Claude Fiifi Hayford

A Computational Analysis of Motif-Negative VDR-DNA Interaction and Possible Explanations of Its Mechanism of Interaction

Master’s Thesis in Molecular Medicine

Trondheim, June 2014

Supervisor: Professor Finn Drabløs

Subject Supervisor: Kjetil Klepper

Norwegian University of Science and Technology
Faculty of Medicine
Department of Cancer Research and Molecular Medicine

NTNU – Trondheim
Norwegian University of
Science and Technology

Abstract

Transcription factor binding to DNA has generally been assumed to be as a result of the recognition of sequence-specific motifs in the transcription factor binding sites. Recent studies have however shown several examples of transcription factor binding where no recognizable motif was identified. The exact mechanisms by which these associations occur remain unclear, although several explanations have been put forward. By employing MotifLab, a computational tool for the analysis of regulatory regions and data from the ENCODE project, this study examines the properties of these motif-positive and motif-negative regions and how they relate or differ from each other to get a better understanding of the factors that lead to transcription factor binding in these cases. Two well-described Vitamin D receptor datasets where there is a mixture of binding sites both with and without a clear motif is utilised.

The results showed that there are differences between motif-negative regions and motif-positive regions in terms of DNA accessibility, the type of regions in which either type of binding takes place, as well as in the types of motifs that are overrepresented in each case. These findings suggest that VDR binding in motif-negative regions does employ a mechanism different from its sequence-specific binding; however a clear elucidation of this mechanism has not been possible.

Acknowledgement

I would like to thank my supervisor, Professor Finn Drabløs for giving me the opportunity to work on this interesting project, for the guidance, feedback and insightful discussions during the course of the project and also for the motivation on occasions when I felt I had hit a brick wall.

I would also like to extend my sincere appreciation to Kjetil Klepper, my co-supervisor for all the help with respect to using MotifLab. Thank you for taking the time to answer my questions. I know I was not always clear what I meant but you were patient and tried to help as best as you could.

Finally, I want to thank all those who in one way or the other believed and supported me through the whole master’s program. To my family and friends I say a big thank you for the unflinching support.

Table of Contents

Abstract iii

Acknowledgement v

List of abbreviations and labels ix

1. Introduction 1

1.1. Genes and gene regulation 1

1.2. The role of epigenetic modifications in gene regulation 2

1.3. Transcription factors and their role in gene regulation 3

1.4. TF’s and cis-regulatory modules 4

1.5. Identification of TF binding sites 4

1.5.1. Systematic Evolution of Ligand by Exponential Enrichment (SELEX) 5

1.5.2. DNAse I footprinting 5

1.5.3. Chromatin Immunoprecipitation (ChIP) 7

1.5.4. Computational analysis of TF binding sites 8

1.5.5. Use of additional information to guide binding site prediction and motif discovery 11

1.6. Workbenches for analysing biological data 11

1.7. Gene Regulation by the Vitamin D Receptor (VDR) 12

1.8. Context and aims of the study 15

2. Methods 19

3. Results and Discussion 29

3.1. VDR binds to an *RXR-VDR*-like motif 29

3.2. VDR binds in HOT (High Occupancy of Transcription Factor) regions 33

3.3. Differences in enriched DNA feature regions in VDR- positive versus negative sequences 37

3.4. DNA feature regions not discriminative for VDR/VDR-like binding sites 40

3.5. Enrichment of *Sp1*-like motif in VDR-negative sequences 41

3.6. VDR- positive and negative sequences have differences in enriched motifs 43

3.7. Weaker VDR binding may be supported by other transcription factor activity/response elements 44

3.8. VDR binding and cluster formation is Cohesin and CTCF independent 46

3.9. VDR ChIP-Seq regions not enriched for direct interaction partners 48

3.10. Identification of putative CRMs modulating VDR action 49

3.11. Motif position distribution 50

4. Conclusion 55

4.1. MotifLab as a workbench for regulatory region analysis 55

4.2. Suggestions for future work 56

References 59

Appendix 65

List of abbreviations and labels

CCDS	Human genome high confidence gene annotations from the Consensus Coding Sequence project
ChIP-Seq	Chromatin Immunoprecipitation followed by high throughput sequencing
Conservation	Evolutionary conservation information from multiple alignment of 44 vertebrate species
CpG Islands	Genomic regions where CpGs are present at significantly higher levels than the rest of the genome as a whole
CRM	Cis regulatory module
DNaseHS_hotspots	DNAse 1 hypersensitivity hotspots from several cell lines indicating general chromatin accessibility and active cis-regulatory sequences
DR3	Direct repeat with a spacer of three nucleotides
ENCODE	Encyclopaedia of DNA elements
Enriched_Heikkinen	Statistically enriched motifs in Heikkinen dataset
Enriched_Ramagopalan	Statistically enriched motifs in Ramagopalan dataset
EnsemblGenes	Gene predictions generated by Ensembl
FAIRE-Seq	Nucleosome depleted regions of the genome identified using Formaldehyde Assisted Isolation of DNA Regulatory Elements
GM12878	Cell line derived from B-lymphocyte from International HapMap Project (similar to that used by Ramagopalan)
H3K27ac	Acetylation of histone 3 on lysine 27
H3K4me1	Mono-methylation of histone 3 on lysine 4
H3K4me3	Tri-methylation of histone 3 on lysine 4
H3K9ac	Acetylation of histone 3 on lysine 9
JSVDR (Classic VDR)	Jaspar_Core representation of the VDR motif
JSVDR_positive	Collection of sequences bearing one or more instances of the JSVDR
K562	Cell line derived from patient with chronic myelogenous leukaemia (similar to that used by Heikkinen)
MEME	Multiple Expectation Maximization for Motif Elicitation
MEMEVDR	*RXR-VDR* motif representation derived from *de novo* motif discovery using MEME
NR4A2_positive	Collection of sequences bearing one or more instances of the motif for *NR4A2*
PSSM	Position specific scoring matrix
PWM	Position weight matrix
Repeat327	Repeat regions detected by Repeatmasker 3.2.7
RXR	Retinoid X receptor
SINE	Small interspersed nuclear element

TF	Transcription factor
TFBS	Transcription factor binding site
TFBS_ChIP-Seq	General regions bound by transcription factors identified by ChIP-Seq
TFBS_observed	Motifs identified from motif scanning
TSS	Transcription start site
VDR	Vitamin D receptor
VDR_negative	Sequences without any *RXR-VDR* or VDR-like motif
VDR_positive	Sequences bearing either *RXR-VDR* or *NR4A2* motif
VDRE	Vitamin D response element

1. Introduction

1.1. Genes and gene regulation

Genes are regions of nucleic acids (DNA) in the genome said to be the basic units of heredity in living organisms that direct the manufacture of specific molecular products (RNA or proteins). These products are essential for the processes in the cell that contribute to growth and development of living organisms. In eukaryotes, the structure of the gene is made complex by the fact that a single gene may be responsible for several gene products. The average eukaryotic gene comprises exons and introns which play different roles. The exons are the regions of the gene that encode the product and the selective removal of the introns (non-coding part of the gene) in posttranscriptional processing such as splicing result in the different molecular end products. Interspersed between these regions of information are also stretches of regulatory regions that afford control over when and how these genes are transcribed.

The interaction of these regulatory DNA regions and regulatory proteins known as transcription factors (TF) constitute one of the mechanisms responsible for the expression and repression of genes. These transcription factors achieve their effect by identifying a specific stretch of nucleotides in the DNA and by hydrogen bonds and Van der Waals forces interact with the DNA. Transcription factors work in a combinatorial manner in which the control of single gene falls under the purview of several TFs acting in concert. This complexity allows for a small number of TFs to regulate a large number of genes and provides for a more smooth regulation. These TFs bind at the regulatory regions of these genes which usually are located at the start of the gene in locations termed as core promoters. These core promoters may be comprised of a TATA-box or TATA-binding region and an initiator region (INR) upstream and proximal to the translation start site (TSS). Other regulatory regions are found in introns, exons and non-coding DNA regions. These are usually termed distal regulatory regions and comprise enhancers, silencers and other modules that are distal to the TSS (1, 2). The function of these regulatory regions is usually dependent on the TFs which bind thus allowing them to act as either activators or repressors of transcription.

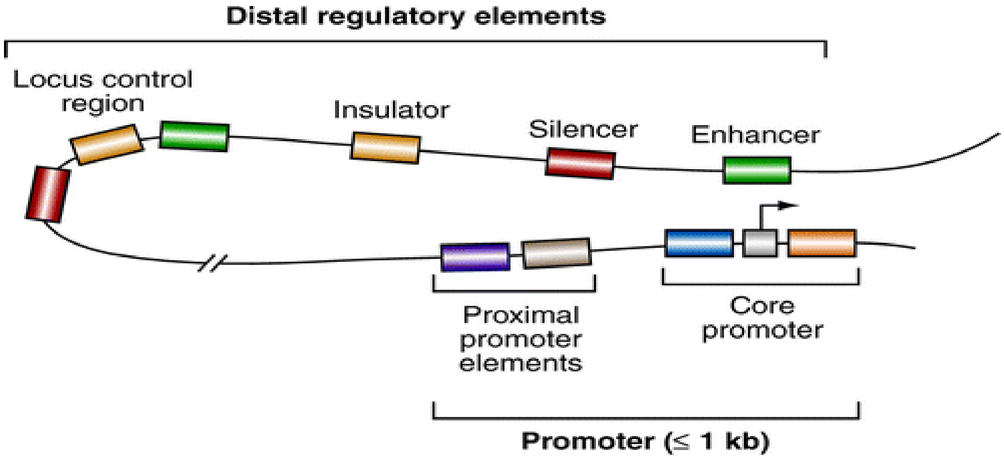


Figure 1 Different types of gene regulatory regions involved in the control of gene expression as shown in (2)

Binding of these factors to the regulatory elements causes the recruitment of other molecules such as members of the DNA polymerase family, histone modifiers and co-activators which form multi-unit complexes responsible for the actual transcription process.

Aside the proteins which interact with DNA, several other factors influence the regulation of these genes. The chromatin state has huge bearing on which genes are available for transcription at any given time (3). The formation of heterochromatin, a dense packaging of DNA and histones into nucleosomes restricts access to DNA and gene regulatory regions and this also contributes to the regulation process as does the presence of euchromatin; a more open region featuring active genes. Epigenetic modifications such as DNA methylation and the different types of histone modifications influence the chromatin state and thus gene regulation in several contexts, be it tissue, cell or even species (4).

1.2. The role of epigenetic modifications in gene regulation

Epigenetic modifications are mechanisms that change patterns of gene activity but do not involve a change in the underlying DNA sequence (5). These modifications mainly comprise DNA methylation and histone modifications. The histone modifications have several roles in the genome ranging from modifying the structure of chromatin to regulating the binding of chromatin factors (3, 4). The predominant modifications are the addition of single to multiple methyl (methylation) groups to lysines and arginines (e.g. H3K4me3, H3K4me1 and H3K27me3) and acetyl (acetylation) to lysines (e.g. H3K27ac and H3K9ac) on histones. DNA has a net negative charge whereas these histone proteins tend to have a positive charge. There exist therefore electrostatic forces of attraction that tends to wrap the DNA tightly around the histone cores in the nucleosome. This tightly packaged region of the genome forms the heterochromatin or closed chromatin. Histone modifications such as acetylation of lysine by histone acetyltransferases (HATs) reduces the positive charge on the histones and this decreases the electrostatic forces between the histone and DNA, leading to a less tightly bound structure. DNA can then be easily assessed by proteins and transcription factors that bind to DNA. These acetylation modifications are reversed by histone deacetylases (HDACs) restoring the charge and thus tightening the chromatin structure (6). Methylation on the other hand does not alter the charge on the histone protein. The presence of these methylated marks on the histone drives the interaction of several chromatin-associated factors with chromatin. These chromatin associated factors bear domains that are able to recognise the methylations and bind to them (3). By binding, they recruit other factors that are responsible for the chromatin remodelling process such as the previously mentioned HATs and HDACs. By influencing the chromatin state of the genome, these modifications are able to control which genomic regions are accessible to the transcription machinery for gene regulation.

1.3. Transcription factors and their role in gene regulation

Transcription factors are a group of protein molecules that act to regulate gene expression either positively or negatively. They have modular functional domains which function in different ways allowing for the regulation of the transcription factor itself and also of the gene they regulate (7, 8). Regulation of TFs is either by ligands which bind or interact with a part of the TF known as the ligand-binding domain or by interaction with other TFs and proteins through the activation function domain (8, 9). Extra regulation of these TFs is achieved by post-translational modifications such as phosphorylation, ubiquitinylation and a host of others which act to activate, suppress or even mark them for destruction (10). One other important part of the TF is the DNA-binding domain (DBD) with which it interacts with DNA. The identification of the response element, a specific string of nucleotides which serves as a template/motif for DNA interaction by the DBD forms the basis of most sequence-specific TF activity (9). TFs may act individually as monomers or in combination with itself or other TFs as mono- or hetero- dimers in binding DNA (11, 12).

The amino composition of the DNA-binding domain allows for some specificity in the regions of DNA these TFs recognize and bind to. These TFs may bind proximally or distally to the regulated region and through several mechanisms exert their effects. In cases where they bind in regions which are distal to the genes they regulate, DNA looping has been suggested to be a mechanism with which these distally located TFs are able to interact with their target genes (13-15). Even in cases where they are located proximally, interaction with the transcription machinery may also involve the activity of co-activators or co-repressors which act as a bridge.

1.4. TF's and cis-regulatory modules

Cis-regulatory modules (CRMs) are long stretches of DNA ranging from some 50-1500bp long that contain the binding sites for several different transcription factors (16). The TF's with binding sites in the CRM allow for the combinatorial control over target genes depending on the levels of the individual TF's and cofactors present at/in a specific time/space. Extra control is also achieved by different CRMs acting together on a single gene. CRMs are believed to exert their influence over their target genes via three main mechanisms the first of which is the looping of DNA to interact with the promoter after TF binding in the CRM (DNA looping model). The second mechanism involves the assembly of TF's and its cofactors at the CRM and subsequent scanning of DNA by the formed complex until it finds the general transcription machinery (DNA scanning model). The third mechanism is thought to be a combination of the other two mechanisms and is termed the facilitated tracking model. In this, assembly of the TF's and their cofactors takes place at the CRM as previously described and although this complex scans the intervening DNA sequence for the promoter, this is done in small steps with the complex still bound to the CRM. The initial step creates a small loop which increases in size as scanning continues until the target promoter is found (16, 17).

1.5. Identification of TF binding sites

The importance of TFs and their role in gene regulation necessitates the ability to recognise the DNA regulatory regions to which they bind. The process of identifying protein-DNA interaction sites are based on experimental and *in silico* methods. The process can begin with either one or the other depending on the information available at the time. These methods tend to be complementary with one serving to augment and help refine the results of the other (18). Several methods exist to determine the binding locations of proteins and transcription factors in the laboratory. Some common methods applied are DNA mobility shift assay, DNAse I footprinting assay, SELEX (Systematic Evolution of Ligand by Exponential Enrichment) and ChIP (Chromatin Immunoprecipitation).

1.5.1. Systematic Evolution of Ligand by Exponential Enrichment (SELEX)

SELEX is an *in vitro* method that is used to select nucleic acids from a large pool based on their selectivity and sensitivity for different molecules (19). In its application in identifying TF binding sites, a pool of DNA oligonucleotides (10^{13} to 10^{15}) comprising sequence nucleotides flanked by constant sequences at each end (19, 20) is partitioned into functional and non-functional sequences based on their affinity for a particular TF of interest. The randomised sequences that are bound to the TF are then selected for an amplification process using a polymerase chain reaction process. The amplified sequences are transformed into a new single stranded DNA pool which is further incubated with the TF and a more stringent selection made for those sequences that have higher affinity for the TF. The selection and amplification steps are repeated until only those nucleotides with the desired binding stringency are left over from the starting population of nucleotides. These are then cloned and each characterized by sequencing (21).

1.5.2. DNAse I footprinting

DNAse I footprinting is a method that employs the DNA cleaving enzyme Deoxyribonuclease I (DNAse I) to identify or create a map of DNA regions of interest that are protected from the action of the enzyme due to specific protein-DNA interactions (22). This method identifies the sequence-specific binding of proteins. The idea behind this approach is that the DNAse I enzyme cleaves DNA randomly at several sites and when the labelled cleaved products are run on an electrophoretic gel, these fragments produce a ladder-like distribution which are visualised in the gel or by using Southern blot assays (23). The presence or inclusion of a ligand or protein that interacts with the DNA however would prevent the DNAse from cutting the site at which the ligand or protein is attached whereas other sites are randomly cleaved as usual. Running this on the gel thus results in a profile similar to the previous run but with gaps where the DNA was protected from the DNAse by the protein-DNA interaction. This gap is what is referred to as the footprint of the protein on the DNA sequence (22, 24).

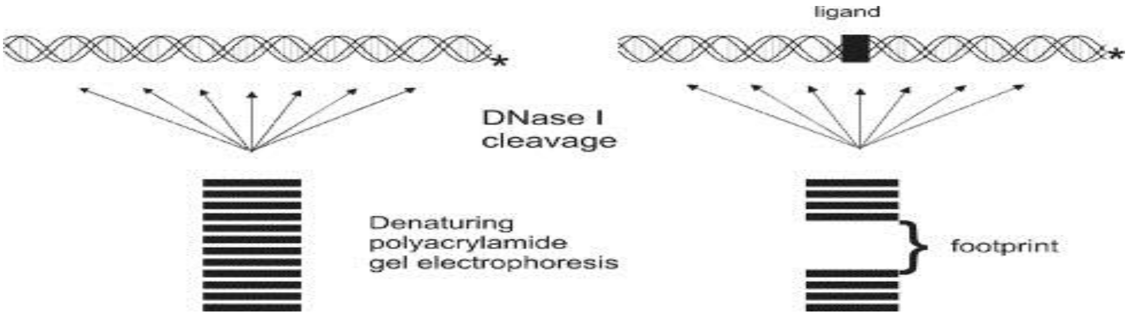


Figure 2 Representation of the DNAse I footprinting experiment. Labelled DNA fragment with bound ligand is digested with DNAse I and run on a polyacrylamide gel. Segment of DNA bound by the ligand appears as a gap. Reprinted with permission from Elsevier: Methods (24), copyright (2007).

In recent times this process has been modified to a high throughput one which allows the simultaneous identification of all such DNAse hypersensitive sites in the genome (25). This DNAse-seq footprinting approach utilizes high throughput sequencing technologies after PCR amplification of the DNAse I digested DNA (23, 26, 27). The idea behind DNAse-seq footprinting is that within the DNAse hypersensitive sites, cleavage of DNA occurs at nucleotides that are not protected by the bound protein and so the distribution of the cleavage sites would not be uniform within the sites. Analysis of individual sites would reveal the presence of peaks and troughs within the signal where the troughs correspond to the binding site of the protein or TF (27, 28).

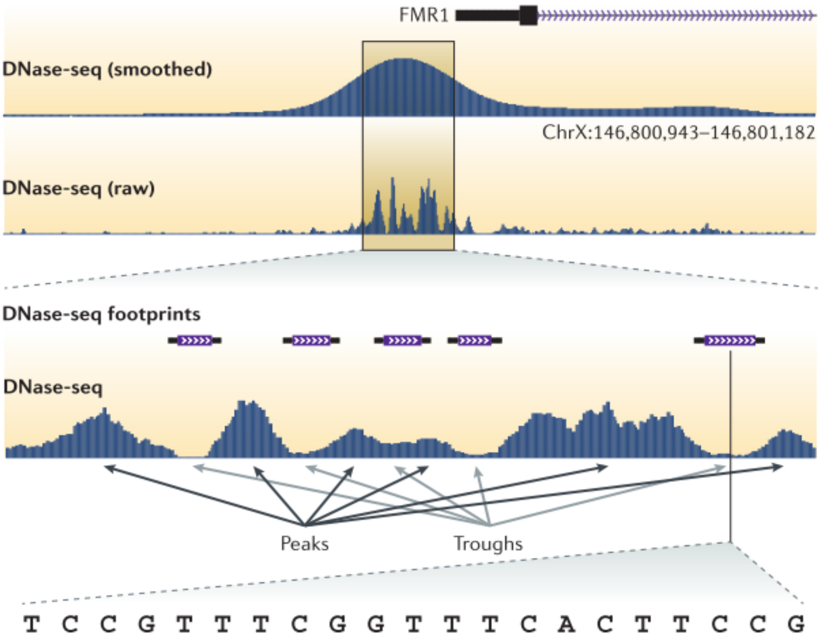


Figure 3 Identification of TF binding sites using DNAse-seq. The troughs represent the sites bound by protein within the hypersensitive site which protects from DNAse cleavage. Adapted with permission from Nature Publishing Group, Nature Reviews Genetics (28), copyright (2012).

1.5.3. Chromatin Immunoprecipitation (ChIP)

Chromatin Immunoprecipitation (ChIP) is a method that allows for the large scale analysis of protein interaction with DNA and thus provides a good means of identifying transcription factor as well as other protein binding sites across the genome of organisms under given conditions. This is achieved by chemically crosslinking proteins to DNA under specific conditions to ensure that the interactions of proteins with DNA are explicitly captured across the whole genome of the cell (29). Following cell lysis and shearing of the DNA into small fragments, an antibody specific to the protein of interest is used to purify and extract DNA fragments to which the cross-linked protein is bound. A reversal of these crosslinks thus yields segments of DNA that under specific conditions are bound by these proteins. Identification of these fragments is achieved by hybridizing with microarrays containing the reference genome as DNA segments (30, 31). This process of ChIP followed by microarray analysis (ChIP-Chip) has however been superseded by ChIP-Seq where the second step of microarray hybridization has been replaced by the more efficient process of next generation sequencing (32). In ChIP-Seq, the purified fragments of DNA obtained after ChIP is sequenced to determine their nucleotide composition. The information obtained from sequencing consists of tens to hundreds of millions of short DNA sequence fragments (reads) of the 5’-ends of both the forward and reverse strands of DNA obtained from ChIP. Comparison of these reads with a reference genome yields regions of overlap (Figure 4). These regions of overlap are subjected to statistical analysis using a control to establish enrichment. Statistically significant regions indicate the regions of the genome where the DNA-interacting protein binds under those specific conditions (33, 34).

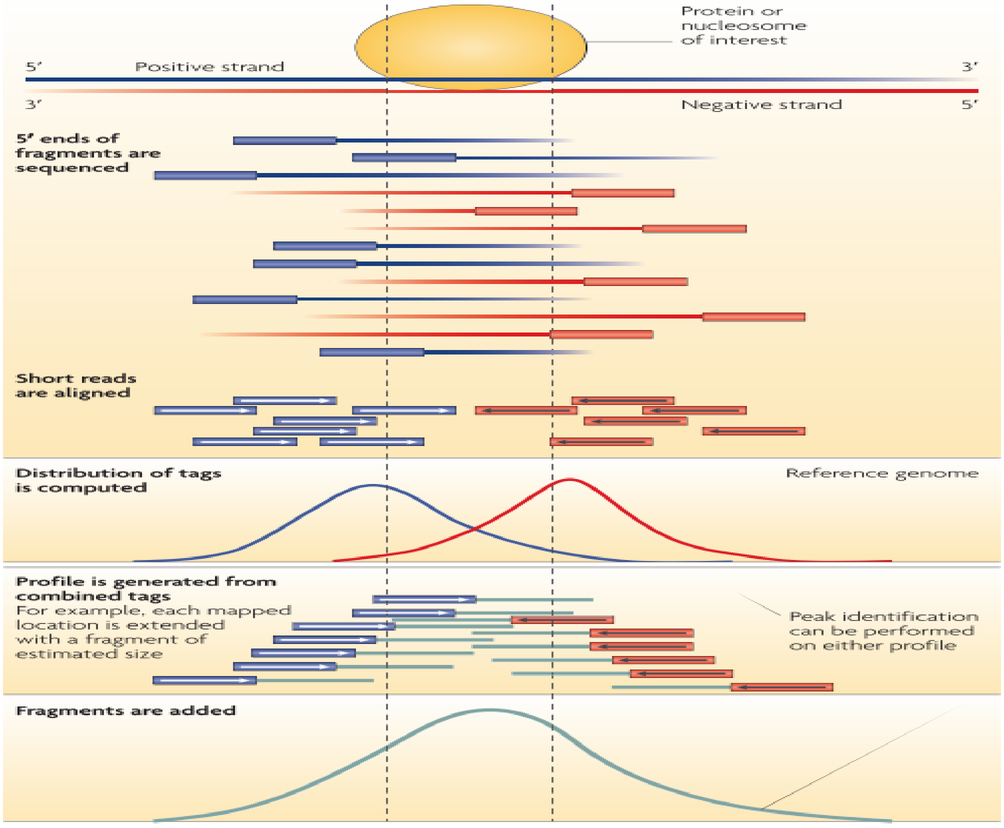

Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

Figure 4 A depiction of the processing steps involved in sequencing genomic data (DNA or RNA) obtained from chromatin immunoprecipitation. Reprinted with permission from Nature Publishing Group: Nature Reviews Genetics (34), copyright (2009)

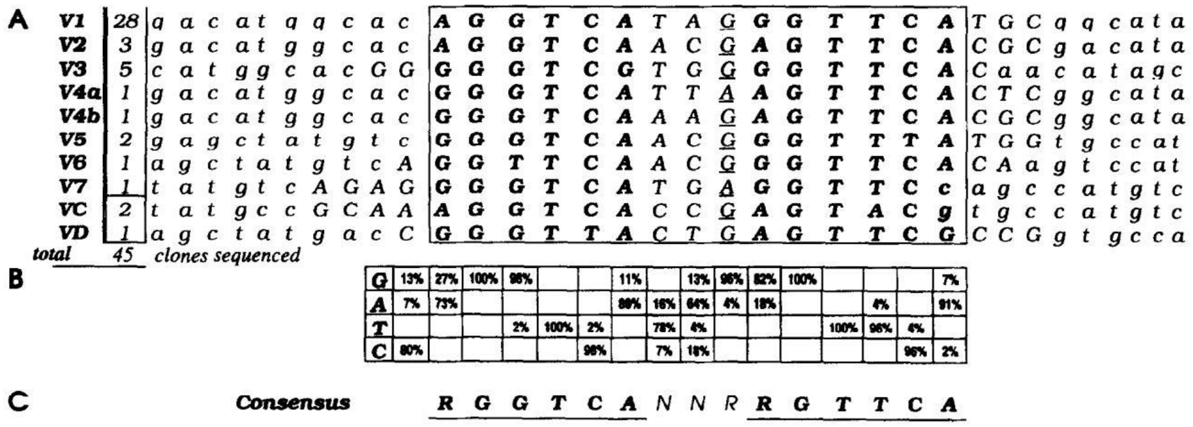
Sequence-specific motifs with which these proteins interact with the DNA can be identified from ChIP-Seq data following a motif scanning or discovery process. ChIP-Seq provides better performance over ChIP-Chip in that it minimizes noise in the data, provides a better resolution and covers a larger region of the genome as compared to the latter which depends on the size of the microarray being used (29, 34, 35).

1.5.4. Computational analysis of TF binding sites

In silico identification strategies for TF binding sites are based on searching for patterns that tend to be overly present in sets of related sequences as opposed to their presence in unrelated sequences (36). Knowledge of genomic regions which are co-regulated or believed to partake in the same processes obtained by the previously described experimental methods and others is used to derive the patterns. These patterns represent the specificity of the transcription factor of interest and may be depicted in either one or several forms namely: a consensus sequence, a position weight matrix (PWM), or a position specific scoring matrix (PSSM).

These patterns are then stored in large databases such as TRANSFAC (37) and JASPAR (38) for use by motif identification algorithms and researchers.

Consensus sequences are obtained by aligning sequences from identified or suspected binding sites. By virtue of the fact that these regions seem to be co-regulated, a pattern of conserved nucleotide positions which run through these sequences is constructed following the alignment which is then used to derive a consensus representation (based on a statistical overrepresentation of the pattern using suitable background frequencies) of the actual region of protein-DNA interaction (39). The derived consensus sequence matches closely in all sites in the sample but this match may not be exact (40) as each position is represented by a consensus nucleotide for all nucleotides present in that position (41). An example of a consensus sequence for the heterodimer of the Vitamin D receptor and Retinoid X receptor is shown in Figure 5. The representation of binding sites by consensus sequences can have variations depending on the number of mismatches allowed or even the positions within which these variations are specified to be allowed in the representation. Although this form of representation may be easy to achieve, obtaining an optimal consensus sequence that is quite capable of predicting new sites may not be too trivial (40).

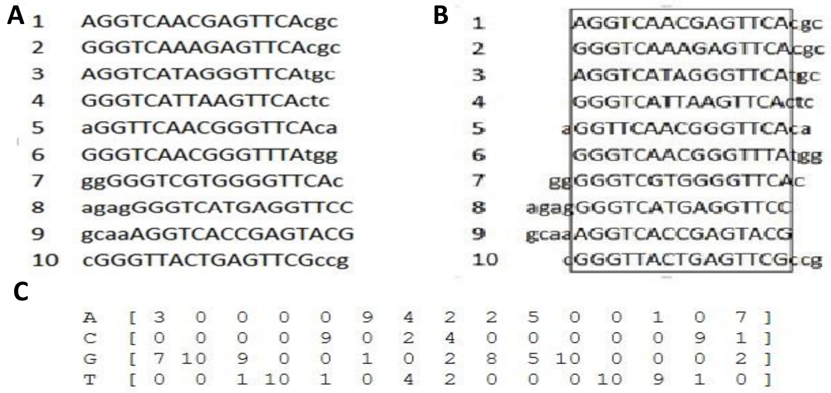

A
V1 28 qacatqqcac AGGTCATAGGGTTCA TGCqqcata
V2 3 gacatggcac AGGTCAACGAGTTCA CGCgacata
V3 5 catggcacGG GGGTCGTGGGGTTCA Caacatagc
V4a 1 gacatggcac GGGTCATTAAGTTCA CTCggcata
V4b 1 gacatggcac GGGTCAAAGAGTTCA CGCggcata
V5 2 gagctatgtc GGGTCAACGGGTTTA TGGtgccat
V6 1 agctatgtcA GGTTCAACGGGTTCA CAagtccat
V7 1 tatgtcAGAG GGGTCATGAGGTTCc agccatgtc
VC 2 tatgccGCAA AGGTCACCGAGTACg tgccatgtc
VD 1 agctatgacC GGGTTACTGAGTTCG CCGgtgcca
total 45 clones sequenced
B
C
Consensus
R G G T C A N N R R G T T C A

G	13%	27%	100%	98%			11%		13%	96%	82%	100%				7%
A	7%	73%					89%	16%	64%	4%	18%			4%		91%
T				2%	100%	2%		78%	4%				100%	96%	4%	
C	80%					98%		7%	18%						96%	2%

Figure 5 A consensus representation of the Vitamin D receptor-Retinoid X receptor heterodimer response element as derived by Colnot et al in (42). (A) An alignment of the cloned sequences used showing the number of times each clone was isolated on the left. (B) Table showing the percentage of each nucleotide present at each position within the response element (region enclosed in the box). (C) Consensus representation of the motif where A, G, T, C are the four nucleotides, N means any nucleotide accepted in that position and R represents an A/G.

Position weight matrices (PWM) and position specific scoring matrices (PSSM) provide an alternative to consensus sequences as a way of representing these TF binding sites. These representations express the probabilities of finding a particular nucleotide at a given position. These representations tend to be more informative about these patterns than the consensus representation as they contain information with regard to the occurrence frequencies of nucleotides in each position of the pattern (41) and allows for inference of how well the TF can bind to such a site based on the idea that the strength of a site is dependent on the contribution of each of the positions making it up.

PWMs are constructed by aligning a set of closely related sequences just as in the case of constructing a consensus motif. A consensus motif can thus be converted to a PWM and vice versa although conversion from a PWM to a consensus representation results in some loss of information (40). A frequency table is then made with each element of the table representing the frequency of each nucleotide at any given position in the alignment. These elements are what are termed the weights. These weights could be expressed as absolute or relative frequencies of each base in a given position. Other methods of computing the weights of the PWM are based on a log likelihoods ratio where the relative frequency of each base in the sequence collection is taken into account. A simple example using 10 sequences for the Vitamin D receptor as shown in the Jaspar_Core database is illustrated in Figure 6. Other implementations may include a pseudocount to correct for small sample sizes.



```
A 1  AGGTCAACGAGTTCAcgc
  2  GGGTCAAAGAGTTCAcgc
  3  AGGTCATAGGGTTCAtgc
  4  GGGTCATTAAGTTCActc
  5  aGGTTCAACGGGTTCAca
  6  GGGTCAACGGGTTTAtgg
  7  ggGGGTCGTGGGGTTCAc
  8  agagGGGTCATGAGGTTCC
  9  gcaaAGGTCACCGAGTACG
  10 cGGGTTACTGAGTTCGccg
```

```
B 1        AGGTCAACGAGTTCAcgc
  2        GGGTCAAAGAGTTCAcgc
  3        AGGTCATAGGGTTCAtgc
  4        GGGTCATTAAGTTCActc
  5       aGGTTCAACGGGTTCAca
  6        GGGTCAACGGGTTTAtgg
  7      ggGGGTCGTGGGGTTCAc
  8    agagGGGTCATGAGGTTCC
  9    gcaaAGGTCACCGAGTACG
  10      cGGGTTACTGAGTTCGccg
```

```
C
A  [ 3  0  0  0  0  9  4  2  2  5  0  0  1  0  7 ]
C  [ 0  0  0  0  9  0  2  4  0  0  0  0  0  9  1 ]
G  [ 7 10  9  0  0  1  0  2  8  5 10  0  0  0  2 ]
T  [ 0  0  1 10  1  0  4  2  0  0  0 10  9  1  0 ]
```

Figure 6 Derivation of a PWM for the RXR::VDR heterodimer using 10 sequences obtained from SELEX experiments as described in (38). (A) The sequence set used. (B) A multiple alignment of VDR binding sequences. (C) Frequency based PWM derived for the *RXR-VDR* binding site.

Identifying a match to the binding site for a known TF would be achieved by calculating scores for the suspected region. The score of a probable binding site is thus the sum of the matrix values for each nucleotide at each position in the sequence in the case of the log-based PWM. A high score indicates close similarity to the consensus indicating that each position may have one of the most common nucleotide in that position for the motif and a threshold may be applied for selection of the most probable sites.

Computational analysis methods can be categorised into two: methods which seek to identify patterns without *a priori* information about the binding sites (motif discovery) and those that use already known patterns of TFs to identify similar in new sets of sequences which are believed to contain binding sites for those TFs (motif scanning/mapping methods) (40). The motif scanning methods use the pattern representations described above in the process. *De novo* motif discovery can as well identify already known motifs in new sequence collections.

1.5.5. Use of additional information to guide binding site prediction and motif discovery

The motif scanning and discovery process is however not as straight forward as it may seem. TFs tend to usually bind to short stretches of DNA sequences (4-10bp) (43) and together with the small DNA alphabet (AGCT), makes the identification of real binding sites difficult. The short motif and small alphabet means that any combination of bases making up a binding site will occur multiple times in the genome by random chance and the highly repetitive nature of nucleotides in the genome also makes it more than likely to identify these short stretches occur at other regions of the genome aside from the actual regulatory or binding sites. This means there is a high likelihood of the sites identified by these methods being false positives (44). In addition, not all genomic regions are available for protein interaction. To help address these challenges several algorithms allow the inclusion of knowledge about these binding sites in the prediction and discovery process. Advances in technology over the past few years means that methods exists that allow the identification of regulatory regions and the features that distinguish them from the surrounding background. The ENCODE (Encyclopedia of DNA Elements) project (45) has yielded a lot of information that has prescribed functions to different parts of the genome and also delineated the features associated with the different parts allowing for a better understanding of the gene regulatory process. Using information such as sequence conservation, histone marks as well as DNAse hypersensitivity information, motif discovery and prediction algorithms are better able to discriminate real binding regions from spurious occurrences (41, 43, 46, 47).

1.6. Workbenches for analysing biological data

The analysis of biological data requires the use of one or several types of resources ranging from databases containing sequence annotations or collections of transcription factor binding profiles to algorithms that enable the processing of raw signals from sequencing experiments.

These resources form an integral part of most analysis pipelines and there has been an increasing number of such resources being made available to help in the solution of the complex questions that arise from biological data. That notwithstanding, there are challenges which arise from having so many resources developed by different groups or even if from the same group, differing in their requirements for use such as input formats. Another challenge in the analysis of biological data using these resources lies in the fact that these resources most often than not are not in one central place requiring that analysis data be moved from one place to another in cases where a multiplicity of these tools are required for a single analysis. The development of workbenches or suites has been of tremendous benefit. These workbenches create a framework that enables the accessibility of these resources from a localized place be it as a standalone application that can be installed on a computer or as cloud-based services that can be accessed from a single web interface. Some notable examples of such workbenches are Galaxy (48) and Mobyle (49).

MotifLab (50) is another example of such a workbench that was developed by researchers at the Norwegian University of Science and Technology. It is a workbench focussing on the integration of tools and data for the analysis of regulatory sequence regions. MotifLab allows for the identification of binding sites for TFs using different motif scanning and discovery tools as well as allowing for the integration of related information from different sources in unrestricted ways in the process. All these processes can be performed from a graphical user interface which makes it easy to use.

1.7. Gene Regulation by the Vitamin D Receptor (VDR)

Vitamin D is a prohormone formed in the skin from 7-dehydrocholesterol with the help of sunlight. Vitamin D is then hydrolysed into a pre-active form, 25-dihydroxy Vitamin D3 by the 25-hydroxylase enzyme in the liver. Further processing in the kidneys by the 25-hydroxyvitamin D-1-α-hydroxylase enzyme results in the formation of active 1,25-dihydroxy Vitamin D3. Several other tissues are capable of synthesizing this active form depending on the levels of pre-active Vitamin D. It is this active form that mediates the pleiotropic effects of Vitamin D uptake including calcium and phosphate absorption in the intestines and bone formation (51-53). The varied effects of this hormone are achieved by binding to its cognate receptor in the cell.

The Vitamin D receptor is a transcription factor belonging to the family of nuclear receptors that mediate the effects of small steroid ligands and molecules on cellular processes (54). The

VDR is however primarily known to mediate the action of 1,25-dihydroxy Vitamin D3 (52, 55). It is composed of 5 principal functional domains which play different roles. The domains include the localization domain which localises the VDR protein to the nucleus after translation, the DNA-binding domain which recognises and binds to the response element, the dimerization domain, the ligand-binding domain to which active 1,25-dihydroxy Vitamin D3 binds and the transcriptional activation (ligand-dependent Activating Function (AF2)) domains (56).

In the absence of its cognate ligand, Vitamin D3, VDR exhibits basal functionality regulating several genes including the gene for the VDR. By ligand binding to its ligand-binding domain, the VDR undergoes conformational changes which activates it and from its monomeric state, forms a heterodimer with the Retinoid X receptor (RXR), another member of the nuclear receptor family (57, 58). Studies have shown the formation of heterodimers of RXR with other nuclear receptors on activation by their cognate stimulatory ligands.

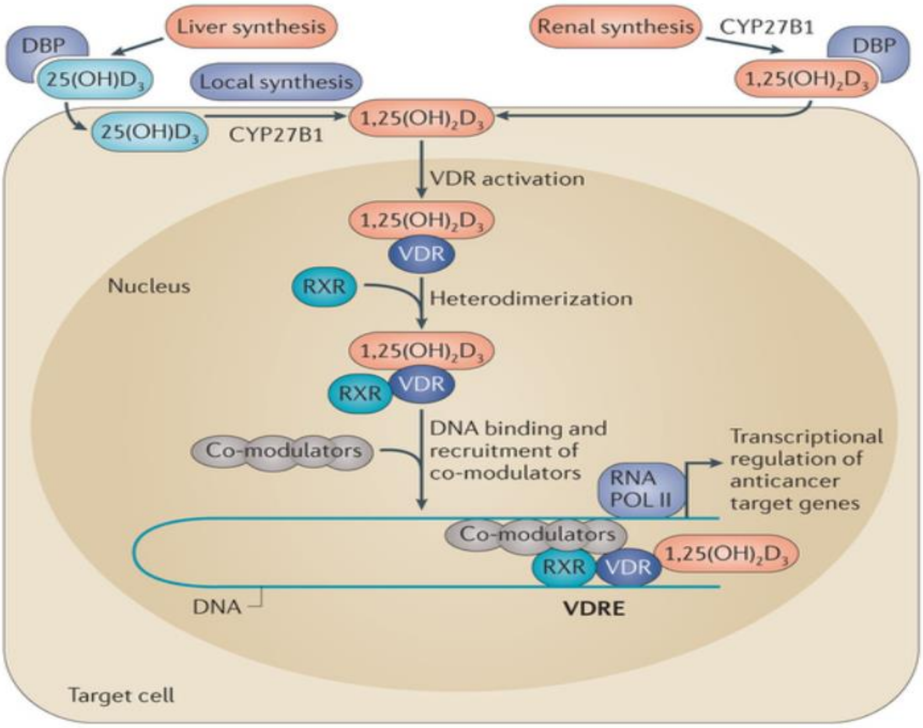


Figure 7 Vitamin D Receptor activation by $1,25(OH)_2$ Vitamin D3 and its mechanism of action. Adapted by permission from Nature Publishing Group: Nature Reviews Cancer (59), copyright (2014)

The VDR can bind to DNA as a monomer or even as dimers but these interactions are not stable. Activation by Vitamin D and heterodimerization with RXR however stabilizes this interaction (11, 12, 60). VDR binds to DNA as described earlier by recognizing the Vitamin D response element (VDRE), a heptad repeat sequence that has a spacer element between the two half-sites. Several values have been given to this spacer region however evidence suggests a preference for a spacer of 3 nucleotides between the half-sites (11, 58). This is known as the DR3-type VDRE where DR indicates that the element is a direct repeat and the value of 3 indicates the spacing between the repeats. As mentioned previously, the VDREs can be located at sites close to the TSS or at distal sites upstream or downstream to the promoter and these sites can contain one to multiple VDREs.

Binding of the VDR heterodimer to the VDRE results in the induction of several genes including the 24-hydoxylase (CYP24A1) gene which encodes the enzyme responsible for catalysing the degradation of 1,25-dihydroxy Vitamin D3 and its precursor (58). To achieve transactivation, the VDR heterodimer acts as a seed which recruits factors responsible for chromatin remodelling such as histone acetyl transferases (HATs) in the form of SRC-1 (Steroid receptor coactivator) or CBP/p300. In addition, TATA binding protein associated factors (TAFs) and the basal transcription machinery are recruited further down in the process. Other factors involved in the transactivating function of the VDR heterodimer include the Vitamin D receptor-interacting protein 205 (DRIP205) which upon binding to the AF2 of VDR, recruits the mediator complex comprising other DRIPs that link the VDR to transcription factor 2B and the RNA polymerase II for transcription initiation. The VDR has been to shown to interact with some members of the mediator complex which serves as a bridge to connect gene-specific regulatory proteins to the RNA polymerase II transcription machinery (56, 61).

Negative regulation by the VDR involves interaction with the VDR-interacting repressor and the recruitment of histone deacetylases (HDACs). In this role VDR has also been shown to interact with the Williams Syndrome Transcription Factor (WSTF) which allows the recruitment of the chromatin remodelling complex WINAC (61-63). A model of gene regulation by the VDR and its co-modulators utilizing chromatin looping as described by Haussler, Jurutka (58) is shown in Figure 8.

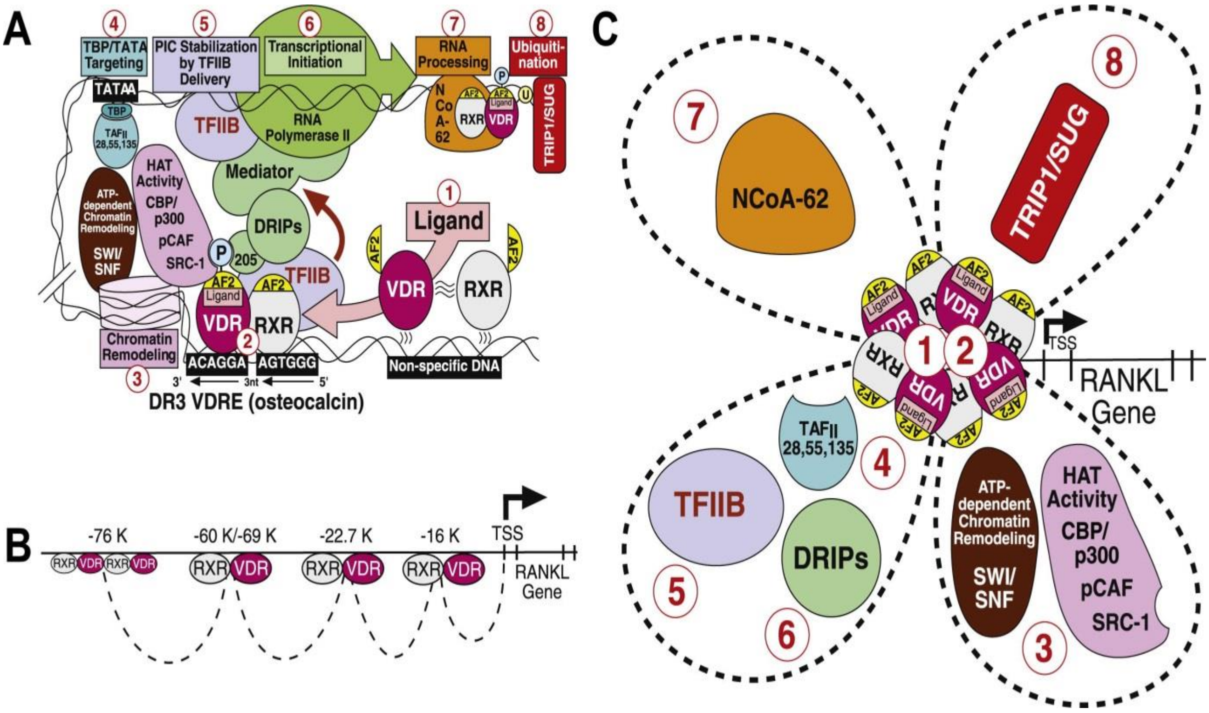


Figure 8 Chromatin looping model of gene regulation by the VDR. (A) Sequential model of gene activation of the rat osteocalcin gene (containing a single VDRE) by the 1α,25(OH)$_2$D$_3$ -bound VDR-RXR heterodimer. Numbers inside circles refer to discreet stages in the activation of a VDR target gene. (B) The presence of several potential VDREs in the 5-prime flanking region of the mouse RANKL gene. (C) Depiction of how these VDREs might cooperate in a chromatin looping model. Multiple VDR-RXR heterodimers may be capable of simultaneously recruiting coactivators to form a regulatory super-complex at the promoter. Adapted by permission from Elsevier: Best Practice & Research Clinical Endocrinology & Metabolism (58), copyright (2011).

1.8. Context and aims of the study

Generally, transcription factors identify a specific string of nucleotides; the response element, which serves as a template/motif with which they interact with the DNA sequence to either activate or repress the transcription of the gene they regulate. In recent times however, it has been observed that some factors do associate with DNA without any clearly recognizable motif for those factors.

One such identified transcription factor is the above-mentioned nuclear hormone receptor for Vitamin D. There is interest in this receptor and its ligand because it has been suggested to play a role in diseases including certain forms of cancer (59, 64). A number of ChIP-Seq experiments carried out for VDR to elucidate and better understand its role and function in gene regulation on a genomic scale have identified sequence regions which have the classic VDR in addition to sequence regions which have no such motif but are indicated by ChIP-Seq as being positive for VDR binding. This observation has not been limited to the VDR as similar observations have been made for other factors of interest (45, 65). Although it has been suggested that these VDR-motif negative regions and others like it could be false positives, they are frequently found in most ChIP-Seq data and the binding events are too strong to be just noise (66). Recent knowledge calls for alternative explanations for the existence of such regions.

Knowledge on transcription related factors and regions of high occupancy of these factors uncovered by the human ENCODE project coupled with recent publications alluding to alternative methods of protein-protein and protein-DNA interactions such as transcription factor co-operativity, motifless binding as well as higher order structure interactions raise several questions and hypotheses with regard to how transcription factors interact with these regions identified in ChIP-Seq experiments. One important question raised by these known-motif-negative regions is the mechanism by which these transcription factors are able to recognise these regions. The closest assumption that may provide insight to this is the presence of a previously unknown alternative motif that is recognised by these transcription factors. The change in recognised binding site could be attributed to conformational changes in the transcription factor protein due to interactions or proximity to other proteins. Co-operativity has also been postulated as a possible explanation for this phenomenon as recent publications describe instances of transcription factor binding to weak sites due to induced stabilization of protein-DNA interactions by cooperative effects of interacting proteins (67). An extension of this hypothesis is the possibility of these transcription factors acting as secondary factors that bind indirectly to these regions by binding to other transcription factors with motifs in this region and piggy-backing. Such a process has previously been described for the VDR where it interacts with the Vitamin D receptor interacting repressor (VDIR) in the transrepression of the Vitamin D 1α hydroxylase gene (68). Some other studies have also shown the estrogen and progesterone nuclear receptors interacting with other transcription factors instead of directly binding to DNA to regulate transcription (69, 70).

This study seeks to provide some explanation of transcription factor binding where there is no clearly recognizable motif as compared to instances from the same experiment where a motif is found by analysing ChIP-Seq data from VDR experiments to identify any characteristics that may be overrepresented in these motif-negative regions.

Through the analyses to be performed, the study would establish whether the motif-negative regions differ from the motif-positive ones with respect to its GC-content, the presence of CpG sites or chromatin marks and modifications that are associated with high probability binding events. In addition, the study would seek to answer the question of whether the motif-negative regions are enriched in other motifs compared to the positive regions indicating the possibility of indirect TF binding and also test the hypothesis of co-operativity by identifying motif pairs enriched in these motif-negative regions.

2. Methods

2.1. Datasets and tools

ChIP-Seq data for VDR in human lymphoblastoid cell lines (LCL) (71) and THP-1 human monocytic leukaemia cells (MCL) (72) subsequently referred to as Ramagopalan and Heikkinen respectively were used in this work. MotifLab versions 1.07 and 1.08 (50) was the main workbench used for analysis.

Data for the CCCTC factor (CTCF) and the Structural Maintenance of Chromosome (*Smc3*) unit, a key component of the Cohesin complex, was retrieved for the GM12878 cell line from the ENCODE database[1]. Information on transcription factor interaction partners in mouse and human from (73) was retrieved as supplementary data[2].

Genomic data used from research articles that were based on the human genome 19 (hg19) assembly were converted to hg18 using the LiftOver tool on the University of California Santa Cruz (UCSC) Genome Bioinformatics website. Other tools and resources used include BioGrid 3.2 (for protein interaction data), UniprotKB (for protein information), NetPath, and STRING. Others are indicated as and when they were used.

2.2. Motif discovery and comparison

Initial motif discovery was carried out using the online version of MEME-ChIP 4.9.0 (74). The *bed* file for each dataset was converted to *fasta* format using MotifLab before submission to MEME. MEME-ChIP was set to report 3 motifs between 6 and 30 bases in length with all other parameters set to default. Jaspar_Core and mouse (UniPROBE) (38) was selected as the motif database to use. Identified motifs were submitted to TOMTOM 4.9.0 (75) for motif comparison using the Pearson correlation coefficient as motif column comparison function and an E-value threshold of less than 10.

2.3. Background modelling

DNA sequence data from ChIP-Seq was imported into MotifLab using information from human genome version 18 (hg18). A fourth order background model was created using the *background model* procedure in MotifLab. The model was created from the imported sequences by setting the model order and strand orientation to 4 and relative respectively.

[1] http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeAwgTfbsUniform

[2] http://www.sciencedirect.com/science/MiamiMultiMediaURL/1-s2.0-S0092867410000796/1-s2.0-S0092867410000796-mmc2.xls/272196/FULL/S0092867410000796/7a324eec378f4ed8963a6eafb6c3b47f/mmc2.xls

Table 1 Parameters for background model generation

Background model (from track)	
Parameter	Value
DNA track	*DNA*
Sequences	*All sequences*
Model order	*4*
Strand orientation	*relative*

The DNA track of input sequences were then masked with random bases from the background model to create five different sets of simulated sequences using the *mask* procedure 5 times. Other parameters for this were kept at default.

Table 2 Parameters for deriving simulated sequences for establishing statistical significance

Transform (mask)	
Parameter	Value
Source	*DNA*
Mask with	*random bases (background model)*
Strand	*relative*
Condition	*none*
In sequence collection	*All sequences*

2.4. Motif scanning

Sequences were scanned using the *SimpleScanner* algorithm in MotifLab at a threshold level of 90% with all other parameters kept at the default. Motifs from a predefined motif collection (of 459 motifs (04/10/13)) from Jaspar_Core_2009 database was used for scanning. The source parameter was varied depending on the transformation carried out on the DNA sequences such as masking repeat regions.

Table 3 Motif scanning parameters for the *SimpleScanner* algorithm

Motif scanning	
Parameter	Value
Source	*DNA (varied)*
Method	*SimpleScanner*
Motif Collection	*varied*
Threshold type	*percentage*
Threshold	*90*
Score	*absolute*
In sequence collection	*All sequences*

2.5. Motif overrepresentation

The occurrence of motifs from motif scanning was calculated for each set of simulated sequences using the *motif numeric map* procedure. The value for each motif in the map was obtained by setting the property parameter of the method to frequency.

Table 4 Parameters for generating motif counts in sequences

Motif numeric map (from track)	
Parameter	Value
Motif Track	*Varied (total of 5 tracks used)*
Property	*frequency*
Sequence Collection	*All sequences*
Within regions	*none*

The average values for each motif was obtained from the set of values for each simulated sequence, using the *increase* procedure and finally dividing the sum by the total number of simulated sequence sets (five). The relative occurrence of motifs in the input sequences was calculated using the *count motif occurrences* procedure with parameters for *motifs* and *significance threshold* specified as Jaspar_Core and 0.05 respectively. Average motif occurrences calculated for the simulated sequences served as background frequencies for reporting p-value and establishing statistical significance of motif overrepresentation. A Bonferroni correction was applied using the number of motifs.

2.6. Region comparison and nucleotide level statistics

Feature annotation tracks containing information for *CpG* regions, *DNAse HotSpots*, general regions bound by transcription factors identified by ChIP-Seq (*TFBS_ChIP-Seq*), gene coding regions (*CCDS*), Repeat regions detected by Repeatmasker 3.2.7 (*RepeatMasker327*), histone modifications (H3K4me1, H3K4me3, H3K9ac and H3K27ac regions for the GM12878 and K562 cell lines depending on the cell line used for acquiring the ChIP-Seq data) and *FAIRE-Seq* regions from the UCSC Genome Browser (76) as well as *Ensembl* gene regions from Ensembl (77) were used. These feature tracks were retrieved using the predefined list of feature tracks in MotifLab. Additional data was retrieved from the genome browser. The feature tracks used are:

wgEncodeBroadChipSeqPeaksGm12878H3k27ac
wgEncodeBroadChipSeqPeaksK562H3k27ac
wgEncodeBroadChipSeqPeaksGm12878H3k4me1
wgEncodeBroadChipSeqPeaksK562H3k4me1
wgEncodeBroadChipSeqPeaksGm12878H3k4me3

wgEncodeBroadChipSeqPeaksK562H3k4me3
wgEncodeBroadChipSeqPeaksGm12878H3k27me3
wgEncodeBroadChipSeqPeaksK562H3k27me3
wgEncodeBroadChipSeqPeaksGm12878H3k9ac
wgEncodeBroadChipSeqPeaksK562H3k9ac
wgEncodeBroadChipSeqPeaksGm12878

Percentage overlap of VDR binding sites with different annotated features was calculated using the *compare region datasets* procedure. All binding sites (VDR and VDR-like) were compared to each feature annotation track. Analysis was performed using only VDR-positive sequences.

Table 5 Parameters for comparing motif sites and DNA feature regions

Compare region datasets	
Parameter	Value
First	*Varied (TFBS tracks for MEMEVDR,JSVDR,NR4A2)*
Second	*Varied (Each feature region track)*
Sequences	*Varied (Specific sequence collection)*
	VDR_positive
	VDR_negative
	MEMEVDR_positive

Enrichment of regions in the different sets of sequences was compared using the *compare region occurrences* procedure of the analysis feature in MotifLab. A hypergeometric test was applied to compute p-values at the 0.05 significance level. The number of region types that were present for each comparison was used for Bonferroni correction. VDR- positive and negative sequence sets were compared using each feature track. The positive set was specified as the control in all instances. Comparisons were also made between classic VDR (JSVDR)-positive and MEMEVDR-positive, NR4A2-positive and MEMEVDR-positive, and NR4A2-positive and JSVDR-positive.

Table 6 Parameters for comparing occurrence of DNA feature regions in VDR positive versus negative sequences

Compare region occurrences	
Parameter	Value
Region track	*Varies (Each feature region track)*
Target set	*VDR_negative*
Control set	*VDR_positive*
Statistical test	*Hypergeometric*
Significance threshold	*0.05*
Bonferroni correction	*Present types*

GC-content for all sequences in the dataset and for the different sequence collections was obtained using *analyse* procedure and the parameters below. The value for the Groups parameter was varied depending on which sequence collection was being analysed.

Table 7 Parameters for the calculation of GC-content

Analyze	
Parameter	Value
Analysis	*GC-content*
DNA track	*DNA*
Groups	*Varied*

2.7. Evaluating the power of feature regions to predict binding sites

The potential of the imported feature tracks to discriminate or predict binding sites in the sequences was assessed by plotting Receiver Operating Characteristic (ROC) curves for each numeric feature track using the observed binding sites for MEMEVDR, JSVDR and *NR4A2* as targets. This was to evaluate whether these numeric feature tracks have higher values inside the observed TF binding sites as compared to the surrounding regions. The *evaluate priors* procedure in MotifLab was called with the following parameter values to achieve this. The priors tracks were the ChIP-Seq signals for these features in the GM12878 and K562 cell lines depending on the dataset being analysed.

Table 8 Parameters for evaluating the predictive power of DNA feature regions for TF binding sites

Analyze: Evaluate priors	
Parameter	Value
Target track	*Varied (TF binding site tracks for JSVDR, MEMEVDR and NR4A2)*
Priors track	*H3K4me1*
	H3K4me3
	H3K27ac
	H3K9ac
Sequences	*JSVDR_positive*
	MEMEVDR_positive
	NR4A2_positive
Threshold	*Above or equal*

2.8. Automatic generation of positional priors

VDR binding site overlap with feature regions were used as inputs to the automatic *priors generator* in MotifLab. The priors generator was specified as a neural network and the input nodes varied using combinations of features above a certain threshold of overlap of binding sites. Three clusters were used; above 50% (ChIP-Seq, DNAse HotSpots, Ensemble Genes, FAIRE-Seq, H3K4me1, H3K4me3 and H3K27ac), above 70% (ChIP-Seq, DNAse HotSpots, FAIRE-Seq, H3K4me1, H3K4me3 and H3K27ac) and above 90% overlap (DNAse HotSpots). The neural network classifier forming the basis of the priors generator was trained on the classic *RXR-VDR* containing set of sequences.

2.9. Manual generation of priors

Histone modification signals were imported into MotifLab. Each signal peak was processed using the *apply* procedure with a sum window of 10 anchored at the start of the sequence. A priors track was generated by summing signal peaks in regions having values greater than 0. The *increase* procedure was used. Generated prior was normalized to change 0 signal peak values in each sequence to 0.01 while maintaining the maximum in each sequence. Tracks whose AUC values indicated bad predictive ability were reversed with a reciprocal function. Combinations of these reversed tracks were also made.

2.10. *De novo* motif discovery

PRORITY and ChIPMunk were used for *de novo* motif discovery. Generated positional priors were used in *de novo* motif discovery using PRIORITY. ChIPMunk was also used on the VDR-positive and negative sets. ChIPMunk was parameterised to find motifs 7 to 19 bases in size, zero or one occurrence per sequence (ZOOPS) using a simple model. ChIPMunk was further used with manually generated priors by setting the method parameters: maximum motif length as 16, minimum motif length as 7, model type as peak and motif to report as one occurrence per sequence (OOPS). The *peaks* parameter was specified as a manually generated prior track.

2.11. Comparison of classic VDR and VDR-like sequences collections

Sequence collections were compared using the *compare sequence collections* procedure in MotifLab. A binomial statistical test using a significance threshold of 0.05 was used in all comparisons made. The threshold was corrected using a Bonferroni correction of the number of all motifs. Motifs overrepresented from the background were specified as the motif collection for use. Analysis was carried out between classic *RXR-VDR* and VDR-like sequence collections to establish whether there was any intersection between the two sequence collections.

2.12. Motif distribution

Having established the overrepresented motifs in all the clusters of sequences, further analysis of how these motifs were distributed with respect to the features of interest in each cluster was performed. This involved calculating average distance of present motifs to other motifs such as those for known or putative interaction partners as well as chromatin opening factors such as *Sp1*. Information about interaction partners were retrieved using BioGrid 3.2, IntAct, Human Protein Reference Database (HPRD) and the UniprotKB database. Additional information was obtained from STRING 9.05, a database of known and predicted protein interactions as well as the supplementary table of transcription factor interactions in humans from (73).

The *distance* procedure was used to calculate the positions of individual nucleotides from VDR binding sites (upstream, downstream and both directions) in VDR-positive sets and sequence upstream ends in VDR-negative sets. This created a numeric data track that was used in the next step.

Table 9 Parameters for esimating motif distribution in sequences

Derive: Distance	
Parameter	Value
Direction	*Both/Upstream/Downstream*
From anchor point	*Region (VDR_TFBS)/Sequence upstream end*
Relative to	Sequence upstream end ()

2.13. Comparing motif positions in VDR positive and negative sequences

To derive distances in the negative sequences, a sequence numeric map was generated using the *statistic* procedure to count the number of bases in each sequence which is equivalent to the length of the sequences. Values in the generated sequence numeric map were then divided by 2 to obtain an approximate location value for the midpoint of each sequence. This map of midpoint positions was then specified as the parameter value of ‘From anchor point’ in the *distance* procedure.

Table 10 Parameters for determining the base count (length) of each sequence

Derive: Statistic	
Parameter	Value
Source	*DNA*

Statistic Function	*Base count*
Strand	*Relative strand*
In sequence collection	*All sequences*

Table 11 Parameters for obtaining the position of each nucleotide relative to the centre of each sequence

Derive: Distance	
Parameter	Value
Direction	*Both*
From anchor point	*Position (numeric map of sequence midpoints)*
Relative to	*Sequence upstream end*

To determine the relative distances of other TFBSs from the identified VDR and VDR-like binding sites, a comparison was made between the average nucleotide position from the sites and the location of these other TFBSs. This was achieved using the *compare motif track to numeric track* procedure.

Table 12 Parameters for calculating the the distance of motifs from a specified motif

Compare motif track to numeric track	
Parameter	Value
Motif track	*Observed_TFBSs*
Motifs	*Enriched_Ramagopalan/Enriched_Heikkinen*
Numeric track	Varied (*Distances calculated for specific motif of interest*)
Sequences	Varied (*Specific motif-positive set of sequences*)
Threshold	*150*

2.14. Motif position distribution

Table 13 Parameters for generating the distribution of motifs in sequences in each dataset

Parameter	Value
Motif track	*Varied*
Motifs	*Enriched_Ramagopalan/Enriched_Heikkinen*
Sequences	*Varied*
Alignment anchor	*centre*
Motif anchor	*centre*
Support	*No*
Bins	*999/383*

The chosen values for the bins are based on the maximum length of sequence in Ramagopalan and Heikkinen datasets which was 9981 and 3821bp respectively.

2.15. Other data processing methods

CTCF and Cohesin data for the GM17828 cell line was retrieved from the UCSC genome browser and further manipulated using bedtools[3] v.2.17. For testing the region overlap between the datasets used and data for different genomic regions, the command used was ‘`bedtools intersect -a inputA.bed -b inputB.bed –u > output.bed`’ where `inputA` was the Ramagopalan/Heikkinen dataset and `inputB` was the region to be compared. The parameter `–u` was specified to report only single instances in `inputA` where there was at least one overlap.

[3] Downloaded from http://code.google.com/p/bedtools/downloads/detail?name=BEDTools.v2.17.0.tar.gz

3. Results and Discussion

3.1. VDR binds to an *RXR-VDR*-like motif

Ligand-bound VDR has a known consensus motif with which it interacts with DNA after heterodimerization with RXR (Figure 5 and Figure 10). Identification of this template in ChIP-Seq data is in most cases an indication of positive binding however actual presence of transcription factor binding sites may be affected by sequence composition and other features resulting in inactive sites or by chance occurrences (36). Using ChIP-Seq data from (71) and (72), the Jaspar_Core database in MotifLab and the MEME-ChIP tool in MEME Suite (74), a search was made for highly represented motifs in the 2776 and 2338 sequences which make up the two datasets as compared to a background.

MEME *de novo* motif discovery yielded three motifs which were substantially enriched (Figure 9) in the submitted sequences from Ramagopalan. The second most enriched ($E\text{-}value=3.1e^{-41}$), on submission to TOMTOM (75) had a high significant similarity to the *RXR-VDR* motif ($p\text{-}value=3.3e^{-12}$). The other two motifs identified by MEME had no similarity to the known VDRE as ascertained by TOMTOM. Two other high scoring motifs similar to that identified by MEME were the *NR4A2* and *usp* motifs (Figure 10) ($p\text{-}value<10^{-3}$ using TOMTOM). Seventeen other motifs were retrieved in total based on the settings used. A visual comparison of the *de novo* motif from MEME and the Jaspar_Core profile for *RXR-VDR* (MA0074.1) referred to as the classic VDR in subsequent mentions showed that the Jaspar_Core profile had a predominant T at positions 4, 12 and 13 unlike the *de novo* motif which allowed for a bit more variability in these positions. A similar observation has been made in ChIP-Seq data from a different cell line (72).

Sequences from the Heikkinen dataset analysed with MEME however, had no clear motif that matched the classic VDR motif. All the motifs returned by MEME when submitted to TOMTOM had no *RXR-VDR* matching result. DREME on the other hand produced a motif that had some similarity to the VDR-like *NR4A2* ($E\text{-}value=1.6e^{-3}$, Figure 9 and Figure 11).

Table 14 Summary of key properties of analysis datasets

	Ramagopalan dataset	Heikkinen dataset
Cell line	GM10855/GM10861 (LCL)	THP-1 MCL
Duration of Vit. D stimulus	36 hours	40 minutes
Concentration of ligand	0.1µM	unknown

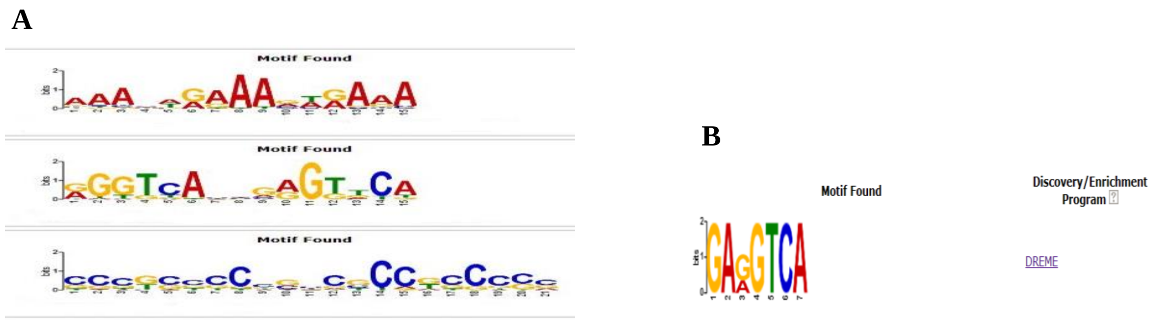
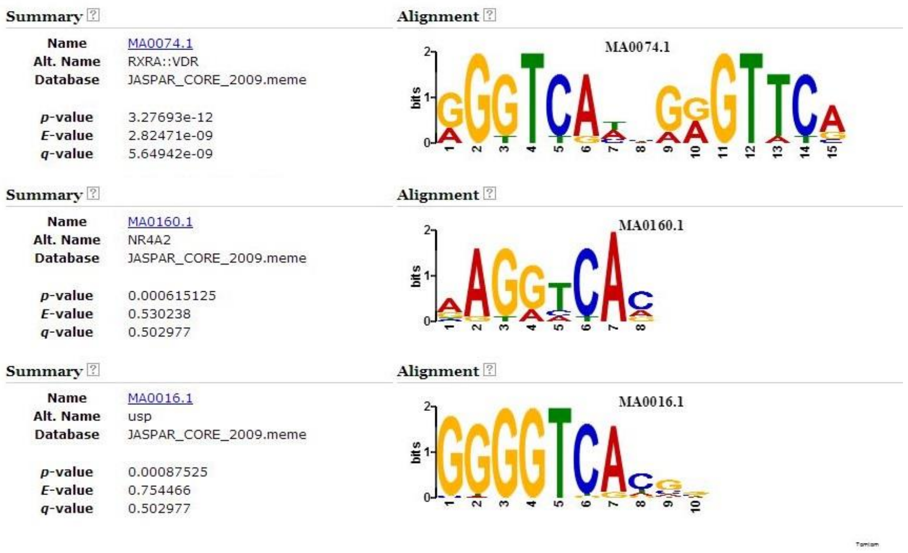


Figure 9 Sequence logos of VDR binding motifs found by *de novo* motif discovery. (A) Top motifs discovered by MEME in Ramagopalan dataset. (B) Motif discovered by DREME in Heikkinen dataset.

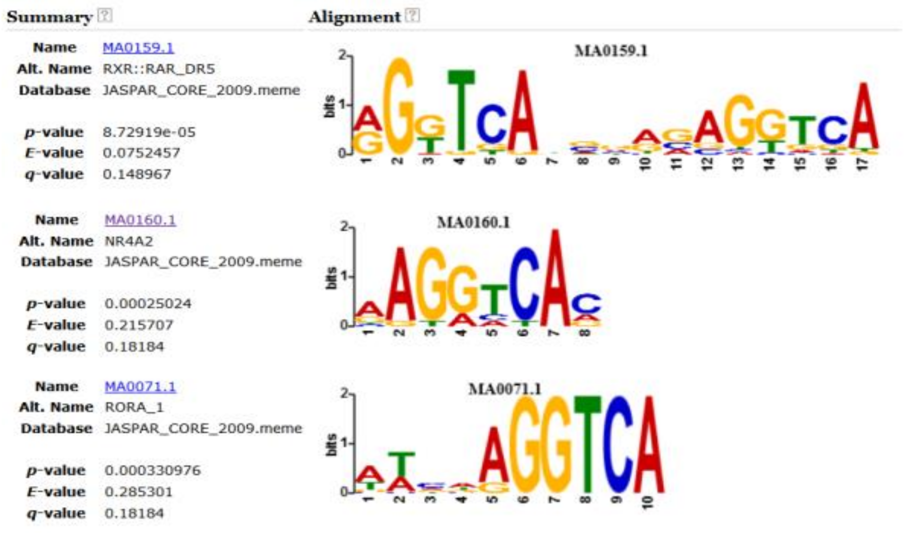


Name	MA0074.1
Alt. Name	RXRA::VDR
Database	JASPAR_CORE_2009.meme
***p*-value**	3.27693e-12
***E*-value**	2.82471e-09
***q*-value**	5.64942e-09

Name	MA0160.1
Alt. Name	NR4A2
Database	JASPAR_CORE_2009.meme
***p*-value**	0.000615125
***E*-value**	0.530238
***q*-value**	0.502977

Name	MA0016.1
Alt. Name	usp
Database	JASPAR_CORE_2009.meme
***p*-value**	0.00087525
***E*-value**	0.754466
***q*-value**	0.502977

Figure 10 Top three motifs similar to motif 2 (Figure 9) from MEME identified by motif comparison using TOMTOM



Name	MA0159.1
Alt. Name	RXR::RAR_DR5
Database	JASPAR_CORE_2009.meme
***p*-value**	8.72919e-05
***E*-value**	0.0752457
***q*-value**	0.148967

Name	MA0160.1
Alt. Name	NR4A2
Database	JASPAR_CORE_2009.meme
***p*-value**	0.00025024
***E*-value**	0.215707
***q*-value**	0.18184

Name	MA0071.1
Alt. Name	RORA_1
Database	JASPAR_CORE_2009.meme
***p*-value**	0.000330976
***E*-value**	0.285301
***q*-value**	0.18184

Figure 11 DREME discovered motif and the top 3 motifs matching the discovered consensus after submitting to TOMTOM for the Heikkinen dataset. The *de novo* motif has some similarity

to the RXR half-site of the *RXR-VDR* motif. The VDR half-site of the heterodimer has a predominant T in the fourth position from the 3' end.

A fourth order background model based on each dataset's sequences was subsequently created to distinguish enriched motifs from those occurring by chance by virtue of the base composition of the sequences. Higher order background models have been shown to improve the results obtained from motif-finding algorithms (78). The background model was scanned using the Jaspar_Core database of transcription factor motifs as was the actual sequences. Occurrence frequencies of motifs in the background were calculated and used to calculate occurrence frequencies of motifs in the actual dataset and thus establish motif overrepresentation. This background model was also used in subsequent motif discovery steps where a background had to be specified.

Out of the 459 motifs that are included in the Jaspar_Core database, 135 were identified as overrepresented (Appendix *A.2*) in sequences from the Ramagopalan dataset as compared to 149 in the Heikkinen set. These motifs formed the focus of the subsequent analyses performed on each dataset respectively. Comparison of the seventeen motifs retrieved by TOMTOM with the 135 motifs observed to be overrepresented in Ramagopalan showed an intersection of only 2 motifs; *RXR-VDR* and *NR4A2*. Using these two motifs for motif scanning within MotifLab as previously described yielded 65% of the scanned sequences showing the presence of one or more of the two motifs (Table 15). This was not too different from results earlier published on the same dataset (71) which had 67% of sequences having the VDR-like sites. This difference may however be attributed to the threshold applied to motif scanning, filtering out of VDR-like motifs which were not indicated as being statistically enriched in these sequences, for example the *usp* motif, as well as a difference in the number of motifs returned by MEME which had similarity to the classic VDR motif. Similarly, motif scanning using MotifLab with only the MEME-derived motif showed multiple to one occurrence per sequence in 18% of sequences in the Ramagopalan dataset (504 out of 2776).

Table 15 Percentage of sequences indicated by motif scanning as being positive for VDR or VDR-like motifs. Most sequences were positive for the VDR-like *NR4A2* motif in both datasets compared to the MEME *de novo* motif or the Jaspar_Core profile for *RXR-VDR*. Some sequences were positive for all three motifs whereas others had different combinations of the three motifs.

Motif	Number of sequences (%) with motif	
	Ramagopalan	Heikkinen
MEME *de novo*	504 (18.2)	1110 (47.5)
JSVDR(*RXR-VDR*)	192 (6.9)	163 (~7.0)

VDR-like (*NR4A2*)	1741 (62.7)	1937 (82.8)
JSVDR+*NR4A2*	1810 (65.2)	1950 (83.4)
De novo+JSVDR+*NR4A2*	1924 (69.3)	2023 (86.5)

Following a similar process of analysis on the Heikkinen dataset showed that about 47% of sequences had the DREME-derived VDR-like *NR4A2* motif (1110 out of 2338). Scanning with the Jaspar_Core *RXR-VDR* however had 163 sequences (~7%) being positive for the motif although MEME *de novo* motif scanning results was negative for this motif. Of the similarities to the DREME-motif identified by TOMTOM, 10 motifs were included in those overrepresented from the background.

The differences observed between the two datasets can be attributed to the fact that samples were from two different cell lines and the treatment conditions were different (36hrs and 40mins after ligand stimulation for samples from the LCL and THP-1 MCL respectively). The longer time before harvest after stimulation may explain the higher fraction of *RXR-VDR* motif-positive regions in the Ramagopalan dataset as compared to Heikkinen in that the VDR has more time to be fully activated by ligand and is thus more able to identify and bind to high affinity sites such as its classic response element. This would however mean that sites bound but without a motif are as a result of inactive VDR or not fully activated VDR. This explanation is supported by evidence that there is a shift in binding sites from those lacking classical VDREs to DR3-type response elements after stimulation (72) but it however does not account for how the VDR binds to these motifless regions. A time course experiment with ChIP-Seq in the same cell line for both VDR and its heterodimerization partner, RXR, may give further insight as to whether these motifless regions are simultaneously occupied by both VDR and RXR under the same conditions and also elucidate whether VDR binds to the DR3 response element in a time dependent manner as the information from these two datasets seem to suggest. It is important to note however that, a study utilizing this suggested approach in a different cell line found that the VDR and RXR did not always co-localize to the same region (79) supporting the view that VDR may bind as a monomer or dimerize with alternative partner(s) and not require RXR even after ligand stimulation.

It was also hypothesised that the huge difference in the type of VDR binding between the two datasets could mean an associated difference in other features between them particularly in terms of presence of motifs for other DNA binding factors. A comparison of enriched motifs between the two datasets showed a significant overlap (*p-value=2.87e*$^{-40}$) of 105 motifs which were present (Appendix *A.6*, Figure 12). Further analysis was performed to understand the role and distribution of these common motifs and those that were unique to each dataset.

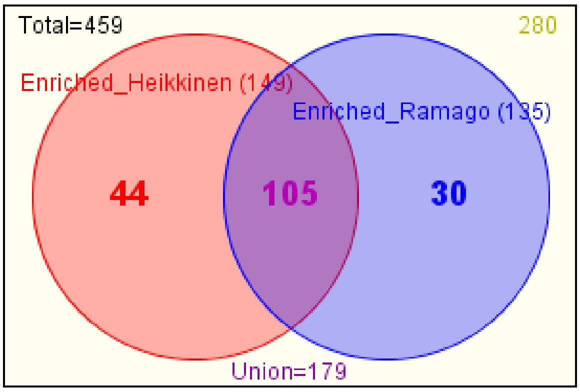


Figure 12 Venn diagram showing the intersection of enriched motifs in Ramagopalan and Heikkinen datasets. A large fraction of enriched motifs are shared between the datasets from LCL and THP-1 MCL

To facilitate the subsequent analysis, a distinction was made between sequences with and without the motif(s) of interest. Sequences indicated by scanning as bearing a VDR motif were selected as the VDR-positive subset. This subset of sequences comprised sequences with the classic *RXR-VDR* motif and those with VDR-like motifs. These two classes were clustered as well. Sequences bearing no VDR or VDR-like motif formed the VDR-negative subset.

3.2. VDR binds in HOT (High Occupancy of Transcription Factor) regions

DNA sequence regions, depending on their base composition, acquired modifications and structure confer properties which in most cases distinguish them from other regions. These properties and characteristics also influence the type and manner of interactions in which these sequence regions are involved. Knowledge of associated properties of regions of interest can therefore be used for predictive purposes. In the case of transcription factor binding, data exists on the human genome which can be used to potentially predict regions of high likelihood of transcription factor binding (65). In addition, knowing these regions can be used to focus the search for transcription factor binding sites (47).

Based on the different human genomic regions defined by Yip, Cheng (65), a comparison was made between the VDR ChIP-Seq datasets and these genomic regions to identify the defined regions within which the VDR datasets fell. VDR peaks from the Ramagopalan dataset had an overlap of 81.6% and 94.1% with High Occupancy of Transcription related factor (HOT) regions and Binding Active Regions (BAR) respectively. HOT regions were defined as genomic “regions bound by many transcription related factors that do not usually co-associate globally in the whole genome” and BARs as “broad genomic regions that transcription related factors tend to bind” (65). On the other hand, there was low overlap with Binding Inactive Regions (BIR) and Low Occupancy of Transcription related factor (LOT) regions. These observations were similar in the Heikkinen dataset but with lower values for the percentage overlap (Table 16).

Table 16 Comparison of VDR binding peaks and defined genomic regions

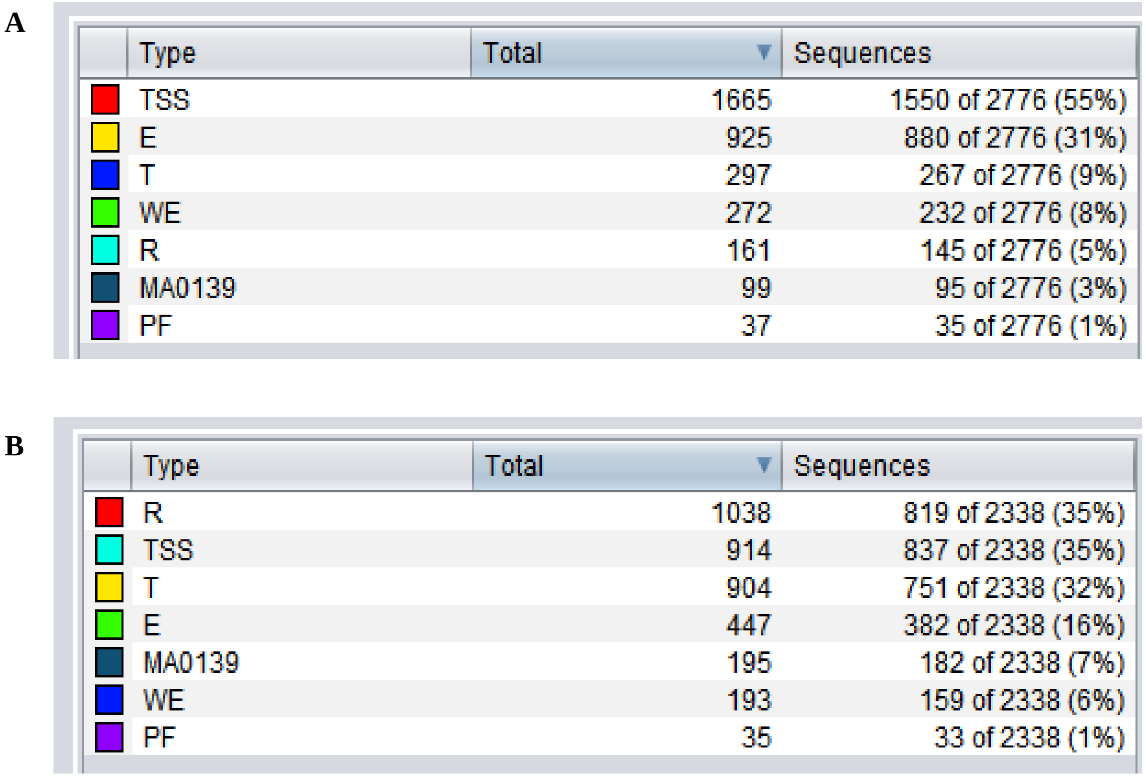
Region types	Number of sequences	
	Ramagopalan dataset	Heikkinen dataset
VDR-binding peaks (Peaks)	2776	2338
HOT regions	92551	81398
LOT regions	199325	166941
BAR regions	213438	175957
BIR regions	2611774	1738767
PRM regions	105147	100781
DRM regions	58222	47733
Peaks∩HOT	2264 (81.6%)	940 (40.2%)
Peaks∩LOT	9 (0.3%)	29 (1.2%)
Peaks∩BAR	2614 (94.1)	1538 (65.8%)
Peaks∩BIR	28 (1.0%)	211 (9.0%)
Peaks∩PRM	1575 (56.7%)	1094 (46.8%)
Peaks∩DRM	635 (22.9%)	327 (14%)

For the Ramagopalan and Heikkinen datasets, defined genomic regions for the GM12878 and K562 cell lines were used. The number of regions is given with the percentage of VDR-binding regions that overlap the different regions in parenthesis. ∩ is intersection. High occupancy of transcription related factors (HOT), Low occupancy of transcription related factors (LOT), Binding active regions (BAR), Binding inactive regions (BIR), Promoter-proximal regulatory modules (PRM) and Gene-distal regulatory modules (DRM).

These observations together with the fact that a large number of TF motifs were enriched in the sequences served to bolster the assumption that most if not all the VDR peak regions in both datasets were real binding sites. These HOT regions have also been noted for having significant overlaps with motifless binding peaks of other TFs (65) and thus our datasets were not different in this regard.

Analysis of the observed binding sites from motif scanning and data of known features for DNA regions of the human genome showed that the binding sites had a high propensity to overlap with regions of different histone modifications; H3K27ac (80.2% overlap), H3K4me1 (77.4% overlap) and H3K4me3 (75.7% overlap) (details provided in Appendix *A.1*) in sequences with VDR-like sites as compared to MEME-derived or classic VDR sites

(Table 17, Appendix *A.7*). The highest overlap was recorded in DNAse HotSpots (97.5%). These were not unexpected because these features and chromatin marks tend to be indicative of active regulatory elements (65) and were also marks that were enriched in the HOT and BAR regions. Of interest however is the fact that these features were predominant in the VDR-like sequences as compared to the others in the Ramagopalan dataset. In the case of the Heikkinen dataset, there were no significant differences observed between the different clusters in the VDR-positive sequences. This can be attributed to the fact that the MEME-derived motif was similar to the VDR-like motif and thus these clusters contained mostly the same sequences as was ascertained by comparing the sequences in the two clusters. This observation was also made for the classic *RXR-VDR* and the VDR-like cluster. Of importance in these analyses however is the fact that these genomic features were commonly enriched at the observed binding sites despite the disparity between the two datasets. H3K4me1 and H3K9ac are found usually in enhancer regions distal to the transcription start site whereas H3K4me3 and H3K27ac are indicative of active promoters. The presence of H3K4me1 marks is also thought to influence the recruitment of chromatin modifiers or pioneer factors (4, 80). The high enrichment of these marks particularly H3K4me1 compared to H3K4me3, suggested the probability of these regions being predominantly enhancer regions. The low overlap between sequences in the Ramagopalan dataset and those in the Heikkinen dataset also indicated some cell-type specificity and this specificity to cell types has been shown to be a feature of enhancers as compared to promoters which show activity in multiple cell types (81). This suggestion was however not in tandem with the earlier results which indicated a large majority of peak regions being in proximal positions with respect to TSSs (Table 16). Additional information in the form of H3K27me3 peak regions was thus added to the analysis feature set. H3K27me3 marks are associated with repressed regions and high frequency of occurrence of this mark together with about 50% observation frequencies for H3K4me1 and H3K4me3 and low occurrence frequencies for H3K27ac have been shown to be associated with inactive/poised promoter states (81). The results however showed relatively low occurrence of H3K27me3 marks in the whole dataset. To ascertain the region types present, the VDR datasets were compared to chromatin segmentation information derived by Hoffman, Ernst (82) which describes different chromatin states based on combinations of chromatin accessibility data and histone modifications learned by applying unsupervised learning methods. The results for this analysis showed the enrichment of active regions in the form of TSSs, flanking regions to TSSs and enhancers in the Ramagopalan dataset supporting the finding of proximal regulatory regions in this dataset. The results also further highlighted the difference between the two datasets in the analysis, with the Heikkinen dataset showing enrichment for repressed regions compared to the Ramagopalan dataset (Figure 13, Appendix *A.10*).



A

	Type	Total	Sequences
🟥	TSS	1665	1550 of 2776 (55%)
🟨	E	925	880 of 2776 (31%)
🟦	T	297	267 of 2776 (9%)
🟩	WE	272	232 of 2776 (8%)
■	R	161	145 of 2776 (5%)
■	MA0139	99	95 of 2776 (3%)
🟪	PF	37	35 of 2776 (1%)

B

	Type	Total	Sequences
🟥	R	1038	819 of 2338 (35%)
■	TSS	914	837 of 2338 (35%)
🟨	T	904	751 of 2338 (32%)
🟩	E	447	382 of 2338 (16%)
■	MA0139	195	182 of 2338 (7%)
🟦	WE	193	159 of 2338 (6%)
🟪	PF	35	33 of 2338 (1%)

Figure 13 Distribution of chromatin segments in the two analysis datasets using segmentation tracks. There is enrichment of TSSs and enhancers in (A) Ramagopalan dataset as compared to a higher overlap with repressed chromatin segments in (B) Heikkinen dataset. Segmentation track used was the consensus combination track of ChromHMM and Segway. TSS=Predicted promoter regions including TSSs, T=Predicted transcribed regions, WE=Weak enhancer regions, MA0139=CTCF enriched regions, PF=Promoter flanking regions, E=Predicted enhancer regions and R=Predicted repressed or low activity regions.

Assuming VDR binding to different genomic regions occurs in a time dependent manner, there seems to be a trend in the results which suggest that at the early stage (Heikkinen dataset), VDR has a preference for inactive gene regions with low levels of open chromatin and over time moves to active chromatin outside of genes in late stages (Ramagopalan dataset). This observation is consistent with the strong representation of repressive regions in the Heikkinen dataset as opposed to the Ramagopalan dataset but drawing substantive conclusions from this may be erroneous due to the differences between the datasets mentioned previously. There was low overlap with regions of repeat and CpG islands and the absence of these low complexity regions also lowers the probability that the observed binding sites could be spurious events.

Table 17 The percentage overlap of identified sites of motifs of interest with different sequence feature regions in the genome. The significance of these features in sequence collections based on these motifs is also shown. Result for the Heikkinen dataset is shown as Appendix A.7

Feature	Percent overlap of binding sites from motif scanning (PPV)		
	MEME *de novo*	JSVDR (*RXR-VDR*)	VDR-like (*NR4A2*)
DNAse HotSpots	96.9	94.8	97.5*/** (*p-value<0.001*)
TFBS_ChIP-Seq	78.7	69.4	91.3*/** (*p-value<0.0009*)
H3K27ac	59.1	49.2	80.2*/** (*p-value<0.05*)
H3K4me1	74.5	68.0	77.4** (*p-value<0.05*)
H3K4me3	53.0	44.1	75.7*/** (*p-value<0.05*)
H3K9ac	45.4	37	72.4*/**(*p-value<0.05)*
H3K27me3	29.8	40.7	15.5
FAIRE-Seq	73.4	65.6	78.8*/** (*p-value<0.003*)
RepeatMasker327	29.3	27.8	26.8
CpG islands	7.4	7.6	9.2
CCDS	32.3	34.5	29.1
Ensembl Genes	54.1	52.3	50.3

*Significant difference in region overrepresentation between MEME *de novo* and VDR-like, **Significant difference in region overrepresentation between classic VDR and VDR-like. These features were similar in MEME-motif sequences and the classic VDR sequences.

3.3. Differences in enriched DNA feature regions in VDR- positive versus negative sequences

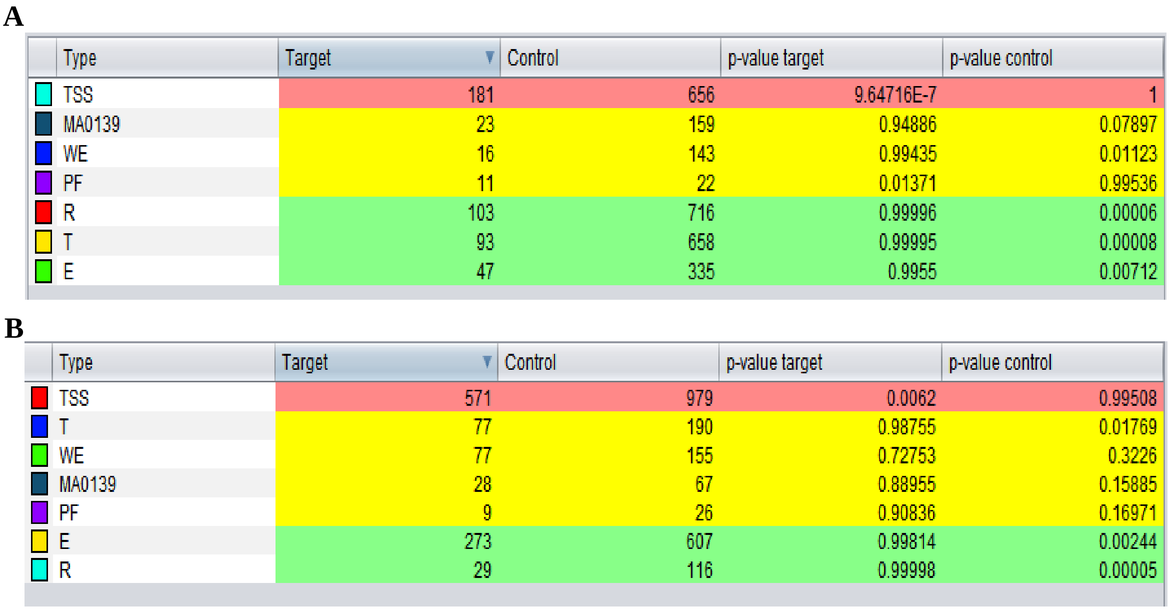
Having clustered the sequences into VDR- positive and negative (based on the Jaspar_Core profile for *RXR-VDR* and *NR4A2*), the sequences in each cluster were analysed for any similarities and differences. It was expected that VDR-positive sequences would have feature sets that distinguished the observed binding sites from their surrounding regions and thus influenced the binding of VDR to its response element. Knowing these features and their occurrence could thus be used to identify such regions in the negative set of sequences and consequently aid the identification of VDREs in them. This could be achieved by the generation of positional priors which would be used to guide motif discovery in the negative set. In addition, any differences unique to the negative set of sequences as compared to the positive could form the basis for explanation for the phenomenon of motifless binding or even form the basis for the identification of an erstwhile unknown motif.

Comparing the different regions of sequences in the VDR-positive and negative sets was not informative for DNAse hypersensitivity sites and FAIRE-Seq for the Gm12878 and K562 cell lines (cell lines comparable to those used to derive ChIP-Seq data for Ramagopalan and Heikkinen respectively), as well as for CCDS and Ensembl genes. These occurred at the same rate in both positive and negative sets. On the other hand, a repeat type (MIRb, *p-value=0.2e*$^{-}$

^{4} *and 6.3e^{-9} in Ramagopalan and Heikkinen respectively*) was overrepresented in the positive set of sequences whereas the negative set had higher GC-rich repeats (*p-value=8.3e^{-7} and 0.3e^{-4} in Ramagopalan and Heikkinen respectively*). The MIRb (Mammalian-wide Interspersed Repeat) is a member of the small interspersed nuclear element (SINE) family of repeats and as the name suggests is found in most mammalian genomes accounting for close to 2% of primate DNA (83). It has been suggested that MIRs located upstream of enhancers play a role in their regulation and also increase their activity (84). Analysis of the position of this repeat showed very little overlap with the binding sites of interest although majority of sites seemed to be within 100bp of the repeat on average.

There was also a difference in the types of histone modifications that were enriched in the two sequence sets; H3K27ac and H3K4me3 were enriched in the negative set (*p-value=0.018 and 0.008 respectively*) whereas H3K4me1 was enriched in the positive set (*p-value=4.7e^{-6}*). Inclusion of the MEME-derived motif in the definition of VDR-positive sequences yielded a similar trend of enrichment but with stronger significance values (*p-value=1.4e^{-4}, 4.2e^{-6}* and *4.7e^{-7}* for H3K27ac, H3K4me3 and H3K4me1 respectively). A similar observation was made for the Heikkinen dataset with the exception of H3K4me1 which was the same in positive and negative sequences. As mentioned previously, H3K4me3 has a role in chromatin remodelling by recruitment of modifiers and pioneer factors and hence it's overrepresentation in sequences without any identifiable VDR consensus motif may be informational with regards to this phenomenon of motifless binding. H3K27ac and H3K4me3 however, tend to be predominantly located in active promoter regions whereas H3K4me1 has a tendency for being in enhancer sites. These modifications have also been observed to show a bimodal distribution around most active transcription factor binding sites in TFBS regions in other datasets (85). This observation could however not be ascertained with regard to the analysis datasets as data for such an analysis was unavailable. Nevertheless, observations of enrichment of these marks at sites without the full consensus sequence for VDR binding could be indicative of a support mechanism which allows for weaker interactions. The *NR4A2* motif which is predominant in most of the sequences analysed corresponds to the *RXR* half-site of the VDR DR3-type motif as do some other motifs which are enriched from the background.

In addition, the analysis of the distribution of different chromatin segments in the different sequence clusters and how they compared to each other was also carried out. The results from this showed that sequences without the known VDR motif were predominantly in TSS regions whereas motif-positive sequences were enriched for enhancer as well as repressed regions (Figure 14). From these observations, it can be argued that TF binding to distal regulatory regions or low activity regions requires a motif that closely matches the consensus to facilitate stronger binding whereas much more flexibility may be allowed in regions proximal to genes. Another suggestion that can be put forward from these results is that VDR binds in these distal regulatory regions and by the mechanism of DNA looping participates in a complex with the transcription machinery and other TFs present at the promoter regions. This suggestion is likely because ChIP-Seq data only represents DNA segments where there are protein-DNA interactions and does not capture the manner of these interactions or interactions between fragments. Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) could be employed to these regions to test the suggestion.



A

	Type	Target	Control	p-value target	p-value control
🟦	TSS	181	656	9.64716E-7	1
🟦	MA0139	23	159	0.94886	0.07897
🟦	WE	16	143	0.99435	0.01123
🟪	PF	11	22	0.01371	0.99536
🟥	R	103	716	0.99996	0.00006
🟨	T	93	658	0.99995	0.00008
🟩	E	47	335	0.9955	0.00712

B

	Type	Target	Control	p-value target	p-value control
🟥	TSS	571	979	0.0062	0.99508
🟦	T	77	190	0.98755	0.01769
🟩	WE	77	155	0.72753	0.3226
🟦	MA0139	28	67	0.88955	0.15885
🟪	PF	9	26	0.90836	0.16971
🟨	E	273	607	0.99814	0.00244
🟦	R	29	116	0.99998	0.00005

Figure 14 Distribution of chromatin segmentation states in VDR-positive (Control) versus VDR-negative (Target) sequences as determined by a combined track derived from the ChromHMM and SegWay software on the (A) Heikkinen and (B) Ramagopalan datasets. TSS=Predicted promoter region including TSSs, T=Predicted transcribed regions, WE=Weak enhancer regions, MA0139=CTCF enriched regions, PF=Promoter flanking regions, E= Predicted enhancer regions and R=Predicted repressed or low activity regions. (Bonferroni corrected p-value threshold= 0.007, Red=overrepresented in target, Yellow=equally present and Green=overrepresented in control)

Analysis of the imported feature tracks was expected to give an indication of the sequence features of the observed VDR binding sites that were distinct from regions without VDR motif. The results of the region comparison showed no clear distinction or trend between actual binding sites and surrounding regions aside the fact that signals of some chromatin marks and features were stronger in motif-negative regions indicating that VDR binding to the classic VDRE may have some preference of sequence features. There is evidence to also suggest that motif-negative peaks are more likely to occur in regions close to TSSs than other regulatory regions.

3.4. DNA feature regions not discriminative for VDR/VDR-like binding sites

As part of the motif discovery process, additional information in the form of chromatin marks and conservation of bases have been shown to improve the efficiency of discovery algorithms and thus some algorithms are implemented to allow the incorporation of such information in the process. In the analyses datasets, it was observed that there was some but not much difference between motif positive and negative regions with respect to some of these features. Nevertheless, it is possible that these feature regions could be used to identify the observed binding sites by having high signal values in regions that overlap with the observed TF binding sites (or vice versa). Should that be the case, then these features could form the basis for positional priors either as individual priors or as combinations of features, depending on their predictive ability. An evaluation of the potentials of these features to discriminate binding sites was thus carried out.

From the ROC curve results shown in Figure 15 for the Ramagopalan dataset, the potential of the selected features to identify the VDR-like *NR4A2* had values no better than that from random guessing. Similar values were obtained for MEME-derived VDR and the classic VDR using H3K4me1 signals (Table 18). For very bad AUC values (H3K27ac and H3K4me3 for MEME-VDR and JSVDR), the signals were inverted using a reciprocal function to try and obtain better values, however this resulted in no appreciable improvement. Different combinations of these two signals were also unable to discriminate positive sites from surrounding regions. The results from the Heikkinen dataset were similar with none of the features able to select positive sites from the background (results not shown).

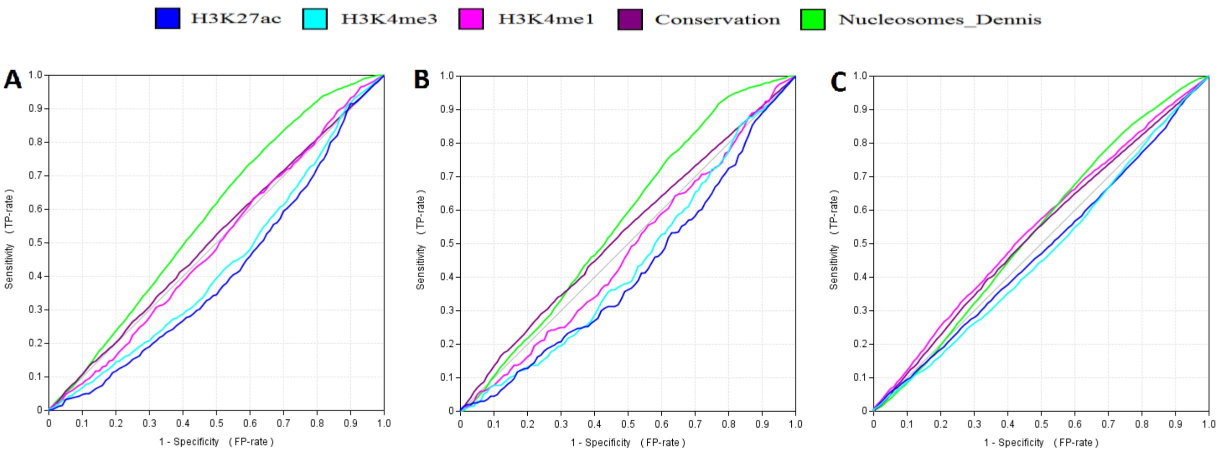


Figure 15 Receiver Operating Characteristic (ROC) curves showing the potential of selected DNA features to discriminate different positive binding sites from the background. Shown are curves for the observed sites for (A) the MEME-derived motif, (B) the Jaspar_Core profile for the VDR and (C) the VDR-like *NR4A2*. Values for AUC given in Table 18

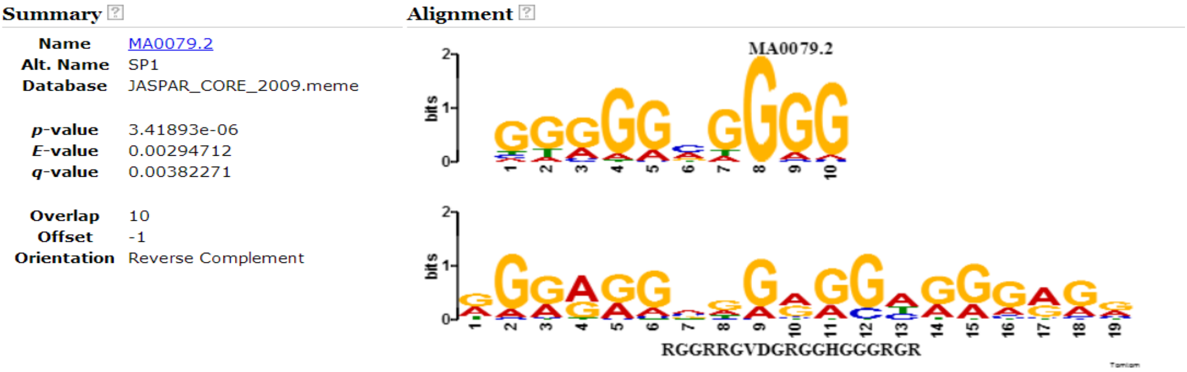
Table 18 Area under the curve values for the ROC curves shown in Figure 15 using DNA features and chromatin marks signals to discriminate positive binding sites from their surrounding background. Values in parenthesis are AUC values after the signals for these features were transformed using a reciprocal function to improve performance

Feature	AUC values for ROC curves		
	MEME-VDR	JSVDR	VDR-like (*NR4A2*)
H3K4me1	0.497	0.479	0.548
H3K4me3	0.437 (0.584)	0.443 (0.560)	0.472
H3K27ac	0.413 (0.602)	0.421(0.571)	0.481
H3K9ac	0.426 (0.608)	0.439 (0.591)	0.472
Conservation	0.512	0.533	0.531

3.5. Enrichment of *Sp1*-like motif in VDR-negative sequences

Despite the negative results from the preceding analysis, some of these features were used to guide the motif discovery process because they showed a high overlap with VDR-binding sites in the VDR-positive sequence sets. Properties that were pro-transactivation such as histone modifications (H3K4me3, H3K27ac, H3K9ac and H3K4me1), DNAse hypersensitivity sites, conservation and gene coding regions were used to generate positional priors using the inbuilt *Priors Generator*. The generated positional prior was used in *de novo* motif discovery using PRIORITY. Several combinations of the feature tracks were also used to derive positional priors which were subsequently used in *de novo* motif discovery. The generated priors were applied to the different subsets of sequences. Motif discovery by ChIPMunk was also used on the VDR-positive and negative sets. The motif discovery process is particularly important in identifying stretches of nucleotides that are consistently present in the analysis set and thus form the basis for extrapolation to a known motif or postulation of a new motif. Motif discovery using positional priors helps in the prediction of motif presence based on characteristics observed from confirmed binding sites.

Using the priors generated for *de novo* motif discovery yielded no general consensus motifs for all the different priors. This result could be attributed to the noisy nature of the priors that were derived based on the region comparisons carried out earlier (the lack of a difference and discriminative potential). Discovery with ChIPMunk and no priors however yielded a consensus motif that had high similarity (*p-value=3.4e*$^{-6}$, TOMTOM) to the *Sp1* motif in the negative set of sequences (Figure 16). Comparison of motif enrichment between VDR-positive and negative sequences also showed the *Sp1* motif as one of the overrepresented in the VDR-negative set (*p-value=1.48e*$^{-40}$, Appendix *A.3*). Heikkinen, Vaisanen (72) have in an earlier study reported identifying a *Sp1*-matching motif from *de novo* motif discovery in a subset of sequences from VDR ChIP-Seq in human monocytic leukaemia cells although there was no specific enrichment in sequences that lacked a VDR-like motif compared to those in which it was present. The fact that *Sp1* is enriched in both datasets, especially in VDR-negative sequences as compared to the positive sequences, suggests a role for *Sp1* in the DNA-binding activity of VDR in the absence of its classic response element. Earlier studies by Huang, Chen (86) have observed *in vitro* and *in vivo* interaction of the VDR with *Sp1* with suggestions that this interaction may modulate the expression of genes lacking the VDRE. Further studies have also proposed a model for this interaction where Sp1 serves as an anchor to which VDR binds and through its transactivation domain, recruit the transcription machinery to effect the regulation of genes (87).



Summary

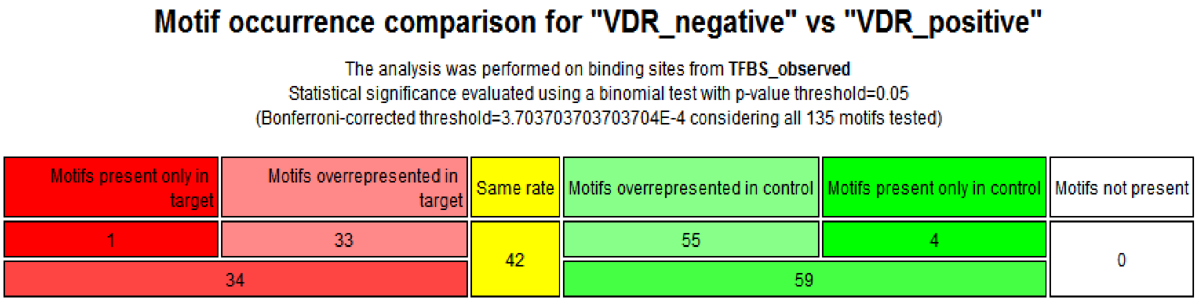
Name	MA0079.2
Alt. Name	SP1
Database	JASPAR_CORE_2009.meme
p-value	3.41893e-06
E-value	0.00294712
q-value	0.00382271
Overlap	10
Offset	-1
Orientation	Reverse Complement

Figure 16 *Sp1* motif identified by TOMTOM as most similar to motif discovered by ChIPMunk in *de novo* motif discovery in VDR-negative sequences

Using similar parameters on the positive set of sequences unsurprisingly resulted in a consensus representation that matched the classic *RXR-VDR* motif and MEME-VDR. Of interest though was the fact that these predicted sites did not always overlap the positive sites identified by motif scanning using the Jaspar_Core database or MEME-derived motif. The motif discovery process was to confirm the presence of the VDR motif and thus it was expected that the identified sites would be direct overlaps. Such differences can be attributed to the matrices and profiles created by the different discovery algorithms which may be different from that contained in the Jaspar_Core database. These identified sites could also be additional VDR binding sites not identified in the motif scanning process as a high threshold of 90% was applied as cut-off. This would not be surprising as it has been observed that multiple VDREs can exist in each region with suggestions that the synergistic activity of these binding sites helps increase the inductive efficiency of the TF(58).

3.6. VDR- positive and negative sequences have differences in enriched motifs

Having discovered no clear motif(s) in sequences without VDR sites present using the motif discovery process with priors, it was reasoned that there may be information gleaned from the motifs that were present in the different sets of sequences. VDR may have interacting partners present which facilitate its binding without a clear response element. A comparison of the motifs present in the two sets of sequences (positive versus negative) as well as the two clusters in the positive set (classic VDR versus VDR-like) showed that some motifs were overrepresented in some groups compared to others (Figure 17 and Figure 18).

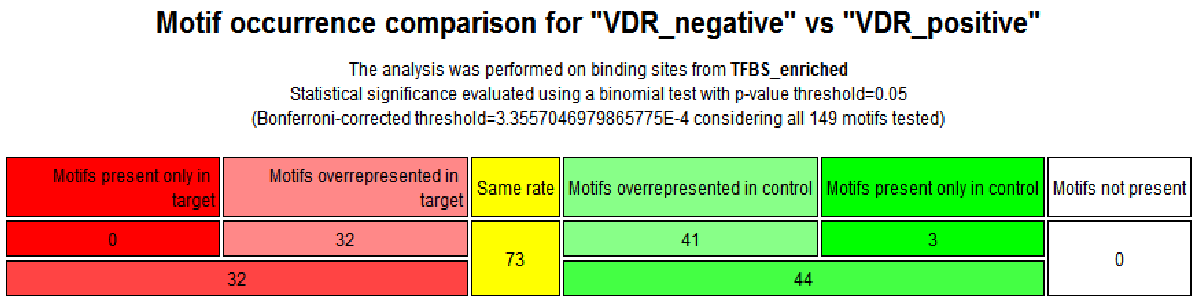


Motif occurrence comparison for "VDR_negative" vs "VDR_positive"

The analysis was performed on binding sites from **TFBS_observed**
Statistical significance evaluated using a binomial test with p-value threshold=0.05
(Bonferroni-corrected threshold=3.7037037037037037E-4 considering all 135 motifs tested)

Motifs present only in target	Motifs overrepresented in target	Same rate	Motifs overrepresented in control	Motifs present only in control	Motifs not present
1	33	42	55	4	0
34			59		

Figure 17 Overrepresentation of motifs in motif-negative (target) and positive (control) sequences in the Ramagopalan dataset.



Motif occurrence comparison for "VDR_negative" vs "VDR_positive"

The analysis was performed on binding sites from **TFBS_enriched**
Statistical significance evaluated using a binomial test with p-value threshold=0.05
(Bonferroni-corrected threshold=3.3557046979865775E-4 considering all 149 motifs tested)

Motifs present only in target	Motifs overrepresented in target	Same rate	Motifs overrepresented in control	Motifs present only in control	Motifs not present
0	32	73	41	3	0
32			44		

Figure 18 Overrepresentation of motifs in motif-negative (target) and positive (control) sequences in the Heikkinen dataset

Visual observation of the motifs overrepresented in the VDR-negative subset of sequences showed that the most significantly enriched of these motifs had stretches of single nucleotides with majority of these being either G or C. Comparing the GC-content of the motifs overrepresented in each also showed that there was a difference between the two sets (Appendix *A.12*). The preponderance of G and C nucleotides was not surprising when taken in the context of the high GC-content of the sequences. Indeed, motif discovery yielded a *Sp1*-like motif which has high GC-content. A point of interest in this observation is the homooligonucleotide stretches present in these motifs and by extension sequences. An earlier model developed by Sela and Lukatsky (88) to explain non-specific transcription factor-DNA binding affinity has suggested that these homooligonucleotide stretches increases the propensity of sequences for non-specific TF binding by lowering the binding free energy. This lowering of binding free energy allows TFs to bind in such regions and by the process of one-dimensional sliding along DNA, search for and bind to the cognate response element. Whether a similar conclusion can be drawn from our observations remains unclear although they further suggest that this principle is applicable genome-wide in yeast regulatory sequences especially in regions of high TF occupancy. Another possible explanation for the presence of these GC-rich motifs is the interaction of VDR with a TF that binds in such regions. Such a mechanism has been reported for some other nuclear receptors, including the estrogen receptor, where it forms a complex with *Sp1* and interacts with GC-rich motifs to induce the retinoic acid receptor alpha 1 gene (69).

3.7. Weaker VDR binding may be supported by other transcription factor activity/response elements

An interesting motif which was overrepresented in the VDR-like set was *h* (a basic Helix Loop Helix (bHLH) type factor). Activated VDR has been reported to exert transrepressive effects indirectly via interaction with a bHLH-type transcription factor bound to DNA (68). Additionally, chromosome opening activity has been reported for *TFE3*, a bHLH-type factor (80). A similar role may be indicated in this case. It was generally expected that there would be a difference in the motifs that were overrepresented from the background in the two clusters particularly with a higher enrichment of motifs for factors that would enhance VDR binding in the VDR-like set. The VDR-like motifs are similar to the *RXR* half-site of the classic *RXR-VDR* motif and thus may more likely represent weaker binding of the formed heterodimer. On the other hand, these single half sites could be binding sites for ligand-bound VDR monomer species. Although this suggestion may not be directly supported, observations by Cheskis and Freedman (11) have indicated the predominant presence of monomeric ligand-bound VDR on the DR3 element.

In addition, the enrichment of *Sp1*, a ubiquitous factor that participates in transactivation and repression and binds with high affinity to GC-rich motifs, in the VDR-like cluster of sequences could also indicate a role of opening up chromatin for transcription factor access to the weaker binding sites (89) thereby acting as a pioneer factor for VDR (60). Observations *in vitro* indicate that some transcription factors are unable to bind to target sequences in nucleosomal DNA and may require cooperative interaction with other factors (80, 90). This same argument could be applied for its presence in the VDR-negative set of sequences. The overrepresentation of *Sp1* could however be attributed to the GC-rich composition of the sequences (50.7% and 53.7% on average (Appendix *A.11*), compared to the human genome-wide average of 41% (91)), in which case it may not play any significant role. The choice of the 4th order model, coupled with using the original sequences as input for background model generation however ensures a higher likelihood that the simulated sequences are similar to the input sequences and thus have comparable GC content. Huang, Chen (86) have also described in an earlier study, the formation of a complex between *Sp1* and VDR that facilitates transactivation of genes without the classic VDRE in their promoter especially after ligand treatment. The functional importance of *Sp1* overrepresentation in these sequences, although uncertain, can therefore not be discounted.

Also, unliganded VDR is known to constitutively bind to DNA forming homodimers to stabilize the interaction and protect VDR from degradation. Activation by Vitamin D disrupts and minimizes this homodimerization, stabilizing the bound monomer and positively influencing the heterodimerization process (11). The presence of single half-sites in some sequences could thus be attributed to unliganded or ligand-bound VDR binding and would have to be analysed by comparison with data from unstimulated VDR to identify region overlaps particularly with the VDR motif-negative sequences. In addition, RXR has been observed to be bound to DNA sites in the absence of ligand, acting as markers for potential VDR binding after stimulation (92). The preponderance of these single half-sites could thus be attributed to this function as well but the heterodimerization of RXR with other members of the nuclear receptor family makes this a non-trivial problem. Heikkinen, Vaisanen (72) report that there is a genomic shift in VDR binding locations from non-DR3 to DR3-type response elements upon ligand stimulation. This could also indicate a potentially indiscriminate VDR-DNA interaction process in the absence of ligand and quite possibly in the ligand stimulated state.

3.8. VDR binding and cluster formation is Cohesin and CTCF independent

Cohesin is a complex structure formed by members of the Structural maintenance of chromosomes (*Smc*) family of proteins in addition to non-*Smc* units. It has been known to play a role in the cell cycle and recent studies have further indicated that aside this role it may also have a role in gene regulation by mediating long range DNA interactions, marking the formation of dense clusters of TFs as an anchor and strengthening the binding of TFs to motifs to which they have low affinity (93, 94). The presence of Cohesin has been also been shown to correlate to the presence of binding sites for the CCCTC binding factor (CTCF) which acts as an insulator separating active chromatin regions from non-active regions. In this analysis, it was hypothesised that sequence regions negative for the VDR or VDR-like motif would have some correlation with the presence of signals for Cohesin and or CTCF which could then be used in the search for motifs in the negative sets of sequences.

Majority of regions in sequences from the Ramagopalan dataset when compared with Cohesin (using ChIP-Seq data for the *Smc3* subunit of the Cohesin complex) and CTCF regions showed very low coverage with classic VDR, VDR-like and VDR-negative sequences having greater than 50% coverage in regions without both Cohesin and CTCF (Table 19). The low coverage by CTCF regions was expected as results from the earlier motif scanning had shown that although the motif for CTCF was enriched, it only occurred in 2% of all the sequences (less than 4% in the Heikkinen dataset) and occurred at the same rate in both motif- positive and negative sequences. The low coverage by Cohesin was however not expected as CTCF-depleted Cohesin regions have been strongly correlated with active enhancer elements (95) and from the earlier analyses carried out, there was evidence suggesting the substantial presence of enhancers. In addition, the observation that a high number of TFs were seemingly overrepresented (from the results of motif scanning) and were closely located in the sequences suggested the possibility of cis regulatory modules formed by different clusters of these TFs. These clusters may or may not be functionally related, however, Faure, Schmidt (96) have in an earlier study described the stabilization of CRMs formed by dense clusters of TFs by Cohesin in the absence of CTCF, and this is believed to help regulate expression in a tissue specific manner in addition to mediating long range looping of DNA. The co-localization of Cohesin and TFs in these sequences would thus have supported the idea that VDR acts in a synergistic manner with varying combinations of these TFs to exert its regulatory control on the target genes as it has also been suggested that the function of CRMs may be achieved through molecules such as Cohesin (97). In addition, the stabilization offered by Cohesin has been suggested to allow TFs to bind to motifs with less similarity to their canonical motifs (96) and this would have been applicable to findings in this analysis. This was however not apparent from the analysis carried out as only 11% of VDR-negative sequences had coverage with CTCF-independent Cohesin.

There was a somewhat higher coverage in regions with both CTCF and Cohesin than in regions with either one or the other (Table 19) in the different sequence collections. CTCF is known to recruit Cohesin (94) serving as a positional marker and this co-localization was not surprising. CTCF is also a marker that segregates euchromatin from heterochromatin by binding in insulator regions and preventing the spread of one type of chromatin state to the other. This action accounts for some of the differences observed in the expression patterns of different cell types and also the differences between cells in different states. The overall low levels of CTCF coverage in these regions may be indicative of the open chromatin state lending credence to the view that these regions are actual receptor binding regions. High levels of CTCF coverage would have however supported the DNA-looping of intermediate DNA between distal regulatory regions and their target promoters as such a role has been attributed to CTCF as well (13, 93).

The analysis was not performed for sequences from the Heikkinen dataset as region information for Cohesin (the Smc3 subunit) in the cell line from which the data was derived that is comparable with that used for the Ramagopalan dataset was unavailable.

Table 19 Overlap of sequences in VDR-positive, VDR-negative and VDR-like collections with regions of Cohesin and CTCF binding obtained from ChIP-Seq. Cohesin and CTCF region datasets processed with BEDtools to obtain intersecting regions and exclusive regions

Region combination	% of sequences with coverage in collection		
	Classic VDR (n=192)	VDR-like (n=1618)	VDR-negative (n=966)
Cohesin^{-}/CTCF^{-}	78.5	54.9	58.3
Cohesin^{-}/CTCF	6.2	10.9	11.3
Cohesin/CTCF	10.4	19.4	18.8
Cohesin/CTCF^{-}	4.7	14.6	11.5

Cohesin and CTCF negative regions (Cohesin^{-}/CTCF^{-}), only CTCF-positive (Cohesin^{-}/CTCF), Cohesin and CTCF positive (Cohesin/CTCF) and only Cohesin-positive (Cohesin/CTCF^{-})

3.9. VDR ChIP-Seq regions not enriched for direct interaction partners

TF’s interacting in a combinatorial manner is said to be crucial for the expression of genes in a temporo-spatial and tissue and/or cell specific manner (73). From the preponderance of TF motifs indicated as enriched in the sequences, it was hypothesised that there could be in this collection, a set of motifs for interacting partners for the VDR, which together with it may form CRMs or even serve as anchors to which the VDR binds as described for the VDR interacting repressor.

Using data from (73) where an atlas of interacting transcription factor proteins in mouse and man was created, a list of 17 TFs was identified as potential candidates based on a criteria of physical interaction observed in the study (Table 20).

Table 20 List of genes for TFs with which VDR is suggested or shown to interact with

Gene 1 ID	Gene 1 Symbol	Gene 2 ID	Gene 2 Symbol
1387	CREBBP	7421	VDR
2959	GTF2B	7421	VDR
5927	JARID1A	7421	VDR
3725	JUN	7421	VDR
5469	PPARBP	7421	VDR
6256	RXRA	7421	VDR
6258	RXRG	7421	VDR
4088	SMAD3	7421	VDR
6772	STAT1	7421	VDR
7421	VDR	7421	VDR
7421	VDR	7704	ZBTB16
7421	VDR	8202	NCOA3
7421	VDR	8648	NCOA1
7421	VDR	10499	NCOA2
7421	VDR	22938	SNW1
7421	VDR	8805	TRIM24

7421	VDR	55806	HR
7421	VDR	23054	NCOA6

Comparing this list to the collection of enriched motifs common to both datasets resulted in nought as the motifs for these interacting partners were not enriched in both datasets. Of note though was the fact that CREB (cyclic AMP-responsive element-binding protein), a TF which is bound by CREBBP (CREB binding protein) was quite enriched in both datasets with a high occurrence in both datasets' sequences (56.4% and 37.2% in Heikkinen and Ramagopalan datasets respectively). This overrepresentation was however stronger in VDR and VDR-like motif-containing sequences than in the motif-negative ones. CREB and CREBBP can therefore not be considered as TFs to which VDR binds to in the absence of its known response element. There was also the presence of members of the MED-1 co-activator complex (*Esrrb*, *HNF4A*, *NR2F1* and *NR4A2*) of which the VDR forms a part.

3.10. Identification of putative CRMs modulating VDR action

Cis regulatory modules as mentioned previously play a vital role in the spatiotemporal regulation of genes. The enrichment of several TFBS in the two analysis datasets necessitates finding out if there is an overrepresentation of motif pairs in sequences without an observable VDR motif. Supporting this need was the fact that there was an enrichment of DNAse hypersensitivity and H3K9ac marks in regions without the classic DR3-type VDR motif. These marks are known to be associated with CRMs (97).

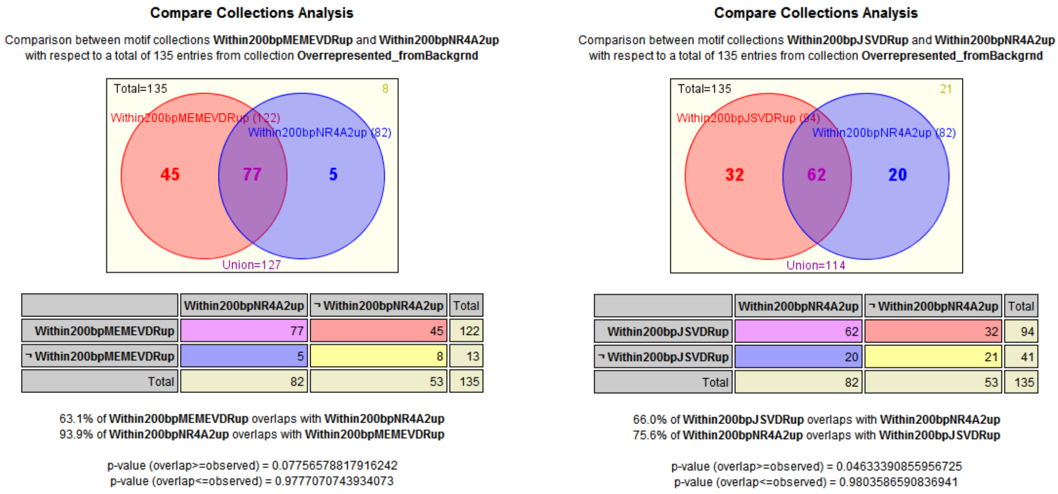
In the VDR-negative set of sequences particularly, it would have been thought that these clusters of TFs forming putative CRMs may contain interacting partners for VDR which would obviate the need for VDREs. This was however not the case and thus motifs which were enriched in the dataset as well as in the negative sequence set as compared to the positive formed the basis for identifying CRMs. The ModuleSearcher program on module discovery in the negative sequence set reported several modules with different combinations of TF motifs. Of interest in these motifs was the fact that the *Ets-1* motif was present in majority of modules identified in the Ramagopalan dataset. This was not the case for the Heikkinen dataset as this motif was not overrepresented in it. Ets-1 has been reported to act as an activator of some nuclear receptors including VDR in the absence of ligand (98) but it does appear that this occurred in the presence of a VDRE. It was also difficult to generate modules based on interaction partners for the enriched motifs in MotifLab as the module collection came back empty anytime the procedure was performed.

Interestingly; there was a difference in results between the two datasets with regard to the CRM-suggestive H3K9ac marks. Comparing the MEME-derived binding sites and the JASPAR core motif sites to the VDR-like (NR4A2) sites in the Ramagopalan dataset showed a significant enrichment (p-value=e^{-12} and e^{-9} respectively) of the H3K9ac marks in the VDR-like sites. This was not the case in the Heikkinen dataset and this can be attributed to the fact that there were only a small fraction of DR3-type VDREs in this dataset. This observation however serves to support the view that there is some subtle difference in binding mechanisms employed in these two cases; where there is the classic DR3 element and where it is absent.

3.11. Motif position distribution

Regulation of transcription is a process that involves several regulatory factors working in concert. These factors may or may not be in close proximity to each other. However factors that are involved in specific pathways of transcription tend to be located some average distance away from other interacting partners. The proximity of other binding sites which may not be interacting partners can also give useful information especially when they are factors that are known to contribute positively or negatively to transcription factor binding. Establishing how these different transcription factor sites are distributed in relation to each other or to a specified locus also helps to identify those that may be involved in the same process.

Analysis of the Ramagopalan data showed that sequences with the MEME-derived motif had the largest number of other motifs within the 200bp region upstream of the identified binding sites (n=122) whereas only 82 motifs were within the same region for the VDR-like sites. Figure 19 below shows the amount of overlap between the classic VDRE (MEME-derived and Jaspar_Core motif) and VDR-like sites.



Compare Collections Analysis

Comparison between motif collections **Within200bpMEMEVDRup** and **Within200bpNR4A2up** with respect to a total of 135 entries from collection **Overrepresented_fromBackgrnd**

	Within200bpNR4A2up	¬ Within200bpNR4A2up	Total
Within200bpMEMEVDRup	77	45	122
¬ Within200bpMEMEVDRup	5	8	13
Total	82	53	135

63.1% of **Within200bpMEMEVDRup** overlaps with **Within200bpNR4A2up**
93.9% of **Within200bpNR4A2up** overlaps with **Within200bpMEMEVDRup**

p-value (overlap>=observed) = 0.07756578817916242
p-value (overlap<=observed) = 0.9777070743934073

Compare Collections Analysis

Comparison between motif collections **Within200bpJSVDRup** and **Within200bpNR4A2up** with respect to a total of 135 entries from collection **Overrepresented_fromBackgrnd**

	Within200bpNR4A2up	¬ Within200bpNR4A2up	Total
Within200bpJSVDRup	62	32	94
¬ Within200bpJSVDRup	20	21	41
Total	82	53	135

66.0% of **Within200bpJSVDRup** overlaps with **Within200bpNR4A2up**
75.6% of **Within200bpNR4A2up** overlaps with **Within200bpJSVDRup**

p-value (overlap>=observed) = 0.04633390855956725
p-value (overlap<=observed) = 0.9803586590836941

Figure 19 Overlap of enriched motifs within the 200bp upstream regions of identified VDR and VDR-like binding sites

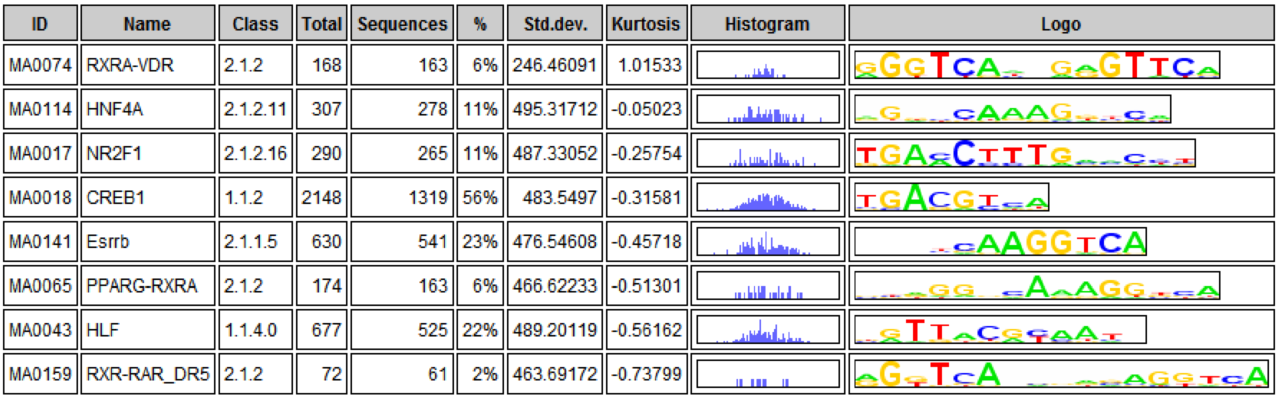
The 5 motifs that were uniquely present in Ramagopalan regions with *NR4A2* binding as compared to the MEME identified sites however, were spread out over average distances between 44 and 188 base pairs from the *NR4A2* sites. In addition, all but one had more than 400 sites in the collection of 1741 sequences representing those with the *NR4A2* sites. Furthermore, this motif for the hunchback factor (Hb), a *Drosophila melanogaster* protein, had no similarity to any of the VDR-like motifs (*Average log-likelihood ratio score* < -*5.460*). A similar observation was made for the 20 motifs unique to *NR4A2* binding sites compared to the Jaspar_Core-derived VDR sites.

In the case of the Heikkinen dataset, it was interesting to note that there was a preponderance of members of the nuclear receptor family within 200bp upstream region of the *NR4A2* sites (Table 21). These were the closest to the *NR4A2* sites and also enriched in the Ramagopalan dataset. The value of 200bp was chosen to approximate the size of a nucleosome and its linker regions (146bp DNA wrapped around the histone octamer) (100). Figure 20 shows the positional distribution of these factors over the length of sequences in which they are found.

Table 21 Enriched motifs 200bp from *NR4A2* sites and their average distance upstream from the *NR4A2* site

ID	Short Name	Classification	Average Distance
MA0141	Esrrb	2.1.1.5	87.4
MA0159	RXR-RAR_DR5	2.1.2	171.6
MA0065	PPARG-RXRA	2.1.2	171.9
MA0017	NR2F1	2.1.2.16	180.7
MA0018	CREB1	1.1.2	181.2
MA0074	RXRA-VDR	2.1.2	182.3
MA0114	HNF4A	2.1.2.11	191.4

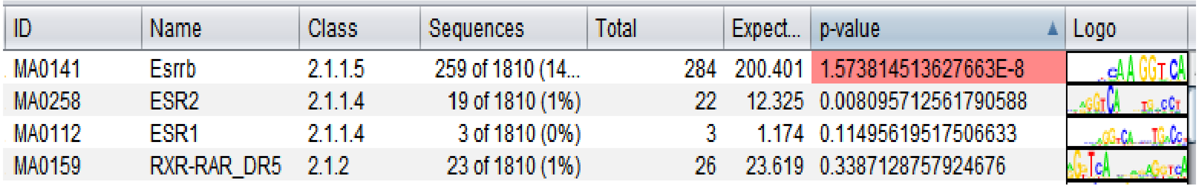
MA0043	HLF	1.1.4.0	198.0



ID	Name	Class	Total	Sequences	%	Std.dev.	Kurtosis	Histogram	Logo
MA0074	RXRA-VDR	2.1.2	168	163	6%	246.46091	1.01533		
MA0114	HNF4A	2.1.2.11	307	278	11%	495.31712	-0.05023		
MA0017	NR2F1	2.1.2.16	290	265	11%	487.33052	-0.25754		
MA0018	CREB1	1.1.2	2148	1319	56%	483.5497	-0.31581		
MA0141	Esrrb	2.1.1.5	630	541	23%	476.54608	-0.45718		
MA0065	PPARG-RXRA	2.1.2	174	163	6%	466.62233	-0.51301		
MA0043	HLF	1.1.4.0	677	525	22%	489.20119	-0.56162		
MA0159	RXR-RAR_DR5	2.1.2	72	61	2%	463.69172	-0.73799		

Figure 20 Positional distribution analysis of motifs enriched in the Heikkinen dataset showing how these motifs are distributed along the lengths of the sequences in which they are found. All sequences included in the analysis are are centred with respect to each other.

Extending the region around the observed VDR and VDR-like binding sites in the Ramagopalan dataset by 20 bases and scanning for any motifs reported as overrepresented in the positive set showed the estrogen related receptor beta (*Esrrb*) as being predominant in this region (*p= 1.57e*$^{-8}$, Figure 21). This is consistent with observations made by Tuoresmaki, Vaisanen (101) where they observed *ESSRB* as one of the most enriched motifs in the ±100bp region of VDR DR3-positive peaks. This motif bears similarity with the *RXR*-half site and is one of the closest to the *NR4A2* motif, one of the VDR-like motifs (Figure 21). The similarity between the two half-sites of the classic *RXR-VDR* could mean an alternative albeit weaker form of binding using motifs with similarity to these half-sites which are in close proximity. This motif was seen to be enriched in the VDR-positive collection as compared to the negative as well (Appendix *A.3*). Another point of note about these motifs is that the profile TGACC and its reverse complement seem to be common among them. Whether these motifs can serve as half-sites or alternative binding sites for the VDR may be a possibility but their number of sites present and distances to the VDR-like *NR4A2* motif makes this speculative.



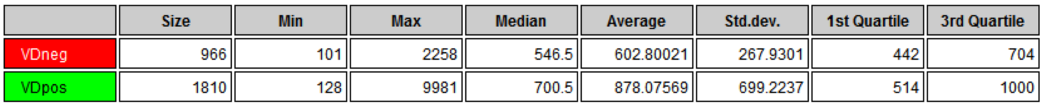
ID	Name	Class	Sequences	Total	Expect...	p-value	Logo
MA0141	Esrrb	2.1.1.5	259 of 1810 (14...	284	200.401	1.573814513627663E-8	
MA0258	ESR2	2.1.1.4	19 of 1810 (1%)	22	12.325	0.008095712561790588	
MA0112	ESR1	2.1.1.4	3 of 1810 (0%)	3	1.174	0.11495619517506633	
MA0159	RXR-RAR_DR5	2.1.2	23 of 1810 (1%)	26	23.619	0.3387128757924676	

Figure 21 Enrichment of *Esrrb* in VDR-positive sequences in 20bp region surrounding binding sites

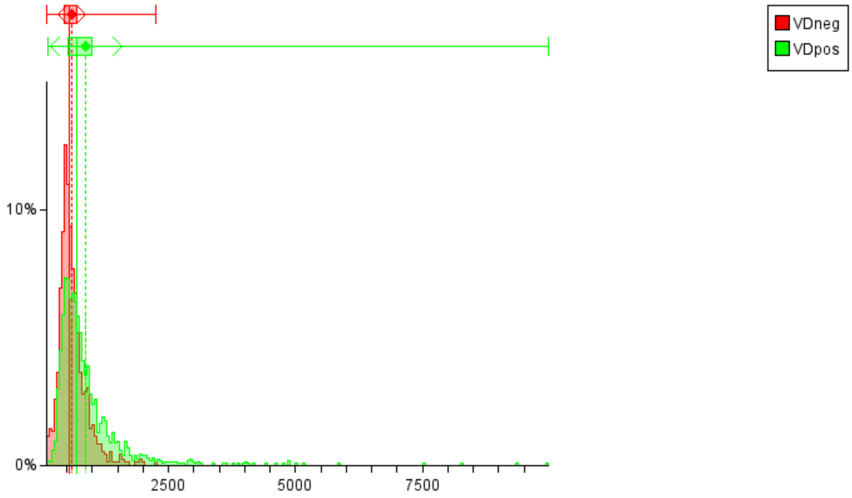
Having observed *Sp1* as being overrepresented in the VDR-like, as well as VDR-negative sequence collections in the Ramagopalan dataset, the relative distances of other motifs from *Sp1* were calculated and it was noted that the two VDR motifs (*RXR-VDR and NR4A2*) were in close proximity with average distances of 92, 176 bases respectively, supporting the view that its role in recruiting chromatin remodelling factors may play a part in weaker VDR binding. It was also expected that VDR-related and or VDR interaction partners would be found in the region of *Sp1* in the negative set of sequences as well. This was however not the case as these factors were absent from the set of binding sites that were overrepresented in the VDR-negative sequence set.

Another interesting observation that was made in both datasets used for the analysis related to the lengths of sequences in the different clusters. It was noted that sequences which were positive for either VDR or any of the VDR-like motifs tended to be longer than their counterparts without any of these motifs particularly in the Heikkinen dataset. This was analysed by plotting the base count of the sequences in each cluster against the number of sequences with the same base count as shown in Figure 22 below. With the binding sites of TFs in these sequences not centrally enriched, this observation raises the question of whether these VDR-negative sequences are representative of the actual binding sites *in vivo* as some information may have been lost in the processing steps after sequencing. Results from the earlier analyses are however firmly indicative of these sites being real binding regions.

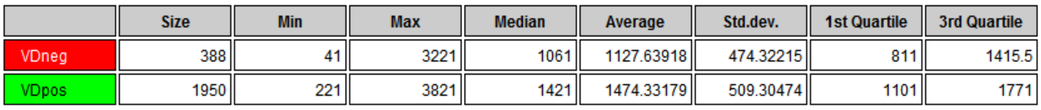
A



	Size	Min	Max	Median	Average	Std.dev.	1st Quartile	3rd Quartile
VDneg	966	101	2258	546.5	602.80021	267.9301	442	704
VDpos	1810	128	9981	700.5	878.07569	699.2237	514	1000



B



	Size	Min	Max	Median	Average	Std.dev.	1st Quartile	3rd Quartile
VDneg	388	41	3221	1061	1127.63918	474.32215	811	1415.5
VDpos	1950	221	3821	1421	1474.33179	509.30474	1101	1771

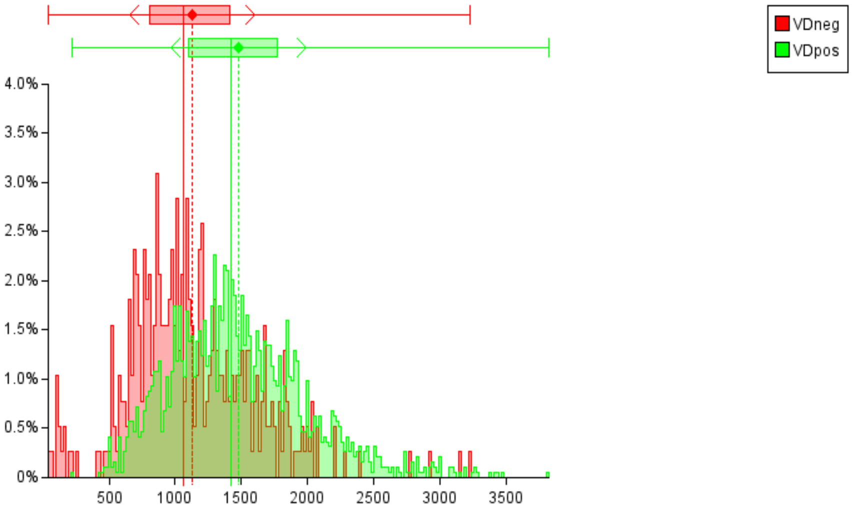


Figure 22 Distribution of sequence lengths in VDR- positive and negative clusters in analysis datasets showing positive sequences as being predominantly longer than the negative counterparts. (A) represents the distribution in the Ramagopalan dataset and (B) the distribution in the Heikkinen dataset. X-axis represents the length of sequences and the Y-axis represents the fraction of sequences with the corresponding length. Value range divided into 200 bins

4. Conclusion

By the computational analysis of motif-negative binding using ChIP-Seq datasets for the Vitamin D receptor, there has been observed, differences in the enrichment of features between motif-positive regions and their motif-negative counterparts suggesting the existence of some influencing mechanisms for motifless binding. In terms of DNAse hypersensitivity sites, FAIRE-Seq regions, regions of coding genes (CCDS), Ensembl genes and CpG islands, no differences were observed between the two regions. These features are known indicators of regulatory regions and this lack of difference thus minimizes the possibility of the motif-negative regions being spurious regions included in the ChIP-Seq data as a result of errors in processing or other factors that influence the ChIP-Seq process. It was also shown that there was an enrichment of motifs for factors that matched one of the half-sites for the *RXR-VDR* heterodimer motif however no other motifs were found in proximity to these to suggest a potential heterodimerization partner for the VDR. Indeed, *de novo* motif discovery using prior information yielded no consensus motif and even when no prior information was utilized to guide the process, only a motif with close similarity to the *Sp1* motif was identified. Seeing as the *Sp1* motif was enriched in both datasets and it has been documented to play a role in motif-negative binding of other nuclear receptors, it is quite probable that it may play a role in the binding of VDR in these motif-negative regions. In addition, despite the observation that majority of the VDR peaks corresponded to what has been termed HOT regions, regions where motif-negative interactions have been reported, no singular motif or collection of motifs were shown to be enriched in the VDR motif negative regions to support the suggestion of VDR binding via interaction with a factor whose motif is present either as a dimer or cis regulatory module.

In conclusion, although several mechanisms have been suggested as explanation for the interaction of VDR with motif-negative sequences, none is able to encompass all the motif-negative regions.

4.1. MotifLab as a workbench for regulatory region analysis

The utility of MotifLab as a workbench for this analysis cannot be over emphasised. From the get go, the graphical interface provided an aesthetically pleasing appeal without a lot of clutter. The labels given to tools were intuitive allowing one to easily guess what operation can be performed with the tool. This was further enhanced by balloons which provided brief descriptions of operations. This was applicable to the parameters for each operation as well and thus made it easy to know what inputs were allowed for the specific operation. In the dialog for each operation was a help button which provided detailed information of the operation and its parameters. This functionality was however dependent on being connected to the internet although most operations were easily performed without the internet. The exception to this was when data from the UCSC browser or other web-based sources had to be loaded. One important feature of MotifLab that proved invaluable was the ability to record all the analysis steps in a protocol script that could be saved and easily loaded again to perform the same analysis on a different dataset. The whole session could also be saved and loaded to continue the analysis at a later date. A note of caution however is that the protocol script must be saved prior to saving the session as any unsaved changes in the script do not show up when the saved session is next loaded. While working on a session, it was also possible to load and execute different saved or previously prepared protocol scripts. I found this handy as it was possible to create scripts for different analyses which were loaded and executed as and when it was required without losing any data from any previous analysis carried out. The results from analysis could also be visualised within MotifLab and also exported in different formats for further analysis or for use in a write-up.

The downside to MotifLab in my case was the memory requirement as I found out that anything less than 1GB was quite insufficient and resulted in operations taking exceedingly long times to complete. Running a full protocol script on large dataset (>2000 sequences on 2GB RAM) was unadvisable. This challenge seems to have been foreseen by the developer and thus portions instead of the whole script can be executed. A click on the memory usage indicator in MotifLab also seems to free up some space. In addition, MotifLab allows the user to specify the amount of RAM to allocate to the application before it starts in the standalone version. Although full documentation of all the procedures was not available at the time of use, most commonly used procedures are catered for on the MotifLab website and there are also tutorial videos for performing several types of analyses.

4.2. Suggestions for future work

For future work, I would first suggest a time series ChIP-Seq experiment in the same cell line to ascertain if VDR binding to the classic VDRE is time-dependent and also determine the proportion of peaks which are positive for the motif over the time course. Additionally, the application of liquid chromatography followed by mass spectrometry of the wash after pull down and reversal of crosslinks during ChIP may be informational with regard to which other proteins are pulled down with VDR. To carry on from this analysis, a comparison of motif-positive ChIP-Seq peaks of TFs for which motifs have been identified as being enriched in both datasets and the motif-negative VDR peaks may indicate putative factors with which VDR may act combinatorially either by direct interactions or through a shared interaction partner or complex. Finally, it has been suggested that non-B DNA associates with TFs to regulate transcription (102) and this could be an interesting line of analysis to explain the phenomenon of motifless binding.

References

1. Walhout AJM. Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping. Genome Res. 2006;16(12):1445-54.

2. Maston GA, Evans SK, Green MR. Transcriptional Regulatory Elements in the Human Genome. Annual Review of Genomics and Human Genetics. 2006;7(1):29-59.

3. Kouzarides T. Chromatin modifications and their function. Cell. 2007;128(4):693-705.

4. Bannister AJ, Kouzarides T. Regulation of chromatin by histone modifications. Cell Res. 2011;21(3):381-95.

5. Jaenisch R, Bird A. Epigenetic regulation of gene expression: how the genome integrates intrinsic and environmental signals. Nat Genet. 2003;33 Suppl:245-54.

6. Eberharter A, Becker PB. Histone acetylation: a switch between repressive and permissive chromatin - Second in review series on chromatin dynamics. Embo Rep. 2002;3(3):224-9.

7. Latchman DS. Transcription factors: an overview. Int J Biochem Cell Biol. 1997;29(12):1305-12.

8. Kadonaga JT. Regulation of RNA polymerase II transcription by sequence-specific DNA binding factors. Cell. 2004;116(2):247-57.

9. Mitchell PJ, Tjian R. Transcriptional regulation in mammalian cells by sequence-specific DNA binding proteins. Science. 1989;245(4916):371-8.

10. Filtz TM, Vogel WK, Leid M. Regulation of transcription factor activity by interconnected post-translational modifications. Trends Pharmacol Sci. 2014;35(2):76-85.

11. Cheskis B, Freedman LP. Ligand modulates the conversion of DNA-bound vitamin D3 receptor (VDR) homodimers into VDR-retinoid X receptor heterodimers. Mol Cell Biol. 1994;14(5):3329-38.

12. Sone T, Kerner S, Pike JW. Vitamin D receptor interaction with specific DNA. Association as a 1,25-dihydroxyvitamin D3-modulated heterodimer. J Biol Chem. 1991;266(34):23296-305.

13. Lee BK, Iyer VR. Genome-wide studies of CCCTC-binding factor (CTCF) and cohesin provide insight into chromatin structure and regulation. J Biol Chem. 2012;287(37):30906-13.

14. Schleif R. DNA looping. Annu Rev Biochem. 1992;61:199-223.

15. Petrascheck M, Escher D, Mahmoudi T, Verrijzer CP, Schaffner W, Barberis A. DNA looping induced by a transcriptional enhancer in vivo. Nucleic Acids Res. 2005;33(12):3743-50.

16. Jeziorska DM, Jordan KW, Vance KW. A systems biology approach to understanding cis-regulatory module function. Semin Cell Dev Biol. 2009;20(7):856-62.

17. Blackwood EM, Kadonaga JT. Going the Distance: A Current View of Enhancer Action. Science. 1998;281(5373):60-3.

18. Kolchanov NA, Merkulova TI, Ignatieva EV, Ananko EA, Oshchepkov DY, Levitsky VG, et al. Combined experimental and computational approaches to study the regulatory elements in eukaryotic genes. Briefings in bioinformatics. 2007;8(4):266-74.

19. Nieuwlandt D. In vitro selection of functional nucleic acid sequences. Curr Issues Mol Biol. 2000;2(1):9-16.

20. Stoltenburg R, Reinemann C, Strehlitz B. SELEX-A (r)evolutionary method to generate high-affinity nucleic acid ligands. Biomol Eng. 2007;24(4):381-403.

21. Wang J, Lu J, Gu G, Liu Y. In vitro DNA-binding profile of transcription factors: methods and new insights. J Endocrinol. 2011;210(1):15-27.

22. Galas DJ, Schmitz A. DNAse footprinting: a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978;5(9):3157-70.

23. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132(2):311-22.

24. Hampshire AJ, Rusling DA, Broughton-Head VJ, Fox KR. Footprinting: a method for determining the sequence selectivity, affinity and kinetics of DNA-binding ligands. Methods. 2007;42(2):128-40.

25. Crawford GE, Holt IE, Whittle J, Webb BD, Tai D, Davis S, et al. Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS). Genome Res. 2006;16(1):123-31.

26. Song L, Crawford GE. DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harbor protocols. 2010;2010(2):pdb prot5384.

27. Boyle AP, Song L, Lee BK, London D, Keefe D, Birney E, et al. High-resolution genome-wide in vivo footprinting of diverse transcription factors in human cells. Genome Res. 2011;21(3):456-64.

28. Furey TS. ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions. Nature reviews Genetics. 2012;13(12):840-52.

29. Johnson DS, Mortazavi A, Myers RM, Wold B. Genome-wide mapping of in vivo protein-DNA interactions. Science. 2007;316(5830):1497-502.

30. Alberts B. Molecular biology of the cell. 5th ed. New York: Garland Science; 2008.

31. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, et al. Genome-Wide Location and Function of DNA Binding Proteins. Science. 2000;290(5500):2306-9.

32. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. Nucleic Acids Res. 2008;36(16):5221-31.

33. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature methods. 2007;4(8):651-7.

34. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nature reviews Genetics. 2009;10(10):669-80.

35. Satoh J-i, Tabunoki H. Molecular network of chromatin immunoprecipitation followed by deep sequencing-based vitamin D receptor target genes. Multiple Sclerosis Journal. 2013;19(8):1035-45.

36. Marchal K, Thijs G, De Keersmaecker S, Monsieurs P, De Moor B, Vanderleyden J. Genome-specific higher-order background models to improve motif detection. Trends Microbiol. 2003;11(2):61-6.

37. Wingender E, Dietze P, Karas H, Knuppel R. TRANSFAC: A database on transcription factors and their DNA binding sites. Nucleic Acids Res. 1996;24(1):238-41.

38. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res. 2004;32(Database issue):D91-4.

39. Cora D, Di Cunto F, Provero P, Silengo L, Caselle M. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs. BMC Bioinformatics. 2004;5:57.

40. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16(1):16-23.

41. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nature reviews Genetics. 2004;5(4):276-87.

42. Colnot S, Lambert M, Blin C, Thomasset M, Perret C. Identification of DNA-Sequences That Bind Retinoid-X-Receptor-1,25(Oh)(2)D-3-Receptor Heterodimers with High-Affinity. Mol Cell Endocrinol. 1995;113(1):89-98.

43. Cuellar-Partida G, Buske FA, McLeay RC, Whitington T, Noble WS, Bailey TL. Epigenetic priors for identifying active transcription factor binding sites. Bioinformatics. 2012;28(1):56-62.

44. Hannenhalli S. Eukaryotic transcription factor binding sites - modeling and integrative search methods. Bioinformatics. 2008;24(11):1325-31.

45. Consortium EP, Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.

46. Ramsey SA, Knijnenburg TA, Kennedy KA, Zak DE, Gilchrist M, Gold ES, et al. Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. Bioinformatics. 2010;26(17):2071-5.

47. Ernst J, Plasterer HL, Simon I, Bar-Joseph Z. Integrating multiple evidence sources to predict transcription factor binding in the human genome. Genome Res. 2010;20(4):526-36.

48. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15(10):1451-5.

49. Neron B, Menager H, Maufrais C, Joly N, Maupetit J, Letort S, et al. Mobyle: a new full web bioinformatics framework. Bioinformatics. 2009;25(22):3005-11.

50. Klepper K, Drablos F. MotifLab: a tools and data integration workbench for motif discovery and regulatory sequence analysis. BMC Bioinformatics. 2013;14:9.

51. Grober U, Spitz J, Reichrath J, Kisters K, Holick MF. Vitamin D: Update 2013: From rickets prophylaxis to general preventive healthcare. Dermatoendocrinol. 2013;5(3):331-47.

52. DeLuca HF. Overview of general physiologic features and functions of vitamin D. Am J Clin Nutr. 2004;80(6 Suppl):1689S-96S.

53. Holick MF. Vitamin D deficiency. N Engl J Med. 2007;357(3):266-81.

54. Jones G, Strugnell SA, DeLuca HF. Current understanding of the molecular actions of vitamin D. Physiol Rev. 1998;78(4):1193-231.

55. Kato S. The function of vitamin D receptor in vitamin D action. J Biochem. 2000;127(5):717-22.

56. Haussler MR, Whitfield GK, Haussler CA, Hsieh JC, Thompson PD, Selznick SH, et al. The nuclear vitamin D receptor: biological and molecular regulatory properties revealed. J Bone Miner Res. 1998;13(3):325-49.

57. Carlberg C, Seuter S, Heikkinen S. The first genome-wide view of vitamin D receptor locations and their mechanistic implications. Anticancer Res. 2012;32(1):271-82.

58. Haussler MR, Jurutka PW, Mizwicki M, Norman AW. Vitamin D receptor (VDR)-mediated actions of 1 alpha,25(OH)(2)vitarnin D-3: Genomic and non-genomic mechanisms. Best Pract Res Cl En. 2011;25(4):543-59.

59. Feldman D, Krishnan AV, Swami S, Giovannucci E, Feldman BJ. The role of vitamin D in reducing cancer risk and progression. Nat Rev Cancer. 2014.

60. Carlberg C, Campbell MJ. Vitamin D receptor signaling mechanisms: integrated actions of a well-defined transcription factor. Steroids. 2013;78(2):127-36.

61. Deeb KK, Trump DL, Johnson CS. Vitamin D signalling pathways in cancer: potential for anticancer therapeutics. Nat Rev Cancer. 2007;7(9):684-700.

62. Barnett C, Krebs JE. WSTF does it all: a multifunctional protein in transcription, repair, and replication. Biochem Cell Biol. 2011;89(1):12-23.

63. Kato S, Kim MS, Yamaoka K, Fujiki R. Mechanisms of transcriptional repression by 1,25(OH)2 vitamin D. Curr Opin Nephrol Hypertens. 2007;16(4):297-304.

64. Haussler MR, Haussler CA, Jurutka PW, Thompson PD, Hsieh J-C, Remus LS, et al. The vitamin D hormone and its nuclear receptor: molecular actions and disease states. J Endocrinol. 1997;154(3 Suppl):S57-S73.

65. Yip KY, Cheng C, Bhardwaj N, Brown JB, Leng J, Kundaje A, et al. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. Genome Biol. 2012;13(9):R48.

66. Kheradpour P, Kellis M. Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments. Nucleic Acids Res. 2013.

67. Martinez GJ, Rao A. Immunology. Cooperative transcription factor complexes in control. Science. 2012;338(6109):891-2.

68. Murayama A, Kim MS, Yanagisawa J, Takeyama K, Kato S. Transrepression by a liganded nuclear receptor via a bHLH activator through co-regulator switching. EMBO J. 2004;23(7):1598-608.

69. Sun G, Porter W, Safe S. Estrogen-induced retinoic acid receptor alpha 1 gene expression: role of estrogen receptor-Sp1 complex. Mol Endocrinol. 1998;12(6):882-90.

70. Owen GI, Richer JK, Tung L, Takimoto G, Horwitz KB. Progesterone regulates transcription of the p21(WAF1) cyclin-dependent kinase inhibitor gene through Sp1 and CBP/p300. J Biol Chem. 1998;273(17):10696-701.

71. Ramagopalan SV, Heger A, Berlanga AJ, Maugeri NJ, Lincoln MR, Burrell A, et al. A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. Genome Res. 2010;20(10):1352-60.

72. Heikkinen S, Vaisanen S, Pehkonen P, Seuter S, Benes V, Carlberg C. Nuclear hormone 1alpha,25-dihydroxyvitamin D3 elicits a genome-wide shift in the locations of VDR chromatin occupancy. Nucleic Acids Res. 2011;39(21):9181-93.

73. Ravasi T, Suzuki H, Cannistraci CV, Katayama S, Bajic VB, Tan K, et al. An atlas of combinatorial transcriptional regulation in mouse and man. Cell. 2010;140(5):744-52.

74. Machanick P, Bailey TL. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics. 2011;27(12):1696-7.

75. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. Quantifying similarity between motifs. Genome Biol. 2007;8(2).

76. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, et al. The UCSC Genome Browser database: extensions and updates 2013. Nucleic Acids Res. 2013;41(D1):D64-D9.

77. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, et al. Ensembl 2008. Nucleic Acids Res. 2008;36(suppl 1):D707-D14.

78. Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouzé P, et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics. 2001;17(12):1113-22.

79. Meyer MB, Goetsch PD, Pike JW. VDR/RXR and TCF4/beta-catenin cistromes in colonic cells of colorectal tumor origin: impact on c-FOS and c-MYC gene expression. Mol Endocrinol. 2012;26(1):37-51.

80. Zaret KS, Carroll JS. Pioneer transcription factors: establishing competence for gene expression. Genes Dev. 2011;25(21):2227-41.

81. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, et al. Mapping and analysis of chromatin state dynamics in nine human cell types. Nature. 2011;473(7345):43-9.

82. Hoffman MM, Ernst J, Wilder SP, Kundaje A, Harris RS, Libbrecht M, et al. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 2013;41(2):827-41.

83. Smit AF, Riggs AD. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. Nucleic Acids Res. 1995;23(1):98-102.

84. Smith AM, Sanchez MJ, Follows GA, Kinston S, Donaldson IJ, Green AR, et al. A novel mode of enhancer evolution: the Tal1 stem cell enhancer recruited a MIR element to specifically boost its activity. Genome Res. 2008;18(9):1422-32.

85. Nie Y, Liu H, Sun X. The patterns of histone modifications in the vicinity of transcription factor binding sites in human lymphoblastoid cell lines. PLoS One. 2013;8(3):e60002.

86. Huang YC, Chen JY, Hung WC. Vitamin D3 receptor/Sp1 complex is required for the induction of p27Kip1 expression by vitamin D3. Oncogene. 2004;23(28):4856-61.

87. Cheng HT, Chen JY, Huang YC, Chang HC, Hung WC. Functional role of VDR in the activation of p27(Kip1) by the VDR/Sp1 complex. J Cell Biochem. 2006;98(6):1450-6.

88. Sela I, Lukatsky DB. DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. Biophys J. 2011;101(1):160-6.

89. Li L, He S, Sun JM, Davie JR. Gene regulation by Sp1 and Sp3. Biochem Cell Biol. 2004;82(4):460-71.

90. Adams CC, Workman JL. Binding of disparate transcriptional activators to nucleosomal DNA is inherently cooperative. Mol Cell Biol. 1995;15(3):1405-21.

91. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

92. Zella LA, Meyer MB, Nerenz RD, Lee SM, Martowicz ML, Pike JW. Multifunctional enhancers regulate mouse and human vitamin D receptor gene transcription. Mol Endocrinol. 2010;24(1):128-47.

93. Merkenschlager M, Odom DT. CTCF and cohesin: linking gene regulatory elements with their targets. Cell. 2013;152(6):1285-97.

94. Yan J, Enge M, Whitington T, Dave K, Liu J, Sur I, et al. Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. Cell. 2013;154(4):801-13.

95. Ball AR, Jr., Chen YY, Yokomori K. Mechanisms of cohesin-mediated gene regulation and lessons learned from cohesinopathies. Biochimica et biophysica acta. 2014;1839(3):191-202.

96. Faure AJ, Schmidt D, Watt S, Schwalie PC, Wilson MD, Xu H, et al. Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules. Genome Res. 2012;22(11):2163-75.

97. Pike JW, Meyer MB. Fundamentals of vitamin D hormone-regulated gene expression. J Steroid Biochem Mol Biol. 2013(0).

98. Tolon RM, Castillo AI, Jimenez-Lara AM, Aranda A. Association with Ets-1 causes ligand- and AF2-independent activation of nuclear receptors. Mol Cell Biol. 2000;20(23):8793-802.

99. Handel AE, Sandve GK, Disanto G, Berlanga-Taylor AJ, Gallone G, Hanwell H, et al. Vitamin D receptor ChIP-seq in primary CD4+ cells: relationship to serum 25-hydroxyvitamin D levels and autoimmune disease. BMC Med. 2013;11:163.

100. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. Nature. 1997;389(6648):251-60.

101. Tuoresmaki P, Vaisanen S, Neme A, Heikkinen S, Carlberg C. Patterns of Genome-Wide VDR Locations. PLoS One. 2014;9(4):e96105.

102. Kumar P, Yadav VK, Baral A, Kumar P, Saha D, Chowdhury S. Zinc-finger transcription factors are associated with guanine quadruplex motifs in human, chimpanzee, mouse and rat promoters genome-wide. Nucleic Acids Res. 2011;39(18):8005-16.

Appendix

A.1 Region Comparison (Ramagopalan dataset)

Comparing 'TFBS_JSVDR' against 'TFBS_ChIP_Seq'

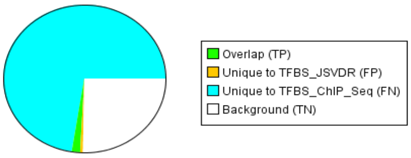
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
2053	905	33978	97292
1.5%	0.7%	25.3%	72.5%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.021	0.974	0.694	0.259	0.02	0.357	0.04	0.268	-0.016



69.4% of **TFBS_JSVDR** overlaps with **TFBS_ChIP_Seq** (PPV)
2.1% of **TFBS_ChIP_Seq** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 2% with respect to both tracks (Performance Coefficient)

Comparing 'TFBS_JSVDR' against 'EnsemblGenes'

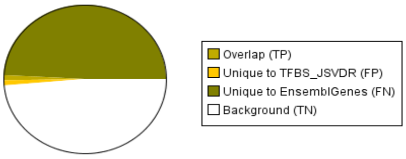
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
1548	1410	65225	66045
1.2%	1.1%	48.6%	49.2%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.023	0.979	0.523	0.497	0.022	0.273	0.044	0.497	0.006



52.3% of **TFBS_JSVDR** overlaps with **EnsemblGenes** (PPV)
2.3% of **EnsemblGenes** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 2.2% with respect to both tracks (Performance Coefficient)

Comparing 'TFBS_JSVDR' against 'CpG_islands'

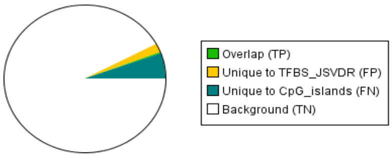
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
225	2733	123858	7412
0.2%	2%	92.3%	5.5%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.029	0.978	0.076	0.944	0.022	0.053	0.042	0.924	0.012



7.6% of **TFBS_JSVDR** overlaps with **CpG_islands** (PPV)
2.9% of **CpG_islands** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 2.2% with respect to both tracks (Performance Coefficient)

Comparing 'TFBS_JSVDR' against 'FAIRE_seq_peaks'

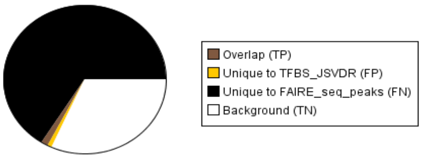
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
1940	1018	42624	88646
1.4%	0.8%	31.8%	66%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.021	0.977	0.656	0.325	0.021	0.339	0.041	0.332	-0.006



65.6% of **TFBS_JSVDR** overlaps with **FAIRE_seq_peaks** (PPV)
2.1% of **FAIRE_seq_peaks** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 2.1% with respect to both tracks (Performance Coefficient)

Comparing 'TFBS_JSVDR' against 'DNaseHS_hotspots'

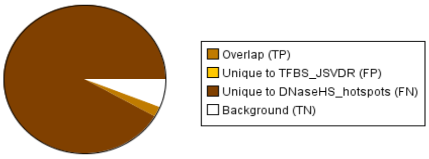
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
2804	154	8119	123151
2.1%	0.1%	6%	91.7%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.022	0.981	0.948	0.062	0.022	0.485	0.044	0.081	0.006



94.8% of **TFBS_JSVDR** overlaps with **DNaseHS_hotspots** (PPV)
2.2% of **DNaseHS_hotspots** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 2.2% with respect to both tracks (Performance Coefficient)

Comparing 'TFBS_JSVDR' against 'H3K27ac_Gm12878_peak'

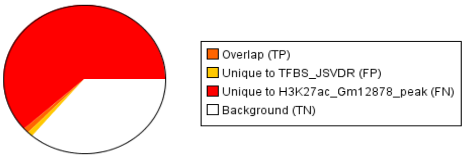
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
1455	1503	48519	82751
1.1%	1.1%	36.1%	61.6%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.017	0.97	0.492	0.37	0.017	0.255	0.033	0.372	-0.042



49.2% of **TFBS_JSVDR** overlaps with **H3K27ac_Gm12878_peak** (PPV)
1.7% of **H3K27ac_Gm12878_peak** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 1.7% with respect to both tracks (Performance Coefficient)

Comparing 'TFBS_JSVDR' against 'H3K4me1_Gm12878_peak'

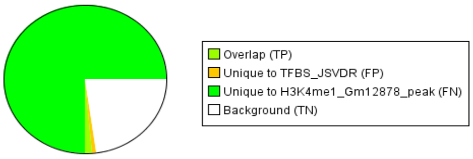
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
2010	948	30682	100588
1.5%	0.7%	22.9%	74.9%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.02	0.97	0.68	0.234	0.019	0.35	0.038	0.244	-0.03



68% of **TFBS_JSVDR** overlaps with **H3K4me1_Gm12878_peak** (PPV)
2% of **H3K4me1_Gm12878_peak** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 1.9% with respect to both tracks (Performance Coefficient)

Comparing 'TFBS_JSVDR' against 'H3K4me3_Gm12878_peak'

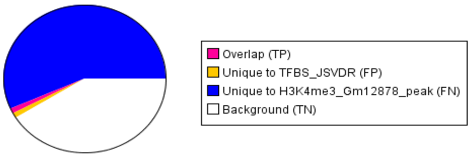
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
1305	1653	55110	76160
1%	1.2%	41.1%	56.7%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.017	0.971	0.441	0.42	0.016	0.229	0.032	0.42	-0.041



44.1% of **TFBS_JSVDR** overlaps with **H3K4me3_Gm12878_peak** (PPV)
1.7% of **H3K4me3_Gm12878_peak** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 1.6% with respect to both tracks (Performance Coefficient)

Comparing 'TFBS_JSVDR' against 'CCDS'

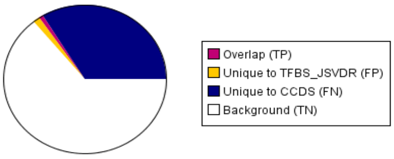
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
1020	1938	86021	45249
0.8%	1.4%	64.1%	33.7%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.022	0.978	0.345	0.655	0.021	0.183	0.041	0.648	0



34.5% of **TFBS_JSVDR** overlaps with **CCDS** (PPV)
2.2% of **CCDS** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 2.1% with respect to both tracks (Performance Coefficient)

Comparing 'TFBS_JSVDR' against 'RepeatMasker327'

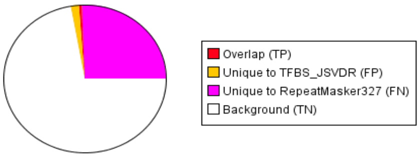
Analysis based on 192 sequences from collection **JSVDR_positive**

Base Counts

TP	FP	TN	FN
822	2136	96945	34325
0.6%	1.6%	72.2%	25.6%

Statistics

SN	SP	PPV	NPV	PC	ASP	F	Acc	CC
0.023	0.978	0.278	0.739	0.022	0.151	0.043	0.728	0.005



27.8% of **TFBS_JSVDR** overlaps with **RepeatMasker327** (PPV)
2.3% of **RepeatMasker327** overlaps with **TFBS_JSVDR** (SN)
The relative overlap is 2.2% with respect to both tracks (Performance Coefficient)

A.2 Motifs enriched from background (Ramagopalan dataset)

Analysis performed with motifs from **Jaspar_Core** and sites from **TFBS_observed** on 2776 sequences. Expected motif frequencies from **Frequencies_expected**.
Statistical significance evaluated using a binomial test with p-value threshold=0.05 (Bonferroni-corrected threshold=1.0893246187363835E-4 considering all 459 motifs tested)

ID	Name	Class	Total	Sequences	%	Expected	p-value
MA0106	TP53	4.3.1.1	2	1	0%	0.000	0.0
MA0415	YAP1	1.1.1	3	3	0%	0.000	0.0
MA0435	YPR015C	2.3	4	4	0%	0.000	0.0
MA0138	REST	2.3.2.2	1	1	0%	0.000	0.0
MA0045	HMG-IY	0.2.1	2402	860	30%	563.800	0.0
MA0120	id1	2.3	4685	1444	52%	1921.000	0.0
MA0079	SP1	2.3.1.0	4557	1617	58%	2091.600	0.0
MA0074	RXRA-VDR	2.1.2	198	192	6%	12.800	2.3434895490765235E-157
MA0039	Klf4	2.3.2.2	2280	1232	44%	1369.200	3.079933178976263E-112
MA0050	IRF1	3.5.3.0	658	546	19%	238.200	1.3636400510714893E-110
MA0123	ABI4	0.5.2.0	2713	1135	40%	1900.400	2.0693969912503605E-69
MA0033	FOXL1	3.3	9541	2038	73%	7974.400	4.4533182913969824E-66
MA0013	br_Z4	2.3	3176	1356	48%	2309.600	5.071725049932308E-66
MA0049	hb	2.3.2.2	6984	1487	53%	5679.400	9.953923860198008E-64
MA0139	CTCF	2.3	96	70	2%	9.000	5.033585985968533E-63
MA0324	LEU3	2.4.1.0	1193	659	23%	714.400	1.6767440605364938E-60
MA0062	GABPA	3.5.2	622	467	16%	304.800	1.891425682401901E-57
MA0453	nub	2.3	305	266	9%	104.400	2.9396150448336844E-57
MA0443	btd	2.3	1796	1001	36%	1219.400	2.2677242938111066E-54
MA0146	Zfx	2.3	876	642	23%	509.400	1.2168006087784364E-49
MA0010	br_Z1	2.3	929	657	23%	553.000	1.286753816779391E-48
MA0057	MZF1_5-13	2.3.2.2	4000	1825	65%	3168.800	1.2303443333859162E-46
MA0041	FOXD3	3.3	1478	821	29%	1003.400	3.317115661652411E-45
MA0404	TBS1	2.4.1	380	187	6%	168.200	6.14271439917845E-45
MA0015	Cf2_II	2.3.2.2	784	324	11%	463.000	1.6934239114326363E-42
MA0277	AZF1	2.3	2650	1352	48%	2015.200	2.516066263630848E-42
MA0082	Squamosa	4.4	837	621	22%	507.200	1.8973859656707396E-41
MA0018	CREB1	1.1.2	1490	1032	37%	1034.000	4.267971837114263E-41
MA0060	NFYA	4.8.1.0	273	213	7%	107.400	4.923166603010312E-41
MA0012	br_Z3	2.3	2347	1211	43%	1762.400	6.309548406452107E-41
MA0055	Myf	1.2.2.0	1219	743	26%	813.200	9.51405283287063E-41
MA0388	SPT23	unknown	6336	2055	74%	5350.400	2.7942716155625338E-40
MA0205	Trl	2.3	3129	1432	51%	2453.000	4.496944111727066E-40
MA0002	RUNX1	4.11	1291	958	34%	880.000	4.501015306198561E-39
MA0073	RREB1	2.3	92	68	2%	19.600	1.9070313290549962E-32
MA0431	YML081W	2.3	6884	2076	74%	5971.200	5.968580130337641E-32
MA0160	NR4A2	2.1.2.17	3268	1741	62%	2658.200	3.440692616764698E-31
MA0042	FOXI1	3.3	1160	768	27%	824.600	6.257027140314789E-29
MA0076	ELK4	3.5.2.0	405	314	11%	222.400	1.4719883273600494E-28
MA0458	slp1	3.3	1521	1002	36%	1135.800	2.4446546657481015E-28
MA0003	TFAP2A	1.6.1	10959	2206	79%	9873.400	3.186534054259598E-28
MA0425	YGR067C	2.3	7256	2091	75%	6388.200	1.3192811794258485E-27

ID	Name	Class	Total	Sequences	%	Expected	p-value
MA0317	HCM1	3.3	4840	1792	64%	4197.000	2.3028075257520702E-23
MA0303	GCN4	1.1.1.5	218	141	5%	103.800	5.953811988462878E-23
MA0081	SPIB	3.5.2.0	7274	2376	85%	6493.400	1.0799826304123732E-22
MA0242	run-Bgb	4.11	607	522	18%	399.600	1.1773887104392253E-22
MA0052	MEF2A	4.4.1.1	383	323	11%	225.600	4.26262308855483E-22
MA0148	FOXA1	3.3	892	674	24%	641.000	1.3254650215193485E-21
MA0375	RSC30	2.4.1	3872	749	26%	3335.400	9.508304098144944E-21
MA0043	HLF	1.1.4.0	384	328	11%	234.400	9.089087224407455E-20
MA0141	Esrrb	2.1.1.5	428	383	13%	273.400	1.311201268585701E-18
MA0088	Znf143	2.3	96	85	3%	34.000	1.8171925476401625E-18
MA0417	YAP5	1.1.1	14830	2541	91%	13852.800	1.179286814756397E-16
MA0285	CRZ1	2.3	8417	2308	83%	7683.800	1.186197905512155E-16
MA0446	fkh	3.3	1142	779	28%	888.600	2.7068379335026756E-16
MA0445	D	4.7	974	710	25%	742.600	2.9424937702270633E-16
MA0346	NHP6B	4.7	106	78	2%	43.600	1.0618828468809963E-15
MA0449	h	1.2.5.1	278	130	4%	167.600	4.243909391532719E-15
MA0173	CG11617	3.1	1122	763	27%	884.000	8.492075151666313E-15
MA0084	SRY	4.7.1.0	3574	1570	56%	3141.800	2.3415634526575686E-14
MA0369	RLM1	4.4	54	40	1%	15.800	4.444284713177931E-14
MA0268	ADR1	2.3	12338	2523	90%	11540.200	9.877732891740435E-14
MA0047	FOXA2	3.3	486	412	14%	344.400	3.8773050442959175E-13
MA0274	ARR1	1.1.1.5	8122	2211	79%	7498.800	6.106991188654272E-13
MA0345	NHP6A	4.7	66	46	1%	23.600	6.511210398978284E-13
MA0096	bZIP910	1.1	391	318	11%	268.400	1.4375004412994859E-12
MA0286	CST6	1.1.2	381	294	10%	260.200	1.4572919269120048E-12
MA0162	Egr1	2.3.2.1	453	377	13%	321.000	2.2959476503181608E-12
MA0156	FEV	3.5.2	2190	1381	49%	1883.400	3.004262469050961E-12
MA0314	HAP3	4.8.1.0	128	108	3%	65.200	4.205054067085168E-12
MA0133	BRCA1	unknown	7159	2343	84%	6597.800	4.6866804227648936E-12
MA0399	SUT1	2.4.1	8200	1580	56%	7628.000	4.8356078498424366E-11
MA0107	RELA	4.1.1.0	578	413	14%	438.600	1.2283850183772545E-10
MA0373	RPN4	2.3	2612	1400	50%	2305.800	2.216390899214843E-10
MA0201	Ptx1	3.1	927	689	24%	750.200	2.600154033374409E-10
MA0368	RIM101	2.3	1646	1132	40%	1409.800	4.707053812327895E-10
MA0244	slbo	2.3	4780	1877	67%	4376.400	9.34357616365718E-10
MA0315	HAP4	4.8.1.0	239	191	6%	158.200	1.3696501731635083E-9
MA0102	CEBPA	1.1.3.0	2057	1185	42%	1799.400	1.5175344729461606E-9
MA0261	lin-14	unknown	6973	2392	86%	6501.600	3.777000276822062E-9
MA0283	CHA4	2.4	2546	1099	39%	2265.600	3.9582867600982674E-9
MA0357	PHO4	1.2.5.3	1618	633	22%	1397.600	4.664260643775308E-9
MA0386	TBP	4.6.1	19	15	0%	3.600	1.0148759873384451E-8
MA0028	ELK1	3.5.2.0	2224	1329	47%	1971.400	1.2892154226518568E-8
MA0361	RDS1	2.4	5595	1142	41%	5200.600	3.316271826754662E-8
MA0124	NKX3-1	3.1.1.15	1921	1042	37%	1694.200	3.58247127672331E-8
MA0260	che-1	2.3	4860	2069	74%	4499.000	5.4685940719606957E-8
MA0114	HNF4A	2.1.2.11	195	185	6%	130.000	6.481166469563176E-8
MA0316	HAP5	4.8.1.0	104	91	3%	59.000	7.665394107650297E-8

ID	Name	Class	Total	Sequences	%	Expected	p-value
MA0356	PHO2	3.1	34209	2387	85%	33255.400	8.800634148901488E-8
MA0058	MAX	1.3.2.2	782	507	18%	647.000	1.4833845191378971E-7
MA0213	brk	unknown	2449	1109	39%	2205.400	1.786429999280868E-7
MA0459	tll	2.1.2.15	283	253	9%	207.200	3.4251065627891277E-7
MA0197	Oct	3.1	4313	1609	57%	4005.200	7.918114244728617E-7
MA0051	IRF2	3.5.3.0	30	30	1%	10.600	8.134144894266692E-7
MA0011	br_Z2	2.3	3703	1603	57%	3418.800	8.260630010421168E-7
MA0117	Mafb	1.1.1.3	5219	2137	76%	4886.000	1.2421749954710365E-6
MA0086	sna	2.3.2.2	3907	1860	67%	3630.200	2.8926622222903902E-6
MA0297	FKH2	3	688	548	19%	576.400	3.449841524529358E-6
MA0157	FOXO3	3.3	2542	1396	50%	2326.000	5.253525003972314E-6
MA0340	MOT3	2.3.3.0	7217	2334	84%	6849.800	5.473649361961844E-6
MA0398	SUM1	unknown	3873	1332	47%	3606.400	5.858108955629919E-6
MA0040	FOXQ1	3.3	440	378	13%	356.200	9.985927929375357E-6
MA0149	EWSR1-FLI1	unknown	8	5	0%	1.000	1.024914513996682E-5
MA0065	PPARG-RXRA	2.1.2	86	84	3%	52.600	1.4764344039704628E-5
MA0333	MET31	2.3	2615	1510	54%	2407.200	1.5153792177712381E-5
MA0289	DAL80	2.2.1.2	838	649	23%	723.000	1.597232191473978E-5
MA0165	Abd-B	3.1.1.1	7046	1978	71%	6705.600	1.8801684283217625E-5
MA0419	YAP7	1.1.1	223	201	7%	167.200	2.243471794803681E-5
MA0336	MGA1	3.3.3.4	29	29	1%	12.000	2.2615612982104616E-5
MA0017	NR2F1	2.1.2.16	192	180	6%	141.000	2.6257235509838956E-5
MA0059	MYC-MAX	1.3.2	191	148	5%	141.000	3.621793464775598E-5
MA0048	NHLH1	1.2	503	309	11%	419.400	4.0428737096361676E-5
MA0332	MET28	1.1	2477	1462	52%	2291.600	6.762559044892259E-5
MA0300	GAT1	2.2.1.2	662	534	19%	570.000	9.079213892947535E-5
MA0239	prd	3.1.1.9	2714	1504	54%	2524.800	1.0183106431865824E-4
MA0126	ovo	2.3.2.2	2714	1504	54%	2524.800	1.0183106431865824E-4

A.3 Motif overrepresentation comparison between VDR motif-positive (control) and motif-negative (target) sequence collections (Ramagopalan dataset)

Motif occurrence comparison for "VDR_negative" vs "VDR_positive"

Statistical significance evaluated using a binomial test with p-value threshold=0.05
(Bonferroni-corrected threshold=3.703703703703704E-4 considering all 135 motifs tested)

<table>
<tr><td>Motifs present only in target</td><td>Motifs overrepresented in target</td><td>Same rate</td><td>Motifs overrepresented in control</td><td>Motifs present only in control</td><td>Motifs not present</td></tr>
<tr><td>1</td><td>33</td><td rowspan="2">42</td><td>55</td><td>4</td><td rowspan="2">0</td></tr>
<tr><td colspan="2">34</td><td colspan="2">59</td></tr>
</table>

ID	Name	Class	Target	Control	p-value target	p-value control
MA0138	REST	2.3.2.2	1	0	0	1
MA0280	CAT8	2.4	4485	6810	5.54E-284	1
MA0399	SUT1	2.4.1	3399	4801	8.48E-266	1

MA0361	RDS1	2.4	2431	3164	2.22E-234	1
MA0375	RSC30	2.4.1	1711	2161	9.52E-178	1
MA0003	TFAP2A	1.6.1	3866	7093	4.62E-122	1
MA0341	MSN2	2.3.3.0	6272	12902	3.25E-104	1
MA0123	ABI4	0.5.2.0	1092	1621	3.08E-76	1
MA0283	CHA4	2.4	1034	1512	3.23E-75	1
MA0213	brk	unknown	935	1514	4.26E-50	1
MA0425	YGR067C	2.3	2412	4844	1.95E-49	1
MA0324	LEU3	2.4.1.0	513	680	2.33E-49	1
MA0404	TBS1	2.4.1	200	180	1.22E-40	1
MA0079	SP1	2.3.1.0	1558	2999	1.48E-40	1
MA0431	YML081W	2.3	2253	4631	9.43E-40	1
MA0285	CRZ1	2.3	2664	5753	2.63E-33	1
MA0443	btd	2.3	666	1130	3.90E-31	1
MA0268	ADR1	2.3	3740	8598	2.12E-26	1
MA0373	RPN4	2.3	894	1718	2.62E-24	1
MA0028	ELK1	3.5.2.0	772	1452	1.10E-23	1
MA0062	GABPA	3.5.2	258	364	1.98E-22	1
MA0362	RDS2	2.4	674	1248	2.76E-22	1
MA0076	ELK4	3.5.2.0	179	226	1.39E-20	1
MA0039	Klf4	2.3.2.2	773	1507	4.47E-20	1
MA0006	Arnt-Ahr	1.2.6	832	1692	1.59E-16	1
MA0162	Egr1	2.3.2.1	176	277	8.87E-12	1
MA0449	h	1.2.5.1	116	162	3.95E-11	1
MA0260	che-1	2.3	1471	3389	7.22E-11	1
MA0146	Zfx	2.3	291	585	2.21E-07	1
MA0088	Znf143	2.3	44	52	5.49E-07	1
MA0117	Mafb	1.1.1.3	1510	3709	0.00002	1
MA0057	MZF1_5-13	2.3.2.2	1163	2837	0.00005	1
MA0156	FEV	3.5.2	652	1538	0.0001	1
MA0096	bZIP910	1.1	132	259	0.00016	1
MA0160	NR4A2	2.1.2.17	0	3268	1	0
MA0074	RXRA-VDR	2.1.2	0	198	1	0
MA0435	YPR015C	2.3	0	4	1	0
MA0106	TP53	4.3.1.1	0	2	1	0
MA0356	PHO2	3.1	7637	26572	1	2.87626E-313
MA0033	FOXL1	3.3	2104	7437	1	8.51E-100
MA0141	Esrrb	2.1.1.5	42	386	1	2.04E-87
MA0165	Abd-B	3.1.1.1	1563	5483	1	1.54E-70
MA0182	CG4328	3.1	2295	7681	1	3.58E-66
MA0231	lbe	3.1	2361	7816	1	5.30E-61
MA0063	Nkx2-5	3.1.1.15	2129	7040	1	2.44E-54
MA0219	ems	3.1.1.6	1813	6062	1	2.45E-52
MA0274	ARR1	1.1.1.5	1880	6242	1	3.70E-50
MA0398	SUM1	unknown	842	3031	1	1.18E-47
MA0195	Lim3	3.1	1690	5600	1	1.17E-44
MA0197	Oct	3.1	956	3357	1	2.27E-44
MA0417	YAP5	1.1.1	3602	11228	1	6.76E-43
MA0049	hb	2.3.2.2	1619	5365	1	2.83E-42
MA0013	br_Z4	2.3	700	2476	1	9.12E-35
MA0132	Pdx1	3.1	1232	4051	1	6.85E-31
MA0244	slbo	2.3	1117	3663	1	3.39E-27
MA0075	Prrx2	3.1	940	3125	1	3.90E-27
MA0317	HCM1	3.3	1138	3702	1	2.26E-25
MA0215	btn	3.1	693	2339	1	1.38E-23
MA0201	Ptx1	3.1	188	739	1	3.11E-21
MA0045	HMG-IY	0.2.1	544	1858	1	2.38E-20
MA0084	SRY	4.7.1.0	843	2731	1	1.52E-18
MA0209	ap	3.1	664	2187	1	4.56E-18
MA0388	SPT23	unknown	1547	4789	1	6.46E-18
MA0120	id1	2.3	1125	3560	1	8.20E-18
MA0082	Squamosa	4.4	172	665	1	1.33E-17
MA0041	FOXD3	3.3	329	1149	1	2.28E-15
MA0086	sna	2.3.2.2	937	2970	1	2.72E-15

MA0313	HAP2	4.8.1.0	3249	9610	1	1.67E-14
MA0458	slp1	3.3	344	1177	0.99999	1.48E-13
MA0037	GATA3	2.2.1.1	2173	6504	0.99999	5.23E-13
MA0340	MOT3	2.3.3.0	1796	5421	0.99999	6.10E-13
MA0173	CG11617	3.1	249	873	0.99998	1.22E-12
MA0446	fkh	3.3	254	888	0.99997	2.11E-12
MA0102	CEBPA	1.1.3.0	481	1576	0.99997	3.10E-12
MA0064	PBF	2.2	3022	8879	0.99997	1.25E-11
MA0042	FOXI1	3.3	262	898	0.99991	7.60E-11
MA0148	FOXA1	3.3	197	695	0.99989	1.09E-10
MA0011	br_Z2	2.3	904	2799	0.99992	1.11E-10
MA0015	Cf2_II	2.3.2.2	172	612	0.99984	3.60E-10
MA0157	FOXO3	3.3	611	1931	0.99986	4.40E-10
MA0012	br_Z3	2.3	563	1784	0.99978	1.49E-09
MA0040	FOXQ1	3.3	92	348	0.99949	7.25E-09
MA0017	NR2F1	2.1.2.16	35	157	0.99934	8.09E-09
MA0047	FOXA2	3.3	103	383	0.99947	8.73E-09
MA0346	NHP6B	4.7	17	89	0.9988	2.83E-08
MA0018	CREB1	1.1.2	352	1138	0.99935	2.83E-08
MA0124	NKX3-1	3.1.1.15	461	1460	0.99936	2.93E-08
MA0297	FKH2	3	157	531	0.9972	1.37E-06
MA0445	D	4.7	231	743	0.99385	0.00001
MA0368	RIM101	2.3	403	1243	0.99299	0.00002
MA0277	AZF1	2.3	661	1989	0.99289	0.00002
MA0314	HAP3	4.8.1.0	25	103	0.9872	0.00007
MA0114	HNF4A	2.1.2.11	41	154	0.9848	0.00012
MA0002	RUNX1	4.11	343	948	0.56084	0.41558
MA0010	br_Z1	2.3	242	687	0.70156	0.20237
MA0043	HLF	1.1.4.0	100	284	0.64598	0.29159
MA0048	NHLH1	1.2	150	353	0.035	0.9986
MA0050	IRF1	3.5.3.0	195	463	0.02529	0.99935
MA0051	IRF2	3.5.3.0	9	21	0.35523	0.80325
MA0052	MEF2A	4.4.1.1	105	278	0.37244	0.72887
MA0055	Myf	1.2.2.0	360	859	0.00491	0.99999
MA0058	MAX	1.3.2.2	216	566	0.26082	0.86533
MA0059	MYC-MAX	1.3.2	45	146	0.88603	0.02606
MA0060	NFYA	4.8.1.0	66	207	0.87053	0.03426
MA0065	PPARG-RXRA	2.1.2	18	68	0.93312	0.00722
MA0073	RREB1	2.3	30	62	0.07432	0.99236
MA0081	SPIB	3.5.2.0	2010	5264	0.02427	0.99944
MA0098	ETS1	3.5.2	7864	20817	0.00167	1
MA0107	RELA	4.1.1.0	177	401	0.0074	0.99996
MA0126	ovo	2.3.2.2	692	2022	0.95678	0.00225
MA0133	BRCA1	unknown	1956	5203	0.10301	0.98211
MA0139	CTCF	2.3	28	68	0.2772	0.8653
MA0149	EWSR1-FLI1	unknown	3	5	0.27351	0.91445
MA0199	Optix	3.1	774	2195	0.8475	0.04651
MA0205	Trl	2.3	890	2239	0.00577	0.99998
MA0239	prd	3.1.1.9	692	2022	0.95678	0.00225
MA0242	run-Bgb	4.11	169	438	0.24354	0.88454
MA0261	lin-14	unknown	1918	5055	0.05159	0.99649
MA0286	CST6	1.1.2	117	264	0.0224	0.99949
MA0289	DAL80	2.2.1.2	207	631	0.94477	0.00419
MA0300	GAT1	2.2.1.2	167	495	0.85494	0.04308
MA0303	GCN4	1.1.1.5	53	165	0.8273	0.06733
MA0315	HAP4	4.8.1.0	57	182	0.8851	0.02604
MA0316	HAP5	4.8.1.0	22	82	0.94205	0.00488
MA0332	MET28	1.1	632	1845	0.95123	0.00306
MA0333	MET31	2.3	761	1854	0.00075	1
MA0336	MGA1	3.3.3.4	11	18	0.06823	0.99382
MA0345	NHP6A	4.7	15	51	0.82209	0.08152
MA0357	PHO4	1.2.5.3	452	1166	0.10577	0.98136
MA0369	RLM1	4.4	19	35	0.05888	0.9954
MA0386	TBP	4.6.1	8	11	0.05038	0.99659

MA0415	YAP1	1.1.1	1	2	0.51586	0.76153
MA0419	YAP7	1.1.1	60	163	0.48654	0.56085
MA0453	nub	2.3	74	231	0.87855	0.0294
MA0459	tll	2.1.2.15	71	212	0.77837	0.11531

Motif occurrence comparison for "NR4A2_positive" vs "JSVDR_positive"

Statistical significance evaluated using a binomial test with p-value threshold=0.05
(Bonferroni-corrected threshold=3.7037037037037037E-4 considering all 135 motifs tested)

Motifs present only in target	Motifs overrepresented in target	Same rate	Motifs overrepresented in control	Motifs present only in control	Motifs not present
4	44	78	8	0	1
48			8		

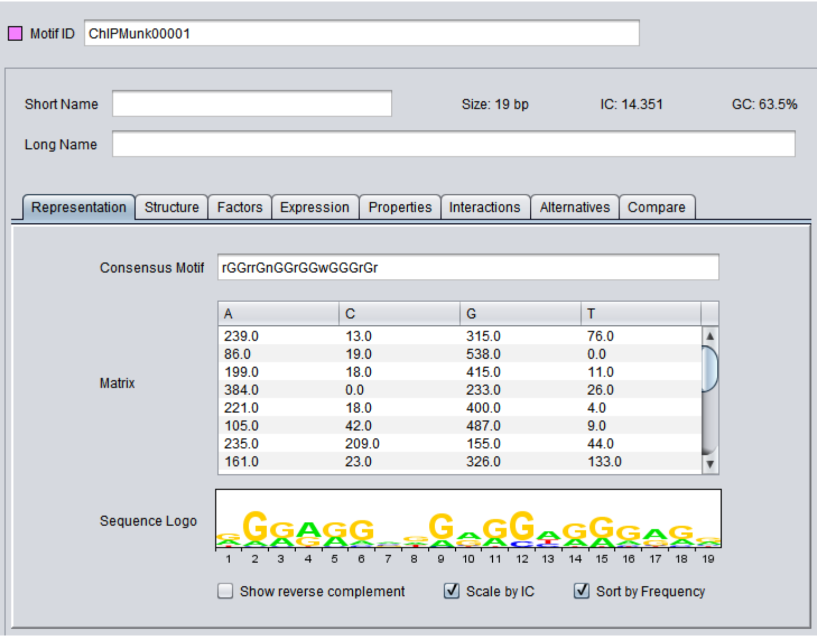
ID	Name	Class	Target	Control	p-value target	p-value control
MA0149	EWSR1-FLI1	unknown	5	0	0	1
MA0435	YPR015C	2.3	4	0	0	1
MA0106	TP53	4.3.1.1	2	0	0	1
MA0415	YAP1	1.1.1	2	0	0	1
MA0340	MOT3	2.3.3.0	5290	460	-?	0.4647
MA0399	SUT1	2.4.1	4697	254	7.64191E-197	1
MA0375	RSC30	2.4.1	2111	91	2.3302E-181	1
MA0341	MSN2	2.3.3.0	12596	843	3.24402E-170	1
MA0361	RDS1	2.4	3094	163	5.47919E-144	1
MA0280	CAT8	2.4	6665	422	6.81734E-131	1
MA0425	YGR067C	2.3	4750	293	1.55898E-107	1
MA0079	SP1	2.3.1.0	2936	165	9.02275E-106	1
MA0431	YML081W	2.3	4540	280	7.08658E-104	1
MA0283	CHA4	2.4	1482	73	5.57458E-88	1
MA0003	TFAP2A	1.6.1	6932	473	3.10568E-81	1
MA0123	ABI4	0.5.2.0	1584	82	1.82602E-79	1
MA0449	h	1.2.5.1	160	2	3.76671E-77	0.99999
MA0268	ADR1	2.3	8400	602	5.58197E-64	1
MA0443	btd	2.3	1103	57	2.05394E-56	0.99999
MA0006	Arnt-Ahr	1.2.6	1650	95	6.41733E-55	0.99999
MA0039	Klf4	2.3.2.2	1471	85	3.37849E-48	0.99997
MA0324	LEU3	2.4.1.0	672	33	1.67112E-41	0.99986
MA0285	CRZ1	2.3	5620	410	2.50879E-36	0.99981
MA0213	brk	unknown	1475	92	1.73401E-33	0.9996
MA0062	GABPA	3.5.2	358	16	7.71044E-30	0.99882
MA0404	TBS1	2.4.1	177	6	1.6646E-27	0.99775
MA0160	NR4A2	2.1.2.17	3268	235	1.20449E-25	0.99837
MA0076	ELK4	3.5.2.0	224	9	8.9237E-25	0.99691
MA0028	ELK1	3.5.2.0	1422	95	1.56324E-21	0.99605
MA0146	Zfx	2.3	577	33	4.21482E-21	0.99527
MA0057	MZF1_5-13	2.3.2.2	2766	200	6.91653E-21	0.99566
MA0058	MAX	1.3.2.2	556	33	1.33116E-17	0.99085

ID	Name	Class	Target	Control	p-value target	p-value control
MA0162	Egr1	2.3.2.1	272	14	2.22604E-15	0.98631
MA0357	PHO4	1.2.5.3	1140	78	1.44998E-14	0.98575
MA0002	RUNX1	4.11	931	64	9.78311E-12	0.97396
MA0362	RDS2	2.4	1218	87	4.91792E-11	0.97001
MA0059	MYC-MAX	1.3.2	145	7	1.00006E-10	0.96598
MA0073	RREB1	2.3	60	2	1.40868E-10	0.9649
MA0260	che-1	2.3	3308	256	1.4385E-10	0.96733
MA0315	HAP4	4.8.1.0	180	10	2.14222E-8	0.94566
MA0369	RLM1	4.4	34	1	7.16866E-8	0.94666
MA0333	MET31	2.3	1808	138	1.09107E-7	0.93613
MA0316	HAP5	4.8.1.0	82	4	1.44057E-6	0.92211
MA0242	run-Bgb	4.11	433	30	4.77502E-6	0.90641
MA0052	MEF2A	4.4.1.1	273	18	0.00001	0.89961
MA0060	NFYA	4.8.1.0	203	13	0.00003	0.88886
MA0373	RPN4	2.3	1668	131	0.00005	0.8757
MA0117	Mafb	1.1.1.3	3613	294	0.00015	0.85874
MA0074	RXRA-VDR	2.1.2	127	198	1	6.15473E-170
MA0033	FOXL1	3.3	7203	780	1	7.29218E-10
MA0356	PHO2	3.1	25772	2516	1	1.35188E-9
MA0417	YAP5	1.1.1	10888	1084	1	3.26111E-6
MA0182	CG4328	3.1	7423	749	1	0.00002
MA0120	id1	2.3	3457	369	1	0.00005
MA0231	lbe	3.1	7550	756	1	0.00005
MA0173	CG11617	3.1	838	107	1	0.00009
MA0010	br_Z1	2.3	665	64	0.99795	0.20811
MA0011	br_Z2	2.3	2708	253	0.99998	0.11652
MA0012	br_Z3	2.3	1723	153	0.87496	0.3782
MA0013	br_Z4	2.3	2390	254	1	0.00076
MA0015	Cf2_II	2.3.2.2	591	72	1	0.00331
MA0017	NR2F1	2.1.2.16	154	14	0.75029	0.45914
MA0018	CREB1	1.1.2	1120	100	0.86079	0.38819
MA0037	GATA3	2.2.1.1	6313	601	1	0.01096
MA0040	FOXQ1	3.3	339	35	0.99966	0.16697
MA0041	FOXD3	3.3	1116	100	0.89348	0.37001
MA0042	FOXI1	3.3	878	77	0.67868	0.46202
MA0043	HLF	1.1.4.0	277	21	0.01768	0.75268
MA0045	HMG-IY	0.2.1	1806	182	1	0.02168
MA0047	FOXA2	3.3	373	44	1	0.0278
MA0048	NHLH1	1.2	339	33	0.98823	0.26929
MA0049	hb	2.3.2.2	5217	489	1	0.03949
MA0050	IRF1	3.5.3.0	444	60	1	0.00073
MA0051	IRF2	3.5.3.0	20	3	0.99744	0.24917
MA0055	Myf	1.2.2.0	837	76	0.92923	0.34685
MA0063	Nkx2-5	3.1.1.15	6844	599	0.83416	0.39312
MA0064	PBF	2.2	8629	795	1	0.04176
MA0065	PPARG-RXRA	2.1.2	64	9	0.99999	0.10757
MA0075	Prrx2	3.1	3032	274	0.99168	0.24586

ID	Name	Class	Target	Control	p-value target	p-value control
MA0081	SPIB	3.5.2.0	5120	469	0.99998	0.11261
MA0082	Squamosa	4.4	647	60	0.96851	0.30655
MA0084	SRY	4.7.1.0	2629	263	1	0.01099
MA0086	sna	2.3.2.2	2896	248	0.28926	0.57384
MA0088	Znf143	2.3	51	3	0.00597	0.81412
MA0096	bZIP910	1.1	253	22	0.54111	0.51873
MA0098	ETS1	3.5.2	20289	1720	0.00173	0.80678
MA0102	CEBPA	1.1.3.0	1531	157	1	0.02002
MA0107	RELA	4.1.1.0	385	36	0.94321	0.34079
MA0114	HNF4A	2.1.2.11	147	16	0.9984	0.20988
MA0124	NKX3-1	3.1.1.15	1408	132	0.99892	0.18878
MA0126	ovo	2.3.2.2	1967	173	0.77972	0.42082
MA0132	Pdx1	3.1	3915	378	1	0.01898
MA0133	BRCA1	unknown	5064	458	0.9993	0.17622
MA0139	CTCF	2.3	64	6	0.7656	0.47375
MA0141	Esrrb	2.1.1.5	381	33	0.52852	0.51678
MA0148	FOXA1	3.3	677	70	1	0.078
MA0156	FEV	3.5.2	1502	139	0.99615	0.22316
MA0157	FOXO3	3.3	1867	193	1	0.00857
MA0165	Abd-B	3.1.1.1	5311	510	1	0.0106
MA0195	Lim3	3.1	5419	528	1	0.00382
MA0197	Oct	3.1	3243	319	1	0.01275
MA0199	Optix	3.1	2141	206	1	0.07098
MA0201	Ptx1	3.1	716	75	1	0.05858
MA0205	Trl	2.3	2182	226	1	0.0044
MA0209	ap	3.1	2111	220	1	0.00394
MA0215	btn	3.1	2256	241	1	0.00084
MA0219	ems	3.1.1.6	5869	567	1	0.00508
MA0239	prd	3.1.1.9	1967	173	0.77972	0.42082
MA0244	slbo	2.3	3569	344	1	0.02509
MA0261	lin-14	unknown	4909	475	1	0.00874
MA0274	ARR1	1.1.1.5	6084	552	0.99992	0.13466
MA0277	AZF1	2.3	1927	190	1	0.04014
MA0286	CST6	1.1.2	260	23	0.65253	0.48407
MA0289	DAL80	2.2.1.2	620	49	0.01377	0.75481
MA0297	FKH2	3	517	53	0.99997	0.12393
MA0300	GAT1	2.2.1.2	488	41	0.26742	0.59408
MA0303	GCN4	1.1.1.5	162	18	0.99968	0.16941
MA0313	HAP2	4.8.1.0	9384	829	0.97642	0.28322
MA0314	HAP3	4.8.1.0	102	10	0.91146	0.38648
MA0317	HCM1	3.3	3588	335	1	0.08582
MA0332	MET28	1.1	1815	150	0.02634	0.72409
MA0336	MGA1	3.3.3.4	17	1	0.08154	0.76872
MA0345	NHP6A	4.7	48	5	0.92045	0.39733
MA0346	NHP6B	4.7	87	8	0.74216	0.47468
MA0368	RIM101	2.3	1207	110	0.96682	0.30506
MA0386	TBP	4.6.1	11	1	0.61071	0.61225

ID	Name	Class	Target	Control	p-value target	p-value control
MA0388	SPT23	unknown	4662	450	1	0.01146
MA0398	SUM1	unknown	2959	284	1	0.04346
MA0419	YAP7	1.1.1	159	14	0.60477	0.5075
MA0445	D	4.7	713	79	1	0.01854
MA0446	fkh	3.3	859	90	1	0.04122
MA0453	nub	2.3	223	17	0.03559	0.72773
MA0458	slp1	3.3	1142	113	1	0.08405
MA0459	tll	2.1.2.15	203	26	1	0.03476
MA0138	REST	2.3.2.2	0	0	1	1

A.4 *De Novo* motif discovery in motif-negative set of sequences and comparison to Jaspar_Core profiles using TOMTOM



Motif ID: ChIPMunk00001

Short Name:

Size: 19 bp

IC: 14.351

GC: 63.5%

Long Name:

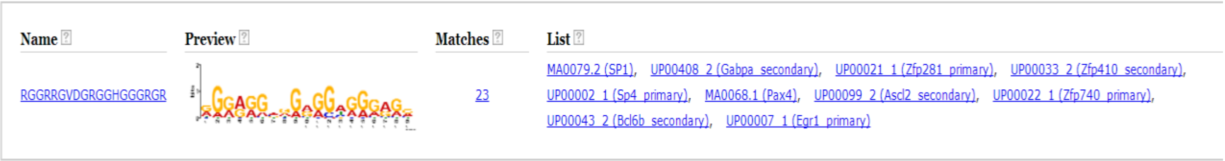
Consensus Motif: rGGrrGnGGrGGwGGGrGr

Matrix

A	C	G	T
239.0	13.0	315.0	76.0
86.0	19.0	538.0	0.0
199.0	18.0	415.0	11.0
384.0	0.0	233.0	26.0
221.0	18.0	400.0	4.0
105.0	42.0	487.0	9.0
235.0	209.0	155.0	44.0
161.0	23.0	326.0	133.0

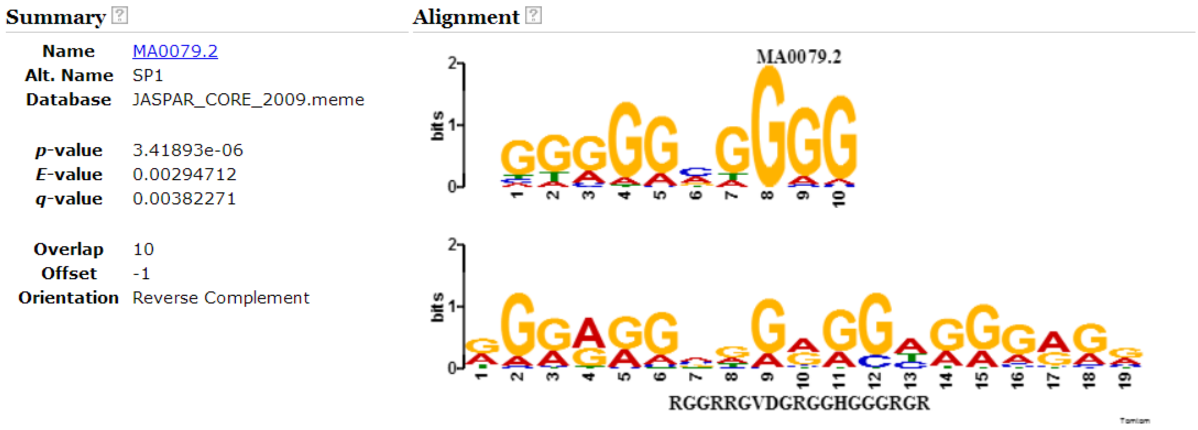
□ Show reverse complement ☑ Scale by IC ☑ Sort by Frequency

A. Sequence logo and weight matrix from de novo motif discovery using ChIPMunk in MotifLab



Name	Preview	Matches	List
RGGRRGVDGRGGHGGGRGR		23	MA0079.2 (SP1), UP00408_2 (Gabpa_secondary), UP00021_1 (Zfp281_primary), UP00033_2 (Zfp410_secondary), UP00002_1 (Sp4_primary), MA0068.1 (Pax4), UP00099_2 (Ascl2_secondary), UP00022_1 (Zfp740_primary), UP00043_2 (Bcl6b_secondary), UP00007_1 (Egr1_primary)

B. Motif comparison using TOMTOM pulls out 23 motifs from Jaspar_Core with similarity to the *de novo* motif



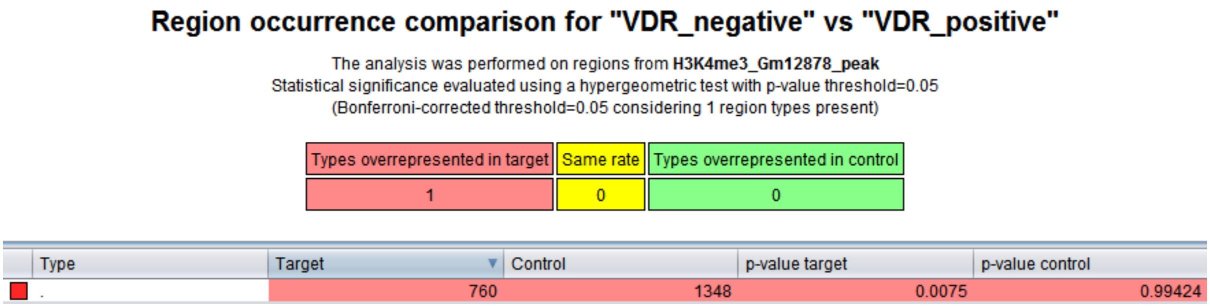
Summary

Name	MA0079.2
Alt. Name	SP1
Database	JASPAR_CORE_2009.meme
p-value	3.41893e-06
E-value	0.00294712
q-value	0.00382271
Overlap	10
Offset	-1
Orientation	Reverse Complement

Alignment

C. *Sp1* motif shows highest similarity to motif from *de novo* motif discovery

A.5 Comparison of enrichment of different chromatin marks between sequence collections in Ramagopalan dataset

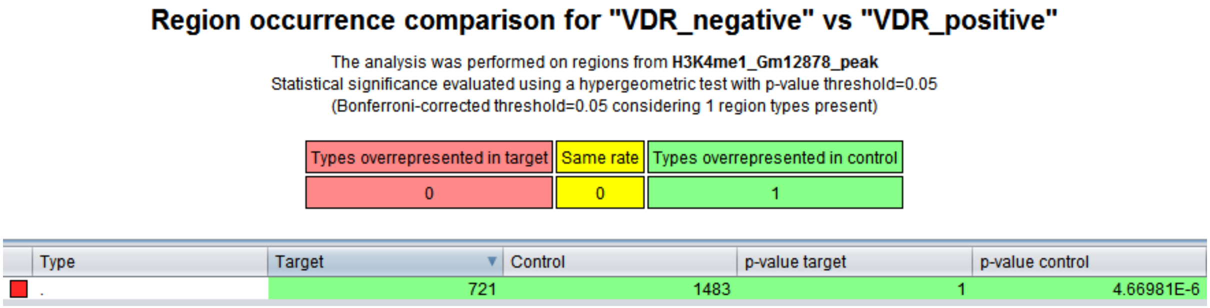


Region occurrence comparison for "VDR_negative" vs "VDR_positive"

The analysis was performed on regions from **H3K4me3_Gm12878_peak**
Statistical significance evaluated using a hypergeometric test with p-value threshold=0.05
(Bonferroni-corrected threshold=0.05 considering 1 region types present)

Types overrepresented in target	Same rate	Types overrepresented in control
1	0	0

	Type	Target	Control	p-value target	p-value control
🟥	.	760	1348	0.0075	0.99424

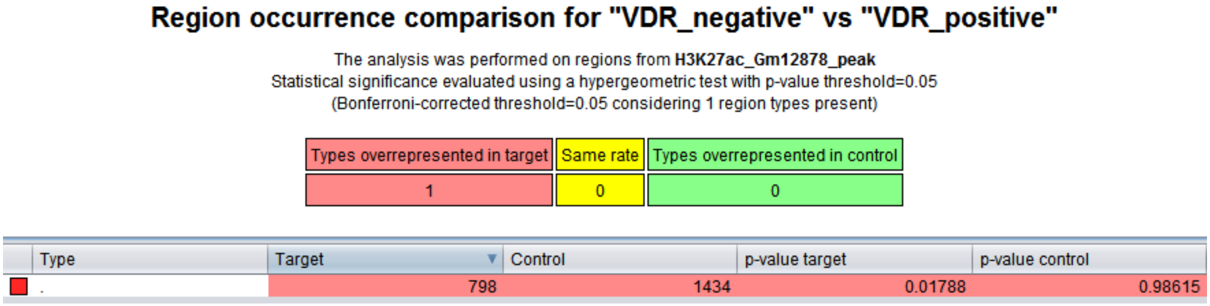


Region occurrence comparison for "VDR_negative" vs "VDR_positive"

The analysis was performed on regions from **H3K4me1_Gm12878_peak**
Statistical significance evaluated using a hypergeometric test with p-value threshold=0.05
(Bonferroni-corrected threshold=0.05 considering 1 region types present)

Types overrepresented in target	Same rate	Types overrepresented in control
0	0	1

	Type	Target	Control	p-value target	p-value control
🟥	.	721	1483	1	4.66981E-6

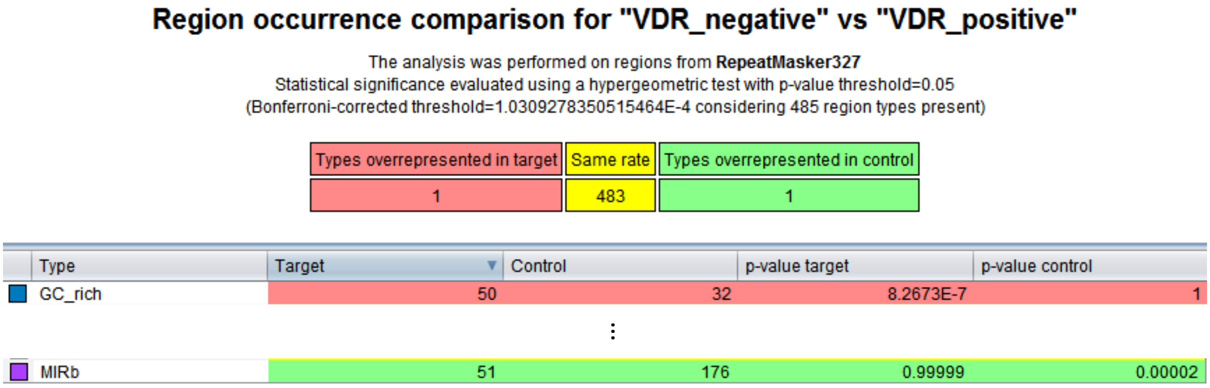


Region occurrence comparison for "VDR_negative" vs "VDR_positive"

The analysis was performed on regions from **H3K27ac_Gm12878_peak**
Statistical significance evaluated using a hypergeometric test with p-value threshold=0.05
(Bonferroni-corrected threshold=0.05 considering 1 region types present)

Types overrepresented in target	Same rate	Types overrepresented in control
1	0	0

	Type	Target	Control	p-value target	p-value control
🟥	.	798	1434	0.01788	0.98615



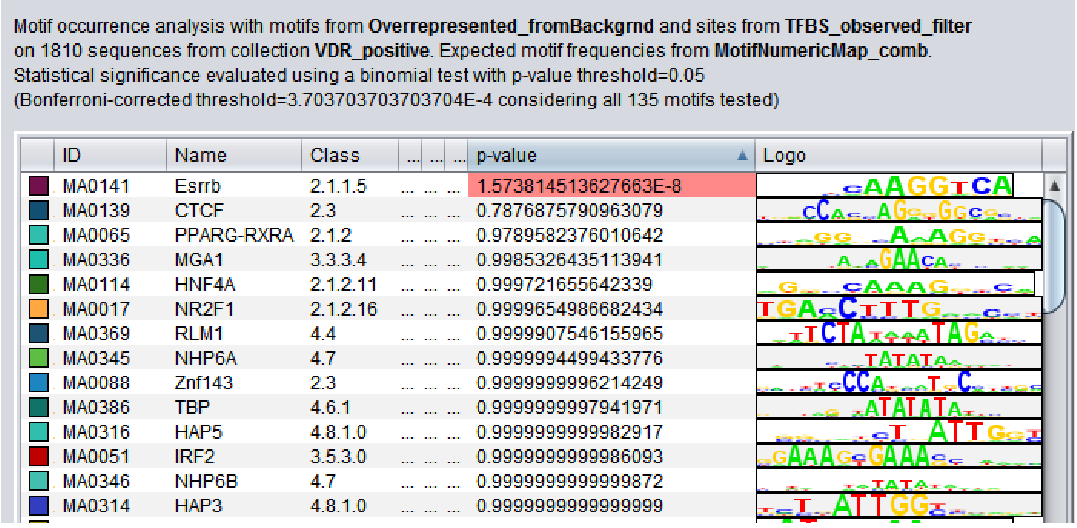
Region occurrence comparison for "VDR_negative" vs "VDR_positive"

The analysis was performed on regions from **RepeatMasker327**
Statistical significance evaluated using a hypergeometric test with p-value threshold=0.05
(Bonferroni-corrected threshold=1.0309278350515464E-4 considering 485 region types present)

Types overrepresented in target	Same rate	Types overrepresented in control
1	483	1

	Type	Target	Control	p-value target	p-value control
🟦	GC_rich	50	32	8.2673E-7	1
			⋮		
🟪	MIRb	51	176	0.99999	0.00002

Motif enrichment in region 20 bases from VDR binding sites



Motif occurrence analysis with motifs from **Overrepresented_fromBackgrnd** and sites from **TFBS_observed_filter** on 1810 sequences from collection **VDR_positive**. Expected motif frequencies from **MotifNumericMap_comb**.
Statistical significance evaluated using a binomial test with p-value threshold=0.05
(Bonferroni-corrected threshold=3.7037037037037E-4 considering all 135 motifs tested)

	ID	Name	Class	...	...	...	p-value	Logo
■	MA0141	Esrrb	2.1.1.5	...	...	...	1.573814513627663E-8	
■	MA0139	CTCF	2.3	...	...	...	0.7876875790963079	
■	MA0065	PPARG-RXRA	2.1.2	...	...	...	0.9789582376010642	
■	MA0336	MGA1	3.3.3.4	...	...	...	0.9985326435113941	
■	MA0114	HNF4A	2.1.2.11	...	...	...	0.99972165564239	
■	MA0017	NR2F1	2.1.2.16	...	...	...	0.9999654986682434	
■	MA0369	RLM1	4.4	...	...	...	0.9999907546155965	
■	MA0345	NHP6A	4.7	...	...	...	0.9999994499433776	
■	MA0088	Znf143	2.3	...	...	...	0.9999999996214249	
■	MA0386	TBP	4.6.1	...	...	...	0.9999999997941971	
■	MA0316	HAP5	4.8.1.0	...	...	...	0.9999999999982917	
■	MA0051	IRF2	3.5.3.0	...	...	...	0.9999999999986093	
■	MA0346	NHP6B	4.7	...	...	...	0.9999999999999872	
■	MA0314	HAP3	4.8.1.0	...	...	...	0.9999999999999999	

A.6 Enriched motifs common to Ramagopalan and Heikkinen analysis datasets

ID	Short Name	Classification
MA0096	bZIP910	1.1
MA0332	MET28	1.1
MA0120	id1	2.3
MA0013	br_Z4	2.3
MA0146	Zfx	2.3
MA0010	br_Z1	2.3
MA0073	RREB1	2.3
MA0277	AZF1	2.3
MA0012	br_Z3	2.3
MA0431	YML081W	2.3
MA0205	Trl	2.3
MA0425	YGR067C	2.3

ID	Short Name	Classification
MA0417	YAP5	1.1.1
MA0117	Mafb	1.1.1.3
MA0274	ARR1	1.1.1.5
MA0303	GCN4	1.1.1.5
MA0018	CREB1	1.1.2
MA0286	CST6	1.1.2
MA0102	CEBPA	1.1.3.0
MA0043	HLF	1.1.4.0
MA0055	Myf	1.2.2.0
MA0449	h	1.2.5.1
MA0357	PHO4	1.2.5.3
MA0003	TFAP2A	1.6.1

MA0443	btd	2.3
MA0285	CRZ1	2.3
MA0139	CTCF	2.3
MA0268	ADR1	2.3
MA0244	slbo	2.3
MA0453	nub	2.3
MA0373	RPN4	2.3
MA0260	che-1	2.3
MA0368	RIM101	2.3
MA0333	MET31	2.3
MA0011	br_Z2	2.3
MA0088	Znf143	2.3
MA0362	RDS2	2.4
MA0280	CAT8	2.4
MA0361	RDS1	2.4
MA0283	CHA4	2.4
MA0297	FKH2	3
MA0199	Optix	3.1
MA0201	Ptx1	3.1
MA0356	PHO2	3.1
MA0173	CG11617	3.1
MA0197	Oct	3.1
MA0033	FOXL1	3.3
MA0041	FOXD3	3.3
MA0042	FOXI1	3.3
MA0458	slp1	3.3
MA0317	HCM1	3.3
MA0148	FOXA1	3.3
MA0446	fkh	3.3
MA0047	FOXA2	3.3
MA0040	FOXQ1	3.3
MA0157	FOXO3	3.3
MA0002	RUNX1	4.11
MA0242	run-Bgb	4.11
MA0082	Squamosa	4.4
MA0346	NHP6B	4.7
MA0345	NHP6A	4.7
MA0445	D	4.7
MA0045	HMG-IY	0.2.1
MA0123	ABI4	0.5.2.0

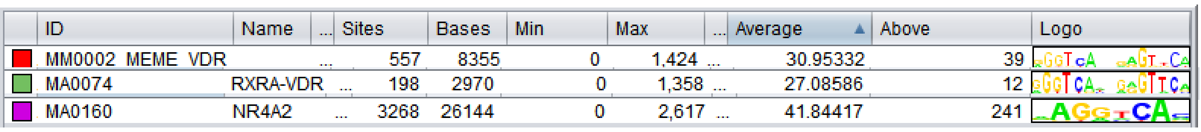
MA0141	Esrrb	2.1.1.5
MA0074	RXRA-VDR	2.1.2
MA0065	PPARG-RXRA	2.1.2
MA0114	HNF4A	2.1.2.11
MA0459	tll	2.1.2.15
MA0017	NR2F1	2.1.2.16
MA0160	NR4A2	2.1.2.17
MA0300	GAT1	2.2.1.2
MA0289	DAL80	2.2.1.2
MA0079	SP1	2.3.1.0
MA0162	Egr1	2.3.2.1
MA0015	Cf2_II	2.3.2.2
MA0039	Klf4	2.3.2.2
MA0057	MZF1_5-13	2.3.2.2
MA0049	hb	2.3.2.2
MA0341	MSN2	2.3.3.0
MA0340	MOT3	2.3.3.0
MA0404	TBS1	2.4.1
MA0375	RSC30	2.4.1
MA0399	SUT1	2.4.1
MA0324	LEU3	2.4.1.0
MA0165	Abd-B	3.1.1.1
MA0124	NKX3-1	3.1.1.15
MA0062	GABPA	3.5.2
MA0081	SPIB	3.5.2.0
MA0076	ELK4	3.5.2.0
MA0028	ELK1	3.5.2.0
MA0050	IRF1	3.5.3.0
MA0052	MEF2A	4.4.1.1
MA0386	TBP	4.6.1
MA0084	SRY	4.7.1.0
MA0060	NFYA	4.8.1.0
MA0314	HAP3	4.8.1.0
MA0315	HAP4	4.8.1.0
MA0316	HAP5	4.8.1.0
MA0149	EWSR1-FLI1	unknown
MA0388	SPT23	unknown
MA0398	SUM1	unknown
MA0213	brk	unknown
MA0261	lin-14	unknown
MA0133	BRCA1	unknown

A.7 Region enrichment in Heikkinen dataset

Feature	Percent overlap of binding sites from motif scanning (PPV)		
	MEME *de novo*	MA0074.1 (RXR-VDR)	VDR-like (NR4A2)
DNAse HotSpots	86.3	94.9	89.3
TFBS_ChIPSeq	57.4	60.6	60.1
H3K27ac	43.5	32.7	44.9
H3K4me1	53	45.8	53.8
H3K4me3	39.9	30.4	39.5
H3K9ac	42.8	31	43.2
H3K27me3	36.4	53	35.5
FAIRE-Seq	48.1	53.8	50.6
RepeatMasker327	37.1	26	30.3
CpG islands	8.3	7.1	9.3
CCDS	41.2	31.5	41.9
Ensembl Genes	61.8	47.6	63.1

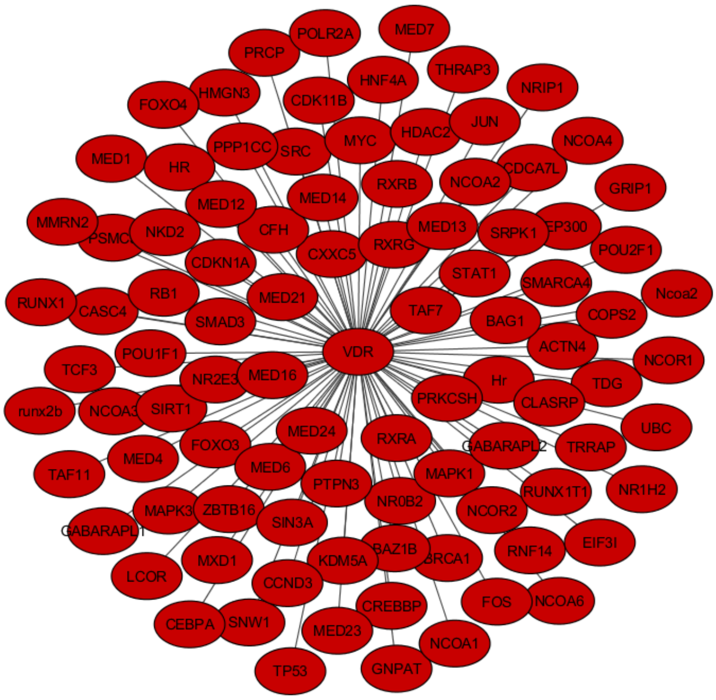
*Significant difference in region overrepresentation between MEME *de novo* and VDR-like, **Significant difference in region overrepresentation between classic VDR and VDR-like. These features were similar in MEME-motif sequences and the classic VDR sequences.

A.8 Distance of repeat MIRb relative to binding sites of motifs of interest

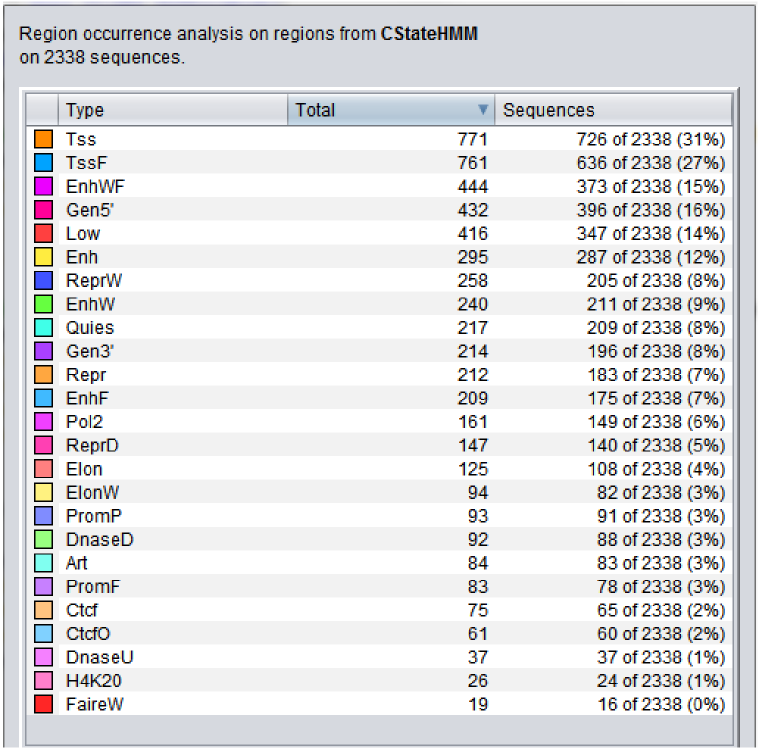


	ID	Name	...	Sites	Bases	Min	Max	...	Average ▲	Above	Logo
🟥	MM0002_MEME_VDR		...	557	8355	0	1,424	...	30.95332	39	
🟩	MA0074	RXRA-VDR	...	198	2970	0	1,358	...	27.08586	12	
🟪	MA0160	NR4A2	...	3268	26144	0	2,617	...	41.84417	241	

A.9 Proteins that interact or are involved in the formation of complexes with VDR retrieved from BioGrid 3.2 and visualised in Cytoscape 3.0.2



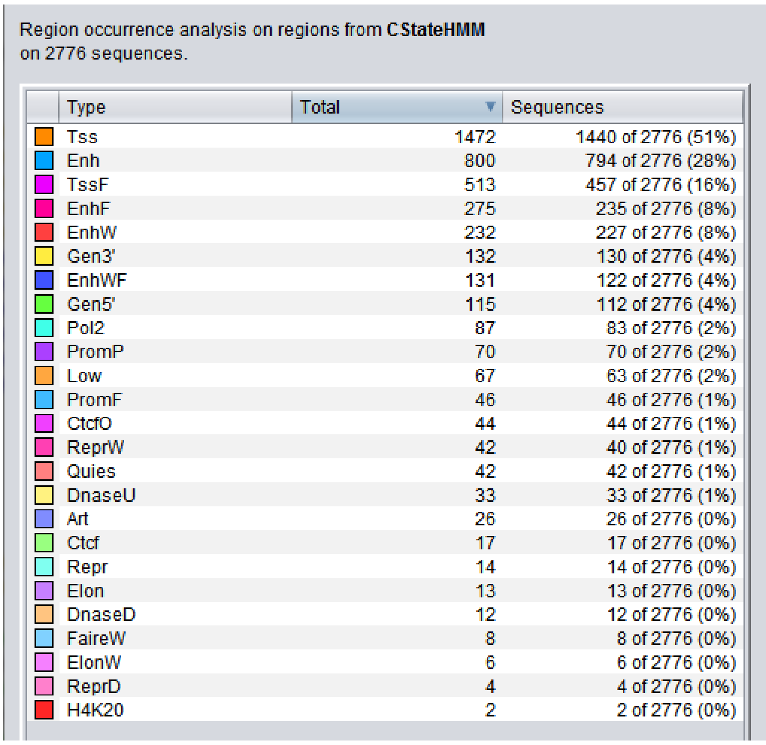
A.10 Region analysis using segmentation data derived using ChromHMM and a combination track derived from a combination of data derived using ChromHMM and Segway on analysis datasets



Region occurrence analysis on regions from **CStateHMM**
on 2338 sequences.

	Type	Total ▼	Sequences
🟧	Tss	771	726 of 2338 (31%)
🟦	TssF	761	636 of 2338 (27%)
🟪	EnhWF	444	373 of 2338 (15%)
🟥	Gen5'	432	396 of 2338 (16%)
🟥	Low	416	347 of 2338 (14%)
🟨	Enh	295	287 of 2338 (12%)
🟦	ReprW	258	205 of 2338 (8%)
🟩	EnhW	240	211 of 2338 (9%)
🟦	Quies	217	209 of 2338 (8%)
🟪	Gen3'	214	196 of 2338 (8%)
🟧	Repr	212	183 of 2338 (7%)
🟦	EnhF	209	175 of 2338 (7%)
🟪	Pol2	161	149 of 2338 (6%)
🟥	ReprD	147	140 of 2338 (5%)
🟥	Elon	125	108 of 2338 (4%)
🟨	ElonW	94	82 of 2338 (3%)
🟦	PromP	93	91 of 2338 (3%)
🟩	DnaseD	92	88 of 2338 (3%)
🟦	Art	84	83 of 2338 (3%)
🟪	PromF	83	78 of 2338 (3%)
🟧	Ctcf	75	65 of 2338 (2%)
🟦	CtcfO	61	60 of 2338 (2%)
🟪	DnaseU	37	37 of 2338 (1%)
🟥	H4K20	26	24 of 2338 (1%)
🟥	FaireW	19	16 of 2338 (0%)

A. Region occurrence in Heikkinen dataset using chromatin segmentation data derived using the ChromHMM showing the enrichment of TSS and TSS flanking regions



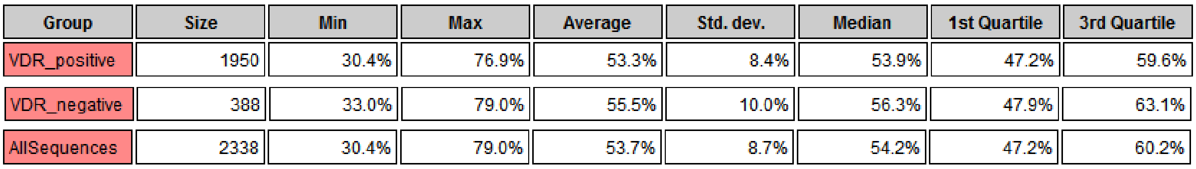
Region occurrence analysis on regions from **CStateHMM**
on 2776 sequences.

	Type	Total ▼	Sequences
🟧	Tss	1472	1440 of 2776 (51%)
🟦	Enh	800	794 of 2776 (28%)
🟪	TssF	513	457 of 2776 (16%)
🟥	EnhF	275	235 of 2776 (8%)
🟥	EnhW	232	227 of 2776 (8%)
🟨	Gen3'	132	130 of 2776 (4%)
🟦	EnhWF	131	122 of 2776 (4%)
🟩	Gen5'	115	112 of 2776 (4%)
🟦	Pol2	87	83 of 2776 (2%)
🟪	PromP	70	70 of 2776 (2%)
🟧	Low	67	63 of 2776 (2%)
🟦	PromF	46	46 of 2776 (1%)
🟪	CtcfO	44	44 of 2776 (1%)
🟥	ReprW	42	40 of 2776 (1%)
🟥	Quies	42	42 of 2776 (1%)
🟨	DnaseU	33	33 of 2776 (1%)
🟦	Art	26	26 of 2776 (0%)
🟩	Ctcf	17	17 of 2776 (0%)
🟦	Repr	14	14 of 2776 (0%)
🟪	Elon	13	13 of 2776 (0%)
🟧	DnaseD	12	12 of 2776 (0%)
🟦	FaireW	8	8 of 2776 (0%)
🟪	ElonW	6	6 of 2776 (0%)
🟥	ReprD	4	4 of 2776 (0%)
🟥	H4K20	2	2 of 2776 (0%)

C. Region occurrence in Ramagopalan dataset using chromatin segmentation data derived using the ChromHMM showing the enrichment of TSS regions

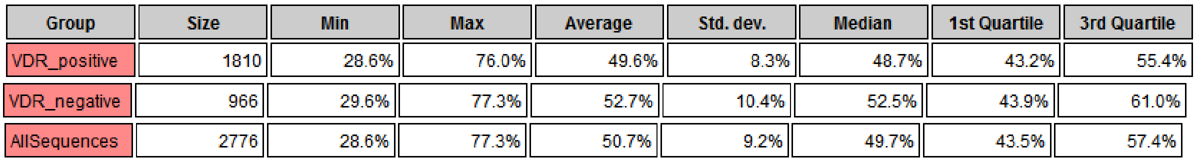
A.11 GC Content analysis of sequences in negative and positive sequence sets

Heikkinen:



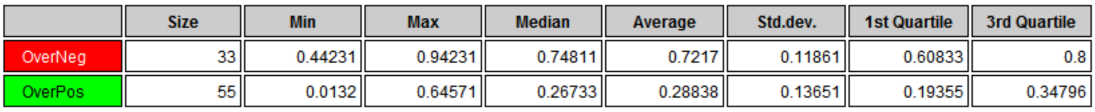
Group	Size	Min	Max	Average	Std. dev.	Median	1st Quartile	3rd Quartile
VDR_positive	1950	30.4%	76.9%	53.3%	8.4%	53.9%	47.2%	59.6%
VDR_negative	388	33.0%	79.0%	55.5%	10.0%	56.3%	47.9%	63.1%
AllSequences	2338	30.4%	79.0%	53.7%	8.7%	54.2%	47.2%	60.2%

Ramagopalan:

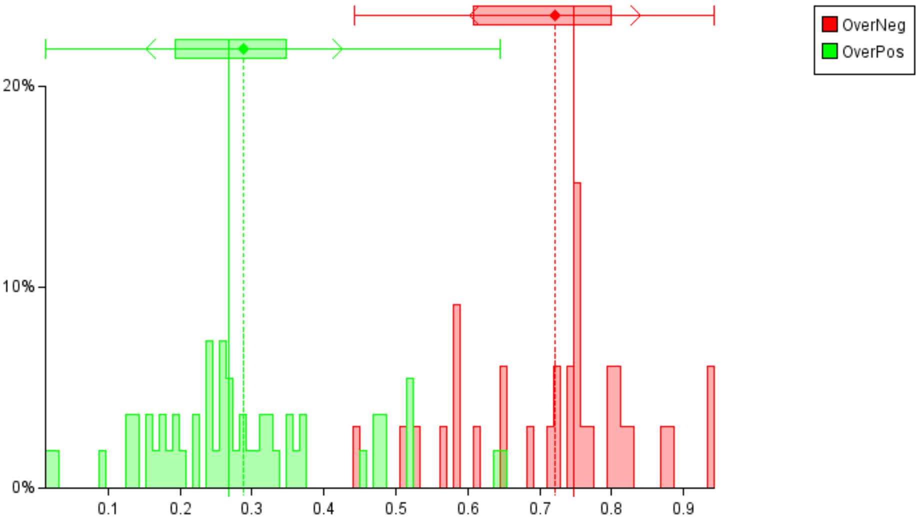


Group	Size	Min	Max	Average	Std. dev.	Median	1st Quartile	3rd Quartile
VDR_positive	1810	28.6%	76.0%	49.6%	8.3%	48.7%	43.2%	55.4%
VDR_negative	966	29.6%	77.3%	52.7%	10.4%	52.5%	43.9%	61.0%
AllSequences	2776	28.6%	77.3%	50.7%	9.2%	49.7%	43.5%	57.4%

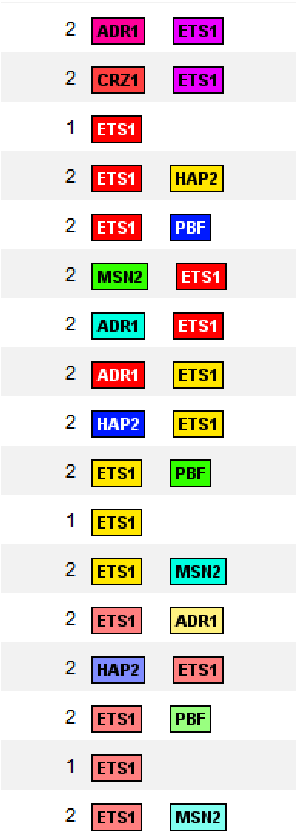
A.12 GC content analysis of motifs enriched in negative and positive sequence sets (Ramagopalan)



	Size	Min	Max	Median	Average	Std.dev.	1st Quartile	3rd Quartile
OverNeg	33	0.44231	0.94231	0.74811	0.7217	0.11861	0.60833	0.8
OverPos	55	0.0132	0.64571	0.26733	0.28838	0.13651	0.19355	0.34796



A.13 Motif pairs identified by ModuleSearcher in the Ramagopalan dataset



2	ADR1	ETS1
2	CRZ1	ETS1
1	ETS1	
2	ETS1	HAP2
2	ETS1	PBF
2	MSN2	ETS1
2	ADR1	ETS1
2	ADR1	ETS1
2	HAP2	ETS1
2	ETS1	PBF
1	ETS1	
2	ETS1	MSN2
2	ETS1	ADR1
2	HAP2	ETS1
2	ETS1	PBF
1	ETS1	
2	ETS1	MSN2