The AgBioData Journey Towards Best Practices Of Data Sharing And Management In Agricultural Research And Education.

Annarita Marrano
amarrano@phoenixbioinformatics.org

The Plant Biology 2024 Conference
June 22, 2024
Honolulu, HI
OUTLINE

- Who we are
- What we do to enhance FAIR data management
- Our recommendations and resources
- Why and how you should join our efforts
THE AGBIODATA CONSORTIUM

- Founded in 2015
- 44 Genetic, Genomics, and Breeding (GGB) resources
- Over 250 members
- Mission: ensure standards and best practices for the acquisition, display, and retrieval of GGB data
"Research data generated with federal funding are **publicly and equitably accessible**" (the Nelson Memo, the Office of Science and Technology Policy – OSTP; 2022)
Aim 1: Recommendations, standards, and implementation plans for FAIR data.

Aim 2: Expand the network to include key stakeholders.

Aim 3: Educational and training materials for researchers.

Aim 4: Roadmap for a sustainable GGB data/database ecosystem.

Award Abstract # 2126334
HOW CAN WE MAKE OUR DATA FAIR?

- Genotype-to-Phenotype (G2P)
- Standards for Genetic Variation (SGV)
- Data Reuse (DR)
- Genome Assembly and Annotation Nomenclature (GAAN)
- Pan-genomes

Recommendations from our working groups (WG)

- Genomic, Genotypic, & Phenotypic Variation
- Genome Assemblies and gene models
- Pan-genomic resources
GENOMIC, GENOTYPIC, & PHENOTYPIC VARIATION

- Submit the data to appropriate data repositories
  - The G2P white paper provides a detailed list of database resources per data-types in plant science.

- Adopt community-based data format and ontologies, if available
  - Variant Call Format (VCF) file for genotypic datasets
  - A new AgBioData WG on phenotypic data standardization and management
  - Seek for help from the community-databases!

- Implement data quality checks before sharing your data

- Submit complete meta-information
  - Used Code, protocols and analysis workflows, etc.

From Deng et al. (2023) https://doi.org/10.1093/database/baad088

https://doi.org/10.12688/f1000research.109080.2
GENOME ASSEMBLIES AND GENE MODELS

• Different labs sequence the genome of the same individual

• Genomic labs continuously generate new versions of the same individual genome assembly and annotation

• Difficulties in
  • Tracking the different versions of a genome assembly and annotation
  • Linking gene models to annotation analyses and assemblies

Identify a nomenclature system that generate STANDARDIZED ASSEMBLY and GENE MODEL NAMES that are both human and machine-readable.
GENOME ASSEMBLIES AND GENE MODELS

Genome assembly identifier

<ToLID>.<sample_identifier>.<consortium>.<assembly_version>.<optional>.fasta

Accession/vety/landrace/breed

Species name
as provided by the Tree of Life project
https://id.tol.sanger.ac.uk/search

Or Project/Group assembling

e.g., haplotype for phased assembly
GENOME ASSEMBLIES AND GENE MODELS

Gene model identifier

<assembly_prefix><annotation_version><chromosome><entity><6-digit ID number><optional>

- sub-genome and chromosome for polyploid genomes
- Haplotype if phased assemblies
- Transcript isoforms for multi-exon genes

https://github.com/AgBioData/Genome-Assembly-and-Annotation-Nomenclature_WG
PAN-GENOMES

- Pan-genome terminology and use
- Analysis software and pipelines
- Visualizing tools

https://github.com/AgBioData/Pan-genomes/blob/main/Pan-genome_resource.md
HOW CAN WE TEACH FAIR DATA MANAGEMENT?

AgBioData curriculum for FAIR Ag Science

**FOCUS:** FAIR data principles within the context of GGB agricultural research

**AUDIENCE:** educators of undergraduate and graduate students

**MODALITIES:** 1. in-person and interactive; 2. self-directed and asynchronous

Open access, easy to use, customizable and remixable

Generated by the Education WG
AgBioData curriculum for FAIR Ag Science

1. What is a biological digital repository?
2. FAIR and databases
3. Bio-databases: types of data, finding and obtaining data
4. Creating and sharing trustworthy data
5. Submitting data to a database
6. How to use your library resources
7. Databases for agriculture

Slides and recording will be accessible at
CURRENT WGs

- Education
- FAIR Scientific Literature
- Phenotypic Data Standardization and Management
- scRNAseq Biocuration
- Standards for Genetic Variation Data
- Sustainability

https://www.agbiodata.org/current-working-groups

Booth # 406
HOW TO PARTICIPATE IN AGBIODATA

• Interested in our activities and working groups? Send an email to agbiodata@gmail.com!
• Become a member! Visit our website www.agbiodata.org
• Join our Slack workplace!
• @AgBioData
• Follow us on LinkedIn
• Monthly meetings/webinars (1st Wed of the month)
• If you have a GGB resource, join the consortium!
ACKNOWLEDGEMENTS

AgBioData SC members:
Carson Ardson
Sarah Dyer
Sunita Kumari
Dorrie Main
John P. McNamara
Sushma Naithani
Monica Poelchau
Leonore Reiser
Peter Selby
Meg Staton

Past AgBioData SC members:
Jacqueline Campbell
Ethy Cannon
Laurel Cooper
Peter Harrison
Lisa Harper
Eva Huala
Sook Jung
Marcela Tello-Ruiz

Past PC:
Darwin Campbell

The AgBioData Working Groups