Searching for, Characterizing, & Analyzing Non-Coding DNA & RNA
Identifying MicroRNAs & Comparing Repetitive DNA: Hands-On Exercise

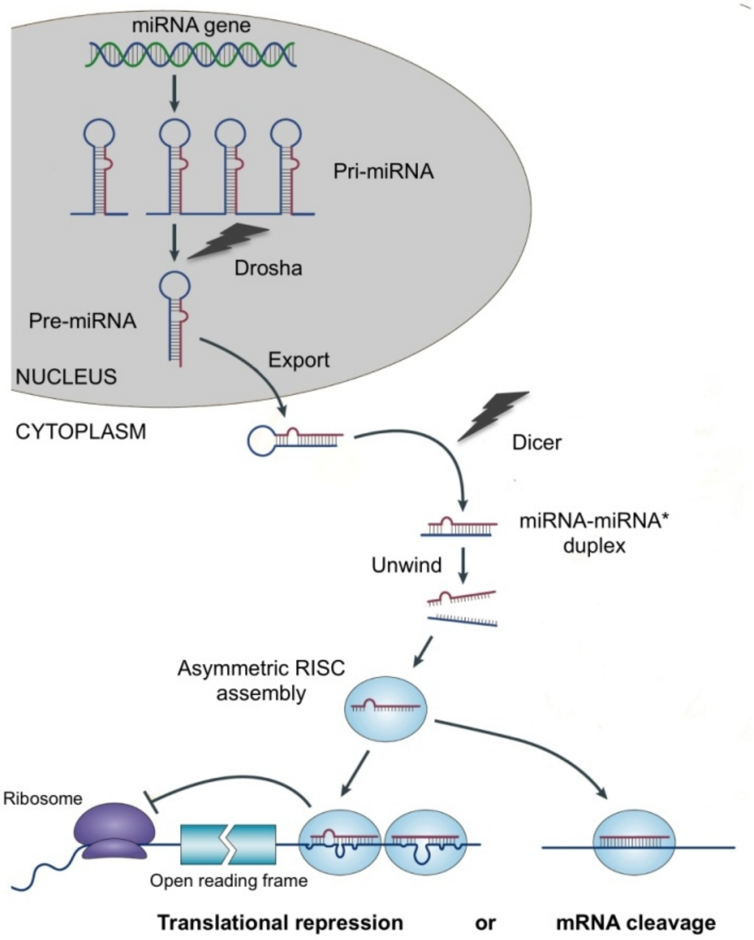
Genomic analysis has revealed that the vast majority of sequence data from multicellular eukaryotes (like humans!) is non-coding, and in many cases repetitive. Non-coding regions can still be transcribed and, in many cases, the transcribed RNA molecules can exhibit secondary, tertiary, and quaternary structures (much like folded proteins!) and perform a function in the genome. Non-coding and/or repetitive regions can be either functional (e.g., microRNAs) or non-functional (e.g., transposable elements [TEs]), with respect to the host genome in which they are found. However, even when they are not currently functional, non-coding regions can often have major consequences over the long-term for a given lineage. For example, new TE insertions in/near genes or non-homologous recombination among recently inserted TEs provide a significant source of mutation (and thus genetic variation) in many species in which TE activity has been closely examined.

Overview: Today, we are going to use several different tools. One tool, MirEval, searches sequence data to find the genetic precursors of microRNAs (miRNA; non-coding but functional RNA molecules, which play a role in gene regulation, primarily by silencing genes through the targeted hybridization of miRNA to mRNA molecules). Second, we will familiarize ourselves with RepBase and tools related to identifying and quantifying various types of repetitive DNA. RepBase is similar to GenBank, but it contains only repeat sequences.

Part I: Finding microRNAs in the human herpesvirus genome that may explain its latent form

Oral and genital herpes results from infection with human herpesviruses 1 and 2 (HHV-1 and HHV-2). In addition to active replication in epithelial cells (producing the characteristic herpes lesions), these viruses can enter neurons where they remain latent (dormant) for long periods of time. The mechanisms involved in establishing and maintaining latency remain poorly understood. However, a non-coding RNA called LAT (latency-associated transcript), which is actually a transcribed intron, has been implicated in down-regulating viral genes during latency. The prominence of this non-coding RNA, together with the demonstrated use of microRNAs by other viruses to maintain latency (e.g., Epstein-Barr virus, SV-40, hepatitis C, HIV), suggests that it may be useful to look for miRNAs in the LAT sequence.

In order to predict potential microRNAs in the LAT region of the human herpesvirus, we must review what is known about the biogenesis of microRNAs – how are they synthesized and processed in metazoans (i.e. animals)? Ultimately, two processing steps occur that trim the primary miRNA (pri-miRNA) transcript into mature miRNA. The unprocessed pri-miRNA transcript consists of a ~33 bp long hairpin structure with tails at both the 5' and 3' ends. In animals, an enzyme called Drosha cleaves the hairpin tails resulting in the pre-miRNA molecule (see figure from Goyvaerts et al. 2013).



The pre-miRNA molecule then undergoes a second processing step, in which an enzyme called Dicer cleaves the terminal loop off of the hairpin shaped molecule, resulting in a mature miRNA duplex (~22 nt). One of the duplex strands will be degraded. The surviving mature miRNA will bind with an Argonaute protein and form a silencing complex that will target nucleic acid molecules (primarily mRNAs) by matching sequence. Base-pair interactions lead to cleavage and degradation of the mRNA, thus silencing the gene from which they came.

microRNA prediction with MirEval

To predict potential miRNAs, the first tool we will use is **MirEval** (http://mimirna.centenary.org.au/mireval/). This program combines up to 4 kinds of miRNA analysis, depending on what is available for the sequence you submit:

(i) RNA secondary structure analysis finds potential pre-miRNA "hairpin" substrates (hairpins already trimmed by Drosha) that contain Dicer cleavage sites
(ii) conservation analysis, based on BLAST alignment and comparison with other vertebrate sequence data
(iii) cluster analysis to look for other miRNAs in flanking sequence (many miRNAs occur in clusters)
(iv) comparison with known miRNAs in **miRBase**.

Go to the **NCBI Nucleotide** database to find the *complete genome* sequence for **Human herpesvirus 1** (the **RefSeq** sequence). There are many ways to find it (you should be an expert by now! it's also linked on the webpage). ***<u>HINT</u>****: Use thorough search terms and search only* ***RefSeq****.*

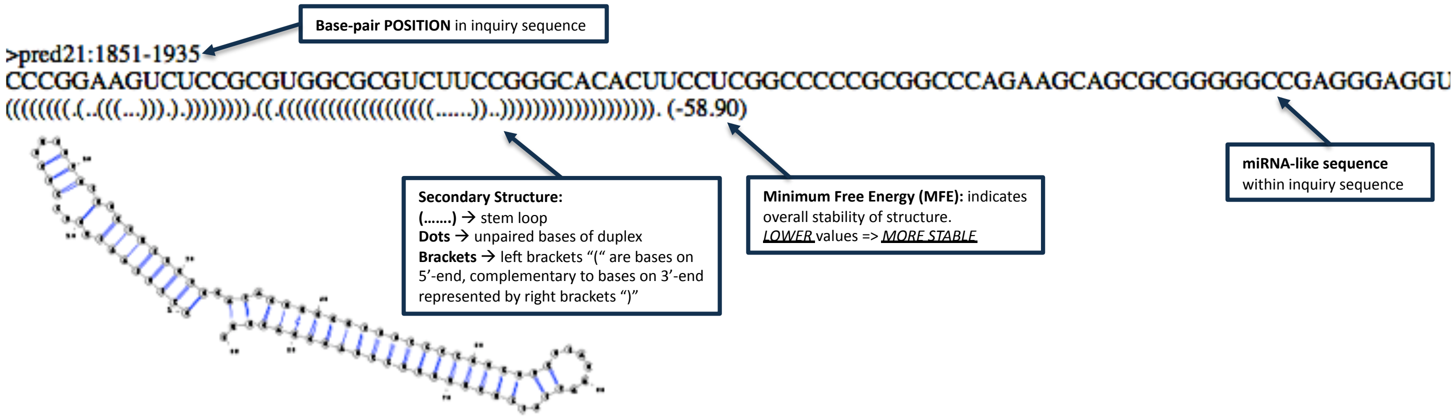
Once you've navigated to the HHV-1 whole genome in NCBI (accession: **NC_001806.1**), search the text of the GenBank record (Command+F or `Control+F`) for "**LAT**" (for ease, add a space after LAT or check **match case**). You should find a feature labeled **stable LAT intron**. There are multiple LAT introns, **select the <u>first one</u>** ***(position: 4953-6907).*** *Note the orientation of this intron within the genome:*

____

Click on the **intron** link on the left side of the screen to see the sequence for just this LAT stable intron. On the bottom right where it says **Display**, click on the link that allows you to view the nucleotide sequence for this intron in **FASTA** format. Highlight and copy this FASTA sequence (Command+C or `Control+C`).

Navigate to the MirEval (**http://mimirna.centenary.org.au/mireval/**) website and paste (Command+V or `Control+V`) your copied sequence into the window. Because there is no option for our "organism" (it is a virus), select **Others** (#32) under **Genome** (this automatically performs the analysis based on **[i]** and **[iv]** from above.) Where it asks **Predict the <u>original/complementary</u> strand of your sequence**, choose **<u>original</u>** and **Submit.**

Scroll through the output, and refer to the example figure below with explanations of the output details. Notice that for each predicted miRNA stem-loop precursor, there is information on the position, the actual sequence, the structure (indicated by brackets and dots), the minimum free energy (**MFE**) score (the lower the score, the better; notice the numbers are usually negative!), and a graphic with the predicted structure. If there are hits to sequences in miRBase, they will be listed at the end of the output. ***<u>NOTE</u>***: Please make sure to keep *any and all* **MirEval** result windows open, as you will need these sequences again later for further analyses.



```
>pred21:1851-1935
CCCGGAAGUCUCCGCGUGGCGCGUCUUCCGGGCACACUUCCUCGGCCCCCGCGGCCCAGAAGCAGCGCGGGGGCCGAGGGAGGU
((((((((.(..(((...))).).))))))))).((.(((((((((((((((((((......))..)))))))))))))))))))). (-58.90)
```

Base-pair POSITION in inquiry sequence

Secondary Structure:
(…….) → stem loop
Dots → unpaired bases of duplex
Brackets → left brackets "(" are bases on 5'-end, complementary to bases on 3'-end represented by right brackets ")"

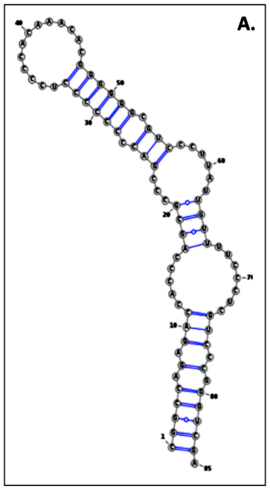
Minimum Free Energy (MFE): indicates overall stability of structure.
<u>LOWER</u> values => <u>MORE STABLE</u>

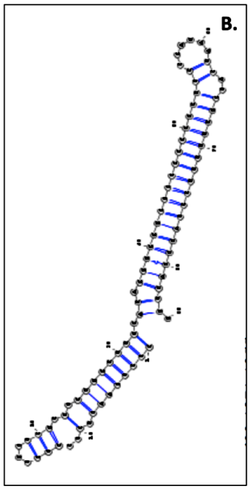
miRNA-like sequence within inquiry sequence

Input data for MirEval must be <10 kb. You can search multiple sequences at once (if they are < 10 kb total and in FASTA format). If you are searching sequence from a species for which there is a lot of genome information, the MirEval output is different because it uses all 4 search strategies listed above. If you would like to see the full capabilities of this program, run through the MirEval tutorial on your own.

Are the predicted miRNA structures listed in an order based on position or score? ___________
How many potential microRNA precursors were identified using your original sequence? _______________
What about when you use the complement? (NOTE: Go back to MirEval and select the <u>complementary</u> strand of your sequence) _________________________________
Among all the predicted miRNAs on both strands, which one has the BEST (lowest) score? (Write the score, orientation, and position) ____
Were any similarities to known microRNAs in miRBase found? ____________

Which of these two structures below (A. or B.) represents a more stable candidate miRNA? _______ *(Why?)*





Evidence for microRNA: conservation and targets
Predicting the existence of a miRNA based on structure is interesting, but it's a long way from showing functionality. Ultimately, wet lab experiments would be necessary, but we can gain additional insight using bioinformatics. If these are real miRNAs, we might expect they:
(1) be ***conserved*** in the LAT of HHV-2 virus, which is very closely related and biologically similar to HHV-1
(2) have ***target*** sequences, which means that the predicted miRNA would have to have at least 1 perfect or near-perfect match to a segment of the human or HHV genome. *Let's try looking for both.*

(1) *Conservation*: Obtain the sequence for the LAT stable intron for HHV-2 from the NCBI Nucleotide database in FASTA format (**NC_001798.1**). Again, there are multiple LAT introns, **select the first one *(position: 4812-7025).*** Use MirEval to predict potential miRNAs in this sequence. Again, identify the lowest MFE scoring (best) predicted miRNA (*remember to use the original AND the complement*) in HHV-2. *Is the best predicted HHV-2 miRNA in roughly the same region as the best predicted HHV-1 miRNA?*_______ *If so, does this increase your confidence that this is a real miRNA?* _______________

(2) *Target*: Go back to your results from HHV-1. Take your *best* predicted miRNA from HHV-1 and copy its sequence from the MirEval results. Use Web BLAST to search using this sequence as your query.
Under **Database**, specify the **RefSeq** database **(refseq_genomic)** and **under Program Selection** select **Highly Similar Sequences** (remember miRNAs only silence targets with nearly *exact* matches). Under **Organism** type "**human**" and select **human (taxid:9606)**. Press **+ button**, to add another box. Type "**herpes**" and select **Herpesviridae (taxid:10292).** Perform the BLAST (limiting your search to these two organisms to save time) and see if there are any other exact or nearly exact hits in the HHV-1 genome or the human genome (other than the hit you expect to see in the LAT stable intron, which you may remember was located between position 4953 and 6907 in the HHV-1 genome).
Are there any other matches to the HHV-1 genome? ______________
Do your BLAST results support the hypothesis that you have identified a real miRNA? _______________________
On your own…

Predicting cleavage sites

Earlier in this module (Part I), you searched for microRNAs encoded in the LAT region of HHV-1 and HHV-2 using structural prediction, sequence conservation across HHV-1 and HHV-2, and by identifying potential target sequences. These predictions of functional miRNA could be validated by wet lab experiments *(what kind of experiments?),* however experiments involving humans are time-consuming, costly, and involve many ethical considerations. Therefore, proceeding with your research you could further boost your confidence that the miRNA you have found is real, and that you should no longer prioritize the other candidate miRNAs identified by mirEval.

Using the set of pre-miRNA sequences generated by MirEval, let's use another program called **PHDcleav** that searches for Dicer cleavage sites on the 5' arm of pre-miRNAs. This will help us to assess the validity of the **MirEval** proposed hairpins, which should contain Dicer cleavage sites on either side of the hairpin loop.
Start with the five lowest scoring (MFE scores) predicted miRNAs from your **MirEval** output (derived from the HHV-1 LAT intron complement).
Navigate to the **PHDcleav** website (http://www.imtech.res.in/raghava/phdcleav/) using the link available on the webpage. Click on the **submission** tab at the top of the page. Paste all five sequences, in order from lowest MFE score to highest, separating each sequence with a semicolon "**;**", into the submission page input window. Leave the **threshold on default of 0** and then select **Run Prediction**. A results window will load a table that displays all the possible Dicer cleavage sites, in all 5 of your sequences; your sequences will be numbered automatically according to the order in which they were pasted in the input window. For example, in the results table "hairpin_5" is the fifth sequence submitted in the input window.
Did the MirEval sequence with the lowest MFE score (most stable) have any Dicer prediction sites? ____________
Which of the top five most stable pre-miRNAs from MirEval have potential Dicer cleavage sites? _____________
Based on this exercise, would you use multiple miRNA prediction programs in the future? *(Why?)*

Part II: Searching for transposable elements using CENSOR

As we have discussed, there are many categories of repetitive elements in the genome. One major category, transposable elements (TEs; also called mobile elements, or mobile genetic elements), account for very large proportions of most multicellular eukaryotic genomes. Many bioinformatic tools have been developed to identify these elements, several of the most useful of which were developed by the Genetic Information Research Institute (GIRI; http://www.girinst.org/). There are 3 main resources available at this website:

1) A database of all known, deposited repetitive sequences.
2) A program called **RepeatMasker--** it can be used to quantify how much of any given genome sequence is composed of repeats if you have a library of repeat sequences with which to "mask" the genome of interest.
3) A web tool called **CENSOR--** basically a program like BLAST that allows you to search your query sequence against a database of known repeats (although you can limit your search by TE type or taxa).

An area of active research is to determine how and why TEs accumulate differentially among species and within local regions of the genome. For example, if TE insertions can disrupt gene function, one might hypothesize that TEs are less likely to accumulate in gene-rich regions. Similarly, if recombination provides a mechanism whereby costly TE insertions can be purged from the genome, we might expect that regions of the genome that have low recombination rates also accumulate a higher number of TEs. We can test this hypothesis by comparing TE content on sex chromosomes in humans because the frequency of recombination is different for the X and the Y. (*Which do think has a higher rate of recombination in humans? Why?*) We can use a randomly selected autosome as a control to see if the TE abundance we see on the X or the Y appears to be more "typical" of non-sex chromosomes.

We'll use **CENSOR** to BLAST our samples from chromosomes and tell us what portion of the sequence are actually TEs. Human chromosomes are pretty large, so instead of doing the whole X and Y, we can take a subsample by downloading one sequenced BAC (bacterial artificial chromosome) clone from each of the chromosomes we are interested in (the X, the Y, and an autosome). This approach will give us a first look, and can tell us whether the project is worth pursuing. *If we had more time, however, how might we improve the experimental design to test this hypothesis?*

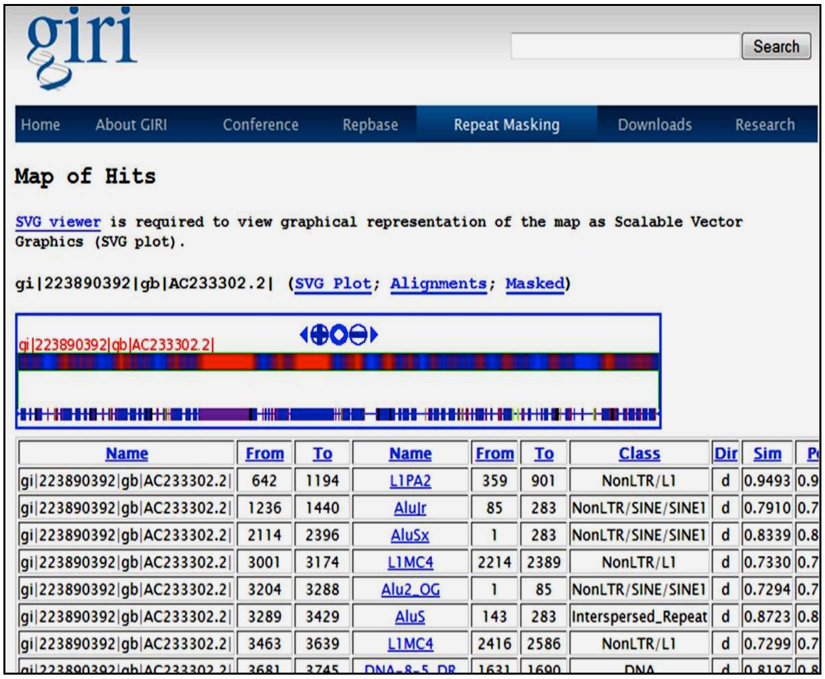
To start, search **Genbank** for the human X chromosome BAC clone **AC233302.2**. Click on the **Display Settings** on the top left of the GenBank record to view the sequence in **FASTA** format. Navigate to the **CENSOR** webpage (http://www.girinst.org/censor/index.php), and copy and paste the GenBank FASTA sequence directly into the box that says **Paste query sequence here**. Select **Sequence Source: <u>.... Homo sapiens (Human)</u>.** There is no need to check any of the boxes on this page at this time, but make sure you read them and know what they are. Hit the **Submit File** or **Submit Sequence** button depending on how you have entered your sequence data.

It will take a few minutes to run, as it is a long sequence. While you are waiting, consider the following: *Which do you think will have more TEs, the X or the Y? What about the autosome? Do you think it will have more, fewer, or the same number of TEs per basepair as the sex chromosomes? Discuss your ideas with your neighbor and come up with a hypothesis and write it here:*

Why is the "per bp" caveat important?

Search **GenBank** for the 2 other sequences you need for your comparison—a BAC from an autosome and the Y chromosome. You can use the following accession numbers to search for suitable sequences (autosome: **AC005690.8** and Y: **AC244170.3**). For each of your searches, the CENSOR output will look like the figure below:

NOTE: Keep track of which result window corresponds to which chromosome!



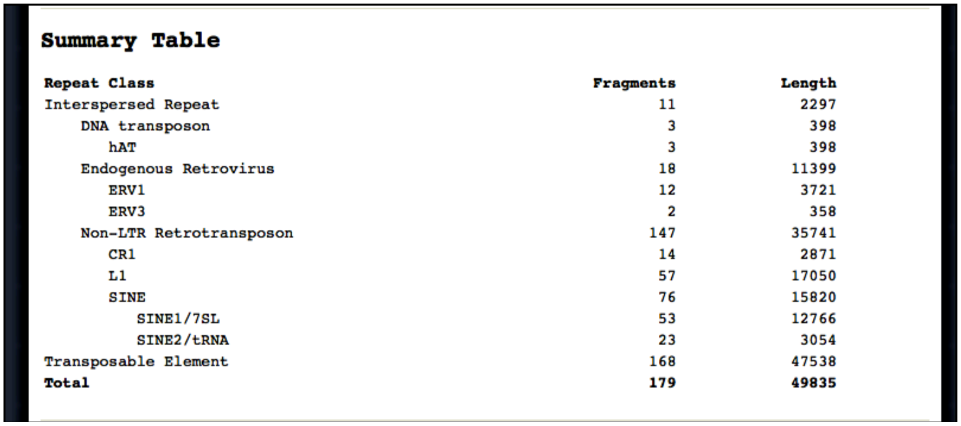
Map of Hits

SVG viewer is required to view graphical representation of the map as Scalable Vector Graphics (SVG plot).

gi|223890392|gb|AC233302.2| (SVG Plot; Alignments; Masked)

Name	From	To	Name	From	To	Class	Dir	Sim	Pe
gi|223890392|gb|AC233302.2|	642	1194	L1PA2	359	901	NonLTR/L1	d	0.9493	0.9
gi|223890392|gb|AC233302.2|	1236	1440	AluJr	85	283	NonLTR/SINE/SINE1	d	0.7910	0.7
gi|223890392|gb|AC233302.2|	2114	2396	AluSx	1	283	NonLTR/SINE/SINE1	d	0.8339	0.8
gi|223890392|gb|AC233302.2|	3001	3174	L1MC4	2214	2389	NonLTR/L1	d	0.7330	0.7
gi|223890392|gb|AC233302.2|	3204	3288	Alu2_OG	1	85	NonLTR/SINE/SINE1	d	0.7294	0.7
gi|223890392|gb|AC233302.2|	3289	3429	AluS	143	283	Interspersed_Repeat	d	0.8723	0.8
gi|223890392|gb|AC233302.2|	3463	3639	L1MC4	2416	2586	NonLTR/L1	d	0.7299	0.7
gi|223890392|gb|AC233302.2|	3681	3745	DNA-8-5_DR	1631	1690	DNA	d	0.8197	0.8

At the top, you will notice a graphical representation of the regions of the query sequence (your BAC clone) that were similar to TEs in the database. What is the first thing that you notice? Probably that there are a lot of hits! As you know by now, this is not a surprise because approximately 66% of the human genome is composed of TEs. Thus, we might expect our randomly selected chromosome segment captured by this BAC clone to also have about 66% TEs (that is, 66% could be our null hypothesis). It is hard to quantify what the percentage is by looking at the graphic, so let's scroll through the output. Most of the output page shows the data for each specific region of similarity (including the region on the query sequence, the starting nucleotide position, the ending nucleotide position, the name of the TE that it matched with, the starting and ending position in that sequence, the class of TE, the direction, the percent similarity, the percent positives [for protein alignments], and the score). Below this table is more detailed information about each hit. However, if you scroll to the very bottom of the page, you will see a summary of all the data that will help us answer our question (an example summary table is shown below).



```
Summary Table

Repeat Class                                    Fragments        Length
Interspersed Repeat                                    11          2297
    DNA transposon                                      3           398
       hAT                                              3           398
    Endogenous Retrovirus                              18         11399
       ERV1                                            12          3721
       ERV3                                             2           358
    Non-LTR Retrotransposon                           147         35741
       CR1                                             14          2871
       L1                                              57         17050
       SINE                                            76         15820
          SINE1/7SL                                    53         12766
          SINE2/tRNA                                   23          3054
Transposable Element                                  168         47538
Total                                                 179         49835
```

The summary table gives you the total number of fragments in your query sequence that hit to all the known repeats (179), as well as giving you the number of basepairs in your query that matched to any repeats (49835). The total number of fragment hits is also separated into specific repeat classes. CENSOR does not classify interspersed repeats as TEs; however, we will include them in our calculations of TEs, as most interspersed repeats are the result of transposition events.

Run CENSOR for *each* of your 3 BAC clones, record the data in the table below, calculate the percentage needed in the last column to perform your comparison, and answer the following questions:

DATA TABLE

Chromosome Type	Accession #	Length of BAC clone (bp)*	Total number of element fragments (hits)	Length of repetitive DNA in basepairs	*Percent of the BAC composed of repetitive sequence*
Autosome	**AC005690.8**				
X	**AC233302.2**				
Y	**AC244170.3**				

*Obtain the BAC clone length from the GenBank record

Which chromosome has the most TE fragments? __________________

Is the type of chromosome with the most TE fragments <u>also</u> the type with the highest proportion of basepairs contributed by TEs? _______

*Which type of chromosome has the highest proportion of sequence contributed by TEs?*________

Does your data support your original hypothesis about where TEs might accumulate based on the frequency of recombination? _________

Is the proportional contribution of TEs to the autosome more similar to that observed on the X or the Y? Why do you think this is the case?

This exercise was developed for a 2014 workshop. The authors give permission for the non-commercial educational use of this material.

Further Reading: *1) Alexander, R.P. et al., 2010. Annotating non-coding regions of the genome. Nature Reviews Genetics 11, 559-571. 2) Bachtrog, D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. Nature Reviews Genetics 14, 113-124. 3) Ghildiyal, M. and Zamora, P.D., 2009. Small silencing RNAs: an expanding universe. Nature Reviews Genetics 10, 94-108. 4) Treangen, T.J. & Salzberg, S. L. 2011. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 13, 36-46.*