I5K Workspace webinar - New and upcoming features and datasets at the i5k Workspace@NAL

Monica Poelchau

National Agricultural Library

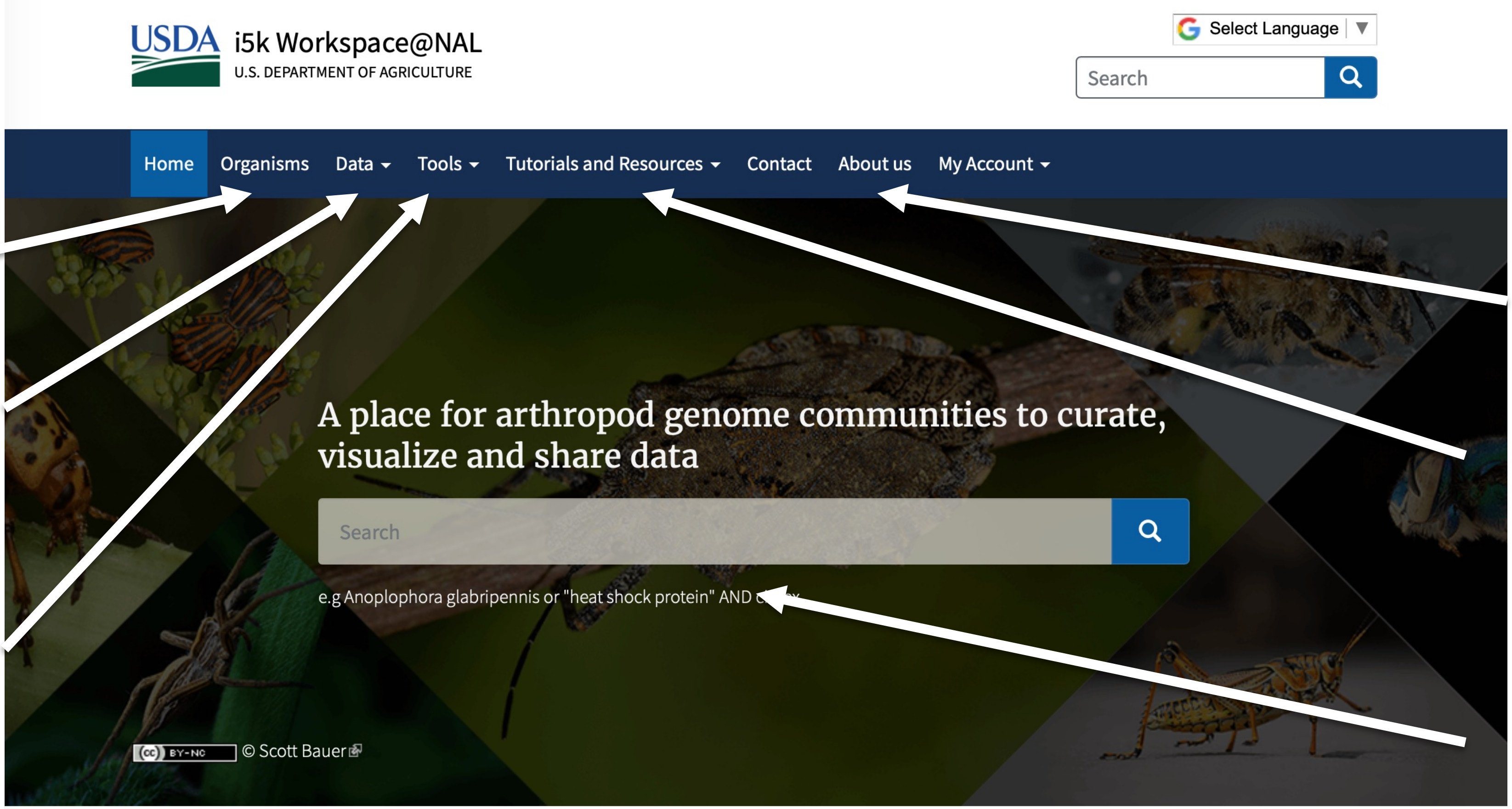
USDA-ARS

May 25th, 2021

Agenda

1. New RNA-Seq tracks in Apollo;
2. New datasets from the Ag100Pest project that are coming soon;
3. New functional annotations of proteins;
4. Upcoming Apollo software updates and new features;
5. Upcoming i5k Workspace software updates and new features.

The i5k Workspace@NAL



A place for arthropod genome communities to curate, visualize and share data

e.g Anoplophora glabripennis or "heat shock protein" AND

Genome data for 79 species

Browse genome assemblies and annotations

Use our tools to curate and search sequence data

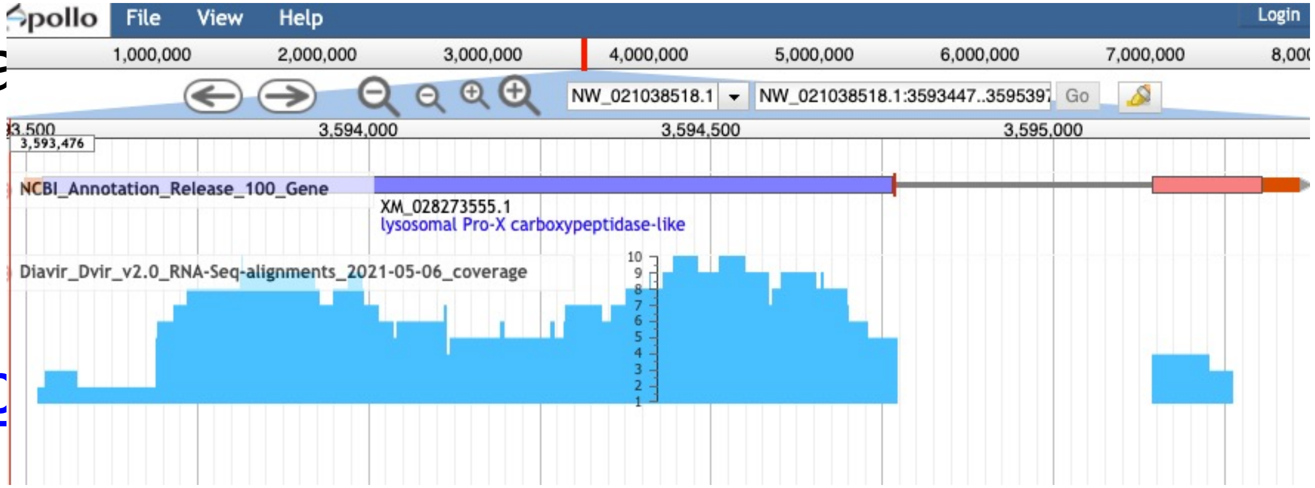
Learn more about the i5k Workspace@NAL

Learn more about genome curation

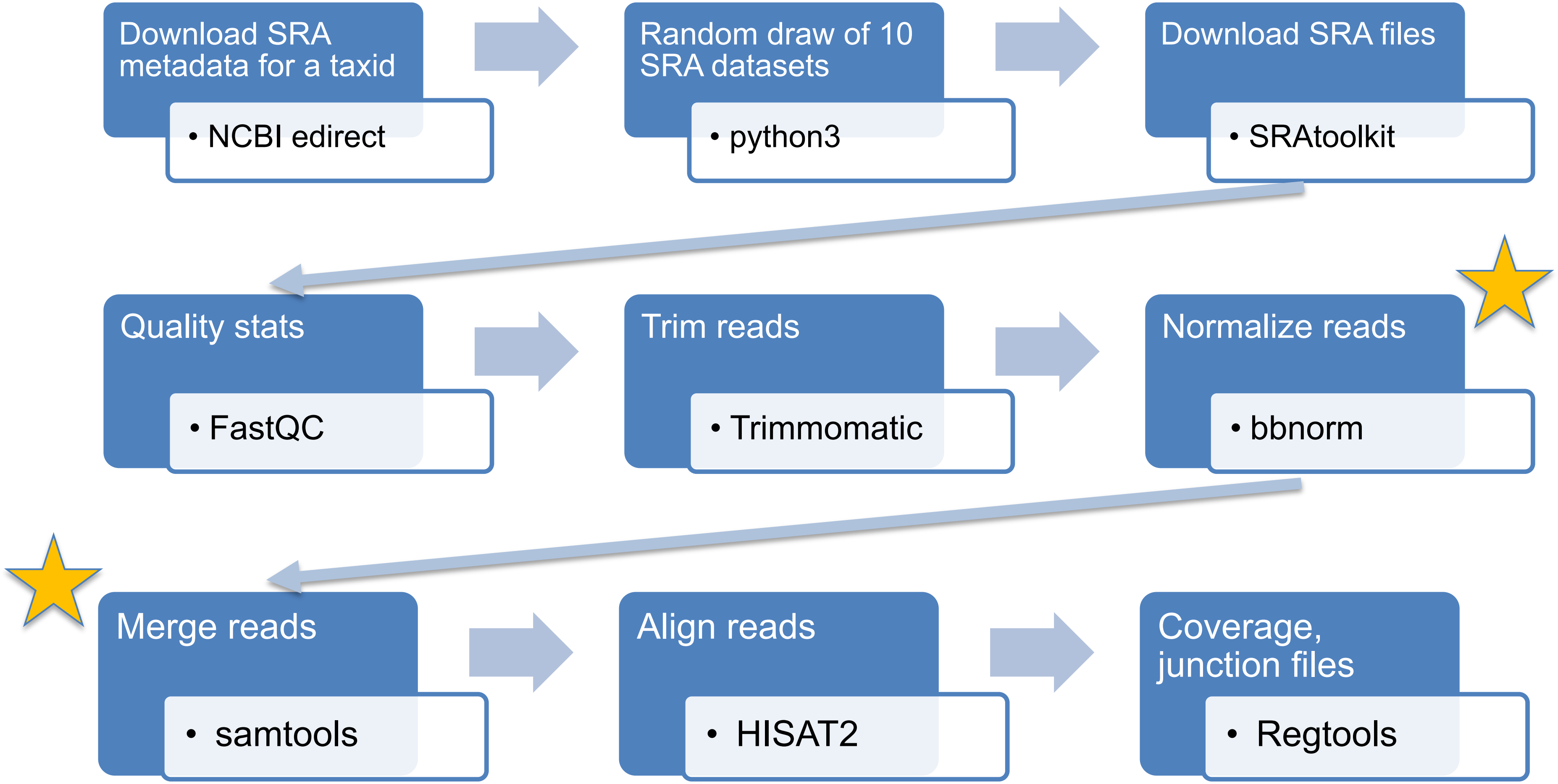
Search available data

RNA-Seq alignment pipeline

- RNA-Seq is critical evidence for manual curation
- Our python pipeline to generate a merged RNA-Seq track from multiple SRA accessions: https://github.com/NAL-i5K/NAL_RNA_seq_annotation_pipeline
- We are running this pipeline for all i5k Workspace organisms



RNA-Seq alignment pipeline



Download SRA metadata for a taxid
- NCBI edirect

Random draw of 10 SRA datasets
- python3

Download SRA files
- SRAtoolkit

Quality stats
- FastQC

Trim reads
- Trimmomatic

Normalize reads
- bbnorm

Merge reads
- samtools

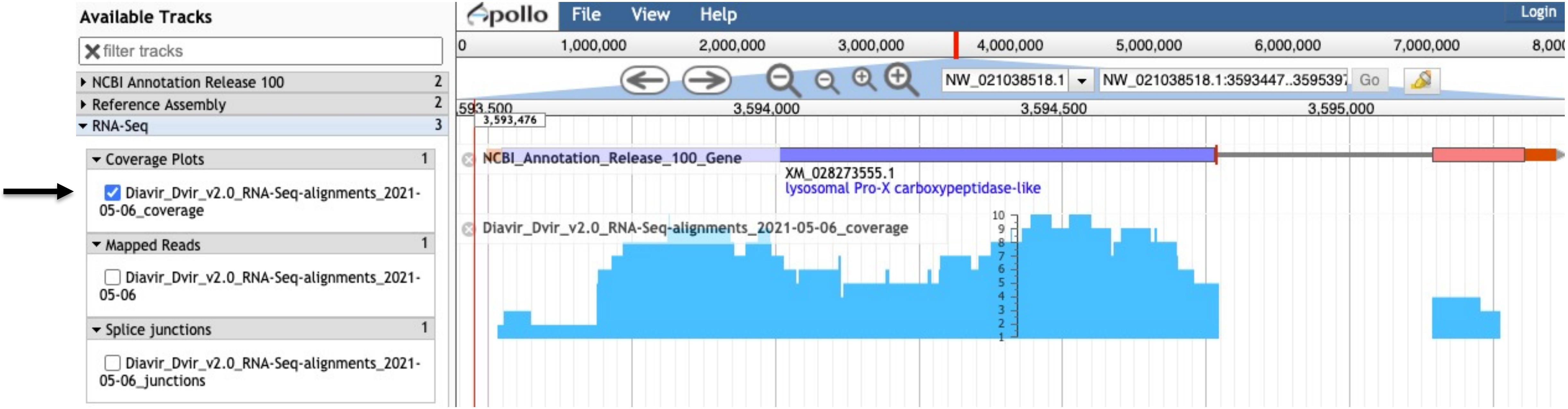
Align reads
- HISAT2

Coverage, junction files
- Regtools

https://github.com/NAL-i5K/NAL_RNA_seq_annotation_pipeline

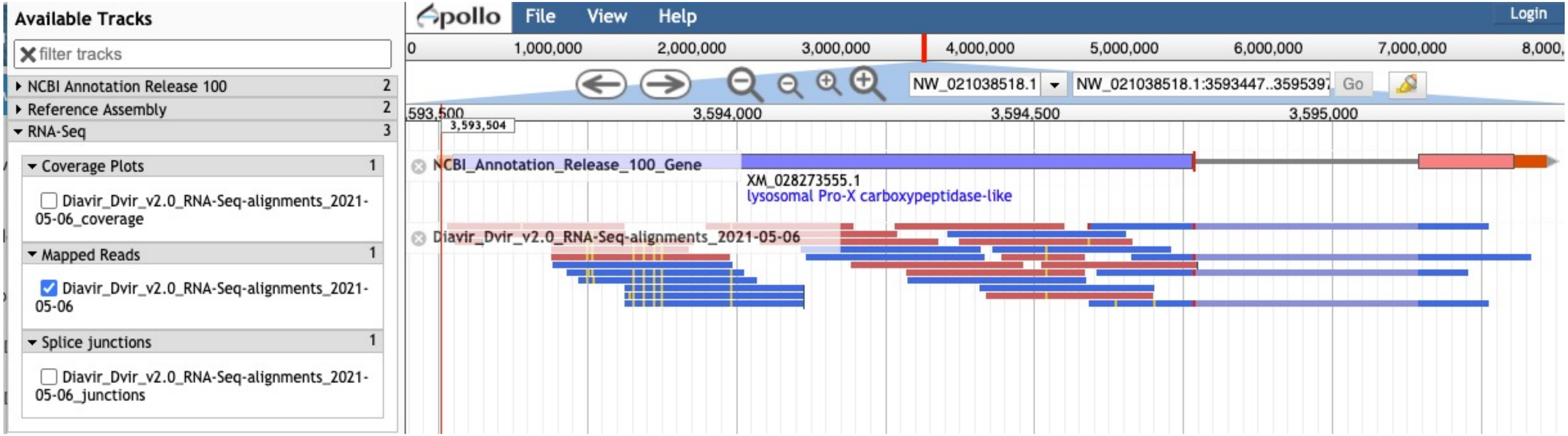
RNA-Seq pipeline outputs

Coverage plots: Histogram of the number of mappings at each nucleotide; hover over the blue area to see the value


Available Tracks
filter tracks
NCBI Annotation Release 100 2
Reference Assembly 2
RNA-Seq 3
Coverage Plots 1
☑ Diavir_Dvir_v2.0_RNA-Seq-alignments_2021-05-06_coverage
Mapped Reads 1
□ Diavir_Dvir_v2.0_RNA-Seq-alignments_2021-05-06
Splice junctions 1
□ Diavir_Dvir_v2.0_RNA-Seq-alignments_2021-05-06_junctions
Apollo
File
View
Help
Login
0
1,000,000
2,000,000
3,000,000
4,000,000
5,000,000
6,000,000
7,000,000
8,000
NW_021038518.1
NW_021038518.1:3593447..3595397
Go
593,500
3,593,476
3,594,000
3,594,500
3,595,000
NCBI_Annotation_Release_100_Gene
XM_028273555.1
lysosomal Pro-X carboxypeptidase-like
Diavir_Dvir_v2.0_RNA-Seq-alignments_2021-05-06_coverage
10
9
8
7
6
5
4
3
2
1

RNA-Seq pipeline outputs

Mapped reads: Individual glyphs of each mapped read. Show mapped and spliced areas, and SNPs/indels. Informative, but hard to work with when zoomed out.



Available Tracks

filter tracks

- NCBI Annotation Release 100 2
- Reference Assembly 2
- RNA-Seq 3
  - Coverage Plots 1
    - □ Diavir_Dvir_v2.0_RNA-Seq-alignments_2021-05-06_coverage
  - Mapped Reads 1
    - ☑ Diavir_Dvir_v2.0_RNA-Seq-alignments_2021-05-06
  - Splice junctions 1
    - □ Diavir_Dvir_v2.0_RNA-Seq-alignments_2021-05-06_junctions

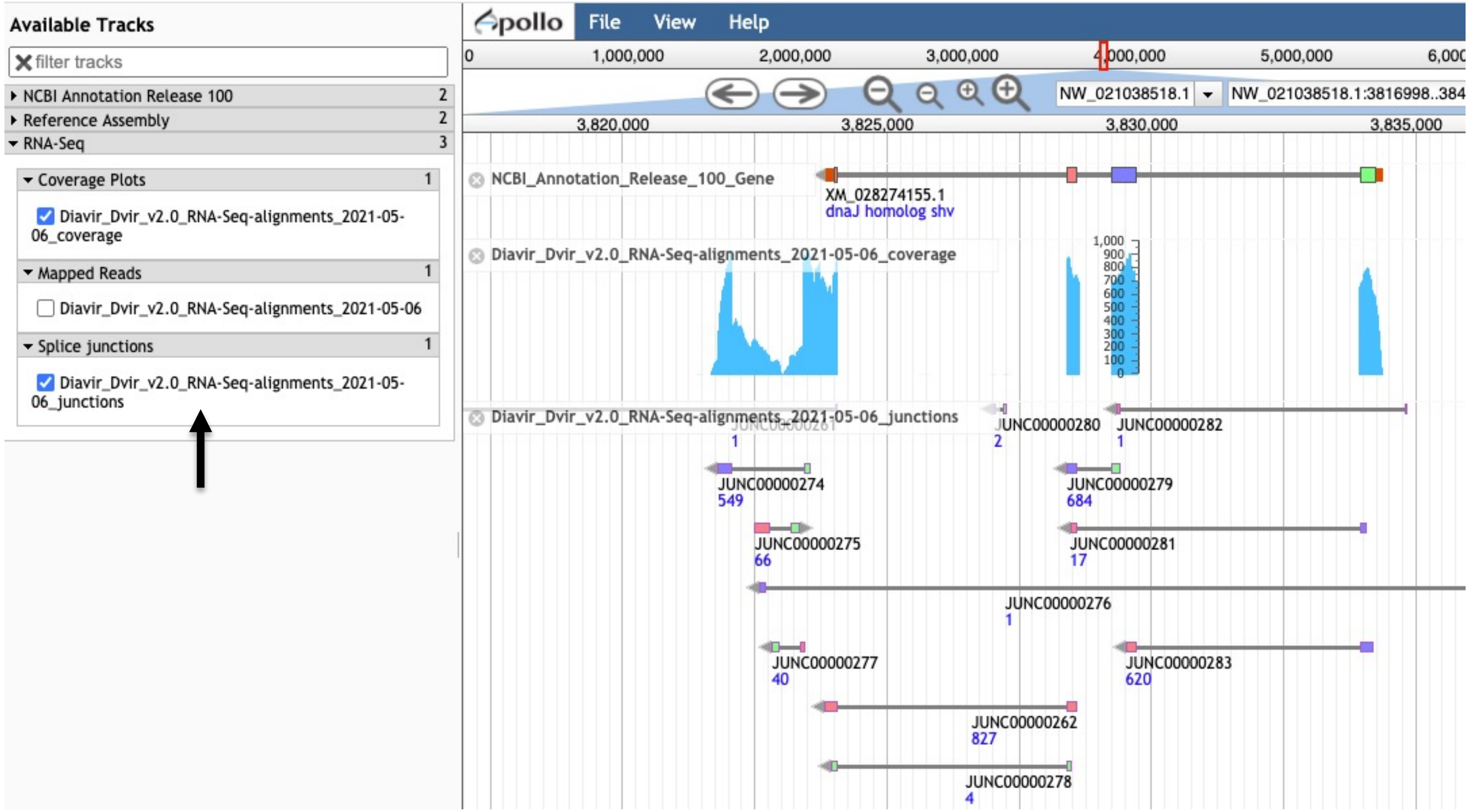
NW_021038518.1

NW_021038518.1:3593447..3595397

RNA-Seq pipeline outputs

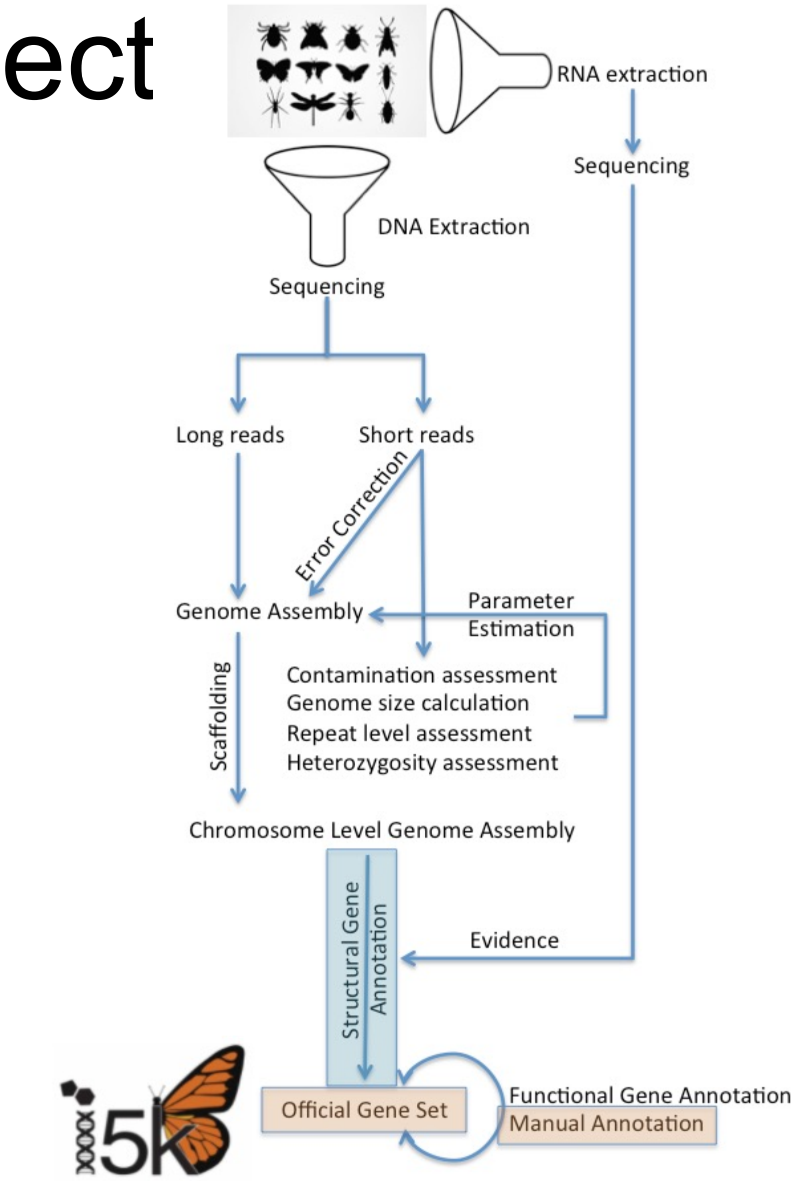
Junction reads:

- Useful combined with coverage plots
- show where mapped reads are spliced.
- Blue numbers show the ‘score’ – the number of mappings that support the splice junction.



The Ag100Pest Project

- USDA-ARS effort to provide reference-quality genome assemblies and annotations for over 100 agricultural pest species relevant to the USA
- USDA-ARS's contribution towards the Earth BioGenome Project (https://www.earthbiogenome.org)
- http://i5k.github.io/ag100pest
- https://www.youtube.com/watch?v=K81AI_ZrQmM
- Executive team members: Anna Childers, Brian Scheffler, Kevin Hackett
- Core team members: Scott Geib, Brad Coates, Tim Smith, Monica Poelchau, Chris Childers



RNA extraction

Sequencing

DNA Extraction

Sequencing

Long reads

Short reads

Error Correction

Genome Assembly

Parameter Estimation

Scaffolding

Contamination assessment
Genome size calculation
Repeat level assessment
Heterozygosity assessment

Chromosome Level Genome Assembly

Structural Gene Annotation

Evidence

Official Gene Set

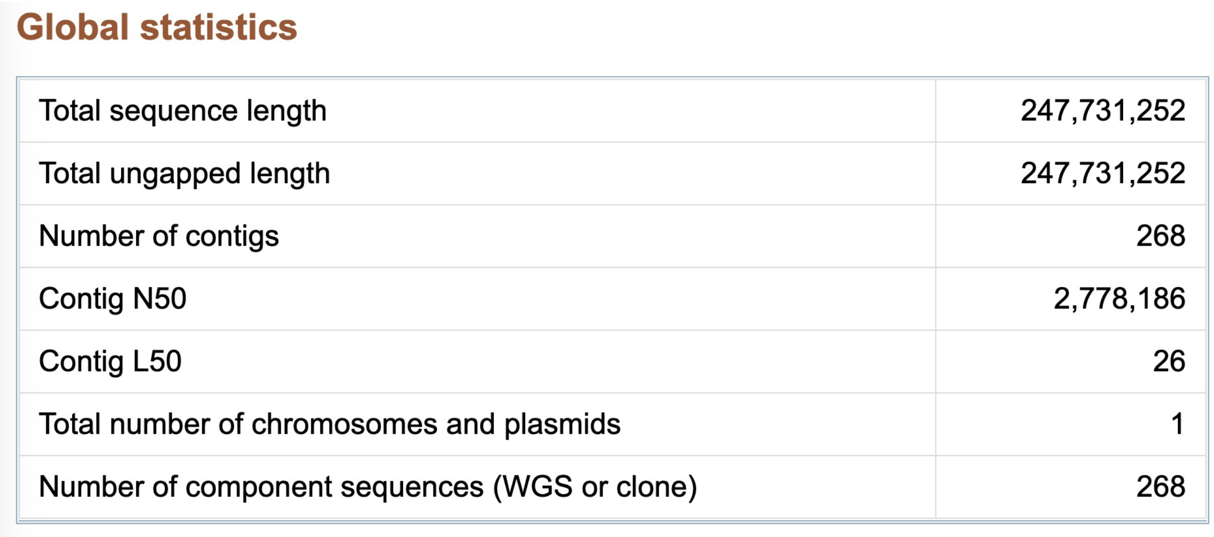
Functional Gene Annotation
Manual Annotation

i5k

Ag100Pest datasets

- Current priority list includes 169 genomes across 8 orders
- 70 genomes have completed sequencing; 25 assemblies should be available at the i5k Workspace by the end of September
- So far, assembly quality is excellent; cf. *Vespa mandarinia*
- The i5k Workspace will be hosting these genomes once they become available

Asian Giant Hornet RefSeq assembly stats



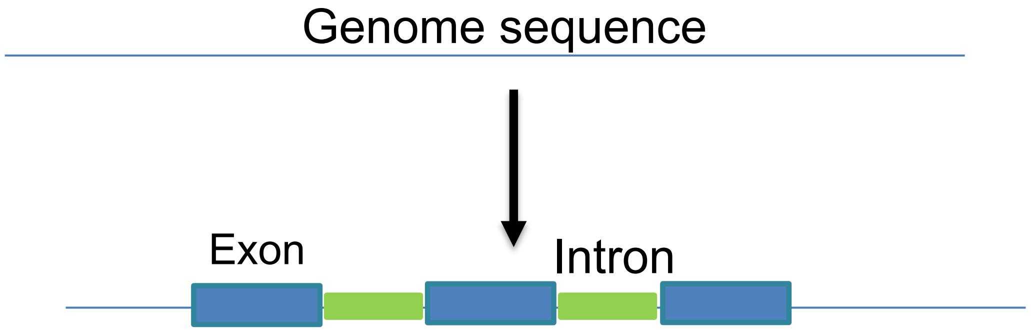
Global statistics

Total sequence length	247,731,252
Total ungapped length	247,731,252
Number of contigs	268
Contig N50	2,778,186
Contig L50	26
Total number of chromosomes and plasmids	1
Number of component sequences (WGS or clone)	268

Functional annotation

Structural annotation

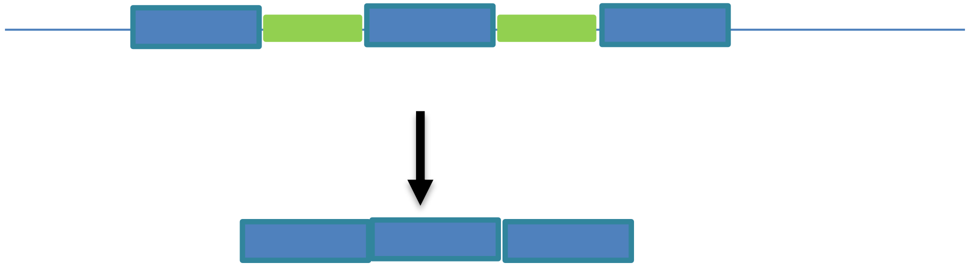
Computationally predict gene structure from genome sequence



Functional annotation

Associate predicted protein sequence with functional terms assigned to homologous protein

- Gene Ontology terms
- Pathway components (e.g. KEGG)

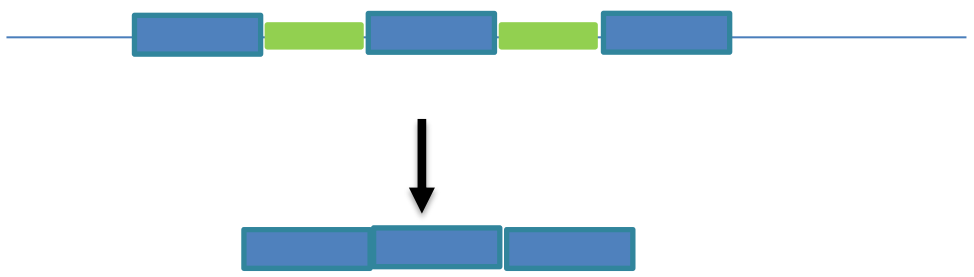


Protein family: short-chain dehydrogenases/reductases (SDR) (P00334)

Functional annotation

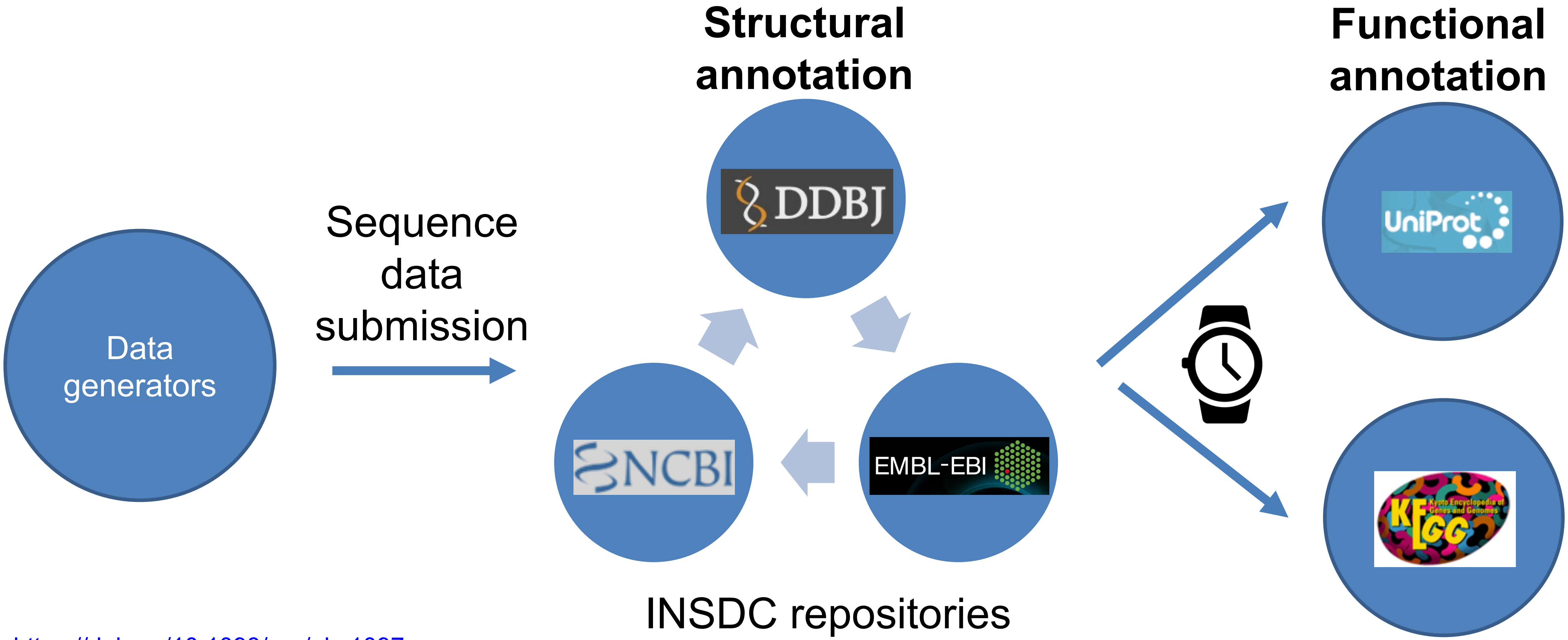
- Useful for first pass at gene function
- Can help prioritize manual annotation efforts by grouping into categories

Functional annotation



Protein family: short-chain dehydrogenases/reductases (SDR) (P00334)

Typical flow of sequence data



Data generators

Sequence data submission

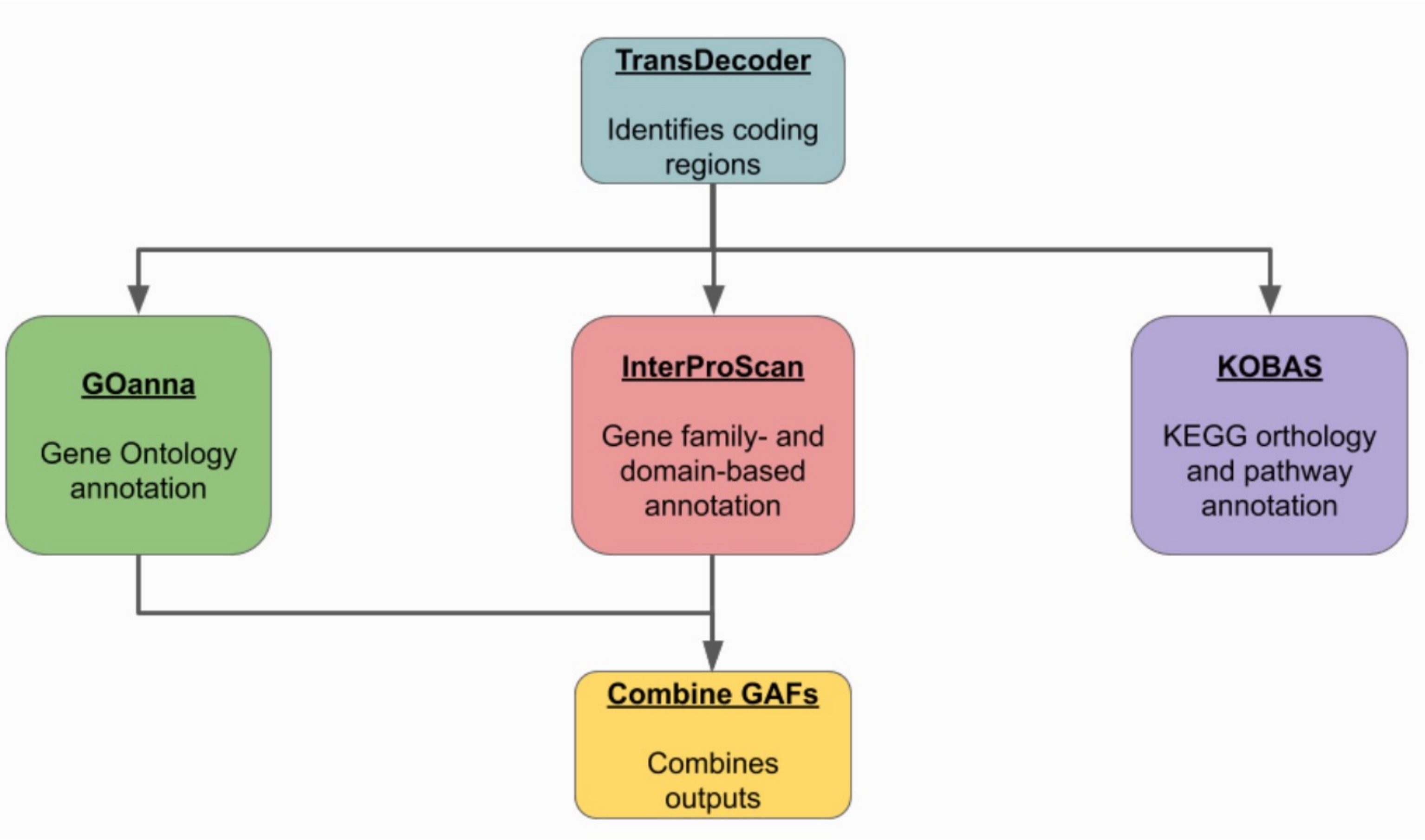
Structural annotation

INSDC repositories

Functional annotation

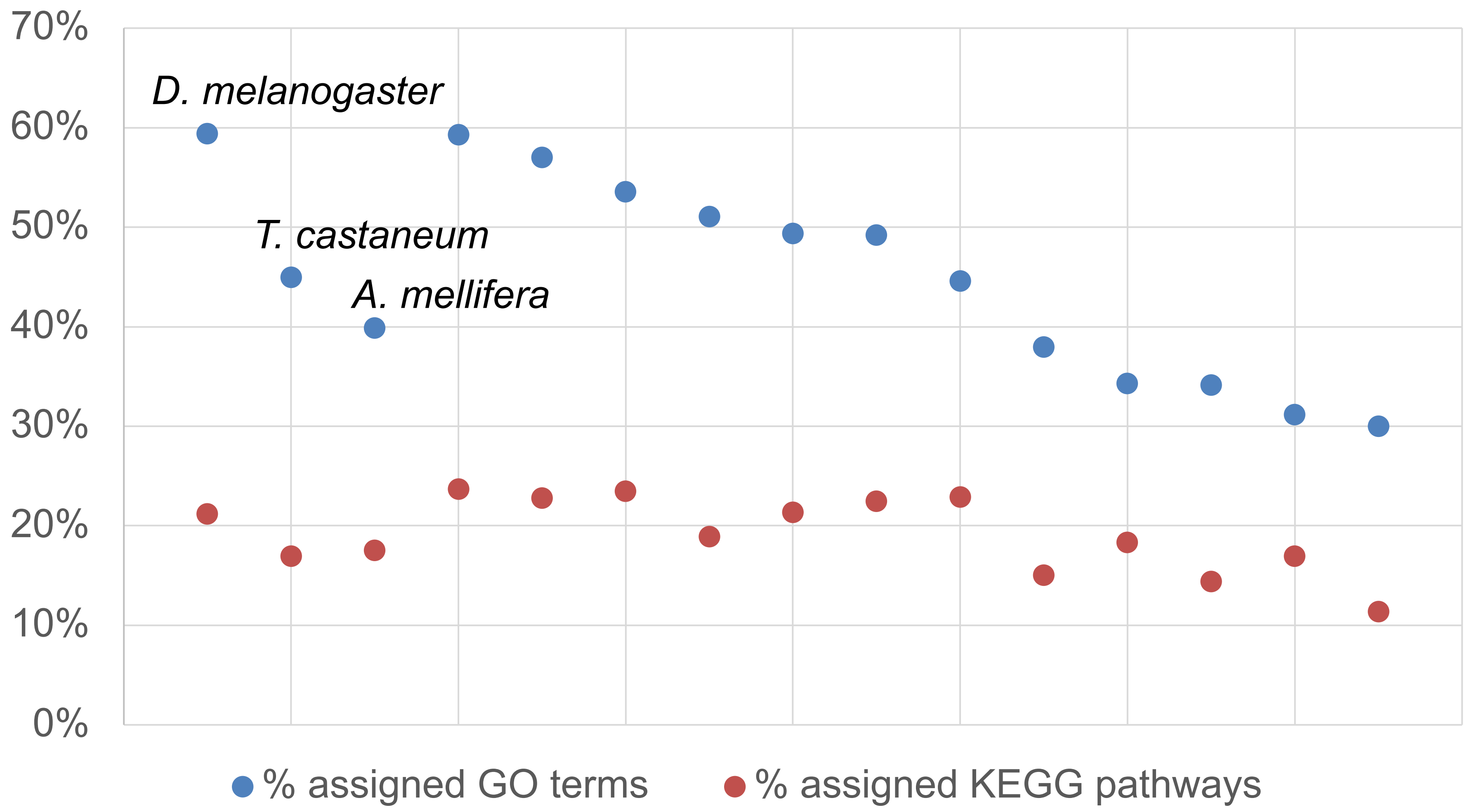
https://doi.org/10.1093/nar/gkx1097

Functional annotation workflow


TransDecoder
Identifies coding regions
GOanna
Gene Ontology annotation
InterProScan
Gene family- and domain-based annotation
KOBAS
KEGG orthology and pathway annotation
Combine GAFs
Combines outputs

- Documentation: https://agbase-docs.readthedocs.io/en/latest/
- Github: https://github.com/AgBase
- Credits: Surya Saha, Amanda Cooksey, Anna Childers, Fiona McCarthy
- Publication currently under development

Functional annotation preliminary results



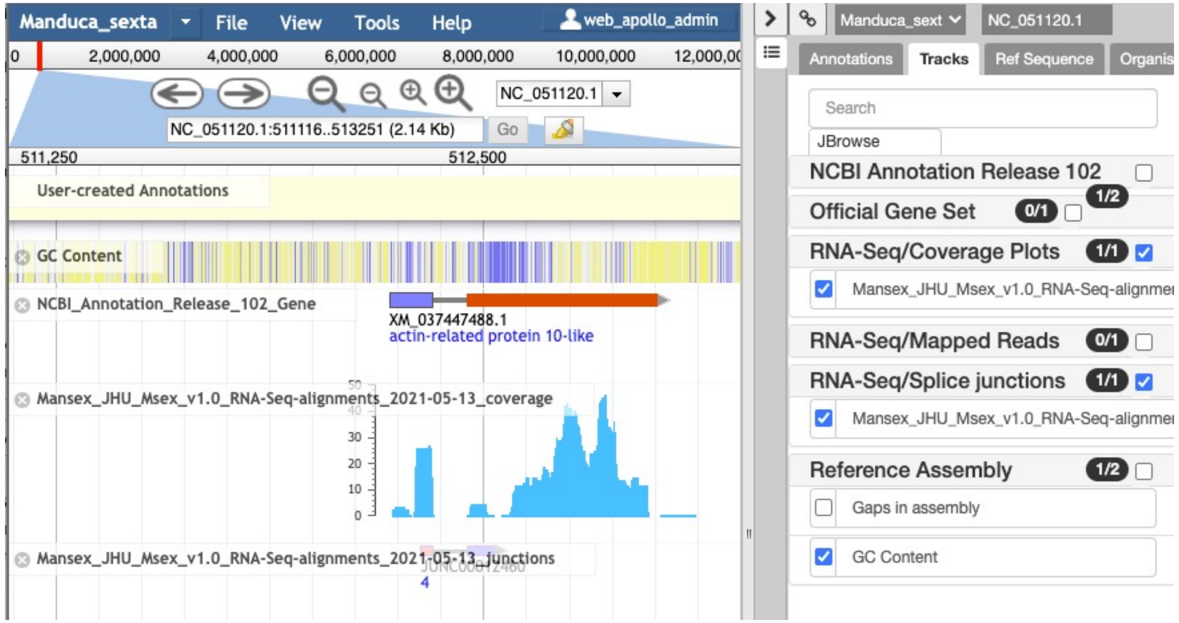
Diaphorina citri
Athalia rosae
Varroa destructor
Agrilus planipennis
Ceratitis capitata
Cimex lectularius
Limnephilus lunatus
Homalodisca vitripennis
Oncopeltus fasciatus
Copidosoma floridanum
Latrodectus hesperus
Eurytemora affinis

Functional annotation workflow

- Links to initial datasets available here: https://i5k.nal.usda.gov/news/functional-annotation-datasets-are-available-11-i5k-workspace-organisms
- Coming next:
  - Publication fully describing the pipeline and results
  - Functional annotations for all i5k Workspace organisms
  - Visualization and search on i5k.nal.usda.gov
  - Training materials for manual annotation

Upcoming Apollo updates and features

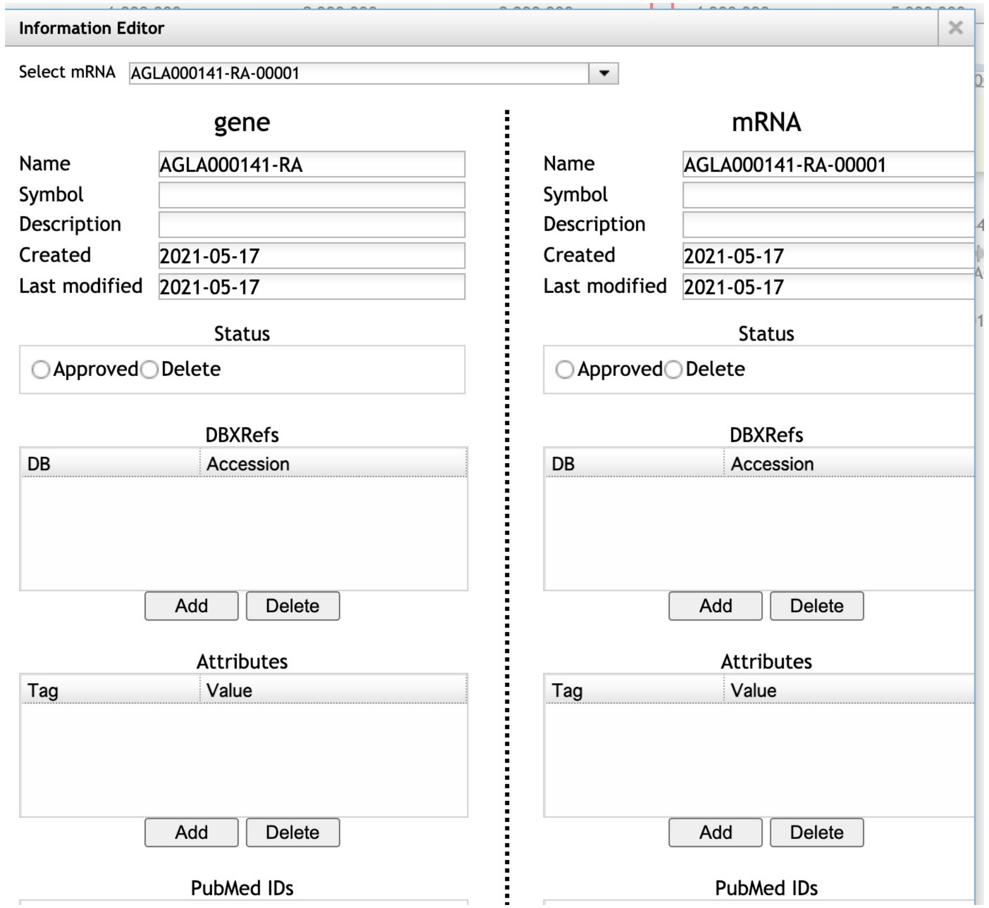
- We will upgrade Apollo2.1 to Apollo2.6 in the upcoming months
- New Features in 2.6.x:
- Information editor looks quite different!
- Blat search is in a different location
- Annotations can be created from blat features


Manduca_sexta
File
View
Tools
Help
web_apollo_admin
0
2,000,000
4,000,000
6,000,000
8,000,000
10,000,000
12,000,00
NC_051120.1
NC_051120.1:511116..513251 (2.14 Kb)
Go
511,250
512,500
User-created Annotations
GC Content
NCBI_Annotation_Release_102_Gene
XM_037447488.1
actin-related protein 10-like
Mansex_JHU_Msex_v1.0_RNA-Seq-alignments_2021-05-13_coverage
50
40
30
20
10
0
Mansex_JHU_Msex_v1.0_RNA-Seq-alignments_2021-05-13_junctions
4
Manduca_sext
NC_051120.1
Annotations
Tracks
Ref Sequence
Organis
Search
JBrowse
NCBI Annotation Release 102 □
Official Gene Set 0/1 □ 1/2
RNA-Seq/Coverage Plots 1/1 ☑
☑ Mansex_JHU_Msex_v1.0_RNA-Seq-alignmer
RNA-Seq/Mapped Reads 0/1 □
RNA-Seq/Splice junctions 1/1 ☑
☑ Mansex_JHU_Msex_v1.0_RNA-Seq-alignmer
Reference Assembly 1/2 □
□ Gaps in assembly
☑ GC Content

USDA National Agricultural Library
U.S. DEPARTMENT OF AGRICULTURE

Apollo updates and features – Information Editor

Current (v2.1.x)



Select mRNA AGLA000141-RA-00001

gene

Name	AGLA000141-RA
Symbol	
Description	
Created	2021-05-17
Last modified	2021-05-17

Status
○ Approved ○ Delete

DBXRefs

DB	Accession

Attributes

Tag	Value

PubMed IDs

mRNA

Name	AGLA000141-RA-00001
Symbol	
Description	
Created	2021-05-17
Last modified	2021-05-17

Status
○ Approved ○ Delete

DBXRefs

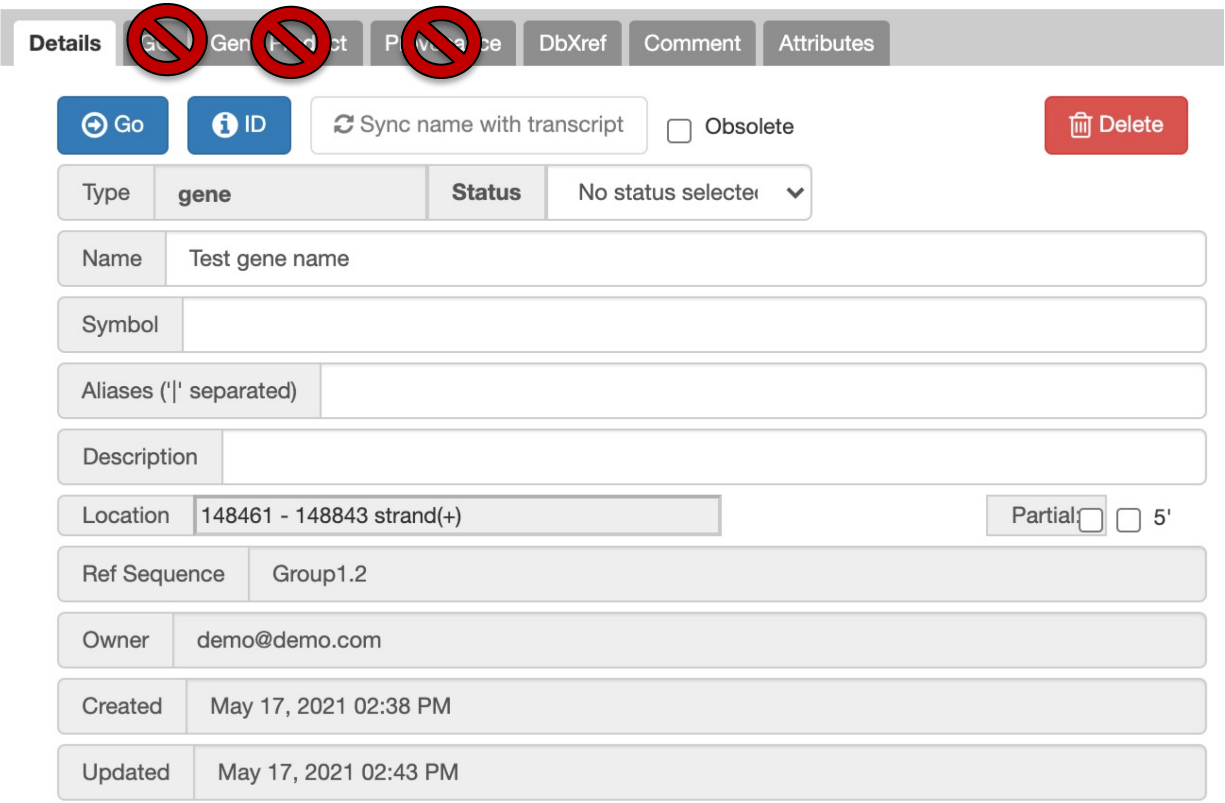
DB	Accession

Attributes

Tag	Value

PubMed IDs

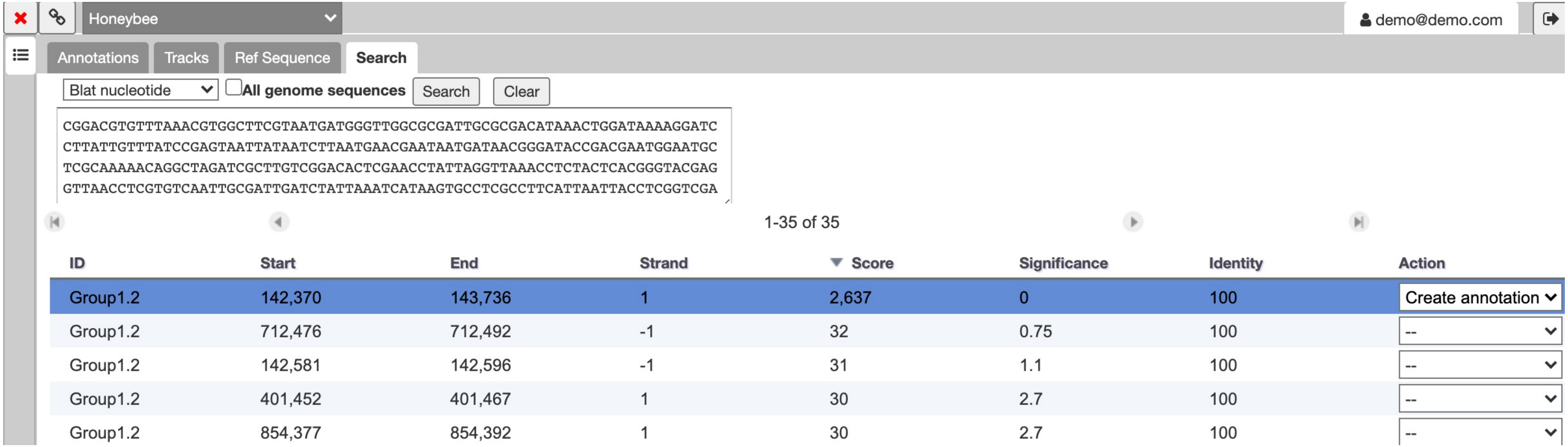
Upcoming (v2.6.x)



□ Obsolete

Type	gene
Status	No status selecte
Name	Test gene name
Symbol	
Aliases ('|' separated)	
Description	
Location	148461 - 148843 strand(+)
Partial	□ □ 5'
Ref Sequence	Group1.2
Owner	demo@demo.com
Created	May 17, 2021 02:38 PM
Updated	May 17, 2021 02:43 PM

Apollo updates and features - Blat search



Blat nucleotide
□ All genome sequences

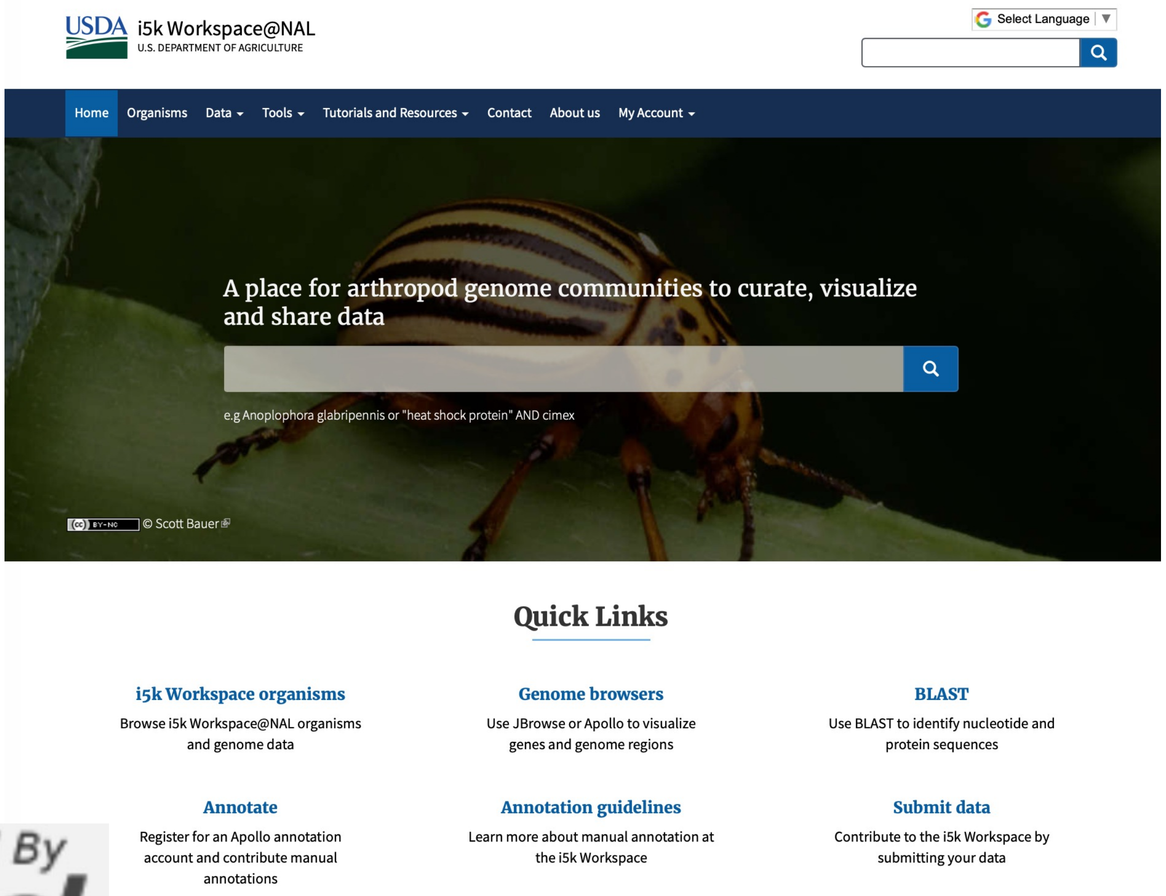
```
CGGACGTGTTTAAACGTGGCTTCGTAATGATGGGTTGGCGCGATTGCGCGACATAAACTGGATAAAAGGATC
CTTATTGTTTATCCGAGTAATTATAATCTTAATGAACGAATAATGATAACGGGATACCGACGAATGGAATGC
TCGCAAAAACAGGCTAGATCGCTTGTCGGACACTCGAACCTATTAGGTTAAACCTCTACTCACGGGTACGAG
GTTAACCTCGTGTCAATTGCGATTGATCTATTAAATCATAAGTGCCTCGCCTTCATTAATTACCTCGGTCGA
```

1-35 of 35

ID	Start	End	Strand	▼ Score	Significance	Identity	Action
Group1.2	142,370	143,736	1	2,637	0	100	Create annotation
Group1.2	712,476	712,492	-1	32	0.75	100	--
Group1.2	142,581	142,596	-1	31	1.1	100	--
Group1.2	401,452	401,467	1	30	2.7	100	--
Group1.2	854,377	854,392	1	30	2.7	100	--

Upcoming i5k Workspace features

- Major upgrade from Tripal v2 to Tripal v3
- Web services – programmatic access to our data
- Different look and feel of our site – in particular organism, gene, analysis pages
- Elasticsearch search engine
- Working on adding gene pages for legacy content



A place for arthropod genome communities to curate, visualize and share data

e.g Anoplophora glabripennis or "heat shock protein" AND cimex

© Scott Bauer

Quick Links

i5k Workspace organisms
Browse i5k Workspace@NAL organisms and genome data

Genome browsers
Use JBrowse or Apollo to visualize genes and genome regions

BLAST
Use BLAST to identify nucleotide and protein sequences

Annotate
Register for an Apollo annotation account and contribute manual annotations

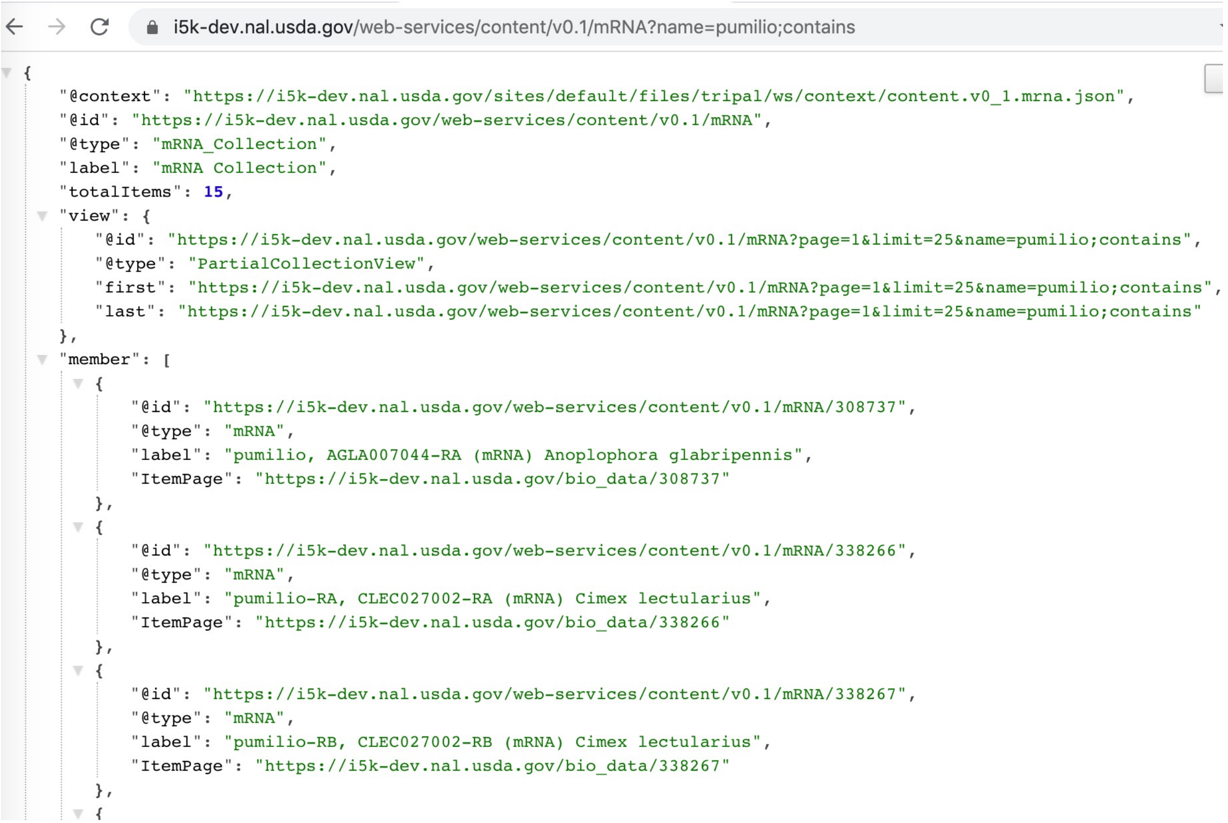
Annotation guidelines
Learn more about manual annotation at the i5k Workspace

Submit data
Contribute to the i5k Workspace by submitting your data

Powered By Tripal

Upcoming i5k Workspace features – web services

- Web services – allow you to programmatically access the i5k Workspace@NAL content
- Full documentation on how to use them: https://tripal.readthedocs.io/en/latest/user_guide/web_services.html
- Example URL: https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA?name=pumilio;contains

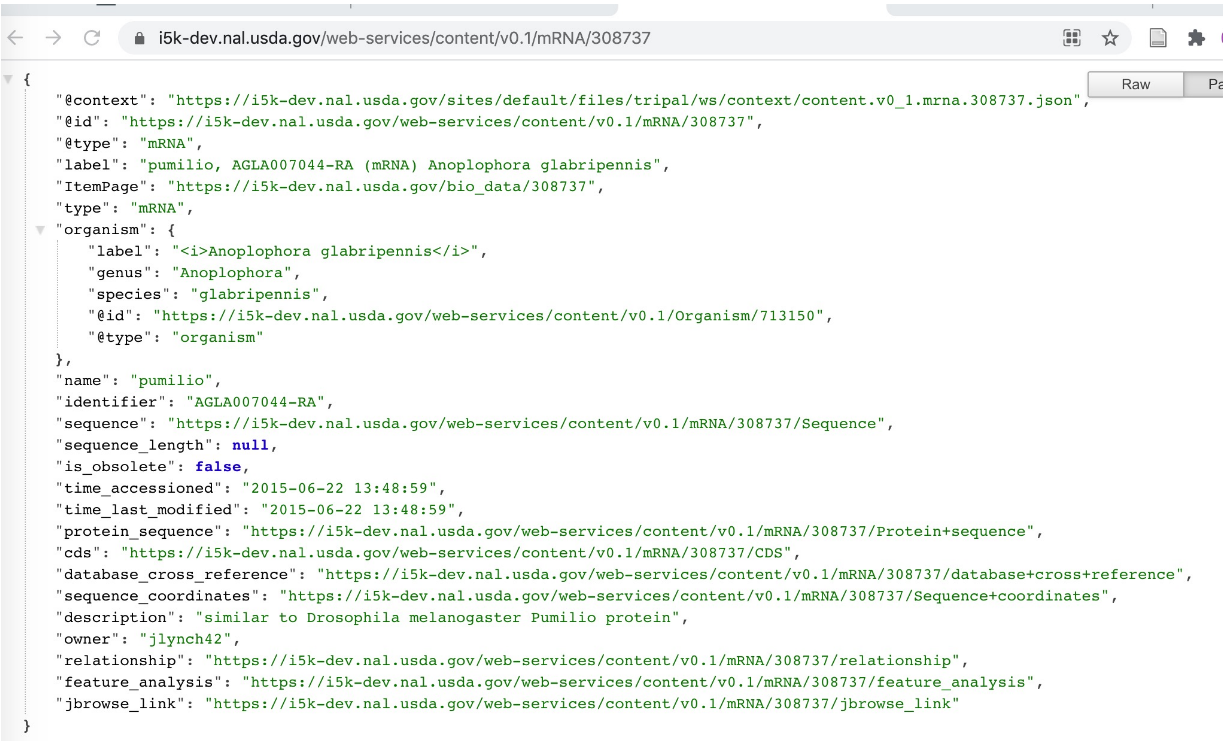


i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA?name=pumilio;contains

```
{
    "@context": "https://i5k-dev.nal.usda.gov/sites/default/files/tripal/ws/context/content.v0_1.mrna.json",
    "@id": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA",
    "@type": "mRNA_Collection",
    "label": "mRNA Collection",
    "totalItems": 15,
    "view": {
        "@id": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA?page=1&limit=25&name=pumilio;contains",
        "@type": "PartialCollectionView",
        "first": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA?page=1&limit=25&name=pumilio;contains",
        "last": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA?page=1&limit=25&name=pumilio;contains"
    },
    "member": [
        {
            "@id": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737",
            "@type": "mRNA",
            "label": "pumilio, AGLA007044-RA (mRNA) Anoplophora glabripennis",
            "ItemPage": "https://i5k-dev.nal.usda.gov/bio_data/308737"
        },
        {
            "@id": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/338266",
            "@type": "mRNA",
            "label": "pumilio-RA, CLEC027002-RA (mRNA) Cimex lectularius",
            "ItemPage": "https://i5k-dev.nal.usda.gov/bio_data/338266"
        },
        {
            "@id": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/338267",
            "@type": "mRNA",
            "label": "pumilio-RB, CLEC027002-RB (mRNA) Cimex lectularius",
            "ItemPage": "https://i5k-dev.nal.usda.gov/bio_data/338267"
        },
        {
```

Upcoming i5k Workspace features – web services

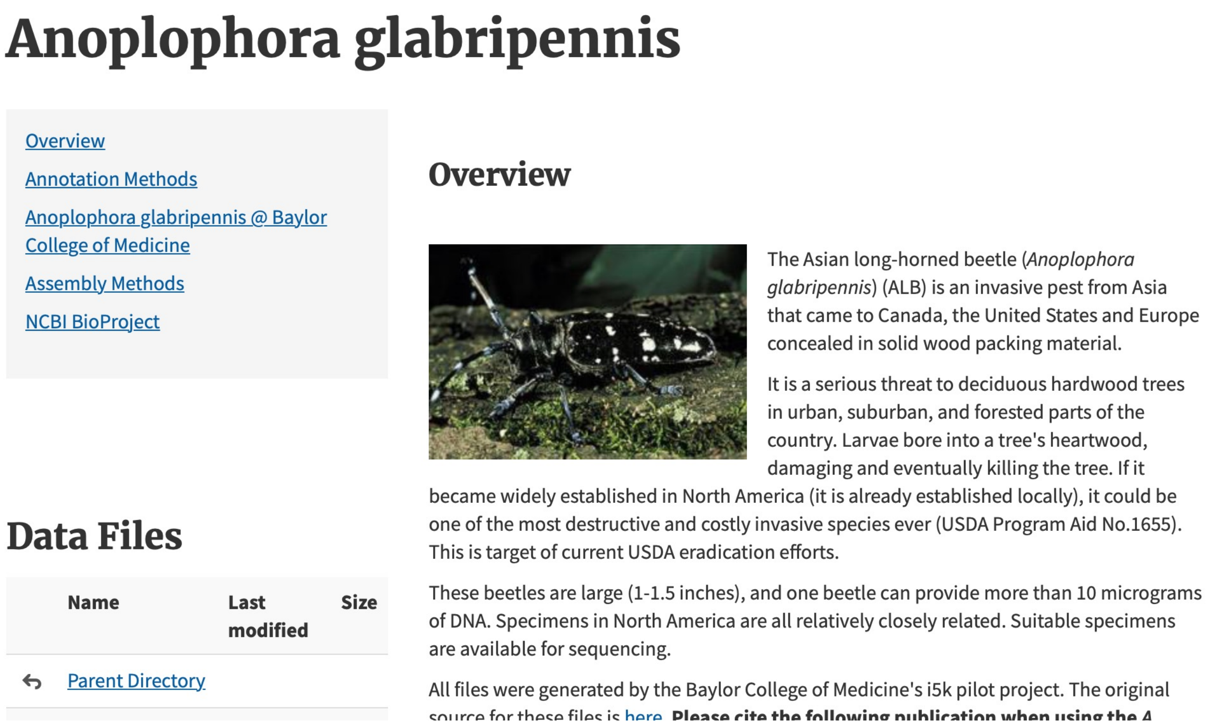
- Full information on one of the mRNAs from the previous search
- https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737



```
{
  "@context": "https://i5k-dev.nal.usda.gov/sites/default/files/tripal/ws/context/content.v0_1.mrna.308737.json",
  "@id": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737",
  "@type": "mRNA",
  "label": "pumilio, AGLA007044-RA (mRNA) Anoplophora glabripennis",
  "ItemPage": "https://i5k-dev.nal.usda.gov/bio_data/308737",
  "type": "mRNA",
  "organism": {
    "label": "<i>Anoplophora glabripennis</i>",
    "genus": "Anoplophora",
    "species": "glabripennis",
    "@id": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/Organism/713150",
    "@type": "organism"
  },
  "name": "pumilio",
  "identifier": "AGLA007044-RA",
  "sequence": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737/Sequence",
  "sequence_length": null,
  "is_obsolete": false,
  "time_accessioned": "2015-06-22 13:48:59",
  "time_last_modified": "2015-06-22 13:48:59",
  "protein_sequence": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737/Protein+sequence",
  "cds": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737/CDS",
  "database_cross_reference": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737/database+cross+reference",
  "sequence_coordinates": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737/Sequence+coordinates",
  "description": "similar to Drosophila melanogaster Pumilio protein",
  "owner": "jlynch42",
  "relationship": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737/relationship",
  "feature_analysis": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737/feature_analysis",
  "jbrowse_link": "https://i5k-dev.nal.usda.gov/web-services/content/v0.1/mRNA/308737/jbrowse_link"
}
```

Upcoming i5k Workspace features – organism pages

Current



Anoplophora glabripennis

- Overview
- Annotation Methods
- Anoplophora glabripennis @ Baylor College of Medicine
- Assembly Methods
- NCBI BioProject

Overview

The Asian long-horned beetle (*Anoplophora glabripennis*) (ALB) is an invasive pest from Asia that came to Canada, the United States and Europe concealed in solid wood packing material.

It is a serious threat to deciduous hardwood trees in urban, suburban, and forested parts of the country. Larvae bore into a tree's heartwood, damaging and eventually killing the tree. If it became widely established in North America (it is already established locally), it could be one of the most destructive and costly invasive species ever (USDA Program Aid No.1655). This is target of current USDA eradication efforts.

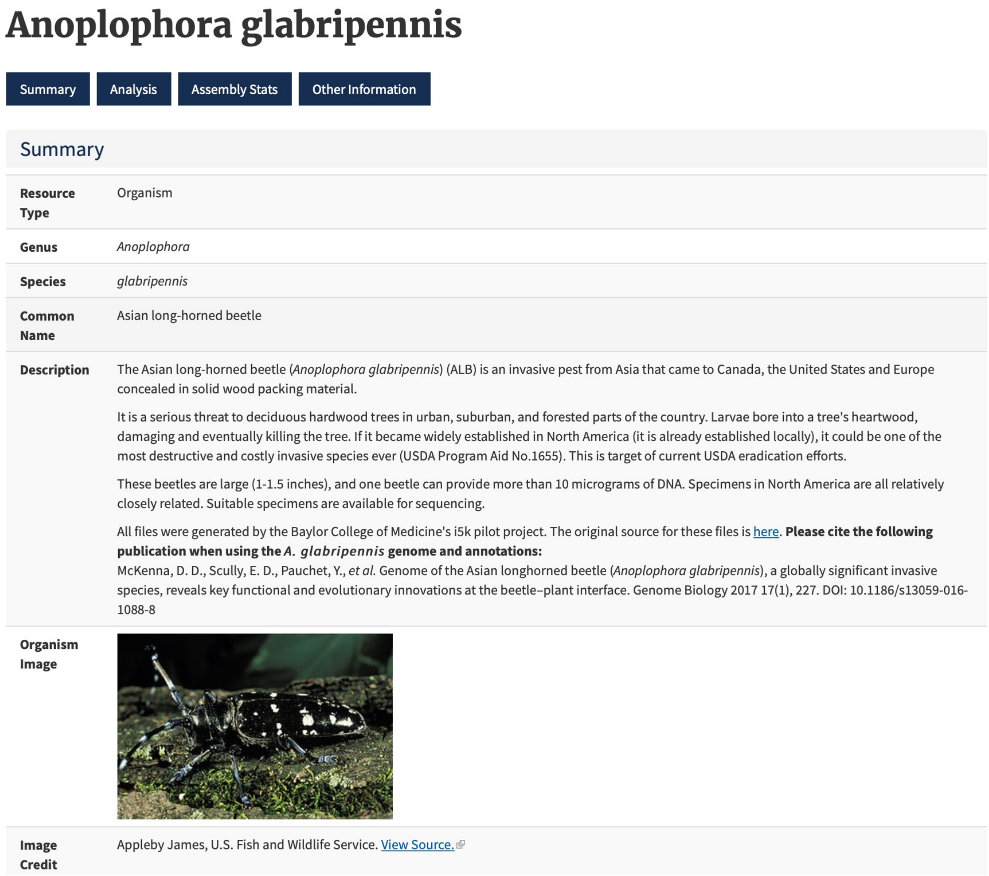
These beetles are large (1-1.5 inches), and one beetle can provide more than 10 micrograms of DNA. Specimens in North America are all relatively closely related. Suitable specimens are available for sequencing.

All files were generated by the Baylor College of Medicine's i5k pilot project. The original source for these files is here. **Please cite the following publication when using the A**

Data Files

	Name	Last modified	Size
↩	Parent Directory		

Upcoming

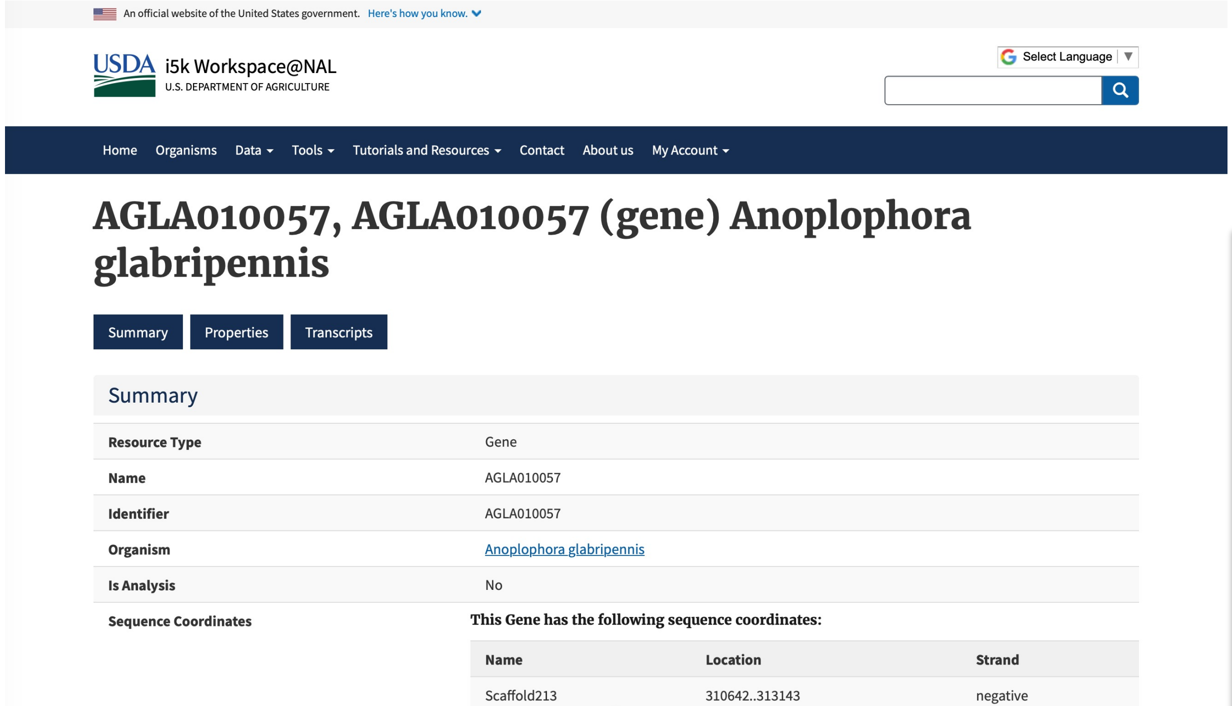


Anoplophora glabripennis

Summary

Resource Type	Organism
Genus	*Anoplophora*
Species	*glabripennis*
Common Name	Asian long-horned beetle
Description	The Asian long-horned beetle (*Anoplophora glabripennis*) (ALB) is an invasive pest from Asia that came to Canada, the United States and Europe concealed in solid wood packing material. It is a serious threat to deciduous hardwood trees in urban, suburban, and forested parts of the country. Larvae bore into a tree's heartwood, damaging and eventually killing the tree. If it became widely established in North America (it is already established locally), it could be one of the most destructive and costly invasive species ever (USDA Program Aid No.1655). This is target of current USDA eradication efforts. These beetles are large (1-1.5 inches), and one beetle can provide more than 10 micrograms of DNA. Specimens in North America are all relatively closely related. Suitable specimens are available for sequencing. All files were generated by the Baylor College of Medicine's i5k pilot project. The original source for these files is here. **Please cite the following publication when using the *A. glabripennis* genome and annotations:** McKenna, D. D., Scully, E. D., Pauchet, Y., *et al.* Genome of the Asian longhorned beetle (*Anoplophora glabripennis*), a globally significant invasive species, reveals key functional and evolutionary innovations at the beetle–plant interface. Genome Biology 2017 17(1), 227. DOI: 10.1186/s13059-016-1088-8
Organism Image	
Image Credit	Appleby James, U.S. Fish and Wildlife Service. View Source.

Upcoming i5k Workspace features – gene pages



An official website of the United States government. Here's how you know.

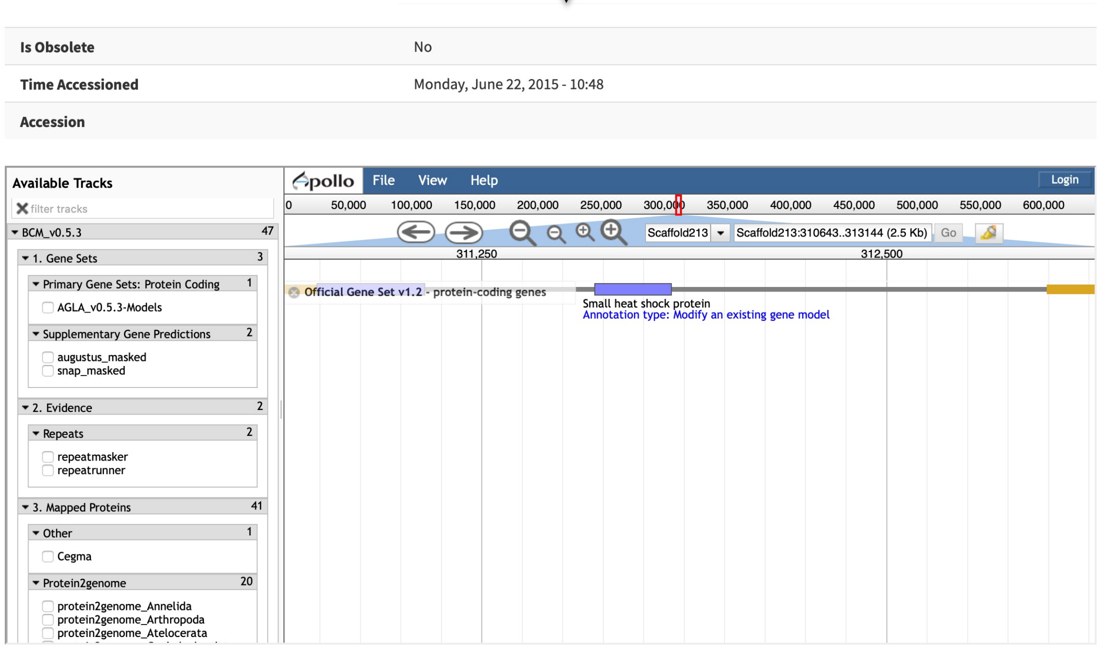
USDA i5k Workspace@NAL
U.S. DEPARTMENT OF AGRICULTURE

AGLA010057, AGLA010057 (gene) Anoplophora glabripennis

Summary

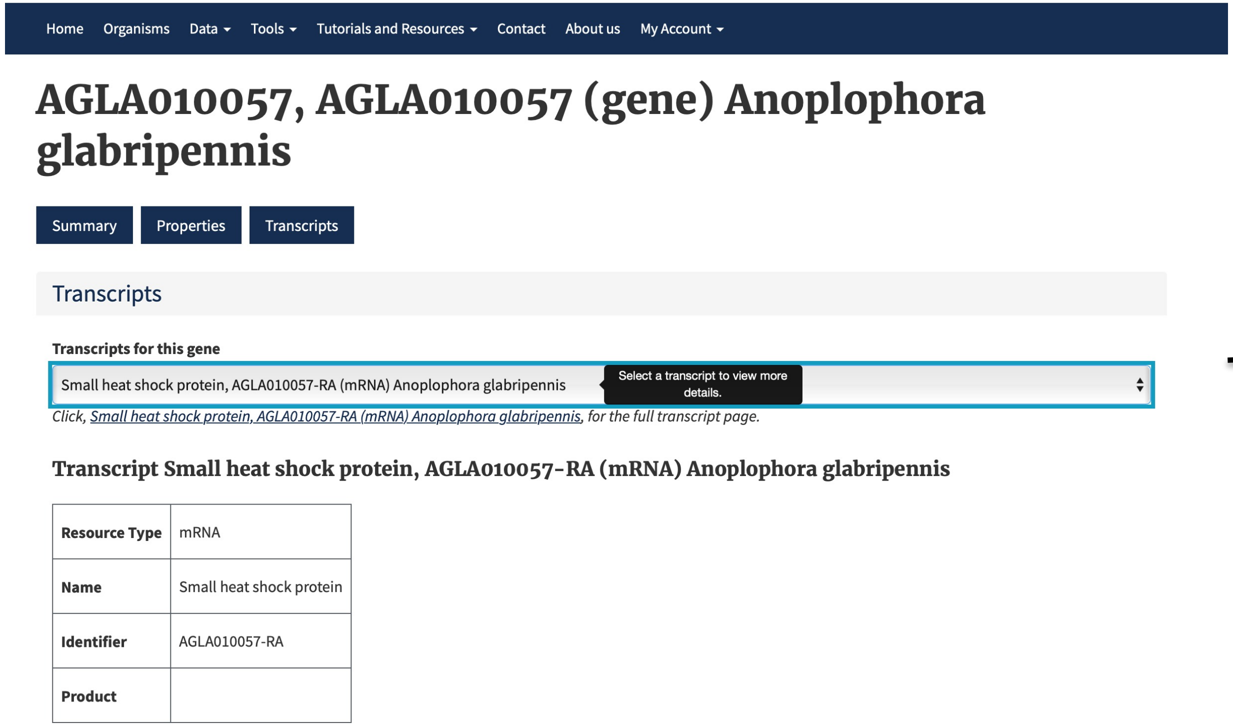
Resource Type	Gene
Name	AGLA010057
Identifier	AGLA010057
Organism	Anoplophora glabripennis
Is Analysis	No
Sequence Coordinates	This Gene has the following sequence coordinates:

Name	Location	Strand
Scaffold213	310642..313143	negative



Is Obsolete	No
Time Accessioned	Monday, June 22, 2015 - 10:48
Accession	

Upcoming i5k Workspace features – gene pages



AGLA010057, AGLA010057 (gene) Anoplophora glabripennis

Transcripts

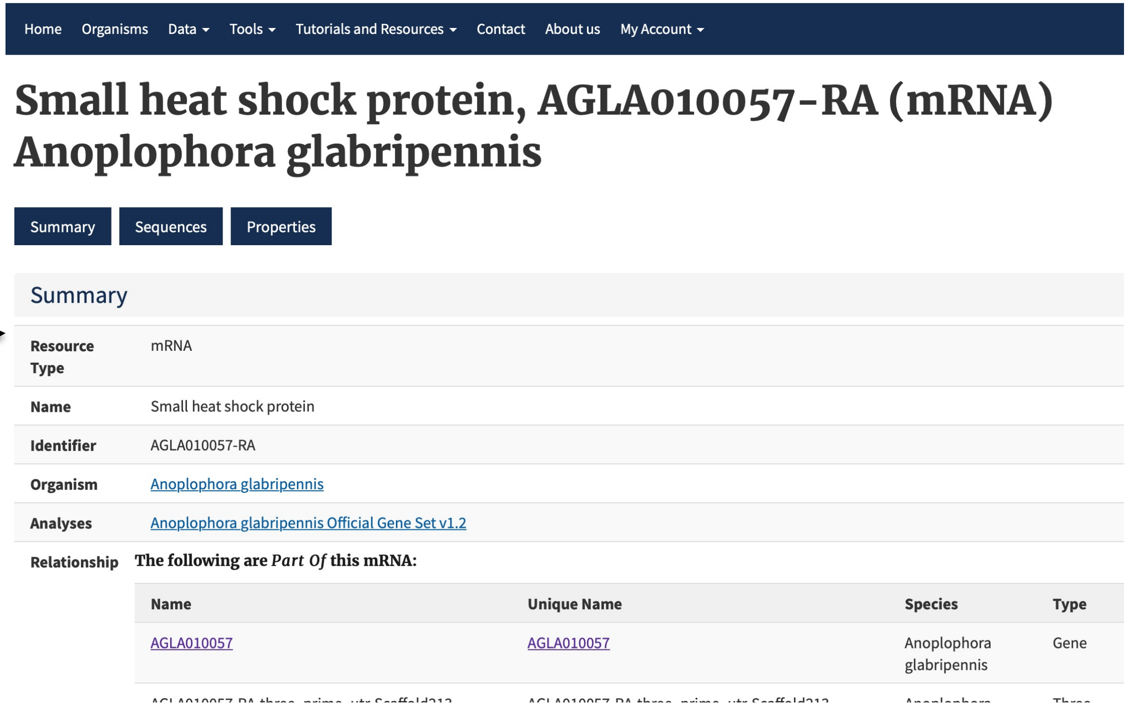
Transcripts for this gene

Small heat shock protein, AGLA010057-RA (mRNA) Anoplophora glabripennis

Click, Small heat shock protein, AGLA010057-RA (mRNA) Anoplophora glabripennis, for the full transcript page.

Transcript Small heat shock protein, AGLA010057-RA (mRNA) Anoplophora glabripennis

Resource Type	mRNA
Name	Small heat shock protein
Identifier	AGLA010057-RA
Product	



Small heat shock protein, AGLA010057-RA (mRNA) Anoplophora glabripennis

Summary

Resource Type	mRNA
Name	Small heat shock protein
Identifier	AGLA010057-RA
Organism	Anoplophora glabripennis
Analyses	Anoplophora glabripennis Official Gene Set v1.2
Relationship	**The following are *Part Of* this mRNA:**

Name	Unique Name	Species	Type
AGLA010057	AGLA010057	Anoplophora glabripennis	Gene
AGLA010057-RA-three_prime_utr-Scaffold213	AGLA010057-RA-three_prime_utr-Scaffold213	Anoplophora	Three

Thank you!

- The NAL Team
- i5k Coordinating Committee
- I5k Workspace working group
- Apollo & JBrowse Development Teams
- GMOD/Tripal community
- All of our users and contributors!

Contact us:

https://i5k.nal.usda.gov/contact

i5k@ars.usda.gov