Insulator Analysis

- Computationally identify CTCF sites across a selected chromosome in human
- Obtain public CTCF ChIP-Seq data from 23 different human cell lines

- Extract sites that overlap between both sets (constitutive CTCF sites)

- Study chromatin states of CTCF occupied regions

- Retrieve closest candidate insulators (up and downstream of gene of interest)

- Analyse syntenic region in chick and mouse genomes

Enhancer Prediction

- Computationally predict enhancers within insulator regions using the software 'DREiVe'

Transcription Factor Binding Site Analysis

- Detect conserved TF binding sites in each enhancer using the RSAT toolkit

Make biological inferences
A. Obtain CTCF Position Frequency Matrix

B. Shuffle chromosome, scan against CTCF PFM and calculate ‘weighted’ scores

\[ w = \log_2 \left( \frac{(n + \sqrt{N}) \times b}{N + \sqrt{N}} / b \right) \]

Where,
- \( n \): no. of occurrences of current nucleotide in current column (e.g. "87" for A in column 1, "414" for G in column 2 etc)
- \( N \): total no. of observations, sum of all nucleotides in a column (913 in this case)
- \( b \): [background] frequency of current nucleotide

C. Scan actual chromosome against CTCF matrix, re-calculate ‘weight’ scores and compare distributions

D. Calculate False Discovery Rate (FDR) and P value

\[ FDR = \frac{V}{V + S} \]
\[ P = \frac{A}{B} \]

Where,
- \( V \): no. of sites in control sample (false positives)
- \( S \): no. of sites in test sample (true positives)
- \( A \): no. of sites of a given weight in control sample
- \( B \): total no. of sites in control sample

In above example, a weight score of “18” has an FDR and P value of 0 (optimal threshold). Sites of weight \( \geq 18 \) likely to be real CTCF sites

<table>
<thead>
<tr>
<th>Start</th>
<th>End</th>
<th>Pattern</th>
<th>Score</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19</td>
<td>GGCTTTTGCAATTAGCCGC</td>
<td>12.5</td>
<td>+</td>
</tr>
</tbody>
</table>