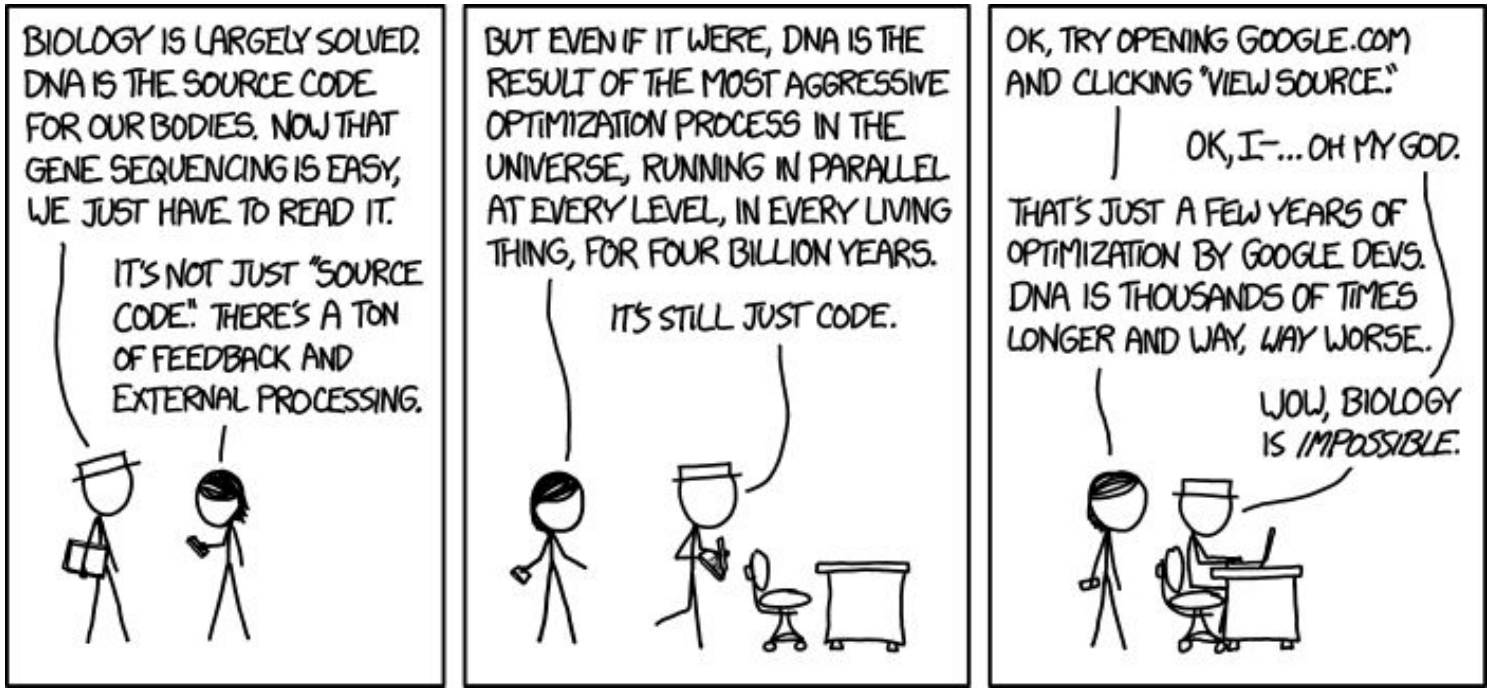


Project #1. "What causes antibiotic resistance?" Alignment to reference, variant calling

AACGCTAACGGTAA

AACCGCGAACTAA

AACGCTAACGGTAA

AACCGCGAACTAA

AAC - GCTAACGGTAA

AACCGCGAAC - - TAA

Course Logistics

One project each 2 weeks - brief introduction, background data and instructions for implementation.

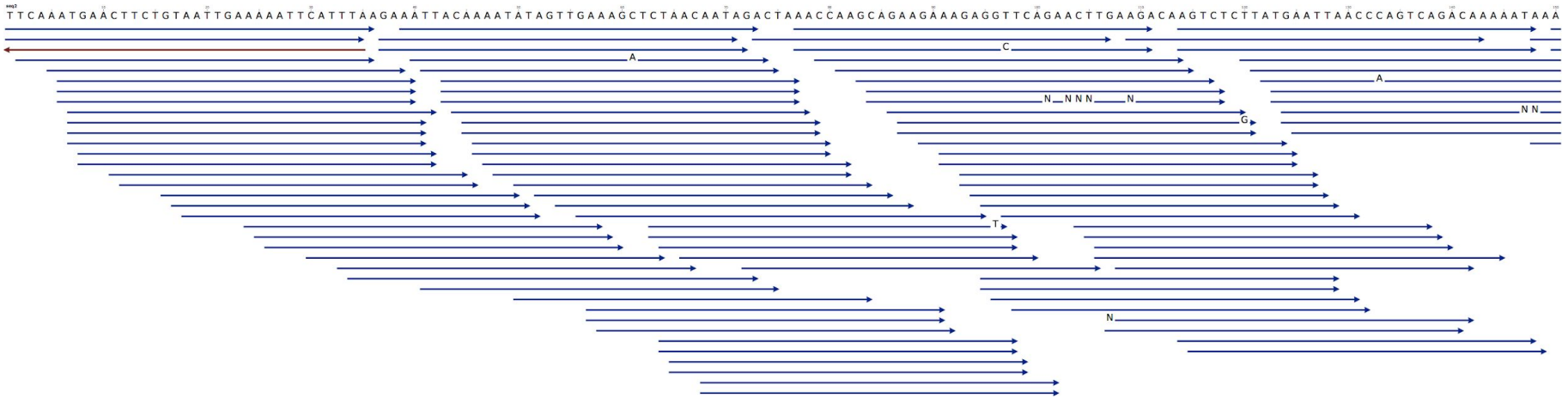
Work in teams of 2.

Report - basically a mini-paper, 2-3 pages long, with introduction, methods, results, etc.

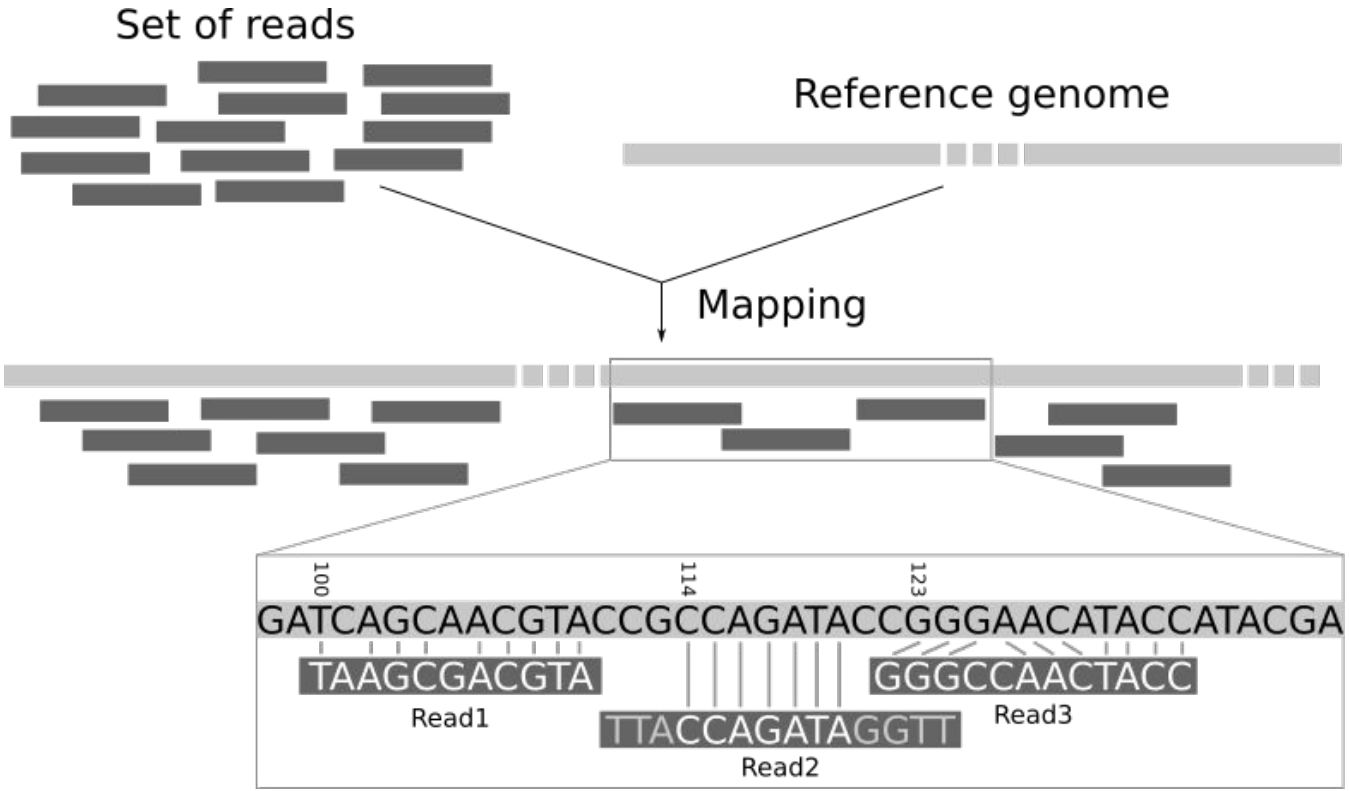
10 points max per project, one lowest score will be dropped.

>70% of the maximum score to pass the course.

Short read alignment



Find the read position in the genome


Set of reads
Reference genome
Mapping
100
114
123
GATCAGCAACGTACCGCCAGATACCGGGAACATACCATACGA
TAAGCGACGTA
Read1
TTACCAGATAGGTT
Read2
GGGCCAACTACC
Read3

Alignment

The earliest alignment algorithms (Smith-Waterman and Needleman-Wunsch) are still used today to compare small pieces of DNA one-by-one.

But the computing power needed to map millions of short reads to large genomes requires special algorithms to speed up the process.

Alignment

Burrows-Wheeler transform allows to map reads to the reference sequence.

The reference is summarized with a special reversible index, the index makes its faster to search.

Reads may contain several mutations, insertions, or deletions.

Alignment applications

- Quality assessment
  - Error rate
  - Insert size distribution
  - Chimeric read/read-pairs
  - Genome fraction
- SNP calling
- Comparative analysis
  - CNVs
- Transcriptomics
  - Gene expression
  - Exon/intron detection

Short read alignment

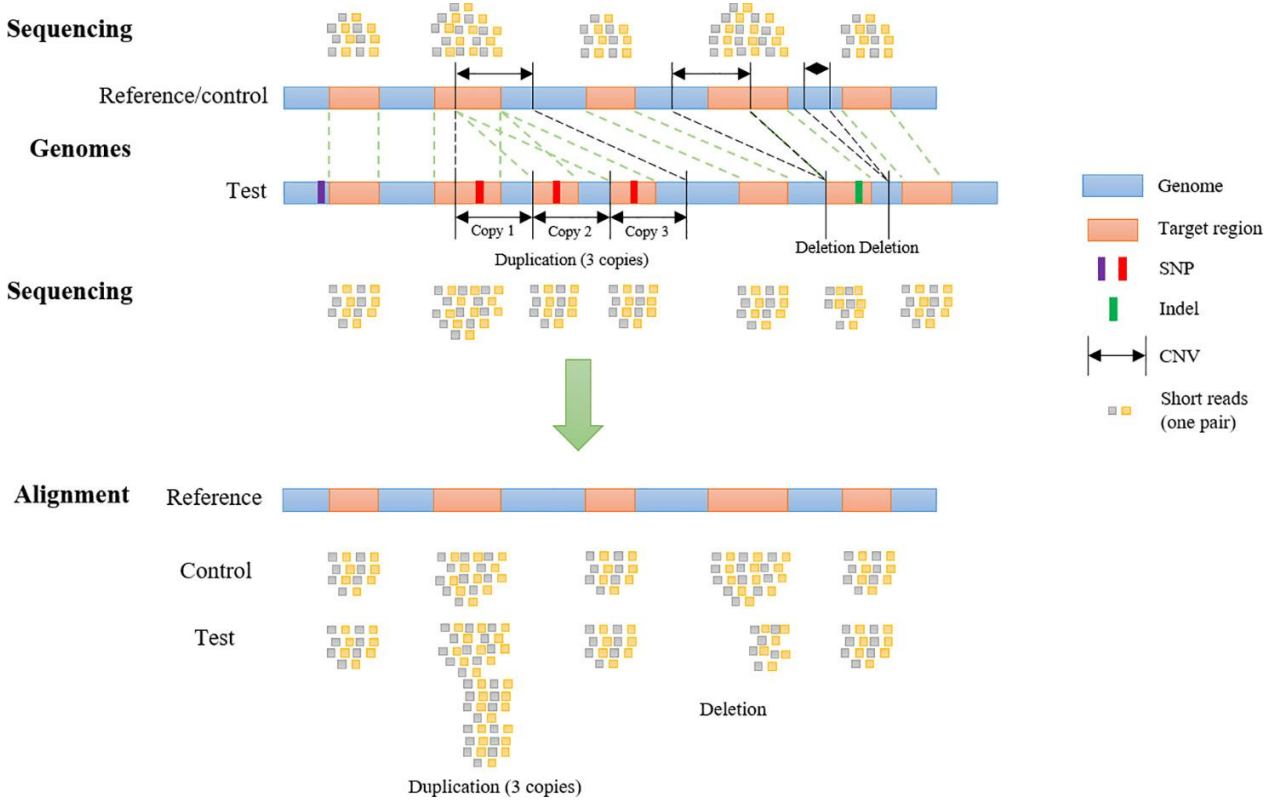
- Challenges

  Small length
  Gigabytes of data
  Different sequencing errors
  SNPs
  Genomic repeats

- Tools

  Bowtie, BWA (Illumina only)
  Bowtie2, BWA-SW, BWA-MEM (Multiple technologies)
  TopHat, STAR (RNA-Seq)
  …and many more

CNV detection



SAM files

- Read ID (QNAME)
- Reference ID (RNAME)
- Mapping position (POS)
- Mate reference ID (RNEXT)
- Mate position (PNEXT)
- Observed insert length (TLEN)
- Read sequence (SEQ)
- Read quality (QUAL)
- CIGAR string
    - 34M 1I 4M 2D 1X 3M

SAM files

```
@HD     VN:1.0  SO:coordinate
@SQ     SN:chr20        LN:64444167
@PG     ID:TopHat       VN:2.0.14       CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-rea
lign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr
20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714       16      chr20   190930  3       100M    *       0       0
       CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCCTCT
C    BBDCCDDCCDDDDCDDDDDDCDCCCDBC?DDDDDDDDDDDDDDDCCDCDDDDDDDDDDCCCCEDDDC?DDDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
    AS:i:-15        XM:i:3  XO:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714   HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961       16      chr20   193953  50      100M    *       0       0
       TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G    DCDDDDEDDDDDDDCDDDDDDDCCCDDDCDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJIJJJJJIHJJJJJJHHHHHFFFFFCCC
    AS:i:-16        XM:i:3  XO:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030       16      chr20   270877  50      100M    *       0       0
       GGCTTTATTGGTAAAAAAGGAATAGCAGATTTAATCAGAAATTCCCACCTGGCCCAGCAGCACCAACCAGAAAGAAGGGAAGAAGACAGGAAAAAACCA
C    DDDDDDDDDCCDDDDDDDDDDEEEEEEEFFFEFFEGHHHHFGDJJIHJJIJIJJJIIIIGGFJJIHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
    AS:i:-11        XM:i:2  XO:i:0  XG:i:0  MD:Z:0A85G13    NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699       0       chr20   271218  50      50M4700N50M     *       0
       0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

BAM files

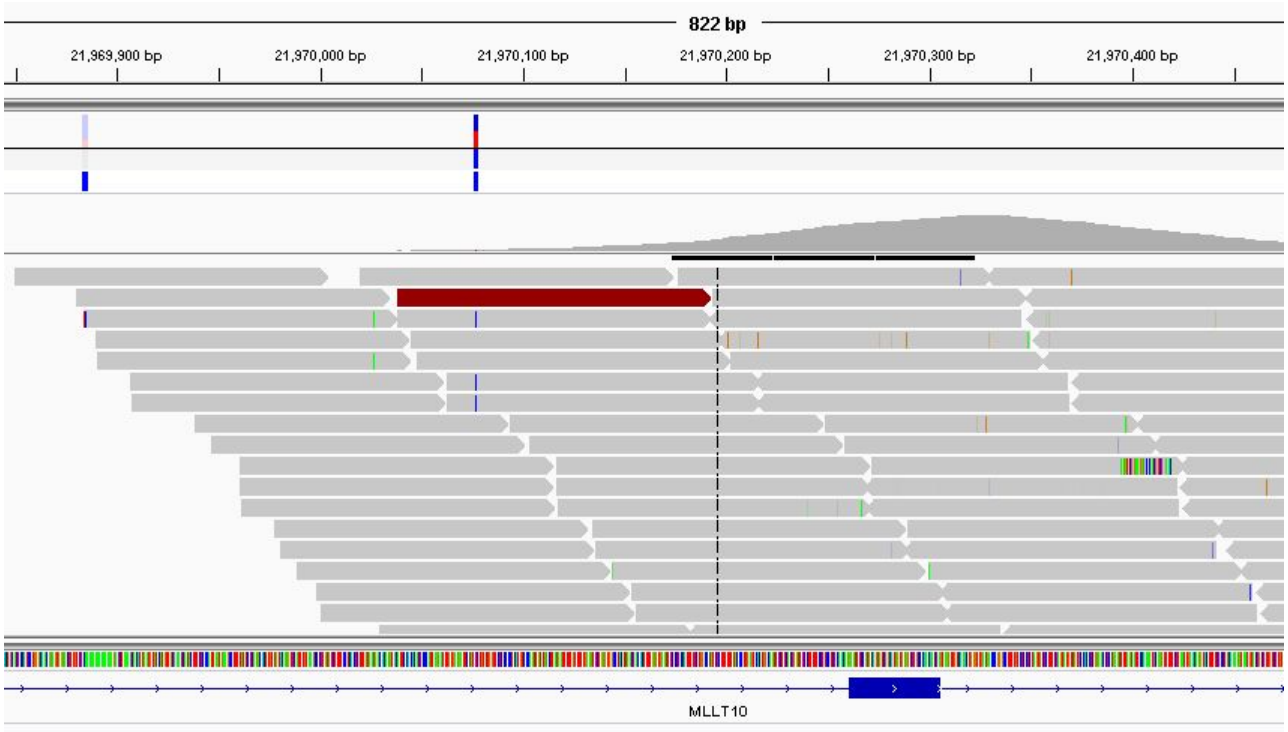
BAM (Binary Alignment Map):

a compressed SAM file, less filesize, not human-readable.

We can convert a sam file to a bam file as follows:

samtools view -S -b alignment.sam > alignment.bam

Alignment visualization with IGV



SNP calling

Process of finding bases in the NGS data that differ from the reference genome

- Typically including an associated statistical confidence score.
- Also known as “variant calling”

We need enough coverage to distinguish real variants from sequencing errors

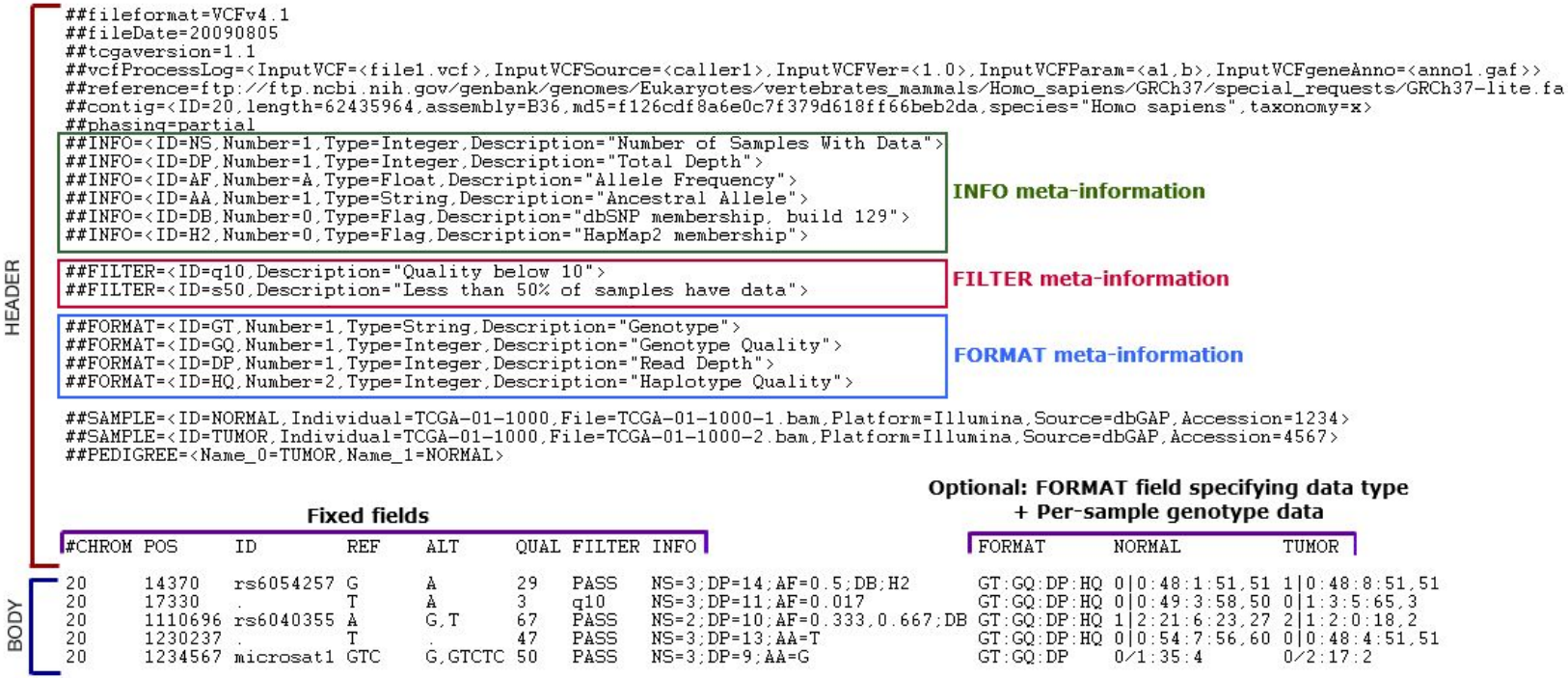
VCF files

- Chromosome (#CHROM)
- Position (POS)
- Unique identifiers where available (ID)
- Reference base(s) (REF)
- Alternate non-reference alleles (ALT)
- Phred quality score for the variant (QUAL)
- Optional filters (FILTER)
- Additional information (INFO)

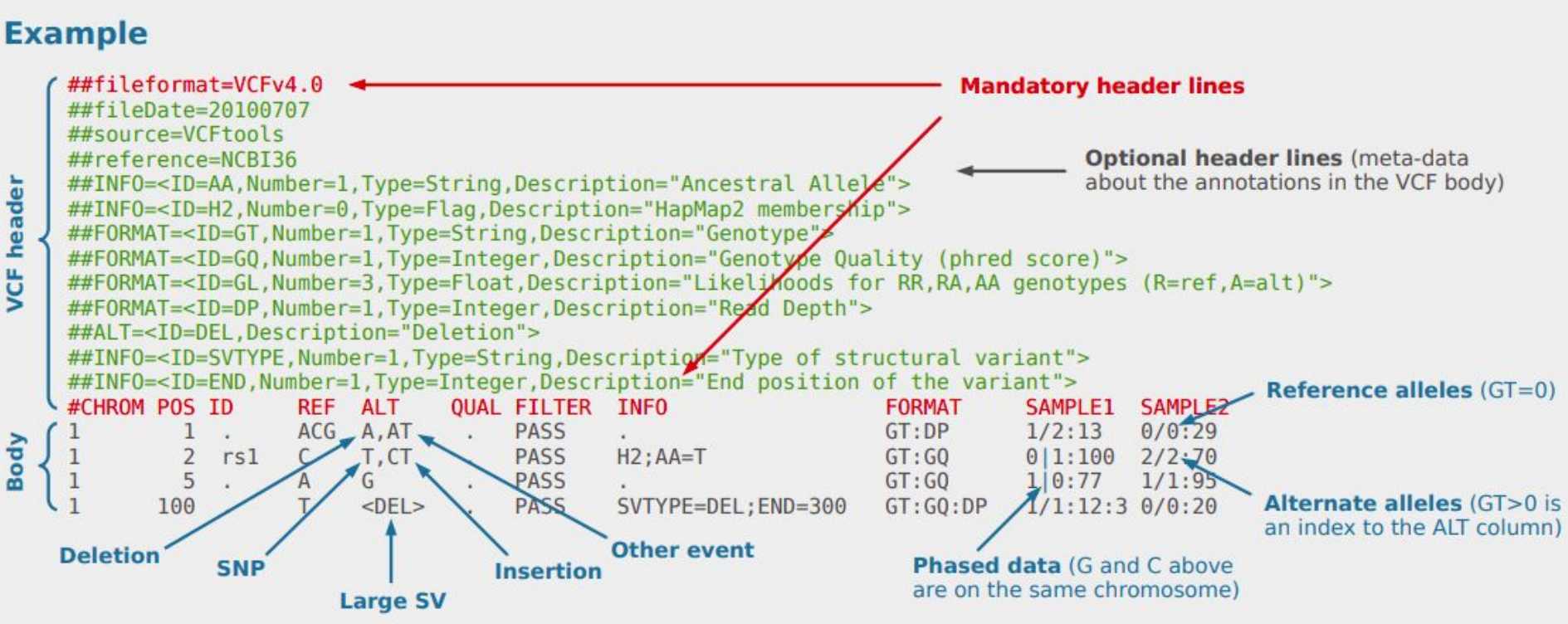
VCF files

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	10177	.	A	AC	100	PASS	AC=2130;AF=0.425319;AN=5008;NS=2504
1	10235	.	T	TA	100	PASS	AC=6;AF=0.00119808;AN=5008;NS=2504
1	10352	rs145072688	T	TA	100	PASS	AC=2191;AF=0.4375;AN=5008;NS=2504
1	10505	.	A	T	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504
1	10506	.	C	G	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504
1	10511	.	G	A	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504
1	10539	.	C	A	100	PASS	AC=3;AF=0.000599042;AN=5008;NS=2504
1	10542	.	C	T	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504
1	10579	.	C	A	100	PASS	AC=1;AF=0.000199681;AN=5008;NS=2504
1	10616	rs376342519	CCGCCGTTGCAAAGGCGCGCCG	C	100	PASS	AC=4973;AF=0.993011;AN=5008;NS=2504
1	10642	.	G	A	100	PASS	AC=21;AF=0.00419329;AN=5008;NS=2504
1	11008	.	C	G	100	PASS	AC=441;AF=0.0880591;AN=5008;NS=2504
1	11012	.	C	G	100	PASS	AC=441;AF=0.0880591;AN=5008;NS=2504
1	11063	.	T	G	100	PASS	AC=15;AF=0.00299521;AN=5008;NS=2504
1	13011	.	T	G	100	PASS	AC=3;AF=0.000599042;AN=5008;NS=2504
1	13110	.	G	A	100	PASS	AC=134;AF=0.0267572;AN=5008;NS=2504

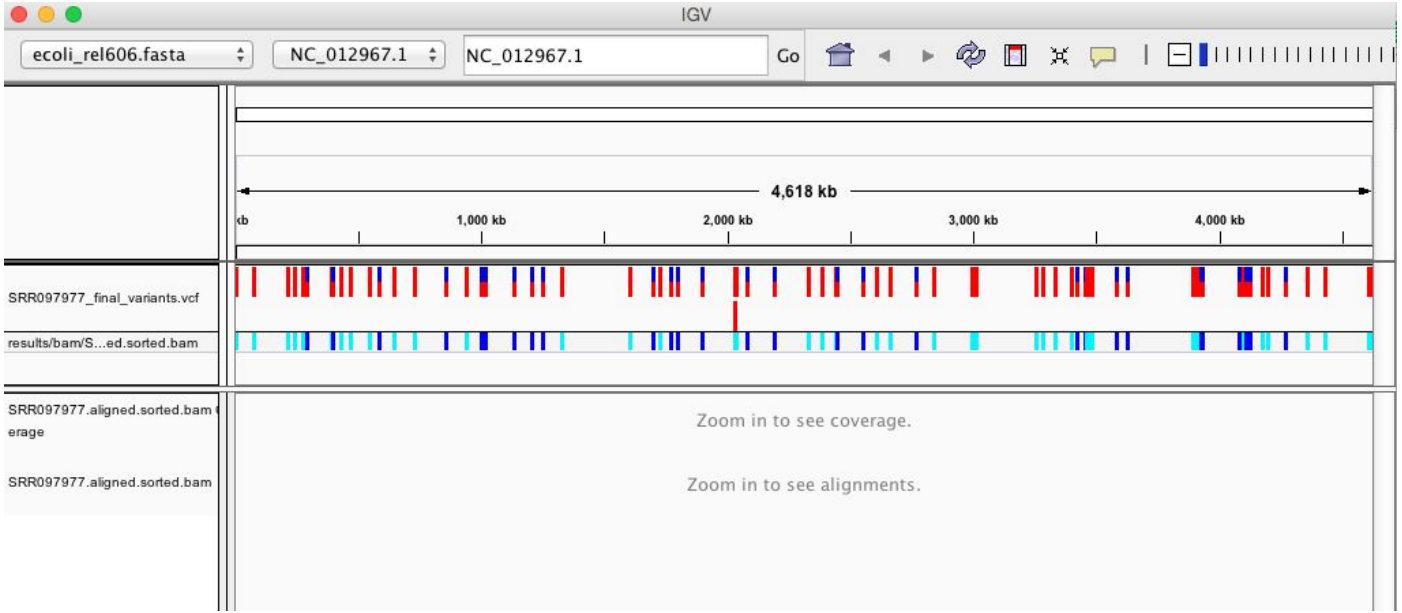
VCF file structure



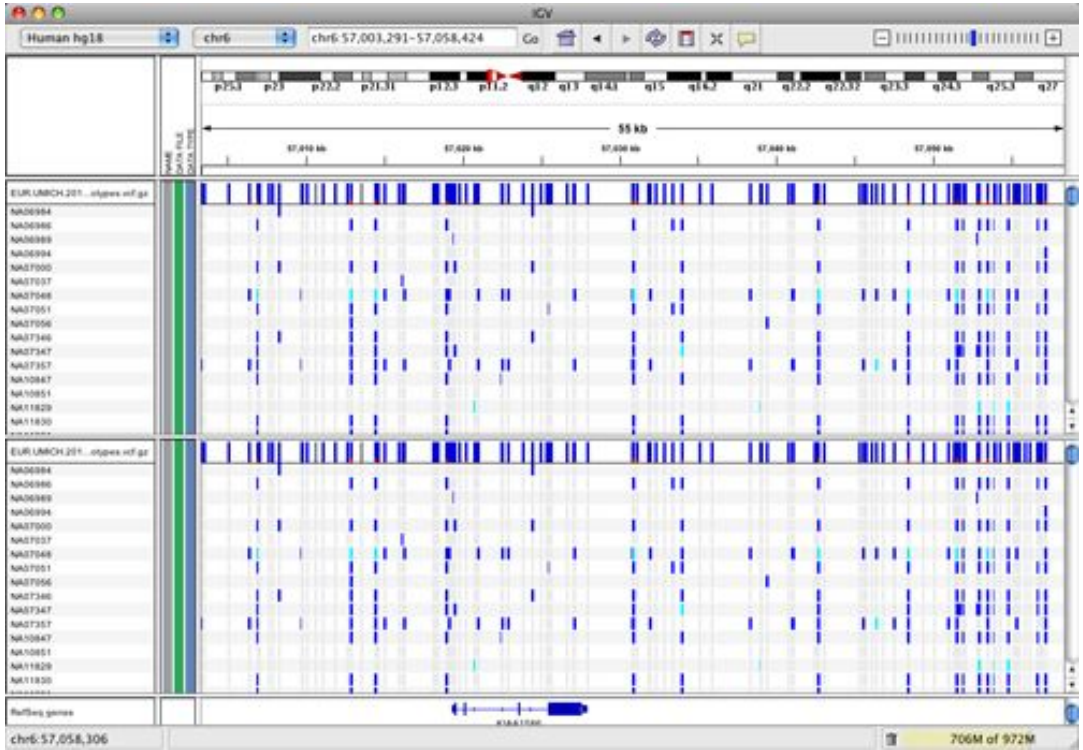
VCF file structure



Alignment visualization with IGV: one sample



Alignment visualization with IGV: multiple samples

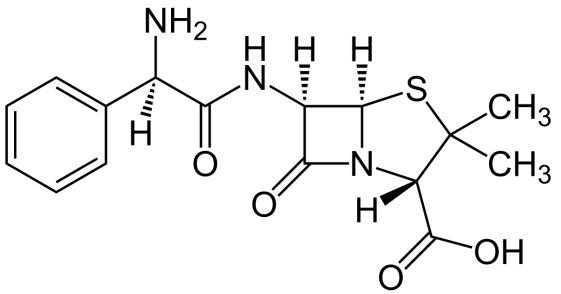


Tools

- Alignment and data processing
  - samtools
  - vcftools
- SNP calling and annotation
  - VarScan
  - SnpEff
- Visualization
  - Tablet
  - IGV
- Pipelines
  - GATK

Case study - antibiotic resistance

Real sequencing data from a strain of *E. coli* resistant to the antibiotic ampicillin.

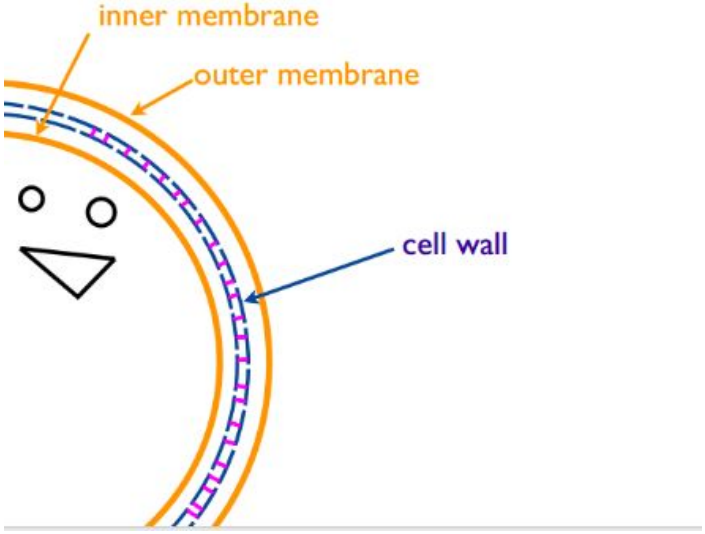


Goals

- Map reads to reference
- Locate mutations
- Figure out what mutations do
- Classify mechanism of resistance
- Make recommendations for alternative treatment

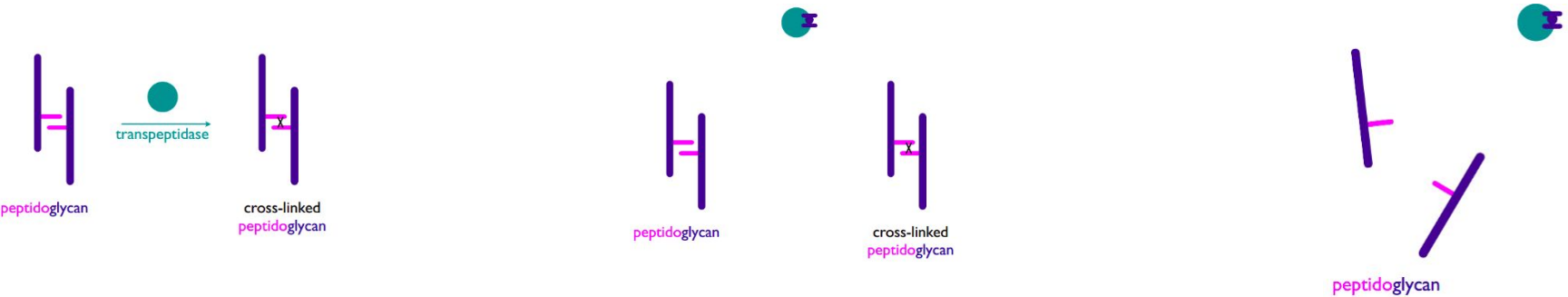
Case study - antibiotic resistance

Ampicillin acts as an irreversible inhibitor of the enzyme transpeptidase, which is needed by bacteria to make the cell wall.

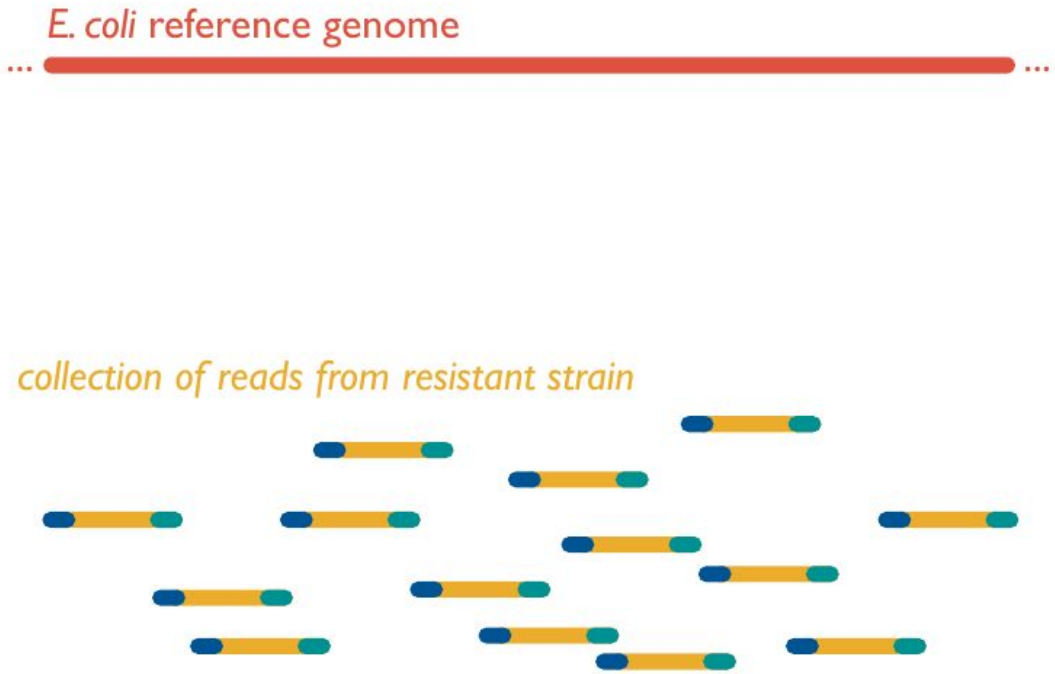


Case study - antibiotic resistance

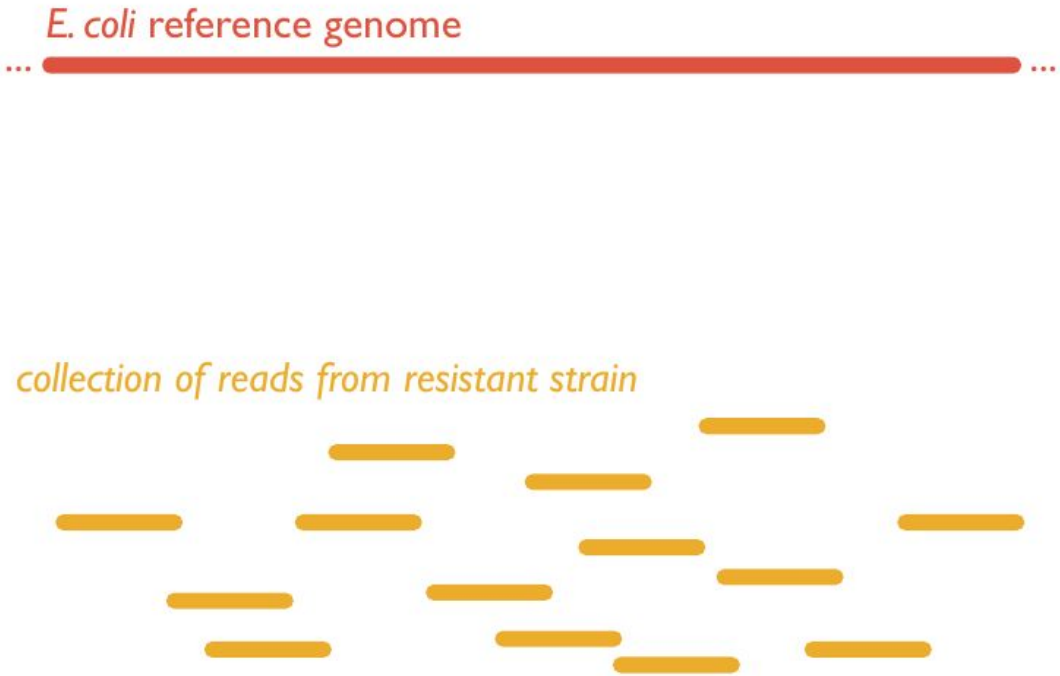
Ampicillin acts as an irreversible inhibitor of the enzyme transpeptidase, which is needed by bacteria to make the cell wall.



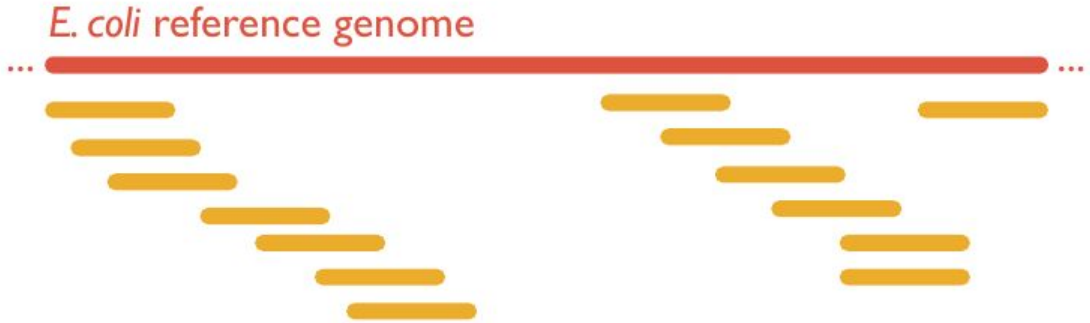
Case study - antibiotic resistance



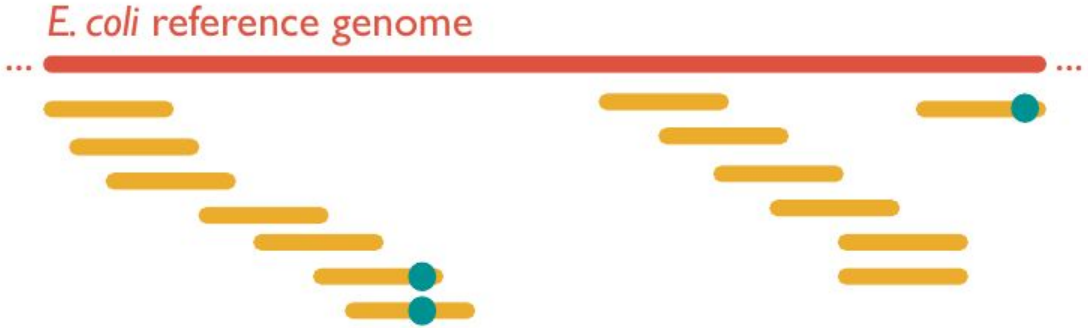
Case study - antibiotic resistance



Case study - antibiotic resistance



Case study - antibiotic resistance



General mechanisms of resistance

1) target site alteration so antibiotic can’t bind

2) inactivation or modification of antibiotic itself

3) alter metabolic pathway to compensate

4) reduce amount of drug in cell

4b) kick drug out of the cell (efflux pumps)

4a) decrease permeability so drug can’t enter (alter pores to block hydrophilic drugs, alter membrane to block hydrophobic drugs)

General mechanisms of resistance

Which one works in our case?

(and what can we do?)

Thank you!

Questions?