I5k Workspace webinar
Changes to gene and protein data access

Monica Poelchau
7/19/2022
Agenda

• Why did we remove the i5k Workspace gene and protein pages?
  – What exactly did we remove?
  – Why?
  – A brief digression on how we keep i5k Workspace data in sync with NCBI

• The data are still accessible – how can you find what you need?
  – Gene and protein information
  – Dataset-level information

• Q&A to address any questions or concerns.
Gene and protein page removal

- What was removed?
  - Pages that showed information at the individual gene and protein level
  - Ability to search for gene and protein information (e.g. “Cimex lectularius heat shock protein”)

CCIN027589, CCIN027589 (gene) Cephus cinctus

Transcripts
The following features are part of this gene:

**CCIN027589-RA**

<table>
<thead>
<tr>
<th>Name</th>
<th>Odorant receptor 58</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>CCIN027589-RA</td>
</tr>
<tr>
<td>Type</td>
<td>mRNA</td>
</tr>
<tr>
<td>Ddbref</td>
<td>NCBI_GP:RLZ02283.1</td>
</tr>
<tr>
<td>Analysis</td>
<td>Cephus cinctus annotations cepcin_065v1.1</td>
</tr>
<tr>
<td></td>
<td>Source: Cephus cinctus genome assembly Ccin1 (GCF_000341935.1)</td>
</tr>
<tr>
<td>Annotator Comments</td>
<td>Note: manually curated model, revised mRNA compared to XM_015753704.2; manually curated model, revised mRNA compared to XM_015753705.2</td>
</tr>
<tr>
<td>owner</td>
<td>hrobertson</td>
</tr>
</tbody>
</table>
Why did we remove the pages?

1. The data are already available elsewhere at the gene/protein level:
   1. NCBI
   2. Other insect databases, e.g. HymenopteraMine, VEuPathDB

2. The data are still available at the dataset level:
   1. i5k Workspace@NAL
   2. Ag Data Commons
   3. NCBI
   4. Other insect databases, e.g. HymenopteraMine, VEuPathDB

3. Not a high priority for users
   1. Structured interview results
   2. Fewer page views

4. Challenging to maintain the code

5. Challenging to keep the content up-to-date and comprehensive
Why are the data at NCBI?

- NCBI, as part of the INSDC, is the primary archive for all sequence data.
- Our policy is to submit all sequence data and metadata that can go to NCBI, to NCBI.

DATA GENERATORS

INSDC REPOSITORIES

[Link: https://doi.org/10.1093/nar/gkx1097]
Manual annotation QA/QC and submission

1. Manual annotation
2. Pull new NCBI metadata
3. QA/QC of GFF3
4. NCBI QA/QC
5. NCBI submission (sqn)
Result of a successful GenBank submission

Odorant receptor 58 [Cephus cinctus]

GenBank: RLZ02283.1

LOCUS RLZ02283 411 aa linear INV 24-OCT-2018
DEFINITION Odorant receptor 58 [Cephus cinctus].
ACCESSION RLZ02283
VERSION RLZ02283.1
DBLINK BioProject: PRJNA168335
            BioSample: SAMN02905554
DBSOURCE accession KB467292.1
KEYWORDS Cephus cinctus (wheat stem sawfly)
ORGANISM Cephus cinctus
            Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;
            Pterygota; Neoptera; Holometabola; Hymenoptera; Cepheoidea;
            Cephidae; Cephus.
REFERENCE 1 (residues 1 to 411)
            AUTHORS Robertson,H.M., Robinson,G.E., Wanner,K.W. and Walden,K.K.O.
            TITLE The Genome of the Wheatstem Sawfly, Cephus cinctus
            JOURNAL Unpublished
REFERENCE 2 (residues 1 to 411)
            AUTHORS Robertson,H.M., Robinson,G.E., Wanner,K.W. and Walden,K.K.O.
            TITLE Direct Submission
            JOURNAL Submitted (31-AUG-2012) Entomology, University of Illinois at
            
How can I find a specific gene?

NCBI Protein search


NCBI Gene search

How can I find a specific gene?

Vectors

https://vectorbase.org/vectorbase/app

Hymenoptera

https://hymenopteramine.rnet.missouri.edu/hymenopteramine/begin.do
How can I find a gene set?

**I5k Workspace@NAL search**

[Image of search results]

**I5k Workspace@NAL browse**

[Image of organism overview]

https://i5k.nal.usda.gov/  
https://i5k.nal.usda.gov/organisms
How can I find a gene set?

I5k Workspace@NAL browse

Ladona fulva annotations Official Gene Set ladful_OGSv1.1

Summary

<table>
<thead>
<tr>
<th>Resource Type</th>
<th>Genome Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Ladona fulva annotations Official Gene Set ladful_OGSv1.1</td>
</tr>
<tr>
<td>Program, Workflow, Method Name</td>
<td>I5k Workspace Official Gene Set workflow, GenBank accessions</td>
</tr>
<tr>
<td>Program Version</td>
<td>NA</td>
</tr>
<tr>
<td>Data Source</td>
<td>Source Name : Ladona fulva genome assembly LfH_v2.0 (GCA_000376725.2)</td>
</tr>
<tr>
<td></td>
<td>Source URI : <a href="https://i5k-nal.usda.gov/data/i5ktesopeda/ladful_h2lf_genome_fulva_v2.0/">https://i5k-nal.usda.gov/data/i5ktesopeda/ladful_h2lf_genome_fulva_v2.0/</a></td>
</tr>
<tr>
<td></td>
<td>Organism : L. fulva (Scarce Chaser)</td>
</tr>
<tr>
<td></td>
<td>Description : The Ladona fulva Official Gene Set ladful_OGSv1.1 is a minor update of Official Gene Set ladful_OGSv1.0. Protein pages for the manual annotations can be accessed at NCBI: <a href="https://www.ncbi.nlm.nih.gov/variation/review/ladful_v2.0/">https://www.ncbi.nlm.nih.gov/variation/review/ladful_v2.0/</a> Laduna fulva genome assembly LfH_v2.0 (GCA_000376725.2). The full dataset is accessible at the Ag Data Commons: <a href="https://doi.org/10.15482/USDA.ACC/1323395/">https://doi.org/10.15482/USDA.ACC/1323395/</a></td>
</tr>
</tbody>
</table>

Analysis

<table>
<thead>
<tr>
<th>Name</th>
<th>Program</th>
<th>Date Constructed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole genome assembly of Ladona fulva</td>
<td>Baylor College of Medicine genome assembly pipeline</td>
<td>Jul 16th, 2013</td>
</tr>
<tr>
<td>BCM annotation of the Ladona fulva assembly using Maker and additional analyses</td>
<td>MAKER</td>
<td>Feb 6th, 2014</td>
</tr>
<tr>
<td>Ladona fulva annotations Official Gene Set ladful_OGSv1.0</td>
<td>I5k Workspace Official Gene Set workflow</td>
<td>Feb 18th, 2021</td>
</tr>
<tr>
<td>Ladona fulva genome assembly LfH_v2.0 (GCA_000376725.2)</td>
<td>AllPaths, ATLAS-link, ATLAS-gpffl, redundans</td>
<td>Oct 31st, 2017</td>
</tr>
<tr>
<td>Ladona fulva annotations Official Gene Set ladful_OGSv1.1</td>
<td>I5k Workspace Official Gene Set workflow, GenBank accessions</td>
<td>Oct 6th, 2021</td>
</tr>
</tbody>
</table>
How can I find a gene set?

**NCBI search**

- How to search:
  - **Genome**
  - **ncbi.nlm.nih.gov/genome/**

**NCBI datasets**

- How to access datasets:
  - **ncbi.nlm.nih.gov/datasets/**
  - FYI, don’t use Safari
How can I find a gene set?

Ag Data Commons search

i5k Workspace

About the i5k Workspace

The i5k Workspace (https://i5k.nal.usda.gov) is an inclusive genome portal for any arthropod genome project that would like to make use of our resources. We provide download services, BLAST, the JBrowse genome browser, and the Apollo manual curation service. Over 50 arthropod genomes are now part of the i5k Workspace, and users are encouraged to browse the genomes that we host, and contribute to the curation of each genome. For more information about the i5k Workspace, you can read our paper on the i5k Workspace, view our posters and talks, and find our software projects on GitHub. The Ag Data Commons is now hosting a growing number of i5k Workspace datasets.

About the i5k Initiative

The i5k Initiative is a transformative project that aims to sequence and analyze the genomes of 5,000 arthropod species. The National Agricultural Library has partnered with the i5k Initiative to create the i5k Workspace@NAL, which serves any ‘orphanned’ arthropod genome project's hosting needs. For more information about the i5k initiative, read the paper and visit the website.

https://data.nal.usda.gov/
Questions?
Thank you!

• Contact us:
  – i5k@usda.gov
  – https://i5k.nal.usda.gov/contact
List of datasets with removed gene pages