GenePalette User Manual

Version 2.1

Table of Contents

New to Version 2.1

♦ The OrthologGrabber module automates the acquisition of orthologous sequences by performing BLAT searches to the UCSC genome browser database

♦ Fixed bugs with sequence alignment functions that were particularly noticeable with small word sizes

♦ Updated GenBank Access protocols to comply with new shift to https

May 2017

New to Version 2.0

- Multiple sequences can now be compared to a reference sequence, and manipulated graphically
    - Compare as many sequences as you want
    - Excellent for both phylogenetic footprinting, and for the navigation of rapidly evolving sequences
    - Dynamic focusing of alignment by clicking on landmarks to center the view
    - Pseudo-alignment of orthologous points (anchor points)
    - Easily detect insertions/deletions/inversions
- To save space on the screen, individual interface components can be hidden
- In program PCR primer calculations make it even easier to design your experiments.

August 2009

New to Version 1.1

Since the release of version 1.04 of GenePalette in June, 20002, we have made many changes. The software is improved in so many ways that we decided it deserved its own tenth. So, here is a list of the major changes introduced with version 1.1 of GenePalette:

- Improved speed of access to fragments of large contigs – human/mouse data is accessed at least 10 times faster!
- Added the ability to import gene annotation from **Ensembl** with our new GenBank format importer
- New space saving layout that is more friendly to smaller screens
- Postscript output of Graphical and Markup views allows users to edit their images in graphics packages such as Adobe Illustrator
- Enhanced support for access to local sequences – you can search your local genome collections by gene-symbol
- Improved organization of local sequences – curate your local GenBank directory by storing collections in their own sub-directory
- New clickability functions allows enhanced connectivity between interface elements

October 2003

CHAPTER 1: GenePalette Tutorials

Introduction

This chapter contains several tutorials that will guide the user through the operation of GenePalette. The first tutorial will give the user the bare-minimum knowledge of how to use the program. The later tutorials will expand on the fundamentals to demonstrate some of the more complex abilities of GenePalette. Another way to become acquainted with the features of the program is to watch the instructional videos located on the documentation page of the website.

In these tutorials, we will take you step-by-step through the use of the program: Program components (menu's, buttons, sliders, etc) are shown in **Bold** text. Actions that you should perform in the program are Underlined.

I should note that some of the images presented are from older versions of the program. However, they are still relevant to the necessary functions being demonstrated.

Tutorial 1: Reconstruction of a published reporter construct

When studying enhancers, it is common to want to understand how a published reporter fragment was constructed. Although it appears trivial, this process can be quite time-consuming and difficult to accomplish. However, using GenePalette, this task is executed quite easily. In this example, the genomic region surrounding a gene of the Enhancer of Split Complex of *Drosophila melanogaster* will be loaded and viewed with respect to a published upstream enhancer. We will use this loaded sequence to highlight the basic features of GenePalette.

Loading a GenBank Sequence

- A copy of the sequence downloaded from GenBank is included in the Sequences directory under the main GenePalette directory. If you are not connected to the internet when doing the tutorial, you can go to the **File** menu and select **Open Sequence**. Select the file named tutorial.seq.
- For a tutorial about constructing effective Entrez queries, look at the "Entrez Help" section featured at www.ncbi.nlm.nih.gov/Entrez/.

- Chapter 2 will provide some helpful hints about finding genome sequence for genes in specific organisms.

To load a sequence from GenBank, you first must go to the **GenomeTools** menu, and select the menu-item **Entrez Nucleotide Query (NLM)**. Clicking this menu item brings up a dialog that asks for an Entrez Query. The text that is entered into the dialog will be sent directly as a nucleotide search to the National Library of Medicine's Entrez server. To load a gene in any sequenced genome, one must know how that gene is referred to in genome annotations. The gene that will be loaded for this exercise is called *Enhancer of Split (E(spl)) mγ.* However, in the annotated fly genome, it appears as HLHmgamma. For the *Drosophila* genome, one can consult FlyBase (www.flybase.org) for the official gene symbol, and in most cases, that symbol will be used in GenBank as well.

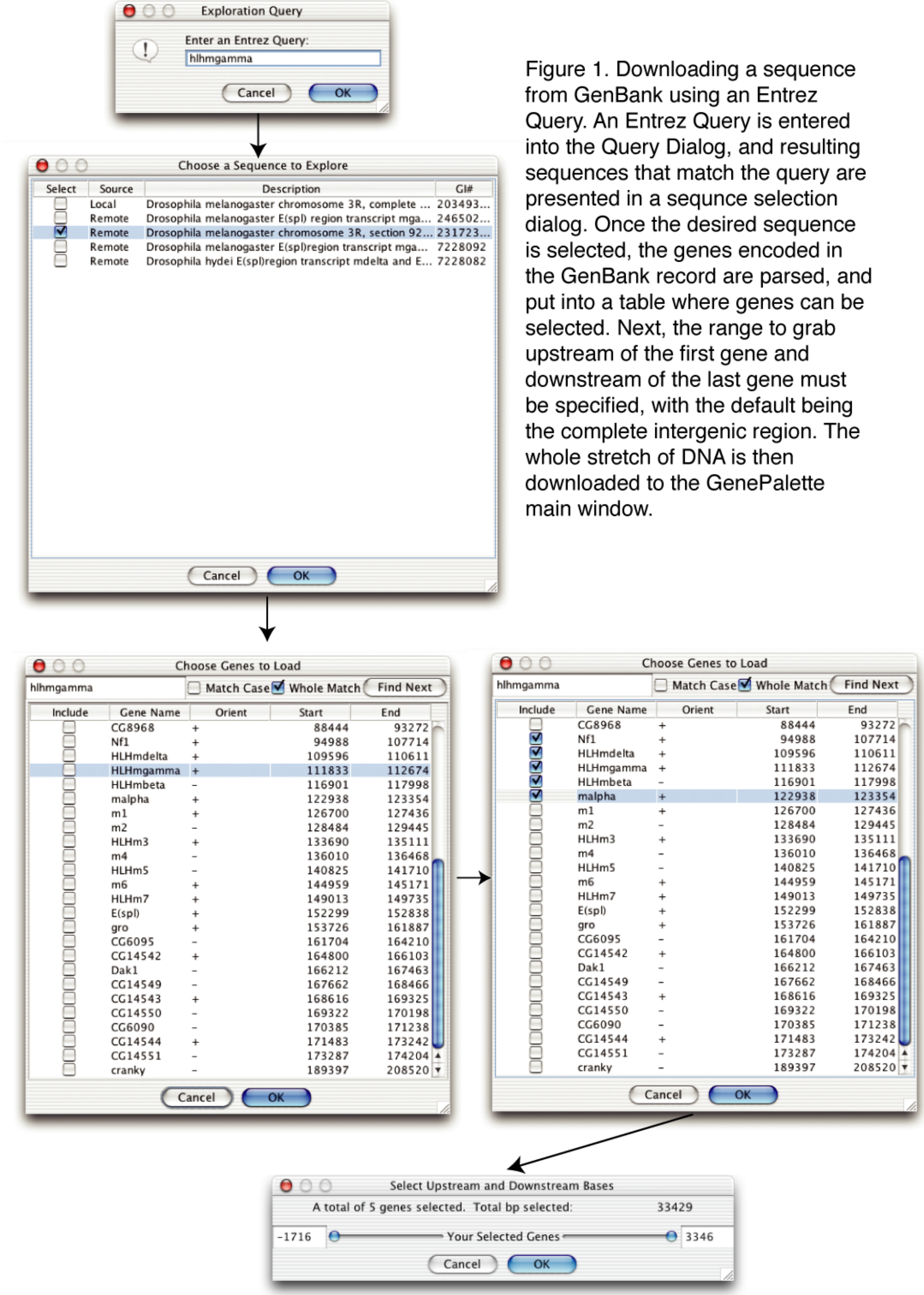
To access the genomic region surrounding *E(spl) mγ*, type "hlhmgamma Drosophila chromosome" into the **Entrez Query dialog box**, and hit OK. A loading dialog appears on screen to update how many sequence matches were found. Once all of the matching sequences are loaded, a selection dialog appears that gives the one-line description for each sequence. If you typed the same query (hlhmgamma) into an Entrez nucleotide search using a web-browser (www.ncbi.nlm.nih.gov/Entrez/), you would get the same sequences that appear in the sequence selection dialog. Click on the checkbox next to the description line labeled "**Drosophila** melanogaster **chromosome** 3R", and hit OK.

Choosing genes to load

Once the sequence has been selected, a loading dialog appears, giving a running update of how many genes associated with the sequence have been loaded. For large chromosomes, this may take a minute. When all of the genes are loaded, the gene annotation data is presented in a selection table. If there is an exact match between a gene name on the table and the first word of the query, the line containing that gene will be highlighted (As is the case in our example). To select a gene from this sequence, click the checkbox in the leftmost column of the table. For this example, we will check HLHmgamma, and the two neighboring genes on either side: Nf1 and HLHmdelta which are before HLHmgamma on the table, and also HLHmbeta and malpha, which appear below HLHmgamma on the gene selection table. After clicking their checkboxes, hit the OK button.

Choosing flanking base-pairs

The next dialog to appear is used to select the upstream and downstream flanking bases that will be downloaded. The dialog consists of two slider bars, the first is used to specify the number of bases upstream of the first gene selected, and the second specifies the number of bases downstream of the last gene selected on the sequence. The default value is set to take the maximum length of intergenic region up to the first base of the next transcript in both directions. In the current dialog, hitting the OK button would result in taking 1716 bp upstream of the first



Enter an Entrez Query: hlhmgamma

Select	Source	Description	GI#
□	Local	Drosophila melanogaster chromosome 3R, complete ...	203493...
□	Remote	Drosophila melanogaster E(spl) region transcript mga...	246502...
☑	Remote	Drosophila melanogaster chromosome 3R, section 92...	231723...
□	Remote	Drosophila melanogaster E(spl)region transcript mga...	7228092
□	Remote	Drosophila hydei E(spl)region transcript mdelta and E...	7228082

hlhmgamma □ Match Case ☑ Whole Match

Include	Gene Name	Orient	Start	End
□	CG8968	+	88444	93272
□	Nf1	+	94988	107714
□	HLHmdelta	+	109596	110611
□	HLHmgamma	+	111833	112674
□	HLHmbeta	-	116901	117998
□	malpha	+	122938	123354
□	m1	+	126700	127436
□	m2	-	128484	129445
□	HLHm3	+	133690	135111
□	m4	-	136010	136468
□	HLHm5	-	140825	141710
□	m6	+	144959	145171
□	HLHm7	+	149013	149735
□	E(spl)	+	152299	152838
□	gro	+	153726	161887
□	CG6095	-	161704	164210
□	CG14542	+	164800	166103
□	Dak1	-	166212	167463
□	CG14549	-	167662	168466
□	CG14543	+	168616	169325
□	CG14550	-	169322	170198
□	CG6090	-	170385	171238
□	CG14544	+	171483	173242
□	CG14551	-	173287	174204
□	cranky	-	189397	208520

hlhmgamma □ Match Case ☑ Whole Match

Include	Gene Name	Orient	Start	End
□	CG8968	+	88444	93272
☑	Nf1	+	94988	107714
☑	HLHmdelta	+	109596	110611
☑	HLHmgamma	+	111833	112674
☑	HLHmbeta	-	116901	117998
☑	malpha	+	122938	123354
□	m1	+	126700	127436
□	m2	-	128484	129445
□	HLHm3	+	133690	135111
□	m4	-	136010	136468
□	HLHm5	-	140825	141710
□	m6	+	144959	145171
□	HLHm7	+	149013	149735
□	E(spl)	+	152299	152838
□	gro	+	153726	161887
□	CG6095	-	161704	164210
□	CG14542	+	164800	166103
□	Dak1	-	166212	167463
□	CG14549	-	167662	168466
□	CG14543	+	168616	169325
□	CG14550	-	169322	170198
□	CG6090	-	170385	171238
□	CG14544	+	171483	173242
□	CG14551	-	173287	174204
□	cranky	-	189397	208520

A total of 5 genes selected. Total bp selected: 33429

-1716 Your Selected Genes 3346

Figure 1. Downloading a sequence from GenBank using an Entrez Query. An Entrez Query is entered into the Query Dialog, and resulting sequences that match the query are presented in a sequnce selection dialog. Once the desired sequence is selected, the genes encoded in the GenBank record are parsed, and put into a table where genes can be selected. Next, the range to grab upstream of the first gene and downstream of the last gene must be specified, with the default being the complete intergenic region. The whole stretch of DNA is then downloaded to the GenePalette main window.

gene (Nf1), and 3346 bp downstream of the last gene (malpha) selected. The minimum value that can be specified by this dialog is no sequence upstream or downstream of the selected genes. In this exercise, it doesn't really matter, since we are generously grabbing a two-gene radius around mγ: go ahead and hit OK to take the complete upstream and downstream intergenic regions.

Interacting with the loaded sequence

After the OK button is pressed, a progress dialog appears that notifies you of download progress. After the download, the sequence data is loaded into the main window, and you can begin to navigate and use the sequence. First note the anatomy of the loaded window. There are four main areas that are separated by resizable dividers (Figure 2). Experiment with resizing the divider spacing.

The Sequence Display

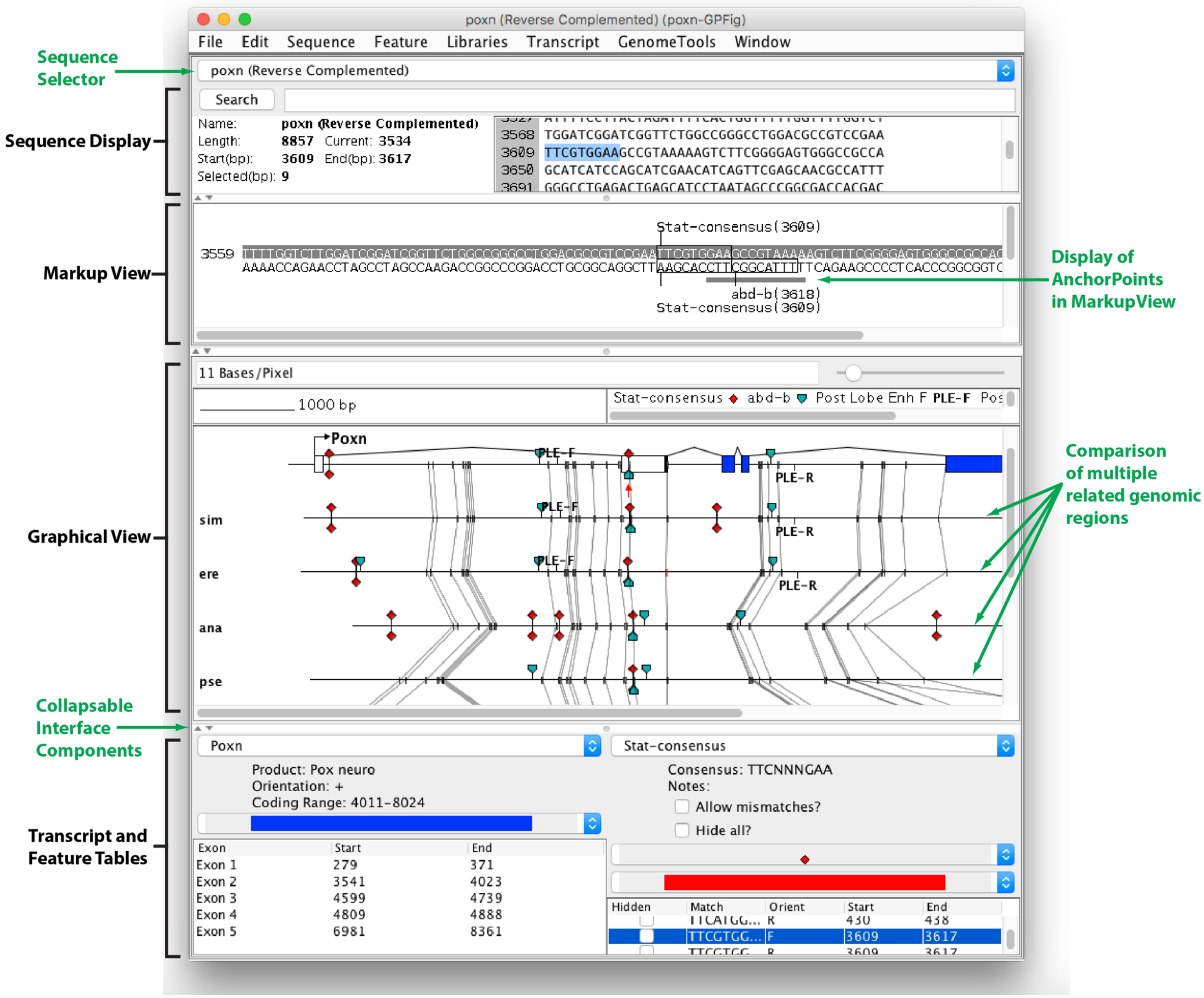
The top area of the loaded GenePalette window contains the nucleotide sequence, along with information about the sequence. As you select portions of the sequence, data pertaining to your selection is reported in panel to the left. You will also notice that a box will appear in the graphical view that shows what you have selected. Select some part of the sequence to try this feature out. Note that you can copy portions of the sequence from this text area.

The Markup Display

The markup view is the next area of the main window, and is initially a blank panel. Later during the session we will be using this area to view features at the nucleotide sequence level. Upon user interaction with the graphical display, regions are displayed with features highlighted on the sequence.

The Graphical Display

The third main area contains a graphical representation of the sequence. The top panels contain data about this graphical view: the base-pair:pixel ratio, a slider bar that allows you to adjust the ratio, a scale bar, and a legend panel that will indicate what each symbol means as we start adding features to this sequence. Move the slider to see how the base pair : pixel ratio affects the graphical view. The main part of the graphical display is the graphical representation itself (**Graphical View**). Note how you can move the scroll bars up/down and left/right to scroll to parts of the sequence that you want to see. Genes appear in the graphical view as boxes. The direction of transcription can be seen by both the direction of the arrow that comes out of the box (arrow pointing right is on the top strand, arrow pointing left is on the bottom strand). The colored boxes represent


Sequence Selector
Sequence Display
Markup View
Graphical View
Collapsable Interface Components
Transcript and Feature Tables
Display of AnchorPoints in MarkupView
Comparison of multiple related genomic regions
poxn (Reverse Complemented) (poxn-GPFig)
File Edit Sequence Feature Libraries Transcript GenomeTools Window
poxn (Reverse Complemented)
Search
Stat-consensus(3609)
abd-b(3618)
Stat-consensus(3609)
11 Bases/Pixel
1000 bp
Stat-consensus abd-b Post Lobe Enh F PLE-F Pos
Poxn
PLE-F
PLE-R
sim
ere
ana
pse
Poxn
Stat-consensus

Name: **poxn (Reverse Complemented)**
Length: **8857** Current: **3534**
Start(bp): **3609** End(bp): **3617**
Selected(bp): **9**

```
3568 TGGATCGGATCGGTTCTGGCCGGGCCTGGACGCCGTCCGAA
3609 TTCGTGGAAGCCGTAAAAAGTCTTCGGGGAGTGGGCCGCCA
3650 GCATCATCCAGCATCGAACATCAGTTCGAGCAACGCCATTT
3691 GGGCCTGAGACTGAGCATCCTAATAGCCCGGCGACCACGAC
```

```
3559 TTTTGGTCTTGGATCGGATCGGTTCTGGCCGGGCCTGGACGCCGTCCGAATTCGTGGAAGCCGTAAAAAGTCTTCGGGGAGTGGGCCGCCAC
     AAAACCAGAACCTAGCCTAGCCAAGACCGGCCCGGACCTGCGGCAGGCTTAAGCACCTTCGGCATTTTTCAGAAGCCCCTCACCCGGCGGTC
```

Product: Pox neuro
Orientation: +
Coding Range: 4011-8024

Exon	Start	End
Exon 1	279	371
Exon 2	3541	4023
Exon 3	4599	4739
Exon 4	4809	4888
Exon 5	6981	8361

Consensus: TTCNNNGAA
Notes:
□ Allow mismatches?
□ Hide all?

Hidden	Match	Orient	Start	End
□	TTCATGG...	R	430	438
□	TTCGTGG...	F	3609	3617

Figure 2. GenePalette Window, showing active interface components. Features new to version 2.1 are labeled in green.

coding portions of exons, and the white boxes represent non-coding exon portions. It is important to observe that the gene Nf1 has three splice variants annotated. One variant is displayed on the line that represents the DNA sequence, while the other two variants are placed above the first. Use the scroll bars on the left and bottom of the graphical view to look at the three alternate transcripts of Nf1. Can you tell what differences exist between the alternates?

Notice that when you click on an exon in this view, several things happen. First, the clicked exon is highlighted in red. Second, the exon will be highlighted in the bottom area, which holds annotation data. Finally, the range of sequence represented by the box you clicked will be highlighted in the sequence display. If you click on a white box, you will select sequence that is in an untranslated region.

If you click on a colored box, you will select sequence that is in the coding region of an exon. If an exon is entirely coding or non-coding, you will select the sequence that spans the whole exon. Click on both coding and non-coding exons in the graphical display to get a feel for how the other components react to this display. Make sure to click on different exons of Nf1 so you can really see how exons are selected in the data table.

Data Tables

The final area of the main GenePalette window is split vertically into two panels. The rightmost panel is currently empty, but will soon contain data about features that will be added to the sequence. The leftmost panel contains data about the genes that currently reside on the sequence. Each transcript has an entry in the Combo-Box at the top of the panel, labeled with the name of the gene (And the number of alternate transcripts in parentheses as in Nf1). To access different transcripts, simply click on the combo-box and highlight the gene you want to see. Directly under the combo-box, there is data about the gene unit (product name, gene orientation, and a range that specifies where coding sequence starts and ends). The final component of this top region is a Combo-Box that designates the color of the gene. Experiment with changing the transcript color and note how this changes the color in the graphical display.

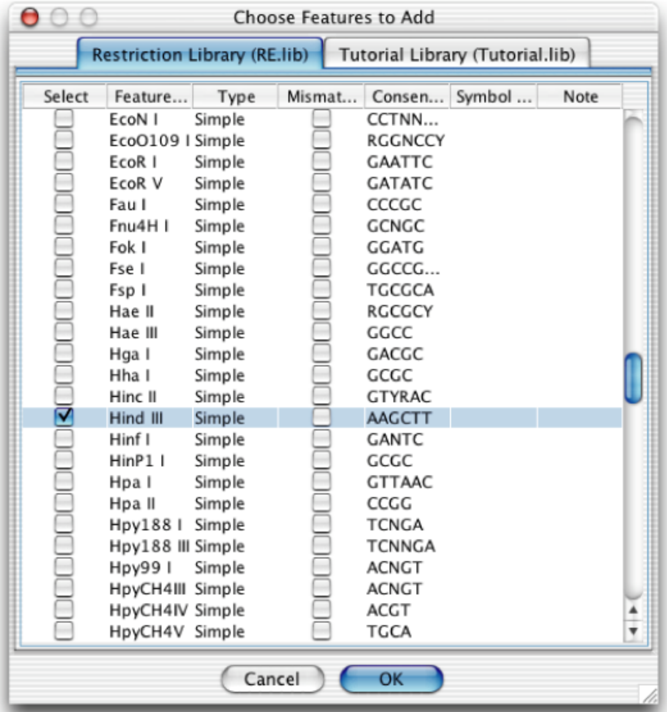
The bottom portion of each transcript's data panel contains a table of all the exons. When you click on a row in the exon table, the exon is boxed in the graphical view. So that you can see how exons overlap, the box is drawn to be as tall as the tallest overlapping exon. For example, clicking on any of the first 17 exons of the blue Nf1 transcript will give a box that surrounds all three transcripts. However, if you click on Exon 18, the box will only surround the first two transcripts. Click on the rows in the exon tables of Nf1 to see how exon-boxing works. Now that we are comfortable with the basic sequence operations of the main window, it is time to explore the regulatory sequence of the gene *E(Spl) mγ*.

Adding Features to a Sequence

In this section we will add some features to the sequence we have loaded. Features are defined as any sequence element that can be described by sequence identity. These would include transcription factor binding sites, primer sequences, restriction enzyme sites, SNPs, mRNA regulatory motifs, or anything else you can think of. In Nellesen et al., 1999[1], the cloning of an enhancer element upstream of *E(Spl) mγ* was described:

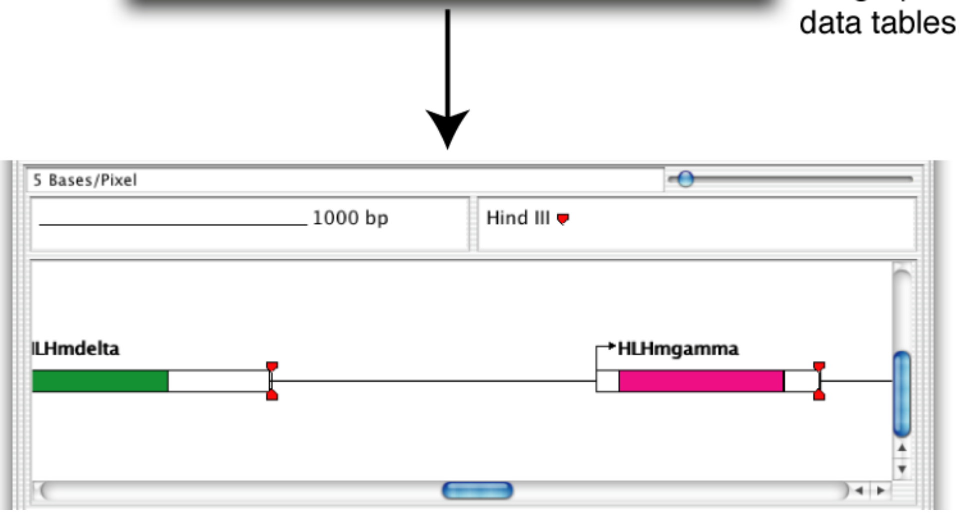
[1] Nellesen DT, Lai EC, Posakony JW. 1999. Discrete enhancer elements mediate selective responsiveness of enhancer of split complex genes to common transcriptional activators. Dev Bio 1;213(1):33-53

mγ**.** From a 2.1-kb *Hin*dIII fragment containing *mγ* (Delidakis and Artavanis-Tsakonas, 1992), a 1243-bp *Ecl*136II–*Hin*d III fragment was subcloned into CaSpeRlacZ (CZmγ1.2). A 234-bp *Kpn*I–*Xba*I fragment containing the mγ enhancer was cloned directly into HZCaSpeRlacZ (HZmgKX).



Select	Feature...	Type	Mismat...	Consen...	Symbol ...	Note
□	EcoN I	Simple	□	CCTNN...		
□	EcoO109 I	Simple	□	RGGNCCY		
□	EcoR I	Simple	□	GAATTC		
□	EcoR V	Simple	□	GATATC		
□	Fau I	Simple	□	CCCGC		
□	Fnu4H I	Simple	□	GCNGC		
□	Fok I	Simple	□	GGATG		
□	Fse I	Simple	□	GGCCG...		
□	Fsp I	Simple	□	TGCGCA		
□	Hae II	Simple	□	RGCGCY		
□	Hae III	Simple	□	GGCC		
□	Hga I	Simple	□	GACGC		
□	Hha I	Simple	□	GCGC		
□	Hinc II	Simple	□	GTYRAC		
☑	Hind III	Simple	□	AAGCTT		
□	Hinf I	Simple	□	GANTC		
□	HinP1 I	Simple	□	GCGC		
□	Hpa I	Simple	□	GTTAAC		
□	Hpa II	Simple	□	CCGG		
□	Hpy188 I	Simple	□	TCNGA		
□	Hpy188 III	Simple	□	TCNNGA		
□	Hpy99 I	Simple	□	ACNGT		
□	HpyCH4III	Simple	□	ACNGT		
□	HpyCH4IV	Simple	□	ACGT		
□	HpyCH4V	Simple	□	TGCA		

Figure 3. Dialog for Adding Features from Feature Libraries. Each library is a tabbed-pane in the dialog, and contains a table of the library's features. To select features to add, click on the tab for a library, and check off the features you want to add in the leftmost column's checkbox. Once you hit OK, the feature(s) is searched across the loaded sequence, and matches are displayed in the graphical view and data tables.


5 Bases/Pixel

Using the restriction library that is packaged with GenePalette, and a special library created for this tutorial, we can visualize the creation of this enhancer, and easily understand its makeup.

About Libraries

Because the addition of features is such a routine operation in GenePalette, we have created a system of Feature Libraries to store profiles for commonly added features. GenePalette comes with a restriction library of 208 enzymes. Additionally, we have included a tutorial library for the purposes of this exercise. We have not included a library of commonly used transcription factor binding sites because there are so many ways to interpret binding data such that we feel it is up to the user to compile libraries of sites that they believe in.

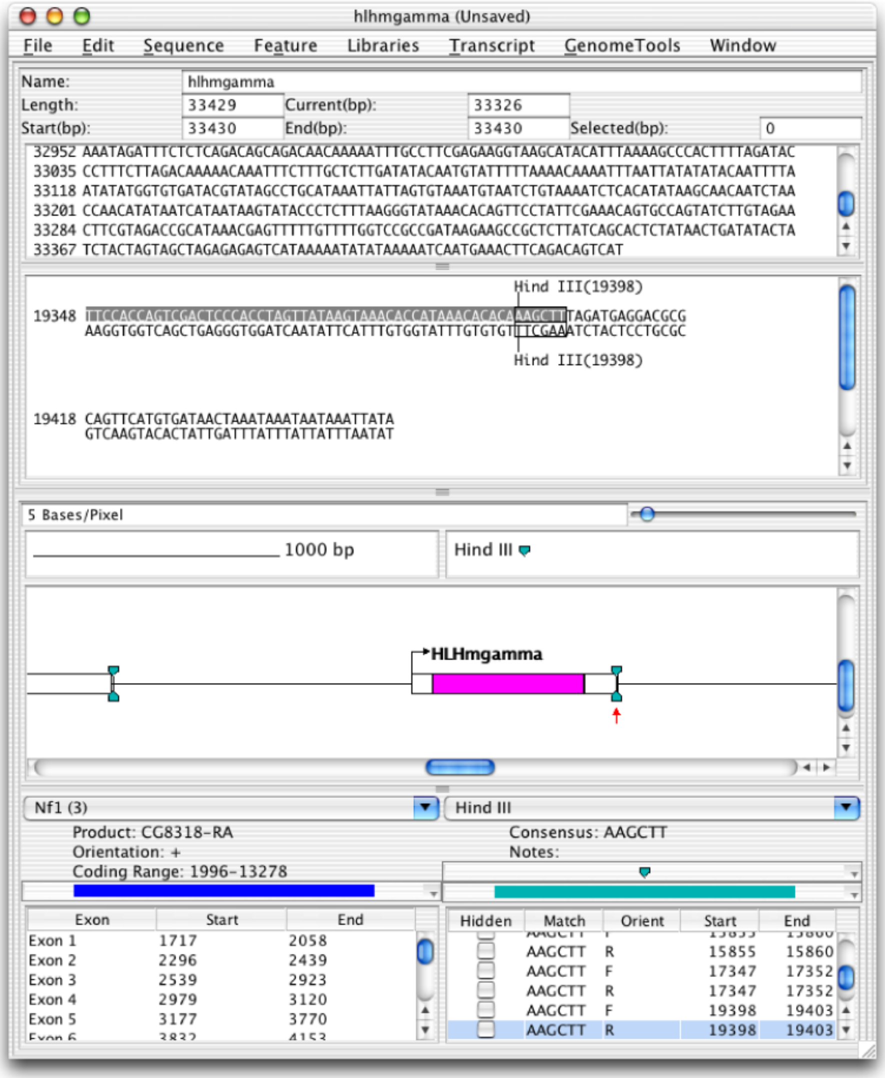
When the GenePalette application is started, all library files contained in the **Libraries** directory under the main application directory are loaded into the library management system. User libraries that are not contained in this directory can be loaded manually.

Adding Library Features

To add a feature from a library to a sequence, go to the **Libraries** menu, and click **Add Feature from Library**. Find the *Hind* III feature in the Restriction Library, and click the checkbox in the leftmost column of the table. Then click OK.

Interacting with added Features

The *Hind* III sites will show up on the graphical view as vertical lines above and below the sequence, terminated with a symbol that designates what feature it is. The feature appears both above and below the line because *Hind* III sites are palindromic: there is a match on both the top and bottom strands. Observe how there are *Hind* III sites flanking mγ in a ~2.0 kb chunk, as described by Nellesen et al., 1999. A panel containing data for the *Hind* III feature appears in the lower right-hand corner of the window. There you can find data about the feature as well as well as modify the appearance of the feature. Use the **shape combo-box** to select a differently shaped symbol for the feature. Use the **color combo-box** to change the color of the feature. Not only can you choose between different shapes, but you can also set the symbol to be a letter, or word. Select the **Text Symbol** option from the shape combo box, and write anything you want in the subsequent dialog that appears (like 'H' or something). Notice how these changes to the feature symbol change the graphical display. Below the shape manipulation portion of the feature panel there is a table of all of the sites that match the feature. If you click a row in this table, a red arrow will appear under the clicked site. If you click in the leftmost column of this table, the match whose row you clicked will be hidden in the graphical view. Experiment with clicking rows and checkboxes of the feature table for *Hind* III.



Name: hlhmgamma
Length: 33429 Current(bp): 33326
Start(bp): 33430 End(bp): 33430 Selected(bp): 0

```
32952 AAATAGATTTCTCTCAGACAGCAGACAACAAAAATTTGCCTTCGAGAAGGTAAGCATACATTTAAAAGCCCACTTTTAGATAC
33035 CCTTTCTTAGACAAAAACAAATTTCTTTGCTCTTGATATACAATGTATTTTTAAAACAAAATTTAATTATATATACAATTTTA
33118 ATATATGGTGTGATACGTATAGCCTGCATAAATTATTAGTGTAAATGTAATCTGTAAAATCTCACATATAAGCAACAATCTAA
33201 CCAACATATAATCATAATAAGTATACCCTCTTTAAGGGTATAAACACAGTTCCTATTCGAAACAGTGCCAGTATCTTGTAGAA
33284 CTTCGTAGACCGCATAAACGAGTTTTTGTTTTGGTCCGCCGATAAGAAGCCGCTCTTATCAGCACTCTATAACTGATATACTA
33367 TCTACTAGTAGCTAGAGAGAGTCATAAAAATATATAAAAATCAATGAAACTTCAGACAGTCAT
```

```
                                                       Hind III(19398)
19348 TTCCACCAGTCGACTCCCACCTAGTTATAAGTAAACACCATAAACACACAAAGCTTTAGATGAGGACGCG
      AAGGTGGTCAGCTGAGGGTGGATCAATATTCATTTGTGGTATTTGTGTGTTTCGAAATCTACTCCTGCGC
                                                       Hind III(19398)

19418 CAGTTCATGTGATAACTAAATAAATAATAAATTATA
      GTCAAGTACACTATTGATTTATTTATTATTTAATAT
```

5 Bases/Pixel

Nf1 (3)
Product: CG8318-RA
Orientation: +
Coding Range: 1996-13278

Hind III
Consensus: AAGCTT
Notes:

Exon	Start	End
Exon 1	1717	2058
Exon 2	2296	2439
Exon 3	2539	2923
Exon 4	2979	3120
Exon 5	3177	3770
Exon 6	3832	4153

Hidden	Match	Orient	Start	End
□	AAGCTT	F	15855	15860
□	AAGCTT	R	15855	15860
□	AAGCTT	F	17347	17352
□	AAGCTT	R	17347	17352
□	AAGCTT	F	19398	19403
□	AAGCTT	R	19398	19403

Figure 4. Clicking a feature. When a feature is clicked in the graphical view, three things happen: (1)A red arrow is displayed under the feature in the graphical view, (2) 50bp upstream and downstream of the site is loaded into the markup view, and (3) the row for that site is highlighted in the data table.

Interacting with Features in the Graphical View

Another way to access and visualize features is through the graphical view. When you click on a feature present in the sequence three things happen in the main window: (1) A red arrow is displayed under the feature in the graphical view, (2) 50bp upstream and downstream of the site is loaded into the markup view, and (3) the row for that site is highlighted in the data table (Figure 3). The red arrow serves as a place marker to remind you what region is presented in the markup view. The selection of the feature table row serves as a convenient way to see what the sequence of the match is, and to quickly hide unwanted matches to the feature consensus. The markup view allows you to see the consensus match in the context of surrounding sequence. Click on a *Hind III* site in the **Graphical View** to generate a Markup view of the site.

Using the Markup View

Clicking on a feature in the **Graphical View** causes the sequence flanking that feature to be loaded into the **Markup View**. This view provides a convenient way to examine features at the nucleotide sequence level. Notice that each feature occurring in the view appears as a box around the matched sequence. The feature name and position of the match start are displayed as a label to the box. Click on the label to the *Hind* III boxes that appear in the markup view. This results in a repositioning of the red arrow, as well as highlighting the match's row in the **Feature Table**. Click on any base in the **Markup View**. This results in a repositioning of the arrow, and selection of that base in the **Sequence Display**. Another feature of the markup view is that it relates information about the genes that are annotated on the sequence. DNA that is not associated with a transcription unit appears in the markup view as black letters. DNA that encodes a non-coding region of a transcription unit appears as white letters on a gray background for the strand upon which the gene resides. Bases that code for a protein portion of a transcript appear as the same color as the coding exons of that transcript.

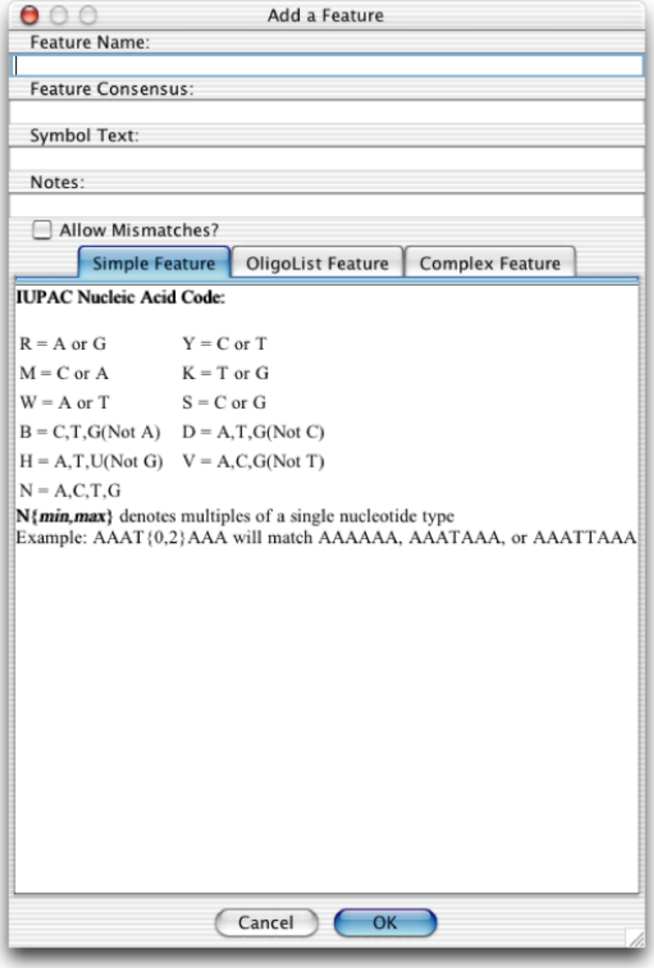
Using the Graphical View to select sequence and create a Markup View

Another way to activate the markup view for a region is to drag out a box of sequence in the graphical view. This operation will generate a **Markup View** of the boxed region, and will select the boxed sequence in the **Sequence Display**. Press the mouse button in the graphical view and drag the mouse across a region of the sequence. Experiment with generating a markup view, and using it to see features and transcription unit details. When you box a region that overlaps an exon that is not on the line of the DNA, that "off the line" exon will be annotated in the **Markup View**. When you drag out a box that will cover more than 5kb, a Markup View is not generated, and instead a button is presented that allows you to see the Markup View. The reason for suppressing this view in large sequences is that it can take a long time to generate the view, and there is a cost in speed to maintain such a large Markup View.

Adding a New Feature to the Sequence

Now that we have added *Hind* III sites to the sequence, we can see how a 2.1 kb *Hind* III fragment could encompass the mγ transcription unit. The next step that was done was to take a 1.2 kb *Ecl*136 II-*Hind* III fragment from the 2.1 kb *Hind* III fragment. Unfortunately our GenePalette restriction library does not include this site, so we are going to have to look it up and add it directly to the sequence. Go to the **Feature** menu, and select **Add Feature**. Type Ecl136 II into the field labeled **Feature Name**, and type GAGCTC into the field labeled **Feature Consensus**. If you are already attached to the idea of using text symbols for features in the graphical view, you can type some text into the **Symbol Text** field ('Ecl' or something). The notes field is optional, so you don't have to worry about putting anything into it. Click OK in the dialog, and you can see that there is an *Ecl*136 II site right between the two *Hind* III sites. If you drag a box between the upstream *Hind* III and *Ecl*136 II sites, you can see that this distance is ~1240 bp by looking at the **Selected (bp)** data field in the top portion of the sequence display.

If you really want to save the *Ecl*136 II site, go to the **Libraries** Menu, and click **Add to Library from History**. From this menu item, you can add any site that has been added to a sequence during the session to a library. Choose the *Ecl*136 II site, and then choose a library to add it to.



IUPAC Nucleic Acid Code:

R = A or G	Y = C or T
M = C or A	K = T or G
W = A or T	S = C or G
B = C,T,G(Not A)	D = A,T,G(Not C)
H = A,T,U(Not G)	V = A,C,G(Not T)
N = A,C,T,G	

N{*min,max*} denotes multiples of a single nucleotide type
Example: AAAT{0,2}AAA will match AAAAAA, AAATAAA, or AAATTAAA

Figure 5. Dialog for adding a feature directly to a sequence.

Trimming a Sequence

Now that we have a specific region of mγ that we are interested in, it would be nice to narrow our search so that we can focus on this gene. To do this, select a box which includes both the mγ locus and the *Hind* III/*Ecl*136 II fragment. Go to the **Sequence** menu, and select **Trim Sequence to Graphical Selection** via the Trim Sequence sub-menu (Or press command-T on a Mac, or Control-T on other platforms). A new GenePalette window is created which contains just the sequence that was selected in the box. The old window is also still there in case you wanted to use it (You will want to keep this window around for the third tutorial).

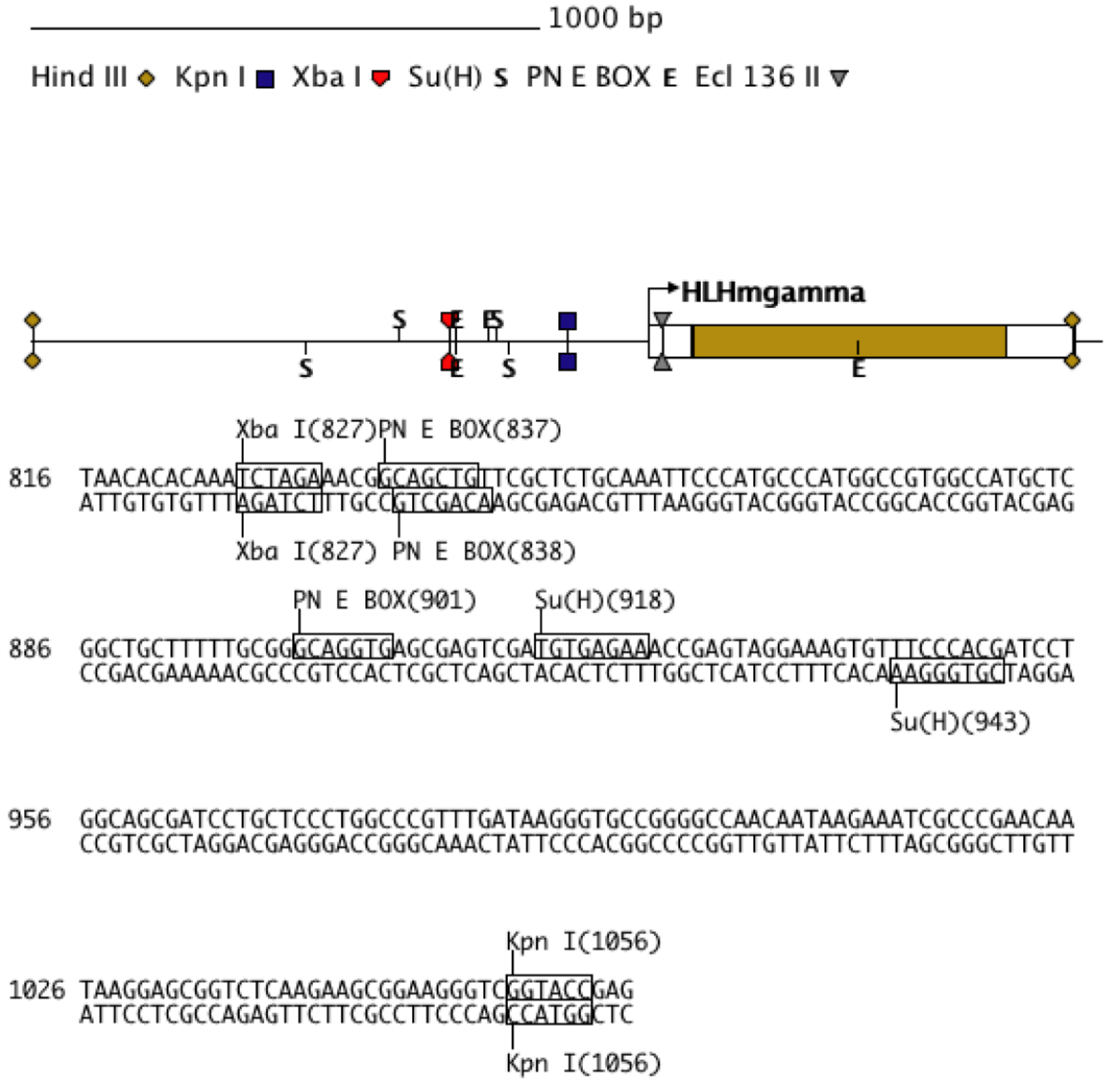
Completing the enhancer analysis of mγ

The enhancer described by Nellesen et al. was contained in a *Kpn* I/*Xba* I subfragment. Add *Kpn* I and *Xba* I from the restriction library, just as we have done for *Hind* III. Finally you can see the small piece of DNA that was used for the mγ reporter gene. As described in the text, this fragment contains binding sites for both proneural basic Helix-Loop-Helix (bHLH) activators and a transcription factor Suppressor of Hairless (Su(H)). Included in the **Tutorial Library** are binding site consensuses for these two (labeled "Su(H)" and "PN E BOX"). Add these features from the tutorial library to the sequence: go to the **Libraries** menu, and click the **Add Feature From Library** menu-item. You can add multiple features from multiple libraries at the same time by clicking them. Now you can see that the *Xba* I/*Kpn* I enhancer fragment contains two high affinity binding sites for Su(H) and two binding sites for proneural bHLH activators (there are 3 in the display, but if you look closely, one site is a semi-palindromic match).

Exporting Images of the current view

GenePalette allows you to export images of both the graphical display and the markup view in JPEG, PNG, or PostScript format (Figure 6). These formats were selected to cover the wide range of uses for an output image. The JPEG/PNG images can be a great way to quickly demonstrate a genomic feature for a lab-meeting presentation, or something to email to a collaborator. The Postscript image is extremely useful for custom-editing your image for presentations, posters, or even publication purposes. The PostScript format produced by GenePalette can be opened in graphical editors such as Adobe Illustrator, and every object in the image can be modified. From the **File** menu, click on **Export Graphical View** and select from **Export PNG** or **Export Postscript**. Supply a filename for your image. The file selection dialog always first points to a directory called Images under the main application directory. This is a convenient place to keep images created by GenePalette. Open the image you exported in a graphical program to make sure it worked. To export the markup view of the *Xba* I/*Kpn* I enhancer fragment, the first thing to do is to drag out a selection box surrounding the *Xba* I and *Kpn* I sites.

This will create a markup view of the boxed region. Once the desired section is marked up, go to the **File** menu, and click the **Export Markup View** submenu, select from **Export JPEG** or **Export Postscript,** and choose a file to export. Figure 6 shows what both JPEG images look like. This concludes the first tutorial. The next tutorial will highlight other main features of GenePalette, using the first sequence loaded (Nf1 – mα).



```
           Xba I(827)PN E BOX(837)
816  TAACACACAAATCTAGAAACGGCAGCTGTTCGCTCTGCAAATTCCCATGCCCATGGCCGTGGCCATGCTC
     ATTGTGTGTTTAGATCTTTGCCGTCGACAAGCGAGACGTTTAAGGGTACGGGTACCGGCACCGGTACGAG
           Xba I(827) PN E BOX(838)

               PN E BOX(901)    Su(H)(918)
886  GGCTGCTTTTTGCGGGCAGGTGAGCGAGTCGATGTGAGAAACCGAGTAGGAAAGTGTTTCCCACGATCCT
     CCGACGAAAAACGCCCGTCCACTCGCTCAGCTACACTCTTTGGCTCATCCTTTCACAAAGGGTGCTAGGA
                                                                 Su(H)(943)

956  GGCAGCGATCCTGCTCCCTGGCCCGTTTGATAAGGGTGCCGGGGCCAACAATAAGAAATCGCCCGAACAA
     CCGTCGCTAGGACGAGGGACCGGGCAAACTATTCCCACGGCCCCGGTTGTTATTCTTTAGCGGGCTTGTT

                                 Kpn I(1056)
1026 TAAGGAGCGGTCTCAAGAAGCGGAAGGGTCGGTACCGAG
     ATTCCTCGCCAGAGTTCTTCGCCTTCCCAGCCATGGCTC
                                 Kpn I(1056)
```

Figure 6. GIFs exported from GenePalette showing the graphical view of the mγ transcription unit, and the marked up view of the Xba I/Kpn I enhancer fragment

Tutorial 2: Creating gel-shift oligos for a mammalian enhancer

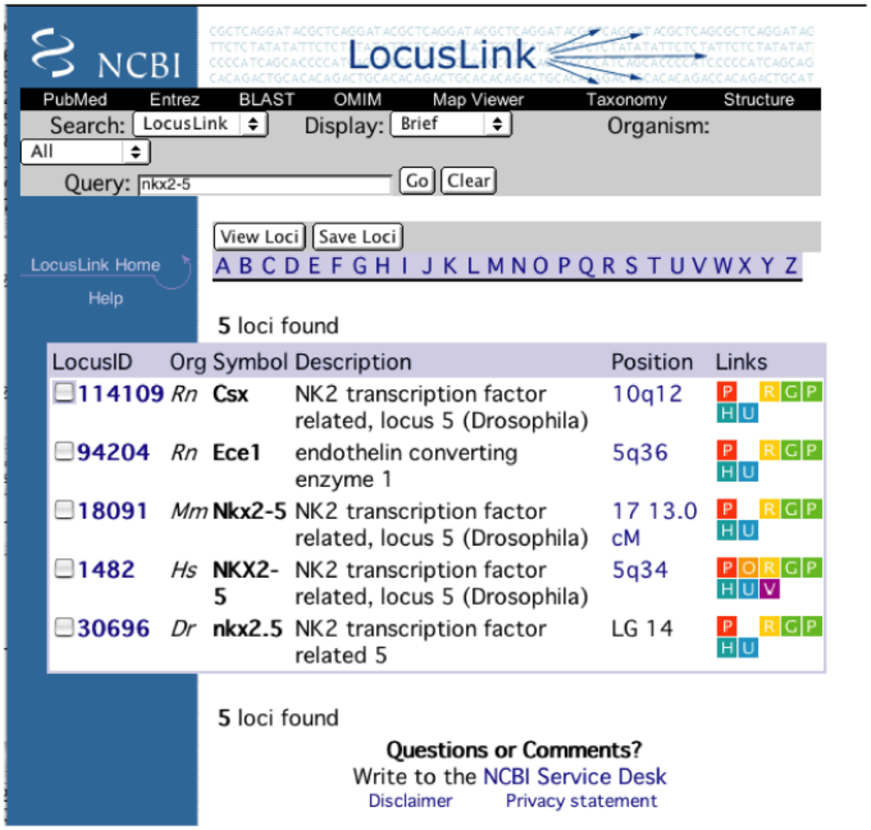
One typical operation when investigating the regulation of an enhancer is to create oligonucleotides that span predicted binding sites for use in an electrophoretic mobility shift assay. The high level of interconnectivity between the interface components of GenePalette makes it a natural for this task. In this tutorial example, we will use GenePalette to access a well-studied enhancer of the mammalian Nkx2-5 gene, and create oligos to test some GATA binding sites. During this tutorial, the user will learn how to access mammalian genome sequences through both GenBank and Ensembl. Also the user will learn how to precisely use the integrated interface to traverse from graphical representation to primary sequence.

- A copy of the sequence downloaded from GenBank is included in the Sequences directory under the main GenePalette directory. If you are not connected to the internet when doing the tutorial, you can go to the **File** menu and select **Open Sequence**. Select the file named tutorial2.seq
- Detailed information on accessing mammalian genomes will be covered in Chapter 2

Loading a Mammalian Gene: Symbols and Data Sources

When using GenePalette, it is very important to know the acknowledged symbol for the gene of interest. For mammalian genomes, the best way to find a gene symbol as it will appear in GenBank is to search for your gene in LocusLink at NCBI (http://www.ncbi.nlm.nih.gov/LocusLink/) (Figure 7). In our case, we will be looking at the gene Nkx2-5, which is the same symbol that is used in GenBank records.

There are two options for loading in sequence and annotations. In the first tutorial, we learned how to load sequence in from GenBank. The alternate route is to load your sequence from an external source that can output a GenBank Flat File, such as Ensembl (http://www.ensembl.org/). As you get used to GenePalette, you will find that each option has its strengths and weaknesses. GenBank can be slow to access, but you might find that you like the annotation better. Ensembl is fast, and sometimes it has better annotation, but symbols are referenced by long gene identifiers like “ENSMUSG00000015579”. For the sake of completeness, we will look at both the GenBank and Ensembl version of the sequence.


NCBI
LocusLink
PubMed Entrez BLAST OMIM Map Viewer Taxonomy Structure
Go Clear
View Loci Save Loci
A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
LocusLink Home
Help
Questions or Comments?
Write to the NCBI Service Desk
Disclaimer Privacy statement

Search: LocusLink Display: Brief Organism: All

Query: nkx2-5

5 loci found

LocusID	Org	Symbol	Description	Position	Links
□114109	*Rn*	Csx	NK2 transcription factor related, locus 5 (Drosophila)	10q12	P R G P H U
□94204	*Rn*	Ece1	endothelin converting enzyme 1	5q36	P R G P H U
□18091	*Mm*	Nkx2-5	NK2 transcription factor related, locus 5 (Drosophila)	17 13.0 cM	P R G P H U
□1482	*Hs*	NKX2-5	NK2 transcription factor related, locus 5 (Drosophila)	5q34	P O R G P H U V
□30696	*Dr*	nkx2.5	NK2 transcription factor related 5	LG 14	P R G P H U

5 loci found

Figure 7. LocusLink query at NCBI is the best way to find the accepted symbol for a gene, as well as what chromosome it is on.

Loading a mouse gene through GenBank

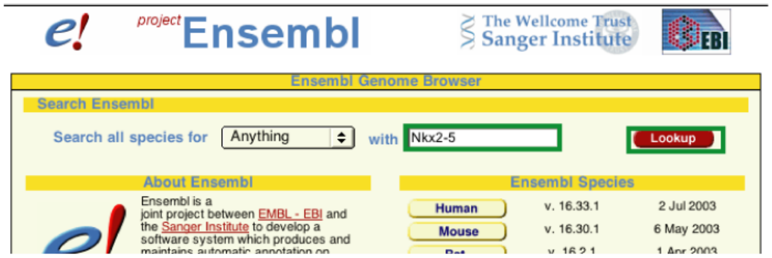
To load the Nkx2-5 gene through GenBank, click on **Entrez Nucleotide Query** from the **GenomeTools** menu. Type “nkx2-5 mus chromosome” into the Entrez query dialog. We added the word “chromosome” to the end of our Entrez query because this is an easy way to narrow our search to the chromosomal contig sequences we want to use. At the time of writing, the “nkx2-5” Entrez search yields 86 records. a mouse genomic contig on chromosome 17, and a human genomic contig on chromosome 5, consistent with the chromosomal positions seen in Fig 7. Select the sequence that is titled “Mus Musculus strain C57BL/6J chromosome 17…” by clicking the checkbox in the leftmost column of the selection table. The program loads genes on this mouse sequence in exactly the same way that the fly mγ region was loaded.

Once the genes are loaded, the gene selection table automatically highlights the row containing the Nkx2-5 gene (because it was the first word of the Entrez Query we used). Click the checkbox for Nkx2-5, and one gene upstream and downstream. Select the maximum range for upstream and downstream sequences (just click **OK** in the range selection dialog). The nucleotide sequence of the selected region is then downloaded, and a new GenePalette window is generated. In the next section, we will go through the same process using the Ensembl database.

Loading a gene through Ensembl

Ensembl (http://www.ensembl.org) is a popular annotation database of particular utility to mammalian genome users. In many instances the annotation by Ensembl is ahead of the annotation used at GenBank. When designing the external compatibilities for GenePalette, we placed highest priority on a system that would support as many genomes as possible. For most purposes, GenBank seemed to be the very best option: one can query by gene name/symbol, GenBank offers access to all public genomes (including bacterial, viral and yeast), and it is all provided through a single location (Entrez Nucleotides) with a single format (the GenBank Flat File). Neither Ensembl, nor the Distributed Annotation System (DAS) seemed to meet these pivotal criteria. However, a helpful GenePalette user mentioned that Ensembl can export regional annotations in GenBank Flat File format.

To load Nkx2-5 from the Ensembl database (Figs 8 – 9), go to the Ensembl website (http://www.ensembl.org) using your favorite web browser. Type "nkx2-5" into the search text field (Fig 8). Next, a list of all matching entries in the database will be shown (Fig 8). The lists are given in alphabetical order by organism. Go down to the matches from within the Mus musculus Gene Index. Click on the hyperlink labeled ENSMUSG00000015579. You will now be taken to a page GeneView page for the mouse Nkx2-5 gene (Fig 8). Go to the bottom row of the Ensembl Gene Report Table and click the hyperlink marked "Export gene data in EMBL, GenBank, or FASTA. This hyperlink brings you to the ExportView page (Fig 9) that allows you to customize a regional export. The ID field in this web form is already filled in with the gene ID (ENSMUSG00000015579). Just below the ID field, is a text-field that allows you to designate the flanking base-pairs to download. Enter 200000 into the "Show context of:…." field. In the middle of the page, are a bunch of export options. Select Export as GenBank, and in the checkboxes below, check the box for Gene Information. Once these fields are filled in, click the **Export** button. A new page is brought up that contains a GenBank flat file of the 20kb upstream and downstream of Nkx2-5 (Fig 9). Copy the whole GenBank flat file into your copy buffer, and then go to the **File** menu in GenePalette, click on the **New Sequence** submenu, and select the **GenBank Flat**



Ensembl Genome Browser

Search Ensembl

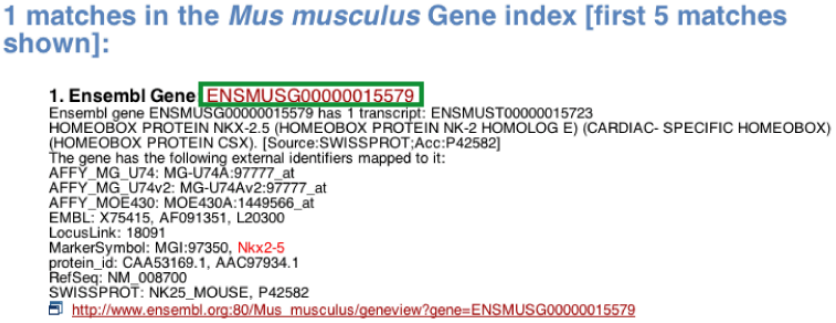
Search all species for Anything with Nkx2-5 Lookup

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on

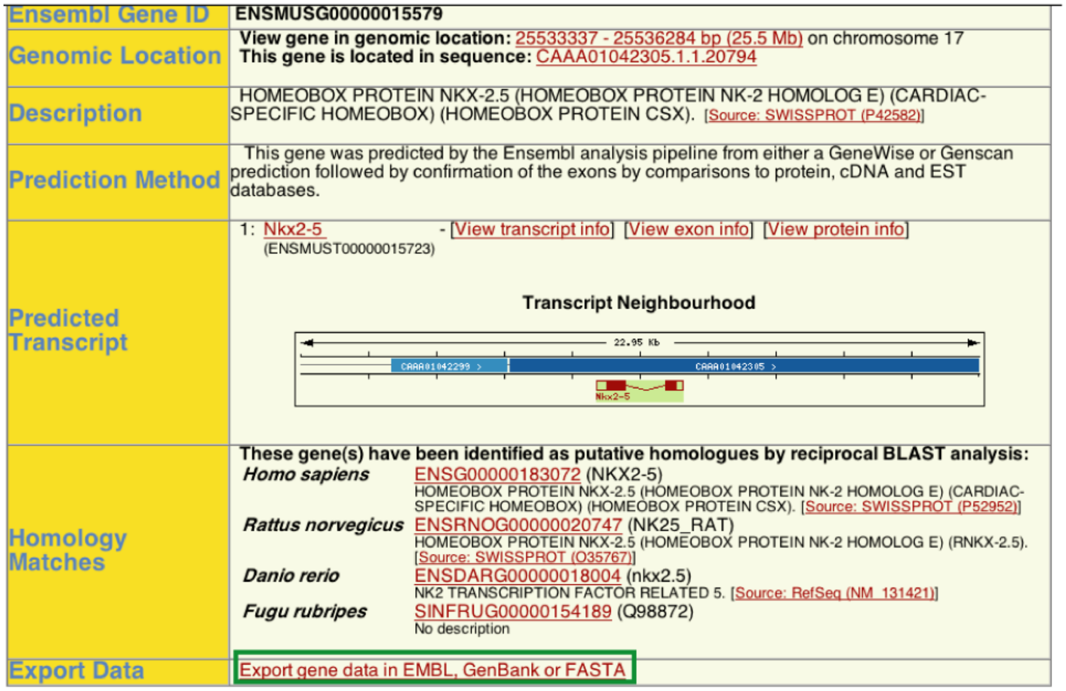
Ensembl Species

Human	v. 16.33.1	2 Jul 2003
Mouse	v. 16.30.1	6 May 2003
Rat	v. 16.2.1	1 Apr 2003



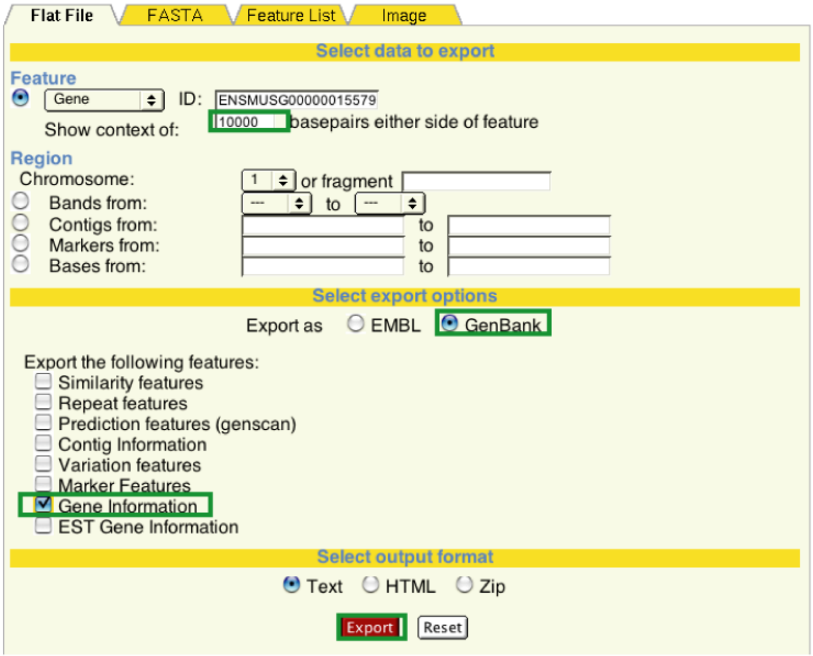
1 matches in the *Mus musculus* Gene index [first 5 matches shown]:

1. Ensembl Gene ENSMUSG00000015579
Ensembl gene ENSMUSG00000015579 has 1 transcript: ENSMUST00000015723
HOMEOBOX PROTEIN NKX-2.5 (HOMEOBOX PROTEIN NK-2 HOMOLOG E) (CARDIAC- SPECIFIC HOMEOBOX) (HOMEOBOX PROTEIN CSX). [Source:SWISSPROT;Acc:P42582]
The gene has the following external identifiers mapped to it:
AFFY_MG_U74: MG-U74A:97777_at
AFFY_MG_U74v2: MG-U74Av2:97777_at
AFFY_MOE430: MOE430A:1449566_at
EMBL: X75415, AF091351, L20300
LocusLink: 18091
MarkerSymbol: MGI:97350, Nkx2-5
protein_id: CAA53169.1, AAC97934.1
RefSeq: NM_008700
SWISSPROT: NK25_MOUSE, P42582
http://www.ensembl.org:80/Mus_musculus/geneview?gene=ENSMUSG00000015579



Ensembl Gene ID	**ENSMUSG00000015579**
Genomic Location	**View gene in genomic location:** 25533337 - 25536284 bp (25.5 Mb) on chromosome 17 **This gene is located in sequence:** CAAA01042305.1.1.20794
Description	HOMEOBOX PROTEIN NKX-2.5 (HOMEOBOX PROTEIN NK-2 HOMOLOG E) (CARDIAC-SPECIFIC HOMEOBOX) (HOMEOBOX PROTEIN CSX). [Source: SWISSPROT (P42582)]
Prediction Method	This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases.
Predicted Transcript	1: Nkx2-5 - [View transcript info] [View exon info] [View protein info] (ENSMUST00000015723) **Transcript Neighbourhood**
Homology Matches	**These gene(s) have been identified as putative homologues by reciprocal BLAST analysis:** ***Homo sapiens*** ENSG00000183072 (NKX2-5) HOMEOBOX PROTEIN NKX-2.5 (HOMEOBOX PROTEIN NK-2 HOMOLOG E) (CARDIAC-SPECIFIC HOMEOBOX) (HOMEOBOX PROTEIN CSX). [Source: SWISSPROT (P52952)] ***Rattus norvegicus*** ENSRNOG00000020747 (NK25_RAT) HOMEOBOX PROTEIN NKX-2.5 (HOMEOBOX PROTEIN NK-2 HOMOLOG E) (RNKX-2.5). [Source: SWISSPROT (O35767)] ***Danio rerio*** ENSDARG00000018004 (nkx2.5) NK2 TRANSCRIPTION FACTOR RELATED 5. [Source: RefSeq (NM_131421)] ***Fugu rubripes*** SINFRUG00000154189 (Q98872) No description
Export Data	Export gene data in EMBL, GenBank or FASTA

Figure 8. Loading a gene on the Ensembl Website. Green boxes highlight buttions, links and text that should be entered to access a gene. After typing a gene name into the search textfield, click the lookup button. In the search results web page, scroll down to where Mus musculus matches were found, and click the hyperlink for the Nxk2-5 gene (ENSMUSG0000....). In the gene report page, click the hyperlink at the" Export Data" portion of the table so that you can export a GenBank Flat File.



Flat File | FASTA | Feature List | Image

Select data to export

Feature

● Gene ID: ENSMUSG00000015579

Show context of: 10000 basepairs either side of feature

Region

Chromosome: 1 or fragment

○ Bands from: --- to ---

○ Contigs from: to

○ Markers from: to

○ Bases from: to

Select export options

Export as ○ EMBL ● GenBank

Export the following features:

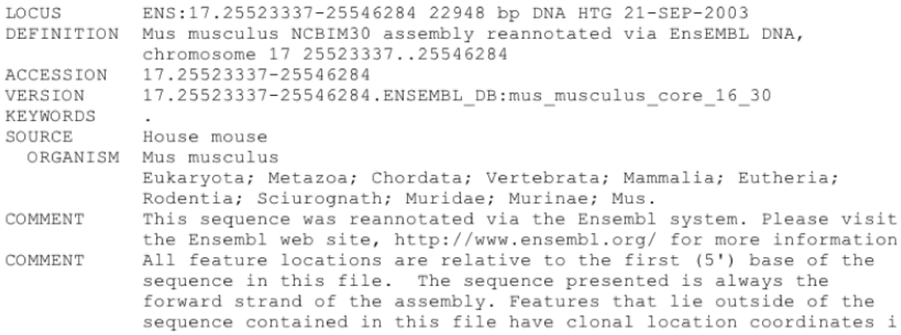
- □ Similarity features
- □ Repeat features
- □ Prediction features (genscan)
- □ Contig Information
- □ Variation features
- □ Marker Features
- ☑ Gene Information
- □ EST Gene Information

Select output format

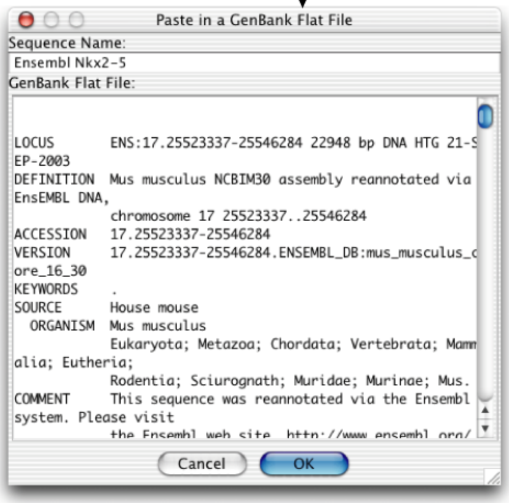
● Text ○ HTML ○ Zip

Export | Reset

Figure 9. Exporting Ensembl annotations into GenPalette using a GenBank Flat File. Green boxes represent selections, checkboxes and textfields that should be adjusted. After entering a value for flanking basepairs, selecting GenBank format, and making sure that the "Gene Information" checkbox is clicked, you are ready to export. A text page is brought up that you can copy, and then paste into the external GenBank Flat File dialog. After completing the dialog, the mouse sequence with Ensembl annotations are brought into GenePalette, and can be used like any other GenePalette sequence.



```
LOCUS       ENS:17.25523337-25546284 22948 bp DNA HTG 21-SEP-2003
DEFINITION  Mus musculus NCBIM30 assembly reannotated via EnsEMBL DNA,
            chromosome 17 25523337..25546284
ACCESSION   17.25523337-25546284
VERSION     17.25523337-25546284.ENSEMBL_DB:mus_musculus_core_16_30
KEYWORDS    .
SOURCE      House mouse
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria;
            Rodentia; Sciurognath; Muridae; Murinae; Mus.
COMMENT     This sequence was reannotated via the Ensembl system. Please visit
            the Ensembl web site, http://www.ensembl.org/ for more information
COMMENT     All feature locations are relative to the first (5') base of the
            sequence in this file.  The sequence presented is always the
            forward strand of the assembly. Features that lie outside of the
            sequence contained in this file have clonal location coordinates i
```



Paste in a GenBank Flat File

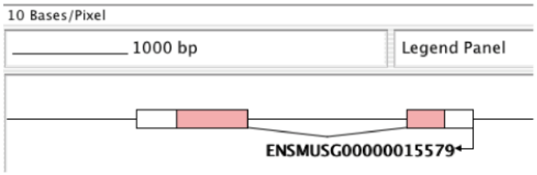
Sequence Name:

Ensembl Nkx2-5

GenBank Flat File:

```
LOCUS       ENS:17.25523337-25546284 22948 bp DNA HTG 21-S
EP-2003
DEFINITION  Mus musculus NCBIM30 assembly reannotated via
EnsEMBL DNA,
            chromosome 17 25523337..25546284
ACCESSION   17.25523337-25546284
VERSION     17.25523337-25546284.ENSEMBL_DB:mus_musculus_c
ore_16_30
KEYWORDS    .
SOURCE      House mouse
  ORGANISM  Mus musculus
            Eukaryota; Metazoa; Chordata; Vertebrata; Mamm
alia; Eutheria;
            Rodentia; Sciurognath; Muridae; Murinae; Mus.
COMMENT     This sequence was reannotated via the Ensembl
system. Please visit
            the Ensembl web site, http://www.ensembl.org/
```

Cancel | OK

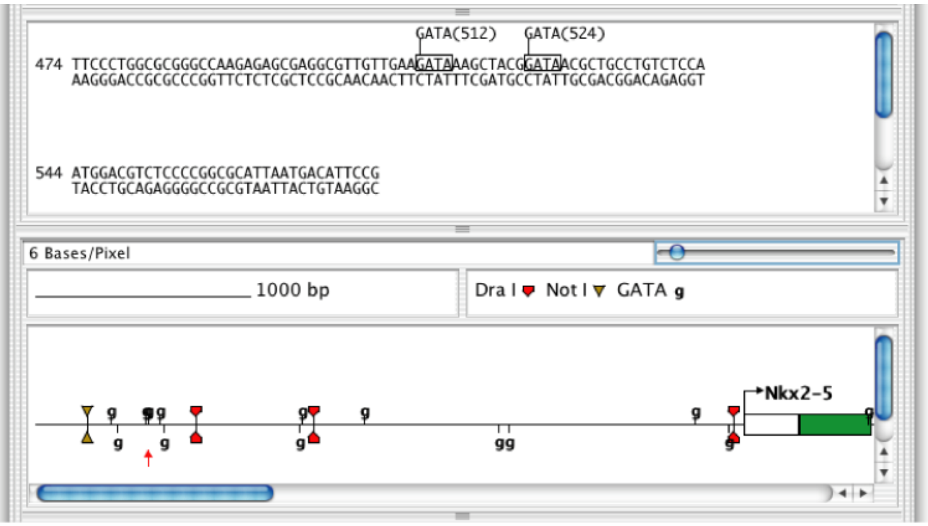


File option. A dialog will appear, which has a **Sequence Name** text field and a **GenBank Flat File** text field. Put a name in the first field so that you can identify the new window. Paste the copied file into the **GenBank Flat File** text-area, and click **OK.** The GenBank flat file will be loaded into GenePalette, just as it would have been loaded from Entrez. Note that instead of having a gene name like Nkx2-5, the gene name is "ENSMUSG0000000…..", but if you look back at the record for Nkx2-5, you will see that the number matches. Now that we have learned how to access the sequence by the two different ways, you can complete the tutorial with either the Entrez, or the Ensembl version. If you are extremely enthusiastic, go ahead and follow the steps using both sequences to see how they are the same.

Reconstruction of the Nkx2-5 distal enhancer

A studied enhancer of the Nkx2-5 gene has been narrowed down to a region between –3059 and –2554 upstream of the transcription start site[2]:

The -3059/-2554 distal *nkx-2.5* regulatory element corresponds to a *Not*I/*Dra*I restriction fragment which was linked directly to the *lacZ*/SV-40 UTR transgene in pBluescript.



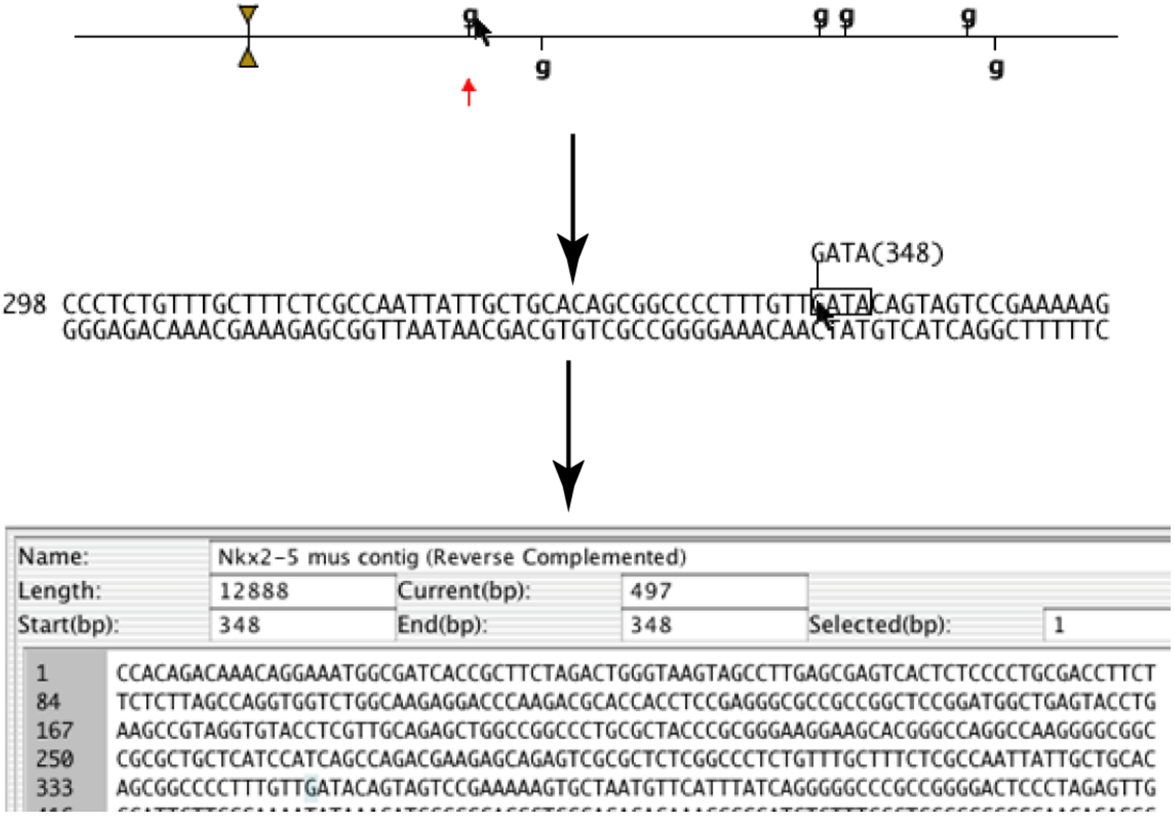
```
                                           GATA(512)   GATA(524)
474 TTCCCTGGCGCGGGCCAAGAGAGCGAGGCGTTGTTGAAGATAAAGCTACGGATAACGCTGCCTGTCTCCA
    AAGGGACCGCGCCCGGTTCTCTCGCTCCGCAACAACTTCTATTTCGATGCCTATTGCGACGGACAGAGGT

544 ATGGACGTCTCCCCGGCGCATTAATGACATTCCG
    TACCTGCAGAGGGGCCGCGTAATTACTGTAAGGC
```

Figure 10. View of the Nkx2-5 distal enhancer. GATA binding sites are viewed within a .5 kb Not I/ Dra I fragment that has enhancer activity.

[2] Searcy RD, Vincent EB, Liberatore CM, Yutzey KE. 1998. A GATA-dependent nkx-2.5 regulatory element activates early cardiac gene expression in transgenic mice. Development. 1998 Nov;125(22):4461-70.

To view the bounds of this element, add *Not I* and *Dra I* from the restriction library to the sequence (via **Add Features from Library** under the **Libraries** menu). About 2.5 kb upstream of Nkx2-5, you will see a *Dra I* site that is the beginning of this distal enhancer (Figure 10). In Searcy *et al*., several matches to the GATA consensus (simply GATA) were documented to be within this enhancer. To see where GATA binding sites are, add a GATA feature to the sequence: click **Add Features** from the **Features** menu. In the Feature dialog, type "GATA" in both the feature name field, and the feature consensus field. To make things easier to see, you can change the symbol for GATA to be the letter "g" (go to **Text Symbol** in the shape-selection **Combo-Box** of the **Feature Panel**). You can see that there are several GATA matches clustered in the *Not I*/ *Dra I* distal enhancer (Figure 10).



```
298 CCCTCTGTTTGCTTTCTCGCCAATTATTGCTGCACAGCGGCCCCTTTGTTGATACAGTAGTCCGAAAAAG
    GGGAGACAAACGAAAGAGCGGTTAATAACGACGTGTCGCCGGGGAAACAACTATGTCATCAGGCTTTTTC
```

Name: Nkx2-5 mus contig (Reverse Complemented)
Length: 12888 Current(bp): 497
Start(bp): 348 End(bp): 348 Selected(bp): 1

```
1    CCACAGACAAACAGGAAATGGCGATCACCGCTTCTAGACTGGGTAAGTAGCCTTGAGCGAGTCACTCTCCCCTGCGACCTTCT
84   TCTCTTAGCCAGGTGGTCTGGCAAGAGGACCCAAGACGCACCACCTCCGAGGGCGCCGCCGGCTCCGGATGGCTGAGTACCTG
167  AAGCCGTAGGTGTACCTCGTTGCAGAGCTGGCCGGCCCTGCGCTACCCGCGGGAAGGAAGCACGGGCCAGGCCAAGGGGCGGC
250  CGCGCTGCTCATCCATCAGCCAGACGAAGAGCAGAGTCGCGCTCTCGGCCCTCTGTTTGCTTTCTCGCCAATTATTGCTGCAC
333  AGCGGCCCCTTTGTTGATACAGTAGTCCGAAAAAGTGCTAATGTTCATTTATCAGGGGGCCCGCCGGGGACTCCCTAGAGTTG
```

Figure 11. Creating Gel Shift Oligos Part I. Clicking on a feature in the Graphical View generates a clickable markup view. By clicking on the first base of the feature match, the user can select sequence around the match because that nucleotide will be selected in the Markup View.

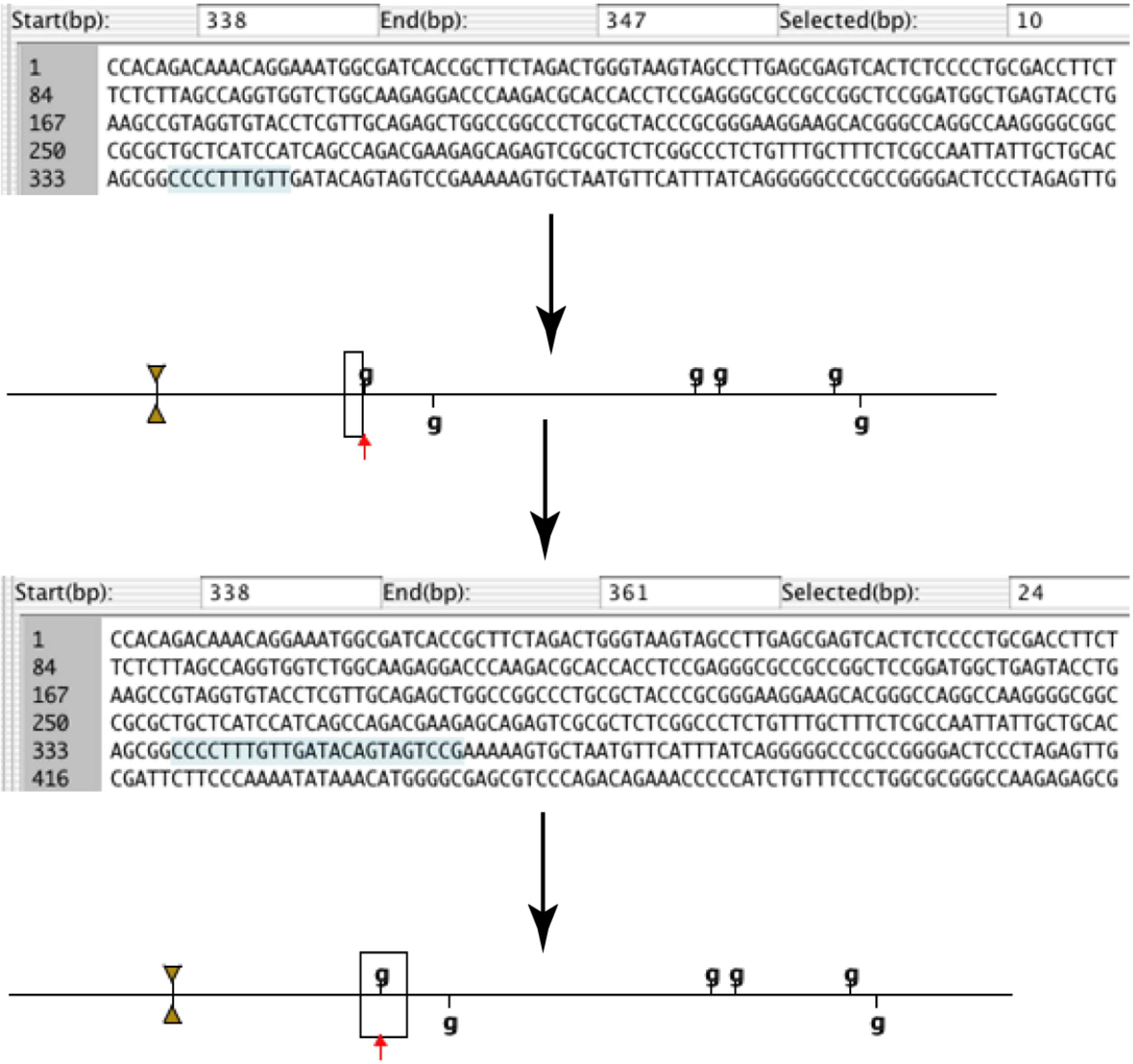
Designing the oligos

To design oligonucleotides for EMSA analysis, we usually choose sequences that are around 20-30 nucleotides in length, centered on the binding site to test. For this example, it will be easiest to design oligos that are 24 nucleotides: you want 10 nucleotides upstream, and 10 nucleotides downstream of the GATA core. In GenePalette, one can select oligos in 5 simple steps (Figures 11, 12):

1. Click on the site in the **Graphical View**: a **Markup View** is generated
2. Click on the first base of the site in the **Markup View**: That base is selected in the **Sequence Display**
3. Decide how many bases upstream you want to include (10 bp), and select that many bases upstream of the first base in the **Sequence Display**: The length of your sequence selection is displayed in the **Sequence Display**, and a graphical representation of your selection is boxed in the **Graphical View**
4. Now that you know the starting point of your oligo, select sequence downstream of the oligo to the length that you have decided (24 bp) to make your oligos: The length of your selection is displayed in the **Sequence Display**, and your selection is boxed in the **Graphical Display**
5. In the **Sequence** menu, select the option labeled "**PCR Primer Stats for selected sequence**". This will bring up a dialog box that displays the forward and reverse complement version of the sequence along with melting temperature and base composition.
6. Copy your selection (and reverse complement sequence) from the PCR primer stats dialog, and paste it into a text file/document

Designed Oligos:

```
CCCCTTTGTTGATACAGTAGTCCG
CGGACTACTGTATCAACAAAGGGG
AATGTTCATTTATCAGGGGGCCCG
CGGGCCCCCTGATAAATGAACATT
CGTTGTTGAAGATAAAGCTACGGA
TCCGTAGCTTTATCTTCAACAACG
TAAAGCTACGGATAACGCTGCCTG
CAGGCAGCGTTATCCGTAGCTTTA
CATTCCGGGTGATAGTTGCAGCTT
AAGCTGCAACTATCACCCGGAATG
AGTTGCAGCTTATCTTTCAATTAA
TTAATTGAAAGATAAGCTGCAACT
```



Start(bp): 338 End(bp): 347 Selected(bp): 10

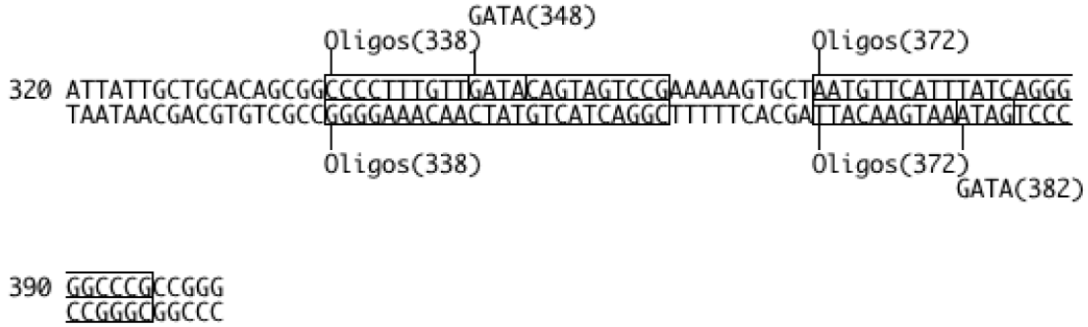
```
1    CCACAGACAAACAGGAAATGGCGATCACCGCTTCTAGACTGGGTAAGTAGCCTTGAGCGAGTCACTCTCCCCTGCGACCTTCT
84   TCTCTTAGCCAGGTGGTCTGGCAAGAGGACCCAAGACGCACCACCTCCGAGGGCGCCGCCGGCTCCGGATGGCTGAGTACCTG
167  AAGCCGTAGGTGTACCTCGTTGCAGAGCTGGCCGGCCCTGCGCTACCCGCGGGAAGGAAGCACGGGCCAGGCCAAGGGGCGGC
250  CGCGCTGCTCATCCATCAGCCAGACGAAGAGCAGAGTCGCGCTCTCGGCCCTCTGTTTGCTTTCTCGCCAATTATTGCTGCAC
333  AGCGGCCCCTTTGTTGATACAGTAGTCCGAAAAAGTGCTAATGTTCATTTATCAGGGGGCCCGCCGGGGACTCCCTAGAGTTG
```

Start(bp): 338 End(bp): 361 Selected(bp): 24

```
1    CCACAGACAAACAGGAAATGGCGATCACCGCTTCTAGACTGGGTAAGTAGCCTTGAGCGAGTCACTCTCCCCTGCGACCTTCT
84   TCTCTTAGCCAGGTGGTCTGGCAAGAGGACCCAAGACGCACCACCTCCGAGGGCGCCGCCGGCTCCGGATGGCTGAGTACCTG
167  AAGCCGTAGGTGTACCTCGTTGCAGAGCTGGCCGGCCCTGCGCTACCCGCGGGAAGGAAGCACGGGCCAGGCCAAGGGGCGGC
250  CGCGCTGCTCATCCATCAGCCAGACGAAGAGCAGAGTCGCGCTCTCGGCCCTCTGTTTGCTTTCTCGCCAATTATTGCTGCAC
333  AGCGGCCCCTTTGTTGATACAGTAGTCCGAAAAAGTGCTAATGTTCATTTATCAGGGGGCCCGCCGGGGACTCCCTAGAGTTG
416  CGATTCTTCCCAAAATATAAACATGGGGCGAGCGTCCCAGACAGAAACCCCCATCTGTTTCCCTGGCGCGGGCCAAGAGAGCG
```

Figure 12. Oligo design Part 2. As you select bases upstream of the site, you can see your selection being boxed in the graphical view. The Sequence Display reports how many bases have been selected. Once you have found the place to begin the oligo, start from this place, and select the number of nucleotides that you want to include (24 in the case of this example).

Above is a list of oligos I designed using the 5 steps. If you want to see how my oligos look on the sequence (Figure 13), you can add them as an OligoList feature. Copy all of the oligo. Go to the **Features** menu, and click **Add Feature**. In the Feature Dialog, give the Feature a name like “oligos”, and click on the tab labeled **OligoList Feature**. Paste the list of oligos into the text area contained under the **OligoList Feature** tab. Click **OK**.



```
                         GATA(348)
                  Oligos(338)                   Oligos(372)
320 ATTATTGCTGCACAGCGGCCCCTTTGTTGATACAGTAGTCCGAAAAAGTGCTAATGTTCATTTATCAGGG
    TAATAACGACGTGTCGCCGGGGAAACAACTATGTCATCAGGCTTTTTCACGATTACAAGTAAATAGTCCC
                  Oligos(338)                   Oligos(372)
                                                      GATA(382)

390 GGCCCGCCGGG
    CCGGGCGGCCC
```

Figure 13. View of designed oligos . After oligos were designed, they were added as an OligoList feature so that they could be verified.

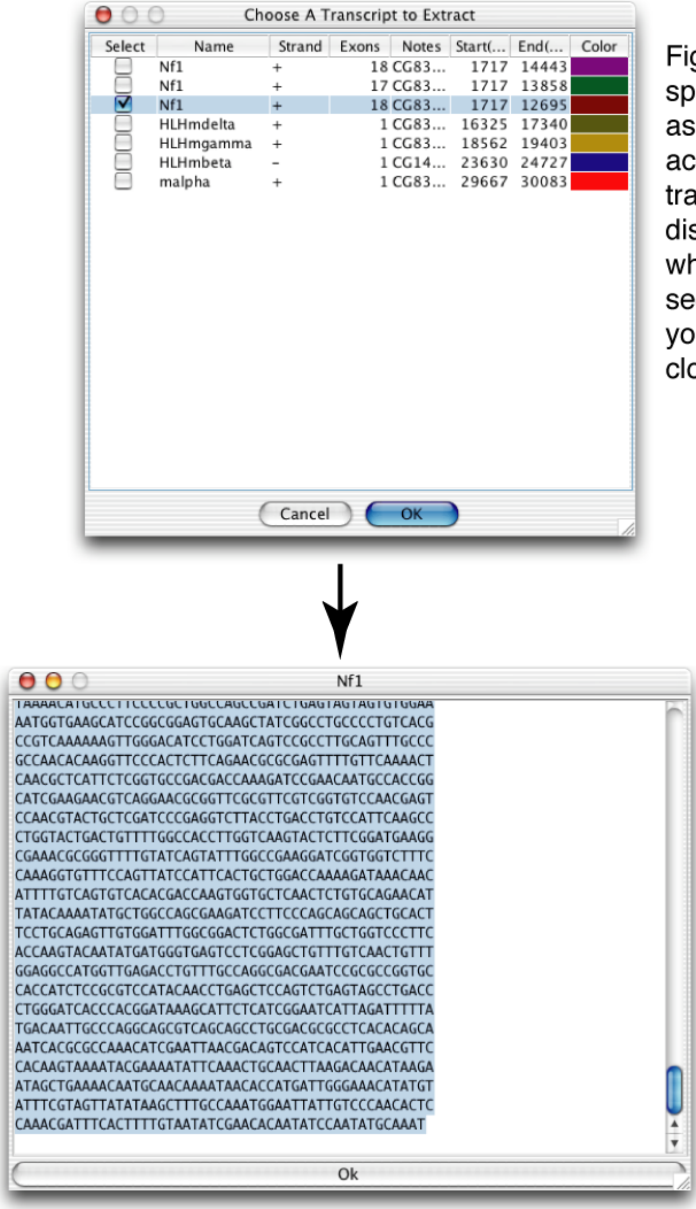
Tutorial 3: Working with Transcription Units

In most cases, a GenePalette user will be interested in using and manipulating the transcription units that are annotated on the sequence. Not only is it important to be able to see where a transcript is located, it is also necessary to be able to modify the annotation so that it matches further information you may have about a transcript. This tutorial will show you how to use the annotation access and editing features of GenePalette. We will use the portion of the Enhancer of split complex that was downloaded during Tutorial 1. A copy of this sequence is available with GenePalette in the file "tutorial.seq" in the **Sequences** directory.

- If you are starting this part from scratch, follow the instructions from **Loading a GenBank Sequence** from Part 1.
- If you are not connected to the internet, you can follow the majority of this tutorial using the "tutorial.seq" file contained in the **Sequences** directory under the main GenePalette Directory
- A glossary of all GenePalette menu-items will be given in Chapter 4

Extracting a cDNA

It is useful to have access to the spliced mRNA implied by the annotation of a transcript on genomic DNA (Figure 14). To access this information, go to the **Sequence** menu, and click **Extract Transcript cDNA Sequence**. A dialog appears that gives you a list of transcripts currently associated with the sequence. To demonstrate how this works, it is most helpful to use Nf1, which is the only gene with multiple exons in the sequence. Select one of the three Nf1 alternates from the transcript selection dialog. Note that you can easily tell between the two alternates through the last column of the table labeled Color. Click the OK button. A sequence output dialog appears that has the spliced sequence in it. From this dialog you can copy the sequence for subsequent use (Blast, strider or whatever). When you are done with the cDNA sequence, click the OK button of the dialog to close the dialog.



Select	Name	Strand	Exons	Notes	Start(...	End(...	Color
□	Nf1	+	18	CG83...	1717	14443	🟪
□	Nf1	+	17	CG83...	1717	13858	🟩
☑	Nf1	+	18	CG83...	1717	12695	🟥
□	HLHmdelta	+	1	CG83...	16325	17340	🟫
□	HLHmgamma	+	1	CG83...	18562	19403	🟨
□	HLHmbeta	-	1	CG14...	23630	24727	🟦
□	malpha	+	1	CG83...	29667	30083	🟥

```
TAAAACATGCCCTTCCCCGCTGGCCAGCCGATCTGAGTAGTAGTGTGGAA
AATGGTGAAGCATCCGGCGGAGTGCAAGCTATCGGCCTGCCCCTGTCACG
CCGTCAAAAAAGTTGGGACATCCTGGATCAGTCCGCCTTGCAGTTTGCCC
GCCAACACAAGGTTCCCACTCTTCAGAACGCGCGAGTTTTGTTCAAAACT
CAACGCTCATTCTCGGTGCCGACGACCAAAGATCCGAACAATGCCACCGG
CATCGAAGAACGTCAGGAACGCGGTTCGCGTTCGTCGGTGTCCAACGAGT
CCAACGTACTGCTCGATCCCGAGGTCTTACCTGACCTGTCCATTCAAGCC
CTGGTACTGACTGTTTTGGCCACCTTGGTCAAGTACTCTTCGGATGAAGG
CGAAACGCGGGTTTTGTATCAGTATTTGGCCGAAGGATCGGTGGTCTTTC
CAAAGGTGTTTCCAGTTATCCATTCACTGCTGGACCAAAAGATAAACAAC
ATTTTGTCAGTGTCACACGACCAAGTGGTGCTCAACTCTGTGCAGAACAT
TATACAAAATATGCTGGCCAGCGAAGATCCTTCCCAGCAGCAGCTGCACT
TCCTGCAGAGTTGTGGATTTGGCGGACTCTGGCGATTTGCTGGTCCCTTC
ACCAAGTACAATATGATGGGTGAGTCCTCGGAGCTGTTTGTCAACTGTTT
GGAGGCCATGGTTGAGACCTGTTTGCCAGGCGACGAATCCGCGCCGGTGC
CACCATCTCCGCGTCCATACAACCTGAGCTCCAGTCTGAGTAGCCTGACC
CTGGGATCACCCACGGATAAAGCATTCTCATCGGAATCATTAGATTTTTA
TGACAATTGCCCAGGCAGCGTCAGCAGCCTGCGACGCGCCTCACACAGCA
AATCACGCGCCAAACATCGAATTAACGACAGTCCATCACATTGAACGTTC
CACAAGTAAAATACGAAAATATTCAAACTGCAACTTAAGACAACATAAGA
ATAGCTGAAAACAATGCAACAAAATAACACCATGATTGGGAAACATATGT
ATTTCGTAGTTATATAAGCTTTGCCAAATGGAATTATTGTCCCAACACTC
CAAACGATTTCACTTTTGTAATATCGAACACAATATCCAATATGCAAAT
```

Figure 14. Extracting a cDNA. The spliced cDNA sequence of any associated transcript can be accessed. After selecting a transcript, the spliced cDNA is displayed in a dialog text area where you can select the sequence, and copy it. When done, you should press the "OK" button to close the dialog.

Extracting Coding and Non-coding Sequences

The **Sequence** menu contains two related submenus that allow you to extract either coding or non-coding sequence from the genomic fragment. Both submenus (**Extract Non-Coding Sequence** and **Extract Coding Sequence**) use an identical interface to select a sub-region of the genomic sequence for extraction. Once a region is selected, the type of sequence that is desired will remain as A's C's T's and G's, while the undesired sequence will be masked into N's. This type of manipulation could prove to be extremely useful for applications such as MEME searches (http://meme.sdsc.edu/meme/website/) for non-coding regulatory motifs, or for highlighting the exon structure for coding sequence. The three ways to extract coding/non coding sequence will be demonstrated. Note that untranslated regions (5' UTR, and 3' UTRs) of transcripts are considered non-coding, and will not be masked by N's in a non-coding extraction, but will be masked in a coding extraction.

Extracting by Numbers

This option displays a dialog that asks for a range to select. Type in a range (a starting base pair, and an ending bp) and see what happens. If you choose a range that overlaps an exon, you can really see how these features work. Try to extract the range of 1 to 3000 as both non-coding and coding. This range overlaps the first 4 exons of Nf1 (If you followed the tutorial instructions for Entrez queries or used the tutorial sequence provided with GenePalette).

Extracting by Gene Boundaries

Selecting this option results in the display of all transcripts on the sequence, much like the dialog displayed for selecting genes from an Entrez query. Choose the genes that you want to include in the masked sequence. Note that if you choose several genes, the outermost genes are selected as the minimum possible range that you can select. Other genes not selected, but which occur in the selected region will be masked anyways. After genes have been selected, you can then choose the range upstream and downstream of the selected genes to include in the masked sequence. **It is important to know that the maximum ranges are not the ranges to the next gene on the sequence, but will instead include the whole sequence.** On the other hand, the minimum range is the shortest distance that will include all selected genes.

Extracting by Graphical Selection

To use this option, choose the region you want to include by dragging a selection box around it in the graphical view. Then go to the menu, and choose **by Graphical Selection** from one of the extraction sub-menus. The sequence represented by the box you selected will be masked and output into the sequence output window.

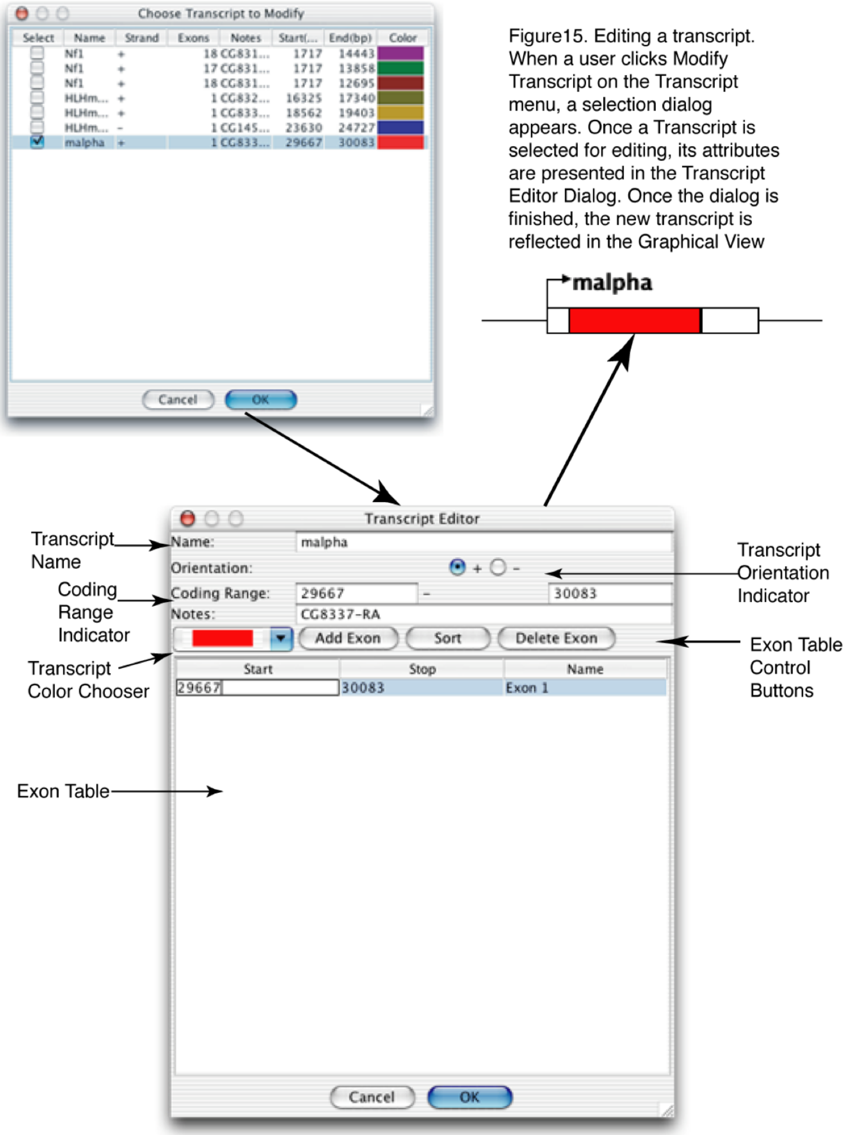
Adding and modifying transcripts

Of all the genes currently loaded into the graphical view, only mα has not had its untranslated region annotated. If we want to completely understand the organization of a transcription unit, we should make sure that the unit is completely annotated, including 5' and 3' ends if possible. Fortunately for us, someone has already described the mα transcription unit. If you go to the GenBank record AJ011140 in your web-browser, you will see that the 3' and 5' UTR of malpha have been identified:

```
FEATURES             Location/Qualifiers
     source          1..1436
                     /organism="Drosophila melanogaster"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:7227"
                     /map="96F11-14"
     protein_bind    6..14
                     /bound_moiety="Su(H)"
     protein_bind    133..141
                     /bound_moiety="Su(H)"
     protein_bind    143..151
                     /bound_moiety="Su(H)"
     protein_bind    426..434
                     /bound_moiety="Su(H)"
     gene            734..1436
                     /gene="E(spl)malpha"
     TATA_signal     734..740
                     /gene="E(spl)malpha"
     prim_transcript 767..1436
                     /gene="E(spl)malpha"
     CDS             837..1253
                     /gene="E(spl)malpha"
                     /codon_start=1
                     /product="Malpha"
                     /protein_id="CAB39164.1"
                     /db_xref="GI:4493348"
                     /db_xref="SWISS-PROT:O97178"

                    /translation="MCQQVVVVANTNNKMKTSYSIKQVLKTLFK
                    KQQKQQQKPQGSLESLESVDNLRNAQVEEAYYAEIDENAANEKL
                    AQLAHSQEFEIVEEQEDEEDVYVPVRFARTTAGTFFWTTNLQPV
                    ASVEPAMCYSMQFQDRWAQA"
     misc_feature    1283..1289
                     /gene="E(spl)malpha"
                     /note="bearded box"
     misc_feature    1396..1402
                     /gene="E(spl)malpha"
                     /note="bearded box"
     polyA_signal   1417..1423
                     /gene="E(spl)malpha"
```

If we look at the record, we can see that the primary transcript goes from 767 to 1436, and that the CDS goes from 837 to 1253. If we do some quick math we can see that the 5'UTR of mα is 70 bp (837 – 767), while the 3' UTR of mα is 183 bp (1436 – 1253). At the moment, the coding range for mα is set to the same range as the first exon. All we have to do to is change the size of ‘s first exon so that it starts 70bp further upstream, and 183 bases further downstream.



Choose Transcript to Modify

Select	Name	Strand	Exons	Notes	Start(...	End(bp)	Color
☐	Nf1	+	18	CG831...	1717	14443	🟪
☐	Nf1	+	17	CG831...	1717	13858	🟩
☐	Nf1	+	18	CG831...	1717	12695	🟥
☐	HLHm...	+	1	CG832...	16325	17340	🟫
☐	HLHm...	+	1	CG833...	18562	19403	🟨
☐	HLHm...	-	1	CG145...	23630	24727	🟦
☑	malpha	+	1	CG833...	29667	30083	🟥

Transcript Editor

Name: malpha

Orientation: ● + ○ -

Coding Range: 29667 – 30083

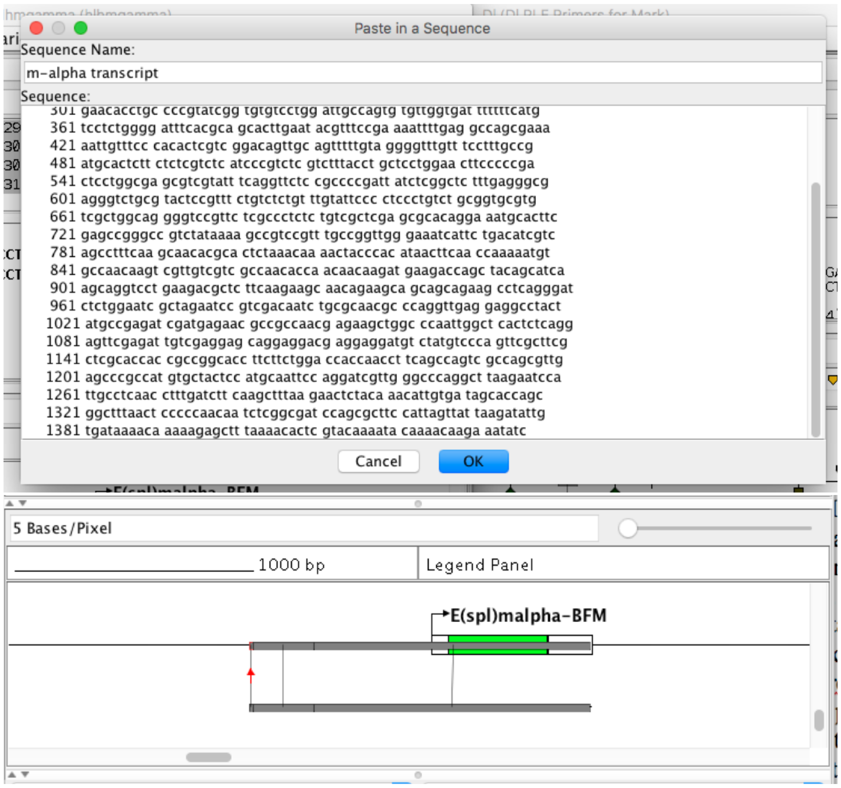
Notes: CG8337-RA

Start	Stop	Name
29667	30083	Exon 1

Figure15. Editing a transcript. When a user clicks Modify Transcript on the Transcript menu, a selection dialog appears. Once a Transcript is selected for editing, its attributes are presented in the Transcript Editor Dialog. Once the dialog is finished, the new transcript is reflected in the Graphical View

Using the new comparison function, we can directly compare this transcript to the annotated sequence, by copying the sequence of the m-alpha transcript from this GenBank record, and selecting **Add a sequence comparison** from the **Sequence** menu. Paste the sequence of the m-alpha transcript into the dialog, and press **OK.** This will launch a sequence comparison which will find the matching sequences within the genomic region of *m-alpha*. You can then click on the most 5' gray "anchor-point" box to find the further upstream nucleotide of the alignment. Using the "hlhmgamma" tutorial sequence file, this position is at basepair 13131.

Go to the **Transcript** menu, and select **Modify Transcript**. You will have to select a transcript to modify via the transcript selection dialog. Select mα, and hit OK. Now you are in the transcript modification dialog (Figure 15). This dialog is used both for adding new transcripts and modifying existing transcripts. The



Sequence Name: m-alpha transcript

Sequence:

```
 301 gaacacctgc cccgtatcgg tgtgtcctgg attgccagtg tgttggtgat ttttttcatg
 361 tcctctgggg atttcacgca gcacttgaat acgtttccga aaattttgag gccagcgaaa
 421 aattgtttcc cacactcgtc ggacagttgc agtttttgta ggggtttgtt tcctttgccg
 481 atgcactctt ctctcgtctc atcccgtctc gtctttacct gctcctggaa cttcccccga
 541 ctcctggcga gcgtcgtatt tcaggttctc cgccccgatt atctcggctc tttgagggcg
 601 agggtctgcg tactccgttt ctgtctctgt ttgtattccc ctccctgtct gcggtgcgtg
 661 tcgctggcag gggtccgttc tcgccctctc tgtcgctcga gcgcacagga aatgcacttc
 721 gagccgggcc gtctataaaa gccgtccgtt tgccggttgg gaaatcattc tgacatcgtc
 781 agcctttcaa gcaacacgca ctctaaacaa aactacccac ataacttcaa ccaaaaatgt
 841 gccaacaagt cgttgtcgtc gccaacacca acaacaagat gaagaccagc tacagcatca
 901 agcaggtcct gaagacgctc ttcaagaagc aacagaagca gcagcagaag cctcagggat
 961 ctctggaatc gctagaatcc gtcgacaatc tgcgcaacgc ccaggttgag gaggcctact
1021 atgccgagat cgatgagaac gccgccaacg agaagctggc ccaattggct cactctcagg
1081 agttcgagat tgtcgaggag caggaggacg aggaggatgt ctatgtccca gttcgcttcg
1141 ctcgcaccac cgccggcacc ttcttctgga ccaccaacct tcagccagtc gccagcgttg
1201 agcccgccat gtgctactcc atgcaattcc aggatcgttg ggcccaggct taagaatcca
1261 ttgcctcaac ctttgatctt caagctttaa gaactctaca aacattgtga tagcaccagc
1321 ggctttaact cccccaacaa tctcggcgat ccagcgcttc cattagttat taagatattg
1381 tgataaaaca aaaagagctt taaaacactc gtacaaaata caaaacaaga aatatc
```

Figure 16. Pasting the transcript sequence of m-alpha into the sequence comparison function's dialog can position the true 5' and of the gene on the DNA.

Name field holds the transcript name as it will appear on both the tab for that transcript as well as in the graphical view. The orientation selector is used to indicate the direction of transcription. The fields labeled "Coding Range" are used to indicate the start and stop of protein coding in the sequence (you can enter zeros into the two text fields for this transcript to indicate that the transcript does not code for protein). The Notes field is an optional field. The Color chooser can be used to change the transcript color, much as you would use the same chooser in the tab for the transcript.

There are 3 buttons next to the color chooser which control how the exon data table works. The **Add Exon** button will add an exon to the end of the exon table. The **Sort** button will sort the exons that have been entered, and name them according to position. Finally, the **Delete Exon** button will delete the exon that is highlighted in the table (if an exon is highlighted). You can enter values for the start and stop for each exon, which exists in the table as a separate row. You cannot edit the Name column, because the program automatically names exons by order in the direction of transcription.. The exon (using the hlhmgamma file) should now go from 13131 to 14567. Because we believe in the coding region designation, we will not alter that part. Click the OK button, and you should have an accurate version of the mα transcript!

This concludes the tutorial portion of the manual. The next chapter will go into depth about downloading sequences from GenBank.

CHAPTER 2: GenBank Access

Introduction

Pivotal to the usefulness of GenePalette is the ability to access sequences and annotation through the Internet. GenePalette uses the Entrez server at the National Library of Medicine (NLM) / National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov/Entrez/) as its portal to the world of genomic sequence. In this chapter, the methodology of GenePalette's GenBank parser is discussed, as well as tips for the most efficient use of GenBank via GenePalette.

Getting to know the GenBank Parser

To have a thorough understanding of how to most effectively use GenePalette, a modest comprehension of its GenBank interaction cascade is needed.

Creating an Entrez query

The most common starting point for access to genomic data through GenePalette will be in the form of an Entrez query, just as one would perform through the web interface at http://www.ncbi.nlm.nih.gov/Entrez/ (see Figure 1, Chapter 1) If one is familiar with the use of genomic GenBank records of their favorite organism, then this step will come naturally (There are some genome-specific tips later on in this chapter). The query that is entered is sent to the Entrez server, and results are received by GenePalette and parsed into a table for user selection.

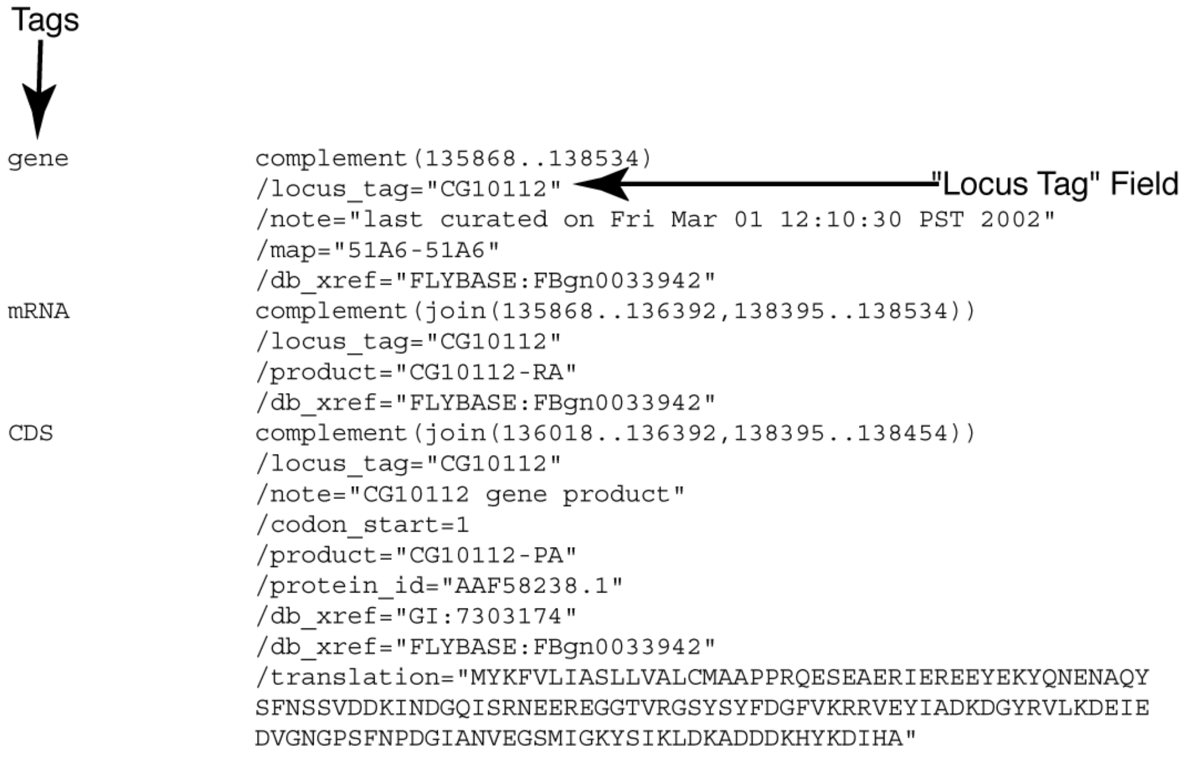
Parsing Entrez query results

The Entrez results page that is parsed is identical to the HTML that underlies the web version of Entrez. If only one sequence results, the program skips the step of asking the user to select a sequence. There are 3 pieces of data associated with each query result. The first piece of information displayed in the sequence selection dialog is whether the sequence is available remotely or locally (see the section below about local storage of sequence). The second is a description line that tells the user what is contained within the sequence. The third is a unique identifier called a gi number. A gi number is associated with only one sequence, and when that sequence record changes in any way, the gi number also changes. This unique identifier will be used in the next step of the program to request the GenBank record for the selected sequence.

It is important to be aware that GenePalette is successful at parsing 99.9% (or at least an overwhelming majority) of query results returned. However, every once in a while, there are sequences not visible in the results table, and therefore can't be selected through an Entrez query. This is because they did not parse properly. The problem is usually due to a non-standard nomenclature or record signature in the results page. At the moment, we have sidestepped this problem by adding a menu-item for loading a sequence directly by gi number (**GenomeTools** menu, **Load a Sequence by GI#**). This way, no sequence is unsearchable.

Parsing Genes

As mentioned in the previous section, once a sequence is selected, the gi number for that sequence is sent in a request to the Entrez server for the text version of the GenBank record. If accepted, GenePalette will begin to parse the data contained in this record line by line. If the GenBank record is very long (length in this case is proportional to the number of genes in the sequence), then the loading and parsing will take a long time (see the section about local sequence access below). A GenBank record is delimited by statements initiated with "keys" that are indented to the left of the annotation data (Figure 1). Each gene in a GenBank record has between one and three keyed entries for each transcription unit contained within. The simplest annotation consists entirely of CDS statements. These statements tell where coding regions begin and end in a comma-delimited list of exons. In genomes with little or no untranslated transcript (microbial, lower eukaryote), this is all you need. In more complex annotation schemes, the record will have statements that describe the CDS, and the mRNA, so that untranslated regions of a transcript are annotated. The parser collects data from mRNA and CDS statements, and cross-references them through the gene name as specified by the "locus_tag=" or "/gene=" field (Figure 1). If a sequence doesn't seem to parse well, it is probably because it is not annotated in a standard way. Please let us know if an important organism's sequences are not parsable by GenePalette.


Tags
"Locus Tag" Field

```
gene            complement(135868..138534)
                /locus_tag="CG10112"
                /note="last curated on Fri Mar 01 12:10:30 PST 2002"
                /map="51A6-51A6"
                /db_xref="FLYBASE:FBgn0033942"
mRNA            complement(join(135868..136392,138395..138534))
                /locus_tag="CG10112"
                /product="CG10112-RA"
                /db_xref="FLYBASE:FBgn0033942"
CDS             complement(join(136018..136392,138395..138454))
                /locus_tag="CG10112"
                /note="CG10112 gene product"
                /codon_start=1
                /product="CG10112-PA"
                /protein_id="AAF58238.1"
                /db_xref="GI:7303174"
                /db_xref="FLYBASE:FBgn0033942"
                /translation="MYKFVLIASLLVALCMAAPPRQESEAERIEREEYEKYQNENAQY
                SFNSSVDDKINDGQISRNEEREGGTVRGSYSYFDGFVKRRVEYIADKDGYRVLKDEIE
                DVGNGPSFNPDGIANVEGSMIGKYSIKLDKADDDKHYKDIHA"
```

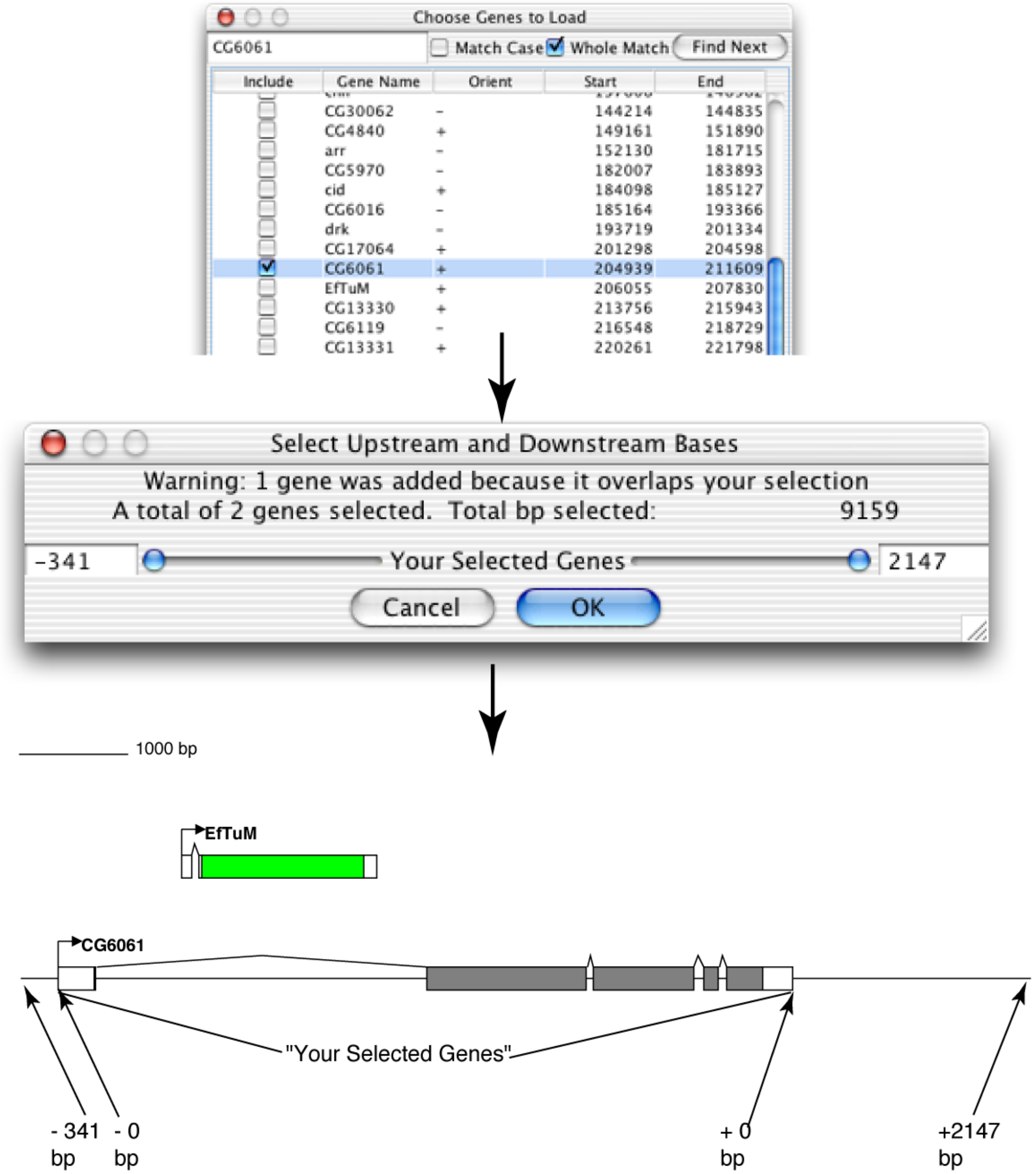
Figure 1. Gene Annotations in a GenBank record. On the left, tags are used to specify different parts of a gene's annotation. All of the gene's data are cross-referenced by the "Locus Tag" field.

Selecting a portion of sequence

Once annotation is loaded and compiled, one of two things happens. If there were no gene annotations on the sequence (unordered working draft sequence or whatever), then the user is prompted to enter a range that they would like to download. In the more common situation, the user is presented with a list of genes that were parsed from the GenBank record (see Figure 1, Chapter 1 for a picture). If the first word of your Entrez query matches a gene on the table, that row will be highlighted. You can click on the genes that you are interested in looking at. If you select a gene that overlaps another gene, all overlapping genes will then be selected (Figure 2).

Choosing upstream and downstream ranges

Once genes from the genomic sequence have been selected, the user must decide how much upstream and downstream sequence to download before the parser can move on. This decision is made with a dialog that has two slider bars, one upstream and one downstream (Figure 2). In between the two sliders is text that symbolizes the genes selected from the previous dialog. The upstream slider will go anywhere from zero bases upstream of the most upstream gene, all the way



Choose Genes to Load

CG6061 □ Match Case ☑ Whole Match (Find Next)

Include	Gene Name	Orient	Start	End
□	CG30062	-	144214	144835
□	CG4840	+	149161	151890
□	arr	-	152130	181715
□	CG5970	-	182007	183893
□	cid	+	184098	185127
□	CG6016	-	185164	193366
□	drk	-	193719	201334
□	CG17064	+	201298	204598
☑	CG6061	+	204939	211609
□	EfTuM	+	206055	207830
□	CG13330	+	213756	215943
□	CG6119	-	216548	218729
□	CG13331	+	220261	221798

Select Upstream and Downstream Bases

Warning: 1 gene was added because it overlaps your selection

A total of 2 genes selected. Total bp selected: 9159

-341 Your Selected Genes 2147

Figure 2. Gene selection and range selection. When a gene is selected which overlaps another gene, all possible overlapping genes are included in the selection and a warning is issued in the Range Selection Dialog. In this case, although CG6051 was selected, EfTuM is also included due to overlap. At the bottom, you can see the extremes that are possible in the range selection dialog. At the maximum range, all sequence to the next gene is selected, while in the minimum, no sequence above or below the gene is included.

to the first base of the next gene (Figure 2). The downstream slider works in a similar way. Above the sliders is contained information about the number of genes and base-pairs selected. If extra genes were included in the selection, a message appears in this dialog to let you know how many genes were added.

Loading the selected sequence

Finally, when the range of sequence to be loaded has been negotiated, the parser will download the sequence from GenBank, using the unique gi number. The sequence retrieval system uses the E-Utilities at Entrez for quick access to sequence fragments.

Entrez query basics

A basic Entrez query is usually going to start with a gene name or symbol. Using the gene symbol as the first word of a query has the added benefit of automatically searching the resulting gene table for that first word. Terms can be added to a query separated by spaces. For example, compare these two queries:

```
egfr
egfr drosophila
egfr AND drosophila [Organism]
```

The first query yields 39,165 sequences, while the second gives 3455, and the third generates 800 sequence results. Terms stringed together with spaces are treated as if the “AND” keyword was used to separate them. This means that results must match both terms: “egfr” and “drosophila”. You can also require that words be seen next to each other by placing a multi-word term into quotes. Another trick of the trade is to use the “[orgn]” tag to specify organism (Caenorhabditis[orgn] chromosome). An extremely useful trick is to designate a range of sequence lengths that you want to search. To see all Arabidopsis chromosomes, here is a query that gives you only 7 results:

```
Arabidopsis thaliana [ORGN] AND 10000000:50000000 [SLEN]
```

This query requires an Arabidopsis sequence that is between 1 and 50 megabases.

Genome-specific tips for Entrez queries

Each organism has its own gene nomenclature and resultantly, has its own best way to get to the sequence through an Entrez query. Here, some tips are listed for a few prominent organisms. Links to organism-specific resources are found on the GenePalette website (www.genepalette.org), and also the Entrez Genomes page (http://www.ncbi.nlm.nih.gov:80/entrez/query.fcgi?db=Genome).

Drosophila melanogaster

The easiest way to get to a fly gene is to use either its gene symbol (if it has one) or the CG number. The CG number is a gene symbol in the form of CGnnnnn (where n is a number, and CG stands for Celera Gene). Even if a gene has a symbol, it will still be associated with a CG number, which you can get from GadFly/FlyBase. If the gene has a really non-specific gene name, you can narrow the search by adding "contig", "section", or "drosophila" after the gene name.

Anopheles gambiae

Currently, the best way to access an *Anopheles* gene is to go to the mosquito search page at Ensembl (http://www.ensembl.org), type in your gene symbol, name, or drosophila CG-number into the search field, and look for hits in the *Anopheles* gene, domain or family index. This will take you to genes in Ensembl that have the ENSANG# that is used in GenBank. Hopefully the web resources for *Anopheles Gambiae* will improve soon.

Caenorhabditis elegans

The worm sequence is currently available as whole chromosome sequences. If you have the www-n (where w is a character, and n is a number) symbol, you are pretty much golden. This symbol alone gives you a limited number of query results to choose from. To access an unnamed gene, you have to find its map element/locus name (old = C36B7.1, new = XH7321). A good cross-reference for getting to this name is on the WormGenes page of the NCBI Acembly website: (http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html?worm). You can also use the Entrez-Genomes page to find these names.

Arabidopsis thaliana

Finding an Arabidopsis gene is extremely easy. Just type in the At number: AtCHRgnnnnnn (CHR = chromosome, n = number; example:At5g67540). You will definitely want to download the whole chromosome collection of *Arabidopsis* (see the Local Storage section below).

Homo sapiens/ Mus Musculus

The easiest way to find a gene in the working draft version of the human or mouse genomes is to use locus link to find approved symbols for the gene of interest. You will most likely find several matches to your symbol, since there are so many copies of things. It may be helpful to add "contig", "homo" or "mus" to the end of the query if there are a lot of results coming back. We now have downloadable genome sets for the human and mouse genomes on our webpage.

Local storage of large sequences

As genome sequence projects finish their assembly and annotation, the sequence data is transformed from small scaffolds to large whole-chromosome contiguous segments (contigs). Although this is the best representation of the sequence *in vivo*, it is not necessarily the most convenient way to look specific genes *in silico*. Large contigs with thousands of genes that span multiple mega bases take a long time to download. To evade this problem of long downloads, we have added the ability to download a sequence once, and access it as needed. To make this process very simple, we have made genome collections from prominent mature genome projects (*Arabidopsis thaliana*, *Drosophila melanogaster*, and *Caenorhabditis elegans*). We will attempt to make new collections available on our website as demand arises (if you really want us to make a collection for your genome, just ask!).

Steps to storing a sequence

If your genome is not present on the GenePalette website as a collection to download, you will have to manually download sequences to store them. Please let us know that we overlooked your favorite organism so that we can put a genome collection on the web.

The interface for storing a sequence locally is very similar to that for the usual exploration of a sequence. You go to the **GenomeTools** menu, and select the option for **Save GenBank Records and Sequences to Disk**. You will get a dialog that asks for an Entrez query. This will work in a way that is identical to the previous Entrez query dialog. However, you will want to construct an Entrez query that is very specific for the genome sequences that you would like to compile. It helps to use multiple AND arguments, as well as the [] specifiers, like [ORGN] for organism or [SLEN] for sequence length. Here are examples of the Entrez queries that use to update our genome sequence collections:

"Drosophila melanogaster"[Organism] AND chromosome [All Fields] AND ("1000000"[SLEN] : "40000000"[SLEN])

“Caenorhabditis elegans” [Organism] AND chromosome[All Fields] AND ("1000000"[SLEN] : "40000000"[SLEN])

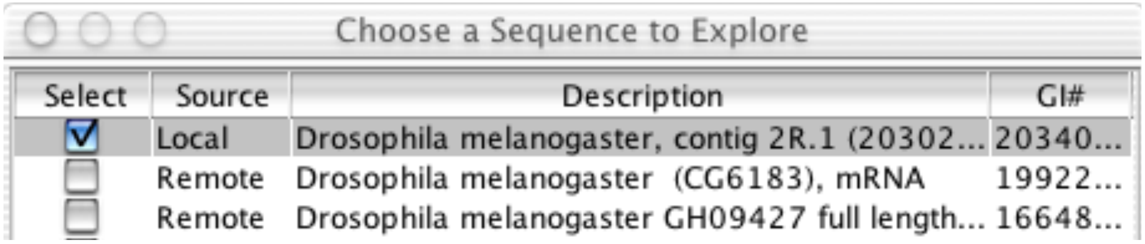
Next, you will get a result dialog that is almost identical to the result dialog for conventional sequence loading. The only difference between these two is that in this dialog, you can select multiple sequences for download. Note that the second column of the table in this dialog (**Source**) indicates whether the sequence is **Remote** or **Local**. If the sequence is **Local**, selecting this box will overwrite the previously saved sequence. You can select one sequence, or many sequences to be loaded. If you selected multiple sequences, they will be loaded one by one. If applicable, the program asks if you would like to overwrite a previously saved sequence. First, the GenBank annotation data is downloaded, compiled, and the Java object data is stored in a file with the extension ".eg" in the GenBank directory under the main program (".eg" stands for EntrezGrabber, which is the name of the device that interacts with GenBank). Next, the sequence is downloaded line-by-line and stored in a flat file with the extension ".nt" (for nucleotide). Both ".nt" and ".eg" files are named according to the gi number that they represent. This is a convenient because the sequence and annotation for every gi is required to stay the same; if changes are made, a new gi is created, and our old one will be ignored.

Steps for installing sequence collections downloaded from *www.genepalette.org*

In many cases, the genome you want to use will be available at our website. The download from the website will be much less painful than doing it through the software. First, download the zip archive for the your favorite genome from our website. Use standard unzipping (WinZip, Zipit, unzip, etc) software to decompress the archive. Each archive will decompress all of the sequences into a directory with the same name as the archive (eg "D_melanogaster_6"). Move this directory which contains all of the ".eg" and ".nt" files into the GenBank subdirectory under the main program directory. The program will search this GenBank directory, and all immediate subdirectories for stored files. This way, you can organize your stored genomes by organism name, which is really useful for the local access process described at the end of the chapter.

How to use a stored sequence

Once a sequence has either been added from a downloaded sequence collection, or manually stored through the program, there are two ways to access it. If you do the standard **Entrez Nucleotide Query (NLM)** from the **GenomeTools** menu. If a sequence that you have stored appears as a query result, the **Source** column will show that it is **Local** (Figure 3). Select the local sequence, and go through the steps for selecting a gene region as you normally would. The big differences are that now loading all of the genes on the sequence is much more quick (5-10 seconds), and loading the sequence into GenePalette is faster (1-2 seconds to access the sequence).



Choose a Sequence to Explore

Select	Source	Description	GI#
☑	Local	Drosophila melanogaster, contig 2R.1 (20302...	20340...
□	Remote	Drosophila melanogaster (CG6183), mRNA	19922...
□	Remote	Drosophila melanogaster GH09427 full length...	16648...

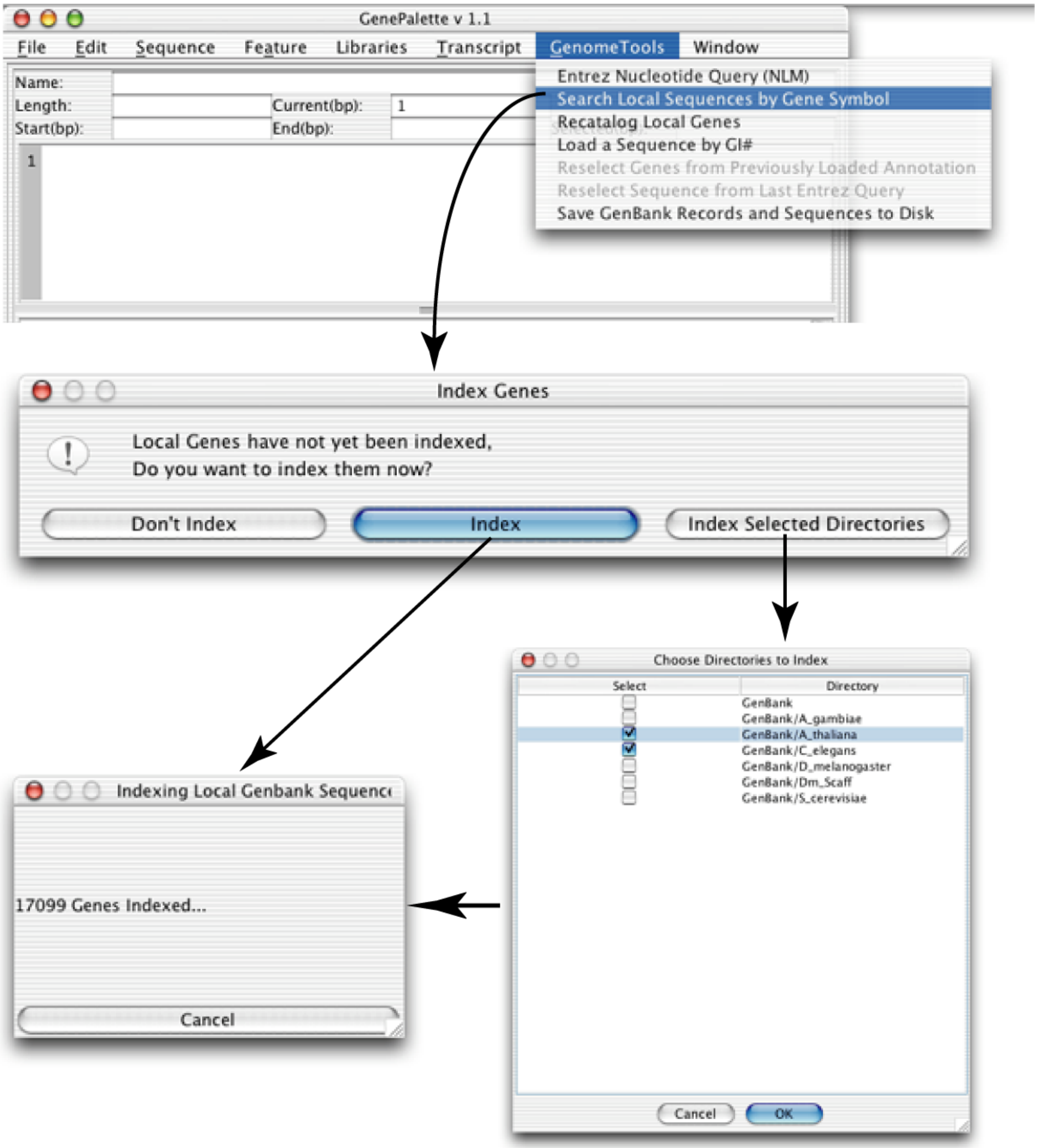
Although collections are stored in the GenBank directory, the Entrez query function is still very useful. This is because the GenBank records are indexed in several ways. Many times you can get to the right genomic contig by typing in the full name of the gene, or by typing in some other identifier. For example, if you know the CG number for a *Drosophila* gene that has a well known symbol (eg *numb* = CG3779), you will be able to get to the *Drosophila* contig for that gene with far fewer Entrez hits than the actual symbol ("numb" gives 246 hits, "CG3779" gives 5). However, it is extremely useful to access locally stored genome collections without ever having to use an internet connection. What if you are using your laptop on a plane? What if you want to browse gene models while lounging on the beach? To perform searches for genes stored locally, one must first index the genes that are stored.

Indexing genes

In order to search for genes contained within your local genome collections, you must first have indexed the sequences contained within your GenBank directory (Figure 4). Once you have added files to your GenBank directory, select **Recatalog Local Genes** under the **GenomeTools** menu. When indexing, you have two choices. You can index everything you have – all files in your GenBank directory, and any files that are in any immediate subdirectory of the GenBank directory. Alternatively, you can index a selected number of subdirectories (Figure 4). Once the indexing process is completed, you are ready to search this local index for your gene.

Searching Local Collections

Once indexing is over, you can search your local sequences for genes by the gene symbol (Figure 5). Simply go to **Search Local Sequences by Gene Symbol** under the **GenomeTools** menu. The difficulty with the local search is that you must know the gene symbol used in your locally stored sequences. The search is not case-sensitive, so you do not have to worry about proper capitalization. When you type in your symbol and press **OK**, a list of perfect matches is returned. For convenience, the subdirectory from which the match was found is presented. In Figure 5, you can see that a search for the EGFR gene yields 2 hits: one in the Dm_scaff directory, and one in the D_melanogaster directory. Once you select a sequence to explore, you are brought to a gene list that is identical to the gene list that you use during an Entrez Query. The symbol that you input is automatically searched along the gene list that is loaded, and if present, that row in the gene table is highlighted.



GenePalette v 1.1
File
Edit
Sequence
Feature
Libraries
Transcript
GenomeTools
Window
Name:
Length:
Current(bp):
Start(bp):
End(bp):
Entrez Nucleotide Query (NLM)
Search Local Sequences by Gene Symbol
Recatalog Local Genes
Load a Sequence by GI#
Reselect Genes from Previously Loaded Annotation
Reselect Sequence from Last Entrez Query
Save GenBank Records and Sequences to Disk
Index Genes
Don't Index
Index
Index Selected Directories
Choose Directories to Index
Cancel
OK
Indexing Local Genbank Sequence
Cancel

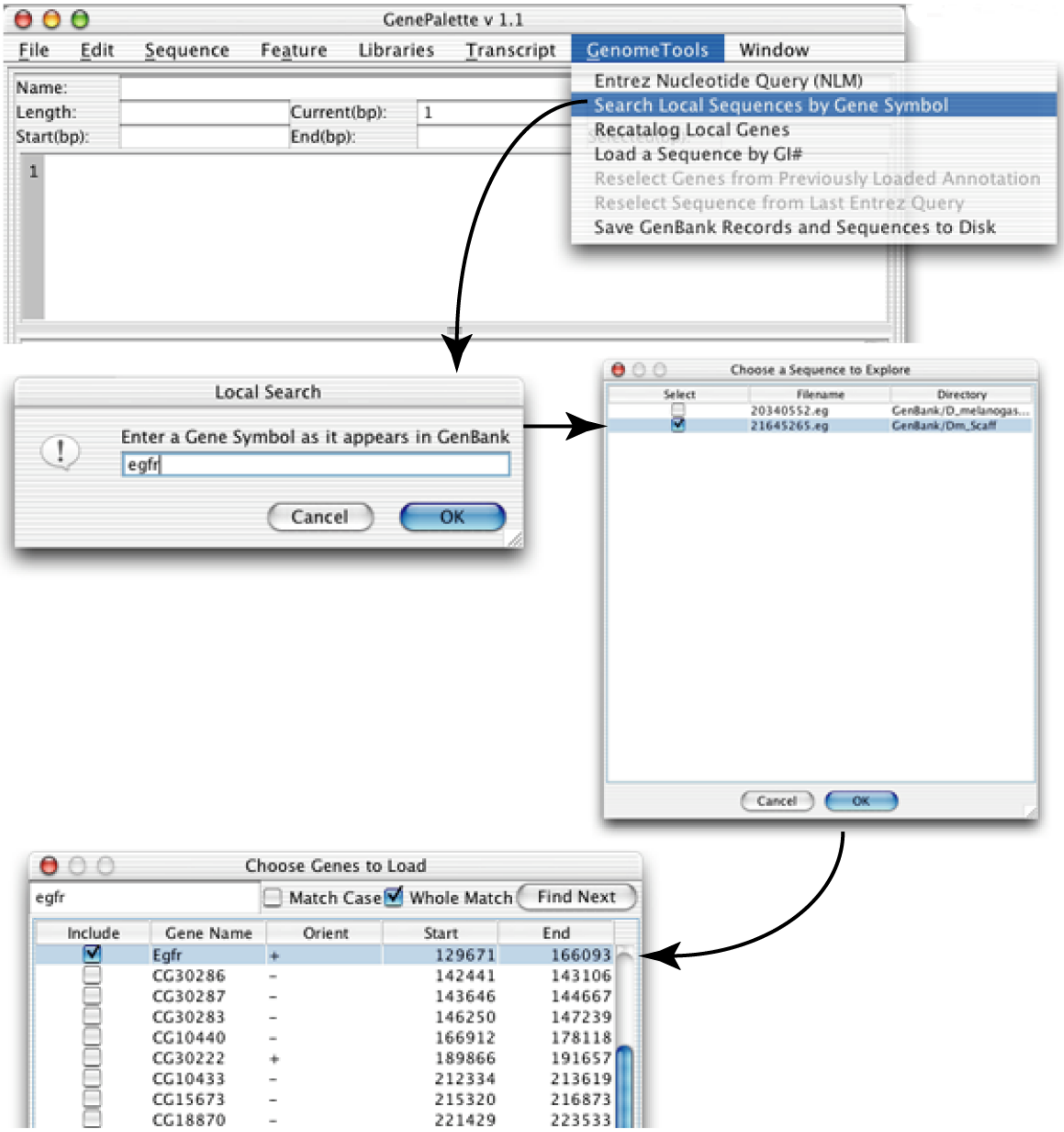
Current(bp): 1

Local Genes have not yet been indexed,
Do you want to index them now?

Select	Directory
□	GenBank
□	GenBank/A_gambiae
☑	GenBank/A_thaliana
☑	GenBank/C_elegans
□	GenBank/D_melanogaster
□	GenBank/Dm_Scaff
□	GenBank/S_cerevisiae

17099 Genes Indexed...

Figure 4. Indexing genes from local sequences. The first time you attempt to search local genes during a GenePalette session, you must index the sequences contained within your GenBank directory. You can either index all sequences, or choose selected subdirectories to index.


GenePalette v 1.1
File
Edit
Sequence
Feature
Libraries
Transcript
GenomeTools
Window
Entrez Nucleotide Query (NLM)
Search Local Sequences by Gene Symbol
Recatalog Local Genes
Load a Sequence by GI#
Reselect Genes from Previously Loaded Annotation
Reselect Sequence from Last Entrez Query
Save GenBank Records and Sequences to Disk
Name:
Length:
Start(bp):
Current(bp):
End(bp):
Local Search
Cancel
OK
Choose a Sequence to Explore
Cancel
OK
Choose Genes to Load
Find Next

Current(bp): 1

Enter a Gene Symbol as it appears in GenBank

egfr

Select	Filename	Directory
□	20340552.eg	GenBank/D_melanogas...
☑	21645265.eg	GenBank/Dm_Scaff

egfr □ Match Case ☑ Whole Match

Include	Gene Name	Orient	Start	End
☑	Egfr	+	129671	166093
□	CG30286	-	142441	143106
□	CG30287	-	143646	144667
□	CG30283	-	146250	147239
□	CG10440	-	166912	178118
□	CG30222	+	189866	191657
□	CG10433	-	212334	213619
□	CG15673	-	215320	216873
□	CG18870	-	221429	223533

Figure 5. Searching local sequences by gene symbol. Once Indexing is over, you can use this option to search through indexed directories by typing in a gene's symbol. Files with matches to your symbol will be displayed, as well as the directory that the file appears in. Once a local file has been selected, you can choose genes to load, just as in a normal Entrez Query.

CHAPTER 3: Features and Feature Libraries

Introduction

The main intent of the GenePalette creators was to provide a way to view sequence elements relative to each other and to gene annotations. To do this, one must have a way to locate and view elements of interest in a sequence. We have provided a basic set of tools for marking up a sequence through consensus descriptions called **Features**. Obviously, when performing routine analysis, it would be most efficient to have a way of re-using features instead of typing them in from memory. Resultantly, we created a set of tools for managing **Feature Libraries**. This chapter will explain the features about **Features**.

Feature Basics

A Feature is defined as any sequence element that can be described by nucleotide sequence identity. This includes transcription factor binding sites, restriction enzyme sites, SNPs, primers, microsatellites, RNA regulatory motifs, promoter elements, etc. The first thing to know is the basic nomenclature used to define a feature consensus.

IUPAC Code

Of course, you can always define a sequence using A's T's C's and G's, but GenePalette also recognizes the single letter code set by the International Union of Pure and Applied Chemistry for matching multiple bases to a single letter. This code below appears in the dialogs for adding features, and also appears as a tool tip in fields that require IUPAC code.

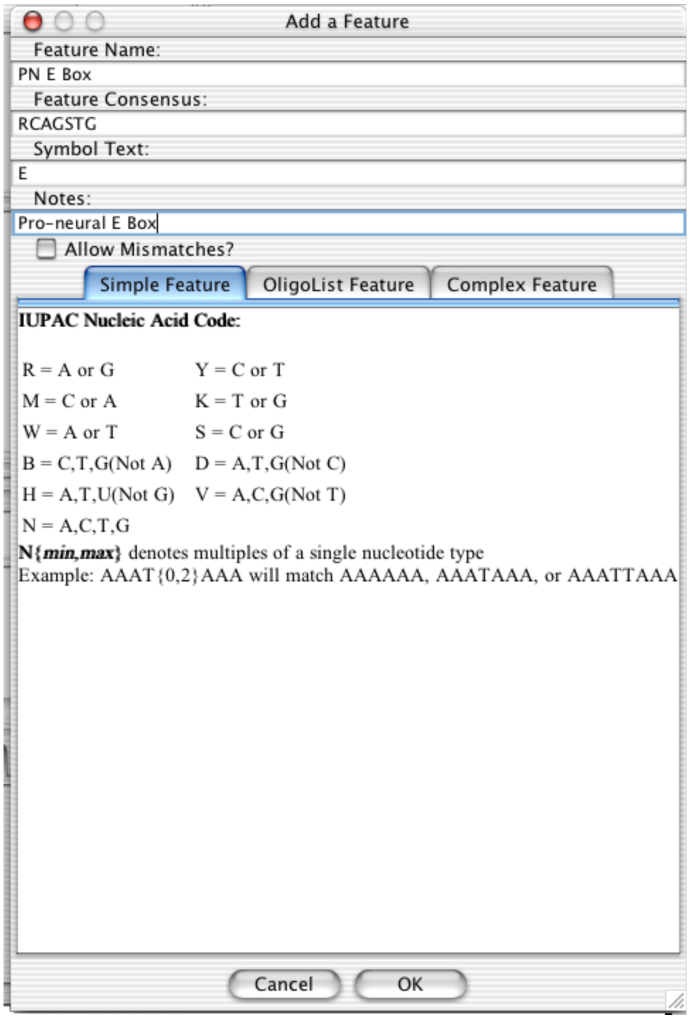
IUPAC Nucleic Acid Code:

R = A or G	Y = C or T
M = C or A	K = T or G
W = A or T	S = C or G
B = C,T,G(Not A)	D = A,T,G(Not C)
H = A,T,U(Not G)	V = A,C,G(Not T)
N = A,C,T,G	

Specifying a variable number of the same base

In addition to the IUPAC code, GenePalette also allows a Feature to search for a variable repeat of the same base. Lets say that you have a transcription factor that binds to two core sequences that are spaced by 0 to 4 nucleotides. The consensus would be expressed normally as GGGCCA N(0-4) TGGCCC. This means that there could be as few as zero, and as many as four occurrences of "N" in between the two binding cores. In GenePalette, the above consensus would be input as GGGCCAN{0,4}TGGCCC. The brackets follow the letter that will be of a variable length repeat, the first number will be the lowest number of occurrences, and the second number will be the maximum number of occurrences. Every possible number of repeats in between 0 and 4 will be allowed:

GGGCCATGGCCC
GGGCCANTGGCCC
GGGCCANNTGGCCC
GGGCCANNNTGGCCC
GGGCCANNNNTGGCCC



Feature Name:
PN E Box
Feature Consensus:
RCAGSTG
Symbol Text:
E
Notes:
Pro-neural E Box
□ Allow Mismatches?

IUPAC Nucleic Acid Code:

R = A or G	Y = C or T
M = C or A	K = T or G
W = A or T	S = C or G
B = C,T,G(Not A)	D = A,T,G(Not C)
H = A,T,U(Not G)	V = A,C,G(Not T)
N = A,C,T,G	

N{*min,max*} denotes multiples of a single nucleotide type
Example: AAAT{0,2}AAA will match AAAAAA, AAATAAA, or AAATTAAA

Figure 1. Adding a Simple Feature to a sequence

Three Types of Feature

Whenever features are being manipulated, there are 3 options for what kind of feature you can add/modify. The same feature-editing dialog is used whether you are adding a feature to a library or a sequence, or modifying a feature that exists in one of those places. The top part of the dialog contains general information about the feature which pertains to all of the 3 types of feature, while the bottom portion of the dialog contains a tabbed-pane which allows you to select which feature sub-type you want to use. If you are not editing a simple feature, this tabbed-pane will require that more data to be entered into the pane.

Simple Features

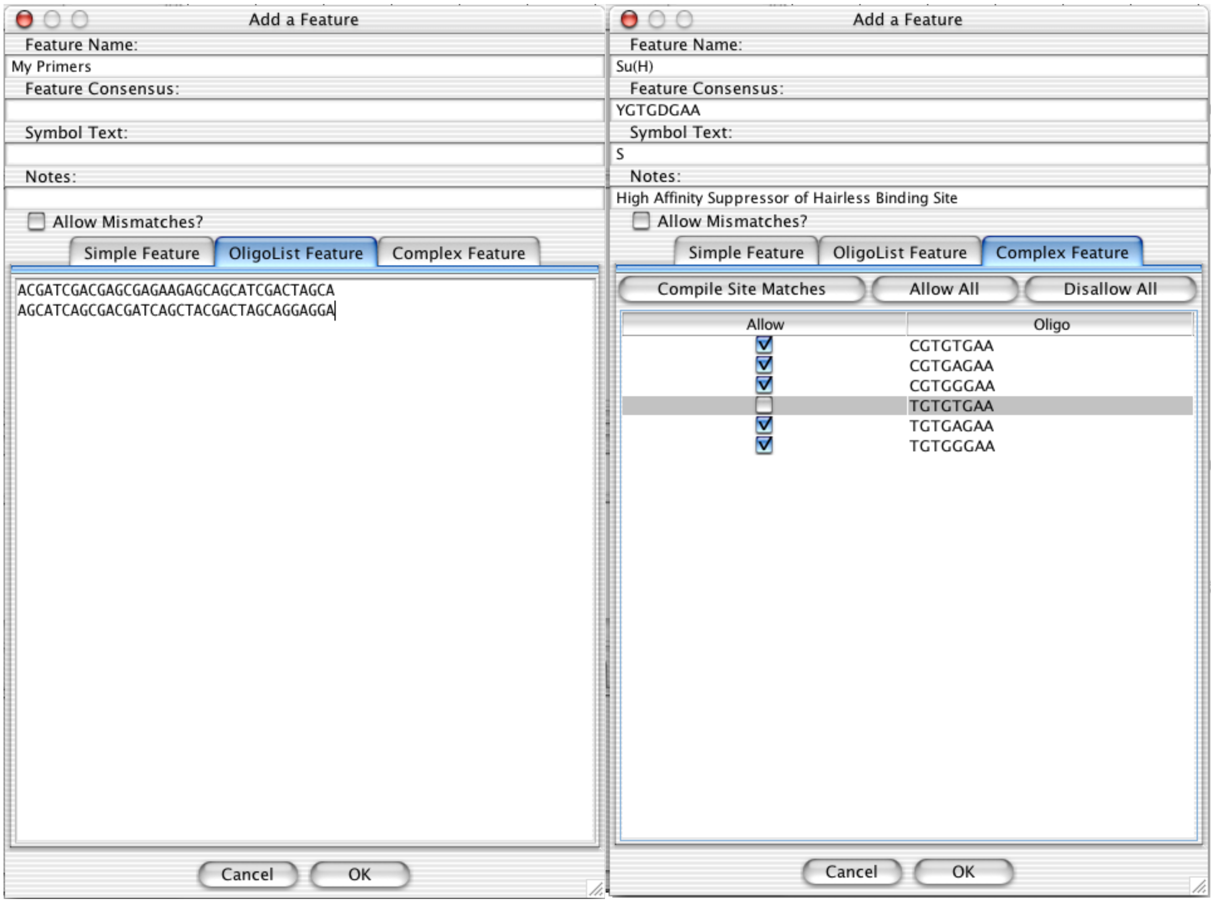
A simple feature is the most common type of feature (Figure 1). It consists of a name, a simple consensus that will be used to search the sequence, any notes that you would like to associate with the feature (optional), and text that you would like to symbolize the feature with in the graphical view (optional). You can click the mismatch checkbox to allow a single mismatch between the sequence and the consensus that you entered. For convenience the bottom panel contains a summary of the IUPAC code.

OligoList Features

This type of feature can be used to find any sequence that matches a list of oligonucleotides (Figure 2). Just simply type all of the oligos or simple consensuses you want matched into the text area under the OligoList tab. You do **not** need to fill in the **Feature Consensus** text field at the top of the dialog if you are adding an OligoList. Each oligo/consensus should have its own line in the text area (separate terms with carriage returns). An OligoList is useful in a multitude of situations. When you have a primer-pair for which you want to have the same symbol, this is a perfect choice. Another use would be if you have random binding site selection (RBSS) data, and would like to search for matching oligos rather than making a core consensus, you can put all of the selected oligos into an OligoList. This adds a different type of specificity to a binding search. The third application for this type of feature is when a feature has multiple simple definitions that cannot be grouped into one simple consensus. For example, in Markstein et al., 2002[3], the binding site for the *Drosophila* transcription factor Dorsal was searched for in the fly genome, with the following consensus:
GGGWWWWCCM
GGGWDWWWCCM

[3] Markstein M, Markstein P, Markstein V, Levine MS. 2002. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. Proc Natl Acad Sci U S A. 22;99(2):763-8.

Although these are related, you cannot make a single feature that matches all of the possible sequences implied by these two consensuses. You could use GGGW{4,5}CCM., but you would miss the version of the second consensus where the 5th position is a 'G'. You could consolidate the two versions into GGGWDW{2,3}CCM, but this would result in matches that you do not think are real Dorsal binding sites. The best solution to this problem is to make an OligoList for which each of the two Dorsal consensuses appear as separate lines in the OligoList text area.



Add a Feature

Feature Name: My Primers

Feature Consensus:

Symbol Text:

Notes:

□ Allow Mismatches?

```
ACGATCGACGAGCGAGAAGAGCAGCATCGACTAGCA
AGCATCAGCGACGATCAGCTACGACTAGCAGGAGGA
```

Add a Feature

Feature Name: Su(H)

Feature Consensus: YGTGDGAA

Symbol Text: S

Notes: High Affinity Suppressor of Hairless Binding Site

□ Allow Mismatches?

Allow	Oligo
☑	CGTGTGAA
☑	CGTGAGAA
☑	CGTGGGAA
□	TGTGTGAA
☑	TGTGAGAA
☑	TGTGGGAA

Figure 2. Adding OligoList and Complex Features to a sequence

Complex Features

The final type of feature is the Complex Feature (Figure 2). This feature allows the user to restrict a consensus to a subset of the matches implied in the IUPAC code. This is useful if you know of certain species that you are aware will match the consensus, but are not pertinent to the feature you are trying to highlight. To make a Complex feature, fill in the top of the dialog. You **must** fill in the **Feature Consensus** text field at the top. Once the top is filled, click on the

Complex Feature tab, and click the button labeled **Compile Site Matches**, this brings up a list of all possible matches implied by the IUPAC code entered into the Feature Consensus text field. Then you can click the check boxes next to each species you want disallowed. One example of a Complex feature is the consensus-binding site for Suppressor of Hairless. We know of 5 octamers that this transcription factor binds to with high affinity:

CGTGGGAA
CGTGAGAA
CGTGTGAA
TGTGGGAA
TGTGAGAA

You could put these all together with the simple consensus YGTGDGAA, but that would allow the species TGTGTGAA, which is not a high affinity binding site. In this situation, the Complex feature shines as an easy way to define a consensus that has a complex definition.

Feature Libraries

Now that we know about the various types of sequence features supported by GenePalette, it is time to learn about how you can store them and reuse them. Feature libraries hold collections of features that can be added to a sequence. All open windows operate on the same set of libraries: If you add to a library in one GenePalette window, the library in all windows is changed. All manipulations of feature libraries are conducted through the **Libraries** menu.

Included Libraries

The GenePalette software comes equipped with 2 libraries. One is the restriction library (RE.lib). This library contains 208 consensuses for commonly used restriction enzymes, listed in alphabetical order. The second included library is a small library that is used with the tutorials in chapter one.

The Libraries Directory

You may have already noticed that when you start up GenePalette, there are already 2 libraries that are available to use (see *Included Libraries* above). These libraries are available because they are in the **Libraries** subdirectory, under the main directory for the program. Any library contained in this directory will be opened automatically when the program starts up.

Creating Libraries

You can create new Libraries by going to the **Libraries Menu**. You must choose a name for the library, and a place to store the library.

Adding to Libraries

To add a feature to a library, you go to the **Libraries** menu, and click **Add Feature to Library**. You then get a selection dialog that asks you which library you want to add to. Select one of the available libraries, and hit **OK**. The normal feature-editor dialog comes up, and if you fill it in and hit **OK**, then that feature will be added to the library you selected. As soon as the feature is added, the library is saved to disk (as with any modification to a library).

Deleting from Libraries

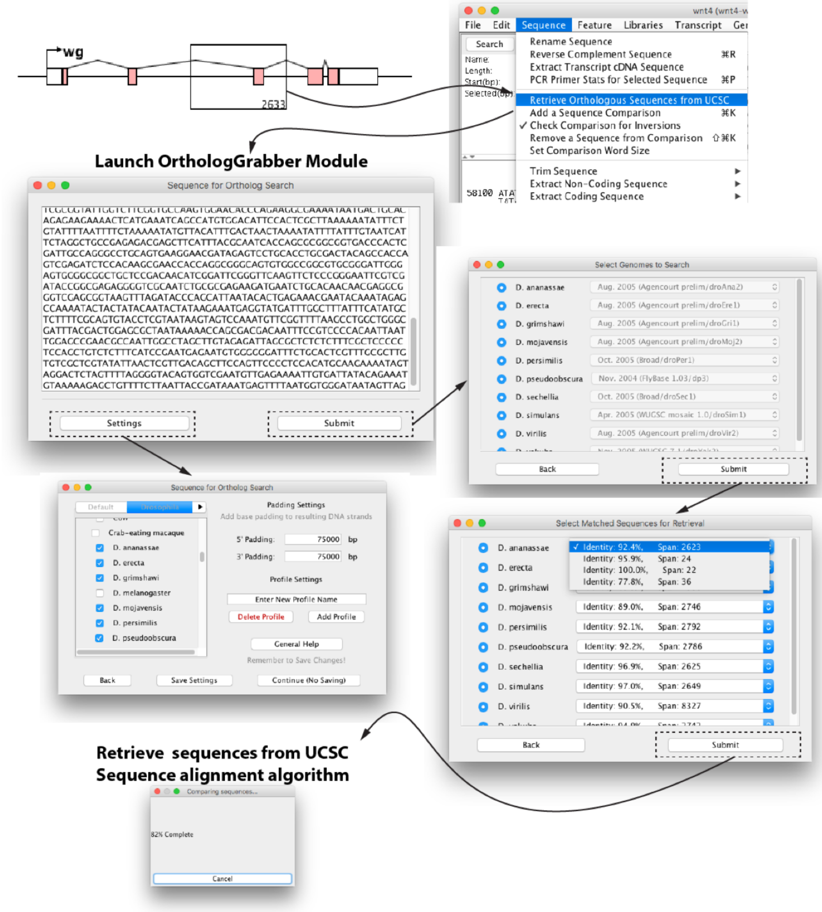
To delete a feature from a library, go to the **Libraries** menu, and click **Delete Feature from Library**. You will be asked to choose a library from which features will be deleted. Once selected, a table of all features in the library is presented, and you can click as many features in the library to delete them. **As soon as you click OK, the selected features are deleted from the library, and the library is saved to disk**. Features deleted from libraries are added to the session history, in case you really did not want to delete them. You can then add these features back to the library if you wish by clicking on the **Libraries** menu, and selecting **Add to Library from History**.

Feature History

All features added to a sequence, or modified in a sequence are added to a list of Features in the session history. Additionally, features deleted from libraries are also added to the session history for safety. You can add features from history to either the current sequence (**Feature -> Add Feature From History)**, or to a library (**Libraries->Add to Library from History**). Another feature of the history is that you can use the history to copy features from one library to another: Add a whole bunch of features from a source library to any kind of sequence. These features are then added to the session history. Then you can add to a different library from the session history, and have basically copied features between libraries. Finally, you can add any feature that is part of a saved sequence to the Feature History (**Feature -> Include all Current Features in History**). This way, if you added features and saved a sequence, these features can be added to the history, and then put in a Feature Library, or added to other sequences via the Feature History.

Chapter 4: Retrieving orthologous sequences and sequence comparisons

Often, when working with genomic sequence, particularly regulatory sequences, one wants to know whether a particular motif or set of motifs are conserved among species. This practice, known as phylogenetic footprinting is quite useful, but at also time-consuming, as it involves multiple blast searches, or database clicks to assemble a set of sequences to compare between. As of 2017, we have added a new function, the **OrthologGrabber (Figure 1),** accessible through the **Sequence menu** that allows users to BLAT sequences against the UCSC database to find homologous regions to compare.


wg
2633
wnt4 (wnt4-w
File
Edit
Sequence
Feature
Libraries
Transcript
Ger
Search
Name:
Length:
Start(bp):
Selected(bp):
Rename Sequence
Reverse Complement Sequence ⌘R
Extract Transcript cDNA Sequence
PCR Primer Stats for Selected Sequence ⌘P
Retrieve Orthologous Sequences from UCSC
Add a Sequence Comparison ⌘K
✓ Check Comparison for Inversions
Remove a Sequence from Comparison ⇧⌘K
Set Comparison Word Size
Trim Sequence
Extract Non-Coding Sequence
Extract Coding Sequence
58100 ATA
Launch OrthologGrabber Module
Sequence for Ortholog Search
Settings
Submit
Select Genomes to Search
Back
Submit
Select Matched Sequences for Retrieval
Back
Submit
Sequence for Ortholog Search
Default
Drosophila
Padding Settings
Add base padding to resulting DNA strands
Profile Settings
Enter New Profile Name
Delete Profile
Add Profile
General Help
Remember to Save Changes!
Back
Save Settings
Continue (No Saving)
Retrieve sequences from UCSC
Sequence alignment algorithm
Comparing sequences...
82% Complete
Cancel

Select Genomes to Search:

- ● D. ananassae Aug. 2005 (Agencourt prelim/droAna2)
- ● D. erecta Aug. 2005 (Agencourt prelim/droEre1)
- ● D. grimshawi Aug. 2005 (Agencourt prelim/droGri1)
- ● D. mojavensis Aug. 2005 (Agencourt prelim/droMoj2)
- ● D. persimilis Oct. 2005 (Broad/droPer1)
- ● D. pseudoobscura Nov. 2004 (FlyBase 1.03/dp3)
- ● D. sechellia Oct. 2005 (Broad/droSec1)
- ● D. simulans Apr. 2005 (WUGSC mosaic 1.0/droSim1)
- ● D. virilis Aug. 2005 (Agencourt prelim/droVir2)

Select Matched Sequences for Retrieval:

- ● D. ananassae ✓ Identity: 92.4%, Span: 2623 / Identity: 95.9%, Span: 24 / Identity: 100.0%, Span: 22 / Identity: 77.8%, Span: 36
- ● D. erecta
- ● D. grimshawi
- ● D. mojavensis Identity: 89.0%, Span: 2746
- ● D. persimilis Identity: 92.1%, Span: 2792
- ● D. pseudoobscura Identity: 92.2%, Span: 2786
- ● D. sechellia Identity: 96.9%, Span: 2625
- ● D. simulans Identity: 97.0%, Span: 2649
- ● D. virilis Identity: 90.5%, Span: 8327

Sequence for Ortholog Search (settings):

- □ Crab-eating macaque
- ☑ D. ananassae
- ☑ D. erecta
- ☑ D. grimshawi
- □ D. melanogaster
- ☑ D. mojavensis
- ☑ D. persimilis
- ☑ D. pseudoobscura

5' Padding: 75000 bp
3' Padding: 75000 bp

Figure 1. Launching the OrthologGrabber module.

Launching the OrthologGrabber Module

To begin the process of launching a sequence comparison, you first much choose what portion of the gene you would like to use to anchor your alignment. To launch the module, you must have a sequence loaded. <u>You can select the portion that will be aligned by drawing a box in the GraphicalView (Figure 1)</u>. There are many reasons that you might want to select only a subportion of the sequence for a sequence alignment:

- ♦ The bigger the sequence you BLAT, the longer everything takes
- ♦ You only need a small portion to center your alignment, as the program allows you to specify how much flanking sequence to retrieve relative to the BLAT hit.
- ♦ There is a limit to 75kb for BLAT searches imposed by the UCSC database.

If you do not box a region in the currently loaded sequence, the OG module will use the entire sequence currently loaded to search the database. Once the menu item is selected (**Retrieve orthologous sequences from UCSC** under the **Sequence** menu), a second java application window will launch (Figure 1). The first panel of this window shows the sequence that was selected to BLAT against the UCSC database. If you press the settings button, you can do several important things:

- You can choose any number of species to perform a BLAT search against
- You can save different profiles of species for rapid access to different organismal groups
- You can specify how much flanking sequence you would like to obtain adjacent to the BLAT hit
- You can save these settings so that you can use them in the future

Once you have selected species, specified the amount of flanking sequences, and (hopefully) saved your settings, you can press the "Continue" or "Back" button to return to the window that displays the sequence to be compared. Press the **Submit** button, and the program will display the next window, which gives the user the opportunity to select which version of the genome will be compared. **Your currently used settings will be used again by the program in subsequent launches of the OrthologGrabber module, so if you always are working with a particular profile of species, you should be able to simply hit "Submit" on this window without altering the settings page.** In the next window, if multiple genome versions are available, you can select them from a drop-down menu. Once you press **Submit**, the program will perform BLAT searches against these selected genomes, and will display the resulting hits that can be imported for alignment. For each species, there is a drop-down menu that lists each significant BLAT hit, displaying the percent identity and length of hit. The best hit is always automatically selected. If no excellent hit was found for a species, it will not be included in the window, and if a sub-par hit was found, you can deselect it by pressing the radio buttons on the left-hand side of the page. Clicking **Submit** will then instruct the program to retrieve the selected regions (as well as their flanking regions as specified on the settings page), and launch a new sequence comparison.

Launching sequence alignments outside of the OrthologGrabber module

The OrthologGrabber module greatly facilitates the job of finding orthologous regions to align, but it is not the only way to do this. There are many more species that have been sequenced than are available via the UCSC database. If the species you are working with is not in the UCSC database, or if you would like more control over how regions of interest are selected, go to "**Add a Sequence Comparison"** in the **Sequence** menu (or press CMD/CTRL-K). This menu-item will bring up a dialog that will allow you to paste in one or more sequences. For convenience, if pasting in just one sequence, you can give it a name in top field. Alternately, one can paste in list of one or more sequences that have FASTA descriptor lines (the descriptor is each line that starts with a '>'). If a FASTA descriptor is available, those descriptors will be set to the name of each sequence pasted in. You can paste in as many sequences as you'd like into this window. Finally, it should be noted that you can add a sequence to a comparison that already has multiple species in it already.

Sequence alignments

Once the OrthologGrabber has been launched, or sequences have been added to the comparison manually, the program will then start its sequence alignment algorithm (Figure 2). Depending on how large of a region is being compared, this may take a few minutes, and a loading dialog will appear that gives a sense of how fast or slow this is going. The algorithm is a brute-force algorithm that combs the shortest sequence for words of a specified length (default is 15bp) that appear in each sequence. Because repetitive sequences abound in non-coding regions, there are many words that will appear more than once, and these are not as useful for determining orthologous segments to visualize in alignments. For this reason, the algorithm is limited to sequences that appear only once in any given sequence being compared. Once the initial sequence comparison is done, one can change several parameters of the alignment to optimize the analysis:

- The minimum word size can be changed by selecting **Set Comparison Word Size** in the **Sequence Menu**
- You can toggle whether the algorithm will look for inverted sequences in the **Sequence Menu**
- Remove problematic sequences (typically ones that are much shorter than the rest) by selecting **Remove a Sequence from Comparison** in the **Sequence Menu**

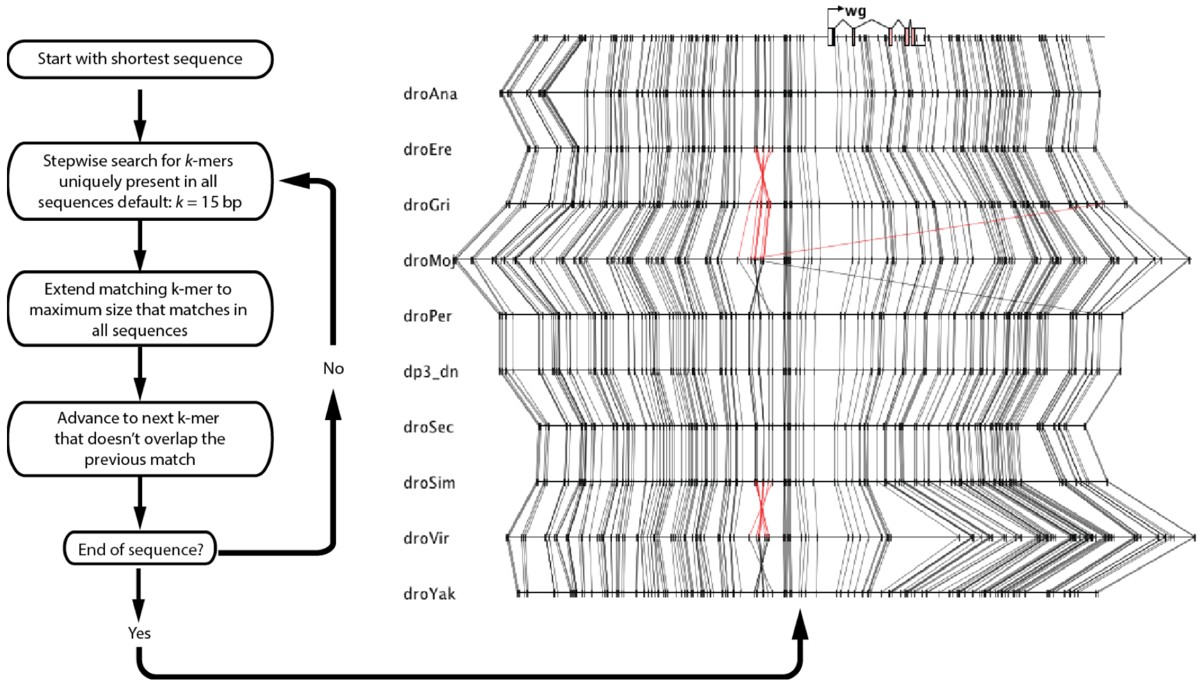

Start with shortest sequence
Stepwise search for *k*-mers uniquely present in all sequences default: *k* = 15 bp
Extend matching k-mer to maximum size that matches in all sequences
Advance to next k-mer that doesn't overlap the previous match
End of sequence?
No
Yes
wg
droAna
droEre
droGri
droMoj
droPer
dp3_dn
droSec
droSim
droVir
droYak

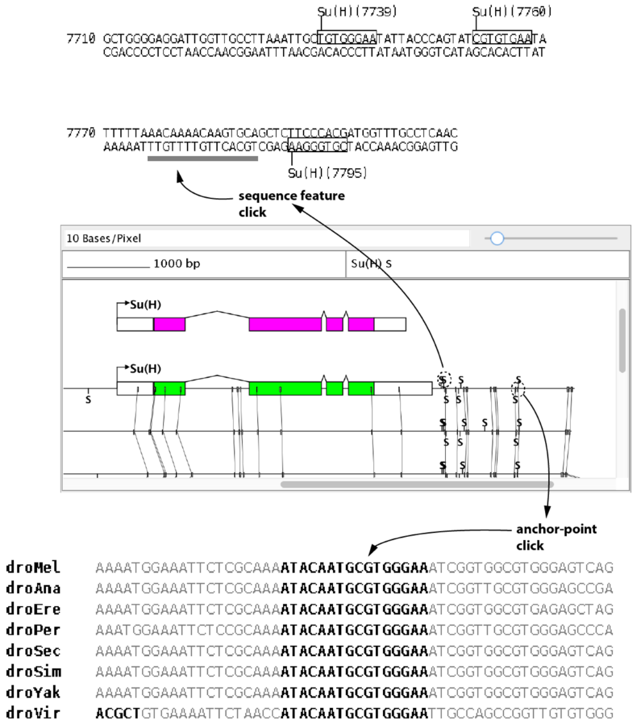
Figure 2. Alignment Algorithm

When the algorithm is finished, regions that are identical between aligned sequences show up in the graphical and markup view as gray boxes that we call **AnchorPoints**. AnchorPoints that match between species are connected by thin gray lines. If inversions are checked, the AnchorPoints are red, connected by red lines. Each added sequence is accessible in the Sequence Display through a drop-down box that lists the name of all sequences in the comparison. This will switch the sequence display to the selected sequence, in which all of the normal SequenceDisplay functions are available.

Interacting with Alignments

GenePalette provides several ways to dynamically interact with sequence alignments once they have been generated. In the graphical view, clicking on an AnchorPoint elicits several responses from the interface:

- The GraphicalView is centered upon the clicked AnchorPoint, and it is highlighted in red
- The sequence of the clicked AnchorPoint is selected and displayed in the SequenceDisplay. This action selects the correct sequence from the drop-down menu in the SequenceDisplay as well
- A MarkupView showing the aligned AnchorPoint is displayed. In this view, AnchorPoint sequences are bolded and colored in black, while non-AnchorPoint sequences are non-bold gray. Inverted AnchorPoints are highlighted in red/pink



```
7710 GCTGGGGAGGATTGGTTGCCTTAAATTGCTGTGGGAATATTACCCAGTATCGTGTGAATA
     CGACCCCTCCTAACCAACGGAATTTAACGACACCCTTATAATGGGTCATAGCACACTTAT

7770 TTTTTAAACAAAACAAGTGCAGCTCTTCCCACGATGGTTTGCCTCAAC
     AAAAATTTGTTTTGTTCACGTCGAGAAGGGTGCTACCAAACGGAGTTG
```

```
droMel    AAAATGGAAATTCTCGCAAAATACAATGCGTGGGAAATCGGTGGCGTGGGAGTCAG
droAna    AAAATGGAAATTCTCGCAAAATACAATGCGTGGGAAATCGGTTGCGTGGGAGCCGA
droEre    AAAATGGAAATTCTCGCAAAATACAATGCGTGGGAAATCGGTGGCGTGAGAGCTAG
droPer    AAATGGAAATTCTCCGCAAAATACAATGCGTGGGAAATCGGTTGCGTGGGAGCCCA
droSec    AAAATGGAAATTCTCGCAAAATACAATGCGTGGGAAATCGGTGGCGTGGGAGTCAG
droSim    AAAATGGAAATTCTCGCAAAATACAATGCGTGGGAAATCGGTGGCGTGGGAGTCAG
droYak    AAAATGGAAATTCTCGCAAAATACAATGCGTGGGAAATCGGTGGCGTGGGAGTCAG
droVir    ACGCTGTGAAAATTCTAACCATACAATGCGTGGGAATTGCCAGCCGGTTGTGTGGG
```

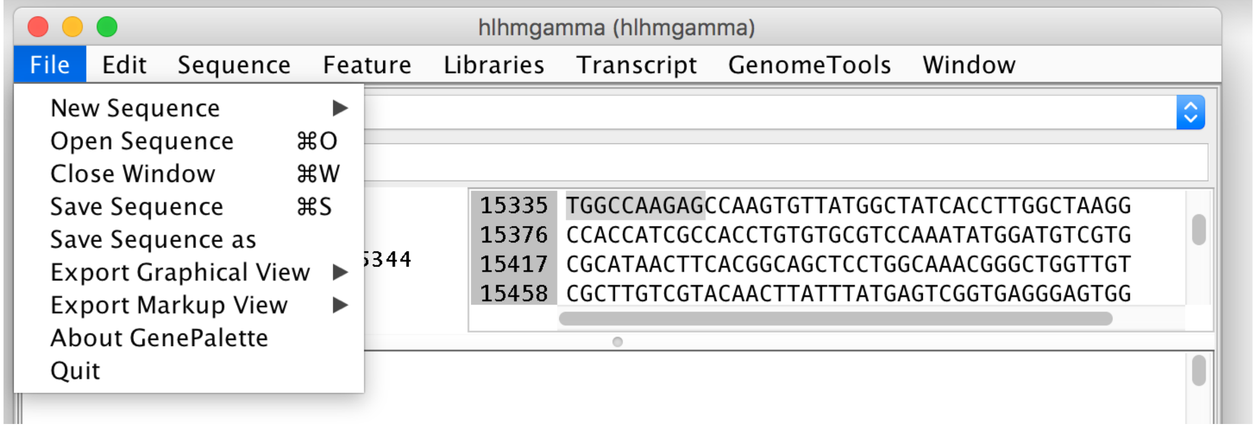
Figure 3. Interacting with alignments in the GraphicalView

Interactions with other interface components are altered as well (Figure 3). For example, dragging a box across a region in the GraphicalView will generate an altered MarkupView that displays AnchorPoints as gray boxes below the DNA-sequence. This allows one to determine whether a binding site or feature in the boxed region is contained within, or overlaps a conserved sequence region. One can also select regions in the sequence display of any sequence in the comparison, and it will be graphically displayed as a boxed region in the GraphicalView.

CHAPTER 5: Index of Menu Items

Introduction

This chapter gives a 1-2 sentence description of every available menu item in GenePalette. Screen shots are on a MAC. Shortcuts/hotkeys in these menu images will be accessible by the CONTROL button on PC and Unix platforms.



```
15335 TGGCCAAGAGCCAAGTGTTATGGCTATCACCTTGGCTAAGG
15376 CCACCATCGCCACCTGTGTGCGTCCAAATATGGATGTCGTG
15417 CGCATAACTTCACGGCAGCTCCTGGCAAACGGGCTGGTTGT
15458 CGCTTGTCGTACAACTTATTATGAGTCGGTGAGGGAGTGG
```

The File Menu

New Sequence-> Sequence Only
Displays a dialog where you can create a new sequence by typing in a sequence name, and pasting the nucleotides into a text area. In the sequence text area, all non-nucleotide characters are filtered out.

New Sequence-> GenBank Flat File
Displays a dialog where you can create a new sequence by typing in a sequence name, and pasting a GenBank Flat File into text area. GenBank Flat Files copied from the Ensembl Exportview will be parsed using this menu-item (see Tutorial 2, Chapter 1).

Open Sequence
Open a sequence that was previously saved in GenePalette.

Save Sequence
Save changes to a sequence that has been saved before, or save a previously unsaved file to disk.

Save Sequence As
Specify and save sequence under a new file name.

Export Graphical View -> Export JPEG
Exports an image of the Graphical Display in JPEG format.

Export Graphical View -> Export PNG
Exports an image of the Graphical Display in PNG format

Export Graphical View -> Export PostScript
Exports an image of the Graphical Display in Postscript format.

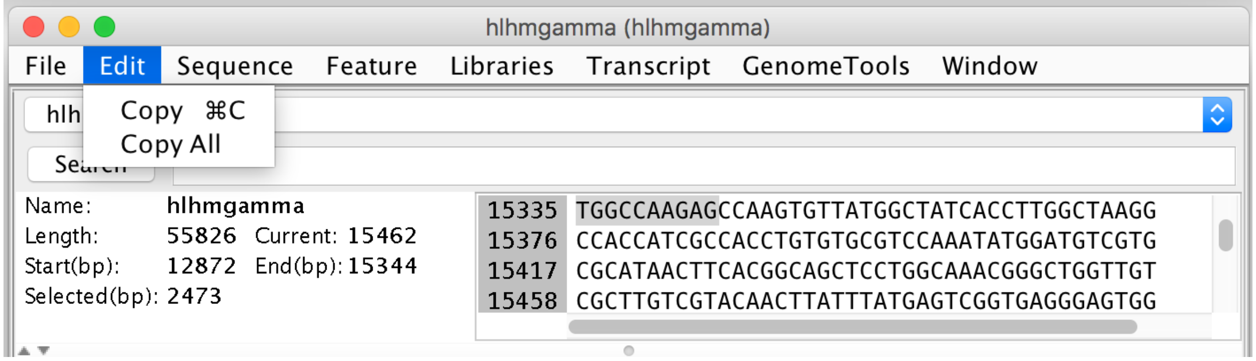
Export Markup View -> Export JPEG
Exports an image of the Markup View in JPEG format.

Export Markup View -> Export PNG
Exports an image of the Markup View in PNG format.

Export Markup View -> Export PostScript
Exports an image of the graphical display in Postscript format.

About GenePalette
Displays the About GenePalette window. The hyperlinks are real!

Quit
Closes the program. All unsaved files will be verified with the user before closing.



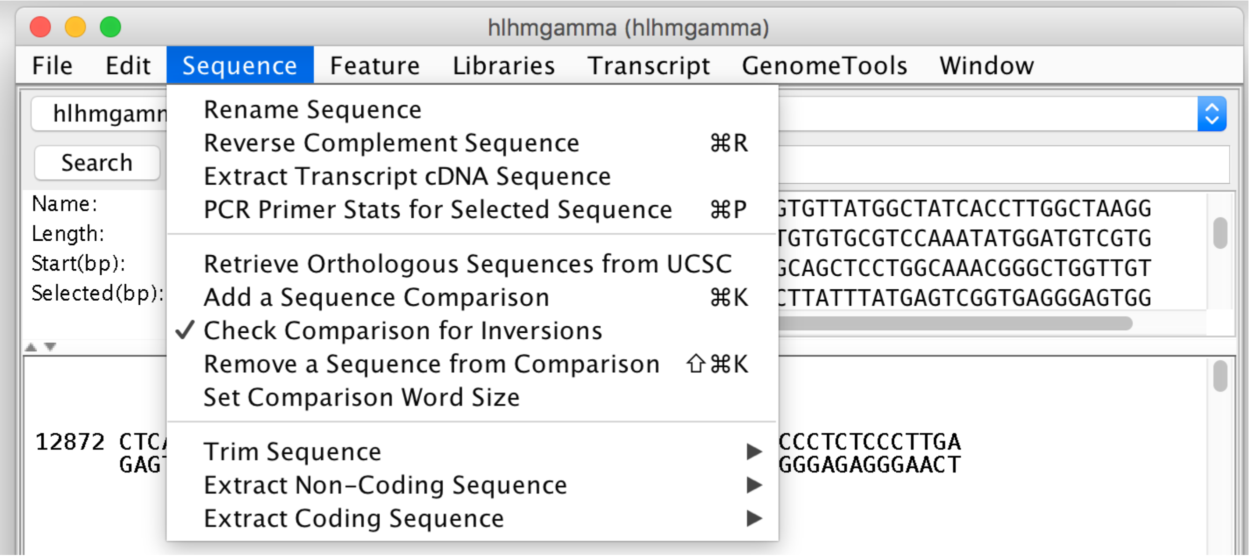
Name: hlhmgamma
Length: 55826 Current: 15462
Start(bp): 12872 End(bp): 15344
Selected(bp): 2473

```
15335  TGGCCAAGAGCCAAGTGTTATGGCTATCACCTTGGCTAAGG
15376  CCACCATCGCCACCTGTGTGCGTCCAAATATGGATGTCGTG
15417  CGCATAACTTCACGGCAGCTCCTGGCAAACGGGCTGGTTGT
15458  CGCTTGTCGTACAACTTATTTATGAGTCGGTGAGGGAGTGG
```

The Edit Menu

Copy
Copies sequence currently selected in the text area of the sequence display into the clipboard.

Copy All
Copies all of the sequence in the text area of the sequence display into the clipboard.


hlhmgamma (hlhmgamma)
File
Edit
Sequence
Feature
Libraries
Transcript
GenomeTools
Window
hlhmgamm
Search
Name:
Length:
Start(bp):
Selected(bp):
Rename Sequence
Reverse Complement Sequence ⌘R
Extract Transcript cDNA Sequence
PCR Primer Stats for Selected Sequence ⌘P
Retrieve Orthologous Sequences from UCSC
Add a Sequence Comparison ⌘K
✓ Check Comparison for Inversions
Remove a Sequence from Comparison ⇧⌘K
Set Comparison Word Size
Trim Sequence ▶
Extract Non-Coding Sequence ▶
Extract Coding Sequence ▶
GTGTTATGGCTATCACCTTGGCTAAGG
TGTGTGCGTCCAAATATGGATGTCGTG
GCAGCTCCTGGCAAACGGGCTGGTTGT
CTTATTTATGAGTCGGTGAGGGAGTGG
12872 CTC
GAG
CCCTCTCCCTTGA
GGGAGAGGGAACT

The Sequence Menu

Rename Sequence
Changes the name of the sequence in both the sequence display data area, and in the title of the window (and in the Window Menu).

Reverse Complement Sequence
Flips the sequence and transcripts and features around so that the reverse-complement strand is the strand displayed in the sequence display. This new sequence appears in a new GenePalette window, leaving the old sequence untouched.

Extract Transcript cDNA Sequence
Splices all exons together into a cDNA sequence, which is displayed in a dialog that allows you to copy and paste the sequence.

PCR Primer Stats for Selected Sequence
If a region of the loaded sequence is selected or boxed, selecting this item will display a dialog window that shows useful information for primer design. The

forward and reverse complement versions of the candidate primer region are show, along with melting temp and base composition.

Retrieve Orthologous Sequences from UCSC
Launches the OrthologGrabber module which searches the region boxed in the GraphicalView (or alternately the entire sequence) against selected species of the UCSC genome browser databse. High ranking BLAT hits can be retrieved. See Tutorial 4 in chapter 1 for details.

Add a Sequence Comparison
Brings up a dialog that allows the user to input one or more sequences (in FASTA format) that will be compared to the currently loaded sequences. T

Check Comparison for Inversions
Menu item that toggles whether sequences that are reverse complemented will be checked during the alignment algorithm. Selecting this item will launch the sequence alignment function (and associated loading dialog box). The default option is to check reverse complemented sequences.

Remove a Sequence from Comparison
Brings up a dialog allowing the user to select sequences to be removed from the current sequence alignment. Selecting this item will launch the sequence alignment function (and associated loading dialog box).

Set Comparison Word Size
Brings up dialog allowing the user to input an integer for the minimum sized word that will be searched for in the alignment algorithm. The default value for this is 15. Settings ranging from 10-20 are recommended. Smaller word sizes cause the algorithm to take longer. Selecting this item will launch the sequence alignment function (and associated loading dialog box).

Trim Sequence
Submenu that gives you three ways to trim the sequence down to a smaller size. All three ways result in a new GenePalette window that contains the shortened sequence, while the old sequence remains in its current window, unaltered.

- *by Numbers* – Allows you to trim by specifying how much sequence to remove from the front and end of the sequence. Entering 5000 into the **Front Cut** text field will remove 5000 bp from the front of the sequence. Entering 5000 into the **Rear Cut** field will result in 5000 bp removed from the end of the sequence.

- *by Gene Boundaries* – Allows the user to select a core set of genes that should be maintained in the trimmed sequence. Once those genes are selected, the user must use a slider to select the number of base pairs upstream and downstream of the genes that will be kept. The default position on the slider is to keep everything, and the user must move the sliders closer to the selected genes to trim more.

- *to Graphical Selection* – If a box is selected within the graphical view of the window (see tutorial, chapter 1), then that region will be trimmed out of the big sequence. For convenience, this item has a hotkey (CMD/CTRL-T).

Extract Non-Coding Sequence
Submenu that allows the user to mask protein-coding sequences for a portion of the currently loaded sequence in three different ways. All of these ways result in the masked sequence being presented in a text-area dialog so that the user can copy the masked sequence for use elsewhere.

- *by Numbers* – Allows the user to enter a range of bases on the current sequence to extract non-coding sequence from. This option should not be confused with the "*by Numbers*" option in the Trim submenu.

- *by Gene Boundaries* – Allows the user to select a core set of genes that should be included in the masked sequence. Once those genes are selected, the user must use a slider to select the number of base pairs upstream and downstream of the selected genes. The default position on the slider is to keep everything, and the user must move the sliders closer to the limit the resulting sequence to a specific region.

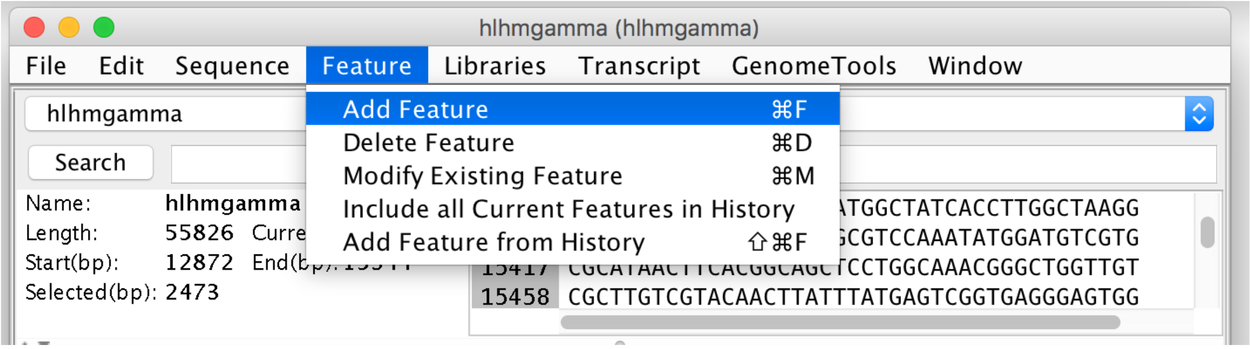
- *to Graphical selection* – If a box is selected within the graphical view of the window (see tutorial, chapter 1), then that region will be selected for masking.

Extract Coding Sequence
Much like the *Extract Non-Coding Sequence* submenu, this submenu allows users to mask non-coding sequence, while maintaining protein-coding sequence.

- *by Numbers* – Allows the user to enter a range of bases on the current sequence to extract non-coding sequence from. This option should not be confused with the “*by Numbers*” option in the Trim submenu.

- *by Gene Boundaries* – Allows the user to select a core set of genes that should be included in the masked sequence. Once those genes are selected, the user must use a slider to select the number of base pairs upstream and downstream of the selected genes. The default position on the slider is to keep everything, and the user must move the sliders closer to the limit the resulting sequence to a specific region.

- *to Graphical selection* – If a box is selected within the graphical view of the window (see tutorial, chapter 1), then that region will be selected for masking.



```
Name:          hlhmgamma
Length:        55826  Curre
Start(bp):     12872  End(b
Selected(bp):  2473

      ATGGCTATCACCTTGGCTAAGG
      GCGTCCAAATATGGATGTCGTG
15417 CGCATAACTTCACGGCAGCTCCTGGCAAACGGGCTGGTTGT
15458 CGCTTGTCGTACAACTTATTTATGAGTCGGTGAGGGAGTGG
```

The Feature Menu

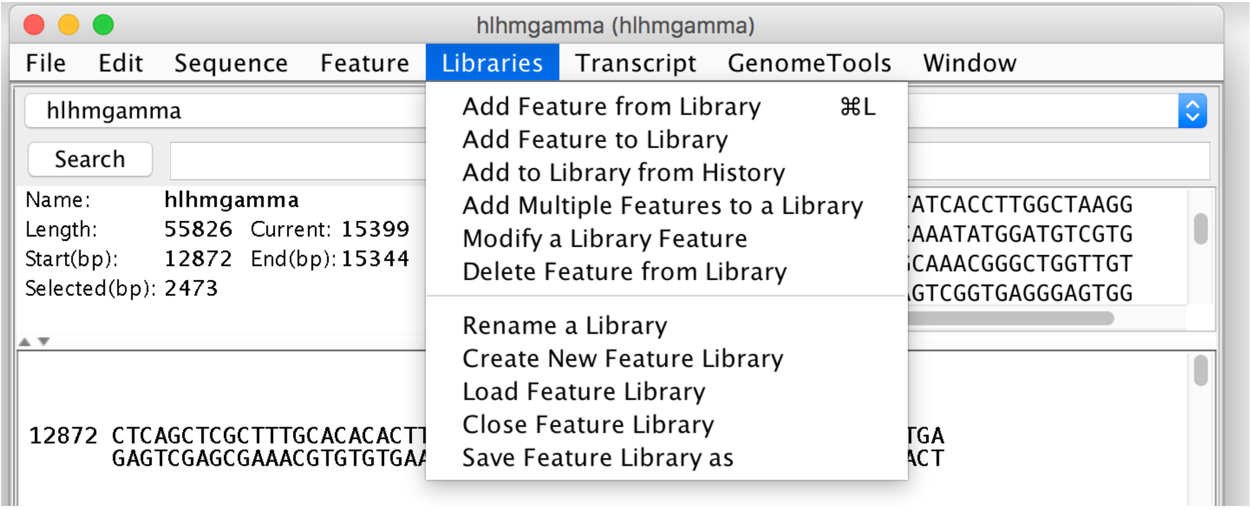
Add Feature
Add a new feature to the currently loaded sequence.

Delete Feature
Delete a feature from the currently loaded sequence.

Modify Existing Feature
Modify a feature that has been added to the sequence.

Include all Current Features in History
Takes all of the features that are loaded on the current sequence, and places them into the session history so that the feature can be added to other sequences, or to a Feature Library

Add Feature From History
Select a feature from the session history to add to the current sequence. The history contains features that have been added or modified in any sequence during the current GenePalette session. It also contains features that have been deleted from libraries during the current session.



Name: hlhmgamma
Length: 55826 Current: 15399
Start(bp): 12872 End(bp): 15344
Selected(bp): 2473

The Libraries Menu

Add Feature from Library
Allows user to add multiple features from any loaded library to the current sequence.

Add Feature to Library
Add a new feature to a loaded library

Add to Library from History
Add a feature to a loaded library from the session history. The history contains features that have been added or modified in any sequence during the current GenePalette session. It also contains features that have been deleted from libraries during the current session.

Modify a Library Feature
Use the feature-editor dialog to change a library feature.

Delete Feature from Library
Delete one or more features from a single library.

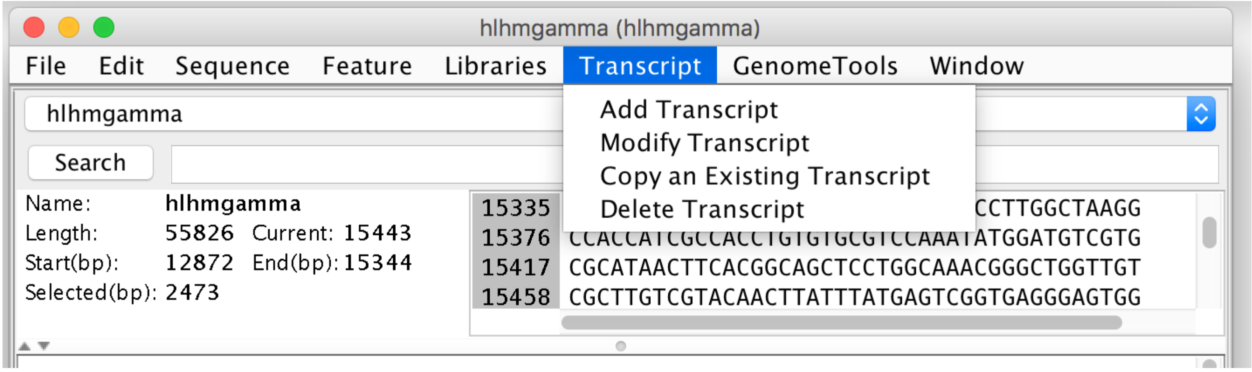
Rename a Library
Change the name of the library as it appears in the library selection tabs (for adding features from libraries) and in the library selection dialog.

Create New Feature Library
Makes a new feature library. The user must specify a name and file under which to save the new library.

Load Feature Library
Load a feature library that has been saved. If a library is in the **Libraries** directory under the main program directory, the library will be automatically loaded when the program starts up.

Close Feature Library
Closes a selected feature library. The library will not be available for modifying or using until it is opened again.

Save Feature Library As
All library changes are saved automatically to the original file that the library was opened with. However, this option lets you save the library as a new file in a new location.



```
Name:          hlhmgamma
Length:        55826   Current: 15443
Start(bp):     12872   End(bp): 15344
Selected(bp):  2473
```

```
15335                                         CCTTGGCTAAGG
15376  CCACCATCGCCACCTGTGTGCGTCCAAATATGGATGTCGTG
15417  CGCATAACTTCACGGCAGCTCCTGGCAAACGGGCTGGTTGT
15458  CGCTTGTCGTACAACTTATTTATGAGTCGGTGAGGGAGTGG
```

The Transcript Menu

Add Transcript
Adds a new transcript to the current sequence. Brings up a blank transcript-editor dialog.

Modify Transcript

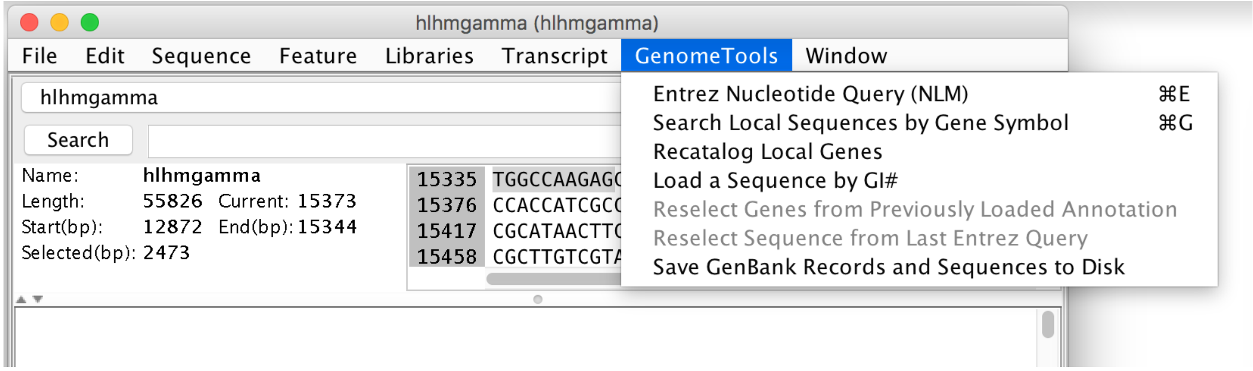
Make changes to a transcript that exists in the sequence. After the target transcript is selected, the data associated with the transcript is loaded into a transcript-editor dialog.

Copy an Existing Transcript

Makes a copy of transcript that is associated with the current sequence. This new transcript has the word "Copy" appended to the old transcript name.

Delete Transcript

Allows the user to select multiple transcripts to delete.


hlhmgamma (hlhmgamma)
File
Edit
Sequence
Feature
Libraries
Transcript
GenomeTools
Window
Search
Entrez Nucleotide Query (NLM) ⌘E
Search Local Sequences by Gene Symbol ⌘G
Recatalog Local Genes
Load a Sequence by GI#
Reselect Genes from Previously Loaded Annotation
Reselect Sequence from Last Entrez Query
Save GenBank Records and Sequences to Disk

hlhmgamma

Name: hlhmgamma
Length: 55826 Current: 15373
Start(bp): 12872 End(bp): 15344
Selected(bp): 2473

```
15335 TGGCCAAGAGC
15376 CCACCATCGCC
15417 CGCATAACTTC
15458 CGCTTGTCGTA
```

The GenomeTools Menu

Entrez Nucleotide Query (NLM)

Starts cascade of dialogs that allows the user to search GenBank via the National Library of Medicine's Entrez Query.

Search Local Genes by Gene Symbol

Allows user to search all of the sequences stored in the GenBank directory and immediate subdirectories therein for genes by gene symbol. If sequences have not been indexed, this item will take the user through the indexing process.

Recatalog Local Genes

Performs the indexing of local genes. This process does not remove previous indexes (You can catalog one genome at one point, and then add another genome at a later point without losing the first genome).

Load a Sequence by GI#

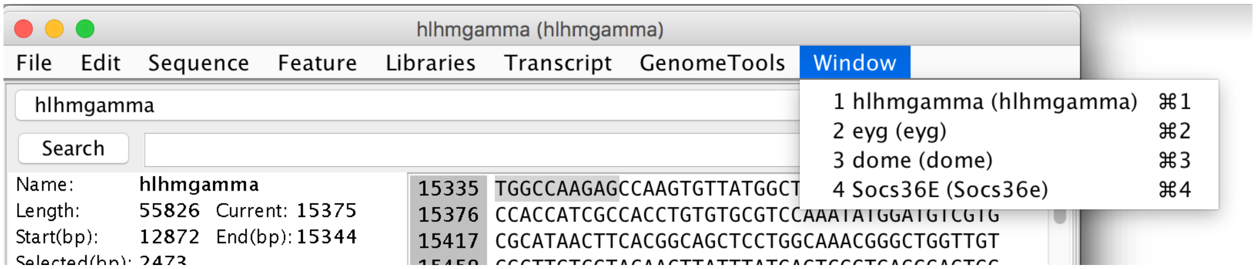
Prompts the user for a GI# (See chapter 2) so that genes associated with the sequence identified by this unique number can be selected for download into

GenePalette.

Reselect Genes from Previously Loaded Annotation
Allows the user to reselect genes from a parsed GenBank record that was loaded by either of the above two ways. A session history of parsed GenBank records is kept. Due to memory limitations, a parsed GenBank record is not added to this session history if the sequence was loaded from a local source.

Reselect Sequence from Last Entrez Query
This accesses the list of sequences from the last Entrez query so that you can go back and select a different sequence. This is especially useful if you want to download both the mouse and human version of a gene.

Save GenBank Records and Sequences to Disk
Search GenBank for records that you would like to download and save on disk so that future access to both the parsed annotation data and sequence is rapidly accelerated.



hlhmgamma

Name: **hlhmgamma**
Length: 55826 Current: 15375
Start(bp): 12872 End(bp): 15344
Selected(bp): 2473

```
15335 TGGCCAAGAGCCAAGTGTTATGGCT
15376 CCACCATCGCCACCTGTGTGCGTCCAAATATGGATGTCGTG
15417 CGCATAACTTCACGGCAGCTCCTGGCAAACGGGCTGGTTGT
```

The Window Menu

This menu contains an item for each window currently open in the GenePalette session. Just click on the line for the window that you want, and it will be brought to the front. Hotkeys for this menu item are CMD/CNTRL-#, for window numbers 1-9.

Filename:	GenePalette2.1.docx
Folder:	/Users/mrebeiz/Library/Containers/com.microsoft.Word/Data/Documents
Template:	/Users/mrebeiz/Library/Group Containers/UBF8T346G9.Office/User Content.localized/Templates.localized/Normal.dotm
Title:	GenePalette
Subject:	
Author:	Jim Posakony
Keywords:	
Comments:	
Creation Date:	3/28/09 8:13:00 PM
Change Number:	15
Last Saved On:	5/16/17 10:56:00 AM
Last Saved By:	Mark Rebeiz
Total Editing Time:	320 Minutes
Last Printed On:	5/16/17 10:56:00 AM
As of Last Complete Printing	
Number of Pages:	70
Number of Words:	17,063 (approx.)
Number of Characters:	84,805 (approx.)