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

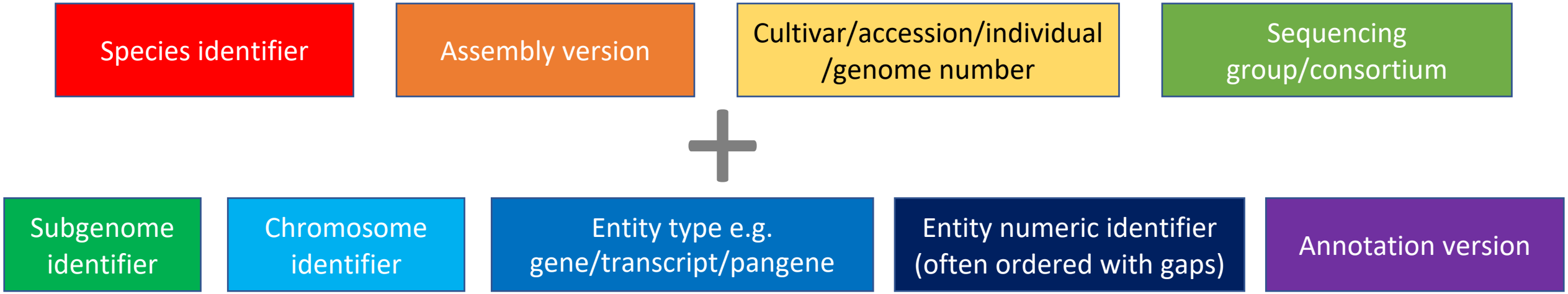
- Components include:

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:



- But human readable and accurate
- Ideally machine readable too

Gene model IDs

Examples:	Species	Assembly version	Accession	Group	Sub-genome	Chromo-some	entity	ID #	Annot. version
C01p010030.1_BnaDAR	B na		DAR		C	01	p	010030	.1
Glyma.01g000100.Wm82.a2.v1	Gly ma	a2	Wm82			01	g	000100	v1
Horvu_BARKE_1H01G000300.1	Hor vu		BARKE			1H	G	000300	.1
TraesCS3D02G273600	Tr aes		CS		D	3	G	273600	02
Vitvi18g12230	Vit vi					18	g	12230	
Zm00001eb000050	Z m	e	00001					000050	b

- 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

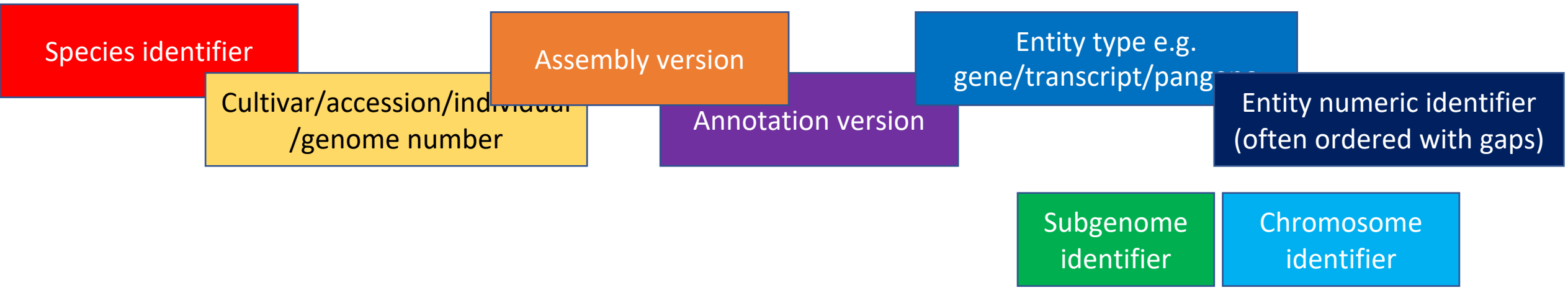
And the survey said…

Species identifier

Cultivar/accession/individual /genome number

Assembly version

Annotation version



With spacer characters…

Recommendations / considerations

Species - use ToLIDs - unique but variable length (7-9 chars)
https://gitlab.com/wtsi-grit/darwin-tree-of-life-sample-naming/-/blob/master/tolids.txt
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?

Original locus IDs:	**New assembly IDs:**	**New locus IDs**
C01p010030.1_BnaDAR	ddBraNapu.DAR.1.1	ddBraNapu.DAR.1.1.01Cp010030
Glyma.01g000100.Wm82.a2.v1	drGlyMaxx.WM82.a2.1	drGlyMaxx.WM82.a2.1.01g000100
Horvu_BARKE_1H01G000300.1	lpHorVulg.BARKE.1.1	lpHorVulg.BARKE.1.1.01g000300
TraesCS3D02G273600	lpTriAest.CS.1.2	lpTriAest.CS.1.2.03Dg273600
Vitvi18g12230	drVitVini.PN40024.1.1	drVitVini.PN40024.1.1.18g012230
Zm00001eb000050	lpZeaMays.00001.e.b	lpZeaMays.00001.e.b.01g000050

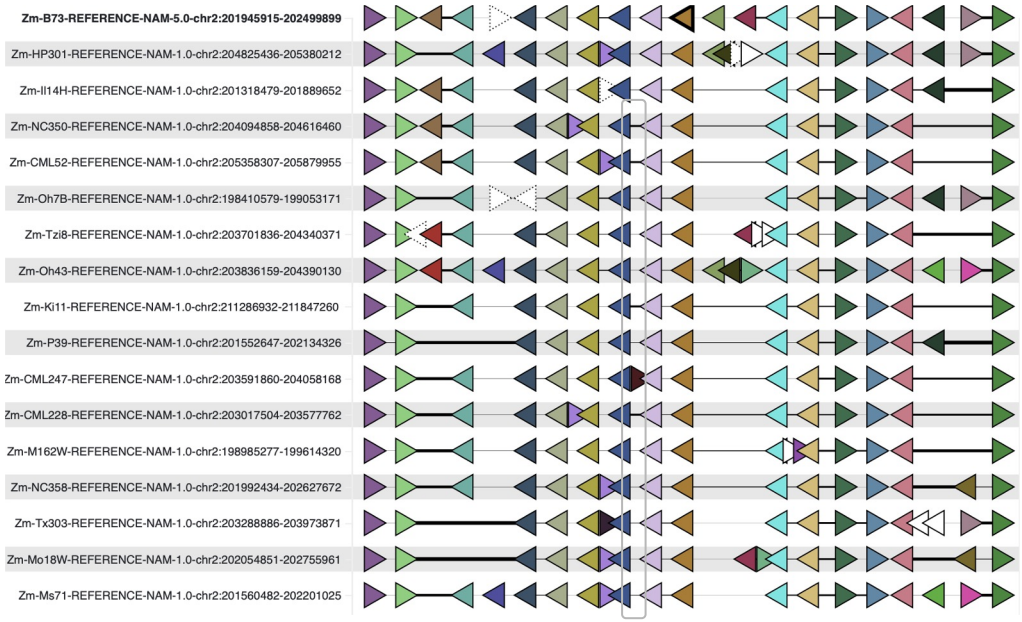
<ToLID>.<variety>.<assembly version>.<annotation version>.<chr><subgenome><entity><numeric ID>

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 syntenty. If one or more gene models has been associated with a classical locus, the locus is also a member.

By syntenty ...



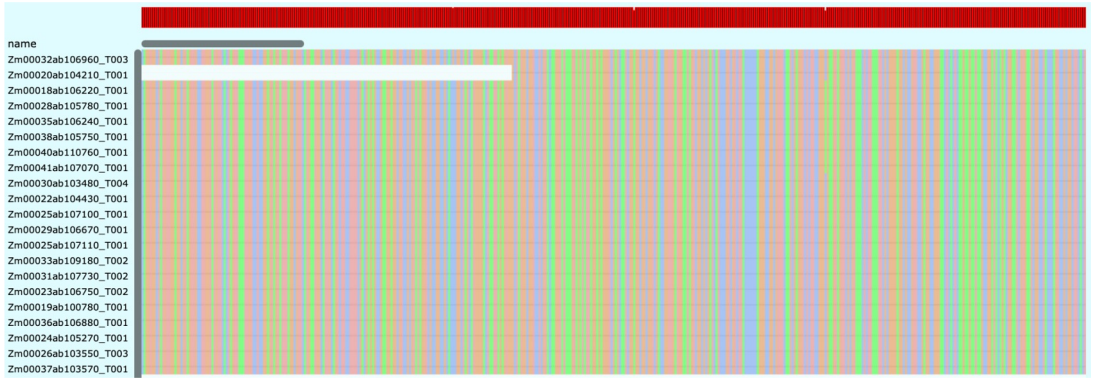
How should it be identified?

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



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

Current members
Kapeel Chougule (chair)
Sarah Dyer(Co-chair)
Ethalinda Cannon
Justin Elser
Huiting Zhang
Yogendra Khedikar
David Molik
Adam Wright
Nathan Grant

More information: https://www.agbiodata.org/node/451

Come and share your thoughts!

Join AgBioData on slack

Or email :

agbiodata@gmail.com.

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!