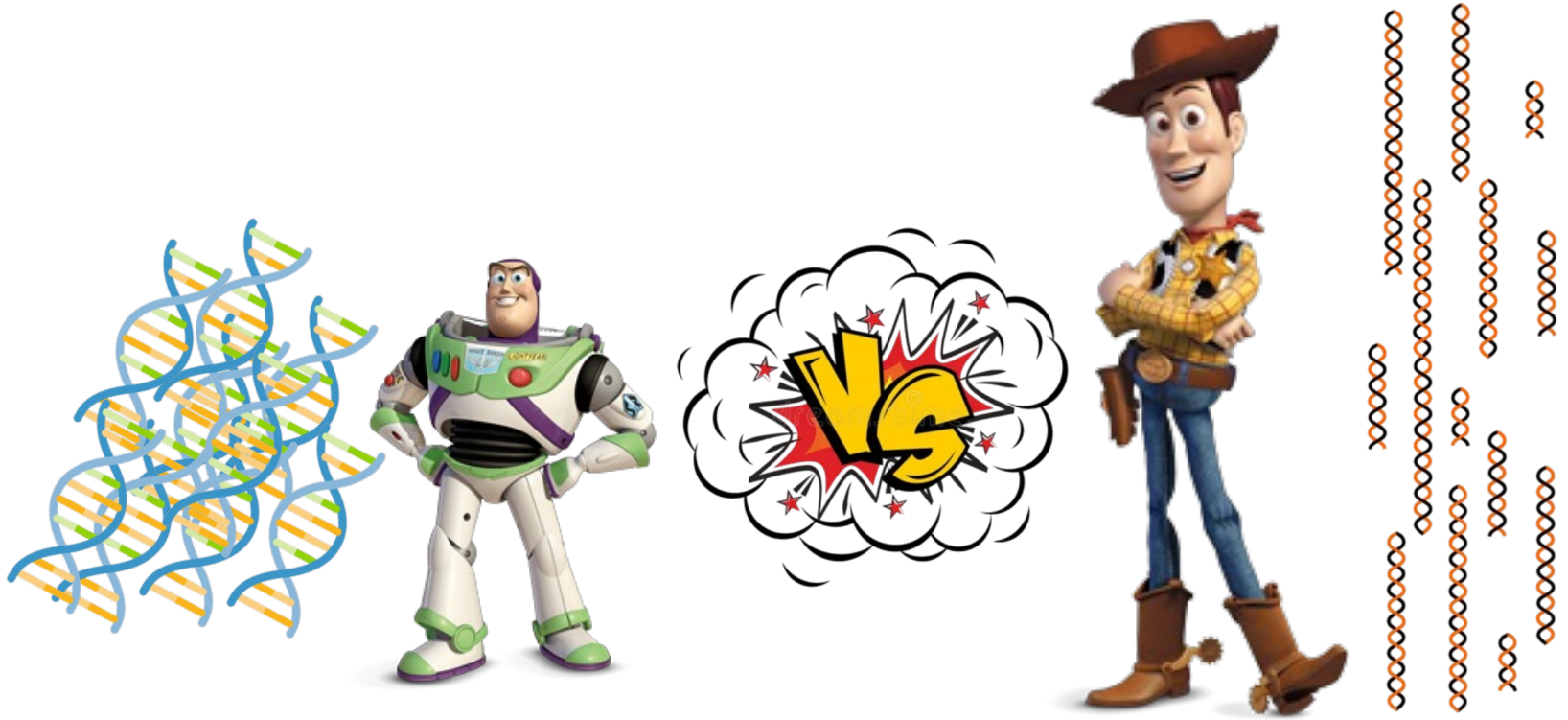
Size matters not. Look at me. Judge me by my size, do you? ← 2 supporting reads

Size matters not. Look at me. Judge me by my s*o*ze, do you? ← 1 supporting read

So far, so good.


ONE DOES NOT SIMPLY
ASSEMBLE A GENOME

Short-read vs long-read assembly



Short read

Long read

Due to the nature of the output raw reads, the algorithms of assembly for short and long read are not the same !!

What makes a jigsaw puzzle hard?

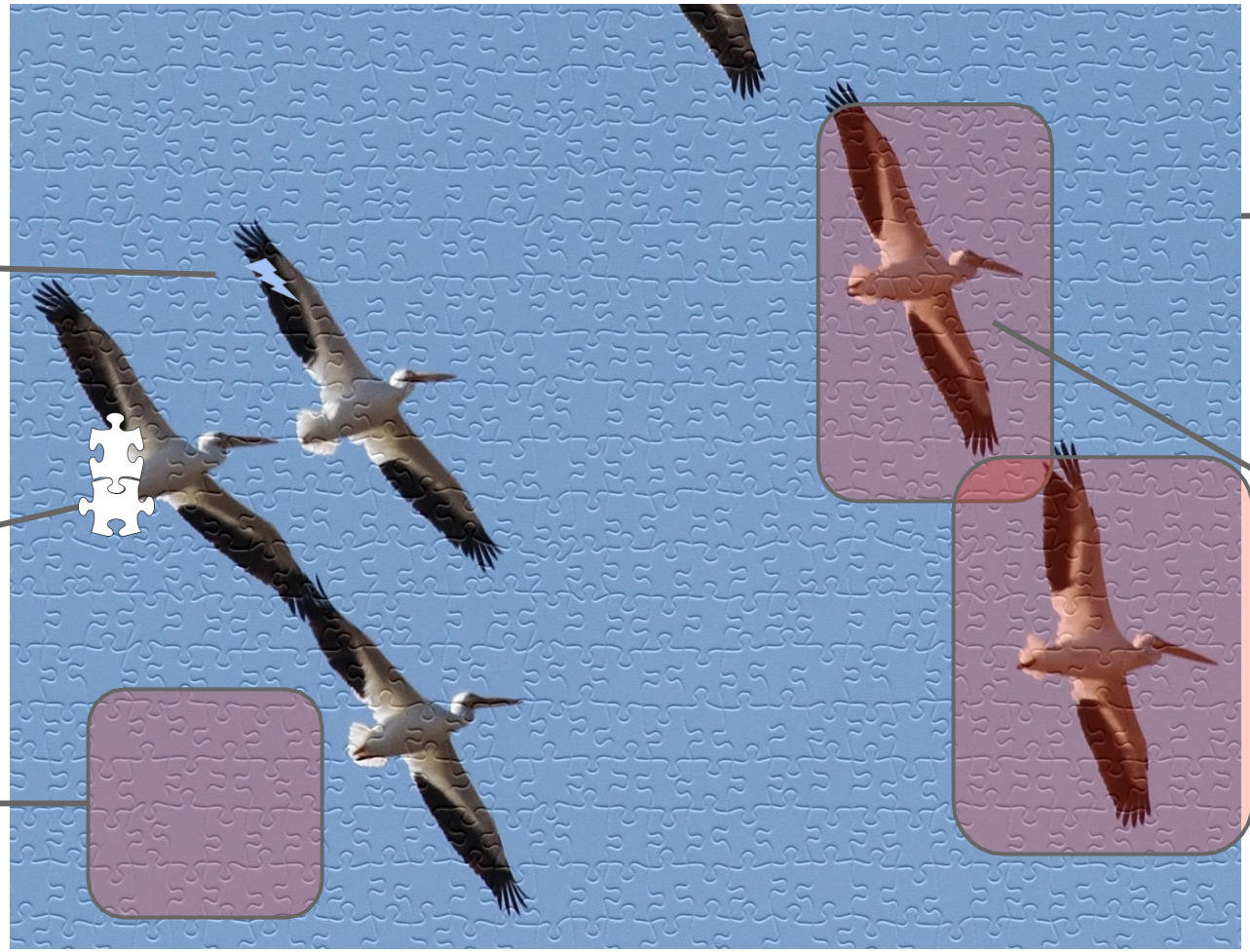
And so is Assembly …

Short read *Long read*
No box
(don't know what we have)

Long read
Frayed pieces
Short read

Missing pieces
Short read *Long read*

Repetitive regions
Short read



Short read
Lots of pieces
Long read

Short read *Long read*
Dirty pieces
Long read *Long read*

Short read
Multiple copies
Long read

Short read *Long read*
No corners
(circular genomes, don't know where is the start nor the end)

Short read assembly algorithms

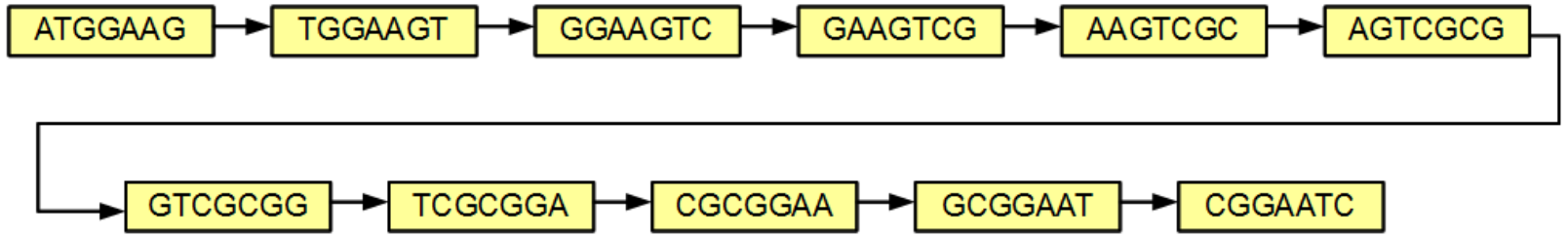
Read Length < Repeat Length

sequence **ATGGAAGTCGCGGAATC**

7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

de Bruijn graph

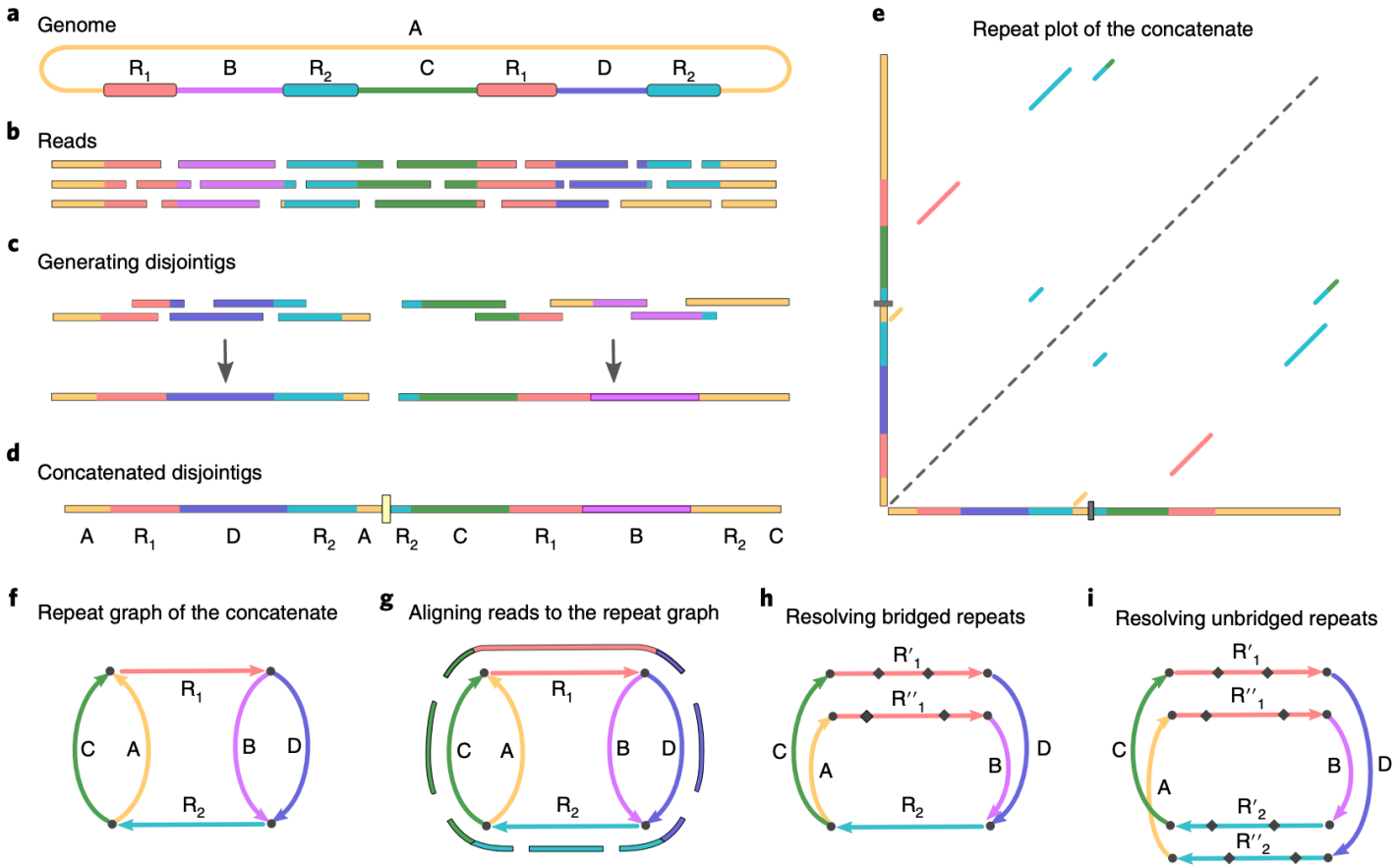


Different K-mer may result different assembly

Long read assembly algorithms

Read Length > Repeat Length

Flye
Repeat graphs

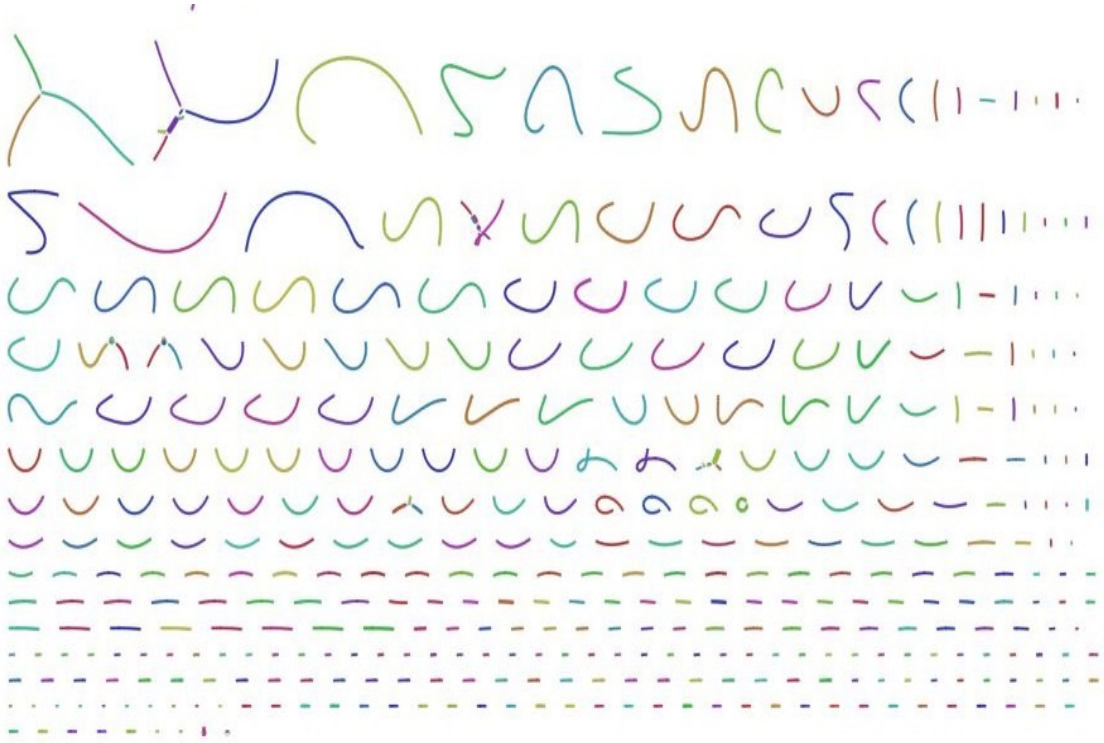


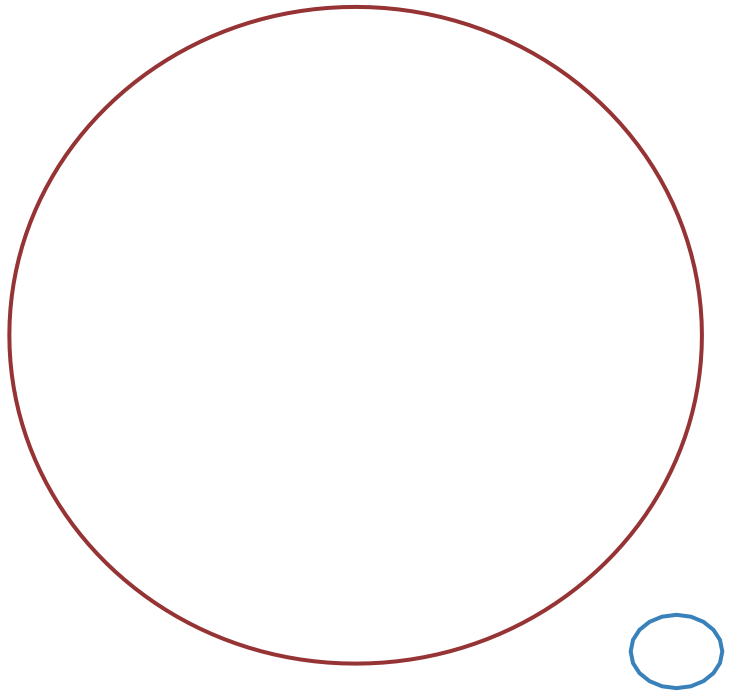
https://doi.org/10.1038/s41587-019-0072-8

Why we have to polish after assembly for long read ?

- As we sequence as much as the input DNA, no extra synthesis reads create to increase the quality of the read
- The longer the reads are, the lower the quality is
- We **can not** discard the low quality read as we will lost ~60-80% of the reads
- For ONT, we have to basecall the raw read from the current signal and error can occur during this step

Draft vs Finished genomes (bacteria)





150 bp - Illumina - $200

10,000 bp - Pacbio - $2000

Assessing the contig

3 “**C**” criteria

- **Contiguity: N50 statistic**

  how many contigs we get ?

- **Completeness: total size**

  how long are the contigs ?

$$\frac{Assembled\ Genome\ Size}{Estimated\ Genome\ Size}$$

- **Correctness**

  how correct are our contigs ?