ANALYSIS OF BINDING SITES OF SOX FAMILY TRANSCRIPTION FACTORS

DEBDUTH BARDHAN PIJUSH

SOX FAMILY

- A family of transcription factors organized together by the presence of the HMG-box binding domain (part of protein)
  - The Sox genes are organized together because it has the HMG-box of a gene responsible for initiation of male sex-determination (SRY)
  - Sox = SRY-related HMG box
- The family regulates cell fate decisions during development
  - Sex determination and neuronal development

TELOMERES

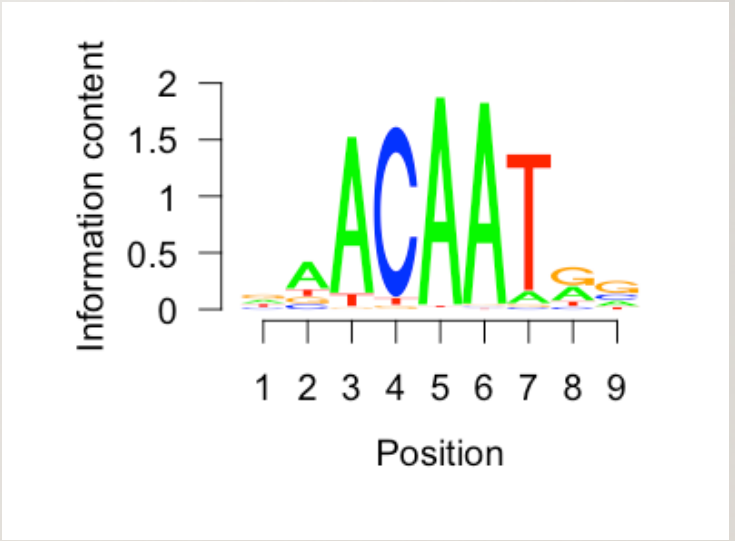
- The region of a repetitive sequence (TTAGGG) at the end of a chromosome
  - Protects the chromosome from fusing with other chromosomes and deterioration
- A decrease in telomere length happens over aging
  - 11 kb at birth to 4 kb in old age
- Sox2 is responsible for pluripotency of embryonic (undifferentiated) and neural stem cells
  - Telomerase, an enzyme that replenishes the telomere cap of the DNA, present only in certain types of cells like stem cells

HYPOTHESES

- Are there binding sites of Sox family transcription factors near the end of the chromosome?
  - If so, are the shorter binding sites in that region?
  - Since these regions go through much insertion and deletions, perhaps shorter binding sites would be present in the telomeric region?
- Is there any link to the location of the transcription factors' last binding site to the number of binding sites on the chromosome for the specific transcription factor?

METHODS

- MotifDb used to gather the position frequency matrices for H.sapiens from the Sox family and the jolma2013 database
  - Only Sox2, 7, 8, 9, 10, 14, 15, 18, 21 present in the database
- Sequence logos used to represent each transcription factor's binding site sequence based on the PFM (Sox9 shown to the right)
- UCSC database wasused for the H.sapiens genome
- Biostrings package was used to match the transcription factor sequences to each respective chromosome's sequence



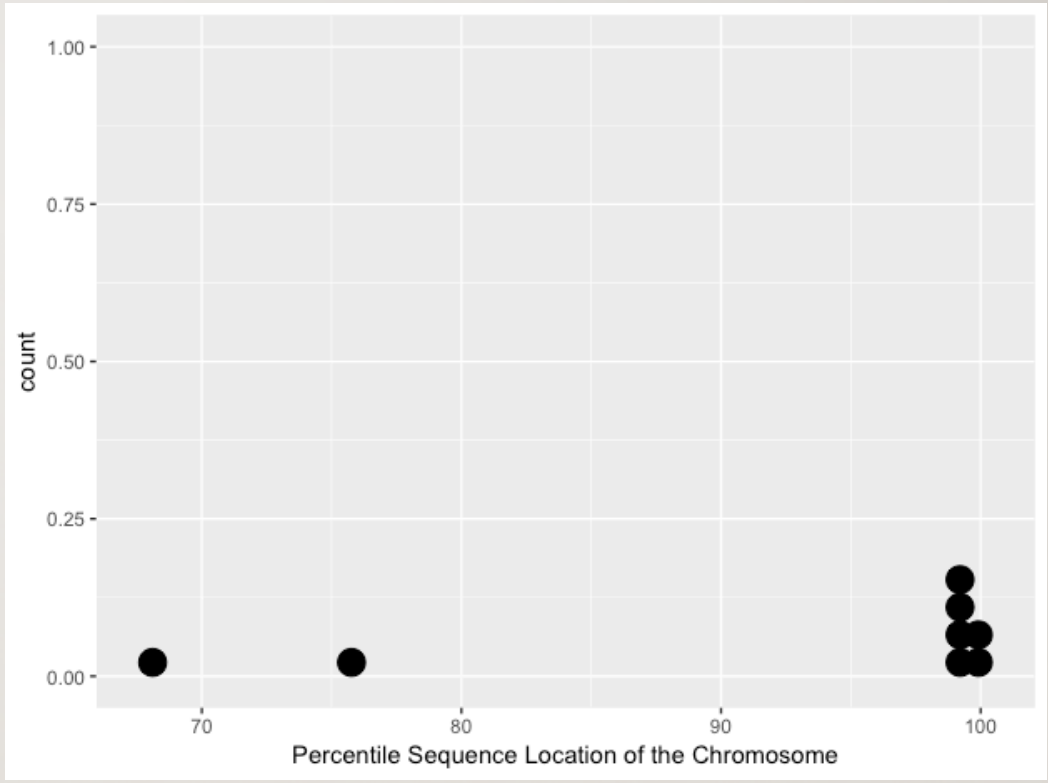
WHAT DATA WAS COLLECTED

- From the matching of the TFBS sequence to the chromosome:
  - The counts of matches were collected (number of binding sites)
  - The start site of the last binding site for each TF was collected – weighted by the length of the chromosome, giving a percentile
  - The length of each transcription factor's binding site

RESULTS

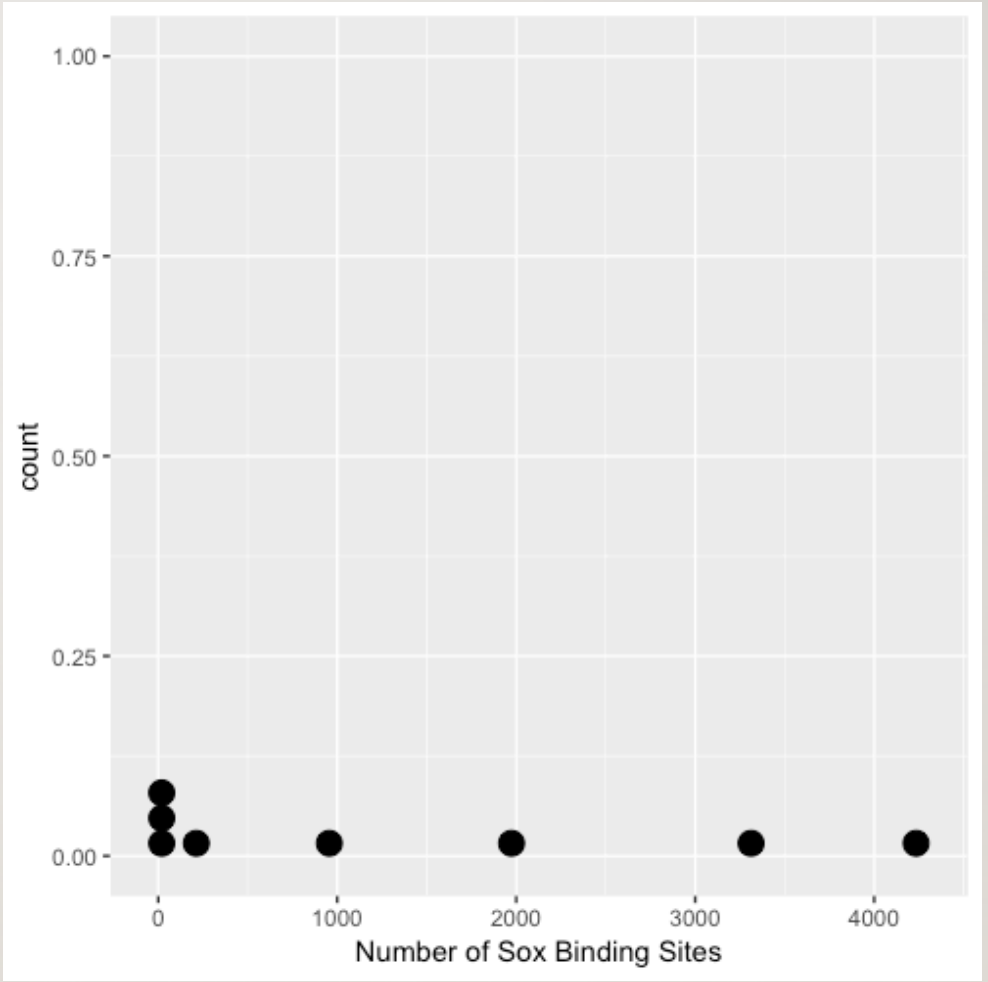
START SITES

- Sox2, Sox14, Sox21, Sox8, Sox9 and Sox18 all had start sites in the last 1 or 2 percentile of the chromosome, compared to the others.
  - This is near the telomere
- A t-test was ran between those 6 transcription factors against the others, giving a p-value of 0.08738 , meaning that it would be significant if the alpha value were considered to be 0.1.
  - Larger sample size needed of all the Sox family (better database)



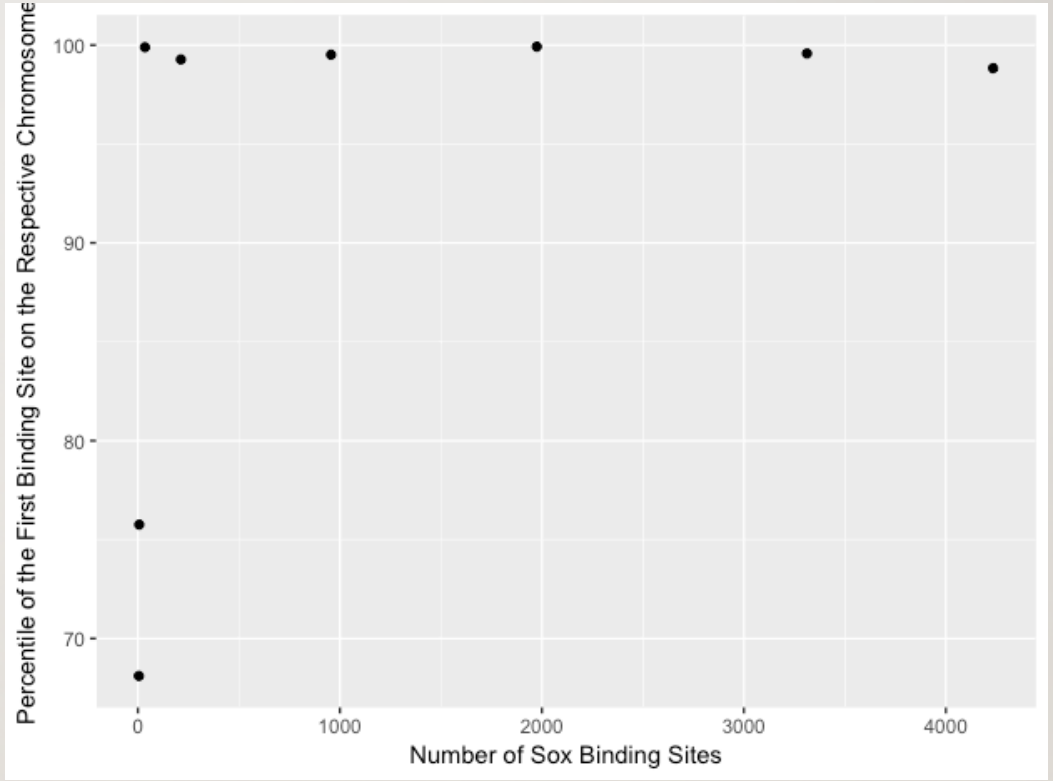
BINDING SITE COUNT

- Sox8, Sox10, and Sox15 had less than 50 binding sites each for their respective chromosomes



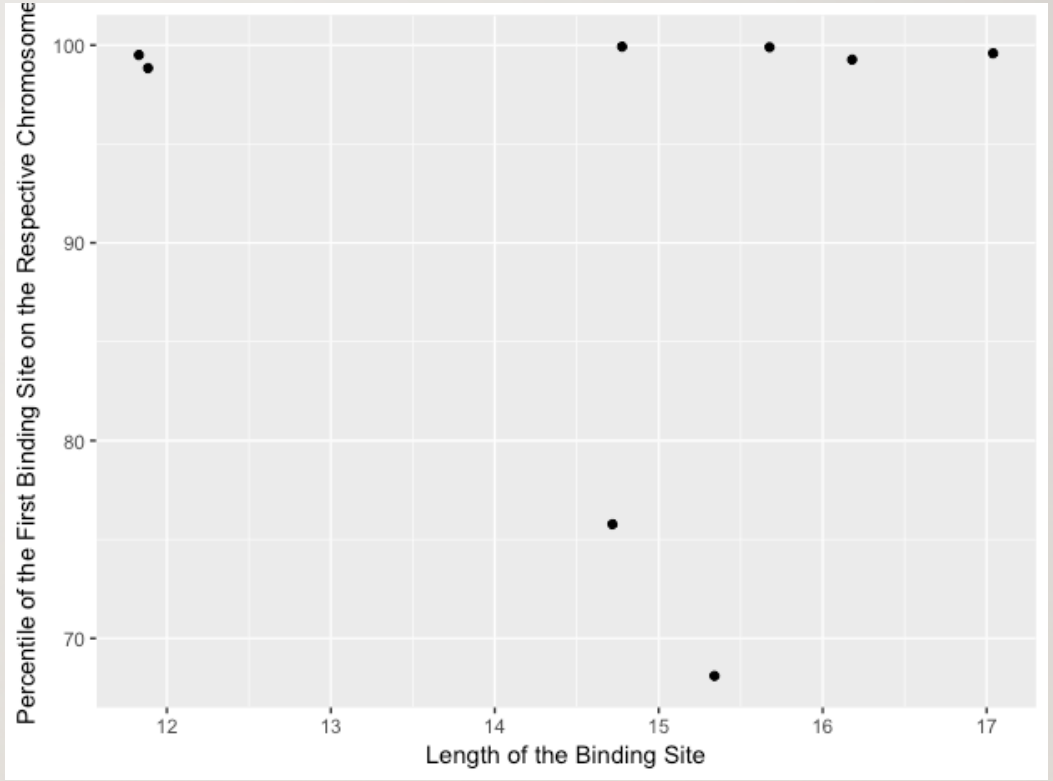
RELATIONSHIP BETWEEN LOCATION ON CHROMOSOME AND NUMBER OF BINDING SITES?

- Does not appear to have much of a relationship
- Coefficient of Determination was found to be 0.228
- The transcription factors with sites near the end of the chromosomes varied in number of binding sites



RELATIONSHIP BETWEEN LOCATION ON CHROMOSOME AND LENGTH OF BINDING SITE?

- R^2 = 0.005, extremely uncorrelated



CONCLUSION

- There are a considerable number of Sox family transcription factors that are located near the telomeres
  - Gives indication that they might be related
  - But, the length of the binding sequences have no correlation with the location of the last binding site
- There is no correlation between the number of binding sites for each transcription factor and the location of the last binding site