**Assignment 3: Learning polyadenylation signals from sequence**

**Download sequence files.**

Download the relevant sequence files for your organism from UCSC genome browser.

<http://hgdownload.soe.ucsc.edu/downloads.html>

Select the correct organism and assembly, and click on “Full Data Set”:

Human, Feb. 2009 (hg19)

Insect, D. melanogaster, Aug. 2014 (dm6)

Nematode, C. elegans, May 2006 (ce6)

Deuterostome, C. intestinalis, Mar. 2005 (ci2)

Then click on:

chromFa.tar.gz (human)

dm6.fa.gz (fly)

chromFa.tar.gz (worm)

ScaffoldFa.zip (ciona)

Untar/ungzip/unzip these archives and note the path containing your files.

**Create bed files defining:**

1. The region 100 bp upstream and downstream of each annotated transcript end (200 nucleotides total).
2. All exons.
3. All introns.

See the bed file specification: <https://genome.ucsc.edu/FAQ/FAQformat.html#format1>

Use the table from Assignment 1 to generate this file.

**Fetch sequence defined in your bed file from your genome files.**

Download this python script to help you fetch defined sequence from fasta files:

<https://www.dropbox.com/s/qhpfthvj4b7m6lf/fetchFromBed.py>

To use this script, you would type:

python [path to script]/fetchFromBed.py --fetch bed\_f genomeDir out\_f

where

bed\_f is the bed file

genomeDir is the path to your genome directory

out\_f is the output file

**Taking this data in aggregate, analyze base composition within these sequences, and identify motifs that are enriched in the vicinity of transcript ends, relative to exons and introns.**

That is, obtain counts of all kmers for each fetched sequence cohort, where k = 1 through 6, and compare frequencies of each kmer. Compare transcript ends versus exons, transcript ends versus introns, and exons versus introns.

1. Plot the data in a representation that is useful and informative to answer the question – what are characteristics of base composition in each region? Why do you think think base composition varies in these regions? (You may need to go to the literature for clues.)
2. There are sequence motifs that are highly enriched at the ends of transcripts that recruit cellular machinery to cleave nascent RNA transcribed by RNA Polymerase. Identify the most likely candidates for these sequence motifs.
3. Once you have identified likely candidate(s) for this motif, plot the frequency of this/these motif(s) as a function of distance relative to the transcript end (e.g. your x-axis will go from -100 to +100, relative to the transcript end).