Chromosome Wide Prediction of Tissue Specific Enhancers Employing Computational and Statistical Approaches

By

Amna Farooq

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-I-Azam, University

Islamabad- Pakistan

2016

Chromosome Wide Prediction of Tissue Specific Enhancers Employing Computational and Statistical Approaches

By

Amna Farooq

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-I-Azam, University

Islamabad- Pakistan

2016

Chromosome Wide Prediction of Tissue Specific Enhancers Employing Computational and Statistical Approaches

By

Amna Farooq

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

MASTERS OF PHILOSOPY

IN

BIOINFORMATICS

National Center for Bioinformatics

Faculty of Biological Sciences

Quaid-I-Azam, University

Islamabad- Pakistan

2016

CERTIFICATE

This thesis is submitted by **Amna Farooq** 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. XYZ
Designation
Corresponding Intuitional Address

Chairperson: _____
Dr. Sajid Rashid
Associate Professor
Quaid-I-Azam University, Islamabad

Date: 11 july, 2016

DEDICATED

To

My parents and siblings!! ☺

ACKNOWLEDGEMENTS

First and foremost sincere thanks to **Allah** for all the things He has both given and denied me for only He knows what's best for me. Gratefully, joyously, and relentlessly, I commend His splendour and generosity, my **Al-Basir**; beyond all beautiful dreams and hopes.

You cannot steer your way through any journey without a guide. I am highly indebted to ***Dr. Amir Ali Abbasi*** for his guidance and constant supervision as well as for providing necessary information regarding the project. His support has been a great source of motivation for me.

I tender my thanks to ***Dr. Sajid Rashid***, Chairperson, National Center for Bioinformatics, Faculty of Biological Sciences, QAU, for being a great mentor and guide in my course work.

My thanks and appreciations also go to my colleague ***Summaira Yasmeen*** in developing the project, ***Rozeena arif*** for being there in thick and thin, and to my other ***research colleagues*** who have willingly helped me out with their abilities.

Last but not the least, my deepest gratitude goes to ***beloved parents and siblings*** for their endless support, prayers and encouragements all the way since the beginning of my studies.

Amna Farooq

ABSTRACT

Eukaryotic genome is a regulatory jungle having various gene circuits or networks. Regulatory elements serve as switches in these gene circuits. Complete understanding of gene regulation can unravel the hidden mysteries of the genome. Seemingly junk or non coding portions of genome like gene deserts and interons can also hold crucial functional duty of gene regulation. Nevertheless, limited understanding of regulatory elements has hindered their identification always. However, various experimental and computational techniques have been developed to capture their presence indirectly if not directly.

Cis-Regulatory Modules function in a highly tissue specific fashion. This tissue specificity is bestowed to CRMs by combinatorial nature of action of transcription factors. Transcription factors behave in combinatorial mode to achieve spatiotemporal specificity. Cis-Regulatory Modules bear binding sites for these transcription factors in same combinatorial way and inherit the spatiotemporal specificity too. We exploited same property of Cis-Regulatory Modules for their prediction. For this purpose we curated a catalogue of transcription factors using strong literature and experimental evidence. Taking into consideration the evolutionary relevance of human forebrain, we attempted to define forebrain specific transcription factor code. Co occupancy of these forebrain specific transcription factors with was used to locate forebrain specific Cis-Regulatory Modules in non-coding, non repetitive segments of chromosome 7 including gene deserts.

We successfully predicted 1453 putative forebrain specific Cis-Regulaory Modules. On their comparison with experimentally verified enhancers from VISTA Enhancer browser, we found some of the enhancers overlapping with our putative CRMs. Developmental stage and conservation independence, low cost and time efficiency make our method of prediction a comparable substitute of already available computational and experimental techniques of Cis-Regulatory module detection.

Table of Contents

LIST OF ABBREVIATIONS

BED	Browser Extensible Data
Bp	base pair
CRE	Cis Regualtory Element
CRMs	Cis Regulatory Modules
CNS	Central Nervous System
ENCODE	ENCyclopedia of DNA elements
IUPAC	International Union of Pure and Applied Chemistry
PSSMs	Position Specific Scoring Matrix
PWM	Position Weight Matrix
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

Human brain is termed as most complex organ being evolved over last 600 million years ago. Having 3 distinct gross parts, brain can be divided in to Forebrain (proesncephelon), Midbrain (mesencephelon) and hindbrain (rhombencephelon). Largest in size among these three is proesncepheolon, the forebrain, which further divides in to telencephelon (cerebrum) and diencephalon (thalamus, hypothalamus, metathalamus, epithalamus and subthalamus). Chief part among these is cerebrum which is divided in to two hemispheres each having a covering of grey matter and cortex.

Human brain though shares general structural divisions (forebrain, mid brain, hind brain) with other mammals but has more developed cerebral cortex. Higher cognitive and behavioral abilities like thinking, intelligence, decision making are attributed to forebrain namely cortex. In other words, our conscious experience can be attributed to neocortex, a highly specialized part of cortex. Neocortex being associated with intelligence and flexible behavior varies in size among different species. Studies of large brained mammals show that neocortex is disproportionately enlarged in size as compared to other parts of brain (Barton & Harvey, 2000). Subdivision of neocortex in six layers with differing functional role allows serial processing of sensory information. This serial processing has allowed complex computations possible in brain, which is distinguishing attribute of higher mammals. Undoubtedly, corticalization of human brain has over shadowed other parts of brain and has facilitated abilities and accomplishments unmatched by any other species (Kaas, 2013).

How did such a brain evolve? Comparative studies of present day vertebrate brain can possibly give answer. Considering the genetic underpinnings of complex function and abilities of brain, a different set of genes is active at each developmental stage of brain. In fact, there is no developmental process, patterning or fate determination event happening without a particular set of genes active for the purpose. For instance, LIM-homeodomain genes are famous for their role in brain development. A study (Bachy et al., 2002) demonstrates their role in evolution of forebrain by implying their combinatorial code and claims that diversity in function of these genes is due to modifications in coding and regulating sequences of these genes which in turn allow diversified forebrain connectivity. Thus, it can be inferred that gene regulation is a more important contributor to forebrain evolution and development as compared to protein coding genes themselves.

1.1 Gene Regulation

Coding genes are responsible for a particular proteome of a cell which in turn specifies identity of the cell. Since genetic content is shared among all cell types in an organism, these genes are dependent on certain regulatory elements for their expression. Moreover, for purpose of organizational maintenance, energy conservation, and generating phenotypic variance, it is imperative that genes are only expressed when they are needed. This fine control of expression is achieved by employing regulatory elements, which in association with their respective genes form gene regulatory networks or gene circuits. One gene can have multiple regulatory elements associated with it and one regulatory element can regulate multiple genes. Regulatory elements of genome often termed as Cis-Regualtory elements have certain binding sties on them. These binding sites have preferences for Trans-acting elements which when assemble on the Cis-Regualtory element, initiate and regulate transcription of the associated gene. Trans-acting elements are usually proteins that bind to DNA sequence synergistically and antagonistically usually in form of a group, thereby regulating transcriptional events (Hu & Gallo, 2010). A notable fact over here is that individual transcription factors are not tissue and gene specific, instead they synergistically control the gene expression, hence can be termed as a group of trans-acting elements that behave tissue specifically but in a combinatorial fashion.(Latchman, 1998)

1.2 Cis-Reulatory Elements

Cis-Regulatory Elements can be broadly classified into two main groups. Distantly acting and proximally acting. Proximally acting regulatory elements usually lie near the transcription start site and bear binding sites for assembly of transcriptional complex. They also define origin and orientation of transcription. This category comprises of promoters and promoter-proximal elements. Core promoter is located near the transcriptional start site, often 30 bp upstream. It includes TATA box and other similar binding sites which serve as assembly point for RNA polymerase. This core promoter often requires promoter-proximal elements for efficient initiation of transcription. Promoter-proximal elements are present within 100-200bp upstream transcription start site. They serve as binding sites for transcription factors which complete transcription initiation complex (figure1.1 a).

Second class of Cis-acting elements is of distantly acting elements which are distance and orientation independent in their mode of action. This class comprises of enhancers, repressors and insulators(Maston et al., 2006). These elements can be found in non coding portion of

DNA, be it intergenic or intragenic. As name indicates, enhancers enhance expression of gene, where as repressor are meant to reduce Though poorly understood, It is hypothesized that enhancers when complexed with relevant transcription factors are recruited to target gene promoter via chromatin looping, which brings the enhancer and promoter in to indirect physical contact (Dekker et al., 2002).

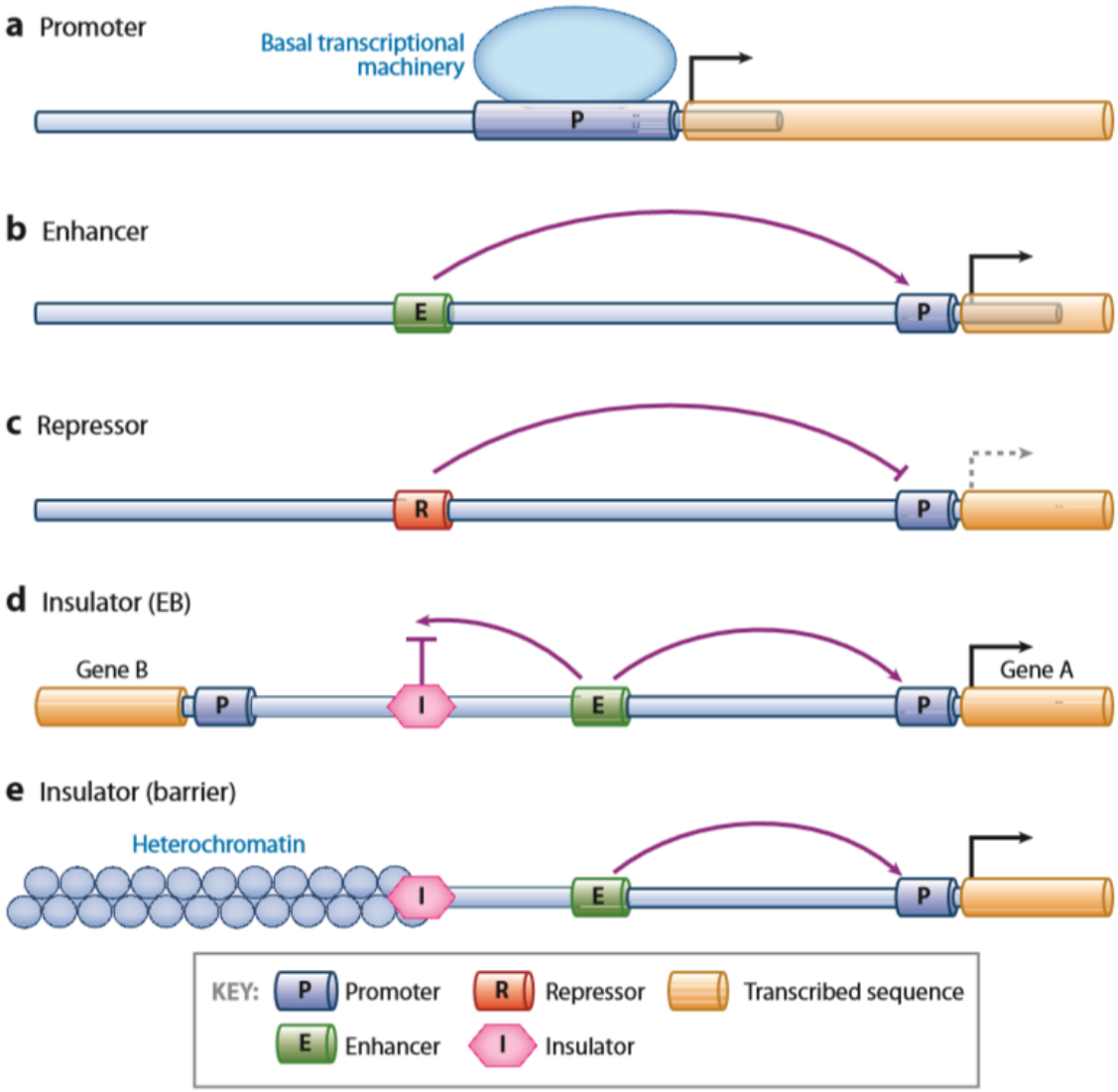

Figure 1.1|Cis-Regulatory Elements. a) Promoter(blue) binding to basal trancriton machienery, to mediate transciption of assosiated gene.**b) Enhnacer** actively up regulating trancription of assosiated gene by rectruiting to basal promoter. **c) Repressor** repressing the activity of assosiated gene by down regulating the expression. **d) Insulator (EB)** blocking the activation of gene B by enhnacer E which is assosiated to gene A. Thus restricting an enhancer to its associated gene only to prevent any misscommunication. **e) Insulator (barrier)** preventing the extending chromatin condensation from itruding the regulatory element controllong the assosiated gene.

Insulators are intergenic elements which uphold transcriptional integrity of neighbouring genes by maintaining distinct regulatory domains. They can either act as enhancer blockers (EB) or can serve as barriers also. In enhancer blocker (EB) mode, inulstor disrupts activity of enhancer in a position dependent manner when placed between enhancer and its respective gene (figure1.1 d). While in barrier mode, it sets boundary between domains of active and repressive chromatin (figure1.1 e).

1.3 Enhancer prediction

Among all the explained types of Cis-Regulstory elements, enhancers are most widely studied and searched yet less understood. Hence, it is hard to predict these elements on basis of primary sequence alone. Enhancers usually recruit a set of transcription factors, to acquire tissue specific regulatory control but transcription factor binding sites are far too short (6-12 bp) and degenerate to sensitively locate presence of enhancer using primary sequence information only. Therefore, current regulatory element predictions rely on indirect methods like comparative genomics and on studies of trans-factor binding and chromatin state.

1.3.1 Experimental Approaches

Experimental techniques exploit TF binding property of enhancers and are mainly divided to *in vivo* and *in vitro* techniques. *In vivo* approaches uncover information of consensus binding sites of TFs, sequence specificity, and the molecular framework of sequence dictated interactions (Hardison & Taylor, 2012). Whereas, in vitro methods reveal the TF consensus binding sites data along with their binding energies, and the physical parameters crucial for these binding events (Geertz & Maerkl, 2010). Following table describes some of known *in vitro* and *in vivo* methods.

Table 1.1| List of available experimental approaches for enhancer prediction.

Type of methods	Name	Method	DNA sequence space	Reference
***In vitro* methods**	SELEX, CASTing	Selection of target	>200,000 sites	(Tuerk & Gold, 1990; Wright et al., 1991)
	HT-SELEX, Bind-n-Seq	One of target couple to NGS	>200,000 sites	(Zhao et al., 2009; Zykovich et al., 2009)

	PBM, CSI	Protein-binding microarray	up to 1 million sites	(Mukherjee et al., 2004; Warren et al., 2006)
	DIP-chip	DNA immunoprecipitation	All genomic sites	(Liu et al., 2005)
	MTIOMI	Mechanical trapping	1000-100 sites	(Maerkl & Quake, 2007)
	EMSA	Gel shift	Around 10 sites	(Hellman & Fried, 2007)
	BIAcore	Surface plasma resonance	Up to 100 sites	(Fägerstam et al., 1992)
***In vivo* methods**	ChIP-chip	ChIP coupled to micro array	All genomic sites	(Buck & Lieb, 2004)
	ChIP-Seq	ChIP coupled to NGS	All genomic sites	(Visel et al., 2009)
	Dam-ID	TF mediated DNA methylation profiling	All genomic sites	(van Steensel & Henikoff, 2000)
	PICh	Resverse ChIP	One genomic sites	(Déjardin & Kingston, 2009)
	DNaseI-seq	DNaseI sensitivity profiling coupled to NGS	All genomic sites	(Hesselberth et al., 2009)

1.3.2 Computational Approaches

Many research groups have diverted their efforts towards computational approaches for identification of enhancers due to high cost and low resource efficiency of experimental procedures. These 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. For instance, SCORE (Sites Clustering Over Random Expectation) incorporates improbable concentrations of transcription factor binding sites to detect cis-regulatory elements (Rebeiz et al., 2002) . Another tool Toucan identifies cis-regulatory elements from set of co-regulated genes using TFBSs position weight matrixes and statistical over representation of TFBS (Aerts et al., 2003). Some de novo methods of prediction have employed various algorithms like Hidden

Markov Model, expectation maximization, problabistic mixture modeling (NestedMica, (Down & Hubbard, 2005)) etc.

One of the famous computational methods is phylogenetic foot printing which exploits filtering power of evolutionary constraint. When coupled with other techniques, it has increased specificity of predictions in various studies (Gibbs et al., 2004). Some online tools like ConSite harness high quality transcription factor models and cross specie conservation for regulatory element prediction(Sandelin et al., 2004). Limitation of phylogenetic foot printing is its insensitivity to weakly conserved regulatory elements. However, inputs and outputs of computational and experimental studies are similar in current context. Keeping their individual pros and cons, both types of methods perform better if used interdepenedently. Figure 1.2 shows the interdependent role of computational and experimental approaches of cis-regulatory element prediction.

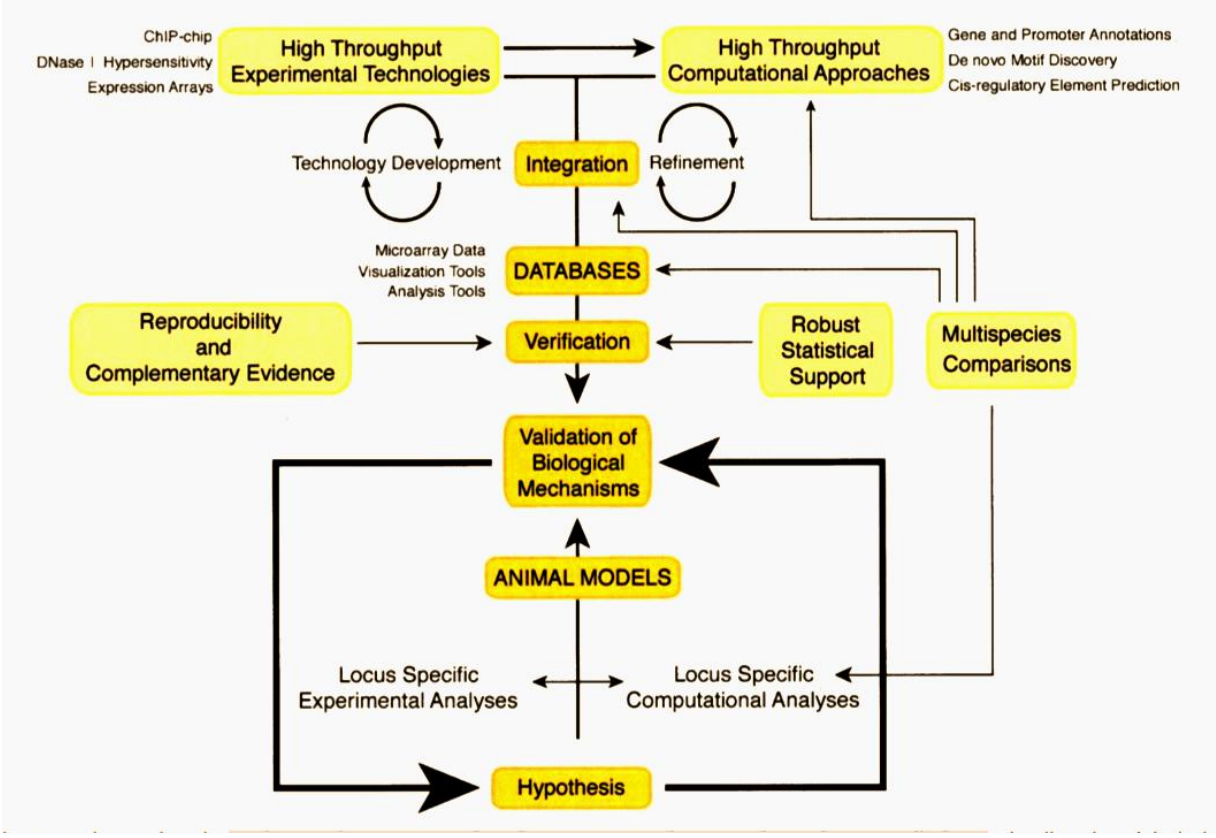

Figure 1.2| Relationship between experimental and computational methods of CRMs prediction. Centeral yellow box labeled as "Validation of Biological Mechanisms" describes the ultimate goal for researchers studying a biological pathway. Methods illustrated on top of image represent high-through put analysis used to predict TFBSs and to elucidate functional relevence of those elements. Methods illustrated from bottom of the image represent conventional "locus/ position specific" analysis that initiates from a narrowly defiened hypothesis of biological function and involves animal model usage also(Elnitski et al., 2006).

1.4 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(Scherer et al., 2003). Up till current annotation status, it contains approximately 1800 genes. Numerous of them hold clinical significance. Developmentally crucial genes like GLI3, TWIST, SHH, and CDK6 also reside on chromosome 7. 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. Genes and related features summary of human chromosome 7 is described in table 1.2. Owing to disease relevance, moderate size and some previously performed studies, we opted chromosome 7 for our study as a test data.

Table 1.2| Human Chromosome 7 gene summary.

Categories of genes	No of genes	Gene length	Transcript length	Exon size	No. of exons per gene
Known genes	863	69877	2639	261	10.1
Novel genes	41	50103	1989	386	5.2
Partial Genes	70	42964	1850	339	5.5
Predicted genes	481	14573	1026	326	3.1
Putative and non coding RNA genes	213	17501	1638	629	2.6
Total	1668				

1.5 Gene Deserts

Genes are not distributing in genome randomly. Some regions have high gene density and some have low. Such considerable large regions which are completely gene deficient are termed as gene deserts. There is no precise definition of "considerable" intergenic distance; however 400-700 kbs is considered minimum intergenic distance in various studies. Some gene deserts, specifically those which show substantial evolutionary conservation do depict functional relevance and play part in gene regulation. Thus, gene deserts can serve as best search space for enhancer hunt. taking into account their importance, many studies have been dedicated to regulatory element search in gene deserts. Enhancers have been identified in gene deserts surrounding *Iroquois* genes (de la Calle-Mustienes et al., 2005). Two highly conserved elements have been identified in a 1400kb large gene desert downstream the human *ISL1* gene (Uemura et al., 2005). In another study, 36 enhancers in gene deserts were identified and functionally confirmed in transgenic mice enhancer assay(Pennacchio et al., 2006).

Considering importance of gene deserts, we also included chromosome 7 known gene deserts in our search.

1.6 Aims of Study

Eukaryotic genomic landscape is more like a regulatory jungle. This spatiotemporally specific regulation of genes has blessed higher eukaryotes with phenotypic complexity. Key drivers of this regulation are cis-regulatory elements which bear binding sites for transcription factors. Transcription factors act in a highly combinatorial fashion to precisely control the time and space of gene expression. Consequently, cis-regulatory elements also bear transcription factor binding sites in same combinatorial manner. In the current study we aim to decipher vocabulary of tissue specific Cis-Regulatory Modules (CRMs) keeping forebrain development as model frame work. After curating transcription factor code, CRMs would be identified in a chromosome wide manner. Human chromosome 7 will be chosen as a search space. Taking into account functional relevance of gene deserts, CRM search will be directed to gene deserts also.

Chapter 2

MATERIAL AND METHODS

2. MATERIAL AND METHODS

As per aims of study, we intend to predict forebrain specific enhancers on basis of transcription factor binding sites present on whole chromosome, thus initial step was to collect genomic sequences of selected human chromosome 7. Chromosome 7 was chosen on basis of its disease relevance, rich gene density and presence of numerous CNS related genes (Hillier et al., 2003).

2.1 Data collection

For enhancer prediction, two inputs were required as raw material according to our study scheme; Transcription factor binding sites and genomic sequences. Numerous online repositories are available which contain binding sites for nearly all transcription factors. These databases have experimental as well as predicted TFBS generated using computational and statistical techniques like Bayesian theorem, genetic algorithm, and Hidden Markov Model. Table 2.1 catalogues some online repositories which provide binding profiles of transcription factors.

Table 2.1|Online repositories providing insight to Transcriptional data.

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

2.1.1 Forebrain Specific Transcription Factor Catalogue

A set of transcription factors crucially expressed in forebrain was designed by extensive literature survey. Arduous statistical and computational analysis was done to downsize the set to six vital transcription factors essentially expressed in forebrain tissue.

2.1.2 Transcription Factor Binding Site Collection

Transcription Factor Binding Sites for each transcription factor were collected from literature survey and over lapped with TRANSFAC results. TRANSFAC is a public data base which provides experimentally verified binding sites, Position Weight Matrixes of consensus binding sites and regulated genes of eukaryotic transcription factors (Matys et al., 2003).

2.1.3 Genomic Sequence Collection

Keeping in view the goals of study, we collected experimentally validated forebrain enhancer data from VISTA Enhancer browser. VISTA Enhancer browser is a publically available database which provides experimentally validated human and mouse enhancers functionally tested in mice (Visel et al., 2007). It chooses candidate enhancers usually from the non-coding portion of genome, on basis of their deep conservation among vertebrates and epigenetic marks associated with enhancer activity. Selected non coding regions are then subjected to experimental validation in transgenic mice assay. Any element exhibiting reproducible expression in same tissue among more than three embryos is considered positive. High resolution images of transgenic embryos for each positive element are also provided. Such elements which fail to exhibit reproducible expression pattern in any three mice embryos out of the five tested are considered negative. For each enhancer following information is provided: Expression pattern with embryonic images, genomic coordinates in Hg19 or mm9, flanking genes, conservation pattern of region in mouse, primers used and DNA sequence. We collected 104 enhancers which exhibited expression pattern in forebrain which were used as training data set in our study.

Second part of sequences collection was retrieval of genomic sequence used as test data set in our study. For testing data set, we chose human chromosome 7. DNA sequence for human chromosome 7 was retrieved from UCSC Genome Browser under GRCh37/hg19 assembly. UCSC Genome Browser is one of the largest and widely used online repository which contains reference sequences and working draft assemblies of various genomes (Kent et al., 2002). Chromosome 7 DNA sequence was downloaded from UCSC by using UCSC DAS server. UCSC DAS server provides facilitates users by providing access to genome annotation data for all assemblies currently available in the Genome Browser. FASTA sequence of each chromosome is available in form of compressed .gz files. Since enhancers cannot be present in repeats, we obtained repeat masked sequence for chromosome7 from UCSC. Repeats from RepeatMasker and Tandem Repeats Finder are masked by character N.

Rest of non repeat sequence is represented in upper case letters. Since UCSC provides positive or sense strand sequence for download, sequence was translated to complementary sequence by a simple perl code to get a copy of negative strand.

2.2 Masking

Enhancers are usually present in non-genic and non-repeat portion of genome. In other words, enhancers are often found in intergenic spaces (other than repeats), and intra-geneic spaces like interons. To reduce our search space and to be more targeted towards regions highly favorable for enhancer activity, we had to mask chromosome 7 further. Exonic portion of chromosome 7 had to be masked prior to enhancer search.

For exon masking, exonic coordinates for all genes present on chromosome 7 were obtained from Table Browser of UCSC Genome Browser for same assembly. Fields of table browser were adjusted likewise; group was set to Genes and Genes Predictions for track RefSeq genes. Output format was set to Browser Extensible Data (BED) and BED records were created pre exon plus zero bases at each end to include non coding exons also.

2.2.1 BED Processing

Genes are present on both strands of chromosome, so both strands were masked for their respective exons. For this purpose, columns other than strand ChrNo, Start, End and Strand information were removed from the BED file. Afterwards, coordinates of positive and negative strand were extracted to separate files.

Genomic sequence of chr7 positive strand was subjected to masking with bedtools masking module with respective BED file. Same procedure was repeated for negative strand. bedtools are a set of UNIX command line tools for various genomic analysis tasks. Bedtools apply set theory to genome and enable user to intersect, merge, count, sort, complement and subtract genomic coordinates. It accepts various file formats like BED, BAM, GFF, VCF and GTF.

2.3 TFBS Mapping on Test Data Set

Forebrain specific TFBS catalogue prepared in previous steps was fed to Transcription Factor Binding Site Mapping Algorithm (TFBSMA) developed in our lab. TFBSMA works employing string searching paradigm. Since one transcription factor can have multiple binding sites and TFBS are degenerate, suitable data structure to hold this multi-dimensional data was hash. A hash for TFBS was created along the hash for nucleotide ambiguity as in

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 FB specific catalogue as an output. Output was generated separately for positive and negative strands o chromosome 7.

Table 2.2|IUPAC Nucleotide ambiguity code (Kozl & Listy, 1978).

Nucleotide code symbol	Mnemonic	Base
A	**A**denine	A
T	**T**hymidine	T
G	**G**uanine	G
C	**C**ytosine	C
W	**W**eak	A or T
S	**S**trong	C or G
Y	P**y**rimidine	C or T
R	Pu**r**ine	A or G
K	**K**eto	G or T
M	A**m**ino	A or C
B	Not A	T or G or C
D	Not C	A or T or G
H	Not G	A or T or C
V	Not T and not U	A or G or C
N	a**n**y	A or T or G or C

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.

2.4.1 Cluster Size Optimization

Cluster size would define size of an enhancer. If a cluster has on average six TFBS with size between range of 1000-4000 bps, this makes an average spacer distance lying in range of 150-600 bps. Thus, for size optimization of clusters, clusters were generated at spacer

sequence of 150, 250, 400, 500 and 600 bps. Results were analyzed to further choose the optimal spacer sequence.

2.4.1.1 Spacer distance Optimization from Experimental Evidence

Apart from optimization by random cluster generation, proof from experimental data was also collected. For this purpose, sequences for functionally validated forebrain specific enhancers were retrieved from VISTA Enhancer browser (as training data set). Sequence for each enhancer was subjected to TFBS occurrence program as an input. Output received was then used to calculated distance between each neighboring TFBS. Average ($\overline{D}$) of this distance was calculated as per the formula mentioned below and was used as optimal spacer sequence in further experimentation and analysis. Where $\overline{D}$ stands for average spacer distance, e stands for enhancer, n for total number of enhancers, i for number of binding site of an enhancer, m total number of binding site within an enhancer and $BS_{e\,i}$ stands for a particular binding site of a particular enhancer. Figure 2.1 explains the formula pictorially.

$$\overline{D} = \frac{\sum_{e=1}^{n} [\sum_{i=1}^{m} [BS_{e\,i+1} - BS_{e\,i}]]}{n}$$

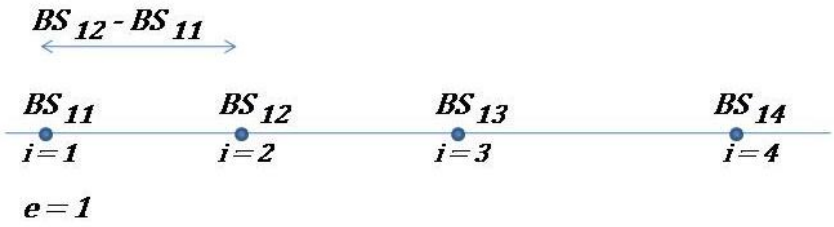

Figure 2.1|Spacer Distance Optiomization: $BS_{e\,i}$ represents a transcription factor binding site on an enhancer e (which in this case is 1). For enhancer 1, four binding sites $i = 1,2,3,4$ are shown. Distance between two binding sites within an enahncer is calculated by subtracting the genomic position of both binding sites ie. *BS 12 - BS 12.*

2.5 Filtration of Clustering Data

Clusters obtained by using 250 bps as spacer sequence were stored in database, where further filtration was applied. As first filter applied was of uniqueness and number of TFBS. Only those instances were chosen where binding sites for 6 out of 6 transcription factors of forebrain specific catalogue were present. Resultant clusters were subjected to size calculation. Average length of enhancer usually varies between 500-4000 bps. A second step for filtration was carried out by applying size filter of 3000. Clusters exceeding the threshold of filter were discarded. Homotypic and heterotypic clusters were kept. All the filters were applied on data obtained for positive strand as well as negative strand.

2.6 Combining Results

Clusters obtained from negative and positive strands of chromosome 7 were combined. Clusters overlapping between positive and negative strand were removed. Cluster Ids were replaced by CRM Ids to avoid redundancy in identifiers.

2.7 Validation from Vista Enhancers

Vista Enhancers browser keeps experimental evidence for enhancers. In an effort attempted to validate our results from vista enhancer browser, we collected all enhancers residing on human chromosome 7, regardless of their expression tissue. Information about coordinates and expression tissue was extracted from data via regular expression and special variables. This information was converted into bed format by another code. Resultant bed coordinates of vista enhancers of chr7 were overlapped with our TFBS based CRMs.

2.8 Locating CRMs in Gene Deserts

Gene deserts are long genomic stretches which are deficient of coding genes. In current study, information of gene deserts residing on human chromosome 7 was adapted from the project of The Center for Applied Genomics (Scherer, et al., 2003). For this study any region on chromosome 7 having no known, novel, or partial gene in more than 500kb was termed as gene desert. Percentage of repeat content and no. of CpG islands was also noted. Our predicted CRMs were then located within the defined gene deserts.

2.9 Data Representation

Circular visualization was chosen for representing chromosome wide results. For this purpose Circos, a circular visualization tool was chosen. Circos is a software package for visualizing data in a circular layout allowing user to analyze relationships between different dimensional data (Krzywinski et al., 2009). By using circular ideograms it helps in visualizing relationships between genomic coordinates. It also allows incorporation of data in form of scatter, heat maps, line and histogram plots, tiles, connectors and text. It is command line based tools which works using multiple statistics, mathematics and graphics modules of perl.

We converted our data into format described by Circos. Configuration files was prepared and linked to data files. Data was represented in two tracks in form of tiles.

Chapter 3

RESULTS

3. RESULTS

In current study we formulated catalogue of transcription factors specifically active in forebrain. For this purpose we relied on functional data reported in multiple studies. After performing some statistical and computational analysis we finalized a catalogue of six transcription factors essentially active in forebrain (Table 3.1).

Table 3.1|Transcription factor catalogue forFB enahncer prediction

Sr.No	Transcription Factor	Known Expression Tissue	Source
1.	*NGN2*	Diencephalon	(Osório et al., 2010)
2.	OTX1	Future Pros-Encephalon	(Gray et al., 2004)
3.	OTX2	Future Pros-Encephalon	(Gray, et al., 2004)
4.	GATA3	Diencephalon	(Gray, et al., 2004)
5.	FOXP2	Diencephalon, Ventricular Layer	(Gray, et al., 2004)
6.	HES5	Diencephalon, Telencephalon	(Gray, et al., 2004)

3.1 TFBS mapping output

Binding sites of these transcription factors collected from literature and verified from TRANSFAC were assembled in a hash to handle the dimensionality of data. After tackling the degeneracy of binding site, binding sites of selected transcription factors were coded to TFBSMA. Exon and repeat masked genomic sequence of human chromosome 7 was subjected to TFBS mapping. Table 3.2 summarizes number of binding sites found against each transcription factor from catalogue. There were total 190,397 TFBS found on positive strand of chromosome 7 and 174,142 on negative strand.

Table 3.2| Transcription factor binding sites found on chromosome 7.

Transcription Factor	Occurrences on +ve strand	Occurrences on -ve strand
FOXP2	41928	29680
HES5	57112	54067
GATA3	34491	41963
NGN2	24186	8900
OTX1	15432	18356
OTX2	17248	21176
Total	**190397**	**174142**

3.2 TFBS Clusters

Different enhancers have different TFBS present on them, depending upon the transcription factor cooperativity. Keeping in mind our FB specific TF catalogue, we assumed that a cluster has on average six TFBS with size between range of 1000-4000 bps, this makes an average spacer distance lying in range of 150-600 bps. In an attempt to optimize length of spacer sequence, we generated clusters at spacer sequence of 150, 250, 400, 500 and 600 bps for both strands of chromosome 7. Each cluster contained 1 or more redundant TFBS depicting homotypic clustering and up to 6 unique TFBS portraying heterotypic clustering. Table 3.3 shows the number of clusters generated for each strand against each spacer distance selected.

Table 3.3| Clusters generated at various spacer distances.

Spacer Distance	Positive Strand	Negative Strand
150	124,645	114,661
250	106,614	99,788
400	86,510	82,699
500	75,909	73,387
600	67,019	65,559

Other then the test data set, functional data set was also used for spacer sequence optimization. By calculating the average distance between each binding site present on 104 forebrain specific enhancers retrieved from VISTA Enhancer Browser, we got a spacer sequence of 250 bps. Thus, Clustering results obtained by setting spacer distance at 250bps were used for further analysis.

3.3 Filtration of Clusters

All such clusters having less than 5 out of six unique transcription factor binding sites were filtered out. Moreover, any clusters having size above 3000 bps were also filtered out. Figure 3.1 shows number of clusters predicted with maximum 1,2,3,4,5 and 6 unique TFBS for negative and positive strands of chromosome 7. Though large number of clusters were predicted with only one or two TFBS, but such results were discarded in coherence with the structure of eukaryotic Cis-regulatory modules. Clusters with 5 TFBSs were used in further studies.

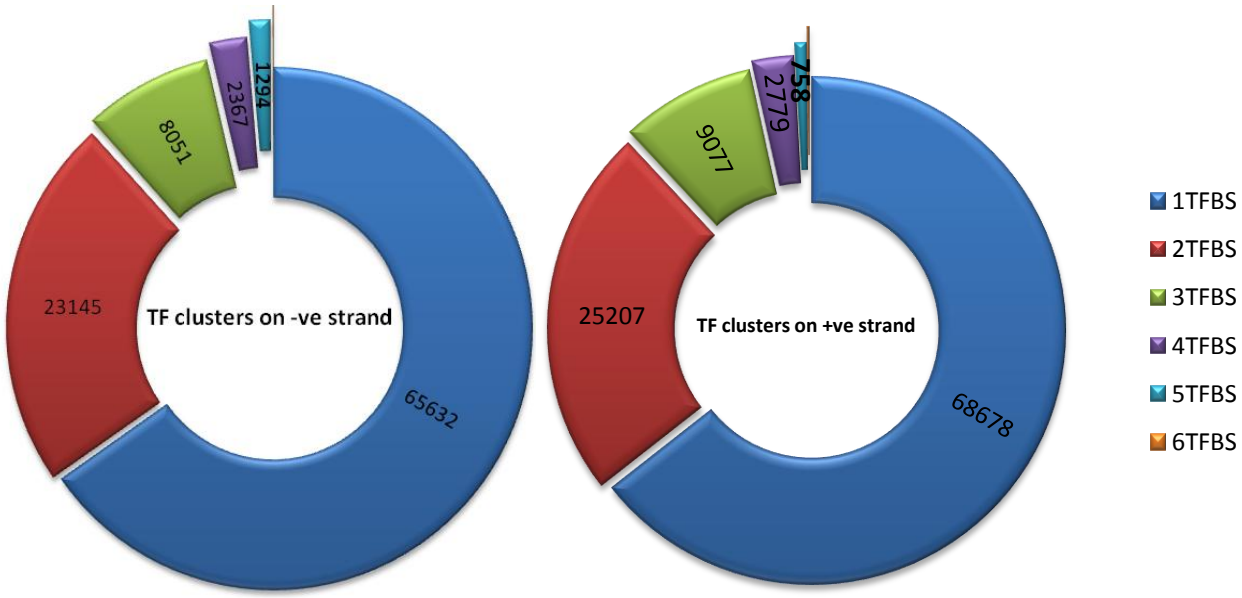

Figure 3.1| Pie charts of clusters on base of maximum unique TFBS.

3.4 Overlaps between Strands

Data was generated and clustered in a way that there were no overlapping clusters with in a strand. To ensure that there are no overlapping clusters between strands, such enhancers which had complete or even partially overlapping coordinates between positive and negative strands were deleted. There were total 13 overlaps found, and those enhancers were deleted from either positive or negative strand results after careful analysis based on length and number of binding sites predicted in each.

After careful removal of overlaps, elements from positive and negative strands were combined and assigned unique identifiers to remove any ambiguity based on cluster number. Table 3.4 and table 3.5 represent our filtered Cis-Regulatory Modules CRMs for chromosome 7 positive and negative strands respectively.

Table 3.4| Cis-Regulatory Modules (CRMs) predicted on positive strand of chromosome 7, keeping spacer sequence distance 250 bps with minimum 5 TFBS.

CRM_id	Coordinates	Size (bps)	CRM_id	Coordinates	Size (bps)
crm1	Chr7:555191-556038	847	**crm2**	Chr7:2069300-2069755	455
crm3	Chr7:2926066-2926562	496	**crm4**	Chr7:3364858-3365633	775
crm5	Chr7:3368219-3368857	638	**crm6**	Chr7:3730053-3730273	220
crm7	Chr7:3764507-3765130	623	**crm8**	Chr7:3771897-3772432	535
crm9	Chr7:4271453-4271583	130	**crm10**	Chr7:5099899-5100299	400
crm11	Chr7:5725180-5725449	269	**crm12**	Chr7:6316719-6317234	515
crm13	Chr7:6369020-6369182	162	**crm14**	Chr7:6473895-6474644	749
crm15	Chr7:7189023-7189258	235	**crm16**	Chr7:7654310-7655092	782
crm17	Chr7:8364472-8364692	220	**crm18**	Chr7:8420461-8421227	766
crm19	Chr7:8424700-8425262	562	**crm20**	Chr7:8516335-8517229	894
crm21	Chr7:8530711-8531459	748	**crm22**	Chr7:8557221-8557655	434
crm23	Chr7:8614202-8614998	796	**crm24**	Chr7:8677609-8678151	542
crm25	Chr7:8901651-8901896	245	**crm26**	Chr7:9265841-9266294	453
crm27	Chr7:9387815-9388196	381	**crm28**	Chr7:9389044-9389764	720
crm29	Chr7:9445281-9445731	450	**crm30**	Chr7:9450839-9451055	216
crm31	Chr7:9452947-9453475	528	**crm32**	Chr7:9535938-9536722	784
crm33	Chr7:9569681-9570324	643	**crm34**	Chr7:9592943-9593259	316
crm35	Chr7:9747677-9747940	263	**crm36**	Chr7:9871954-9872366	412
crm37	Chr7:9965422-9965862	440	**crm38**	Chr7:10308819-10309226	407
crm39	Chr7:10322218-10322752	534	**crm40**	Chr7:10364940-10365422	482
crm41	Chr7:10469207-10469867	660	**crm42**	Chr7:10520249-10520507	258
crm43	Chr7:10676242-10676553	311	**crm44**	Chr7:10685350-10685688	338

crm45	Chr7:10829738-10830280	542	**crm46**	Chr7:11286958-11287401	443
crm47	Chr7:11379135-11379845	710	**crm48**	Chr7:11491422-11492435	1013
crm49	Chr7:11653824-11654219	395	**crm50**	Chr7:11751991-11752902	911
crm51	Chr7:11757548-11758321	773	**crm52**	Chr7:11914015-11914842	827
crm53	Chr7:11968389-11968856	467	**crm54**	Chr7:12083831-12084431	600
crm55	Chr7:12159544-12159881	337	**crm56**	Chr7:12443320-12444124	804
crm57	Chr7:12861409-12861708	299	**crm58**	Chr7:12985554-12985822	268
crm59	Chr7:13046707-13047178	471	**crm60**	Chr7:13330969-13331530	561
crm61	Chr7:13370850-13371229	379	**crm62**	Chr7:13538557-13539607	1050
crm63	Chr7:13764452-13765166	714	**crm64**	Chr7:13890124-13891258	1134
crm65	Chr7:13903154-13903507	353	**crm66**	Chr7:14129602-14130263	661
crm67	Chr7:14210716-14211266	550	**crm68**	Chr7:14299573-14300190	617
crm69	Chr7:14580805-14581491	686	**crm70**	Chr7:14592210-14592703	493
crm71	Chr7:14664655-14665599	944	**crm72**	Chr7:14733432-14733862	430
crm73	Chr7:14767706-14767930	224	**crm74**	Chr7:14871939-14872217	278
crm75	Chr7:15169461-15170214	753	**crm76**	Chr7:15185431-15186006	575
crm77	Chr7:15215029-15215737	708	**crm78**	Chr7:15358085-15358176	91
crm79	Chr7:15539639-15540072	433	**crm80**	Chr7:15596207-15596765	558
crm81	Chr7:15705199-15706111	912	**crm82**	Chr7:15724413-15724925	512
crm83	Chr7:15804741-15805066	325	**crm84**	Chr7:16049466-16050131	665
crm85	Chr7:16090703-16091469	766	**crm86**	Chr7:16098972-16099515	543
crm87	Chr7:16111299-16111896	597	**crm88**	Chr7:16226772-16227074	302
crm89	Chr7:16564749-16565340	591	**crm90**	Chr7:16566812-16567240	428
crm91	Chr7:16724264-16725101	837	**crm92**	Chr7:16832070-16832467	397
crm93	Chr7:17073723-17074236	513	**crm94**	Chr7:17171270-17171456	186
crm95	Chr7:17336006-17336410	404	**crm96**	Chr7:17417620-17418061	441
crm97	Chr7:17501599-17502475	876	**crm98**	Chr7:17599170-17599525	355

crm99	Chr7:17667852-17668150	298	**crm100**	Chr7:17832762-17833297	535
crm101	Chr7:17833964-17834945	981	**crm102**	Chr7:17867082-17867246	164
crm103	Chr7:18281375-18282121	746	**crm104**	Chr7:18340932-18341495	563
crm105	Chr7:18384127-18385006	879	**crm106**	Chr7:18516959-18517601	642
crm107	Chr7:18591556-18592136	580	**crm108**	Chr7:18592516-18592699	183
crm109	Chr7:18729021-18729888	867	**crm110**	Chr7:18769020-18769558	538
crm111	Chr7:19042983-19043645	662	**crm112**	Chr7:19142133-19142640	507
crm113	Chr7:19178722-19179391	669	**crm114**	Chr7:19395325-19395604	279
crm115	Chr7:19412552-19412854	302	**crm116**	Chr7:19537515-19538158	643
crm117	Chr7:19607919-19608726	807	**crm118**	Chr7:19894266-19894604	338
crm119	Chr7:19978368-19978833	465	**crm120**	Chr7:19979980-19980376	396
crm121	Chr7:20121512-20122120	608	**crm122**	Chr7:20342295-20342799	504
crm123	Chr7:20477599-20478732	1133	**crm124**	Chr7:20512599-20513323	724
crm125	Chr7:20515124-20515635	511	**crm126**	Chr7:20524129-20524583	454
crm127	Chr7:20726552-20727097	545	**crm128**	Chr7:21002813-21003638	825
crm129	Chr7:21039629-21040151	522	**crm130**	Chr7:21086622-21087323	701
crm131	Chr7:21100226-21100836	610	**crm132**	Chr7:21112705-21113133	428
crm133	Chr7:21253669-21254488	819	**crm134**	Chr7:21299101-21299840	739
crm135	Chr7:21313627-21314098	471	**crm136**	Chr7:21357145-21357897	752
crm137	Chr7:21413082-21413493	411	**crm138**	Chr7:21528437-21528667	230
crm139	Chr7:21696281-21696751	470	**crm140**	Chr7:22089581-22090216	635
crm141	Chr7:22168847-22169442	595	**crm142**	Chr7:22174917-22175254	337
crm143	Chr7:22212198-22212590	392	**crm144**	Chr7:22243598-22244384	786
crm145	Chr7:22249208-22249677	469	**crm146**	Chr7:22288133-22288646	513
crm147	Chr7:22347941-22348452	511	**crm148**	Chr7:22403916-22404309	393
crm149	Chr7:22534264-22534637	373	**crm150**	Chr7:22639577-22640219	642
crm151	Chr7:22857246-22857774	528	**crm152**	Chr7:22891521-22891933	412

crm153	Chr7:23016974-23017310	336	**crm154**	Chr7:23098072-23098421	349
crm155	Chr7:23146548-23147143	595	**crm156**	Chr7:23338454-23338828	374
crm157	Chr7:23418769-23419023	254	**crm158**	Chr7:23469132-23469660	528
crm159	Chr7:23507633-23508458	825	**crm160**	Chr7:24193611-24193839	228
crm161	Chr7:24232543-24233146	603	**crm162**	Chr7:24784974-24785233	259
crm163	Chr7:24937381-24937748	367	**crm164**	Chr7:24960472-24960696	224
crm165	Chr7:25215314-25215712	398	**crm166**	Chr7:25264129-25264345	216
crm167	Chr7:25300401-25300912	511	**crm168**	Chr7:25382466-25383118	652
crm169	Chr7:25406556-25407122	566	**crm170**	Chr7:25601942-25603038	1096
crm171	Chr7:25772996-25773708	712	**crm172**	Chr7:25887504-25887770	266
crm173	Chr7:26073235-26073423	188	**crm174**	Chr7:26231647-26231889	242
crm175	Chr7:26232635-26233168	533	**crm176**	Chr7:26432137-26432567	430
crm177	Chr7:26695285-26695911	626	**crm178**	Chr7:26701998-26702615	617
crm179	Chr7:27334683-27334966	283	**crm180**	Chr7:28032857-28033329	472
crm181	Chr7:28068589-28069467	878	**crm182**	Chr7:28140072-28140673	601
crm183	Chr7:28195041-28195680	639	**crm184**	Chr7:28228486-28229220	734
crm185	Chr7:28367427-28367937	510	**crm186**	Chr7:28536064-28536700	636
crm187	Chr7:28583869-28584756	887	**crm188**	Chr7:28613617-28614096	479
crm189	Chr7:28618558-28619313	755	**crm190**	Chr7:28647655-28648450	795
crm191	Chr7:28939422-28940124	702	**crm192**	Chr7:29110393-29111620	1227
crm193	Chr7:29111874-29112461	587	**crm194**	Chr7:29264294-29264858	564
crm195	Chr7:29611456-29611777	321	**crm196**	Chr7:29875812-29876251	439
crm197	Chr7:29893579-29894155	576	**crm198**	Chr7:30062229-30062639	410
crm199	Chr7:30573627-30574199	572	**crm200**	Chr7:30817021-30817324	303
crm201	Chr7:31381520-31382224	704	**crm202**	Chr7:31641418-31642164	746
crm203	Chr7:31783929-31784735	806	**crm204**	Chr7:31920411-31920955	544
crm205	Chr7:31961822-31962231	409	**crm206**	Chr7:32062301-32063065	764

crm207	Chr7:32065412-32065948	536	**crm208**	Chr7:32075484-32075733	249
crm209	Chr7:32086911-32087131	220	**crm210**	Chr7:32527342-32527538	196
crm211	Chr7:32972211-32972413	202	**crm212**	Chr7:33597224-33597651	427
crm213	Chr7:33913038-33913662	624	**crm214**	Chr7:34068813-34069118	305
crm215	Chr7:34288922-34289436	514	**crm216**	Chr7:34626097-34626596	499
crm217	Chr7:34711885-34712447	562	**crm218**	Chr7:35052752-35053060	308
crm219	Chr7:35290675-35291064	389	**crm220**	Chr7:35325018-35325790	772
crm221	Chr7:35360367-35361017	650	**crm222**	Chr7:35426692-35427221	529
crm223	Chr7:35489891-35490504	613	**crm224**	Chr7:35709264-35709732	468
crm225	Chr7:35754766-35754937	171	**crm226**	Chr7:35807574-35808156	582
crm227	Chr7:36109558-36110281	723	**crm228**	Chr7:36375841-36376086	245
crm229	Chr7:36944497-36944930	433	**crm230**	Chr7:37052115-37052653	538
crm231	Chr7:37066496-37067204	708	**crm232**	Chr7:37151608-37152435	827
crm233	Chr7:37184943-37185319	376	**crm234**	Chr7:37214185-37214953	768
crm235	Chr7:37430192-37430575	383	**crm236**	Chr7:37585079-37585796	717
crm237	Chr7:38107726-38108491	765	**crm238**	Chr7:38381449-38382641	1192
crm239	Chr7:38516513-38516958	445	**crm240**	Chr7:38881650-38882241	591
crm241	Chr7:39174050-39174272	222	**crm242**	Chr7:39191354-39192303	949
crm243	Chr7:39492723-39493593	870	**crm244**	Chr7:39581860-39582064	204
crm245	Chr7:39805470-39805730	260	**crm246**	Chr7:40592860-40593464	604
crm247	Chr7:40598664-40599420	756	**crm248**	Chr7:40668834-40669482	648
crm249	Chr7:40693280-40693915	635	**crm250**	Chr7:40797300-40798089	789
crm251	Chr7:40830808-40831621	813	**crm252**	Chr7:40834741-40835242	501
crm253	Chr7:40911787-40912089	302	**crm254**	Chr7:40953421-40953608	187
crm255	Chr7:41052418-41053016	598	**crm256**	Chr7:41084603-41084999	396
crm257	Chr7:41123538-41124109	571	**crm258**	Chr7:41168626-41169008	382
crm259	Chr7:41383693-41384569	876	**crm260**	Chr7:41551988-41552549	561

crm261	Chr7:41711657-41712302	645	**crm262**	Chr7:42052268-42052851	583
crm263	Chr7:42184368-42184834	466	**crm264**	Chr7:42249863-42250914	1051
crm265	Chr7:42263089-42263599	510	**crm266**	Chr7:42507551-42507982	431
crm267	Chr7:42661681-42662166	485	**crm268**	Chr7:42749637-42749979	342
crm269	Chr7:42783107-42783403	296	**crm270**	Chr7:42947976-42948598	622
crm271	Chr7:42971024-42971264	240	**crm272**	Chr7:43150581-43150960	379
crm273	Chr7:43319589-43320470	881	**crm274**	Chr7:43490745-43491026	281
crm275	Chr7:43565000-43565659	659	**crm276**	Chr7:43828618-43829093	475
crm277	Chr7:45014438-45014863	425	**crm278**	Chr7:45284545-45285115	570
crm279	Chr7:45581297-45581741	444	**crm280**	Chr7:45604111-45604236	125
crm281	Chr7:45608123-45608464	341	**crm282**	Chr7:45716900-45717333	433
crm283	Chr7:45826755-45827082	327	**crm284**	Chr7:46472579-46473109	530
crm285	Chr7:46523120-46523706	586	**crm286**	Chr7:46995553-46996309	756
crm287	Chr7:47014652-47015244	592	**crm288**	Chr7:47018711-47019246	535
crm289	Chr7:47125242-47125653	411	**crm290**	Chr7:47191632-47191870	238
crm291	Chr7:47498627-47498880	253	**crm292**	Chr7:47547296-47548128	832
crm293	Chr7:47605349-47605665	316	**crm294**	Chr7:47613262-47613776	514
crm295	Chr7:48121638-48122022	384	**crm296**	Chr7:48141934-48142709	775
crm297	Chr7:48196850-48197112	262	**crm298**	Chr7:48210020-48210663	643
crm299	Chr7:48387070-48387789	719	**crm300**	Chr7:48694619-48695482	863
crm301	Chr7:48709574-48710027	453	**crm302**	Chr7:48883051-48883568	517
crm303	Chr7:48956402-48957388	986	**crm304**	Chr7:49184578-49184886	308
crm305	Chr7:49287120-49287425	305	**crm306**	Chr7:49293265-49293940	675
crm307	Chr7:49744312-49744736	424	**crm308**	Chr7:50451386-50452319	933
crm309	Chr7:50473990-50474769	779	**crm310**	Chr7:50511589-50511872	283
crm311	Chr7:50730669-50731453	784	**crm312**	Chr7:50831366-50831972	606
crm313	Chr7:51089767-51090090	323	**crm314**	Chr7:51299832-51300172	340

crm315	Chr7:51631058-51631520	462	**crm316**	Chr7:51654381-51654770	389
crm317	Chr7:52075355-52075954	599	**crm318**	Chr7:52209433-52209765	332
crm319	Chr7:52221254-52221880	626	**crm320**	Chr7:52480300-52480583	283
crm321	Chr7:52528921-52529225	304	**crm322**	Chr7:52594454-52594765	311
crm323	Chr7:52603679-52604223	544	**crm324**	Chr7:52652302-52652949	647
crm325	Chr7:52654330-52654856	526	**crm326**	Chr7:52746993-52747487	494
crm327	Chr7:52826371-52826891	520	**crm328**	Chr7:52944866-52945131	265
crm329	Chr7:53049149-53049888	739	**crm330**	Chr7:53111397-53111838	441
crm331	Chr7:53158195-53158456	261	**crm332**	Chr7:53180615-53181033	418
crm333	Chr7:53208294-53208715	421	**crm334**	Chr7:53220118-53220622	504
crm335	Chr7:53445893-53446467	574	**crm336**	Chr7:53492487-53493002	515
crm337	Chr7:53838929-53839454	525	**crm338**	Chr7:54073253-54074042	789
crm339	Chr7:54221881-54222307	426	**crm340**	Chr7:54365667-54366183	516
crm341	Chr7:54400328-54400614	286	**crm342**	Chr7:54712188-54712735	547
crm343	Chr7:54870849-54871108	259	**crm344**	Chr7:55140384-55140717	333
crm345	Chr7:55229345-55230250	905	**crm346**	Chr7:57893806-57894860	1054
crm347	Chr7:62923522-62924027	505	**crm348**	Chr7:63764576-63765207	631
crm349	Chr7:63765620-63765785	165	**crm350**	Chr7:63899092-63899801	709
crm351	Chr7:63906656-63907391	735	**crm352**	Chr7:64571763-64572236	473
crm353	Chr7:64909565-64910110	545	**crm354**	Chr7:65056577-65057022	445
crm355	Chr7:65072868-65073731	863	**crm356**	Chr7:65468786-65469505	719
crm357	Chr7:66332969-66333146	177	**crm358**	Chr7:66788599-66789017	418
crm359	Chr7:67774699-67775258	559	**crm360**	Chr7:67885739-67886107	368
crm361	Chr7:67924495-67924912	417	**crm362**	Chr7:68430650-68430820	170
crm363	Chr7:68575761-68576084	323	**crm364**	Chr7:69026399-69026821	422
crm365	Chr7:69400538-69401061	523	**crm366**	Chr7:69547570-69548411	841
crm367	Chr7:69840555-69841197	642	**crm368**	Chr7:70132104-70132845	741

crm369	Chr7:70189919-70190571	652	**crm370**	Chr7:70252880-70253788	908
crm371	Chr7:70431476-70431840	364	**crm372**	Chr7:70470370-70470736	366
crm373	Chr7:71505016-71505161	145	**crm374**	Chr7:71909120-71909385	265
crm375	Chr7:72883728-72884149	421	**crm376**	Chr7:73452850-73453507	657
crm377	Chr7:74604508-74604735	227	**crm378**	Chr7:75528464-75528838	374
crm379	Chr7:75959587-75959925	338	**crm380**	Chr7:76542316-76543456	1140
crm381	Chr7:76545261-76545867	606	**crm382**	Chr7:76945707-76946165	458
crm383	Chr7:76946889-76947261	372	**crm384**	Chr7:77431547-77431879	332
crm385	Chr7:77655442-77656050	608	**crm386**	Chr7:77783541-77784265	724
crm387	Chr7:78120146-78121128	982	**crm388**	Chr7:78388568-78388880	312
crm389	Chr7:78407297-78407930	633	**crm390**	Chr7:78544882-78545400	518
crm391	Chr7:78911858-78912534	676	**crm392**	Chr7:78966663-78967361	698
crm393	Chr7:79071338-79072024	686	**crm394**	Chr7:79123565-79124213	648
crm395	Chr7:79301668-79302444	776	**crm396**	Chr7:79415217-79415320	103
crm397	Chr7:79448307-79448854	547	**crm398**	Chr7:80047217-80047596	379
crm399	Chr7:80218455-80218854	399	**crm400**	Chr7:80232776-80233249	473
crm401	Chr7:80276672-80277007	335	**crm402**	Chr7:80297525-80298105	580
crm403	Chr7:80377368-80377748	380	**crm404**	Chr7:80437090-80437674	584
crm405	Chr7:80534960-80535545	585	**crm406**	Chr7:80553986-80554532	546
crm407	Chr7:80557672-80558129	457	**crm408**	Chr7:80597244-80597616	372
crm409	Chr7:80648844-80649112	268	**crm410**	Chr7:80665081-80665695	614
crm411	Chr7:81062495-81063381	886	**crm412**	Chr7:81189848-81190759	911
crm413	Chr7:81312258-81312908	650	**crm414**	Chr7:81325061-81325414	353
crm415	Chr7:81346522-81347120	598	**crm416**	Chr7:81451919-81452032	113
crm417	Chr7:81636681-81637326	645	**crm418**	Chr7:81788581-81788988	407
crm419	Chr7:81803970-81804764	794	**crm420**	Chr7:81829862-81830615	753
crm421	Chr7:81997047-81997637	590	**crm422**	Chr7:82192039-82192465	426

crm423	Chr7:82196899-82197504	605	**crm424**	Chr7:82288889-82289991	1102
crm425	Chr7:82456484-82456962	478	**crm426**	Chr7:82497035-82497448	413
crm427	Chr7:82714121-82714513	392	**crm428**	Chr7:82911919-82912323	404
crm429	Chr7:82959278-82959473	195	**crm430**	Chr7:82984995-82985857	862
crm431	Chr7:83139162-83139691	529	**crm432**	Chr7:83184726-83185336	610
crm433	Chr7:83526446-83526899	453	**crm434**	Chr7:83635432-83636104	672
crm435	Chr7:84250087-84250578	491	**crm436**	Chr7:84342212-84342533	321
crm437	Chr7:84563053-84563650	597	**crm438**	Chr7:84704320-84704623	303
crm439	Chr7:84985865-84986444	579	**crm440**	Chr7:85041948-85042260	312
crm441	Chr7:85323130-85323788	658	**crm442**	Chr7:85965062-85965526	464
crm443	Chr7:86097908-86098387	479	**crm444**	Chr7:86240255-86241048	793
crm445	Chr7:86289718-86290106	388	**crm446**	Chr7:86846642-86847385	743
crm447	Chr7:87161116-87161440	324	**crm448**	Chr7:87173138-87173784	646
crm449	Chr7:87590805-87591519	714	**crm450**	Chr7:87599134-87599666	532
crm451	Chr7:87823199-87823622	423	**crm452**	Chr7:87979281-87979895	614
crm453	Chr7:88030461-88030631	170	**crm454**	Chr7:88386526-88387137	611
crm455	Chr7:88433909-88434864	955	**crm456**	Chr7:88546988-88547574	586
crm457	Chr7:88595201-88595720	519	**crm458**	Chr7:88854975-88855396	421
crm459	Chr7:88960405-88960529	124	**crm460**	Chr7:89137721-89138158	437
crm461	Chr7:89242439-89243047	608	**crm462**	Chr7:89333427-89333514	87
crm463	Chr7:89572863-89573266	403	**crm464**	Chr7:89596645-89597271	626
crm465	Chr7:89840164-89840610	446	**crm466**	Chr7:89984978-89985687	709
crm467	Chr7:90245386-90246035	649	**crm468**	Chr7:90567199-90567835	636
crm469	Chr7:90675093-90675857	764	**crm470**	Chr7:90727258-90727531	273
crm471	Chr7:90755441-90756057	616	**crm472**	Chr7:90773634-90774318	684
crm473	Chr7:90822342-90822929	587	**crm474**	Chr7:90951221-90951668	447
crm475	Chr7:91052806-91053014	208	**crm476**	Chr7:91146973-91147330	357

crm477	Chr7:91198629-91199051	422	**crm478**	Chr7:91441275-91441642	367
crm479	Chr7:91752246-91752768	522	**crm480**	Chr7:92036099-92036585	486
crm481	Chr7:92137924-92138522	598	**crm482**	Chr7:93080133-93080640	507
crm483	Chr7:93082477-93083001	524	**crm484**	Chr7:93108289-93108478	189
crm485	Chr7:93114299-93114986	687	**crm486**	Chr7:93163751-93164414	663
crm487	Chr7:93300911-93301172	261	**crm488**	Chr7:93563665-93564373	708
crm489	Chr7:93623458-93623735	277	**crm490**	Chr7:93695778-93696408	630
crm491	Chr7:93773306-93773515	209	**crm492**	Chr7:93791500-93792101	601
crm493	Chr7:93852952-93854024	1072	**crm494**	Chr7:93887809-93888089	280
crm495	Chr7:93952496-93952793	297	**crm496**	Chr7:94012798-94013719	921
crm497	Chr7:94144949-94145889	940	**crm498**	Chr7:94221644-94222635	991
crm499	Chr7:94225964-94226425	461	**crm500**	Chr7:94246962-94247683	721
crm501	Chr7:94283572-94284101	529	**crm502**	Chr7:94326527-94327287	760
crm503	Chr7:94671818-94672487	669	**crm504**	Chr7:94755780-94756165	385
crm505	Chr7:94895062-94895322	260	**crm506**	Chr7:94944086-94945102	1016
crm507	Chr7:95377329-95377825	496	**crm508**	Chr7:95498665-95498885	220
crm509	Chr7:95567533-95567809	276	**crm510**	Chr7:95613092-95613448	356
crm511	Chr7:95915881-95916586	705	**crm512**	Chr7:96030624-96030959	335
crm513	Chr7:96175470-96175942	472	**crm514**	Chr7:96468944-96469227	283
crm515	Chr7:96533595-96533956	361	**crm516**	Chr7:96565771-96565997	226
crm517	Chr7:96575558-96576354	796	**crm518**	Chr7:97135738-97136554	816
crm519	Chr7:97483172-97483716	544	**crm520**	Chr7:97740035-97741024	989
crm521	Chr7:98726534-98726698	164	**crm522**	Chr7:98824531-98824938	407
crm523	Chr7:99168822-99169048	226	**crm524**	Chr7:99546084-99546438	354
crm525	Chr7:100809307-100809537	230	**crm526**	Chr7:101546525-101547228	703
crm527	Chr7:102332238-102332471	233	**crm528**	Chr7:102411894-102412229	335
crm529	Chr7:102482848-102483523	675	**crm530**	Chr7:102773986-102774358	372

crm531	Chr7:103180370-103180993	623	**crm532**	Chr7:103182253-103183266	1013
crm533	Chr7:103217075-103217655	580	**crm534**	Chr7:103330758-103331558	800
crm535	Chr7:103374417-103374913	496	**crm536**	Chr7:103535690-103535956	266
crm537	Chr7:103788108-103788539	431	**crm538**	Chr7:104061346-104062172	826
crm539	Chr7:104215270-104215446	176	**crm540**	Chr7:104484721-104485153	432
crm541	Chr7:104515681-104516311	630	**crm542**	Chr7:104740259-104741060	801
crm543	Chr7:104839371-104839710	339	**crm544**	Chr7:105422605-105422831	226
crm545	Chr7:105766374-105767008	634	**crm546**	Chr7:106037926-106038293	367
crm547	Chr7:106147944-106148343	399	**crm548**	Chr7:106175609-106175941	332
crm549	Chr7:106333768-106334371	603	**crm550**	Chr7:106526783-106527272	489
crm551	Chr7:106758166-106758731	565	**crm552**	Chr7:106955710-106956173	463
crm553	Chr7:107084039-107084886	847	**crm554**	Chr7:107121249-107121599	350
crm555	Chr7:107498036-107498426	390	**crm556**	Chr7:107573513-107574253	740
crm557	Chr7:107634824-107635177	353	**crm558**	Chr7:107940507-107940600	93
crm559	Chr7:108423569-108423854	285	**crm560**	Chr7:108985040-108985219	179
crm561	Chr7:109050306-109050556	250	**crm562**	Chr7:109275183-109275910	727
crm563	Chr7:109319508-109319865	357	**crm564**	Chr7:109323683-109324245	562
crm565	Chr7:109504170-109504573	403	**crm566**	Chr7:110059045-110059471	426
crm567	Chr7:110090977-110091313	336	**crm568**	Chr7:110279938-110280473	535
crm569	Chr7:110363253-110363613	360	**crm570**	Chr7:110638205-110638772	567
crm571	Chr7:110707895-110708320	425	**crm572**	Chr7:110744194-110744837	643
crm573	Chr7:110867912-110868599	687	**crm574**	Chr7:111026317-111026754	437
crm575	Chr7:111077986-111078375	389	**crm576**	Chr7:111227222-111227990	768
crm577	Chr7:111286361-111287046	685	**crm578**	Chr7:111424124-111424214	90
crm579	Chr7:111425889-111426729	840	**crm580**	Chr7:111508947-111509473	526
crm581	Chr7:111545263-111545808	545	**crm582**	Chr7:111711574-111711982	408
crm583	Chr7:111733868-111734548	680	**crm584**	Chr7:111737848-111738185	337

crm585	Chr7:111779182-111779659	477	**crm586**	Chr7:111933685-111934319	634
crm587	Chr7:112487166-112487389	223	**crm588**	Chr7:112504491-112504805	314
crm589	Chr7:112641579-112642444	865	**crm590**	Chr7:112724551-112724893	342
crm591	Chr7:112798131-112798446	315	**crm592**	Chr7:112886832-112887404	572
crm593	Chr7:113020284-113020963	679	**crm594**	Chr7:113052119-113052318	199
crm595	Chr7:113089539-113089800	261	**crm596**	Chr7:113200965-113201340	375
crm597	Chr7:113222830-113223216	386	**crm598**	Chr7:113321872-113322523	651
crm599	Chr7:113333493-113333779	286	**crm600**	Chr7:113413370-113414055	685
crm601	Chr7:113452652-113453895	1243	**crm602**	Chr7:113801259-113801818	559
crm603	Chr7:113856811-113857184	373	**crm604**	Chr7:114074506-114075009	503
crm605	Chr7:114175343-114175890	547	**crm606**	Chr7:114177467-114178169	702
crm607	Chr7:114196119-114196689	570	**crm608**	Chr7:114210679-114211486	807
crm609	Chr7:114232214-114232631	417	**crm610**	Chr7:114491298-114491950	652
crm611	Chr7:114558151-114558892	741	**crm612**	Chr7:114733855-114734137	282
crm613	Chr7:114862086-114862601	515	**crm614**	Chr7:114876560-114877296	736
crm615	Chr7:114946658-114947569	911	**crm616**	Chr7:114964062-114964693	631
crm617	Chr7:114982907-114983216	309	**crm618**	Chr7:115192539-115193118	579
crm619	Chr7:115262250-115262661	411	**crm620**	Chr7:115309574-115309843	269
crm621	Chr7:115415726-115415954	228	**crm622**	Chr7:115550136-115550533	397
crm623	Chr7:115598280-115598779	499	**crm624**	Chr7:115665511-115666087	576
crm625	Chr7:115694861-115695569	708	**crm626**	Chr7:115893136-115893602	466
crm627	Chr7:115988159-115988530	371	**crm628**	Chr7:116085851-116086235	384
crm629	Chr7:116116309-116116587	278	**crm630**	Chr7:116305226-116305640	414
crm631	Chr7:116367841-116368403	562	**crm632**	Chr7:116379274-116379915	641
crm633	Chr7:116420252-116420901	649	**crm634**	Chr7:116487858-116488250	392
crm635	Chr7:116654947-116655357	410	**crm636**	Chr7:116718454-116718963	509
crm637	Chr7:116851027-116851944	917	**crm638**	Chr7:117003082-117003723	641

crm639	Chr7:117285454-117285971	517	**crm640**	Chr7:117356167-117357249	1082
crm641	Chr7:117538972-117539347	375	**crm642**	Chr7:117678716-117679229	513
crm643	Chr7:117681334-117681703	369	**crm644**	Chr7:117686726-117687065	339
crm645	Chr7:117959031-117959755	724	**crm646**	Chr7:117966793-117967235	442
crm647	Chr7:118004565-118005256	691	**crm648**	Chr7:118717182-118717727	545
crm649	Chr7:118813137-118813828	691	**crm650**	Chr7:118893972-118894781	809
crm651	Chr7:119017815-119018299	484	**crm652**	Chr7:119077920-119078173	253
crm653	Chr7:119144537-119144990	453	**crm654**	Chr7:119292026-119292462	436
crm655	Chr7:119358459-119359382	923	**crm656**	Chr7:119731054-119731675	621
crm657	Chr7:119903938-119904093	155	**crm658**	Chr7:120008478-120009239	761
crm659	Chr7:120059200-120059631	431	**crm660**	Chr7:120073170-120073529	359
crm661	Chr7:120281164-120281544	380	**crm662**	Chr7:120304339-120304715	376
crm663	Chr7:120644431-120645094	663	**crm664**	Chr7:120734734-120735478	744
crm665	Chr7:120884615-120885090	475	**crm666**	Chr7:121164597-121165387	790
crm667	Chr7:121519538-121519728	190	**crm668**	Chr7:121642583-121643159	576
crm669	Chr7:121660790-121661404	614	**crm670**	Chr7:121695278-121695701	423
crm671	Chr7:121778997-121779462	465	**crm672**	Chr7:122061756-122062739	983
crm673	Chr7:122078569-122078877	308	**crm674**	Chr7:122261332-122262122	790
crm675	Chr7:122290403-122290883	480	**crm676**	Chr7:122334577-122335438	861
crm677	Chr7:122425705-122426474	769	**crm678**	Chr7:122430506-122430765	259
crm679	Chr7:122755019-122755342	323	**crm680**	Chr7:122760585-122760867	282
crm681	Chr7:123083779-123084370	591	**crm682**	Chr7:123114101-123114522	421
crm683	Chr7:123378178-123378613	435	**crm684**	Chr7:123725395-123725904	509
crm685	Chr7:124058674-124059343	669	**crm686**	Chr7:124207929-124208369	440
crm687	Chr7:124519169-124519310	141	**crm688**	Chr7:124591949-124592343	394
crm689	Chr7:124682897-124683509	612	**crm690**	Chr7:124803787-124804555	768
crm691	Chr7:124986471-124986919	448	**crm692**	Chr7:125574917-125575563	646

crm693	Chr7:125813442-125813569	127	**crm694**	Chr7:125901051-125901618	567
crm695	Chr7:126313431-126313814	383	**crm696**	Chr7:126414791-126415161	370
crm697	Chr7:126436224-126436759	535	**crm698**	Chr7:126483103-126483615	512
crm699	Chr7:126545511-126546066	555	**crm700**	Chr7:126563620-126564195	575
crm701	Chr7:126607171-126607824	653	**crm702**	Chr7:126721926-126722245	319
crm703	Chr7:126786531-126786799	268	**crm704**	Chr7:126787196-126787690	494
crm705	Chr7:127072171-127072666	495	**crm706**	Chr7:127365896-127366635	739
crm707	Chr7:127471253-127471929	676	**crm708**	Chr7:127516323-127516676	353
crm709	Chr7:128503735-128504066	331	**crm710**	Chr7:128866494-128866859	365
crm711	Chr7:128924511-128924794	283	**crm712**	Chr7:129119767-129120802	1035
crm713	Chr7:129263271-129263914	643	**crm714**	Chr7:129657841-129658277	436
crm715	Chr7:130012729-130013286	557	**crm716**	Chr7:130028127-130028450	323
crm717	Chr7:130066911-130067452	541	**crm718**	Chr7:130480282-130480559	277
crm719	Chr7:130716868-130717086	218	**crm720**	Chr7:130780853-130781630	777
crm721	Chr7:131410168-131410419	251	**crm722**	Chr7:131606490-131606930	440
crm723	Chr7:131930880-131931292	412	**crm724**	Chr7:132682121-132682495	374
crm725	Chr7:132856403-132857619	1216	**crm726**	Chr7:132884525-132885219	694
crm727	Chr7:132901182-132901463	281	**crm728**	Chr7:133195481-133195841	360
crm729	Chr7:133260727-133262192	1465	**crm730**	Chr7:133338027-133338707	680
crm731	Chr7:133522481-133523183	702	**crm732**	Chr7:133633909-133634407	498
crm733	Chr7:133642248-133642963	715	**crm734**	Chr7:133661296-133661644	348
crm735	Chr7:133690043-133690139	96	**crm736**	Chr7:133725315-133725811	496
crm737	Chr7:133754879-133755311	432	**crm738**	Chr7:133993117-133993875	758
crm739	Chr7:134139457-134139858	401	**crm740**	Chr7:134469180-134469866	686
crm741	Chr7:134506933-134507352	419	**crm742**	Chr7:134545007-134545663	656
crm743	Chr7:134567295-134567572	277	**crm744**	Chr7:134615842-134616964	1122
crm745	Chr7:134878956-134879413	457	**crm746**	Chr7:135093979-135094588	609

crm747	Chr7:135137789-135138288	499	**crm748**	Chr7:135264900-135265112	212
crm749	Chr7:135369744-135370513	769	**crm750**	Chr7:135448379-135448787	408
crm751	Chr7:135525488-135525808	320	**crm752**	Chr7:135715266-135715591	325
crm753	Chr7:135754203-135754424	221	**crm754**	Chr7:135891416-135891764	348
crm755	Chr7:135936228-135937150	922	**crm756**	Chr7:136067109-136067593	484
crm757	Chr7:136092928-136093424	496	**crm758**	Chr7:136163707-136164591	884
crm759	Chr7:136393796-136394476	680	**crm760**	Chr7:137075110-137075730	620
crm761	Chr7:137200746-137201044	298	**crm762**	Chr7:137289048-137289356	308
crm763	Chr7:137425986-137426206	220	**crm764**	Chr7:137565337-137565856	519
crm765	Chr7:137797392-137797781	389	**crm766**	Chr7:137827483-137828338	855
crm767	Chr7:138261351-138261729	378	**crm768**	Chr7:138730947-138731611	664
crm769	Chr7:138862269-138862771	502	**crm770**	Chr7:139691576-139692258	682
crm771	Chr7:139787918-139788601	683	**crm772**	Chr7:139827992-139828672	680
crm773	Chr7:139854945-139855956	1011	**crm774**	Chr7:140455737-140456273	536
crm775	Chr7:140589305-140589647	342	**crm776**	Chr7:140870095-140870671	576
crm777	Chr7:141003934-141004413	479	**crm778**	Chr7:141079391-141080008	617
crm779	Chr7:141248505-141248895	390	**crm780**	Chr7:141396733-141397515	782
crm781	Chr7:141925745-141926054	309	**crm782**	Chr7:141997606-141997970	364
crm783	Chr7:142187026-142187920	894	**crm784**	Chr7:142272364-142273000	636
crm785	Chr7:142465654-142466302	648	**crm786**	Chr7:142968338-142968882	544
crm787	Chr7:143031767-143032430	663	**crm788**	Chr7:143671930-143672436	506
crm789	Chr7:143767217-143767592	375	**crm790**	Chr7:143813707-143814251	544
crm791	Chr7:143932449-143932872	423	**crm792**	Chr7:143955713-143956491	778
crm793	Chr7:144177111-144177606	495	**crm794**	Chr7:144188215-144188485	270
crm795	Chr7:144194461-144194971	510	**crm796**	Chr7:144332462-144333162	700
crm797	Chr7:144367102-144367611	509	**crm798**	Chr7:144432347-144432851	504
crm799	Chr7:144518461-144519606	1145	**crm800**	Chr7:144740528-144740961	433

crm801	Chr7:144798510-144799149	639	**crm802**	Chr7:144855079-144855647	568
crm803	Chr7:145142629-145143393	764	**crm804**	Chr7:145632566-145633253	687
crm805	Chr7:145707755-145707852	97	**crm806**	Chr7:145765936-145766431	495
crm807	Chr7:145824725-145825174	449	**crm808**	Chr7:145874383-145875056	673
crm809	Chr7:145896090-145896581	491	**crm810**	Chr7:145911950-145912598	648
crm811	Chr7:146264802-146265590	788	**crm812**	Chr7:146267355-146267713	358
crm813	Chr7:146278541-146279222	681	**crm814**	Chr7:146377603-146378178	575
crm815	Chr7:146540676-146541285	609	**crm816**	Chr7:146569712-146570150	438
crm817	Chr7:146712626-146713745	1119	**crm818**	Chr7:146831506-146831800	294
crm819	Chr7:146950069-146950618	549	**crm820**	Chr7:146974747-146975112	365
crm821	Chr7:147191443-147191461	18	**crm822**	Chr7:147523305-147523969	664
crm823	Chr7:147524355-147524620	265	**crm824**	Chr7:147869807-147870663	856
crm825	Chr7:148284414-148284734	320	**crm826**	Chr7:148291763-148292176	413
crm827	Chr7:148601647-148601986	339	**crm828**	Chr7:148624758-148625006	248
crm829	Chr7:149574633-149574896	263	**crm830**	Chr7:149756735-149757214	479
crm831	Chr7:149786679-149787200	521	**crm832**	Chr7:150415871-150416534	663
crm833	Chr7:150855390-150855541	151	**crm834**	Chr7:150911107-150911320	213
crm835	Chr7:151117370-151117874	504	**crm836**	Chr7:151609328-151609764	436
crm837	Chr7:151615201-151615527	326	**crm838**	Chr7:151945748-151946216	468
crm839	Chr7:152004519-152004683	164	**crm840**	Chr7:152881363-152881800	437
crm841	Chr7:153018329-153018870	541	**crm842**	Chr7:153147235-153147833	598
crm843	Chr7:153507728-153508066	338	**crm844**	Chr7:153649501-153650049	548
crm845	Chr7:153776653-153777380	727	**crm846**	Chr7:153857467-153858261	794
crm847	Chr7:153914067-153914499	432	**crm848**	Chr7:153919923-153920664	741
crm849	Chr7:154028699-154029111	412	**crm850**	Chr7:154088506-154088925	419
crm851	Chr7:154144444-154144645	201	**crm852**	Chr7:154750340-154750858	518
crm853	Chr7:154763401-154763758	357	**crm854**	Chr7:154860227-154860885	658

crm855	Chr7:155814564-155814994	430	**crm856**	Chr7:155965163-155965501	338
crm857	Chr7:156101434-156101735	301	**crm858**	Chr7:156148749-156149288	539
crm859	Chr7:156360367-156360887	520	**crm860**	Chr7:156380650-156381403	753
crm861	Chr7:156481814-156482459	645	**crm862**	Chr7:156583792-156585020	1228
crm863	Chr7:156989882-156990420	538	**crm864**	Chr7:158084526-158085204	678
crm865	Chr7:158372151-158372921	770	**crm866**	Chr7:158479835-158480333	498
crm867	Chr7:158612497-158612788	291	**crm868**	Chr7:158879582-158880710	1128
crm869	Chr7:158891893-158892569	676	**crm870**	Chr7:159042494-159043250	756

Table 3.5| Cis-Regulatory Modules (CRMs) predicted on negative strand of chormosome 7, keeping spacer distance 250 bps, with minimum 5 TFBS.

CRM_id	Coordinates	Size (bps)	CRM_id	Coordinates	Size (bps)
crm871	Chr7:1083457-1083872	415	**crm872**	Chr7:1281723-1282007	284
crm873	Chr7:1632674-1633285	611	**crm874**	Chr7:3167739-3168080	341
crm875	Chr7:3367522-3367955	433	**crm876**	Chr7:3473562-3474064	502
crm877	Chr7:3684581-3684958	377	**crm878**	Chr7:3971758-3972148	390
crm879	Chr7:4166639-4167323	684	**crm880**	Chr7:6378245-6378746	501
crm881	Chr7:7704800-7705016	216	**crm882**	Chr7:7775527-7776172	645
crm883	Chr7:7917151-7917408	257	**crm884**	Chr7:7963144-7963780	636
crm885	Chr7:8003380-8004180	800	**crm886**	Chr7:8005030-8005413	383
crm887	Chr7:8090711-8091036	325	**crm888**	Chr7:8166078-8166582	504
crm889	Chr7:8189175-8189655	480	**crm890**	Chr7:8514808-8515448	640
crm891	Chr7:8629975-8630752	777	**crm892**	Chr7:8637151-8637811	660
crm893	Chr7:8640137-8640438	301	**crm894**	Chr7:8989397-8989839	442
crm895	Chr7:9174306-9174909	603	**crm896**	Chr7:9490657-9491047	390
crm897	Chr7:9515821-9516078	257	**crm898**	Chr7:9638909-9639408	499
crm899	Chr7:9784010-9784298	288	**crm900**	Chr7:9976755-9977064	309
crm901	Chr7:10020664-10021376	712	**crm902**	Chr7:10232886-10233682	796
crm903	Chr7:10446913-10447332	419	**crm904**	Chr7:10679085-10679694	609
crm905	Chr7:10927947-10928253	306	**crm906**	Chr7:11070233-11070890	657
crm907	Chr7:11127255-11127628	373	**crm908**	Chr7:11144732-11145366	634
crm909	Chr7:11182530-11182826	296	**crm910**	Chr7:11215753-11216423	670
crm911	Chr7:11256382-11256951	569	**crm912**	Chr7:11449762-11450087	325
crm913	Chr7:11492918-11493355	437	**crm914**	Chr7:11660575-11661150	575
crm915	Chr7:11790916-11791384	468	**crm916**	Chr7:11888259-11888786	527

crm917	Chr7:11889862-11890646	784	**crm918**	Chr7:12092230-12093125	895
crm919	Chr7:12266098-12266713	615	**crm920**	Chr7:12564785-12565472	687
crm921	Chr7:12657044-12657522	478	**crm922**	Chr7:12809232-12809576	344
crm923	Chr7:12840820-12841314	494	**crm924**	Chr7:12914315-12914452	137
crm925	Chr7:13026836-13027258	422	**crm926**	Chr7:13606842-13607156	314
crm927	Chr7:13729719-13730307	588	**crm928**	Chr7:13776343-13776915	572
crm929	Chr7:13797218-13797767	549	**crm930**	Chr7:13868413-13868911	498
crm931	Chr7:13920690-13920887	197	**crm932**	Chr7:14180418-14180673	255
crm933	Chr7:14203731-14204375	644	**crm934**	Chr7:14366412-14366666	254
crm935	Chr7:14369948-14370401	453	**crm936**	Chr7:14654162-14654891	729
crm937	Chr7:15062536-15063297	761	**crm938**	Chr7:15227347-15227599	252
crm939	Chr7:15236892-15237155	263	**crm940**	Chr7:15342889-15343495	606
crm941	Chr7:15433806-15434472	666	**crm942**	Chr7:15733478-15734139	661
crm943	Chr7:15805148-15805654	506	**crm944**	Chr7:15817435-15817885	450
crm945	Chr7:15917806-15919339	1533	**crm946**	Chr7:15948435-15949101	666
crm947	Chr7:16196993-16197372	379	**crm948**	Chr7:16247916-16248070	154
crm949	Chr7:16648869-16649226	357	**crm950**	Chr7:17226531-17227095	564
crm951	Chr7:17268668-17268922	254	**crm952**	Chr7:17757605-17758212	607
crm953	Chr7:17845205-17845500	295	**crm954**	Chr7:18264766-18265773	1007
crm955	Chr7:18511186-18511833	647	**crm956**	Chr7:18671653-18672297	644
crm957	Chr7:18915780-18916094	314	**crm958**	Chr7:19200319-19200825	506
crm959	Chr7:19344201-19344720	519	**crm960**	Chr7:19371907-19372500	593
crm961	Chr7:19624670-19625236	566	**crm962**	Chr7:19854183-19854967	784
crm963	Chr7:20168539-20169349	810	**crm964**	Chr7:20255088-20255885	797
crm965	Chr7:20423633-20424044	411	**crm966**	Chr7:20429408-20429908	500
crm967	Chr7:20901105-20901473	368	**crm968**	Chr7:21981964-21983345	1381
crm969	Chr7:23213228-23213867	639	**crm970**	Chr7:23516184-23516558	374

crm971	Chr7:24007227-24007405	178	**crm972**	Chr7:24255758-24256248	490
crm973	Chr7:24961226-24961558	332	**crm974**	Chr7:24964087-24964739	652
crm975	Chr7:24992994-24993233	239	**crm976**	Chr7:25023729-25023996	267
crm977	Chr7:26458691-26459370	679	**crm978**	Chr7:26508158-26508707	549
crm979	Chr7:26686275-26686668	393	**crm980**	Chr7:27029153-27030345	1192
crm981	Chr7:27555868-27556609	741	**crm982**	Chr7:27717163-27717439	276
crm983	Chr7:28208961-28209557	596	**crm984**	Chr7:28453441-28454331	890
crm985	Chr7:28508884-28509108	224	**crm986**	Chr7:28556026-28556570	544
crm987	Chr7:28811085-28811581	496	**crm988**	Chr7:29295448-29296328	880
crm989	Chr7:29956377-29956721	344	**crm990**	Chr7:30203655-30204339	684
crm991	Chr7:30353265-30353659	394	**crm992**	Chr7:31205715-31205962	247
crm993	Chr7:31452931-31453464	533	**crm994**	Chr7:31668449-31669112	663
crm995	Chr7:31685629-31685833	204	**crm996**	Chr7:31825778-31826398	620
crm997	Chr7:32061251-32061659	408	**crm998**	Chr7:33077546-33078513	967
crm999	Chr7:33737951-33738221	270	**crm1000**	Chr7:33857272-33857849	577
crm1001	Chr7:34163244-34163745	501	**crm1002**	Chr7:34944410-34945318	908
crm1003	Chr7:35120204-35120485	281	**crm1004**	Chr7:35284676-35285311	635
crm1005	Chr7:35346948-35347508	560	**crm1006**	Chr7:35683471-35683709	238
crm1007	Chr7:36942687-36943309	622	**crm1008**	Chr7:37020078-37020884	806
crm1009	Chr7:37095570-37096337	767	**crm1010**	Chr7:37336326-37336833	507
crm1011	Chr7:37348022-37348533	511	**crm1012**	Chr7:37684383-37684856	473
crm1013	Chr7:38010636-38011327	691	**crm1014**	Chr7:38290886-38291129	243
crm1015	Chr7:38374173-38374594	421	**crm1016**	Chr7:38384140-38384583	443
crm1017	Chr7:38419532-38419756	224	**crm1018**	Chr7:38461209-38461807	598
crm1019	Chr7:38660323-38660809	486	**crm1020**	Chr7:38740469-38740928	459
crm1021	Chr7:38814502-38815259	757	**crm1022**	Chr7:38864433-38864885	452
crm1023	Chr7:39097739-39098306	567	**crm1024**	Chr7:39151932-39152553	621

crm1025	Chr7:39164007-39164444	437	**crm1026**	Chr7:39243215-39243527	312
crm1027	Chr7:40238790-40239077	287	**crm1028**	Chr7:40537325-40538613	1288
crm1029	Chr7:40618748-40620011	1263	**crm1030**	Chr7:40622257-40622507	250
crm1031	Chr7:40684165-40684716	551	**crm1032**	Chr7:41745566-41745987	421
crm1033	Chr7:42019150-42019697	547	**crm1034**	Chr7:42332406-42333184	778
crm1035	Chr7:43031536-43031869	333	**crm1036**	Chr7:43573198-43574304	1106
crm1037	Chr7:44085247-44085945	698	**crm1038**	Chr7:44679719-44680494	775
crm1039	Chr7:46071765-46072256	491	**crm1040**	Chr7:46253660-46254015	355
crm1041	Chr7:46670485-46670733	248	**crm1042**	Chr7:47121681-47121993	312
crm1043	Chr7:47273512-47274323	811	**crm1044**	Chr7:48283700-48284148	448
crm1045	Chr7:48775228-48775739	511	**crm1046**	Chr7:49285244-49285967	723
crm1047	Chr7:49716743-49717126	383	**crm1048**	Chr7:49952058-49952277	219
crm1049	Chr7:51108947-51109173	226	**crm1050**	Chr7:51170828-51171063	235
crm1051	Chr7:51316779-51317036	257	**crm1052**	Chr7:51732495-51732937	442
crm1053	Chr7:51735041-51735708	667	**crm1054**	Chr7:51763952-51764364	412
crm1055	Chr7:51764668-51765012	344	**crm1056**	Chr7:51821156-51821554	398
crm1057	Chr7:51845675-51847049	1374	**crm1058**	Chr7:52137182-52137965	783
crm1059	Chr7:52232895-52233378	483	**crm1060**	Chr7:52356286-52356756	470
crm1061	Chr7:52488204-52488559	355	**crm1062**	Chr7:52532033-52532451	418
crm1063	Chr7:52558091-52559097	1006	**crm1064**	Chr7:52999104-52999933	829
crm1065	Chr7:53080810-53081574	764	**crm1066**	Chr7:53083790-53084392	602
crm1067	Chr7:53134072-53134755	683	**crm1068**	Chr7:53521119-53521626	507
crm1069	Chr7:53653578-53653954	376	**crm1070**	Chr7:53870989-53871650	661
crm1071	Chr7:54444089-54444595	506	**crm1072**	Chr7:55059082-55059631	549
crm1073	Chr7:56659226-56659683	457	**crm1074**	Chr7:56798108-56798623	515
crm1075	Chr7:57145570-57146048	478	**crm1076**	Chr7:57343362-57344184	822
crm1077	Chr7:57784711-57785175	464	**crm1078**	Chr7:61671885-61672169	284

crm1079	Chr7:62711861-62712339	478	**crm1080**	Chr7:62877678-62878021	343
crm1081	Chr7:63418567-63419099	532	**crm1082**	Chr7:63868044-63868831	787
crm1083	Chr7:63886344-63886850	506	**crm1084**	Chr7:63918648-63918768	120
crm1085	Chr7:63957157-63957342	185	**crm1086**	Chr7:64033742-64033944	202
crm1087	Chr7:64312342-64313067	725	**crm1088**	Chr7:65688908-65689053	145
crm1089	Chr7:66765936-66766217	281	**crm1090**	Chr7:66913126-66913426	300
crm1091	Chr7:67611940-67612471	531	**crm1092**	Chr7:68064807-68065220	413
crm1093	Chr7:68182483-68182975	492	**crm1094**	Chr7:68532841-68533149	308
crm1095	Chr7:69080662-69081352	690	**crm1096**	Chr7:69626022-69626543	521
crm1097	Chr7:69652186-69652358	172	**crm1098**	Chr7:69705646-69706171	525
crm1099	Chr7:69732990-69733191	201	**crm1100**	Chr7:70032105-70032284	179
crm1101	Chr7:70067607-70068250	643	**crm1102**	Chr7:70238216-70238557	341
crm1103	Chr7:70583068-70584021	953	**crm1104**	Chr7:70665523-70665817	294
crm1105	Chr7:70992268-70992591	323	**crm1106**	Chr7:72411675-72412161	486
crm1107	Chr7:72710044-72710530	486	**crm1108**	Chr7:72900961-72901329	368
crm1109	Chr7:74131752-74132172	420	**crm1110**	Chr7:75248941-75249271	330
crm1111	Chr7:76482521-76483358	837	**crm1112**	Chr7:76702178-76703057	879
crm1113	Chr7:76888845-76889213	368	**crm1114**	Chr7:77768369-77768861	492
crm1115	Chr7:78191366-78191759	393	**crm1116**	Chr7:78241251-78241634	383
crm1117	Chr7:78312575-78313281	706	**crm1118**	Chr7:78957774-78958208	434
crm1119	Chr7:79401205-79402112	907	**crm1120**	Chr7:79432359-79433149	790
crm1121	Chr7:79449236-79449564	328	**crm1122**	Chr7:79631373-79632734	1361
crm1123	Chr7:79671008-79671658	650	**crm1124**	Chr7:79727570-79727890	320
crm1125	Chr7:79817459-79817676	217	**crm1126**	Chr7:79818337-79818513	176
crm1127	Chr7:79835222-79835930	708	**crm1128**	Chr7:79837000-79837535	535
crm1129	Chr7:79964710-79965486	776	**crm1130**	Chr7:80113492-80114008	516
crm1131	Chr7:80144604-80145294	690	**crm1132**	Chr7:80262266-80262755	489

crm1133	Chr7:80265622-80266329	707	**crm1134**	Chr7:80355393-80355831	438
crm1135	Chr7:80498194-80499087	893	**crm1136**	Chr7:80570551-80571488	937
crm1137	Chr7:80655550-80656021	471	**crm1138**	Chr7:80659736-80660094	358
crm1139	Chr7:80951213-80951419	206	**crm1140**	Chr7:81024316-81024949	633
crm1141	Chr7:81132586-81133060	474	**crm1142**	Chr7:81612185-81612515	330
crm1143	Chr7:81796515-81797267	752	**crm1144**	Chr7:81854222-81854485	263
crm1145	Chr7:83165567-83165850	283	**crm1146**	Chr7:83335426-83335823	397
crm1147	Chr7:83351137-83351804	667	**crm1148**	Chr7:83445227-83445731	504
crm1149	Chr7:83490969-83491549	580	**crm1150**	Chr7:83690716-83691094	378
crm1151	Chr7:83707305-83707982	677	**crm1152**	Chr7:83879488-83880193	705
crm1153	Chr7:83912454-83913097	643	**crm1154**	Chr7:83936194-83936933	739
crm1155	Chr7:84086515-84087031	516	**crm1156**	Chr7:84108021-84108402	381
crm1157	Chr7:84111281-84111721	440	**crm1158**	Chr7:84126425-84127200	775
crm1159	Chr7:84443014-84443700	686	**crm1160**	Chr7:84488214-84488985	771
crm1161	Chr7:84525862-84526159	297	**crm1162**	Chr7:84737068-84737380	312
crm1163	Chr7:84862471-84862905	434	**crm1164**	Chr7:84908252-84908762	510
crm1165	Chr7:84919489-84920283	794	**crm1166**	Chr7:84982551-84982952	401
crm1167	Chr7:85158844-85159738	894	**crm1168**	Chr7:85221244-85221570	326
crm1169	Chr7:85329640-85330274	634	**crm1170**	Chr7:85643138-85643482	344
crm1171	Chr7:85755024-85755391	367	**crm1172**	Chr7:85809491-85809754	263
crm1173	Chr7:85933786-85934087	301	**crm1174**	Chr7:86047401-86048104	703
crm1175	Chr7:86642299-86642818	519	**crm1176**	Chr7:86750571-86751042	471
crm1177	Chr7:86803536-86804298	762	**crm1178**	Chr7:86998645-86999280	635
crm1179	Chr7:87231644-87231992	348	**crm1180**	Chr7:87586299-87586653	354
crm1181	Chr7:87593415-87593764	349	**crm1182**	Chr7:87666868-87667267	399
crm1183	Chr7:87946971-87947824	853	**crm1184**	Chr7:88203987-88204669	682
crm1185	Chr7:88326980-88327672	692	**crm1186**	Chr7:88376995-88377948	953

crm1187	Chr7:88477961-88478337	376	**crm1188**	Chr7:88490107-88490593	486
crm1189	Chr7:88710344-88710695	351	**crm1190**	Chr7:90366168-90366488	320
crm1191	Chr7:90419915-90420847	932	**crm1192**	Chr7:90529513-90530261	748
crm1193	Chr7:90779426-90780283	857	**crm1194**	Chr7:90829509-90830069	560
crm1195	Chr7:91575238-91575422	184	**crm1196**	Chr7:91685785-91686339	554
crm1197	Chr7:91694890-91695393	503	**crm1198**	Chr7:91860045-91860200	155
crm1199	Chr7:91923808-91924770	962	**crm1200**	Chr7:91945595-91945978	383
crm1201	Chr7:92018857-92019813	956	**crm1202**	Chr7:92381521-92382688	1167
crm1203	Chr7:92438379-92438889	510	**crm1204**	Chr7:93239041-93239237	196
crm1205	Chr7:93252589-93252968	379	**crm1206**	Chr7:93314431-93314761	330
crm1207	Chr7:93540310-93540727	417	**crm1208**	Chr7:93541676-93542206	530
crm1209	Chr7:93607749-93608292	543	**crm1210**	Chr7:94042139-94043648	1509
crm1211	Chr7:94657876-94658031	155	**crm1212**	Chr7:94658649-94659089	440
crm1213	Chr7:94906217-94906727	510	**crm1214**	Chr7:94908957-94909448	491
crm1215	Chr7:94913854-94914412	558	**crm1216**	Chr7:95002084-95002635	551
crm1217	Chr7:95207923-95208274	351	**crm1218**	Chr7:95258407-95258747	340
crm1219	Chr7:95437388-95437679	291	**crm1220**	Chr7:95605249-95605557	308
crm1221	Chr7:95824348-95825002	654	**crm1222**	Chr7:96179242-96179622	380
crm1223	Chr7:96563724-96564249	525	**crm1224**	Chr7:96583923-96584380	457
crm1225	Chr7:96631827-96632491	664	**crm1226**	Chr7:97230120-97230621	501
crm1227	Chr7:97429549-97429977	428	**crm1228**	Chr7:97488739-97489419	680
crm1229	Chr7:97809583-97809740	157	**crm1230**	Chr7:98508345-98509046	701
crm1231	Chr7:99348630-99349726	1096	**crm1232**	Chr7:102432660-102433169	509
crm1233	Chr7:102632073-102632673	600	**crm1234**	Chr7:102729719-102730266	547
crm1235	Chr7:102941519-102942109	590	**crm1236**	Chr7:103195583-103195927	344
crm1237	Chr7:103329267-103329987	720	**crm1238**	Chr7:103967316-103967735	419
crm1239	Chr7:104083918-104084268	350	**crm1240**	Chr7:104268447-104269340	893

crm1241	Chr7:105751199-105751901	702	**crm1242**	Chr7:106028493-106029161	668
crm1243	Chr7:106226019-106226393	374	**crm1244**	Chr7:106232735-106232969	234
crm1245	Chr7:107245874-107246409	535	**crm1246**	Chr7:107690999-107691291	292
crm1247	Chr7:107769007-107769412	405	**crm1248**	Chr7:108387295-108387724	429
crm1249	Chr7:108489267-108489692	425	**crm1250**	Chr7:108677560-108678080	520
crm1251	Chr7:108756686-108757289	603	**crm1252**	Chr7:109214802-109215776	974
crm1253	Chr7:109343905-109344813	908	**crm1254**	Chr7:109947468-109947909	441
crm1255	Chr7:110127874-110128264	390	**crm1256**	Chr7:110302604-110302926	322
crm1257	Chr7:110544941-110545403	462	**crm1258**	Chr7:110606572-110606966	394
crm1259	Chr7:110729030-110729363	333	**crm1260**	Chr7:110804296-110804705	409
crm1261	Chr7:111159442-111159871	429	**crm1262**	Chr7:111624532-111625202	670
crm1263	Chr7:111884729-111885005	276	**crm1264**	Chr7:111914596-111915155	559
crm1265	Chr7:111915933-111916428	495	**crm1266**	Chr7:111974244-111974957	713
crm1267	Chr7:112112484-112113356	872	**crm1268**	Chr7:112125644-112126422	778
crm1269	Chr7:112745922-112746442	520	**crm1270**	Chr7:113060547-113061098	551
crm1271	Chr7:113092994-113093330	336	**crm1272**	Chr7:113093843-113094089	246
crm1273	Chr7:113403862-113404432	570	**crm1274**	Chr7:113531343-113531484	141
crm1275	Chr7:113544703-113545529	826	**crm1276**	Chr7:113714673-113715229	556
crm1277	Chr7:113728662-113729247	585	**crm1278**	Chr7:113770475-113770704	229
crm1279	Chr7:114072563-114072863	300	**crm1280**	Chr7:114128564-114129117	553
crm1281	Chr7:114259457-114259678	221	**crm1282**	Chr7:114413911-114414504	593
crm1283	Chr7:114455167-114455594	427	**crm1284**	Chr7:114482917-114483411	494
crm1285	Chr7:114956736-114957061	325	**crm1286**	Chr7:115271386-115271913	527
crm1287	Chr7:115722590-115723133	543	**crm1288**	Chr7:115805243-115805903	660
crm1289	Chr7:115873013-115873434	421	**crm1290**	Chr7:116212527-116212849	322
crm1291	Chr7:116754238-116754692	454	**crm1292**	Chr7:116915747-116916152	405
crm1293	Chr7:117267072-117267346	274	**crm1294**	Chr7:117406011-117406488	477

crm1295	Chr7:117680666-117681123	457	**crm1296**	Chr7:117750680-117751150	470
crm1297	Chr7:118254110-118254651	541	**crm1298**	Chr7:118713677-118713996	319
crm1299	Chr7:118766839-118767159	320	**crm1300**	Chr7:118890177-118890899	722
crm1301	Chr7:118901646-118902498	852	**crm1302**	Chr7:118936103-118936451	348
crm1303	Chr7:119038758-119038959	201	**crm1304**	Chr7:119126748-119127423	675
crm1305	Chr7:119313684-119314168	484	**crm1306**	Chr7:119377924-119378569	645
crm1307	Chr7:119997459-119997627	168	**crm1308**	Chr7:120319318-120319529	211
crm1309	Chr7:120344959-120345382	423	**crm1310**	Chr7:120375959-120376376	417
crm1311	Chr7:120685828-120686432	604	**crm1312**	Chr7:120775966-120776349	383
crm1313	Chr7:120778838-120779283	445	**crm1314**	Chr7:120812158-120812580	422
crm1315	Chr7:120879415-120880106	691	**crm1316**	Chr7:120928310-120928823	513
crm1317	Chr7:120982903-120983591	688	**crm1318**	Chr7:121124428-121125177	749
crm1319	Chr7:121443823-121444483	660	**crm1320**	Chr7:122106158-122106587	429
crm1321	Chr7:122108059-122108503	444	**crm1322**	Chr7:122287147-122287687	540
crm1323	Chr7:122353090-122353823	733	**crm1324**	Chr7:122366428-122366907	479
crm1325	Chr7:122520899-122521107	208	**crm1326**	Chr7:122603013-122603412	399
crm1327	Chr7:122646493-122646827	334	**crm1328**	Chr7:122679377-122679685	308
crm1329	Chr7:123711185-123711843	658	**crm1330**	Chr7:123856912-123857165	253
crm1331	Chr7:123886068-123886483	415	**crm1332**	Chr7:124483336-124483963	627
crm1333	Chr7:124722146-124722875	729	**crm1334**	Chr7:125014095-125015496	1401
crm1335	Chr7:125050034-125050525	491	**crm1336**	Chr7:125435952-125436271	319
crm1337	Chr7:125536599-125537105	506	**crm1338**	Chr7:125588759-125589530	771
crm1339	Chr7:125913930-125914443	513	**crm1340**	Chr7:125926616-125926934	318
crm1341	Chr7:125952820-125953552	732	**crm1342**	Chr7:126003275-126003948	673
crm1343	Chr7:126099546-126099946	400	**crm1344**	Chr7:126297567-126297988	421
crm1345	Chr7:126655399-126656331	932	**crm1346**	Chr7:126739351-126739829	478
crm1347	Chr7:126927922-126928830	908	**crm1348**	Chr7:127156808-127157415	607

crm1349	Chr7:127504246-127504738	492	**crm1350**	Chr7:127527924-127528528	604
crm1351	Chr7:127622907-127623392	485	**crm1352**	Chr7:127663462-127663698	236
crm1353	Chr7:128172306-128173570	1264	**crm1354**	Chr7:128228164-128228475	311
crm1355	Chr7:130416460-130417106	646	**crm1356**	Chr7:130823543-130824158	615
crm1357	Chr7:130917264-130917800	536	**crm1358**	Chr7:131060381-131060762	381
crm1359	Chr7:131130823-131131281	458	**crm1360**	Chr7:131476916-131477465	549
crm1361	Chr7:132238449-132238976	527	**crm1362**	Chr7:132242379-132242502	123
crm1363	Chr7:132748809-132749452	643	**crm1364**	Chr7:132995966-132996413	447
crm1365	Chr7:133154513-133154911	398	**crm1366**	Chr7:133431796-133432239	443
crm1367	Chr7:133445782-133446123	341	**crm1368**	Chr7:133931497-133931909	412
crm1369	Chr7:134600060-134600292	232	**crm1370**	Chr7:134718448-134718990	542
crm1371	Chr7:134720701-134721105	404	**crm1372**	Chr7:134783743-134784082	339
crm1373	Chr7:135421388-135421742	354	**crm1374**	Chr7:136171493-136172448	955
crm1375	Chr7:136189829-136190100	271	**crm1376**	Chr7:136451601-136452007	406
crm1377	Chr7:136545769-136546085	316	**crm1378**	Chr7:136561736-136562275	539
crm1379	Chr7:136564522-136565251	729	**crm1380**	Chr7:136944316-136944631	315
crm1381	Chr7:137036825-137037286	461	**crm1382**	Chr7:137077847-137078892	1045
crm1383	Chr7:137277810-137278180	370	**crm1384**	Chr7:137391683-137391955	272
crm1385	Chr7:137399041-137399905	864	**crm1386**	Chr7:137628921-137629602	681
crm1387	Chr7:138140795-138140897	102	**crm1388**	Chr7:138279834-138280360	526
crm1389	Chr7:138574645-138574811	166	**crm1390**	Chr7:138765493-138765694	201
crm1391	Chr7:138979659-138980375	716	**crm1392**	Chr7:139026050-139026354	304
crm1393	Chr7:139095403-139095739	336	**crm1394**	Chr7:139341109-139341457	348
crm1395	Chr7:140024293-140024854	561	**crm1396**	Chr7:140714521-140715168	647
crm1397	Chr7:141019000-141019203	203	**crm1398**	Chr7:141328987-141329349	362
crm1399	Chr7:141477862-141478536	674	**crm1400**	Chr7:141618233-141618662	429
crm1401	Chr7:141731178-141731632	454	**crm1402**	Chr7:142541671-142542140	469

crm1403	Chr7:143748555-143749624	1069	**crm1404**	Chr7:143900028-143900338	310
crm1405	Chr7:144064900-144065142	242	**crm1406**	Chr7:144104292-144104940	648
crm1407	Chr7:144167729-144168465	736	**crm1408**	Chr7:144741490-144742018	528
crm1409	Chr7:144995043-144995634	591	**crm1410**	Chr7:145425562-145425989	427
crm1411	Chr7:145687985-145688279	294	**crm1412**	Chr7:145719263-145719961	698
crm1413	Chr7:146064528-146065307	779	**crm1414**	Chr7:146118112-146118644	532
crm1415	Chr7:146334924-146335552	628	**crm1416**	Chr7:146465726-146466123	397
crm1417	Chr7:146793942-146794379	437	**crm1418**	Chr7:146829008-146829990	982
crm1419	Chr7:146976750-146977256	506	**crm1420**	Chr7:147039586-147040101	515
crm1421	Chr7:147048620-147049068	448	**crm1422**	Chr7:147140918-147141446	528
crm1423	Chr7:147283966-147284268	302	**crm1424**	Chr7:147623179-147623664	485
crm1425	Chr7:147784122-147784504	382	**crm1426**	Chr7:147983420-147983854	434
crm1427	Chr7:148015709-148016419	710	**crm1428**	Chr7:148032684-148033421	737
crm1429	Chr7:148070821-148071053	232	**crm1430**	Chr7:148479265-148479354	89
crm1431	Chr7:148610335-148610614	279	**crm1432**	Chr7:149155493-149155797	304
crm1433	Chr7:149428609-149428757	148	**crm1434**	Chr7:149743046-149743413	367
crm1435	Chr7:149748705-149748941	236	**crm1436**	Chr7:149864478-149864944	466
crm1437	Chr7:149879753-149880358	605	**crm1438**	Chr7:150824769-150825179	410
crm1439	Chr7:151791045-151791693	648	**crm1440**	Chr7:151886048-151886741	693
crm1441	Chr7:152768942-152769600	658	**crm1442**	Chr7:153260598-153260782	184
crm1443	Chr7:153614336-153614702	366	**crm1444**	Chr7:154681209-154681592	383
crm1445	Chr7:154770663-154771207	544	**crm1446**	Chr7:155530920-155531434	514
crm1447	Chr7:155921304-155922207	903	**crm1448**	Chr7:156146412-156147083	671
crm1449	Chr7:156557440-156557737	297	**crm1450**	Chr7:157576725-157577254	529
crm1451	Chr7:157842625-157842957	332	**crm1452**	Chr7:158495137-158495506	369
crm1453	Chr7:158598045-158598592	547			

3.5 Chromosome 7 VISTA Enhancers

To compare and validate TFBS based CRM prediction, functionally tested enhancers present on chromosome 7 were acquired. There were total 133 enhancers present on human chromosome 7 having expressions in various tissues like forebrain, mid brain, eye, ear etc. Any enhancer can be positive if it shows reproducible expression in same tissue among more than three embryos, where as a negative enhancers fails to exhibit reproducible expression pattern in any three mice embryos out of the five tested. Table 3.6 summarizes detailed information about VISTA enhancers on human chromosome 7. Each row describes element id, coordinates and expression pattern of enhancer as either negative or positive. For each positive enhancer, information of expression tissue is also provided in table. Since enhancers show spatiotemporally specific expression, so each enhancer can show activity in multiple tissues. For each positive enhancer, all expression tissues are mentioned along with the probability of expression $p(E)$ seen according to formula mentioned below.

$$p(E) = \frac{no\ of\ expressed\ embryos}{Total\ no\ of\ embryos\ tested}$$

Tissue wise distribution of these experimentally verified enhancers is depicted in figure3.2, depicting highest number of enhancers expressing in brain. Tissue category brain in this figure includes all enhancers reported in forebrain, midbrain and hindbrain.

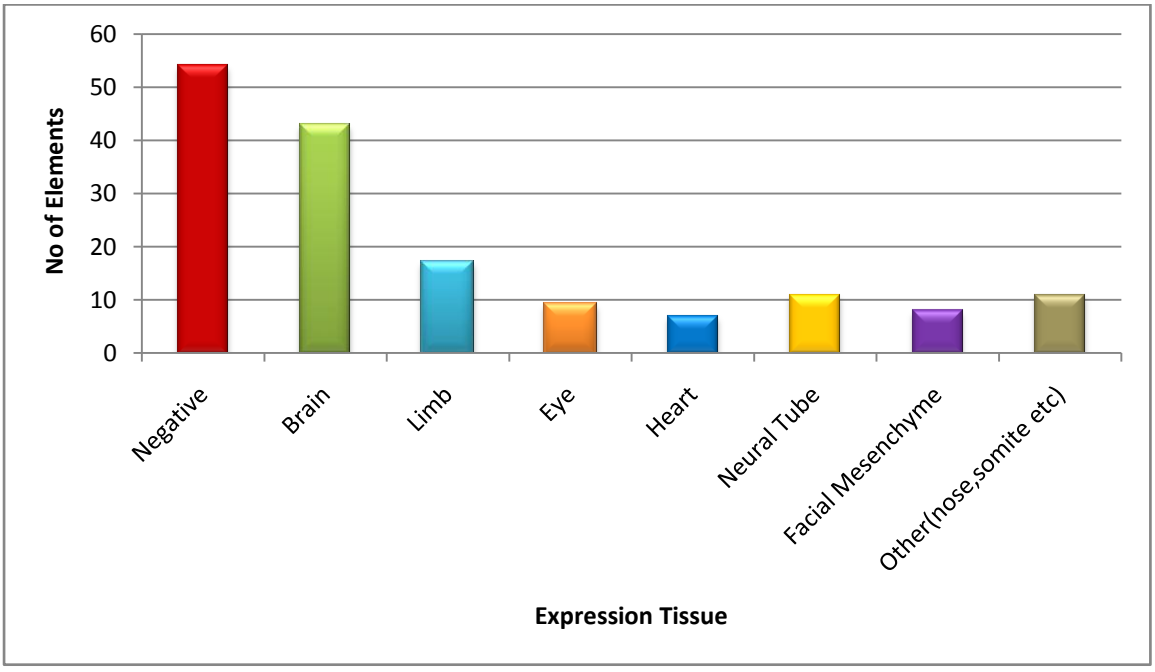

Figure 3.2|Tissue wise distribution of VISTA enhancers on human chromosome 7.

Since our concerned tissue in this study was forebrain, to assess sensitivity of our model (in later stages), we filtered out elements expressing exclusively in forebrain (figure 3.3). Elements expressing in forebrain as well as other tissues are represented as forebrain inclusive elements in figure 3.3. Remaining positive elements exhibiting expression in tissues other than forebrain are represented as others.

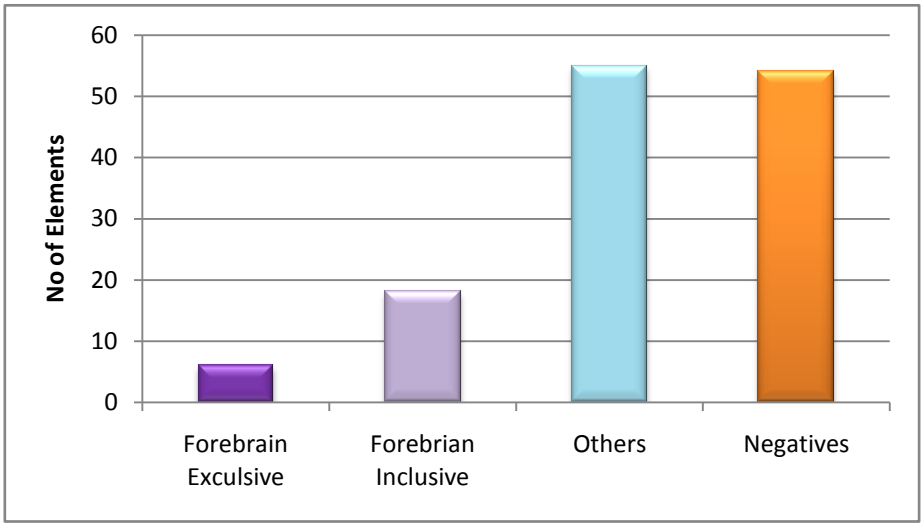

Figure 3.3|Forebrain wise distribution of VISTA enhancers.

Table 3.6 | VISTA elements residing on human chromosme 7.

Sr.No	VISTA Element ID	Coordinates	Expression	Expression Tissue
1	element 110	chr7:21003280-21004750	positive	forebrain[4/4]
2	element 111	chr7:42191728-42193638	positive	forebrain[4/4]
3	element 173	chr7:1267832-1269169	negative	
4	element 190	chr7:115313238-115315056	negative	
5	element 218	chr7:114056847-114058647	positive	forebrain[5/7]
6	element 219	chr7:114062615-114063919	negative	
7	element 220	chr7:114070779-114073266	negative	
8	element 221	chr7:114208952-114210310	negative	
9	element 222	chr7:114295109-114296373	positive	heart[3/6]
10	element 239	chr7:20829390-20830407	negative	
11	element 293	chr7:1265154-1266318	positive	midbrain (mesencephalon)[8/10] , forebrain[8/10] , branchial arch[3/10]
12	element 294	chr7:21811370-21812002	positive	neural tube[3/5]
13	element 296	chr7:26728697-26729802	positive	hindbrain (rhombencephalon)[4/5]
14	element 297	chr7:31401914-31403016	positive	melanocytes[4/5]
15	element 298	chr7:96633582-96634303	positive	midbrain (mesencephalon)[3/6]
16	element 299	chr7:115115427-115117325	negative	
17	element 300	chr7:115133829	negative	

		-115135428		
18	element 301	chr7:115582121 -115582622	negative	
19	element 461	chr7:35454800-35456218	positive	eye[9/9]
20	element 463	chr7:35505184-35506342	positive	heart[5/6]
21	element 464	chr7:35320121-35321085	negative	
22	element 465	chr7:35590614-35591629	negative	
23	element 466	chr7:35268514-35269380	positive	facial mesenchyme[3/4]
24	element 467	chr7:35275271-35276121	negative	
25	element 468	chr7:35331175-35332120	negative	
26	element 469	chr7:35359004-35359854	negative	
27	element 470	chr7:35236863-35237925	negative	
28	element 471	chr7:35254518-35255251	negative	
29	element 472	chr7:35458551-35459522	positive	hindbrain (rhombencephalon)[5/6] , eye[6/6] , heart[6/6]
30	element 473	chr7:35342478-35343110	negative	
31	element 474	chr7:35388314-35389035	negative	
32	element 501	chr7:27249858-27251057	negative	
33	element 550	chr7:13506207-13507276	positive	forebrain[3/5] , branchial arch[3/5]
34	element 629	chr7:27288268-27289491	positive	dorsal root ganglion[4/8]
35	element 635	chr7:50449432-	negative	

		50450191		
36	element 644	chr7:18885971-18887230	positive	hindbrain (rhombencephalon)[4/9] , forebrain[6/9] , limb[6/9]
37	element 658	chr7:69702231-69703615	positive	neural tube[9/15] , midbrain (mesencephalon)[11/15]
38	element 684	chr7:115113551-115114543	negative	
39	element 701	chr7:21084342-21085460	positive	midbrain (mesencephalon)[4/8]
40	element 720	chr7:113793545-113794562	positive	midbrain (mesencephalon)[5/7]
41	element 738	chr7:115295267-115296597	negative	
42	element 749	chr7:13450920-13451719	positive	hindbrain (rhombencephalon)[9/10] , midbrain (mesencephalon)[9/10]
43	element 769	chr7:26524662-26525636	positive	neural tube[5/7]
44	element 785	chr7:114381553-114382837	negative	
45	element 792	chr7:114395170-114396591	negative	
46	element 793	chr7:10684318-10685359	positive	midbrain (mesencephalon)[10/13]
47	element 799	chr7:9271308-9272358	positive	forebrain[12/13]
48	element 807	chr7:22091362-22092557	positive	forebrain[8/9]
49	element 816	chr7:14379627-14380740	positive	hindbrain (rhombencephalon)[9/10]
50	element 827	chr7:69881761-69882691	negative	
51	element 833	chr7:69980613-69981299	negative	
52	element 844	chr7:20832628-20833902	positive	forebrain[5/6] , facial mesenchyme[3/6]
53	element 846	chr7:69891243-	negative	

		69892028		
54	element 861	chr7:115262750 -115264769	negative	
55	element 877	chr7:50219464- 50220288	negative	
56	element 925	chr7:70103123- 70104331	positive	neural tube[8/15] , limb[8/15]
57	element 954	chr7:114287987 -114290557	negative	
58	element 956	chr7:114299711 -114302078	positive	midbrain (mesencephalon)[3/7] , forebrain[6/7]
59	element 961	chr7:50333048- 50334464	negative	
60	element 966	chr7:114326912 -114329772	positive	hindbrain (rhombencephalon)[6/11]
61	element 972	chr7:50357638- 50358644	negative	
62	element 999	chr7:114142132 -114143527	positive	midbrain (mesencephalon)[4/6] , facial mesenchyme[4/6]
63	element 1007	chr7:20997668- 20999102	positive	forebrain[3/3] , facial mesenchyme[3/3]
64	element 1019	chr7:20838843- 20840395	positive	forebrain[4/6]
65	element 1069	chr7:114052100 -114053403	negative	
66	element 1080	chr7:114261073 -114263089	positive	ear[3/5]
67	element 1132	chr7:42432266- 42433365	negative	
68	element 1148	chr7:21019551- 21021200	positive	limb[4/4]
69	element 1167	chr7:114055419 -114056748	negative	
70	element 1192	chr7:114463797 -114464462	positive	hindbrain (rhombencephalon)[5/5] , midbrain (mesencephalon)[3/5]
71	element 1213	chr7:42252831-	positive	hindbrain (rhombencephalon)[5/6] ,

		42254560		midbrain (mesencephalon)[5/6] , forebrain[3/6]
72	element 1215	chr7:114292900 -114293972	negative	
73	element 1223	chr7:21239775-21240922	negative	
74	element 1226	chr7:21080801-21082086	positive	neural tube[12/16] , hindbrain (rhombencephalon)[12/16] , midbrain (mesencephalon)[13/16] , forebrain[14/16] , eye[12/16] , cranial nerve[10/16]
75	element 1268	chr7:155547616 -155548788	positive	limb[4/6]
76	element 1308	chr7:127174386 -127177546	positive	midbrain (mesencephalon)[7/7] , forebrain[7/7]
77	element 1325	chr7:25791903-25794282	positive	hindbrain (rhombencephalon)[5/5] , midbrain (mesencephalon)[5/5] , forebrain[5/5]
78	element 1336	chr7:34097962-34100011	positive	neural tube[4/6] , forebrain[6/6]
79	element 1401	chr7:108014729 -108017176	negative	
80	element 1418	chr7:155264047 -155265809	positive	hindbrain (rhombencephalon)[4/5] , midbrain (mesencephalon)[4/5]
81	element 1425	chr7:69596979-69598263	positive	midbrain (mesencephalon)[4/6]
82	element 1430	chr7:28114903-28117771	positive	limb[11/12]
83	element 1465	chr7:25919000-25922541	positive	somite[9/9] , limb[9/9]
84	element 1508	chr7:135456908 -135459236	negative	
85	element 1586	chr7:42185602-42187508	positive	limb[3/3]
86	element 1600	chr7:25695037-25697616	positive	forebrain[4/5] , limb[3/5] , branchial arch[4/5] , nose[4/5] , facial mesenchyme[4/5]

87	element 1604	chr7:28630662-28634883	positive	limb[10/10] , branchial arch[10/10] , nose[10/10] , genital tubercle[9/10] , facial mesenchyme[10/10]
88	element 1623	chr7:5671918-5676544	positive	midbrain (mesencephalon)[4/5] , branchial arch[5/5] , facial mesenchyme[5/5] , other[4/5]
89	element 1626	chr7:93975962-93978880	positive	branchial arch[6/6] , facial mesenchyme[6/6]
90	element 1627	chr7:90753895-90758080	positive	midbrain (mesencephalon)[20/20] , forebrain[11/20] , limb[13/20] , eye[18/20] , ear[18/20]
91	element 1631	chr7:90797871-90800712	positive	somite[10/13]
92	element 1633	chr7:90777214-90780836	positive	forebrain[8/8]
93	element 1642	chr7:95874847-95878101	positive	neural tube[7/8] , hindbrain (rhombencephalon)[7/8] , forebrain[8/8]
94	element 1659	chr7:35412725-35416349	positive	heart[9/9]
95	element 1660	chr7:69399070-69400682	positive	ear[5/8]
96	element 1677	chr7:25533607-25536880	positive	neural tube[6/6] , hindbrain (rhombencephalon)[6/6] , midbrain (mesencephalon)[3/6]
97	element 1703	chr7:27156051-27156664	negative	
98	element 1740	chr7:16043133-16045391	negative	
99	element 1753	chr7:158888320-158891362	positive	heart[14/18]
100	element 1798	chr7:115451531-115454796	positive	neural tube[6/9] , hindbrain (rhombencephalon)[6/9] , midbrain (mesencephalon)[6/9] , forebrain[9/9] , eye[6/9]
101	element 1809	chr7:121967528-121971078	positive	neural tube[5/9] , hindbrain (rhombencephalon)[7/9] , midbrain

				(mesencephalon)[7/9] , forebrain[9/9]
102	element 1831	chr7:95236622-95240458	positive	heart[10/15] , other[11/15]
103	element 1876	chr7:113558123-113559105	negative	
104	element 1878	chr7:19157694-19159713	negative	
105	element 1880	chr7:19146921-19150251	negative	
106	element 1908	chr7:150665549-150670065	negative	
107	element 1947	chr7:2651286-2654820	negative	
108	element 1988	chr7:120536767-120540791	negative	
109	element 2059	chr7:50564242-50565706	positive	heart[8/9]
110	element 2071	chr7:115993663-115997109	positive	other[6/9]
111	element 2077	chr7:135742628-135745716	positive	ear[5/9]
112	element 2079	chr7:42532168-42534364	negative	
113	element 2081	chr7:107587162-107588892	positive	dorsal root ganglion[4/9] , limb[5/9] , trigeminal V (ganglion, cranial)[6/9] , cranial nerve[5/9]
114	element 2082	chr7:139910588-139914247	positive	dorsal root ganglion[3/7] , trigeminal V (ganglion, cranial)[3/7]
115	element 2198	chr7:151439778-151443512	negative	
116	element 2200	chr7:151453949-151457105	positive	midbrain (mesencephalon)[8/12]
117	element 2249	chr7:99594233-99596427	negative	
118	element 2258	chr7:147972273-147975381	negative	

119	element 2306	chr7:18868716-18870405	positive	limb[4/4]
120	element 2307	chr7:18874272-18875369	positive	limb[2/2] , branchial arch[2/2]
121	element 2308	chr7:95705089-95705696	positive	limb[8/8] , genital tubercle[4/8]
122	element 2309	chr7:95726279-95727680	positive	limb[3/3]
123	element 2310	chr7:96221075-96221574	negative	
124	element 2311	chr7:96124918-96125415	positive	branchial arch[8/11]
125	element 2312	chr7:96075405-96075991	positive	branchial arch[3/3]
126	element 2313	chr7:95765880-95766473	positive	forebrain[2/2] , limb[2/2] , branchial arch[2/2] , ear[2/2]
127	element 2314	chr7:95696530-95697029	negative	
128	element 2315	chr7:69325868-69327684	positive	hindbrain (rhombencephalon)[4/13] , midbrain (mesencephalon)[12/13] , forebrain[5/13] , eye[11/13]
129	element 2316	chr7:69646374-69649029	positive	eye[3/5]
130	element 2317	chr7:69720457-69721867	positive	eye[6/6] , other[3/6]
131	element 2318	chr7:69731835-69734096	positive	Neural tube[4/4] , hindbrain (rhombencephalon) [4/4] , midbrain (mesencephalon) [4/4] , forebrain [4/4]
132	element 2387	chr7:130012949-130014460	negative	
133	element 2419	chr7:82039621-82041108	positive	somite[12/12] ,limb[11/12], branchial arch [9/12] , eye[10/12]

3.6 Results Validation

Comparison of our predicated CRMs with functionally tested VISTA Enhancers from chromosome 7 reveled 8 experimentally validated VISTA enhancers overlapping with CRMs. Out of total eight, 2 enhancers were forebrain exclusive while 2 were forebrain inclusive enhancers. Following table 3.7 provides more information about these results.

Table 3.7| VISTA Enhancers overlapping with CRMs (CRMs bearing * represent overlap with forebrain exclusive enhancer).

No.	CRM Id	Coordinates	Vista Element Id	Expression Tissue
1	**crm44**	chr7:10685350-10685359	element 793	Midbrain
2	**crm128***	chr7:21003280-21003638	element 110	Forebrain
3	**crm365**	chr7:69400538-69400682	element 1660	Ear
4	**crm1099**	chr7:69732990-69733191	element 2318	Forebrain(4/4 embryos), neural tube, midbrain, hindbrain
5	**crm471**	chr7:90755441-90756057	element 1627	Forebrain (11/20 embryos) midbrain, eye, ear
6	**crm1193***	chr7:90779426-90780283	element 1633	Forebrain (8/8 embryos)
7	**crm1279**	chr7:114072563-114072863	element 220	Negative
8	**crm715**	chr7:130012949-130013286	element 2387	Negative

3.7 CRMs in Gene Deserts

Total of 20 gene deficient regions were identified on chromosome 7 which exceeded in size from 500kbps. Generally, these regions had high repeat content and low CpG density. Table 3.8 encompasses CRMs located in these gene deserts along with other characteristics of these regions.

Table 3.8|Gene Deserts on chromosome 7 and CRMs residing in gene deserts

Gene Desert	Location	Size (Bps)	No. of over lapping CRMs	Overlapping CRM Ids	Flanking Reference Genes		Genes and Models in Region	CpG Island	Repetitive Content	
									LINEs	SINEs
1	7q11.22-q11.23	2,277,393	12	crm358, crm1090, crm1091, crm359, crm360, crm361, crm1092, crm1093, crm362, crm1094, crm363, crm364	FLJ13195/ STAG3L4	AUTS2	2 predicted	1	10.8%	25.8%
2	7q31.1	2,089,310	19	crm1248, crm559, crm1249, crm1250, crm1251, crm560, crm561, crm1252, crm562, crm563, crm564, crm1253, crm565, crm1254, crm566, crm567, crm1255, crm568, crm1256	THC12014 70/ DNAJB9	IMMP2L	3 predicted; 1 pseudogene	0	32.9%	5.2%
3	7p12.2-p12.1	1,718,821	30	crm315, crm316, crm1052, crm1053, crm1054, crm1055, crm1056, crm1057, crm317, crm1058, crm318, crm319, crm1059, crm1060, crm320, crm1061, crm321, crm1062, crm1063, crm322, crm323, crm324, crm325, crm326, crm327, crm328, crm1064, crm329, crm1065, crm1066	KIAA0633/ COBL	FLJ40449/ POM121L12	2 predicted; 1 putative; 1 pseudogene	4	26.0%	8.1%
4	7p22.1-p21.3	2,026,979	31	crm25, crm894, crm895, crm26, crm27, crm28, crm29, crm30, crm31, crm896, crm897, crm32, crm33, crm34, crm898, crm35, crm899, crm36, crm37, crm900, crm901, crm902, crm38, crm39, crm40, crm903, crm41, crm42, crm43, crm904, crm44	NXPH1	IMAGE:3605453/BC002644	2 predicted; 1 putative; 1 pseudogene	3	31.2%	5.3%

5	7q31.31-q31.32	2,030,939	23	crm645, crm646, crm647, crm1297, crm1298, crm648, crm1299, crm649, crm1300, crm650, crm1301, crm1302, crm651, crm1303, crm652, crm1304, crm653, crm654, crm1305, crm655, crm1306, crm656, crm657	ANKRD7	hCT1816883/ KCND2	1 pseudogene	2	32.2%	4.2%
6	7p21.3	1,200,299	19	crm922, crm923, crm57, crm924, crm58, crm925, crm59, crm60, crm61, crm62, crm926, crm927, crm63, crm928, crm929, crm930, crm64, crm65, crm931	ARL4	ETV1	1 pseudogene	2	24.0%	8.7%
7	7q21.11-q21.13	1,601,976	20	crm438, crm1162, crm1163, crm1164, crm1165, crm1166, crm439, crm440, crm1167, crm1168, crm441, crm1169, crm1170, crm1171, crm1172, crm1173, crm442, crm1174, crm443, crm444	IMAGE:5272175/ SEMA3D	GRM3	1 predicted	1	33.8%	4.9%
8	7p12.3-p12.2	1,445,770	12	crm300, crm301, crm1045, crm302, crm303, crm304, crm1046, crm305, crm306, crm1047, crm307, crm1048	MGC26484 /ABCA13	ZPBP	none	1	31.0%	6.0%
9	7q31.33-q31.2	1,508,616	17	crm688, crm689, crm1333, crm690, crm691, crm1334, crm1335, crm1336, crm1337, crm692, crm1338, crm693, crm694, crm1339, crm1340, crm1341, crm1342	THC1079110/POT1	GRM8	1 predicted	0	28.7%	5.1%
10	7q36.1-q36.2	1,031,719	6	crm1441, crm840, crm841, crm842, crm1442, crm843	ARP3BETA/ACTR3B	DPP6	2 predicted	2	24.3%	8.4%

11	7p13-p12.3	1,353,882	12	crm1039, crm1040, crm284, crm285, crm1041, crm286, crm287, crm288, crm1042, crm289, crm290, crm1043	IGFBP3	PRO1866/ TNS3	2 predicted; 1 pseudogene	1	31.4%	5.8%
12	7q31.2-q31.31	915,233	13	crm612, crm613, crm614, crm615, crm1285, crm616, crm617, crm618, crm619, crm1286, crm620, crm621, crm622	IMAGE:4276820/ MDFIC	TFEC	1 putative	0	19.9%	6.1%
13	7q21.2	817,344	5	crm460, crm461, crm462, crm463, crm464	FLJ32110/ ZNF804B	IMAGE:5295327/ STEAP1	1 predicted ; 1 pseudogene	1	33.6%	5.2%
14	7q31.2	789,050	16	crm1269, crm591, crm592, crm593, crm594, crm1270, crm595, crm1271, crm1272, crm596, crm597, crm598, crm599, crm1273, crm600, crm601	GPR85	PPP1R3/ PPP1R3A	none	1	31.4%	5.1%
15	7q35	1,280,308	12	crm800, crm1408, crm801, crm802, crm1409, crm803, crm1410, crm804, crm1411, crm805, crm1412, crm806	TPK1	THC12035 97/ CNTNAP	3 pseudogenes	1	29.7%	6.3%
16	7p14.1-p13	679,845	6	crm1034 ,crm266 ,crm267 ,crm268 ,crm269 ,crm270	GLI3	MGC2821	1 predicted	0	31.1%	6.5%
17	7q21.11	681,251	10	crm394 ,crm395 ,crm1119 ,crm396 ,crm1120 ,crm397 ,crm1121 ,crm1122 ,crm1123 ,crm1124	AIP1/ MAGI2	GNAI1	1 predicted; 1 pseudogene	0	30.0%	6.3%
18	7p21.2-p21.1	550,042	8	crm958 ,crm959 ,crm960 ,crm114 ,crm115 ,crm116 ,crm117 ,crm961	FERD3L	LOC221830/ TWISTNB	1 predicted	0	35.4%	5.2%
19	7p14.1	550,042	3	crm259, crm260, crm261	BC033981	INHBA	none	0	16.6%	9.7%
20	7q21.3-q22.1	550,197	2	crm518, crm1226	DC11/ ACN9	TAC1	2 predicted	0	28.7%	12.2%
	Total	25Mb	264			**Average**	0.8	4.2%	28.1%	7.5%
					Standard	**Deviation**	0.8	2.3%	6.3%	4.7%

Plotting the size of gene deserts against their respective repeat content percentage and CRMs predicted shows interesting results evident in figure 3.4. No of CRMs represented as red line in figure seems directly proportional to size represented in blue bars. Whereas, a decrease in number of CRMs is seen with increase in repeat content percentage depicted inversely proportional relation between the two. Since repeat content is represented in percentage, so it seems almost linear as compared to decreasing size of gene deserts if seen left to right. Gene desert1, 2,4 and 5 being larger in size among other regions had most number of CRMs with moderate repeat content. Gene desert 20 being smallest in size and having highest LINE and SINE percentage had lowest number of CRM predictions.

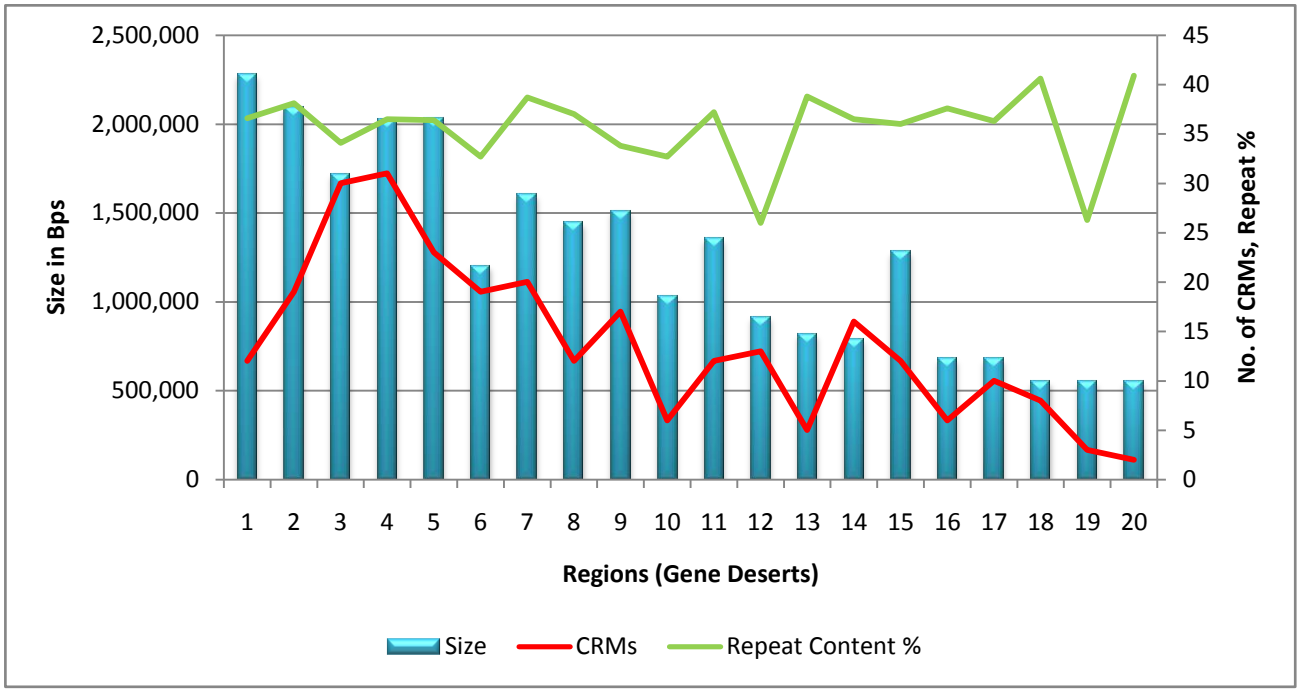

Figure 3.4| Comparative analysis of size, CRMs and repeat content of gene deserts.

3.8 Final representation

Circos, a data representation an analysis tool was used to summarize all the results of study in one circular diagram. Outer most orange circle in fugure 3.5 represents ideogram of human chromosome 7. Inner purple blocks represent gene deserts and the red marks are predicted CRMs. Inner most circle is for validation from vista. All marks in this track represent forebrain specific vista enhancers of chromosome 7. Green marks represent CRM non overlapping and red represent CRM overlapping vista enhancers.

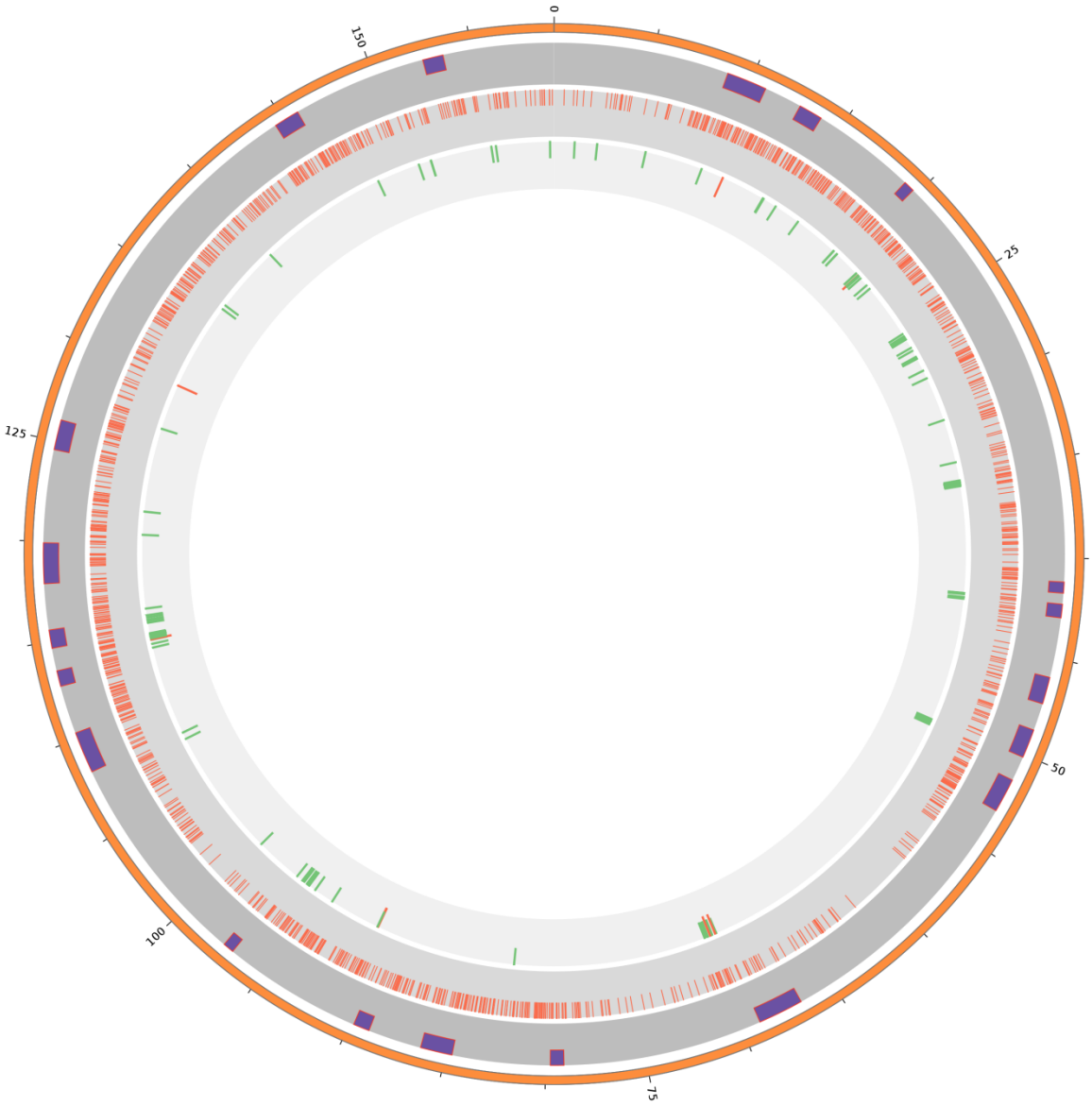

Figure 3.5|Chromosome 7 view with gene deserts (purple), CRMs (red) and vista enhancers (red).

Chapter 4

DISCUSSION

4. DISCUSSION

Since completion of human genome project and invention of next generation sequencing techniques, scientific community has been subjected to massive data explosion making annotation a key challenge. Scientific community has been able to annotate almost all the coding portion of genome, yet some questions are unanswered. Expanding the search space, attention of researchers shifted towards the long "junk" considered non-coding portion of the DNA. Seemingly simple stretch of As, Ts, Gs and Cs turn out to be a complex repertoire of genes, transposable elements and Cis-Regulatory Modules like promoters, enhancers, silencers and various other non coding elements. Individual identity of each cell is beholden to the regulatory elements which apparently reside in non-coding region of genome.

Among all regulatory elements, enhancers hold a prime responsibility of gene regulation. Owing their importance, many efforts have been invested in their identification, prediction and association with target genes. Lack of particular vocabulary, long range action, short and degenerated transcription factor binding sites are the properties of enhancers which make their recognition more difficult. However, some experimental and computational techniques are developed which predict enhancers with their own limitations and cost.

Comparative genomics is used in various studies to predict enhancers with a limited sensitivity (Grad et al., 2004; Pennacchio, et al., 2006). Since this technique is based on ultra conservation of elements, it generally misses out least conserved and organism specific enhancers. Recently, experimental techniques like ChIP-Seq have been developed to overcome the deficiencies of phylogenetic foot printing by providing conservation independent platform (Blow et al., 2010; Creyghton et al., 2010; Visel, et al., 2009). It is capable of producing high throughput data elevating the amount of information received finally. However, it fails to cater spatio-temporal specific nature of enhancers properly. Moreover, its histological specificity and high cost adds to its limitations. These limitations prove computational techniques and pipelines as indispensable substitute.

Most eukaryotic enhancers have more than one transcription factor binding sites on them thus often named as Cis-Regulatory Modules. Combinatorial binding of transcription factors on enhancer grants it spatio-temporal specificity as well as leaves some important marks for its identification. We exploited same property of CRMs in our study for their identification. Limiting our scope to forebrain, we compiled a catalogue of transcription factors having high

expression probability in forebrain. This catalogue was generated and purified using strong functional, statistical and literature evidence. Average size of human forebrain enhancer varies between 500-4000bps and usually contains 5-8 unique transcription factor binding sites distributed unevenly on enhancer. Distance between TFBSs is important for proper binding of transcription factors. Considering the average size and number of binding sites, we computed average distance between TFBS using functionally validated forebrain enhancers from VISTA enhancer browser. 104 forebrain specific enhancers were involved to compute average spacer distance, which turned out to be 250 bps.

Human chromosome 7 was chosen as our search space. Cis-acting elements usually reside in non-coding and non-repetitive part of genome. To reduce the search space and back ground noise, we masked repeats and exons on chromosome 7 prior to our search. TFs rarely bind to enhancers as single unit thus transcriptional regulation of gene is directed by combinatorial programming of TFs that interact as clusters of localized domains (Spitz & Furlong, 2012).To identify such clusters of localized domains on chromosome 7, we combined each set of TFBSs into a cluster that had lower than average spacer distance between any two TFBSs. Initially a huge number of clusters similar to CRMs were predicted. To increase sensitivity of the method we needed to capture maximum number of binding sites within a CRM, for this purpose we filtered out any clusters having binding sites less than 5. As a result, we report a set of 1453 regions on chromosome 7, which have at least 5 unique binding sites for forebrain specific transcription factors and can serve as potential Cis-Reglatory Module. Genomic regions having binding preferences for multiple transcription factors are more likely to endure chromatin remodeling (Heintzman et al., 2009) and DNAse hypersensitivity sites (Heintzman, et al., 2009) demonstrating functional relevance of the regions predicted in our study.

Conventional methods of enhancer identification involve use of epigenetic marks and phylogenetic foot printing as explained earlier. One such repository of functionally validated enhancers is VISTA Enhancer browser. Our predicted CRMs were compared with 133 functionally validated enhancers inhabiting chromosome 7 from VISTA enhancer browser. Out of 6 forebrain specific VISTA enhancers, we were able to catch 2, which depict significant sensitivity despite of difference in mode of prediction of both methods.

Gene deserts are long gene deficient regions. Being usually conserved, they can serve as best place for CRM or enhancer hunt.(Nobrega et al., 2003). Many studies have been dedicated to

enhancer search in gene deserts (de la Calle-Mustienes, et al., 2005; Sotelo et al., 2010). We collected 20 gene deserts on chromosome 7 from literature survey and mapped our CRMs in those regions. 6.3% of human chromosome 7 is gene desert. Out of 1453 CRMs, 264 lied unevenly in the 20 gene deserts resulting 22% of total CRMs in gene deserts. Since these gene deserts had varied percentage of repeat content, number of CRMs also varied. However, low number of CRMs was seen in gene deserts having high repeat content showing inversely proportional relation.

Conclusively, this study presents a computational platform supported by literature and experimental evidence, which can be used for locating potential CRMs. Employing Transcription factor binding site specificity and expression data systematically, transcriptional regulatory network of a specific tissue, can be analyzed in much less time and cost as compared to other techniques available today.

Chapter 5

REFERENCES

5. REFERENCES

Aerts, S., Thijs, G., Coessens, B., Staes, M., Moreau, Y., & De Moor, B. (2003). Toucan: deciphering the cis-regulatory logic of coregulated genes. *Nucleic acids research, 31*(6), 1753-1764.

Bachy, I., Failli, V., & Rétaux, S. (2002). A LIM-homeodomain code for development and evolution of forebrain connectivity. *Neuroreport, 13*(2), A23-A27.

Barton, R. A., & Harvey, P. H. (2000). Mosaic evolution of brain structure in mammals. *Nature, 405*(6790), 1055-1058.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., . . . Chen, F. (2010). ChIP-Seq identification of weakly conserved heart enhancers. *Nature genetics, 42*(9), 806-810.

Buck, M. J., & Lieb, J. D. (2004). ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics, 83*(3), 349-360.

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., . . . Sharp, P. A. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences, 107*(50), 21931-21936.

de la Calle-Mustienes, E., Feijóo, C. G., Manzanares, M., Tena, J. J., Rodríguez-Seguel, E., Letizia, A., . . . Gómez-Skarmeta, J. L. (2005). A functional survey of the enhancer activity of conserved non-coding sequences from vertebrate Iroquois cluster gene deserts. *Genome research, 15*(8), 1061-1072.

Déjardin, J., & Kingston, R. E. (2009). Purification of proteins associated with specific genomic Loci. *Cell, 136*(1), 175-186.

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science, 295*(5558), 1306-1311.

Down, T. A., & Hubbard, T. J. (2005). NestedMICA: sensitive inference of over-represented motifs in nucleic acid sequence. *Nucleic acids research, 33*(5), 1445-1453.

Elnitski, L., Jin, V. X., Farnham, P. J., & Jones, S. J. (2006). Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome research, 16*(12), 1455-1464.

Fägerstam, L. G., Frostell-Karlsson, Å., Karlsson, R., Persson, B., & Rönnberg, I. (1992). Biospecific interaction analysis using surface plasmon resonance detection applied to kinetic, binding site and concentration analysis. *Journal of Chromatography A, 597*(1-2), 397-410.

Geertz, M., & Maerkl, S. J. (2010). Experimental strategies for studying transcription factor–DNA binding specificities. *Briefings in functional genomics, 9*(5-6), 362-373.

Gibbs, R. A., Weinstock, G. M., Metzker, M. L., Muzny, D. M., Sodergren, E. J., Scherer, S., . . . Burch, P. E. (2004). Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature, 428*(6982), 493-521.

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

Gray, P. A., Fu, H., Luo, P., Zhao, Q., Yu, J., Ferrari, A., . . . Cai, Z. (2004). Mouse brain organization revealed through direct genome-scale TF expression analysis. *Science, 306*(5705), 2255-2257.

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.

Hellman, L. M., & Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nature protocols, 2*(8), 1849-1861.

Hesselberth, J. R., Chen, X., Zhang, Z., Sabo, P. J., Sandstrom, R., Reynolds, A. P., . . . Noble, W. S. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. *Nature methods, 6*(4), 283-289.

Hillier, L. W., Fulton, R. S., Fulton, L. A., Graves, T. A., Pepin, K. H., Wagner-McPherson, C., . . . Walker, R. (2003). The DNA sequence of human chromosome 7. *Nature, 424*(6945), 157-164.

Kaas, J. H. (2013). The evolution of brains from early mammals to humans. *Wiley Interdisciplinary Reviews: Cognitive Science, 4*(1), 33-45.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome research, 12*(6), 996-1006.

Kozl, K., & Listy, C. (1978). Biochemical nomenclature and related documents. *Chem. Listy, 72*, 288-305.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., . . . Marra, M. A. (2009). Circos: an information aesthetic for comparative genomics. *Genome research, 19*(9), 1639-1645.

Latchman, D. (1998). Eukaryotic transcription factors Academic Press. *San Diego, CA*.

Liu, C. L., Kaplan, T., Kim, M., Buratowski, S., Schreiber, S. L., Friedman, N., & Rando, O. J. (2005). Single-nucleosome mapping of histone modifications in S. cerevisiae. *PLoS biol, 3*(10), e328.

Maerkl, S. J., & Quake, S. R. (2007). A systems approach to measuring the binding energy landscapes of transcription factors. *Science, 315*(5809), 233-237.

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.

Matys, V., Fricke, E., Geffers, R., Gößling, E., Haubrock, M., Hehl, R., . . . Kel-Margoulis, O. V. (2003). TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic acids research, 31*(1), 374-378.

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.

Nobrega, M. A., Ovcharenko, I., Afzal, V., & Rubin, E. M. (2003). Scanning human gene deserts for long-range enhancers. *Science, 302*(5644), 413-413.

Osório, J., Mueller, T., Rétaux, S., Vernier, P., & Wullimann, M. F. (2010). Phylotypic expression of the bHLH genes Neurogenin2, Neurod, and Mash1 in the mouse embryonic forebrain. *Journal of Comparative Neurology, 518*(6), 851-871.

Pennacchio, L. A., Ahituv, N., Moses, A. M., Prabhakar, S., Nobrega, M. A., Shoukry, M., . . . Lewis, K. D. (2006). In vivo enhancer analysis of human conserved non-coding sequences. *Nature, 444*(7118), 499-502.

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.

Sandelin, A., Wasserman, W. W., & Lenhard, B. (2004). ConSite: web-based prediction of regulatory elements using cross-species comparison. *Nucleic acids research, 32*(suppl 2), W249-W252.

Scherer, S. W., Cheung, J., MacDonald, J. R., Osborne, L. R., Nakabayashi, K., Herbrick, J.-A., . . . Khaja, R. (2003). Human chromosome 7: DNA sequence and biology. *Science, 300*(5620), 767-772.

Sotelo, J., Esposito, D., Duhagon, M. A., Banfield, K., Mehalko, J., Liao, H., . . . Wu, X. (2010). Long-range enhancers on 8q24 regulate c-Myc. *Proceedings of the National Academy of Sciences, 107*(7), 3001-3005.

Spitz, F., & Furlong, E. E. (2012). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics, 13*(9), 613-626.

Tuerk, C., & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science, 249*(4968), 505-510.

Uemura, O., Okada, Y., Ando, H., Guedj, M., Higashijima, S.-i., Shimazaki, T., . . . Okamoto, H. (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.

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.

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., . . . Chen, F. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature, 457*(7231), 854-858.

Visel, A., Minovitsky, S., Dubchak, I., & Pennacchio, L. A. (2007). VISTA Enhancer Browser—a database of tissue-specific human enhancers. *Nucleic acids research, 35*(suppl 1), D88-D92.

Warren, C. L., Kratochvil, N. C., Hauschild, K. E., Foister, S., Brezinski, M. L., Dervan, P. B., . . . Ansari, A. Z. (2006). Defining the sequence-recognition profile of DNA-binding molecules. *Proceedings of the National Academy of Sciences of the United States of America, 103*(4), 867-872.

Wright, W. E., Binder, M., & Funk, W. (1991). Cyclic amplification and selection of targets (CASTing) for the myogenin consensus binding site. *Molecular and cellular biology, 11*(8), 4104-4110.

Zhao, Y., Granas, D., & Stormo, G. D. (2009). Inferring binding energies from selected binding sites. *PLoS Comput Biol, 5*(12), e1000590.

Zykovich, A., Korf, I., & Segal, D. J. (2009). Bind-n-Seq: high-throughput analysis of in vitro protein–DNA interactions using massively parallel sequencing. *Nucleic acids research*, gkp802.