

Predicting tissue-specific transcription factor binding sites with chromatin modification information

Tom Whittington¹

t.whittington@imb.uq.edu.au

Andrew Perkins¹

a.perkins@imb.uq.edu.au

Tim Bailey¹

t.bailey@imb.uq.edu.au

¹ Institute for Molecular Bioscience, The University of Queensland, Brisbane, QLD 4072, Australia

Abstract

In silico prediction of transcription factor binding sites (TFBSs) is central to the task of gene regulatory network elucidation. Genomic DNA sequence information provides a basis for these predictions, due to the sequence-specificity of transcription factor binding events. However, DNA sequence alone is an impoverished source of information for the task of TFBS prediction in eukaryotes, as additional factors such as chromatin structure regulate binding events. We show that incorporating high-throughput chromatin modification estimates can greatly improve the accuracy of *in silico* prediction of *in vivo* binding for a wide range of transcription factors in human and mouse. Importantly, predictions made with the use of chromatin structure information are tissue-specific. This result suggests that the use of chromatin modification information can lead to accurate tissue-specific transcriptional regulatory network elucidation.

Keywords: Chromatin, Transcription Factor, Position-weight matrix

1 Introduction

We evaluated the use of chromatin modification information for improving predictions of transcription factor binding sites (TFBSs) *in silico*. We considered the chromatin modification H3K4me3 (trimethylation of lysine 4 of histone H3), which has long been regarded as a marker for open chromatin and actively transcribed genes [3]. The genome-wide distribution of this mark was recently characterised in several mouse and human tissues [4, 1, 2].

We show that this genome-wide chromatin modification information can be used to greatly improve the accuracy of genome-scale TFBS prediction for all 14 mouse transcription factors (TFs) and all ten human TFs considered. The improvement gained is consistently highest when the chromatin modification data is derived from that same tissue in which the TFBS predictions are being made, which indicates that our approach yields tissue-specific TFBS predictions. In addition, chromatin modification information yields better performance than simple filtering using either transcriptional start site (TSS) or phylogenetic conservation information, indicating that our approach represents a significant advance on existing methods for refining TFBS prediction.

2 Method and Results

We evaluate the usefulness of H3K4me3 distribution information when applied as a filter to predictions generated with a position-weight matrix model of TF binding. We employ gold-standard datasets based on available high-throughput estimates of *in vivo* TF binding locations. These gold-standard

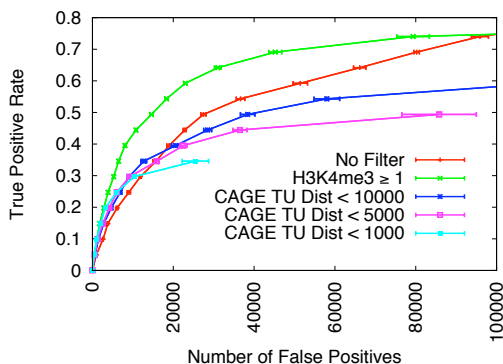


Figure 1: **Superiority of H3K4me3 information.** We compare H3K4me3 and TSS proximity filters for prediction of binding sites for TF Klf4. A subset of all considered TSS distance thresholds are presented for clarity.

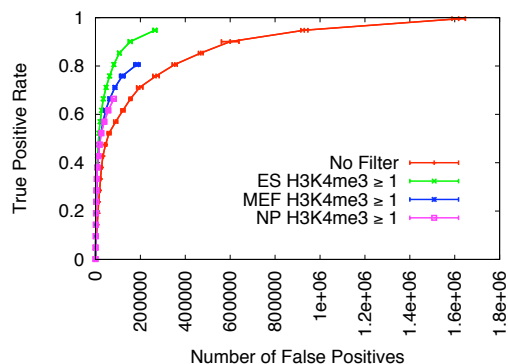


Figure 2: **Tissue-specificity TFBS predictions.** H3K4me3 information derived from ES cells proves more useful than MEF or NP information for predicting cMyc binding sites. The transcription factor gold-standard data are derived from mouse ES cells.

datasets allow us to evaluate predictive performance. Our accuracy metric is an ROC-like curve that plots the number of false positives (FPs) incurred at each in a range of true positive (TP) rates.

Performance of H3K4me3 filtering is significantly superior to TSS proximity filtering over a large range of sensitivities. Fig. 1 shows that a H3K4me3 filter of 2.0 attains an equal or higher specificity than a CAGE-based TSS filter, over all sensitivity rates, for the transcription factor Klf4.

We find that predictive accuracy is greatest when the H3K4me3 data is derived from mouse embryonic stem (ES) cells, rather than mouse embryonic fibroblasts (MEFs) or neural precursors (NPs). This indicates that binding site predictions are tissue-specific. The mouse cMyc TF exemplifies this outcome (Fig. 2).

3 Discussion

The ability to predict tissue-specific TFBSs is a clear advantage of our method, given that TFs can act in different regulatory frameworks in different tissue types. We expect that proliferating high-throughput chromatin modification datasets will facilitate tissue-specific regulatory network characterisation.

References

- [1] Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., *et al.*. High-resolution profiling of histone methylations in the human genome.. *Cell*, **129**(4), 823–837, 2007.
- [2] Guenther, M. G., Levine, S. S., Boyer, L. A., *et al.*. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, **130**(1), 77–88, 2007.
- [3] Kouzarides, T. Chromatin modifications and their function.. *Cell*, **128**(4), 693–705, 2007.
- [4] Mikkelsen, T. S., Ku, M., Jaffe, D. B., *et al.*. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.. *Nature*, **448**(7153), 553–560, 2007.