What can we do with microbial WGS data?

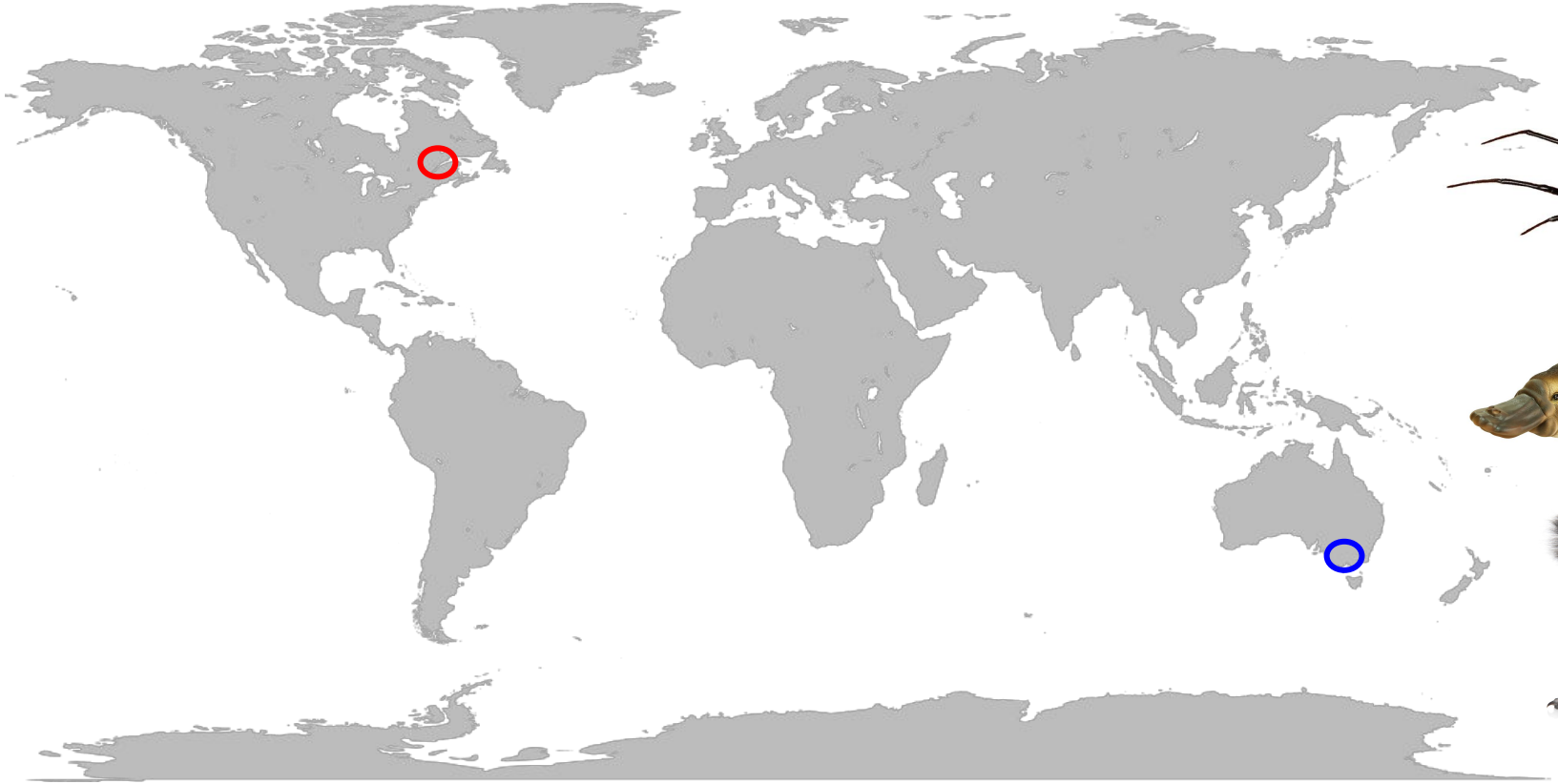
A/Prof Torsten Seemann

Victorian Life Sciences Computation Initiative (VLSCI)
Microbiological Diagnostic Unit Public Health Laboratory (MDU PHL)
Doherty Applied Microbial Genomics (DAMG)

The University of Melbourne

About me

Melbourne, Australia



Microbial genomics + bioinformatics

Doherty Centre for
Applied Microbial Genomics

Microbiological Diagnostic Unit

:: Oldest public health lab in Australia
  : established 1897 in Melbourne
  : large historical isolate collection back to 1950s
:: National reference laboratory
  : *Salmonella*, *Listeria*, EHEC
:: WHO regional reference lab
  : vaccine preventable invasive bacterial pathogens

Bacterial genomics

Small genome

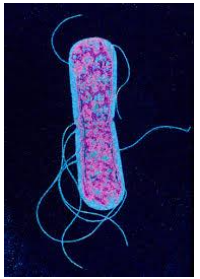
6,000,000,000

letters

30,000 genes

Genome

A T G C

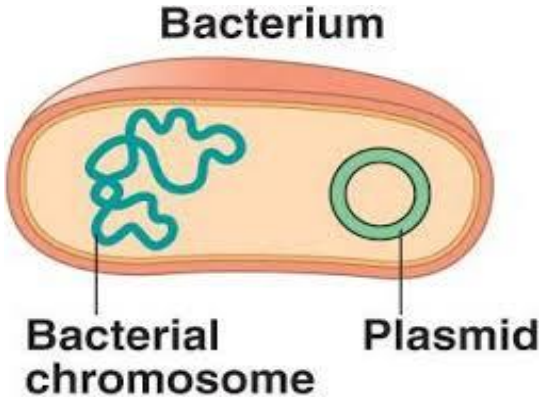


3,000,000

letters

3,000 genes

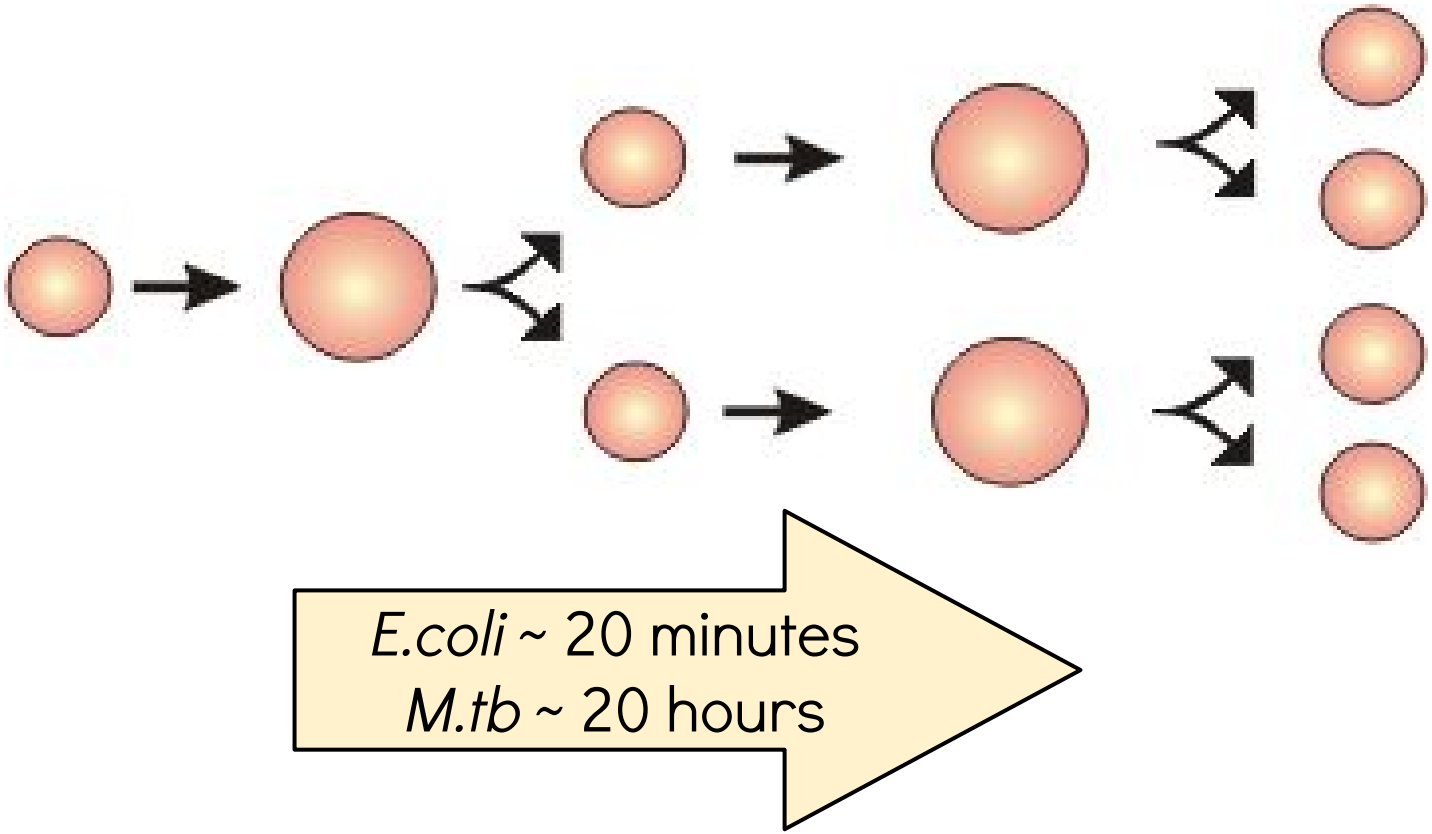
Replicons



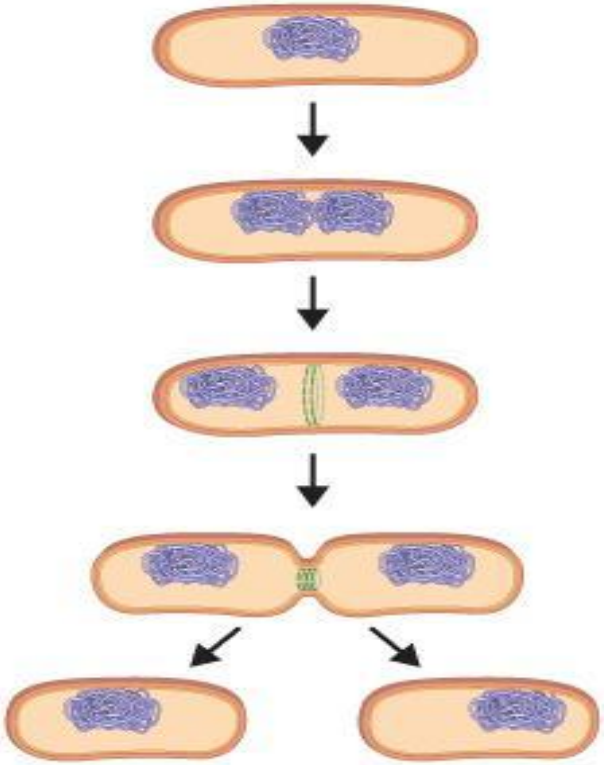
Usually 1 big
circular chromosome
(1M to 10M bases)

Sometimes 1-6
"mini" chromosomes
(4k - 300k bases)

(Relatively) fast growers



Vertical transfer of DNA



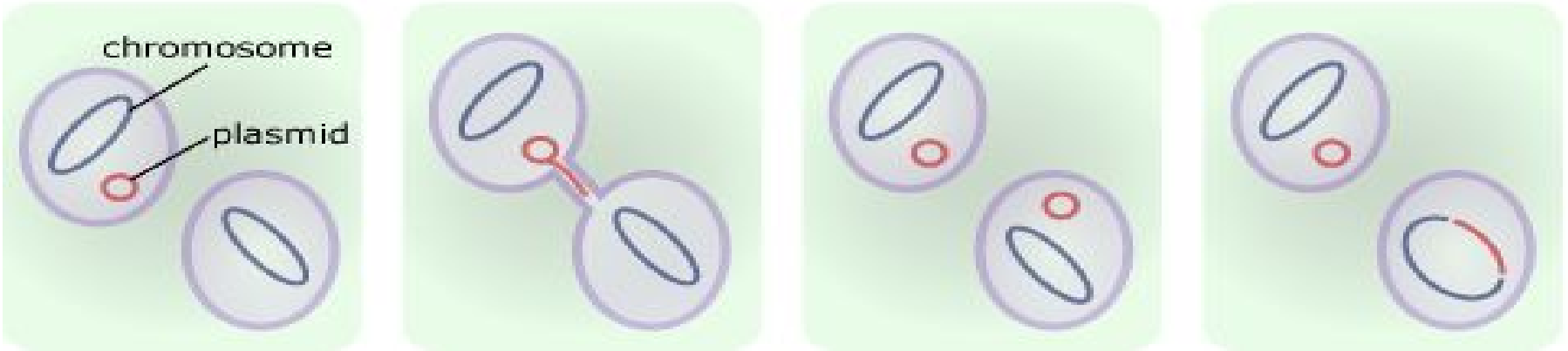
Occurs during cell division

Sometimes it makes an <u>error</u> copying the DNA

eg. A → T

Horizontal transfer of DNA

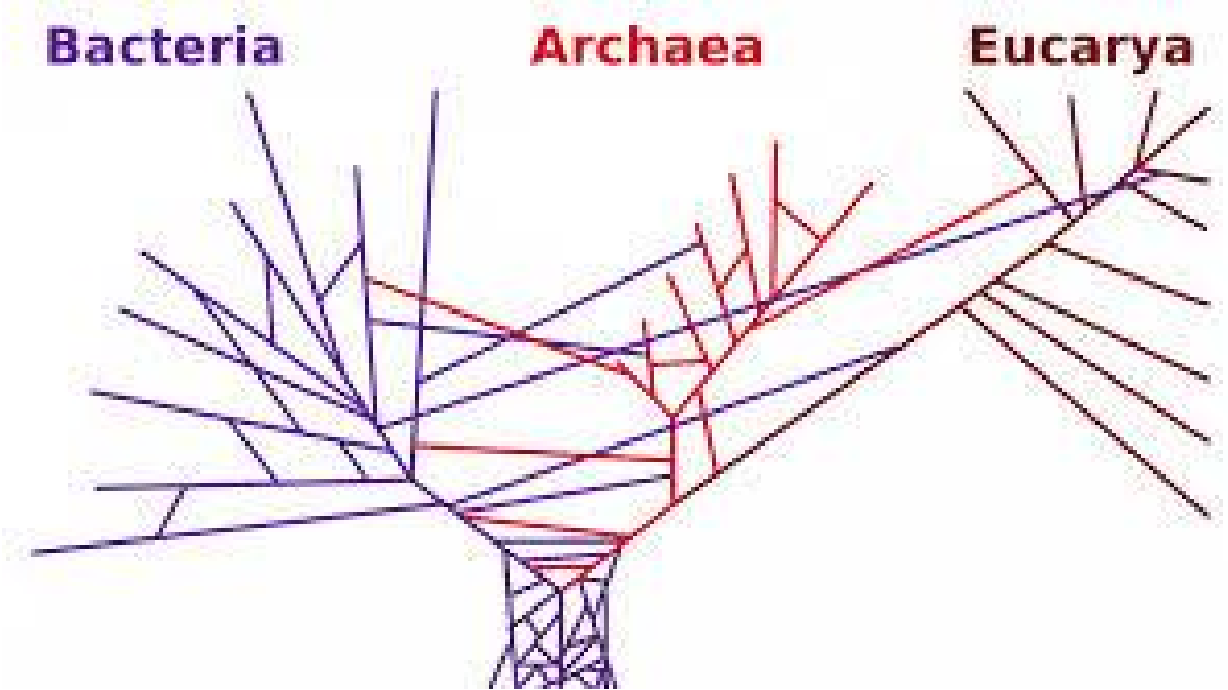
Occurs *between* bacterial cells



Plasmids: virulence & antibiotic resistance genes!

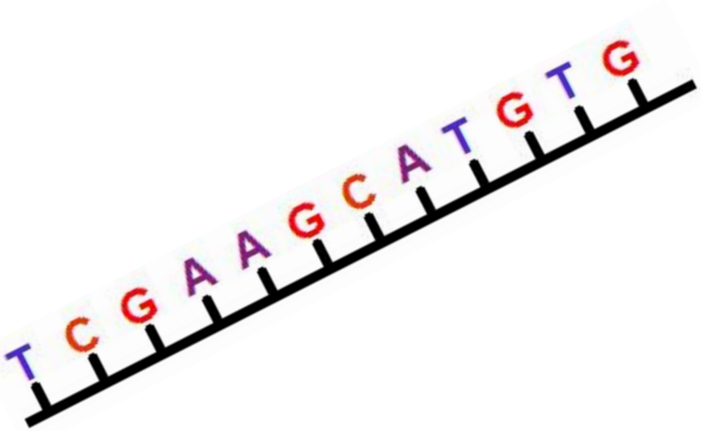
The *real* tree of life

A mixture of horizontal & vertical transmission.

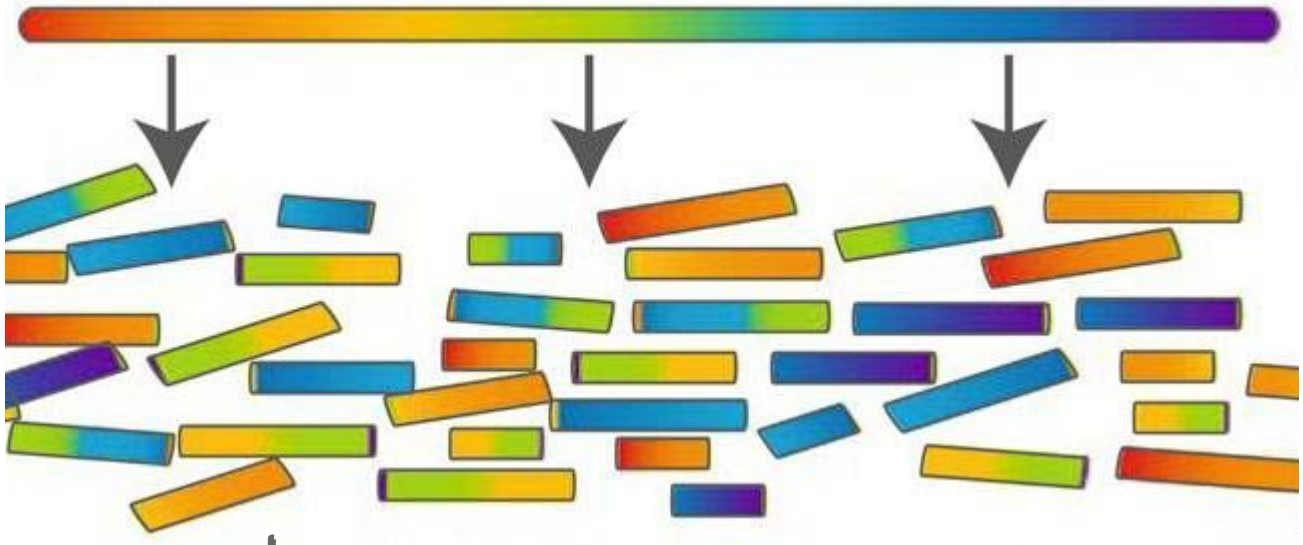


Whole genome sequencing (WGS)

In an ideal world



The real world (for now)



Genome

Shred

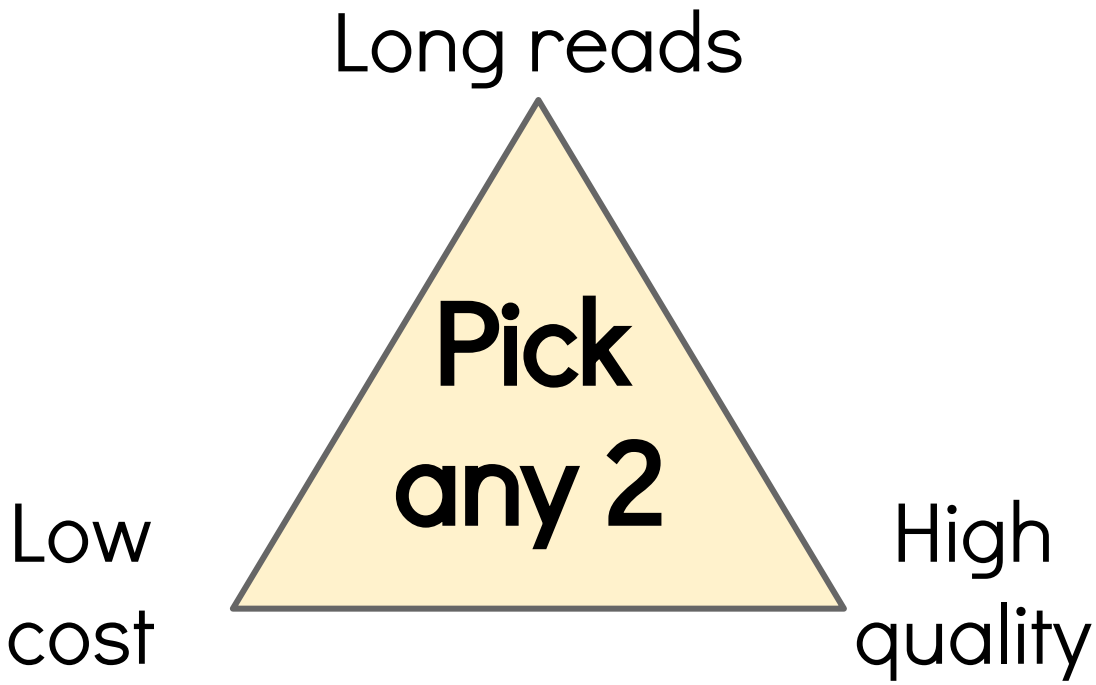
Sequence

Reads are stored in “FASTQ” files

WGS technologies

illumina®	100 - 300 bp
ion torrent	100 - 400 bp
PACIFIC BIOSCIENCES®	5,000 - 25,000+ bp
Oxford NANOPORE Technologies	5,000 - 150,000+ bp

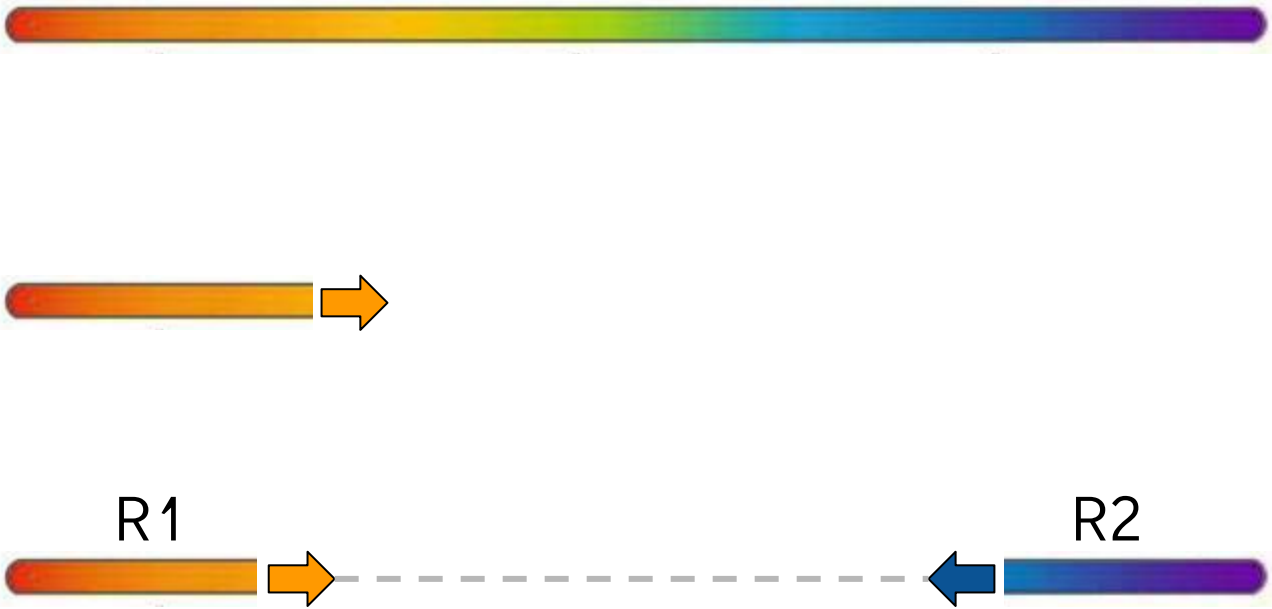
Which sequencing platform?



Workhorse for bacterial sequencing

illumina®	100 - 300 bp
ion torrent	100 - 400 bp
PACIFIC BIOSCIENCES®	5,000 - 25,000+ bp
Oxford NANOPORE Technologies	5,000 - 150,000+ bp

Types of reads

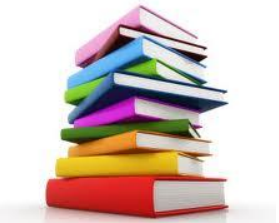


DNA fragment
~500 bp

"Single end"

"Paired end"

What you get back



Millions to billions of reads (big files):

```
ATGCTTCTCCGCCTTTAATTAAAATTCCATTTCGTGCACCAACACCCGTTCCTACCATAATAGCTGTTGGAGTCGCTAAACCTAATGCACATGGACACGC        <- 1st read
CTAAGATACTGCCATCTTCTTCCAACGTAAATTGTACGTGATTTTCGATCCATTTTCTTCGAGGTTCTACTTTGTCACCCATTAGTGTGGTTACTCGACG        <- 2nd read
GAATATGCGTGGACAGATGACGAATTGGCAGCAATGATTAAAAAAGTCGGCAAAGGATATATGCTACAGCGATATAAAGGACTTGGAGAGATGAATGCGG
ATCAATGCAAATACAAGATGTGACAATGCGCGCAATGCAATGATAACTGGTGTTGTCAAAAAGAAACCGAATGTCGTACCTAGTGCAACAGCCACTGCAA
GGAAAAAATGAGAAAAAATTCAGTTCGAAAACTAACGATTTCTGCTTTATTGATTGGGATGGGGGTCATTATCCCAATGGTTATGCCTAAAATCATGATC
GATGAAACAATCCAACAAATACCATTCAATAATTTCACAGGGGAAAATGAGACNCTAAGTTTCCCCGTATCAGAAGCAACAGAAAGAATGGTGTTTCGCT
...
...                                    <--- 100 to 300 bp --->
...
AATTGGACTTTGTCACCGATTTTCAGTTCATCTATGTCCACGCTTATTTTTTCAGCAGTAGCATTCAAAATCACTCCGTCATTGCTGAATGATGTCCCCA
CTCCTGTTTCTTTATCTATAATTGAACTGTAAACATGAGGAATCACTTTTTTTACACCTGCATCGATTGCAATTTTCAGAATTTCTTCAAAGTTTGAAAG
AAACTGCCATTCAAATGCTGCAAGACATGGGAGGTACTTCAATCAAGTATTTCCCGATGAAAGGCTTAGCACATAGGGAAGAATTTAAAGCAGTTGCGGA
ATCATTCCTACGCCAGTCATTTCGCGTAGTTCTTTTACCATTTTAGCTGTAACGTCTGCCATGTTTAACTCCTCCTGTGTGTGTTCTTTTTAAAAAAAGC        <- last read
```

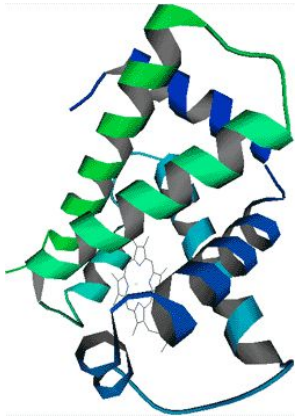
Applications of WGS

- Diagnostics
  - species ⇒ subspecies ⇒ strain identification
  - *in silico* antibiogram and virulence profile
- Surveillance
  - *in silico* genotyping - MLST, serotyping, VNTR, MLVA
  - what's lurking in our hospital/community?
- Forensics
  - outbreak detection & source tracking
  - phylogenomics

Not just genomes!

If you can transform your assay into sequencing lots of short pieces of DNA, then NGS is applicable.

- exome (targeted subsets of genomic DNA)
- RNA-Seq (transcripts via cDNA)
- ChIP-Seq (protein:DNA binding sites)
- HITS-CLIP (protein:RNA binding sites)
- methylation (bisulphite treatment of CpG)

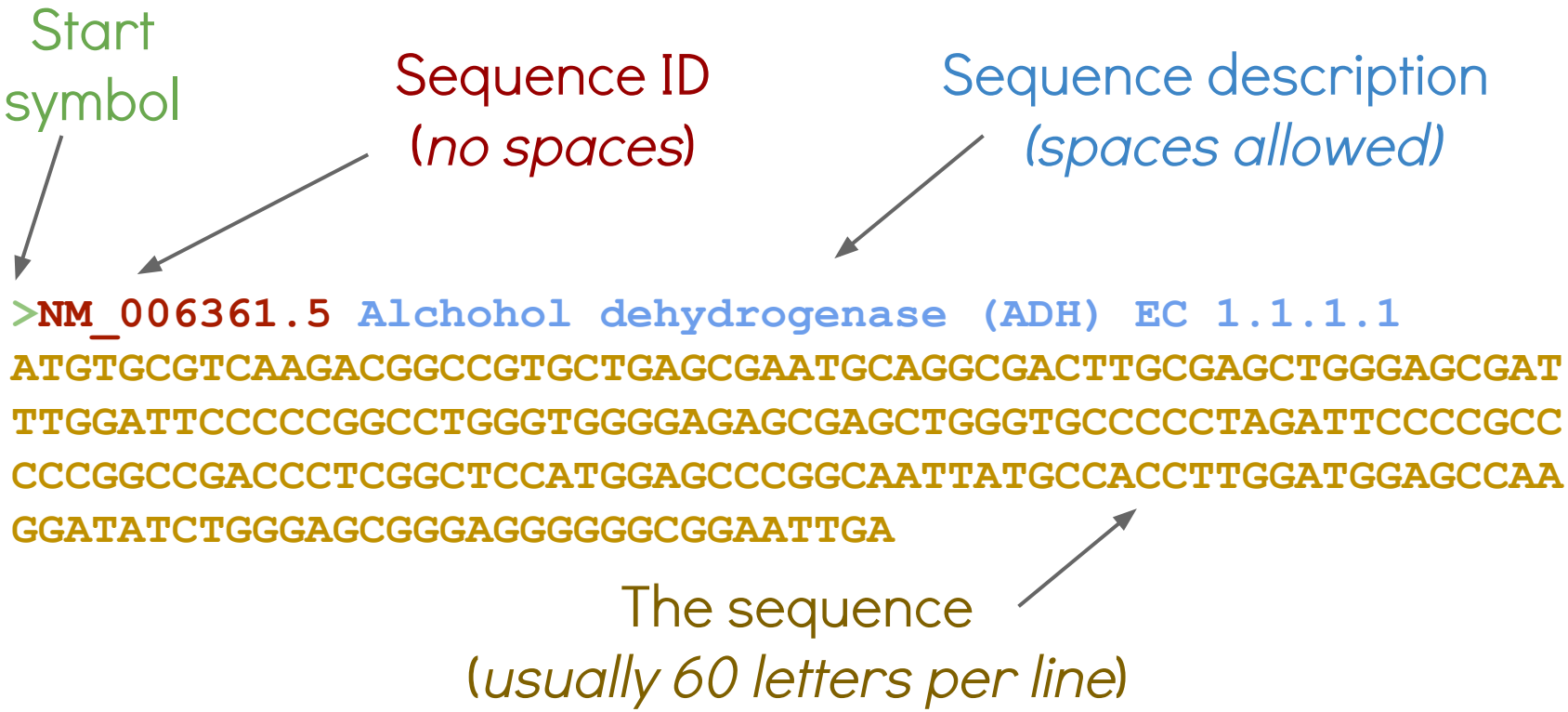


FASTA sequence file format

FASTA

```
>NM_006361.5 Alchohol dehydrogenase (ADH) EC 1.1.1.1
ATGTGCGTCAAGACGGCCGTGCTGAGCGAATGCAGGCGACTTGCGAGCTGGGAGCGAT
TTGGATTCCCCCGGCCTGGGTGGGGAGAGCGAGCTGGGTGCCCCCTAGATTCCCCGCC
CCCGGCCGACCCTCGGCTCCATGGAGCCCGGCAATTATGCCACCTTGGATGGAGCCAA
GGATATCTGGGAGCGGGAGGGGGGCGGAATTGA
```

FASTA components



Multi-FASTA

Concatenation of individual FASTA entries, using ">" as an entry separator

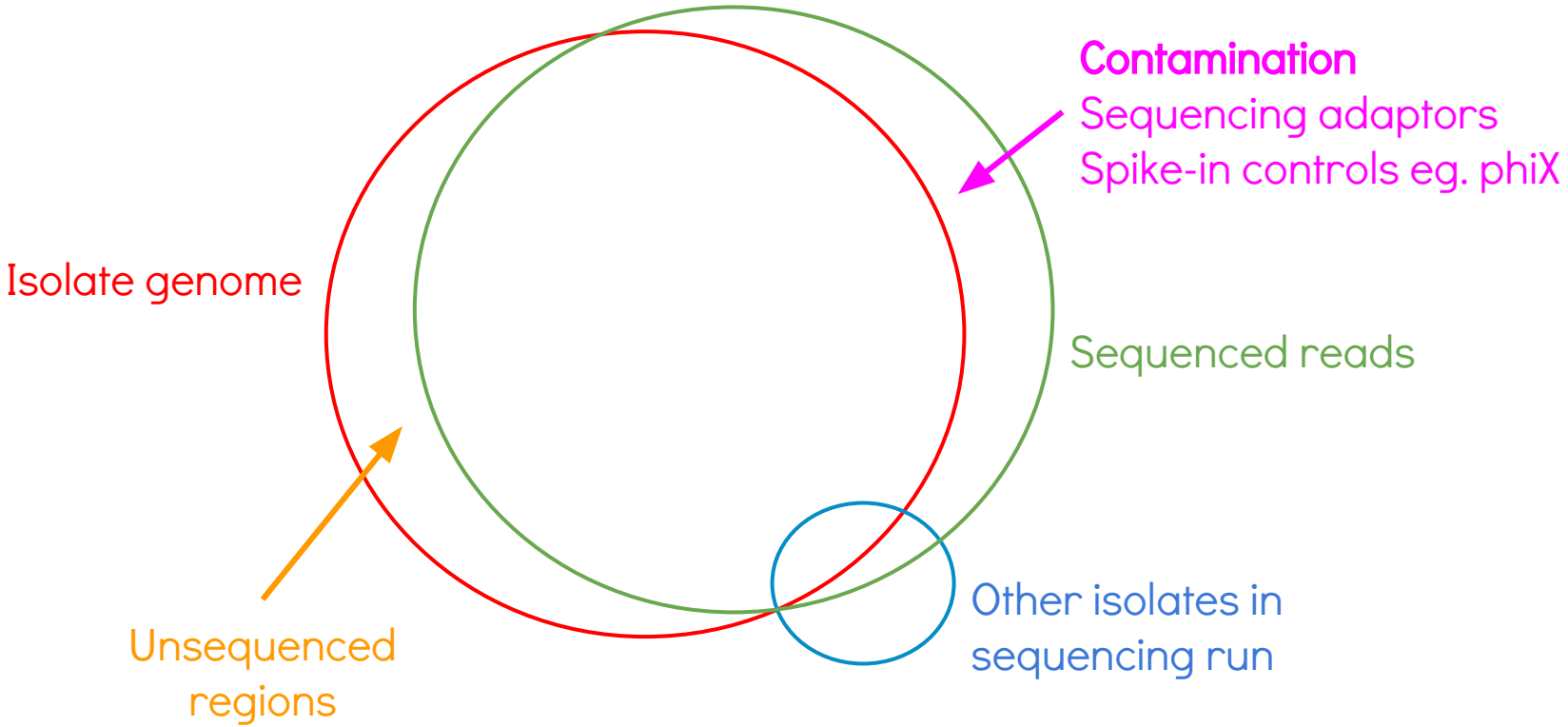
```
>read00001
TCTTGCGTCAAGACGGCCGTGCTGAGCGAATGCAGGCGACTTGCGAGCTGGGAGCGA
>read00002
TGGATTCCCCCGGCCTGGGTGGGGAGAGCGAGCTGGGTGCCCCCTAGATTCCCCGCC
>read00003
GGCCGACCCTCGGCTCCATGGAGCCCGGCAATTATGCCACCTTGGATGGAGCCAAGG
>read00004
TCTGGGAGCGGGAGGGGGGCGGAATCTGGAGCGAGCTGGGTGCCCCCTAGATTCCCC
>read00004
GCGGAATCTGGAGCGAGCTGGGTGCCCCCTAGATTCCCCGCATCGTAGATTAGATAT
```

The DNA alphabet

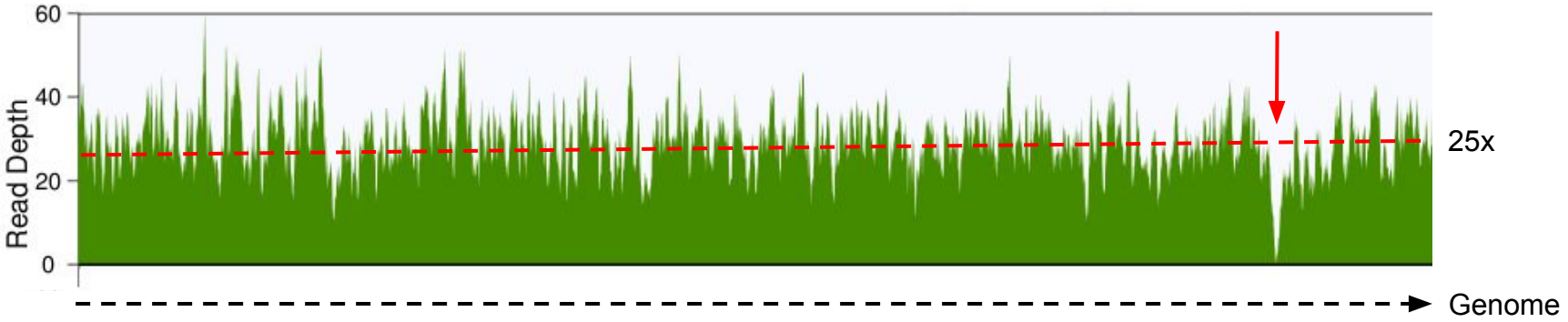
- Standard
  - A G T C
- Extended
  - adds N (unknown base)
- Full
  - adds R Y M S W K V H D B (ambiguous bases)
  - R = A *or* G (puRine)
  - Y = C *or* T (pYrimidine)
  - *... and so on for all the combinations*

Sequence Quality

What data do we really have?



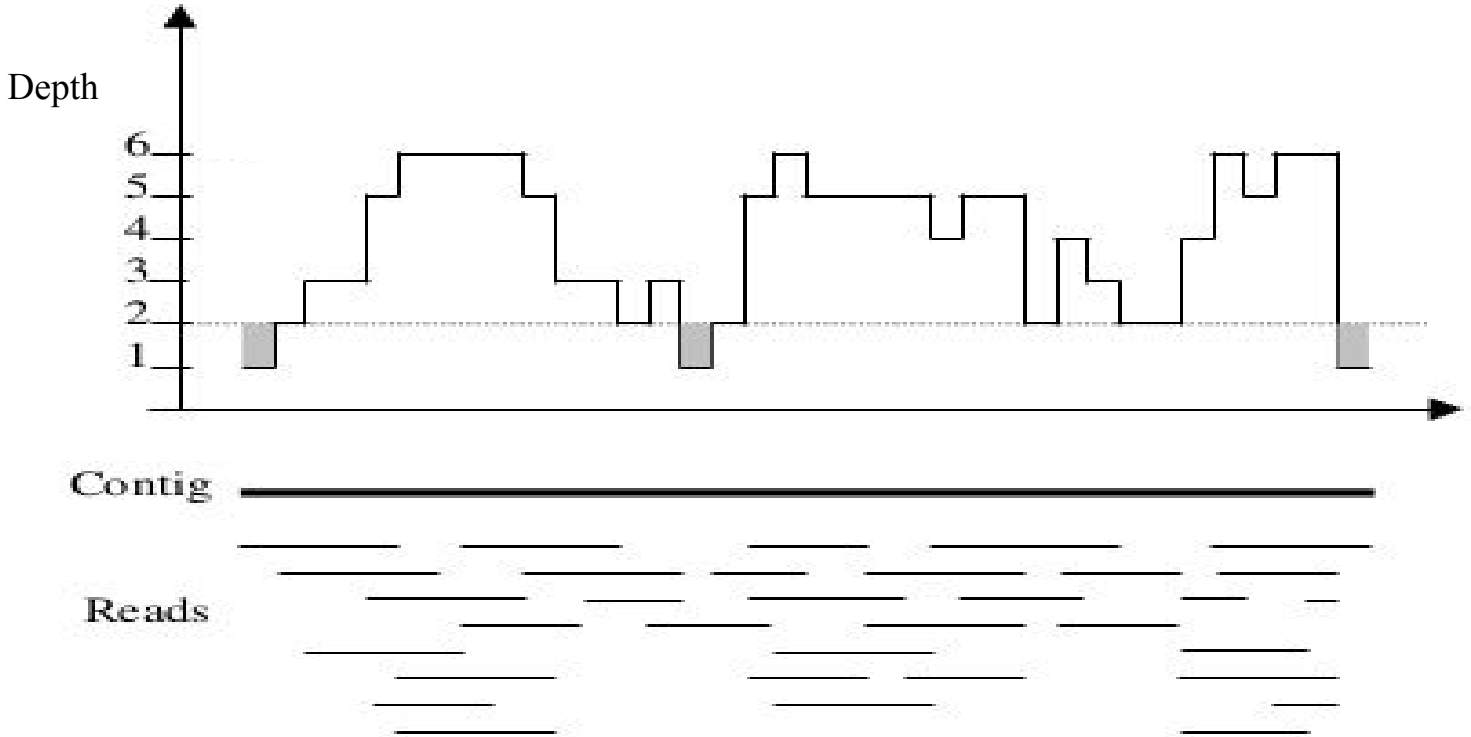
Do we have enough data?



:: Depth
  : expressed as fold-coverage of genome *eg.* 25x
  : means each base sequenced 25 times (on average)

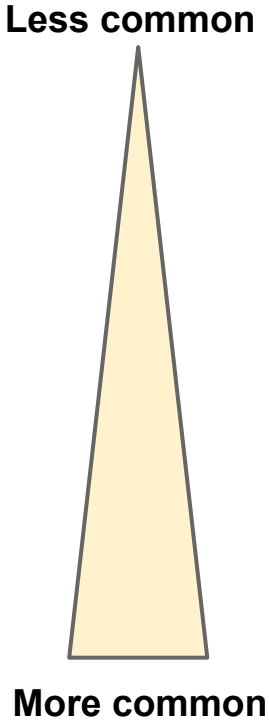
:: Coverage
  : the % of genome sequenced with depth > 0

Depth (of coverage)



Sequences have errors

- nonsense reads — *instrument oddness*
- duplicate reads — *amplify a low complexity library*
- adaptor read-through — *fragment too short*
- indel errors — *skipping bases, inserting extra bases*
- uncalled base — *couldn't reliably estimate, replace with "N"*
- substitution errors — *reading wrong base*



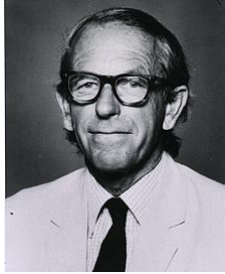
Illumina reads

- Usually 100 - 300 bp
- Indel errors are rare
- Substitution errors < 1%
  - Error rate higher at 3' end
- Very high quality overall

DNA base quality

- DNA sequences often have a *quality value* associated with each nucleotide
- A measure of reliability for each base
- Derived from a physical process
  - chromatogram (Sanger sequencing)
  - pH reading (Ion Torrent sequencing)
  - Voltage measurement (Oxford Nanopore)

“Phred” qualities

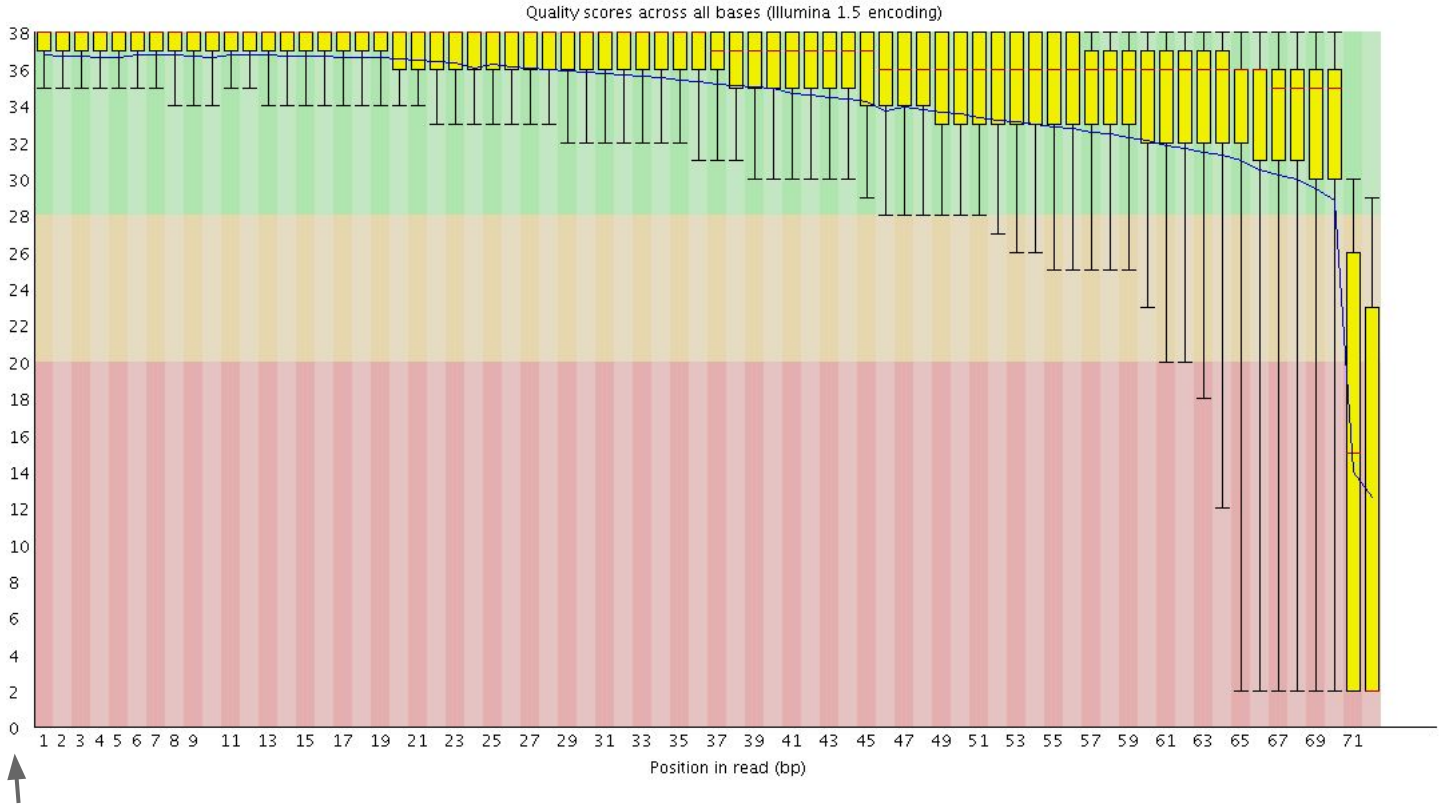


Quality	Chance it's wrong	Accuracy	Description
10	1 in 10	90%	Bad
20	1 in 100	99%	Maybe
30	1 in 1000	99.9%	OK
40	1 in 10,000	99.99%	Very good
50	1 in 100,000	99.999%	Excellent

$$Q = -10 \log_{10} P \quad <=> \quad P = 10^{-Q/10}$$

Q = Phred quality score P = probability of base being incorrect

Quality plot (*FastQC*)



Y-axis is "Phred" quality values (higher is better)

Quality filtering

- Keep all reads
  - let the downstream software cope
- Reject some reads
  - average quality below some threshold
  - contain any ambiguous bases
- Trim reads
  - remove low quality bases from end(s)
  - keep longest "sub-read" that is acceptable
- Best strategy is analysis dependent

FASTQ files

FASTQ

FASTQ sequence entry looks like this:

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

FASTQ components

Start symbol

Sequence ID

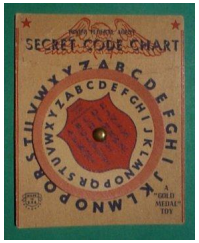
Sequence

Separator line

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

Encoded quality values, one symbol per nucleotide

FASTQ quality encoding



Uses letters/symbols to represent numbers:

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

Multi-FASTQ

Same as multi-FASTA, just concatenate:

@M00267:3:15997:1501

CTCGTGCTCTACTTTAGAAGCTAATGATTCTGTTTGTAGAACATTTTCTACCACTACATCTTTTTCTTGCTTCGCATCTT

+

:=?DD:BDDF>FFHI>E>B9AE>4C<4CCAE+AEG3?EAGEHCGIIIIIIIIIIIIGIIIEIIIIGGIDGIID/;4C<EE

@M00267:3:15997:1505

GCCTATAGTAGAAGAAAAAGAAGTGGCTCAAGAAATGAGTGCACCGCAGGAAGTTCCAGCGGCTGAATTACTTCATGAAA

+

<@@FFF?DHFHGHIIIFGIIGIGICDGEGCHIIIIIIIIIGIHIIFG<DA7=BHHGGIEHDBEBA@CECDD@CC>CCCAC

@M00267:3:14073:1508

GTCTTGCTAAATTTAAATAATCTGAAATAATTTGTTCTGCCCGGTCCAATTCAGCTAATACGAGACGCATATAATCCTTA

+

:?DDDDD?84CFHC><F>9EEH>B>+A4+CEH4FFEHFHIIIIIIIIIIIIIGGIIIIIIIIG>B7BBEBBB@CDDCFC

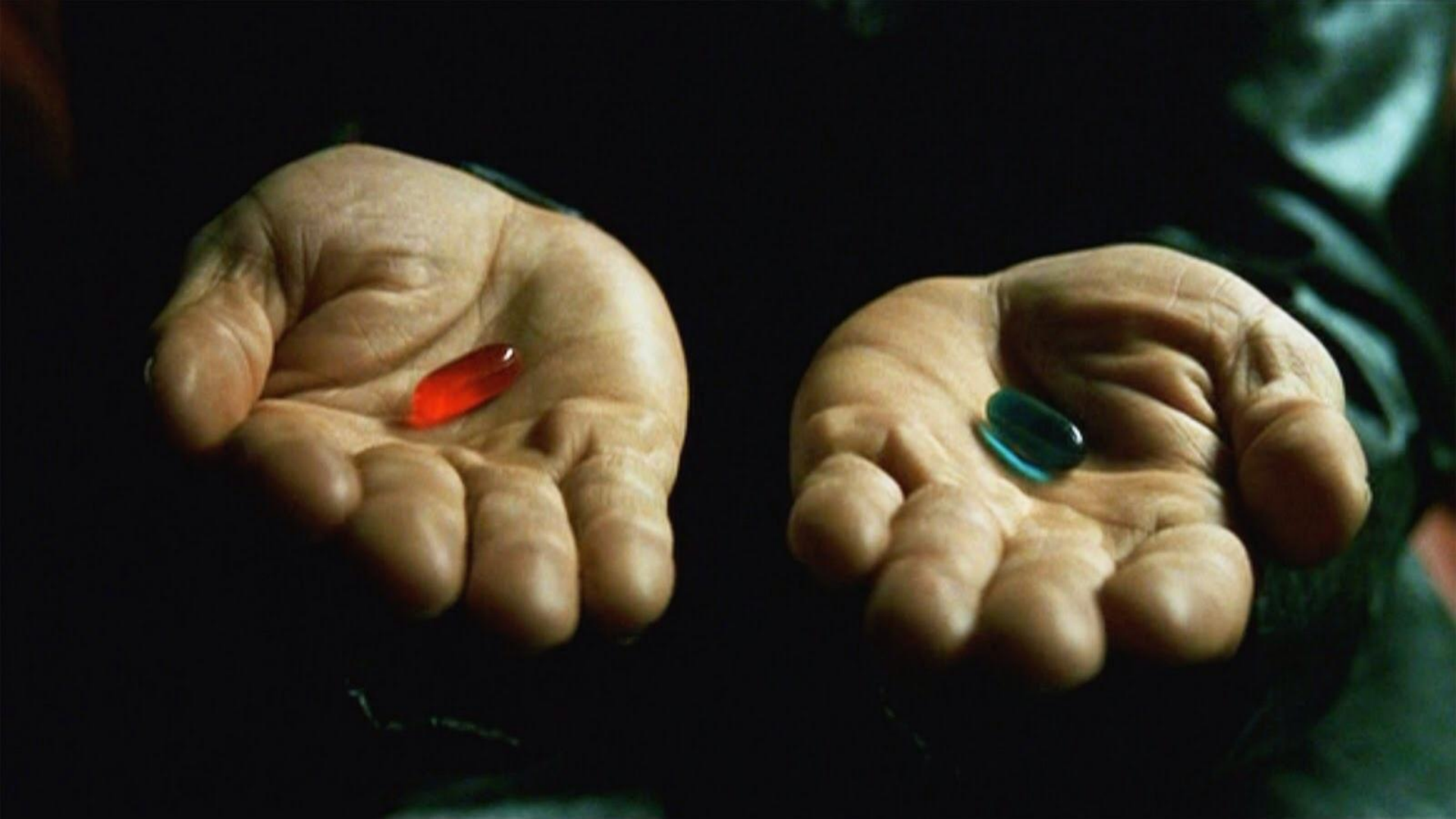
@M00267:3:14073:1513

ACGTACAGAGATGCAAAAGTCAGAGAAACTTAATATTGTAAGTGAGTTAGCAGCAAGTGTTGCACATGAGGTTCGAAATC

+

1@@DDDADHGDF?FBGGAFHHCHGGCGGFHIECHGIIGIGFGHGHIIHHEGCCFCB>GEDF=FCFBGGGD@HEHE9=;AD

Got my reads, now what?



Two options

:: *De novo* genome assembly
   : reconstruct original sequence from reads alone
   : like a giant jigsaw puzzle
   : *Create*

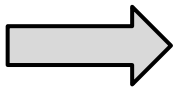
:: Align to reference
   : identify where each read fits on a related genome
   : can not always be uniquely placed
   : *Compare*

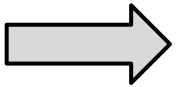
Genome assembly

(the red pill)

De novo genome assembly

De novo = without reference to other genomes





De novo assembly

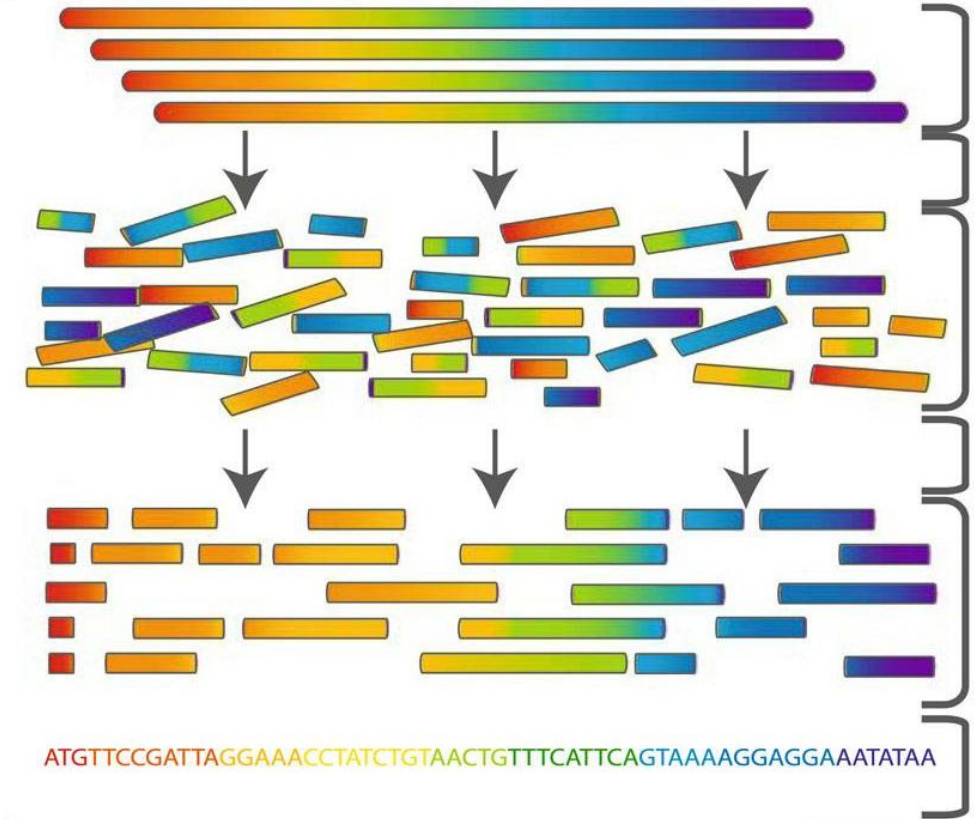
Reconstruct the original genome sequence from the sequence reads only

Millions of short sequence (**reads**)

A few long sequences (**contigs**)

Ideally, one sequence per chromosome.

Overlap - Layout - Consensus



Amplified DNA

Shear DNA

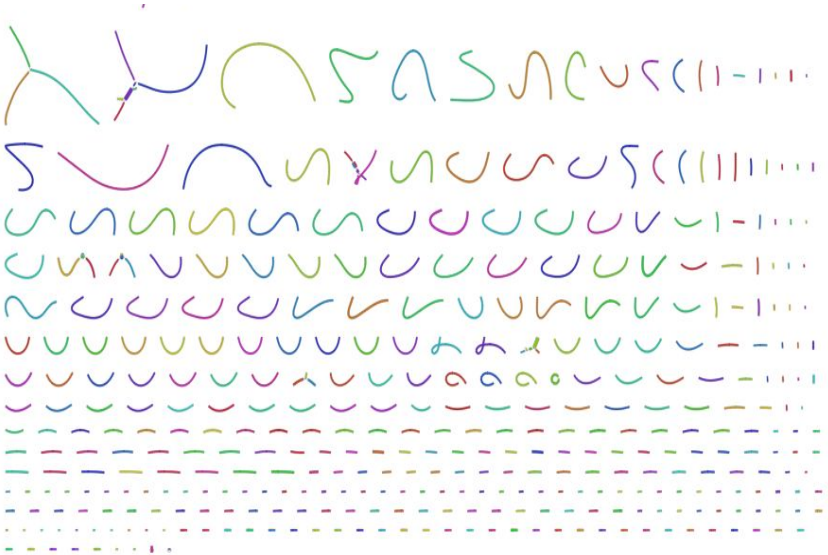
Sequenced reads

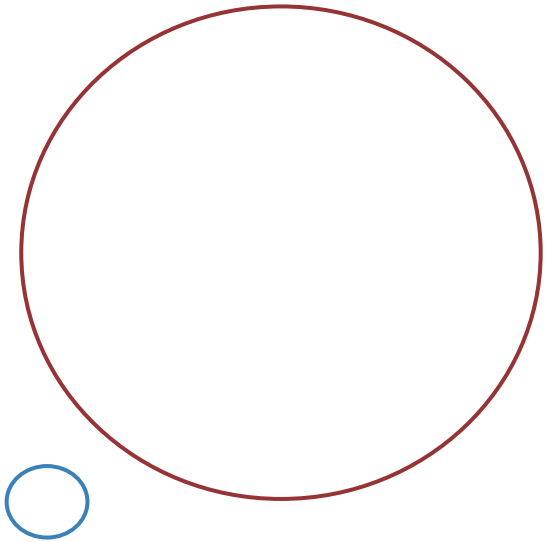
Overlaps

Layout

Consensus ↠ “Contigs”

Draft vs Finished genomes

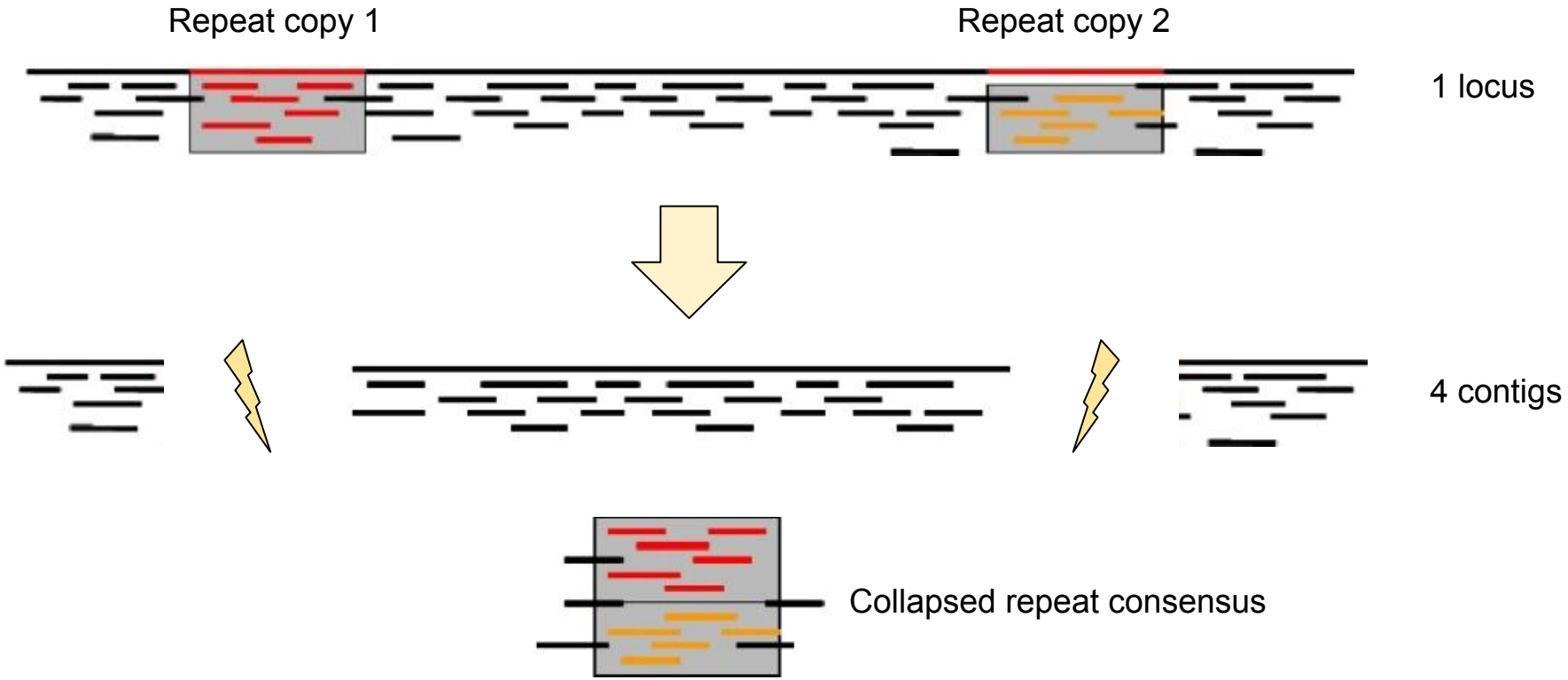




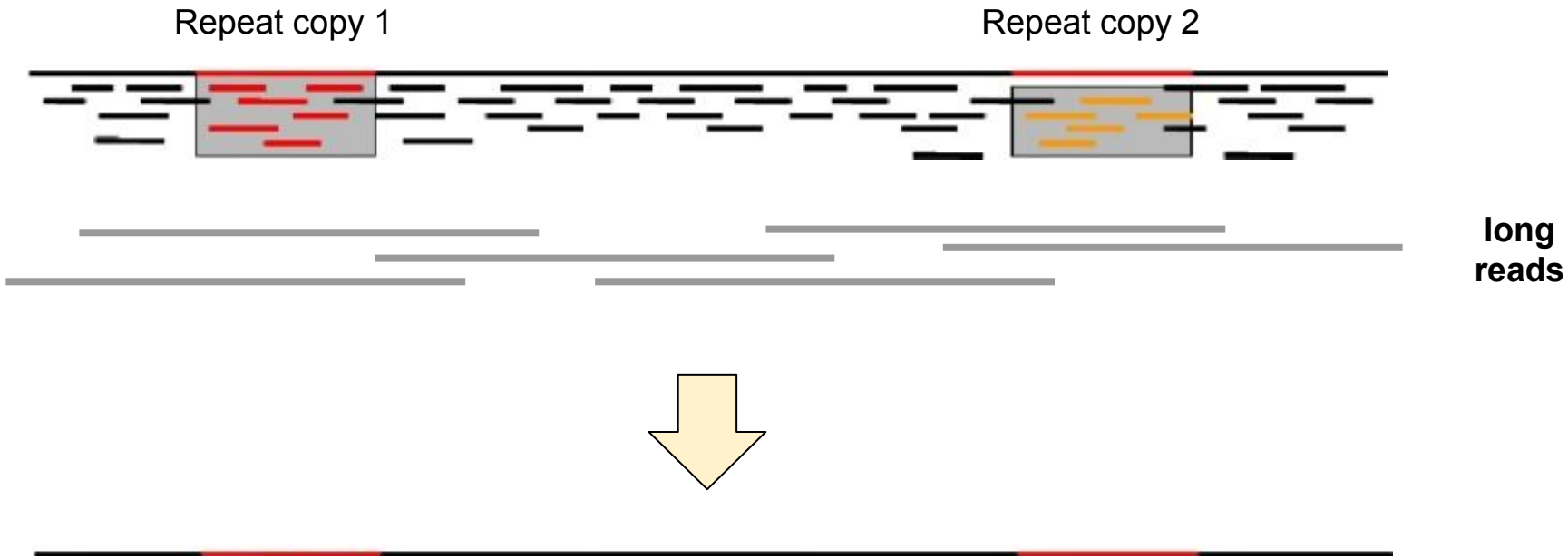
250 bp - Illumina - $250

8000 bp - Pacbio - $2500

Repeats


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

Long reads can span repeats



The law of repeats

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

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

Align to reference

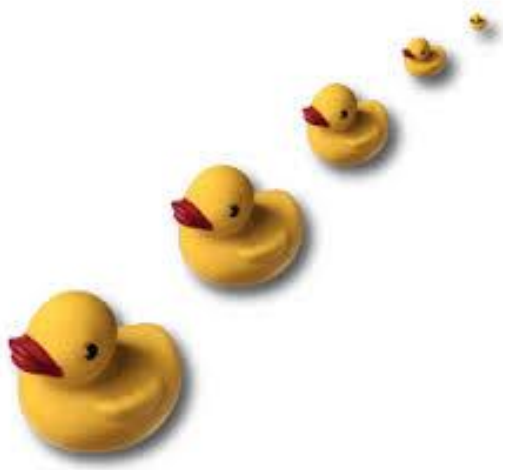
(the blue pill)

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

Look at differences

AGTCTGATTAGCTTAGCTTGTAGCGCTATATTAT Reference

```
AGTCTGATTAGCTTAGAT
      ATTAGCTTAGATTGTAG
           CTTAGATTGTAGC-C
    TGATTAGCTTAGATTGTAGC-CTATAT
       TAGCTTAGATTGTAGC-CTATATT
           TAGATTGTAGC-CTATATTA
           TAGATTGTAGC-CTATATTAT
```

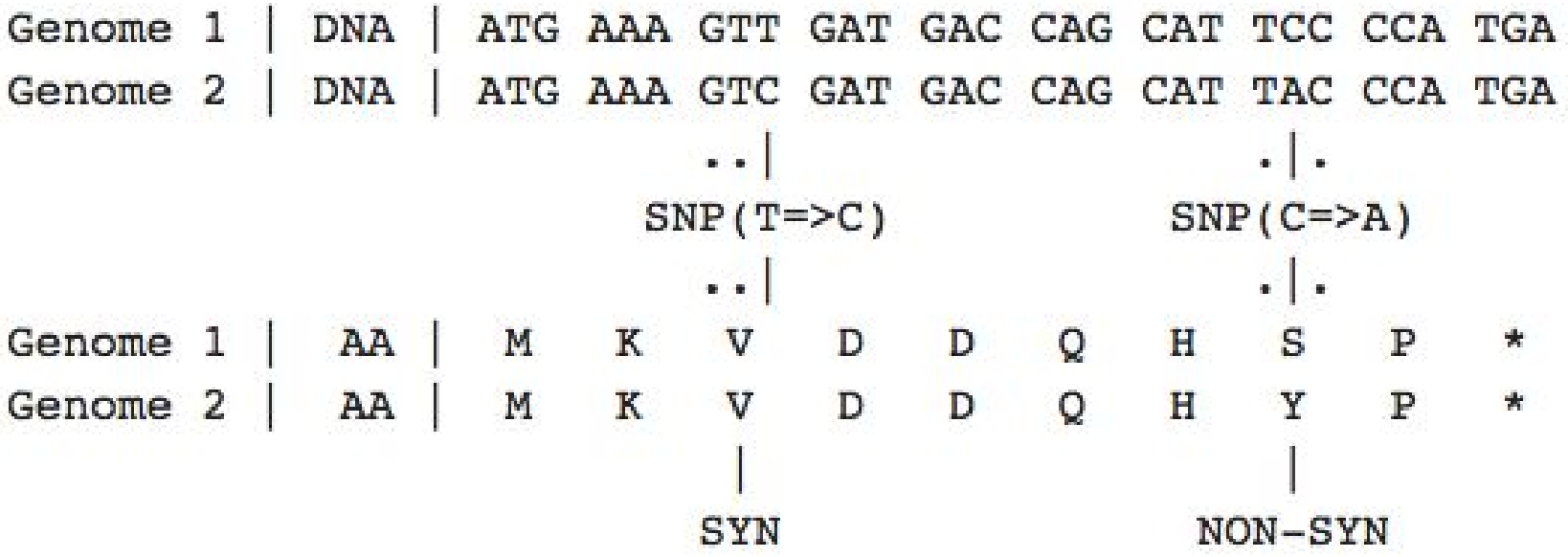
Reads

Look at differences

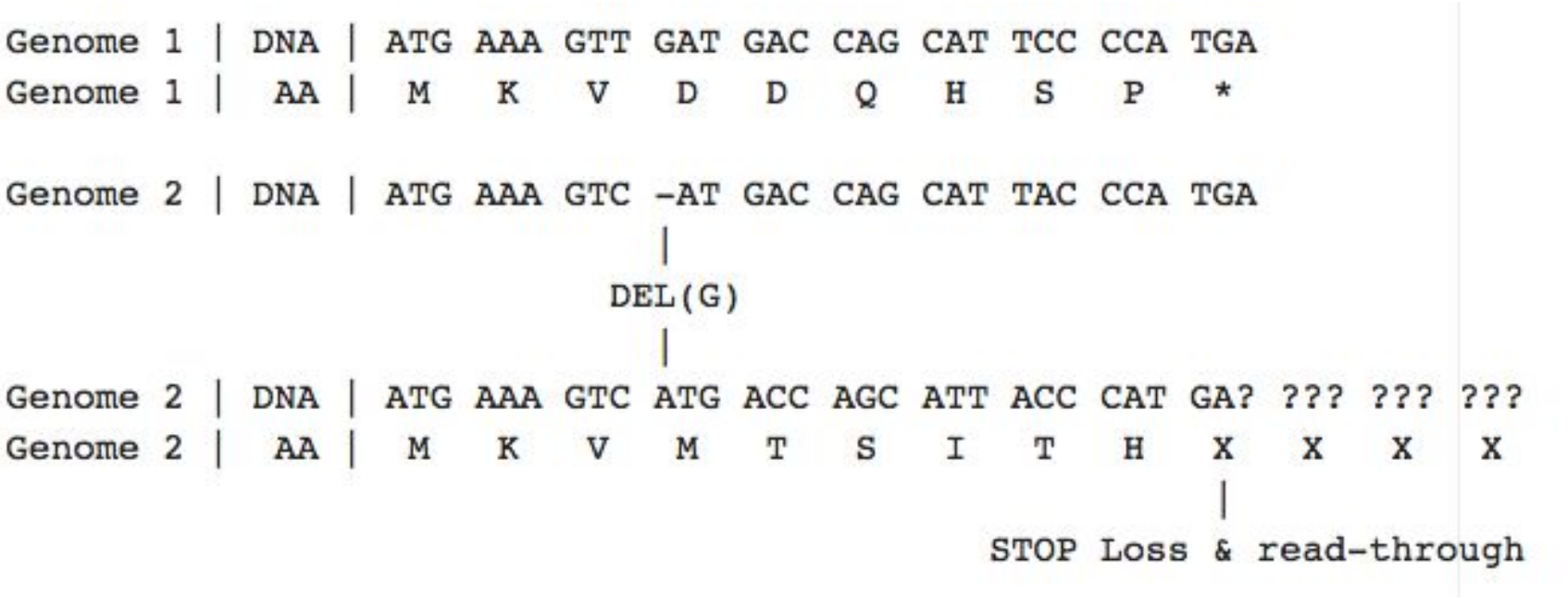
Types of variants we can detect

Type	Reference	Alternate
SNP (single)	T	G
MNP (multiple)	TA	GC
Insertion	AGT	A**C**GT
“	ATCGGG	ATC**TGA**GGG
Deletion	ACGT	AGT
“	ATCTGAGGG	ATCGGG

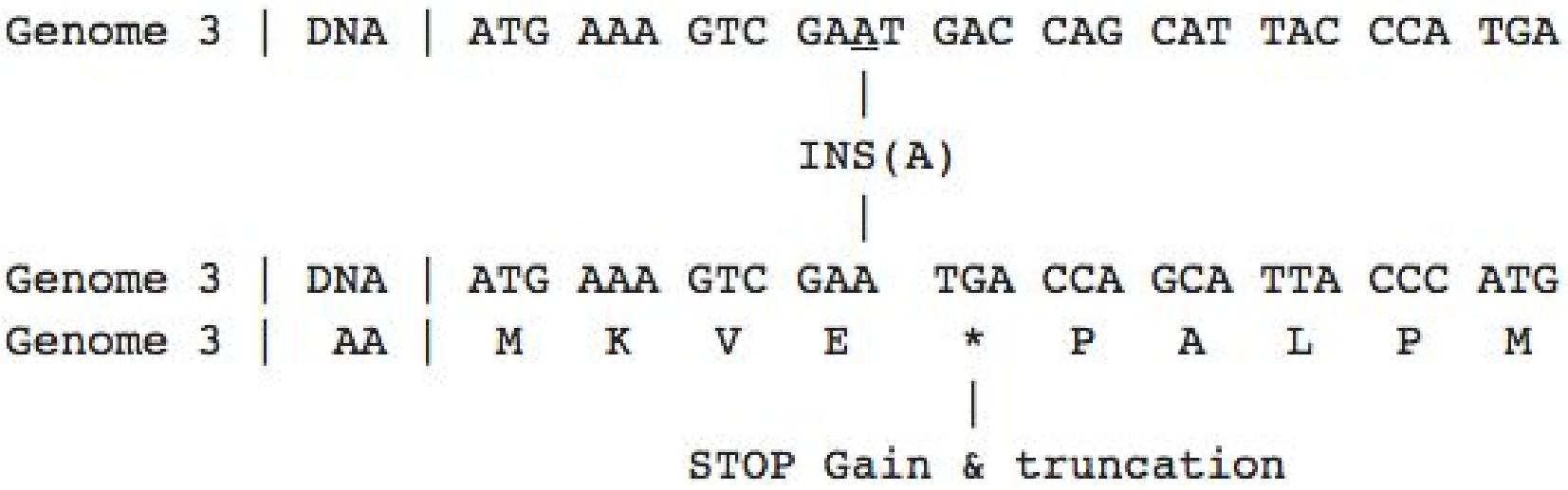
Synonymous & non-synonymous SNPs



Indel causing frame-shift



Indel causing truncation



Best practice

- Use both approaches
  - reference-based + *de novo*
- Best of both worlds
  - and worst of both worlds - interpretation is non-trivial
- Still need
  - good epidemiology, metadata and domain knowledge!

Downstream bioinformatics

Correcting contigs

- Pacbio - long reads
  - ~10 x 1bp homopolymer insertion errors per Mbp
  - Cause frame-shift errors in coding genes
- Illumina - short reads
  - Don't suffer from indel issues
- Use Illumina to correct Pacbio!
  - It works well

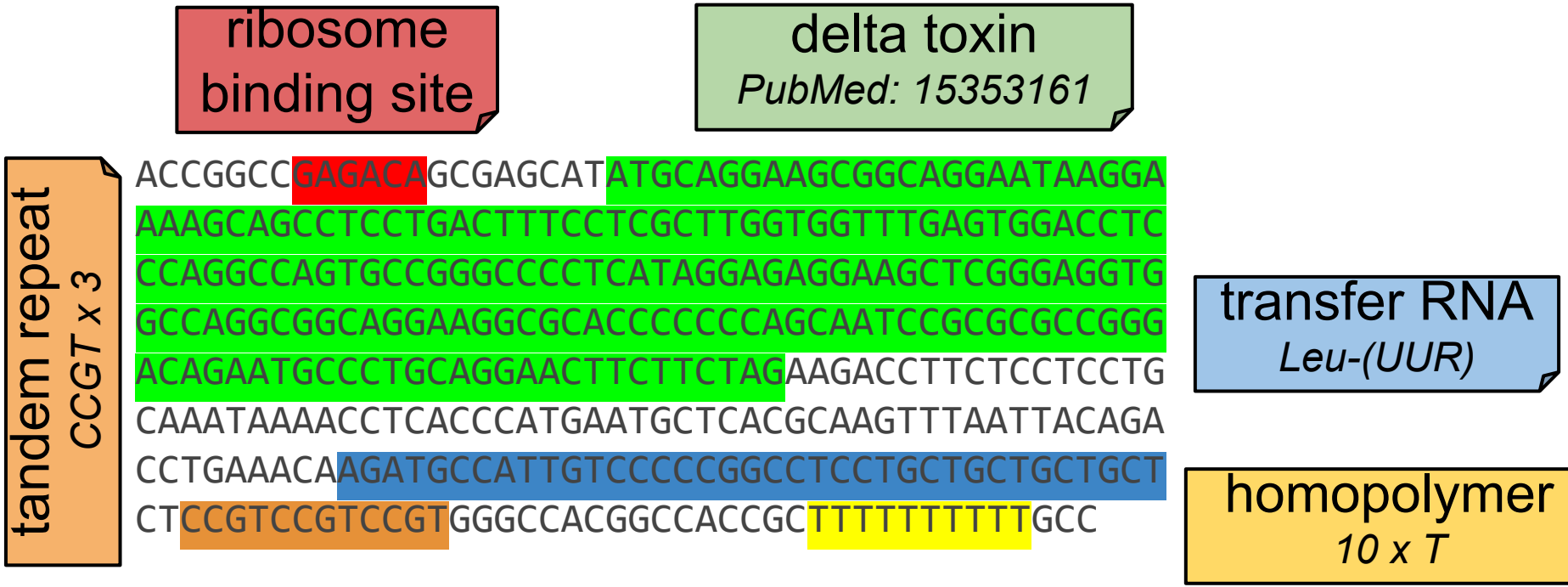
Species identification

Name	Readcount (% of classified reads)
Mycobacterium abscessus	959 (11.65%)
Vibrio cholerae	937 (11.38%)
Salmonella enterica subsp. enterica	738 (8.96%)
Enterobacter cloacae subsp. cloacae NCTC 9394	704 (8.55%)
Rhodobacter sphaeroides 2.4.1	613 (7.44%)
Staphylococcus aureus	567 (6.89%)
Klebsiella	563 (6.84%)
Bacillus cereus ATCC 10987	434 (5.27%)
Bacillus cereus	347 (4.21%)
Rhodobacter sphaeroides	334 (4.06%)
(Remaining organisms)	2038 (24.75%)

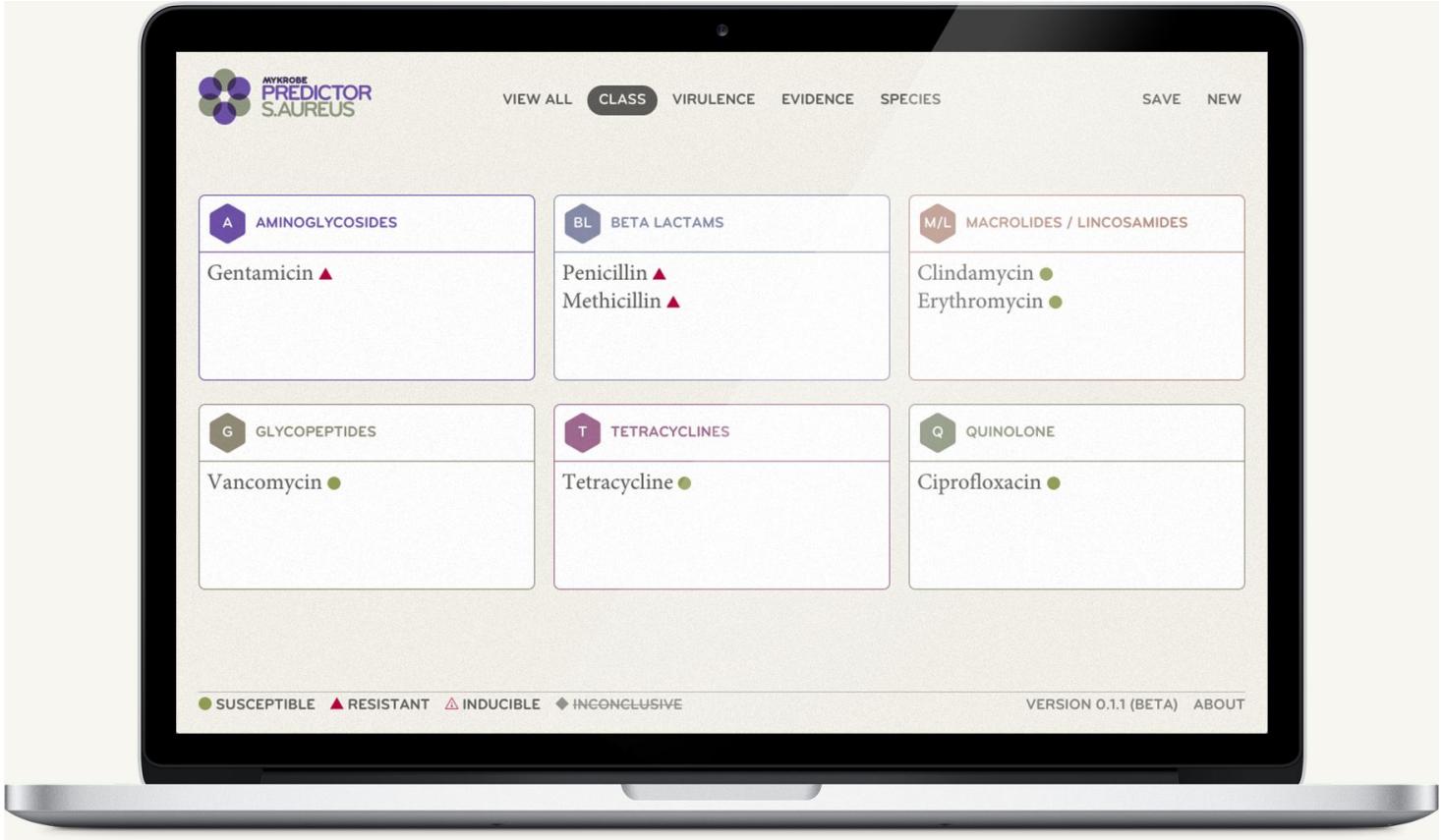
[show more / less]

Annotation

Adding biological information to sequences.



Antimicrobial resistance

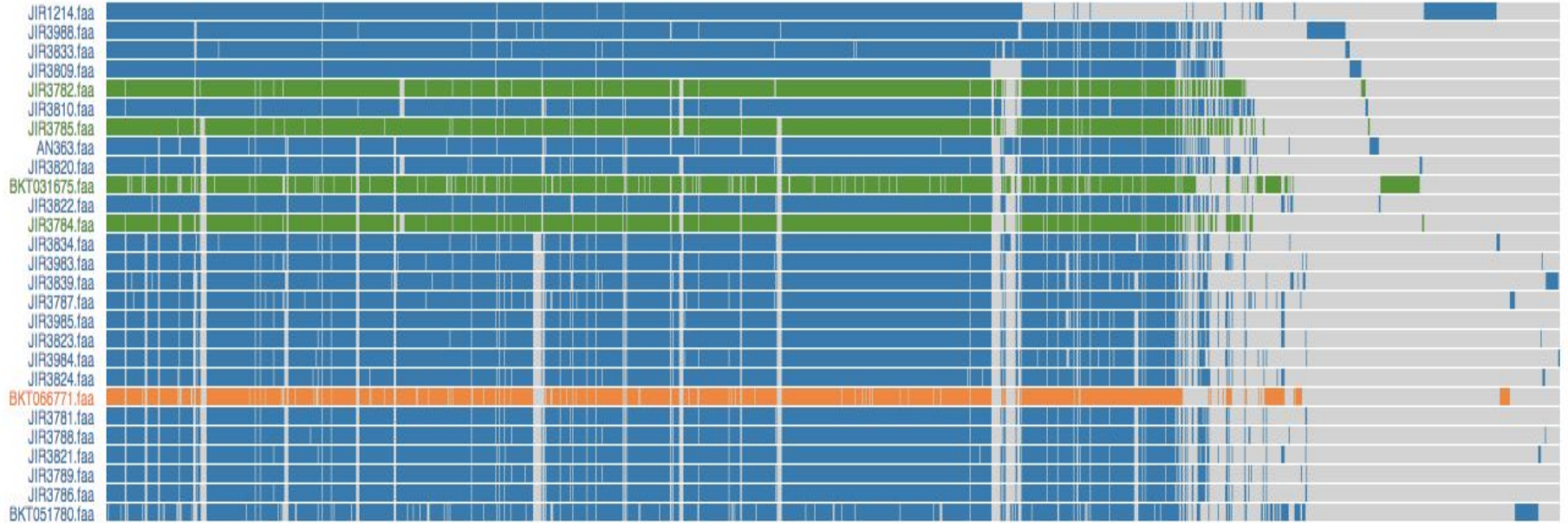


Phylogenetics

Core genome

Core is common to all & has similar sequence.

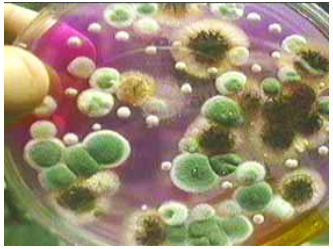
Pan genome

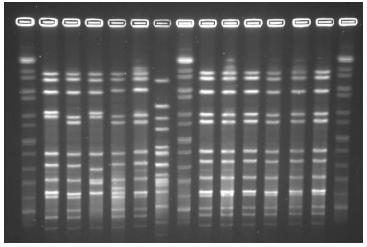


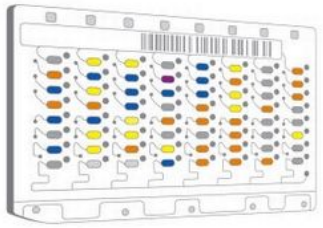
Rows are genomes, columns are genes.

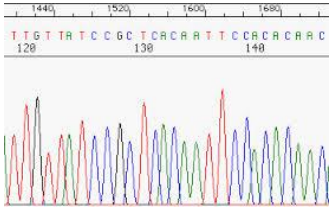
Application to public health and clinical microbiology

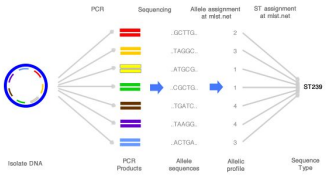
Traditional workflow

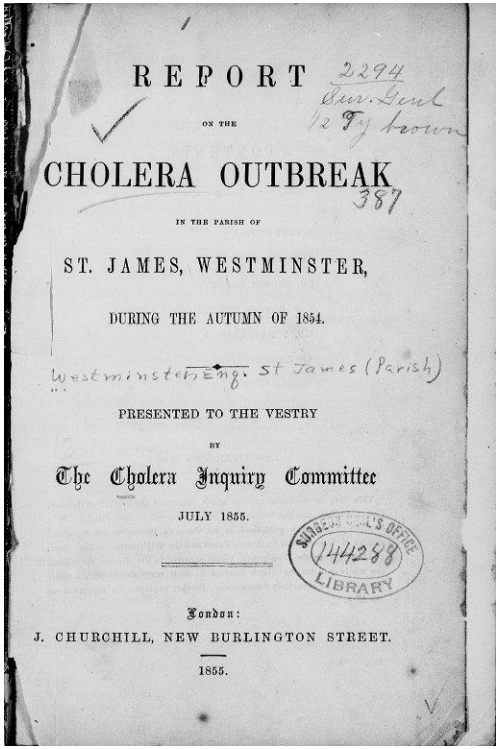






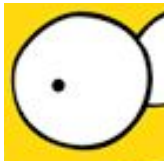






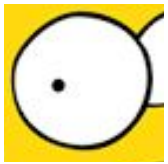
A bacterial isolate

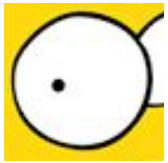
Focus on a small “informative” section

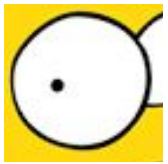


e.g. MLST, VNTR, PFGE, <insert genotyping method here>

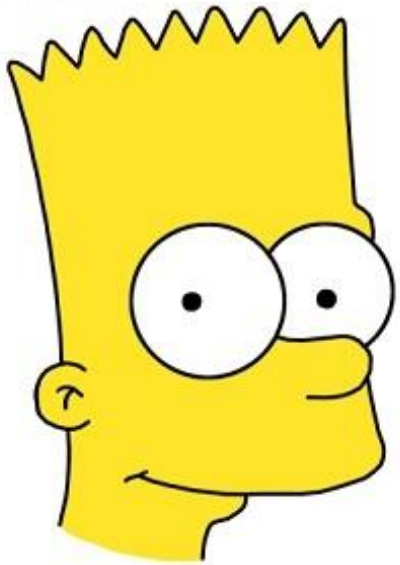
Genotype shows isolates are related



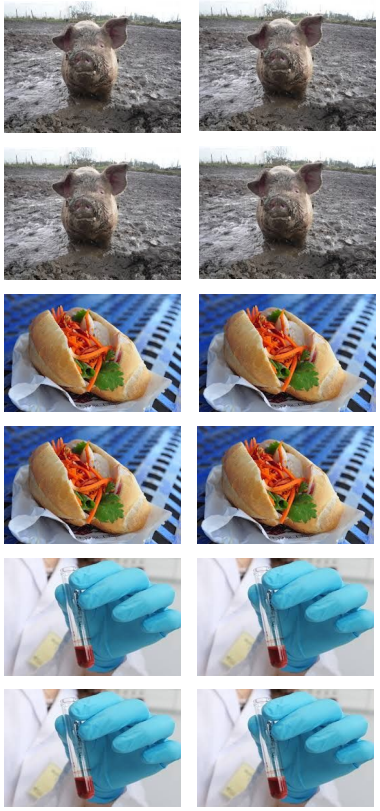




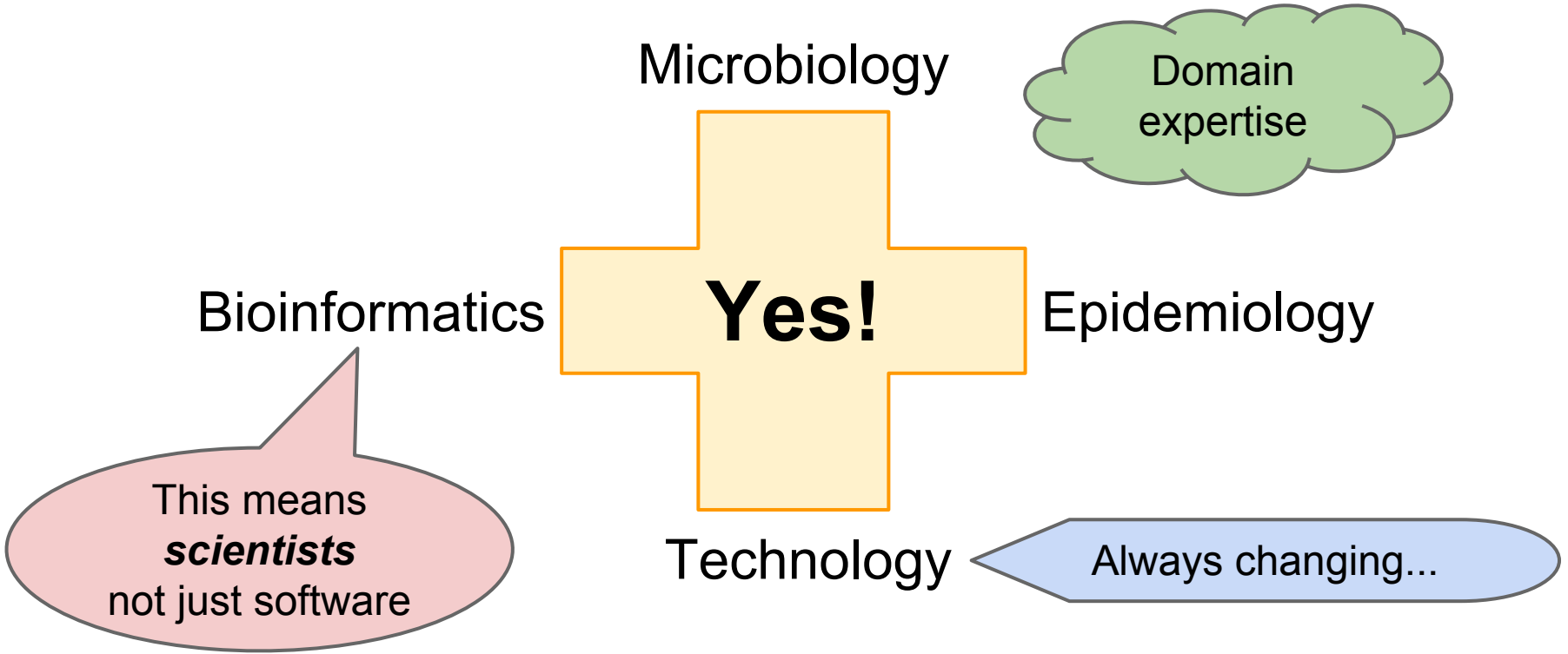
D’oh!



Modern workflow

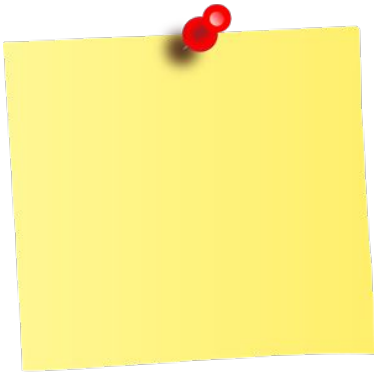


Does WGS deliver?



Conclusions

Take home messages



- Shorts reads & repeats
  - Can't reconstruct genome fully
  - Can't always align unambiguously
- Bioinformatics is a skill
  - not just pushing buttons
  - be nice to us and we will enjoy helping you

Acknowledgements

Anne Syme

Simon Gladman

Anders da Silva

Google

Marcel Behr

Lynn Dery Capes

Robyn Lee

Ines Levade

Fin. Merci.

That’s all folks!