Gene and genome nomenclature

Kapeel Chougule
Cold Spring Harbor Laboratory, NY

2023 AgBioData Community Workshop
May 1 & 2nd 2023
Motivation

Importance of accurate and persistent identifiers for assemblies and gene models in the public domain

This will help users:

• Understand multiple assemblies and annotations per species
• Replicate results and understand differences
• Compare gene models across assemblies
• Track citation and downstream use
Genome / assembly naming conventions

• Components include:

Species identifier  Assembly version  Cultivar/accession/individual  Sequencing group/consortium

e.g. fCotGob3.1 = 1st assembly version of 3rd individual of fish (ToLID prefix f) Cottoperca gobio (CotGob) from DToL project

• We would like to identify best practice recommendations for Agbio communities
Gene model ID naming conventions

• Components include:

<table>
<thead>
<tr>
<th>Subgenome identifier</th>
<th>Chromosome identifier</th>
<th>Entity type e.g. gene/transcript/pangene</th>
<th>Entity numeric identifier (often ordered with gaps)</th>
<th>Annotation version</th>
</tr>
</thead>
</table>

e.g. C01p010030.1 = C genome, chromosome 1, type=pangene, identifier=010030, version=1

• A need to capture transcript isoform, annotation version of gene model and assembly version without confusion
Putting it all together

• Very long identifiers:
  - Species identifier
  - Assembly version
  - Cultivar/accession/individual
  - Genome number
  - Sequencing
  - Group/consortium
  - Subgenome
  - Identifier
  - Chromosome
  - Identifier
  - Entity type e.g.
  - Gene/transcript/pangene
  - Entity numeric identifier
  - (often ordered with gaps)
  - Annotation version

• But human readable and accurate
• Ideally machine readable too
# Gene model IDs

<table>
<thead>
<tr>
<th>Examples:</th>
<th>Species</th>
<th>Assembly version</th>
<th>Accession</th>
<th>Group</th>
<th>Sub-genome</th>
<th>Chromosome</th>
<th>entity</th>
<th>ID #</th>
<th>Annot. version</th>
</tr>
</thead>
<tbody>
<tr>
<td>C01p010030.1_BnaDAR</td>
<td>B na</td>
<td>DAR</td>
<td>C</td>
<td>01</td>
<td>p</td>
<td>010030</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glyma.01g000100.Wm82.a2.v1</td>
<td>Gly ma</td>
<td>a2</td>
<td>Wm82</td>
<td>01</td>
<td>g</td>
<td>000100</td>
<td>v1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Horvu_BARKE_1H01G000300.1</td>
<td>Hor vu</td>
<td>BARKE</td>
<td>1H</td>
<td>G</td>
<td>000300</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TraesCS3D02G273600</td>
<td>Tr aes</td>
<td>CS</td>
<td>D</td>
<td>3</td>
<td>G</td>
<td>273600</td>
<td>02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vitvi18g12230</td>
<td>Vit vi</td>
<td></td>
<td></td>
<td>18</td>
<td>g</td>
<td>12230</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zm000001eb0000050</td>
<td>Z m</td>
<td>e</td>
<td>00001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Element order varies - which part relates to which element?
- Conventions vary e.g. 1-3 letter abbreviations for species
  - *Vitis vinifera* as **Vitvi** or **Vivin** or **Vvi** or **Vv**
- Special characters
  - letters and digits safest
  - dashes, full stops and underscores may cause unexpected parsing outcomes
Community Survey Feedback

AgBioData Genome Assembly and Annotation Nomenclature Working Group survey

This survey is designed to
1) Gather feedback regarding genome assembly and gene model identifier naming preferences for AgBioData species
2) Explore metrics used for assessing genome assembly quality

Total 11 respondents
And the survey said...

Species identifier
Cultivar/accession/individual /genome number
Assembly version
Annotation version

Species identifier
Cultivar/accession/individual /genome number
Assembly version
Annotation version
Entity type e.g. gene/transcript/pangen
Entity numeric identifier (often ordered with gaps)
Subgenome identifier
Chromosome identifier

With spacer characters...
Recommendations / considerations

Species - use ToLIDs - unique but variable length (7-9 chars)
https://id.tol.sanger.ac.uk/search

Variety / accession / individual - variable length

Subgenome - may or may not be present

Versions could be >1 character over time
What does this look like in reality?

<table>
<thead>
<tr>
<th>Original locus IDs:</th>
<th>New assembly IDs:</th>
<th>New locus IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>C01p010030.1_BnaDAR</td>
<td>ddBraNapu.DAR.1.1</td>
<td>ddBraNapu.DAR.1.1.01Cp010030</td>
</tr>
<tr>
<td>Glyma.01g000100.Wm82.a2.v1</td>
<td>drGlyMaxx.WM82.a2.1</td>
<td>drGlyMaxx.WM82.a2.1.01g000100</td>
</tr>
<tr>
<td>Horvu_BARKE_1H01G000300.1</td>
<td>lpHorVulg.BARKE.1.1</td>
<td>lpHorVulg.BARKE.1.1.01g000300</td>
</tr>
<tr>
<td>TraesCS3D02G273600</td>
<td>lpTriAest.CS.1.2</td>
<td>lpTriAest.CS.1.2.03Dg273600</td>
</tr>
<tr>
<td>Vitvi18g12230</td>
<td>drVitVini.PN40024.1.1</td>
<td>drVitVini.PN40024.1.1.18g012230</td>
</tr>
<tr>
<td>Zm000001eb000050</td>
<td>lpZeaMays.00001.e.b</td>
<td>lpZeaMays.00001.e.b.01g000050</td>
</tr>
</tbody>
</table>

<ToLiD>.<variety>.<assembly version>.<annotation version>.<chr><subgenome><entity><numeric ID>
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 synteny. If one or more gene models has been associated with a classical locus, the locus is also a member.

By synteny ...

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
Assembly Quality Control (QC) metrics

The ability to understand and compare the quality and completeness of genome assemblies and annotations.

• Catalog common, existing QC metrics
• Keep in mind that older metrics may not work well for newer assemblies which are increasingly telomere-to-telomere
• Recommend a minimum set of metrics to permit comparing assemblies and annotations to each other
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.
Summary & Future directions

- Active engagement with communities
  - ID components we are missing / have not considered from our communities?
  - Are long IDs acceptable or can / should some components be sacrificed?
  - Do the IDs need to be human readable at all?
  - Which QC metrics for assemblies and annotations?

- Next 6 months
  - Community feedback and engagement
  - White paper

More information: [https://www.agbiodata.org/node/451](https://www.agbiodata.org/node/451)
Come and share your thoughts!

Or email: agbiodata@gmail.com.

Join AgBioData on [Slack](#)
Points for breakout session

• Would you adopt our standard?
• How to raise awareness and encourage adoption? (Target audience?)
• If there was a service to mint IDs would you use it? If not, why not?

Come and share your thoughts!