GTF (General Transfer Format)

Bradford Condon

2017-05-24

This post is part 3 of a series on file formats, written for the 2017 UK-KBRIN Essentials of Next Generation Sequencing Workshop at the University of Kentucky. The conference website is hosted [here](http://www.endophyte.uky.edu/ngs/).

General transfer format (GTF), also known as General Feature Format (GFF) 2.0, is the format for transcripts in exercise 4, **RNAseq**. For more details, please see the [ensembl guide to GFF](www.ensembl.org/info/website/upload/gff.html).

GTF/GFF2.0 is fairly straightforward. Understand that is can describe a variety of sequence features, not just transcripts. All components of a gene's structure, such as introns, exons, and protein CDS, can be described with GTF using the correct feature type (column 3).

## Format

Each feature takes up one line, with **9** columns per line (plus optional track definition lines).

### Columns

1. **seqname** Chromosome or scaffold name.
2. **source** Database, project, or program name.
3. **feature** Feature type (eg Gene)
4. **start** Start position
5. **end** End position
6. **score**
7. **strand** + (forward) or - (reverse)
8. **frame** 0, 1, 2
9. **attribute** Tag "value" ; tag "value"

1 transcribed\_unprocessed\_pseudogene gene 11869 14409 . + . gene\_id "ENSG00000223972"; gene\_name "DDX11L1"; gene\_source "havana"; gene\_biotype "transcribed\_unprocessed\_pseudogene";

In the above sample GTF file from ensembl, for example, the name is *(chromosome) 1*, the source is *transcribed\_unprocessed\_pseudogene*, the feature type is *gene*, the start-end is *11869-14409*, there is no applicable score (.), it's on the forward (+) strand, there is no frame data, and there are four attributes. The first attribute, for example, is gene\_name with the value *DDX11L1*.

### Track lines

The optional track lines start with the word 'track', followed by space-seperated key=value pairs.