

**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.

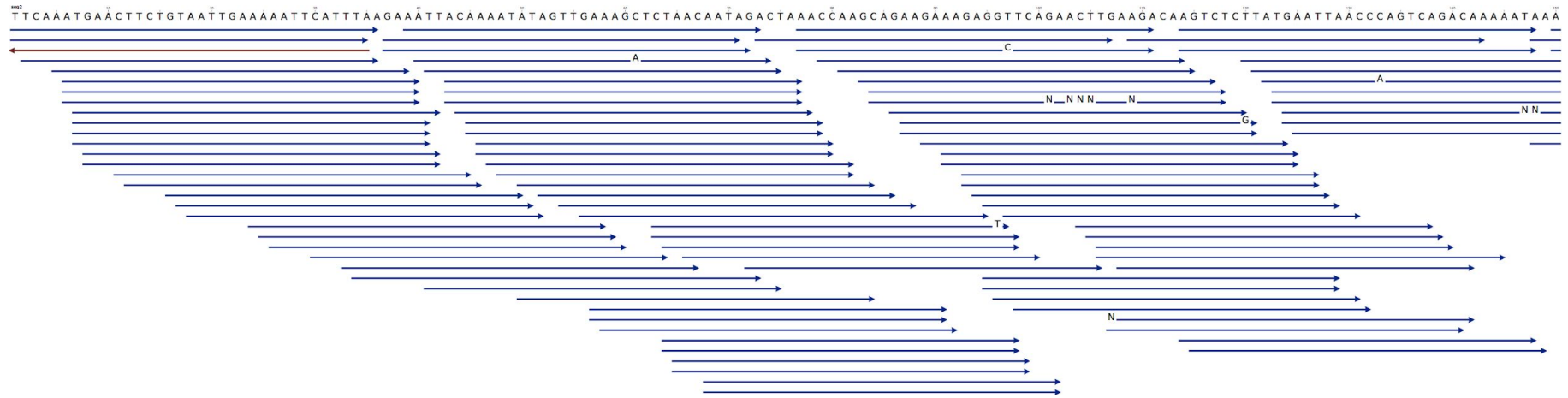
Work in teams of 2.

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

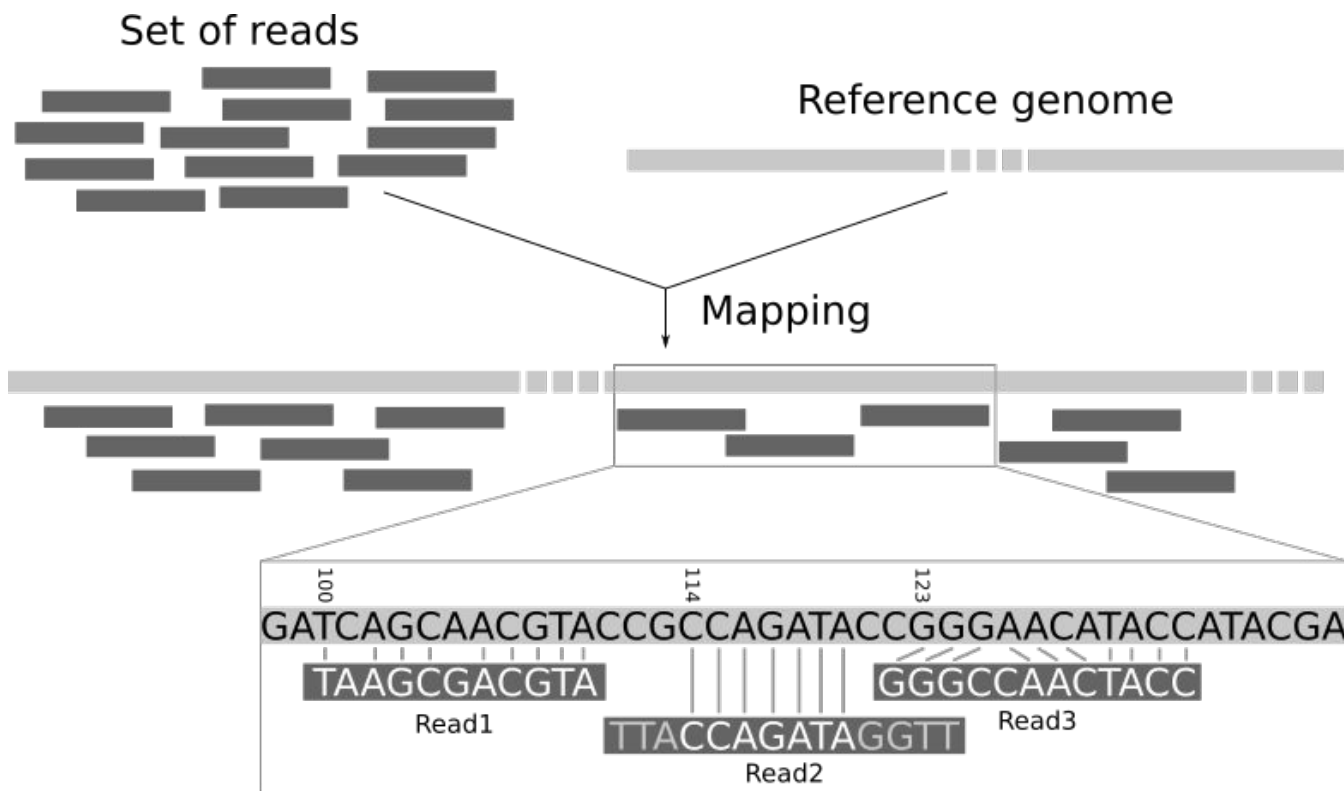
10 points max per project, one lowest score will be dropped.

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

# Short read alignment



**Find the read position in the genome**



# 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

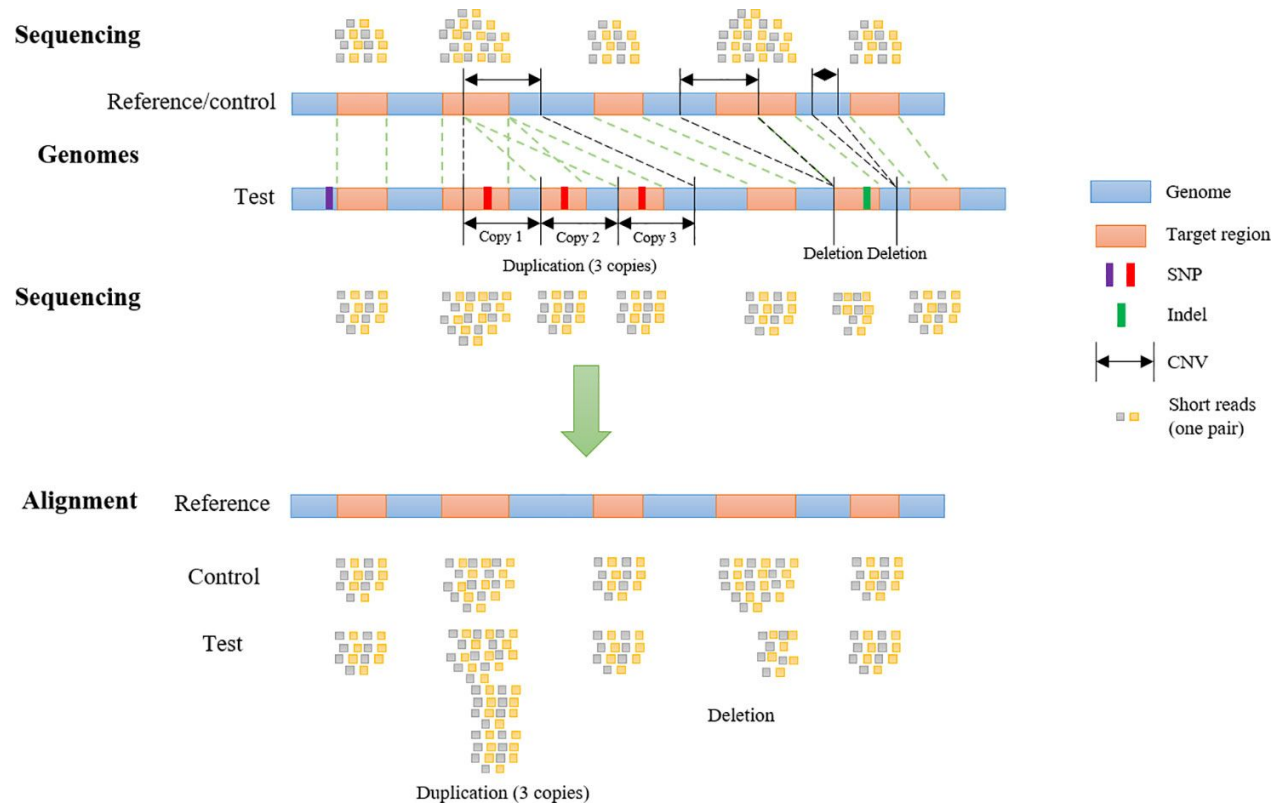
- Bowtie, BWA (Illumina only)

- 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 11 4M 2D 1X 3M

# SAM files

```
@HD      VN:1.0  S0: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-realign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr20 /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
      CCGTGTTTAAAGGTGGATGCGGTCACCTTCCCAGCTAGGCTTAGGGATTCTTAGTTGGCCTAGGAAATCCAGCTAGTCCTGTCTCTCAGTCCCCCTCT
C      BBDCCDDCCDDDDCDCCCCDBC?DDDDDDDDDDDDDDCCDDDDDDDDDDCCCCEDDC?DDDDDDDDDDDDDDDDDDDDBDHFFFFDC@@
      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
      TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTTCGTCCTGGGGCAGTGGACCTTCCAGTGATTCCCCTGACATAAGGGGCATGGACGA
G      DCDDDDDEDDDDDDCDDDDDDDCCDDDDDDDEEC>DFFFEJJJJJIGJJJJIHGBHHGJIJJJJJGJJJJJJJJJJHJJJJJJHHHHHHFFFFCCC
      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
      GGCTTTATTGGTAAAAAGGAATAGCAGATTTAATCAGAAATCCACCTGGCCCAGCAGACCAACCAGAAAGAAGGAAGAAGACAGGAAAAAACCA
C      DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHFGDJJIHJJJIJJJJIIIGFJJIIIIJJJJJJJJGHHFAHGFHJHFGGHFFFD@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      GTGGCTCTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTCG
accepted_hits.sam
```

# BAM files

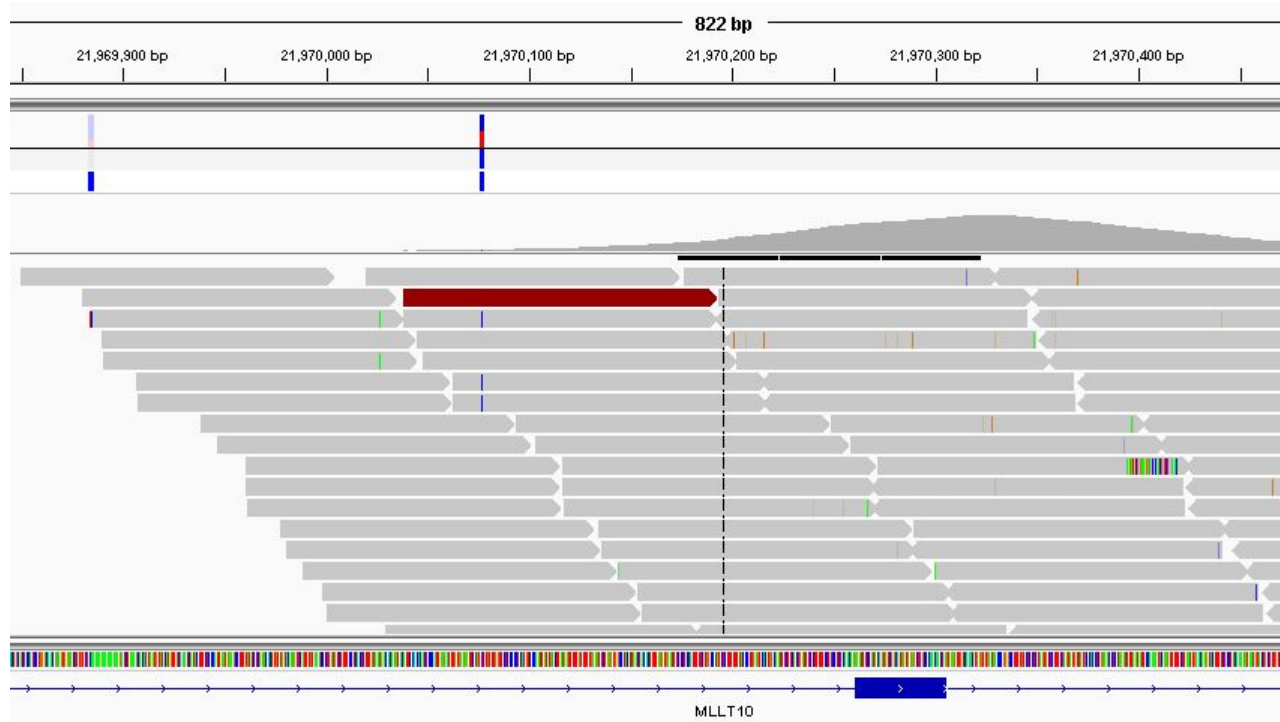
BAM (Binary Alignment Map):

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

```
##fileformat=VCFv4.1
##fileDate=20090805
##tcgaversion=1.1
##vcfProcessLog=<InputVCF=<file1.vcf>,InputVCFSource=<caller1>,InputVCFVer=<1.0>,InputVCFParam=<a1,b>,InputVCFgeneAnno=<anno1.gaf>>
##reference=ftp://ftp.ncbi.nih.gov/genbank/genomes/Eukaryotes/vertebrates_mammals/Homo_sapiens/GRCh37/special_requests/GRCh37-lite.fa
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">

##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">

##SAMPLE=<ID=NORMAL,Individual=TCGA-01-1000,File=TCGA-01-1000-1.bam,Platform=Illumina,Source=dbGAP,Accession=1234>
##SAMPLE=<ID=TUMOR,Individual=TCGA-01-1000,File=TCGA-01-1000-2.bam,Platform=Illumina,Source=dbGAP,Accession=4567>
##PEDIGREE=<Name_0=TUMOR,Name_1=NORMAL>
```

INFO meta-information

FILTER meta-information

FORMAT meta-information

Optional: FORMAT field specifying data type  
+ Per-sample genotype data

## Fixed fields

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
20	14370	rs6054257	G	A	29	PASS	NS=3;DP=14;AF=0.5;DB;H2
20	17330	.	T	A	3	q10	NS=3;DP=11;AF=0.017
20	1110696	rs6040355	A	G,T	67	PASS	NS=2;DP=10;AF=0.333,0.667;DB
20	1230237	.	T	.	47	PASS	NS=3;DP=13;AA=T
20	1234567	microsat1	GTC	G,GTCTC	50	PASS	NS=3;DP=9;AA=G

FORMAT	NORMAL	TUMOR
GT:GQ:DP:HQ	0 0:48:1:51,51	1 0:48:8:51,51
GT:GQ:DP:HQ	0 0:49:3:58,50	0 1:3:5:65,3
GT:GQ:DP:HQ	1 2:21:6:23,27	2 1:2:0:18,2
GT:GQ:DP:HQ	0 0:54:7:56,60	0 0:48:4:51,51
GT:GQ:DP	0/1:35:4	0/2:17:2

# VCF file structure

## Example

**VCF header**

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
```

**Mandatory header lines**

**Optional header lines** (meta-data about the annotations in the VCF body)

**Body**

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	SAMPLE1	SAMPLE2
1	1	.	ACG	A,AT	.	PASS	.	GT:DP	1/2:13	0/0:29
1	2	rs1	C	T,CT	.	PASS	H2;AA=T	GT:GQ	0/1:100	2/2:70
1	5	.	A	G	.	PASS	.	GT:GQ	1/0:77	1/1:95
1	100	.	T	<DEL>	.	PASS	SVTYPE=DEL;END=300	GT:GQ:DP	1/1:12:3	0/0:20

**Reference alleles** (GT=0)

**Alternate alleles** (GT>0 is an index to the ALT column)

**Deletion**

**SNP**

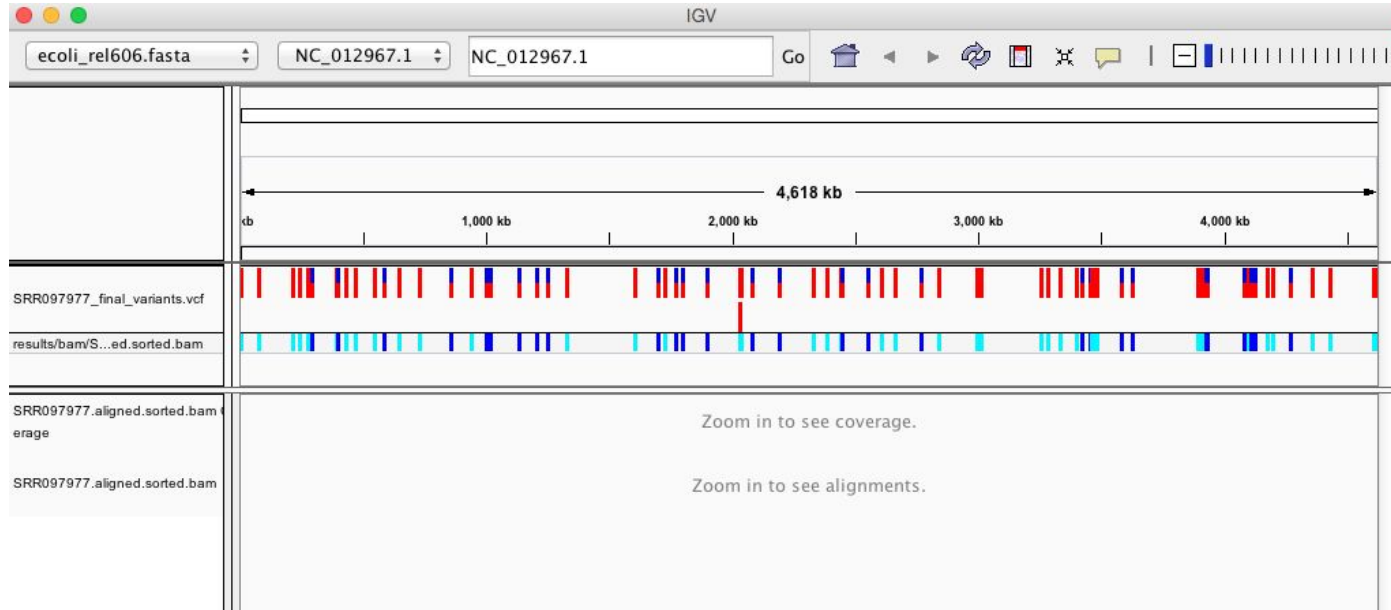
**Large SV**

**Insertion**

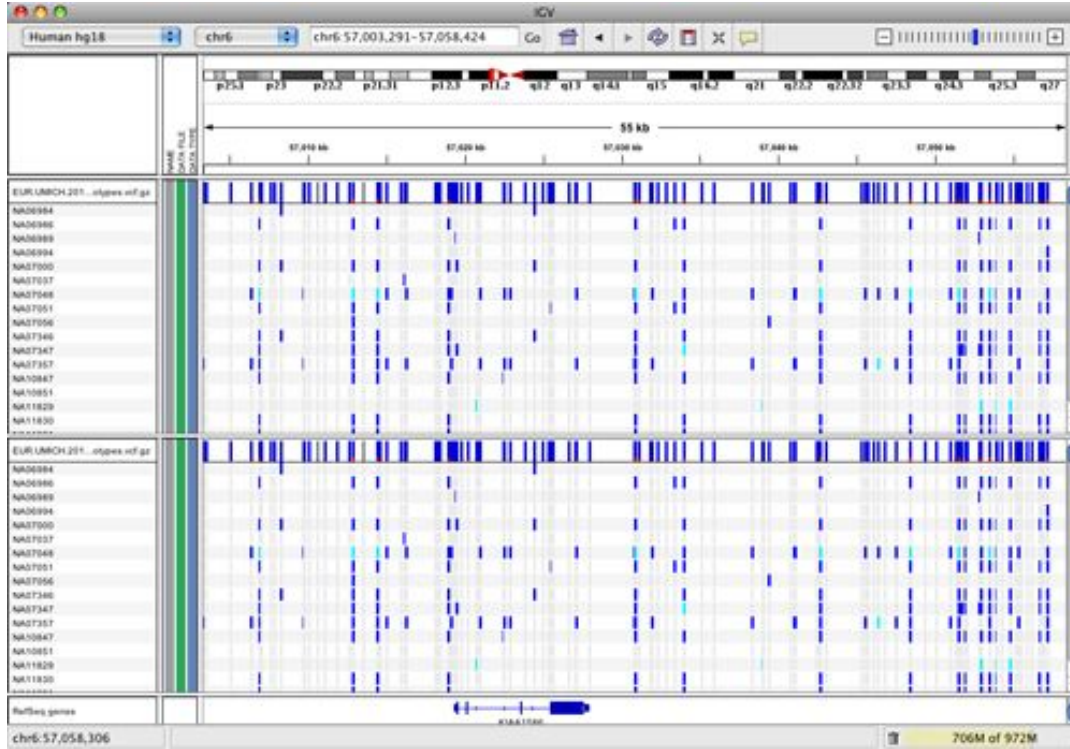
**Other event**

**Phased data** (G and C above are on the same chromosome)

# Alignment visualization with IGV: one sample



# Alignment visualization with IGV: multiple samples



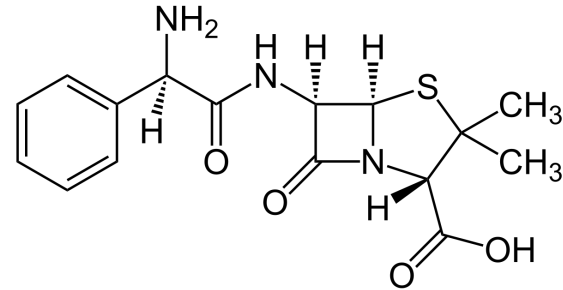
# Tools

- 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.



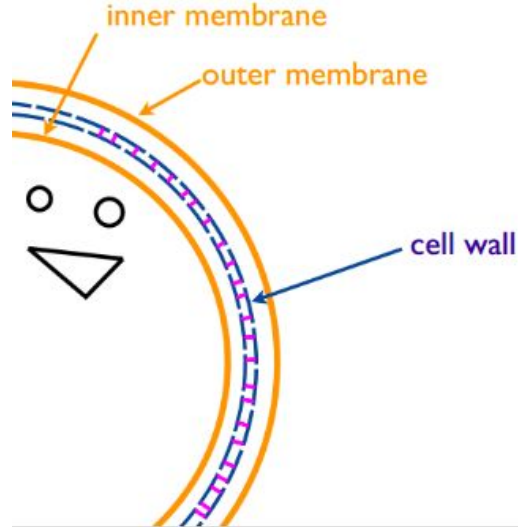


# Goals

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

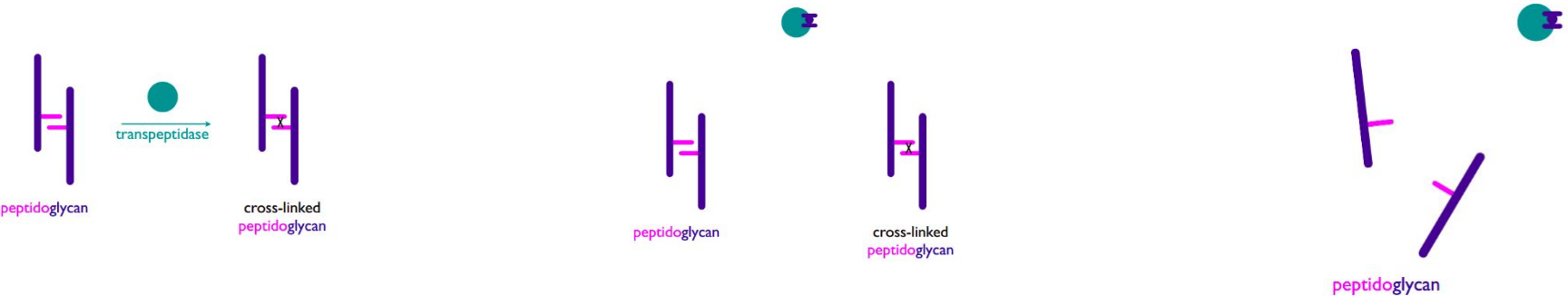
# Case study - antibiotic resistance

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.

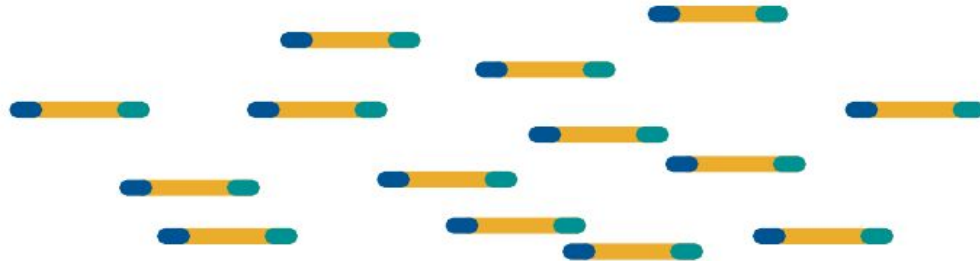


# Case study - antibiotic resistance

*E. coli* reference genome



*collection of reads from resistant strain*



# Case study - antibiotic resistance

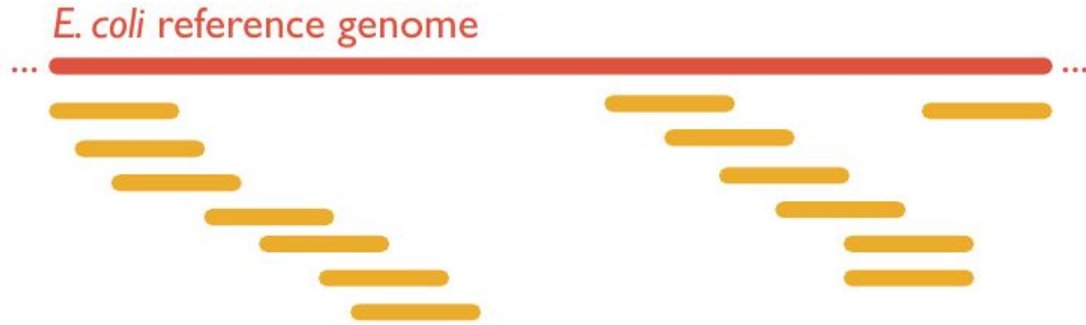
*E. coli* reference genome



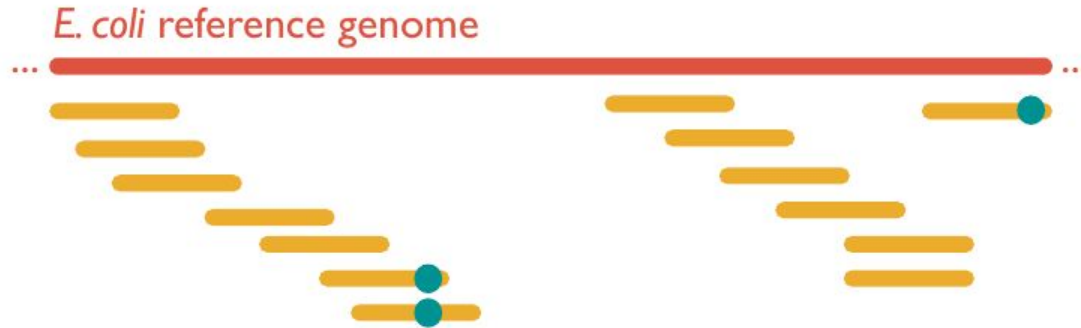
*collection of reads from resistant strain*



# 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?**