**RNA-seq pipeline for repeat browser—from SQuIRE output**

1. **Heatmap: (all/unique reads)**
2. RPKM/FPKM =

Probably need use TPM instead of RPKM/FPKM because TPM could be compared directly

TEcounts.txt file

1. Count up the total reads in a sample and divide that number by 1,000,000 – this is our “per million” scaling factor. (-2 column of subFcounts.txt)
2. Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK). (-2 column)/(TEend-TEstart)
3. Divide the RPK values by the “per million” scaling factor. This gives you TPM.
4. **Consensus view:**

TEcounts.txt file

1. Read lines from the test\_hg38\_TEcounts.txt
2. Find TEsubfamily in *rmsk file* (consensus length)
3. In rmsk file, -4 ~ -2 column.

(https://genome.ucsc.edu/cgi-bin/hgTables?db=mm10&hgta\_group=varRep&hgta\_track=rmsk&hgta\_table=rmsk&hgta\_doSchema=describe+table+schema)

1. End + abs(#base pair)=total consensus length of TEsubfamily

(or could get it from *hg38TEsubF.size.txt.txt*

1. Start – End = **consensus read length** (from start to end) from rmsk.file
2. Deem the consensus regions as 1, do cumulative sum to get density plot
3. **Genome view (TE subfamily locus)**

TEcounts.txt file

1. Note: there is only all\_reads genome view for SQuIRE output
2. For each subfamily, find the TE\_start position and TE\_chr, draw plot