AgBioData GFF3 discussion

March 4th, 2020

Upcoming AgBioData schedule:

- April 8th (NOTE: Second Wednesday of the month)
  - Guest speaker: Dr. Madha Devare from CGIAR talking about the Gardian platform
- May 6th
  - Group discussion: Pan-genomes
    - A number of people have asked for a discussion about pan-genome visualization
- June:
  - Guest Speaker: Dr. Julie Dunning Hotopp talking about secondary data usage and analysis
    - (2018 Research Parasite winner)

Agenda

- Brief introduction:
  - What is the gff3 format? Is it problematic?
- Round table/examples
  - Scott Cain; Chris Elsik; Andrew Farmer; Nathan Weeks; Maggie Woodhouse; Monica Poelchau; Philip Bayer; insert your name here
- Discussion:
  - Are there common issues that AgBioData could work together to fix?
    - E.g. agree on best practices to solve a particular problem
    - Page on AgBioData website with tools to fix common gff3 problems?
- Gathering document for notes:
  - https://docs.google.com/document/d/1B7QGGIWGM9u8EbCD9jeuBEM72avPOFE0_F4Ow7qwpbI/edit?usp=sharing

The GFF3 format

- The specification: https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md
  - Although there are many richer ways of representing genomic features via XML and in relational database schemas, **the stubborn persistence of a variety of ad-hoc tab-delimited flat file formats declares the bioinformatics community's need for a simple format that can be modified with a text editor and processed with shell tools like grep**.
  - The GFF format, although widely used, has fragmented into multiple incompatible dialects.
  - When asked why they have modified the published Sanger specification, bioinformaticists frequently answer that the format was insufficient for their needs, and they needed to extend it.
  - The proposed GFF3 format addresses the most common extensions to GFF, while preserving backward compatibility with previous formats.

The GFF3 format

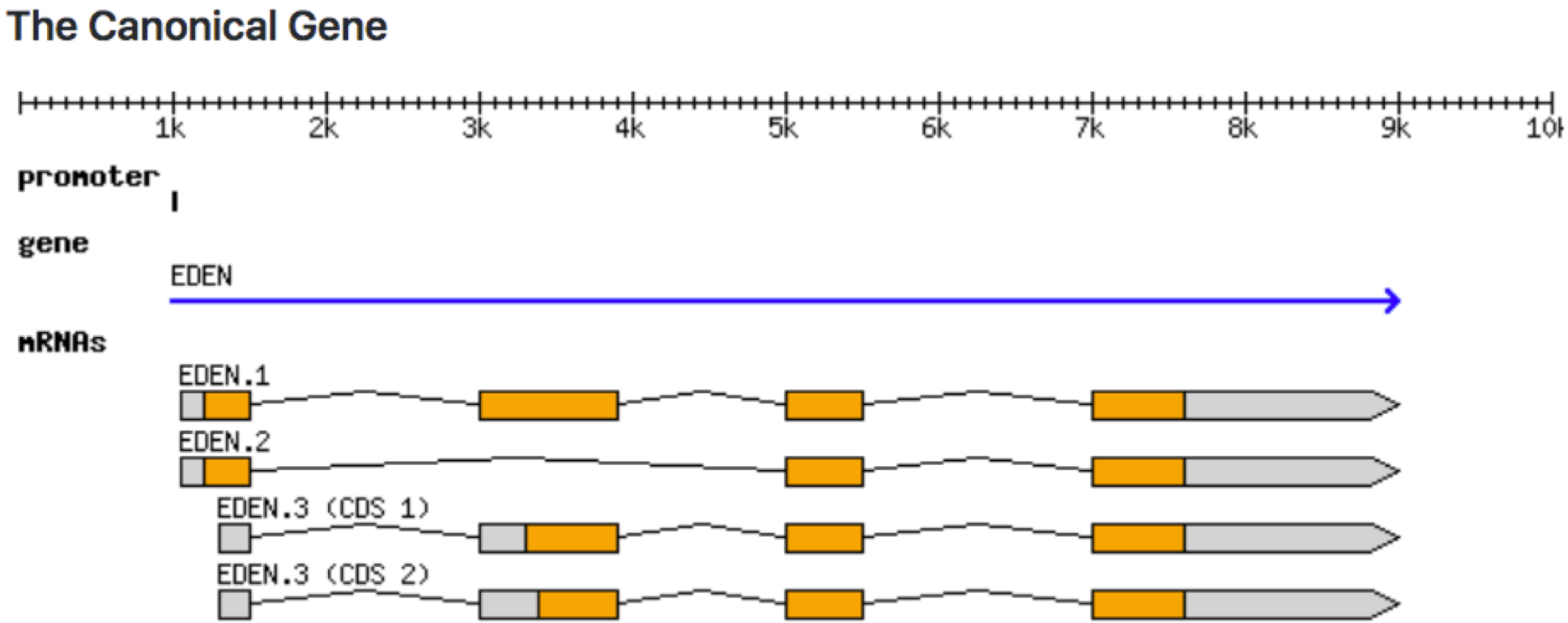

FIGURE 1

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

The GFF3 format

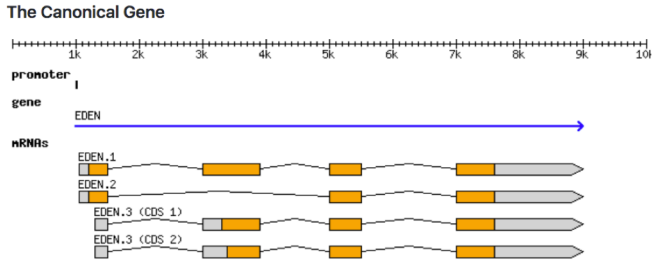

FIGURE 1

```
 0  ##gff-version 3.2.1
 1  ##sequence-region ctg123 1 1497228
 2  ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
 3  ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
 4  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
 5  ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
 6  ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
 7  ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
 8  ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
 9  ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
10  ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11  ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12  ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13  ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14  ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15  ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16  ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17  ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18  ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19  ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20  ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21  ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22  ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23  ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24  ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

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

Common issues

- Interoperability
  - Ingesting a gff3 into your database
  - Using another group's gff3 in a standard tool for your database
  - Providing gff3 files for your users that they can combine with other databases' gff3 files without reformatting
  - My pet peeve – not interpreting the CDS phase column correctly can break the protein coding sequence
- Fixing broken gffs
  - Easy to manipulate, therefore easy to break!
- There are 2 types of annotation represented in gff3 files; these may require 2 separate discussions
  - Structural annotations
  - Functional annotations

Examples from AgBioData members

- Scott Cain; Chris Elsik; Andrew Farmer; Nathan Weeks; Maggie Woodhouse; Philip Bayer; Monica Poelchau; and YOU!

Comments from Philipp Bayer

Forrest Fellow

University of Western Australia

AUGUSTUS and exonerate

- AUGUSTUS' gff output is weirdly non-standard and many downstream tools like EvidenceModeler crash.
  - This script fixes that: https://github.com/jorvis/biocode/blob/master/gff/convert_augustus_to_gff3.py
- Newer AUGUSTUS versions are even worse, they put the authors' names into the gff3 comment section
  - one guy has an umlaut in his name and instead of typing 'oe' they put in the umlaut, which crashes the above script
- Same goes for the exonerate gff: convert_exonerate_gff_to_std_gff.pl

Broken gff files

- Many, many plant genomes have broken gff files. The first B. napus Darmor-bzh genome for example has a weird gff file missing the ‘gene’ rows
  - https://www.biostars.org/p/275813/
  - Nobody ever fixed that because most people working on it moved on to other positions

I5k Workspace example issues

- QA/QC
  - Of user-submitted gff3 files
  - Of user-curated gene models (e.g. from Apollo)
  - Things are mainly challenging when the files are modified by a human and not a program
- We created a python toolkit to deal with some (not all) of the ways that gff3s can break, especially with manual annotation
  - https://github.com/NAL-i5K/gff3toolkit
- NCBI submission of gff3 files

Discussion – possible solutions

- Standardizing the representation of the most common structural and functional information
  - Expand SO's 'canonical gene model'?
  - Adopt NCBI's gff3 submission standard?
- Maintain a page with tools to solve common gff3 formatting problems?

GFF3 Standards at the Alliance of Genome Resources

Scott Cain
scott@scottcain.net
GMOD project coordinator
WormBase senior developer
AGR genome features working group lead
March 4, 2020

ALLIANCE of GENOME RESOURCES

GMOD

The Alliance of Genome Resources (AGR)

An NIH funded project to create a single resource for identifying model resources for human health research. Composed of:

- Yeast (SGD)
- Fruitfly (FlyBase)
- Worm (WormBase)
- Zebrafish (ZFIN)
- Mouse (MGI)
- Rat (RGD)
- Gene Ontology

ALLIANCE of GENOME RESOURCES

GMOD

Organization/Working Groups

Progress is primarily made through working groups centered around a single aspect of the resource, like

- Anatomy
- Expression
- User Interface
- Data wrangling (this was were most of the standard dev work happened)
- Genome features (JBrowse, in page widgets)

ALLIANCE of GENOME RESOURCES

Attempt at a standard

At first, protein coding genes only:

- First agree on how to represent the central dogma (in SO terms):
  - gene (NOT is_a children like protein_coding_gene)
    - mRNA (NOT other transcript types)
      - CDS (required)
      - exon (optional)
      - three_ and five_prime_utr (optional)
- No restriction on column 2 (source)

ALLIANCE of GENOME RESOURCES

GMOD

Other standard items

Required tags:

- Required tag for gene features: curie: a resource-wide universal identifier, eg curie=WB:WBGene00023193
- CDS features can be grouped using only a Parent tag but in the case where more than one CDS derives from a transcript, explicit ID attributes should be used to group them.
- A comment near to the top of the document should indicate the build that the GFF file derives from, eg `# Genome build: GRCm38-C57BL/6J`

ALLIANCE of GENOME RESOURCES

GMOD

Generalizing to non-coding genes

- All gene features still have SO type “gene” but have a transcript type that corresponds to the type of gene it is (like tRNA, pre_miRNA).
- Required ninth column tag: so_term_name where the value is that of the SO term name that is most correct for the gene (tRNA_gene, miRNA_gene, protein_coding_gene). When a gene has both coding and non-coding transcripts, the value should be protein_coding_gene.
- Except as noted elsewhere, there are no restrictions on tag/value pairs in the ninth column--they will be ignored.

ALLIANCE of GENOME RESOURCES

GMOD

Positive results

- Building consistent JBrowse instances for all species is “easy”
  - Made easier still by dockerizing our JBrowse server
  - https://github.com/alliance-genome/agr_jbrowse_container
  - https://hub.docker.com/r/gmod/
- A written standard:
  https://docs.google.com/document/d/1yjQ7lozyETeoGkPfSMTAT8IN3ZIAuy5YkbsBdjGeLww/edit?usp=sharing
- At the moment, that’s mostly it. GFF is not an input format for the Alliance data store; it’s only used to drive JBrowse and the in-page widget (via a server side Apollo-driven translator)

ALLIANCE of GENOME RESOURCES

GMOD

Downsides

- Not much. Minor growing pains, some confusion initially between groups as the standard developed.
- No validator yet--everybody agrees having one would be good, nobody has the time to write it.
- Every MOD had to spend some time rewriting their GFF creation scripts, some more than others.

ALLIANCE of GENOME RESOURCES

GMOD

Acknowledgements

Members of the Genome Features working group:

- Nathan Dunn (Apollo, GO)
- Paul Hale (MGI)

Members of the Data Quartermasters working group:

- Sierra Moxon (Leader extraordinaire, ZFIN)
- Kevin Howe (emeritus, WormBase)
- Stacia Engle (emeritus, SGD)
- Joel Richardson (MGI)
- Jennifer Smith (RGD)
- Chris Tabone (emeritus, FlyBase)

ALLIANCE of GENOME RESOURCES