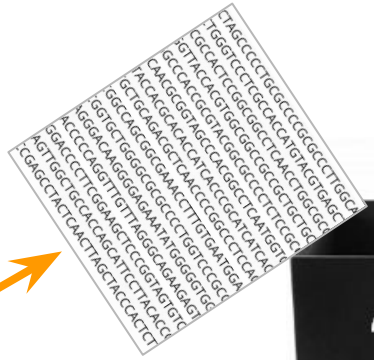
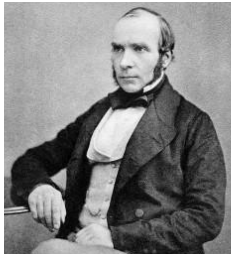
Introduction to bioinformatics for microbial genomics

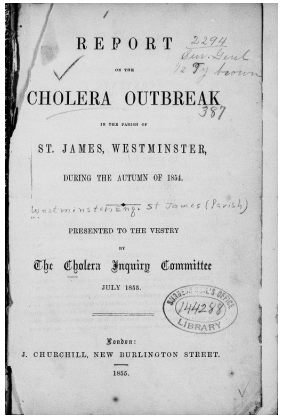
A/Prof Torsten Seemann

Introduction

Bioinformatics







What's inside the black box?

It's black boxes, all the way down.

Data / file formats

Software / analysis

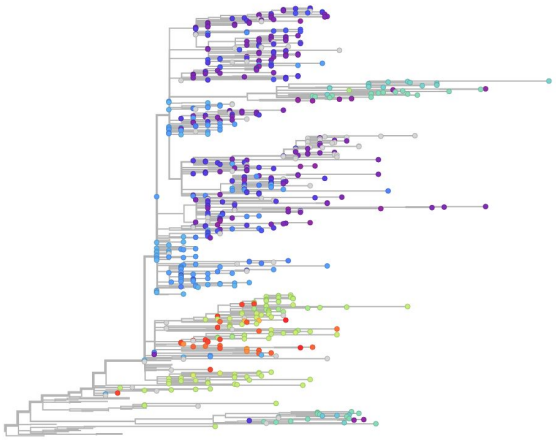
Results / interpretation

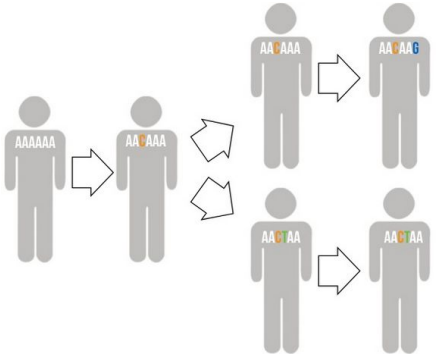
Reporting / action

We use real black boxes too!

Traditional public health and clinical microbiology

The roles of a public health laboratory





Diagnostics

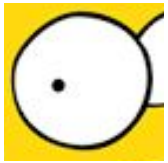
Surveillance

Outbreak response

Tracking a diverse range of pathogens

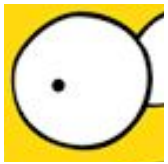
Pathogen: *Homerbacter simpsonii*

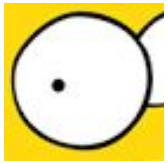
Focus on a small informative section



Genotyping via one or more marker genes.

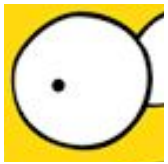
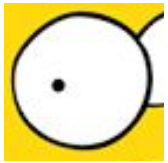
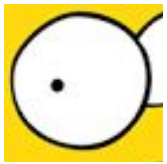
Another sample arrives at the lab





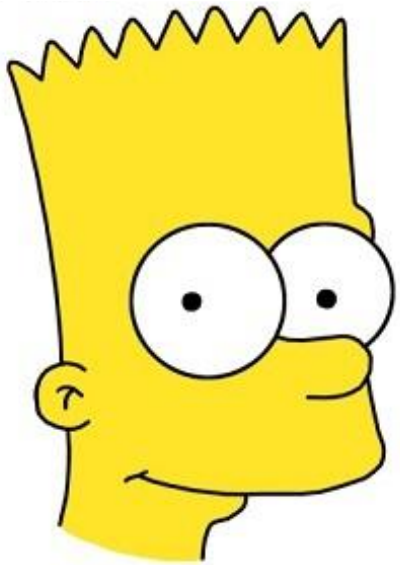
Looks related.

Another sample from same city

Possible outbreak ?

D’oh !



How can we improve this?

Of course you know the answer.

Whole genome sequencing

:: Use 100% of the genome

:: Single nucleotide resolution
   : Infer source and transmission of infection

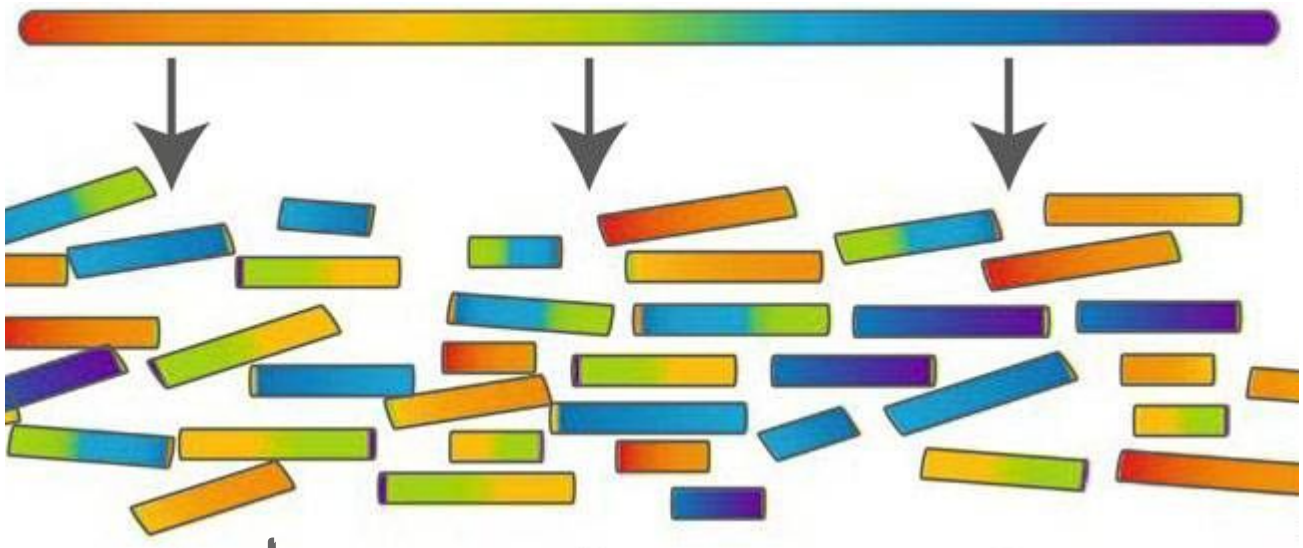
:: Full complement of genes
   : Track mobile elements and antimicrobial resistance

Routine sequencing of isolates

Can sequence 100s of genomes per week

Whole genome sequencing

The currency of genomics



Reads are stored in FASTQ files

Types of sequence reads

illumina®

100 - 300 bp

PACIFIC BIOSCIENCES®

5,000 - 45,000 bp

Oxford NANOPORE Technologies

5,000 - 900,000+ bp

FASTQ files

The FASTQ format

A single sequence read looks like this:

```
@read00179
AGTCTGATATGCTGTACCTATTATAATTCTAGGCGCTCAT
+
8;ACCCD?DD???@B9<9<871CAC@=A;0.-&*,(0**#
```

FASTQ components

Start symbol

Sequence ID

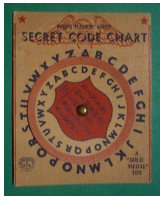
Sequence

Separator line

```
@read00179
AGTCTGATATGCTGTACCTATTATAATTCTAGGCGCTCAT
+
8;ACCCD?DD???@B9<9<871CAC@=A;0.-&*,(0**#
```

Encoded quality values,
one symbol per nucleotide

FASTQ quality encoding

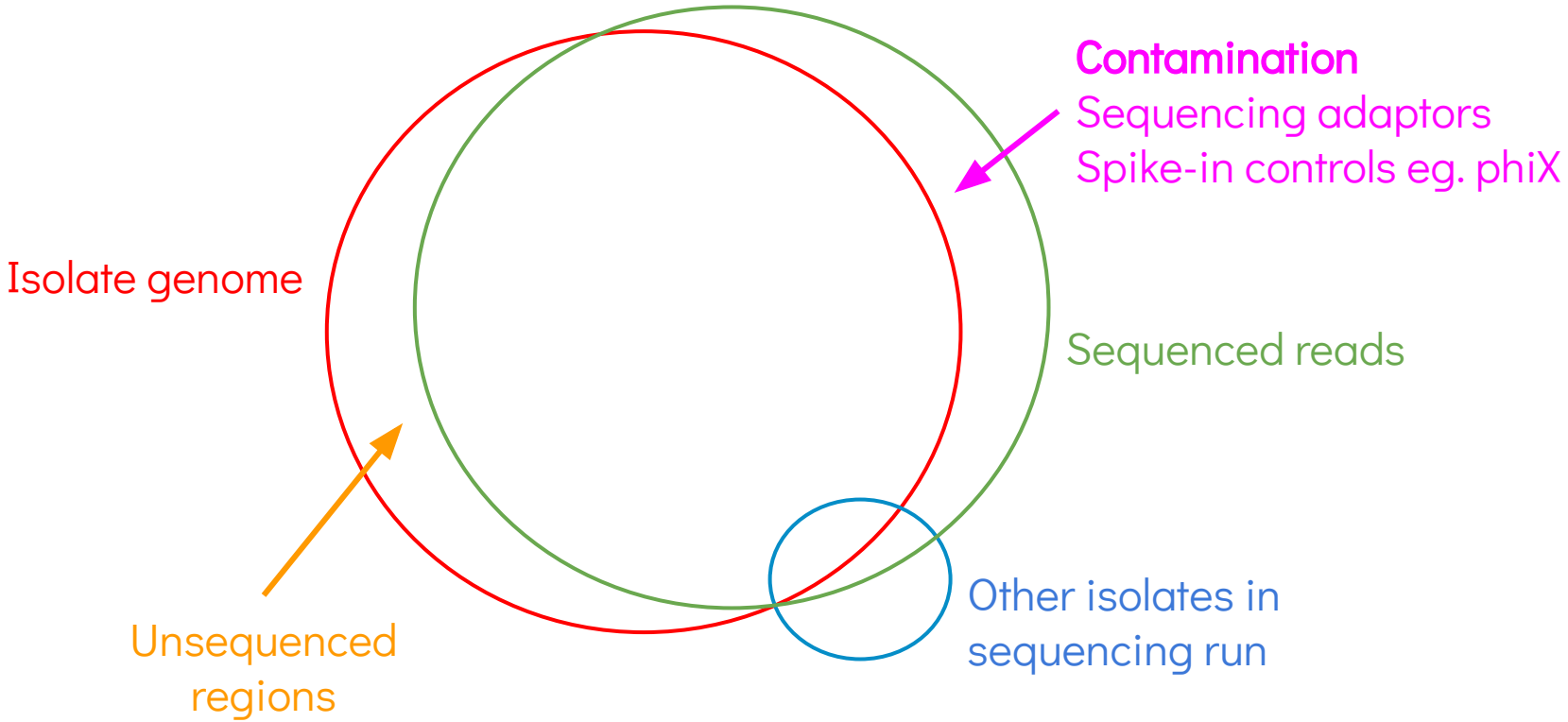



```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
|         |         |         |         |
Q0        Q10       Q20       Q30       Q40
bad       maybe     ok        good      excellent
```

- Use characters to represent numbers
  - numbers measure quality
  - a reliability estimate for each base
  - estimated from the physical measuring process
  - formalized during the Human Genome Project

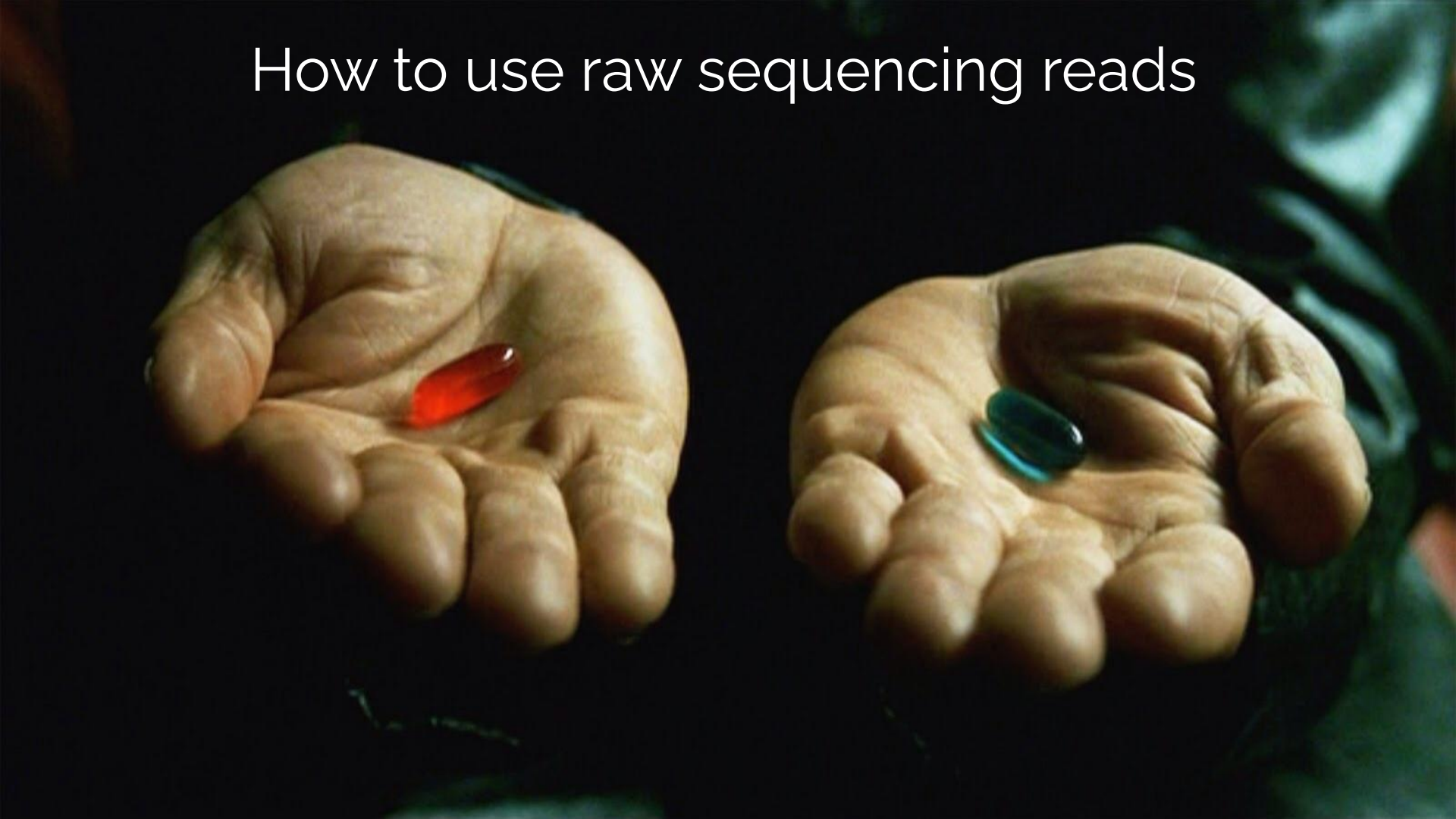
Quality control

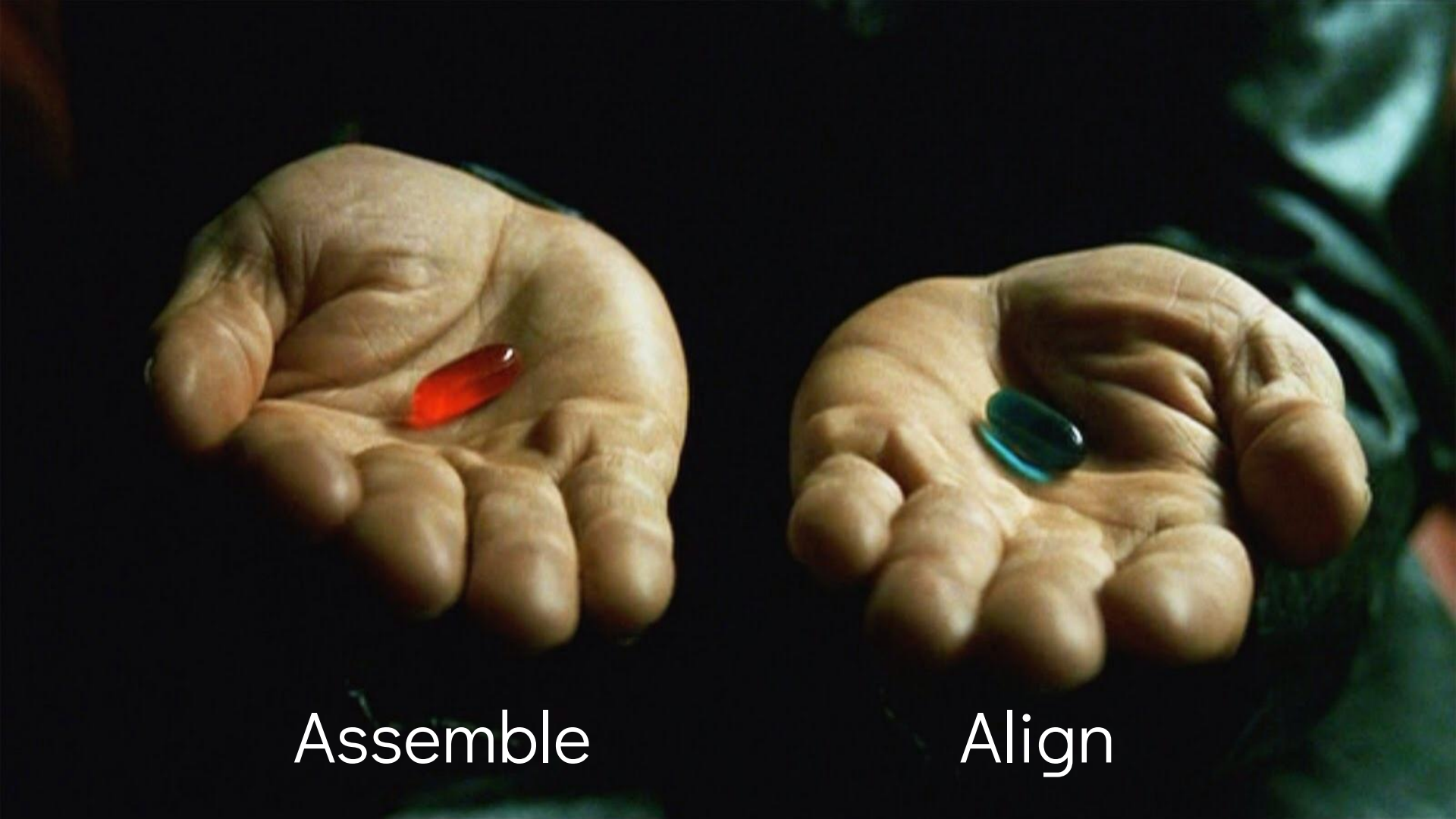
What data do we really have?



Got my reads, now what?

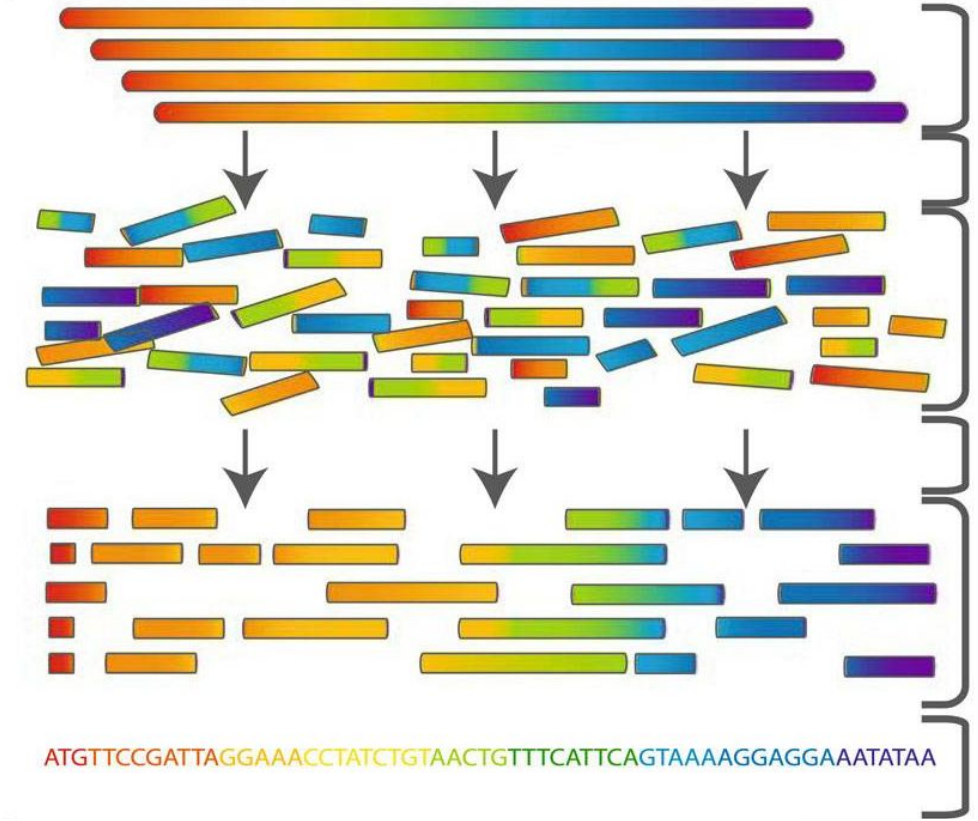
How to use raw sequencing reads




Assemble
Align

Genome assembly

De novo genome assembly



Amplified DNA

Shear DNA

Sequenced reads

Overlaps

Layout

Consensus ↠ "Contigs"

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

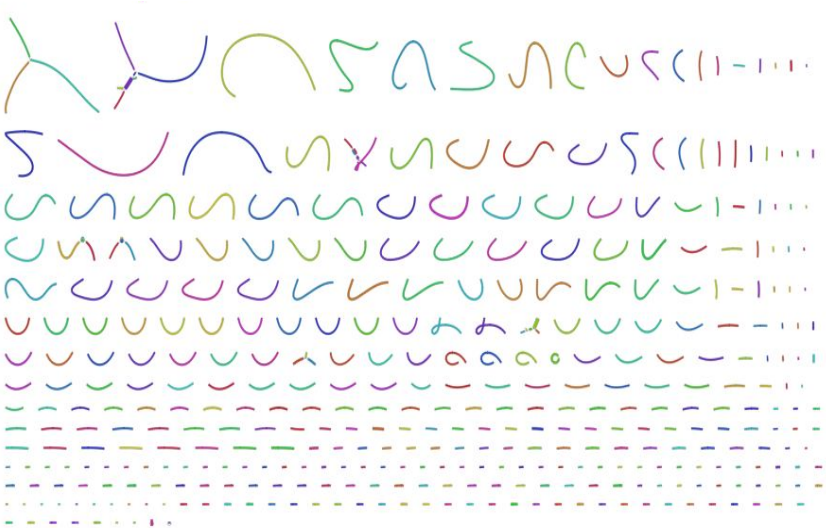
Ex-spy chief stands by SBS mission

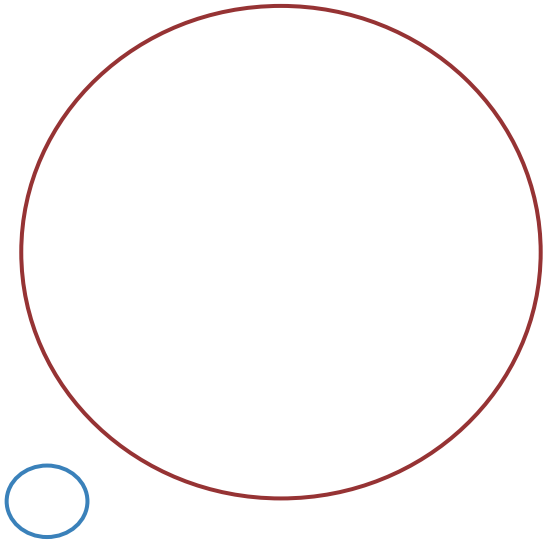
"From scratch"

De novo genome assembly

Usually sequencing a *population* of cells

The effect of read length





150 bp - Illumina (short)

8,000 bp - Nanopore (long)

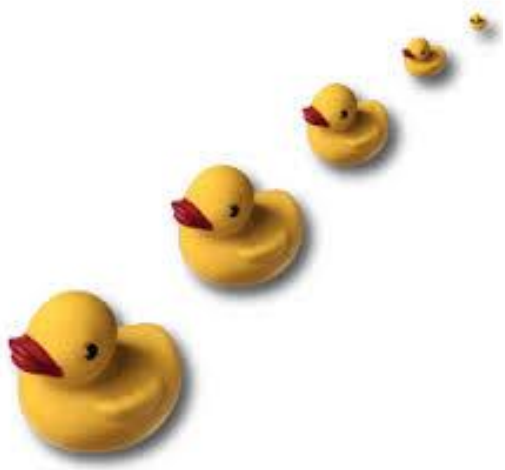
Align to reference

Align to reference

Seven short 4bp reads

AGTC TTAC GGGA CTTT

TAGG TTTA ATAG



Aligned to 31bp reference

AGTCTTTATTATAGGGAGCCATAGCTTTACA

AGTC TAGG ATAG TTAC

TTTA GGGA CTTT

Ambiguous alignment

Eight short 4bp reads

AGTC TTAC GGGA CTTT

TAGG TTTA ATAG **TTAT**

Aligned to 31bp reference

AGTCTTTATTATAGGGAGCCATAGCTTTACA

AGTC TAGG ATAG TTAC

TTTA GGGA CTTT

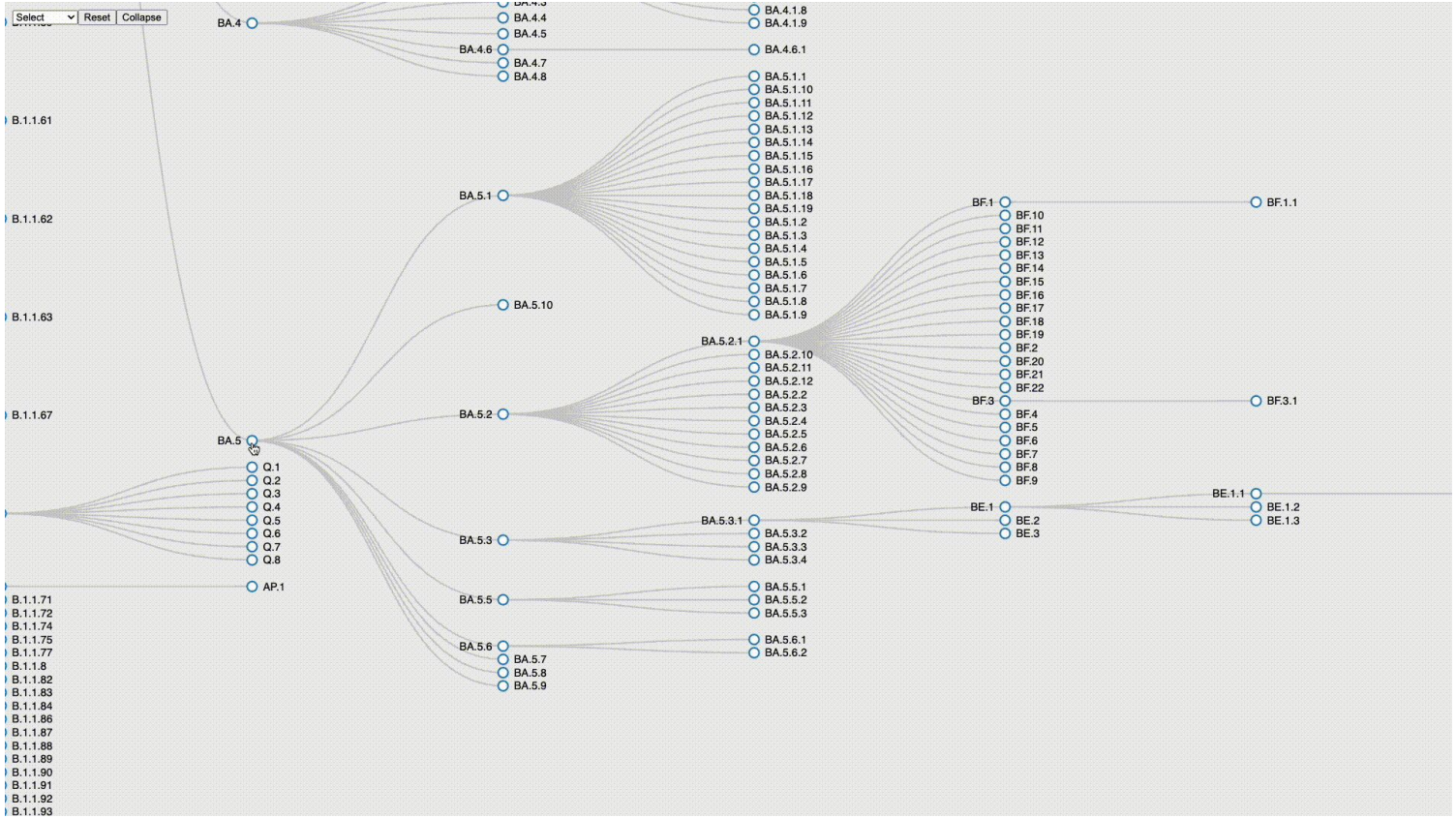
TTAT

TTAT

Using reads, assemblies, and alignments

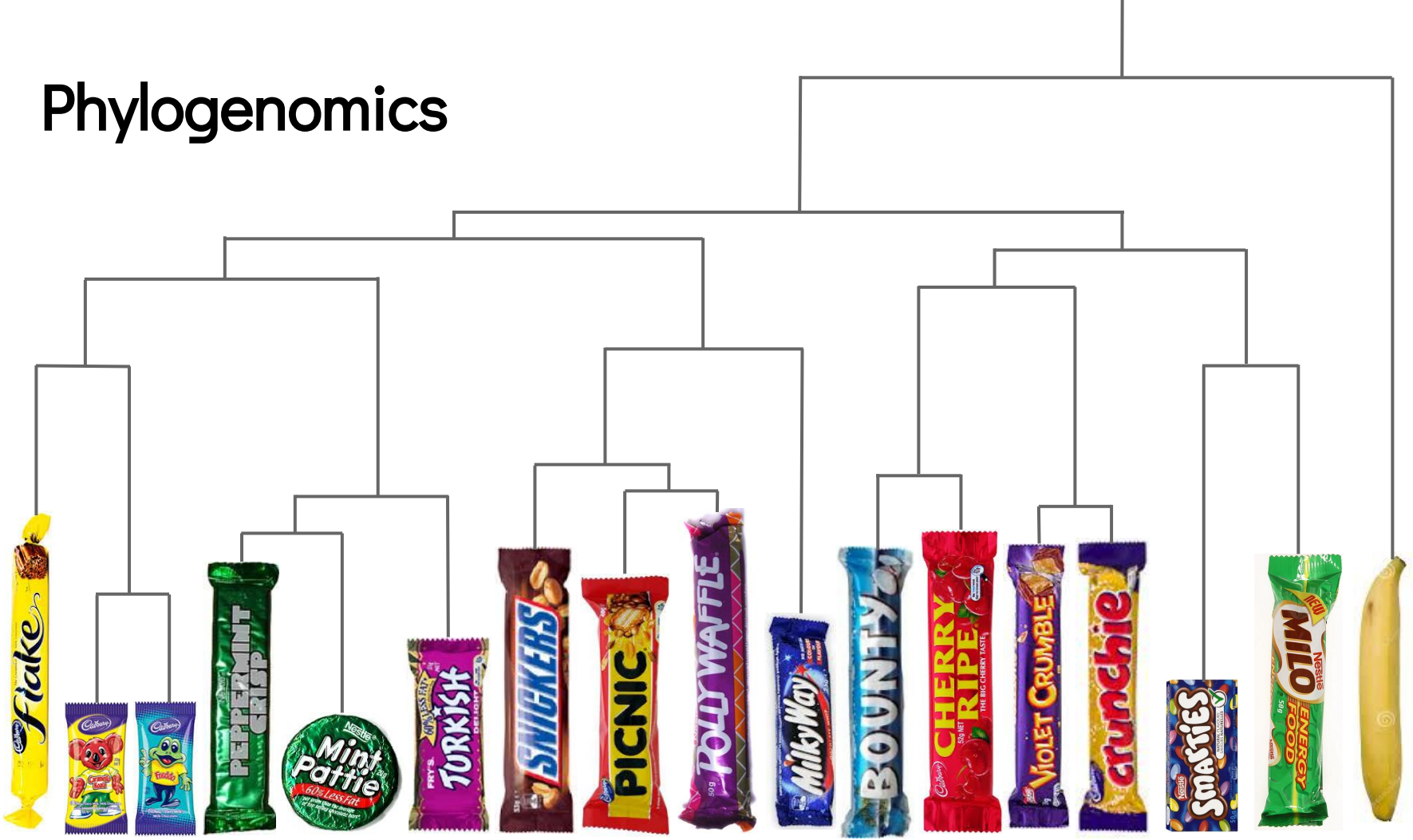
Finding differences

Genotyping / lineages



Antimicrobial resistance

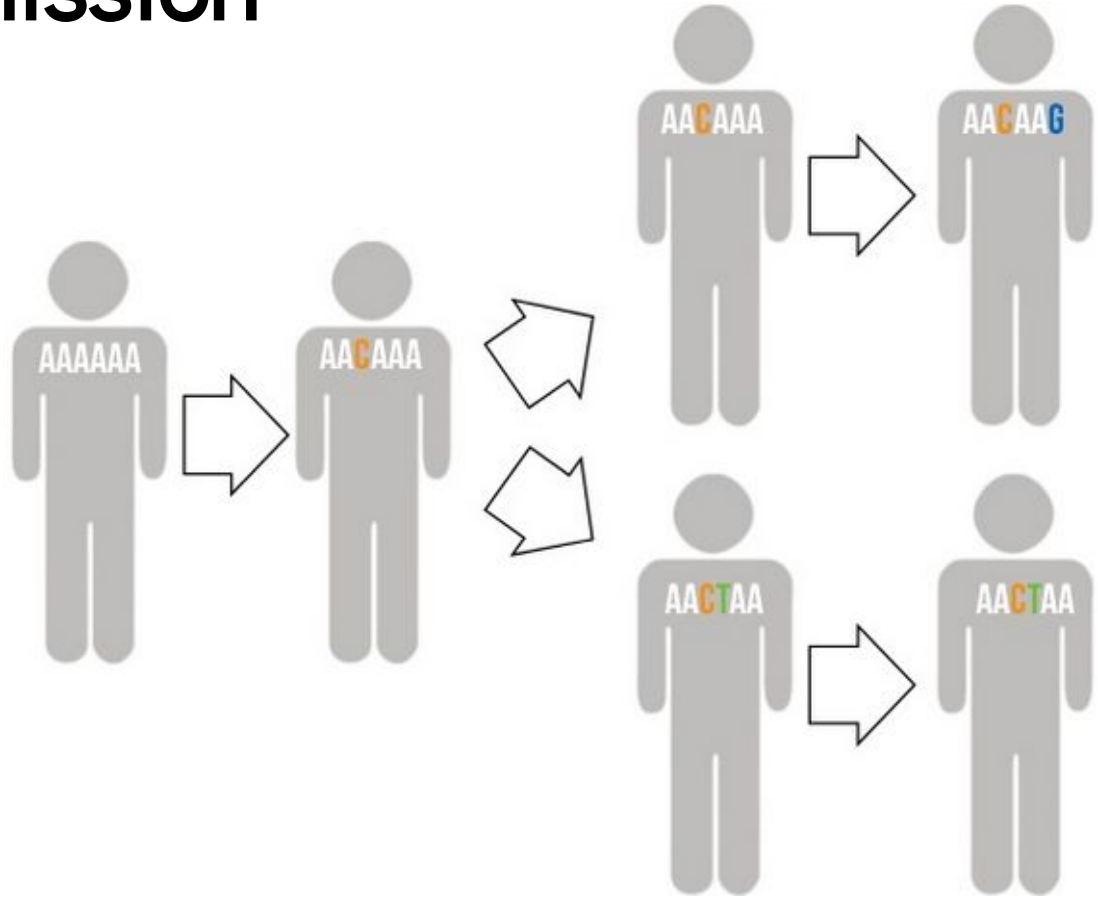
Phylogenomics



Inferring transmission

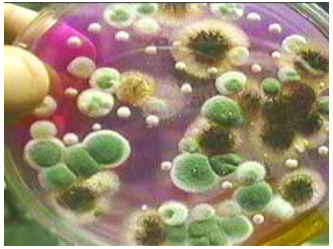
:: Identical sequence does not imply transmission

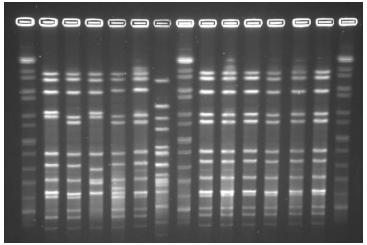
:: Easier to rule out than in

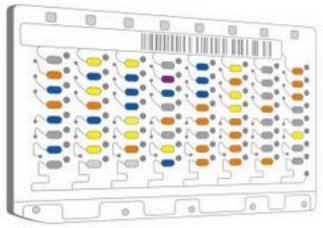


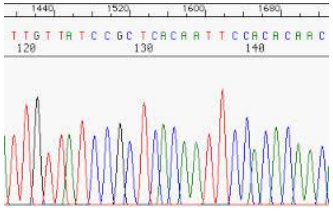
Conclusions

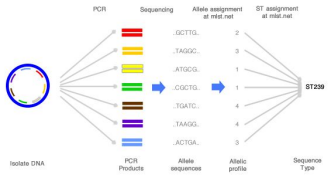
Traditional workflow

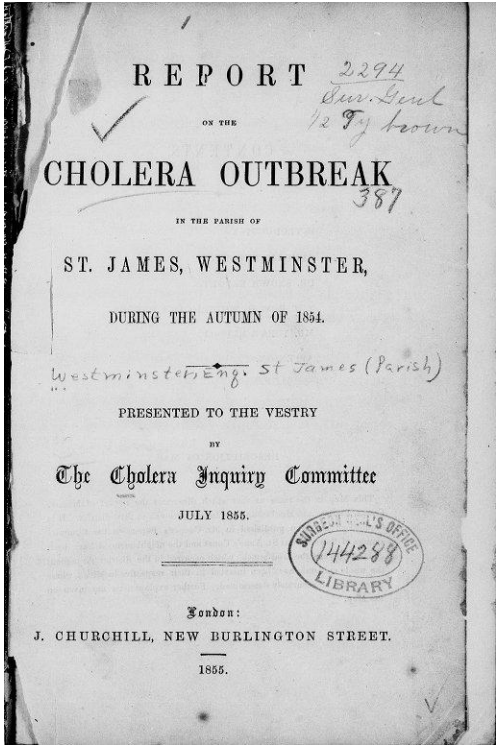




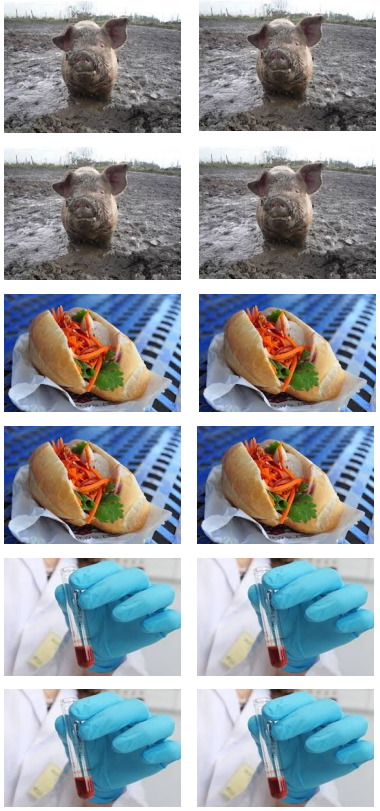








Modern workflow



Thank you for listening