GFF3 Format

The official documentation for the GFF3 format can be found here

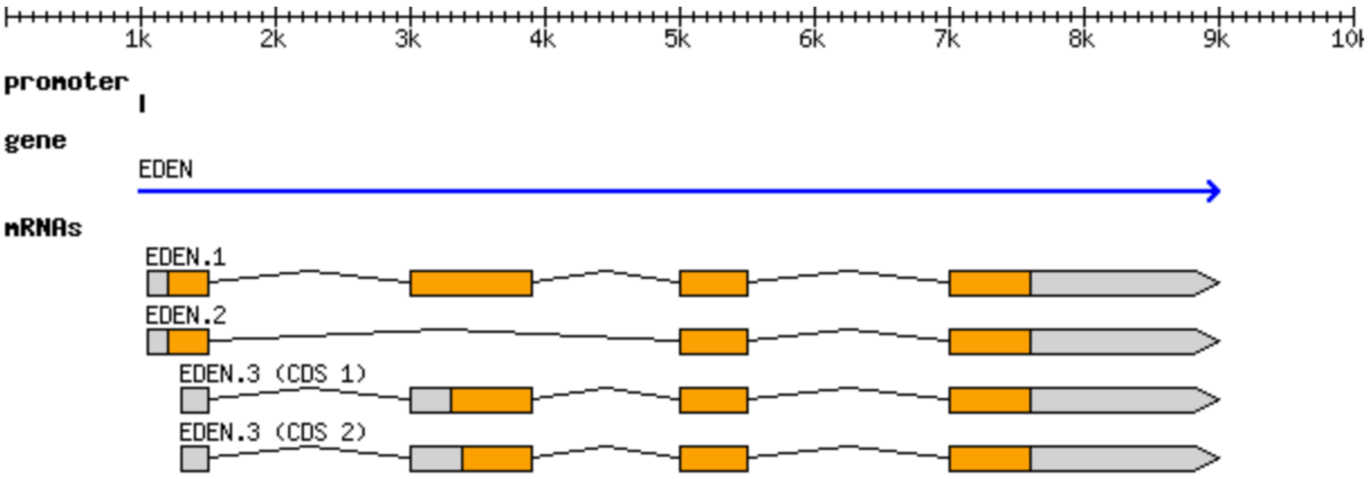
General Feature Format (GFF) is a tab-delimited text file that holds information concerning any and every feature that can be applied to a nucleic acid or protein sequence. Everything from CDS, microRNAs, binding domains, ORFs, and more can be handled by this format. Unfortunately there have been many variations of the original GFF format and many have since become incompatible with each other. The latest accepted format (GFF3) has attempted to address many of the issues that were missing from previous versions.

GFF3 has 9 required fields, though not all are utilized (either blank or a default value of '.').

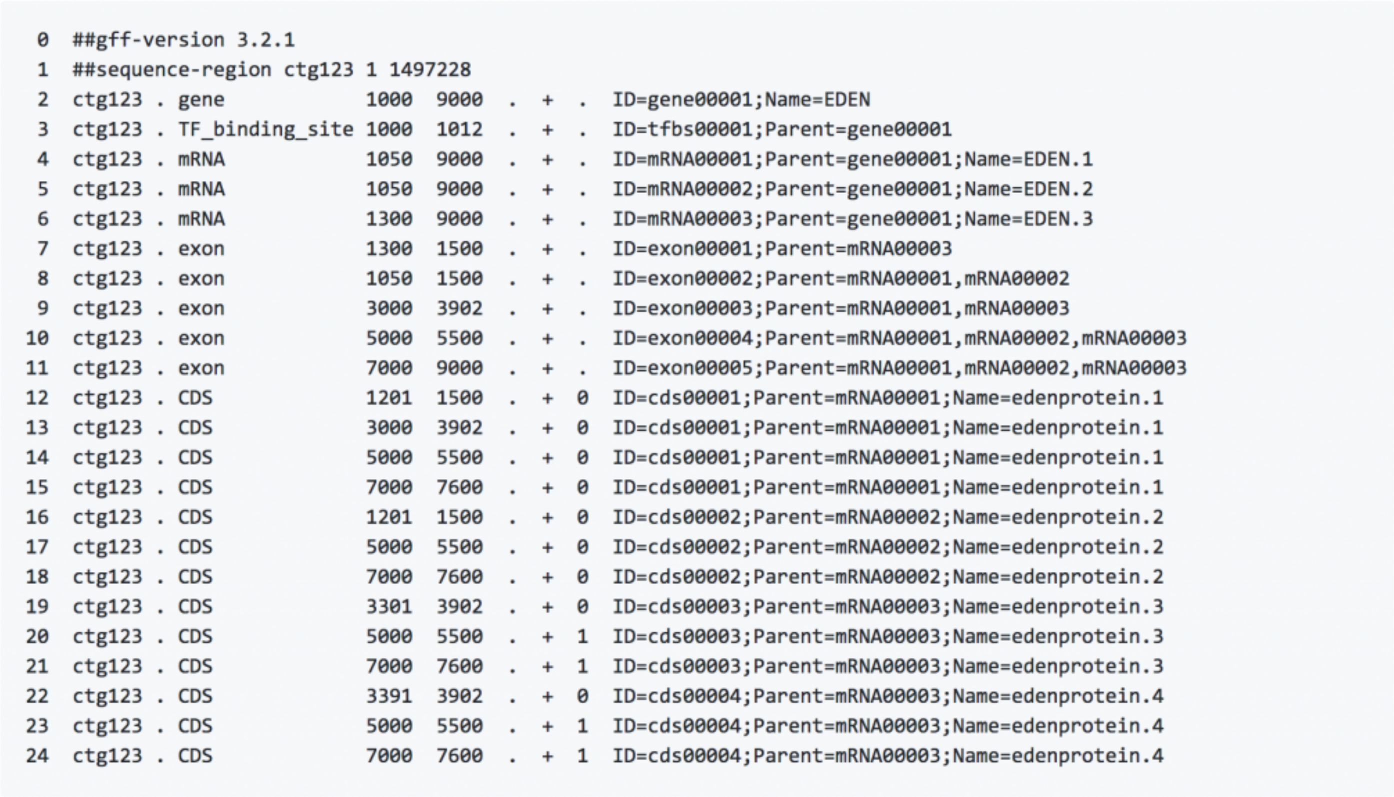
1. Sequence ID
2. Source
   - Describes the algorithm or the procedure that generated this feature. Typically Genescane or Genebank, respectively.
3. Feature Type
   - Describes what the feature is (mRNA, domain, exon, etc.).
   - These terms are constrained to the [Sequence Ontology terms](http://www.sequenceontology.org/).
4. Feature Start
5. Feature End
6. Score
   - Typically E-values for sequence similarity and P-values for predictions.
7. Strand
8. Phase
   - Indicates where the feature begins with reference to the reading frame. The phase is one of the integers 0, 1, or 2, indicating the number of bases that should be removed from the beginning of this feature to reach the first base of the next codon.
9. Atributes
   - A semicolon-separated list of tag-value pairs, providing additional information about each feature. Some of these tags are predefined, e.g. ID, Name, Alias, Parent . You can see the full list [here](https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md).

GFF3 Example

The canonical gene can be represented by the following figure



https://github.com/The-Sequence-Ontology/Specifications/blob/master/gff3.md
The same information can be represented in GFF3 format:



```
##gff-version 3.2.1
##sequence-region ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . CDS             1201  1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             3000  3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             7000  7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             1201  1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             5000  5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             7000  7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             3301  3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             5000  5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             7000  7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS             3391  3902  .  +  0  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             5000  5500  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS             7000  7600  .  +  1  ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

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

What Software uses GFF3?

Any tool that requires information about gene position for analysis such as:

- Mapping RNA-seq such as [Tophat](https://ccb.jhu.edu/software/tophat/index.shtml), [HTSeq](https://htseq.readthedocs.io/en/release_0.9.1/)
- Genome Browsers like [IGV](http://software.broadinstitute.org/software/igv/), [Gbrowse](http://gmod.org/wiki/GBrowse), [UCSC](https://genome.ucsc.edu/)

How is this file generated?

- Feature identification software report motifs/features in this format.
- Almost all sequence annotation databases report in this format.