AGS Curation Clinic – How to name annotations

Monica Poelchau
National Agricultural Library
USDA-ARS
June 12th, 2019

Agenda

- The bigger data picture – what happens to the data that you generate
  - Why there are standards and rules
- Naming genes and proteins
  - Naming definitions
  - I5k Workspace naming guidelines
  - Adding names, etc. with the Apollo information editor
  - Much of this information will apply to other databases - check their guidelines, though

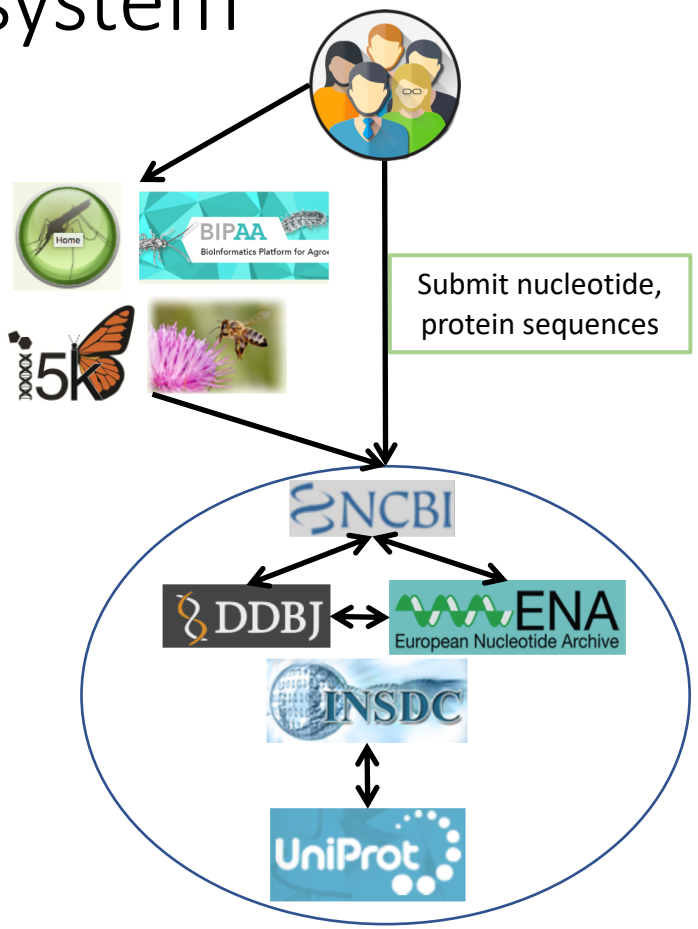
Background

The i5k Workspace@NAL (https://i5k.nal.usda.gov)

- The i5k Workspace@NAL is **the ARS database for arthropod genes and genomes**
- Led by Chris Childers and Monica Poelchau (NAL)
- **Provides access** to **71** arthropod genome projects and counting;
- Facilitates community-driven, manual gene annotation **curation** of over **15,000** gene models;
- Provides webinars, tutorials, and training for the i5k community
- See more at Poster #18 on Friday

A generalized and idealized overview of the sequence ‘data ecosystem’

- User creates data and submits to NCBI
- Data gets propagated through INSDC (ENA, DDBJ, NCBI)
- UniProt ingests protein data
- All of these large databases provide value-added information
  - Archiving
  - Tools
  - QC
  - Functional information
- Smaller community databases add data in to the larger ecosystem, or consume it
  - Improved data and metadata
  - Services and tools tailored towards specific communities
  - Serve as an interface between the user and larger DBs
- Standards and guidelines ease the transitions and movement of data throughout these databases



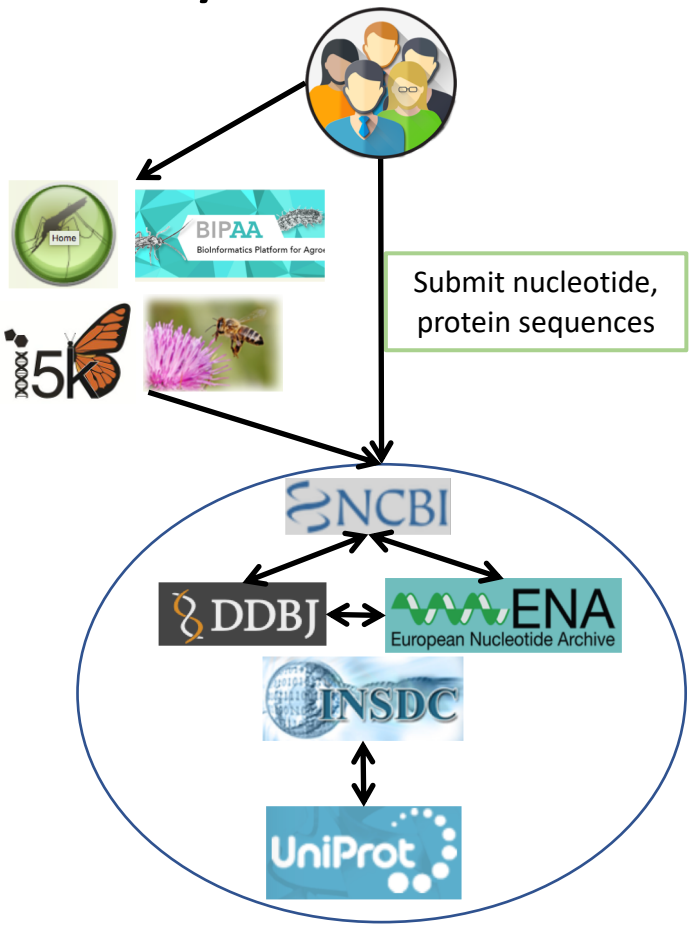
USDA

Naming standards

- Several larger genome communities have committees (sometimes funded) for naming standard development and enforcement
  - E.g. in human, vertebrates, fly, maize
- I5k Workspace doesn't have such a committee.
  - Your name gatekeepers are mainly me and NCBI
- We recommend and use the "International Protein Nomenclature Guidelines" (IPNG), tailored towards Apollo use
  - https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/

Naming standards – why?

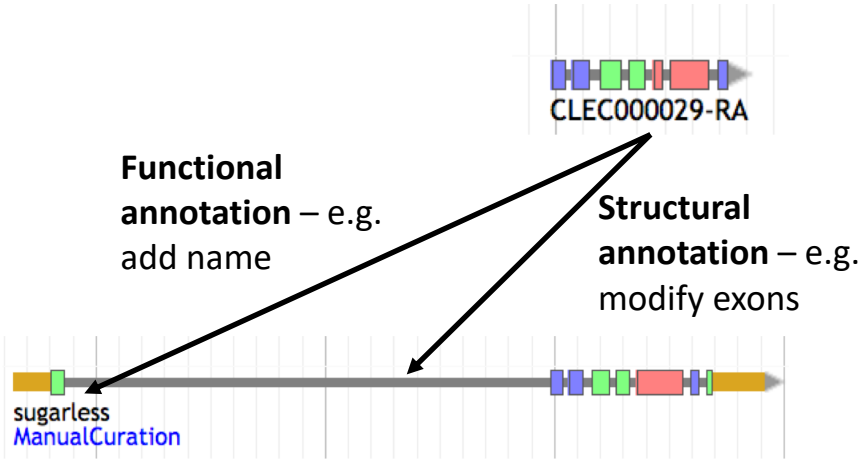
- Name carries important information about protein or gene function
- Name will often be propagated to other species – needs to make sense in their context, as well
- Helps to improve consistency across taxa/databases



Definitions

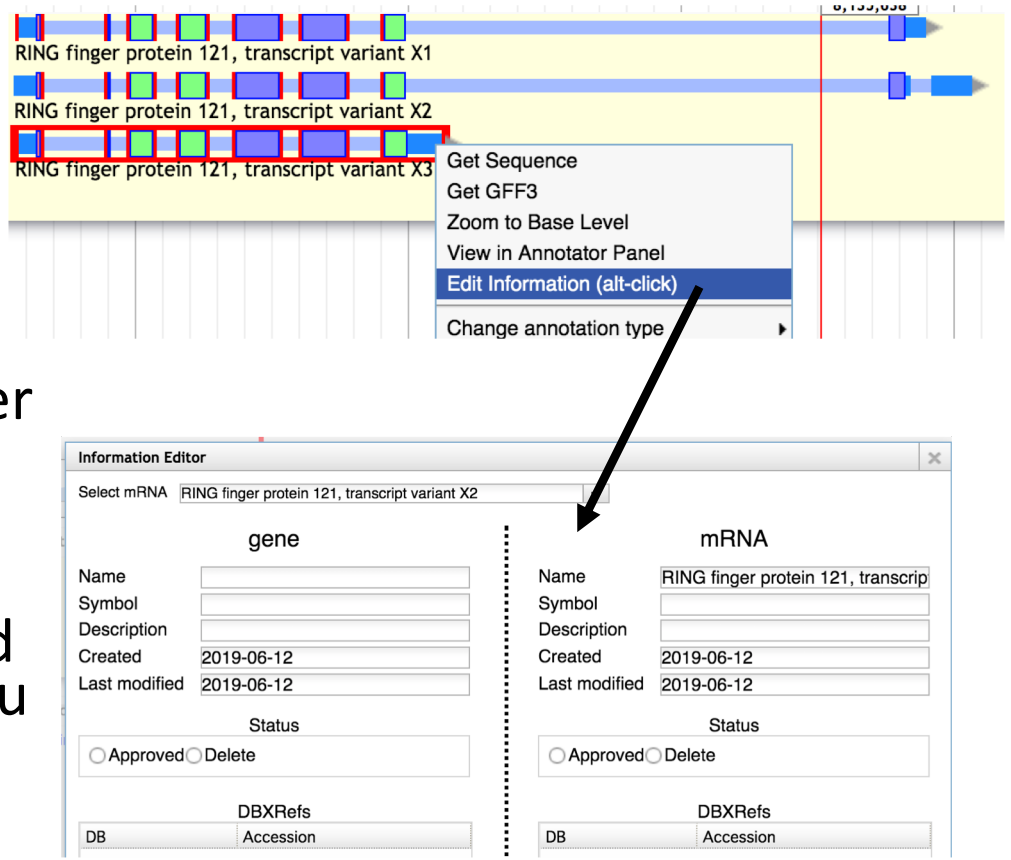
Structural vs. functional annotation

- Annotations describe the structure and function of genes in a genome.
- **Structural**: describe the structure of the gene
- **Functional**: describe the function of a gene
- **Manual**: Update gene structure and function



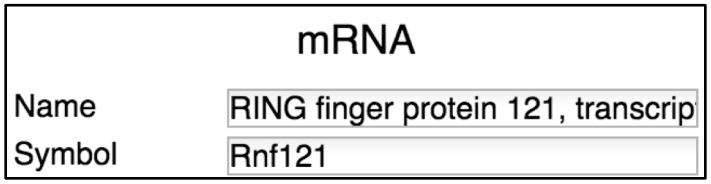
Genes vs. proteins

- Names can be applied to both genes and proteins
- Whether you name the gene or the protein depends on the research community
- At the i5k Workspace, enter naming information in the Apollo mRNA panels (equivalent to protein)
- You can add both gene and protein information – if you are pressed for time, just do protein.



Names vs. symbols

- Name:
  - Describes the function of a gene or protein
  - “A good protein name is one which is unique, unambiguous, can be attributed to orthologs from other species” (IPNG)
  - Should not describe a phenotype, anatomical features, or taxon-specific characteristics
- Symbol:
  - a short form of the name
  - We don’t recommend coining new symbols – okay to adopt existing ones, though



Descriptions vs. Comments

- Descriptions:
  - Use this field if you have a longer description of the protein
  - Will show up as a Note in NCBI
  - E.g. “Putative Phosphoenolpyruvate Carboxykinase”
- Comments:
  - Used for general comments on the annotation process, or caveats on the annotation
  - We keep the comments at the i5k Workspace, but these don’t make it into NCBI
  - E.g. “Added Name based on 89% blastp similarity with XP_123571.2”

I5k Workspace Guidelines

I5k Workspace Guidelines - Names

Are you adopting a name from a homolog?

- You can re-use existing, established names (e.g. from *Drosophila melanogaster*)
- Don't add a species prefix (although okay to use in your manuscript for clarity)
- If you want to imply uncertainty, you can append '-like' to the name
- Good: "Ultraspiracle" 
- Okay: "Ultraspiracle-like"
- Bad: "Clec-ultraspiracle" or "similar to ultraspiracle" 

https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines

USDA

I5k Workspace Guidelines - Names

- Are you naming an isoform?
  - use the suffix “isoform A”, “isoform B”, etc.
- Are you naming a fragmented gene?
  - include a comment 'Part X of Y', where Y is the total number of fragments, and X is the ordinal number for that gene.
  - Don’t add ‘partial’ or ‘part of’ to the name.

https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines

I5k Workspace Guidelines - Names

- Are you naming a ‘new’ gene?
  - Choose a name that could be propagated to all orthologous proteins; try not to make it species- or tissue-specific
    - **Good: “magnesium transporter”** 
    - Bad: “diapause-associated protein” 
- Are you naming a gene from a gene family?
  - Check if a naming system already exists: http://www.uniprot.org/docs/nomlist.txt
  - Use Arabic numbers to specify the different members encoded by a multigene family.

https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines

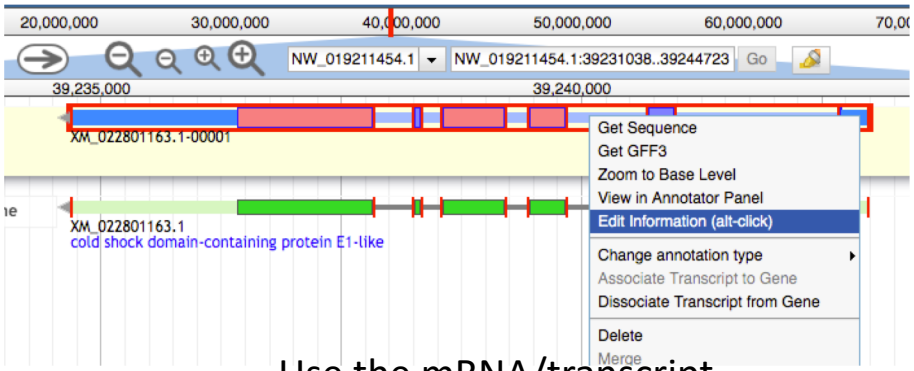
I5k Workspace Guidelines - Symbols

- Are abbreviations of the descriptive gene name.
- We do not recommend coining new symbols for newly named genes.
- However, if a name from an orthologous gene was adopted, you may use this gene's symbol, as well.
- Don't use species prefixes (e.g. Clec-Pepck)
- Examples: Pepck, Ser12

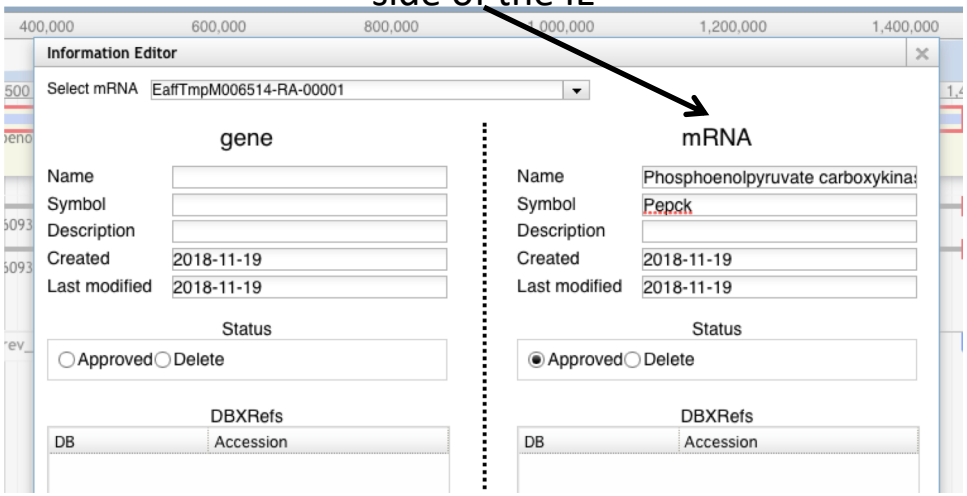
https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines

Using the Information Editor at the i5k Workspace

- Select the model in Apollo, then right-click, and select ‘Edit Information’ from the drop-down menu
- Use the ‘mRNA’ section for name and symbol
- Comments – Document what changes you performed, and your justification for the name (e.g. “Name based on 88% sequence similarity via blastp to D. melanogaster pepck P20007”)



Use the mRNA/transcript side of the IE



What happens to my annotation when I'm done?

- This depends on the genome project that you're working on.
- If the genome coordinator has asked us to generate an OGS (Official Gene Set), we will do so
- We are working on a pipeline to submit Official Gene Sets to GenBank, where they will be archived/accessioned
- Otherwise, don't assume that your annotation will be archived.
  - If you need it to be, get in touch with us and we'll figure out what to do.
- Get in touch with us and the genome project coordinator if you're not sure about the status of a genome project.

Other naming resources

- I5k Workspace: https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines
- AphidBase: https://bipaa.genouest.org/is/how-to-annotate-a-genome/
- VectorBase: https://www.vectorbase.org/content/gene-metadata-form
- HGD: http://hymenopteragenome.org/
- NCBI: https://www.ncbi.nlm.nih.gov/genome/doc/internatprot_nomenguide/

Thank you!

- The NAL Team
  - Chris Childers
  - Min-Chen Hsu
  - Chun-Hung Lin
  - Chia-Tung Wu
- i5k Coordinating Committee
- i5k Pilot Project
- Apollo & JBrowse Development Teams
- GMOD/Tripal community
- AgBioData
- All of our users and contributors!

Contact us:

- https://i5k.nal.usda.gov/contact
- i5k@ars.usda.gov
- Monica.Poelchau@ars.usda.gov
- Christopher.Childers@ars.usda.gov

USDA