Genome assembly and annotation nomenclature

Ethy Cannon
USDA-ARS CICGRU
Ames, IA

2024 AgBioData Community Workshop

April 29, 30, & May 2 2024

The problem:

The naming of things matters!

Naming of genome assemblies, annotations and gene models is out of control. Analyses across multiple genome datasets is usually difficult and time-consuming, requiring special case handling, normalization of names and cross-references to original names.

Sometimes different groups give different names to the same thing.

Researchers tend to not realize the importance of names, or the difficulty of developing a nomenclature, and too often believe their way of naming things is better.

The solution?

Change is difficult!

Standards are difficult to enforce.

No nomenclature standard is perfect, so researchers will want to "improve" it.

Different research "cultures" may favor different approaches to nomenclatures.

Different organisms may have special needs (polyploidy, haplotype assemblies, organelles)

The AgBioData Genomic Data Nomenclature Working Group

Formed in 2021, we're are still struggling with the problem, looking for workable solutions.

We have finalized a white paper which we will submit to bioRxiv within the next few weeks.

Current active members:

Ethy Cannon *(USDA-ARS, maize)*
Sara Dyer *(EMBL-EBI, Non-Vertebrates)*
Kapeel Chougule *(CSHL, plants)*
David Molik *(USDA-ARS, arthropods)*
Adam Wright *(Ontario Institute for Cancer Research)*
Huiting Zhang *(Washington State University & USDA-ARS, fruit trees)*

The Long and Short of It

Nomenclature matters!

Use your community's nomenclature. If you don't like it, work with your community to fix it.

Human and machine readable.

Gene models should identify the source assembly.

Consider using our recommendations.

Examples of formal nomenclature standards

Note that most existing nomenclature rules apply to classical genes (loci rather than gene models), repeats, proteins, et cetera, **not** assemblies, annotations, and gene models.

Maize:

assembly name:
[Genus-species]-[cultivar]-[quality]-[group]-[version]
Zm-B73-REFERENCE-NAM-5.0

assembly identifier/annotation name:
[Genus-species][assembly id][version][annotation version]
Zm00001eb

gene model:
[annotation name][6-digit gene model number]
Zm00001eb0002040

Examples of formal nomenclature standards

Legumes (Legume Information System)

assembly: [gensp].**[cultivar]**.gnm[assembly version].[identifier]
*cerca.**ISC453364**.gnm3.3N1M* (original name listed in metadata)
annotation: [gensp].**[cultivar]**.gnm.assmbly version].ann**[annot version]**.[identifier]
*cerca.**ISC453364**.gnm3.ann**1**.3N1M*
gene model: [annotation].**[original gene model identifier]**

Examples of formal nomenclature standards

Apple (Genome Database for Rosaceae)

Assembly:

[species] **[sample identifier]** [consortium]v[version].[subversion]

Malus x domestica ***SuperCrisp*** *ABC v1.1*

Short assembly:

[ToLID].[sample identifier].[consortium].v[version].[subversion].[optional]

drMalDome.SC.ABC.v1.1

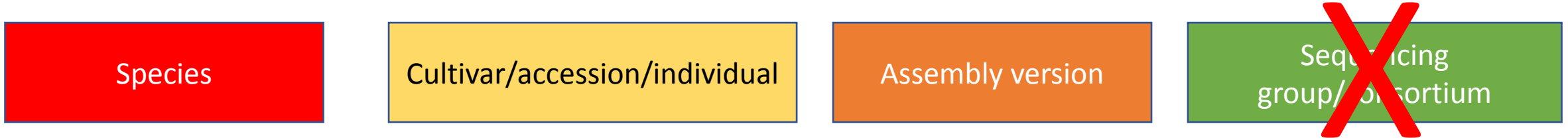
Gene model:

[short assembly]a[**annot version**][organelle][chromosome][entity][number][optional]

Nuclear gene: drMalDome.SC.ABC.v1.1a**1**.chr01Ag000010
Mitochondria gene: drMalDome.SC.ABC.v1.1a**1m**01g000010
Plastid gene: drMalDome.SC.ABC.v1.1a**1p**01.g000010

Our specific recommendations

Genome assembly name



Use Tree of Life for species name (ToLID)

[high level taxonomic rank][clade][one upper, two lower case letters for genus][one upper, three lower case letters for species]

ipZeaMays, ddBraNapu, drMalDome

Examples:

lpZeaMays.**00001**.**5.0**, ddBraNapu.**DAR**.**1.0**, drMalDome.**Honeycrisp**.**1.0***

* This is a haplotype assembly

Gene model naming

Assembly name	Annotation version	Chromosome identifier	Entity type e.g. gene/transcript/pangene	6-digit identifier (often ordered with gaps)

Annotation version is an integer (no subversion)

Chromosome identifier includes optional subgenome and haplotype

Entities: **g**=gene, **p**=pan-gene, **t**=transcript

The 6-digit identifier can be ordered along chromosomes, and can contain gaps, e.g., number by 10s.

Gene model naming

gene models:

lpZeaMays.00001.5.0.2.**01g**000050
ddBraNapu.DAR.1.0.1.**01Cg**010030
drMalDome.Honeycrisp.1.1.1.**03Hap1g**031896

Isoforms:

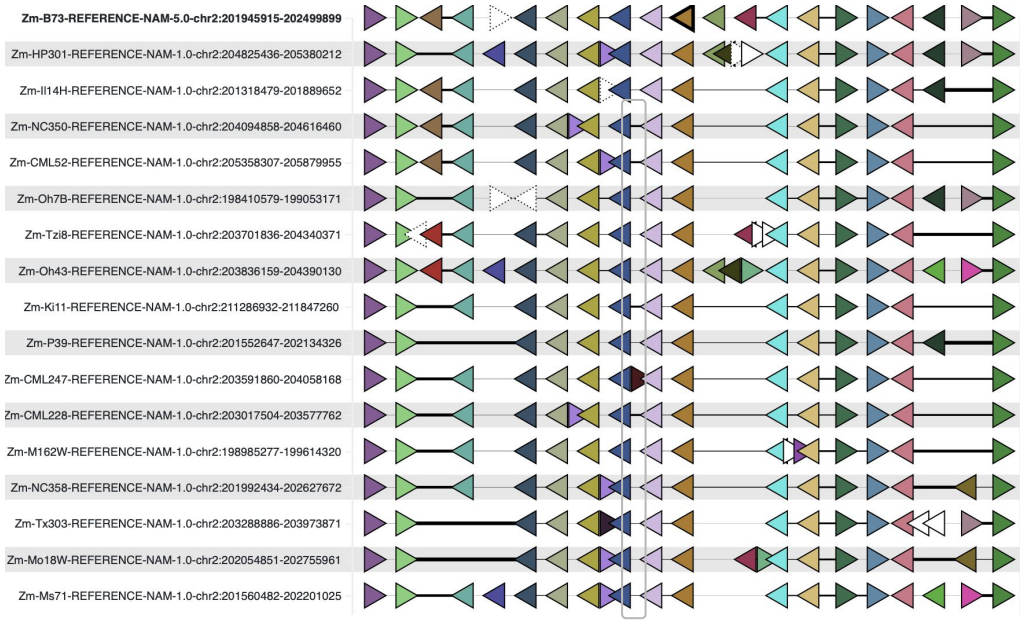
lpZeaMays.00001.5.0.2.**01t**000050.3
ddBraNapu.DAR.1.0.1.**01Ct**010030.1
drMalDome.Honeycrisp.1.1.1.**03Hap1t**031896.1

Pan-gene nomenclature

What is a pan-gene?

A possible definition for a pan-gene is **the set of all gene models in a set of annotations that appear to be the same thing.** This is determined by sequence similarity and syntenty. If one or more gene models has been associated with a classical locus, the locus is also a member.

By syntenty ...



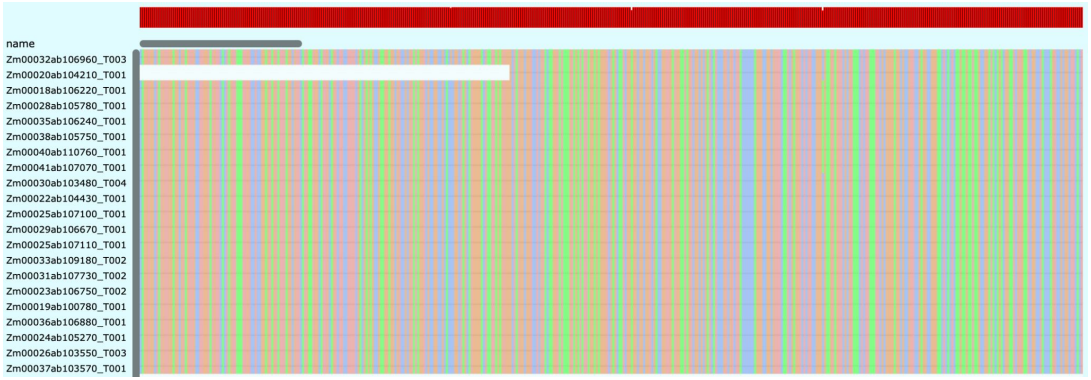
How should it be identified?

Annotation and pan-gene methods are still evolving, so permanent identifiers should not be defined. A pan-gene should be identified by any of its members or associated locus.

Naming of analysis-specific pan-genes could be:

1. [clade].[version].pandddddd
2. [clade].[version].[pan-position].pandddddd
3. [clade].[version].[chr*].pandddddd
4. [clade].**official**.pandddddd OR [clade].[group].pandddddd

... and by sequence similarity



Pan-gene nomenclature

How should it be identified?

Annotation and pan-gene methods are still evolving, so permanent identifiers should not be defined. A pan-gene should be identified by any of its members or associated locus.

Naming of analysis-specific pan-genes could be:

1. [clade].[version].pandddddd
2. [clade].[version].[pan-position].pandddddd
3. [clade].[version].[chr*].pandddddd
4. [clade].**official**.pandddddd OR [clade].[group].pandddddd
5. [clade].**official**.[ver].pandddddd OR [clade].[group].[ver].pandddddd

Summary of working group outcomes

- Final draft of white paper will be submitted to bioRxiv
- Formal submission of the white paper will follow shortly
- Letter emphasizing the importance of standard nomenclature will be submitted in a co-submission.

Summary of working group outcomes

Past and present working group members:

Kapeel Chougule (chair)
Sarah Dyer* (Co-chair)
Sunil K Kenchanmane* (past chair)
Ethy Cannon*
Justin Elser
Nathan Grant
Lucia Hoffmann
Sachiko Isobe
Pankaj Jaiswal
Yogendra Khedikar
David Molik*
Paul Otyama
Tara Rickman
Patrice Salomé
Brian Smith-White
Adam Wright*
Huiting Zhang*

* currently active members

Assembly (and Annotation) QC metrics

- The nomenclature WG was unable to make much progress on this topic.
- Nonetheless, it is important.
- New metrics are emerging as assembly and annotation methods improve.

Examples of formal nomenclature standards

Brassica:

gene model: [genome letter*][chromosome]['p' if pan-gene][5 digit gene model number].[version number]_[genus letter][species 2 letters][3 LETTER genotype]

C01p010030.1_BnaDAR

* X = cross species