

# Artemis

**Learning objectives:**

- Loading files into Artemis
- Search options in Artemis
- Selecting and extracting sequences and features in Artemis
- Optional exercise: Structural and functional annotation in Artemis

## Introduction

Artemis is a free DNA viewer and annotation tool written by Kim Rutherford (Rutherford *et al.*, 2000). It has been routinely used by the Parasite Genomics Group at the Wellcome Sanger Institute 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, 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

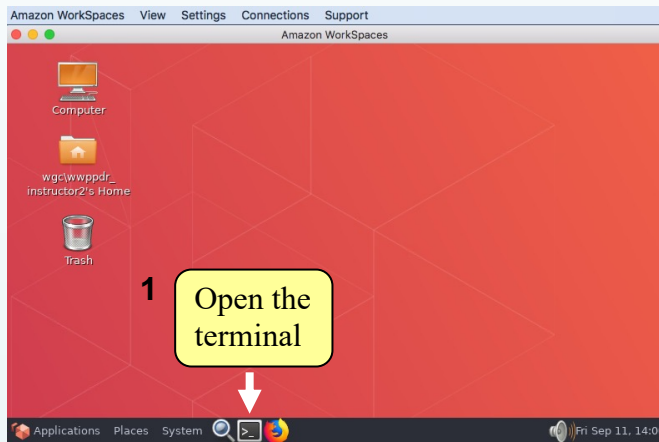
The aim of this Module is for you to become familiar with the basic functions of Artemis using a series of worked examples. These examples are designed to take you through the most immediately useful functions. However, there will be time, and encouragement, for you to explore other menus; nooks and 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

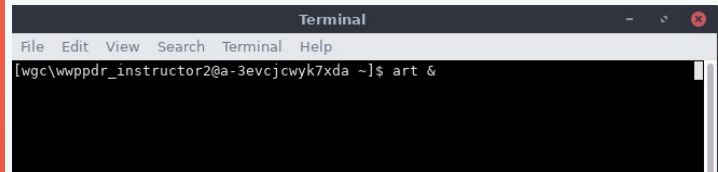
## 1. Starting up the Artemis software

Open the terminal on workspaces and type “art” then hit return.

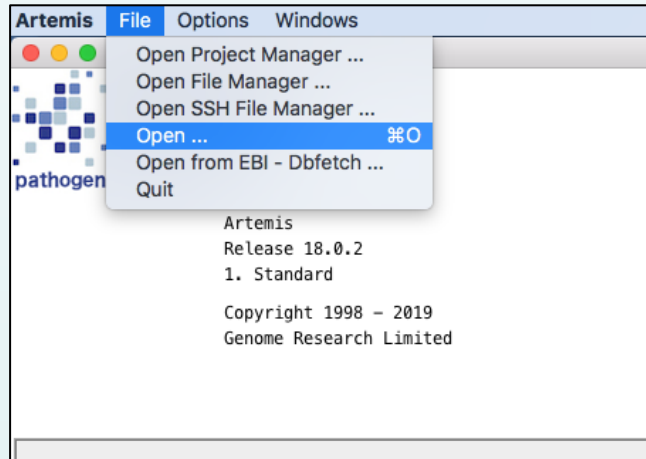
Navigate to the directory Module\_1\_Artemis, exercise\_1 containing the file Pf3D7\_03.fasta.



2 Type “art &” in the terminal

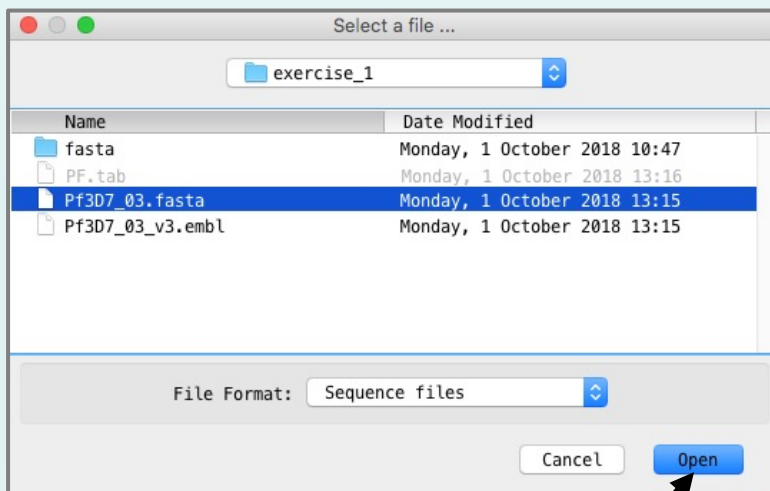


3 Click  
4 ‘File’ then  
‘Open’



5

Single click  
to select the  
DNA file

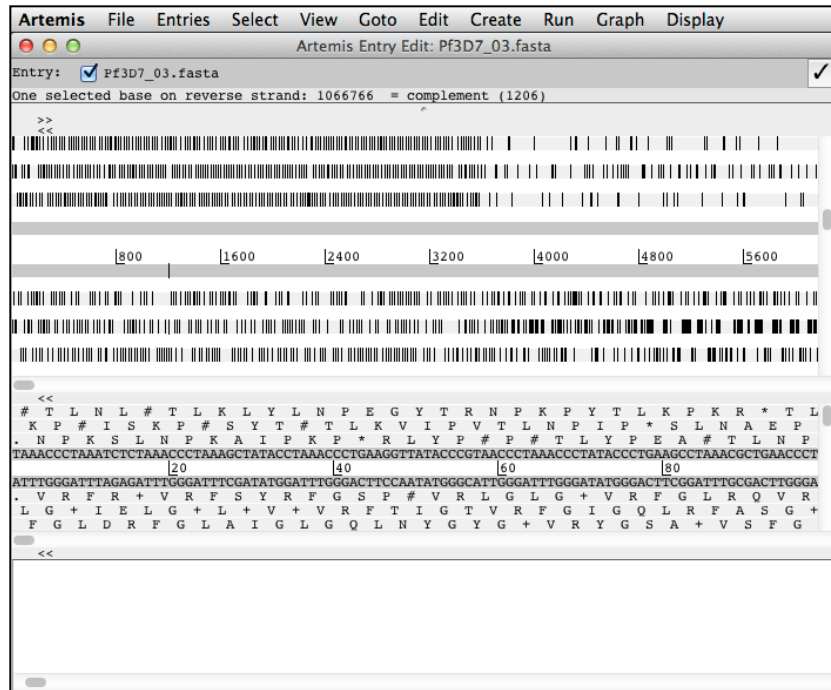


6 Single click to open file in Artemis then wait

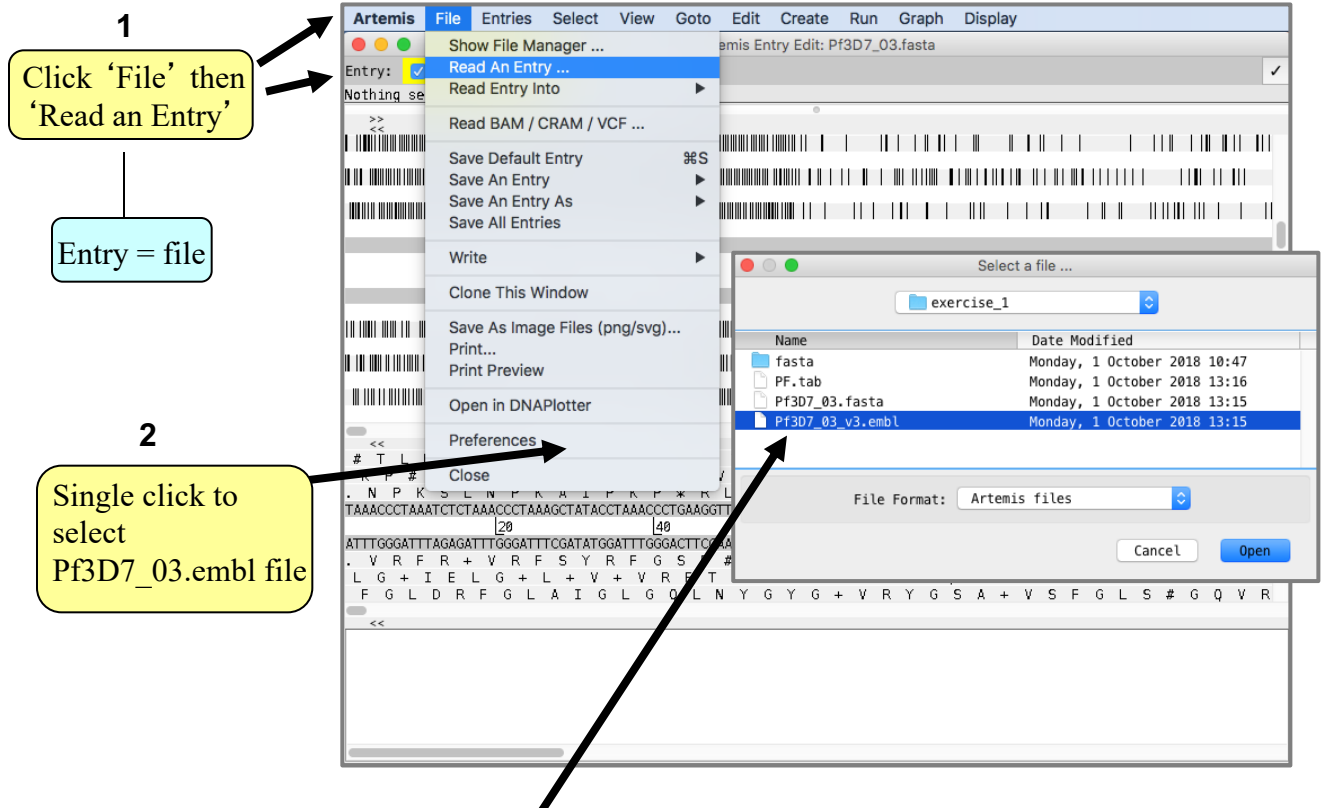
DNA sequence files will have the suffix ‘.fasta’. Annotation files end with ‘.embl’, or ‘.tab’. Use this feature to select the type of file displayed in this panel.

## 2. Loading annotation files (entries) into Artemis

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



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



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.



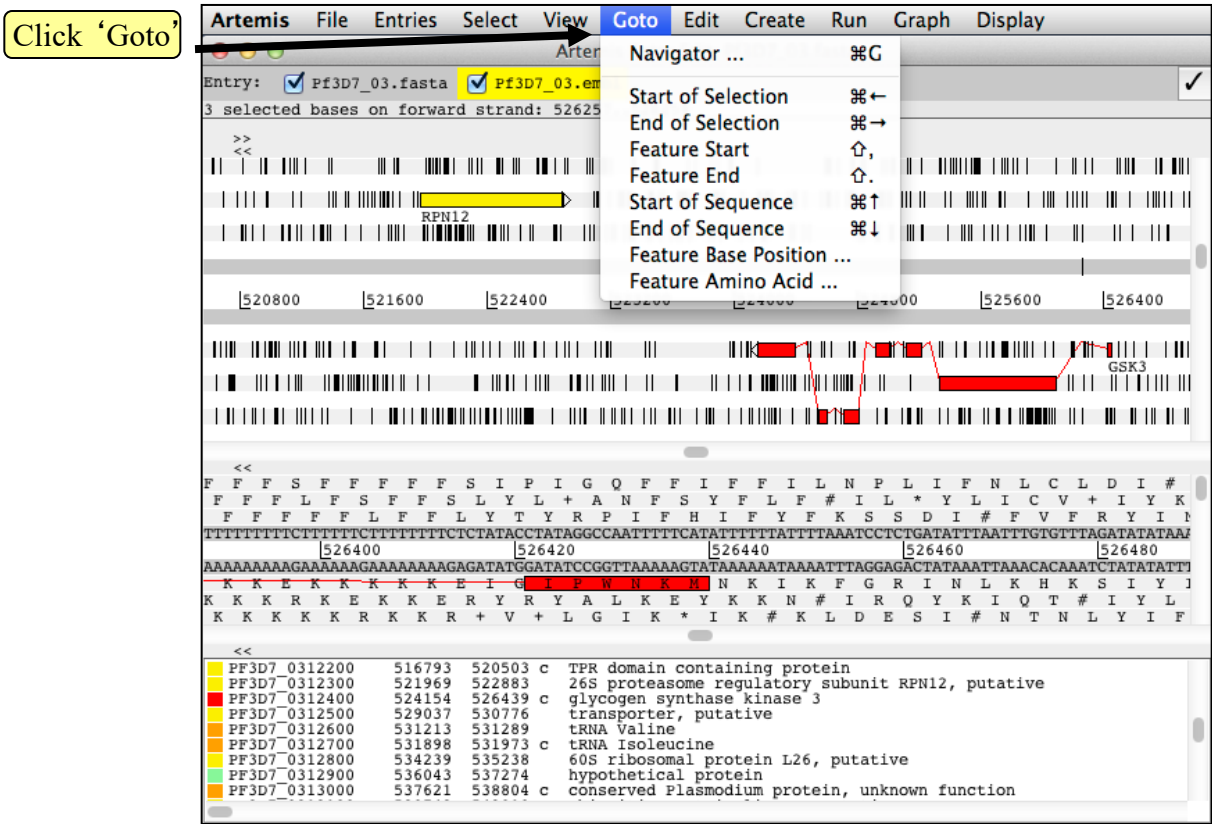
1. Drop-down menus. There's lots in there so don't worry about them right now.
2. Shows what entries are currently loaded (bottom line) and gives details regarding the feature selected in the window below; in this case an acyl-CoA synthetase (selected line).
3. This is the main sequence view panel. The 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.

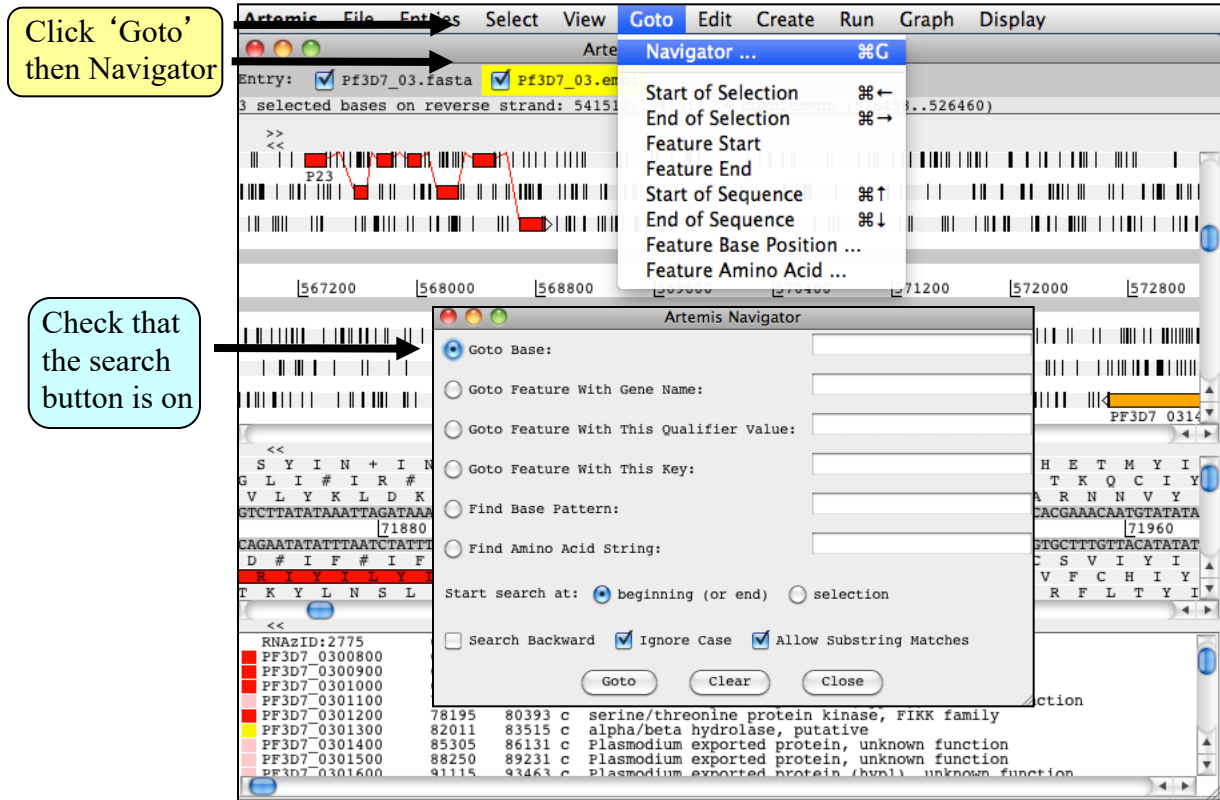


It may seem that ‘Goto’ ‘Start of Selection’ and ‘Goto’ ‘Feature Start’ do the same thing. Well they do if you have a feature selected but ‘Goto’ ‘Start of Selection’ will also work for a region which you have highlighted by click-dragging in the main window. So yes, give it a try! This is a very commonly used feature, so it is worth memorizing the keyboard shortcuts 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.



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**

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.

**Artemis** File Entries Select View Goto Edit Create Run Graph Display  
Artemis Entry Edit: Pf3D7\_03.fasta

Entry: ☒ Pf3D7\_03.fasta ☒ Pf3D7\_03.embl ☒ PF.tab

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

>>

Diagram showing genomic features (CLAG3.1) and protein domains (misc\_feature, misc\_feature ure) across a sequence range from 133600 to 140800.

<<

Sequence alignment view showing multiple sequence alignments (T Y N M H I Y I N I N 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 ...).

<<

Table listing gene annotations and their coordinates:

Gene Name	Start	End	Description
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	CGSH iron-sulfur domain-containing protein, putative
Pf3D7_0302800	148046	149305	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 1, putative
Pf3D7_0303400	178420	181127	palmityl transferase
Pf3D7_0303500	182731	189411	spindle pole body protein, putative
Pf3D7_0303600	190180	190719	plasmoredoxin
RNA2t5:2791	191187	191244	
Pf3D7_0303700	191245	192591 c	dihydroilpoamide acyltransferase, putative

## Misc features

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'. You can also run Pfam by going to the Run menu and selecting 'Pfam search'. For this you need to select one CDS.

### **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.

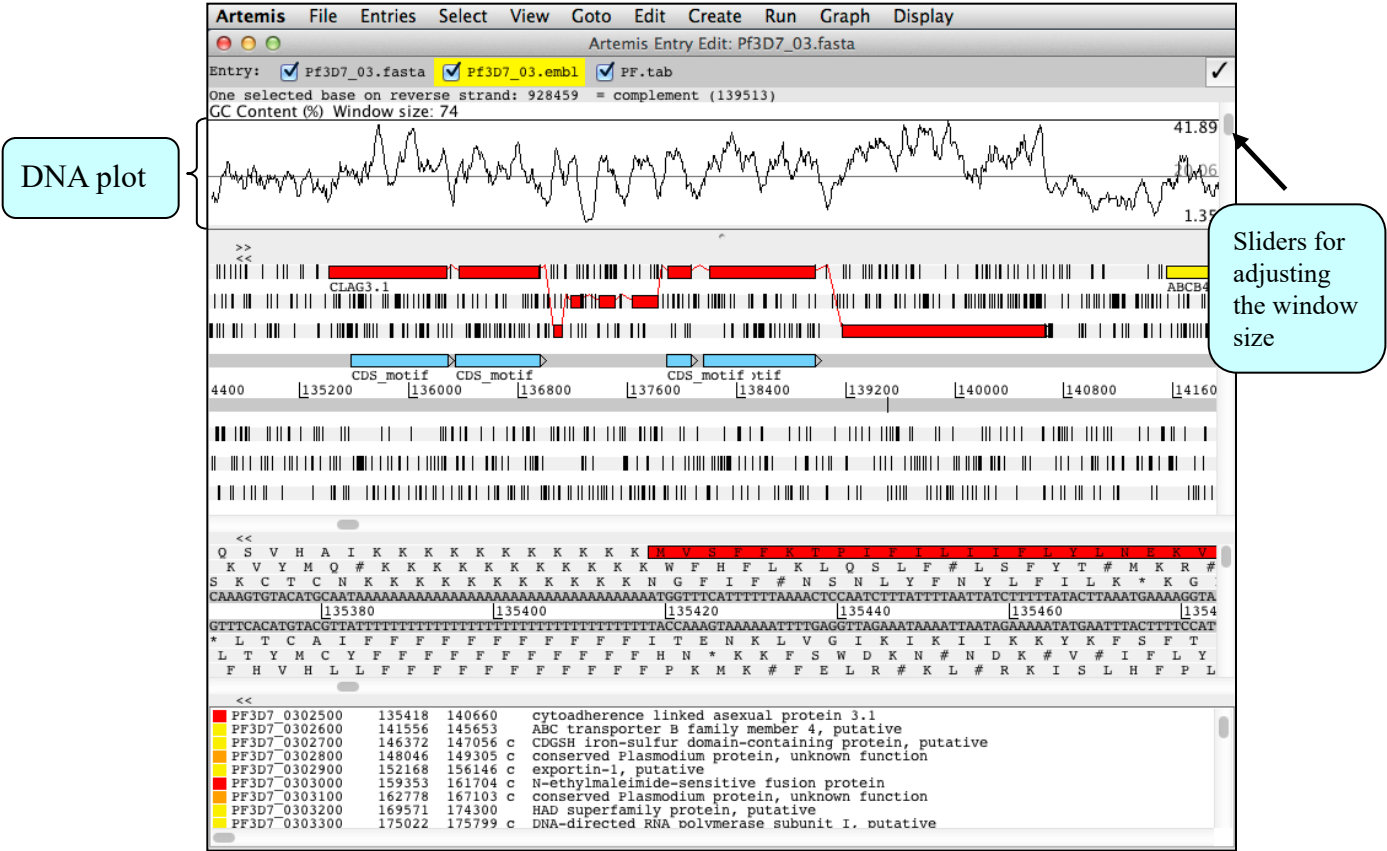
Further information on specific Pfam entries can be found on the web at <http://pfam.xfam.org/>



In addition to looking at the fine details 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 ‘Graph’ menu to see all those available. Some of the most useful plots for *P. falciparum* is the ‘GC Content (%)’ 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.



## 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 *rif<sup>R</sup>*, as shown below.

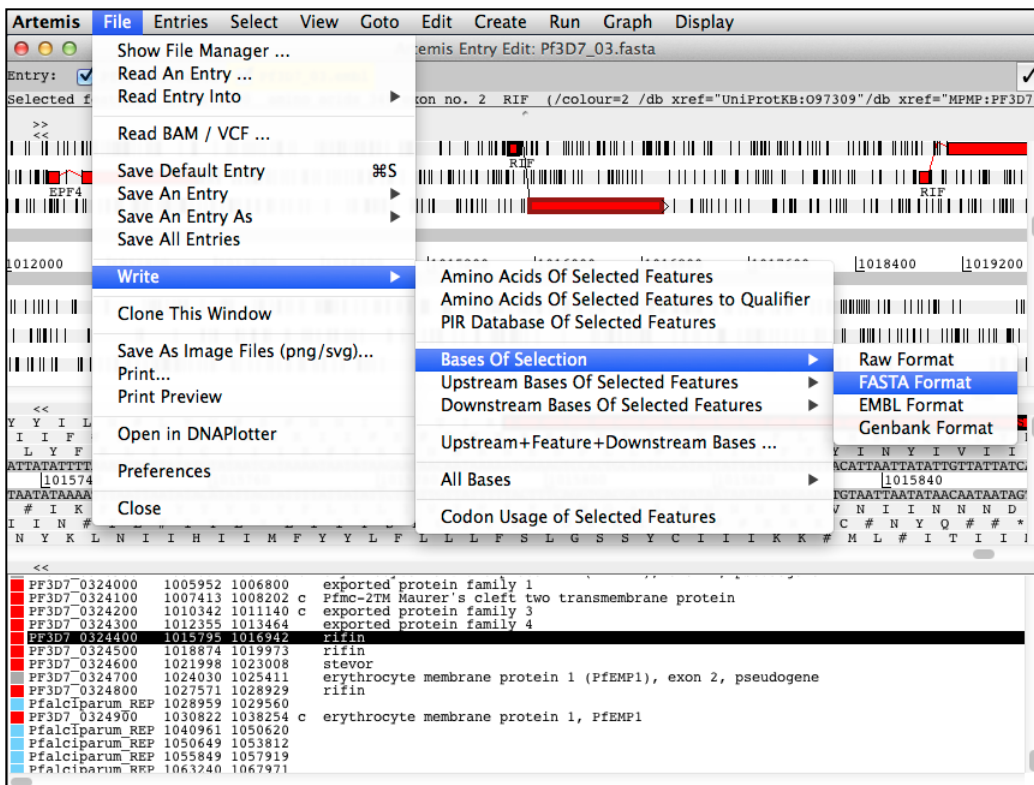
- Click 'Select' then 'Feature Selector'
- Make sure the buttons are down  
Set Key to 'CDS' and Qualifier to 'product'
- Type search term
- Click to select features containing search term
- Click to view selected features
- Double click to bring features into main view window.

The screenshot shows the Artemis software interface. The main window displays a sequence view with various features. The 'Feature Selector' dialog box is open, showing the search criteria. The 'Key' is set to 'CDS' and the 'Qualifier' is set to 'product'. The search term 'rifin' is entered in the 'Containing this text:' field. The 'Allow Partial Match' checkbox is checked. The 'Select' button is highlighted. The background shows the Artemis main window with a sequence view and a list of features.

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

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.



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

Select View Close

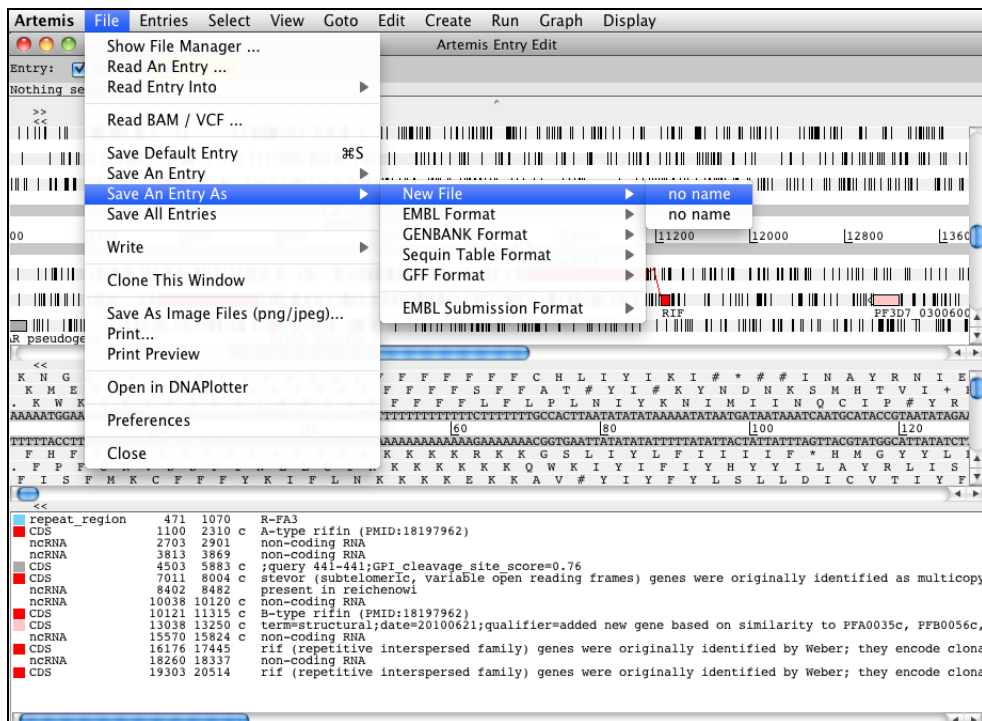
Space for a search term or amino acid motif

In the next part of the exercise we will be looking at the region containing the *rif<sup>r</sup>* 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.

The screenshot displays the Artemis Entry Edit interface. At the top, there are two tabs labeled "Entry:" and "no name". Below them, a section titled "Nothing selected" shows a genomic map with various features. A red bar highlights a region around position 9600, labeled "RNAzID:2755". Another red bar highlights a region around position 10400, labeled "RIF". A blue bar highlights a region around position 11200, labeled "PF3D7\_030040". The map also shows other features like "R pseudogene" and "PF3D7\_030040". Below the map, a sequence alignment is shown with positions 20, 40, 60, 80, 100, and 120. The sequence includes nucleotide bases (A, C, G, T) and gaps (-). At the bottom, a table lists genomic features and their coordinates:

Feature	Start	End	Description
repeat_region	471	1070	R-FA3
CDS	1100	2310	A-type rifin (PMID:18197962)
ncRNA	2703	2901	non-coding RNA
ncRNA	3813	3869	non-coding RNA
CDS	4503	5883	query 481-441; GPI cleavage site score=0.76
CDS	7011	8004	stevor (subtelomeric, variable open reading frames) genes were originally identified as multicopy present in reichenowi
ncRNA	8402	8482	non-coding RNA
ncRNA	10038	10120	B-type rifin (PMID:18197962)
CDS	10121	11315	term=structural,date=20100621;qualifier=added new gene based on similarity to PFA0035c, PFB0056c,
ncRNA	13038	13252	non-coding RNA
ncRNA	15570	15824	rif (repetitive interspersed family) genes were originally identified by Weber; they encode clones
ncRNA	16176	17445	non-coding RNA
ncRNA	18260	18337	rif (repetitive interspersed family) genes were originally identified by Weber; they encode clones
CDS	19303	20514	

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.



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

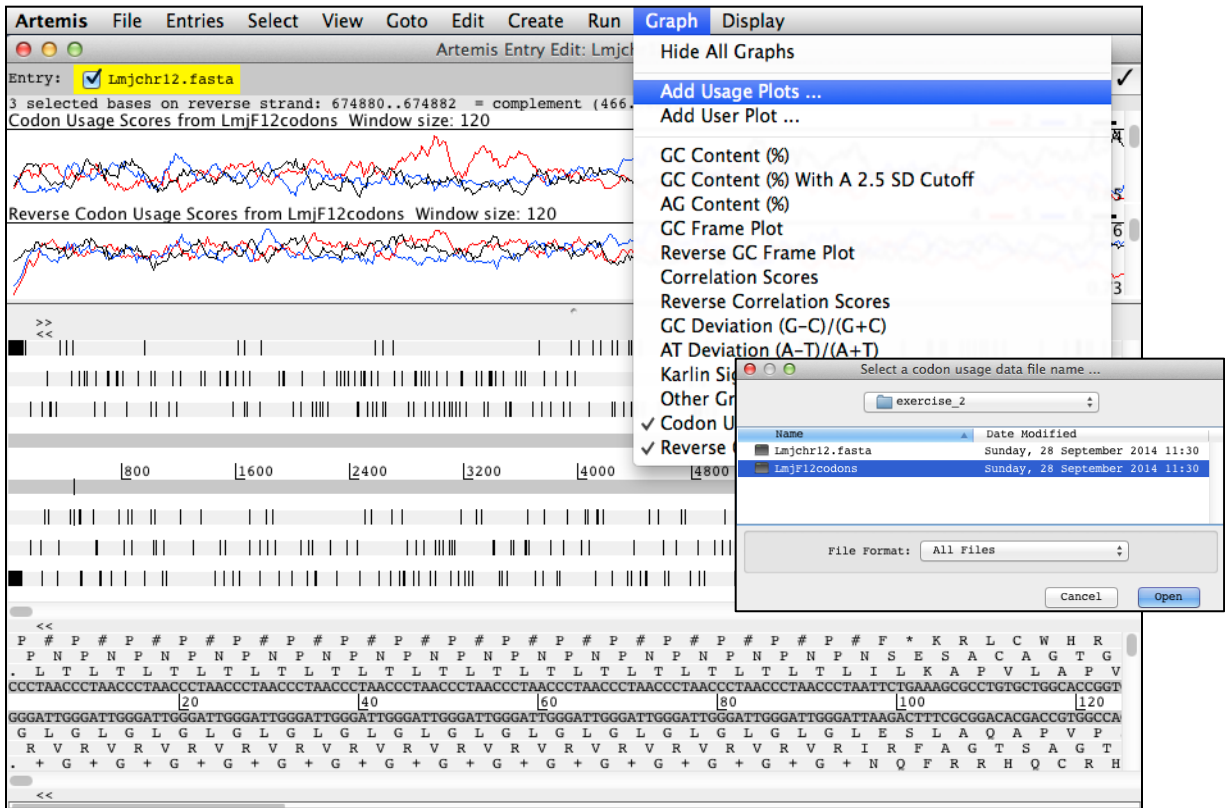
# Optional exercise

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, optional\_exercise 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.





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 rapidly 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.

The screenshot shows the Artemis genome browser interface. The 'Create' menu is open, and 'Mark ORFs in Range ...' is selected. The 'New Entry' label points to the 'ORFS\_100+' entry in the 'Entries' list. The 'Predicted ORF' label points to a blue bar representing an ORF in the sequence window. The bottom window shows the predicted peptide sequence and a table of ORF statistics.

Feature	Start	End	Score
CDS	460	960	none
CDS	964	1617	none
CDS	1188	1583	none
CDS	1783	2577	none
CDS	2707	3735	none
CDS	3986	4798	none
CDS	4965	6044	none
CDS	5917	6288	none
CDS	6405	8081	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 the Comparative Genomics Module), 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.

conserved hypothetical protein [Leishmania major strain Friedlin]  
Sequence ID: ref|XP\_001681812.1| Length: 620 Number of Matches: 1  
[See 1 more title\(s\)](#)

Range: 1 to 620 [GenPlot](#) [Graphics](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
1238 bits (3203)	0.0		Compositional matrix adjust.	620/620 (100%)	0/620 (0%)	+1
Query 40	MHTTPTTFSFSPFPFVAPLSPIAHLAHAGLLRQPVQLRMSGEFSLQLGLVTV					219
Sbjct 1	MHTTPTTFSFSPFPFVAPLSPIAHLAHAGLLRQPVQLRMSGEFSLQLGLVTV					60
Query 220	DKCDSADLTPTTSAKAARFVWAPTHPPSKCOTAEZTCVLEKRRGAHPVADPTGE					399
Sbjct 61	DKCDSADLTPTTSAKAARFVWAPTHPPSKCOTAEZTCVLEKRRGAHPVADPTGE					120
Query 400	VLLRRKNGVEQIVYNSKIPSVGVNKLAKMKEREENSPLFKYPLAVGEAQGE					579
Sbjct 121	VLLRRKNGVEQIVYNSKIPSVGVNKLAKMKEREENSPLFKYPLAVGEAQGE					180
Query 580	ARRVLQELRRCHQEQAMRKKEGLRERARLAERAVVQKAGTAREADARRKIMD					759
Sbjct 181	ARRVLQELRRCHQEQAMRKKEGLRERARLAERAVVQKAGTAREADARRKIMD					240
Query 760	GEAVSETAAKALTREERADANVDARVASGNCKEEDDRRLAAERTQLAEENFRA					939
Sbjct 241	GEAVSETAAKALTREERADANVDARVASGNCKEEDDRRLAAERTQLAEENFRA					300
Query 940	AQQRARAKAQEQAEARARVEMELQGLAQERVKLEGIHRRNAELKGAQDSARERRW					1119
Sbjct 301	AQQRARAKAQEQAEARARVEMELQGLAQERVKLEGIHRRNAELKGAQDSARERRW					360
Query 1120	RANSADVHLQAMPNSLFDVARRQRDEANRAQQRMEDTAVNVRLAQKRAQAQRDR					1299
Sbjct 361	RANSADVHLQAMPNSLFDVARRQRDEANRAQQRMEDTAVNVRLAQKRAQAQRDR					420
Query 1300	DRQYAEYAKALENFQREVEHARQRQERQELQDAEATAVRQAHADAARRREGSVV					1479
Sbjct 421	DRQYAEYAKALENFQREVEHARQRQERQELQDAEATAVRQAHADAARRREGSVV					480
Query 1480	PLLPWPAQSGCAEKAKIDANRRFREDLRQAQQRDERAQEEAERADRALVEYDTL					1659
Sbjct 481	PLLPWPAQSGCAEKAKIDANRRFREDLRQAQQRDERAQEEAERADRALVEYDTL					540
Query 1660	ARAVERRERKRRKAEHLRRLELAQIAEKRRGAVGDRACAAADVTHVPATEAMRLTC					1839
Sbjct 541	ARAVERRERKRRKAEHLRRLELAQIAEKRRGAVGDRACAAADVTHVPATEAMRLTC					600
Query 1840	PVTGELLPASAYDFGVQR					1899
Sbjct 601	PVTGELLPASAYDFGVQR					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

Entry: ☒ Lmjchr12.fasta ☒ ORFS\_100+

Selected feature: bases 1902 amino acids 633 CDS (/score=51 /colour=128 128 255 /note="none")

Codon Usage Scores from LmjF12codons Window size: 120

Reverse Codon Usage Scores from LmjF12codons

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.

Artemis Feature Edit: CDS

Key: CDS Add Qualifier: note

Location: 12745..14646

Complement Grab Range Remove Range Goto Feature Tidy TAT ObjectEdit User Qualifiers

systematic\_id="LmjF12.0070"  
product="hypothetical protein, conserved"  
colour=10

OK Cancel Apply

Feature	Start	End	Score
CDS	12745	14646	none
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	16444	none
CDS	16731	17000	none

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 File Entries Select View Goto Edit Create Run Graph Display

Artemis Entry Edit: Lmjchr12.fasta

Entry: ☒ Lmjchr12.fasta ☒ ORFS\_100+

Selected feature: bases 1902 amino acids 633 LmjF12.0070 (/systematic id="LmjF12.0070" /product="hypothetical protein, conserved")

Codon Usage Scores from LmjF12codons Window size: 172

Reverse Codon Usage Scores from LmjF12codons Window size: 150

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

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

Sequence: << N S L V L R T A L A S S C T P M D R T R \* K R T T R W R P S I Y A V M H T H T F E T A I L S S C A O L L R L R V L Q W T E H V R E L H V G D L L C M P S C T H T R L S I O F S R P A H S S C V F V Y S N G O N T K E N Y T L A T F F V C R H A H T H A F H C A A T T C T C T C G T C C T G C G C A C A G C T T G C G T C T T C G T G T A C T C C A A T G G A C A G A A C A G T C A A A G A G A A C T A C A C G T T G G C G A C C T C T T T G T A T G C C G T C A T G C A C A C A C A C G C C T T C A C T 2680 12700 12720 12740 12760 12780 12800 G T T A A G A G A C A G G A C G C G T G T C G A A A C G A A G C A C A T G A G G T T A C C T G T C T T G T G C A C T T T C T C T T G A T G T G C A A C C G C T G G A A G A A C A T A C G G C A G T A C G T G T G T G T G C G G A A A G T G A C N E R G A C I R E D Q A C S K R R R T S W H V S C S L S S C T P S R R Q I G D H V C V R R E S L E R T R R V V R Q R G E K Y A T M C V C V G K V +

CDS	Start	End	Product
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

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.

The screenshot shows the Artemis genome browser interface. The 'Select' menu is open, and 'Features Matching Qualifier' is selected. A yellow callout box with an arrow points to this menu item, labeled '2. Select Features Matching Qualifier from Select menu'. Below the menu, a 'Select a qualifier name ...' dialog box is open, showing 'colour' in the dropdown. A yellow callout box with an arrow points to this dialog, labeled '3. Select colour as a qualifier. This will select all features of the same colour.' The background shows a genomic track with various features, including CDS (Coding DNA Sequences) and Lm12.0080. A yellow callout box with an arrow points to a specific gene model in the track, labeled '1. Select an annotated gene model'. The bottom of the screen shows a list of features with their coordinates and qualifiers.

Feature	Start	End	Qualifier
CDS	11445	12026	none
CDS	12504	12809	none
CDS	12745	14646	none
CDS	12923	13297	none
CDS	12969	13457	none
CDS	14793	15299	none
CDS	15055	15471	none
CDS	15300	15665	none
CDS	15821	16201	none

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

5. After the features have been copied to Lmjchr12.fasta, de-select ORFS\_100+. Only annotated ORFs should remain.

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