Explanation of *Hydractinia* RNA annotation GFF files

There are eight GFF files:

Four are for the **Hydractinia echinata** genome:

- `hech.trna.detailed.gff`: tRNA predictions, with all metadata.
- `hech.trna.minimal.gff`: tRNA predictions, without metadata.
- `hech.trna.detailed.gff`: Rfam predictions (all RNAs except tRNA), with all metadata.
- `hech.trna.minimal.gff`: Rfam predictions (all RNAs except tRNA), without metadata.

And four for the **Hydractinia symbiolongicarpus** genome:

- `hsym.trna.detailed.gff`: tRNA predictions, with all metadata.
- `hsym.trna.minimal.gff`: tRNA predictions, without metadata.
- `hsym.trna.detailed.gff`: Rfam predictions (all RNAs except tRNA), with all metadata.
- `hsym.trna.minimal.gff`: Rfam predictions (all RNAs except tRNA), without metadata.

Tab-delimited columns in `detailed` and `minimal` files:

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

1. `seqid`: sequence name
2. `source`: prediction method (`tRNAscan-SE` or `cmsearch`)
3. `type`: model (`tRNA-<isotype-model>`, or Rfam family accession)
4. `start`: start coordinate of prediction (always `<= end`)
5. `end`: end coordinate of prediction (always `>= start`)
6. `score`: bit score of prediction
7. `strand`: `+` if on positive strand, `-` if on negative strand
8. `phase`: uninformative in these files, always `.`
9. `attributes`: one or more `<key>=<value>;` strings, with possible `<key>`s listed below with explanations

---

Explanation of `<key>=<value>;` strings in `attributes` column:

Below are explanations of the `<value>` for each possible `<key>` you will find in the `attributes` column, separated into categories:

Software/database version:

- `tRNAscan-SE`: version of tRNAscan-SE used, always `2.0.5` when present, only present in `trna` files
- `Infernal`: version of Infernal used, always `1.1.2` when present, only present in `rfam` files
- `Rfam`: Rfam release used, always `14.1` when present, only present in `rfam` files

These three keys are the only keys present in the ***`attributes`*** *column in the* ***`minimal`*** *files, they are also present in the* ***`detailed`*** *files.*

Target sequence:

- `seqlen`: full length in nucleotides of target sequence

Data from cmsearch `--tblout` output file:

- `evalue`: E-value of hit, in search of full genome *H. echinata* or *H. symbiolongicarpus* v1.0 assembly sequence file
- `mdlaccn`: Rfam model accession
- `mdlcoords`: `<mdlstart>-<mdlend>` model start and end positions of hit

Data from tRNAscan-SE `-o` output file:

- `type`: isotype model (tRNAscan-SE `tRNA type` column)
- `anticodon`: anticodon (tRNAscan-SE `Anticodon` column)
- `ibegin`: beginning position of intron, `0` for none (tRNAscan-SE `Intron Bounds - Begin` column)
- `iend`: ending position of intron, `0` for none (tRNAscan-SE `Intron Bounds - End` column)
- `pseudo`: `yes` if flagged as pseudogene, else `no` (`yes` if tRNAscan-SE `Note` column contains `pseudo`)

Repeat overlaps:

- `repeat_overlap`: `HydSINE1(<start>-<stop>` if this prediction overlaps by 10 or more nucleotides with a RepeatMasker HydSINE1 prediction from `<start>` to `<stop>` (`<start> <= <stop>`), else `no`

Note: only tRNA/HydSINE1 repeat overlaps are reported in this way. No overlaps between other repeat models and tRNAs are reported, nor any overlaps between repeat models and Rfam models.

Hit score data:

- `scF`: `score/<max-score>`, where `max-score` is the maximum `score` in the file for this `model`
- `highscoring`: `yes` if `scF >= 0.9`, else `no`

Hit length data:

- `lenF`: `<length>/<max-length>`, where `<length>` is `stop-start+1`, and `<max-length>` is the `<length>` for the prediction with the maximum score in the file for this `model`
- `fragment`: `no` if `lenF >= 0.9`, else `yes`

Tandem array data:

- `tandem_array`: `yes(<start>-<stop>)` if this prediction is in a tandem array that begins at position `<start>` and ends at position `<stop>` on strand `strand` (`<start> <= <stop>`); `ineligible` if this prediction occurs on a sequence and strand with less than 10 predictions for this `model` and is consequentally not eligible to be in a tandem array; `no` if this prediction is eligible to be in a tandem array and is not;
- `tandem_array_X`: number of predictions in the tandem array this prediction is in, *only exists if* ***`tandem_array`*** *value starts with* ***`yes`***
- `tandem_array_N`: number of spacings between predictions in the tandem array this prediction is in that are within the range of `[tandem_array_Dmin..tandem_array_Dmax]`, *only exists if* ***`tandem_array`*** *value starts with* ***`yes`***
- `tandem_array_Dmin`: minimum spacing length for the tandem array this prediction is in *only exists if* ***`tandem_array`*** *value starts with* ***`yes`***
- `tandem_array_Dmax`: maximum spacing length for the tandem array this prediction is in *only exists if* ***`tandem_array`*** *value starts with* ***`yes`***

Contact eric.nawrocki@nih.gov with questions