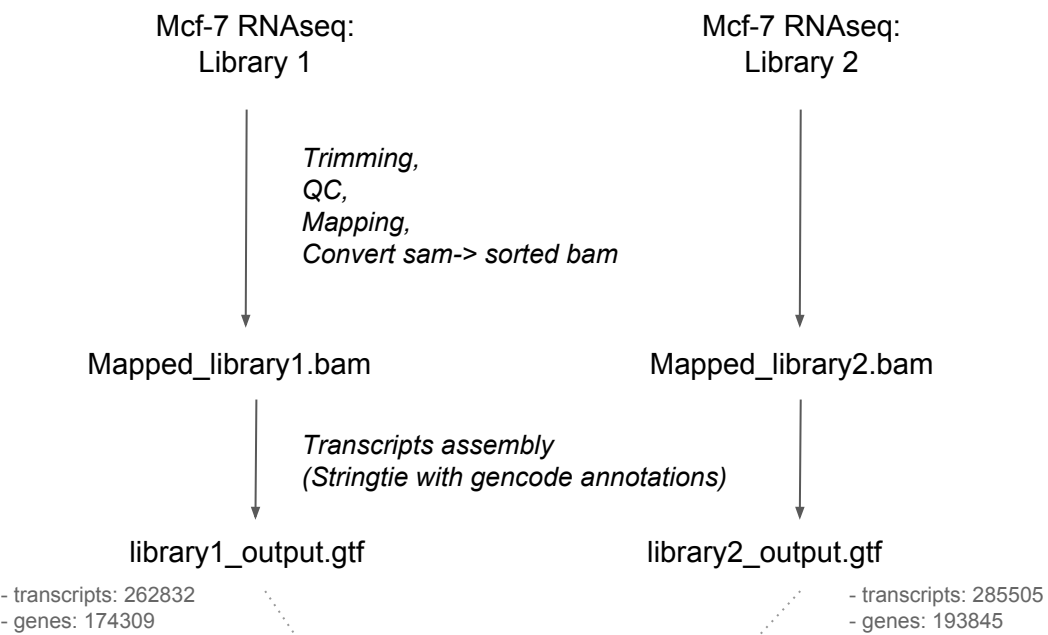
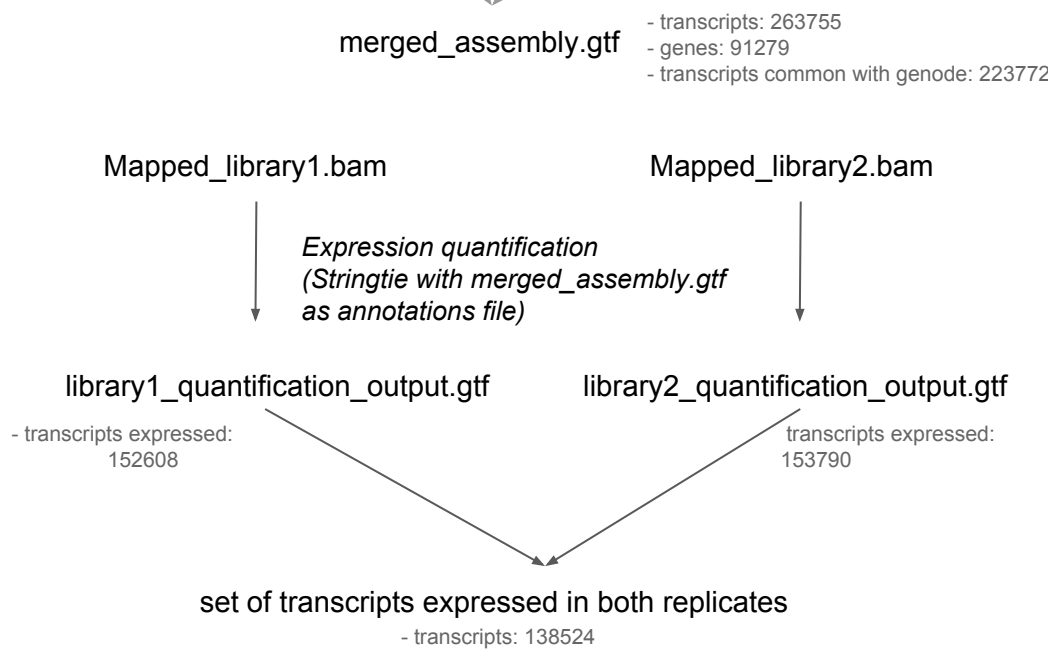


# Part 1: Annotate Mcf7-expressed lincRNAs and PCGs

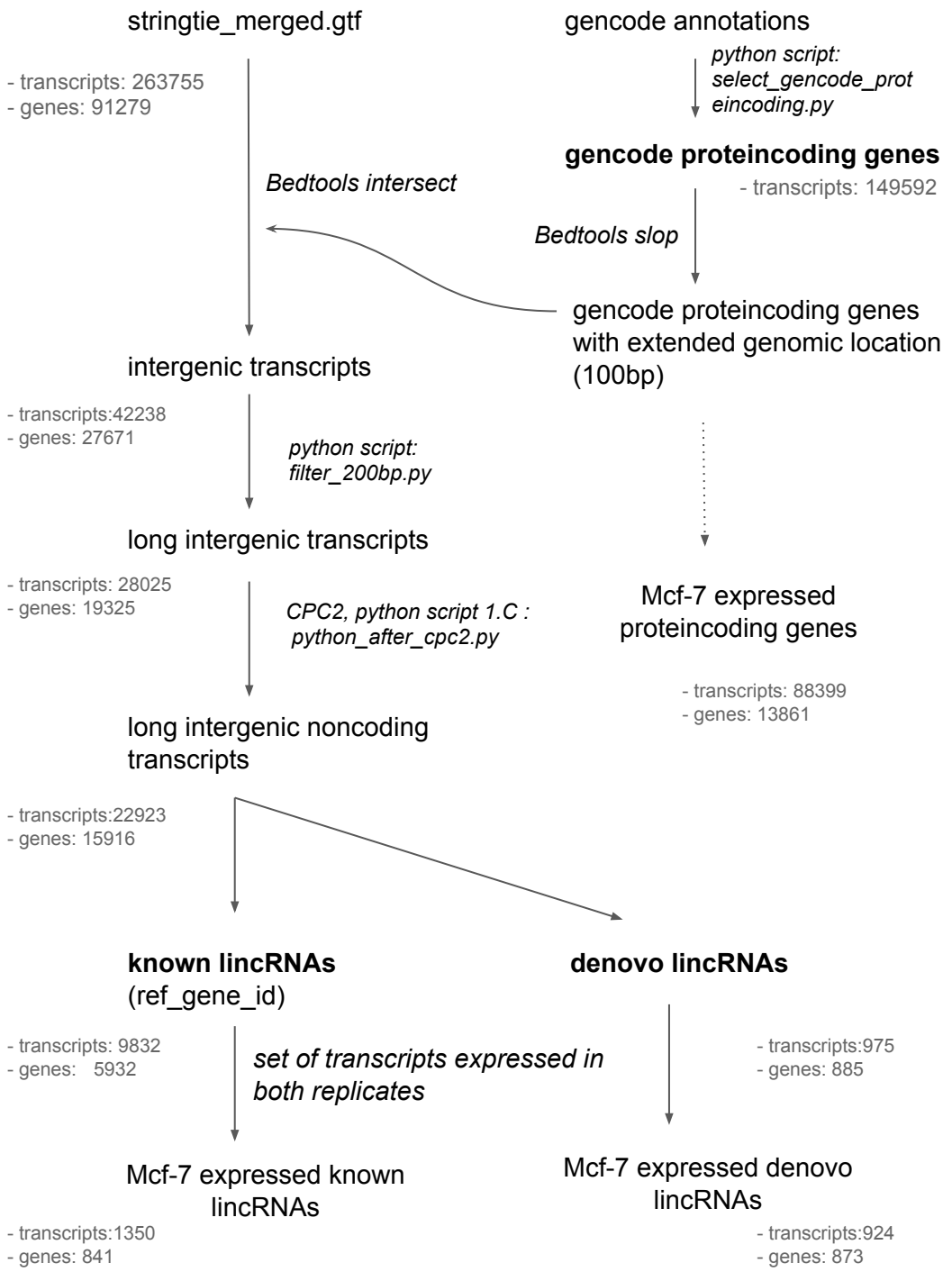
## 1.A. Build a set of expressed transcripts with libraries



## 1.B. Quantify expressed transcripts in both libraries



1.C. Identify lincRNAs and PCGs



Final .gtf

Gencode PCGs: 149592 transcripts

Known lincRNAs: 9832 transcripts

Denovo lincRNAs: 975 transcripts

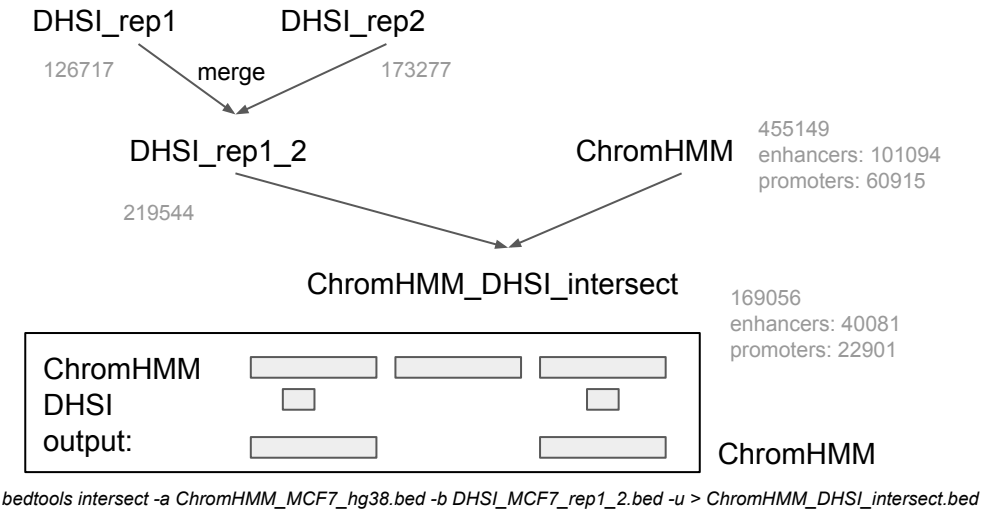
## Scripts

### A. Mapping Mcf-7 RNAseq raw data

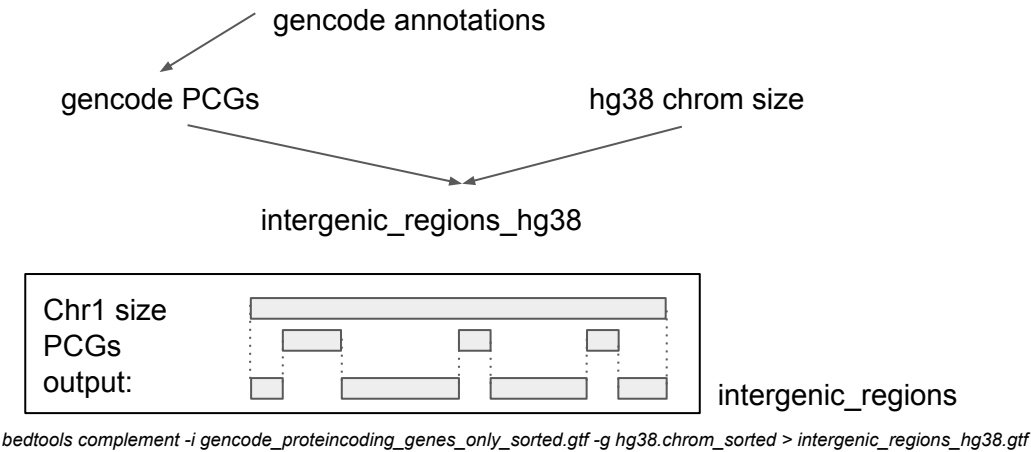
- |      |                     |       |                       |
|------|---------------------|-------|-----------------------|
| 1.   | bash_annotations.sh | ————→ | python_annotations.py |
| 1.A1 | bash_library1.sh    | ————→ | python_library1.py    |
| 1.A2 | bash_library2.sh    | ————→ | python_library2.py    |
| 1.B  | bash_merge.sh       | ————→ | python_merge.py       |
| 1.C  | bash_after_cpc2.sh  | ————→ | python_after_cpc2.py  |

# Part 2: Annotate elincRNAs and plincRNAs

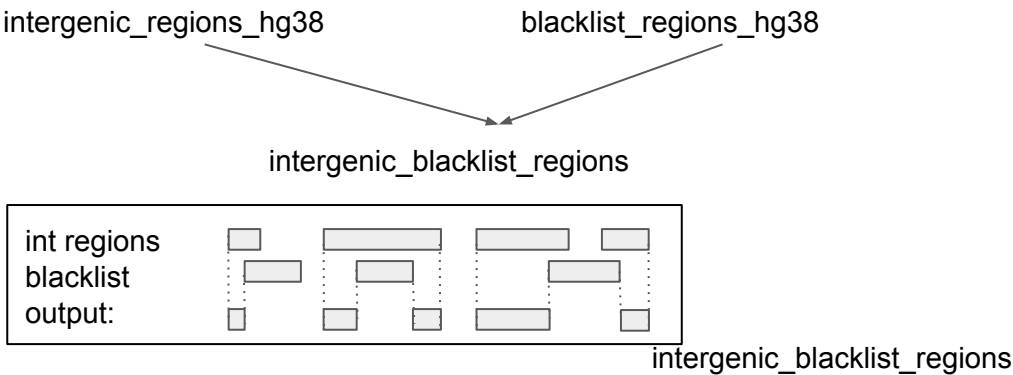
## 2.A. Overlap enhancers with DHSI regions to get accessible enhancers



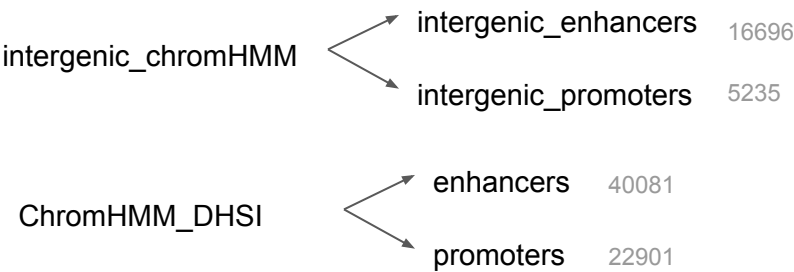
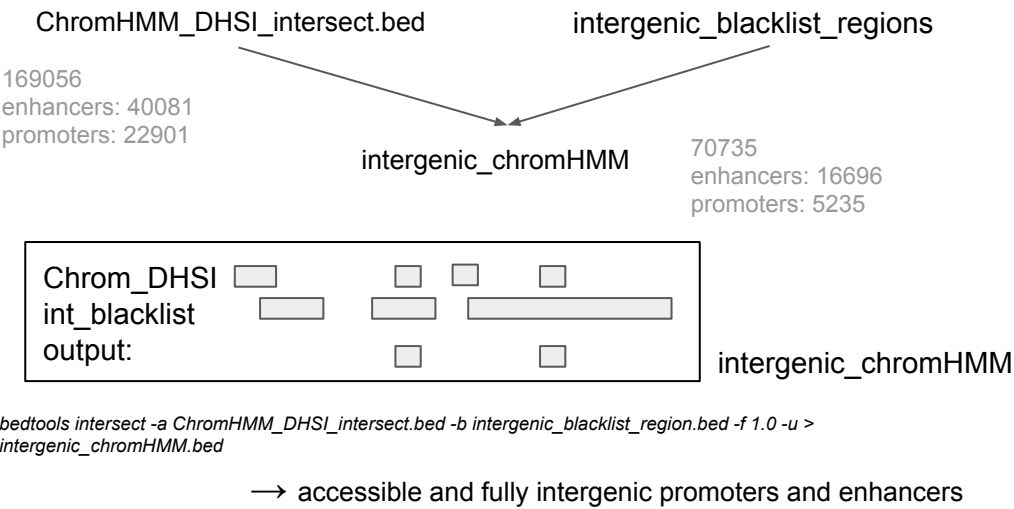
## 2.B. Get intergenic regions of the genome



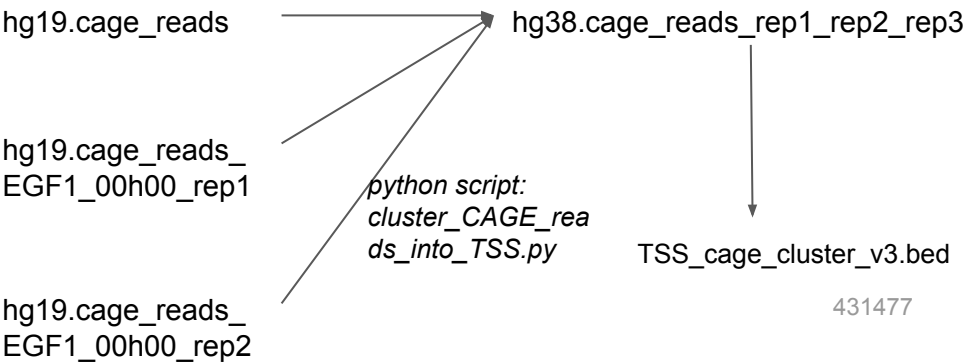
## 2.C. Exclude blacklist regions from hg38 intergenic region



2.D. Overlap accessible chromHMM with intergenic region to get intergenic chromHMM (promoters and enhancers)



2.E. Get TSS from MCF7 CAGE reads (detailed pipeline: 'CAGE clustering')



**2.F. Overlap intergenic enhancers with CAGE TIRs to get enhancer-associated TIRs (eTIRs)**

intergenic eTIRs: 5956            (4453 enhancers)

```
bedtools intersect -a TSS_cage_cluster.bed -b intergenic_enhancers.bed
-u > intergenic_enhancer_associated_TIRs.bed
```

intergenic pTIRs: 4504            (3070 promoters)

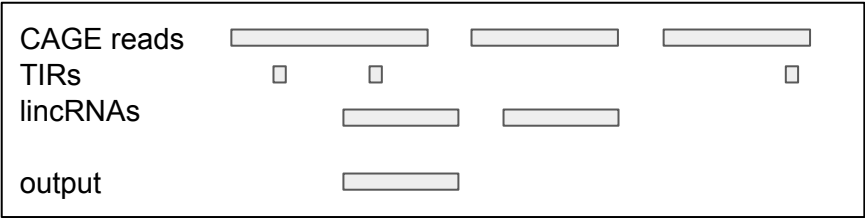
```
bedtools intersect -a TSS_cage_cluster.bed -b intergenic_promoters.bed
-u > intergenic_promoter_associated_TIRs.bed
```

pTIRs: 29391

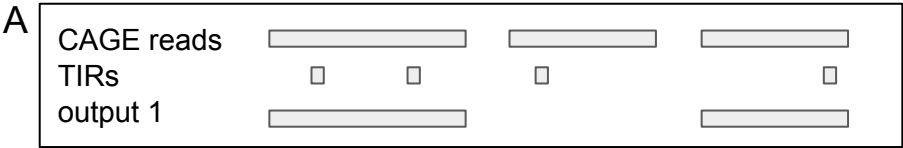
```
bedtools intersect -a TSS_cage_cluster.bed -b promoters.bed -u >
promoter_associated_TIRs.bed
```

**2.G. Determine expressed elincRNAs/plincRNAs/PCGs using CAGE data**

Method: determine whether there is a read that overlap both a TIR and a lincRNA



Commands written for elincRNAs. Repeted for plincRNAs and PCGs



```
bedtools intersect -a cage_reads_hg38_sorted.bed -b intergenic_enhancer_associated_TIRs.bed
-s -u > cage_associated_eTIRs.bed
```



```
bedtools intersect -a breast_exp_lincrnas_first_exon.gtf -b cage_associated_eTIRs.bed -u -s >
elincRNAs.gtf
```

|                                    |       |
|------------------------------------|-------|
| breast_exp_lincrnas_first_exon.gtf | 2274  |
| breast_exp_PCGs_first_exon.gtf     | 88399 |

---

|   |         |
|---|---------|
| TIRs                                    | 431477  |
| intergenic_enhancer_associated_TIRs.bed | 5956    |
| intergenic_promoter_associated_TIRs.bed | 4504    |
| promoter_associated_TIRs.bed            | 29391   |
| cage_associated_eTIRs.bed               | 555103  |
| cage_associated_pTIRs.bed               | 2082626 |
| cage_associated_pcg_TIRs.bed            | 8537013 |
| elincRNAs.gtf                           | 104     |
| plincRNAs.gtf                           | 252     |
| pPCGs.gtf                               | 25521   |

---

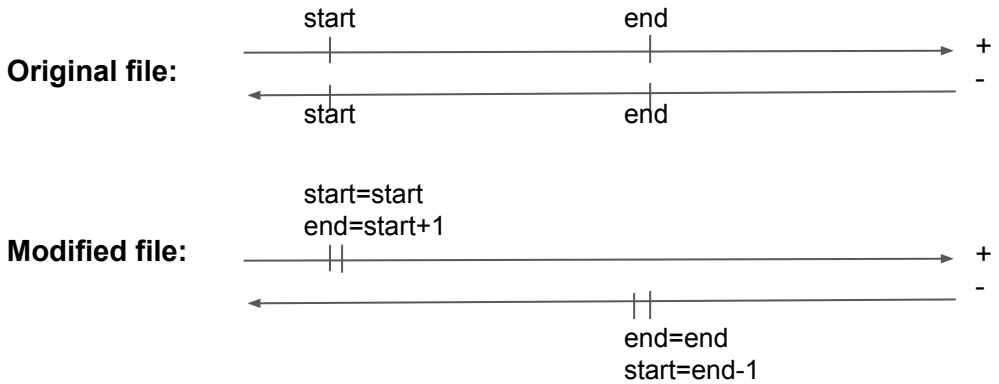
*Analysis V3 transcript exon count: 1 exon / 2 exons / more than 2 exons*

|            |                    |              |
|------------|--------------------|--------------|
| elincRNAs: | 25 / 38 / 41       | ratio: 3.16  |
| plincRNAs: | 82 / 68 / 102      | ratio: 2.07  |
| pPCG:      | 384 / 2452 / 22685 | ratio: 65.46 |

---

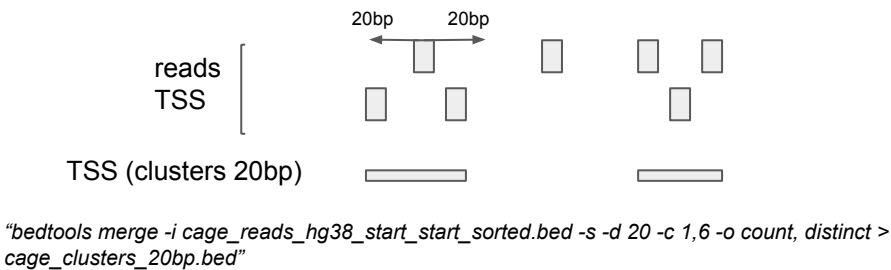
# CAGE clustering

## 1. extract the 5' end position of each read (read TSSs)



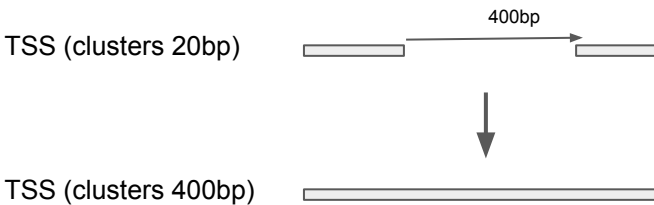
→ file: "cage\_reads\_hg38\_start\_start.bed"

## 2. cluster reads TSS closer than 20bp



→ keep only clusters that contain 2 reads or more

## 3. cluster TSS closer than 400bp



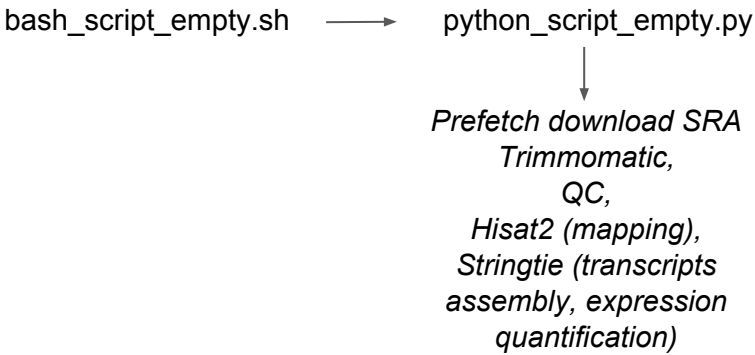
```
"bedtools merge -i cage_cluster_20bp_pairs_sorted.bed -s -d 400 -c 1,5,6 -o count, sum, distinct > cage_clusters_400bp.bed"
```

→ keep only clusters that contain 5 reads or more  
→ TSS\_cage\_cluster.bed



## Part 3: Quantify lincRNAs and PCGs expression level in GTEx breast RNAseq

### Mapping SRA scripts:



### ***make\_scripts.py :***

- for each SRA:
  - replace in python\_script: SRRxxx by SRA number
  - replace in bash\_script: SRRxxx by SRA number
  - run bash\_script → run python\_script

3.A. Store results to save space

Implemented in the *python\_script*, a function modifies the output (gtf format ~260M) to keep only the necessary information (~5.5M).

3.a. Transcripts expression level file for one indiv.

| transcript_id     | Fpkm     | Tpm      |
|-------------------|----------|----------|
| ENST00000473358.1 | 0.000000 | 0.000000 |
| ENST00000417324.1 | 0.000000 | 0.000000 |
| ENST00000461467.1 | 0.000000 | 0.000000 |
| ENST00000641515.2 | 0.000000 | 0.000000 |
| ENST00000335137.4 | 0.000000 | 0.000000 |
| ENST00000477740.5 | 0.000000 | 0.000000 |
| ENST00000453576.2 | 0.000000 | 0.000000 |
| ENST00000496488.1 | 0.000000 | 0.000000 |
| MSTRG.78.1        | 3.591500 | 7.411879 |

3.B. Verify, debug, re-run scripts

1. Implemented in the script *make\_scripts.py*, a function create a file (*done.txt*)that list all SRR to mapp and quantify.

3.b.1 Original *done.txt* file

|            |               |
|------------|---------------|
| SRR1068977 | not_processed |
| SRR1068999 | not_processed |
| SRR1070208 | not_processed |
| SRR1070260 | not_processed |
| SRR1070738 | not_processed |
| SRR1071084 | not_processed |
| SRR1071905 | not_processed |
| SRR1074860 | not_processed |
| SRR1075484 | not_processed |
| SRR1076219 | not_processed |
| SRR1076441 | not_processed |
| SRR1077139 | not_processed |
| SRR1077920 | not_processed |
| SRR1078258 | not_processed |

2. Implemented in the script *python\_script.py*, a function modifies *done.txt* to replace 'not\_processed' by the size of the output.

3.b.2 *done.txt* file after the first run

|            |               |
|------------|---------------|
| SRR1068977 | 5711028       |
| SRR1068999 | 5672836       |
| SRR1070208 | not_processed |
| SRR1070260 | 5710846       |
| SRR1070738 | 5671154       |
| SRR1071084 | 5690314       |
| SRR1071905 | not_processed |
| SRR1074860 | 5682464       |
| SRR1075484 | 5702848       |
| SRR1076219 | 5687583       |
| SRR1076441 | not_processed |
| SRR1077139 | 5709827       |
| SRR1077920 | 5706540       |
| SRR1078258 | not_processed |

3. script: *get\_not\_processed.py*  
Reads the file *done.txt* and re-run script for SRR number labelled as 'not\_processed'

5. script:  
*extract\_err\_file\_done\_v2.py*  
Once all jobs are completed, this script reads extract %trimmed and %mapped reads and add them to *done.txt*

4. script:  
*get\_not\_processed\_vdb-cache.py*  
After bug identified for unprocessed jobs, this script remove bugging lines considered not important and re-run scripts.

3.b.2 *done\_v2.txt* file one all jobs completed

| SRA        | file_size | %mapping | trimming:both surv/dropped |
|------------|-----------|----------|----------------------------|
| SRR1068977 | 5711028   | 98.63%   | ['86.53%/3.62%']           |
| SRR1068999 | 5672836   | 98.60%   | ['86.91%/5.03%']           |
| SRR1070208 | 5686865   | 98.56%   | ['88.71%/4.62%']           |
| SRR1070260 | 5710846   | 98.34%   | ['87.55%/4.63%']           |
| SRR1070738 | 5671154   | 98.71%   | ['87.80%/4.61%']           |
| SRR1071084 | 5690314   | 97.99%   | ['90.26%/3.10%']           |
| SRR1071905 | 5681685   | 98.60%   | ['92.35%/2.52%']           |
| SRR1074860 | 5682464   | 98.67%   | ['89.50%/4.25%']           |
| SRR1075484 | 5702848   | 98.08%   | ['89.65%/3.42%']           |
| SRR1076219 | 5687583   | 98.58%   | ['87.06%/5.09%']           |
| SRR1076441 | 5677204   | 98.66%   | ['89.67%/2.66%']           |
| SRR1077139 | 5709827   | 98.38%   | ['88.72%/4.23%']           |
| SRR1077920 | 5706540   | 98.03%   | ['91.41%/3.13%']           |
| SRR1078258 | 5705450   | 98.42%   | ['88.82%/4.02%']           |

### 3.C. Build phenotype table to call cis-eQTL

script: *make\_table\_phenotype.py*

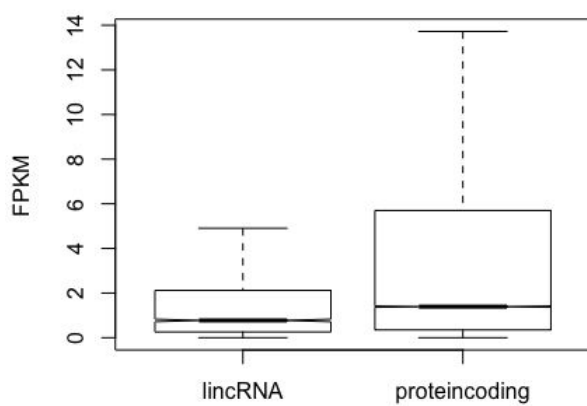
This script reads the 224 output file and write a phenotype table that contains as rows transcripts and as columns their chr, start, end, and expression level in each of the 224 individuals.

| transcript_id     | Fpkm     | Tpm      |
|-------------------|----------|----------|
| ENST00000473358.1 | 0.000000 | 0.000000 |
| ENST00000417324.1 | 0.000000 | 0.000000 |
| ENST00000461467.1 | 0.000000 | 0.000000 |
| ENST00000641515.2 | 0.000000 | 0.000000 |
| ENST00000335137.4 | 0.000000 | 0.000000 |
| ENST00000477740.5 | 0.000000 | 0.000000 |
| ENST00000453576.2 | 0.000000 | 0.000000 |
| ENST00000496488.1 | 0.000000 | 0.000000 |
| MSTRG.78.1        | 3.591500 | 7.411879 |

x 224

| #chr | start  | end    | transcript_id     | GTEx-XBED | GTEx-U8XE | GTEx-XV70 | GTEx-X0T4 | GTEx-RUSQ |
|------|--------|--------|-------------------|-----------|-----------|-----------|-----------|-----------|
| chr1 | 65419  | 71585  | ENST00000641515.2 | 0.000000  | 0.000000  | 0.000000  | 0.000000  | 0.000000  |
| chr1 | 69055  | 70108  | ENST00000335137.4 | 0.075827  | 0.000000  | 0.000000  | 0.000000  | 0.000000  |
| chr1 | 450703 | 451697 | ENST00000426406.3 | 0.040425  | 0.028647  | 0.055988  | 0.021569  | 0.018998  |
| chr1 | 685679 | 686673 | ENST00000332831.4 | 0.038286  | 0.028647  | 0.055988  | 0.021569  | 0.018998  |
| chr1 | 923928 | 939291 | ENST00000420190.6 | 1.530093  | 0.000000  | 0.000000  | 0.801416  | 0.000000  |
| chr1 | 925150 | 935793 | ENST00000437963.5 | 0.063177  | 0.000000  | 0.000000  | 0.000000  | 0.000000  |
| chr1 | 925738 | 944575 | ENST00000342066.7 | 0.000000  | 0.000000  | 0.021964  | 0.000000  | 0.133415  |
| chr1 | 925741 | 944581 | ENST00000617307.4 | 0.030687  | 0.000000  | 0.000000  | 0.000000  | 0.000000  |
| chr1 | 925741 | 944581 | ENST00000618181.4 | 0.000000  | 0.000000  | 0.000000  | 0.000000  | 0.000000  |
| chr1 | 925741 | 944581 | ENST00000622503.4 | 0.020048  | 0.021790  | 1.098797  | 0.030740  | 0.101302  |

### Mcf7 expressed transcripts - expression level



### Breast expressed lincRNAs expression level

