ENCODE long-read RNA-seq paper references processing document

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

FANTOM CAGE

Download

FANTOM CAGE peaks for human were obtained from https://fantom.gsc.riken.jp/5/datafiles/latest/extra/CAGE_peaks/hg19.cage_peak_phase1and2combined_coord.bed.gz.

Preprocessing

- Lifted over from hg19 to hg38
- Filtered out regions that were duplicated during liftover
- Filtered out regions that were mapped to a different chromosome than originally during liftover
- Filtered out regions that went from *n* bp long to >= *n*+10 bp long during liftover
- Filtered out regions that are <= 1 bp long

GTEx transcriptome

Download

The GTEx long-read transcriptome was obtained from https://storage.googleapis.com/gtex_analysis_v9/long_read_data/flair_filter_transcripts.gtf.gz.

Preprocessing

- Novel genes were filtered out

PolyA Atlases

Download

The PolyA site atlases were obtained from the following links

- Human: https://www.polyasite.unibas.ch/download/atlas/2.0/GRCh38.96/atlas.clusters.2.0.GRCh38.96.bed.gz
- Mouse: https://www.polyasite.unibas.ch/download/atlas/2.0/GRCm38.96/atlas.clusters.2.0.GRCm38.96.bed.gz

Preprocessing

For both species:

- Regions on noncanonical chromosomes (those beginning with ‘GL’ or ‘KI’) were filtered out
- The ‘chr’ prefix was added to the chromosome of each region