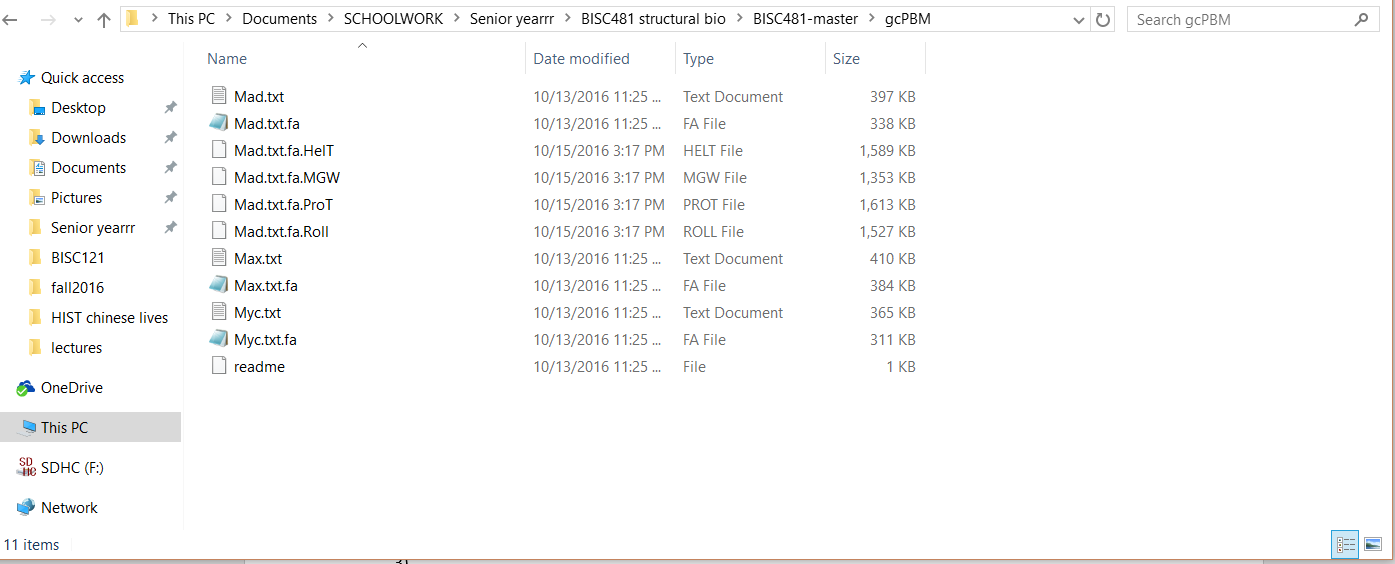
Colin Kunze BISC 481

Homework 3 10/25/16

1. See README.md
2. 1. Systematic evolution of ligands by exponential enrichment (SELEX-seq) is a technique that helps identify small DNA sequences that have the highest binding affinities for a particular protein. It begins by subjecting the protein to many random short sequences of DNA. The few sequences that do bind to the protein are kept, amplified, and used as a starting point for future cycles. The further cycles involve continuing to select the highest affinity sequences and sometimes altering them by deletions to find the best matching sequence.

Protein binding microarray (PBM) is a technique that allows you to determine the affinities of binding sites for a particular transcription factor. The transcription factor under question is tagged with an epitope and tested on an array of thousands of short known dsDNA sequences. A fluorophore-tagged antibody then binds to the epitope that is on the transcription factor. By measuring the intensity of the fluorescence in every cell, we can quantitatively determine the affinity of the transcription factor for that dsDNA sequence.

* 1. In Chromatin Immunoprecipitation (ChIP-seq), DNA is crosslinked and then sonicated. An antibody for the target protein is added and immunoprecipitated. Thus, we have complexes of the target protein and the DNA sequences nearest to it (that were crosslinked to it). We reverse the crosslinks and sequence all the DNA fragments that remained. Using motif discovery algorithms, we align the sequences and determine which nucleotides are probabilistically important in the protein-DNA complex.
  2. SELEX-seq is powerful in finding a strong match for the binding sequence, but it takes a lot of time and resources. The protein binding matrix is a good tool because it is able to track multiple proteins at the same time. However, both the protein binding microarray and SELEX-seq must be done in vitro. This means that the protein and DNA have been removed from a biologically active environment. The advantage of ChIP-seq is that we get a freeze frame of where the proteins might have been interacting with the DNA in a biologically active environment. The disadvantages of ChIP-seq include that we only look at one point in time, changes in the cell’s stresses may change the DNA that the protein interacts most with or has access to.

1. 
2. See code. Results (printed to console):

[1] "R-squared value of using '1-mer' model for Mad.txt"

[1] 0.7758688

[1] "R-squared value using '1mer' +'1-shape' model for Mad.txt"

[1] 0.863375

[1] "R-squared value of using '1-mer' model for Max.txt"

[1] 0.7853121

[1] "R-squared value using '1mer' +'1-shape' model for Max.txt"

[1] 0.8640762

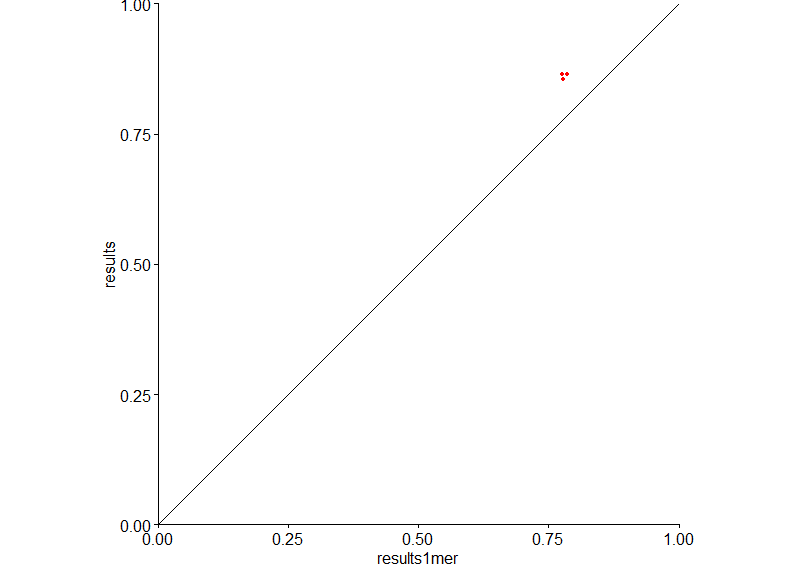
[1] "R-squared value of using '1-mer' model for Myc.txt"

[1] 0.7782286

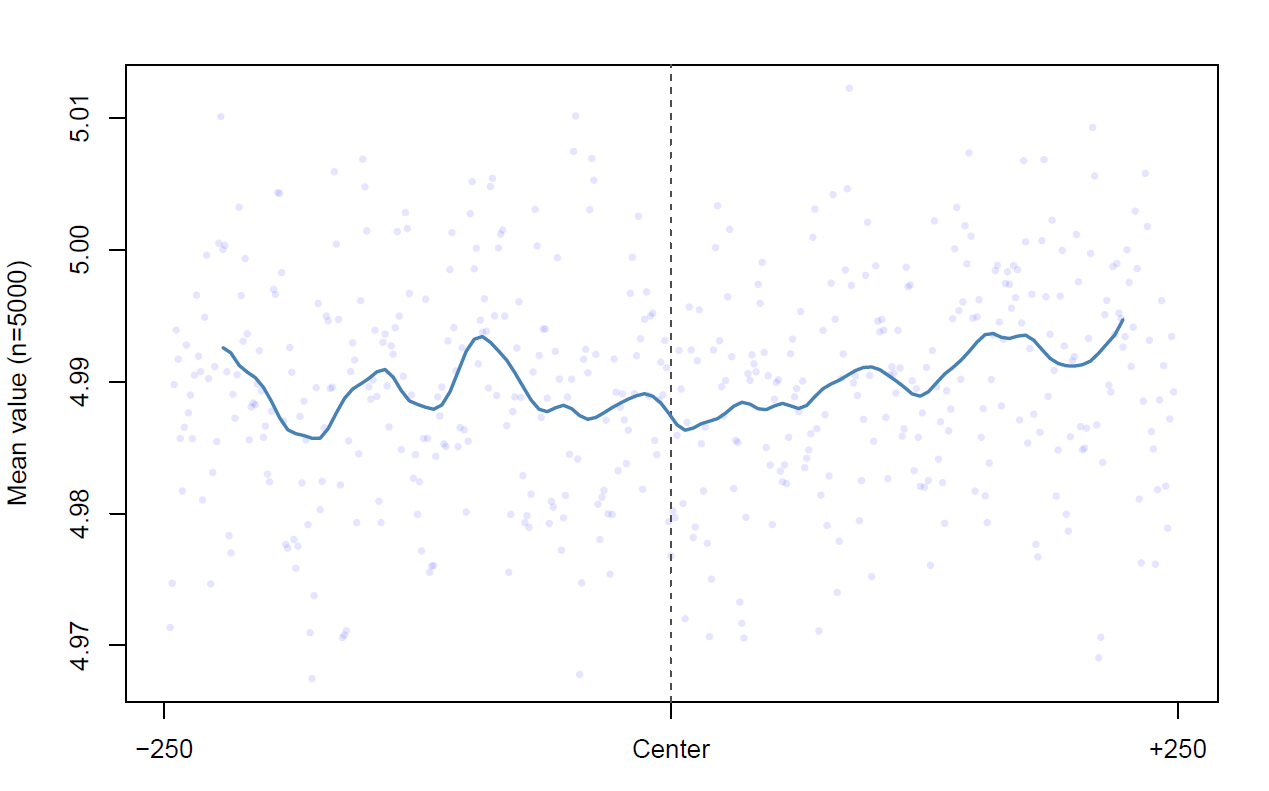
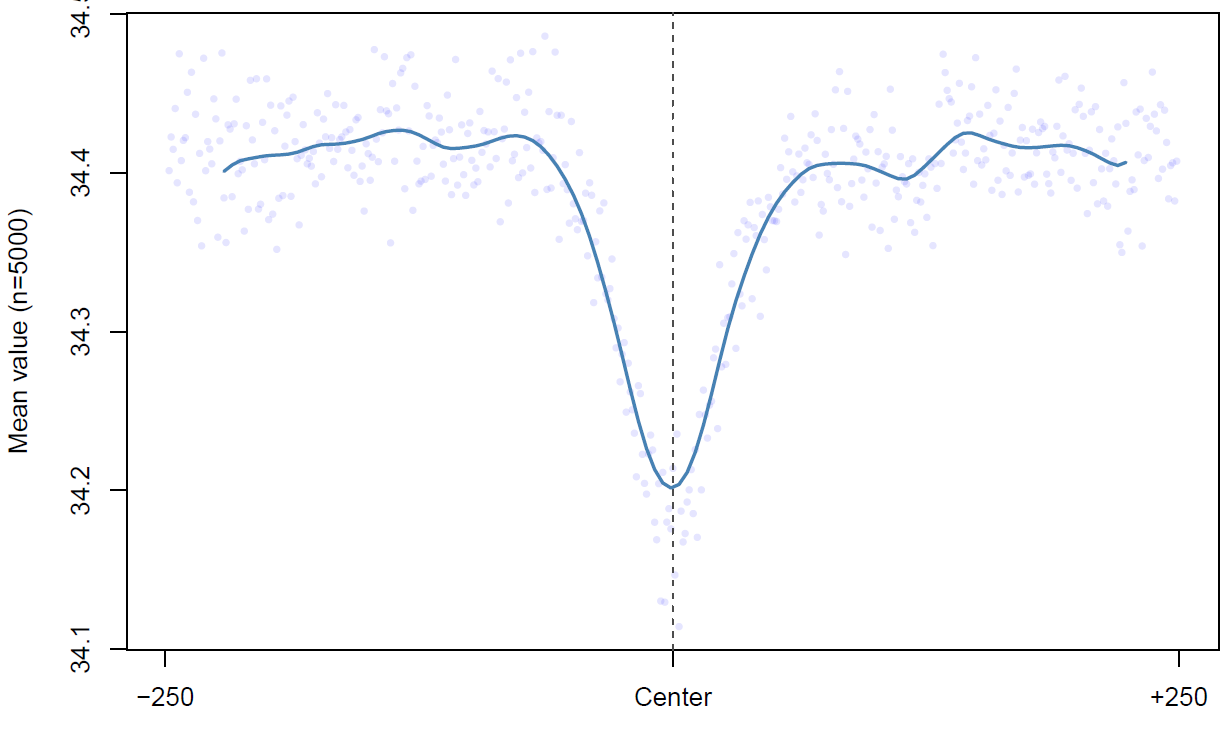
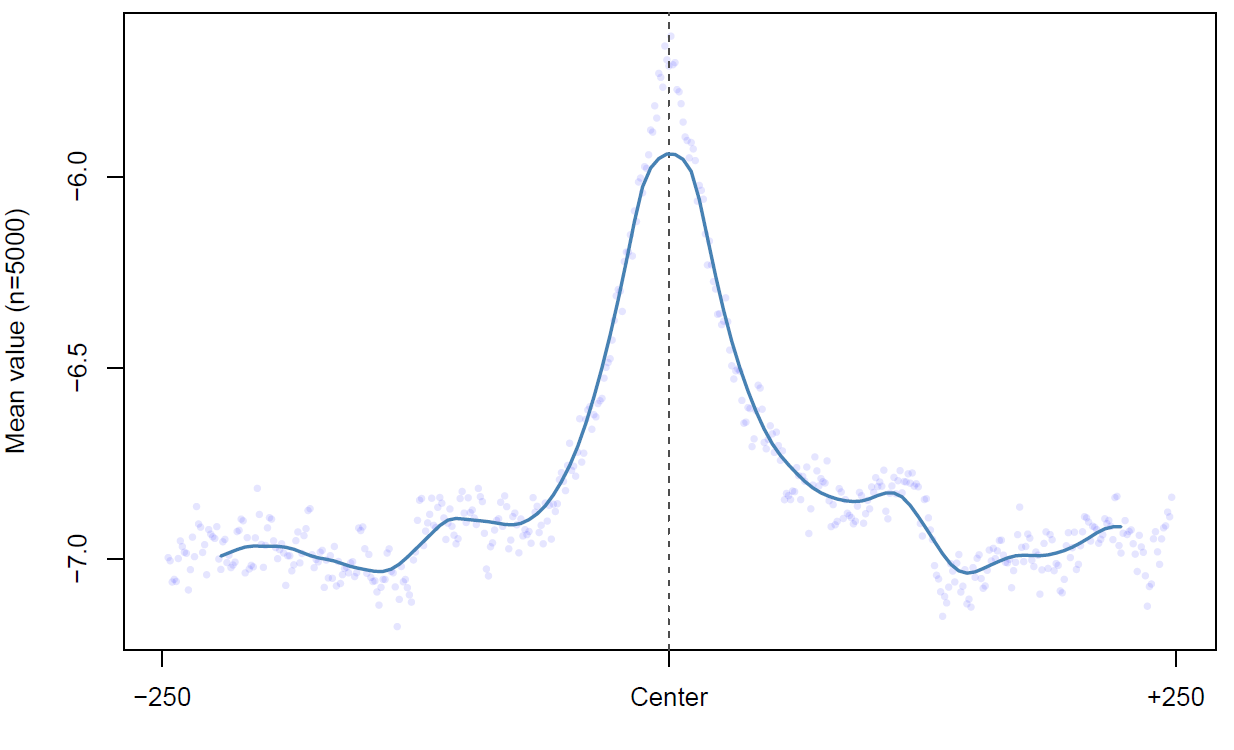
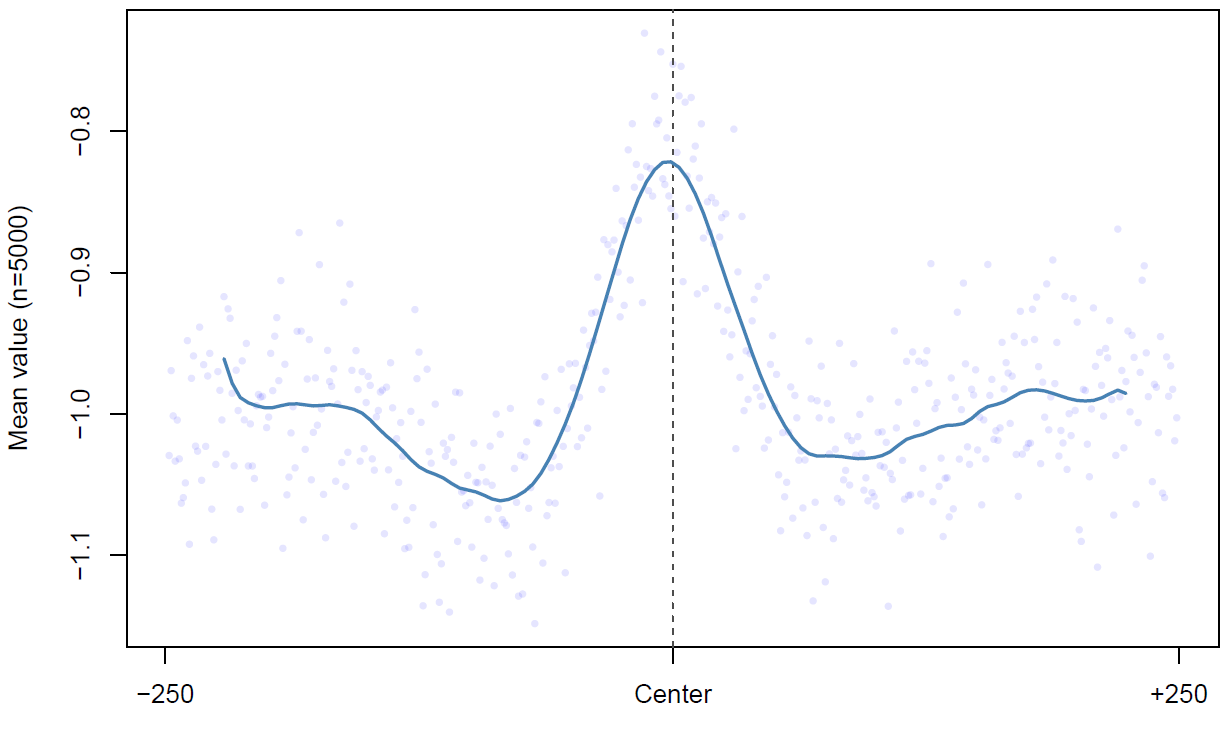
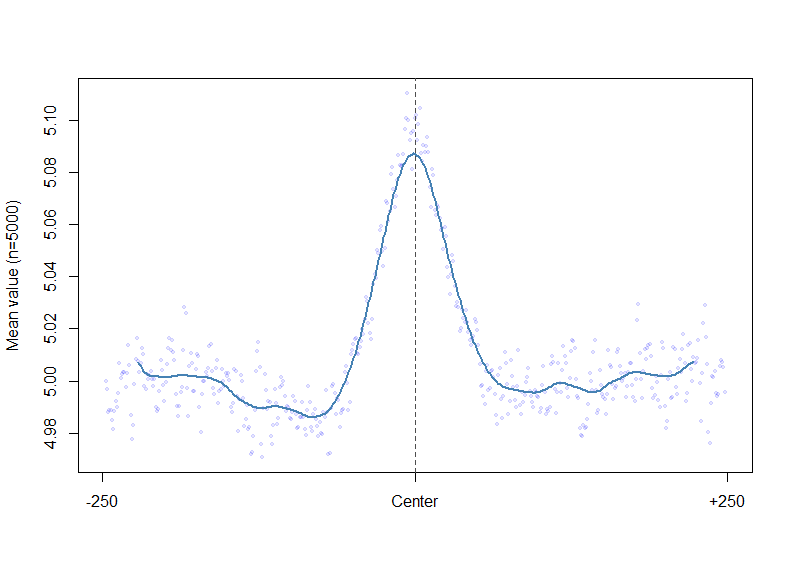
[1] "R-squared value using '1mer' +'1-shape' model for Myc.txt"

[1] 0.8545246

1. The plot shows that the r-squared values are higher in the “1-mer + 1-shape” model than in the “1-mer” model. If more sets of data confirm this comparison, then it is safe to say that the added parameters of shape also play a role in the transcription binding factor association.



R-squared value of the model with shape (results) versus the r-squared value of the model without shape (results1mer)



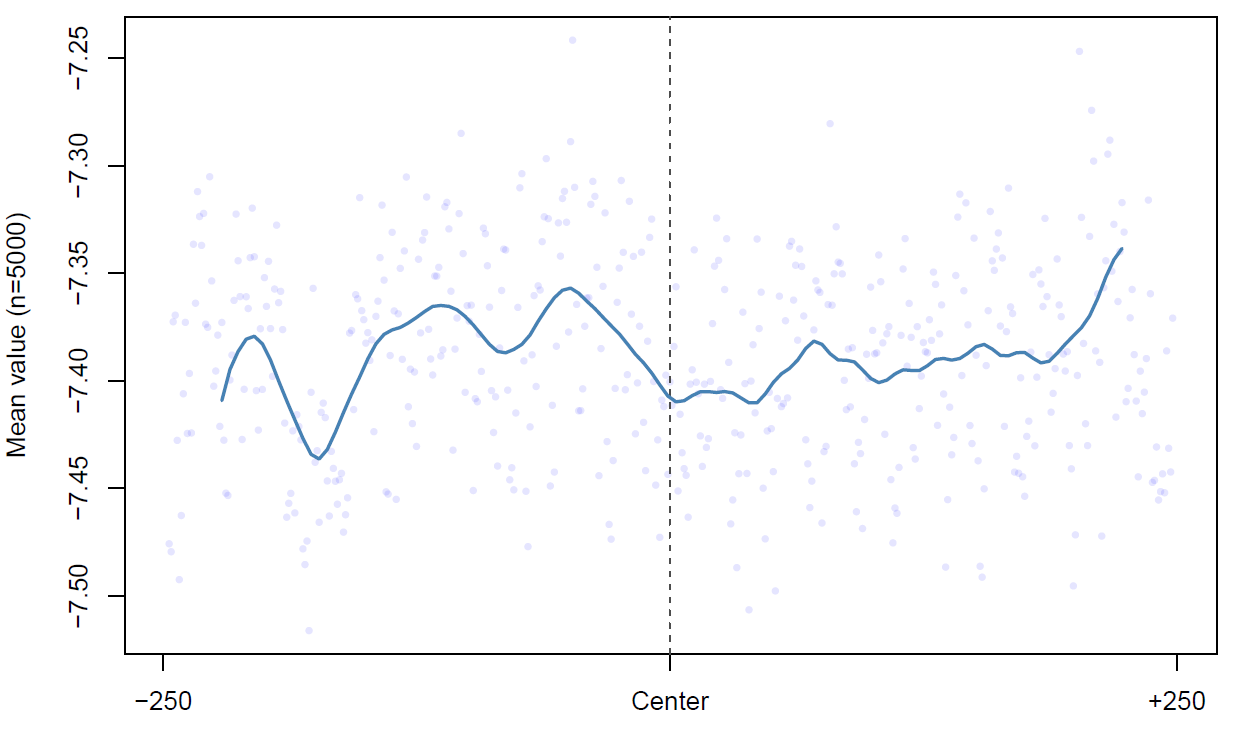
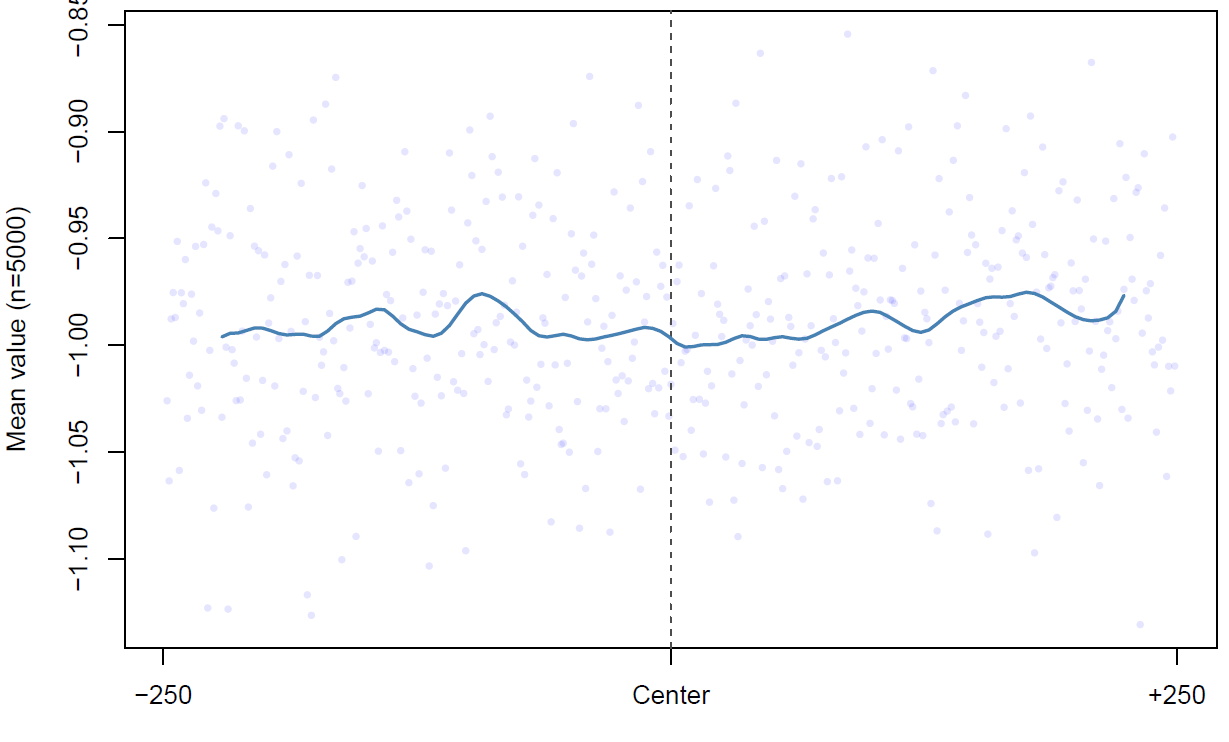
Bound Roll

Bound Minor Groove Width

Bound Propeller Twist

Bound Helix Twist

Unbound Minor Groove Width

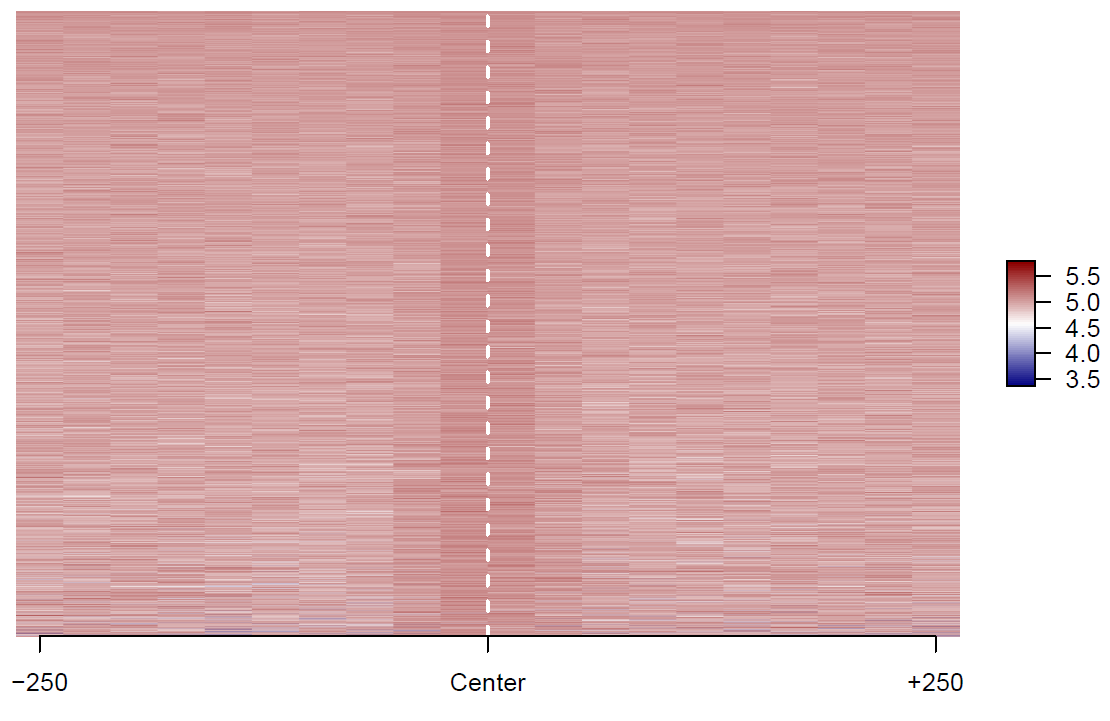


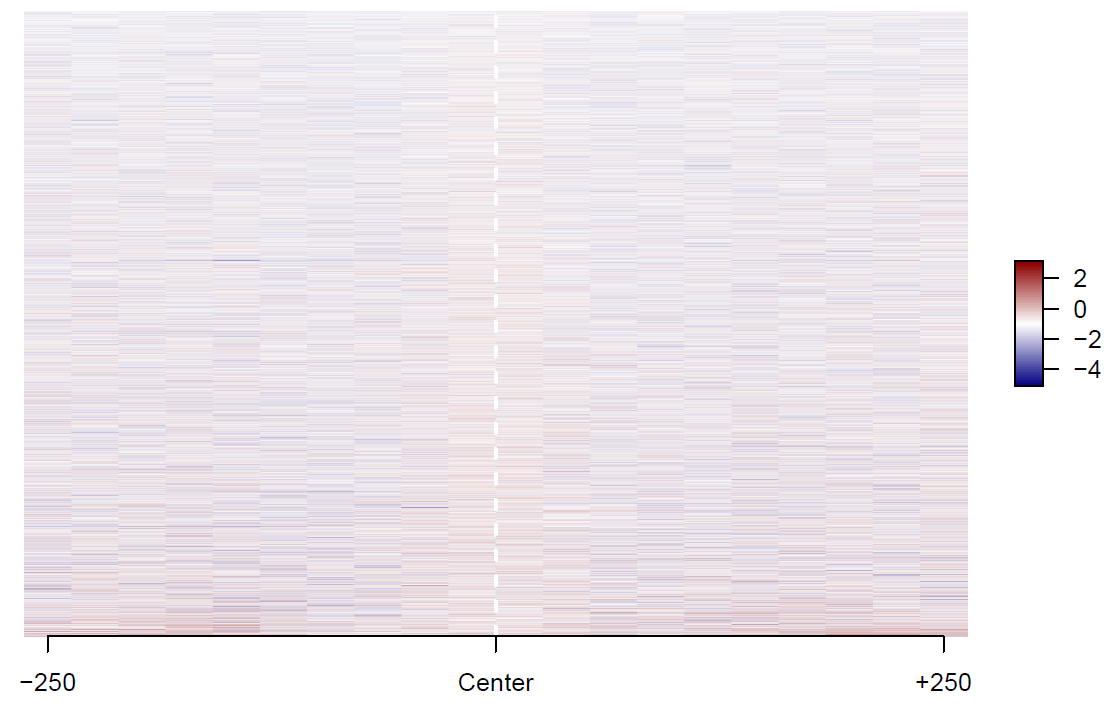
Unbound Roll

Unbound Propeller Twist

Bound Heat Map of Minor Groove Width

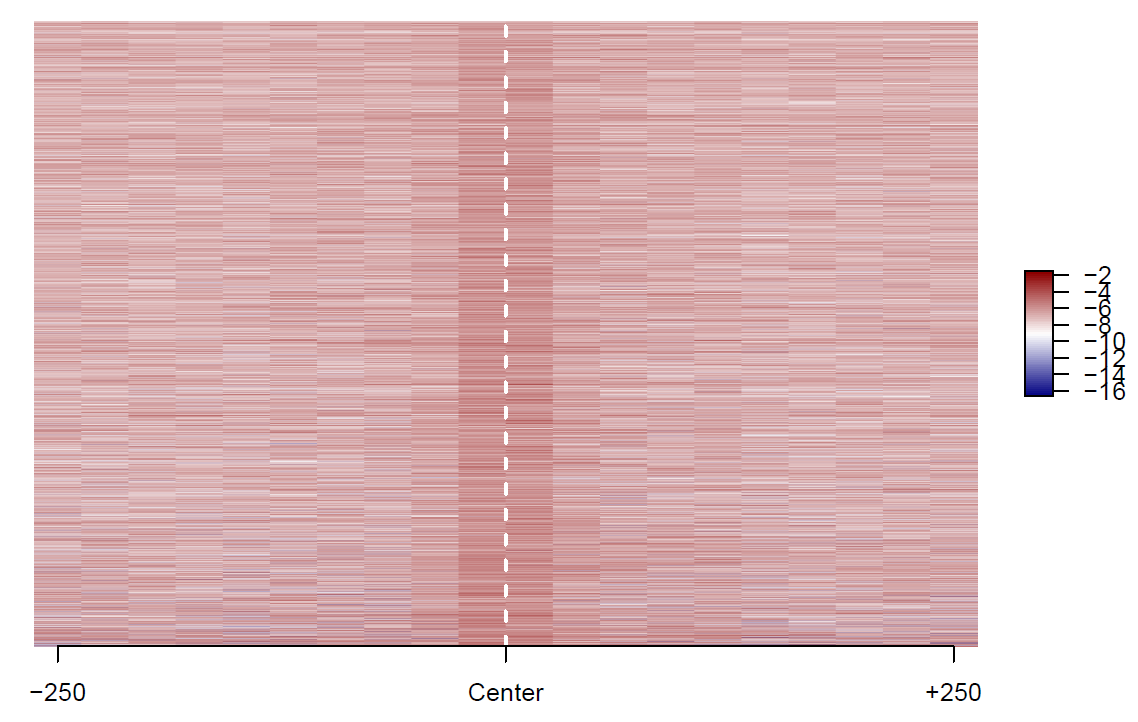
Unbound Helix Twist

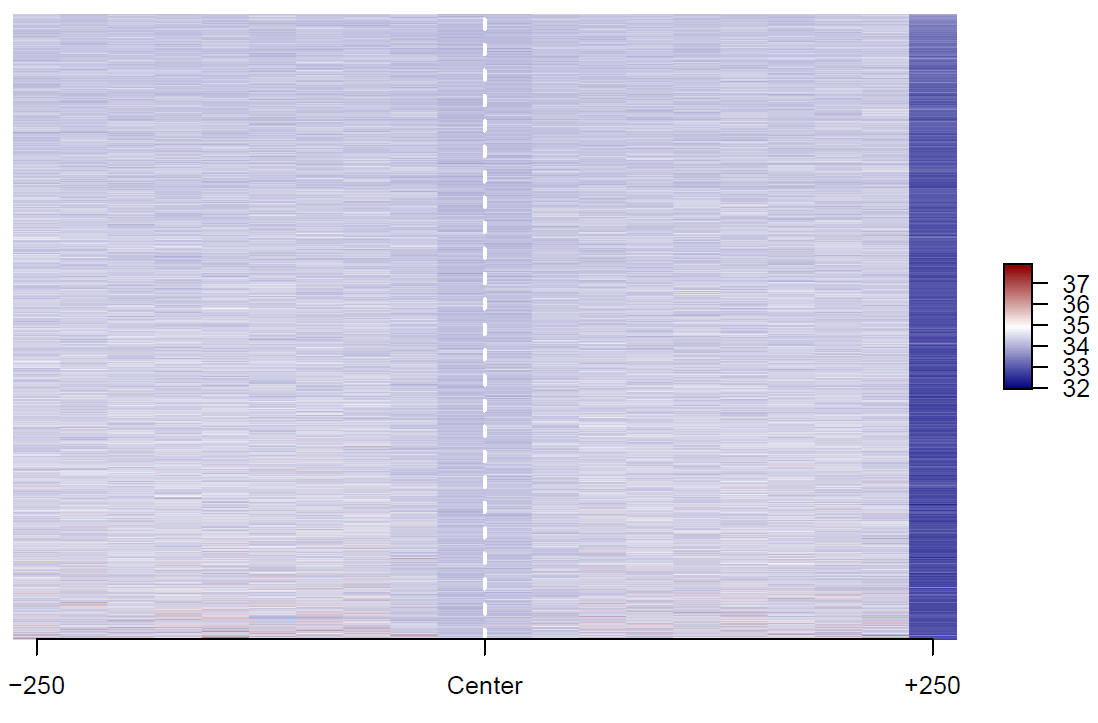




Bound Propeller Twist Heat Map

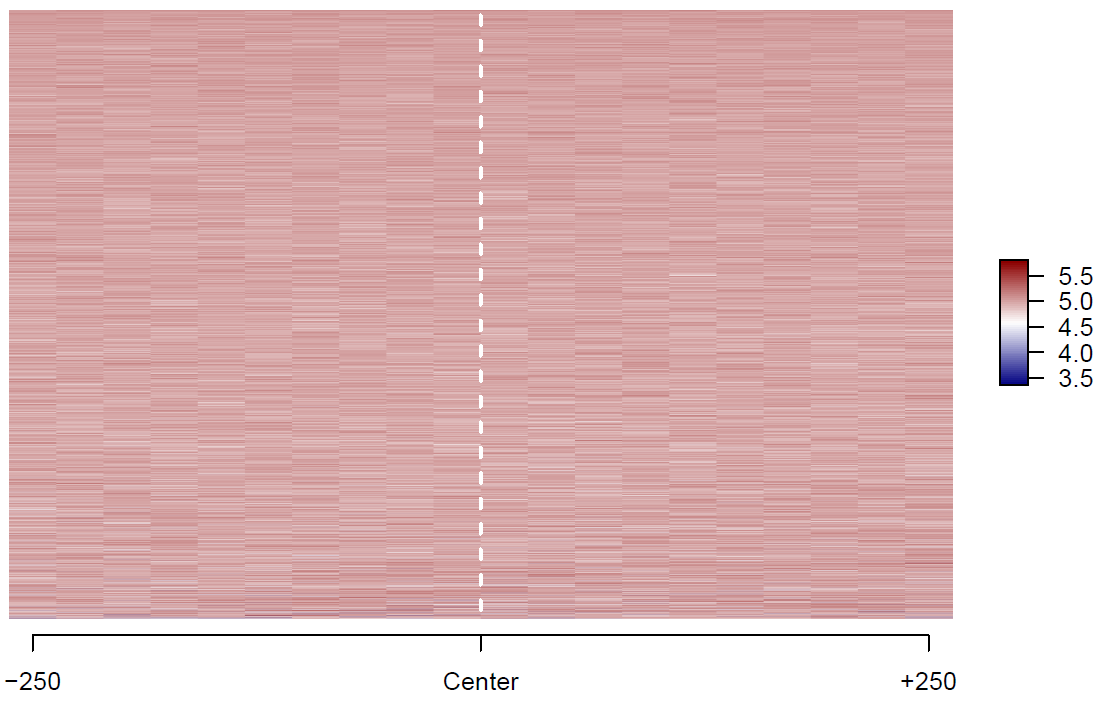
Bound Roll Heat Map

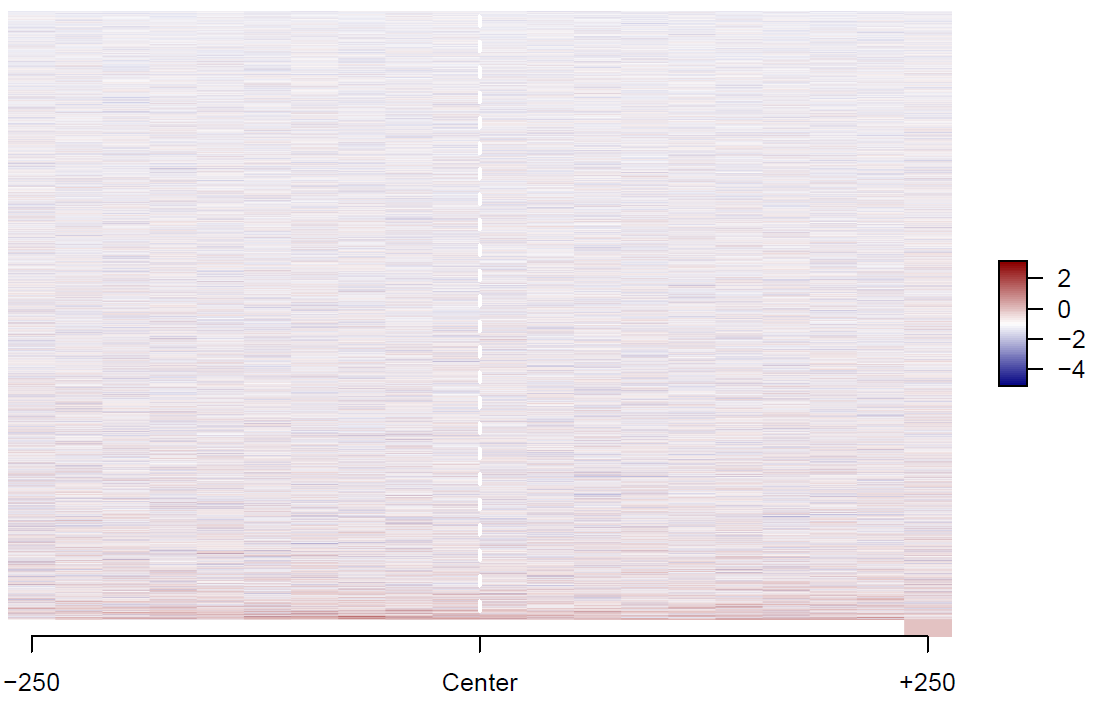




Unbound Minor Groove Width Heat Map

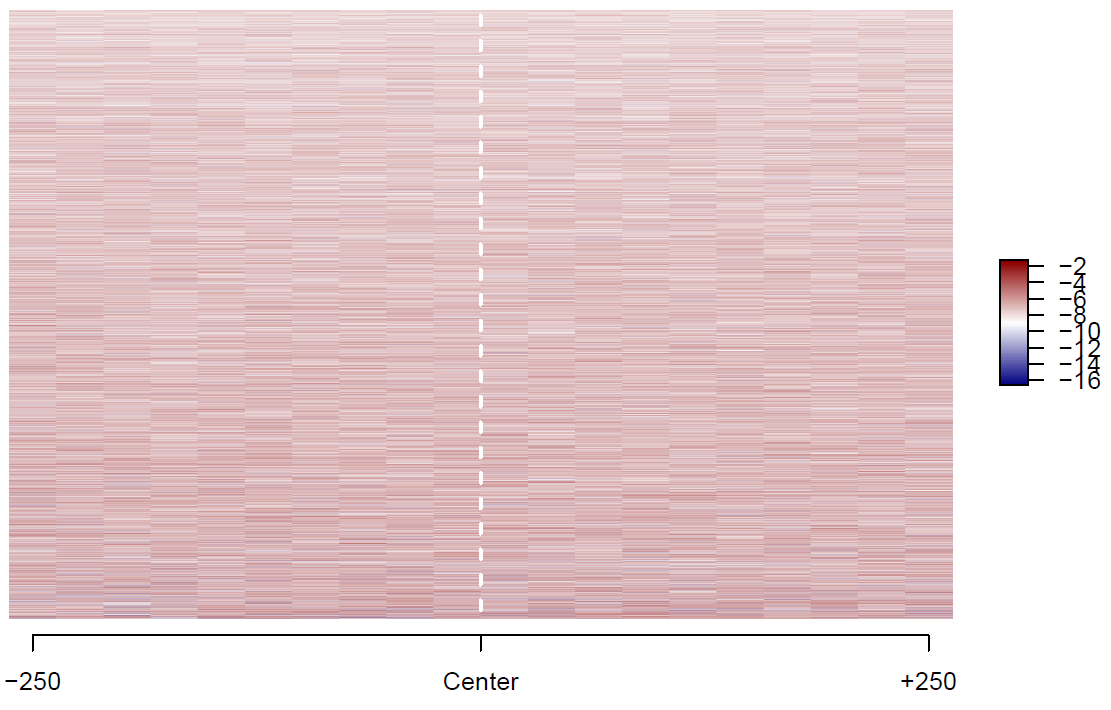
Bound Helical Twist Heat Map

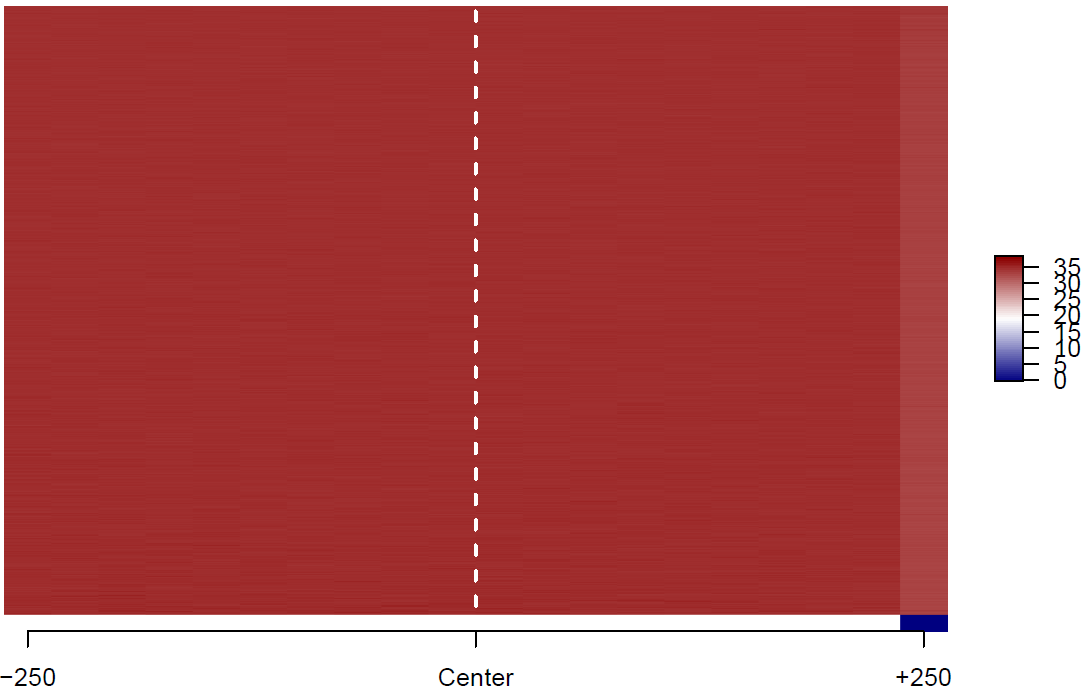




Unbound Propeller Twist Heat Map

Unbound Roll Heat Map

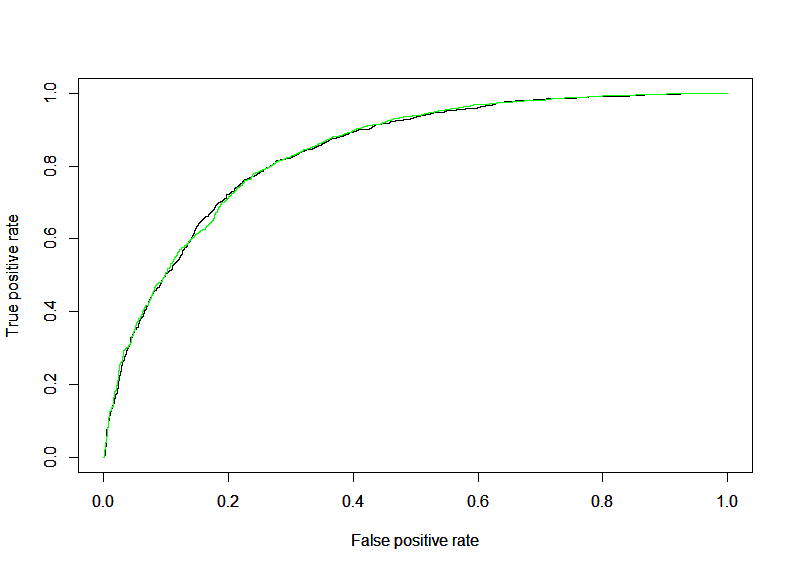




Unbound Helical Twist Heat Map

The results of the plots and the heat maps show that the shape of the DNA is very clearly non-randomly distorted during transcription binding. It implicates that the shape of DNA has a lot to do with its interaction with transcription factors. The plots all show some spike around the center of the transcription factor binding site while the heat maps also show some distinct difference in physical properties closest to the site of binding.

Note: for the helical twist and roll heat maps, a dummy column of 0s was joined to the end of the matrix to produce 500 values that could fit in 20 bins. Thus, the last bin may appear skewed.



AUROC for CTCF

0.8415 – 1mer

0.8398 – 1mer and 1-shape

The AUC values do not differ much for the bound\_500 versus unbound\_500 sequences.

The curve here represents the quality of our logistic regression. Maximizing the area under the curve (maximum theoretical area = 1) would give us a perfect logistic regression. Neither model shows a distinctive advantage in this category, showing that the shape of the DNA cannot help predict binding.