Recommendations for managing your genome project data

I5k Workspace@NAL webinar series
July 3rd, 2018
Outline

• Genome project life cycles
• Data management:
  • What it is
  • Why it’s important
• Data management components in the genome project life cycle
• Data management resources at the i5k Workspace@NAL
Resources

- CSU’s guidelines on data management: [https://lib.colostate.edu/services/data-management/](https://lib.colostate.edu/services/data-management/)
- The Open Science Framework: [https://osf.io/](https://osf.io/)
- Software to format assembly and annotations for NCBI submission: [http://genomeannotation.github.io/GAG/](http://genomeannotation.github.io/GAG/)
- Fort Lauderdale agreement: [https://www.genome.gov/pages/research/wellcomereport0303.pdf](https://www.genome.gov/pages/research/wellcomereport0303.pdf)
- Toronto agreement: [https://dx.doi.org/10.1038%2F461168a](https://dx.doi.org/10.1038%2F461168a)

- Software for QC and merging of manual annotations: [https://github.com/NAL-i5K/GFF3toolkit](https://github.com/NAL-i5K/GFF3toolkit)
- Other perspective on the genome project lifecycle: [https://dx.doi.org/10.12688%2Ff1000research.7559.1](https://dx.doi.org/10.12688%2Ff1000research.7559.1)
- A non-exhaustive list of arthropod genome databases: [http://i5k.github.io/share](http://i5k.github.io/share)
  - Any arthropod: [i5k Workspace@NAL](http://i5k.github.io/share)
  - Hymenoptera: [Hymenoptera Genome Database](http://i5k.github.io/share)
  - Ants: [Fourmidable](http://i5k.github.io/share)
  - Insect vectors of disease: [VectorBase](http://i5k.github.io/share)
  - Aphids: [AphidBase](http://i5k.github.io/share)
  - Lepidoptera: [LepBase](http://i5k.github.io/share)
- I5k Workspace submission information: [https://i5k.nal.usda.gov/data-submission-overview](https://i5k.nal.usda.gov/data-submission-overview)
- The Ag Data Commons: [https://data.nal.usda.gov/](https://data.nal.usda.gov/)
- Resources for genomics methods:
  - Genome Curation Communities site: [http://genomescuration.github.io/](http://genomescuration.github.io/)
  - I5k webinar series: [http://i5k.github.io/webinar](http://i5k.github.io/webinar)
Genome Project Data management

• Genome projects have a life cycle
• Here, I will cover **when** and **how** to manage your (arthropod) genome data
• Take-home message: manage your data during the genome project, not at the end

[Diagram of the genome project lifecycle with stages including Design, Sequencing, Genome Assembly, Structural and Functional Annotation, Community Curation, Publication, and Data management.]

cf. [https://dx.doi.org/10.12688%2Ff1000research.7559.1](https://dx.doi.org/10.12688%2Ff1000research.7559.1)
What is data management, and why is it important?

• What it is:
  • Best practices for creating, organizing, storing, and sharing your data products

• Why do it:
  • You have to. (Usually - compliance with funding agencies).
  • Increases the impact of your research
  • Improves reproducibility

https://lib.colostate.edu/services/data-management/
Data Management Plan Components

1. **What** data types?
2. **Where** should the data be deposited?
3. **When** should the data be deposited and released?
4. What kind of access should the data have?
5. Who is responsible?

0. Experimental Design

- **What**: Usually no data involved at this point

- **Where** (optional): the Open Science Framework ([https://osf.io/](https://osf.io/))
  - OSF allows you to manage the organization and content of your research project
  - Can also share with other researchers

- **When** (optional): at the start of the project
1. DNA/RNA Sequencing – data management

- **What**: File formats can include: fastq, fasta, sra
- **Where**: NCBI or other INSDC repository
- **When** to deposit: ASAP
1. DNA/RNA Sequencing – typical problems

1. Files are big – need to have a decent internet connection to upload to a repository

2. NCBI’s database structure – BioProject, BioSample, Experiment, can initially be confusing to navigate

3. Metadata – how much to submit?
1. DNA/RNA Sequencing - advice

1. CyVerse can help with submitting to SRA:

2. BioSample metadata – you’ll probably want to choose the Invertebrate or “Genome, metagenome or marker sequences (MIxS compliant)” packages
2. Assembly – data management

- **What**: fasta, agp files
- **Where**: NCBI/INSDC, *domain-specific repository*
- **When to deposit at NCBI**: As soon as it is ‘stable’. Don’t wait for all downstream analyses to be completed first
- **When to deposit at i5k Workspace**: once your assembly and annotations are stable and ready for the public
2. Assembly - problems

1. NCBI performs QC on your assembly, and you will probably have to perform corrections based on the QC results
   1. Because of this, a successful submission can take some time
   2. If you don’t have good command-line skills, fixing the genome assembly can be difficult
   3. There is an alpha release of NCBI’s contamination screen on GitHub if you’d like to try it out for yourself: https://github.com/NCBI-Hackathons/ContamFilter
2. Assembly - advice

1. Why submit the genome assembly to NCBI early on during your genome project?
   1. Most publications require submission of the genome assembly
   2. NCBI’s QC can improve the quality of your assembly
   3. Stable accession numbers for your sequences allow for better reproducibility and less confusion in your manuscript
   4. If you perform downstream analyses on your pre-NCBI assembly, you may have to re-analyze if the NCBI QC requires a lot of changes
   5. If of sufficient quality, and if RNA-Seq from the same species is available, NCBI can annotate the assembly for you
3. Structural and functional annotation – data management

- **What**: gff3, fasta, tbl, gtf
- **Where**: *domain-specific repository, NCBI/INSDC*
- **When to deposit at NCBI**: 
  - Along with genome assembly 
  - When the annotations are stable
- **When to deposit at i5k**
  - **Workspace**: once your assembly and annotations are stable and ready for the public
3. Structural and functional annotation – problems and advice

• Submission to NCBI requires some reformatting and QC
• This is manageable if your annotations are automated
  • Software to format assembly and annotations for NCBI submission: [http://genomeannotation.github.io/GAG/](http://genomeannotation.github.io/GAG/)
• NCBI will perform structural and some functional annotation for your genome if the assembly quality is sufficient
  • Your annotations will automatically deposited at NCBI when they’re done – no data management necessary
4. Community curation

• What is community curation?
  • Scientists collectively examine and improve gene models (usually computationally predicted)

• Community curation at the i5k Workspace:
  • Via the Apollo software
  • Access to a large community of curators
  • Tutorials, guidelines, webinars
  • Registration mechanism for new annotators
  • One-on-one support
  • Over 400 registered annotators have curated over 10,000 gene models using the Apollo software
4. Community curation – data management

- **What**: gff3, fasta
  - NOT word documents!!!
- **Where**: *domain-specific repository, NCBI/INSDC*
- **When to deposit at NCBI**: *Once community curation effort is complete*
4. Community curation – data management

• Submission to NCBI can require substantial reformatting and QC
• This is quite difficult if your annotations are manual
  • Non-standard formatting of functional annotations make the submission difficult
  • The i5k Workspace is working on tools for this
5. Publication

- Most of the data that you analyze in your paper should already be submitted in a repository – cite it accordingly!
  - SRA accession numbers
  - NCBI Genome assembly accession number
  - GenBank annotation accession number, if available
- Gene names – if you’re referring to a gene in your publication, it’s best to also refer to its stable identifier
5. Publication

• Other data files – if there’s an appropriate repository, deposit them there, and not in the supplemental data as a pdf.
  • Supplemental files are usually not machine-readable
  • Therefore, they are not easy to find and analyze
  • Repositories make it much easier to share your data with others later on when you’ve forgotten the experimental details
What kind of access should the data have?

• Are you a federal employee?
  • Data generated by federal employees has either US Public Domain or Creative Commons Zero status

• Otherwise:
  • federally-funded data and non-federal data may vary depending on funder requirements

• See https://creativecommons.org/ for more explanations about licenses
Who should be in charge of data submission?

• Someone with access to the relevant metadata (usually someone involved in the study design)
• Someone with sufficient time and patience for submission
• Ultimately, if you’re the PI, you are responsible for successful management of the research data from your grant or project
When should the data be released (as opposed to deposited)?

• This depends.
  • Most of the problems that we’ve seen so far are due to communication breakdown (in part because large consortia are involved)

• In general, we encourage:
  • The early release of all genome data to speed along scientific discovery
  • Use of the Fort Lauderdale and Toronto agreements to communicate your intent to publish
  • Genome project leads to communicate with their group/consortium on publication plans
Domain-specific repositories

• A non-exhaustive list: http://i5k.github.io/share
  • *Any arthropod: i5k Workspace@NAL
  • Hymenoptera: Hymenoptera Genome Database
  • Ants: Fourmidable
  • Insect vectors of disease: VectorBase
  • Aphids: AphidBase
  • Lepidoptera: LepBase

• Provide value-added curation services and tools to make clade-specific data easier to find and use

• Not a replacement for NCBI submission
The i5k Workspace@NAL

- The i5k initiative tasked itself with coordinating the sequencing and assembly of 5000 insect or related arthropod genomes
- International effort to prioritize insect genomes for sequencing; provide guidelines for genome sequencing and curation; and seek funding.
- The i5k Workspace@NAL is available to help any i5k (arthropod) project with genome hosting needs
Why join the i5k Workspace?

• Gain access to a large diverse community
  • A diversity of organisms
    • 64 species and counting
  • Large user community with many different interests
    • People versed in the biology of specific systems
    • Experts in a species or group of species
• A common interface for accessing data, tools and search
• Curation tools to improve annotation quality – in particular for community curation
• Help with data management
I5k Workspace Project Basics

• The i5k Workspace centers around *projects*.
  • A project is a collection of data based on the genome assembly of an arthropod
  • All data is used in the context of the genome assembly

• Each project has a *project coordinator*.
  • Serves as the point of contact for questions about the project
  • Main responsibility: approve or reject new Apollo users

• All of our data is user-submitted
What do we need for a project?

• Your project metadata
  • Information about your organism
  • Metadata for submitted data files (the more the better)
    • What tools or methods were used
    • Software versions and options set
    • When and where the data were generated
    • Other information (location collected, life-stage, etc.)

• Your data files
  • Genome assembly needs to be in GenBank/ENA/DDBJ
  • Data should be open access (no private repositories)
  • Additional datasets need to be mapped to the same assembly
What do we do with your data?

• Create resources
  • Organism and gene pages
  • Data downloads

• Integrate your data with our tools
  • Genome browser
  • BLAST, Clustal, HMMer
  • Apollo for gene curation

• Offer post curation services
  • Annotation QC and Official Gene Set (OGS) Creation
  • In progress – re-map OGS to updated assembly
Submission

Workspace@NAL
https://i5k.nal.usda.gov/

Resources
- Organism Information Page
- Bulk data downloads
- Tutorials

Tools
- Custom BLAST interface
- JBrowse genome browser
- Apollo manual curation tool
- HMMer
- Clustal

Services
- Manual annotation quality control
- Official gene set generation

Challenges
- Non-standard data formatting
- Failure to submit all metadata (e.g., sample origin; analysis methods)

- ‘Frozen’ genome assembly
- Automated annotations
- Ancillary datafiles (e.g., RNA-Seq alignments)
What don’t we do with your data?

• Computationally intense analyses such as
  • Gene prediction
  • Raw RNAseq mapping

• We are not a long-term archive or repository
  • NCBI
  • Ag Data Commons
  • Dryad Digital Repository
  • CyVerse Data commons
  • Many other options available
Considerations before submitting

• You need to have an *arthropod* genome assembly, accessioned by NCBI (or another INSDC member)

• *All data submitted to the i5k Workspace is public.*
  • However, we do state whether Ft. Lauderdale/Toronto agreements of data sharing should apply

• Is your genome an ‘orphan’, or is there another suitable database?
  • We can host genomes that are already hosted elsewhere, and actively communicate with other database providers
  • All manual annotation efforts need to be at one database
Submitting data to the i5k Workspace: 1. Register for an account

- Apply for a dataset submission account: https://i5k.nal.usda.gov/register/project-dataset/account

- Once your account is approved, you can submit projects, assemblies or other datasets
Submitting data to the i5k Workspace: 2. Start a Project

• Log in
  • [https://i5k.nal.usda.gov/user](https://i5k.nal.usda.gov/user)

• From menu, select ‘Data -> Submit data -> Request a new i5k Workspace Project’
  • [https://i5k.nal.usda.gov/datasets/request-project](https://i5k.nal.usda.gov/datasets/request-project)

• We’ll review your submission and will get in touch with you
Submitting data to the i5k Workspace: 3. Submit your data

- All information submitted through this form will be re-formatted for display at the i5k Workspace (except for email address and file checksum)

- From menu, select ‘Data -> Submit data -> Submit a dataset’
  - https://i5k.nal.usda.gov/datasets/submit-a-dataset
Other resources at the NAL: the Ag Data Commons

- Hosts any dataset funded by the USDA
- Landing page
- Citable DOI
- 36 i5k datasets already available
Thank you!

The NAL Team
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Gary Moore
- Susan McCarthy

i5k Workspace alumni
- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Yu-yu Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

- i5k Coordinating Committee
- i5k Pilot Project
- 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
- Monica.Poelchau@ars.usda.gov
- Christopher.Childers@ars.usda.gov