

# Module 1

# Artemis

## Introduction

Artemis is a DNA viewer and annotation tool, free to download and use, written by Kim Rutherford from the Sanger Institute (Rutherford *et al.*, 2000). The program allows the user to view a range of files, from simple sequence files (e.g. fasta format) to EMBL/Genbank entries, as well as the results of sequence analyses, in a highly interactive and intuitive graphical format. Artemis is routinely used by the Pathogen Genomics group for annotation and analysis of both prokaryotic and eukaryotic genomes, and can also be used to visualize mapped data from next generation sequencing. Several types/sets of information can be viewed simultaneously within different contexts. For example, Artemis gives you the two views of the same genome region, so you can zoom in to inspect detailed DNA sequence motifs, and also zoom out to view local gene architecture (e.g. operons), or even an entire chromosome or genome, all within one screen. It is also possible to perform analyses within Artemis and save the output for future reference.

## Aims

The aim of this Module is for you to become familiar with the basic functions of Artemis using a series of worked examples. These examples are designed to take you through the most immediately useful functions. However, there will be time, and encouragement, for you to explore other menus; features of Artemis that are not described in the exercises in this manual, but which may be of particular interest to some users. Like all the Modules in this workshop, please remember:

**IF YOU DON' T UNDERSTAND, PLEASE ASK!**

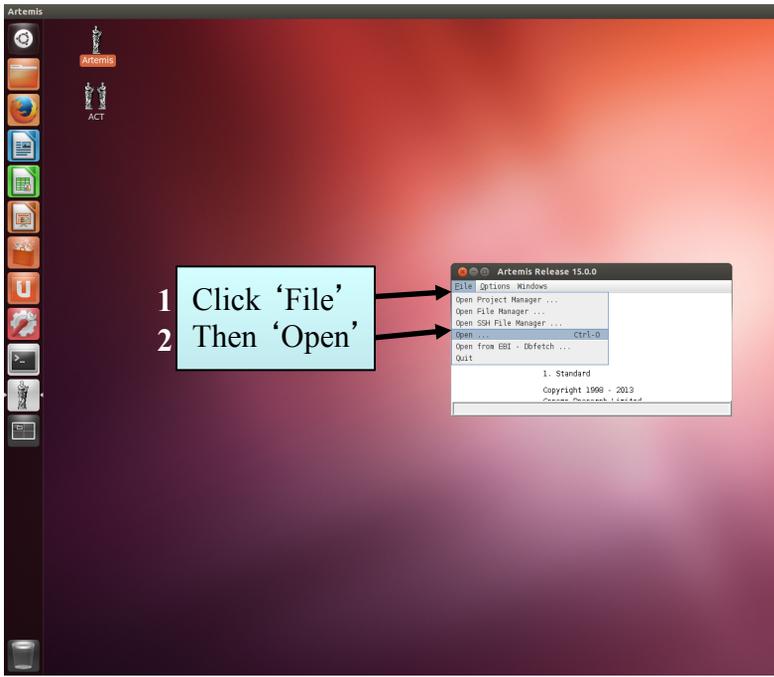
## Artemis Exercise 1

### 1. Starting up the Artemis software

Double click the Artemis icon on the desktop.

A small start-up window will appear (see below). The directory **Module\_1\_Artemis** contains all files you will need for this module.

Now follow the sequence of numbers to load up the *Salmonella* Typhi chromosome sequence. Ask a demonstrator for help if you have any problems.

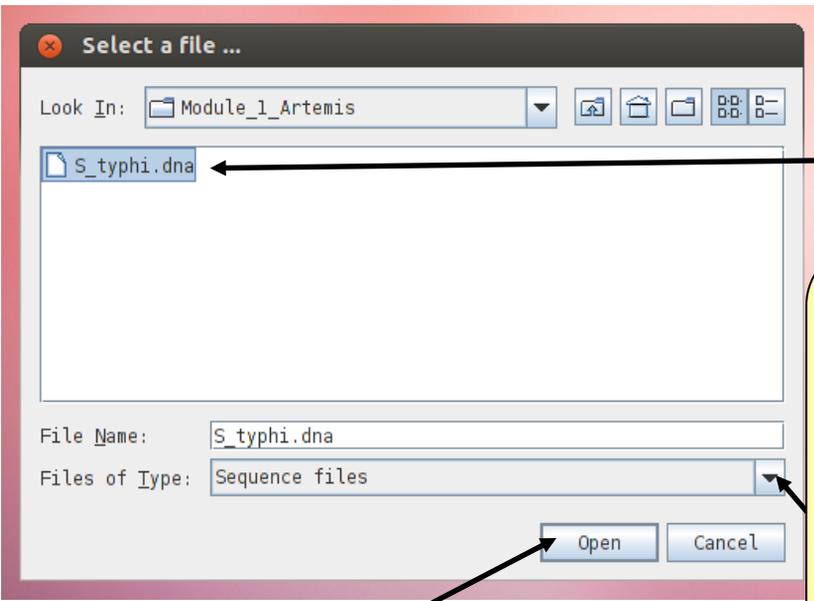


1 Click 'File'  
2 Then 'Open'

In the 'Options' menu you can switch between prokaryotic and eukaryotic mode.

You can also start Artemis from the terminal window by typing 'art'.

For simplicity it is a good idea to open a new start up window for each Artemis session and close down any sessions once you have finished an exercise.



3 Single click to select file S\_typhi.dna

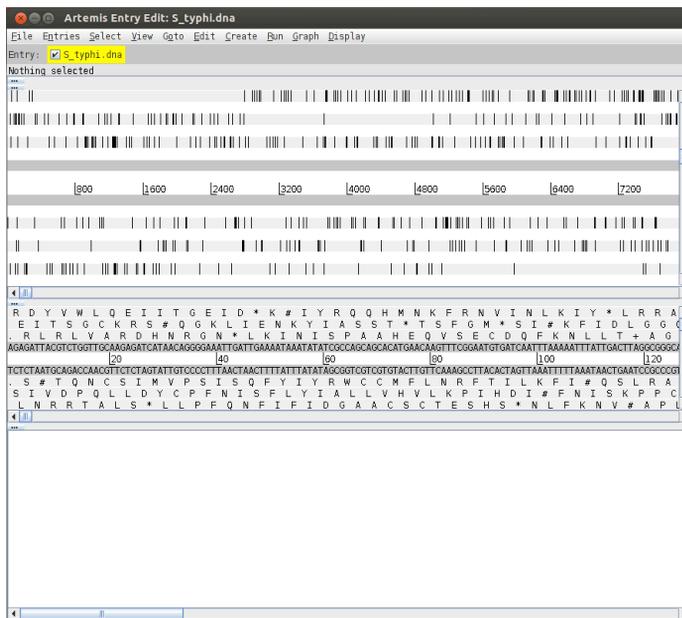
4 Single click to open file in Artemis then wait

Change to 'All Files' if you want to display all the files in the directory.

Use this feature to choose the type of file to be displayed in this panel. DNA sequence files will have the suffix '.dna'. Annotation files end with '.tab'. You can also open '.embl' files.

## 2. Loading an annotation file (entry) into Artemis

Hopefully you will now have an Artemis window like this! If not, ask a demonstrator for assistance.



Now follow the numbers to load the annotation file for the *Salmonella* Typhi chromosome.

**1**

Click 'File' then 'Read an Entry'

Entry = file

What's an "Entry"? It's a file of DNA and/or features which can be overlaid onto the sequence information displayed in the main Artemis view panel.

**2**

Single click to select file S\_typhi.tab

**3**

Single click to open file in Artemis then wait (click 'no' if an error window pops up)

### 3. The basics of Artemis

Now you have an Artemis window open let's look at what's in there.

The screenshot shows the Artemis genome browser interface. At the top, there are menu options: File, Entries, Select, View, Goto, Edit, Create, Run, Graph, Display. Below the menu, the 'Entry' field shows 'S\_typhi.dna' and 'S\_typhi.tab'. The 'Selected feature' section displays details for 'STY0004' (bases 1287, amino acids 428, gene 'thrC', product 'threonine synthase').

The main panel shows a DNA sequence with various features represented by colored boxes. The sequence is displayed in two strands: the top strand is the forward strand (5' to 3') and the bottom strand is the reverse strand (3' to 5'). Stop codons are marked with black vertical bars. Genes and other annotated features are displayed as colored boxes above and below the sequence. The features include STY0001, STY0003, STY0004, STY0007, STY0005, and STY0006.

The bottom panel shows a list of features with their coordinates and descriptions. The selected feature, STY0004, is highlighted in yellow. The list includes:

Feature	Start	End	Description
CDS	190	255	Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit to LPT_ECOLI (21 aa), 98% identity in 21 aa overlap
CDS	337	2799	Orthologue of E. coli thrA (AKIH_ECOLI); Fasta hit to AKIH_ECOLI (820 aa), 94% identity in 820 aa overlap
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_ECOLI (310 aa), 94% identity in 308 aa overlap
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_ECOLI (428 aa), 93% identity in 428 aa overlap
misc_feature	4022	4066	PS00165 Serine/threonine dehydratases pyridoxal-phosphate attachment site
CDS	5114	5887	Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_ECOLI (258 aa), 86% identity in 257 aa overlap
CDS	5966	7396	Similar to Bacillus subtilis amino acid carrier protein alst ALST 5W;ALST_BACSU (045068; P40743) fasta
misc_feature	7091	7138	PS00873 Sodium:alanine symporter family signature
CDS	7665	8319	Fasta hit to TALA_ECOLI (316 aa), 65% identity in 311 aa overlap
misc_feature	7755	7781	PS01054 Transaldolase signature 1
misc_feature	8049	8102	PS00958 Transaldolase active site
CDS	8729	9319	Orthologue of E. coli mog (MOG_ECOLI); Fasta hit to MOG_ECOLI (195 aa), 94% identity in 192 aa overlap
misc_feature	8933	8974	PS01078 Molybdenum cofactor biosynthesis proteins signature 1

Numbered callouts in the image point to the following elements:

1. Drop-down menus (File, Entries, Select, View, Goto, Edit, Create, Run, Graph, Display)
2. Entry (top line) showing 'S\_typhi.dna' and 'S\_typhi.tab'
3. Main sequence view panel showing DNA strands and features
4. Zoomed-in panel showing nucleotides and amino acids
5. Feature panel showing details of various features
6. Sliders for zooming view panels
7. Sliders for scrolling along the DNA
8. Slider for scrolling feature list

1. **Drop-down menus:** There's lots in there so don't worry about all the details right now.
2. **Entry (top line):** shows which entries are currently loaded with the default entry highlighted in yellow (this is the entry into which newly created features are created). Selected feature: the details of a selected feature are shown here; in this case gene STY0004 (yellow box surrounded by thick black line).
3. This is the main **sequence view panel**. The central 2 grey lines represent the forward (top) and reverse (bottom) DNA strands. Above and below those are the 3 forward and 3 reverse reading frames. Stop codons are marked on the reading frames as black vertical bars. Genes and other annotated features (eg. Pfam and Prosite matches) are displayed as coloured boxes. We often refer to predicted genes as coding sequences or CDSs.
4. This panel has a similar layout to the main panel but is zoomed in to show nucleotides and amino acids. Double click on a CDS in the main view to see the zoomed view of the start of that CDS. Note that both this and the main panel can be scrolled left and right (7, below) zoomed in and out (6, below).
5. **Feature panel:** This panel contains details of the various features, listed in the order that they occur on the DNA. Any selected features are highlighted. The list can be scrolled (8, below).
6. **Sliders** for zooming view panels.
7. **Sliders** for scrolling along the DNA.
8. **Slider** for scrolling feature list.



## 4.2 Navigator

The Navigator panel is fairly intuitive so open it up and give it a try.

Click 'Goto'  
then Navigator

The screenshot shows the Artemis software interface. The main window displays a genomic map with various features like CDS, misc\_feature, and STY0001. A 'Goto' menu is open, showing options like 'Goto Base', 'Goto Feature With Gene Name', etc. A separate 'Artemis Navigator' dialog box is also visible, with the 'Goto Base' option selected. The dialog box contains a list of features with their coordinates and a search field.

Feature	Start	End	Strand
CDS	190	255	Ort
CDS	337	2799	Ort
misc_feature	343	369	PS0
misc_feature	2314	2382	PS0
CDS	2801	3730	Ort
misc_feature	3068	3103	PS0
CDS	3734	5020	Ort
misc_feature	4022	4066	PS0
CDS	5114	5887	c Ort
CDS	5966	7396	c Sim
misc_feature	7091	7138	c PS0
CDS	7665	8618	Fas
misc_feature	7755	7781	PS0
misc_feature	8049	8102	PS00956 Transaldolase active site
CDS	8729	9319	Orthologue of E. coli mog (MOG_ECOLI); Fasta hit to MOG_ECOLI (195 aa), 94% identity in
misc_feature	8933	8974	PS01078 Molybdenum cofactor biosynthesis proteins signature 1
CDS	8933	8933	Fasta hit to MOG_ECOLI (195 aa), 94% identity in

Check that the  
appropriate search  
button is on

Suggestions about where to go:

1. Think of a number between 1 and 4809037 and go to that base (notice how the cursors on the horizontal sliders move with you).
2. Your favourite gene name (it may not be there so you could try '*fts*').
3. Use '**Goto Feature With This Qualifier value**' to search the contents of all qualifiers for a particular term. For example using the word 'pseudogene' will take you to the next feature with the word 'pseudogene' in any of its qualifiers. Note how repeated clicking of the 'Goto' button takes you to the following pseudogene in the order that they occur on the chromosome.
4. Look at **Appendix VIII** which is a functional classification scheme used for the annotation of *S. Typhi*. Each CDS has a class qualifier best describing its function. Use the '**Goto Feature With This Qualifier value**' search to look for CDSs belonging to a class of interest by searching with the appropriate class values.
5. tRNA genes. Type 'tRNA' in the '**Goto Feature With This Key**'.
6. Regulator-binding DNA consensus sequence (real or made up!). Note that degenerate base values can be used (**Appendix X**).
7. Amino acid consensus sequences (real or made up!). You can use 'X's. Note that it searches all six reading frames regardless of whether the amino acids are encoded or not.

What are Keys and Qualifiers? See **Appendix IV**

Clearly there are many more features of Artemis which we will not have time to explain in detail. Before getting on with this next section it might be worth browsing the menus. Hopefully you will find most of them easy to understand.

## Artemis Exercise 2

This part of the exercise uses the files and data you already have loaded into Artemis from Part I. By a method of your choice go to the region from bases 2188349 to 2199512 on the DNA sequence. This region is bordered by the *fbaB* gene which codes for fructose-bisphosphate aldolase. You can use the Navigator function discussed previously to get there. The region you arrive at should look similar to that shown below (maybe you have to use the zoom sliders).

The screenshot shows the Artemis genome browser interface. The main display area shows a DNA sequence with various features represented by colored bars and labels. Two callout boxes on the right point to specific features: 'CDS features' points to a green bar labeled 'STY2365', and 'Misc features' points to a yellow bar labeled 'misc\_feature'. Below the sequence, a list of features is displayed with their coordinates and descriptions.

Feature Type	Start	End	Description
CDS	190	255	Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit to LPT_ECOLI (21 aa), 86% identity in 21 aa overlap
CDS	337	2799	Orthologue of E. coli thrA (AKIH_ECOLI); Fasta hit to AKIH_ECOLI (820 aa), 94% identity in 820 aa overlap
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_ECOLI (310 aa), 94% identity in 308 aa overlap
misc_feature	3068	3103	PS00627 GMP kinases putative ATP-binding domain
CDS	3734	5020	Orthologue of E. coli thrC (THRC_ECOLI); Fasta hit to THRC_ECOLI (428 aa), 93% identity in 428 aa overlap
misc_feature	4022	4066	PS00165 Serine/threonine dehydratases pyridoxal-phosphate attachment site
CDS	5114	5887	Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_ECOLI (258 aa), 86% identity in 257 aa overlap
misc_feature	5966	7396	c Similar to Bacillus subtilis amino acid carrier protein alst ALST SW:ALST_BACSU (Q45068; P40743) fast
CDS	7091	7138	c PS00873 Sodium:alanine symporter family signature
misc_feature	7665	8618	Fasta hit to TALA_ECOLI (316 aa), 65% identity in 311 aa overlap
misc_feature	7755	7781	PS01054 Transaldolase signature 1
misc_feature	8049	8102	PS00958 Transaldolase active site
CDS	8729	9319	Orthologue of E. coli mog (MOG_ECOLI); Fasta hit to MOG_ECOLI (195 aa), 94% identity in 192 aa overlap
misc_feature	8933	8974	PS01078 Molybdenum cofactor biosynthesis proteins signature 1

Once you have found this region have a look at some of the information available:

### **Information to view:**

#### **Annotation**

If you click on a particular feature you can view the annotation associated with it: select a CDS feature (or any other feature) and click on the 'Edit' menu and select 'Selected Feature in Editor'. A window will appear containing all the annotation that is associated with that CDS. The format for this information is constrained by that which can be submitted to the EMBL database.

#### **Viewing amino acid or protein sequence**

Click on the 'View' menu and you will see various options for viewing the bases or amino acids of the feature you have selected, in two formats i.e. EMBL (view -> selection) or fasta (view -> bases or view -> amino acids). This can be very useful when using other programs that are not integrated into Artemis e.g. those available on the Web that require you to cut and paste sequence into them.

#### **Plots/Graphs**

Feature plots can be displayed by selecting a CDS feature then clicking 'View' and 'Feature Plots'. The window which appears shows plots predicting hydrophobicity, hydrophilicity and coiled-coil regions for the protein product of the selected CDS.

In addition to looking at the fine detail of the annotated features it is also possible to look at the characteristics of the DNA covering the region displayed. This can be done by adding various plots to the display, showing different characteristics of the DNA. Some of the plots can be used to look at the protein coding potential of translation frames within the DNA, and others can be used to search for horizontally acquired DNA (such as GC frame plot).



1. To make this process faster and clearer, **switch off stop codons** by clicking with the right mouse button in the main view panel. A menu will appear with an option to de-select 'Stop Codons' (see below).
2. You will also need to temporarily **remove all of the annotated features** from the Artemis display window. In fact if you leave them on, which you can, they would be too small to see when you zoomed out to display the entire genome. To remove the annotation click on the S\_typhi.tab entry button on the grey entry line of the Artemis window shown above.

2 To de-select the annotation click here.

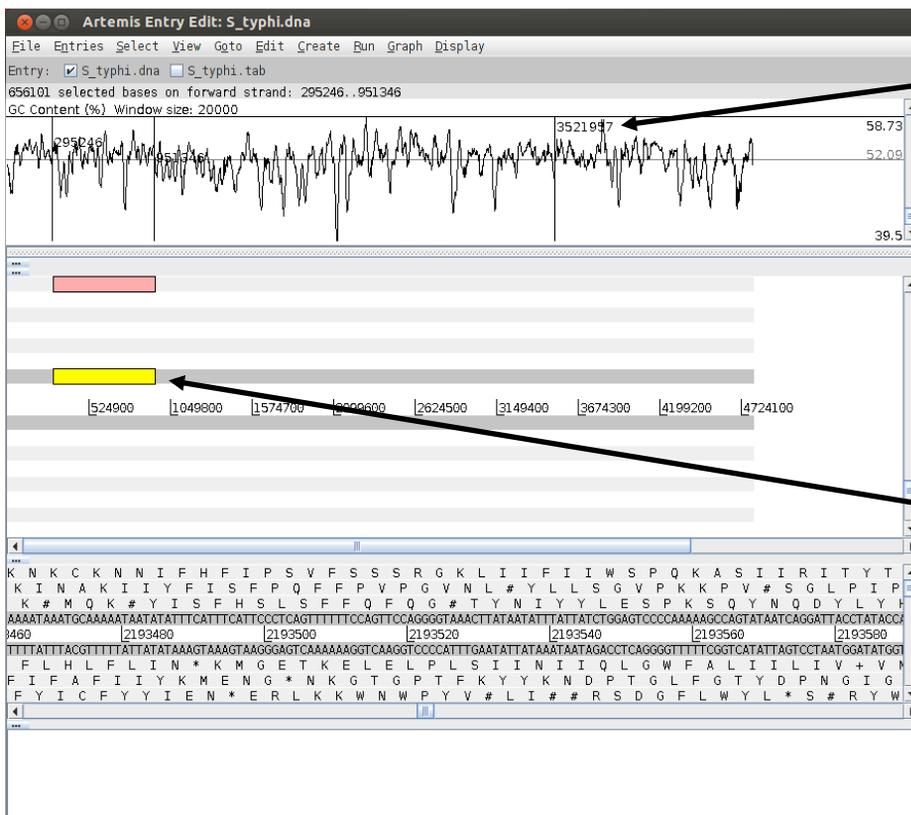
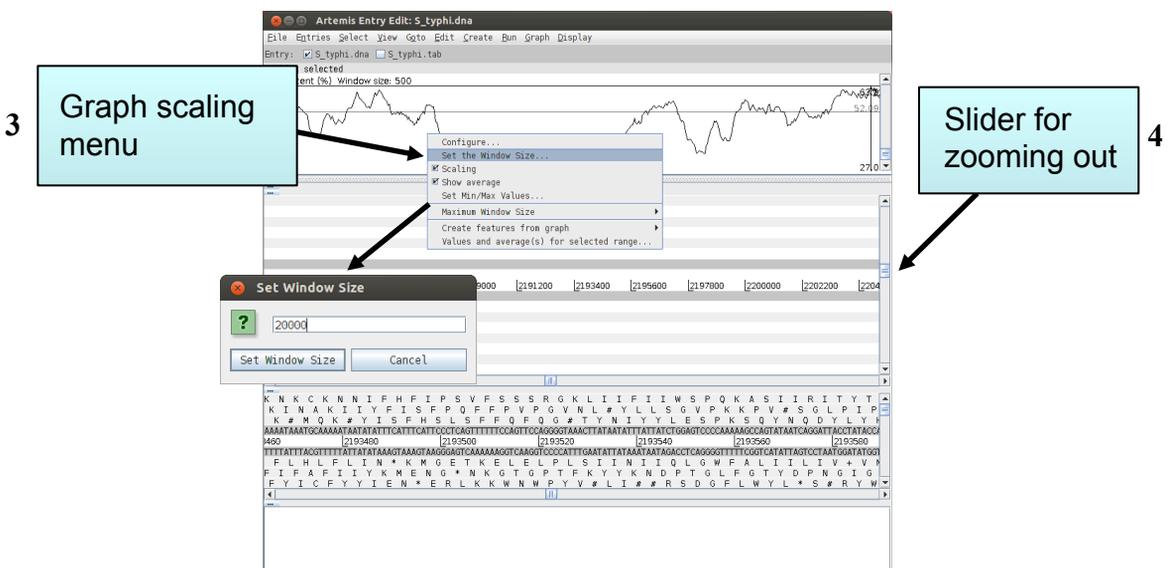
The screenshot shows the Artemis genome browser interface. The main view panel displays a genomic map with various features represented by colored arrows and boxes. A context menu is open over the main view panel, listing various options for feature display. The 'Stop Codons' option is highlighted, and a callout box points to it with the text 'Menu item for de-selecting stop codons'. Another callout box points to the 'S\_typhi.tab' entry button in the top left corner with the text 'No stop codons shown on frame lines'. The bottom left corner shows a legend for the features displayed in the main view panel.

No stop codons shown on frame lines

Menu item for de-selecting stop codons

Feature Type	Count
CDS	190
CDS	337
misc_feature	343
misc_feature	2314
CDS	2601

- One final tip is to **adjust the scaling** for each graph displayed before zooming out. This increases the maximum window size over which a single point for each plot is calculated. To adjust the scaling click with the right mouse button over a particular graph window. A menu will appear with an option "Set the Window size" (see above), set the window size to '20000'. You should do this for each graph displayed (if you get an error message press continue).
- You are now ready to zoom out by dragging or clicking the slider indicated below. Once you have zoomed out fully to see the entire genome you will need to adjust the smoothing of the graphs using the vertical graph sliders as before, to have a similar view to that shown below.



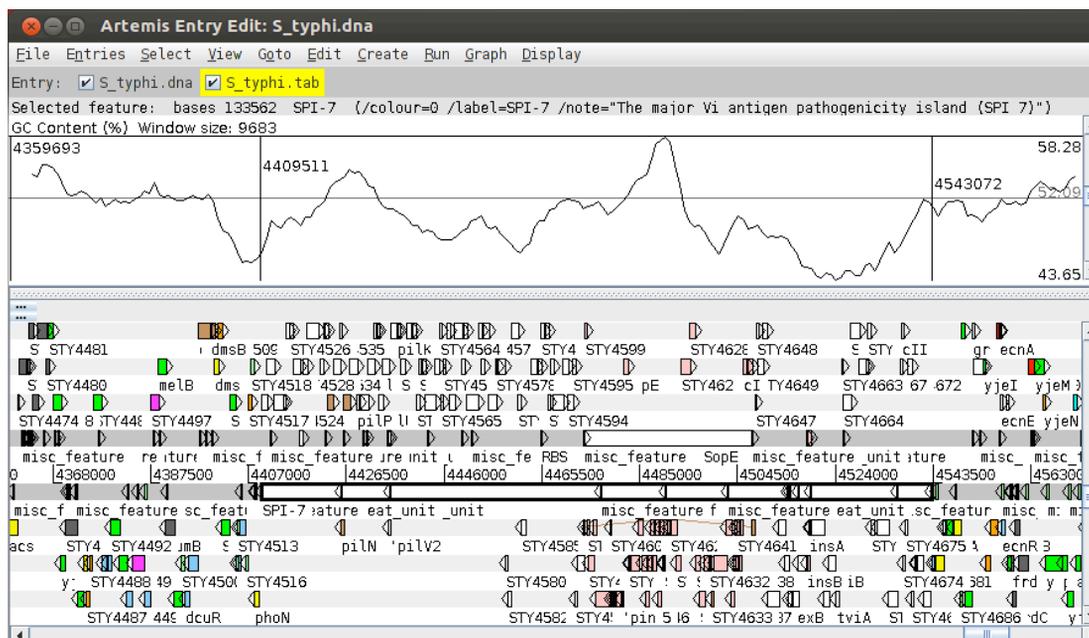
## Artemis Exercise 3

Now go to position 4409511. The next region we are looking at is defined as a *Salmonella* pathogenicity island (SPI). SPI-7, or the major Vi pathogenicity island, is ~134 kb in length and contains ~30 kb of integrated bacteriophage.

The region you should be looking at is shown below and is a classical example of a *Salmonella* pathogenicity island (SPI). The definitions of what constitutes a pathogenicity island are quite diverse. However, below is a list of characteristics which are commonly seen within these regions, as described by Hacker *et al.*, 1997.

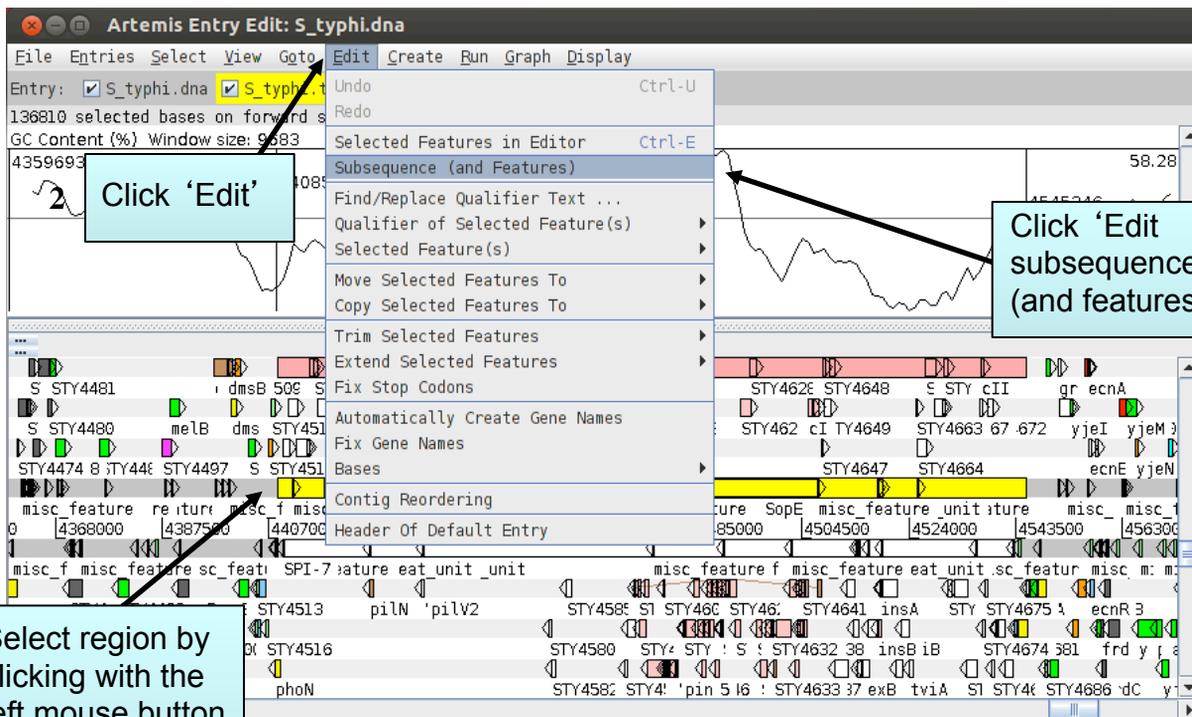
1. Often inserted alongside stable RNAs
2. Atypical G+C contents.
3. Carry virulence-related functions
4. Often carry genes encoding transposase or integrase-like proteins
5. Unstable and self-mobilisable
6. Of limited phylogenetic distribution

Have a look in and around this region and look for some of these features.



We are going to extract this region from the whole genome sequence and perform some more detailed analysis on it. We will aim to write and save new EMBL format files which will include just the annotations and DNA for this region.

Follow the numbers on the next page to complete the task.



1  
Select region by clicking with the left mouse button & dragging

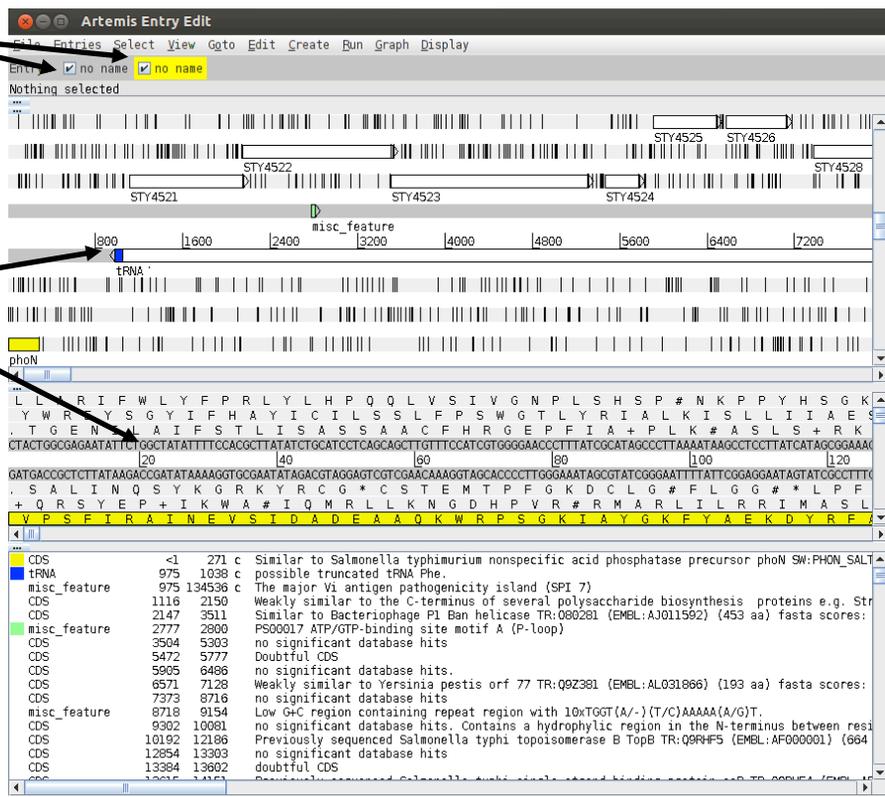
2  
Click 'Edit'

3  
Click 'Edit subsequence (and features)'

A new Artemis window will appear displaying only the region that you highlighted

Note the entry names have changed

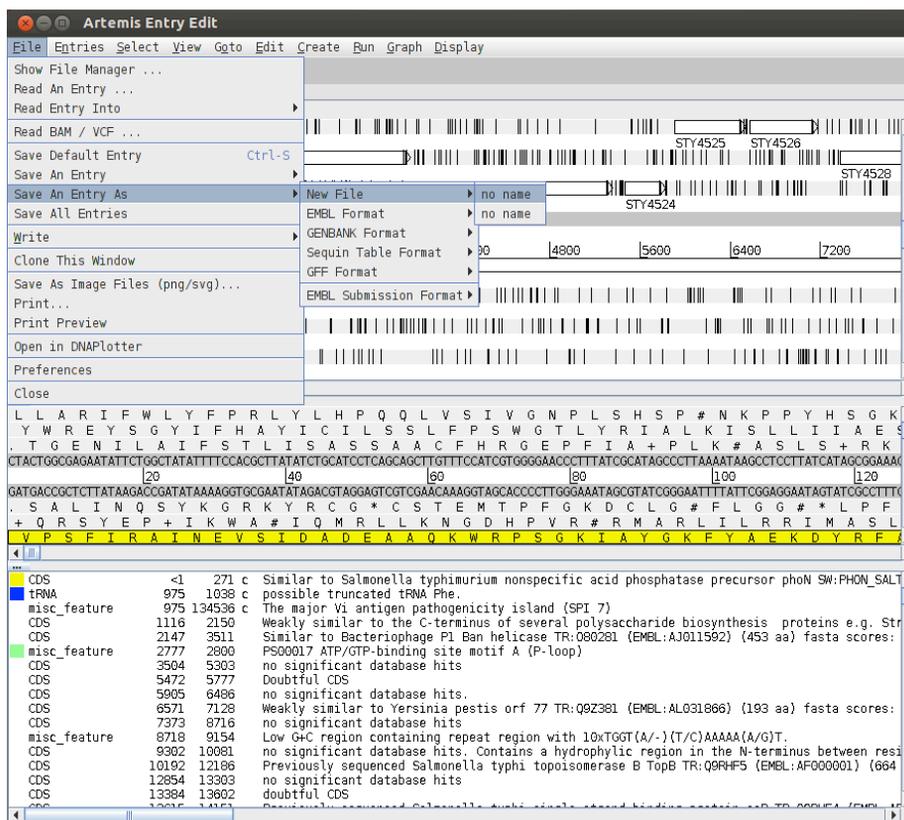
Note the bases have been renumbered from the first base you selected.



Note that the two entries on the grey 'Entry' line are now denoted 'no name'. They represent the same information in the same order as the original Artemis window but simply have no assigned 'Entry' names. As the sub-sequence is now viewed in a new Artemis session, this prevents the original files (S\_typhi.dna and S\_typhi.tab) from being over-written.

We will save the new files with relevant names to avoid confusion. So click on the 'File' menu then 'Save An Entry As' and then 'New File'. Another menu will ask you to choose one of the entries listed. At this point they will both be called 'no name'. Left click on the top entry in the list. A window will appear asking you to give this file a name. Save this file as spi7.dna

Do the same again for the second unnamed entry and save it as spi7.tab



We are going to look at this region in more detail and to attempt to define the limits of the bacteriophage that lies within this region. Luckily for us all the phage-related genes within this region have been given a colour code number 12 (pink; for a list of the other numerical values that Artemis will display as colours for features see **Appendix IX**). We are going to use this information to select all the relevant phage genes using the Feature selector as shown below and then define the limits of the bacteriophage.

First we need to create a new entry (click 'Create' then 'New Entry'). Another entry will appear on the entry line called, you guessed it, 'no name'. We will eventually copy all our phage-related genes into here.

1 Click 'Select' then 'Feature Selector'

Make sure the buttons are selected

2 Set Key to 'CDS' and Qualifier to 'colour'

3 Type search term

4 Click to select features containing search term

5 Click to view selected features in a list

6 feature list

The screenshot shows the Artemis Entry Edit window with the Feature Selector dialog open. The dialog has 'CDS' selected for the Key and 'colour' for the Qualifier. The search text is '12'. The 'Select' button is highlighted. Below the dialog, a window titled 'All features with key "CDS" with qualifier "colour" containing text "12"' is shown, displaying a list of features with columns for ID, coordinates, and description.

ID	Coordinates	Description
CDS	65714 66091	no significant database hits
CDS	66179 66396	Similar to Escherichia coli prophage P2 Ogr protein SW:09PK
CDS	66494 67964	Similar to Bacteriophage P2 late gene control protein D SW:VPL
CDS	67561 68046	Similar to Bacteriophage P2 complete genome U essential tail
CDS	68948 70826	Similar to Bacteriophage 186 protein G TR:Q37848 (EMBL:U32222)
CDS	70819 70938	Similar to Bacteriophage 186 Orf52 H TR:O80316 (EMBL:U32222)
CDS	70953 71255	Similar to Bacteriophage P2 complete genome E essential tail
CDS	71310 71625	Similar to Bacteriophage P2 major tail tube protein FI SW:VPL
CDS	71835 73007	Similar to Bacteriophage P2 major tail sheath protein FI SW:VPL
CDS	73542 74264	Similar to Salmonella typhimurium invasion-associated secret
CDS	74482 74969	Similar to Bacteriophage P2 probable tail fiber assembly prot
CDS	74576 76495	Similar to Bacteriophage P2 probable tail fiber protein SW:VPL
CDS	76492 77097	Similar to Bacteriophage P2 tail protein I SW:VPL_BFP2 (P2670)
CDS	77090 77998	Similar to Bacteriophage P2 baseplate assembly protein J SW:VPL
CDS	77895 78344	Similar to Bacteriophage P2 baseplate assembly protein W SW:VPL
CDS	78341 78919	Similar to Bacteriophage P2 baseplate assembly protein V SW:VPL
CDS	78986 79434	Similar to Bacteriophage P2 tail completion protein S SW:VPL
CDS	79427 79658	Similar to Bacteriophage P2 tail completion protein R SW:VPL
CDS	79954 80379	Similar to Bacteriophage P2 protein LysB protein involved in
CDS	80379 80750	no significant database hits. Contains possible membrane span
CDS	80761 81231	Similar to Serratia marcescens putative phage lysoczyme NucD TR

The genes listed in (6) are only those fitting your selection criteria. They can be copied or cut / moved in to a new entry so we can view them in isolation from the rest of the information within spi7.tab.

Firstly in window (6) select all of the CDSs shown by clicking on the 'Select' menu and then selecting 'All'. All the features listed in window (6) should now be highlighted. To copy them to another entry (file) click 'Edit' then 'Copy Selected Features To' then 'no name'. Close the two smaller feature selector windows and return to the SPI-7 Artemis window. You could rename the 'no name' entry as phage.tab, as you did before. Temporarily remove the features contained in 'spi7.tab' file by left clicking on the entry button on the grey entry line. Only the phage genes should remain.

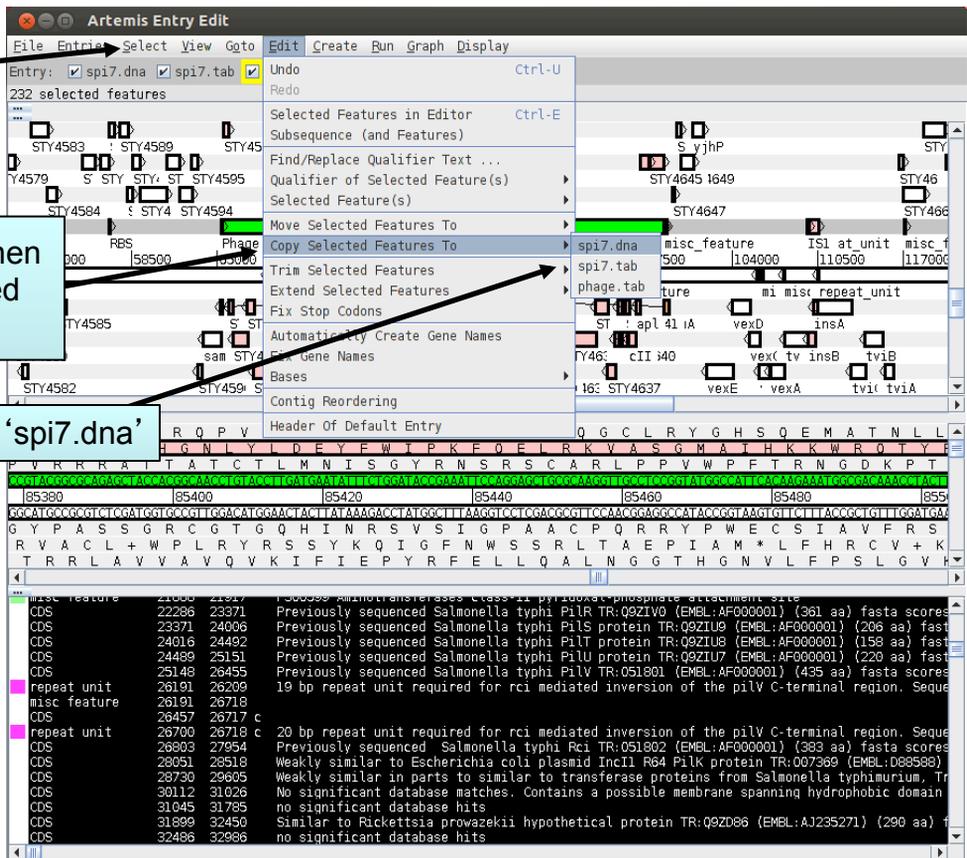


Your final task is to write out the spi7 files in EMBL submission format, and create a merged annotation and sequence file in EMBL submission format. In Artemis you are going to copy the annotation features from the '.tab' file into the '.dna' file, and then save this entry in EMBL format. Don't worry about error messages popping up. This is because not all entries are accepted by the EMBL database.

1 Click 'Select' then 'All'

2 Click 'Edit', then 'Copy Selected Features To'

3 Select 'spi7.dna'

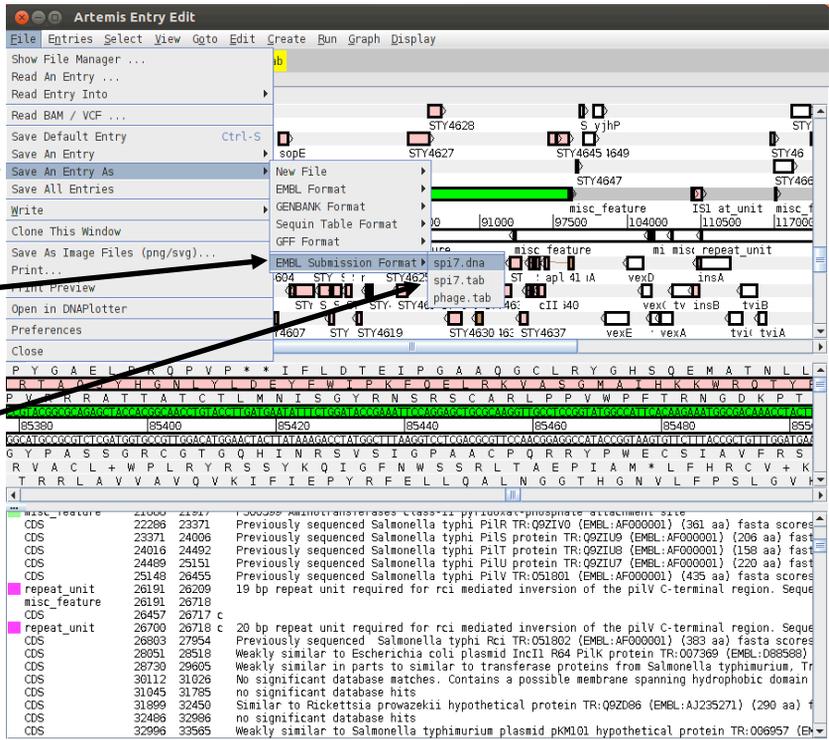


4 Click 'File' then 'Save An Entry As'

5 'EMBL Submission Format'

6 Select 'spi7.dna'

7 Save file as spi7.embl



Now open the EMBL format file that you have just created in Artemis.

tRNA	1	64	c	possible truncated tRNA Phe.
misc_feature	1	133562	c	The major Vi antigen pathogenicity island (SPI 7)
CDS	142	1176		Weakly similar to the C-terminus of several polysaccharide biosynthesis proteins e.g. Str
CDS	1173	2537		Similar to Bacteriophage P1 Ban helicase TR:080281 (EMBL:AJ011592) (453 aa) fasta scores:
misc_feature	1803	1826		PS00017 ATP/GTP-binding site motif A (P-loop)
CDS	2530	4329		no significant database hits
CDS	4498	4803		Doubtful CDS
CDS	4931	5512		no significant database hits.
CDS	5597	6154		Weakly similar to Yersinia pestis orf 77 TR:Q9Z381 (EMBL:AL031866) (193 aa) fasta scores:
CDS	6399	7742		no significant database hits
misc_feature	7744	8180		Low G+C region containing repeat region with 10xTGGT(A/-)(T/C)AAAA(A/G)T.
CDS	8328	9107		no significant database hits. Contains a hydrophilic region in the N-terminus between resi
CDS	9218	11212		Previously sequenced Salmonella typhi topoisomerase B TopB TR:Q9RHF5 (EMBL:AF000001) (664
CDS	11880	12329		no significant database hits
CDS	12410	12628		doubtful CDS
CDS	12641	13177		Previously sequenced Salmonella typhi single strand binding protein ssB TR:Q9RHE4 (EMBL:AF

You will see that the colours of the features have now changed. This is because not all the qualifiers in the previous entry are accepted by the EMBL database, so some have not been saved in this format. This includes the '/colour' qualifier, so Artemis displays the features with default colours.

When you download sequence files from EMBL and visualize them in Artemis you will notice that they are displayed using default colours. You can customize your own annotation files with the '/colour' qualifier and chosen number (**Appendix IX**), to differentiate features. To do this you can use the Feature Selector to select certain features and annotate them all using the 'Edit', 'Change Qualifiers of Selected' function.

## Artemis Exercise 4

This exercise will introduce you to database searches and will give you a first insight in the annotation of genes.

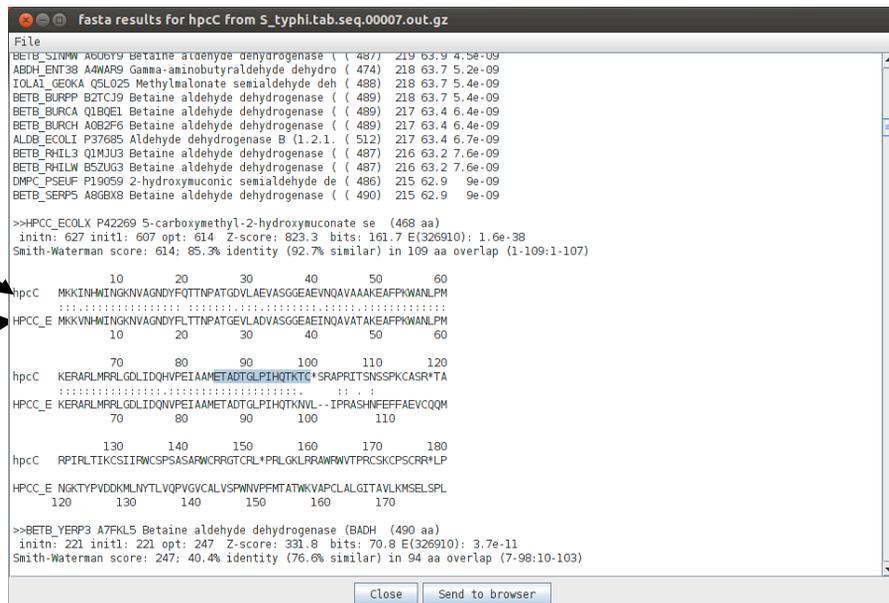
The gene you will work on is *hpcC* (STY1136). Go to this gene by using one of the different methods you have learned so far.

As you can see the gene is full with stop codons indicating that we are looking at a pseudogene. To correct the annotation we are going to use database search. Follow now the numbers in the figure below to start a database search. The search may take a couple of minutes to run; a banner will pop up to tell you when its complete (3).

The screenshot shows the Artemis software interface for editing the *S. typhi* dna entry. The main window displays a gene map with features like *hpcC* and *hpcB*. A search menu is open, showing options for various database searches. A 'fasta pro' dialog box is open, indicating the search is complete. The search results table is visible at the bottom of the screen.

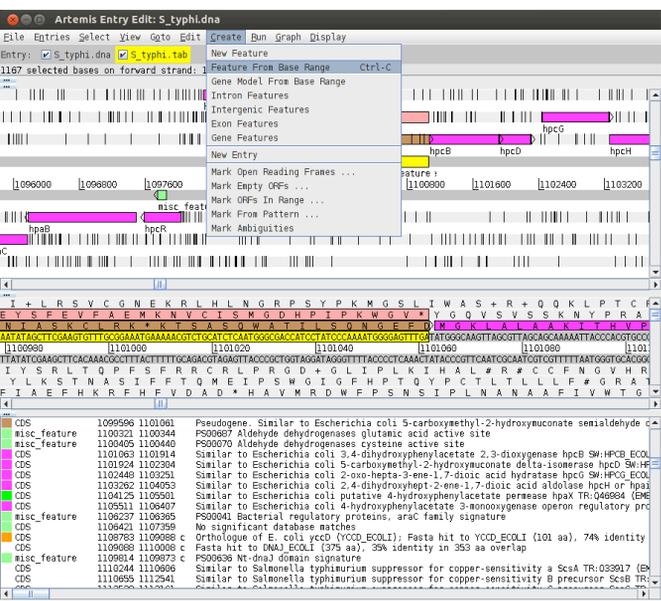
Feature	Start	End	Description
CDS	1099596	1101061	Pseudogene. Similar to Escherichia coli 5-carboxymethyl-2-hydroxyumconate semialdehyde d
misc_feature	1100321	1100344	PS00687 Aldehyde dehydrogenases glutamic acid active site
misc_feature	1100405	1100440	PS00070 Aldehyde dehydrogenases cysteine active site
CDS	1101063	1101914	Similar to Escherichia coli 3,4-dihydroxyphenylacetate 2,3-dioxygenase hpcB SW:HPCB ECOL
CDS	1101924	1102304	Similar to Escherichia coli 5-carboxymethyl-2-hydroxyumconate delta-isomerase hpcD SW:HF
CDS	1102448	1103251	Similar to Escherichia coli 2-oxo-hepta-3-ene-1,7-dioic acid hydratase hpcG SW:HPGC ECOL
CDS	1103262	1104053	Similar to Escherichia coli 2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase hpcH or hpa
CDS	1104125	1105501	Similar to Escherichia coli putative 4-hydroxyphenylacetate permease hpaX TR:Q46984 (EME
CDS	1105511	1106407	Similar to Escherichia coli 4-hydroxyphenylacetate 3-monooxygenase operon regulatory pro
misc_feature	1106237	1106395	PS00041 Bacterial regulatory proteins, araC family signature
CDS	1106421	1107359	No significant database matches
CDS	1108783	1109088 c	Orthologue of E. coli yccD (YCCD_ECOLI); Fasta hit to YCCD_ECOLI (101 aa), 74% identity
CDS	1109088	1110098 c	Fasta hit to DNAI_ECOLI (375 aa), 35% identity in 353 aa overlap
misc_feature	1109814	1109873 c	PS00636 Nt-dnaJ domain signature
CDS	1110244	1110606	Similar to Salmonella typhimurium suppressor for copper-sensitivity a ScaA TR:Q33917 (EM
CDS	1110655	1112541	Similar to Salmonella typhimurium suppressor for copper-sensitivity B precursor ScaB TR:
CDS	1112550	1113101	Similar to Salmonella typhimurium suppressor for copper-sensitivity C precursor ScaC TR:

To view the search results click 'View', then 'Search Results', then 'fasta results'. The results will appear in a scrollable window. Scroll down to the first sequence comparison and you should see the results as shown in the next figure.

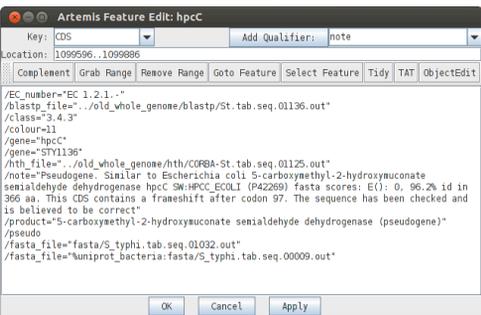


Can you see where the stop codon has been introduced into the sequence of our gene of interest? Search for the highlighted amino acid sequence in *hpcC*. Have a look if you can find the subsequent amino acids of the database hit in any of the three reading frames. You will see the sequence can be found in the second frame! What has happened? The last amino acid in common is a K then the amino acids start to differ till the stop codon. The amino acid in common is a K then the next base is an A, too. This little homopolymeric region can cause trouble during DNA replication if the polymerase slips and introduces an additional 'A'. This shifts the proper reading frame into the second frame.

To correct the annotation we have to edit the CDS now. Left click on the right amino acid continuing the amino acid sequence on the second frame (have a look in the fasta title and look at the sequence of the gene in the database when you are not sure) and drag till the end of the gene. Then click 'Create' 'Feature from base range' and 'OK'. A new blue CDS feature will appear on the appropriate frame line.



As the original gene annotation is too long we have to shorten it. Click on the original *hpcC* CDS, 'Edit' 'Selected features in Editor'. A window will pop up and you can change the end position in 'location' (the end position is the last base of the stop codon).



The new CDS feature can then be merged with the original gene as shown below (1-3).

A small window will appear asking you whether you are sure you want to merge these features. Another window will then ask you if you want to 'delete old features'. If you click 'yes' the CDS features you have just merged will disappear leaving the single merged CDS. If you select 'no' all of the three CDS features (the two CDSs you started with plus the merged feature) will be retained.

2

Click 'Edit'

3

'Selected Features' 'Merge'

1

Select both the original gene-model and the new CDS feature, which is to be merged with it to form a new gene

Artemis Entry Edit: S\_typhi.dna

File Entries Select View Goto Edit Create Run Graph Display

Entry: S\_typhi.dna S\_typhi.tab

2 selected features total bases 1

Undo Ctrl-U

Redo

Selected Features in Editor Ctrl-E

Subsequence (and Features)

Find/Replace Qualifier Text ...

Qualifier of Selected Feature(s)

Selected Feature(s)

Duplicate Ctrl-D

Merge Ctrl-M

Unmerge

Unmerge All Segments

Delete Ctrl-Delete

Delete Exons

Remove Introns

Convert Keys ...

Trim Selected Features

Extend Selected Features

Fix Stop Codons

Automatically Create Gene Names

Fix Gene Names

Bases

Contig Reordering

Header Of Default Entry

K P P T P A C L F T R L K R A D P A R L A # L R I L L R R S V P A D E R O D L S G \* R

N R R L R P A I S P D # N V I T P R A S H N F F F A E V C O O M N G K T Y P V D D

T A D I G L P I H O T K T C \* S R A P R I T S N S S P K C A S R \* T A R P I R L T I

A A C C G C C G A C C G G C C T A T T C A C C A G A C T A A A A C G T G C T G A T C C C G C G C G C C T C G C A T A A C T T C G A A T T C T C G C G A A G T G T G C C A G C A G A T G A A C G G C A A G A C C T A T C C G G T T G A C G A T

099860 099880 1099800 1099900 1099920 1099940 1099960

T T G C C G C T G T G G C C G A C G G A T A A G T G T C T G A T T T T G C A C G A C T A G G C G C G C G G A G C G T A T T G A A G C T T A A G A A G C G G C T T C A C A C G G T C G T C T A C T T G C C G T T C T G G A T A G G C C A A C T G C T A

G G V G A Q R N V L S F R A S G A R R A Y S R I R R R L T G A S S R C S R D P Q R Y

F R R O R G A A E G S + F T S I G R A E C L K S N K A S T H W C I F P L V + G T S S

V A S V P R G I \* W V L V H Q D R A G R M V E F E E G F H A L L H V A L G I R N V I

hydrogenase hpcC SW:PCC\_ECOLI (P42269) fasta scores

Artemis Entry Edit: S\_typhi.dna

File Entries Select View Goto Edit Create Run Graph Display

Entry: S\_typhi.dna S\_typhi.tab

result

Selected feature: bases 1461 amino acids 498 hpcC (/EC number="EC 1.2.1.1" /class="3.4.3" /colour=11 /gene="hpcC"/gene="STY1143")

T

hpcG

hpcC

hpcB

hpcD

hpcE

misc\_feature

0 1096000 1099000 1097600 1098400 1099200 1100000 1100800 1101600 1102400 1103

misc\_feature

hpcB

hpcR

hpcC

I (Q05353) fasta scores: E(): 0, 92.3% id in 259 aa

CD\_ECOLI (Q05354) fasta scores: E(): 0, 81.6% id in 1

I (P42270) fasta scores: E(): 0, 90.6% id in 267 aa

TR:Q47093 [EMBL:Z47799] fasta scores: E(): 0, 87.8%

L:Z37980) fasta scores: E(): 0, 87.3% id in 458 aa

tein HpaI TR:Q46985 [EMBL:Z37980) fasta scores: E():

in 101 aa overlap

BL\_117594) fasta scores: E(): 0, 99.7% id in 120 aa

**Tip:** To select more than one feature (of any type) you must hold the shift key down.

## Artemis Exercise 4 - Second part

In the first part of the exercise you have learned how to correct a gene annotation. But what if you think a gene is missing?

Remember that there are loads of genomes that were submitted to the databases several years ago and in general the annotation is not updated to take into account new data. Sometimes it is worth checking regions which look strange to you.

Go to position 2,248,400 by using one of the different methods you have learned so far. If you look carefully you will notice a region shown below which there is no predicted gene. This type of non-coding region in *Salmonella* is very unusual (this is also true for other bacteria). To determine if this non-coding region is truly as published, load the codon usage information for *Salmonella* into Artemis by following the figure below.

The file 'S\_typhi.cod' contains codon usage information taken from a public website (see below).

1

Click Graph

2

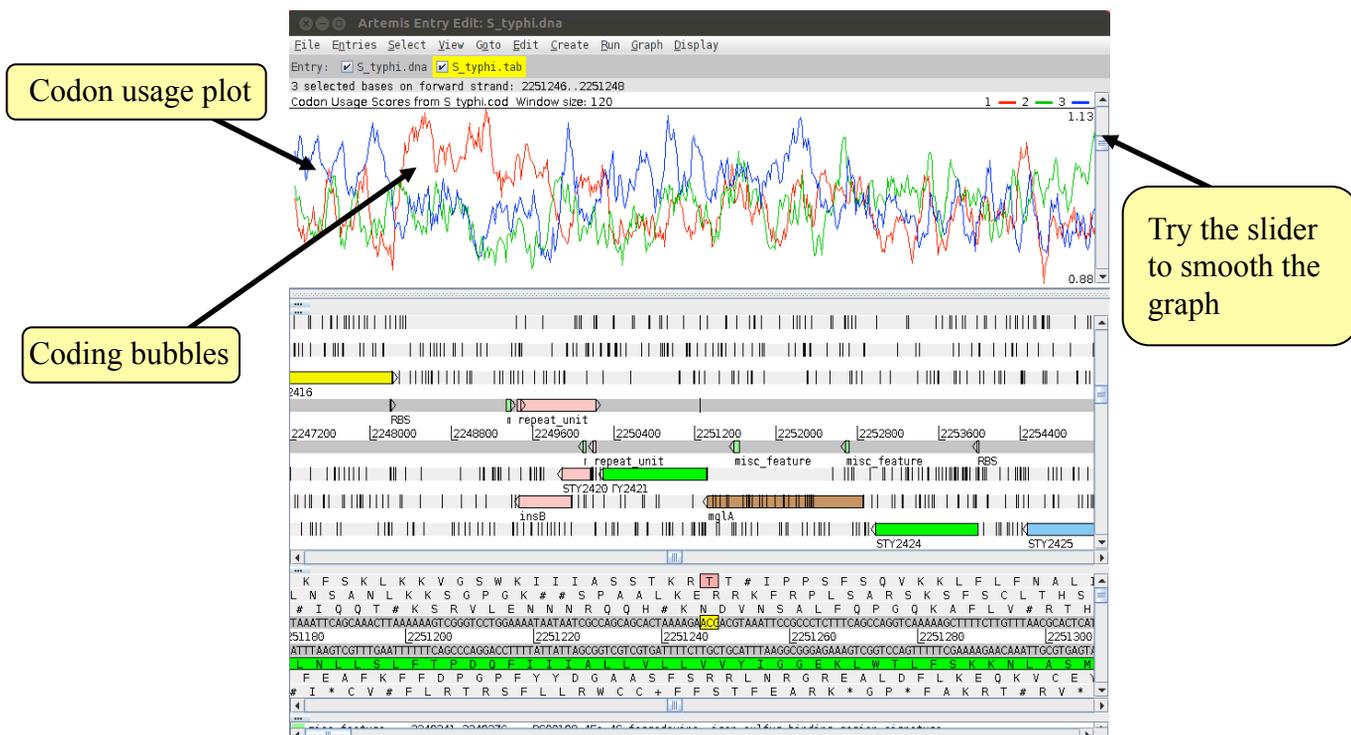
Click 'Add usage plots' and select 'S\_typhi.cod'

Non-coding

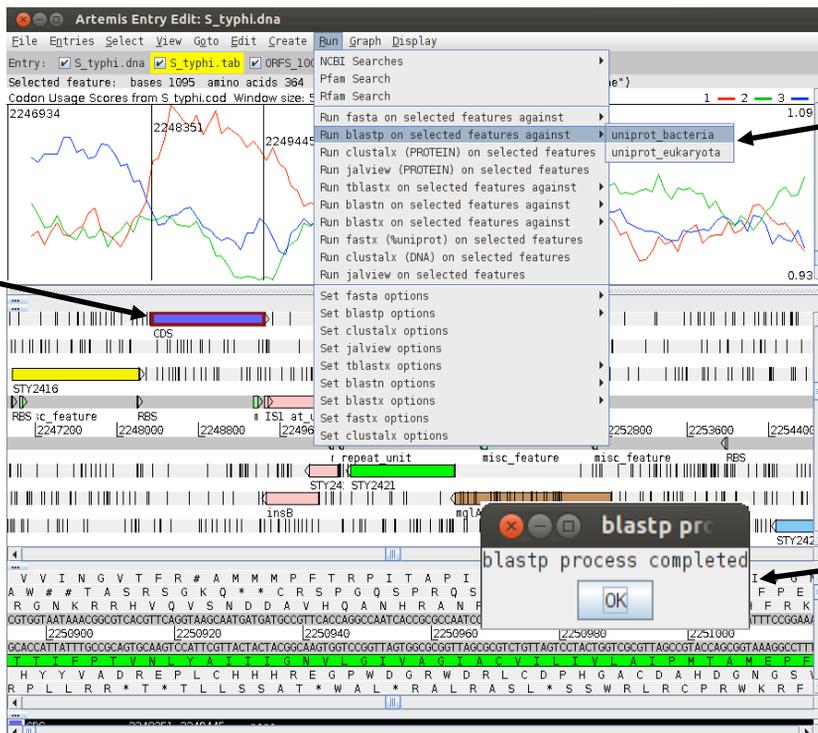
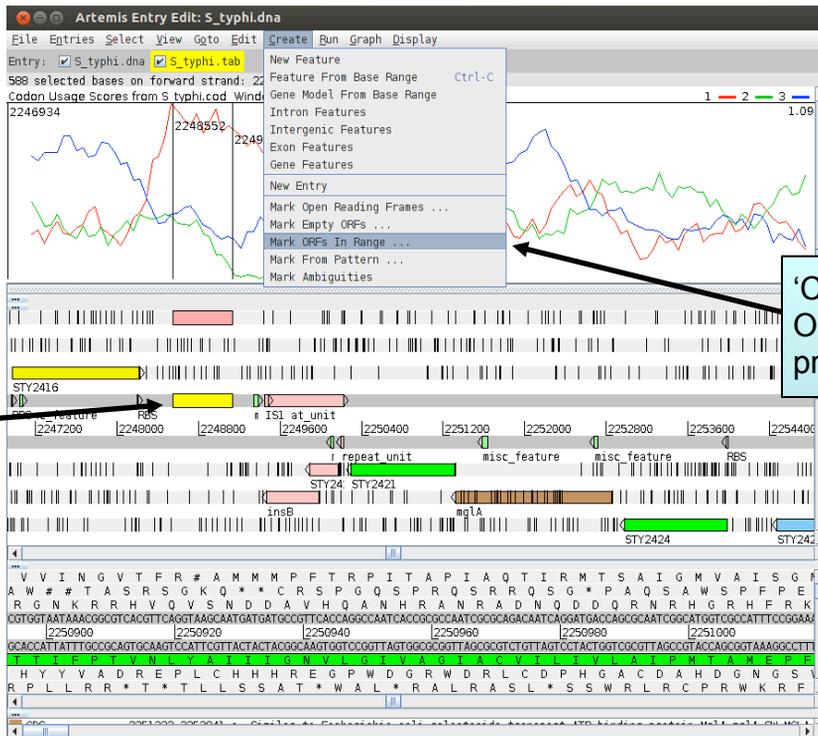
UUU	24.4	2445	UCU	12.1	1212	UAU	18.6	1864	UGU	6.5	651
UUC	15.6	1563	UCC	18.5	1854	UAC	13.3	1335	UGC	5.8	585
UUA	15.8	1582	UCA	15.3	1534	UAA	1.6	158	UGA	1.4	145
UUG	12.2	1222	UGC	9.5	952	UAG	0.5	47	UGG	13.3	1326
CUU	16.1	1614	CCU	18.5	1855	CAU	11.4	1143	CGU	14.3	1426
CUC	11.2	1128	CCC	6.8	681	CAC	7.2	725	CGC	11.7	1178
CUA	6.6	656	CCA	9.3	938	CAA	13.8	1381	CGA	6.7	665
CUG	34.1	3411	CCG	14.1	1415	CAG	27.2	2716	CGG	8.2	822
AUU	27.3	2729	ACU	14.8	1476	AAU	27.8	2699	AGU	13.2	1319
AUC	28.4	2837	ACC	28.1	2812	AAC	22.7	2266	AGC	16.4	1639
AUA	9.9	988	ACA	14.4	1445	AAA	35.4	3546	AGA	6.8	599
AUG	26.8	2597	ACG	15.5	1552	AAG	17.5	1752	AGG	4.7	472
GUU	28.3	2833	GCU	17.7	1778	GAU	33.9	3394	GGU	19.5	1949
GUC	15.4	1541	GCC	21.6	2168	GAC	28.8	2888	GGC	21.8	2898
GUA	12.8	1281	GCA	21.8	2185	GAA	34.8	3481	GGA	12.6	1259
GUG	19.4	1941	GCG	19.3	1931	GAG	28.8	2888	GGG	13.6	1359

Codon usage table taken from:  
[www.kazusa.or.jp/codon](http://www.kazusa.or.jp/codon)

When you first load the codon table into Artemis the graphs calculated for both upper and lower strands will be displayed (not shown). To add/remove one of these to/from the view click on the Graph menu and check the box alongside the option 'codon usage scores from S\_typhi.cod' (the reverse plot is also represented in this list).



Based on the codon usage table Artemis calculates for each triplet in succession a score based on how well it matches the commonly used codons in that organism. The three lines shown above represent the scores for each reading frame. If the codons for a particular frame match those of the calculated codon usage table a high score is given. Practically speaking this manifests itself as a 'coding bubble' where a gap opens up in the plot indicating that this region is likely to be coding (see above). The plot suggests that this empty region actually encodes a product. So now we have to create the open reading frame (ORF), blast the amino acid sequence and add the annotation. Follow the instruction on the next page to do this.



To view the search results click 'View', then 'Search Results', then 'blastp results'. The results will appear in a scrollable window. You see that the product of the gene is "NAD-dependent dihydropyrimidine dehydrogenase subunit PreA). To add the product to the annotation follow the instruction in the next page.

1 Click on ORF

2 'Edit', 'Selected feature in editor'

3 Select 'product' from dropdown list

4 Click 'Add qualifier'

5 Add result from blast search

The annotation of the ORF is now complete. You can add as much information as you want. Have a look at the other qualifiers if some time is left. The last thing you have to do is copy the annotated feature to `S_typhi.tab`. To do that select the feature and go to 'Edit', 'Copy selected features to' and click '`S_typhi.tab`'. Don't forget to save the tab file.