ANNOTATING A GENOME SEQUENCE: GENE MODELS

Alexie Papanicolaou & Monica Munoz-Torres

NEXT-GEN Sequencing | NESCent Academy

Durham, NC. August 15-26, 2011

Generation of gene models.

by Alexie Papanicolaou.

MAKER

Manual curation of automated gene models.

by Monica Munoz-Torres

APOLLO

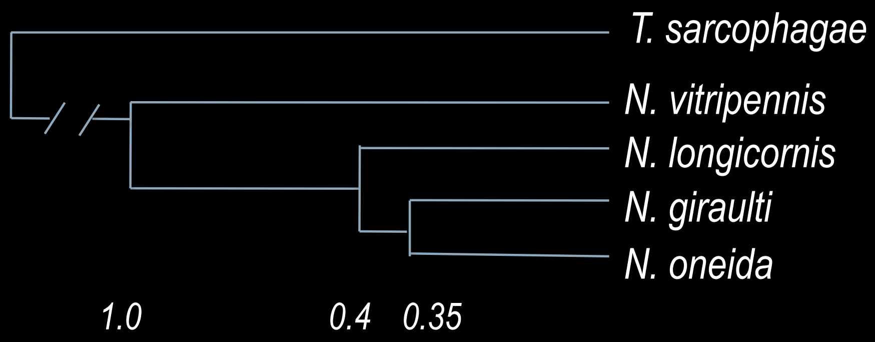
INTRODUCTION

- When annotating a genome, you are looking at a ‘frozen photograph’ of the assembly.
- At HGD a gene model in the *Nasonia vitripennis* genome may be supported by any combination of models from NCBI RefSeq, NCBI *ab* initio and Fgenesh, Fgenesh++, Augustus and GLEAN. Additional evidence alignments with EST and/or gene models from other species of *Nasonia* and other model and non-model insects; these alignments were calculated using algorithms from either Exonerate or tBLASTx.

INTRODUCTION

- RefSeq entries constitute a high-quality gene prediction set; because of this, in many cases a RefSeq gene model can be the starting point for your manual annotation efforts. When editing an existing gene model, please be sure to identify the corresponding RefSeq model (RefSeq ID: XP, NP for protein sequences; XM, NM for mRNAs) that you are planning to replace with your manual annotation.
- There may be more than one transcript per RefSeq gene, so please check them all. You may also wish to annotate additional transcripts rather than replacing a RefSeq transcript. In some cases, annotation may involve adding UTRs without modifying the CDS.

NASONIA
THE JEWEL WASP



… IN THE LAB

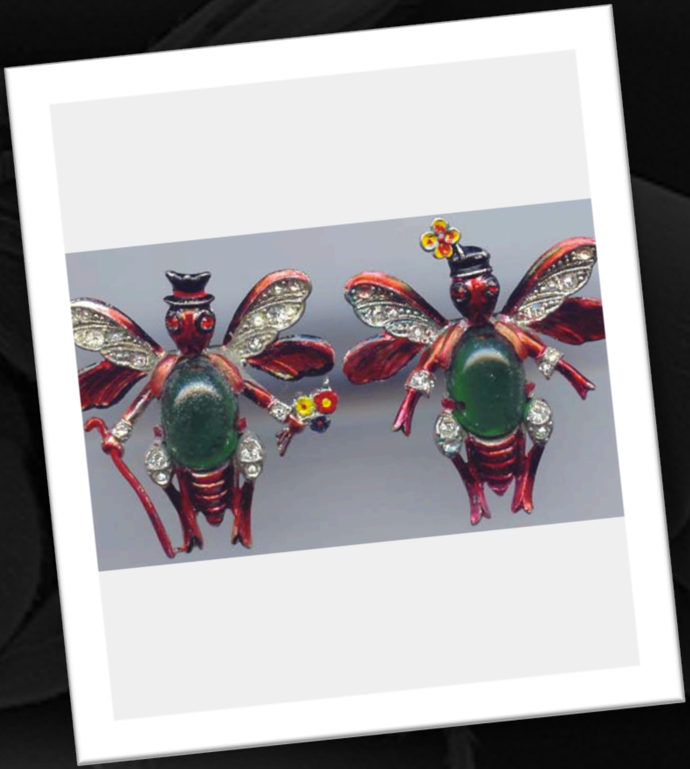
- Tractability
- Easy to rear
- Haplodiploidy
- Healthy isogenic inbred lines
- Visible & molecular markers
- Mutation screening in haploid sex: track genes (even complex genetics traits), positional cloning

… IN THE FIELD

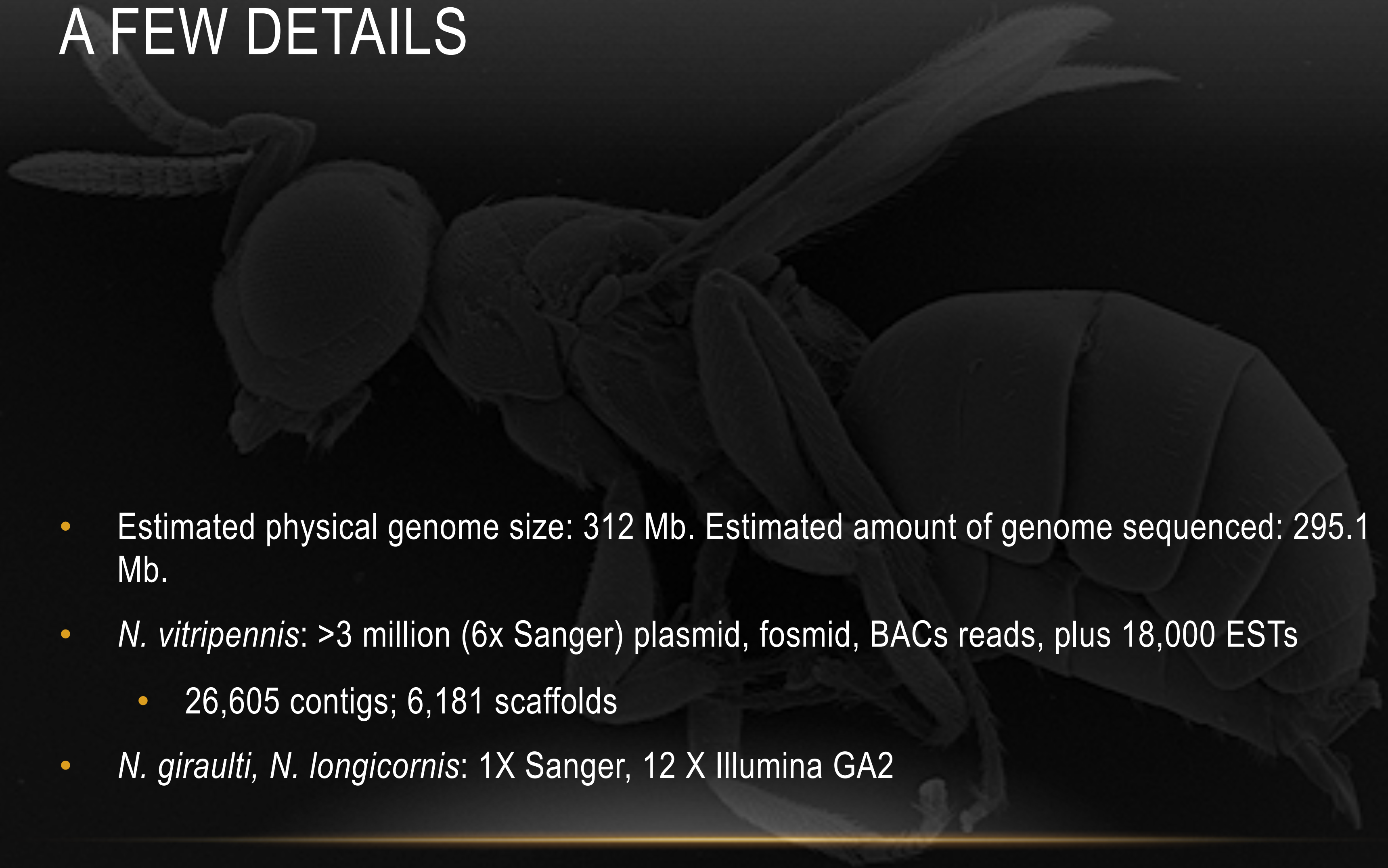
- Allopatry & microsympatry
- Parasitoid wasps belong in a group with more beneficial insects to humans than any other group!
- Hosts: *N. vitripennis* is a generalist; attacks blow flies (*L. cuprinea*), flesh flies (*Sarcophaga*), house flies. Other spp: *Protocalliphora* or ‘bird nest’ blow flies.
- Insect pests control

‘VOODOO’

- *Nasonia*’s venoms are not meant to kill, but to induce a zombie-like stage. A plethora of 79 venoms orchestrates paralysis and altered growth, creating a nursery-like environment for *Nasonia*’s larvae to thrive on.

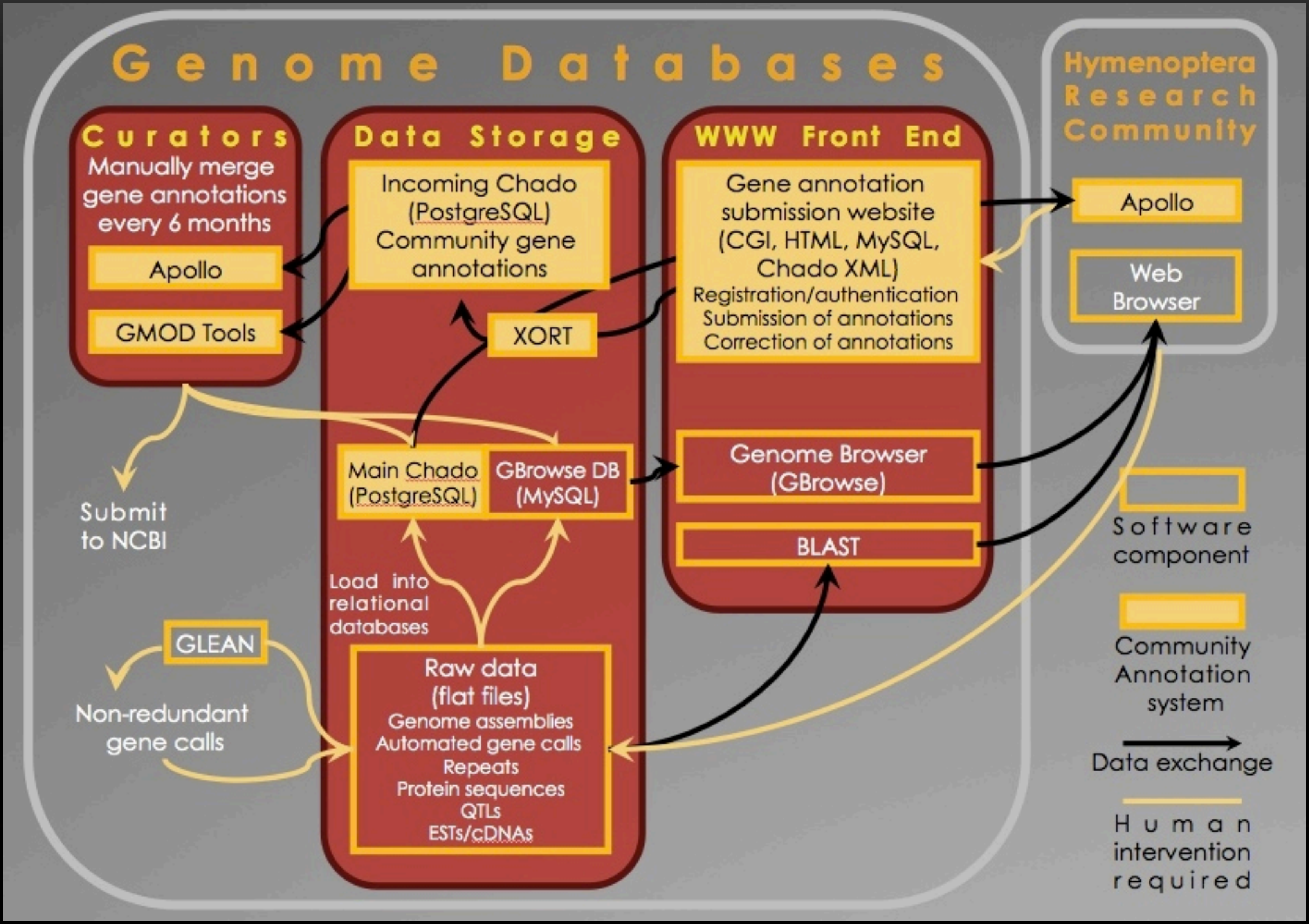


A FEW DETAILS



- Estimated physical genome size: 312 Mb. Estimated amount of genome sequenced: 295.1 Mb.
- *N. vitripennis*: >3 million (6x Sanger) plasmid, fosmid, BACs reads, plus 18,000 ESTs
    - 26,605 contigs; 6,181 scaffolds
- *N. giraulti, N. longicornis*: 1X Sanger, 12 X Illumina GA2

DATA EXCHANGE IN HGD


Genome Databases
Curators
Manually merge gene annotations every 6 months
Apollo
GMOD Tools
Data Storage
Incoming Chado (PostgreSQL) Community gene annotations
XORT
Main Chado (PostgreSQL)
GBrowse DB (MySQL)
Load into relational databases
Raw data (flat files)
Genome assemblies
Automated gene calls
Repeats
Protein sequences
QTLs
ESTs/cDNAs
WWW Front End
Gene annotation submission website (CGI, HTML, MySQL, Chado XML)
Registration/authentication
Submission of annotations
Correction of annotations
Genome Browser (GBrowse)
BLAST
Submit to NCBI
GLEAN
Non-redundant gene calls
Hymenoptera Research Community
Apollo
Web Browser
Software component
Community Annotation system
Data exchange
Human intervention required

METHODS | FINDING A HOMOLOG

- 1. Find a protein homolog of your gene of interest in GenBank, Ensembl or Uniprot.

  A heterologous homolog will help you most efficiently retrieve a sequence in your recently sequenced genome of interest. Alternatively, you may start with an EST or cDNA. In this example we will retrieve a homolog of the protein Dicer-1.

  Find a homolog of Dicer-1 using UniProt (http://www.uniprot.org/), Ensembl (http://www.ensembl.org/) or NCBI (http://www.ncbi.nlm.nih.gov/).

METHODS | KEEP IN MIND

When looking for heterologous homologs:

- It is best to use protein sequences to query the databases.
- Protein sequences diverge more slowly than DNA sequences, allowing for more sensitive blast searches. Human and Drosophila are the best species to start with, because they have the best annotated protein sets to date, thanks to large full-length cDNA projects and exhaustive curation.
- Additionally, draft genome sequences are available for six ant species and honey bee. In most cases, manual annotation efforts were also conducted.
- Swissprot (part of UniProt) is one of the best starting places to search protein sequences, thanks to their stringent curation parameters. Other databases include Ensembl and NCBI.

METHODS | GENOME COORDINATES

- 2. Locate your gene of interest in the *Nasonia* genome assembly using NasoniaBase BLAST (http://www.hymenopteragenome.org/nasonia/?q=blast).
  - A. Carefully choose one of the BLAST algorithms and determine the existence of an official gene model, homologous to your gene of interest.

    The standard BLAST output includes a summary of each hit in the database (describing the gene model identifier and E value), followed by alignments for each hit. Be sure to look both at scores and alignments to make the most informed decision about your wasp homolog.

METHODS | GENOME COORDINATES

- 2. Locate your gene of interest in the *Nasonia* genome assembly using NasoniaBase BLAST (http://www.hymenopteragenome.org/nasonia/?q=blast).
  - B. Carefully choose the appropriate BLAST algorithm to query the ‘Nasonia Scaffolds Assembly Nvit_1.0’ database and identify the exact location on the assembly.

    On the BLAST output, a link to the right of each alignment points to “See complete hit in GBrowse”. If the link is not active, visit the GBrowse pages and enter the gene model identifier as displayed on the BLAST results page. The most up to date links for GBrowse on NasoniaBase can be found under the ‘Tools’ tab.

METHODS | ACCESS TO APOLLO

- Download Apollo free of charge from http://apollo.berkeleybop.org/current/install.html
  - Available for Windows, Mac OS X, and Linux
  - Online and offline install options available
  - Requires Java (available for download)

METHODS | ACCESS TO APOLLO

- Get the *Nasonia* configuration files from the VM, located under /nescent/sessions/day4/curation/rest-of-the-directory

  Extract and open the 'what's-the-name-of-config-files-tar' folder and move its contents to the conf directory, replacing existent files if prompted. (If the extension '.txt' has been added, remove it).

  Mac OS X: go to Applications, "right-click" on Apollo icon and click on 'Show Package Contents'. Move all files to /Apollo/Contents/Resources/app/conf/. Alternatively, using a terminal mv files to /Applications/Apollo.app/Contents/Resources/app/conf

  Windows: conf directory located inside Apollo directory under Program Files.

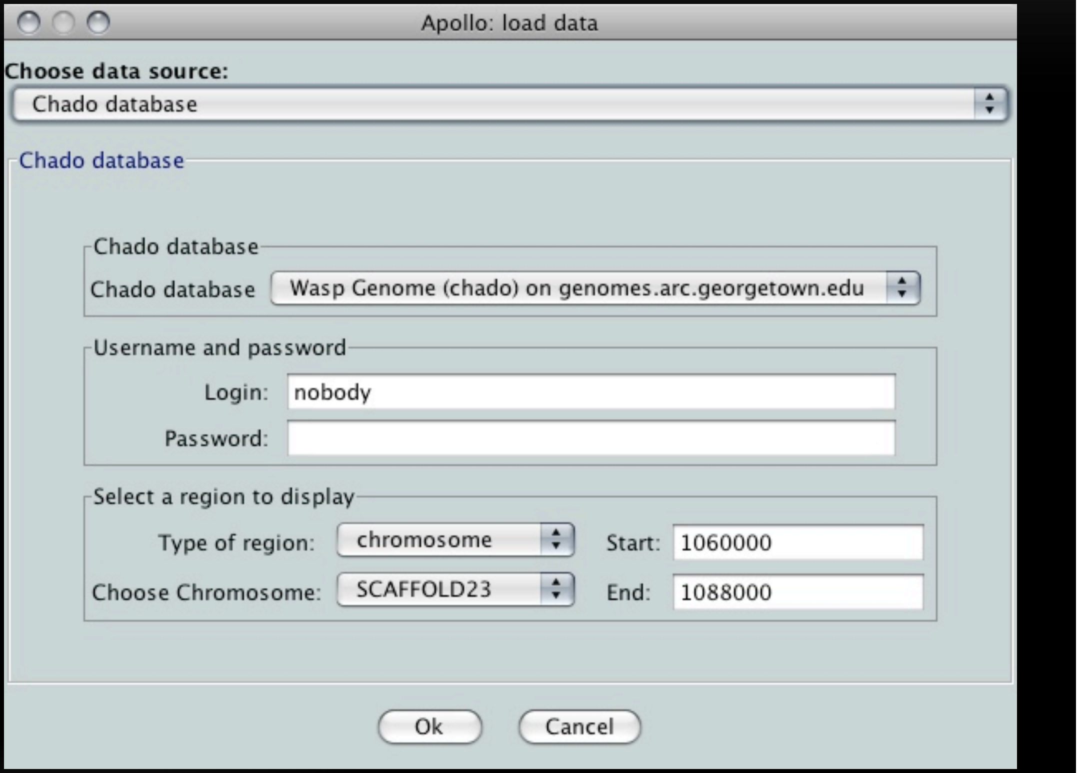
  Linux (on the VM): find the Apollo directory under and replace the extant 'conf' directory with the one you just decompressed.

METHODS | APOLLO CONFIGURATION FILES

- Include databases from other species (That’ s ok.)
- Requires some modification (Already done for you.)
    - Customized, species-specific configuration files
        - Connects to individual databases
        - Different types of evidence
        - Different colors for each evidence tracks

METHODS | LOADING DATA INTO APOLLO

- Choose ‘*Chado database*’ as your data source and ‘*Wasp Genome*’ as your Chado database.
- Login as ‘nobody’.
  No *Password* required.
- Choose ‘*chromosome*’ as your type of region. Choose ‘*SCAFFOLD####*’ as your chromosome of interest and enter the coordinates around your gene of interest.



Apollo: load data

Choose data source: Chado database

Chado database

Chado database

Chado database: Wasp Genome (chado) on genomes.arc.georgetown.edu

Username and password

Login: nobody

Password:

Select a region to display

Type of region: chromosome

Start: 1060000

Choose Chromosome: SCAFFOLD23

End: 1088000

METHODS | WHEN LOADING THE DATA

- Apollo loads 25Kbp of flanking sequence
  - good for extending gene models
- If it is too close to the edge of the scaffold:
  - Problem: The annotation is not displayed correctly
  - Solution: Reopen the region by manually entering scaffold coordinates that extend about 1 – 5 Kb away from both sides of the model.

METHODS | INITIAL RECONNAISSANCE AND ADJUSTMENTS

- Evidence information panel
- Resize window, reposition splitters
- ‘Working Area’: Blue area in the middle, marked by red arrowheads. ‘Drag’ the gene model/exons/regions you want to modify; all changes to a gene will be done while working on this ‘temporary’ model.
- **Important**: protein alignments do not necessarily reflect the entire length of the similar protein; non-conserved regions simply do not show up. Results: a short protein alignment or one with missing internal exons. Protein alignments may also be problematic in regions with tandem closely related genes; for example, aligning in part to one gene and then skipping over to align the rest to a second gene.

METHODS | ANNOTATING A SIMPLE CASE

- WHEN “The RefSeq prediction is correct, or nearly correct, assuming that no aligned data extends beyond the RefSeq and if so, it is not likely to be coding sequence, and/or the RefSeq prediction matches what you know about the gene”:
  - Can you add UTRs?
  - Check exon structures.
  - Check splice sites.
  - Check the ‘start’ and ‘stop’ sites.
  - Check the predicted protein product(s).

If the protein product still does not look correct, go on to “Annotating more complex cases”.

METHODS | ADDITIONAL FUNCTIONALITY

- You may also need to learn how to:
  - Search for a specific sequence
  - Get genomic sequence
  - Merge exons
  - Add/Delete an exon
  - Create an exon de novo (within an intron or outside existing annotations).
  - Right clicking or apple-clicking on a feature to get feature ID and additional information
  - Looking up homolog descriptions going to the accession web page at UniProt/ Swissprot

METHODS | ANNOTATING MORE COMPLEX CASES

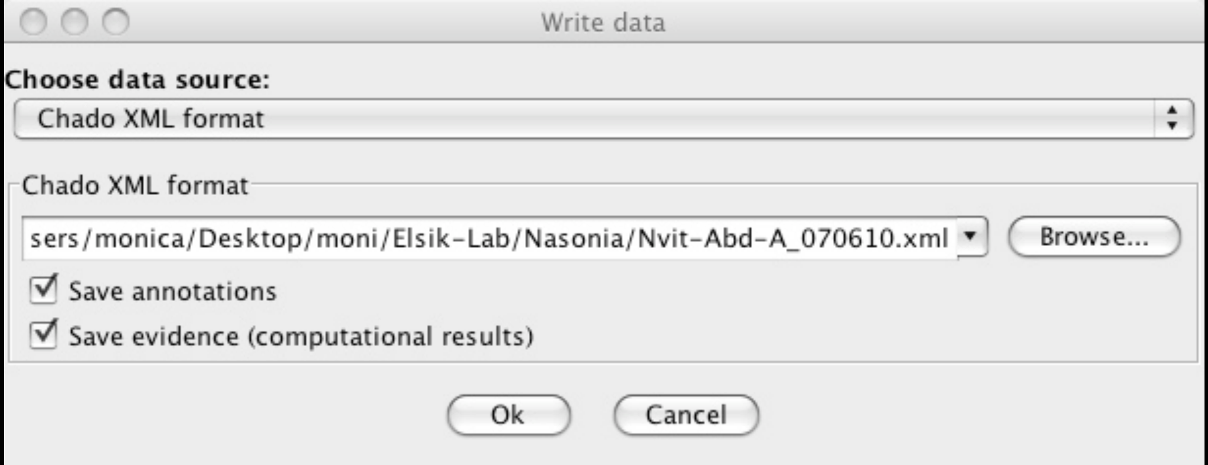
- Incomplete annotation: protein checks indicate gaps, missing 5’ sequences or missing 3’ sequences.
- Merge of 2 RefSeq predictions on same scaffold
- Merge of 2 RefSeq predictions on different scaffolds (uh-oh).
- Split of a RefSeq prediction
- Selenocysteine, single-base error, frameshift, and other inconvenient phenomena

METHODS | ADDING IMPORTANT PROJECT INFORMATION

- Canned and Customized Comments
- NCBI ID, RefSeq ID, gene symbol(s), common name(s), synonyms, top BLAST hits (GenBank IDs), orthologs with species names, and **anything else you can think of**, because you are the expert.
- Type of annotation (e.g. whether or not the gene model was changed)
- Data source (for example if the Fgeneshpp predicted gene was the starting point for your annotation)
- The kinds of changes you made to the RefSeq gene model, e.g: split, merge
- Functional description
- Whether you would like for a NasoniaBase curator to check the splice sites
- Whether part of your gene is on a different scaffold

METHODS | SAVING YOUR WORK

- You may save local copies of your **unfinished** work as GAME-XML files. Check the ‘Save evidence’ feature. (THIS IS VERY CLUNKY AND OFTEN DOESN’T WORK!!)
- Save Chado XML and GFF3 copies of your **finished** work.



Choose data source:

Chado XML format

Chado XML format

sers/monica/Desktop/moni/Elsik-Lab/Nasonia/Nvit-Abd-A_070610.xml

☑ Save annotations

☑ Save evidence (computational results)

ASSIGNMENT |
SMALL RNA-PROCESSING GENES IN NASONIA

Below is a list of genes known to be involved in processing small RNA molecules:

Dicer, Argonaute, Piwi, Aubergine, Spindle-E, Rm62, r2d2, Fmr1, vig, Tudor-SN, Gawky, Armitage, Belle, CG10883-PA, CG17265-PA, Dcp, Drosha, Headcase, Loquacious, Pasha, Pacman, , SID, eri-1