- This document provides an overview of the analysis workflow -

(1) Run the pipeline script:

```
ipython fast_iCLIP.py <target>
```

This will produce all the source data files needed for figures. By default, the pipeline will generate figures. However, several python notebooks are provided to make this easier.

(2) Running python notebooks:

There are three notebooks provided. Details for installing the python notebook environment can be easily found (ipython.org/notebook.html). It is convenient to access the notebooks remotely (e.g., from your laptop) by directing the notebook to a specific port on your cluster and tunneling to that port. This is done using the below (using lmartin@changrila as the cluster we tunnel to).

```
ssh -N -f -L localhost:8889:localhost:7000 lmartin@changrila
```

This detailed link (http://wisdomthroughknowledge.blogspot.com/2012/07/accessing-ipython-notebook-remotely.html) provides more information. Tunneling allows your to access the cluster hosted notebooks (and all the cluster associated datafiles) from your local (e.g., your laptop) browser using the port specified (e.g., http://localhost:8889).

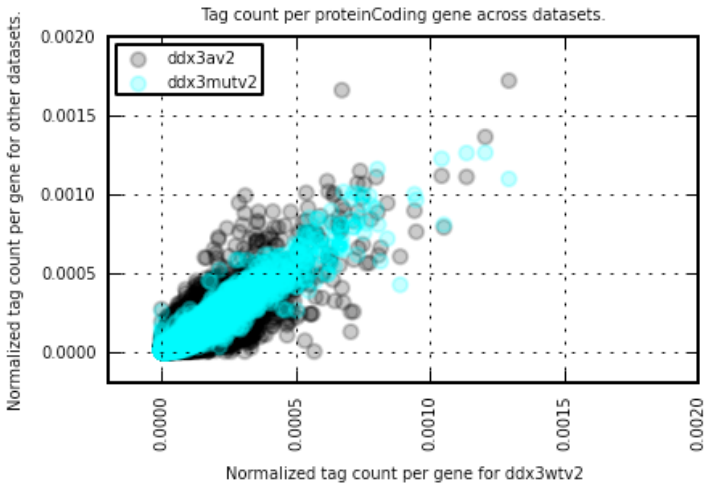
(3) fast_iCLIP.ipynb

This walks through the code used to generate each figure (1-6) output by the basic pipeline. Using this notebook, one can modify the code and / or change figure parameters easily.

(4) fast_iCLIP_metaanalysis.ipynb

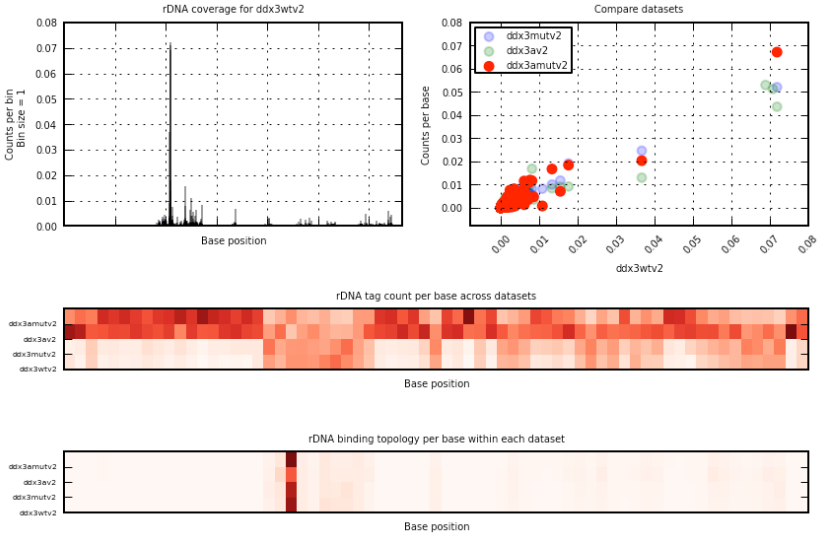
This notebook makes it easy to compare results across datasets. There are two primary analyses. (1) The first is a scatter plot of tag counts (normalized by sequencing depth) across genes of a specific type (e.g, protein coding) between datasets. See below:

```
# *** INPUT ***
plt.subplot(2,2,1)
target='ddx3wtv2' # Dataset (x-axis) against which others will be compared
toCompare=['ddx3av2','ddx3mutv2'] # Datasets (y-axis) for comparison
colors=['black','cyan'] # Colors
DF1=makeScatter(target,toCompare,'proteinCoding',colors) # Specify the gene type to compare
plt.tight_layout()
```



The second kind of metaanalysis compares the topology of binding across a specified class of RNA (e.g., any repeat RNA class analyzed). For example, examining tag count per position across datasets for rRNA is done using a scatterplot and two heat maps (one normalized per position for comparison of binding intensity at specific positions between datasets and the other normalized per dataset, providing a comparison of binding topology within each dataset).

```
# *** INPUT ***
repeatName='rDNA'
target='ddx3wtv2'
toCompare=['ddx3mutv2','ddx3av2','ddx3amutv2']
```



(5) fast_iCLIP_metaanalysis.ipynb

The final kind of analysis that is typically performed per dataset explores binding motifs using HOMER (http://homer.salk.edu/homer/motif/), which must be installed to perform the analysis. The first section of the notebook will automatically perform motif analysis on all reads in the dataset and / or reads from selected mRNA regions (e.g., 5' UTR). The second section of the notebook allows selection of specific mRNA regions (e.g., below) from the average binding topology diagram (Figure 2). Here, regions are the start codon are selected. The reads will be isolated and run through motif analysis (results are output to a specified directory).