**Assignment 1: Get friendly with genomes and genome tables**

**Preparation for Assignment:**

First, download the relevant table for your organism from UCSC genome browser.

<http://genome.ucsc.edu>

Go to Tools -> Table Browser

Select the correct clade, organism, assembly, and track:

Mammal, human, Feb. 2009 (hg19), Refseq genes

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

Nematode, C. elegans, May 2006 (ce6), Refseq genes

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

Enter a file name in the “output file” field.

Then click on “get output” to download the table.

These tables are annotations for all known transcripts in each organism. Each row is a separate transcript. Due to RNA splicing, some genes may have multiple transcripts. The meaning of the columns is as follows:

bin: ignore this

name: transcript name

chrom: chromosome

strand: transcribed strand direction (+ or -)

txStart: start coordinate of the transcript

txEnd: end coordinate of the transcript

cdsStart: start coordinate of coding sequence

cdsEnd: end coordinate of coding sequence

exonCount: number of exons

exonStarts: start coordinates of exons (note the extra “,” at the end)

exonEnds: end coordinates of exons (note the extra “,” at the end)

score: ignore this

name2: gene name

cdsStartStat: ignore

cdsEndStat: ignore

exonFrames: reading frame of each exon (0, 1, or 2)

There are some important features of the table to note:

1. The “start” and “end” coordinates are in ascending order numerically, not in strand direction. So while the genomic coordinate for the 3' end of a plus strand transcript is denoted by txEnd, the coordinate for the 5' end of a minus strand transcript would also be denoted by txEnd. Likewise, the coordinate for the 5' end of a plus strand transcript is the txStart, and the coordinate for the 3' end of a minus strand transcript is also denoted by txStart. The same convention goes for exonStarts and exonEnds.
2. The coordinates all are 0-based starts and 1-based ends. What this means is that if an exon actually starts at base 60 and ends at base 70, what would be listed in the table is 59 and 70 for the exon start and exon end, respectively. UCSC does this to allow for easy arithmetic of lengths, e.g. the length of this exon is 70 - 59 = 11 bases. The same convention goes for txStart, txEnd, cdsStart, cdsEnd, exonStarts, and exonEnds. Starts are always 0-based and ends are always 1-based. See illustration for clarification:

…55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75…

-----XXXXXXXXXXX-----

Here, the exon is denoted by “X’s” and the introns by “-‘s”. This 11 nucleotide exon spans bases 60 to 70, inclusive. However, in the UCSC table, the exonStart would be 59, and the exonEnd would be 70.

1. The UCSC genome browser will show you these tables in graphical format. Click on “Genomes” and select the appropriate group, genome, and assembly. You can type in gene names or coordinates, and look at the transcripts annotated in the table. Make sure that your table (Refseq or Ensembl) is visible (select “full”) in the “Genes and Gene Predictions” menu.

**Assignment:**

1) Plot distributions (histograms) of:

Number of transcripts per gene

Number of introns per transcript

Number of exons per transcript

Intron lengths across the entire genome

Exon lengths across the entire genome

Coding sequence length across all transcripts (note that these should all be multiples of 3)

3' UTR lengths

5' UTR lengths

Also output the mean, median, minimum, and maximum, and standard deviation for each metric. For some of these metrics, it will be more useful to plot on a log scale.

2) Compute the fraction of exons that are a multiple of 3

**Suggested modules, functions, and hints:**

You will need to be able to open and read files – see “Reading and Writing files” here: <https://docs.python.org/2/tutorial/inputoutput.html> (you will need to use the “open” function).

You will need to be able to split strings by tab stops and commas – look up basic string operations here: <https://docs.python.org/2/library/string.html> (you will need “strip” and “split” functions). A tab character is “\t”.

You will need to be able to plot data and save figures – see the pylab/matplotlib library: <http://matplotlib.org/>. These functions will be handy: hist, xlabel, ylabel, savefig