Module 1
Artemis

Introduction

Artemis is a free DNA viewer and annotation tool written by Kim Rutherford (Rutherford *et al*., 2000). It is routinely used by the Pathogen Genomics Group for annotation and analysis of both prokaryotic and eukaryotic genomes. The program allows the user to view simple sequence files, EMBL/Genbank entries and the results of sequence analyses in a highly interactive and intuitive graphical format. Artemis is designed to present multiple sets/types of information within a single context. This manifests itself as the ability to zoom in to inspect DNA sequence motifs and zoom out to view local gene architecture (e.g. operons), several kilobases of a genome or even an entire genome in one screen. It is also possible to perform some analyses within Artemis with the output stored for later access.

Aims

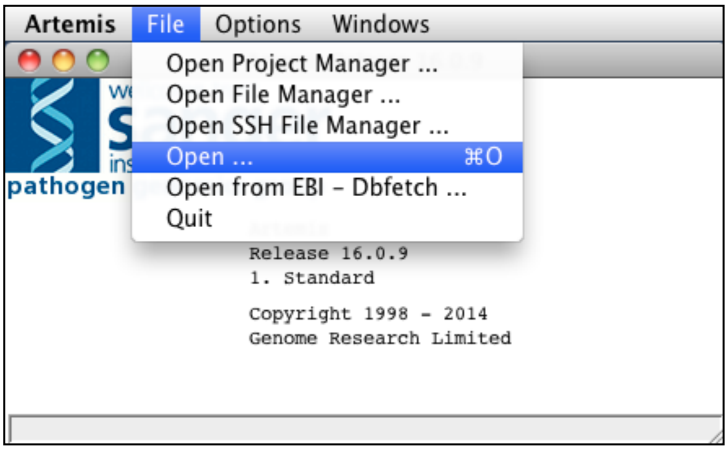
The aim of this Module is for you to become familiar with the basic functions of Artemis using a series of worked examples. These examples are designed to take you through the most immediately useful functions. However, there will be time, and encouragement, for you to explore other menus; nooks and crannies of Artemis that are not featured in the exercises in this manual. Like all the Modules in this workshop, the key is ‘if you don’t understand please ask’.

Artemis Exercise 1 Part I

1. Starting up the Artemis software

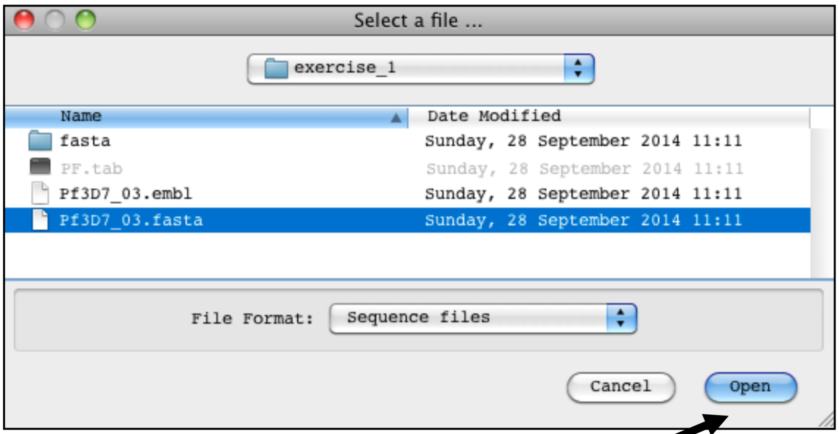
Double click the ARTEMIS Icon on your Desktop
A small start-up window will appear (see below).
Navigate to the directory Module_1_Artemis, exercise_1 containing the file Pf3D7_03.fasta using the file manager.

1
2
Click 'File' then 'Open'



```
Release 16.0.9
1. Standard

Copyright 1998 - 2014
Genome Research Limited
```

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.



Name	Date Modified
fasta	Sunday, 28 September 2014 11:11
PF.tab	Sunday, 28 September 2014 11:11
Pf3D7_03.embl	Sunday, 28 September 2014 11:11
Pf3D7_03.fasta	Sunday, 28 September 2014 11:11

File Format: Sequence files

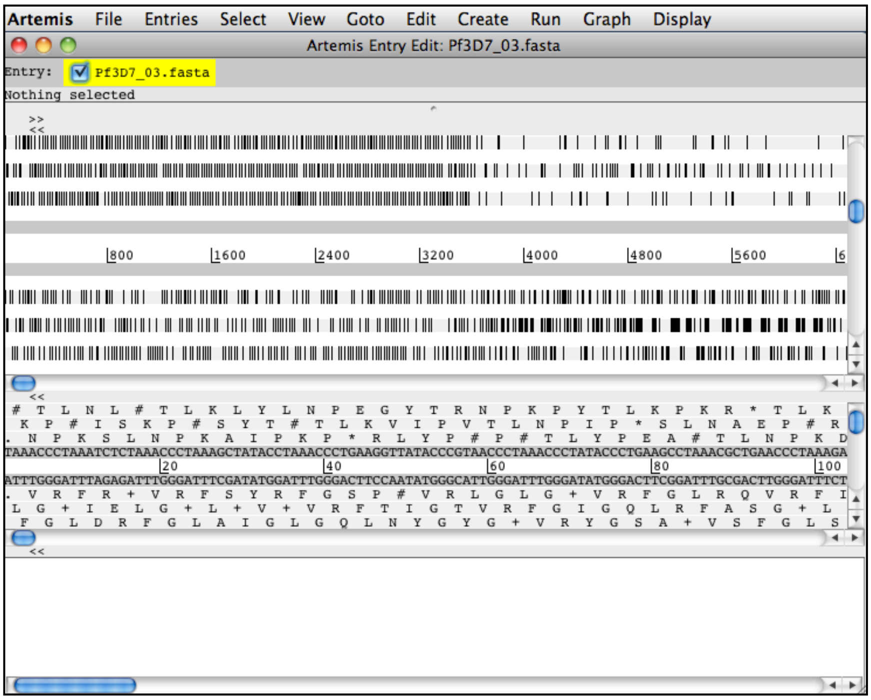
3
Single click to select the DNA file

4
Single click to open file in Artemis then wait

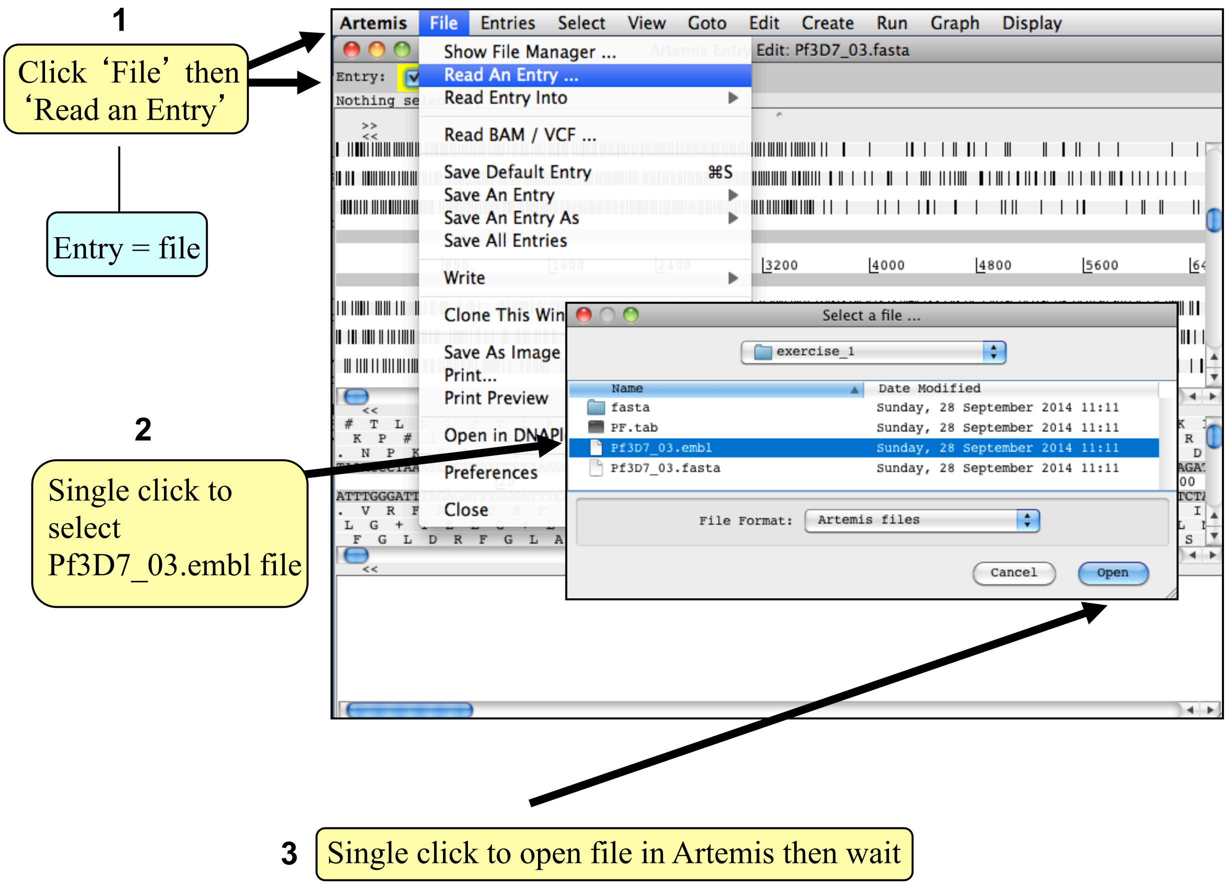
DNA sequence files will have the suffix '.fasta'. Annotation files end with '.embl', or '.tab'. Use this feature to select the type of file displayed in this panel.

2. Loading annotation files (entries) into Artemis

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


Artemis File Entries Select View Goto Edit Create Run Graph Display
Artemis Entry Edit: Pf3D7_03.fasta
Entry: ☑ Pf3D7_03.fasta
Nothing selected
800 1600 2400 3200 4000 4800 5600

Now follow the numbers to load up the annotation file for *Plasmodium falciparum* 3D7 chromosome 3.



1 Click 'File' then 'Read an Entry'

Entry = file

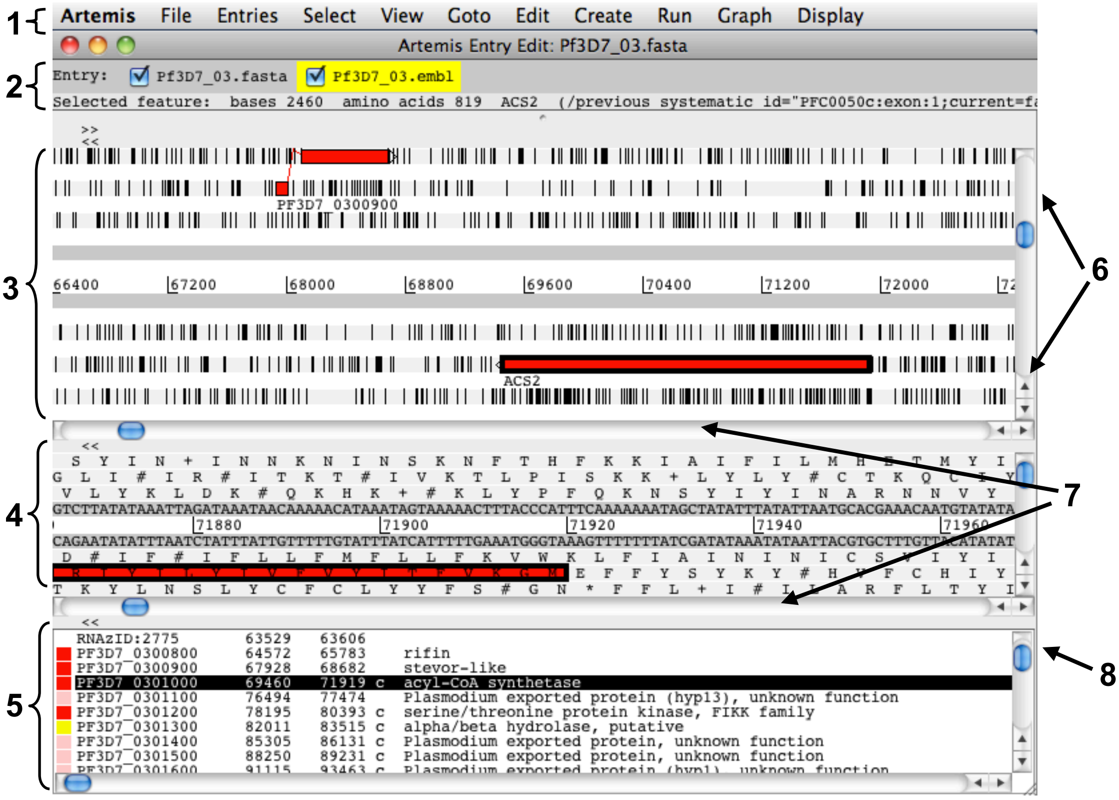
2 Single click to select Pf3D7_03.embl file

3 Single click to open file in Artemis then wait

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.

3. The basics of Artemis

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



```
Entry:  ☑ Pf3D7_03.fasta  ☑ Pf3D7_03.embl
Selected feature:  bases 2460  amino acids 819  ACS2  (/previous systematic id="PFC0050c:exon:1;current=fa
```

```
  RNAzID:2775       63529   63606
🟥 PF3D7_0300800     64572   65783     rifin
🟥 PF3D7_0300900     67928   68682     stevor-like
🟥 PF3D7_0301000     69460   71919 c   acyl-CoA synthetase
  PF3D7_0301100     76494   77474     Plasmodium exported protein (hyp13), unknown function
🟥 PF3D7_0301200     78195   80393 c   serine/threonine protein kinase, FIKK family
🟨 PF3D7_0301300     82011   83515 c   alpha/beta hydrolase, putative
  PF3D7_0301400     85305   86131 c   Plasmodium exported protein, unknown function
  PF3D7_0301500     88250   89231 c   Plasmodium exported protein, unknown function
  PF3D7_0301600     91115   93463 c   Plasmodium exported protein (hyp1), unknown function
```

1. Drop-down menus. There's lots in there so don't worry about them right now.
2. Shows what entries are currently loaded (bottom line) and gives details regarding the feature selected in the window below; in this case an acyl-CoA synthetase (selected 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 as black vertical bars. Genes and other features (eg. Pfam matches) are displayed as coloured boxes. We will refer to genes as coding sequences or CDSs from now on.
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 gene in the main view to see the zoomed view of the start of that gene. Note that both this and the main panel can be scrolled left and right (7, below) zoomed in and out (6, below).
5. This panel lists the various features in the order that they occur on the DNA with the selected gene 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.

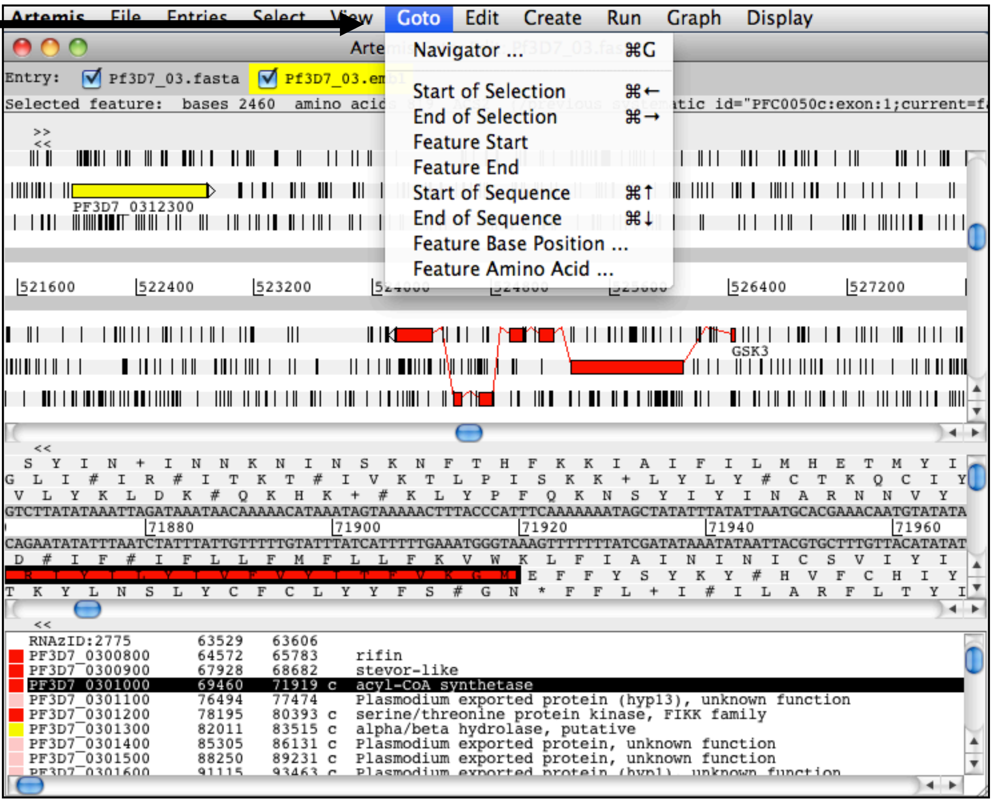
4. Getting around in Artemis

The 3 main ways of getting to where you want to be in Artemis are the Goto drop-down menu, the Navigator and the Feature Selector. The best method depends on what you're trying to do and knowing which one to use comes with practice.

4.1 The 'Goto' menu

The functions on this menu (ignore the Navigator for now) are shortcuts for getting to locations within a selected feature or for jumping to the start or end of the DNA sequence. Most are self-explanatory, so feel free to try any of them.

Click 'Goto'



```
RNAzID:2775        63529   63606
PF3D7_0300800      64572   65783     rifin
PF3D7_0300900      67928   68682     stevor-like
PF3D7_0301000      69460   71919 c   acyl-CoA synthetase
PF3D7_0301100      76494   77474     Plasmodium exported protein (hyp13), unknown function
PF3D7_0301200      78195   80393 c   serine/threonine protein kinase, FIKK family
PF3D7_0301300      82011   83515 c   alpha/beta hydrolase, putative
PF3D7_0301400      85305   86131 c   Plasmodium exported protein, unknown function
PF3D7_0301500      88250   89231 c   Plasmodium exported protein, unknown function
PF3D7_0301600      91115   93463 c   Plasmodium exported protein (hyp1), unknown function
```

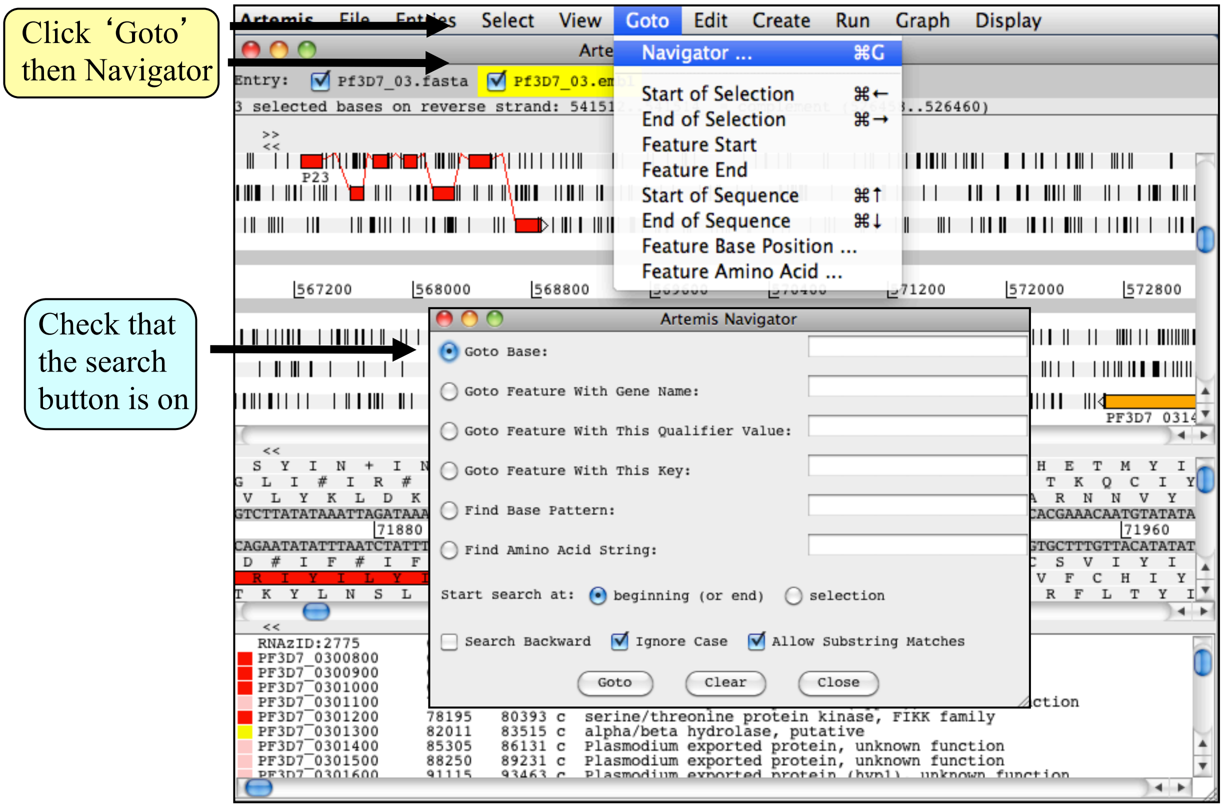
It may seem that 'Goto' 'Start of Selection' and 'Goto' 'Feature Start' do the same thing. Well they do if you have a feature selected but 'Goto' 'Start of Selection' will also work for a region which you have highlighted by click-dragging in the main window. So yes, give it a try! This is a very commonly used feature, so it is worth memorizing the keyboard short-cuts for these, ctrl<left arrow> and ctrl <right arrow> respectively.

Suggested tasks:

1. Zoom out, highlight a large region of sequence by clicking the left hand button and dragging the cursor then go to the start and end of the highlighted region.
2. Select a gene then go to the start and end.
3. Go to the start and end of the genome sequence.
4. Select a gene. Within it, go to a base (nucleotide) and/or amino acid of your choice.

4.2 Navigator

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



Click 'Goto' then Navigator

Check that the search button is on

Entry: ☑ Pf3D7_03.fasta ☑ Pf3D7_03.embl

● 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

□ Search Backward ☑ Ignore Case ☑ Allow Substring Matches

```
PF3D7_0301200   78195   80393 c  serine/threonine protein kinase, FIKK family
PF3D7_0301300   82011   83515 c  alpha/beta hydrolase, putative
PF3D7_0301400   85305   86131 c  Plasmodium exported protein, unknown function
PF3D7_0301500   88250   89231 c  Plasmodium exported protein, unknown function
```

Suggestions of where to go:

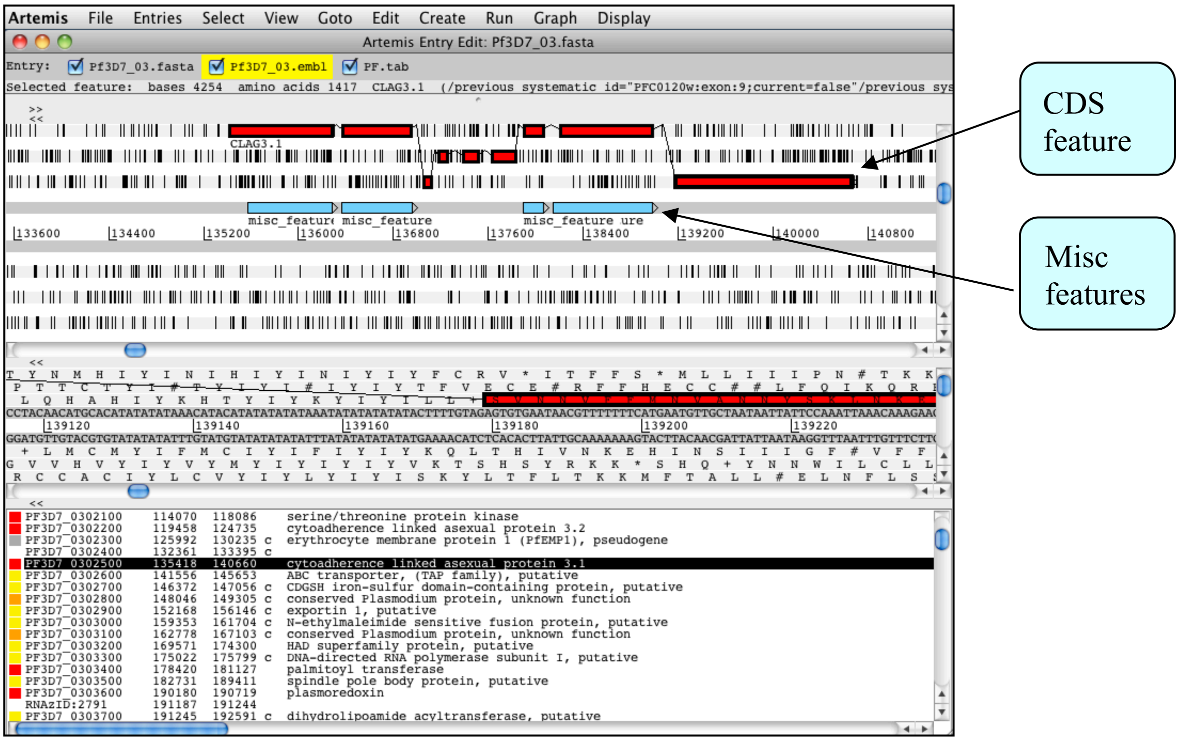
1. Think of a number between 1 and 1067971 and go to that base (notice how the cursors on the horizontal sliders move with you).
2. Your favourite gene name (it may not be there so you could try 'VAR').
3. Use 'Goto Feature With This Qualifier value' to search the contents of all qualifiers for a particular 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 through the pseudogenes as they occur on the chromosome.
4. tRNA genes. Type 'tRNA' in the 'Goto Feature With This Key'.
5. 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 1 Part II

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 located between bases 134000 to 141000 on the DNA sequence. This region encodes the *CLAG3.1* gene which codes for cytoadherence linked asexual protein. You can use either the Navigator, Feature Selector or Goto functions discussed previously to get there. The region you arrive at should look similar to that shown below.



```
Selected feature:  bases 4254  amino acids 1417  CLAG3.1  (/previous systematic id="PFC0120w:exon:9;current=false"/previous sys
```

```
T Y N M H I Y I N I H I Y I N I Y I Y F C R V * I T F F S * M L L I I I P N # T K K
 P T T C T Y I # T Y I Y I # I Y I Y T F V E C E # R F F H E C C # # L F Q I K Q R
  L Q H A H I Y K H T Y I Y K Y I Y I L L + S V N N V F F M N V A N N Y S K L N K F
CCTACAACATGCACATATATATAAACATACATATATATATAAATATATATATATACTTTTGTAGAGTGTGAATAACGTTTTTTCATGAATGTTGCTAATAATTATTCCAAATTAAACAAAGAAC
      |139120              |139140              |139160              |139180              |139200              |139220
GGATGTTGTACGTGTATATATATTTGTATGTATATATATATTTATATATATATATGAAAACATCTCACACTTATTGCAAAAAAAGTACTTACAACGATTATTAATAAGGTTTAATTTGTTTCTTG
  + L M C M Y I F M C I Y I F I Y I Y K Q L T H I V N K E H I N S I I I G F # V F F
G V V H V Y I Y V Y M Y I Y I Y I Y V K T S H S Y R K K * S H Q + Y N N W I L C L L
 R C C A C I Y L C V Y I Y L Y I Y I S K Y L T F L T K K M F T A L L # E L N F L S
```

```
PF3D7_0302100    114070  118086    serine/threonine protein kinase
PF3D7_0302200    119458  124735    cytoadherence linked asexual protein 3.2
PF3D7_0302300    125992  130235 c  erythrocyte membrane protein 1 (PfEMP1), pseudogene
PF3D7_0302400    132361  133395 c
PF3D7_0302500    135418  140660    cytoadherence linked asexual protein 3.1
PF3D7_0302600    141556  145653    ABC transporter, (TAP family), putative
PF3D7_0302700    146372  147056 c  CDGSH iron-sulfur domain-containing protein, putative
PF3D7_0302800    148046  149305 c  conserved Plasmodium protein, unknown function
PF3D7_0302900    152168  156146 c  exportin 1, putative
PF3D7_0303000    159353  161704 c  N-ethylmaleimide sensitive fusion protein, putative
PF3D7_0303100    162778  167103 c  conserved Plasmodium protein, unknown function
PF3D7_0303200    169571  174300    HAD superfamily protein, putative
PF3D7_0303300    175022  175799 c  DNA-directed RNA polymerase subunit I, putative
PF3D7_0303400    178420  181127    palmitoyl transferase
PF3D7_0303500    182731  189411    spindle pole body protein, putative
PF3D7_0303600    190180  190719    plasmoredoxin
RNAzID:2791      191187  191244
PF3D7_0303700    191245  192591 c  dihydrolipoamide acyltransferase, putative
```

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

Information to view:

Annotation

If you click on a particular feature you can view the annotation attached to it: select a CDS feature (or any other feature) and click on the 'Edit' menu and select 'Selected Feature in Editor', or simply push 'E'. A window will appear containing all the annotation that is associated with that CDS.

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 or FASTA. 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.

Load additional files

The results from the Pfam protein motif searches are not shown, but can be viewed by loading the appropriate file. Click on 'File' then 'Read an Entry' and select the file PF.tab. Each Pfam match will appear as a coloured blue feature in the main display panel on the grey DNA lines. To see the details click the feature then click 'View' then 'Selection' or click 'Edit' then 'Selected Features in Editor'. Please ask if you are unsure about Pfam.

Viewing the results of database searches

Click the 'View' menu, then select 'Search Results' and then 'Fasta results'. The results of the database search will appear in a scrollable window.

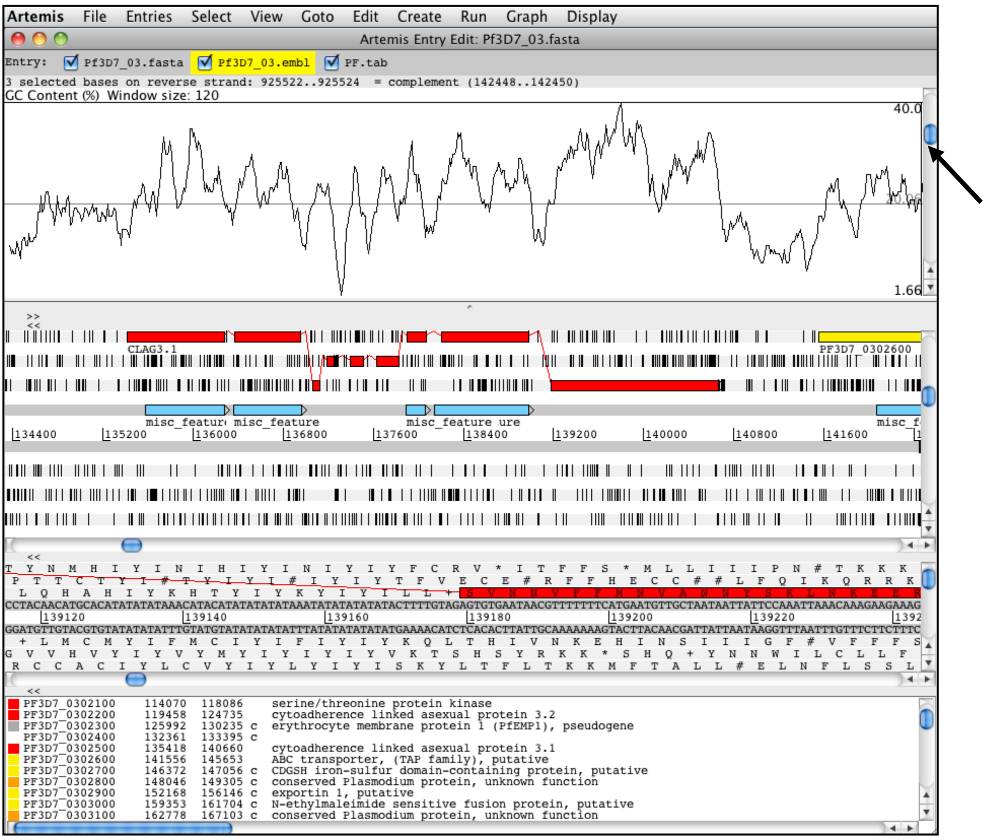
Further information on specific Pfam entries can be found on the web at http://pfam.xfam.org/

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 to the display various plots showing different characteristics of the DNA.

To view the graphs:

Click on the ʻGraphʼ menu to see all those available. Some of the most useful plots for *P. falciparum* is the ʻGC Content (%)ʼ as shown below. G+C content is a very good indicator of coding capacity in Malaria. On average, the coding regions are ~23% G+C and the non-coding regions are ~19%. Have a look at the G+C content for this region by selecting the appropriate graph. Left click within the graph window and then select by clicking on the exons to see how this relates to the G+C peaks on the graph.

DNA plot



```
3 selected bases on reverse strand: 925522..925524  = complement (142448..142450)
```

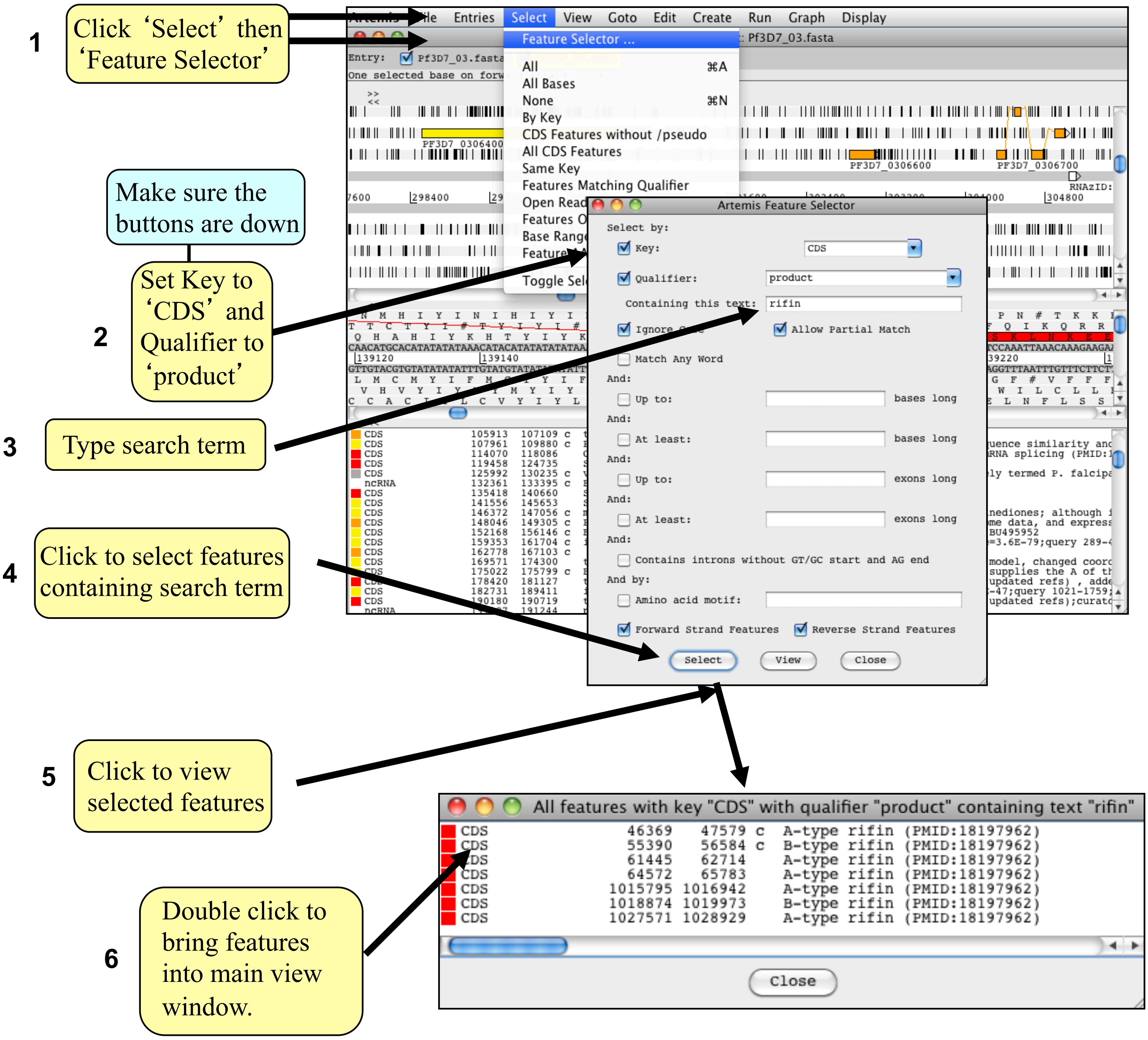
```
PF3D7_0302100   114070  118086    serine/threonine protein kinase
PF3D7_0302200   119458  124735    cytoadherence linked asexual protein 3.2
PF3D7_0302300   125992  130235 c  erythrocyte membrane protein 1 (PfEMP1), pseudogene
PF3D7_0302400   132361  133395 c
PF3D7_0302500   135418  140660    cytoadherence linked asexual protein 3.1
PF3D7_0302600   141556  145653    ABC transporter, (TAP family), putative
PF3D7_0302700   146372  147056 c  CDGSH iron-sulfur domain-containing protein, putative
PF3D7_0302800   148046  149305 c  conserved Plasmodium protein, unknown function
PF3D7_0302900   152168  156146 c  exportin 1, putative
PF3D7_0303000   159353  161704 c  N-ethylmaleimide sensitive fusion protein, putative
PF3D7_0303100   162778  167103 c  conserved Plasmodium protein, unknown function
```

Sliders for adjusting the window size

Artemis Exercise 1 Part III

In this part of the Module we will be looking at methods of selecting and extracting features. We are going to extract different genes and regions and perform some more detailed analysis on it. We will aim to write and save new EMBL format files which will include just the annotation and DNA for this region.

In Artemis you can select genes fitting different search criteria. One possibility is to look for a specific product, for example *rifin*, as shown below.



1 Click ‘Select’ then ‘Feature Selector’

Make sure the buttons are down

2 Set Key to ‘CDS’ and Qualifier to ‘product’

3 Type search term

4 Click to select features containing search term

5 Click to view selected features

6 Double click to bring features into main view window.

Select by:

☑ Key: CDS

☑ Qualifier: product

Containing this text: rifin

☑ Ignore Case ☑ Allow Partial Match

□ Match Any Word

And: □ Up to: bases long

And: □ At least: bases long

And: □ Up to: exons long

And: □ At least: exons long

And: □ Contains introns without GT/GC start and AG end

And by: □ Amino acid motif:

☑ Forward Strand Features ☑ Reverse Strand Features

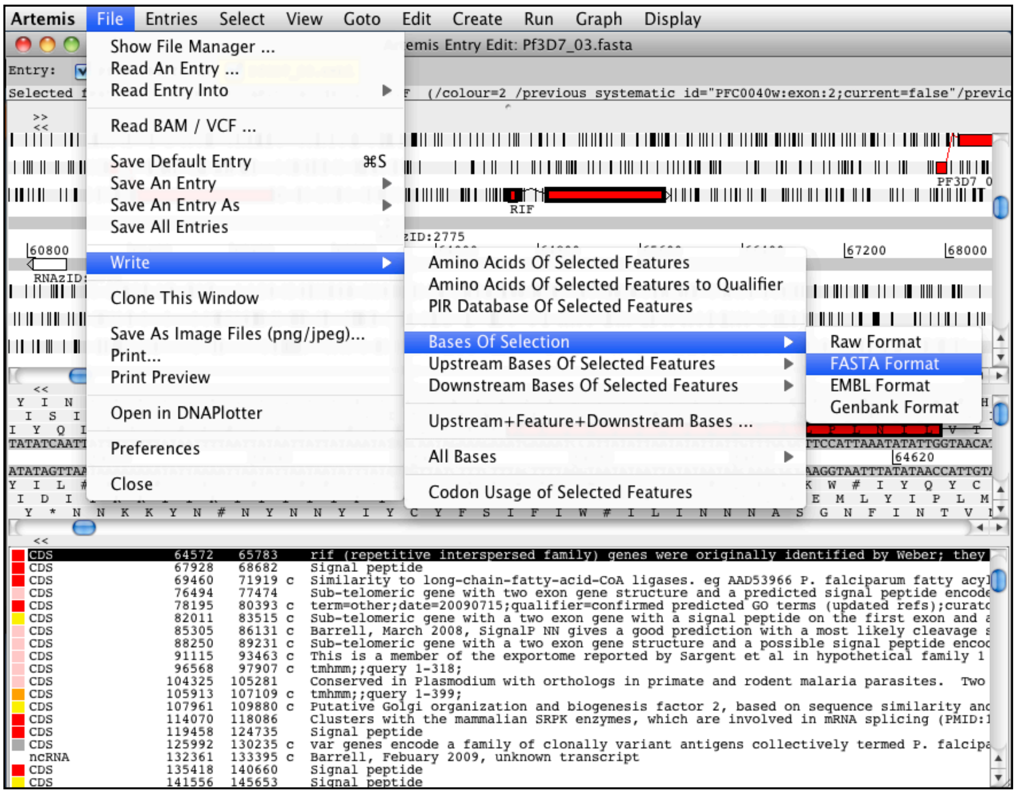
All features with key "CDS" with qualifier "product" containing text "rifin"

```
CDS       46369    47579 c  A-type rifin (PMID:18197962)
CDS       55390    56584 c  B-type rifin (PMID:18197962)
CDS       61445    62714    A-type rifin (PMID:18197962)
CDS       64572    65783    A-type rifin (PMID:18197962)
CDS     1015795  1016942    A-type rifin (PMID:18197962)
CDS     1018874  1019973    B-type rifin (PMID:18197962)
CDS     1027571  1028929    A-type rifin (PMID:18197962)
```

Close

The genes listed in **6** (on the previous page) are only those fitting your selection criterion. They can be copied or moved in to a new entry so they can be viewed in isolation from the rest of the information within Pf3D7_03.embl. To create a new entry go to 'Create' and choose 'New Entry'.

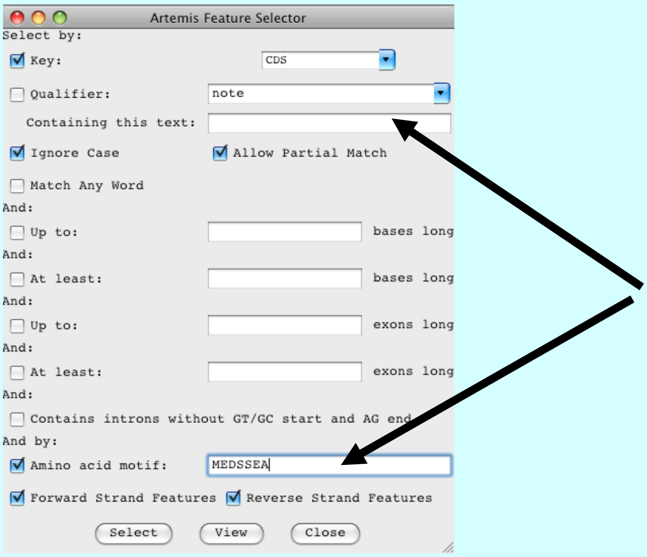
In the next step of the exercise choose one of the selected genes and write out a fasta-file of the sequence.



```
CDS        64572   65783     rif (repetitive interspersed family) genes were originally identified by Weber; they
CDS        67928   68682     Signal peptide
CDS        69460   71919 c   Similarity to long-chain-fatty-acid-CoA ligases. eg AAD53966 P. falciparum fatty acyl
CDS        76494   77474     Sub-telomeric gene with two exon gene structure and a predicted signal peptide encode
CDS        78195   80393 c   term=other;date=20090715;qualifier=confirmed predicted GO terms (updated refs);curatc
CDS        82011   83515 c   Sub-telomeric gene with a two exon gene with a signal peptide on the first exon and a
CDS        85305   86131 c   Barrell, March 2008, SignalP NN gives a good prediction with a most likely cleavage s
CDS        88250   89231 c   Sub-telomeric gene with a two exon gene structure and a possible signal peptide encod
CDS        91115   93463 c   This is a member of the exportome reported by Sargent et al in hypothetical family 1
CDS        96568   97907 c   tmhmm;;query 1-318;
CDS       104325  105281     Conserved in Plasmodium with orthologs in primate and rodent malaria parasites.  Two
CDS       105913  107109 c   tmhmm;;query 1-399;
CDS       107961  109880 c   Putative Golgi organization and biogenesis factor 2, based on sequence similarity and
CDS       114070  118086     Clusters with the mammalian SRPK enzymes, which are involved in mRNA splicing (PMID:1
CDS       119458  124735     Signal peptide
CDS       125992  130235 c   var genes encode a family of clonally variant antigens collectively termed P. falcipa
ncRNA     132361  133395 c   Barrell, Febuary 2009, unknown transcript
CDS       135418  140660     Signal peptide
CDS       141556  145653     Signal peptide
```

Click 'File' then 'Write 'Bases of Selection' 'FASTA Format'

Additional methods of 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 such as text or amino acid searches.



Artemis Feature Selector

Select by:

☑ Key: CDS

☐ Qualifier: note

Containing this text:

☑ Ignore Case ☑ Allow Partial Match

☐ Match Any Word

And:

☐ Up to: bases long

And:

☐ At least: bases long

And:

☐ Up to: exons long

And:

☐ At least: exons long

And:

☐ Contains introns without GT/GC start and AG end

And by:

☑ Amino acid motif: MEDSSEA

☑ Forward Strand Features ☑ Reverse Strand Features

Space for a search term or amino acid motif

In the next part of the exercise we will be looking at the region containing the *rifin* genes in more detail. They are located at the end of the chromosomes, in the subtelomeric region. We are going to extract this region from the whole chromosome sequence. Then we will aim to write and save new EMBL format files which will include just the annotation and DNA for this region.

3

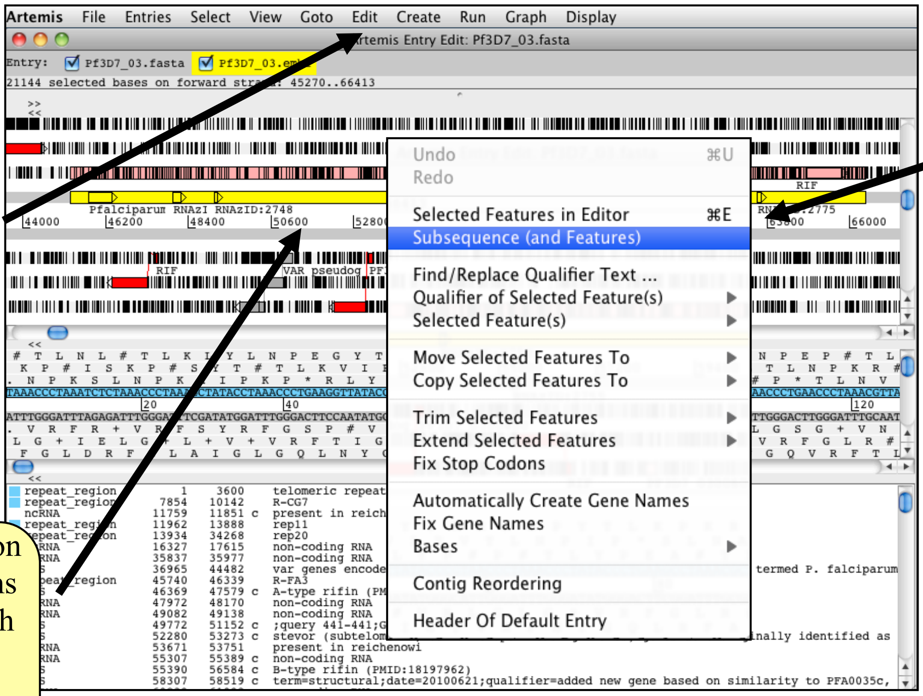
Click ‘Edit Subsequence (and Features)’

2

Click ‘Edit’

1

Select the region containing rifins by clicking with the left mouse button and dragging.

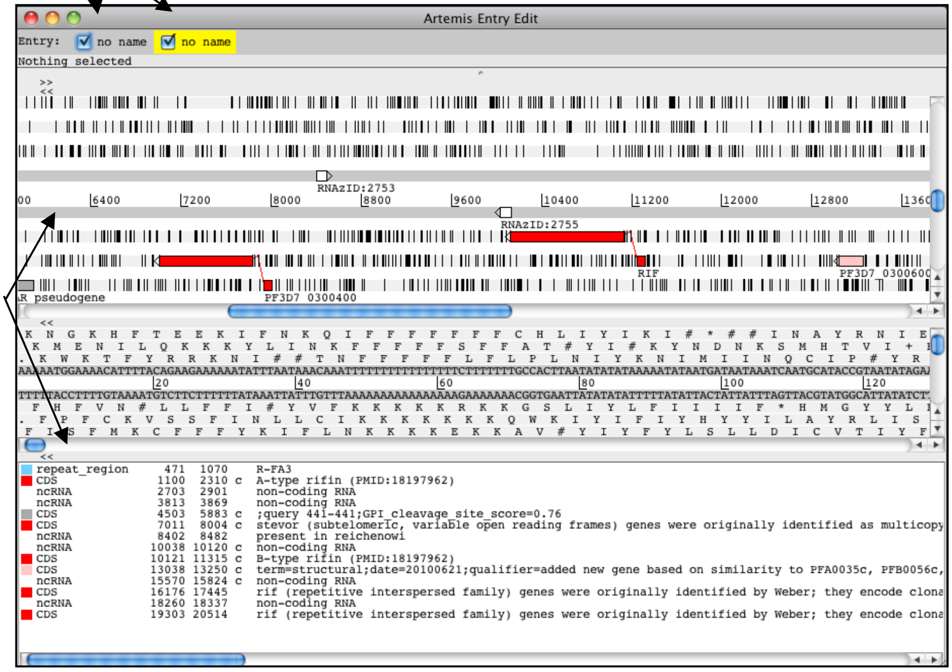

Artemis File Entries Select View Goto Edit Create Run Graph Display
Artemis Entry Edit: Pf3D7_03.fasta
Entry: ☑ Pf3D7_03.fasta ☑ Pf3D7_03.em
21144 selected bases on forward strand: 45270..66413
Undo ⌘U
Redo
Selected Features in Editor ⌘E
Subsequence (and Features)
Find/Replace Qualifier Text ...
Qualifier 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
Pfalciparum RNAzI RNAzID:2748
RIF
VAR pseudog
44000
46200
48400
50600
66000

4

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

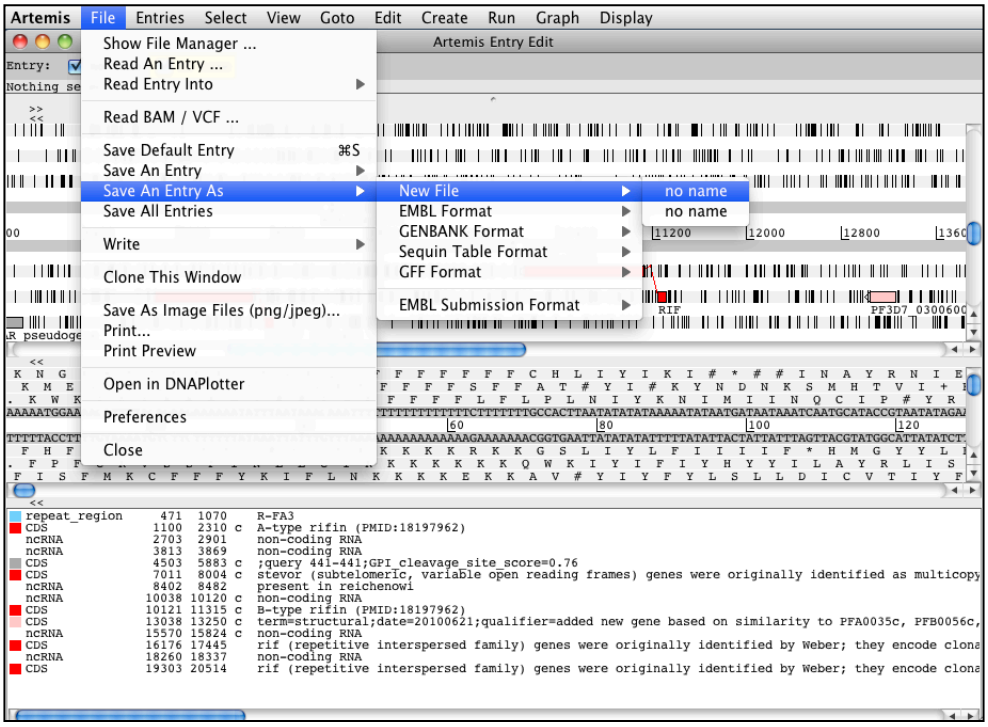
Note the entry names have changed

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



```
repeat_region     471  1070    R-FA3
CDS              1100  2310 c  A-type rifin (PMID:18197962)
ncRNA            2703  2901    non-coding RNA
ncRNA            3813  3869    non-coding RNA
CDS              4503  5883 c  ;query 441-441;GPI_cleavage_site_score=0.76
CDS              7011  8004 c  stevor (subtelomeric, variable open reading frames) genes were originally identified as multicopy
ncRNA            8402  8482    present in reichenowi
ncRNA           10038 10120 c  non-coding RNA
CDS             10121 11315 c  B-type rifin (PMID:18197962)
CDS             13038 13250 c  term=structural;date=20100621;qualifier=added new gene based on similarity to PFA0035c, PFB0056c,
ncRNA           15570 15824 c  non-coding RNA
CDS             16176 17445    rif (repetitive interspersed family) genes were originally identified by Weber; they encode clona
ncRNA           18260 18337    non-coding RNA
CDS             19303 20514    rif (repetitive interspersed family) genes were originally identified by Weber; they encode clona
```

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 name. 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. The new files can be saved in different formats.



```
🟦 repeat_region     471  1070     R-FA3
🟥 CDS              1100  2310 c   A-type rifin (PMID:18197962)
   ncRNA            2703  2901     non-coding RNA
   ncRNA            3813  3869     non-coding RNA
▨ CDS              4503  5883 c   ;query 441-441;GPI_cleavage_site_score=0.76
🟥 CDS              7011  8004 c   stevor (subtelomeric, variable open reading frames) genes were originally identified as multicopy
   ncRNA            8402  8482     present in reichenowi
   ncRNA           10038 10120 c   non-coding RNA
🟥 CDS             10121 11315 c   B-type rifin (PMID:18197962)
   CDS             13038 13250 c   term=structural;date=20100621;qualifier=added new gene based on similarity to PFA0035c, PFB0056c,
   ncRNA           15570 15824 c   non-coding RNA
🟥 CDS             16176 17445     rif (repetitive interspersed family) genes were originally identified by Weber; they encode clona
   ncRNA           18260 18337     non-coding RNA
🟥 CDS             19303 20514     rif (repetitive interspersed family) genes were originally identified by Weber; they encode clona
```

Once you have finished this exercise remember to close this Artemis session down completely before starting the next exercise.

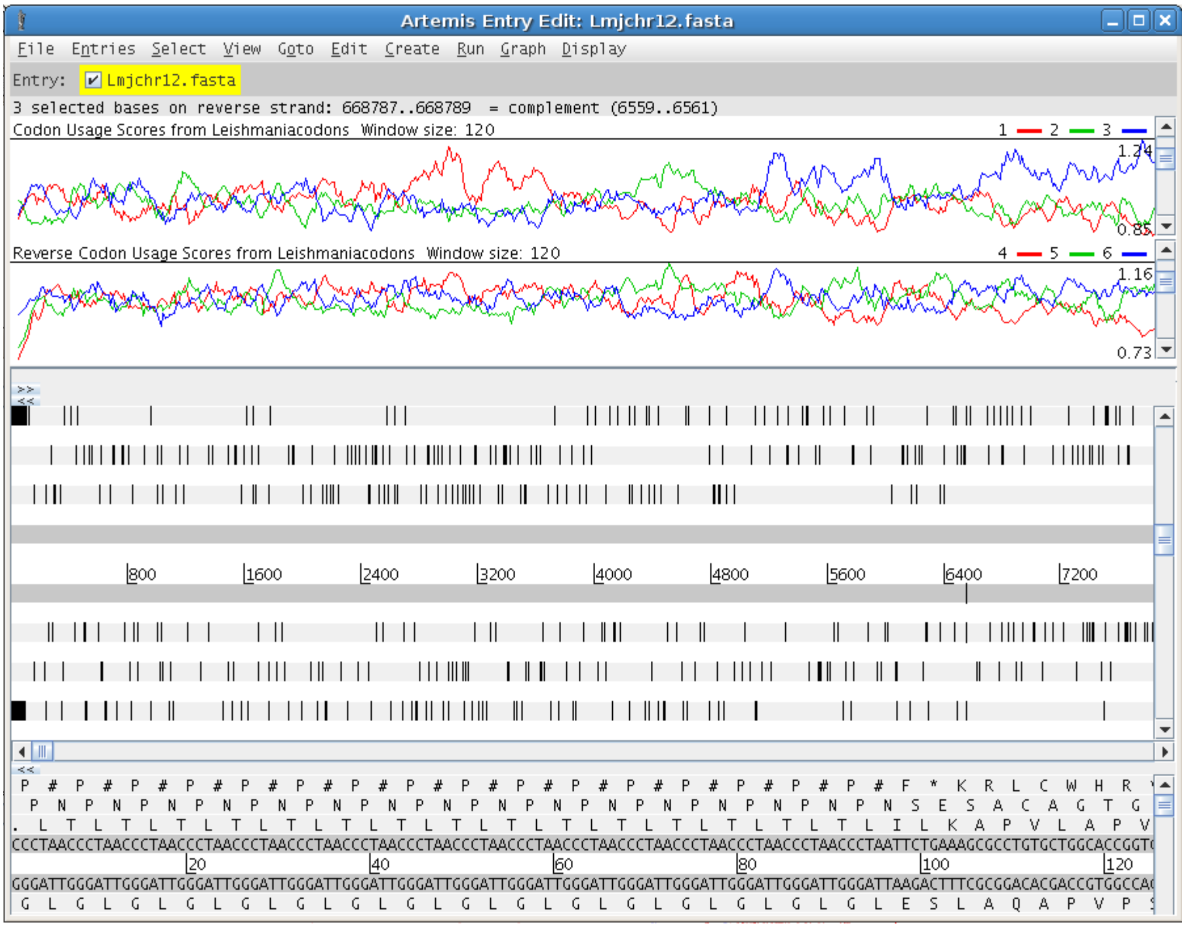
Artemis Exercise 2

We are now switching to a different organism. The following exercise demonstrates how to use Artemis as a tool for structural annotation. Given a length of chromosome with no existing annotation Artemis can mark up ORFs above a given size. This also shows how codon usage plots can be exploited in gene model prediction.

If you haven’t already closed the previous session of Artemis, do so now. Double click the ARTEMIS Icon on your Desktop and navigate to the directory Module_1_Artemis, exercise_2 and open the sequence file Lmjchr12.fasta.

Next, open the codon usage table file LmjF12codons by selecting ‘Add Usage Plots’ from the Graph menu. Codon usage is a very good indicator of coding capacity in *Leishmania* genomes where there is a much more prominent codon bias for some amino acids.

Note, we will cover the use of RNAseq data in gene prediction later on during the course.


Artemis Entry Edit: Lmjchr12.fasta
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ Lmjchr12.fasta
Codon Usage Scores from Leishmaniacodons Window size: 120
1 2 3
1.74
0.85
Reverse Codon Usage Scores from Leishmaniacodons Window size: 120
4 5 6
1.16
0.73
800
1600
2400
3200
4000
4800
5600
6400
7200

```
3 selected bases on reverse strand: 668787..668789  = complement (6559..6561)
```

```
P # P # P # P # P # P # P # P # P # P # P # P # P # P # P # F * K R L C W H R
 P N P N P N P N P N P N P N P N P N P N P N P N P N P N P N S E S A C A G T G
. L T L T L T L T L T L T L T L T L T L T L T L T L T L T L I L K A P V L A P V
CCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAACCCTAATTCTGAAAGCGCCTGTGCTGGCACCGGT
          20          40          60          80          100          120
GGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTGGGATTAAGACTTTCGCGGACACGACCGTGGCCA
 G L G L G L G L G L G L G L G L G L G L G L G L G L G L G L E S L A Q A P V P
```

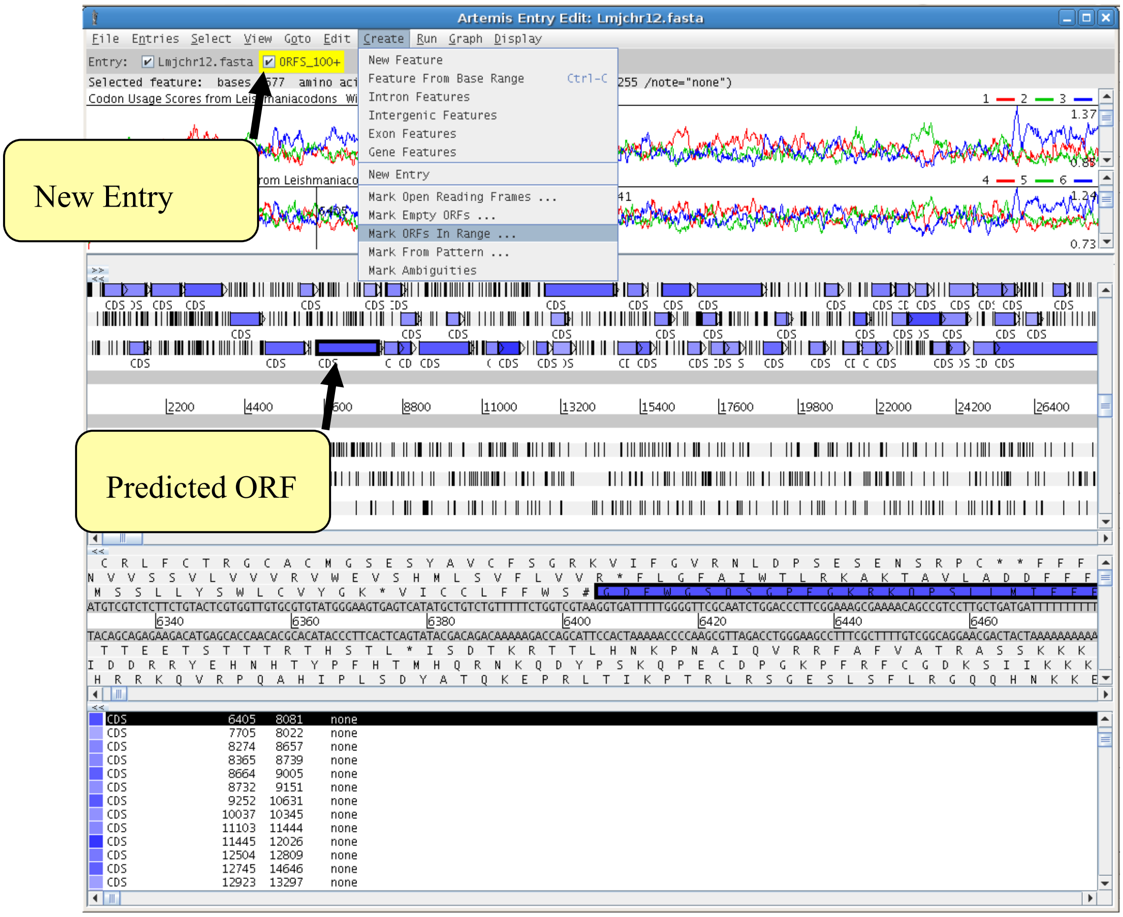
Select the first 100 kbs of sequence on the positive strand either by highlighting the sequence in the sequence window (use shift and click to select the final base) or choose the ‘Base Range’ option in the select menu and enter ‘1..100000’.

With this region selected, select ‘Mark ORFs in Range’ from the Create menu. When prompted for minimum ORF size enter 100. Note that this results in the creation of a new entry called ‘ORFS_100+’. You can experiment with a range of ORF sizes by de-selecting this entry and repeating the first steps in this process.

Note that the marked up ORFs vary in colour from pale to navy blue. This colouring reflects the codon usage support for this model with darker blue being highly supported by codon usage.

Try selecting some of the newly created features in the gene window. Double clicking on one of these will bring up the predicted peptide sequence in the bottom window. You can rapdily move to the N- or C-terminus of the predicted peptide by holding down ctrl, and then left or right arrow respectively.

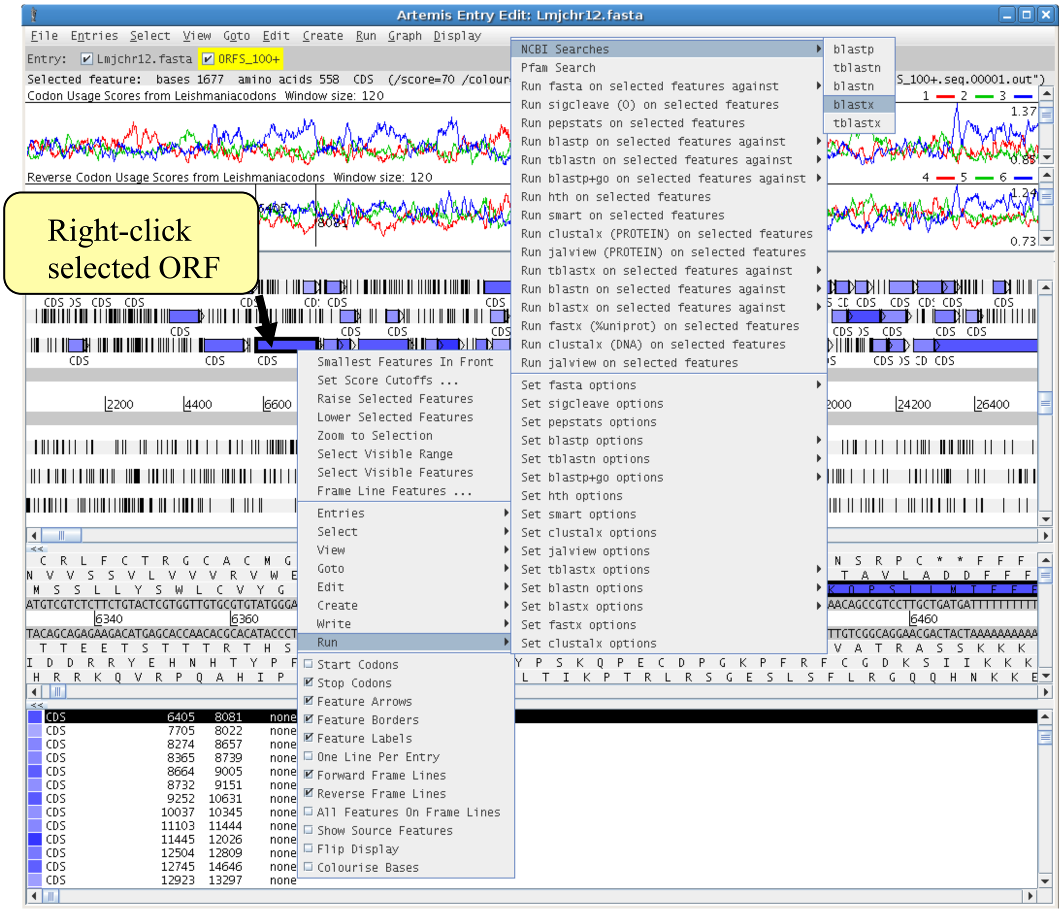
Note that we have chosen only to generate ORFs for the positive strand for this example. In a genome not organized into transcription units we would normally do likewise for the reverse strand as well.



Artemis Entry Edit: Lmjchr12.fasta
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ Lmjchr12.fasta ☑ ORFS_100+
New Feature
Feature From Base Range Ctrl-C
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
Codon Usage Scores from Leishmaniacodons
New Entry
Predicted ORF

```
CDS     6405   8081   none
CDS     7705   8022   none
CDS     8274   8657   none
CDS     8365   8739   none
CDS     8664   9005   none
CDS     8732   9151   none
CDS     9252  10631   none
CDS    10037  10345   none
CDS    11103  11444   none
CDS    11445  12026   none
CDS    12504  12809   none
CDS    12745  14646   none
CDS    12923  13297   none
```

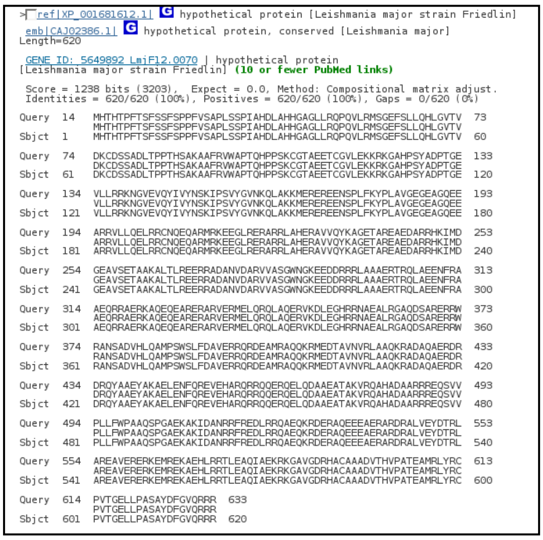
Although some of these predictions are likely to be correct, there is considerable overlap between predicted ORFs, and many are small and unsupported by codon usage. To validate/negate our predicted models we need to do further sequence comparison. This can be done with a tool such as ACT (to be discussed later in Module 2), or with one of the integrated Blast options in Artemis. Select the ORF at position 12745, click on it, then select RUN>NCBI Searches>blastx. This will open a browser window with NCBI results.


Artemis Entry Edit: Lmjchr12.fasta
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ Lmjchr12.fasta ☑ ORFS_100+
Codon Usage Scores from Leishmaniacodons Window size: 120
Reverse Codon Usage Scores from Leishmaniacodons Window size: 120
CDS
2200
4400
6600
22000
24200
26400
6340
6360
6460
NCBI Searches
Pfam 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 blastp+go 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 (%uniprot) on selected features
Run clustalx (DNA) on selected features
Run jalview on selected features
Set fasta options
Set sigcleave options
Set pepstats options
Set blastp options
Set tblastn options
Set blastp+go 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
blastp
tblastn
blastn
blastx
tblastx
Smallest Features In Front
Set Score Cutoffs ...
Raise Selected Features
Lower Selected Features
Zoom to Selection
Select Visible Range
Select Visible Features
Frame Line Features ...
Entries
Select
View
Goto
Edit
Create
Write
Run
☐ Start Codons
☑ Stop Codons
☑ Feature Arrows
☑ Feature Borders
☑ Feature Labels
☐ One Line Per Entry
☑ Forward Frame Lines
☑ Reverse Frame Lines
☐ All Features On Frame Lines
☐ Show Source Features
☐ Flip Display
☐ Colourise Bases

```
Selected feature:  bases 1677  amino acids 558  CDS  (/score=70 /colour ... S_100+.seq.00001.out")
```

```
CDS    6405   8081  none
CDS    7705   8022  none
CDS    8274   8657  none
CDS    8365   8739  none
CDS    8664   9005  none
CDS    8732   9151  none
CDS    9252  10631  none
CDS   10037  10345  none
CDS   11103  11444  none
CDS   11445  12026  none
CDS   12504  12809  none
CDS   12745  14646  none
CDS   12923  13297  none
```

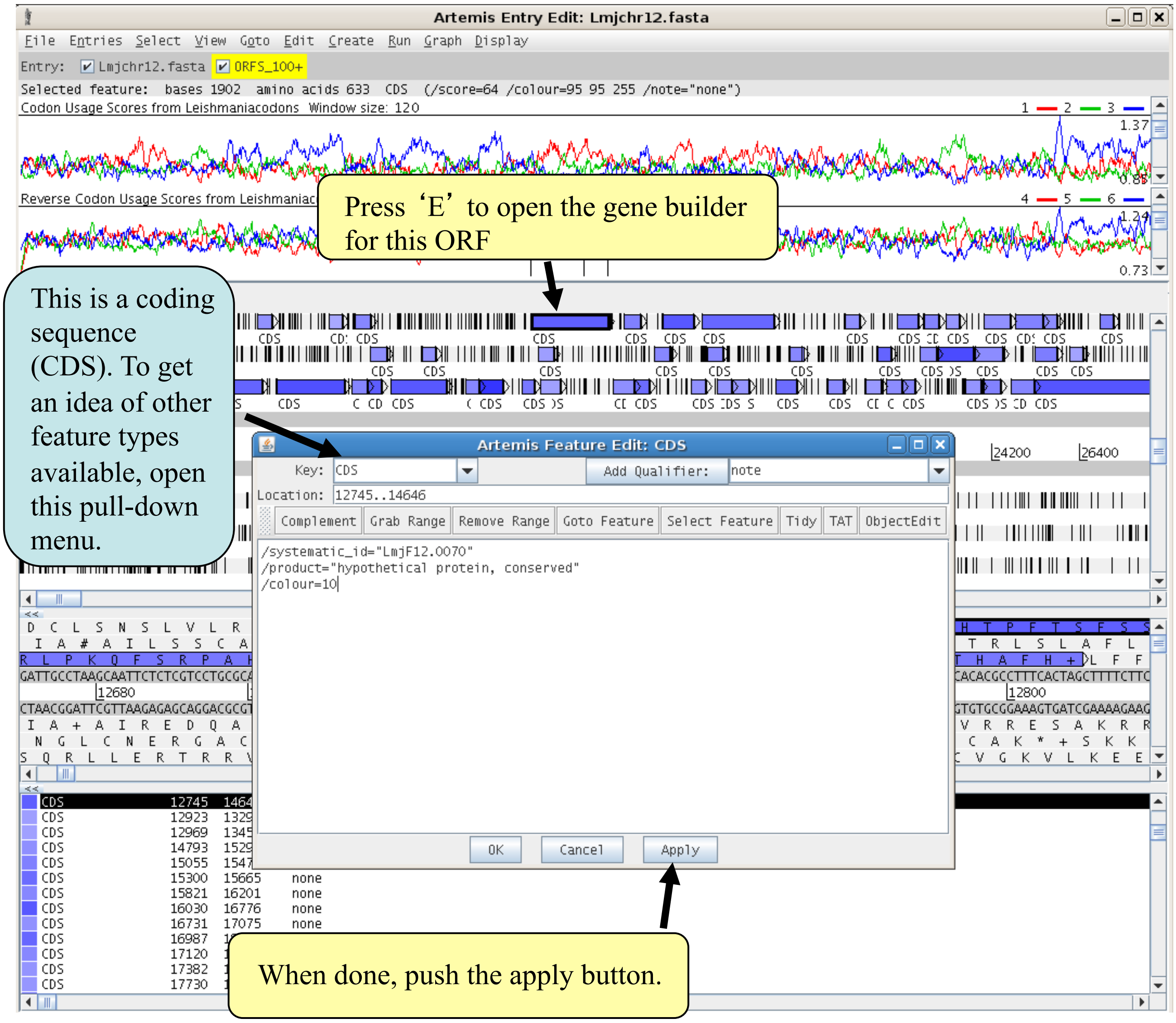
Right-click selected ORF



```
>ref|XP_001681612.1| G hypothetical protein [Leishmania major strain Friedlin]
 emb|CAJ02386.1| G hypothetical protein, conserved [Leishmania major]
Length=620

 GENE ID: 5649892 LmjF12.0070 | hypothetical protein
[Leishmania major strain Friedlin] (10 or fewer PubMed links)

 Score = 1238 bits (3203),  Expect = 0.0, Method: Compositional matrix adjust.
 Identities = 620/620 (100%), Positives = 620/620 (100%), Gaps = 0/620 (0%)

Query  14   MHTHTPFTSFSSFSPPFVSAPLSSPIAHDLAHHGAGLLRQPQVLRMSGEFSLLQHLGVTV  73
            MHTHTPFTSFSSFSPPFVSAPLSSPIAHDLAHHGAGLLRQPQVLRMSGEFSLLQHLGVTV
Sbjct  1    MHTHTPFTSFSSFSPPFVSAPLSSPIAHDLAHHGAGLLRQPQVLRMSGEFSLLQHLGVTV  60

Query  74   DKCDSSADLTPPTHSAKAAFRWWAPTQHPPSKCGTAEETCGVLEKKRKGAHPSYADPTGE  133
            DKCDSSADLTPPTHSAKAAFRWWAPTQHPPSKCGTAEETCGVLEKKRKGAHPSYADPTGE
Sbjct  61   DKCDSSADLTPPTHSAKAAFRWWAPTQHPPSKCGTAEETCGVLEKKRKGAHPSYADPTGE  120

Query  134  VLLRRKNGVEVQYIVYNSKIPSVYGVNKQLAKKMEREREENSPLFKYPLAVGEGEAGQEE  193
            VLLRRKNGVEVQYIVYNSKIPSVYGVNKQLAKKMEREREENSPLFKYPLAVGEGEAGQEE
Sbjct  121  VLLRRKNGVEVQYIVYNSKIPSVYGVNKQLAKKMEREREENSPLFKYPLAVGEGEAGQEE  180

Query  194  ARRVLLQELRRCNQEQARMRKEEGLRERARRLAHERAVVQYKAGETAREAEDARRHKIMD  253
            ARRVLLQELRRCNQEQARMRKEEGLRERARRLAHERAVVQYKAGETAREAEDARRHKIMD
Sbjct  181  ARRVLLQELRRCNQEQARMRKEEGLRERARRLAHERAVVQYKAGETAREAEDARRHKIMD  240

Query  254  GEAVSETAAKALTLREERRADANVDARVVASGWNGKEEDDRRRLAAAERTRQLAEENFRA  313
            GEAVSETAAKALTLREERRADANVDARVVASGWNGKEEDDRRRLAAAERTRQLAEENFRA
Sbjct  241  GEAVSETAAKALTLREERRADANVDARVVASGWNGKEEDDRRRLAAAERTRQLAEENFRA  300

Query  314  AEQRRAERKAQEQEARERARVERMELQRQLAQERVKDLEGHRRNAEALRGAQDSARERRW  373
            AEQRRAERKAQEQEARERARVERMELQRQLAQERVKDLEGHRRNAEALRGAQDSARERRW
Sbjct  301  AEQRRAERKAQEQEARERARVERMELQRQLAQERVKDLEGHRRNAEALRGAQDSARERRW  360

Query  374  RANSADVHLQAMPSWSLFDAVERRQRDEAMRAQQKRMEDTAVNVRLAAQKRADAQAERDR  433
            RANSADVHLQAMPSWSLFDAVERRQRDEAMRAQQKRMEDTAVNVRLAAQKRADAQAERDR
Sbjct  361  RANSADVHLQAMPSWSLFDAVERRQRDEAMRAQQKRMEDTAVNVRLAAQKRADAQAERDR  420

Query  434  DRQYAAEYAKAELENFQREVEHARQRRQQERQELQDAAEATAKVRQAHADAARRREQSVV  493
            DRQYAAEYAKAELENFQREVEHARQRRQQERQELQDAAEATAKVRQAHADAARRREQSVV
Sbjct  421  DRQYAAEYAKAELENFQREVEHARQRRQQERQELQDAAEATAKVRQAHADAARRREQSVV  480

Query  494  PLLFWPAAQSPGAEKAKIDANRRFREDLRRQAEQKRDERAQEEEAERARDRALVEYDTRL  553
            PLLFWPAAQSPGAEKAKIDANRRFREDLRRQAEQKRDERAQEEEAERARDRALVEYDTRL
Sbjct  481  PLLFWPAAQSPGAEKAKIDANRRFREDLRRQAEQKRDERAQEEEAERARDRALVEYDTRL  540

Query  554  AREAVERERKEMREKAEHLRRTLEAQIAEKRKGAVGDRHACAAADVTHVPATEAMRLYRC  613
            AREAVERERKEMREKAEHLRRTLEAQIAEKRKGAVGDRHACAAADVTHVPATEAMRLYRC
Sbjct  541  AREAVERERKEMREKAEHLRRTLEAQIAEKRKGAVGDRHACAAADVTHVPATEAMRLYRC  600

Query  614  PVTGELLPASAYDFGVQRRR  633
            PVTGELLPASAYDFGVQRRR
Sbjct  601  PVTGELLPASAYDFGVQRRR  620
```

Not surprisingly, the top hit is to a gene on chromosome 12 in *L. major,* a hypothetical protein.

Now that we know that this is a real gene we can make a few adjustments. First, open the gene builder window by selecting the ORF and pressing E. This will open a text window where we can add annotations on the gene. Start by deleting the current 'automatic' annotations in this window. Try entering the text in the gene builder shown below to record gene ID, predicted product and a colour code that will distinguish this gene from the automatically generated ORFs.


Artemis Entry Edit: Lmjchr12.fasta
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ Lmjchr12.fasta ☑ ORFS_100+
Codon Usage Scores from Leishmaniacodons Window size: 120
Reverse Codon Usage Scores from Leishmaniac
CDS
24200
26400
12680
12800
Artemis Feature Edit: CDS
Key: CDS
Add Qualifier: note
Complement Grab Range Remove Range Goto Feature Select Feature Tidy TAT ObjectEdit
OK Cancel Apply

Selected feature: bases 1902 amino acids 633 CDS (/score=64 /colour=95 95 255 /note="none")

Location: 12745..14646

```
/systematic_id="LmjF12.0070"
/product="hypothetical protein, conserved"
/colour=10
```

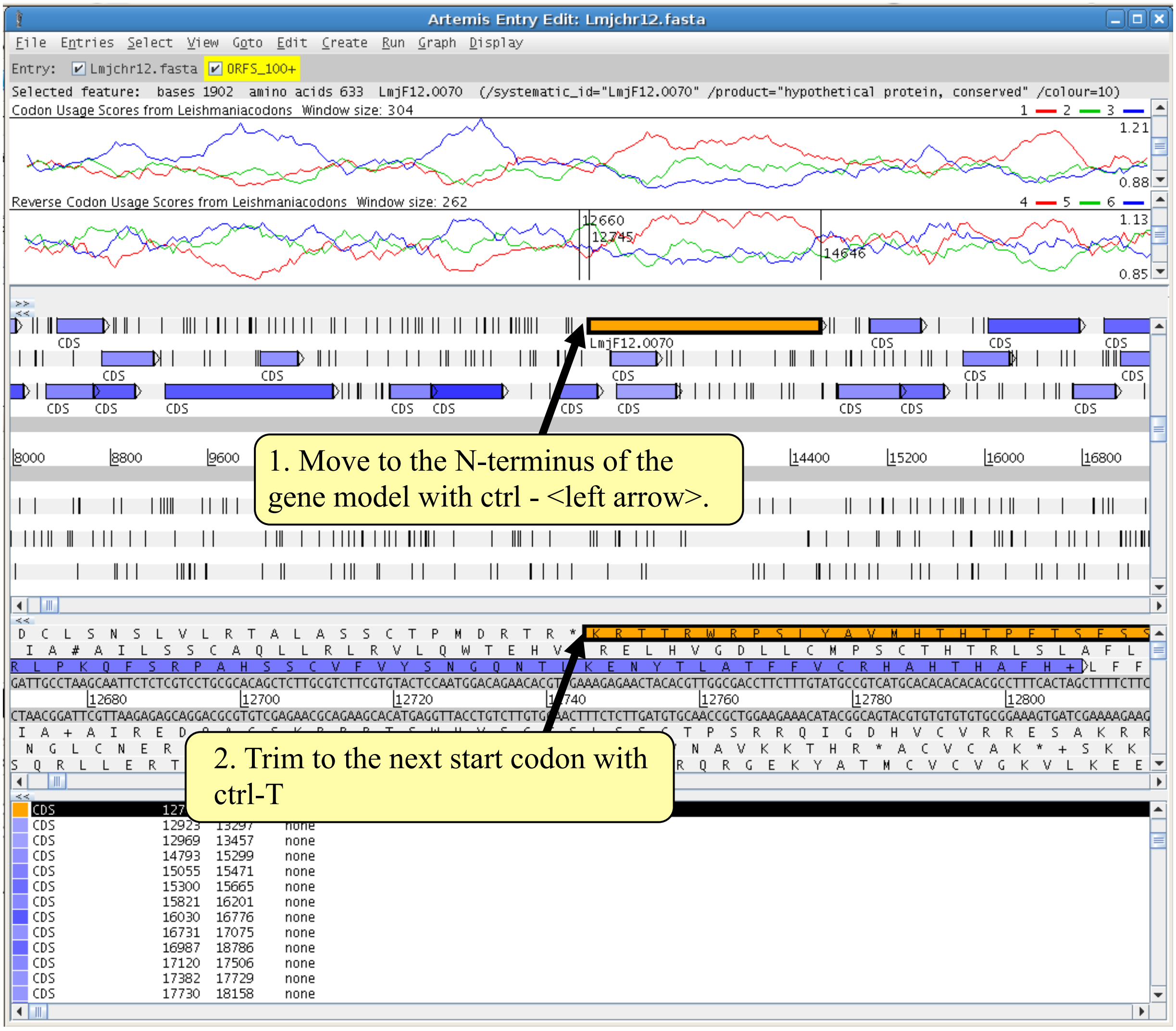
Press 'E' to open the gene builder for this ORF

This is a coding sequence (CDS). To get an idea of other feature types available, open this pull-down menu.

When done, push the apply button.

Based on the NCBI blast results we can adjust the N-terminus of this model to the correct start codon. To automatically position the sequence window at the N-terminus of the gene model push ctrl-<left arrow>.

Go to Edit>Trim Selected Feature>To Next Met (or ctrl-T), then reposition the sequence window at the new start as described above. Continue until the start resembles the NCBI blast results. If trimmed passed the desired start codon the model can be reset through Edit>Extend Selected Feature>To Previous Stop Codon, or ctrl-Q.


Artemis Entry Edit: Lmjchr12.fasta
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ Lmjchr12.fasta ☑ ORFS_100+
Codon Usage Scores from Leishmaniacodons Window size: 304
Reverse Codon Usage Scores from Leishmaniacodons Window size: 262
LmjF12.0070
CDS

```
Selected feature:  bases 1902  amino acids 633  LmjF12.0070  (/systematic_id="LmjF12.0070" /product="hypothetical protein, conserved" /colour=10)
```

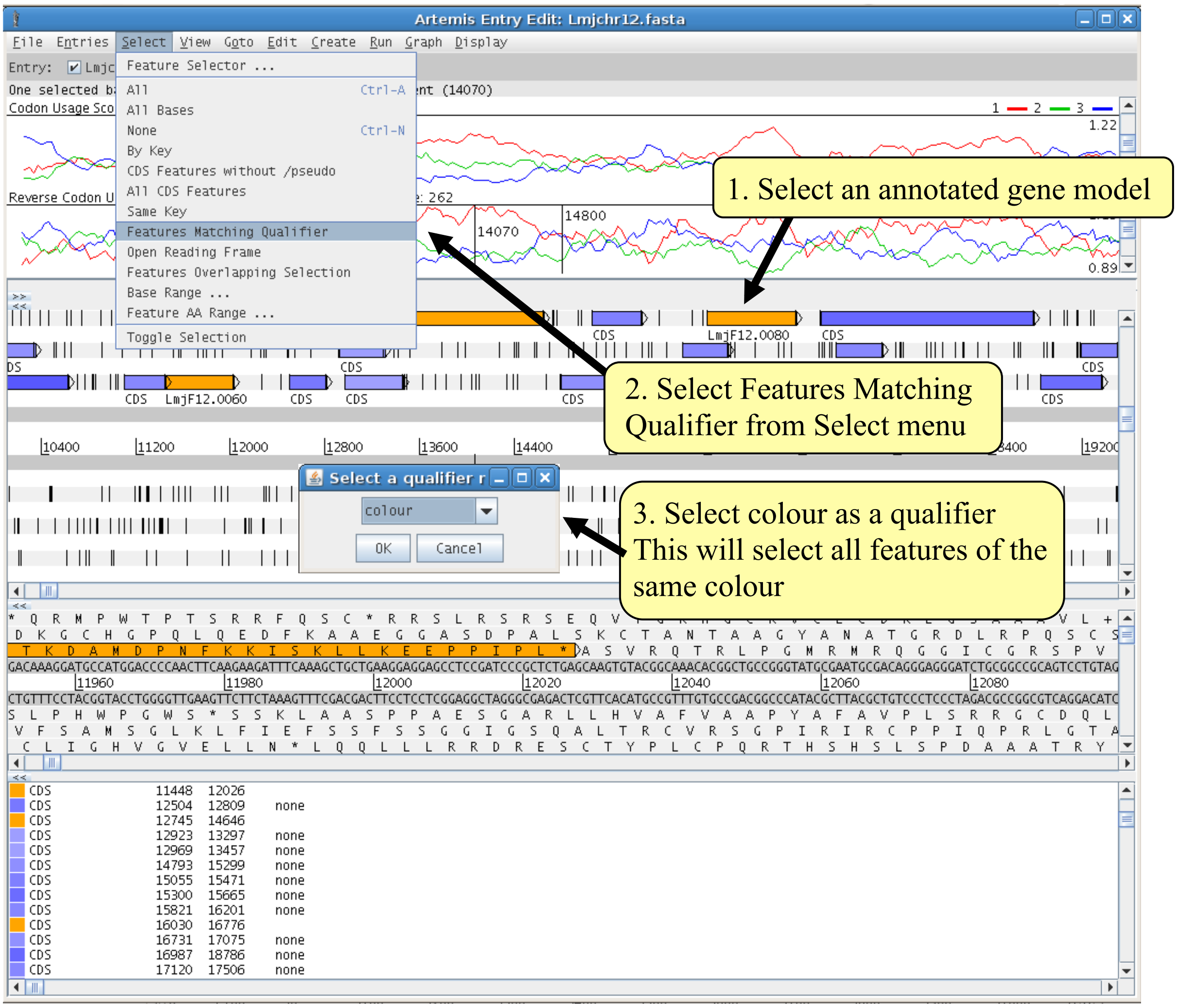
```
CDS    127
CDS    12923  13297   none
CDS    12969  13457   none
CDS    14793  15299   none
CDS    15055  15471   none
CDS    15300  15665   none
CDS    15821  16201   none
CDS    16030  16776   none
CDS    16731  17075   none
CDS    16987  18786   none
CDS    17120  17506   none
CDS    17382  17729   none
CDS    17730  18158   none
```

1. Move to the N-terminus of the gene model with ctrl - <left arrow>.

2. Trim to the next start codon with ctrl-T

There are more than 20 protein coding genes in the first 100 kbs of chromosome 12. See how many of these you can find by repeating the steps in the past slides.

IMPORTANT!! Any changes made to the predicted ORFs will be written to an entry file called ORFS_100+. When you're done with gene predictions follow the steps below to save these entries to the sequence file instead. Make sure all of the annotated features have a /colour=10 in their gene builder window.

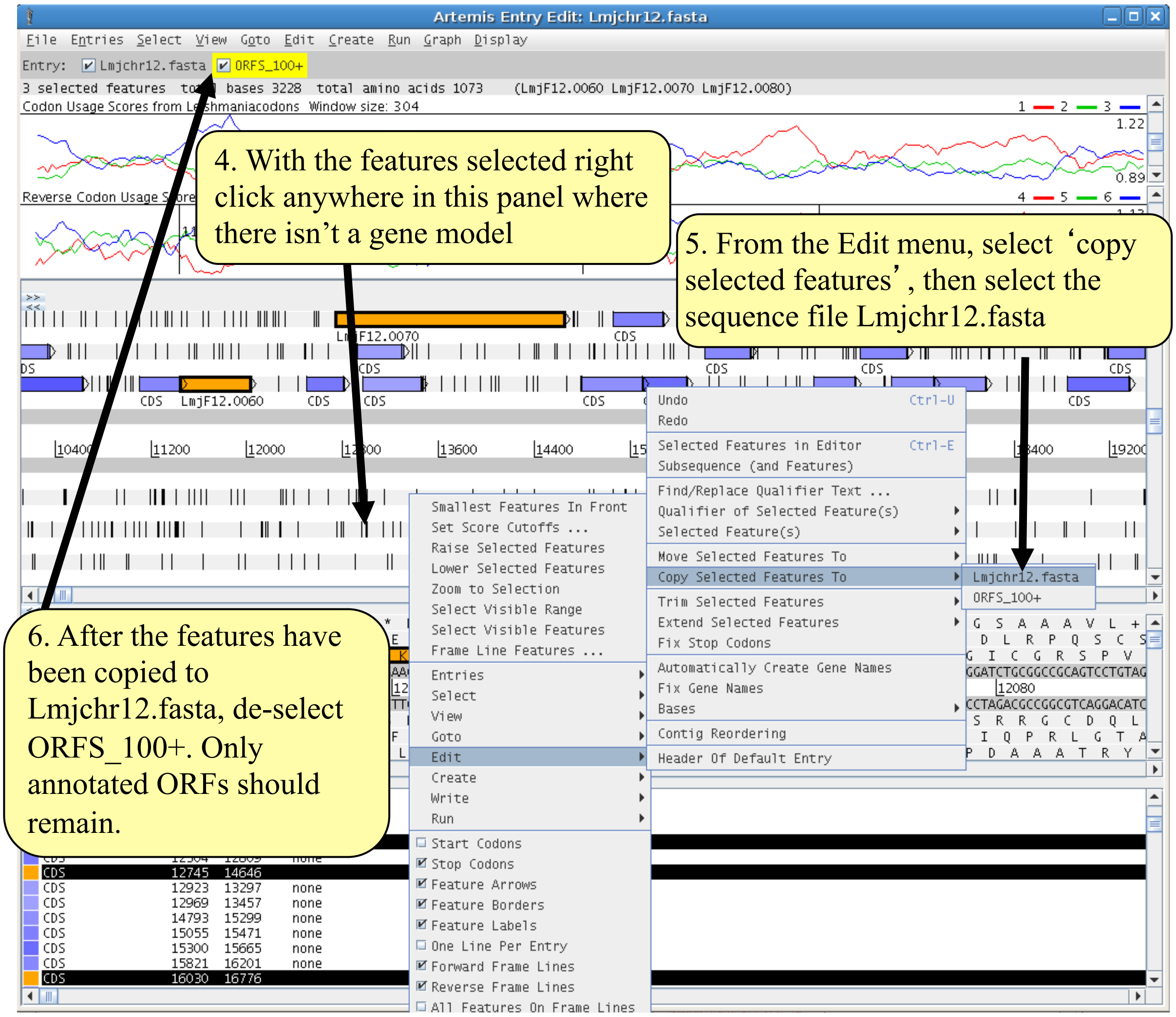


1. Select an annotated gene model

2. Select Features Matching Qualifier from Select menu

3. Select colour as a qualifier This will select all features of the same colour

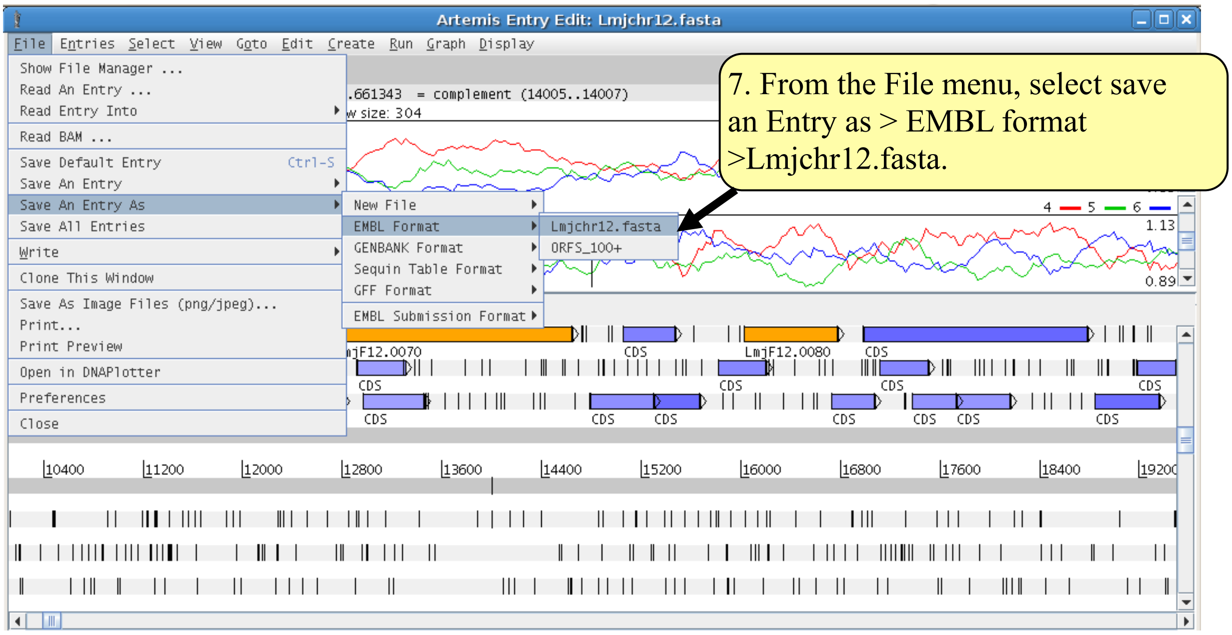
```
CDS   11448  12026
CDS   12504  12809   none
CDS   12745  14646
CDS   12923  13297   none
CDS   12969  13457   none
CDS   14793  15299   none
CDS   15055  15471   none
CDS   15300  15665   none
CDS   15821  16201   none
CDS   16030  16776
CDS   16731  17075   none
CDS   16987  18786   none
CDS   17120  17506   none
```


Artemis Entry Edit: Lmjchr12.fasta
File Entries Select View Goto Edit Create Run Graph Display
Entry: ☑ Lmjchr12.fasta ☑ ORFS_100+
3 selected features total bases 3228 total amino acids 1073 (LmjF12.0060 LmjF12.0070 LmjF12.0080)
Codon Usage Scores from Leishmaniacodons Window size: 304
Reverse Codon Usage Score
LmjF12.0070
LmjF12.0060
CDS
Smallest Features In Front
Set Score Cutoffs ...
Raise Selected Features
Lower Selected Features
Zoom to Selection
Select Visible Range
Select Visible Features
Frame Line Features ...
Entries
Select
View
Goto
Edit
Create
Write
Run
☐ Start Codons
☑ Stop Codons
☑ Feature Arrows
☑ Feature Borders
☑ Feature Labels
☐ One Line Per Entry
☑ Forward Frame Lines
☑ Reverse Frame Lines
☐ All Features On Frame Lines
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
Lmjchr12.fasta
ORFS_100+
Trim Selected Features
Extend Selected Features
Fix Stop Codons
Automatically Create Gene Names
Fix Gene Names
Bases
Contig Reordering
Header Of Default Entry

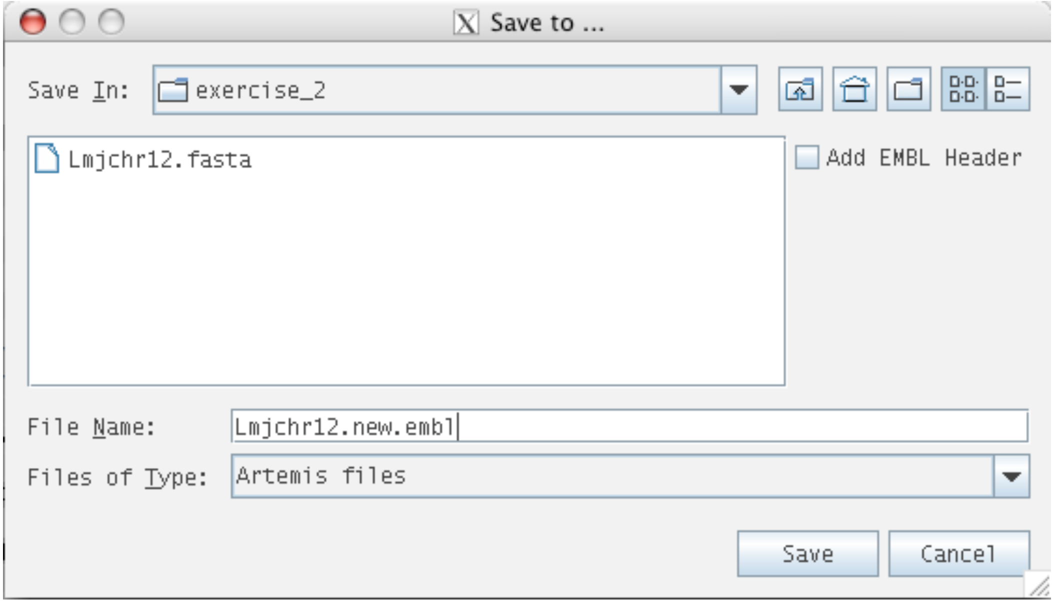
4. With the features selected right click anywhere in this panel where there isn't a gene model

5. From the Edit menu, select 'copy selected features', then select the sequence file Lmjchr12.fasta

6. After the features have been copied to Lmjchr12.fasta, de-select ORFS_100+. Only annotated ORFs should remain.



7. From the File menu, select save an Entry as > EMBL format >Lmjchr12.fasta.



Save In: exercise_2

Lmjchr12.fasta

□ Add EMBL Header

File Name: Lmjchr12.new.embl

Files of Type: Artemis files

Save the sequence file as Lmjchr12.new.embl.

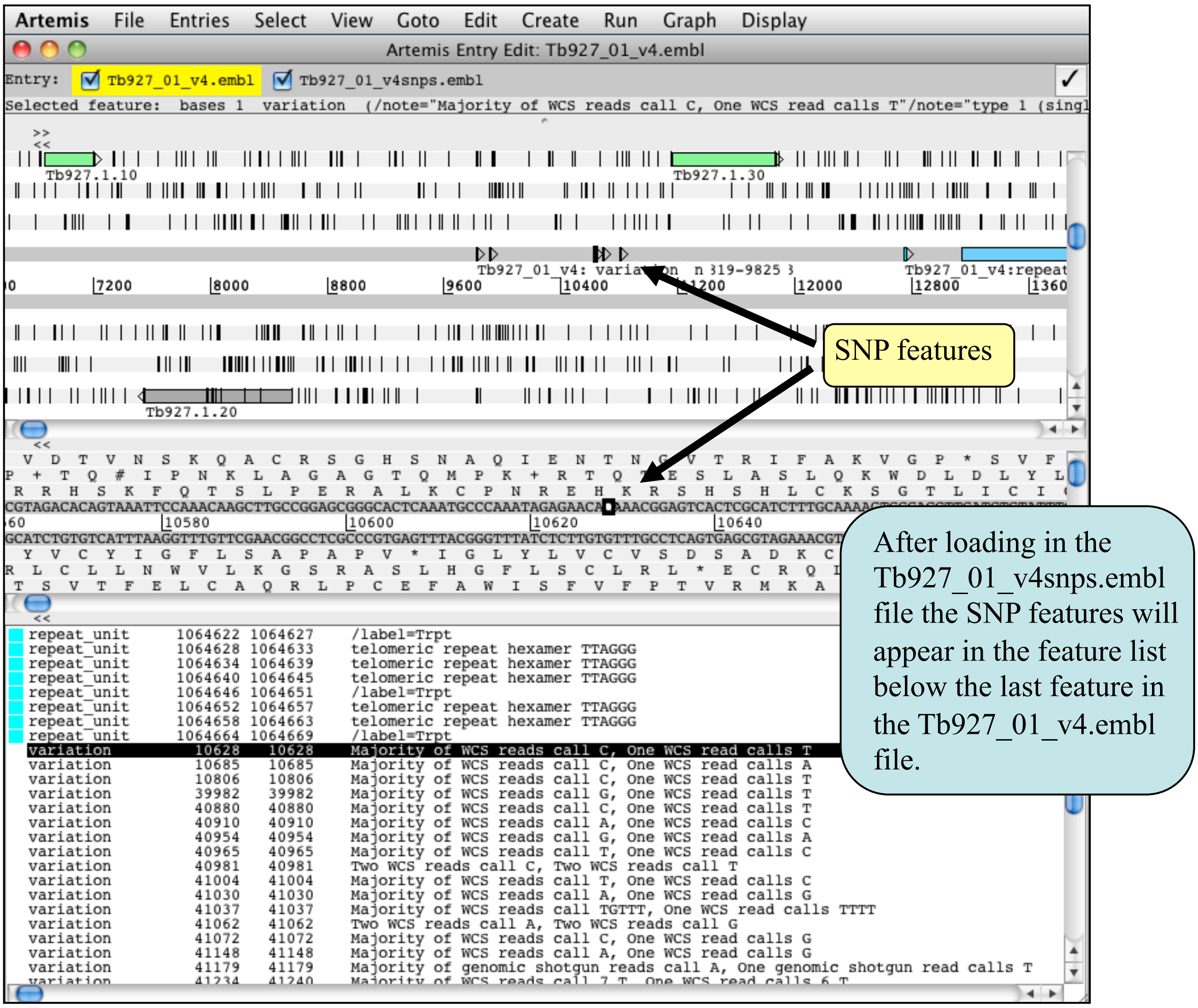
Optional exercise

This optional exercise demonstrates how to use Artemis to construct queries for features within features. In the exercise below we will load a file containing SNP data, then retrieve a list of all CDS' that overlap with SNP features.

Files required:

Tb927_01_v4.embl - Contains sequence and annotation for *T. brucei* chromosome 1
Tb927_01_v4snps.embl - Contains SNP features for *T. brucei* chromosome 1

Use the file manager to open Tb927_01_v4.embl, then as shown in previous exercises select File>Read Entry >Tb927_01_v4snps.embl

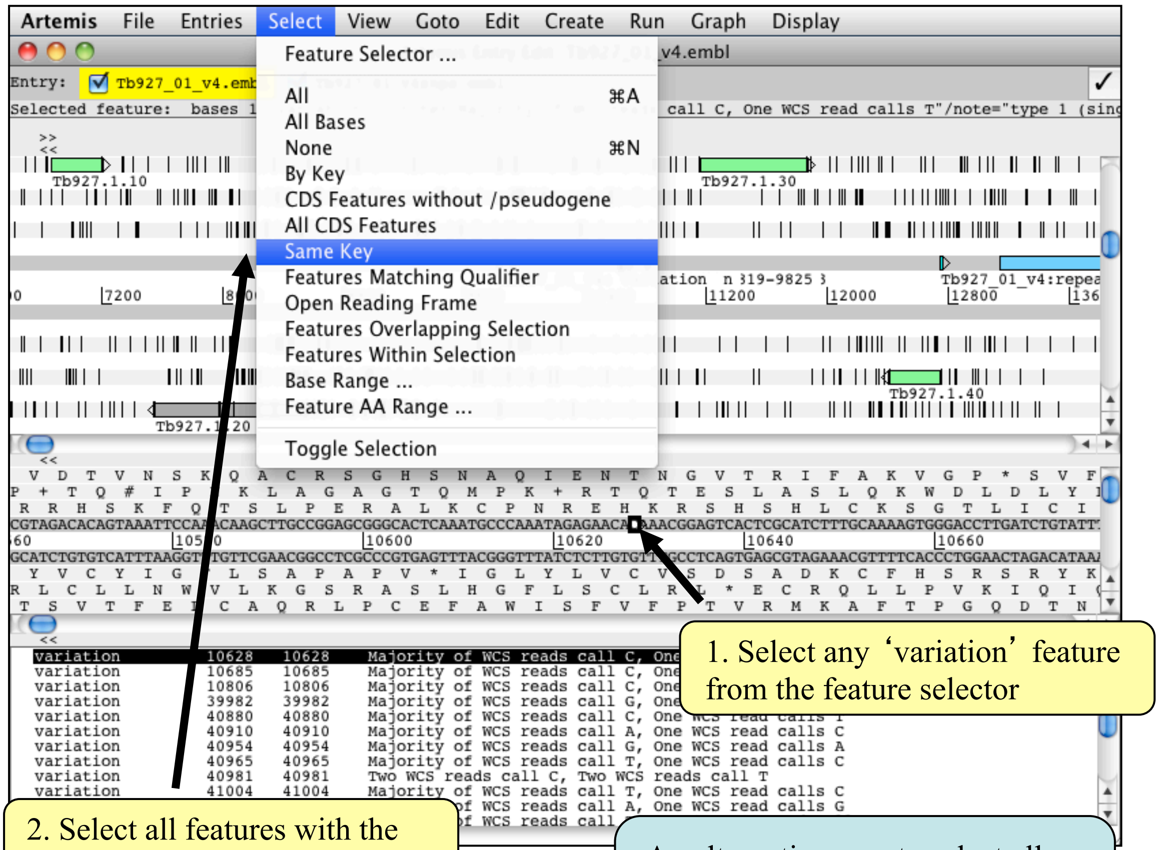


```
Entry:  ☑ Tb927_01_v4.embl  ☑ Tb927_01_v4snps.embl
Selected feature:  bases 1  variation  (/note="Majority of WCS reads call C, One WCS read calls T"/note="type 1 (singl
```

```
repeat_unit     1064622 1064627    /label=Trpt
repeat_unit     1064628 1064633    telomeric repeat hexamer TTAGGG
repeat_unit     1064634 1064639    telomeric repeat hexamer TTAGGG
repeat_unit     1064640 1064645    telomeric repeat hexamer TTAGGG
repeat_unit     1064646 1064651    /label=Trpt
repeat_unit     1064652 1064657    telomeric repeat hexamer TTAGGG
repeat_unit     1064658 1064663    telomeric repeat hexamer TTAGGG
repeat_unit     1064664 1064669    /label=Trpt
variation         10628   10628    Majority of WCS reads call C, One WCS read calls T
variation         10685   10685    Majority of WCS reads call C, One WCS read calls A
variation         10806   10806    Majority of WCS reads call C, One WCS read calls T
variation         39982   39982    Majority of WCS reads call G, One WCS read calls T
variation         40880   40880    Majority of WCS reads call C, One WCS read calls T
variation         40910   40910    Majority of WCS reads call A, One WCS read calls C
variation         40954   40954    Majority of WCS reads call G, One WCS read calls A
variation         40965   40965    Majority of WCS reads call T, One WCS read calls C
variation         40981   40981    Two WCS reads call C, Two WCS reads call T
variation         41004   41004    Majority of WCS reads call T, One WCS read calls C
variation         41030   41030    Majority of WCS reads call A, One WCS read calls G
variation         41037   41037    Majority of WCS reads call TGTTT, One WCS read calls TTTT
variation         41062   41062    Two WCS reads call A, Two WCS reads call G
variation         41072   41072    Majority of WCS reads call C, One WCS read calls G
variation         41148   41148    Majority of WCS reads call A, One WCS read calls G
variation         41179   41179    Majority of genomic shotgun reads call A, One genomic shotgun read calls T
variation         41234   41240    Majority of WCS reads call 7 T, One WCS read calls 6 T
```

SNP features

After loading in the Tb927_01_v4snps.embl file the SNP features will appear in the feature list below the last feature in the Tb927_01_v4.embl file.

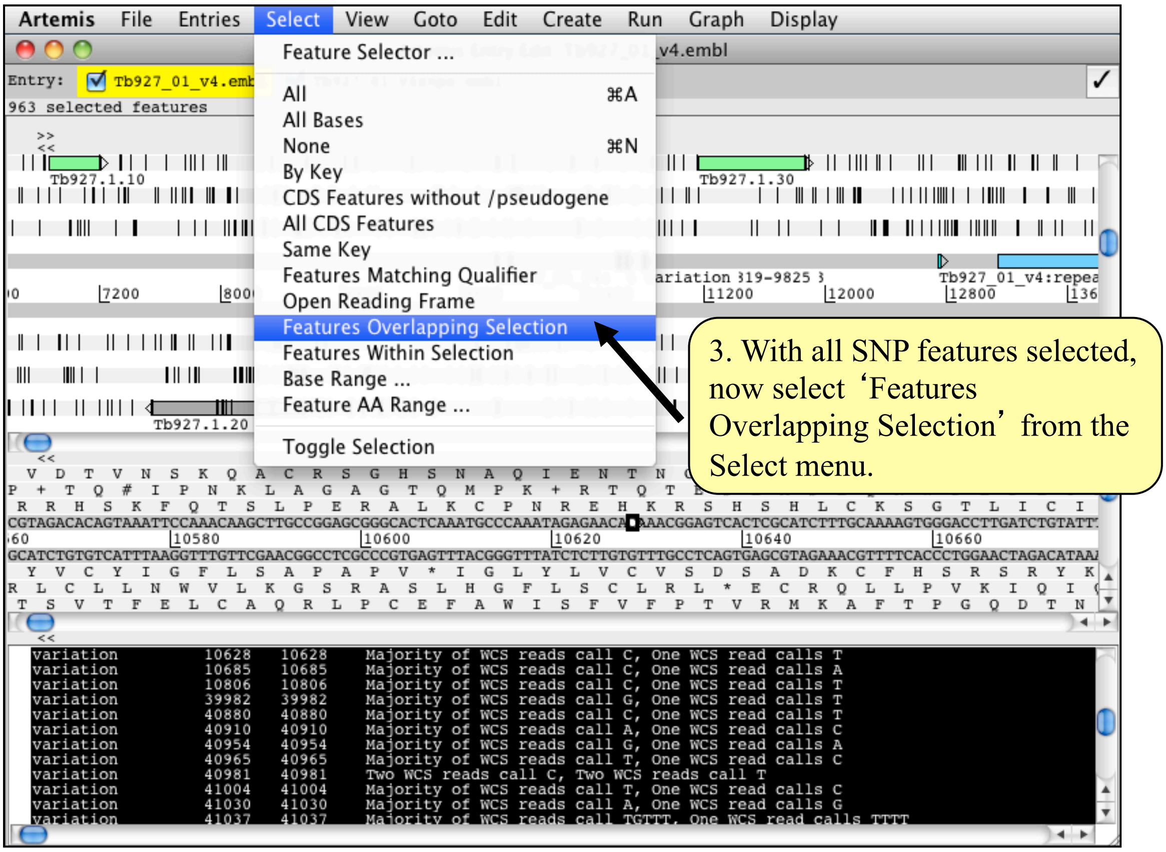


```
variation    10628   10628   Majority of WCS reads call C, One
variation    10685   10685   Majority of WCS reads call C, One
variation    10806   10806   Majority of WCS reads call C, One
variation    39982   39982   Majority of WCS reads call G, One
variation    40880   40880   Majority of WCS reads call C, One
variation    40910   40910   Majority of WCS reads call A, One WCS read calls C
variation    40954   40954   Majority of WCS reads call G, One WCS read calls A
variation    40965   40965   Majority of WCS reads call T, One WCS read calls C
variation    40981   40981   Two WCS reads call C, Two WCS reads call T
variation    41004   41004   Majority of WCS reads call T, One WCS read calls C
                                      of WCS reads call A, One WCS read calls G
```

1. Select any ‘variation’ feature from the feature selector

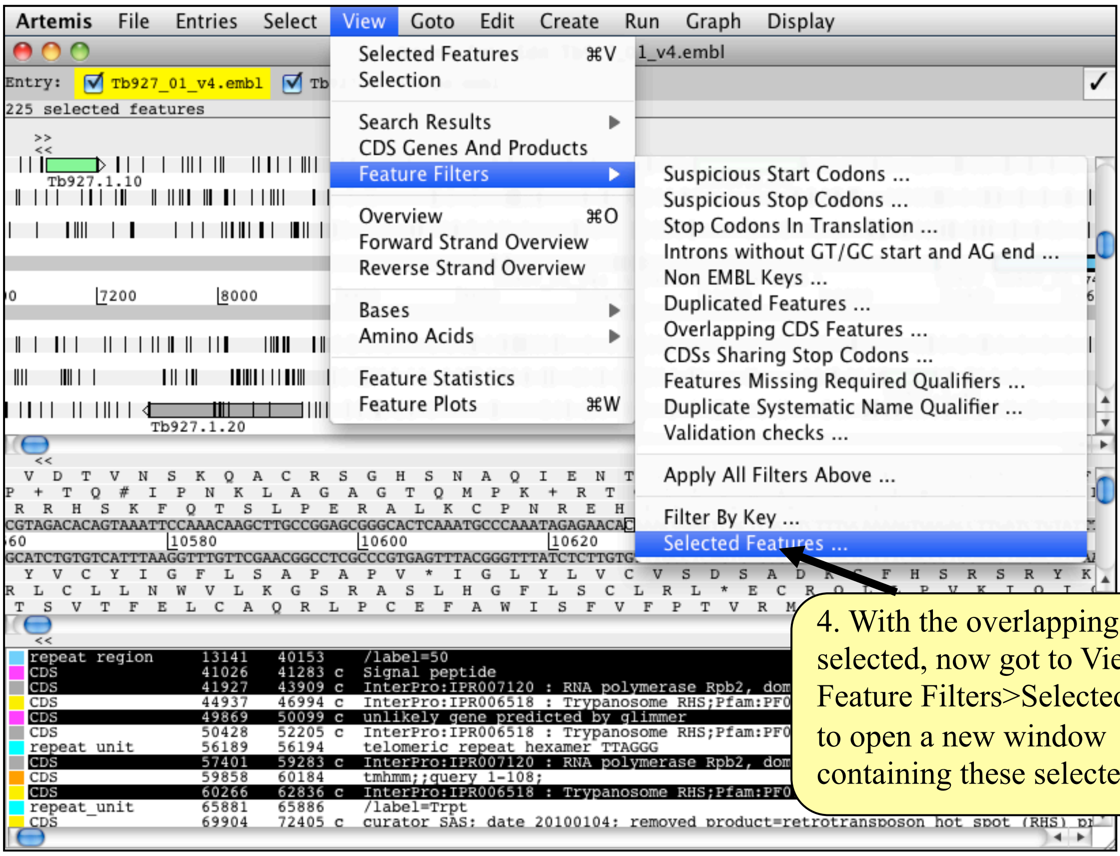
2. Select all features with the same key (variation)

An alternative way to select all SNP features is to select ‘By Key’, then select ‘variation’.



```
variation    10628   10628   Majority of WCS reads call C, One WCS read calls T
variation    10685   10685   Majority of WCS reads call C, One WCS read calls A
variation    10806   10806   Majority of WCS reads call C, One WCS read calls T
variation    39982   39982   Majority of WCS reads call G, One WCS read calls T
variation    40880   40880   Majority of WCS reads call C, One WCS read calls T
variation    40910   40910   Majority of WCS reads call A, One WCS read calls C
variation    40954   40954   Majority of WCS reads call G, One WCS read calls A
variation    40965   40965   Majority of WCS reads call T, One WCS read calls C
variation    40981   40981   Two WCS reads call C, Two WCS reads call T
variation    41004   41004   Majority of WCS reads call T, One WCS read calls C
variation    41030   41030   Majority of WCS reads call A, One WCS read calls G
variation    41037   41037   Majority of WCS reads call TGTTT, One WCS read calls TTTT
```

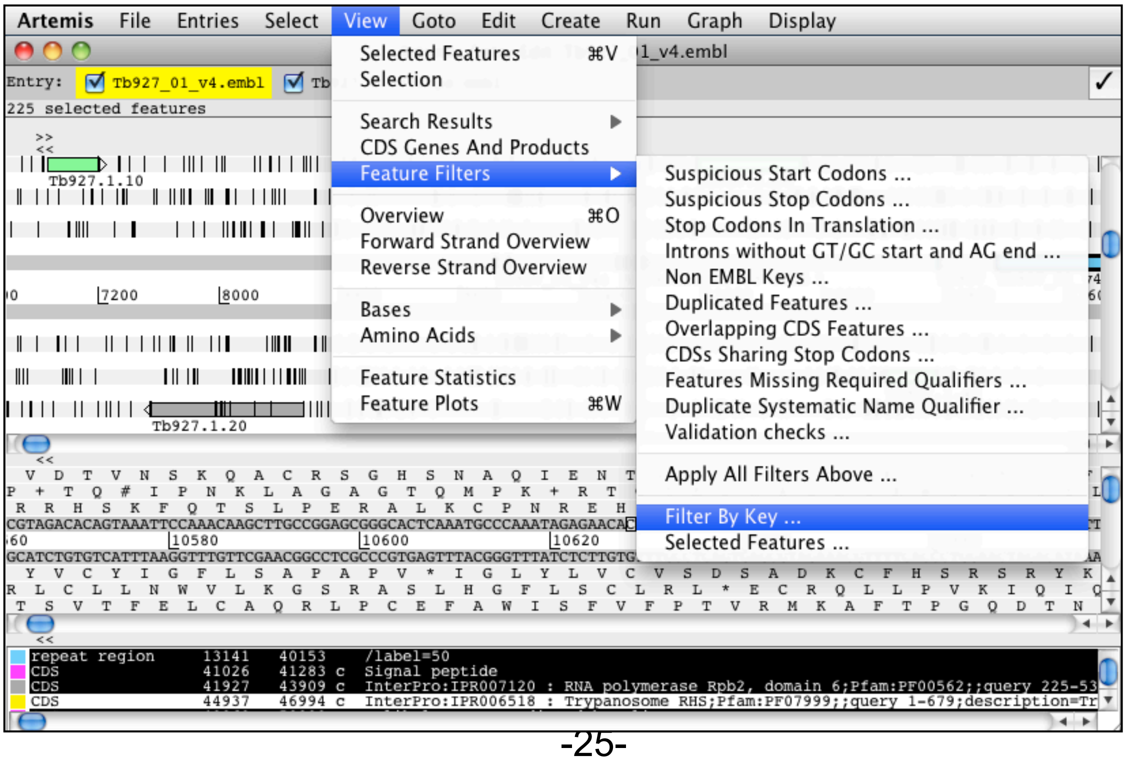
3. With all SNP features selected, now select ‘Features Overlapping Selection’ from the Select menu.



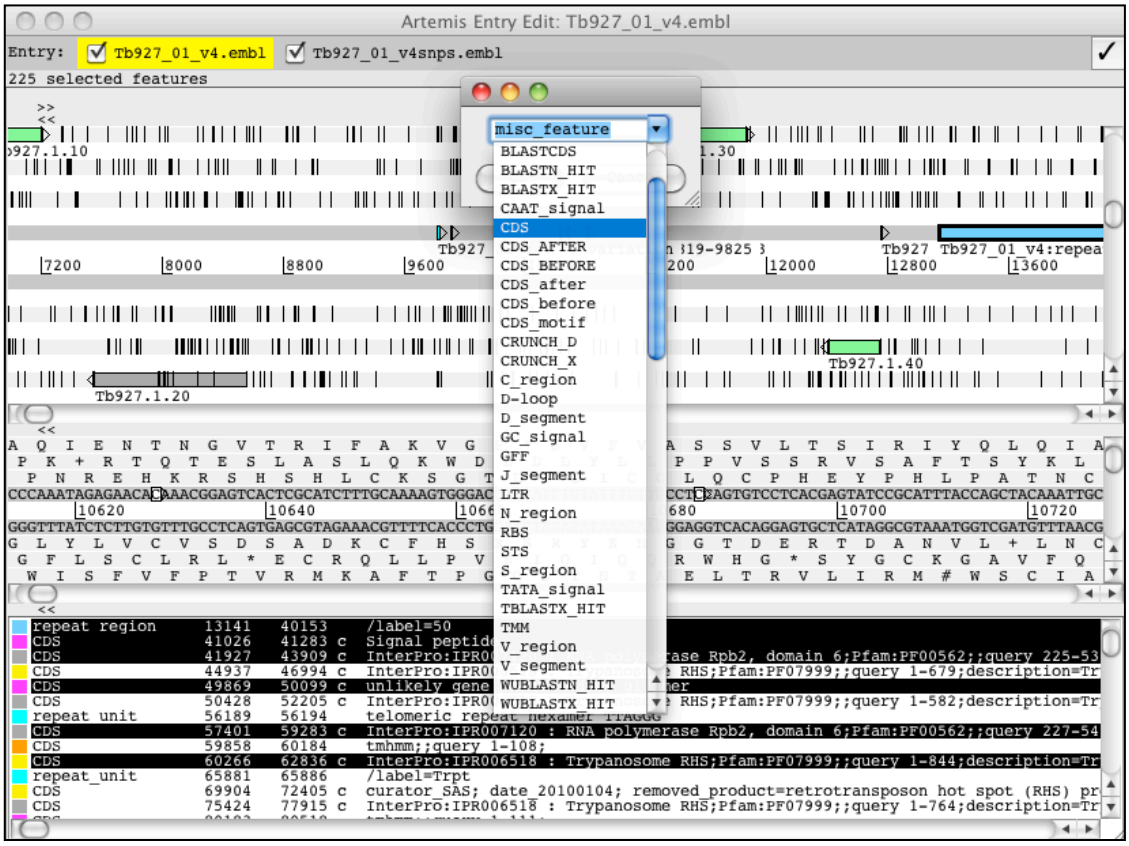
```
repeat region  13141   40153    /label=50
CDS            41026   41283 c  Signal peptide
CDS            41927   43909 c  InterPro:IPR007120 : RNA polymerase Rpb2, dom
CDS            44937   46994 c  InterPro:IPR006518 : Trypanosome RHS;Pfam:PF0
CDS            49869   50099 c  unlikely gene predicted by glimmer
CDS            50428   52205 c  InterPro:IPR006518 : Trypanosome RHS;Pfam:PF0
repeat unit    56189   56194    telomeric repeat hexamer TTAGGG
CDS            57401   59283 c  InterPro:IPR007120 : RNA polymerase Rpb2, dom
CDS            59858   60184    tmhmm;;query 1-108;
CDS            60266   62836 c  InterPro:IPR006518 : Trypanosome RHS;Pfam:PF0
repeat_unit    65881   65886    /label=Trpt
CDS            69904   72405 c  curator SAS: date 20100104: removed product=retrotransposon hot spot (RHS) pr
```

4. With the overlapping features selected, now got to View select Feature Filters>Selected Features to open a new window containing these selected results.

All the features overlapping with the SNP features should now appear in a new window. Note that this window contains not only CDS features, but features such as 5' UTRs, repeat regions and other miscellaneous features that overlap with SNPs. To see only CDS features we need to apply a second filter. With the non-overlapping features still selected, select View>Feature Filters>Filter by Key. Select CDS for the Key, and only CDS' containing SNPs should appear in the filter window.



```
repeat region  13141   40153    /label=50
CDS            41026   41283 c  Signal peptide
CDS            41927   43909 c  InterPro:IPR007120 : RNA polymerase Rpb2, domain 6;Pfam:PF00562;;query 225-53
CDS            44937   46994 c  InterPro:IPR006518 : Trypanosome RHS;Pfam:PF07999;;query 1-679;description=Tr
```



```
repeat region   13141   40153     /label=50
CDS             41026   41283 c   Signal peptid
CDS             41927   43909 c   InterPro:IPR0                   ase Rpb2, domain 6;Pfam:PF00562;;query 225-53
CDS             44937   46994 c   InterPro:IPR0                   RHS;Pfam:PF07999;;query 1-679;description=Tr
CDS             49869   50099 c   unlikely gene                   her
CDS             50428   52205 c   InterPro:IPR0                   RHS;Pfam:PF07999;;query 1-582;description=Tr
repeat unit     56189   56194     telomeric repeat hexamer TTAGGG
CDS             57401   59283 c   InterPro:IPR007120 : RNA polymerase Rpb2, domain 6;Pfam:PF00562;;query 227-54
CDS             59858   60184     tmhmm;;query 1-108;
CDS             60266   62836 c   InterPro:IPR006518 : Trypanosome RHS;Pfam:PF07999;;query 1-844;description=Tr
repeat_unit     65881   65886     /label=Trpt
CDS             69904   72405 c   curator_SAS; date_20100104; removed_product=retrotransposon hot spot (RHS) pr
CDS             75424   77915 c   InterPro:IPR006518 : Trypanosome RHS;Pfam:PF07999;;query 1-764;description=Tr
```

Other Queries to Try:

1. Try performing the ‘reverse’ query, select all SNPs that overlap with CDS features.
2. Save a list of features to a file by right clicking on the feature filter window and Selecting ‘Save List to File’
3. Use the Select Menu to select all features with the same ‘key’
4. Use the Filter menu to look for suspicious gene models - missing start codons, missing stop codons, stop codons in translation and duplicated features.
5. Search for a qualifier value (try ‘hypothetical protein’), in the Edit menu, select ‘Find/Replace Qualifier Text’. Try doing a boolean search in the same way (try ‘hypothetical AND conserved, or ‘hypothetical AND unlikely).
6. Using the same option, find features containing duplicate qualifiers (more than one qualifier with the same value)