

Lab 5.2:

Apollo: Gene Annotation Tool

Sanja Rogic

PhD student in Computer Science Department, UBC

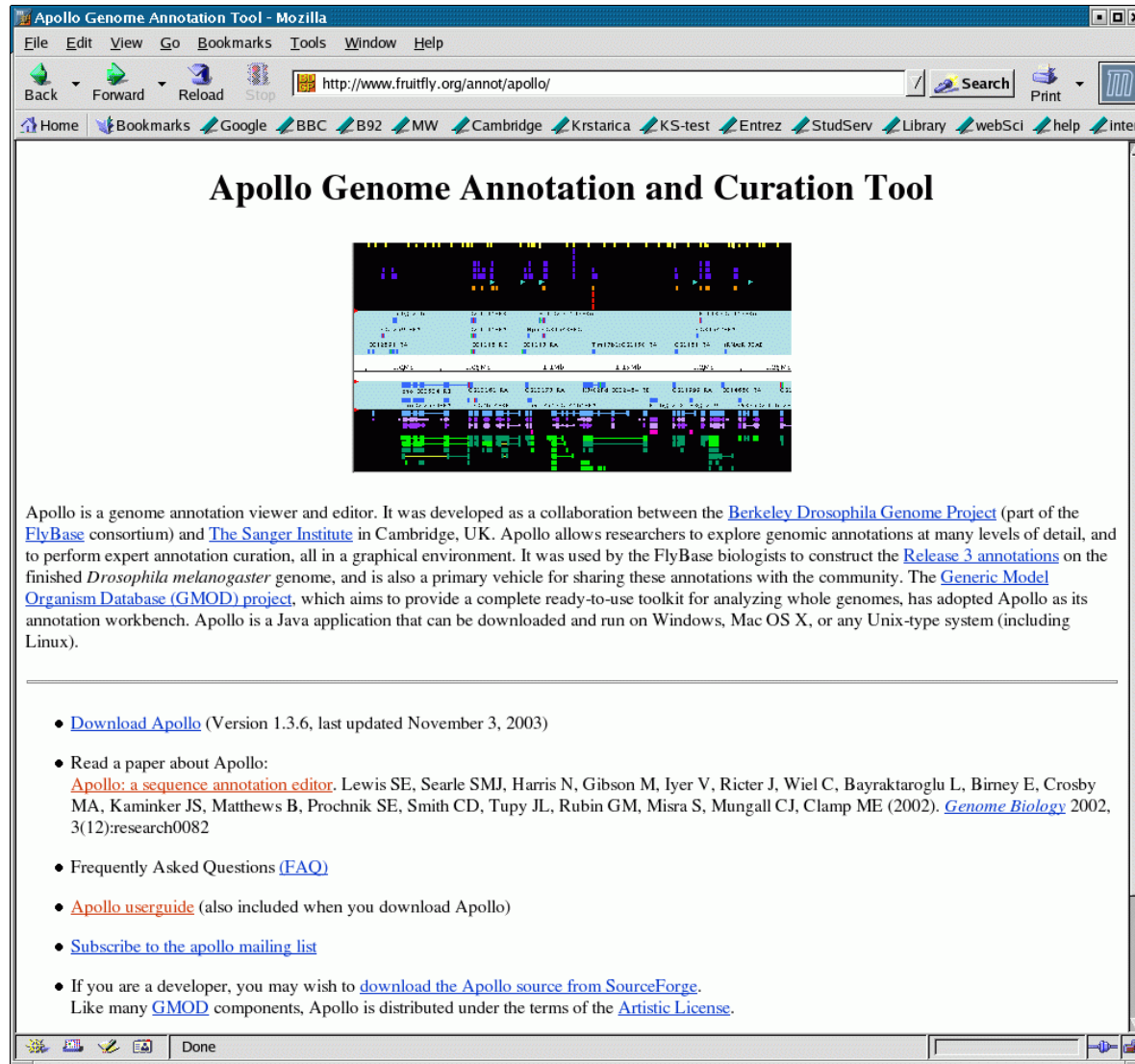
rogic@cs.ubc.ca

Outline

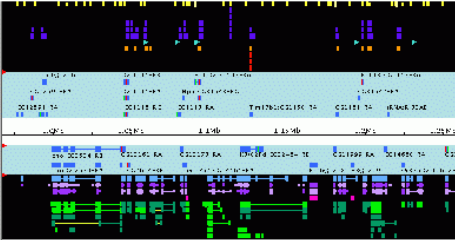
- Finding and installing Apollo
- Loading data
- Using Apollo as genome browser
- Using Apollo as annotation tool

Apollo - background

- developed as a collaboration between the Berkeley Drosophila Genome Project and The Sanger Institute in Cambridge, UK
- goal was to develop a tool to annotate the fly but which is also going to be capable to annotate and browse any larger eukaryotic genome



Apollo Genome Annotation and Curation Tool



Apollo is a genome annotation viewer and editor. It was developed as a collaboration between the [Berkeley Drosophila Genome Project](#) (part of the [FlyBase](#) consortium) and [The Sanger Institute](#) in Cambridge, UK. Apollo allows researchers to explore genomic annotations at many levels of detail, and to perform expert annotation curation, all in a graphical environment. It was used by the FlyBase biologists to construct the [Release 3 annotations](#) on the finished *Drosophila melanogaster* genome, and is also a primary vehicle for sharing these annotations with the community. The [Generic Model Organism Database \(GMOD\) project](#), which aims to provide a complete ready-to-use toolkit for analyzing whole genomes, has adopted Apollo as its annotation workbench. Apollo is a Java application that can be downloaded and run on Windows, Mac OS X, or any Unix-type system (including Linux).

- [Download Apollo](#) (Version 1.3.6, last updated November 3, 2003)
- Read a paper about Apollo:
[Apollo: a sequence annotation editor](#). Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, Kaminker JS, Matthews B, Prochnik SE, Smith CD, Tupy JL, Rubin GM, Misra S, Mungall CJ, Clamp ME (2002). *Genome Biology* 2002, 3(12):research0082
- Frequently Asked Questions ([FAQ](#))
- [Apollo userguide](#) (also included when you download Apollo)
- [Subscribe to the apollo mailing list](#)
- If you are a developer, you may wish to [download the Apollo source from SourceForge](#). Like many [GMOD](#) components, Apollo is distributed under the terms of the [Artistic License](#).

Installing Apollo

- Java application
- versions for:
 - Windows
 - Mac OS X
 - any Unix-type system
- code is open source and freely downloadable
- flexible and extendable
- still under development

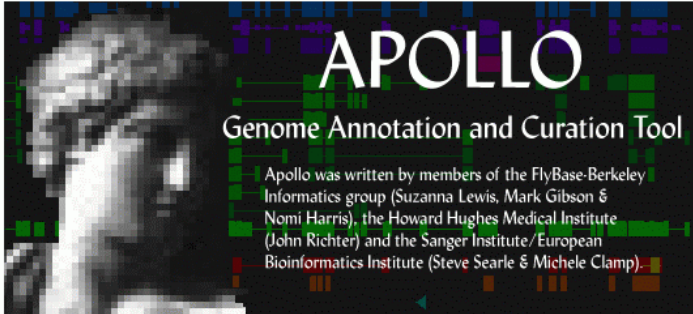
Install Apollo Genome Annotation Curation Tool - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop <http://www.fruitfly.org/annot/apollo/install.html> Search Print

Home Bookmarks Google BBC B92 MW Cambridge Krstarica KS-test Entrez StudServ Library webSci help intern

Download and Install Apollo Genome Annotation Tool



Apollo was written by members of the FlyBase-Berkeley Informatics group (Suzanna Lewis, Mark Gibson & Nomi Harris), the Howard Hughes Medical Institute (John Richter) and the Sanger Institute/European Bioinformatics Institute (Steve Searle & Michele Clamp).

Version 1.3.6 (November 3, 2003)

Click the "Start Installer" button below to automatically start the Apollo installation process. If the "Start Installer" button does not appear, or if it doesn't launch the installer, you can use the links below "Available Installers" to manually download the installer to your computer and then launch it.

Recommended Installation for Your Platform:

[Start Installer for Windows...](#)

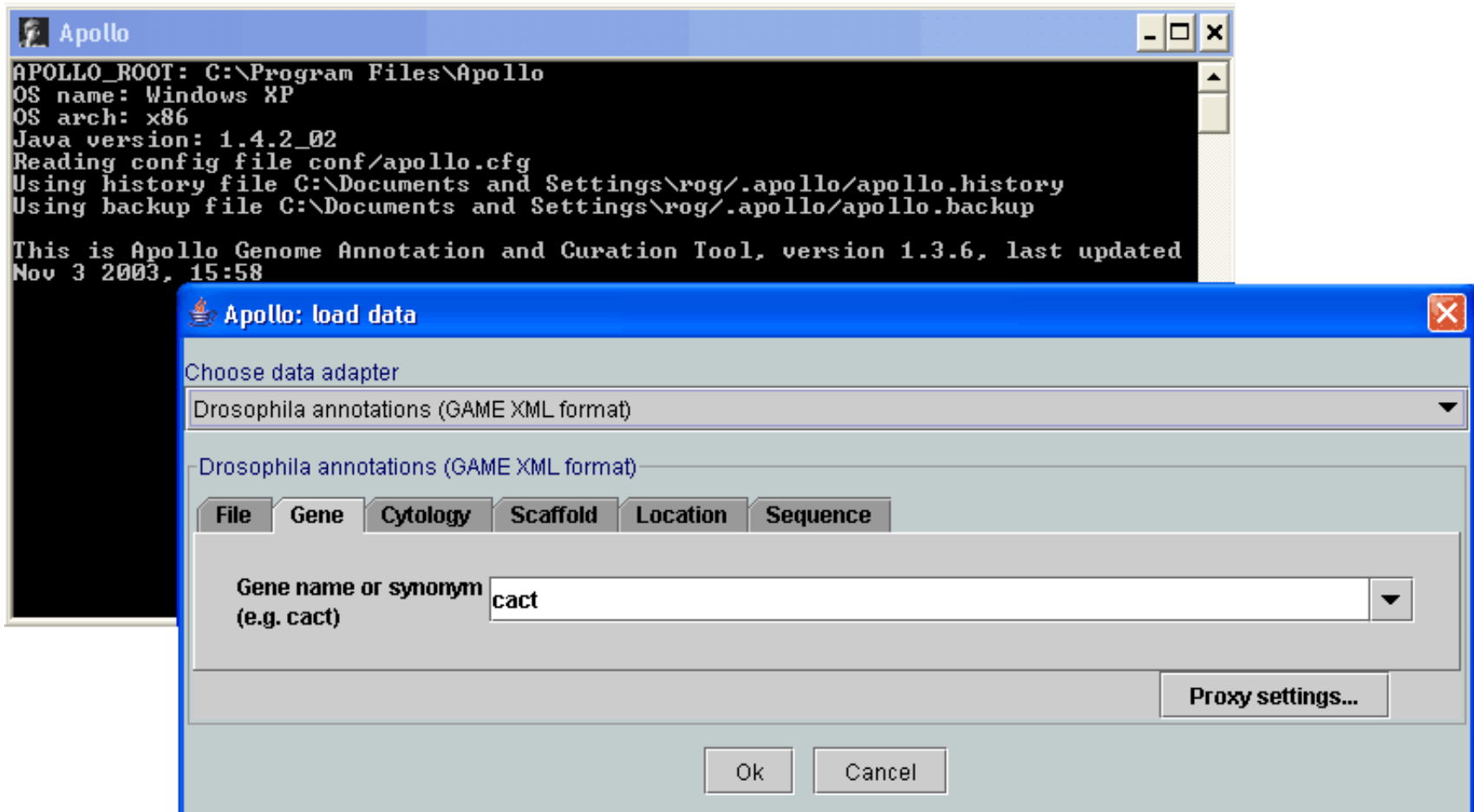
Installer created with [InstallAnywhere](#)® by Zero G Software, Inc. Copyright 2002. www.ZeroG.com

Available Installers

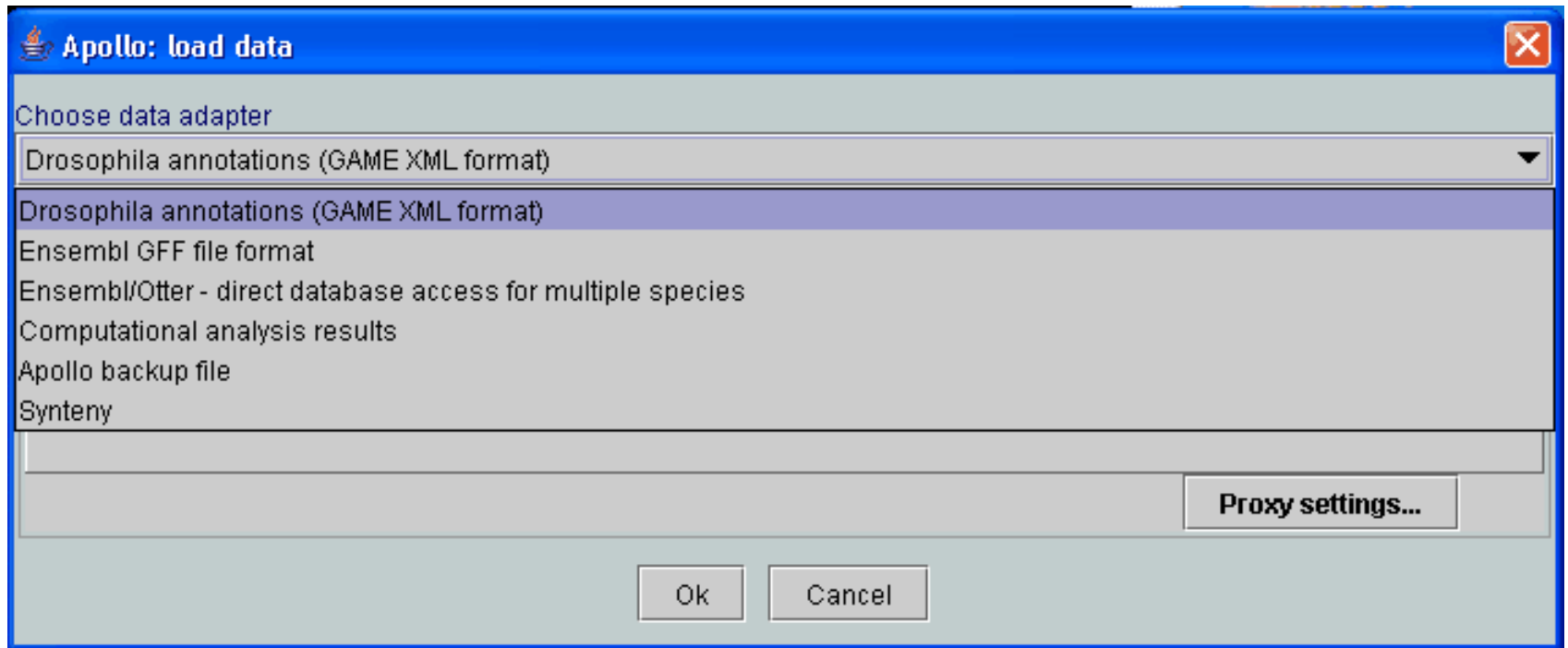
Platform	includes Java VM	without Java VM	Instructions
Windows	Download (22.6M)		View
Mac OS X		Download (9.1M)	View
Solaris	Download (40.7M)		View
Linux	Download (41.6M)		View
Any Unix Platform		Download (9.5M)	View

Done

Running Apollo

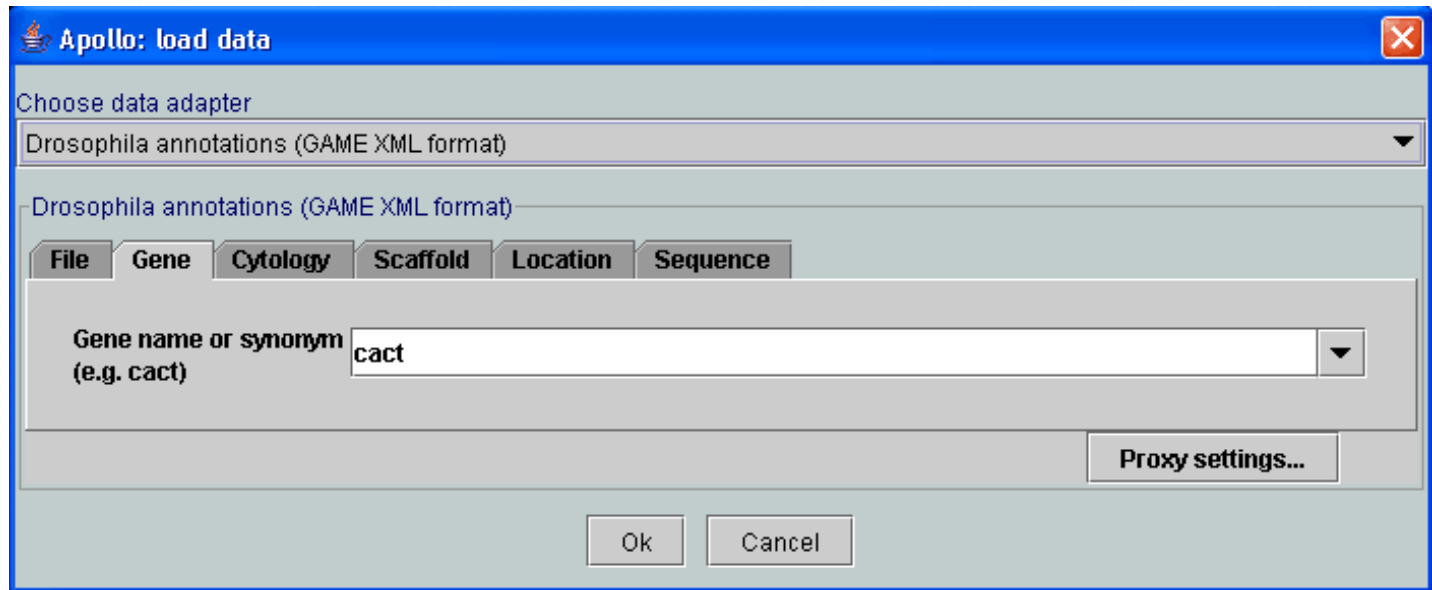


Loading data



Reading Drosophila annotations

- the annotations of the Drosophila genome are stored in a format called GAME XML
 - GAME (Genome Annotation Markup Elements) is a syntax for exchange of genomic information
 - XML - eXtensible Markup Language for interchange of structured data
- GAME XML can be read from a file or pulled across the network from the GadFly database

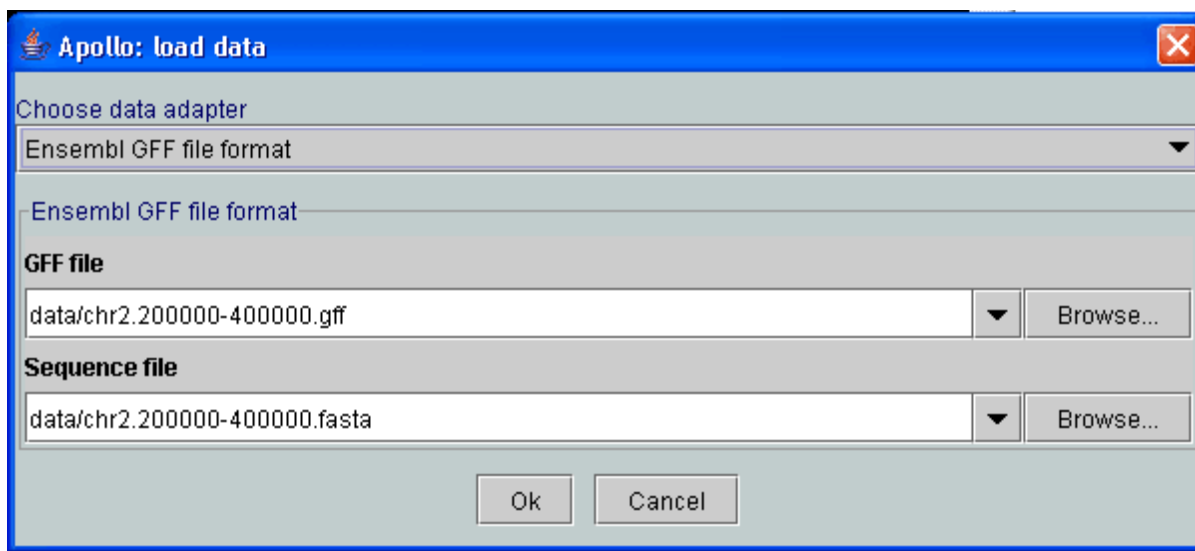


- local file
- gene name
- cytological band (e.g., 40B)
- scaffold accession – gets a ~350Kb chunk of genome
- chromosome arm, start and end position (e.g., 3R 10000 30000)
- sequence (using BLAST search) – finds the scaffold that includes the best match to the query sequence

} locally

} over the network

Reading Ensembl GFF files



- GFF – General Feature Format
- load local GFF file (Ensemble format)
- you can also read a FASTA-format sequence file to go with the GFF data
- Ensemble GFF not rich enough to support curated annotation

Reading Ensembl databases

The screenshot displays the 'Apollo: load data' dialog box. The 'Choose data adapter' dropdown is set to 'Ensembl/Otter - direct database access for multiple species'. The 'Region' section has 'Stable ID' selected with the value 'ENSG00000157399'. The 'Chromosome' section is set to 'Chr 1 1 50000'. The 'Accession' field contains 'AC003663.1.1.132070'. The 'Tracks' section is expanded, showing a list of tracks with checkboxes: Genes (checked), Dna Protein Alignments (checked), Dna Dna Alignments (checked), Features (simple) (checked), Simple Peptides (unchecked), Repeats (unchecked), Prediction Transcripts (checked), and Variations (e.g. SNPs) (unchecked). The 'Data Source' section is expanded, showing a 'Data Source' dropdown set to 'default'. The 'Configuration for: default data source' section contains fields for 'Host' (kaka.sanger.ac.uk), 'Port' (3306), 'User' (anonymous), and 'Password'. The 'Ensembl Database Name' dropdown is open, showing a list of database names: homo_sapiens_core_18_34, homo_sapiens_core_18_34, homo_sapiens_core_19_34a, homo_sapiens_core_19_34b, homo_sapiens_disease_18_34, homo_sapiens_disease_19_34a, homo_sapiens_disease_19_34b, homo_sapiens_est_18_34, and homo_sapiens_est_19_34a. A red arrow points from the 'Show Data Source Configuration' button to the 'Data Source' dialog box. Another red arrow points from the 'Tracks' section to the 'Tracks' list. The 'bioinformatics.ca' logo is visible at the bottom center.

Apollo: load data

Choose data adapter
Ensembl/Otter - direct database access for multiple species

Ensembl/Otter - direct database access for multiple species

Region

Stable ID: ENSG00000157399

Chromosome: Chr 1 1 50000

Clone Fragment: Accession: AC003663.1.1.132070

Tracks

Show Tracks

Data Source: Show Data Source Configuration

Ok Cancel

Data Source

Hide Data Source Configuration

Data Source: default

Configuration for: default data source

Host: kaka.sanger.ac.uk

Port: 3306

User: anonymous

Password:

Ensembl Database Name: homo_sapiens_core_18_34

Find...

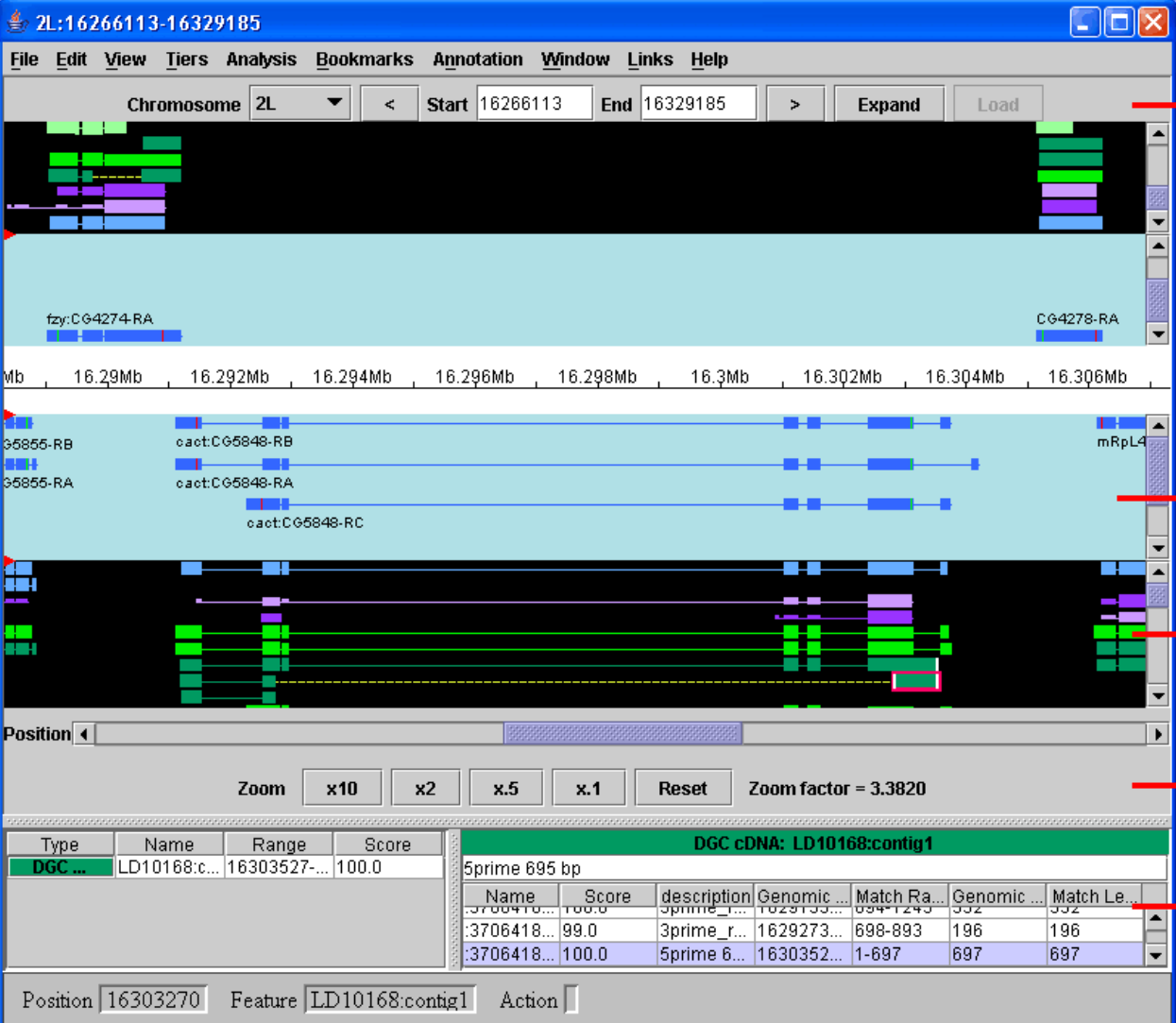
homo_sapiens_core_18_34
homo_sapiens_core_19_34a
homo_sapiens_core_19_34b
homo_sapiens_disease_18_34
homo_sapiens_disease_19_34a
homo_sapiens_disease_19_34b
homo_sapiens_est_18_34
homo_sapiens_est_19_34a

bioinformatics.ca

Try this - Start up Apollo

- start up the application
- choose GAME XML data adapter
- connect to GadFly Database:
 - choose **gene** option in the load panel
 - gene name: *cact*

Main window



navigation

coordinate line

annotation panel

result panel

zoom in/out

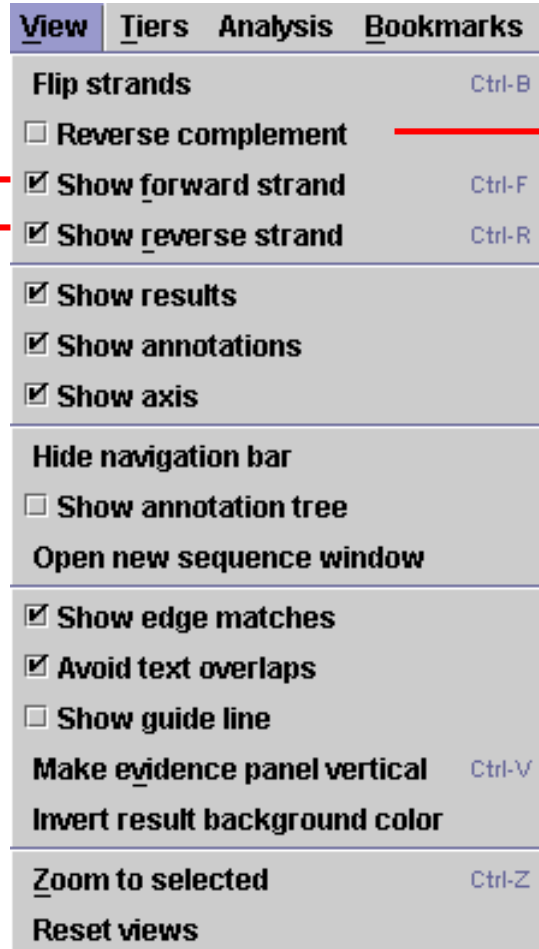
feature detail panel

Navigation

- select chromosome arm, start and end position
- request the region immediately upstream or downstream
- **expand** button extend the current region by 50%
- hit **load** button every time
- data fetched over the internet
- zooming and horizontal scrolling
- centering the display – middle mouse click or dual mouse click

Forward and reverse strands

displaying
forward/reverse
strand

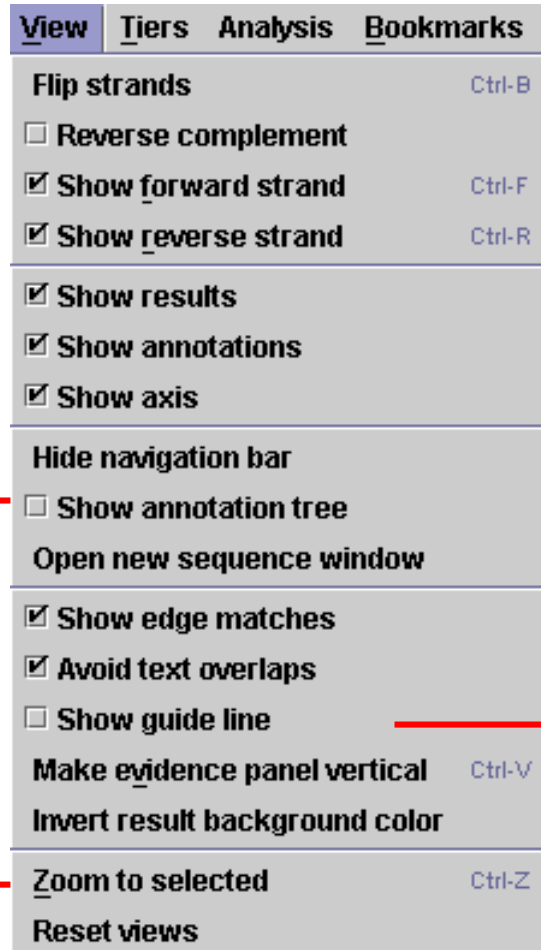


reverse complement
the features and
sequence

Try this – reverse complement and navigation

- turn off the forward strand
- select reverse complement
- center display to 2nd exon of transcript *cact:CG5848-RC*
- zoom in until sequence shows up
- read the first five nucleotides at the 5' end of the exon
- return to the original view (use reset)

Other options in View menu



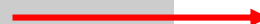
list of annotated genes



zooms and centers selected feature



mark current location



Feature detail panel

Type	Name	Range	Score
Com...	L04964-AE...	16303708-...	100.0

Community GB: L04964 (length=2011)				
Drosophila melanogaster cactus mRNA exons 1-7, complete cds.				
Score	Genomic Range	Match Range	Genomic Length	Match Length
100.0	16291534-16...	1590-2011	422	422
100.0	16292792-16...	1332-1589	258	258
100.0	16292942-16...	1243-1331	89	89
100.0	16301254-16...	1032-1242	211	211
100.0	16301601-16...	854-1031	178	178
99.0	16303138-16...	132-853	722	722

- detailed info for selected feature(s)
- left panel: type, name, range, score
- right panel: coordinates and other info
- each feature set displayed only once in left panel

Types panel

Tiers Analysis Bookmarks A

- Show types panel
- Collapse all tiers Ctrl-C
- Expand all tiers Ctrl-E
- Show all tiers
- Hide all tiers
- Invert tier order
- Decrease tier height (-)
- Increase tier height (+)

Types

<input type="checkbox"/> Sort	Gene Prediction	Show <input checked="" type="checkbox"/>
<input type="checkbox"/> Label		Expand <input checked="" type="checkbox"/>
<input type="checkbox"/> Sort	Transposon	Show <input checked="" type="checkbox"/>
<input type="checkbox"/> Label		Expand <input checked="" type="checkbox"/>
<input type="checkbox"/> Sort	Fly Sequence	Show <input checked="" type="checkbox"/>
<input type="checkbox"/> Label		Expand <input checked="" type="checkbox"/>
<input type="checkbox"/> Sort	Fly EST	Show <input checked="" type="checkbox"/>
<input type="checkbox"/> Label		Expand <input checked="" type="checkbox"/>
<input type="checkbox"/> Sort	Non-coding RNA	Show <input checked="" type="checkbox"/>
<input type="checkbox"/> Label		Expand <input checked="" type="checkbox"/>
<input type="checkbox"/> Sort	BLASTX Similarity to Fly	Show <input checked="" type="checkbox"/>
<input type="checkbox"/> Label		Expand <input checked="" type="checkbox"/>
<input type="checkbox"/> Sort	BLASTX Similarity to Other ...	Show <input checked="" type="checkbox"/>
<input type="checkbox"/> Label		Expand <input checked="" type="checkbox"/>
<input type="checkbox"/> Sort	Insertion Site	Show <input checked="" type="checkbox"/>
<input type="checkbox"/> Label		Expand <input checked="" type="checkbox"/>

right
click

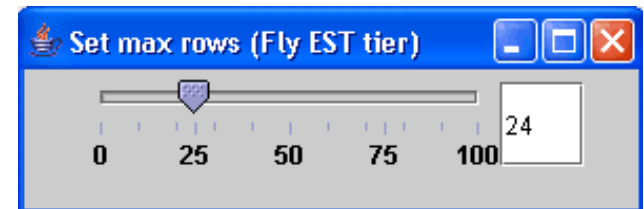
Genie
Genscan

right
click

tRNA-result
Non-coding RNA

Types panel

- **show** – switch tier on/off
- **expand** – features shown in different rows
- **sort** – by score
- **label** – show label
- limit number of rows in a tier:
 - left-click when sort on (number of row)
 - middle-click (threshold)
- change order of tiers in result panel: select a feature + shift + right-click + drag
- change the colour of feature type: right-click on a tier + select a feature + choose a colour from a colour box



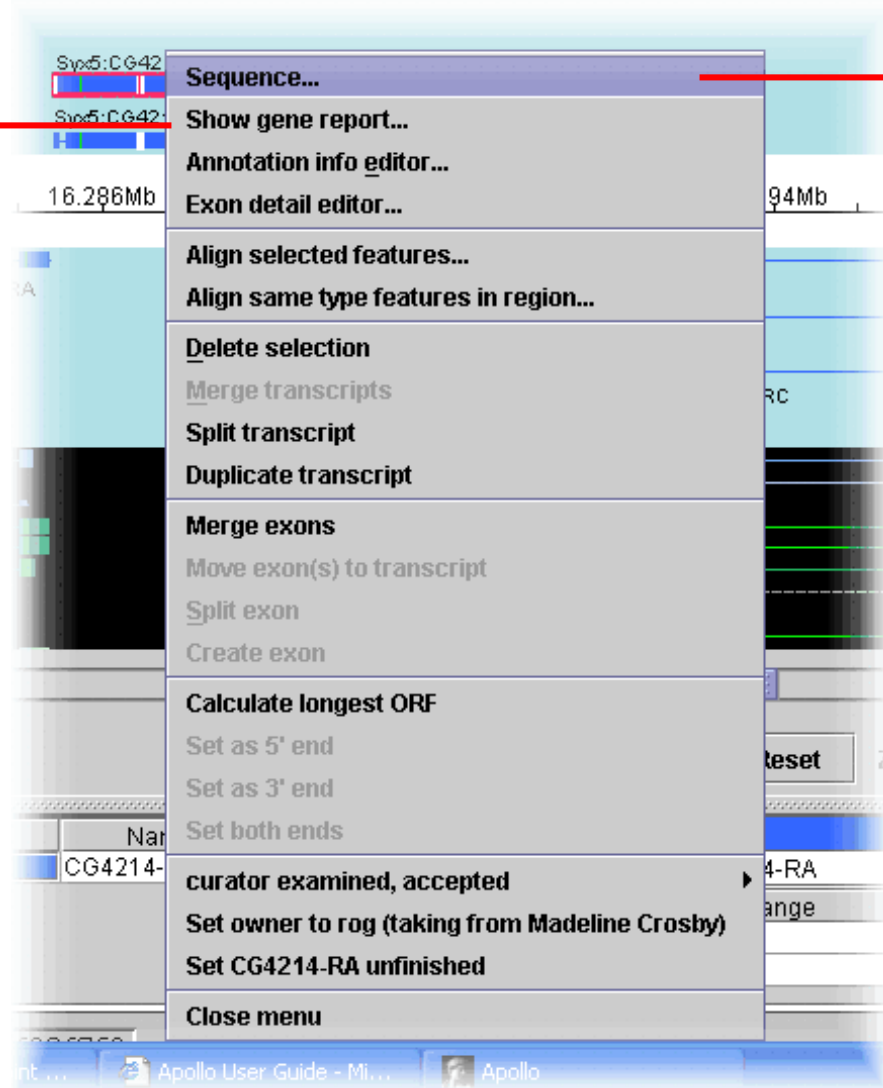
Try this – playing with tiers

- collapse **Blastix similarity to fly** tier
- set max number of rows to 2 for **Fly sequence** tier
- change order of tiers – move **Fly EST** closest to the annotation panel
- change colour of **Fly EST** tier

Selecting multiple features

- selecting a transcript:
 - click on an intron
 - double-click on an exon
- multiple features:
 - add to a selection by shift-clicking with left mouse button
 - rubber-banding – press and drag middle button around the features

Right-clicking on a transcript



open Sequence window

operations on the transcript

operations on the exons

get report from the GadFly database

Annotation info

Syx5 Annotation Information

Annotation

BG:DS02740.18
BG:DS02740.5
BG:DS02740.8
BG:DS02740.9
c(2)M
cact
CG13258
CG31737
CG31818
CG4278
CG4440
CG5861
chif
cni
Cyp303a1
fzy
heix
mRpL4
pkAAP
Syx5

Annotation name: Syx5
Annotation synonyms:
Is problematic?
Type: gene
Is dicistronic?
Evaluation of peptide: curator examined, accepted

Type	ID Value	DB Name
id	GO:0005486	GO
id	GO:0016083	GO

Gene Transcript: CG4214-RB
SwissProt comment: Imperfect match to REAL SP with corresponding FBgn
Gene Transcript: CG4214-RA
SwissProt comment: Imperfect match to REAL SP with corresponding FBgn

Annotation comments

Select an author/date pair to edit an existing comment, or click 'Add' to add a new comment.

Author and date:

Gene Transcript

CG4214-RB
CG4214-RA

Gene Transcript name: CG4214-RB
Gene Transcript synonyms:
Is problematic?
Finished?
Owned by: Madeline Crosby

Gene Transcript comments

Select an author/date pair to edit an existing comment, or click 'Add' to add a new comment.

Author and date:

Follow external selection

Search functions – from edit menu

Find [Close]

Position: 16266399 [Goto]

Name: [Find]

Sequence: [Find]

Search reverse strand?

Use Regular Expressions?

Results: Centering on base 16266399

Find [Close]

Position: [Goto]

Name: [Find]

Sequence: GATTACA [Find]

Search reverse strand?

Use Regular Expressions?

Position	Sequence
16285783-16285789	GATTACA
16299281-16299287	GATTACA
16315057-16315063	GATTACA
16320463-16320469	GATTACA

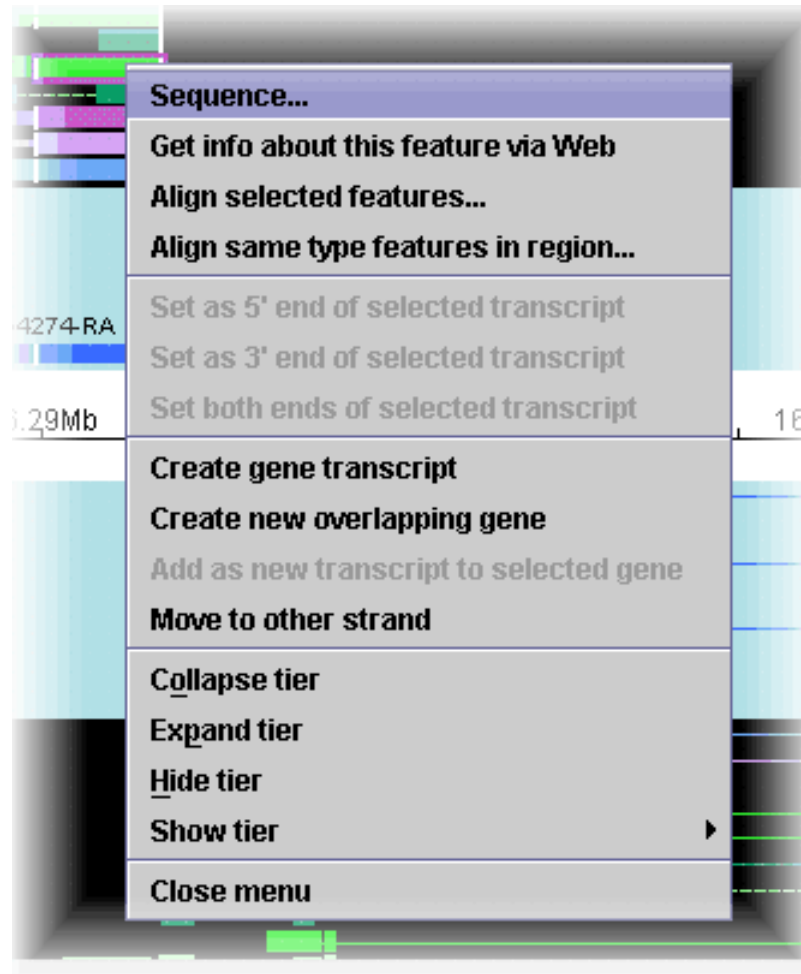
Results: Found sequence 4 times

Try this – finding a sequence

- find a sequence: ACATTAG
- which of the found matches is located in an annotated gene
- what is the name of that transcript?

Right-clicking on a feature

multiple sequence alignment



operations on the tiers

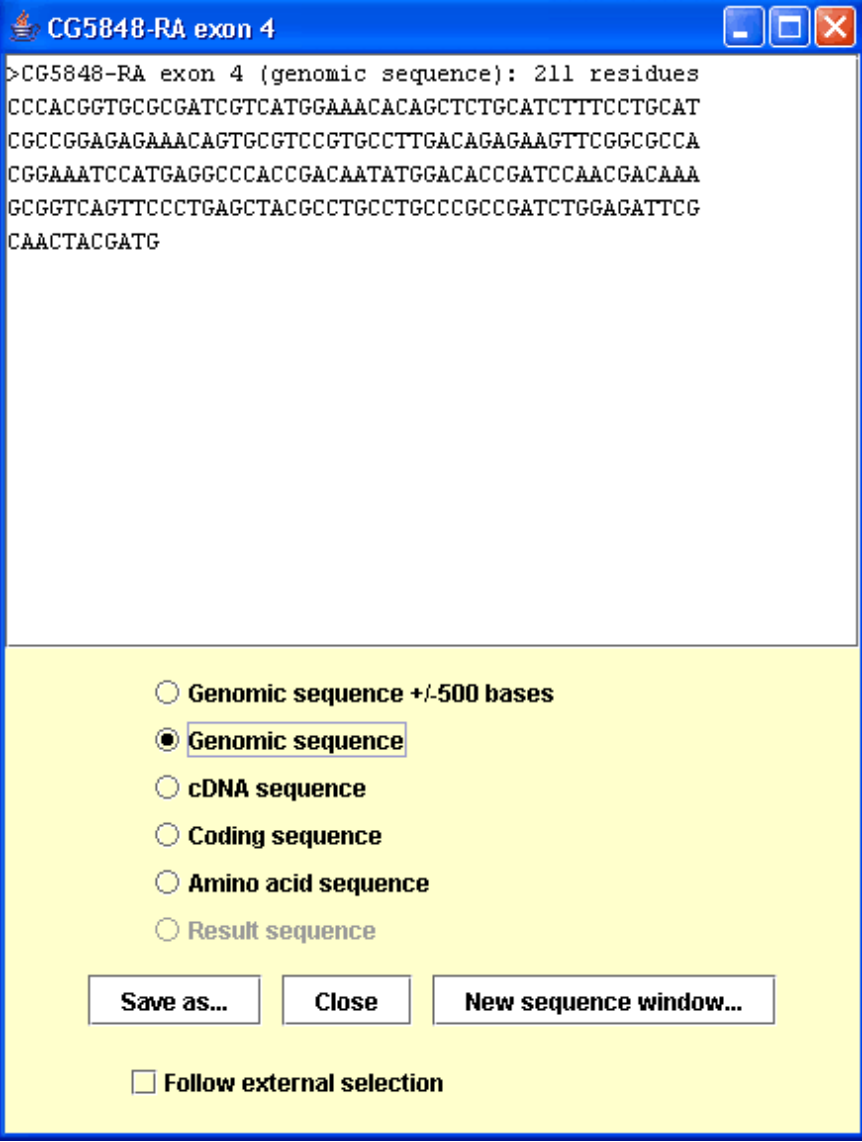


(also in Tiers menu)

Sequence level features

- zoom in to see the sequence
- start and stop codons displayed for all six frames
- right-click on a feature and choose “Sequence” to open a Sequence windows

Sequence window



CG5848-RA exon 4 (genomic sequence): 211 residues

```
CCCACGGTGCGCGATCGTCATGGAAACACAGCTCTGCATCTTTCTGCAT  
CGCCGGAGAGAAACAGTGCCTCCGTGCCTTGACAGAGAAGTTCCGGCGCCA  
CGGAAATCCATGAGGCCACCGACAATATGGACACCGATCCAACGACAAA  
GCGGTCAGTTCCTGAGCTACGCCTGCCTGCCCGCCGATCTGGAGATTCC  
CAACTACGATG
```

Genomic sequence +/-500 bases

Genomic sequence

cDNA sequence

Coding sequence

Amino acid sequence

Result sequence

Follow external selection

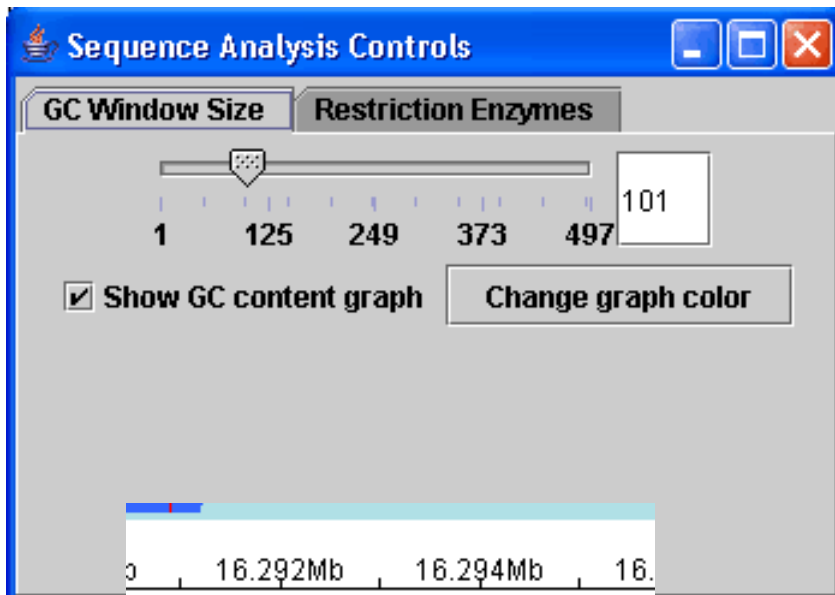
Try this – using Sequence window

- use sequence window to determine:
 - what is the genomic sequence length of the transcript *cact:CG5848-RB*?
 - how many amino acid residues are there in the same transcript?

Analysis menu

GC content

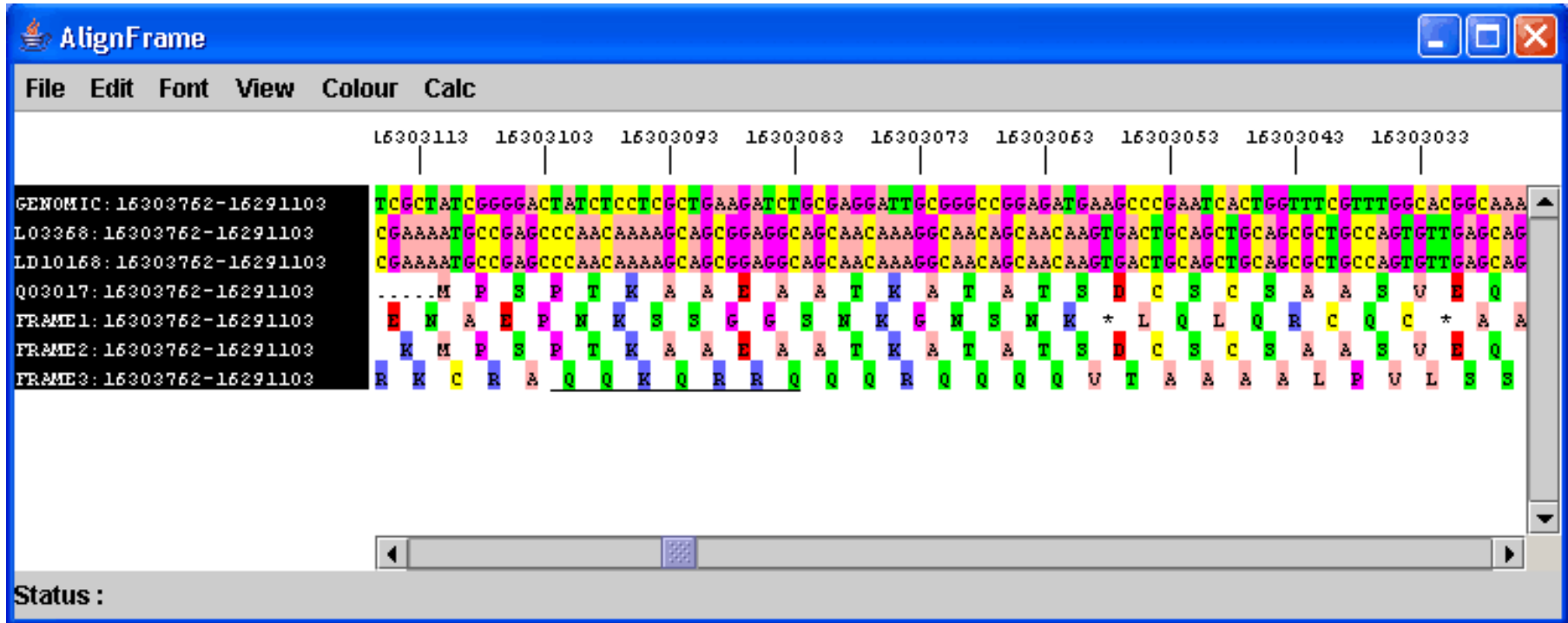
restriction sites



The screenshot shows the 'Sequence Analysis Controls' window with the 'Restriction Enzymes' tab selected. A list of enzymes is shown on the left, and a table of restriction sites is displayed on the right.

	Position	Sequence
AatII	16266687-16266692	GACGTC
AcellI	16273765-16273770	GACGTC
AcilI	16298644-16298649	GACGTC
Acil	16299267-16299272	GACGTC
AclI	16300466-16300471	GACGTC
AflII	16305230-16305235	GACGTC
AgeI	16306093-16306098	GACGTC
AhaII		
AluI		

Viewing alignments - Jalview



- select features to align + right-click + select “Align selected features”
- sequences will show up in Jalview along with genomic sequence and three-frame translation

Try this – viewing alignments

- look at the alignment of the supporting evidence for *cact* gene
 - “rubberband” features you want to align
 - right-click
 - select “Align selected features” from drop-down menu

What is annotation?

- biological interpretation of a specific region on a nucleic acids sequence
- any feature that can be anchored to a sequence is an annotation (exon, promoter, transposable element, regulatory region, CpG island)

Apollo as annotation tool

- only GAME XML data can be edited
- GFF format not rich enough to support annotation
- annotation can be saved (“Save as...”) in GAME XML or GFF format
- Apollo does not compute any of the features
 - all computational evidence needs to be pre-computed and imported in Apollo

Creating a gene model

- select results on which to base the gene annotation and drag them into annotation panel
- if there is an overlap with existing gene a new transcript will be created
- “Create gene transcript” another way to do it
- “Create new overlapping gene” will create a new gene
- exons can be added to an existing transcript by shift + left-click + drag exon to the transcript
- to create an exon without support right-click and select “Create exon”
- to delete exon/transcript choose “Delete feature”

Try this – creating a gene model

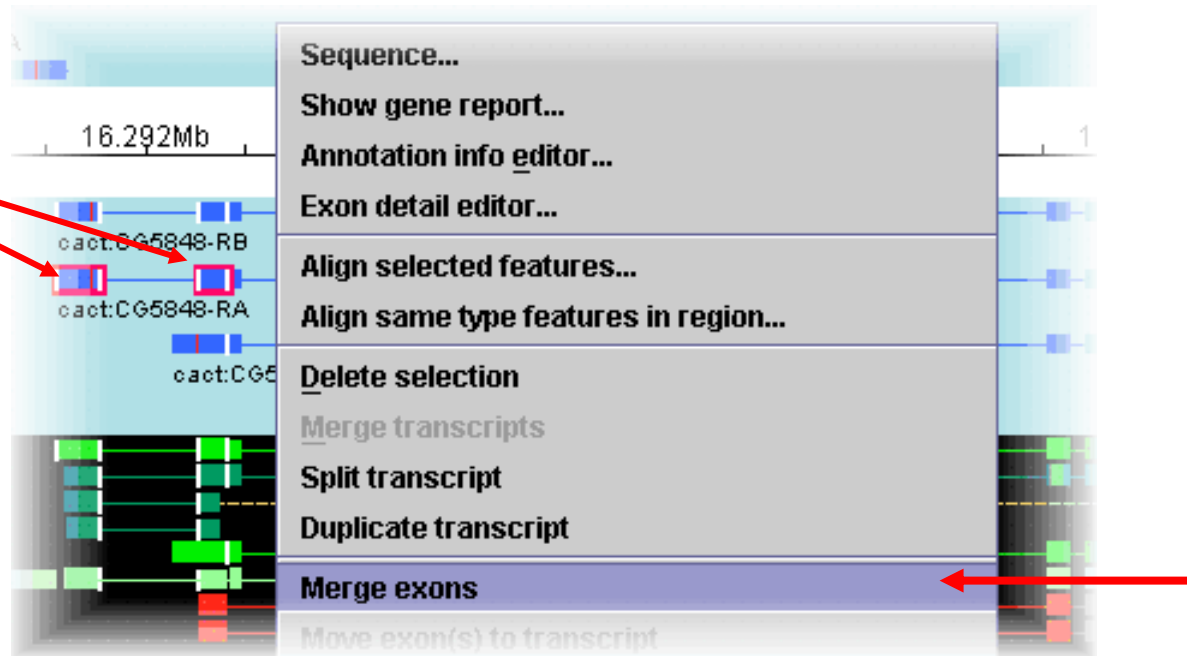
- lets suppose that we believe Genie's prediction for *cact* gene - create a new transcript based on this computational evidence
- next: delete one of the exons from any of the *cact* transcript and create it again by dragging it from result panel

Merge exons

1. select exon
2. shift-click another exon in the same transcript
3. right-click to bring up popup menu and select “Merge exons”

 all introns between two exons will disappear

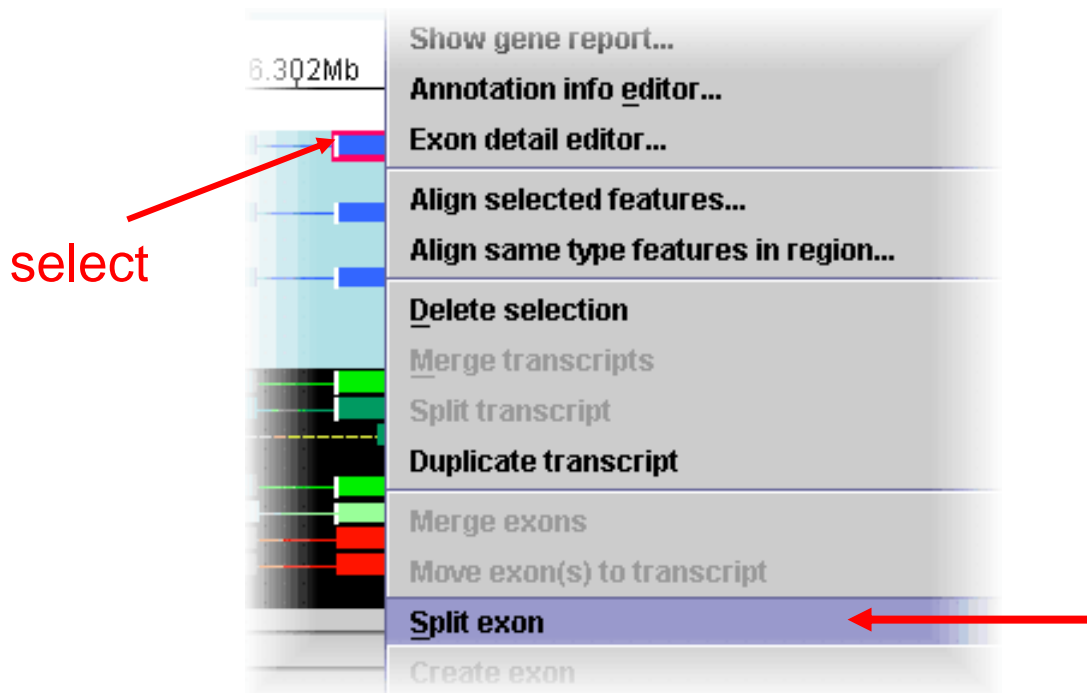
select



Split exon

1. select exon
2. put cursor on exon where you want the intron to be and right-click
3. select “Split exon”

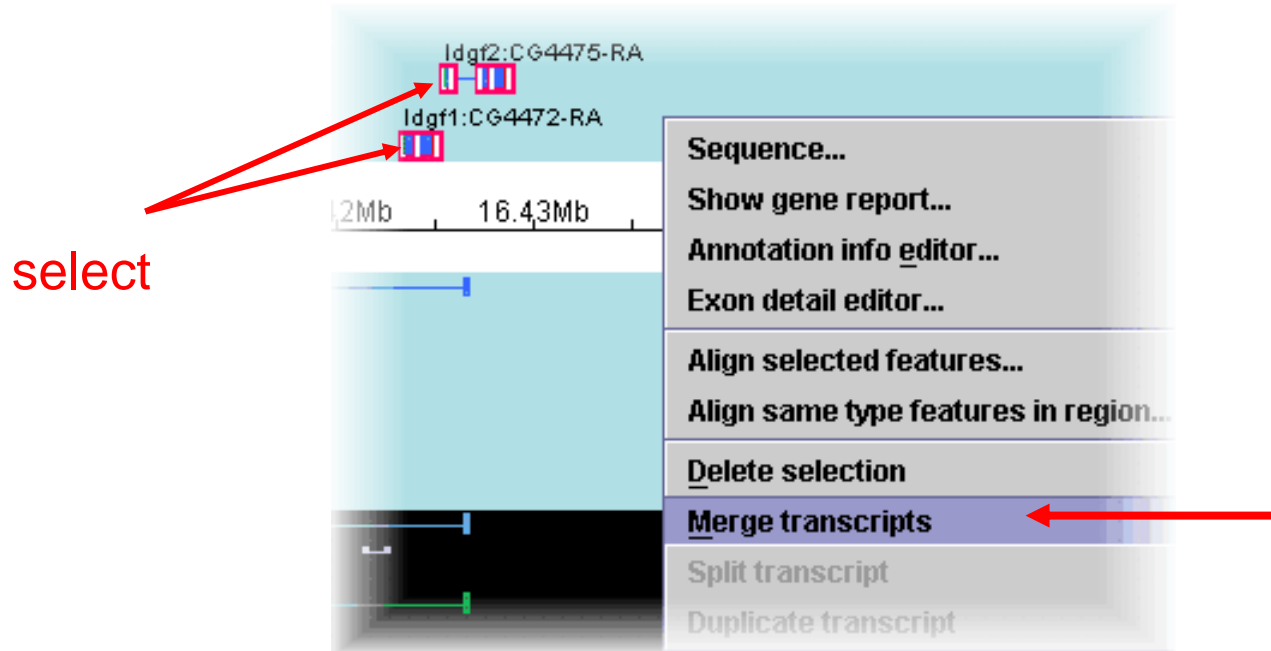
 one-base intron will be created



Merge transcripts

1. select exon or all of transcript A
2. shift-click select an exon or all of transcript B
3. right-click to bring up popup menu and select “Merge transcripts”

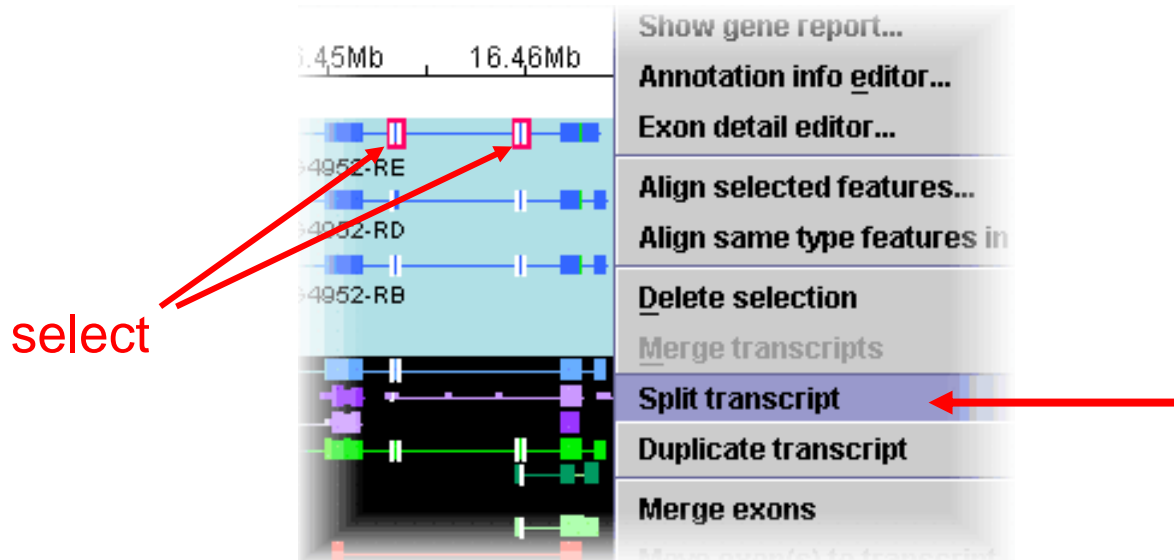
 two transcripts will be merged into one



Split transcript

1. select exon at one end of split location
2. shift-click on the exon on the other side of the intron
3. right-click and select “Split transcript”

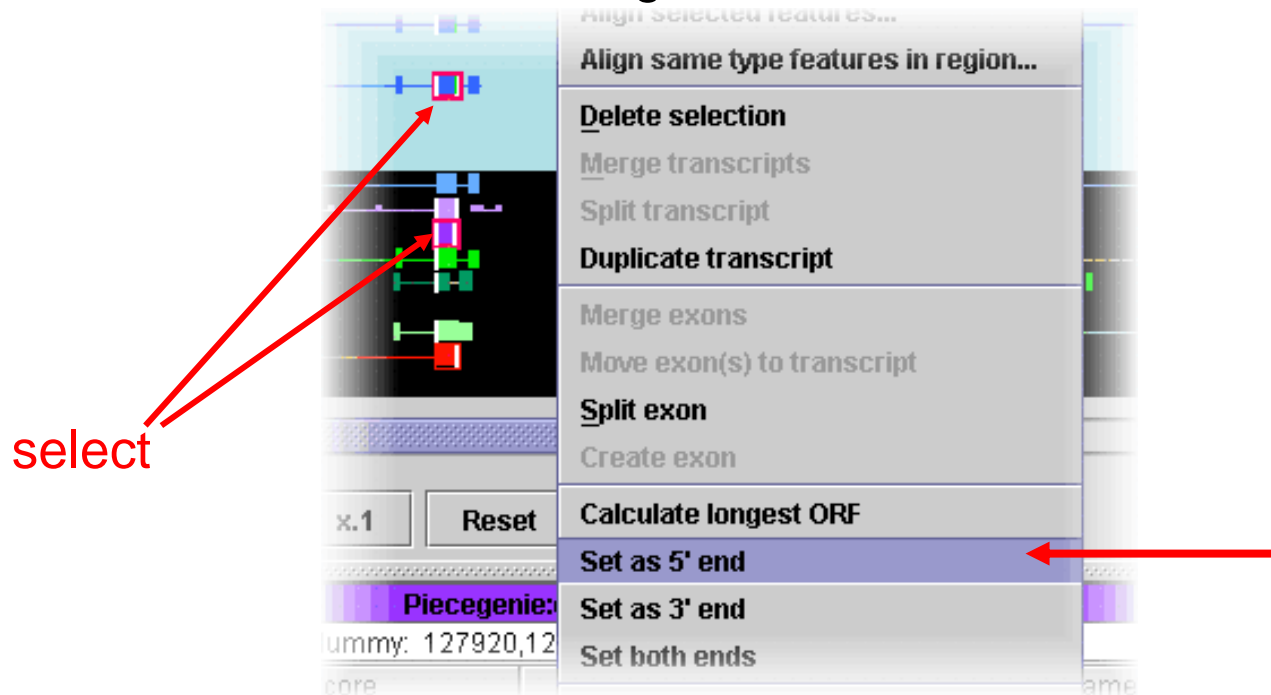
 will remove the intron connecting two exons and create two genes



Set as 5' (3') end

1. select exon you wish to modify
2. shift-click on an exon in result panel whose boundaries you wish to adopt
3. right-click and select "Set as 5' end"

 exon boundaries will change



Exon detail editor

- right-click on an exon and select “Exon detail editor...”

Forward Strand Exon Editor

16289951
L I S S R F K S Y G L E C R Q Y S G C V G
D S L F P Q D I L N I L W D I W G S V P D T N I I V W A L W
TTGATCTCTTCCAGATTTAAATCTTATGGATTGGAGTGCCGACAATATAGTGGCTGTG
TTGATCTCTTCCAGATTTAAATCTTATGGATTGGAGTGCCGACAATATAGTGGCTGTG

16290009
L G Q L R L F V E R T D R K Y R A A Y
A L G S L C V S Y L W N T A H Q T P G N I E S L T
GCCTTGGGCAGTTGCGTCTATTTGTGGAACGCACAGACCGGAAATATCGAGCAGCTTA
GCCTTGGGCAGTTGCGTCTATTTGTGGAACGCACAGACCGGAAATATCGAGCAGCTTA

16290067
G V F G G R L R R L A I V D P G G A D
R E S L E R E R G A D Y T A G S R L S W I S Q E R G Q R I Y
CGGAGTTTGAGGAGGGCGACTACGCAGGCTCGCTATCGTGGATCCAGGAGGGGCAGAT
CGGAGTTTGAGGAGGGCGACTACGCAGGCTCGCTATCGTGGATCCAGGAGGGGCAGAT

16290125
T C H R Q Q S H R C R G A V E A L W G D L L Q S E A
L A P S G A N S A T P V P W E S C G T C A S P K V K S
ACTTGCCATCGGCAACAGCAACGGTGCCGTGGAGCTGTGGGACTGCTCCAAGTGAAAG
ACTTGCCATCGGCAACAGCAACGGTGCCGTGGAGCTGTGGGACTGCTCCAAGTGAAAG

3 3 2

Transcript: CG4274-RA Translation length: 526

Go to next 5' gene (CG4214-RB)

Go to next 3' gene (CG4278-RA)

Find sequence... Clear search hits

Show introns in translation viewer Follow external selection

Exon detail editor

- separate line of sequence for each transcript
- three-frame translation also shown
- exons denoted in blue with successive exons shown in alternating light and dark blue shades
- gene structure shown on the bottom of the window

Exon detail editor

- green and red line represent translation start and stop
- numbers on the exons indicate the translation reading frame (1-top, 2-middle, 3-bottom)
- yellow box indicates the region of sequence that is currently visible
- clicking on the graphic of the transcript will center the sequence around that region

Operations in exon detail editor

makes one-base
intron

Location = 16289869
Base 280 of exon 2 mRNA base 764

- Split
- Make intron
- Create exon
- Delete exon
- Merge with 5' exon
- Merge with 3' exon
- Set as 5' end
- Set as 3' end
- Set start of translation
- Find sequence...
- Sequence...

creates one-base exon

- right-click on a nucleotide to get a popup menu
- merge - deletes intron and merges with adjacent exon
- set – right-click on a base in an exon to set it as 5' (3') end

Try this – changing gene structure in EDE

- open EDE for any of the exons of *syx5:CG4214-RB* transcript
- split exon 2 and make an intron of arbitrary length by dragging the exon boundaries
- split new exon 3 and make an intron using “Set as 5’(3’) end”
- merge exons 4 and 5

Configuring Apollo

- `apollo.cfg` - main configuration file in `apollo/conf`
- `.style` - file for each data source (`game.style`, `ensembl.style`)
- `.tier` - (`game.tiers`, `ensembl.tiers`)
- personal `.cfg`, `.style`, and `.tiers` files should be in `.apollo` directory

Summary

- Apollo as genome browser
 - GadFly database
 - Ensemble database
- Apollo as annotation tool
 - sequence analysis has to be done independently of Apollo and then imported using appropriate format
 - annotation can be saved in GAME or GFF format