



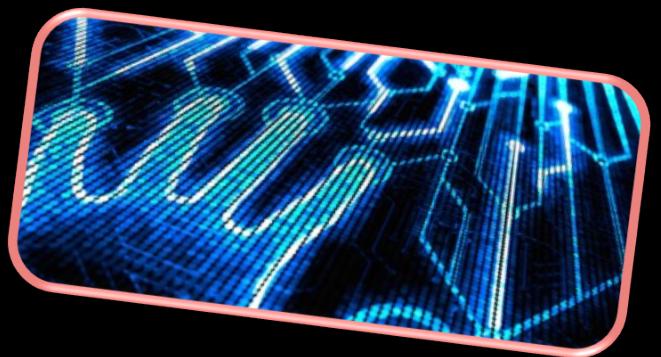
# Apollo:

## Collaborative Genome Annotation Editing

Monica Munoz-Torres, PhD | @monimunozto  
Phoenix Bioinformatics - for Lawrence Berkeley National Laboratory  
A workshop for EMBL-ABR. 02 November, 2017

# Today...

We will learn effective ways to extract valuable information about a genome through curation efforts.



# After this workshop, you will:

- Better understand curation in the context of genome annotation:  
**assembled genome → automated annotation → manual annotation**
- Become familiar with Apollo's environment and functionality.
- Learn to identify homologs of known genes of interest in your newly sequenced genome.
- Learn how to corroborate and modify automatically annotated gene models using all available evidence in Apollo.



# Schedule

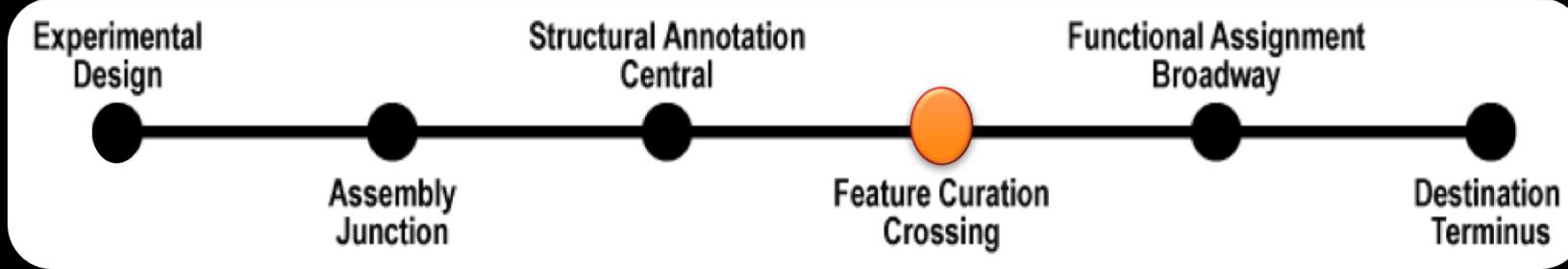
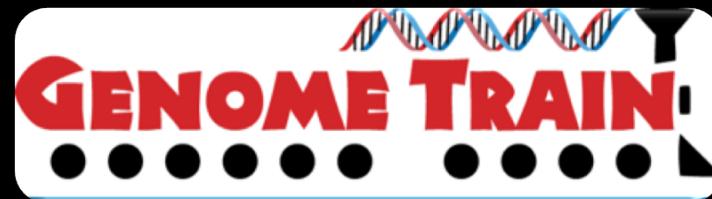


|                                   |           |
|-----------------------------------|-----------|
| 1. Genome Curation:               | 7 minutes |
| 2. Predicting & annotating genes: | 7 min.    |
| 3. Apollo - intro & examples:     | 20 min.   |
| 4. Hands-on practice              | 40 min.   |
| 5. Break                          | 6 min.    |
| 6. Hands-on practice (ctd.)       | 40 min.   |

# Slides

<http://bit.ly/apollo-ags-intro2>

<http://bit.ly/apollo-ags-edit2>

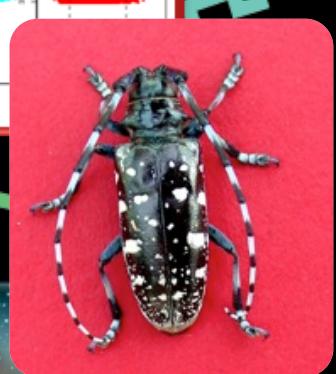
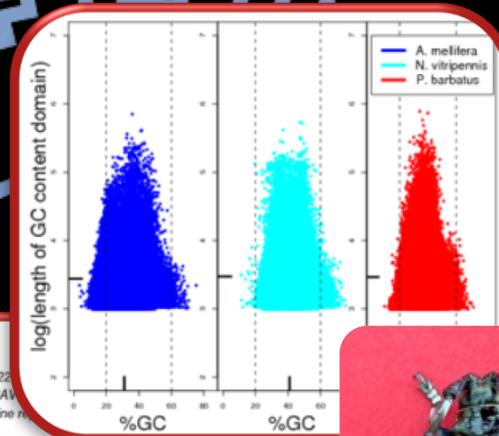
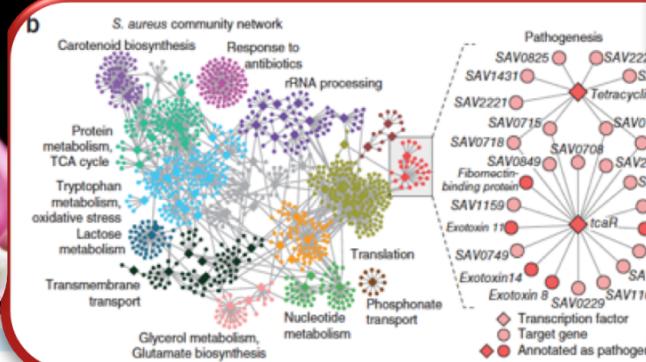
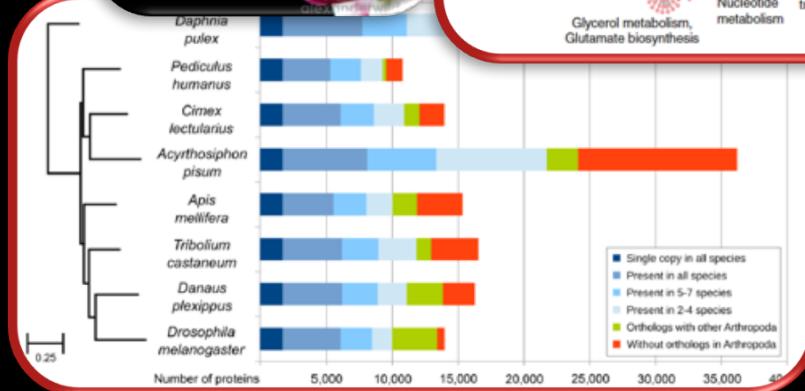




# Genome Curation

Extracting knowledge from data

# Unlocking genomes



Marbach et al. 2011. *Nature Methods* | Shutterstock.com | Alexander Wild



# Good genes are required!



## 1. Generate gene models

- A few rounds of gene prediction.

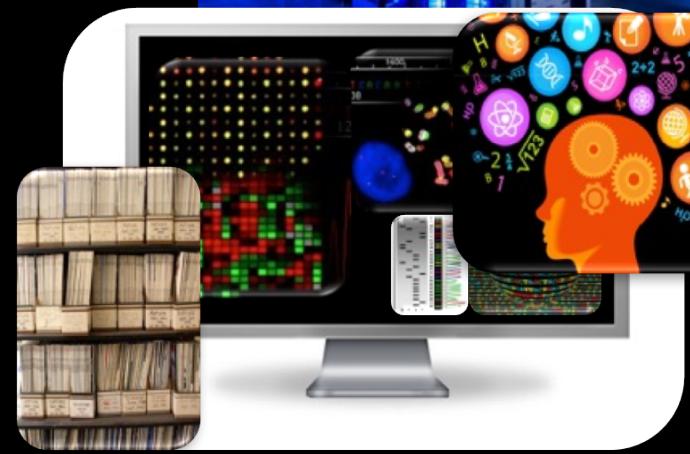


## 2. Annotate gene models

- Function, expression patterns, metabolic network memberships.

## 3. Manually review them

- Structure & Function.



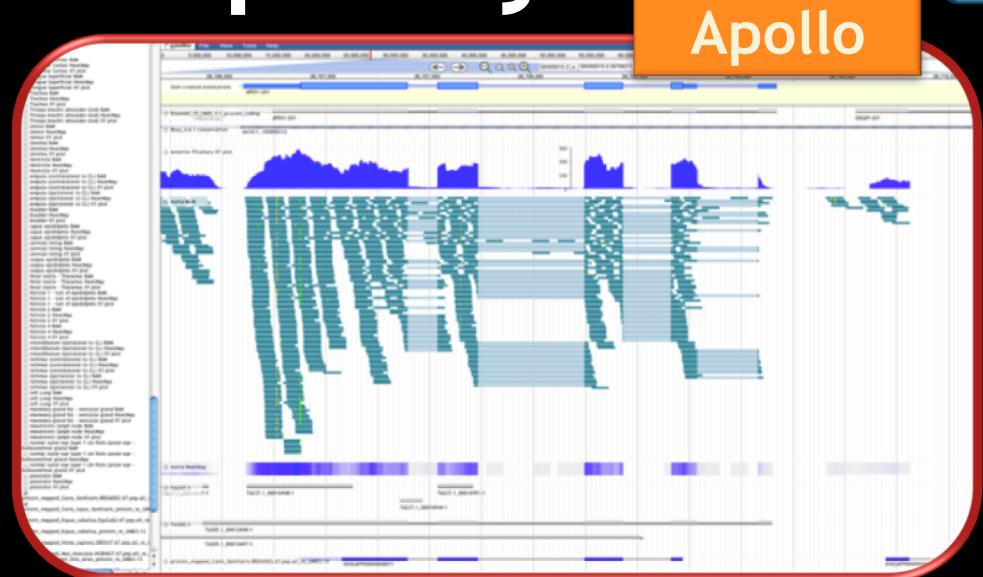
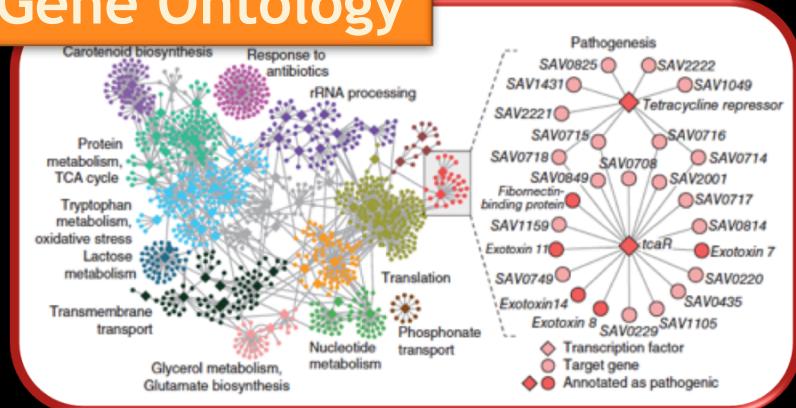
# Curation improves quality



Apollo

Best representation of biology & removal of elements reflecting errors in automated analyses.

## Gene Ontology



Functional assignments through comparative analysis using literature, databases, and experimental data.

# Curation is valuable:



- To make accurate orthology assessments
- To accurately annotate expanded / contracted gene families
- To identify novel genes, species-specific isoforms
- To efficiently take advantage of transcriptomic analyses

# Curation is inherently collaborative



- It is impossible for a single individual to curate an entire genome with precise biological fidelity.
- Curators need second opinions and insights from colleagues with domain and gene family expertise.

SCALE

A white funnel icon with blue liquid inside, positioned next to the word "SCALE".

EXPERTISE

A yellow lightbulb icon with rays of light, positioned next to the word "EXPERTISE".

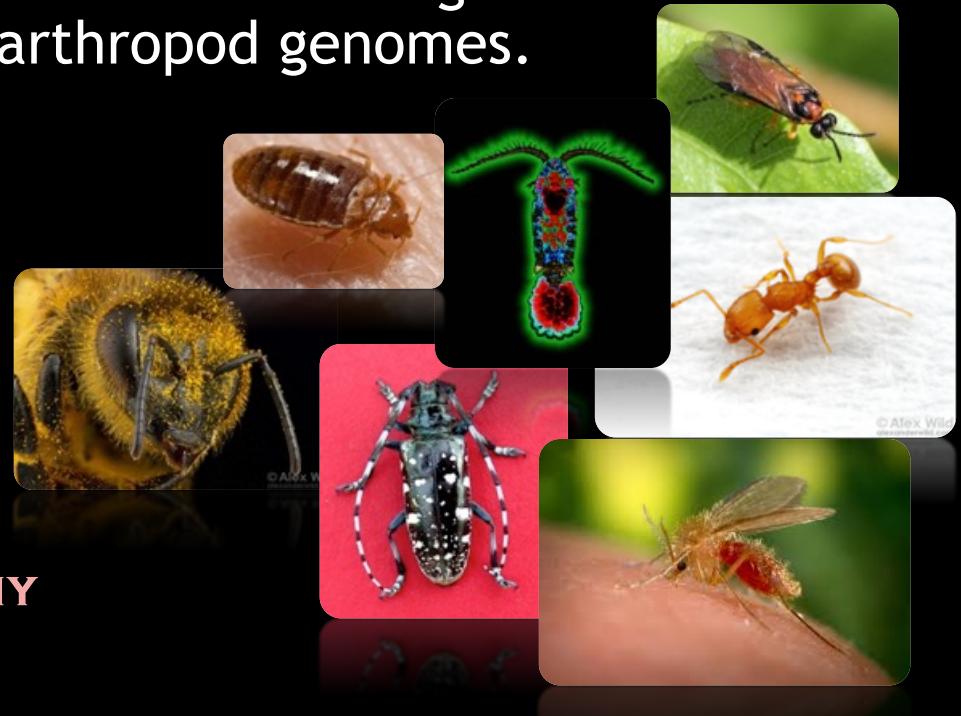


# i5k - five thousand arthropod genomes

<http://i5k.github.io>

- Transformative, broad, & inclusive initiative to organize sequencing and analysis of 5,000 arthropod genomes.

- **WORLDWIDE AGRICULTURE**
- **FOOD SAFETY**
- **MEDICINE**
- **ENERGY PRODUCTION**
- **MODELS IN BIOLOGY**
- **MOST ECOSYSTEMS**
- **EVERY BRANCH OF THE PHYLOGENY**





A transformative, broad, & inclusive initiative to organize sequencing and analysis of 5,000 arthropod genomes

#### **FOCUSES ON SPECIES KNOWN TO BE IMPORTANT TO:**

- **WORLDWIDE AGRICULTURE**
- **FOOD SAFETY**
- **MEDICINE**
- **ENERGY PRODUCTION**
- **MODELS IN BIOLOGY**
- **Most Ecosystems**
- **EVERY BRANCH OF THE PHYLOGENY**



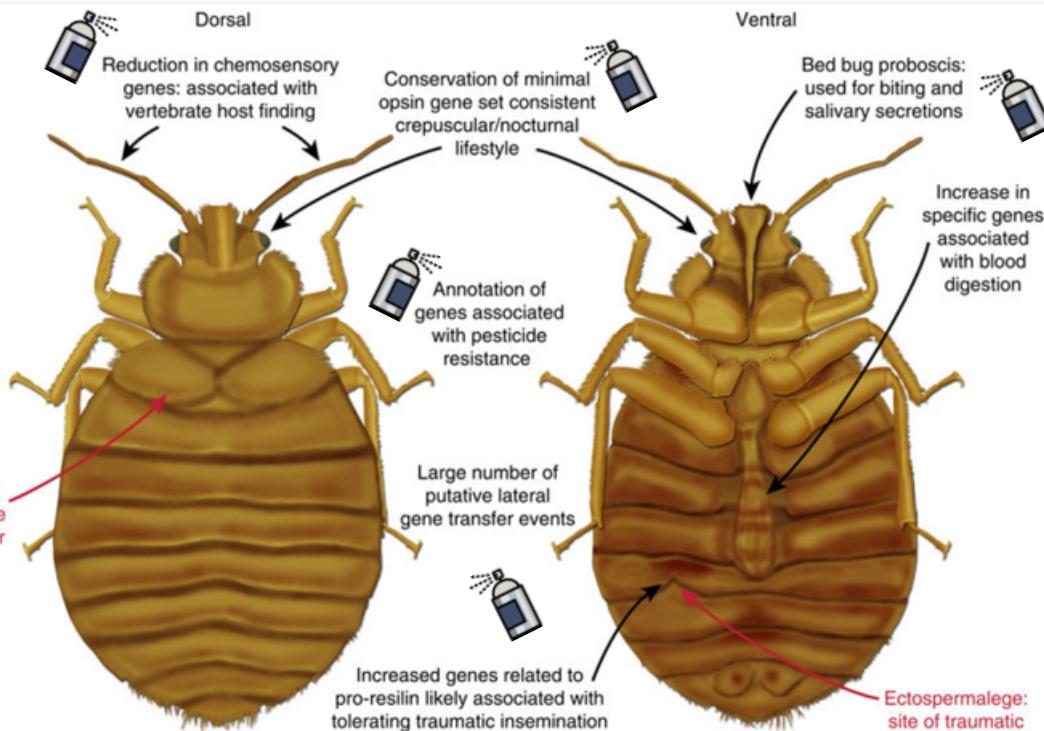
# The bed bugs, they're back!

Benoit et al. (2015) *Nature Communications*. doi:10.1038/ncomms10165

~80 Curators!

- International Travel and Commerce
- Increased Insecticide Resistance

<http://i5k.github.io>



Red, general characteristics of bed bugs; black, key aspects identified and expanded by genome sequencing and manual curation.

- Timely resource for biology of human ectoparasites.
- Discovery of new targets for control.
- Common lab strain collected before introduction of pyrethroid insecticides.
  - What triggered the current bed bug resurgence?
  - Did bed bugs originate from one or multiple sources?
- Studies on mechanisms that hinder vertebrate pathogen survival & proliferation and transmission.





# Predicting & annotating gene structures



# Gene Prediction & Gene Annotation

Identification and annotation of genomic elements:

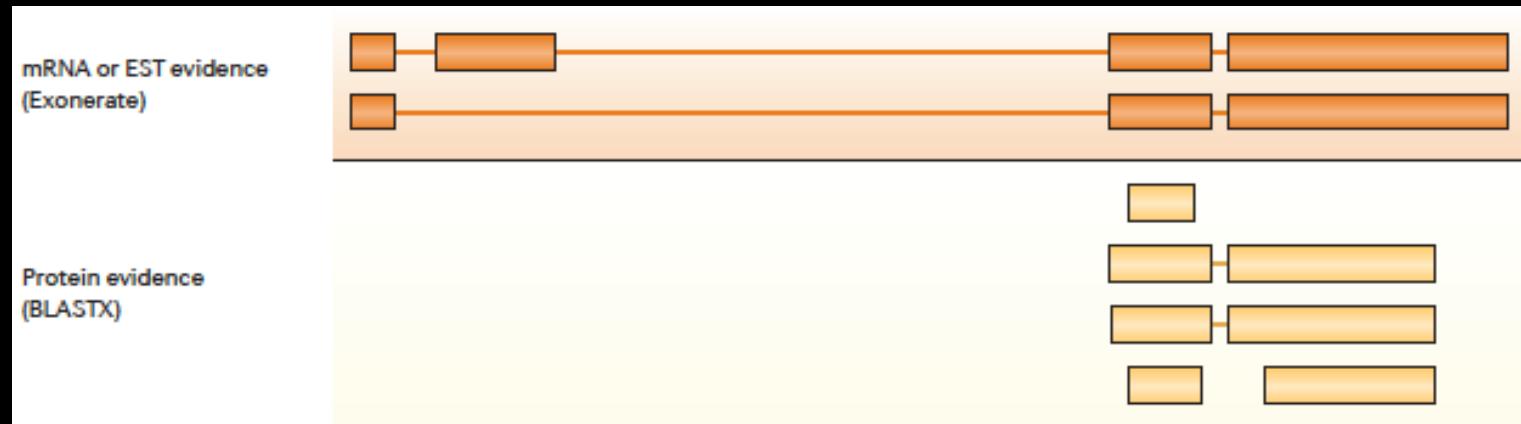
- Primarily focuses on protein-coding genes.
- Also identifies RNAs (tRNA, rRNA, long and small non-coding RNAs (ncRNA)), regulatory motifs, repetitive elements, etc.
- Happens in 2 steps:
  - Computation phase
  - Annotation phase



# Computation Phase

## 1) Experimental data are aligned to the genome:

RNA-sequencing reads, proteins, etc.



Yandell & Ence. *Nature Rev* 2012 doi:10.1038/nrg3174



# Computation Phase

## 2) Gene predictions are generated:

2a) *Ab initio*: based on nucleotide sequence and composition  
e.g. Augustus, fgenesh, etc.

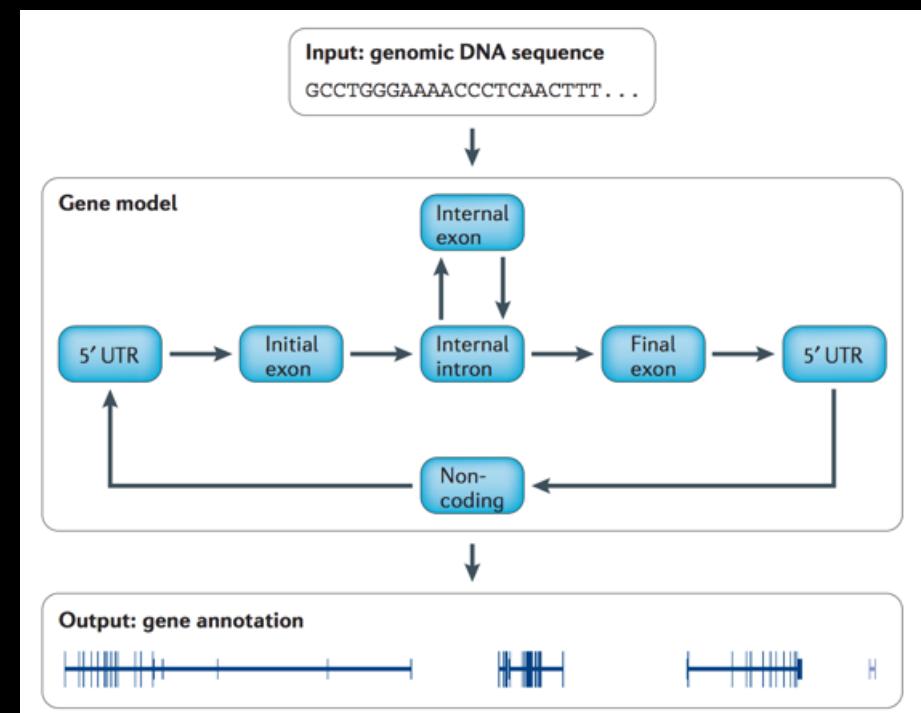
2b) Using experimental evidence: identifying domains and motifs  
e.g. SGP2, JAMg, fgenesh++, etc.



# Gene Prediction - methods for discovery

## 2a) *Ab initio*:

- Based on DNA composition
- Deals strictly with genomic sequences
- Makes use of statistical approaches (e.g. HMM) to search for coding regions and typical gene signals
  - E.g. Augustus, fgenesh, etc.



# Gene Prediction - methods for discovery

## 2b) Evidence-based:

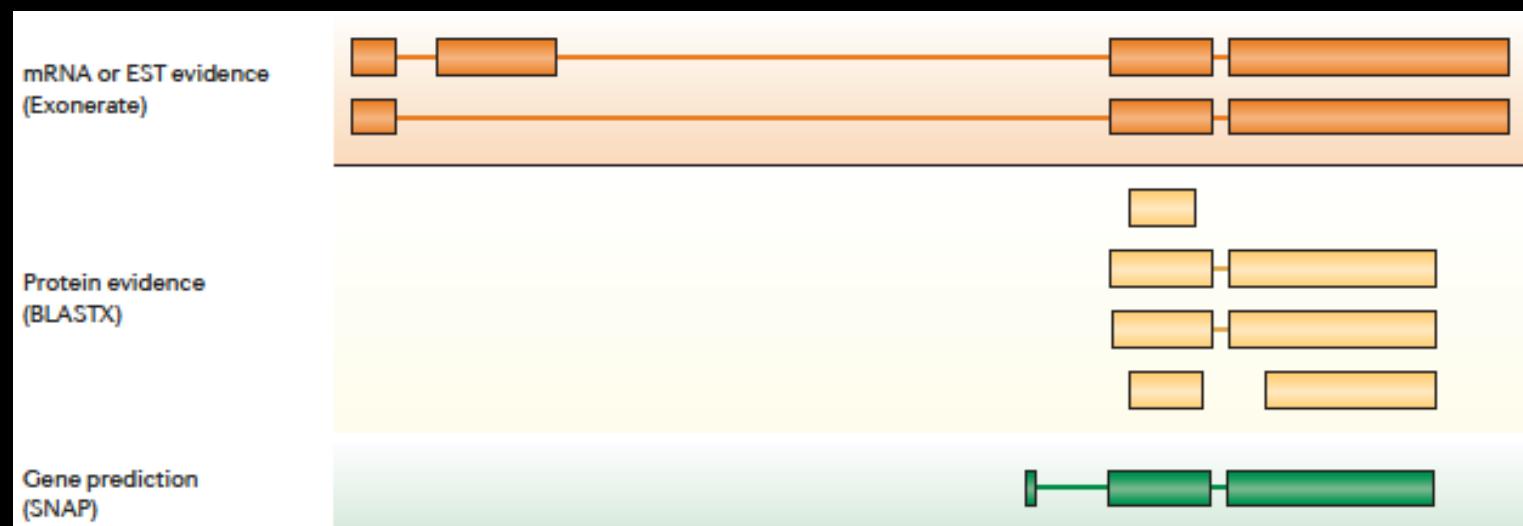
Finds genes using either similarity searches against public databases or other experimental data sets e.g. RNAseq.

E.g: SGP2, fgenesh++, JAMg, etc.



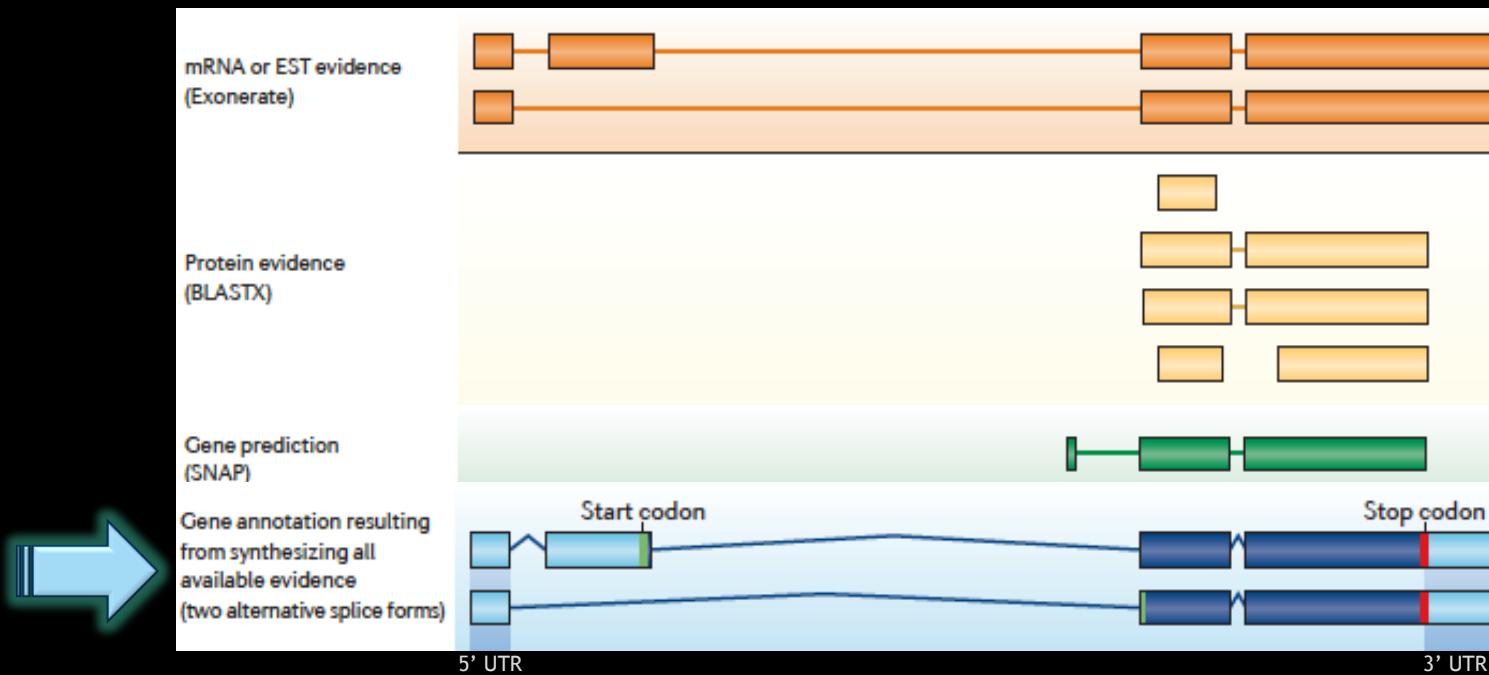
# Computation Phase: result

- The single most likely coding sequence, no UTRs, no isoforms.



# Annotation Phase

- Data from experimental evidence and prediction tools are synthesized into a reliable set of structural gene annotations.



**Result:** gene models that generally include UTRs, isoforms, evidence trails.

# Consensus Gene Sets

Gene models may be organized into sets using:

- Combiners for automatic integration of predicted sets
  - e.g: GLEAN, EvidenceModeler, etc.
- Tools packaged into pipelines
  - e.g: MAKER, PASA, Gnomon, Ensembl, etc.



# Challenges



## *Ab initio*

- + can capture species-specific or highly-divergent genes
- false positive predictions (incomplete predictions, readthrough predictions)
- not enough on its own to establish orthology

## Reference-guided

- + uses reliable gene orthologs from better-annotated species
- can miss species-specific genes and other sequences
- not enough on its own to establish orthology

# Some suggestions



- Hybrid reference-guided & *ab initio* gene prediction
- Generate transcriptomic data to confirm predictions, extend & improve models, identify new expressed loci.
  - the more tissues, the better!
- Review synteny to verify orthologous assignments
  - largely manual for now.



# Annotating gene functions



# Functional Annotation

Attaching metadata to structural annotations for the purpose of assigning a particular function.

- Assignments do not necessarily have to be supported by your own experimental data.
- Sequence similarity approaches must be informed and validated by evolutionary theory, not just a score value.



# Gene Ontology

GeneOntology.org

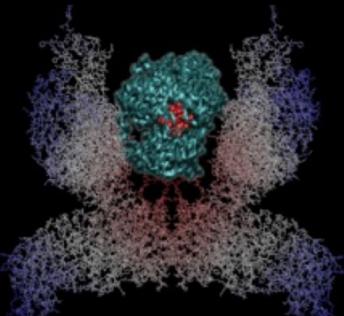


Terms (classes) arranged in a graph: molecular functions, biological processes, cellular locations, and the relationships connecting them all, in a species-independent manner.

## 1. Molecular Function

An elemental activity or task or job

- protein kinase activity
- insulin receptor activity



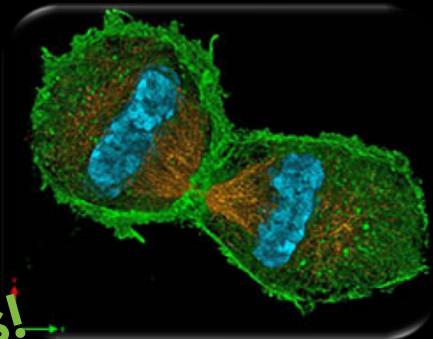
Insulin Receptor  
Petrus et al, 2009, *ChemMedChem*

~150 Contributors!

## 2. Biological Process

A commonly recognized series of events

- cell division

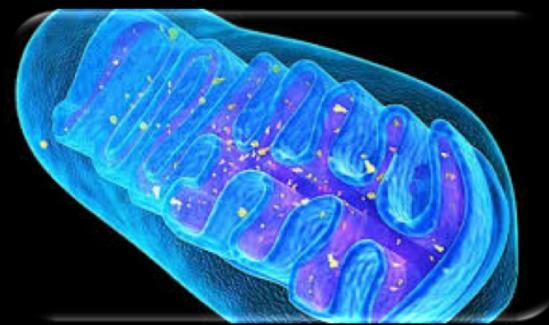


End of Telophase.  
Lothar Schermerle

## 3. Cellular Component

Where a gene product is located

- mitochondria
- mitochondrial matrix
- mitochondrial inner membrane



Mitochondrion.  
PaisekaScience Photo Library



**Collaboratively  
curating gene structures**



# General process of curation

1. Select or find a **region of interest** (e.g. scaffold).
2. Select appropriate **evidence** tracks to review the genome element to annotate (e.g. gene model).
3. Determine whether a feature in an existing evidence track will provide a reasonable **gene model** to start working.
4. If necessary, **adjust** the gene model.
5. Check your edited gene model for **integrity and accuracy** by comparing it with available homologs.
6. **Comment** and finish.





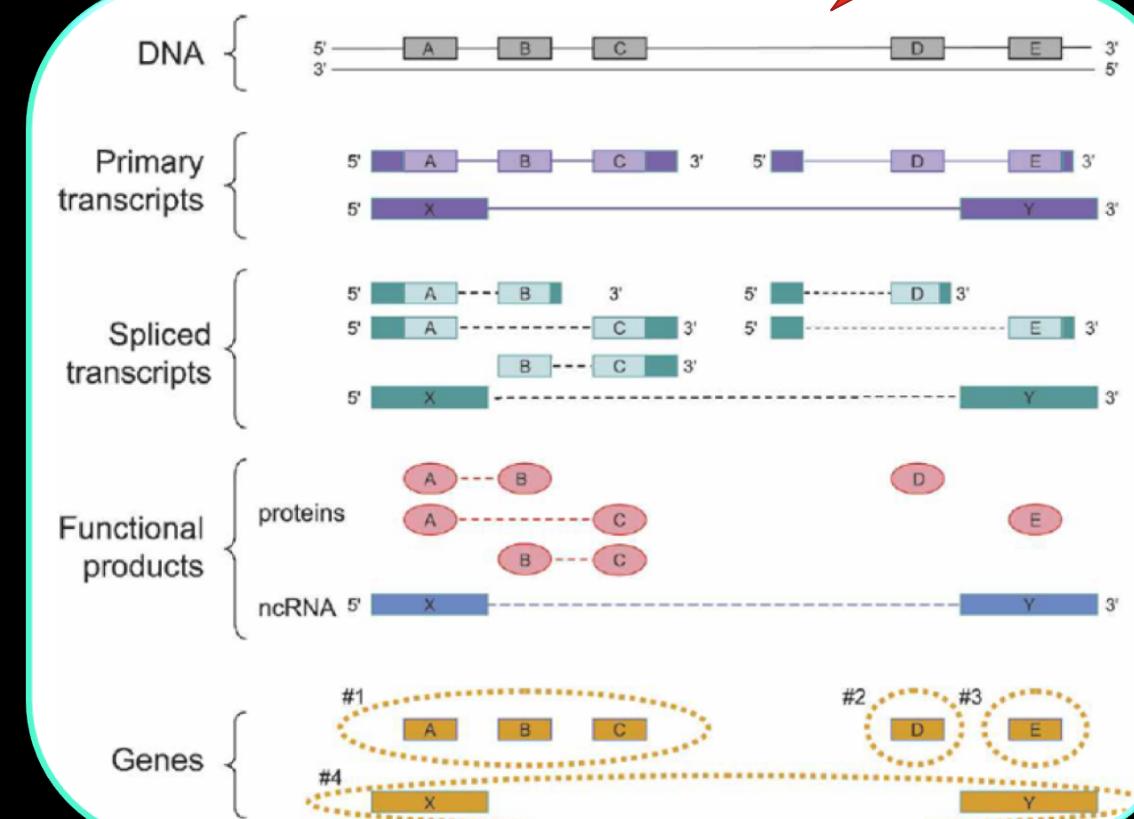
Bioreresher

# A brief refresher

# The gene: *a moving target*

Biorefresher

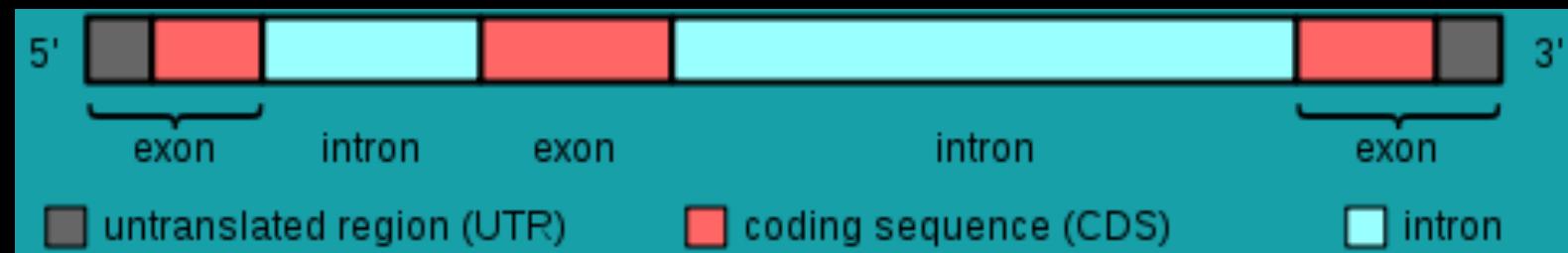
“The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.”



Gerstein et al., 2007. Genome Res

Bioreresher

# mRNA



"Gene structure" by Daycd- Wikimedia Commons

# Reading frames

In eukaryotes, only one reading frame per section of DNA is biologically relevant at a time: can be transcribed into RNA and translated into protein.

## OPEN READING FRAME (ORF)

ORF = Start signal + coding sequence (divisible by 3) + Stop signal

# Splice sites

Splicing “signals” (from the point of view of an intron):

- 5' end splice “signal” (site): usually GT (less common: GC)
- 3' end splice site: usually AG

**...] $5'$  - GT / AG -  $3'$ [...**

Alternatively bringing exons together produces more than one protein from the same genic region: isoforms.

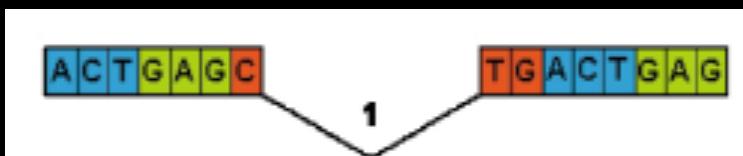
# Exons and Introns

Bioreresher

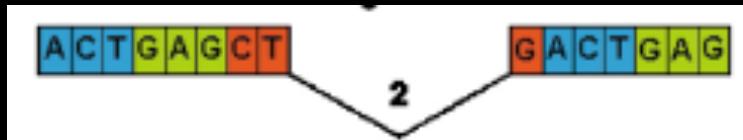
- Introns can interrupt the reading frame of a gene by inserting a sequence between two consecutive codons



- Between the first and second nucleotide of a codon



- Or between the second and third nucleotide of a codon



# Obstacles to transcription and translation

- Premature *Stop* codons in the message: A process called **non-sense mediated decay** checks and corrects them to avoid incomplete splicing, DNA mutations, transcription errors, and leaky scanning of ribosome - which can cause changes in the reading frame (frame shifts).
- Insertions and deletions (**indels**) can cause frame shifts when the indel is not divisible by three. As a result, the peptide can be abnormally long, or abnormally short - depending on when the first in-frame *Stop* signal is located.



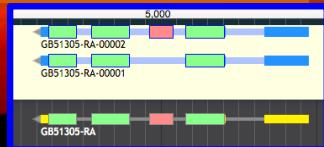
# Functionality overview

# Apollo Genome Annotation Editor

Collaborative, instantaneous,  
web-based, built on top of JBrowse.

GenomeArchitect.org

★ Color by CDS frame, toggle strands,  
set color scheme and highlights.

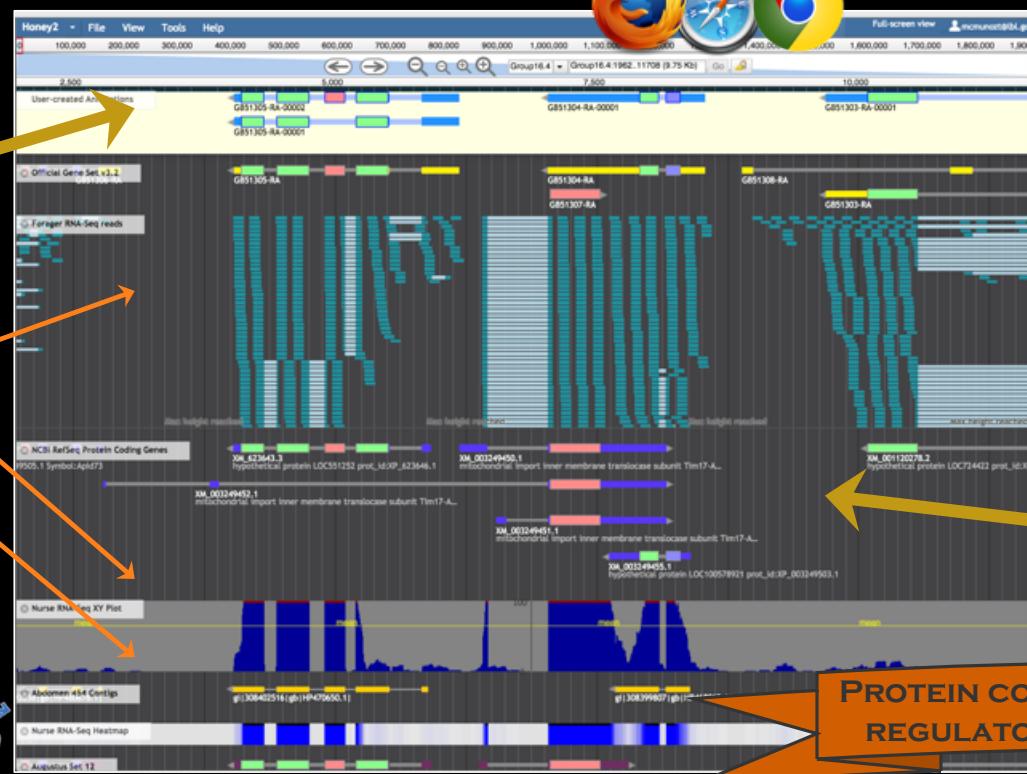


★ Query the genome using BLAT.

★ Navigate and zoom.

★ Search for a gene model  
or a scaffold.

★ Upload evidence files (GFF3, BAM, BigWig),  
add combination and sequence search tracks.

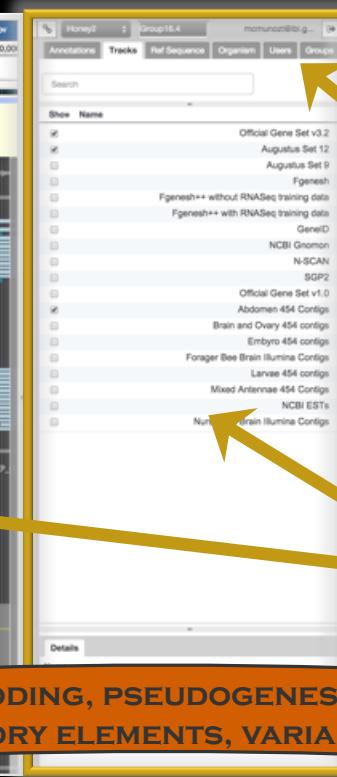


★ User-created annotations.

★ Stage and cell-type  
specific transcription  
data.



PROTEIN CODING, PSEUDOGENES, ncRNAs,  
REGULATORY ELEMENTS, VARIANTS, ETC.



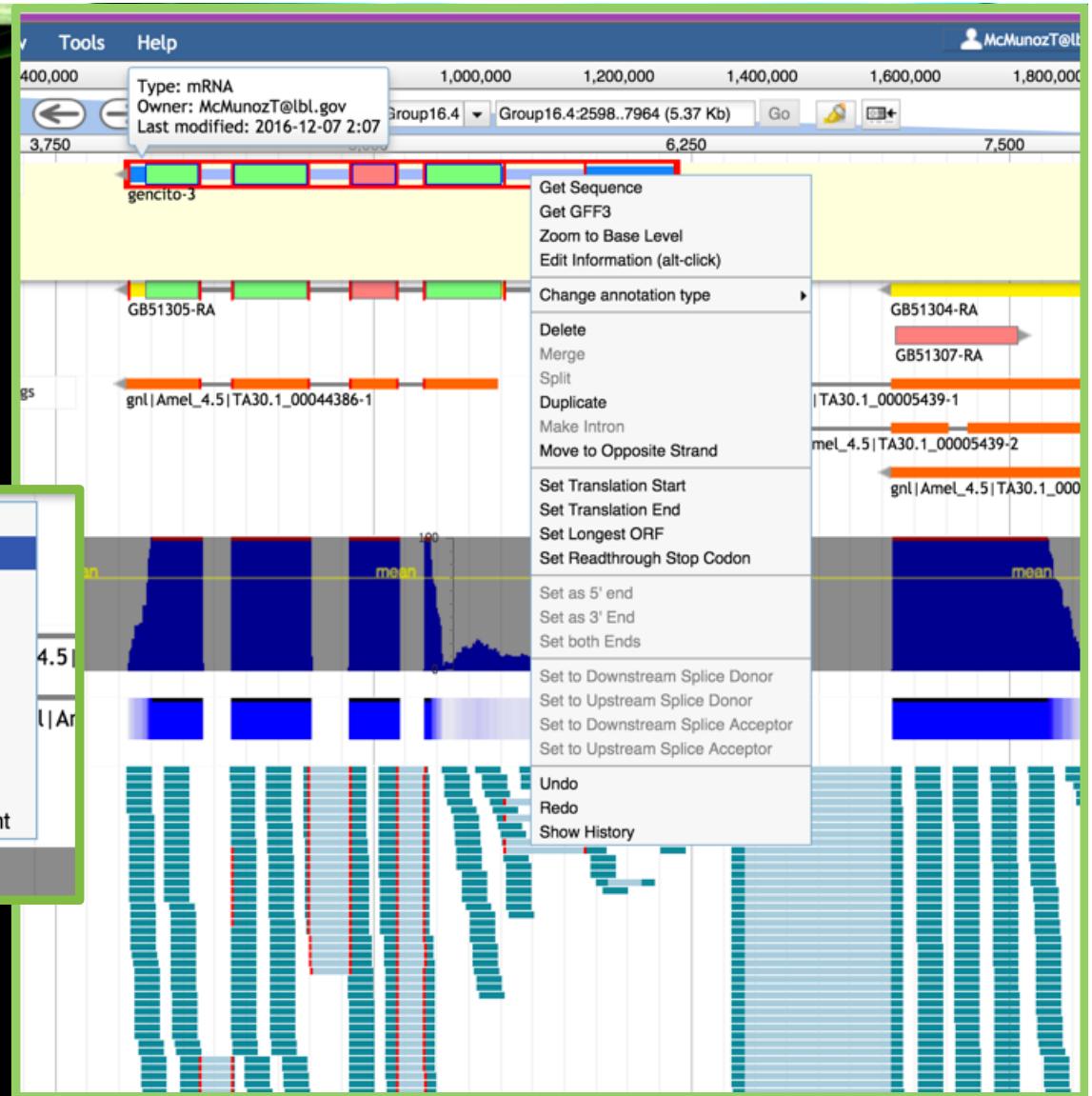
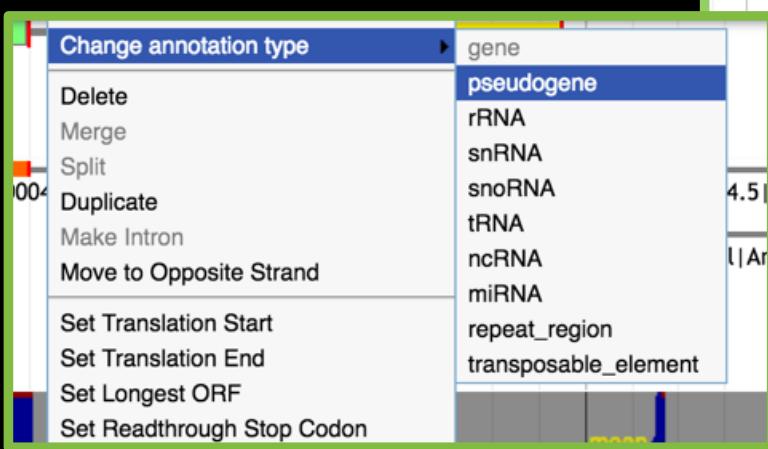
Admin

★ Annotator panel.



Evidence Tracks. ★

# Right-click functionality



GenomeArchitect.org

# Apollo

## Export



Annotations Tracks Ref Sequence Organism Users Groups Admin

Search

Length Minimum Maximum

Export All Selected (2) None

GFF3 FASTA

Honey2  
2 exported  
Type: GFF3

GFF3  GFF3 with FASTA Export Annotations Close

Export

Honey2  
2 exported  
Type: FASTA

Genomic  cDNA  CDS  Peptide Export Annotations Close

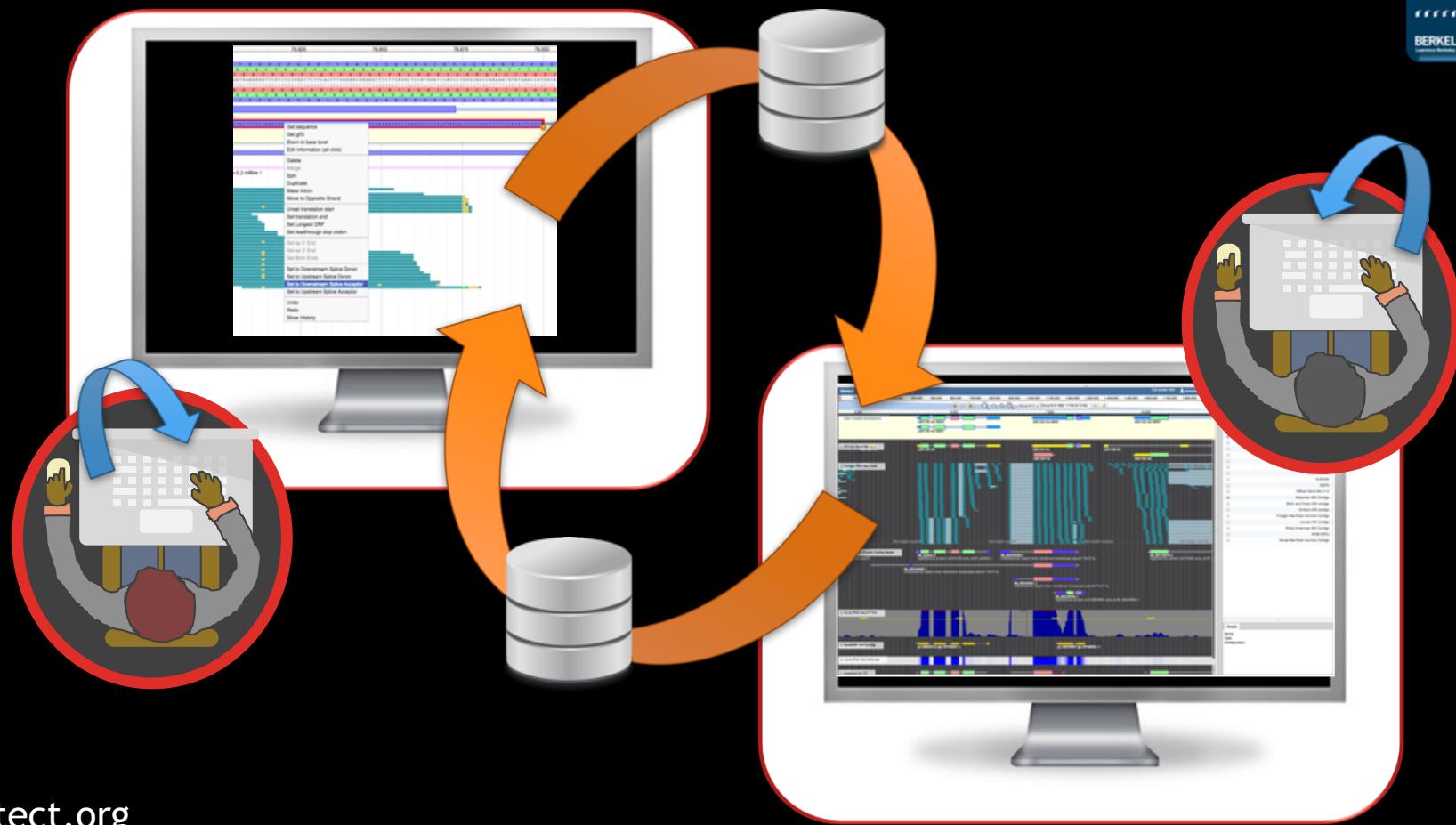
GroupUn5044 ▲ Length 540 560 564



GenomeArchitect.org

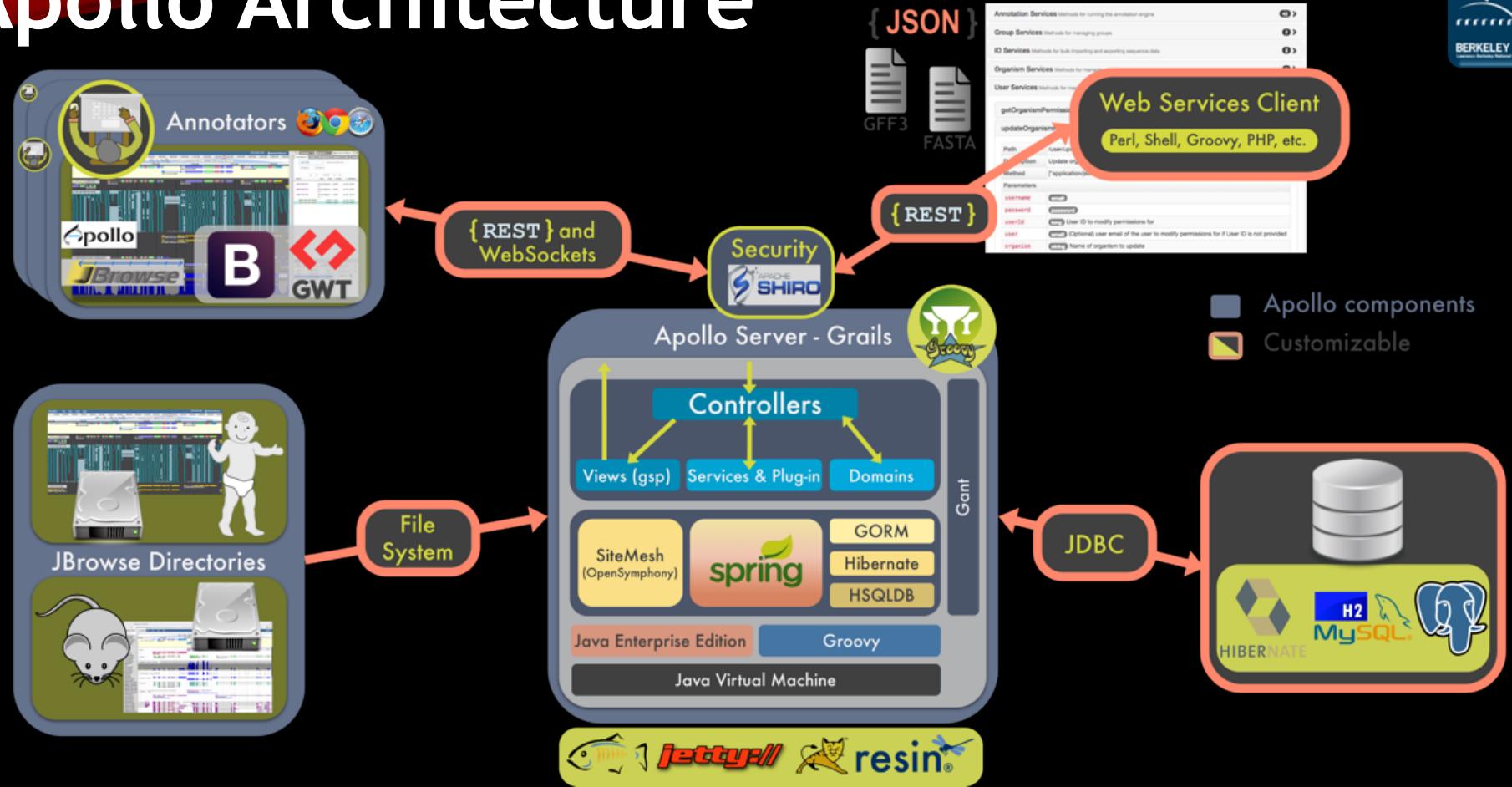
# Apollo

Collaboration in real time

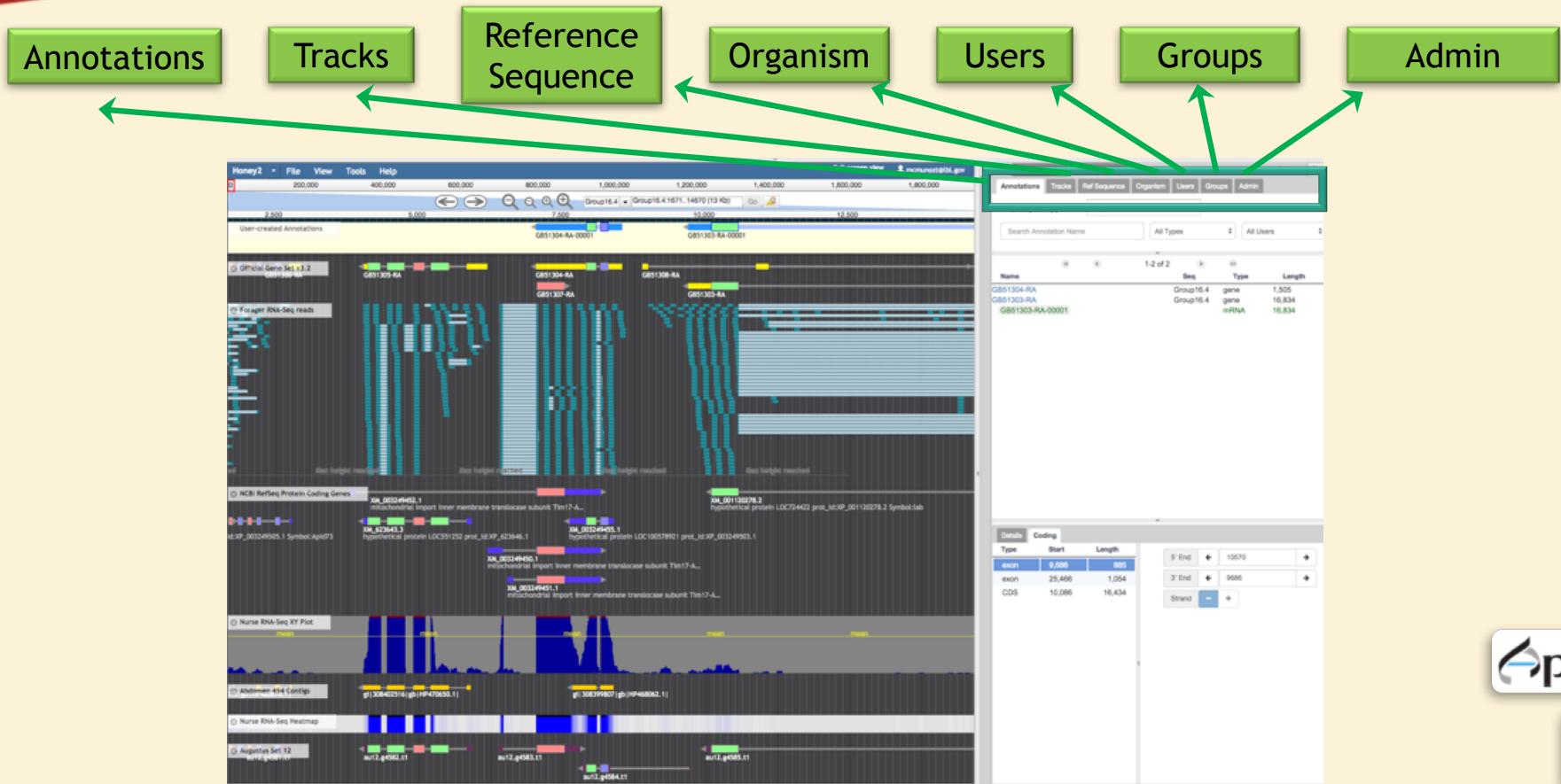


GenomeArchitect.org

# Apollo Architecture



# Removable Annotator Panel



# Annotation details & exon boundaries

Annotations

Honeybee Group16.4 pepita@mendiet...

Annotations Tracks Ref Sequence

Annotation Name All Types Reference Sequence All Users Go to Annotation

1-2 of 2

| Name     | Seq       | Type | Length | Updated      |
|----------|-----------|------|--------|--------------|
| spel1    | Group1.33 | gene | 10,934 | Mar 03, 2017 |
| spel1-RA |           | mRNA | 10,934 | Mar 03, 2017 |
| test1-RA | Group16.4 | gene | 49,826 | Mar 21, 2017 |

gene

spel1 RA

mRNA

1

Details Coding

Name spel1-RA

Description

Location 206752 - 217685 strand(+)

Ref Sequence Group1.33

Owner McMunozT@lbl.gov

2

Coding

| Type | Start   | Length |
|------|---------|--------|
| exon | 214,881 | 395    |
| exon | 217,109 | 93     |
| exon | 213,802 | 124    |
| exon | 214,334 | 168    |
| exon | 206,876 | 50     |

5' End 217109  
3' End 217201  
Strand - +



# Navigating to an annotation

## Annotations

The screenshot shows the Apollo genome annotation tool interface. At the top, there is a header with a back arrow, the species "Honeybee", the group "Group16.4", and a user "pepita@mendiet...". Below the header, there are three tabs: "Annotations" (which is selected and highlighted in red), "Tracks", and "Ref Sequence".  
The main area contains search and filter fields: "Annotation Name" (empty), "All Types" (set to "All"), "Reference Sequence" (empty), "All Users" (set to "All"), and a "Go to Annotation" button.  
A modal window is open, showing a dropdown menu for "All Types" which includes "Gene", "Pseudogene", "Transposable Element", and "Repeat Region".  
The results table displays two entries:

| Name     | Seq       | Type | Length | Updated      |
|----------|-----------|------|--------|--------------|
| spel1    | Group1.33 | gene | 10,934 | Mar 03, 2017 |
| spel1-RA |           | mRNA | 10,934 | Mar 03, 2017 |
| test1-RA | Group16.4 | gene | 49,826 | Mar 21, 2017 |

  
Annotations are highlighted with green boxes:

- "gene" is highlighted in a green box above the first row.
- "mRNA" is highlighted in a green box below the second row.
- "spel1" is highlighted in a blue box in the "Name" column of the first row.
- "spel1-RA" is highlighted in a green box in the "Name" column of the second row.

A blue arrow points from the "spel1" highlight to the "Name" column of the first row. A green arrow points from the "mRNA" highlight to the "Name" column of the second row.



# Displaying tracks with supporting data

The screenshot shows the JBrowse interface for the Honeybee genome. The top navigation bar includes 'Honeybee' and 'Group16.4'. The 'Tracks' tab is selected, highlighted with a red border. A callout box points to the 'Available Tracks' panel, which lists numerous tracks such as 'Abdomen 454 Contigs', 'Apis cerana reads', 'Augustus Set 12', 'Brain and Ovary 454 contigs', 'Cfpo\_00013.3', 'Embryo 454 contigs', 'Fgenesh', 'Fgenesh++ with RNASeq training data', 'Fgenesh++ without RNASeq training data', 'Forager RNA-Seq HeatMap', 'Forager RNA-Seq XY Plot', and 'Forager RNA-Seq reads'. The 'Official Gene Set v3.2' track is selected and displayed in the main viewer area, showing gene models and annotations. The viewer area also includes a 'User-created Annotations' section and a zoom control.

Available Tracks

- Abdomen 454 Contigs
- Acsp\_00012.2
- Al\_00013.3
- Apis cerana reads
- Augustus Set 12
- Augustus Set 9
- Brain and Ovary 454 contigs
- Cfpo\_00013.3
- Embryo 454 contigs
- Fgenesh
- Fgenesh++ with RNASeq training data
- Fgenesh++ without RNASeq training data
- Forager RNA-Seq HeatMap
- Forager RNA-Seq XY Plot
- Forager RNA-Seq reads

Honeybee - File View Tools Help

Group16.4

User-created Annotations

Official Gene Set v3.2

JBrowse Track Selector Show Hide JBrowse Track Selector

Show Name

| Show                                | Name                                   |
|-------------------------------------|--|
| <input checked="" type="checkbox"/> | Official Gene Set v3.2                 |
| <input type="checkbox"/>            | Augustus Set 12                        |
| <input type="checkbox"/>            | Augustus Set 9                         |
| <input type="checkbox"/>            | Fgenesh                                |
| <input type="checkbox"/>            | Fgenesh++ without RNASeq training data |
| <input type="checkbox"/>            | Fgenesh++ with RNASeq training data    |
| <input type="checkbox"/>            | GenID                                  |
| <input type="checkbox"/>            | NCBI Gnomon                            |



# Navigating to ‘Reference Sequence’ (i.e. assembly fragments: scaffolds, chromosomes, etc.)

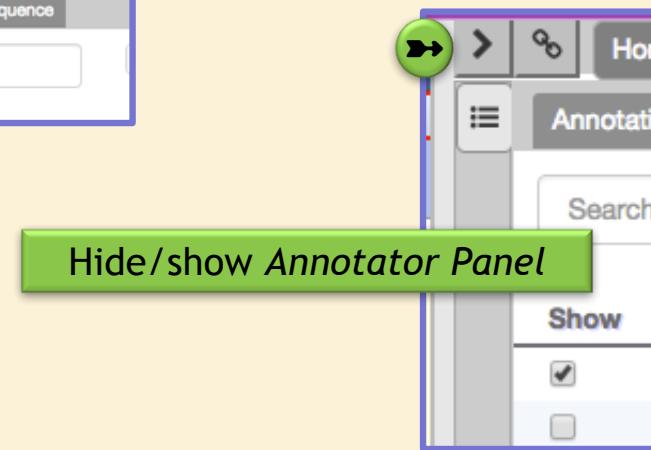
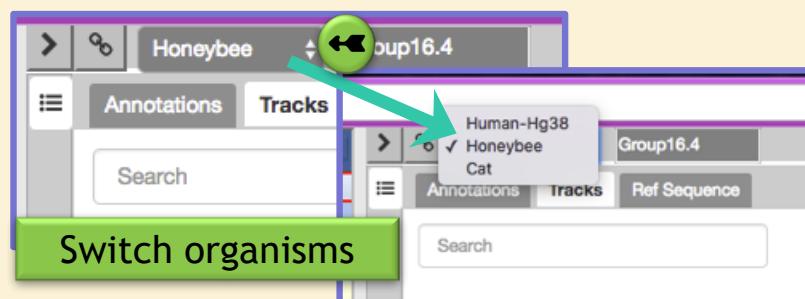
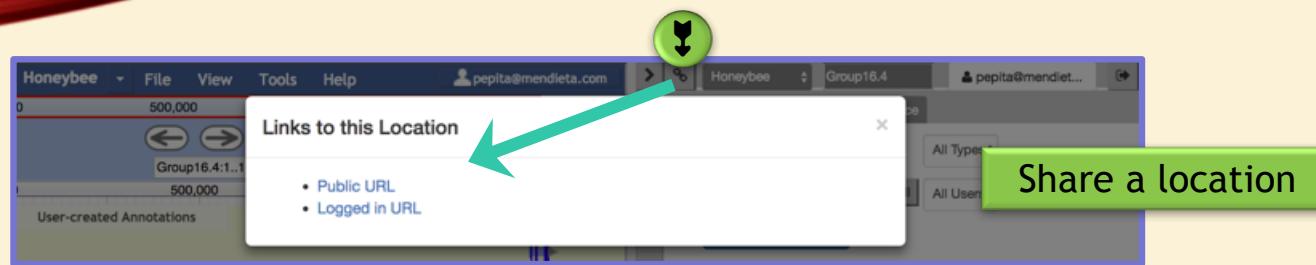
Ref Sequence

The screenshot shows the Apollo genome browser interface. At the top, there is a header with a back arrow, the project name "Honeybee", the assembly version "Group16.4", and a user account. Below the header, there are tabs for "Annotations", "Tracks", and "Ref Sequence", with "Ref Sequence" being the active tab and highlighted with a red box. On the left, there is a search bar with a back arrow icon, a "Length" filter with "Minimum" and "Maximum" buttons, and an "Export" section with "GFF3" and "FASTA" options. In the main panel, a table lists 1-50 of 5,644 entries, with columns for "Name", "Length", and "Annotations". The first few entries are: "Group11.18" (4,736,299), "Group9.10" (4,726,012), "Group15.19" (3,997,324), "Group2.19" (3,883,383), and "Group12.13" (2,883,042). To the right of the table are two zoomed-in views of the genome tracks. The top view shows a track for "Group1.33" with a gene "Group1.12" highlighted. The bottom view shows a track for "Group1.33" with a gene "Group1.11" highlighted. Both views include a "Tools" menu and a "Help" button.

| Name       | Length    | Annotations |
|------------|-----------|-------------|
| Group11.18 | 4,736,299 |             |
| Group9.10  | 4,726,012 |             |
| Group15.19 | 3,997,324 |             |
| Group2.19  | 3,883,383 |             |
| Group12.13 | 2,883,042 |             |



# Additional functionality



# Slides

<http://bit.ly/apollo-emblabr-intro>

<http://bit.ly/apollo-emblabr-edit>

# Follow along



# Access Apollo

| Your number | Email                      | Password      | Server | Organism | Begin at |
|-------------|----------------------------|---------------|--------|----------|----------|
| 1           | user.one@example.com       | userone       | 1      | Honey0   | 1        |
| 2           | user.two@example.com       | usertwo       | 2      | Honey0   | 1        |
| 3           | user.three@example.com     | userthree     | 3      | Honey0   | 1        |
| 4           | user.four@example.com      | userfour      | 4      | Honey0   | 1        |
| 5           | user.five@example.com      | userfive      | 5      | Honey0   | 1        |
| 6           | user.six@example.com       | usersix       | 1      | Honey1   | 7        |
| 7           | user.seven@example.com     | userseven     | 2      | Honey1   | 7        |
| 8           | user.eight@example.com     | usegereight   | 3      | Honey1   | 7        |
| 9           | user.nine@example.com      | usernine      | 4      | Honey1   | 7        |
| 10          | user.ten@example.com       | usereten      | 5      | Honey1   | 7        |
| 11          | user.eleven@example.com    | useleven      | 1      | Honey2   | 1        |
| 12          | user.twelve@example.com    | usertwelve    | 2      | Honey2   | 1        |
| 13          | user.thirteen@example.com  | userthirteen  | 3      | Honey2   | 1        |
| 14          | user.fourteen@example.com  | userfourteen  | 4      | Honey2   | 1        |
| 15          | user.fifteen@example.com   | userfifteen   | 5      | Honey2   | 1        |
| 16          | user.sixteen@example.com   | usersixteen   | 1      | Honey3   | 7        |
| 17          | user.seventeen@example.com | userseventeen | 2      | Honey3   | 7        |
| 18          | user.eighteen@example.com  | usereighteen  | 3      | Honey3   | 7        |

| Your number | Email                | Password | Server | Organism | Begin at |
|-------------|----------------------|----------|--------|----------|----------|
| 1           | user.one@example.com | userone  | 1      | Honey0   | 1        |
| 2           | user.two@example.com | usertwo  | 2      | Honey0   | 1        |

|    |                              |                 |   |        |   |
|----|------------------------------|-----------------|---|--------|---|
| 24 | user.twentyfour@example.com  | usertwentyfour  | 4 | Honey4 | 1 |
| 25 | user.twentyfive@example.com  | usertwentyfive  | 5 | Honey4 | 1 |
| 26 | user.twentysix@example.com   | usertwentysix   | 1 | Honey5 | 7 |
| 27 | user.twentyseven@example.com | usertwentyseven | 2 | Honey5 | 7 |
| 28 | user.twentyeight@example.com | usertwentyeight | 3 | Honey5 | 7 |
| 29 | user.twentynine@example.com  | usertwentynine  | 4 | Honey5 | 7 |
| 30 | user.twentynine@example.com  | usertwentynine  | 5 | Honey5 | 7 |



# Files

<http://bit.ly/apollo-emblabr-exercises1>

<http://bit.ly/apollo-emblabr-exercises2>

# Thank You.

Berkeley Bioinformatics Open-Source Projects,  
Environmental Genomics & Systems Biology,  
Lawrence Berkeley National Laboratory

Suzanna Lewis & Chris Mungall

Seth Carbon (GO - Noctua / AmiGO)

Eric Douglas (GO / Monarch Initiative)

Nathan Dunn (Apollo)



## Collaborators

- Ian Holmes, Eric Yao, UC Berkeley (JBrowse)
- Chris Elsik, Deepak Unni, U of Missouri (Apollo)
- Paul Thomas, USC (Noctua)
- Monica Poelchau, USDA/NAL (Apollo)
- Gene Ontology Consortium (GOC)
- i5k Community

## Funding

- Work for GOC is supported by NIH grant 5U41HG002273-14 from NHGRI.
- Apollo is supported by NIH grants 5R01GM080203 from NIGMS, and 5R01HG004483 from NHGRI.
- BBOP is also supported by the Director, Office of Science, Office of Basic Energy Sciences, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231

[berkeleybop.org](http://berkeleybop.org)



Berkeley  
UNIVERSITY OF CALIFORNIA



# BBOP Projects



- **GeneOntology.org (GO)**
  - Assigning function to genes in all organisms (including Noctua)
- **GenomeArchitect.org (Apollo)**
  - Collaborative curation of genomes and gene models
- **MonarchInitiative.org**
  - Using comparative phenomics to illuminate human diseases
- **INCA**
  - Intelligent Concept Assistant for application of metadata
- **Planteome.org**
  - (Prime: OSU) Common reference ontologies & annotations
- **AllianceGenome.org (AGR)**
  - Unified Model Organism Databases
- **NCATS Translator**
  - Automating the translation of mechanistic biological knowledge to clinical applications



[berkeleybop.org](http://berkeleybop.org)