Evolutionary dynamics of CTCF binding sites in mammalian genomes

Lipika Pradeepkumar Taneja
MS15126

A dissertation submitted for the partial fulfilment of
BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali
May 2020

Certificate of Examination

This is to certify that the dissertation titled "**Evolutionary dynamics of CTCF binding sites in mammalian genomes**" submitted by **Lipika Pradeepkumar Taneja** (MS15126) for the partial fulfilment of BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr Shashi B. Pandit Dr Rajesh Ramachandran Dr Kuljeet S. Sandhu

(Supervisor)

Dated: June 15, 2020

Declaration

The work presented in this dissertation has been carried out by me under the guidance of **Dr. Kuljeet Singh Sandhu** at the Indian Institute of Science Education and Research Mohali. This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Lipika Pradeepkumar Taneja

(Candidate)

Dated: June 15, 2020

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr. Kuljeet Singh Sandhu

(Supervisor)

Acknowledgement

I would like to express my sincere gratitude to Dr Kuljeet Singh Sandhu for giving me the opportunity to work in his laboratory for my MS thesis and for his guidance and support throughout the year.

I also thank my thesis committee members – Dr Shashi B. Pandit and Dr Rajesh Ramachandran. A big thanks to IISER Mohali for providing the academic facilities and comfortable environment that ensured the time here went by smoothly, and also for the opportunity to be exposed to the research world through its curriculum. I would also like to thank DST for providing the INSPIRE fellowship to students like me to pursue Science.

I would like to thank all members of the Genome Biology Lab – Ken, Yachna, Meenakshi, Jui, Mohan and Smriti for always being ready to help and giving their valuable inputs. I am especially grateful to Sunandini for being immensely supportive both inside and outside lab and for project discussions and tea breaks when I needed them the most.

I would like to extend my gratitude to Dr. N.G Prasad for being an extremely encouraging mentor and for his help and advice in times of uncertainty.

A very special thanks to all the friends I have made here, for always being there to help, encourage, listen and inspire, for everything I have learnt from each one of them and for the most memorable experiences these five years.

Lastly, but most importantly, I would like to thank my family for being my greatest support system. For all the faith they've had in me, for always encouraging me to do my best, for providing me with a comfortable life and for ensuring that I had the liberty to make my own choices that helped me become the person I am today.

List of databases and software

1.	RSAT	Matrix scan algorithm for obtaining CTCF binding sites in hg19
2.	JASPAR	CTCF binding motif position specific scoring matrix (PSSM)
3.	UCSC Genome Browser	Human genome assembly (hg19), chain files from hg19 to other species
4.	ChIP-seq web server (ChIP-cor analysis module)	Base-wise conservation scores of CTCF binding sites in placental mammals
5.	NCBI Assembly database	Assembly information of mammalian genomes
6.	UCSC RepeatMasker	hg19 repeat sequences
7.	UCSC LiftOver standalone tool	Lifting over human CTCF binding sites to other species
8.	TimeTree	Divergence times of mammalian species from human
9.	ENSEMBL Biomart	Orthologs of human genes in other species
10.	ENCODE project	Hi-C obtained coordinates of TAD boundaries, chromHMM chromatin states in hg19
11.	BGEE	Gene expression values in adult tissues in human and mouse
12.	Expression Atlas	Gene expression values across developmental stages in human and mouse

List of tables

List of figures

Contents

Abstract

CTCF is a ubiquitously expressed protein, which over the years, has been associated with several functions, initially that of a transcriptional repressor and activator. Subsequently, it was identified as an insulator protein with enhancer-blocking functions. Research on the three-dimensional genome revealed its involvement in genome organization, mainly through the formation of chromatin loops, leading to long-range communication between genes and regulatory elements and also blocking interactions, aligning with its role as an insulator. Evidence for these functions of CTCF has come from Hi-C maps and experiments showing the enrichment of TAD boundaries with CTCF binding sites. Studies have shown changes in gene expression at specific loci as a result of changes in chromatin loops and disruption of TAD structure. However, the mechanism and extent of this effect is not understood.

In this thesis, we study the evolutionary dynamics of CTCF binding, focussing on the CTCF binding sites in human that are lost in other species. We hypothesize that evolutionary differences in CTCF binding could reflect in changes in gene expression and lineage or species specific phenotypes We also do a more detailed analysis of the sites that are lost in mouse. We look at the effect of the presence of CTCF binding sites on the expression of the gene nearest to it. While small changes are seen in gene expression across developmental stages, differences in chromatin states of these sites do not show enough difference to validate these changes.

Chapter 1

Introduction

1.1 CTCF structure and functions

CTCF, also known as the CCCTC-binding protein is a highly conserved protein in higher eukaryotes. It consists of an N-terminal domain, a C-terminal domain and a central DNA-binding domain with 11 zinc-fingers, which is especially highly conserved.

Over time, CTCF has been implicated with several diverse functions in the regulation of gene expression. Studies involving CTCF at specific loci have indicated its role in transcriptional regulation, imprinting and insulator activity. The discovery of CTCF and other early studies described the protein as a transcriptional repressor of the c-myc oncogenes in human and chicken (Filippova et al., 1996; Lobanenkov et al., 1990).

Another study later revealed the role of CTCF as a transcriptional activator for the amyloid beta-protein precursor (APP) gene promoter (Vostrov & Quitschke, 1997).

Subsequently, CTCF was found to behave as an insulator protein with enhancer blocking activity, in vertebrates. An insulator is a sequence of DNA with binding sites for DNA binding proteins. As an enhancer blocker, it blocks interaction between regulatory elements or between enhancer and gene promoters. One of the first studies to report this used enhancer blocking assays which showed the role of CTCF as an insulator at the chicken β-globin locus (Bell et al., 1999). Similarly, a study showed that CTCF mediated enhancer blocking controls imprinting at the mammalian H19/Igf2 locus. CTCF binding sites were found in the imprinted control region (ICR) of the locus (Bell & Felsenfeld, 2000).

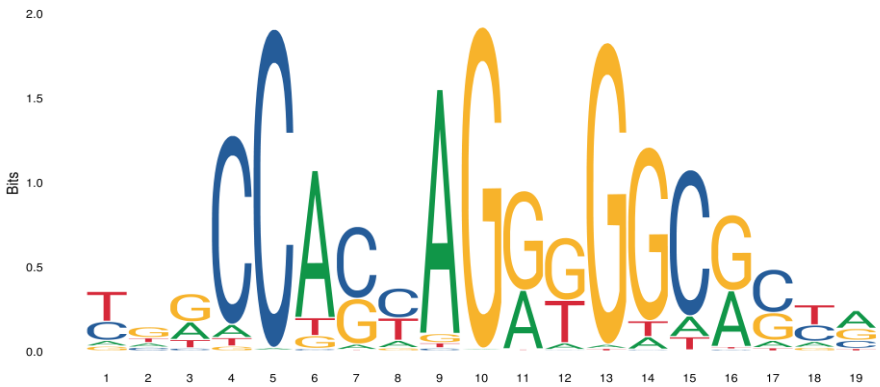

Fig.1.1 Sequence logo of CTCF binding motif obtained from JASPAR database (Fornes et al., 2019)

1.2 CTCF and the 3D genome

Following more recent advances in genome technologies, mainly in 3C (chromosome conformation capture) (Dekker et al., 2002) based methods have resulted in studies delving into the three-dimensional organization of the genome. The studies have revealed the genome wide role of CTCF in maintaining genome architecture.

Owing to their large size, the genomes of higher eukaryotes are tightly packed inside the limited space of the nucleus, forming loops and folds. These genomes not only contain information in the linear sequence of their DNA but also in space. The chromatin is organized hierarchically in three-dimension and this organization is closely linked to its function. In the interphase stage, the chromosomes in the nucleus exists in the form of distinct chromosome territories (Cremer et al., 2006; Cremer & Cremer, 2010). At a smaller scale, the chromatin forms two compartments - trancriptionally inactive regions tend to interact with each other towards the nuclear periphery and form the B compartment, whereas, trancriptionally active regions interact preferentially towards the interior of the nucleus, forming the A compartment. Locally, these compartments consist of self-interacting domains called TADs (topologically associated domains) . TADs were identified using Hi-C (Lieberman-Aiden et al., 2009) – a 3C technology combined with high throughput sequencing, used to investigate all-to-all interactions in the genome. The result of a Hi-C experiment is a two-dimensional contact map showing pairwise interaction frequencies. High resolution Hi-C maps led to the identification of chromatin loops throughout the genome. Some of these loops were found to be anchored at TAD boundaries and are therefore involved in the formation of TAD structures. Within a TAD, these loops link enhancers to promoters of genes and are associated with gene expression. They insulate the regions within a TAD from regions of another, maintaining the integrity of the three dimensional genome. These chromatin loops bring about long range interactions, bringing regions in contact in space even when they are distantly located in the linear DNA sequence (Rao et al., 2014).

Long range interactions connect multiple genes and distal regulatory elements as well as regulatory elements to each other, creating complex three-dimensional networks.

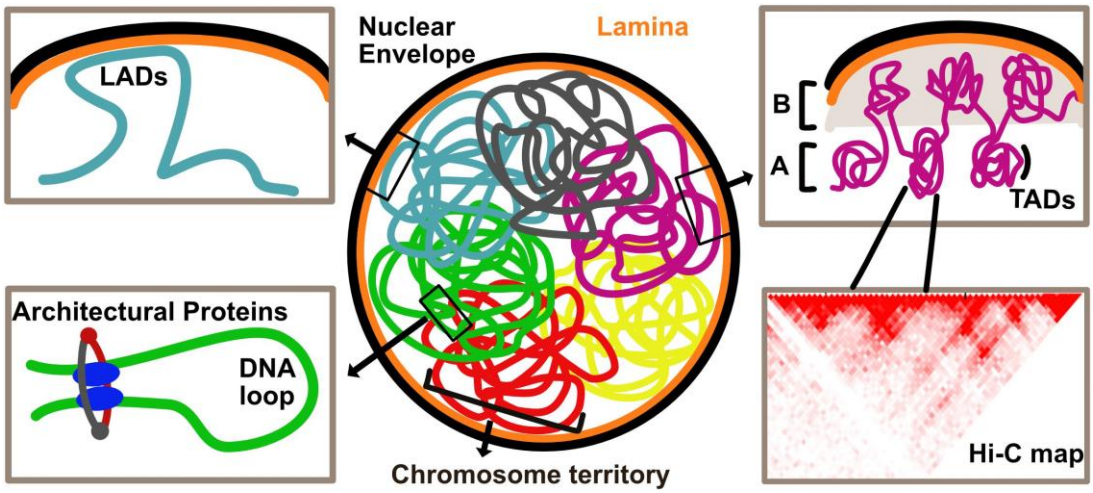

Fig.1.2 Three-dimensional organization of the genome (Sivakumar et al., 2019)

The boundaries of these TADs were found to be enriched with CTCF, transfer RNA, housekeeping genes and short interspersed elements (SINEs) (Dixon et al., 2012).
A study used an auxin-inducible degron system to deplete CTCF in mouse embryonic cells (mESCs). Reversible degradation of CTCF resulted in the loss of insulation at TAD boundaries as revealed by Hi-C maps, indicating the importance of CTCF in the formation of TADs. The study also revealed the disruption of chromatin loops between the CTCF boundaries on depletion of CTCF. These effects were reversed on the restoration of CTCF (Nora et al., 2017). These studies highlighted the role of CTCF in both blocking as well as mediating long-range interactions, forming the basis of its diverse functions and as a result its involvement in the three-dimensional organization of the genome.

A model proposed for the formation of these loops is the loop extrusion model wherein, factors most likely cohesin extrude portions of the DNA in the form of loops until they encounter bound CTCF (Fudenberg et al., 2016). The formation of chromatin loops is not only dictated by the presence of CTCF binding sites but also by the polarity of CTCF binding at these sites, requiring them to be oriented in a convergent manner. The deletion of CTCF binding sites results in disruption of chromatin loops which are not restored on insertion of the same site with an opposite orientation (de Wit et al., 2015). The deletion, inversion or repositioning of these sites can therefore destabilise higher order chromatin structures and lead to the loss of insulation at TAD boundaries, resulting in fusion of TADs. This disruption of the TAD structure can lead to new enhancer-promoter

interactions, affecting the expression of genes (Lupiáñez et al., 2015; Narendra et al., 2015).

CTCF mediated chromatin loops have also been shown to regulate alternative splicing, possibly by bringing splicing factors in proximity with exons (Ruiz-Velasco et al., 2017).

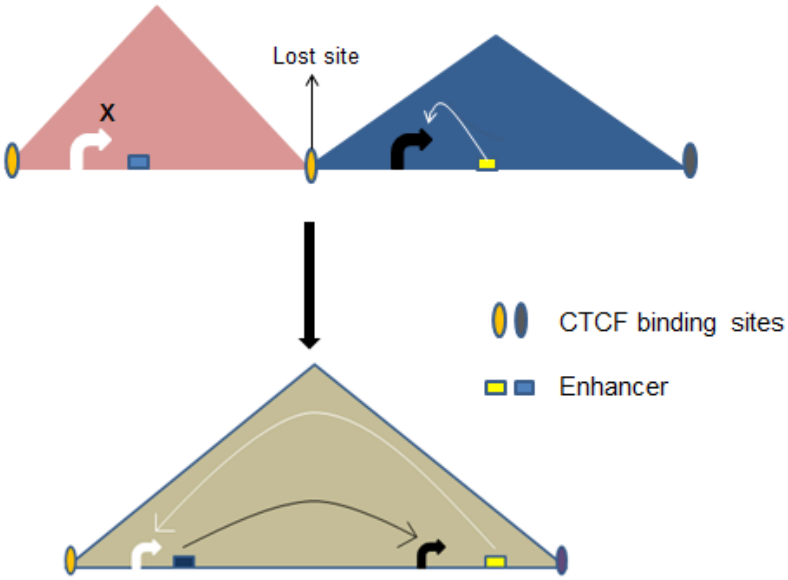

Fig.1.3 Schematic showing changes in TAD structure and long range interaction

Although evidence for this model exists and suggests that changes in long range interactions possibly lead to differences in gene expressions, the extent of this effect is unknown. While TAD boundaries are found to be enriched with CTCF binding sites, these sites are also found all throughout the genome. Therefore the exact role of CTCF in the genome remains in question. Since CTCF is ubiquitous and essential to cell viability, genome-wide disruption of CTCF to study its function is not feasible. An evolutionary analysis of CTCF binding in-silico provides an approach free of the limitations posed by experimental studies. This approach exploits the natural genetic and phenotypic variation in organisms and provides a system to study the effect of CTCF on gene expression. This would not only help in studying the contribution of CTCF-mediated gene expression differences to the evolution of lineage specific traits, but it could also provide an insight into the mechanism by which CTCF binding causes these differences.

Chapter 2

Materials and Methods

2.1 Identification of CTCF binding sites

The CTCF binding sites in the human genome were predicted by a matrix scan algorithm using RSAT (Turatsinze et al., 2008). The CTCF binding motif in human was obtained from JASPAR (Fornes et al., 2019) in the form of a position-specific scoring matrix (PSSM). The human genome (hg19) was obtained from the UCSC Genome Browser (Kent et al., 2002).

In this algorithm, the genome is scanned for the matrix in a sliding manner and a score is calculated for the sequence starting from each position.

$$\text{Score, } W_S = \ln\left(\frac{P(S|M)}{P(S|B)}\right)$$

where,

P(S|M) = probability of generating the sequence segment given the matrix.

P(S|B) = probability of generating the sequence segment given the background model.

A p-value is obtained for each match based on the score. A significance value is also obtained for each site where,

$$\text{Significance} = -\log_{10}(\text{p} - \text{value})$$

The p-value cut-off for obtaining the highest scoring matches was obtained from a distribution of p-values. These sites were selected for further analysis.

The sites were filtered by removing human-specific repeats. The repeat sequences for hg19 were obtained from UCSC's RepeatMasker track (Kent et al., 2002).

2.2 Mapping sites to other genomes

The CTCF binding sites obtained in the human genome were lifted over to 33 other mammalian genomes. Assembly information and statistics were obtained from the NCBI Assembly database (NCBI Resource Coordinators, 2018). The chain files for each genome were obtained from the UCSC Genome Browser (Kent et al., 2002). UCSC's

LiftOver standalone tool was used to lift over the sites. Sites lifted over to each of the genomes were obtained as well as the sites in the human genome that were not matched to each genome. Divergence times of each species from human were obtained from the TimeTree web resource (Kumar et al., 2017).

2.3 Expression divergence

An expression divergence analysis was done for mouse-specific losses, taking cow and dog as outgroups. Mouse-specific losses were defined as the CTCF-binding sites in the human genome that were mapped to cow and dog genomes but were absent in mouse. For each site, the gene nearest to the CTCF-binding site in human (hg19), and its ortholog in mouse (mm10) were considered for the analysis.

The gene expression values for eight adult tissues – heart, kidney, skeletal muscle, brain, colon, liver, lung and testis - in human and mouse were obtained from the BGEE database (Bastian et al., 2008). The data was already processed and available in FPKM (fragments per kilobase per million). Human and mouse orthologous genes were obtained from ENSEMBL's Biomart server (Cunningham et al., 2019). The expression values of these orthologous genes were taken for the divergence analysis.

The data was first quantile normalized and the Spearman correlation between the vector containing expression values for the human tissues and the vector containing expression values for mouse tissues was calculated for each gene. Following this, the expression divergence value for each gene was found where,

$$\text{Divergence} = 1 - \text{Correlation}$$

The expression divergence values of the genes nearest to CTCF-binding sites present in mouse and of genes nearest to sites absent in mouse were taken and difference between the two sets was tested by a Mann-Whitney U test.

2.4 Time course expression

The time series RNA-seq data for human and mouse was obtained for 7 tissues – forebrain, hindbrain, heart, liver, kidney, testis and ovary from the Expression Atlas database (Papatheodorou et al., 2018). The expression values in FPKM were obtained for genes in each tissue across developmental stages, including both embryonic and fetal stages.

The same CTCF binding sites representing mouse-specific losses and the genes nearest to them were used for the analysis along with the sites present in all three species. The time course expression of the genes nearest to sites present in mouse and those nearest to the sites absent in mouse were compared for each human and mouse tissue.

2.5 Analysis of TAD boundary density around CTCF binding sites

The TAD boundary density around the CTCF binding sites was plotted in human.
The coordinates of TAD boundaries for hg19 were obtained from the ENCODE portal (Davis et al., 2018) using data from Hi-C projects for the IMR90 cell line (Dixon et al., 2012). 1Mb region was taken both upstream and downstream of the CTCF binding site, the region was divided into 10Kb bins and the number of TAD boundaries in each bin averaged across the sites was found.

2.6 Analysis of chromatin states of CTCF binding sites in human

The chromatin states mapping to the CTCF binding sites were found. The chromatin states information generated by a hidden Markov model analysis was downloaded from the ENCODE portal (Davis et al., 2018). The data was obtained for 4 cell lines – GM1878, H1-hESC, HepG2 and K562 with the following accession numbers – ENCFF001TDH, ENCFF001TDI, ENCFF001TDJ, ENCFF001TDN, ENCSR655GEL, ENCSR381BVE, ENCSR399GZH and ENCSR052IZW.

Python was used as the programming language for file handling. R was used for statistical analysis and data visualization.

Chapter 3

Results

3.1 Obtaining CTCF binding sites in human

We obtained a total of 7,378,237 CTCF binding sites after applying the matrix scan algorithm to the human genome (hg19) using the CTCF PSSM. Using a frequency distribution of the scores of the sites (Fig.3.1A) and the distribution in the log10 scale (Fig. 3.1B), we decided the threshold score to select high-scoring sites to use for further analysis.

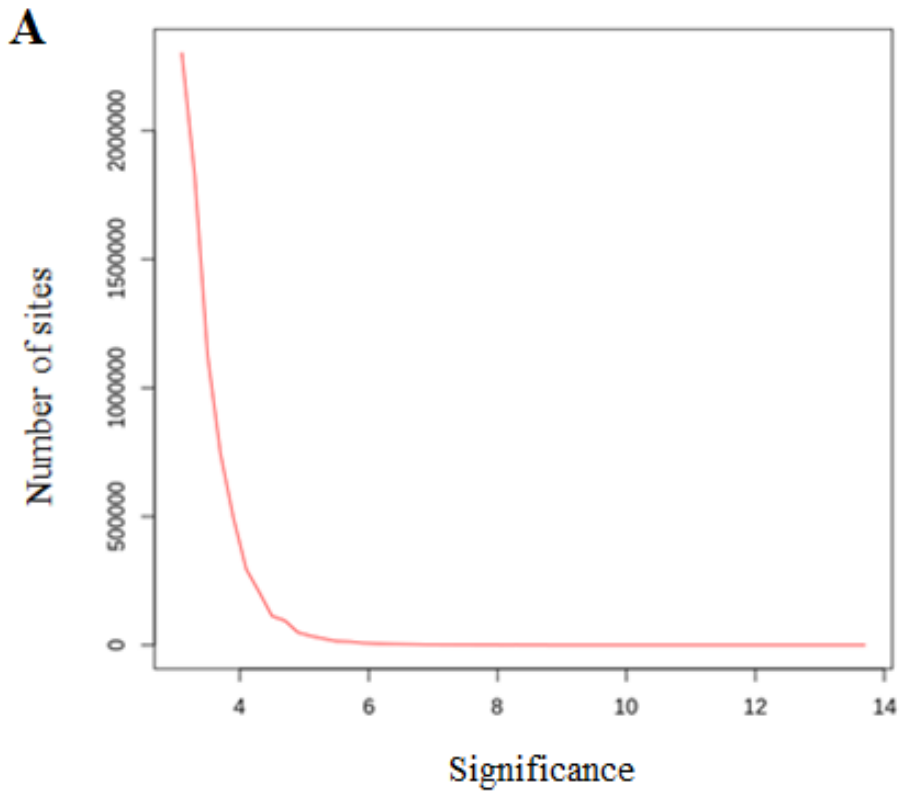

Fig.3.1 (A) Frequency distribution of scores

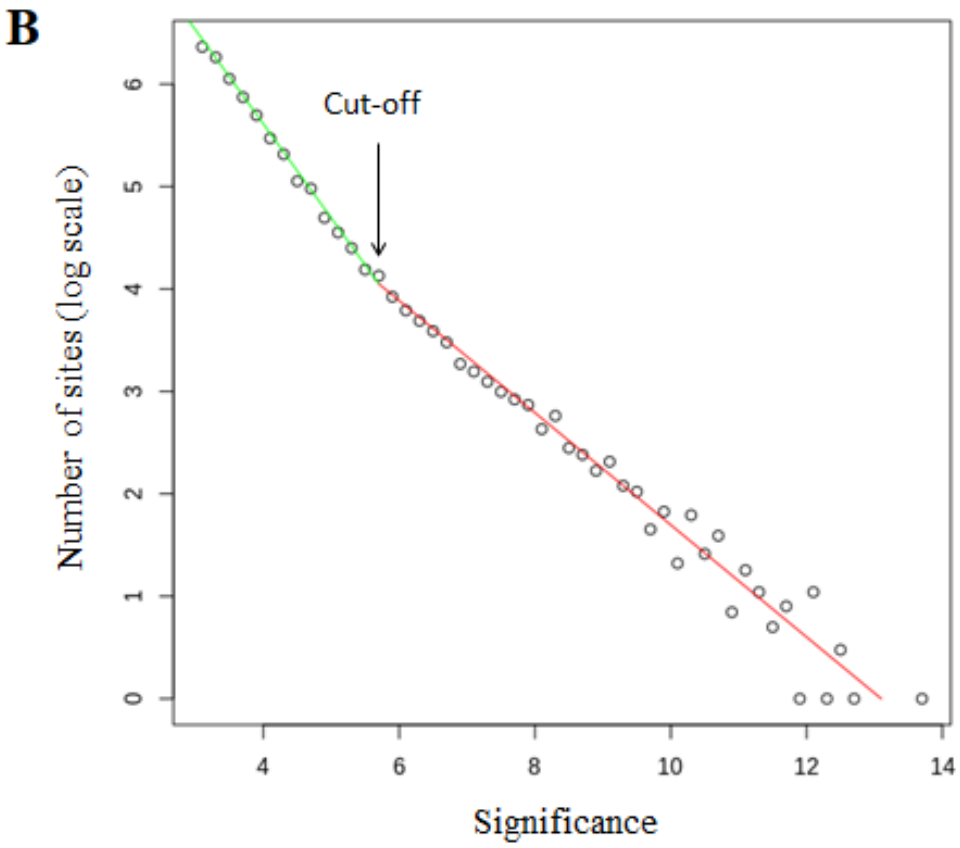

Fig.3.1 (B) Distribution of scores with frequency in log scale

The score is the negative logarithm of the p-value for each site match. A score of 5.5 (p-value = 3.2e-6) was decided as the threshold and 42,296 sites with a score higher than 5.5 and after filtering for human specific repeats were used for the study.

3.2 Mapping sites to mammalian genomes

The CTCF binding sites were found to be highly conserved in placental mammals as seen in Fig.3.2 based on PhyloP base-wise conservation scores (Ambrosini et al., 2014). The figure shows the cross-species conservation of the CTCF binding sites and +1000bp and -1000bp of flanking region. The spike in the plot indicates high sequence conservation at the binding site itself.

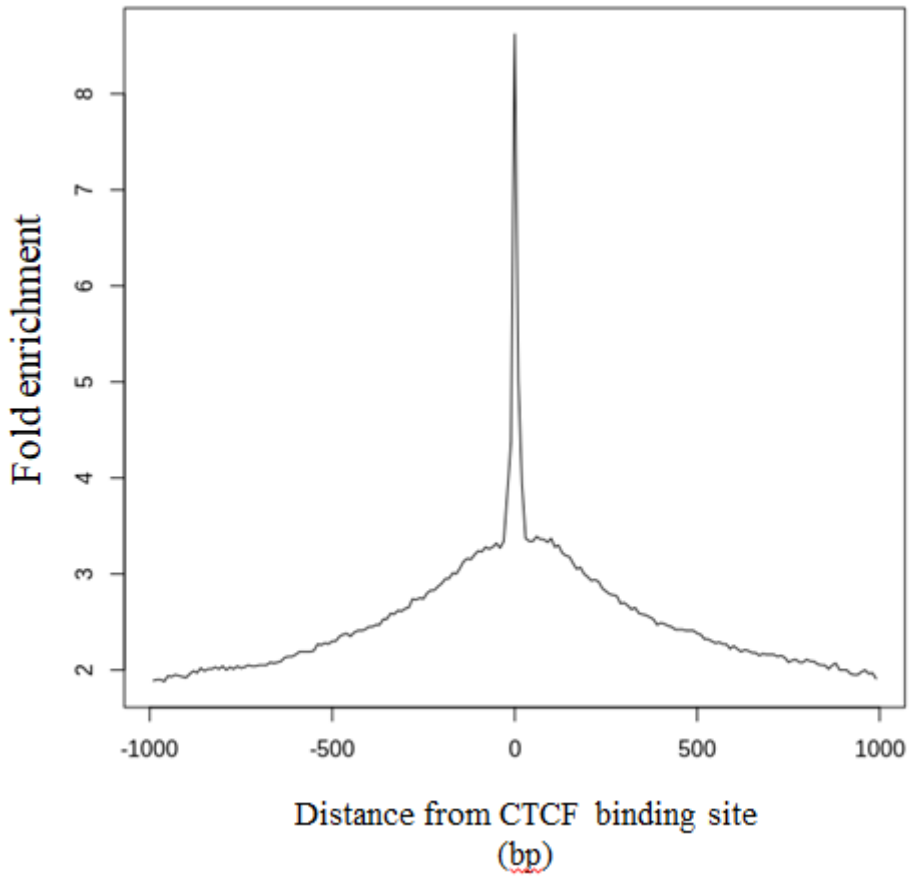

Fig.3.2 Sequence conservation of CTCF binding sites in placental vertebrates

We mapped the CTCF binding sites in human to 33 mammalian species out of a total of 46 species with available chin files from hg19, choosing species with either a chromosome level of genome assembly or a scaffold level assembly with a scaffold N50 value greater than 1Mb. Table A1 contains details of the genome assembly of all 46 mammalian species for which chain files from hg19 were available.

After lifting over the CTCF binding sites in human to the other mammalian species, we obtained the sites that appeared to be absent in each species. Fig.3.3 shows the phylogenetic tree of the species that were used for the analysis. Table A2 contains a summary of the lift over of hg19 CTCF binding sites to the 33 species selected.

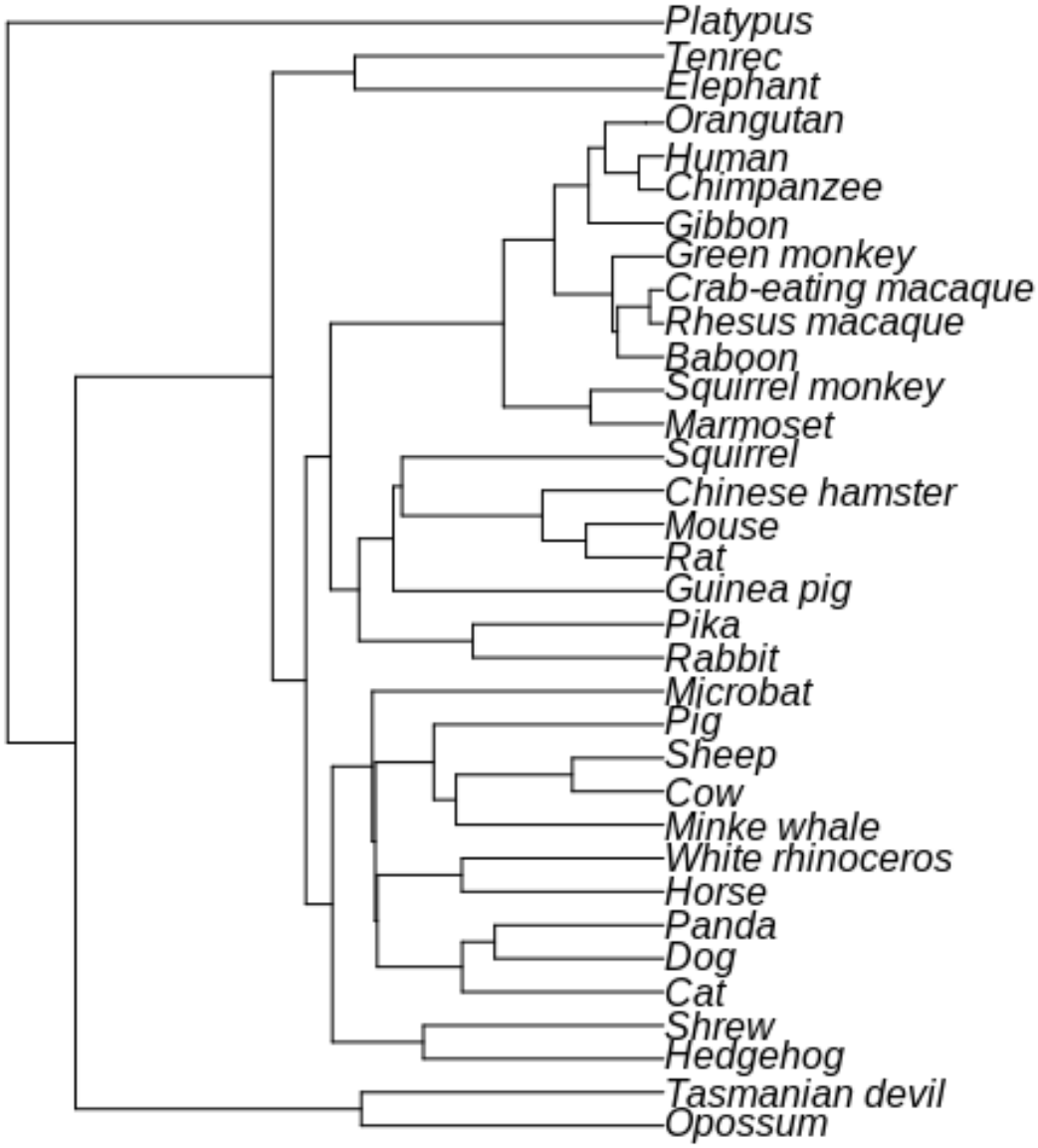

Fig.3.3 Phylogenetic tree of 34 mammalian species (including human) used in the analysis.

The bar plot in Fig.3.4 shows the percentage of the CTCF binding sites in human that were lifted over and lost in each of the other mammalian species. We see that the there is a great variation in the percentage of sites lifted over in these species.

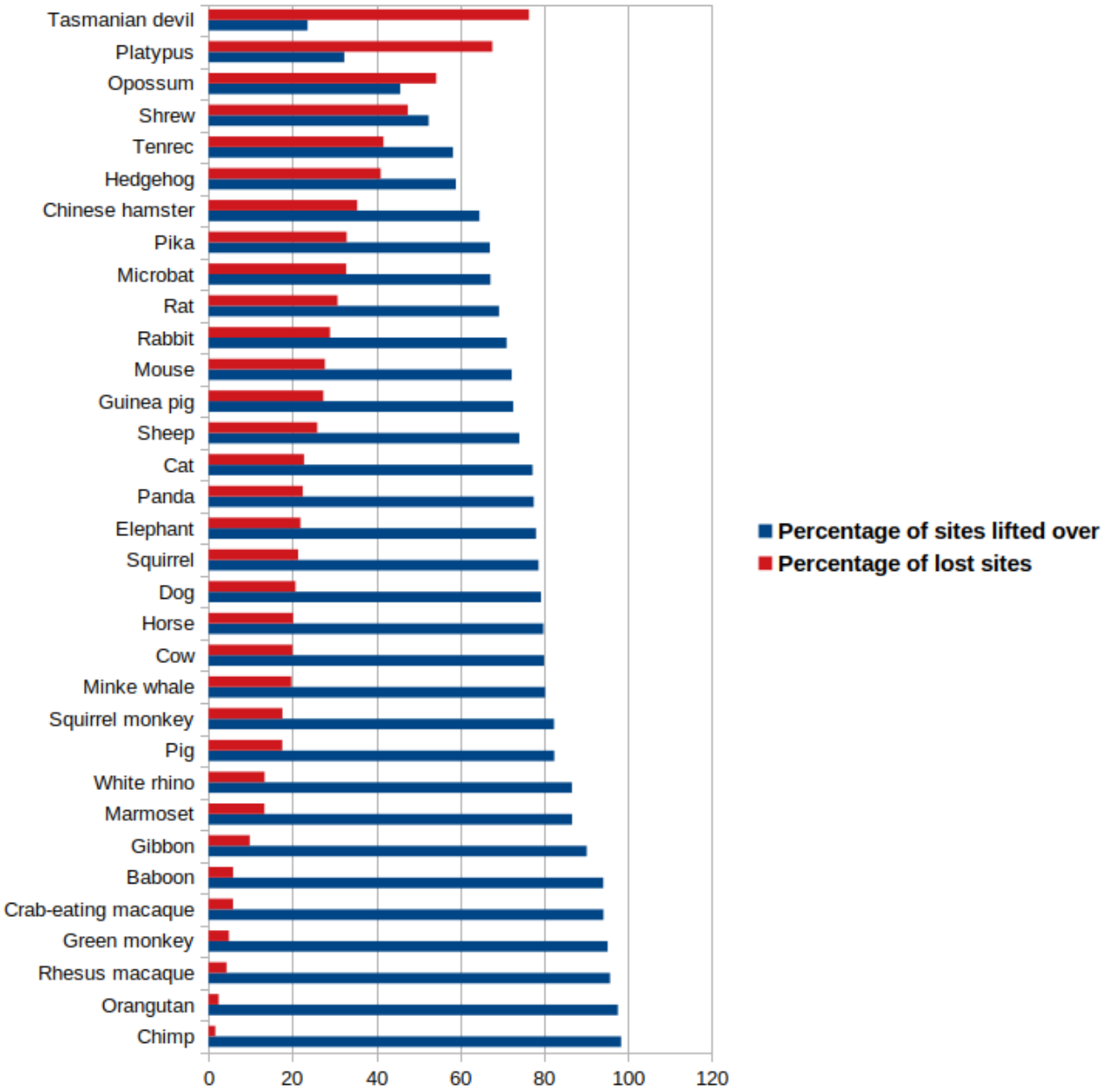

Fig.3.4 Percentage of sites in human that were lifted over or lost in each species

Next, we wanted to see the relationship between the proportion of sites in human that are absent in a particular species and the divergence time of the species from human. We see in Fig.3.5 that there is a positive correlation between the divergence time between human and an organism, and the percentage of CTCF binding sites in human that were not lifted over. It appears that the more diverged a species is from human, more is the number of sites that seem to be lost in the species.

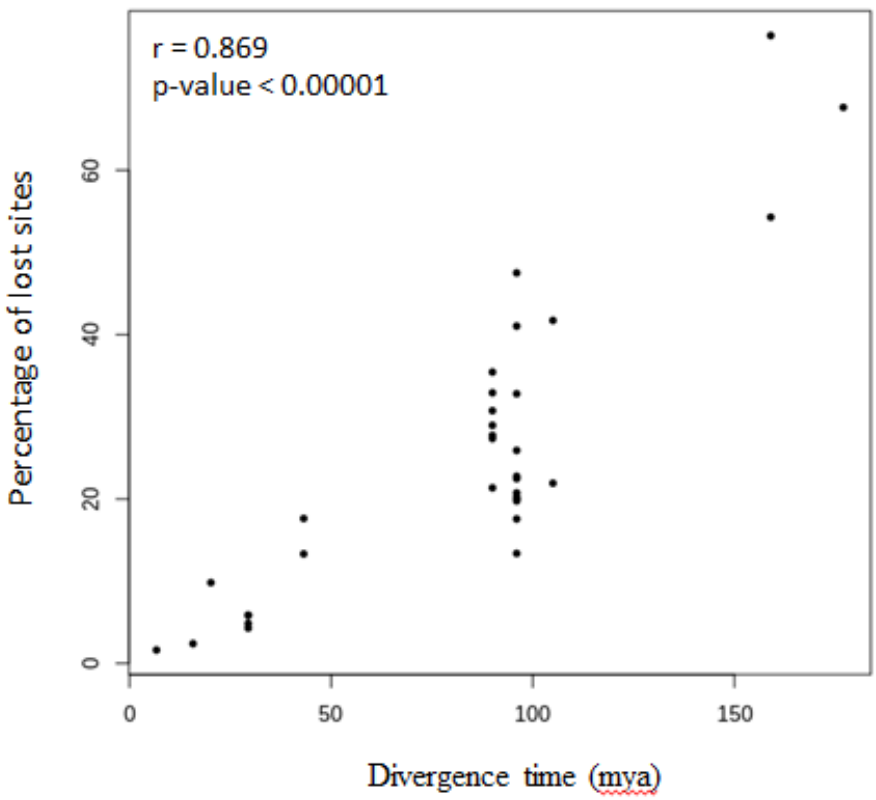

Fig.3.5 Scatter plot of percentage of lost sites v/s divergence time. It shows the Pearson correlation coefficient I and the p-value representing the significance of the test.

3.3 Effect of CTCF binding sites on gene expression

To test our hypothesis, we looked at how gene expression is possibly affected by the presence or absence of CTCF binding sites in mouse. Fig.3.6 shows the expression divergence in human (hg19) and mouse (mm10) for genes nearest to the CTCF binding sites, both present and absent in mouse. The gene expression values are for 8 adult tissues – heart, kidney, skeletal muscle, brain, liver, lung, colon and testis.

As seen in Fig.3.6, there is no significant difference between the expression divergence in genes nearest to CTCF binding sites present in mouse and nearest to those absent in mouse.

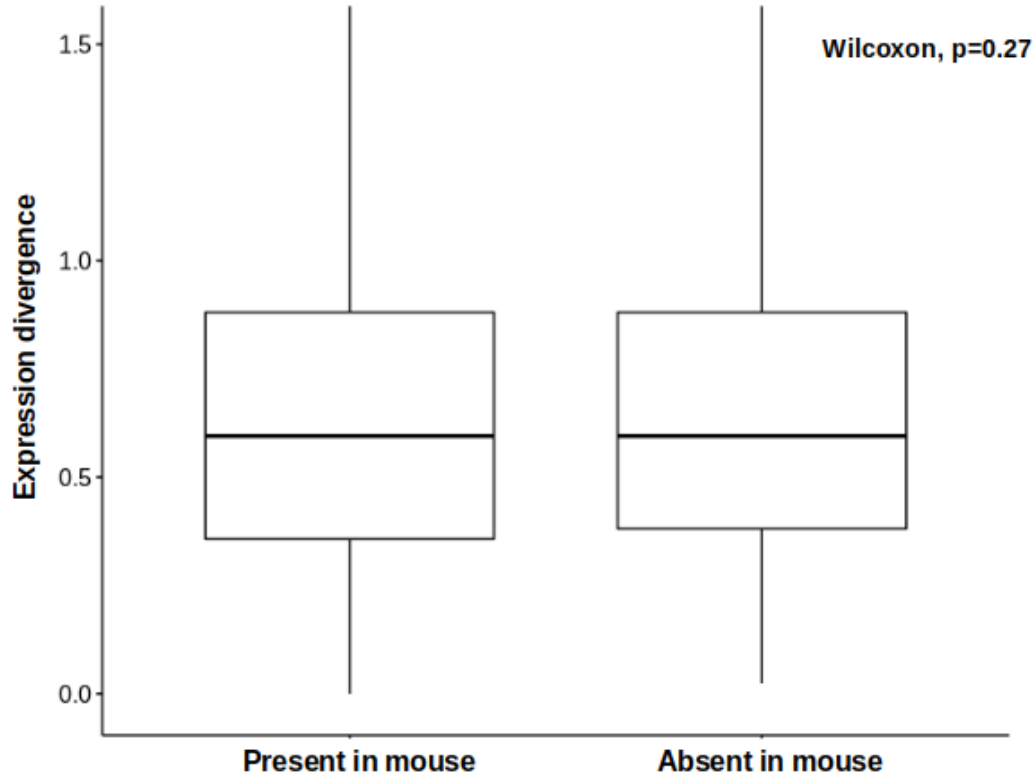

Fig.3.6 Expression divergence in human and mouse of genes nearest to CTCF binding sites present and absent in mouse.

Seeing that there was no effect of presence or absence of CTCF binding sites on the gene expression in adult tissue, we wanted to check if a difference arises in other developmental stages of the organism. CTCF has been found to be a major player in the development of several tissues. Looking at the trajectories of the expression of genes nearest to the CTCF binding sites across development could help in seeing if there is any difference arising due to the presence of the CTCF binding site.

Fig.3.7-3.10 show the time course expression, i.e. across developmental stages (both embryonic and post-natal), of genes in four tissues – forebrain, kidney, liver and heart, and a comparison of the genes nearest to sites present in mouse and those that are absent in mouse.

In Fig.3.7, we see that in the mouse forebrain, the median expression of genes nearest to CTCF sites absent in mouse is greater than the median expression of genes nearest to sites present in mouse. This difference is not seen in the human forebrain.

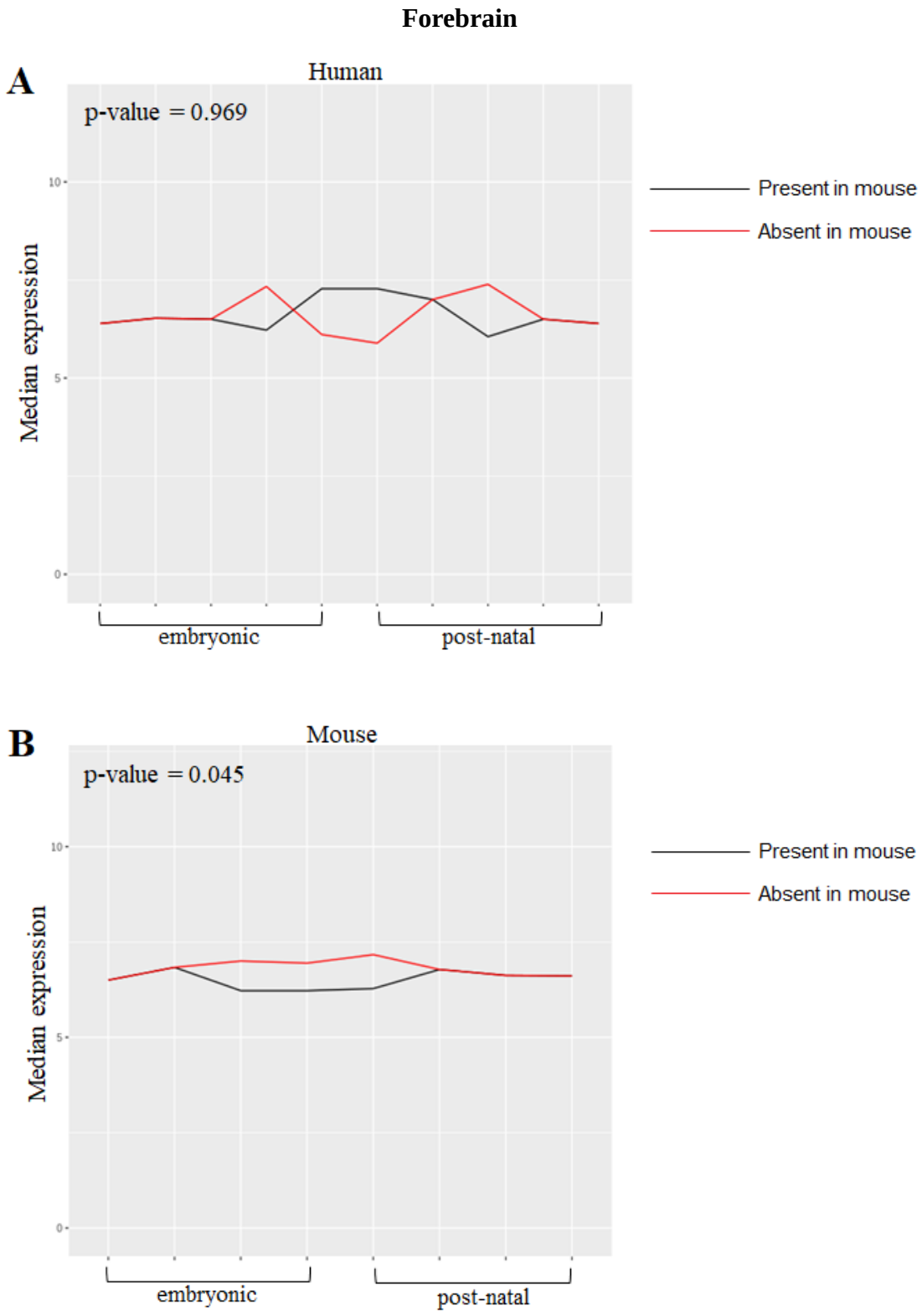

Fig.3.7 Median expression of genes nearest to CTCF binding sites in forebrain across development. The black line represents the expression of genes nearest to sites present in mouse and the red line represents the expression of genes nearest to sites absent in mouse. A Mann-Whitney U test was done to test the differences between the two.

(A) Median expression of genes in human (B) Median expression of genes in mouse

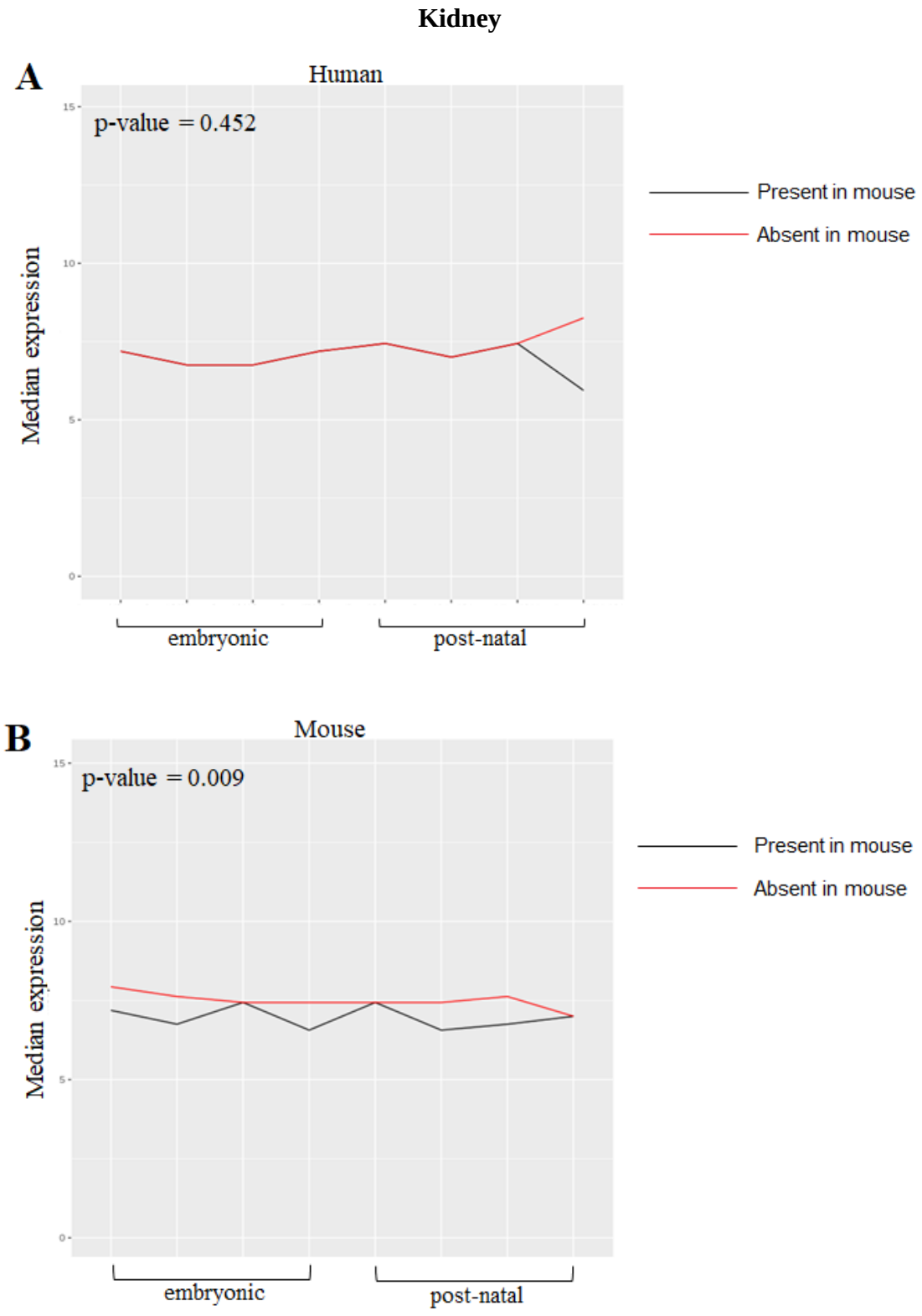

Fig.3.8 Median expression of genes nearest to CTCF binding sites in kidney across development. The black line represents the expression of genes nearest to sites present in mouse and the red line represents the expression of genes nearest to sites absent in mouse. A Mann-Whitney U test was done to test the differences between the two.

(A) Median expression of genes in human (B) Median expression of genes in mouse

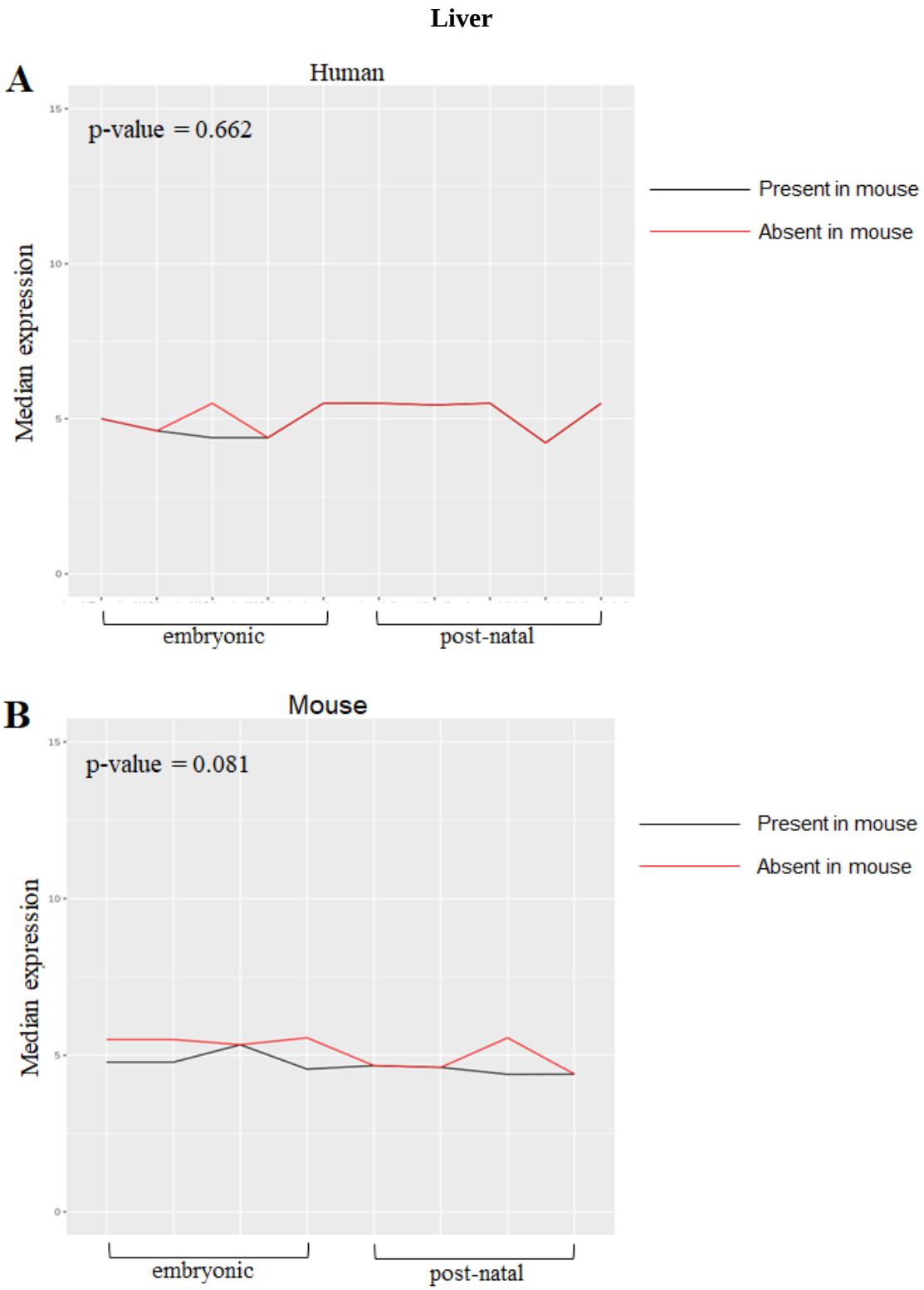

Fig.3.9 Median expression of genes nearest to CTCF binding sites in liver across development. The black line represents the expression of genes nearest to sites present in mouse and the red line represents the expression of genes nearest to sites absent in mouse. A Mann-Whitney U test was done to test the differences between the two.

(A) Median expression of genes in human (B) Median expression of genes in mouse

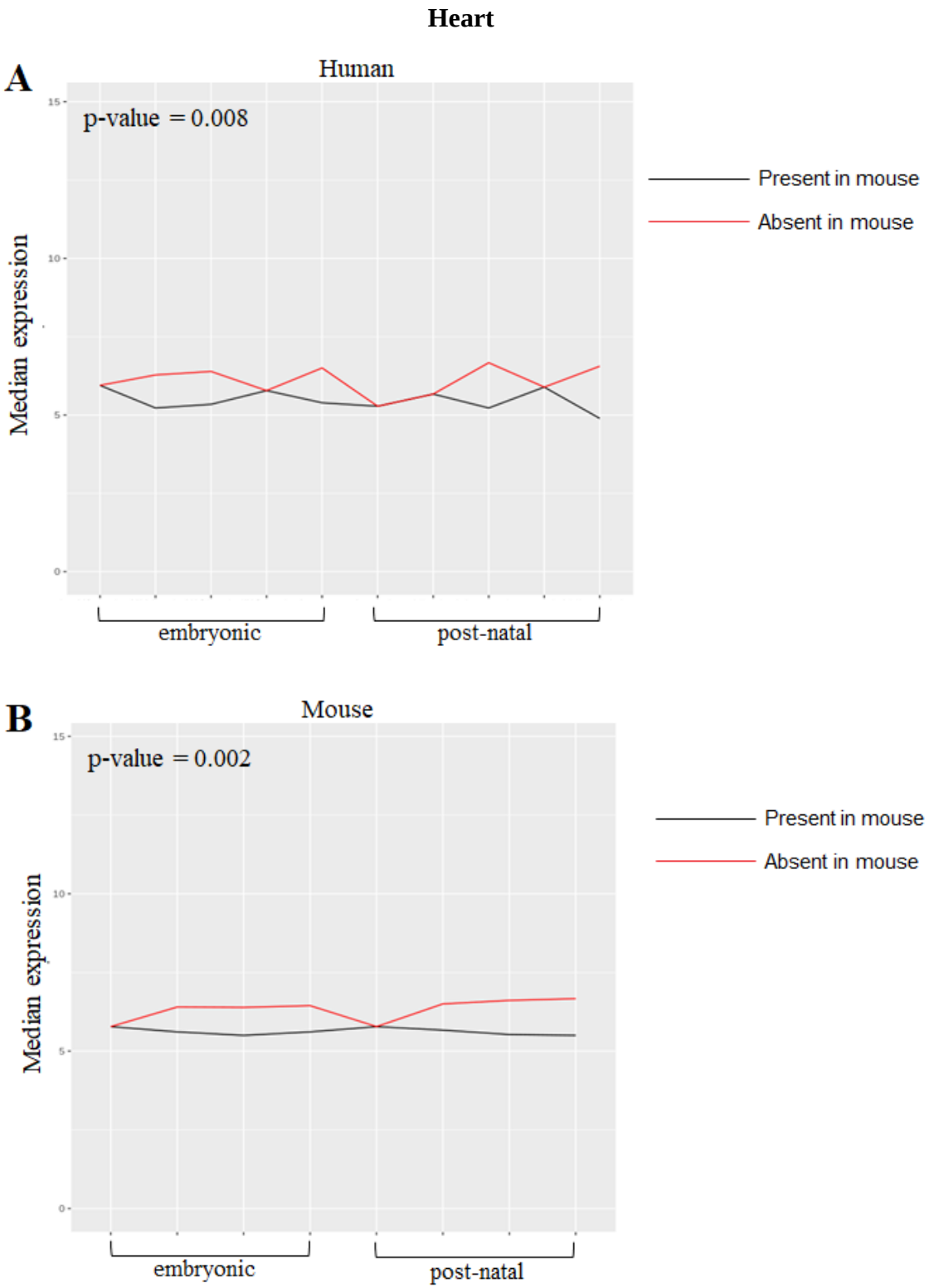

Fig.3.10 Median expression of genes nearest to CTCF binding sites in heart across development. The black line represents the expression of genes nearest to sites present in mouse and the red line represents the expression of genes nearest to sites absent in mouse. A Mann-Whitney U test was done to test the differences between the two.

(A) Median expression of genes in human (B) Median expression of genes in mouse

Similarly, we see in kidney and liver (Fig.3.8 and Fig.3.9), in human tissue there is little to no difference between the median expression of genes nearest to sites present in mouse and those nearest to sites absent in mouse. However, in mouse, the genes nearest to sites absent in mouse show a higher expression.

In heart tissue (Fig.3.10), the difference in expression is seen in both human and mouse.

This could probably mean that the CTCF binding sites that are absent in mouse are insulator-associated in human. When absent in mouse, they could lead to a higher expression of the genes nearest to them as seen in the above results.

We also looked for patterns in gene expression across CTCF binding sites in human and mouse in adult brain tissue shown in Fig.A1.

3.4 Association of CTCF binding sites with chromatin states

Following this, we see if the chromatin states associated with the present and absent CTCF binding sites show any difference in human. We check for this difference in 4 cell lines and for 3 chromatin states – insulator, enhancer and strong enhancer.

Table A3 contains the number of human CTCF binding sites mapping to each chromatin state.

In Fig.3.11, we see a small difference, in that the sites absent in mouse have a higher percentage of insulator sites as compared to the sites present in mouse.

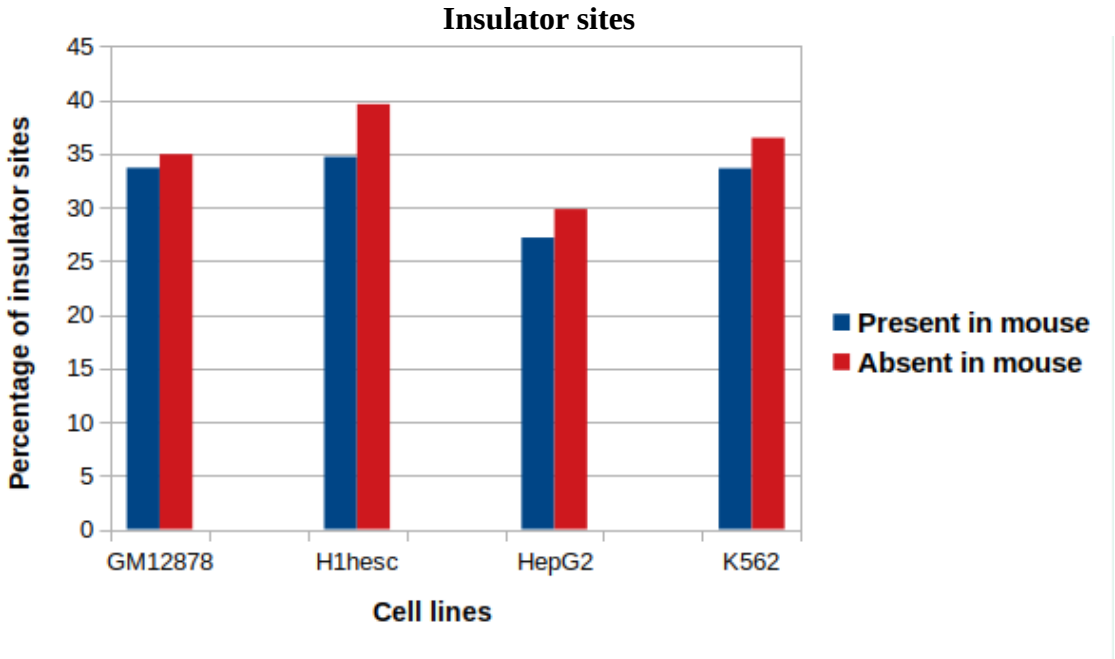

Fig.3.11 Percentage of CTCF binding sites mapping to insulator states

Fig.3.12 and Fig.3.13 show the percentages of enhancer sites. There is almost no difference between the percentage of sites present in mouse and those absent in mouse that map to strong enhancer regions. Similarly, the percentage of weak enhancer sites are almost the same in sites present and absent in mouse.

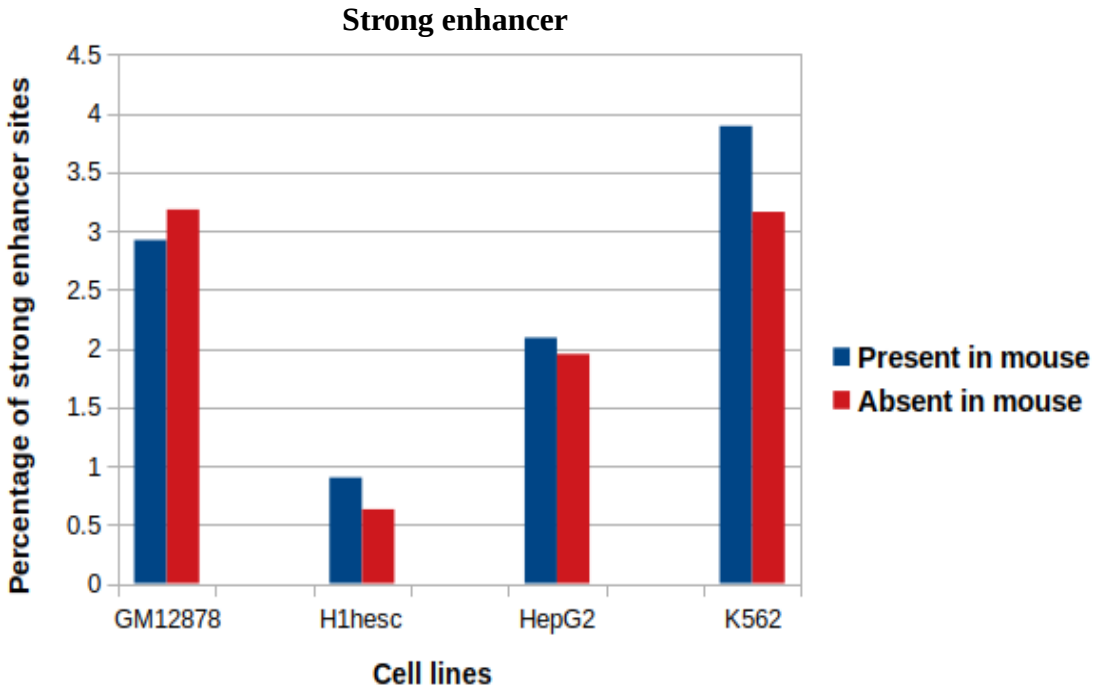

Fig.3.12 Percentage of CTCF binding sites mapping to strong enhancers

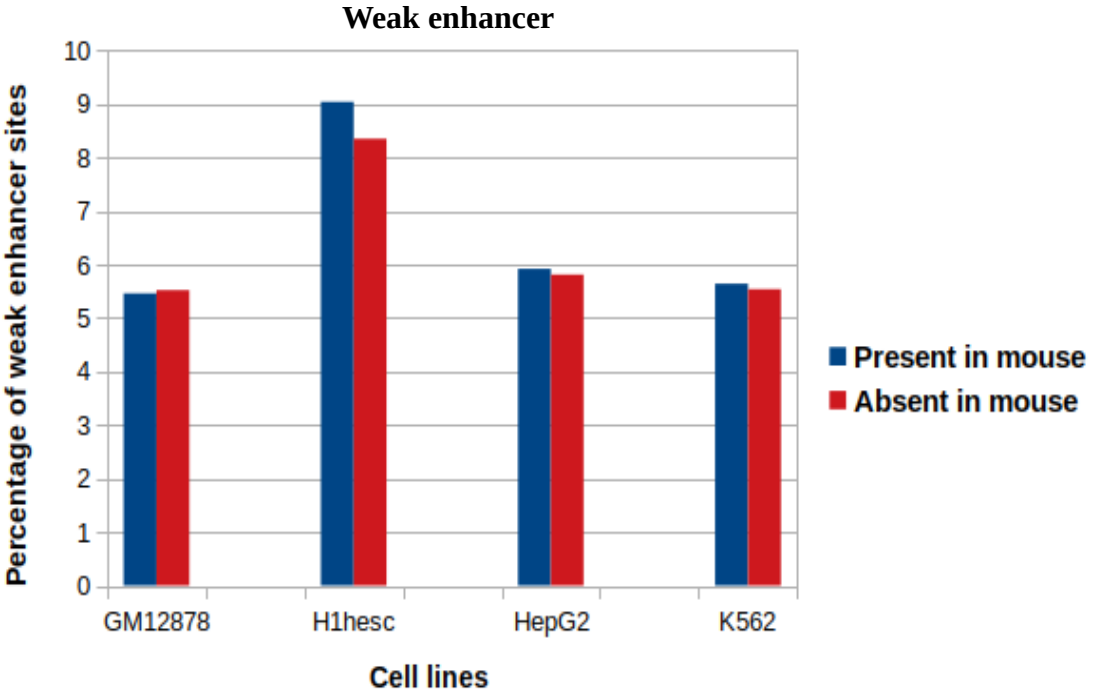

Fig.3.13 Percentage of CTCF binding sites mapping to weak enhancer

3.5 Enrichment of TAD boundaries

Finally, we wanted to see the relation of these sites with the three-dimensional organization of the genome. While CTCF is known to be found at TAD boundaries, a large proportion of these sites are not associated with boundaries (Dixon et al., 2012) and possibly mediate long-range interactions within TADs.

Fig.3.14 shows the density of TAD boundaries around the CTCF binding site. We see that the sites that are present in mouse have a higher number of TAD boundaries around them as compared to sites that are absent in mouse. The sites that are lost in mouse are probably not associated with TAD boundaries and might serve some other function.

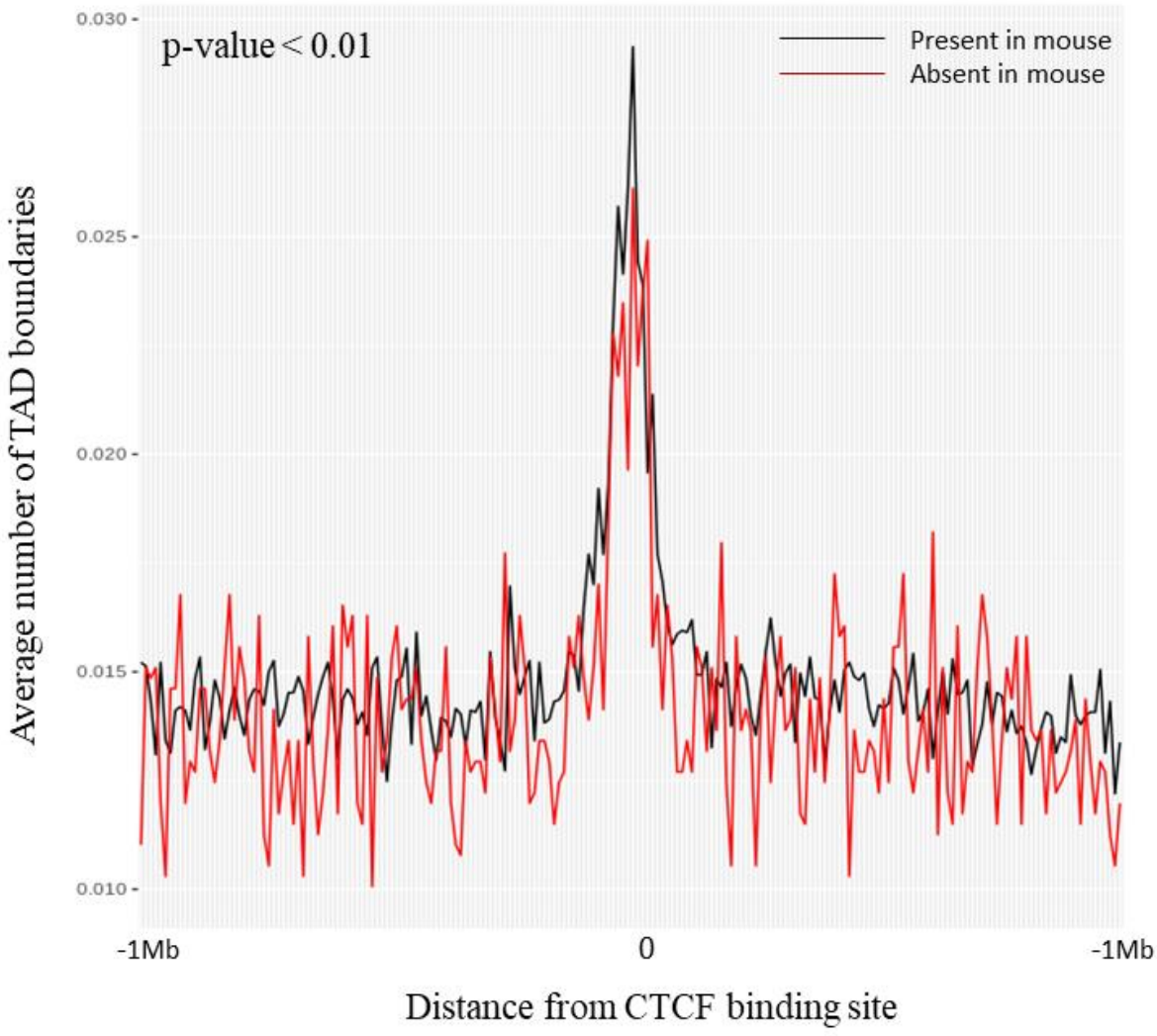

Fig.3.14 Average number of TAD boundaries in 10Kb bins in a 2Mb region (-1Mb/+1Mb from the CTCF binding site). The black line represents the sites present in mouse and the red line represents the sites absent in mouse.

Chapter 4

Discussion

We see from these results that there appears to be a turnover of human CTCF binding sites in other mammalian species and the loss of sites is positively correlated with the divergence time between the species and human. It is expected that with an increase in divergence between species, the conservation of sites will decrease.

The presence or absence of a CTCF binding site does not seem to cause any difference in expression of the gene nearest to it in the adult tissue. These differences might arise in earlier stages of development which would align with the known involvement of CTCF in development. There appears to be a little difference in expression of genes nearest to CTCF binding sites present and absent in mouse across development. In three mouse tissues – forebrain, kidney and liver, we see that the median expression of genes nearest to sites absent in mouse is a little higher than that of genes nearest to sites present in mouse. In contrast to this, the expression of these two sets of genes is nearly the same across development in the human counterparts of these tissues. It appears that the loss of CTCF binding sites in mouse leads to a loss in insulation, however little. The distribution of these sites across different chromatin states does not shed much light on this hypothesis. Sites present and absent in mouse map to enhancer states in nearly equal proportions. While there is some difference in the percentage of present and absent insulator sites – with the sites absent in mouse showing a slightly higher percentage of insulator sites- the difference is too small to make a conclusive claim about this theory. It remains to be checked if there indeed is an effect of the loss of sites on gene expression and if so, the mechanism by which this difference occurs.

From the differences in TAD boundary density around the CTCF binding sites, we see that the sites present in mouse have a greater number of TAD boundaries around them as compared to the sites absent in mouse. This seems likely as large scale TADs have been found to be conserved in different species. The sites that are lost in mouse are probably involved in the establishment of more local chromatin loops and interactions, rather than in large scale genome organization.

A deeper analysis of these sites and the region around them is required to ascertain their functional significance, if any. More analysis of the expression across CTCF binding sites in different stages of development could provide more useful information.

A similar analysis with different species and comparative analyses of the sites with respect to lineage specific traits would also help in understanding the evolutionary dynamics of CTCF binding.

It could also be extended to the study of position effect of these sites, by studying the sites present in different species and their position relative to particular genes. A difference in gene expression could occur not only by the presence or absence of a CTCF binding site, but also by a change in the nature (for instance, the chromatin state) of the site. This kind of evolutionary analysis could also complement molecular studies investigating the function of CTCF.

References

1. Ambrosini, G., Dreos, R., & Bucher, P. (2014, April 6). *Principles of ChIP-seq Data Analysis Illustrated with Examples*. https://doi.org/10.13140/2.1.4608.4807
2. Bastian, F., Parmentier, G., Roux, J., Moretti, S., Laudet, V., & Robinson-Rechavi, M. (2008). Bgee: Integrating and Comparing Heterogeneous Transcriptome Data Among Species. In A. Bairoch, S. Cohen-Boulakia, & C. Froidevaux (Eds.), *Data Integration in the Life Sciences* (pp. 124–131). Springer. https://doi.org/10.1007/978-3-540-69828-9_12
3. Bell, A. C., & Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. *Nature*, *405*(6785), 482–485. https://doi.org/10.1038/35013100
4. Bell, A. C., West, A. G., & Felsenfeld, G. (1999). The Protein CTCF Is Required for the Enhancer Blocking Activity of Vertebrate Insulators. *Cell*, *98*(3), 387–396. https://doi.org/10.1016/S0092-8674(00)81967-4
5. Cremer, T., & Cremer, M. (2010). Chromosome territories. *Cold Spring Harbor Perspectives in Biology*, *2*(3), a003889. https://doi.org/10.1101/cshperspect.a003889
6. Cremer, T., Cremer, M., Dietzel, S., Müller, S., Solovei, I., & Fakan, S. (2006). Chromosome territories—A functional nuclear landscape. *Current Opinion in Cell Biology*, *18*(3), 307–316. https://doi.org/10.1016/j.ceb.2006.04.007
7. Cunningham, F., Achuthan, P., Akanni, W., Allen, J., Amode, M. R., Armean, I. M., Bennett, R., Bhai, J., Billis, K., Boddu, S., Cummins, C., Davidson, C., Dodiya, K. J., Gall, A., Girón, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., … Flicek, P. (2019). Ensembl 2019. *Nucleic Acids Research*, *47*(D1), D745–D751. https://doi.org/10.1093/nar/gky1113
8. Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M.

(2018). The Encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801. https://doi.org/10.1093/nar/gkx1081

9. Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing Chromosome Conformation. *Science*, *295*(5558), 1306–1311. https://doi.org/10.1126/science.1067799
10. de Wit, E., Vos, E. S. M., Holwerda, S. J. B., Valdes-Quezada, C., Verstegen, M. J. A. M., Teunissen, H., Splinter, E., Wijchers, P. J., Krijger, P. H. L., & de Laat, W. (2015). CTCF Binding Polarity Determines Chromatin Looping. *Molecular Cell*, *60*(4), 676–684. https://doi.org/10.1016/j.molcel.2015.09.023
11. Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–380. https://doi.org/10.1038/nature11082
12. Filippova, G. N., Fagerlie, S., Klenova, E. M., Myers, C., Dehner, Y., Goodwin, G., Neiman, P. E., Collins, S. J., & Lobanenkov, V. V. (1996). An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes. *Molecular and Cellular Biology*, *16*(6), 2802–2813. https://doi.org/10.1128/mcb.16.6.2802
13. Fornes, O., Castro-Mondragon, J. A., Khan, A., van der Lee, R., Zhang, X., Richmond, P. A., Modi, B. P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., & Mathelier, A. (2019). JASPAR 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, gkz1001. https://doi.org/10.1093/nar/gkz1001
14. Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, *15*(9), 2038–2049. https://doi.org/10.1016/j.celrep.2016.04.085
15. Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, and D. (2002). The Human Genome Browser at UCSC. *Genome Research*, *12*(6), 996–1006. https://doi.org/10.1101/gr.229102

16. Kumar, S., Stecher, G., Suleski, M., & Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution*, *34*(7), 1812–1819. https://doi.org/10.1093/molbev/msx116
17. Lieberman-Aiden, E., Berkum, N. L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369
18. Lobanenkov, V. V., Nicolas, R. H., Adler, V. V., Paterson, H., Klenova, E. M., Polotskaja, A. V., & Goodwin, G. H. (1990). A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5'-flanking sequence of the chicken c-myc gene. *Oncogene*, *5*(12), 1743.
19. Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., … Mundlos, S. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, *161*(5), 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004
20. Narendra, V., Rocha, P. P., An, D., Raviram, R., Skok, J. A., Mazzoni, E. O., & Reinberg, D. (2015). CTCF establishes discrete functional chromatin domains at the Hox clusters during differentiation. *Science*, *347*(6225), 1017–1021. https://doi.org/10.1126/science.1262088
21. NCBI Resource Coordinators. (2018). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *46*(D1), D8–D13. https://doi.org/10.1093/nar/gkx1095
22. Nora, E. P., Goloborodko, A., Valton, A.-L., Gibcus, J. H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L. A., & Bruneau, B. G. (2017). Targeted Degradation of CTCF

Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, *169*(5), 930-944.e22. https://doi.org/10.1016/j.cell.2017.05.004

23. Papatheodorou, I., Fonseca, N. A., Keays, M., Tang, Y. A., Barrera, E., Bazant, W., Burke, M., Füllgrabe, A., Fuentes, A. M.-P., George, N., Huerta, L., Koskinen, S., Mohammed, S., Geniza, M., Preece, J., Jaiswal, P., Jarnuczak, A. F., Huber, W., Stegle, O., … Petryszak, R. (2018). Expression Atlas: Gene and protein expression across multiple studies and organisms. *Nucleic Acids Research*, *46*(D1), D246–D251. https://doi.org/10.1093/nar/gkx1158
24. Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021
25. Ruiz-Velasco, M., Kumar, M., Lai, M. C., Bhat, P., Solis-Pinson, A. B., Reyes, A., Kleinsorg, S., Noh, K.-M., Gibson, T. J., & Zaugg, J. B. (2017). CTCF-Mediated Chromatin Loops between Promoter and Gene Body Regulate Alternative Splicing across Individuals. *Cell Systems*, *5*(6), 628-637.e6. https://doi.org/10.1016/j.cels.2017.10.018
26. Sivakumar, A., de las Heras, J. I., & Schirmer, E. C. (2019). Spatial Genome Organization: From Development to Disease. *Frontiers in Cell and Developmental Biology*, *7*. https://doi.org/10.3389/fcell.2019.00018
27. Turatsinze, J.-V., Thomas-Chollier, M., Defrance, M., & van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nature Protocols*, *3*(10), 1578–1588. https://doi.org/10.1038/nprot.2008.97
28. Vostrov, A. A., & Quitschke, W. W. (1997). The zinc finger protein CTCF binds to the APBbeta domain of the amyloid beta-protein precursor promoter. Evidence for a role in transcriptional activation. *The Journal of Biological Chemistry*, *272*(52), 33353–33359. https://doi.org/10.1074/jbc.272.52.33353

Appendix

Table A1 Assembly information of mammalian genomes

Species	Species (common name)	Assembly	Assembly level	Scaffold N50 (kb)
Papio anubis	Baboon	papAnu2	Chromosme	528.927
Otolemur garnetti	Bushbaby	otoGar1	Scaffold	101.35
Pan troglodytes	Chimp	panTro6	Chromosme	53103.722
Macaca fascicularis	Crab-eating macaque	macFas5	Chromosme	88649.475
Nomascus leucogenys	Gibbon	nomLeu3	Chromosme	52956.88
Chlorocebus sabaeus	Green monkey	chlSab1	Chromosme	81825.804
Callithrix jacchus	Marmoset	calJac3	Chromosme	5167.444
Microcebus murinus	Mouse lemur	micMur1	Scaffold	140.884
Pongo abelii	Orangutan	ponAbe3	Chromosme	98475.126
Macaca mulatta	Rhesus macaque	rheMac10	Chromosme	82346.004
Saimiri boliviensis boliviensis	Squirrel monkey	saiBol1	Scaffold	18744.88
Tarsius syrichta	Tarsier	tarSyr1	Scaffold	12.214
Cricetulus griseus	Chinese hamster	criGri1	Scaffold	1558.295
Cavia porcellus	Guinea pig	cavPor3	Scaffold	27942.054
Dipodomys ordii	Kangaroo rat	dipOrd1	Scaffold	36.427
Mus musculus	Mouse	mm10	Chromosme	54517.951
Oryctolagus cuniculus	Rabbit	oryCun2	Chromosme	35972.871
Rattus norvegicus	Rat	rn6	Chromosme	14986.627
Spermophilus tridecemlineatus	Squirrel	speTri2	Scaffold	8192.786
Vicugna pacos	Alpaca	vicPac1	Scaffold	230.521
Bos taurus	Cow	bosTau7	Chromosme	2599.288
Tursiops truncatus	Dolphin	turTru1	Scaffold	166.056
Balaenoptera acutorostrata	Minke whale	balAcu1	Scaffold	12843.668
Sus scrofa	Pig	susScr11	Chromosme	88231.837
Ovis aries	Sheep	oviAri3	Chromosme	100079.507
Equus caballus	Horse	equCab2	Chromosme	46749.9
Ceratotherium simum simum	White rhino	cerSim1	Scaffold	26277.727
Felis catus	Cat	fetCat5	Chromosme	4658.941
Canis lupus	Dog	canfam3	Chromosme	45876.61
Ailuropoda melanoleuca	Panda	ailMel1	Scaffold	1281.781
Dasypus novemcinctus	Armadillo	dasNov3	Scaffold	46.559
Loxodonta africana	Elephant	loxAfr3	Scaffold	46401.353
Erinaceus europaeus	Hedgehog	eriEur2	Scaffold	3264.618
Pteropus vampyrus	Megabat	pteVam1	Scaffold	124.06
Myotis lucifugus	Microbat	myoLuc2	Scaffold	4293.315
Ochotona princeps	Pika	ochPri3	Scaffold	26863.993
Procavia capensis	Rock hyrax	proCap1	Scaffold	24.297
Sorex araneus	Shrew	sorAra2	Scaffold	22794.405
Choloepus hoffmanni	Sloth	choHof1	Scaffold	9.667

Echinops telfairi	Tenrec	echTel2	Scaffold	45764.842
Tupaia belangeri	Tree shrew	tupBel1	Scaffold	88.86
Ornithorhynchus anatinus	Platypus	ornAna1	Chromosme	991.605
Macropus eugenii	Wallaby	macEug1	Scaffold	36.602
Sarcophilus harrisii	Tasmanian devil	sarHar1	Scaffold	1847.106
Monodelphis domestica	Opossum	monDom5	Chromosome	59809.81

Scaffold N50 is the length of scaffold such that the 50% of the genome comprises of scaffolds of this length or longer.

Table A2 Summary of lift over of human CTCF binding sites to other mammalian genomes

Species	Divergence time from human (mya)	Number of sites lifted over	Number of sites lost
Chimp	6.7	41605	691
Orangutan	15.76	41278	1018
Rhesus macaque	29.44	40478	1818
Green monkey	29.44	40241	2055
Crab-eating macaque	29.44	39820	2476
Baboon	29.44	39810	2486
Gibbon	20.19	38138	4158
Marmoset	43.2	36661	5635
White rhino	96	36635	5661
Pig	96	34864	7432
Squirrel monkey	43.2	34838	7458
Minke whale	96	33935	8361
Cow	96	33820	8476
Horse	96	33756	8540
Dog	96	33522	8774
Squirrel	90	33262	9034
Elephant	105	33022	9274
Panda	96	32787	9509
Cat	96	32669	9627
Sheep	96	31335	10961
Guinea pig	90	30723	11573
Mouse	90	30563	11733
Rabbit	90	30046	12250
Rat	90	29292	13004
Microbat	96	28420	13876
Pika	90	28369	13927
Chinese hamster	90	27305	14991
Hedgehog	96	24940	17356
Tenrec	105	24648	17648
Shrew	96	22202	20094
Opossum	159	19334	22962
Platypus	177	13685	28611
Tasmanian devil	159	9990	32306

Total number of CTCF binding sites obtained in human = 42296

Table A3 Number of human CTCF binding sites in different chromatin states

Cell line	CTCF binding sites	Number of insulator sites	Number of strong enhancer sites	Number of weak enhancer sites
GM12878	Present in mouse	8737	758	1418
	Absent in mouse	1557	142	246
H1hesc	Present in mouse	9000	235	2345
	Absent in mouse	1765	28	372
HepG2	Present in mouse	7050	542	1535
	Absent in mouse	1329	87	259
K562	Present in mouse	8723	1010	1464
	Absent in mouse	1625	141	247

Number of CTCF binding sites present in mouse = 25946
Number of CTCF binding sites absent in mouse = 4456

Expression across CTCF binding sites

Fig.A1 shows a heatmap of expression, in adult brain tissue, across the CTCF binding sites absent in mouse. We see the expression of genes in a -1Mb/+1Mb region centred around the CTCF binding sites in human and the expression of the orthologous genes in mouse. The plot does not seem to show any pattern indicating a difference in expression. This analysis when modified and adopted for other tissues and developmental stages could be useful in a more robust study of the effects of these sites on gene expression.

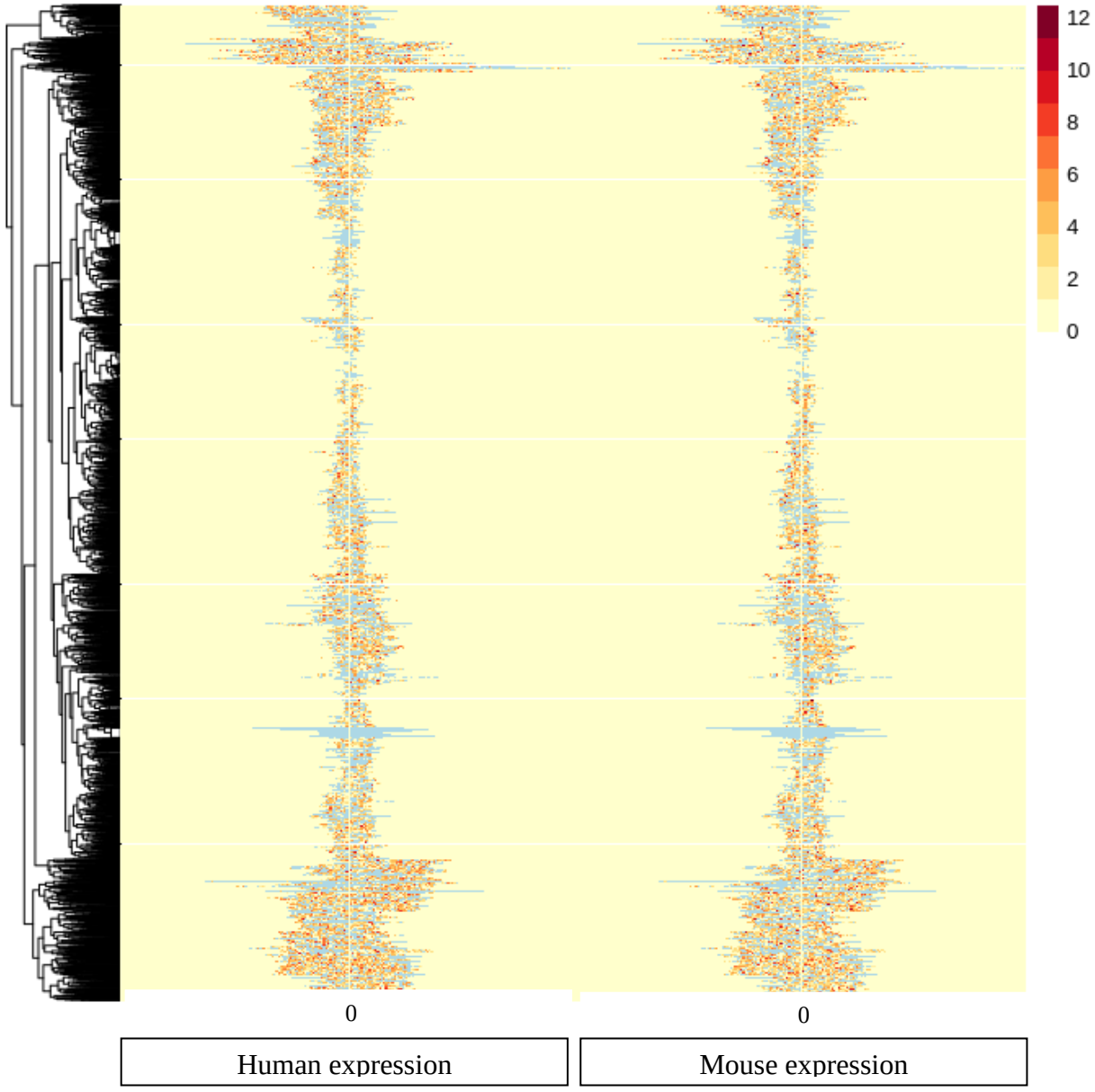

Fig. A1 Heat map showing expression across CTCF binding sites (-1Mb/+1Mb from the site) absent in mouse in adult brain tissue. The expression values are in log2 scale.