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

1. DNA extracted
2. Illumina small reads
3. PacBio long reads
4. Hi-C data
5. Pseudo-chromosomes

A. A few community cDNAs
B. RNA-seq data, genome annotation
C. A few accessions sequenced
D. Many accessions sequenced
E. Pan-genome
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?

Yes 81.8%

No 18.2%

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:

- 81.8% Human readable (i.e. confer information about the gene model)
- 18.2% 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?

- Variety / line / accession / individual identifier
- Entity e.g. 'g' for genes, 'p' for pangenes, 't' for transcript
- Genome annotation version
- Genome assembly version
- Unique numeric identifier e.g. 00000123
- Characters to space the components for readability e.g. dot, dash or underscore
- Chromosome
- Sub-genome (e.g. for polyploids)
- Gene annotation version (if different from genome annotation version)
Do AgBioData databases provide adequate assembly and annotation quality metrics?
11 responses

- Yes: 72.7%
- No: 9.1%
- Don't know: 18.2%
Which of the following metrics would you like to use to help you gauge genome quality?

- N50, L50
- BUSCO
- NG50, LG50
- genome LTR Assembly Index (LAI)
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

- Human readable and accurate
- Ideally machine readable too
## Gene model IDs

- **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

### Examples:

<table>
<thead>
<tr>
<th>Species</th>
<th>Assembly version</th>
<th>Accession</th>
<th>Group</th>
<th>Sub-genome</th>
<th>Chr</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>.1</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>
</tr>
<tr>
<td>Horvu_BARKE_1H01G000300.1</td>
<td>Hor vu</td>
<td>BARKE</td>
<td>1H01</td>
<td>G</td>
<td>000300</td>
<td>.1</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>
</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>
</tr>
<tr>
<td>Zm00001ea036589</td>
<td>Z m</td>
<td>e</td>
<td>00001</td>
<td></td>
<td></td>
<td></td>
<td>a</td>
<td></td>
</tr>
</tbody>
</table>
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

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

Is N50 enough to measure assembly quality?

How do we assess completeness of genome assembly?

What about organellar genomes?
## Proposed standards & metrics in literature

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

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

- 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

- **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.

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

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

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

USDA-ARS 8062-21000-041-00D