**Assignment #3 GFF feature files and visualization**

Your biologist collaborators are going to do a site-directed/insertional mutagenesis in Arabidopsis, using the list of 167 genes from your last assignment as the desired targets.  Inserts will be targeted to the repeat CTTCTT, and they want inserts to go into EXONS.

**Tasks:  for 10% (easy)**

**1:**  Using BioRuby, examine the sequences of the ~167 Arabidopsis genes from the last assignment by retrieving them from whatever database you wish

**2:** Loop over every **exon** feature, and scan it for the CTTCTT sequence

**3:**  Take the coordinates of every CTTCTT sequence and create a new Sequence Feature (you can name the feature type, and source type, whatever you wish; the start and end coordinates are the first ‘C’ and the last ‘T’ of the match.).  Add that new Feature to the EnsEMBL Sequence object.  (YOU NEED TO KNOW:  When you do regular expression matching in Ruby, use RegEx/MatchData objects; there are methods that will tell you the starting and ending coordinates of the match in the string)

**4a:**  Once you have found them all, and added them all, loop over each one of your CTTCTT features (using the #features method of the EnsEMBL Sequence object) and create a GFF3-formatted file of these features.

**4b:**  Output a report showing which (if any) genes on your list do NOT have exons with the CTTCTT repeat

Biologists always want to “see the data”... and usually that means they want to see it in the context of all other genomic features.  ENSEMBL allows this.

Unfortunately, “Coordinate systems” are a grano en el culo!  The squence files you retrieve (e.g. using dbfetch) indicates the feature start/stop relative to the small segment of DNA in that file.  Online databases like ENSEMBL use whole-chromosome start/end coordinates (and sometimes contig coordinates).

**Tasks:  for 10% (harder)**

**5:**   Re-execute your GFF file creation so that the CTTCTT regions are now in the full chromosome coordinates used by EnsEMBL.  Save this as a separate file -***I will grade you on both 4a and 5 so be sure to submit both!*** You have to calculate the chromosome coordinates yourself, somehow… hint - look in the information at the beginning of the sequence file.  NOTE THAT THIS REQUIRES YOU TO CHANGE COLUMN 1 OF THE GFF FILE ALSO!  NOT JUST THE START/END COORDINATES (see below)

**6:**   Prove that your GFF file is correct by uploading it to ENSEMBL and adding it as a new “track” to the genome browser of Arabidopsis (see http://plants.ensembl.org/info/website/upload/index.html  - there are also links here that will help you understand GFF format → tells you what information should appear in Column 1!)

Along with your code, for this assignment please submit a screen-shot of your GFF track beside the **AT2G46340** gene on the ENSEMBL website to show me that you were successful.  (note that the menu in the top-left of the ENSEMBL track-browser has an “export as image” option - or you can just take a screenshot - whichever you prefer)

**PLEASE LISTEN TO ME!   Create a small sequence (~20nt) that has CTTCTT on both the + and - strands, then test your code on that - make sure that the results are correct before you test it on large sequences.  Counting the positions on the minus strand is hard!!!!!  Be careful!!!**