1. See GitHub

2. **<u>High-throughput binding assays:</u> Briefly describe (a) the in vitro experiments SELEX-seq and PBM, and (b) the in vivo experiment of ChIP-seq. (c) Compare and discuss the advantage and disadvantage of these methods.**

Answer: (a) **<u>SELEX-seq</u>** measures the binding of a single protein to millions of oligonucleotides over several enrichment cycles. When the set of bound oligonucleotides is retrieved, the protein is detached and amplified to make a new pool and a sample from that pool is then sequenced. The output is several sets of sequences in the end, each of which correspond to a different cycle. The **advantage** includes higher accuracy in predicting *in vivo* binding, owing to their ability to measure binding to longer k-mers. and the **disadvantage** includes results of advanced cycles that may suffer from over-specification, where high affinity k-mers are represented at the expense of low-affinity k-mers. **<u>PBM</u>**, or protein binding microarrays, are designed to measure protein-DNA binding intensity in a high-throughput and unbiased manner. A large array contains dsDNA probes with unique sequences designed to collectively cover all DNA 10-mers. The final output is the set of probe sequences and the binding intensity of a protein to each one. The **advantage** for this method is that it provides accurate measurements for studying TF specificity. It allows better ranking of k-mers according to their binding intensities, compared with HT-SELEX data. (b) **<u>CHIP-seq</u>** or chromatin immunoprecipitation combines the strategy of ChIP and high-throughput sequencing technologies and allows the detection of BSs throughout the entire genome. The **disadvantage** is that in vitro models are cheaper and easier, and therefore more appealing. CHIP-seq requires a large number of probes. However, *in vivo* is useful when combined with in vitro experiments. The **challenge** is that BSs are short and degenerate, and DNA probes are longer than typical BSs and thus may contain many putative sites, so it is difficult to distinguish between specific and nonspecific background binding (Orenstein, Shamir 2016). The **advantage** of CHIP-Seq is that it does not suffer from noise due to the hybridization step in ChIP-chip. However, intensity signals measured on arrays may not be linear in its entire range and its dynamic range is limited below and above saturation points. Lastly, the significant **advantage** is that the genome covering is not limited by the probe sequences fixed on the array; this is particularly important for analysis of repetitive regions of the gnome, which are typically masked out on arrays.

3. See R code

4. See R code

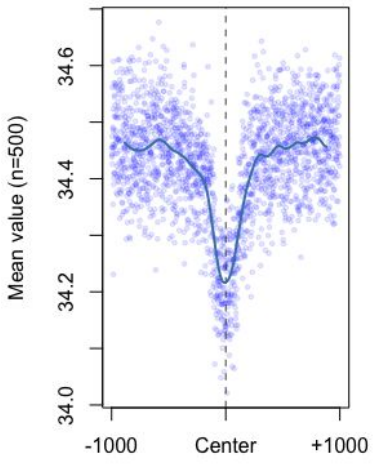
5. **<u>High throughput in vitro data analysis:</u>**

Answer: For this question I substituted “1-mer” for “sequence.” Sequence is on the x-axis and sequence-and-shape is on the y-axis. The 3 points representing R^2 discussed for *Mad, Max,* and *Myc* are (0.775,0.863), (0.785,0.865), and (0.778,0.855). In calculating the comparison between the two different models, the R^2 of the sequence-and-shape model is consistently higher than that of the sequence model, leading me to believe that the sequence-and-shape model is a better model to use in terms of accuracy and specificity. This

makes sense because additional information about shape should lead to a model with more information and better predictive methods.

6. See R code

7. <u>High-throughput in vivo data analysis:</u> plotShape() functions of DNAshapeR to generate ensemble plots for the DNA shape parameters of minor groove width, propellor twist, roll, and helix twist based on the sequences downloaded for question 6.



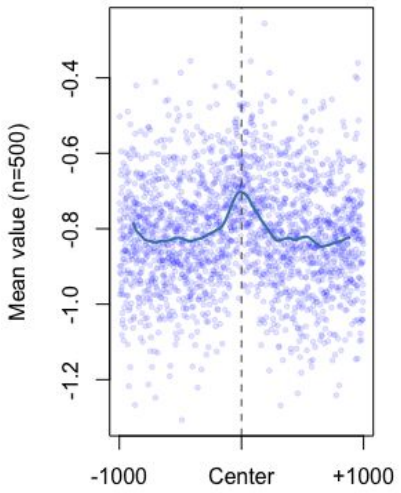
<u>HelT: helix twist</u>

For these graphs, the horizontal line indicates sequence. The center of the helix twist evaluation sees an inversion of the center. The center of the minor groove twist shows a mean value

around 5.10, where the propellor twist shows a -5.5 maximum, with a higher intensity than other graphs owing to the flexibility of the propellor twist sequence. The Roll graph shows a lower intensity and center of about -0.7.

MGW: minor groove twist

ProT: propellor twist



Roll

8. Build prediction models for in vitro data: Build logistic regression models for “1-mer” and “1 mer+shape” features, draw a plot of the ROC curves, and calculate the AUC score for each curve. Briefly discuss what you have learned from the results.
The AUC score for “1-mer” is 0.8405 and for “1 mer+shape” it is 0.8411. From a range of 0.5 to 1, the AUC scores 0.8405 and 0.8411 are both closer to 1 than they are to 0.5. This indicates that the results show a more “true positive rate” than a “false positive rate”