Gene regulation in three dimensions

Kristian Thoresen Wittek

June 2016

MASTER THESIS

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

Supervisor: Finn Drabløs
Co supervisor: Morten Rye

Acknowledgement

This thesis was written at the department of molecular medicine of the Norwegian University of Science and technology (NTNU). I would like to thank my supervisor, professor Finn Drabløs for being a vitally important reference and source of advices and guiding me during difficulties. I would also like to thank my co supervisor Morten Rye. Additionally, I would like to thank Kjetil Klepper for helping me solve problems relating to Motiflab. Thank you for taking the time to answer my questions.

Last, but not least I would like to thank Julie my fiancé who has supported me and been there for me when I needed it the most.

K.T.W

Abstract

Gene regulation has traditionally mainly been viewed as a 1D and possibly 2D process. In the 1D view the genome is seen as a linear string of nucleotides, where one or more transcription factors (TFs) bind to transcription factor binding sites (TFBSs), and thereby regulate the expression of genes that are nearby in the linear genomic sequence. The 2D process can be described with DNA looping where proteins bind to distal binding sites and bring them in close proximity of the transcription start site (TSS). However, in reality such interactions take place in 3D space meaning multiple interactions, possibly also between separate chromosomes. The binding of TFs to genomic DNA is experimentally studied using mainly the ChIP-Seq protocol. In some cases, the motif for transcription factor binding is found only in a subset of peaks. In this project we looked at the phenomenon of motifless (ML) binding in Vitamin D Receptor (VDR) ChIP-Seq experiments where interestingly as much as 50% of the identified peaks seem to be ML. These motifless binding sites may be caused by 3D interactions or DNA looping where the DNA strand folds back and interacts with itself or distally with another chromosome. Here we use statistics and computer software to measure localization of ML sites in 3D as well as 1D space. This is completed using previously generated genomic annotation data from ENCODE and other datasets to measure differences between ML peaks and regions containing known TF binding sites.

Table of Contents

List of Figures

List of Tables

List of abbreviations and labels

Bp	Base pairs
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
ChIP-Seq	Chromatin Immunoprecipitation followed by high throughput sequencing
ChIP-seq	Chromatin immunoprecipitation sequencing
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
DHSs	DNase 1 hypersensitivity sites
DNA	Deoxyribonucleic acid
ENCODE	Encyclopaedia of DNA 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
HOT	High Occupancy target
LOT	Low Occupancy Target
MB	motifbound
MCFDR	Monte Carlo False Discovery Rate
MEME	Multiple Expectation Maximization for Motif Elicitation
ML	motifless
PSSM	Position specific scoring matrix
PWM	Position weight matrix
RAD21, SMC3, CTCF	Cohesin associated proteins
RXR	Retinoid X receptor
TAD	Topological Associating Domain
TF	Transcription factor
TFBS	observed Motifs identified from motif scanning
TFBS	Transcription factor binding site
TSS	Transcription start site
VDR	Vitamin D receptor
VDRMB	Vitamin D receptor motifbound
VDRML	Vitamin D receptor motifless

1 Introduction

In all living cells there is a recipe that describes the building blocks of life called deoxyribonucleic acid or DNA. This molecule carries most of the genetic instructions used in the development, functioning and reproduction of all known living organisms. The nucleotides adenine, thymine, cytosine and guanine are placed on a linear strand of alternating 2-deoxyribose and phosphate that makes up DNA. Genes are segments of the DNA strand that direct the manufacture of specific molecular end products that are essential in growth, development and maintenance of all biological processes in living things. The complete set of genes and DNA are wrapped around histones and located inside the cell nucleus. The entire DNA strand is often referred to as the genome where eukaryotic genomes usually comprise of coding and non-coding DNA. Genes consist of introns and exons where exons make up the coding and introns the non-coding regions of genes. In between genes are a subset of noncoding DNA called intergenic regions as well as regulatory regions that affect genes and how they are transcribed. Eukaryotes are made more complex as one gene can be responsible for more than one product by posttranscriptional processing where introns are removed which result in expression of exons that can be rearranged and can result in different molecular end products. The entire DNA sequence is then organized into 23 chromosome pairs that altogether make up the human genome. Genes are then recognized by a set of proteins that interpret the genetic code and produce a string of mRNA that is later used as a recipe to create long chains of amino acids. These chains are then usually folded by chaperone enzymes into proteins that have a specific function in the body. Regulated gene expression is crucial to the proper growth and survival of an organism. In general, gene expression is the process by which the production of individual, genetically-encoded proteins are uniquely regulated in response to specific environmental and developmental signals. Without it, the genetic code of a chromosome is the equivalent of a computer data file without the appropriate application to run it.

Because of the large number of applications included in this project all tools and workbenches are written in *italic*.

1.1 Gene regulation

1.1.1 Transcription factors and their role in gene regulation

A group of proteins called transcription factors act to regulate gene expression either positively or negatively. To better achieve this modular functional domains exist in these proteins and act in different ways to regulate the activity of the transcription factor itself and also the gene they regulate [1]. Regulation of TFs are either by the ligand binding in the ligand-binding domain in the TF or by interaction with other TFs through the activation function domain. Other regulatory factors of TFs include posttranslational modifications like phosphorylation, ubiquitination and a host of others which activates or suppresses or even marks them for destruction. Another important part of the TF is the DNA binding domain (DBD) which interacts with DNA. A specific string of nucleotides which serves as template/motif for DNA, called response element serves as a template for DNA by the DBD and forms the basis for most sequence specific TF activity. TFs work either alone or with others as mono- or hetero- dimers in binding DNA.

The DBD in the TF contains amino acids that allow for some specificity regarding nucleotide binding. The TFs bind proximally or distally to regulated regions and will through several mechanisms exert their effects. In cases where TFs bind regions that are distal to the genes being regulated, DNA looping has been suggested to be a mechanism with which these distally located TFs are able to interact with their target genes. Also in cases where they are located proximally, co-activators or co-repressors interact with the transcription machinery and act as a bridge between them to initiate transcription [2].

Other distal regulatory factors are silencers, insulators and other DNA regions that are distal to the TSS, but in some aspect influence transcription initiation. The immediate function of regulatory regions on transcription are dependent on the type of TFs that bind and can act as either activators or repressors of transcription. Regulatory DNA regions together with regulatory proteins make up the expression and repression of genes. Regulatory factors achieve their interaction with DNA by hydrogen bonds and Van der Waals forces.

Transcription factors work in a combinatorial manner where the control of single genes falls under the influence of several TFs acting together. This allows for a small number of TFs to regulate a large number of genes. To be able to regulate gene transcription, TFs bind to regions that are usually located at the start of genes called response elements [3]. The response element is a small part of a larger core promotor which is the minimal portion of the promoter required to properly initiate transcription. Core promotors may consist of TATA-binding region (TATA-box) and or an initiator region (INR) located proximal to the translation start site (TSS). More regulatory regions are found in introns, exons and non-coding regions of DNA. These are usually termed enhancers and can influence gene expression from many thousand base pairs away [4]. Enhancer genes and activator proteins promote transcription by bending the DNA strand bringing them in close proximity of the promotor site that in a complex with other TFs allow for RNA polymerase to bind and initiate transcription of RNA. Bending the DNA strand putting enhancer genes and activator proteins in close proximity of the promotor site is called DNA looping and is an important DNA regulatory function. This feature makes it possible for distally located genes to control gene expression directly through close interaction [5].

1.1.2 TFs and cis regulatory modules

Cis regulatory modules (CRMs) are stretches of DNA usually 100-1000 DNA base pairs in length where different TFs can bind to the CRM binding domains [6]. TFs with binding sites within CRM allow for combinatorial control over target genes dependent on the cofactors and concentration of specific TFs at a specific time and place. Control is also exerted through different CRMs acting together on single genes. The functional role of CRMs on gene regulation are believed to work by three known mechanisms. One mechanism is called the DNA scanning model and involves the assembly of TFs and its cofactors at the CRM and subsequent scanning of DNA until the general transcription machinery complex is found. The second mechanism is believed to be a combination of the two first mechanisms termed the facilitated tracking model [5]. The assembly of the TFs and their cofactors takes place in the CRM. Although, this complex scans for intervening DNA sequence for the promotor, 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 promotor is found. The third mechanism is the previously mentioned DNA looping where the promotor, after TF binding, interacts with DNA in the CRM and this is usually referred to as the DNA looping model. While the underlying chemistry of DNA looping is common to all systems, the precise biochemical mechanism of DNA looping in different systems can vary [2].

1.1.3 The role of genomic 3D interactions in gene regulation

1.1.3.1 DNA looping and cohesin

DNA looping appears to have been chosen by nature in such a variety because it solves problems both of binding and of geometry because this enables binding of a protein to multiple sites on the DNA strand that again enables high occupancy of DNA sites by a low number of proteins [2].

DNA looping is facilitated by the cohesion complex which is known to influence the genomic 3D structure [7, 8]. Cohesin consists of four core subunits being SMC1A, SMC3, RAD21 and STAG1 or STAG2 [9]. One emerging aspect of cohesin is its ability to regulate gene transcription by binding to CTCF regions, rearranging the genome 3D structure to regulate transcription [10]. More specifically it has been confirmed that CTCF and cohesion facilitates inter domain looping, based on actively transcribed DNA sites, at the immunoglobulin, HoxA, β-globin, and interferon gamma loci [11, 12]. In contrast CTCF has been linked with local intra domain interactions in B-cells [13]. To make this even more complex Carbon-Copy Chromosome Conformation Capture (5C) analysis revealed that long range interactions frequently span CTCF binding sites suggesting that CTCF could be working as a barrier or insulator in repressed regions as well as in a positive manner in active regions looping enhancers and promotors together [14]. Despite being commonly linked to CTCF, cohesin has recently been given a CTCF-independent role in transcriptional regulation and chromatin looping [15].

1.1.3.2 TADs and HOT regions

Other features that affect DNA 3D structure are called topologically associating domains or TADs. These domains show high local DNA interaction frequencies compared to other DNA regions. TADs are conserved across mammalian species and form clusters of insulated neighbourhoods that show low interaction frequency between the different domains [16]. These neighbourhoods are defined by cohesin-associated CTCF-CTCF loops that make up the structural framework for TADs. These structures are rarely affected by human sequence variation, but are frequently altered by somatic mutations in cancer [17-19].

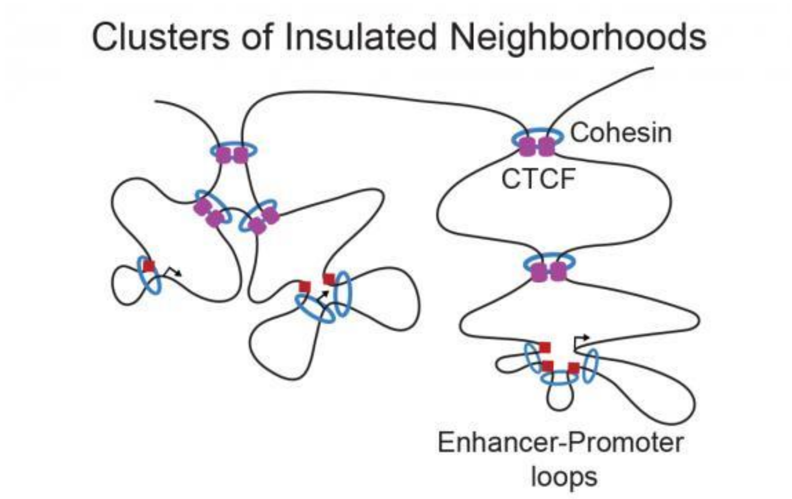


Figure 1 in this schematic of a chromosome's 3D structure, the DNA loops (black) are anchored by CTCF proteins (purple) at their bases. Regulatory elements, called enhancers (red) only affect the genes (black arrows) that reside in the loops with them. The protein cohesin (blue rings) forms the looped structures. The entire structure forms an insulated neighbourhood with decreased interaction rates with down or upstream insulated neighbourhoods. Adapted by permission from Elsevier: Cell Press [16], copyright (2016).

However, TADs are not to be confused with high-occupancy target (HOT) regions that are compact genome loci that is highly transcribed together in the genome as well as binding many different TFs. On the opposite hand there are LOT regions consisting of intergenic regions (IGR) that are stretches of DNA sequences located between genes and a subset of noncoding DNA [20-22]. Specifically, mammalian HOT regions are regulatory hubs that integrate the signals from diverse regulatory pathways to quantitatively tune the promoter for RNA polymerase II recruitment. TADs and HOT regions both play an important part in TF gene regulation by affecting genome structure as well as being integrating hubs for gene regulation.

1.1.3.3 The role of epigenetic modifications in gene regulation

Recruitment of TFs to regulatory elements like members of the DNA polymerase family, histone modifiers and co-activators form a multi-unit complex responsible for the actual transcription process. Aside from this other proteins interact with DNA influencing regulation by rearranging the chromatin state and also the three dimensional structure of the genome [23, 24]. DNA stretches of approximately 150 bp are wrapped around octets of histone proteins to form nucleosomes. These and other DNA associated proteins make up chromatin which is a complex structure that works as a scaffold that again directs DNA activity [25]. Combinations of these DNA associated proteins make up different chromatin formations that is functionally linked to the expression state of the DNA they bind. One of these states are heterochromatin which is a dense packaging of DNA and histones into nucleosomes restricting access to DNA and regulatory regions that also contributes to the regulation process. A different form of chromatin is euchromatin which is a more open and accessible form of chromatin usually associated with active genes. Epigenetic modifications such DNA methylation and histone modifications regulate the chromatin state by altering the combination of proteins binding to DNA. These chromatin factors regulate the chromatin structure by altering histone modifications by adding single to multiple methyl (methylation) groups to lysine and arginine (H3K27 and H3K9ac) on histones. These histone proteins tend to have positive charge and DNA have negative charge and therefore electrostatic forces of attraction tends to wrap DNA tightly around the histone core in the nucleosome forming heterochromatin. On the other hand, histone modifications like acetylation of lysine by histone acetyltransferases (HATs) removes the positive charges on histones resulting in a more open chromatin formation. DNA can now be freely accessed by transcription factors and initiation of DNA transcription can occur. This is reversed by histone deacetylases (HDACs) which restores the electrostatic force and tightening the chromatin structure. Combinations of chromatin proteins define five principal chromatin types as are shown in Figure 2. More detailed descriptions of chromatin types exists, but in this project we will limit us to the five main types as described by colour in Figure 2 [26].

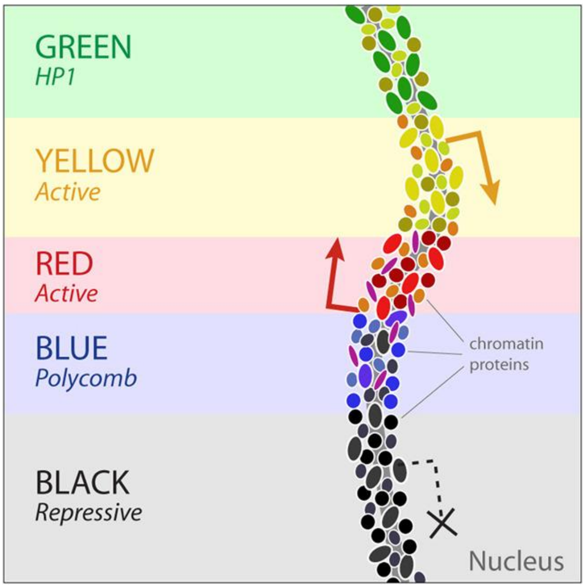


Figure 2 Systematic protein location mapping reveals five principal chromatin types. Red and yellow chromatin mark different types of genes and would in our case represent the euchromatin states. Black chromatin marks a distinct type of repressive chromatin where no or little transcriptional activity is present. Blue and green chromatin correspond to known types of heterochromatin. Different chromatin types include distinct compositions of chromatin proteins defining the different types. Reprinted with permission from Elsevier [26], copyright (2010).

A well-established method of detecting open and closed states of chromatin are by defining DHSs. These sites are specific regions of the genome where chromatin has lost its condensed structure, exposing the DNA and making it accessible for TF binding [27]. Because of the open chromatin state of the DNA it will be sensitive to degradation by the DNase I enzyme. Since this remodelled state is necessary for the binding of proteins such as transcription factors DNase I hypersensitivity sites (DHSs) define in a large way active and inactive DNA in regards to gene transcription [28].

1.2 Experimental approaches

1.2.1 Identification of TF binding sites by ChIP-seq

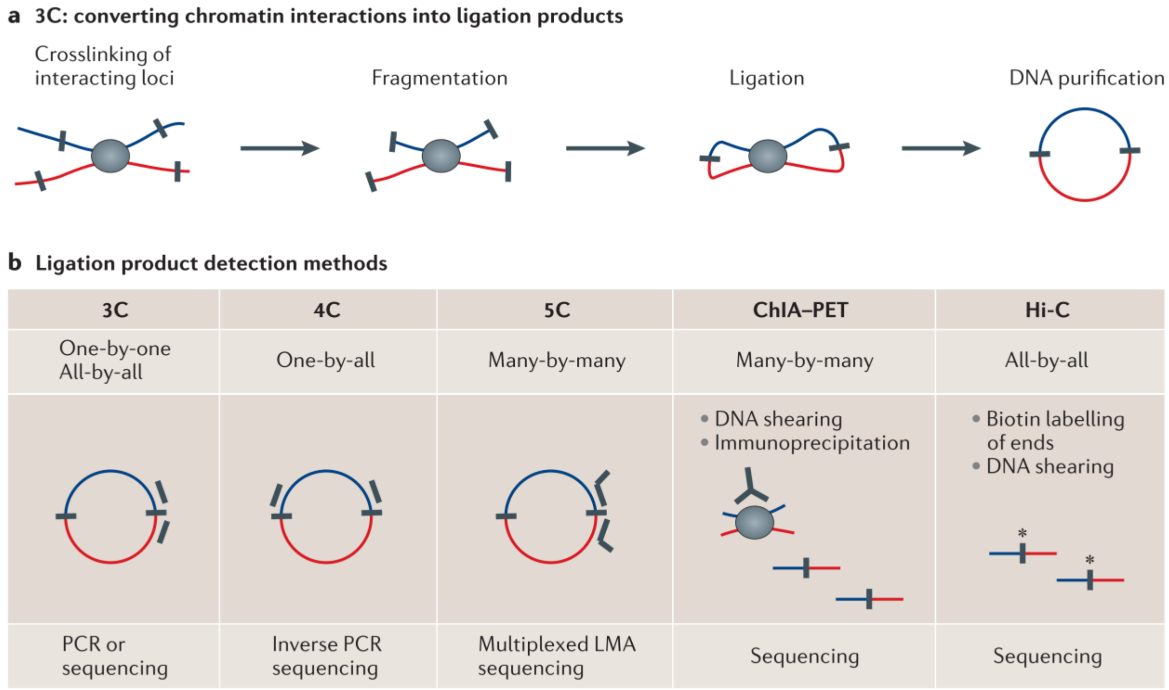
The important role that TFs play in gene regulation dictates the necessity to recognize the DNA regulatory regions to which they bind. Both experimental and in silico methods are used to achieve this. These methods tend to be complementary to each other where one is used to augment and help refine the results of the other. Several methods exist to determine the different binding locations of proteins and transcription factors in the laboratory and one of them are Chromatin immunoprecipitation sequencing (ChIP-seq).

Chromatin immunoprecipitation (ChIP) is a type of immunoprecipitation experimental technique that allows for large scale analysis of protein interaction with DNA and thus provides a good means of identifying transcription factor binding sites. This is achieved by TFs binding to DNA by either formaldehyde or other crosslinking compounds. Then the crosslinked DNA are fragmented into ~500 bp long segments by sonication or nuclease digestion. Crosslinked DNA fragments associated with the protein of interest are collected from the cell debris using an appropriate protein-specific antibody. Associated DNA fragments are purified and sequenced before they are mapped to a reference genome. The amount of reads associated with each DNA segment represents the number of interactions between the DNA and that specific protein, in the cell sample, and are usually referred to as ChIP-seq peaks. The information obtained from sequencing consists of tens of hundreds of millions of short DNA sequence fragments (reads) of the 5‘-ends of both the forward and reverse strands of DNA [29, 30]. Comparison of reads with a reference genome yields regions of overlap. 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.

1.2.2 Chromosome conformational capture

In 2002 Job Dekker published an article describing a method that would allow him to detect in vivo 3D DNA structure. He called this method chromosome conformation capture or 3C.

The 3C method aimed at identifying, locating and mapping physical interactions between genetic elements located throughout the human genome. In the years after Dekker published his article, new techniques were developed as described in Figure 3. The common goal of all the different 3C technologies was the same. Create a library of interacting segments in chromosomal DNA and locate the different interactions in the genome [31, 32].


a 3C: converting chromatin interactions into ligation products
Crosslinking of interacting loci
Fragmentation
Ligation
DNA purification
b Ligation product detection methods
*
*

3C	4C	5C	ChIA–PET	Hi-C
One-by-one All-by-all	One-by-all	Many-by-many	Many-by-many	All-by-all
			• DNA shearing • Immunoprecipitation	• Biotin labelling of ends • DNA shearing
PCR or sequencing	Inverse PCR sequencing	Multiplexed LMA sequencing	Sequencing	Sequencing

Figure 3 a) Showing the main initial steps of 3C technology being crosslinking of interacting loci, fragmentation, ligation, and DNA purification that are similar in 3C, 4C, 5C, ChIA-PET and Hi-C. b) Shows the different chromosome conformation capture assays and the main differences between 3C, 4C, 5C, ChIA-PET and Hi-C. Figure 3 was acquired from Dekker et al, and modified to fit this article. Reprinted by permission from Nature publishing group [32], copyright (2013)

The initial stepwise methods for a 3C analysis are usually similar in 3C, 4C, 5C, ChIA-PET and Hi-C. The first step is isolating crosslinking DNA and proteins and this is usually done by DNA fragmentation. The DNA fragments are then ligated, purified and sequenced. The number of sequence reads for each ligated product can then be correlated to the number DNA-protein crosslinking occurrences. The sequencing library generated now consists of ligated intersecting DNA segments that can be mapped to a reference genome to reveal interacting DNA segments. ChIA-PET data are acquired by similar methods where antibodies are created that have high affinity to DNA binding proteins and consequently immunoprecipitation. The DNA bound by the different proteins can be sequenced and mapped to a reference genome where the result is a library containing protein specific genome interactions [33]. This library can then be used to explore the genome organization at a few hundred kilo base pair resolution. Because of the large amount of data produced by the 3C methods it is best analysed using different computational methods.

1.3 Computational approaches

1.3.1 Computational analysis of TF binding sites

In silico methods for identification of TF binding sites are based on searching for patterns that are overly presented in a set of related sequences, for example a ChIP-seq dataset, as opposed to unrelated sequences. These sequences make up related and co- regulated genomic regions that are believed to partake in the same process. These patterns describe the TF of interest and its specificity and can be represented in different ways namely: a consensus sequence, a position weight matrix (PWM) or a position specific scoring matrix (PSSM). This pattern can be used to detect similar binding sites in other sequences that are believed to be active in the same process. Patterns are then stored in large databases such as TRANSFAC_public and JASPAR_core for use in motif identification algorithms or by researchers [34, 35].

By aligning DNA sequences from suspected binding sites it is possible to make a consensus sequence based on each base position and its conservation across all the sequences. This consensus sequence is derived in the belief that the sequences used are co- regulated as well as actual protein DNA interactions. This is verified by statistical overrepresentation of the pattern based on background frequencies. The consensus sequence matches closely all positions in the sequence, but as the consensus sequence is based on a consensus nucleotide all the different nucleotides in that position might not be exact. The consensus sequence can have variations dependent on the type of mismatches allowed or even the positions within which these variations are specified to be allowed in the representation.

An alternative to consensus sequences, when representing TF binding sites, are position weight matrices (PWM) and position specific scoring matrices (PSSM). These methods provide occurrence probabilities for each nucleotide in each position based on occurrence frequency in the pattern. This usually yields more information about the pattern than a consensus sequence as it 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 much like a consensus pattern by relating a set of sequences. PWMs are made by collecting a set of sequences and making a frequency table with each element of the table representing the frequency of each nucleotide at any given position in the alignment.

The base count for each position is called weights meaning that the higher the count the more likely it is for that nucleotide to occur in that specific position. A different representation for measuring weight is by log likelihoods ratio where the relative frequency of each base in the sequence collection is taken into account [36].

A PWM can be used in calculating the likelihood of one suspected TF binding region to be an actual binding region. The score of a probable TF is then the sum of the matrix values for each nucleotide at each position in the sequence. Higher scores will indicate that the TF is likely to bind to the sequence region and lower scores will indicate that it is less likely that this is an actual binding site.

Computational analysis methods use one of two methods to detect TF binding sites. One is to seek for patterns without a priori information about the binding patterns often referred to as motif discovery. The second is using known patterns to discover TFs usually referred to as motif scanning/ mapping methods. The motif scanning methods use the pattern representations described above in the process and this is usually less demanding computer memory wise.

1.3.2 Use additional information to guide binding site prediction and motif discovery

Motif scanning and discovery are not as straight forward as it might seem. The difficulty lies in the fact that motifs tend to be short sequences about four to ten bp (base pairs) and the DNA library only consists of four nucleotides (ACGT). Because of this the motif will occur multiple times throughout the genome by chance resulting in many false positives. Advances in technology in the past few years have made technology available that makes it possible to distinguish real motifs and identify features that distinguish them from the background. The ENCODE project has revealed many features of the genome which helps in understanding the gene regulatory process [37]. This information is used to limit the number of false motifs by using information such as DHSs, histone marks as well as sequence conservation data implemented in motif discovery/scanning algorithms. This will make in silico analysis better able to discriminate real binding regions from noise [38].

1.4 Workbenches for analysing biological data

As sequence information and annotations are continually updated to the human genome project new versions are released sequentially [39]. This has resulted in the release of many different versions of the human genome project over the years. This presents problems when doing bioinformatics as datasets will be made using different genome builds. However, a tool exist that is called *USCS genome lift over* and enables the user to take a dataset made using one version of the human genome build and converting it.

Analysing biological data often requires multiple resources from multiple databases that contain sequence annotations and collections of transcription factor binding profiles. This data is often used in algorithms that enable the processing of raw signals from sequencing experiments. Over the last decade many resources have been made available to help solve the questions that arise from biological data. One challenge is the use of different data formats by different groups, which makes it difficult to incorporate in to computer algorithms without first formatting the data. The answer to these challenges has been in large the development of workbenches that create a framework which enables the accessibility and localization of data in one place. This makes it possible to use multiple applications to analyse one dataset and its features. These workbenches are often available in a cloud based service or can be accessed from the web interface. Examples of workbenches are *MotifLab2* developed at the Norwegian university of Science and Technology (NTNU), *HiBrowse* that is a galaxy based tool made at the University of Oslo (UiO) and *Bedtools* which is freely available online [38, 40, 41].

1.4.1 *MotifLab2* version 2.0.-2

MotifLab2 is a workbench that focuses on the integration of tools and data for the analysis of regulatory regions. *MotifLab2* can be used to detect different binding sites for TFs and uses motif scanning and discovery tools as well as integrating different applications and databases. All the *MotifLab2* processes can be run from a graphical user interface that makes it easy to use (Figure 4). *MotifLab2* integrates multiple motif discovery tools including *AlignAce, BioProspector, ChipMunk, MEME, MotifSampler, Priority* and *Weeder*. This coupled with different types of data such as phylogenetic conservation, epigenetic marks, DHSs, ChIP-Seq data, positional binding preferences of transcription factors, TF-TF interactions, TF-expression and target gene expression makes this a powerful tool in motif site prediction. *MotifLab2* is the perfect tool for analysing ChIP-seq datasets for known TF binding sites. Another feature in *MotifLab2* is the protocol function that lets the user define a list of operations to be executed in order. Protocols can be used to document the steps you perform during an analysis session, and they can describe workflows that can be automatically executed in *MotifLab2*. If you like, you can specify exactly which sequences to perform the analyses on in the protocol itself, and the protocol will then always perform the analysis on these sequences.

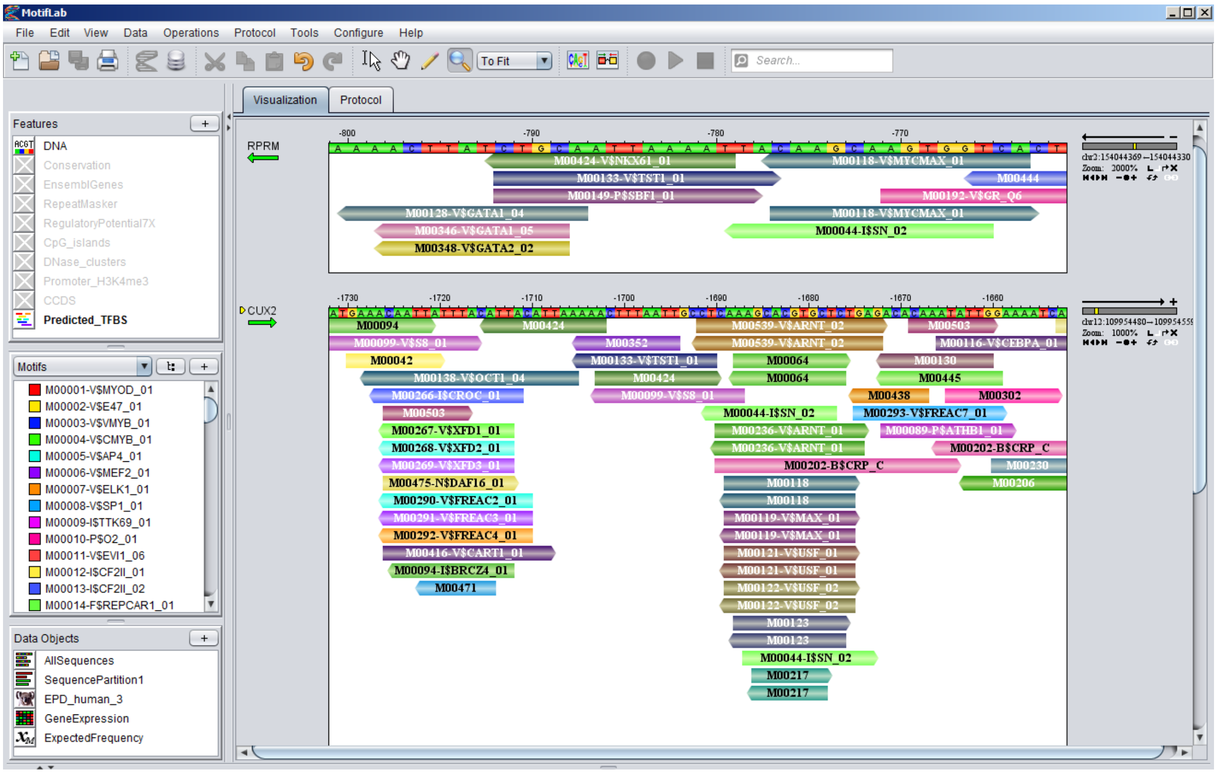
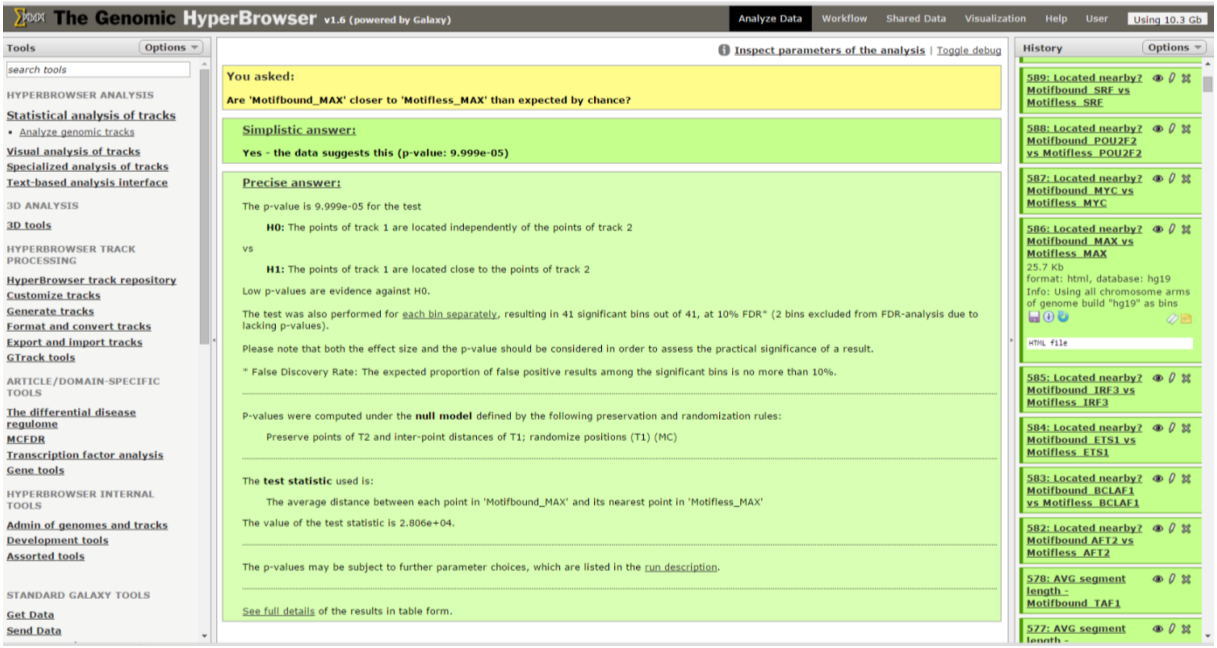


Figure 4 Example screenshot from *MotifLab2* showing how predicted motifs on two different DNA segments are visualized, their strand orientation as well as each motifs bp position and direction in the genetic segment in question [38].

Motiflab2 comes in different ram versions ranging between 256 MB to 4 GB and is optimized for UNIX systems. Another version of *Motiflab2* exists that runs on a minimal GUI platform that enables the user to run more demanding computer tasks regarding RAM usage. However, to use the *Minimal GUI platform* the user need to provide a protocol as mentioned earlier.

1.4.2 *HiBrowse* version 1.6

Previously described 3C technology coupled with next generation sequencing has allowed for characterization of genome- wide chromatin 3D structure. To better understand gene regulation and how genome 3D structure can affect this, methods for analysing such data is needed. One method is *HiBrowse* that uses hypothesis-testing and realistic assumptions in null models. *HiBrowse* uses tracks to refer to a series of data units positioned on a line-based coordinate system. Another similar tool is called the *Genomic Hyper Browser* and it is built on the same platform as *HiBrowse* and the two will probably be merged in the future [42]. Some of the tools that are used in this project was originally published as tools in *the Genomic Hyperbrowser*, but has also been implemented in the newer *HiBrowse* version. In this project, only the *HiBrowse* version has been used.


The Genomic HyperBrowser v1.6 (powered by Galaxy)
Analyze Data
Workflow
Shared Data
Visualization
Help
User
Using 10.3 Gb
Tools
Options
search tools
HYPERBROWSER ANALYSIS
Statistical analysis of tracks
Analyze genomic tracks
Visual analysis of tracks
Specialized analysis of tracks
Text-based analysis interface
3D ANALYSIS
3D tools
HYPERBROWSER TRACK PROCESSING
HyperBrowser track repository
Customize tracks
Generate tracks
Format and convert tracks
Export and import tracks
GTrack tools
ARTICLE/DOMAIN-SPECIFIC TOOLS
The differential disease regulome
MCFDR
Transcription factor analysis
Gene tools
HYPERBROWSER INTERNAL TOOLS
Admin of genomes and tracks
Development tools
Assorted tools
STANDARD GALAXY TOOLS
Get Data
Send Data
Inspect parameters of the analysis | Toggle debug
History
Options
589: Located nearby? Motifbound_SRF vs Motifless_SRF
588: Located nearby? Motifbound_POU2F2 vs Motifless_POU2F2
587: Located nearby? Motifbound_MYC vs Motifless_MYC
586: Located nearby? Motifbound_MAX vs Motifless_MAX
25.7 Kb
format: html, database: hg19
Info: Using all chromosome arms of genome build "hg19" as bins
HTML file
585: Located nearby? Motifbound_IRF3 vs Motifless_IRF3
584: Located nearby? Motifbound_ETS1 vs Motifless_ETS1
583: Located nearby? Motifbound_BCLAF1 vs Motifless_BCLAF1
582: Located nearby? Motifbound AFT2 vs Motifless_AFT2
578: AVG segment length - Motifbound_TAF1
577: AVG segment length -

You asked:

Are 'Motifbound_MAX' closer to 'Motifless_MAX' than expected by chance?

Simplistic answer:

Yes - the data suggests this (p-value: 9.999e-05)

Precise answer:

The p-value is 9.999e-05 for the test

H0: The points of track 1 are located independently of the points of track 2

vs

H1: The points of track 1 are located close to the points of track 2

Low p-values are evidence against H0.

The test was also performed for each bin separately, resulting in 41 significant bins out of 41, at 10% FDR* (2 bins excluded from FDR-analysis due to lacking p-values).

Please note that both the effect size and the p-value should be considered in order to assess the practical significance of a result.

* False Discovery Rate: The expected proportion of false positive results among the significant bins is no more than 10%.

P-values were computed under the **null model** defined by the following preservation and randomization rules:

Preserve points of T2 and inter-point distances of T1; randomize positions (T1) (MC)

The **test statistic** used is:

The average distance between each point in 'Motifbound_MAX' and its nearest point in 'Motifless_MAX'

The value of the test statistic is 2.806e+04.

The p-values may be subject to further parameter choices, which are listed in the run description.

See full details of the results in table form.

Figure 5 Example screenshot from *HiBrowse* showing the output of a statistical analysis measuring closeness using the *located nearby* tool for ML and MB ChIP-seq datasets for the MAX TF, including the estimated p- value and a simplistic answer, for the hypothesis tested [40].

The genomic track represents genome features as track elements or set of track elements comprising a biological feature for example CTCF binding sites from a ChIP-seq experiment [43]. Next generation sequencing and ChIP-seq experiments creates high-resolution data along the genome. This large amount of data can be interpreted by tools like HiBrowse and to predict the significance of the outputs produced *HiBrowse* uses a Monte Carlo False Discovery Rate (MCFDR). This is done by randomizing the assignment of regions multiple times and then measure if the initial score is significant or not based on the score distribution. This MCFDR algorithm iterates between adding MC samples across tests and calculating intermediate FDR values for the collection of tests. Monte Carlo sampling in *HiBrowse* is stopped either by the number of Monte Carlo values or by the FDR threshold. The output of a statistical test in *HiBrowse* is often a simplistic answer including a p- value for the statement given. This p-value is the highest possible detected p-value based on the MCFDR resampling depth [44].

HiBrowse also offers a 3D colocalization tool that based on a reference model predicts the spatial closeness between segments of two different datasets. The user can ask whether all the genomic elements in the BED-file are more/less co-localized in 3D, in an all-versus-all fashion, than what would be expected by chance. In this case, the mean of the observed standardized interaction frequencies is compared to the expected value estimated from the permuted positions in representative regions in the rest of the Hi-C (3D) track [40].

1.4.3 *Bedtools* version 2.25.0.

Bedtools is a UNIX based command line tool that enables the user to a wide range of different genomic analyses. In this project the tools *Intersect* and *Closest* was used. The *intersect* tool lets the user find overlapping intervals using many different parameters. The tool *Closest* lets the user find the closest intervals among multiple datasets. In *Bedtools* the user specifies –a files that works as a query and potentially multiple –b files that work as databases for the query.

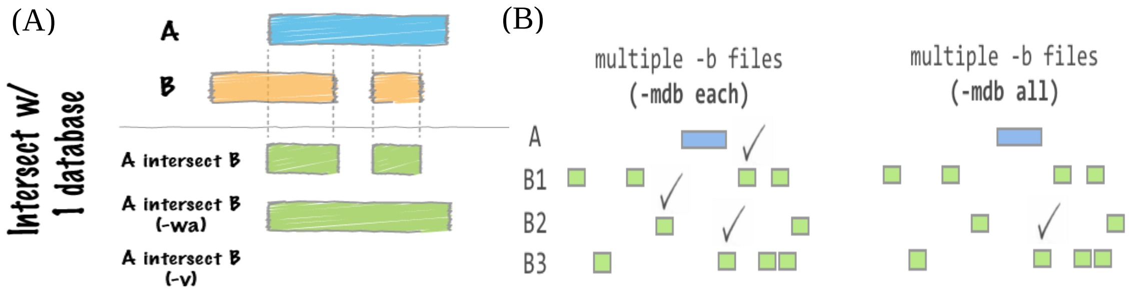


Figure 6 (A) shows a visual presentation of the *intersect* tool from *Bedtools* and how it estimates overlapping segments between A and B files. The options –wa can be used if the user wants to write the original entry in A for each overlap and –v can be used if the user wants to only report those entries in A that have no overlap in B. (B) shows a visual representation of how the *closest* tool measures intersegment distance for multiple –b files and the different -mdb settings can be used to specify how multiple databases should be resolved. -each reports the closest records for each database and - all reports the closest records among all databases [41].

Different tools in *Bedtools* version 2.25.0 can be assigned different options to affect the analysis and how the output is formatted as described in Figure 6.

1.5 Gene regulation by the vitamin D receptor - an interesting biological system

When UV-B radiation hits the skin a two-step transformation reaction occurs that converts 7-dehydo-cholesterol ultimately into the active form of vitamin D or calcidol (1α,25$(OH)_2$D3). In the first step 7-dehydrocholesterol is reduced by ultraviolet light in a ring-opening reaction where the product is previtamin D3 (Pre-D3). Second, previtamin D3 spontaneously isomerizes to cholecalciferol or vitamin D3 which are the same form that is obtained through nutrients. Cholecalciferol then travels through the bloodstream all the way to the liver where it is converted to calcidol (25-hydroxyvitamin-D3) and later in the kidneys to the active form, Calcitriol or 1α,25$(OH)_2$D3 [45].

Vitamin D receptor (VDR) is a transcription factor belonging to the family of nuclear receptors and are composed of 5 functional domains. The localization domain guides transportation of VDR to the nucleus after translation, the DNA binding domain recognize and bind to the response elements, the dimerization domain and the ligand binding domain where calcidol binds. When this happens the VDR changes its conformation and forms a heterodimer with retinoid X receptor (RXR) through its dimerization domain.

The VDR can bind DNA as monomers or even dimers, but these interactions are not stable. Activation by calcidol and heterodimerization with RXR however stabilize this interaction. VDR binds to DNA as described earlier by recognizing the vitamin D response element (VDRE), a heptad repeat sequence with a spacer element between two half sites. The spacer region usually consists of three nucleotides, but this varies to some degree. This is known as the DR3-type VDRE where D3 indicates that the element is a direct repeat and the value 3 indicates a spacing between the repeats.

After DNA binding by the VDR-RXR heterodimer a wide variety of other proteins are recruited to the complex as well as initiating the translation machinery. One of the activated genes are 24-hydroxylase (CYP24A1) which are responsible for catalysing the degradation of 1,25-dihydroxy Vitamin D3 and its precursor. To achieve transactivation, the VDR heterodimer acts as an initiator 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 transactivation 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. Negative regulation by VDR involves interactions with VDR interacting repressor and recruitment of histone deacetylases (HDACs) [45]. Genes in the vitamin D signalling system, such as those coding for vitamin D receptor (VDR) and the enzymes 25-hydroxylase (CYP2R1), 1α-hydroxylase (CYP27B1), and 24-hydroxylase (CYP24A1) have large CpG islands in their promoter regions and therefore can be silenced by DNA methylation. Additionally, VDR protein physically interacts with coactivator and corepressor proteins, which in turn are in contact with chromatin modifiers, such as HATs, HDACs, HMTs, and also with chromatin remodelers. Further, a number of genes encoding for chromatin modifiers and remodelers, are primary targets of VDR and its ligands [46].

1.6 Aims of the study

In a ChIP-seq experiment, it would in principle be expected to find a known TF binding site in all resulting ChIP-seq peaks. In practice some sequences will lack this TF binding site due to experimental noise. However, it is expected to appear in most of the peaks from the ChIP-seq analysis. Interestingly, in VDR ChIP-seq experiments, as much as 50% of the identified peaks seem to lack known TF binding sites. In other words, they are motifless. These ML peaks seem to be real binding sites even though they have no clear explanation for their binding capacity. One possible explanation is 3D interactions, including DNA looping where two DNA segments are in closer physical proximity to each other than to intervening sequences. This may lead to a TF being crosslinked not only to the genomic region it actually is recognizing, but also to additional regions that are close to the TF, but without actually binding to it through a TFBS. This may then lead to false positive or “ML” binding sites. This hypothesis can be investigated by testing statistically for association between false binding sites (without motif) and short physical 3D distance to real binding sites (with motif). A figure representing the workflow in this project can be found in Figure 7.

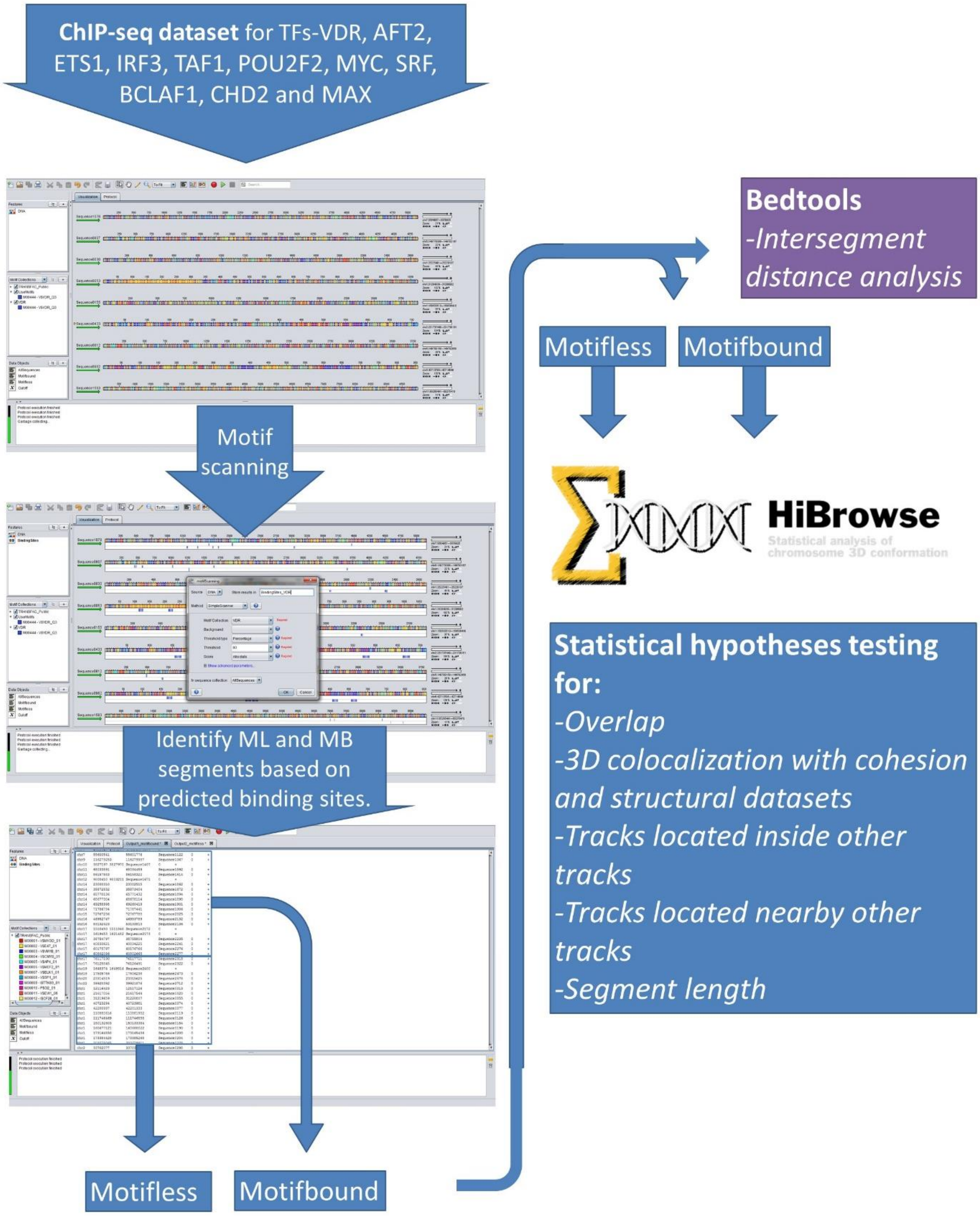

ChIP-seq dataset for TFs-VDR, AFT2, ETS1, IRF3, TAF1, POU2F2, MYC, SRF, BCLAF1, CHD2 and MAX
Motif scanning
Identify ML and MB segments based on predicted binding sites.
Motifless
Motifbound
Bedtools
-Intersegment distance analysis
Motifless
Motifbound
HiBrowse
Statistical analysis of chromosome 3D conformation
Statistical hypotheses testing for:
-Overlap
-3D colocalization with cohesion and structural datasets
-Tracks located inside other tracks
-Tracks located nearby other tracks
-Segment length

Figure 7 TF ChIP-seq datasets were uploaded to *MotifLab2* and appropriate motifs were selected according to the TF in question, described in Table 9. Then they were scanned for known binding sites using simple scanner at indicated thresholds Table 11. Sequences containing one or more motifs were determined as MB and the rest as ML. Sequences were then analysed using different tools from *Bedtools* version 2.25.0. Then datasets were uploaded to *HiBrowse* where different hypothesis and descriptive analyses were completed to describe the different properties of ML and MB segments.

2 Materials and methods

2.1 VDR

Table 1 VDR ChIP-seq data from Ramagopalan used in this project and relevant sources.

Data	Type	PMID	Lab	Reference	Url	Dataset
VDR ChIP-seq*	ChIP-seq peaks	20736230	Sreeram V. Ramagopalan,	[47]	http://genome.cshlp.org/content/20/10/1352.full	http://genome.cshlp.org/content/suppl/2010/08/24/gr.107920.110.DC1/Supplementary_Table_1.xls

*Using the *Liftover* tool from UCSC 0.95 was set as the minimum ration of bases that must remap.

2.2 Cohesin

ChIP-seq data given in Table 2 was analysed by Encode consortiums participating labs using the ENCODE and modENCODE Guidelines For Experiments Generating ChIP, DNase, FAIRE, and DNA Methylation Genome Wide Location Data protocol, Version 2.0, July 20, 2011.

Table 2 Optimal IDR threshold ChIP-seq data for the cohesin subunits and CTCF from ENCODE in cell type GM12878 as well as relevant sources.

Data	Type	Accession	Lab	Reference	Url	Dataset
RAD21	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000BMY	Richard Myers, HAIB	[48]	https://www.encodeproject.org/experiments/ENCSR000BMY/	https://www.encodeproject.org/files/ENCFF002CHR/@@download/ENCFF002CHR.bed.gz
SMC3	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000DZP	Michael Snyder, Stanford	[48]	https://www.encodeproject.org/experiments/ENCSR000DZP/	https://www.encodeproject.org/files/ENCFF002CPN/@@download/ENCFF002CPN.bed.gz
CTCF	optimal idr thresh	ENCSR000DZN	Michael Snyder,	[48]	https://www.encodeproject.org/experim	https://www.encodeproject.org/files/ENCFF002COQ/@@download/ENCFF002COQ.bed.gz

	olded peaks		Stanford		ents/ENCSR000DZN/	

2.3 ChIA-PET

ChIA-PET dataset made using RAD21 as target given in Table 3 were produced using the ENCODE and modENCODE Guidelines For Experiments Generating Data using RNA-Binding Proteins (RBPs): ENCODE and modENCODE Standards for RIP-Chip and RIP-Seq Experiments Version 2.0, 9 January 2012.

Table 3 ChIA-PET data from ENCODE for cell type GM12878 used in this project and relevant sources.

Data	Type	Accession	Lab	Reference	Url	Dataset
ChIA-PET	peaks	ENCSR752QCX	Michael Snyder, Stanford	[48]	https://www.encodeproject.org/experiments/ENCSR752QCX/	https://www.encodeproject.org/files/ENCFF002EMR/@@download/ENCFF002EMR.bed.gz

2.4 HOT and LOT

Table 4 HOT and LOT datasets from ENCODE for cell type GM12878 used in this project and relevant sources.

Data	PMID	Lab	Reference	Url	Datasets
HOT Whole genome HOT intergenic LOT whole genome LOT intergenic	22950945	Yale	[21]	http://encodenets.gersteinlab.org/metatracks/HOT_Gm12878_merged.bed.gz	http://encodenets.gersteinlab.org/metatracks/

2.5 DNase

Table 5 DHSs dataset from ENCODE for cell type GM12878 used in this project and relevant sources.

Data	Type	PMID/ Accession	Lab	Reference	Url	Dataset
DNase	DNase-seq	22955617/ENCSR000EMT	John Stamatoyannopoulos, UW	[27]	https://www.encodeproject.org/experiments/ENCSR000EMT/	https://www.encodeproject.org/files/ENCFF001WFU/@@download/ENCFF001WFU.bed.gz

2.6 TAD

Table 6 TAD dataset from Ji et al 2016 for embryonic stem cells (hESCs) used in this project and relevant sources.

Data	Type	PMID	Reference	Url	Dataset
TADs	Mango-Called High-Confidence SMC1 ChIA-PET	26686465	[16]	http://www.cell.com/action/showImagesData?pii=S1934-5909%2815%2900505-6	http://www.cell.com/cms/attachment/2045959822/2057172164/mmc7.xlsx

2.7 Other transcription factors

ChIP-seq data in Table 7 was analysed by Encode consortiums participating labs using the ENCODE and modENCODE Guidelines For Experiments Generating ChIP, DNase, FAIRE, and DNA Methylation Genome Wide Location Data protocol, Version 2.0, July 20, 2011.

Table 7 Optimal IDR threshold GM12878 ChIP-seq data for AFT2, ETS1, IRF3, TAF1, POU2F2, MYC, SRF, BCLAF1, CHD2 and MAX from ENCODE used in this project and relevant sources.

Data	Type	Accession	Lab	Ref	Url	Dataset
AFT2	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000BQK	Richard Myers, HAIB	[48]	https://www.encodeproject.org/experiments/ENCSR000BQK/	https://www.encodeproject.org/files/ENCFF002CGO/@@download/ENCFF002CGO.bed.gz
ETS1	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000BKA	Richard Myers, HAIB	[48]	https://www.encodeproject.org/experiments/ENCSR000BKA/	https://www.encodeproject.org/files/ENCFF002CGY/@@download/ENCFF002CGY.bed.gz
IRF3	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR408JQO	Michael Snyder, Stanford	[48]	https://www.encodeproject.org/experiments/ENCSR408JQO/	https://www.encodeproject.org/files/ENCFF103BPB/@@download/ENCFF103BPB.bed.gz

TAF1	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000BGS	Richard Myers, HAIB	[48]	https://www.encodeproject.org/experiments/ENCSR000BGS/	https://www.encodeproject.org/files/ENCFF002CHY/@@download/ENCFF002CHY.bed.gz
POU2F2	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000BGP	Richard Myers, HAIB	[48]	https://www.encodeproject.org/experiments/ENCSR000BGP/	https://www.encodeproject.org/files/ENCFF002CHP/@@download/ENCFF002CHP.bed.gz
MYC	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000DKU	Vishwanath Iyer, UTA	[48]	https://www.encodeproject.org/experiments/ENCSR000DKU/	https://www.encodeproject.org/files/ENCFF002DAI/@@download/ENCFF002DAI.bed.gz
SRF	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000BGE	Richard Myers, HAIB	[48]	https://www.encodeproject.org/experiments/ENCSR000BGE/	https://www.encodeproject.org/files/ENCFF002CHW/@@download/ENCFF002CHW.bed.gz
BCLAF1	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000BJZ	Richard Myers, HAIB	[48]	https://www.encodeproject.org/experiments/ENCSR000BJZ/	https://www.encodeproject.org/files/ENCFF002CGT/@@download/ENCFF002CGT.bed.gz
CHD2	ChIP-seq, Encode optimal idr	ENCSR000DZR	Michael Snyder, Stanford	[48]	https://www.encodeproject.org/expe	https://www.encodeproject.org/files/ENCFF002C

	thresholded peaks				riments/ENCSR000DZR/	OO/@@download/ENCFF002COO.bed.gz
MAX	ChIP-seq, Encode optimal idr thresholded peaks	ENCSR000DZF	Michael Snyder, Stanford	[48]	https://www.encodeproject.org/experiments/ENCSR000DZF/	https://www.encodeproject.org/files/ENCFF002COW/@@download/ENCFF002COW.bed.gz

2.8 Workbenches

Motiflab2 version 2.0.-2. Was used to find ML peaks for VDR, AFT2, BCLAF1, ETS1, IRF3, MAX, MYC, POU2F2, SRF, TAF1, CHD2. Then *HiBrowse* was used to do different statistical tests on the mentioned TFs. Lastly, *Bedtools* version 2.25.0 was used to find the distance between closest pairs of MB and ML. Additionally, *Excel* was used to do computations on the *Bedtools* outputs.

Table 8 Different workbenches and associated tools used in this project.

Application	**Version**	**Tools**	**References**
MotifLab2 *	2.0.-2.	Simple Scanner	[38]
	Simple Scanner is a motif scanning program based on simple PWM matching using log-odds ratios and a zero-order background model and was used to determine ML sequences.		
HiBrowse	1.6	**Descriptive statistics:** Avg. segment length, Avg. segment distance **Hypothesis testing:** Overlap, Located inside, Located nearby (S-S) **3D analysis:** Colocalization between two point tracks tool	[40]
	Avg. segment distance measures the average distance between elements in track. **Avg. segment length** measures the average length of segments in track **Segment distance** measures the distribution of distances from each element in track 1 to the nearest element in track 2 **Overlap (S-S),** Are track 1 overlapping track 2 more than expected by chance? **Located inside (P-P),** Are track 1 falling inside track 2 more than expected by chance? **Located nearby (S-S),** Are track 1 closer to track 2 more than is expected by chance?		

	Colocalization between two point tracks tool measures whether all the genomic elements in the BED-file are more/less co-localized in 3D, in an all-versus-all fashion, than what would be expected by chance.		
Bedtools	2.25.0.	Closest, intersect	[41]
	Closest will return the nearest segment from A to the closest feature in all B files where the least genomic distance from the start or end of A to feature in B is returned for all segments in A. **Intersect** will return overlapping segments between multiple files.		

* A Unix based system was used to be able to run *MotifLab2* in 4GB ram.

2.9 Matrices used in motifs scanning

Table 9 shows the different motifs used in ML detection in all TF datasets.

Table 9 PWMs\ motifs from JASPAR_core, TRANSFAC_public and Wang et al. 2012 used by *Simple scanner* in *MotifLab2 [34, 35, 38, 49]*.

TF	Motifs used
VDR	M00444
AFT2	M00040, M00179, M00041, MA0270, MA0269
BCLAF1	M00341, MA0062, MA0081, MA0098
ETS1	M00032, M00074, M00339, MA0098
IRF3	M00118, M00119, M00123, M00322, M00615, MA0055, MA0058, MA0059, MA0091, MA0093, MA0104, MA0147, PB0043, PB0147,MA0050, MA0051, MA0158, M00453, M00063, M00062
MAX	M00118, M00119, M00123, M00322, M00615, MA0055, MA0058, MA0059, MA0091, MA0093, MA0104, MA0147, PB0043, PB0147
MYC	MA0104, MA0147, M00322, M00118, M00491, M00055, MA0059, MA0058, MA0055, MA0093, MA0091, M00615, MA0438, M00123, MM0001*
POU2F2	M00210, MA0142, MA0197, MM0002*
SRF	M00152, M00186, M00215, MA0083, MM0003*
TAF1	M00369
CHD2	MA0088, MM0004*

*Matrices not found in JASPAR_core or TRANSFAC_public, were found in Wang et al 2012 and uploaded to *MotifLab2* see attachments (Table 39, Table 40, Table 41, Table 42)

3 Results

3.1 Summary of the work completed in this project

In this brief summary, we describe the different procedures and why they were completed. For a more detailed description, see the different sections 3.2-3.10.

Before motif scanning could be done, it was important to have all datasets in the same genome build. All but one datasets were already in the hg19 version and the VDR dataset was formatted to hg19 from hg 18 using USCS genome lift over. Motif scanning was then completed to identify ML and MB peaks for the following TFs VDR, AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1. They were identified as having potential for ML peaks, based on motif enrichment in *Factorbook* www.factorbook.org (F. Drabløs, personal communication). MB and ML datasets were then uploaded to *HiBrowse*. The *3D colocalization* tool requires datasets to be in a GT track format and datasets were formatted using the *create GT track* tool. After data formatting the physical 3D distance between MB and ML segments was measured. Doing this analysis would indicate if the ML and MB are located close in 3D space in the genome more than is expected by chance. Another way to test for 3D physical distance is to use ChIA-PET data. Finding overlaps between MB and ML with ChIA-PET could indicate that the overlapping segment is close to or part of a DNA crosslink. ML peaks could occur because of DNA looping, as hypothesized, and this was investigated with an overlap analysis using ChIP-seq data from proteins associated with the cohesin complex (RAD21, SMC3 and CTCF). Promotors receive a multitude of signals that work to activate or repress TFs and many of these promotor binding sites can be found inside HOT regions. MB segments contain known binding sites and were expected to be located inside HOT regions more often than ML segments. This was because ML was believed to overlap with enhancers because of DNA looping. If ML peaks are associated to enhancers and MB to promotors it would be expected to find them located inside DHSs and this was tested using a located inside analysis for ML and MB inside DHSs. Another way that DNA interacts through genome 3D structure is by TADs where one sub hypothesis is that MB segments binds to close ML segments in pairs within the same TAD. This could be investigated using *Bedtools closest* and *intersect* tools. Finding that ML and MB form inter TAD pairwise interactions could indicate TAD structure being important structural boundaries for ML interactions.

3.2 Motif scanning using *MotifLab2* and *USCS Genome Lift Over*

The complete VDR dataset of 2776 segments was successfully converted from hg18 to hg19 using UCSC annotations lift over tool. Sequences were sorted based on Hg18 start sites, where the start and end positions from different versions were plotted in Figure 8. This figure shows the start and end positions before and after genome lift over. This was completed to identify potentially large changes between different genome builds and in what way it could have affected the VDR dataset.

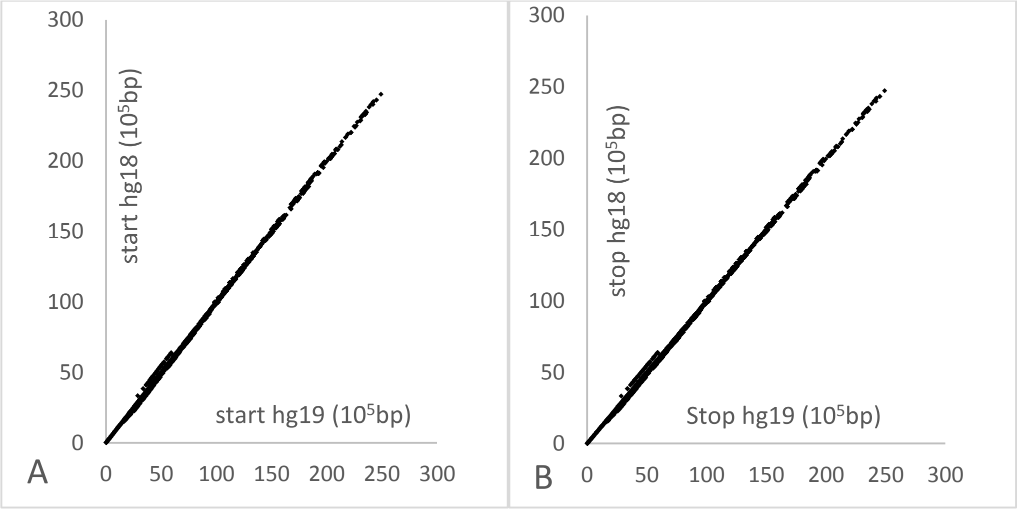


Figure 8 Dot plots comparing start (A) and stop (B) sites before and after *UCSC genome lift over* from hg18 to hg19 in the VDR dataset and units(X,Y) are displayed in milions(10^5) bp. In this figure the chromosome ID was ignored.

Datasets were uploaded to *MotifLab2* on a UNIX based system which was necessary to run the 4 GB Ram version of *MotifLab2*. The application was lunched using the web start application from http://tare.medisin.ntnu.no/motiflab/. In *MotifLab2* PWMs are called motifs, but the two are essentially the same. Motifs for the TFs examined were selected from TRANSFAC_public, JASPAR_core and Wang 2012 and used as motif collection in *MotifLab2* [34, 35, 49]. Motifs originating from TRANSFAC_public are denoted with their two first letters as M0, JASPAR_core as MA and the generated motifs made based on Wang 2012 are denoted as MM (Table 9). A tool called *Simple Scanner* was then used to scan all the TF datasets for known binding sites. The motif detection threshold was estimated by doing comparisons of the number of sequences predicted to be MB and ML for the different thresholds and the score was set to absolute. Additional parameters used for motif scanning can be found in Table 10.

Table 10 *Simple Scanner* and parameters used for motif scanning using different threshold for different TF datasets and PWMs from TRANSFAC_public, JASPAR_core and Wang 2012. Threshold was set to percentage similarity needed for sequence to match and the sequence scoring value was set to absolute meaning a log odds ratio. Different parameters were set for different datasets and these are presented in *italic*.

Motif scanning	
Parameter	Value
Source	DNA
Method	*Simple Scanner*
Motif collection	*See Table 9*
Threshold type	Percentage
Threshold	*82-95%*
Score	Absolute
In sequence collection	All sequences

Sequences that did not contain any known binding sites for the TF in question were identified as ML. The sequences that contained one or more binding sites were identified as MB. The motif scanning reviled 109 MB and 2666 ML VDR peaks at threshold 95%, 991 MB and 1784 ML at threshold 90% and 2309 MB and 466 ML peaks at threshold 85%. For the ten other TFs the threshold was chosen based on a random pre scanning using a subset of the sequences from the original dataset. The top 1000 sequences in an unsorted dataset was picked and scanned using different thresholds to find the optimal threshold for each dataset. The optimal threshold was estimated based on two arguments. The first was that the number of sequences predicted ML and MB should be close to even. The second was to keep the threshold high enough to reduce the amount of false positives. Thresholds, ML and MB predicted sequences for the different TFs are shown in Table 11. For some TFs, multiple PWMs were used in motif scanning, see Table 9. Some of the TFs mentioned have more than one motifs known to be related binding sites so adding these to the motif collection used in motif scanning would better predict true ML segments.

Table 11 Thresholds used to estimated ML and MB for VDR, AFT2, BCLAF1, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1 as well as the predicted number of MB and ML segments.

TF	Threshold %	ML	MB
VDR	90	991	1784
AFT2	90	13305	10001
BCLAF1	90	531	5575
CHD2	82	9230	6366
ETS1	95	857	3263
IRF3	90	956	2586
MAX	95	7458	5084
MYC	90	234	3456
POU2F2	90	14542	8293
SRF	90	6996	1546
TAF1	90	13093	1178

3.3 GT tracks and formats used by *HiBrowse*

After the ML and MB datasets had been predicted they were uploaded to *HiBrowse*. The *3D colocalization* tools in *HiBrowse* requires datasets to be formatted as GT tracks and this was done using the *Create GT track tool from unstructured tabular data* tool. To convert bed files into GT tracks in *HiBrowse* each tabular file was converted individually. All datasets used by the *3D colocalization* tool were formatted in the same way as described here. No lines were skipped and columns were selected individually. Each column was named as indicated in Table 12. No dense track type or indexing standard was used. The file was not auto corrected or cropped in any way. For all non 3D experiments the original BED files were used.

Table 12 Parameters used in the *Create GT track file from unstructured tabular data* tool for formatting data previous to *3D colocalization* analysis. The input data was all in tabular hg19 bed format and the three first columns of the dataset (Chromosome, sequence start and sequence end) were selected to be converted into a GT track.

Create GT track file from unstructured tabular data	
Select input source	Tabular file from history
Select tubular file	VDR ChIP-seq
Character to use to split lines into columns	Tab
Number of lines to skip (from front)	0
Column selection method	Select individual columns
Select the name for column #1	Seqid
Select the name for column #2	Start
Select the name for column #3	End
Select a specific genome build	Yes
genome build	Human Feb. 2009 (hg19/GRCh37)
Create dense track type (i.e. function, step function, or genome partition)	No
Indexing standard used for start and end coordinates	0-indexed, end exclusive
Auto-correct the sequence id (‘seqid’) column	No
Crop segments crossing sequence ends	No

All segments were converted successfully and no data was lost during formatting.

3.4 3D genome localization and the cohesin complex

The GT tracks generated in 3.3 was then used to measure 3D closeness between points on GT track 1 and their location on the segments of GT track 2. This analysis was completed using the 3D version of the Genomic Hyperbrowser called *HiBrowse*. Interactions were set to inter- and intra-chromosomal interactions and the cell line used was GM12878 with a resolution of 1 million base pairs. GT Track 1 was randomized and GT track 2 was preserved in the null model and the number of resampling’s was set to 1000 and tail was set to more than expected by chance. P-value was selected as statistic and the analysis was set to compare in bounding regions. The parameters for this analysis is presented in Table 13.

Table 13 Parameters used for estimating 3D colocalization between two GT tracks with a resolution of 1 million bp and a 1000 numbers of resampling's estimating if the middle points of track 1 localise in 3D genome space with the middle points of track 2 more than is expected by chance.

3D Colocalization	
Genome build	Human Feb. 2009 (hg19/GRCh37)
Interactions	Using inter- and intra-chromosomal interactions
Cell line	GM12878
Dataset and resolution	GM12878-HiC-HindIII-R1-1 million bp
Randomization options	Randomize Track 1(conserve consecutive distances), preserve track 2
Number of resampling's	1000
Tail	More
Statistic	p-value
Compare in	Bounding regions
null hypothesis	The points of track 1 are localized independently of the segments of track 2 in the 3D genome space
alternative hypothesis	The points of track 1 tend to Localize with the segments of track 2 in 3D genome space.

The *3D colocalization* analysis described above was initiated for VDRML-VDRML, VDRMB-VDRMB and VDRMB-VDRML GT tracks. Then the analysis was completed for the VDR GT tracks and RAD21, SMC3 and CTCF GT tracks (Table 14). CTCF, RAD21, and SMC3 GT tracks was then compared to AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1 ML and MB GT tracks (Table 30).

Table 14 *3D Colocalization* analysis between VDRML, VDRMB, CTCF, RAD21 and SMC3 GT tracks and the p-value estimated using *HiBrowse* for the alternative hypothesis presented in Table 13.

Analysis	Track 1	Track 2	P-value
3D colocalization between two point tracks	VDRMB	VDRMB	0.0009990
3D colocalization between two point tracks	VDRML	VDRML	0.0009990
3D colocalization between two point tracks	VDRMB	VDRML	0.0009990
3D colocalization between two point tracks	VDRML	CTCF	0.0009990
3D colocalization between two point tracks	VDRML	RAD21	0.0009990
3D colocalization between two point tracks	VDRML	SMC3	0.0009990
3D colocalization between two point tracks	VDRMB	RAD21	0.0009990
3D colocalization between two point tracks	VDRMB	SMC3	0.0009990

The colocalization analysis showed that the points of GT track 1 are expected by chance to more often than not be located nearby in 3D space to segments in GT track 2 for the GT tracks presented in Table 14. The VDRMB and VDRML GT track seems to be localized in 3D genome space with CTCF, RAD21 and SMC3 GT tracks more often than expected by chance. However, no difference in p- value was detected for *3D colocalization* between different analyses presented in Table 14.

3.5 Chia PET and MB/ML overlap

A ChIA-PET dataset from Heidari et al. was uploaded to *HiBrowse* and the tool overlap was used to detect overlapping genetic segments between TF and ChIA-PET datasets. The same parameters as described here was used for all the following overlap experiments described in this report (Table 15). The null model was set to Preserve segments (track 2), segments lengths and inter segment gaps (track 1) randomize positions (track 1). A Monte Carlo model was used. The sampling depth Monte Carlo false discovery rate was set to fixed 10 000 samples.

Table 15 Parameters for detecting overlap between two datasets with a Monte Carlo model using a Monte Carlo False Discovery Rate sampling depth of 10000 samples.

Overlap	
Genome Build	Human Feb. 2009 (hg19/GRCh37)
Null model	Preserve segments (track 2), segments lengths and inter segment gaps (track 1) randomize positions (track 1) (MC)
MCFDR Sampling depth	Fixed 10 000 samples
null hypothesis	The segments of track 1 are located independently of the segments of track 2 with respect to overlap
Alternative hypothesis	The segments of track 1 tend to overlap the segments of track 2.

The overlap analysis was completed for ChIA-PET, RAD21, SMC3, CTCF against the VDRML/MB datasets as well as AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1 ML and MB datasets. The analysis was also completed for heterochromatin as control.

Table 16 *Overlap* analysis from *HiBrowse* between CTCF, SMC3, RAD21 and ChIA-PET against VDRML and VDRMB datasets with a p-value estimated using *HiBrowse* for the alternative hypothesis presented in Table 15.

Analysis	Track 1	Track 2	P-value
Overlap?	ChIA-PET	VDRMB	9.999e-05
Overlap?	ChIA-PET	VDRML	9.999e-05
Overlap?	RAD21	VDRML	9.999e-05
Overlap?	RAD21	VDRMB	9.999e-05
Overlap?	SMC3	VDRMB	9.999e-05
Overlap?	SMC3	VDRML	9.999e-05
Overlap?	CTCF	VDRML	9.999e-05
Overlap?	CTCF	VDRMB	9.999e-05
Overlap?	CTCF	RAD21	9.999e-05
Overlap?	CTCF	SMC3	9.999e-05
Overlap?	VDRML	Heterochromatin	1
Overlap?	VDRMB	Heterochromatin	1

The *overlap analysis* shows overlap between ChIA-PET, RAD21, SMC3 and CTCF to VDRMB and VDRML than would be expected by chance. Significant overlap was also measured between ChIA-PET data and ML, MB datasets AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1 (Table 35). However, there was no difference in significance relating to estimated overlap between the measured tracks.

3.6 Enhancers and promotors association to ML and MB

To see if the ML DNA segments would be more associated with enhancers and the MB segments to the promotors an overlap analysis between VDRMB/VDRML to enhancers and promotors was completed. Overlap was also measured for repressed regions as a control.

Table 17 *Overlap* analysis from *HiBrowse* for VDRMB/VDRML, strong/weak enhancer and Active/weak promotor and repressed regions as well as the p-value estimated using *HiBrowse* for the alternative hypothesis represented in Table 15.

Analysis	Track 1	Track 2	p-value
Overlap?	VDRML	Strong enhancer	9.999e-05
Overlap?	VDRMB	Strong enhancer	9.999e-05
Overlap?	VDRML	Weak enhancer	9.999e-05
Overlap?	VDRMB	Weak enhancer	9.999e-05
Overlap?	VDRML	Active promotor	9.999e-05
Overlap?	VDRMB	Active promotor	9.999e-05
Overlap?	VDRML	Weak promotor	9.999e-05
Overlap?	VDRMB	Weak promotor	9.999e-05
Overlap?	VDRML	Repressed	1.0
Overlap?	VDRMB	Repressed	1.0

The *overlap analysis* shows overlap between VDRMB/VDRML and all enhancers and promotor datasets, but not for repressed. However, no difference between the measured tracks was observed.

The next analysis that was completed was the hypothesis testing tool *located nearby* from *HiBrowse*. A Monte Carlo model was used and the MCFDR sampling depth was set to 10 000 samples. The alternative hypothesis was set to closer to and the null model was set to preserve points of T2 and inter-point distances of T1; randomize positions (T1). A Monte Carlo model was used and the random seed was set to random. Details are described in Table 18.

Table 18 Parameters used to detect nearby segments using the *located nearby* tool from *HiBrowse* for two tracks that estimates if points on track 1 lies closer to points on track 2 more than is expected by chance.

Located nearby	
Genome build	Human Feb. 2009 (hg19/GRCh37)
MCFDR sampling depth	Fixed 10 000 samples
Alternative hypothesis	Closer to
Test statistic	Geometric mean of distances (bp)/ Arithmetic mean of distances ($\log_{10}$)
Null model	Preserve points of T2 and inter-point distances of T1; randomize positions (T1) (MC)
Random seed	Random

null hypothesis	The points of track 1 are located independently of the points of track 2
alternative hypothesis	The points of track 1 are located close to the points of track 2

The *located nearby* analysis estimated the localization of VDRML and VDRMB in regards to enhancers and promotors and a p-value for the alternative hypothesis presented in Table 18 was estimated.

Table 19 *Located nearby* analysis estimating if the points of track one are located close to points of track 2 more than is expected by chance. The results are given as average log distance between neighbours and a p- values based on the null and alternative hypothesis presented in Table 18.

Analysis	Track 1	Track 2	Test statistic: Average log-distance	p- value
Located nearby	VDRML	Strong enhancer	5.711	9.999e-05
Located nearby	VDRMB	Strong enhancer	5.456	9.999e-05
Located nearby	VDRML	Active promotor	6.745	9.999e-05
Located nearby	VDRMB	Active promotor	6.635	9.999e-05

Results presented in Table 19 show that VDRMB and VDRML are both located close to enhancers and promotors more often than is expected by chance. Additionally, it could seem like VDRMB are closer to enhancers and promotors than VDRML. However, no difference between the different experiments was measured.

3.7 DHSs

Doing an *overlap analysis* between DHSs and VDRMB/ML would indicate whether or not the VDR datasets are located close to DNA with open chromatin structure. This was estimated using the *overlap analysis* described previously in this thesis. Overlapping regions was estimated between VDRMB/ML and DHSs regions.

Table 20 Overlap analysis from *HiBrowse* for VDRML, VDRMB and DHSs as well as the p-value estimated using *HiBrowse* for the alternative hypothesis presented in Table 15.

Analysis	Track 1	Track 2	P-value
Overlap?	VDRMB	DHSs	9.999e-05
Overlap?	VDRML	DHSs	9.999e-05

The DNA chromatin states in or around VDRMB and VDRML tracks are based on the previous test most likely in an open chromatin state. However, no significant DHSs difference was measured in overlap between VDRML and VDRMB.

3.8 HOT and LOT regions relating to ML and MB

Another way to measure the relative positions of TFs inside MB or ML segments are too look for HOT regions which are occupied by many TFs. To estimate if ML and MB segments are located inside HOT regions more than expected by chance the *Located Inside* tool from *HiBrowse* was used. The analysis was set to “Do track 1 fall inside track 2, more than expected by chance?” and the MCFDR sampling depth to 10 000 samples. The alternative hypothesis was set to more than expected by chance. Random seed was set to random and null model was set to “Do track 1 fall inside track 2, more than expected by chance?”

Table 21 The *Located Inside* tool from *HiBrowse* was used to measure if the middle points of each VDRML and VDRMB was located inside HOT regions more than is expected by chance.

Located inside	
Genome build	Human Feb. 2009 (hg19/GRCh37)
MCFDR sampling depth	Fixed 10 000 samples
Alternative hypothesis	more
Treat track 1 as	The middle point of every segment
Treat track 2 as	Original format
Test statistic	Arithmetic mean of differences
Null model	Preserve points (T1), segment lengths and inter-segment gaps (T2); randomize positions (T2) (MC)
Random seed	Random
null hypothesis	The points of track 1 are located independently of the segments of track 2 with respect to whether they fall inside or outside
alternative hypothesis	The points of track 1 tend to fall inside the segments of track 2

The test was completed for VDRML and VDRMB against HOT and LOT datasets for whole genome and intergenic datasets. It was also tested with LOT regions as a control

Table 22 *Located inside* analysis between VDRML, VDRMB, HOT whole genome, HOT, intergenic, LOT whole genome and LOT intergenic datasets using the parameters described in Table 21.

Analysis	Track 1	Track 2	P-value
Located inside?	VDRML	HOT Whole Genome	9.999e-05
Located inside?	VDRMB	HOT Whole Genome	9.999e-05
Located inside?	VDRML	HOT intergenic	9.999e-05
Located inside?	VDRMB	HOT intergenic	9.999e-05
Located inside?	VDRML	LOT intergenic	1.0
Located inside?	VDRMB	LOT intergenic	1.0
Located inside?	VDRML	LOT whole genome	1.0
Located inside?	VDRMB	LOT whole genome	1.0

The *located inside* analysis showed significant results between VDRML and VDRMB inside HOT whole genome as well as HOT intergenic, but not for the LOT tracks. However, it was not possible to detect any difference between MB and ML enrichment.

3.9 TADs, -a structural framework for gene regulation

TADs are DNA 3D structures and within TADs there are many DNA loops bound together by cohesin that all together creates insulated neighbourhoods. Inside these TADs there are HOT regions that contain genes that because of the 3D structure of the TAD can be regulated by only a few TFs. Doing a *Located Inside* analysis with VDRML and VDRMB inside TADs would indicate if VDRML or VDRMB are located inside TADs more often than expected by chance.

Table 23 *Located Inside* analysis for VDRML and VDRMB measuring if they are located inside TADs more often than what is expected by chance.

Analysis	Track 1	Track 2	P-value
Located inside?	VDRML	TAD	9.999e-05
Located inside?	VDRMB	TAD	9.999e-05

The result presented in Table 23 indicates that VDRML and VDRMB are located within TADs more often than is expected by chance. However, no difference was measured for ML-TAD or MB-TAD.

To further investigate the matter of ML and MB interacting in pairs inside TADs, The *Bedtools* tool *Intersect* was used to estimate the number of times each VDRML or VDRMB overlaps the same TAD. In this case TADs was used as -a and VDRMB and VDRML were used as –b files. This way TAD regions are treated as query and both VDRML and VDRMB as input looking for overlapping segments with TADs. The TAD data was clustered into chromosomes. The parameters used are described in Table 24.

Table 24 Options used in the *intersect* analysis from *Bedtools* version 2.25.0 and their function.

Bedtools intersect	
-a	TAD
-b	VDRMB, VDRML
-wa	Write the original entry in A for each overlap.
-wb	Write the original entry in B for each overlap. Useful for knowing what A overlaps. Restricted by -f and -r.
-filenames	When using multiple databases (-b), show each complete filename instead of a field when also printing the DB record.

The *Intersect* tool from *Bedtools* was used to measure how often each VDRMB and VDRML *intersects* with TADs. The result of this analysis is shown below.

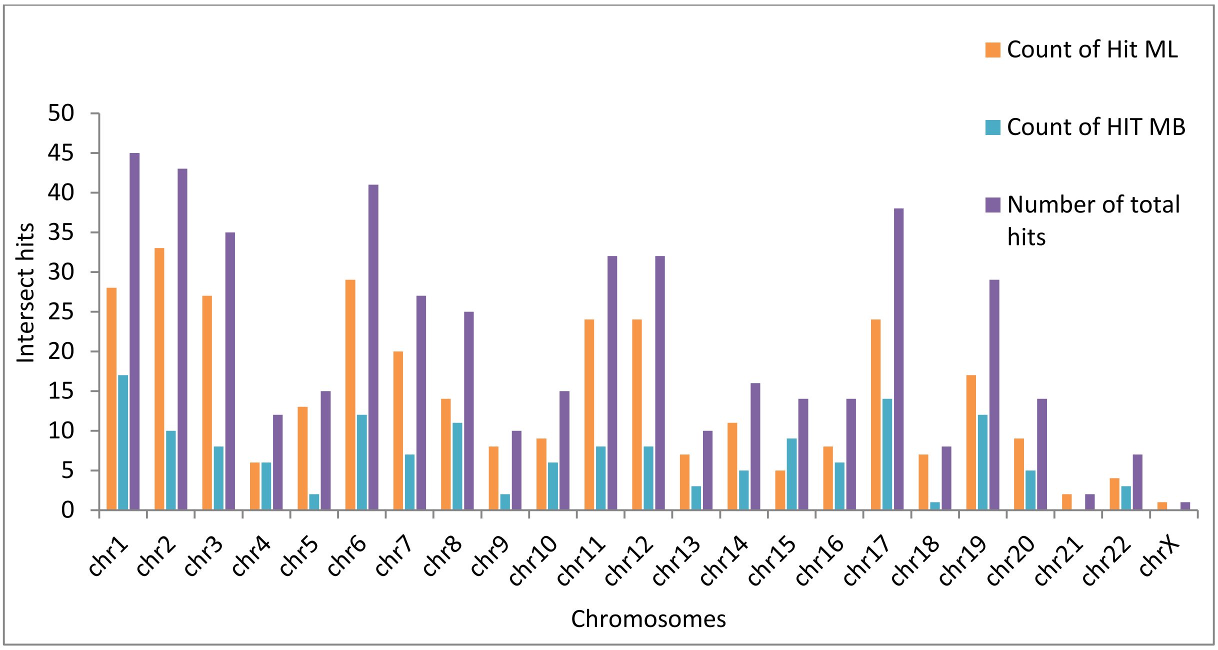


Figure 9 Showing a histogram that show counts of how often VDRML and VDRMB overlaps TADs in different chromosomes. In this figure the number of overlapping ML (orange) and MB (blue) with TADs are given for each chromosome. The total number of hits for each chromosome is given in purple.

Only 280 cases of VRML and VDRMB showed overlap with TADs. There was a total of 1731 TADs in the original dataset and only 86 of them shows overlap with either VDRMB or VDRML.

3.10 Are ML and MB interacting locally in pairs?

The next subject that was examined was if there were specific interactions between ML and MB segments. It was hypothesized MB and ML could work as stems for loops inside TAD structures. The idea was that MB and ML, would inside TADs form loops by interacting locally with each other, forming stem like looped structures. As can be seen in Figure 1 the idea was that ML and MB could be the binding sites of cohesin and also facilitators of the TAD structures. To examine this the length of an average TAD was measured to see if the total length of a VDRML segment combined with the intersegment distance between VDRML and VDRMB and the length of an VDRMB would exceed the average TAD length. This was examined using a descriptive tool from *HiBrowse* called *AVG Segment Length*. The analysis was set to the average length of the track in question and overlaps was handled as clusters and described in Table 25.

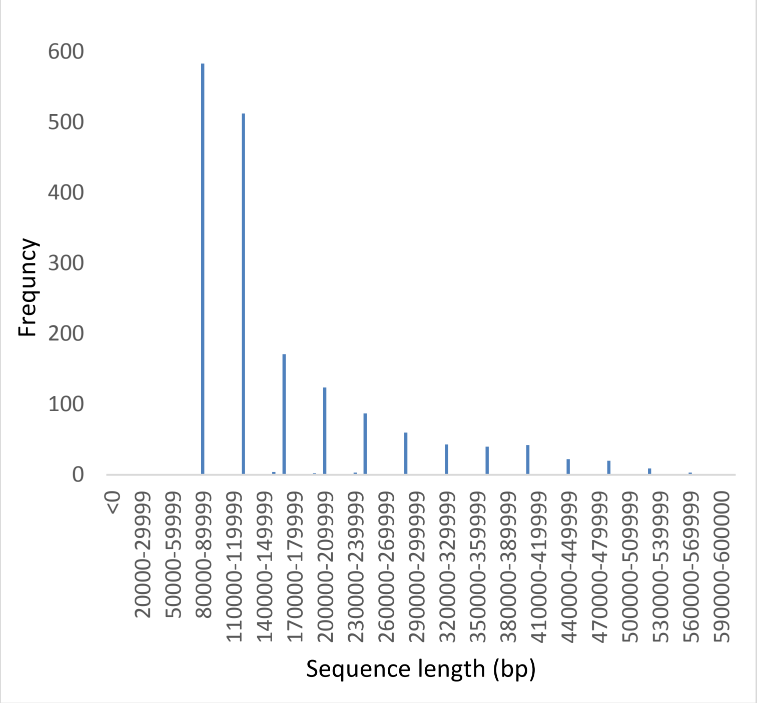
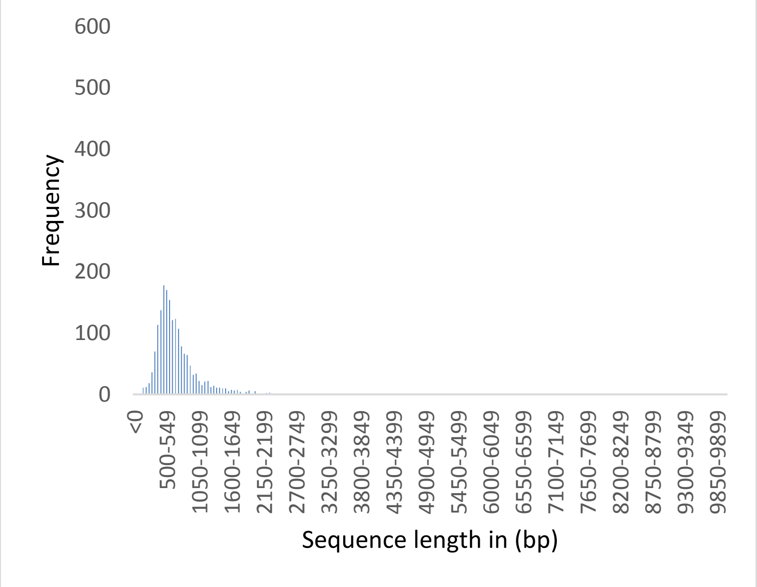
Table 25 Parameters used by the descriptive tool *Average Segment Length* in *HiBrowse* to obtain information about the average segment length of datasets, TADs, VDRML and VDRMB.

Avg. segment length	
Genome build	Human Feb. 2009 (hg19/GRCh37)

Analysis	The average length of Track
Overlap handling	Cluster overlaps

The average segment length was estimated for TADs, VDRML and VDRMB. The sequence length frequency distribution was also included in the results.

Table 26 ***Average Segment Length*** **analysis for TADs, VDRML and VDRMB including the sequence length frequency distribution.**

Analysis	Track 1	Avg. Length	Sequence length frequency distribution
Average Segment Length	TADs	1.573e+05 bp	
Average Segment Length	VDRML	694.1 bp	

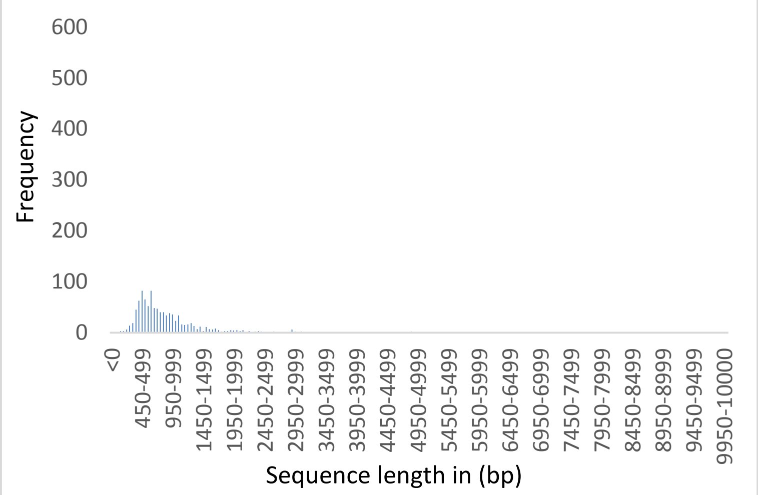
Average Segment Length	VDRMB	930.8 bp	

Table 26 shows that TADs are distributed in units of 40 kilo base pairs (kbp) with an average length of 1.573 million bp. VDRML show a peak around 700 bp and VDRMB a peak around 950 bp.

After the average segment length was estimated the *Segment Distance* analysis between the closest pairs VDRML and VDRMB datasets was completed in *HiBrowse*. The analysis was made between each VDRML to the nearest VDRMB segment.

Table 27 The *Segment Distances* tool from *HiBrowse* measures the average segment distance between closest pairs of segments and returns a frequency distribution of these intersegment distances.

Segment distances	
Genome build	Human Feb. 2009 (hg19/GRCh37)
Analysis	The distribution of distances from each segment of Track 1' to the nearest segment of Track 2
Overlap handling	Cluster overlaps of track 1

Inter segment distances viewed in Figure 10 are in Log$_{10}$ values.

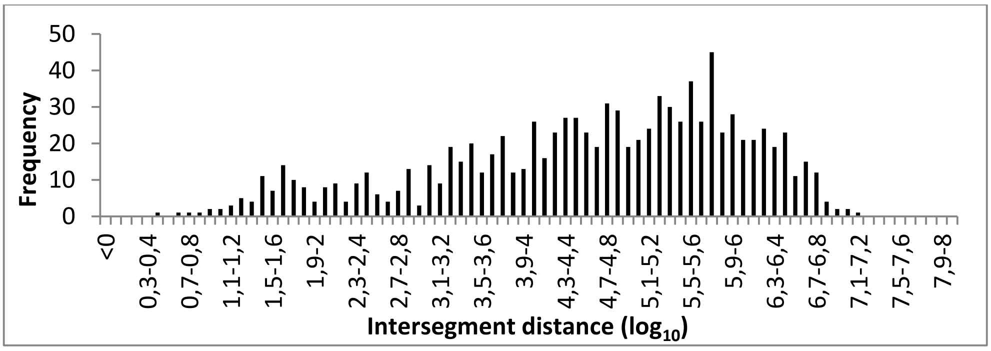


Figure 10 Frequency distribution of distances in log$_{10}$ from each VDRMB to the nearest VDRML made using the *Segment Distance* tool from *HiBrowse*.

Then the data for TADs presented in Table 26 was reproduced using Log$_{10}$ values. Results are shown in Figure 11.

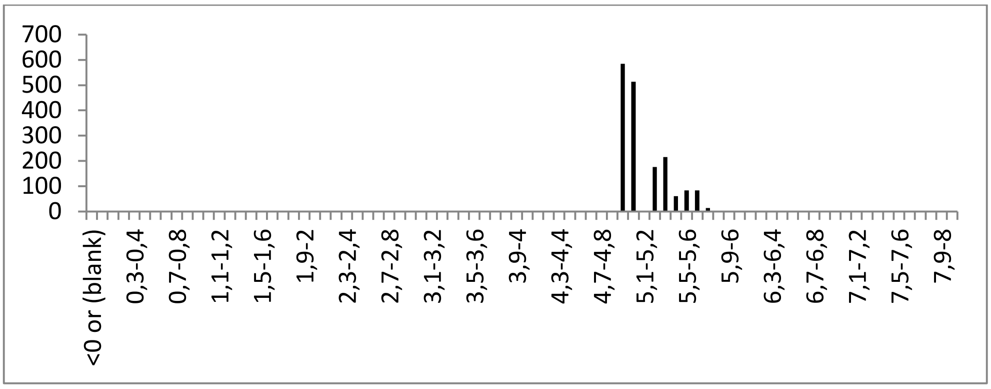


Figure 11 Showing the frequency distribution of TADs in $\log_{10}$ values.

Both Figure 10 and Figure 11 show a peak around log 5,5 which could indicate that the average intersegment distance between VDRMB and VDRML are located within the boundaries of a TAD. However, this is unlikely as the majority of TADs are below log 5,5.

If MB and ML work in pairs forming local interactions, it would be expected to find them in MB-ML pairs located close together on the genome strand. To investigate the distance, the tool *closest* was used on MB and ML pairs from TFs VDR, AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1. The analysis was completed using both MB and ML as –a files and both as b- files meaning two runs for each TF. A tool from *Bedtools* version 2.25.0 was used to measure the intersegment distance from each MB to the nearest ML segment using the tool *closest* using option –N, -mdb all, and –filenames and both ML and MB as –b files.

Table 28 Parameters used in the *closest* analysis from *Bedtools* version 2.25.0 and a short description of the settings used. The test was completed in two runs for the VDR datasets using different a files but, both as b files all times.

Bedtools closest	
-a	*[TF]MB/[TF]ML*
-b	*[TF]MB,[TF]ML*
-d	In addition to the closest feature in B, report its distance to A as an extra column. The reported distance for overlapping features will be 0.
-N	Require that the query and the closest hit have different names. For BED, the 4th column is compared.
-mdb all	Specify how multiple databases should be resolved. All reports closest records among all databases.
-filenames	When using multiple databases (-b), show each complete filename instead of a field when also printing the DB record.

Excel was used to do calculation on the *Bedtools closest* dataset and used to calculate the average intersegment distance for all non-overlapping segments. To limit the number of false positives a cut-off at 200 kbp was used. Figure 12 shows the total amount of VDRMB and VDRML pairs estimated using different –a files.

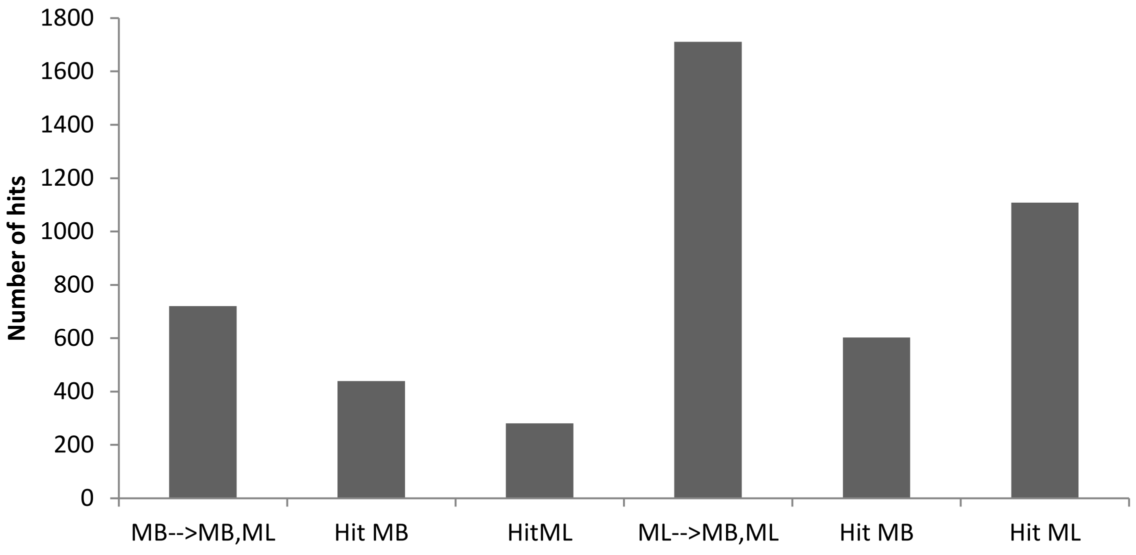


Figure 12 Histogram showing the closest (in bp) VDRML/VDRMB segment for each VDRMB and VDRML across both datasets. Each column denoted Hit MB and Hit ML represents the number of times VDRMB (Hit MB) or VDRML (Hit ML) are located closer to the query than the other. The columns MB →MB,ML and ML→MB,ML shows the total number of cases found for each situation.

The same analysis was completed for the ten other TFs described in Figure 13 also using a cut-off of 200 kbp and the same parameters.

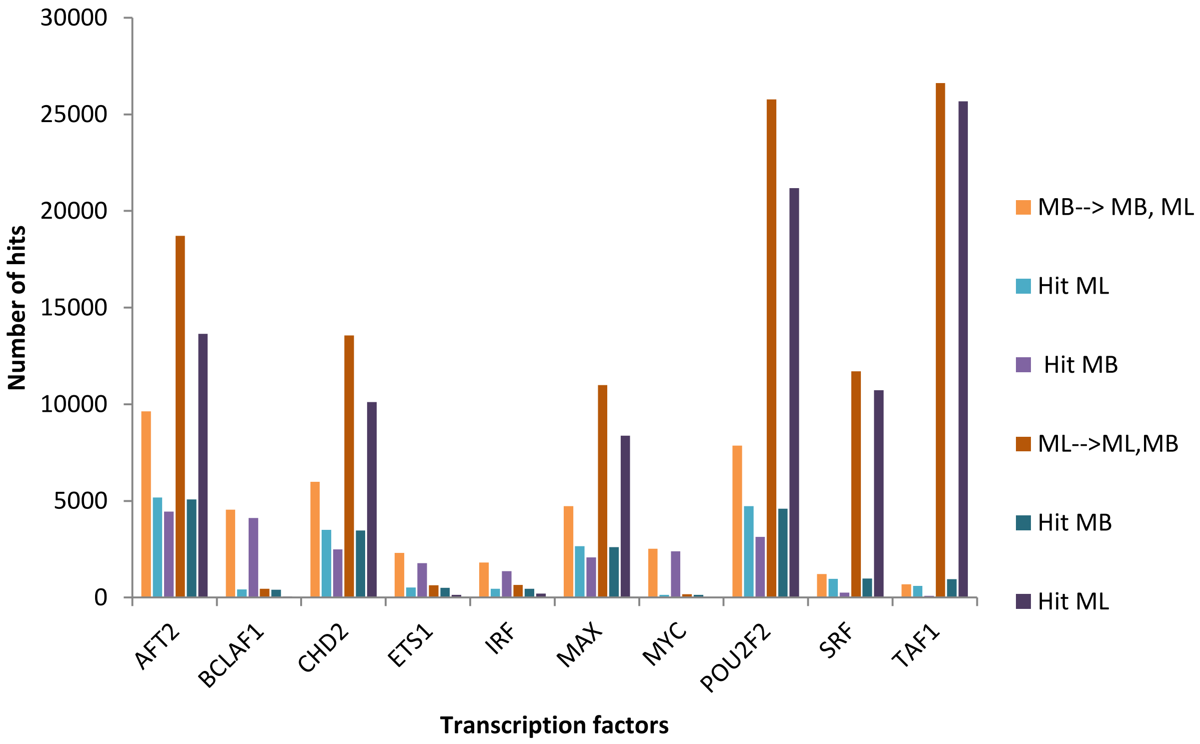


Figure 13 in the figure above MB →MB, ML and ML→ML,MB is the total number of hits found for sequentially MB and ML datasets given parameters described in Table 28 . Hit ML is the number of ML found and Hit MB is the number of MB found for both ML→ML,MB and MB→MB,ML.

Figure 12 does not indicate that MB and ML are located close to each other in pairs. However, there are indications of ML being located close to ML more often than it is located to MB.

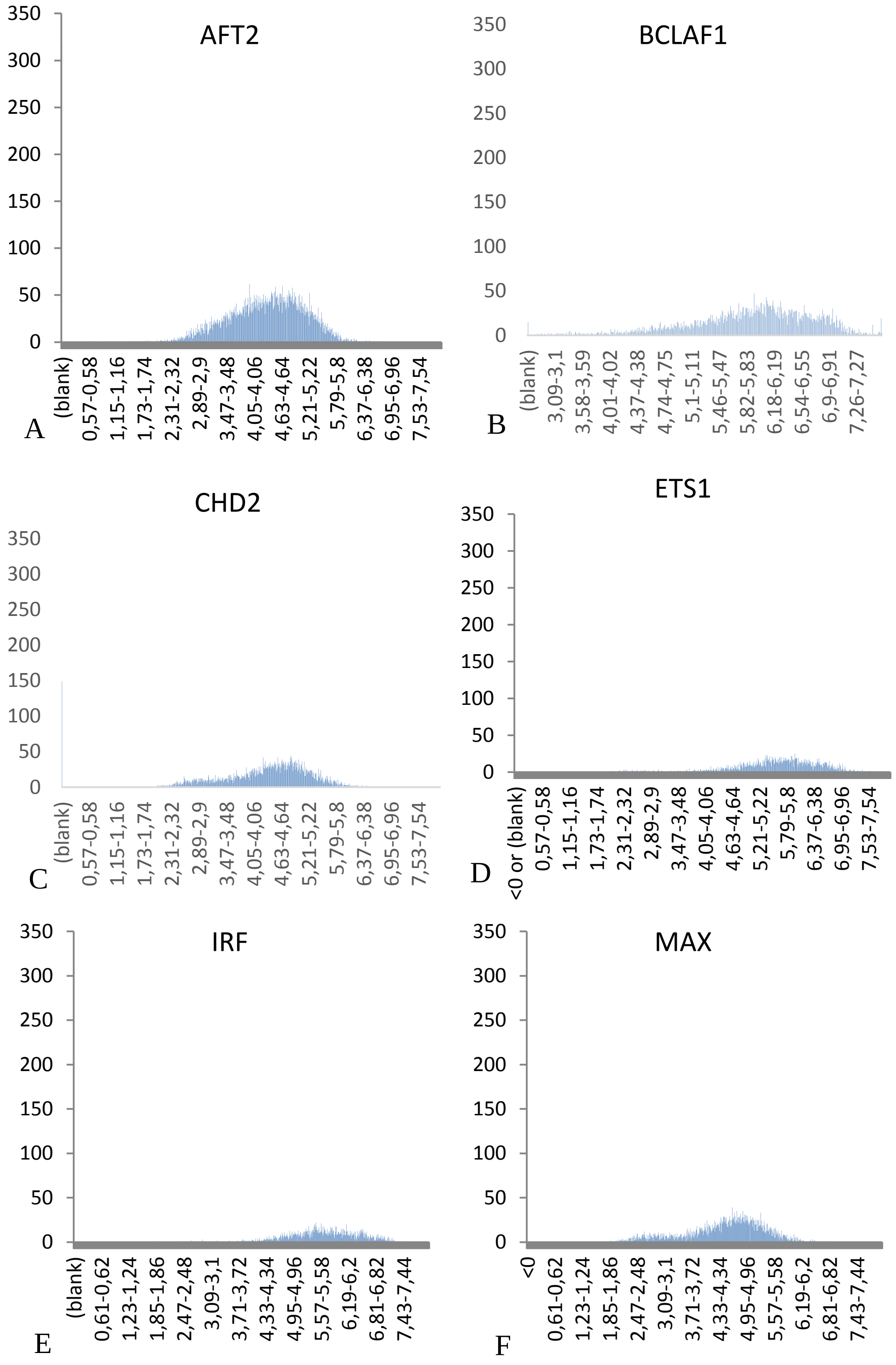
Table 29 The tools *Average Segment Length*, *Located Nearby* (aritmic and geometric) and *Segment Distance* from *HiBrowse* was used to estimate the closest located pairs of ML and MB segments for VDR, AFT2, BCLAF1, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1 using previously described parameters described in Table 18, Table 26, Table 27.

	Average segment length – ML (bp)	Average segment lengths MB (bp)	Located nearby geometrics (p-value)	Avg. arithmetic distance between MB and ML (bp)	Avg. geometric distance between MB and ML ($\log_{10}$)
VDR	694.1	930.8	0.04762	6.002e+05	10.91
AFT2	430.5	469.2	9.999e-05	7.738e+04	9.896
BCLAF1	377,9	403,8	9.999e-05	*	13.36.
CHD2	376,9	379,8	9.999e-05	9.681e+04	10.02
ETS1	210.9	220.7	9.999e-05	1.573e+06	12.84.
IRF3	249,3	279,4	9.999e-05	*	12.84
MAX	445,9	442,5	9.999e-05	1.243e+05	10.17
MYC	596	617,7	9.999e-05	*	13.85
POU2F2	280,5	294,8	9.999e-05	9.359e+04	9.958
SRF	210,2	207,2	9.999e-05	2.195e+05	10.87
TAF1	303.2	314,6	9.999e-05	5.627e+04	9.154

*For the *located nearby* analysis using arithmetic distance *HiBrowse* was not able to estimate a p-value and a collection of FDR-corrected p-values per bin were computed instead. For the alternative hypothesis used by *HiBrowse* these simplistic answers were returned for the following TFs. BCLAF1, Yes - the data supports H1 at least in some bins (36 significant bins out of 40, at 10% FDR). ETS1 Yes - the data supports H1 this at least in some bins (35 significant bins out of 41, at 10% FDR). MYC Yes - the data supports H1 at least in some bins (25 significant bins out of 37, at 10% FDR).

The results presented in Table 29 show that ML and MB segments for TFs VDR, AFT2, BCLAF1, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1 are more often than expected by chance located close to each other.

The *closest* tool was then used to find the distance between the closest pairs of MB and ML using only MB as –a file and only ML as –b file using –filenames as only parameter. This was done to see if there was a favourable intersegment distance between MB and its closest ML segments. MB was believed to be the facilitator of the looping and therefore it was used as –a file and ML as –b file. Figure 14 below shows the distribution of distances between MB and its closest ML neighbour. This analysis was completed for TFs AFT2, BCLAF1 CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF, TAF1 and VDR. One segment located at the X chromosome in the IRF dataset were excluded, as it had no neighbour on the same chromosome. All datasets have a pseudo count of +1 to avoid the problem of log (0).



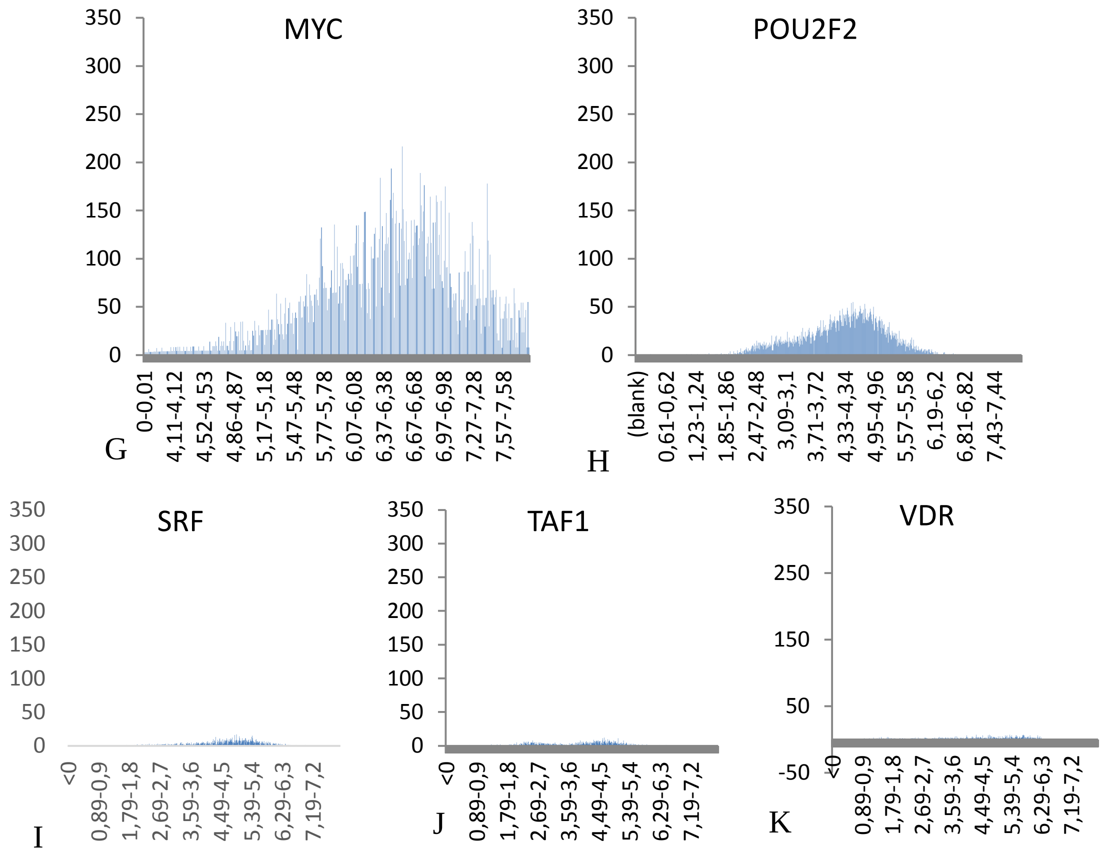


Figure 14 Frequency distributions of intersegment distance, in $\log_{10}$ bp, between closest pairs of ML and MB for TFs AFT2(A), BCLAF1(B), CHD2(C), ETS1(D), IRF3(E), MAX(F), MYC(G), POU2F2(H), SRF(I), TAF1(J) and VDR (K) with pseudocount +1. The y-axis in figures A-K show a frequency distribution and the x-axis show the intersegment distance between MB and ML in $\log_{10}$(bp).

Results presented in Figure 14 does not indicate a favourable intersegment distance between MB and ML for any of the TFs presented in Figure 14.

Finally the complete set of data regarding the position of VDRMB and VDRML datasets in regards to ChIA-PET, HOT regions, DNase hypersensitivity sites, CTCF, Enhancers, Promotors, SMC3 and Rad21 were combined in Figure 15 using the estimated p-values for the different test completed in this project.

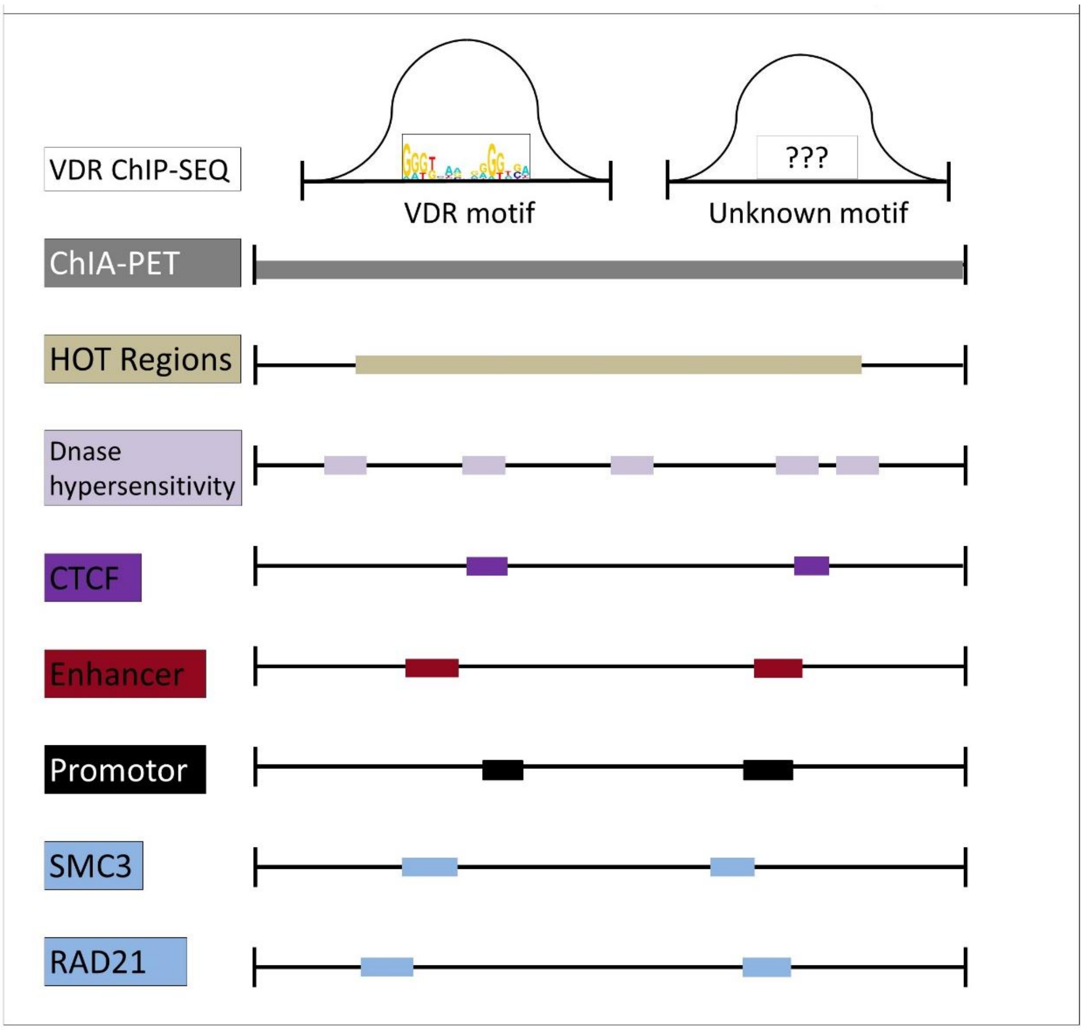


Figure 15 shows a visual presentation of some of the *overlap* experiments completed in this project and how the results can be used to visualise the positional relationship between VDRMB/VDRML peaks in relation to HOT, DHSs, CTCF, Enhancers, Promotors, SMC3, RAD21. The two peaks presented represents the ChIP-seq results obtained from Ramagopalan et al and beneath are the datasets that show significant overlap with VDRML and VDRMB. The black line between the separators represents the human DNA strand and the coloured areas in between represents genetic features as indicated on the left. Adapted by permission from Elsevier: Cell Press [16], copyright (2016).

Figure 15 shows, based on the p-values estimated by *HiBrowse*, how VDRMB and VDRML could relate to the other genomic features presented in the figure above.

4 Discussion

4.1 *USCS Genome Liftover* and motif scanning by *MotifLab2*

Eleven TFs known to show signs of ML binding was examined in this project and they were AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF, TAF and VDR. All TF datasets were predicted, by *MotifLab2*, to include ML sequences, see Table 11. However, before any motif predictions could be completed, the VDR dataset needed to be changed from NCBI36/hg18 to GRCh37/hg19. This was necessary as coordinates of different genome builds can change from one assembly to the next as gaps are closed and duplications are reduced. Occasionally, a chunk of sequence may be moved to an entirely different chromosome as the genome build is updated. To be able to compare datasets it is therefore important to use the same genome build, as coordinates will be different in different builds. This was also the case for the VDR dataset where the original dataset was made using hg18. Because all other datasets used in this project was already in the hg 19 build the VDR dataset needed to be changed to hg19 using the *USCSs liftover* tool. Figure 8 compares the start and stop positions before and after lift over. All segment intervals were conserved during the liftover except one. A segment on chromosome 4 with start site 103 967 094 and stop site 103 969 025 with a length of 1 931bp was changed to chromosome 4 start site 103 747 959 and stop site 103 749 887 with a length of 1 928bp. However, the segment was predicted ML in both hg18 and hg19. As Figure 8 shows there is little difference between Ramagopalans VDR data before and after *USCS liftover* and it is also worth mentioning that the same number of segments was predicted ML in both VDR-hg18 and VDR-hg19.

The VDR sequence motif M00444 from the motif collection TRANSFAC_public was used to predict the VDRML genetic segments from the VDR dataset. *Simple scanner* requires a percentage threshold of similarity for identifying positive matches along the DNA. By lowering this threshold, more segments are likely to be predicted MB so finding a threshold with high enough stringency and still predicting both ML and MB segments was necessary. VDR ChIP-seq peaks scanning in *MotifLab2* returned an uneven number of VDRML and VDRMB sequences where 1784 peaks were predicted to be ML and 992 were predicted to be MB at 90% similarity threshold. Different scans using the same motifs were completed but, with different percentages to find the optimal threshold for each TF, see Table 11. The VDRML and VDRMB DNA segments predicted at 90% were selected as this threshold predicted a high number of sequences in each VDRML and VDRMB dataset without losing much stringency. Ten other TFs were also examined for ML binding and these were AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1.

Here we assume that there are ML cases in all the TFs above without knowing that this is actually the case. In this project, a trial and error method was used to find thresholds for motif scanning. By doing comparisons of results from motif scans using different thresholds a number was selected to use, see section 3.2. However, a more advanced method could have been applied. It is possible to use a more objective basis for motif discovery by using for example, *Find Individual Motif Occurrences* (FIMO) that measures the significance of the motifs found in the sequence dataset. Additionally, this would enable us to use a p- value as threshold for finding significant motifs ensuring that only significant motifs are included. This means that the MB dataset would include only segments that have significant motifs and the ML dataset would not.

Another way to do motif discovery could have been to use a saturation approach. Then the numbers of segments included would have been plotted as the threshold was gradually lowered. By sequentially lowering the threshold and plot the number of motifs against the current threshold we could identify the real ML segments. A curve would appear that initially grew slowly until there were motifs in for example 50% of the segments before it would almost stop changing and then start again at a lower threshold. At this point there would be mostly noise. However, this curve will only appear in TF ChIP-seq dataset that include true ML segments. Using one of these approaches would have given a better prediction of actual ML cases in the ten TF examined. The VDR dataset however, have been tested for significant ML segments using FIMO [50, 51].

In a paper published in 2012 by Foley, ML binding is described as indirect recruitment by another TF whose motif is present [22]. This implies that the ML peaks are in fact MB peaks that contain a binding site for another TF. However, it would be likely that this TFs motif would show up in de novo motif discovery of the ML peaks. This was done by Handel in 2013 where he used MEME and de novo motif detection to find TF binding sites in a VDR dataset. After completing this Handel et. al. identified cases of ML peaks in a VDR ChIP-seq dataset [52]. Handel further argues that VDRML ChIP-seq peaks are caused by gene-environment interactions, which supports the 3D colocalization results presented in this thesis (Table 14).

4.2 ML-MB distance in 3D genome space and cohesin

After the TF datasets had been scanned by *MotifLab2*, ML and MB datasets were uploaded to *HiBrowse* for further analysis. The results from section 3.4 indicates that ML and MB sequences are located closely together in 3D genome space. *HiBrowse* measures physical distance in 3D between two DNA points and returns whether or not these two points are localised in 3D. However, it does not measure actual interactions between two segments. The results from Table 16 shows that the analysis done in *HiBrowse* predicts ML and MB to be localised in 3D genome space more often than is expected by chance. On the other hand, the interactions might not be MB-ML as initially expected, but rather ML-ML or MB-MB. Table

14 shows that there is significant colocalization between ML-ML, MB-MB and MB-ML which indicates that there could be a point in X,Y,Z dimensions inside the cell nucleus where these DNA segments localize.

Another dataset used in this project that contains information about 3D genome structure was a ChIA-PET dataset for cell line GM12878 using RAD21 as target. Hi-C or 5C was expected to best represent 3D genome structure, but no such dataset was found for cell line GM12878 at the time of the project. However in an article published by Li et al in 2012, Li argues that ChIA-PET data are more detailed and with a higher resolution in regards to specific protein factor-mediated chromatin interactions than the 3C assays [33]. Li argues that when looking for TF specific interactions, ChIA-PET datasets are more detailed and gives a better resolution than 3C assays. On the other hand, it was necessary to use an all to all approach to look for 3D genomic interactions and not only RAD21 specific chromatin interactions. Additionally, RAD21 is a subunit in the cohesion complex and is known to be an important player in facilitating 3D genome structure through the cohesion complex. Therefore, the fact that the ChIA-PET dataset used in this project contained RAD21 specific chromatin interactions was expected not to influence the results of this project. However, using a Hi-C dataset for GM12878 would have been favourable.

Both VDRMB and VDRML were estimated to overlap with the ChIA-PET dataset more than expected by chance (Table 16). This indicates that ChIA-PET DNA crosslinks are close in bp distance to both VDRML and VDRMB. This again could mean that VDRML and VDRMB are close enough in space to form interactions. However, both VDRML and VDRMB show significant overlap with the ChIA-PET dataset and therefore the interactions could also be VDRML-VDRML or VDRMB-VDRMB. It is also important to consider that even though ChIA-PET data contains a library of interacting segments this does not mean that VDRML and VDRMB are active in DNA interactions just because they show significant overlap with ChIA-PET. It could however be the case that they happen to be located close to the interacting segments, in bp distance, without actually taking part in the interaction.

The results presented in Table 14 predicts VDRML and VDRMB to be localised in 3D with CTCF, RAD21, SMC3 and each other more than is expected by chance. These results could imply that cohesin works to create interactions between segments in the VDR datasets. Another *3D colocalization* experiment that is presented in the same table involves the different proteins in the cohesin complex as well as CTCF that all show significant 3D colocalization with the VDR datasets. The hypothesis presented in the introduction of this thesis is that DNA looping can cause the DNA strand to bend and interact with itself, or distally to another chromosome, resulting in ML peaks. Cohesin has been shown to affect the genome 3D structure and this is also believed to affect gene regulation [7, 10]. Tang Z. writes in an article published in 2015 *"We find that CTCF/cohesin-mediated interaction anchors serve as structural foci for spatial organization of constitutive genes concordant with CTCF-motif orientation, whereas RNAPII interacts within these structures by selectively drawing cell-type-specific genes toward CTCF foci for coordinated transcription"*[53]. Here TANG et al. describes a 3D genome model that by CTCF motif orientation creates open and closed compartments of chromatin. This compartmentalization of genetic regions using CTCF as boundaries are similar to TAD structures. These findings correlate well with the results presented in Ji X. article from 2015 presenting TADs and how they form insulated neighbourhoods [16]. In this project *HiBrowse* predicted significant 3D colocalization and overlap between the cohesin subunits, CTCF, and ChIA-PET datasets with MB and ML datasets for AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1, see Table 30 in attachments. The results from the *3D colocalization* analysis supports the hypothesis of 3D interactions between segments in the TF datasets and cohesin. However, the fact that there was not measured any differences in p-values for MB-MB, ML-ML and MB-ML, for any of the previously mentioned TFs makes these results inconclusive in regards to the main hypothesis in this report.

4.3 TADs, possible inter domain position of MB and ML

The results presented in section 3.9 show that VDRML and VDRMB are located within TAD regions more often than is expected by chance. The TAD data used in this project was created by looking for long range interactions between 40 kbp bins from a ChIA-PET experiment using SMC1 as target. This is why the TADs presented in Table 26 show peaks around multiplications of 40 kbp [16]. Additionally, the TADs were predicted for embryonic stem cells (hESCs) and not lymphoblastoid cells (GM12878) which have been used in this project. Even though TADs are conserved throughout mammals the fact that they are described for a different cell line could affect the results predicted by *HiBrowse*. Therefore, a dataset focusing on TADs in GM12878, and using better resolution, is needed before an in depth analysis of the TADs and ML relationship can be determined.

In an article written by Ji X et al 2016 he describes TADs as important factors for making up the 3D regulatory genome landscape [16]. Finding significantly overlapping DNA regions between TADs, VDRMB and VDRML could indicate that VDRML and VDRMB are both located inside TADs. *HiBrowse* predicts, as showed in Table 23, that VDRML and VDRMB both locate inside TADs more often than is expected by chance. However, it does not indicate that they overlap TADs in pairs. This was further investigated using the *intersect* tool from *Bedtools* identifying overlapping VDRML and VDRMB with TADs. In this analysis 486 cases of overlapping TADs were found from a total of 1731 initial TADs. 155 of these overlaps was with VDRMB and 330 with VDRML and the results can be seen in Figure 9. This result does not indicate that MB and ML are located inside TADs in pairs. This is because the overall overlap frequency is low. However, the results could be affected by the lack of a detailed TAD dataset as mentioned earlier. On the other hand, the results show multiple ML overlaps with each MB segment inside each TAD. This could indicate possibly multiple ML interactions with each MB inside the same TADs. This would, if true, correlate well with the idea of functional chromatin topological domains described as open chromatin regions by Tang Z. and TADs presented by Ji X. [16, 53]. However, this could also be explained by false positives from the initial motif scanning where these appear as ML while they are really just noise making the total ML count to high.

Another way to test for MB-ML pairs located closely in bp distance is by using the *Segment Distance* tool from *HiBrowse*. If ML and MB form locally interacting pairs one possible explanation could be interactions inside TAD complexes between pairs of MB and ML segments. One hypothesis was that MB and ML could be working as binding sites for cohesin in creating the CTCF loops as shown in Figure 1. This was based on the results from Table 23 that shows that VDRML and VDRMB both overlap TADs as well as the overlap measured between VDRML and VDRMB with CTCF in Table 16. If true, the closest pairs of VDRML and VDRMB including intersegment distance should be within the boundaries of an average length of a TAD. The tools *Average Segment Length, Located nearby and Intersegment Distance* from *HiBrowse* was used to measure the length of the ML and MB segments for all the eleven TFs used in this project. The results from these experiments are shown in Table 29. Additionally, the average length of each segment as well as providing a frequency distribution of sequence length and intersegment distance between the closest pairs of VDRML and VDRMB is shown in Figure 10, Figure 11, and Table 26. The average intersegment distance between the closest pairs of VDRML and VDRMB, as predicted in *HiBrowse* suggests that this could be possible. However, this is unlikely as most TADs seem to be shorter than the average intersegment VDRML-VDRMB distance. The same analysis was completed for MB and ML datasets for AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1 and they all showed similar results as the VDR datasets.

4.4 HOT, LOT, promotor and enhancers position to MB/ML

Another sub hypothesis that was examined during this project was that MB would associate with promotors and MLs would associate with enhancers. Therefore, it was expected to find MB inside HOT regions more often than ML segments. This was because HOT regions are associated with a high number of promotor binding sites. This was expected because MB segments contain binding sites for the TF. Additionally, multiple tests estimating MB and MLs relative bp distance to promotors and enhancer was completed and can be found in the attachments (Table 31 and Table 32). The results from Table 17 was predicted using *HiBrowse* and the results indicate that both ML an MB show overlap with HOT regions more than is expected by chance.

Enhancers are located far away from the promotor that they regulate and one sub hypothesis that was tested during this project was that ML binding sites could be enhancer related, meaning that they would together with the enhancer loop back to the promotor. Therefore, MB sites are expected to be located nearby the promotor more often than ML. This was tested by measuring how often VDRML overlapped enhancers and VDRMB overlapped promotors (Table 17). These results showed statistically significant results for overlapping regions between enhancers and ML as well as for promotors and MB. Additionally, significant overlap was measured for ML and promotors as well as for MB and enhancers. Next the average distance between ML and enhancers as well as MB and promotes was measured with the *located nearby* tool, and the results can be seen in Table 19. Further, descriptive statistics from *HiBrowse* was used to see if enhancers more often than not would fall inside VDRML and promotors inside VDRMB. A test from *HiBrowse* called *counts* measured this and the results are presented in attachments (Table 32). Finally, the tool *point distance* was used to measure the frequency distribution of distances to nearest promotor or enhancer from the closest MB or promotor (attachments Table 31). However, no supporting evidence of enhancers being associated more with ML or promotors with MB segments was found during this project. On the other hand, there was an indication of VDRMB being located closer to enhancers and promotors than VDRML as can be seen in Table 19. However, this could be the results of there being more ML than MB segments.

Given that VDRML and VDRMB more often than not are found inside HOT regions, it was expected that they would also be located inside DHSs more often than was expected by chance. This was done because these accessible chromatin regions are functionally related to transcriptional activity, as this remodelled state is necessary for the binding of the transcriptional machinery. The results presented in Table 20 indicates that both VDRML and VDRMB more often than is expected by chance locates inside DHSs. These results indicate that if there are interactions between ML and MB these interactions are most likely interactions that occur close to actively transcribed DNA. Further, ML and MB segments are localized in 3D. Additionally, the intersegment distance results show no support for close local interactions. Both these results can be used to support the hypothesis of long range DNA interactions between segments in the TF ChIP-seq datasets. Also possibly even interactions between different chromosomes.

The significant overlap between ML and HOT regions were also discovered by Yip K. 2012 et al [20]. They found using, 100 TFs and whole sets of binding peaks of all TRFs in each cell line as background, that ML binding peaks have very significant overlaps with HOT regions. This was true no matter whether they consider all TRF peaks in the whole genome, or only those in intergenic regions. In all cases, their estimated z-score was more than 25, which corresponds to a P-value < 3×10^{-138}. These results are similar to what was found in this project and indicates that VDRML segments do overlap with HOT regions. On the other hand, we also found that VDRMB significantly overlaps HOT regions as can be seen in Table 22.

To summarise the *overlap* experiments a visual presentation was made and can be seen in Figure 15. Here we show some of the different overlap experiments and how they were predicted to overlap the VDRMB and VDRML datasets. This figure is an interpretation of these results and are based on the simplistic answer returned by *HiBrowse*. As Figure 15 shows both VDRML and VDRMB datasets overlaps ChIA-PET, HOT, DHS, CTCF, Enhancers, Promotors, SMC3 and RAD21 more than expected by chance.

4.5 ML and MB relationship

Continuing to pursue possible local MB-ML interactions the *closest tool* from *Bedtools* was used to detect the closest neighbours in bp distance for either VDRML or VDRMB from either dataset. This was completed two times, first for VDRMB and then for VDRML were both was used as *–a* files detecting the closest segment from both datasets. If there were close local interactions between VDRML and VDRMB this test would reveal that more often than not each VDRML would have a VDRMB as its closest neighbour and vice versa. The same analysis was completed for the closest MB and ML segments from TFs AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1. The results presented in Figure 12 and Figure 13 on the other hand gives no support for this hypothesis. On the other hand, there seem to be the case that ML are located close to ML more often than MB segments. This could indicate that there are local ML-ML interactions. However, it could also be a result of there being significantly more ML than there are MB segments. Additionally, it is possible for a loop to form with for example two MLs on one side and two MBs on the other. Then the analysis would indicate that ML is closest to ML and MB is closest to MB, even if they actually could interact as MB-ML pairs.

For the same TFs the bp distance between all MB and their closest ML segments where measured using the *Closest* tool and showed in Figure 14. A pseudo count was used to correct for overlapping segments that affected the results by increasing all intersegment distances with +1. This was done to prevent log (0) as overlapping segments returned a bp distance of zero. Because all the figures (a-k) seem to show a peak around 1 000 000 bp the pseudo count was expected to have little or no effect on the visual presentation. The results from Figure 14 does not indicate that there is a favourable intersegment distance between MB- ML for the TFs, VDR, AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1 as was initially thought.

It is important to mention that we have not been able to measure any differences for ML or MB segments in regards to 3D colocalization, overlap to cohesin related proteins, overlap with DHSs, overlap with TADs or overlaps with HOT regions. This makes the results from these experiments somewhat inconclusive in regards to what we can infer about the main hypothesis. Another strategy than used here needs to be applied in further investigations in this subject. The tools used in this project are proven suboptimal for finding differences between MB and ML peaks.

5 Conclusion

We have not been able to either prove or reject the original hypothesis. No strong basis was found for the fact that ML cases are the result of non-DBD interactions caused by DNA looping. However, we have identified differences between MB and ML peaks, but what these differences are we do not know. A more detailed approach using more advanced methods is needed to accurately describe the differences between MB and ML peaks.

6 Practical uses and suggestions for future work

The results found could be relevant in future treatment of chronic vitamin D deficiency. ML 3D interactions could be important in vitamin D related gene regulation. The ML peaks seems to also be common in the other TFs examined in this project meaning that ML interactions could be an until now neglected aspect of gene regulation.

For future work I would suggest using FIMO or another method that estimates motif significance for motif detection. Then I would suggest using a Hi-C dataset for estimating possible ML 3D interactions. Additionally, it would be interesting to see if the new tool that is currently under development at UIO called GSuite Hyper Browser could better identify features of ML peaks than HiBrowse.

References

1. Latchman, D.S., *Transcription factors: an overview.* Int J Biochem Cell Biol, 1997. **29**(12): p. 1305-12.
2. Schleif, R., *DNA looping.* Annu Rev Biochem, 1992. **61**: p. 199-223.
3. Smale, S.T. and J.T. Kadonaga, *The RNA polymerase II core promoter.* Annu Rev Biochem, 2003. **72**: p. 449-79.
4. Walhout, A.J., *Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping.* Genome Res, 2006. **16**(12): p. 1445-54.
5. Saiz, L. and J.M. Vilar, *DNA looping: the consequences and its control.* Curr Opin Struct Biol, 2006. **16**(3): p. 344-50.
6. Jeziorska, D.M., K.W. Jordan, and K.W. Vance, *A systems biology approach to understanding cis-regulatory module function.* Semin Cell Dev Biol, 2009. **20**(7): p. 856-62.
7. Sofueva, S. and S. Hadjur, *Cohesin-mediated chromatin interactions--into the third dimension of gene regulation.* Brief Funct Genomics, 2012. **11**(3): p. 205-16.
8. Jin, F., et al., *A high-resolution map of the three-dimensional chromatin interactome in human cells.* Nature, 2013. **503**(7475): p. 290-4.
9. Schmidt, D., et al., *A CTCF-independent role for cohesin in tissue-specific transcription.* Genome Res, 2010. **20**(5): p. 578-88.
10. Panigrahi, A.K., et al., *A cohesin-RAD21 interactome.* Biochem J, 2012. **442**(3): p. 661-70.
11. Ong, C.T. and V.G. Corces, *Enhancer function: new insights into the regulation of tissue-specific gene expression.* Nat Rev Genet, 2011. **12**(4): p. 283-93.
12. Holwerda, S. and W. de Laat, *Chromatin loops, gene positioning, and gene expression.* Front Genet, 2012. **3**(217).
13. Lin, Y.C., et al., *Global changes in the nuclear positioning of genes and intra- and interdomain genomic interactions that orchestrate B cell fate.* Nat Immunol, 2012. **13**(12): p. 1196-204.
14. Sanyal, A., et al., *The long-range interaction landscape of gene promoters.* Nature, 2012. **489**(7414): p. 109-13.

15. Smallwood, A. and B. Ren, *Genome organization and long-range regulation of gene expression by enhancers.* Curr Opin Cell Biol, 2013. **25**(3): p. 387-94.
16. Ji, X., et al., *3D Chromosome Regulatory Landscape of Human Pluripotent Cells.* Cell Stem Cell, 2016. **18**(2): p. 262-75.
17. Dixon, J.R., et al., *Topological domains in mammalian genomes identified by analysis of chromatin interactions.* Nature, 2012. **485**(7398): p. 376-80.
18. Dixon, J.R., et al., *Chromatin architecture reorganization during stem cell differentiation.* Nature, 2015. **518**(7539): p. 331-6.
19. Phillips-Cremins, J.E., et al., *Architectural protein subclasses shape 3D organization of genomes during lineage commitment.* Cell, 2013. **153**(6): p. 1281-95.
20. Yip, K.Y., 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): p. 2012-13.
21. Kvon, E.Z., et al., *HOT regions function as patterned developmental enhancers and have a distinct cis-regulatory signature.* Genes Dev, 2012. **26**(9): p. 908-13.
22. Foley, J.W. and A. Sidow, *Transcription-factor occupancy at HOT regions quantitatively predicts RNA polymerase recruitment in five human cell lines.* BMC Genomics, 2013. **14**(720): p. 1471-2164.
23. de Graaf, C.A. and B. van Steensel, *Chromatin organization: form to function.* Curr Opin Genet Dev, 2013. **23**(2): p. 185-90.
24. de Laat, W. and D. Duboule, *Topology of mammalian developmental enhancers and their regulatory landscapes.* Nature, 2013. **502**(7472): p. 499-506.
25. Baker, M., *Making sense of chromatin states.* Nat Methods, 2011. **8**(9): p. 717-22.
26. Filion, G.J., et al., *Systematic protein location mapping reveals five principal chromatin types in Drosophila cells.* Cell, 2010. **143**(2): p. 212-24.
27. Thurman, R.E., et al., *The accessible chromatin landscape of the human genome.* Nature, 2012. **489**(7414): p. 75-82.
28. John, S., et al., *Genome-scale mapping of DNase I hypersensitivity.* Curr Protoc Mol Biol, 2013. **Chapter 27**: p. Unit 21.27.
29. Johnson, D.S., et al., *Genome-wide mapping of in vivo protein-DNA interactions.* Science, 2007. **316**(5830): p. 1497-502.
30. Jothi, R., et al., *Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data.* Nucleic Acids Res, 2008. **36**(16): p. 5221-31.
31. Dekker, J., et al., *Capturing chromosome conformation.* Science, 2002. **295**(5558): p. 1306-11.
32. Dekker, J., M.A. Marti-Renom, and L.A. Mirny, *Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data.* Nat Rev Genet, 2013. **14**(6): p. 390-403.
33. Li, G., et al., *Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.* Cell, 2012. **148**(1-2): p. 84-98.
34. Matys, V., et al., *TRANSFAC: transcriptional regulation, from patterns to profiles.* Nucleic Acids Res, 2003. **31**(1): p. 374-8.
35. Portales-Casamar, E., et al., *JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles.* Nucleic Acids Res, 2010. **38**(Database issue): p. D105-10.
36. Stormo, G.D., *DNA binding sites: representation and discovery.* Bioinformatics, 2000. **16**(1): p. 16-23.
37. *An integrated encyclopedia of DNA elements in the human genome.* Nature, 2012. **489**(7414): p. 57-74.
38. Klepper, K. and F. Drablos, *MotifLab: a tools and data integration workbench for motif discovery and regulatory sequence analysis.* BMC Bioinformatics, 2013. **14**(9): p. 1471-2105.
39. Lander, E.S., et al., *Initial sequencing and analysis of the human genome.* Nature, 2001. **409**(6822): p. 860-921.
40. Paulsen, J., et al., *HiBrowse: multi-purpose statistical analysis of genome-wide chromatin 3D organization.* Bioinformatics, 2014. **30**(11): p. 1620-2.

41. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics, 2010. **26**(6): p. 841-2.

42. Sandve, G.K., et al., *The Genomic HyperBrowser: inferential genomics at the sequence level.* Genome Biol, 2010. **11**(12): p. 2010-11.

43. Gundersen, S., et al., *Identifying elemental genomic track types and representing them uniformly.* BMC Bioinformatics, 2011. **12**(494): p. 1471-2105.

44. Sandve, G.K., E. Ferkingstad, and S. Nygard, *Sequential Monte Carlo multiple testing.* Bioinformatics, 2011. **27**(23): p. 3235-41.

45. Norman, A.W., *From vitamin D to hormone D: fundamentals of the vitamin D endocrine system essential for good health.* Am J Clin Nutr, 2008. **88**(2): p. 491S-499S.

46. Fetahu, I.S., J. Hobaus, and E. Kallay, *Vitamin D and the epigenome.* Front Physiol, 2014. **5**: p. 164.

47. Ramagopalan, S.V., et al., *A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution.* Genome Res, 2010. **20**.

48. *A user's guide to the encyclopedia of DNA elements (ENCODE).* PLoS Biol, 2011. **9**(4): p. e1001046.

49. Wang, J., et al., *Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors.* Genome Res, 2012. **22**(9): p. 1798-812.

50. Handel, A.E., 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**: p. 163.

51. Grant, C.E., T.L. Bailey, and W.S. Noble, *FIMO: scanning for occurrences of a given motif.* Bioinformatics, 2011. **27**(7): p. 1017-8.

52. Handel, A.E., et al., *Vitamin D receptor ChIP-seq in primary CD4+ cells: relationship to serum 25-hydroxyvitamin D levels and autoimmune disease.* BMC Medicine, 2013. **11**: p. 163-163.

53. Tang, Z., et al., *CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription.* Cell, 2015. **163**(7): p. 1611-27.

Attachments

Multiple tests for overlap as well as colocalization in 3D for MB and ML, ChIP-seq peaks for transcription factors AFT2, BCLAF1, CHD2, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1.

Table 30 ***Overlap*** **and** ***Colocalization*** **analysis for MB and ML datasets, AFT2 BCLAF1, ETS1, IRF3, MAX, MYC, POU2F2, SRF and TAF1.**

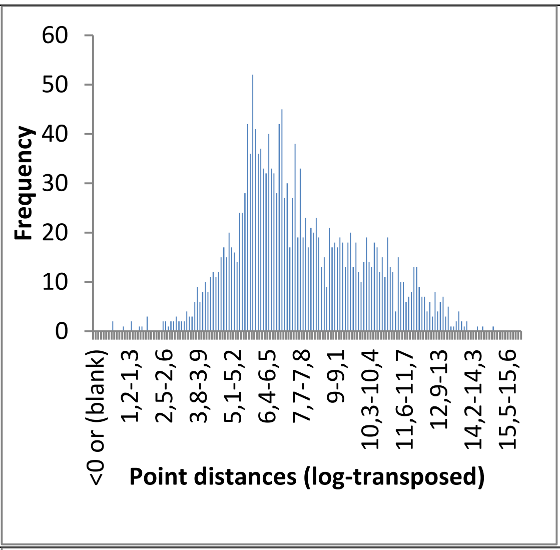
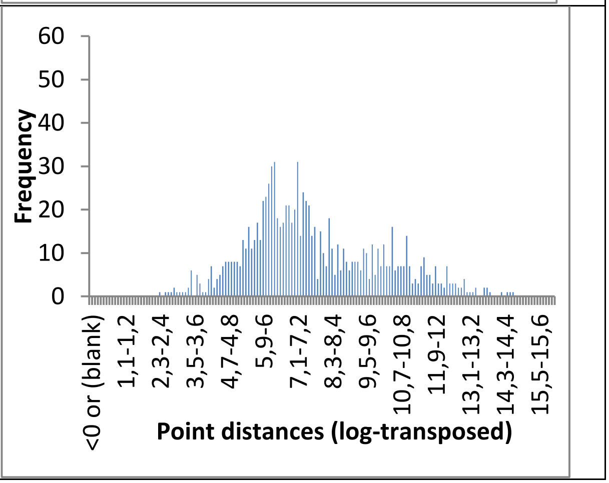
	Overlap?		Colocalization?		
	ML	MB	ML	MB	
ChIA-PET	9.999e-05	9.999e-05	9.999e-04	9.999e-04	AFT2
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	BCLAF1
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	ETS1
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	IRF3

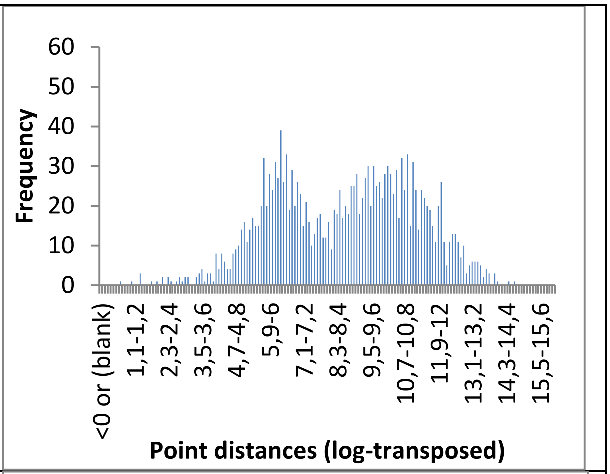
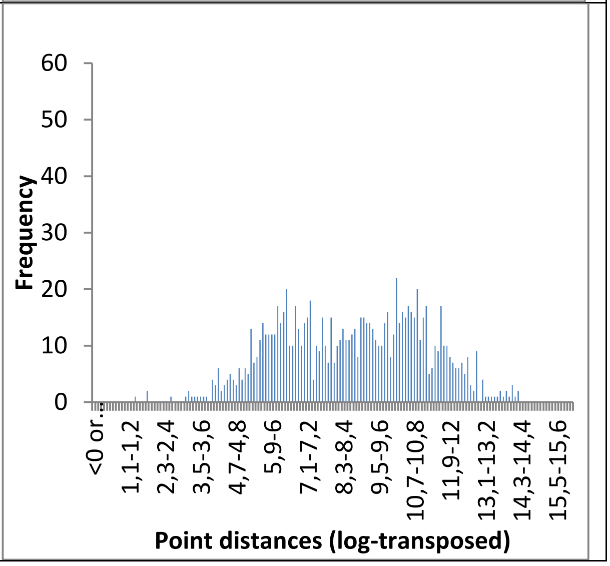
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	MAX
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	MYC
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	CHD2
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	POU2F2
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	SRF
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	TAF1
RAD21	9.999e-05	9.999e-05	9.999e-04	9.999e-04	AFT2
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	BCLA
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	ETS1
	9.999e-05	0.001998	0.001998	9.999e-04	IRF3
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	MAX
	9.999e-05	0.003996	0.003996	9.999e-04	MYC
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	CHD2
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	POU2F2
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	SRF
	9.999e-05	9.999e-05	9.999e-04	0.002997	TAF1
SMC3	9.999e-05	9.999e-05	9.999e-04	9.999e-04	AFT2
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	BCLA
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	ETS1
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	IRF3
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	MAX
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	MYC
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	CHD2
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	POU2F2
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	SRF
	9.999e-05	9.999e-05	9.999e-04	9.999e-04	TAF1
CTCF	9.999e-05	9.999e-05	9.999e-04	0.001998	AFT2
	9.999e-05	9.999e-05	0.01399	0.003996	BCLAF1
	9.999e-05	9.999e-05	0.01598	0.008991	ETS1
	9.999e-05	9.999e-05	0.01798	0.008991	IRF3
	9.999e-05	9.999e-05	0.003996	0.004995	MAX
	9.999e-05	9.999e-05	0.01698	0.01099	MYC
	9.999e-05	9.999e-05	0.001998	0.001998	CHD2

	9.999e-05	9.999e-05	0.002997	9.999e-04	POU2F2
	9.999e-05	9.999e-05	0.002997	0.004995	SRF
	9.999e-05	9.999e-05	9.999e-04	0.006993	TAF1

Point distances test for VDR ML and VDR MB to enhancers and promotors.

Table 31 output from the *HiBrowse* test *Point Distances* between VDRML/ VDRMB and enhancers/promotors

Analysis	Track 1	Track 2	
Point Distances	ML	Strong enhancer	
Point Distances	MB	Strong enhancer	

Point Distances	ML	Active promotor	
Point Distances	MB	Active promotor	

Counts test showing elements of promotor and enhancer tracks falling inside and outside VDR MB and ML.

Table 32 *Counts* inside/outside analysis from *HiBrowse* detecting association between promotors and enhancers against VDR motifbound/less tracks.

Analysis	Track 1	Track 2	Number of track 1	Number of track 2	Number of track 2 elements falling inside track 1	Proportion of Track 2 falling inside track 1
Counts, inside/outside	ML	Weak promotor	1783	35 060	113	0.003223
Counts, inside/outside	MB	Weak promotor	989	35 060	59	0.001683
Count, inside/outside	ML	Active promotor	1783	15 276	253	0.01656
Count, inside/outside	MB	Active promotor	989	15 276	174	0,01139
Counts, inside/outside	ML	Weak enhancer	1783	69 103	171	0.002475

Counts, inside/outside	MB	Weak enhancer	989	69 103	78	0.001129
Count, inside/outside	ML	Strong enhancer	1783	25 486	385	0.01511
Count, inside/outside	MB	Strong enhancer	989	25 486	244	0.009574

Protocol used in *MotifLab2* for ML detection in TFs.

```
DNA = new DNA Sequence Dataset(DataTrack:DNA)
TRANSFAC_Public = new Motif Collection(Collection:TRANSFAC Public)
UseMotifs = new Motif Collection(insert motif codes)
prompt UseMotifs
Cutoff=new Numeric Variable(insert cuttof treshold)
prompt Cutoff
BindingSites = motifScanning in DNA with SimpleScanner [Motif Collection=UseMotifs,Threshold type=”Percentage”,Threshold=Cutoff,Score=”Absolute”]
Motifbound = new Sequence Collection(Statistic:(“region count” in BindingSites)>=1)
Motifless = new Sequence Collection(Statistic:(“region count” in BindingSites)<1)
Output1_motifbound = output Motifbound in BED format [Add CHR prefix=”yes”]
Output2_motifless = output Motifless in BED format
```

Beneath are some outputs from *HiBrowse* obtained in this project that was not included in the project report.

Table 33 Output of *Overlap* analysis from *HiBrowse* between VDRMB and ChIA-PET datasets.

Results	Global analysis
P-value	0.0009990
FDR-adjusted p-values	None
Test statistic: Observed base pair overlap	52 066
Mean of null distribution	3 584.
Median of null distribution	3 524.
Standard deviation of null distribution	1 091.
Difference from mean	4.848e+04
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples with extreme test statistic	0
Number of elements in ‘ChIA-PET ‘	30 711
Number of elements in ‘ChIP-Seq, motifbound’	989
Assembly gap coverage	0.006139

Table 34 *Overlap* analysis from *HiBrowse* between ML and ChIA-PET datasets.

Results	Global analysis
P-value	0.0009990

FDR-adjusted p-values	None
Test statistic: Observed base pair overlap	102 189
Mean of null distribution	4 563.
Median of null distribution	4 538.
Standard deviation of null distribution	1 166.
Difference from mean	9.763e+04
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples with extreme test statistic	0

Table 35 ***3D colocalization*** **test in** ***HiBrowse*** **of VDRMB with ChIA-PET.**

Results	Global analysis
P-value	0.0009990
FDR-adjusted p-values	None
Test statistic: Main result	0.2412
Mean of null distribution	0.08949
Median of null distribution	0.08963
Standard deviation of null distribution	0.008158
Difference from mean	0.1518
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples with extreme test statistic	0
Number of elements in ‘ChIP-Seq Motifless ‘	1 783
Number of elements in ‘ChIP-Seq, motifbound’	989
Assembly gap coverage	0.006139

Table 36 Output of ***Located Inside*** **analysis from** ***HiBrowse*** **of ChIP-Seq ML inside CTCF binding regions.**

Results	Global analysis
P-value	9.999e-05
FDR-adjusted p-values	None
Test statistic: Number of ‘ChIP-Seq Motifless ‘ inside ‘CTCF_binding sites_seq’	216
Mean of null distribution	11.mai
Median of null distribution	11.0
Standard deviation of null distribution	3.329
Difference from mean	204.9
Number of Monte Carlo samples	10 000
Number of Monte Carlo samples	10 000
Number of Monte Carlo samples with extreme test statistic	0
Number of elements in ‘ChIP-Seq Motifless ‘	1 783
Number of elements in ‘CTCF_binding sites_seq’	112 386

Assembly gap coverage	0.006139

Table 37 Output of *Overlap* analysis in *HiBrowse* between RAD21 and VDRML.

Results	Global analysis
P-value	0.0009990
FDR-adjusted p-values	None
Test statistic: Observed base pair overlap	141 894
Mean of null distribution	5 106.
Median of null distribution	4984.0
Standard deviation of null distribution	1 398.
Difference from mean	1.368e+05
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples with extreme test statistic	0
Number of elements in ‘ChIP-Seq, RAD21’	23 945
Number of elements in ‘ChIP-Seq Motifless ‘	1 783
Assembly gap coverage	0.006139

Table 38 Output of *Overlap* analysis in *HiBrowse* between SMC3 and VDRML.

Results	Global analysis
P-value	0.0009990
FDR-adjusted p-values	None
Test statistic: Observed base pair overlap	141 894
Mean of null distribution	5 106.
Median of null distribution	4984.0
Standard deviation of null distribution	1 398.
Difference from mean	1.368e+05
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples	1 000
Number of Monte Carlo samples with extreme test statistic	0
Number of elements in ‘ChIP-Seq, RAD21’	23 945
Number of elements in ‘ChIP-Seq MotifLess ‘	1 783
Assembly gap coverage	0.006139

Beneath are the motifs and scoring matrix obtained from Wang et al 2012.

MYC motif (MM0001)

Table 39 Matrix from Wang et al used for motif scanning by *Simplescanner* on the MYC ChIP-seq dataset [49].

A	C	G	T
0.155388	0.295739	0.401003	0.147870

0.345865	0.295739	0.135338	0.223058
0.092732	0.413534	0.448622	0.045113
0.000000	1.000000	0.000000	0.000000
0.954887	0.000000	0.045113	0.000000
0.000000	0.9724 31	0.000000	0.027569
0.157895	0.000000	0.842105	0.000000
0.077694	0.055138	0.000000	0.867168
0.000000	0.000000	1.000000	0.000000
0.000000	0.057644	0.739348	0.203008
0.107769	0.471178	0.200501	0.220551
0.190476	0.273183	0.182957	0.353383
0.140351	0.305764	0.340852	0.213033
0.127820	0.436090	0.228070	0.208020

POU2F2 MM0002

Table 40 Matrix from Wang et al used for motif scanning by *Simplescanner* on the POU2F2 ChIP-seq dataset [49].

A	C	G	T
0.154450	0.374346	0.183246	0.287958
0.217277	0.162304	0.253927	0.366492
0.256545	0.332461	0.206806	0.204188
0.748691	0.102094	0.073298	0.075916
0.000000	0.002618	0.000000	0.997382
0.146597	0.041885	0.000000	0.811518
0.028796	0.000000	0.000000	0.971204
0.057592	0.000000	0.903141	0.039267
0.026178	0.971204	0.000000	0.002618
0.994764	0.005236	0.000000	0.000000
0.007853	0.000000	0.000000	0.992147
0.583770	0.034031	0.264398	0.117801
0.277487	0.180628	0.120419	0.421466

SRF MM0003

Table 41 Matrix from Wang et al used for motif scanning by *Simplescanner* on the SRF ChIP-seq dataset [49].

A	C	G	T
0.256410	0.282051	0.123077	0.338462
0.074359	0.125641	0.030769	0.769231
0.020513	0.082051	0.256410	0.641026
0.187179	0.169231	0.364103	0.279487
0.000000	1.000000	0.000000	0.000000
0.000000	0.987179	0.000000	0.012821
0.207692	0.035897	0.000000	0.756410
0.066667	0.000000	0.000000	0.933333
0.843590	0.010256	0.038462	0.107692

0.030769	0.010256	0.000000	0.958974
0.579487	0.000000	0.002564	0.417949
0.202564	0.007692	0.007692	0.782051
0.025641	0.000000	0.974359	0.000000
0.000000	0.002564	0.997436	0.000000

CHD2 (UA1) MM0004

Table 42 Matrix from Wang et al used for motif scanning by *Simplescanner* on the CHD2 ChIP-seq dataset [49].

A	C	G	T
0.041045	0.078358	0.029851	0.850746
0.000000	1.000000	0.000000	0.000000
0.996269	0.003731	0.000000	0.000000
0.000000	0.291045	0.003731	0.705224
0.955224	0.003731	0.022388	0.018657
0.007463	0.962687	0.026119	0.003731
0.145522	0.044776	0.026119	0.783582
0.026119	0.003731	0.958955	0.011194
0.041045	0.026119	0.914179	0.018657
0.735075	0.022388	0.093284	0.149254
0.029851	0.014925	0.955224	0.000000
0.958955	0.007463	0.014925	0.018657
0.257463	0.026119	0.690299	0.026119
0.962687	0.014925	0.022388	0.000000
0.910448	0.011194	0.074627	0.003731