Be'ware of genomic nucleotide composition while hunting transcription factor binding sites in ChIP-based experiments

R. Hunt Newbury[1,2,4], A. Kwon[1,3], D. Arenillas[1,3], and W.W. Wasserman[1,3]
[1]Centre for Molecular Medicine and Therapeutics / Child and Family Research Centre, University of British Columbia, Vancouver, BC, Canada.
[2]Bioinformatics Program, University of British Columbia, Vancouver, BC, Canada.
[3]Department of Medical Genetics, University of British Columbia, Vancouver, BC, Canada
[4]Corresponding author: crebecca@cmmt.ubc.ca

The regulation of gene expression is a complex process fundamental to how a cell develops and interacts with its environment. One class of proteins involved in the regulation of genes are transcription factors (TFs), which recognize and bind to short sequences of DNA (TF binding sites – TFBS). ChIP-based experiments, such as ChIP-Seq, are commonly used to isolate sequences that TFs putatively bind within. After the sequences have been obtained, bioinformatics methods attempt to predict the presence of TFBSs within the target sequences using motif models, which exhibit poor specificity (i.e. high false positive rates). Statistical over-representation of predicted TFBSs within a set of ChIP-defined regions relative to the genomic background is often assessed as a means to link a TF to the regulation of a set of genes. However, nucleotide composition of the analyzed sequences can have confounding effects. We address the factors influencing the success of motif over-representation analysis and demonstrate the impact of different strategies to correct for the problem.