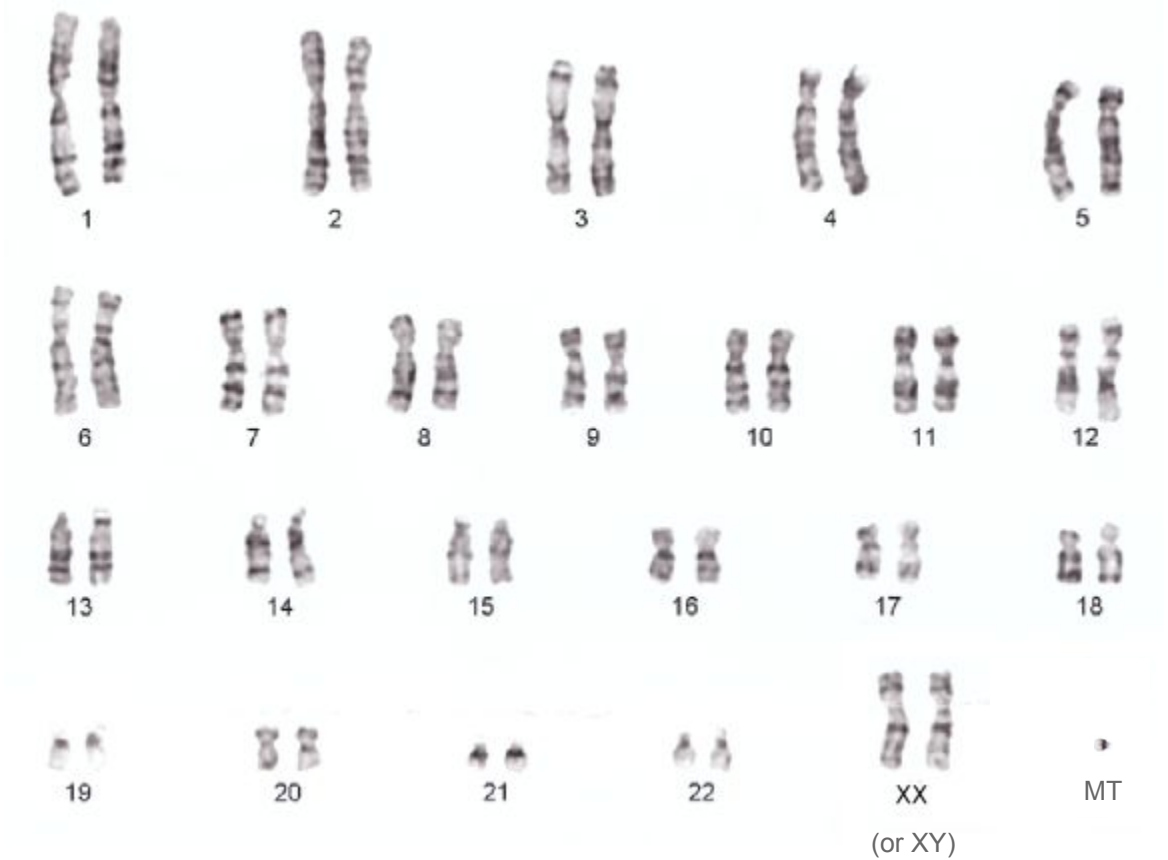
De novo Genome Assembly

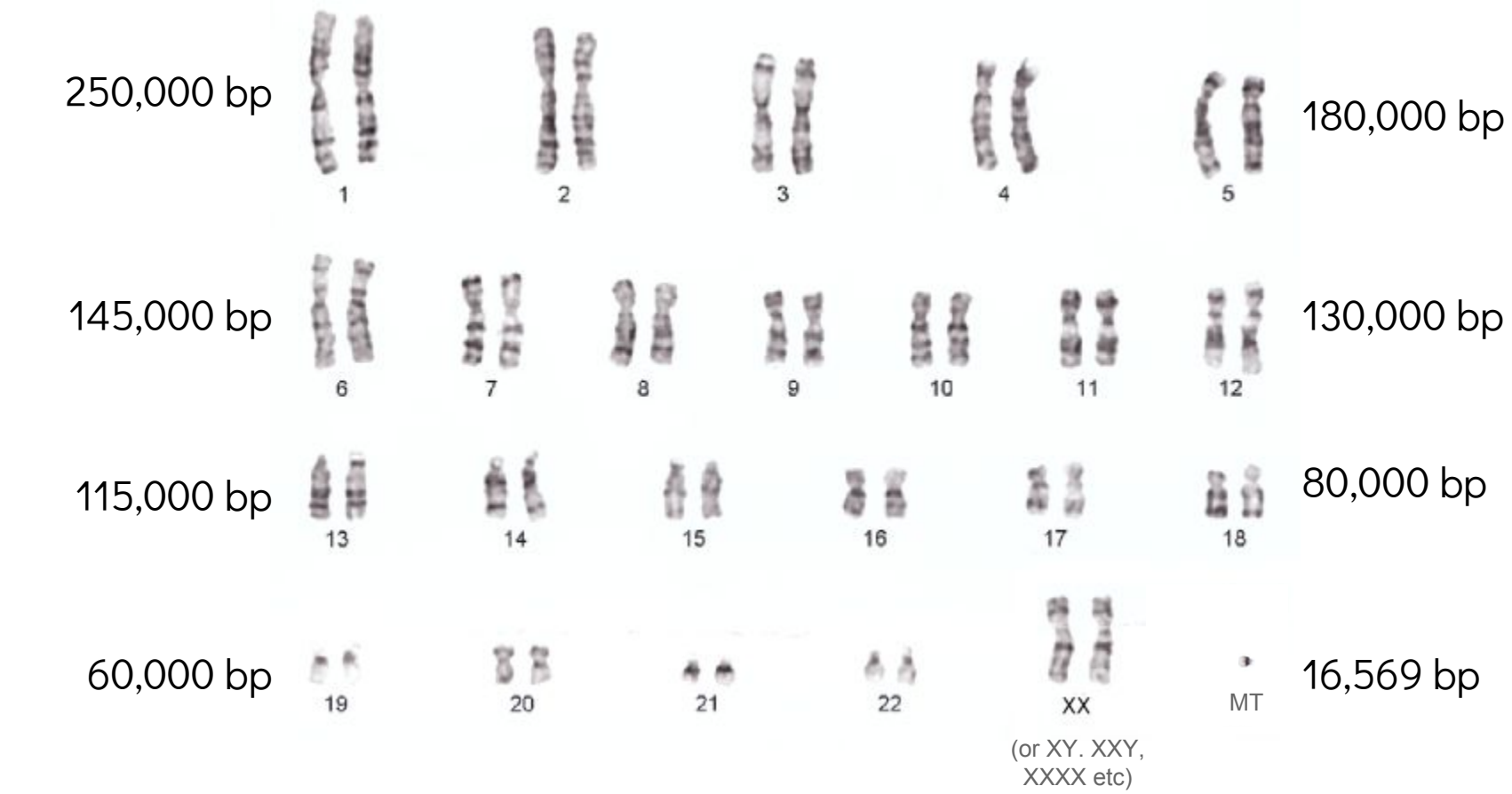
A/Prof Torsten Seemann

Introduction

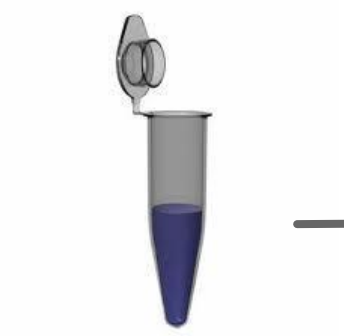
The human genome has 47 pieces



We want the DNA sequence of all 47 pieces



In an ideal world ...



```
AGTCTAGGATTCGCTATAG
ATTCAGGCTCTGATATATT
TCGCGGCATTAGCTAGAGA
TCTCGAGATTCGTCCCAGT
CTAGGATTCGCTAT
AAGTCTAAGATTC...
```

Human DNA

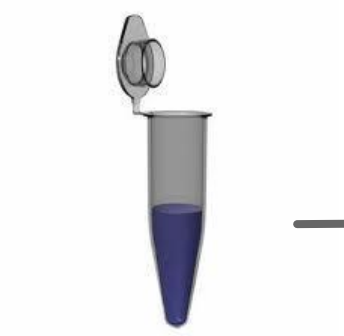
iSequencer ™

46 chromosomal seqs

+

1 mitochondrial seq

Sooner than we think ?



```
AGTCTAGGATTCGCTATAG
ATTCAGGCTCTGATATATT
TCGCGGCATTAGCTAGAGA
TCTCGAGATTCGTCCCAGT
CTAGGATTCGCTAT
AAGTCTAAGATTC...
```

Human DNA

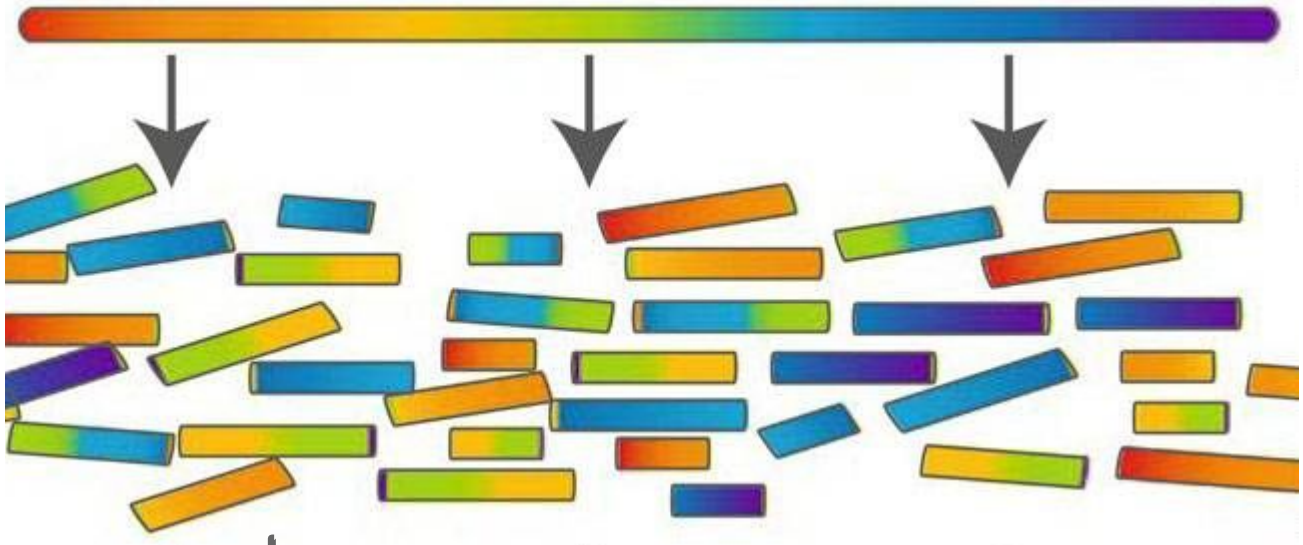
ONT SmidgION ™

46 chromosomal seqs

+

1 mitochondrial seq

The real world (for now)



Genome

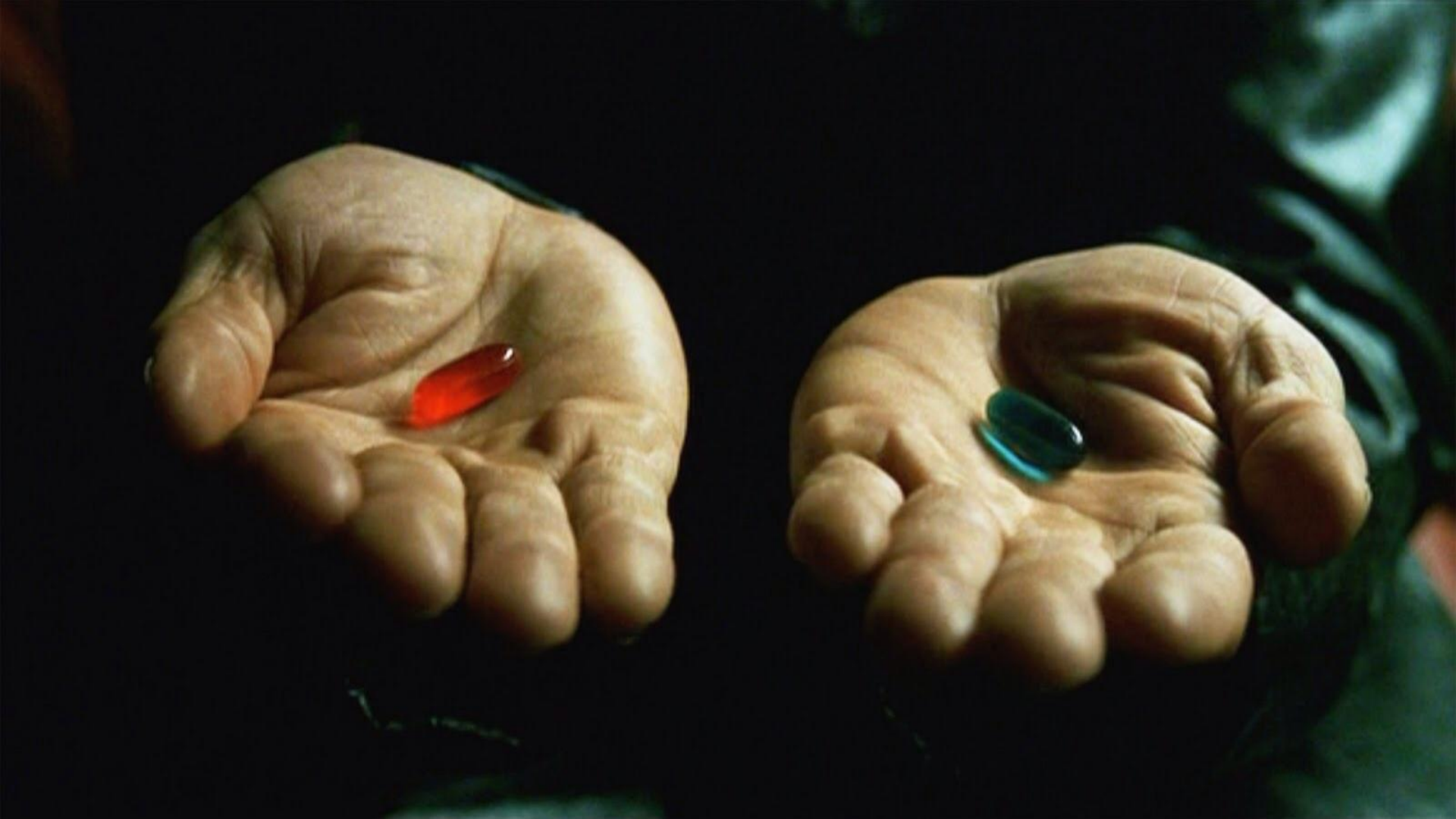
Short fragments

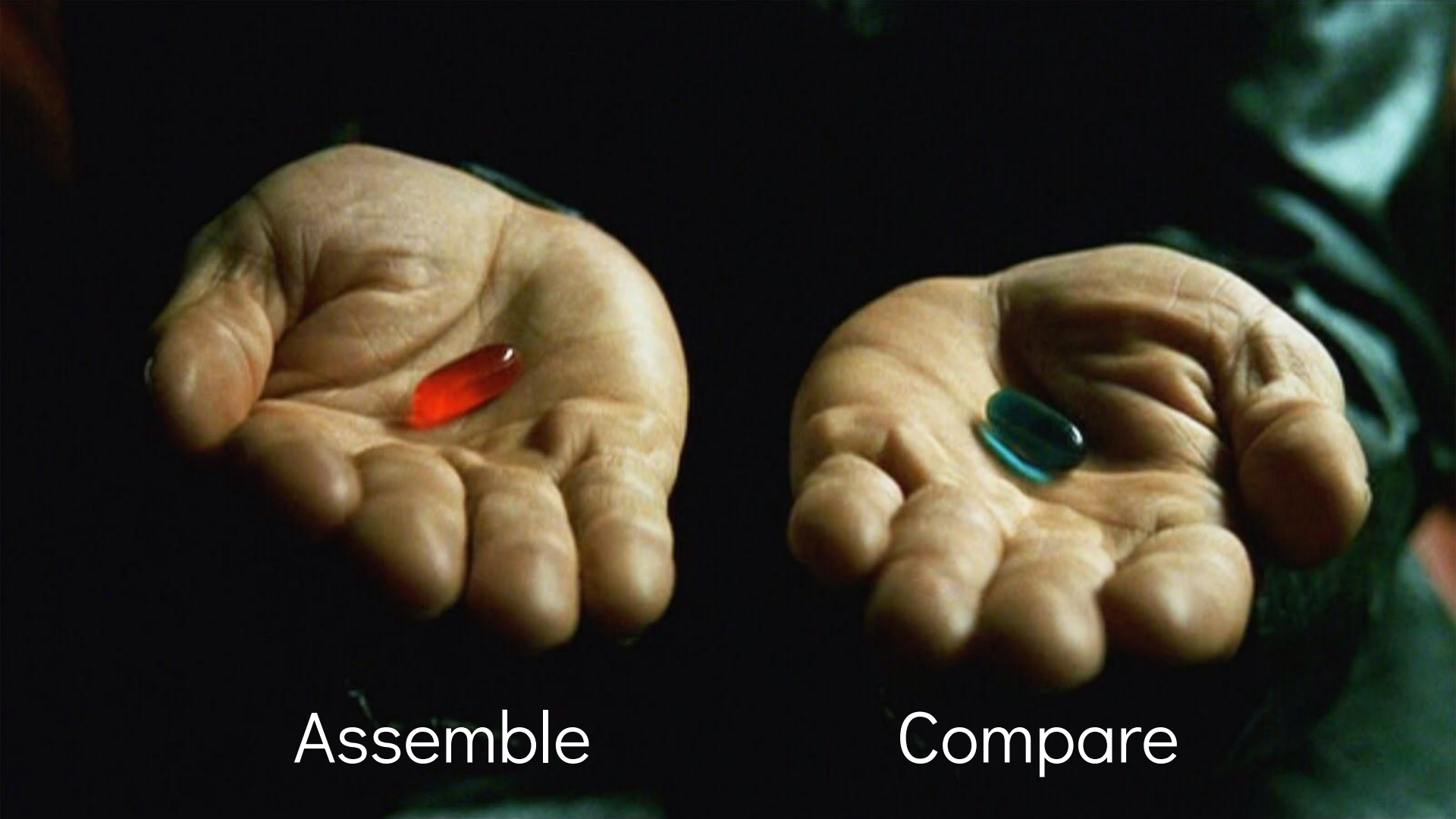
Sequencing

Reads are stored in “FASTQ” files

Read lengths

illumina®	50 - 300 bp
ion torrent	100 - 400 bp
PACIFIC BIOSCIENCES®	5,000 - 40,000+ bp
Oxford NANOPORE Technologies	1,000 - 1,000,000+ bp



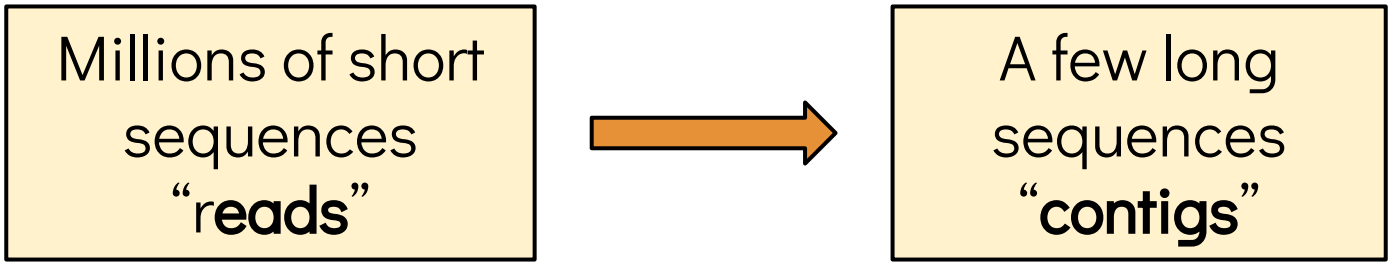

Assemble
Compare

Genome assembly

(the red pill)

De novo genome assembly

Reconstruct the original genome from the sequence reads only



Ideally, one sequence per replicon.

De novo genome assembly


THE AUSTRALIAN

TRADING PLACES

HENRY ERGAS

Risks rise in market downturn

Stan 'deserves' to be a grand slam man

Rio details blueprint for indigenous jobs

Ex-spy chief stands by SBS mission

"From scratch"

De novo genome assembly


THE AUSTRALIAN

TRADING PLACES

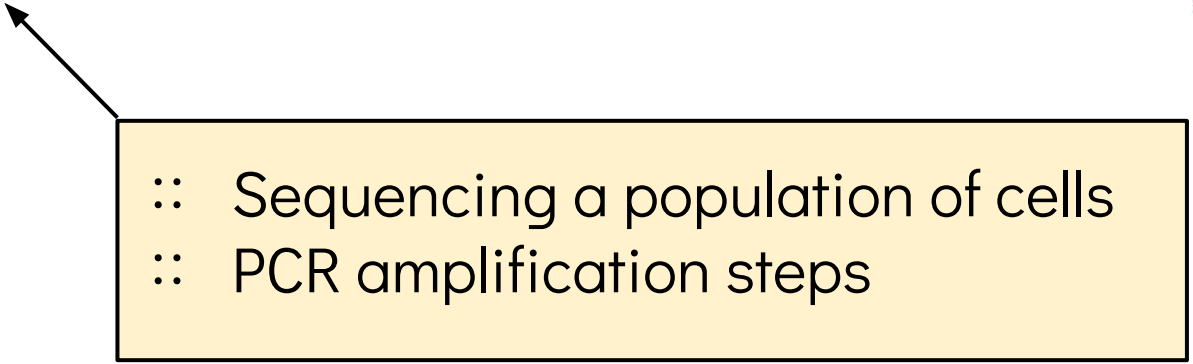
HENRY ERGAS

Risks rise in market downturn

Stan 'deserves' to be a grand slam man

Rio details blueprint for indigenous jobs

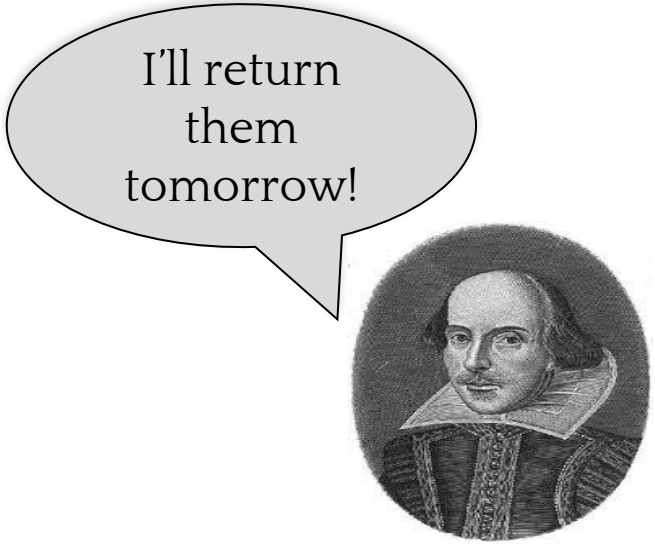
Ex-spy chief stands by SBS mission



:: Sequencing a population of cells
:: PCR amplification steps

An example

A small "genome"

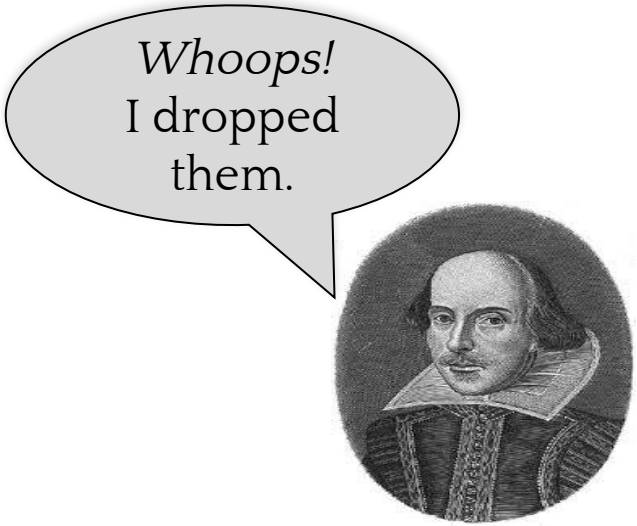



```
Friends,
 Romans,
    countrymen,
       lend me your ears;
```

Shakespearomics

- ***Reads***

```
ds, Romans, count
ns, countrymen, le
Friends, Rom
send me your ears;
crymen, lend me
```



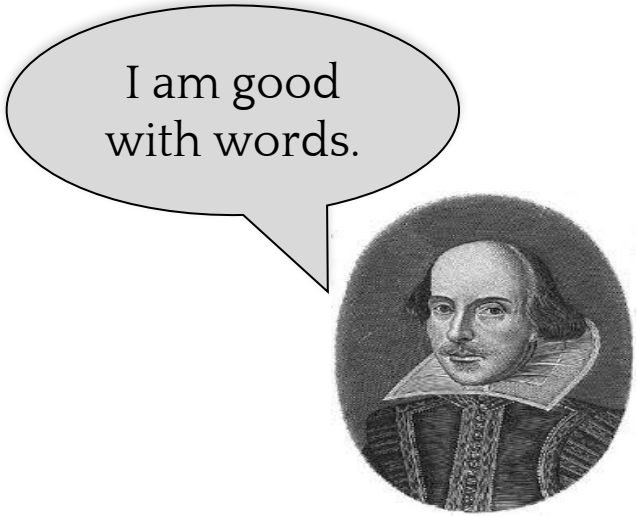
Shakespearomics

- ***Reads***

```
ds, Romans, count
ns, countrymen, le
Friends, Rom
send me your ears;
crymen, lend me
```

- ***Overlaps***

```
Friends, Rom
     ds, Romans, count
               ns, countrymen, le
                        crymen, lend me
                                send me your ears;
```



Shakespearomics

- ***Reads***

```
ds, Romans, count
ns, countrymen, le
Friends, Rom
send me your ears;
crymen, lend me
```

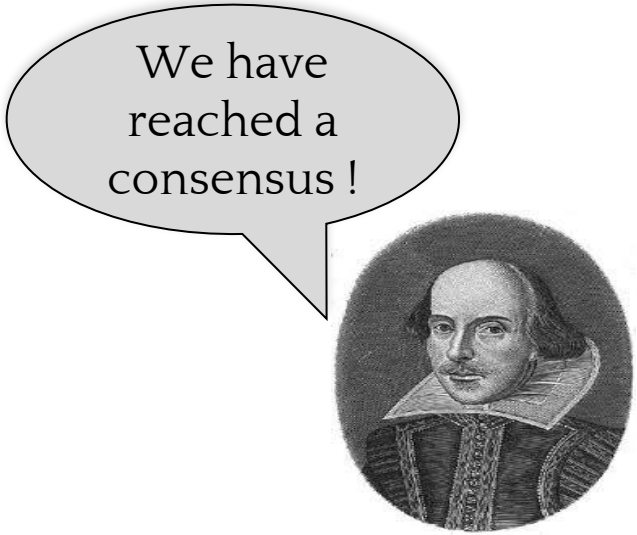
- ***Overlaps***

```
Friends, Rom
     ds, Romans, count
               ns, countrymen, le
                     crymen, lend me
                             send me your ears;
```

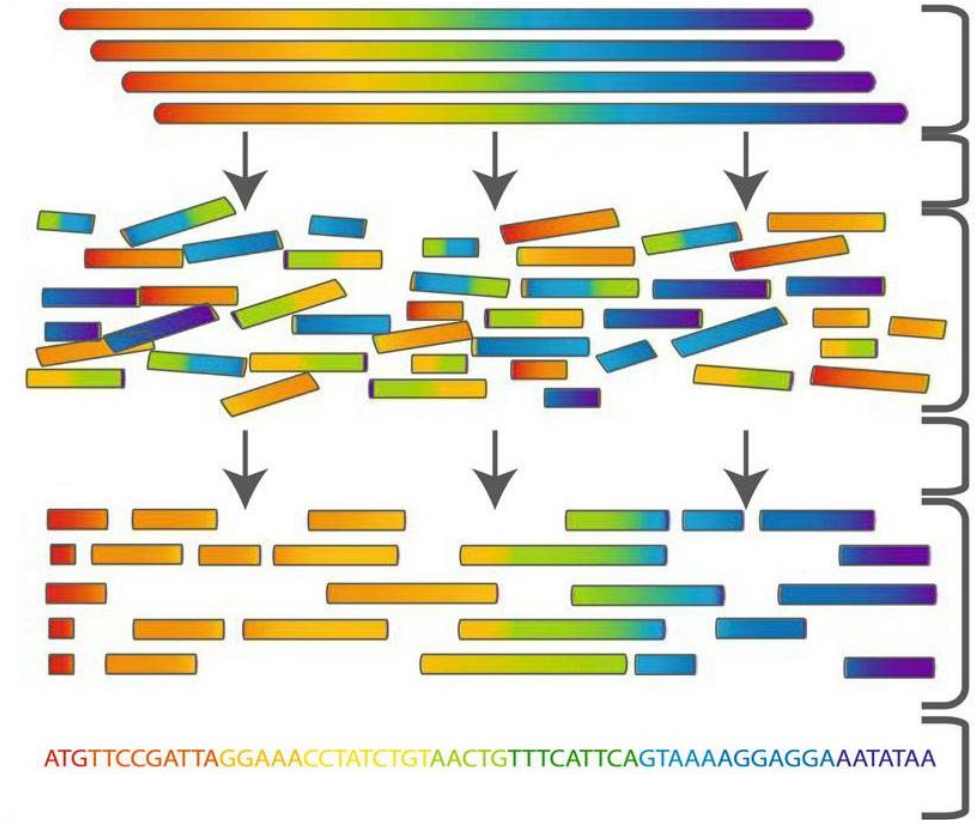
- ***Majority consensus***

```
Friends, Romans, countrymen, lend me your ears;
```

(1 contig)



Overlap - Layout - Consensus



Amplified DNA

Shear DNA

Sequenced reads

Overlaps

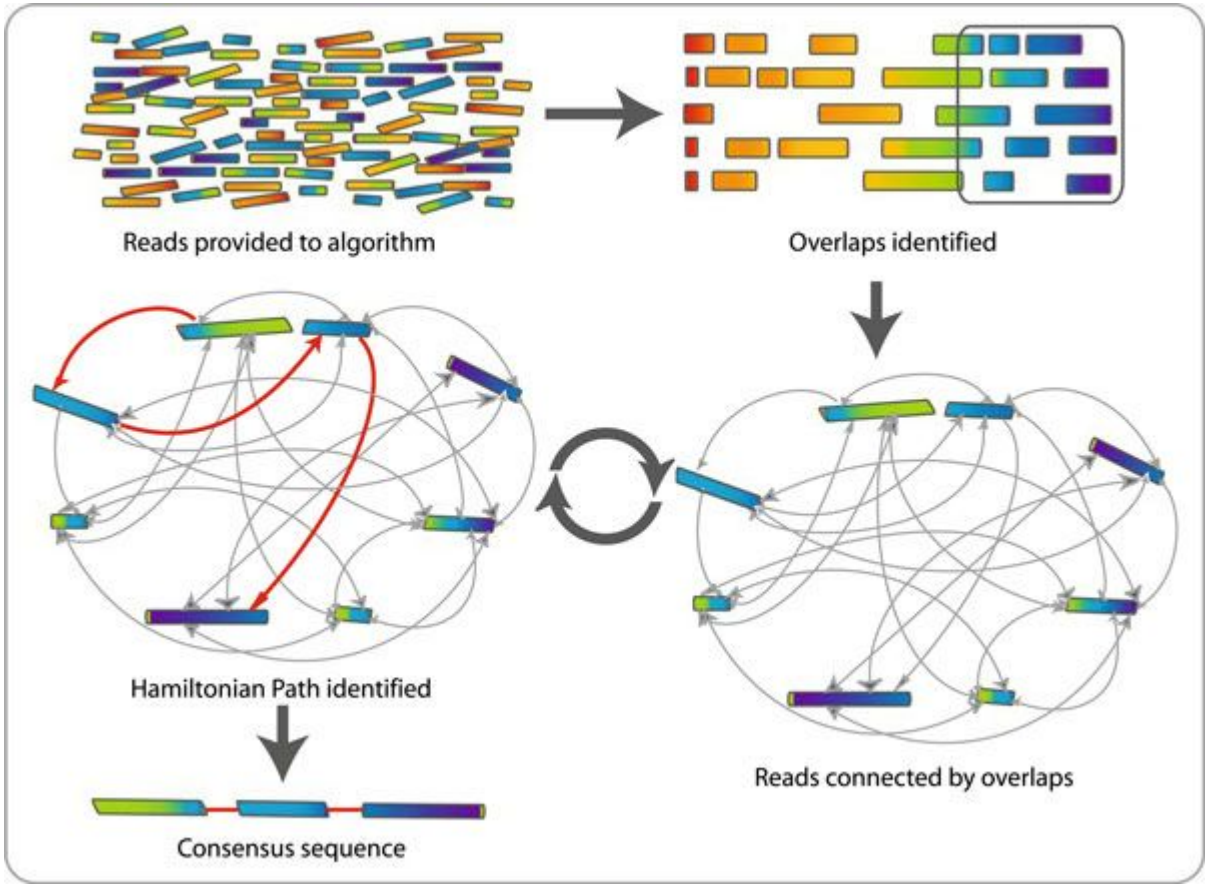
Layout

Consensus ↠ “Contigs”

Assembly graphs

(not Excel bar charts)

Overlap graph



Another example is this

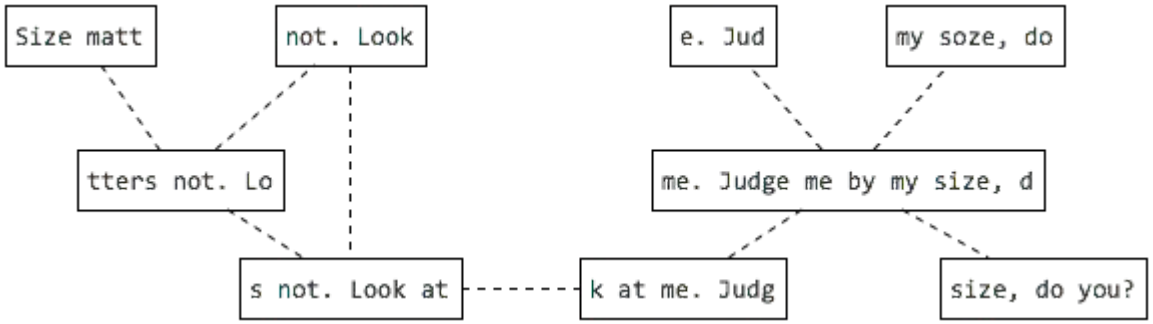
not. Look | s not. Look at | size, do you?

Size matt | e. Jud | my soze, do

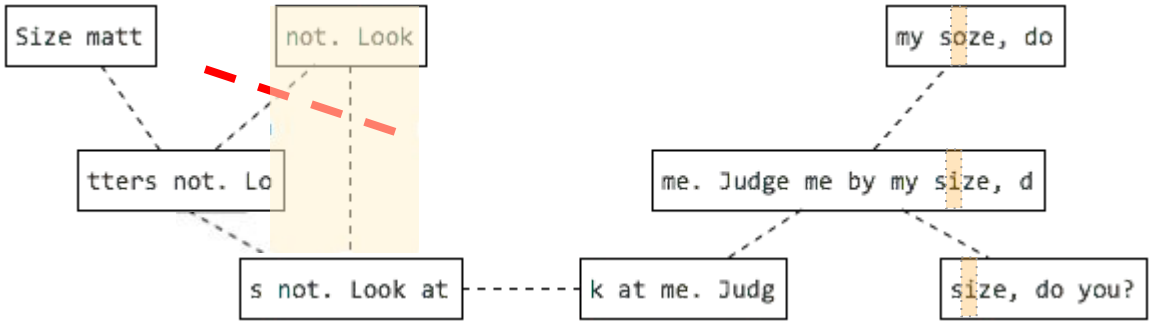
tters not. Lo | k at me. Judg

me. Judge me by my size, d

Overlaps find

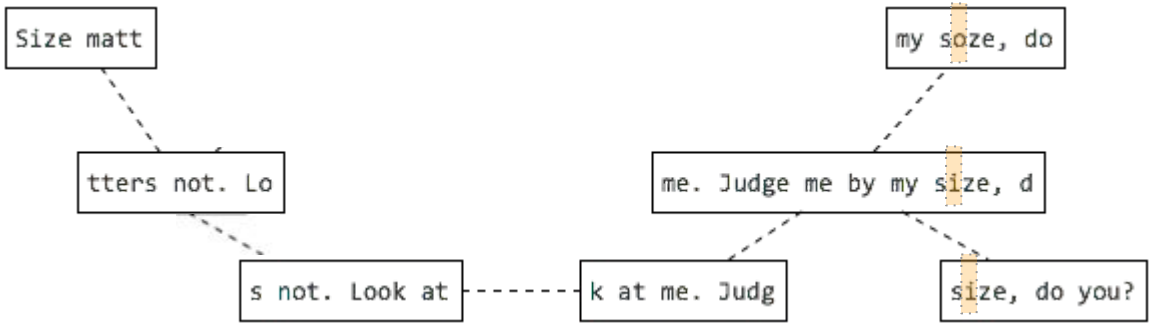


The graph one can simplify



"`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

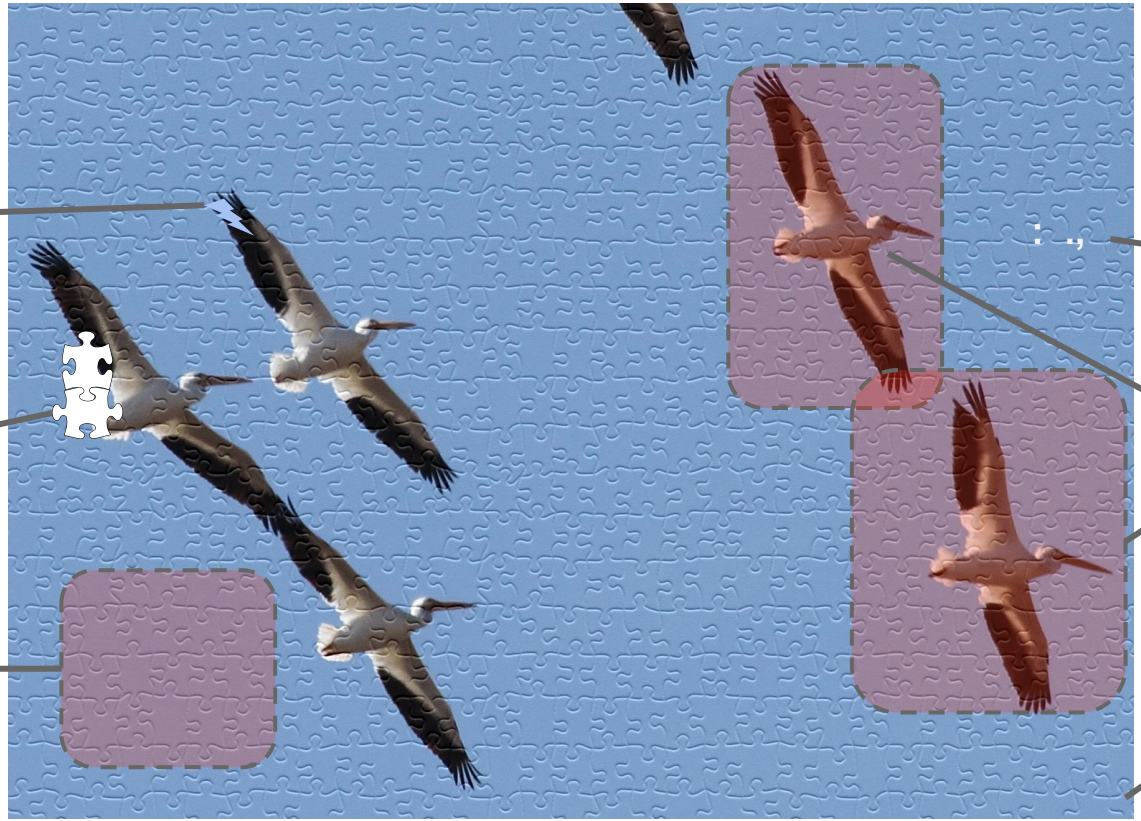
Size matters not. Look at me. Judge me by my soze, do you? ← 1 supporting read

So far, so good.


ONE DOES NOT SIMPLY
ASSEMBLE A GENOME

Why is it so hard?

What makes a jigsaw puzzle hard?



No box

Frayed pieces

Missing pieces

Repetitive regions

Lots of pieces

Dirty pieces

Multiple copies

No corners (circular genomes)

What makes genome assembly hard

1. **Many pieces (read length is very short compared to the genome)**

Size of the human genome = 3.2×10^9 bp (3,200,000,000)

Typical short read length = 10^2 bp (100)

A puzzle with millions to billions of pieces

Storing in RAM is a challenge ⇢ "succinct data structure"

What makes genome assembly hard

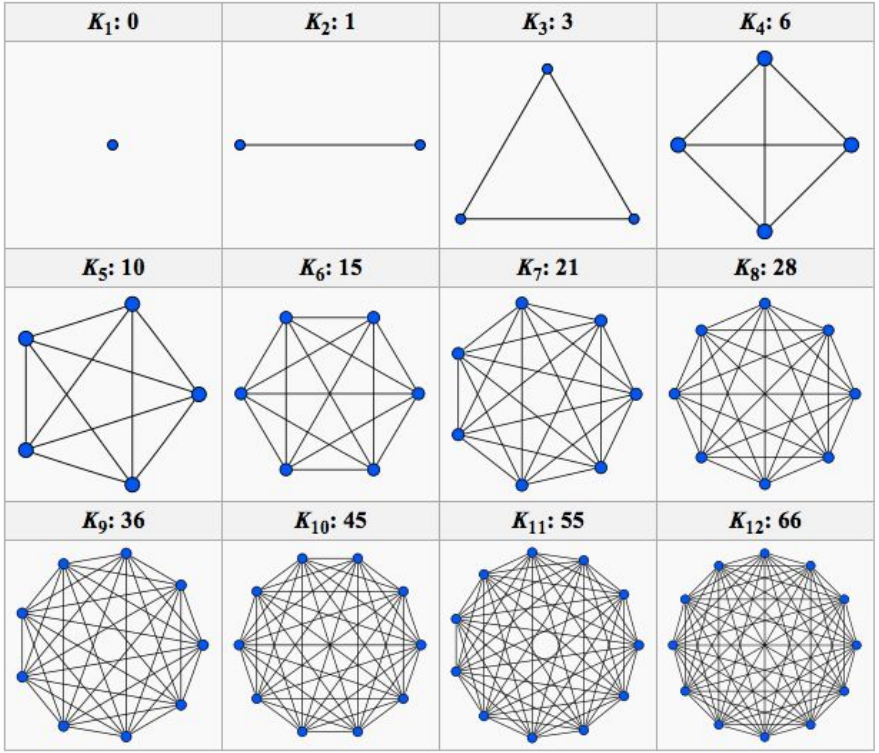
2. Lots of overlaps

Finding overlaps means examining every pair of reads

Comparisons = $N \times (N-1)/2$

$\sim N^2$

Lots of smart tricks to reduce this close to ~N



What makes genome assembly hard

3. Lots of sky (short repeats)

ATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATATA

```
   TATATATA  TATATATA         TATATATA       TATATATA     TATATATA
                  TATATATA  TATATATA                 TATATATA
 TATATATA     TATATATA         TATATATA
```

How could we possibly assemble this segment of DNA?
All the reads are the same!

What makes genome assembly hard

4. Dirty pieces (sequencing errors)

```
Read 1: GGAACCTTTGGCCCTGT
Read 2:          GGCGCTGTCCATTTTAGAAACC
```

What counts as “overlapping” ?

- Minimum overlap length
- Minimum DNA identity

What makes genome assembly hard

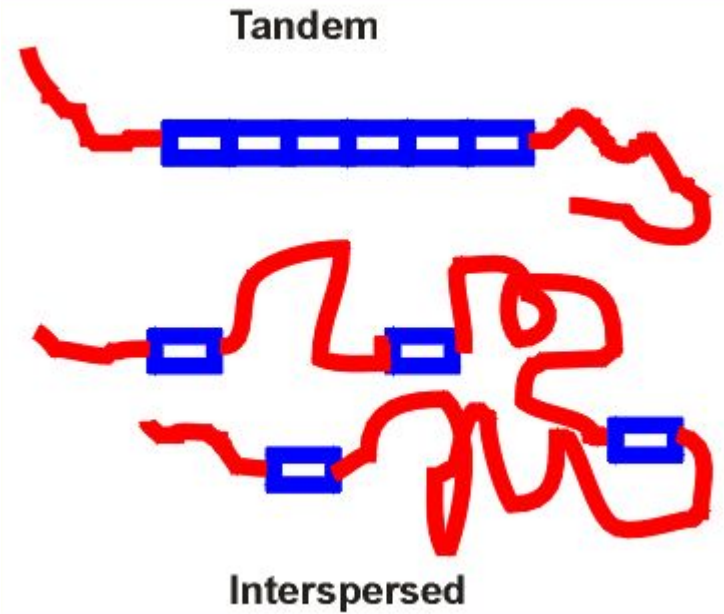
5. Multiple copies (long repeats)

Our old nemesis the REPEAT !

Repeats

What is a repeat?

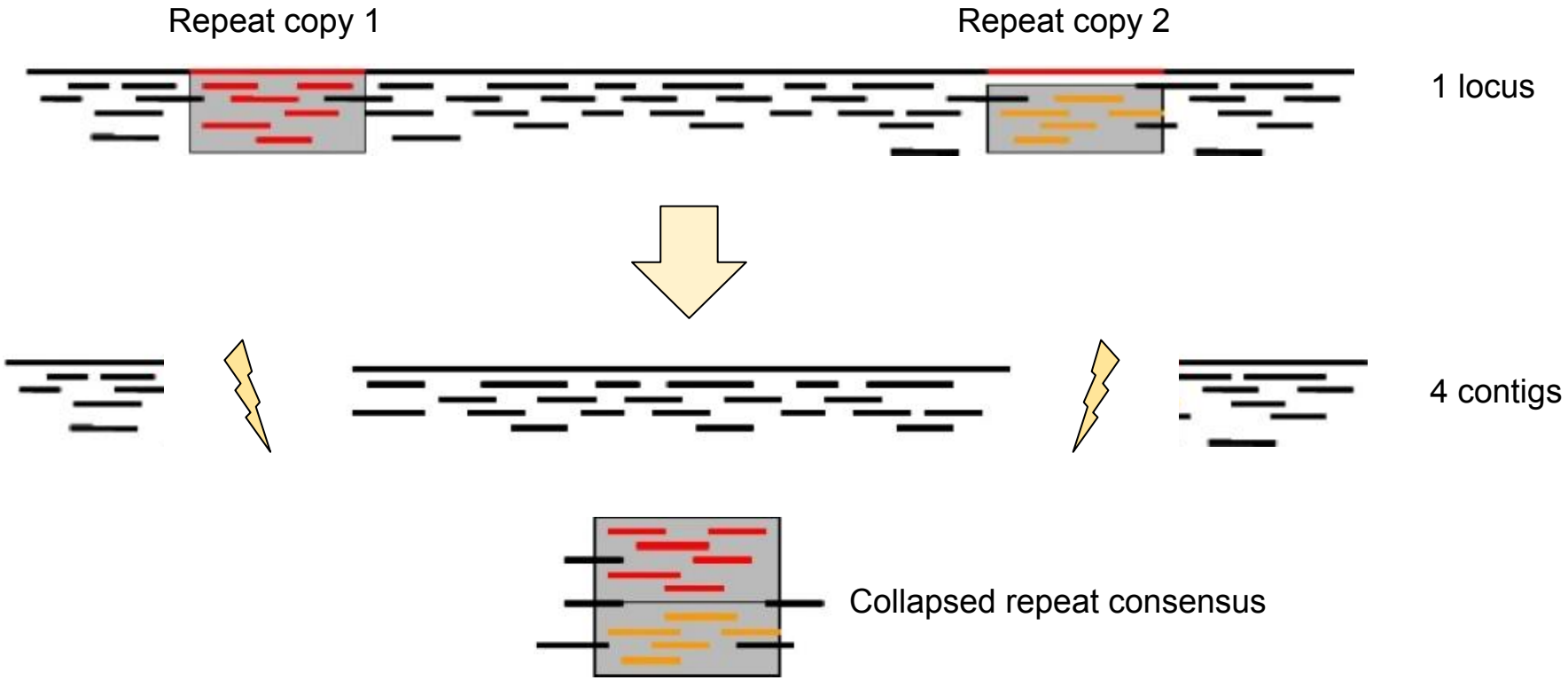
A segment of DNA
that occurs *more than once*
in the genome



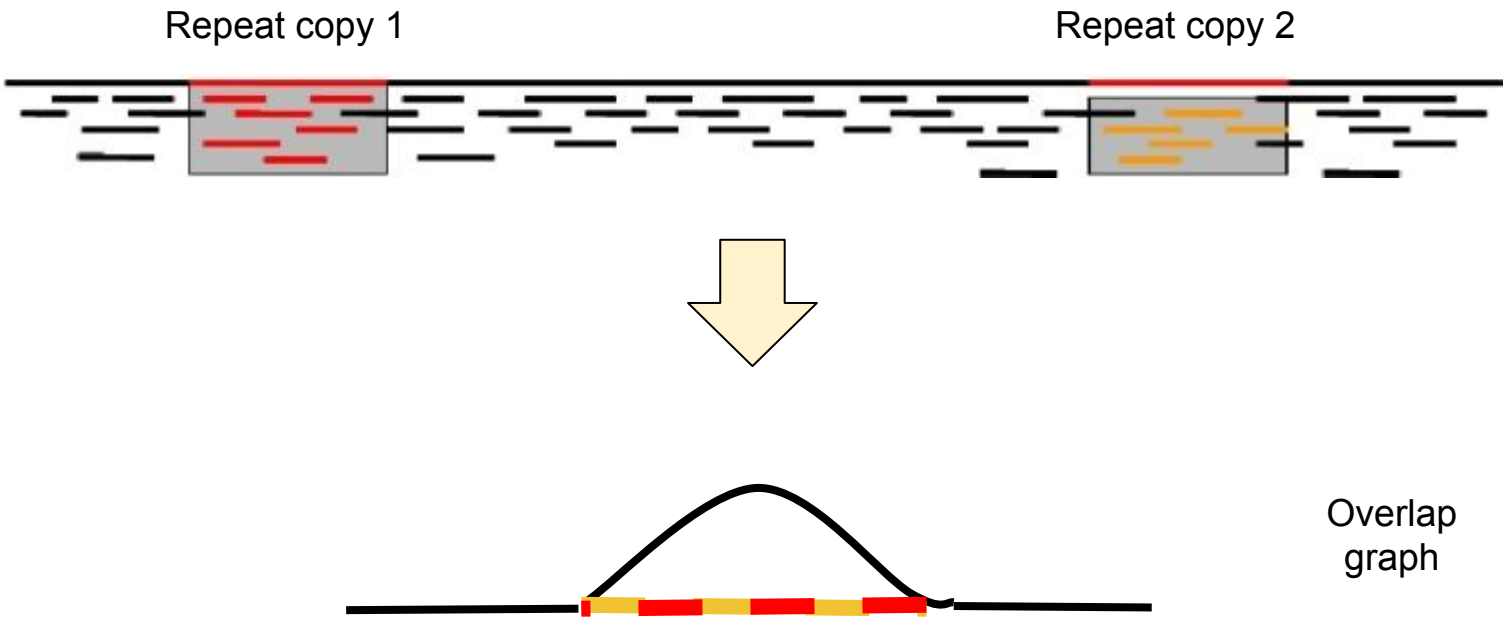
Major classes of repeats in the human genome

Repeat Class	Arrangement	Coverage (Hg)	Length (bp)
Satellite (micro, mini)	Tandem	3%	2-100
SINE	Interspersed	15%	100-300
Transposable elements	Interspersed	12%	200-5k
LINE	Interspersed	21%	500-8k
rDNA	Tandem	0.01%	2k-43k
Segmental Duplications	Tandem or Interspersed	0.2%	1k-100k

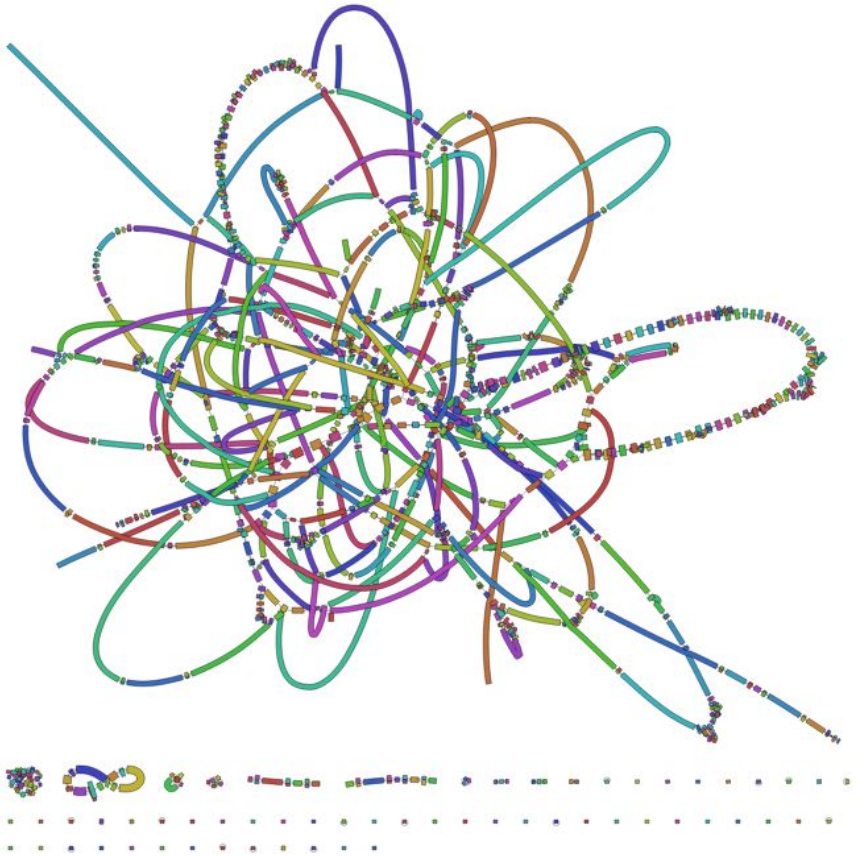
Repeats


Repeat copy 1
Repeat copy 2
1 locus
4 contigs
Collapsed repeat consensus

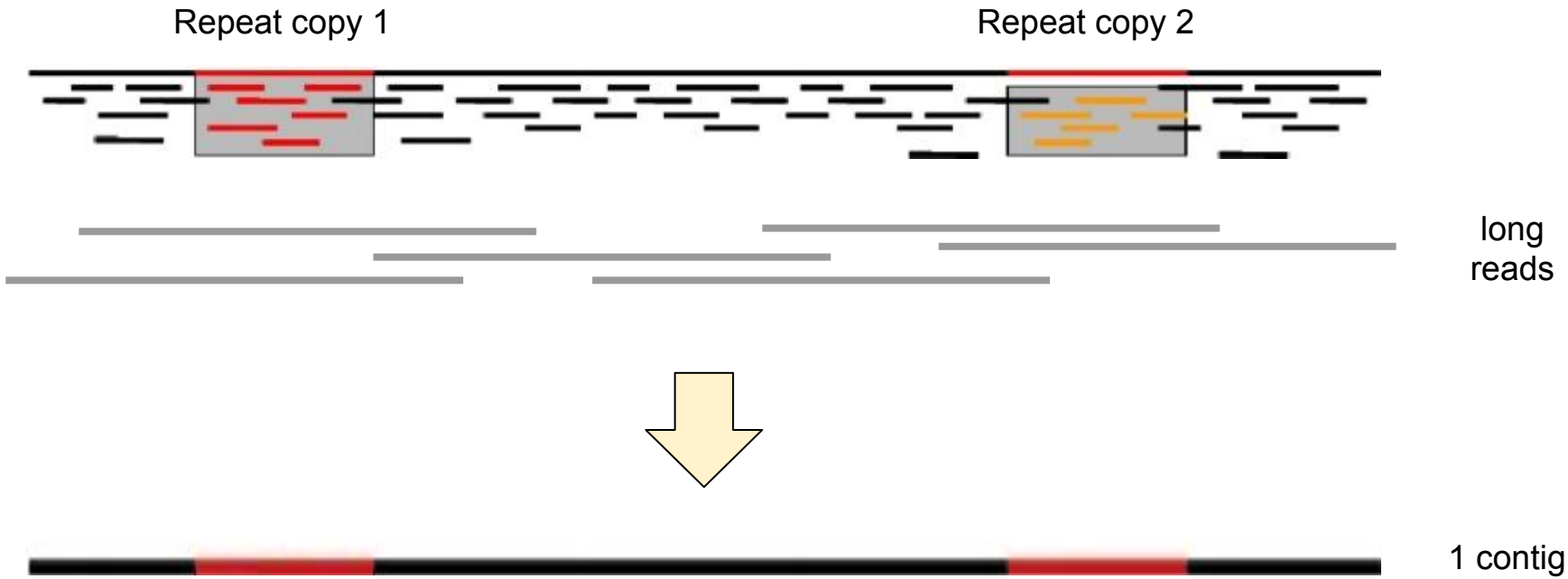
Repeats cause graph ambiguity



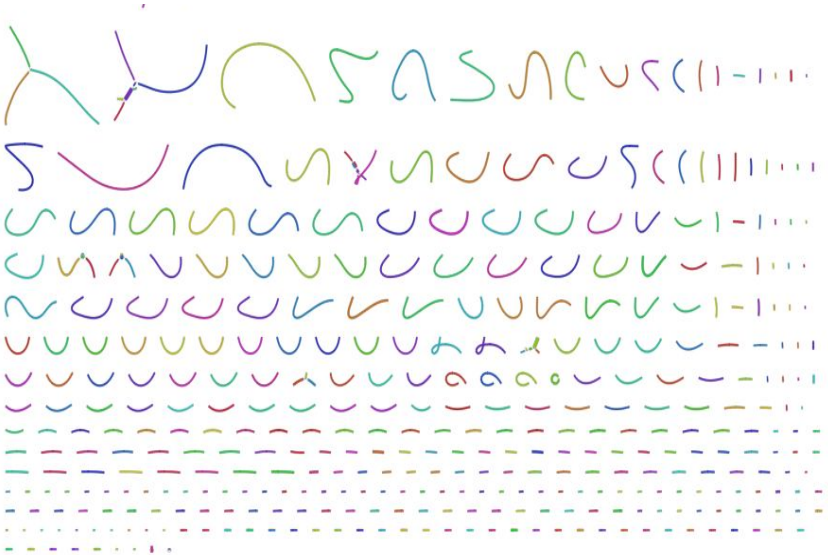
Repeats are hubs in the graph



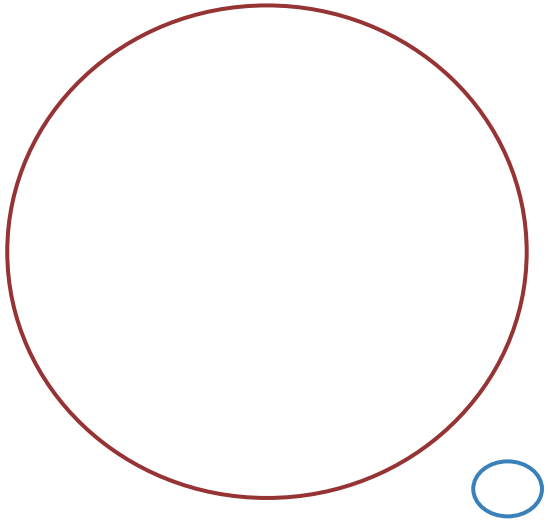
Long reads can span repeats



Draft vs Finished genomes (bacteria)



150 bp - Illumina - $200



10,000 bp - Pacbio - $2000

The two laws of repeats

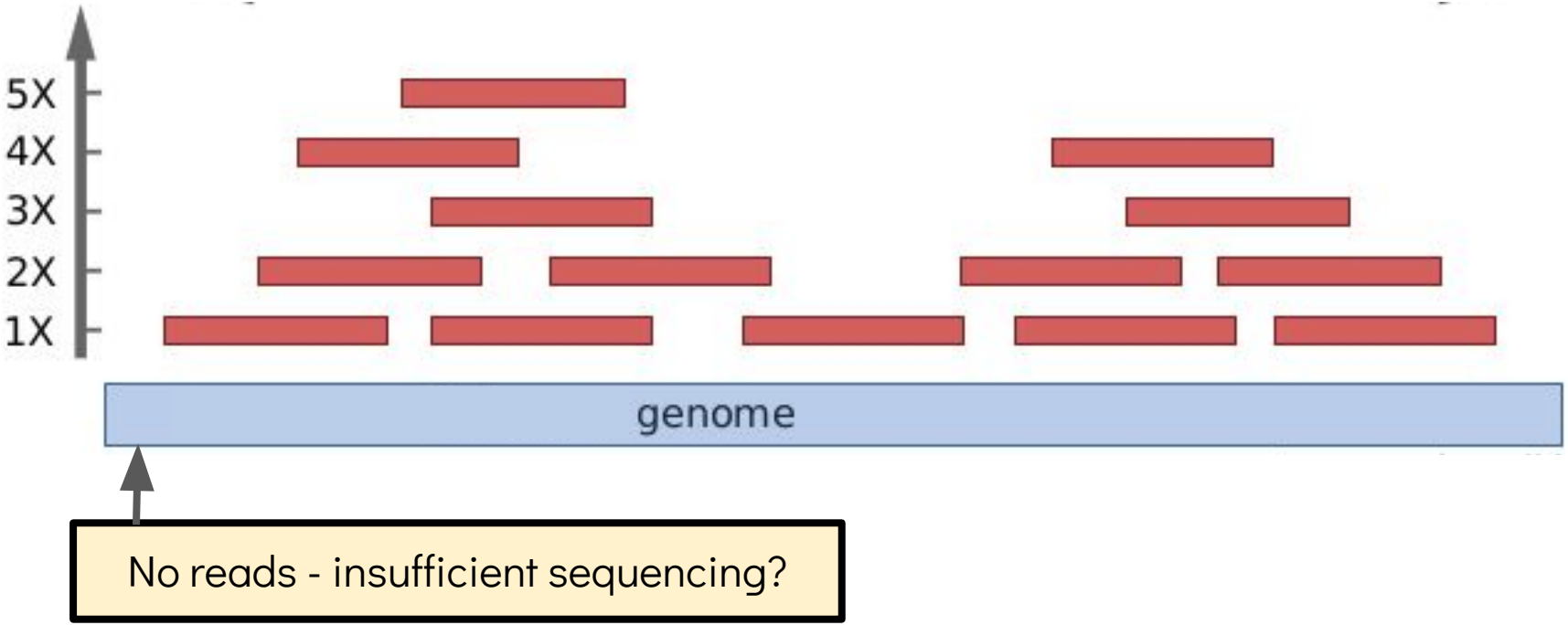
1. It is impossible to resolve repeats of length L unless you have reads longer than L.

2. It is impossible to resolve repeats of length L unless you have reads longer than L.

Assumptions

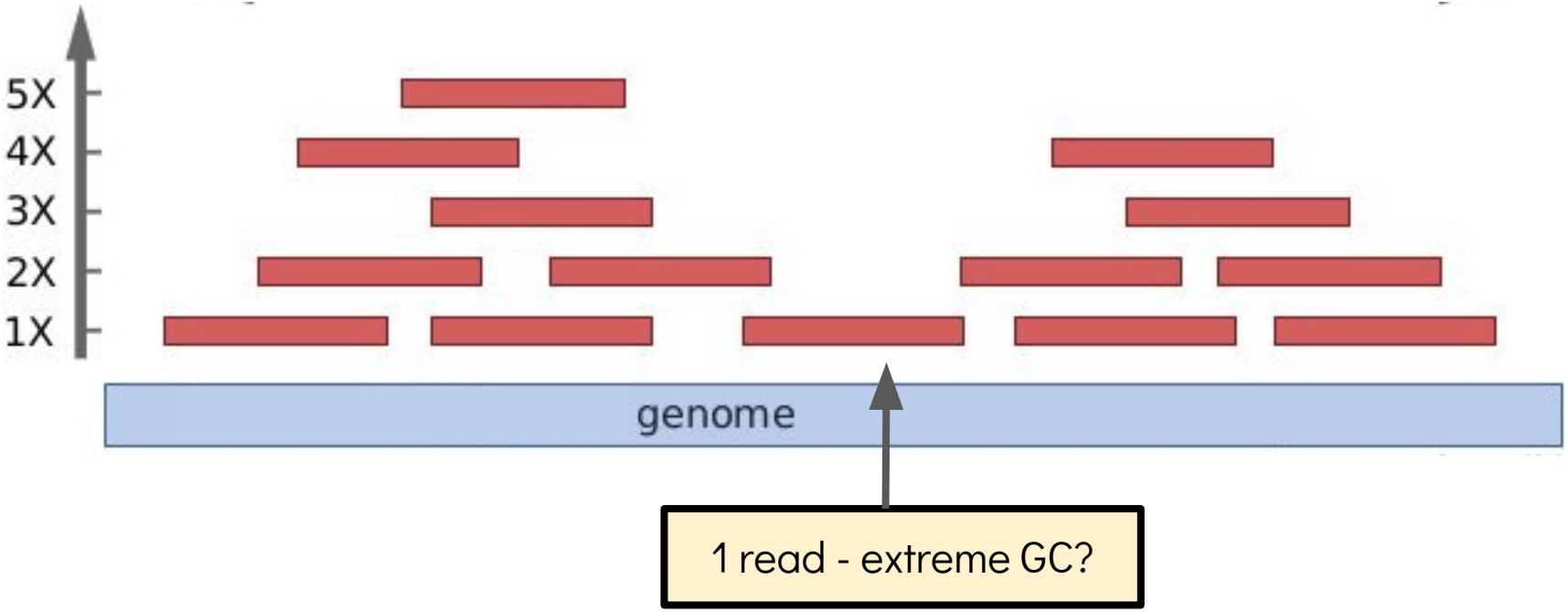
Full coverage

Each base in the genome was sequenced



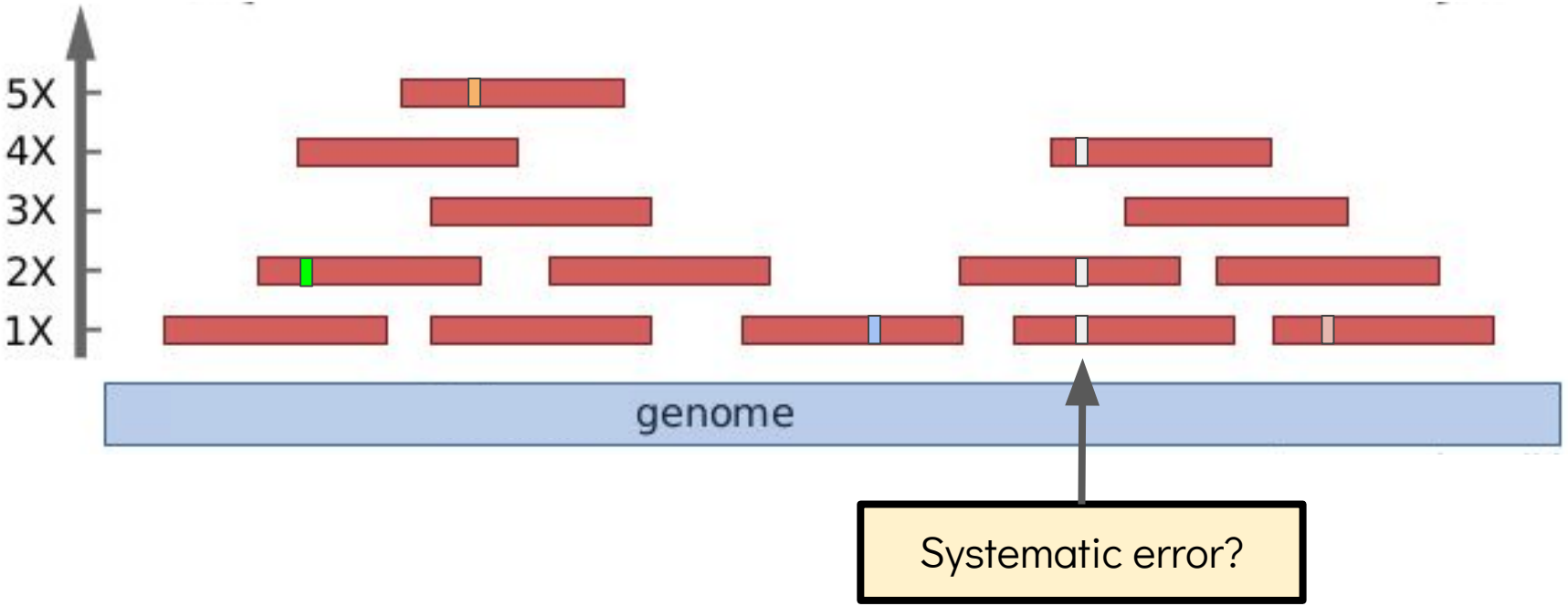
Sufficient depth

Each base was covered by enough independent reads



Random errors

Sequencing read errors are random so consensus wins out

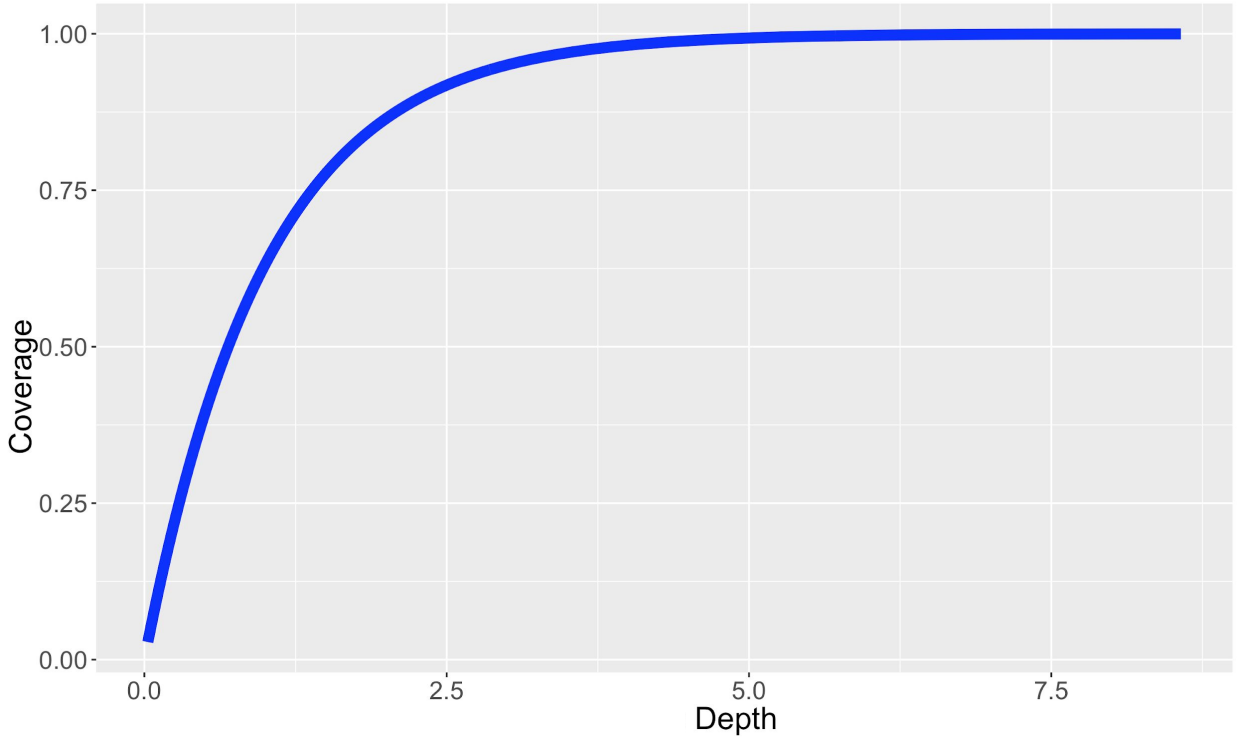


How much data do we need?

Coverage and depth are related

Approximate formula for coverage assuming random reads

$$\text{Coverage} = 1 - e^{-\text{Depth}}$$



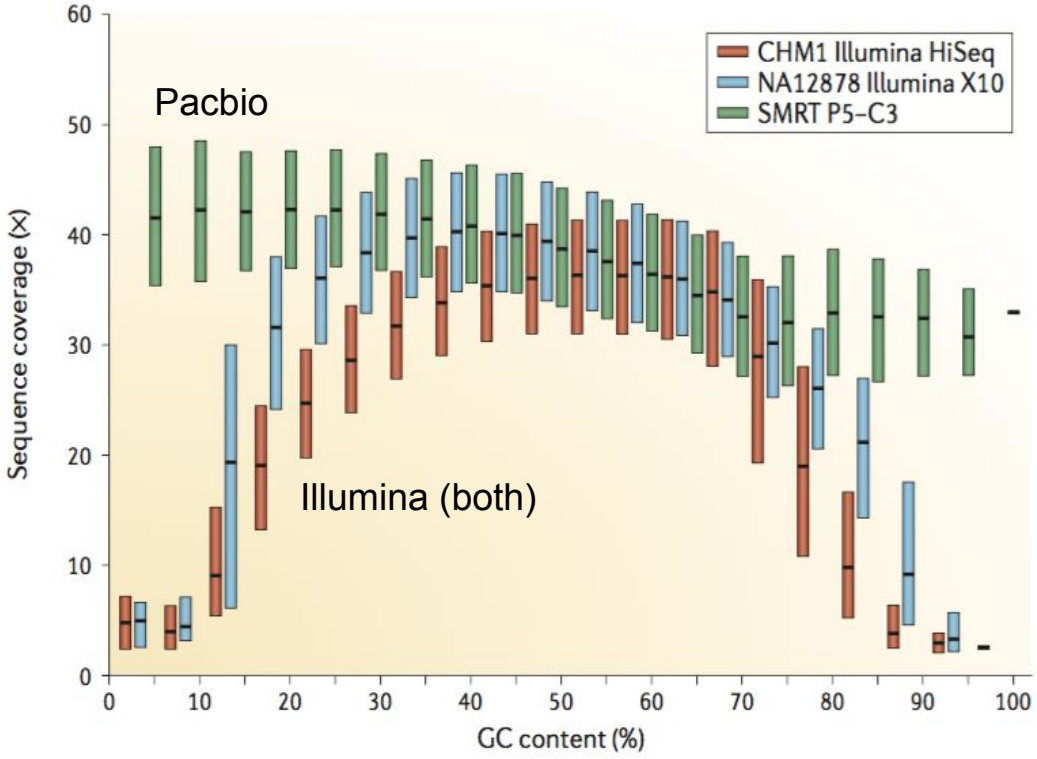
Much more sequencing needed in reality

Sequencing is not random

- GC and AT rich regions are under represented
- Other chemistry quirks

More depth needed for:

- sequencing errors
- polyploid organisms
- mixed population
- cancer



Assessing assemblies

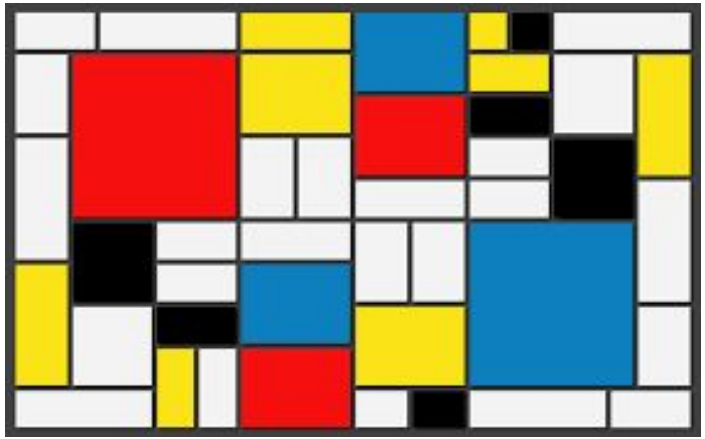
Contiguity

Completeness

Correctness

Contiguity

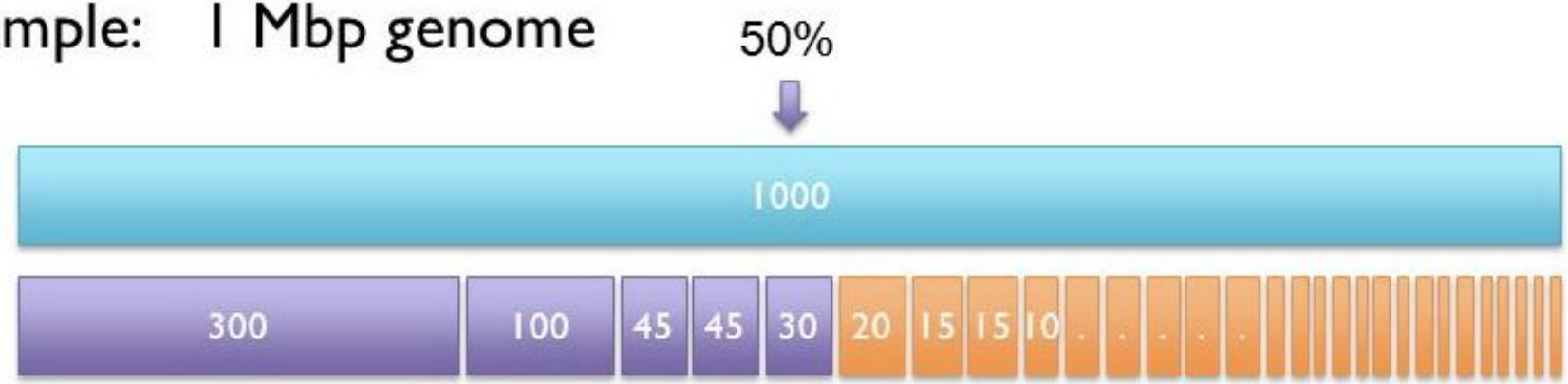
- Desire
  - Fewer contigs
  - Longer contigs



- Metrics
  - Number of contigs
  - Average contig length
  - Median contig length
  - Maximum contig length
  - “N50”, “NG50”, “D50”

Contiguity: the N50 statistic

Example: 1 Mbp genome



N50 size = 30 kbp

(300k+100k+45k+45k+30k = 520k >= 500kbp)

Completeness : Total size

Proportion of the original genome represented by the assembly

Can be between 0 and 1

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

Proportion of estimated genome size

... but estimates are not perfect

Completeness: core genes

Proportion of coding sequences can be estimated based on known core genes thought to be present in a wide variety of organisms.

Assumes that the proportion of assembled genes is equal to the proportion of assembled **core** genes.

In the past this was done with a tool called CEGMA

There is a new tool for this called BUSCO

$$\frac{\text{Number of Core Genes in Assembly}}{\text{Number of Core Genes in Database}}$$

Correctness

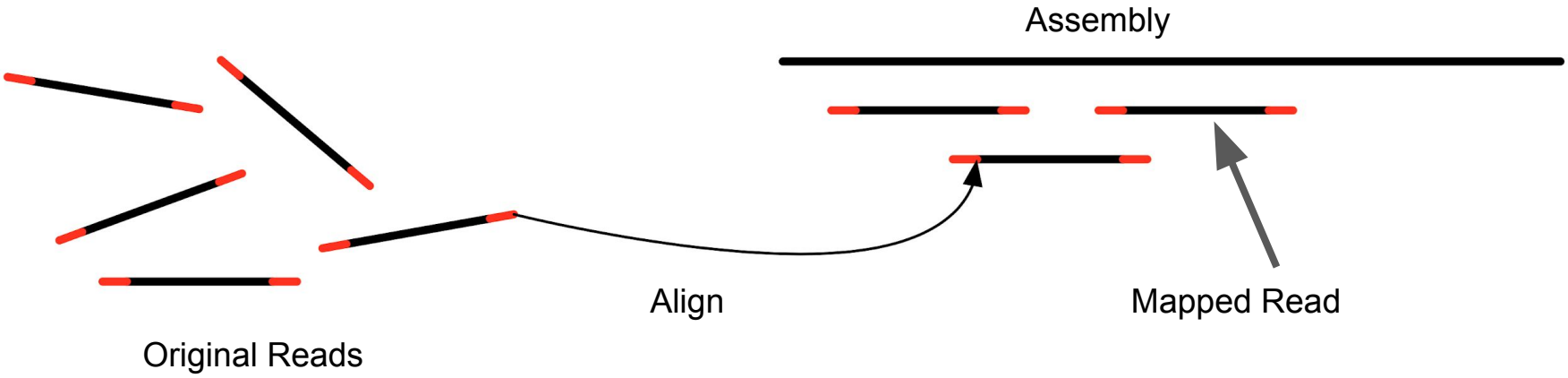
Proportion of the assembly that is free from mistakes

Errors include

1. Mis-joins
2. Repeat compressions
3. Unnecessary duplications
4. Indels / SNPs caused by assembler

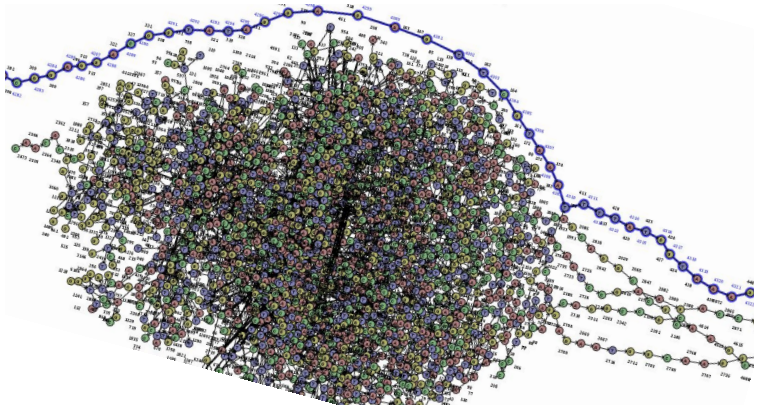
Correctness: check for self consistency

- Align all the reads back to the contigs
- Look for inconsistencies



Assemble ALL the things

Not just genomes



- Transcriptomes
  - One contig for every isoform
  - Do not expect uniform coverage
- Meta-genomes
  - Mixture of different organisms
  - Host, bacteria, virus, fungi all at once
  - All different depths
- Meta-transcriptomes
  - Combination of above!

Conclusions

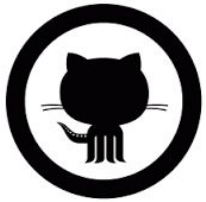
Take home points

- *De novo* assembly is the process of reconstructing a long sequence from many short ones

- Represented as a mathematical “overlap graph”

- Assembly is very challenging (“impossible”) because
  - sequencing bias under represents certain regions
  - Reads are short relative to genome size
  - Repeats create tangled hubs in the assembly graph
  - Sequencing errors cause detours and bubbles in the assembly graph

Contact



tseemann.github.io

torsten.seemann@gmail.com

@torstenseemann

The End