Identify functional patterns in high throughput binding assays

Project Report

Bioc7018

University of Queensland

Author:
Alex Essebier
Masters Student 42369305,
UQ

Supervisor:
Dr. Mikael Bodén
Associate Professor SCMB,
UQ

November 27, 2017

Contents

1 Introduction

Transcription factors (TFs) are a set of proteins with the ability to bind specific DNA sequences and regulate transcription [19]. They play a key role in regulating genes through the developmental process and maintaining homeostasis allowing the existence of more complex organisms. TFs and their ability to provide fine scale control over transcription of genes is a topic of significant interest to better understand the regulatory processes that take place within an organism. A number of epigenetic features including histone marks and chromatin state also influence the regulatory network of the genome. Both types of features can be explored through binding assays to investigate the influence they have on regulation genome-wide [10]. Mapping binding sites is the first step in better understanding TFs however, there are more layers of information not provided by the knowledge that one TF binds in a certain location. Understanding the binding mode or functional outcome of a TF binding at different binding sites is a more complex task that has yet to be well resolved. Identifying ways to obtain this knowledge for TFs and potentially other DNA-binding proteins is vital for deciphering the gene regulatory networks that allow for a wide range of biological processes [40].

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is an *in vivo* experiment to determine where a protein binds the genome. Proteins include transcription factors, DNA-binding enzymes, histones, chaperones or nucleosomes [2]. This approach is well established as a means of identifying where a single TF (or other DNA-binding proteins) will bind the DNA in a single cell type at a specific point in time. The binding sites are identified by analysing the read depth across the genome and identifying significantly enriched regions using peak calling algorithms [23, 28]. The results from ChIP-seq analysis are often treated as binary in the sense that a TF has or has not bound to a certain location. The other use of ChIP-seq data is to identify differentially bound locations when comparing two TFs or conditions. DiffBind is an R program that identifies significantly differentially bound sites by measuring differences in read densities [49]. Although this analyses density and magnitude of peak shape, it does so only to identify differences in binding patterns. No further exploration of the shape, magnitude and density of the read depth across all peaks has previously been conducted.

After ChIP-seq peak calling, all peaks above the required statistical threshold are treated equally for motif discovery, target gene discovery, location analysis, epigenetic analysis and other downstream analyses. It is unlikely that all peaks represent the same TF binding mode and investigating peak shape is an approach to identify this functionally relevant information.

ChIP-seq is a powerful resource and has provided invaluable information about a number of TFs and epigenetic marks. However, it is possible that there is information locked in ChIP-seq data that we are not yet taking advantage of. Recently, Sherwood et al. (2014) [45], showed that modelling the magnitude and shape of genome-wide DNase I hypersensitivity profiles allowed identification of TF binding sites. Several other algorithms also exist to infer TF binding from DNase-seq data [17, 38, 5]. These approaches are examples of new information being gathered

from a technique not designed to provide that specific information.

We propose that ChIP-seq peaks from a TF experiment can be clustered based on their shape, density and magnitude leading to identification of different binding modes and functional patterns of a TF.

1. Develop a modelling technique to identify functionally relevant clusters, based on peak shape and magnitude.
    - 1.1. Investigate available clustering approaches to identify one suitable for clustering ChIP-seq data
    - 1.2. Identify and implement the algorithms required to perform clustering, using the selected technique, on ChIP-seq data
    - 1.3. Process ChIP-seq peak data to create a dataset appropriate for clustering using the selected technique
    - 1.4. Explore peak shapes represented across clusters and across TFs to identify any common shapes or patterns
2. Identify biologically relevant data associated with cluster groups to expand knowledge of TF binding modes, functional patterns and interactions
    - 2.1. Investigate the biological significance of peak shape on TF binding location
        - 2.1.1. Compare the distributions of all peaks and their locations within each cluster
        - 2.1.2. Compare the enrichment of individual locations within each cluster
        - 2.1.3. Identify global patterns linking peak shape to genomic locations
    - 2.2. Investigate the relationship between peak shape and epigenetic data
        - 2.2.1. Explore how peaks within each cluster are annotated by the Broad ChromHMM data
            - 2.2.1.1. Compare the distributions of chromatin annotations within each cluster
            - 2.2.1.2. Compare the enrichment of individual chromatin annotations within each cluster
        - 2.2.2. Explore whether peaks within each cluster are enriched for specific epigenetic marks according to the ENCODE dataset (e.g. H3K27ac)
        - 2.2.3. Identify global patterns linking peak shape to epigenetic data
    - 2.3. Investigate the relationship between peak shape and sequence by analysing motifs of different clusters

1.1 Role of transcription factors

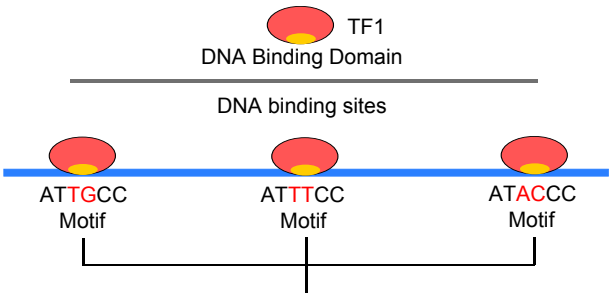
TFs play an important role in gene expression, cell differentiation and homeostasis. TFs are also key metabolic and developmental regulators. A TF is defined as having one or more DNA-binding domains (DBDs) which encode a sequence-specific DNA-binding module. Each TF is classified by the type of DBD present in the protein. There are four superclasses of TF based on broad structural similarities: basic, zinc-coordinating, helix-turn-helix and β-scaffold. Domains outside these four classes are referred to as 'other' [53]. Depending on the TF, hundreds to tens of thousands of binding sites can occur throughout the genome. The two most common locations for TF binding sites (TFBS) are promoter and enhancer regions. Genes consist of a number of elements when observed in the 5' to 3' direction beginning with the promoter, 5'UTR, transcription start site (TSS), exons and introns, transcription end site (TES), 3'UTR and poly-adenylation tail. Genes are regulated by enhancer and promoter regions which recruit the transcription machinery according to TF binding and other signals. Promoter regions occur in the 5' region near the TSS. Enhancers can be located at a greater distance from the TSS, can be upstream or downstream and can even occur within introns. They can operate from a distance using DNA looping.

When a TF binds, there are two outcomes for gene expression; activation or repression. Each TF has the potential to either activate or repress its target gene and this outcome is dependent on many of the factors that also influence TF binding.

The interaction between a TF and DNA is complex. To understand the relationship, multiple variables have to be taken into account. A number of these variables and their impact on TF binding are discussed including some existing experimental approaches available to tease apart the complex network of interactions. These variables include chromatin structure and nucleosome occupancy, and interactions between multiple TFs. TFs do not act alone but form a TF complex that allows activation or repression of the target gene. The basal transcription machinery, containing general TFs and Pol II, is required to control transcription. Through another complex called Mediator, the basal transcriptional machinery connects with TFs binding the DNA [7]. TFs can bind a broad range of DNA sequences yet manage to control regulation and transcription on a very fine scale. It is the network of interactions between TFs and the DNA structure which allow for such fine scale control [46].

1.2 Transcription factor binding

Sequence-specific DNA-binding activity is mediated through the DBD however, not all DBDs have been classified and sequence-specific DNA-binding can occur without a known DBD. The DNA sequences recognized by TFs are degenerate and relatively short; between 4 and 20 base pairs (bps). The sequence is referred to as a motif and can be represented by a consensus sequence, position weight matrix (PWM) or tables of affinities (or relative affinities) to individual sequences as seen in Figure 1. There is no 'gold standard' to represent the actual sequence preferences of a TF and current motif models may not be sufficient to do so [19].



Position Weight Matrix (PWM)

Position	A	T	C	G
1	1	0	0	0
2	0	1	0	0
3	0.33	0.66	0	0
4	0	0.33	0.33	0.33
5	0	0	1	0
6	0	0	1	0

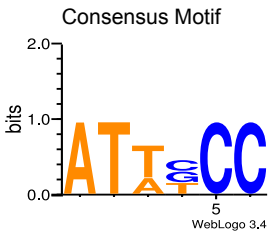


Figure 1: A simple schematic demonstrating a hypothetical TF, how it may bind to DNA and how those binding events are generalised. Three hypothetical binding sites are demonstrated each with variations in the binding sequence. A PWM is shown based on the frequency of each letter at each position in the three different sequences. A consensus motif then provides a visual representation based on the data represented in the PWM.

To understand transcription in an organism requires detailed knowledge of all TFs, and their binding affinities to all possible DNA sequences [25]. Knowledge of any co-operative binding interactions is also required. Currently, this information is mostly unknown. For example, mammals have an estimated 1300-2000 TFs but databases such as Jaspar and Uniprobe, which contain *in vivo* and *in vitro* data about DNA-binding motifs and preferences of TFs, only list 590 and 515 non-redundant proteins respectively [20, 25, 36]. Very few bound sequences have been identified and for many that have been, only a consensus sequence exists which lacks detail. A consensus sequence represents the DNA sequence bound by the TF with the highest affinity for that sequence while hiding low affinity or uncommon sequences. Binding between sites and TFs is not binary and some sites will be bound more strongly and/or frequently than others. The affinity between a TF and DNA sequence can be affected by a range of factors. This includes flanking sequence of domains in the TF, sequence polymorphisms in the target DNA sequence, site accessibility within chromatin regions, cooperation and competition between TFs, concentration of TF and cell type [19, 53, 25, 13, 16]. Aim 2.3. will explore variations in the motifs of different peak shapes and separate binding events based on sequence preference. This will highlight how not all TFs bind consistently to one sequence but instead a degenerate set of sequences that can be separated based on peak shape.

Two interactions allow the relatively high affinity of a TF to a specific DNA sequence. Non-sequence specific interactions with the DNA backbone are low affinity and allow the TF to slide along the DNA until the high affinity, sequence specific interactions with DNA bases immobilizes the TF to its target site. This immobilization occurs for a sufficient amount of time to allow

regulation of transcription to occur [25]. The major groove of DNA has substantial potential for hydrogen bonding and van der Waals' interactions. Many TFs interact with the major groove of DNA, but minor groove and phosphate and sugar backbone interactions are also frequent [53, 26].

A TF can be shown to have many predicted binding sites with equal predicted binding affinity based on *in vitro* experiments. *In vivo*, few of these binding sites will be occupied. For a TF to occupy a site, the site must be accessible. Chromatin higher order structure and nucleosome occupancy can prevent TF binding by restricting access to sites throughout the genome. An open or permissive chromatin landscape will allow access of TFs to DNA. This landscape can be influenced by a range of epigenetic features as well as TF interactions which can recruit chromatin modifying factors [16, 19]. Aims 2.2. will investigate variations in the epigenetic landscape around different peak clusters and determine how peak shape is related to site accessibility, chromatin landscape and other epigenetic marks such as H3K27ac.

1.3 Transcription factor families

Within the superclasses of TFs there exist TF families that use related structural motifs for recognition [39]. Members of the same TF family will often have highly similar DBDs and higher order structure but they are rarely identical. They typically have similar DNA binding specificities yet are capable of executing unique functions *in vivo*. However, when assayed *in vitro*, they generally do not show large differences in binding specificity with consensus motifs consistent between family members [47]. This indicates that small changes in the target sequence carry significant meaning when comparing TF binding events.

1.4 Transcription factor domains and interactions

TFs can contain multiple DBDs as well as other domains such as effector domains. TF effector domains can interact with a number of different partners and play a range of roles in transcriptional regulation. Unlike DBDs, effector domains are more physically unstructured and far less conserved increasing their binding potential. They can interact with basal transcription machinery, other TFs, and directly or indirectly recruit histone and chromatin modifying enzymes. Cooperative interactions between TFs expand the possible DNA sequences recognised *in vivo* as well as the binding energetics of the protein-DNA interaction. This influences both the affinity and the outcome (e.g. target gene expression levels) of TF binding [13, 47]. Exploring the motifs of different clusters according to Aim 2.3. will highlight interactions between TFs by identifying secondary upstream or downstream binding sites using spaced motif analysis (SpaMo) [54].

1.4.1 Dimerisation

A number of TF DBD domains allow binding as monomers including zinc finger and homeodomain containing TFs. Other domains require dimerization to bind to DNA such as TFs containing basic

helix-loop-helix (bHLH), basic leucine zipper (bZIP) or Rel homology domains. Oligomers or higher order complexes of three or more TFs can also exist [22, 14].

Formation of dimers between TFs is mediated, in some cases, through effector domains. More generally, the interaction between two TFs often occurs through the DBDs or adjacent domains [14, 13]. Dimers are protein-protein interactions between TFs that generally form in solution prior to binding DNA. Heterodimers form between two different TFs while homodimers exist between two of the same TFs. Dimerization can both stabilize the structure and induce conformational changes. In the case of TFs which require dimerization, such as those containing bHLH domains, dimerization between the DBDs on the two TFs stabilises the dimer allowing contact with the major groove of DNA to be made. bZIP dimers also function in a similar way [14]. Dimerization increases the control possible from a limited number of TFs by allowing a large number of combinations to exist and potentially play different roles. One way dimers can alter transcription is through the recruitment of different cofactors and transcriptional machinery [46, 29]. The binding sequences recognized by TFs in dimers can have similarities to the single TF binding sequences but variations are expected. Heterodimers can recognize half-binding sites comparable to the individual TF's motifs arranged in head to tail formation. Heterodimers can also form after DNA binding has produced allosteric changes in the protein structure of one TF allowing dimer formation [29].

1.4.2 Cooperation, competition and cofactors

Cooperation between TFs is a mechanism that contributes to the fine scale control exerted by TFs. It allows an increased number of regulatory changes to be made by different combinations of TFs working together. Three classes of cooperative interactions have been proposed in TFs. The first, and simplest, involves direct contact between two or more TFs and is described as mutually cooperative binding. The second class describes two TFs that do not bind cooperatively to DNA but both bind to a third TF which confers cooperativity and is described as indirect. The third mechanism describes the additive effect of TFs which bind near one another but lack protein-protein interactions. This class has an unclear mechanism. [37].

Competition between TFs for binding sites is dependent on the TF copy number, the number of available binding sites and the affinities to the available binding sites. For example, when two TFs A and B can bind to the same site, which one will be preferred? If A has a depleted copy number or other available binding sites with higher affinity, B will be more competitive and more likely to bind to the site than A [6].

Transcription cofactors can function both independently and in cooperation to fine-tune promoter activity. They achieve this by linking a TF to the transcription complex without requiring direct binding to the DNA [50, 42]. Latent specificity describes a cofactor induced change in recognition of sequences. It is a theory which describes how differences in amino acid sequence within a TF family may only impact DNA recognition and binding specificity when bound with cofactors [47].

2 Deconvoluting transcription factor binding

All of this evidence paints a complex picture of TF binding in an organism with intricate interactions and a large number of variables. There is not a straightforward process in which a TF binds to a promoter using a specific, non-degenerate DNA sequence, recruits transcription machinery and influences the outcome of gene expression. Rather, a TF can bind as a monomer, homodimer or heterodimer, or oligodimer to a variety of degenerate DNA sequences. This can only occur when the chromatin landscape allows access and no other competitive features are present on the DNA. Attempts to unravel some of this information have involved use of both *in vitro* and *in vivo* experimental approaches.

2.1 *In vitro* approaches

In vitro experiments use purified and/or synthetic components in a test tube. Footprinting and electrophoretic mobility shift assay (EMSA) are two *in vitro* techniques that are used to identify TF binding sites where the motif or binding sequence is unknown [25]. Protein binding microarrays (PBMs) and one-hybrid interaction analyses are *in vitro* methods to investigate specific DNA sequences that proteins bind rather than locations. By investigating TF affinity or specificity to a specific set of sequences, a clearer picture of what the TF can and cannot bind is provided. PBMs allow all possible DNA sequence variants of a given length k to be assigned a binding specificity for a TF using a single microarray [4]. This allows in depth analysis of motif variation, nucleotide preference and binding site preference at very high resolution. The biggest downside to *in vitro* methods is that they do not represent the cellular environment which is known to have a significant impact on where, and with what affinity, TFs will bind.

2.2 ChIP-seq

ChIP-seq experiments provide a snapshot of genomic sites occupied by the protein of interest, in this case a TF, in a specific cell line or sample. Once the sites have been obtained, the sequence of each of those sites can be analysed to discover a consensus sequence to which the TF may bind. ChIP-seq is an important validation tool for *in vitro* methods because it describes the natural environment in which the protein-DNA interaction is occurring [25]. A summary of the ChIP-seq experimental approach, data processing, quality control and peak callers is included in Appendix A.

Due to the nature of peak calling algorithms and their stringent cut offs, a significant amount of data is lost when ChIP-seq data processing is performed. This includes true binding sites with low affinity or specificity or that may only be present in a small population of cells in the full sample. Also, smoothing of the peak profile can eliminate two distinct peaks that were located near one another [2]. The biggest problem with ChIP-seq experiments is that the experimental steps produce false positive because DNA is pulled down in the immunoprecipitation (IP) step

when a true binding event is not present. Then, the stringent data processing steps introduce false negatives to counteract these false positives. ChIP-exo is an *in vivo* experimental approach that addresses these concerns as well as problems related to degenerate motifs and low occupancy binding [44]. As a newer technology there are not as many datasets available as for ChIP-seq and the approach has not yet become mainstream in investigating TF binding. Exploring ChIP-seq instead will allow more meaningful information to be extracted from the large number of existing datasets including ENCODE [9].

2.3 Challenges to ChIP-seq results

ChIP-seq experiments provide information on where TFs bind and based on these binding regions, motifs can be determined from the sequence. It has also been shown that tag densities at binding sites indicate binding affinity with higher densities showing higher affinity [27]. A number of factors can confound the ChIP-seq process. Many of these, specific to TFs, have been mentioned previously and include dimer formation, chromatin landscape, co-localization, DNA looping to promote interactions, co-factors and in general, the fact that TFs do not act independently on transcription. TFs and histone modifications as well as other DNA-bound proteins are known to work together to carry out cellular functions. When comparing multiple experiments, if there are indications that different proteins or modifications are binding at the same genomic location, distinguishing between true co-binding events is difficult. The proteins or modifications may be present in the same genomic location but in different cells and may appear to be working together when they are not [15]. MEME is a tool that combines a number of programs all designed to process ChIP-seq peaks and perform motif analysis [3]. SpaMo is one element of the suite that attempts to infer physical interactions between the given TF and any TFs bound at neighbouring sites [54]. Aim 2.3. will investigate how different peak shapes relate to different motifs or binding partners.

Other examples of confounding effects are more complex to solve. Situations exist where the TF of interest is bound by its antibody and is not itself bound to DNA but instead to another TF. The peak describing this interaction will not contain sequence specific to the TF of interest but rather the TF interacting with DNA. A similar situation can occur when the TF of interest is bound to an enhancer but has promoted looping in the DNA to bring the enhancer region and promoter of interest together. In this situation two peaks, one from the true binding event and one from the interaction between multiple TF, can be present in the ChIP-seq results creating confusion [10].

In a situation where a dimer is formed, the antibody may be specific to one half of the dimer pulling down information in which unexpected motifs may be observed. MEME can help to deconvolute the interaction in some of these cases by reporting motifs for the full dimer or the two half sites [3, 54]. The antibody could also lose specificity when the TF of interest is in a dimer preventing identification of a range of sites.

3 Extracting information from ChIP-seq peaks

Using an unsupervised model which clusters peaks based on the similarity of their peak shapes will allow exploration of how peak shape relates to a number of biologically relevant factors addressed by Aim 2. Following Aim 1.3., a dataset will be created that describes each peak and the associated read depths across a segmented window of set size. By analysing read depth the problem is very similar to clustering gene expression patterns.

3.1 Clustering approaches

Common methods for clustering gene expression data include K-means clustering and self organizing maps (SOM). Using a Dirichlet approach to cluster data is something that has not previously been tested. The method is, however, suited to the problem of clustering the peak shapes within ChIP-seq data. Dirichlet distributions have previously been used to perform protein multiple sequence alignment (MSA) but not clustering of gene expression data [55]. Aim 1.1. will determine which clustering approach is optimal for this problem. Hierarchical approaches were not considered because Dirichlet clustering performs a full partition and the nested aspects of hierarchical clustering restrict conclusions being drawn about specific and unique peak shapes.

3.1.1 K-means

K-means clustering is a common, but basic, method for clustering gene expression data. It is a partition based method which uses an objective function to assign data points to cluster centres which are data points themselves. It is available in R using Euclidean distance as the method to calculate distance between data points and cluster centres. To be effective at clustering, normalization of the data is required. K-means is also sensitive to noise in the data [12, 24].

3.1.2 SOM

SOM are a partition based approach which use a single layered neural network to map each data point to an output neuron with the closest reference vector. Output neurons are organized using a grid neighbourhood structure so similar clusters are also closer neighbours. There are a large number of variables in SOM including the output neuron grid structure, the distance function and the input data. SOM are very sensitive to noise, particularly datasets that contain invariant or flat patterns. Data also requires normalization prior to clustering [30, 24]. A SOM algorithm is available in R through the Kohonen package [51].

Both methods are capable of clustering the read depth data that describes the ChIP-seq peaks. The performance of Dirichlet clustering will be compared to these two existing methods according to Aim 1.1.. The benefits of Dirichlet clustering are attractive if it can perform as well as, or better, than existing methods. K-means and SOM approaches require normalization of data

prior to clustering so the first benefit of using a Dirichlet approach is that it does not requiring normalization.

3.1.3 Dirichlet

Dirichlet clustering using a Gibbs sampling approach creates a Dirichlet mixture model based on a set of histograms. This mixture model is made up of Dirichlet distributions representing each cluster. Each Dirichlet distribution in the model generates a probability distribution which will always add to 1. The probability distribution is discrete and multinomial constraining the original input data to that of a histogram or set of counts. This approach to clustering histograms is well suited to ChIP-seq peak data where read depth can be counted across a peak region.

Each peak is represented by a set of counts based on the read depth at intervals across the peak (restricted to a 500bp window around the peak summit). This creates a set of count vectors or histograms across all peaks. After each round of clustering, each Dirichlet distribution has a set of alpha values which represent the 'shape' of the cluster and act as the centre of the cluster. This distribution changes as clustering proceeds and each data point is categorically clustered by calculating the probability of it belonging to each cluster and assigning it to the cluster with the highest probability. Updating the model based on new cluster assignments and repeating this process moves data toward the 'best' clustering outcome based on a likelihood value (eq. 7 [55]).

The second key benefit of Dirichlet clustering is that the sum of the alpha values for each peak indicates how much evidence or support is available in the original data. This allows peaks to not only be separated by shape but also based on evidence and support. For example, two peaks X and Y can share a similar shape but peak X has a raw sum of 200 counts and peak Y only 30. Peak X has more evidence to support the shape and will be placed into a different cluster to peak Y. In K-means or SOM clustering, these two peaks would be treated equally as normalisation prior to clustering will mask the raw counts. In Dirichlet clustering, the raw counts are available to the algorithm and this information is taken into account during clustering. A high alpha sum indicates a strong concentration of data around the centre of the cluster suggesting peaks with strong evidence belong to the cluster. A low alpha sum indicates a weak concentration or low level of evidence in the peaks belonging to the cluster.

To calculate the alpha values and perform clustering, a Gibbs sampling approach is used where the description length (DL) of the data given the model is minimised as convergence is reached [55]. DL is used in two capacities in this implementation of the Dirichlet algorithm; DL of the data given the model and DL of the model itself. It is a statistical measure which can be thought of as representing the complexity of the model based on the information available. The quantity of data passed to the model has the most influence on the complexity outcome. To measure convergence, we are calculating the DL of the data given the model as the model parameters are updated [55]. That is, we are attempting to identify the best state of the current model for the data provided.

3.1.4 Optimal cluster number

In any clustering technique the identification of optimal cluster number is the biggest challenge. Most techniques require the user to specify the cluster number including all three approaches discussed here. Without a priori knowledge of the data or having an expected number of clusters making this selection is difficult. Algorithms exist to calculate the optimal number of clusters for each of the three approaches.

For Dirichlet clustering the optimisation method is known as the Minimum Description Length (MDL) principle. It requires training of multiple models for comparison to identify the one that minimizes the DL of the data given the model, plus the DL of the model itself. For each model, the best state to represent the data is identified and summed with how complex the model, described by a Dirichlet mixture, is. The complexity of the model itself is also related to the data but relies more on analysing the data parameters and content where as the first DL looks at data probabilities. The complexity or DL of the model will always increase as more clusters introduces more information that require more complex explanations. The model that minimises the two DL calculations is the one with the optimal cluster number [55].

3.1.5 Semantic similarity

When clustering gene expression patterns it is common to perform functional analysis. This provides an insight into the function of the resulting gene sets. A successful clustering will result in clusters that have different functional patterns. It is therefore possible to compare clustering approaches by comparing the functional patterns of the resulting gene sets using semantic similarity. Semantic similarity provides a quantification of the pairwise similarity of every cluster in every approach based on enriched GO terms. Ideally, clusters will show less similarity when the gene expression profiles have been clustered effectively. GOSemSim is a package available for R that takes two gene sets (i.e. cluster 1 and 2 from the Dirichlet approach or cluster 1 from Dirichlet and K-means), discovers enriched GO terms for each gene set then performs a comparison of the enriched GO terms taking into account the GO hierarchy and semantic meaning of the words. This test will form the basis of the comparison for Aim 1.1..

3.2 Peak analysis

Features that could influence cluster formation through peak shape include affinity, location, chromatin landscape, epigenetic markers, motif, physical interactions, dimers and more. After successfully clustering peaks using a Dirichlet approach, three key analyses will help identify the functional information linked to peak shape in individual TFs according to the Aims of the project: location analysis, epigenetic analysis and motif analysis. Understanding these three features and how they relate to different clusters as well as each other will provide significant functional information.

Identifying functional patterns from ChIP-seq peaks will have widespread applications to a number of datasets and experiments. It would lead to valuable new information being extracted

from the thousands of ChIP-seq datasets that already exist. It would refine our understanding of how TF binding modes play a role in gene expression, development and disease.

4 Methodology

4.1 Data collection

Eight TFs from the ENCODE dataset were selected for study each from a single cell type except for two, RAD21 and MAX, which had two cell types investigated. TFs included: GABP (H1Hesc), MAFK (HepG2), MAXGm (GM12878), MAXH1 (H1Hesc),)RAD21 (GM12878), RAD21 (H1Hesc), RXRA (H1Hesc), SP1 (H1Hesc), SRF (H1Hesc) and TBP (H1Hesc). For selection they required raw data in the form of bam files, processed narrow peak files, a reported motif and a cell type that had been included in the Broad Chromatin HMM analysis. Bam files and narrow peak files were downloaded from Factorbook. Where multiple replicas were available, one was selected for download randomly. The Chromatin HMM track was downloaded from UCSC.

4.2 Clustering approaches

A dataset containing gene expression data across 10 conditions was obtained for analysis. K-means, SOM and Dirichlet clustering were all used to generate four clusters, enough clusters to find meaningful patterns while keeping the required number of comparisons low. Investigating these three approaches addresses Aim 1.1.. The dataset was also split randomly into 4 different groups to demonstrate an outcome when clustering was uninformed. A K-means algorithm has been implemented in the R statistics package. The SOM algorithm is available through the Kohonen package on R. Both approaches require normalization of data prior to clustering. For K-means, the data was normalized by dividing each data point in the row by the row sum. The Kohonen package provides its own normalization function which was used prior to clustering. The Dirichlet algorithm is implemented in the bnkit package developed by myself and other members of the Bodén group.

The gene sets were isolated from each of the four clusters using the annotation assigned to each gene expression profile. The semantic similarity algorithm will accept Entrez gene IDs as input for the analysis so each gene symbol from the data was converted to the appropriate ID using the Python MyGene package. Each pairwise semantic similarity comparison was performed by the GOSemSim package and a matrix was constructed. Each GO term hierarchy (biological process (BP), cellular component (CC) and molecular function (MF)) was explored separately resulting in three matrices comparing the similarity of functional patterns across clusters and techniques.

4.3 Dirichlet clustering

A Dirichlet clustering algorithm had previously been implemented in bnkit by myself and other members of the Bodén group. This algorithm follows the description in [55] and its implementation addresses Aim 1.2.. The Gibbs sampling approach was discussed in the Introductory material.

4.3.1 Algorithm

One concern not addressed by the paper was the occurrence and handling of empty clusters. Empty clusters occur when the available data does not comply with the current distributions of the clusters. The distribution representing the cluster becomes so unfavourable that no data points are assigned to it. This prevents clustering reaching completion. [55] states a random starting point should be used to initialise each distribution however, using a data point from the data itself reduced the number of situations where empty bins occurred. This approach is similar to how K-means clustering first identifies the cluster 'centres'. To consistently handle the occurrence of empty bins, the code was implemented to identify the data point with the lowest probability of belonging to its current cluster and move it into the empty bin. With n empty bins, the n worst data points were used. This eliminated the problem of empty clusters allowing clustering to always run to completion.

4.3.2 ChIP-seq peak strandedness

ChIP-seq peaks are not strand specific as reads from both strands are analysed to identify peak locations. The orientation of each peak is not restricted and symmetry can be expected as a result of strandedness. To account for this symmetry in the Dirichlet algorithm, when assigning a data point to a cluster, the probabilities of both orientations are tested and the orientation found to have the highest probability is assigned to the cluster. This doubles the number of calculations required to place a data point in a cluster but ensures that no bias is introduced by forcing an orientation onto a peak.

4.3.3 Optimal cluster number

Each different data set will have a different optimal number of clusters. To identify this number the minimum description length algorithm proposed by [55] was implemented. This approach is described in the introductory material and relies on the sum of two DL calculations culminating in eq 5 in [55].

4.4 ChIP-seq peak processing

To address Aim 1.3. and to be clustered using the Dirichlet algorithm, each peak must be represented by a histogram. In this case, the counts for each bin of the histogram are based on the read depth around the peak. To generate consistent histograms with equal bin sizes and numbers,

each peak was given a uniform size of 500bps around the summit of the peak as identified in the narrow peaks file. A TF peak is ideally less than 500bp so this window will capture the required read depth information. A constant number of columns is required for the count vectors that make up the Dirichlet distribution. Each uniform window is broken into segments and the read depth is counted for each segment to capture the shape and magnitude of a peak. This requires the narrow peak file to identify peak locations and the bam file to count read depth. It is possible that reads are counted more than once. A smaller segment size (e.g. 10bp) provides more sensitivity to the shape of the peak. It also takes longer to cluster and process. Segments of 20bps in a 500bp window were selected for this investigation however both variables can be changed.

4.5 Genetic location analysis

To address Aim 2.1.1. and 2.1.2.; ChIPSeeker, an R package, was used to annotate ChIP-seq peaks to both a genomic location and the nearest gene. The TSS region was set to 3000bps upstream and 1000bps downstream and annotations were made for all peaks in each cluster. Six locations were used to describe all peaks with each peak having one assigned location: exon, intron, 3'UTR, 5'UTR, promoter and distal intergenic. This reduced the annotations made by ChIPSeeker grouping the multiple exon, intron and promoter annotations into a single annotation group respectively. Each cluster had a set of counts for each location forming a location distribution. A chi-squared test was then performed with the null hypothesis that the distribution of peak locations is independent of the cluster assignment. To determine the contribution of each location category to the variations in location distributions between clusters, individual enrichment tests were performed for each cluster and location combination using Fisher's exact test. The null hypothesis being that the count of binding in a specific location is independent of cluster assignment. Locations identified as enriched by the Fisher's exact test were annotated as over or under depending on whether the count was higher or lower than the expected value respectively. The expected value for each cell in a contingency table was calculated using (row total * column total)/N.

4.6 Epigenetic analysis

This methodology addressed Aim 2.2.1. and Aim 2.2.2.. The Broad ChromHMM track was downloaded from the UCSC genome browser for H1Hesc, GM12878 and HepG2 cell types (hg19). Each peak was assigned a chromatin state based on its location according to this track. For each cluster, the counts of each chromatin state were recorded creating a contingency table with rows for clusters and columns for each chromatin state. A chi-squared test was performed on the full contingency table with the null hypothesis that each chromatin state profile is independent of cluster allocation. Fisher exact tests were also performed for enrichment of each state in each cluster against all other states and clusters. In this case the null hypothesis was that the count of a specific epigenetic annotation is independent of cluster assignment. Annotations were annotated as over or under following the same approach discussed previously.

4.7 MEME analysis

A MEME analysis was performed on each cluster group of each TF according to Aim 2.3.. MEME requires centred and uniform data to perform its analysis so the bed file describing the peaks centred around the summit, calculated prior to clustering, was used for each cluster. Bedtools was used with the hg19 unmasked reference to obtain fasta files for each cluster in each TF. These fasta files were then passed to MEME using the meme-chip algorithm with default parameters. The top DREME motif was selected from each cluster for analysis. DREME limits its motif searches to 8bp.

5 Results

The full summary of results for each TF can be seen in Appendix B.

5.1 Clustering comparison

Each clustering approach was successfully applied to the gene expression dataset containing ten different conditions. The following results fulfil Aim 1.1.. All were able to generate four different clusters on which a semantic similarity analysis could be performed.

When comparing the three clustering approaches using semantic similarity, the K-means approach was able to provide the lowest average similarity scores between clusters across all three GO term categories as shown in Table 1. Dirichlet clustering had the second lowest average similarity scores across all three clusters. In the MF and CC categories, Dirichlet clusters had the lowest pairwise similarity scores across all pairwise tests, 0.79 and 0.86 respectively. The SOM approach and the randomly separated data points had higher similarities. All clustering approaches were able to separate data points more effectively than what we would expect to see by chance. Dirichlet clustering and K-means had the best performance but the benefits of Dirichlet clustering outweigh those of K-means. Dirichlet distributions do not require normalization so no data is lost during the clustering process. The alpha values resulting from Dirichlet clusters provide clear information about the distributions of each cluster or the centres and the evidence or support within each cluster. It is a flexible approach that is not as sensitive to noise as K-means clustering and is therefore the clustering approach that will be used in this investigation.

When comparing the three clustering approaches theoretically, the Dirichlet approach is able to provide the most information in the clustering outcome. This is because the biggest pro of the Dirichlet algorithm is that it takes raw counts rather than normalized data like K-means and SOM. Normalisation is the biggest con of both K-means and SOM. More information is available to the Dirichlet algorithm allowing it to make more informed decisions about how data is clustered.

Table 1: The average and minimum semantic similarity scores for all pairwise comparisons between clusters within each approach. Scores across three GO term categories for K-means, Dirichlet and SOM clustering approaches including a random model for comparison.

	MF		BP		CC	
Approach	Mean	Min.	Mean	Min.	Mean	Min.
Dirichlet	0.85	0.79	0.92	0.85	0.93	0.86
K-means	0.84	0.80	0.90	0.84	0.92	0.88
SOM	0.88	0.88	0.93	0.09	0.95	0.93
Random	0.92	0.91	0.95	0.94	0.97	0.95

5.2 Dirichlet algorithm

Adaptations to the Gibbs sampling algorithm that performs Dirichlet clustering were successfully completed in line with Aim 1.2.. All clusters are populated during clustering removing the issue of empty clusters (see Section 4.3.1). The probabilities of both orientations of a ChIP-seq peak belonging to a cluster are checked to ensure the optimal cluster allocation for every peak (see Section 4.3.2).

The optimal cluster number was successfully identified for each of the ten different TF datasets. The change in DL as cluster number increases can be viewed graphically in Figure 2 using SRF as an example. The DL plot was similar for all TF datasets with a gradual decline to the minima followed by a sharp incline in description length after the optimal cluster number has been reached. After the incline, two trends were apparent, either the DL would continue to gradually increase or it would begin to gradually decrease again. In the cases where a decrease was observed, a second minima would never be reached due to the increasing complexity of a model with such high cluster numbers. The magnitude of the description length value and the number of clusters identified as optimal varied depending on the number of peaks in the dataset and the variation in peak shape across the peaks.

5.3 Peak clusters

Data processing was successfully completed (Aim 1.3.) and the clustering process optimised (Aim 1.2. for every TF dataset. Each TF had a set of clusters that could be visualised using the alpha values of its Dirichlet distribution from the mixture model. For each TF there was little similarity observed between clusters. The shape of each cluster peak was either unique, or distinguished by a varied peak height or a change in the alpha sum value. A variety of different peak shapes were observed with some similarities across TFs. Each TF had one peak with a single summit in the centre of the plot which levelled off evenly at both sides. This shape can be seen in cluster 1 of SRF and cluster 2 of GABP in Figure 3 and is reminiscent of a bell curve. A bimodal peak was also a common feature among the different TFs and an example can be seen in cluster 2 of

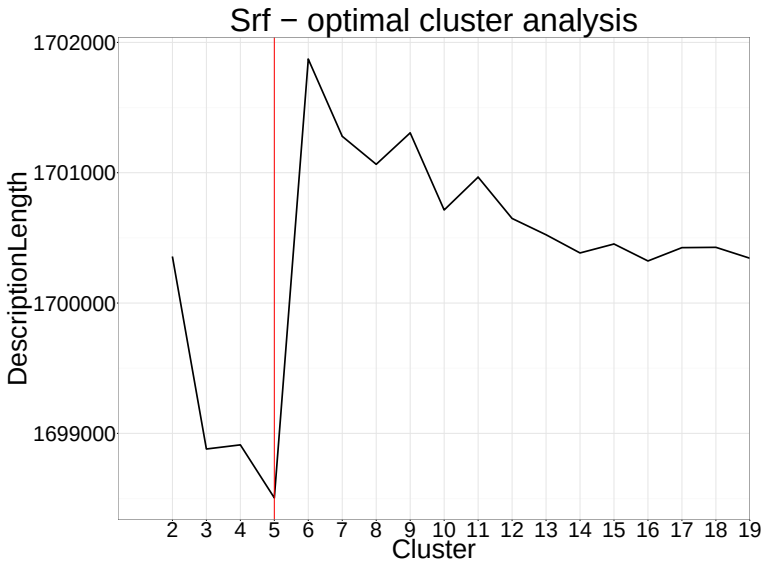


Figure 2: A graphical representation of the MDL principle demonstrating how the sum of DLs changes as cluster number increases. The optimal cluster number is highlighted with a red line where the value is minimised. This model represents the best representation of the data without introducing unnecessary complexity with additional clusters.

SRF and cluster 5 of GABP in Figure 3. Other peak shapes included peaks with relatively flat profiles, peaks with increasing height across the window and peaks with summits not centred in the window. In each TF, a variety of these shapes could be observed, however clusters with similar peak shapes also occurred. Where a similar peak shape occurred, for example two centred bell like shapes, they were always distinguished by height or alpha sum or both. One peak would be higher than the other indicating increased read depth across the peaks belonging to that cluster. The alpha sum value for a cluster was often varied between two clusters with a similar shape indicating that members of the cluster with a lower alpha sum had less evidence or support even though the same shape was observed. This is most evident in clusters 4 and 5 of MAXH1 in Figure 9.

Although similar shapes were observed across all ten TF datasets, the magnitudes or heights of these peaks varied and no identical examples of peaks could be seen between any TFs. This provided the first indications that TFs behave too differently to identify consistent patterns between different experiments.

Not only do different TFs demonstrate different peak shapes but the same TF in a different cell type also behaves differently. MAX and RAD21 both had data from H1Hesc and Gm12878 cells analysed. All four sets of results can be found in Appendix B. Neither TF shared the same dataset size or optimised cluster number. For MAX, there were no clear similarities between any peak shapes in the two sets of clusters. For RAD21, some similarities in cluster shape were observed between clusters 0 and 6 and clusters 2 and 5 from Gm12878 and H1Hesc respectively. Although the shapes shared similarities, the heights of the peaks varied and the H1Hesc result contained nine clusters compared to four. This suggests that cell type and the accompanying biological factors

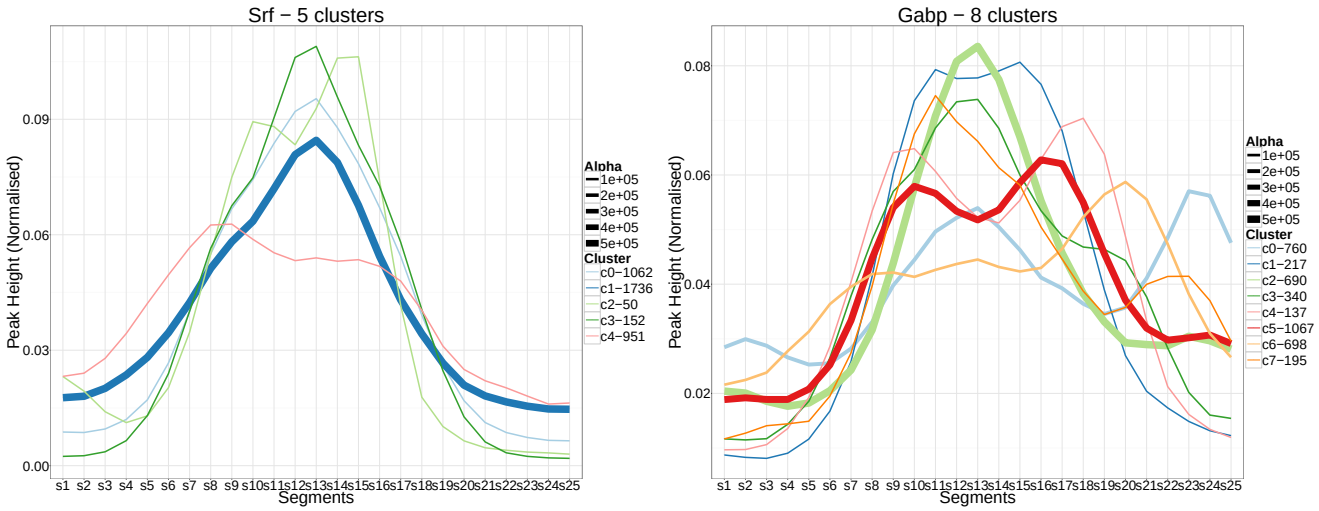


Figure 3: A visualisation of the SRF and GABP clustering results after identification of the optimal cluster number. Note that the axes are not equal. Each has a unique set of peak shapes represented by each cluster. The alpha values for each peak were normalised to show the shape. The changes in alpha sum are also represented by the thickness of the line identifying clusters which are based on more evidence. For each cluster, the number of peaks in the cluster has been reported in the legend.

from different cells affect the peak shapes generated by ChIP-seq.

5.4 Location analysis

For each TF, each individual cluster group demonstrated a set of locations that differed from all other clusters. This can be visualised in Figure 4 where the location profiles of GABP and SRF are shown. A clear visual relationship between peak shape and binding location was observed within each TF providing a result for Aim 2.1.1.. A Chi-squared test was able to demonstrate this in some TFs, for example SRF with a p-value of 8.39e-06. However, where the changes were not as severe, the significance was lost in the high degrees of freedom required for the test. Where the variations of each location distribution were visually obvious but not significant according to the Chi-squared test, the individual analyses of each location within clusters provided a clearer description of which locations were varied in which clusters. This secondary set of location results follows Aim 2.1.2. and is referred to as individual locations.

5.4.1 Location profiles

For almost every TF, one cluster had a profile of locations that was highly similar to the distribution we would expect to see according to the full set of peaks. In Figure 4 this can be seen in cluster 2 and 5 of GABP and cluster 1 of SRF when compared to the column showing the expected distribution. After identifying all clusters that best represented the expected distribution, their peak shapes were analysed to identify any patterns. Four TFs showed a bell shaped peak and four showed a flatter and lower peak with no identical matches in either set. GABP had two cluster

groups that were highly similar to the expected distribution - one with a bell like shape and one with a clear bimodality. MAX in the Gm12878 cell type had no cluster group that represented the expected distribution well. No distinct pattern was evident linking the cluster showing the expected distribution of locations and peak shape.

Location profiles demonstrating the most unique distributions or lacking peaks in specific locations (e.g. no binding in 3'UTR) were also compared to peak shape to determine patterns that could be observed across multiple TFs. Again, a variety of peak shapes were associated with highly variable location profiles.

When comparing MAX and RAD21 results between the two cell types, RAD21 again shared more similarities than MAX. For MAX, the Gm12878 dataset had a higher percentage of promoter binding while the H1hesc dataset showed a higher percentage of distal intergenic binding. The clusters previously identified to have similar peak shapes in RAD21 did not share similar binding location profiles indicating that comparing between TF results is not the best use of this modelling approach. Instead, it provides added layers of information to individual experimental datasets.

5.4.2 Individual locations

Analysing individual locations identifies more specific patterns within each cluster for a TF. It also allows identification of significantly over or under represented locations. Figure 5 shows the individual tests for the each TF. In the SRF results, each group has a distinct pattern of enriched locations. Cluster 2, that best representing the expected location profile, has no significantly enriched locations. The other four clusters show over or under enriched intronic and/or promoter regions. In GABP, we saw two clusters following the profile of the expected distribution, the individual tests have not clarified this outcome as both clusters show no enriched locations. Clusters 3 and 4 also demonstrate the same pattern of individually enriched locations however the location profiles appear to differ.

Cluster 0 in MAXH1 and cluster 6 in SP1 both have promoter under represented and intron and distal intergenic over represented however the peak shapes associated with either cluster do not share any similarity in shape. Very few clusters shared the same significantly over and under enriched peaks.

RAD21 and MAX had their individually enriched locations compared and neither TF shared similarities between the two cell types. In RAD21, the H1Hesc cell type data had clusters that generally had distal intergenic regions over enriched and promoter regions under enriched. In the Gm12878 cell type data the clusters had a mix of over and under enriched results for the same two locations.

Within a TF, each location profile is dependent on its cluster assignment providing a result for Aim 2.1.1.. The individual locations also vary depending on cluster assignment providing a result for Aim 2.1.2.. Within a TF, peak shape is linked to binding locations. A global pattern linking peak shape to an expected location profile or set of enriched locations could not be found providing a negative result for Aim 2.1.3..

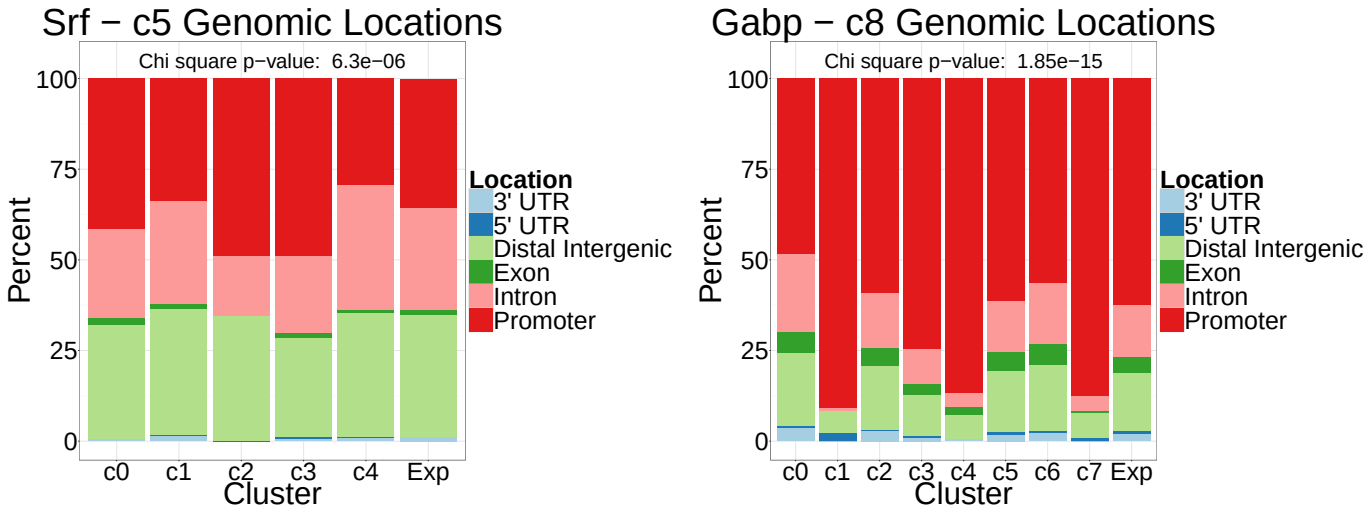


Figure 4: A visualisation of the genomic location profiles for the binding sites represented by the peaks in each cluster from SRF and GABP. The profiles demonstrate that each cluster has a different set of genomic locations and the Chi-square test results both show that these profiles are dependent on cluster assignment. The column to the left represents the profile that would be expected across all peaks.

5.5 Epigenetic analysis

A similar pattern of results to the location analysis was observed for the epigenetic analysis. Within a TF, each individual cluster demonstrated a profile of chromatin states that differed from all others providing a result for Aim 2.2.1.1.. This relationship can be observed in Figure 6 where profiles for GABP and SRF are shown. Visually, it is apparent that there is a relationship between cluster assignment and epigenetic profile but the Chi-squared test is not effective in demonstrating this due to such high degrees of freedom with so many chromatin states assessed (15 states). Following Aim 2.2.1.2., the enrichment of individual chromatin annotation within clusters provided a secondary result that was combined with the full profile analysis.

5.5.1 Epigenetic profile

In the location profiles, one cluster tended to have a distribution similar to the expected distribution for every TF. Looking at the same cluster and comparing the epigenetic profile to the expected profile, fewer similarities were observed. In Figure 6 for SRF, cluster 1 remained the most similar to the expected profile and this pattern was observed in five other clusters. In GABP, cluster 5 shows more similarity to the expected distribution than cluster 2 providing a means of differentiating the two clusters. Cluster 2 is lacking peaks annotated as Repetitive/CNV indicating a link between the two unique peak shapes (bell shaped and bimodal) and epigenetic state. Although the two clusters had similar location profiles, the differences in the epigenetic profiles provide an explanation for the variations in peak shape. This is an example of information provided by this modelling approach that can clarify TF interactions and binding modes. The remaining five TFs did not contain

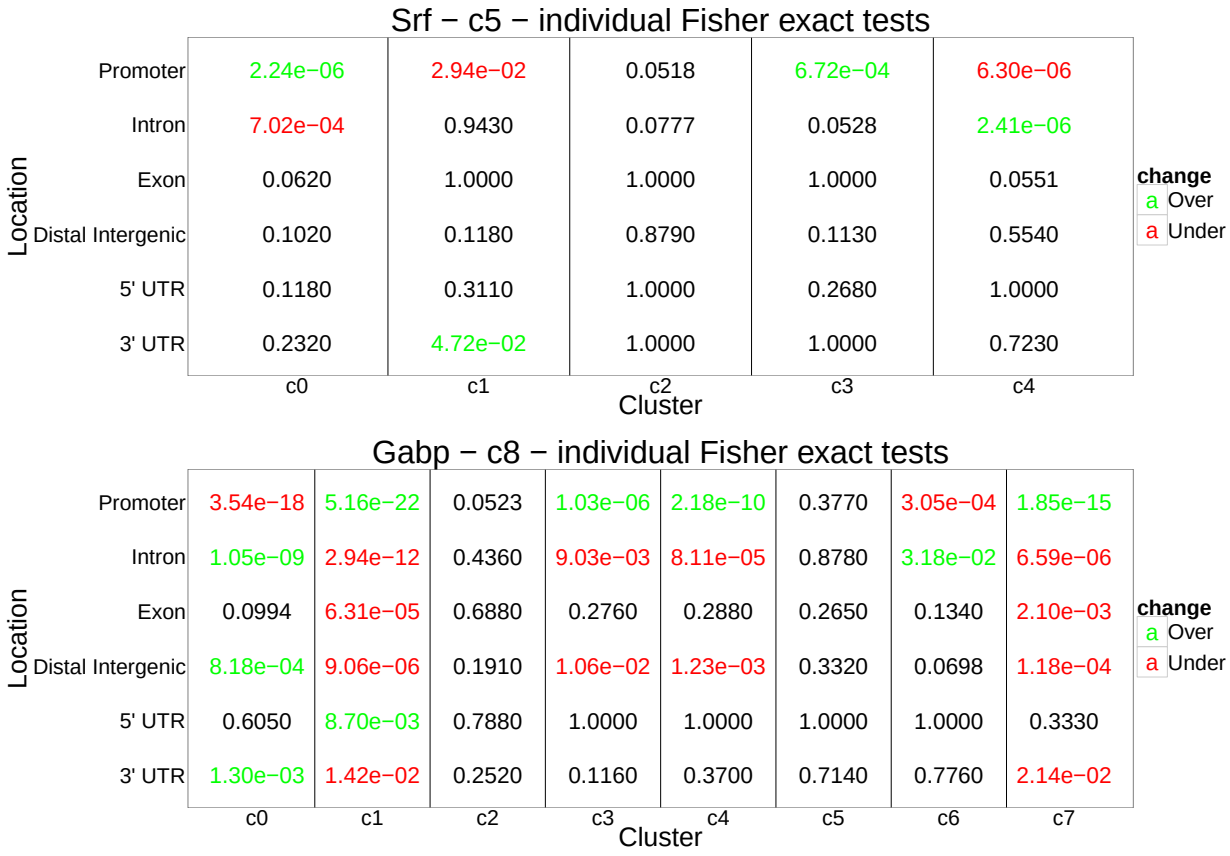


Figure 5: A visualisation of the individual Fisher exact test results for the SRF and GABP clusters. The individual tests show which specific locations are linked to the dependency between cluster assignment and genomic location. They further demonstrate the differences between each cluster. Significant results are green if the count was higher than the expected value and red if the count was lower than the expected value.

clusters with profiles that looked similar to the expected profile.

With a large number of annotations, the variations observed within TFs and between TFs were difficult to observe and compare. No global patterns based on the full distribution of chromatin annotations were identified providing a negative result for Aim 2.2.3..

Comparing the two MAX datasets, in the Gm12878 cell type, all clusters show high percentages of heterochromatin, transcription elongation and weak transcription annotations. In contrast, the H1Hesc cell type shows lower percentages of a larger number of annotations. The exception being cluster 6 where the repetitive annotations show a striking majority. Comparing the two RAD21 datasets returned similar results to MAX in the Gm12878 cell type showing high percentages of the same annotations . The RAD21 H1Hesc dataset had a higher percentage of insulator and poised promoter annotations. Cell type plays a significant role in the epigenetic profile of the same TFs.

In the MAX H1Hesc results, cluster 6 showed a majority of repetitive annotations which was not observed in any other cluster. The peak shape of this cluster was sharp and higher than any other peak. In SRF, cluster 2 had a noticeably higher percentage of weak enhancer and repetitive

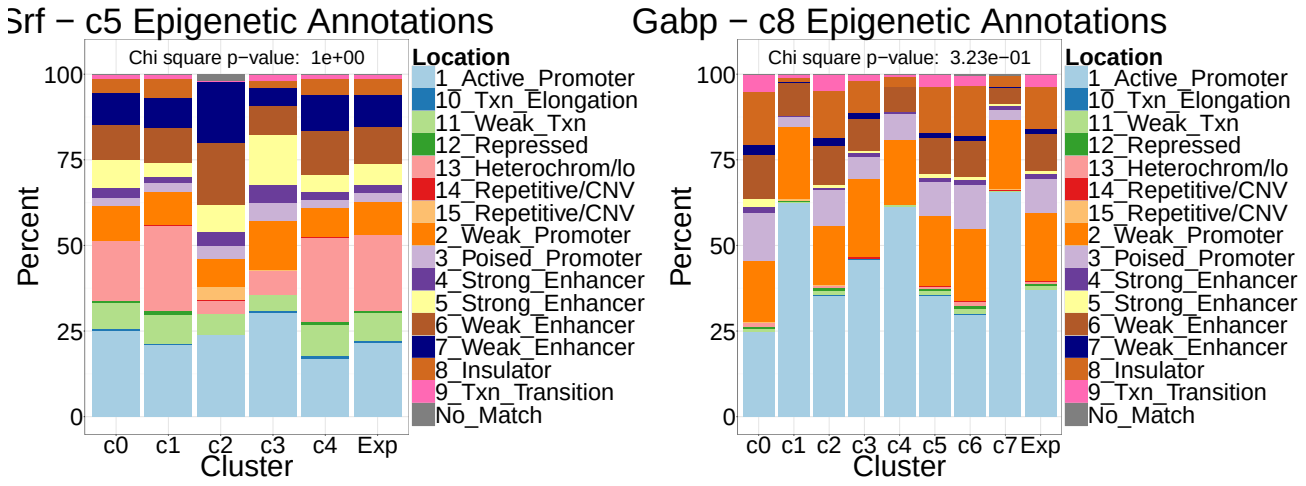


Figure 6: A visualisation of the epigenetic annotation profiles of SRF and GABP. The profiles demonstrate that each cluster is defined by a different epigenetic environment. The Chi-square test was not successful in demonstrating the dependence of epigenetic annotation to cluster due to the high degrees of freedom. The visualisation is, however, striking. The column to the left represents the profile that would be expected across all peaks.

annotations while also showing the lowest percentage of heterochromatin. The peak shape of cluster 2 was bimodal while all others showed a single peak or a wide shape. In both cases, the epigenetic profile identified features unique to a cluster that could quickly be linked to peak shape. Although both TFs saw an overrepresentation of repetitive annotation, the two associated peak shapes were not similar. Within each TF, the peak shape was distinct with a clear biological outcome in chromatin state providing a result for Aim 2.

5.5.2 Individual annotations

Differences in chromatin annotations were observed when looking at the epigenetic profiles. Following Aim 2.2.1.2., individual annotations allowed analysis of which annotations were over or under represented. Each cluster group for each TF showed differences in which annotations were over or under represented. This is evident in Figure 7 where the individual results for SRF and GABP are reported. In both TFs, each cluster has a unique set of enriched annotations. Where the location could not completely separate the clusters, the epigenetic results are able to identify differences. Clustering peaks based on shape leads to significant variations in chromatin annotations which are linked to TF binding and function.

Cluster 2 in SRF, 0 in RXRA and 2 in TBP all showed over enriched repetitive annotations and a bimodal peak shape. Cluster 4 and 5 in GABP and 1 and 6 in SP1 also have bimodal shapes but are not enriched in repetitive annotations. No pattern could be found that held true across all TFs indicating Aim 2.2.3. has a negative outcome.

In section 5.4.2, cluster 0 in MAXH1 and 6 in SP1 were identified as having the same individually enriched locations but no peak shape similarity. Combining the individual enrichment

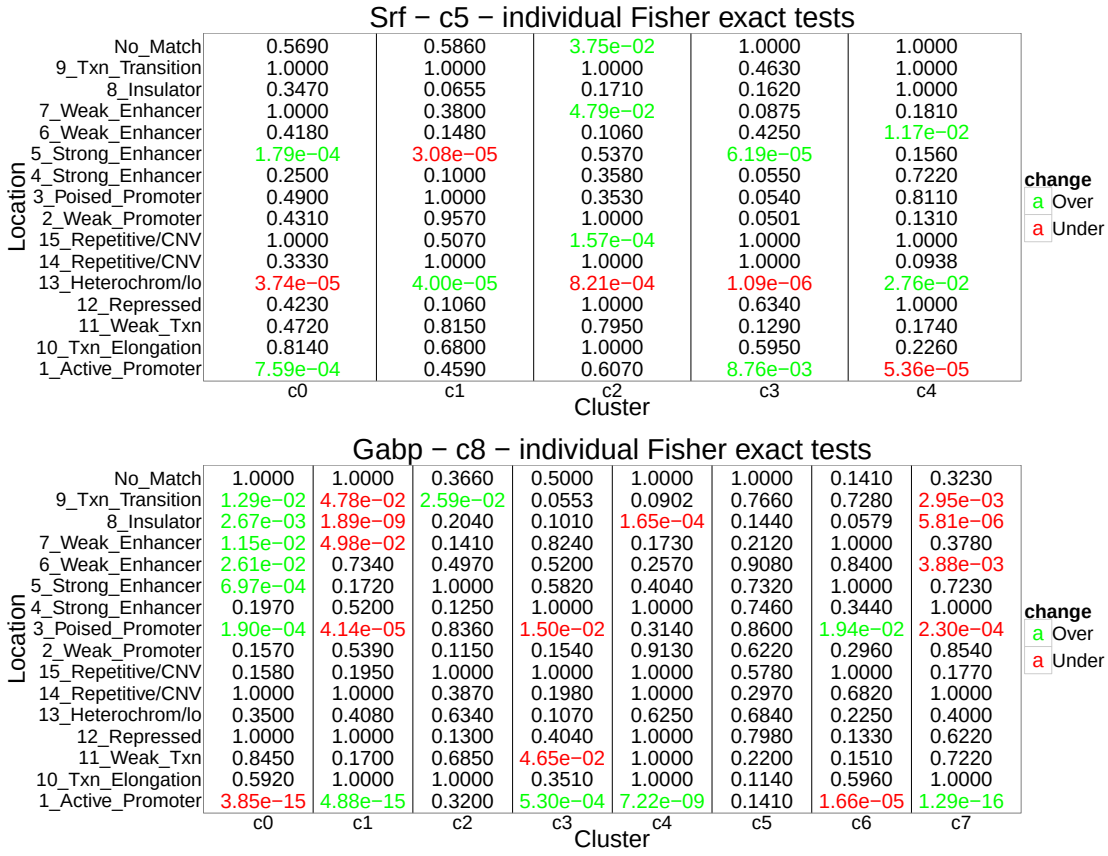


Figure 7: A visualisation of the significance values for each individual Fisher exact test on the annotations. The individual tests show that enrichment of specific annotations is dependent on cluster assignment. Each cluster shows a distinct set of enriched annotations for both TFs. Significant results are green if the count was higher than the expected value and red if the count was lower than the expected value.

results for chromatin annotations, we see that cluster 0 in MAXH1 is over enriched for active promoter and strong enhancer while cluster 6 in SP1 has no enrichment of active promoter and is under enriched for strong enhancer. The only similarity between the two cluster groups is an over enrichment of weak promoter. This is an example of the intricate layers of information that affect the outcome of TF binding.

The disparity between the two cell types for RAD21 and MAX continued into the individual annotation analysis. The H1Hesc cell type showed a number of enriched annotations across all clusters while the Gm12878 cell type had fewer enriched annotations for both TFs.

For Aim 2.2.1.1. the epigenetic profiles within each TF were dependent on cluster assignment. This dependency is further observed in the individual enrichment tests of each annotation indicating which specific annotations are linked to each cluster assignment. Epigenetic analysis also provided clarification where a relationship between peak shape and location could not be found. No global patterns were identified linking chromatin annotations to peak shape or location.

5.6 MEME analysis

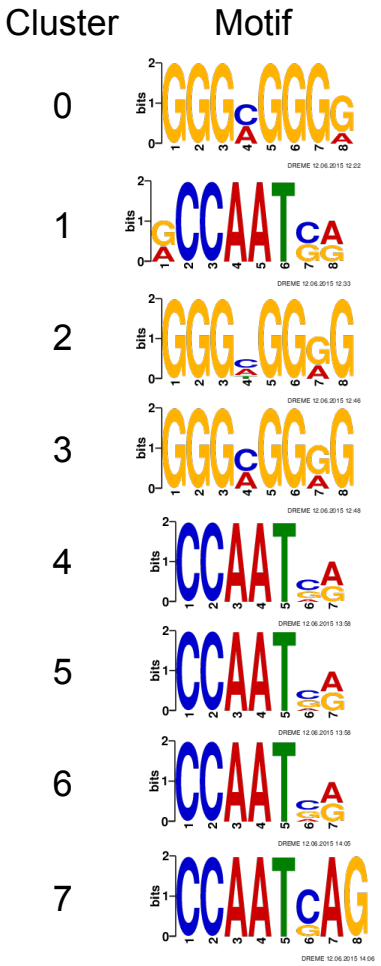


Figure 8: The motifs generated by MEME for each cluster in the SP1 dataset. Clusters 0, 2 and 3 show the SP1 motif while the others show NFY as the most enriched motif in the DREME results.

Variations in sequence were captured by identifying a motif for every cluster addressing Aim 2.3.. Every TF experiment contained at least one cluster which returned the expected motif for the TF. The only exception being TBP which has a complex motif that is not well resolved. How sequence is related to peak shape is not clear from comparing motifs. Consensus motifs lack specific detail about the variety of enriched sequences. The observed changes in motifs within a TF indicate that different clusters are enriched for different sequences. At least one cluster in every TF had a motif with a distinct difference. In SRF, the motif for cluster 0 was longer than any other motif. As seen in Figure 10, clusters 0-4 contained the MAX motif with cluster 2 the only one to represent it as a 6bp motif. Cluster 5 and 6 did not show the MAX motif.

The remaining clusters in each TF showed position specific changes within the motif. In SP1 in Figure 8, cluster 1 and 4 share the same 5bp core of their motifs: CCAAT. In cluster 1, this core is preceded by an enriched [GA] then followed by an enriched [CG]. In cluster 4, the core is not preceded by anything and is followed by an enriched [CGA]. It is minor changes to the motif, like identification of a third nucleotide present in the sequence at a specific position, that can indicate variation in sequence that will influence binding. Similar changes can be seen in Figure 10 for the MAXH1 results in clusters 0-4. Small changes like this were present in all TF results including comparisons between the RAD21 and MAX sets of results.

SP1 showed two distinct motifs between all 8 clusters. In cluster 0, 2 and 3 the SP1 motif is the top hit while the other clusters contained the NFY motif as the top hit. This indicates an interaction between the two TFs. Cluster 0 and 2 also share the same enriched locations and similar enriched epigenetic patterns while cluster 3 is different. Cluster 3 shows over enrichment in transcription elongation and active promoter while cluster 0 and 2 show under enrichment as well as other variations. The SpaMo results indicated that in cluster 0 and 2, the SP1 and NFY motifs were spaced significantly close to one another. This result was not observed in cluster 3. The clusters with NFY as the top hit were more likely to be spaced near another NFY site than an SP1 site.

Cluster 6 of MAXH1 had a high proportion of binding in repetitive regions. The motif associated with this cluster is unique from all other results and is not recognized in a TOMTOM search.

Repetitive sequence has a confounding effect on motif searches when using an unmasked fasta file.

Using a motif to represent changes in sequence is not the most effective approach as consensus motifs, such as those provided by MEME, are known to mask the detail represented within the sequence. A means of exploring the sets of significantly enriched sequences within a fasta file is under development within the Bodén group but was not used in this analysis.

5.7 The impact of alpha sum

The MAXH1 results summarised in Figure 9 contained an example of two clusters sharing similar shape but different alpha sum. Comparing the location profiles and individual location results of cluster 4 and 5, clear differences can be observed. Cluster 5 is over enriched for promoter binding and cluster 4 is under enriched for promoter binding. Cluster 4 is also more likely to have binding in intronic and distal intergenic regions.

The same can be seen in the epigenetic set of results with cluster 4 over enriched for weak enhancer and transcription elongation and under enriched for poised and weak promoter, repetitive annotations and active promoter. In contrast, cluster 5 was over enriched for active and weak promoter, and strong enhancer annotations.

The motifs for cluster 4 and 5 also vary significantly. Cluster 4, with more evidence, contains the expected MAX motif as the top result. Cluster 5, contains a motif with no relation to MAX and an E-value that is only just significant. The peaks identified by cluster 5 appear to be noise or unrelated to MAX binding.

6 Discussion

TFs are regularly studied using ChIP-seq experiments but no one has previously explored the biological significance of ChIP-seq peak shapes. We implemented and applied a Dirichlet model to cluster ChIP-seq peaks as a novel approach to exploring peak shapes. Using this model, we made a number of specific observations providing new levels of detail about TF binding. Specifically, we successfully clustered ChIP-seq peaks based on their shape, density and magnitude then demonstrated how each cluster contains unique, biologically relevant, features. We were also able to explore these features in depth using statistical tests.

We showed that Dirichlet clustering is a novel modelling approach that performs comparably to K-means clustering and outperforms SOM when applied to gene expression data (Aim 1.1.). Our implementation has improved sensitivity through its ability to analyse both orientations of a ChIP-seq peak to account for the strandedness and potential symmetry of ChIP-seq results (Aim 1.2.). Our approach results in evidence based clusters providing a new way to identify and interpret noise in the data. Existing clustering approaches are sensitive to noise. A means of identifying optimal cluster number is also a benefit to any modelling approach related to clustering.

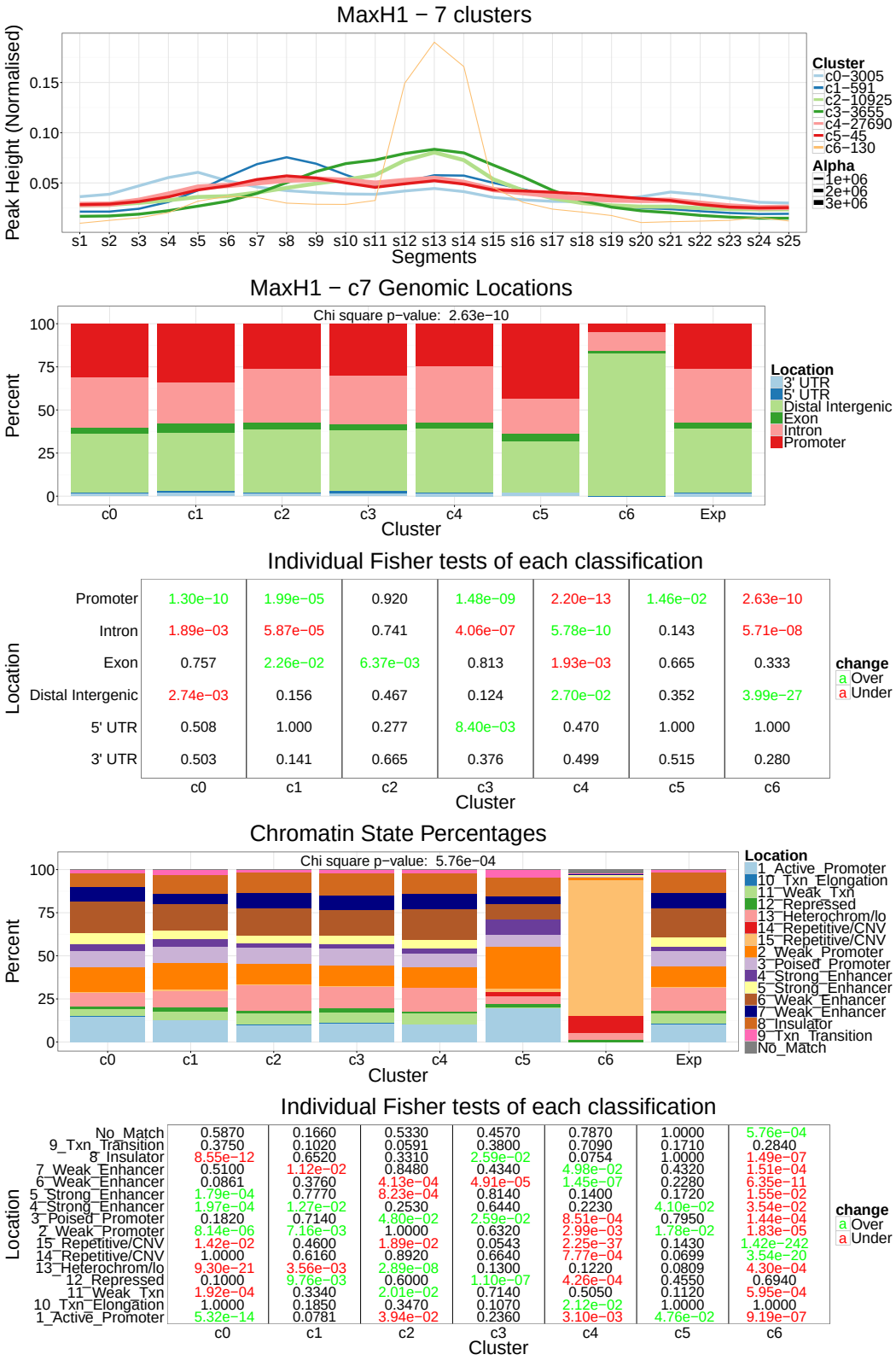


Figure 9: A summary containing the MAX peak shapes, genomic location analyses and epigenetic analyses in the H1Hesc cell type. This figure helps illustrate how all the tests relate to one another to help identify the different binding modes of MAX. For example, cluster 6 has a distinguished peak, a high percentage of distal intergenic binding and is over enriched for distal intergenic binding, a high percentage of repetitive annotations and over enrichment for repetitive annotations. The peaks in this cluster share a unique set of biological features.

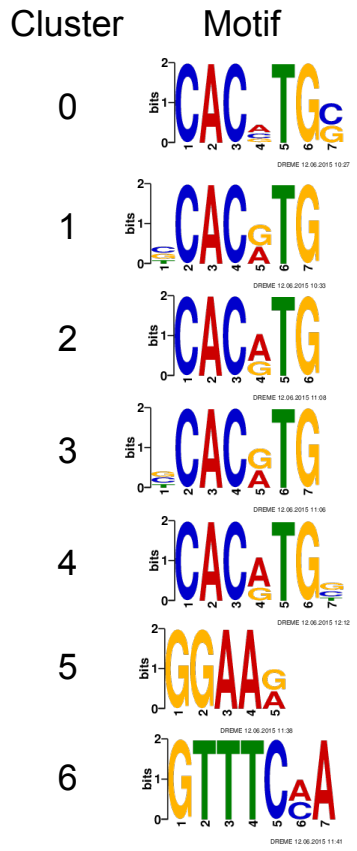


Figure 10: The motifs generated by MEME for each cluster in the MAX dataset in the H1Hesc cell type. Clusters 0-4 show the MAX motif while clusters 5 and 6 show very distinct results. The motif from cluster 5 is common and general and has a large number of significant hits when a match was searched for using TOMTOM (MEME). Cluster 6's motif had three significant hits, none from TFs known to interact with MAX.

Our model shows that a variety of peak shapes exist in each dataset that can be successfully clustered (Aim 1.4.). Similarities in peak shapes were observed between TFs but we noticed that although these similarities exist, there is too much variation for these shapes to be comparable across TFs. Biologically, where height was the differentiating feature between two cluster shapes, the variations in read depth can be linked to binding affinity with a higher peak indicating higher affinity [27]. Our model can therefore separate binding events based on affinity as well as peak shape. Identifying which binding events occur with higher or lower affinity will allow researchers to analyse them in isolation resulting in new theories about TF binding modes.

We also saw examples of peaks that were differentiated by alpha sum demonstrating our model's ability to use the support or evidence of a peak to further separate them. We observed in MAXH1 that cluster 4 and 5 were differentiated by alpha sum leading to variations in genomic location, epigenetic annotations and motif. We identified the peaks in cluster 5 as noise and irrelevant to the binding outcomes of MAX in H1Hesc. In previous experiments, these two sets of peaks would have been treated equally however our model has shown that they represent different events and should be analysed in isolation to determine their true link to binding.

In our analysis, we show that genomic location is dependent on peak shape (Aim 2.1.). We did not observe any global patterns linking peak shape and genomic location across different TFs. Where identical patterns of individual location enrichment were observed between TF clusters, the peaks shapes were not consistent. Instead, we observed that each cluster group has a different set of locations where a TF is more likely to bind. Our model allows analysis of the different location patterns unique to each TF based on clustering. It is now possible to understand where different binding events, defined by our clusters, occur; painting a clearer picture of a TFs interactions with DNA.

Our model is also able to show that the epigenetic environment around the peaks is dependent on cluster assignment or peak shape (Aim 2.2.). Similar to the location analysis, no global patterns were observed linking peak shape and epigenetic environment. We saw that where similarities existed between location profiles of different clusters but not the peak shape, the epigenetic profile

of the clusters was different providing evidence that the epigenetic environment influences peak shape more than the general location of binding. A researcher investigating a TF can use this epigenetic data to describe the different epigenetic states surrounding subsets of peaks and explore the effect state has on outcomes such as activation or repression of target genes.

Clusters identified by our model demonstrate variations in enriched motifs (Aim 2.3.. We observed both minor variations in motif and changes that indicate enriched binding of different motifs. We showed that motif variations could be linked back to epigenetic profiles to help explain variations. We could not link motif changes to specific peak shapes. Small variations in the motif indicate presence of different nucleotides at specific positions in the sequence. Single nucleotide changes influence the binding outcomes of a TF. Our model is able to separate peaks into clusters that demonstrate variation in sequence that is significant enough to alter the motif. Consensus motifs are known to mask features of the sequence and a means of exploring sequence enrichment without using a motif is required to further explore this relationship between peak shape and sequence content.

It has been shown that TF binding is complex and relies on many features. Our model supports that claim by identifying no global patterns between TFs and showing how no single biological feature (e.g. genomic location) can predict the peak shape or vice versa. We also observed that the same TF in a different cell type will exhibit remarkable differences in all aspects reported by the model. What our model does achieve is a means of isolating binding events based on peak shape and demonstrating the biologically significant differences between the binding events. It allows exploration of subsets of peaks that have been shown to behave differently. By providing this detailed information, TFs can now be explored in more depth based on existing experiments using our new approach.

6.1 Future directions

The current model is limited in its exploration of biological features potentially linked to peak shape. The scope of this project can be extended to explore many more features and provide more refined information about TFs. The concept of clustering by peak shape has more applications than what has been explored here. Clustering could be applied to TF families to break down each set of peaks, identify the biological features of each cluster then compare between family members to identify similarities or differences. TFs known to bind as dimers could have their peaks clustered and the results combined with *in vitro* experiments to explore sequence differences and determine whether peak shape is linked to binding of different dimers. TFs which are known to bind cooperatively could be explored to identify a link between peak shape and which cooperative interaction is occurring.

The model itself could also be refined by exploring the window size used around the peak summit and the segment size used to bin read depth counts. Expanding the window will allow more unique shapes to be identified by exploring further to either side of the summit. This could

also help identify other TFs binding close to the TF of interest. Decreasing the segment size could improve sensitivity and increase the variety of peak shapes identified. The data supplied to the model could also be varied. For example, the threshold for significant peaks could be dropped allowing more peaks to be analysed. The clustering approach can then help identify noisy peaks as noisy peaks will contain less evidence and be separated into their own cluster groups. This would reduce false positive results in the final set of peaks. There is a possibility that low affinity peaks with better evidence would then be identified as true binding events improving the false negative rate of peak calling.

Sequence content is key to TF binding due to its sequence specific nature. Incorporating sequence into the model would allow the shape and sequence to be modelled concurrently resulting in clusters that are more specific to TF binding events.

The potential applications and future directions of this model are significant and clearly indicate the importance of what we have developed. Our model has the potential to refine the way we study TFs and generally improve our knowledge of regulation in organisms.

7 Conclusion

TF binding is a complex process that is currently studied using techniques that do not provide all the information required to completely understand the process. We created a model that clusters ChIP-seq peaks based on shape, density and magnitude then successfully links each cluster to its unique biological features. The model has not yet reached its full potential but already the potential applications are widespread. New modelling approaches, such as this one, are essential for unlocking the complex aspects of regulation in the genome.

Appendices

A ChIP-seq

A.1 Experimental approach

In ChIP-seq experiments, cells are first treated with formaldehyde to crosslink DNA-associated proteins to the DNA. DNA is then fragmented using sonication and protein bound fragments are targeted by a specific antibody. Immunoprecipitation (IP) collects fragments that are bound by the antibody. The crosslinks between the protein and DNA are then reversed and the fragments create a library that is analysed using high-throughput sequencing. The protein of interest is referred to as 'ChIPed' after the immunoprecipitation step has occurred. Sequencing requires ligation of oligonucleotide linkers or adapters to both ends of the fragments followed by next-generation sequencing (NGS) [2, 35, 32, 43, 21, 34, 18, 52]. For the IP to be successful, a highly specific antibody against the DNA-binding protein of interest is required. This requires prior knowledge of the existence of a DNA-binding protein or histone modification. If the antibody is not specific enough, the resulting data will be noisy with non-specific proteins being pulled down with the true binding events [15, 32]. ChIP-seq experiments also require a very large number of cells. This limits the types of cells on which ChIP experiments can be performed. The large cell requirement also has the effect of masking interactions between the protein and target regions which are only present in a small number of cells.

The application of NGS allows higher resolution, lower noise and higher genomic coverage when compared to the ChIP-chip assay. ChIP-chip uses microarray hybridization after amplification of fragments and is an older, but still frequently used technique [18].

A.2 Data processing and quality control

ChIP-seq experiments produce sequencing tags from the ChIP library across the whole genome [10]. Library complexity must be sufficient with low-complexity libraries, those containing a large number of redundant reads, indicating a failed experiment where insufficient DNA was collected. The failure could be due to antibody quality, amount of cell material, over-amplification of PCR or over-cross-linking. In low-complexity libraries the same PCR-amplified products are sequenced repeatedly leading to many small peaks detected causing a high false-positive rate. Removal of redundant reads is one approach to correct low-complexity [2, 15].

Often, single end 25-35bp reads are used in ChIP-seq studies however paired end reads are also used and can resolve biases particularly related to repetitive sequence [8]. To effectively identify protein binding locations, the sequencing experiment must provide sufficient coverage by sequencing reads (sequencing depth). The sequencing depth required to effectively analyse ChIP-seq data depends on the size of the genome and the number and size of binding sites of the protein

[2, 15]. For example, a sequencing depth of 20 million reads should be adequate for a mammalian TF with thousands of narrow binding sites. Histones are proteins with broader binding sites and require more depth, between 40 and 60 million reads, for effective ChIP-seq analysis [2]. Increased sequencing depth can allow detection of sites with lower levels of enrichment. After sequencing is complete, target regions or binding sites are identified by the aggregation of reads or tags at specific locations in the genome. To identify these regions, a number of processing steps with associated quality control measures must be undertaken. There are two key steps in ChIP-seq analysis; read mapping and peak calling. Prior to read mapping, the quality of reads produced by sequencing must be explored. This quality control is to identify possible sequencing errors or biases that could have negative effects on read mapping and all consequent analyses. Sequencing technologies, such as Illumina, are capable of filtering poor quality reads as part of the sequencing process. FastQC is a tool that analyses a number of features to determine the quality of a set of reads. If there are features in the FastQC report that are not meeting quality thresholds, steps should be taken to trim or filter the problem reads [2, 1].

Once the quality of reads has been validated, they can be mapped back to the reference genome. Mapping tools available including Bowtie, BWA, SOAP or MAQ. Each has their own strengths and weaknesses particularly related to speed and memory requirements. Most of the short-read mapping tools are effective for mapping ChIP-seq data with limited differences in results. Bowtie2 is a popular tool which has been validated for mapping short read sequences and used in ChIP-seq experiments previously [33, 31]. It will be used for mapping in this experiment. Mapped reads should only have 2-3 mismatches however this can vary depending on the accuracy of the sequencing technology in use [40, 35]. Ideally above 70% of the reads will be uniquely mapped. Less than 50% uniquely mapped reads can indicate problems with the experimental approach including inadequate read length or problems with the sequencing platform. There exist ChIPed proteins that will always have an unavoidably low percentage of uniquely mapped reads, for example if the protein binds frequently in repetitive DNA. In this situation, paired end reads have been shown to improve mapping across repetitive regions of DNA. Prior to peak calling, mapped reads must be assessed using quality metrics such as strand cross-correlation analysis (SCCA) or IP enrichment estimation. This quality control step will detect experimental failures such as insufficient sequencing depth or insufficient enrichment by immunoprecipitation. Duplicate sequences need to be removed as well as non-unique mapped reads.

SCCA measures the degree of immunoprecipitated fragment clustering to assess data quality. The main principle behind ChIP-seq is that high-quality target regions will show clustering of sequence tags. SCCA uses this principle and the fact that enriched tags on the forward and reversed strands are separated by a distance from the binding site centre dependent on the fragment size distribution. ChIP DNA fragments are sequenced from the 5' end and when aligned to the genome will result in two peaks, one on each strand. The two peaks will flank the binding location of the ChIPed protein or modification. Cross-correlation of the tags will provide the optimal detection of target regions as well as provide data for quality control [40]. The tags on one strand are shifted by

k to overlap with the opposite strand and the Pearson correlation of the two read density profiles is taken. The cross-correlation score peaks at the shift corresponding to the fragment length and the shift corresponding to the read length. These two peaks indicate the ideal combination of reads for peak calling. From the cross-correlation results, normalized strand cross-correlation coefficient (NSC) is measured as the ratio between cross-correlation at the fragment length and the background cross-correlation. The ratio between cross-correlation at the fragment length the read length is known as the relative strand cross-correlation coefficient (RSC). Together, they reflect the signal-to-noise ratio in ChIP-seq data. A successful ChIP experiment would generally have NSC ≤ 1.05 and RSC ≤ 0.8 [2, 32].

To predict the target regions for the ChIPed protein or modification in the genome, regions with significant numbers of mapped reads or peaks must be found. A number of peak calling algorithms exist to make these predictions. In most situations, accurate peak predictions rely on the use of a control sample as a background. There are a number of ways a control sample can be generated. The most popular is to perform the ChIP experiment on a second biological replicate but to skip the IP step. This way, the reads generated will represent the expected background levels of DNA in the sample. A negative control can also be generated by repeating the ChIP experiment on a sample with the DNA-binding protein or modification knocked out.

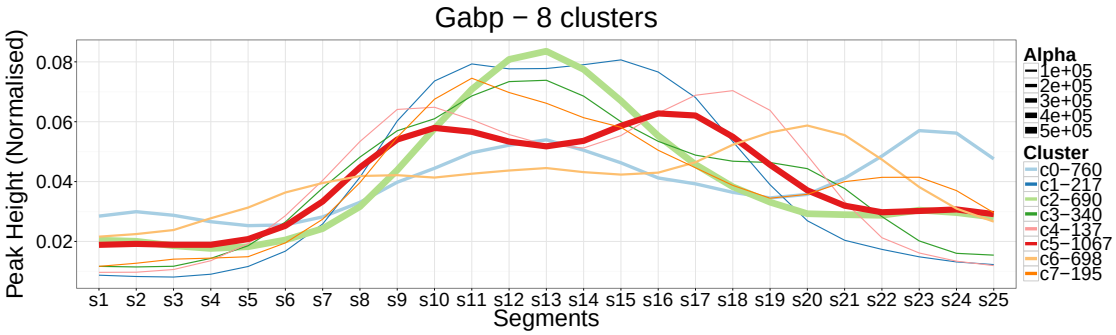
A.3 Peak callers

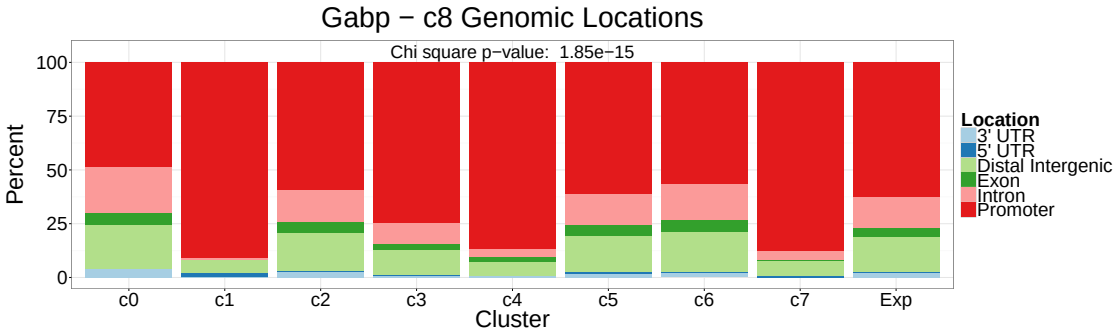
Peak callers search for regions in which tags are enriched in the target sample and not the control sample. The key differences between the different peak callers is the statistical method applied to identify enriched regions and the processing of read counts prior to enrichment analysis. Peak calling programs all follow the same basic approach to calling peaks. First, a profile is created from the sample and control data (if available) in which the tag data is smoothed. For example, CisGenome and spp use a sliding window of fixed width and replace each strand specific site with the tag count summed over the entire window, centred at the site [23, 28]. Tag aggregation and kernel density estimation are other examples of approaches to profiling the tag data. Next, a statistical model is used to identify enriched regions. It is possible to do this without background or control data however it is not as accurate. The simplest models analyse height based on the number of reads such as CisGenome. Model-based analysis of ChIP-seq (MACS) and spp both use variations of a Poisson model [11, 28]. More complex statistical models can be also be used. For example, BayesPeak is a peak caller which uses a fully Bayesian hidden Markov model to detect enriched locations [48]. Based on the statistical model, peaks are ranked by number of reads, a p-value or a q-value and a cut off is applied to select the best or most likely results [41, 40, 2]. Due to the diversity of DNA-binding proteins, cell types, conditions, modifications and factors being assayed; common guidelines for designing, performing and processing ChIP-seq experiments and their resulting data will not be appropriate for all situations [15]. A challenge to peak calling is different DNA-binding proteins and their associated antibodies result in three

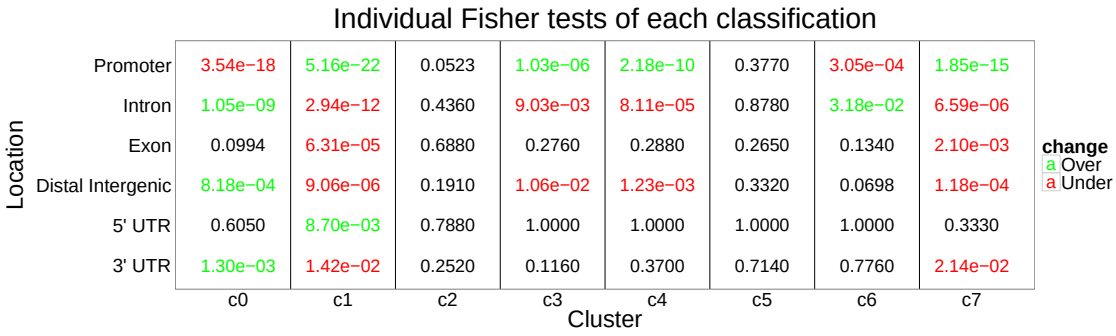
different types of enriched regions: sharp/punctate/narrow, broad and mixed [40]. Sharp peaks are typical of TFs or histone modifications at regulatory elements. Broad regions are more typical of histone modifications that mark domains such as repressed or transcribed regions. RNA polymerase generates peak data that can contain both broad and narrow regions due to the dynamic nature of its interactions with DNA. Current peak calling tools require prior knowledge of the type of peak expected based on the tag data. MACS and spp can handle both broad and narrow peaks but require the distinction to be specified by the user. Both use a Poisson model, with slight variations, which allows consistent p-values to be obtained across peaks by using both the ratio of mapped reads between the target and control samples as well as the absolute tag numbers [40, 28, 11]. MACS is optimised for single end reads while spp is optimised for paired end reads. MACS also shifts the tags to create a uniform peak for each location rather than strand specific peaks. MACS is also more user friendly and was selected as the best peak calling tool for this experiment.

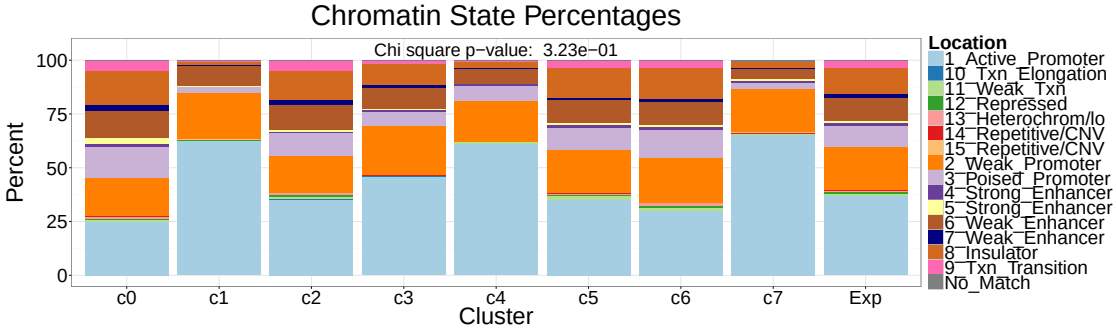
B TF results

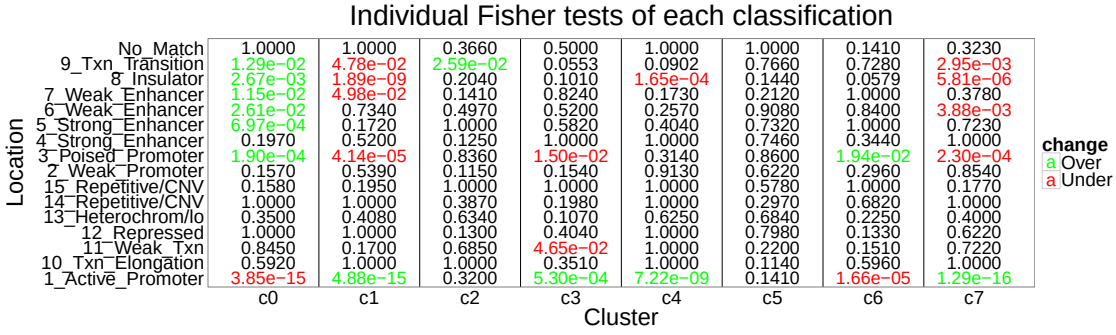
B.1 GABP



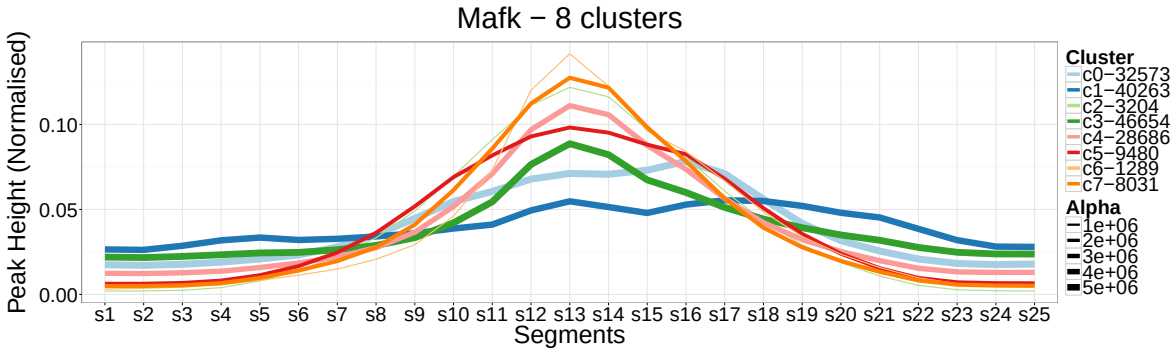


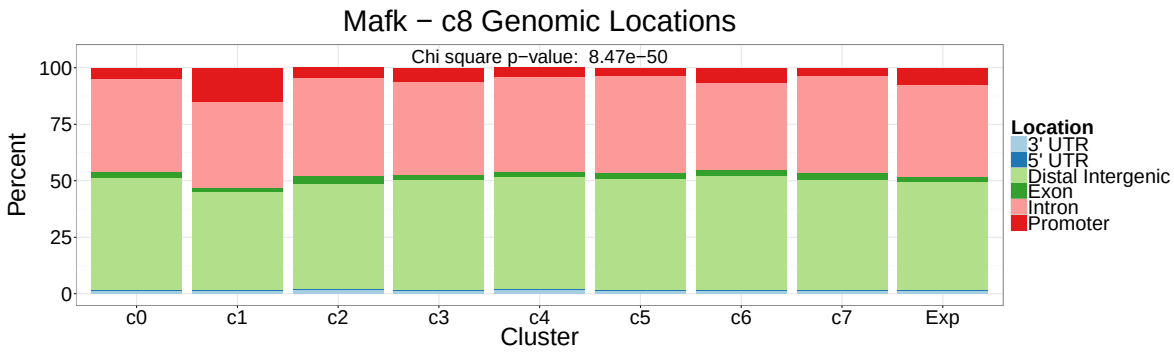






B.2 MAFK



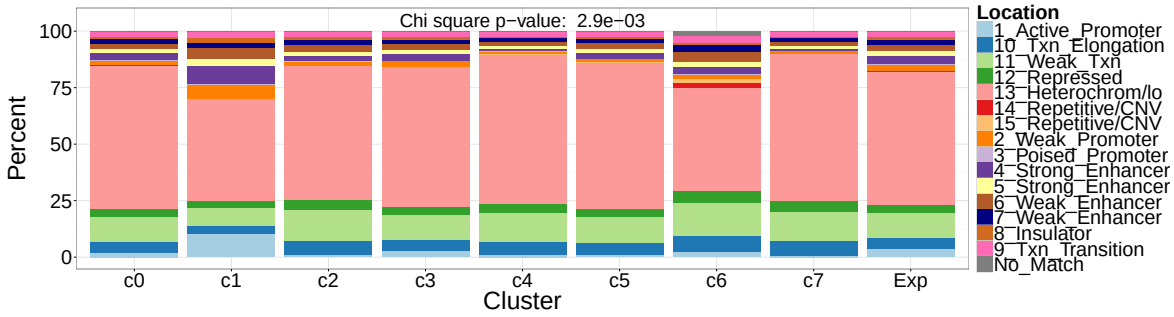


Individual Fisher tests of each classification

Location	c0	c1	c2	c3	c4	c5	c6	c7
Promoter	1.10e−93	0.00e+00	6.94e−11	1.37e−29	9.64e−143	5.24e−61	0.2380	8.47e−50
Intron	1.17e−02	5.42e−40	4.62e−03	3.92e−02	1.96e−05	1.39e−05	0.1810	1.02e−04
Exon	0.9050	5.03e−06	2.95e−05	7.28e−04	1.33e−02	1.63e−03	1.0000	8.10e−07
Distal Intergenic	4.46e−15	9.04e−95	0.1380	1.43e−07	3.19e−12	3.34e−02	3.82e−02	0.2380
5' UTR	0.3260	0.0621	0.1750	0.3160	0.1150	0.0757	0.8110	3.45e−02
3' UTR	0.9590	0.5860	0.1590	0.1830	0.2940	0.4560	0.7280	0.4480

Cluster

change: a Over, a Under



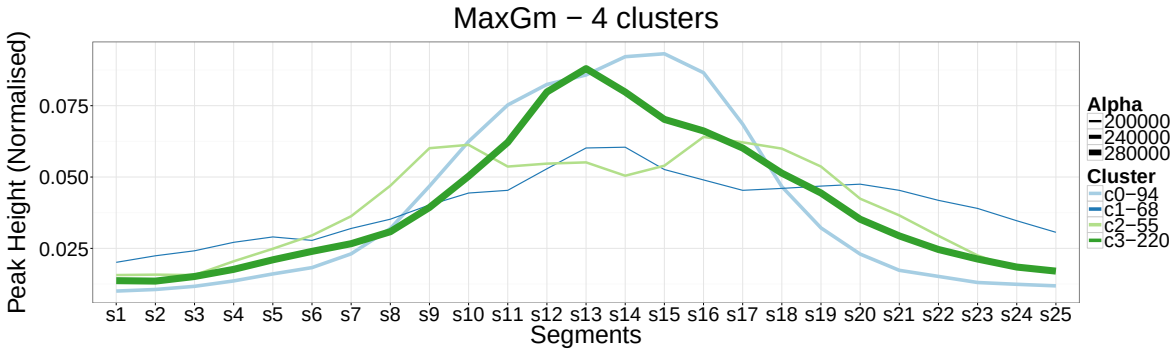
Individual Fisher tests of each classification

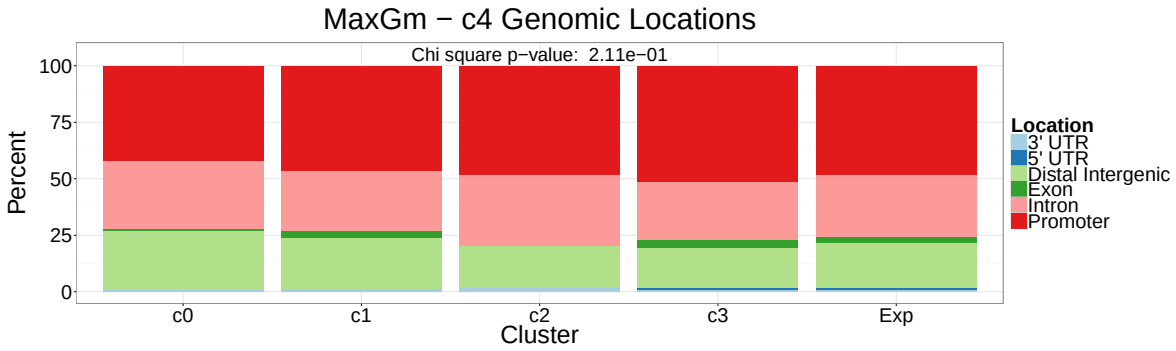
Location	c0	c1	c2	c3	c4	c5	c6	c7
No_Match	0.5370	0.3660	6.56e−03	2.62e−02	0.6270	0.0516	3.65e−12	2.90e−03
9_Txn_Transition	0.1070	1.59e−04	0.0619	4.26e−03	0.0582	0.0771	2.49e−02	0.5770
8_Insulator	3.30e−03	4.36e−86	0.3660	0.1840	1.76e−33	1.81e−12	0.7220	2.59e−12
7_Weak_Enhancer	0.0945	3.12e−05	0.2660	0.3290	8.82e−04	0.1750	5.29e−03	0.4760
6_Weak_Enhancer	5.58e−07	1.28e−119	0.7960	7.75e−03	9.41e−45	7.75e−07	7.22e−03	3.25e−13
5_Strong_Enhancer	6.09e−03	2.20e−86	0.5160	5.69e−05	2.68e−46	0.7290	0.6090	8.49e−11
4_Strong_Enhancer	7.09e−13	0.00e+00	6.16e−08	2.10e−32	3.27e−152	3.19e−09	0.1810	4.51e−55
3_Poised_Promoter	2.91e−02	1.79e−19	0.2890	0.2920	7.44e−08	1.64e−04	0.8100	0.3360
2_Weak_Promoter	2.04e−31	0.00e+00	5.42e−06	1.30e−10	1.13e−77	1.54e−22	0.1230	9.95e−26
15_Repetitive/CNV	3.36e−03	0.1790	0.7380	0.3240	1.0000	0.5620	5.26e−19	0.5320
14_Repetitive/CNV	0.3270	0.8200	1.0000	0.6660	0.1980	3.41e−02	3.82e−32	0.2230
13_Heterochrom/lo	3.65e−74	0.00e+00	0.6370	4.58e−44	1.42e−150	5.67e−31	7.65e−23	5.30e−28
12_Repressed	0.8830	2.42e−36	0.0975	7.34e−03	9.46e−10	0.5010	5.91e−03	4.67e−10
11_Weak_Txn	6.73e−03	9.91e−95	1.07e−08	0.8670	9.89e−25	2.38e−04	1.65e−05	7.39e−11
10_Txn_Elongation	0.2470	1.92e−48	5.43e−03	2.00e−02	3.45e−17	0.6630	1.33e−03	1.32e−10
1_Active_Promoter	3.29e−114	0.00e+00	7.46e−22	2.77e−59	4.87e−233	4.88e−61	1.05e−02	4.16e−83

Cluster

change: a Over, a Under

B.3 MAXGm

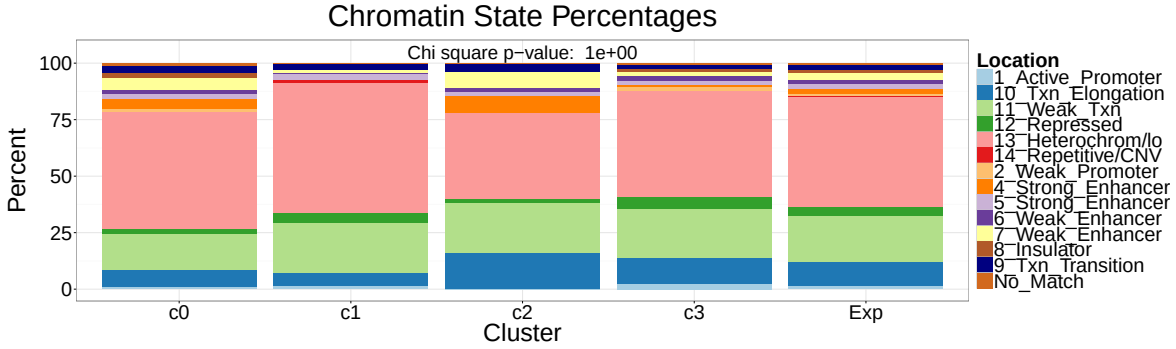




Individual Fisher tests of each classification

Location	c0	c1	c2	c3
Promoter	0.199	0.791	1.000	0.211
Intron	0.601	1.000	0.518	0.453
Exon	0.697	0.657	0.621	0.338
Distal Intergenic	0.147	0.624	0.857	0.192
5' UTR	1.000	1.000	1.000	0.499
3' UTR	1.000	0.570	0.488	0.682

Cluster



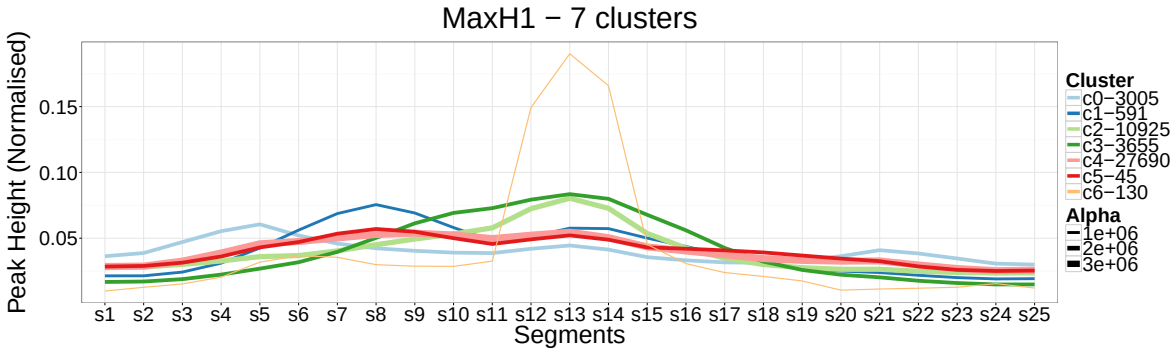
Individual Fisher tests of each classification

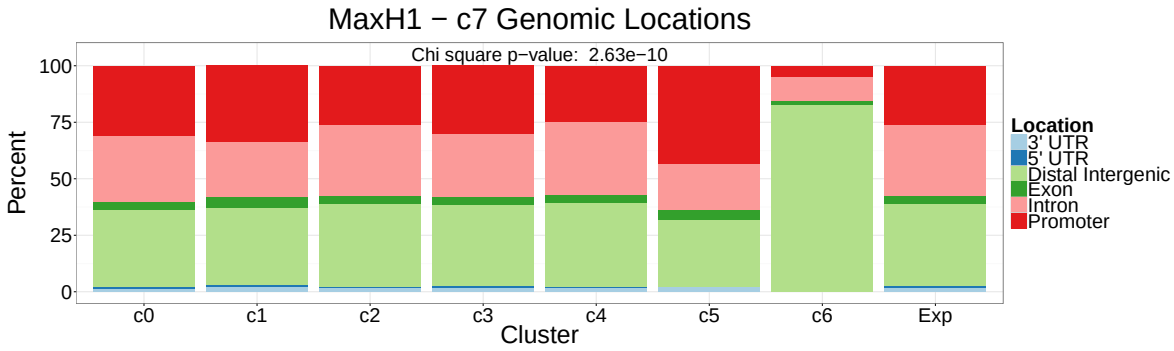
Location	c0	c1	c2	c3
No_Match	0.3840	1.0000	1.0000	1.0000
9_Txn_Transition	0.7090	0.6830	0.6360	0.3790
8_Insulator	0.2940	1.0000	1.0000	1.0000
7_Weak_Enhancer	0.1930	0.7060	0.0858	0.1110
6_Weak_Enhancer	0.6830	0.6160	1.0000	0.7240
5_Strong_Enhancer	1.0000	0.6360	1.0000	0.7500
4_Strong_Enhancer	0.2330	0.3730	0.0265	0.0610
2_Weak_Promoter	1.0000	1.0000	1.0000	0.3720
14_Repetitive/CNV	1.0000	0.1560	1.0000	0.4970
13_Heterochrom/lo	0.4850	0.1160	0.1130	0.5030
12_Repressed	0.3850	0.7500	0.7130	0.2280
11_Weak_Txn	0.2510	0.7430	0.7240	0.6360
10_Txn_Elongation	0.3440	0.2030	0.1550	0.4370
1_Active_Promoter	1.0000	1.0000	0.6030	0.4490

Cluster

change
a Over

B.4 MAXH1



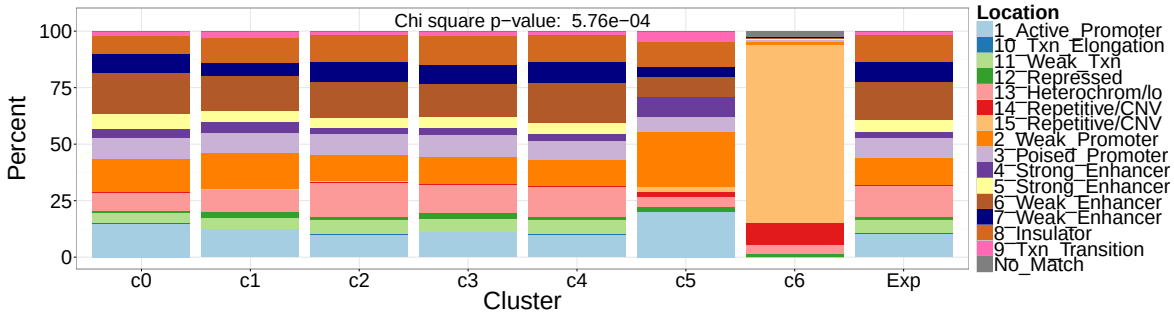


Individual Fisher tests of each classification

Location	c0	c1	c2	c3	c4	c5	c6
Promoter	1.30e−10	1.99e−05	0.920	1.48e−09	2.20e−13	1.46e−02	2.63e−10
Intron	1.89e−03	5.87e−05	0.741	4.06e−07	5.78e−10	0.143	5.71e−08
Exon	0.757	2.26e−02	6.37e−03	0.813	1.93e−03	0.665	0.333
Distal Intergenic	2.74e−03	0.156	0.467	0.124	2.70e−02	0.352	3.99e−27
5' UTR	0.508	1.000	0.277	8.40e−03	0.470	1.000	1.000
3' UTR	0.503	0.141	0.665	0.376	0.499	0.515	0.280

Cluster

change: Over, Under



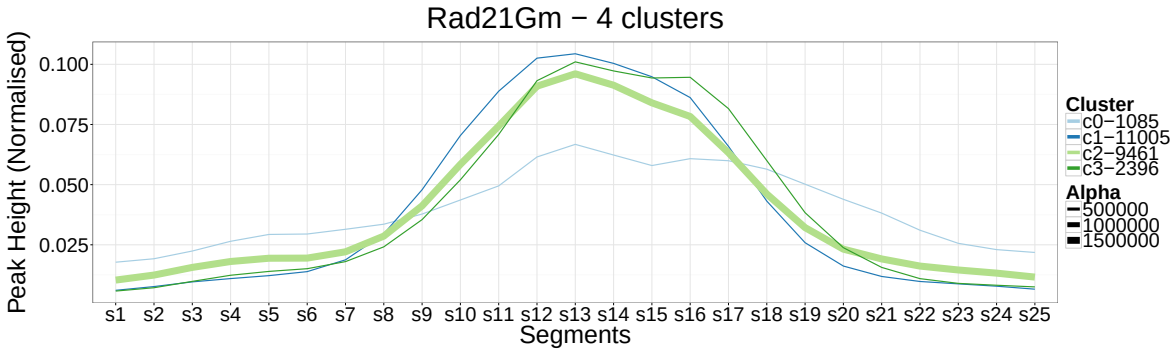
Individual Fisher tests of each classification

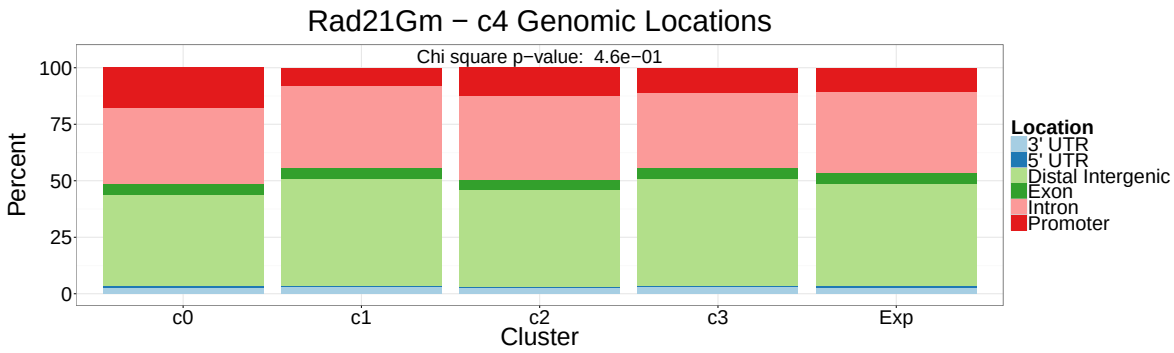
Location	c0	c1	c2	c3	c4	c5	c6
No_Match	0.5870	0.1660	0.5330	0.4570	0.7870	1.0000	5.76e−04
9_Txn_Transition	0.3750	0.1020	0.0591	0.3800	0.7090	0.1710	0.2840
8_Insulator	8.55e−12	0.6520	0.3310	2.59e−02	0.0754	1.0000	1.49e−07
7_Weak_Enhancer	0.5100	1.12e−02	0.8480	0.4340	4.98e−02	0.4320	1.51e−04
6_Weak_Enhancer	0.0861	0.3760	4.13e−04	4.91e−05	1.45e−07	0.2280	6.35e−11
5_Strong_Enhancer	1.79e−04	0.7770	8.23e−04	0.8140	0.1400	0.1720	1.55e−02
4_Strong_Enhancer	1.97e−04	1.27e−02	0.2530	0.6440	0.2230	4.10e−02	3.54e−02
3_Poised_Promoter	0.1820	0.7140	4.80e−02	2.59e−02	8.51e−04	0.7950	1.44e−04
2_Weak_Promoter	8.14e−06	7.16e−03	1.0000	0.6320	2.99e−03	1.78e−02	1.83e−05
15_Repetitive/CNV	1.42e−02	0.4600	1.89e−02	0.0543	2.25e−37	0.1430	1.42e−242
14_Repetitive/CNV	1.0000	0.6160	0.8920	0.6640	7.77e−04	0.0699	3.54e−20
13_Heterochrom/lo	9.30e−21	3.56e−03	2.89e−08	0.1300	0.1220	0.0809	4.30e−04
12_Repressed	0.1000	9.76e−03	0.6000	1.10e−07	4.26e−04	0.4550	0.6940
11_Weak_Txn	1.92e−04	0.3340	2.01e−02	0.7140	0.5050	0.1120	5.95e−04
10_Txn_Elongation	1.0000	0.1850	0.3470	0.1070	2.12e−02	1.0000	1.0000
1_Active_Promoter	5.32e−14	0.0781	3.94e−02	0.2360	3.10e−03	4.76e−02	9.19e−07

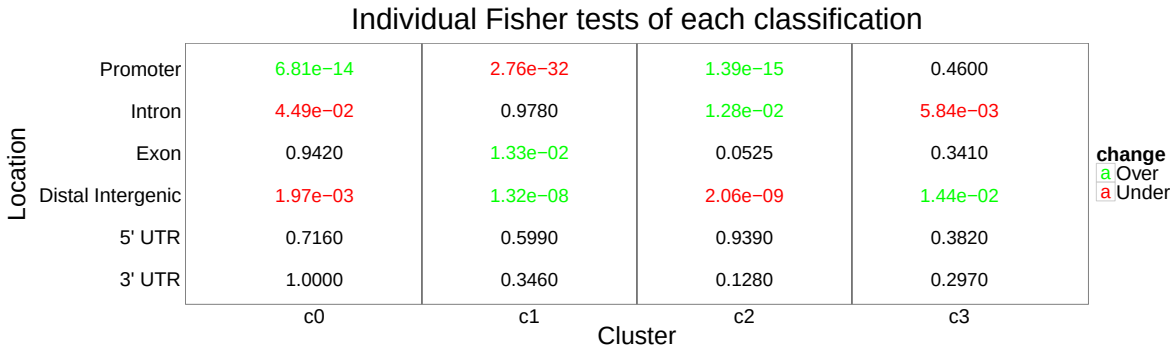
Cluster

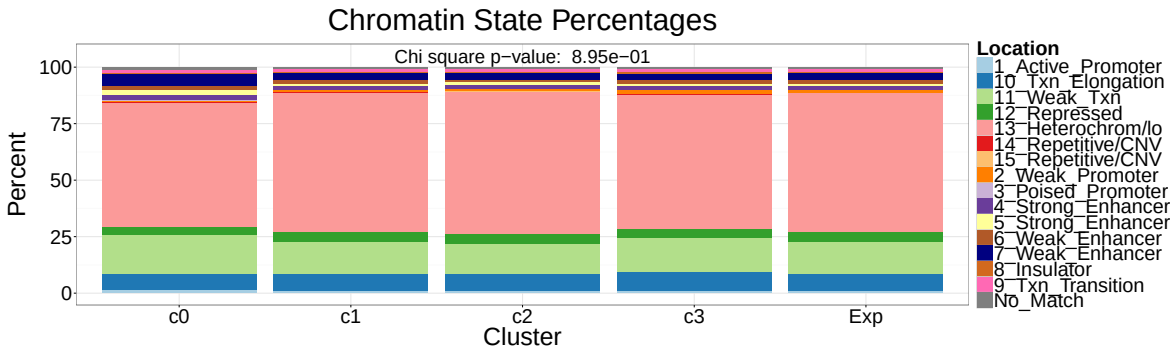
change: Over, Under

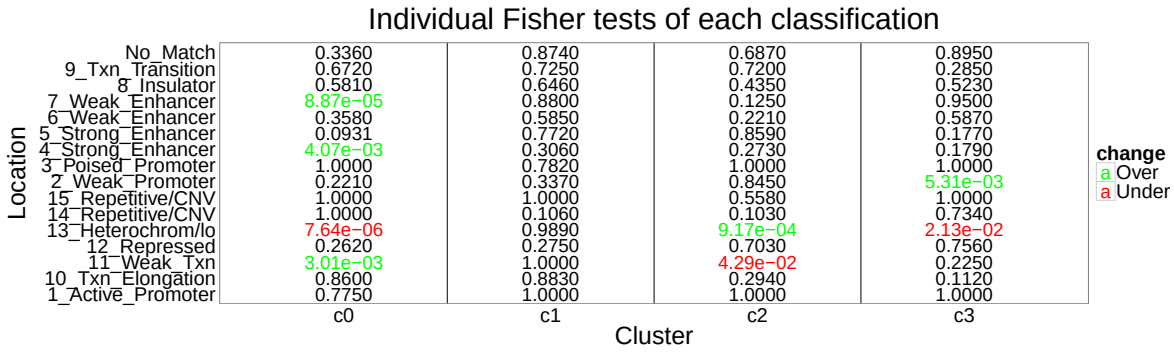
B.5 RAD21Gm


Rad21Gm − 4 clusters
Peak Height (Normalised)
0.100
0.075
0.050
0.025
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22 s23 s24 s25
Segments
Cluster
c0−1085
c1−11005
c2−9461
c3−2396
Alpha
500000
1000000
1500000

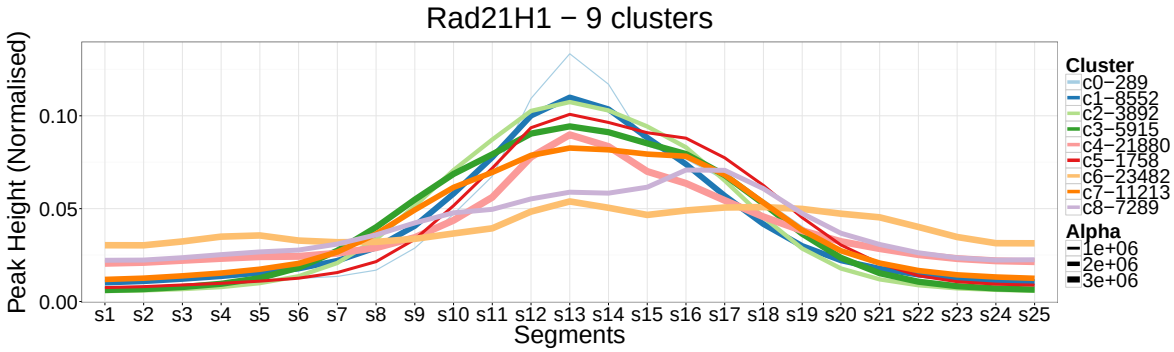

Rad21Gm − c4 Genomic Locations
Chi square p−value: 4.6e−01
Percent
100
75
50
25
0
c0 c1 c2 c3 Exp
Cluster
Location
3' UTR
5' UTR
Distal Intergenic
Exon
Intron
Promoter

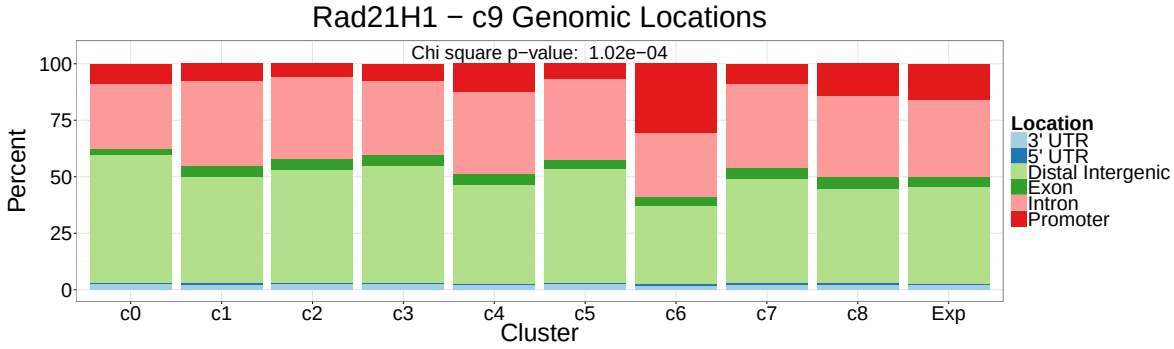

Individual Fisher tests of each classification
Location
Promoter 6.81e−14 2.76e−32 1.39e−15 0.4600
Intron 4.49e−02 0.9780 1.28e−02 5.84e−03
Exon 0.9420 1.33e−02 0.0525 0.3410
Distal Intergenic 1.97e−03 1.32e−08 2.06e−09 1.44e−02
5' UTR 0.7160 0.5990 0.9390 0.3820
3' UTR 1.0000 0.3460 0.1280 0.2970
c0 c1 c2 c3
Cluster
change
Over
Under


Chromatin State Percentages
Chi square p−value: 8.95e−01
Percent
100
75
50
25
0
c0 c1 c2 c3 Exp
Cluster
Location
1_Active_Promoter
10_Txn_Elongation
11_Weak_Txn
12_Repressed
13_Heterochrom/lo
14_Repetitive/CNV
15_Repetitive/CNV
2_Weak_Promoter
3_Poised_Promoter
4_Strong_Enhancer
5_Strong_Enhancer
6_Weak_Enhancer
7_Weak_Enhancer
8_Insulator
9_Txn_Transition
No_Match


Individual Fisher tests of each classification
Location
No_Match 0.3360 0.8740 0.6870 0.8950
9_Txn_Transition 0.6720 0.7250 0.7200 0.2850
8_Insulator 0.5810 0.6460 0.4350 0.5230
7_Weak_Enhancer 8.87e−05 0.8800 0.1250 0.9500
6_Weak_Enhancer 0.3580 0.5850 0.2210 0.5870
5_Strong_Enhancer 0.0931 0.7720 0.8590 0.1770
4_Strong_Enhancer 4.07e−03 0.3060 0.2730 0.1790
3_Poised_Promoter 1.0000 0.7820 1.0000 1.0000
2_Weak_Promoter 0.2210 0.3370 0.8450 5.31e−03
15_Repetitive/CNV 1.0000 1.0000 0.5580 1.0000
14_Repetitive/CNV 1.0000 0.1060 0.1030 0.7340
13_Heterochrom/lo 7.64e−06 0.9890 9.17e−04 2.13e−02
12_Repressed 0.2620 0.2750 0.7030 0.7560
11_Weak_Txn 3.01e−03 1.0000 4.29e−02 0.2250
10_Txn_Elongation 0.8600 0.8830 0.2940 0.1120
1_Active_Promoter 0.7750 1.0000 1.0000 1.0000
c0 c1 c2 c3
Cluster
change
Over
Under

B.6 RAD21H1



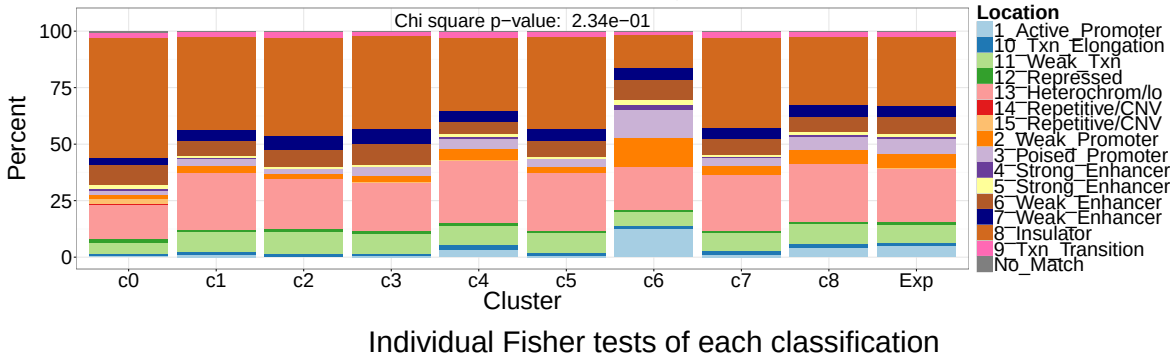


Individual Fisher tests of each classification

Location	c0	c1	c2	c3	c4	c5	c6	c7	c8
Promoter	4.82e−04	3.05e−129	3.62e−87	1.44e−90	1.80e−59	2.88e−32	0.00e+00	2.61e−125	1.02e−04
Intron	0.0707	3.92e−13	1.60e−03	0.1850	6.86e−18	0.0879	4.76e−118	1.84e−16	5.73e−04
Exon	0.1600	0.1380	0.3930	0.8480	0.2980	0.0867	1.07e−04	0.2210	0.0677
Distal Intergenic	2.13e−06	1.18e−16	5.02e−21	7.63e−46	2.00e−03	1.89e−11	1.70e−187	1.69e−14	0.2110
5' UTR	1.0000	0.4080	0.0906	0.6840	0.7400	0.3050	0.5770	0.9030	0.6570
3' UTR	0.4160	0.1580	0.0801	0.0705	0.9350	0.0977	1.15e−05	0.2360	0.9330

Cluster

change: a Over, a Under



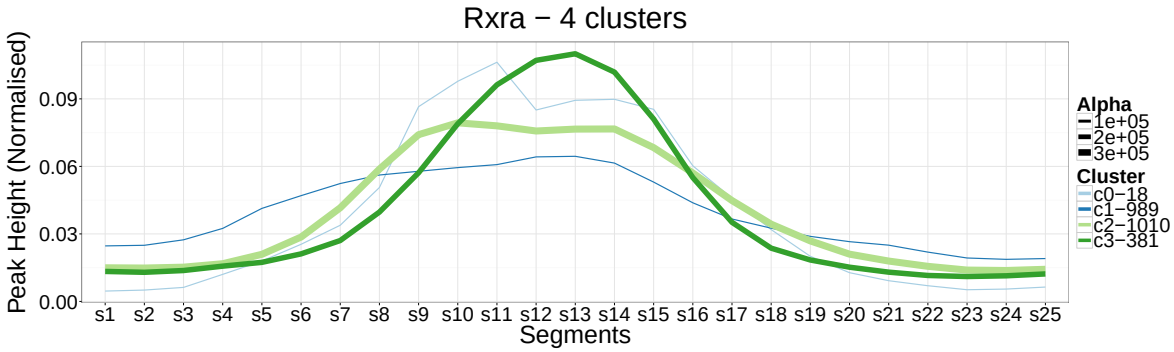
Individual Fisher tests of each classification

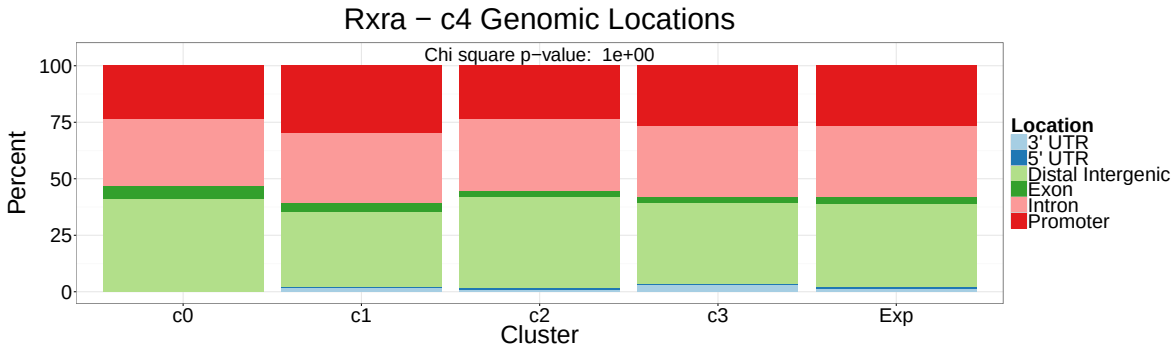
Location	c0	c1	c2	c3	c4	c5	c6	c7	c8
No_Match	0.0879	0.8900	0.8410	0.8700	3.58e−02	0.7730	0.7800	0.2680	0.2340
9_Txn_Transition	1.0000	0.2730	0.1570	0.3480	6.61e−06	0.8730	6.95e−17	4.72e−05	0.5430
8_Insulator	4.64e−16	7.87e−107	6.43e−71	1.24e−72	7.91e−14	2.36e−20	0.00e+00	2.00e−113	0.6890
7_Weak_Enhancer	0.1120	0.8780	0.0658	4.93e−06	5.53e−06	1.0000	0.1020	0.1460	0.2720
6_Weak_Enhancer	0.3660	1.40e−03	0.1760	6.61e−09	3.23e−27	0.9260	1.50e−25	0.2680	1.45e−02
5_Strong_Enhancer	0.2890	1.26e−05	5.61e−07	3.72e−04	8.40e−05	2.93e−03	5.08e−46	9.17e−07	0.9570
4_Strong_Enhancer	1.0000	5.61e−10	1.23e−09	5.84e−12	1.29e−05	4.32e−04	6.51e−76	1.60e−10	0.4890
3_Poised_Promoter	2.71e−04	4.32e−45	1.59e−35	6.90e−23	2.61e−32	7.43e−10	0.00e+00	5.01e−45	3.88e−03
2_Weak_Promoter	1.04e−03	7.14e−54	4.43e−40	9.83e−35	1.56e−32	3.05e−12	0.00e+00	1.49e−42	0.3300
15_Repetitive/CNV	6.87e−10	0.1120	0.6440	0.7200	0.5280	1.0000	0.4130	0.4190	0.5110
14_Repetitive/CNV	0.3120	0.3400	1.0000	0.1280	0.6630	0.0785	0.0865	0.8880	1.0000
13_Heterochrom/lo	5.04e−04	7.86e−03	3.70e−02	2.26e−05	3.24e−56	0.0999	3.00e−79	2.25e−03	7.78e−04
12_Repressed	0.2330	0.4650	0.0516	3.32e−02	0.3530	0.6350	1.30e−05	0.4540	0.1840
11_Weak_Txn	4.97e−02	6.48e−04	2.78e−04	4.94e−02	2.64e−07	0.1970	2.21e−34	0.6130	2.83e−03
10_Txn_Elongation	0.4530	0.7750	0.4100	1.89e−03	3.51e−11	1.0000	1.56e−07	0.7670	0.1830
1_Active_Promoter	7.34e−05	7.60e−104	3.84e−73	1.44e−93	2.03e−47	1.62e−28	0.00e+00	9.86e−112	4.06e−04

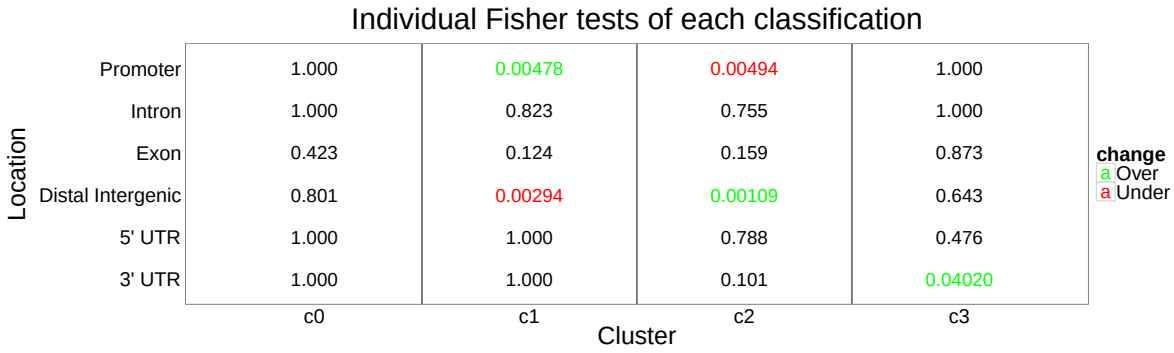
Cluster

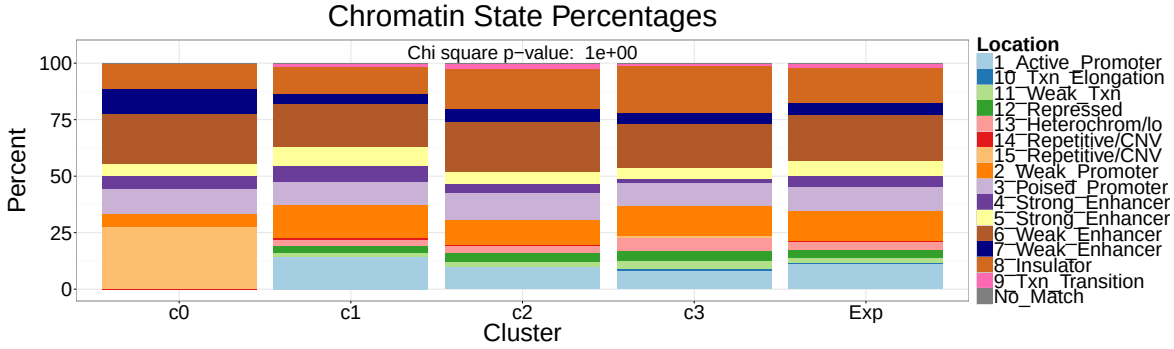
change: a Over, a Under

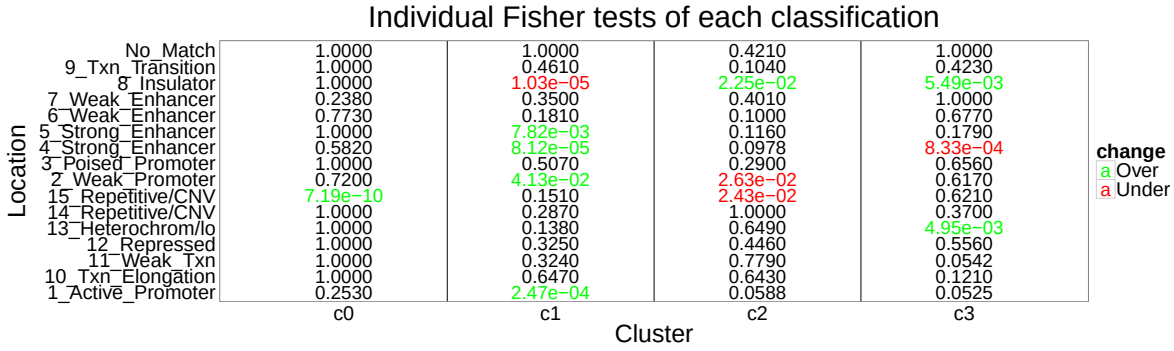
B.7 RXRA


Rxra − 4 clusters
Peak Height (Normalised)
0.00
0.03
0.06
0.09
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22 s23 s24 s25
Segments
Alpha
1e+05
2e+05
3e+05
Cluster
c0−18
c1−989
c2−1010
c3−381

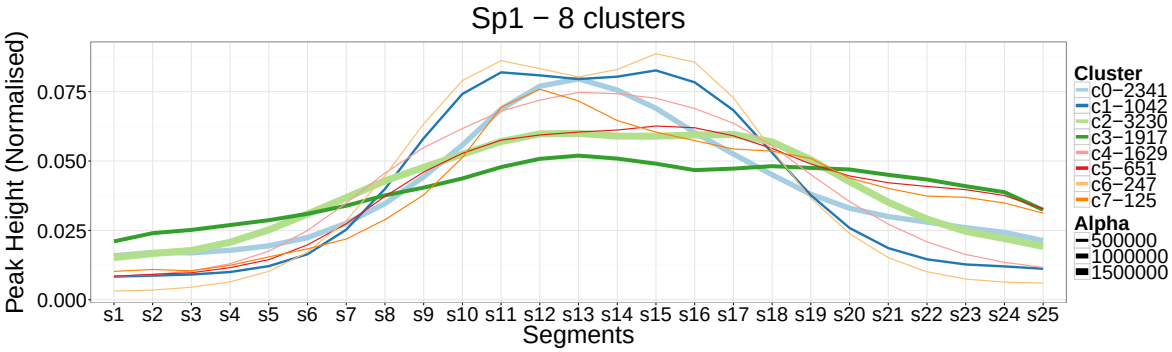

Rxra − c4 Genomic Locations
Chi square p−value: 1e+00
Percent
0
25
50
75
100
c0 c1 c2 c3 Exp
Cluster
Location
3' UTR
5' UTR
Distal Intergenic
Exon
Intron
Promoter

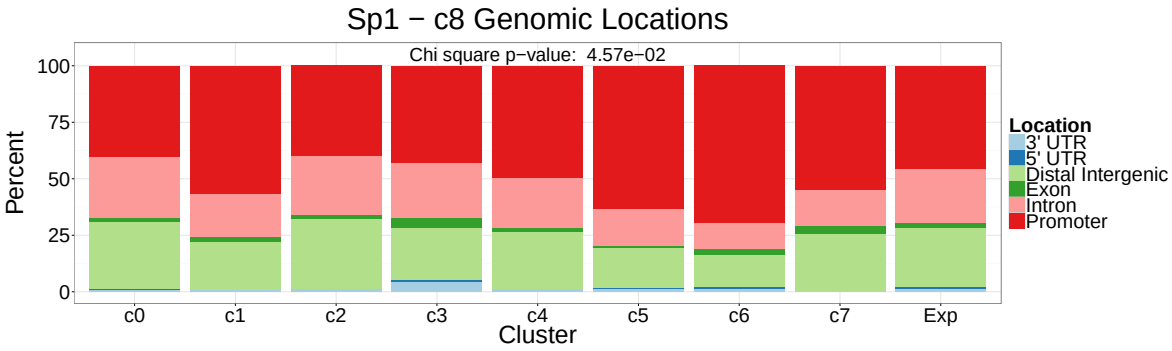

Individual Fisher tests of each classification
Location
Promoter 1.000 0.00478 0.00494 1.000
Intron 1.000 0.823 0.755 1.000
Exon 0.423 0.124 0.159 0.873
Distal Intergenic 0.801 0.00294 0.00109 0.643
5' UTR 1.000 1.000 0.788 0.476
3' UTR 1.000 1.000 0.101 0.04020
c0 c1 c2 c3
Cluster
change
Over
Under


Chromatin State Percentages
Chi square p−value: 1e+00
Percent
0
25
50
75
100
c0 c1 c2 c3 Exp
Cluster
Location
1_Active_Promoter
10_Txn_Elongation
11_Weak_Txn
12_Repressed
13_Heterochrom/lo
14_Repetitive/CNV
15_Repetitive/CNV
2_Weak_Promoter
3_Poised_Promoter
4_Strong_Enhancer
5_Strong_Enhancer
6_Weak_Enhancer
7_Weak_Enhancer
8_Insulator
9_Txn_Transition
No_Match


Individual Fisher tests of each classification
Location
No_Match 1.0000 1.0000 0.4210 1.0000
9_Txn_Transition 1.0000 0.4610 0.1040 0.4230
8_Insulator 1.0000 1.03e−05 2.25e−02 5.49e−03
7_Weak_Enhancer 0.2380 0.3500 0.4010 1.0000
6_Weak_Enhancer 0.7730 0.1810 0.1000 0.6770
5_Strong_Enhancer 1.0000 7.82e−03 0.1160 0.1790
4_Strong_Enhancer 0.5820 8.12e−05 0.0978 8.33e−04
3_Poised_Promoter 1.0000 0.5070 0.2900 0.6560
2_Weak_Promoter 0.7200 4.13e−02 2.63e−02 0.6170
15_Repetitive/CNV 7.19e−10 0.1510 2.43e−02 0.6210
14_Repetitive/CNV 1.0000 0.2870 1.0000 0.3700
13_Heterochrom/lo 1.0000 0.1380 0.6490 4.95e−03
12_Repressed 1.0000 0.3250 0.4460 0.5560
11_Weak_Txn 1.0000 0.3240 0.7790 0.0542
10_Txn_Elongation 1.0000 0.6470 0.6430 0.1210
1_Active_Promoter 0.2530 2.47e−04 0.0588 0.0525
c0 c1 c2 c3
Cluster
change
Over
Under

B.8 SP1



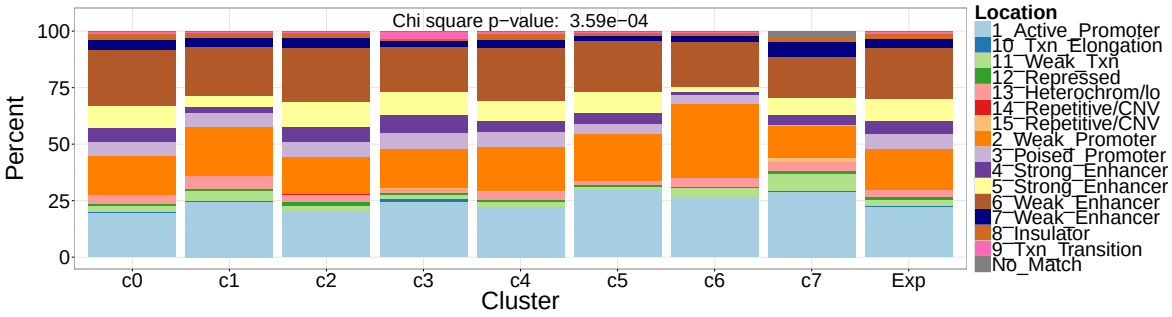


Individual Fisher tests of each classification

Location	c0	c1	c2	c3	c4	c5	c6	c7
Promoter	3.44e−09	1.60e−13	1.95e−14	6.50e−03	5.69e−04	1.40e−20	2.09e−14	4.57e−02
Intron	3.61e−05	4.24e−04	1.09e−03	0.4090	0.1080	4.71e−06	9.98e−07	0.0551
Exon	0.1760	0.6740	0.0510	1.99e−10	0.1940	5.76e−03	0.6740	0.5510
Distal Intergenic	9.12e−05	3.91e−05	1.22e−12	2.31e−04	0.2110	1.47e−07	3.19e−06	1.0000
5' UTR	0.8650	0.0879	0.1660	2.42e−03	1.0000	1.0000	0.1060	1.0000
3' UTR	6.97e−03	0.0913	4.65e−03	6.13e−20	0.0542	0.5220	1.0000	0.2710

Cluster

change: a Over, a Under



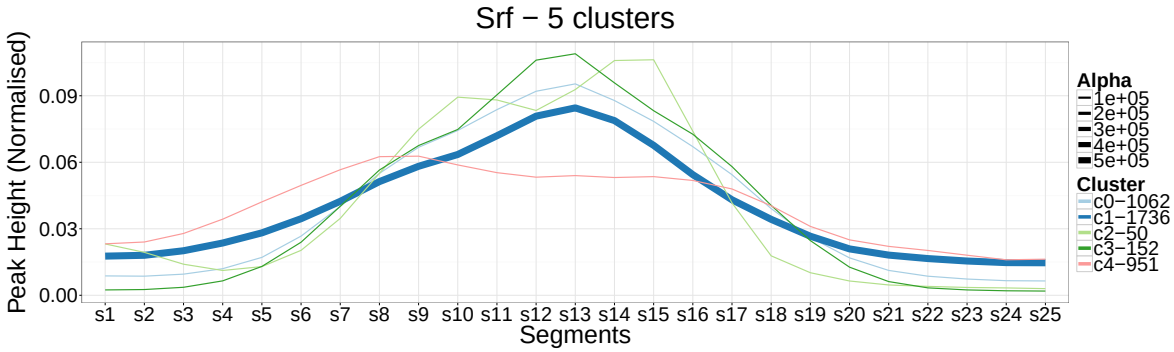
Individual Fisher tests of each classification

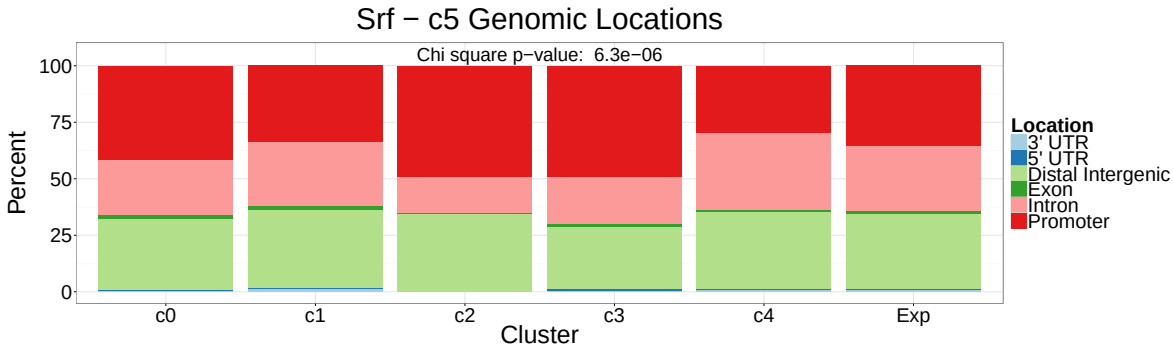
Location	c0	c1	c2	c3	c4	c5	c6	c7
No_Match	0.3250	0.6250	0.5400	0.4770	0.7070	1.0000	1.0000	3.59e−04
9_Txn_Transition	0.6060	0.0810	1.22e−02	4.33e−12	0.4050	4.88e−02	0.7730	0.4140
8_Insulator	0.0538	0.7250	0.5480	2.00e−03	4.21e−02	0.7700	0.6380	0.7390
7_Weak_Enhancer	3.25e−02	0.2660	2.82e−02	5.49e−05	0.5740	5.56e−03	0.6100	0.1470
6_Weak_Enhancer	3.57e−02	0.2460	0.1120	1.73e−03	0.4830	1.0000	0.2520	0.2410
5_Strong_Enhancer	0.1510	5.19e−09	5.19e−05	0.3900	0.2150	0.8350	4.26e−06	0.5360
4_Strong_Enhancer	0.9220	2.97e−06	0.0878	5.49e−06	0.1770	0.3100	1.08e−03	0.7060
3_Poised_Promoter	0.9620	0.9460	0.4130	0.1200	0.7400	3.66e−02	0.1090	5.66e−04
2_Weak_Promoter	0.0705	6.88e−03	2.22e−03	0.2160	0.1100	0.0526	2.24e−08	0.2960
15_Repetitive/CNV	0.3570	0.4960	0.6810	0.1030	0.6030	1.0000	1.0000	2.51e−03
14_Repetitive/CNV	1.0000	0.5430	0.2400	0.6320	0.6130	1.0000	1.0000	1.0000
13_Heterochrom/lo	0.0972	7.38e−05	0.9090	1.31e−05	0.6590	1.03e−02	0.5940	0.6220
12_Repressed	0.4530	0.3690	4.21e−02	0.2440	0.9010	0.4510	0.3750	0.6610
11_Weak_Txn	0.8230	5.96e−05	0.5930	1.66e−03	0.7310	0.0683	0.0591	4.04e−03
10_Txn_Elongation	3.48e−03	0.5740	7.41e−06	5.82e−18	7.10e−03	0.2700	1.0000	0.3330
1_Active_Promoter	8.29e−04	0.1100	2.10e−03	1.90e−02	0.8220	7.99e−06	0.1430	0.1050

Cluster

change: a Over, a Under

B.9 SRF



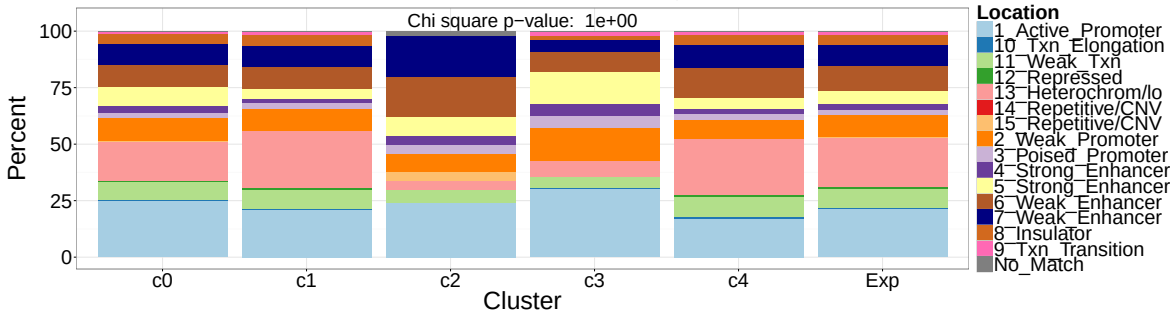


Individual Fisher tests of each classification

Location	c0	c1	c2	c3	c4
Promoter	2.24e−06	2.94e−02	0.0518	6.72e−04	6.30e−06
Intron	7.02e−04	0.9430	0.0777	0.0528	2.41e−06
Exon	0.0620	1.0000	1.0000	1.0000	0.0551
Distal Intergenic	0.1020	0.1180	0.8790	0.1130	0.5540
5' UTR	0.1180	0.3110	1.0000	0.2680	1.0000
3' UTR	0.2320	4.72e−02	1.0000	1.0000	0.7230

Cluster

change: a Over, a Under



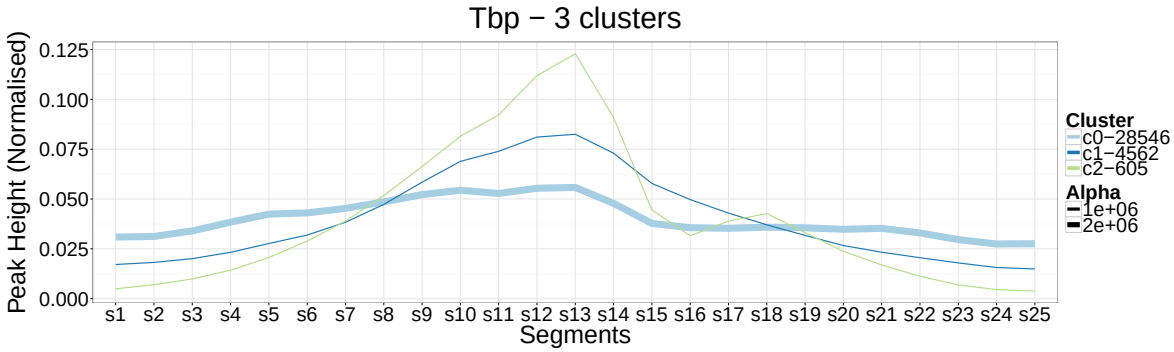
Individual Fisher tests of each classification

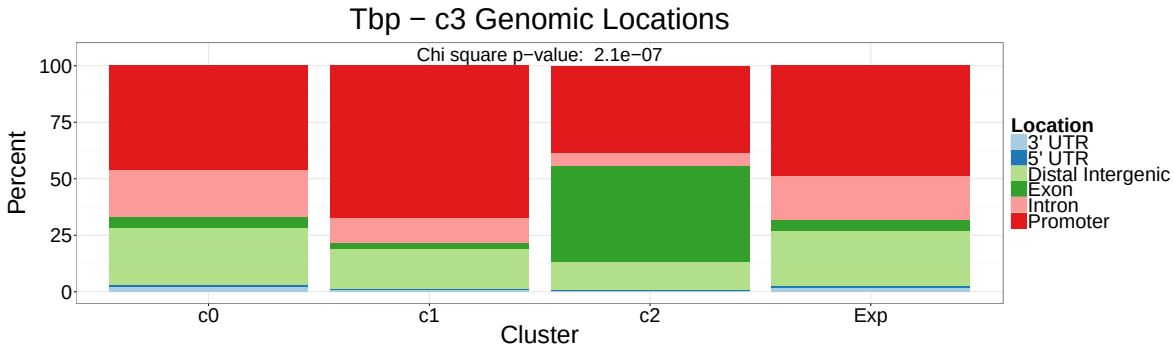
Location	c0	c1	c2	c3	c4
No_Match	0.5690	0.5860	3.75e−02	1.0000	1.0000
9_Txn_Transition	1.0000	1.0000	1.0000	0.4630	1.0000
8_Insulator	0.3470	0.0655	0.1710	0.1620	1.0000
7_Weak_Enhancer	1.0000	0.3800	4.79e−02	0.0875	0.1810
6_Weak_Enhancer	0.4180	0.1480	0.1060	0.4250	1.17e−02
5_Strong_Enhancer	1.79e−04	3.08e−05	0.5370	6.19e−05	0.1560
4_Strong_Enhancer	0.2500	0.1000	0.3580	0.0550	0.7220
3_Poised_Promoter	0.4900	1.0000	0.3530	0.0540	0.8110
2_Weak_Promoter	0.4310	0.9570	1.0000	0.0501	0.1310
15_Repetitive/CNV	1.0000	0.5070	1.57e−04	1.0000	1.0000
14_Repetitive/CNV	0.3330	1.0000	1.0000	1.0000	0.0938
13_Heterochrom/lo	3.74e−05	4.00e−05	8.21e−04	1.09e−06	2.76e−02
12_Repressed	0.4230	0.1060	1.0000	0.6340	1.0000
11_Weak_Txn	0.4720	0.8150	0.7950	0.1290	0.1740
10_Txn_Elongation	0.8140	0.6800	1.0000	0.5950	0.2260
1_Active_Promoter	7.59e−04	0.4590	0.6070	8.76e−03	5.36e−05

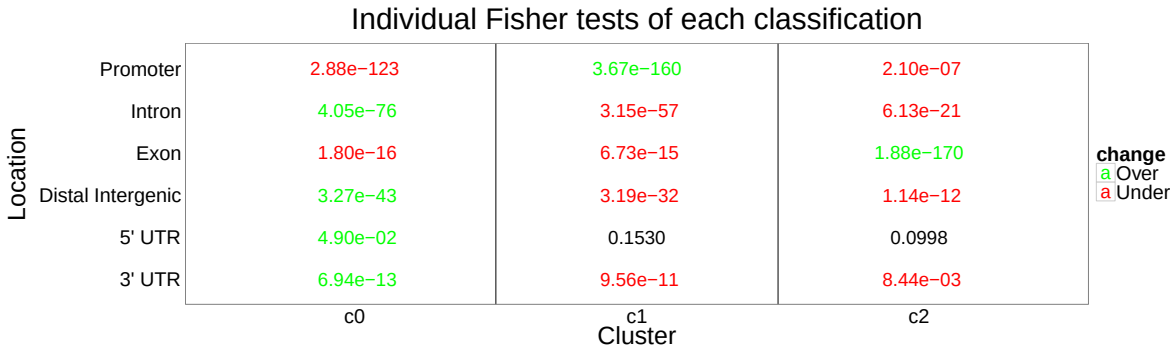
Cluster

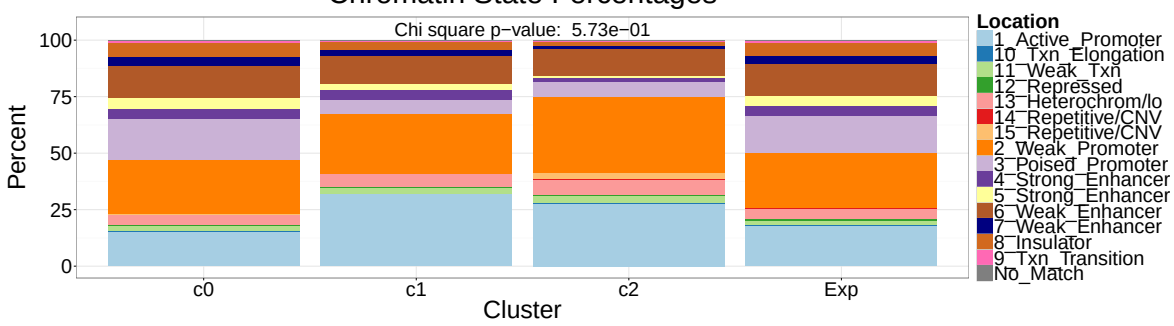
change: a Over, a Under

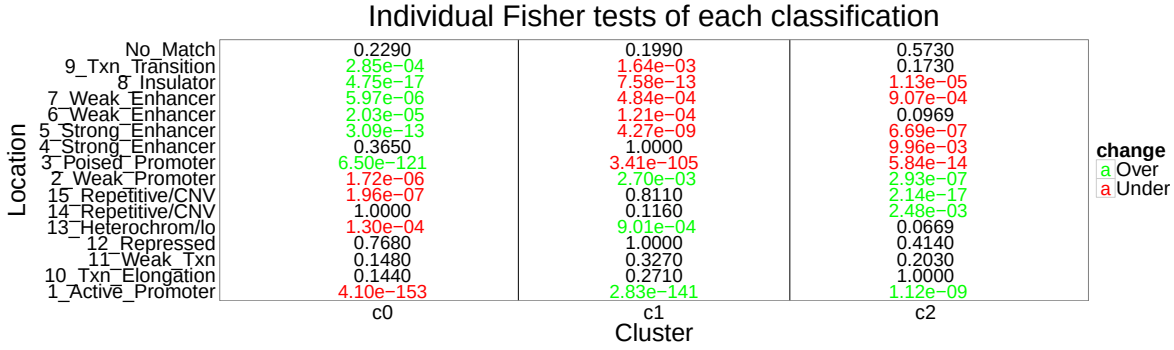
B.10 TBP


Tbp − 3 clusters
Peak Height (Normalised)
0.125
0.100
0.075
0.050
0.025
0.000
s1 s2 s3 s4 s5 s6 s7 s8 s9 s10 s11 s12 s13 s14 s15 s16 s17 s18 s19 s20 s21 s22 s23 s24 s25
Segments
Cluster
c0−28546
c1−4562
c2−605
Alpha
1e+06
2e+06


Tbp − c3 Genomic Locations
Chi square p−value: 2.1e−07
Percent
100
75
50
25
0
c0
c1
c2
Exp
Cluster
Location
3' UTR
5' UTR
Distal Intergenic
Exon
Intron
Promoter


Individual Fisher tests of each classification
Location
Promoter 2.88e−123 3.67e−160 2.10e−07
Intron 4.05e−76 3.15e−57 6.13e−21
Exon 1.80e−16 6.73e−15 1.88e−170
Distal Intergenic 3.27e−43 3.19e−32 1.14e−12
5' UTR 4.90e−02 0.1530 0.0998
3' UTR 6.94e−13 9.56e−11 8.44e−03
c0
c1
c2
Cluster
change
Over
Under


Chromatin State Percentages
Chi square p−value: 5.73e−01
Percent
100
75
50
25
0
c0
c1
c2
Exp
Cluster
Location
1_Active_Promoter
10_Txn_Elongation
11_Weak_Txn
12_Repressed
13_Heterochrom/lo
14_Repetitive/CNV
15_Repetitive/CNV
2_Weak_Promoter
3_Poised_Promoter
4_Strong_Enhancer
5_Strong_Enhancer
6_Weak_Enhancer
7_Weak_Enhancer
8_Insulator
9_Txn_Transition
No_Match


Individual Fisher tests of each classification
Location
No_Match 0.2290 0.1990 0.5730
9_Txn_Transition 2.85e−04 1.64e−03 0.1730
8_Insulator 4.75e−17 7.58e−13 1.13e−05
7_Weak_Enhancer 5.97e−06 4.84e−04 9.07e−04
6_Weak_Enhancer 2.03e−05 1.21e−04 0.0969
5_Strong_Enhancer 3.09e−13 4.27e−09 6.69e−07
4_Strong_Enhancer 0.3650 1.0000 9.96e−03
3_Poised_Promoter 6.50e−121 3.41e−105 5.84e−14
2_Weak_Promoter 1.72e−06 2.70e−03 2.93e−07
15_Repetitive/CNV 1.96e−07 0.8110 2.14e−17
14_Repetitive/CNV 1.0000 0.1160 2.48e−03
13_Heterochrom/lo 1.30e−04 9.01e−04 0.0669
12_Repressed 0.7680 1.0000 0.4140
11_Weak_Txn 0.1480 0.3270 0.2030
10_Txn_Elongation 0.1440 0.2710 1.0000
1_Active_Promoter 4.10e−153 2.83e−141 1.12e−09
c0
c1
c2
Cluster
change
Over
Under

References

[1] S Andrews et al. Fastqc: A quality control tool for high throughput sequence data. *Reference Source*, 2010.

[2] Timothy Bailey, Pawel Krajewski, Istvan Ladunga, Celine Lefebvre, Qunhua Li, Tao Liu, Pedro Madrigal, Cenny Taslim, and Jie Zhang. Practical guidelines for the comprehensive analysis of chip-seq data. *PLoS computational biology*, 9(11):e1003326, 2013.

[3] Timothy L Bailey, Mikael Boden, Fabian A Buske, Martin Frith, Charles E Grant, Luca Clementi, Jingyuan Ren, Wilfred W Li, and William S Noble. Meme suite: tools for motif discovery and searching. *Nucleic acids research*, page gkp335, 2009.

[4] Michael F Berger, Anthony A Philippakis, Aaron M Qureshi, Fangxue S He, Preston W Estep, and Martha L Bulyk. Compact, universal dna microarrays to comprehensively determine transcription-factor binding site specificities. *Nature biotechnology*, 24(11):1429–1435, 2006.

[5] Alan P Boyle, Lingyun Song, Bum-Kyu Lee, Darin London, Damian Keefe, Ewan Birney, Vishwanath R Iyer, Gregory E Crawford, and Terrence S Furey. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. *Genome research*, 21(3):456–464, 2011.

[6] Robert C Brewster, Franz M Weinert, Hernan G Garcia, Dan Song, Mattias Rydenfelt, and Rob Phillips. The transcription factor titration effect dictates level of gene expression. *Cell*, 156(6):1312–1323, 2014.

[7] Carsten Carlberg and Ferdinand Molnár. The basal transcriptional machinery. pages 37–54, 2014.

[8] Aaron Diaz, Kiyoub Park, Daniel A Lim, and Jun S Song. Normalization, bias correction, and peak calling for chip-seq. *Statistical applications in genetics and molecular biology*, 11(3), 2012.

[9] ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, sep 2012.

[10] Peggy J. Farnham. Insights from genomic profiling of transcription factors. *Nature Reviews Genetics*, 10(9):605–616, sep 2009.

[11] Jianxing Feng, Tao Liu, Bo Qin, Yong Zhang, and Xiaole Shirley Liu. Identifying chip-seq enrichment using macs. *Nature protocols*, 7(9):1728–1740, 2012.

[12] E. W. Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[13] Seth Frietze and Peggy J Farnham. Transcription factor effector domains. In *A Handbook of Transcription Factors*, pages 261–277. Springer, 2011.

[14] Alister PW Funnell and Merlin Crossley. Homo-and heterodimerization in transcriptional regulation. In *Protein Dimerization and Oligomerization in Biology*, pages 105–121. Springer, 2012.

[15] Terrence S Furey. Chip–seq and beyond: new and improved methodologies to detect and characterize protein–dna interactions. *Nature Reviews Genetics*, 13(12):840–852, 2012.

[16] Michael J Guertin and John T Lis. Chromatin landscape dictates hsf binding to target dna elements. *PLoS genetics*, 6(9):e1001114, 2010.

[17] Jay R Hesselberth, Xiaoyu Chen, Zhihong Zhang, Peter J Sabo, Richard Sandstrom, Alex P Reynolds, Robert E Thurman, Shane Neph, Michael S Kuehn, William S Noble, et al. Global mapping of protein-dna interactions in vivo by digital genomic footprinting. *Nature methods*, 6(4):283–289, 2009.

[18] Christine E Horak and Michael Snyder. Chip-chip: a genomic approach for identifying transcription factor binding sites. *Methods in enzymology*, 350:469–483, 2002.

[19] TR Hughes. Introduction to a handbook of transcription factors. In *A Handbook of Transcription Factors*, pages 1–6. Springer, 2011.

[20] Maxwell A Hume, Luis A Barrera, Stephen S Gisselbrecht, and Martha L Bulyk. Uniprobe, update 2015: new tools and content for the online database of protein-binding microarray data on protein–dna interactions. *Nucleic acids research*, page gku1045, 2014.

[21] Vishwanath R Iyer, Christine E Horak, Charles S Scafe, David Botstein, Michael Snyder, and Patrick O Brown. Genomic binding sites of the yeast cell-cycle transcription factors sbf and mbf. *Nature*, 409(6819):533–538, 2001.

[22] Aleksander Jankowski, Ewa Szczurek, Ralf Jauch, Jerzy Tiuryn, and Shyam Prabhakar. Comprehensive prediction in 78 human cell lines reveals rigidity and compactness of transcription factor dimers. *Genome research*, 23(8):1307–1318, 2013.

[23] Hongkai Ji, Hui Jiang, Wenxiu Ma, David S Johnson, Richard M Myers, and Wing H Wong. An integrated software system for analyzing chip-chip and chip-seq data. *Nature biotechnology*, 26(11):1293–1300, 2008.

[24] Daxin Jiang, Chun Tang, and Aidong Zhang. Cluster analysis for gene expression data: A survey. *Knowledge and Data Engineering, IEEE Transactions on*, 16(11):1370–1386, 2004.

[25] Arttu Jolma and Jussi Taipale. Methods for analysis of transcription factor dna-binding specificity in vitro. In *A Handbook of Transcription Factors*, pages 155–173. Springer, 2011.

[26] Steven R Jordan and Carl O Pabo. Structure of the lambda complex at 2.5 a resolution: details of the repressor-operator interactions. *Science*, 242(4880):893–899, 1988.

[27] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. Genome-wide identification of in vivo protein–dna binding sites from chip-seq data. *Nucleic acids research*, 36(16):5221–5231, 2008.

[28] Peter V Kharchenko, Michael Y Tolstorukov, and Peter J Park. Design and analysis of chip-seq experiments for dna-binding proteins. *Nature biotechnology*, 26(12):1351–1359, 2008.

[29] Jennifer J Kohler and Alanna Schepartz. Effects of nucleic acids and polyanions on dimer formation and dna binding by bzip and bhlhzip transcription factors. *Bioorganic & medicinal chemistry*, 9(9):2435–2443, 2001.

[30] Teuvo Kohonen. Self-organization and associative memory. *Self-Organization and Associative Memory, 100 figs. XV, 312 pages.. Springer-Verlag Berlin Heidelberg New York. Also Springer Series in Information Sciences, volume 8*, 1, 1988.

[31] Mathieu Lajoie, Yu-Chih Hsu, Richard M Gronostajski, and Timothy L Bailey. An overlapping set of genes is regulated by both nfib and the glucocorticoid receptor during lung maturation. *BMC genomics*, 15(1):231, 2014.

[32] Stephen G Landt, Georgi K Marinov, Anshul Kundaje, Pouya Kheradpour, Florencia Pauli, Serafim Batzoglou, Bradley E Bernstein, Peter Bickel, James B Brown, Philip Cayting, et al. Chip-seq guidelines and practices of the encode and modencode consortia. *Genome research*, 22(9):1813–1831, 2012.

[33] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[34] Jason D Lieb, Xiaole Liu, David Botstein, and Patrick O Brown. Promoter-specific binding of rap1 revealed by genome-wide maps of protein–dna association. *Nature genetics*, 28(4):327–334, 2001.

[35] Wenxiu Ma and Wing Hung Wong. The analysis of chip-seq data. *Methods Enzymol*, 497:51–73, 2011.

[36] Anthony Mathelier, Xiaobei Zhao, Allen W Zhang, François Parcy, Rebecca Worsley-Hunt, David J Arenillas, Sorana Buchman, Chih-yu Chen, Alice Chou, Hans Ienasescu, et al. Jaspar 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic acids research*, page gkt997, 2013.

[37] Joanna A Miller and Jonathan Widom. Collaborative competition mechanism for gene activation in vivo. *Molecular and cellular biology*, 23(5):1623–1632, 2003.

[38] Shane Neph, Jeff Vierstra, Andrew B Stergachis, Alex P Reynolds, Eric Haugen, Benjamin Vernot, Robert E Thurman, Sam John, Richard Sandstrom, Audra K Johnson, et al. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature*, 489(7414):83–90, 2012.

[39] Carl O Pabo and Robert T Sauer. Transcription factors: structural families and principles of dna recognition. *Annual review of biochemistry*, 61(1):1053–1095, 1992.

[40] Peter J. Park. ChIP-seq: Advantages and challenges of a maturing technology. *Nature Reviews Genetics*, 10(10):669–680, oct 2009.

[41] Shirley Pepke, Barbara Wold, and Ali Mortazavi. Computation for chip-seq and rna-seq studies. *Nature methods*, 6:S22–S32, 2009.

[42] G P Redei. *Encyclopedia of Genetics, Genomics, Proteomics, and Informatics*. Springer, 2008.

[43] Bing Ren, François Robert, John J Wyrick, Oscar Aparicio, Ezra G Jennings, Itamar Simon, Julia Zeitlinger, Jörg Schreiber, Nancy Hannett, Elenita Kanin, et al. Genome-wide location and function of dna binding proteins. *Science*, 290(5500):2306–2309, 2000.

[44] Ho Sung Rhee and B Franklin Pugh. Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. *Cell*, 147(6):1408–1419, 2011.

[45] Richard I Sherwood, Tatsunori Hashimoto, Charles W O'Donnell, Sophia Lewis, Amira A Barkal, John Peter van Hoff, Vivek Karun, Tommi Jaakkola, and David K Gifford. Discovery of directional and nondirectional pioneer transcription factors by modeling dnase profile magnitude and shape. *Nature biotechnology*, 32(2):171–178, 2014.

[46] Trevor Siggers, Michael H Duyzend, Jessica Reddy, Sidra Khan, and Martha L Bulyk. Non-dna-binding cofactors enhance dna-binding specificity of a transcriptional regulatory complex. *Molecular systems biology*, 7(1), 2011.

[47] Matthew Slattery, Todd Riley, Peng Liu, Namiko Abe, Pilar Gomez-Alcala, Iris Dror, Tianyin Zhou, Remo Rohs, Barry Honig, Harmen J Bussemaker, et al. Cofactor binding evokes latent differences in dna binding specificity between hox proteins. *Cell*, 147(6):1270–1282, 2011.

[48] Christiana Spyrou, Rory Stark, Andy G Lynch, and Simon Tavaré. Bayespeak: Bayesian analysis of chip-seq data. *BMC bioinformatics*, 10(1):299, 2009.

[49] Rory Stark and Gordon Brown. Diffbind: differential binding analysis of chip-seq peak data. *R package version*, 100, 2011.

[50] Mary C Thomas and Cheng-Ming Chiang. The general transcription machinery and general cofactors. *Critical reviews in biochemistry and molecular biology*, 41(3):105–178, 2006.

[51] R. Wehrens and L.M.C. Buydens. Self- and super-organising maps in r: the kohonen package. *J. Stat. Softw.*, 21(5), 2007.

[52] Amy S Weinmann, Pearlly S Yan, Matthew J Oberley, Tim Hui-Ming Huang, and Peggy J Farnham. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and cpg island microarray analysis. *Genes & development*, 16(2):235–244, 2002.

[53] Matthew T Weirauch and TR Hughes. A catalogue of eukaryotic transcription factor types, their evolutionary origin, and species distribution. In *A Handbook of Transcription Factors*, pages 25–73. Springer, 2011.

[54] Tom Whitington, Martin C Frith, James Johnson, and Timothy L Bailey. Inferring transcription factor complexes from chip-seq data. *Nucleic acids research*, 39(15):e98–e98, 2011.

[55] Xugang Ye, Yi-Kuo Yu, and Stephen F Altschul. On the inference of dirichlet mixture priors for protein sequence comparison. *Journal of Computational Biology*, 18(8):941–954, 2011.