Guidelines for Gene and Genome Assembly Nomenclature (GAAN)

Advancing Clarity and Utility in Genome Assembly and Gene Annotation Naming

AgBioData

Toward enhanced genomics, genetics, and breeding research outcomes through standardization of practices and protocols across agricultural databases

Genome Assembly Nomenclature Working Group

Presenters: Adam Wright & Dr. David Molik

Date: February 5th 2025

GENETICS

GSA Genetics Society of America

JOURNAL ARTICLE ACCEPTED MANUSCRIPT

Guidelines for Gene and Genome Assembly Nomenclature

Ethalinda K S Cannon, David C Molik, Adam J Wright, Huiting Zhang, Loren Honaas, Kapeel Chougule, Sarah Dyer

Genetics, iyaf006, https://doi.org/10.1093/genetics/iyaf006
Published: 15 January 2025 **Article history**

Abstract

The rapid increase in the number of reference-quality genome assemblies presents significant new opportunities for genomic research. However, the absence of standardized naming conventions for genome assemblies and annotations across datasets creates substantial challenges. Inconsistent naming hinders the identification of correct assemblies, complicates the integration of bioinformatics pipelines, and makes it difficult to link assemblies across multiple resources. To address this, we developed a specification for standardizing the naming of reference genome assemblies, to improve consistency across datasets and facilitate interoperability. This specification was created with FAIR (Findable, Accessible, Interoperable, and Reusable) practices in mind, ensuring that reference assemblies are easier to locate, access, and reuse across research communities. Additionally, it has been designed to comply with primary genomic data repositories, including members of the International Nucleotide Sequence Database Collaboration (INSDC) consortium, ensuring compatibility with widely used databases. While initially tailored to the agricultural genomics community, the specification is adaptable for use across different taxa. Widespread adoption of this standardized nomenclature would streamline assembly management, better enable cross-species analyses, and improve the reproducibility of research. It would also enhance natural language processing applications that depend on consistent reference assembly names in genomic literature, promoting greater integration and automated analysis of genomic data. This is a good time to consider more consistent genomic data nomenclature as many research communities and data resources are now finding themselves juggling multiple datasets from multiple data providers.

10.1093/genetics/iyaf006

The Importance of Nomenclature:

- Clear, consistent naming enhances the usability and reproducibility of genome assemblies and gene annotations.
- A well-structured framework supports collaboration, data sharing, and cross-referencing across repositories and tools.

Objective:

- Introduce the GAAN framework, a standardized approach for genome assembly nomenclature.

Problems Identified:

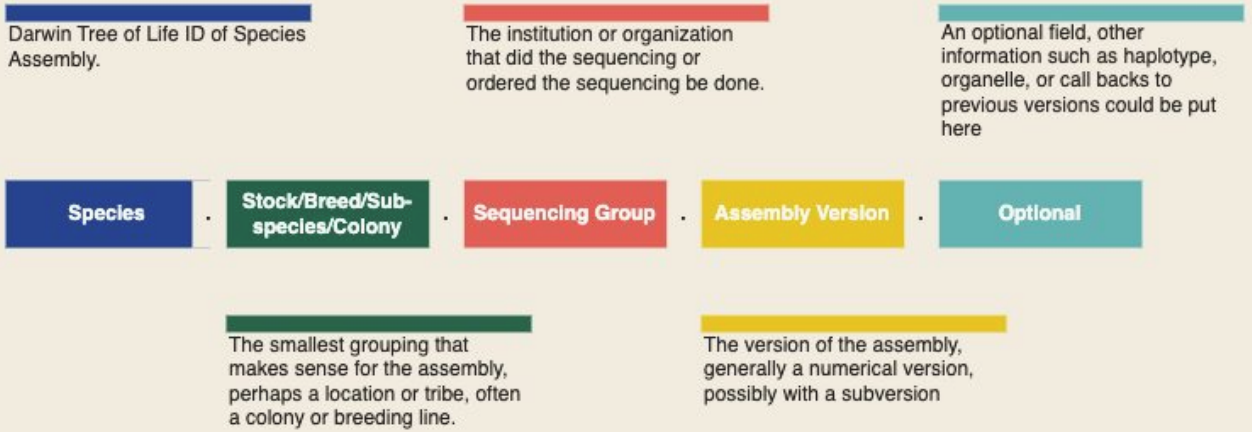
- Lack of consistency across research groups and databases.
- Difficulty in integrating data from different sources due to ambiguous naming.
- Impact on computational tools requiring predictable naming structures.

Core Features of the Framework:

- Incorporates species, sequencing group, colony/breed/strain, version, and other critical metadata.
- Ensures compatibility with standards from AgBioData discussions.
- Aligns with major repositories like International Nucleotide Sequence Database Collaboration (INSDC) for seamless integration.

Detailed Format:

- Naming components include:
  - **Species:** Scientific name or common identifier.
  - **Sequencing Group:** Institution or group responsible for sequencing.
  - **Colony/Breed/Strain:** Specifies population or genetic background.
  - **Version:** Reflects updates or improvements to assemblies.
  - **Optional Metadata:** Unique identifiers for additional context.



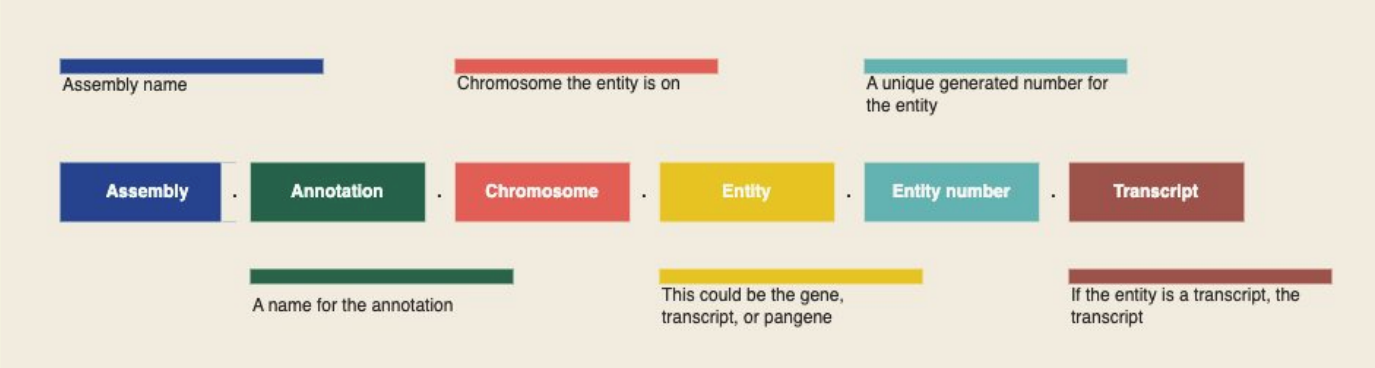
A Computationally legible name

Fields of the name are separated by periods, making it easy for software to denote translate the name into its component parts, a special rule around assembly version and the optional field insures this: if hypothetical name reading software detects six fields, both assembly sub-version and optional field will be assumed, if five fields are detected sub-version will be detected if the last field is only numeric and optional if it contains letters. This is why subversions should always only be numeric.

10.1093/genetics/iyaf006

idCulSono.KS.ABADRU.1.0.female

Species . Colony . Sequencing Group . Version . SubVersion . Misc



10.1093/genetics/iyaf006

Original gene model ID	New assembly ID	New gene model ID
C01p010030.1_BnaDAR	ddBraNapu.DAR.1.0	ddBraNapu.DAR.1.1.01C.p010030
Glyma.01g000100.Wm82.a2.v1	drGlyMax.WM82.2.0	drGlyMax.WM82.2.1.01..g000100
Horvu_BARKE_1H01G000300.1	lpHorVulg.BARKE.1.0	lpHorVulg.BARKE.1.1.01..g000300
TraesCS3D02G273600	lpTriAest.CS.1.0	lpTriAest.CS.1.1.03D.g273600
Vitvi18g12230	drVitVini.PN40024.1.0	drVitVini.PN40024.1.1.18.g012230
Honeycrisp_HAP1_v1.0.031896	drMalDome.Honeycrisp.1.1.HAP1	drMalDome.Honeycrisp.1.1.3Hap1.g031896

10.1093/genetics/iyaf006

Integration with Existing Standards:

- The GAAN framework aligns with repositories like INSDC (GenBank, EMBL-EBI, DDBJ).
- Designed to support future data sharing and interoperability with platforms like Darwin Tree of Life Identifiers and Vertebrate Breed Ontology.

Purpose:

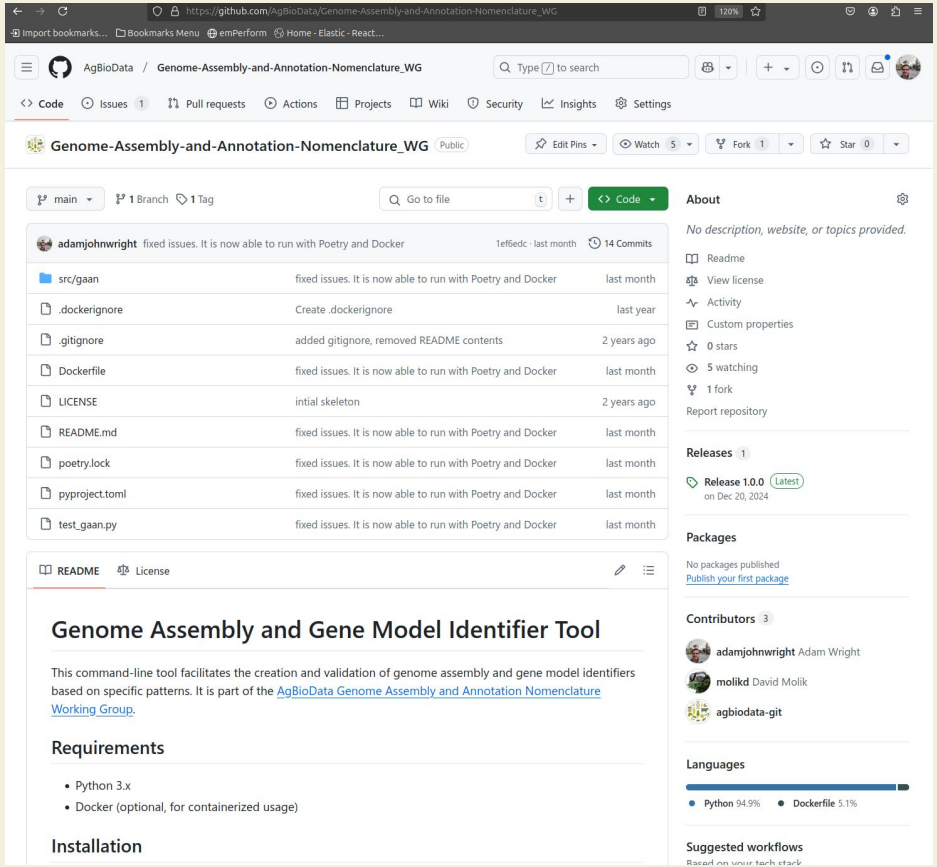
- Validates genome assembly names against the GAAN guidelines.
- Provides feedback on compliance and suggests corrections.

Current Features:

- Automated checks for required components.
- Validation against existing naming standards.

Future Plans:

- Integration with external databases for enhanced validation.
- Compatibility with evolving naming ontologies.



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

Key Advantages:

- Standardized naming improves clarity, reproducibility, and data integration.
- Simplifies collaboration across research groups and institutions.
- Facilitates computational analysis with predictable naming structures.
- Encourages adoption of global standards.

- Strengthens collaboration between researchers and institutions.
- Ensures datasets are valuable not just today but in future studies.
- Encourages adoption of GAAN across diverse communities by prioritizing openness and usability

1. Aligning with FAIR Principles

- **Findable:**
  - Standardized naming ensures genome assemblies can be easily located in global repositories (e.g., INSDC, Darwin Tree of Life).
- **Accessible:**
  - Consistent and well-documented nomenclature facilitates seamless access across platforms.
- **Interoperable:**
  - GAAN framework integrates with ontologies (e.g., Vertebrate Breed Ontology) and databases, enabling cross-platform compatibility.
- **Reusable:**
  - Clear versioning and metadata enhance data reusability for downstream analysis and future research.

2. Building Trust Through Transparency

- **Framework Transparency:**
  - Openly documented guidelines and validation criteria foster user confidence.
- **Compatibility with Standards:**
  - Alignment with established repositories and naming standards ensures reliability and widespread acceptance.
- **Tool Validation:**
  - GAAN tool automates checks and corrections, minimizing errors and maintaining data integrity.
- **Future Integration:**
  - Plans to incorporate external databases enhance long-term trust and adaptability.

- **Next Steps:**
  - Promote the GAAN framework through workshops, publications, and community discussions.
  - Enhance the GAAN tool with support for additional database and ontology lookups.
  - Continuous feedback from users to refine and update the framework.

Summary:

- GAAN provides a robust, standardized approach to naming genome assemblies and gene annotations.
- The framework aligns with global standards and supports interoperability.
- Adoption of GAAN will enhance the utility and accessibility of genomic data.

Call to Action:

- Encourage researchers and institutions to adopt the GAAN framework and tool.


AgBioData
Toward enhanced genomics, genetics, and breeding research outcomes through standardization of practices and protocols across agricultural databases