Pseudogene Flat-File Format

The pseudogene information will be split between two files:

- gtf files (conforming to the gtf format)
- pgn files (our own creation)

In both file types, all fields are tab-delimited and fields to be left blank contain a single period.

gtf files: Tab delimited, "." used to denote a blank field. Note that some seemingly unnecessary fields are required to conform to the gtf format specification.

1. **Chromosome number** (e.g. chr22)
2. **Source**: Group / program that found the feature ("pseudogene.org" if found by us).
3. **Feature**: "pseudogene".
4. **Pgene start**: coordinate of starting base on the chromosome (relative to the positive strand, counting from 1).
5. **Pgene end**: coordinate of final base on the chromosome (relative to the positive strand, counting from 1).
6. **Score**: "." – will be used to record our "confidence" in the pseudogene, once we have a scoring system worked out.
7. **Strand**: "+" or "-" to indicate the strand.
8. **Frame**: "." – not relevant to us right now.
9. **Other information**: this field contains the following information, (using field name, and deliminated by semi-colons).
   (a) *pgene_type processed* or *pgene_type nonprocessed*.
   (b) *protein X* where X is the Swissprot/TrEmbl accession number for the protein from which this pgene was derived.
   (c) *exons* $X : S_1 - F_1 : S_2 - F_2 : \ldots : S_X - F_X$ where X is the number of exons, S_i is the starting coordinate of the ith exon, and F_i is the final coordinate of the ith exon.
   (d) *name: <Swissprot protein name>.<organism>.mb<rounded megabase position>.kb<rounded kilobase position>.<pgene accession number>.v<version number>*. Will substitute the sprot name with the protein acc if the sprot name is not known.
   (e) *alternate_name: <protein acc>.mb<organism>.kb<rounded megabase position>.<rounded kilobase position>.<pgene accession number>.v<version number>*.
   (f) *pgene_id: PGO.<GenBank Org. Taxanomy Number>.<sequentially assigned unique id>*. Unique to each pseudogene version.

(g) *transcript_id: PGO.<GenBank Org. Taxanomy Number>.<sequentially assigned unique id>*. Same as pgene_id – necessary to conform to gtf specifications.

Sample line: chr1 <tab> pseudogene.org <tab> pseudogene <tab> 166209 <tab> 166358 <tab> . <tab> - <tab> . <tab> pgene_type processed; protein P02404; exons 1:166209-166358; name RL39_HUMAN.human.m0.k166.1.v1; alt_name P02404.human.0.166.1.v1; gene_id PGO.9606.1; transcript_id PGO.9606.1

pgn files: Records all other information we have on the pgenes. Three lines per entry:

- **Line 1**: Information described below.
- **Line 2**: The predicted amino acid sequence of the pseudogene, aligned to protein (i.e. contains gap, stop codon and frame-shift symbols).
- **Line 3**: The amino acid sequence of the protein, aligned to predicted pgene sequence (i.e. contains gap, stop codon and frame-shift symbols).

First Line:

1. **pgene_id**: PGO.<GenBank Org. Tax. Number>.<unique id>. Same as from gtf file.
2. **Chromosome** (e.g. "chr19")
3. **Query start**: Coordinate of first protein amino acid in alignment (relative to protein, counting from 1).
4. **Query end**: Coordinate of last protein aa in alignment.
5. **E-value**: Expect value of the pseudogene fragment in the TBLASTX search.
6. **AA_ident**: amino acid sequence identity between the pseudogene and query protein.
7. **DNA_ident**: nucleotide sequence identity between the pseudogene and the query protein, coding region only. Some query proteins don't have coding sequence available.
8. **Cytogenic_band**: chromosomal band as predicted by Ensembl.
9. **Completeness**: sequence completeness of the pseudogene compared with the query protein.
10. **Polya**: "0" or "1" or "2" or "3".
    - **0**: No polyA tail (> 30 A in 50 bp window) detected of the pseudogene.
    - **1**: Has polyA tail and also polyadenilation signal with 50 bp of the beginning of the tail.
    - **2**: Has polyA tail and polyadenilation signal within 50-100 bp of the beginning of the tail.
    - **3**: Has polyA tail but no polyadenilation detected.

11. **Disable**: "0" or "d" or "h".
    - **0**: Indicates no disablement (only for RP pseudogenes).
    - **d**: Indicates disablement in a region of low sequence identity.
    - **h**: Indicates disablement in region of high sequence identity.
12. **GC_Pgene**: GC content of the pseudogene sequence.
13. **GC_Isochore**: GC content of the 100K bp window on the chromosome.
14. **Isochore_class**: isochore class where the pseudogene resides. L1, L2, HJ1, H2 H3
15. **Kimura_Distance**: Evolution distance of the pseudogene sequence from the present day sequence.
16. **Comment**: cytoplasmic ribosomal protein pseudogenes are labeled as "RP".
17. **MIM**: Entry of the query protein in the MIM database (Mendelian Inheritance in Man).

Sample first line: PGO.9606.1 <tab> chr1 <tab> 17 <tab> 128 <tab> 1e-24 <tab> 0.634 <tab> 0.82 <tab> 1p36.330.88 <tab> 3 <tab> h <tab> 0.57 <tab> 0.61 <tab> H3 <tab> 0.2092 <tab> . <tab> 603840