Task 2

Nathan LeRoy

4/16/2022

0. Calculate the pairwise Jaccard similarity for this set of 4 experiments:

```r
filesList = c(
  "data/helas3_ctcf.narrowPeak.gz",
  "data/helas3_jun.narrowPeak.gz",
  "data/hepg2_ctcf.narrowPeak.gz",
  "data/hepg2_jun.narrowPeak.gz"
)

files = lapply(filesList, read.table)
granges = lapply(files, function(x) {
  GRanges(seqnames=x$V1, ranges=IRanges(x$V2, x$V3))
})

pairwiseJaccard(granges)
```

```
##        [,1]   [,2]   [,3]   [,4]
## [1,] 1.0000 0.0204 0.6070 0.0234
## [2,] 0.0204 1.0000 0.0131 0.1650
## [3,] 0.6070 0.0131 1.0000 0.0265
## [4,] 0.0234 0.1650 0.0265 1.0000
```

1. Which two interval sets are the most similar?

The **HeLa S3 - CTCF** (1) sets and the **Hep G2 - CTCF** (3) sets are the most similar according to the Jaccard similarity matrix.

2. Which two interval sets are the most different?

The **HeLA S3 Jun** (2) and **Hep G2 - CTCF** (3) sets are the the most different according to the Jaccard similarity matrix.

3. Based on these results, which factor, CTCF or Jun, would you predict varies more across cell types?

Based on these results, I would expect Jun to vary more across cells. The two different cell lines with the Jun transcription factor have a much smaller Jaccard similarity value (0.165) than the two cell lines with CTCF (0.607).

4. Based on these results, do the genomic locations found by ChIP-seq experiments depend more on the cell-type, or on the transcription factor being assayed?

From these results, it would seem that the genomic locations found in the ChIP-seq experiments depend more on the transcription factor (TF) being assayed. The Jaccard similarity is **higher** for identical TFs than it is for identical cell-lines.