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

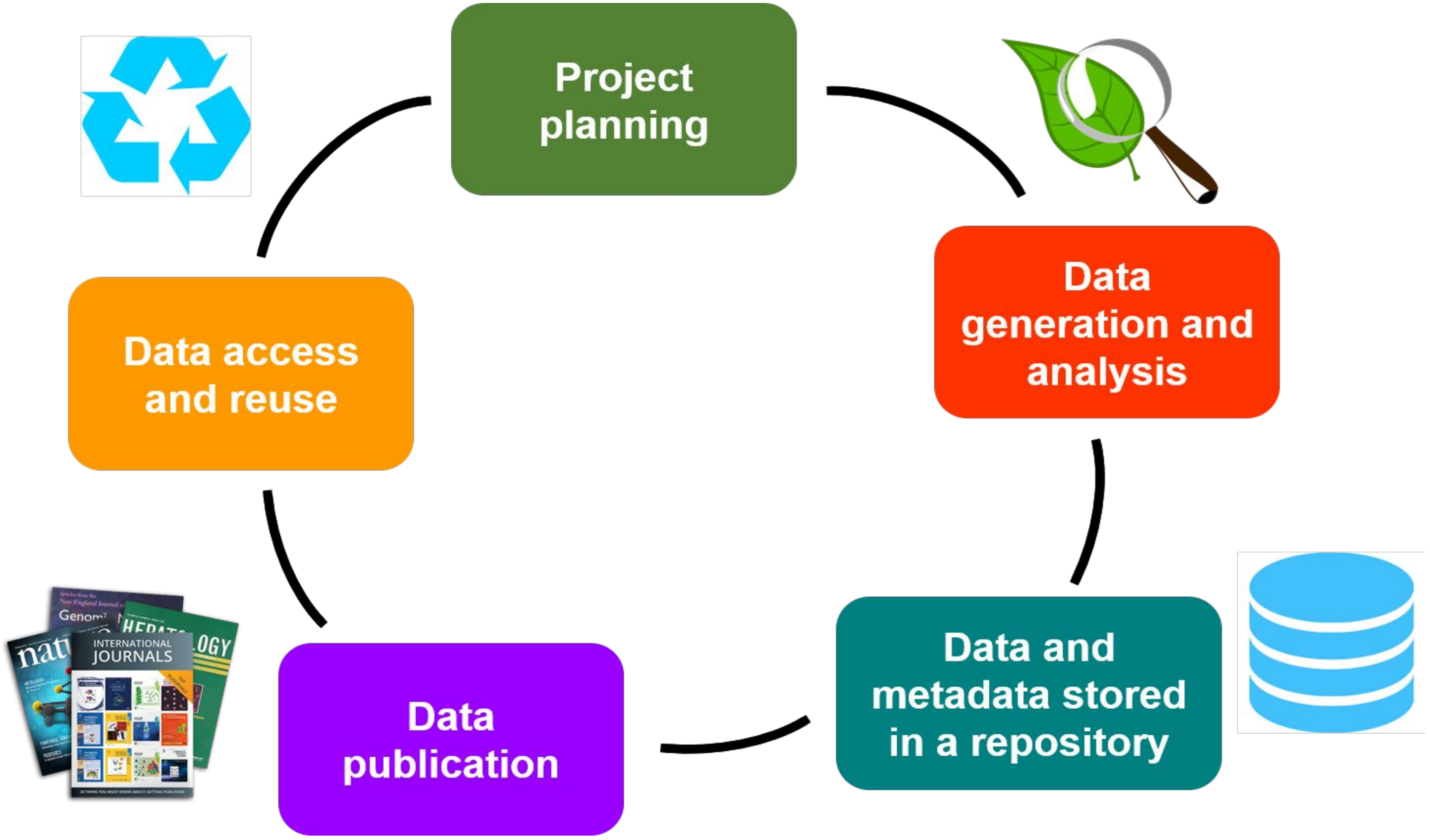
OUTLINE

- 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*

PeanutBase
Genetic and genomic data to enable m
TreeGenes
A Forest Tree Genome Database
Planteome
Ontologies for Plant Biology
i5k
maizeGDB
MAIZE GENETICS AND GENOMICS DATABASE
CuGenDB
µP
Gramene
AgBase
Enterprise Breeding System
Sweet GAINS
AnimalQTLdb
Alfalfa Breeder's Toolbox
AgroPortal
LIRMM
CYVERSE
tair
CITRUS GENOME DATABASE
COTTONGEN
COTTON DATABASE RESOURCES
KitBase
GENOME DATABASE FOR ROSACEAE
BAR
GOBii
USDA
Hymenoptera Genome Database
AfricaYam
FAANG
Functional Annotation of Animal Genomes
PMN
Plant Metabolic Network
GrainGenes
A Database for Triticeae and Avena
Crop Ontology
for agricultural data
USDA ARS-GRIN
WheatIS
e!Ensembl



Project planning

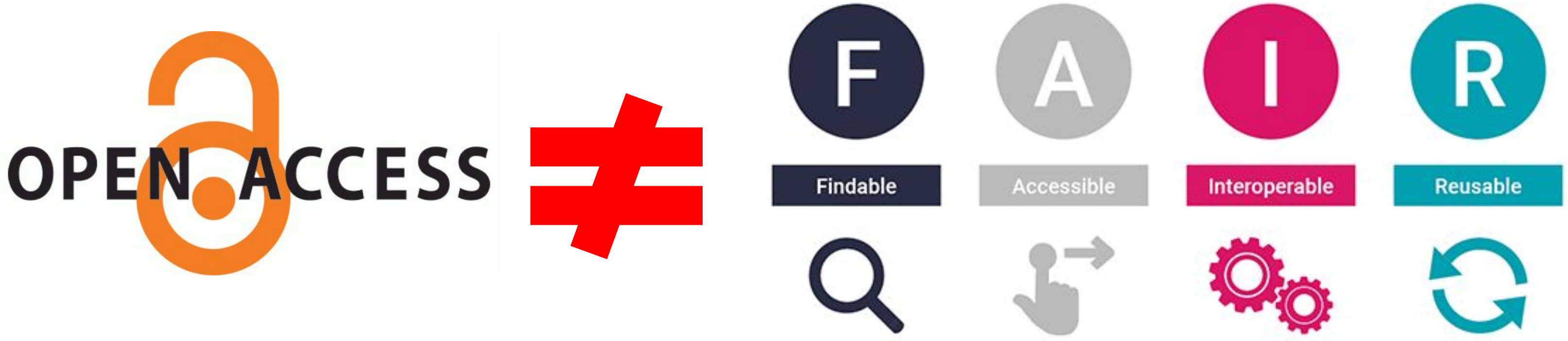
Data generation and analysis

Data and metadata stored in a repository

Data publication

Data access and reuse

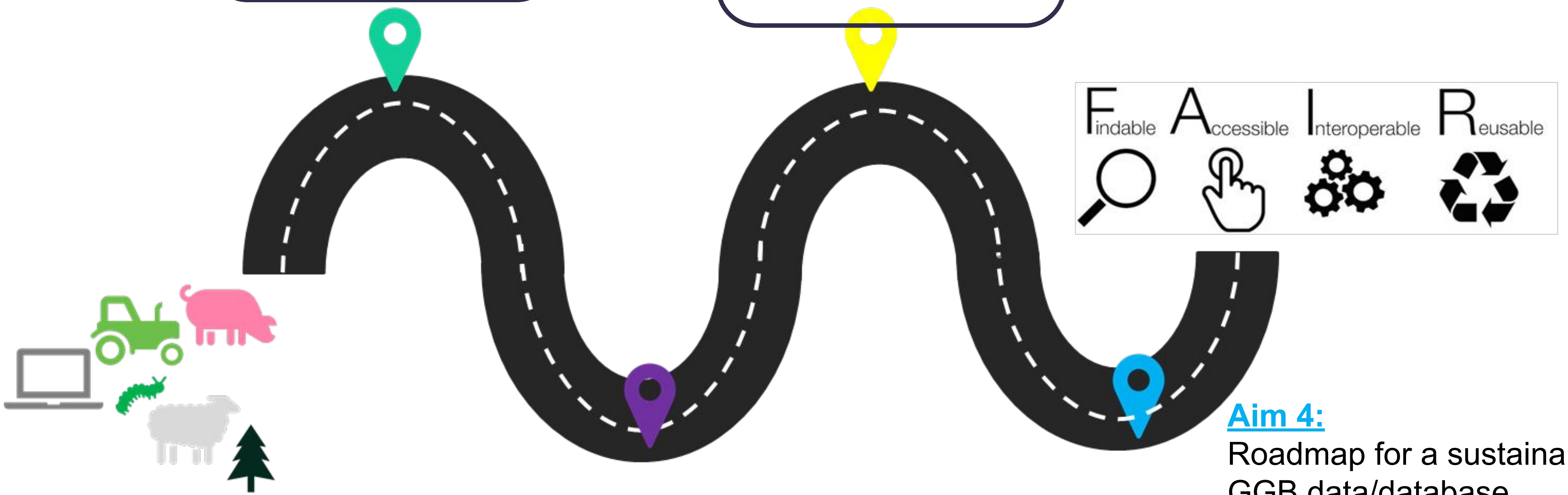
”Research data generated with federal funding are **publicly and equitably accessible**” (the Nelson Memo, the Office of Science and Technology Policy – OSTP; 2022)



Award Abstract
# 2126334

Aim 1:
Recommendations, standards, and implementation plans for FAIR data.

Aim 3:
Educational and training materials for researchers.



Aim 2:
Expand the network to include key stakeholders.

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

HOW CAN WE MAKE OUR DATA FAIR?

Recommendations from
our working groups (WG)

- Genotype-to-Phenotype (G2P)
- Standards for Genetic Variation (SGV)
- Data Reuse (DR)

Genomic, Genotypic, &
Phenotypic Variation

- Genome Assembly and Annotation Nomenclature (GAAN)

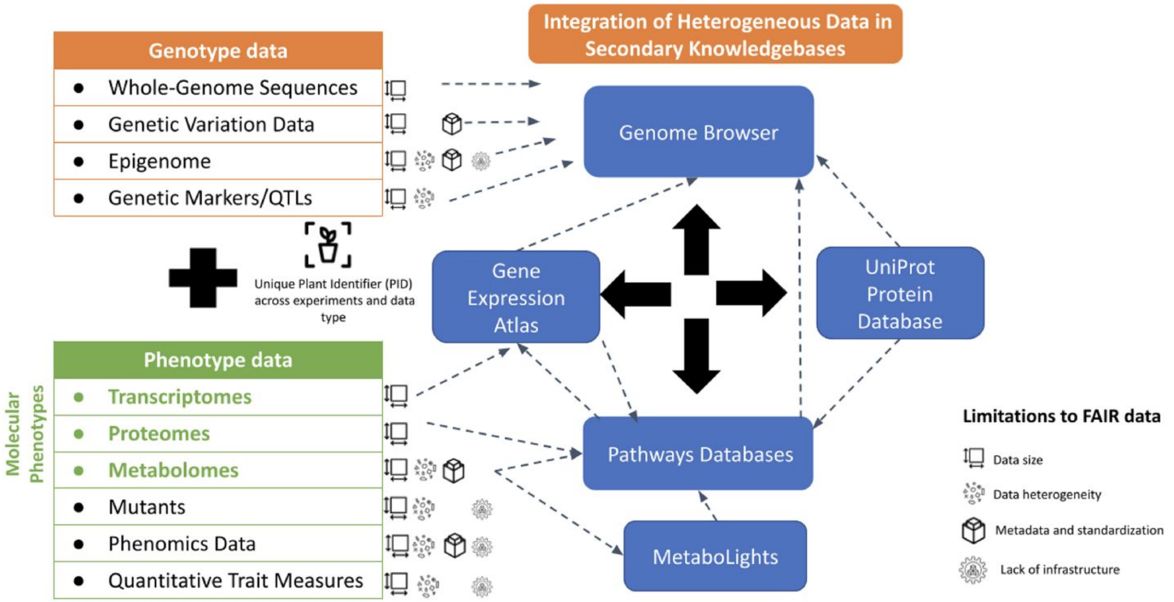
Genome Assemblies and
gene models

- Pan-genomes

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

REVISED Recommendations for the formatting of Variant Call Format (VCF) files to make plant genotyping data FAIR [version 2; peer review: 2 approved]

Sebastian Beier[1,2], Anne Fiebig[1], Cyril Pommier[3], Isuru Liyanage[4], Matthias Lange[1], Paul J. Kersey[5], Stephan Weise[1], Richard Finkers[6,7], Baron Koylass[4], Timothee Cezard[4], Mélanie Courtot[4,8], Bruno Contreras-Moreira[9], Guy Naamati[4], Sarah Dyer[4], Uwe Scholz[1]

https://doi.org/10.12688/f1000research.109080.2

GENOME ASSEMBLIES AND GENE MODELS

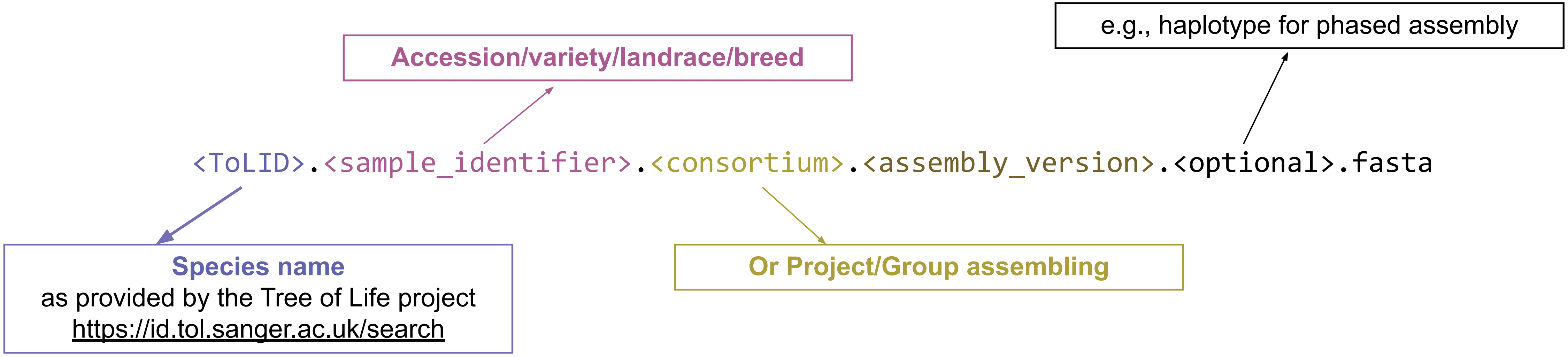
Nomenclature issues

- 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



e.g., haplotype for phased assembly

Accession/variety/landrace/breed

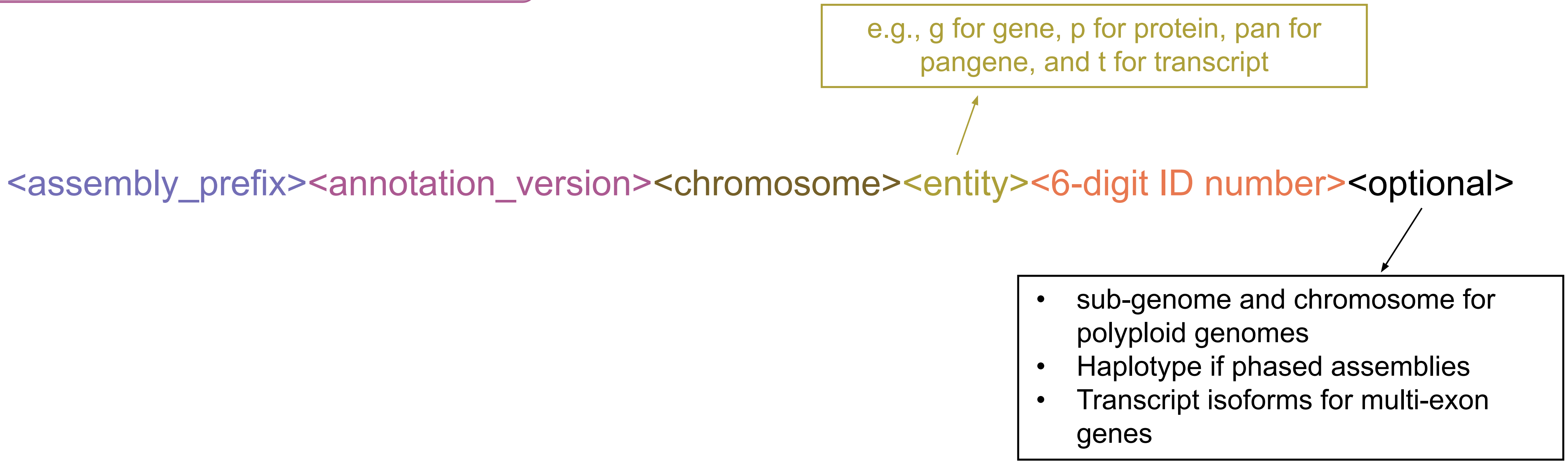
```
<ToLID>.<sample_identifier>.<consortium>.<assembly_version>.<optional>.fasta
```

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

Or Project/Group assembling

GENOME ASSEMBLIES AND GENE MODELS

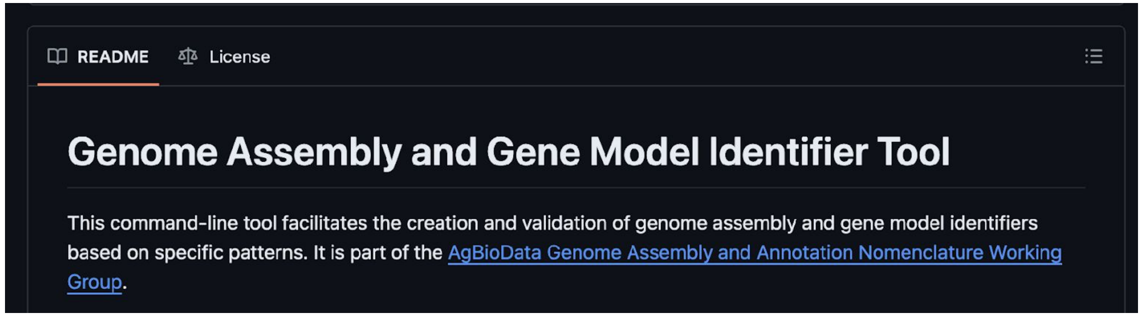
Gene model identifier



e.g., g for gene, p for protein, pan for pangene, and t for transcript

<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



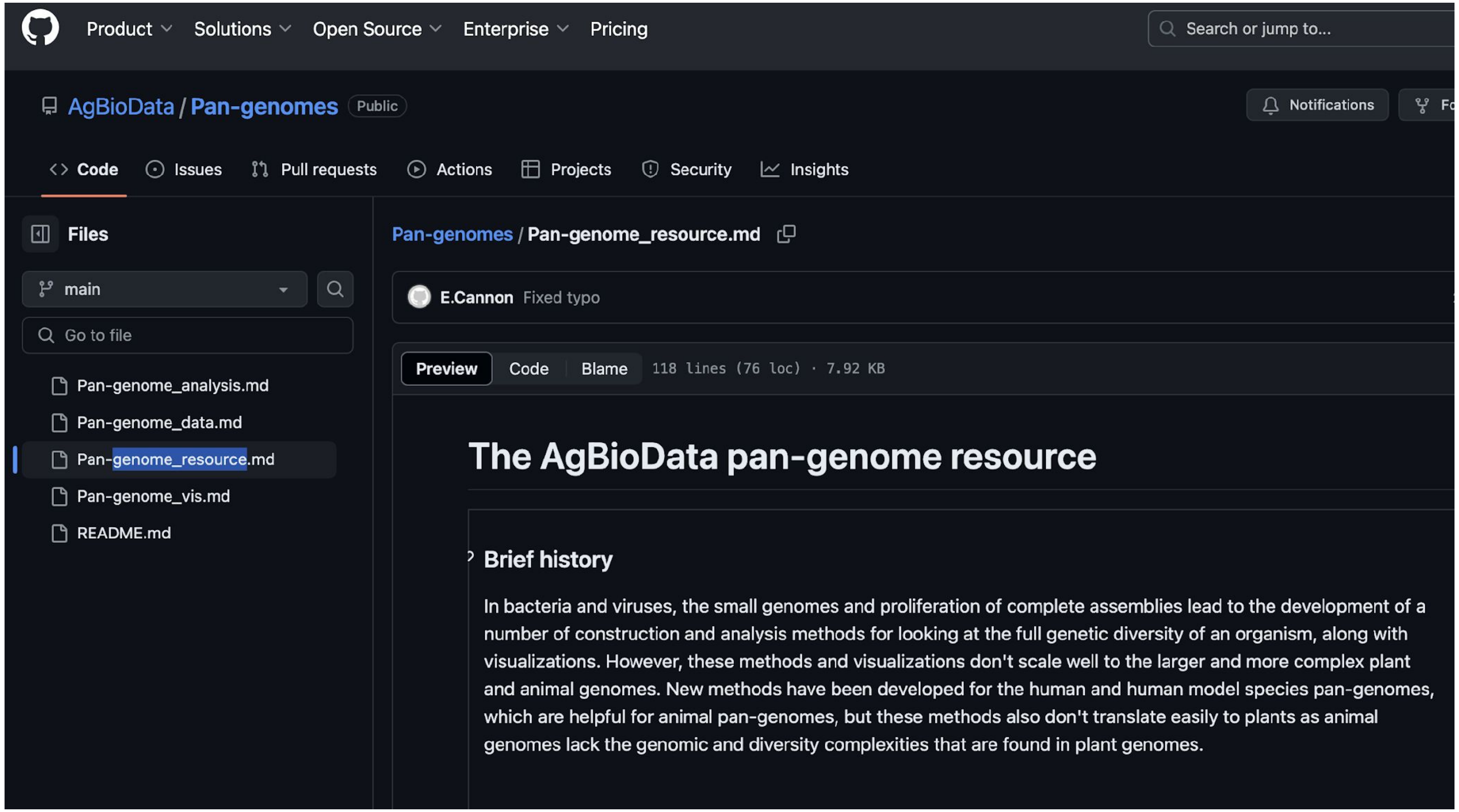
Genome Assembly and Gene Model Identifier Tool

This command-line tool facilitates the creation and validation of genome assembly and gene model identifiers based on specific patterns. It is part of the AgBioData Genome Assembly and Annotation Nomenclature Working Group.

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

PAN-GENOMES

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



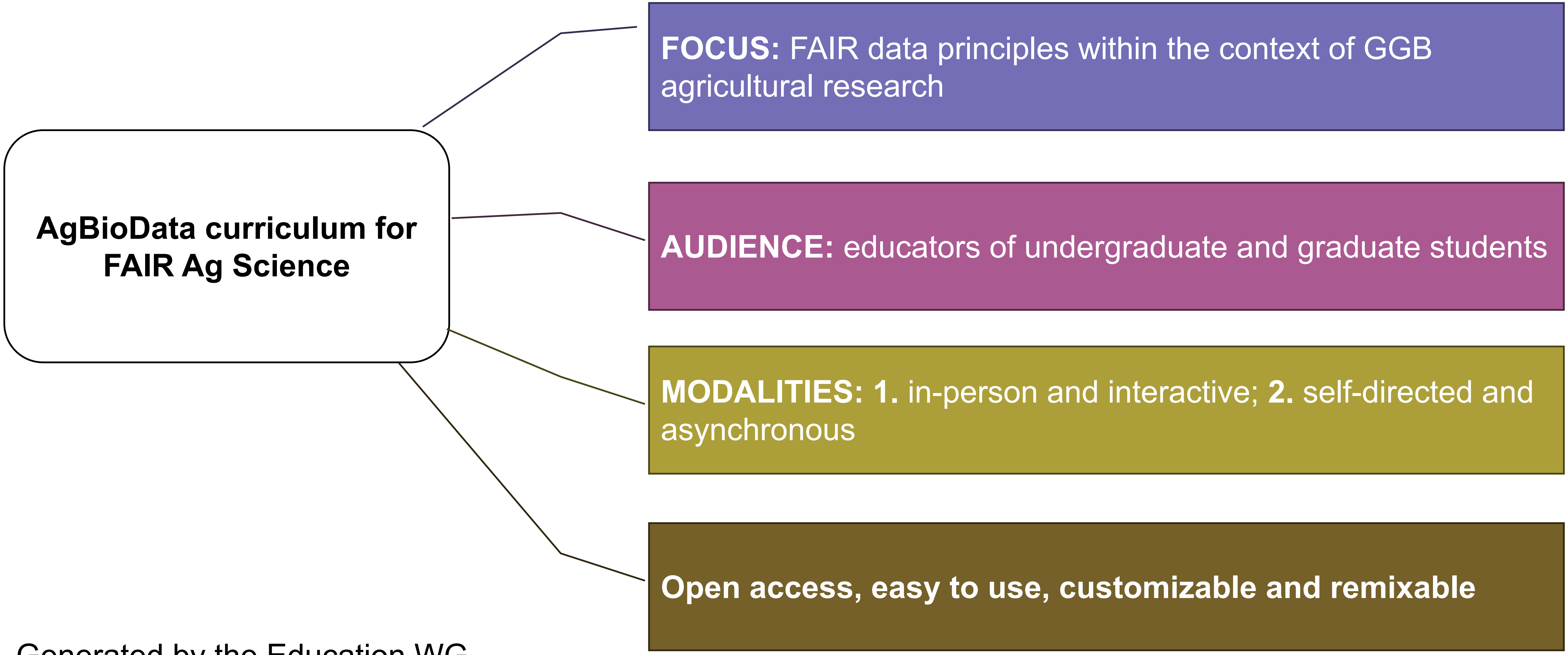
The AgBioData pan-genome resource

Brief history

In bacteria and viruses, the small genomes and proliferation of complete assemblies lead to the development of a number of construction and analysis methods for looking at the full genetic diversity of an organism, along with visualizations. However, these methods and visualizations don't scale well to the larger and more complex plant and animal genomes. New methods have been developed for the human and human model species pan-genomes, which are helpful for animal pan-genomes, but these methods also don't translate easily to plants as animal genomes lack the genomic and diversity complexities that are found in plant genomes.

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

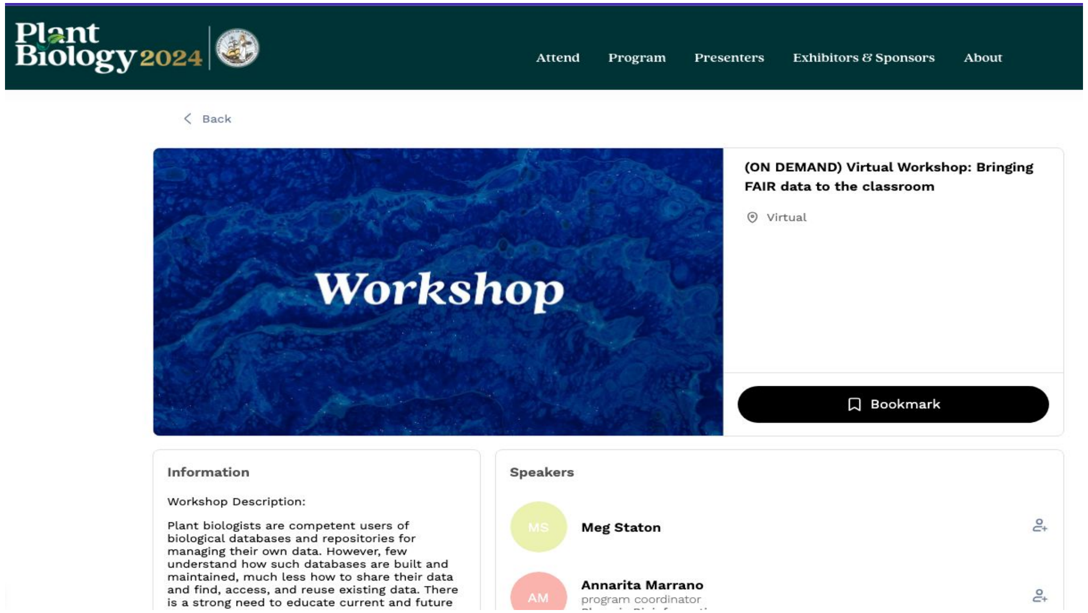
Generated by the Education WG

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

zenodo



(ON DEMAND) Virtual Workshop: Bringing FAIR data to the classroom

Virtual

Information

Workshop Description:

Plant biologists are competent users of biological databases and repositories for managing their own data. However, few understand how such databases are built and maintained, much less how to share their data and find, access, and reuse existing data. There is a strong need to educate current and future

Speakers

Meg Staton

Annarita Marrano
program coordinator

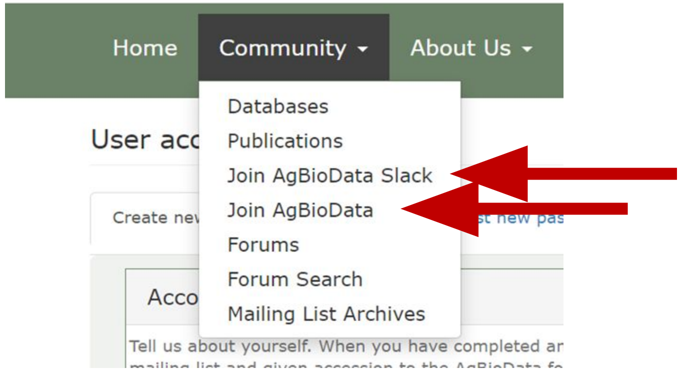
CURRENT WGs

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

Booth
# 406

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

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!**

Booth # 406

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

Award Abstract # 2126334

The AgBioData Working Groups