Whole-Genome Assembly and Annotation Nomenclature for *Zea mays*

2021 update

Introduction

As the number of genome assemblies, gene model sets and gene models increase, unambiguous identification of each of these entities is necessary for clarity and precision. Our objective is to define identifiers that are unique across genomes, will not change, and will follow a nomenclature that is both machine and human readable. To ensure that identifiers are unique, MaizeGDB will be the single naming authority for the assignment of identifiers. The intent here is to provide a system for assigning unique and stable identifiers, so that researchers can be confident a genome and its gene models are all uniquely identified. Genome assembly and annotation Project Personnel should work with MaizeGDB Personnel to acquire genome and annotation names that comply with the guidelines herein. In addition to the need for unique persistent identifiers, metadata is required to properly identify and describe genome assembly datasets. Project Personnel from genome assembly and annotation groups will also work with MaizeGDB Personnel to provide required metadata, as outlined here: https://www.maizegdb.org/contribute_data.

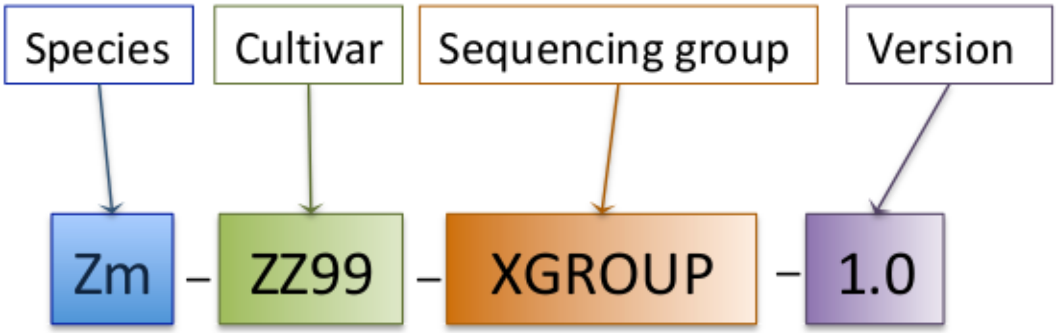
1. Genome Assembly Names

1a. Preamble:

A genome assembly is defined here as a group of sequences representing an entire genome, their relative order and metadata that allows unambiguous identification of the genome. A genome assembly dataset contains several entities including; a unique name, unambiguous identification of the sample from which DNA was taken, the type and quality of the assembly, the program and methods used to assemble the sequences, and the contig, scaffold and/or chromosomal sequences. (more details here: https://www.maizegdb.org/contribute_data). Each new public genome assembly will be given a globally unique, persistent name and identifier. MaizeGDB will be the naming authority for these names. Genome assembly Project Personnel preparing to deposit genome assemblies into NCBI/EBI/DDBJ should first contact MaizeGDB to be assigned official identifiers.

1b. Genome assembly name

<species id>- <cultivar>-<project>-<assembly version> .<sub-version>



The genome assembly name

The name for the first assembly version of a cultivar named ZZ99, sequenced by group XGROUP:
Zm-ZZ99- XGROUP-1.0

1c. Definition of each element in the name

<species id>: an uppercase letter "Z" and lowercase letter "m" to indicate the maize species *Zea mays*; further information about the species will be included in the metadata of the assembly. Additional two letter codes can be used for other species in the genus Zea as follows:

Zea mays subspecies mays	Zm
Zea mays subspecies parviglumis	Zv
Zea mays subspecies mexicana	Zx
Zea mays subspecies huehuetenangensis	Zh
Zea diploperennis	Zd
Zea perennis	Zp
Zea luxurians	Zi
Zea nicaraguensis	Zn

<cultivar>: a short, descriptive name of the cultivar/inbred/accession/land race used. The name must be descriptive and distinguishable from lines with similar names. Note: To further uniquely identify the **<cultivar>**, seed representing the cultivar/inbred/accession/land race (often bulk lots of sibling seeds) should be made publicly available by depositing seed at, e.g. Plant Introduction Station or Maize Genetics COOP Stock Center), and their unique accession must be given in the metadata.

<project>: A short name or abbreviation that uniquely identifies the project responsible for the assembly. The project identifies the original generators of the sequence and the owners of that sequence in NCBI. More information to uniquely identify the sequence generators must be given in the metadata.

<assembly version>: A numerical value representing the version of the assembly. A new assembly version will be given when a sequence assembly is significantly improved by the original Project Personnel, or when the exact accession is sequenced again and a new assembly is created by the original Project Personnel. If the sequenced cultivar is not the exact same as previous versions, a new genome identifier is given. If a different group sequences the same exact cultivar, a new genome identifier is given.

<sub-version>: Indicates a minor update rather than a complete resequencing and/or re-assembly.

NOTE: Between 2016 and 2021, genome names included a **quality** field, which could contain either REFERENCE or DRAFT. In short order, given the rapid improvement of genome sequencing technologies, nearly all new maize genomes were of quality 'REFERENCE'. Thus it is no longer necessary to include quality in all assembly names. This provides an added benefit of shortening genome assembly names. **This change will be made for assemblies completed in 2021 and beyond and will not apply retroactively to older assemblies.**

1d. Genome assembly code

Each genome assembly will also get a “shorthand” **genome assembly code**. This short code will be a synonym for the Genome Assembly Name and will be the connection between a genome assembly and its annotations and gene models. MaizeGDB will be the single naming authority to assign genome identifiers.

The genome assembly code is:

<species id><genome id #><genome assembly version letter>

Examples: (Note that older assemblies use 2016 naming rules)

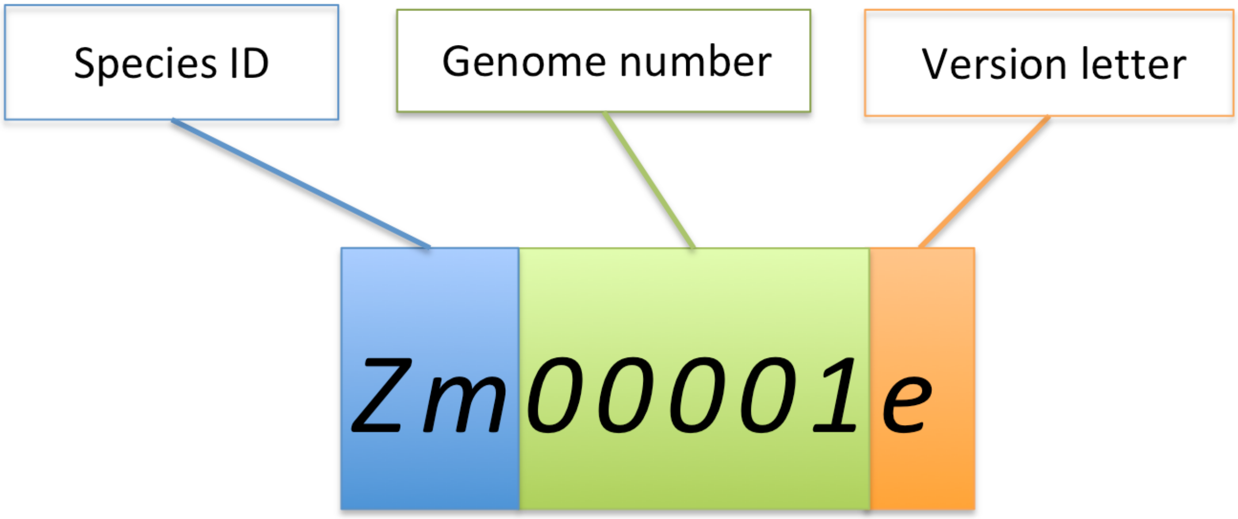
Zm00001e = Zm-B73-REFERENCE-NAM-5.0
Zm00004b = Zm-W22-REFERENCE-NRGENE2.0
Zm00018a = Zm-B97-REFERENCE-NAME-1.0

1d.1 Definition of each element in the genome assembly code

<species id>: see <species id> explanation above

<genome id #> is a 5 digit number that represents a single genome, linked to a single genome assembly name. **This number is assigned by MaizeGDB**. If the 100,000 unique 5 digit numbers IDs run out, IDs can include alphabetic characters (with the only restriction that there cannot be more than 2 consecutive alphabetic characters, to avoid words). Adding alphabetic characters in the assembly code will produce an identifier range from 0000A – ZZ9ZZ, providing 50 million additional identifiers.

<genome assembly version letter>: an alphabetic representation of the assembly version number, incrementing the letter for each major or minor update. The first assembly version would be indicated with an 'a', the next with a 'b', and so on. For example, a set of assembly versions might be: 1.0=a, 1.1=b, 2.0=c. If an assembly has more than 26 versions, then it will be assigned a new assembly ID #, starting at version 27='a', and will be linked to the original assembly ID in its metadata.



The genome assembly identifier

1e. Process required for official assignment of a genome ID

To receive a Genome Assembly Name and Genome Assembly Code, Project Personnel should contact MaizeGDB before submitting a genome sequence assembly to NCBI/EBI/DDBJ. All genomes served and supported through MaizeGDB will be required to comply with this nomenclature convention.

2. Gene Model Set IDs

2a. Preamble

A Gene Model Set is defined here as the complete group of predicted genes from a single genome assembly. A Gene Model Set is the outcome of an annotation pipeline that calculates the structure of potential genes based on their sequence, and includes genome coordinates for each predicted gene.

Gene model sets are specific to the genome assembly from which they were derived. Each Gene Model Set resulting from the annotation of a specific genome assembly will be given a globally unique, persistent identifier. MaizeGDB will be the naming authority to assign identifiers for Gene Model Sets.

The identification of two types of Gene Model Sets are described in this document. The first type is the "**project gene model set.**" This is the set of gene models produced by the genome assembly Project Personnel, or by a group collaborating closely with the Project Personnel. The second type is a "**2nd party gene model set,**" which is a set of gene models from annotation carried out by a group not associated with the original genome assembly Project Personnel. No indication of quality is inherent in either group. Distinguishing these two groups allows for the coexistence of independent gene model sets that do not require integration with each other.

2b. ID convention for the Project Gene Model Set

Since the 2016 recommendations for naming gene model sets, it became necessary to add a character to indicate the version of gene model sets in cases when an assembly is re-annotated by the sequencing group using new methods.

Used for B73 v4 (Zm-B73-REFERENCE-GRAMENE-4.0; Zm00001d) and earlier:
<genome assembly code>.<annotation minor version number>

B73 v5 (Zm-B73-REFERENCE-NAME-5.0) and later:
<genome assembly code><annotation major version letter>.<annotation minor version number>

Major version: all gene model identifiers change.
Minor version: a few gene models might be added, removed or changed but identifiers are otherwise retained.

Examples (2016 style):
Zm00001c.1 - Gene model set for B73 RefGen_v3
Zm00001d.1 – Gene model set for Zm-B73-REFERENCE-GRAMENE-4.0
Zm00001d.2 – Revised Gene model set for Zm-B73-REFERENCE-GRAMENE-4.0
Zm00004b.1 - Gene model set for Zm-W22-REFERENCE-NRGENE-2.0

Examples of 2021 update:
Zm00001eb.1 - 2nd annotation of Zm00001e (Zm-B73-REFERENCE-NAM-5.0), 1st release
Zm00001eb.2 - 2nd annotation of Zm00001e (Zm-B73-REFERENCE-NAM-5.0), 2nd release (gene model ids retained)
Zm00018ab.1 = 2nd annotation of Zm00018a (Zm-B97-REFERENCE-NAM-1.0), 1st release.
Zm00056aa.1 = 1st annotation of Zm00056a (Zm-A188-REFERENCE-KSU-1.0), 1st release

2c. Definition of each element in the ID convention for the Representative Gene Model Set

<genome assembly code>: See section 1.d above.

<annotation major version letter>: A letter indicating the major version number. a=1st full annotation, b=the 2nd full annotation (a complete reannotation that does not retain ids from the 1st), et cetera.

<annotation minor version number>: A sequential number indicating the minor version of the gene model annotation set, which is assigned by the annotation group. New versions should be assigned when the group makes a change to the set, which includes only new, improved, and/or deleted gene models within the set. Gene models in common between the minor versions retain their ids. New versions should be communicated to MaizeGDB.

2d. Recommendation for Alternative Gene Model Set IDs

As annotation methods improve, there are likely to be increasing numbers of independent annotation analyses for any genome assembly. Naming of the resulting Alternative Gene Model Sets should not resemble the naming of the Representative Gene Model Set, but should have a similar number of characters, and should be flexible. The important issue for Alternative Gene Model Sets is to create a name that can be used as a prefix for the individual gene models in the Alternative Gene Model Set. We recognize that this means that the Genome Assembly Code will thus not feature in The Alternative Gene Model Set name, and linking the set to the Assembly will be done through the metadata. MaizeGDB will be the naming authority to assign IDs for Alternative Gene Model Sets.

<alternative gene model set ID>: 8 alphanumeric characters that meet the following criteria:

- does not start with 'Z' or 'z',
- contains no spaces, punctuation, or special characters,
- will be used as a prefix for gene model ids (see section 3a. below),
- does not end with a digit.

Examples:

ISU2019a
ISU19gpg
B73v4ISU

A version annotation code can also be added, at the discretion of the annotation group. New versions should be communicated to MaizeGDB.

Note: Exceptions are made for long-standing gene annotation pipelines with their own established naming conventions, such as the NCBI Eukaryotic Genome Annotation Pipeline (https://www.ncbi.nlm.nih.gov/genome/annotation_euk/process/).

3. Gene Model IDs

3a Preamble: A gene model is here defined as the genomic DNA sequence of a predicted gene defined by all of its alternative transcripts and 5' and 3' regulatory region. As a gene model is a computed entity, it may or may not be associated with a functional gene. Gene models result from annotation of individual genome assemblies; and thus, each gene model is tied to the genome assembly from which it was calculated.

3b. ID convention for Gene Models from the Representative Gene Model Set:

<genome assembly code><annotation major version letter><six digits>

For transcripts and proteins:
<genome assembly code><annotation major version letter><six digits><transcript/protein>

Examples (2016 style; see section 2.b above):

Zm00001a459310
Zm00001d459384
Zm00004a845733

Examples (Updated 2021 style)

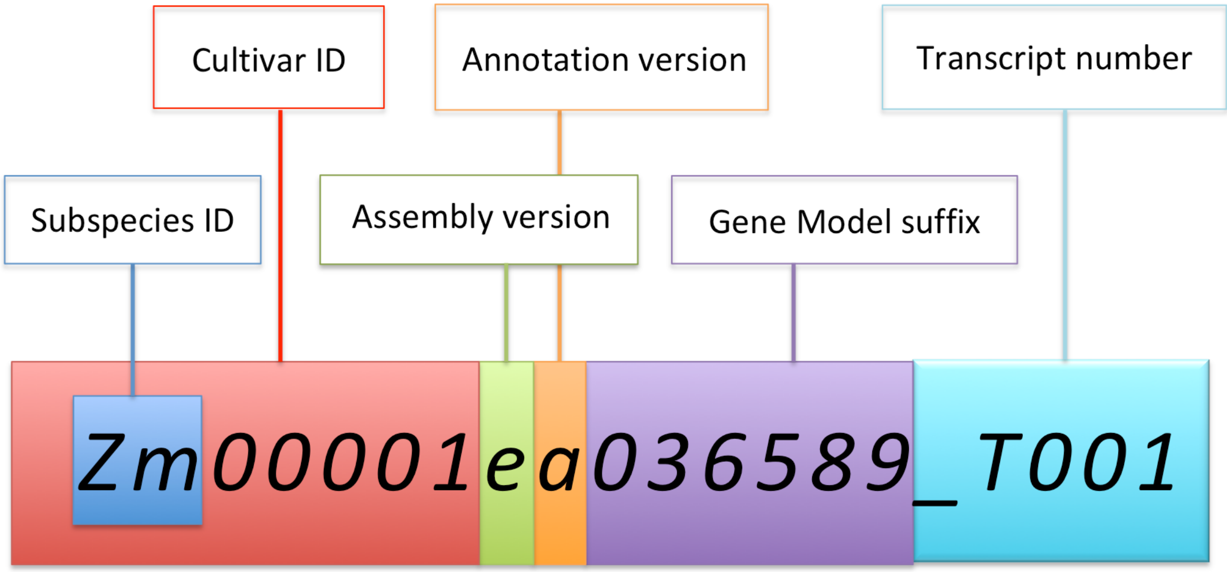
Zm00001ea000353 - first annotation
Zm00001eb00040 - second annotation

<genome assembly code>: see details in section 1b. above.

<annotation major version>: see details in section 2b. above.

<six digits>: a random six-digit number that is unique per gene model within the assembly. As genome assemblies appear to now be sufficiently stable to have some confidence in gene order, gene models should be named sequentially, with gaps of at least 10 to permit inserted gene models that are found after the annotation was completed, or to accommodate merged gene models that should be split into two or more separate gene models. **If there are no sequential numbers available when adding a new, split, or merged gene model, choose the closest available number.**

<transcript/protein>: a transcript is indicated by appending Tddd, where ‘ddd’ is the number of the isoform, for example Zm00001e002340_T001, Zm00001e002340_T002. For proteins, ‘P’ is substituted: Zm00001e002340_P001, Zm00001e002340_P002.



Zm -> Zea mays
Zm00001 -> Zea mays B73 Reference Genome (CSHL / NAM project)
Zm00001e -> Zm-B73-REFERENCE-NAM-5.0 (B73 RefGen_v5)
Zm00001ea - > Zm-B73-REFERENCE-NAM-5.0 Annotation set 1.0
Zm00001ea036589 -> Gene model (wx1) in B73 v5 a1
Zm00001ea036589_T001 ->Transcript 001 for gene model (wx1) in B73 v5 a1

The transcript id

Additional gene model ID rules:

- Merged gene models will retain the name of the upstream gene model.
- Split gene models: the first retains the original gene model identifier, subsequent models get sequential numbers, with even gaps, if possible.
- Gene models that have minor improvements (improved exon/intron structure, sequence error corrections, changes that don't significantly change the length of the gene model) keep their ID.
- When the ownership of an annotation set is transferred to another group it will retain the same assembly code, but subsequent versions will reflect the new ownership in the metadata.
- Metadata about the sequenced entity is attached to the assembly version, and will follow the MIxS recommendation (http://gensc.org/projects/mixs-gsc-project/). This information will be available for any given gene model through its membership in a gene model set..
- Transcript and translation IDs are not conserved within assemblies or across gene model set versions. Transcripts and translations are renumbered for each new assembly.

- Coding and noncoding genes are numbered the same and this information is not encoded in the ID. This information may be in the metadata. It is acceptable to only submit coding genes to GenBank, but this information needs to be in the metadata.

3c. ID convention for Gene Models from an Alternative Gene Model Set:

<alternative gene model set ID><gene model number>

<alternative gene model set ID>: Defined as 8 alphanumeric characters in 2d above.

<gene model number>: 6 additional characters unique to each gene model.

Additional Guidelines:

- Gene model IDs must be 14 characters in length.
- The Gene model ID will be made up of an 8 alphanumeric prefix and 6 digit gene model number. No spaces, punctuation, or special characters are allowed. The prefix represents the project and genome assembly and is constant for each gene model. This prefix must be preregistered with MaizeGDB to guarantee uniqueness.
- The prefix cannot start with 'Z' or 'z' as these leading letters are reserved for official annotation sets.
- The last 6 characters are numeric only. Each gene model has a unique 6 digit number.
- All transcripts IDs begin with the Gene Model ID and ends with '_T' followed by a three-digit number (e.g., _T001). The three-digit number starts with 001 and increments with each additional transcript.
- All translation IDs begin with the Gene Model ID and ends with '_P' followed by a three-digit number (e.g., _P001). The three-git number starts with 001 and increments with each additional transcript.

Exceptions are made for long-standing gene annotation pipelines, such as the NCBI Eukaryotic Genome Annotation Pipeline.