



Apollo:

Collaborative Genome Annotation Editing

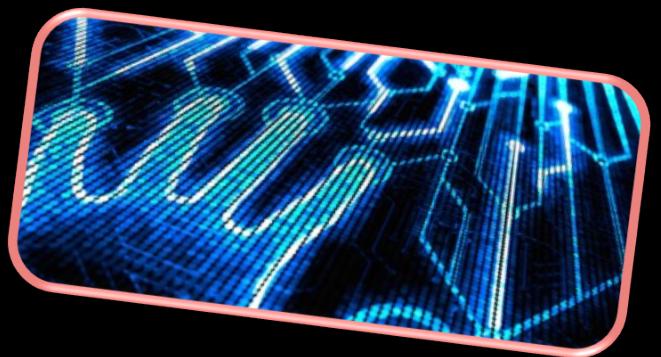
Monica Munoz-Torres, PhD | @monimunozto

Berkeley Bioinformatics Open-Source Projects
Environmental Genomics and Systems Biology Division
Lawrence Berkeley National Laboratory

A workshop for AGS X. University of Notre Dame, South Bend, IN. 08 June, 2017

Today...

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



After this workshop, you will:

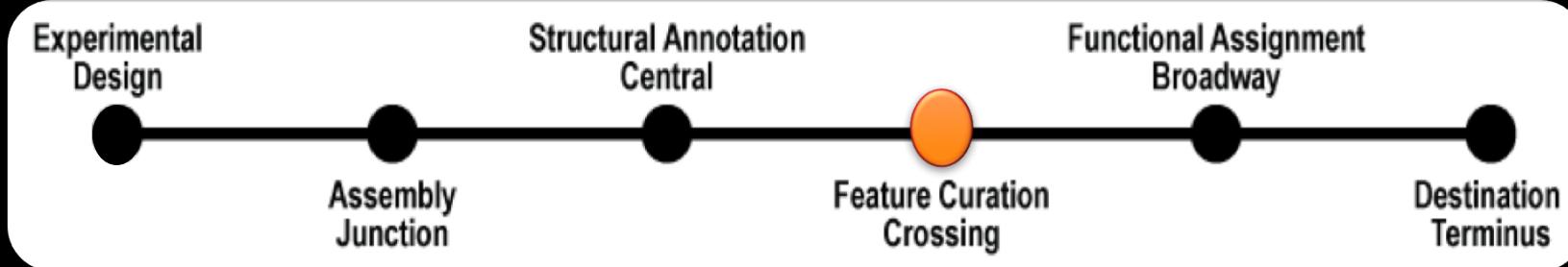


- Better understand curation in the context of genome annotation:
assembled genome → automated annotation → manual annotation
- Become familiar with Apollo's environment and functionality.
- Learn to identify homologs of known genes of interest in your newly sequenced genome.
- Learn how to corroborate 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.







Knowledge

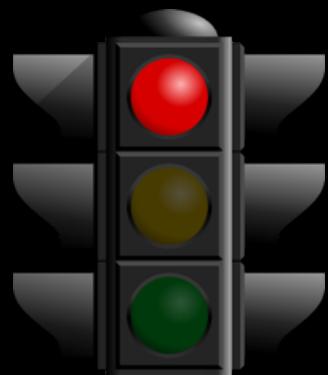
Extracting knowledge from data

Red: FF0000

Data



Extracting knowledge from data



Information

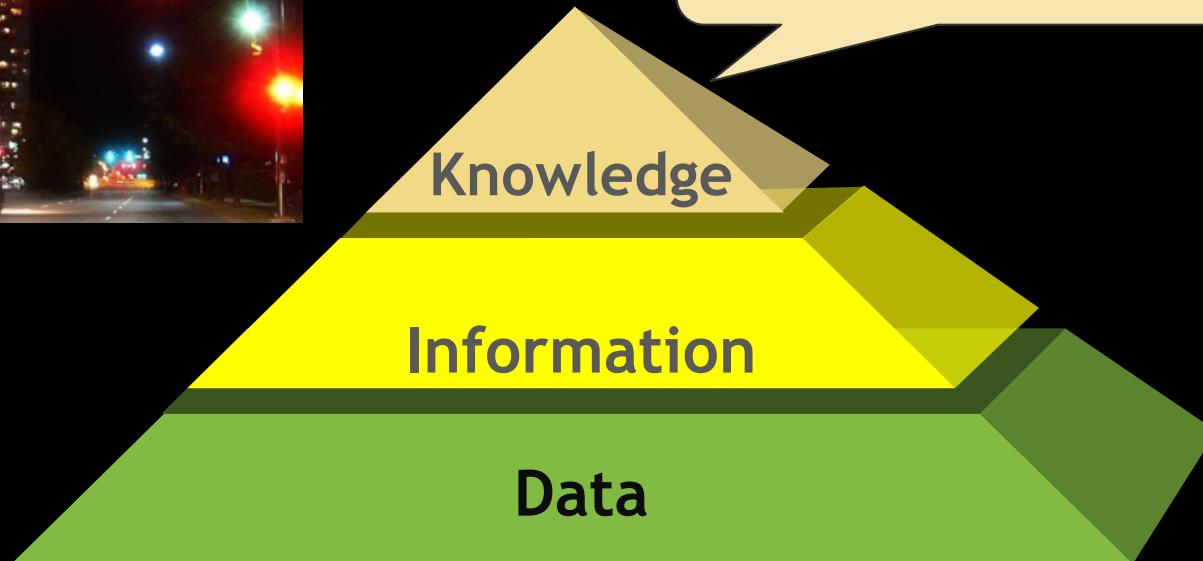
Data

South-facing traffic light at
St Mary's Rd. has turned red

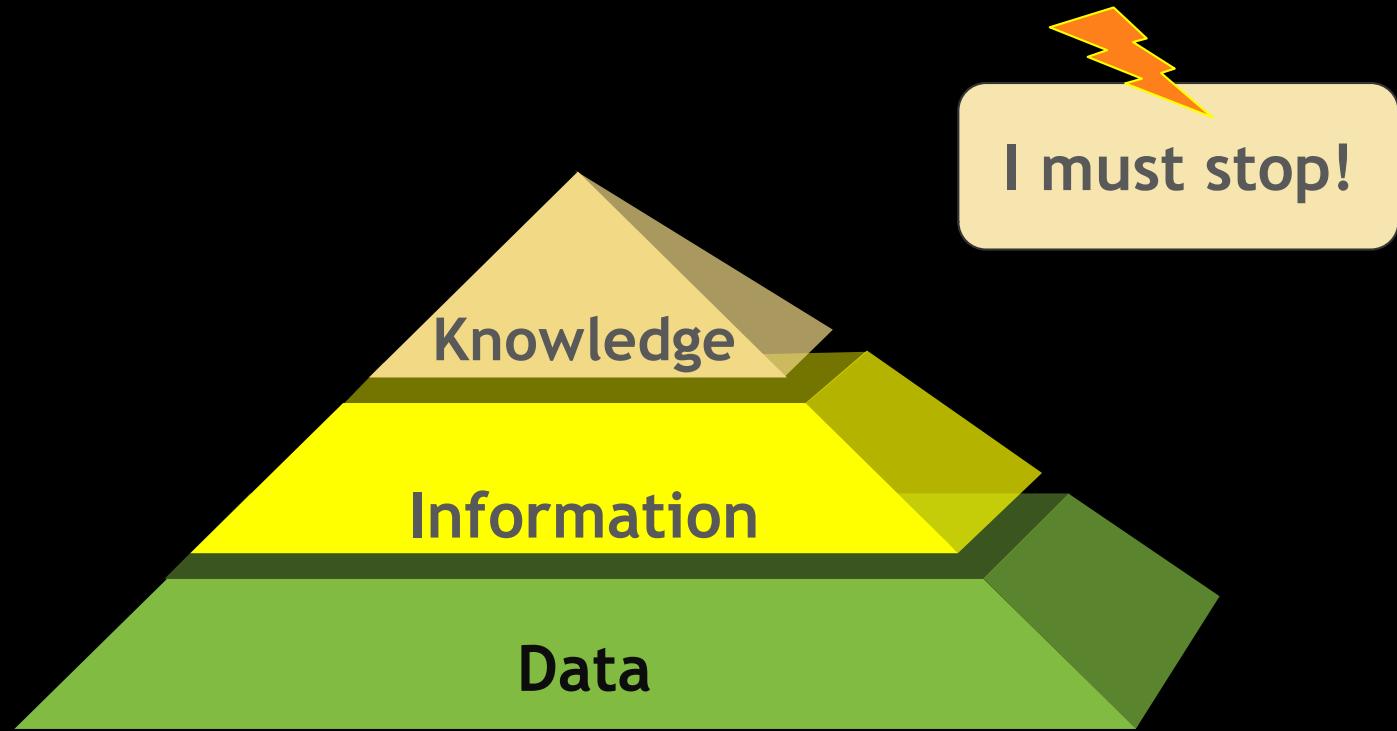
Extracting knowledge from data



The light I am driving towards has just turned red



Extracting knowledge from data

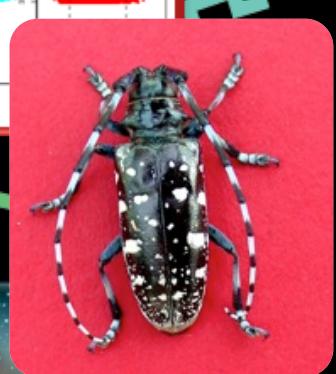
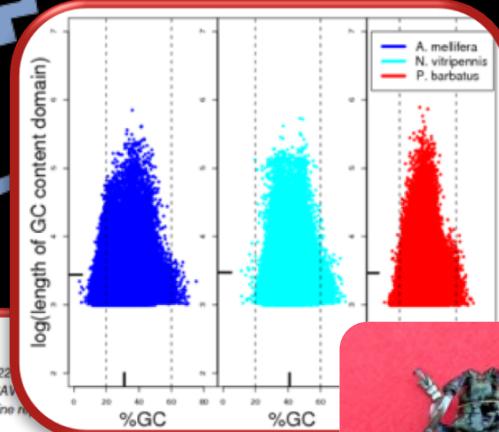
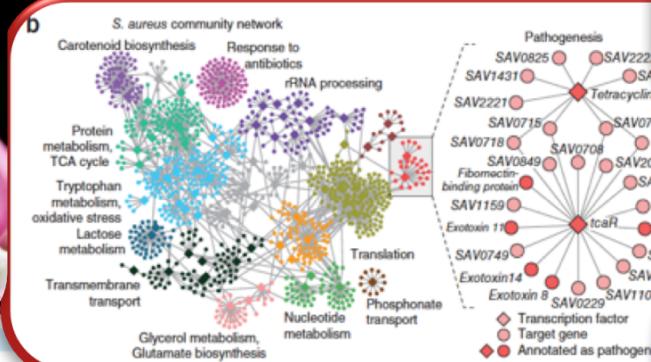
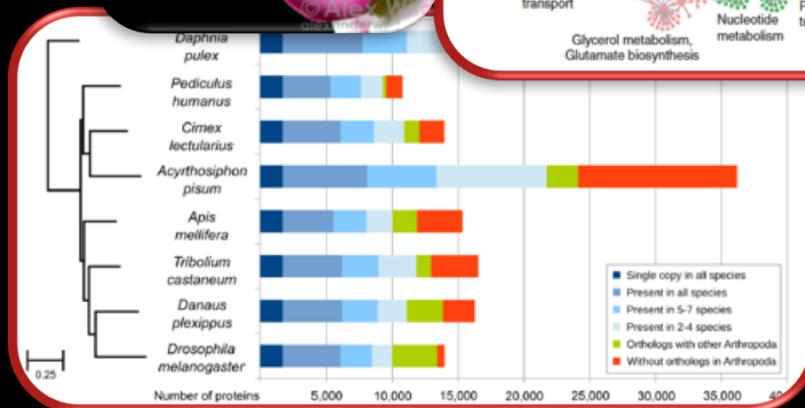




Genome Curation

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.



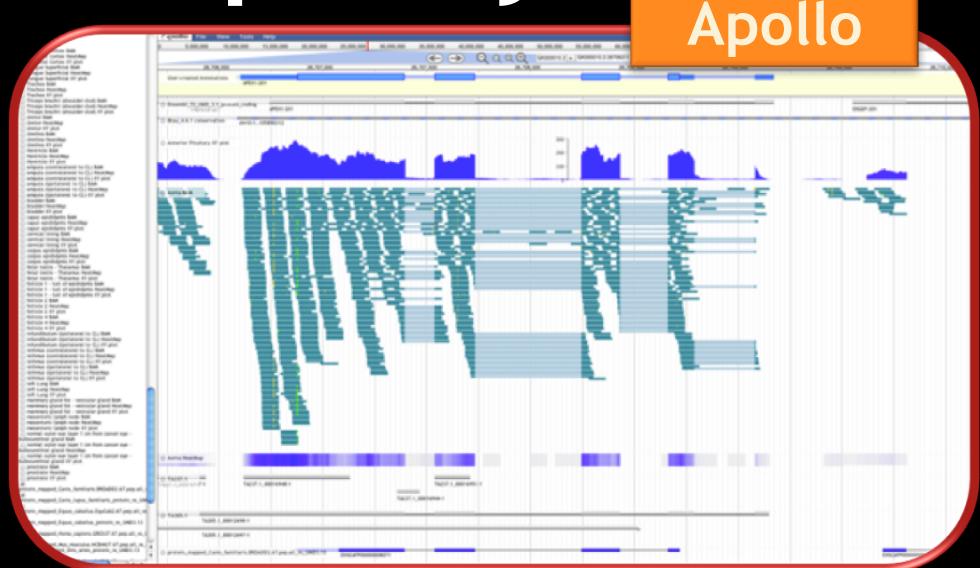
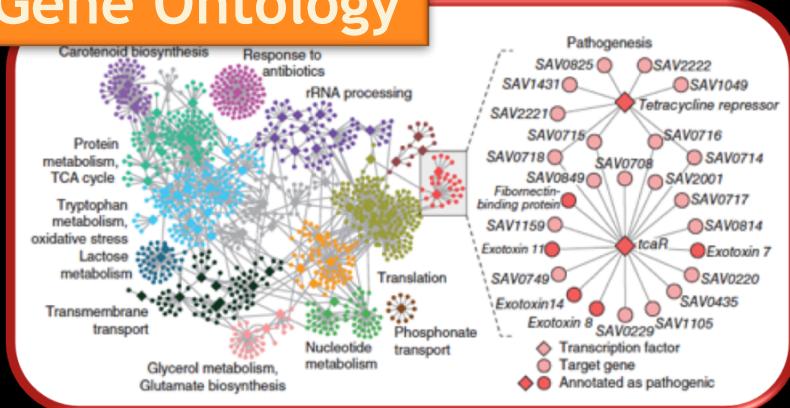
Curation improves quality



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".

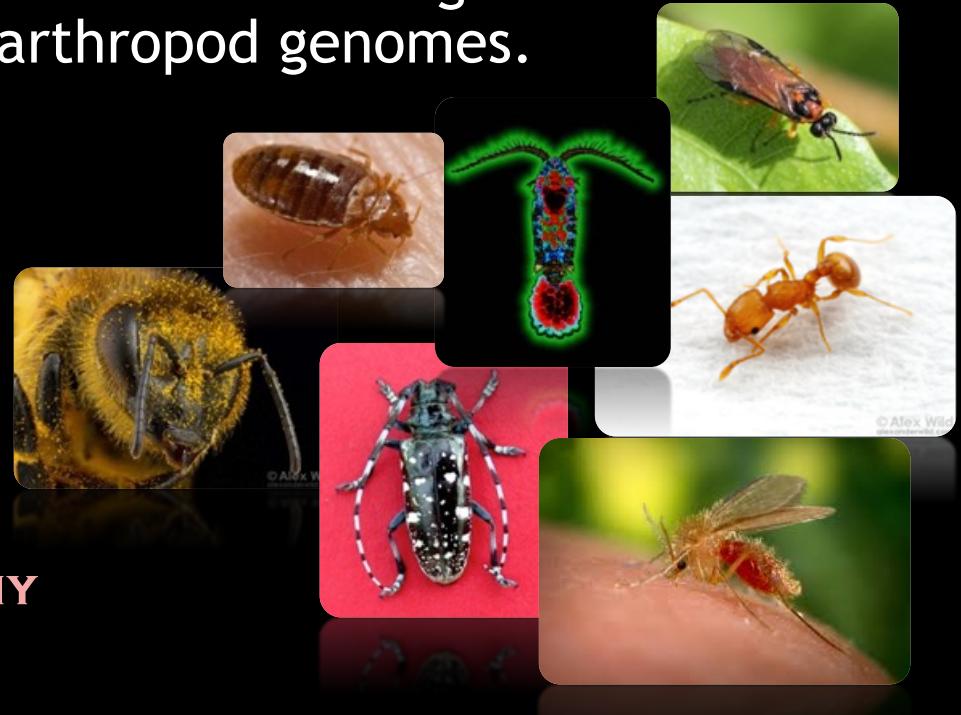


i5k - five thousand arthropod genomes

<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**



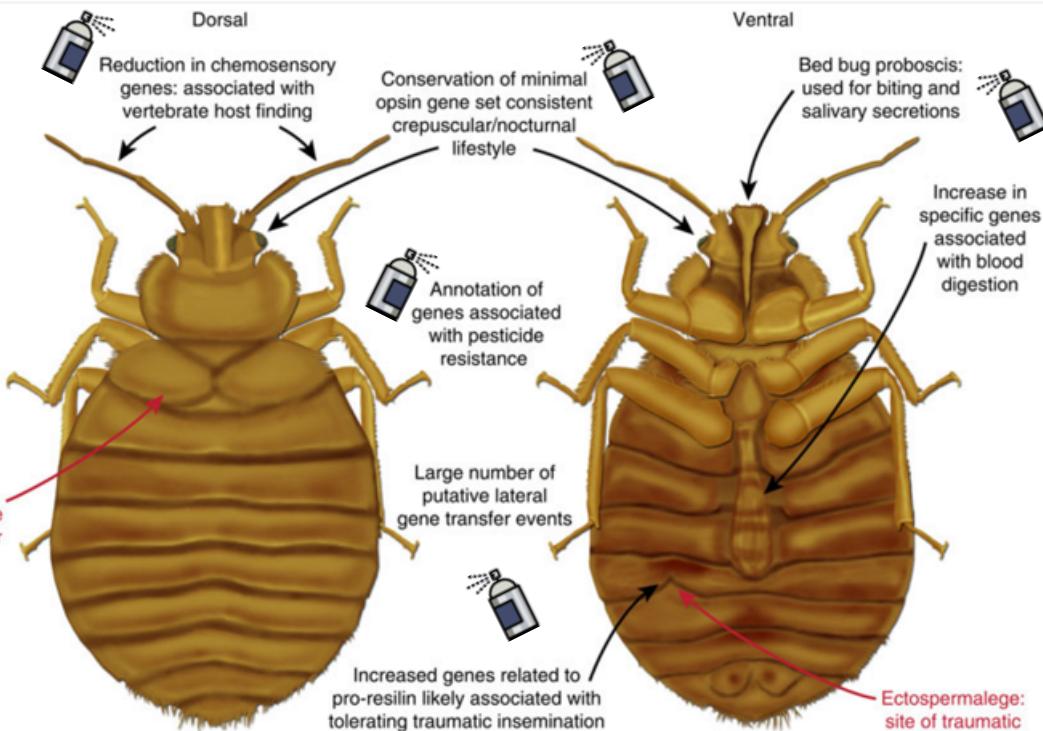
The bed bugs, they're back!

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:

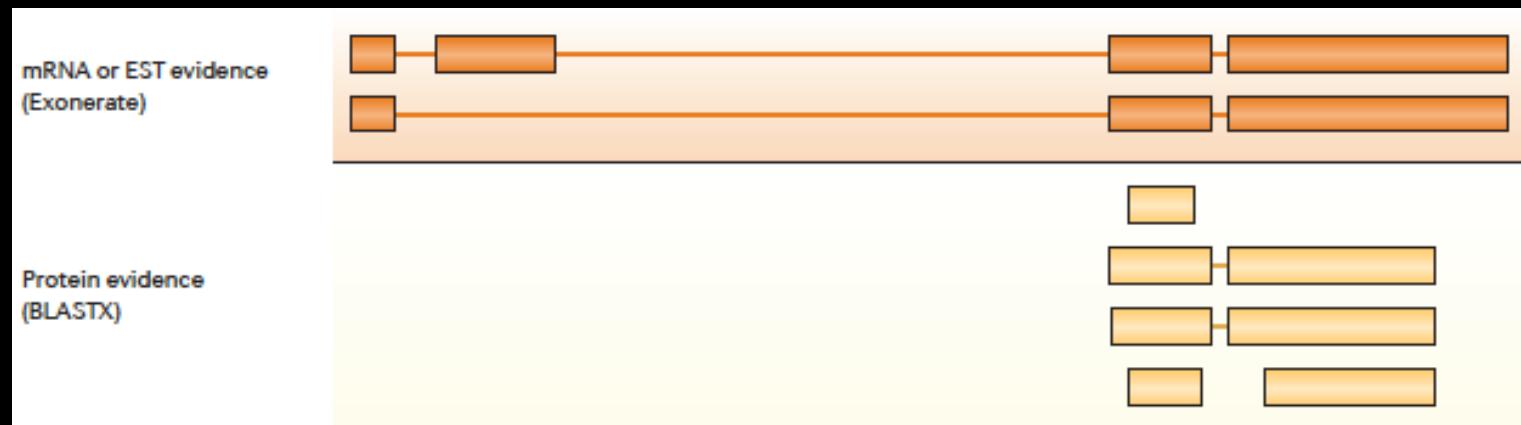
- Primarily focuses on protein-coding genes.
- Also identifies RNAs (tRNA, rRNA, long and small non-coding RNAs (ncRNA)), regulatory motifs, repetitive elements, etc.
- Happens in 2 steps:
 - Computation phase
 - Annotation phase



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.

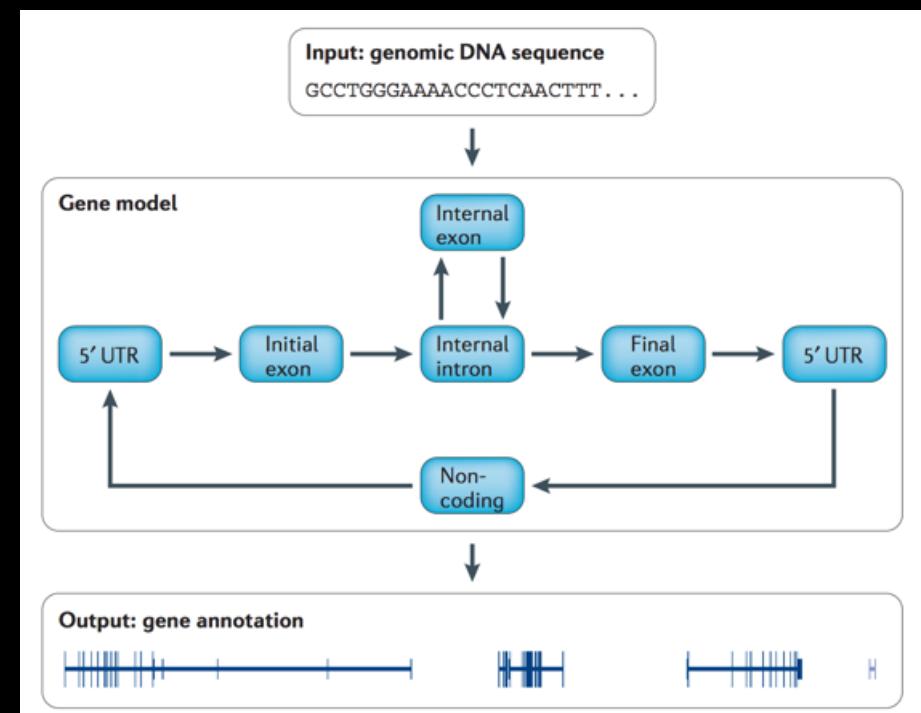
2b) Using experimental evidence: identifying domains and motifs
e.g. SGP2, JAMg, fgenesh++, etc.



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:

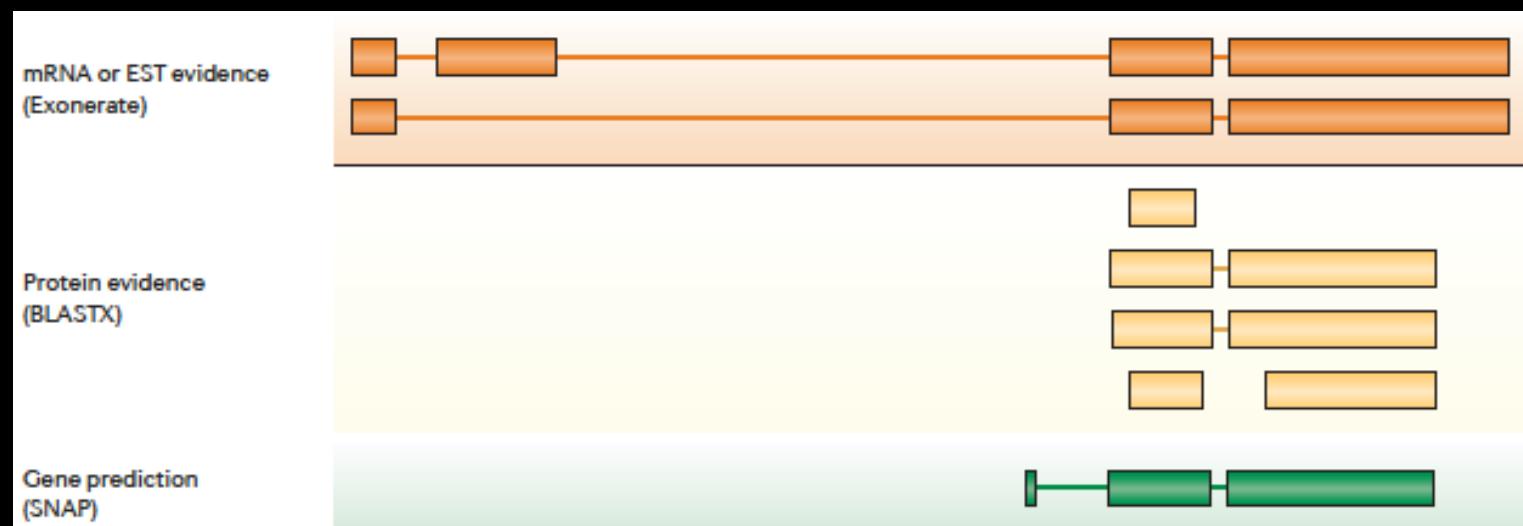
Finds genes using either similarity searches against public databases or other experimental data sets e.g. RNAseq.

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



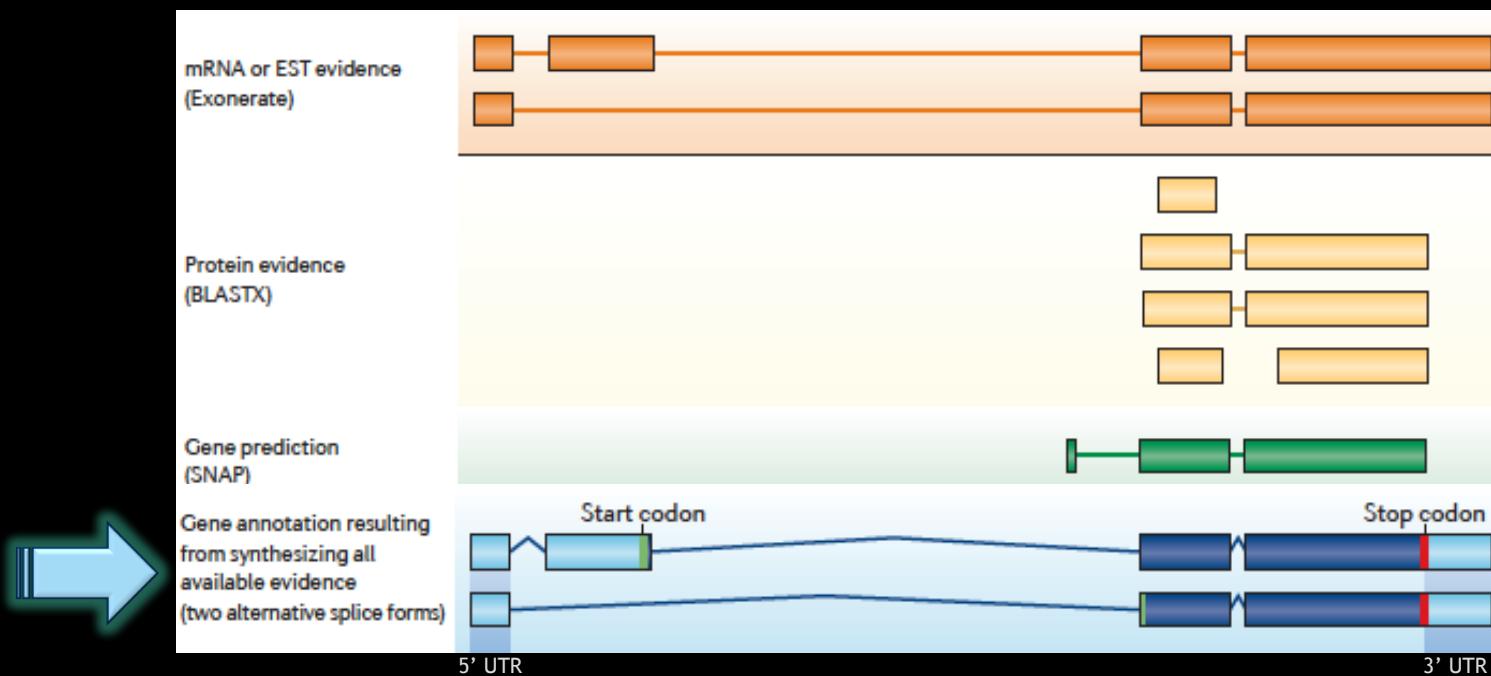
Computation Phase: result

- 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

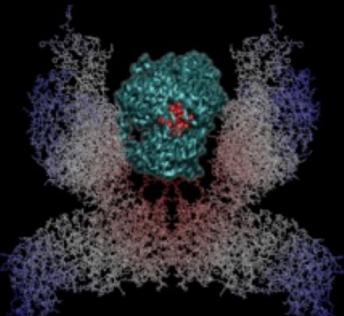


Terms (classes) arranged in a graph: molecular functions, biological processes, cellular locations, and the relationships connecting them all, in a species-independent manner.

1. Molecular Function

An elemental activity or task or job

- protein kinase activity
- insulin receptor activity

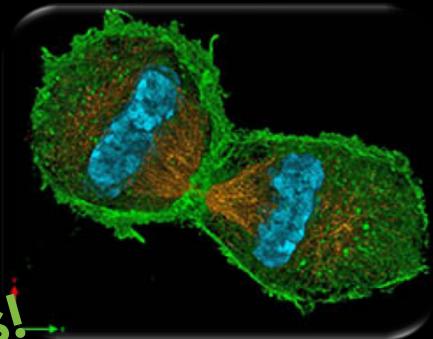


Insulin Receptor
Petrus et al, 2009, *ChemMedChem*

2. Biological Process

A commonly recognized series of events

- cell division



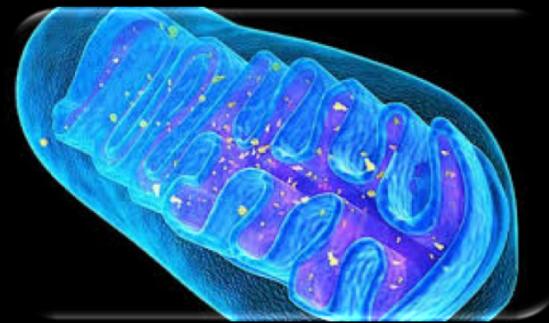
End of Telophase.
Lothar Schermerle

~150 Contributors!

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.





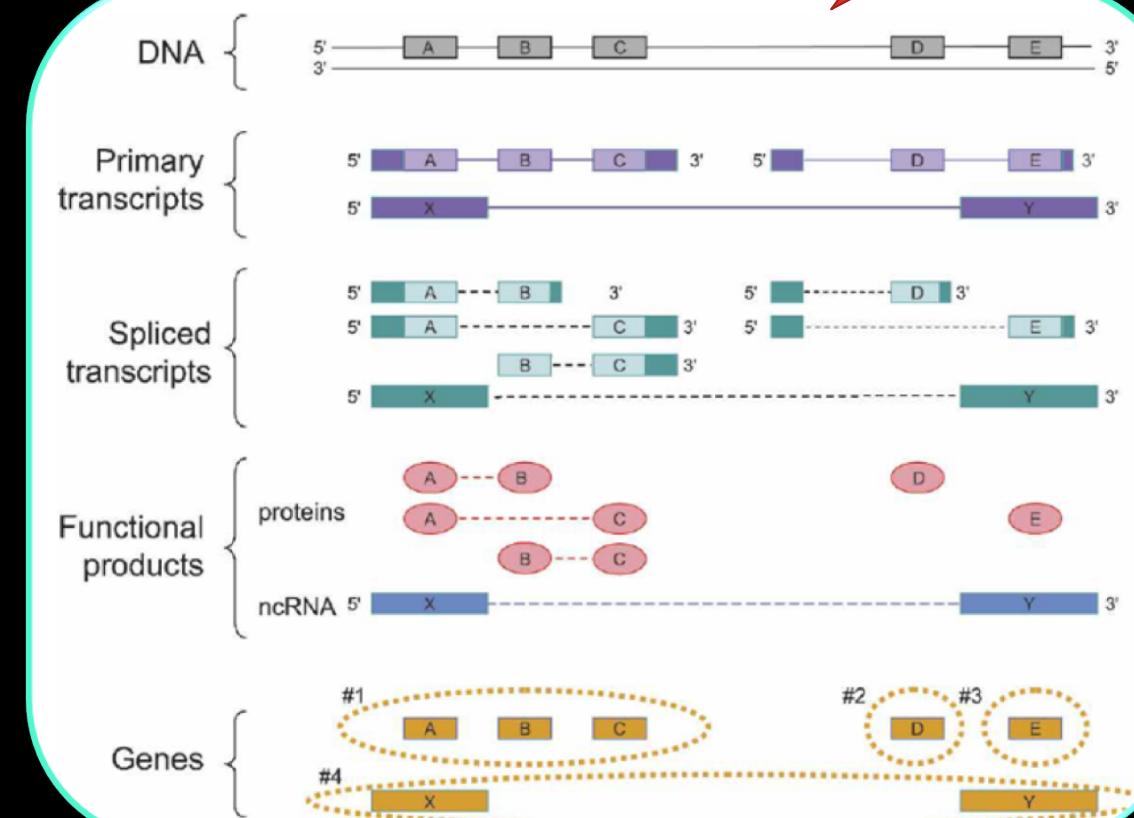
Bioreresher

A brief refresher

The gene: *a moving target*

Biorefresher

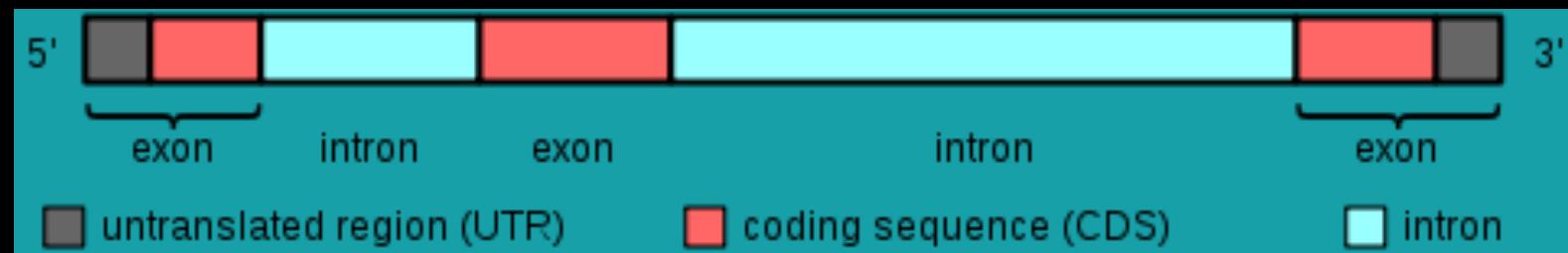
“The gene is a union of genomic sequences encoding a coherent set of potentially overlapping functional products.”



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.

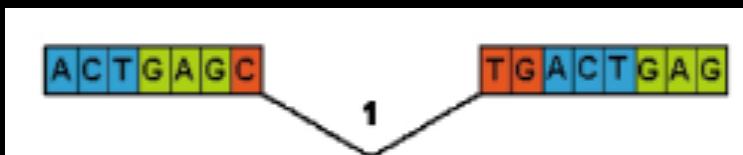
Exons and Introns

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.



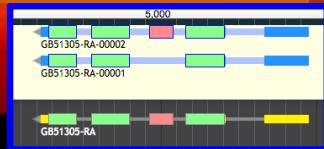
Functionality overview

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.

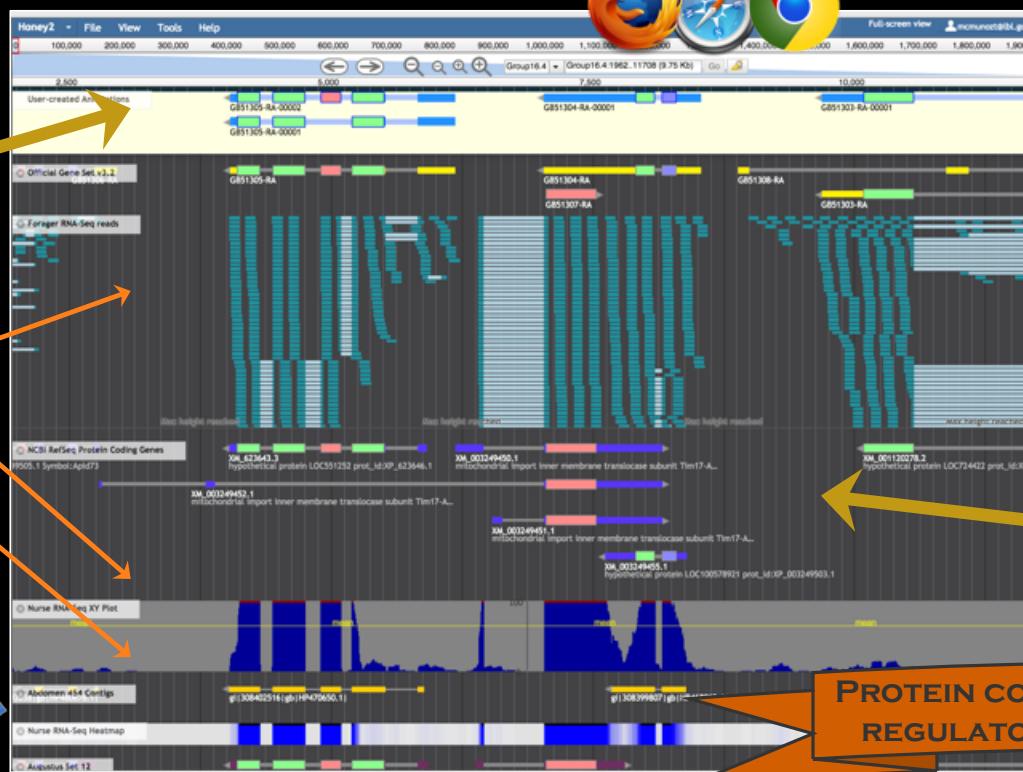
★ Navigate and zoom.

★ Search for a gene model
or a scaffold.

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



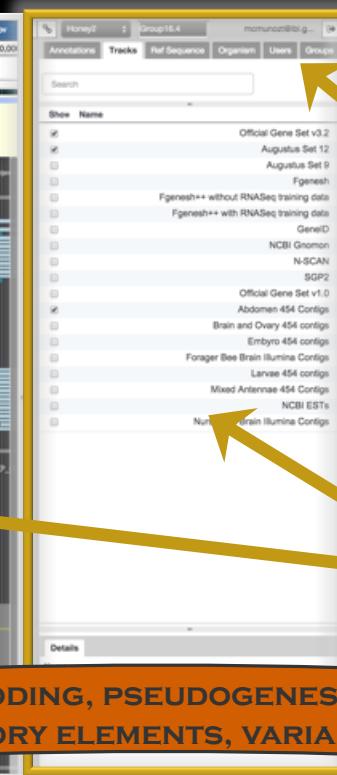
★ User-created annotations.



★ Stage and cell-type
specific transcription
data.

Admin

★ Annotator panel.

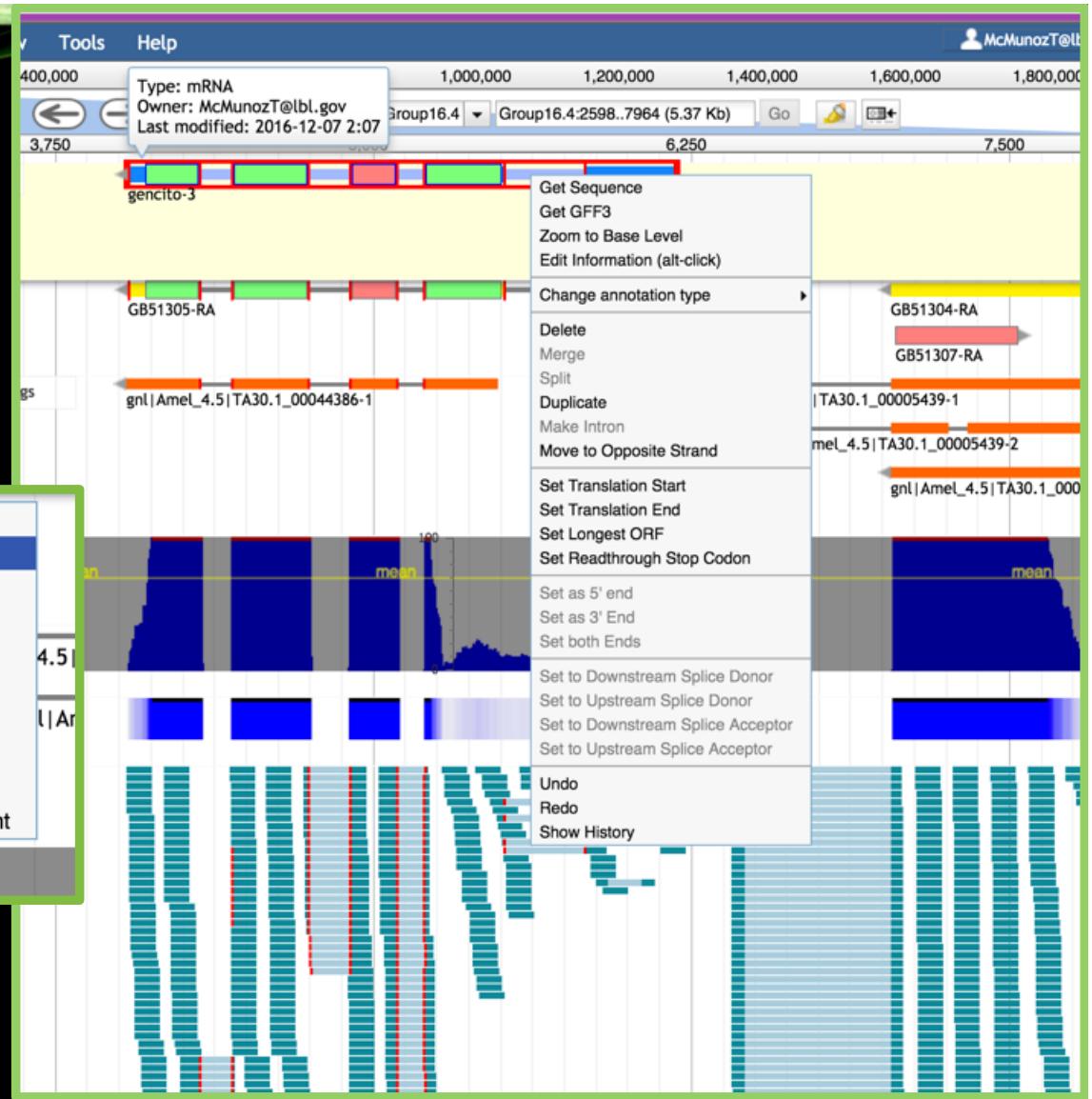
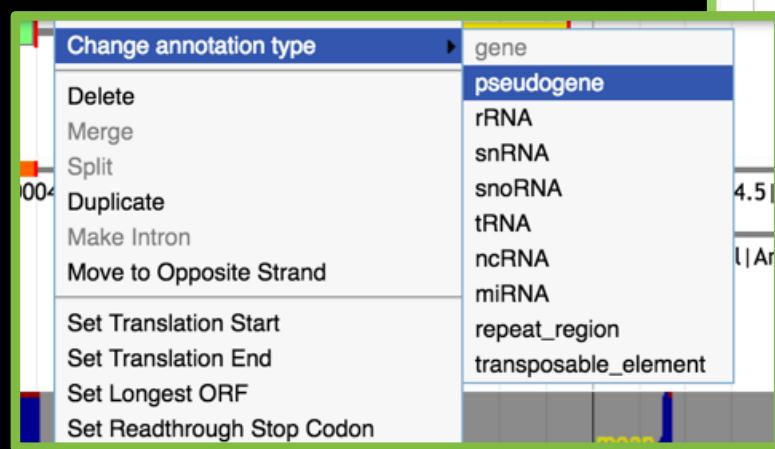


Evidence Tracks. ★

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



Right-click functionality



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

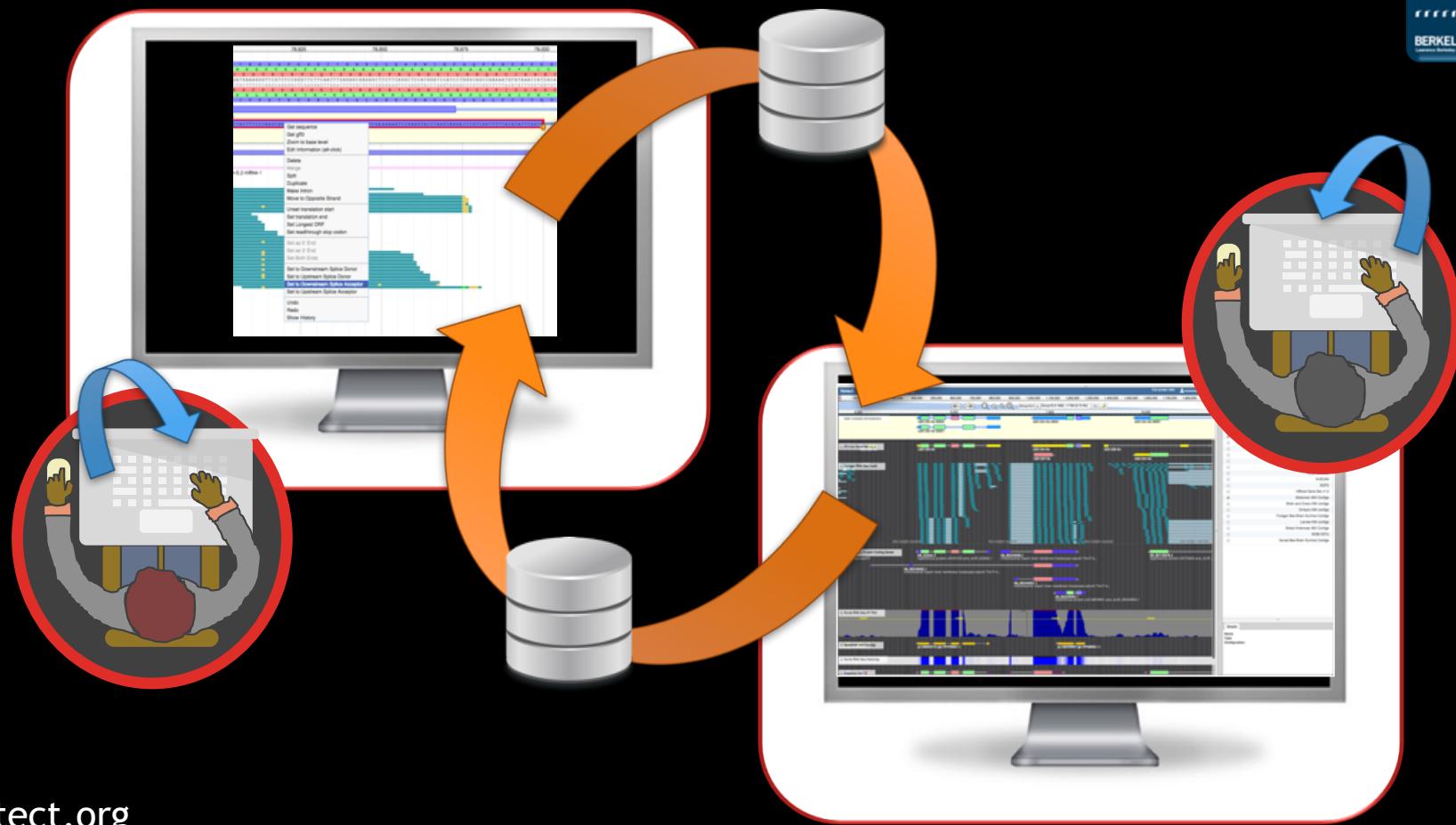
GroupUn5044 ▲ Length 540 560 564



GenomeArchitect.org

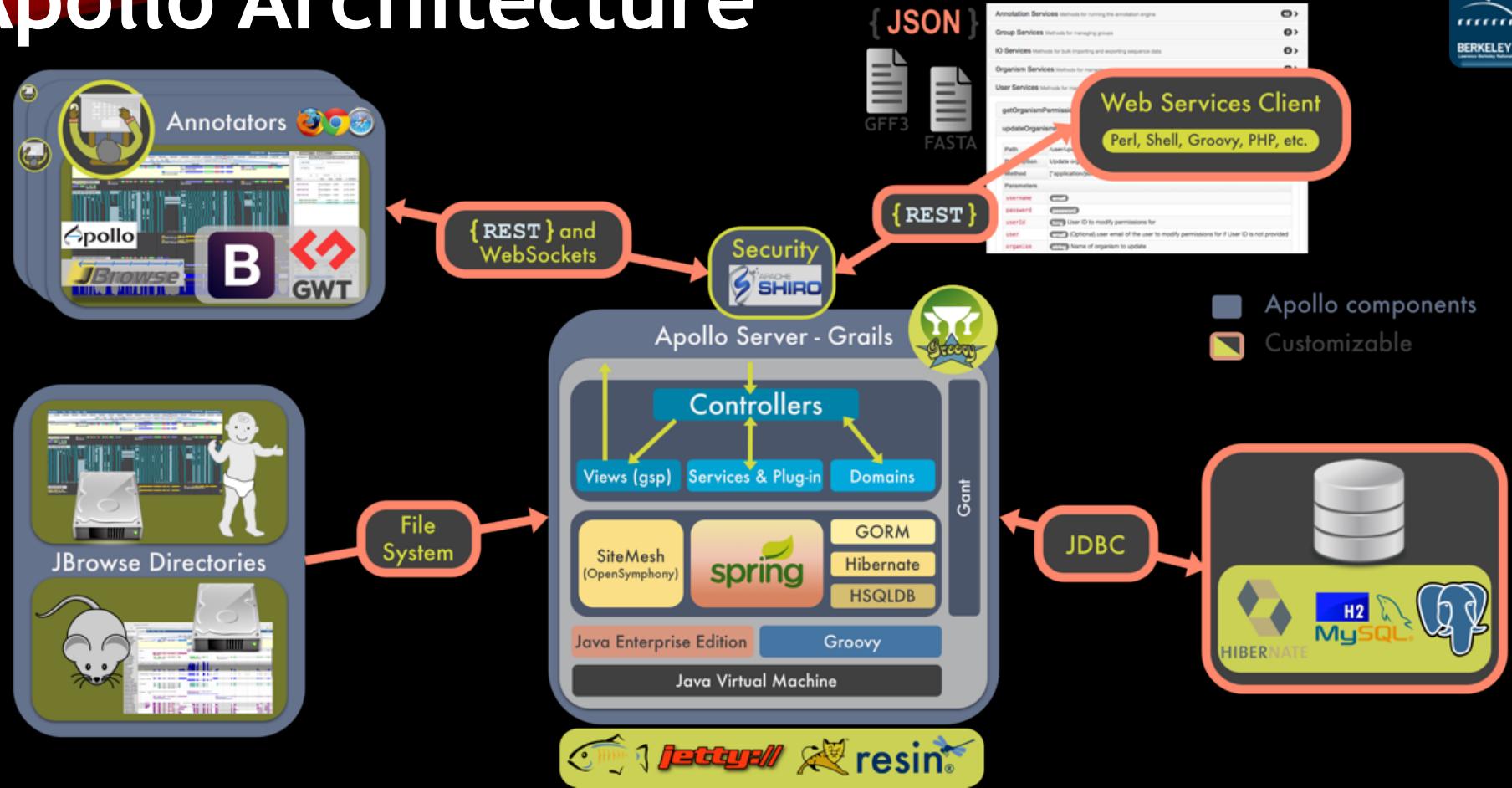
Apollo

Collaboration in real time

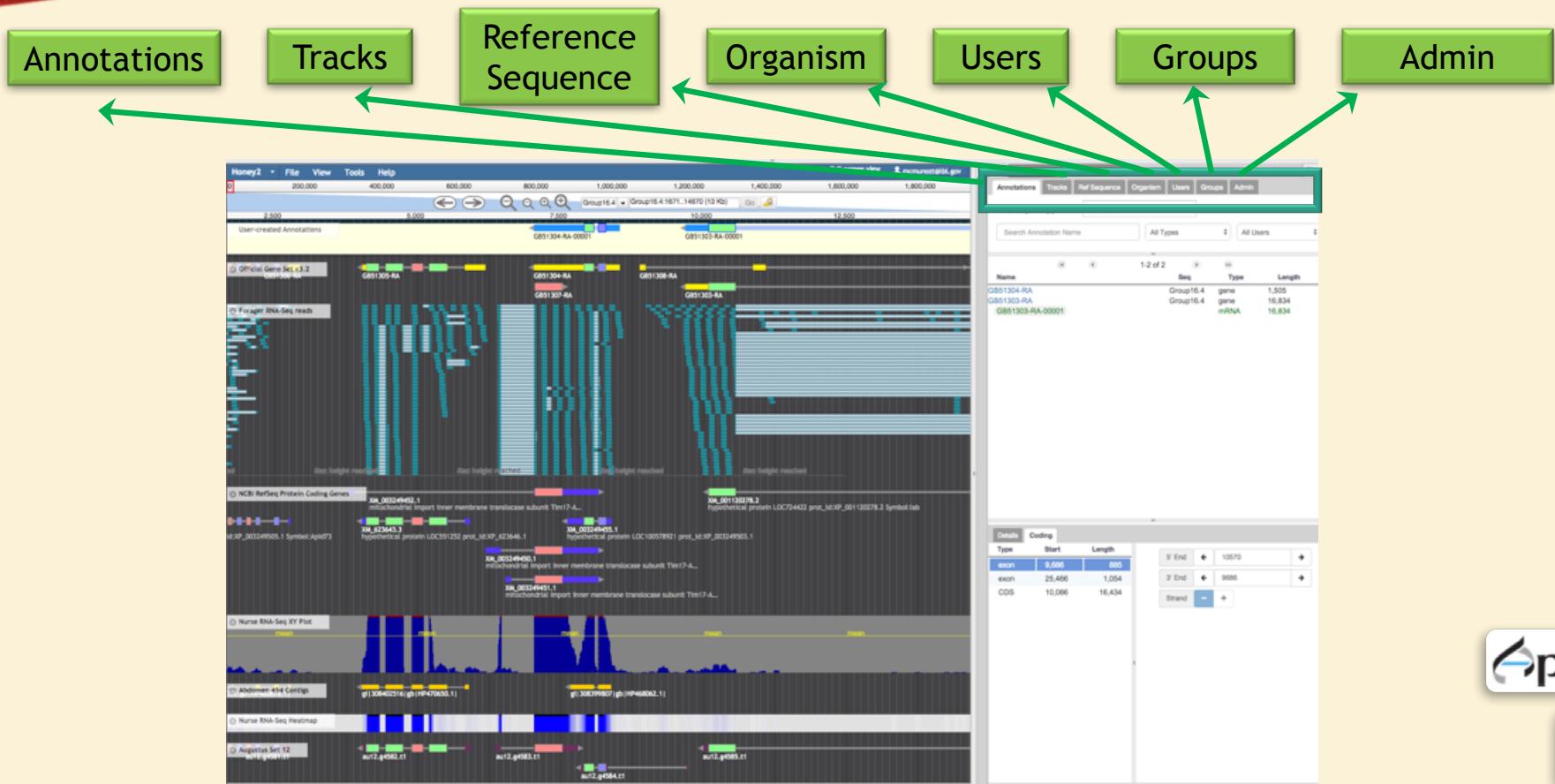


GenomeArchitect.org

Apollo Architecture



Removable Annotator Panel



Annotation details & exon boundaries

Annotations

The screenshot shows the Apollo annotation interface. A blue arrow points from the 'Annotations' tab in the top navigation bar to a detailed view of the 'spel1-RA' annotation. Another blue arrow points from the 'mRNA' entry in the table below to the exon boundary details in the bottom right panel.

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

mRNA

Details Coding

1

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. The top navigation bar includes 'Annotations' (highlighted in green), 'Tracks', and 'Ref Sequence'. The main search area has fields for 'Annotation Name' (with placeholder 'All Types'), 'Reference Sequence' (with placeholder 'All'), and 'All Users'. A large blue button labeled 'Go to Annotation' with a circular arrow icon is prominent. Below this, a search result 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' over the first row, 'mRNA' over the second row, and 'spel1' over the first entry in the list. A green arrow points from the 'mRNA' label to the 'spel1-RA' row. A small inset window shows a zoomed-in view of the 'All Types' dropdown menu, which includes options like 'Gene', 'Pseudogene', 'Transposable Element', and 'Repeat Region', with 'Gene' selected.



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 highlighted with a red box. Below the tabs is a search bar and a 'JBrowse Track Selector' button. The main content area displays a list of available tracks under the heading 'Available Tracks'.

Available Tracks

- Abdomen_454 Contigs
- Acceptor-5' GSSv1.2
- Apis_454v8
- Apis_cerana reads
- Augustus Set 12
- Augustus Set 9
- Brain and Ovary 454 contigs
- Cfpo_00003.3
- Cfpo_00004.3
- Embryo_454 contigs
- Fgenesh
- Fgenesh++ with RNASeq training data
- Fgenesh++ without RNASeq training data
- Forager_Beacon_HeatMap
- Forager_RNA-Seq XY Plot
- Forager_RNA-Seq reads

The main genome track area shows a yellow 'User-created Annotations' track and an 'Official Gene Set v3.2' track with various genomic features like exons and genes.

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 "Group1.33:189668..234268 (44.6 K)" and its annotations. The bottom view shows a track for "Group1.33" with a gene "Group1.11" and its annotations. Both views have a green circular arrow icon in the top right corner.

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

The screenshot displays the Apollo genome annotation tool interface with several highlighted features:

- Share a location**: A green button with a download icon and a tooltip pointing to a "Links to this Location" dialog box containing options for "Public URL" and "Logged in URL".
- Switch organisms**: A green button with a switch icon and a tooltip pointing to the organism selection dropdown menu, which includes "Human-Hg38", "Honeybee", and "Cat".
- Leave a session**: A green button with a logout icon and a tooltip pointing to the user session information at the top of the screen.
- Hide/show Annotator Panel**: A green button with a double arrow icon and a tooltip pointing to the panel control buttons on the right side of the interface.

Follow along



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	userten	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
19	user.nineteen@example.com	usernineteen	4	Honey3	7
20	user.twenty@example.com	usertwenty	5	Honey3	7
21	user.twentyone@example.com	usertwentyone	1	Honey4	1
22	user.twentytwo@example.com	usertwentytwo	2	Honey4	1
23	user.twentythree@example.com	usertwentythree	3	Honey4	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

Access Apollo

Server	URL
1	http://tinyurl.com/apollo-agss1
2	http://tinyurl.com/apollo-agss2
3	http://tinyurl.com/apollo-agss3
4	http://tinyurl.com/apollo-agss4
5	http://tinyurl.com/apollo-agss5

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



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)

Monica Munoz-Torres (Apollo / GO)

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



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