Chromosome Wide Prediction of Human Limb Specific Cis-Regulatory Modules Employing Computational and Statistical Approaches

By

Hadia Haider Abbasi

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University

Islamabad, Pakistan

2020

Chromosome Wide Prediction of Human Limb Specific Cis-Regulatory Modules Employing Computational and Statistical Approaches

By

Hadia Haider Abbasi

A thesis submitted in partial fulfilment of the requirements for the degree of

MASTER OF PHILOSOPHY

IN

BIOINFORMATICS

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-i-Azam University

Islamabad, Pakistan

2020

CERTIFICATE

This thesis is submitted by **Hadia Haider Abbasi** from National Center for Bioinformatics, Faculty of Biological Sciences, Quaid-I-Azam University, Islamabad- Pakistan, is accepted in its present form as satisfying the thesis requirements for the Degree of Master of Philosophy in Bioinformatics.

:

Internal Examiner: _____
Dr. Amir Ali Abbasi
Associate Professor & Supervisor
Quaid-I-Azam University, Islamabad

External Examiner: _____

Dr.
Designation
Corresponding Intuitional Address

Chairperson: _____

Dr. Amir Ali Abbasi
Associate Professor
Quaid-I-Azam University, Islamabad

Date: February 07, 2020

DEDICATED

To

Almighty ALLAH and the Holy Prophet Muhammad (P.B.U.H)

&

My parents

ACKNOWLEDGEMENTS

In the name of Allah, the most Gracious and the Most Merciful Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesis. With heads down in esteem and thankfulness for His munificence and generosity, I show all gratitude to our Creator for always leading me with His helping Hand and showering all His blessings, then to His Prophet (P.B.U.H) of whom we are the followers.

I would like to express deepest gratitude to my supervisor Dr. Amir Ali Abbasi, for his guidance, supervision and constant support. His guidance always kept me on course to glide over all obstacles on the way and his ingenuity helped me to put all the right colors on the canvas to portray the exact picture that was primarily quite naive in mind.

Special appreciation goes to my parents for supporting me spiritually throughout my life and their restless efforts always pushed me to achieve my goals, without them nothing was possible and nothing will be.

I would also like to thank my husband and his family for their support and encouragement.

My deepest gratitude goes to my friend Huma Shireen for her guidance, encouragement and support throughout the project.

Special thanks to my friend Washaakh for being there in thick and thin and for all the unconditional support in this very intense academic year.

Thank you very much.

Hadia Haider Abbasi

ABSTRACT

Gene function is defined not only by its product but also by its specific pattern of expression, which needs to be adjusted to a variety of conditions (developmental timing, cell type, environmental factors). A large number of cis-acting elements contribute to this regulation both transcriptionally and post- transcriptionally. Elements that act on gene regulation at the transcriptional level fall in different functional categories, including enhancers, silencers, insulators and other architectural elements. Because of their decisive influence on transcription, enhancers appear as major contributors in regulation of gene expression. Giving their importance for gene expression, considerable efforts have been devoted to identify these cis-regulatory elements. Recruitment of sequence-specific transcription factors (TFs) is an essential hallmark of enhancers. Transcription factors (TFs) are thought to act in a combinatorial way, by competing and collaborating to regulate common target genes. Enhancers possess binding sites for these transcription factors in same combinatorial way and inherit the spatiotemporal specificity too. Due to lack of proper language, mode of action and degeneracy of binding sites, they are very hard to investigate, however different techniques both experimental and computational are devised to understand these cis-regulatory elements. Keeping the transcription factor cooperativity as rationale, a pipeline to predict Human limb specific enhancers was devised. Transcription factors can have one or more binding sites which means that TFBSs are degenerate. Utilizing this we endeavored to characterize limb specific transcription factor code through extensive literature survey and statistical analysis. The devised pipeline effectively predicted putative 844 limb specific enhancers in non-coding, non-repetitive segments of human chromosome 3, 4 and 7 respectively. Predicted enhancers were then validated by means of experimentally verified histone modification marks, DNase hypersensitive sites and limb related clinical variants to assure their significance and reliability. Conservation analysis was also performed for these predicted enhancers which revealed that most of these enhancers are conserved till mammals, while a small number goes down to amphibians and fishes. Cost and time proficiency, literature evidence and validation make this strategy of prediction a comparable substitute of already available computational and experimental approaches used to identify cis-regulatory modules.

Table of Contents

LIST OF ABBREVIATIONS

BED	Browser Extensible Data
Bp	base pair
CRE	Cis Regulatory Element
CRMs	Cis Regulatory Modules
DHS	DNase Hypersensitive sites
ENCODE	Encyclopedia of DNA elements
IUPAC	International Union of Pure and Applied Chemistry
TFBSs	Transcription Factor Binding Sites
GRCH	Genome Reference Consortium Human genome
TF	Transcription Factor
TFBS	Transcription Factor Binding Sites
TFBSMA	Transcription Factor Binding Site Mapping Algorithm
TSS	Transcription Start Site

List of Tables

List of Figures

Chapter 1
INTRODUCTION

1. INTRODUCTION

The limbs are projecting paired appendages of an animal body utilized particularly for movement and grasping, for example, wings, arms, and legs. In the human body, the arms and the legs are generally known as upper limbs and lower limbs respectively, to include part of the shoulder and hip girdles. Arms and legs are connected to torso or trunk. Many animals use limbs for locomotion, such as walking, running, or climbing. Some animals can use their forelimbs (which are homologous to arms in humans) to carry and manipulate objects. Hind limbs can also be used by some animals for manipulation (Conte et al. 2015).

The development of the limb bud is often taken as a paradigm for a cellular and molecular comprehension of the common principles of organogenesis and pattern formation. Limb pattern is specified along three principal axes: anterior-posterior (A-P) (e.g., thumb to little finger), dorsal-ventral (D-V) (e.g., back of hand to palm) and proximal-distal (P-D) (e.g., shoulder to nails). The limb skeleton is laid down as five cartilage skeletal elements, not just the three referred to as stylopod (humerus/femur), zeugopod (radius-ulna/tibia-fibula), and autopod (digits); in fact two carpal regions between zeugopod and autopod are present, that initially have the same size as the other segments but then grow substantially less (Luisa et al. 2015).

The genetic processes that control development of the limb, in both vertebrates and invertebrates, are complicated and are still not fully understood. The temporal and spatial expression of individual genes and gene families has been studied. These studies have identified genes essential for limb bud positioning and initiation. For example, evidence indicates that the precise position at which the fore- and hindlimb buds emerge is controlled by a transcriptional network that includes Hox transcriptional regulators, which function in determining the AP body axis of the embryo. Genetic inactivation of *Hox8* paralogues alters the AP position at which the hindlimb bud emerges from the flank. A small set of other genes, including *Gdf11*, *Tbx3* and *Dicer1* is also involved in limb bud positioning. Formation of the nascent limb bud also requires fibroblast growth factors (FGFs) and the transcription factor TBX5. Identification of many genes critical for limb development and patterning have affirmed the idea that despite of morphological and functional diversification from fish fin to tetrapod limb, vertebrate limb architecture is built upon a fairly similar repertoire of regulatory gene (Barham et al. 2008).

1.1 Gene Regulation

Cells express (transcribe and translate) only a subset of their genes. Cells respond and adapt to environmental signals by turning on or off expression of appropriate genes. Gene expression changes, the driving forces for cellular diversity in multicellular organisms, are regulated by a diverse set of gene regulatory elements that direct transcription in specific cells. The components of regulatory control in the human genome include cis-acting elements that act across immense genomic distances to influence the spatial and temporal distribution of gene expression. Regulatory elements including promoters and enhancers are the primary regulatory components of the genome. They interact with site-specific transcription factors to establish cell type identity and regulate gene expression and are associated with specific, epigenetically established and maintained chromatin features such as histone modifications, DNA methylation and DNA looping. Since all genes including transcription factors themselves are governed by regulatory elements, interactions between transcription factors, co-factors and chromatin regulators with regulatory DNA at specific loci effectively establish molecular regulatory networks. The core functional unit of DNA regulatory elements are transcription factor binding sites (TFBS). These binding sites have preferences for Trans-acting elements which when assemble on the Cis-Regulatory element, initiate and regulate transcription of the associated gene (Doane et al.2017).

1.2 Cis-Regulatory Modules (CRMs)

Temporal and spatial regulation of gene expression is a common process in eukaryotic organisms. Transcription factor-mediated control of gene expression has been studied for decades and involves complex interplays between DNA and proteins. Transcription factors bind to CREs, i.e. short sequences that are usually situated upstream of coding sequences, and affect the set-up of the transcriptional machinery. However, CREs not only function as single elements, they also combine with other CREs. The sum of all CREs that convey specific gene expression is called *cis*-regulatory module (CRM). The gene expression patterns regulated by CRMs are highly dependent on the composition of these CRMs, i.e. the number of repeats of a specific CRE, the combination of CREs present, the spacing between CREs, and the CREs' positions within the CRM (Bekiaris et al. 2018).

Promoters, enhancers, silencers, and insulators are among the key cis-regulatory elements. Containing the transcription start sites (TSSs) of a gene, a promoter functions like a switch to turn on or off the transcription of the target gene. An enhancer (or silencer) can dynamically control the expression level of its target gene(s) through its interaction with promoters, even if they are far

away from their target genes in the linear sequence space. An enhancer may reside in the intergenic region upstream or downstream of its target gene(s), and may also be embedded in an intronic region of a gene. Although distal to its target promoter(s) in linear space, a transcriptionally active enhancer is brought close to its target promoter by DNA looping in 3D nuclear space. Two insulators can establish the boundaries of a regulatory domain within which an enhancer is unable to act beyond the insulator, blocking influence on the genes outside the domain (figure 1) (Li et al. 2015).

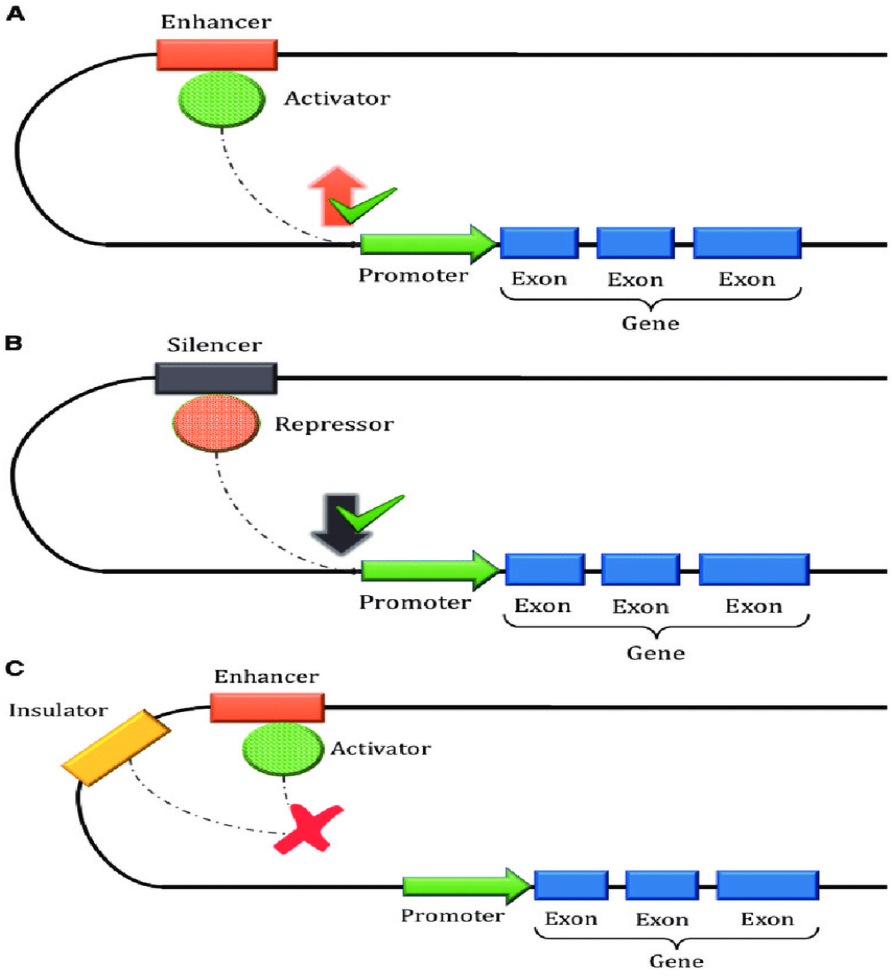

Figure 1.1| *Representation of the effects of cis-regulatory elements: Enhancers (A), the enhancer region binds to a protein (activator) that joins to a specific transcription factor binding site (TFBS) in the promoter region, upregulating the target gene. Silencers (B), the silencer region binds to another protein (repressor) that binds to a specific TFBS in the promoter region, leading to reduced gene expression. Insulators (C), the insulator region interacts with the activator protein of an enhancer, blocking its binding to the promoter and inhibiting gene expression (Rojano et al. 2019).*

CREs play essential roles in determining which genes are specifically active in a cell type, quantitatively controlling the expression levels of these genes at the right times, and confining the regulatory domains of certain functions. Variations in the cis-regulatory regions have been reported to cause assorted abnormal phenotype changes. Thus, identifying and annotating the CREs in the human genome is an important goal for clinical genetics (Li et al. 2015).

1.3 Enhancers and their genomic features

Enhancers are classically defined as TF-binding cis-regulatory regions distal to gene transcription start sites having the ability to increase the transcriptional output from target genes. They are typically of few hundred base pairs length and they serve as platforms to recruit transcription factors through short specific sequences of DNA also known as sequence motifs in order to regulate gene transcription. They are known to be distantly acting elements which are distance and orientation independent in their mode of action (Andersson et al.2015).

Historically, the prediction of enhancers has proved challenging for several reasons. To begin with, enhancers are dispersed over the 98% of the human genome that does not encode proteins, resulting in a large search space (billions of base pairs of DNA). Second, while it is known that they regulate genes in cis, their location relative to their target gene is highly variable; they can either present upstream or downstream of genes but also within introns. Moreover, it is not necessary for an enhancer to act on the respective closest promotor yet can bypass neighboring genes to regulate genes located more distantly along a chromosome. It has also been found that an individual enhancer may regulate multiple genes, adding more complexity to their functional annotation. Also, there is very poorly understood sequence code for enhancers. Lastly, another challenge faced in prediction is that enhancer activity can be restricted to a tissue or cell type, a particular time period in life, or some specific physiological, environmental or pathological conditions (Uemura et al. 2005). As indicated by some ongoing research, the human genome contains a huge number of enhancers and hence it is a significant objective to understand these gene regulatory regions. There are some fundamentally important questions regarding enhancers being raised, for example, prediction of enhancers, mechanism of their action and the manners by which they contribute to evolution and disease (Blow et al. 2010).

1.4 Genome wide efforts to identify enhancers

Most enhancers likely recruit transcription factors to effect tissue-specific expression of their target genes, however known TFBS (transcription factor binding sites) are far too degenerate and are too widely distributed all through the genome to be helpful for enhancer prediction in the absence of additional information. Genome-wide scans for putative regulatory elements rely either upon indirect measures such as evolutionary conservation or, on studies of transcription-factor binding sites (Noonan et al. 2010).

1.5 Analysis of Transcription factors and their binding sites

Generally, four to eight distinct TFs bind within an enhancer, and each factor can bind to multiple sites within it. Copy number, spacing, combination and order of TFBSs greatly effect gene expression. Advancements have been made to predict higher-order site clustering's. Such clusters can be homotypic, containing multiple sites for one particular TF, or heterotypic, containing one or more binding sites for multiple TFs (Vavouri et al. 2005).

Homotypic clustering of TFBSs refers to clustering of multiple TFBSs for the same TF. This kind of clustering can be used to study the effect of binding site strength, orientation and positioning. Homotypic clustering of TFBSs possess the most prevalent organizations of 3-5 closely spaced binding sites allowing the TFs to stabilize each other's binding. Binding sites of homotypic clusters are more conserved then space between the sites.

Heterotypic clustering of TFBSs refers to clustering of multiple TFBSs for different TFs. Enhancers possessing heterotypic clustering are stronger as compared to their homotypic analogs-favors specific TFBS combinations and mimicking putative native enhancers. Exhaustive testing of binding sites permutations suggests that there is a flexibility in binding site order. Whether the gene regulation is heterotypic TFBS cluster (flexible mechanism) or specific patterns of spacing, combination and order of binding sites and TFs are necessary for enhancer function.

1.6 Computational approaches towards identifying enhancers

For post sequencing genome annotation, it is very important to identify transcriptional enhancers and other cis-regulatory modules (CRMs). Computational approaches give a valuable complement to empirical methods for CRM discovery, but it is critical that we develop effective means to

evaluate their performance in terms of estimating their sensitivity and specificity.

Computational approaches generally rely on huge and complex data sets like gene expression profiles, positions of transcription factor binding sites, experimentally verified transcription factor target genes and sequence conservation. Some de novo methods of prediction have employed various algorithms like Hidden Markov Model, expectation maximization, problabistic mixture modeling etc (Halfon et al. 2019).

Sophisticated computational methods have been developed to predict enhancer locations from histone modifications and the majority fit into two categories: discriminative and generative model. The discriminative category is inherently supervised and requires a large training set, usually collected from coactivator binding sites, such as p300. Examples of computational tools in this category are: CSI–ANN, ChromaGenSVM, and RFECS (Bekiaris et al. 2018). (Table1)

Table 1.1| *Brief overview of computational tools used for enhancer identification.*

Category	**Approach**	**Statistical Model**	**Program**
Sequence motif	De novo motifs	Support vector machine	*kmer*-SVM
	TF binding motifs	n/a	EEL- enhancer element locator
Chip-seq	P300/mediator	n/a	Peak profile scanning
DNA methylation	n/a	Support vector machine	SVMmap
		Genomic window based	methylSeekR
Histone modification	Discriminative model	Time delay neural network	CSI-ANN
			ChromaGenSVM
		Support vector machine Random forest	RFECS
	Generative model	Dynamic Bayesian network	Segway
		Hidden markov model	ChromHMM

Table 1.1 *showing multiple computational approaches both from discriminative and generative categories. De novo approaches falls in to sequence motif category that uses SVM model for prediction. Chip-seq and DNA methylation utilizes p300/mediator approach of identifying enhancers. Discriminative models include random forest algorithm, support vector machines and neural networks whereas Hidden markov model belongs to generative category of predicting enhancers.*

In the generative category, multiple methods use hidden Markov model (HMM) or dynamic bayesian network (DBN), including: Chromia, ChromHMM, Segway and ChroModule. Besides histone modification, DNA methylation – the addition of a methyl group to the nucleotide cytosine – is another epigenomic feature that can predict enhancer locations (Li et al. 2015).

1.7 The Search Space

1.7.1 Genomic Features of Human Chromosome 3

Chromosome 3 spans about 198 million base pairs (the building blocks of DNA) and represents approximately 6.5 percent of the total DNA in cells. Chromosome 3 likely contains 1,000 to 1,100 genes that provide instructions for making proteins. These proteins perform a variety of different roles in the body. Developmentally crucial genes like HGD, BTD, CPOX, MLH1, ZNF9, TMIE, BCCB also reside on Chromosome 3. There are several chromosomal aberrations related to chromosome 3 which result into syndromes like Alkaptonuria, Biotinidase deficiency, hereditary coproporphyria, hereditary nonpolyposis colorectal cancer (HNPCC) and type 2 myotonic dystrophy etc.

1.7.2 Genomic Features of Human Chromosome 4

Chromosome 4 spans about 191 million DNA building blocks (base pairs) and represents more than 6 percent of the total DNA in cells. Chromosome 4 likely contains 1200 to 1,500 genes that are very crucial in terms of development and disease. These genes include ANK2, CRMP1, EVC, FGFR3, FGF2, UCHL1, PHOX2B and CXCL1 etc. Gene mutations on chromosome 4 have been linked to genetic disorders and identified in several types of cancer. Examples of conditions associated with gene mutations on chromosome 4 include neurological and neurodegenerative disorders such as Parkinson's disease, Huntington's disease and narcolepsy. Facioscapulohumeral muscular dystrophy is caused by genetic changes involving the long (q) arm of chromosome 4. This condition is characterized by muscle weakness and wasting (atrophy) that worsens slowly over time.

1.7.3 Genomic Features of Human Chromosome 7

Chromosome 7 has almost 159 Mbs large, representing above 5 percent of the total DNA of human cell. Annotation of sequenced genomes is an active area of current research. Chromosome 7 was annotated initially by The Center for Applied Genomics in Toronto. Chromosome 7 likely contains 900 to 1000 genes that are very crucial in terms of development and disease. These genes include

GLI3, TWIST, SHH, CFTR and CDK6. There are several chromosomal aberrations related to chromosome 7 which result into syndromes like Split-hand split-foot syndrome (7q21.3 deletion), Williams-Beuren syndrome and Shwachman-Diamond syndrome. Abnormalities of chromosome 7 are responsible for some cases of Greig cephalopolysyndactyly syndrome, a disorder that affects development of the limbs, head, and face. These chromosomal changes involve a region of the short (p) arm of chromosome 7 that contains the GLI3 gene (Pennacchio et al. 2006).

Owing to disease relevance, limb enriched gene regions, moderate size and some previously performed studies, we opted chromosome 3, 4 and 7 for our study as a test data.

1.8 Aim of Study

The developing limb is one of the best described vertebrate systems for understanding how coordinated gene expression during embryogenesis leads to the structures present in the mature organism. The fact that many essential processes are still largely undescribed. Much of the dynamic transcriptional activity occurring during development is regulated by distal cis-regulatory elements. Modern genomic tools have provided new approaches for studying the function of cis-regulatory elements. The reliable prediction of cis-regulatory elements and, in particular, the prediction of individual transcription factor binding sites has proven to be a very difficult task. Computational biology has been very fruitful in the development of algorithms designed to predict elements that control transcription. In the present study we aim to interpret vocabulary of tissue specific Cis-Regulatory Modules (CRMs) keeping Human Limb Development as model frame work. Subsequent to curating transcription factor code, CRMs would be predicted in a chromosome wide manner. Human chromosome 3, 4 and 7 will be chosen as a search space.

Chapter 2

MATERIALS AND METHODS

2. MATERIAL AND METHODS

As per aim of this research, we intend to identify Human Limb specific enhancers on the basis of transcription factor binding sites residing on entire chromosome. In this manner, the first step was to retrieve the genomic sequences of selected Human chromosomes. Human chromosome 3, 4 and 7 were chosen on premise of their rich gene density, disease relevance and presence of limb related genes.

2.1 Retrieval of data

To predict Human limb specific enhancers, two datasets were required as input as indicated by our research workflow; Transcription factor binding sites and genomic sequences. Various online databases are available which contain transcription factor binding sites for nearly all transcription factors. Computational and statistical techniques such as Bayesian theorem, genetic algorithm, and Hidden Markov Model were used by these repositories to predict transcription factor binding sites. Table 2.1 shows some available online databases which provide binding profiles (TFBS) of transcription factors.

Table 2.1|Overview of Online databases providing insight to Transcription factor binding profiles.

Database	Description
JASPAR	The high-quality transcription factor binding profile database
TRANSFAC	A comprehensive source of eukaryotic gene regulation data
TRED	Transcriptional Regulatory Element Database
TRSDB	A Proteome Database of Transcription Factors
DBD	A database of predicted transcription factors in completely sequenced genomes.
PAZAR	A Database of Transcription Factor and Regulatory Sequence Annotation
ECRBASE	The Database of Evolutionary Conserved Regions (ECRs), Promoters,
TFCAT	A curated catalogue of mouse and human transcription factors
TFCONES	Comparative Genomics of Transcription Factor-Encoding Genes

Table 2.1 showing online repositories possessing TFBSs information. First column contains the abbreviations followed by the details in second column. These repositories contain detailed information about the binding sites for different transcription factors.

2.1.1 Limb specific catalogue of Transcription Factors

A set of transcription factors crucially expressed and specific to human limbs was devised by extensive literature survey. Computational and statistical analysis proved productive in downsizing the set of 20 transcription factors specifically expressed in Human limbs.

2.1.2 Selection of Transcription Factor Binding Sites (TFBS)

Transcription factor binding sites were gathered for each transcription factor. Extensive literature review was performed for the collection of TFBSs. Binding sites retrieved from literature were then overlapped with TRANSFAC and JASPAR results. TRANSFAC is an open source database which provide transcription factor binding sites data validated through experimental investigations. Whereas JASPAR is also the open access repository containing curated and non-redundant TFBS profiles.

2.1.3 Genomic Sequence Collection

Retrieving genomic sequences is the primary requirement of our study. Genomic sequences for Human chromosome 3, 4 and 7 were retrieved from UCSC genome browser particularly for GRCh37/hg19 genome assembly. UCSC genome browser is widely utilized and is one of largest online repository which contains reference sequences and working draft assemblies of respective genomes. UCSC DAS server, particularly aids the user to route the genome annotation data procurable in the genome browser, was used for downloading sequences of chromosome 3, 4 and 7 from UCSC genome browser. FASTA sequence of each chromosome is available in form of compressed .gz files. Knowing that enhancers cannot be present in repeats we downloaded repeat masked (masked repeats are represented by character N) sequences of chromosomes. The downloaded sequences are for the sense or positive strand. To get a copy of negative strand, the sequence data of first strand was translated to complementary strand using a devised Perl script.

2.2 Masking

Enhancers are generally present in non-genic and non-repeat segment of the genome. As such, enhancers mostly reside on intergenic spaces (other than repeats) and intra-genic spaces like introns. To be more specific and targeted towards regions likely to be involved in Enhancers activity and also to reduce our search space, we masked exons (the coding region of genome). Therefore prior to enhancers hunt, we further processed selected chromosomes and masked the exons.

For exon masking, we first fetched the exonic coordinates of genes residing on chromosome 3, 4 and 7 from UCSC table browser. Adjustment of fields was done as per requirement; group was set as Genes and gene predictions for RefSeq genes track. Coordinates were obtained in Bed Extensible data (BED) format creating the records pre exon plus zero bases at each end to include the non-coding exons in the targeted chromosomes.

2.2.1 Bed Processing

Bedtools is a package working on the UNIX command lines and is designed on the basis of set theory. It enables user applying various arithmetic operations like count, sort, intersect, merge, complement and subtract genomic coordinates facilitating the wide scale genome analysis. Bedtools is capable of dealing with different file formats for example BED, BAM, GFF, VCF and GTF.

Exonic regions from both positive and negative strands of targeted chromosomes were masked using bedtools. The input file contained columns composed of chromosome number, Start and End coordinate of respective exon, and strand type. Moving ahead coordinates of positive and negative strand were extracted in two different files.

2.3 TFBS Mapping on Test Data Set

Previously our lab has designed a Transcription Factor Binding Site Mapping Algorithm (TFBSMA). Keeping in mind the complexity of the problem, TFBSMA was programmed using Perl as a programming language. The algorithm works on the principle of string searching paradigm. Data for transcription factors binding sites is multidimensional. One transcription factor

can have multiple transcription factor binding sites and TFBS can degenerate too (Chen et al., 2015). Keeping in mind this multi-dimensional aspect of TFBS, suitable data structure used was hash. A hash of TFBS was created along the hash for nucleotide ambiguity (Table 2.2). Choice of programming language was also done as Perl, on the base of complexity of problem. Program took repeat and exon masked genomic sequence as an input (produced in previous sections), and provided all the occurrences of all the binding sites of transcription factors from limb specific catalogue as an output. Output was generated separately for positive and negative strands of selected chromosomes.

Table 2.2| *Ambiguity code for nucleotides (IUPAC).*

Nucleotide code symbol	Mnemonic	Base
A	Adenine	A
T	thymidine	T
G	guanine	G
C	cytosine	C
W	weak	A or T
S	strong	C or G
Y	pyrimidine	C or T
R	purine	A or G
K	keto	G or T
M	amino	A or C
B	Not A	T or G or C
D	Not C	A or T or G
V	Not T and not U	A or G or C
N	any	A or T or G or C

Table 2.2 constitutes the ambiguity code for nucleotides. First column contains the symbol for each nucleotide. For each nucleotide exists a mnemonic (easily remembered acronym) represented in second column. Third column shows the base against each mnemonic.

2.4 Clustering

Clustering the output from TFBS code was one major step. Each cluster would be a stretch of DNA of specific length that would have multiple TFBS, thus entitling it as an enhancer. Any group of consecutive TFBS having spacer distance between each TFBS less than the defined threshold would become a cluster. Formulated strategy incorporates the Distance based clustering of binding sites. For this reason, a spacer distance between binding site was defined. Binding sites that are falling in between the specified threshold were clustered. This cluster then qualified as an enhancer. The distance selected was performed by optimizing the spacer distance from experimental data from vista enhancer browser (Visel et al.2007). The process of clustering was performed for both positive and negative strand of targeted chromosomes.

2.5 Filtration of clustered output

Further filtration was applied on both positive and negative strands of resulting clusters. The reason was to exclude any sort of redundancy and to get unique clusters. Only those clusters were selected where binding sites for at least 5 out of 6 transcription factors of limb specific catalogue were present. Selected clusters were then exposed to size calculation. Size threshold of 3000bp was defined for further cut down. Enhancers exceeding this defined range were discarded. Enhancers showing both homotypic and heterotypic clustering were kept.

2.6 Concatenating Results

Clusters acquired from positive and negative strands of targeted chromosomes (3, 4 and 7) were combined. Overlapping clusters among positive and negative chromosomal strands were expelled. CRM Ids were assigned to each individual enhancer to resist redundancy in identifiers.

2.7 Validation

To validate the predicted CRMs, previously published Histone modification marks and DNase Hypersensitive sites data (specific to limbs) that is experimentally validated was utilized. Some disease variants specific to limbs were also used for validation.

2.7.1 Overlapping CRMSs with Histone Modification Marks

The H3 acetylation and H3K4me1 signals outside of promoter regions have been correlated with functional enhancers in various cell types. Histone modification data specific to limbs that was previously published was retrieved from ENCODE database (Davis et al., 2018). This data was overlapped with the predicted enhancers using bedtools intersect function.

2.7.2 Overlapping CRMs with DNase Hypersensitive Marks

DNase hypersensitive sites (DHSs) can be considered as markers to identify the *cis*-regulatory elements (enhancers) as they highly correlate with active gene expression (Sullivan et al.2015). DHSs data specific to limbs was fetched from ENCODE project database. The acquired DHSs data was then overlapped with the predicted enhancers using bedtools intersect function.

2.7.3 Overlapping CRMs with clinical variants

Clinical variants specific to limbs were downloaded from NCBI. These variants include limb related disease data that turns out to be very helpful in validating the predicted elements. This data was further overlapped with the predicted CRMs using bedtools intersect command.

2.8 Conservation Analysis of predicted CRMs

After the prediction and validation of limb specific CRMs, their conservation pattern was analyzed using phastcon scoring scheme. For each enhancer, phastcon score was calculated using ftp server. Phastcon scores are actually the probability scores for each base to occur at position. To check the conservation depth of elements, present on targeted chromosomes, phastcon scores were downloaded first from ftp data server. After fetching the scores, a Perl script was designed to

calculate the mean score for mammals, birds, amphibians and fishes. The mean score was set as range for each class in a Perl script. The Perl script took the phastcon scores as input which in output categorized elements and put them in respective classes, hence showing their conservation depths.

2.9 Data Visualization

2.9.1 Heatmap

To make our data more understandable, a heatmap displaying the limb specific transcription factor binding sites residing on the limb enriched genomic region was generated. For this purpose, a genomic region enriched with limb specific genes was retrieved from UCSC genome browser. Targeted transcription factor binding sites were then mapped on the extracted region utilizing a Perl script. Occurrence and cooperativity (both homotypic and heterotypic) of TFBSs was analyzed and the output was represented in the form of Heatmap. R studio software was utilized to generate the heatmap.

2.9.2 Circos Figure

Circos tool was utilized to represent data in circular form. Circos is a tool for visualizing data and information in circular manner allowing the user to analyze the relationship between different dimensional data (Silvain et al. 2014). It helps plotting the data in the form of scatters, line plots, histograms, tiles, connector and text. Circos is basically a command line software that runs the script with .conf extension via commands. The input data was first converted in to a format acceptable by circos. Afterwards, configuration files were generated and input files were linked to .conf files to create circos images.

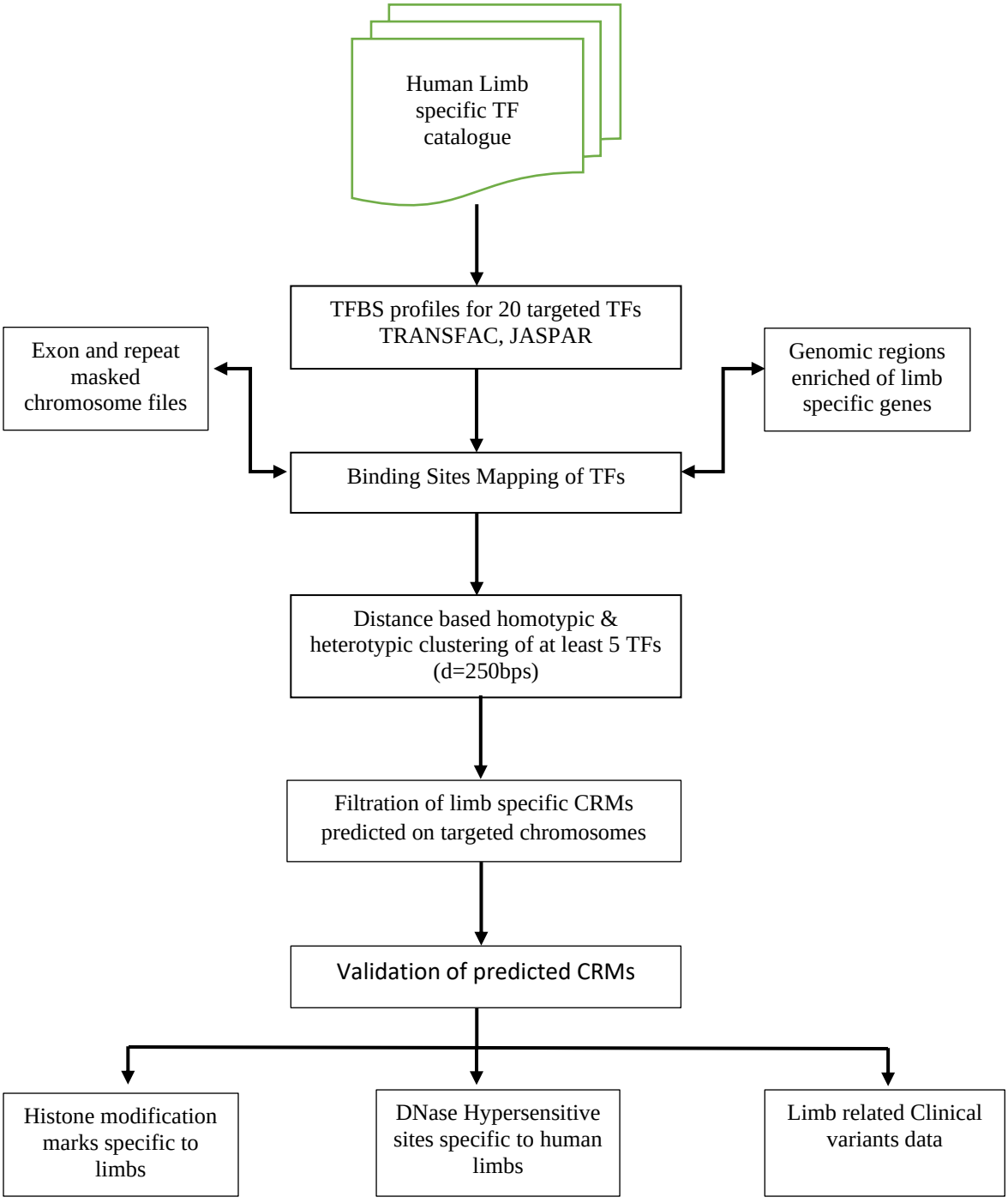

Figure 2.1 showing devised pipeline for prediction of enhancers. In the first step, a TF catalogue specific to human limbs was designed. TFBSs for the selected TFs were retrieved from literature and online available tools i.e. JASPAR and TRANSFAC. Mapping of these binding sites on chromosome files was done further. After mapping, distance-based clustering of mapped sites was performed keeping the distance 250bps. Clusters were then filtered to omit any redundancy. The predicted CRMs were then validated utilizing histone modification marks, DNase hypersensitive sites and limb related clinical variants.

Chapter 3
RESULTS

3. RESULTS

In current study, a catalogue of Transcription factors expressing specifically in Human limbs was formulated. In this manner, extensive literature survey was performed to retrieve functional data that was previously reported. Several computational and statistical approaches were utilized to down size the number of transcription factors. After this analysis, a set of 20 transcription factors was finalized by keeping it under consideration that these TFs are specific to limbs (Table 3.1) (Visel et al. 2007).

Table 3.1| Set of 20 TFs specifically expressing in limbs.

Sr.No	Transcription Factor	Known Expression Tissue
1.	POU3F2	Forelimb, hindlimb
2.	POU6FF1	Limb mesenchyme
3.	YY1	Limb mesenchyme
4.	VMYB	Limb mesenchyme, forelimb, hindlimb
5.	VDR	Forelimb, hindlimb
6.	PLZF	Forelimb, hindlimb, limb mesenchyme
7.	FOXJ2	Forelimb, hindlimb
8.	HNF-4alpha	Forelimb, hindlimb
9.	PBX	Limb, limb mesenchyme
10.	PBX-1	Limb, limb mesenchyme
11.	TBX5	Forelimb bud
12.	SREBP	Forelimb, hindlimb
13.	PAX6	Forelimb, hindlimb
14.	PAX3	Forelimb bud mesenchyme, forelimb
15.	NKX2	Limb mesenchyme, forelimb, hindlimb
16.	MYOGENIN	Forelimb bud mesenchyme
17.	MAZ	Forelimb, hindlimb
18.	MEIS-1	Forelimb bud presumptive ectoderm ridge
19.	CEBP	Limb, limb mesenchyme
20.	CDP	Hindlimb

Table 3.1 demonstrates the catalogue of transcription factors which are specific to limbs. Name of each TF is given in second column. Each of these TFs is specific to limb tissues mentioned in column three.

3.1 TFBS Mapping Outcome

Transcription factor binding sites for selected TF catalogue retrieved from literature and further verified from JASPAR and TRANSFAC were gathered in a hash. TFBS were mapped on targeted chromosomes (3, 4 & 7) using TFBSMA. Genomic sequences of Human chromosome 3, 4 and 7 with exon and repeat masking were then subjected to a Perl script based on TFBSMA. As an output, this script identified the binding sites of selected transcription factors residing on chromosome 3, 4 and 7 along with their positions and cooperativity.

3.1.1 Mapping outcome for Human chromosome 3

The designed Perl script identified the binding sites of selected transcription factors on Human chromosome 3 (HSA 3). Out of whole binding sites catalogue, some of the TFs co-occurred abundantly and cooperatively. On the positive strand of chromosome 3, there were present 249,753 transcription factor binding sites whereas on the negative strand, 199,793 binding sites were mapped. Outcome of transcription factor binding site mapping is shown in figure 3.1.

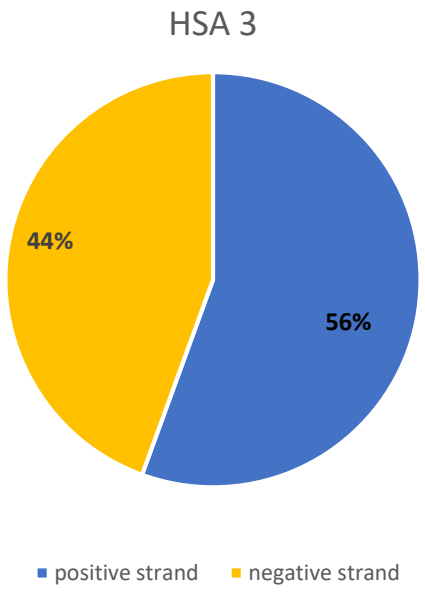

Figure 3.1| *TFs binding sites Mapping on human chromosome 3. On the repeat and exon masked sequence, the pie chart shows that 56% of the total TFBSs occurred on positive strand of HSA 3, while on the negative strand, 44% TFBSs were found.*

3.1.2 Mapping outcome for Human chromosome 4

The Perl script designed previously identified the TFs binding sites on Human chromosome 4 (HSA 4). On the positive strand of chromosome 4, there were present 244,651 transcription factor binding sites whereas on the negative strand, 196,929 binding sites were mapped. Figure 3.2 shows the outcome of transcription factor binding sites on chromosome 4.

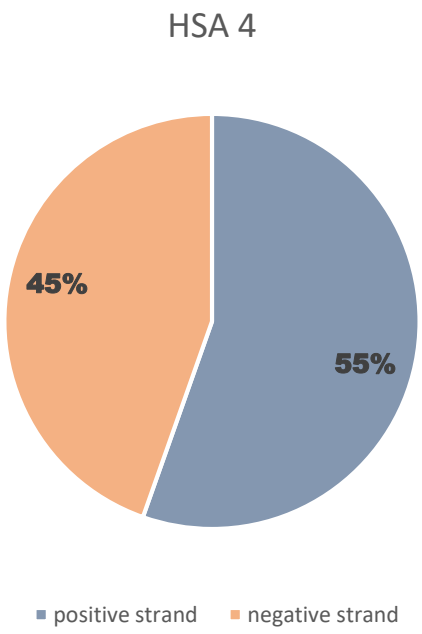

Figure 3.2| *TFs binding site Mapping on human chromosome 4. On the repeat and exon masked sequence, the pie chart shows that 55% of the total TFBSs occurred on positive strand of HSA 4, while on the negative strand, 45% TFBSs were found.*

3.1.3 Mapping outcome for Human chromosome 7

Transcription factor binding sites residing on human chromosome 7 were mapped using previously designed Perl script. There were found 196,007 TFBSs on the positive strand of HSA 7 whereas on the negative strand, 114,950 binding sites were mapped. Figure 3.3 shows the TFBSs outcome for human chromosome 7.

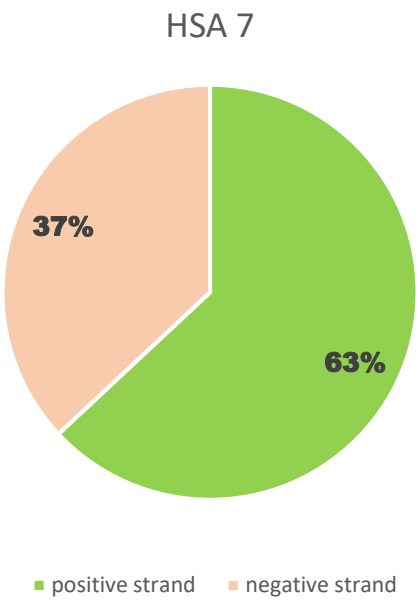

Figure 3.3| *TFs binding site Mapping on human chromosome 7. On the repeat and exon masked sequence, the above chart shows that 63% of the total TFBSs occurred on positive strand of HSA 7, while on the negative strand, 37% TFBSs were found.*

3.2 Predicted clusters from TFBS

As discussed earlier, enhancers are composed of TFBSs clustered homotypically or heterotypically with some spacer distance in between the binding sites. These enhancers have variable transcription factor binding sites present on them. In current pipeline devised, it was assumed that enhancers contain binding sites of at least five to six discrete TFs. These can either be redundant showing homotypic clustering or they can be distinct and non-repetitive showing heterotypic clustering. There exists a spacer distance between two consecutive binding sites. Those TFs which have spacer distance ranging from 0-250 bps and size up to 3000bps were clustered together and filtered through database to optimize the results. Clusters were further processed utilizing bedtools. Each cluster was assigned a unique cluster Id. After applying the filtration strategy on clusters of positive and negative strands of targeted chromosomes (HSA 3, HSA 4 and HSA 7), the number of limb specific enhancers turned out to be 318 on chromosome 3, 300 on chromosome 4 and 226 on chromosome 7 respectively (Figure 3.4).

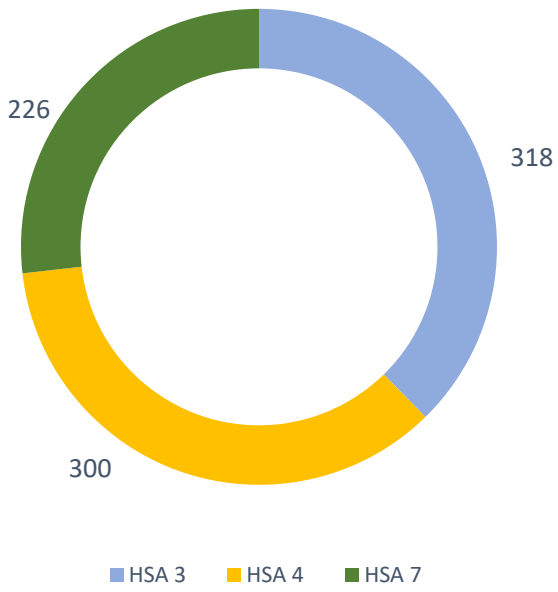

Figure 3.4| Pie chart representing the total number of predicted limb specific enhancers on Human chromosome 3, 4 and 7 respectively. 318 putative enhancers were predicted for chromosome 3(in blue), 300 for chromosome 4(in orange) and 226 enhancers were predicted for chromosome 7(in green) respectively.

3.2.1 Predicted Enhancers on HSA 3

On Human chromosome 3, there were found total 318 enhancers. On the positive strand of chromosome 3, 296 enhancers were present whereas on the negative strand, 22 enhancers were found. To make sure that there is no redundant data, files of both positive and negative strand were overlapped with each other using bedtools. The purpose of overlapping was to assure that no enhancer from the data should be present superfluously. The ones with greater length, maximum cooperativity and heterotypic clustering were retained. There was no overlap found on HSA 3. Table 3.2 shows the predicted limb specific enhancers along with their start/end positions, CRM Ids and size on the positive and negative strands of Human Chromosome 3.

Table 3.2| Predicted Limb specific enhancers on positive and negative strand of chromosome 3 (HSA 3.)

CRM ID	START	END	SIZE (bps)
chr3_crm_1	1411243	1411783	540
chr3_crm_2	12701152	12701744	592
chr3_crm_3	43199164	43199404	240
chr3_crm_4	59321957	59322562	605
chr3_crm_5	77164071	77164427	356
chr3_crm_6	79304554	79305175	621
chr3_crm_7	85662636	85664181	1545
chr3_crm_8	89667330	89667736	406
chr3_crm_9	109107716	109108164	448
chr3_crm_10	124047364	124047919	555
chr3_crm_11	131201396	131201654	258
chr3_crm_12	132810444	132811056	612
chr3_crm_13	140636910	140637326	416
chr3_crm_14	147092959	147093377	418
chr3_crm_15	159634210	159634927	717
chr3_crm_16	161248046	161248534	488
chr3_crm_17	164619253	164620119	866
chr3_crm_18	173178901	173179390	489
chr3_crm_19	173838725	173839617	892
chr3_crm_20	173992949	173993262	313
chr3_crm_21	178936725	178937846	1121
chr3_crm_22	185150527	185151014	487
chr3_crm_23	177540285	177540607	322
chr3_crm_24	88274274	88275168	894
chr3_crm_25	36137973	36138943	970
chr3_crm_26	52929736	52930120	384
chr3_crm_27	102092893	102093666	773
chr3_crm_28	30897565	30898137	572
chr3_crm_29	75182782	75183324	542

CRM ID	START	END	SIZE (bps)
chr3_crm_30	173394042	173394900	858
chr3_crm_31	173568909	173569362	453
chr3_crm_32	19870923	19871118	195
chr3_crm_33	77980128	77980702	574
chr3_crm_34	55501208	55502140	932
chr3_crm_35	100782437	100783744	1307
chr3_crm_36	11806312	11806775	463
chr3_crm_37	172388368	172389080	712
chr3_crm_38	179852101	179852910	809
chr3_crm_39	149786811	149787286	475
chr3_crm_40	165866807	165867104	297
chr3_crm_41	192069861	192070448	587
chr3_crm_42	149557686	149558043	357
chr3_crm_43	1561154	1561779	625
chr3_crm_44	18838350	18839139	789
chr3_crm_45	159028081	159029247	1166
chr3_crm_46	187741542	187742052	510
chr3_crm_47	149682309	149682728	419
chr3_crm_48	69169738	69170995	1257
chr3_crm_49	175762205	175762573	368
chr3_crm_50	89828162	89828749	587
chr3_crm_51	67782928	67784212	1284
chr3_crm_52	75986950	75987589	639
chr3_crm_53	134157151	134157792	641
chr3_crm_54	86643929	86645030	1101
chr3_crm_55	141465662	141466191	529
chr3_crm_56	96015398	96016034	636
chr3_crm_57	62049851	62050242	391
chr3_crm_58	115226303	115227057	754
chr3_crm_59	98631673	98632520	847
chr3_crm_60	104670338	104671271	933

CRM ID	START	END	SIZE (bps)
chr3_crm_61	157800656	157800992	336
chr3_crm_62	133650479	133651080	601
chr3_crm_63	189317118	189317691	573
chr3_crm_64	88207403	88208441	1038
chr3_crm_65	21026324	21027409	1085
chr3_crm_66	31418651	31419267	616
chr3_crm_67	68852210	68852640	430
chr3_crm_68	137599152	137600839	1687
chr3_crm_69	6804025	6804596	571
chr3_crm_70	1581928	1582247	319
chr3_crm_71	6553982	6554575	593
chr3_crm_72	153366891	153367627	736
chr3_crm_73	173150982	173151559	577
chr3_crm_74	85465086	85465505	419
chr3_crm_75	55734964	55735331	367
chr3_crm_76	81383275	81383767	492
chr3_crm_77	84340996	84341359	363
chr3_crm_78	87627011	87627385	374
chr3_crm_79	95224096	95224766	670
chr3_crm_80	48386269	48386725	456
chr3_crm_81	77499099	77499438	339
chr3_crm_82	115872806	115873243	437
chr3_crm_83	131785776	131786338	562
chr3_crm_84	68596548	68597605	1057
chr3_crm_85	23634722	23635326	604
chr3_crm_86	37999286	37999932	646
chr3_crm_87	71065638	71066048	410
chr3_crm_88	149837656	149838139	483
chr3_crm_89	157550616	157551976	1360
chr3_crm_90	192850672	192851229	557
chr3_crm_91	196627013	196627421	408

CRM ID	START	END	SIZE (bps)
chr3_crm_92	75139262	75139951	689
chr3_crm_93	191757150	191758001	851
chr3_crm_94	63772980	63773564	584
chr3_crm_95	177243613	177244079	466
chr3_crm_96	19231254	19231634	380
chr3_crm_97	34656249	34656916	667
chr3_crm_98	64584800	64585236	436
chr3_crm_99	170297588	170298207	619
chr3_crm_100	130085055	130086153	1098
chr3_crm_101	95968103	95968797	694
chr3_crm_102	19669103	19670137	1034
chr3_crm_103	8003440	8004369	929
chr3_crm_104	2413599	2414325	726
chr3_crm_105	157357633	157358454	821
chr3_crm_106	173771177	173771648	471
chr3_crm_107	109819547	109819924	377
chr3_crm_108	157494332	157495526	1194
chr3_crm_109	192436144	192436938	794
chr3_crm_110	154477552	154477975	423
chr3_crm_111	21259544	21260553	1009
chr3_crm_112	76541944	76543027	1083
chr3_crm_113	97295141	97296153	1012
chr3_crm_114	147050058	147050482	424
chr3_crm_115	161723472	161723791	319
chr3_crm_116	78719181	78719886	705
chr3_crm_117	107766500	107767132	632
chr3_crm_118	65020276	65021129	853
chr3_crm_119	167249249	167250048	799
chr3_crm_120	81955272	81956083	811
chr3_crm_121	27367373	27367992	619
chr3_crm_122	34886434	34886950	516

CRM ID	START	END	SIZE (bps)
chr3_crm_123	89096159	89097612	1453
chr3_crm_124	24449502	24450542	1040
chr3_crm_125	75930740	75931116	376
chr3_crm_126	83872804	83873295	491
chr3_crm_127	148446114	148446658	544
chr3_crm_128	159244467	159244989	522
chr3_crm_129	6250310	6250741	431
chr3_crm_130	69558232	69558736	504
chr3_crm_131	81221394	81222243	849
chr3_crm_132	131377244	131378271	1027
chr3_crm_133	69782289	69782611	322
chr3_crm_134	159535783	159536190	407
chr3_crm_135	73624314	73624713	399
chr3_crm_136	81092444	81093238	794
chr3_crm_137	113673554	113674373	819
chr3_crm_138	115522390	115522958	568
chr3_crm_139	144589104	144589945	841
chr3_crm_140	106279036	106279553	517
chr3_crm_141	59396541	59397366	825
chr3_crm_142	61371309	61371839	530
chr3_crm_143	72987474	72988359	885
chr3_crm_144	173214027	173214417	390
chr3_crm_145	134283915	134284488	573
chr3_crm_146	133876641	133877360	719
chr3_crm_147	171157820	171158439	619
chr3_crm_148	34449279	34449749	470
chr3_crm_149	56003537	56004232	695
chr3_crm_150	121517583	121518101	518
chr3_crm_151	144262638	144263299	661
chr3_crm_152	39001083	39001656	573
chr3_crm_153	156135737	156136926	1189

CRM ID	START	END	SIZE (bps)
chr3_crm_154	74543247	74543864	617
chr3_crm_155	94207035	94207914	879
chr3_crm_156	27005352	27006007	655
chr3_crm_157	125006464	125007390	926
chr3_crm_158	102097221	102097752	531
chr3_crm_159	54159943	54161761	1818
chr3_crm_160	65209835	65210460	625
chr3_crm_161	194365374	194365885	511
chr3_crm_162	50709157	50710012	855
chr3_crm_163	120151533	120152039	506
chr3_crm_164	21665378	21666246	868
chr3_crm_165	149981644	149982551	907
chr3_crm_166	106901389	106902078	689
chr3_crm_167	151779418	151780267	849
chr3_crm_168	66045759	66046031	272
chr3_crm_169	160367887	160368435	548
chr3_crm_170	181541678	181542261	583
chr3_crm_171	190163630	190164405	775
chr3_crm_172	154917949	154918373	424
chr3_crm_173	111201350	111202149	799
chr3_crm_174	152073223	152073733	510
chr3_crm_175	102271057	102271693	636
chr3_crm_176	123958014	123958454	440
chr3_crm_177	121729977	121730607	630
chr3_crm_178	85906840	85907506	666
chr3_crm_179	95174332	95174734	402
chr3_crm_180	132942361	132942944	583
chr3_crm_181	168006193	168007097	904
chr3_crm_182	6391733	6392389	656
chr3_crm_183	28080381	28081078	697
chr3_crm_184	113652680	113653095	415

CRM ID	START	END	SIZE (bps)
chr3_crm_185	156497407	156498151	744
chr3_crm_186	38257227	38257674	447
chr3_crm_187	75310722	75311365	643
chr3_crm_188	145451656	145451983	327
chr3_crm_189	61943052	61943539	487
chr3_crm_190	77131395	77131835	440
chr3_crm_191	116092778	116093598	820
chr3_crm_192	144391660	144392185	525
chr3_crm_193	54062790	54063193	403
chr3_crm_194	24837314	24837713	399
chr3_crm_195	190165267	190165588	321
chr3_crm_196	29490980	29491495	515
chr3_crm_197	30875384	30875927	543
chr3_crm_198	55301869	55302500	631
chr3_crm_199	115930684	115931598	914
chr3_crm_200	113622346	113622853	507
chr3_crm_201	29168620	29168988	368
chr3_crm_202	119934836	119935299	463
chr3_crm_203	52938547	52939812	1265
chr3_crm_204	103243234	103243952	718
chr3_crm_205	53794381	53795005	624
chr3_crm_206	71538416	71538665	249
chr3_crm_207	107470369	107470778	409
chr3_crm_208	155176838	155177675	837
chr3_crm_209	190583534	190584000	466
chr3_crm_210	18959698	18960383	685
chr3_crm_211	50416274	50416999	725
chr3_crm_212	106302265	106302845	580
chr3_crm_213	123722233	123723008	775
chr3_crm_214	105257030	105257483	453
chr3_crm_215	177806453	177806888	435

CRM ID	START	END	SIZE (bps)
chr3_crm_216	28145008	28145285	277
chr3_crm_217	177063881	177064640	759
chr3_crm_218	197021726	197022190	464
chr3_crm_219	29408364	29408837	473
chr3_crm_220	145085959	145086846	887
chr3_crm_221	46790491	46790960	469
chr3_crm_222	157929989	157930822	833
chr3_crm_223	160782706	160783053	347
chr3_crm_224	180282546	180282971	425
chr3_crm_225	114142395	114143053	658
chr3_crm_226	178618064	178618387	323
chr3_crm_227	175918120	175919124	1004
chr3_crm_228	152011300	152012013	713
chr3_crm_229	155664991	155665328	337
chr3_crm_230	97017886	97018683	797
chr3_crm_231	106108611	106109077	466
chr3_crm_232	17972029	17972563	534
chr3_crm_233	146830955	146831976	1021
chr3_crm_234	68780917	68781461	544
chr3_crm_235	174911677	174912501	824
chr3_crm_236	94847849	94848363	514
chr3_crm_237	29425893	29426729	836
chr3_crm_238	15518023	15518600	577
chr3_crm_239	56657897	56658620	723
chr3_crm_240	67482599	67483141	542
chr3_crm_241	29977178	29977731	553
chr3_crm_242	50480576	50481152	576
chr3_crm_243	94720546	94721420	874
chr3_crm_244	3175411	3176050	639
chr3_crm_245	6676697	6677243	546
chr3_crm_246	50128488	50129189	701

CRM ID	START	END	SIZE (bps)
chr3_crm_247	114994894	114995923	1029
chr3_crm_248	77134180	77134603	423
chr3_crm_249	70346916	70347388	472
chr3_crm_250	17349679	17350479	800
chr3_crm_251	179253834	179254509	675
chr3_crm_252	56329677	56330128	451
chr3_crm_253	99295583	99296325	742
chr3_crm_254	154054800	154055783	983
chr3_crm_255	17085263	17086306	1043
chr3_crm_256	79267797	79268441	644
chr3_crm_257	42052010	42052469	459
chr3_crm_258	86991111	86991512	401
chr3_crm_259	141884445	141884898	453
chr3_crm_260	119933250	119934055	805
chr3_crm_261	104548512	104549183	671
chr3_crm_262	71614834	71615584	750
chr3_crm_263	97079669	97080245	576
chr3_crm_264	97375493	97376166	673
chr3_crm_265	165732955	165733431	476
chr3_crm_266	76823587	76824879	1292
chr3_crm_267	114954937	114955606	669
chr3_crm_268	3621352	3621880	528
chr3_crm_269	30715985	30716664	679
chr3_crm_270	115027416	115028108	692
chr3_crm_271	102171305	102172449	1144
chr3_crm_272	117177887	117178537	650
chr3_crm_273	188329732	188330274	542
chr3_crm_274	77606262	77606871	609
chr3_crm_275	148422489	148423165	676
chr3_crm_276	179621678	179622179	501
chr3_crm_277	108838138	108839151	1013

CRM ID	START	END	SIZE (bps)
chr3_crm_278	152476852	152477308	456
chr3_crm_279	29283939	29284591	652
chr3_crm_280	84653217	84653704	487
chr3_crm_281	145285397	145286368	971
chr3_crm_282	83321788	83322506	718
chr3_crm_283	151562144	151562479	335
chr3_crm_284	192431223	192431751	528
chr3_crm_285	29953629	29954309	680
chr3_crm_286	44369727	44370599	872
chr3_crm_287	61050452	61051331	879
chr3_crm_288	69962412	69962744	332
chr3_crm_289	76972030	76972646	616
chr3_crm_290	74449523	74449817	294
chr3_crm_291	174364302	174364849	547
chr3_crm_292	197020220	197020570	350
chr3_crm_293	79376810	79377487	677
chr3_crm_294	168623987	168625203	1216
chr3_crm_295	45519763	45520831	1068
chr3_crm_296	117107071	117108193	1122
chr3_crm_297	6699519	6700026	507
chr3_crm_298	150876200	150877085	885
chr3_crm_299	192168532	192169280	748
chr3_crm_300	110495019	110495887	868
chr3_crm_301	175047031	175047866	835
chr3_crm_302	103355331	103355993	662
chr3_crm_303	190845930	190846369	439
chr3_crm_304	59688178	59689529	1351
chr3_crm_305	18445609	18446126	517
chr3_crm_306	36936903	36937212	309
chr3_crm_307	141328489	141329135	646
chr3_crm_308	191144474	191144902	428

CRM ID	START	END	SIZE (bps)
chr3_crm_309	94029437	94029684	247
chr3_crm_310	76892076	76892770	694
chr3_crm_311	85501743	85502941	1198
chr3_crm_312	164297254	164298541	1287
chr3_crm_313	176512613	176512941	328
chr3_crm_314	133282078	133283237	1159
chr3_crm_315	55347651	55348148	497
chr3_crm_316	75330952	75331727	775
chr3_crm_317	118023395	118023874	479
chr3_crm_318	174058049	174058473	424

Table 3.1 comprises of 318 enhancers predicted on positive and negative strand of chromosome 3 (HSA 3). Each enhancer was assigned a unique crm_Id, where crm stands for cis-regulatory module, as exhibited in first column. Start and End coordinate of predicted enhancers is mentioned as START and END in second and third column respectively. These enhancers vary in size depending upon the occurrences of TFBSs within the range of spacer distance of 250bps. Size of each predicted limb specific enhancer is stated in fourth column. Predicted enhancers have at least 5 to 6 distinct TFBSs that are heterotypically clustered.

3.2.2 Predicted Enhancers on HSA 4

The total number of Human limb specific enhancers predicted on Chromosome 4 was 300. Out of these 300 enhancers, 250 were present on positive strand of Chromosome 4 and 50 were found on the negative strand of HSA 4. The enhancers residing on positive and negative strands were then overlapped to check if there is any redundancy in the data. On HSA 4, there were no overlaps found. Enhancers were then assigned unique identifiers to remove any ambiguity. Table 3.3 represent the filtered Cis-regulatory modules for chromosome 4 positive and negative strands respectively.

Table 3.3| Cis-Regulatory modules (CRMs) predicted on Homo sapiens chromosome 4 (HSA 4).

CRM ID	START	END	SIZE (bps)
chr4_crm_1	20646833	20647269	436
chr4_crm_2	20902012	20902376	364
chr4_crm_3	29738076	29739290	1214
chr4_crm_4	30044697	30044967	270
chr4_crm_5	32557942	32558594	652
chr4_crm_6	42591547	42592982	1435
chr4_crm_7	49511404	49511814	410
chr4_crm_8	59678320	59678731	411
chr4_crm_9	64462448	64463082	634
chr4_crm_10	65959042	65959710	668
chr4_crm_11	67561916	67562890	974
chr4_crm_12	73227994	73228529	535
chr4_crm_13	74310497	74310987	490
chr4_crm_14	78939565	78940161	596
chr4_crm_15	82585913	82586412	499
chr4_crm_16	97001808	97002020	212
chr4_crm_17	98504097	98504701	604
chr4_crm_18	106200397	106201083	686
chr4_crm_19	108065180	108065485	305
chr4_crm_20	109364219	109364571	352
chr4_crm_21	111434517	111435052	535
chr4_crm_22	113837644	113838027	383
chr4_crm_23	117557028	117557654	626
chr4_crm_24	117989098	117989501	403
chr4_crm_25	119684285	119684523	238
chr4_crm_26	120070991	120071391	400
chr4_crm_27	120382296	120382624	328
chr4_crm_28	121758256	121759356	1100
chr4_crm_29	123355525	123356363	838

CRM ID	START	END	SIZE (bps)
chr4_crm_30	126004110	126004921	811
chr4_crm_31	136261276	136261595	319
chr4_crm_32	141963271	141963700	429
chr4_crm_33	146579941	146580669	728
chr4_crm_34	149085919	149086378	459
chr4_crm_35	149197922	149198546	624
chr4_crm_36	154215444	154216081	637
chr4_crm_37	156169093	156169501	408
chr4_crm_38	158859633	158860073	440
chr4_crm_39	164586571	164587337	766
chr4_crm_40	165015076	165015502	426
chr4_crm_41	166551082	166551674	592
chr4_crm_42	166748708	166749313	605
chr4_crm_43	166907247	166908650	1403
chr4_crm_44	166969622	166970240	618
chr4_crm_45	169247217	169247664	447
chr4_crm_46	170530603	170530836	233
chr4_crm_47	174838209	174838973	764
chr4_crm_48	177071499	177072443	944
chr4_crm_49	182023035	182023651	616
chr4_crm_50	184912251	184912543	292
chr4_crm_51	61014396	61014873	477
chr4_crm_52	31340176	31340581	405
chr4_crm_53	68370713	68371076	363
chr4_crm_54	54687516	54687865	349
chr4_crm_55	106797260	106798010	750
chr4_crm_56	146074783	146075427	644
chr4_crm_57	26257708	26258349	641
chr4_crm_58	28779104	28779814	710
chr4_crm_59	151549087	151549752	665
chr4_crm_60	186416362	186416620	258

CRM ID	START	END	SIZE (bps)
chr4_crm_61	35329539	35330337	798
chr4_crm_62	94540211	94540950	739
chr4_crm_63	16076718	16076959	241
chr4_crm_64	22444287	22444703	416
chr4_crm_65	93666304	93666671	367
chr4_crm_66	161047672	161048529	857
chr4_crm_67	120290667	120291397	730
chr4_crm_68	120226763	120227071	308
chr4_crm_69	61102209	61103133	924
chr4_crm_70	166688549	166689832	1283
chr4_crm_71	175439656	175440371	715
chr4_crm_72	46646615	46648222	1607
chr4_crm_73	88247379	88248042	663
chr4_crm_74	139399002	139399653	651
chr4_crm_75	136791844	136792160	316
chr4_crm_76	33471418	33472157	739
chr4_crm_77	116592809	116593487	678
chr4_crm_78	136317934	136318984	1050
chr4_crm_79	148950681	148951177	496
chr4_crm_80	174518233	174518811	578
chr4_crm_81	43690186	43690826	640
chr4_crm_82	34077064	34077757	693
chr4_crm_83	37115621	37116490	869
chr4_crm_84	127989797	127990349	552
chr4_crm_85	19384475	19384823	348
chr4_crm_86	144328434	144328840	406
chr4_crm_87	149125905	149126435	530
chr4_crm_88	136162168	136162621	453
chr4_crm_89	42966962	42967672	710
chr4_crm_90	158623422	158624167	745
chr4_crm_91	173069178	173069635	457

CRM ID	START	END	SIZE (bps)
chr4_crm_92	24818378	24818641	263
chr4_crm_93	125634575	125635189	614
chr4_crm_94	151418505	151419598	1093
chr4_crm_95	141175254	141176175	921
chr4_crm_96	163801043	163801829	786
chr4_crm_97	121485297	121485808	511
chr4_crm_98	152061098	152061519	421
chr4_crm_99	180820384	180820734	350
chr4_crm_100	69535530	69536682	1152
chr4_crm_101	112461807	112462395	588
chr4_crm_102	124109827	124110198	371
chr4_crm_103	142735604	142736451	847
chr4_crm_104	19279137	19279890	753
chr4_crm_105	141875697	141876225	528
chr4_crm_106	67958091	67958601	510
chr4_crm_107	102632694	102633519	825
chr4_crm_108	29351426	29352238	812
chr4_crm_109	143924861	143925233	372
chr4_crm_110	167129125	167129600	475
chr4_crm_111	33880370	33880959	589
chr4_crm_112	34976540	34976672	132
chr4_crm_113	180201371	180201914	543
chr4_crm_114	150807463	150807774	311
chr4_crm_115	22646541	22646940	399
chr4_crm_116	134396138	134396712	574
chr4_crm_117	156989368	156989929	561
chr4_crm_118	88617254	88617922	668
chr4_crm_119	162367290	162367693	403
chr4_crm_120	155023547	155024067	520
chr4_crm_121	27932092	27933670	1578
chr4_crm_122	56500014	56500380	366

CRM ID	START	END	SIZE (bps)
chr4_crm_123	12233729	12234924	1195
chr4_crm_124	68862910	68863337	427
chr4_crm_125	158057862	158058510	648
chr4_crm_126	182818617	182819123	506
chr4_crm_127	11816963	11817276	313
chr4_crm_128	39913622	39914704	1082
chr4_crm_129	36853105	36853758	653
chr4_crm_130	45471052	45471628	576
chr4_crm_131	134055183	134056235	1052
chr4_crm_132	154271157	154271784	627
chr4_crm_133	24770379	24770699	320
chr4_crm_134	161437177	161437467	290
chr4_crm_135	23437642	23438417	775
chr4_crm_136	21968008	21968334	326
chr4_crm_137	118190755	118191165	410
chr4_crm_138	172860176	172860509	333
chr4_crm_139	70254868	70255256	388
chr4_crm_140	31879695	31880420	725
chr4_crm_141	93622524	93623157	633
chr4_crm_142	95349858	95350921	1063
chr4_crm_143	96829717	96830872	1155
chr4_crm_144	152885197	152886242	1045
chr4_crm_145	182378403	182379034	631
chr4_crm_146	189371935	189372647	712
chr4_crm_147	57281089	57281814	725
chr4_crm_148	150510530	150511083	553
chr4_crm_149	154807764	154808907	1143
chr4_crm_150	165445898	165446402	504
chr4_crm_151	46739502	46740465	963
chr4_crm_152	76129970	76130573	603
chr4_crm_153	24171488	24171951	463

CRM ID	START	END	SIZE (bps)
chr4_crm_154	182608500	182609021	521
chr4_crm_155	179665500	179666712	1212
chr4_crm_156	87080354	87081445	1091
chr4_crm_157	100504916	100505630	714
chr4_crm_158	41406919	41407832	913
chr4_crm_159	16452433	16452727	294
chr4_crm_160	21896340	21897025	685
chr4_crm_161	48647550	48647873	323
chr4_crm_162	117613712	117614241	529
chr4_crm_163	46957079	46957705	626
chr4_crm_164	19938093	19939020	927
chr4_crm_165	166574862	166575595	733
chr4_crm_166	21193558	21193885	327
chr4_crm_167	44176391	44176802	411
chr4_crm_168	147171574	147172003	429
chr4_crm_169	107849192	107849501	309
chr4_crm_170	133584794	133585663	869
chr4_crm_171	57064899	57065341	442
chr4_crm_172	132872235	132872872	637
chr4_crm_173	39553045	39553247	202
chr4_crm_174	187489990	187490545	555
chr4_crm_175	38049817	38050156	339
chr4_crm_176	27700400	27701077	677
chr4_crm_177	90631715	90632719	1004
chr4_crm_178	89382242	89383022	780
chr4_crm_179	90122182	90122628	446
chr4_crm_180	161399609	161400357	748
chr4_crm_181	78880645	78881459	814
chr4_crm_182	112036224	112036569	345
chr4_crm_183	18557669	18558361	692
chr4_crm_184	94145600	94146256	656

CRM ID	START	END	SIZE (bps)
chr4_crm_185	102136855	102137468	613
chr4_crm_186	146524376	146525238	862
chr4_crm_187	86535834	86536712	878
chr4_crm_188	175372626	175373357	731
chr4_crm_189	117350935	117351553	618
chr4_crm_190	28935152	28935986	834
chr4_crm_191	65028976	65029868	892
chr4_crm_192	61975754	61976489	735
chr4_crm_193	33943169	33944394	1225
chr4_crm_194	135368992	135369918	926
chr4_crm_195	152536732	152537249	517
chr4_crm_196	17869649	17870494	845
chr4_crm_197	86946676	86946905	229
chr4_crm_198	95622358	95623334	976
chr4_crm_199	156822884	156823443	559
chr4_crm_200	28108283	28108892	609
chr4_crm_201	179701020	179701700	680
chr4_crm_202	126914172	126914913	741
chr4_crm_203	141422631	141423122	491
chr4_crm_204	127384475	127385233	758
chr4_crm_205	109367628	109368235	607
chr4_crm_206	121345555	121346293	738
chr4_crm_207	42826633	42827573	940
chr4_crm_208	93849858	93850782	924
chr4_crm_209	77815900	77817142	1242
chr4_crm_210	149038609	149039592	983
chr4_crm_211	70125291	70125604	313
chr4_crm_212	84820780	84821020	240
chr4_crm_213	110837995	110838769	774
chr4_crm_214	27446652	27447510	858
chr4_crm_215	141055005	141055219	214

CRM ID	START	END	SIZE (bps)
chr4_crm_216	151742871	151743530	659
chr4_crm_217	86378076	86378656	580
chr4_crm_218	124871435	124872678	1243
chr4_crm_219	146057851	146058657	806
chr4_crm_220	31134037	31134685	648
chr4_crm_221	10540821	10541519	698
chr4_crm_222	93219869	93220587	718
chr4_crm_223	42613068	42613718	650
chr4_crm_224	79217966	79218360	394
chr4_crm_225	161647166	161647459	293
chr4_crm_226	57461589	57462263	674
chr4_crm_227	61541281	61541781	500
chr4_crm_228	135983733	135984872	1139
chr4_crm_229	68852000	68852615	615
chr4_crm_230	85021690	85021961	271
chr4_crm_231	113662518	113662804	286
chr4_crm_232	26373183	26373849	666
chr4_crm_233	91629779	91630321	542
chr4_crm_234	98066976	98068095	1119
chr4_crm_235	179599720	179600733	1013
chr4_crm_236	21935337	21935982	645
chr4_crm_237	30577059	30577970	911
chr4_crm_238	103022190	103022785	595
chr4_crm_239	63576680	63577318	638
chr4_crm_240	17759850	17760416	566
chr4_crm_241	43365028	43365907	879
chr4_crm_242	70233875	70234341	466
chr4_crm_243	84952787	84953815	1028
chr4_crm_244	116140474	116141058	584
chr4_crm_245	72839352	72839722	370
chr4_crm_246	73041157	73041911	754

CRM ID	START	END	SIZE (bps)
chr4_crm_247	89673986	89674853	867
chr4_crm_248	16592600	16592914	314
chr4_crm_249	33206913	33207704	791
chr4_crm_250	125998984	125999458	474
chr4_crm_251	158917911	158918326	415
chr4_crm_252	189213756	189214026	270
chr4_crm_253	16583795	16584793	998
chr4_crm_254	46931112	46931838	726
chr4_crm_255	112457365	112458255	890
chr4_crm_256	10161093	10162406	1313
chr4_crm_257	97806399	97807323	924
chr4_crm_258	85643603	85643922	319
chr4_crm_259	101337792	101338279	487
chr4_crm_260	53813130	53813499	369
chr4_crm_261	130809233	130810068	835
chr4_crm_262	131601507	131602438	931
chr4_crm_263	167221040	167221761	721
chr4_crm_264	35519825	35520218	393
chr4_crm_265	29801177	29801962	785
chr4_crm_266	143077790	143078160	370
chr4_crm_267	142958159	142958635	476
chr4_crm_268	46245932	46246938	1006
chr4_crm_269	118434021	118435081	1060
chr4_crm_270	85269919	85270401	482
chr4_crm_271	141485113	141485839	726
chr4_crm_272	135140714	135141473	759
chr4_crm_273	182137638	182137863	225
chr4_crm_274	80084633	80085391	758
chr4_crm_275	164513142	164513732	590
chr4_crm_276	27201403	27201813	410
chr4_crm_277	31968029	31968272	243

CRM ID	START	END	SIZE (bps)
chr4_crm_278	31048364	31049092	728
chr4_crm_279	157276097	157276551	454
chr4_crm_280	69201301	69201873	572
chr4_crm_281	137468995	137469343	348
chr4_crm_282	31355035	31355625	590
chr4_crm_283	142949720	142950142	422
chr4_crm_284	162375640	162376401	761
chr4_crm_285	97502614	97503338	724
chr4_crm_286	94338091	94338572	481
chr4_crm_287	67077568	67078121	553
chr4_crm_288	128038580	128039183	603
chr4_crm_289	158575618	158576542	924
chr4_crm_290	124188601	124189122	521
chr4_crm_291	129781331	129781729	398
chr4_crm_292	28904892	28905503	611
chr4_crm_293	62534912	62535224	312
chr4_crm_294	136805925	136807413	1488
chr4_crm_295	144782340	144783129	789
chr4_crm_296	177818150	177819230	1080
chr4_crm_297	176673796	176675000	1204
chr4_crm_298	57888417	57889473	1056
chr4_crm_299	89858901	89859543	642
chr4_crm_300	106856815	106857550	735

Table 2.3 comprises of 300 enhancers predicted on positive and negative strand of chromosome 4 (HSA 4). Each enhancer was assigned a unique crm_Id, where crm stands for cis-regulatory module, as exhibited in first column. Start and End coordinate of predicted enhancers is mentioned as START and END in second and third column respectively. These enhancers vary in size depending upon the occurrences of TFBSs within the range of spacer distance of 250bps. Size of each predicted limb specific enhancer is stated in fourth column. Predicted enhancers have at least 5 to 6 distinct TFBSs that are heterotypically clustered.

3.2.3 Predicted Enhancers on HSA 7

On Human chromosome 7, total number of predicted Cis-Regulatory modules (CRMs) was 226. Positive strand of HSA 7 possesses 193 enhancers whereas 33 were found on the negative strand respectively. Both the strand files were then overlapped to remove repeating elements. One overlap was found for chr7_crm_17. The enhancer with greater length was retained and the other one was removed. The predicted CRMs for Human chromosome 7 were shown in Table 3.4 respectively.

Table 3.4| *Predicted CRMs on positive and negative strand of Homo sapiens chromosome 7 (HSA 7).*

CRM ID	Start	End	Size (bps)
chr7_crm_1	11857551	11858193	642
chr7_crm_2	13054319	13055134	815
chr7_crm_3	15216961	15217301	340
chr7_crm_4	19070177	19070956	779
chr7_crm_5	20421147	20421750	603
chr7_crm_6	22176143	22176633	490
chr7_crm_7	31956595	31957172	577
chr7_crm_8	34987718	34988972	1254
chr7_crm_9	38240111	38240719	608
chr7_crm_10	38868773	38869689	916
chr7_crm_11	39333000	39334232	1232
chr7_crm_12	47273986	47274428	442
chr7_crm_13	79777274	79777745	471
chr7_crm_14	82085996	82086449	453
chr7_crm_15	84935035	84935900	865
chr7_crm_16	88164127	88164727	600
chr7_crm_17	88389446	88390277	831
chr7_crm_18	93161000	93161674	674
chr7_crm_19	94897175	94897504	329
chr7_crm_20	95437334	95437757	423

CRM ID	START	END	SIZE (bps)
chr7_crm_21	1.03E+08	1.03E+08	526
chr7_crm_22	1.07E+08	1.07E+08	372
chr7_crm_23	1.09E+08	1.09E+08	282
chr7_crm_24	1.1E+08	1.1E+08	908
chr7_crm_25	1.12E+08	1.12E+08	449
chr7_crm_26	1.17E+08	1.17E+08	500
chr7_crm_27	1.23E+08	1.23E+08	680
chr7_crm_28	1.25E+08	1.25E+08	287
chr7_crm_29	1.26E+08	1.26E+08	342
chr7_crm_30	1.27E+08	1.27E+08	771
chr7_crm_31	1.36E+08	1.36E+08	471
chr7_crm_32	1.37E+08	1.37E+08	874
chr7_crm_33	1.58E+08	1.58E+08	440
chr7_crm_34	18630162	18630694	532
chr7_crm_35	27554371	27555167	796
chr7_crm_36	11406422	11406854	432
chr7_crm_37	1.05E+08	1.05E+08	128
chr7_crm_38	1.33E+08	1.33E+08	1043
chr7_crm_39	1.07E+08	1.07E+08	856
chr7_crm_40	1.31E+08	1.31E+08	752
chr7_crm_41	1.11E+08	1.11E+08	1348
chr7_crm_42	22465003	22465713	710
chr7_crm_43	26987992	26989116	1124
chr7_crm_44	15656569	15657029	460
chr7_crm_45	9555122	9555763	641
chr7_crm_46	18812789	18813257	468
chr7_crm_47	15877800	15878347	547
chr7_crm_48	26722001	26723113	1112
chr7_crm_49	38983049	38983723	674
chr7_crm_50	89623361	89623901	540
chr7_crm_51	25623624	25624387	763

CRM ID	START	END	SIZE (bps)
chr7_crm_52	85612605	85613224	619
chr7_crm_53	86942337	86942908	571
chr7_crm_54	1.54E+08	1.54E+08	1057
chr7_crm_55	96037888	96038109	221
chr7_crm_56	1.27E+08	1.27E+08	784
chr7_crm_57	1.21E+08	1.21E+08	520
chr7_crm_58	1.54E+08	1.54E+08	801
chr7_crm_59	78839421	78840016	595
chr7_crm_60	1.25E+08	1.25E+08	728
chr7_crm_61	1.52E+08	1.52E+08	872
chr7_crm_62	53191017	53191193	176
chr7_crm_63	23904675	23905304	629
chr7_crm_64	80489894	80490415	521
chr7_crm_65	1.55E+08	1.55E+08	762
chr7_crm_66	40681064	40682504	1440
chr7_crm_67	1.23E+08	1.23E+08	580
chr7_crm_68	83837808	83838582	774
chr7_crm_69	91002815	91003832	1017
chr7_crm_70	41384768	41385637	869
chr7_crm_71	88924492	88924951	459
chr7_crm_72	90247457	90248122	665
chr7_crm_73	52925773	52926314	541
chr7_crm_74	22114565	22115440	875
chr7_crm_75	73121545	73121941	396
chr7_crm_76	19509393	19509939	546
chr7_crm_77	47915172	47915593	421
chr7_crm_78	77740070	77740445	375
chr7_crm_79	50428376	50429201	825
chr7_crm_80	1.11E+08	1.11E+08	276
chr7_crm_81	33245479	33245673	194
chr7_crm_82	1.11E+08	1.11E+08	728

CRM ID	START	END	SIZE (bps)
chr7_crm_83	1.54E+08	1.54E+08	555
chr7_crm_84	88633295	88633935	640
chr7_crm_85	1.27E+08	1.27E+08	339
chr7_crm_86	23502434	23503244	810
chr7_crm_87	3541027	3541821	794
chr7_crm_88	91957352	91958172	820
chr7_crm_89	43248814	43249263	449
chr7_crm_90	84519283	84519730	447
chr7_crm_91	7771881	7773225	1344
chr7_crm_92	1.1E+08	1.1E+08	470
chr7_crm_93	46399664	46400099	435
chr7_crm_94	52872681	52873660	979
chr7_crm_95	13926026	13926507	481
chr7_crm_96	32128342	32129066	724
chr7_crm_97	53007614	53008842	1228
chr7_crm_98	1.03E+08	1.03E+08	770
chr7_crm_99	1.33E+08	1.33E+08	453
chr7_crm_100	1.18E+08	1.18E+08	652
chr7_crm_101	21342602	21343157	555
chr7_crm_102	41226185	41226694	509
chr7_crm_103	1.34E+08	1.34E+08	512
chr7_crm_104	41123735	41124594	859
chr7_crm_105	93249770	93250631	861
chr7_crm_106	28748372	28749180	808
chr7_crm_107	1.29E+08	1.29E+08	589
chr7_crm_108	19070615	19071056	441
chr7_crm_109	1.3E+08	1.3E+08	648
chr7_crm_110	1.2E+08	1.2E+08	992
chr7_crm_111	1.19E+08	1.19E+08	691
chr7_crm_112	1.2E+08	1.2E+08	730
chr7_crm_113	1.15E+08	1.15E+08	552

CRM ID	START	END	SIZE (bps)
chr7_crm_114	3505292	3505624	332
chr7_crm_115	14790815	14791241	426
chr7_crm_116	97008987	97009298	311
chr7_crm_117	16132930	16133452	522
chr7_crm_118	25415249	25416137	888
chr7_crm_119	88617733	88618198	465
chr7_crm_120	8513543	8514351	808
chr7_crm_121	18159223	18159716	493
chr7_crm_122	90082055	90082789	734
chr7_crm_123	1.22E+08	1.22E+08	1104
chr7_crm_124	37769853	37770208	355
chr7_crm_125	90922331	90922844	513
chr7_crm_126	1.47E+08	1.47E+08	1797
chr7_crm_127	1.23E+08	1.23E+08	273
chr7_crm_128	62856282	62857112	830
chr7_crm_129	21738996	21739164	168
chr7_crm_130	43523972	43524973	1001
chr7_crm_131	27022004	27022887	883
chr7_crm_132	80551985	80552484	499
chr7_crm_133	9155363	9156101	738
chr7_crm_134	8751391	8751734	343
chr7_crm_135	1.33E+08	1.33E+08	492
chr7_crm_136	71542714	71543340	626
chr7_crm_137	64149842	64150366	524
chr7_crm_138	84083771	84085454	1683
chr7_crm_139	96325313	96326351	1038
chr7_crm_140	1.04E+08	1.04E+08	731
chr7_crm_141	1.54E+08	1.54E+08	970
chr7_crm_142	25332282	25333224	942
chr7_crm_143	1.23E+08	1.23E+08	765
chr7_crm_144	1.27E+08	1.27E+08	609

CRM ID	START	END	SIZE (bps)
chr7_crm_145	1.52E+08	1.52E+08	733
chr7_crm_146	1.53E+08	1.53E+08	481
chr7_crm_147	1.07E+08	1.07E+08	451
chr7_crm_148	1.11E+08	1.11E+08	1342
chr7_crm_149	1.37E+08	1.37E+08	720
chr7_crm_150	1.2E+08	1.2E+08	669
chr7_crm_151	88438767	88439302	535
chr7_crm_152	92920014	92920982	968
chr7_crm_153	1.22E+08	1.22E+08	599
chr7_crm_154	27501483	27502611	1128
chr7_crm_155	1.16E+08	1.16E+08	796
chr7_crm_156	1.34E+08	1.34E+08	811
chr7_crm_157	1.2E+08	1.2E+08	784
chr7_crm_158	26410520	26411082	562
chr7_crm_159	84423333	84423966	633
chr7_crm_160	1.04E+08	1.04E+08	543
chr7_crm_161	25633449	25634350	901
chr7_crm_162	2817766	2818315	549
chr7_crm_163	35531151	35531591	440
chr7_crm_164	38624504	38625087	583
chr7_crm_165	92459104	92460629	1525
chr7_crm_166	1.44E+08	1.44E+08	935
chr7_crm_167	9117446	9118063	617
chr7_crm_168	37478883	37479508	625
chr7_crm_169	1.21E+08	1.21E+08	515
chr7_crm_170	1.58E+08	1.58E+08	197
chr7_crm_171	52854027	52854442	415
chr7_crm_172	50220908	50221374	466
chr7_crm_173	1.15E+08	1.15E+08	706
chr7_crm_174	26475923	26476709	786
chr7_crm_175	93072490	93073658	1168

CRM ID	START	END	SIZE (bps)
chr7_crm_176	1.2E+08	1.2E+08	646
chr7_crm_177	71642013	71642375	362
chr7_crm_178	53602233	53602677	444
chr7_crm_179	95262432	95262989	557
chr7_crm_180	70039975	70040908	933
chr7_crm_181	19349622	19350609	987
chr7_crm_182	66264553	66265179	626
chr7_crm_183	1.24E+08	1.24E+08	1078
chr7_crm_184	11455857	11456199	342
chr7_crm_185	1.28E+08	1.28E+08	798
chr7_crm_186	3448140	3449015	875
chr7_crm_187	28673631	28673928	297
chr7_crm_188	36935941	36936357	416
chr7_crm_189	17080803	17081081	278
chr7_crm_190	69268027	69268735	708
chr7_crm_191	1.08E+08	1.08E+08	692
chr7_crm_192	1.14E+08	1.14E+08	810
chr7_crm_193	31447126	31447828	702
chr7_crm_194	1.34E+08	1.34E+08	589
chr7_crm_195	1.04E+08	1.04E+08	454
chr7_crm_196	1.36E+08	1.36E+08	686
chr7_crm_197	54537045	54537481	436
chr7_crm_198	94932128	94932628	500
chr7_crm_199	82924616	82925391	775
chr7_crm_200	1.34E+08	1.34E+08	310
chr7_crm_201	1.44E+08	1.44E+08	648
chr7_crm_202	1.55E+08	1.55E+08	343
chr7_crm_203	33060817	33061263	446
chr7_crm_204	1.35E+08	1.35E+08	597
chr7_crm_205	1.21E+08	1.21E+08	205
chr7_crm_206	1.54E+08	1.54E+08	706

CRM ID	START	END	SIZE (bps)
chr7_crm_207	12479071	12479454	383
chr7_crm_208	83584225	83584580	355
chr7_crm_209	80814885	80815690	805
chr7_crm_210	13954518	13955357	839
chr7_crm_211	1.11E+08	1.11E+08	758
chr7_crm_212	8603607	8604202	595
chr7_crm_213	88462774	88463057	283
chr7_crm_214	63357244	63357383	139
chr7_crm_215	1.21E+08	1.21E+08	853
chr7_crm_216	30326820	30327583	763
chr7_crm_217	38424542	38425291	749
chr7_crm_218	16110387	16110835	448
chr7_crm_219	1.47E+08	1.47E+08	665
chr7_crm_220	76484222	76484646	424
chr7_crm_221	97880249	97880586	337
chr7_crm_222	1.39E+08	1.39E+08	506
chr7_crm_223	27797236	27797544	308
chr7_crm_224	21115257	21116217	960
chr7_crm_225	11496814	11497304	490
chr7_crm_226	25425026	25425874	848

Table 3.4 comprises of 226 enhancers predicted on positive and negative strand of chromosome 7 (HSA 7). Each enhancer was assigned a unique crm_Id, where crm stands for cis-regulatory module, as exhibited in first column. Start and End coordinate of predicted enhancers is mentioned as START and END in second and third column respectively. These enhancers vary in size depending upon the occurrences of TFBSs within the range of spacer distance of 250bps. Size of each predicted limb specific enhancer is stated in fourth column. Predicted enhancers have at least 5 to 6 distinct TFBSs that are heterotypically clustered.

3.3 Validation of predicted limb specific enhancers

The prediction Cis-Regulatory modules that are specific to limbs were validated through already published histone modification marks, DNase hypersensitive sites and clinical disease variants specific to limbs.

3.3.1 Validation through Histone modification marks

For the validation of currently predicted limb specific enhancers, previously published histone modification marks data was utilized. Predicted CRMs were overlapped with the histone modification data and common marks were generated by applying arithmetic operations. Targeted human chromosome 3, 4 and 7 shows the presence of histone marks on them.

3.3.1.1 Outcome for chromosome 3

For human chromosome 3, there were found total 412 overlaps out of which 385 were present on positive strand whereas 27 were found on the negative strand of human chromosome 3 (Figure 3.5).

3.3.1.2 Outcome for chromosome 4

There were found total 399 overlaps of Histone modification marks on chromosome 4. Out of these 399, positive strand of HSA 4 possesses 334 overlaps whereas on the negative strand, 65 marks were found (Figure 3.5).

3.3.1.3 Outcome for chromosome 7

For chromosome 7, total 283 overlaps for histone modification marks were found. Out of these 283, positive strand contains 240 histone marks while 43 were found on negative strand of HSA 7 (Figure 3.5).

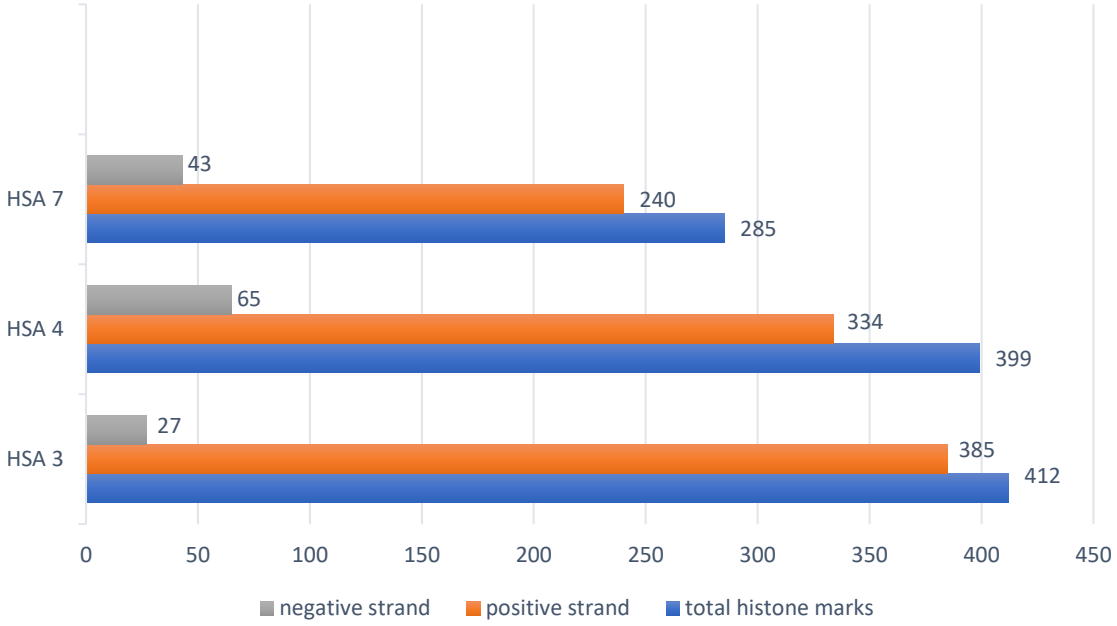

Figure 3.5| Bars representing the outcome of Histone modification marks found on Homo sapiens chromosome 3, 4 and 7 respectively. Blue bars showing total no. of overlaps found on each chromosome, orange showing marks found on positive strand and grey bars showing histone marks on negative strand of all three chromosomes respectively.

3.3.2 Validation through DNase Hypersensitive sites

Predicted CRMs were further validated using DNase hypersensitive sites data of different embryonic stages specific to limbs. All the three targeted chromosomes show overlaps with collected data of DHSs respectively.

3.3.2.1 Outcome for chromosome 3

For Human chromosome 3, 2925 overlaps of DNase hypersensitive sites were generated in total. Out of these, 1826 were found on positive strand whereas on the negative strand of HSA 3, 1099 DHSs were found (Figure 3.6).

3.3.2.2 Outcome for chromosome 4

For Human chromosome 4, there were found total 2132 overlaps of DHSs respectively. Out of this total, 1490 overlaps were mapped on the positive strand and 642 overlaps of DNase hypersensitive sites were present on the negative chromosome of HSA 4 (Figure 3.6).

3.3.2.3 Outcome for chromosome 7

Total number of DNase hypersensitive sites mapped on human chromosome 7 was 2528. Out of this total, 1623 sites were mapped on the positive strand and remaining 905 were found on the negative strand of HSA 7 respectively (Figure 3.6).

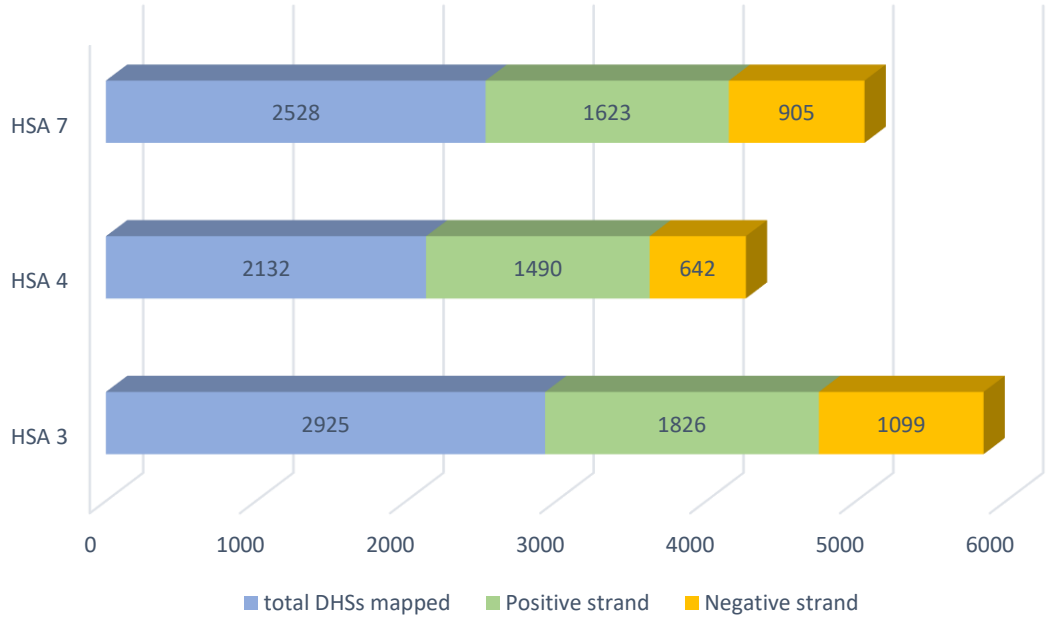

Figure 3.6| Bars representing the mapped DNase hypersensitive sites on Homo sapiens chromosome 3, 4 and 7 respectively. Blue bars showing total number of DHSs mapped, green representing the sites found on positive strands of all three chromosomes and orange displaying the number of DHSs overlapped on negative strand of HSA 3, 4 and 7 respectively.

3.3.3 Validation through Disease variants

The predicted CRMs were further validated by overlapping them with the clinical variants collected from already published data. All the three targeted chromosomes were overlapped with the variant data that is specific to limbs utilizing bedtool commands. Out of three Human chromosomes that were under study, only two showed disease relevance i.e. HSA 3 and HSA 4. For HSA 3, there were found 7 overlapping variants while for HSA 4, there were only 2 matches found. There was no overlap found for chromosome 7. The overlapped variants include limb related diseases such as Hereditary hyperekplexia and Robinow syndrome, a genetic disorder characterized by short limb dwarfism.

Table 3.5| Overlapping disease variants on Homo sapiens chromosome 3 and 4 respectively.

Chromosome	Start	End	CRM ID
HSA3	55501241	55501243	chr3_crm_34
HSA3	55501308	55501310	chr3_crm_34
HSA3	55501464	55501466	chr3_crm_34
HSA3	55501634	55501636	chr3_crm_34
HSA3	55502066	55502068	chr3_crm_34
HSA3	55502108	55502110	chr3_crm_34
HSA3	55502109	55502111	chr3_crm_34
HSA4	158058023	158058025	chr4_crm_125
HSA4	158058042	158058044	chr4_crm_125

Table 3.4 representing the overlapping disease variants for chromosome 3, 4 and 7 respectively. Name of the chromosome is mentioned in the first column (where HSA stands for human sex autosomes). Start and End coordinates of the overlapped enhancers are provided in the second and third column. CRM Id (cis-regulatory module) of each overlapped enhancer is given in the fourth column. Total 7 overlaps were found on chromosome 3. All of the seven variants overlapped with CRM 34. On Chromosome 4, there were two overlaps found each on the same enhancer i.e. CRM 125 respectively.

3.4 Conservation Analysis of predicted Enhancers

Conservation pattern of all the predicted enhancers was performed using phastcon scores retrieved via ftp server. The predicted enhancers were categorized by means of designed Perl script and hence put them in to respective classes i.e. mammals, birds, amphibians and fishes. Many of the predicted CRMs were found to be conserved till mammals while few of the elements goes down till amphibians and fishes. 78% of the total CRMS predicted were conserved till mammals, 19.4% till birds and 0.03% were found to be conserved till amphibians and fishes (Figure 3.7).

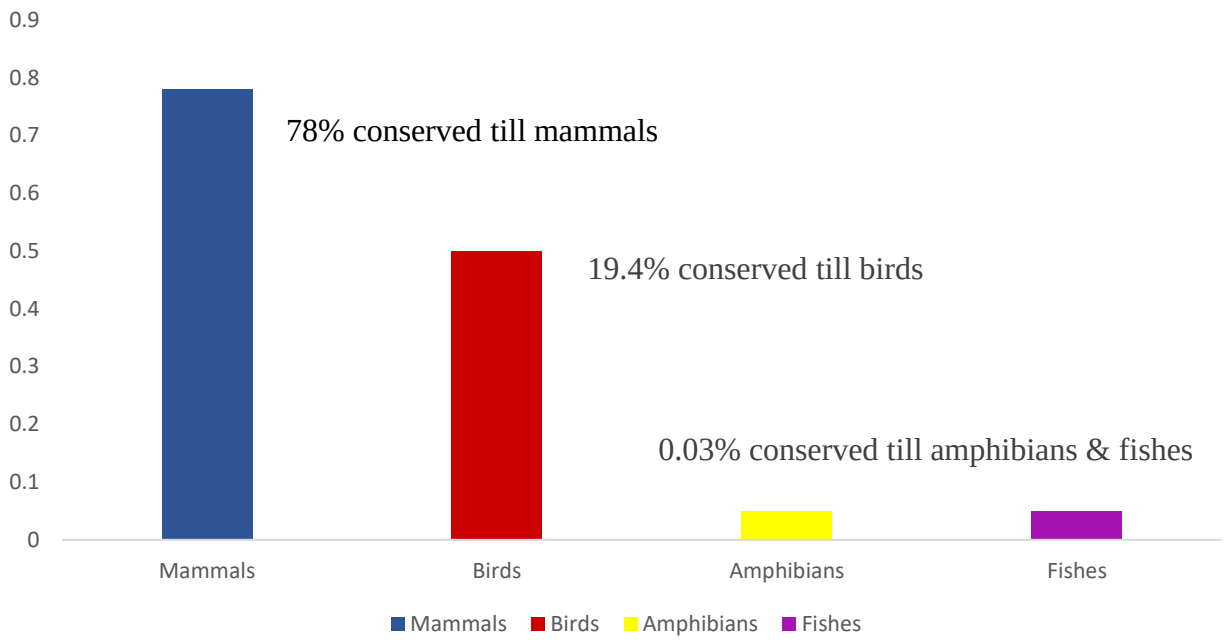

Figure 3.7| Bar chart displaying the conservation pattern governed by predicted Cis-Regulatory modules (CRMs). Blue bar showing that 78% CRMS are conserved till mammals, red representing that 19.4% enhancers are conserved till birds, and only 0.03 % of the total predicted CRMs are conserved till amphibians and fishes as shown above respectively.

3.5 Data Visualization

3.5.1 Heatmap

To better understand the occurrences and cooperativity of transcription factor binding sites, a heatmap was generated that displays the homotypic and heterotypic clustering pattern of TFBSs. Limb enriched genomic region was utilized to map the binding sites of targeted transcription factors. Heatmap generated through R studio software can be seen in figure 3.8.

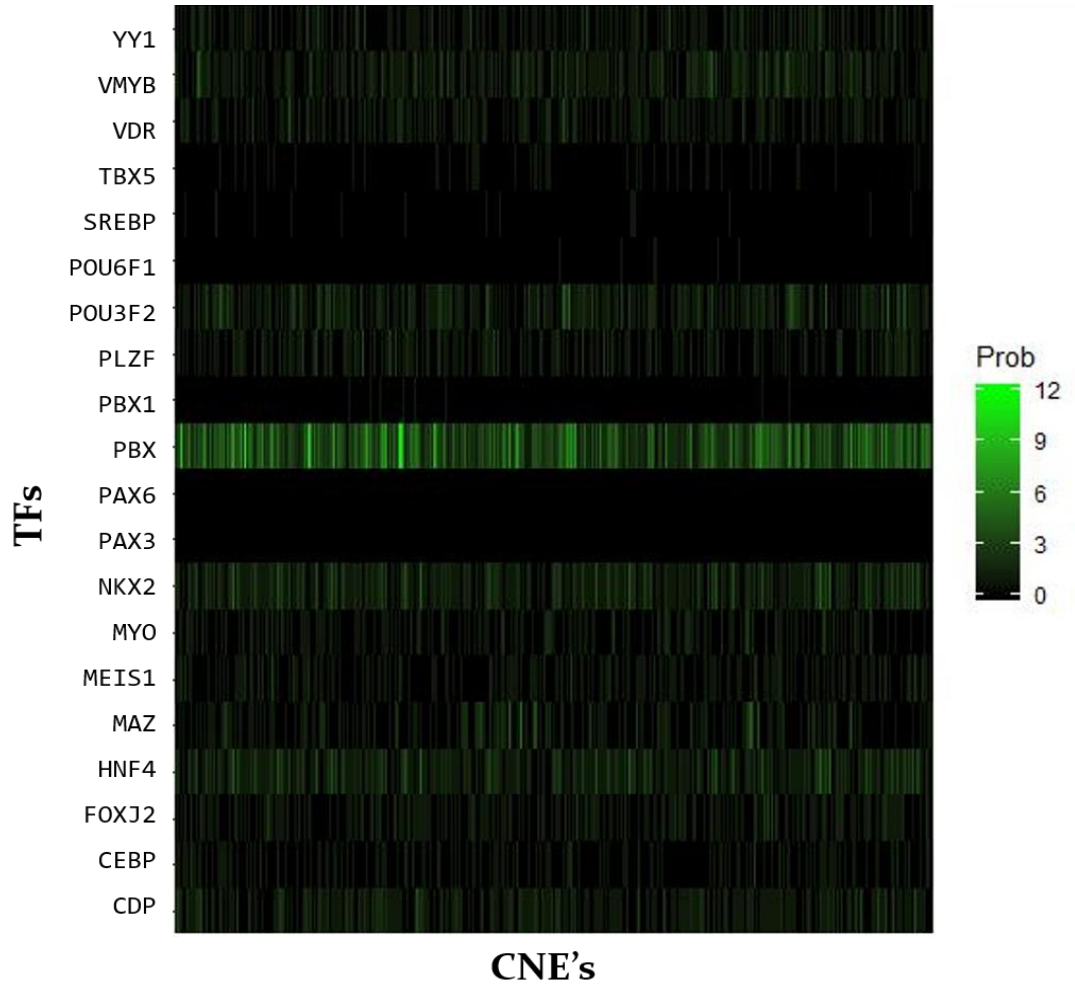

Figure 3.8| Heatmap displaying the transcription factor binding sites (TFBSs) of the 20 Transcription factors (TFs) that were under study with a probability scale (Prob) ranging from 0-12. TFBSs of selected TFs were mapped on a genomic region enriched with limb specific genes. It can be seen that transcription factor PBX is showing maximum homotypic clustering among all. Strong heterotypic clustering pattern can be observed between POU3F2, VDR, PLZF, PBX, NKX2 and HNF4 respectively

3.5.2 Circos Figures

The predicted CRMs on all the three chromosomes (HSA 3, HSA 4 and HSA 7) along with their conservation analysis and validation results were represented in a circular form using Circos data visualization software. Circos figure for chromosome 3 showing the overall summary of predicted limb specific enhancers, overlaps found from validation data and conservation analysis performed, can be seen below (Figure 3.9).

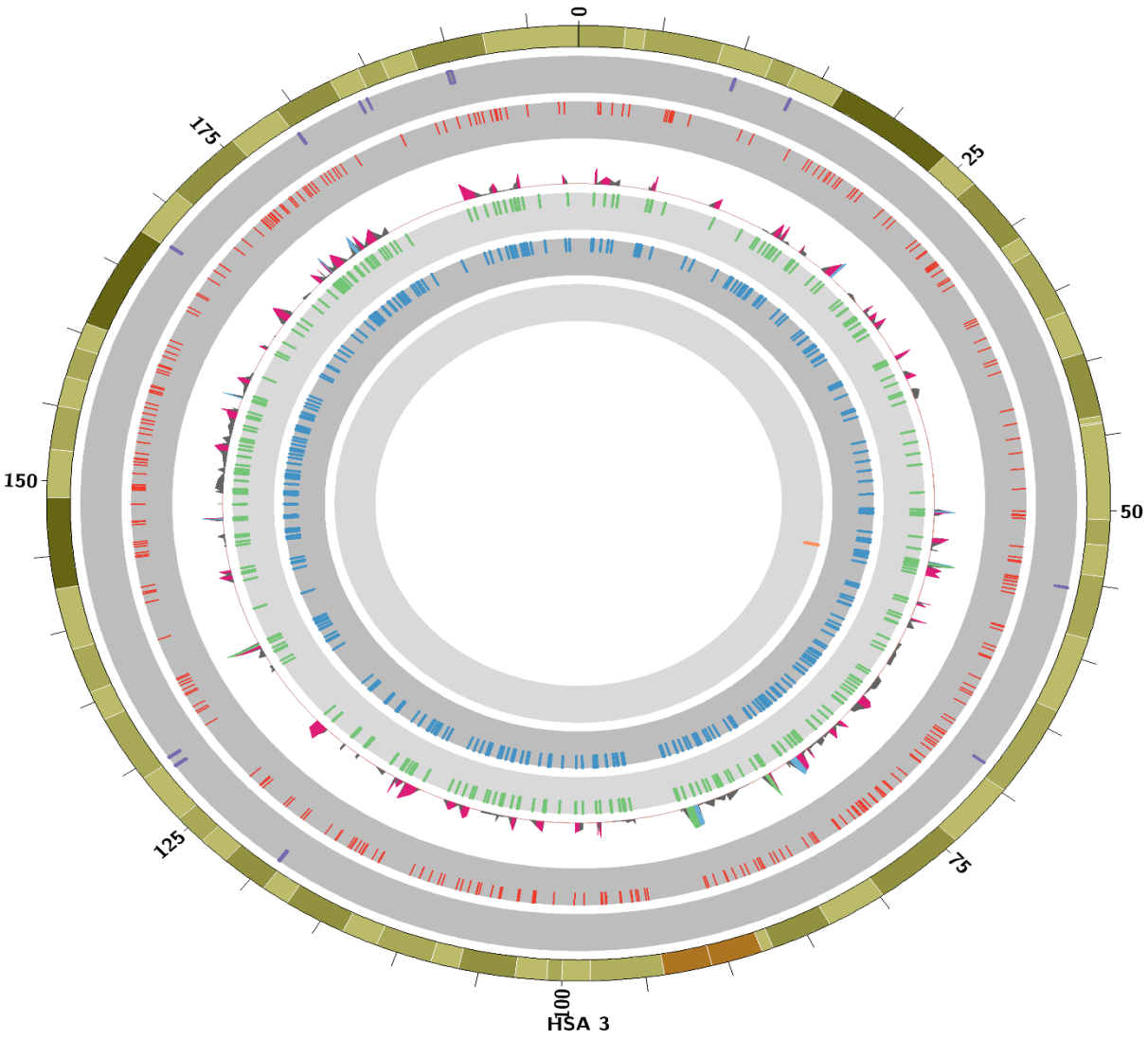

Figure 3.9| Human Chromosome 3 (HSA; human sex autosomes) depiction with genomic regions having limb enriched genes (purple), predicted CRMs (red), conservation pattern (third circle with multi colors representing mammals, birds, amphibians and fishes), histone modification marks (green), hypersensitive sites (blue) and disease variants (orange). The outer boundary displaying labelled numbers represents the circumference of the circle to equally distribute the figure.

Circos figure for chromosome 4 demonstrating the overall outcome of predicted limb specific enhancers, overlaps found from validation data and conservation analysis performed, can be seen below (Figure 3.10).

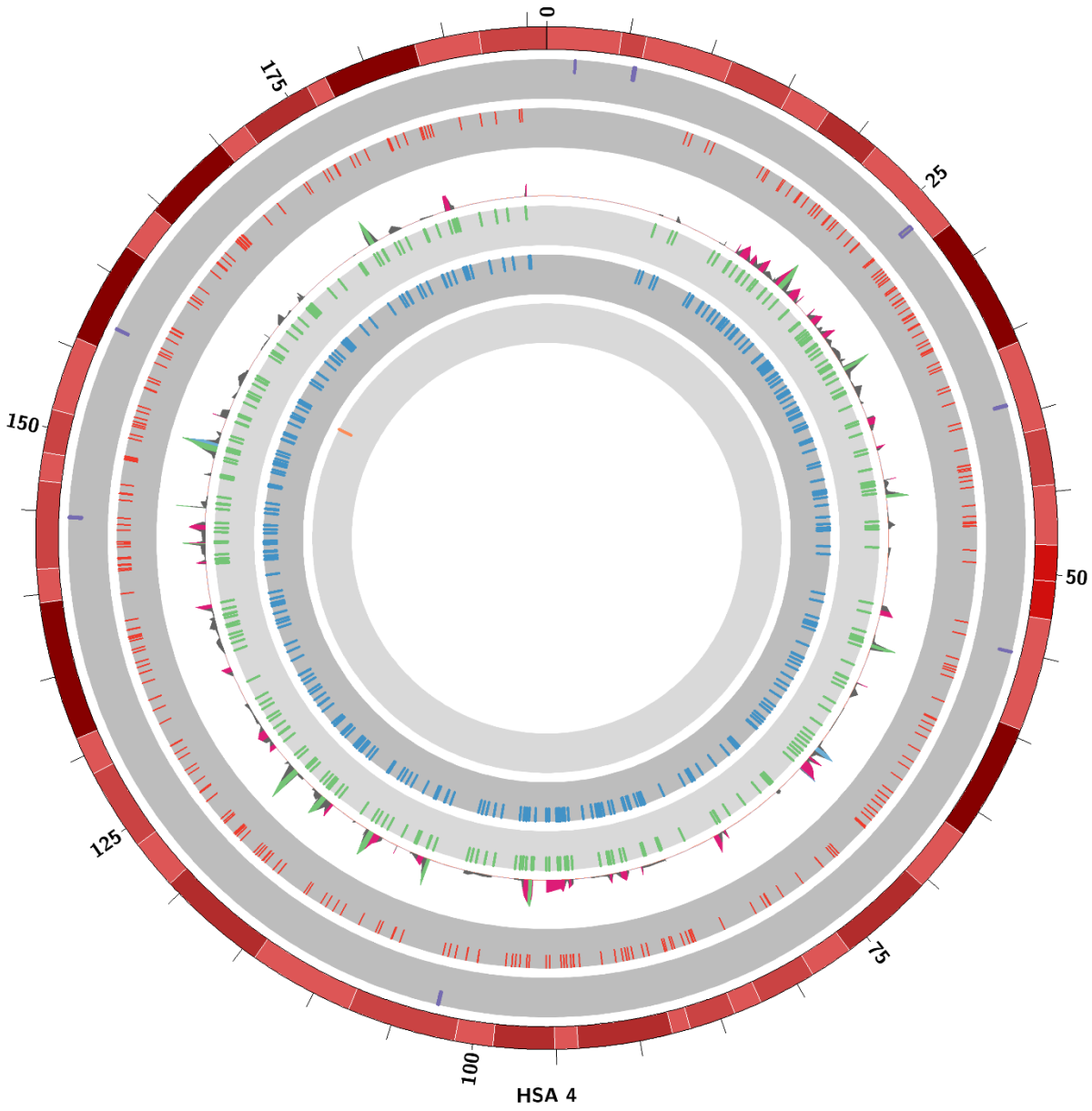

Figure 3.10| Human chromosome 4 (HSA; human sex autosomes) view with overlapping genomic regions in purple, predicted enhancers in red, conservation depth analysis in third circle with different colors, histone modification marks in green, DHSs in blue, and overlapped disease variant in orange. The outer boundary with labelled numbers shows the circumference of the circle to evenly divide the figure in to parts.

Circos figure for chromosome 7 representing the output of followed strategy i.e. predited limb specific enhancers, overlaps found from validation data and conservation analysis performed, can be seen below (Figure 3.11).

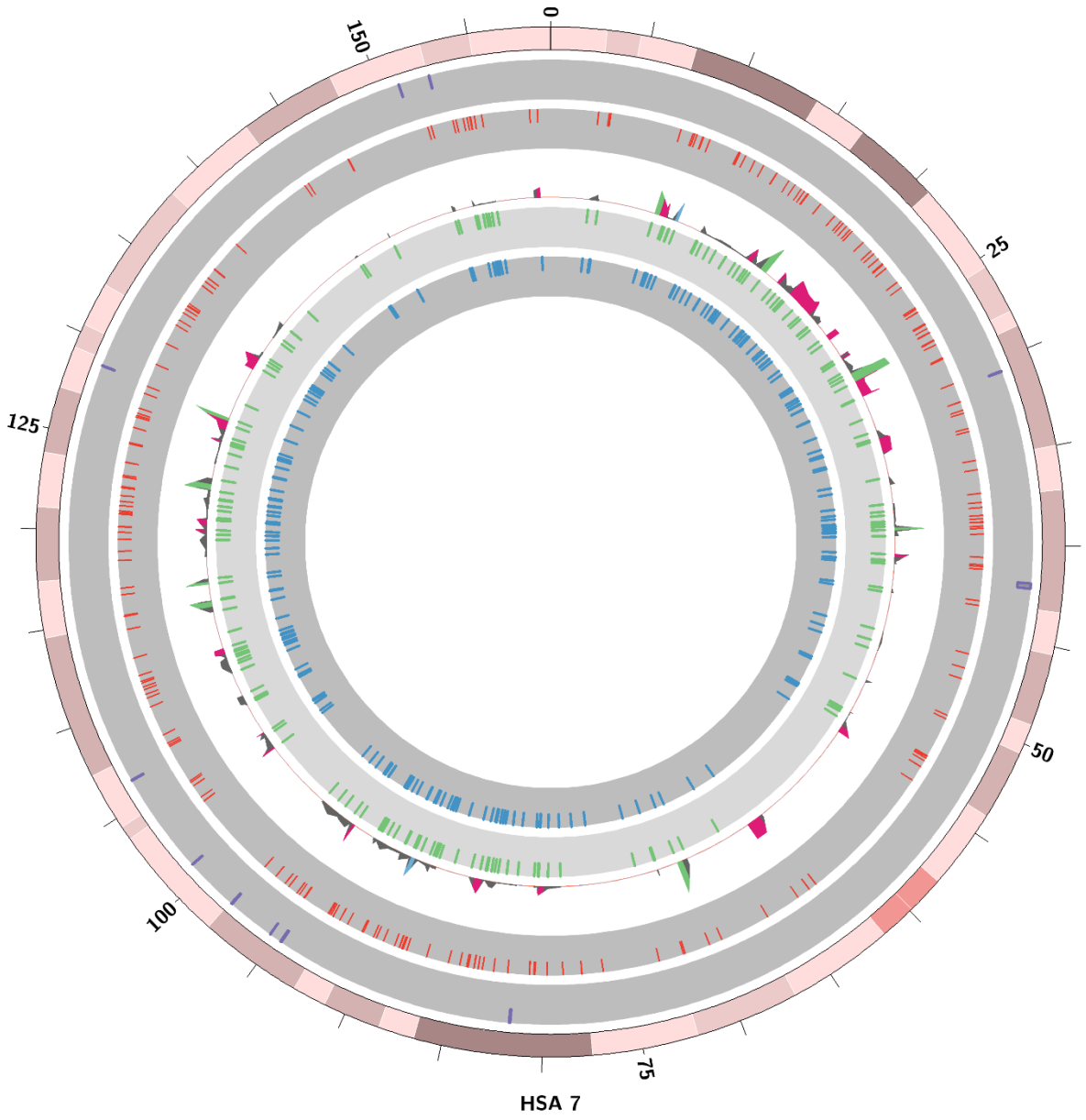

Figure 3.11| *Upshot of Human chromosome 7 (HSA; human sex autosomes). First circle with purple bars showing limb rich gene regions, second circle with red bars showing predicted enhancers, third circle with zigzag pattern showing conservation analysis, fourth circle displaying overlapped histone marks, DHSs overlapped with predicted CRMs can be seen in fifth circle with blue bars respectively. The numbers labelled outside the circle represents the circumference of circle to evenly divide the figure into parts.*

Chapter 4
DISCUSSION

4. DISCUSSION

With the advancement of various computational techniques and biological information available, it has now become easy to identify coding sequences relatively precisely and efficiently from primary DNA sequence information. On the other hand, it is as yet hard to recognize and predict non-coding sequences that incorporate critical cis-regulatory modules (CRMs) such as promotors, enhancers, locus control regions etc. These cis-acting elements are under great consideration because of their regulatory aspects. (Saurabh et al.2009).

Among all cis-regulatory modules (CRMs), enhancers play a critical role in modulating and fine-tuning gene expression in a cell type and developmentally-regulated manner. Due to their involvement in regulation, enhancers are under great emphasis for investigation but due to lack of proper known language, degeneracy of short binding sites, their mode of action, also they are very hard to investigate (Pennacchio et al.2013). Various experimental and computational approaches are devised for the prediction of enhancers each with its own pros and cons.

Enhancers, based on their subversive conservation pattern, get identified by utilizing comparative genomics approach. Since this strategy depends on ultra-conservation of enhancers, it misses the specie specific enhancers which can be more fundamental in the development and specificity of species (Leung et al.2009). Recently, various experimental techniques such as Chip-Seq have been developed to overcome the shortcomings of already known methods by introducing conservation independent platform. Chip-Seq misses one of the most important aspect of enhancers that is spatio-temporal behavior i.e. enhancers only get activated when required. Another con of this technique is its high cost and tissue specificity which makes it less utilizable. These restrictions demonstrate computational strategies and pipelines as imperative substitute to predict and annotate enhancers more efficiently.

Keeping the transcription factors cooperativity as rationale, a pipeline to predict enhancers is devised. Transcription factors can have more than one binding sites. This combinatorial binding help in identifying enhancer marks. Keeping this in mind, a set of transcription factors specific to limbs is retrieved through extensive literature review and statistical methods. Average size of human limb enhancer ranges between 500-4000bps possessing at least 5-8 distinct transcription factor binding sites residing unevenly on the enhancer region. The distance between the transcription factor binding sites matters for the proper binding of TFs (Ezer et al.2014). The

devised pipeline includes distance-based clustering of transcription factors, so the spacer distance is optimized to 250bps for each element. For this purpose, experimentally verified enhancers were utilized retrieved from VISTA enhancer browser. Average spacer distance between the transcription factor binding was calculated through statistical analysis and this distance turns out to be 250bps respectively.

Homo sapiens chromosome 3, 4 and 7 (HSA 3, HSA 4 and HSA 7) were selected as our search space on premise of their rich gene density, disease relevance and presence of limb related genes. As enhancers are absent in repeat regions, we reduced our search space by masking repeats. To be more targeted towards enhancers, exonic regions were also masked. Transcriptional regulation of gene is controlled by combinatorial manner of transcription factors because TFs can hardly bind to enhancers as single unit. These TFs interact as clusters of localized domains. To predict such clusters on targeted chromosomes, each set of transcription factor binding sites was gathered in to a cluster. Many clusters were found this way but to increase the sensitivity of our pipeline, we removed the clusters having less than 5 binding sites from our data. Each enhancer predicted can be defined as a cluster of transcription factor binding sites with at least 5 distinct transcription factors that are specifically linked to limbs. A CRM Id was given to each predicted element as identifier.

Afterwards, the predicted enhancers were validated via already published data of histone modification marks, DNase hypersensitive sites and limb related clinical variants to assure the significance of predicted enhancers. Enhancers have histone modification marks on them (Calo et al. 2013). To validate the presence of histone marks on predicted CRMs, we overlapped the already published data of these marks with our predicted enhancers. Similarly, DHSs are also considered as markers to identify CRMs. Therefore, predicted CREs were also overlapped with limb specific DHSs data. Limb related disease variants fetched from literature were also utilized to validate our results. This validation strategy makes our pipeline more authentic and reliable to predict enhancers.

Finally, we report a set of 844 limb specific enhancers predicted on three targeted human chromosomes i.e. HSA 3, HSA 4 and HSA 7. Out of these 844 enhancers, 318 resides on chromosome 3, 300 on chromosome 4 and 226 enhancers found on chromosome 7 respectively.

Conclusively, the pipeline devised in this study provides a computational platform assisted by literature and experimental proofs, which can be utilized to identify the Cis-Regulatory elements i.e. enhancers. Transcription factor binding sites cooperativity was under consideration in this study to analyze the tissue specific transcriptional regulatory network and gene expression pattern in considerably less time and cost when contrasted with different methods accessible today.

Chapter 5
REFERENCES

5. REFERENCES

Andersson, Robin, Albin Sandelin, and Charles G. Danko. 2015. "A Unified Architecture of Transcriptional Regulatory Elements." *Trends in Genetics* 31(8): 426–33.

Asma, Hasiba, and Marc S. Halfon. 2019. "Computational Enhancer Prediction: Evaluation and Improvements." *BMC Bioinformatics* 20(1): 174.

Barham, Guy, and Nicholas M. P. Clarke. 2008. "Genetic Regulation of Embryological Limb Development with Relation to Congenital Limb Deformity in Humans." *Journal of Children's Orthopaedics* 2(1): 1–9.

Bekiaris, Pavlos Stephanos, Tobias Tekath, Dorothee Staiger, and Selahattin Danisman. 2018. "Computational Exploration of Cis-Regulatory Modules in Rhythmic Expression Data Using the 'Exploration of Distinctive CREs and CRMs' (EDCC) and 'CRM Network Generator' (CNG) Programs." *PLOS ONE* 13(1): e0190421.

Blow, Matthew J. et al. 2010. "ChIP-Seq Identification of Weakly Conserved Heart Enhancers." *Nature Genetics* 42(9): 806–10.

Conte, Daniele, Luisa Guerrini, and Giorgio R. Merlo. 2015. "Novel Cellular and Molecular Interactions During Limb Development, Revealed from Studies on the Split Hand Foot Congenital Malformation." In *New Discoveries in Embryology*, ed. Bin Wu. InTech.

Doane, Ashley S., and Olivier Elemento. 2017. "Regulatory Elements in Molecular Networks." *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 9(3).

Li, Yifeng, Chih-Yu Chen, Alice M. Kaye, and Wyeth W. Wasserman. 2015. "The Identification of Cis-Regulatory Elements: A Review from a Machine Learning Perspective." *Bio Systems* 138: 6–17.

Noonan, James P., and Andrew S. McCallion. 2010. "Genomics of Long-Range Regulatory Elements." *Annual Review of Genomics and Human Genetics* 11: 1–23.

Pennacchio, Len A. et al. 2006. "In Vivo Enhancer Analysis of Human Conserved Non-Coding Sequences." *Nature* 444(7118): 499–502.

Rojano, Elena, Pedro Seoane, Juan A G Ranea, and James R Perkins. 2019. "Regulatory Variants: From Detection to Predicting Impact." *Briefings in Bioinformatics* 20(5): 1639–54.

Uemura, Osamu et al. 2005. "Comparative Functional Genomics Revealed Conservation and Diversification of Three Enhancers of the Isl1 Gene for Motor and Sensory Neuron-Specific Expression." *Developmental Biology* 278(2): 587–606.

Vavouri, Tanya, and Greg Elgar. 2005. "Prediction of Cis-Regulatory Elements Using Binding Site Matrices — the Successes, the Failures and the Reasons for Both." *Current Opinion in Genetics & Development* 15(4): 395–402.

Chen, H., Li, H., Liu, F., Zheng, X., Wang, S., Bo, X., & Shu, W. (2015). An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. *Scientific Reports*, *5*.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., et al. (2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, *46*(Database issue), D794–D801.

Naquin, D., d'Aubenton-Carafa, Y., Thermes, C., & Silvain, M. (2014). CIRCUS: A package for Circos display of structural genome variations from paired-end and mate-pair sequencing data. *BMC Bioinformatics*, *15*(1), 198.

Sullivan, A. M., Bubb, K. L., Sandstrom, R., Stamatoyannopoulos, J. A., & Queitsch, C. (2015). DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. *Current Plant Biology*, Transcriptome Networks (including Y1H and single-cell or cell type-specific profiles), *3–4*, 40–47.

Visel, A., Minovitsky, S., Dubchak, I., & Pennacchio, L. A. (2007). VISTA Enhancer Browser—A database of tissue-specific human enhancers. *Nucleic Acids Research*, *35*(Database issue), D88-92.

Calo, E., & Wysocka, J. (2013). Modification of enhancer chromatin: What, how, and why? *Molecular Cell*, *49*(5), 825–837.

Ezer, D., Zabet, N. R., & Adryan, B. (2014). Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. *Computational and Structural Biotechnology Journal*, *10*(17), 63–69.

Pennacchio, L. A., Bickmore, W., Dean, A., Nobrega, M. A., & Bejerano, G. (2013). Enhancers: Five essential questions. *Nature Reviews. Genetics*, *14*(4), 288–295.

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., . . . Ching, C. W. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature, 459(7243), 108-112.

Mukherjee, S., Berger, M. F., Jona, G., Wang, X. S., Muzzey, D., Snyder, M., . . . Bulyk, M. L. (2004). Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nature genetics, 36(12), 1331-1339.

Rebeiz, M., Reeves, N. L., & Posakony, J. W. (2002). SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data. Proceedings of the National Academy of Sciences, 99(15), 9888-9893.

van Steensel, B., & Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nature biotechnology, 18(4), 424-428.

Maston, G. A., Evans, S. K., & Green, M. R. (2006). Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet., 7, 29-59.

Grad, Y. H., Roth, F. P., Halfon, M. S., & Church, G. M. (2004). Prediction of similarly acting cisregulatory modules by subsequence profiling and comparative genomics in Drosophila melanogaster and D. pseudoobscura. Bioinformatics, 20(16), 2738-2750.