Announcements

• ASPB meeting information – please contact Marcela if you need more information!
  – Plant Bioinformatics Resources Workshop
    Sunday July 18, 2021
    4:45 PM – 6:45 PM EDT
  – Virtual booth exhibit! Anyone can attend the booth – please sign up!

• Code of conduct
  – We will send out the code of conduct for comment – you have the opportunity to provide feedback for one week
  – We are looking for two volunteers outside the SC to serve as contacts to report harassment to (ombud).

• BOSC 2021 - call for abstracts. Submission deadline is May 6th (tomorrow).

• RDA: comments on ‘Challenges of Curating for Reproducible and FAIR Research Output’ are requested until Friday, 21 May, 2021
GFF3 format recommendations

Monica Poelchau, on behalf of the AgBioData GFF3 working group
May 5\textsuperscript{th}, 2021
Poll

• Do you use GFF3 in your database/software project?
  – Yes/No

• If yes - would your database/software be willing to change the way you ingest/export GFF3?
  – Yes
  – Maybe with some modifications
  – No
  – I don’t have enough information
The GFF3 format

<table>
<thead>
<tr>
<th>Column 9</th>
<th>Contains reserved and unreserved attributes</th>
</tr>
</thead>
</table>

Pragmas/directives

<table>
<thead>
<tr>
<th>9-column format</th>
</tr>
</thead>
</table>

---

https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md
The problem

Flybase:
2L FlyBase gene 14615552 14618902 . + . ID=FBgn0000055;Name=Adh;fullname=Alcohol dehydrogenase;

Ensembl Metazoa:
2L Ensembl gene 14615552 14618902 . + . ID=FBgn0000055;Name=Adh;biotype=protein_coding

NCBI RefSeq:
NT_033779.5 RefSeq gene 14615552 14618902 . + . ID=gene-Dmel_CG3481;Dbxref=FLYBASE:FBgn0000055;Name=Adh;description=Alcohol dehydrogenase;gene=Adh;gene_biotype=protein_coding; locus_tag=Dmel_CG3481
The GFF3 working group

- **AgBioData**: Ethalinda Cannon, Andrew Farmer, Zhiliang Hu, Rex Nelson, Monica Poelchau, Surya Saha

- **Alliance of Genome Resources**: Scott Cain, Nathan Dunn, Sierra Moxon

- **NCBI**: Vamsi Kodali, Terence Murphy
Goals of this webinar

• We need your feedback on whether these recommendations are useful.
• We need to know whether you would be willing to implement them (acknowledging that they are still under development)
• We need to know what changes we can work on to make these recommendations more useful.
GFF3 working group goals

1. Ultimate goal - to use a GFF3 file from any software or any database in downstream processing tools or applications (e.g. VEP, Tripal, Apollo, ...) WITHOUT having to modify it

1. Databases and software export their GFF3 files in (a) standard way(s)
2. Databases and software know how to import standard information from a GFF3
GFF3 working group priorities

• At first:
  – get standard representation of certain data types (e.g. protein-coding genes)

• After we got started:
  – scrutinize each of the 9 columns, and the reserved attributes in column 9.
  – You’d be surprised how much discussion each column engendered…
  – Primary focus is gene structure and function
Results overview

• For each column and reserved attribute, we provide the following results from our discussions:
  – Change level: The level of change relative to the specification. Values are 'No change', 'Recommendation only', 'minor', 'moderate', 'major'
  – Summary: A summary of the GFF3 working group's findings.
  – Proposed changes to specification: A list of the proposed changes to the specification.
  – Rationale: The rationale behind these changes.
  – Best Practices: Recommended best practices for this field.
  – Validation: How software would validate whether the field is used correctly.
  – Example: An example implementation of the field.
Results overview

• Types of changes that we recommend:
  – No change: 5
  – Recommendation only: 9
  – Minor: 1
  – Moderate: 1
  – Major: 1

• We primarily have recommendations on how to:
  – interpret the specification;
  – model standard data types.
## Results overview

<table>
<thead>
<tr>
<th>Column</th>
<th>Change level</th>
<th>Attributes</th>
<th>Change level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seqid (column 1)</td>
<td>Recommendation</td>
<td>*ID</td>
<td>Recommendation</td>
</tr>
<tr>
<td>Source (column 2)</td>
<td>No change</td>
<td>Name</td>
<td>Recommendation</td>
</tr>
<tr>
<td>Type (column 3)</td>
<td>No change</td>
<td>Alias</td>
<td>Recommendation</td>
</tr>
<tr>
<td>Start, end (column 4, 5)</td>
<td>No change</td>
<td>*Dbxref</td>
<td>Recommendation</td>
</tr>
<tr>
<td>Score (column 6)</td>
<td>Moderate</td>
<td>Derives_from</td>
<td>Minor</td>
</tr>
<tr>
<td>Strand (column 7)</td>
<td>No change</td>
<td>Note</td>
<td>No change</td>
</tr>
<tr>
<td>Phase (column 8)</td>
<td>Recommendation</td>
<td>*Ontology_term</td>
<td>Recommendation</td>
</tr>
<tr>
<td>*Modeling protein-coding genes</td>
<td>Recommendation</td>
<td>Target, Gap</td>
<td>Recommendation</td>
</tr>
<tr>
<td>*Functional annotations</td>
<td></td>
<td></td>
<td>Major change</td>
</tr>
</tbody>
</table>
Results overview

- Recommendation file: https://docs.google.com/document/d/180g1rfC5n_cR6sioG_LFGaUPNmQyDqTsPafVu4gM018
- Comments are welcome in suggesting mode
- This is still a work in progress, and our recommendations still need discussion and firming up
GFF3 implementation

##gff-version 3.1.26
##gff3 implementation 1.25
##Dbxref URL: https://identifiers.org/
##score AED MAKER-P 3.0 0-1 increases
##NCBITaxon:9606
##sequence-region Scaf1 1 500000
Scaf1 MyDB gene 1 5000 . + . ID=1;gene_id=AD:ADGene001;so_term_name=protein_coding_gene; Dbxref=OD:Gene1234;
Scaf1 MyDB mRNA 1 5000 . + . ID=2;Parent=1;transcript_ID=AD:ADTrans001;Name=alcohol dehydrogenase;rank=1;go_annotations=term%3DGO:0004381%3Bevidence%3DECO:0000315;
Scaf1 MyDB CDS 1 5000 . + 0 ID=3;Parent=2;protein_ID=AD:ADProt001;
Scaf1 MyDB exon 1 5000 . + . Parent=2;
Scaf1 MyDB polypeptide 1 5000 . + . ID=3;Derives_from=3;Ontology_term=GO:0046703
Scaf1 MAKER-P mRNA 1 5000 0.38 + . ID=45221;
###
The ID attribute

• The problem: ID often does double duty as the *locally unique identifier* within the file, AND the *globally unique persistent identifier*. It is only designed to handle the former.

• The solution:
  – Use the ID attribute as the unique identifier within the file.
  – Use the following for the globally unique persistent accession number (CURIE):
    • `gene_id`, `transcript_id`, and `protein_id` (AGR), OR
    • `gene`, `transcript_id`, and `protein_id` (NCBI)
The ID attribute

ID is locally unique, maintains parent-child relationships

Globally unique, persistent accession number:
Attribute tag: gene_id
Attribute value: CURIE with a defined namespace

Scaf1  MyDB  gene  1  5000  .  + .  ID=1;gene_id=AD:ADGene001;
Scaf1  MyDB  mRNA  1  5000  .  + .  ID=2;Parent=1;transcript_ID=AD:ADTrans001;
Scaf1  MyDB  CDS  1  5000  .  + 0  ID=3;Parent=2;protein_ID=AD:ADProt001;
Scaf1  MyDB  exon  1  5000  .  + .  Parent=2;

Discussion – do we need one solution, instead?
Functional annotations

• The problem: Functional annotations, for example GO annotations, should be associated with an evidence code.

• The solution(s):
  1. Use a different file format designed for functional annotations, e.g. GPAD or GAF
  2. For simple use cases – e.g. only one program was used to generate the functional annotation – the evidence code or provenance could be specified in a pragma
  3. Adopt the Apollo complex metadata format

• Questions:
  – How will the validator/data wrangler know which is being used?
  – Should we compromise on one recommendation?
Functional annotations

##gff-version 3.1.26
##gff3 implementation 1.25
##Dbxref URL: https://identifiers.org/
##score AED MAKER-P 3.0 0-1 increases
##NCBITaxon:9606
##sequence-region Scaf1 1 500000
Scaf1 MyDB gene 1 5000. + . ID=1;gene_id=AD:ADGene001;so_term_name=protein_coding_gene;Dbxref=OD:Gene1234;
Scaf1 MyDB mRNA 1 5000. + . ID=2;Parent=1;transcript_ID=AD:ADTrans001;Name=alcohol dehydrogenase;rank=1;go_annotations=term%3DGO:0004381%3Bevidence%3DECO:0000315;
Scaf1 MyDB CDS 1 5000. + 0 ID=3;Parent=2;protein_ID=AD:ADProt001;
Scaf1 MyDB exon 1 5000. + . Parent=2;
Scaf1 MyDB polypeptide 1 5000. + . ID=3;Derives_from=3;Ontology_term=GO:0046703
Scaf1 MAKER-P mRNA 1 5000 0.38 + . ID=45221;
###

**term=GO:0004381,evidence=ECO:0000315;**
Modeling protein-coding genes

• The problem: there are many different ways to represent protein-coding genes.

• The solution:
  – Only one parent per feature
  – Child features should be listed after parent features
  – Do not list multiple values in column 1 (for features split across scaffolds)
  – Polypeptide features are not required or recommended
  – Type should be a valid SO term
GFF3 implementation recs

Use appropriate SO name

Use gene or pseudogene

Optional – use so_term_name at gene level

Don’t use a polypeptide feature

Note that this uses Derives_from
Cross-references (Dbxref)

• The problem:
  – The Dbxref should result in a resolvable URL

• The solution:
  – Specify in the pragma what identifier resource the Dbxref should point to (e.g. identifiers.org)
  – The format for a Dbxref is dbxref=database:identifier. The combination of database and identifier should exist in the identifier resource
Cross-references (Dbxref)

New pragma that can tell a validator which identifier list to build a URL from

```plaintext
##Dbxref URL: https://identifiers.org/
Scaf1 MyDB gene 1 5000 . + . ID=1;gene_id=AD:ADGene001;
Dbxref=MaizeGDB.locus:12098
```

Database string  Accession number

This information combined builds https://identifiers.org/MaizeGDB.locus:12098, which resolves to https://www.maizegdb.org/gene_center/gene/12098
Next steps

• Gather and incorporate feedback from you
• Gather feedback from additional stakeholders
• Validator development ([https://github.com/The-Sequence-Ontology/Specifications/issues/18#issuecomment-812158189](https://github.com/The-Sequence-Ontology/Specifications/issues/18#issuecomment-812158189))
• How are we going to implement this? Depends on feedback. Ideally, it would be an additional implementation standard (…) that is an extension to the existing GFF3 standard maintained by the SO – with version control.
• Add recommendations for more data types (e.g. QTL, miRNAs)
Poll

• Would your database/software be willing to change the way you GFF3 ingest/export based on our guidelines?
  – Yes
  – Maybe with some modifications
  – No
  – I don’t have enough information
Thank you!

• Maggie Woodhouse for initiating this effort
• All initial discussion and working group participants: Maggie Woodhouse, Daniel Lawson, Jeongwoon Kim, Keith Decker
• All reviewers