Cerberus annotation file specifications (Mouse)

Prepared by Fairlie Reese
January 25, 2022
Mortazavi Lab, University of California, Irvine

Contact Information

Fairlie Reese
2300 Biological Sciences III
University of California Irvine
Irvine, CA 92697-2300
Telephone: (949) 824-8393
Email: freese@uci.edu

Software used

- cerberus (https://github.com/fairliereese/cerberus/releases/tag/v1) v1

Cerberus annotation object

After installing cerberus (from the above link), one can read in the annotation object using the following Python code

```
import cerberus
ca = cerberus.read(<cerberus annotation>.h5)
```

Fields of the CerberusAnnotation object

TSS / TES tables

Accessible using `ca.tss` and `ca.tes` respectively. These tables hold information of each TSS or TES region used to form the cerberus reference. File format roughly follows BED format with the following column specifications:

- Chromosome
- Start
- End
- Strand

- Name
    - Formed from the gene ID that the region is associated with and a number to make it unique. For example, ENSG00000000460_1 or ENSG00000000460_2
- source
    - Comma-separated list of sources that support the use of each region as a TSS or TES that were used as input into cerberus
- novelty
    - Novelty of region with respect to the sources used as references in the cerberus run. Will be either ‘Known’ or ‘Novel’
- gene_id
    - Gene ID that the region is associated with
- tss or tes
    - Number that is concatenated with gene_id to form the Name column

IC table

Accessible using `ca.ic`. This table holds information about each intron chain used to form the cerberus reference. File format is as follows:

- Chromosome
- Strand
- Coordinates
    - Hyphen-concatenated list of intron starts and end coordinates
    - Coordinates do NOT include the start of the first exon or end of the last exon
    - Monoexonic transcripts have a single hyphen as their coordinates
- Name
    - Formed from the gene ID that the IC is associated with and a number to make it unique. For example, ENSG00000000460_1 or ENSG00000000460_2
- source
    - Comma-separated list of sources that support the use of each IC that were used as input into cerberus
- novelty
    - Novelty of IC with respect to the sources used as references in the cerberus run. Will be one of the following: ‘Known’, ‘Unspliced’, ‘NNC’, ‘NIC’, ‘ISM’
    - Novelty categories based on terminology coined by SQANTI: https://github.com/ConesaLab/SQANTI
- gene_id
    - Gene ID that the IC is associated with
- ic
    - Number that is concatenated with gene_id to form the Name column

TSS / TES maps

Accessible using `ca.tss_map` or `ca.tes_map`. This table has information from the `cerberus agg_tss or agg_tes` run about which cerberus region each input region was assigned to. File format is as follows:

- Chromosome
- Start
- End
- Strand
- source
    - Name of the source that the input region came from
- Name
    - Name of the TSS or TES that the input region was assigned to. These regions are the ones named in `ca.tss` or `ca.tes`.

Transcript maps

Accessible using `ca.t_map`. This table has information from the `cerberus annotate_transcriptome` steps about which TSS, TES, and IC was assigned to each transcript from an input GTF. File format is as follows:

- original_transcript_id
    - Original ID of transcript used in input GTF
- ic
    - Number of IC used in this transcript
- ic_id
    - Name of IC used in this transcript, these names are in the `ca.ic.Name` column
- tss
    - Number of TSS used in this transcript
- tss_id
    - Name of TSS used in this transcript, these names are in the `ca.tss.Name` column
- tes
    - Number of TES used in this transcript
- tes_id
    - Name of TES used in this transcript, these names are in the `ca.tes.Name` column
- gene_id
    - Gene ID that input transcript is associated with
- gene_name
    - Name of gene that input transcript is associated with
- original_transcript_name
    - Original name of transcript used in input GTF
- transcript_triplet

    - Transcript triplet associated with this transcript based on the TSS, IC, and TES it was assigned
    - Formed as concatenation of other columns: [tss,ic,tes]
- transcript_id
    - New ID for transcript computed by cerberus based on the TSS, IC, and TES it was assigned
    - Formed as concatenation of other columns: gene_name[tss,ic,tes]
- transcript_name
    - New name for transcript computed by cerberus based on the TSS, IC, and TES it was assigned
    - Formed as concatenation of other columns: gene_id[tss,ic,tes]
- source
    - Nickname assigned to input GTF from which the input transcript was derived

Gene triplets

Accessible using `ca.triplets`. This table has information about the various gene triplets calculated for the ENCODE LR-RNA-seq project. File format is as follows:

- source
    - Name of the source used to compute the gene triplets
    - Can either be the name of a GTF that was annotated with `cerberus annotate_transcriptome` or another name that a different set of triplets was calculated from
        - Sources from annotated GTFs: lapa (representative of the ENCODE mouse LR-RNA-seq dataset), vM25
        - all: All isoforms annotated across all the sources
        - obs_det: All detected isoforms from the ENCODE mouse LR-RNA-seq data (>=1 TPM in at least one library)
        - sample_det: Sample-level detected isoforms from the ENCODE mouse LR-RNA-seq data (>= 1 TPM) in a given sample
        - sample_major: Sample-level major isoforms from the ENCODE mouse LR-RNA-seq data. Major isoforms must be detected in the sample and together, are responsible for a cumulative 90% of the gene's expression in the given sample to be part of this set.
        - obs_major: The union of all major isoforms across every sample
        - tissue_det: Tissue-level detected isoforms from the ENCODE mouse LR-RNA-seq data (>= 1 TPM) in a given tissue (different from sample because the tissue metadata aggregates across different time points of the same tissue)
        - tissue_major: Tissue-level major isoforms from the ENCODE mouse LR-RNA-seq data. Major isoforms must be detected in the tissue and together, are responsible for a cumulative 90% of the gene's expression in the given tissue to be part of this set.

- tissue_adult_det: Tissue-level detected isoforms from the ENCODE mouse LR-RNA-seq data in only adult (>= 2mo old) tissues (>= 1 TPM) in a given adult tissue (different from sample because the tissue metadata aggregates across different time points of the same tissue)
- tissue_adult_major: Tissue-level major isoforms from the ENCODE mouse LR-RNA-seq data in only adult (>= 2mo old) tissues. Major isoforms must be detected in the tissue and together, are responsible for a cumulative 90% of the gene's expression in the given tissue to be part of this set.

- gid
  - Gene ID of gene triplet
- n_tss
  - Number of TSSs found for this gene triplet
- n_tes
  - Number of TESs found for this gene triplet
- n_ic
  - Number of ICs found for this gene triplet
- n_iso
  - Number of isoforms used to compute this gene triplet
- splicing_ratio
  - The splicing ratio (calculated as (2*n_ic)/(n_tss+n_tes))
- tss_ratio
  - The simplex coordinates for TSS usage calculated as n_tss/(n_tss+splicing_ratio+n_tes)
- spl_ratio
  - The simplex coordinates for TSS usage calculated as splicing_ratio/(n_tss+splicing_ratio+n_tes)
- tes_ratio
  - The simplex coordinates for TSS usage calculated as n_tes/(n_tss+splicing_ratio+n_tes)
- sector
  - The sector on the simplex that this gene falls in. One of 'tss', 'splicing', 'tes', 'mixed' or 'simple'
- gname
  - Gene name of gene triplet
- sample
  - For the sample level triplets, the sample that the triplet was calculated in; otherwise NaN
- tissue
  - For the tissue level triplets, the tissue that the triplet was calculated in; otherwise NaN
- tissue_adult
  - For the adult tissue level triplets, the adult tissue that the triplet was calculated in; otherwise NaN

- gene_tpm
    - For the sample level, tissue level, and adult tissue level triplets, the TPM of the gene in the sample/tissue/adult tissue that the triplet was calculated in; otherwise NaN