EXTRACTION AND IDENTIFICATION OF FREQUENT SEQUENTIAL PATTERNS IN TRANSCRIPTION FACTOR BINDING SITE ORGANIZATION OF ENHANCERS

BY

PAUL MARCOTTE BISSONNETTE

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Computer Science
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2014

Urbana, Illinois

Adviser:

Assistant Professor Jian Ma

Abstract

The advent of laboratory techniques to assess protein-DNA interactions, and chromatin post-translational modifications has created vast arrays of data correlating binding sites and regulatory elements with genome regions. There has been great interest in developing new computational approaches that leverage these annotations in order to better understand the layout and organization of eukaryotic genomes and gene regulation. In particular, while tools exist to accurately map the coding regions of the genome, few methods exist to leverage new annotations in the non-coding regions of the genome that may hold the key to many important questions about gene regulation. By exploiting the results of large scale annotation studies such as ENCODE to identify cell-type specific enhancers, and transcription factor binding sites, it is possible to better understand enhancer regions of the genome responsible for promoting transcriptional activity. Using frequent sequential pattern mining techniques from the classical data mining field, the importance of linear order in enhancer role and structure was explored in this thesis. Common orderings of binding sites within enhancers were identified across nine cell lines. Putative targets of enhancers exhibiting these patterns were then determined and clustered based on functional classification. Examination of detected patterns indicated that while the choice of transcription factors in the pattern often correlated with the overall function of putative targets, the ordering of these binding sites might be an effective classifier of more specific functional activity. Additionally, findings suggested that the arrangement of binding sites within enhancers was more likely to be cell specific than the transcription factor binding sites themselves. The knowledge that binding site pattern is more strongly linked to target function than binding site co-association alone could lead to important advances in our understanding of the diverse regulatory roles of different enhancers as well as new functional annotations for enhancers, genes, and transcription factors that have not yet been the focus of intensive studies to elucidate their roles.

Acknowledgments

This work was made possible by the kind support of many mentors and colleagues. In particular the guidance and good counsel of my advisor Dr. Jian Ma has been invaluable. Additionally, many of the data mining principles incorporated in this work would not have been possible without the help of Dr. Jiawei Han, and his wonderful data mining seminar. Support from Chao Zhang was critical in applying data mining techniques to biological systems. Yang Zhang provided datasets and helpful insight into genomics and regulatory networks.

Contents

1. Introduction

The study of regulatory genomic elements is an important and unresolved problem in genomic biology. It has been determined that more than 98% of the human genome does not directly code for proteins, and while the structure and function of coding regions that is better understood represent only a small fraction of the genome [1]. Understanding the nature of gene expression can lead to important breakthroughs in mechanisms of cancer and human disease. At the level of gene transcription, gene expression is controlled via transcription factor (TF) proteins that selectively bind to cis-regulatory elements in a sequence-specific manner. These proteins typically bind to specific DNA sequences within enhancer regions, which act on genes to promote or repress expression [2, 3]. Modern techniques in molecular biology have led to the wide availability of data from high-throughput mapping of the interaction between proteins and non-coding regions of the genome that harbor regulatory functions, as well as the epigenetic regulatory marks that guide the transcriptional and regulatory machinery in accessing the genome [4-6]. The wealth of data generated has rapidly outpaced available bioinformatics techniques to dissect and analyze the complex regulatory relationships between epigenetic marks, DNA sequences, and TF binding sites (TFBS). Tools to explore the binding of TF proteins responsible for regulating gene expression and mapping areas of regulatory significance have been in high demand to explore the structure and function of the vast non-coding regions of the genome [7-9]. The Encyclopedia of Genome Elements (ENCODE) project has been instrumental in systematically mapping the genome and disseminating the results of these genome-wide landscapes [5]. The mechanism by which TFs modulate gene expression has been of particularly great interest due to their ubiquitous role in transcriptional regulation.

Recent studies of these factors and their associated regions have focused on elucidating common combinations of TFs that tend to co-associate within the same regions of the genome [7-10]. These groupings can indicate common regulatory patterns in which TFs act

independently or by protein association to affect gene transcription. Common pairing of factors can indicate a protein-protein interaction that may be required to facilitate DNA binding and transcriptional regulation. Studies of these combinatorial patterns have utilized techniques in data mining for frequent itemset analysis to efficiently detect associations using experimental data from high-throughput assays [9]. Due to the rapidly evolving landscape of genomic mapping many of these studies have not incorporated protein DNA association data only recently made available by ENCODE. In addition only one published study has explored the linear arrangement of factors on the genome [11]. In this study the authors did not use ENCODE data to guide the selection of enhancer regions and binding sites, nor did they consider the potential significance of overlapping and interacting binding sites. In this thesis we perform a screen of several databases of publically available binding site and enhancer region information to examine order-aware combinatorial patterns of binding sites within regulatory regions.

1.1 Transcription Factor Binding Sites

Transcription factors (TFs) play a critical role in transcriptional regulation through cis-acting elements present near genes [2, 3, 12]. TFs are proteins capable of recognizing and associating with specific genomic sequences (motifs) [12, 13]. Once associated TFs can mediate the recruitment of important transcriptional proteins, other TFs, and members of the polymerase holoenzyme [3, 12]. A particular TF will recognize a variety of different sequences with varying efficiency. The motif recognized by a TF is generally represented using a position weight matrix (PWM) indicating its preference for particular nucleotides at each position in the transcription factor binding site (TFBS) [13]. Understanding the layout and organization of TFBS in the genome is a crucial step towards elucidating the complex regulatory mechanisms, which control gene expression and ultimately cellular function.

Tools for locating TFBS based on genomic sequences often score putative TFBS based on closeness to known motifs and return high scoring motifs [14]. More sensitive searches will often attempt to calculate p-values and incorporate estimated false discovery rates (FDRs) [15]. Ultimately these methods may fail to detect TFBS or output false TFBS because of topological or structural features of the surrounding genomic sequence not represented in sequence motifs. The most common screens for TFBS require chromatin immunoprecipitation (ChIP) of sequence bound TFs. Sequences are then ascertained through high throughput sequencing (ChIP-seq), or less sensitive array based detection [6, 16]. These techniques are more costly and time consuming than *in silico* identification but can positively determine which areas of the genome are bound by particular TFs. Several curated databases of TFBS motifs and experimentally confirmed TFBS locations exist including TRANSFAC and JASPAR, the latter is open access [17, 18]. The ENCODE project has been instrumental in performing and distributing high throughput screens of the genome for TFBS as well as more general protein association [5]. These ChIP-seq data have focused on identifying windows of high protein occupancy within the genome. The windows detected can be as large as 0.5-1kb, therefore, further processing is required to obtain base-pair resolution maps of protein binding. Additionally these association maps include proteins believed to not bind directly with DNA but rather with other proteins which may themselves preferentially associate with DNA motifs. Studies of these regions have been focused on either identifying new binding motifs in a *de novo* search of the genomic regions, or mapping existing motifs to these locations. This type of mapping is far more specific and reliable than *in silico* mapping of binding sites because of the high false positive rate of *in silico* mapping techniques [19, 20].

Occasionally TFBS will act on other nearby TFBS allowing a particular TF to control the presence of other TFs [12]. In the case of positive control a certain TF will act to strongly recruit other TFs increasing their binding efficiency to nearby sequences. Alternatively in the case of

overlapping TFBS only a single TF can be bound at a time, allowing for TFs to competitively inhibit each other through binding. Lastly if TFBS are too close together they can cause competitive inhibition without overlapping through steric interference. The distance required for this type of interference will be a function of the size and structure of the associated TFs. Identifying constructively and destructively interfering TFBS can add important details to our understanding of TFBS structure and function.

1.2 Enhancers

Many studies have demonstrated that TFBS that target particular genes are generally grouped together into cis-regulatory modules (CRMs) [21, 22]. These modules are generally 50 bp to 1.5 kb in length and there may be several that act on the same gene [21]. An enhancer is a CRM that when bound increases transcription of a target gene, and can be located either proximally or distally to the gene. Enhancers can be as close as just upstream of the promoter region or even within the introns of the target or so far away that they are closer to other genes that they do not regulate or located on different chromosomes [23, 24]. The ability of distal enhancers to act on target genes is due to a three-dimensional proximity of bound TFs to the transcriptional machinery located at the promoter region that can be impossible to capture in a linear sequence. Though tools exist to estimate three-dimensional genomic structure only experimental analysis can confirm structural proximity and more importantly regulatory relationship [6, 16].

A variety of approaches exist to locate enhancers *in silico* based on dense groupings of TFBS, phylogenetic sequence conservation, and proximity to genes or other annotations [25]. Studies have demonstrated that enhancers often exhibit discontinuous conservation where some regions within the enhancers are conserved but not others [11]. More recent advancements based on DNase hypersensitivity and chromatin modifications have shown promising results for detecting enhancer regions based solely on histone modification

information from ChIP-seq experiments. The ChromHMM software constructs a hidden Markov model (HMM) to classify regions of the genome based on histone modifications [26]. Prior work has shown that chromatin marks are an effective measure of cell-type specific enhancer activity [27, 28]. In mapping chromatin state across nine human cell types the histone modifications H3K4me3, H3K4me2, H3K4me1, H3K9ac, H3K27ac, H3K36me3, H4K20me1, H3K27me3, and H3K9me3 were considered for classifying chromatin regions [26]. H3K4 methylation and trimethylation were associated with enhancers and promoters respectively, while dimethylation was associated with both [26, 27, 29]. Both acetyl marks have been shown to correlate with active regulatory regions, while H3K27me3 is associated with Polycomb repressed regions [26, 27, 29]. Lastly, H3K36me3 and H4K20me1 are both associated with transcribed regions [26]. Annotations based on these classifications are included as a track in the UCSC genome browser. The organization and structure of enhancers is of great interest as advancements in this area may lead to better classifiers for enhancer regions and a more complete understanding of transcriptional regulation [22]. The annotations available were generated by the ENCODE project using high throughput ChIP-seq screens of multiple cell lines for a large variety of histone marks.

1.3 Mining Transcription Factor Binding Sites

While previous studies have considered the co-occurrence of TFBS in the same enhancer region, there have been few studies to analyze potentially common orderings of TFBS, and in particular the possibility of patterns containing overlapping TFBS. By applying techniques for FSPM using enhancers identified by ChromHMM, and TFBS from the JASPAR database common TFBS orderings can be identified without losing overlapping TFBS. Both TFBS verified in prior literature through ChIP-seq experiments and curated by JASPAR, and those identified by FIMO using motifs from JASPAR can be analyzed in this manner. Because overlapping TFBS can only be bound by a single TF at a time, and nearby TFBS will lead to

steric occlusion when bound by large TFs, nearby TFBS should be treated differently from more distal TFBS. Ultimately, in frequent sequences that contain sites that may occlude each other these sites should be placed in the same itemset to indicate that they occupy a single position available for binding. Using classic techniques from FSPM it is possible to mine sequences containing such sets of overlapping TFBS. By using enhancer regions detected based on chromatin modifications and allowing for the possibility of overlapping TFBS we provided a more complete analysis of the frequent patterns of TFBS present in the genome. Additionally, the use of experimentally validated TFBS greatly reduces the number of sites requiring processing while simultaneously increasing the accuracy of the results.

2. Literature Review

2.1 Frequent Sequential Pattern Mining

FSPM is a powerful analytical tool originally developed to research consumer purchasing habits [30]. FSPM algorithms process databases of transactions to find common transaction orderings (e.g. individuals who purchase printers will purchase ink in future transactions). Each record in the input database is an ordered sequence of transactions, where each transaction is an unordered set of items (itemset). Define a sequence, $x = a_1a_2a_3a_4a_5a_6 \cdots$ where a_t is a subset of A, the set of all items (alphabet). Then the sequence y = $b_1b_2b_3b_4b_5b_6\ldots$ is a subsequence of x if for every b_i and b_j with $i < j$, there exists a_n and a_m such that b_i is a subset of a_n, b_j is a subset of a_m, and $n < m$. For any sequence x, the number of sequences in the database that are a superset of x is the support of x. Any sequence with support greater than or equal to min_sup is a frequent sequence. Classic FSPM seeks to identify all frequent sequences for a particular database and min_sup (Table. 1). Many algorithms have been proposed for mining all possible frequent patterns from a database including the generalized sequential pattern (GSP) algorithm, the Apriori algorithm, and PrefixSpan [31-33]. GSP constructs all possible patterns in order of increasing length, selecting patterns of sufficient support at each step [33]. Apriori relies on the knowledge that a frequent pattern of length k will contain a frequent pattern of length k-1 to perform additional pruning of putative patterns [31]. As a further improvement, the PrefixSpan algorithm constructs a projected copy of the input database for each frequent pattern found to make selecting longer patterns easier [32]. Sequences in the projected database are all suffixes of sequences in the original database whose corresponding prefix is a superset of the current frequent pattern.

Sequence ID	Sequence
1	<a(bc)d>
2	<ab(cd)>
3	<a(bd)c>
4	<abc>

Table. 1 Example of a sequence database containing four frequent sequences. Sequences are denoted with angle brackets and itemsets are enclosed with parenthesis. With *min_sup* = 3 the patterns a, b, c, ab, ac, ad, bc, and abc are frequent. Additionally with *min_sup* = 2 abd is frequent, notice that abd is a subsequence of **1** and **2**.

For any sufficiently large database, the set of all frequent sequences will be massive, having an exponential upper bound based solely on the size of the alphabet. Many of the identified sequences will contribute little additional information to any meaningful analysis. FSPM can be restricted to the set of closed frequent sequences, or the set of maximally frequent sequences [30]. A frequent sequence, x, is said closed if there is no frequent sequence, y, such that x is a subset of y and the support of x is equal to the support of y. A frequent sequence is maximal if it is not a subsequence of any other frequent sequence regardless of any difference in support. The set of closed frequent sequences is a lossless compression as any subsequence of a frequent sequence is itself frequent, and no support information was lost. The set of maximal frequent sequences preserves information about all frequent sequences but support information is lost in the compression. Heuristic algorithms for more efficiently locating closed and maximal patterns have been previously described including CloSpan [30, 34]. Rather than attempting to find all closed patterns, CloSpan will find closed patterns with high likelihood by merging frequent patterns of increasing length to create new putative patterns. By combining existing patterns rather than extending patterns an element at a time this approach can much more rapidly arrive at very long closed patterns.

Additional requirements may be imposed on frequent sequences including maximum and minimum item distance [35-37]. In the case of maximum distance the definition of subsequences from above is altered to further require that $m - n \leq max\_dist$, and for minimum

distance $m - n \geq min\_dist$ is required. Rather than simply using sequence position a distance function $f_a(n, m)$ can be defined to compute item distance. In order to make pattern mining feasible the distance metric must preserve ordering, so if $n < m < k$ then $f_a(n, m) < f_a(n, k)$. Additionally the triangle inequality can be imposed so that $f_a(n, k) \leq f_a(n, m) + f_a(m, k)$. These additional requirements reduce the efficacy of CloSpan because the combined databases of two merged patterns will need to be pruned to remove sequences violating the maximum or minimum distance constraints [35].

2.2 Bio-sequence Mining

FSPM has been applied in a number of studies to detect frequent patterns of nucleotide and amino acid sequences [38-40]. In these studies the size of the alphabet was generally very small (e.g., A, T, C, and G for nucleotides), while the sequences were very large (potentially entire genomes). Classically FSPM is much more suited to short sequences over large alphabets, largely because of the heavy reliance on pruning inherent in most of the popular algorithms. With large sequences and small alphabets many shorter sequences are likely to have sufficient support to be considered frequent making the space required to perform mining prohibitive. In studies of TFBS and enhancers itemsets are large, and sequences are kept relatively short (0.5 – 2 kb).

Recent studies of enhancer regions have focused on the co-association of different TFBS across many enhancers [7-10]. These studies are attractive due to the interest in combinatorial effects of groups of TFs and the existence of multi-TF complexes that work together to promote transcription. The adaptation of frequent itemset mining (FIM) has proven particularly effective for this type of analysis [41]. FIM is a classic data-mining tool used in market analysis to detect items commonly purchased together in the same transaction by consumers. The support of a particular itemset is defined as the number of transactions containing the itemset, an itemset is frequent if it is support by at least min_sup transactions

(the minimum support). In these studies enhancers were treated as transactions and TFBS as items. Ha *et al.* [7] instead employs a pattern mining approach, but transformed input data so that patterns were lost and only association information was recovered. Only Teng *et al.* [9] used ChIP-seq peaks to map TFBS, and chromatin marks to predict enhancers. Morgan *et al.* relies on Patser [42] to predict TFBS based on motifs taken from TRANSFAC [43]. In their analysis, enhancers were contiguous windows of 100 bp, the high number of TFBS detected using Patser motivated this relatively small window choice. Similarly, Sun *et al.* [8] used Clover to detect TFBS from motifs in TRANSFAC. Rather than using a fixed window size, FIM was performed with a distance constraint requiring all items in an itemset to be close together [44]. In Ha *et al.* TFBS were selected using log scaled scores calculated from PWMs curated by JASPAR and TRANSFAC. Enhancer regions in this study were selected by searching for regions bounded by particular TFBS of interest. Teng *et al.* showed the most promising results based on their comparison with experimentally validated enhancer regions and known TF interactions. In their analysis rather than considering membership of each TFBS in an enhancer as a binary property they used a probabilistic model of membership. Additionally, the enhancers themselves were scored and a probability to assess the likelihood of an enhancer existing was incorporated into itemset scoring. Therefore, rather than searching for a minimum support of each itemset they computed the probability that an itemset had at least minimum support. Finally unlike previous studies, Teng *et al.* attempted to detect an appropriate minimum support for each itemset based on the frequency of each TFBS in the genome.

While FIM is a powerful tool for evaluating TF association, information about the linear configuration of TFBS is lost in these studies. Similar methods for market analysis exist to detect common patterns of serial purchases. These techniques for frequent sequential pattern mining (FSPM) have been shown previously to offer interesting insights into enhancer organization [11]. Cai *et al.* [11] attempted to detect enhancers using sequence conservation

and motifs from TRANSFAC. They focused their analysis on conserved domains within non-coding regions of the genome, and extracted putative enhancers that show common patterns of TFBS using FSPM. ChIP-seq peaks were used to validate *in silico* results from their study but were not incorporated into the analysis. In their study they chose to ignore overlapping and extremely close TFBS, always preferentially selecting the highest scoring binding site. All previous studies have taken this approach to overlapping binding sites despite the potentially interesting interactions between TFs in these regions.

3. Methods

In order to study the importance of binding site order in enhancer regions a modified version of PrefixSpan was designed and used to process enhancer and TFBS data derived from a variety of sources, extracted patterns were then analyzed for functional and regulatory significance (Fig. 1). Each dataset provided putative TFBS mapped to the genome. Sequences of TFBS were constructed for pattern mining by mapping TFBS to putative enhancer regions detected using the ChromHMM software [45, 46]. These sequences were then pre-processed to collapse multiple occluding TFBS into single sites. Each such site would form an itemset, where items were the constituent TFBS. The modified pattern mining performed enforced a minimum distance between itemsets in any frequent pattern. In addition itemset distance was calculated based on actual base-pair offsets of constituent TFBS. The algorithm was also modified to perform additional filtering, and to ultimately only extract maximal patterns. This was done both to remove redundant and extraneous patterns from the output and to ease the processing of large datasets. Following pattern mining the putative targets of each pattern were determined and clustered by functional annotation. These annotated clusters were then compared to regulatory annotations available for TFs in their associated patterns.

3.1 Extracting Enhancer Sequences

Enhancer regions used for analysis were selected with ChromHMM. Nine human cell lines were assayed for nine histone marks commonly associated with promoters, enhancers, and other functional regions [45, 46]. Of the states learned fifteen were selected as likely associated with biological functions and consistently present. Based on association with known annotations these regions were labeled for functional relevance. In our analysis, regions annotated as being either strong or weak enhancers were chosen. Sequences for FSPM were derived through mapping of TFBS to enhancers. Absolute base-pair offsets of TFBS were

included in the metadata for each sequence to differentiate between interfering and non-interfering TFBS.

The TFBS used in enhancer sequences were taken from JASPAR [17, 47]. The sequences were either from the curated set of binding sequences used to construct the JASPAR profiles, or from computationally identified binding sites using FIMO. The curated sequences in the JASPAR database were compiled from published ChIP-seq and SELEX experiments [17, 47]. The quality of each study was manually assessed and only sites determined to be functionally inactive were removed [47]. The JASPAR sequences were available as locations in the hg19 reference genome, in order to map these sequences to enhancers from ChromHMM, the liftOver tool was used to transform the coordinates onto the hg18 reference [17, 46, 48]. The TFBS found using FIMO [15] were selected with a p-value cutoff of 10^{-6} and matches with q-value greater than 0.1 were removed [45, 46]. Each identified TFBS was mapped to an enhancer discovered through ChromHMM and sequences were constructed based on the order of occurrence of TFBS in enhancers.

Additional TFBS/enhancer sequences were derived using ChIP-seq data generated by the ENCODE project [4, 5]. Putative TFBS were selected using FIMO and motifs available through JASPAR on peaks identified by ENCODE as having TF binding activity. False positive TFBS were selected via the same method restricted to random regions of the genome that did not intersect with any ENCODE peaks [20]. The p-value scores of these negative controls were used to construct cutoffs for the positive TFBS. Only putative binding sites with p-value greater than 99.9% of all negative sites selected were used. In addition, for each ENCODE peak only the highest scoring putative TFBS was kept. These sites were then intersected with the ChromHMM enhancer regions to construct a set of sequences for mining. This dataset has the advantage of *in vitro* verification and generally contains far fewer false-positive binding sites than the purely *in silico* datasets generated using FIMO and other motif finders.

In constructing input sequences for FSPM special care was taken to handle overlapping and poorly spaced TFBS within the same enhancer. Rather than eliminating TFBS or simply ignoring potential physical interaction between TF, the data was transformed to allow occluding TFBS to appear within the same itemset in the sequence. To construct each sequence an itemset was created for each TFBS in the order in which they appear in the enhancer. Into these itemsets any TFBS close enough to occlude the corresponding TFBS was also inserted. Lastly the base-pair offset of each associated TFBS was stored along with the itemsets used to construct the enhancer sequences. During FSPM the distance between two itemsets was then computed using the difference in base-pair offsets of the associated TFBS. By selecting an appropriate minimum distance, itemsets containing the same TFBS could be selectively excluded from any detected frequent sequences. Both the cutoff for occlusion and twice that cutoff were selected as minimum distances. In the former instance it was possible that the same TFBS might appear in two itemsets if it fell between two other TFBS that were themselves at least the minimum distance for occlusion apart. In the latter instance while it was impossible for the same TFBS to appear in two itemsets it was possible to miss a TFBS entirely if it fell too close to its neighbors while still falling outside the maximum distance for occlusion. In experiments little difference was seen between these two measures as most TFBS were outside the maximum occlusion distance.

3.2 FSPM: PrefixSpan

TFBS sequences were interrogated using a modified version of the PrefixSpan algorithm. PrefixSpan recursively constructs frequent patterns of increasing length employing a technique known as database projection to construct new copies of the sequence database containing only sequences prefixed by the current pattern [32]. Database projection occurs in linear time and can be done in tandem with computing the support for a new pattern. The projected database contains only the suffixes following the current pattern rather than the entire

sequences. Here a suffix is defined as the final itemset in the current pattern plus all those following it. In this way several suffixes from the same sequence can occur in the projected database if the current pattern occurs multiple times in the sequence (Table 2). Once a projected database has been constructed new patterns can be found quickly in a linear pass over the new database. The two methods of pattern extension employed by PrefixSpan are extension and expansion. Pattern extension occurs when a new itemset with a single item is appended to the current frequent pattern resulting in a new frequent pattern, one itemset longer than the current pattern. Pattern expansion adds a new item to the final itemset in the current pattern yielding a new frequent pattern of the same length. Items are selected for extension by counting the occurrence of items in the current projected database, ignoring those items in the first position of sequences. Items with frequency greater than or equal to the minimum support are candidates for pattern extension. Those items in the first position of sequences in the projected database are considered for pattern expansion. Items in this position occurring with frequency greater than or equal to the minimum support can be added to the final itemset and still result in a frequent sequence.

Sequence ID	**Sequence**
1	(bc)d
2	b(cd)
3	(bd)c
4	bc

Table. 2 The projected database for the prefix ab is given above. The database is a projection of the database from Table 1. Clearly with *min_sup* = 2 no new items are available for expansion, while d and c are both available for extension.

The PrefixSpan algorithm can be easily modified to discard the current pattern if any of the recursively generated super-patterns were frequent (with the same support), more efficiently identifying maximal (closed) patterns. These pruning techniques will not catch all sub-patterns and an additional pass over the final set of patterns will be required to remove sub-patterns

discovered before their super-patterns were identified if the set of only closed or maximal patterns is desired. Additionally, a maximum itemset distance constraint can be trivially enforced by terminating the linear search of patterns in the projected database once maximum item distance from the first itemset in the sequence has been achieved (the first itemset in each sequence is also the final itemset in the current pattern). In the same manner a minimum distance may be imposed by not counting items in itemsets at the beginning of each sequence until an itemset satisfying them minimum distance has been reached.

3.3 FSPM: Details

Enhancer sequences were loaded into a database in memory for processing by FSPM using a modified PrefixSpan. In the initial database each sequence was given a unique identifier and stored as an ordered list containing sets of TFBS. Each set was associated with a particular base-pair offset in the enhancer and contained all TFBS within interacting range of that offset. Prefix projection performed against the initial database during pattern mining was used to construct projected databases containing sub-sequences from the initial database of sequences. In projected databases it was possible for several subsequences of the same sequence to occur, however, the support of a particular pattern was based on the total number of sequences containing the pattern, without regard to the number of occurrences of the pattern within any particular sequence in the support. This was done because the primary use of the pattern support was to determine the number of different enhancers containing the pattern as a cutoff for pattern significance. During recursive pattern extension the current prefix and database is track in memory and new projections are constructed on for each prefix extension for use in further recursive calls.

The PrefixSpan implementation used in this study incorporated a minimum distance constraint and the pruning techniques for selecting maximal sequences during prefix extension/expansion. The prefixspan routine recursively extends a prefix, initially called with an

empty prefix (Algorithm. 3). To determine which items to choose to extend the sequence the subroutine locally_frequent_extend will linearly scan sequences in the database counting the number of sequences each item occurs in, where only occurrences at least min_dist from the start of the sequence are considered (Algorithm. 1). Items to use in sequence extension are selected by locally_frequent_expand, which will linearly scan the database counting the number of times each item occurs at the beginning of a sequence in the database (Algorithm. 1). Each item identified will be used to create a new prefix and project_extend or project_expand will be used to create a database projection for the new seed prefix and a recursive call will be made to prefixspan to continue the search (Algorithm. 2). Frequent sequences of TFBS as well as those enhancers containing such sequences are returned in the result set.

Recursive database projection can lead to exceedingly high memory footprints when processing large databases containing long sequences or many transcription factors. To reduce memory pressure database projection is modified to create new databases containing only offset values in the original database. These lightweight projections reuse the same primary database and can be rapidly constructed, as each entry requires only two integer values, a sequence identifier, and the offset from the beginning of the sequence. Furthermore, during the projection process it is possible for duplicate entries to be inserted into the database if multiple entries already exist corresponding to the same sequence. By ensuring uniqueness of database entries using unique keying the size of all recursively constructed projections and the time required to scan for projection can both be greatly decreased.

Following the completion of pattern mining the set of reported patterns was pruned to include only the maximal patterns. Sorting the patterns in order of decreasing length and then continually selecting the next pattern not already a sub-pattern of a selected pattern was used to collect the final set of patterns. Each pattern was written to a list of frequent patterns and bed files were constructed indicating the support of detected patterns. These files give the genome

coordinates of enhancer regions in the support of each frequent pattern annotated with the locations of binding sites within the aforementioned enhancer. Frequent patterns were used to compare TFBS combinations between cell lines and examine the potential significance of TF associations. Coordinates obtained from bed files were used to dissect the conservation of enhancer regions with frequent patterns, and similarity of regions containing the same frequent pattern.

3.4 Frequent Pattern Analysis

To analyze the patterns detected by FSPM, cell line specific patterns, and patterns common to multiple cell lines were selected. Patterns were limited to those sequences of TFBS containing at least three binding sites. To validate results patterns containing several unique binding sites were searched against existing literature. These searches were used to confirm that TFs known to associate were found in common patterns. Finally, patterns were analyzed to assess the importance of linear ordering of TFBS. In particular the set of detected patterns was scanned to see if multiple possible arrangements of the same pattern were seen following pattern mining of a particular cell line. These results as well as a summary analysis of the difference between the TFBS databases used were incorporated.

Patterns presenting across all cell lines and patterns containing at least four unique TFs were extracted to narrow the list of frequent patterns. A set of putative target genes for each frequent pattern was constructed from the genes nearest to the enhancers containing the pattern. For each pattern a review of prior literature about associated TFs revealed possible regulatory roles for the group of TFs. Using the Database for Annotation Visualization and Integrated Discovery (DAVID), putative targets were clustered by functional annotation [49, 50]. For each pattern the highest enriched functional clusters were compared with the documented regulatory roles of associated TFs.

This type of analysis was used to demonstrate that the patterns detected correlated to previously documented regulatory functions of their constituent TFs, lending credence to their functional relevance. In addition, functional clusters of different patterns containing the same set of TFs were compared to explore the relationship between TFBS order within enhancers and functional significance of genes regulated.

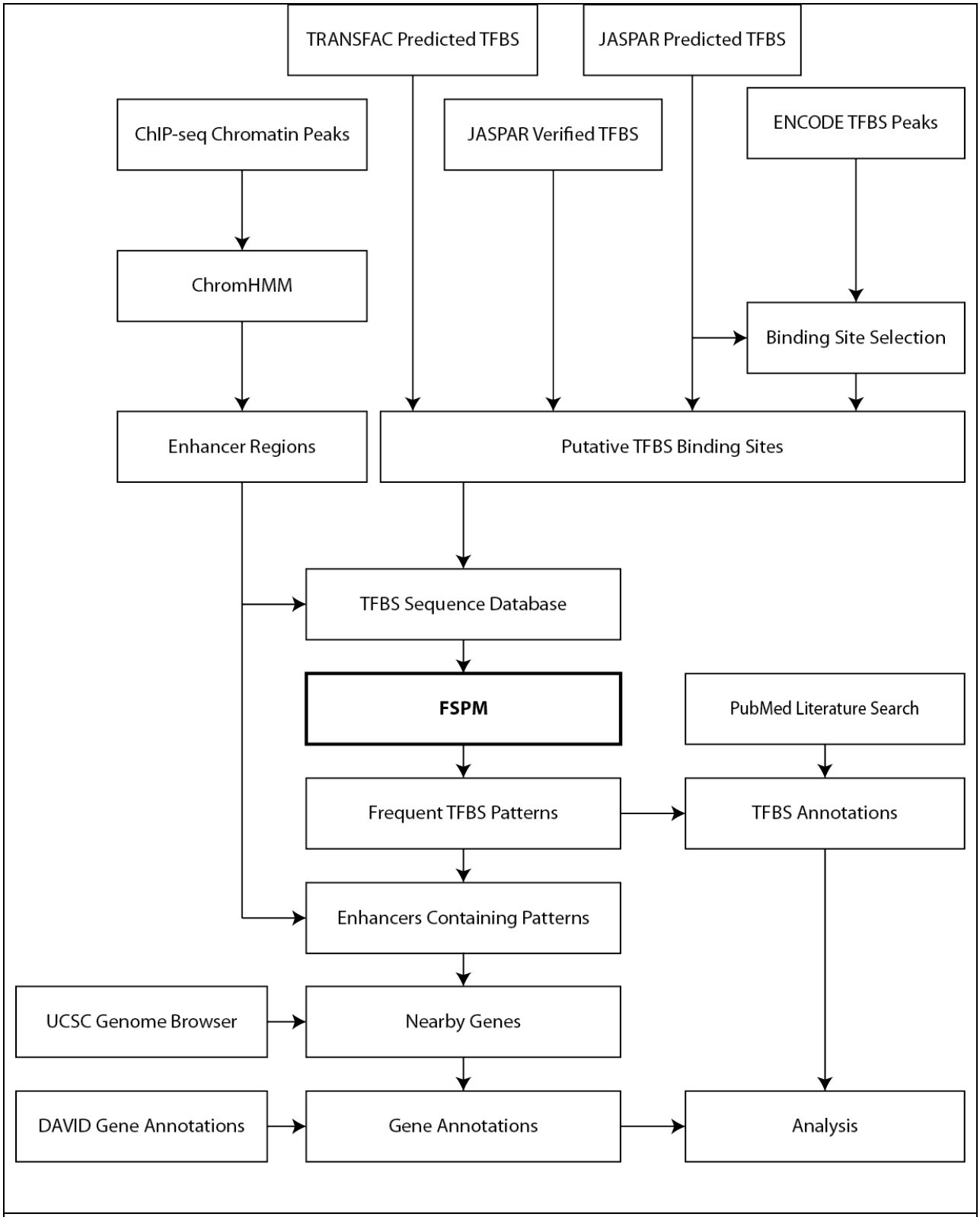

Fig 1. Analysis pipeline for dataset analysis, pattern mining, and annotation of frequent patterns and putative targets for regulatory and functional significance.

```
1:  Function locally_frequent_extend(db, min_dist, min_sup)
2:    Input: Database db
3:    Input: Integer min_dist
4:    Input: Integer min_sup
5:    Output: List items
6:
7:  item_counts = Map([item -> 0 for item in Alphabet])
8:  For seq in db:
9:    Integer curPos = bp_offset(seq[0])
10:   Set found_items = Set()
11:   For i from 1 to |seq|:
12:     If bp_offset(seq[i]) – curPos ≥ min_dist:
13:       found_items = found_items union Set(seq[i])
14:   For item in found_items:
15:     item_counts[item]++
16: Return item for item in item_counts if item_counts[item] > min_sup
17:
18: Function locally_frequent_expand(db, min_sup)
19:   Input: Database db
20:   Input: Integer min_sup
21:   Output: List items
22:
23: item_counts = Map([item -> 0 for item in Alphabet])
24: For seq in db:
25:   For item in seq[0]:
26:     item_counts[item]++
27: Return item for item in item_counts if item_counts[item] > min_sup
```

Listing. 1 Routines for finding frequent items to use in expanding or extending the current prefix. In both cases frequent items are counted once per sequence in the database. Due to the nature of the database projection routine multiple instances of the sequence can occur in the database so during projection a second check for support is performed. The locally_frequent_extend routine has been modified to add a minimum distance constraint.

```
1:  Function project_extend(prefix, db, min_dist, min_sup)
2:    Input: Sequence prefix
3:    Input: Database db
4:    Input: Integer min_dist
5:    Input: Integer min_sup
6:    Output: Database new_db or None
7:
8:  Database new_db = Database()
9:  Integer new_sup = 0
10: Set visited = Set()
11: TFBS suffix = prefix[-1]
12:
13: For seq in db:
14:   Integer curPos = bp_offset(seq[0])
15:   For i from 1 to |seq|:
16:     If bp_offset(seq[i]) – curPos ≥ min_dist and suffix in seq[i]:
17:       new_db[] = seq[i:]
18:       If name(seq) not in visited_sequences:
19:         new_sup++
20:         visited = visited union Set([name(seq)])
21:
22: If new_sup ≥ min_sup:
23:   Return new_db
24: Return None
25:
26: Function project_expand(prefix, db, min_sup)
27:   Input: Sequence prefix
28:   Input: Database db
29:   Input: Integer min_sup
30:   Output: Database new_db or None
31:
32: Database new_db = Database()
33: Integer new_sup = 0
34: TFBS suffix = prefix[-1]
35:
36: For seq in db:
37:   If suffix in seq[0]:
38:     new_db[] = copy(seq)
39:     new_sup++
40:
41: If new_sup ≥ min_sup:
42:   Return new_db
43: Return None
44:
```

Listing. 2 Routines for projecting the sequence database following extending or expanding the current seed prefix. In both cases the new support is computed in addition to constructing a new database. If the new database lacks sufficient support then no database is returned. The project_extend routine has been modified to implement a minimum distance constraint.

```
1:  Function prefixspan(prefix, db, min_dist, min_sup)
2:    Input: Sequence prefix
3:    Input: Database db
4:    Input: Integer min_dist
5:    Input: Integer min_sup
6:    Output: List freq_pat
7:    Output: List freq_pat_sup
8:
9:  patterns = List()
10: freq_items = locally_frequent_expand(db, min_sup)
11: found = False
12:
13: For item in freq_items:
14:   new_pat = copy(prefix)
15:   new_pat[-1].add(item)
16:   new_db  = project_expand(new_pat, db, min_sup)
17:   If new_db is None:
18:     Continue
19:   patterns.extend(prefixspan(new_db, k, new_pat))
20:   found = True
21:
22: freq_items = locally_frequent_extend(db, min_dist, min_sup)
23:
24: For item in freq_items:
25:   new_pat = copy(prefix)
26:   new_pat.append(Set([item]))
27:   new_db  = project_expand(new_pat, db, min_dist, min_sup)
28:   If new_db is None:
29:     Continue
30:   patterns.extend(prefixspan(new_db, k, new_pat))
31:   found = True
32:
33: If not found:
34:   patterns.append(prefix)
35: Return patterns
```

Listing. 3 Main prefixspan routine. Modified so that only patterns that cannot be extended are appended to the list of frequent patterns. Recursive calls build successively longer sequences.

4. Results

Frequent sequence data was extracted from sequences generated using JASPAR, and TRANSFAC binding sites as determined by FIMO. Additionally patterns were extracted from sequences constructed from experimentally verified JASPAR sites and JASPAR motifs mapped onto the ENCODE ChIP-seq peaks as previously described. The JASPAR data generated through FIMO and experimental validation were both found to be too sparse to detect a diverse set of frequent patterns with high support. TFBS identified using FIMO were made available on the UCSC genome browser by the authors of FIMO. Sites were filtered with a p-value cutoff of 10^{-6} and a maximum q-value of 0.1. In each case the binding sites were mapped to enhancer regions identified by the ENCODE project using data from multiple ChIP-seq datasets for different histone modifications. These regions were selected across nine different cell lines using the hg18 coordinate system. The cell lines used in this study were Gm12878, H1hesc, Hepg2, Hmec, Hsmm, Huvec, K562, Nhek, and Nhlf. There were an average of 602,240.67 regions in each cell line and of those 250,767.11 were enhancer regions.

4.1 JASPAR Sites

Sequences were constructed from JASPAR using two distinct datasets. A first set of TFBS was taken using the database of sequences used to construct motif PWMs for JASPAR. This set of sequences has been experimentally verified through either SELEX or ChIP-seq. While less likely to contain false positives than *in silico* binding sites this data set was also extremely small, lacking many common factors. The second set of factors was detected using FIMO to process the hg18 genome. Cutoffs were set as previously described and sites were mapped to ChromHMM enhancer regions. This analysis contained fewer unique sites, likely due to the relatively small size of the JASPAR database as compared to TRANSFAC. Processing of these TFBS revealed many short enhancer sequences and several very long ones. The average length of the enhancer sequences was two for every cell line in both cases.

In performing FSPM the minimum sequence support was chosen manually for each dataset. Values ranging from 10 to 30 were used for the JASPAR dataset as these sequences showed a high degree of heterogeneity, preventing larger values from identifying any meaningful patterns. Due to the reduced number of unique sites identified using FIMO sequences tended to be more similar allowing for support values between 25 and 500. In all cases sequences reported were restricted to maximal frequent patterns. Summary statistics were collected, and interesting patterns were extracted for more detailed analysis. In addition frequent patterns detected were checked for rearrangements to measure the importance of binding site order. To quantify the significance of ordering the frequent patterns detected were transformed into itemsets, removing all ordering information. Then each itemset was checked for the number of times it occurred as a pattern in the original set. In both the FIMO and the JASPAR datasets the many patterns were found to have very few rearrangements, indicating that the ordering of items in the pattern was significant. In particular for FIMO patterns between 36% and 100% of patterns were unique depending on support, and between 68% and 76% for JASPAR patterns.

4.2 TRANSFAC Sites

As in the case of JASPAR, FIMO was used to extract sites using the TRANSFAC data with the aforementioned cutoff values. The TRANSFAC database contains more TFBS motifs than the JASPAR database, leading to more complex patterns. Unlike JASPAR, reference site locations used in constructing motifs was not publically available for TRANSFAC so this analysis was omitted. Because the TRANSFAC database was more complete patterns were easier to detect, therefore a minimum support between 250 and 500 was selected. Once again good evidence was found for the significance of ordering with roughly 55% of patterns being unique.

4.3 ENCODE Sites

The results of high throughput ChIP-seq experiments performed to identify regions of high protein-DNA affinity were used to construct another dataset for pattern analysis. Using FIMO high scoring binding sites were selected within each ChIP-seq peak and mapped to enhancers identified previously by ChromHMM. These sequences were mined for frequent sequences using a minimum support between 150 and 250. Between 57% and 63% of patterns were found to be unique itemsets, demonstrating a preference for particular orderings.

4.4 Patterns Shared Across Cell Lines

As many prior studies have focused on common associations between TFBS using frequent itemset analysis the importance of pattern ordering was of great interest. By transforming frequent patterns into frequent itemsets it was possible to ascertain the number of patterns representing the same itemset as well as the effect of minimum support on both patterns and itemsets. By counting the number of itemsets and patterns observed in various numbers of cell lines it was possible to measure the importance of cell type in TFBS patterns. While both patterns and itemsets were found to often be cell type specific this was much more the case for patterns. The most strike example was observed in the JASPAR/FIMO dataset where most itemsets were found across all cell lines while the organization of those TFBS correlated strongly with the cell lines.

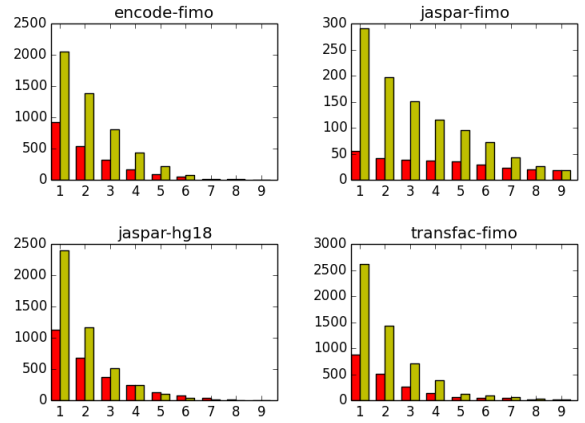

Fig 2. Graphs showing the number of patterns (yellow) or itemsets (red) found in multiple cell lines (x-axis). In both cases most patterns and itemsets are not seen in all cell lines, except FIMO/JASPAR where itemsets were shared between nearly all lines while patterns were far more specific to cell type.

The length and diversity of patterns detected was also compared with prevalence across multiple cell types. Pattern diversity was defined as the number of unique TFs participating in the pattern, and was measured by assessing the average size of frequent itemsets. While a pattern can be a series of repetitive binding sites for the same group of TFs the length of the corresponding itemset will be only the number of unique TFs in the pattern. From this analysis it was determined that in general patterns that were more common had less complexity, however, in FIMO/JASPAR and FIMO/TRANSFAC datasets showed similar levels of pattern complexity regardless of pattern frequency across cell lines.

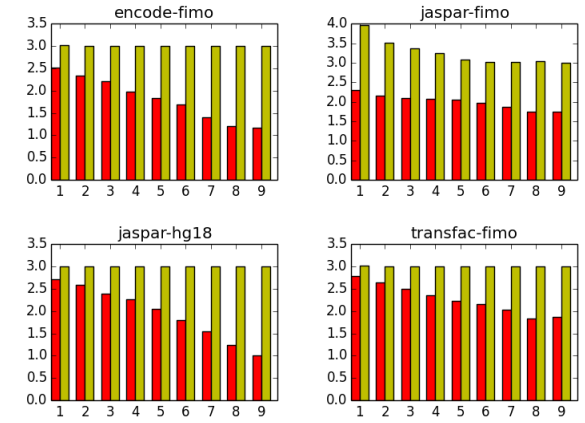

Fig 3. For each dataset the average length of of patterns (yellow) or itemsets (red) present across multiple cell lines was assessed. The plots show average length vs. number of cell lines. Interestingly only the FIMO/JASPAR and FIMO/TRANSFAC datasets showed a relatively consistent diversity of items regardless of pattern frequency.

4.5 Patterns Detected Across All Lines

For each dataset a list of patterns detected across all cell lines was collected. Enhancers containing these patterns were located and the nearest genes in both the upstream and downstream directions were extracted from the UCSC genome browser. Three sets of unique gene IDs were created for each pattern, nearest, upstream, and downstream. Each set was analyzed using DAVID to create a list of highly enriched functional annotations. An additional clustering analysis with DAVID was used to create subgroups for each pattern based on functional similarities. Many of the most common patterns were merely repetitive instances of the same binding site, however, four patterns from the FIMO/JASPAR dataset and two from the TRANSFAC dataset contained a unique set of TFs (Table 5). Interestingly, the patterns identified in TRANSFAC showed no reordering, while the JASPAR patterns were rearrangements of the same set of three TFs, suggesting that the grouping was important though not the order.

Patterns found in all cell lines were filtered to extract those patterns containing at least three unique TFBS. Six such patterns were found, four in the JASPAR/FIMO dataset and two in the TRANSFAC/FIMO dataset. The patterns identified through the TRANSFAC database contained the Sp1, Sp4, Muscle, and KROX TFBS. Both Sp1 and Sp4 are Kruppel-like factors and interact with DNA through a zinc-finger binding domain [51, 52]. The Sp1 TF can function as both a transcriptional activator and a repressor and has been implicated in regulating a variety of cellular events including growth, differentiation, apoptosis, chromatin remodeling, the immune response, and the response to DNA damage [52]. Sp4 has been found to interact with the E2F1 transcription factor, which has been identified as a mediator of both apoptosis and proliferation [53]. The KROX transcription factors play important roles in both development and terminal differentiation [53]. Clustering analysis of the first TRANSFAC pattern revealed an enrichment of targets with positive and negative transcriptional control, as well as actin binding,

cellular adhesion, and structural morphology (Table 7). The role of Sp1, and KROX in controlling development and differentiation correlates well with these cluster annotations. The second set of clusters was very similar to the first though both sets contained so many genes that there was a very diverse set of annotations.

The sets of patterns taken from JASPAR/FIMO were each rearrangements of RREB1, Sp1, and YGR067C. The RREB1 protein is a zinc-finger transcription factor that recognizes the RAS-response element (RRE) in the promoters of target proteins. This TF has been implicated in the Ras signal transduction pathway and may play a role in cellular adhesion and migration [54]. YGR067C is a putative TF found in Saccharomyces cerevisiae, its role in these patterns was not immediately clear from the available literature. Each of the clusters identified in an analysis of all four patterns identified enrichment of transcriptional regulation proteins both positive and negative (Table 7). There were also clusters implicating related proteins in a variety of roles ranging from cellular mobility to chromatin remodeling. As in the previous group the set of genes clustered as so large that the clusters spanned a wide variety of functional roles.

4.6 Patterns Containing at Least 4 TFs

In a second functional analysis, patterns were extracted based on length and associated genes were mapped for functional significance. Long patterns also displayed a high degree of repetitively, reflecting the importance of repetitive groupings of TFBS for strong binding potential. Interestingly a functional analysis of longer patterns demonstrated a stronger correlation between TFBS order and function than shorter more prevalent patterns. The patterns selected for this analysis contained at least four distinct binding sites and four unique TFs. The only three patterns detected meeting this criteria were from the TRANSFAC/FIMO dataset (Table 6). Unlike the more frequent patterns detected across multiple cell lines these longer patterns were only detected in one line.

Most datasets were found to be enriched with short and repetitive patterns, removing these yielded three patterns with at least four binding sites and at least four unique binding factors all from the TRANSFAC/FIMO dataset. All three patterns contained the TFBS Sp1, Sp4, KROX, and GC. The GC-box TFBS is a common motif in eukaryotic enhancers and acts as a promoter of gene expression [55]. The roles of these transcription factors suggest an overall significance of this set of patterns in controlling cellular differentiation and apoptosis. Clustering analysis by DAVID revealed 52 distinct clusters for the first pattern, the largest of which indicated transcriptional regulation, DNA binding, and nuclear localization annotations (Table 8). Proteins with these annotations will themselves serve as TFs potentially effecting broader cellular regulatory events. The next most significant cluster exhibited annotations for actin binding and cytoskeletal protein binding. These annotations would indicate that the proteins in this cluster are involved in constructing the type of rigid cellular structures common in terminally differentiated cells. An additional cluster showed enrichment of factors implicated in *Wnt* signaling and colorectal cancer. The *Wnt* pathway is plays an important role in terminal differentiation of colonocytes [56]. The second pattern showed enrichment for cellular junction and cytoskeletal proteins in addition to those in the first cluster. These types of proteins would also be important for terminal differentiation by increasing cellular adhesion and creating rigid cellular structures. The final pattern showed enrichment for the same clusters as the first pattern as well as a number of clusters specific to neuron development. Proteins in these clusters control neural cell differentiation, and morphology.

Cell Line	Total Regions	Enhancers
Gm12878	570,580	242,768
H1hesc	618,287	240,205
Hepg2	545,647	205,781
Hmec	608,568	291,155
Hsmm	638,307	272,987
Huvec	549,178	234,697
K562	621,678	249,997
Nhek	627,623	279,560
Nhlf	640,298	239,754
Average	**602,240.67**	**250,767.11**

Table 3. Enhancer regions detected using ChromHMM to process ChIP-seq data taken from ENCODE screens.

Dataset	Factors	TFBS	min_sup	Patterns	Itemsets	Sites	TFs
FIMO/	22.22	54,498	25	77.65	28.00	3.55	2.21
JASPAR			100	11.33	6.44	4.06	1.59
			500	1.11	1.11	3.83	1.00
JASPAR	66	78,720.89	10	46.00	28.00	3.30	2.65
			20	6.00	4.00	3.31	2.31
			30	2.00	2.00	3.35	2.15
FIMO/	70.56	101,995.78	250	168.56	89.44	3.00	2.37
TRANSAFAC			300	101.56	51.56	3.00	2.13
			500	19.89	10.89	3.00	2.23
FIMO/	71	106,668.22	150	124.78	71.33	3.01	2.20
ENCODE			200	68.78	42.00	3.00	2.00
			250	33.78	21.33	3.00	1.82

Table 4. For each dataset the number of TFs/TFBS were counted, and for varying min_sup values the average number of patterns, itemsets derived from those patterns, binding sites in each pattern, and TFs involved in each pattern were ascertained across all cell types.

ID	Sequence	Count	Genes	Dataset
1	<Muscle, KROX, Sp1>	3,295	1,557	TRANSFAC/FIMO
2	<KROX, SP4, Sp1>	3,771	1,786	TRANSFAC/FIMO
3	<RREB1, YGR067C, SP1>	432	292	JASPAR/FIMO
4	<YGR067C, SP1, RREB1>	531	240	JASPAR/FIMO
5	<YGR067C, RREB1, SP1>	425	220	JASPAR/FIMO
6	<RREB1, SP1, YGR067C>	414	271	JASPAR/FIMO

Table 5. Patterns found across all cell lines containing a unique set of TFBS and the number of times they were observed, and the number of unique putative targets identified.

ID	Sequence	Count	Genes	Dataset
1	<(GC, SP4), SP4, (SP4, KROX), Sp1>	204	1,557	TRANSFAC/FIMO
2	<(GC, Sp1, SP4), (Sp1, SP4), KROX, Sp1>	200	1,786	TRANSFAC/FIMO
3	<Sp1, Sp1, (SP4, GC, KROX), SP4>	204	292	TRANSFAC/FIMO

Table 6. Patterns found to contain at least 4 distinct TFBS.

ID	Score	Annotations
1	5.72	Poly-Ala (68), Poly-Ser (66), Poly-Pro (57), Poly-Gly (52)
1	3.67	Tetraspanin, subgroup (11), Tetraspanin, conserved site (11), Tetraspanin (11), CD9 antigen (9), 73.Integrins_and_other_cell-surface_receptors (5)
1	3.58	nucleus (412), regulation of transcription (309), DNA binding (258), transcription (254), Transcription (251), transcription regulation (246), non-membrane-bounded organelle (229), intracellular non-membrane-bounded organelle (229), dna-binding (218), regulation of RNA metabolic process (217)
2	7.07	Poly-Ala (79), Poly-Pro (77), Poly-Gly (60), Poly-Gln (27)
2	6.08	nucleus (504), regulation of transcription (379), transcription (318), Transcription (315), transcription regulation (309), DNA binding (301), non-membrane-bounded organelle (274), intracellular non-membrane-bounded organelle (274), regulation of RNA metabolic process (268), regulation of transcription, DNA-dependent (261)
2	3.57	sequence-specific DNA binding (99), activator (90), protein dimerization activity (73), Basic motif (36), Leucine-zipper (27), Basic-leucine zipper (bZIP) transcription factor (14), BRLZ (14), bZIP transcription factor, bZIP-1 (8), leucine zipper (6), Basic leucine zipper (4)
3	2.37	striated muscle cell differentiation (5), muscle cell differentiation (5), syncytium formation by plasma membrane fusion (4), syncytium formation (4), myotube differentiation (4), myoblast fusion (4)
3	1.94	Poly-Ser (17), Poly-Pro (17), Ser-rich (14), Poly-Ala (10), Poly-Gly (9)
3	1.46	Transcription factor jumonji/aspartyl beta-hydroxylase (3), Transcription factor jumonji, JmjN (3), Transcription factor jumonji (3), JmjN (3), JmjC (3)
4	2.11	Poly-Ala (14), Poly-Pro (12), Poly-Gly (9)
4	1.75	Pleckstrin homology-type (8), WH1 (3), EVH1 (3)
4	1.56	developmental protein (19), differentiation (14), neurogenesis (3)
5	2.77	cytoskeletal protein binding (19), protein domain specific binding (11), SH3 domain binding (5)
5	2.27	Pleckstrin homology-type (12), WH1 (3), EVH1 (3)
5	1.98	cytoskeleton (32), cytoskeletal protein binding (19), actin binding (15), actin-binding (14), regulation of organelle organization (5), regulation of cellular component size (5), regulation of protein polymerization (4), regulation of protein complex assembly (4), regulation of cytoskeleton organization (4), regulation of cellular component biogenesis (4)
6	2.08	Poly-Ser (19), Poly-Pro (14), Poly-Ala (12), Poly-Gly (9), Poly-Arg (7)
6	1.86	muscle cell differentiation (7), striated muscle cell differentiation (6), syncytium formation by plasma membrane fusion (3), syncytium formation (3), myotube differentiation (3), myoblast fusion (3)
6	1.68	cell junction (17), structural molecule activity (11), triple helix (4)

Table 7. Functional clusters for putative targets of patterns found across all cell lines.

ID	Score	Annotations
1	1.32	ank repeat (6), Ankyrin (6), ANK 3 (6), ANK 2 (6), ANK 1 (6), ANK (6), ANK 4 (5), ANK 5 (4), ANK 6 (3)
1	1.29	Pleckstrin homology (9), PH (9), PH (8), nucleoside-triphosphatase regulator activity (7), GTPase regulator activity (7), small GTPase regulator activity (6), regulation of small GTPase mediated signal transduction (6), regulation of Ras protein signal transduction (6), enzyme activator activity (5), guanyl-nucleotide exchange factor activity (4)
1	1.07	regulation of transcription (33), chromosome organization (6), chromatin regulator (5), chromatin organization (5), chromatin modification (5)
1	1.04	mutagenesis site (24), disease mutation (17), membrane organization (7)
1	1.02	vesicle-mediated transport (7), sh3 domain (5), Src homology-3 domain (5), SH3 (5), Variant SH3 (4), SH3 (4)
2	2.93	phosphoprotein (82), splice variant (69), alternative splicing (69)
2	1.46	cell junction (12), PDZ/DHR/GLGF (4), PDZ (4)
2	1.26	developmental protein (13), cell surface receptor linked signal transduction (10), Wnt signaling pathway (6), Pathways in cancer (6), Wnt receptor signaling pathway (4), wnt signaling pathway (3), beta-catenin binding (3), Colorectal cancer (3), Basal cell carcinoma (3)
2	1.12	negative regulation of signal transduction (4), negative regulation of cell communication (4), Regulator of G protein signalling (3), RGS (3)
2	1.11	cytoplasm (34), non-membrane-bounded organelle (30), intracellular non-membrane-bounded organelle (30), cytoskeleton (19), cytoskeletal part (15), microtubule cytoskeleton (8), cytoskeletal protein binding (8), cell projection (7), actin-binding (7), actin cytoskeleton (7)
3	1.33	cell surface receptor linked signal transduction (15), developmental protein (10), Pathways in cancer (7), Wnt signaling pathway (5), Wnt receptor signaling pathway (4), Colorectal cancer (4), wnt signaling pathway (3), beta-catenin binding (3), Basal cell carcinoma (3)
3	1.17	mutagenesis site (24), endomembrane system (10), membrane organization (6)
3	1.10	regulation of neurogenesis (5), regulation of nervous system development (5), regulation of cell development (5), regulation of neuron differentiation (4), regulation of neuron projection development (3), regulation of cell projection organization (3), regulation of cell morphogenesis involved in differentiation (3), regulation of cell morphogenesis (3), regulation of axonogenesis (3)
3	1.01	regulation of neurogenesis (5), regulation of nervous system development (5), regulation of cell development (5), positive regulation of developmental process (4), positive regulation of neurogenesis (3), positive regulation of cell differentiation (3), positive regulation of cell development (3)
3	0.91	nucleus (47), ion binding (32), cation binding (32), metal ion binding (31), regulation of transcription (30), transcription regulation (29), transcription (29), Transcription (29), DNA binding (25), metal-binding (24)

Table 8. Functional clusters for putative targets of patterns with at least four unique TFs.

5. Discussion

Patterns extracted from multiple datasets revealed interesting insights into the organization of enhancer regions and transcriptional regulation. Analysis of pattern ordering demonstrated the potential importance of linear sequencing to TFBS function. By incorporating ChIP-seq data into both the enhancer and TFBS identification stages the risk of detecting false enhancers or binding sites was greatly reduced. Interestingly frequent patterns were generally short, containing between three and four positions. Additionally many sequences contained repetitive binding sites for the same TF. In general increasing minimum pattern support had the effect of reducing the number of pattern rearrangements observed. Finally the most commonly observed patterns were those that were predominated by duplicate binding sites.

5.1 Functional Characterization

Functional characterization of putative enhancer targets was used to gain important insights into the potential regulatory functions of TFBS patterns detected. The patterns analyzed were selected manually from among the patterns occurring in all cell lines, and the lengthiest patterns extracted. Both the length and complexity of the patterns selected was of particular interest. As many previous studies focused on pairs of commonly associated TFBS we looked primarily to groups of at least three TFBS spaced sufficiently to make occlusion less likely. Annotation was performed using DAVID to construct clusters of functionally similar genes within each set of putative targets. Targets were selected based on proximity to the enhancer regions containing the pattern. While this type of selection is not ideal it is a commonly used approximation to ascertain the targets of an enhancer.

5.1.1 Patterns Across All Cell Lines

These patterns were extremely frequent appearing across all cell types and in many different enhancer regions in each type. The patterns themselves were short, containing at most

three sites, and only in a limited number of cases three distinct TFs. It was also apparent that ordering did not play an important role in the regulatory function of these patterns. Many orderings of the same set of binding sites appeared in the analysis and little difference was found in the putative targets of these reordered patterns. The value of the clustering analysis was impeded by the extremely large number of putative targets, which resulted in similar clusters among all identified patterns.

5.1.2 Longest Patterns

Unlike the set of extremely common patterns the functional groups for putative targets of long patterns were better differentiated. Average pattern length was reasonably well preserved across patterns regardless of frequency across cell lines, however, more frequent patterns tended to have fewer unique components. Perhaps the most interesting trait observed within these patters was that the differing arrangements of the observed TFBS correlated strongly with target protein function. In one pattern enrichment for actin-binding proteins was observed, while in the second a cluster was detected indicating the presence of cellular adhesion and structural proteins, and in the third, proteins were often important for neuronal cell development and morphology. In all three instances the roles of proteins targeted by these patterns fit the documented roles of the transcription factors well.

5.2 Pattern Rearrangement

While the JASPAR/FIMO dataset provided strong evidence that the itemsets detected in frequent mining are fairly immune to changes across cell types the other datasets showed a decreased number of itemsets observed in multiple cell types. Even in these cases itemsets were far more likely to be extant across cell lines than the frequent patterns they were derived from. From these observations we can conclude both that patterns are a more nuanced tool for both functional analysis and cell type differentiation.

5.3 Future Directions

There are many interesting future extensions to this project that could help further elucidate the role of binding site patterns in transcriptional regulation. While this study has focused on enhancers detected through ChIP-seq analysis new techniques including STARR-seq have shown great promise in detecting putative enhancers and assessing their strength [57]. In addition, new studies have identified *de novo* motifs from ChIP-seq peaks for TFBS made available through ENCODE [19]. These studies have created enormous datasets for which new pruning and noise filtering techniques will be required to perform efficient pattern analysis.

5.3.1 Integration of ChIP-seq/STARR-seq Peaks

Integration of raw ChIP-seq or STARR-seq data into the pattern-mining algorithm could provide a much more finely tuned search for putative binding patterns. This type of analysis could fuse enhancer and binding site identification with pattern analysis. Similar studies have used pattern analysis to search for new enhancer regions but have not leveraged the wealth of ChIP-seq data now available in selecting enhancers, or the far more powerful STARR-seq data in assessing the strength of putative enhancers. This combined approach could allow the pattern-mining process to guide the selection of enhancers and ultimately detect with greater accuracy those binding sites within the enhancer of greatest significance.

5.3.2 Probabilistic Pattern Mining

In Teng *et al.* the authors considered a probabilistic approach to itemset membership based on intensity of ChIP-seq peaks for enhancers, and closeness of putative binding sites to known motifs. Techniques exist for constructing frequent patterns using similar probabilistic measures for pattern membership and could be extended to this work. In particular it might also be possible to incorporate the intensity of ChIP-seq peaks for TFBS from the ENCODE dataset into the analysis. This type of probabilistic approach may better capture the thermodynamic

complexity of TFBS and enhancers. It may also better differentiate weak and strong signals, elucidating new patterns not yet observed.

References

1. Elgar, G. and T. Vavouri, *Tuning in to the signals: noncoding sequence conservation in vertebrate genomes.* Trends in Genetics. **24**(7): p. 344-352.
2. Latchman, D.S., *Transcription factors: An overview.* The International Journal of Biochemistry & Cell Biology, 1997. **29**(12): p. 1305-1312.
3. Ptashne, M. and A. Gann, *Transcriptional activation by recruitment.* Nature, 1997. **386**(6625): p. 569-577.
4. Consortium, E.P., *A user's guide to the encyclopedia of DNA elements (ENCODE).* PLoS Biol, 2011. **9**(4): p. e1001046.
5. Consortium, E.P., et al., *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
6. Johnson, D.S., et al., *Genome-Wide Mapping of in Vivo Protein-DNA Interactions.* Science, 2007. **316**(5830): p. 1497-1502.
7. Ha, N., M. Polychronidou, and I. Lohmann, *COPS: Detecting Co-Occurrence and Spatial Arrangement of Transcription Factor Binding Motifs in Genome-Wide Datasets.* PLoS ONE, 2012. **7**(12): p. e52055.
8. Sun, H., et al., *Unveiling combinatorial regulation through the combination of ChIP information and in silico cis-regulatory module detection.* Nucleic Acids Research, 2012. **40**(12): p. e90.
9. Teng, L., et al., *Discover context-specific combinatorial transcription factor interactions by integrating diverse ChIP-Seq data sets.* Nucleic Acids Research, 2013.
10. Morgan, X.C., et al., *Predicting combinatorial binding of transcription factors to regulatory elements in the human genome by association rule mining.* BMC Bioinformatics, 2007. **8**: p. 445.
11. Cai, X., et al., *Systematic identification of conserved motif modules in the human genome.* BMC Genomics, 2010. **11**: p. 567.
12. Mitchell, P. and R. Tjian, *Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins.* Science, 1989. **245**(4916): p. 371-378.
13. Stormo, G.D., et al., *Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli.* Nucleic Acids Research, 1982. **10**(9): p. 2997-3011.
14. Tompa, M., et al., *Assessing computational tools for the discovery of transcription factor binding sites.* Nat Biotech, 2005. **23**(1): p. 137-144.
15. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: Scanning for occurrences of a given motif.* Bioinformatics, 2011.
16. Collas, P., *The current state of chromatin immunoprecipitation.* Mol Biotechnol, 2010. **45**(1): p. 87-100.
17. Mathelier, A., et al., *JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles.* Nucleic Acids Research, 2013.
18. Matys, V., et al., *TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.* Nucleic Acids Res, 2006. **34**(Database issue): p. D108-10.
19. Kheradpour, P. and M. Kellis, *Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments.* Nucleic Acids Res, 2014. **42**(5): p. 2976-87.
20. Whitfield, T.W., et al., *Functional analysis of transcription factor binding sites in human promoters.* Genome Biol, 2012. **13**(9): p. R50.
21. Blackwood, E.M. and J.T. Kadonaga, *Going the Distance: A Current View of Enhancer Action.* Science, 1998. **281**(5373): p. 60-63.
22. Pennacchio, L.A., et al., *Enhancers: five essential questions.* Nat Rev Genet, 2013. **14**(4): p. 288-295.

23. Shashikant, C.S., et al., *Comparison of diverged Hoxc8 early enhancer activities reveals modification of regulatory interactions at conserved cis-acting elements.* J Exp Zool B Mol Dev Evol, 2007. **308**(3): p. 242-9.
24. Spilianakis, C.G., et al., *Interchromosomal associations between alternatively expressed loci.* Nature, 2005. **435**(7042): p. 637-645.
25. Hardison, R.C. and J. Taylor, *Genomic approaches towards finding cis-regulatory modules in animals.* Nat Rev Genet, 2012. **13**(7): p. 469-83.
26. Ernst, J. and M. Kellis, *ChromHMM: automating chromatin-state discovery and characterization.* Nat Meth, 2012. **9**(3): p. 215-216.
27. Heintzman, N.D., et al., *Histone modifications at human enhancers reflect global cell-type-specific gene expression.* Nature, 2009. **459**(7243): p. 108-112.
28. Visel, A., et al., *ChIP-seq accurately predicts tissue-specific activity of enhancers.* Nature, 2009. **457**(7231): p. 854-8.
29. Heintzman, N.D., et al., *Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome.* Nat Genet, 2007. **39**(3): p. 311-318.
30. Mabroukeh, N.R. and C.I. Ezeife, *A taxonomy of sequential pattern mining algorithms.* ACM Comput. Surv., 2010. **43**(1): p. 1-41.
31. Agrawal, R. and R. Srikant, *Fast Algorithms for Mining Association Rules in Large Databases*, in *Proceedings of the 20th International Conference on Very Large Data Bases*. 1994, Morgan Kaufmann Publishers Inc. p. 487-499.
32. Jian, P., et al. *PrefixSpan,: mining sequential patterns efficiently by prefix-projected pattern growth*. in *Data Engineering, 2001. Proceedings. 17th International Conference on*. 2001.
33. Srikant, R. and R. Agrawal, *Mining Sequential Patterns: Generalizations and Performance Improvements*, in *Proceedings of the 5th International Conference on Extending Database Technology: Advances in Database Technology*. 1996, Springer-Verlag. p. 3-17.
34. Yan, X., J. Han, and R. Afshar, *CloSpan: Mining: Closed Sequential Patterns in Large Datasets*, in *Proceedings of the 2003 SIAM International Conference on Data Mining*. p. 166-177.
35. Ji, X., J. Bailey, and G. Dong, *Mining minimal distinguishing subsequence patterns with gap constraints.* Knowledge and Information Systems, 2007. **11**(3): p. 259-286.
36. Méger, N. and C. Rigotti, *Constraint-based mining of episode rules and optimal window sizes*, in *Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases*. 2004, Springer-Verlag New York, Inc.: Pisa, Italy. p. 313-324.
37. Zaki, M.J., *Sequence mining in categorical domains: incorporating constraints*, in *Proceedings of the ninth international conference on Information and knowledge management*. 2000, ACM: McLean, Virginia, USA. p. 422-429.
38. Brazma, A., et al., *Approaches to the automatic discovery of patterns in biosequences.* J Comput Biol, 1998. **5**(2): p. 279-305.
39. Liu, W. and L. Chen, *An Efficient and Fast Algorithm for Mining Frequent Patterns on Multiple Biosequences*, in *Computer and Computing Technologies in Agriculture IV*, D. Li, Y. Liu, and Y. Chen, Editors. 2011, Springer Berlin Heidelberg. p. 178-194.
40. Wang, K., Y. Xu, and J.X. Yu, *Scalable sequential pattern mining for biological sequences*, in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*. 2004, ACM: Washington, D.C., USA. p. 178-187.
41. Agrawal, R., T. Imieliński, and A. Swami, *Mining association rules between sets of items in large databases.* SIGMOD Rec., 1993. **22**(2): p. 207-216.

42. Hertz, G.Z. and G.D. Stormo, *Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.* Bioinformatics, 1999. **15**(7-8): p. 563-77.
43. Hertz, G.Z., G.W. Hartzell, and G.D. Stormo, *Identification of consensus patterns in unaligned DNA sequences known to be functionally related.* Computer applications in the biosciences : CABIOS, 1990. **6**(2): p. 81-92.
44. Frith, M.C., et al., *Detection of functional DNA motifs via statistical over□representation.* Nucleic Acids Research, 2004. **32**(4): p. 1372-1381.
45. Ernst, J. and M. Kellis, *Discovery and characterization of chromatin states for systematic annotation of the human genome.* Nat Biotech, 2010. **28**(8): p. 817-825.
46. Ernst, J., et al., *Mapping and analysis of chromatin state dynamics in nine human cell types.* Nature, 2011. **473**(7345): p. 43-49.
47. Sandelin, A., et al., *JASPAR: an open□access database for eukaryotic transcription factor binding profiles.* Nucleic Acids Research, 2004. **32**(suppl 1): p. D91-D94.
48. Kuhn, R.M., D. Haussler, and W.J. Kent, *The UCSC genome browser and associated tools.* Briefings in Bioinformatics, 2013. **14**(2): p. 144-161.
49. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources.* Nat Protoc, 2009. **4**(1): p. 44-57.
50. Huang da, W., B.T. Sherman, and R.A. Lempicki, *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.* Nucleic Acids Res, 2009. **37**(1): p. 1-13.
51. Black, A.R., J.D. Black, and J. Azizkhan-Clifford, *Sp1 and kruppel-like factor family of transcription factors in cell growth regulation and cancer.* J Cell Physiol, 2001. **188**(2): p. 143-60.
52. Kaczynski, J., T. Cook, and R. Urrutia, *Sp1- and Kruppel-like transcription factors.* Genome Biol, 2003. **4**(2): p. 206.
53. Khan, S., et al., *Role of specificity protein transcription factors in estrogen-induced gene expression in MCF-7 breast cancer cells.* J Mol Endocrinol, 2007. **39**(4): p. 289-304.
54. Melani, M., et al., *Regulation of cell adhesion and collective cell migration by hindsight and its human homolog RREB1.* Curr Biol, 2008. **18**(7): p. 532-7.
55. Blake, M.C., et al., *Transcriptional initiation is controlled by upstream GC-box interactions in a TATAA-less promoter.* Mol Cell Biol, 1990. **10**(12): p. 6632-41.
56. Andreu, P., et al., *A genetic study of the role of the Wnt/beta-catenin signalling in Paneth cell differentiation.* Dev Biol, 2008. **324**(2): p. 288-96.
57. Arnold, C.D., et al., *Genome-wide quantitative enhancer activity maps identified by STARR-seq.* Science, 2013. **339**(6123): p. 1074-7.