Guidelines for standardizing gene model nomenclature and genome assembly quality metrics

Kapeel Chougule

Ware Laboratory

Computational Science Developer

Cold Spring Harbor Laboratory, NY

PAG-30 AgBioData Workshop

Jan 13, 2023

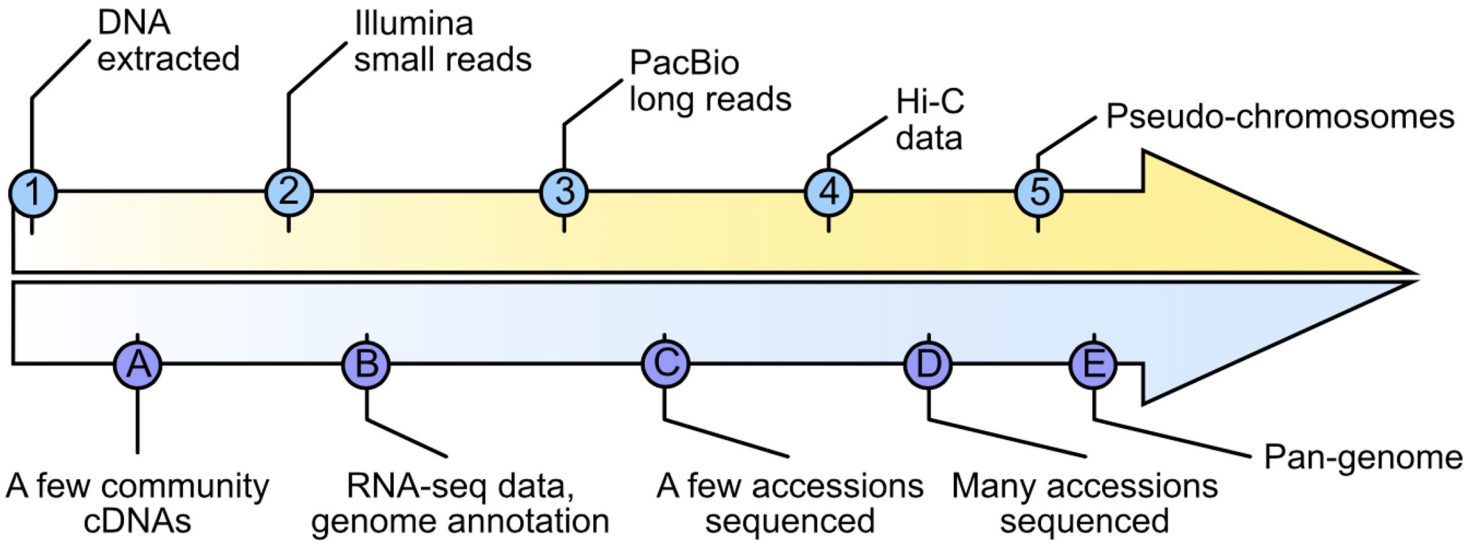
Motivation

Importance of accurate and persistent identifiers for genome 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 timeline



Glossary – before we begin:

- Gene names – e.g. *Reduced height-1*
- Gene symbols – e.g. *Rht-1*
- Gene model/locus – a genomic feature which is predicted to produce a product
- Gene model/locus ID – e.g. TraesCS4A02G271000
- Pangene – a gene model/locus predicted in all assemblies for a given species and which appears to be producing the same product

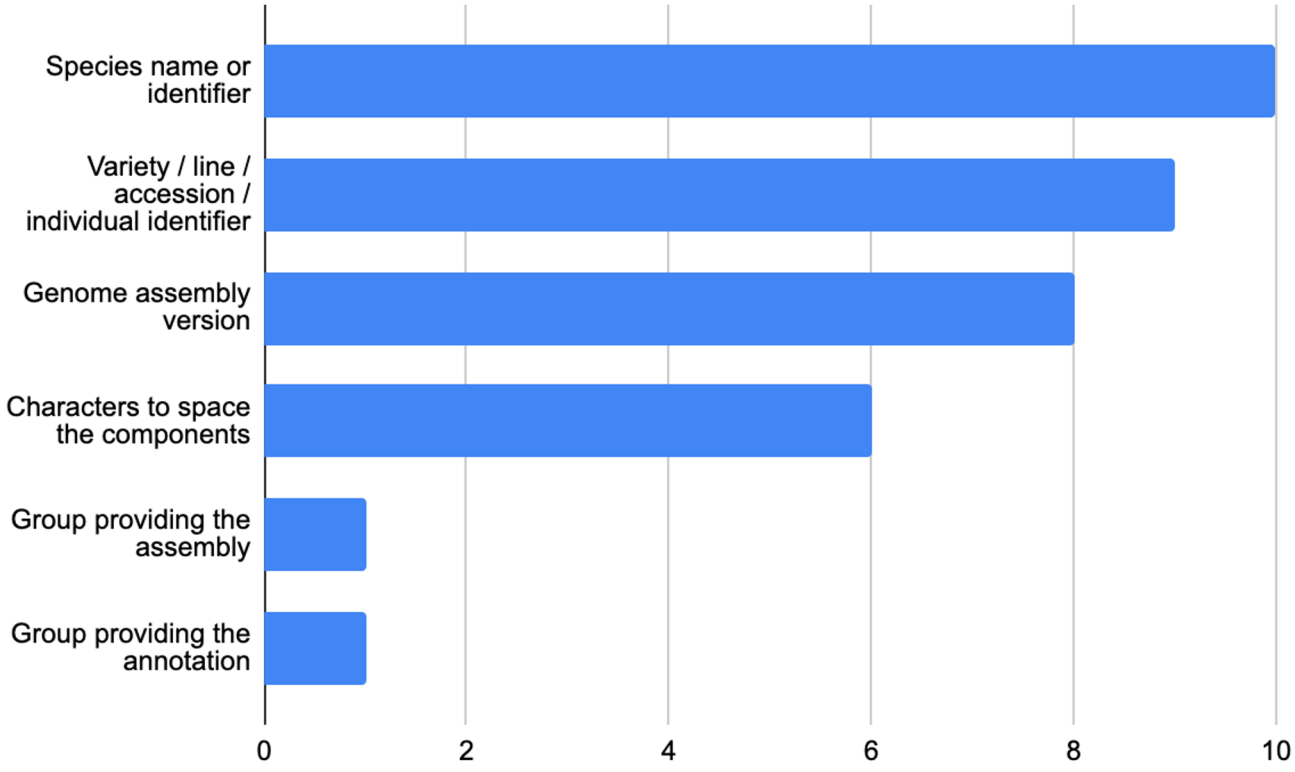
Community Survey Feedback

AgBioData Genome Assembly and Annotation Nomenclature Working Group survey

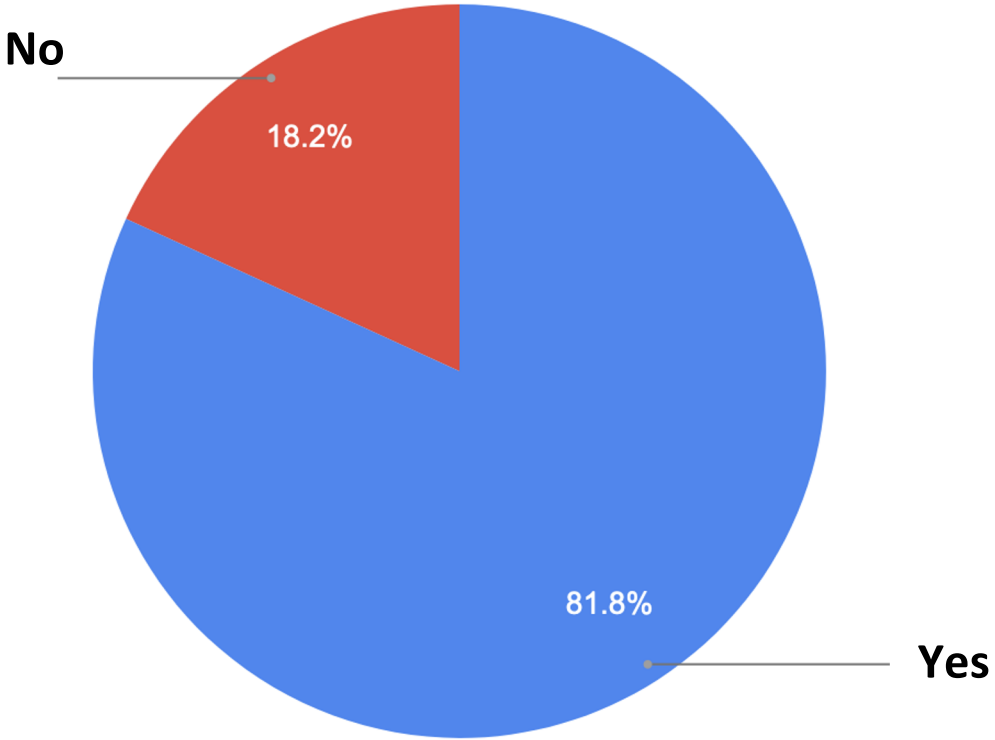
This survey is designed to
1) Gather feedback regarding genome assembly and gene model identifier naming preferences for AgBioData species
2) Explore metrics used for assessing genome assembly quality

Total 11 respondents

Which components should a genome assembly identifier include?



For your main species of interest, is there a gene model nomenclature system defined?

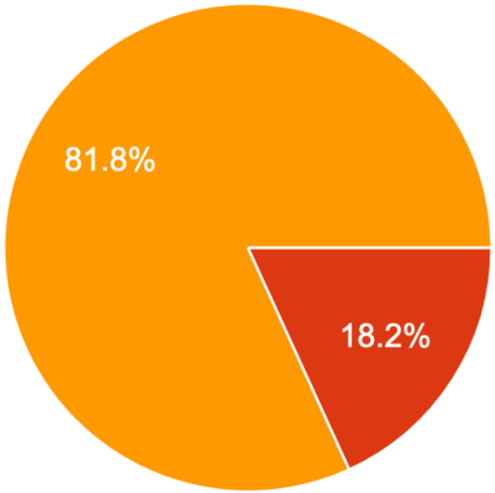


Likes and dislikes in your current nomenclature system

- Simple and easy to understand
- Difficult to deal with assembly for the same subspecies or species
- Long identifiers
- Assigning unique names and moving annotations over between versions

Should gene model identifiers be:

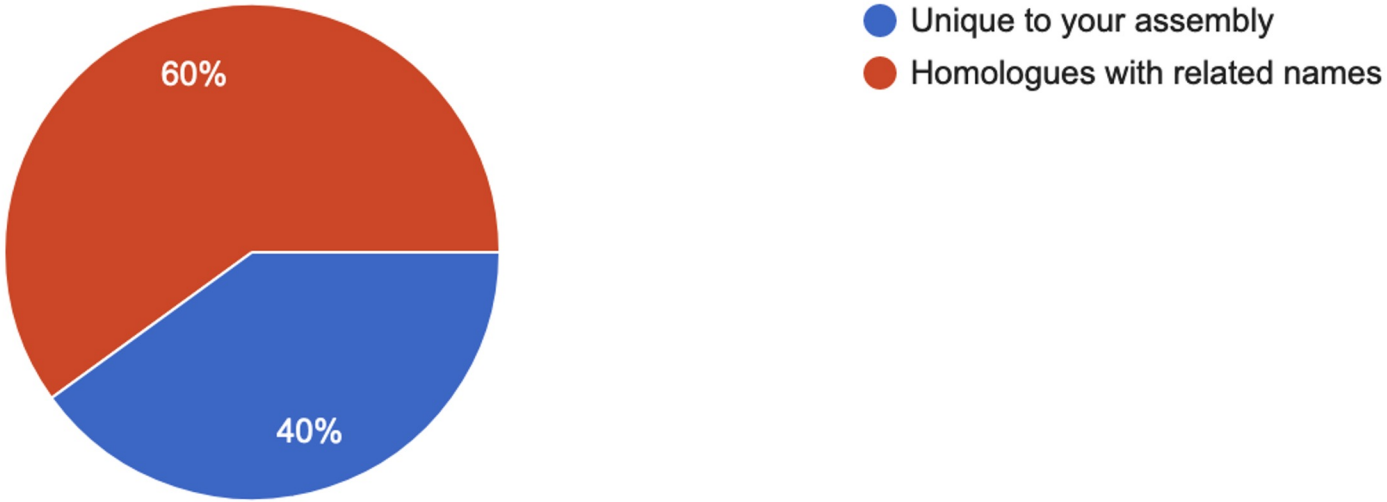
11 responses



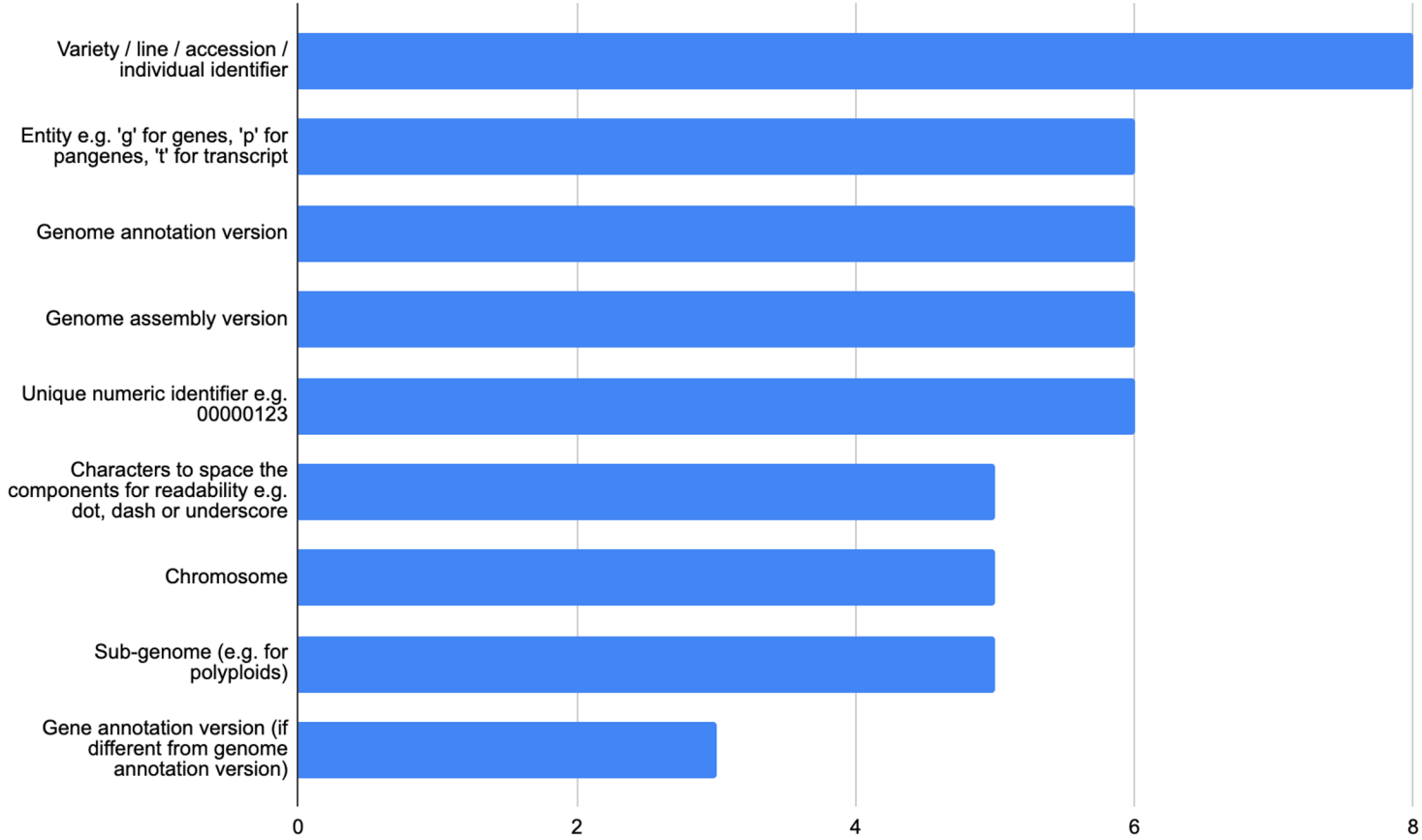
- Human readable (i.e. confer information about the gene model)
- Machine readable (e.g. a numeric representation)
- Ideally both

If you were to annotate a genome for a species where other annotated genomes are already available, would you like to develop your own independent gene model identifiers or assign identifiers based on the existing annotations? e.g. gene000001 in assembly A and assembly B would be homologues

10 responses

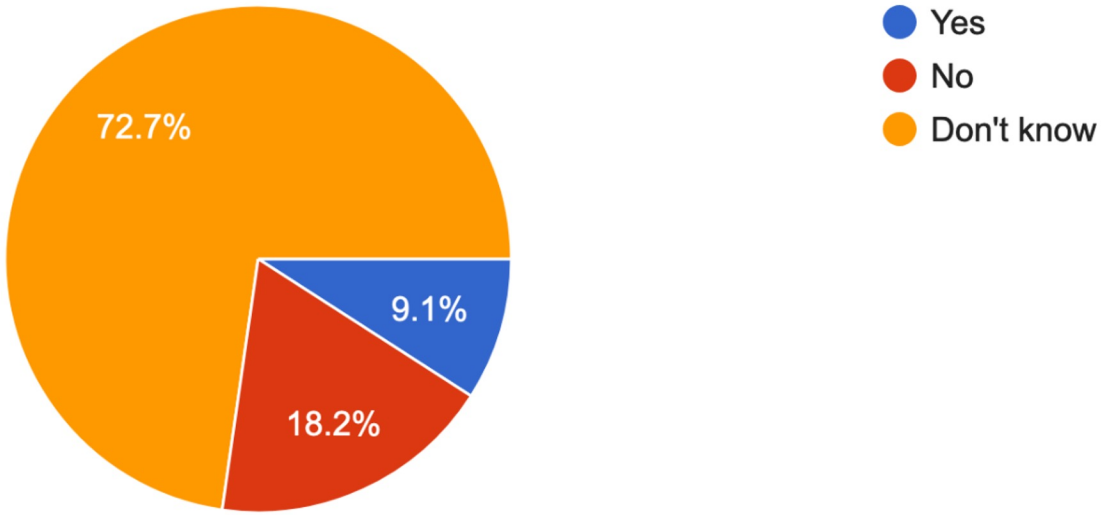


Which components should a gene model identifier include?

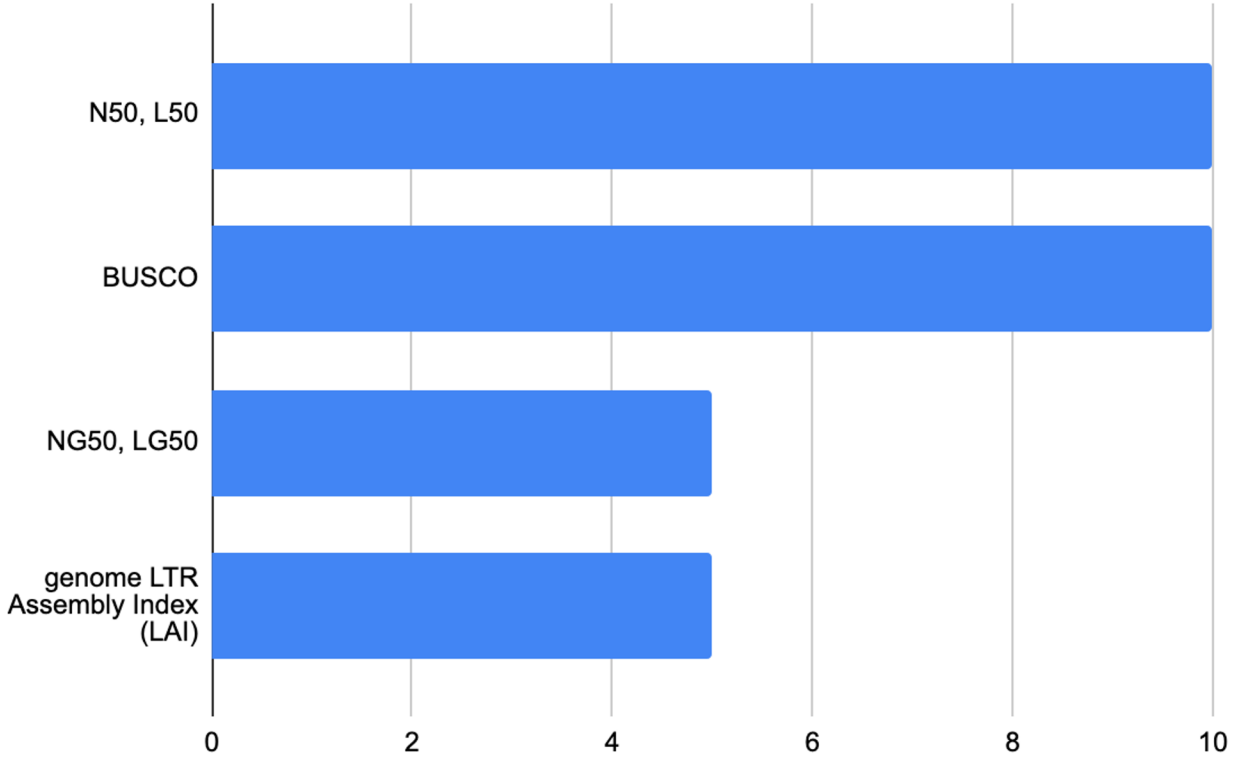


Do AgBioData databases provide adequate assembly and annotation quality metrics?

11 responses



Which of the following metrics would you like to use to help you gauge genome quality?



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:

Subgenome identifier	Chromosome identifier	Entity type e.g. gene/transcript/pangene	Entity numeric identifier (often ordered with gaps)	Annotation version

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

- human readable and accurate
- Ideally machine readable too

Gene model IDs

Examples:	Species	Assembly version	Accession	Group	Sub-genome	Chr	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			1H01	G	000300	.1
TraesCS3D02G273600	Tr aes		CS		D	3	G	273600	02
Vitvi18g12230	Vit vi					18	g	12230	
Zm00001ea036589	Z m	e	00001					036589	a

- 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

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

Commonly used assembly metrics

assembly metrics		
MaizeGDB/GenomeQC	**NCBI/GenBank**	**Others**
N50	Number of contigs	LTR assembly index(LAI)
L50	Largest contig	BUSCO
NG50	Total length	auN
LG50	Nx	Pairwise Distance
Num scaffolds	NGx	Reconstruction (PDR)
Total size of scaffolds	No. of misassemblies	etc…
Total scaff length as % of genome size	No. of misassembled contigs	
Useful scaffold sequences (>=25K nt)	Misassembled contigs length	
Longest scaffold	No. of unaligned contigs	
Shortest scaffold	No. of ambiguously mapped contigs	
Number of scaffolds > 1K nt	Genome fraction (%)	
Number of scaffolds > 10K nt	Duplication ratio	
Number of scaffolds > 100K nt	GC (%)	
Number of scaffolds > 1M nt	No. of mismatches per 100 kb	
Number of scaffolds > 10M nt	No. of indels per 100 kb	
%A		
%C		
%G		
%T		
%N		

Is N50 enough to measure assembly quality ?

How do we assess completeness of genome assembly?

What about organellar genomes?

Proposed standards & metrics in literature

Dimension	Metric	Score for a finished assembly
I. Contiguity	N50	Chromosome N50
	CC ratio[a]	1
II. Completeness	Overall completeness: *k*-mer-based	100%
	Gene space completeness: BUSCO[b]	Near 100%[c]
	Tandem repeat completeness: telomeric and subtelomeric satellite arrays, centromeric satellite arrays, ribosomal DNA loci	100%
	Complete organelle genomes	100%
III. Correctness	Base-level error rate	0%
	Structural error: collapse, inversion, false duplication, chimeric joins	0%
IV. Organellar genomes	Contiguity: (organelle contig)/(organelle genomes)	1
	Completeness	100%
	Correctness: error rate	0%
V. Heterozygosity	Contiguity: Phase block N50	Chromosome N50[d]
	Completeness	100%
	Correctness: error rate	0%

- Provides metric set for assembly evaluation
- 3C: contiguity, completeness & correctness
- A score for each metric

Wang et al, Trend in Genetics (2022)
A proposed metric set for evaluation of genome assembly quality

Proposed standards & metrics in literature

Quality Category	Quality Metric	Finished	7.C.Q50	6.7.Q40	4.5.Q30	VGP
Continuity	Contig (NG50)	= Chr. NG50	>10 Mbp	>1 Mbp	>10 kbp	1-25 Mbp
	Scaffolds (NG50)	= Chr. NG50	= Chr. NG50	>10 Mbp	>100 kbp	23-480 Mbp
	Gaps / Gbp	No gaps	<200	<1,000	<10,000	75-1500
Structural accuracy	False duplications	0%	<1%	<5%	<10%	0.2-5.0%
	Reliable blocks	= Chr. NG50	>90% of Scaffold NG50	>75% of Scaffold NG50	>50% of Scaffold NG50	2-75%
	Curation improvements	All conflicts resolved	Automated + Manual	Automated	No requirement	Automated + Manual
Base accuracy	Base pair QV	>60	>50	>40	>30	39-43
	k-mer completeness	100% complete	>95%	>90%	>80%	87-98%
Haplotype phasing	Phased block (NG50)	= Chr. NG50	>1 Mbp	>100 kbp	No requirement	1.6 Mbp*
Functional completeness	Genes	>98% complete	>95% complete	>90%	>80%	82-98%
	Transcript mappability	98%	>90%	>80%	>70%	96%
Chromosome status	Assigned %	98%	>90%	>80%	No requirement	94.4-99.9%
	Sex chromosomes	Right order, no gaps	Localized homo pairs	At least 1 shared (e.g. X or Z)	Fragmented	At least 1 shared
	Organelles (e.g. MT)	1 Complete allele	1 Complete allele	Fragmented	No requirement	1 Complete allele

- Recommendations for draft to finished qualities assemblies
- notation "**6.7.Q40**"= log-scaled **contig NG50 size**, log-scaled **scaffold NG50 size**, and the QV as **Phred-scaled base accuracy**
- "C" character to denote "complete" contigs or scaffolds that reach telomere-to-telomere continuity.

Rhie et al, bioRxiv (2020) : https://doi.org/10.1101/2020.05.22.110833

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

Come and share your thoughts!

Join AgBioData on slack

Or email :
agbiodata@gmail.com.

Acknowledgments

AgBioData

Toward enhanced genomics, genetics, and breeding research outcomes through standardization of practices and protocols across agricultural databases

USDA-ARS 8062-21000-041-00D

Genome Assembly and Annotation Nomenclature Working group

Chair: Kapeel Chougule

Co-Chair: Sarah Dyer

Members:

- Ethalinda Cannon
- Patrice Salomé
- Loren Honas
- Huiting Zhang
- Nathaniel Jue
- Paul Otyama
- Pankaj Jaiswal
- Brian Smith White
- Tara Rickman
- Maria Skrabisova
- Cecilia Deng
- Yogendra Khedikar