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 Infection Genomics group for annotation and analysis of both prokaryotic and eukaryotic genomes, and can also be used to visualise 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:

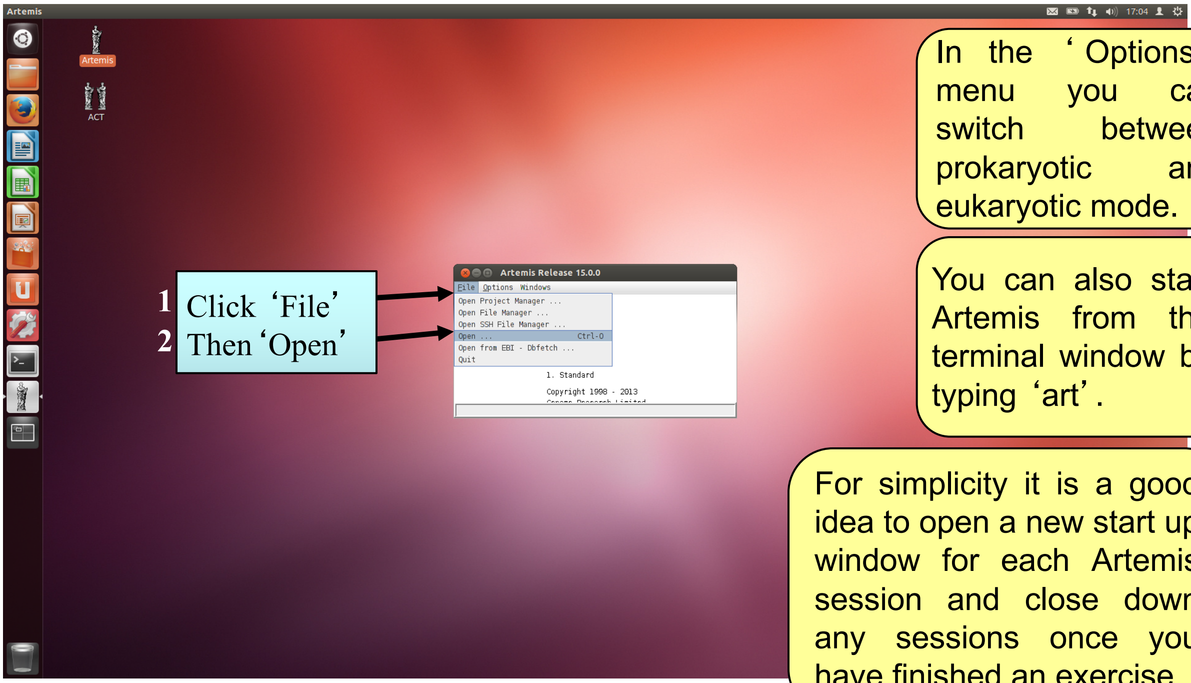
IF YOU DON’ T UNDERSTAND, PLEASE ASK!

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.

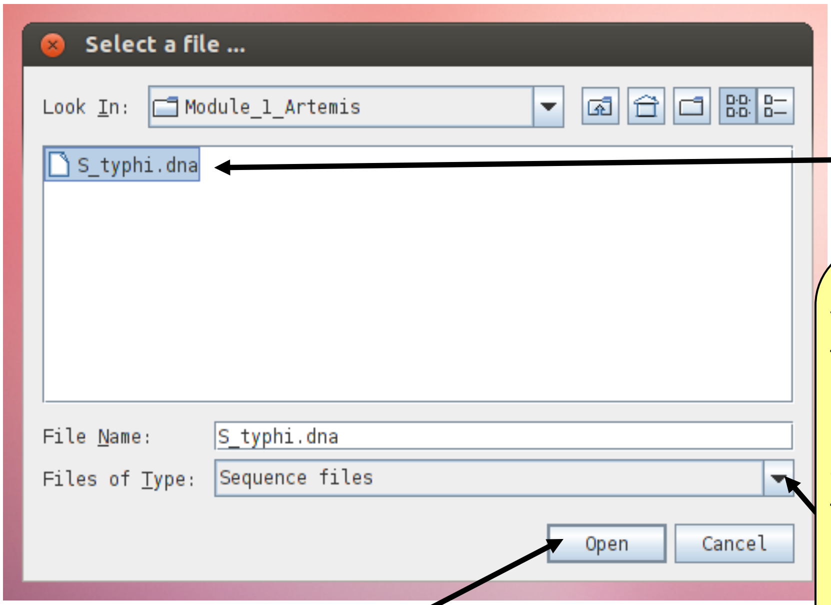


1 Click ‘File’
2 Then ‘Open’

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.



Look In: Module_1_Artemis

File Name: S_typhi.dna

Files of Type: Sequence files

3 Single click to select file S_typhi.dna

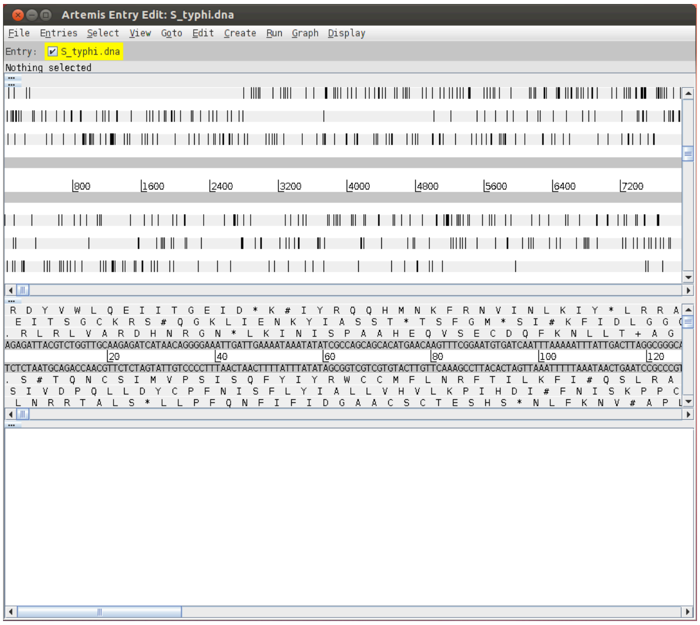
Change to ‘All Files’ if you want to display all the files in the directory.

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.

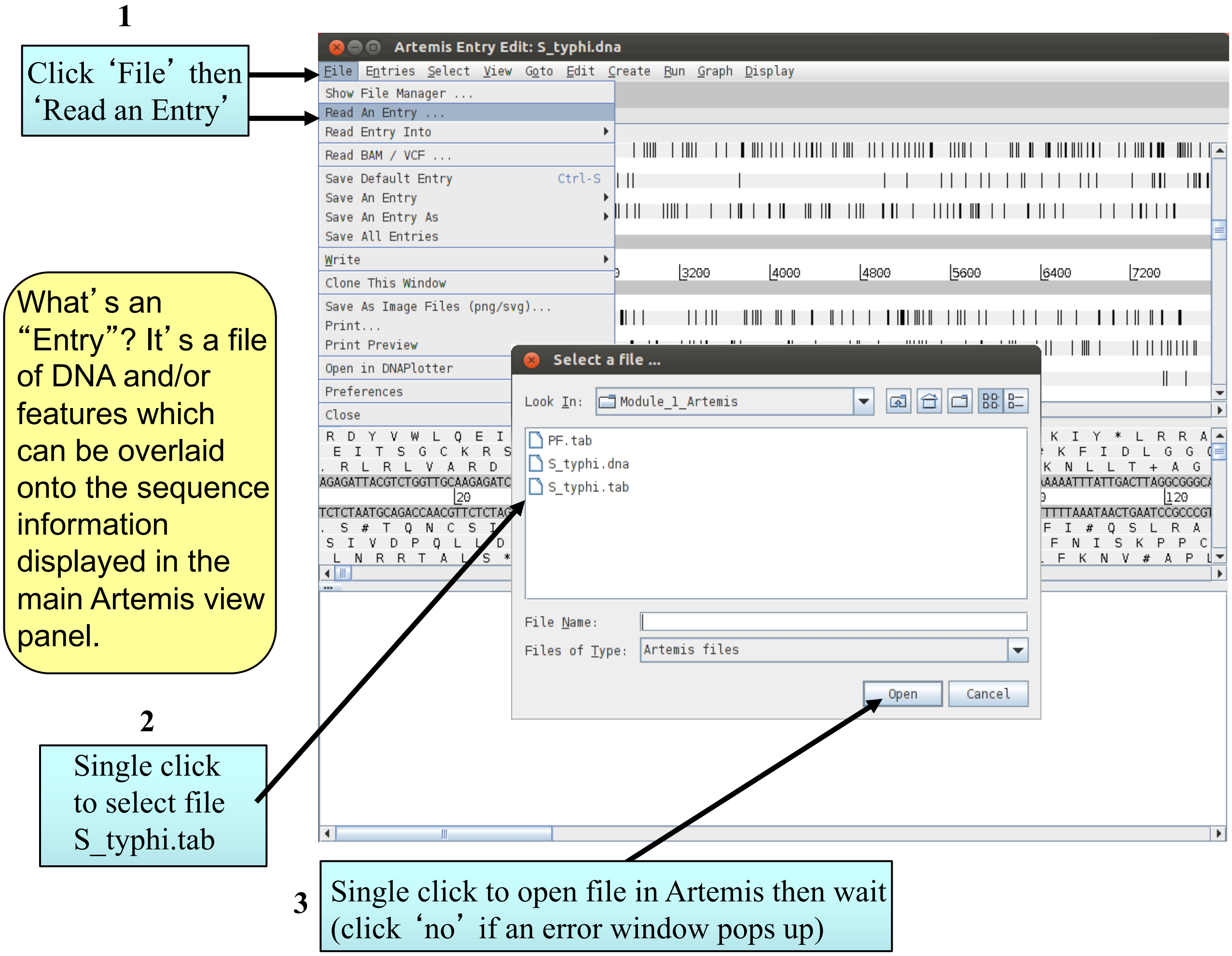
4 Single click to open file in Artemis then wait

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'

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

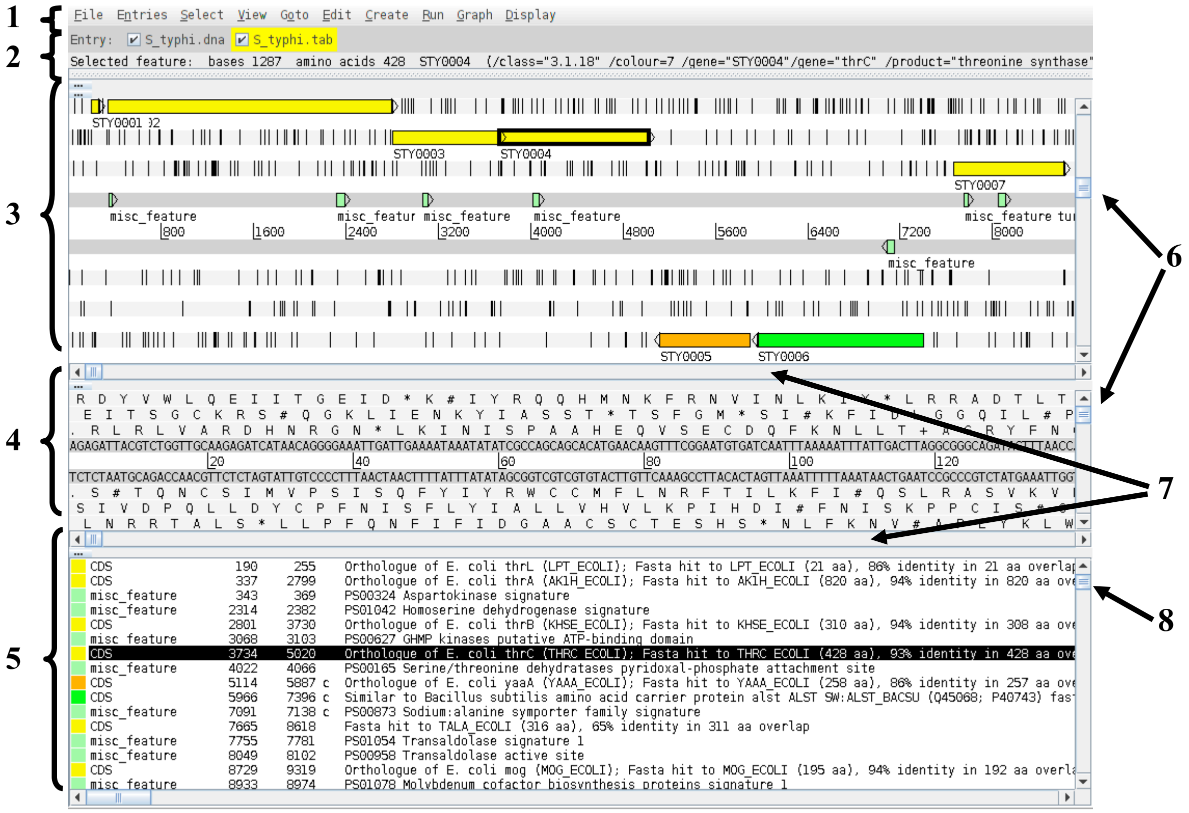
Single click to select file S_typhi.tab

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.



```
Selected feature:  bases 1287  amino acids 428  STY0004  (/class="3.1.18" /colour=7 /gene="STY0004"/gene="thrC" /product="threonine synthase"
```

```
 R  D  Y  V  W  L  Q  E  I  I  T  G  E  I  D  *  K  #  I  Y  R  Q  Q  H  M  N  K  F  R  N  V  I  N  L  K  I  Y  *  L  R  R  A  D  T  L  T
  E  I  T  S  G  C  K  R  S  #  Q  G  K  L  I  E  N  K  Y  I  A  S  S  T  *  T  S  F  G  M  *  S  I  #  K  F  I  D     G  G  Q  I  L  #  P
.  R  L  R  L  V  A  R  D  H  N  R  G  N  *  L  K  I  N  I  S  P  A  A  H  E  Q  V  S  E  C  D  Q  F  K  N  L  L  T  +  A     R  Y  F  N
AGAGATTACGTCTGGTTGCAAGAGATCATAACAGGGGAAATTGATTGAAAATAAATATATCGCCAGCAGCACATGAACAAGTTTCGGAATGTGATCAATTTAAAAATTTATTGACTTAGGCGGGCAGA   TTTAACC
                  |20                 |40                 |60                 |80                 |100                |120
TCTCTAATGCAGACCAACGTTCTCTAGTATTGTCCCCTTTAACTAACTTTTATTTATATAGCGGTCGTCGTGTACTTGTTCAAAGCCTTACACTAGTTAAATTTTTAAATAACTGAATCCGCCCGTCTATGAAATTGG
.  S  #  T  Q  N  C  S  I  M  V  P  S  I  S  Q  F  Y  I  Y  R  W  C  C  M  F  L  N  R  F  T  I  L  K  F  I  #  Q  S  L  R  A  S  V  K  V
 S  I  V  D  P  Q  L  L  D  Y  C  P  F  N  I  S  F  L  Y  I  A  L  L  V  H  V  L  K  P  I  H  D  I  #  F  N  I  S  K  P  P  C  I  S  #
  L  N  R  R  T  A  L  S  *  L  L  P  F  Q  N  F  I  F  I  D  G  A  A  C  S  C  T  E  S  H  S  *  N  L  F  K  N  V  #  A  P  L  Y  K  L  W
```

```
🟨 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 (AK1H_ECOLI); Fasta hit to AK1H_ECOLI (820 aa), 94% identity in 820 aa ove
🟩 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 ove
🟩 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 ove
🟩 misc_feature    4022   4066    PS00165 Serine/threonine dehydratases pyridoxal-phosphate attachment site
🟧 CDS             5114   5887 c  Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_ECOLI (258 aa), 86% identity in 257 aa ove
🟩 CDS             5966   7396 c  Similar to Bacillus subtilis amino acid carrier protein alst ALST SW:ALST_BACSU (Q45068; P40743) fas
🟩 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 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 overla
🟩 misc_feature    8933   8974    PS01078 Molybdenum cofactor biosynthesis proteins signature 1
```

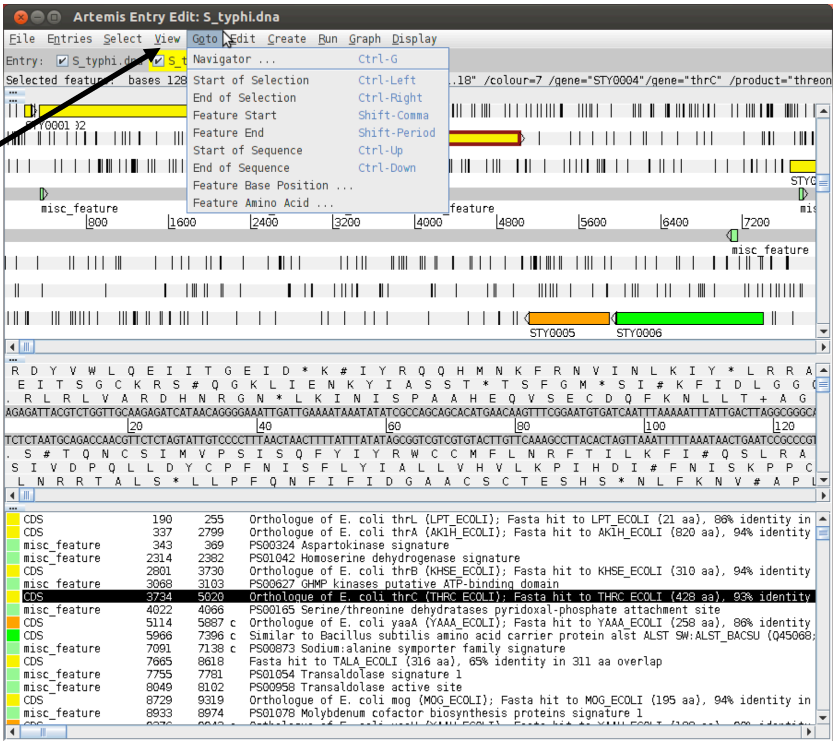
1. **Drop-down menus:** There are lots in there so don’t worry about all the details right now.
2. **Entry (top line):** shows which entries are currently loaded with the default entry highlighted in yellow (this is the entry into which newly created features are created). Selected feature: the details 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. Getting around in Artemis

There are three main ways of getting to a particular DNA region in Artemis:
-the Goto drop-down menu
-the Navigator and
-the Feature Selector (which we will use in Part IV)
The best method depends on what you' re trying to do. Knowing which one to use comes with practice.

4.1 The 'Goto' menu

The functions on this menu (below the Navigator option) are shortcuts for getting to locations within a selected feature or for jumping to the start or end of the DNA sequence. This is really intuitive so give it a try!


Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: S_typhi.dna
Navigator ... Ctrl-G
Start of Selection Ctrl-Left
End of Selection Ctrl-Right
Feature Start Shift-Comma
Feature End Shift-Period
Start of Sequence Ctrl-Up
End of Sequence Ctrl-Down
Feature Base Position ...
Feature Amino Acid ...
misc_feature
800
1600
2400
3200
4000
4800
5600
6400
7200
STY0005
STY0006
20
40
60
80
100
120

```
CDS            190    255     Orthologue of E. coli thrL (LPT_ECOLI); Fasta hit to LPT_ECOLI (21 aa), 86% identity in
CDS            337   2799     Orthologue of E. coli thrA (AK1H_ECOLI); Fasta hit to AK1H_ECOLI (820 aa), 94% identity
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
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
misc_feature  4022   4066     PS00165 Serine/threonine dehydratases pyridoxal-phosphate attachment site
CDS           5114   5887 c   Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_ECOLI (258 aa), 86% identity
CDS           5966   7396 c   Similar to Bacillus subtilis amino acid carrier protein alst ALST SW:ALST_BACSU (Q45068;
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 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
misc_feature  8933   8974     PS01078 Molybdenum cofactor biosynthesis proteins signature 1
```

Click 'Goto'

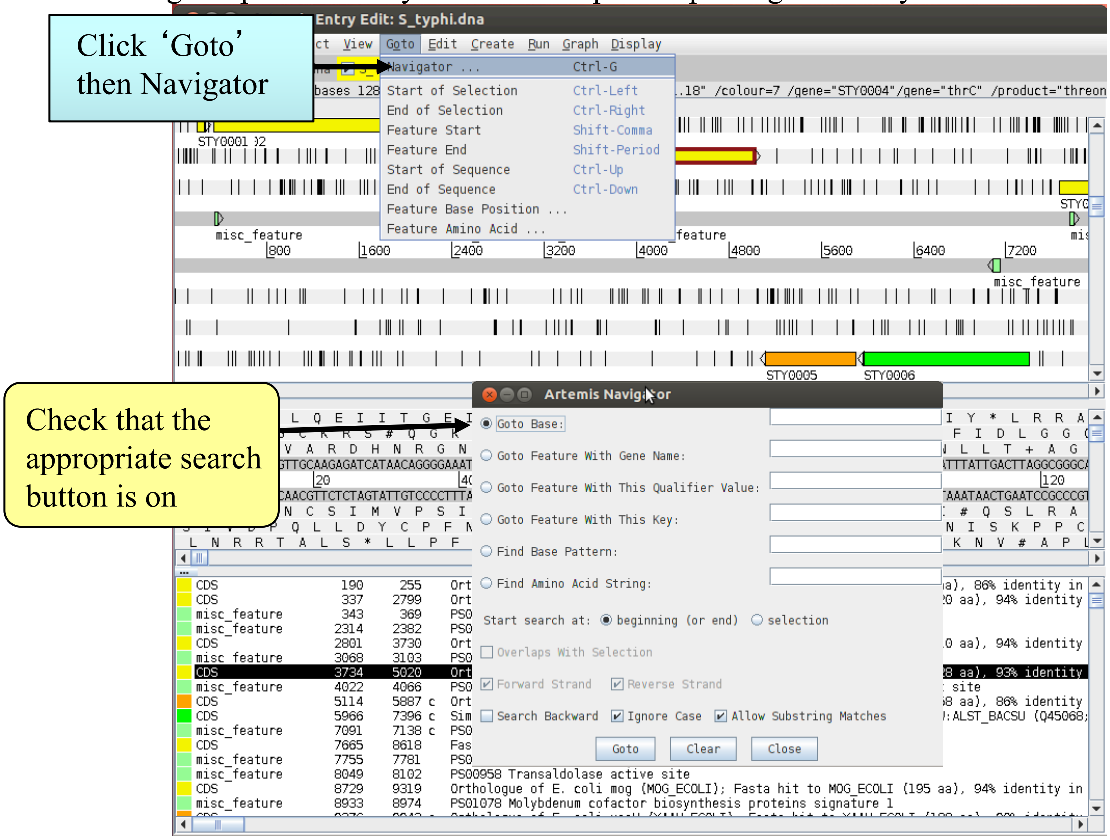
It may seem that 'Goto' 'Start of Selection' and 'Goto' 'Feature Start' do the same thing. Well they do if you have a feature selected but 'Goto' 'Start of Selection' will also work for a region which you have selected by click-dragging in the main window.
So yes, give it a try!

Suggested tasks:

1. Zoom out, select / highlight a large region of sequence by clicking the left hand button and dragging the cursor then go to the start and end of this selected region.
2. Select a CDS then go to the start and end.
3. Go to the start and end of the genome sequence.
4. Select a CDS. Within it, go to a base (nucleotide) and/or amino acid of your choice.
5. Highlight a region then, from the right click menu, select 'Zoom to Selection'.

4.2 Navigator

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


Entry Edit: S_typhi.dna
ct View Goto Edit Create Run Graph Display
Navigator ... Ctrl-G
Start of Selection Ctrl-Left
End of Selection Ctrl-Right
Feature Start Shift-Comma
Feature End Shift-Period
Start of Sequence Ctrl-Up
End of Sequence Ctrl-Down
Feature Base Position ...
Feature Amino Acid ...
bases 128
..18" /colour=7 /gene="STY0004"/gene="thrC" /product="threon
STY0001
misc_feature
800
1600
2400
3200
4000
4800
5600
6400
7200
feature
misc_feature
STY0005
STY0006
Artemis Navigator
Goto
Clear
Close

Click 'Goto' then Navigator

Check that the appropriate search button is on

● Goto Base:

○ Goto Feature With Gene Name:

○ Goto Feature With This Qualifier Value:

○ Goto Feature With This Key:

○ Find Base Pattern:

○ Find Amino Acid String:

Start search at: ● beginning (or end) ○ selection

□ Overlaps With Selection

☑ Forward Strand ☑ Reverse Strand

□ Search Backward ☑ Ignore Case ☑ Allow Substring Matches

```
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    PS00958 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
```

Suggestions about where to go:

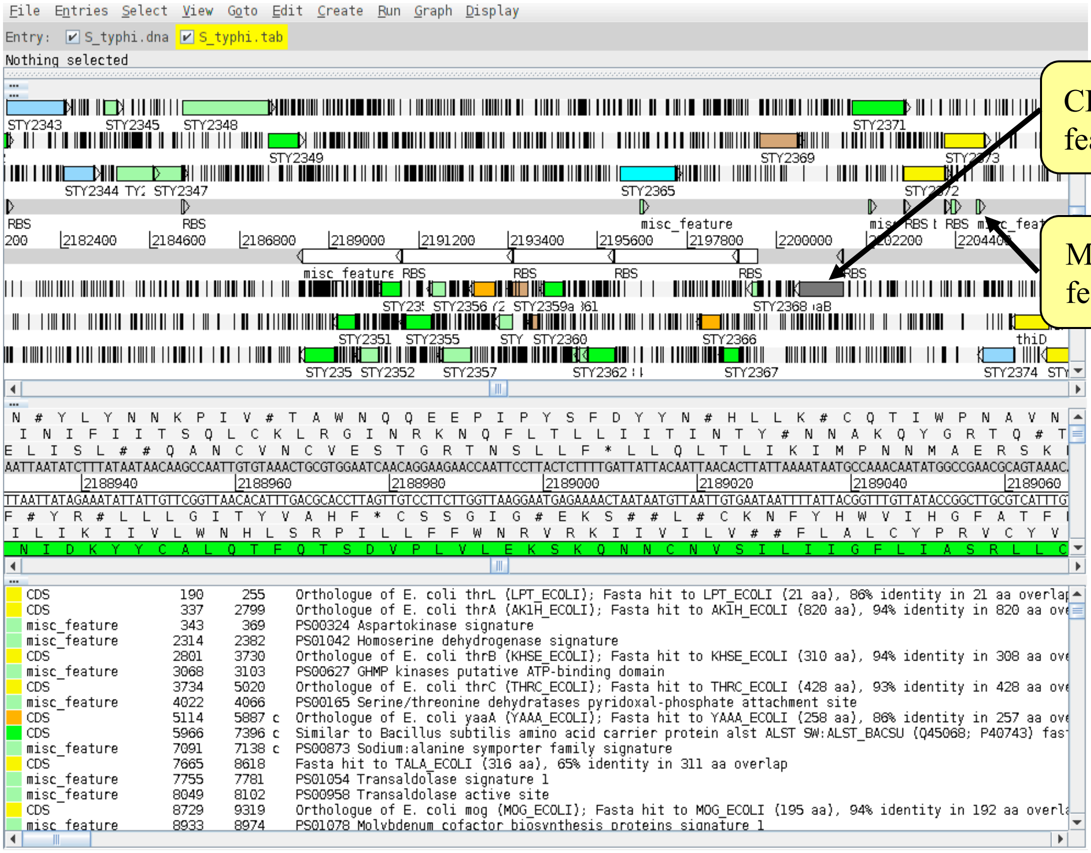
1. Think of a number between 1 and 4809037 and go to that base (notice how the cursors on the horizontal sliders move with you).
2. Your favourite gene name (it may not be there so you could try '*fts*').
3. Use '**Goto Feature With This Qualifier value**' to search the contents of all qualifiers for 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 Xa**).
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).



```
N  #  Y  L  Y  N  N  K  P  I  V  #  T  A  W  N  Q  Q  E  E  P  I  P  Y  S  F  D  Y  Y  N  #  H  L  L  K  #  C  Q  T  I  W  P  N  A  V  N
 I  N  I  F  I  I  T  S  Q  L  C  K  L  R  G  I  N  R  K  N  Q  F  L  T  L  L  I  I  T  I  N  T  Y  #  N  N  A  K  Q  Y  G  R  T  Q  #  T
E  L  I  S  L  #  #  Q  A  N  C  V  N  C  V  E  S  T  G  R  T  N  S  L  L  F  *  L  L  Q  L  T  L  I  K  I  M  P  N  N  M  A  E  R  S  K
AATTAATATCTTTATAATAACAAGCCAATTGTGTAAACTGCGTGGAATCAACAGGAAGAACCAATTCCTTACTCTTTTGATTATTACAATTAACACTTATTAAAATAATGCCAAACAATATGGCCGAACGCAGTAAAC
       2188940          2188960          2188980          2189000          2189020          2189040          2189060
TTAATTATAGAAATATTATTGTTCGGTTAACACATTTGACGCACCTTAGTTGTCCTTCTTGGTTAAGGAATGAGAAAACTAATAATGTTAATTGTGAATAATTTTATTACGGTTTGTTATACCGGCTTGCGTCATTTG
F  #  Y  R  #  L  L  L  G  I  T  Y  V  A  H  F  *  C  S  S  G  I  G  #  E  K  S  #  #  L  #  C  K  N  F  Y  H  W  V  I  H  G  F  A  T  F
 I  L  I  K  I  I  V  L  W  N  H  L  S  R  P  I  L  L  F  F  W  N  R  V  R  K  I  I  V  I  L  V  #  #  F  L  A  L  C  Y  P  R  V  C  Y  V
  N  I  D  K  Y  Y  C  A  L  Q  T  F  Q  T  S  D  V  P  L  V  L  E  K  S  K  Q  N  N  C  N  V  S  I  L  I  I  G  F  L  I  A  S  R  L  L  C
```

```
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 (AK1H_ECOLI); Fasta hit to AK1H_ECOLI (820 aa), 94% identity in 820 aa ove
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 ove
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 ove
misc_feature   4022   4066    PS00165 Serine/threonine dehydratases pyridoxal-phosphate attachment site
CDS            5114   5887 c  Orthologue of E. coli yaaA (YAAA_ECOLI); Fasta hit to YAAA_ECOLI (258 aa), 86% identity in 257 aa ove
CDS            5966   7396 c  Similar to Bacillus subtilis amino acid carrier protein alst ALST SW:ALST_BACSU (Q45068; P40743) fas
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 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 overla
misc_feature   8933   8974    PS01078 Molybdenum cofactor biosynthesis proteins signature 1
```

CDS features

Misc features

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

Information to view:

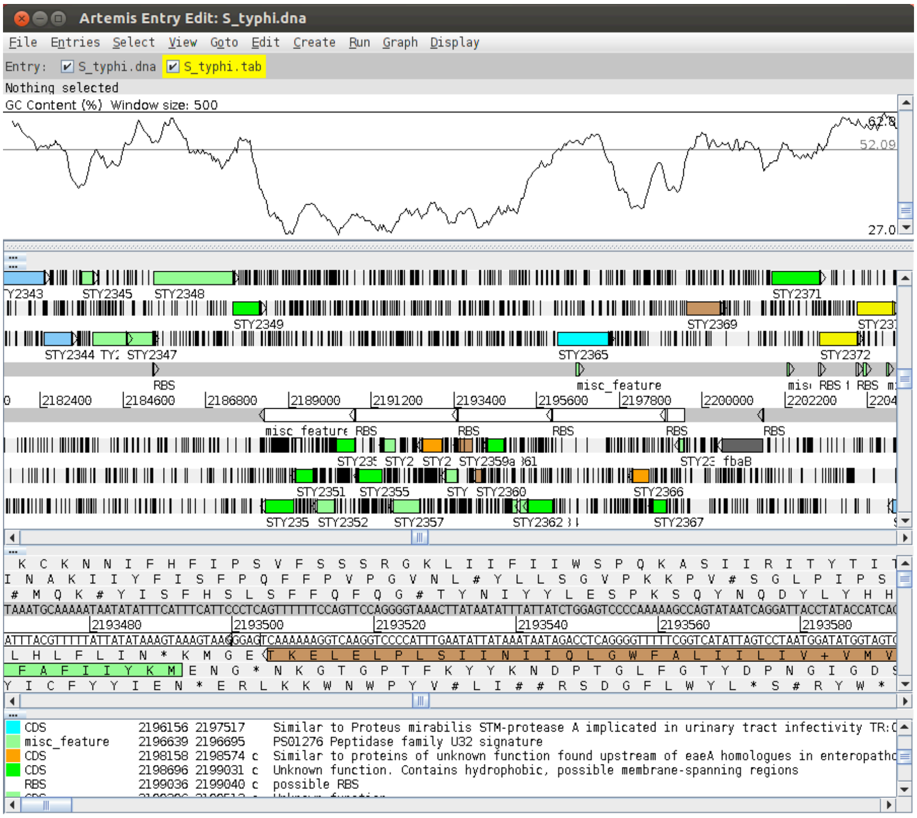
Annotation
If you click on a particular feature you can view the annotation associated with it: select a CDS feature (or any other feature) and click on the ‘Edit’ menu and select ‘Selected Feature in Editor’. A window will appear containing all the annotation that is associated with that CDS. The format 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).

To view the graphs:
Click on the 'Graph' menu to see all those available and then tick the box for 'GC Content (%)'. To adjust the smoothing of the graph you change the window size over which the points on the graph are calculated, using the slider shown below.


Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ S_typhi.dna ☑ S_typhi.tab
Nothing selected
GC Content (%) Window size: 500
62.8
52.09
27.0
Y2343 STY2345 STY2348 STY2371
STY2349 STY2369 STY23
STY2344 TY: STY2347 STY2365 STY2372
RBS misc_feature misc RBS RBS m
0 2182400 2184600 2186800 2189000 2191200 2193400 2195600 2197800 2200000 2202200 2204
misc_feature RBS RBS RBS RBS RBS
STY235 STY2 STY2 STY2359a 361 STY23 fbaB
STY2351 STY2355 STY STY2360 STY2366
STY235 STY2352 STY2357 STY2362 STY2367

```
 K C K N N I F H F I P S V F S S S R G K L I I F I I W S P Q K A S I I R I T Y T I T
I N A K I I Y F I S F P Q F F P V P G V N L # Y L L S G V P K K P V # S G L P I P S
 # M Q K # Y I S F H S L S F F Q F Q G # T Y N I Y Y L E S P K S Q Y N Q D Y L Y H H
TAAATGCAAAAATAATATATTTCATTTCATTCCCTCAGTTTTTTCCAGTTCCAGGGGTAAACTTATAATATTTATTATCTGGAGTCCCCAAAAAGCCAGTATAATCAGGATTACCTATACCATCAC
          2193480          2193500          2193520          2193540          2193560          2193580
ATTTACGTTTTTATTATATAAAGTAAAGTAAGGGAGTCAAAAAAGGTCAAGGTCCCCATTTGAATATTATAAATAATAGACCTCAGGGGTTTTTCGGTCATATTAGTCCTAATGGATATGGTAGTC
 L H L F L I N * K M G E T K E L E L P L S I I N I I Q L G W F A L I I L I V + V M V
 F A F I I Y K M E N G * N K G T G P T F K Y Y K N D P T G L F G T Y D P N G I G D S
Y I C F Y Y I E N * E R L K K W N W P Y V # L I # # R S D G F L W Y L * S # R Y W *
```

```
CDS            2196156 2197517    Similar to Proteus mirabilis STM-protease A implicated in urinary tract infectivity TR:C
misc_feature   2196639 2196695    PS01276 Peptidase family U32 signature
CDS            2198158 2198574 c  Similar to proteins of unknown function found upstream of eaeA homologues in enteropatho
CDS            2198696 2199031 c  Unknown function. Contains hydrophobic, possible membrane-spanning regions
RBS            2199036 2199040 c  possible RBS
```

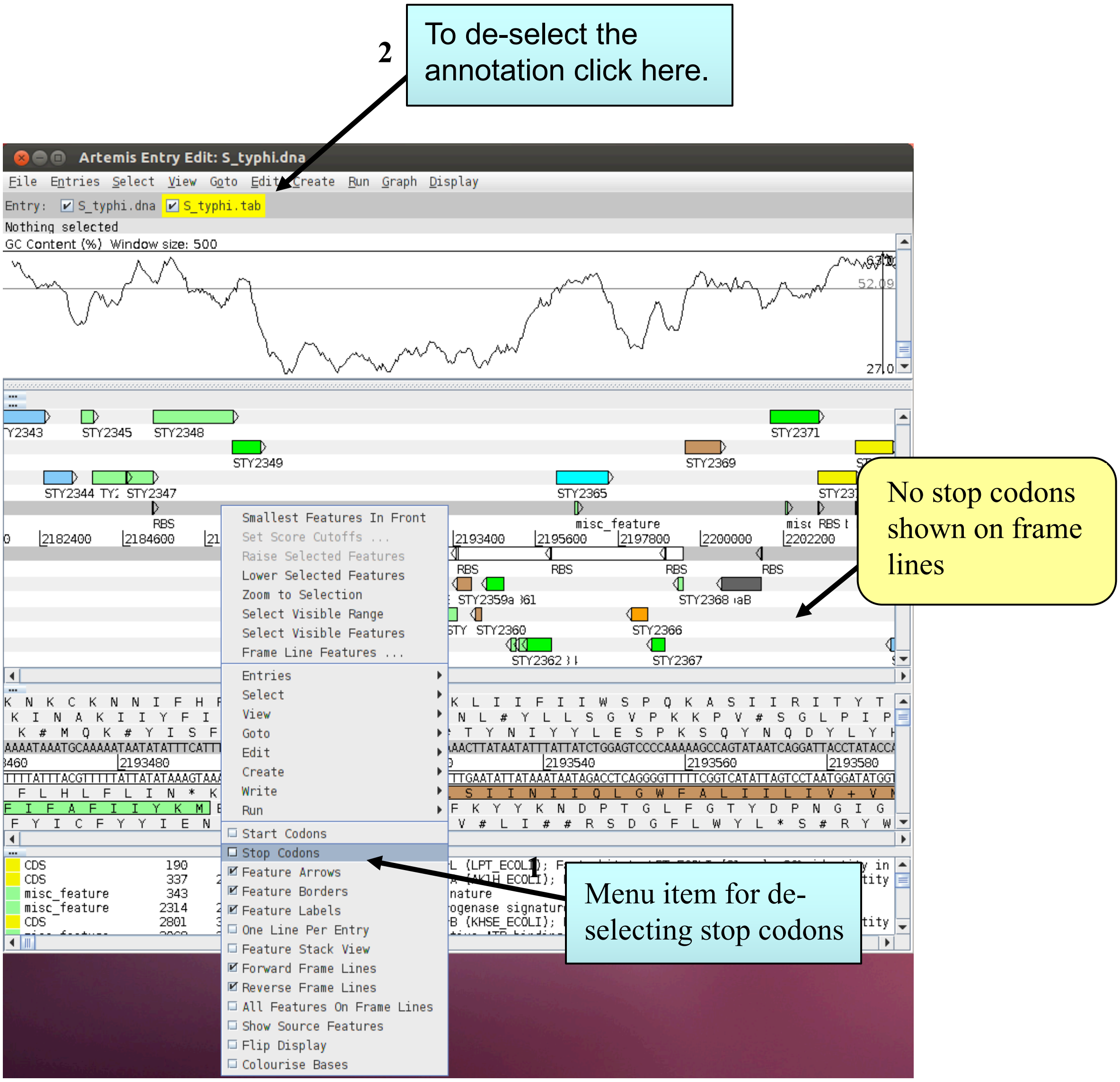
DNA plot

Slider for smoothing

Notice how the plot show a marked deviation around the region you are currently looking at. To fully appreciate how anomalous this region is move the genome view by scrolling to the left and right of this region. The apparent unusual nucleotide content of this region is indicative of laterally acquired DNA that has inserted into the genome.

As well as looking at the characteristics of small regions of the genome, it is possible to zoom out and look at the characteristics of the genome as a whole. To view the entire genome you can use the sliders indicated above. However, be careful zooming out quickly with all the features being displayed, as this may temporarily lock up the computer. Read further so see how to zoom out.

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.



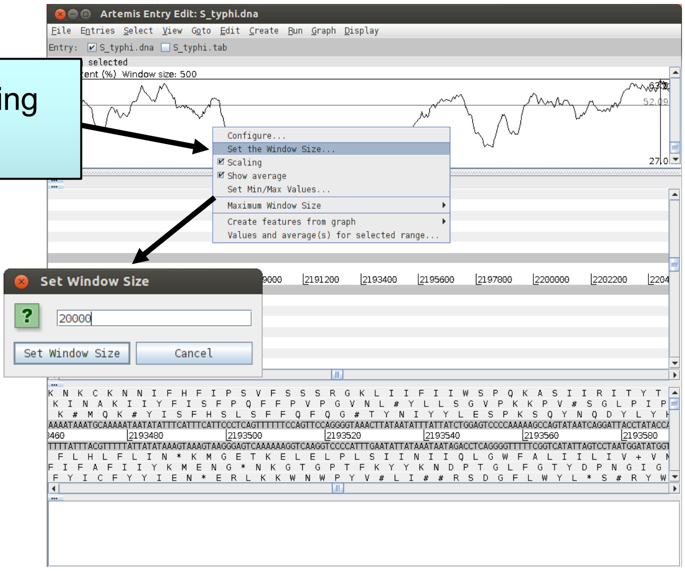
To de-select the annotation click here.

No stop codons shown on frame lines

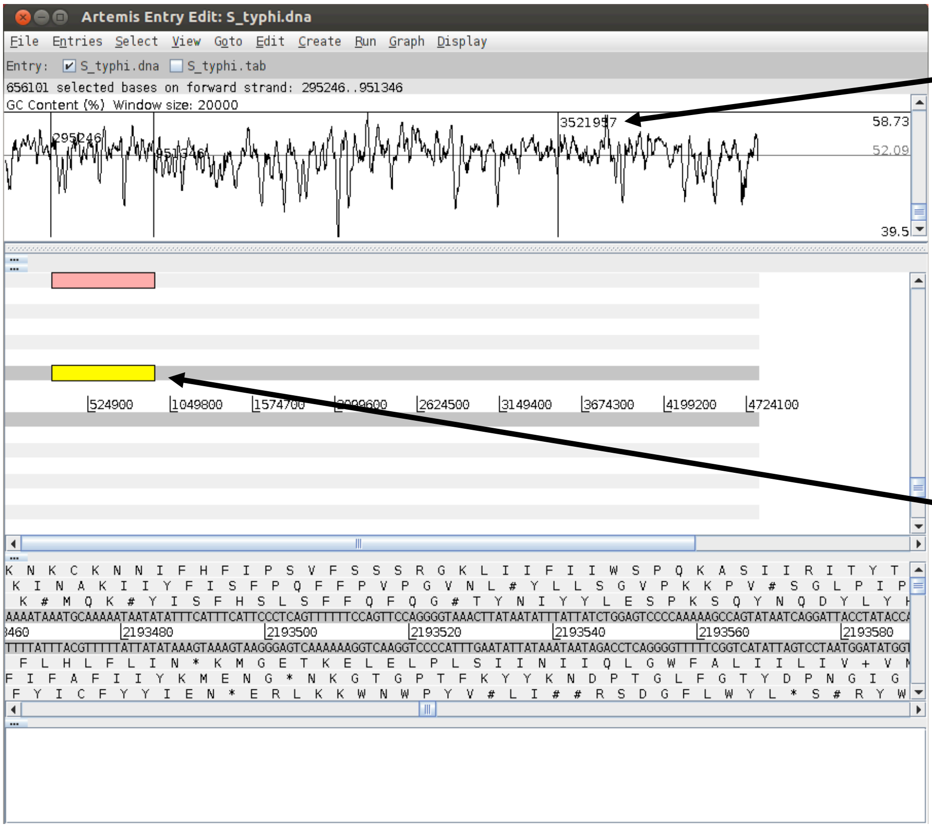
Menu item for de-selecting stop codons

3. 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).
4. 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.

3 Graph scaling menu


Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ S_typhi.dna □ S_typhi.tab
Configure...
Set the Window Size...
☑ Scaling
☑ Show average
Set Min/Max Values...
Maximum Window Size
Create features from graph
Values and average(s) for selected range...
Set Window Size
20000
Set Window Size
Cancel

Slider for zooming out 4


Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ S_typhi.dna □ S_typhi.tab
656101 selected bases on forward strand: 295246..951346
GC Content (%) Window size: 20000
3521957
295246
951346
58.73
52.09
39.5
524900
1049800
1574700
2099600
2624500
3149400
3674300
4199200
4724100

Click with the left mouse button in a graph window. A line and a number will appear. The number is the relative position within the genome (bps).

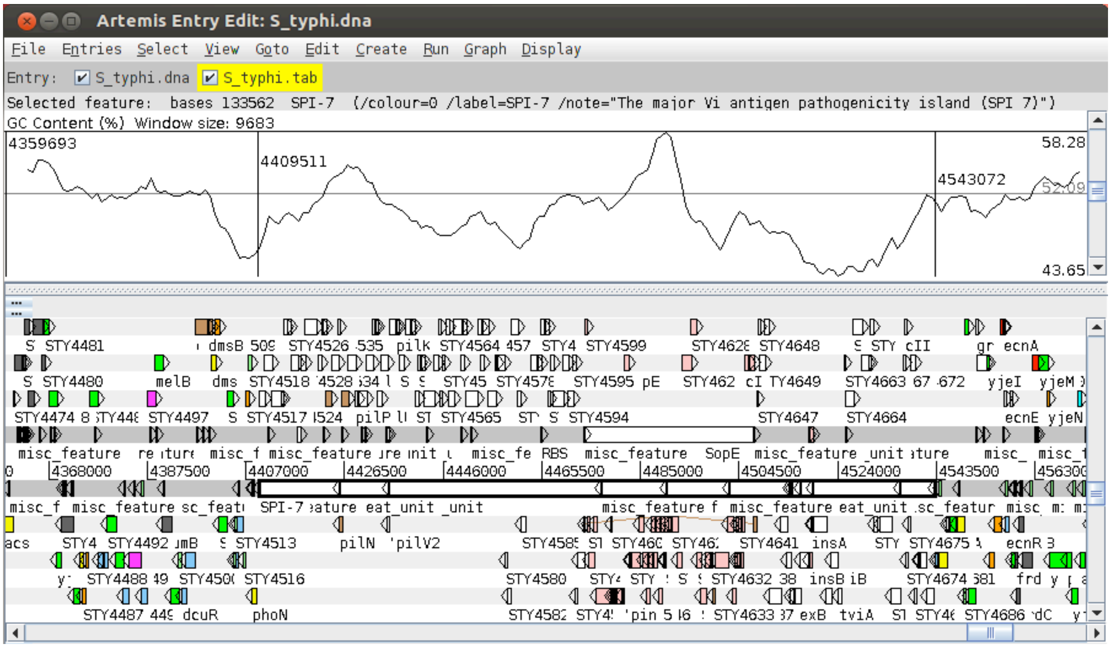
Click and drag to highlight a region on the main DNA line. Notice that the boundaries of this region are now marked in the graph windows that you previously clicked in.

Artemis Exercise 3

Now select the ‘S.typhi.tab’ entry box to switch on the annotation and 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.



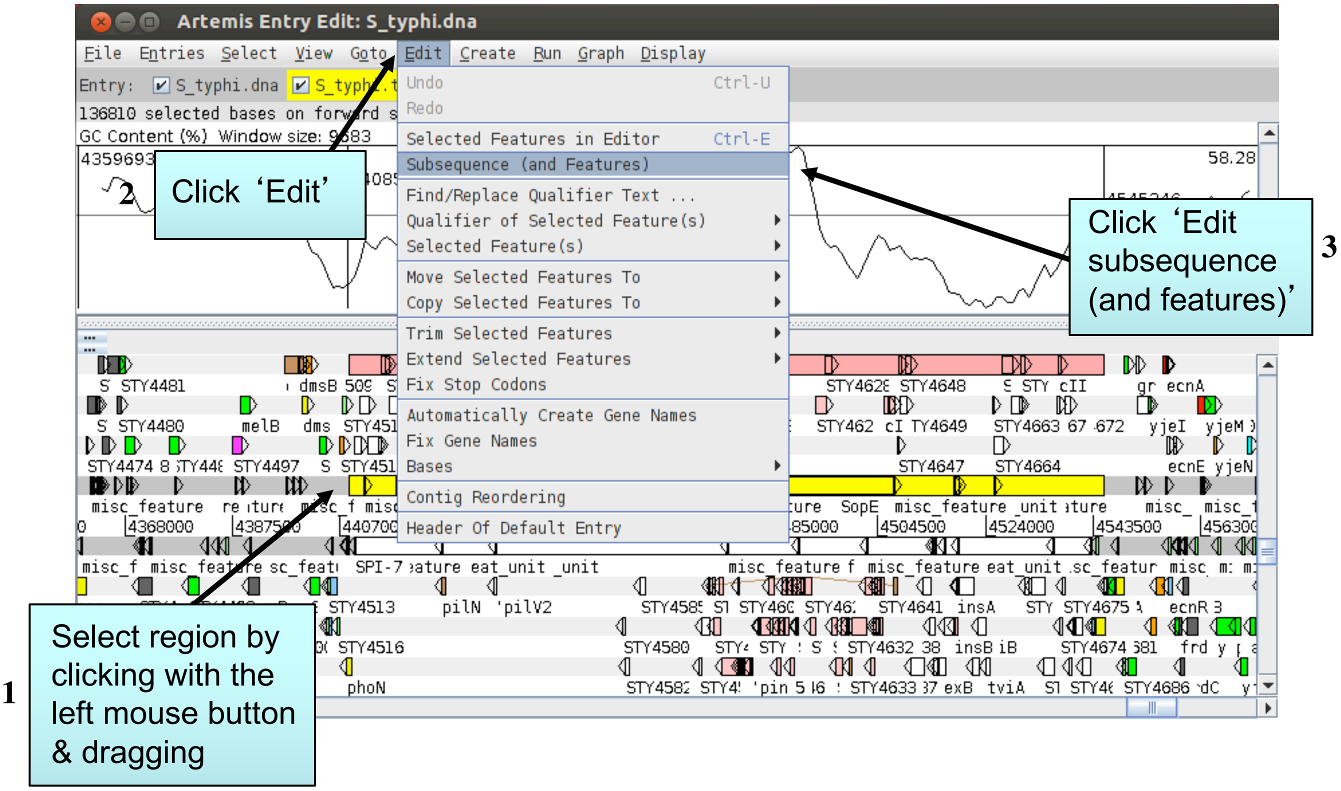
Artemis Entry Edit: S_typhi.dna

Entry: ☑ S_typhi.dna ☑ S_typhi.tab

Selected feature: bases 133562 SPI-7 (/colour=0 /label=SPI-7 /note="The major Vi antigen pathogenicity island (SPI 7)")

GC Content (%) Window size: 9683

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.

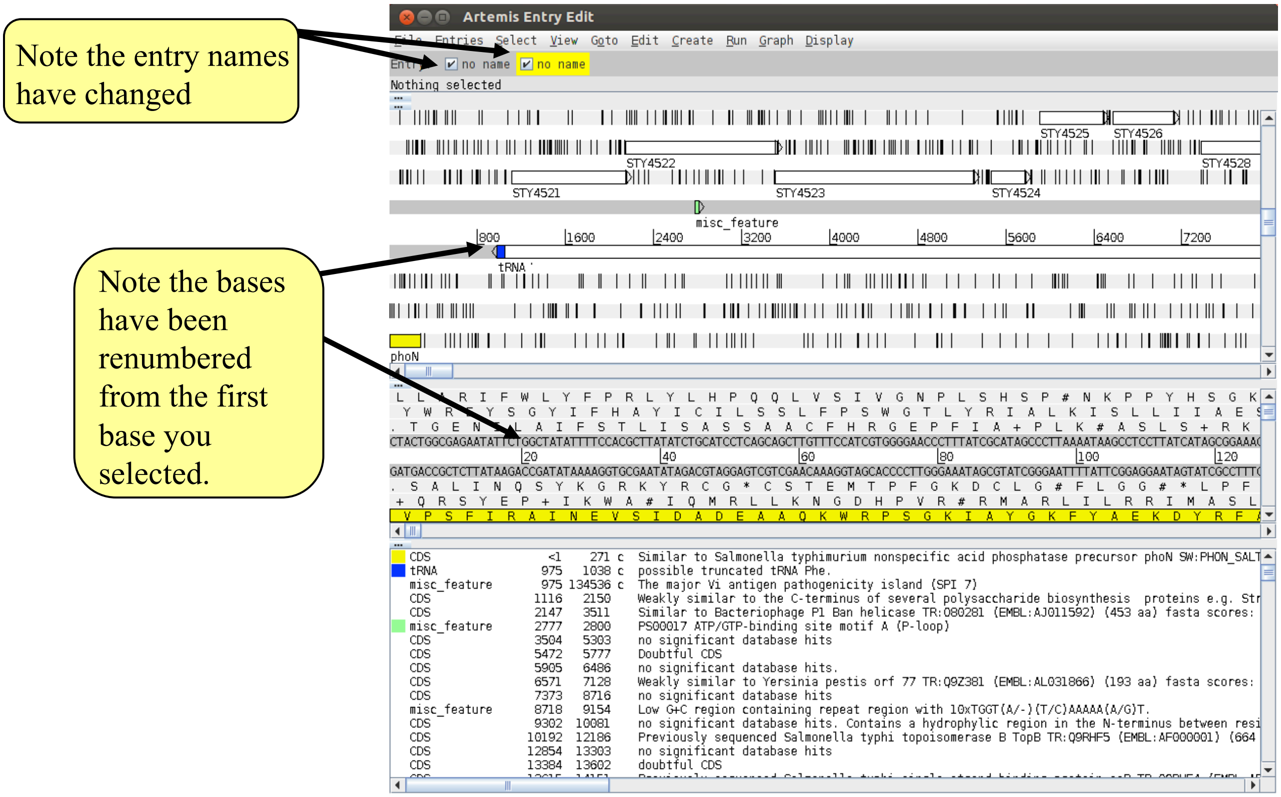

Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ S_typhi.dna ☑ S_typh
136810 selected bases on forward s
GC Content (%) Window size: 9683
Undo Ctrl-U
Redo
Selected Features in Editor Ctrl-E
Subsequence (and Features)
Find/Replace Qualifier Text ...
Qualifier of Selected Feature(s)
Selected Feature(s)
Move Selected Features To
Copy Selected Features To
Trim Selected Features
Extend Selected Features
Fix Stop Codons
Automatically Create Gene Names
Fix Gene Names
Bases
Contig Reordering
Header Of Default Entry
58.28

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


Artemis Entry Edit
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ no name ☑ no name
Nothing selected
STY4525 STY4526 STY4522 STY4528 STY4521 STY4523 STY4524
misc_feature
800 1600 2400 3200 4000 4800 5600 6400 7200
tRNA
phoN

Note the entry names have changed

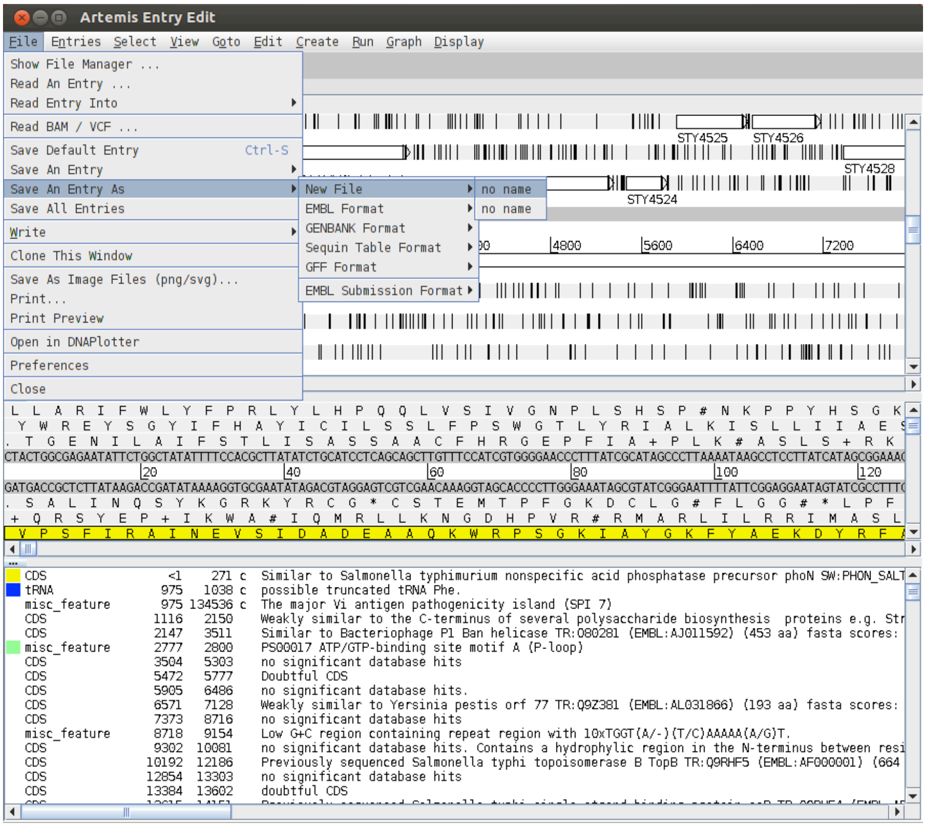
Note the bases have been renumbered from the first base you selected.

```
CDS                <1     271 c  Similar to Salmonella typhimurium nonspecific acid phosphatase precursor phoN SW:PHON_SALT
tRNA              975    1038 c  possible truncated tRNA Phe.
misc_feature      975  134536 c  The major Vi antigen pathogenicity island (SPI 7)
CDS              1116    2150    Weakly similar to the C-terminus of several polysaccharide biosynthesis  proteins e.g. Str
CDS              2147    3511    Similar to Bacteriophage P1 Ban helicase TR:O80281 (EMBL:AJ011592) (453 aa) fasta scores:
misc_feature     2777    2800    PS00017 ATP/GTP-binding site motif A (P-loop)
CDS              3504    5303    no significant database hits
CDS              5472    5777    Doubtful CDS
CDS              5905    6486    no significant database hits.
CDS              6571    7128    Weakly similar to Yersinia pestis orf 77 TR:Q9Z381 (EMBL:AL031866) (193 aa) fasta scores:
CDS              7373    8716    no significant database hits
misc_feature     8718    9154    Low G+C region containing repeat region with 10xTGGT(A/-)(T/C)AAAAA(A/G)T.
CDS              9302   10081    no significant database hits. Contains a hydrophylic region in the N-terminus between resi
CDS             10192   12186    Previously sequenced Salmonella typhi topoisomerase B TopB TR:Q9RHF5 (EMBL:AF000001) (664
CDS             12854   13303    no significant database hits
CDS             13384   13602    doubtful CDS
```

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

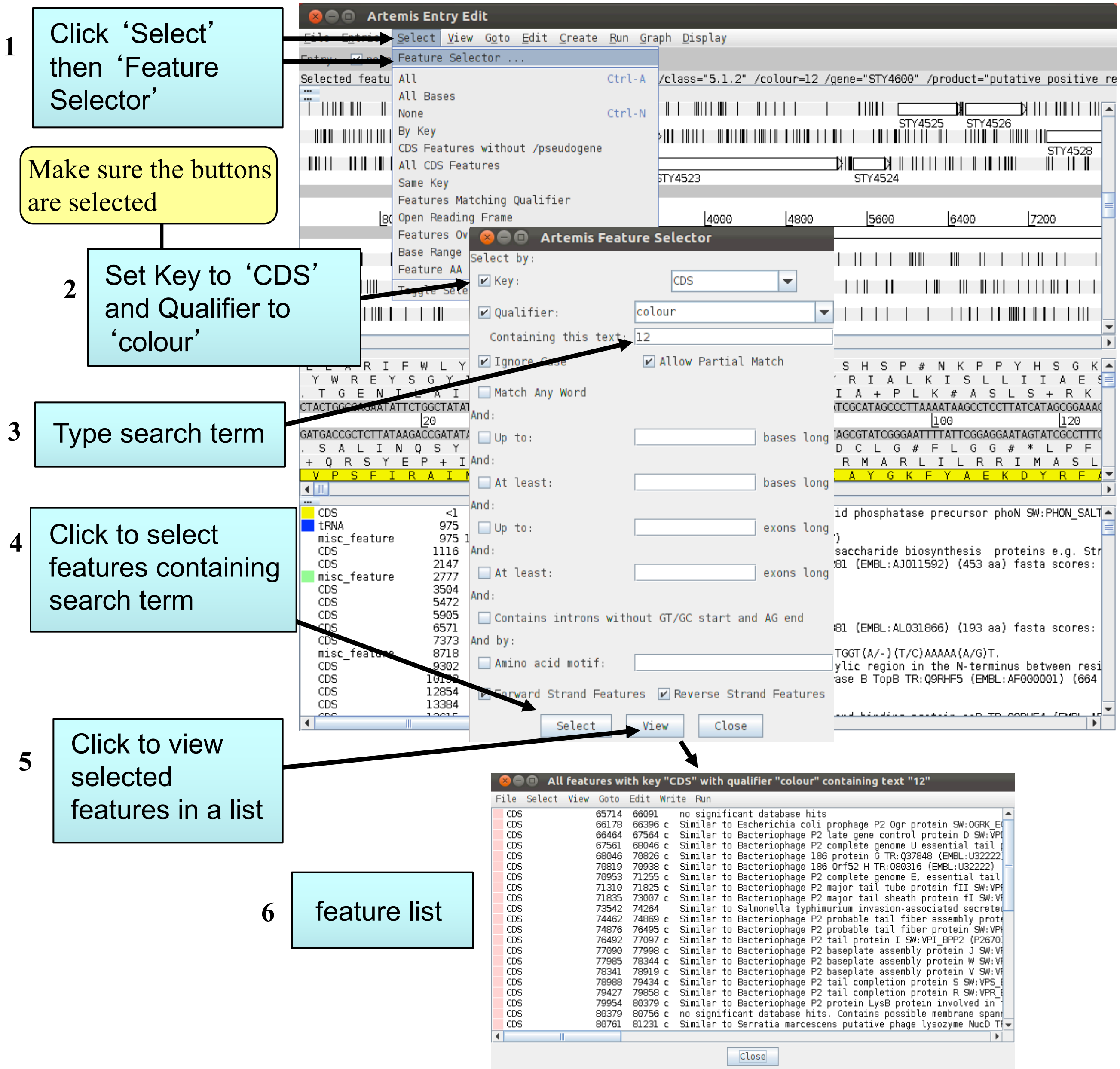


```
 L  L  A  R  I  F  W  L  Y  F  P  R  L  Y  L  H  P  Q  Q  L  V  S  I  V  G  N  P  L  S  H  S  P  #  N  K  P  P  Y  H  S  G  K
  Y  W  R  E  Y  S  G  Y  I  F  H  A  Y  I  C  I  L  S  S  L  F  P  S  W  G  T  L  Y  R  I  A  L  K  I  S  L  L  I  I  A  E
. T  G  E  N  I  L  A  I  F  S  T  L  I  S  A  S  S  A  A  C  F  H  R  G  E  P  F  I  A  +  P  L  K  #  A  S  L  S  +  R  K
CTACTGGCGAGAATATTCTGGCTATATTTTCCACGCTTATATCTGCATCCTCAGCAGCTTGTTTCCATCGTGGGGAACCCTTTATCGCATAGCCCTTAAAATAAGCCTCCTTATCATAGCGGAAAC
GATGACCGCTCTTATAAGACCGATATAAAAGGTGCGAATATAGACGTAGGAGTCGTCGAACAAAGGTAGCACCCCTTGGGAAATAGCGTATCGGGAATTTTATTCGGAGGAATAGTATCGCCTTTG
. S  A  L  I  N  Q  S  Y  K  G  R  K  Y  R  C  G  *  C  S  T  E  M  T  P  F  G  K  D  C  L  G  #  F  L  G  G  #  *  L  P  F
+ Q  R  S  Y  E  P  +  I  K  W  A  #  I  Q  M  R  L  L  K  N  G  D  H  P  V  R  #  R  M  A  R  L  I  L  R  R  I  M  A  S  L
 V  P  S  F  I  R  A  I  N  E  V  S  I  D  A  D  E  A  A  Q  K  W  R  P  S  G  K  I  A  Y  G  K  F  Y  A  E  K  D  Y  R  F
```

```
CDS                <1    271 c  Similar to Salmonella typhimurium nonspecific acid phosphatase precursor phoN SW:PHON_SALT
tRNA              975   1038 c  possible truncated tRNA Phe.
misc_feature      975 134536 c  The major Vi antigen pathogenicity island (SPI 7)
CDS              1116   2150    Weakly similar to the C-terminus of several polysaccharide biosynthesis  proteins e.g. Str
CDS              2147   3511    Similar to Bacteriophage P1 Ban helicase TR:O80281 (EMBL:AJ011592) (453 aa) fasta scores:
misc_feature     2777   2800    PS00017 ATP/GTP-binding site motif A (P-loop)
CDS              3504   5303    no significant database hits
CDS              5472   5777    Doubtful CDS
CDS              5905   6486    no significant database hits.
CDS              6571   7128    Weakly similar to Yersinia pestis orf 77 TR:Q9Z381 (EMBL:AL031866) (193 aa) fasta scores:
CDS              7373   8716    no significant database hits
misc_feature     8718   9154    Low G+C region containing repeat region with 10xTGGT(A/-)(T/C)AAAAA(A/G)T.
CDS              9302  10081    no significant database hits. Contains a hydrophylic region in the N-terminus between resi
CDS             10192  12186    Previously sequenced Salmonella typhi topoisomerase B TopB TR:Q9RHF5 (EMBL:AF000001) (664
CDS             12854  13303    no significant database hits
CDS             13384  13602    doubtful CDS
```

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

Artemis Feature Selector

Select by:
☑ Key: CDS
☑ Qualifier: colour
Containing this text: 12
☑ Ignore Case ☑ Allow Partial Match
☐ Match Any Word
☑ Forward Strand Features ☑ Reverse Strand Features

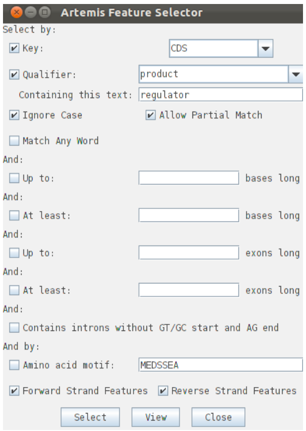
All features with key "CDS" with qualifier "colour" containing text "12"

```
CDS   65714  66091    no significant database hits
CDS   66178  66396 c  Similar to Escherichia coli prophage P2 Ogr protein SW:OGRK_E
CDS   66464  67564 c  Similar to Bacteriophage P2 late gene control protein D SW:VPI
CDS   67561  68046 c  Similar to Bacteriophage P2 complete genome U essential tail 
CDS   68046  70826 c  Similar to Bacteriophage 186 protein G TR:Q37848 (EMBL:U32222)
CDS   70819  70938 c  Similar to Bacteriophage 186 Orf52 H TR:O80316 (EMBL:U32222)
CDS   70953  71255 c  Similar to Bacteriophage P2 complete genome E, essential tail
CDS   71310  71825 c  Similar to Bacteriophage P2 major tail tube protein fII SW:VPF
CDS   71835  73007 c  Similar to Bacteriophage P2 major tail sheath protein fI SW:VF
CDS   73542  74264    Similar to Salmonella typhimurium invasion-associated secreted
CDS   74462  74869 c  Similar to Bacteriophage P2 probable tail fiber assembly prote
CDS   74876  76495 c  Similar to Bacteriophage P2 probable tail fiber protein SW:VPH
CDS   76492  77097 c  Similar to Bacteriophage P2 tail protein I SW:VPI_BPP2 (P2670
CDS   77090  77998 c  Similar to Bacteriophage P2 baseplate assembly protein J SW:VF
CDS   77985  78344 c  Similar to Bacteriophage P2 baseplate assembly protein W SW:VF
CDS   78341  78919 c  Similar to Bacteriophage P2 baseplate assembly protein V SW:VF
CDS   78988  79434 c  Similar to Bacteriophage P2 tail completion protein S SW:VPS_E
CDS   79427  79858 c  Similar to Bacteriophage P2 tail completion protein R SW:VPR_E
CDS   79954  80379 c  Similar to Bacteriophage P2 protein LysB protein involved in 
CDS   80379  80756 c  no significant database hits. Contains possible membrane spann
CDS   80761  81231 c  Similar to Serratia marcescens putative phage lysozyme NucD TF
```

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 (if you can’t remember how to do it have a look at page 14). 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.

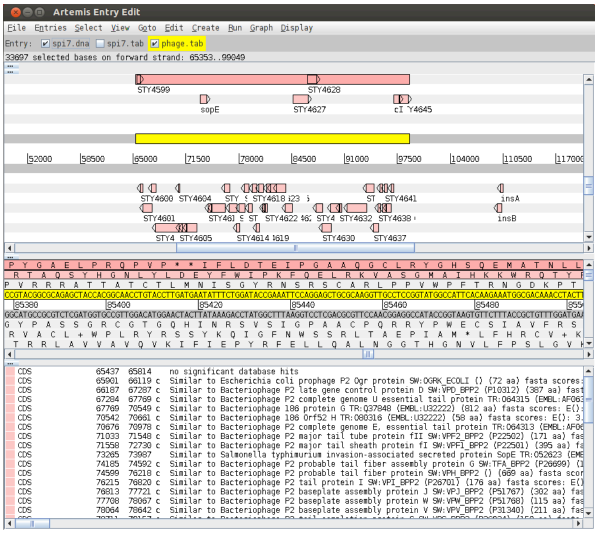
Additional methods for selecting/extracting features using the Feature Selector
It is worth noting that the Feature Selector can be used in many other ways to select and extract subsets of features from the genome, using eg text or amino acid searches.



☑ Key: CDS
☑ Qualifier: product
Containing this text: regulator
☑ Ignore Case ☑ Allow Partial Match
□ Match Any Word
□ Up to: bases long
□ At least: bases long
□ Up to: exons long
□ At least: exons long
□ Contains introns without GT/GC start and AG end
□ Amino acid motif: MEDSSEA
☑ Forward Strand Features ☑ Reverse Strand Features

Space for a search term or amino acid motif

Defining the extent of the prophage
Even from this preliminary analysis it is clear that the prophage occupies a fairly discrete region within SPI-7 (see below). It is often useful to create a new DNA feature to define the limits of this type of genome landmark. To do this switch off stop codons, then use the left mouse button to click and drag over the region that you think defines the prophage.


Artemis Entry Edit
File Entries Select View Goto Edit Create Run Graph Display
STY4599
STY4628
sopE
STY4627
cI Y4645
52000
58500
65000
71500
78000
84500
91000
97500
104000
110500
117000
STY4600 STY4604 STY STY4618 ST STY4641 insA
STY4601 STY461 ST STY4622 STY4 STY4632 STY4638 insB
STY4 STY4605 STY4614 STY4630 STY4637
85380
85400
85420
85440
85460
85480

Entry: ☑ spi7.dna □ spi7.tab ☑ phage.tab

```
33697 selected bases on forward strand: 65353..99049
```

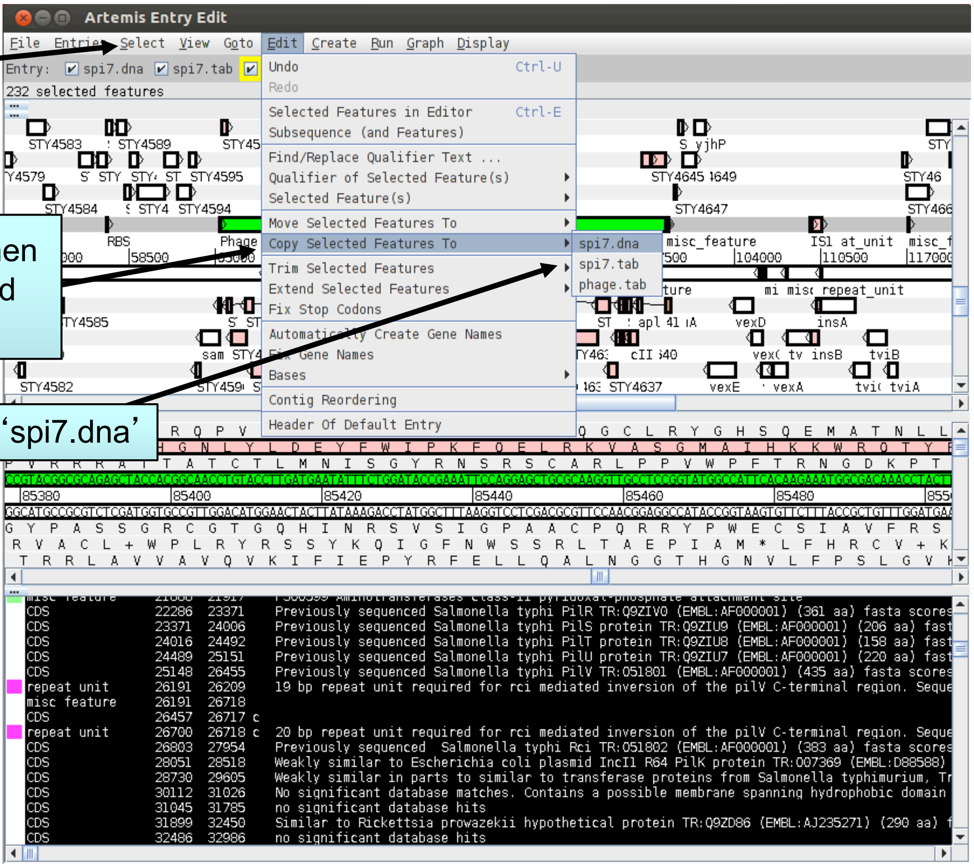
```
P  Y  G  A  E  L  P  R  Q  P  V  P  *  *  I  F  L  D  T  E  I  P  G  A  A  Q  G  C  L  R  Y  G  H  S  Q  E  M  A  T  N  L  L
 R  T  A  Q  S  Y  H  G  N  L  Y  L  D  E  Y  F  W  I  P  K  F  Q  E  L  R  K  V  A  S  G  M  A  I  H  K  K  W  R  Q  T  Y
P  V  R  R  R  A  T  T  A  T  C  T  L  M  N  I  S  G  Y  R  N  S  R  S  C  A  R  L  P  P  V  W  P  F  T  R  N  G  D  K  P  T
CCGTACGGCGCAGAGCTACCACGGCAACCTGTACCTTGATGAATATTTCTGGATACCGAAATTCCAGGAGCTGCGCAAGGTTGCCTCCGGTATGGCCATTCACAAGAAATGGCGACAAACCTACTT
GGCATGCCGCGTCTCGATGGTGCCGTTGGACATGGAACTACTTATAAAGACCTATGGCTTTAAGGTCCTCGACGCGTTCCAACGGAGGCCATACCGGTAAGTGTTCTTTACCGCTGTTTGGATGAA
G  Y  P  A  S  S  G  R  C  G  T  G  Q  H  I  N  R  S  V  S  I  G  P  A  A  C  P  Q  R  R  Y  P  W  E  C  S  I  A  V  F  R  S
 R  V  A  C  L  +  W  P  L  R  Y  R  S  S  Y  K  Q  I  G  F  N  W  S  S  R  L  T  A  E  P  I  A  M  *  L  F  H  R  C  V  +  K
  T  R  R  L  A  V  V  A  V  Q  V  K  I  F  I  E  P  Y  R  F  E  L  L  Q  A  L  N  G  G  T  H  G  N  V  L  F  P  S  L  G  V
```

```
CDS            65437  65814    no significant database hits
CDS            65901  66119 c  Similar to Escherichia coli prophage P2 Ogr protein SW:OGRK_ECOLI () (72 aa) fasta scores:
CDS            66187  67287 c  Similar to Bacteriophage P2 late gene control protein D SW:VPD_BPP2 (P10312) (387 aa) fast
CDS            67284  67769 c  Similar to Bacteriophage P2 complete genome U essential tail protein TR:O64315 (EMBL:AF063
CDS            67769  70549 c  Similar to Bacteriophage 186 protein G TR:Q37848 (EMBL:U32222) (812 aa) fasta scores: E():
CDS            70542  70661 c  Similar to Bacteriophage 186 Orf52 H TR:O80316 (EMBL:U32222) (58 aa) fasta scores: E(): 3.
CDS            70676  70978 c  Similar to Bacteriophage P2 complete genome E, essential tail protein TR:O64313 (EMBL:AF06
CDS            71033  71548 c  Similar to Bacteriophage P2 major tail tube protein fII SW:VPF2_BPP2 (P22502) (171 aa) fas
CDS            71558  72730 c  Similar to Bacteriophage P2 major tail sheath protein fI SW:VPF1_BPP2 (P22501) (395 aa) fa
CDS            73265  73987    Similar to Salmonella typhimurium invasion-associated secreted protein SopE TR:O52623 (EME
CDS            74185  74592 c  Similar to Bacteriophage P2 probable tail fiber assembly protein G SW:TFA_BPP2 (P26699) (1
CDS            74599  76218 c  Similar to Bacteriophage P2 probable tail fiber protein SW:VPH_BPP2 () (669 aa) fasta scor
CDS            76215  76820 c  Similar to Bacteriophage P2 tail protein I SW:VPI_BPP2 (P26701) (176 aa) fasta scores: E()
CDS            76813  77721 c  Similar to Bacteriophage P2 baseplate assembly protein J SW:VPJ_BPP2 (P51767) (302 aa) fas
CDS            77708  78067 c  Similar to Bacteriophage P2 baseplate assembly protein W SW:VPW_BPP2 (P51768) (115 aa) fas
CDS            78064  78642 c  Similar to Bacteriophage P2 baseplate assembly protein V SW:VPV_BPP2 (P31340) (211 aa) fas
```

While the region is highlighted, click on the ‘Create’ menu and select ‘Create feature from base range’. A feature edit window will appear. The default ‘Key’ value given by Artemis when creating a new feature is ‘CDS’. With this ‘Key’ the newly created feature would automatically be put on the translation line. However, if we change this to ‘misc_feature’ (an option in the ‘Key’ drop down menu in the top left hand corner of the Edit window), Artemis will place this feature on the DNA line. This is perhaps more appropriate and is easier to visualise. You can also add a qualifier, such as ‘/label’: select ‘label’ from the ‘Add Qualifier’ list and click ‘Add Qualifier’, ‘/label=’ will appear in the text window; add text of your choice, then click ‘OK’. That text will be used as a feature label to be displayed in the main sequence view panel.

To see how well you have done, tick the little box to turn on spi7.tab.

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.

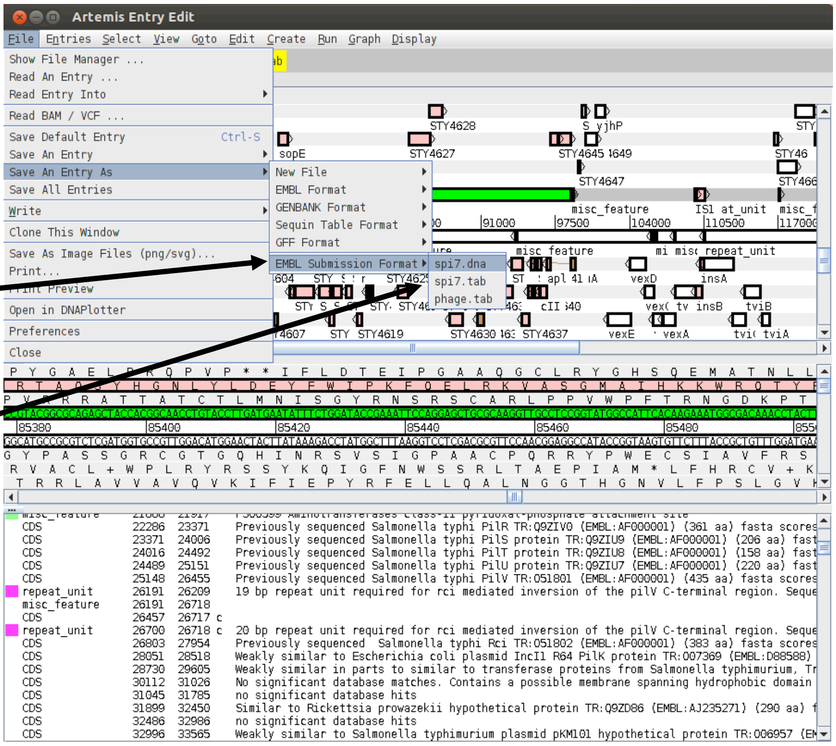

Artemis Entry Edit
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ spi7.dna ☑ spi7.tab
232 selected features
Undo Ctrl-U
Redo
Selected Features in Editor Ctrl-E
Subsequence (and Features)
Find/Replace Qualifier Text ...
Qualifier of Selected Feature(s)
Selected Feature(s)
Move Selected Features To
Copy Selected Features To
Trim Selected Features
Extend Selected Features
Fix Stop Codons
Automatically Create Gene Names
Fix Gene Names
Bases
Contig Reordering
Header Of Default Entry
spi7.dna
spi7.tab
phage.tab

1 Click 'Select' then 'All'

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

3 Select 'spi7.dna'

```
CDS            22286  23371    Previously sequenced Salmonella typhi PilR TR:Q9ZIV0 (EMBL:AF000001) (361 aa) fasta scores
CDS            23371  24006    Previously sequenced Salmonella typhi PilS protein TR:Q9ZIU9 (EMBL:AF000001) (206 aa) fast
CDS            24016  24492    Previously sequenced Salmonella typhi PilT protein TR:Q9ZIU8 (EMBL:AF000001) (158 aa) fast
CDS            24489  25151    Previously sequenced Salmonella typhi PilU protein TR:Q9ZIU7 (EMBL:AF000001) (220 aa) fast
CDS            25148  26455    Previously sequenced Salmonella typhi PilV TR:O51801 (EMBL:AF000001) (435 aa) fasta scores
repeat unit    26191  26209    19 bp repeat unit required for rci mediated inversion of the pilV C-terminal region. Seque
misc feature   26191  26718
CDS            26457  26717 c
repeat unit    26700  26718 c  20 bp repeat unit required for rci mediated inversion of the pilV C-terminal region. Seque
CDS            26803  27954    Previously sequenced  Salmonella typhi Rci TR:O51802 (EMBL:AF000001) (383 aa) fasta scores
CDS            28051  28518    Weakly similar to Escherichia coli plasmid IncI1 R64 PilK protein TR:O07369 (EMBL:D88588)
CDS            28730  29605    Weakly similar in parts to similar to transferase proteins from Salmonella typhimurium, Tr
CDS            30112  31026    No significant database matches. Contains a possible membrane spanning hydrophobic domain
CDS            31045  31785    no significant database hits
CDS            31899  32450    Similar to Rickettsia prowazekii hypothetical protein TR:Q9ZD86 (EMBL:AJ235271) (290 aa) f
CDS            32486  32986    no significant database hits
```



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

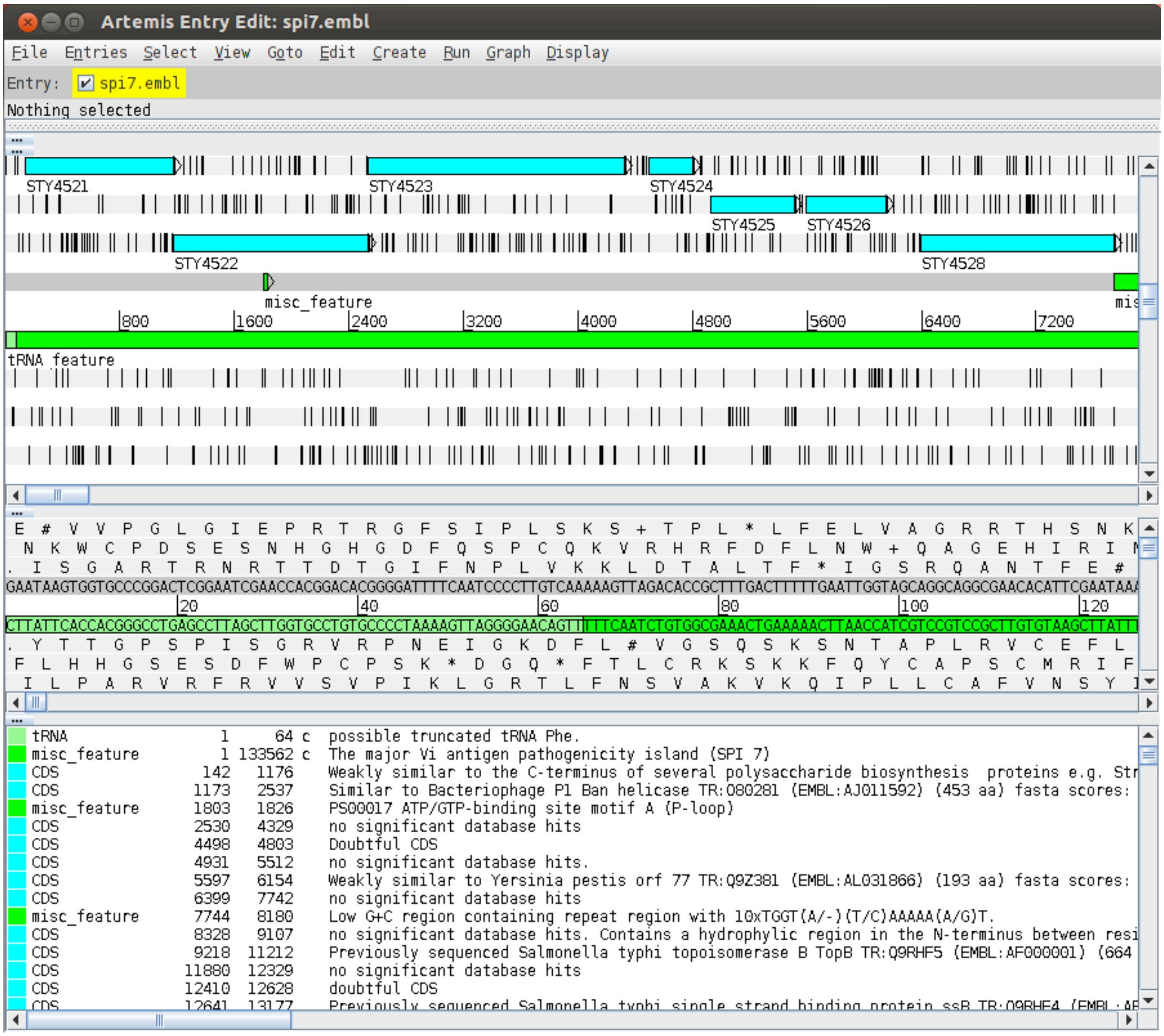
5 'EMBL Submission Format'

6 Select 'spi7.dna'

7 Save file as spi7.embl

```
CDS            22286  23371    Previously sequenced Salmonella typhi PilR TR:Q9ZIV0 (EMBL:AF000001) (361 aa) fasta scores
CDS            23371  24006    Previously sequenced Salmonella typhi PilS protein TR:Q9ZIU9 (EMBL:AF000001) (206 aa) fast
CDS            24016  24492    Previously sequenced Salmonella typhi PilT protein TR:Q9ZIU8 (EMBL:AF000001) (158 aa) fast
CDS            24489  25151    Previously sequenced Salmonella typhi PilU protein TR:Q9ZIU7 (EMBL:AF000001) (220 aa) fast
CDS            25148  26455    Previously sequenced Salmonella typhi PilV TR:O51801 (EMBL:AF000001) (435 aa) fasta scores
repeat_unit    26191  26209    19 bp repeat unit required for rci mediated inversion of the pilV C-terminal region. Seque
misc_feature   26191  26718
CDS            26457  26717 c
repeat_unit    26700  26718 c  20 bp repeat unit required for rci mediated inversion of the pilV C-terminal region. Seque
CDS            26803  27954    Previously sequenced  Salmonella typhi Rci TR:O51802 (EMBL:AF000001) (383 aa) fasta scores
CDS            28051  28518    Weakly similar to Escherichia coli plasmid IncI1 R64 PilK protein TR:O07369 (EMBL:D88588)
CDS            28730  29605    Weakly similar in parts to similar to transferase proteins from Salmonella typhimurium, Tr
CDS            30112  31026    No significant database matches. Contains a possible membrane spanning hydrophobic domain
CDS            31045  31785    no significant database hits
CDS            31899  32450    Similar to Rickettsia prowazekii hypothetical protein TR:Q9ZD86 (EMBL:AJ235271) (290 aa) f
CDS            32486  32986    no significant database hits
CDS            32996  33565    Weakly similar to Salmonella typhimurium plasmid pKM101 hypothetical protein TR:O06957 (EM
```

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



```
E # V V P G L G I E P R T R G F S I P L S K S + T P L * L F E L V A G R R T H S N K
 N K W C P D S E S N H G H G D F Q S P C Q K V R H R F D F L N W + Q A G E H I R I N
. I S G A R T R N R T T D T G I F N P L V K K L D T A L T F * I G S R Q A N T F E #
GAATAAGTGGTGCCCGGACTCGGAATCGAACCACGGACACGGGGATTTTCAATCCCCTTGTCAAAAAGTTAGACACCGCTTTGACTTTTTGAATTGGTAGCAGGCAGGCGAACACATTCGAATAAA
                    |20                |40                |60                |80                |100               |120
CTTATTCACCACGGGCCTGAGCCTTAGCTTGGTGCCTGTGCCCCTAAAAGTTAGGGGAACAGTTTTTCAATCTGTGGCGAAACTGAAAAACTTAACCATCGTCCGTCCGCTTGTGTAAGCTTATTT
. Y T T G P S P I S G R V R P N E I G K D F L # V G S Q S K S N T A P L R V C E F L
 F L H H G S E S D F W P C P S K * D G Q * F T L C R K S K K F Q Y C A P S C M R I F
  I L P A R V R F R V V S V P I K L G R T L F N S V A K V K Q I P L L C A F V N S Y I
```

```
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:O80281 (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)AAAAA(A/G)T.
CDS             8328   9107    no significant database hits. Contains a hydrophylic 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
```

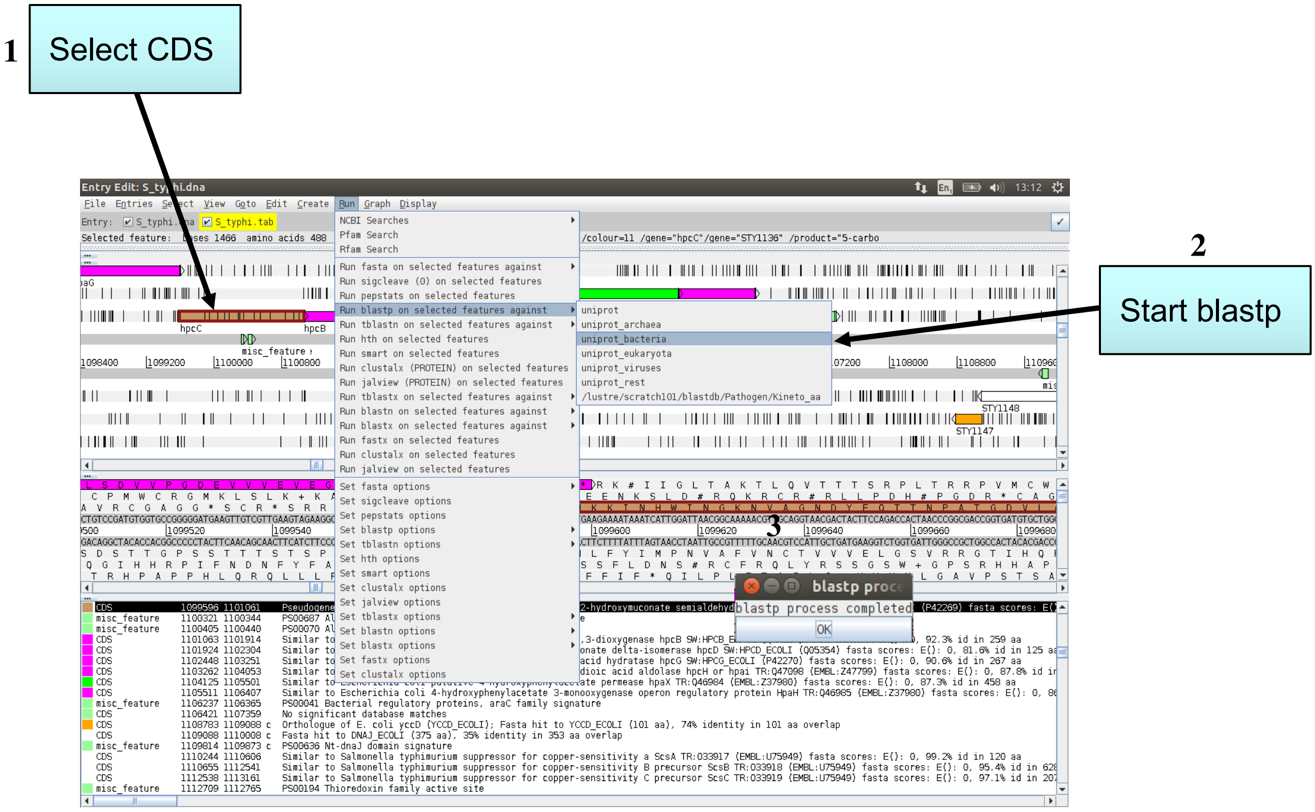
You will see that the colours of the features have now changed. This is because not all the qualifiers in the previous entry are accepted by the EMBL database, so some have not been saved in this format. This includes the '/colour' qualifier, so Artemis displays the features with default colours.

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.

Return to the *S*. Typhi window. The gene you will work on is *hpcC* (STY1136). Go to this gene by using one the different methods you have learned so far. You will now analyse the reading frame of *hpcC* gene. To do this switch on the stop codons (also, switch off the GC content graph if still displayed).
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).


Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ S_typhi.dna ☑ S_typhi.tab
hpcC
hpcB
misc_feature
STY1148
STY1147
NCBI Searches
Pfam Search
Rfam Search
Run fasta on selected features against
Run sigcleave (0) on selected features
Run pepstats on selected features
Run blastp on selected features against
Run tblastn on selected features against
Run hth on selected features
Run smart on selected features
Run clustalx (PROTEIN) on selected features
Run jalview (PROTEIN) on selected features
Run tblastx on selected features against
Run blastn on selected features against
Run blastx on selected features against
Run fastx on selected features
Run clustalx on selected features
Run jalview on selected features
Set fasta options
Set sigcleave options
Set pepstats options
Set blastp options
Set tblastn options
Set hth options
Set smart options
Set clustalx options
Set jalview options
Set tblastx options
Set blastn options
Set blastx options
Set fastx options
Set clustalx options
uniprot
uniprot_archaea
uniprot_bacteria
uniprot_eukaryota
uniprot_viruses
uniprot_rest
/lustre/scratch101/blastdb/Pathogen/Kineto_aa
blastp process
OK
3

1 Select CDS

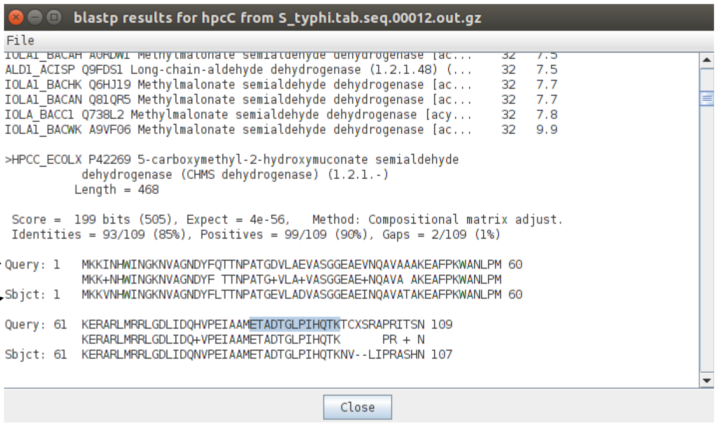
2 Start blastp

Selected feature: bases 1466 amino acids 488 /colour=11 /gene="hpcC"/gene="STY1136" /product="5-carbo

blastp process completed

```
CDS            1099596 1101061    Pseudogene 2-hydroxymuconate semialdehyd (P42269) fasta scores: E()
misc_feature   1100321 1100344    PS00687 Al
misc_feature   1100405 1100440    PS00070 Al
CDS            1101063 1101914    Similar to ,3-dioxygenase hpcB SW:HPCB_E , 92.3% id in 259 aa
CDS            1101924 1102304    Similar to onate delta-isomerase hpcD SW:HPCD_ECOLI (Q05354) fasta scores: E(): 0, 81.6% id in 125 aa
CDS            1102448 1103251    Similar to acid hydratase hpcG SW:HPCG_ECOLI (P42270) fasta scores: E(): 0, 90.6% id in 267 aa
CDS            1103262 1104053    Similar to dioic acid aldolase hpcH or hpai TR:Q47098 (EMBL:Z47799) fasta scores: E(): 0, 87.8% id in
CDS            1104125 1105501    Similar to ate permease hpaX TR:Q46984 (EMBL:Z37980) fasta scores: E(): 0, 87.3% id in 458 aa
CDS            1105511 1106407    Similar to Escherichia coli 4-hydroxyphenylacetate 3-monooxygenase operon regulatory protein HpaH TR:Q46985 (EMBL:Z37980) fasta scores: E(): 0, 86
misc_feature   1106237 1106365    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 in 101 aa overlap
CDS            1109088 1110008 c  Fasta hit to DNAJ_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 ScsA TR:O33917 (EMBL:U75949) fasta scores: E(): 0, 99.2% id in 120 aa
CDS            1110655 1112541    Similar to Salmonella typhimurium suppressor for copper-sensitivity B precursor ScsB TR:O33918 (EMBL:U75949) fasta scores: E(): 0, 95.4% id in 628
CDS            1112538 1113161    Similar to Salmonella typhimurium suppressor for copper-sensitivity C precursor ScsC TR:O33919 (EMBL:U75949) fasta scores: E(): 0, 97.1% id in 207
misc_feature   1112709 1112765    PS00194 Thioredoxin family active site
```

To view the search results click 'View', then 'Search Results', then 'blastp 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.



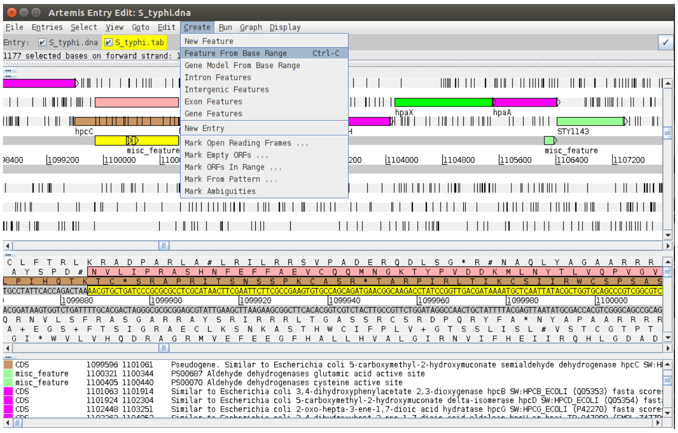
```
IOLA1_BACAH A0RDW1 Methylmalonate semialdehyde dehydrogenase [ac...    32   7.5
ALD1_ACISP Q9FDS1 Long-chain-aldehyde dehydrogenase (1.2.1.48) (...    32   7.5
IOLA1_BACHK Q6HJ19 Methylmalonate semialdehyde dehydrogenase [ac...    32   7.7
IOLA1_BACAN Q81QR5 Methylmalonate semialdehyde dehydrogenase [ac...    32   7.7
IOLA_BACC1 Q738L2 Methylmalonate semialdehyde dehydrogenase [acy...    32   7.8
IOLA1_BACWK A9VF06 Methylmalonate semialdehyde dehydrogenase [ac...    32   9.9

>HPCC_ECOLX P42269 5-carboxymethyl-2-hydroxymuconate semialdehyde
           dehydrogenase (CHMS dehydrogenase) (1.2.1.-)
          Length = 468

 Score =  199 bits (505), Expect = 4e-56,   Method: Compositional matrix adjust.
 Identities = 93/109 (85%), Positives = 99/109 (90%), Gaps = 2/109 (1%)

Query: 1   MKKINHWINGKNVAGNDYFQTTNPATGDVLAEVASGGEAEVNQAVAAAKEAFPKWANLPM 60
            MKK+NHWINGKNVAGNDYF TTNPATG+VLA+VASGGEAE+NQAVA AKEAFPKWANLPM
Sbjct: 1   MKKVNHWINGKNVAGNDYFLTTNPATGEVLADVASGGEAEINQAVATAKEAFPKWANLPM 60

Query: 61  KERARLMRRLGDLIDQHVPEIAAMETADTGLPIHQTKTCXSRAPRITSN 109
            KERARLMRRLGDLIDQ+VPEIAAMETADTGLPIHQTK      PR + N
Sbjct: 61  KERARLMRRLGDLIDQNVPEIAAMETADTGLPIHQTKNV--LIPRASHN 107
```

Our gene

Gene in database

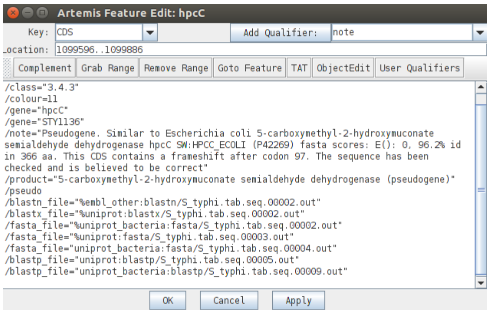
Can you see where the stop codon has been introduced into the sequence of our gene of interest? A stop codon is commonly marked with a * symbol. However, based on our blastp results it is marked as X. Search for the highlighted amino acid sequence in *hpcC*. Have a look if you can find the subsequent 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! The last amino acid in common is a K then the amino acids start to differ. The amino acid K is coded by AAA. 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 on the second frame (this will be the first amino acid after K, have a look at the blastp results 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.



```
CDS           1099596 1101061   Pseudogene. Similar to Escherichia coli 5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase hpcC SW:H
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_ECOLI (Q05353) fasta score
CDS           1101924 1102304   Similar to Escherichia coli 5-carboxymethyl-2-hydroxymuconate delta-isomerase hpcD SW:HPCD_ECOLI (Q05354) fast
CDS           1102448 1103251   Similar to Escherichia coli 2-oxo-hepta-3-ene-1,7-dioic acid hydratase hpcG SW:HPCG_ECOLI (P42270) fasta score
```

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’ (in our case this will be determined by the position of the last shared amino acid, K).



Key: CDS

Add Qualifier: note

Location: 1099596..1099886

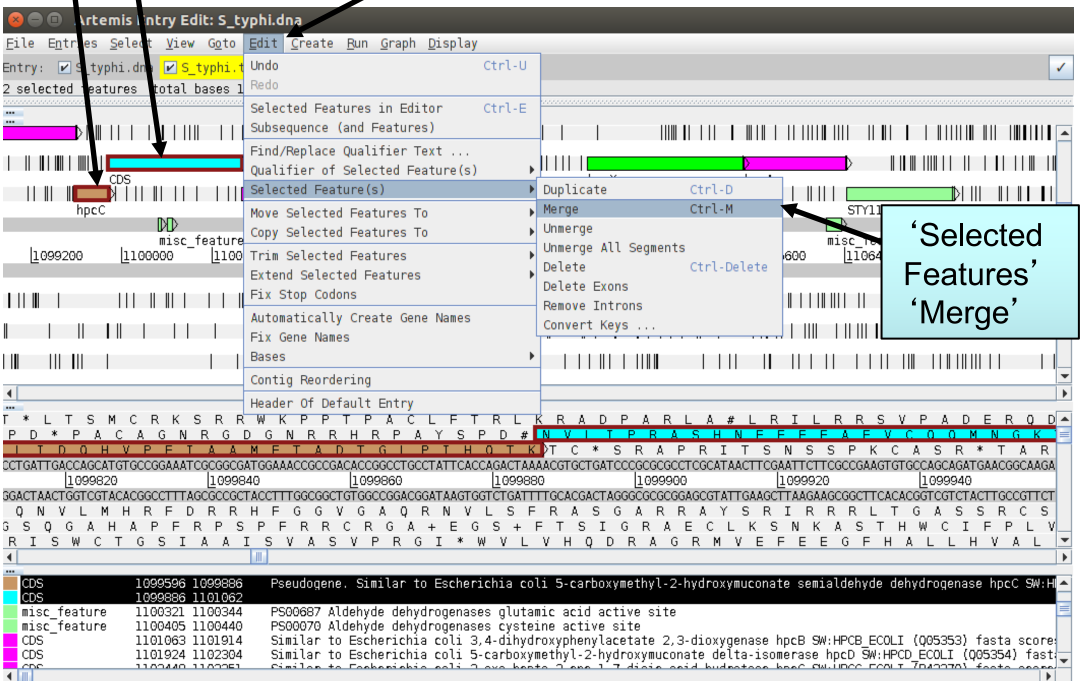
```
/class="3.4.3"
/colour=11
/gene="hpcC"
/gene="STY1136"
/note="Pseudogene. Similar to Escherichia coli 5-carboxymethyl-2-hydroxymuconate
semialdehyde dehydrogenase hpcC SW:HPCC_ECOLI (P42269) fasta scores: E(): 0, 96.2% id
in 366 aa. This CDS contains a frameshift after codon 97. The sequence has been
checked and is believed to be correct"
/product="5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase (pseudogene)"
/pseudo
/blastn_file="%embl_other:blastn/S_typhi.tab.seq.00002.out"
/blastx_file="%uniprot:blastx/S_typhi.tab.seq.00002.out"
/fasta_file="%uniprot_bacteria:fasta/S_typhi.tab.seq.00002.out"
/fasta_file="%uniprot:fasta/S_typhi.tab.seq.00003.out"
/fasta_file="uniprot_bacteria:fasta/S_typhi.tab.seq.00004.out"
/blastp_file="uniprot:blastp/S_typhi.tab.seq.00005.out"
/blastp_file="uniprot_bacteria:blastp/S_typhi.tab.seq.00009.out"
```

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.

1 Select both the original gene-model and the new CDS feature, which is to be merged with it to form a new gene (to select more than one feature you must hold the shift key down).

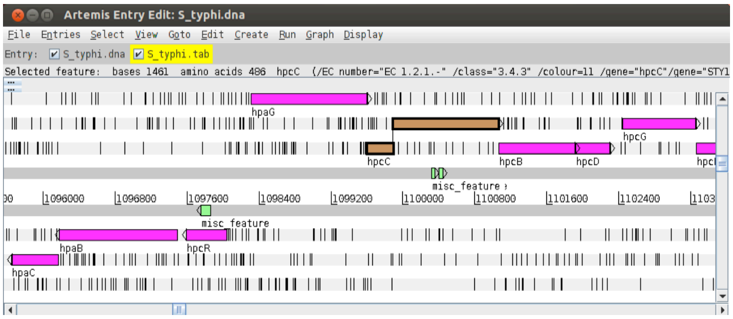
2 Click 'Edit'


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)
Move Selected Features To
Copy Selected Features To
Trim Selected Features
Extend Selected Features
Fix Stop Codons
Automatically Create Gene Names
Fix Gene Names
Bases
Contig Reordering
Header Of Default Entry
Duplicate Ctrl-D
Merge Ctrl-M
Unmerge
Unmerge All Segments
Delete Ctrl-Delete
Delete Exons
Remove Introns
Convert Keys ...
CDS
hpcC
misc_feature
STY11
1099200
1100000
1099820
1099840
1099860
1099880
1099900
1099920
1099940

```
CDS            1099596 1099886    Pseudogene. Similar to Escherichia coli 5-carboxymethyl-2-hydroxymuconate semialdehyde dehydrogenase hpcC SW:H
CDS            1099886 1101062
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_ECOLI (Q05353) fasta score
CDS            1101924 1102304    Similar to Escherichia coli 5-carboxymethyl-2-hydroxymuconate delta-isomerase hpcD SW:HPCD_ECOLI (Q05354) fast
```

3 'Selected Features' 'Merge'

result

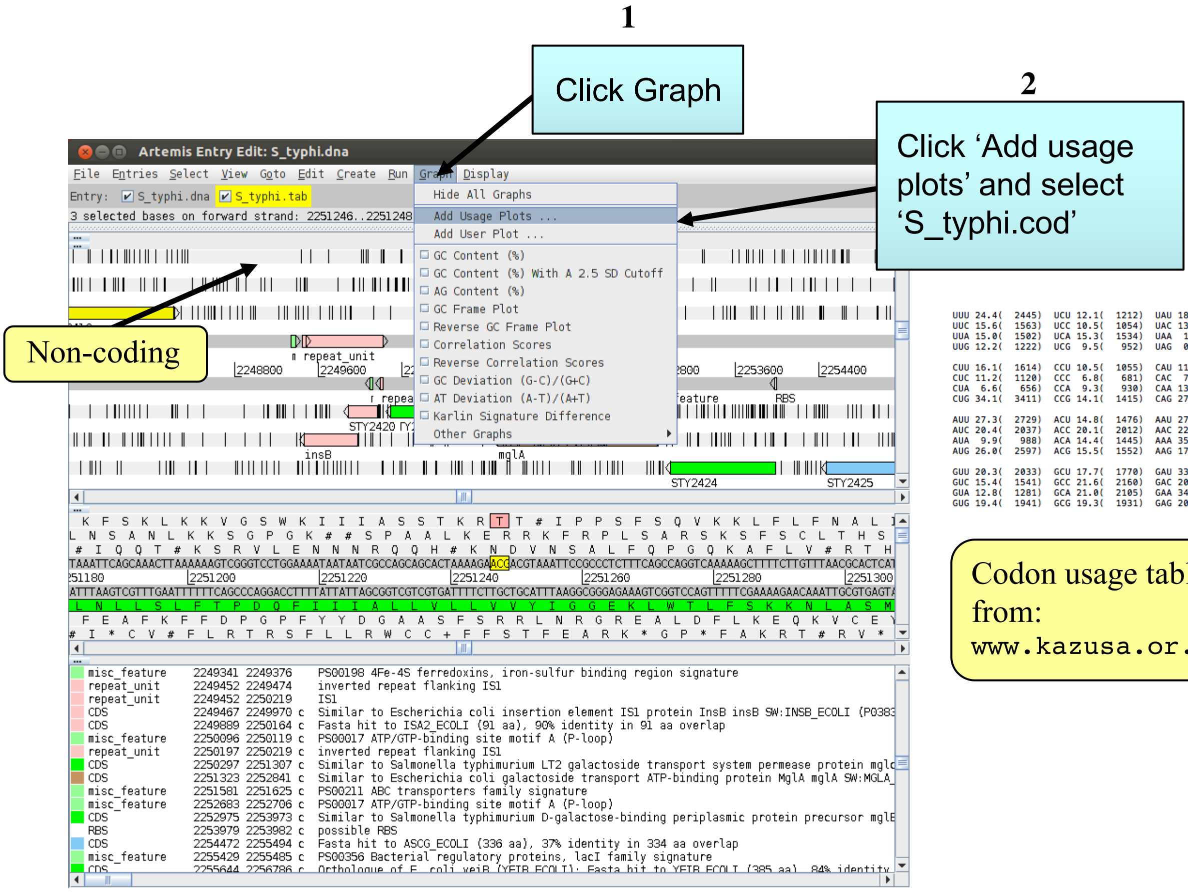

Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ S_typhi.dna ☑ S_typhi.tab
hpaG
hpcG
hpcC
hpcB
hpcD
misc_feature
1096000
1096800
1097600
1098400
1099200
1100000
1100800
1101600
1102400
misc_feature
hpaB
hpcR
hpaC

```
Selected feature:  bases 1461  amino acids 486  hpcC  (/EC number="EC 1.2.1.-" /class="3.4.3" /colour=11 /gene="hpcC"/gene="STY1
```

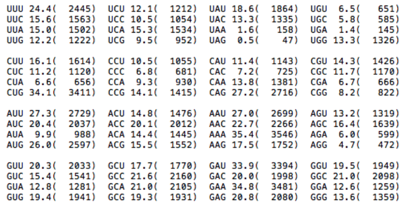
Artemis Exercise 4 - Second part

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 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
Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ S_typhi.dna ☑ S_typhi.tab
3 selected bases on forward strand: 2251246..2251248
Hide All Graphs
Add Usage Plots ...
Add User Plot ...
□ GC Content (%)
□ GC Content (%) With A 2.5 SD Cutoff
□ AG Content (%)
□ GC Frame Plot
□ Reverse GC Frame Plot
□ Correlation Scores
□ Reverse Correlation Scores
□ GC Deviation (G-C)/(G+C)
□ AT Deviation (A-T)/(A+T)
□ Karlin Signature Difference
Other Graphs
repeat_unit
2248800
2249600
2253600
2254400
feature
RBS
STY2420
insB
mglA
STY2424
STY2425
2251180
2251200
2251220
2251240
2251260
2251280
2251300

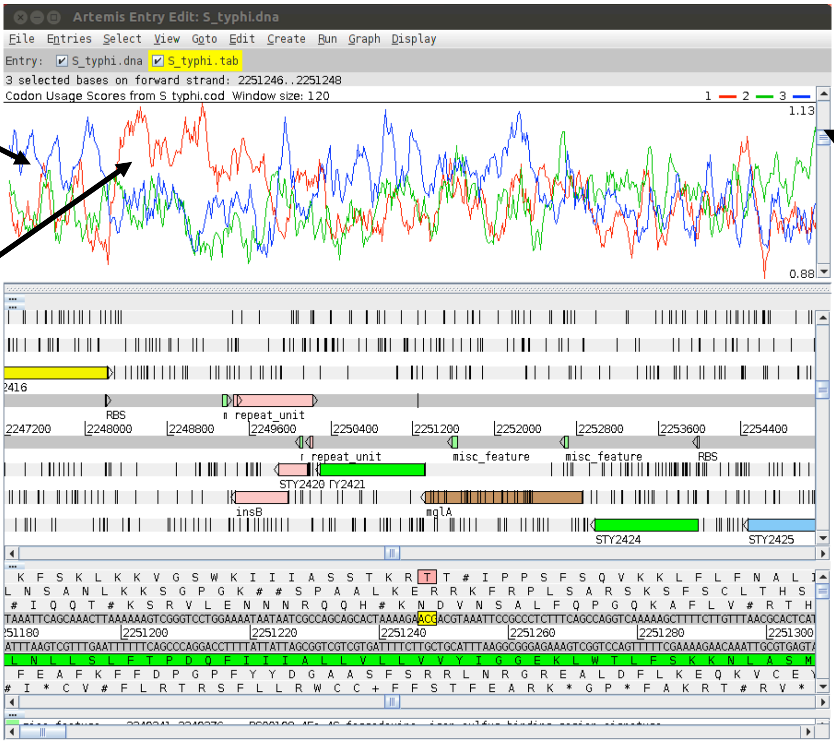
```
misc_feature   2249341 2249376    PS00198 4Fe-4S ferredoxins, iron-sulfur binding region signature
repeat_unit    2249452 2249474    inverted repeat flanking IS1
repeat_unit    2249452 2250219    IS1
CDS            2249467 2249970 c  Similar to Escherichia coli insertion element IS1 protein InsB insB SW:INSB_ECOLI (P0383
CDS            2249889 2250164 c  Fasta hit to ISA2_ECOLI (91 aa), 90% identity in 91 aa overlap
misc_feature   2250096 2250119 c  PS00017 ATP/GTP-binding site motif A (P-loop)
repeat_unit    2250197 2250219 c  inverted repeat flanking IS1
CDS            2250297 2251307 c  Similar to Salmonella typhimurium LT2 galactoside transport system permease protein mglC
CDS            2251323 2252841 c  Similar to Escherichia coli galactoside transport ATP-binding protein MglA mglA SW:MGLA_
misc_feature   2251581 2251625 c  PS00211 ABC transporters family signature
misc_feature   2252683 2252706 c  PS00017 ATP/GTP-binding site motif A (P-loop)
CDS            2252975 2253973 c  Similar to Salmonella typhimurium D-galactose-binding periplasmic protein precursor mglB
RBS            2253979 2253982 c  possible RBS
CDS            2254472 2255494 c  Fasta hit to ASCG_ECOLI (336 aa), 37% identity in 334 aa overlap
misc_feature   2255429 2255485 c  PS00356 Bacterial regulatory proteins, lacI family signature
CDS            2255644 2256786 c  Orthologue of E. coli yeiB (YEIB_ECOLI); Fasta hit to YEIB_ECOLI (385 aa), 84% identity
```



```
UUU 24.4(  2445)  UCU 12.1(  1212)  UAU 18.6(  1864)  UGU  6.5(   651)
UUC 15.6(  1563)  UCC 10.5(  1054)  UAC 13.3(  1335)  UGC  5.8(   585)
UUA 15.0(  1502)  UCA 15.3(  1534)  UAA  1.6(   158)  UGA  1.4(   145)
UUG 12.2(  1222)  UCG  9.5(   952)  UAG  0.5(    47)  UGG 13.3(  1326)

CUU 16.1(  1614)  CCU 10.5(  1055)  CAU 11.4(  1143)  CGU 14.3(  1426)
CUC 11.2(  1120)  CCC  6.8(   681)  CAC  7.2(   725)  CGC 11.7(  1170)
CUA  6.6(   656)  CCA  9.3(   930)  CAA 13.8(  1381)  CGA  6.7(   666)
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 27.0(  2699)  AGU 13.2(  1319)
AUC 20.4(  2037)  ACC 20.1(  2012)  AAC 22.7(  2266)  AGC 16.4(  1639)
AUA  9.9(   988)  ACA 14.4(  1445)  AAA 35.4(  3546)  AGA  6.0(   599)
AUG 26.0(  2597)  ACG 15.5(  1552)  AAG 17.5(  1752)  AGG  4.7(   472)

GUU 20.3(  2033)  GCU 17.7(  1770)  GAU 33.9(  3394)  GGU 19.5(  1949)
GUC 15.4(  1541)  GCC 21.6(  2160)  GAA 34.8(  3481)  GGA 12.6(  1259)
GUA 12.8(  1281)  GCA 21.0(  2105)  GAC 20.0(  1998)  GGC 21.0(  2098)
GUG 19.4(  1941)  GCG 19.3(  1931)  GAG 20.8(  2080)  GGG 13.6(  1359)
```

Codon usage table taken from:
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 remove one of these from the view click on the Graph menu and uncheck the box alongside the option ‘Reverse Codon Usage Scores from S_typhi.cod’.


Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ S_typhi.dna ☑ S_typhi.tab
1 2 3
1.13
0.88
RBS
repeat_unit
misc_feature
STY2420 STY2421
insB
mglA
STY2424
STY2425

3 selected bases on forward strand: 2251246..2251248

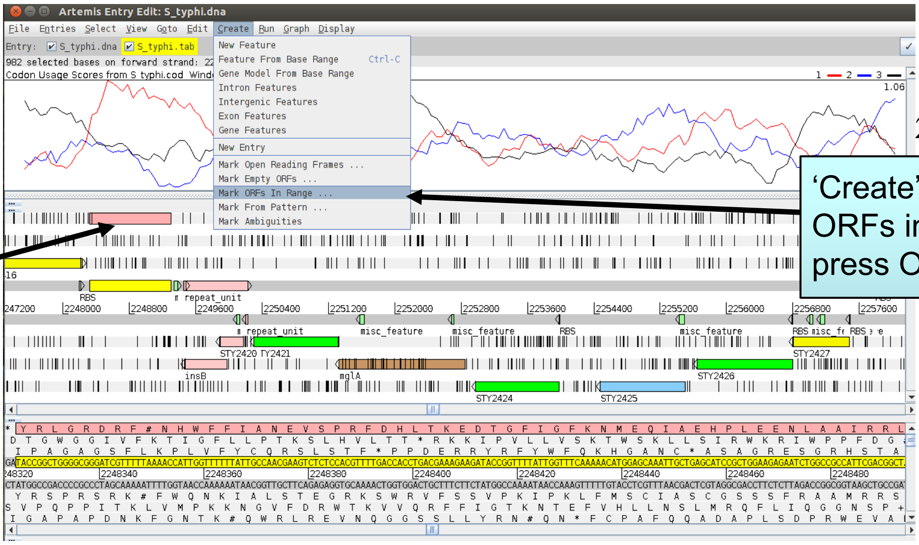
Codon Usage Scores from S_typhi.cod Window size: 120

Codon usage plot

Coding bubbles

Try the slider to smooth the graph

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.

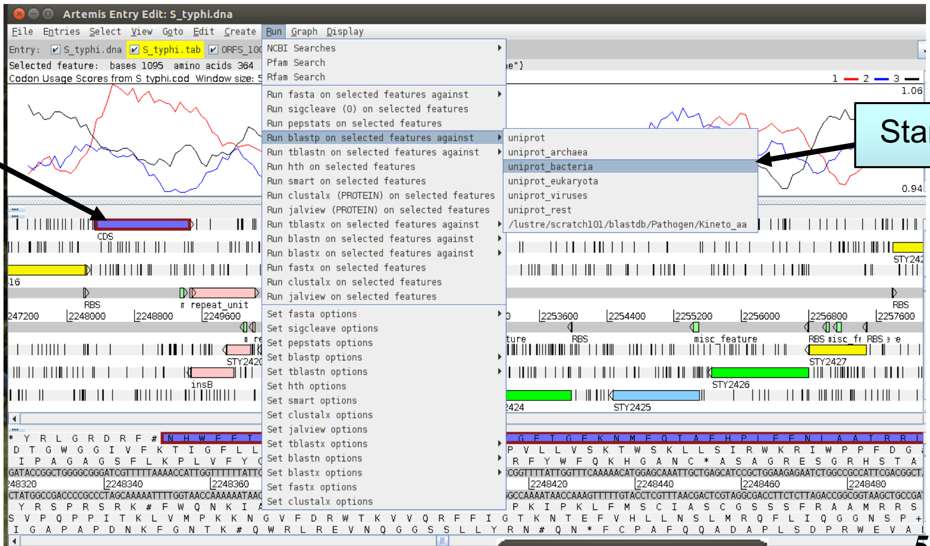

Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
New Feature
Feature From Base Range Ctrl-C
Gene Model From Base Range
Intron Features
Intergenic Features
Exon Features
Gene Features
New Entry
Mark Open Reading Frames ...
Mark Empty ORFs ...
Mark ORFs In Range ...
Mark From Pattern ...
Mark Ambiguities
Entry: ☑ S_typhi.dna ☑ S_typhi.tab
982 selected bases on forward strand
Codon Usage Scores from S typhi.cod
RBS
repeat_unit
misc_feature
insB
mglA
STY2420 TY2421
STY2424
STY2425
STY2426
STY2427

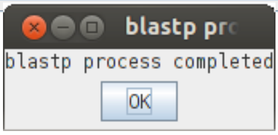
1

Click and drag to highlight

2

'Create', 'Mark ORFs in range', press OK


Artemis Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ S_typhi.dna ☑ S_typhi.tab ☑ ORFS_10
Selected feature: bases 1095 amino acids 364
Codon Usage Scores from S typhi.cod Window size: 5
NCBI Searches
Pfam Search
Rfam Search
Run fasta on selected features against
Run sigcleave (0) on selected features
Run pepstats on selected features
Run blastp on selected features against
Run tblastn on selected features against
Run hth on selected features
Run smart on selected features
Run clustalx (PROTEIN) on selected features
Run jalview (PROTEIN) on selected features
Run tblastx on selected features against
Run blastn on selected features against
Run blastx on selected features against
Run fastx on selected features
Run clustalx on selected features
Run jalview on selected features
Set fasta options
Set sigcleave options
Set pepstats options
Set blastp options
Set tblastn options
Set hth options
Set smart options
Set clustalx options
Set jalview options
Set tblastx options
Set blastn options
Set blastx options
Set fastx options
Set clustalx options
uniprot
uniprot_archaea
uniprot_bacteria
uniprot_eukaryota
uniprot_viruses
uniprot_rest
/lustre/scratch101/blastdb/Pathogen/Kineto_aa
CDS
RBS
repeat_unit
insB
STY2426
STY2427
STY2425



blastp process completed

3

Click on newly created ORF

4

Start blastp

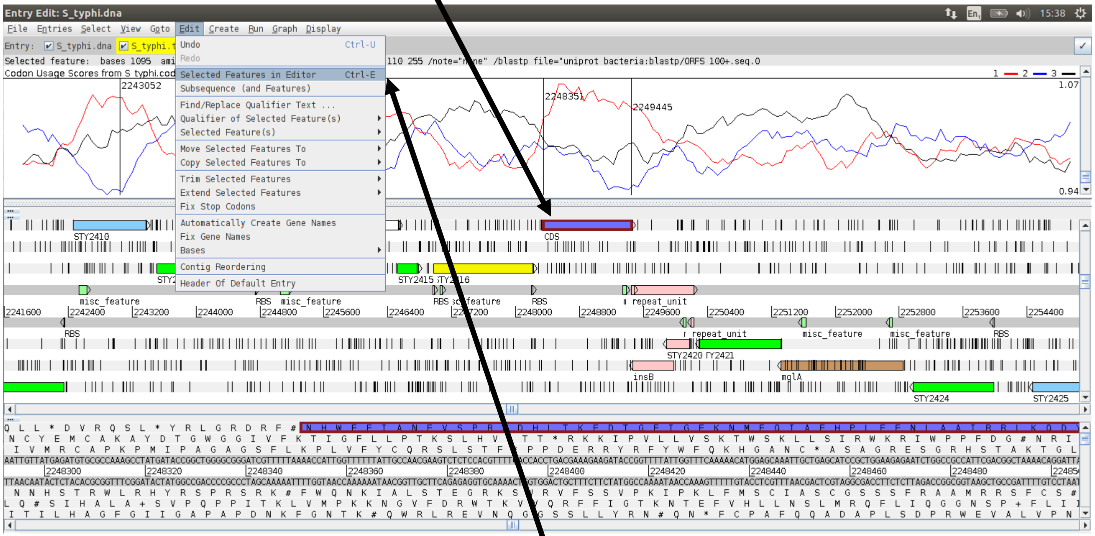
5

OK

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.

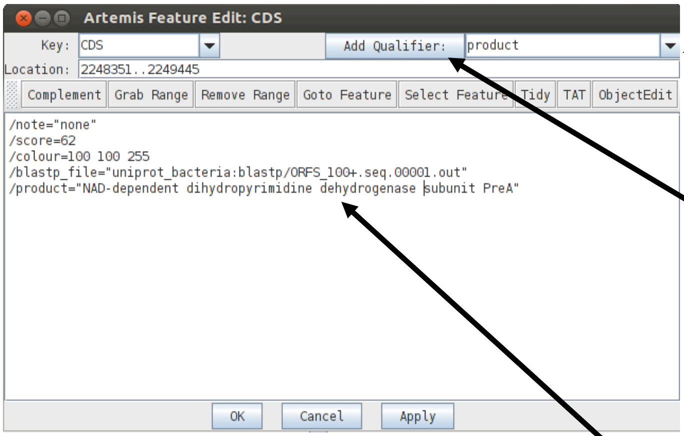
1

Click on ORF


Entry Edit: S_typhi.dna
File Entries Select View Goto Edit Create Run Graph Display
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)
Move Selected Features To
Copy Selected Features To
Trim Selected Features
Extend Selected Features
Fix Stop Codons
Automatically Create Gene Names
Fix Gene Names
Bases
Contig Reordering
Header Of Default Entry
2248351
2249445
CDS

2

‘Edit’, ‘Selected features in editor’



Key: CDS

Add Qualifier: product

Location: 2248351..2249445

```
/note="none"
/score=62
/colour=100 100 255
/blastp_file="uniprot_bacteria:blastp/ORFS_100+.seq.00001.out"
/product="NAD-dependent dihydropyrimidine dehydrogenase subunit PreA"
```

3

Select ‘product’ from dropdown list

4

Click ‘Add qualifier’

5

Add result from blastp 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.