Pathogen Genomics: Introduction to genome sequencing and analysis

Adam Reid & Steve Doyle

Wellcome Sanger Institute/LSHTM

LSHTM Pathogen Genomics

Summary

- What is the point of a genome sequence?
- Genome sequencing technologies
- Sequence data files
- Viewing genomes
- Computer practical 1: Viewing genome sequences
- Computer practical 2: Analysis of sequence variants

Why do genome sequencing?

- Reference for molecular biology
  - *Tropheryma whipplei* causes the potentially fatal Whipple's disease. Could not easily be grown. Genome revealed it had lost genes involved in producing amino acids.
- Identify all the genes that determine the function of the organism
  - *Neisseria meningitidis*, a major cause of meningitis. The first vaccine for a particular form of meningitis for identified by looking for candidates in its genome.
  - *Rickettsia prowazekii* is the cause of epidemic typhus, which killed millions in the early 20th century. It cannot reproduce outside of these cells. It was found to have just over 800 genes.
- Examine evolution by comparative genomics
- Track spread of pathogens
- Identify antimicrobial/drug resistance genes and drug targets
  - Mtb researchers made bacteria resistant to a new drug. Genome sequencing identified the gene involved in resistance.
- Basis for other omics technologies – RNA-seq, ChIP-seq, Methylome etc.

Why do genome sequencing? - Video of Wellcome Sanger Institute researchers

Technology overview

- Sanger sequencing produces ~500bp reads
  - Pros: Highly accurate
  - Cons: Expensive, laborious
  - Uses: High quality reference genomes
- Illumina's sequencing-by-synthesis 75-250bp
  - Pros: cheap, lots of reads (e.g. 500 million per run)
  - Cons: short reads
  - Uses: Resequencing, draft genomes, RNA-seq
- Pacific Biosciences Single-Molecule Real Time (SMRT) reads of 5000bp-40000bp
  - Pros: long reads
  - Cons: Fewer reads than Illumina - ~1 million, low accuracy
  - Uses: Reference genomes

Genome sequencing technologies – Interview with Mike Quail

Sequence data

Fasta

```
>yneN
TTAATGCCTCTTCTCATTTCTTCTGCTGTCATCCGCACAGCAGAAGAATTCCTCATTGAC
TATTATTTCGCAATTTGCTCACATGGATTAAATTAAACTACATACTATAAGATATAAACT
TCTGCCTACAGCTGTAAGAAACTCCGCTCAGTACTGAAGCACCAGTCCTATTTCCTCTTT
TCTCCAGCCTGTTATATTAAGCATACTGATTAACGATTTTTAACGTTATCCGCTAAATAA
ACATATTTGAAATGCATGCGACCACAGTGAAAAACAAAATCACGCAAAGAGACAACTATA
A
>yegR
ACTAACGGCTGCCACCGATAAATTTCAAAAAAGAGCATATACCTAATATTCAACTAAACA
GTGGCATCTTCAATATAATATATTAAAGCCCCCATGGAGTTACCCTGAAGGGCCTCAATG
TCCGTAATTCCTACTTATGTAGGAAATGTTGTACAGAACATTTATTATAATCCTATTCAA
TTATAATAATCATGCCATTATTATATTTAAACACTAGAGAGTGTCGTTGGTATTTAATGG
GGGAAGGTGAGATGAAAAAGATAGCTGCTATATCATTAATTAGTATTTTTATTATGTCTG
G
>emrK
AAATCAGGGATTGTACCGATGATTTATAGTTTCAAGTTGGCACTATAAGTCTTCTTACTA
ATCCTACAGGCGTAAGAATTGTATTGCAAAAGCCACGGTTTAGTCCTCTGTTGTTTTTTT
TGCACCTCATTTAAATTAGGCCTCCAACGTTCCTGGGATAATGTGCAACACATGCACTGT
GTTTGATATGAAGAATGAATGCTCTTTTCATTCAATTCATAAATTTCATCTATGAGAAAT
GAGAGATAATAGTGGAACAGATTAATTCAAATAAAAAACATTCTAACAGAAGAAAATACT
T
>evgA
AATACAATTCTTACGCCTGTAGGATTAGTAAGAAGACTTATAGTGCCAACTTGAAACTAT
AAATCATCGGTACAATCCCTGATTTTATTGTTGACATTTCATTTATGCCGACTATTTATA
TGGTATACTTGTCGAATTATCTTAAAGGAAGCTCAGATTTTCTTATTTTTATTGAGAAAA
TGAGATGACGCCTTATGTCTGTATTACTACAGGGAGAAGGGAGATGCTTCATTGCAAAGG
GAATAATCTATGAACGCAATAATTATTGATGACCATCCTCTTGCTATCGCAGCAATTCGT
>yfdX
TGGCTGTATTTACATTTAATTAATCAGTATTTACATCGATATAATAAATGACATCTCTTT
GTGGTATATAAGAATAGTTCTCTGCGACAGGAAGCATATTCCTACAATTGTAAGACTAAA
ATACTTCTTGCGATAATAACTACAACTGTAAGATAACCCTTTCAAAATGACCGTTGCTCT
CTGATTTCTCATTTCATGCTCACCCAATATGATGGCGGCGTTTTCTAAAACTGTTAAAGA
ATGAGGTAAGTATGAAACGTTTAATTATGGCCACGATGGTCACAGCAATTCTGGCATCTT
C
```

Fastq

```
@HS34_24228:8:1101:1116:7158/2
ATGCAGATTTTTTTACTATAAAAAATCATCATAAGGATAATANNGATAACAATGANNNNNATGTGAATGTAATAA
+
BBBBBFFFBFFFFF<FFFFFFFFFFBFFFFFFFFFFFFFFF!!<<<FFFFFFFF!!!!!<<<<FFFFFFFFFF
@HS34_24228:8:1101:1116:7461/2
AATATTAAAAAAATGGTTGAAAATCCAGAAGCATTTGCCGCANNTGCTGCACCTGNNNNNGCAGCCAAAGCTGAT
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFB<F/BBBBFB<F!!</</<FB/B<!!!!!////<B<FFFF/</F/
@HS34_24228:8:1101:1116:8637/2
TGGTGTTTTTATTTATTTTAAATATATTCTAATAATAATATANNATAAATAAAAANNNNNAATATATATATATAT
+
BB/BBFFFF</F<FFFFFF<F/FFBBFFBF//FFFFFBFBBF!!<<<<//<F/FF!!!!!<</</</<FBFFF/<
@HS34_24228:8:1101:1116:52646/2
CCGAATTATGCCTAGAAATTCACGAAGATGAGGAGGAAGGCGNNGGAAACGACGANNNNNACGACGACCACGAAG
+
BBBBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF!!/<<BFFFFFFF!!!!!<<<FFFFFFFFFFF
@HS34_24228:8:1101:1116:52943/2
AATATAAAACATTGATTGAAGTGGTATTAAAACCAAGCGAACNNACATATGATATNNNNNTTACACATGCGTATT
+
BBBBBFBBFBFFFF<FFF/FFFFFFFFFFFF/<FFFFFFFF!!<<FFFFFBFFF!!!!!/<<FFF/FFFFFFFF
@HS34_24228:8:1101:1116:65353/2
ACGAAATAAATAAAAAGGTATTTAAAACCAAAAATGATAATANNCAATATGTTTANNNNNCATTTAATATTTATT
+
//<BB/FFFFFFFF/FFFFFFFFBFBFB<BBFFFFFFFB<<<!!//<<<BFFBB/!!!!!<<<<FFBBFFFBFFB
@HS34_24228:8:1101:1117:7618/2
TGCTCACCTGTTGATAAAAAATATAATAAAGTTAGAGTTACANNACAACAACAACNNNNNCAACAAGAAGCATAC
+
B//</F///////B/BFF/FF/F/B//BFF///</B///B//!!<<<F/FF/<F/!!!!!/<</FFFFFFF/F/
```

EMBL

```
ID   AF115338   standard; DNA; PRO; 591 BP.
AC   AF115338;
SV   AF115338.1
DT   03-JUN-1999 (Rel. 59, Created)
DT   23-AUG-1999 (Rel. 60, Last updated, Version 2)
DE   Pseudomonas fluorescens ECF sigma factor SigX (sigX) gene, complete cds.
KW   .
OS   Pseudomonas fluorescens
OC   Bacteria; Proteobacteria; gamma subdivision; Pseudomonadaceae; Pseudomonas.
RN   [1]
RP   1-591
RX   MEDLINE; 99369842.
RA   Brinkman F.S., Schoofs G., Hancock R.E., De Mot R.;
RT   "Influence of a putative ECF sigma factor on expression of the major outer
RT   membrane protein, OprF, in Pseudomonas aeruginosa and Pseudomonas
RT   fluorescens";
RL   J. Bacteriol. 181(16):4746-4754(1999).
RN   [2]
RP   1-591
RA   De Mot R.;
RT   ;
RL   Submitted (04-DEC-1998) to the EMBL/GenBank/DDBJ databases.
RL   F.A. Janssens Laboratory of Genetics, Applied Plant Sciences, K.
RL   Mercierlaan 92, Heverlee B-3001, Belgium
DR   SPTREMBL; Q9X4L7; Q9X4L7.
FH   Key             Location/Qualifiers
FH
FT   source 1..591
FT                   /db_xref="taxon:294"
FT                   /organism="Pseudomonas fluorescens"
FT                   /strain="M114"
FT   CDS 1..591
FT                   /codon_start=1
FT                   /db_xref="SPTREMBL:Q9X4L7"
FT                   /transl_table=11
FT                   /gene="sigX"
FT                   /product="ECF sigma factor SigX"
FT                   /protein_id="AAD34329.1"
FT                   /translation="MNKAQTLSTRYDPRELSDEELVARSHTELFHVTRAYEELMRRYQR
FT                   TLFNVCARYLGNDRDADDVCQEVMLKVLYGLKNLEGKSKFKTWLYSITYNECITQYRKE
FT                   RRKKRLMDALSLDPLEEASEEKLLQPEEKGGLDRWLVYVNPIDRGILVLRFVAELEFQE
FT                   IADIMRMSLSATMSRVRKALDKLREKFAGETET"
SQ   Sequence 591 BP; 157 A; 133 C; 170 G; 131 T; 0 other;
     atgaataaag cccaaacgct atccacgcgc tacgaccccc gcgagctctc tgatgaggag        60
     ttggtcgcgc gctcgcatac cgagctttct cacgtaacgc gcgcctatga agaactgatg       120
     cggcgttacc agcgacatt atttaacgtt tgtgcgagat atcttgggaa cgatcgcgac        180
     gcagacgatg tctgtcagga agtcatgttg aaggtgctgt atggcgtgaa gaaccgcag        240
     gggaatcgca agttcaaaac gtggctctac agcatcacgt acaacgatg tattacgcag        300
     tatcggaagg aacggcgaaa gcgtcgcttg atggacgcat tgagtcttga ccccctcgag        360
```

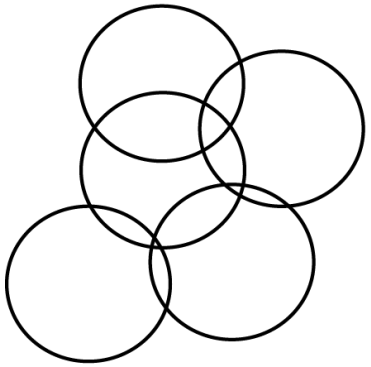
EMBL Flat File

- **Header**
  - Title
  - Taxonomy
  - Citation
- **Features (AA seq)**
- **DNA Sequence**

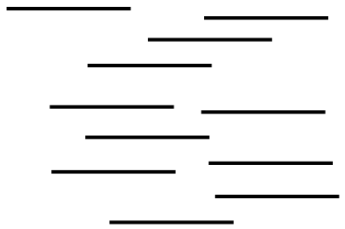
SAM/BAM

```
HS34_24228:8:2109:7462:34726    65      Pf3D7_01_v3     21      0       23M140N52M      Pf3D7_11_v3     2037615 0       CTAAACCCTAAACCCTGAAC
CCTAACCCTGAACCCTGAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAAC BBBBBFFFFFFFFBFBB/FF/FBBFF<FFFFFFFFFFFFFFFFFFF<<B//BB/<7/BFF//B<FBFF<<//<B/     AS:i
:-12    ZS:i:-12        XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YT:Z:UP XS:A:-  NH:i:3
HS34_24228:8:1213:10197:86106   65      Pf3D7_01_v3     48      255     43M229N32M      Pf3D7_07_v3     1444731 0       CCTGAACCCTAAACCCTAAA
CCCTAAACCCTAAACCCTAAACCCTAAACCTAAAACCTAAAACCCAAAACCCTGA BBBBBFFFFFFFFFFFFFFFF<FFFFFFFFFF<FFFFFFBFBFFFFF<//FFBBB/FFFFFF/FBF/<//FFB///     AS:i
:-10    XN:i:0  XM:i:2  XO:i:0  XG:i:0  NM:i:2  MD:Z:24G39T10    YT:Z:UP XS:A:-  NH:i:1
HS34_24228:8:2109:7462:34726    321     Pf3D7_01_v3     56      0       23M105N52M      Pf3D7_11_v3     2037615 0       CTAAACCCTAAACCCTGAAC
CCTAACCCTGAACCCTGAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAAC BBBBBFFFFFFFFBFBB/FF/FBBFF<FFFFFFFFFFFFFFFFFFF<<B//BB/<7/BFF//B<FBFF<<//<B/     AS:i
:-12    ZS:i:-12        XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YT:Z:UP XS:A:-  NH:i:3
HS34_24228:8:2109:7462:34726    321     Pf3D7_01_v3     77      0       23M84N52M       Pf3D7_11_v3     2037615 0       CTAAACCCTAAACCCTGAAC
CCTAACCCTGAACCCTGAACCCTAAACCCTGAACCCTAAACCCTAAACCCTGAAC BBBBBFFFFFFFFBFBB/FF/FBBFF<FFFFFFFFFFFFFFFFFFF<<B//BB/<7/BFF//B<FBFF<<//<B/     AS:i
:-12    ZS:i:-12        XN:i:0  XM:i:0  XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YT:Z:UP XS:A:-  NH:i:3
HS34_24228:8:1209:16347:84469   99      Pf3D7_01_v3     144     255     66M9S   =       242     173     ACCCTAACCCTAAACCCTAAACCTAAAACCCTGAAC
CCTAAACCCTGAACCCTGAACCCTAAACCCAAACCGCCA BBBBB<FFFFFF</FFFFFBFFFFBFF</FFBFBFFFFB//FFFF<///<FFFBBBFFFFB/<BBF/B//<////     AS:i:-9 XN:i:0  XM:i
:0      XO:i:0  XG:i:0  NM:i:0  MD:Z:66 YS:i:0  YT:Z:CP XS:A:-  NH:i:1
HS34_24228:8:1209:16347:84469   147     Pf3D7_01_v3     242     255     75M     =       144     -173    CCCTGAACCCTAAACCCTAAACCCTGAACCCTGAAC
CCTAAAACCTAAACCCTAAACCCTAAACCCTAAACCCTG <F<///BF<</</B<</</BB<//FBFB<<BF<FBF/FFF/FB</FBF<FBBFFB//BFFFFF<<FB/<BB<B     AS:i:0  XN:i:0  XM:i
:0      XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YS:i:-9 YT:Z:CP XS:A:-  NH:i:1
HS34_24228:8:2105:17855:88452   99      Pf3D7_01_v3     265     255     75M     =       675     485     CTGAACCCTGAACCCTAAAACCTAAACCCTAAACCC
TAAACCCTAAACCCTGAACCTAAACCTAAAACCTAAAAC BBBBB<BFFB<F//F//BFFFF/</FFFFF<F</<//</BFF<F<<BFFF///FFFF//F/F//BB/BBF/<<</F     AS:i:0  XN:i:0  XM:i
:0      XO:i:0  XG:i:0  NM:i:0  MD:Z:75 YS:i:-9 YT:Z:CP XS:A:-  NH:i:1
```

What do we do with these data?



1
Break many copies of the chromosome into small pieces



2
Sequence the pieces to find out the order of DNA letters in each one

```
ACTGTCTTCGCG
        TACGACTACGAC

    GATAGCATACGA
CTTCGCGCGCAG

        TCTTCGCGCGCA
            ATAGCATACGAC
CGCAGATAGCAT
```

3a
Find overlaps between these sequences to make a reference genome

```
                CGCAGATAGCAT
    TCTTCGCGCGCA ATAGCATACGAC
     CTTCGCGCGCAG       TACGACTACGAC
ACTGTCTTCGCG     GATAGCATACGA
```

```
ACTGTCTTCGCGCGCAGATAGCATACGACTACGAC
```

Annotation (find the genes)

3b
Map these sequences to a reference genome for resequencing to find differences

```
                CGCATATAGCAT
    TCTTCGCGCGCA ATAGCATAGGAC
     CTTCGCGCGCAT       TAGGACTACGAC
ACTGTCTTCGCG     TATAGCATAGGA
ACTGTCTTCGCGCGCAGATAGCATACGACTACGAC
```

- Look for interesting differences
- Build a tree

How are we going to do our bioinformatics?

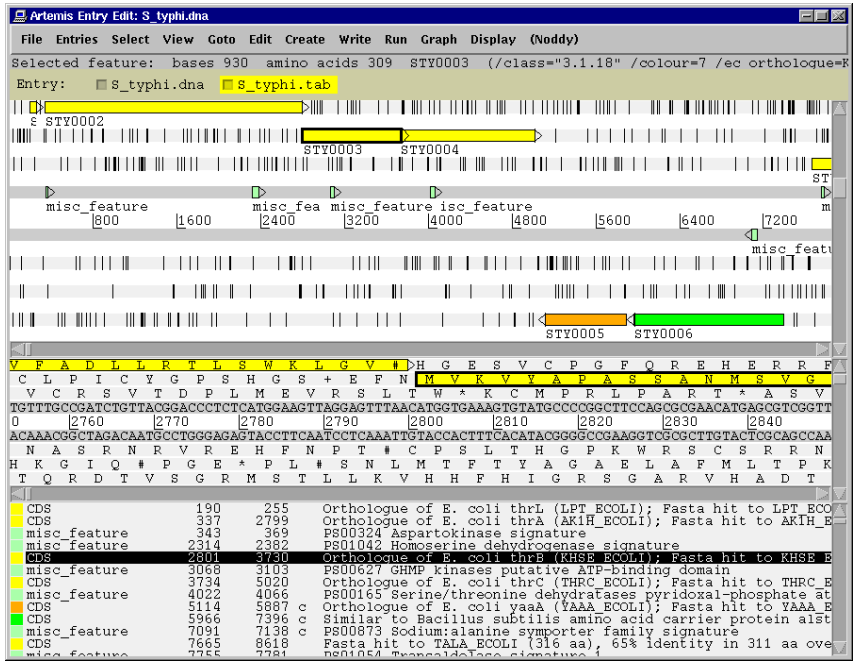
- Virtual machine with Linux
- Artemis for viewing genomes
- Various command line tools for mapping, assembling etc.
- Web-based applications

What will we do in the practicals?

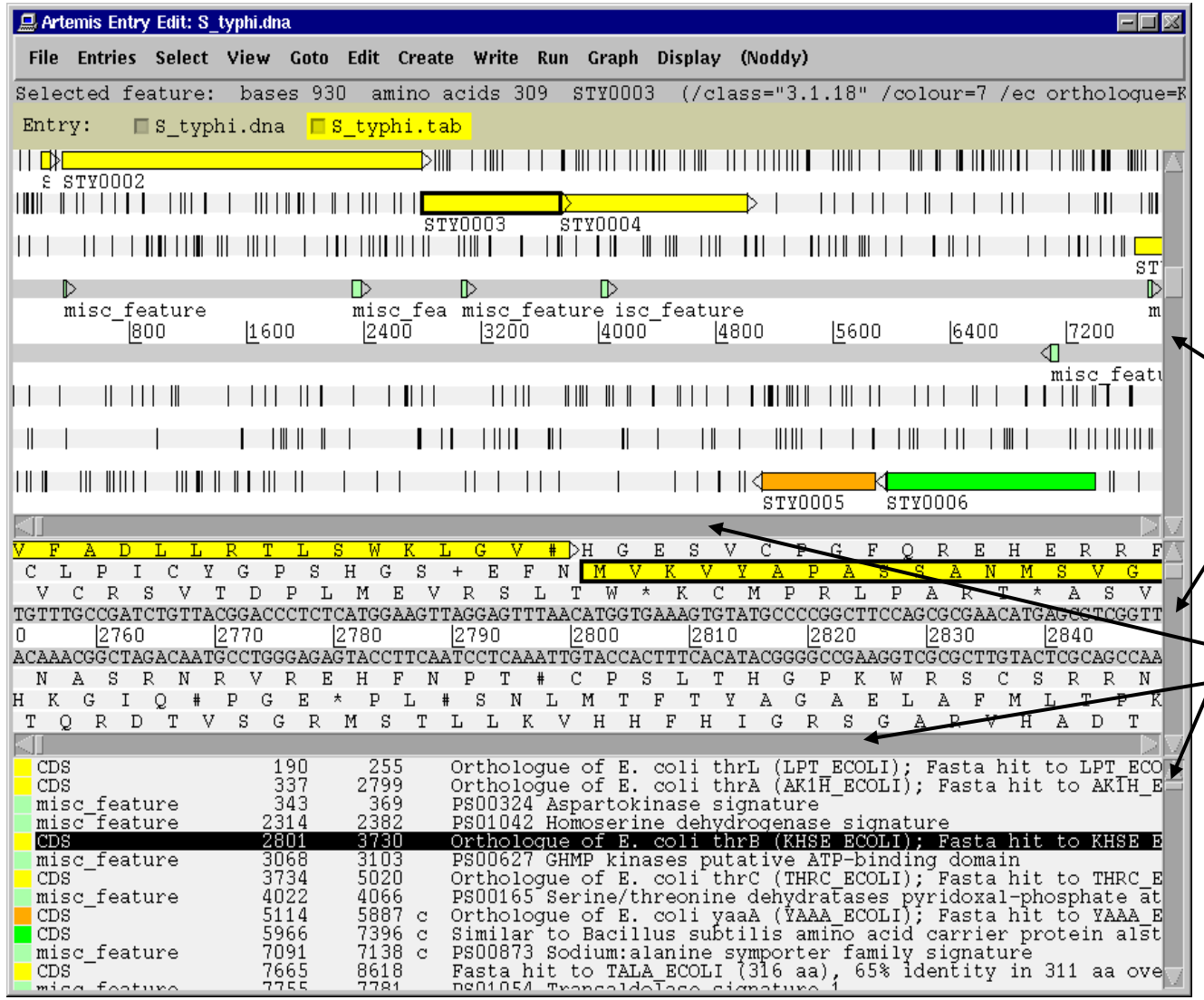
- Get familiar with the Virtual Machine
- Computer practical 1: Use Artemis to get familiar with looking at genomes (morning)
- Computer practical 2: Map short-read genome sequencing data to identify differences between closely related bacteria (afternoon)

Genome browser and annotation tool

- visualization of sequence
  - DNA
  - six frame translation
  - Panoramic and sequence view
- Annotation
  - Features
  - Mapped and listed
  - Editable
  - In layers (entry)
- perform and view analysis
  - basic analysis
  - Basic stats & index can be plotted
  - import and view the results of other searches/analysis
  - Different lines of evidence can be seen together

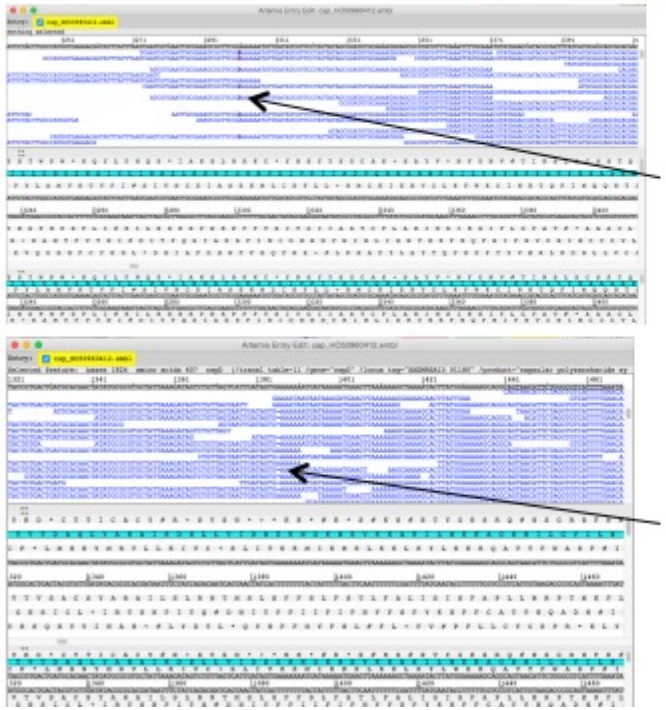


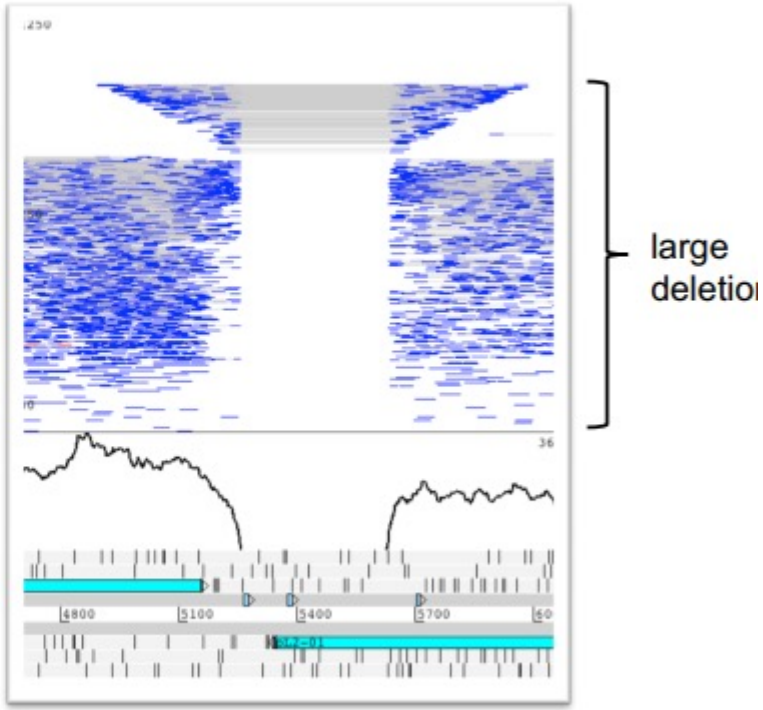
Artemis


Drop Down Menus
Entry Button Line
Main Sequence View Panel
Magnified Sequence View Panel
Feature Menu
Sliders
Sliders
Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Write Run Graph Display (Noddy)
Selected feature: bases 930 amino acids 309 STY0003 (/class="3.1.18" /colour=7 /ec orthologue=K
Entry: S_typhi.dna S_typhi.tab
STY0002
STY0003
STY0004
STY0005
STY0006
misc_feature
misc_fea
misc_feature
isc_feature
800
1600
2400
3200
4000
4800
5600
6400
7200
2760
2770
2780
2790
2800
2810
2820
2830
2840
CDS 190 255 Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit to LPT_ECO
CDS 337 2799 Orthologue of E. coli thrA (AK1H_ECOLI); Fasta hit to AK1H_E
misc_feature 343 369 PS00324 Aspartokinase signature
misc_feature 2314 2382 PS01042 Homoserine dehydrogenase signature
CDS 2801 3730 Orthologue of E. coli thrB (KHSE_ECOLI); Fasta hit to KHSE_E
misc_feature 3068 3103 PS00627 GHMP kinases putative ATP-binding domain
CDS 3734 5020 Orthologue of E. coli thrC (THRC_ECOLI); Fasta hit to THRC_E
misc_feature 4022 4066 PS00165 Serine/threonine dehydratases pyridoxal-phosphate at
CDS 5114 5887 c Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_E
CDS 5966 7396 c Similar to Bacillus subtilis amino acid carrier protein alst
misc_feature 7091 7138 c PS00873 Sodium:alanine symporter family signature
CDS 7665 8618 Fasta hit to TALA_ECOLI (316 aa), 65% identity in 311 aa ove

Viewing mapped reads in Artemis

- Illumina data (bam files)
  - identification of indels

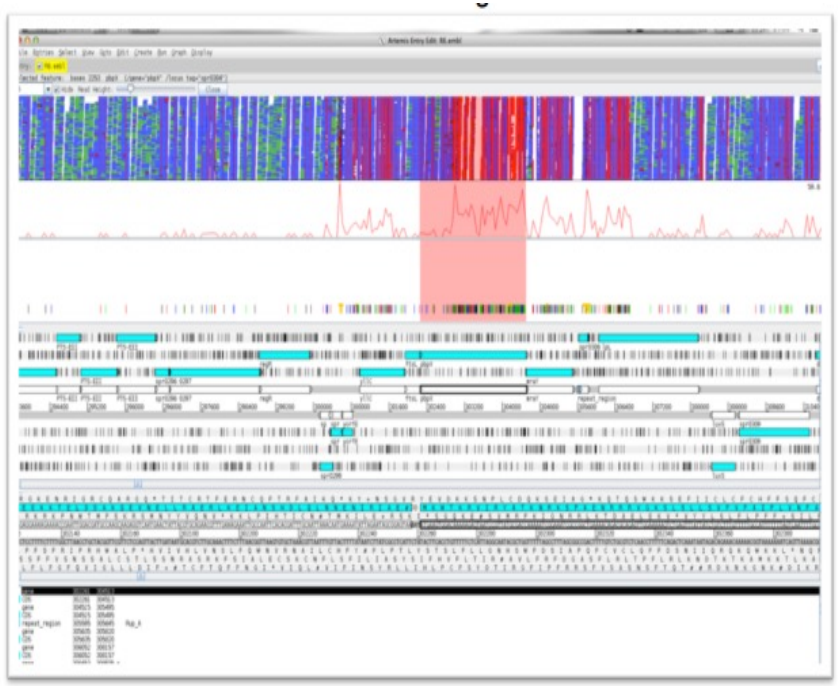




- we will see it on Module 4

Viewing mapped reads in Artemis

Single nucleotide variants (SNVs)



RNAseq data

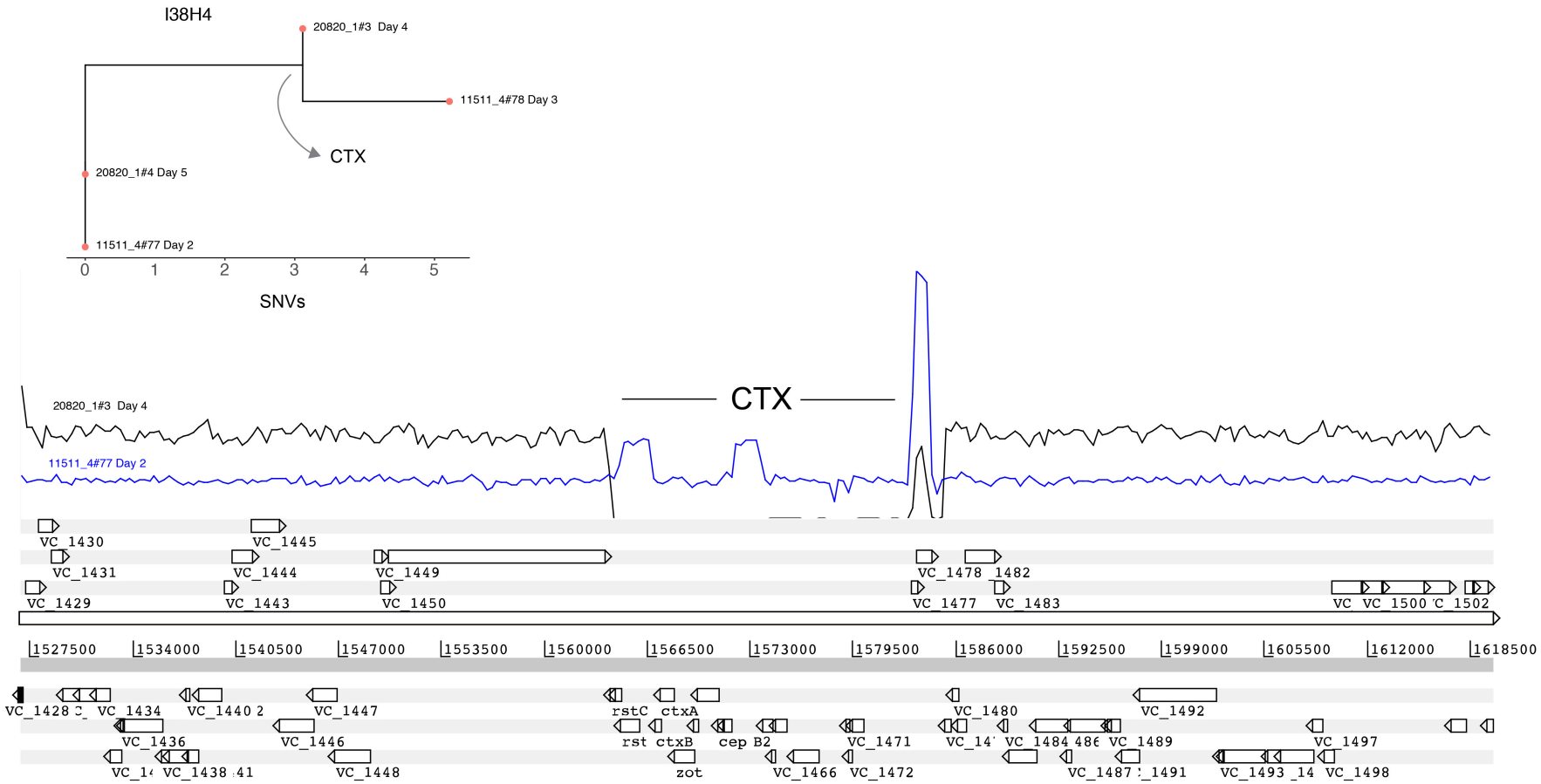
Illumina data (bam files)

Resequencing and mapping

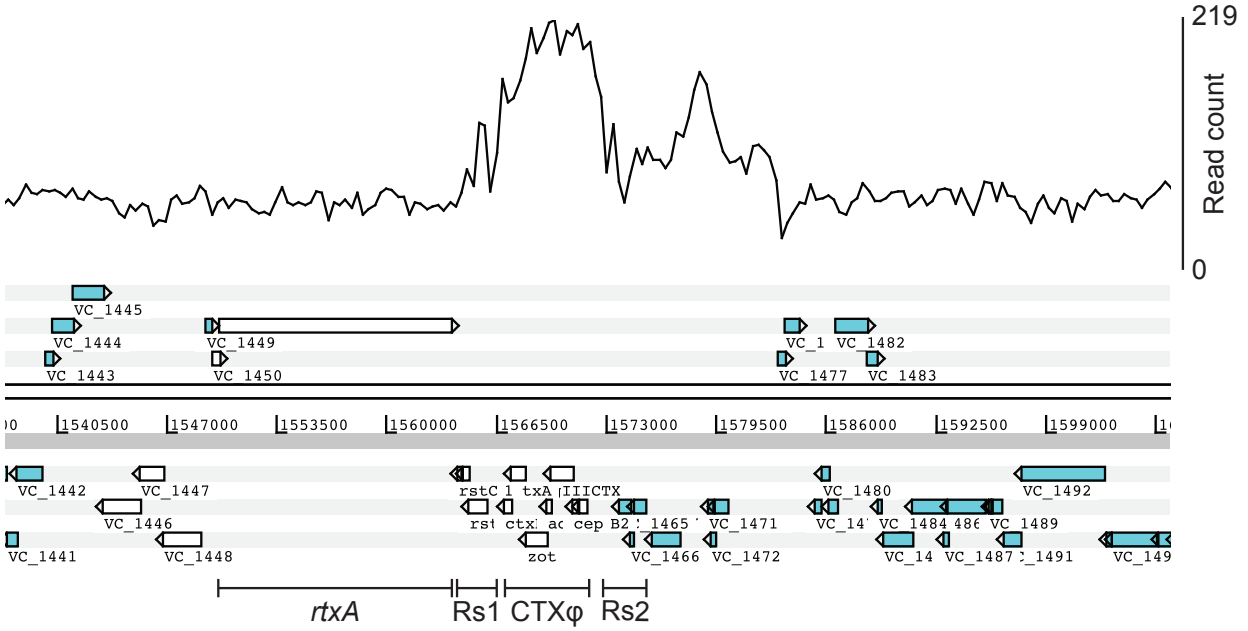
- Aims to capture information on:
  - Single Nucleotide Variants (SNVs/SNPs),
  - insertions and deletions (indels)
  - Copy Number Variants (CNVs) between individuals of a species.

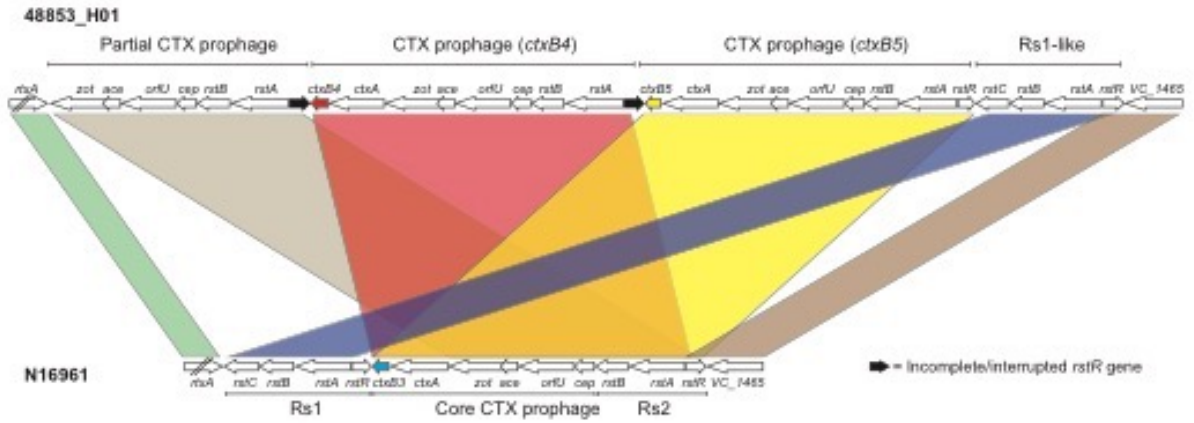
- As sequences diverge from the reference, mapping becomes progressively less effective

Gene presence / absence



Copy number variation





In this session, you will:

- Use Artemis to view genomes
- Understand genome data files
- Understand relationship between the sequence and annotation
- Understand what bacterial genomes look like and how they are arranged