Whole-Genome Assembly and Annotation nomenclature

1. Genome Assembly IDs

<assembly> = <species id>, "-", <cultivar>, "-", <quality>, "-", <project>, "-", <version>

Examples: Zm-B73-REFERENCE-GRAMENE-4.0
Zm-W22-REFERENCE-NRGENE-1.0

<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 title and metadata of the assembly. Additional two letter codes can be used for other species in the genus Zea (e.g. *Zea diploperennis* = Zd, *Zea luxurians* = Zl, *Zea nicaraguensis* = Zn, and *Zea perennis* = Zp).

<cultivar>: a short, descriptive name of the cultivar used. The name must be descriptive in such that it can be distinguished from lines with similar names. The cultivar must be made available (e.g. Plant Introduction Station, Maize Genetics COOP Stock Center)

<quality>: a categorical description based on the quality of the assembly. Values include:

- REFERENCE, contig order is provided as full pseudomolecules, the AGP file may have been submitted to GenBank, and the underlying contigs have been submitted to GenBank;
- REFERENCE_NS, contig order is provided as full pseudomolecules, but underlying contigs are not submitted to GenBank;
- DRAFT, no or limited contig order is provided and underlying contigs have been submitted to GenBank;
- PATCH, an intermediate update of an existing assembly with multiple updates or improvements but not enough to justify a full new assembly version. An assembly that has not been deposited into GenBank will have a “_NS” suffix added to the quality term to designate it as “not submitted.”

<project>: A short name or abbreviation that uniquely identifies the project responsible for the assembly.

<version>: A numerical value representing the version of the assembly.

2. Gene Model Set IDs

<assembly_version_code> = <assembly code><version code>.<sub-version code>

Examples: Zm00001c.1 - B73 RefGen_v3
Zm 00001d.1 – Zm-B73-REFERENCE-GRAMENE-4.0
Zm 00001d.2 – Zm-B73-REFERENCE-GRAMENE-4.0 revised annotation
Zm 00002a.1 - Palomero toluqueno from CINVESTAV
Zm 00003a.1 - Mo17 from JGI
Zm 00004a.1 - W22 from Brutnell et al

<assembly code>: Assembly/project codes are 5 alphanumeric characters long. Groups doing single genomes (or even a handful to a dozen genomes) are assigned IDs ranging from 00001 – 99999. The five digits are numeric and represent a single assembly. This provides us with 100,000 unique IDs to represent different assemblies. When these IDs runs out or a group is able to sequence, assemble, and annotated 100s to 1000s of genomes at a time then they are assigned IDs that include alphabetic characters (with the only restriction is that there cannot be more than 2 consecutive alphabetic characters – to avoid words). The alphabetic characters in the assembly code will be uppercase. These assemblies would range from 0000A – ZZ9ZZ. This would provide us around 50 million additional IDs. In the case where there are over 50 million genomes, we could then possibly expand to six characters (over a billion IDs).

<version code>: The last character of the assembly code is always alphabetic and lowercase. It is associated with the version of the assembly. If an assembly has more than 26 versions, than it will be assigned a new assembly ID, starting at version 27, and will be linked to the original assembly ID in the metadata.

<sub-version code>: If a gene model set corresponding to a specific assembly is modified, the sub-version number is incremented.

3. Gene Model IDs

A **gene model** is defined to be a consensus gene model plus all of its full set of alternative transcripts. If any changes are made to the consensus gene model or any of its transcripts, including adding or removing one or more transcripts, the version number of the gene model is incremented.

<genemodel> =<speciesid><assembly_code><version_code><sixdigits>.<version>

Examples: Zm00001a459384.1
Zm00001d459384.1
Zm00001d459384.2
Zm00004a845733.1
Zm05GG7d832948.1

<species id>: an uppercase letter “Z” and lowercase letter “m” to indicate the maize species Zea mays.

<assembly_code>: 5 alphanumeric characters representing the assembly (alphabetic characters are in uppercase) - see details above.

<six digits>: a random six-digit number that is unique per gene model within the assembly. The order of these numbers DOES NOT indicate the sequential order of a gene along a chromosome.

<version>: if a gene model changes in a new gene model set, its version number is incremented. Changes might include revised transcripts, or the removal or addition of transcripts. Note that if a gene model is split or multiple gene models are merged, they will get new IDs.

Additional gene model ID rules:

- Merged gene models will get new IDs.
- Split gene models will get new IDs.
- Gene models that are improved keep their ID, but get new version within an assembly.
- Gene models that are associated with each other across different assemblies/lines will have different unique IDs, but will be linked using a linking table.
- 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.
- Patches will receive an additional alphabetic character at the end of the ID to reference the version number of the patch. Only gene models updated in a patch version will receive an updated ID. (e.g. Zm0001a459384a)
- Option 1: If an assembly will release patch updates, it will receive an additional assembly code that will represent the patches. Option 2: If an assembly will release patch updates, the patches will be versioned within the original assembly code.
- Gene models should only be updated when an assembly version / patch version is updated.
- A cultivar ID should be linked to an accession number assigned to a germplasm collection.
- 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 (structural annotation version).
- Transcripts and translations IDs will be represented with “_T” or “_P” followed by a three-digit ID appended onto the gene model ID (e.g. Zm00001d459384_T001 and Zm00001d459384_P001). Corresponding transcripts and proteins have the same ID number.

- 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 will be in the metadata. It is acceptable to only submit coding genes to GenBank, but this information needs to be in the metadata.
- Nomenclature for gene families is not addressed in this document, but should follow similar conventions.

4. Notes:

Advantages to this nomenclature include:

- A 14-character gene model identifier is a compromise between ease of use and the flexibility of a larger identifier. This identifier allows for around 50 million different assemblies, 26 versions per assembly, and 1 million gene models per version.
- The gene model is fixed-length with clear rules on what each character represents. This will allow for easy parsing.
- The gene model is human readable. The use of digits and characters allow for a human to quickly distinguish between the species ID, assembly code, and unique gene model ID.
- There is no biological information built into the identifier.
- The biological information is in the required metadata, which can be updated without affecting the gene model ID.
- Relationship information between assemblies is not built into the identifier. This information can be created through linkage tables.

Disadvantages to this nomenclature include:

- Fixed length identifiers do put a limit on the number of unique IDs that can be created. It is possible in the future that the availability of the number of assemblies and versions per assembly could exceed the capacity of this nomenclature model. The assembly code could be expanded to 6-characters if needed to allow for over a billion IDs.
- The 14-character ID is slightly larger than other species (e.g. Rice is 12 characters, Arabidopsis is 9 characters).
- There is some convenience in having biological information stored in the identifier. An additional step is required to use the metadata.